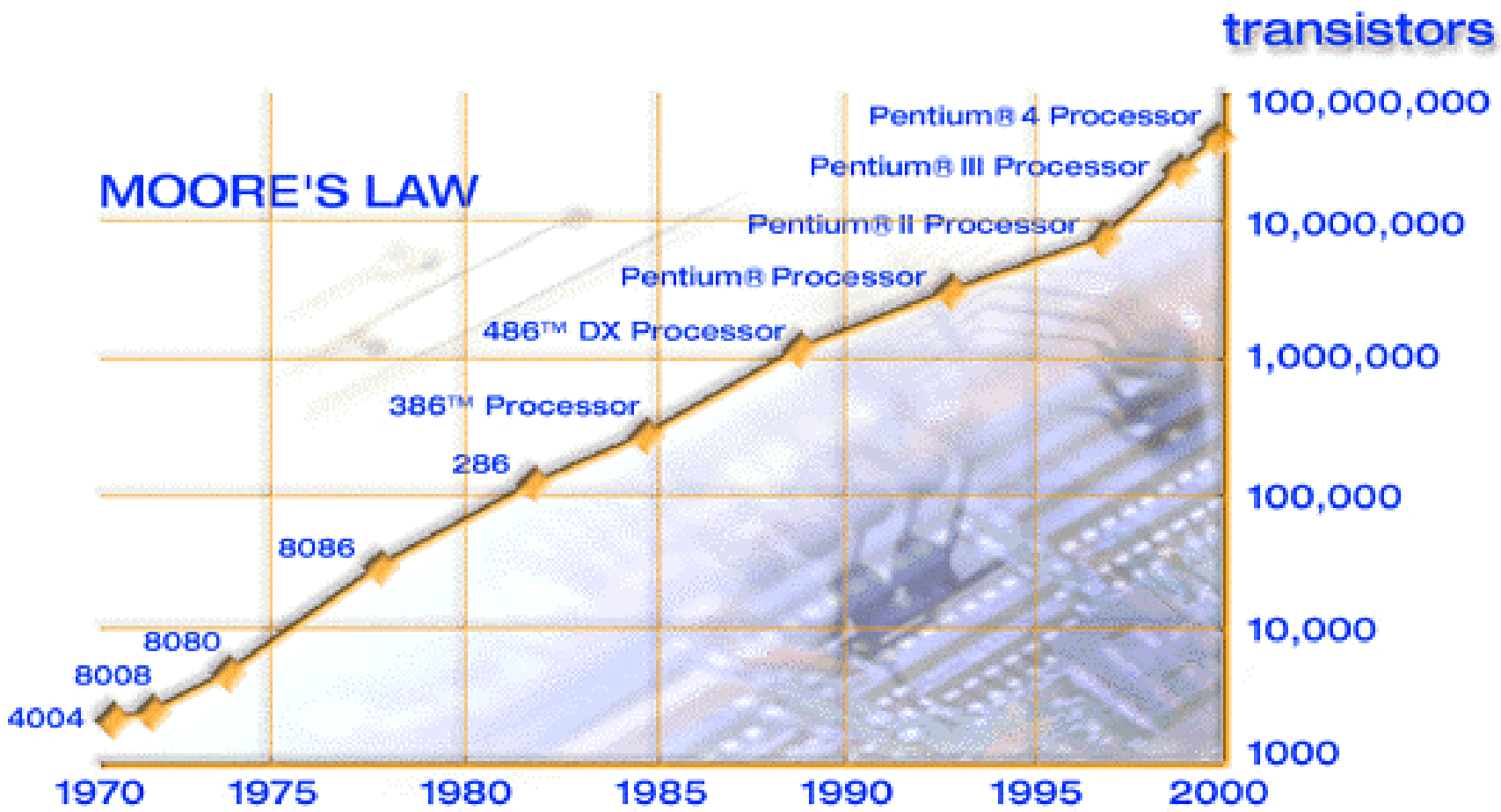
Different kinds of robustness in genetic and metabolic networks

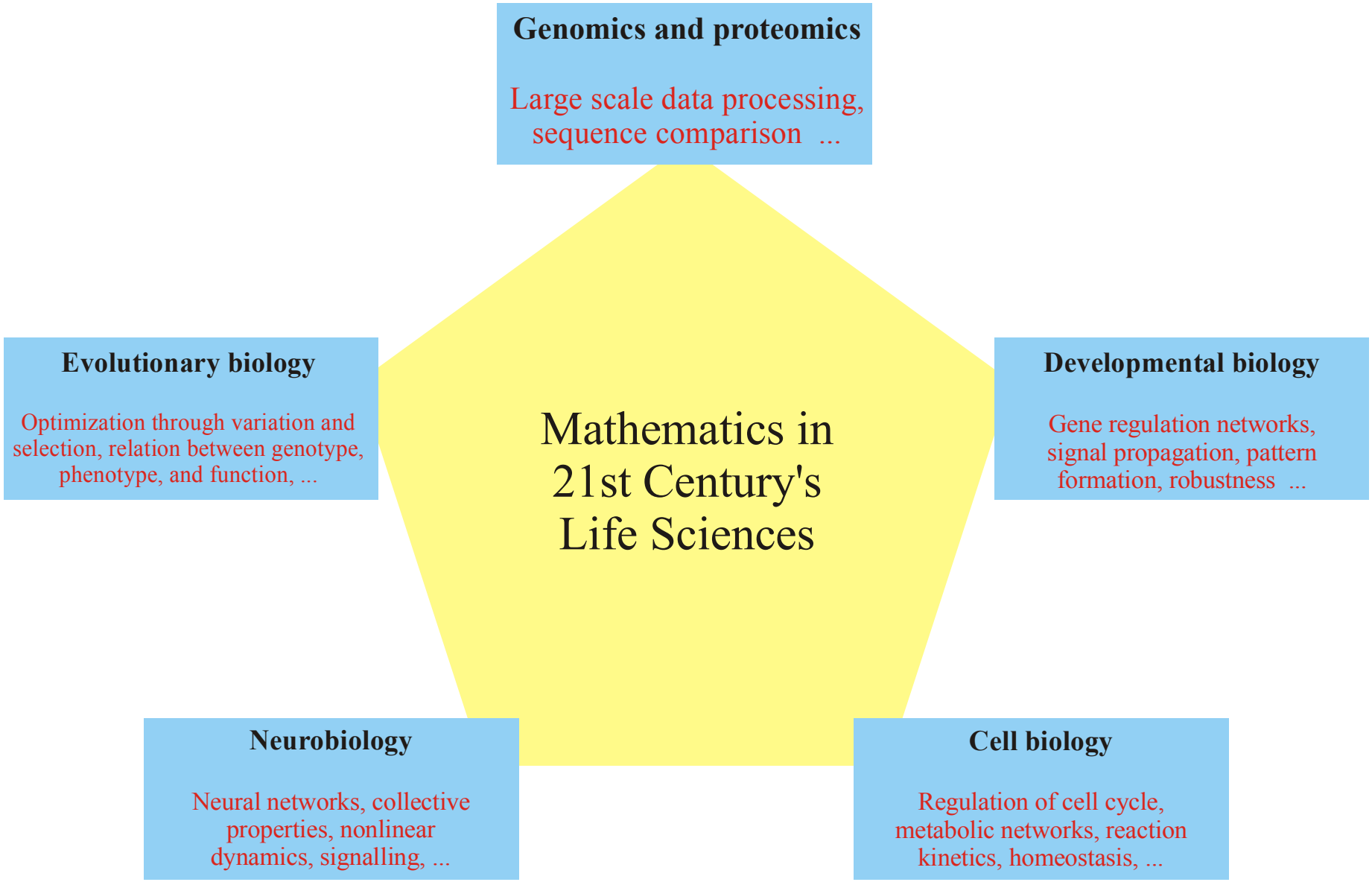
Peter Schuster

Institut für Theoretische Chemie und Molekulare Strukturbiologie der Universität Wien

Seminar lecture

Linz, 15.12.2003

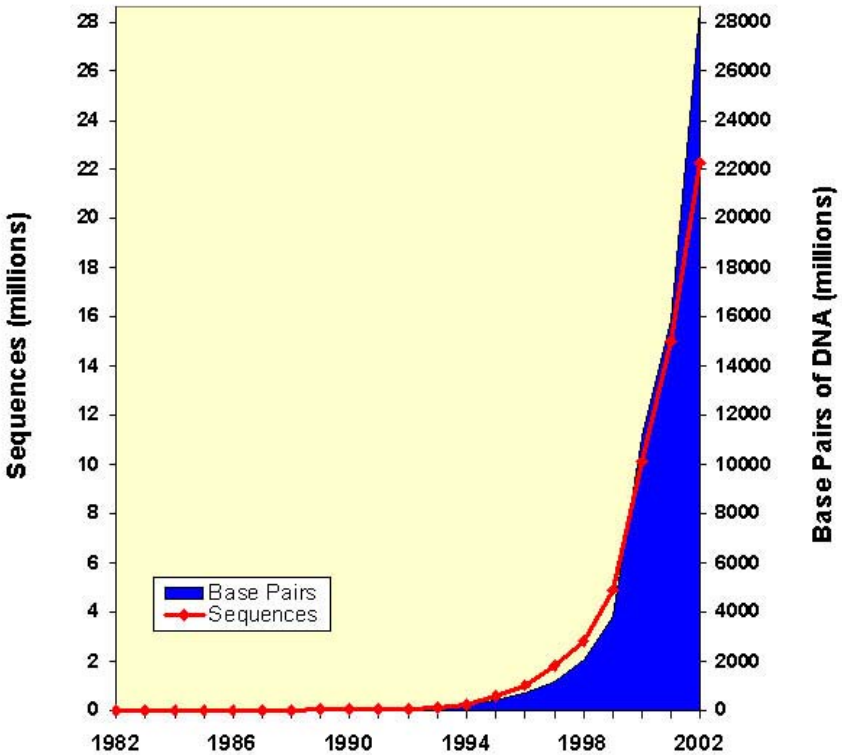

transistors
MOORE'S LAW
Pentium® 4 Processor
Pentium® III Processor
Pentium® II Processor
Pentium® Processor
486™ DX Processor
386™ Processor
286
8086
8080
8008
4004
100,000,000
10,000,000
1,000,000
100,000
10,000
1000
1970
1975
1980
1985
1990
1995
2000


Genomics and proteomics
Large scale data processing, sequence comparison ...
Evolutionary biology
Optimization through variation and selection, relation between genotype, phenotype, and function, ...
Mathematics in 21st Century's Life Sciences
Developmental biology
Gene regulation networks, signal propagation, pattern formation, robustness ...
Neurobiology
Neural networks, collective properties, nonlinear dynamics, signalling, ...
Cell biology
Regulation of cell cycle, metabolic networks, reaction kinetics, homeostasis, ...

Genomics and proteomics

Large scale data processing, sequence comparison ...

E. coli:	Length of the Genome	4×10^6 Nucleotides
	Number of Cell Types	1
	Number of Genes	4 000
Man:	Length of the Genome	3×10^9 Nucleotides
	Number of Cell Types	200
	Number of Genes	30 000 - 100 000



Source: NCBI

Fully sequenced genomes

- *Organisms* 751 projects

153 complete (16 A, 118 B, 19 E)

(*Eukarya* examples: mosquito (pest, malaria), sea squirt, mouse, yeast, homo sapiens, arabidopsis, fly, worm, …)

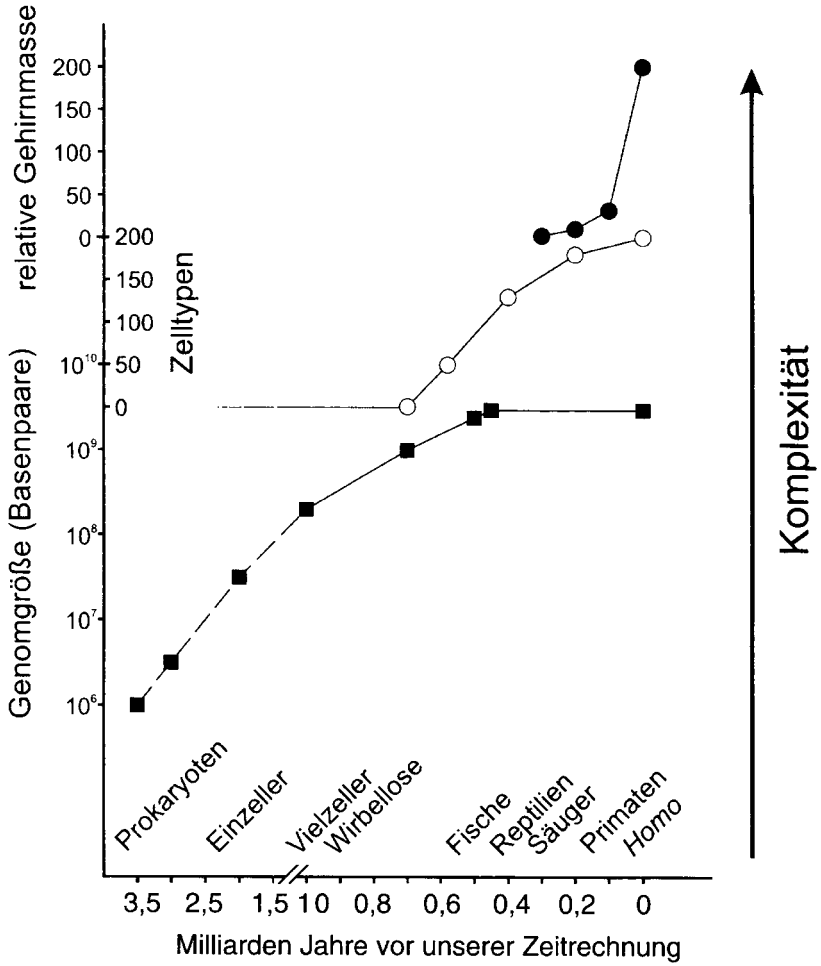
598 ongoing (23 A, 332 B, 243 E)

(*Eukarya* examples: chimpanzee, turkey, chicken, ape, corn, potato, rice, banana, tomato, cotton, coffee, soybean, pig, rat, cat, sheep, horse, kangaroo, dog, cow, bee, salmon, fugu, frog, …)

- *Other structures with genetic information*

68 phages
1328 viruses
35 viroids
472 organelles (423 mitochondria, 32 plastids, 14 plasmids, 3 nucleomorphs)

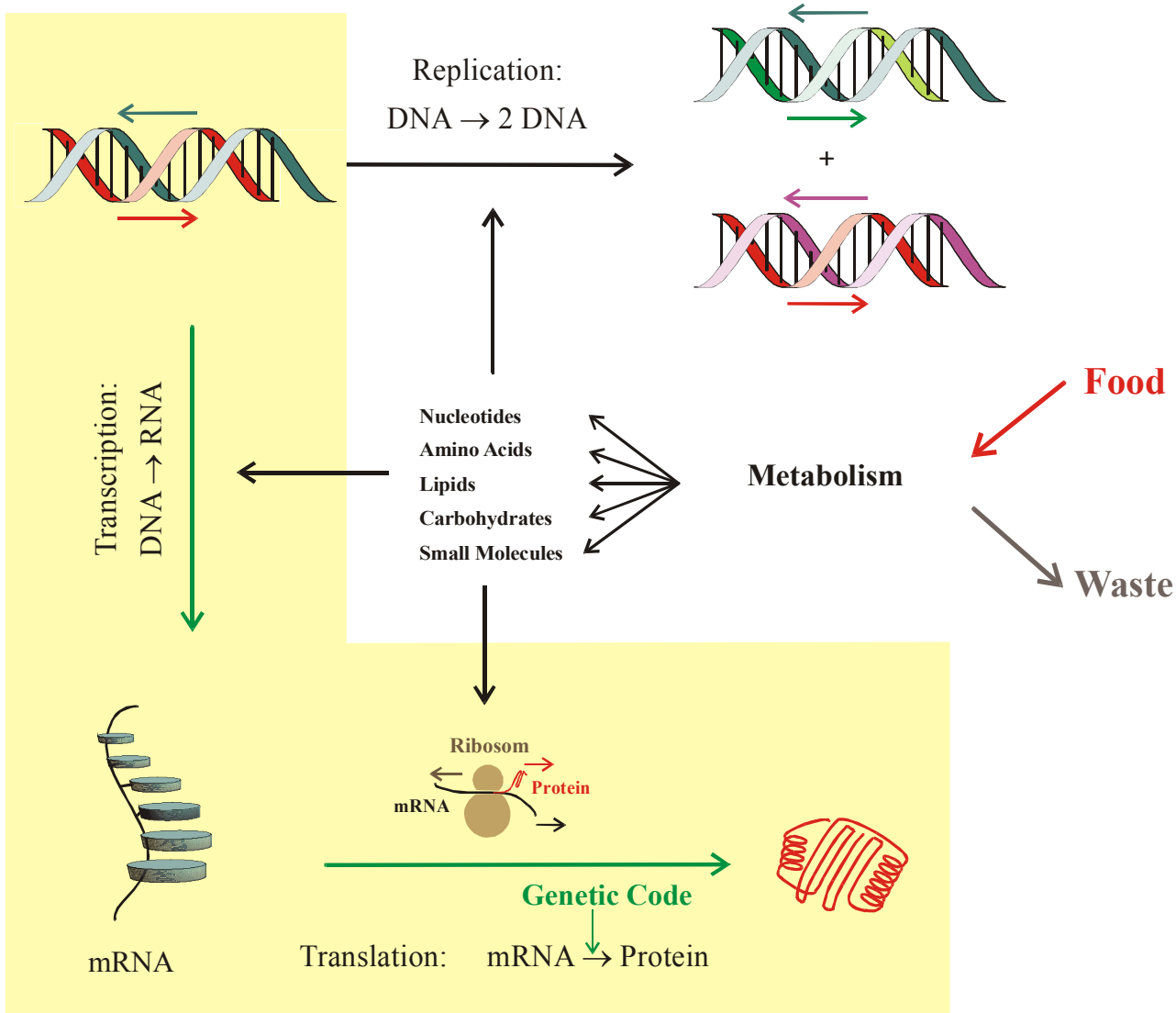
Source: Integrated Genomics, Inc.
August 12th, 2003



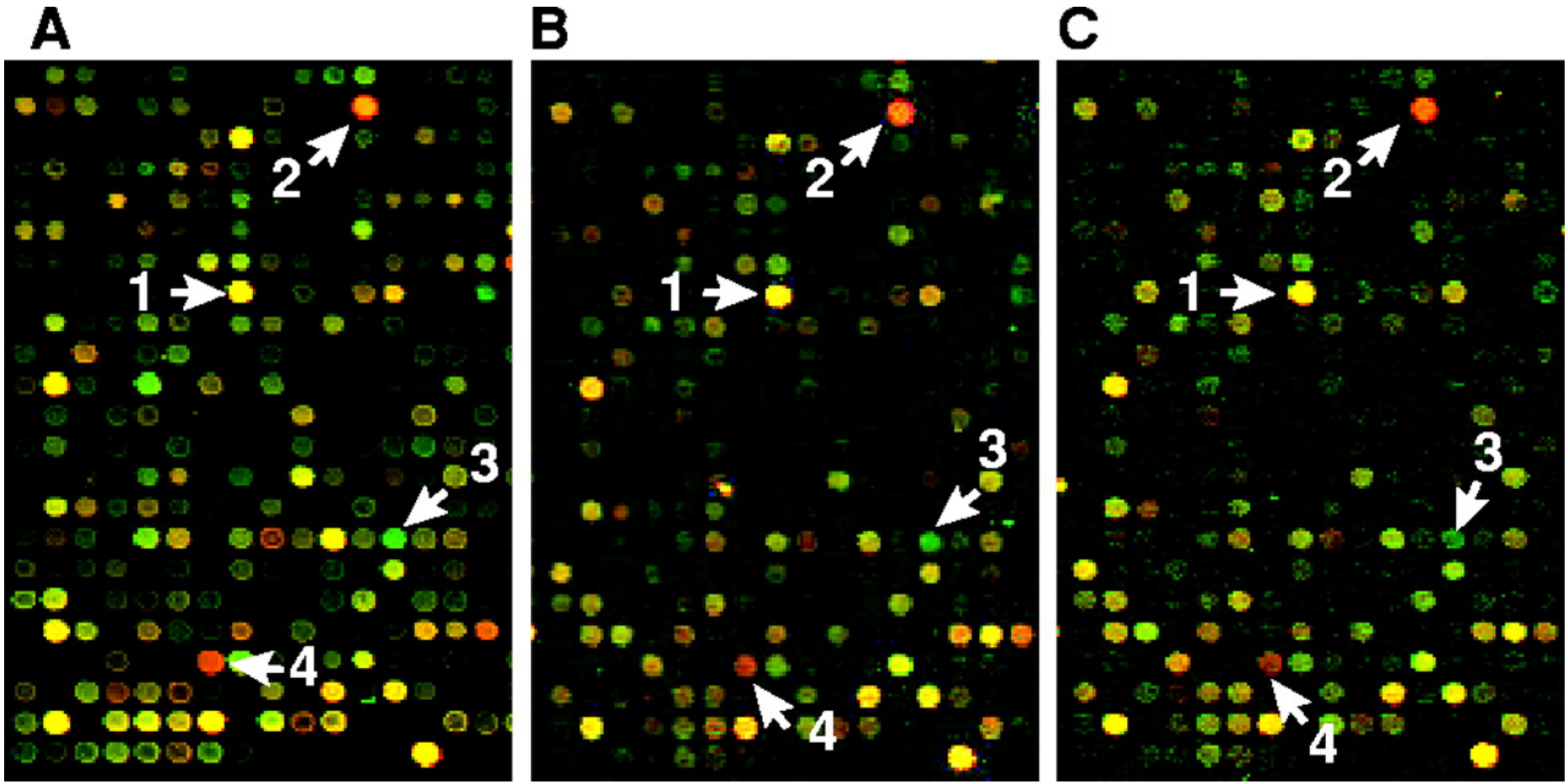
4.10 Die Zunahme der Komplexität ist ein wesentlicher Aspekt der biologischen Evolution, wobei höhere Komplexität sowohl durch Vergrößerung der Zahl von miteinander in Wechselwirkung stehenden Elementen als auch durch Differenzierung der Funktionen dieser Elemente entstehen kann. In dieser Abbildung wird zwischen drei Phasen oder Strategien der Evolution von Komplexität unterschieden. *Untere Kurve*: Zunahme der Genomgröße; logarithmische Auftragung der Zahl der Basenpaare im Genom von Zellen seit Beginn der biologischen Evolution (Daten aus Abbildung 2.3). *Mittlere Kurve*: Zunahme der Zahl der Zelltypen in der Evolution der Metazoa (Daten aus Abbildung 4.8). *Obere Kurve*: Zunahme des relativen Gehirngewichts (bezogen auf die Körperoberfläche) bei Säugetieren (Daten aus Wilson 1985). Für die Abszisse wurden zwei Skaleneinteilungen verwendet, eine für den Zeitraum $>10^9$ Jahre, eine andere für den Zeitraum $<10^9$ Jahre vor der Gegenwart. Oberhalb der Abszisse sind die Namen einiger wichtiger taxonomischer Einheiten angeführt, deren Evolution in etwa beim jeweiligen Wortbeginn einsetzt.

Wolfgang Wieser. Die Erfindung der Individualität oder die zwei Gesichter der Evolution. Spektrum Akademischer Verlag, Heidelberg 1998.

A.C.Wilson. The Molecular Basis of Evolution. Scientific American, Oct.1985, 164-173.



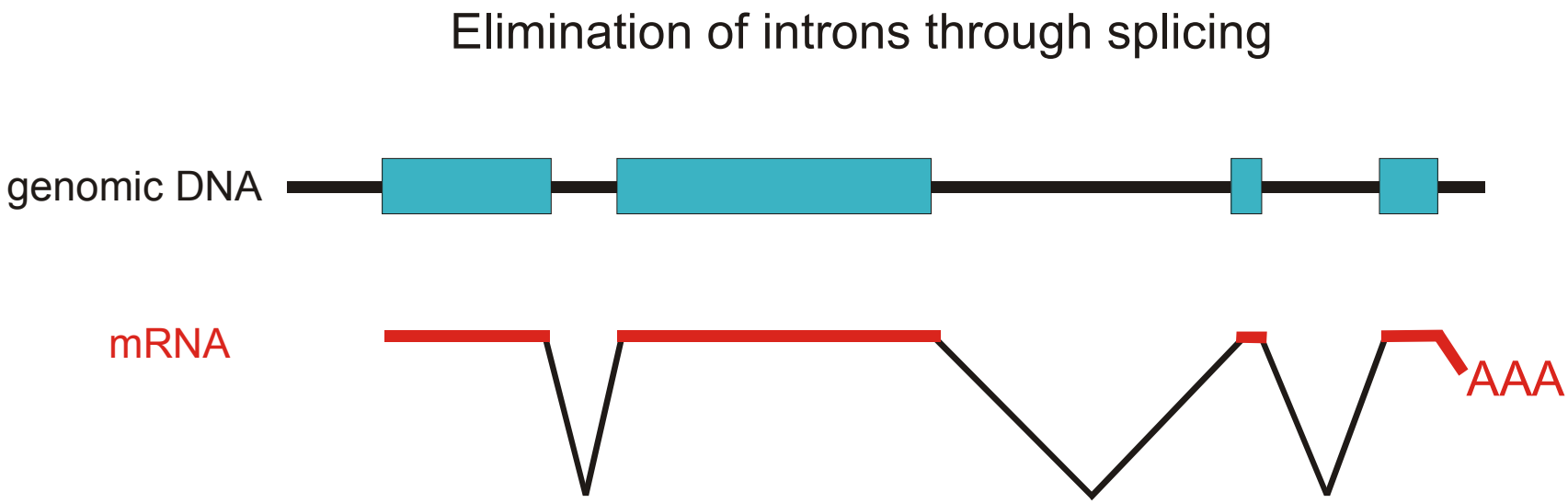
The gene is a stretch of DNA which after transcription gives rise to a mRNA



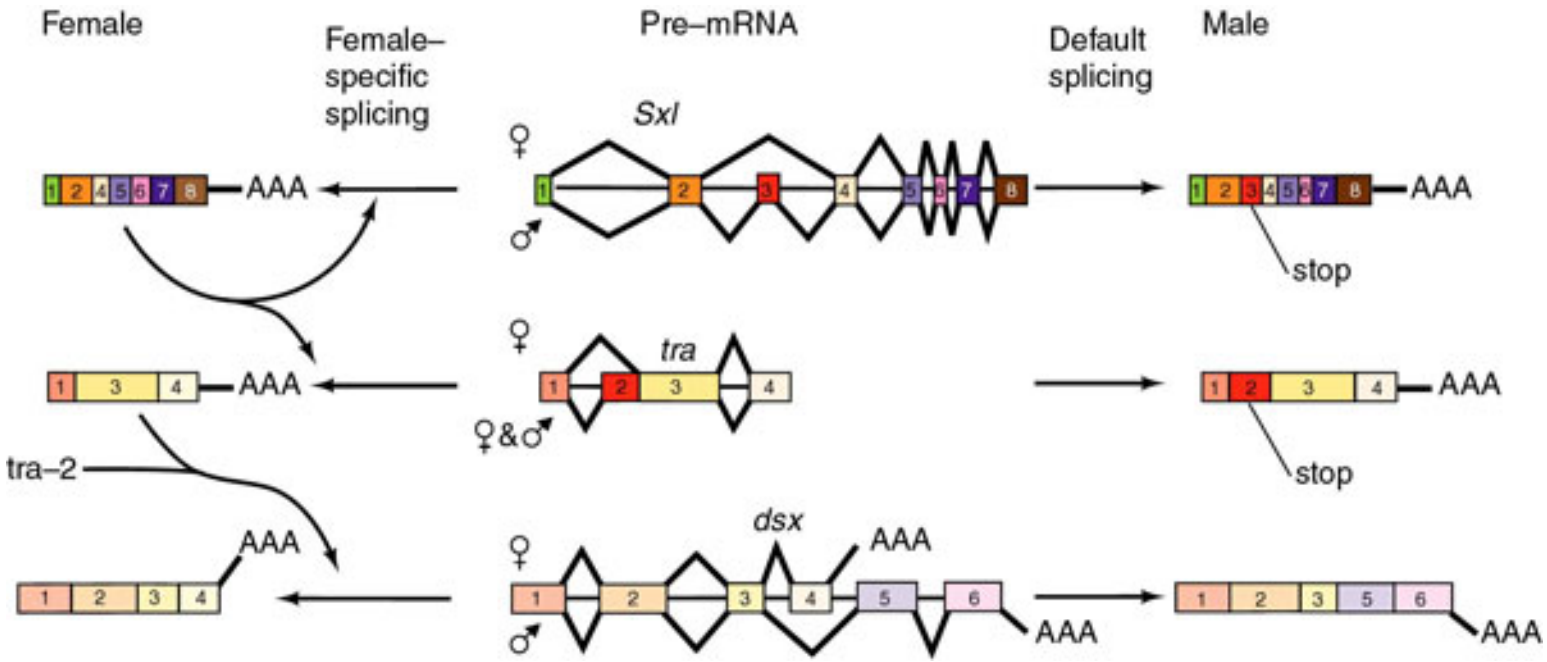
The same section of the microarray is shown in three independent hybridizations. Marked spots refer to: (1) protein disulfide isomerase related protein P5, (2) IL-8 precursor, (3) EST AA057170, and (4) vascular endothelial growth factor

Gene expression DNA microarray representing 8613 human genes used to study transcription in the response of human fibroblasts to serum

V.R.Iyer *et al*., *Science* **283**: 83-87, 1999

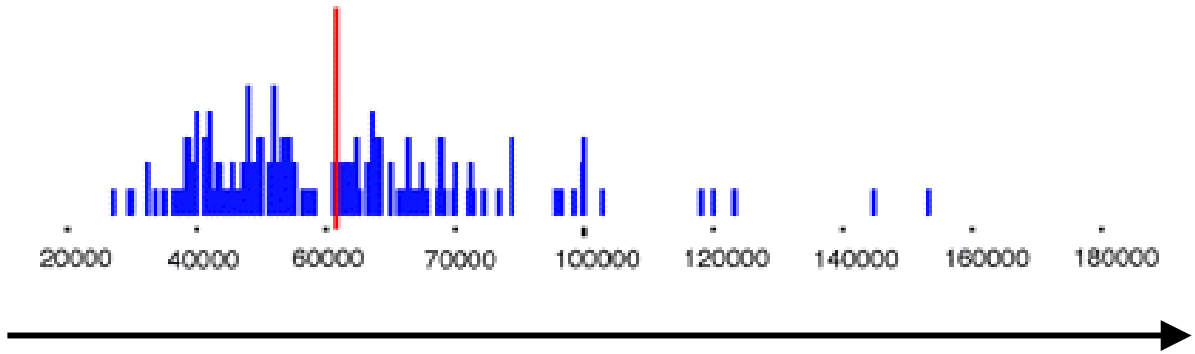


The gene is a stretch of DNA which after transcription and processing gives rise to a mRNA



Sex determination in ***Drosophila*** through alternative splicing

The process of protein synthesis and its regulation is now understood but the notion of the gene as a stretch of DNA has become obscure. The gene is essentially associated with the sequence of unmodified amino acids in a protein, and it is determined by the nucleotide sequence as well as the dynamics of the the process eventually leading to the m-RNA that is translated.

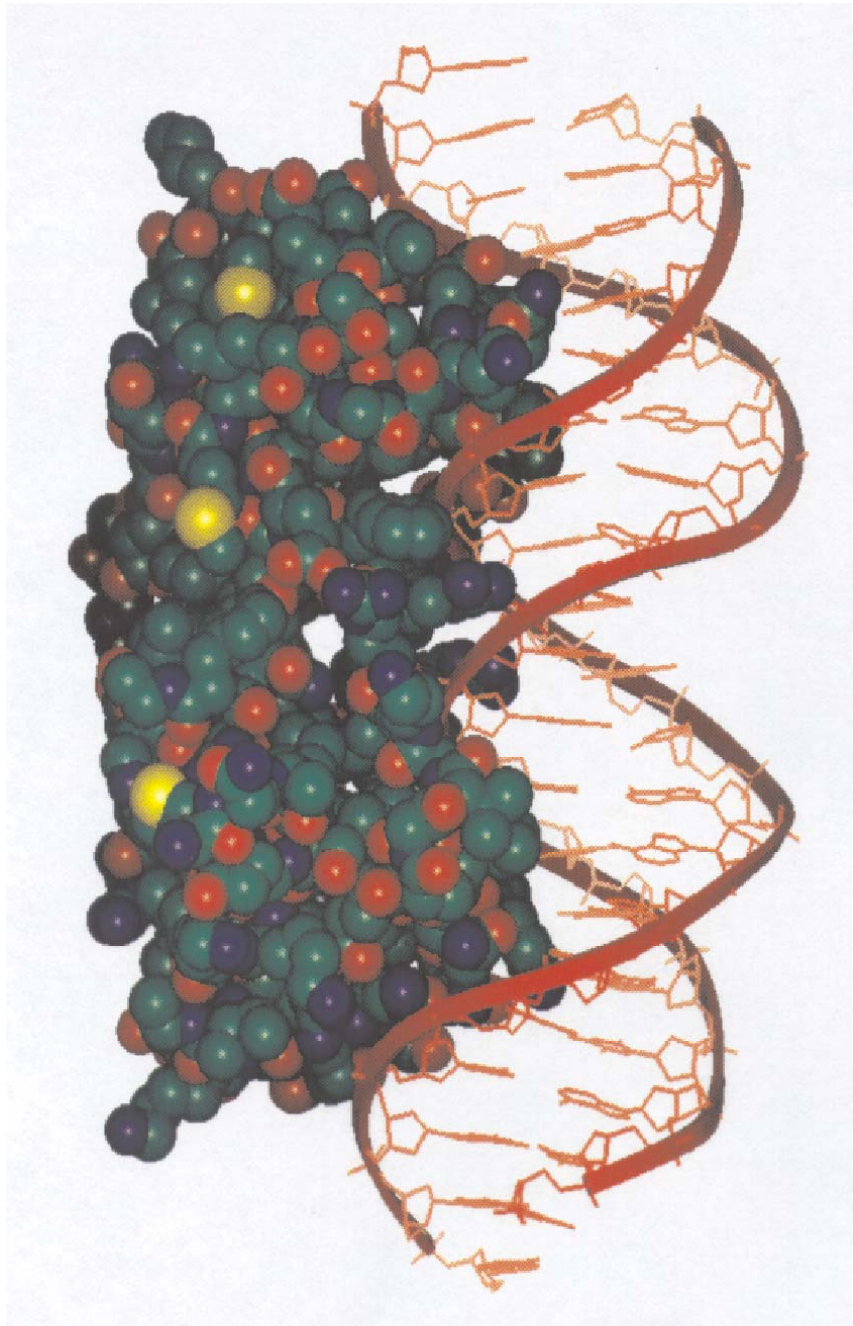


Number of genes in the human genome

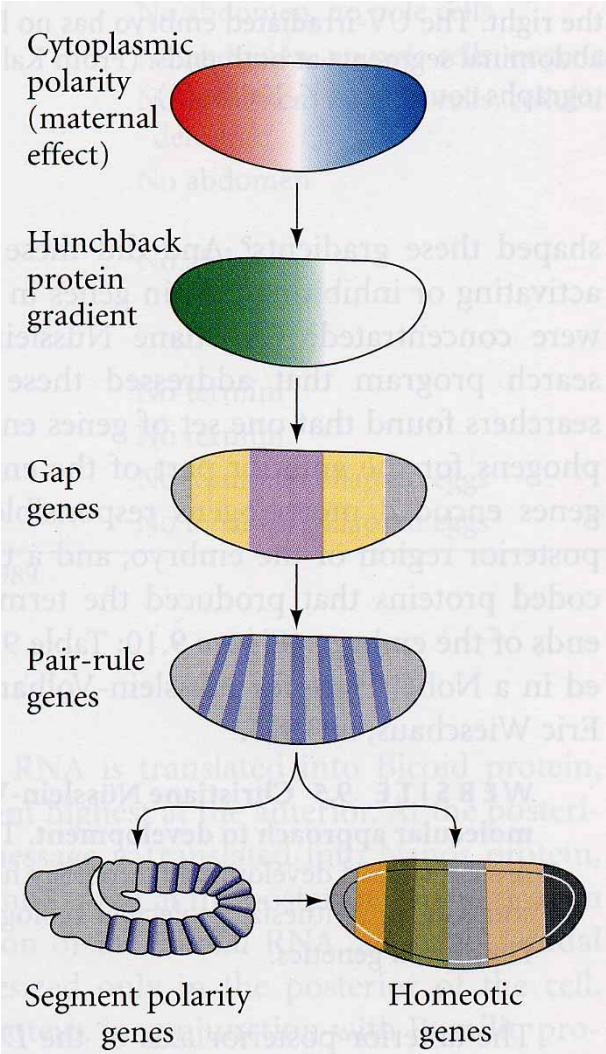
The number of genes in the human genome is still only a very rough estimate

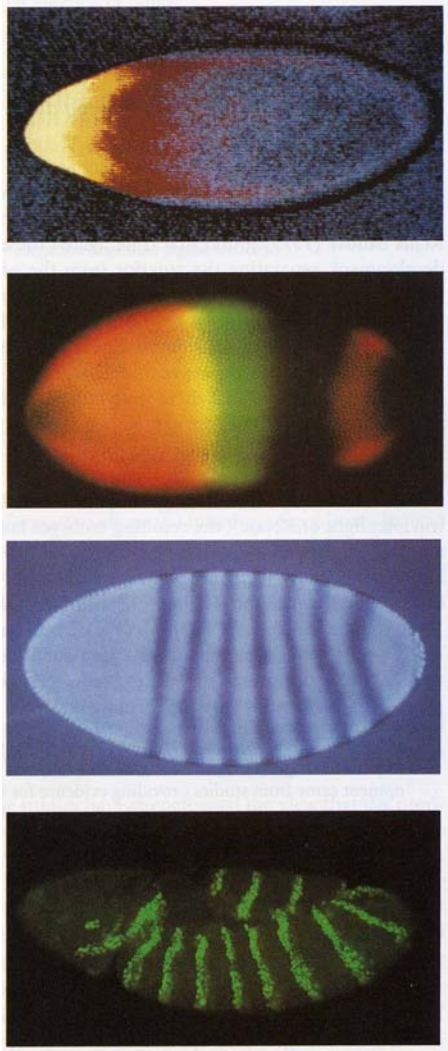
Developmental biology

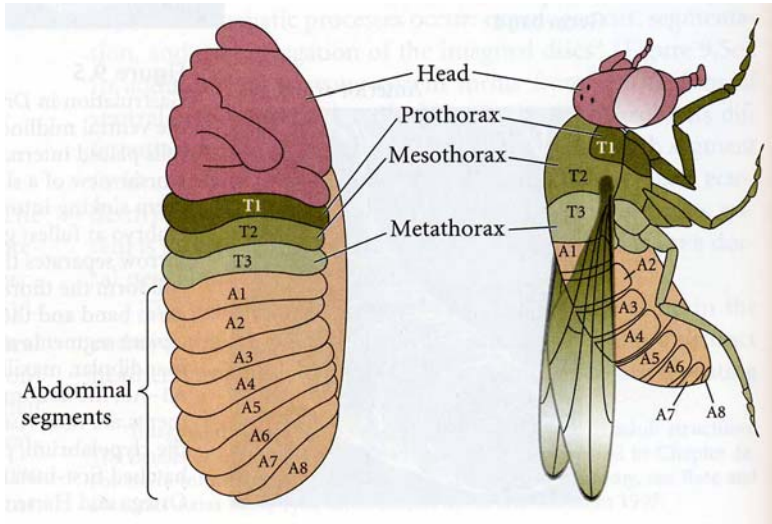
Gene regulation networks, signal propagation, pattern formation, robustness ...



Three-dimensional structure of the complex between the regulatory protein **cro-repressor** and the binding site on λ-phage **B-DNA**

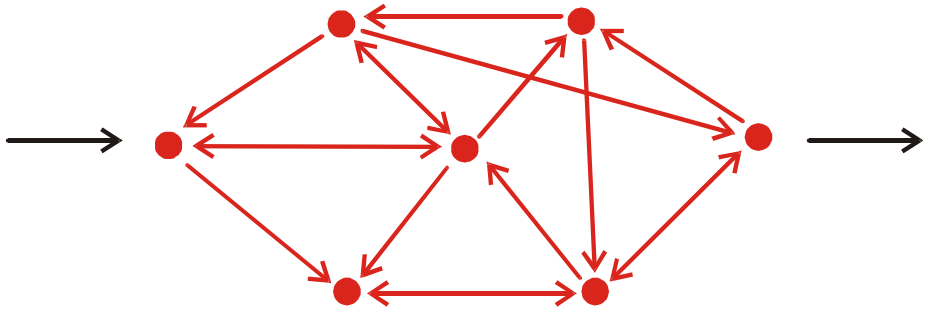






Development of the fruit fly *drosophila melanogaster*: Genetics, experiment, and imago

Linear chain



Network

Processing of information in cascades and networks

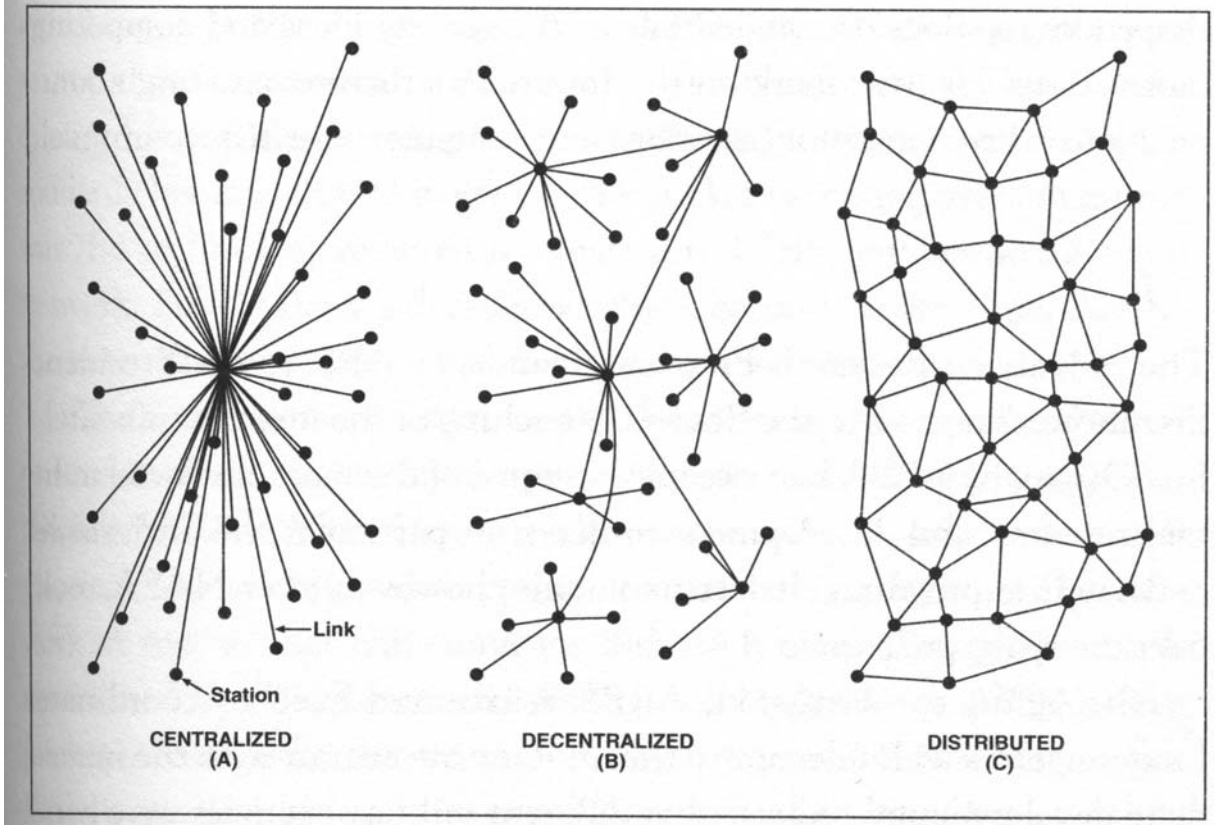


Figure 11.1 Paul Baran's Networks. *In 1964, Paul Baran began thinking about the optimal structure of the Internet. He suggested that there were three possible architectures for such a network—centralized, decentralized, and distributed—and warned that both the centralized and decentralized structures that dominated communications systems of the time were too vulnerable to attack. Instead, he proposed that the Internet should be designed to have a distributed, mesh-like architecture. (Reproduced with permission of Paul Baran.)*

Albert-László Barabási, Linked – The New Science of Networks. Perseus Publ., Cambridge, MA, 2002

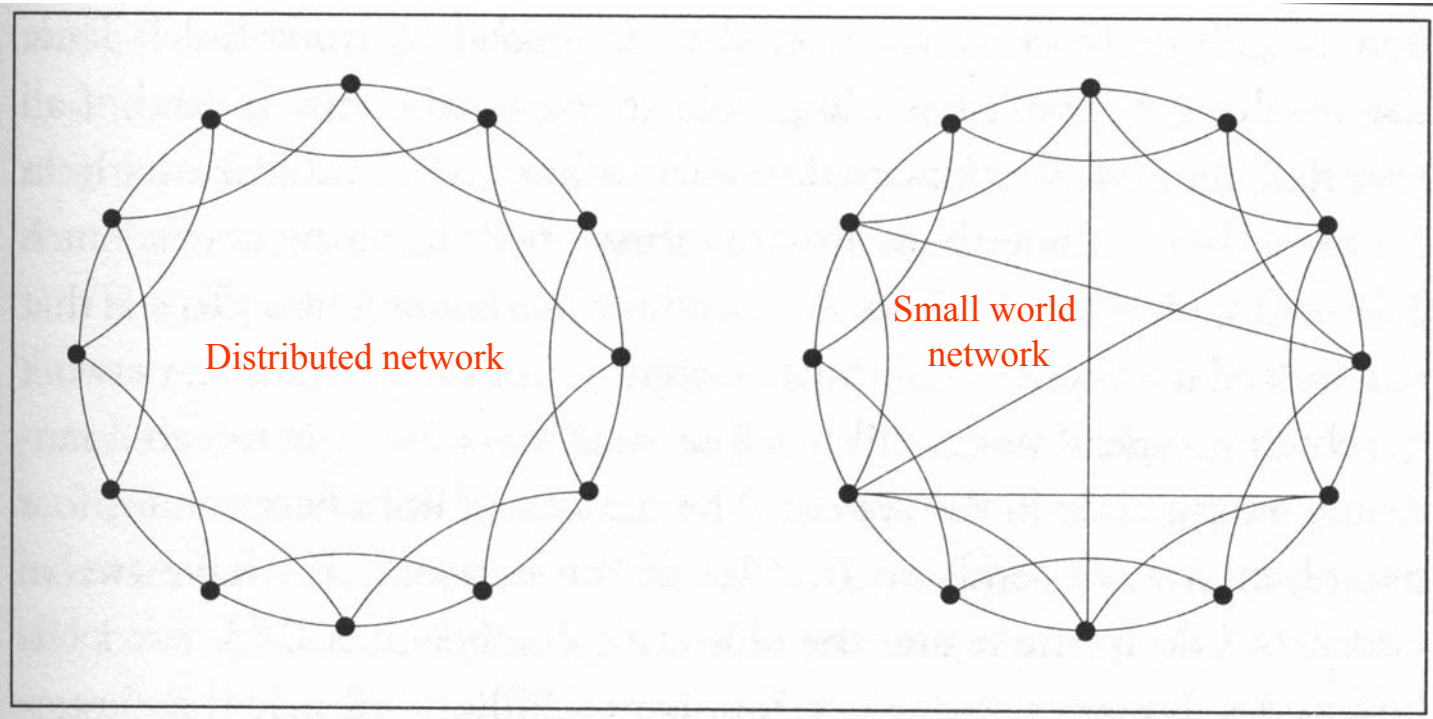


Figure 4.2 A Small and Clustered World. *To model networks with a high degree of clustering, Duncan Watts and Steven Strogatz started from a circle of nodes, where each node is connected to its immediate and next-nearest neighbors (left). To make this world a small one, a few extra links were added, connecting randomly selected nodes (right). These long-range links offer the crucial shortcuts between distant nodes, drastically shortening the average separation between all nodes.*

Albert-László Barabási, Linked – The New Science of Networks. Perseus Publ., Cambridge, MA, 2002

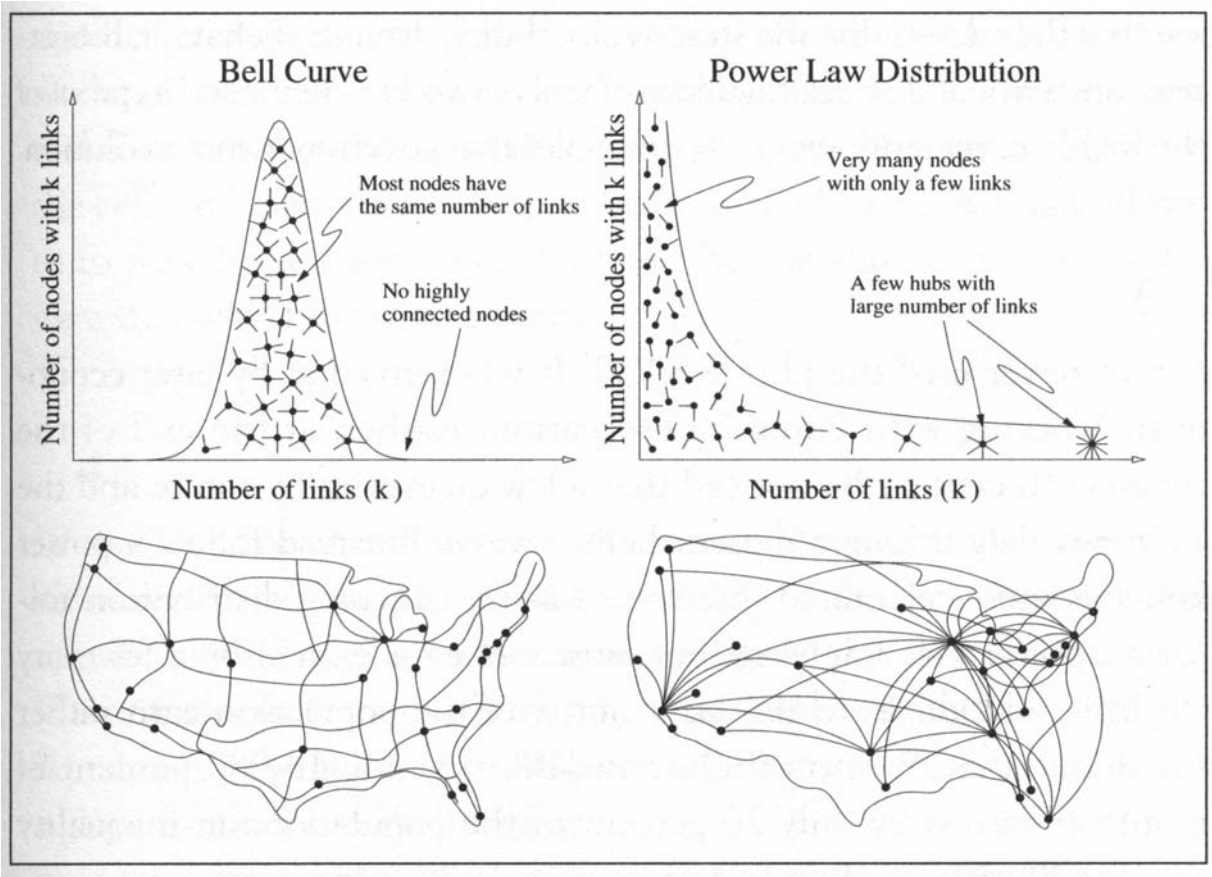
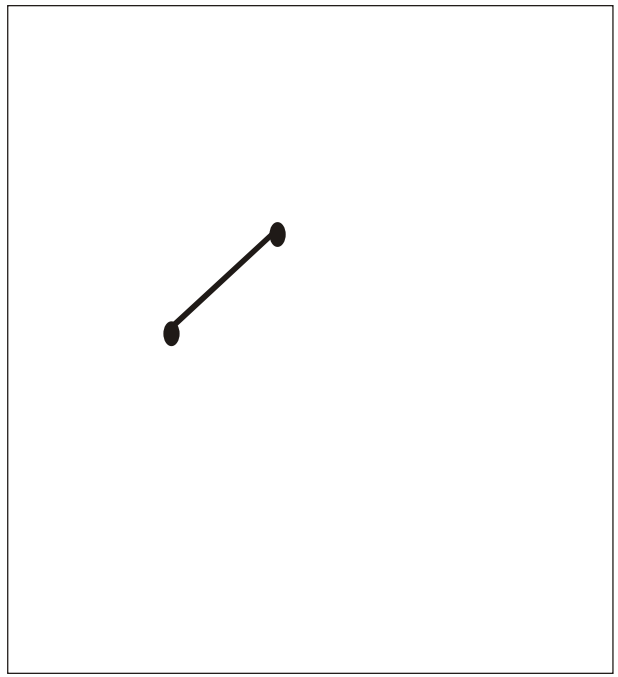
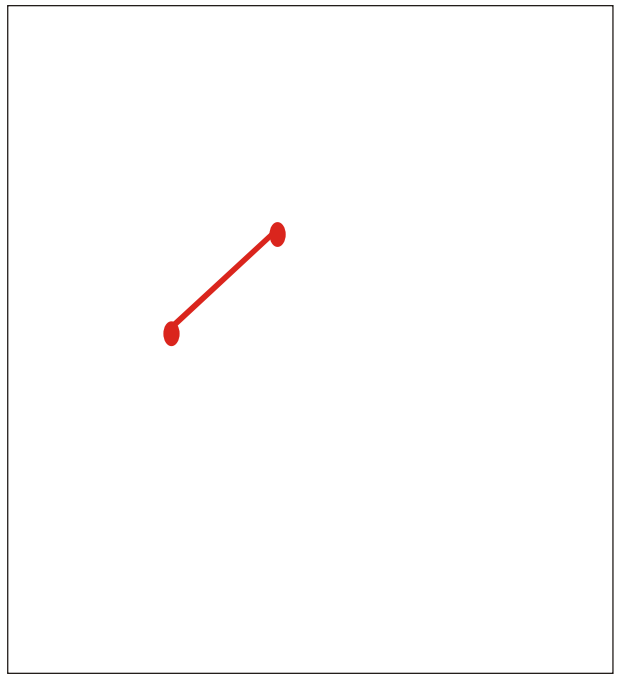


Figure 6.1 Random and Scale-Free Networks. *The degree distribution of a random network follows a bell curve, telling us that most nodes have the same number of links, and nodes with a very large number of links don't exist (top left). Thus a random network is similar to a national highway network, in which the nodes are the cities, and the links are the major highways connecting them. Indeed, most cities are served by roughly the same number of highways (bottom left). In contrast, the power law degree distribution of a scale-free network predicts that most nodes have only a few links, held together by a few highly connected hubs (top right). Visually this is very similar to the air traffic system, in which a large number of small airports are connected to each other via a few major hubs (bottom right).*

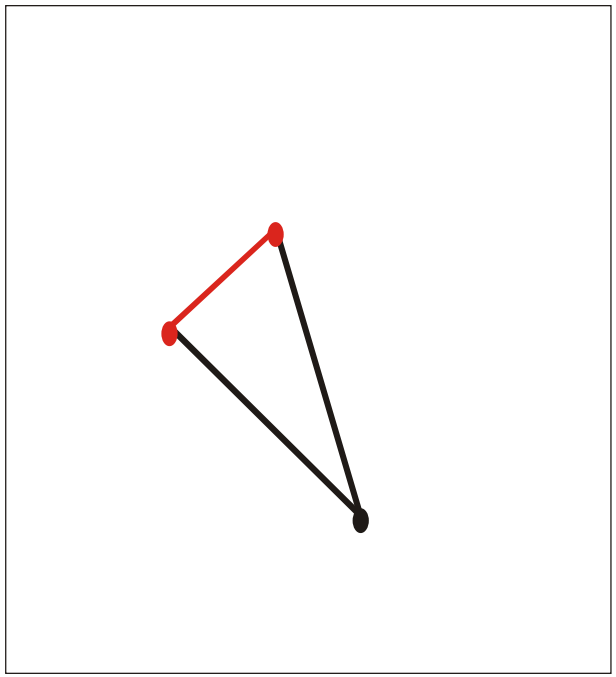
Albert-László Barabási,
Linked – The New Science of Networks
Perseus Publ., Cambridge, MA, 2002



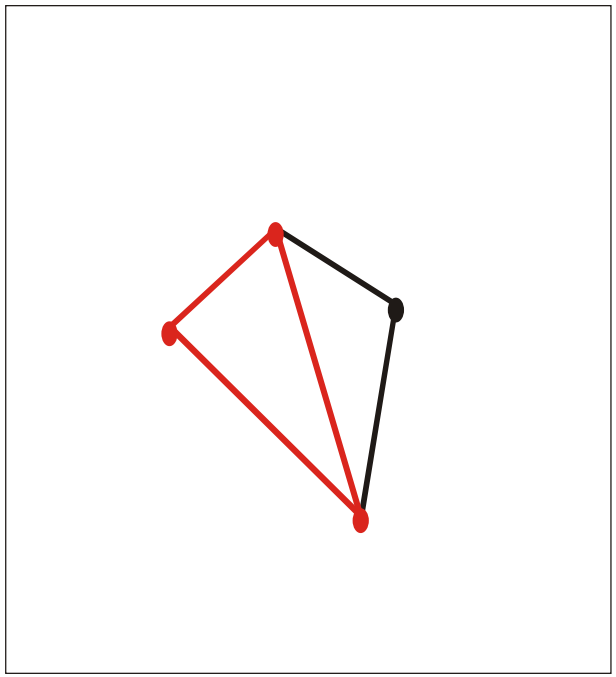
Formation of a scale-free network through evolutionary point by point expansion: Step 000



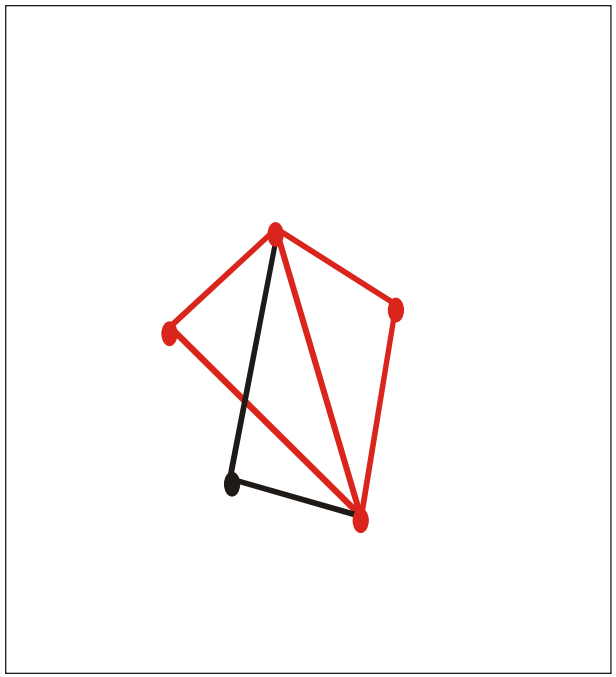
Formation of a scale-free network through evolutionary point by point expansion: Step 001



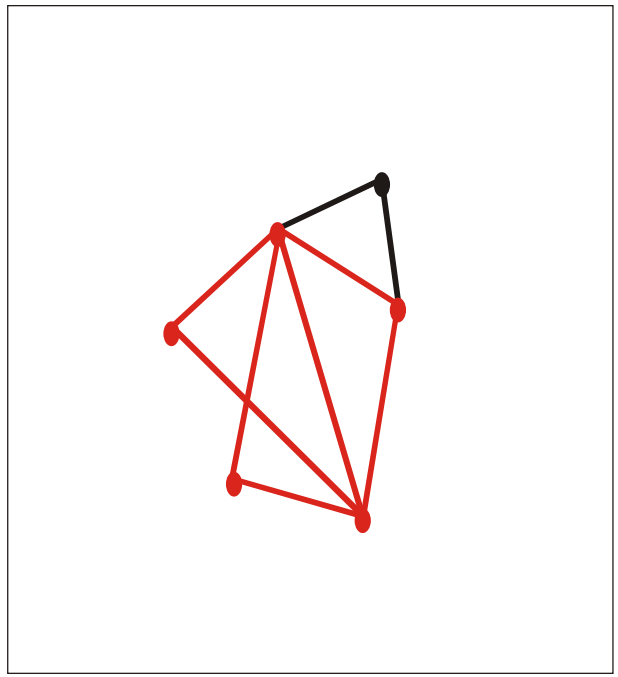
Formation of a scale-free network through evolutionary point by point expansion: Step 002



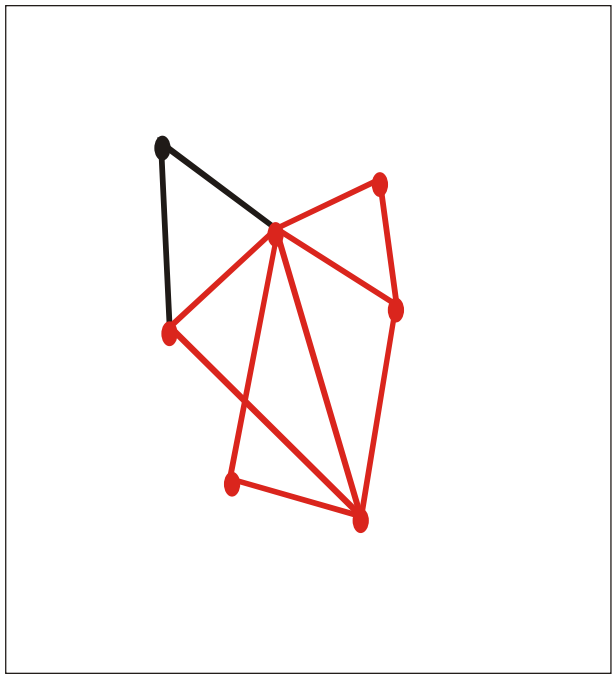
Formation of a scale-free network through evolutionary point by point expansion: Step 003



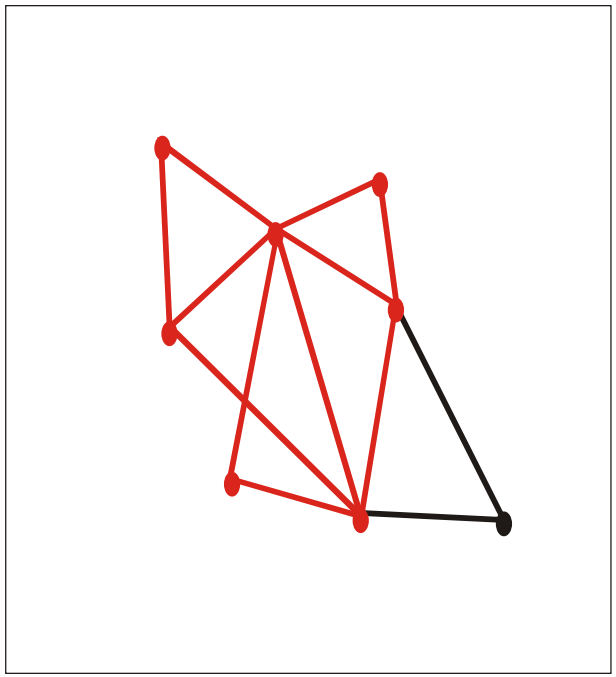
Formation of a scale-free network through evolutionary point by point expansion: Step 004



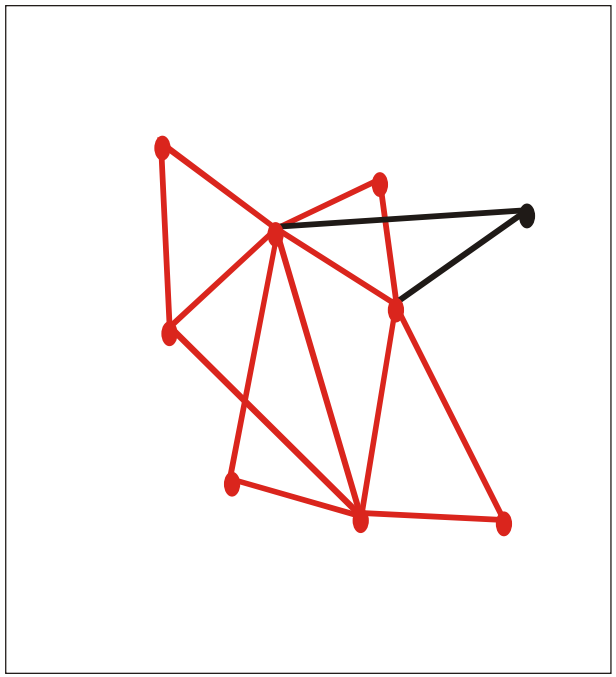
Formation of a scale-free network through evolutionary point by point expansion: Step 005



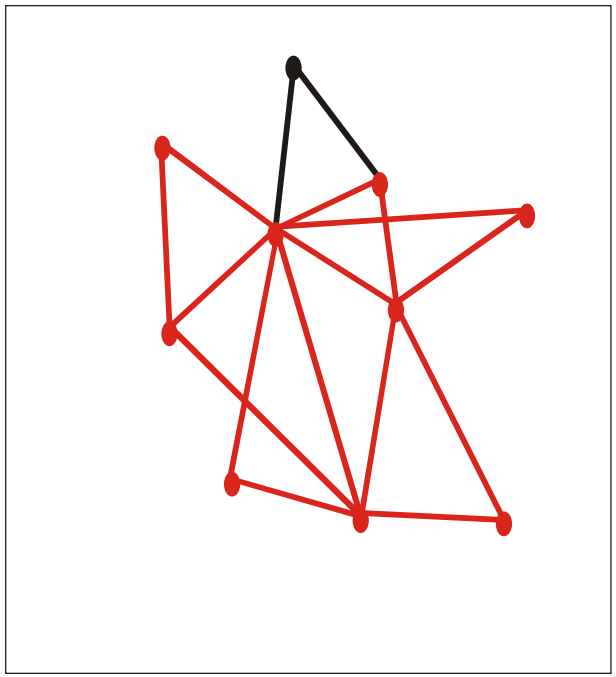
Formation of a scale-free network through evolutionary point by point expansion: Step 006



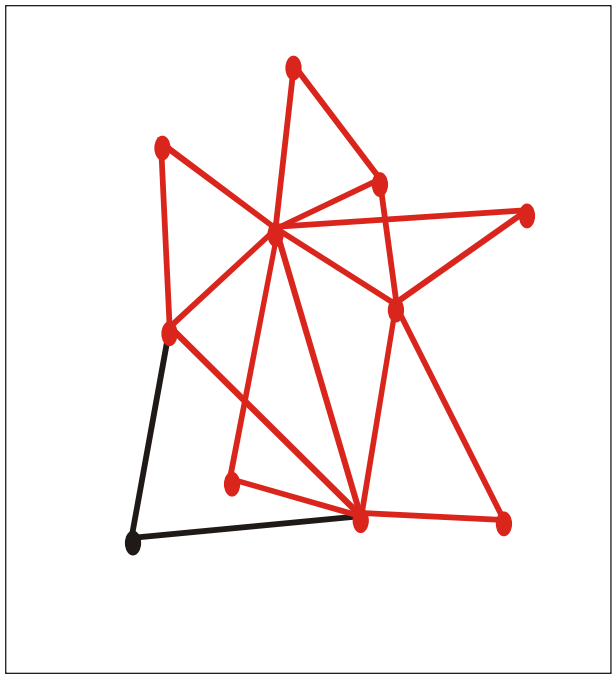
Formation of a scale-free network through evolutionary point by point expansion: Step 007



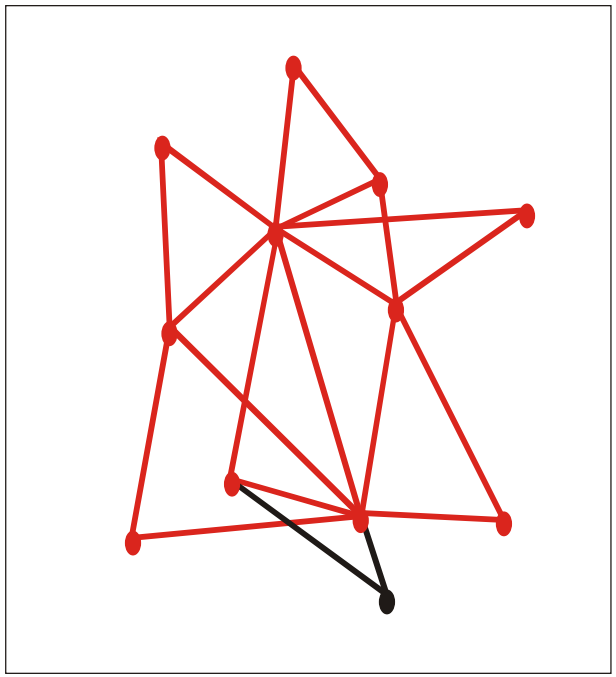
Formation of a scale-free network through evolutionary point by point expansion: Step 008



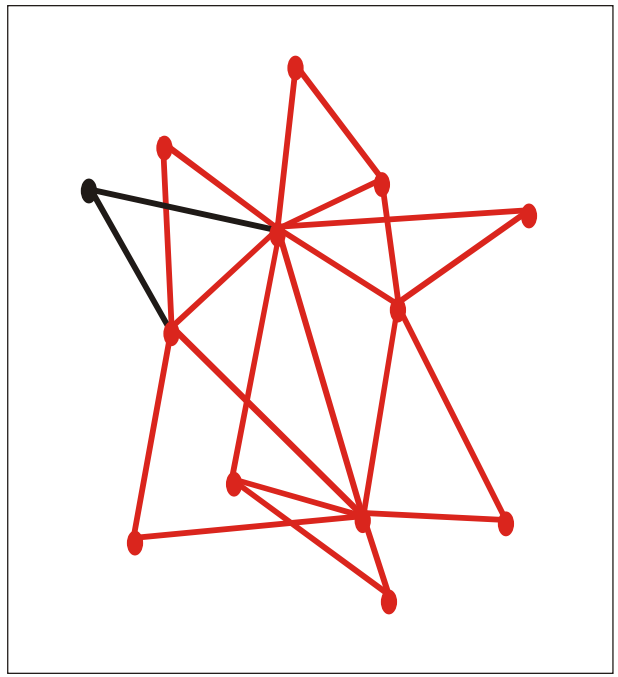
Formation of a scale-free network through evolutionary point by point expansion: Step 009



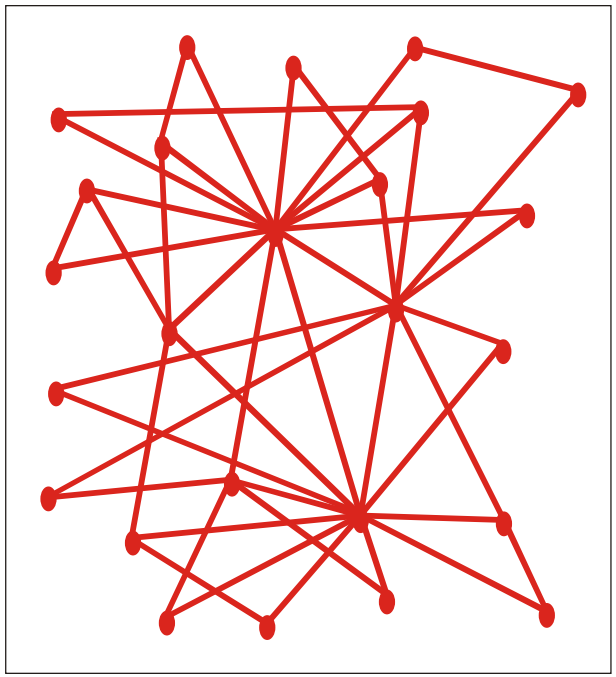
Formation of a scale-free network through evolutionary point by point expansion: Step 010



Formation of a scale-free network through evolutionary point by point expansion: Step 011

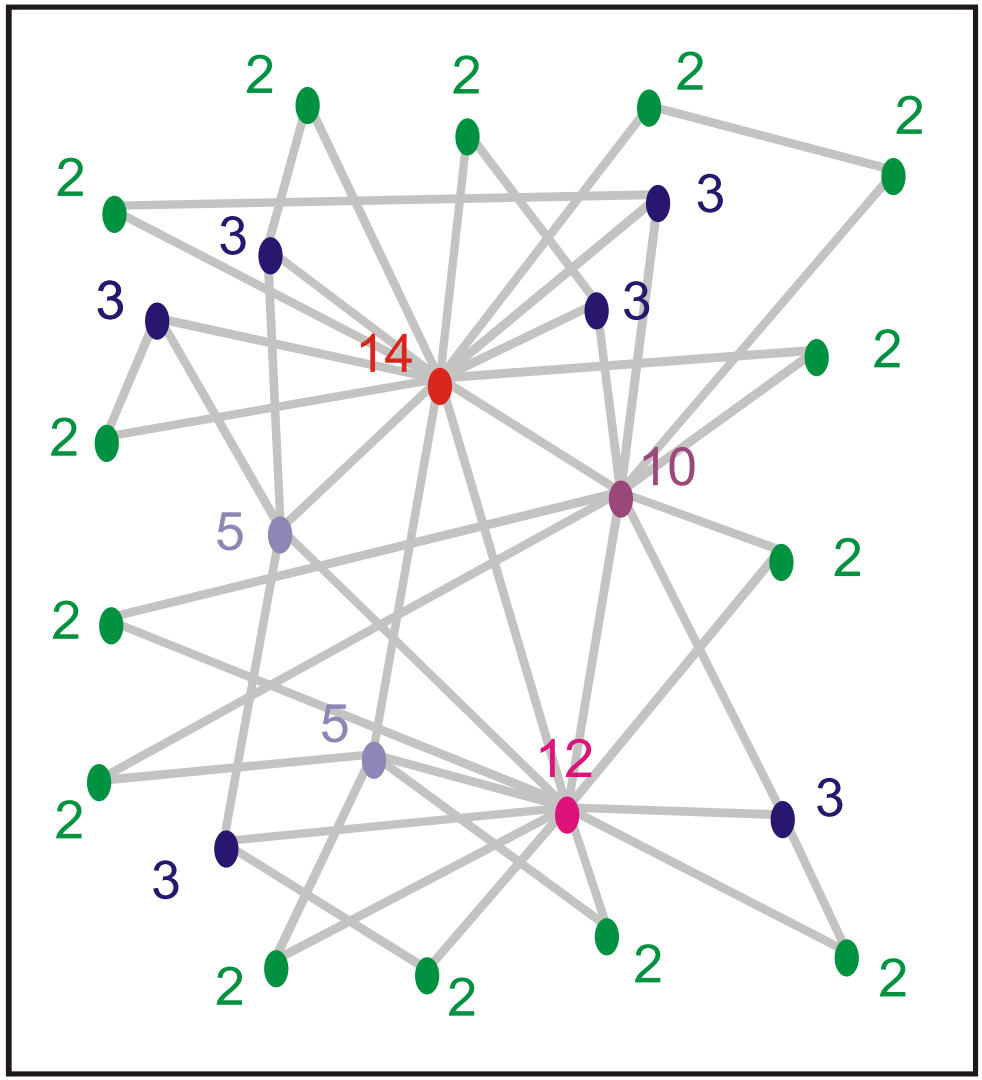


Formation of a scale-free network through evolutionary point by point expansion: Step 012



Formation of a scale-free network through evolutionary point by point expansion: Step 024

links	# nodes
2	14
3	6
5	2
10	1
12	1
14	1



Analysis of nodes and links in a step by step evolved network

Structures in **Directed Networks**

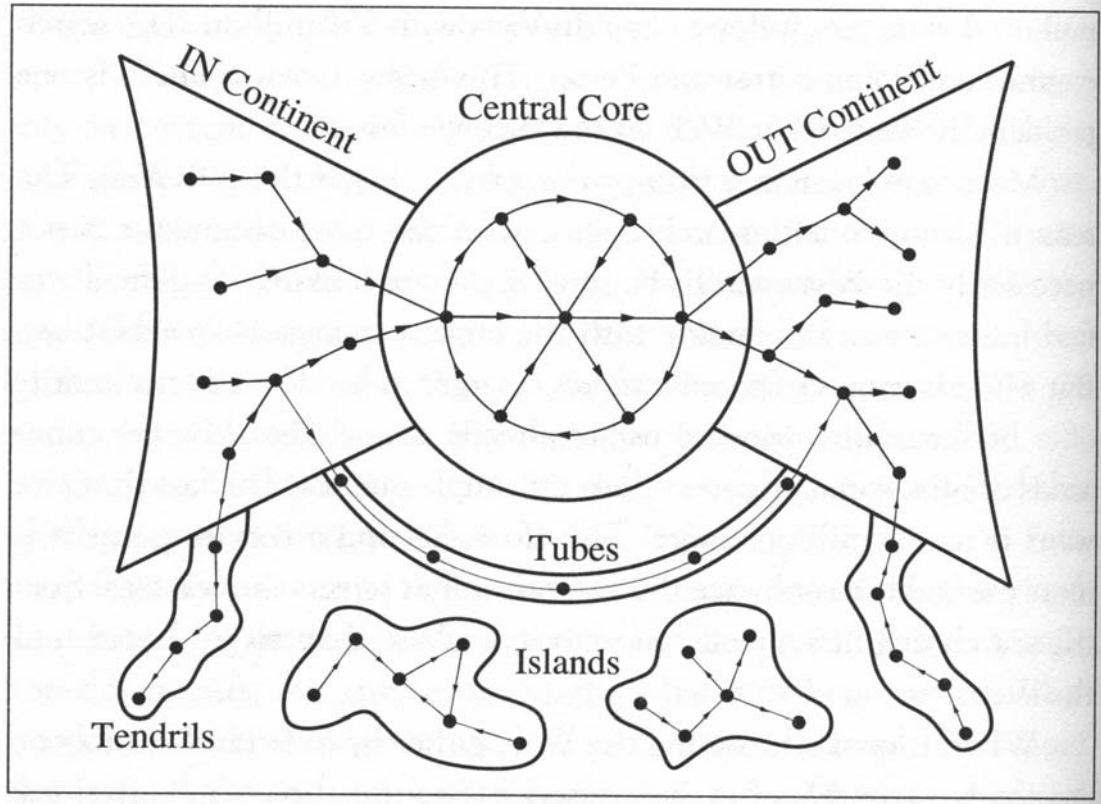


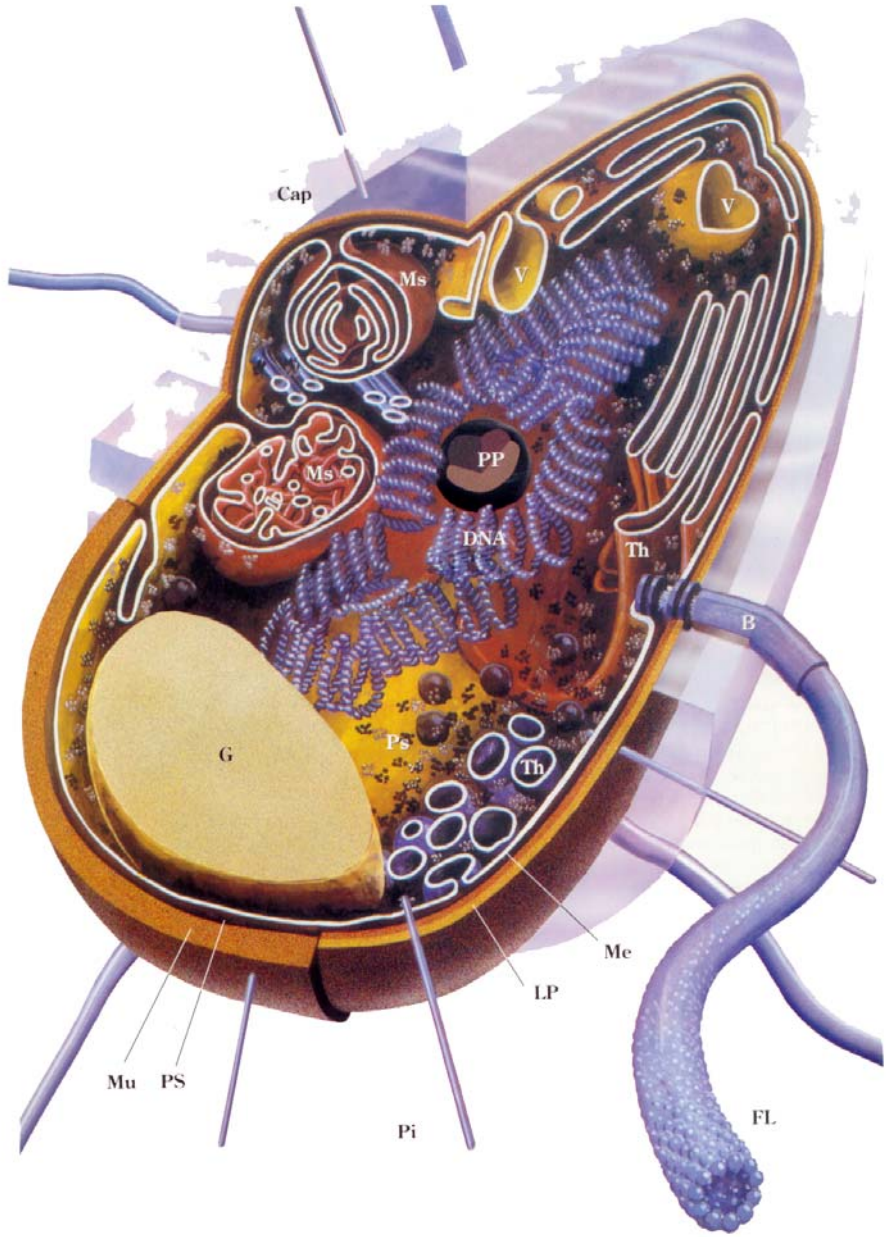
Figure 12.1 The Continents of a Directed Network. *Directed networks such as the World Wide Web naturally break down into several easily identifiable continents. In the central core each node can be reached from every other node. Nodes in the IN continent are arranged such that following the links eventually brings you back to the central core, but starting from the core doesn't allow you to return to the IN continent. In contrast, all nodes of the OUT continent can be reached from the core, but once you've arrived, there are no links taking you back to the core. Finally, tubes directly connect the IN to the OUT continent; some nodes form tendrils, attached only to the IN and OUT continents; and a few nodes form isolated islands that can't be accessed from the rest of the nodes.*

Albert-László Barabási,
Linked – The New Science of Networks.
Perseus Publ., Cambridge, MA, 2002

Cell biology

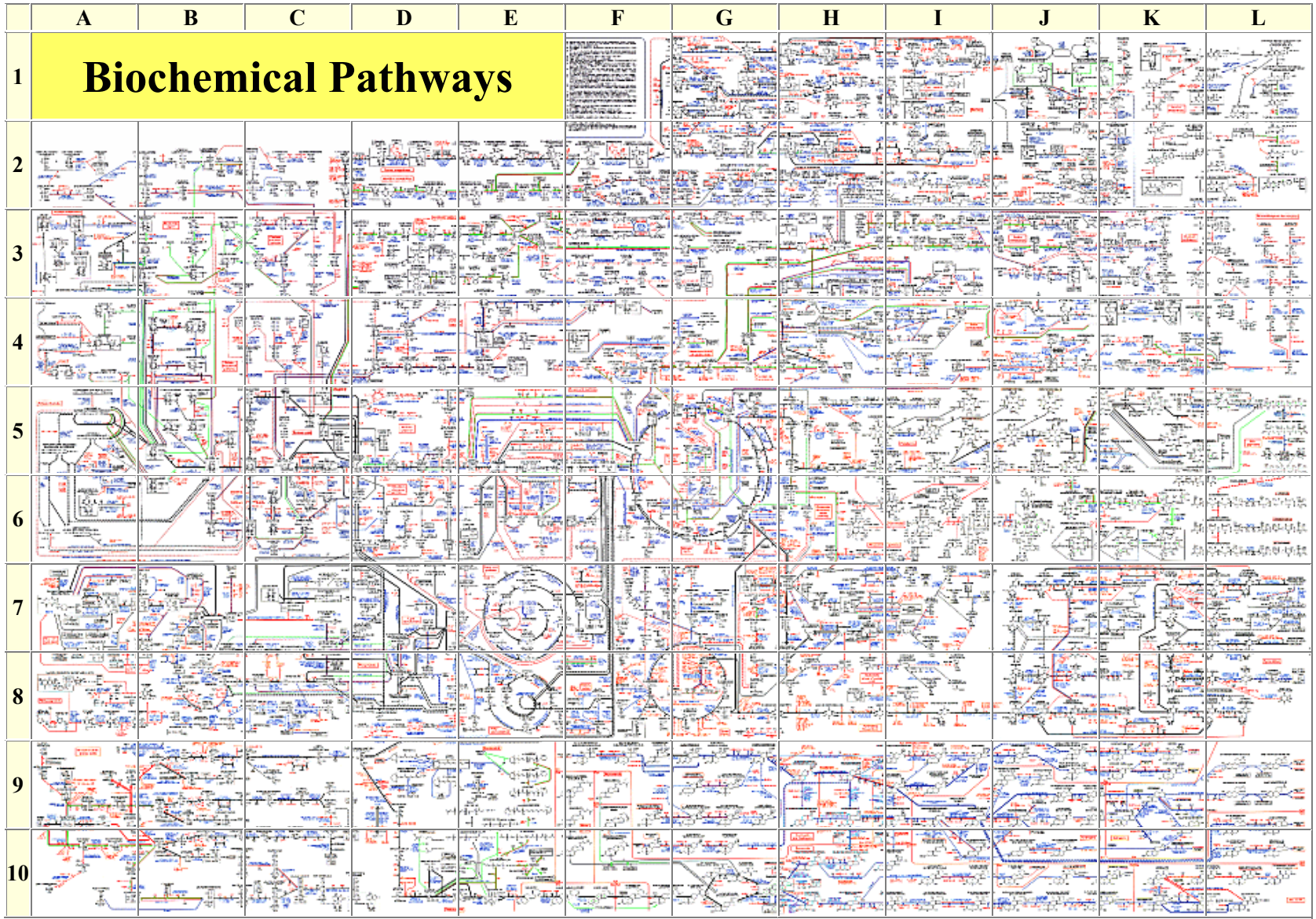
Regulation of cell cycle, metabolic networks, reaction kinetics, homeostasis, ...

The bacterial cell as an example for the simplest form of autonomous life



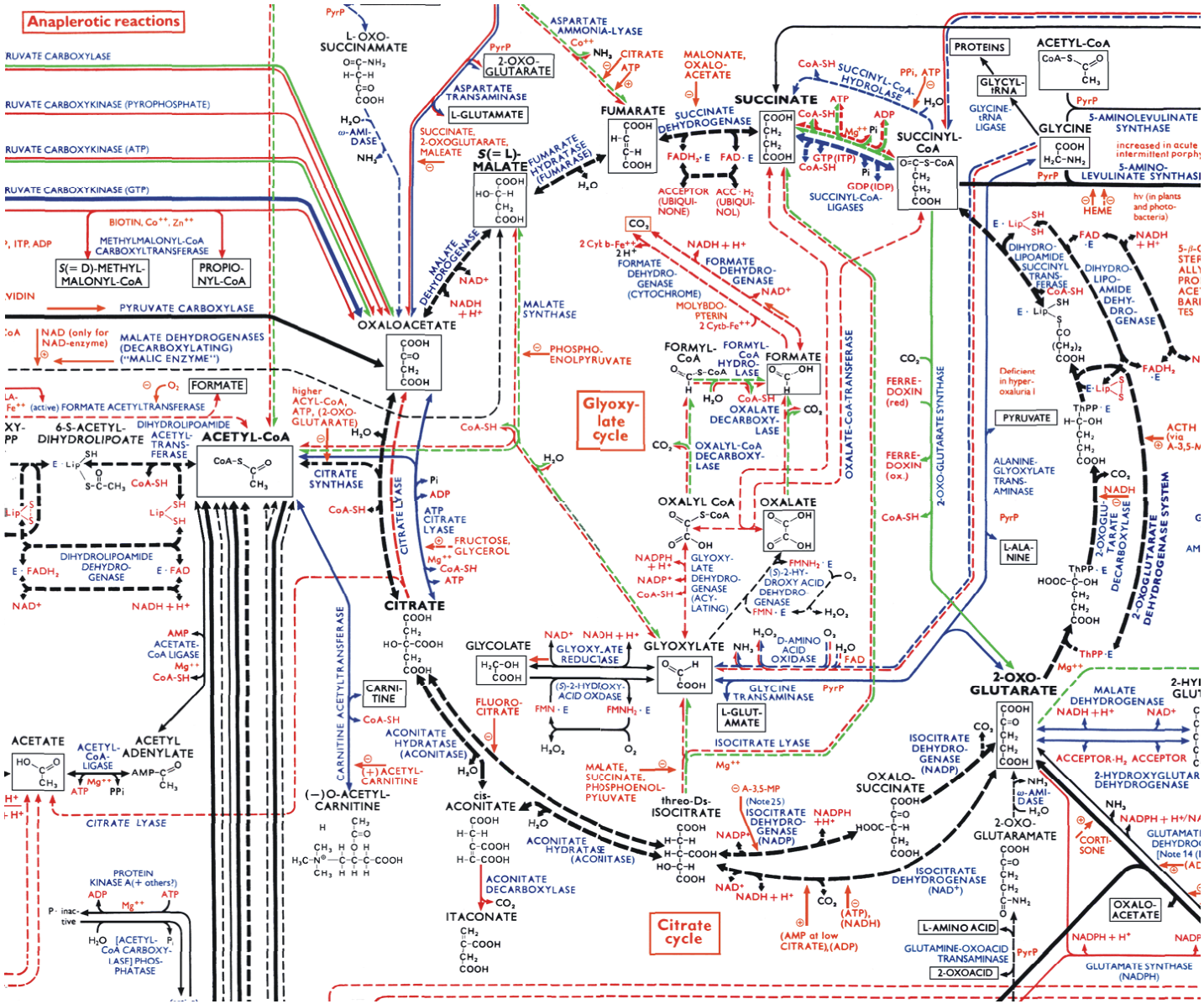
The human body:

10^{14} cells = 10^{13} eukaryotic cells + Î 9Ç10^{13} bacterial (prokaryotic) cells, and Î 200 eukaryotic cell types



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

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



Kinetic differential equations

$$\frac{d\,x_i}{d\,t} = f(x_1, x_2, \ldots, x_n; k_1, k_2, \ldots, k_m);\ i = 1, 2, \ldots, n$$

Reaction diffusion equations

$$\frac{\partial x_i}{\partial t} = D_i \nabla^2 x_i + f(x_1, x_2, \ldots, x_n; k_1, k_2, \ldots, k_m);\ i = 1, 2, \ldots, n$$

Parameter set

$$k_j(T, p, pH, I, \ldots; x_1, x_2, \ldots, x_n);\ j = 1, 2, \ldots, m$$

General conditions: T , p , pH , I , ...

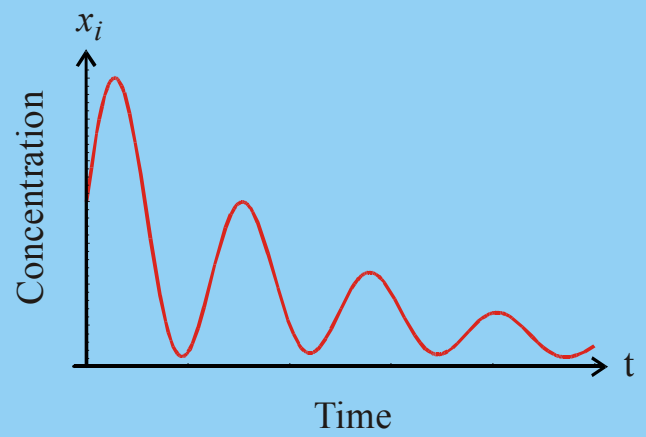
Initial conditions: $x_i(0);\ i = 1, 2, \ldots, n$

Boundary conditions: boundary ... $\vec{s}$
normal unit vector ... $\vec{u}$

Dirichlet , $x_i^{\vec{s}} = f(\vec{r}, t);\ i = 1, 2, \ldots, n$

Neumann , $\frac{\partial x_i}{\partial u} = \hat{u} \cdot \nabla x_i^{\vec{s}} = f(\vec{r}, t);\ i = 1, 2, \ldots, n$

Solution curves: $x_i(t);\ i = 1, 2, \ldots, n$



The forward-problem of chemical reaction kinetics

Kinetic differential equations

$$\frac{dx_i}{dt} = f(x_1, x_2, \ldots, x_n; k_1, k_2, \ldots, k_m);\ i = 1, 2, \ldots, n$$

Reaction diffusion equations

$$\frac{\partial x_i}{\partial t} = D_i \nabla^2 x_i + f(x_1, x_2, \ldots, x_n; k_1, k_2, \ldots, k_m);\ i = 1, 2, \ldots, n$$

General conditions: T, p, pH, I, ...

Initial conditions: $x_i(0);\ i = 1, 2, \ldots, n$

Boundary conditions: boundary ... $\hat{s}$
normal unit vector ... $\hat{u}$

Dirichlet, $x_i^{\hat{s}} = f(\vec{r}, t);\ i = 1, 2, \ldots, n$

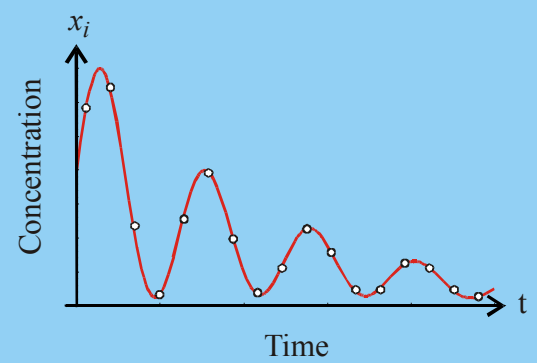
Neumann, $\frac{\partial x_i}{\partial u} = \hat{u} \cdot \nabla x_i^{\hat{s}} = f(\vec{r}, t);\ i = 1, 2, \ldots, n$

Parameter set

$$k_j(T, p, pH, I, \ldots; x_1, x_2, \ldots, x_n);\ j = 1, 2, \ldots, m$$

Data from measurements

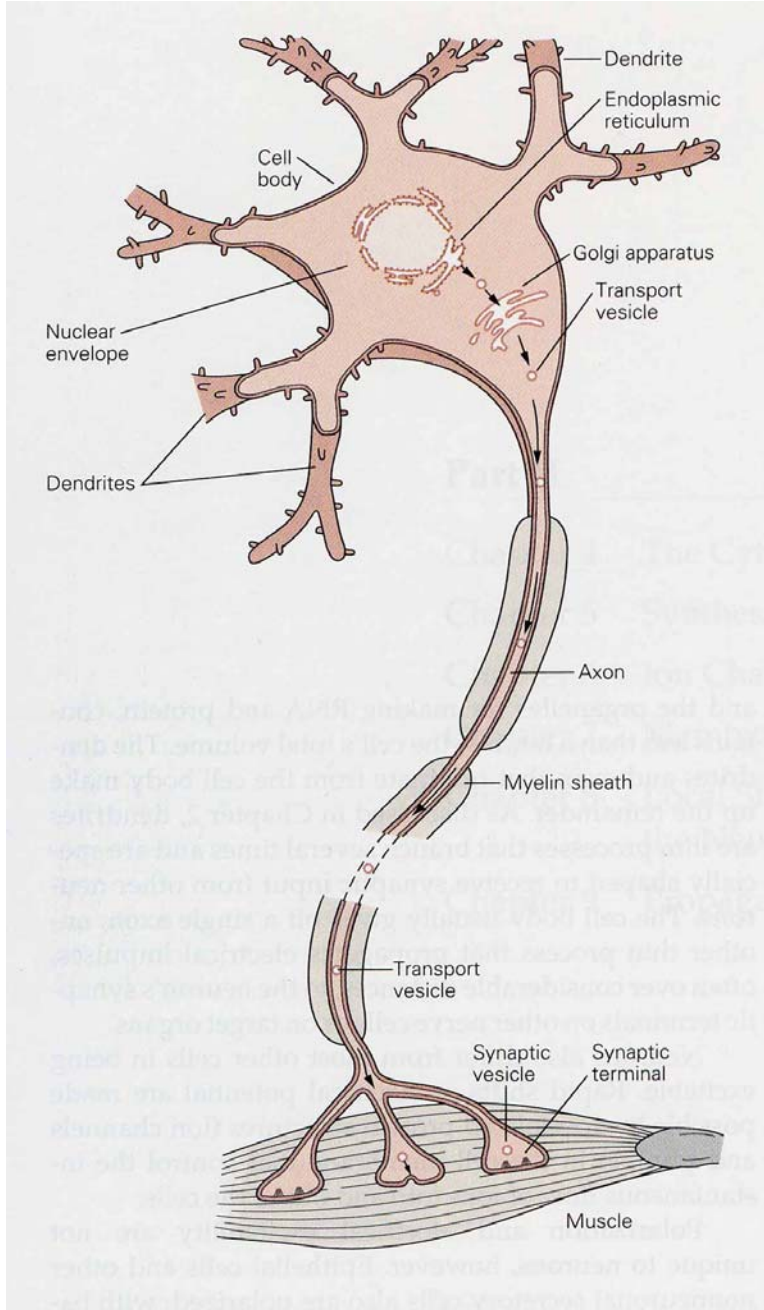
$$x_i(t_k);\ i = 1, 2, \ldots, n;\ k = 1, 2, \ldots, N$$



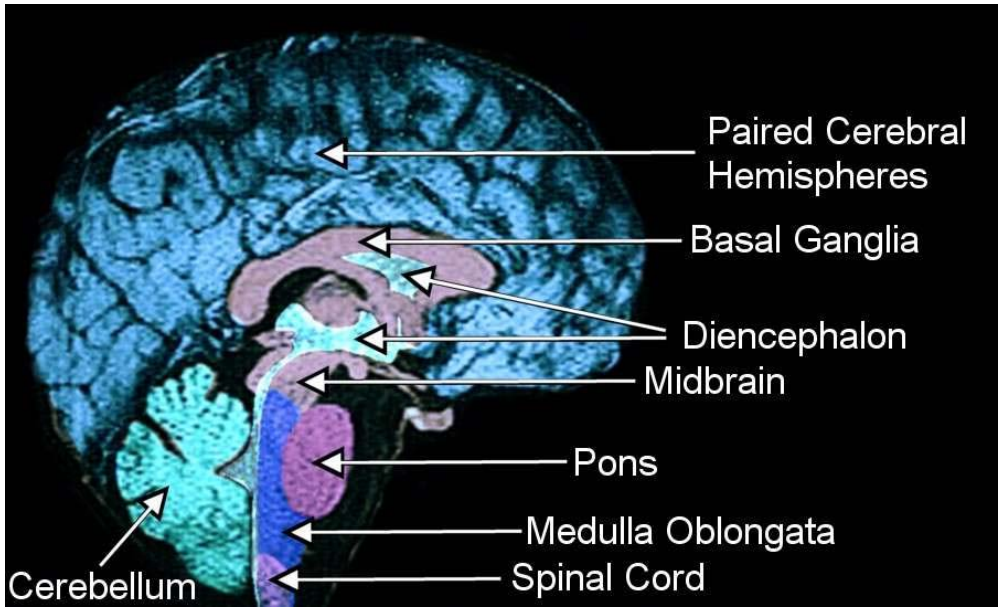
The inverse-problem of chemical reaction kinetics

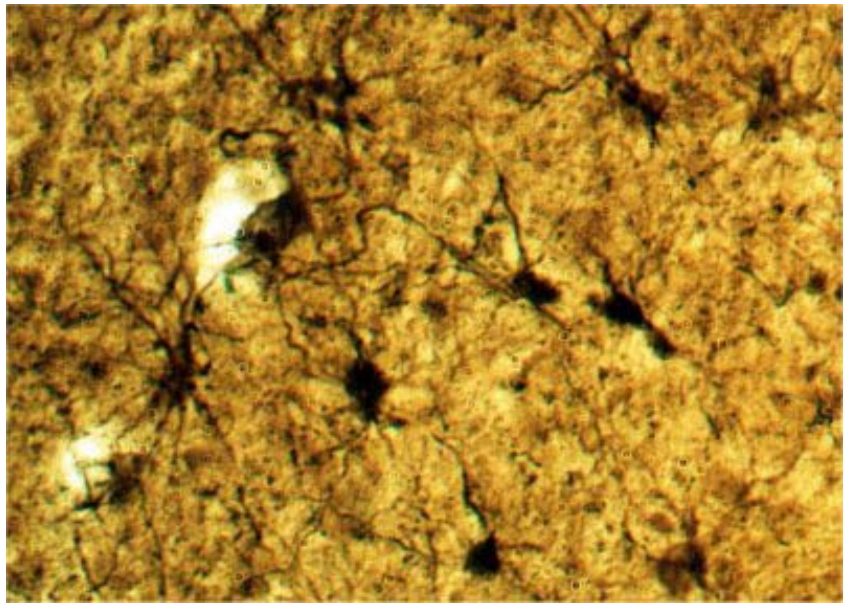
Neurobiology

Neural networks, collective properties, nonlinear dynamics, signalling, ...



A single neuron signaling to a muscle fiber





The human brain

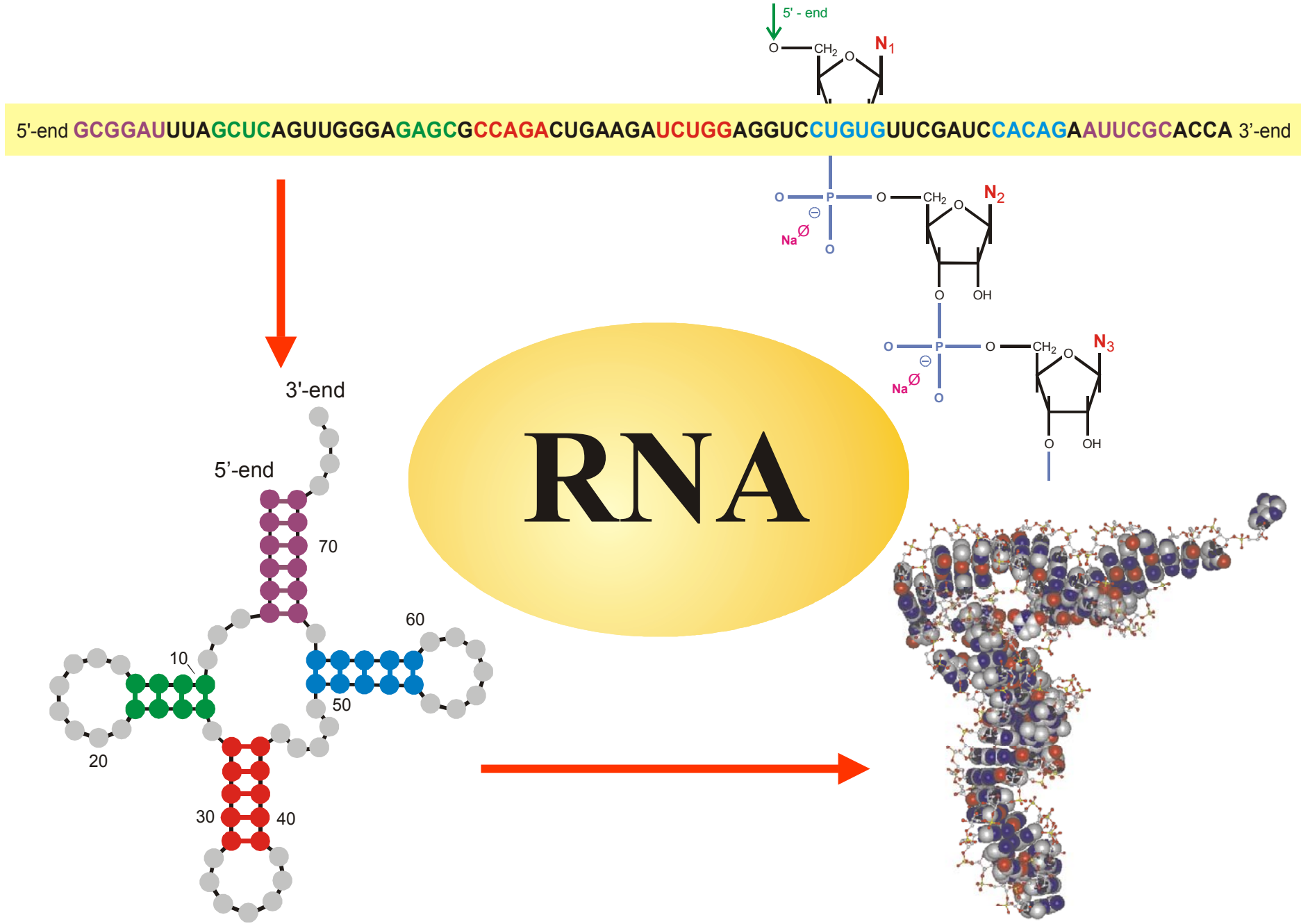
10^{11} neurons connected by Î 10^{13} to 10^{14} synapses

Evolutionary biology

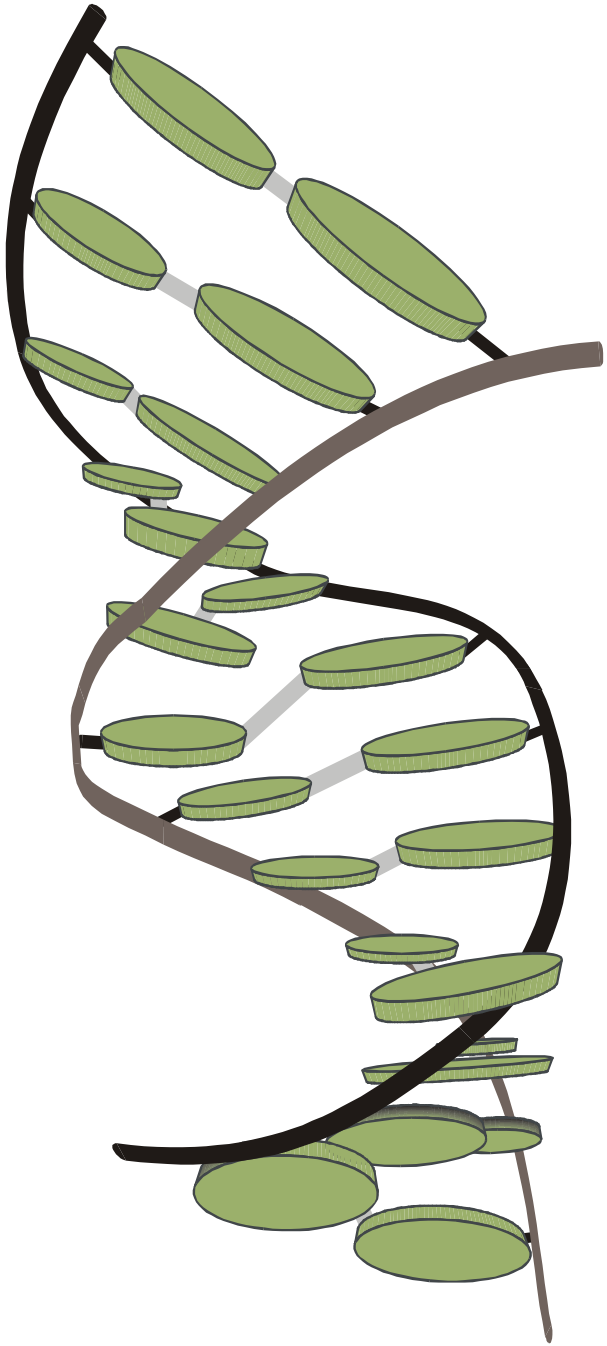
Optimization through variation and selection, relation between genotype, phenotype, and function, ...

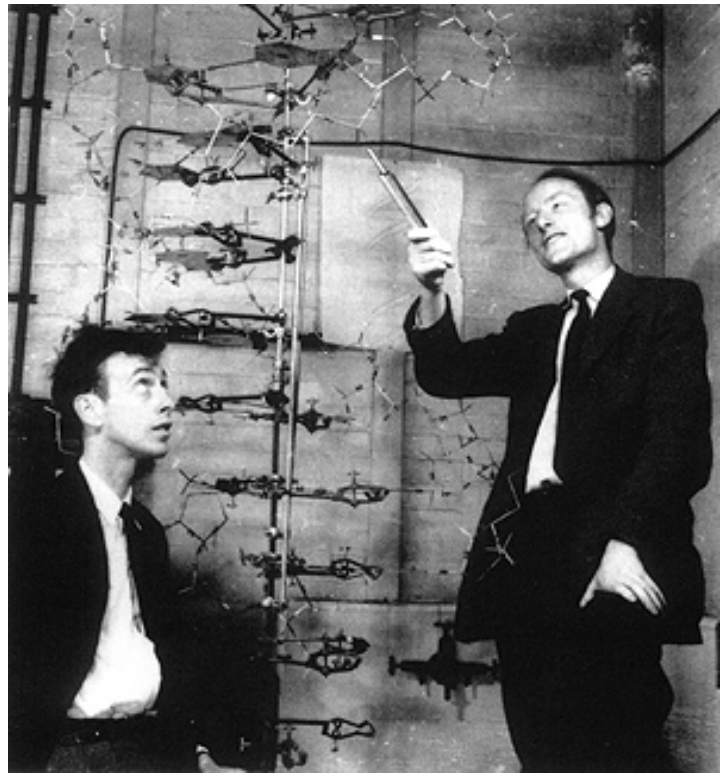
	Generation time	10 000 generations	10^6 generations	10^7 generations
RNA molecules	10 sec 1 min	27.8 h = 1.16 d 6.94 d	115.7 d 1.90 a	3.17 a 19.01 a
Bacteria	20 min 10 h	138.9 d 11.40 a	38.03 a 1 140 a	380 a 11 408 a
Higher multicelluar organisms	10 d 20 a	274 a 20 000 a	27 380 a 2×10^7 a	273 800 a 2×10^8 a

Time scales of evolutionary change



Definition of RNA structure





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

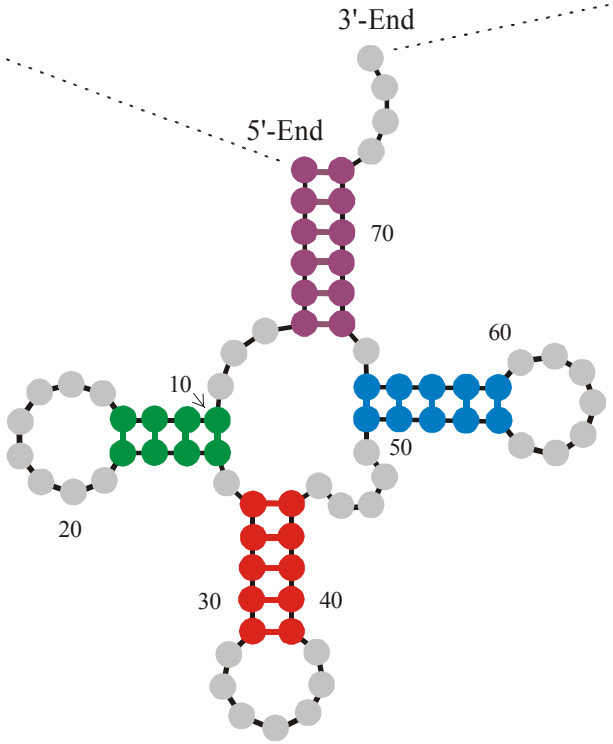
Sequence

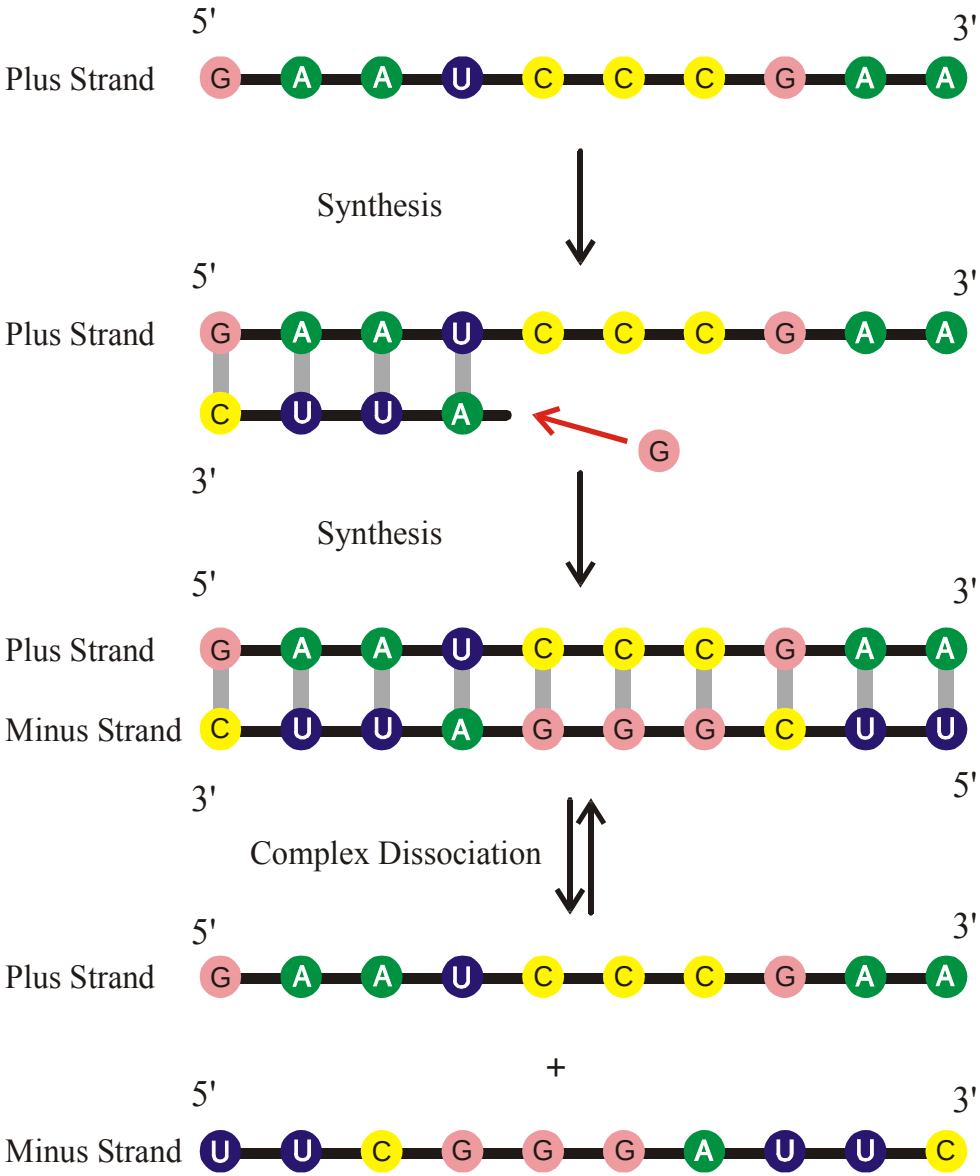
5'-End

3'-End

GCGGAUUUAGCUCAGDDGGGAGAGCMCCAGACUGAAYAUCUGGAGMUCCUGUGTPCGAUCCACAGAAUUCGCACCA

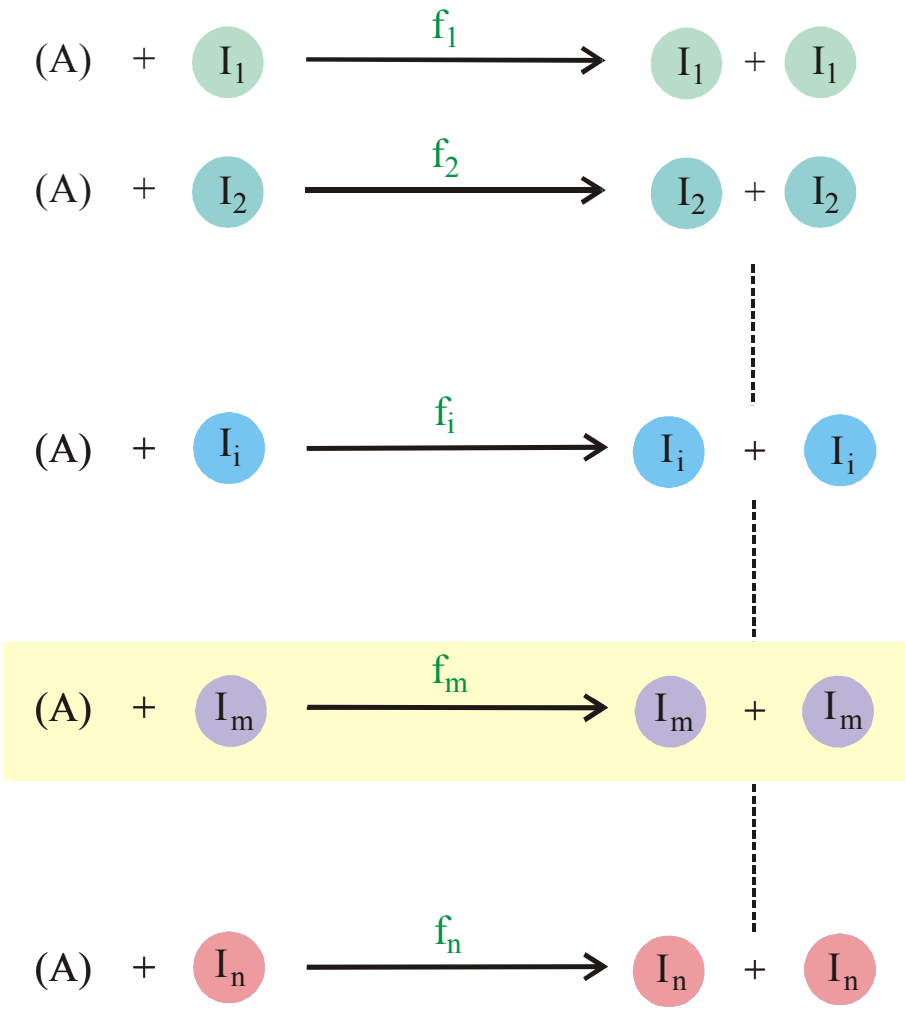
Secondary structure





Complementary replication as the simplest copying mechanism of RNA Complementarity is determined by Watson-Crick base pairs:

G Í C and A=U



$$dx_i / dt = f_i x_i - x_i \Phi = x_i (f_i - \Phi)$$

$\Phi = \Sigma_j f_j x_j$; $\Sigma_j x_j = 1$; i,j =1,2,...,n

$[I_i] = x_i$ Æ 0 ; i =1,2,...,n ;

[A] = a = constant

$f_m = \max \{f_j; j=1,2,...,n\}$

$x_m(t)$ Á 1 for t Á ¸

Reproduction of organisms **or replication** of molecules as the basis of selection

Selection equation: $[I_i] = x_i \neq 0\ ,\ f_i > 0$

$$\frac{dx_i}{dt} = x_i\left(f_i - \phi\right), \quad i=1,2,\cdots,n; \quad \sum_{i=1}^{n} x_i = 1; \quad \phi = \sum_{j=1}^{n} f_j x_j = \overline{f}$$

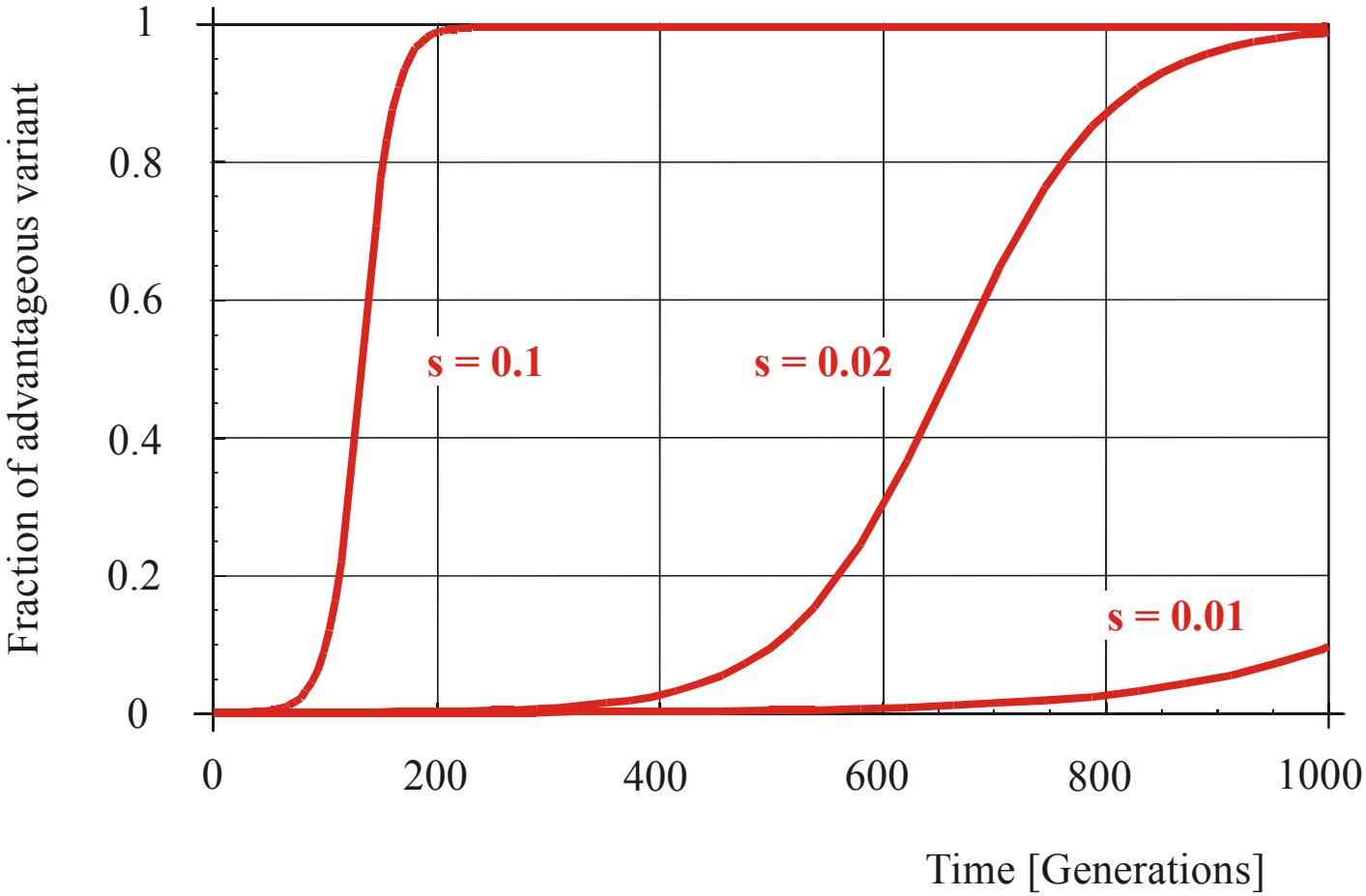
Mean fitness or dilution flux, ϕ(t), is a **non-decreasing function** of time,

$$\frac{d\phi}{dt} = \sum_{i=1}^{n} f_i \frac{dx_i}{dt} = \overline{f^2} - \left(\overline{f}\right)^2 = \mathrm{var}\{f\} \geq 0$$

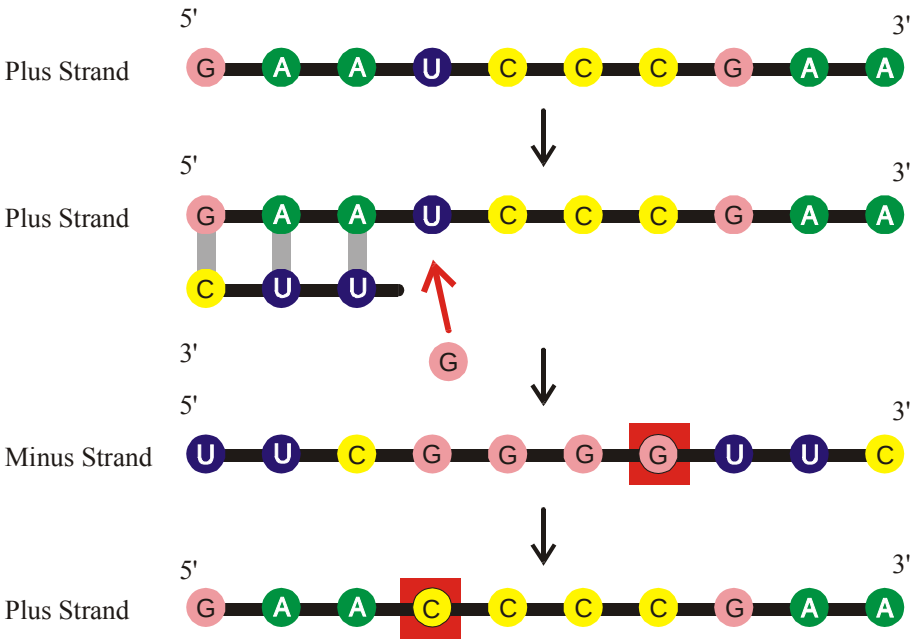
Solutions are obtained by integrating factor transformation

$$x_i(t) = \frac{x_i(0)\cdot\exp(f_i t)}{\sum_{j=1}^{n} x_j(0)\cdot\exp(f_j t)}; \quad i = 1,2,\cdots,n$$

$$\mathbf{s} = (f_2 - f_1) / f_1;\ f_2 > f_1\ ;\ x_1(0) = 1 - 1/N\ ;\ x_2(0) = 1/N$$



Selection of advantageous mutants in populations of N = 10 000 individuals



Point Mutation

GAAUCCCGAA → GAAUCCCGUCCCGAA

Insertion

GAAUCCCGAA → GAAUCCGA

Deletion

Mutations in nucleic acids represent the mechanism of **variation** of **genotypes**.

Theory of molecular evolution

M.Eigen, ***Self-organization of matter and the evolution of biological macromolecules***. Naturwissenschaften **58** (1971), 465-526

C.J. Thompson, J.L. McBride, ***On Eigen's theory of the self-organization of matter and the evolution of biological macromolecules***. Math. Biosci. **21** (1974), 127-142

B.L. Jones, R.H. Enns, S.S. Rangnekar, ***On the theory of selection of coupled macromolecular systems.*** Bull.Math.Biol. **38** (1976), 15-28

M.Eigen, P.Schuster, ***The hypercycle. A principle of natural self-organization. Part A: Emergence of the hypercycle***. Naturwissenschaften **58** (1977), 465-526

M.Eigen, P.Schuster, ***The hypercycle. A principle of natural self-organization. Part B: The abstract hypercycle***. Naturwissenschaften **65** (1978), 7-41

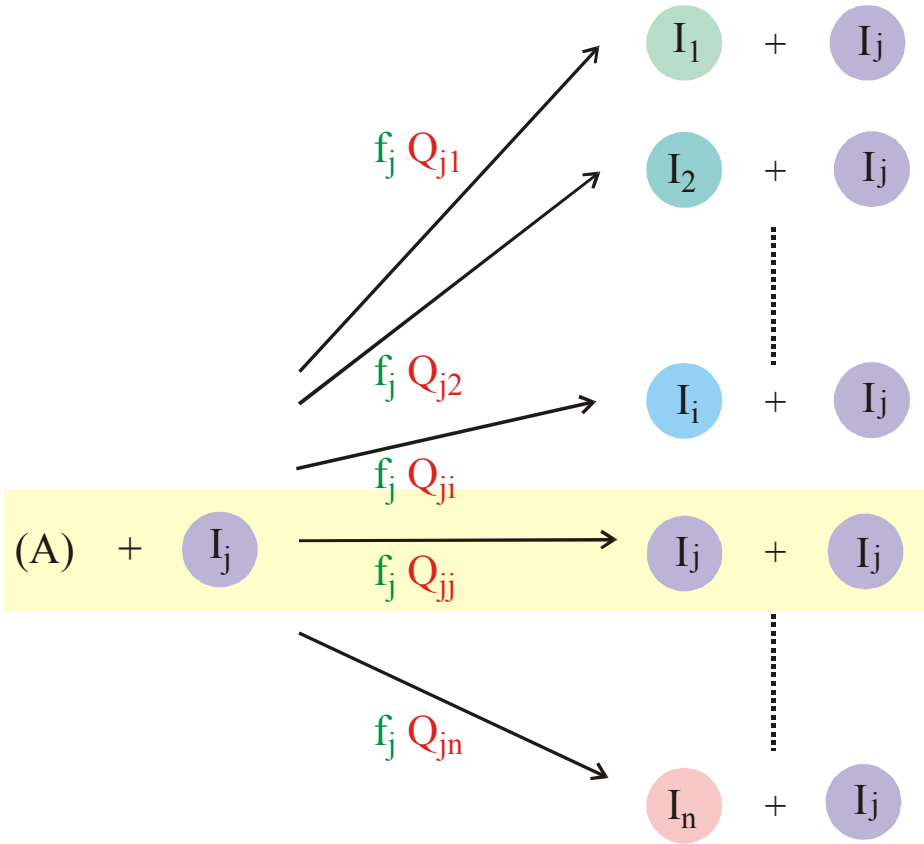
M.Eigen, P.Schuster, ***The hypercycle. A principle of natural self-organization. Part C: The realistic hypercycle***. Naturwissenschaften **65** (1978), 341-369

J. Swetina, P. Schuster, ***Self-replication with errors - A model for polynucleotide replication.*** Biophys.Chem. **16** (1982), 329-345

J.S. McCaskill, ***A localization threshold for macromolecular quasispecies from continuously distributed replication rates***. J.Chem.Phys. **80** (1984), 5194-5202

M.Eigen, J.McCaskill, P.Schuster, ***The molecular quasispecies***. Adv.Chem.Phys. **75** (1989), 149-263

C. Reidys, C.Forst, P.Schuster, ***Replication and mutation on neutral networks***. Bull.Math.Biol. **63** (2001), 57-94



$$dx_i / dt = \Sigma_j f_j Q_{ji} x_j - x_i \Phi$$

$$\Phi = \Sigma_j f_j x_i ; \quad \Sigma_j x_j = 1 ; \quad \Sigma_i Q_{ij} = 1$$

$[I_i] = x_i \neq 0 ; \quad i = 1,2,...,n ;$

$[A] = a = \text{constant}$

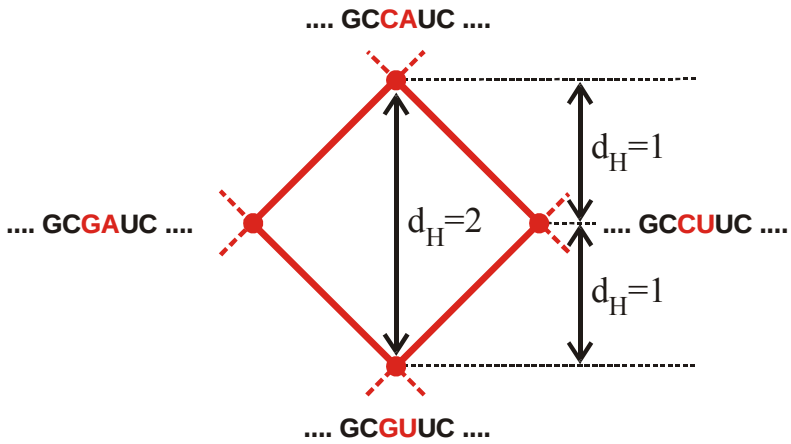
$$Q_{ij} = (1-p)^{\ell - d(i,j)} p^{d(i,j)}$$

p ..... Error rate per digit

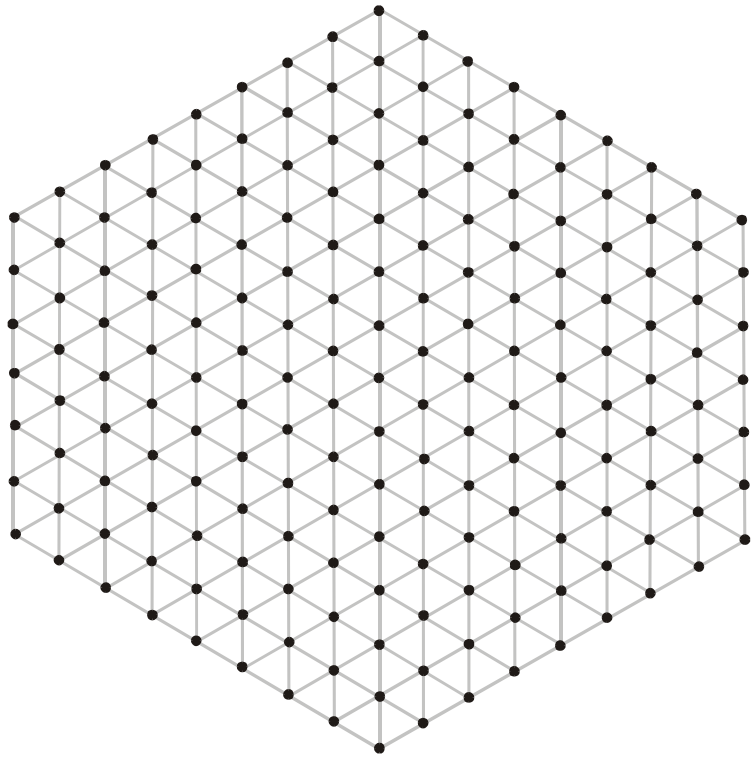
ℓ ..... Chain length of the polynucleotide

d(i,j) .... Hamming distance between I_i and I_j

Chemical kinetics of replication and mutation as parallel reactions

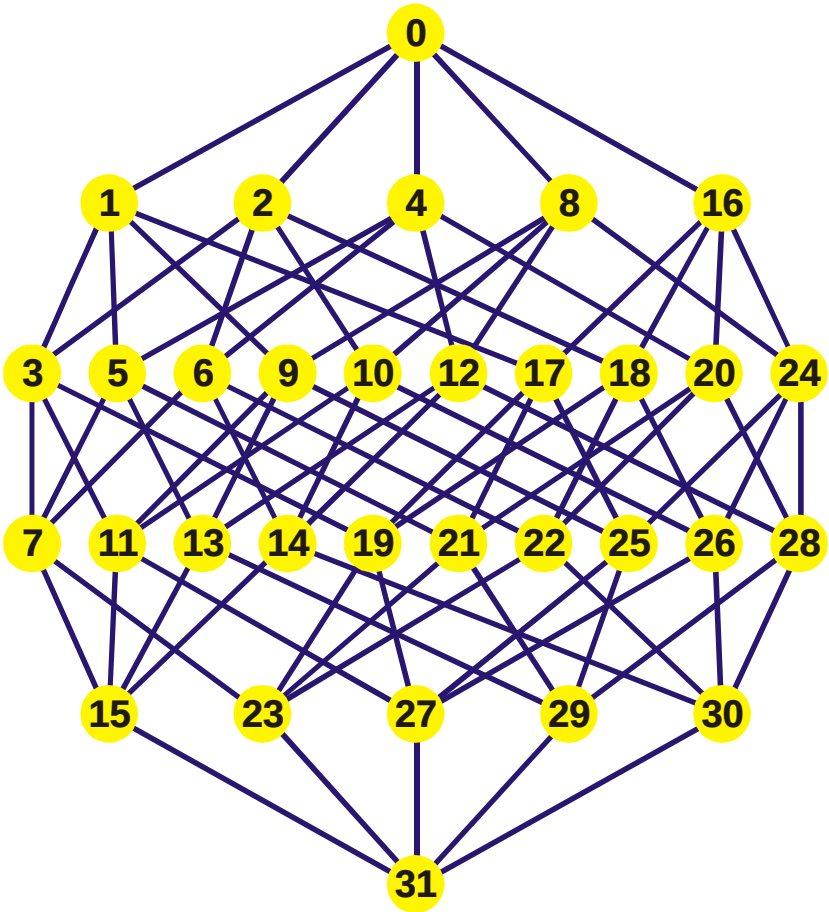


City-block distance in sequence space



2D Sketch of sequence space

Single point mutations as moves in sequence space



Mutant class

0

1

2

3

4

5

Binary sequences are encoded by their decimal equivalents:

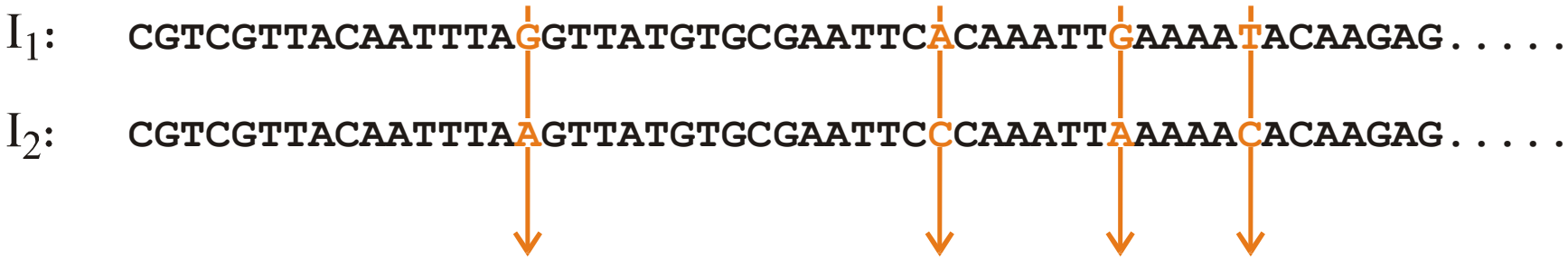
C = 0 and G = 1, for example,

"0" ≡ 00000 = CCCCC,

"14" ≡ 01110 = CGGGC,

"29" ≡ 11101 = GGGCG, etc.

Sequence space of binary sequences of chain lenght n=5



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

Mutation-selection equation: $[\mathrm{I}_i] = x_i \text{ Æ } 0,\ f_i > 0,\ Q_{ij} \text{ Æ } 0$

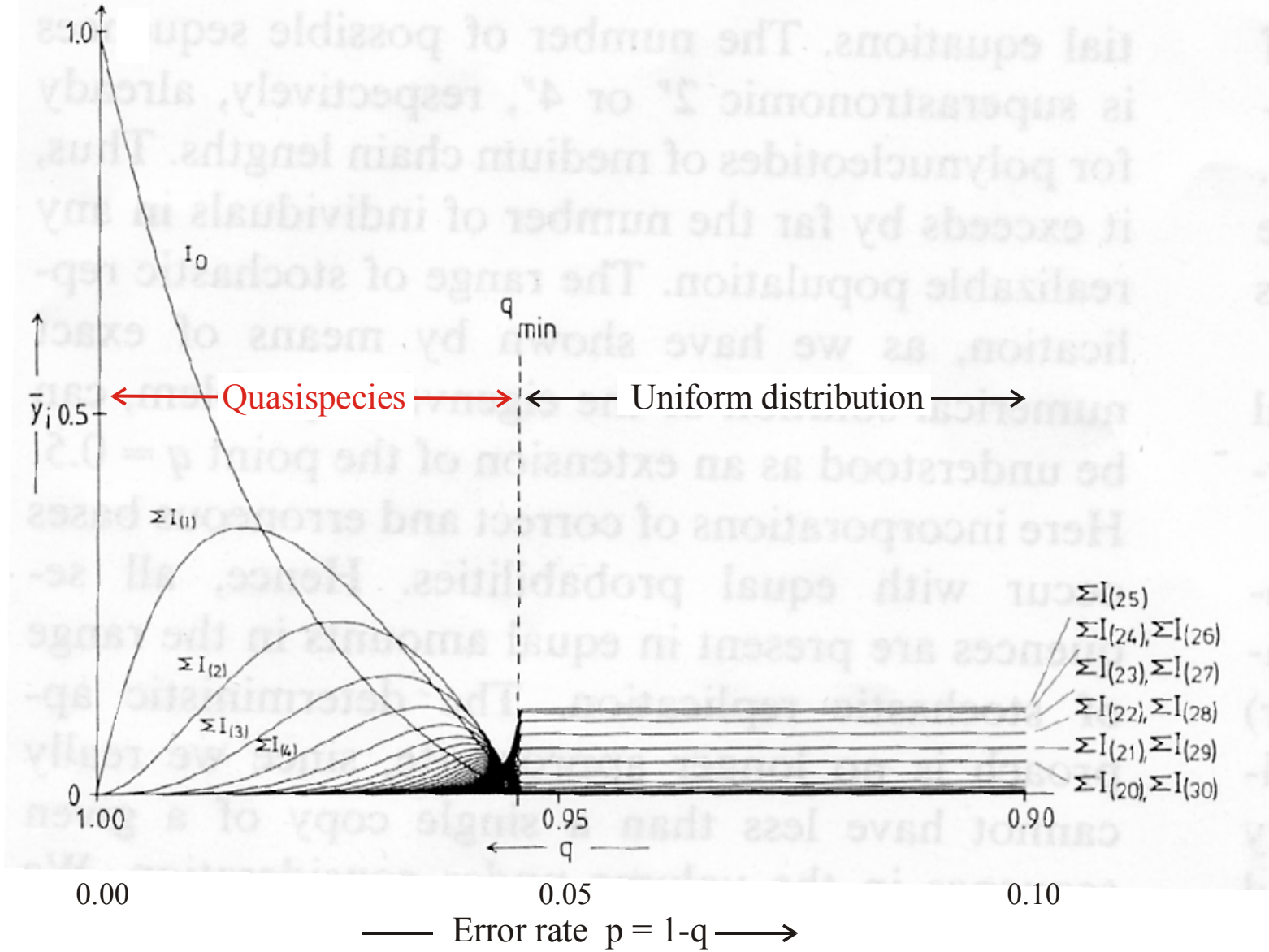
$$\frac{dx_i}{dt} = \sum_{j=1}^{n} f_j Q_{ji}\, x_j - x_i\, \phi, \quad i = 1,2,\cdots,n; \quad \sum_{i=1}^{n} x_i = 1; \quad \phi = \sum_{j=1}^{n} f_j x_j = \overline{f}$$

Solutions are obtained after integrating factor transformation by means of an eigenvalue problem

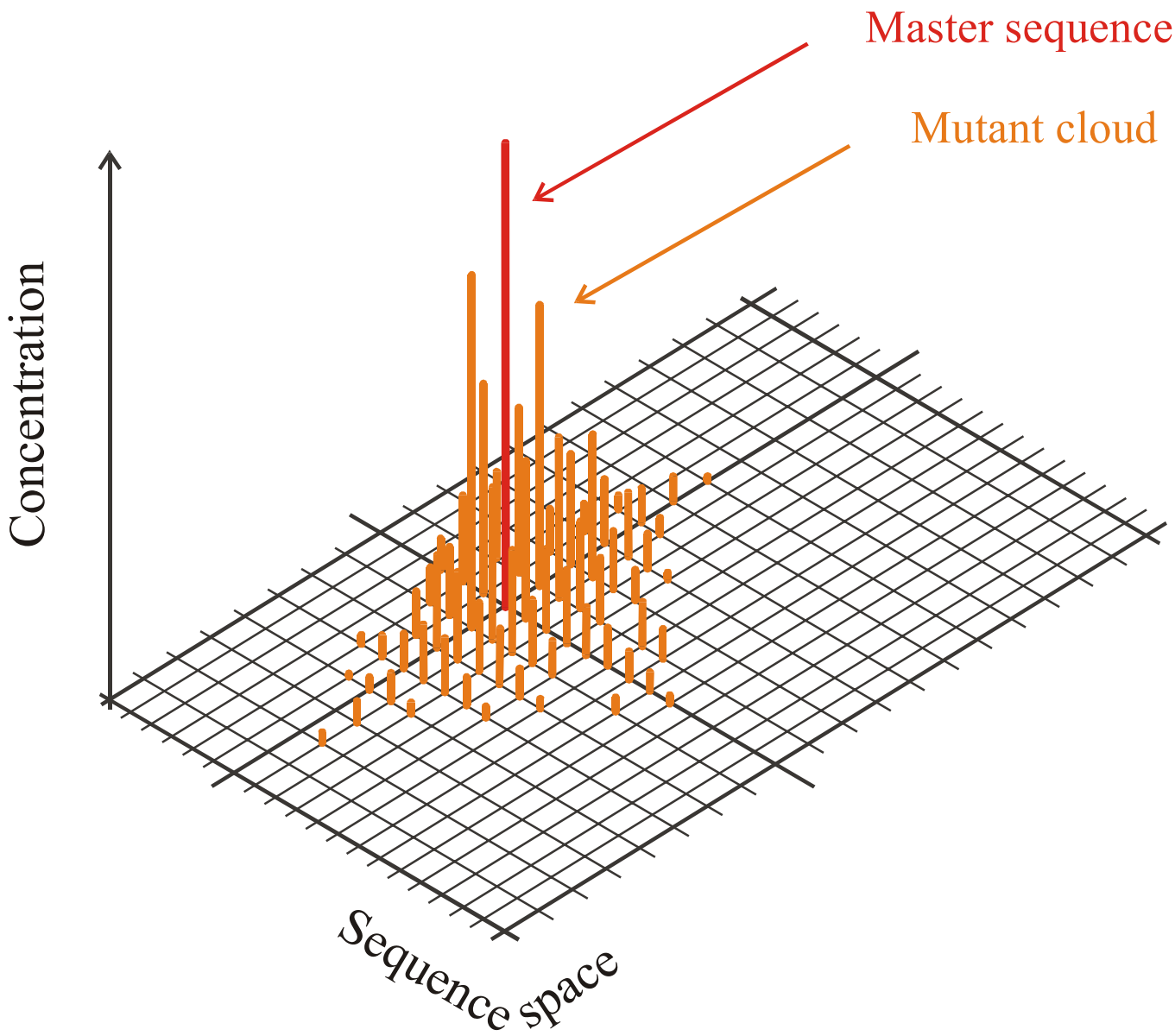
$$x_i(t) = \frac{\sum_{k=0}^{n-1} \ell_{ik} \cdot c_k(0) \cdot \exp(\lambda_k t)}{\sum_{j=1}^{n} \sum_{k=0}^{n-1} \ell_{jk} \cdot c_k(0) \cdot \exp(\lambda_k t)}; \quad i = 1,2,\cdots,n; \quad c_k(0) = \sum_{i=1}^{n} h_{ki}\, x_i(0)$$

$$W \div \left\{ f_i Q_{ij};\ i, j = 1,2,\cdots,n \right\};\ L = \left\{ \ell_{ij};\ i, j = 1,2,\cdots,n \right\};\ L^{-1} = H = \left\{ h_{ij};\ i, j = 1,2,\cdots,n \right\}$$

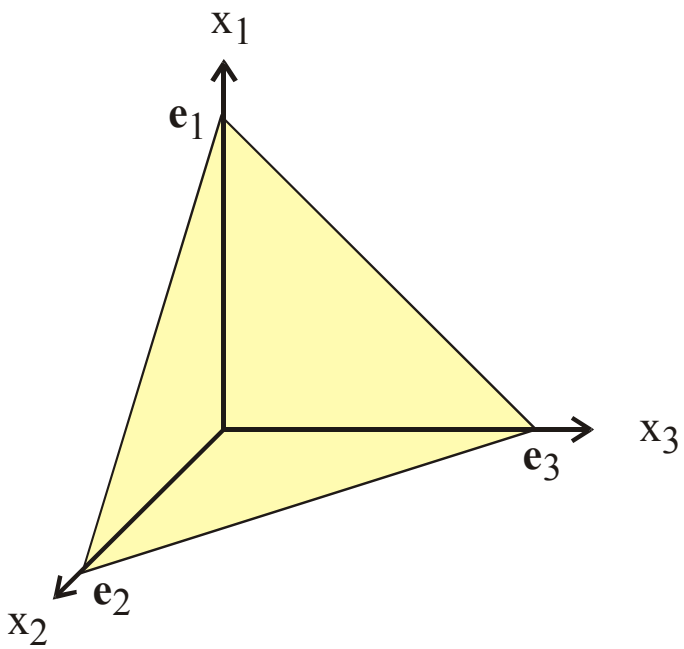
$$L^{-1} \cdot W \cdot L \quad = \quad \Lambda \quad = \quad \left\{ \lambda_k; k = 0, 1, \cdots, n-1 \right\}$$

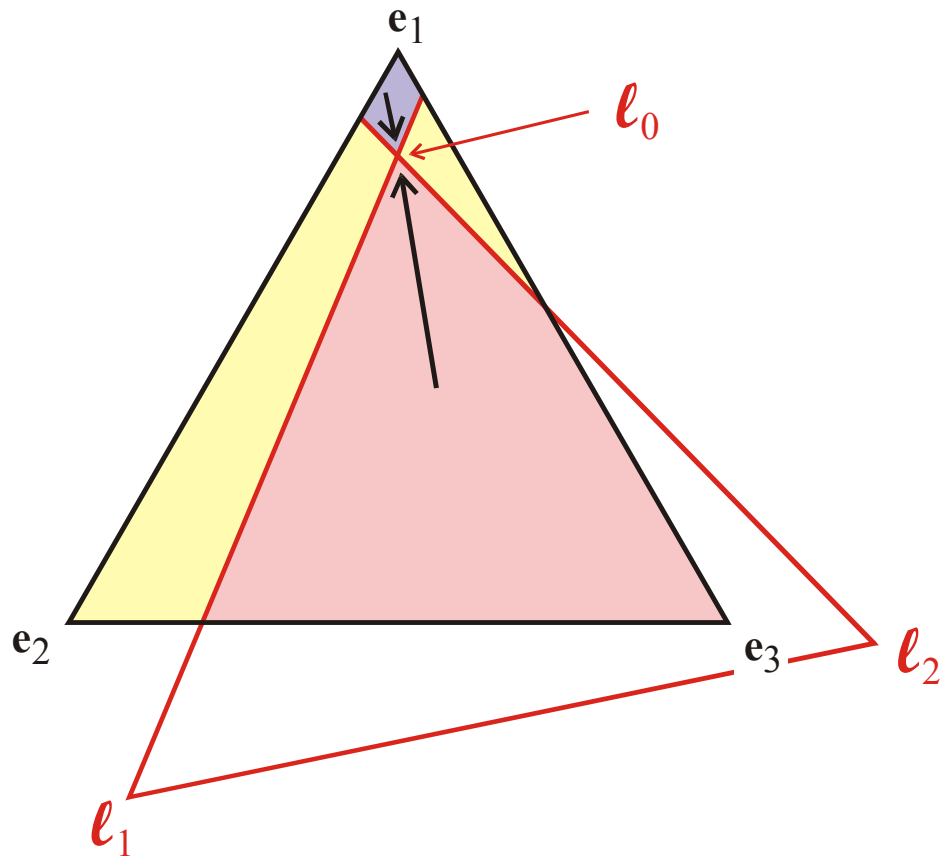


Quasispecies as a function of the replication accuracy q



The molecular quasispecies in sequence space



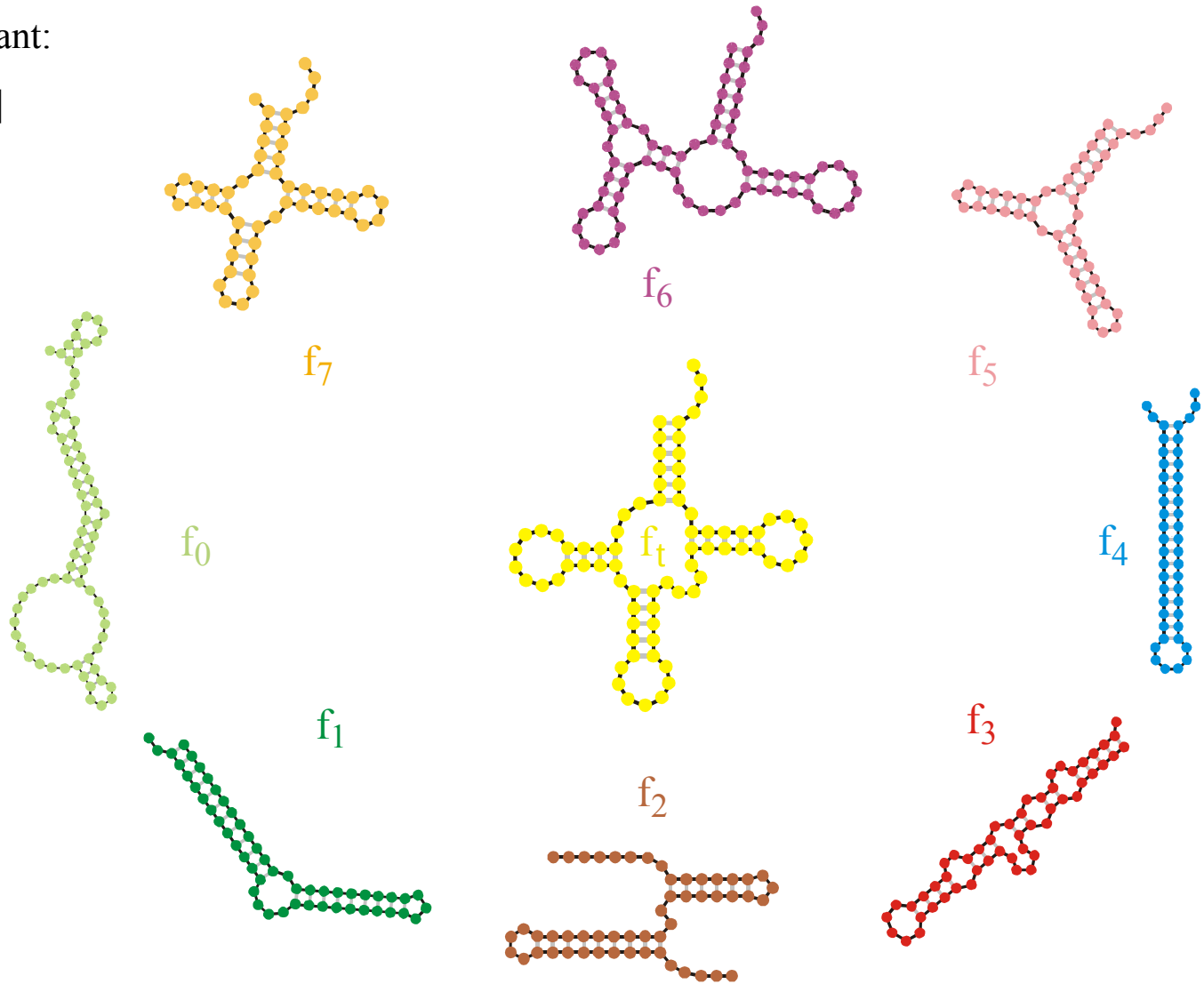


The quasispecies on the concentration simplex $\mathbb{S}_3 = \left\{ x_i \geq 0,\ i = 1,2,3;\ \sum_{i=1}^{3} x_i = 1 \right\}$

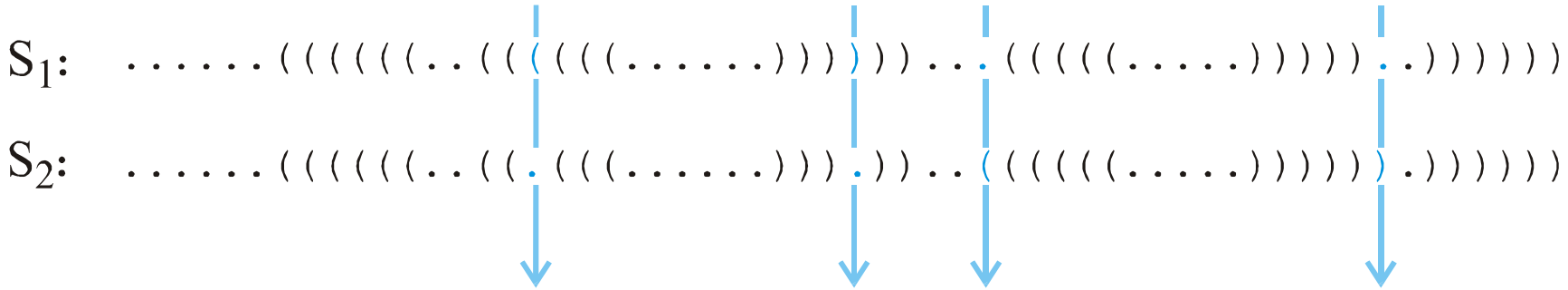
Replication rate constant:

$$f_k = g / [a + Dd_S^{(k)}]$$

$$Dd_S^{(k)} = d_H(S_k, S_t)$$



Evaluation of RNA secondary structures yields replication rate constants



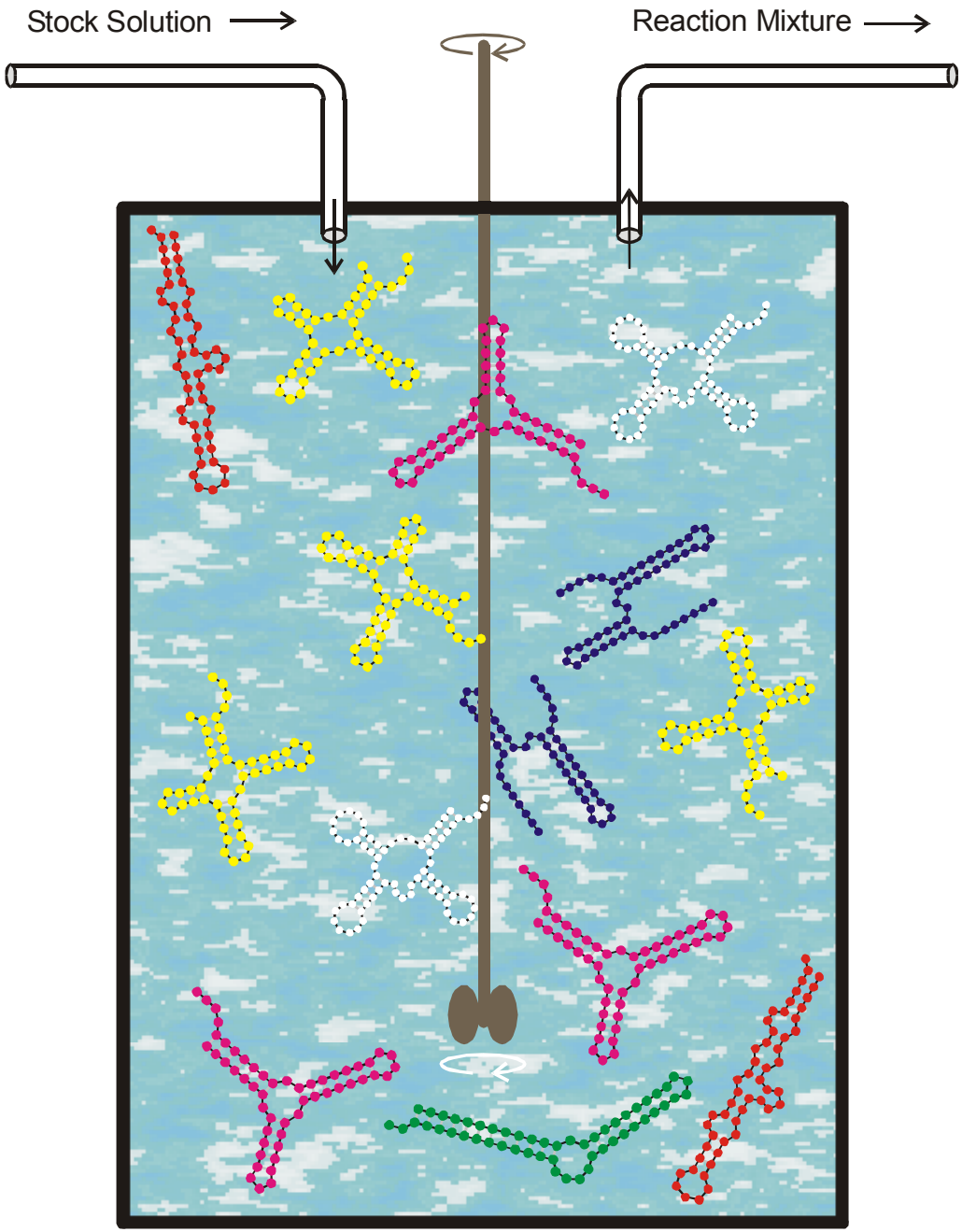
Hamming distance $d_H(S_1,S_2) = 4$

(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



Replication rate constant:

$$f_k = g / [a + Dd_S^{(k)}]$$

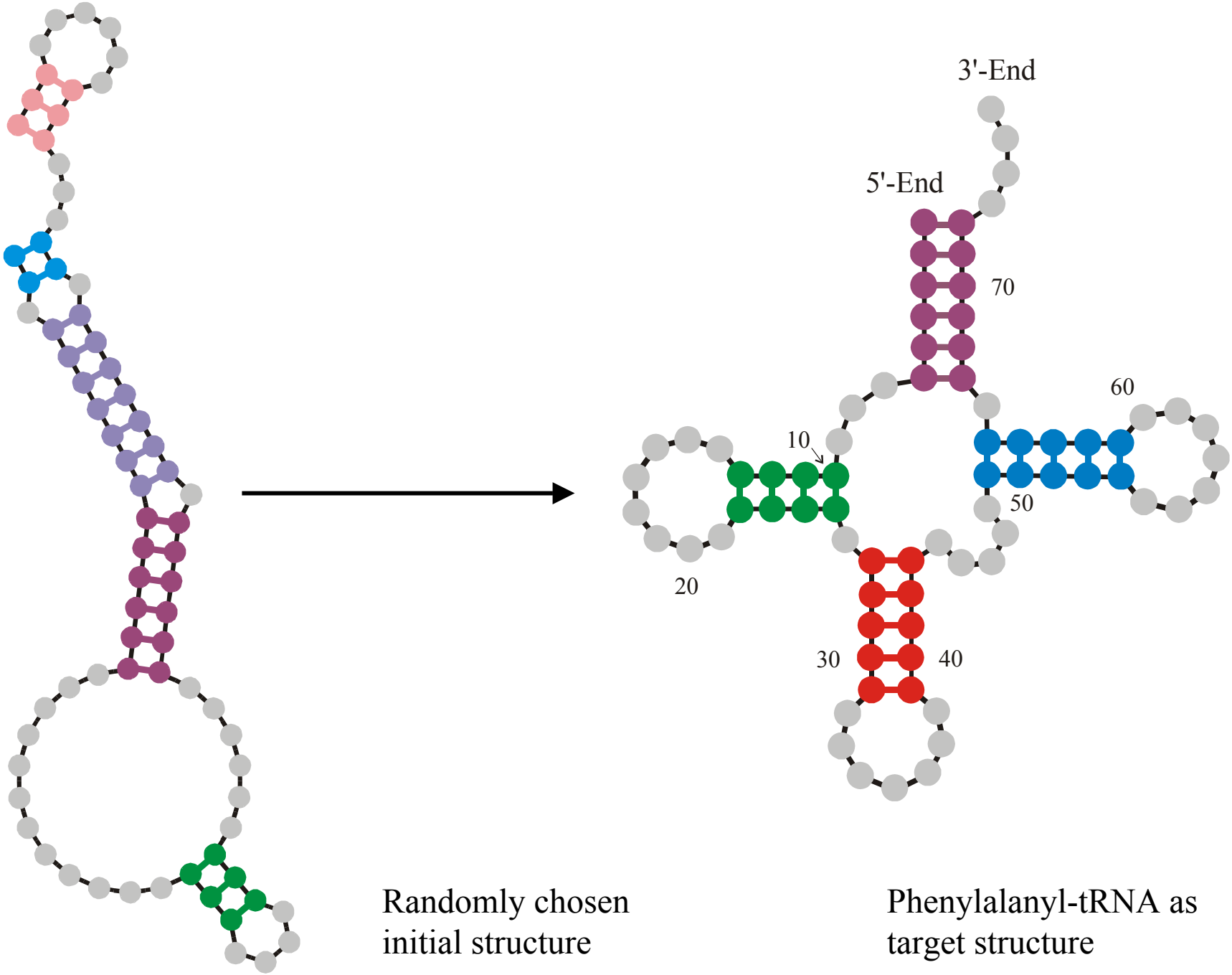
$$Dd_S^{(k)} = d_H(S_k, S_t)$$

Selection constraint:

RNA molecules is controlled by the flow

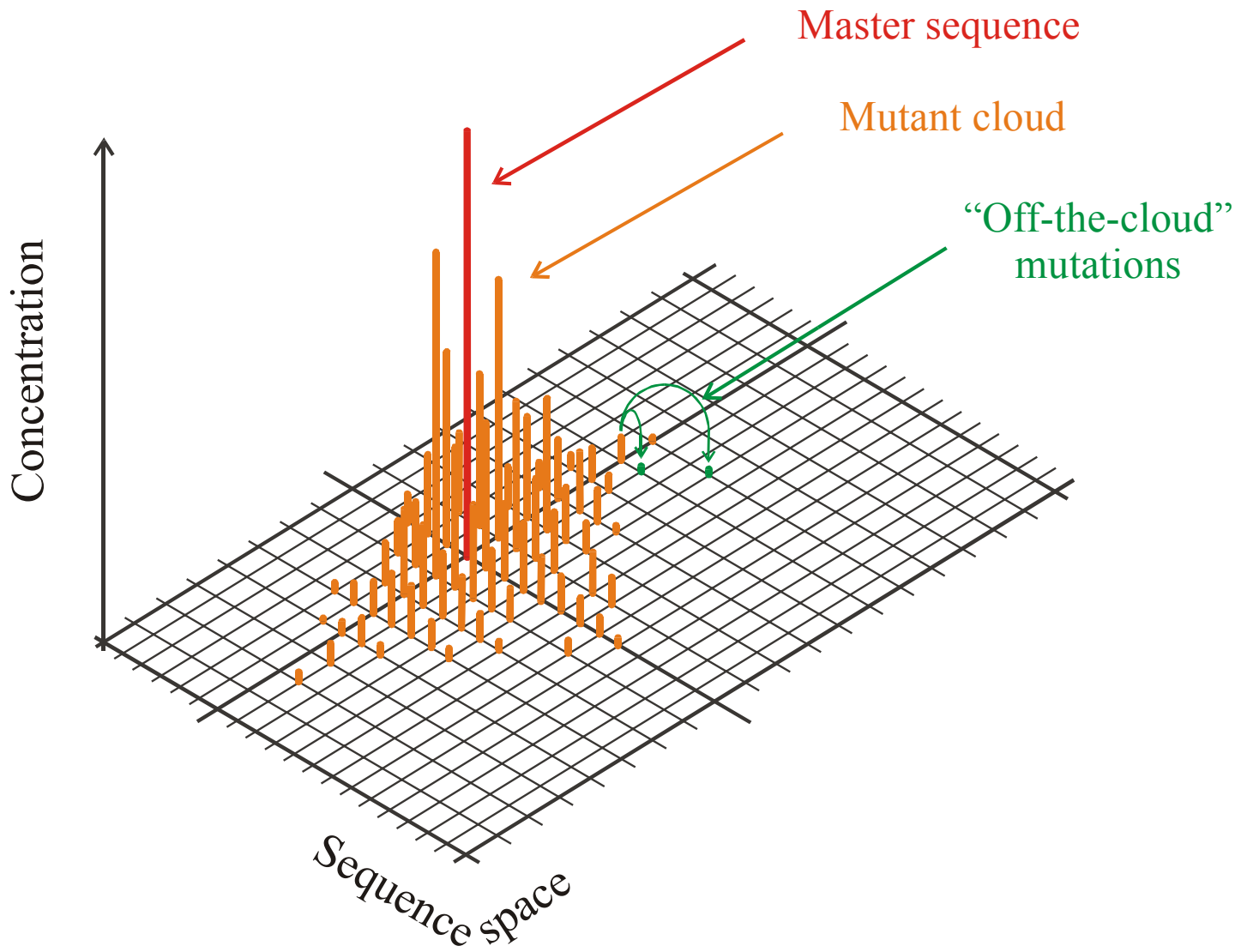
$$N(t) \approx \overline{N} \pm \sqrt{\overline{N}}$$

The flowreactor as a device for studies of evolution ***in vitro*** and ***in silico***

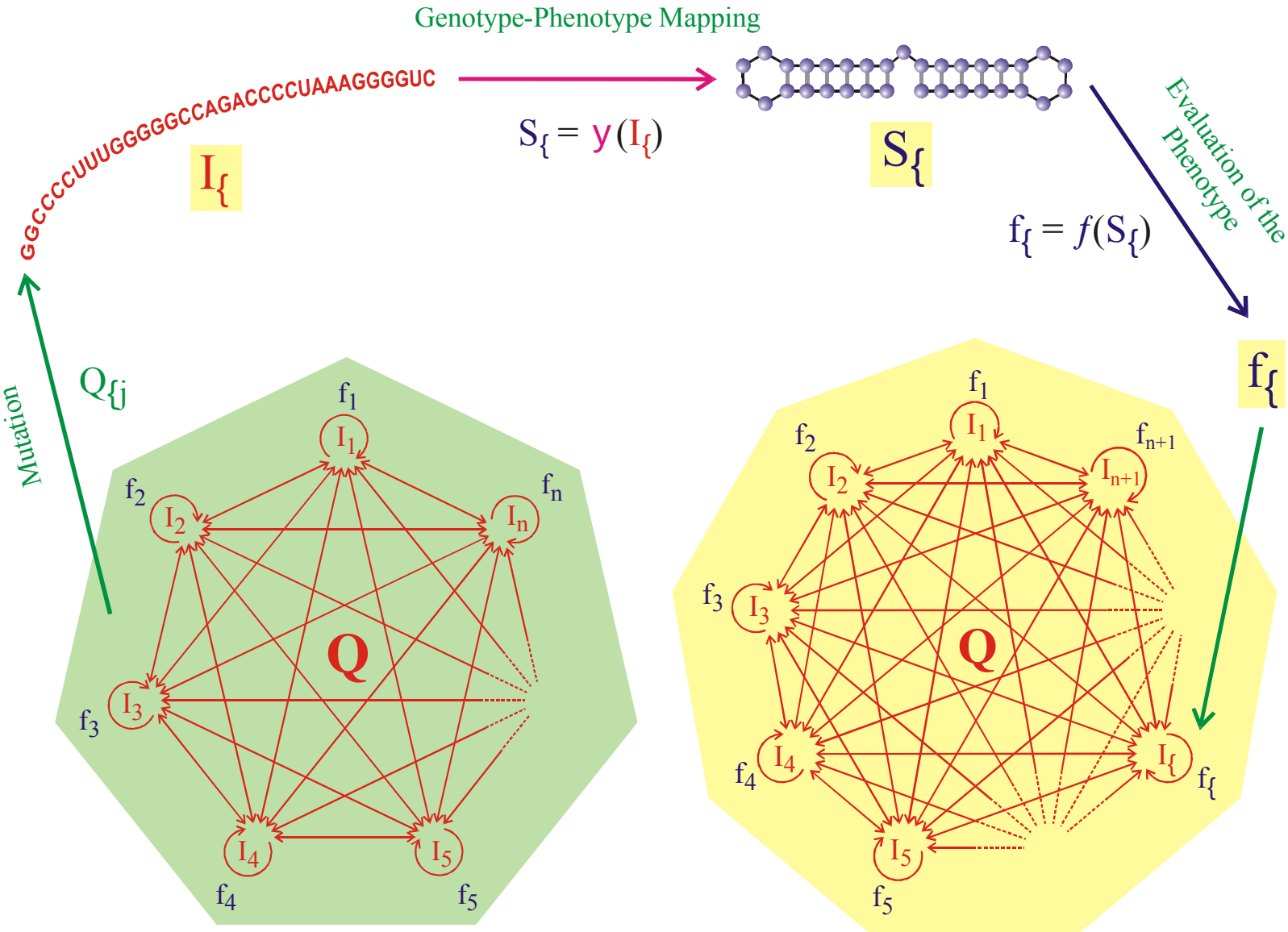


Randomly chosen
initial structure

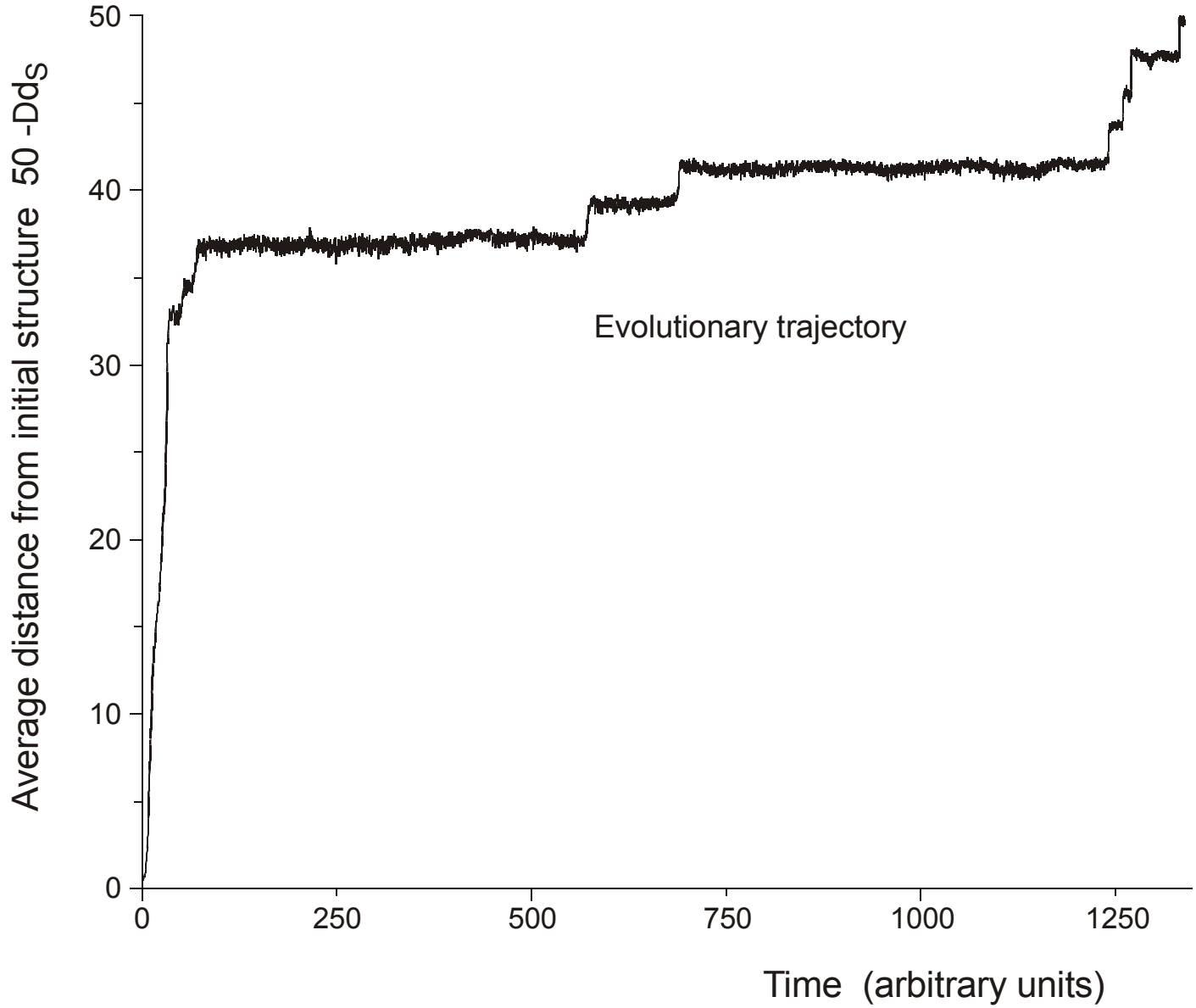
Phenylalanyl-tRNA as
target structure



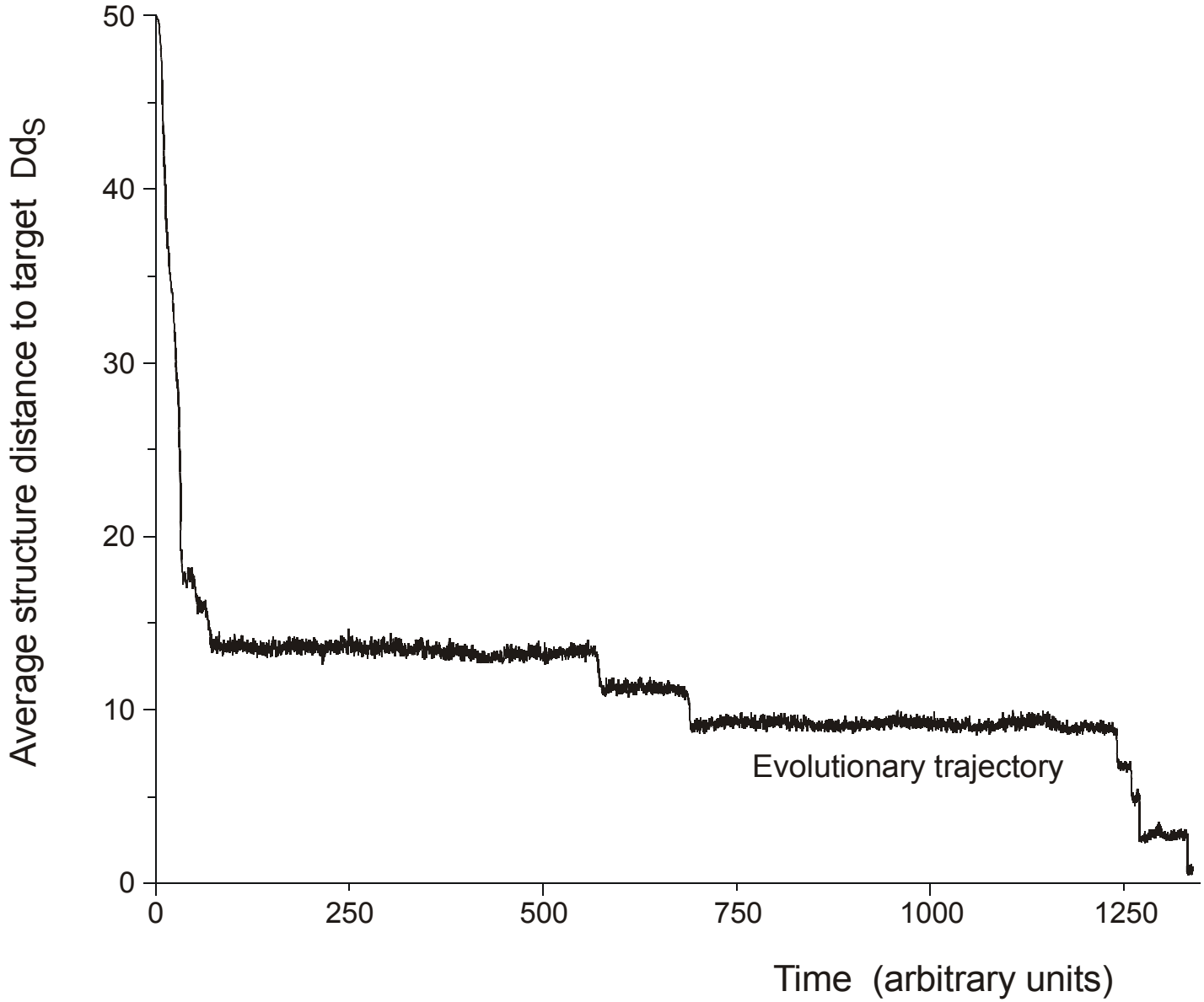
The molecular quasispecies in sequence space



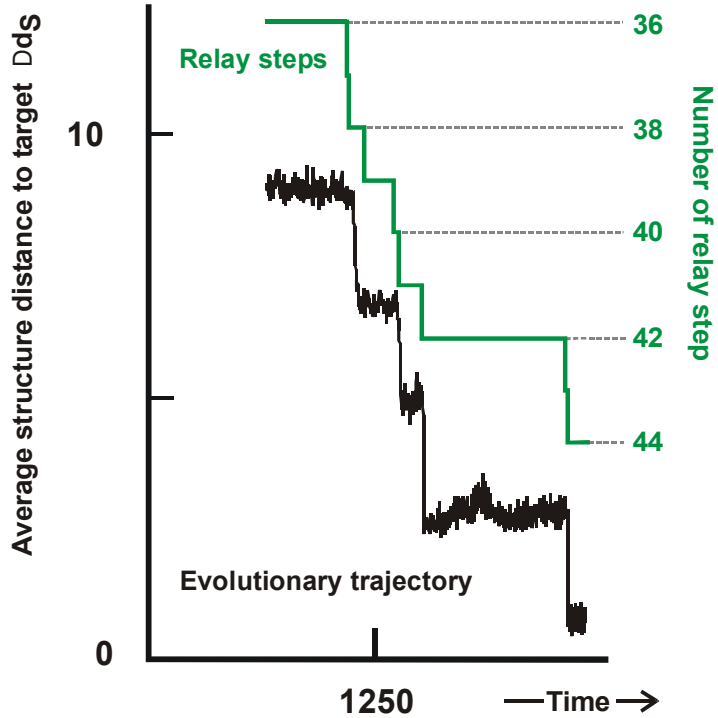
Evolutionary dynamics
including molecular phenotypes

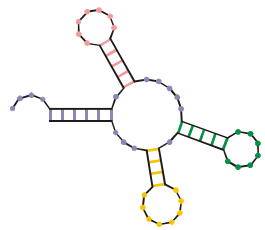


In silico optimization in the flow reactor: Trajectory (**biologists‘ view**)



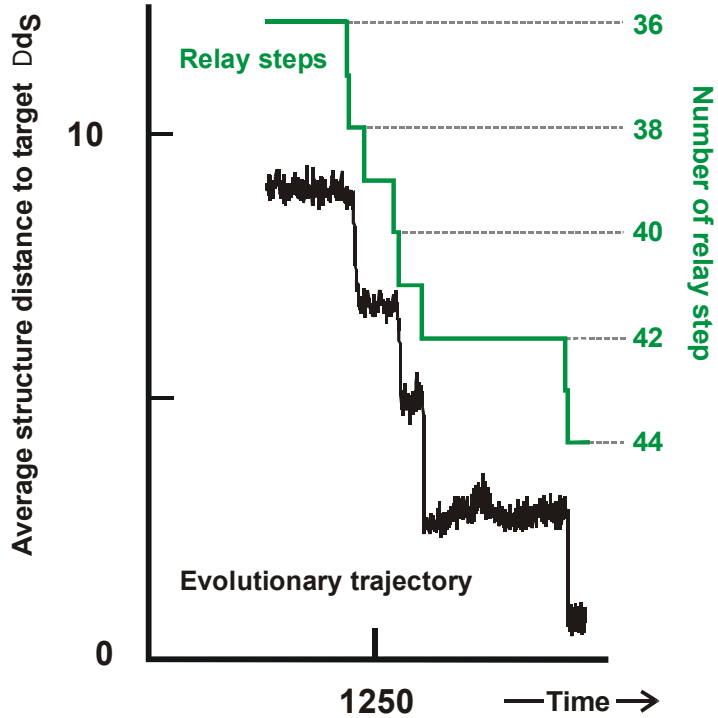
In silico optimization in the flow reactor: Trajectory (**physicists‘ view**)

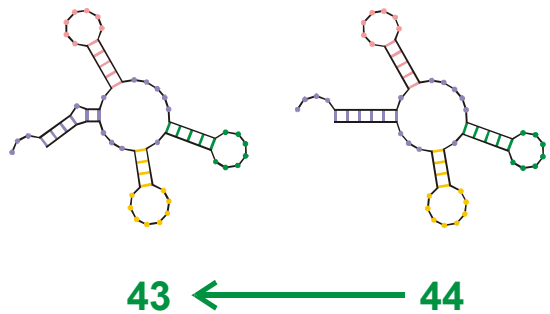




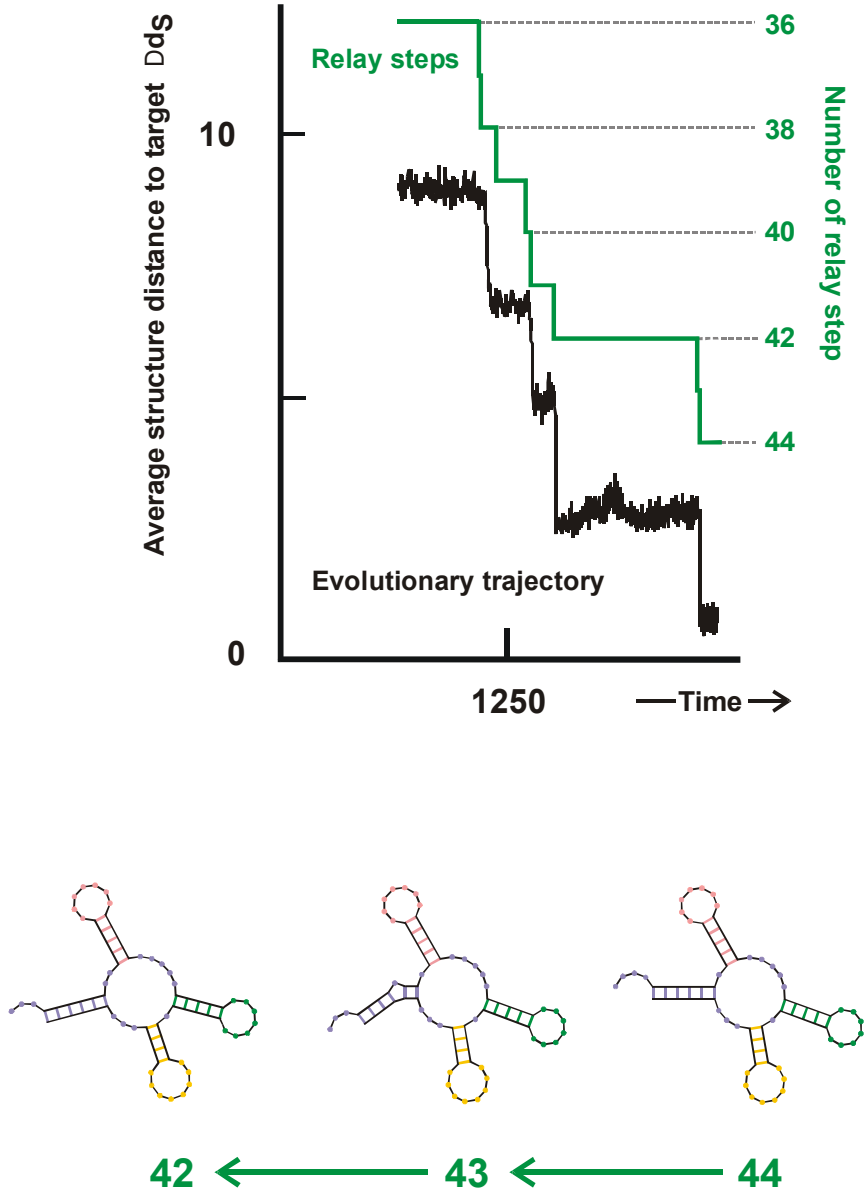
44

Endconformation of optimization

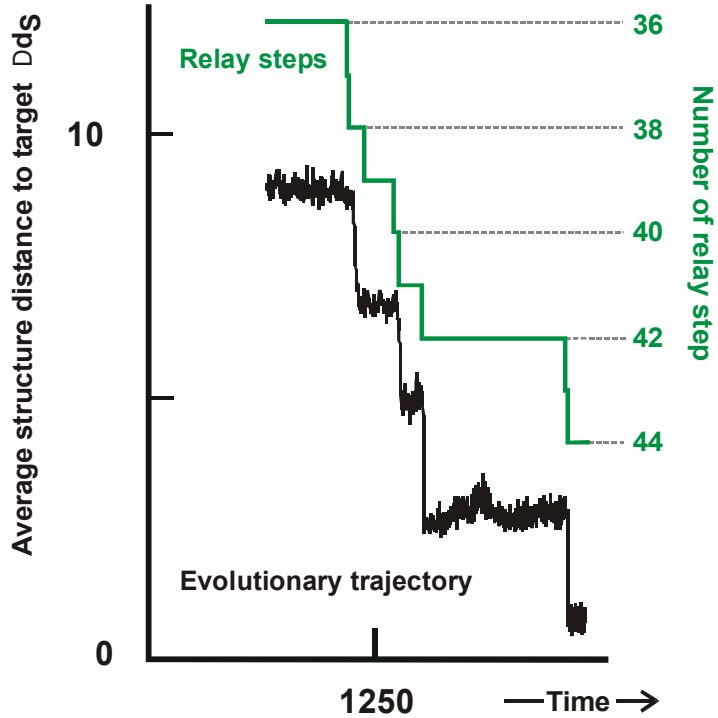


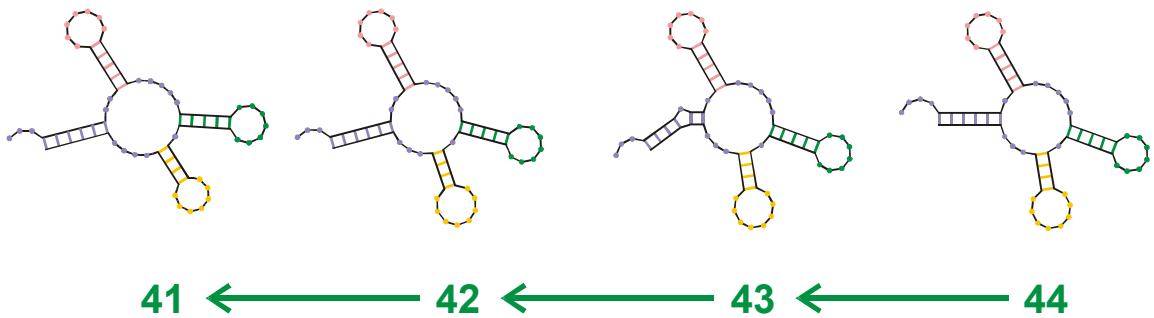


Reconstruction of the last step 43 Á 44

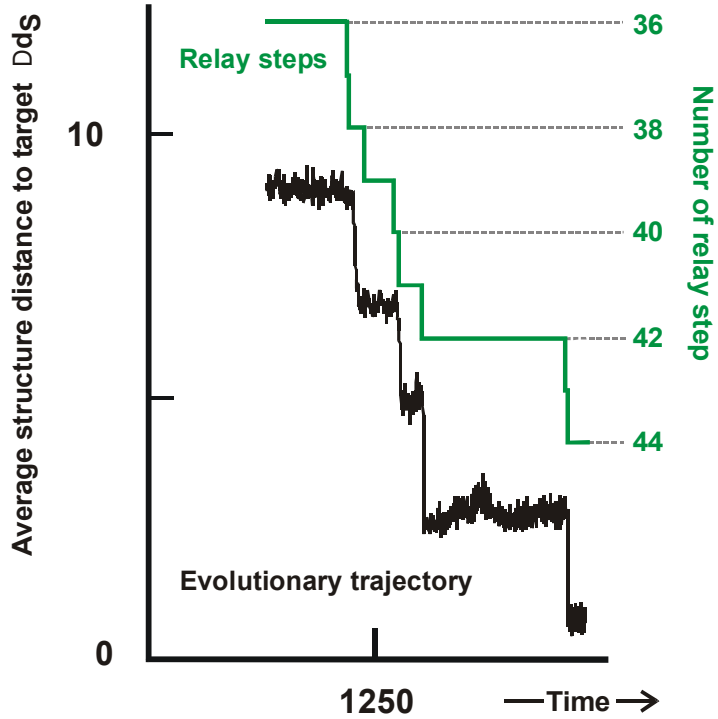


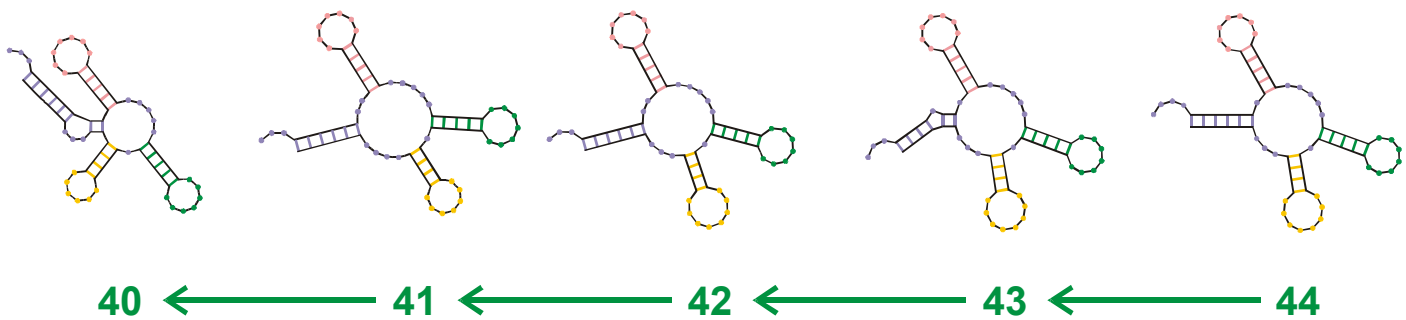
Reconstruction of last-but-one step 42 Á 43 (Á 44)



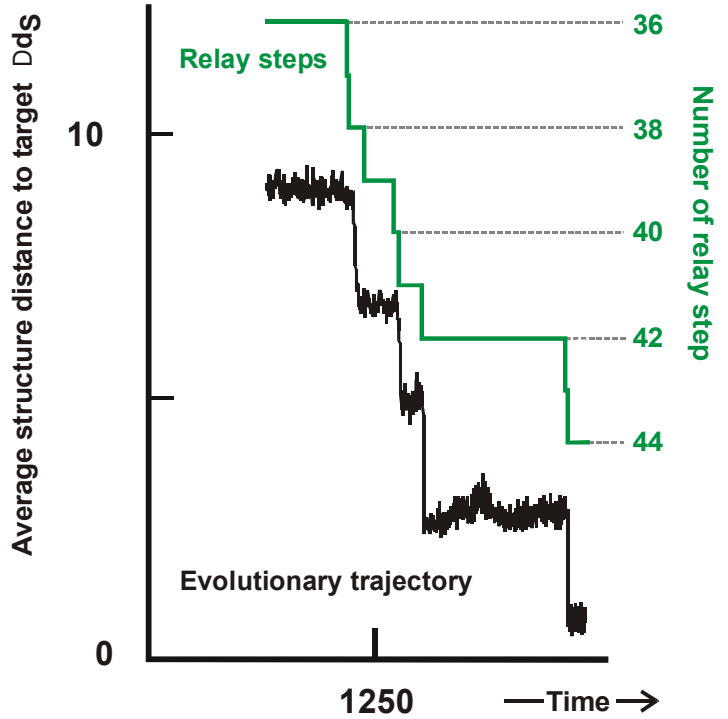


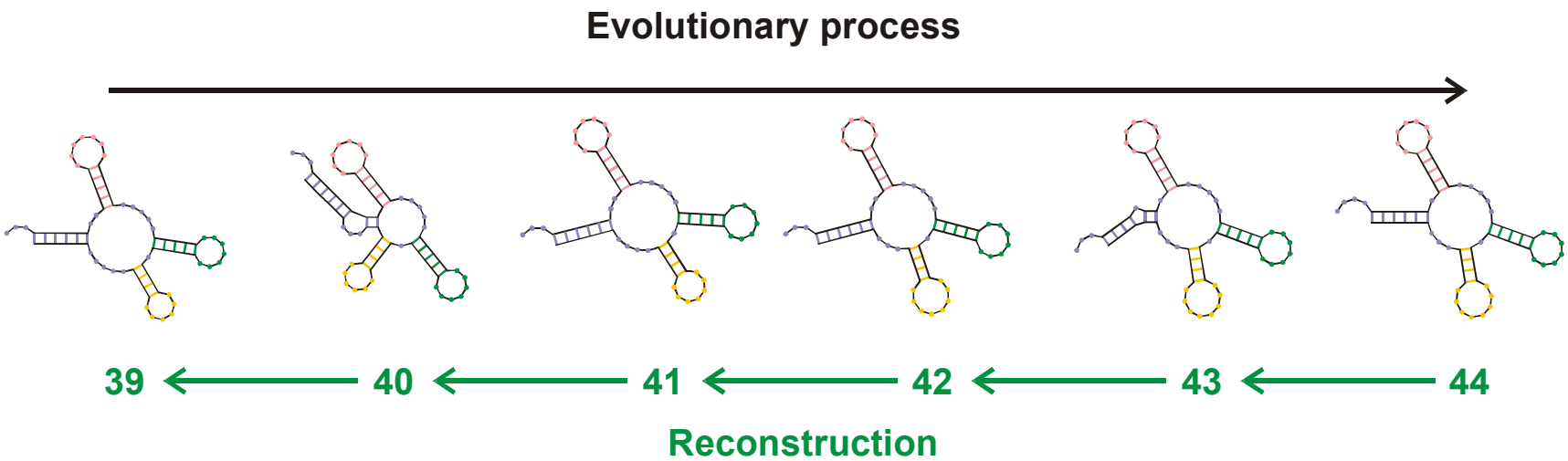
Reconstruction of step 41 Á 42 (Á 43 Á 44)



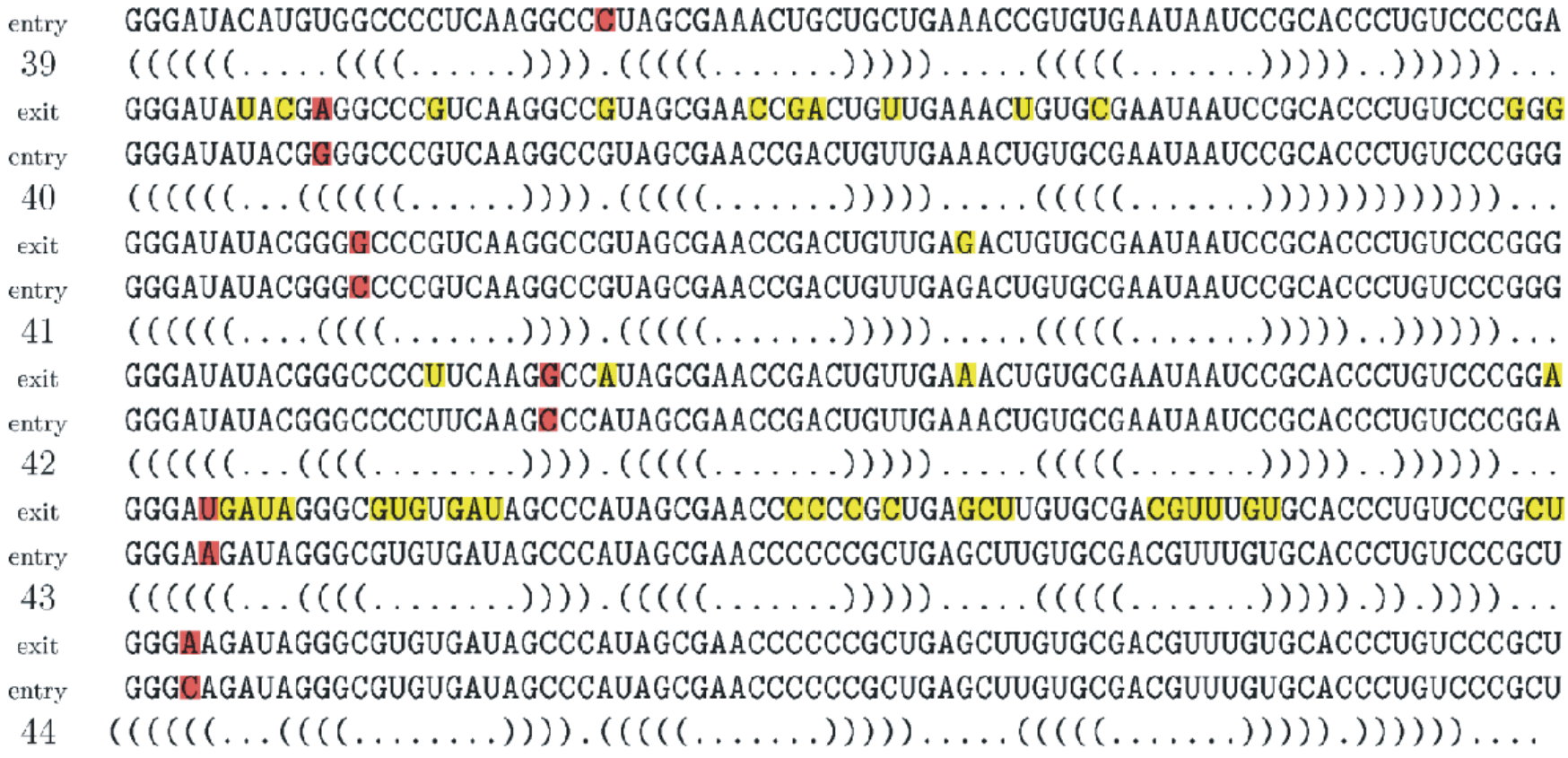


Reconstruction of step 40 Á 41 (Á 42 Á 43 Á 44)



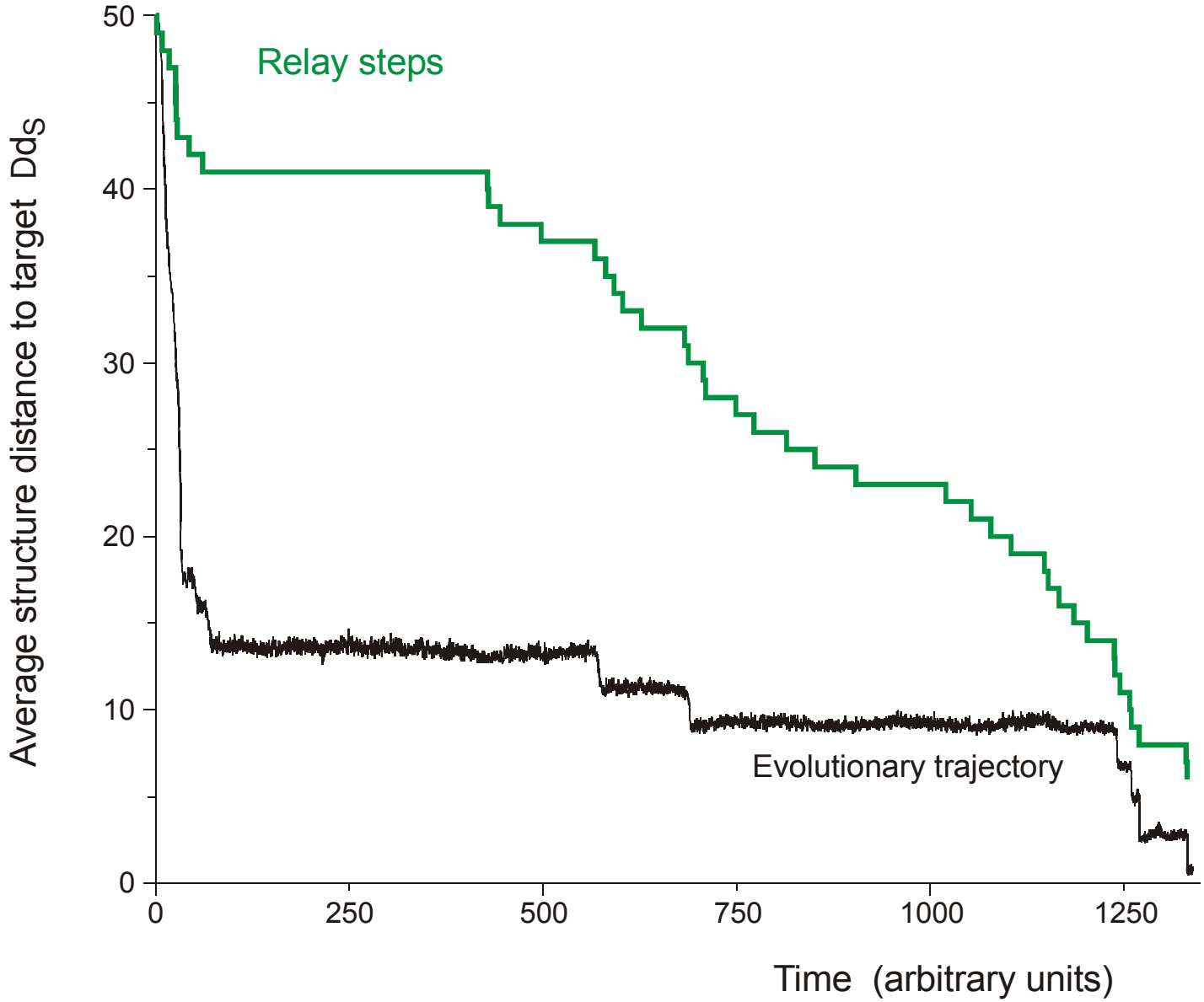


Reconstruction of the relay series



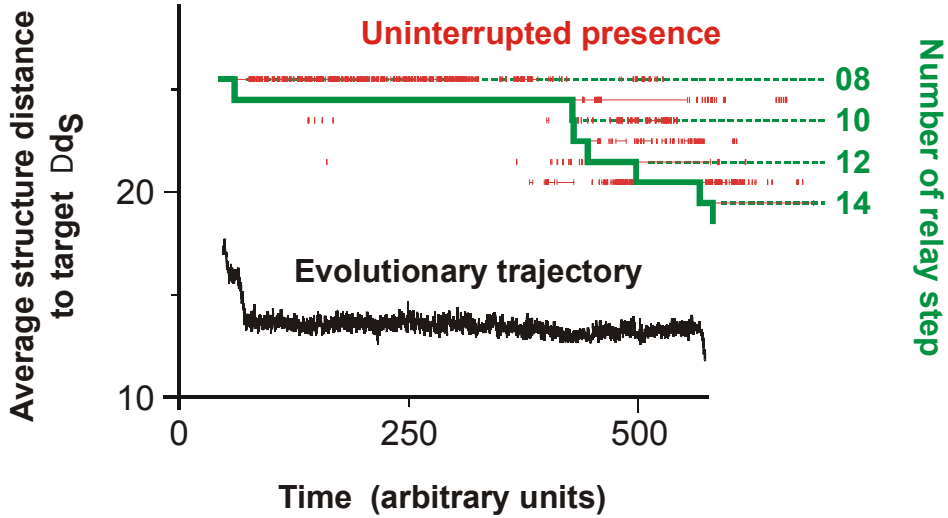
Transition inducing point mutations **Neutral point mutations**

Change in RNA sequences during the final five relay steps 39 Á 44



In silico optimization in the flow reactor: Trajectory and relay steps

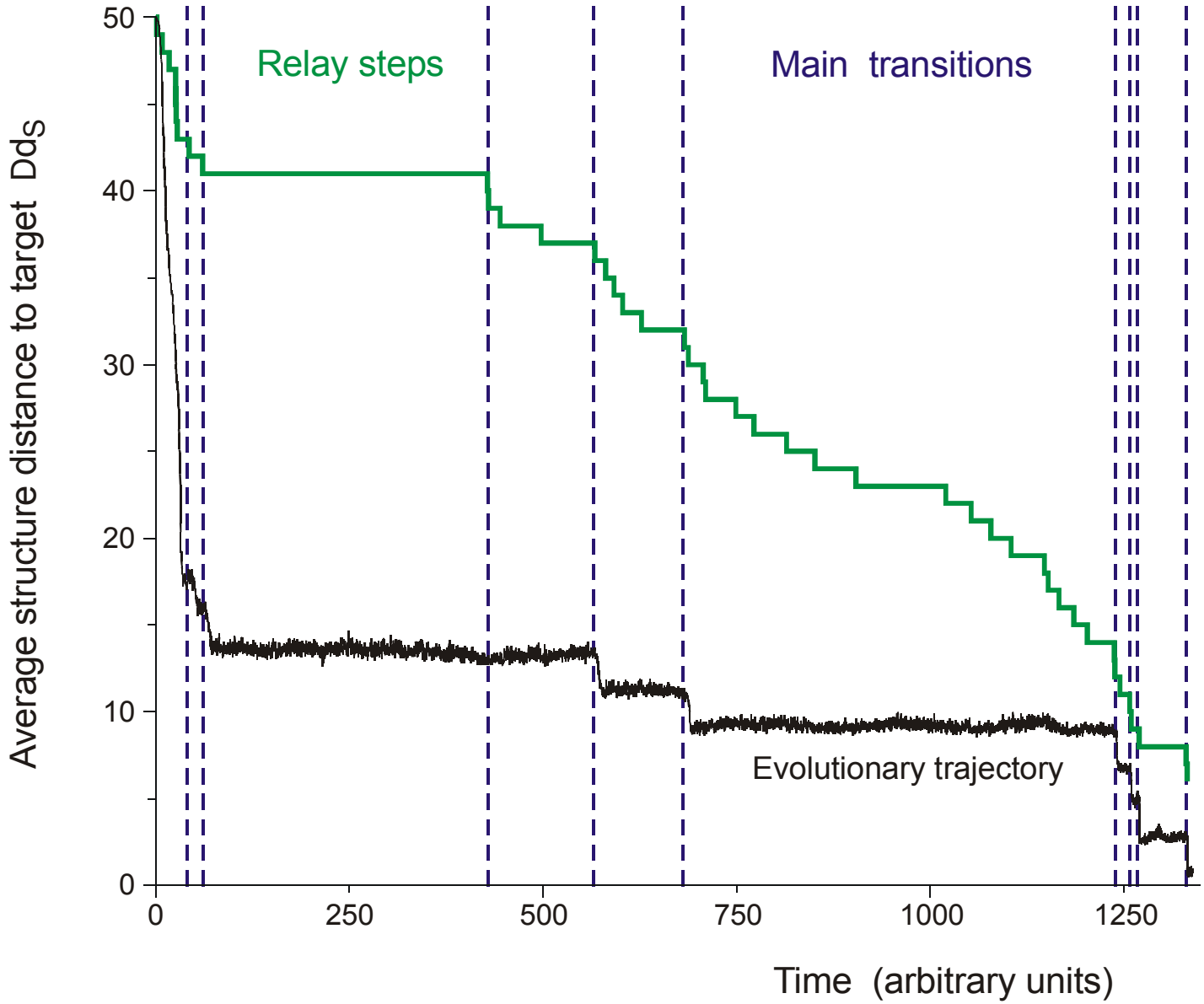
28 neutral point mutations during a long quasi-stationary epoch



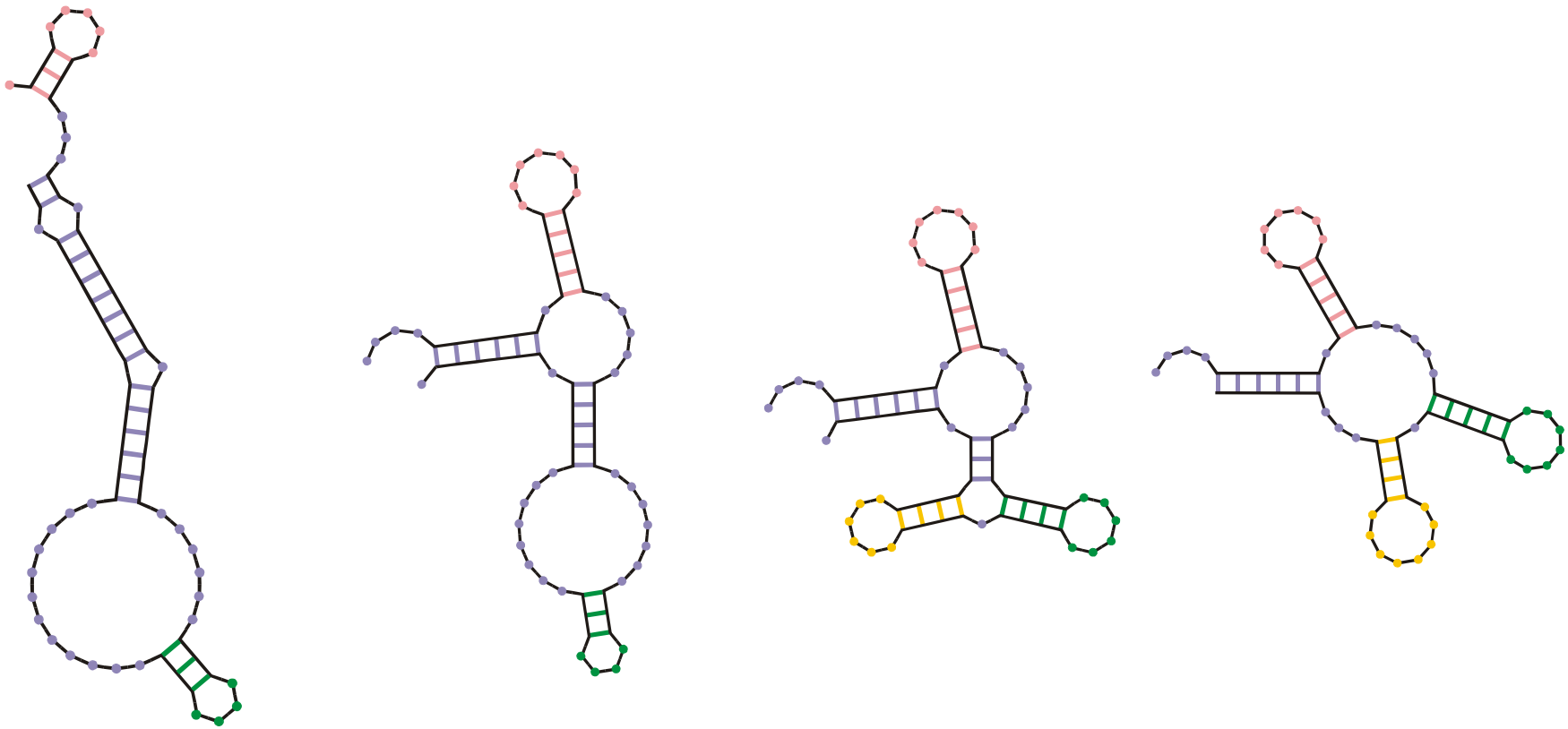

```
entry   GGUAUGGGCGUUGAAUAGUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUCCCGUACAGAA
8       .((((((((((((.........(((....)))........)))))).....((((.........))))))))))....
exit    GGUAUGGGCGUUGAAUAAUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUCCCAUACAGAA
entry   GGUAUGGGCGUUGAAUAAUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUACCAUACAGAA
9       .((((((.((((.........(((....)))........)))))).....((((.........))))).))))))....
exit    UGGAUGGACGUUGAAUAACAAGGUAUCGACCAAAACAACCAACGAGUAAGUGUGUACGCCCCACACACCGUCCCAAG
entry   UGGAUGGACGUUGAAUAACAAGGUAUCGACCAAACAACCAACGAGUAAGUGUGUACGCCCCACACAGCGUCCCAAG
10      .(((((..((((.........(((....)))........)))))).....((((.........))))).)))))....
exit    UGGAUGGACGUUGAAUAACAAGGUAUCGACCAAACAACCAACGAGUAAGUGUGUACGCCCCACACAGCGUCCCAAG
```

Transition inducing point mutations — **Neutral point mutations**

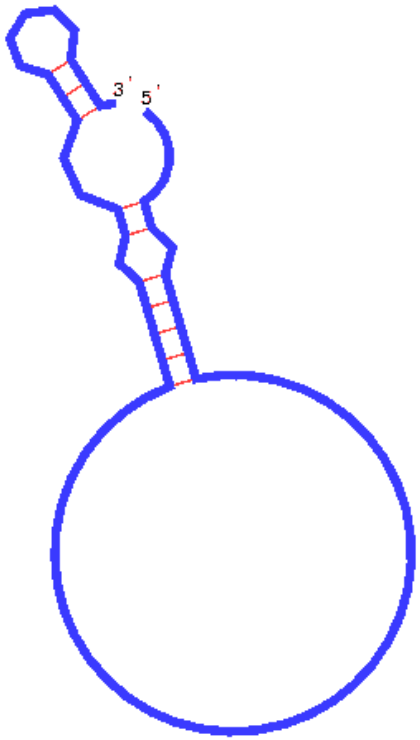
Neutral genotype evolution during phenotypic stasis



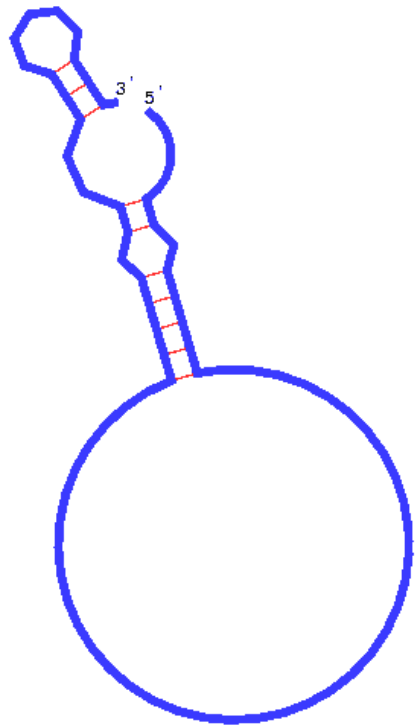
In silico optimization in the flow reactor: Main transitions



Three important steps in the formation of the tRNA clover leaf from a randomly chosen initial structure corresponding to three **main transitions**.

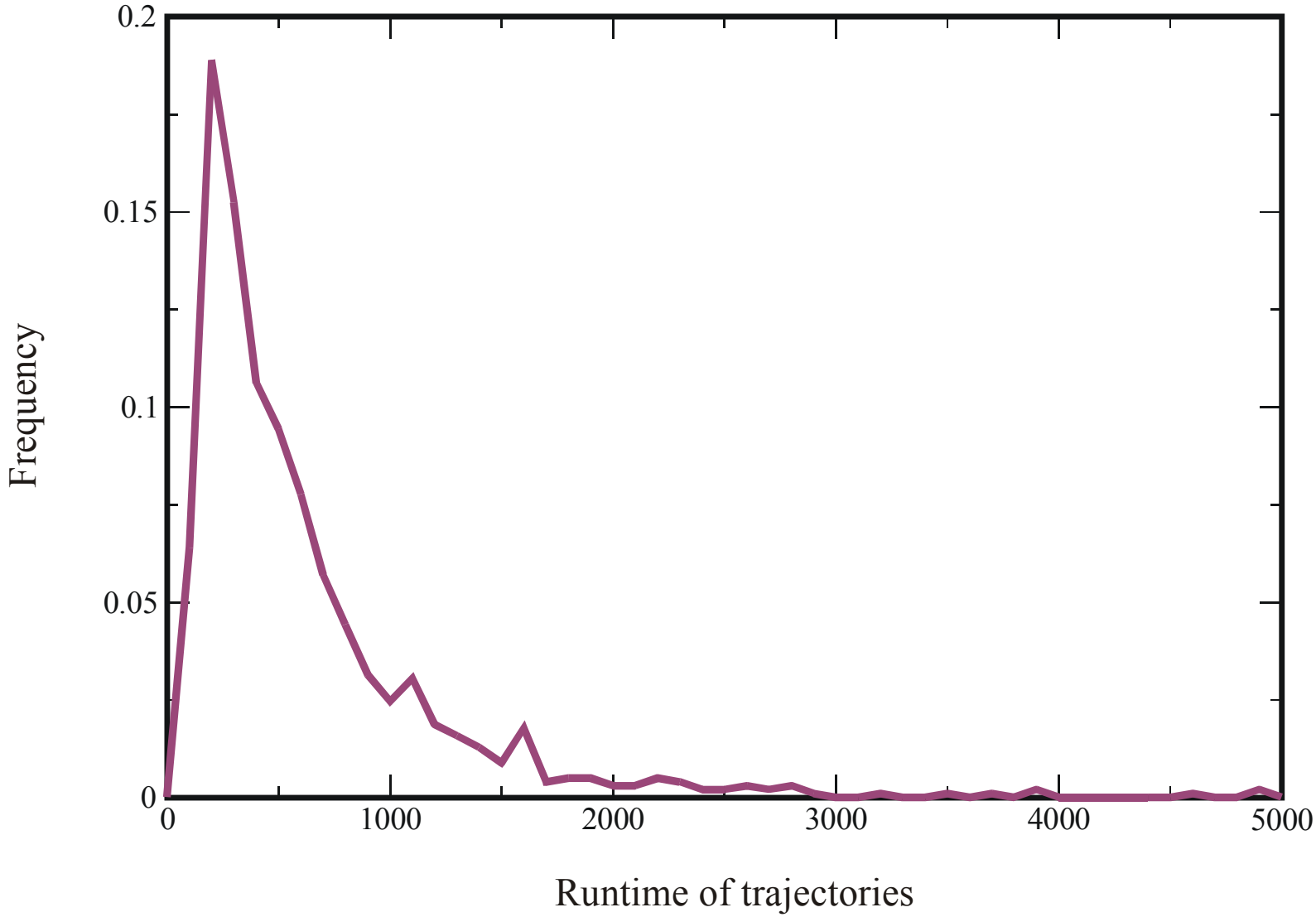


AUGC



GC

Movies of optimization trajectories over the **AUGC** and the **GC** alphabet

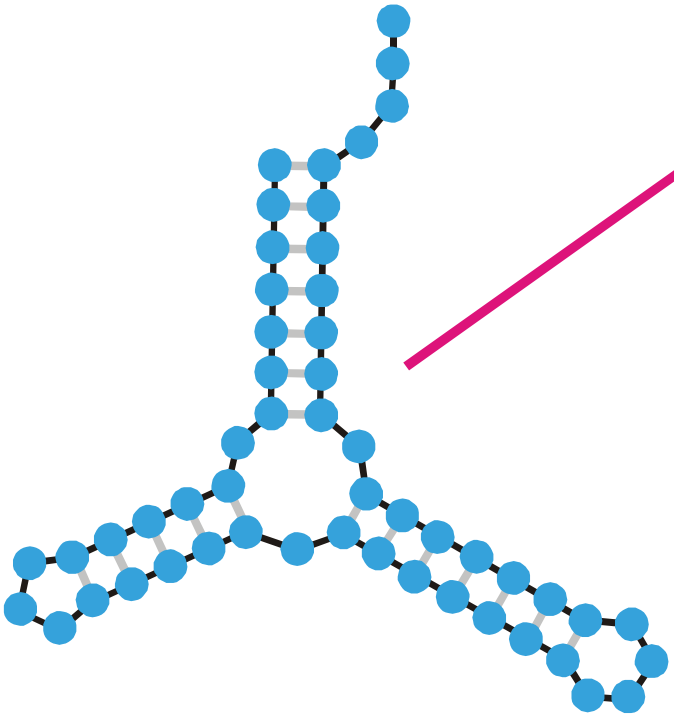


Statistics of the lengths of trajectories from initial structure to target (**AUGC**-sequences)

Alphabet	Runtime	Transitions	Main transitions	No. of runs
AUGC	385.6	22.5	12.6	1017
GUC	448.9	30.5	16.5	611
GC	2188.3	40.0	20.6	107

Statistics of trajectories and relay series (mean values of log-normal distributions)

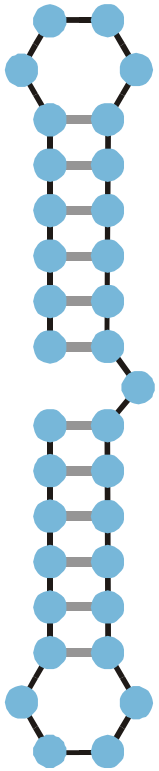
GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA



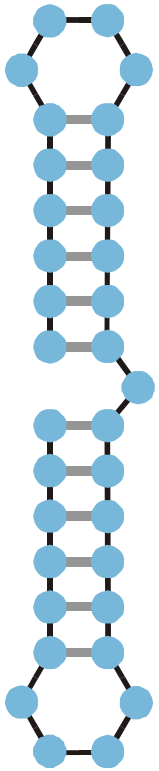
Minimum free energy criterion

Inverse folding of RNA secondary structures

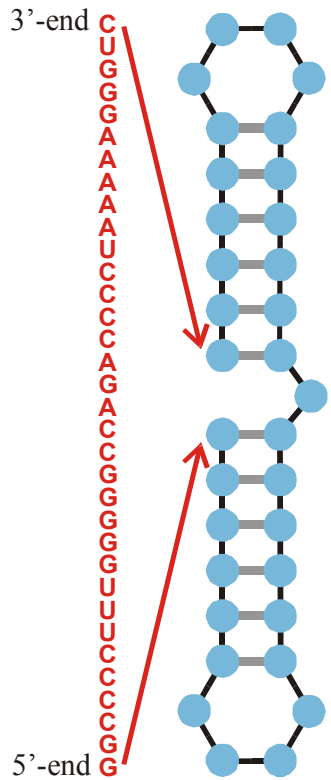
The idea of inverse folding algorithm is to search for sequences that form a given RNA secondary structure under the minimum free energy criterion.



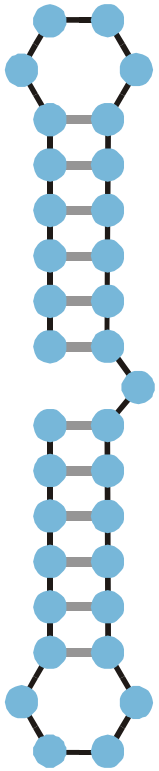
Structure



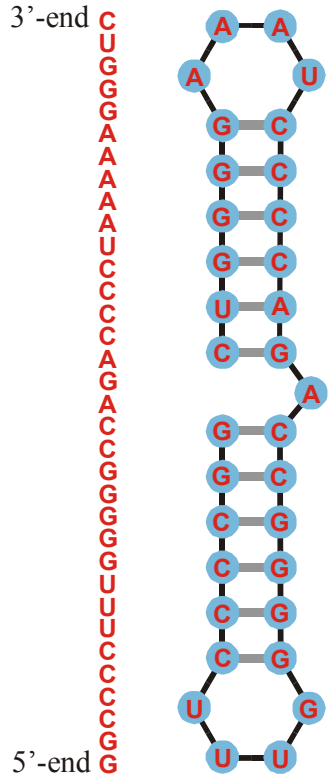
Structure



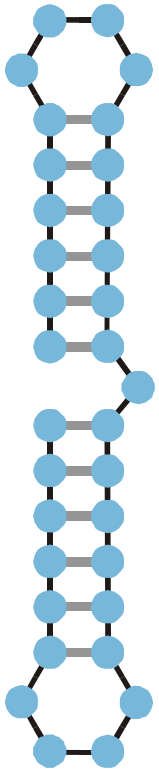
Compatible sequence



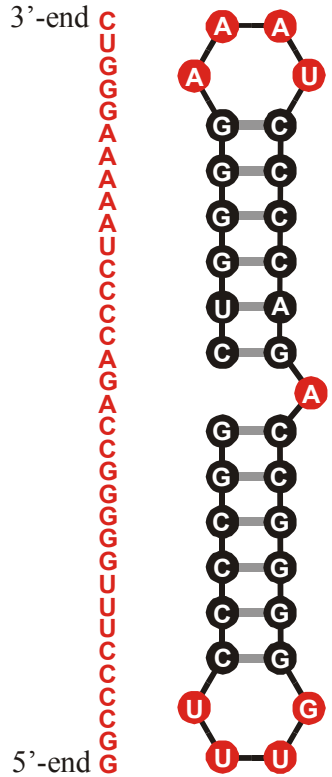
Structure



Compatible sequence



Structure

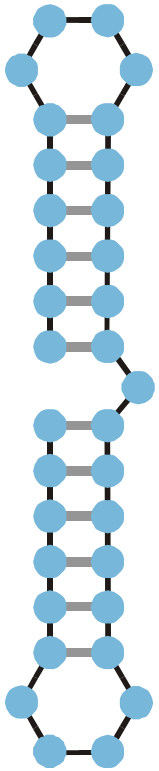


Single nucleotides: **A,U,G,C**

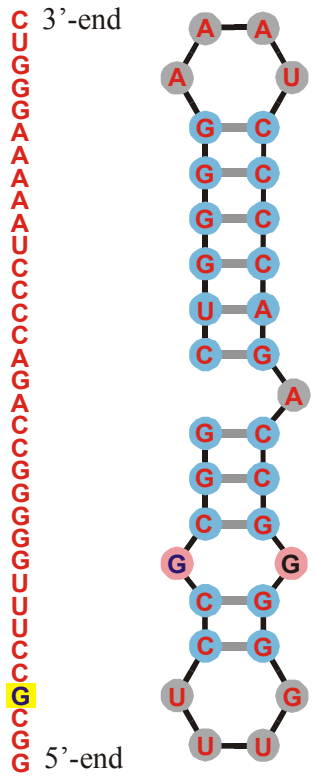
Base pairs:

AU , UA
GC , CG
GU , UG

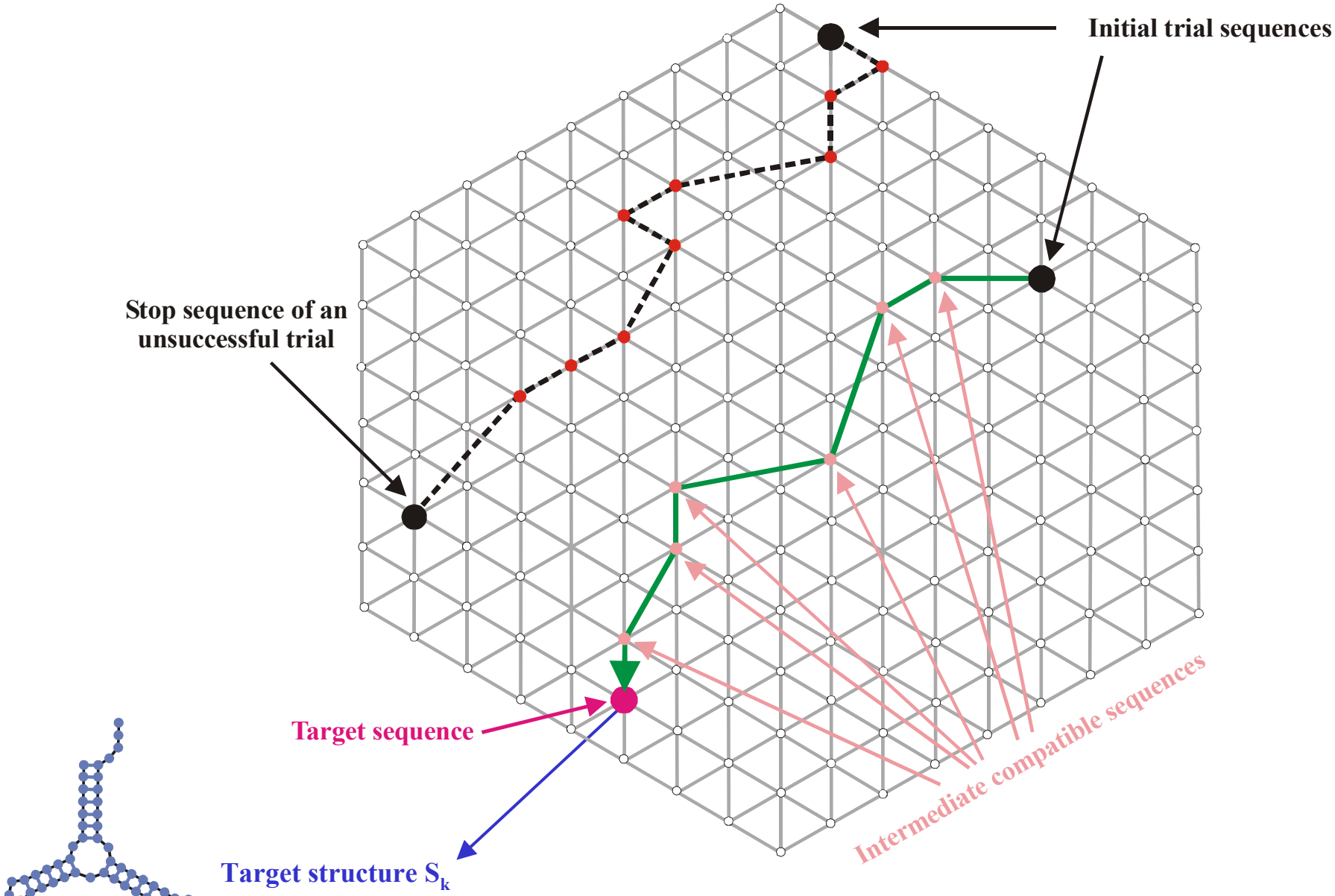
Compatible sequence



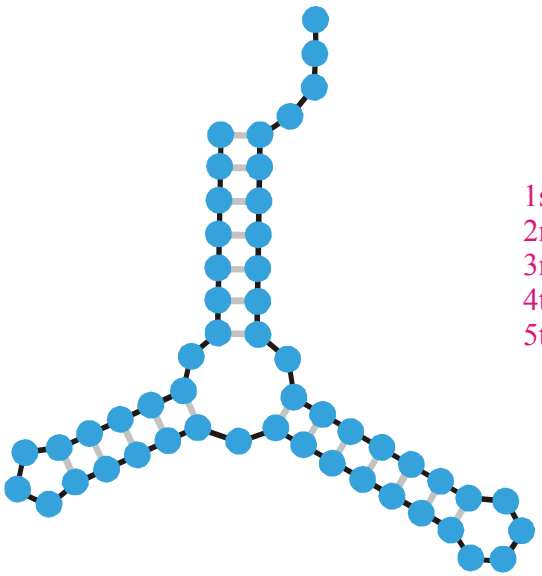
Structure



Incompatible sequence



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.

Theory of genotype – phenotype mapping

P. Schuster, W.Fontana, P.F.Stadler, I.L.Hofacker, ***From sequences to shapes and back: A case study in RNA secondary structures***. Proc.Roy.Soc.London **B 255** (1994), 279-284

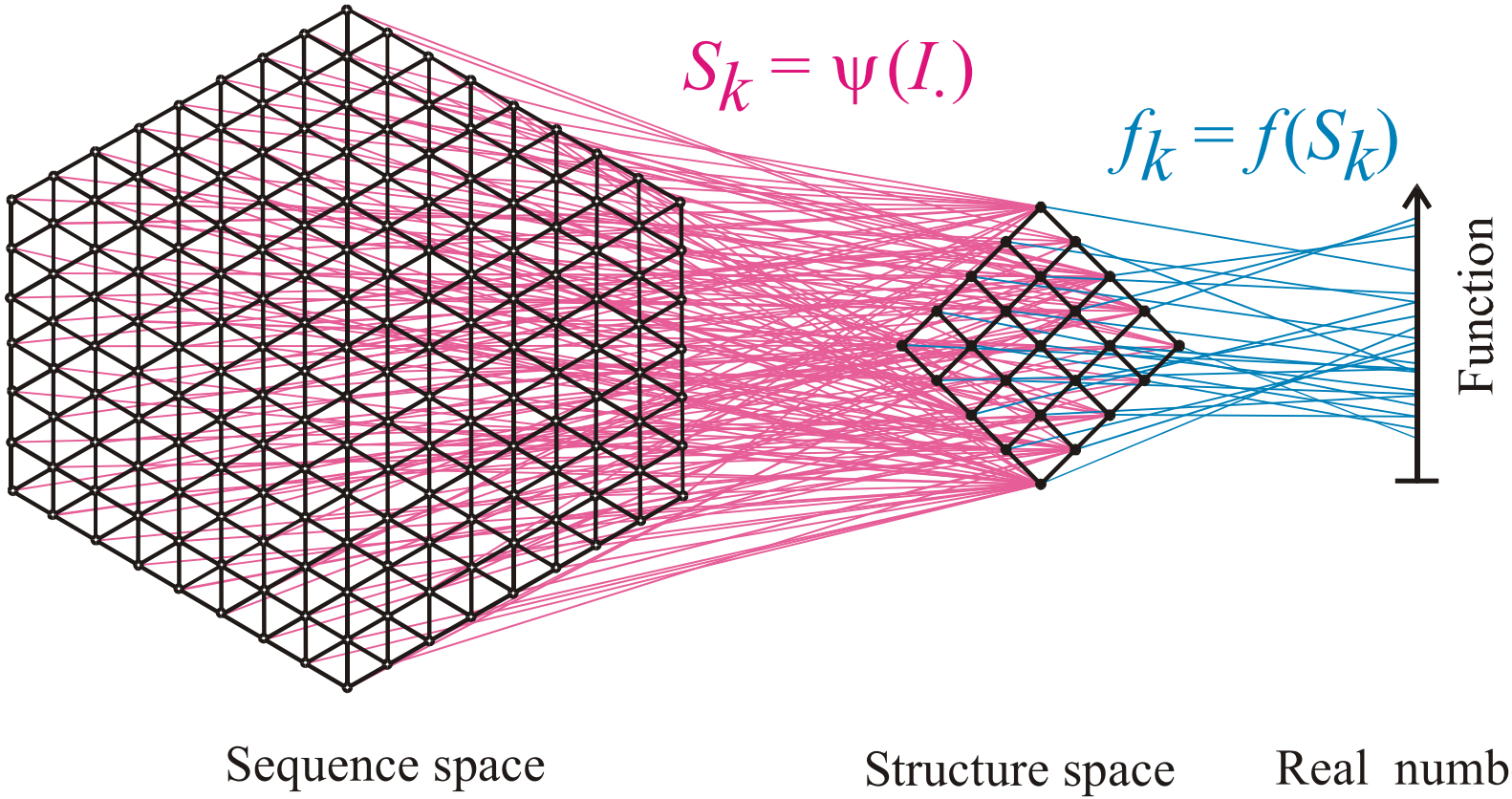
W.Grüner, R.Giegerich, D.Strothmann, C.Reidys, I.L.Hofacker, P.Schuster, ***Analysis of RNA sequence structure maps by exhaustive enumeration. I. Neutral networks***. Mh.Chem. **127** (1996), 355-374

W.Grüner, R.Giegerich, D.Strothmann, C.Reidys, I.L.Hofacker, P.Schuster, ***Analysis of RNA sequence structure maps by exhaustive enumeration. II. Structure of neutral networks and shape space covering***. Mh.Chem. **127** (1996), 375-389

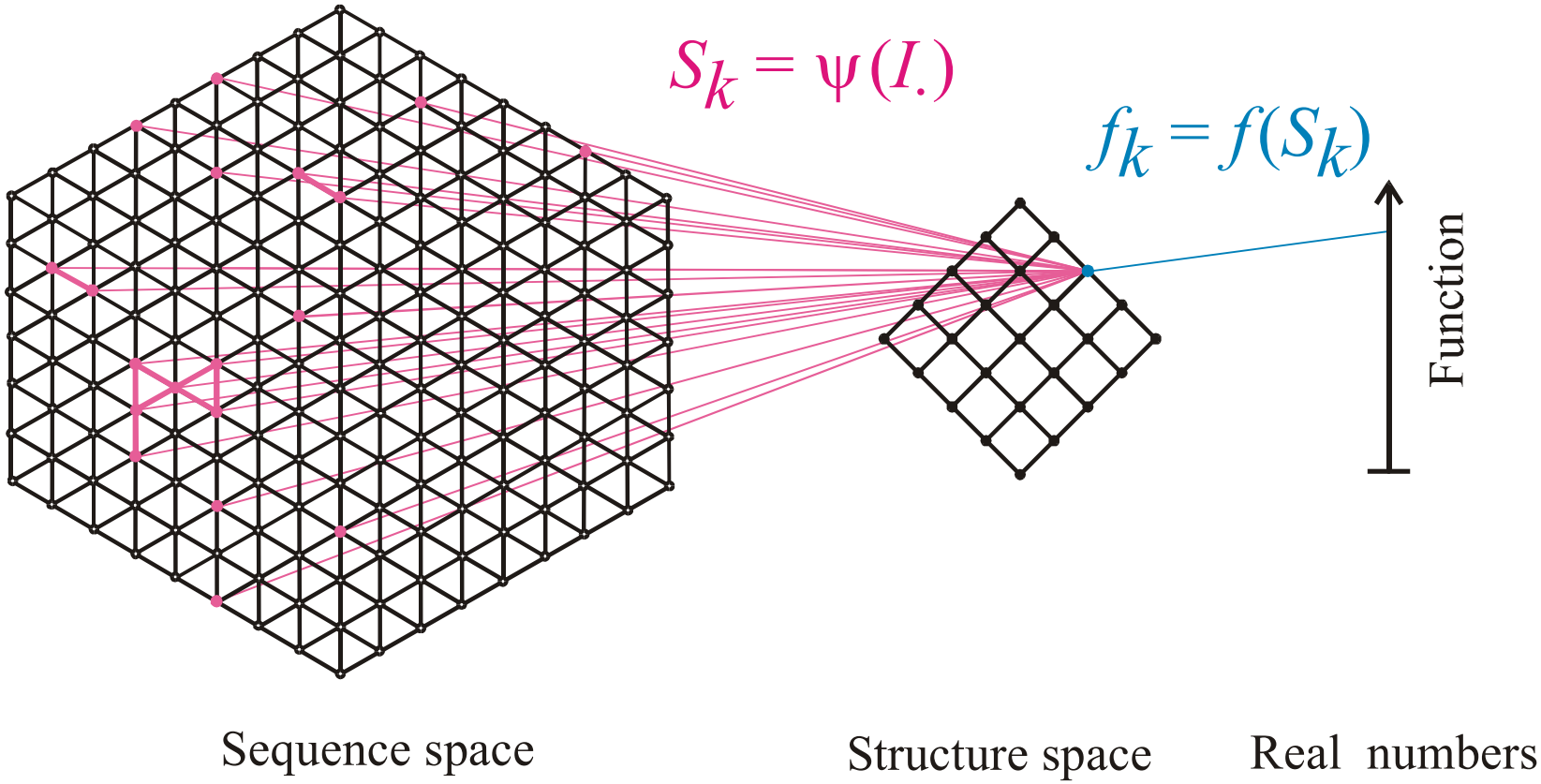
C.M.Reidys, P.F.Stadler, P.Schuster, ***Generic properties of combinatory maps***. Bull.Math.Biol. **59** (1997), 339-397

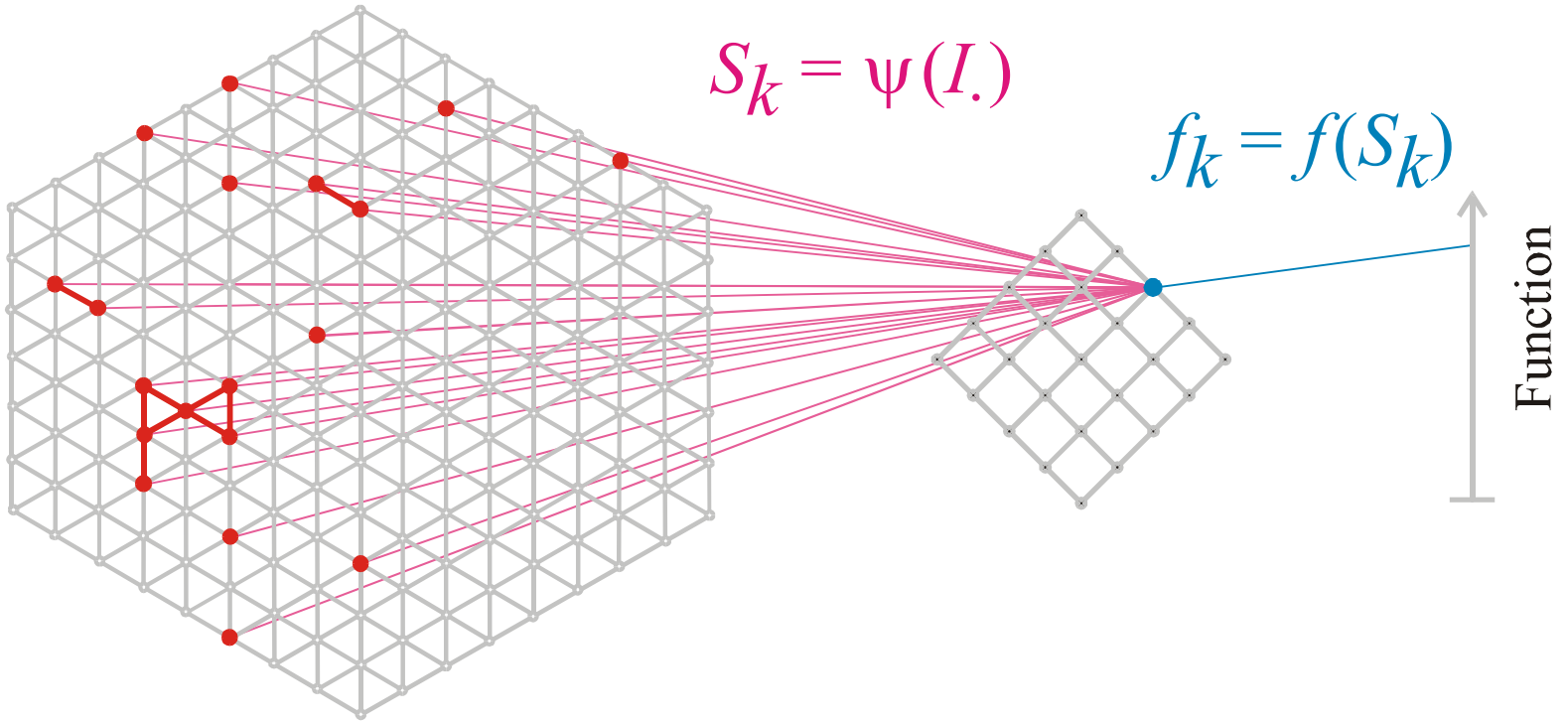
I.L.Hofacker, P. Schuster, P.F.Stadler, ***Combinatorics of RNA secondary structures***. Discr.Appl.Math. **89** (1998), 177-207

C.M.Reidys, P.F.Stadler, ***Combinatory landscapes***. SIAM Review **44** (2002), 3-54



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**

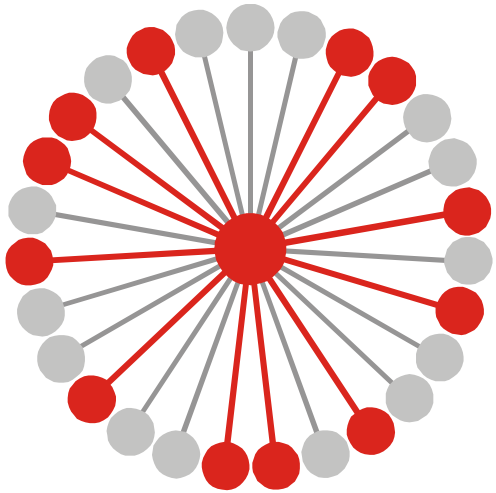
Neutral networks are sets of sequences forming the same structure. $\mathbf{G_k}$ is the pre-image of the structure S_k in sequence space:

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

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.



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

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

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

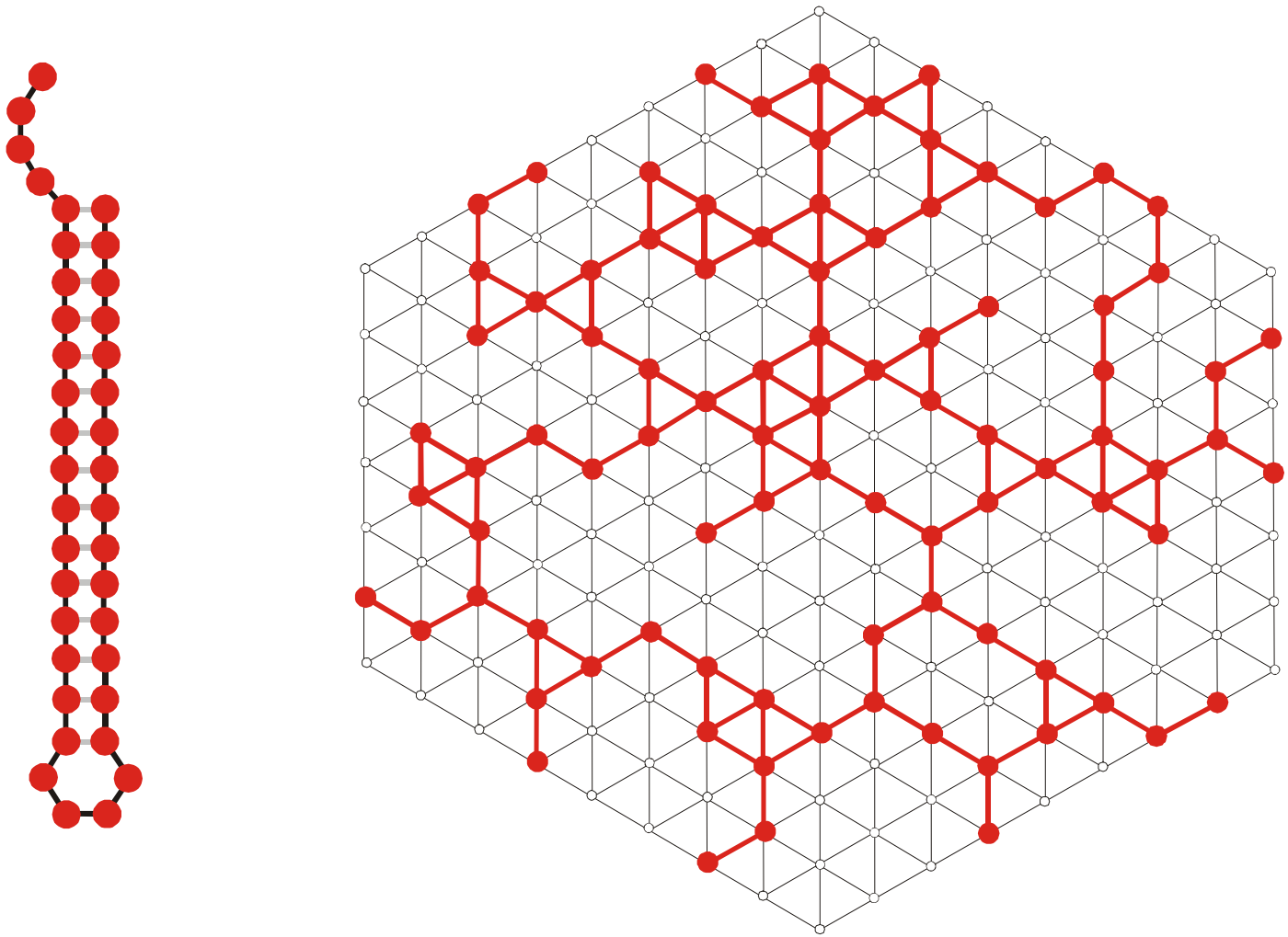
Alphabet size k : **AUGC** ñ k = 4

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

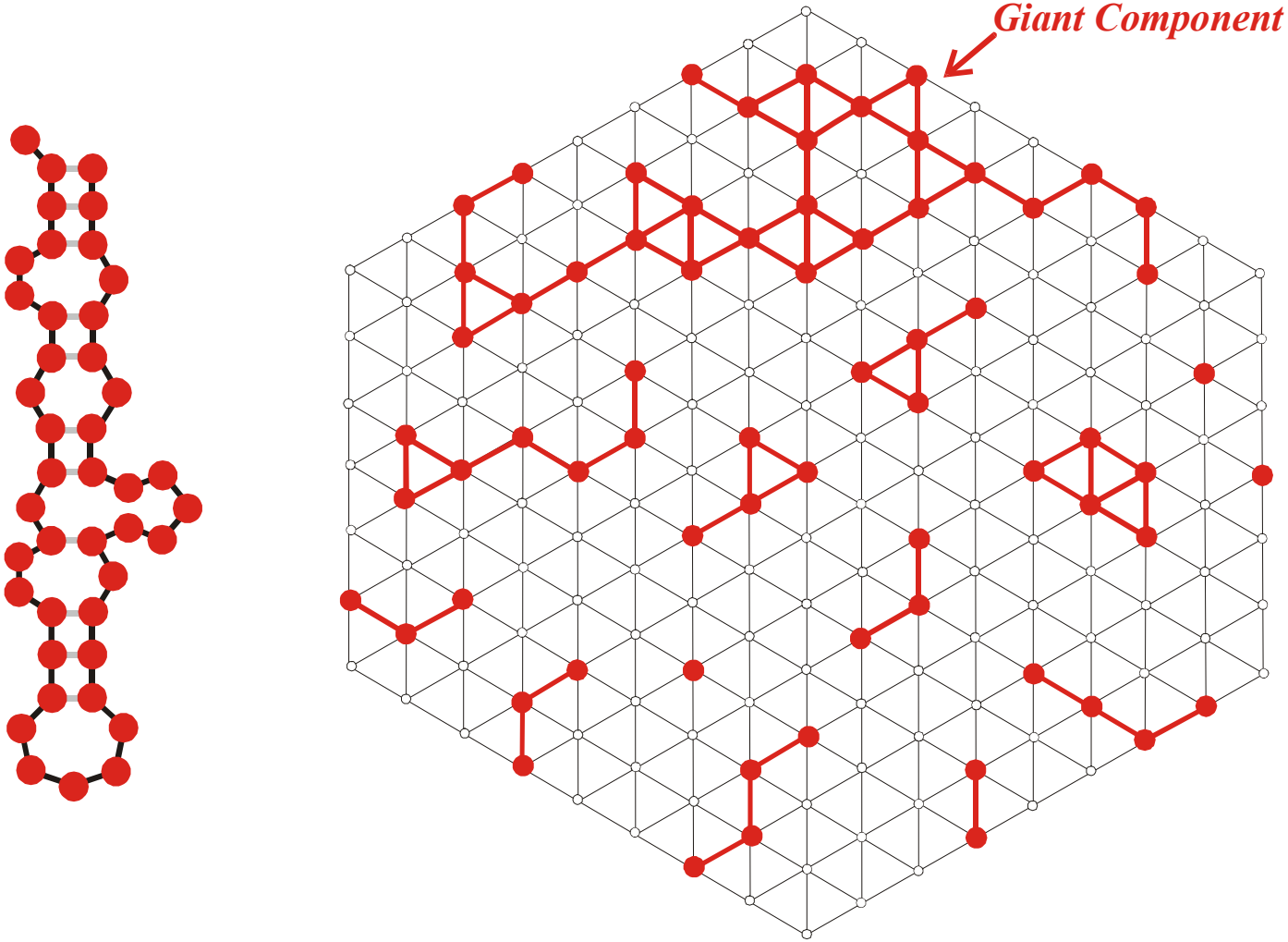
$\bar{\lambda}_k < \lambda_{cr}$ .... network G_k is **not** connected

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

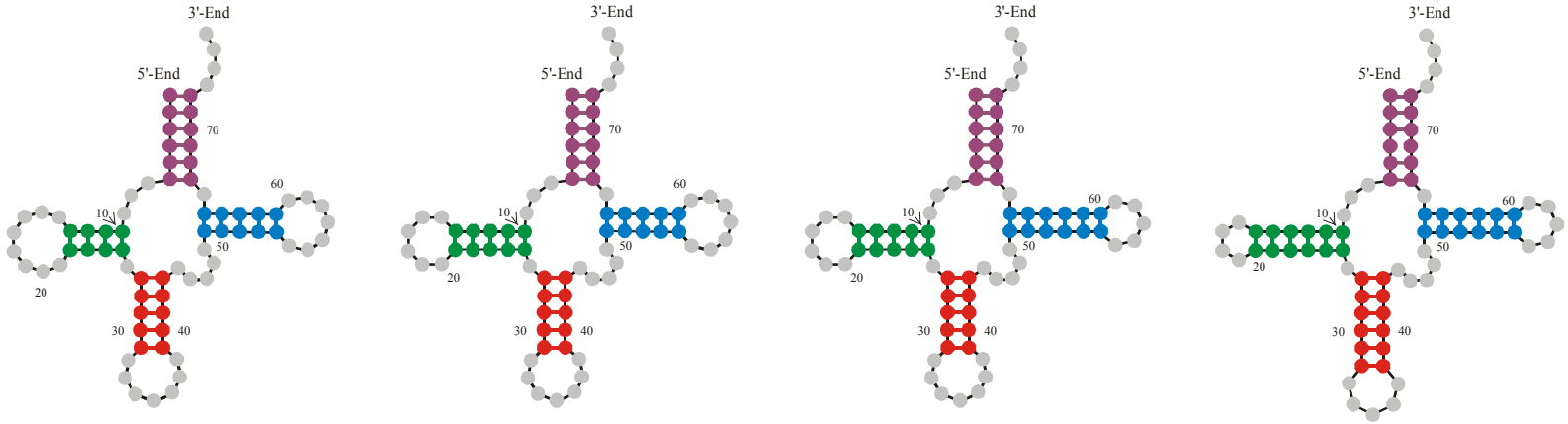
Mean degree of neutrality and connectivity of neutral networks



A connected neutral network



A multi-component neutral network



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

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.

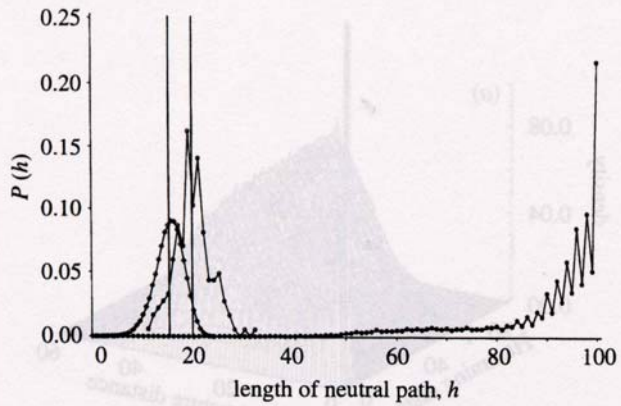
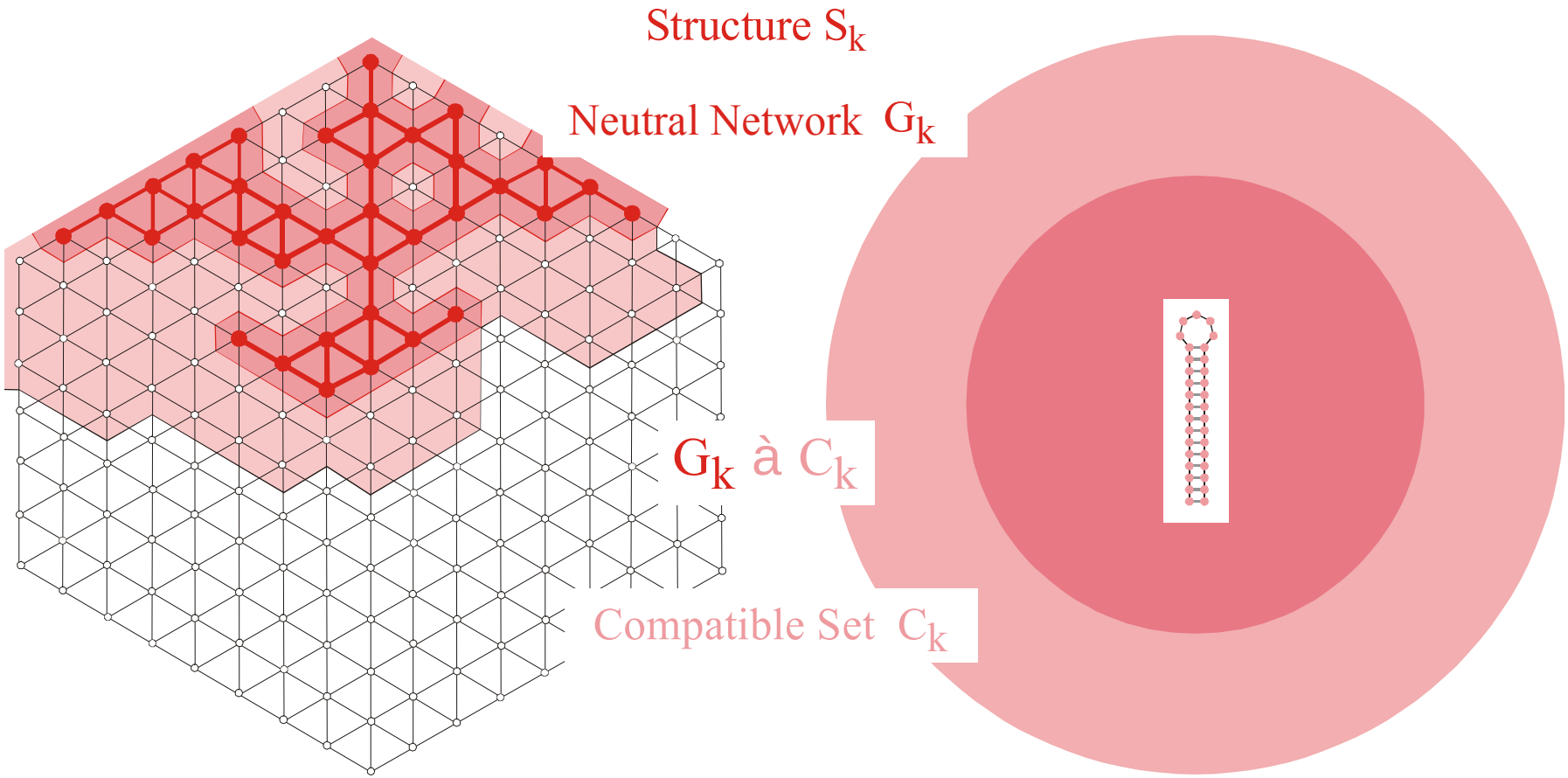


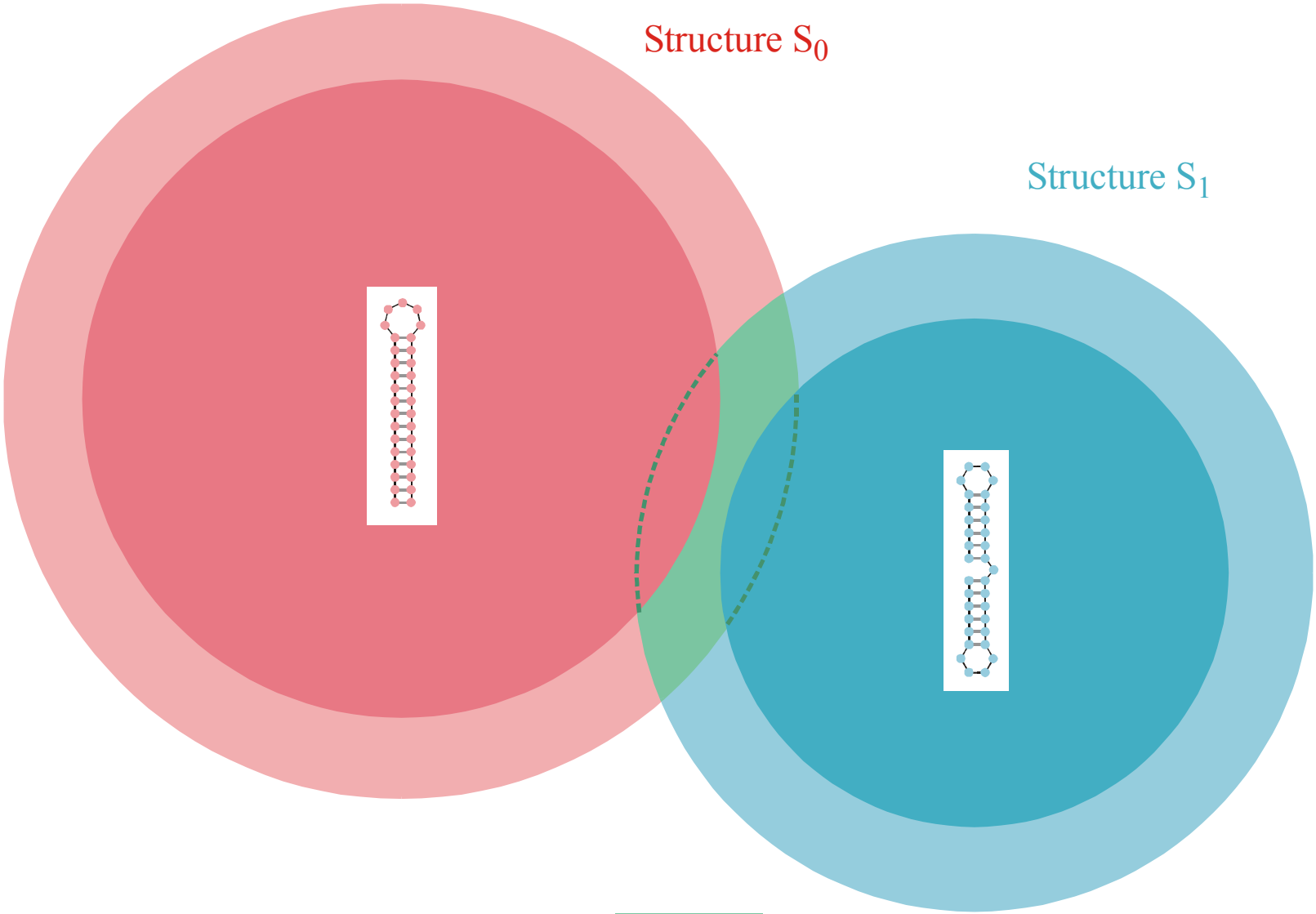
Figure 4. Neutral paths. A neutral path is defined by a series of nearest neighbour sequences that fold into identical structures. Two classes of nearest neighbours are admitted: neighbours of Hamming distance 1, which are obtained by single base exchanges in unpaired stretches of the structure, and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



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.

0092-8240/97 $17.00 + 0.00

S0092-8240(96)00089-4

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

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

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

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

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

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

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

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

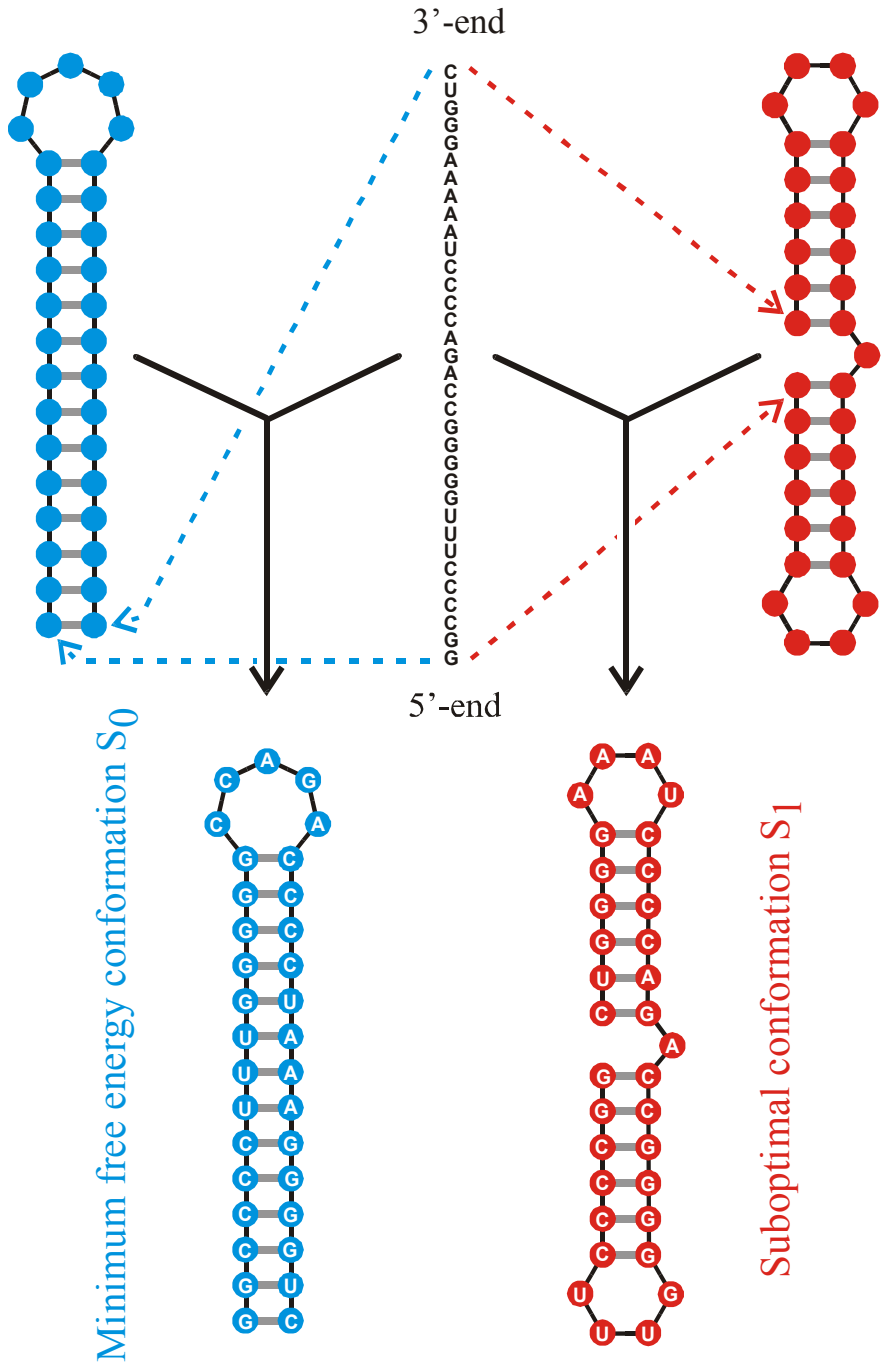
$$\mathbf{C}[s] \cap \mathbf{C}[s'] \neq \varnothing.$$

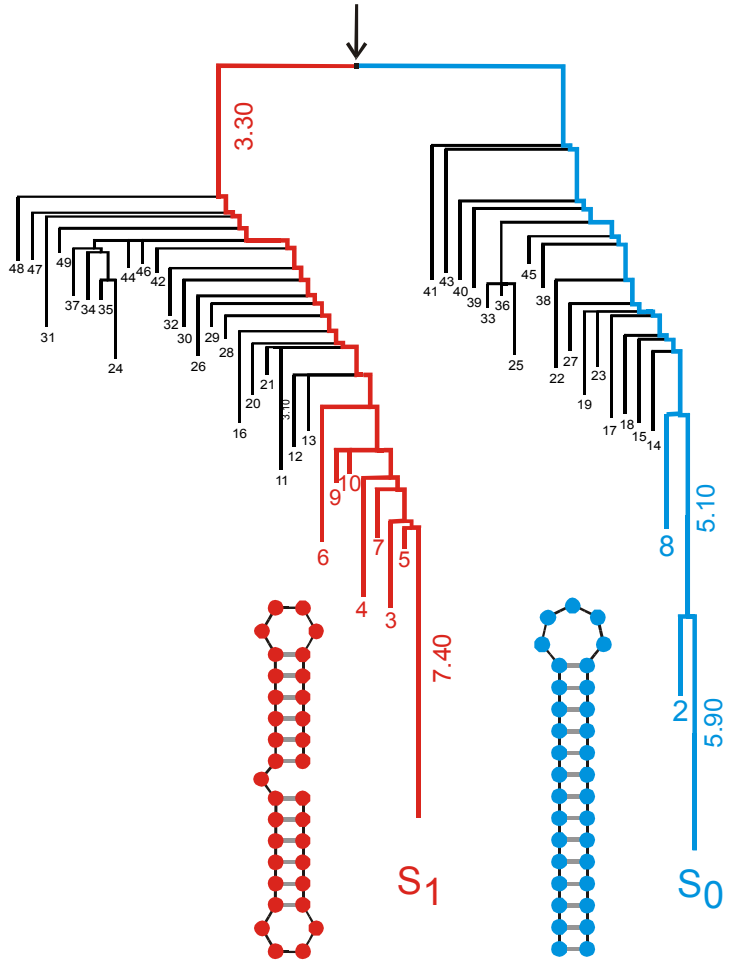
Proof. Suppose that the alphabet admits only the complementary base pair $[XY]$ and we ask for a sequence x compatible to both s and s'. Then $\gamma(s, s') \cong D_m$ operates on the set of all positions $\{x_1, \ldots, x_n\}$. Since we have the operation of a dihedral group, the orbits are either cycles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

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

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

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





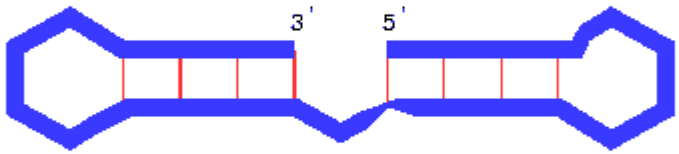
basin '1'

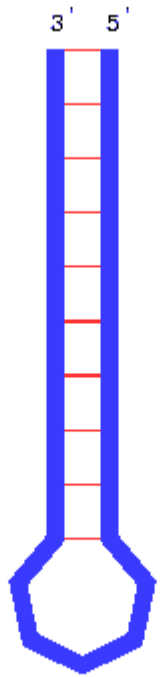
basin '0'

Barrier tree for two
long living structures

long living
metastable structure

minimum free energy
structure





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

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