A few Thoughts on Graphs in Chemistry and Biology

Peter Schuster

19th LL-Seminar on Graph Theory

ÖAW, 25.04.2002

Graphs are seen as valuable tools to order and classify information in various scientific disciplines at an **intermediate stage** of knowledge or level of approximation. Such stages are, for example,

- collection or harvesting of data,

- **ordering of data according to new categories and development of models for qualitative analysis**

- development of model for quatitative analysis and accurate predictions.

Graphs are considered here as tools to

- distiguish chemical isomers,
- describe the flux in chemical reaction networks,

- define biological species by their phylogenetic descent, and
- model genotype-phenotype maps in case of neutrality.

Chemists use graphs to distinguish isomers since the second half of the ninteenth century. Atoms are nodes and chemical bonds are edges. In case of hydrocarbons containing exclusively carbon and hydrogen atoms the position of the atom is sufficient to predict its nature:

H atoms form one bond and are attached to one edge, whereas **C** atoms form always four bonds and are connected to four edges.

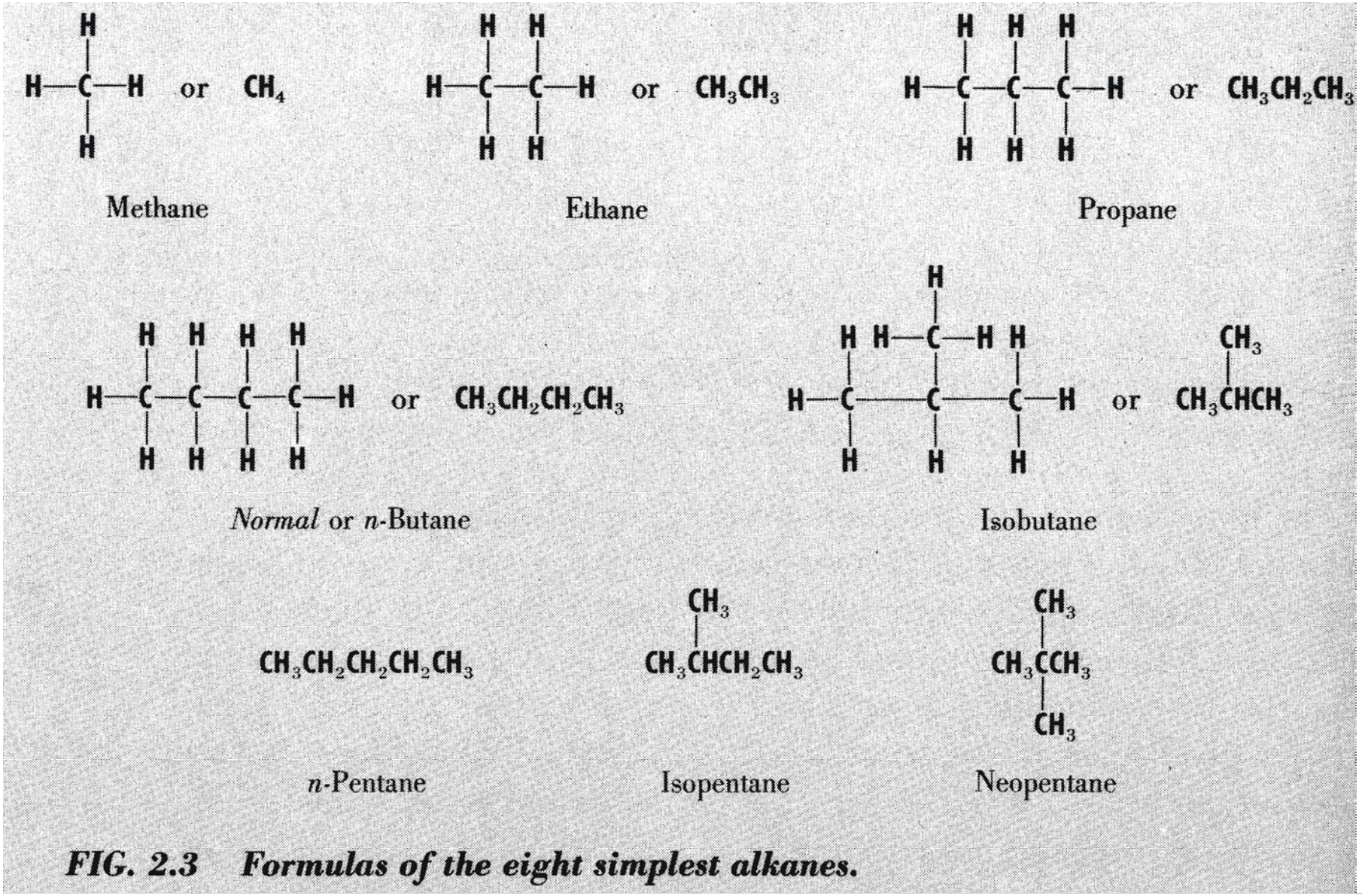
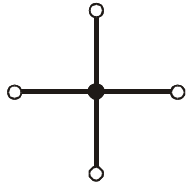


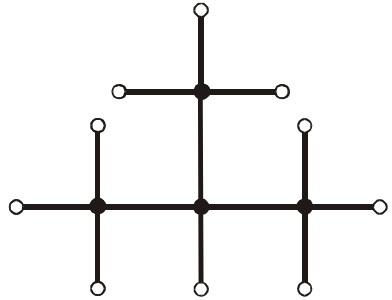
FIG. 2.3 Formulas of the eight simplest alkanes.

D.J.Cram and G.S.Hammond, *Organic Chemistry*, McGraw-Hill, New York 1959, p.18

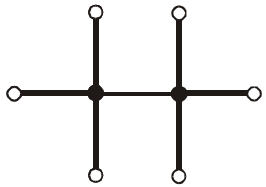
C_nH_{2n+2}, n = 1,2,3,4,5



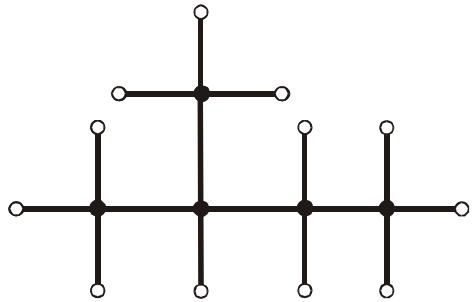
methane



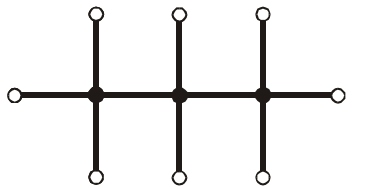
isobutane



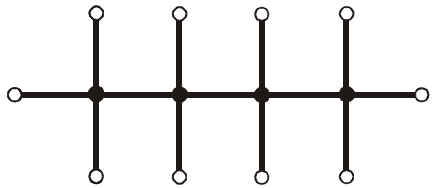
ethane



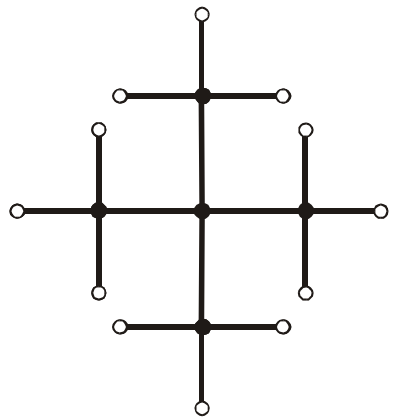
isopentane



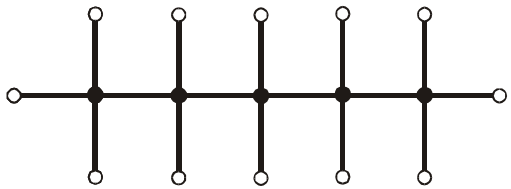
propane



n-butane



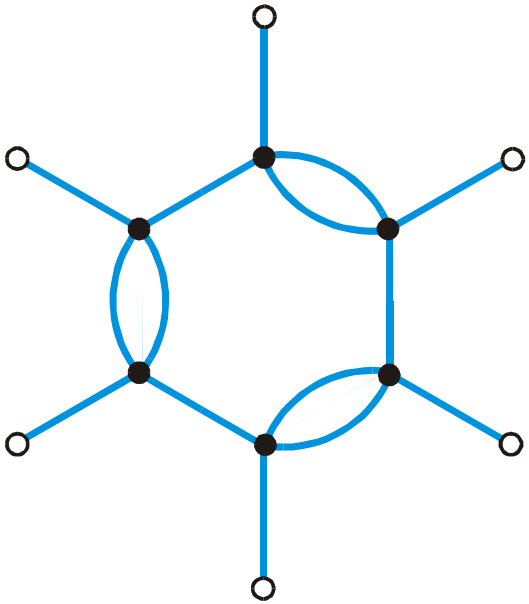
neopentane



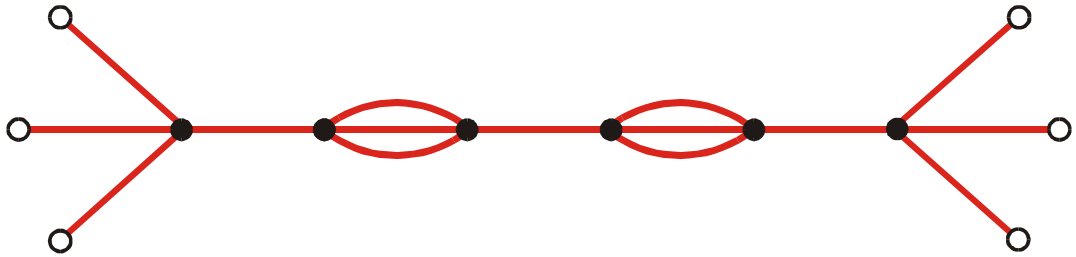
n-pentane

Formulas of the eight simplest alkanes as graphs, which allow for the distinction of isomers, e.g. n- and isobutane, n-, iso- and neo-pentane

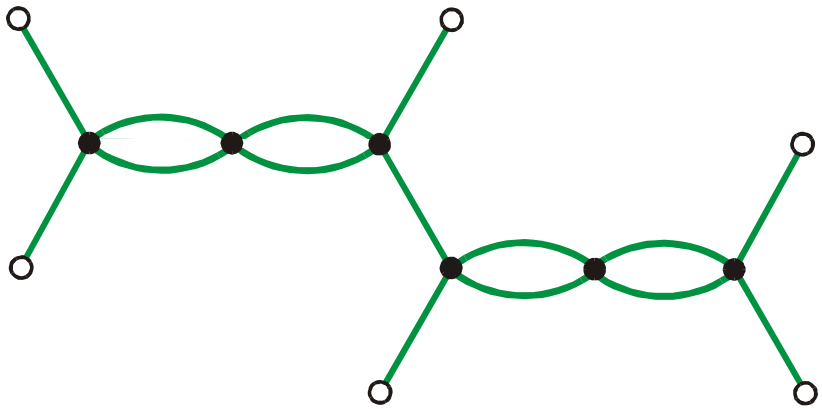
C_6H_6



benzene



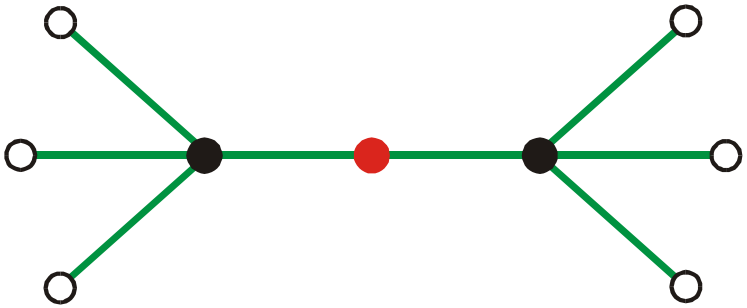
hexa-2,4-diyne (dimethyl-diacetylene)



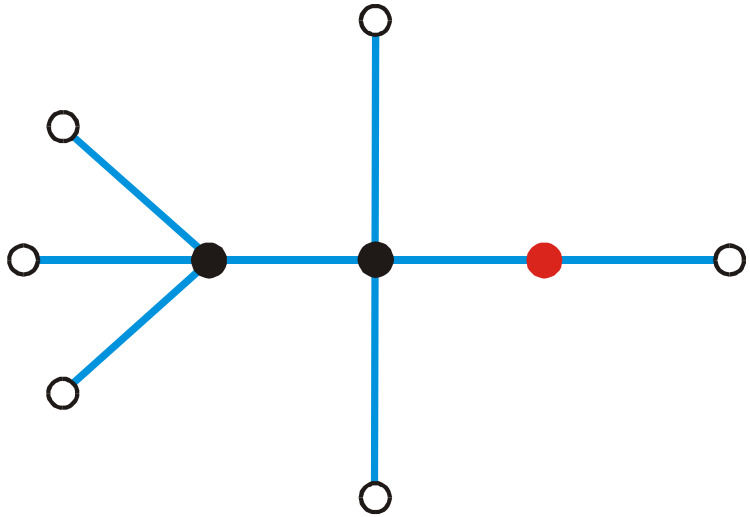
hexa-1,2,4,5-tetraene (diallene)

Graphs allow for a distinction of single-, double- and triple bonds

C_2H_6O



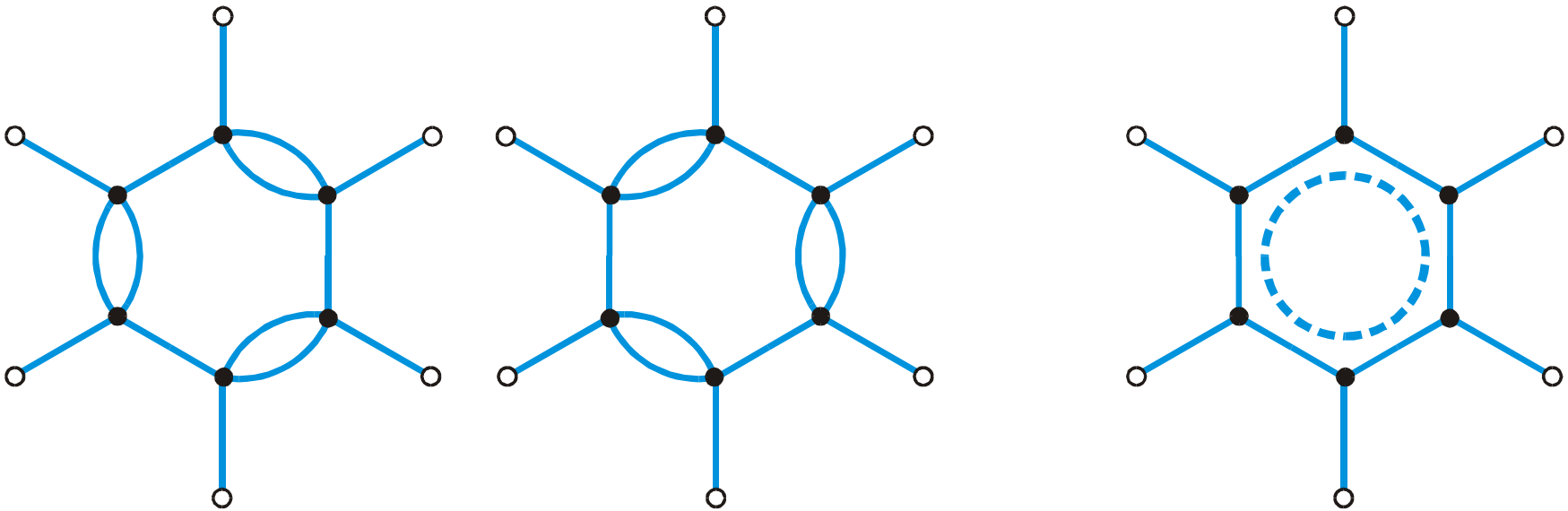
dimethylether



ethanol

Carbon, hydrogen and oxygen atoms are distinguished by the degree of the corresponding nodes: d(H) = 1, d(O) = 2, and d(C) = 4.

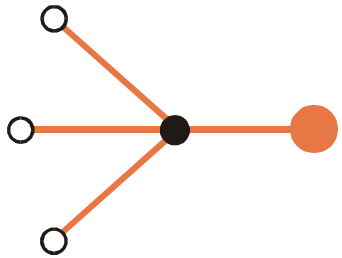
C_6H_6



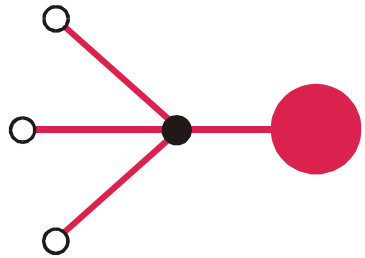
benzene

The benzene molecule cannot be described by a single graph.

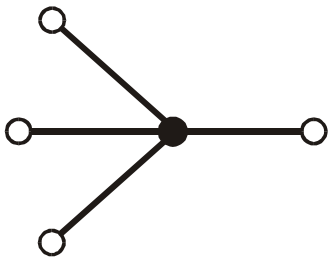
CH_3X



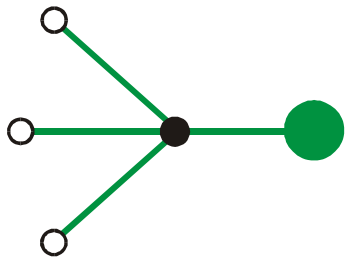
methyl fluoride: X = F



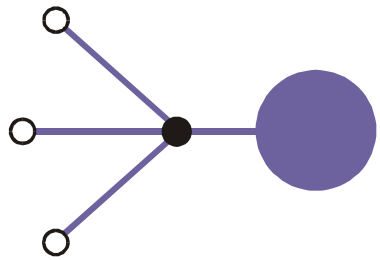
methyl bromide: X = Br



methane: X = H



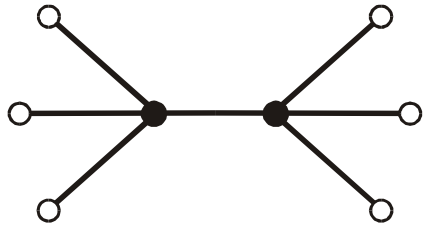
methyl chloride: X = Cl



methyl iodide: X = I

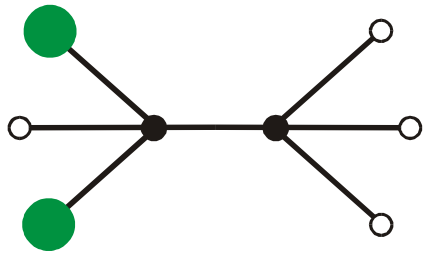
Different atoms forming one bond: H, F, Cl, Br, and I

C_2H_6

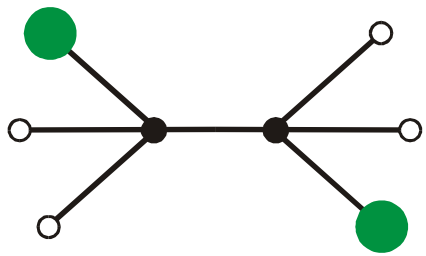


ethane

$C_2H_4Cl_2$

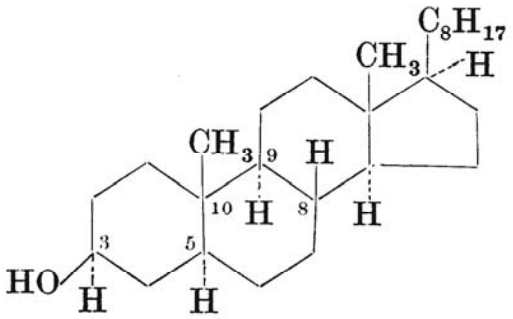


1,1-dichloro ethane



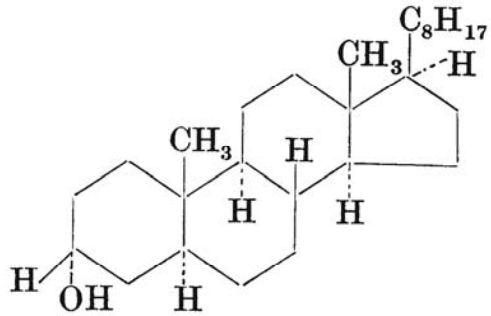
1,2-dichloro ethane

Two isomers that cannot be distinguished by means of their graphs.



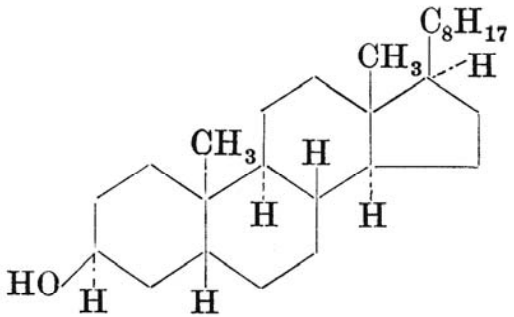
Dihydrocholesterin
Cholestanol

3β OH, 5α H, 10β CH_3, 9α H, 8β H,
14α H, 13β CH_3, 17β Seitenkette,
A/B trans, B/C trans, C/D trans



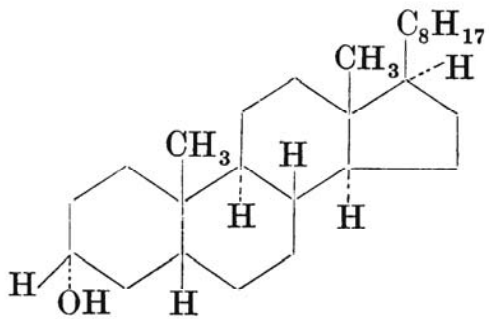
Epidihydrocholesterin
Epicholestanol

3α OH, 5α H, 10β CH_3, 9α H, 8β H,
14α H, 13β CH_3, 17β Seitenkette
A/B trans, B/C trans, C/D trans



Koprosterin
Koprostanol

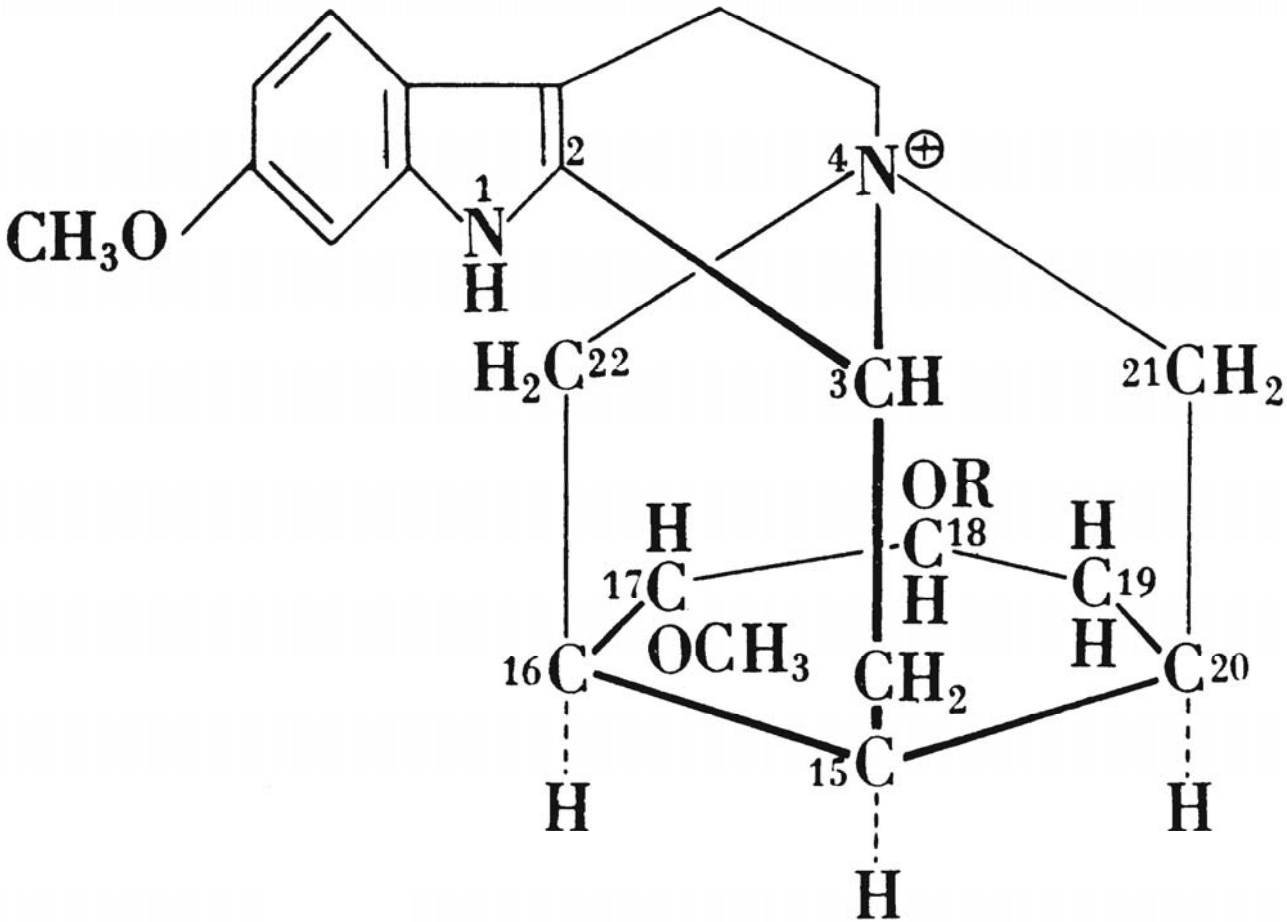
3β OH, 5β H, 10β CH_3, 9α H, 8β H,
14α H, 13β CH_3, 17β Seitenkette
A/B cis, B/C trans, C/D trans



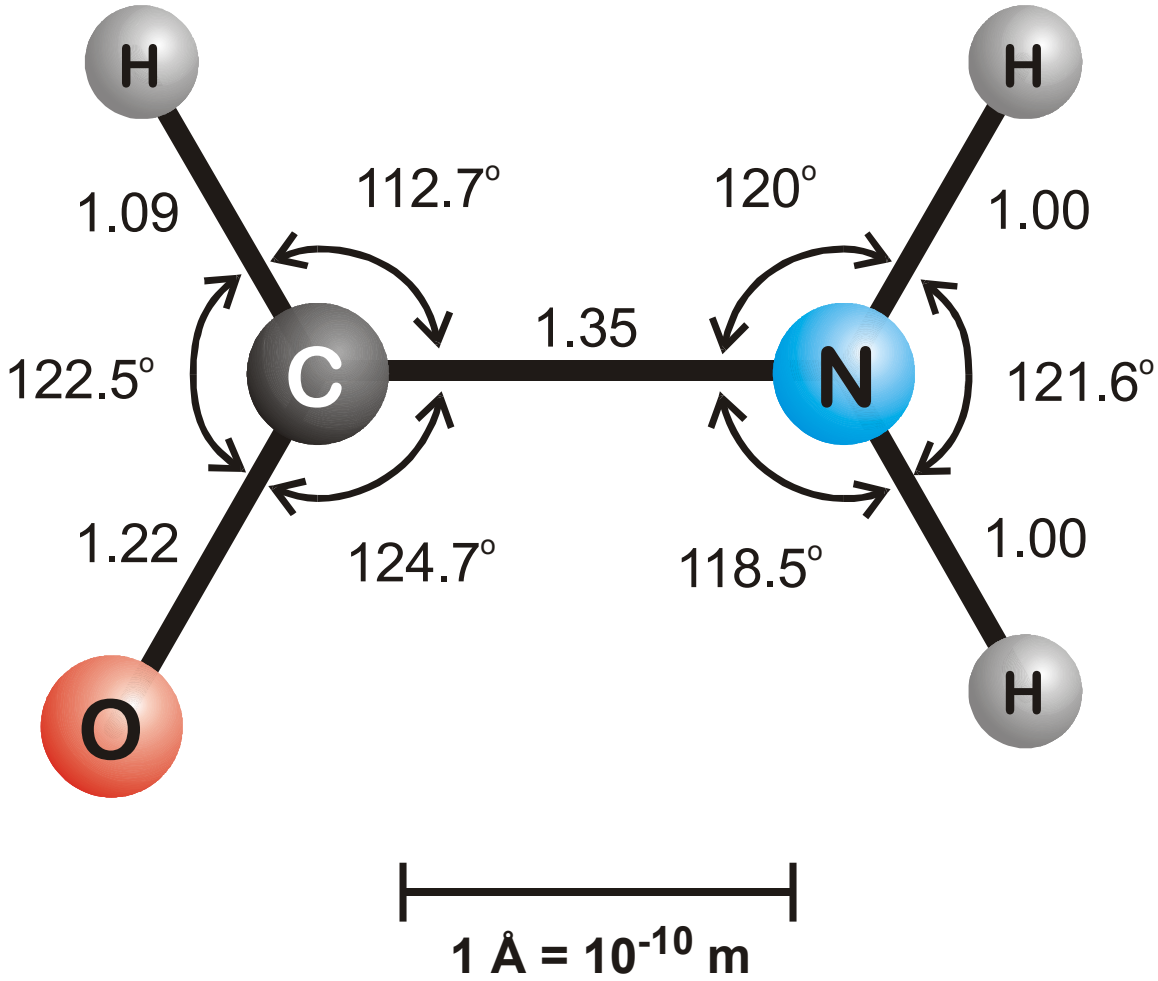
Epikoprosterin
Epikoprostanol

3α OH, 5β H, 10β CH_3, 9α H, 8β H,
14α H, 13β CH_3, 17β Seitenkette
A/B cis, B/C trans, C/D trans

Paul Karrer, Lehrbuch der organischen Chemie,
Georg Thieme Verlag, Stuttgart 1959, p.737

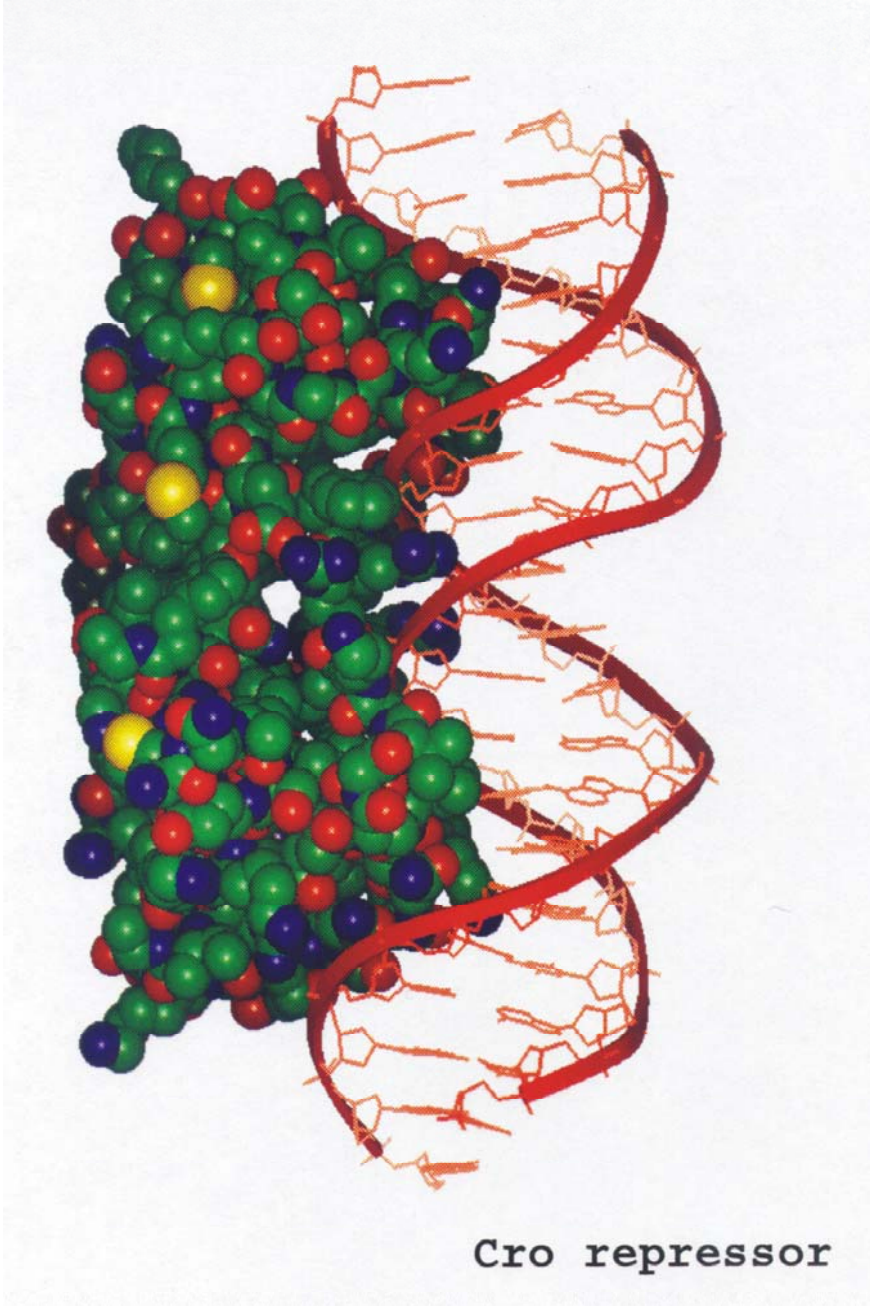


Paul Karrer, Lehrbuch der organischen Chemie, Georg Thieme Verlag, Stuttgart 1959, p.949



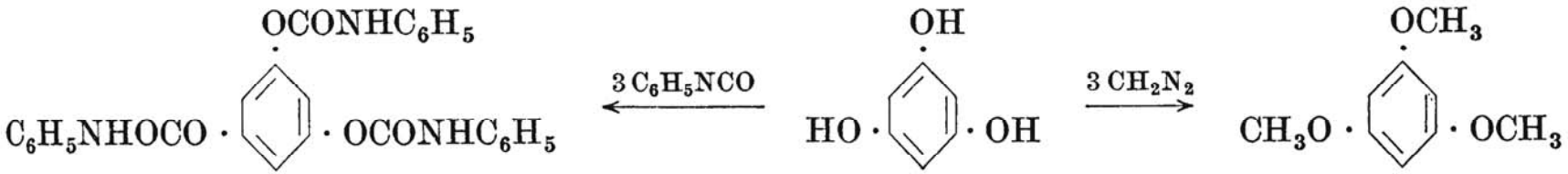
Molecular structure of the formamide molecule

Molecular structure of an association complex between a protein an a nucleic acid



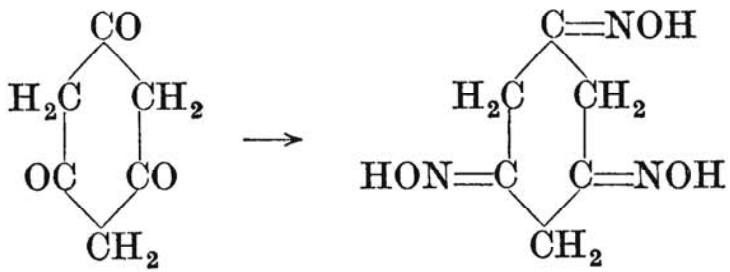
Chemists use directed graphs to model reaction mechanisms in chemical kinetics.

Als Phenol reagiert Phloroglucin z. B. mit Diazomethan und mit Phenylisocyanat, wobei O-Derivate gebildet werden:

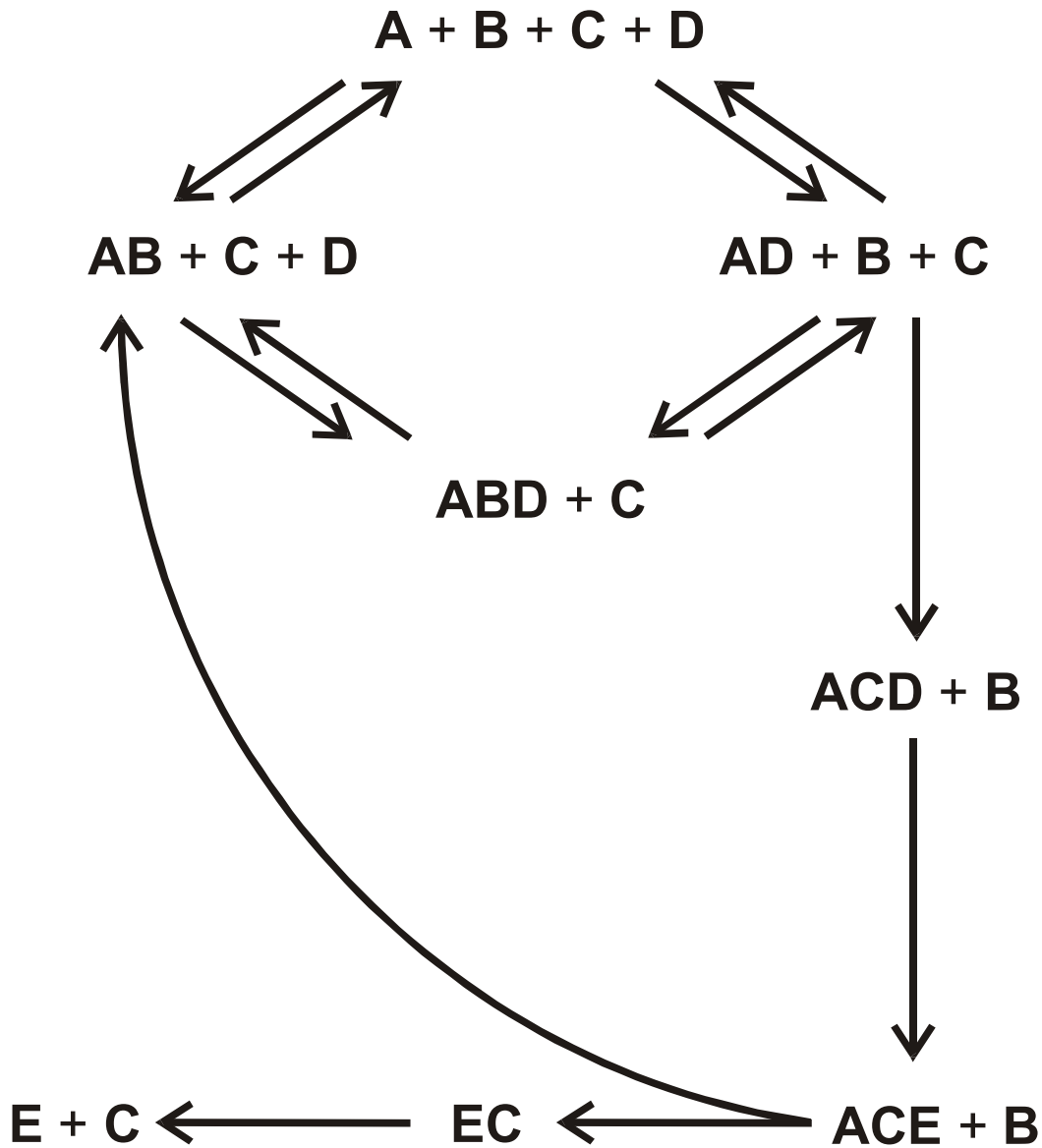


Phloroglucin-trimethyläther.

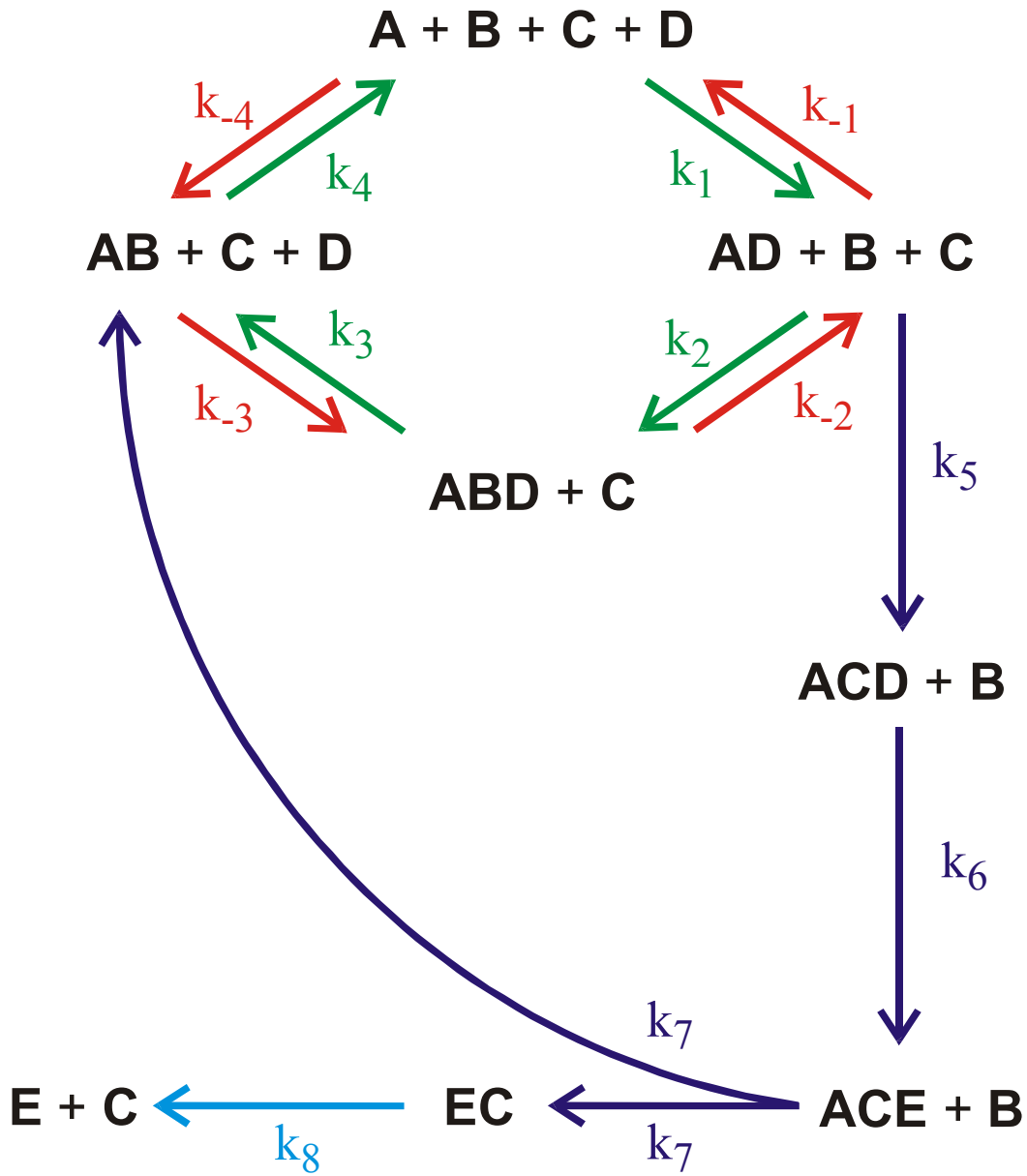
Als Keton setzt es sich mit Hydroxylamin um; es entsteht ein Trioxim:



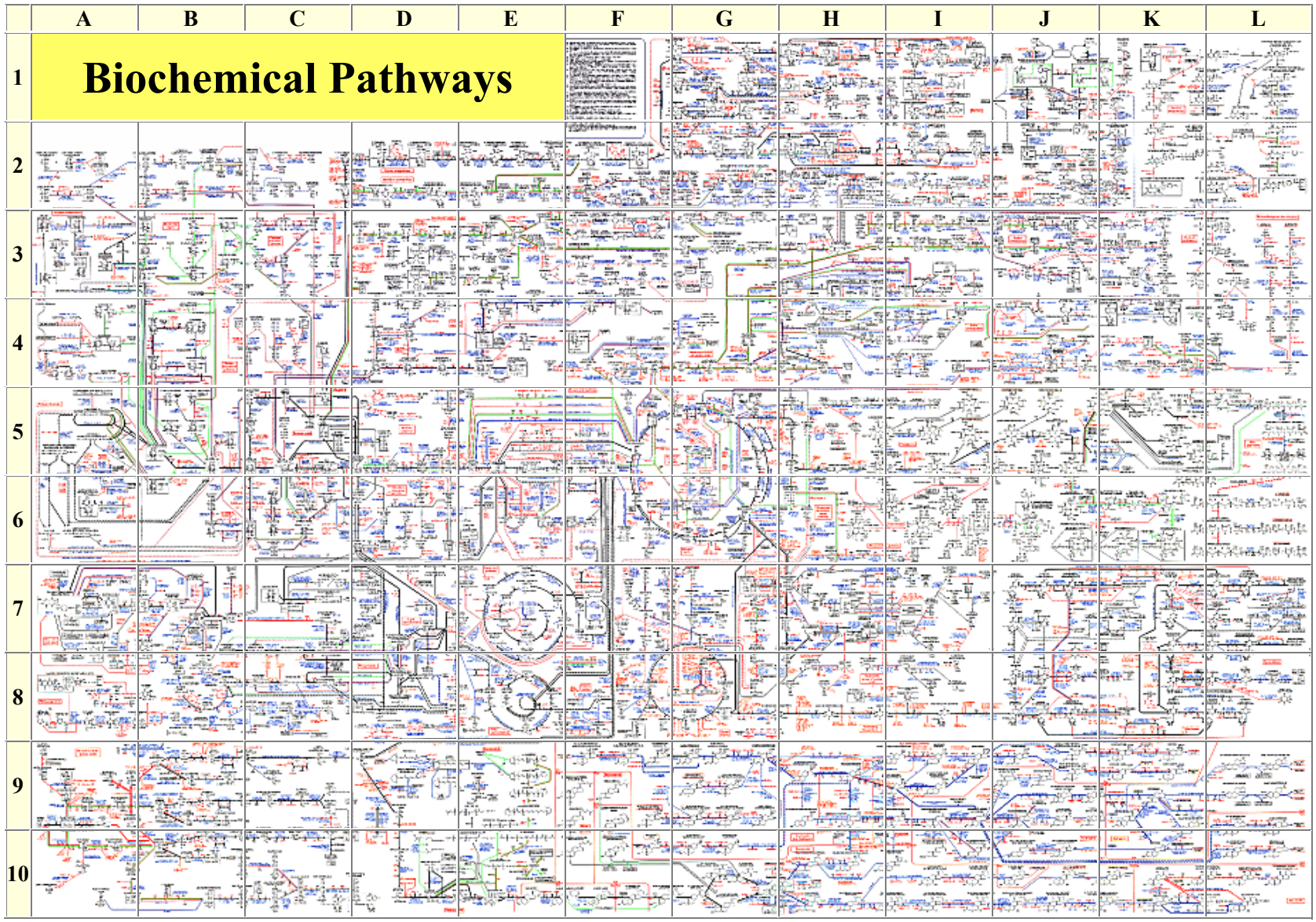
Paul Karrer, Lehrbuch der organischen Chemie, Georg Thieme Verlag, Stuttgart 1959, p.479



Reaction graph of a kinetic mechanism

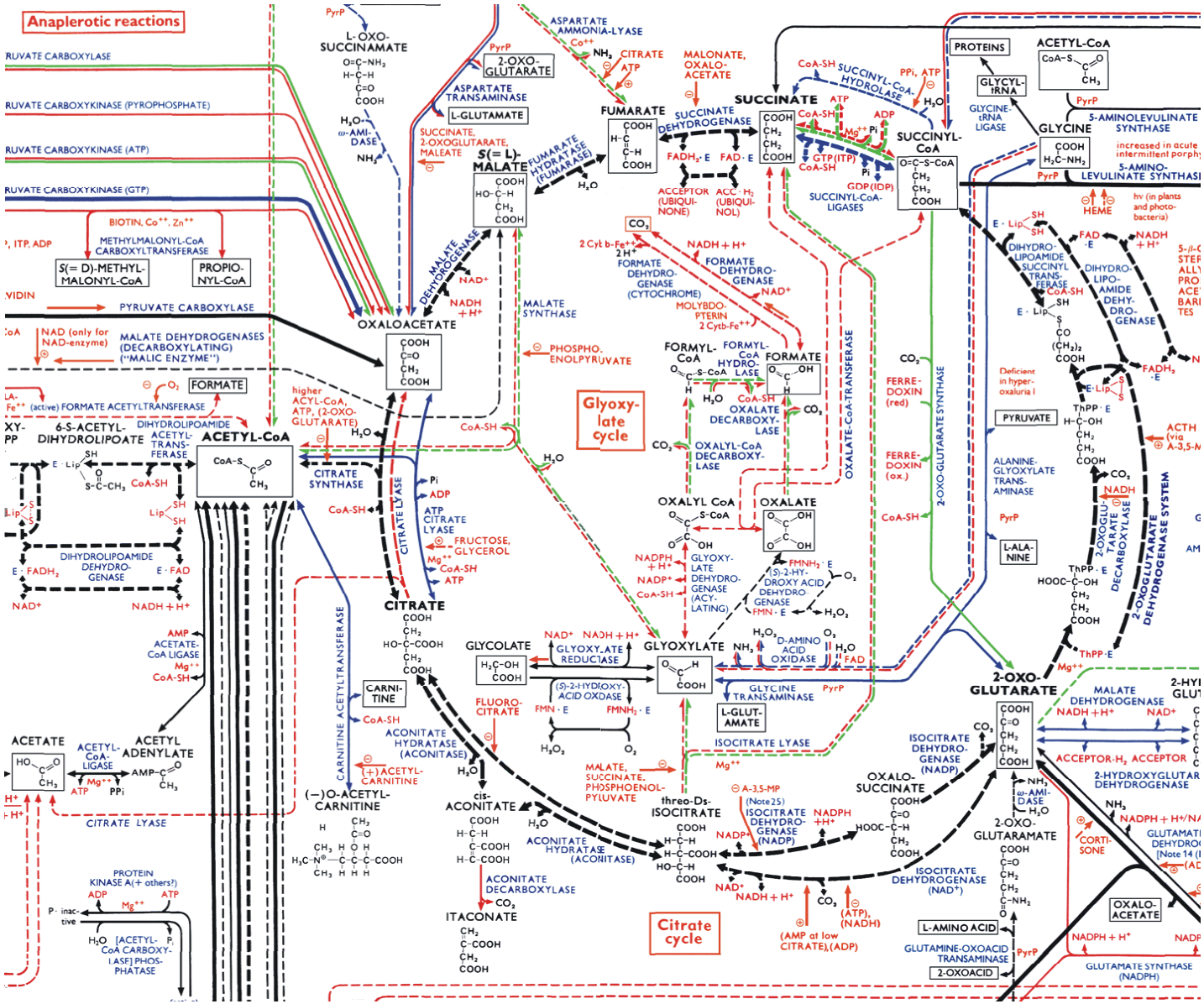


Reaction graph of a kinetic mechanism with rate constants



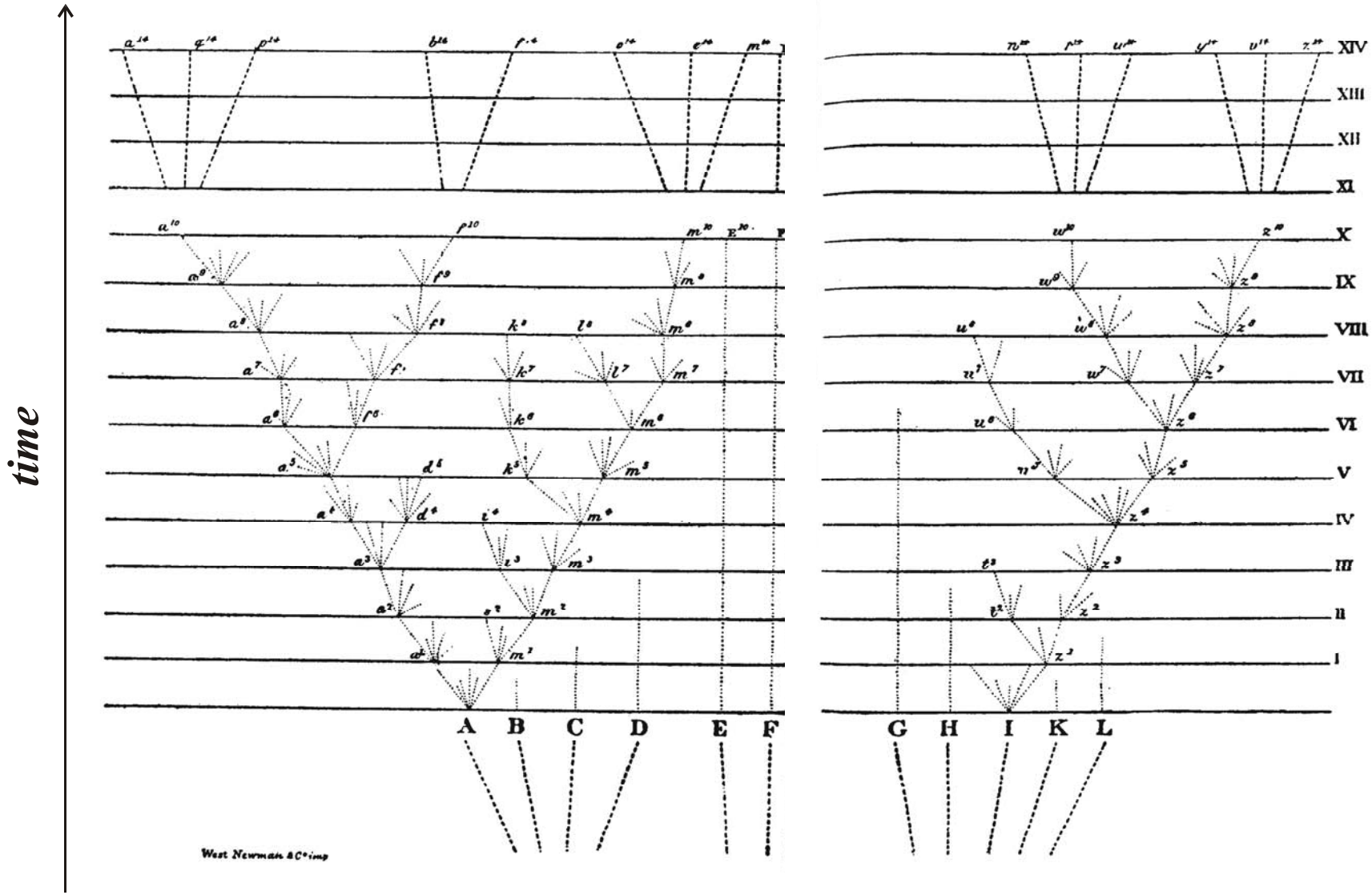
The reaction network of cellular metabolism published by Boehringer-Ingelheim.

The citric acid or Krebs cycle (enlarged from previous slide).

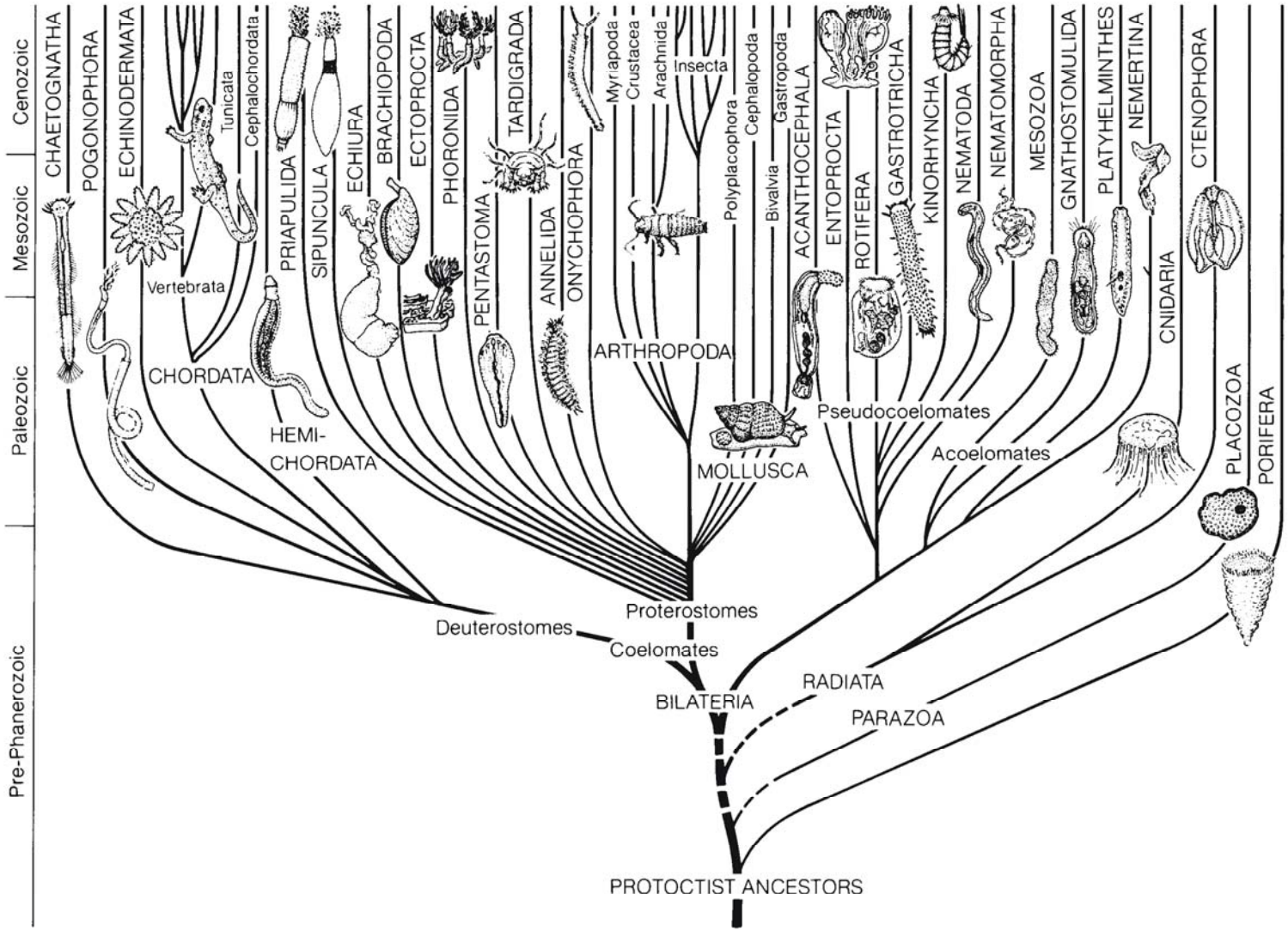


Biologists use directed graphs in the form of trees to distinguish biological species by their descent. The concept of evolution allows for ordering the wealth of species by means of phylogenetic relation.

Direction of development and **time** ordering is introduced by the fossil record.

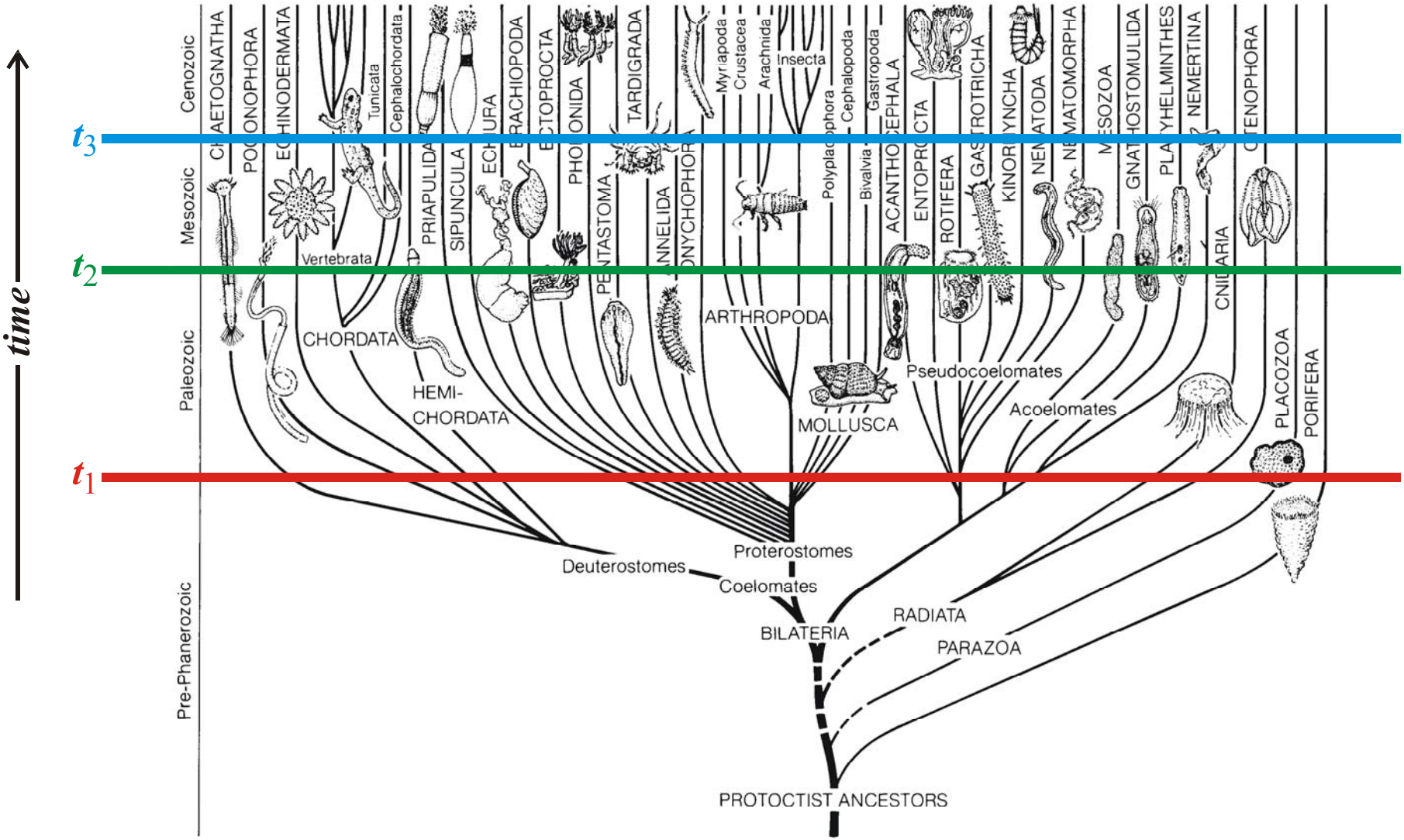


Charles Darwin, *The Origin of Species*, 6th edition.
Everyman‘s Library, Vol.811, Dent London, pp.121-122.



Phylogenetic tree of animal kingdom

Lynn Margulis & Karlene V. Schwarz, *Five Kingdoms. An illustrated guide to the Phyla of Life on Earth*. W.H. Freeman & Co., San Francisco, 1982, p. 160.



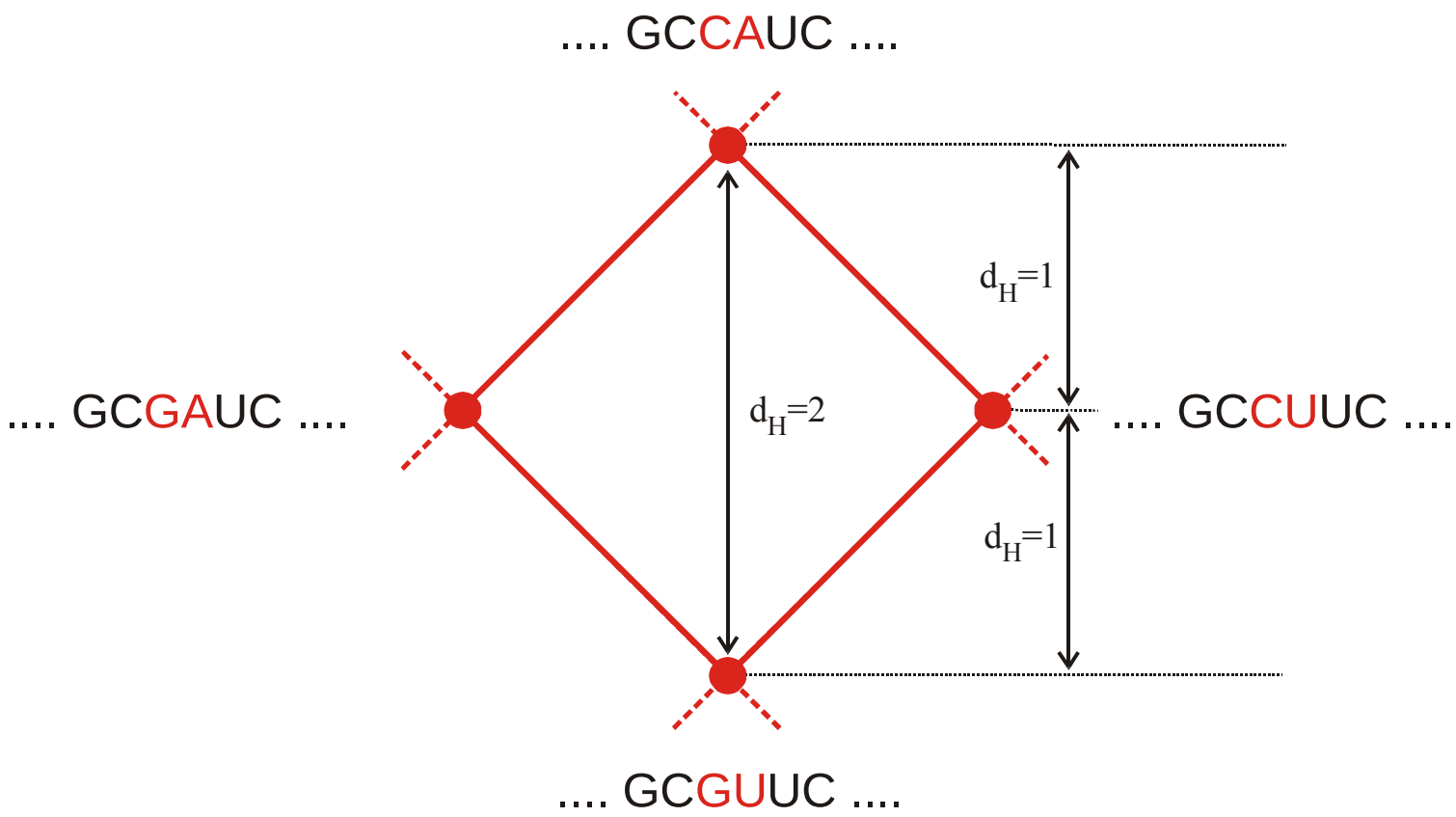
Phylogenetic tree of animal kingdom

Lynn Margulis & Karlene V. Schwarz, *Five Kingdoms. An illustrated guide to the Phyla of Life on Earth*. W.H. Freeman & Co., San Francisco, 1982, p. 160.

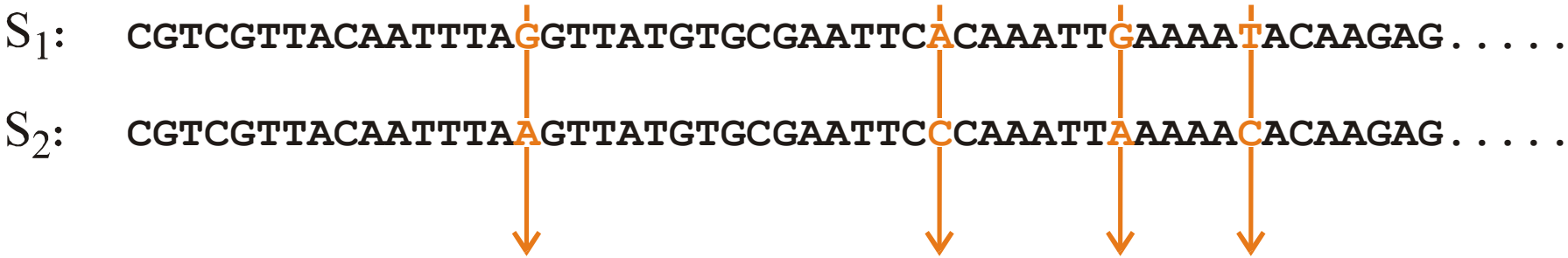
The **genotypes** or **genomes** of individuals and species, being reproductively related ensembles of individuals, are DNA sequences. They are changing from generation to generation through mutation and recombination.

Genotypes unfold into **phenotypes** or organisms, which are the targets of the evolutionary selection process.

Point mutations are single nucleotide exchanges. The **Hamming distance** of two sequences is the minimal number of single nucleotide exchanges that mutually converts the two sequence into each other.



Point mutations as moves in sequence space



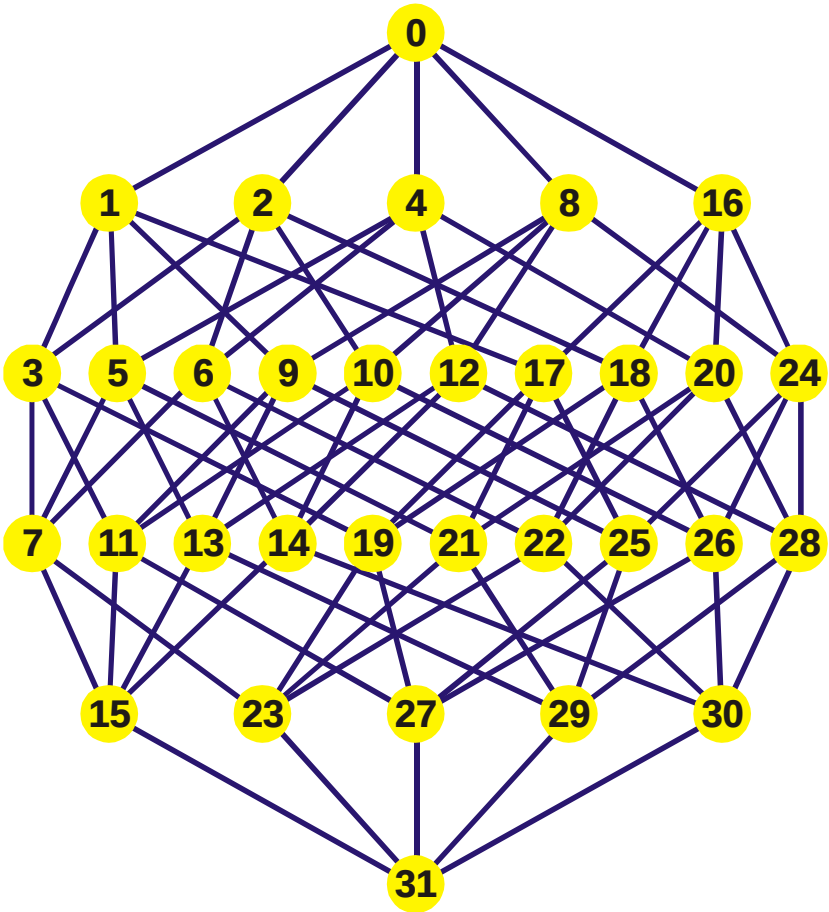
Hamming distance $d_H(S_1,S_2) = 4$

(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 induces a metric in sequence space



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,

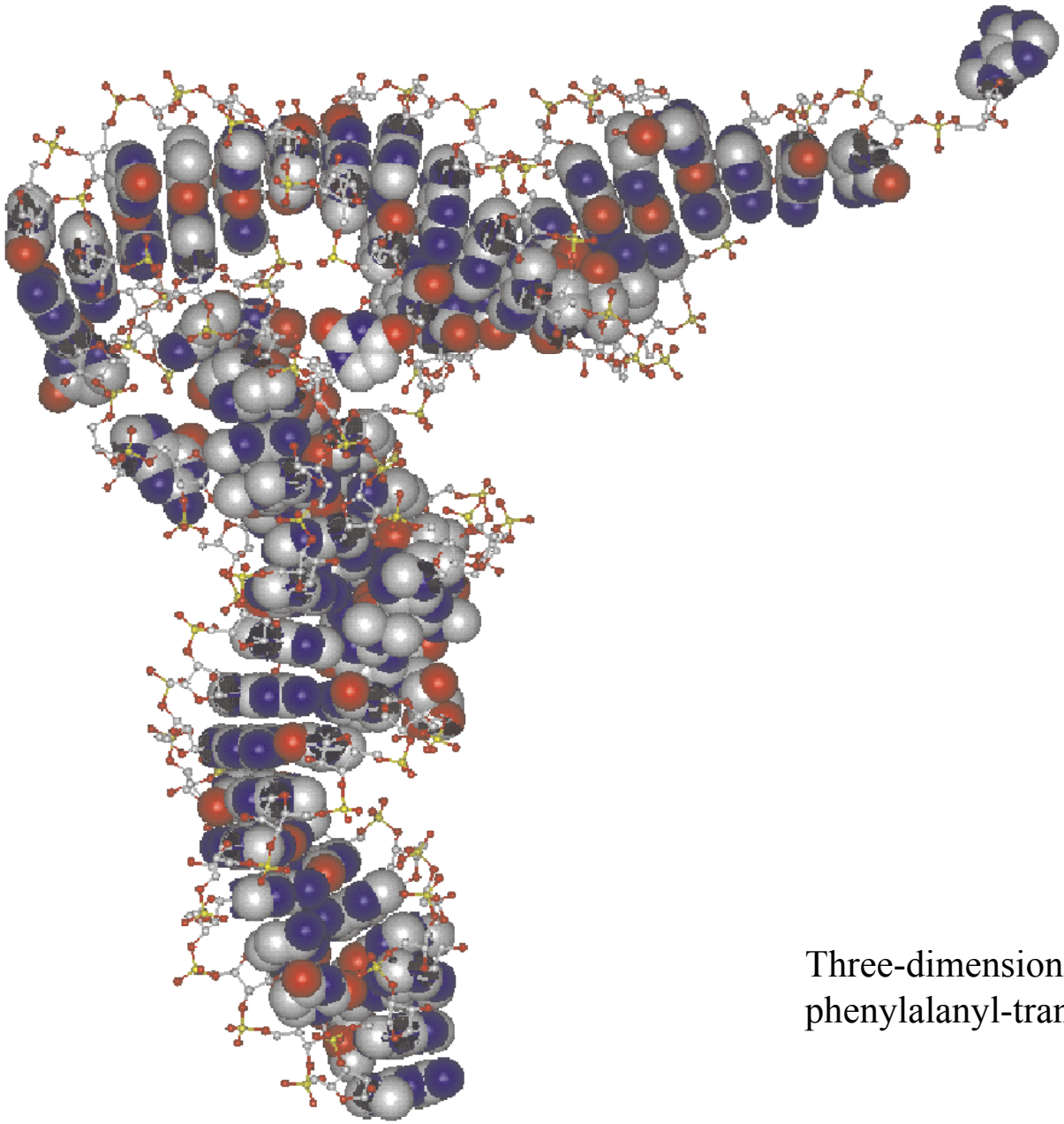
"29" ≡ 11101 = GGGCG, etc.

Sequence space of binary sequences of chain lenght n=5

The **RNA model** considers RNA sequences as genotypes and simplified RNA structures, called secondary structures, as phenotypes.

The **mapping** from genotypes into phenotypes is many-to-one. Hence, it is redundant and not invertible.

Genotypes, i.e. RNA sequences, which are mapped onto the same phenotype, i.e. the same RNA secondary structure, form **neutral networks**. Neutral networks are represented by graphs in sequence space.

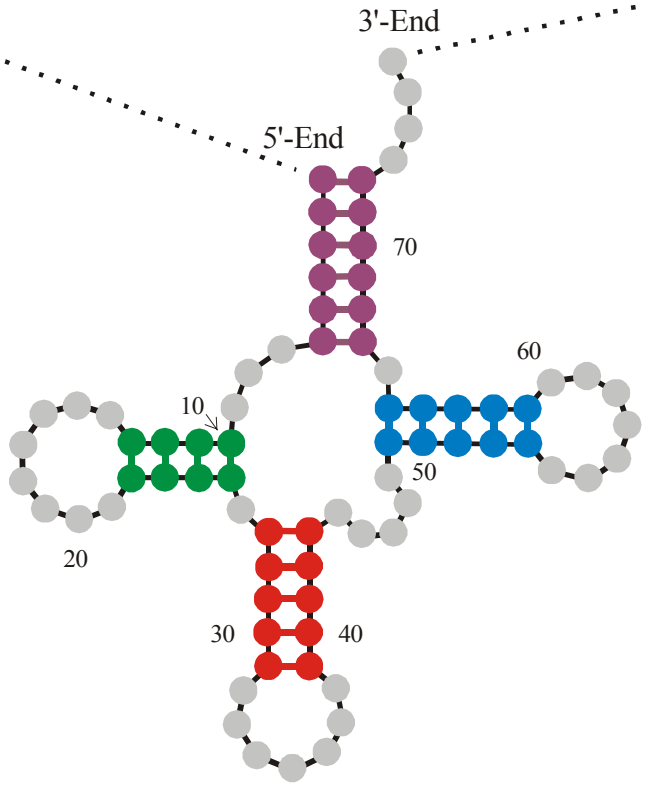


Three-dimensional structure of
phenylalanyl-transfer-RNA

Sequence

5'-End GCGGAUUUAGCUCAGDDGGGAGAGCMCCAGACUGAAYAUCUGGAGMUCCUGUGTPCGAUCCACAGAAUUCGCACCA 3'-End

Secondary structure



Symbolic notation

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

Definition and formation of the secondary structure of phenylalanyl-tRNA

Criterion of
Minimum Free Energy

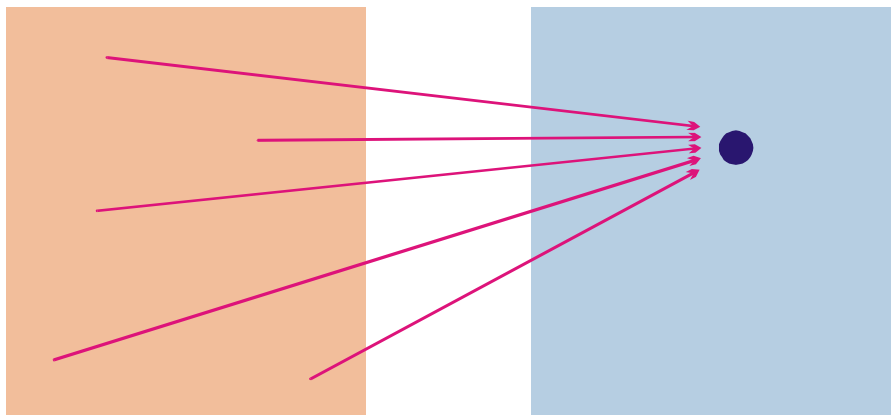
UUUAGCCAGCGCGAGUCGUGCGGACGGGGUUAUCUCUGUCGGGCUAGGGCGC

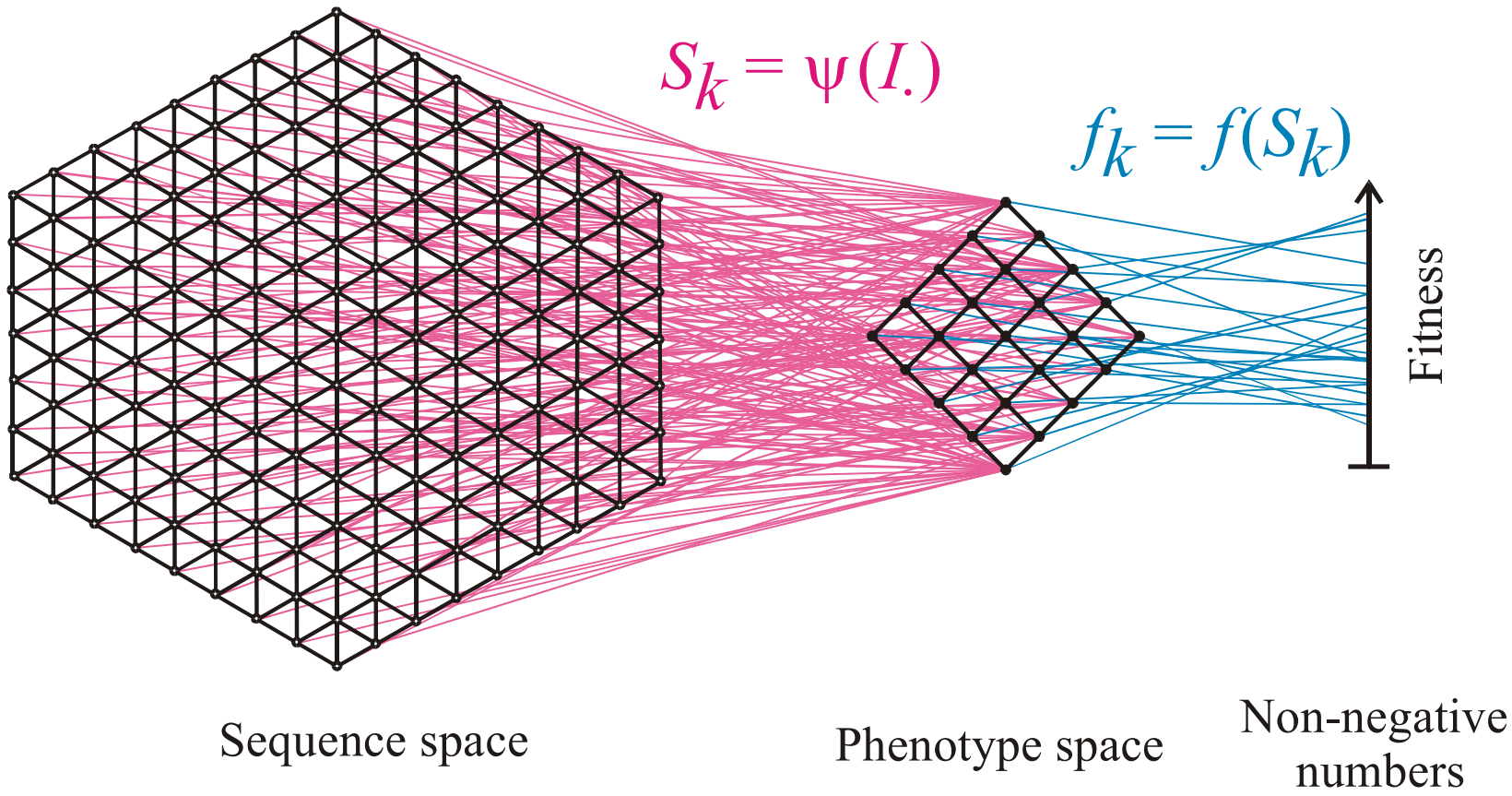
GUGAGCGCGGGGCACAGUUUCUCAAGGAUGUAAGUUUUUGCCGUUUAUCUGG

UUAGCGAGAGAGGAGGCUUCUAGACCCAGCUCUCUGGGUCGUUGCUGAUGCG

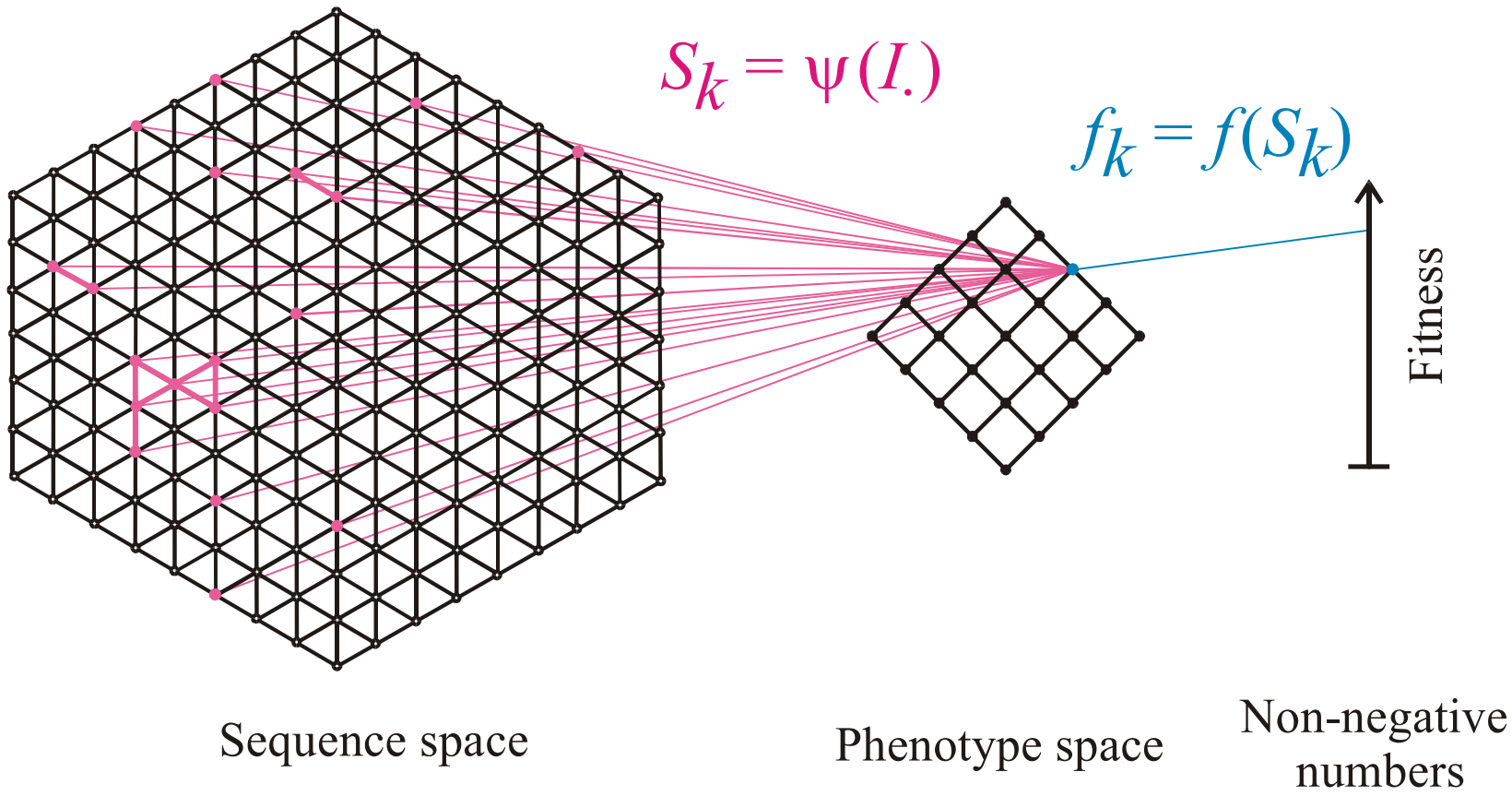
CAUUGGUGCUAAUGAUAUUAGGGCUGUAUUCCUGUAUAGCGAUCAGUGUCCG

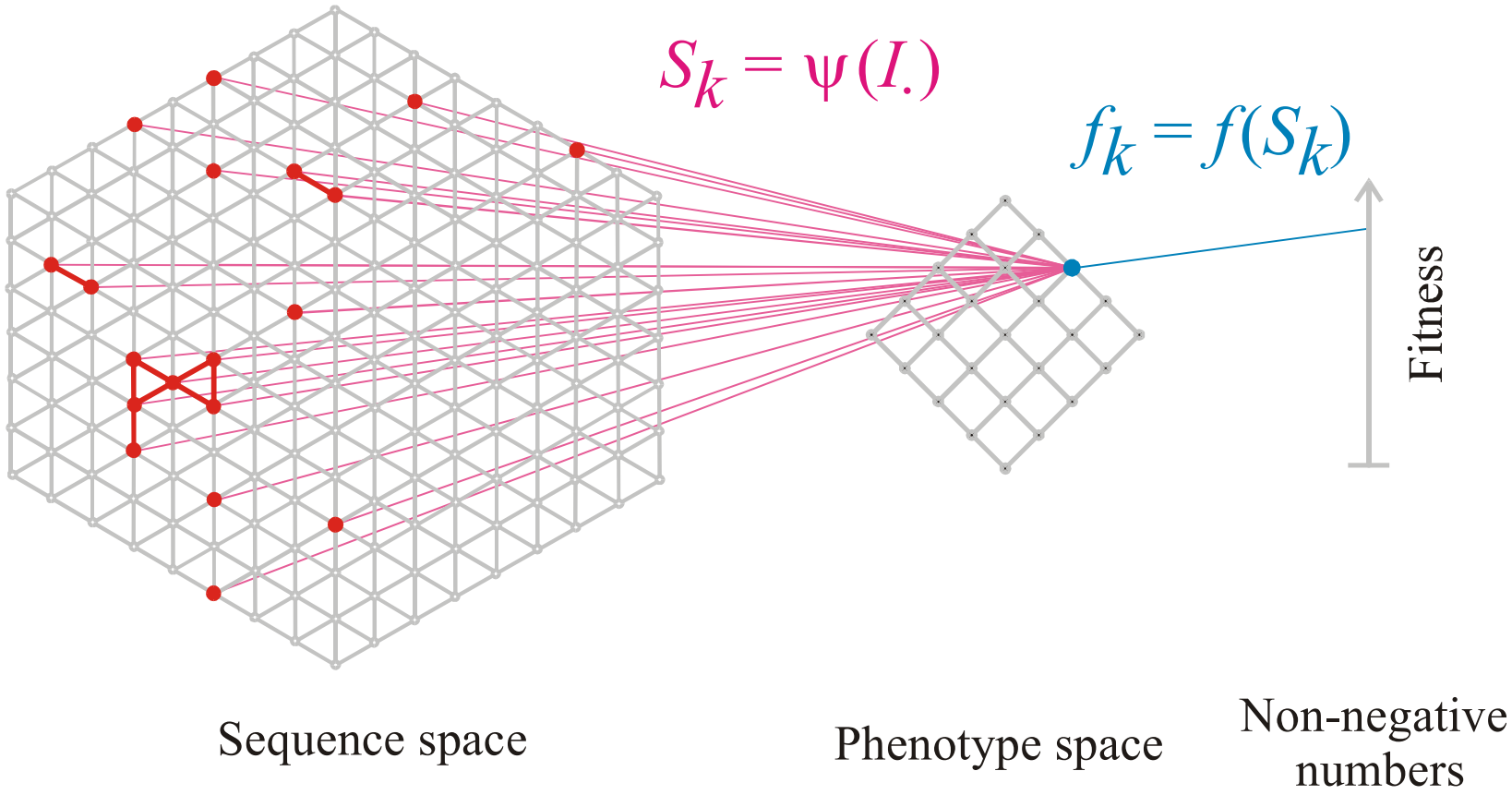
GUAGGCCCUCUUGACAUAAGAUUUUUCCAAUGGUGGGAGAUGGCCAUUGCAG





Mapping from sequence space into phenotype space and into fitness values


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

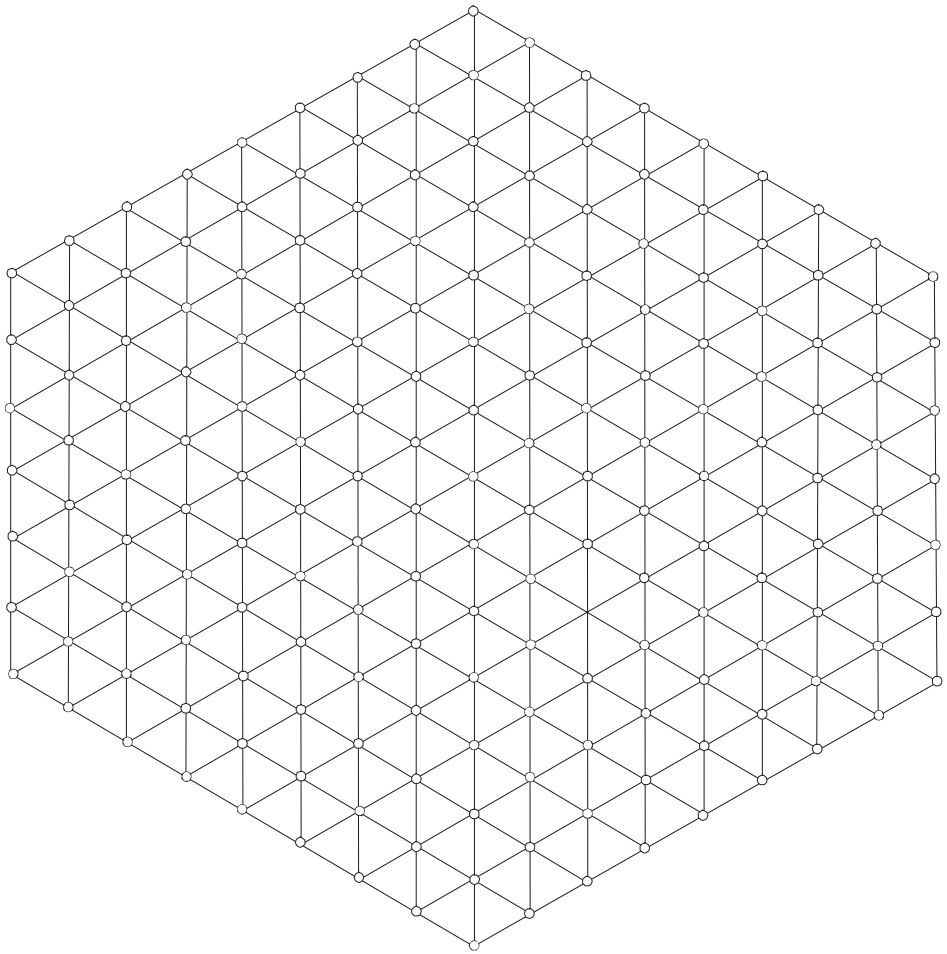

$S_k = \psi(I_\cdot)$
$f_k = f(S_k)$
Fitness
Sequence space
Phenotype space
Non-negative numbers

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

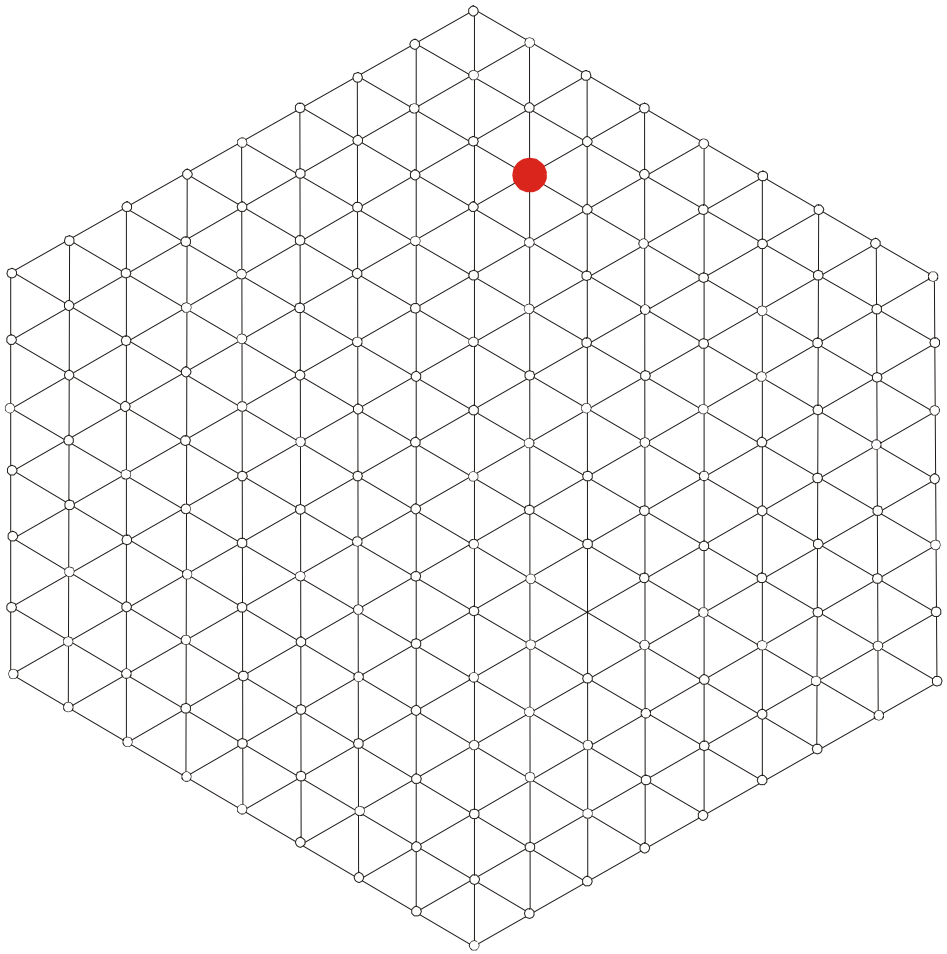
Sketch of sequence space



Random graph approach to neutral networks

Step 01

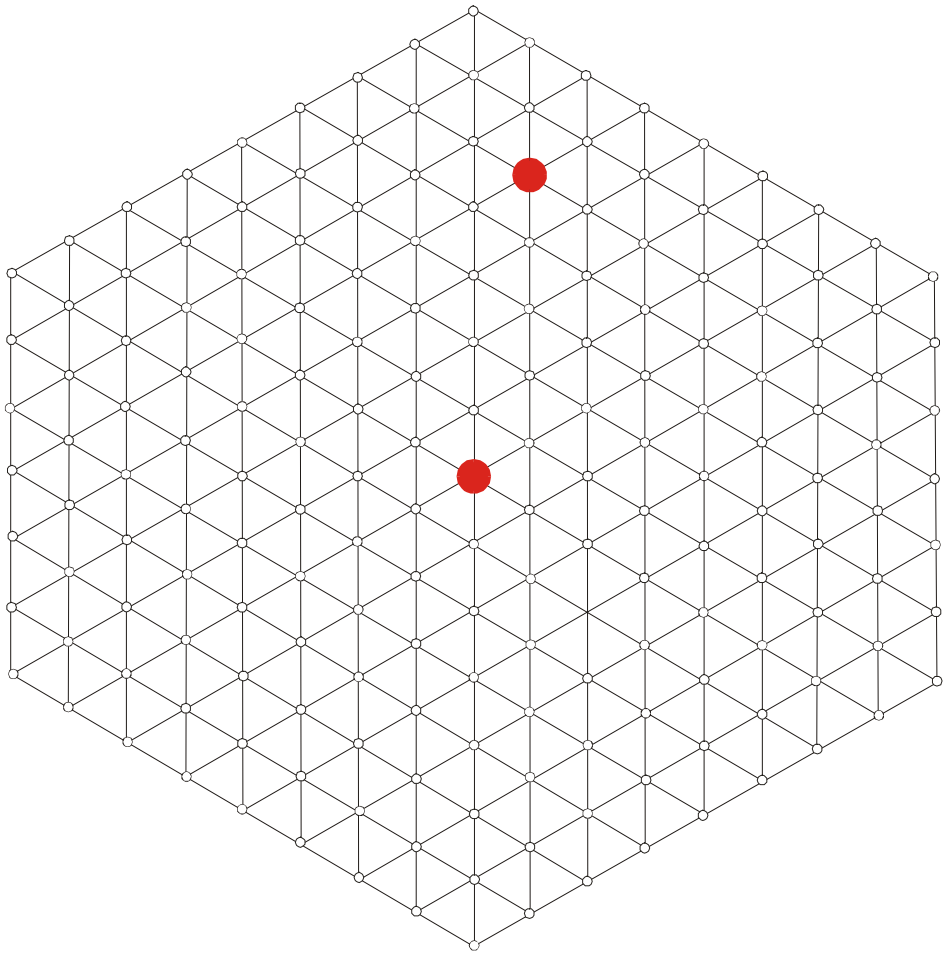
Sketch of sequence space



Random graph approach to neutral networks

Step 02

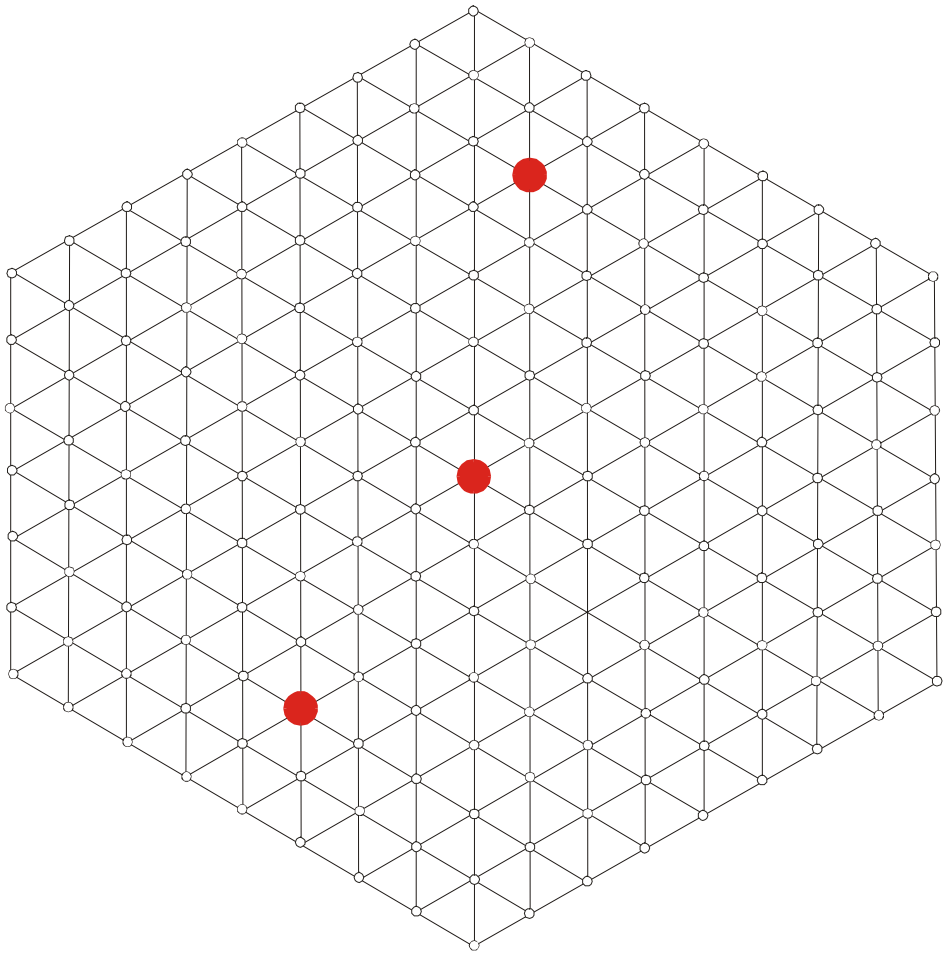
Sketch of sequence space



Random graph approach to neutral networks

Step 03

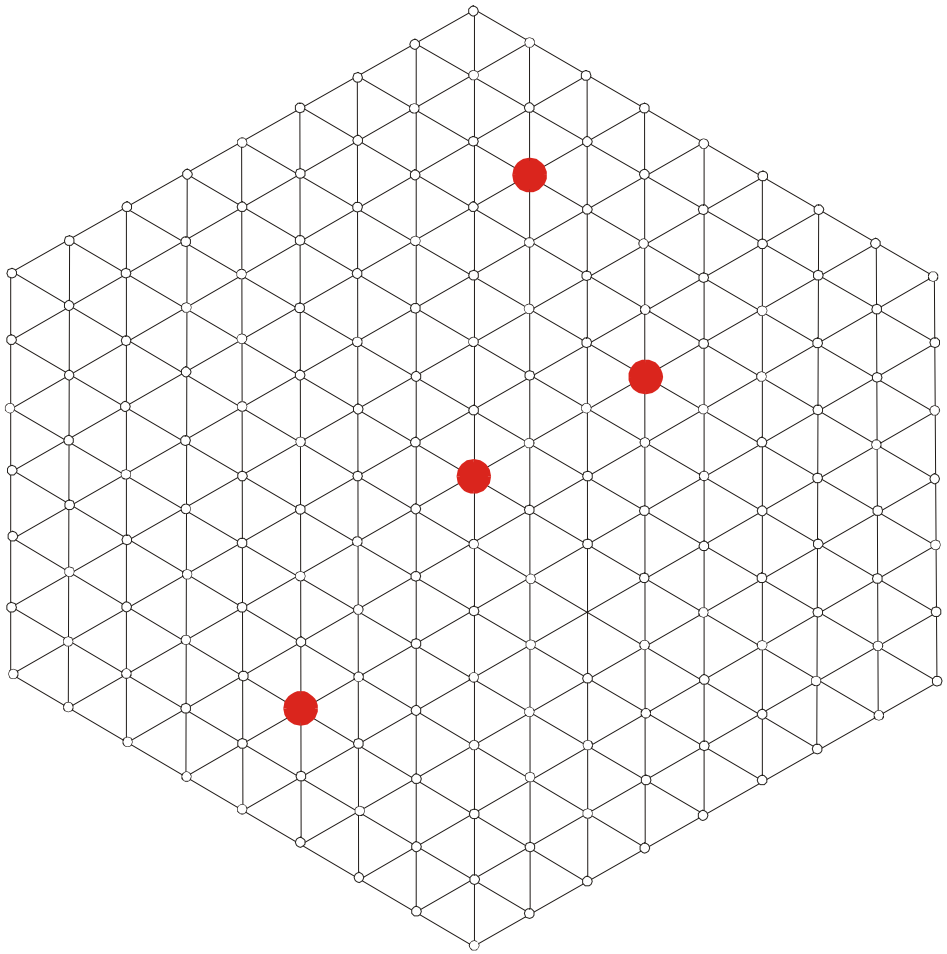
Sketch of sequence space



Random graph approach to neutral networks

Step 04

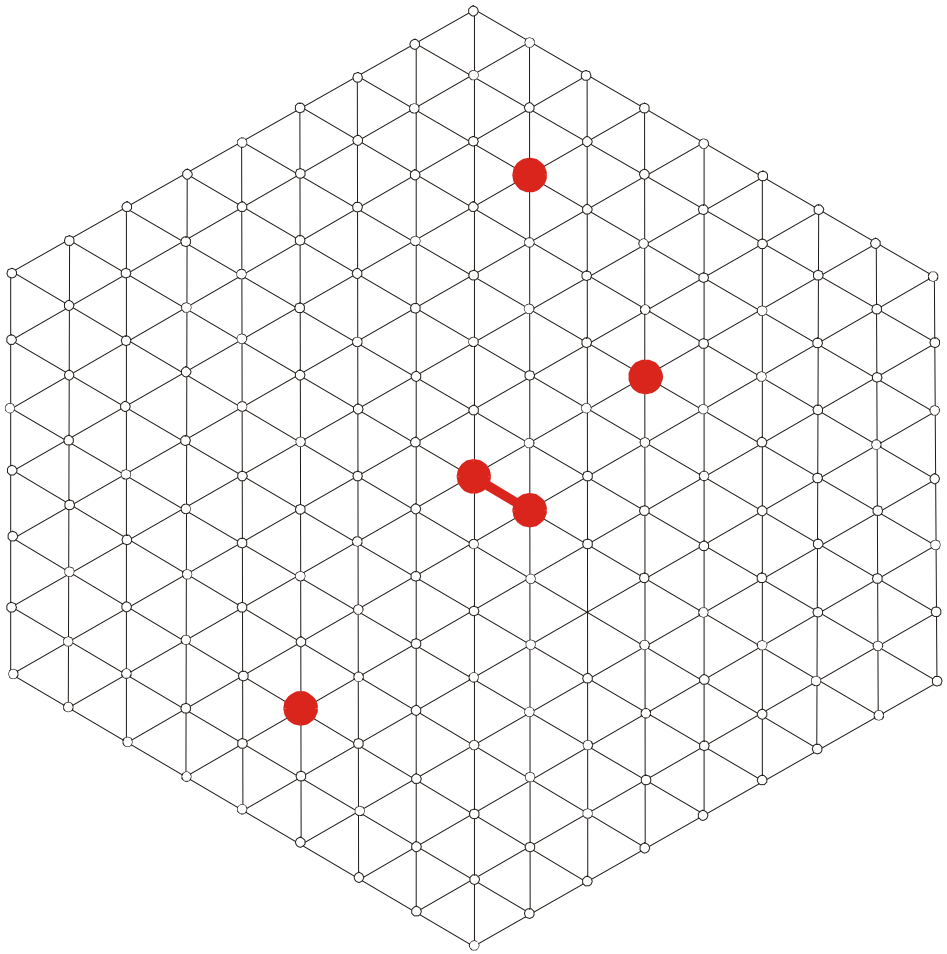
Sketch of sequence space



Random graph approach to neutral networks

Step 05

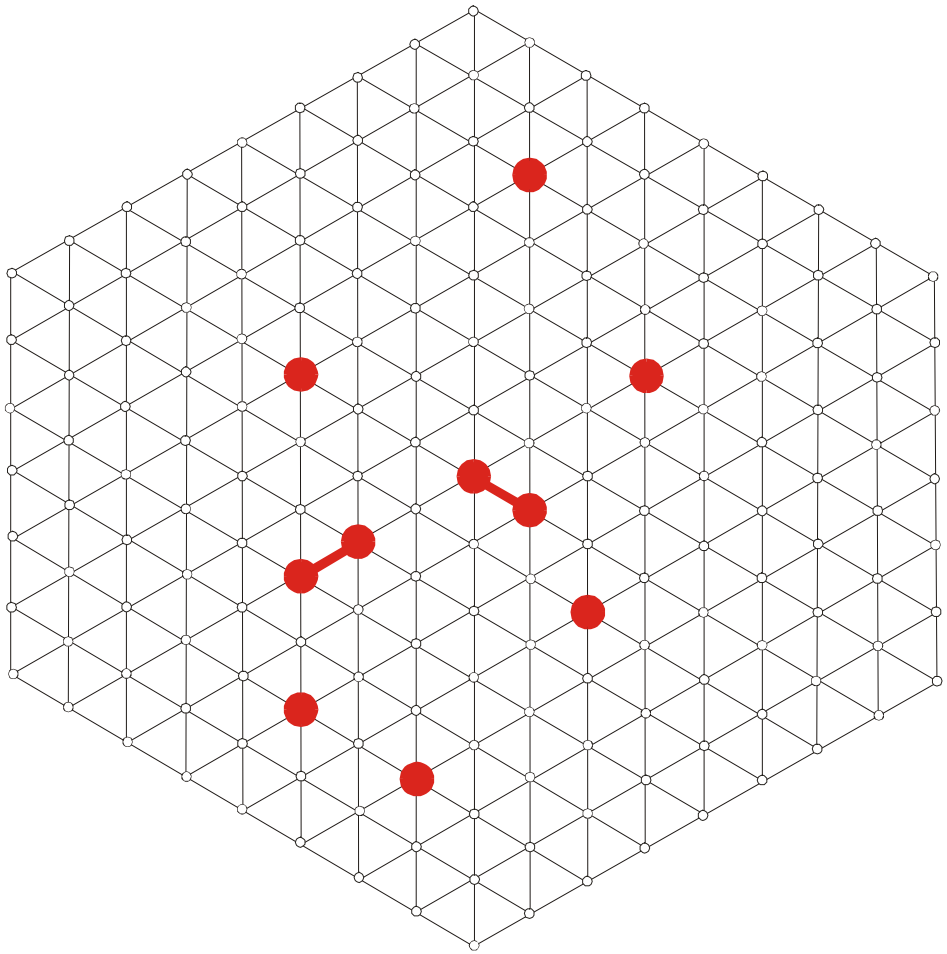
Sketch of sequence space



Random graph approach to neutral networks

Step 10

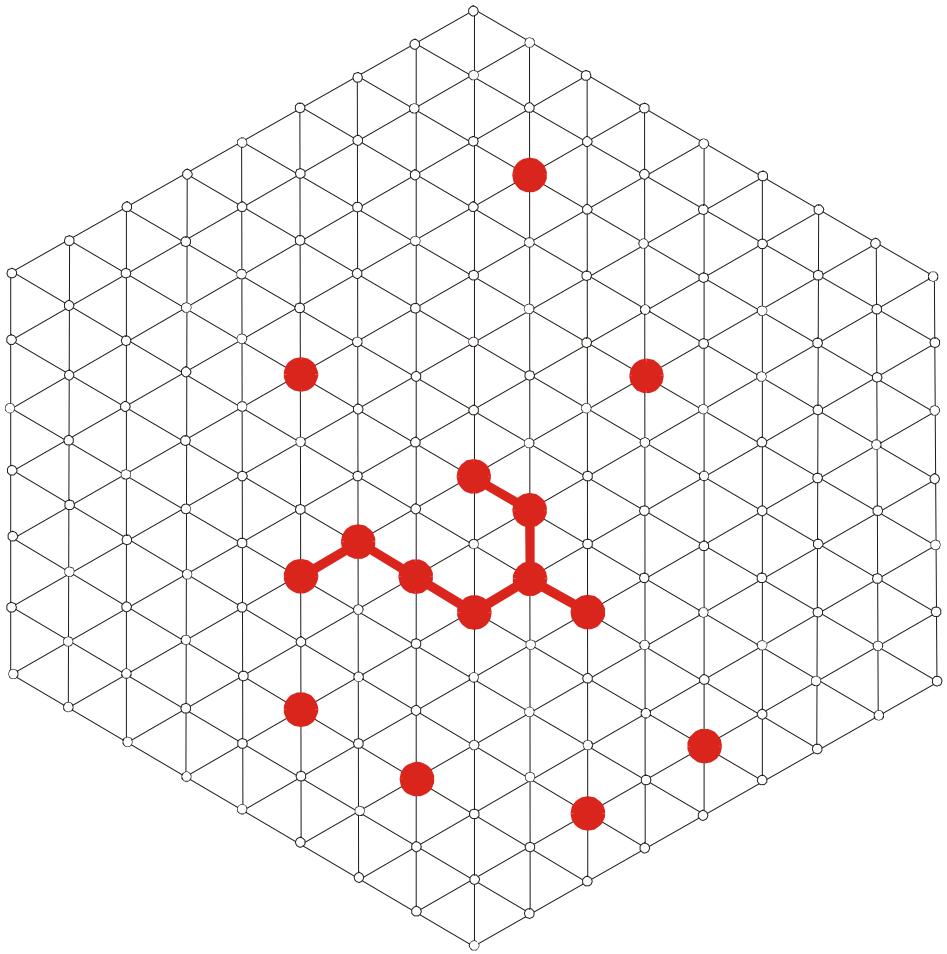
Sketch of sequence space



Random graph approach to neutral networks

Step 15

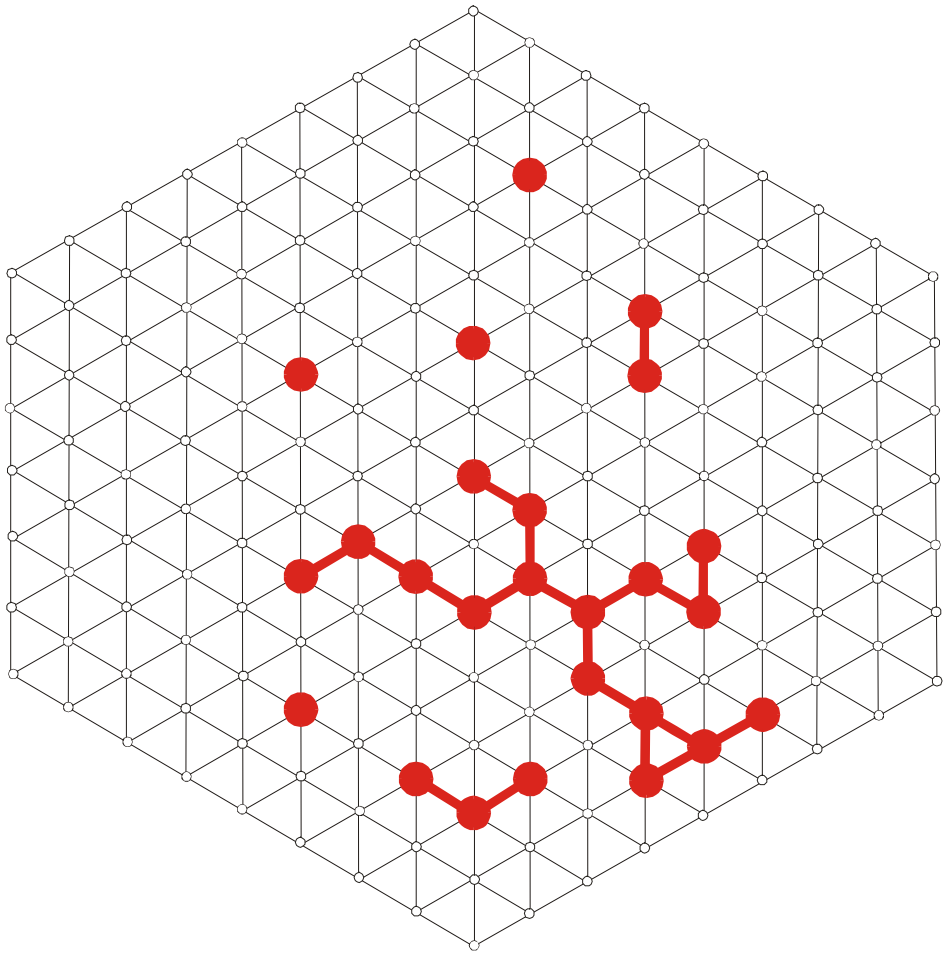
Sketch of sequence space



Random graph approach to neutral networks

Step 25

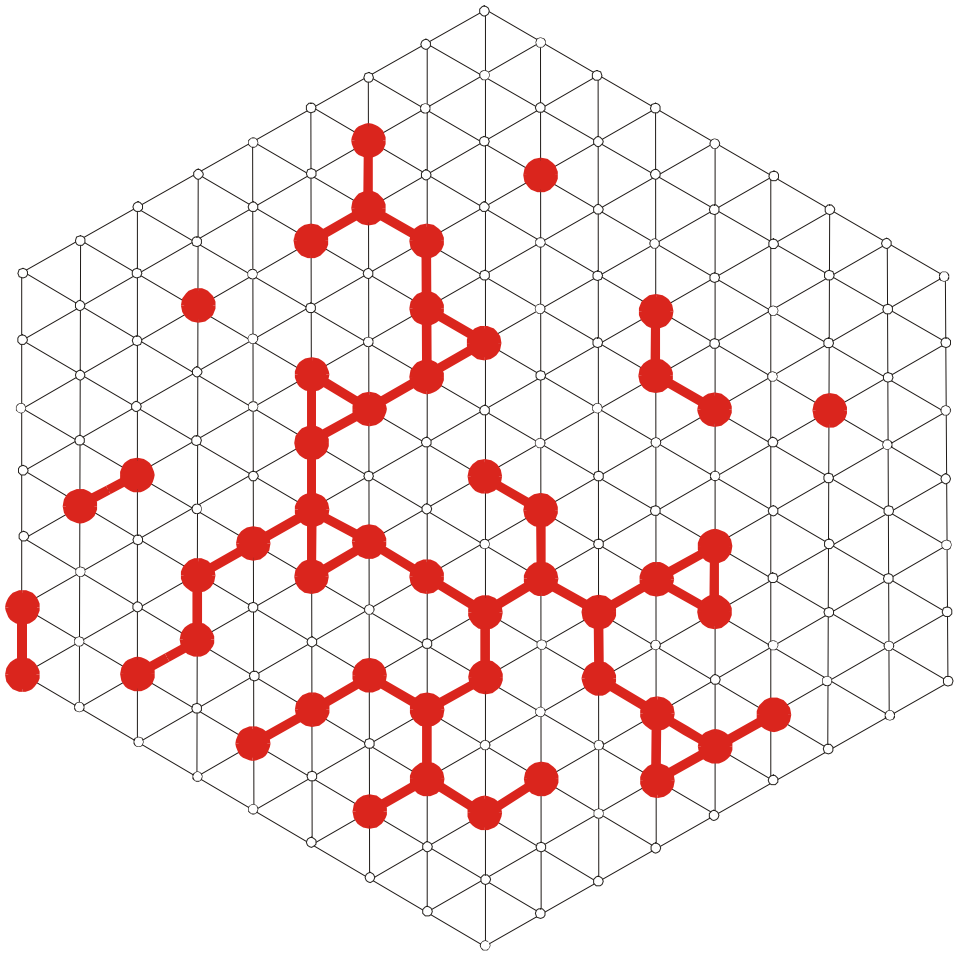
Sketch of sequence space



Random graph approach to neutral networks

Step 50

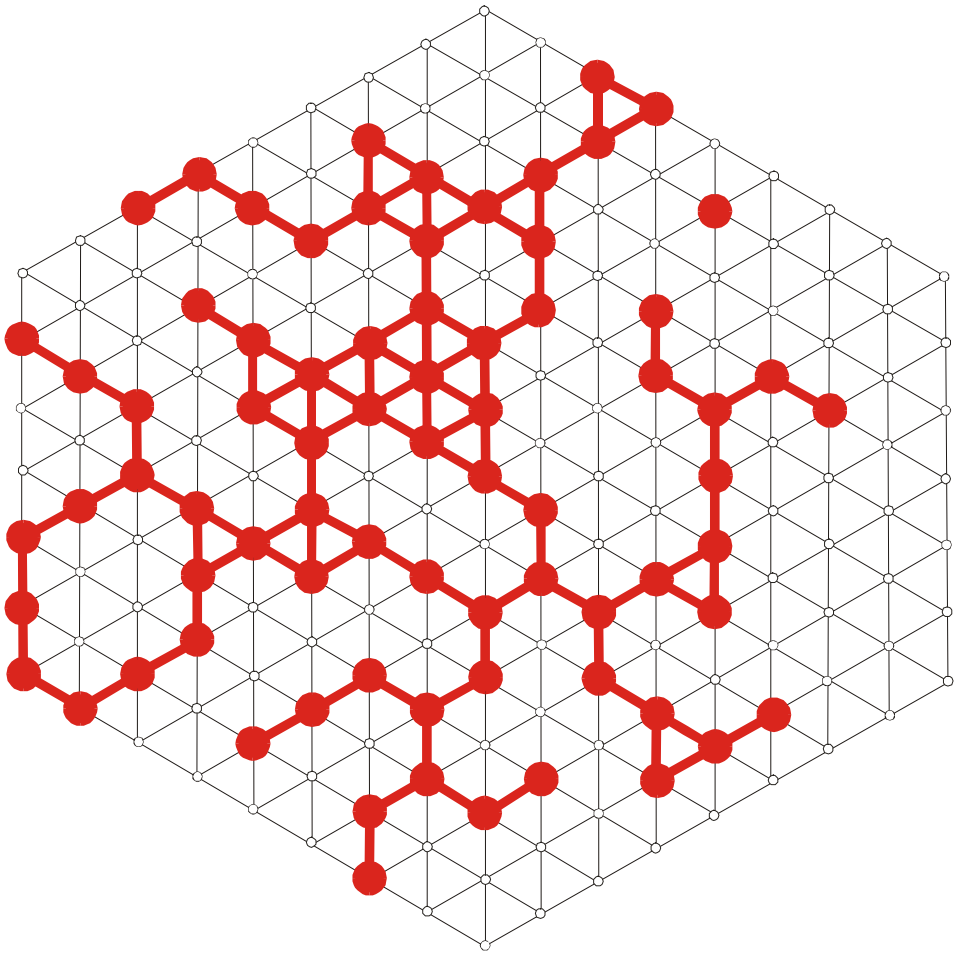
Sketch of sequence space



Random graph approach to neutral networks

Step 75

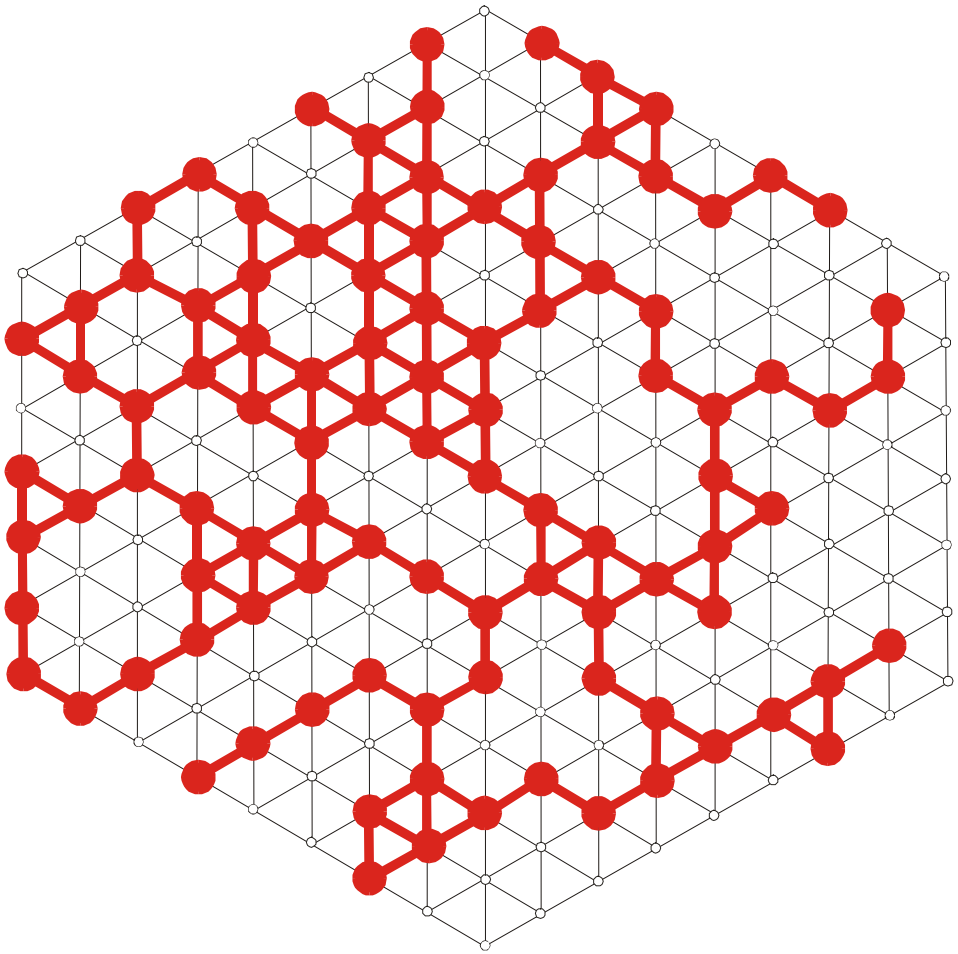
Sketch of sequence space



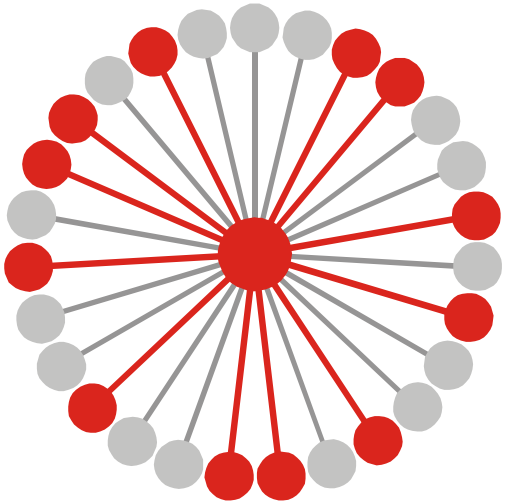
Random graph approach to neutral networks

Step 100

Sketch of sequence space



Random graph approach to neutral networks



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

$$\lambda_j = 12 / 27 , \qquad \bar{\lambda}_k = \frac{\underset{j \in |G_k|}{\text{ø}} \; l_j(k)}{|G_k|}$$

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

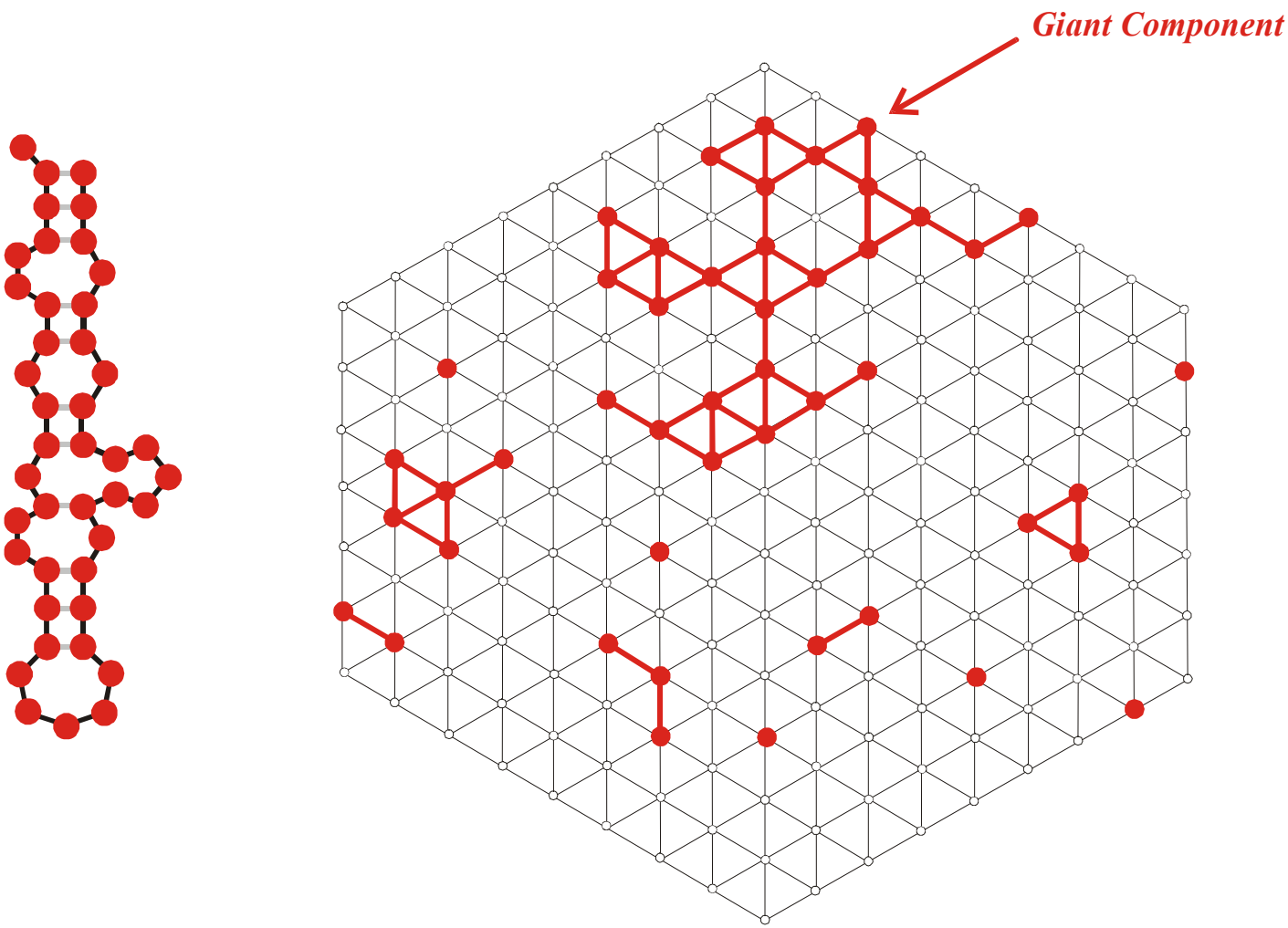
Alphabet size k : **AUGC** ñ k = 4

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

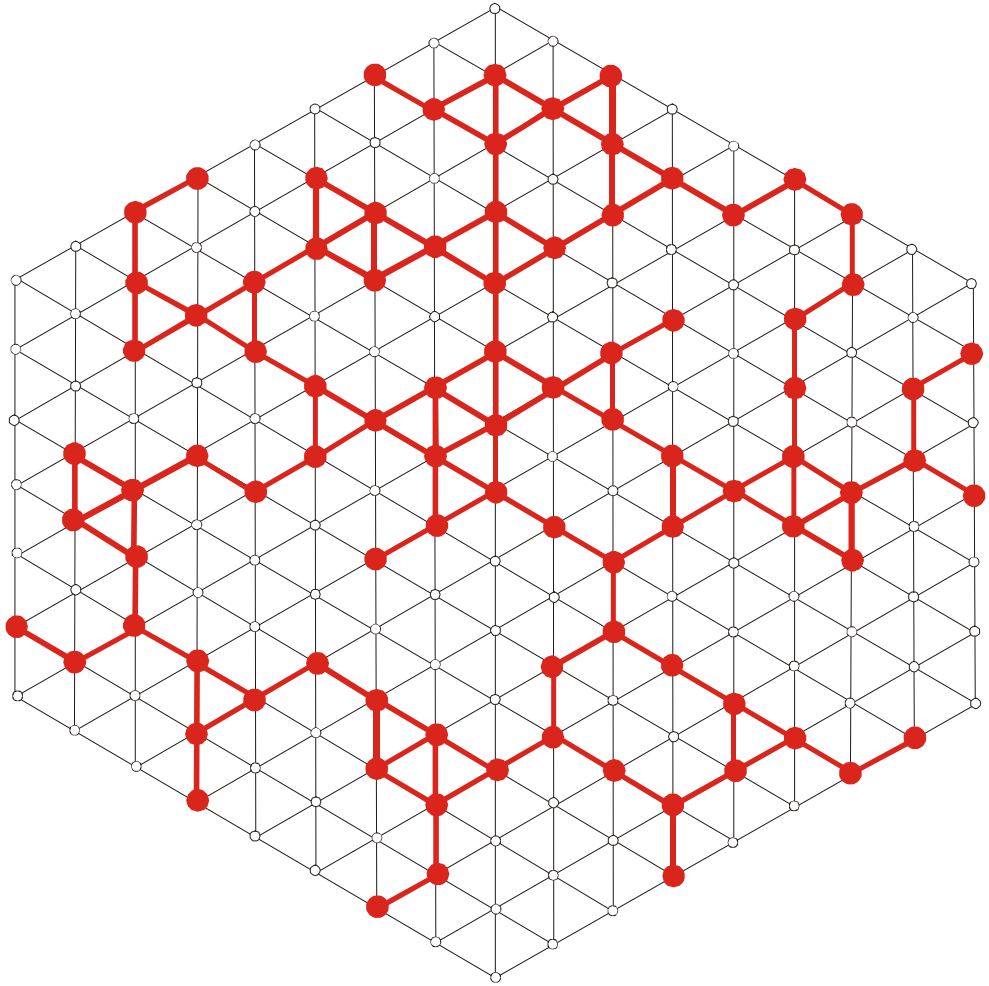
$\bar{\lambda}_k < \lambda_{cr}$ .... network G_k is **not** connected

k	l_{cr}
2	0.5
3	0.4226
4	0.3700

Mean degree of neutrality and connectivity of neutral networks



A multi-component neutral network



A connected 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.

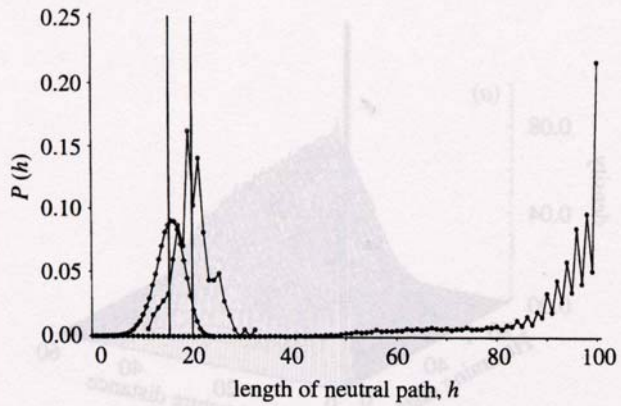
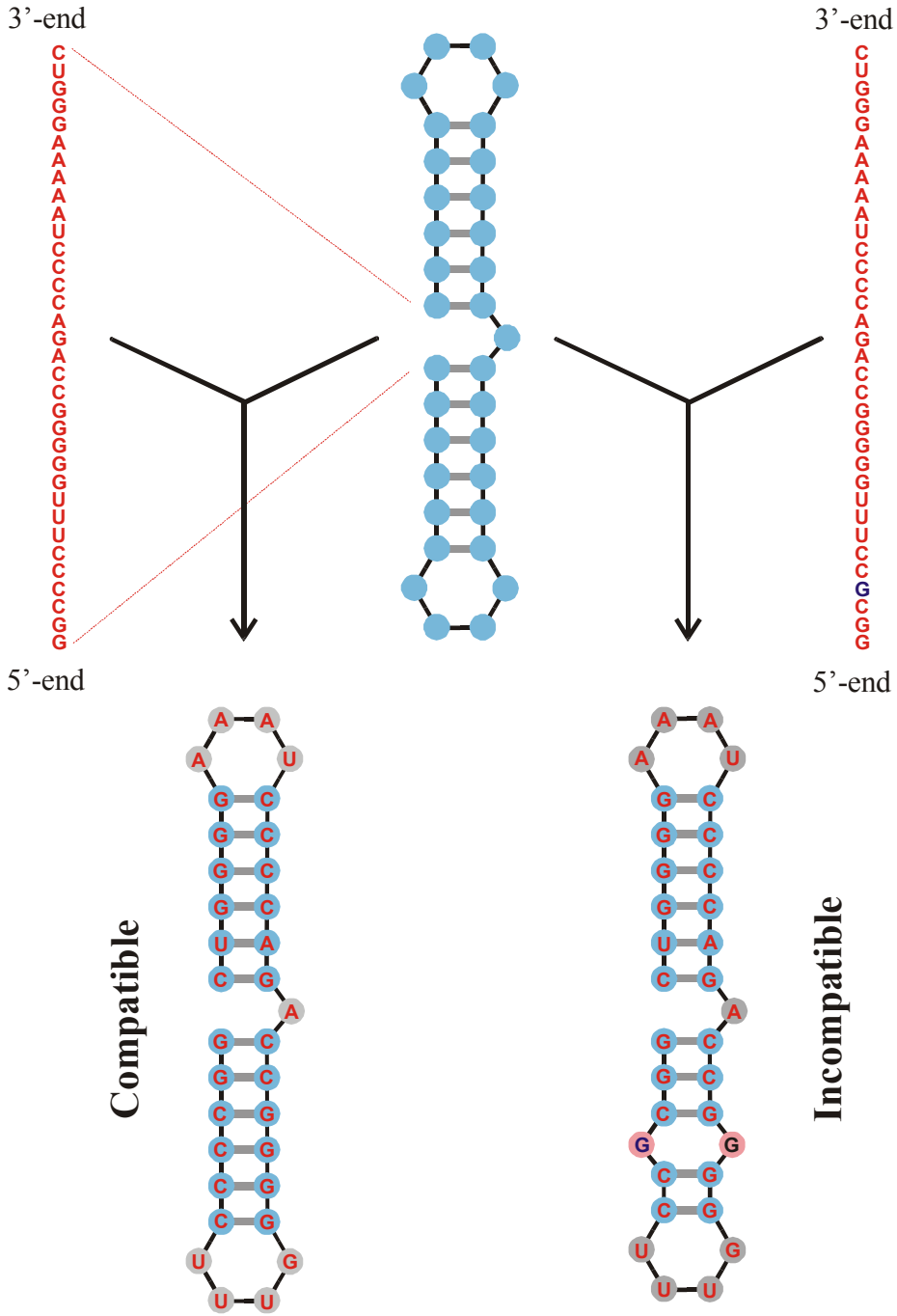


Figure 4. Neutral paths. A neutral path is defined by a series of nearest neighbour sequences that fold into identical structures. Two classes of nearest neighbours are admitted: neighbours of Hamming distance 1, which are obtained by single base exchanges in unpaired stretches of the structure, and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



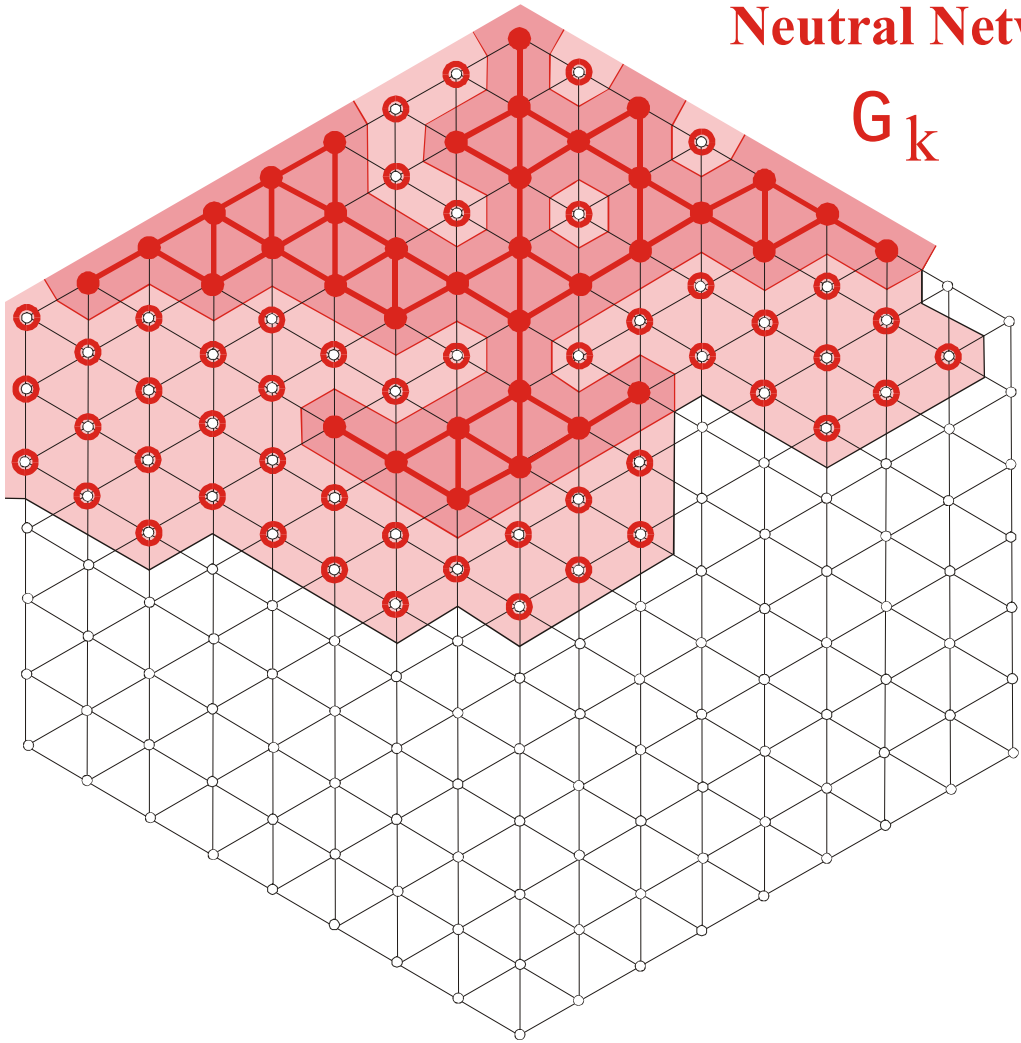
Sequences are **compatible** or incompatible with structures

G_k à C_k

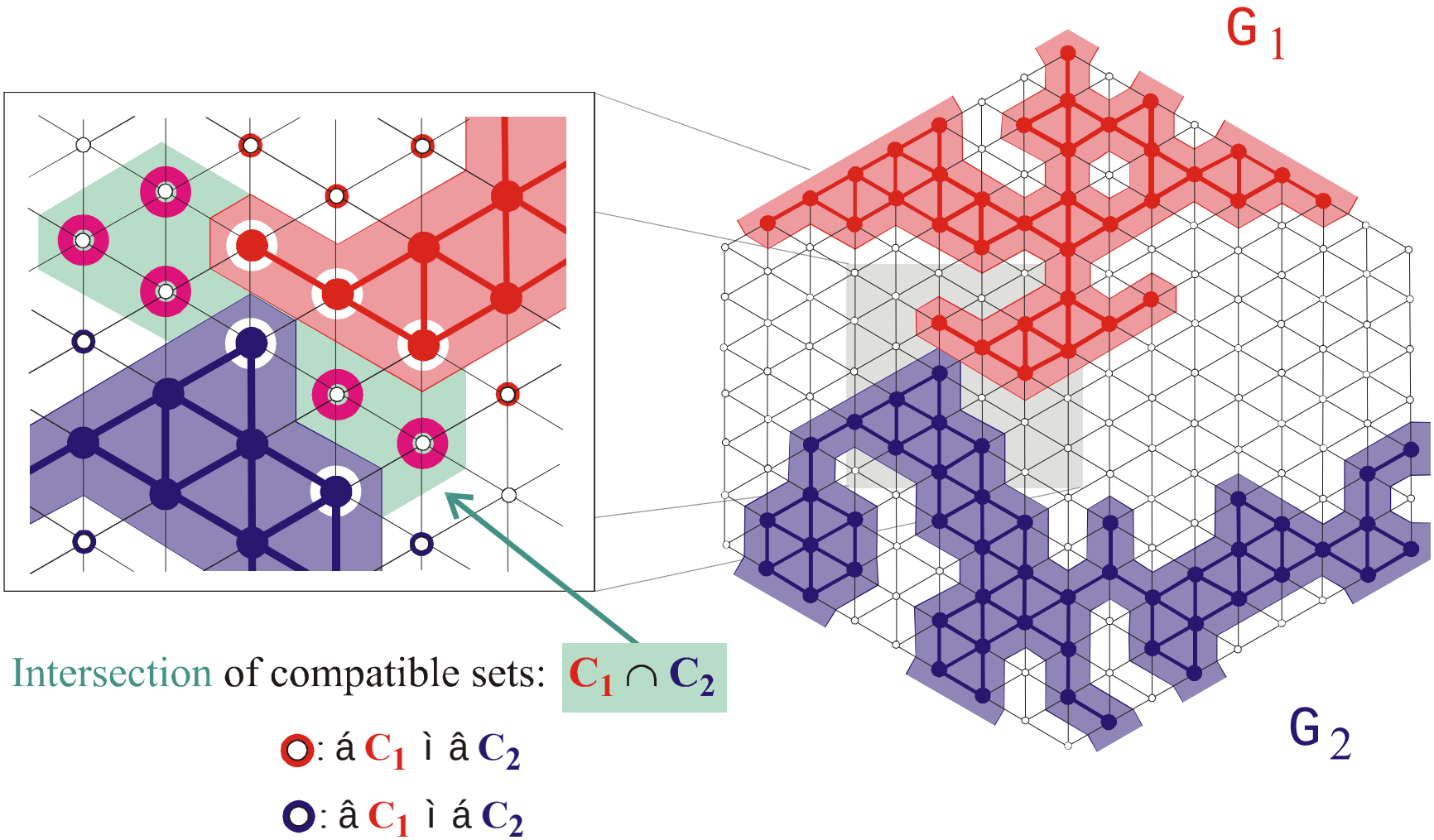
Neutral Network

G_k

Compatible Set C_k

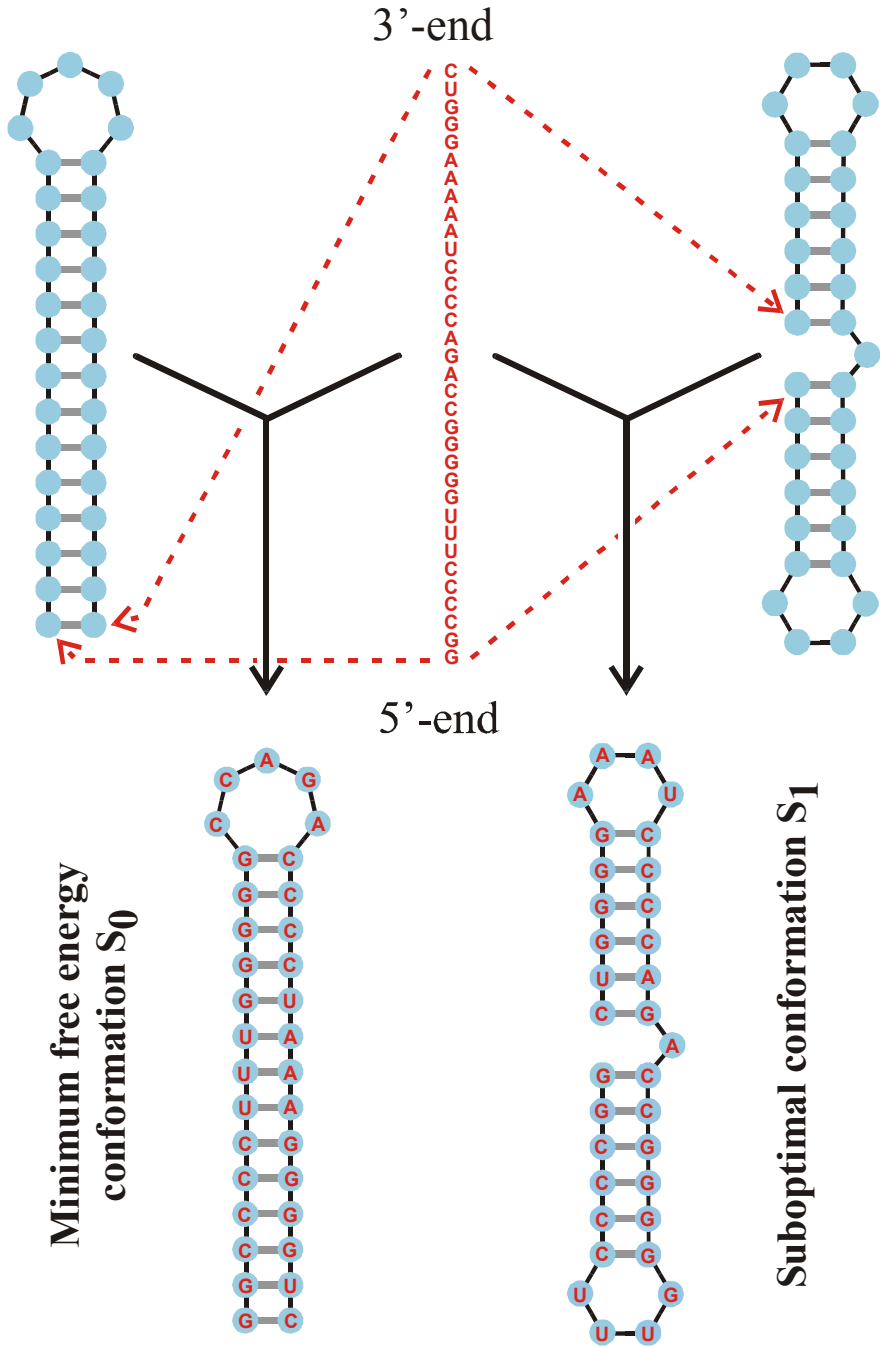


Neutral networks G_k are embedded in sets of compatible sequences C_k.



Two neutral networks, G_1 and G_2, are embedded in compatible sets, C_1 and C_2, respectively. The compatible sets form an intersection consisting of sequences that can form both structures.

A sequence at the **intersection** of two structures



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

0092-8240/97 $17.00 + 0.00

S0092-8240(96)00089-4

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

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

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

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

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

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

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

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

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

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

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

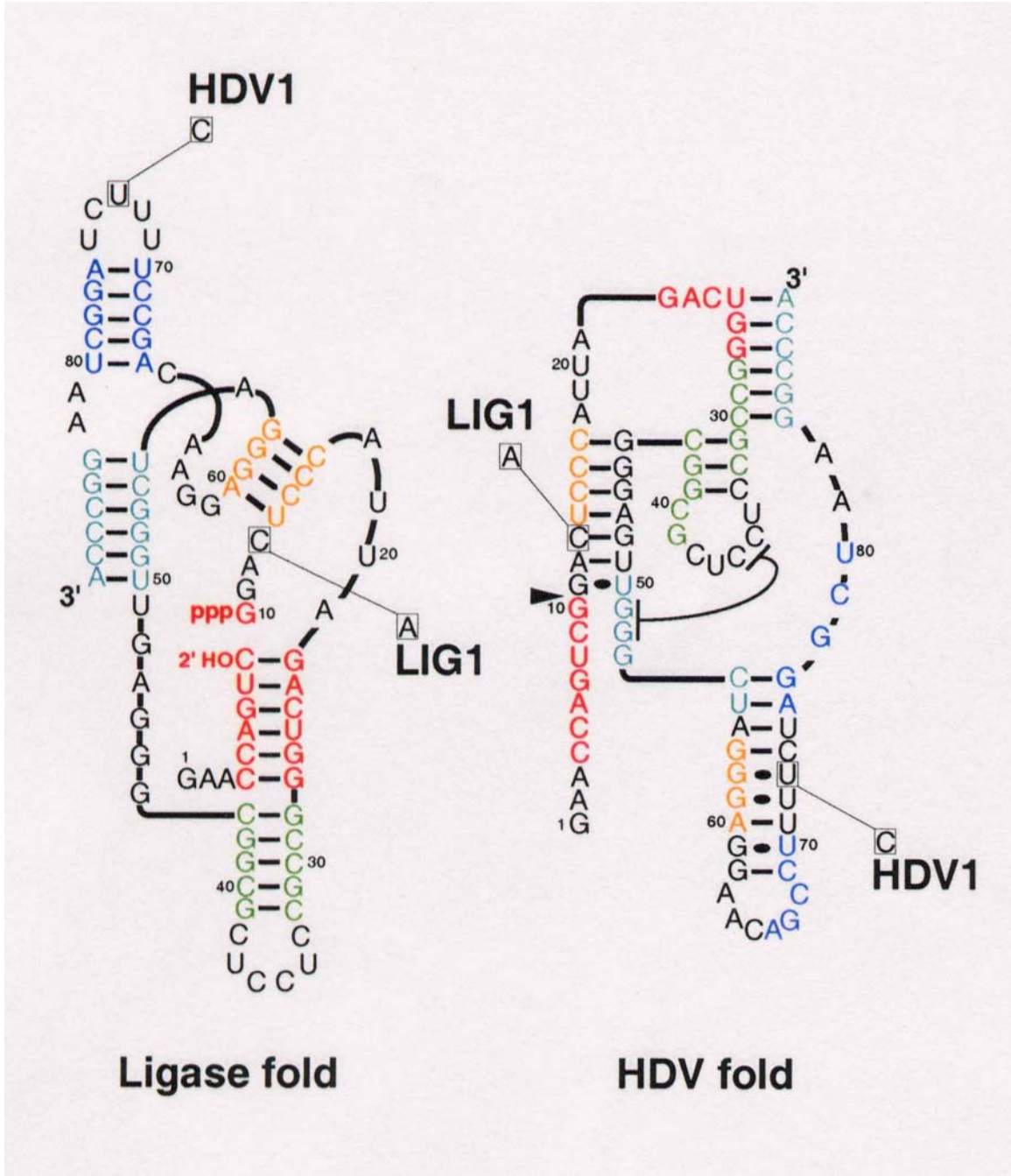
Reference for the definition of the intersection and the proof of the **intersection theorem**

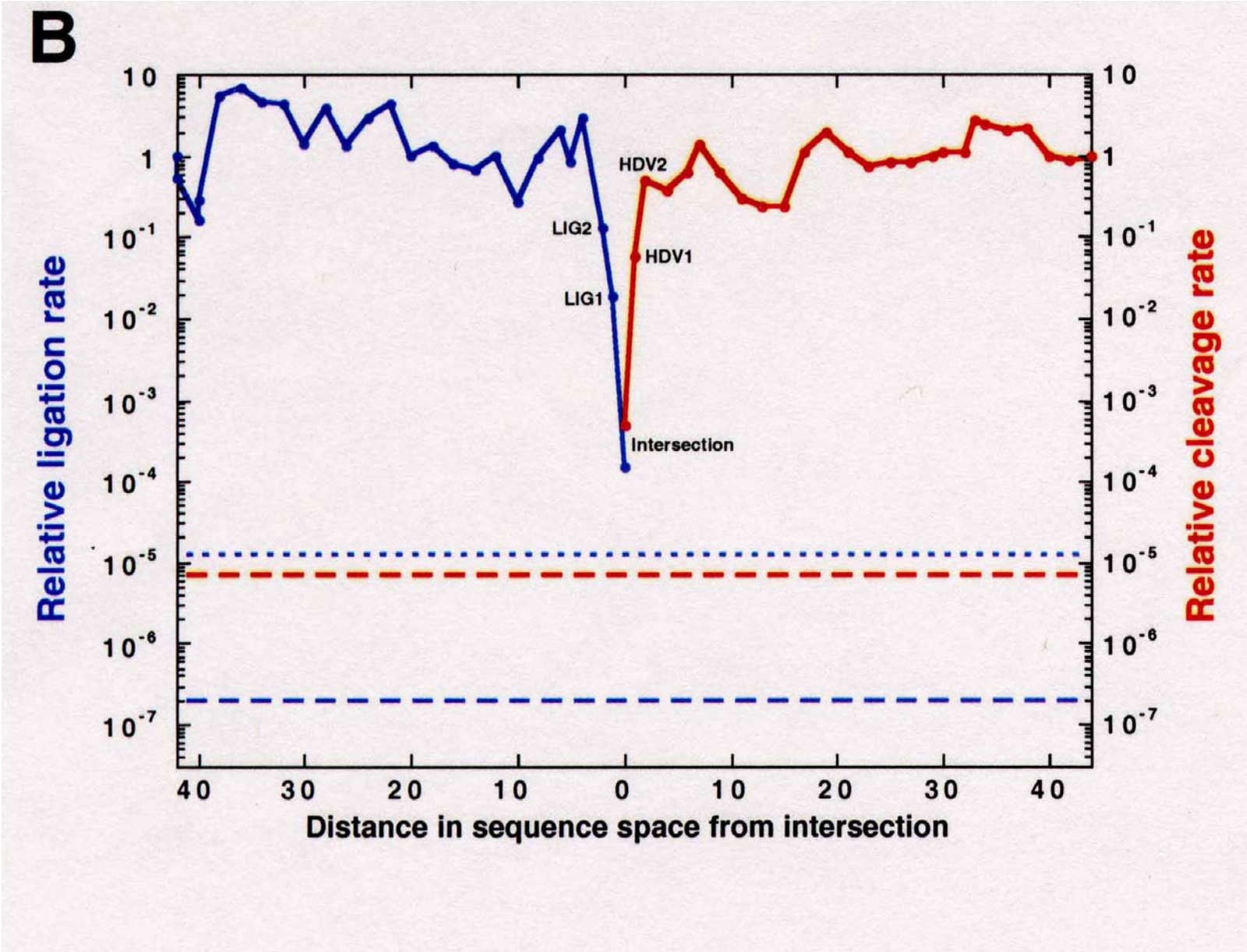
A ribozyme switch

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

The 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

Coworkers

Walter Fontana, Santa Fe Institute, NM

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

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

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

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