Complexity in Molecular Systems

Peter Schuster

Institut für Theoretische Chemie, Universität Wien, Austria
and
The Santa Fe Institute, Santa Fe, New Mexico, USA

Academia Europaea – Klaus Tschira Foundation
„Complexity“

Heidelberg, 25.– 26.04.2008

Web-Page for further information:

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

1. Autocatalytic chemical reactions
2. Replication and biological information
3. Quasispecies and error thresholds
4. Neutral networks in evolution
5. Evolutionary optimization
6. Genetic regulation and metabolism

1. **Autocatalytic chemical reactions**
2. Replication and biological information
3. Quasispecies and error thresholds
4. Neutral networks in evolution
5. Evolutionary optimization
6. Genetic regulation and metabolism

$* \rightarrow A$

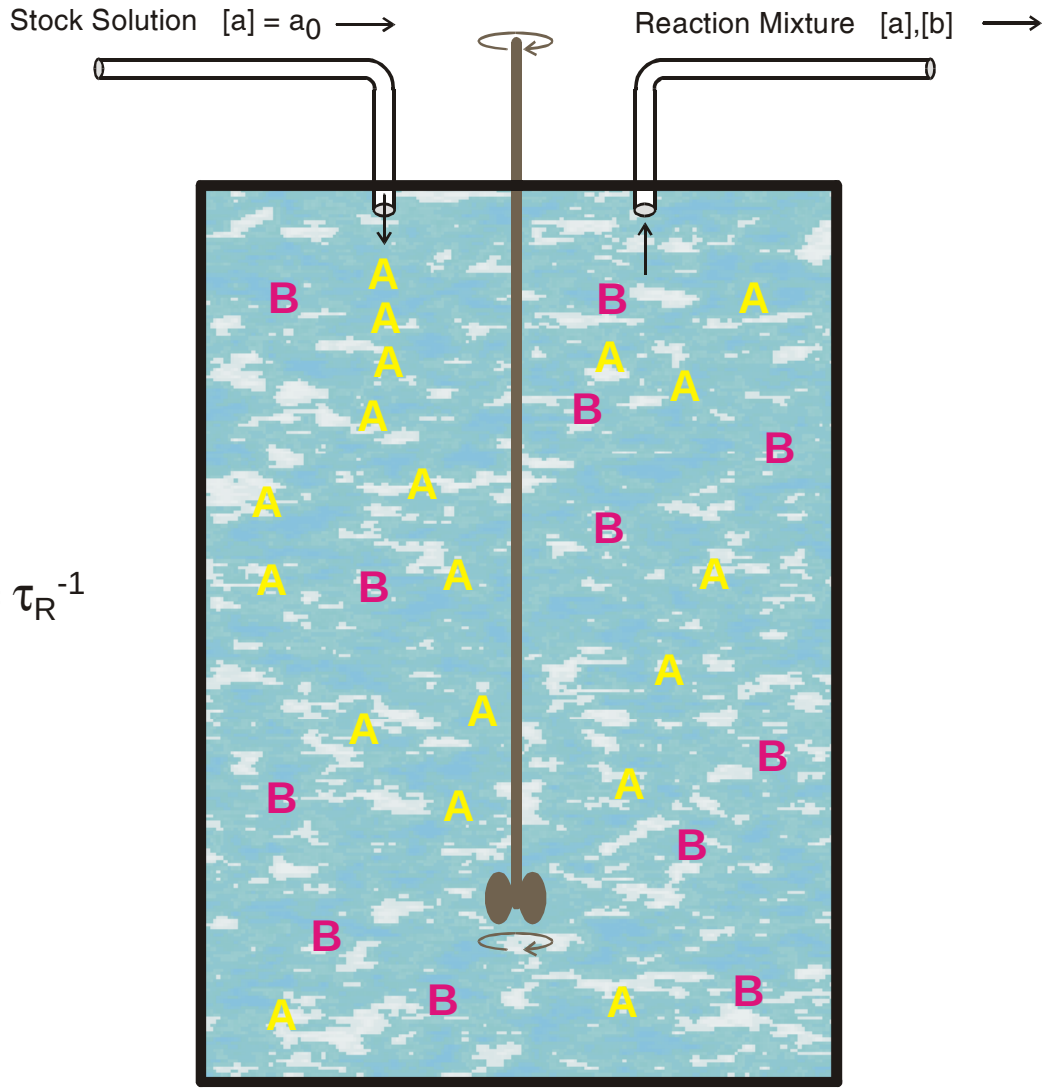
$A \rightarrow B$

$A \rightarrow \varnothing$

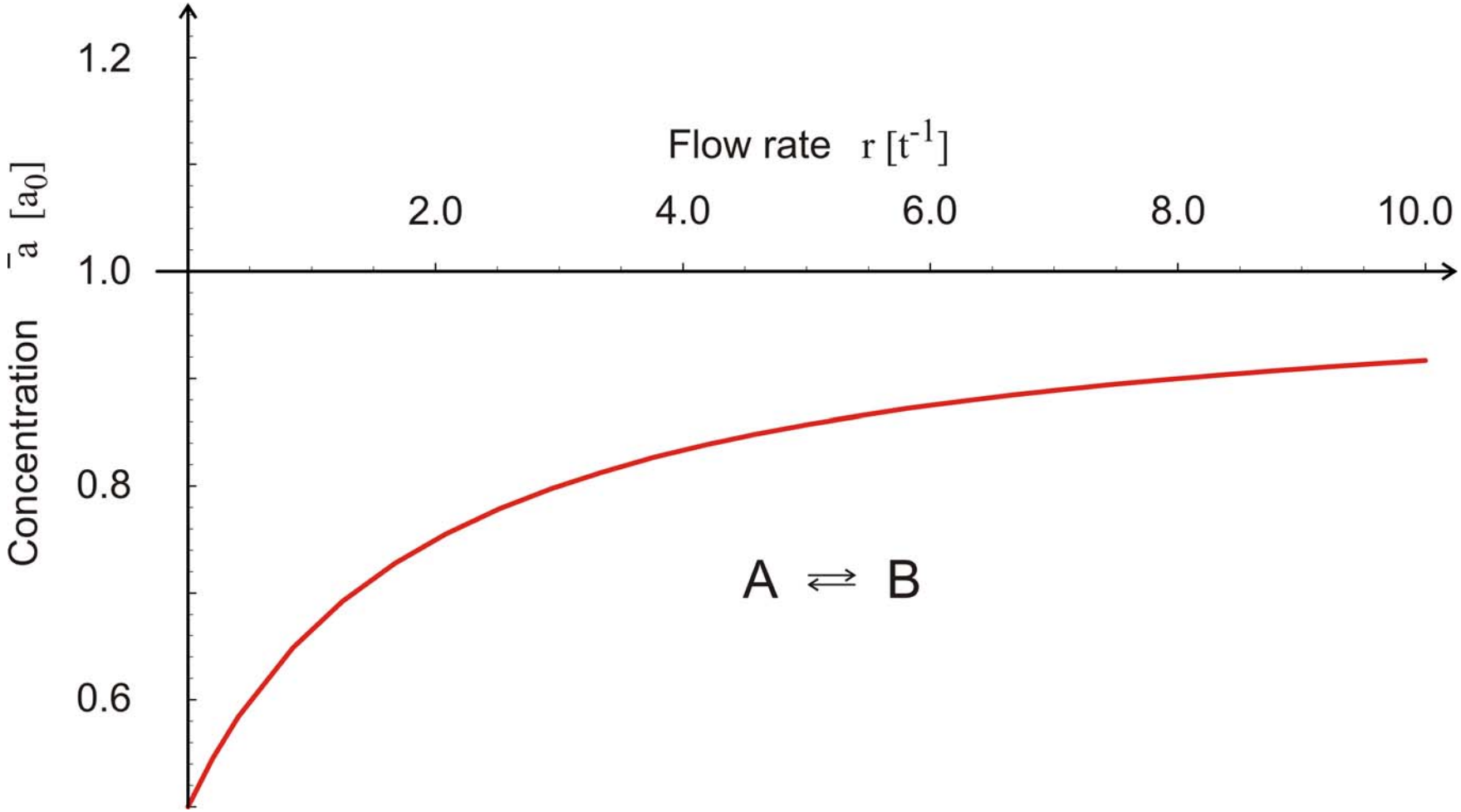
$B \rightarrow A$

$B \rightarrow \varnothing$

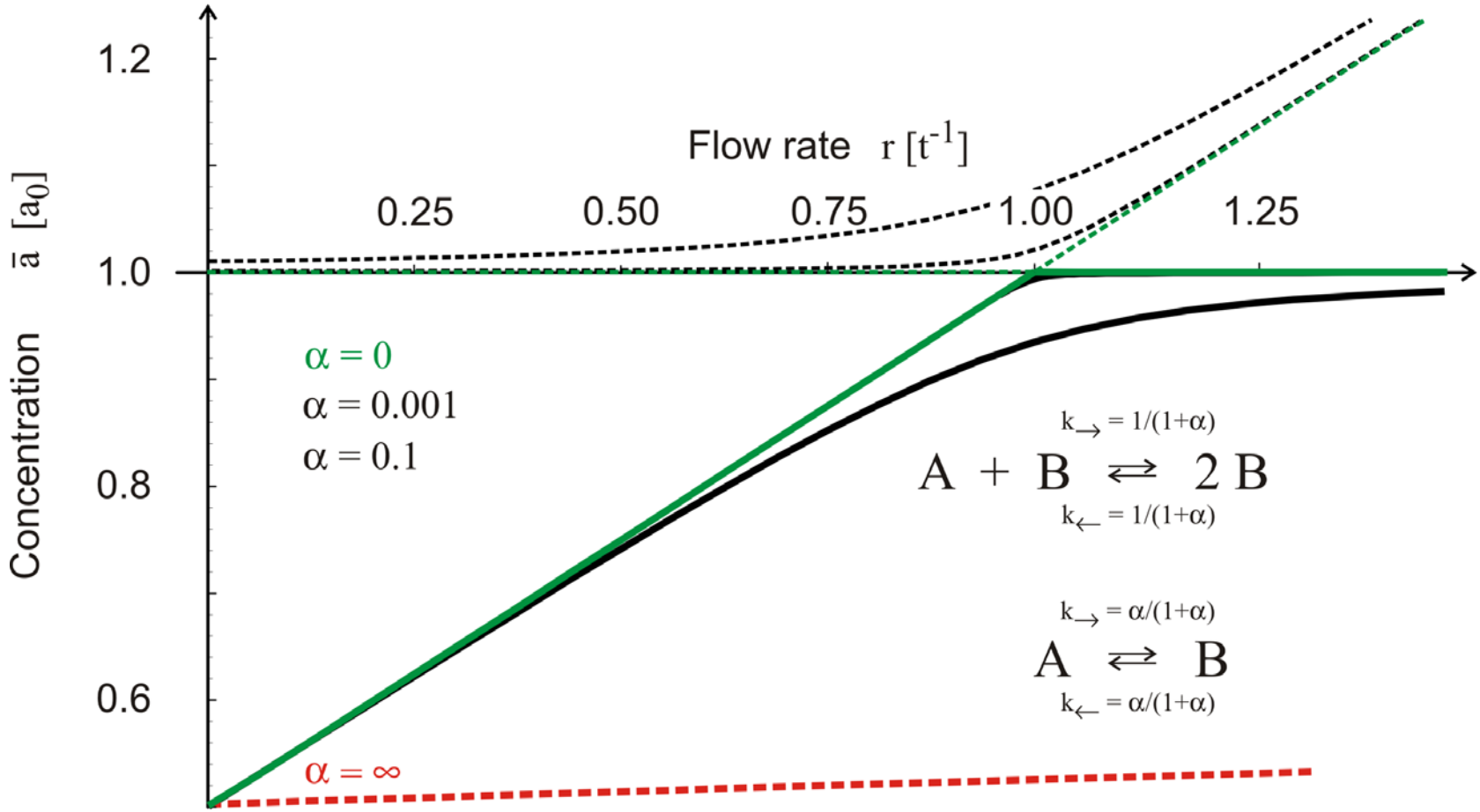
Flow rate $r = \tau_R^{-1}$



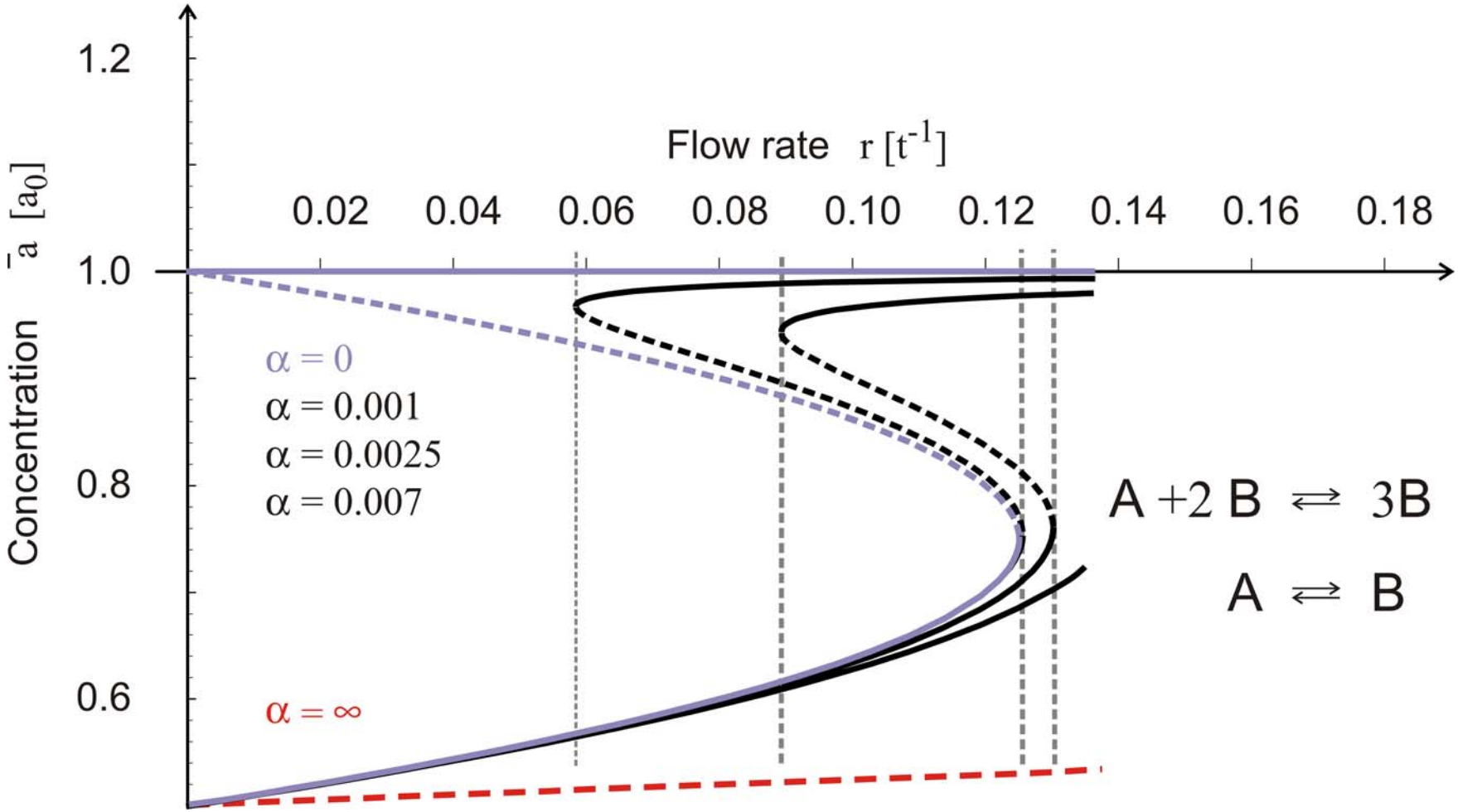
Reactions in the continuously stirred tank reactor (CSTR)



Reversible first order reaction in the flow reactor



Autocatalytic second order reaction and uncatalyzed reaction in the flow reactor



Autocatalytic third order reaction and uncatalyzed reaction in the flow reactor

$$
\begin{aligned}
A &\longrightarrow X \\
2X + Y &\longrightarrow 3X \\
B + X &\longrightarrow Y + C \\
X &\longrightarrow D
\end{aligned}
$$

The Brusselator model

G. Nicolis, I. Prigogine.
Self-organization in nonequilibrium systems. From dissipative structures to order through fluctuations.
John Wiley & Sons, New York 1977

$$
\begin{aligned}
\dot{X} &= 1 - (b+1)X + aX^2Y \\
\dot{Y} &= bX - aX^2Y
\end{aligned}
$$

Mechanism for the bromate–sulfite–ferrocyanide reaction

Number	Reaction
(B1)	$BrO_3^- + HSO_3^- \longrightarrow HBrO_2 + SO_4^{2-}$
(B2)	$HBrO_2 + Br^- + H^+ \longrightarrow 2HOBr$
(B3)	$HOBr + Br^- + H^+ \longrightarrow Br_2 + H_2O$
(B4)	$Br_2 + H_2O \longrightarrow HOBr + Br^- + H$
(B5)	$2HBrO_2 \longrightarrow BrO_3^- + HOBr + H^+$
(B6)	$Br_2 + HSO_3^- + H_2O \longrightarrow 2Br^- + SO_4^{2-} + 3H^+$
(B7)	$H^+ + SO_3^{2-} \longrightarrow HSO_3^-$
(B8)	$HSO_3^- \longrightarrow H^+ + SO_3^{2-}$
(B9)	$BrO_3^- + 2Fe(CN)_6^{4-} + 3H^+ \longrightarrow HBrO_2 + 2Fe(CN)_6^{3-} + H_2O$

Reaction mechanism of an autocatalytic reaction

F. Sagués, I.R. Epstein. *Dalton Trans.* **2003**:1201-1217.

Mechanism of the chlorite–iodide and related reactions[a]

Number	Reaction	Rate law
(M1)	$ClO_2 + I^- \longrightarrow ClO_2^- + \frac{1}{2}I_2$	$v_1 = 6 \times 10^3[ClO_2][I^-]$
(M2)	$I_2 + H_2O \rightleftharpoons HOI + I^- + H^+$	$v_{2a} = 1.98 \times 10^{-3}[I_2]/[H^+] - 3.67 \times 10^9[HOI][I^-]$
		$v_{2b} = 5.52 \times 10^{-2}[I_2] - 3.48 \times 10^9[H_2OI^+][I^-]$
(M3)	$HClO_2 + I^- + H^+ \longrightarrow HOI + HCl$	$v_3 = 7.8[HClO_2][I^-]$
(M4)	$HClO_2 + HOI \longrightarrow HIO_2 + HOCl$	$v_4 = 6.9 \times 10^7[HClO_2][HOI]$
(M5)	$HClO_2 + HIO_2 \longrightarrow IO_3^- + HOCl + H^+$	$v_5 = 1.5 \times 10^6[HClO_2][HIO_2]$
(M6)	$HOCl + I^- \longrightarrow HOI + Cl^-$	$v_6 = 4.3 \times 10^8[HOCl][I^-]$
(M7)	$HOCl + HIO_2 \longrightarrow IO_3^- + Cl^- + 2H^+$	$v_7 = 1.5 \times 10^3[HOCl][HIO_2]$
(M8)	$HIO_2 + I^- + H^+ \rightleftharpoons 2HOI$	$v_8 = 1.0 \times 10^9[HIO_2][I^-][H^+] - 22[HOI]^2$
(M9)	$2HIO_2 \longrightarrow IO_3^- + HOI + H^+$	$v_9 = 25[HIO_2]^2$
(M10)	$HIO_2 + H_2OI^+ \longrightarrow IO_3^- + I^- + 3H^+$	$v_{10} = 110[HIO_2][H_2OI^+]$
(M11)	$HOCl + Cl^- + H^+ \rightleftharpoons Cl_2 + H_2O$	$v_{11} = 2.2 \times 10^4[HOCl][Cl^-][H^+] - 22[Cl_2]$
(M12)	$Cl_2 + I_2 + 2H_2O \longrightarrow 2HOI + 2Cl^- + 2H^+$	$v_{12} = 1.5 \times 10^5[Cl_2][I_2]$
(M13)	$Cl_2 + HOI + H_2O \longrightarrow HIO_2 + 2Cl^- + 2H^+$	$v_{13} = 1.0 \times 10^6[Cl_2][HOI]$
	Rapid equilibria	
(M14)	$HClO_2 \rightleftharpoons ClO_2^- + H^+$	$K_{14} = [ClO_2^-][H^+]/[HClO_2] = 2.0 \times 10^{-2}$
(M15)	$H_2OI^+ \rightleftharpoons HOI + H^+$	$K_{15} = [HOI][H^+]/[H_2OI^+] = 3.4 \times 10^{-2}$
(M16)	$I_2 + I^- \rightleftharpoons I_3^-$	$K_{16} = [I3^-]/[I_2][I^-] = 7.4 \times 10^2$

[a] All concentrations in M, times in s.

Reaction mechanism of an autocatalytic reaction

F. Sagués, I.R. Epstein. *Dalton Trans.* **2003**:1201-1217.

1. $Br^- + 2H^+ + BrO_3^- \rightleftharpoons HOBr + HBrO_2$

2. $Br^- + HBrO_2 + H^+ \rightleftharpoons 2HOBr$

3. $Br^- + HOBr + H^+ \rightleftharpoons Br_2 + H_2O$

4. $H^+ + CH_2(COOH)_2 \rightleftharpoons (OH)_2C=CHCOOH + H^+$

5. $Br_2 + (OH)_2C=CHCOOH \rightleftharpoons H^+ + Br^- + BrCH(COOH)_2$

6. $HOBr + (OH)_2C=CHCOOH \rightleftharpoons H_2O + BrCH(COOH)_2$

7. $HBrO_2 + BrO_3^- + H^+ \rightleftharpoons 2BrO_2\cdot + H_2O$

8. $BrO_2\cdot + Ce(III) + H^+ \rightleftharpoons Ce(IV) + HBrO_2$

9. $Ce(IV) + BrO_2\cdot + H_2O \rightleftharpoons BrO_3^- + 2H^+ + Ce(III)$

10. $2HBrO_2 \rightleftharpoons HOBr + BrO_3^- + H^+$

11. $Ce(IV) + CH_2(COOH)_2 \rightarrow \cdot CH(COOH)_2 + Ce(III) + H^+$

12. $\cdot CH(COOH)_2 + Ce(IV) + H_2O \rightarrow HOCH(COOH)_2 + Ce(III) + H^+$

13. $Ce(IV) + BrCH(COOH)_2 + H_2O \rightarrow Br^- + HOC(COOH)_2 + Ce(III) + H^+$

14. $\cdot CH(COOH)_2 + BrCH(COOH)_2 + H_2O \rightarrow HOC(COOH)_2 + CH_2(COOH)_2 + Br^- + H^+$

15. $HOC(COOH)_2 + Ce(IV) \rightarrow O=C(COOH)_2 + Ce(III) + H^+$

16. $HOC(COOH)_2 + BrCH(COOH)_2 + H_2O \rightarrow HOCH(COOH)_2 + Br^- + HOC(COOH)_2 + H^+$

17. $Ce(IV) + HOCH(COOH)_2 \rightarrow HOC(COOH)_2 + Ce(III) + H^+$

18. $Ce(IV) + O=C(COOH)_2 \rightarrow O=C(COO\cdot)(COOH) + Ce(III) + H^+$

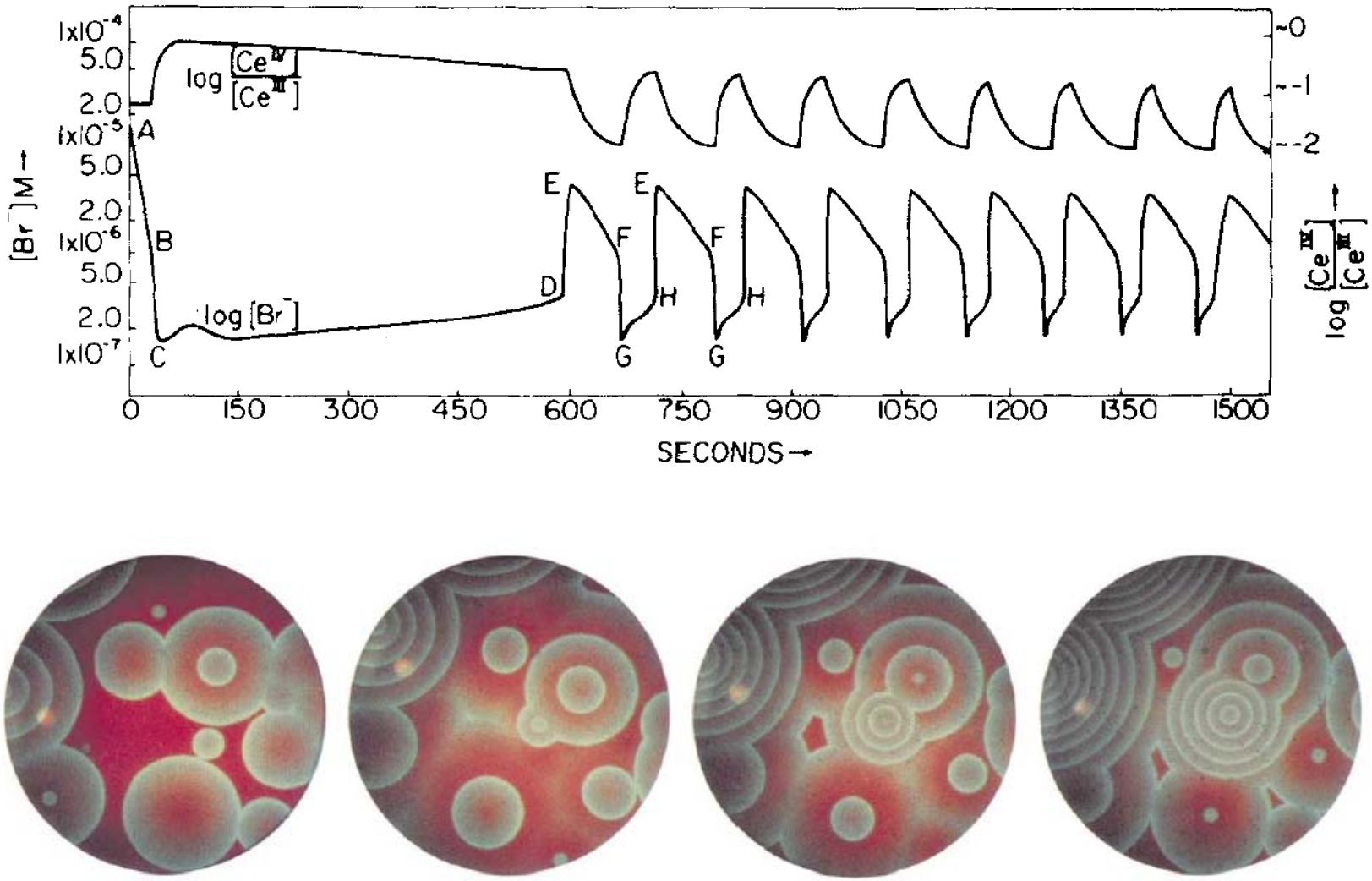
19. $O=C(COO\cdot)(COOH) + Ce(IV) + H_2O \rightarrow HCOOH + Ce(III) + H^+ + 2CO_2$

20. $O=C(COO\cdot)(COOH) + BrCH(COOH) + H_2O \rightarrow HOC(COOH)_2 + O=C(COOH)_2 + Br^- + H^+$

Reaction mechanism of the Belousov-Zhabotinskii reaction

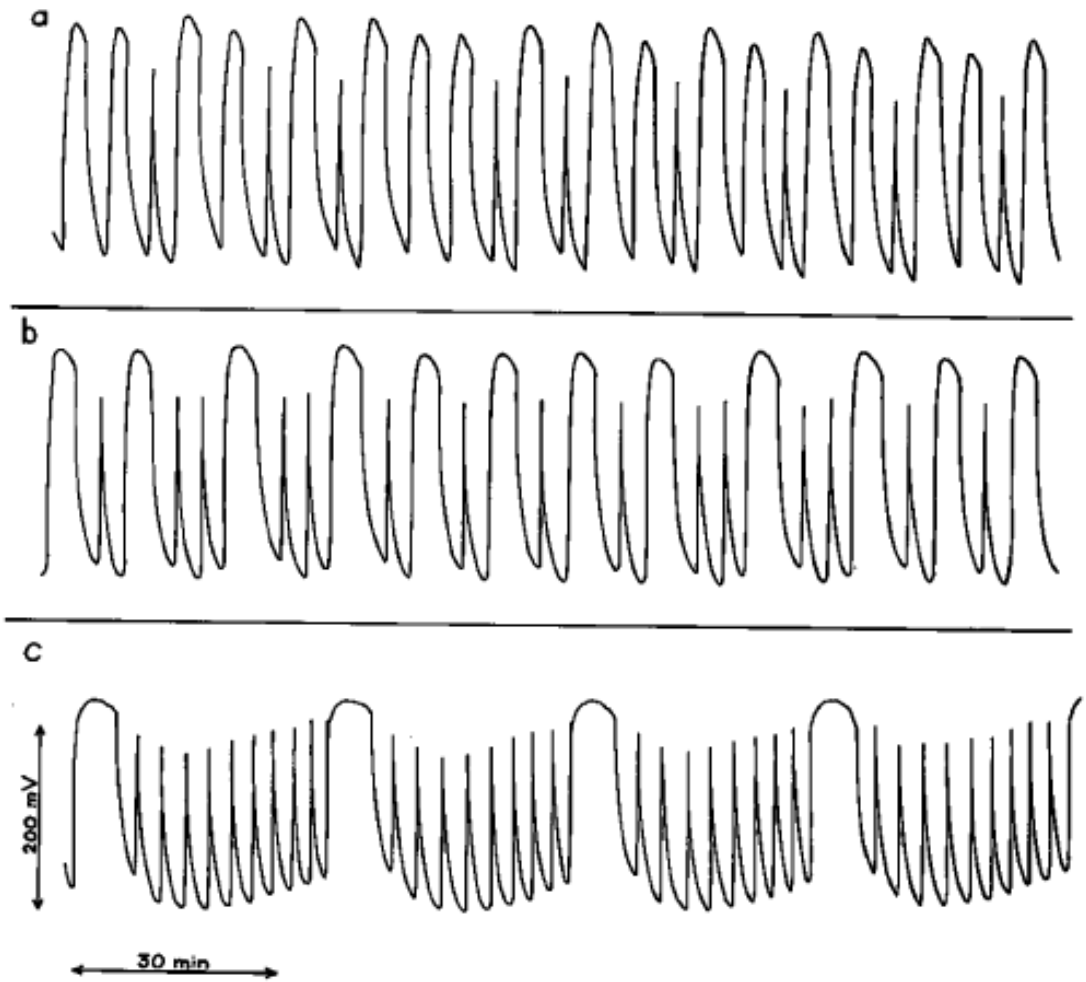
D. Edelson, R.J. Field, R. M. Noyes.

Internat.J.Chem.Kin. **7**:417-432, 1975



Pattern formation in the Belousov-Zhabotinskii reaction

F. Sagués, I.R. Epstein. *Dalton Trans.* **2003**:1201-1217.

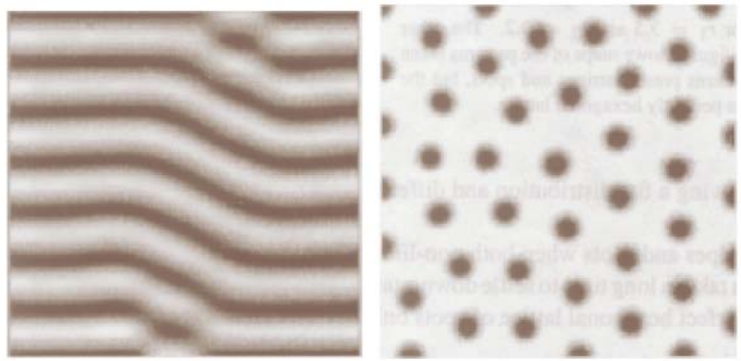


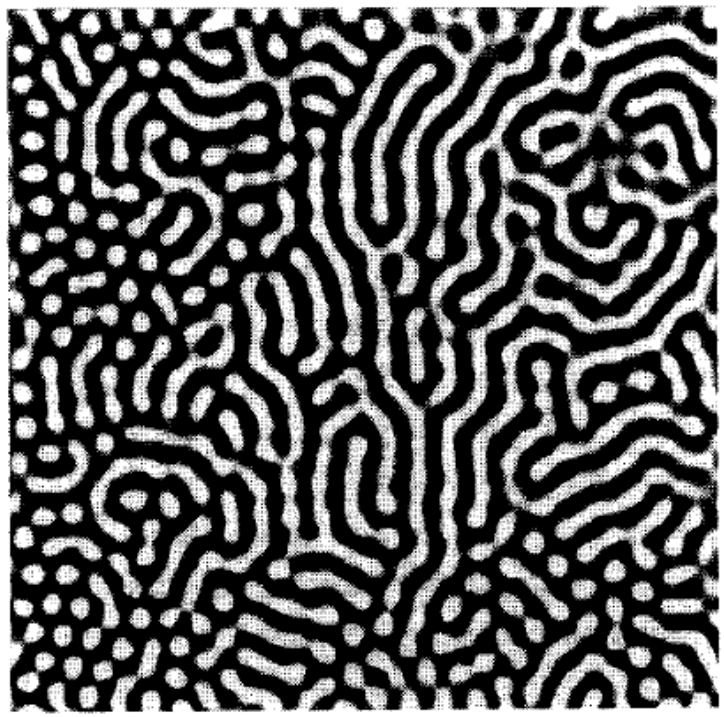
Deterministic chaos in a chemical reaction

F. Sagués, I.R. Epstein. *Dalton Trans.* **2003**:1201-1217.

$$\frac{\partial c_i}{\partial t} = D_i \triangle c_i + F_i(c_1, c_2, \ldots, c_n)\,;\; i = 1, 2, \ldots, n$$

D_i ... diffusion coefficient of substance "i"





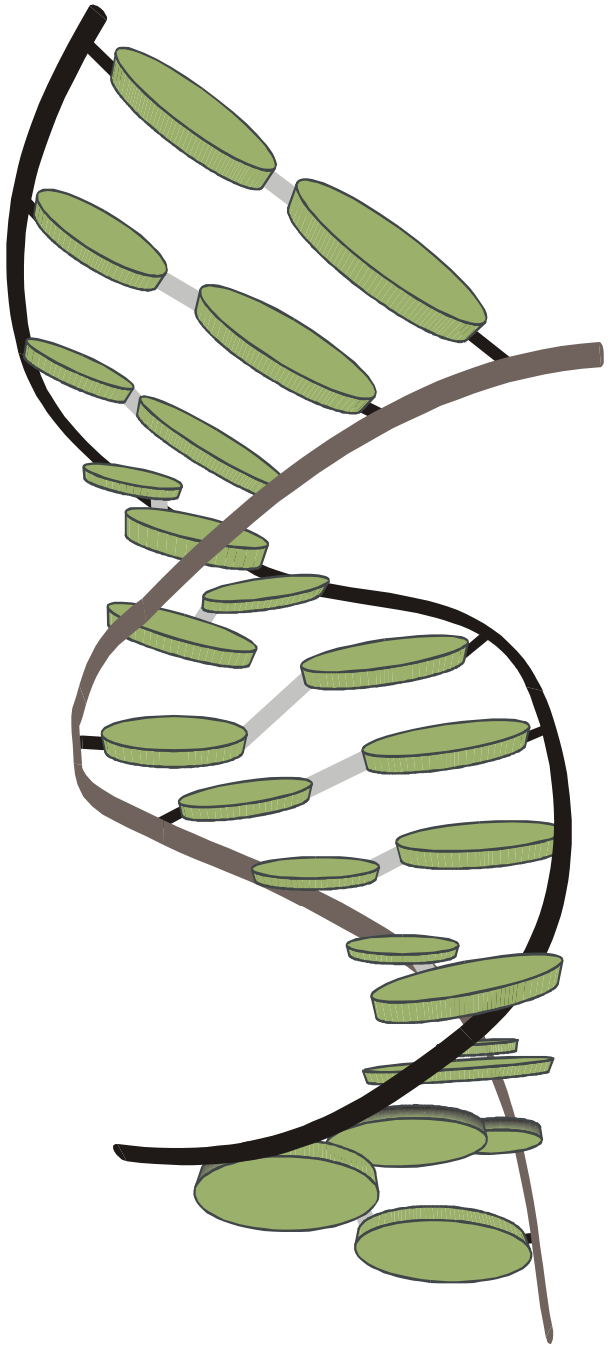
Calculated and experimental Turing patterns

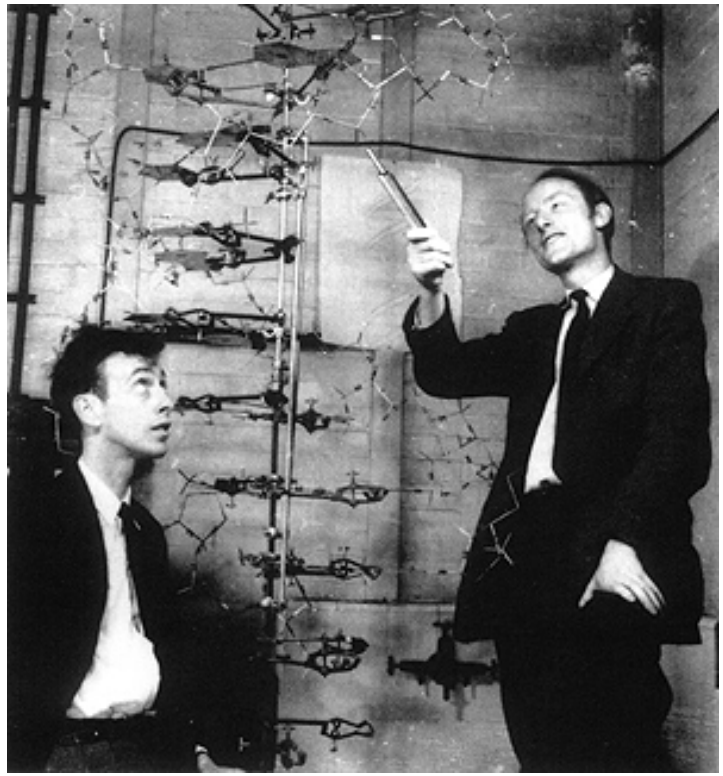
R.A. Barrio, C. Varea, J.L. Aragón, P.K. Maini,
Bull.Math.Biol. **61**:483-505, 1999

R.D. Vigil, Q. Ouyang, H.L. Swinney,
Physica A **188**:15-27, 1992

V. Castets, E. Dulos, J. Boissonade, P. De Kepper,
Phys.Rev.Letters **64**:2953-2956, 1990

1. Autocatalytic chemical reactions
2. **Replication and biological information**
3. Quasispecies and error thresholds
4. Neutral networks in evolution
5. Evolutionary optimization
6. Genetic regulation and metabolism



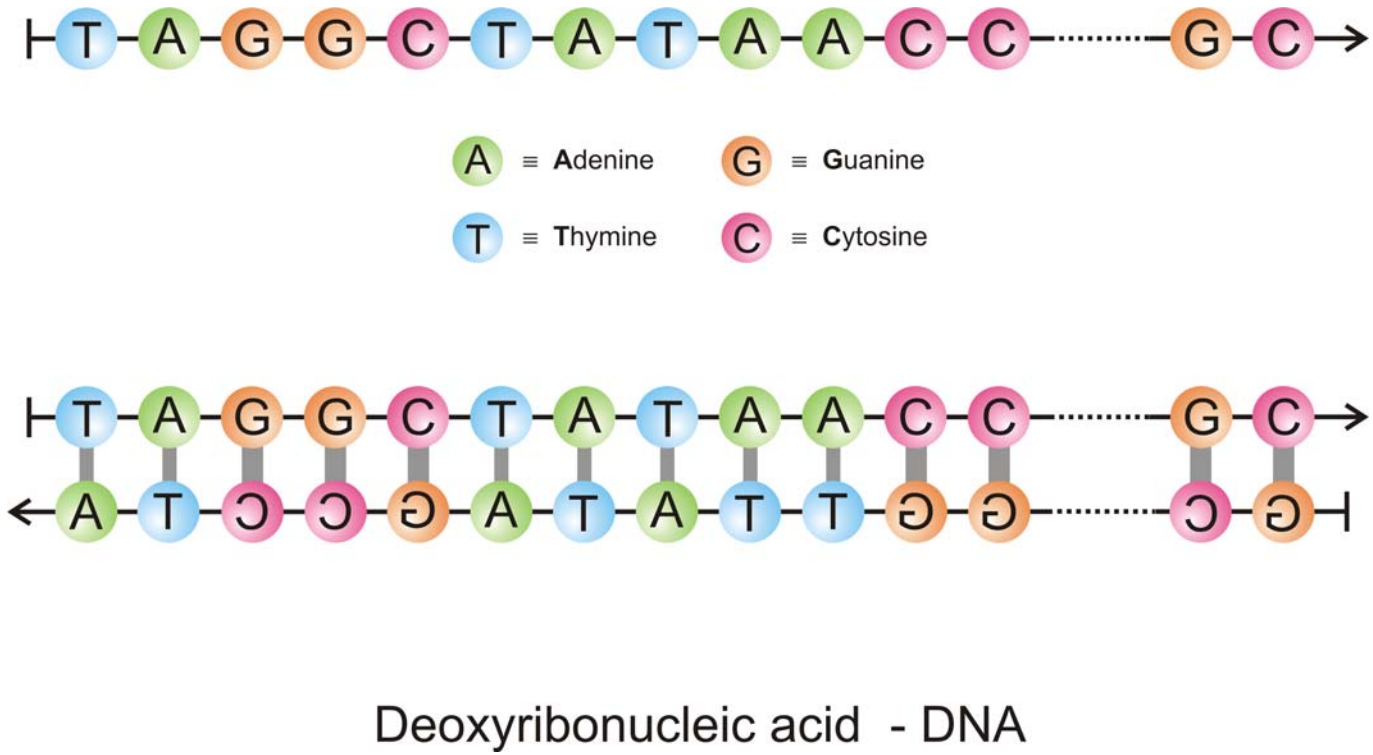


James D. Watson, 1928-, and Francis H.C. Crick, 1916-2004

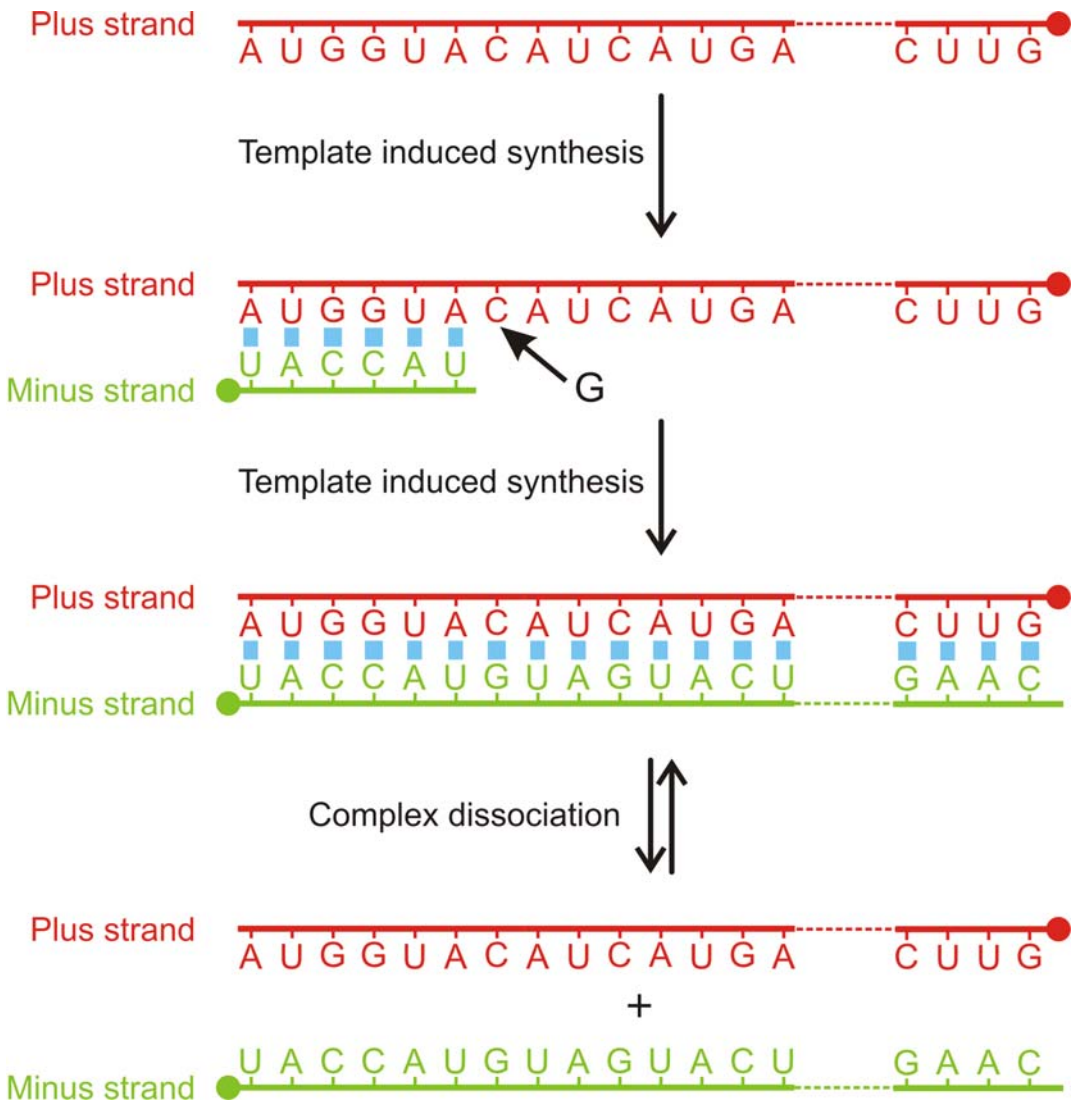
Nobel prize 1962

1953 – 2003 fifty years double helix

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

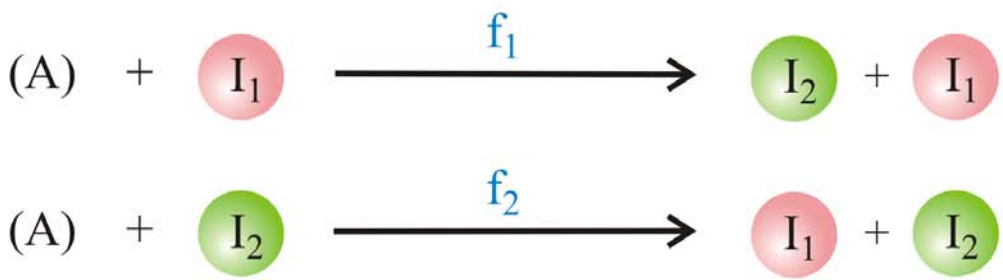


Base complementarity and conservation of genetic information



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

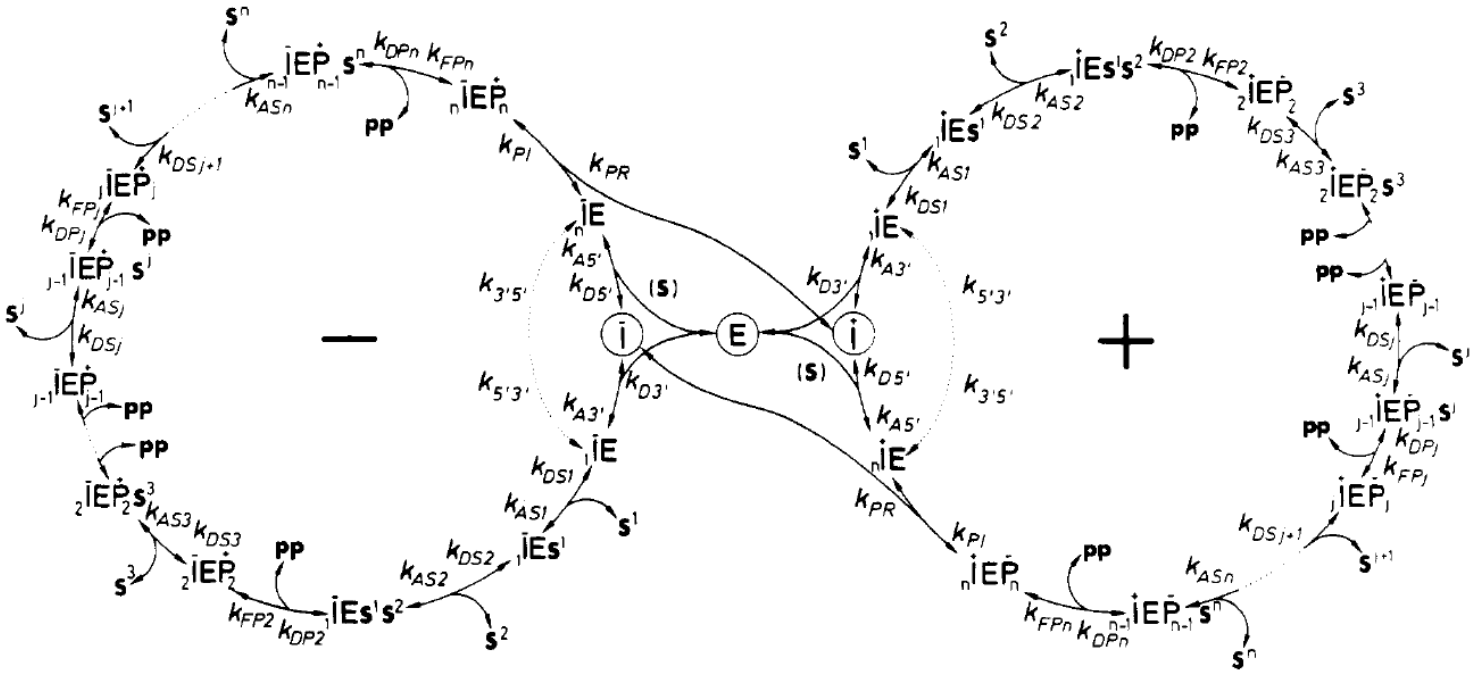
G≡C and **A=U**

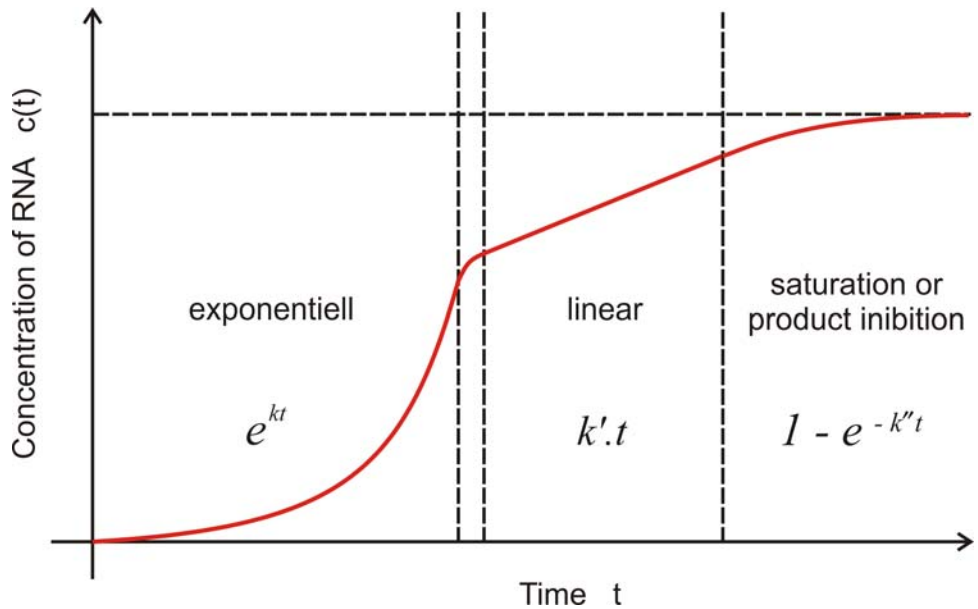


$$dx_1 / dt = f_2 x_2 - x_1 \Phi$$
$$dx_2 / dt = f_1 x_1 - x_2 \Phi$$

$$\Phi = \Sigma_i f_i x_i ; \quad \Sigma_i x_i = 1 ; \quad i = 1,2$$

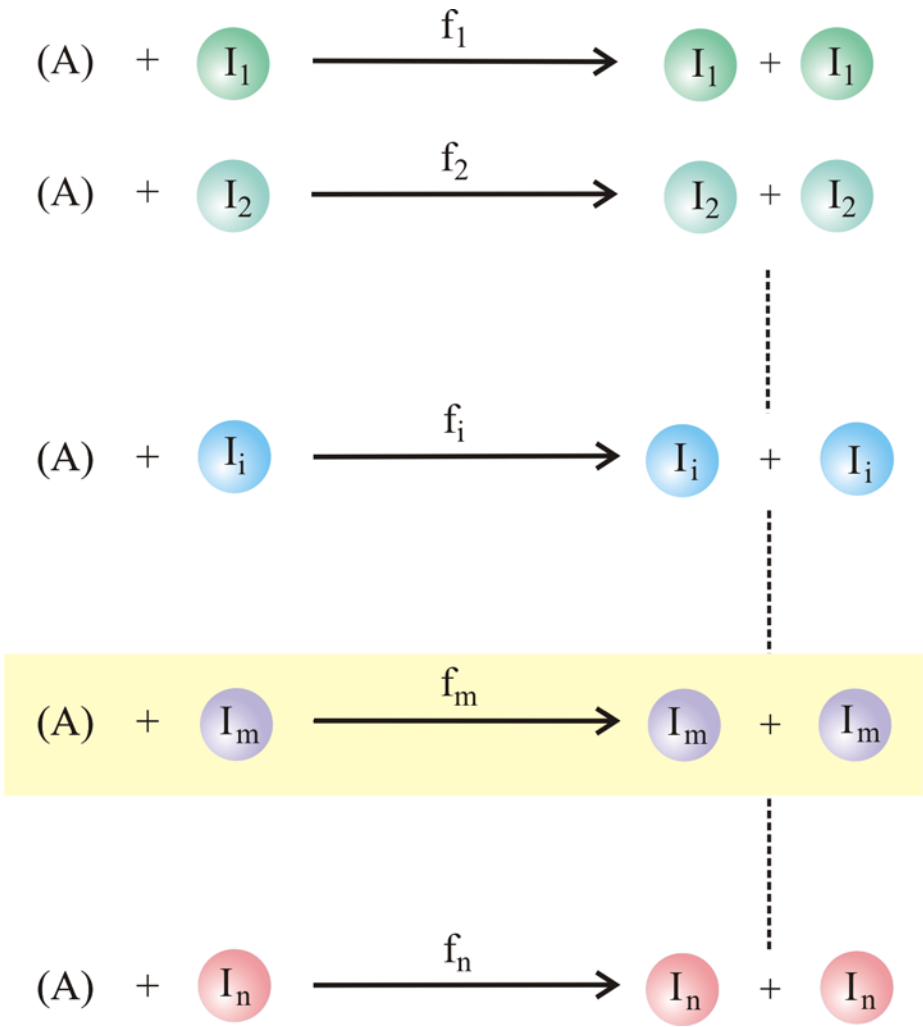
Complementary replication as the simplest molecular mechanism of reproduction





Kinetics of RNA replication

C.K. Biebricher, M. Eigen, W.C. Gardiner, Jr.
Biochemistry **22**:2544-2559, 1983



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

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

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

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

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

$$x_m(t) \rightarrow 1 \quad \text{for} \quad t \rightarrow \infty$$

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

Selection equation: $[\mathrm{I}_i] = x_i \geq 0\ ,\ f_i > 0$

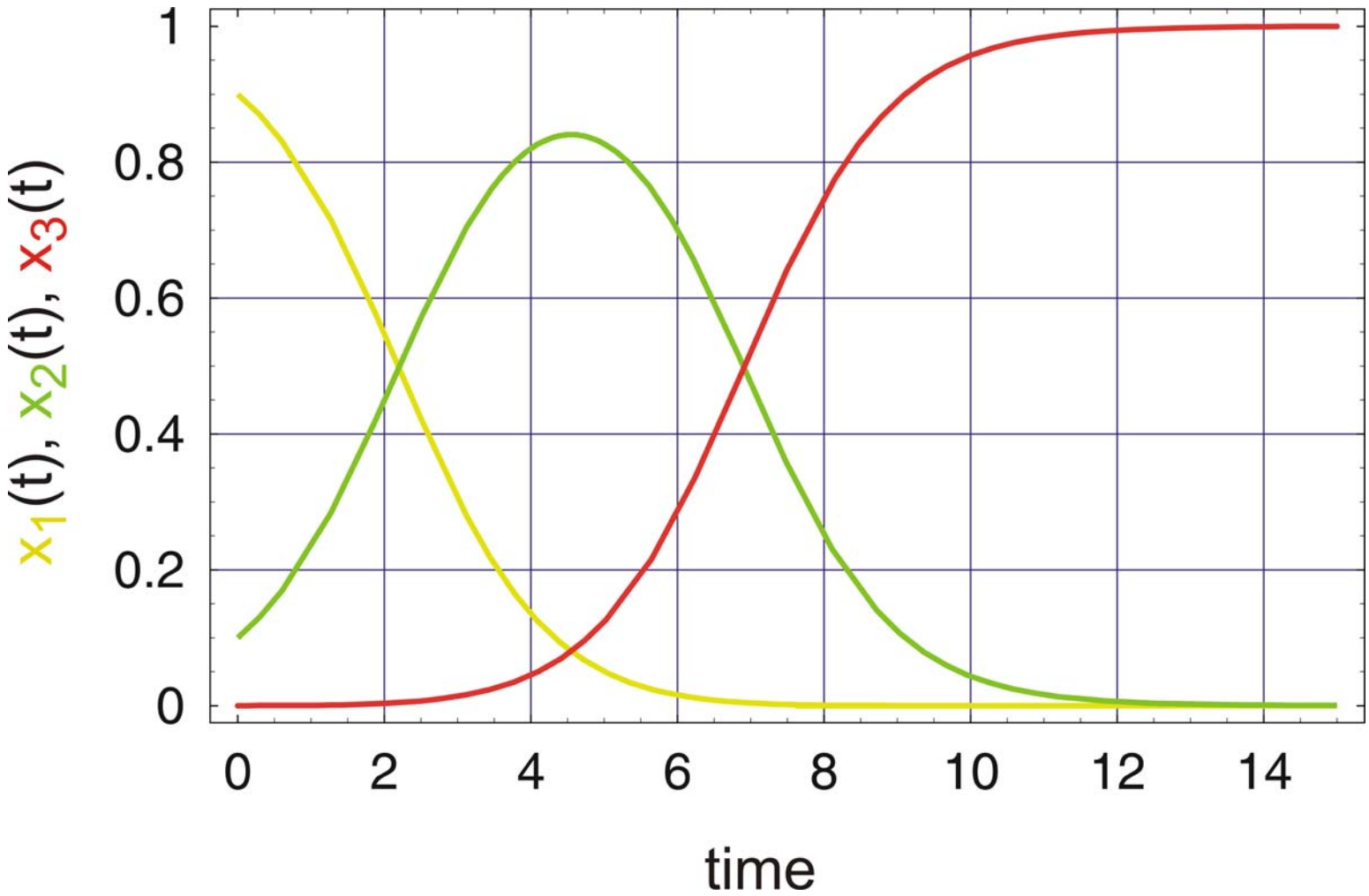
$$\frac{dx_i}{dt} = x_i\left(f_i - \phi\right), \quad i=1,2,\cdots,n; \quad \sum_{i=1}^{n} x_i = 1; \quad \phi = \sum_{j=1}^{n} f_j x_j = \overline{f}$$

Mean fitness or dilution flux, $\phi(\mathrm{t})$, is a non-decreasing function of time,

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

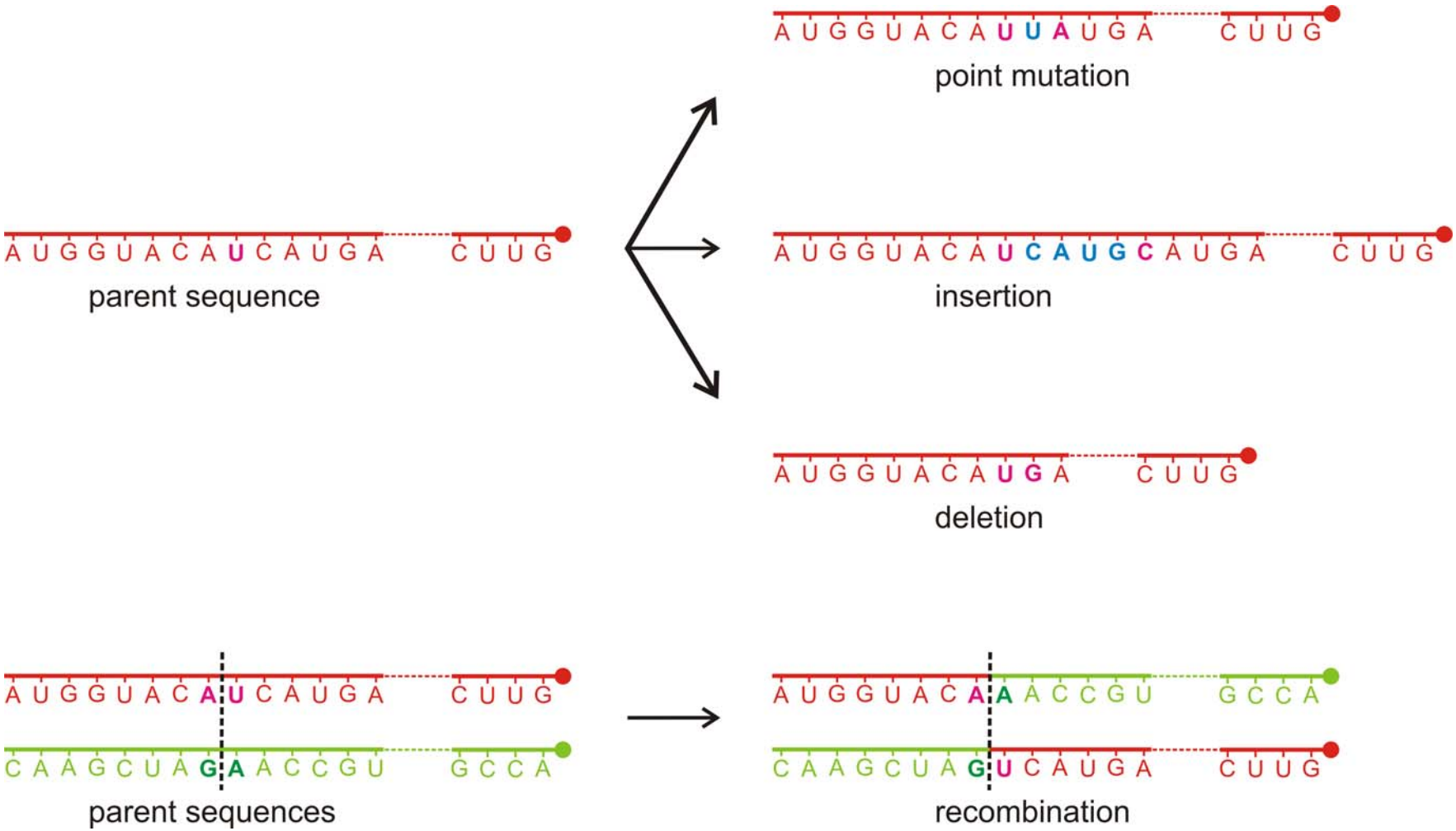
Solutions are obtained by integrating factor transformation

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

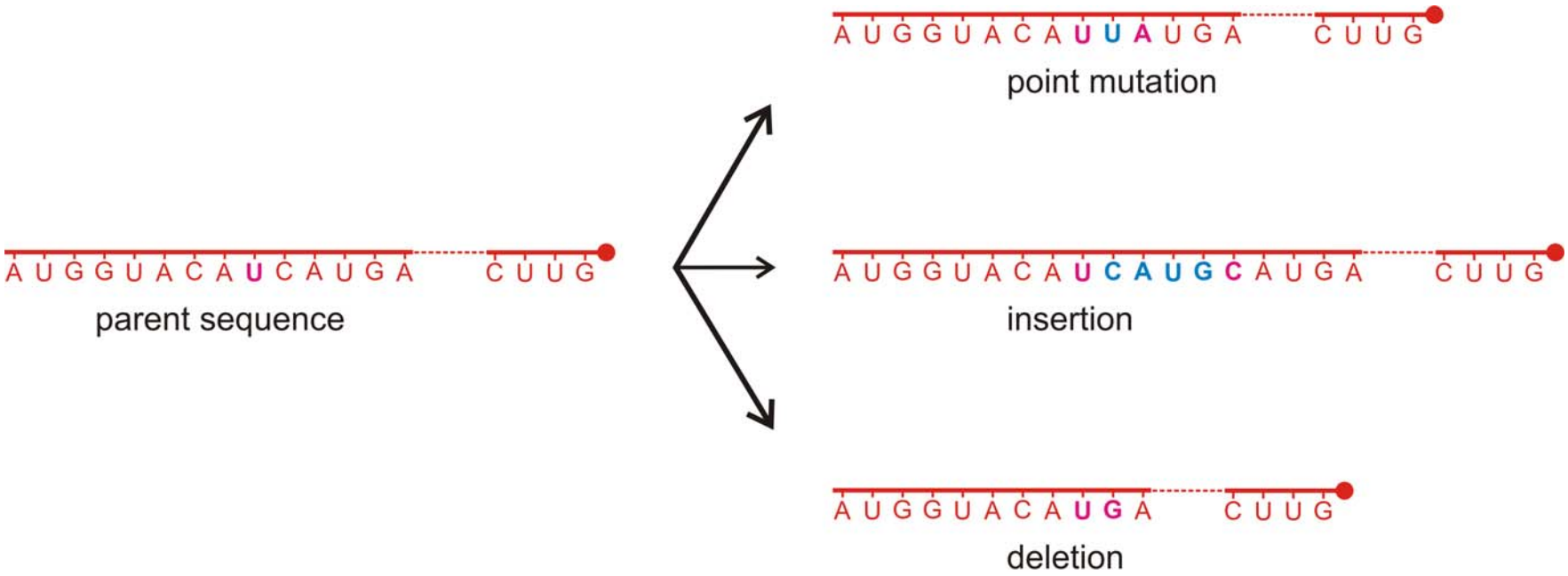


Selection between three species with $f_1 = 1$, $f_2 = 2$, and $f_3 = 3$

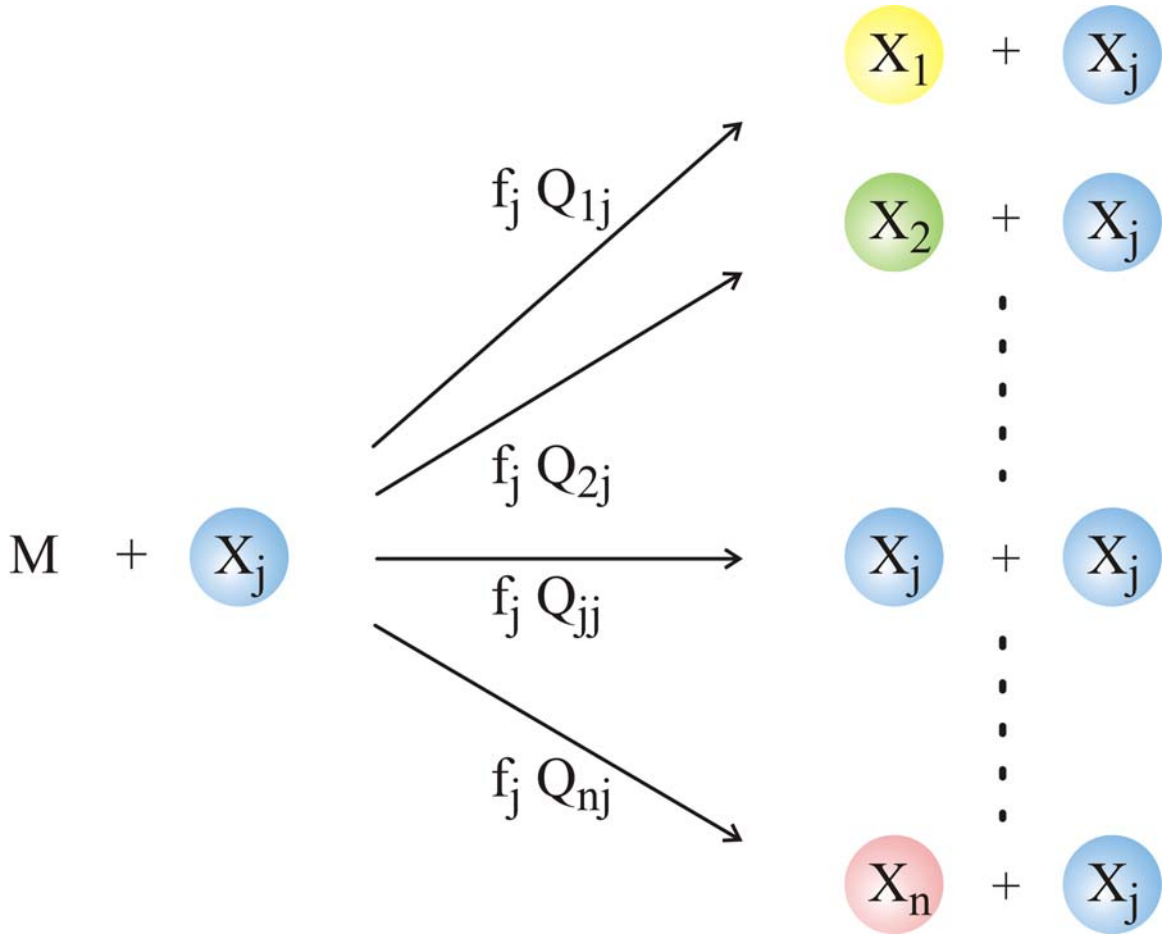
1. Autocatalytic chemical reactions
2. Replication and biological information
3. **Quasispecies and error thresholds**
4. Neutral networks in evolution
5. Evolutionary optimization
6. Genetic regulation and metabolism



Variation of genotypes through mutation and recombination



Variation of genotypes through mutation



Chemical kinetics of replication and mutation as parallel reactions

$$\frac{dx_i}{dt} = \sum_{j=1}^{n} Q_{ij} f_j x_j - x_i \Phi$$

with $\Phi = \sum_{i=1}^{n} f_i x_i$ and $\sum_{i=1}^{n} x_i = 1$

$$\sum_{i=1}^{n} Q_{ij} = 1$$

The replication-mutation equation

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

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

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

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

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

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

Matrix W and Frobenius theorem:

$$\mathrm{W} = \begin{pmatrix} w_{11} & w_{12} & \dots & w_{1n} \\ w_{21} & w_{22} & \dots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \dots & w_{nn} \end{pmatrix}$$

Primitive matrix W:

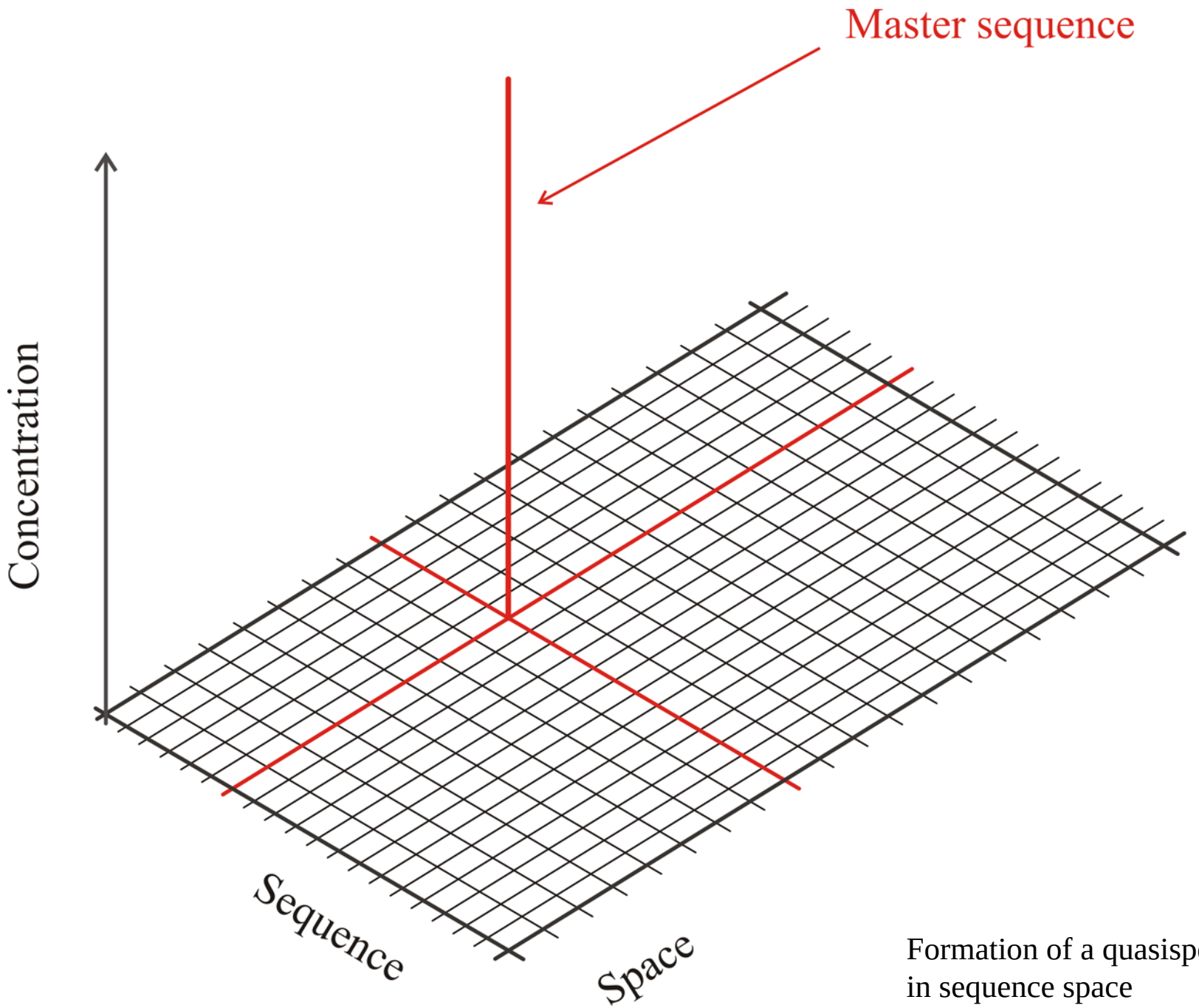
A nonnegative square matrix $\mathrm{W} = \{w_{ij}\}$ is said to be a primitive matrix if there exists k such that $\mathrm{W}^k \gg 0$, i.e., if there exists k such that for all i, j, the (i, j) entry of W^k is positive.

Variation of genotypes through point mutation

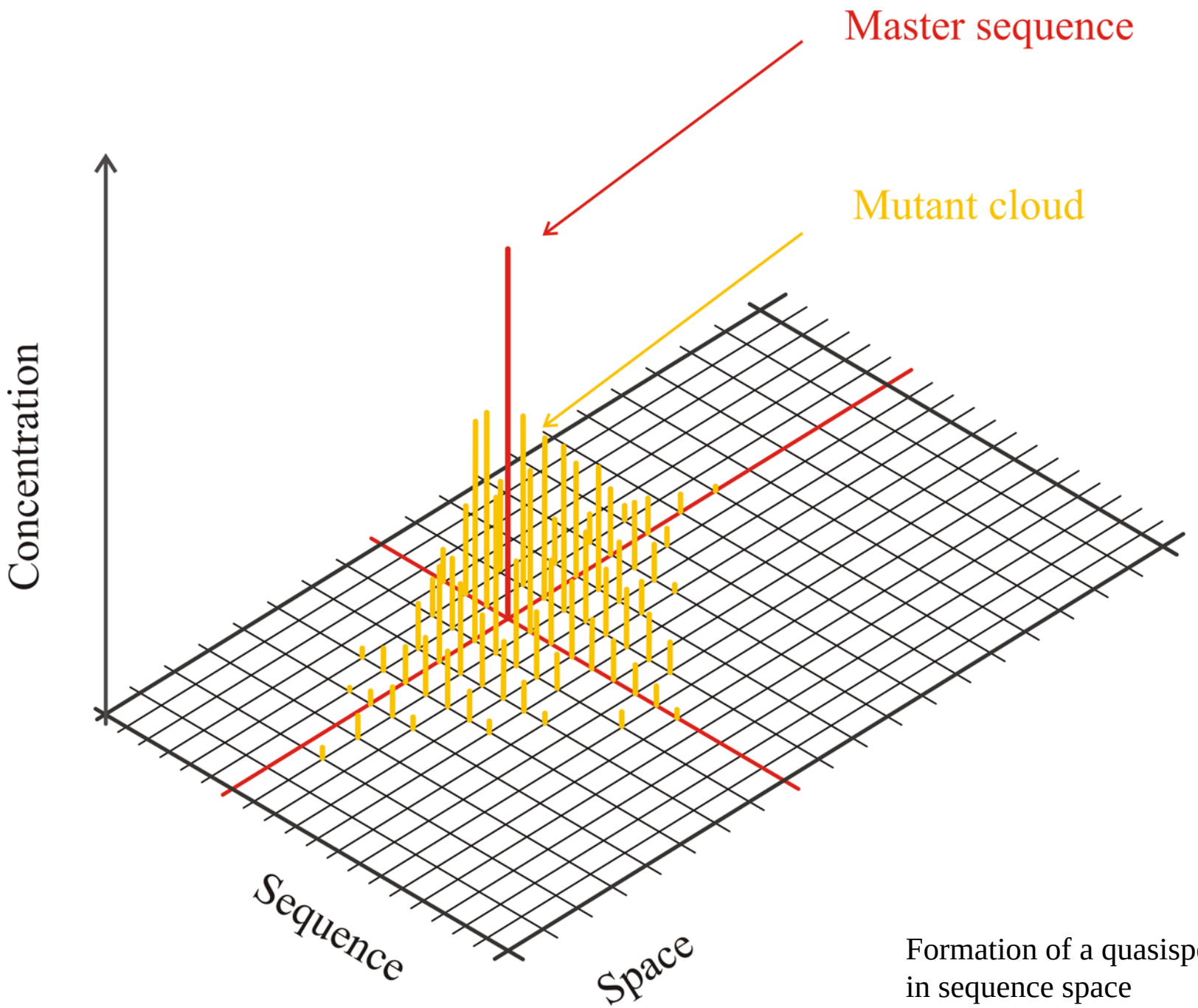
Uniform error rate model:

$$Q_{ij} \;=\; p^{d_H(\mathbf{X}_i,\mathbf{X}_j)}\,(1-p)^{\left(n-d_H(\mathbf{X}_i,\mathbf{X}_j)\right)}$$

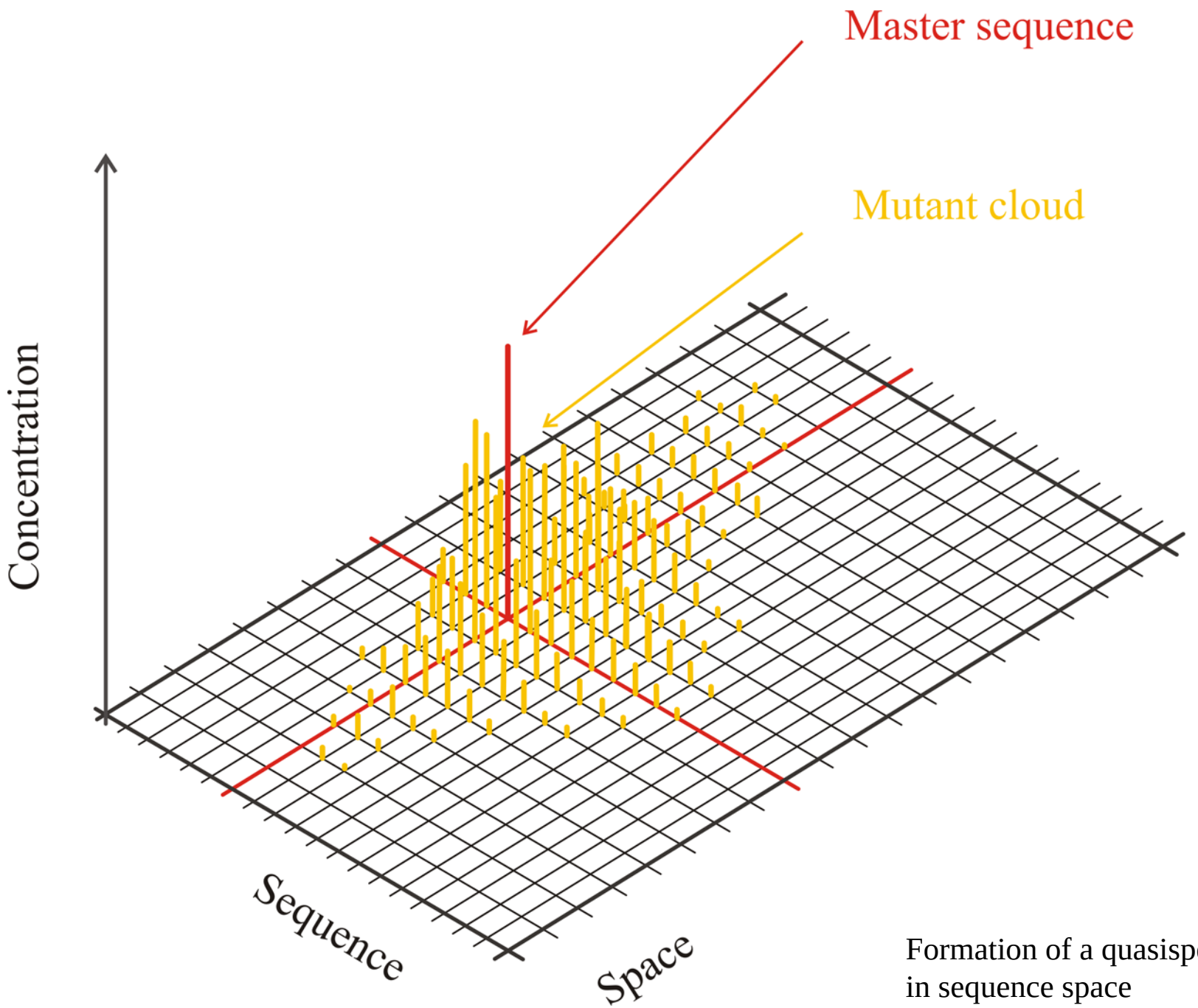
$d_H(\mathbf{X}_i, \mathbf{X}_j)$ ... Hamming distance



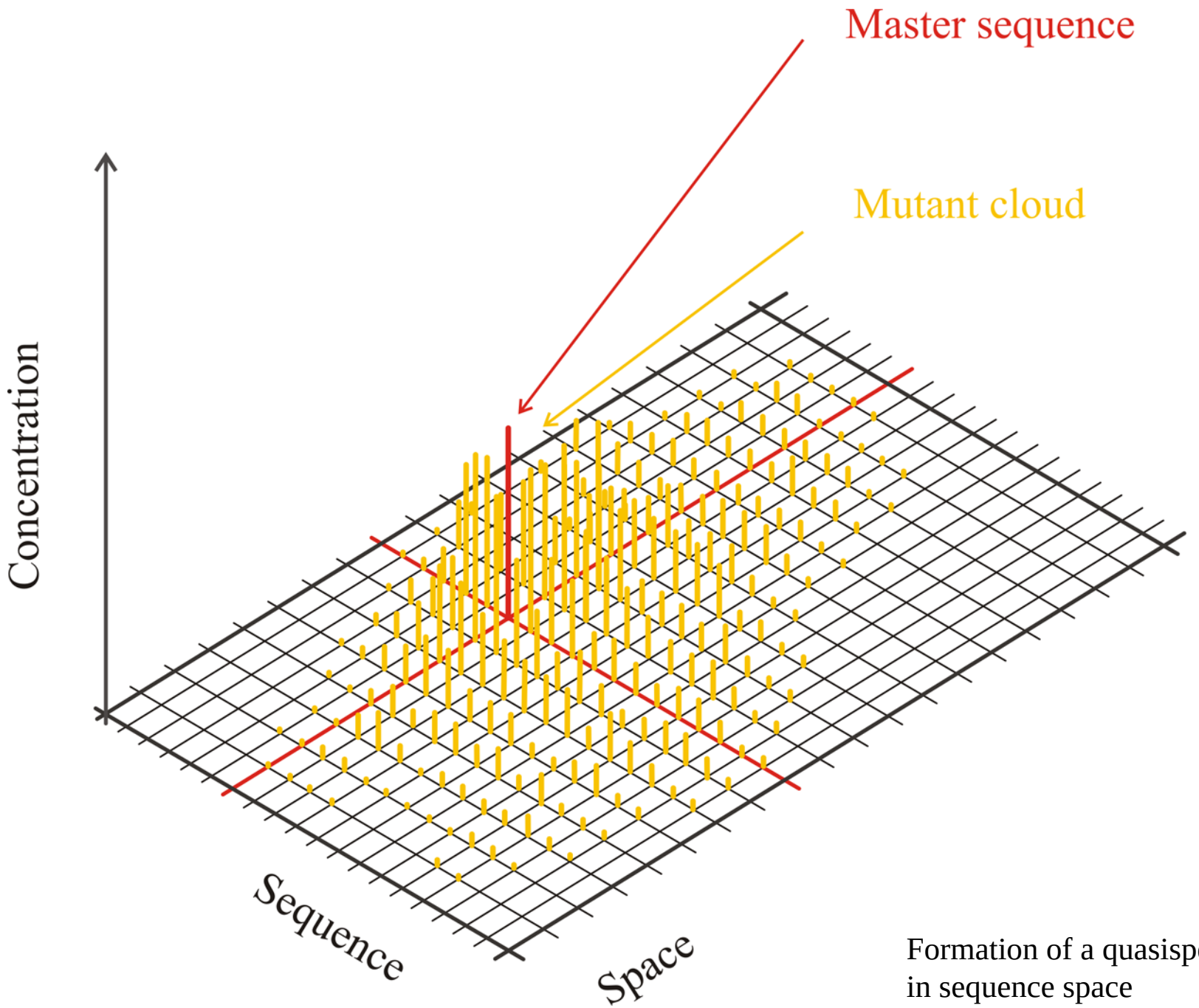
Formation of a quasispecies in sequence space



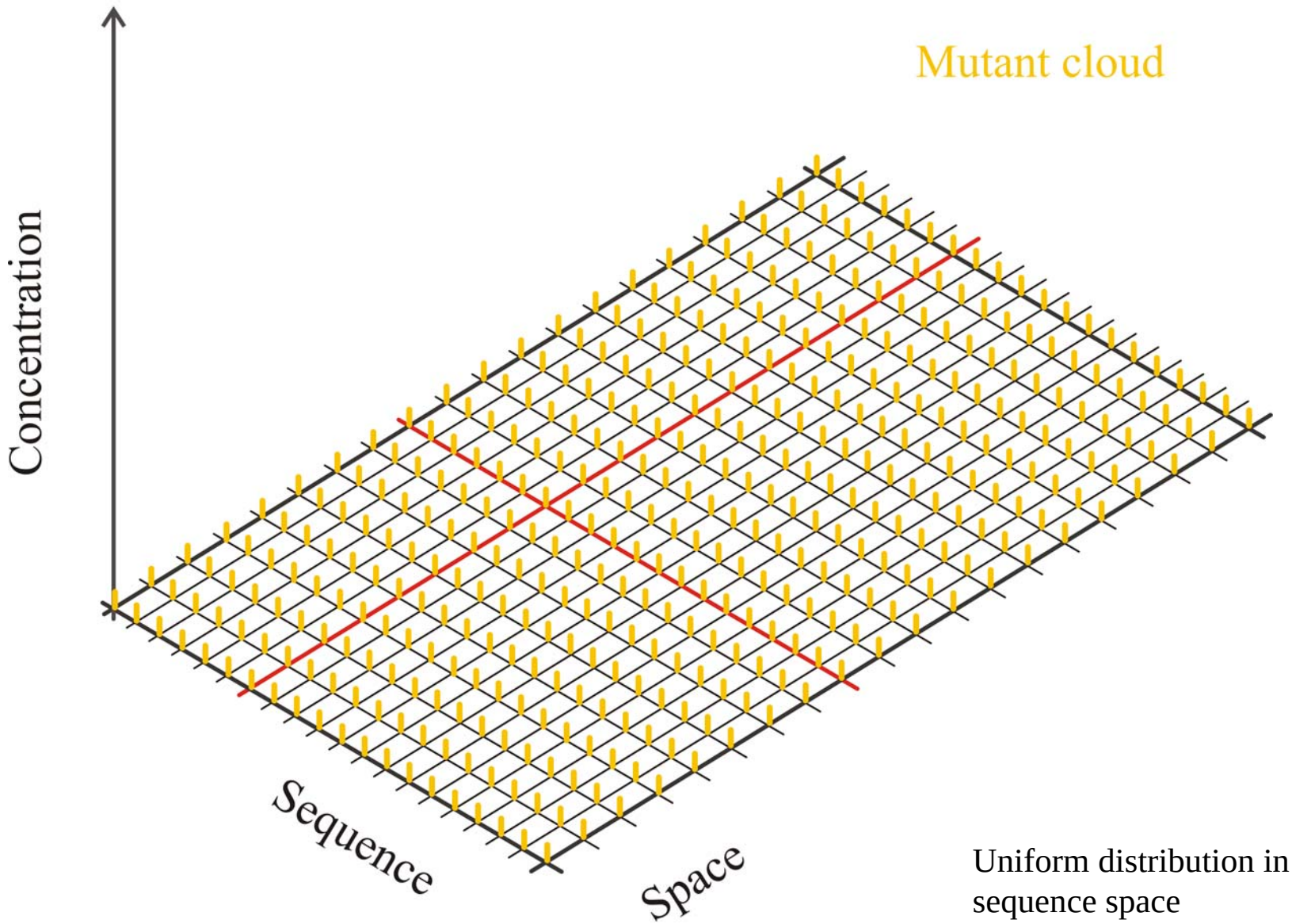
Formation of a quasispecies in sequence space



Formation of a quasispecies in sequence space



Formation of a quasispecies in sequence space



Uniform distribution in sequence space

Biophysical Chemistry 16 (1982) 329–345
Elsevier Biomedical Press

SELF-REPLICATION WITH ERRORS

A MODEL FOR POLYNUCLEOTIDE REPLICATION **

Jörg SWETINA and Peter SCHUSTER *
Institut für Theoretische Chemie und Strahlenchemie der Universität, Währingerstraße 17, A-1090 Wien, Austria

Received 4th June 1982
Revised manuscript received 23rd August 1982
Accepted 30th August 1982

Key words: Polynucleotide replication; Quasi-species; Point mutation; Mutant class; Stochastic replication

A model for polynucleotide replication is presented and analyzed by means of perturbation theory. Two basic assumptions allow handling of sequences up to a chain length of $\nu \approx 80$ explicitly: point mutations are restricted to a two-digit model and individual sequences are subsumed into mutant classes. Perturbation theory is in excellent agreement with the exact results for long enough sequences ($\nu > 20$).

1. Introduction

Eigen [8] proposed a formal kinetic equation (eq. 1) which describes self-replication under the constraint of constant total population size:

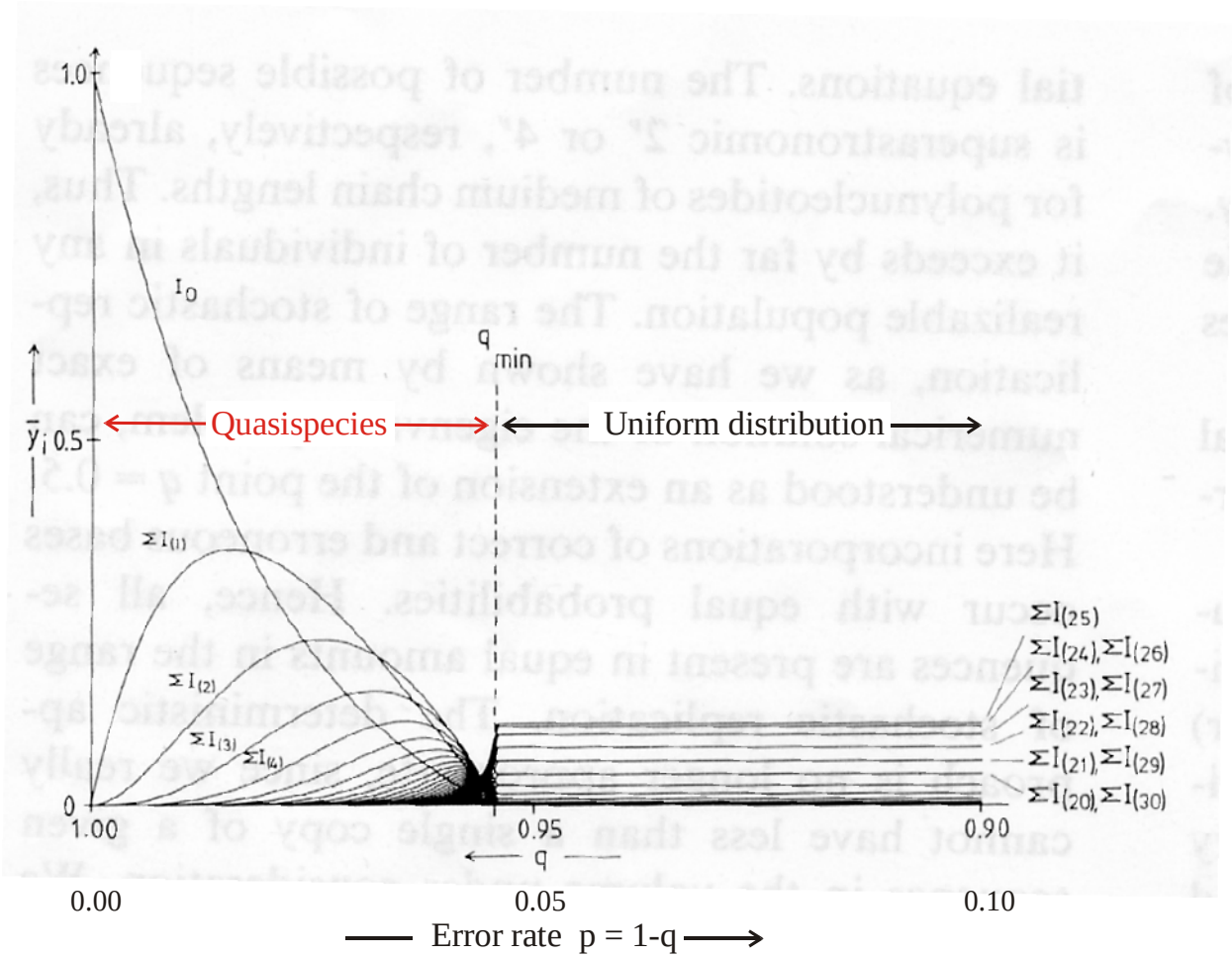
$$\frac{dx_i}{dt} = \dot{x}_i = \sum_j w_{ij} x_j - \frac{x_i}{c}\phi; \; i = 1, \ldots, n \; ^{\dagger} \qquad (1)$$

By x_i we denote the population number or concentration of the self-replicating element I_i, i.e., $x_i = [I_i]$. The total population size or total concentration $c = \sum_i x_i$ is kept constant by proper adjustment of the constraint ϕ: $\phi = \sum_i \sum_j w_{ij} x_j$. Characteristically, this constraint has been called 'constant organization'. The relative values of diagonal (w_{ii}) and off-diagonal (w_{ij}, $i \neq j$) rates, as we shall see in detail in section 2, are related to the accuracy of the replication process. The specific properties of eq. 1 are essentially based on the fact that it leads to exponential growth in the absence of constraints ($\phi = 0$) and competitors ($n = 1$).

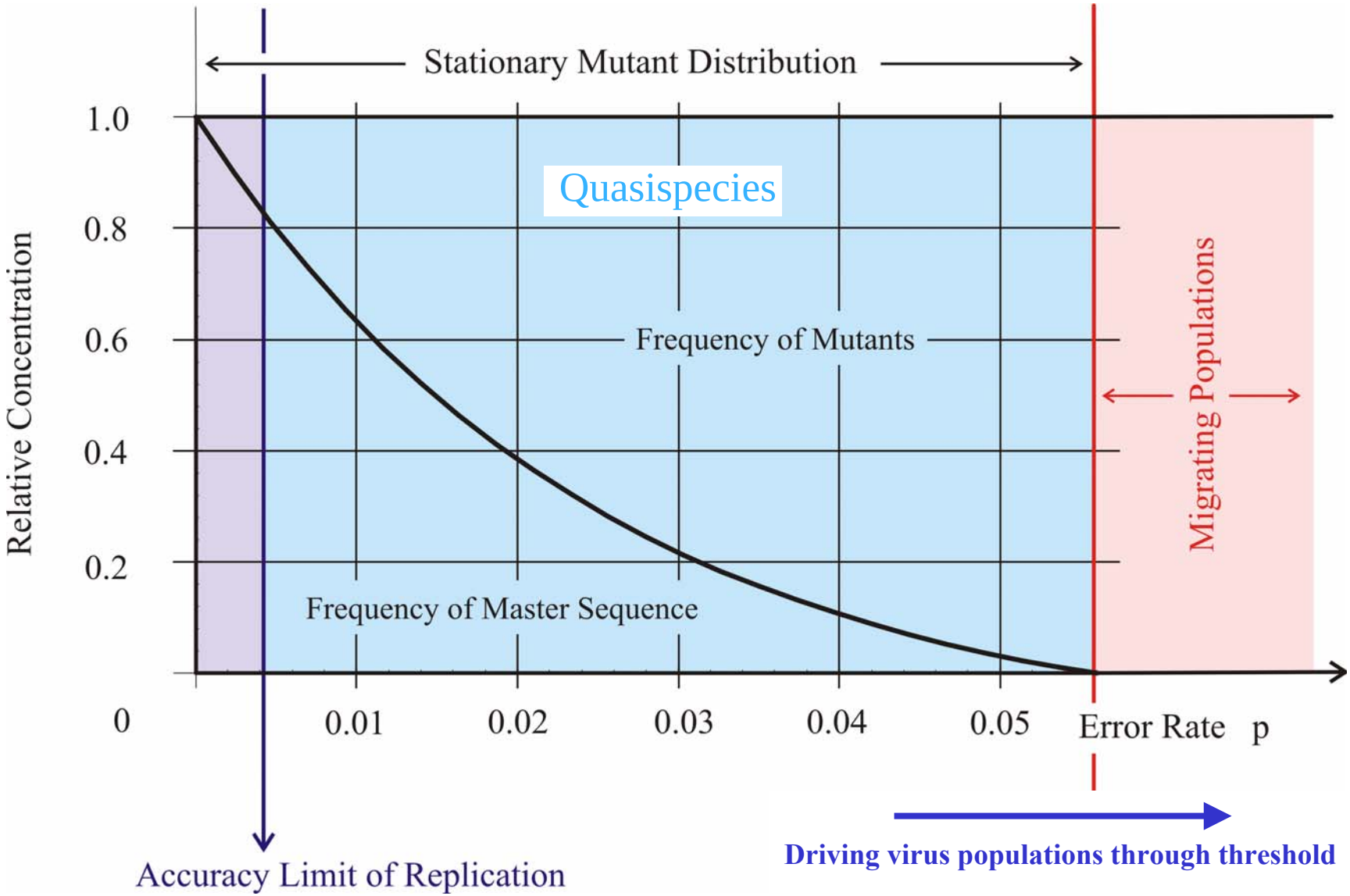
The non-linear differential equation, eq. 1 – the non-linearity is introduced by the definition of ϕ at constant organization – shows a remarkable feature: it leads to selection of a defined ensemble of self-replicating elements above a certain accuracy threshold. This ensemble of a master and its most frequent mutants is a so-called 'quasi-species' [9]. Below this threshold, however, no selection takes place and the frequencies of the individual elements are determined exclusively by their statistical weights.

Rigorous mathematical analysis has been performed on eq. 1 [7,15,24,26]. In particular, it was shown that the non-linearity of eq. 1 can be removed by an appropriate transformation. The eigenvalue problem of the linear differential equation obtained thereby may be solved approximately by the conventional perturbation technique

* Dedicated to the late Professor B.L. Jones who was among the first to do rigorous mathematical analysis on the problems described here.
** This paper is considered as part II of Model Studies on RNA replication. Part I is by Gassner and Schuster [14].
† All summations throughout this paper run from 1 to n unless specified differently: $\sum_i = \sum_{i=1}^{n}$ and $\sum_{i,i\neq j} = \sum_{i=1}^{j-1} + \sum_{i=j+1}^{n}$, respectively.



Quasispecies as a function of the replication accuracy q



The error threshold in replication

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Virus Research 107 (2005) 115–116
Virus Research
www.elsevier.com/locate/virusres

Preface

Antiviral strategy on the horizon

Error catastrophe had its conceptual origins in the middle of the XXth century, when the consequences of mutations on enzymes involved in protein synthesis, as a theory of aging. In those times biological processes were generally perceived differently from today. Infectious diseases were regarded as a fleeting nuisance which would be eliminated through the use of antibiotics and antiviral agents. Microbial variation, although known in some cases, was not thought to be a significant problem for disease control. Variation in differentiated organisms was seen as resulting essentially from exchanges of genetic material associated with sexual reproduction. The problem was to unveil the mechanisms of inheritance, expression of genetic information and metabolism. Few saw that genetic change is occurring at present in all organisms, and still fewer recognized Darwinian principles as essential to the biology of pathogenic viruses and cells. Population geneticists rarely used bacteria or viruses as experimental systems to define concepts in biological evolution. The extent of genetic polymorphism among individuals of the same biological species came as a surprise when the first results on comparison of electrophoretic mobility of enzymes were obtained. With the advent of in vitro DNA recombination, and rapid nucleic acid sequencing techniques, molecular analyses of genomes reinforced the conclusion of extreme inter-individual genetic variation within the same species. Now, due largely to spectacular progress in comparative genomics, we see cellular DNAs, both prokaryotic and eukaryotic, as highly dynamic. Most cellular processes, including such essential information-bearing and transferring events as genome replication, transcription and translation, are increasingly perceived as inherently inaccurate. Viruses, and in particular RNA viruses, are among the most extreme examples of exploitation of replication inaccuracy for survival.

Error catastrophe, or the loss of meaningful genetic information through excess genetic variation, was formulated in quantitative terms as a consequence of quasispecies theory, which was first developed to explain self-organization and adaptability of primitive replicons in early stages of life. Recently, a conceptual extension of error catastrophe that could be defined as "induced genetic deterioration" has emerged as a possible antiviral strategy. This is the topic of the current special issue of ***Virus Research***.

Few would nowadays doubt that one of the major obstacles for the control of viral disease is short-term adaptability of viral pathogens. Adaptability of viruses follows the same Darwinian principles that have shaped biological evolution over eons, that is, repeated rounds of reproduction with genetic variation, competition and selection, often perturbed by random events such as statistical fluctuations in population size. However, with viruses the consequences of the operation of these very same Darwinian principles are felt within very short times. Short-term evolution (within hours and days) can be also observed with some cellular pathogens, with subsets of normal cells, and cancer cells. The nature of RNA viral pathogens begs for alternative antiviral strategies, and forcing the virus to cross the critical error threshold for maintenance of genetic information is one of them.

The contributions to this volume have been chosen to reflect different lines of evidence (both theoretical and experimental) on which antiviral designs based on genetic deterioration inflicted upon viruses are being constructed. Theoretical studies have explored the copying fidelity conditions that must be fulfilled by any information-bearing replication system for the essential genetic information to be transmitted to progeny. Closely related to the theoretical developments have been numerous experimental studies on quasispecies dynamics and their multiple biological manifestations. The latter can be summarized by saying that RNA viruses, by virtue of existing as mutant spectra rather than defined genetic entities, remarkably expand their potential to overcome selective pressures intended to limit their replication. Indeed, the use of antiviral inhibitors in clinical practice and the design of vaccines for a number of major RNA virus-associated diseases, are currently presided by a sense of uncertainty. Another line of growing research is the enzymology of copying fidelity by viral replicases, aimed at understanding the molecular basis of mutagenic activities. Error catastrophe as a potential new antiviral strategy received an important impulse by the observation that ribavirin (a licensed antiviral nucleoside analogue) may be exerting, in some systems, its antiviral activity through enhanced mutagenesis. This has encouraged investigations on new mutagenic base analogues, some of them used in anticancer chemotherapy. Some chapters summarize these important biochemical studies on cell entry pathways and metabolism of mutagenic agents, that may find new applications as antiviral agents.

This volume intends to be basically a progress report, an introduction to a new avenue of research, and a realistic appraisal of the many issues that remain to be investigated. In this respect, I can envisage (not without many uncertainties) at least three lines of needed research: (i) One on further understanding of quasispecies dynamics in infected individuals to learn more on how to apply combinations of virus-specific mutagens and inhibitors in an effective way, finding synergistic combinations and avoiding antagonistic ones as well as severe clinical side effects. (ii) Another on a deeper understanding of the metabolism of mutagenic agents, in particular base and nucleoside analogues. This includes identification of the transporters that carry them into cells, an understanding of their metabolic processing, intracellular stability and alterations of nucleotide pools, among other issues. (iii) Still another line of needed research is the development of new mutagenic agents specific for viruses, showing no (or limited) toxicity for cells. Some advances may come from links with anticancer research, but others should result from the designs of new molecules, based on the structures of viral polymerases. I really hope that the reader finds this issue not only to be an interesting and useful review of the current situation in the field, but also a stimulating exposure to the major problems to be faced.

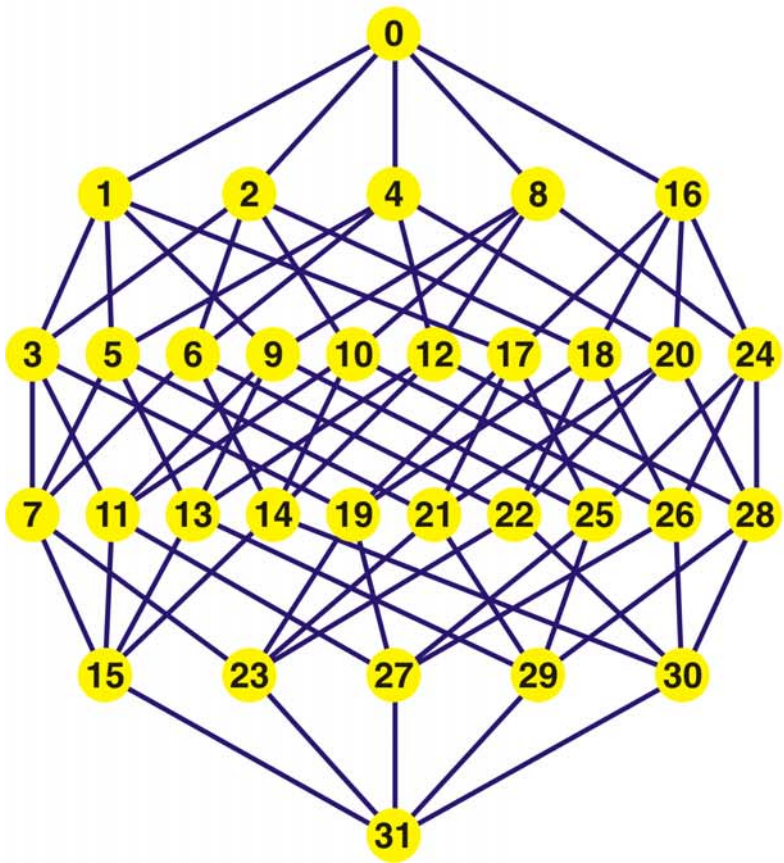
The idea to prepare this special issue came as a kind invitation of Ulrich Desselberger, former Editor of ***Virus Research***, and then taken enthusiastically by Luis Enjuanes, recently appointed as Editor of ***Virus Research***. I take this opportunity to thank Ulrich, Luis and the Editor-in-Chief of ***Virus Research***, Brian Mahy, for their continued interest and support to the research on virus evolution over the years.

My thanks go also to the 19 authors who despite their busy schedules have taken time to prepare excellent manuscripts, to Elsevier staff for their prompt responses to my requests, and, last but not least, to Ms. Lucía Horrillo from Centro de Biología Molecular "Severo Ochoa" for her patient dealing with the correspondence with authors and the final organization of the issue.

Esteban Domingo
Universidad Autónoma de Madrid
Centro de Biología Molecular "Severo Ochoa"
Consejo Superior de Investigaciones Científicas
Cantoblanco and Valdeolmos
Madrid, Spain
Tel.: +34 91 497 84858/9; fax: +34 91 497 4799
E-mail address: edomingo@cbm.uam.es

Available online 8 December 2004

0168-1702/$ – see front matter
doi:10.1016/j.virusres.2004.11.001



Mutant class

0

1

2

3

4

5

Binary sequences can be encoded by their decimal equivalents:

C = 0 and G = 1, for example,

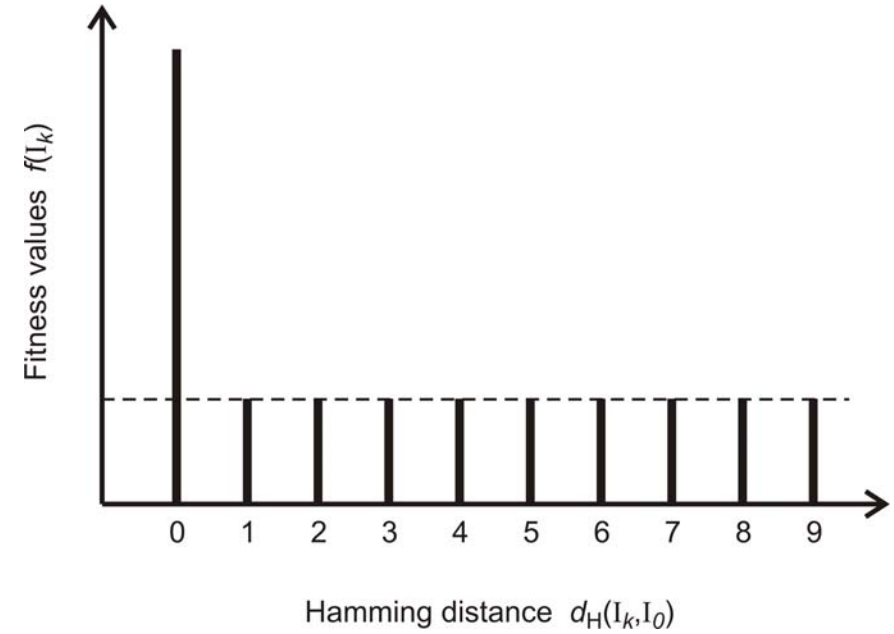
"0" ≡ 00000 = CCCCC,

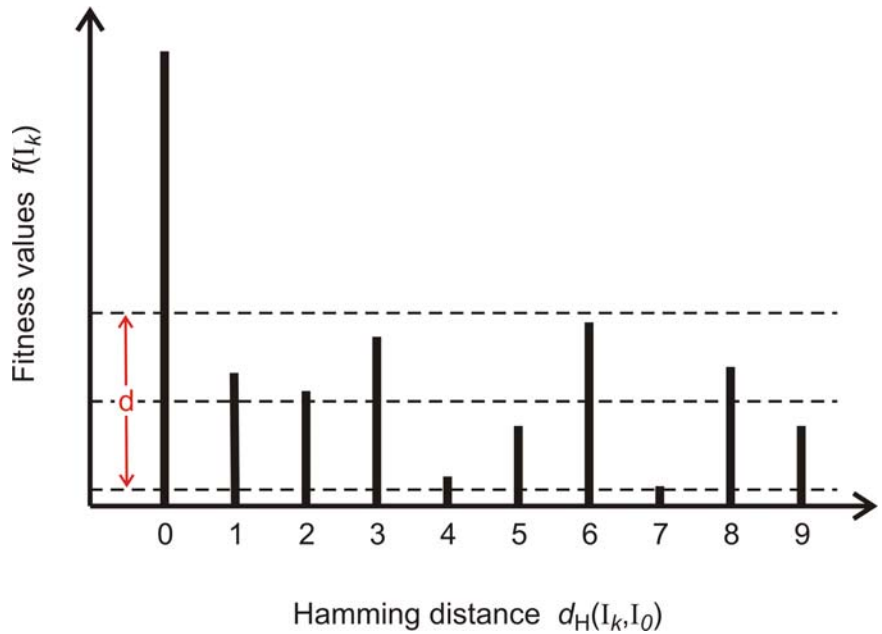
"14" ≡ 01110 = CGGGC,

"29" ≡ 11101 = GGGCG, etc.

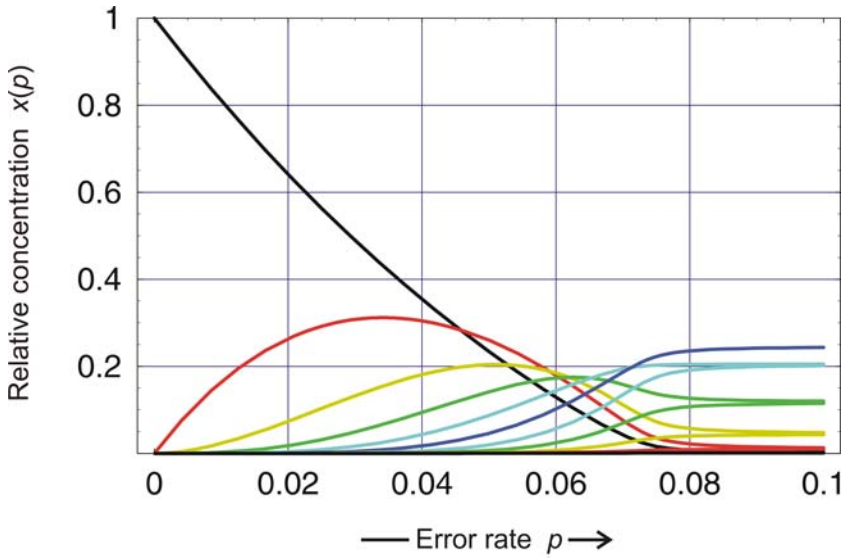
Every point in sequence space is equivalent

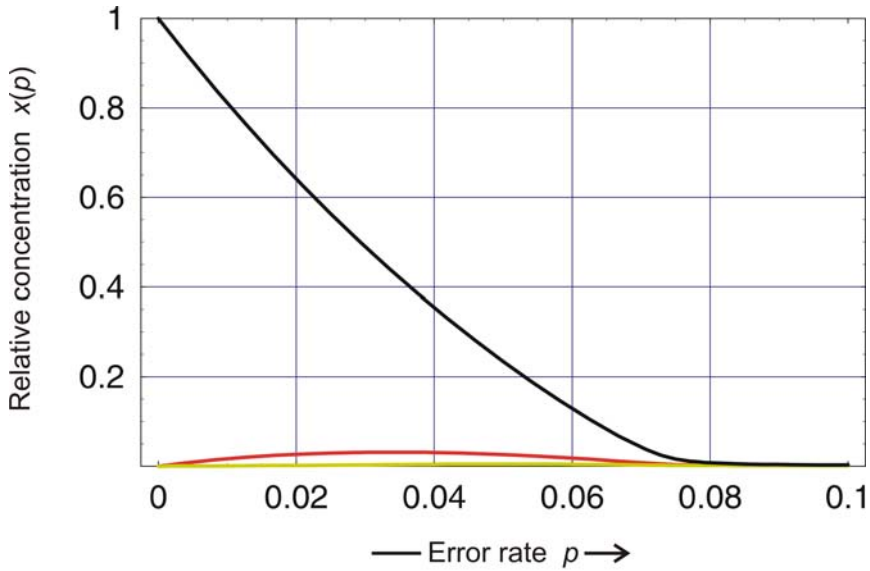
Sequence space of binary sequences with chain length $n = 5$

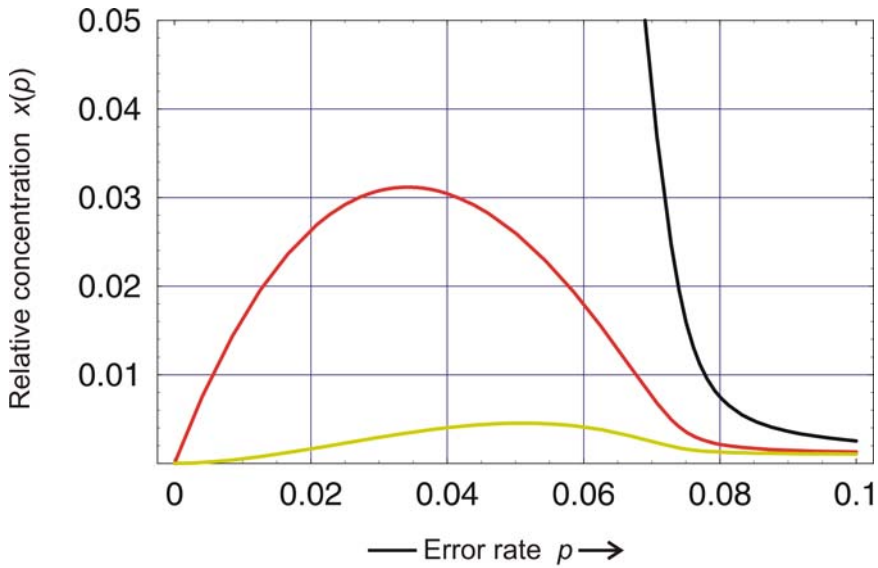




Fitness landscapes showing error thresholds

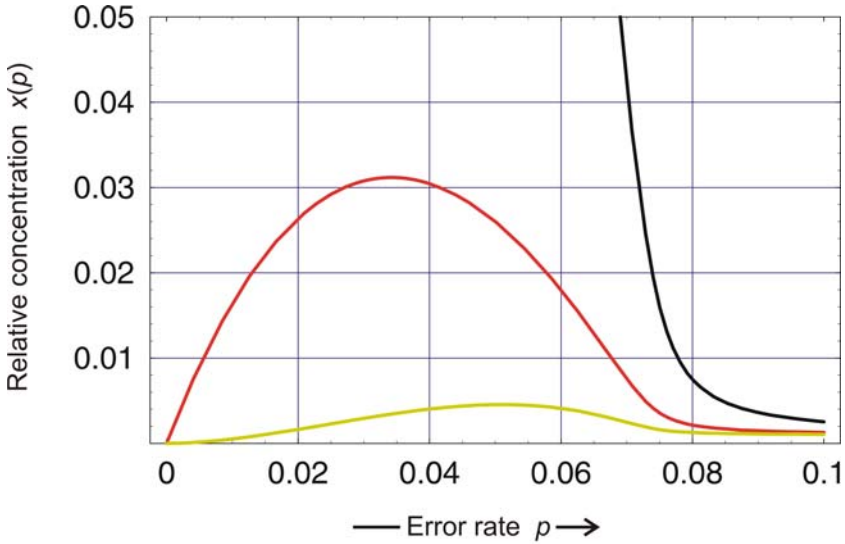


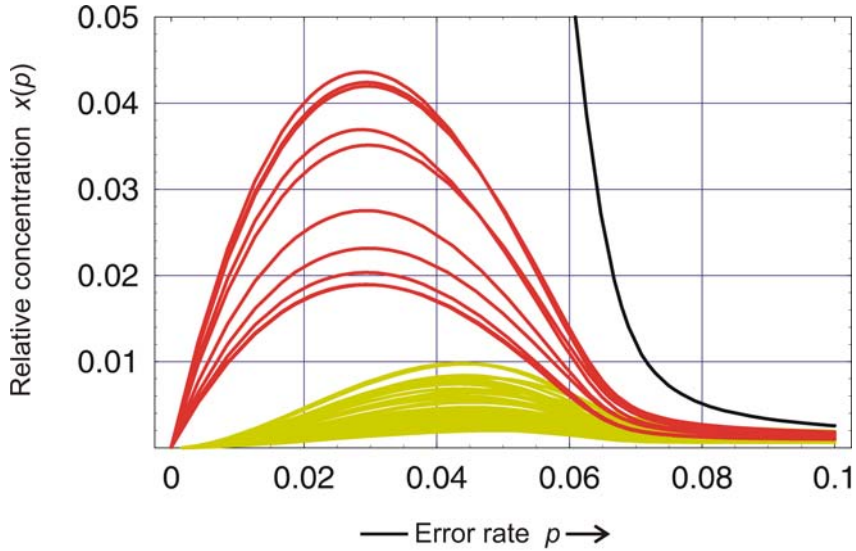


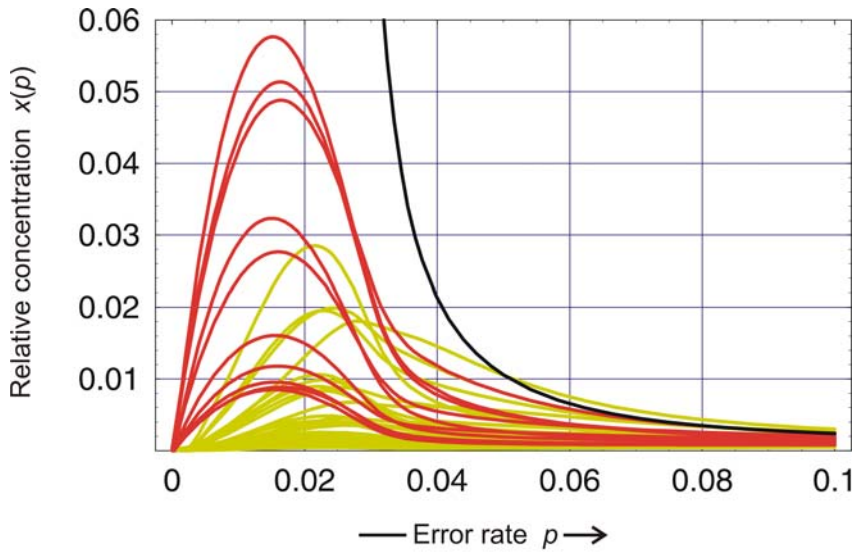


Error threshold: Error classes and individual sequences

n = 10 and $\sigma = 2$



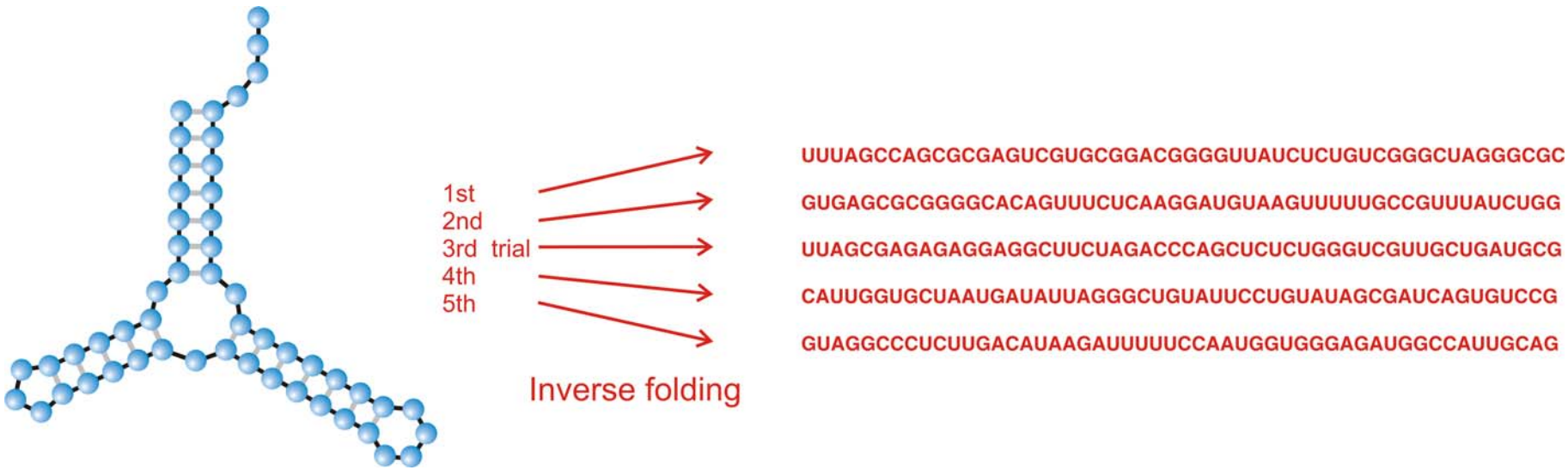




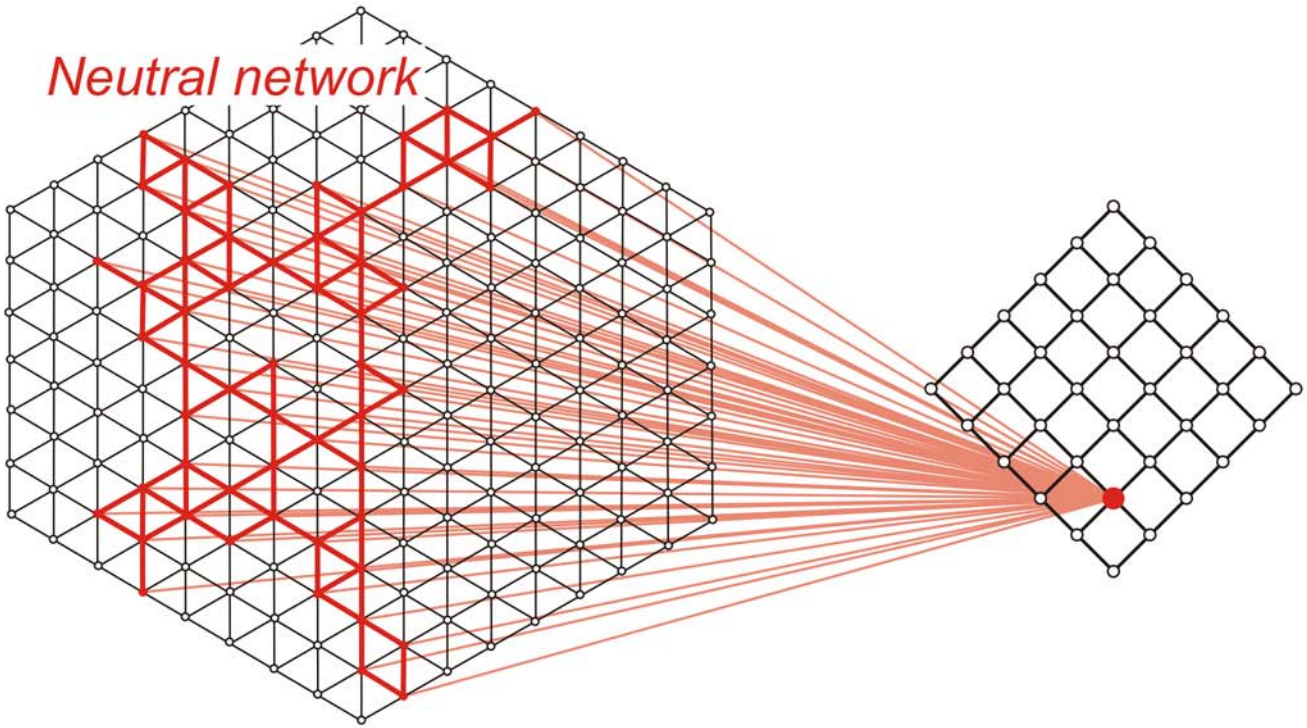
Error threshold: Individual sequences

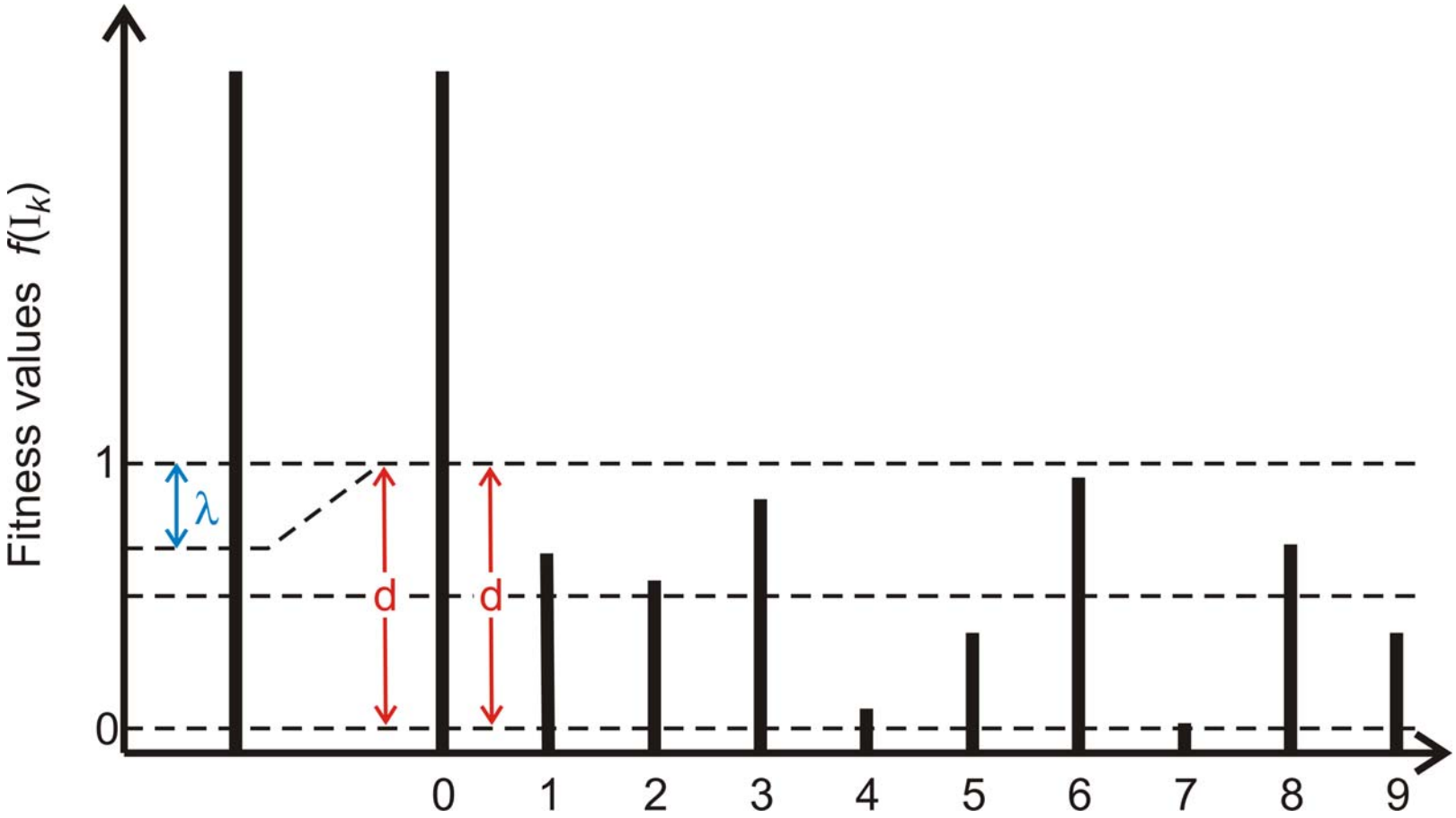
n = 10, σ = 2 and d = 0, 1.0, 1.85

1. Autocatalytic chemical reactions
2. Replication and biological information
3. Quasispecies and error thresholds
4. **Neutral networks in evolution**
5. Evolutionary optimization
6. Genetic regulation and metabolism



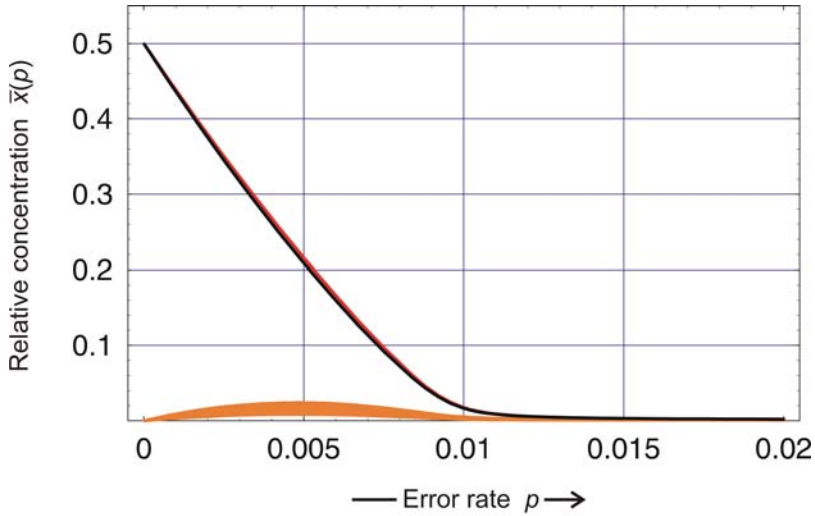
The inverse folding algorithm searches for sequences that form a given RNA structure.

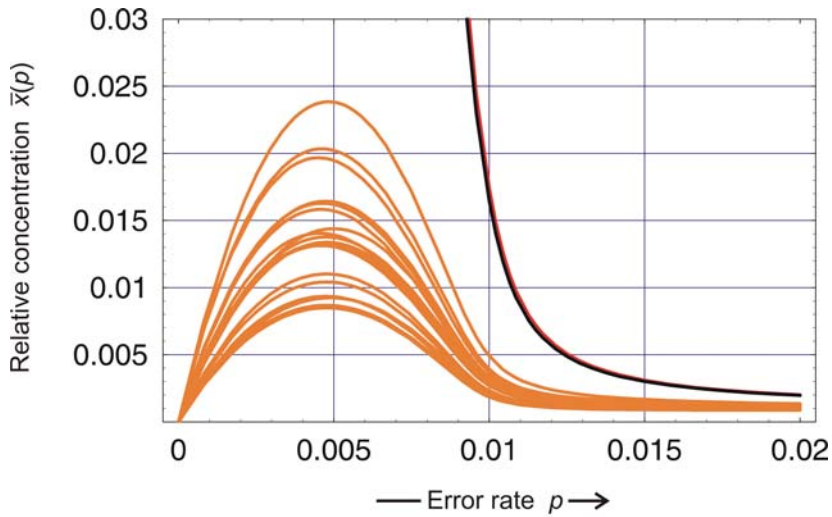

Neutral network
Sequence space
Structure space


Fitness values $f(I_k)$
1
0
λ
d
d
0
1
2
3
4
5
6
7
8
9

Neutral network

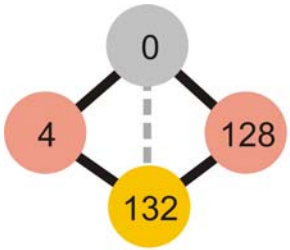
$\lambda = 0.01,\ s = 367$





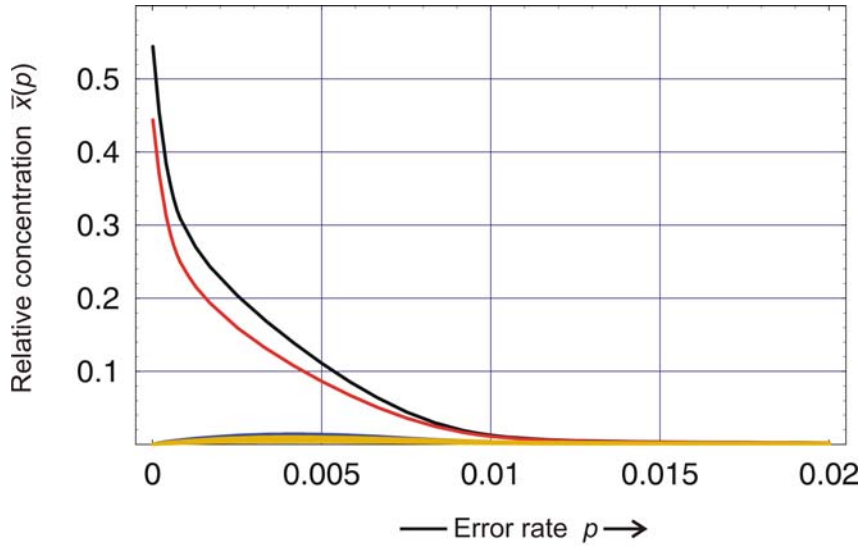
Error threshold: Individual sequences

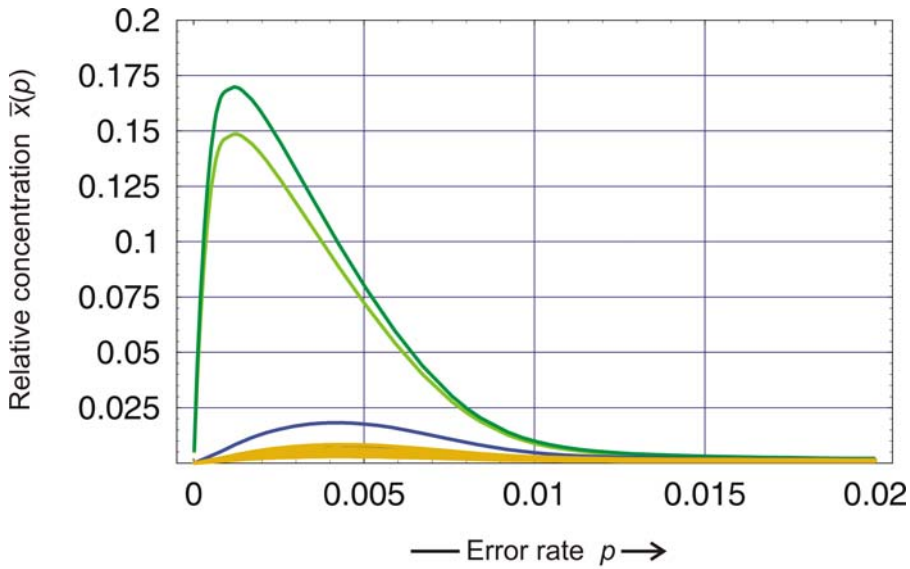
$n = 10,\ \sigma = 1.1,\ d = 1.0$



Neutral networks

λ = 0.01, s = 877





Error threshold: Individual sequences

n = 10, σ = 1.1, d = 1.0

Bulletin of Mathematical Biology Vol. 50, No. 6, pp. 635–660, 1988.
Printed in Great Britain.

0092-8240/88$3.00 + 0.00
Pergamon Press plc
Society for Mathematical Biology

STATIONARY MUTANT DISTRIBUTIONS AND EVOLUTIONARY OPTIMIZATION

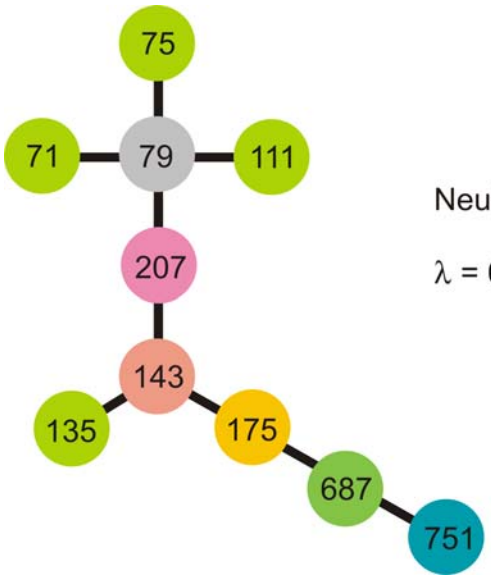
■ PETER SCHUSTER and JÖRG SWETINA
Institut für theoretische Chemie
und Strahlenchemie der Universität Wien,
Währingerstraße 17,
A 1090 Wien,
Austria

Molecular evolution is modelled by erroneous replication of binary sequences. We show how the selection of two species of equal or almost equal selective value is influenced by its nearest neighbours in sequence space. In the case of perfect neutrality and sufficiently small error rates we find that the Hamming distance between the species determines selection. As the error rate increases the fitness parameters of neighbouring species become more and more important. In the case of almost neutral sequences we observe a critical replication accuracy at which a drastic change in the "quasispecies", in the stationary mutant distribution occurs. Thus, in frequently mutating populations fitness turns out to be an ensemble property rather than an attribute of the individual.

In addition we investigate the time dependence of the mean excess production as a function of initial conditions. Although it is optimized under most conditions, cases can be found which are characterized by decrease or non-monotonous change in mean excess productions.

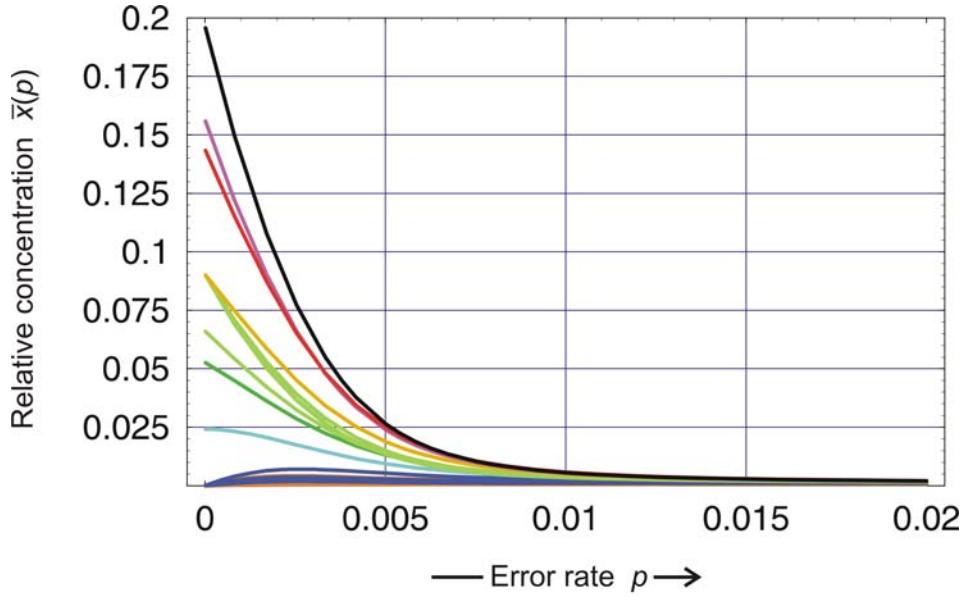
1. Introduction. Recent data from populations of RNA viruses provided direct evidence for vast sequence heterogeneity (Domingo *et al.*, 1987). The origin of this diversity is not yet completely known. It may be caused by the low replication accuracy of the polymerizing enzyme, commonly a virus specific, RNA dependent RNA synthetase, or it may be the result of a high degree of selective neutrality of polynucleotide sequences. Eventually, both factors contribute to the heterogeneity observed. Indeed, mutations occur much more frequently than previously assumed in microbiology. They are by no means rare events and hence, neither the methods of conventional population genetics (Ewens, 1979) nor the neutral theory (Kimura, 1983) can be applied to these virus populations. Selectively neutral variants may be close with respect to Hamming distance and then the commonly made assumption that the mutation backflow from the mutants to the wilde type is negligible does not apply.

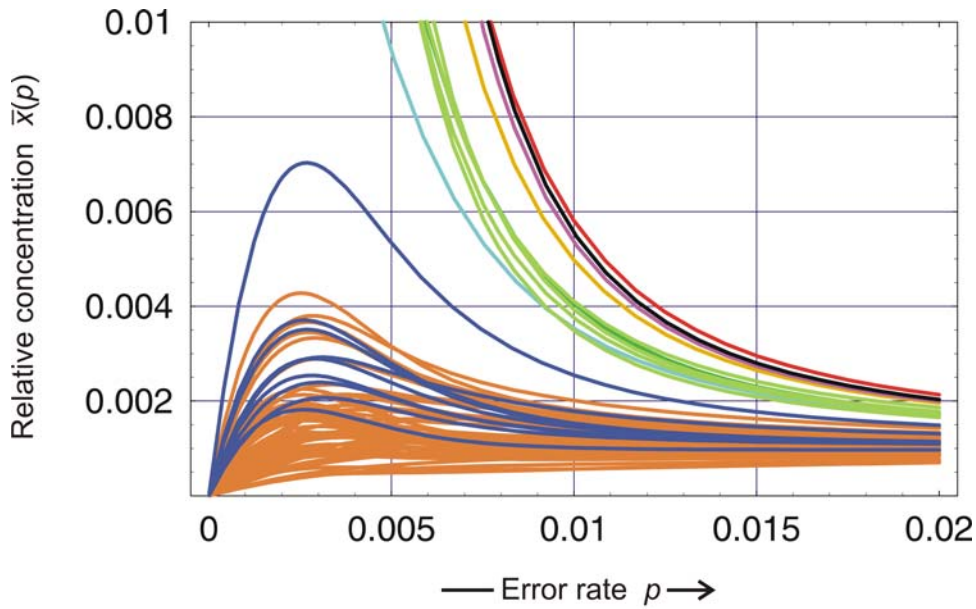
A kinetic theory of polynucleotide evolution which was developed during the past 15 years (Eigen, 1971; 1985; Eigen and Schuster, 1979; Eigen *et al.*, 1987; Schuster, 1986); Schuster and Sigmund, 1985) treats correct replication and mutation as parallel reactions within one and the same reaction network



Neutral network

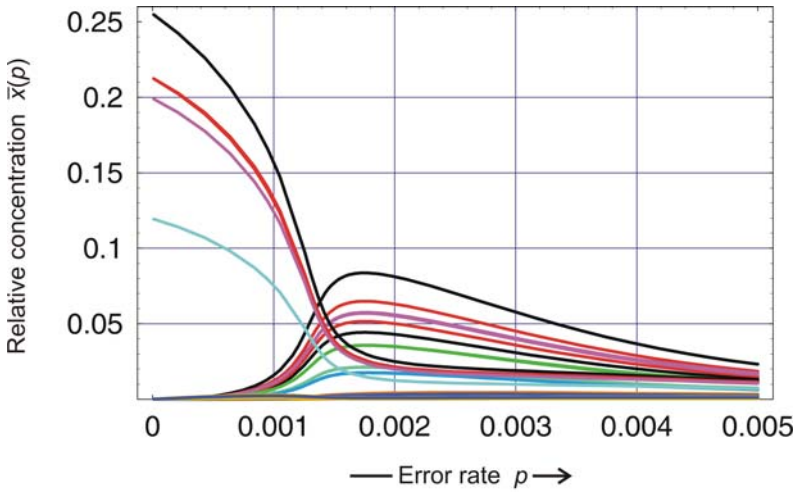
$\lambda = 0.10,\ s = 367$

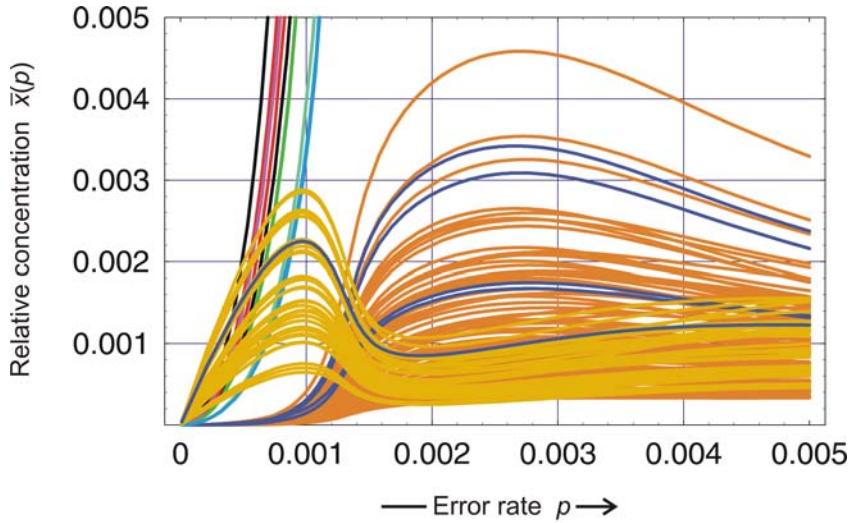


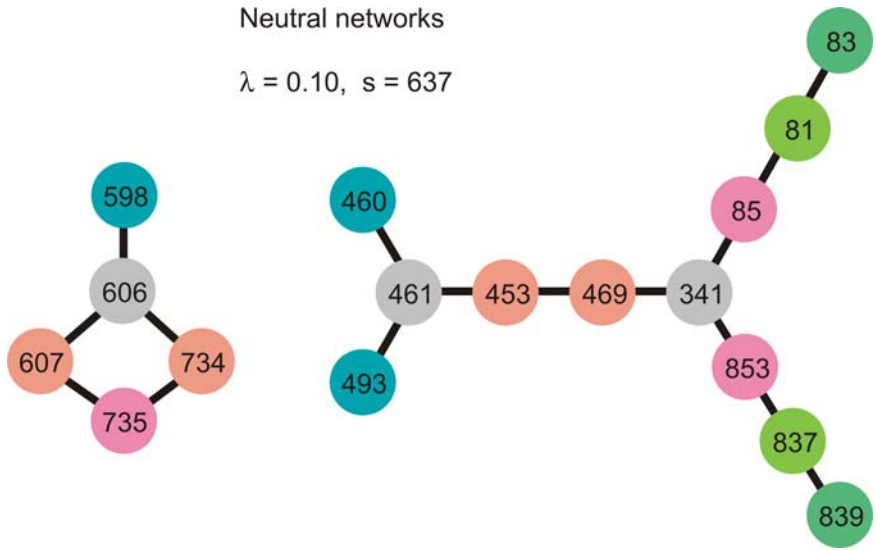


Error threshold: Individual sequences

$n = 10,\ \sigma = 1.1,\ d = 1.0$

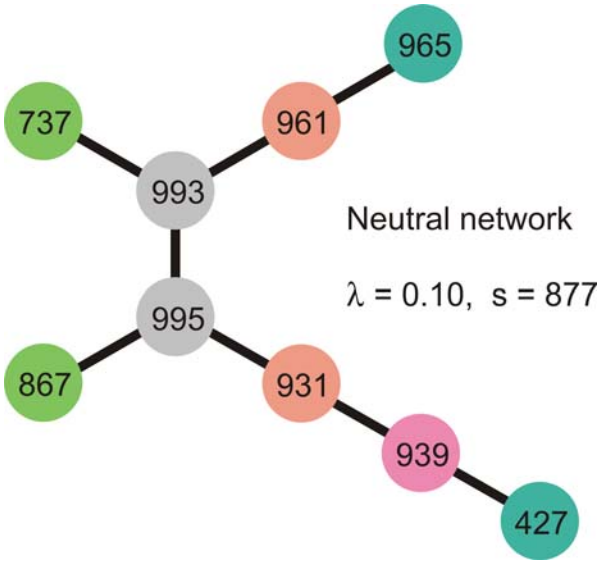






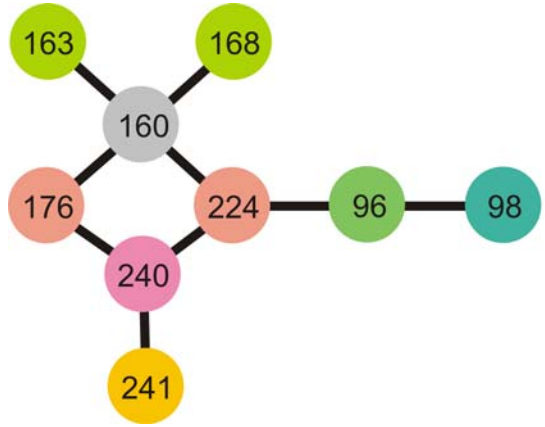
Error threshold: Individual sequences

n = 10, σ = 1.1, d = 1.0



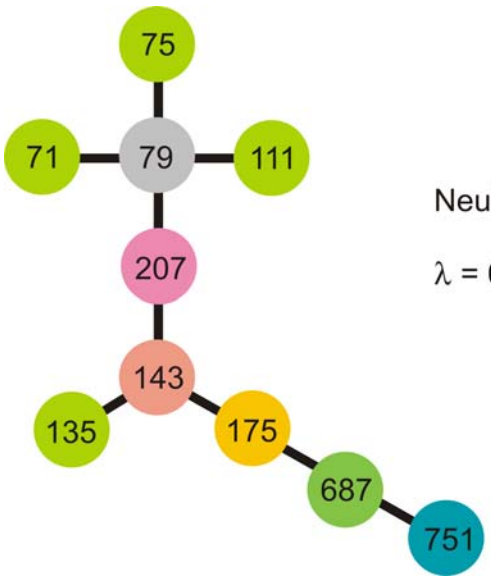
Neutral network

$\lambda = 0.10,\ s = 877$



Neutral network

$\lambda = 0.10,\ s = 491$

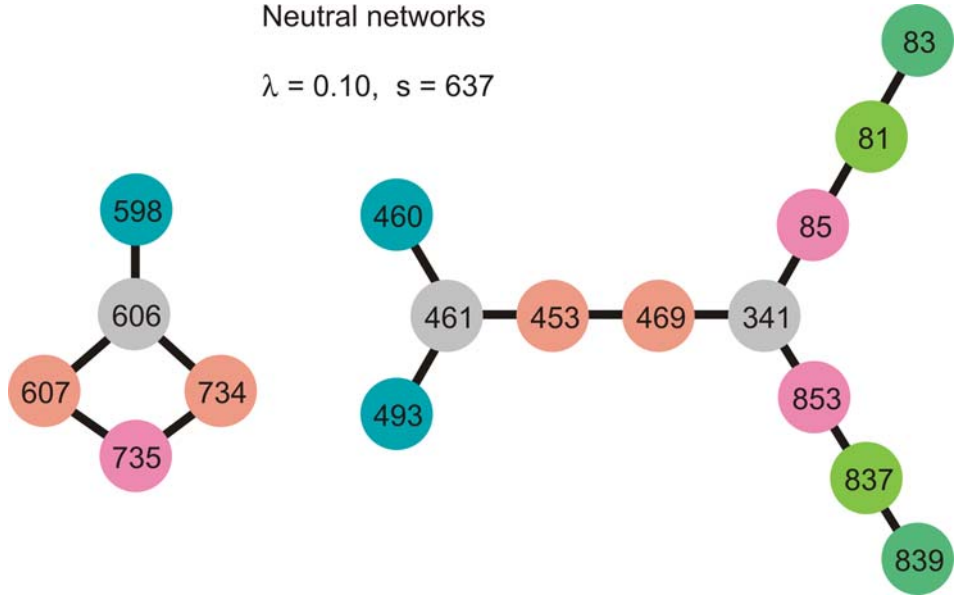


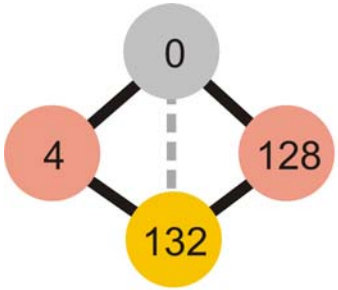
Neutral network

$\lambda = 0.10,\ s = 367$

Neutral networks

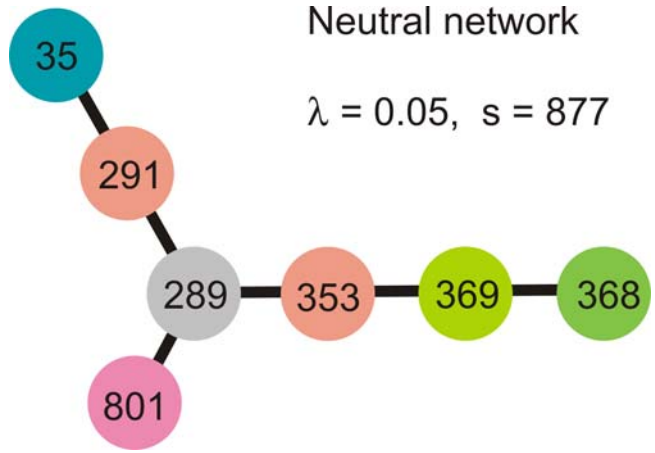
$\lambda = 0.10,\ s = 637$

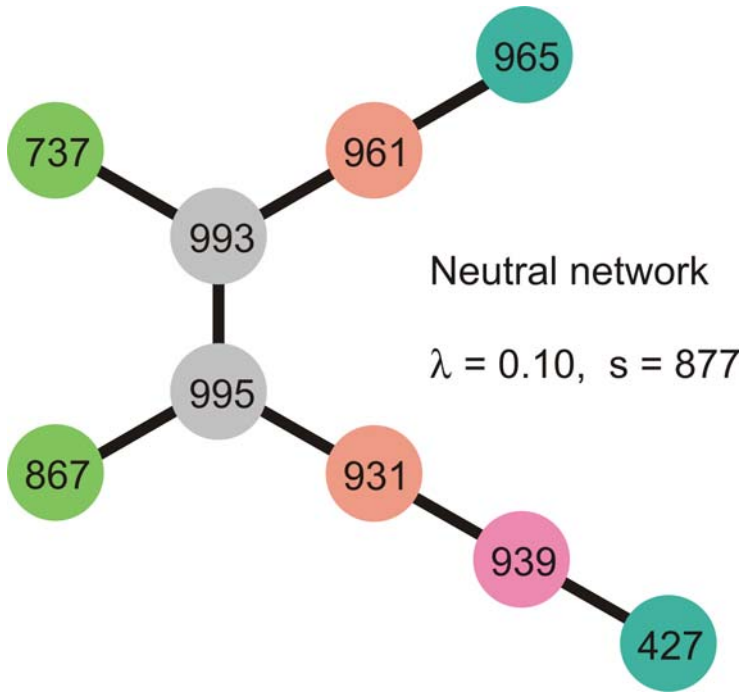




Neutral networks

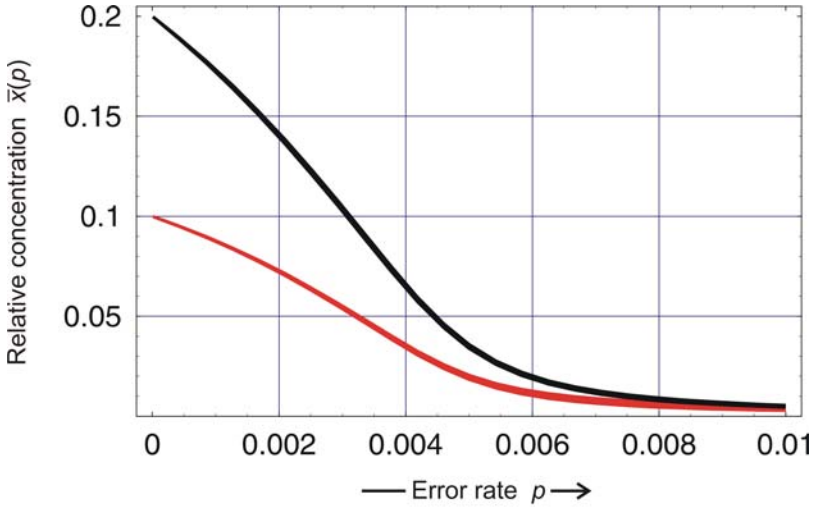
$\lambda = 0.01,\ s = 877$

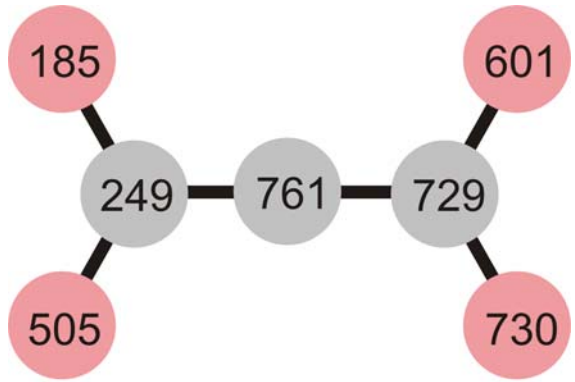




Neutral networks with increasing λ

$n = 10,\ \sigma = 1.1,\ d = 1.0$





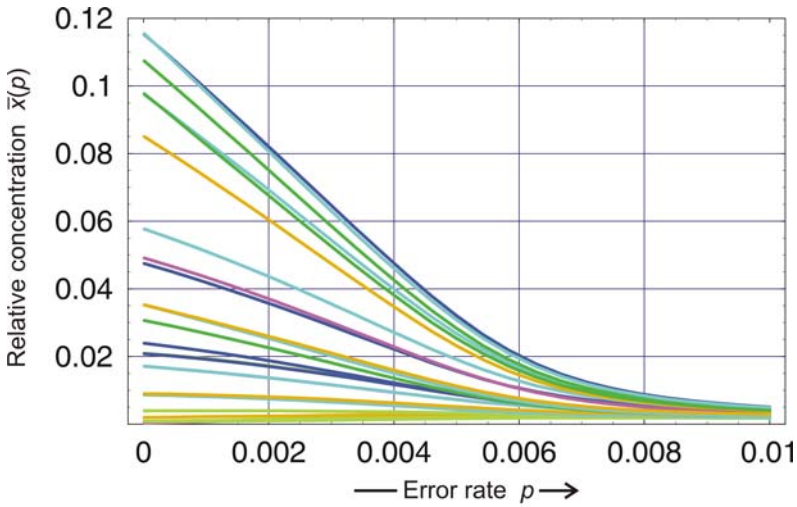
Neutral network

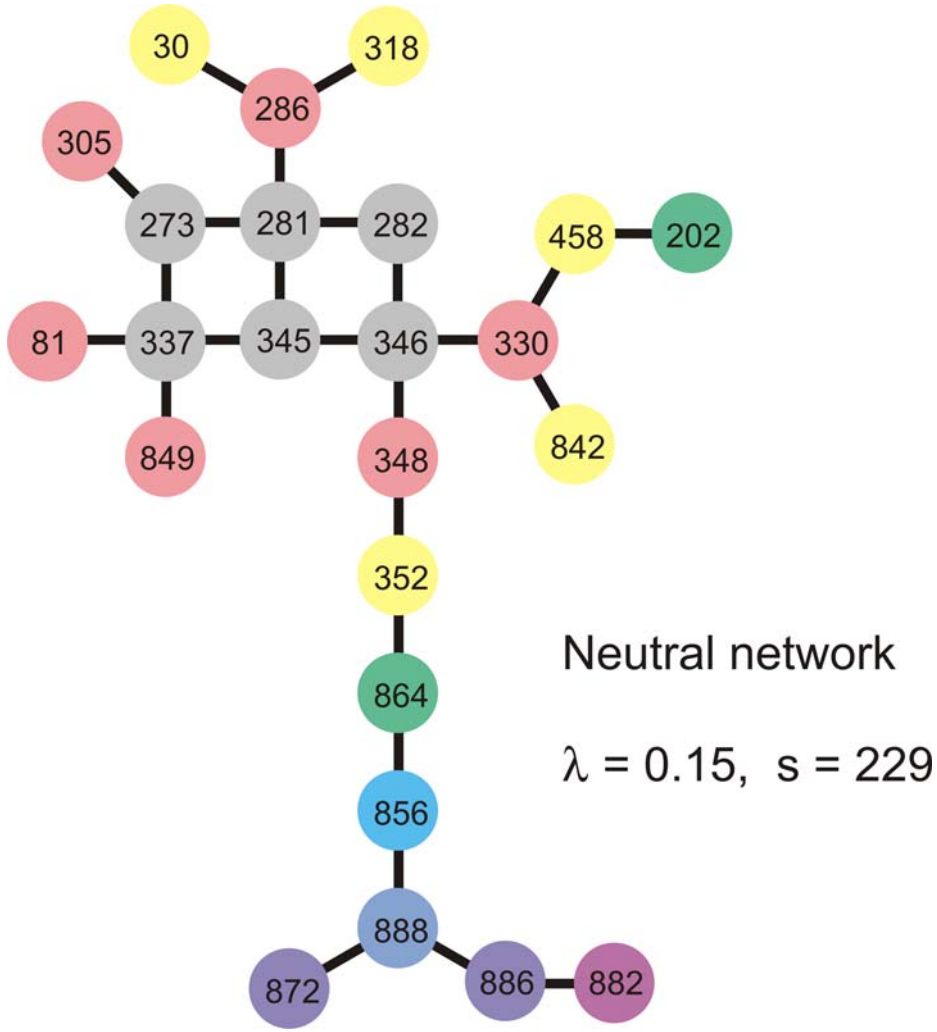
$\lambda = 0.10,\ s = 229$

N = 7

Neutral networks with increasing λ

n = 10, σ = 1.1, d = 1.0

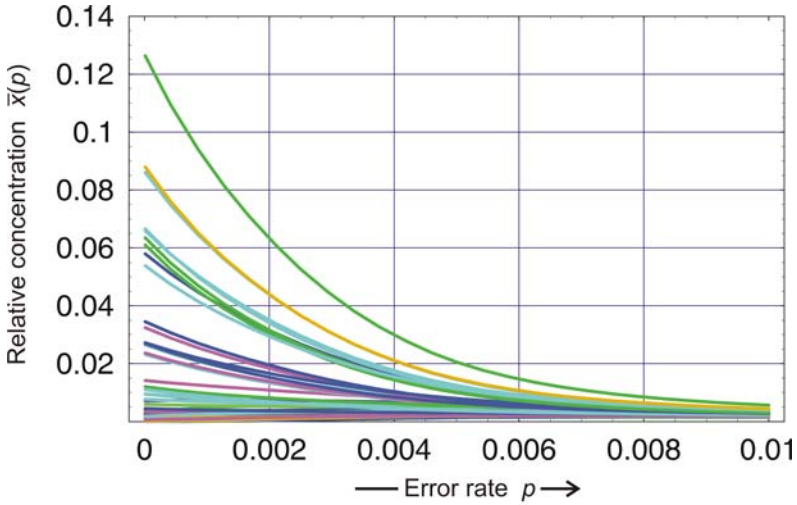


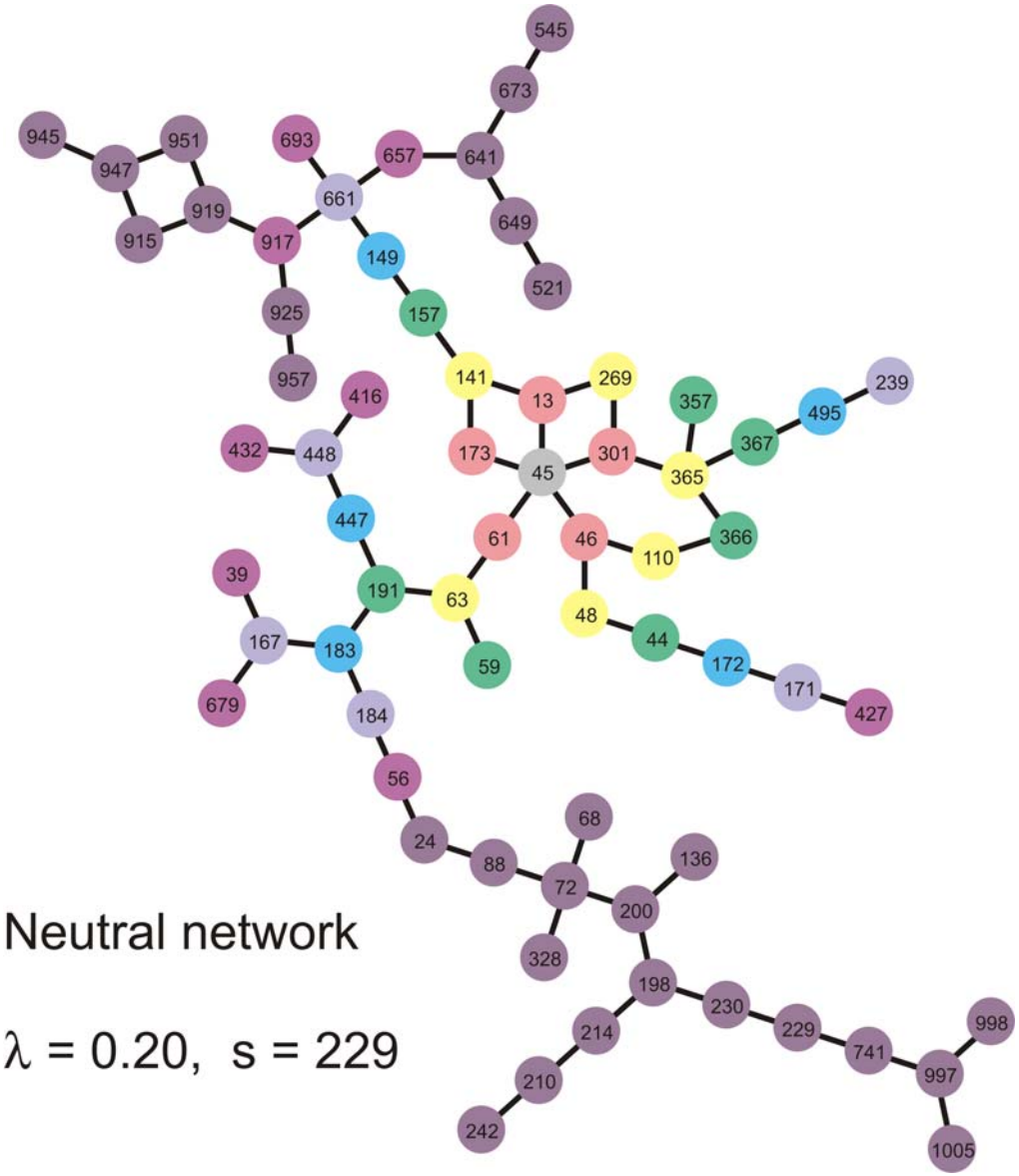


N = 24

Neutral networks with increasing λ

n = 10, σ = 1.1, d = 1.0





Neutral network

$\lambda = 0.20, \; s = 229$

N = 68

Neutral networks with increasing λ

n = 10, σ = 1.1, d = 1.0

1. Autocatalytic chemical reactions
2. Replication and biological information
3. Quasispecies and error thresholds
4. Neutral networks in evolution
5. **Evolutionary optimization**
6. Genetic regulation and metabolism

random individuals. The primer pair used for genomic DNA amplification is 5′-TCTCCCTGGATTCT-CATTTA-3′ (forward) and 5′-TCTTTGTCTTCTGT-TCCACC-3′ (reverse). Reactions were performed in 25 μl using 1 unit of Taq DNA polymerase with each primer at 0.4 μM; 200 μM each dATP, dTTP, dGTP, and dCTP; and PCR buffer [10 mM tris-HCl (pH 8.3), 50 mM KCl_2, 1.5 mM $MgCl_2$] in a cycle condition of 94°C for 1 min and then 35 cycles of 94°C for 30 s, 55°C for 30 s, and 72°C for 30 s followed by 72°C for 6 min. PCR products were purified (Qiagen), digested with Xmn I, and separated in a 2% agarose gel.

32. A nonsense mutation may affect mRNA stability and result in degradation of the transcript [L. Maquat, *Am. J. Hum. Genet.* 59, 279 (1996)].
33. Data not shown; a dot blot with poly $(A)^+$ RNA from 50 human tissues (The Human RNA Master Blot, 7770-1, Clontech Laboratories) was hybridized with a probe from exons 29 to 47 of *MYO15* using the same condition as Northern blot analysis (*13*).
34. Smith-Magenis syndrome (SMS) is due to deletions of 17p11.2 of various sizes, the smallest of which includes *MYO15* and perhaps 20 other genes [S. K-S Chen, L. Potocki, J. R. Lupski, *MMDD Res. Rev.* 2, 122 (1996)]. *MYO15* expression is easily detected in the pituitary gland (data not shown). Haploinsufficiency for *MYO15* may explain a portion of the SMS phenotype such as short stature. Moreover, a few SMS patients have sensorineural hearing loss, possibly because of a point mutation in *MYO15* in trans to the SMS 17p11.2 deletion.
35. R. A. Fridell, data not shown.
36. K. B. Avraham *et al.*, *Nature Genet.* 11, 369 (1995); X-Z. Liu *et al.*, *ibid.* 17, 268 (1997); F. Gibson *et al.*, *Nature* 374, 62 (1995); D. Weil *et al.*, *ibid.*, p. 60.
37. RNA was extracted from cochlea (membranous labyrinths) obtained from human fetuses at 18 to 22 weeks of development in accordance with guidelines established by the Human Research Committee at the Brigham and Women's Hospital. Only samples without evidence of degradation were pooled for poly $(A)^+$ selection over oligo(dT) columns. First-strand cDNA was prepared using an Advantage RT-for-PCR kit (Clontech Laboratories). A portion of the first-strand cDNA (4%) was amplified by PCR with Advantage cDNA polymerase mix (Clontech Laboratories) using human *MYO15*-specific oligonucleotide primers (forward, 5′-GCATGAGCTGCCGGCTAAT-GGG-3′; reverse, 5′-CTCACGGCTTCTGCATGGT-GCTCGGCTGGC-3′). Cycling conditions were 40 s at 94°C; 40 s at 66°C (3 cycles), 60°C (5 cycles), and 55°C (29 cycles); and 45 s at 68°C. PCR products were visualized by ethidium bromide staining after fractionation in a 1% agarose gel. A 688-bp PCR product is expected from amplification of the human *MYO15* cDNA. Amplification of human genomic DNA with this primer pair would result in a 2908-bp fragment.
38. We are grateful to the people of Bengkala, Bali, and the two families from India. We thank J. R. Lupski and K.-S. Chen for providing the human chromosome 17 cosmid library. For technical and computational assistance, we thank N. Dietrich, M. Ferguson, A. Gupta, E. Sorbello, R. Torkzadeh, C. Varner, M. Walker, G. Bourland, and S. Blackstrom-Sternberg (National Institutes of Health Intramural Sequencing Center). We thank J. T. Himant, I. N. Arhya, and S. Winata for assistance in Bali, and T. Barber, S. Sullivan, E. Green, D. Drayna, and J. Battey for helpful comments on this manuscript. Supported by the National Institute on Deafness and Other Communication Disorders (NIDCD) (Z01 DC 00035-01 and Z01 DC 00039-01 to T.B.F. and E.R.W. and R01 DC 03402 to C.C.M.), the National Institute of Child Health and Human Development (R01 HD30428 to S.A.C.) and a National Science Foundation Graduate Research Fellowship to F.J.P. This paper is dedicated to J. B. Snow Jr. on his retirement as the Director of the NIDCD.

9 March 1998; accepted 17 April 1998

Continuity in Evolution: On the Nature of Transitions

Walter Fontana and Peter Schuster

To distinguish continuous from discontinuous evolutionary change, a relation of nearness between phenotypes is needed. Such a relation is based on the probability of one phenotype being accessible from another through changes in the genotype. This nearness relation is exemplified by calculating the shape neighborhood of a transfer RNA secondary structure and provides a characterization of discontinuous shape transformations in RNA. The simulation of replicating and mutating RNA populations under selection shows that sudden adaptive progress coincides mostly, but not always, with discontinuous shape transformations. The nature of these transformations illuminates the key role of neutral genetic drift in their realization.

A much-debated issue in evolutionary biology concerns the extent to which the history of life has proceeded gradually or has been punctuated by discontinuous transitions at the level of phenotypes (*1*). Our goal is to make the notion of a discontinuous transition more precise and to understand how it arises in a model of evolutionary adaptation.

We focus on the narrow domain of RNA secondary structure, which is currently the simplest computationally tractable, yet realistic phenotype (*2*). This choice enables the definition and exploration of concepts that may prove useful in a wider context. RNA secondary structures represent a coarse level of analysis compared with the three-dimensional structure at atomic resolution. Yet, secondary structures are empirically well defined and obtain their biophysical and biochemical importance from being a scaffold for the tertiary structure. For the sake of brevity, we shall refer to secondary structures as "shapes." RNA combines in a single molecule both genotype (replicatable sequence) and phenotype (selectable shape), making it ideally suited for in vitro evolution experiments (*3*, *4*).

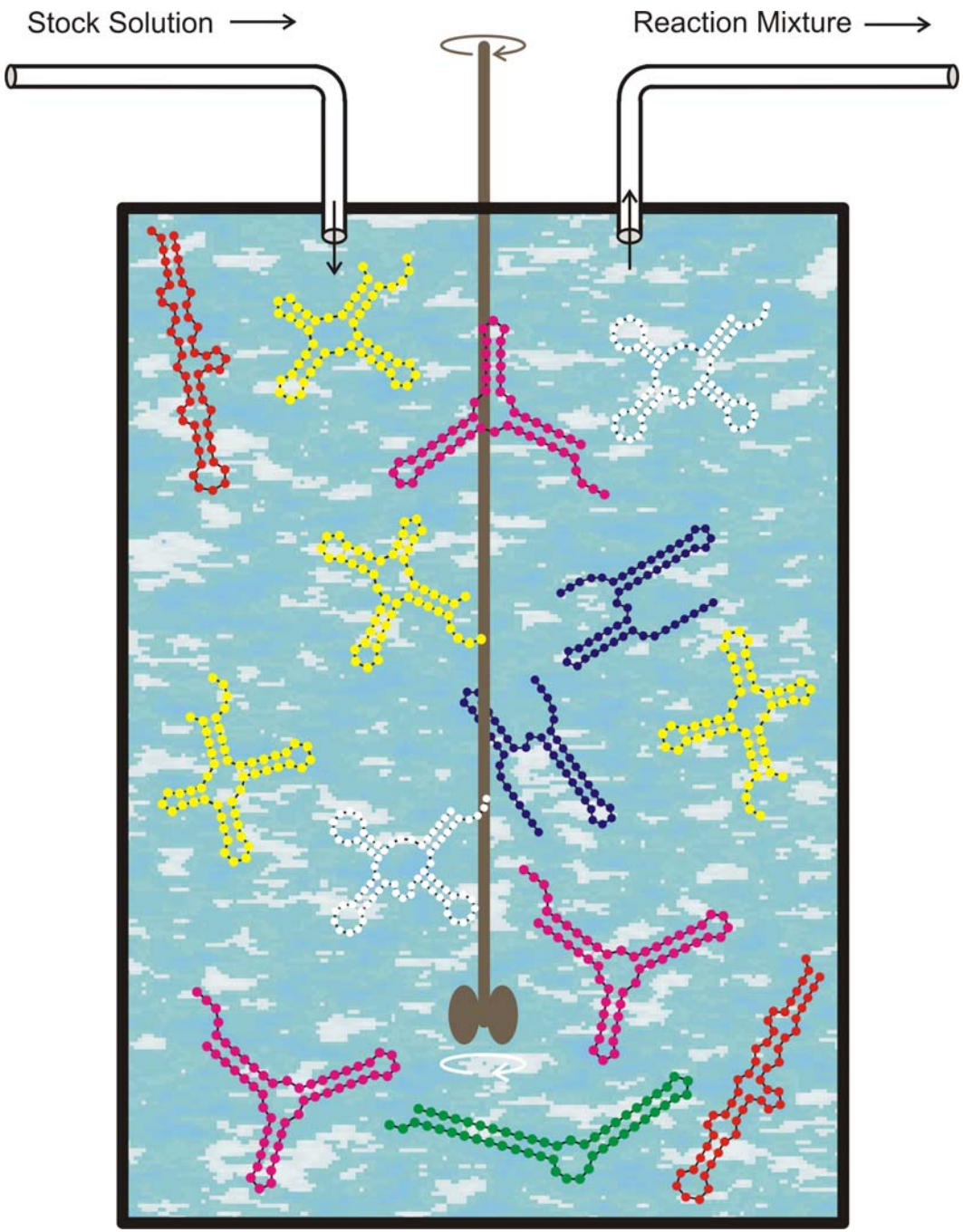
To generate evolutionary histories, we used a stochastic continuous time model of an RNA population replicating and mutating in a capacity-constrained flow reactor under selection (*5*, *6*). In the laboratory, a goal might be to find an RNA aptamer binding specifically to a molecule (*4*). Although in the experiment the evolutionary end product was unknown, we thought of its shape as being specified implicitly by the imposed selection criterion. Because our intent is to study evolutionary histories rather than end products, we defined a target shape in advance and assumed the replication rate of a sequence to be a function of the similarity between its shape and the target. An actual situation may involve more than one best shape, but this does not affect our conclusions.

An instance representing in its qualitative features all the simulations we performed is shown in Fig. 1A. Starting with identical sequences folding into a random shape, the simulation was stopped when the population became dominated by the target, here a canonical tRNA shape. The black curve traces the average distance to the target (inversely related to fitness) in the population against time. Aside from a short initial phase, the entire history is dominated by steps, that is, flat periods of no apparent adaptive progress, interrupted by sudden approaches toward the target structure (*7*). However, the dominant shapes in the population not only change at these marked events but undergo several fitness-neutral transformations during the periods of no apparent progress. Although discontinuities in the fitness trace are evident, it is entirely unclear when and on the basis of what the series of successive phenotypes itself can be called continuous or discontinuous.

A set of entities is organized into a (topological) space by assigning to each entity a system of neighborhoods. In the present case, there are two kinds of entities: sequences and shapes, which are related by a thermodynamic folding procedure. The set of possible sequences (of fixed length) is naturally organized into a space because point mutations induce a canonical neighborhood. The neighborhood of a sequence consists of all its one-error mutants. The problem is how to organize the set of possible shapes into a space. The issue arises because, in contrast to sequences, there are

Institut für Theoretische Chemie, Universität Wien, Währingerstrasse 17, A-1090 Wien, Austria, Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA, and International Institute for Applied Systems Analysis (IIASA), A-2361 Laxenburg, Austria.

Stochastic simulation of evolution of RNA molecules



Replication rate constant:

$$f_k = \gamma / [\alpha + \Delta d_S^{(k)}]$$

$$\Delta d_S^{(k)} = d_H(S_k, S_\tau)$$

Selection constraint:

Population size, N = # RNA molecules, is controlled by the flow

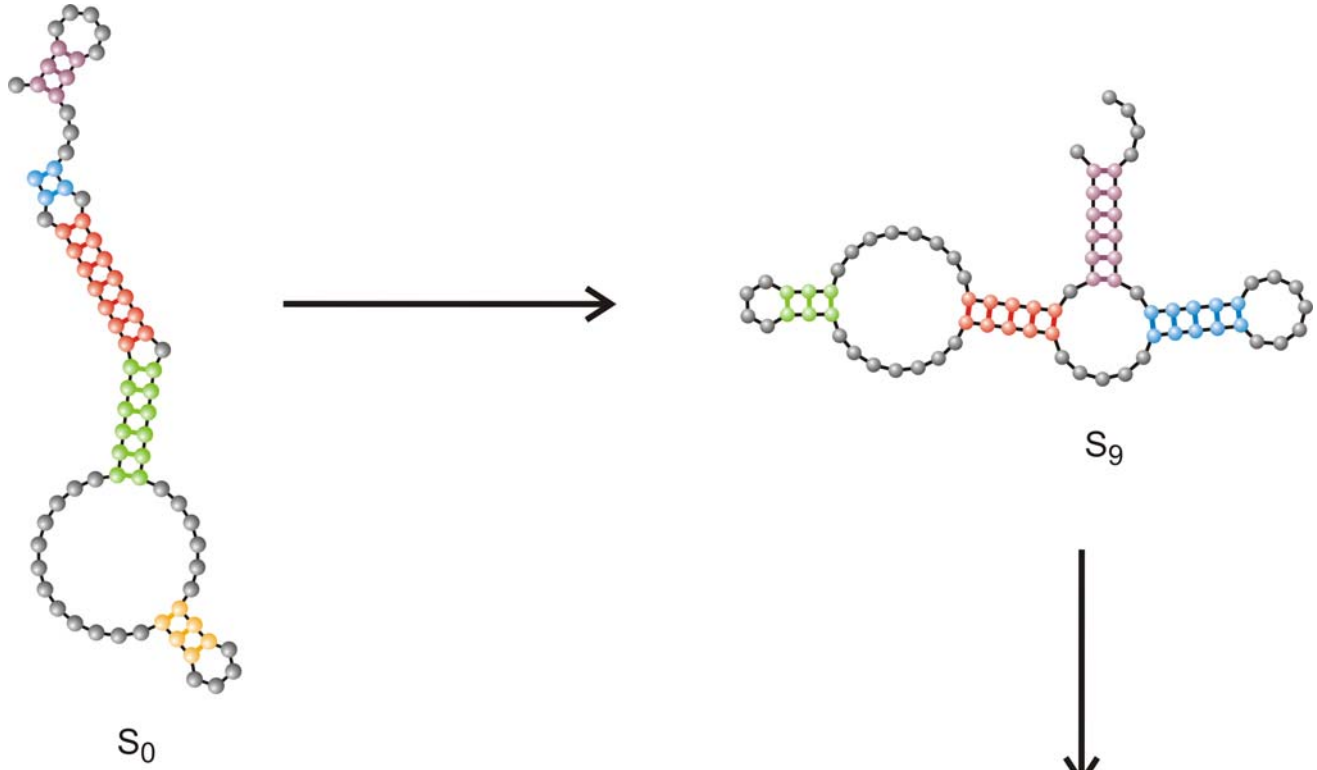
$$N(t) \approx \overline{N} \pm \sqrt{\overline{N}}$$

Mutation rate:

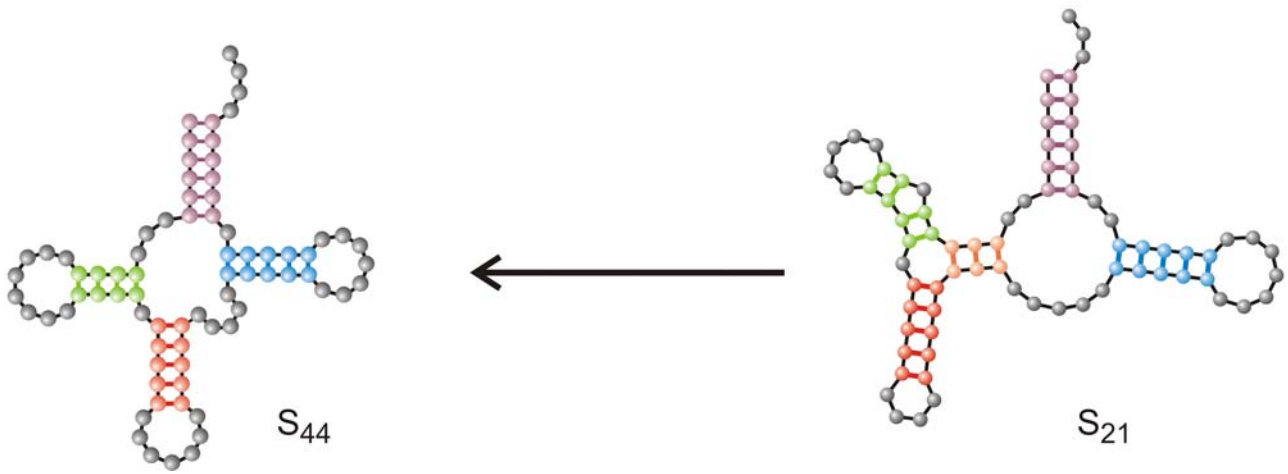
p = 0.001 / site × replication

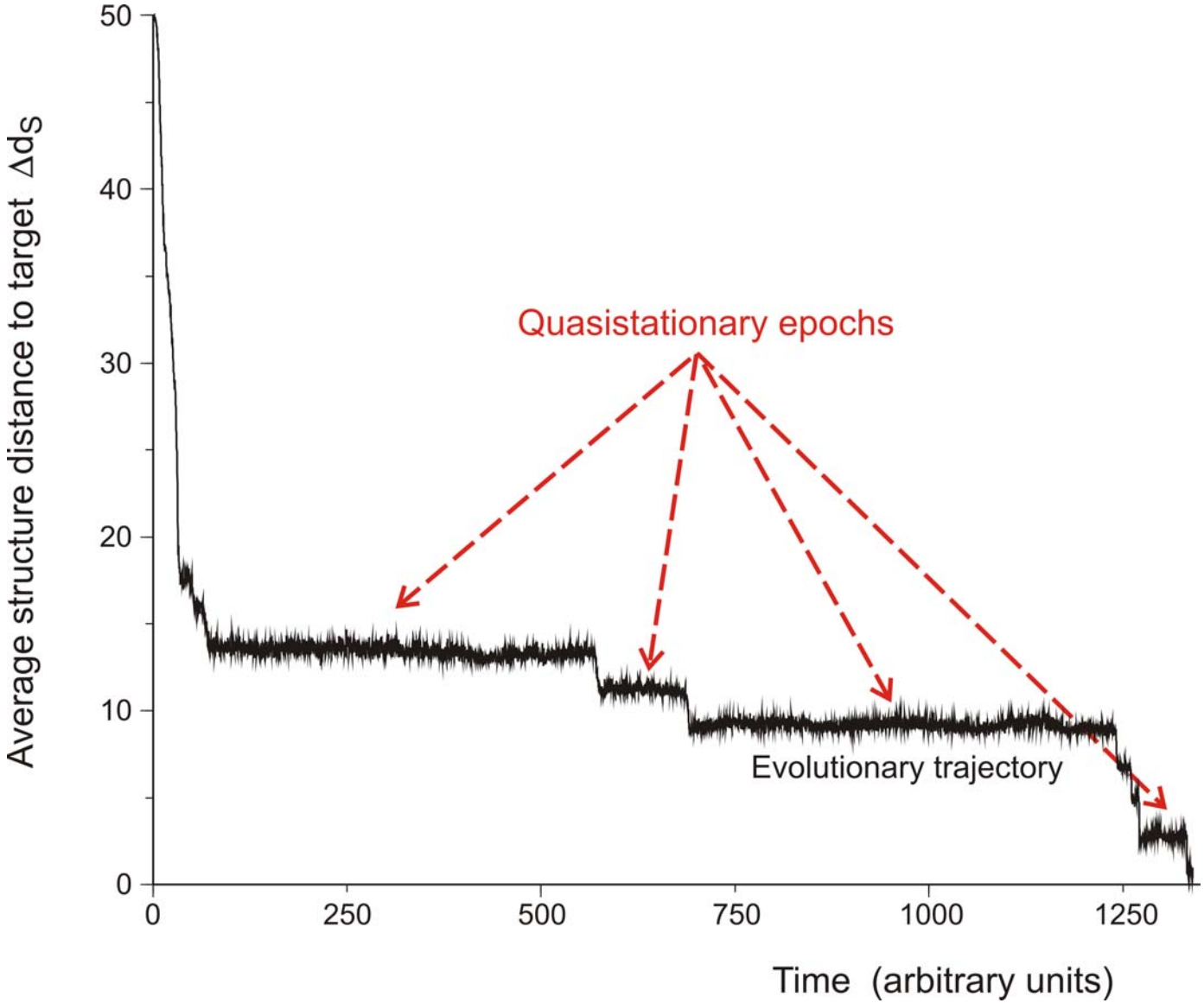
The flowreactor as a device for studies of evolution ***in vitro*** and ***in silico***

Randomly chosen
initial structure



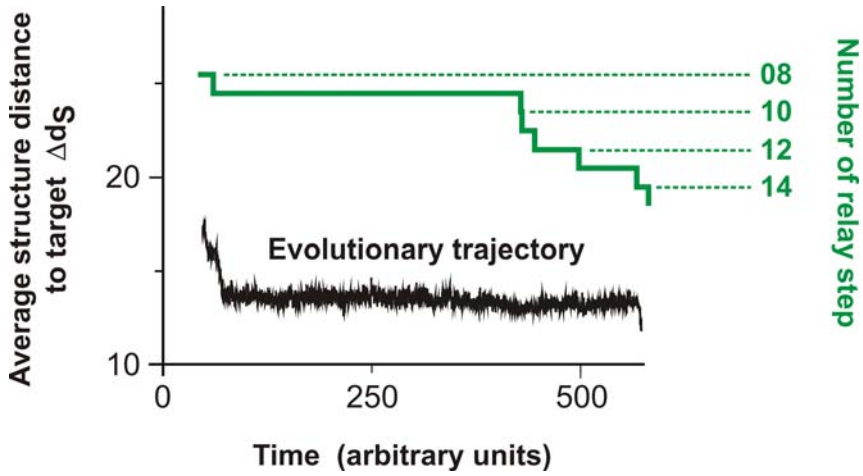
Phenylalanyl-tRNA
as target structure





In silico optimization in the flow reactor: Evolutionary Trajectory

28 neutral point mutations during a long quasi-stationary epoch

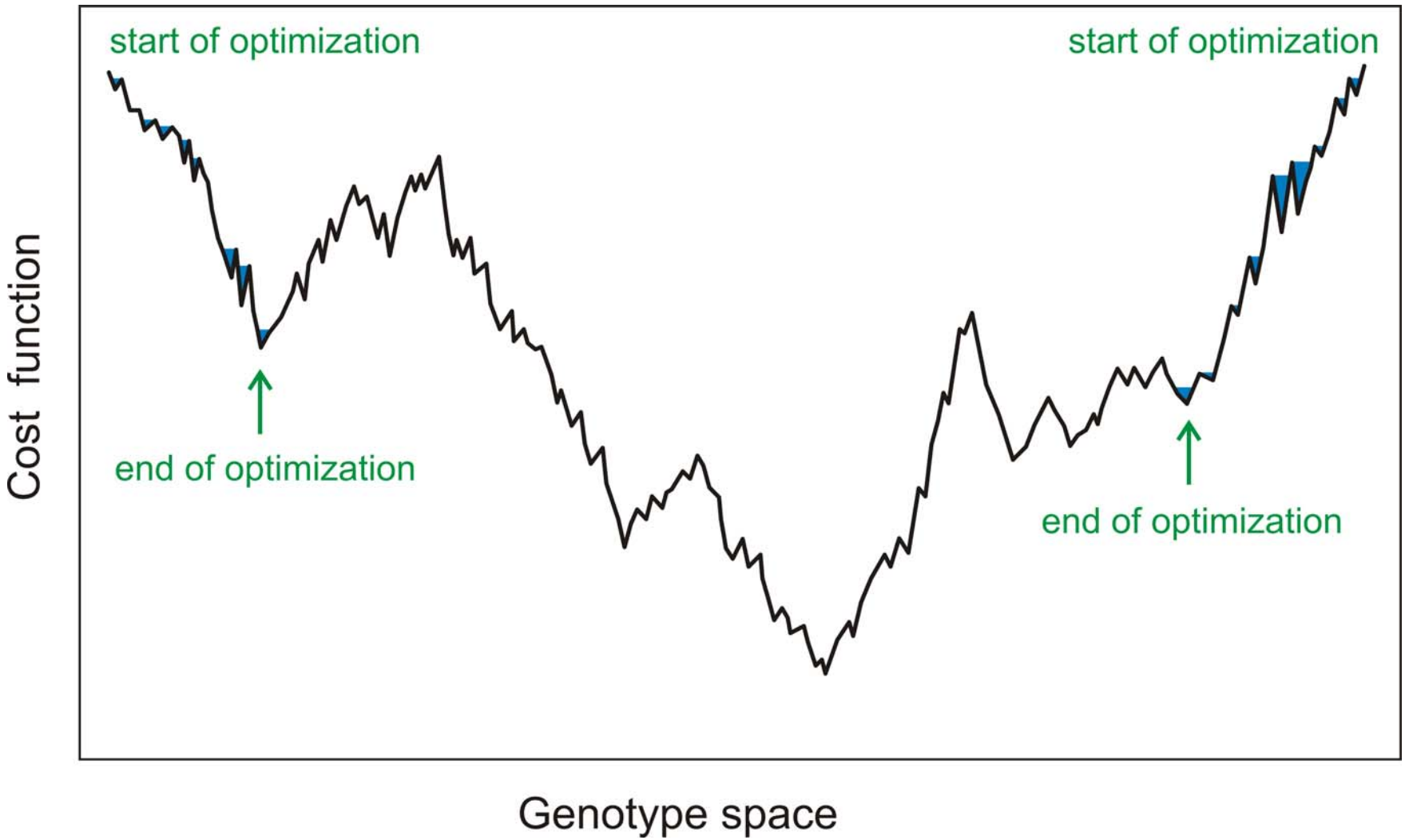


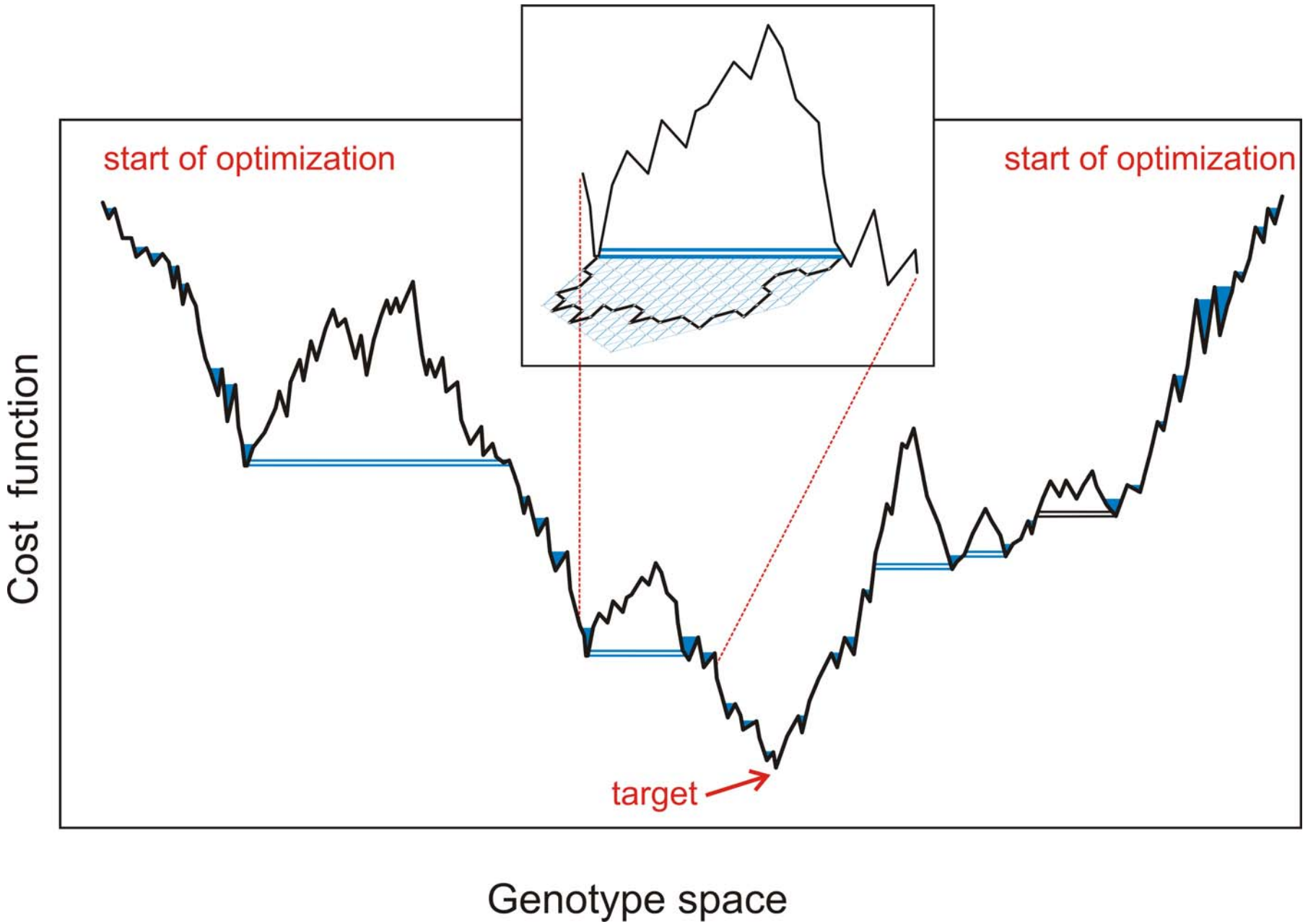

```
entry   GGUAUGGGCGUUGAAUAGUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUCCCGUACAGAA
8       .((((((((((((.........(((....)))).......)))))).....((((.......))))))))))....
exit    GGUAUGGGCGUUGAAUAAUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUCCCAUACAGAA
entry   GGUAUGGGCGUUGAAUAAUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUACCAUACAGAA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

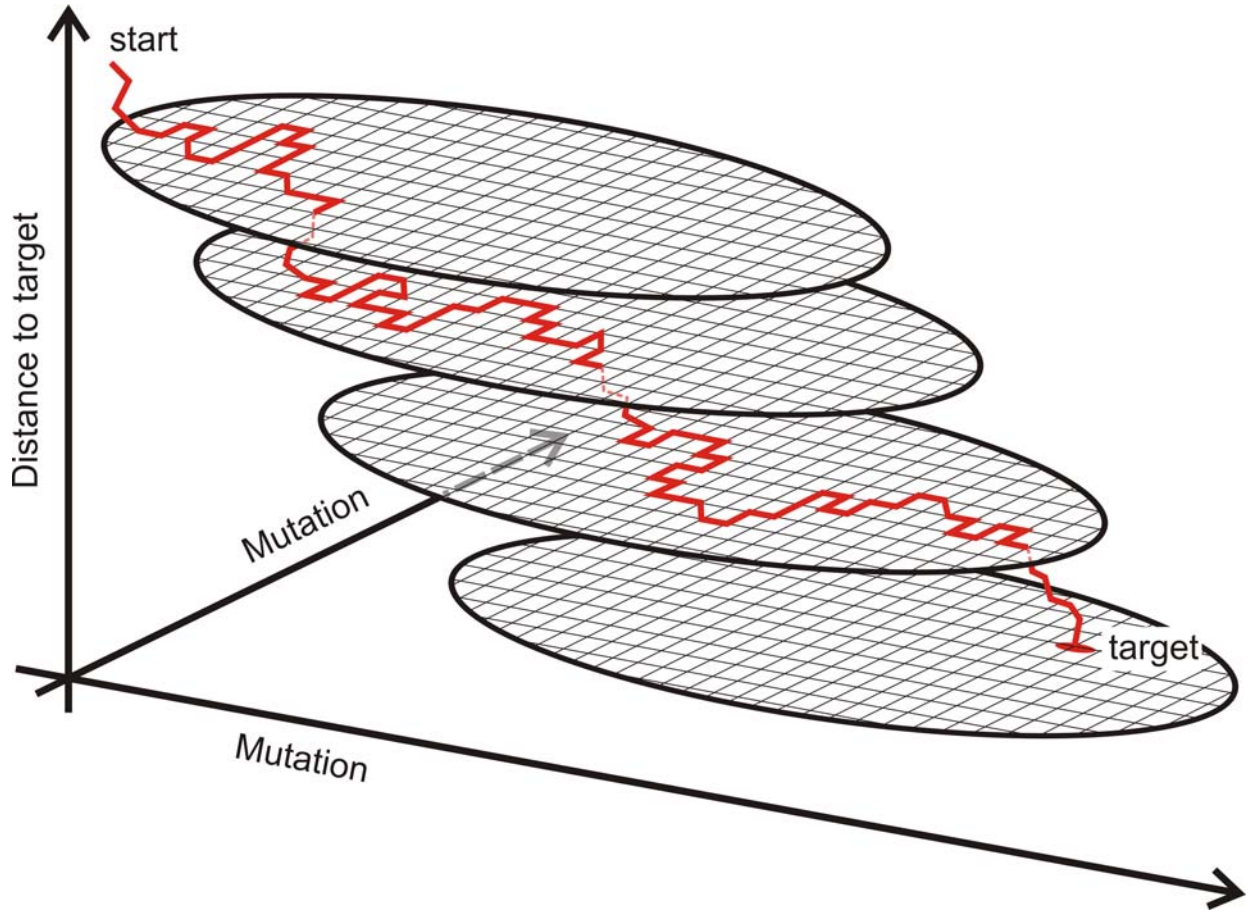
Transition inducing point mutations change the molecular structure

Neutral point mutations leave the molecular structure unchanged

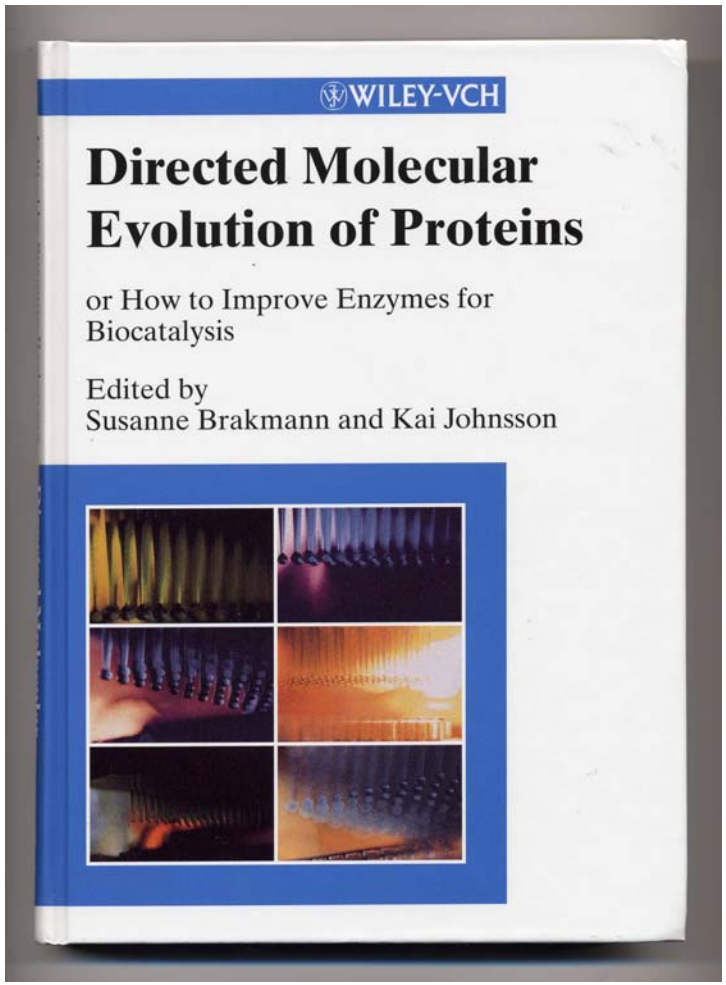
Neutral genotype evolution during phenotypic stasis

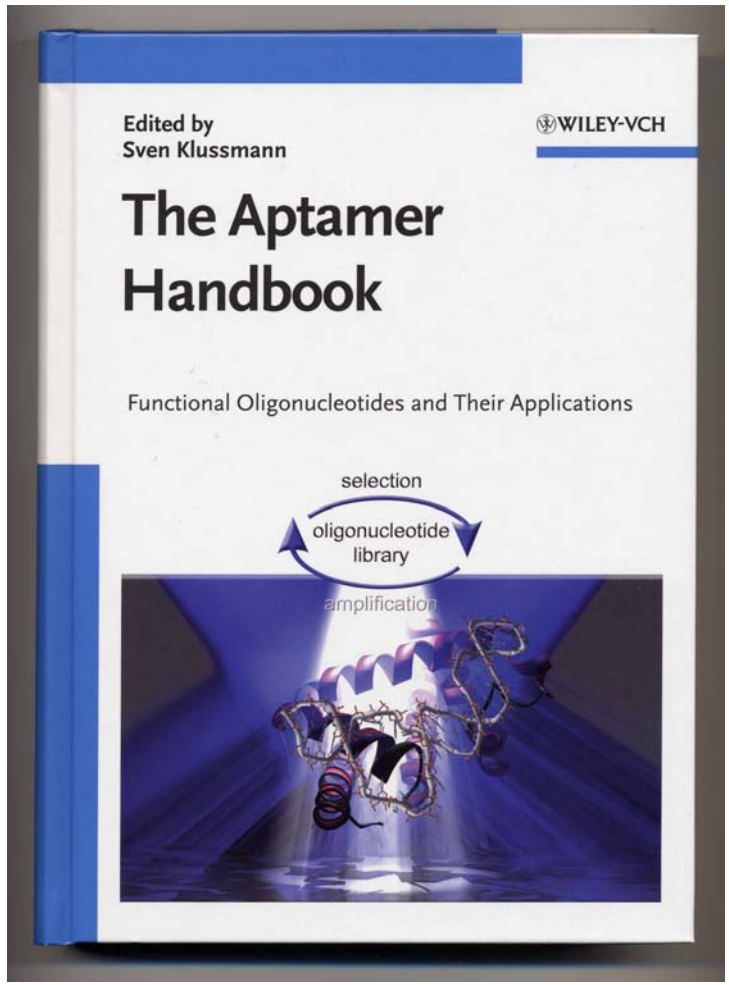

start of optimization
start of optimization
end of optimization
end of optimization
Cost function
Genotype space


start of optimization
start of optimization
target
Cost function
Genotype space



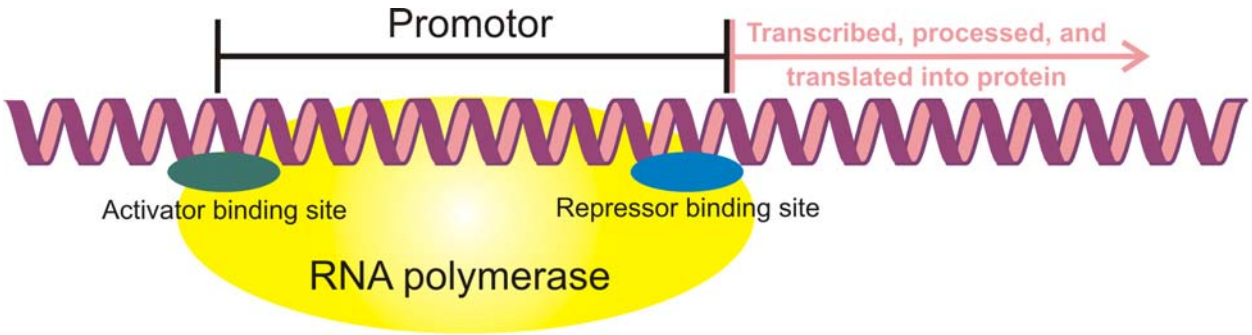
A sketch of optimization on neutral networks





Application of molecular evolution to problems in biotechnology

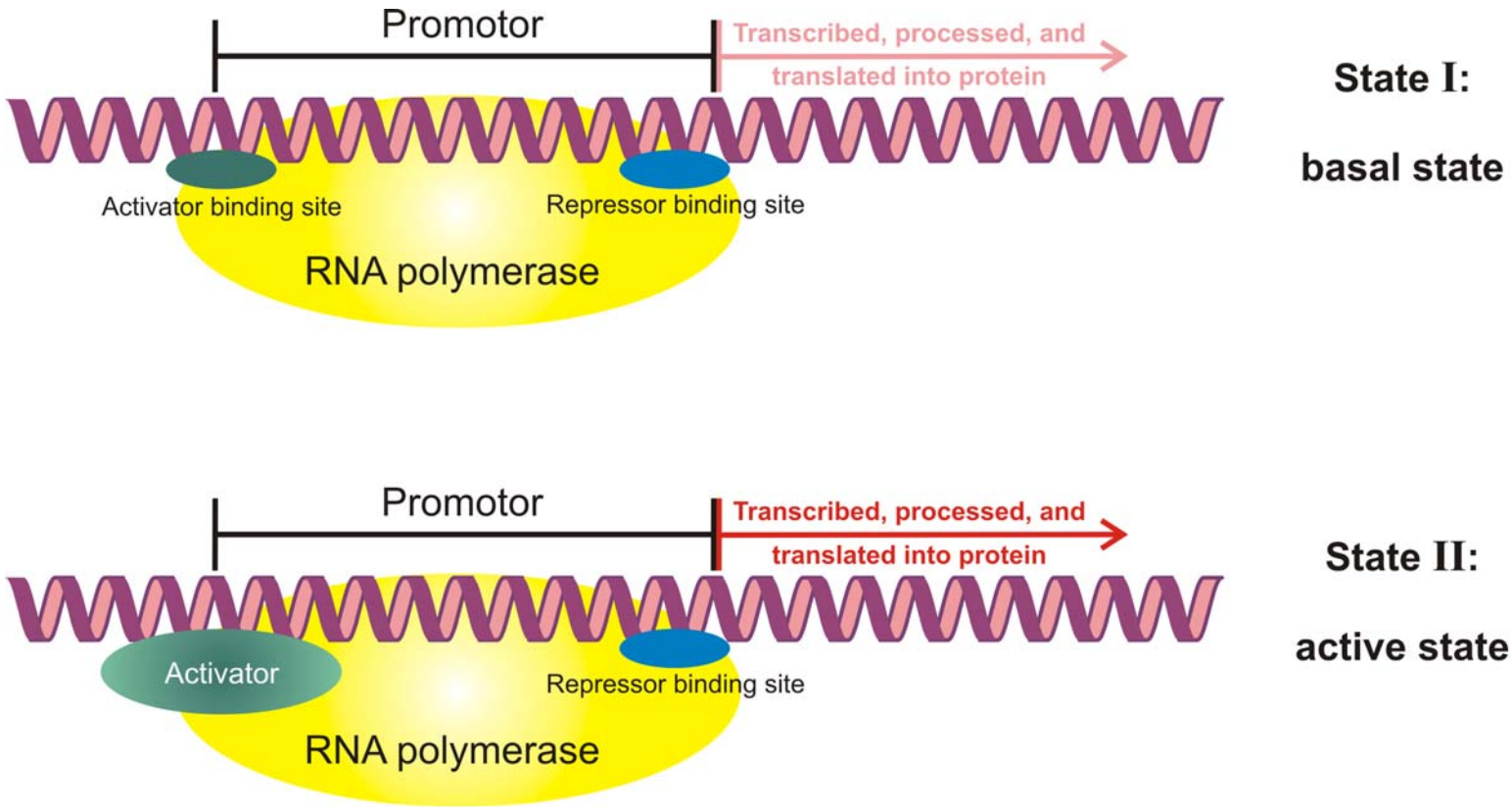
1. Autocatalytic chemical reactions
2. Replication and biological information
3. Quasispecies and error thresholds
4. Neutral networks in evolution
5. Evolutionary optimization
6. **Genetic regulation and metabolism**



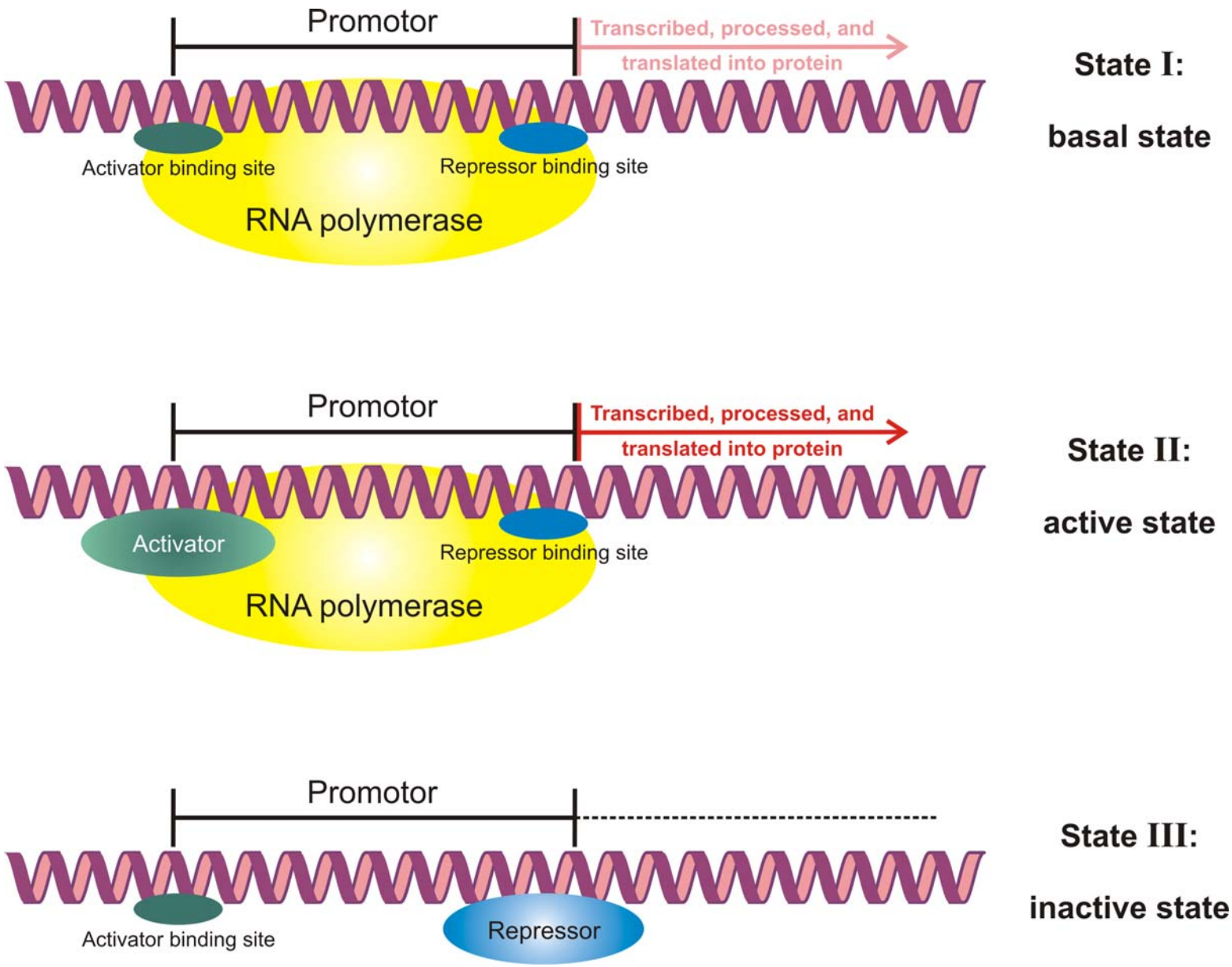
State I:

basal state

States of gene regulation in a bacterial expression control system



States of gene regulation in a bacterial expression control system



States of gene regulation in a bacterial expression control system

Available online at www.sciencedirect.com
ScienceDirect
Journal of Theoretical Biology 246 (2007) 395–419
Journal of Theoretical Biology
www.elsevier.com/locate/yjtbi

Dynamic patterns of gene regulation I: Simple two-gene systems

Stefanie Widder[a], Josef Schicho[b], Peter Schuster[a,c,*]

[a]*Institut für Theoretische Chemie der Universität Wien, Währingerstraße 17, A-1090 Wien, Austria*
[b]*RICAM—Johann Radon Institute for Computational and Applied Mathematics of the Austrian Academy of Sciences, Altenbergerstraße 69, A-4040 Linz, Austria*
[c]*Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA*

Received 24 February 2006; received in revised form 7 January 2007; accepted 8 January 2007
Available online 16 January 2007

Abstract

Regulation of gene activities is studied by means of computer assisted mathematical analysis of ordinary differential equations (ODEs) derived from binding equilibria and chemical reaction kinetics. Here, we present results on cross-regulation of two genes through activator and/or repressor binding. Arbitrary (differentiable) binding function can be used but systematic investigations are presented for gene–regulator complexes with integer valued Hill coefficients up to $n = 4$. The dynamics of gene regulation is derived from bifurcation patterns of the underlying systems of kinetic ODEs. In particular, we present analytical expressions for the parameter values at which one-dimensional (transcritical, saddle-node or pitchfork) and/or two-dimensional (Hopf) bifurcations occur. A classification of regulatory states is introduced, which makes use of the sign of a 'regulatory determinant' D (being the determinant of the block in the Jacobian matrix that contains the derivatives of the regulator binding functions): (i) systems with $D<0$, observed, for example, if both proteins are activators or repressors, to give rise to one-dimensional bifurcations only and lead to bistability for $n \geqslant 2$ and (ii) systems with $D>0$, found for combinations of activation and repression, sustain a Hopf bifurcation and undamped oscillations for $n>2$. The influence of basal transcription activity on the bifurcation patterns is described. Binding of multiple subunits can lead to richer dynamics than pure activation or repression states if intermediates between the unbound state and the fully saturated DNA initiate transcription. Then, the regulatory determinant D can adopt both signs, plus and minus.

Keywords: Basal transcription; Bifurcation analysis; Cooperative binding; Gene regulation; Hill coefficient; Hopf bifurcation

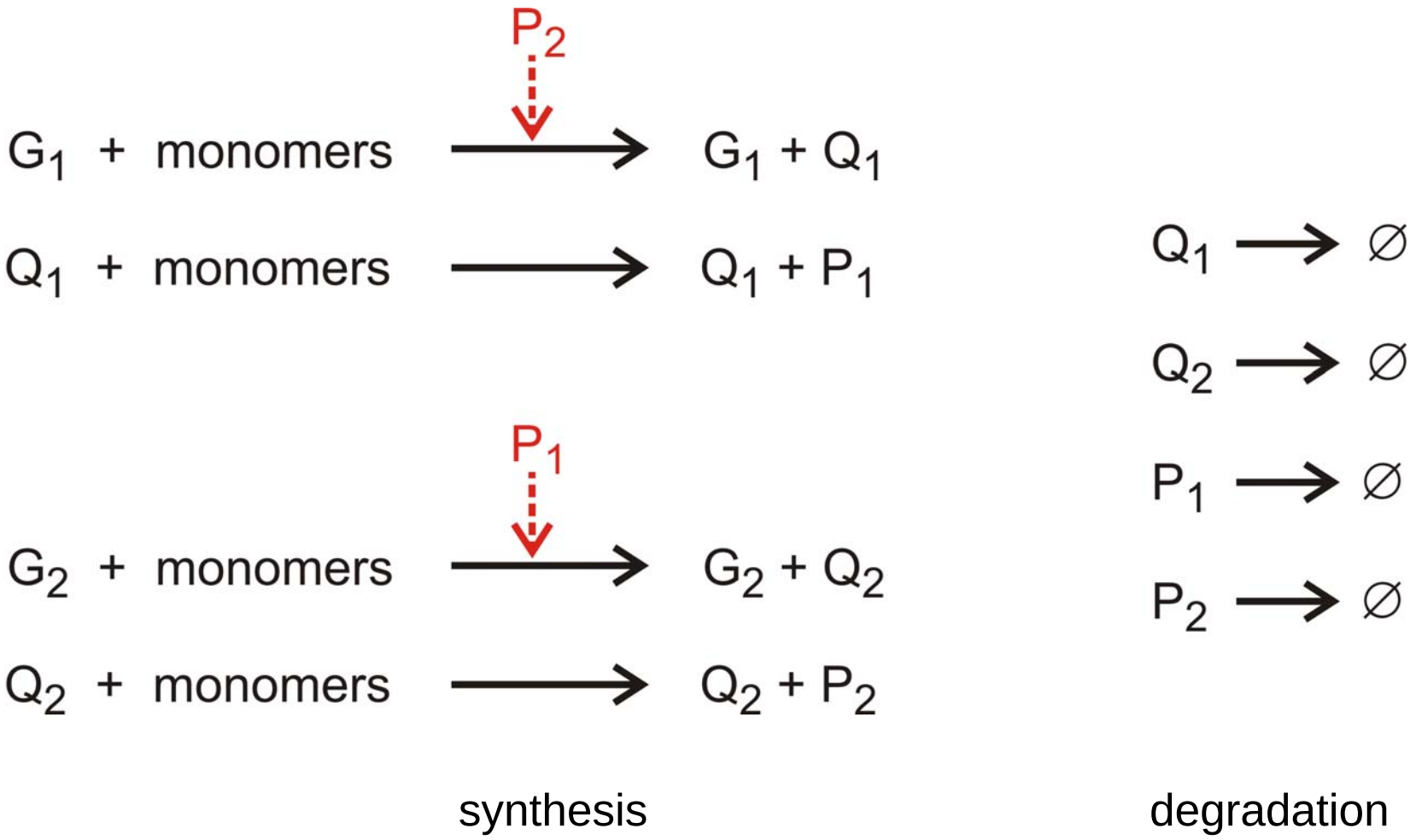
1. Introduction

Theoretical work on gene regulation goes back to the 1960s (Monod et al., 1963) soon after the first repressor protein had been discovered (Jacob and Monod, 1961). A little later the first paper on oscillatory states in gene regulation was published (Goodwin, 1965). The interest in gene regulation and its mathematical analysis never ceased (Tiwari et al., 1974; Tyson and Othmer, 1978; Smith, 1987) and saw a great variety of different attempts to design models of genetic regulatory networks that can be used in systems biology for computer simulation of *gen*(etic and met)*abolic* networks.[1] Most models in the literature aim at a minimalist dynamic description which, nevertheless, tries to account for the basic regulatory functions of large networks in the cell in order to provide a better understanding of cellular dynamics. A classic in general regulatory dynamics is the monograph by Thomas and D'Ari (1990). The currently used mathematical methods comprise application of Boolean logic (Thomas and Kaufman, 2001b; Savageau, 2001; Albert and Othmer, 2003), stochastic processes (Hume, 2000) and deterministic dynamic models, examples are Cherry and Adler (2000), Bindschadler and Sneyd (2001) and Kobayashi et al. (2003) and the recent elegant analysis of bistability (Craciun et al.,

*Corresponding author. Institut für Theoretische Chemie der Universität Wien, Währingerstraße 17, A-1090 Wien, Austria.
Tel.: +43 1 4277 52743; fax: +43 1 4277 52793.
E-mail address: pks@tbi.univie.ac.at (P. Schuster).

[1]Discussion and analysis of combined genetic and metabolic networks has become so frequent and intense that we suggest to use a separate term, *genabolic networks*, for this class of complex dynamical systems.

0022-5193/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2007.01.004



Cross-regulation of two genes

$[\mathrm{G}_1]=[\mathrm{G}_2]=g_0=\text{const.}$

$[\mathrm{Q}_1]=q_1,[\mathrm{Q}_2]=q_2,$

$[\mathrm{P}_1]=p_1,[\mathrm{P}_2]=p_2$

$$\text{Activation}: \quad F_i(p_j)=\frac{p_j^{\mathrm{n}}}{K+p_j^{\mathrm{n}}}$$

$$\text{Repression}: \quad F_i(p_j)=\frac{K}{K+p_j^{\mathrm{n}}}$$

$$i,\ j=1,2$$

$$\frac{dq_1}{dt}=k_1^{\mathrm{Q}}F_1(p_2)-d_1^{\mathrm{Q}}q_1$$

$$\frac{dq_2}{dt}=k_2^{\mathrm{Q}}F_2(p_1)-d_2^{\mathrm{Q}}q_2$$

$$\frac{dp_1}{dt}=k_1^{\mathrm{P}}q_1-d_2^{\mathrm{P}}p_1$$

$$\frac{dp_2}{dt}=k_2^{\mathrm{P}}q_2-d_2^{\mathrm{P}}p_2$$

Stationary points: $\bar{p}_1-\vartheta_1F_1(\vartheta_2F_2(\bar{p}_1))=0,\ \bar{p}_2=\vartheta_2F_2(\bar{p}_1)$

$$\vartheta_1=\frac{k_1^{\mathrm{Q}}k_1^{\mathrm{P}}}{d_1^{\mathrm{Q}}d_1^{\mathrm{P}}},\ \vartheta_2=\frac{k_2^{\mathrm{Q}}k_2^{\mathrm{P}}}{d_2^{\mathrm{Q}}d_2^{\mathrm{P}}}$$

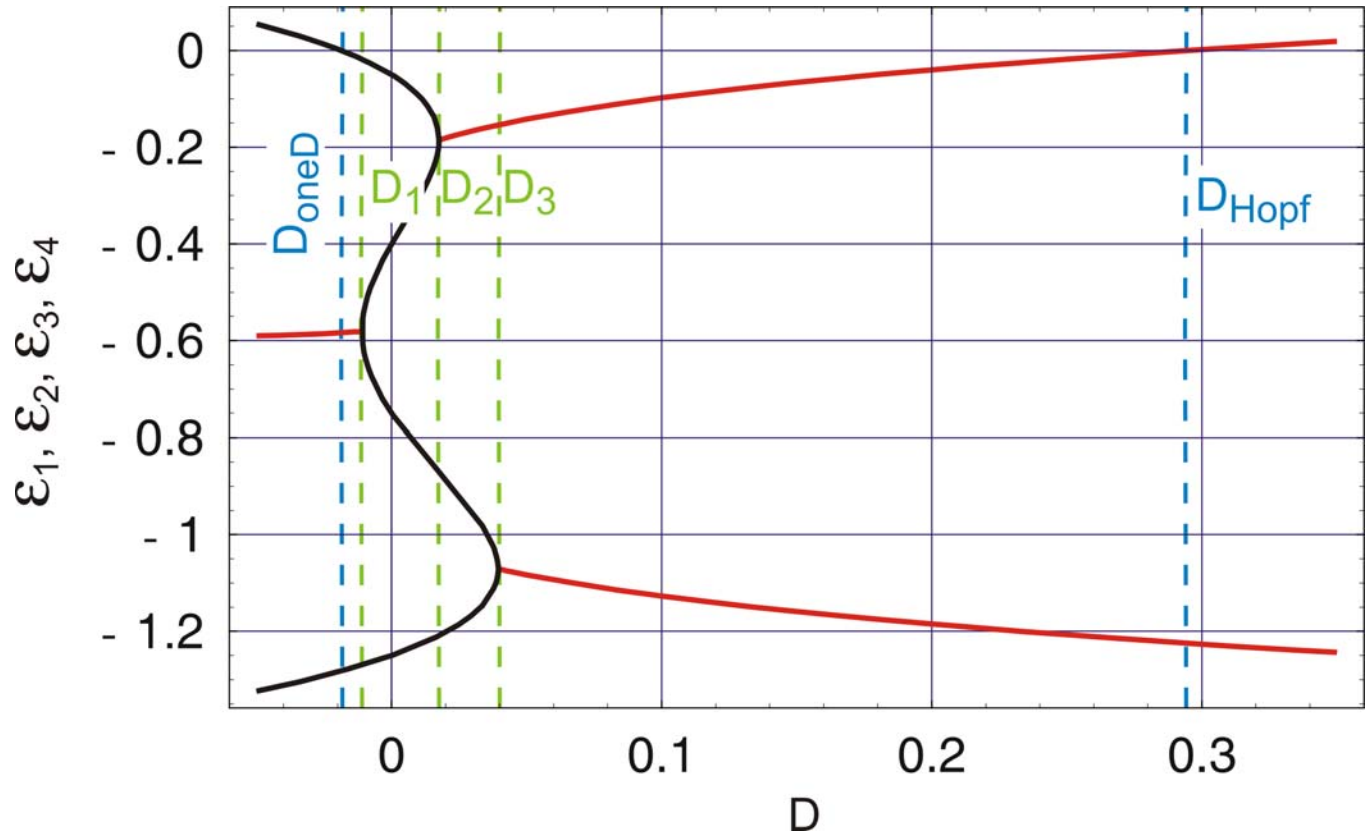
Qualitative analysis of **cross-regulation** of two genes: Stationary points

$$\mathrm{A} = \left\{ a_{ij} = \frac{\partial \dot{x}_i}{\partial x_j} \right\} = \begin{pmatrix} -d_1^Q & 0 & k_1^Q \frac{\partial F_1}{\partial p_1} & k_1^Q \frac{\partial F_1}{\partial p_2} \\ 0 & -d_2^Q & k_2^Q \frac{\partial F_2}{\partial p_1} & k_2^Q \frac{\partial F_2}{\partial p_2} \\ k_1^P & 0 & -d_1^P & 0 \\ 0 & k_2^P & 0 & -d_2^P \end{pmatrix}$$

$$\text{Cross regulation}: \quad \frac{\partial F_1}{\partial p_1} = \frac{\partial F_2}{\partial p_2} = 0$$

$$\left|\mathrm{A} - \varepsilon \mathrm{I}\right| = \begin{vmatrix} -d_1^Q - \varepsilon & 0 & 0 & k_1^Q \frac{\partial F_1}{\partial p_2} \\ 0 & -d_2^Q - \varepsilon & k_2^Q \frac{\partial F_2}{\partial p_1} & 0 \\ k_1^P & 0 & -d_1^P - \varepsilon & 0 \\ 0 & k_2^P & 0 & -d_2^P - \varepsilon \end{vmatrix} = \begin{vmatrix} Q_D & Q_K \\ P_D & P_K \end{vmatrix}$$

Qualitative analysis of **cross-regulation** of two genes: Jacobian matrix



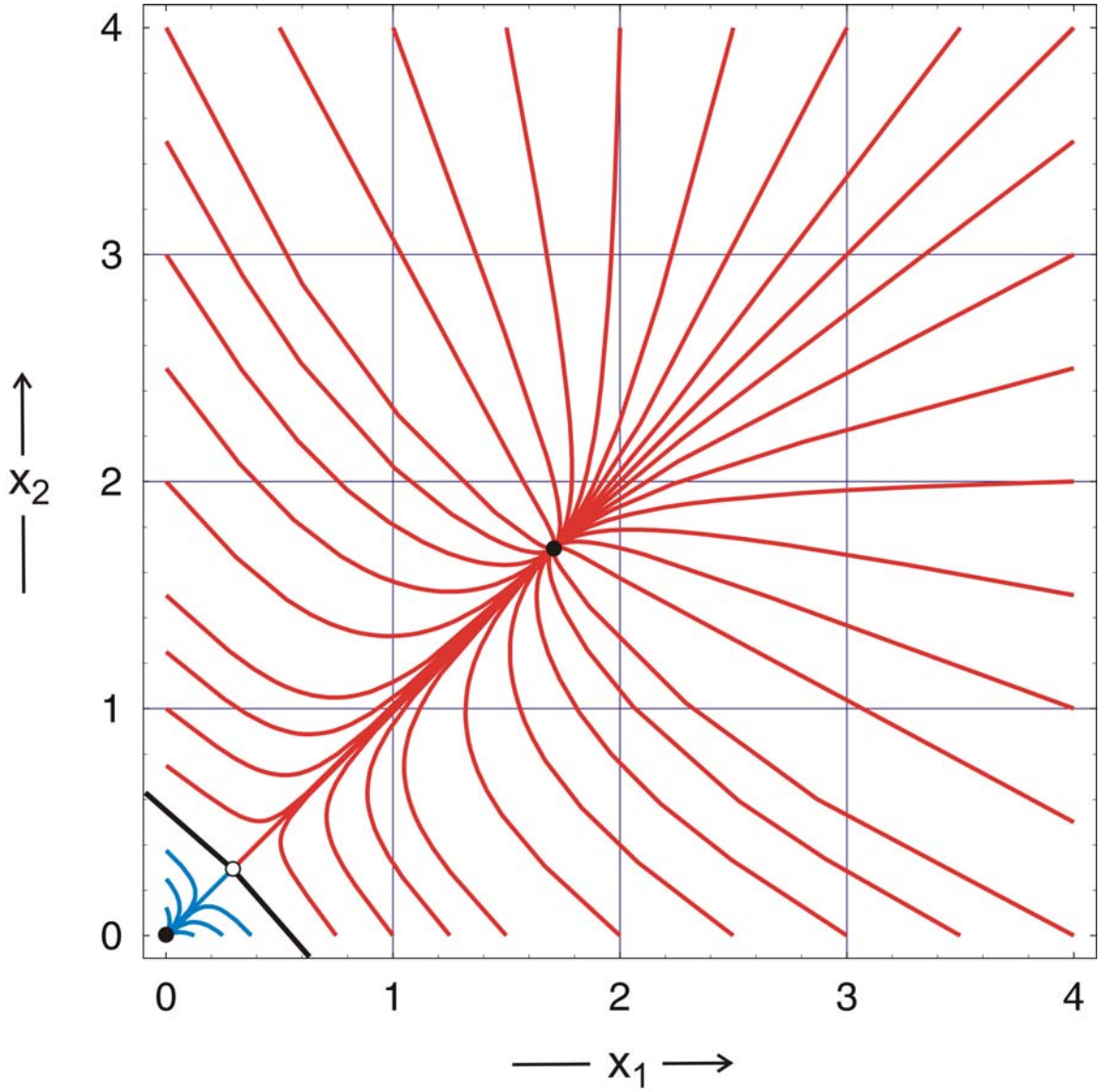
$$D_{\mathrm{OneD}} = -d_1^{\mathrm{Q}} d_2^{\mathrm{Q}} d_1^{\mathrm{P}} d_2^{\mathrm{P}}$$

$$D_{\mathrm{Hopf}} = \frac{(d_1^{\mathrm{Q}}+d_2^{\mathrm{Q}})(d_1^{\mathrm{Q}}+d_1^{\mathrm{P}})(d_1^{\mathrm{Q}}+d_2^{\mathrm{P}})(d_2^{\mathrm{Q}}+d_1^{\mathrm{P}})(d_2^{\mathrm{Q}}+d_2^{\mathrm{P}})(d_1^{\mathrm{P}}+d_2^{\mathrm{P}})}{(d_1^{\mathrm{Q}}+d_2^{\mathrm{Q}}+d_1^{\mathrm{P}}+d_2^{\mathrm{P}})^2}$$

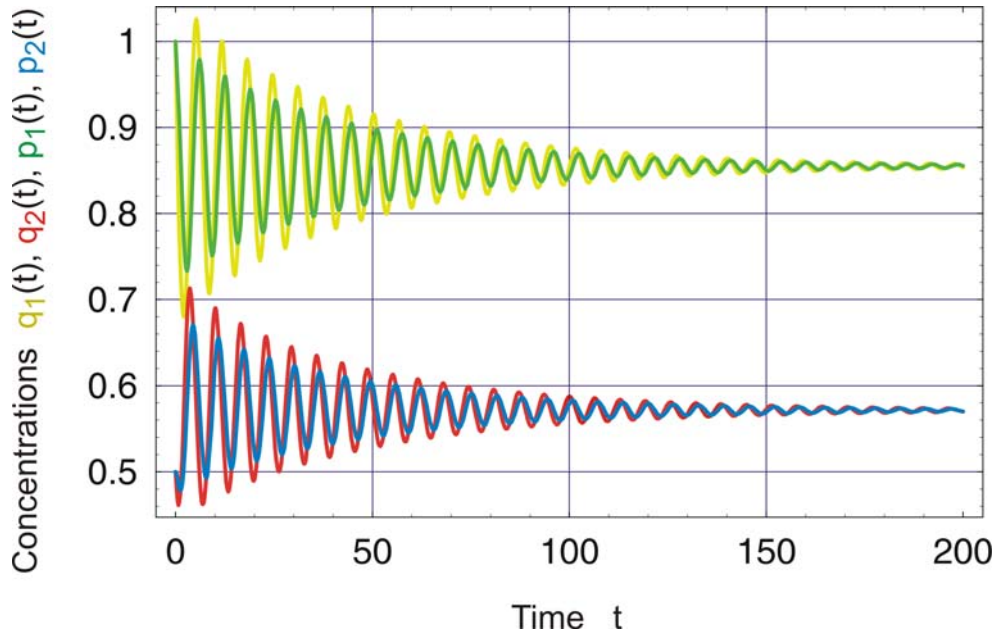
two stable states

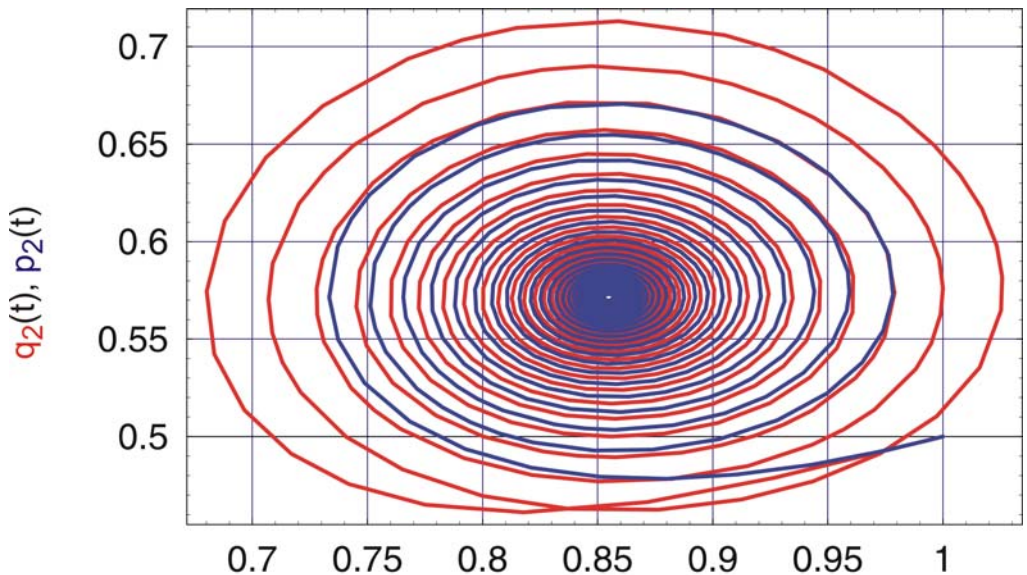
E: both genes off

P: both genes on

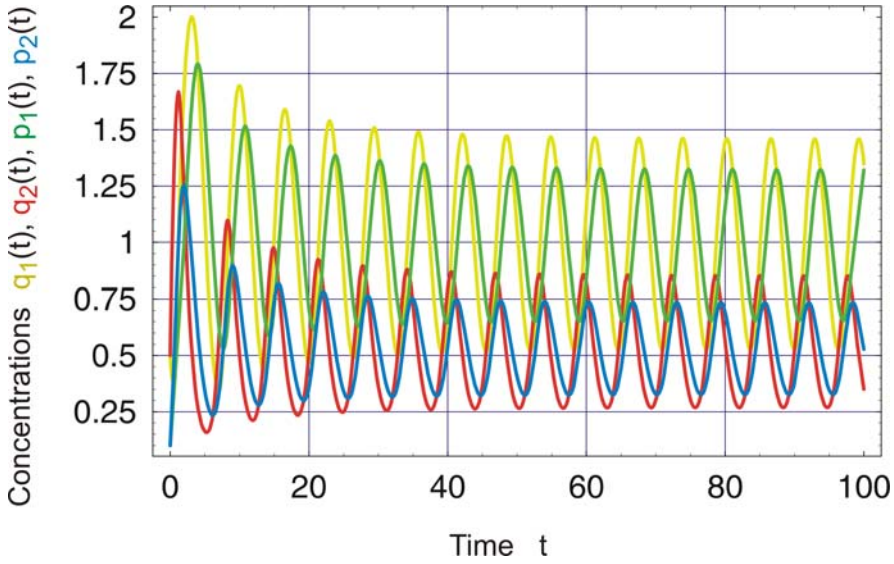


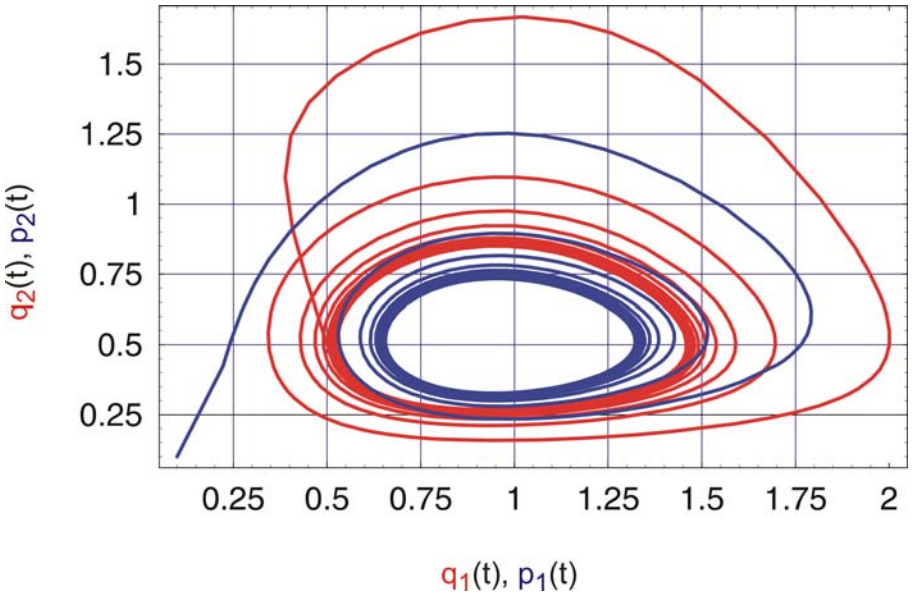
Simplified two gene system (x_1,x_2): act2-act2

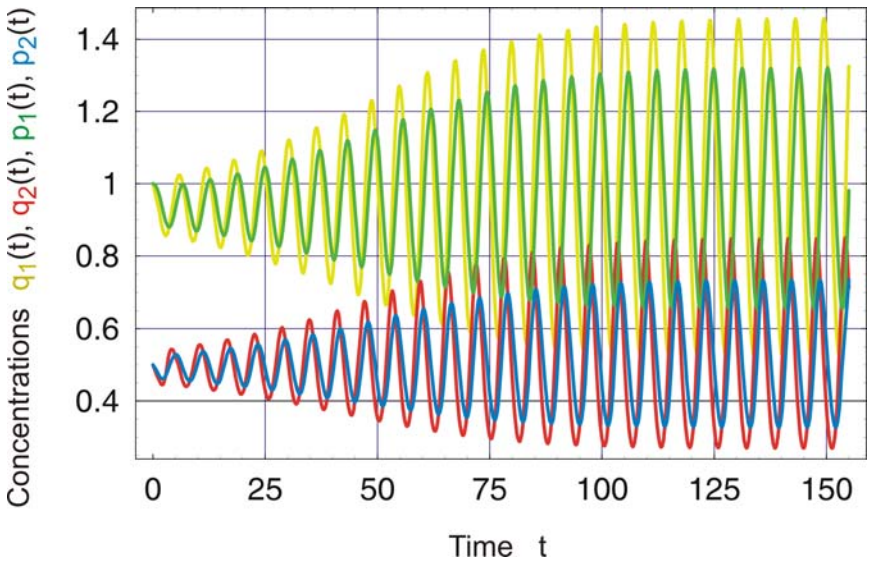


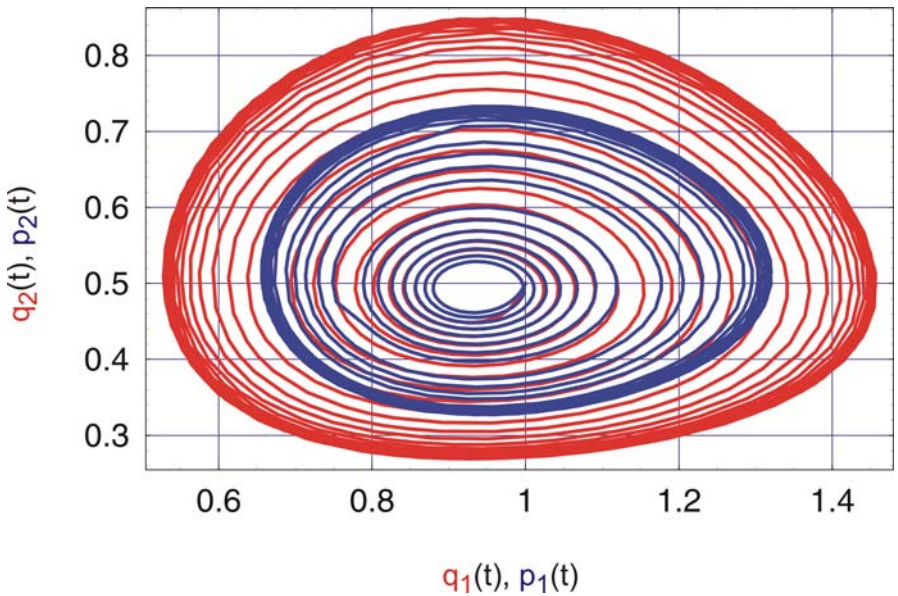


Regulatory dynamics at $D < D_{Hopf}$, act.-repr., n=3







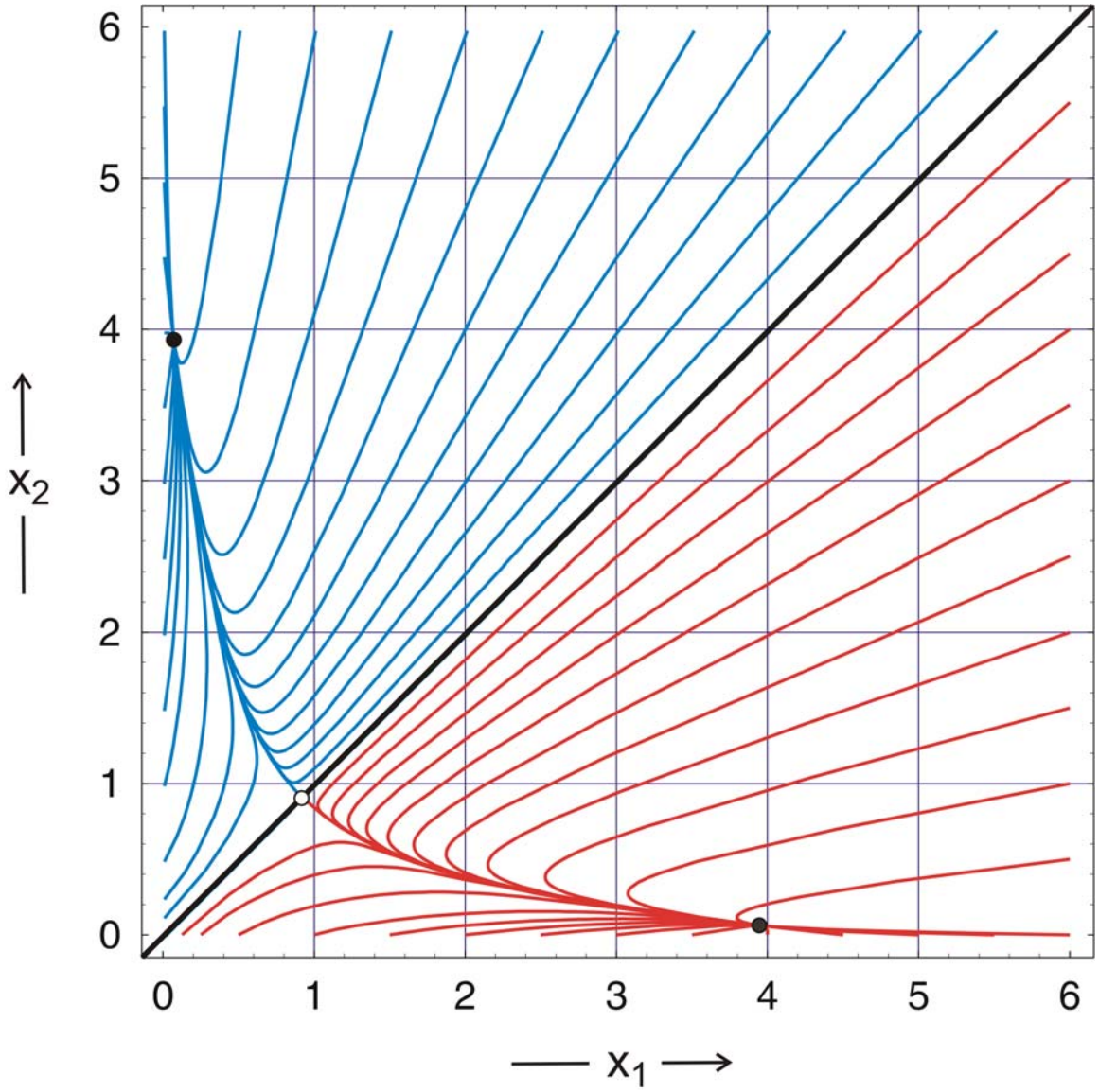


Regulatory dynamics at $D > D_{Hopf}$, act.-repr., n=3

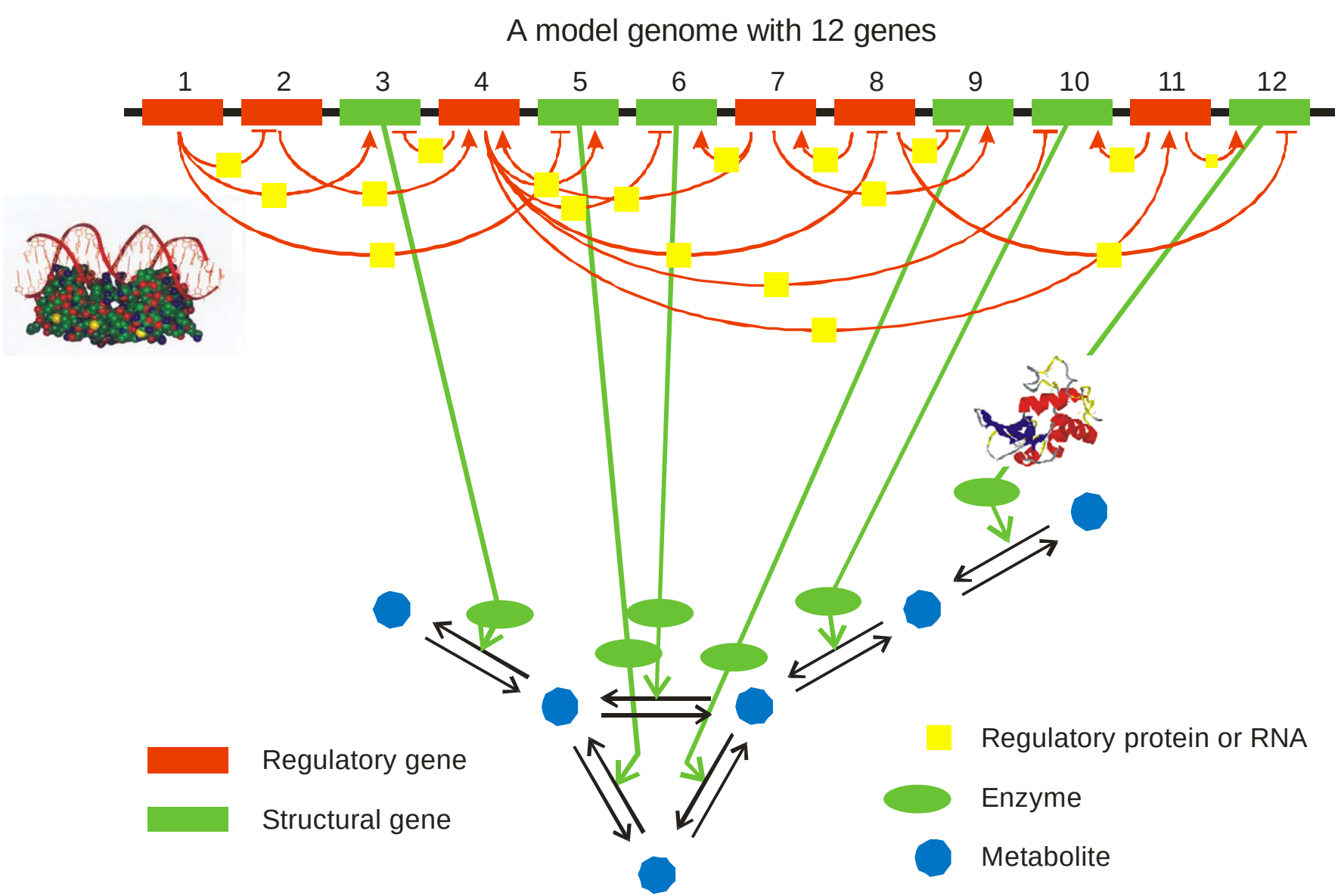
two stable states

P_1: gene 1 on, gene 2 off

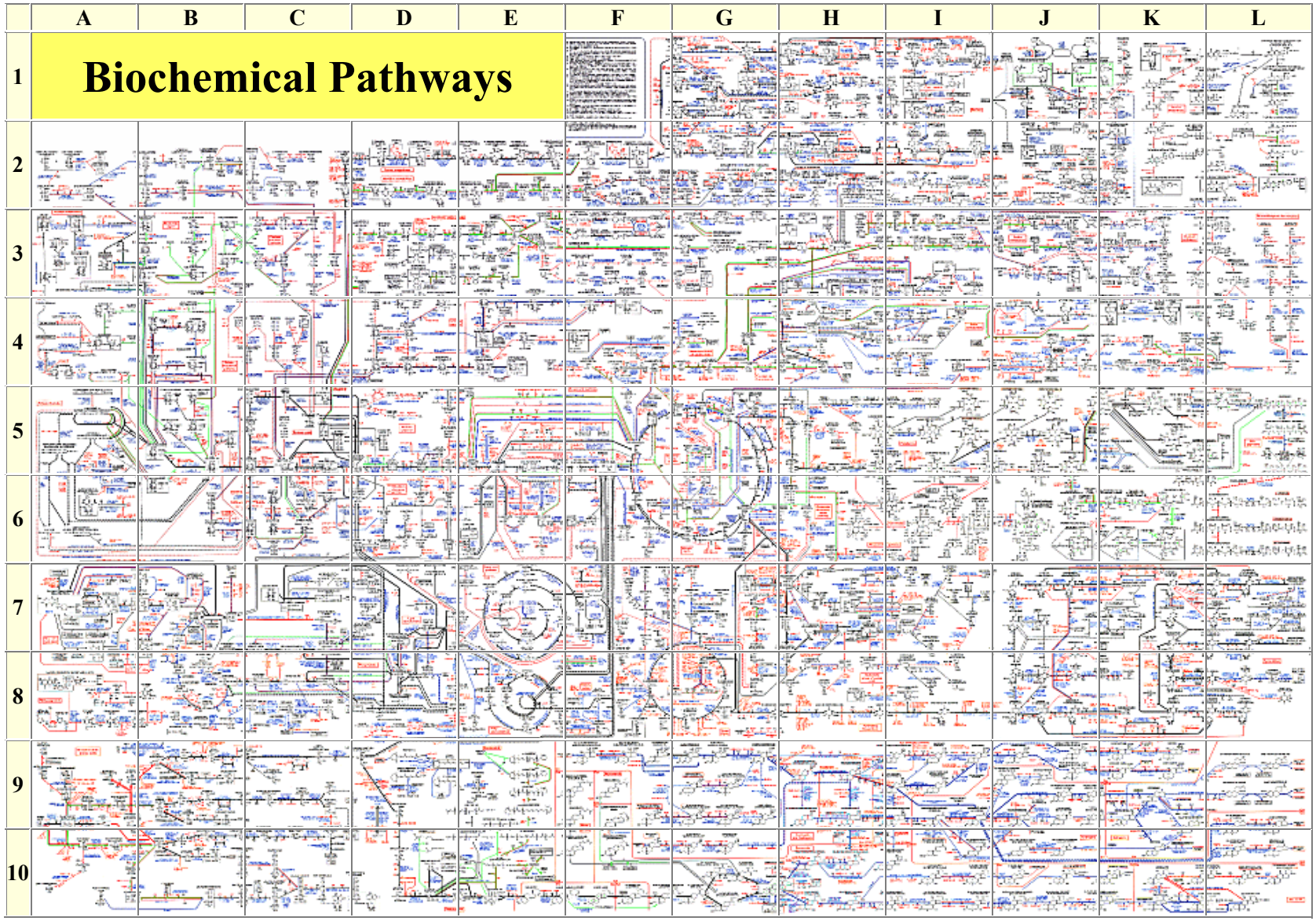
P_2: gene 1 off, gene 2 on



Simplified two gene system (x_1,x_2): rep2-rep2

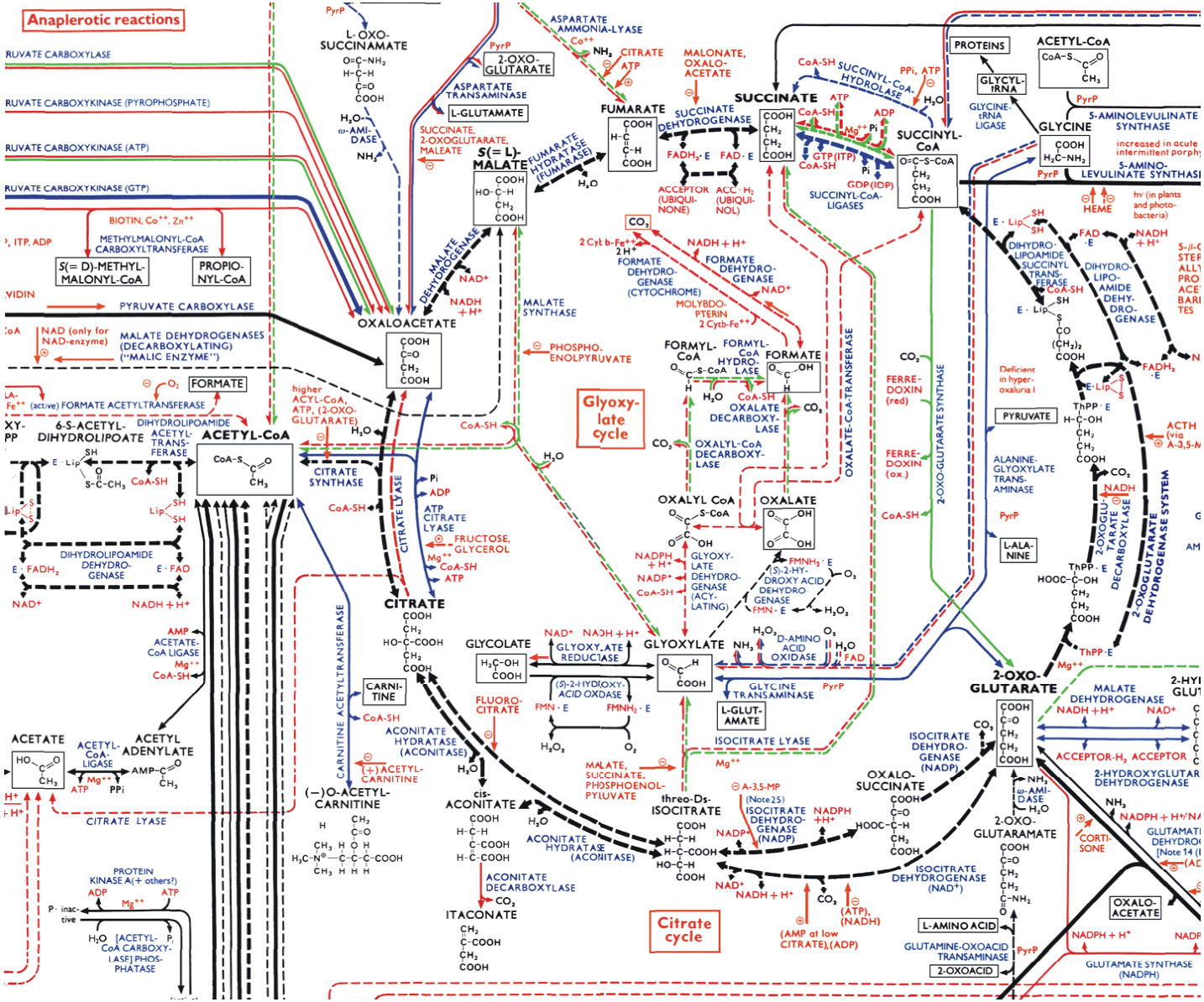


Sketch of a genetic and metabolic network



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

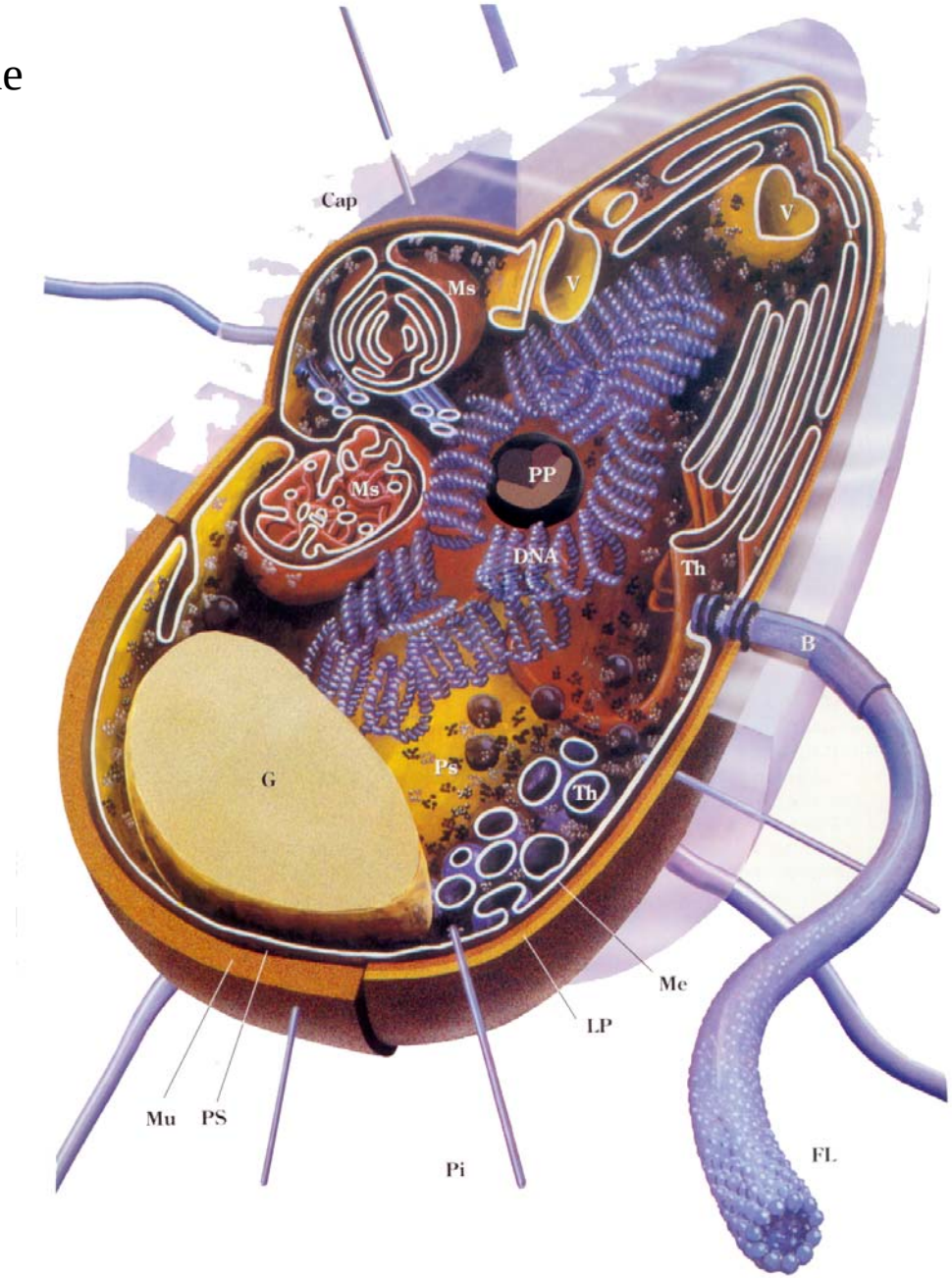
The citric acid or Krebs cycle (enlarged from previous slide).



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

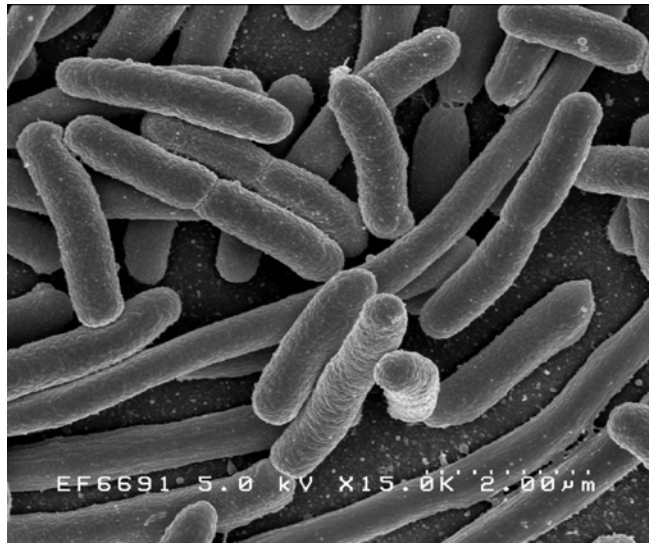
The human body:

10^{14} cells = 10^{13} eukaryotic cells +
$\approx 9\times10^{13}$ bacterial (prokaryotic) cells,
and $\approx$ 200 eukaryotic cell types



The spatial structure of the bacterium *Escherichia coli*

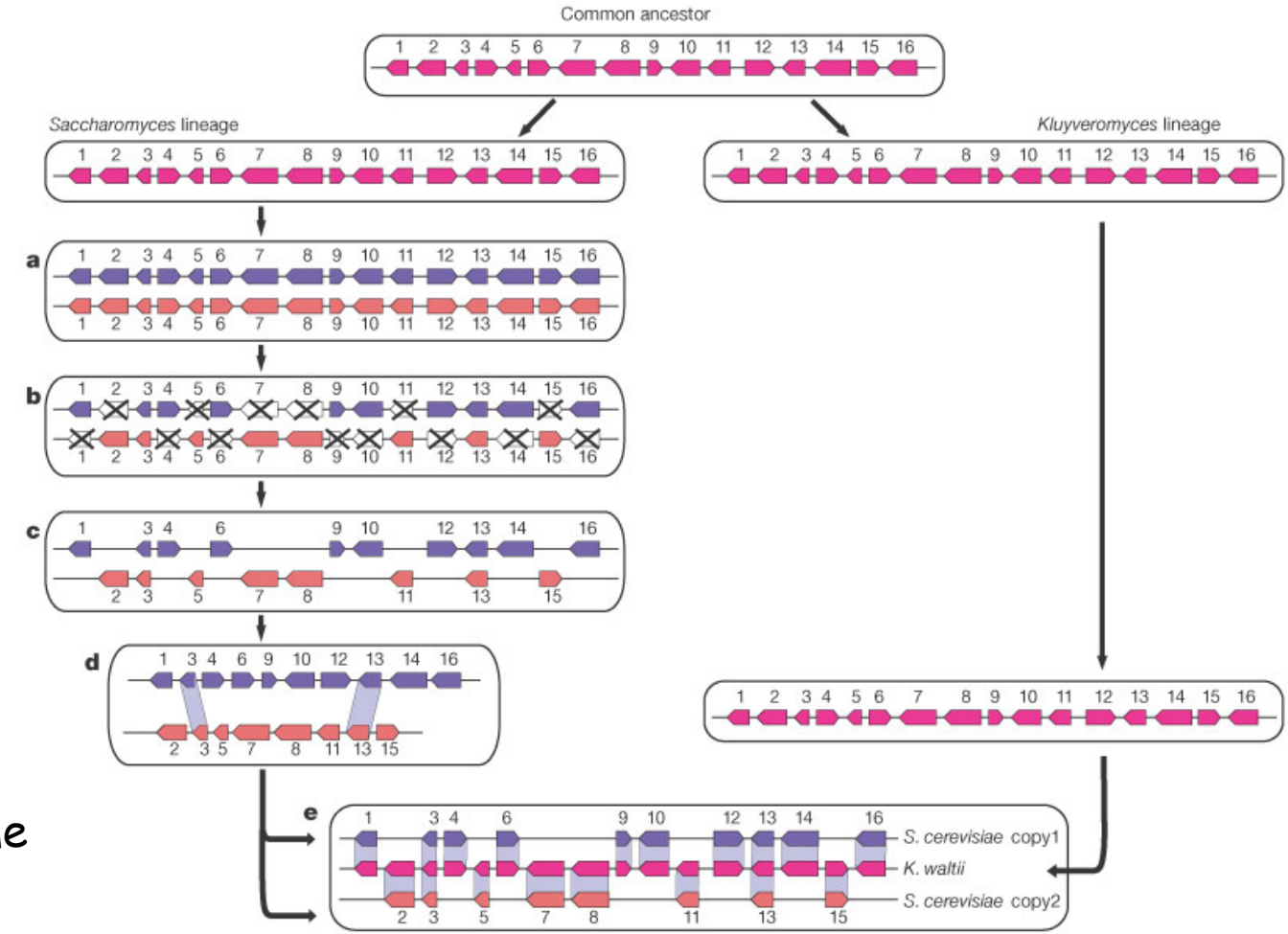
E. coli:	Genome length	4×10^6 nucleotides
	Number of cell types	1
	Number of genes	4 460



Man:	Genome length	3×10^9 nucleotides
	Number of cell types	200
	Number of genes	≈ 30 000

Complexity in biology

A model for a genome duplication in yeast, $\approx 1 \times 10^8$ Jahren ago



Manolis Kellis, Bruce W. Birren, and Eric S. Lander. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* **428**: 617-624, 2004

The difficulty to define the notion of „gene".

Helen Pearson,
Nature **441**: 399-401, 2006

WHAT IS A GENE?

The idea of genes as beads on a DNA string is fast fading. Protein-coding sequences have no clear beginning or end and RNA is a key part of the information package, reports **Helen Pearson**.

'Gene' is not a typical four-letter word. It is not offensive. It is never bleeped out of TV shows. And where the meaning of most four-letter words is all too clear, that of gene is not. The more expert scientists become in molecular genetics, the less easy it is to be sure about what, if anything, a gene actually is.

Rick Young, a geneticist at the Whitehead Institute in Cambridge, Massachusetts, says that when he first started teaching as a young professor two decades ago, it took him about two hours to teach fresh-faced undergraduates what a gene was and the nuts and bolts of how it worked. Today, he and his colleagues need three months of lectures to convey the concept of the gene, and that's not because the students are any less bright. "It takes a whole semester to teach this stuff to talented graduates," Young says. "It used to be we could give a one-off definition and now it's much more complicated."

In classical genetics, a gene was an abstract concept — a unit of inheritance that ferried a characteristic from parent to child. As biochemistry came into its own, those characteristics were associated with enzymes or proteins, one for each gene. And with the advent of molecular biology, genes became real, physical things — sequences of DNA which when converted into strands of so-called messenger RNA could be used as the basis for building their associated protein piece by piece. The great coiled DNA molecules of the chromosomes were seen as long strings on which gene sequences sat like discrete beads.

This picture is still the working model for many scientists. But those at the forefront of genetic research see it as increasingly old-fashioned — a crude approximation that, at best, hides fascinating new complexities and, at worst, blinds its users to useful new paths of enquiry.

Information, it seems, is parceled out along chromosomes in a much more complex way than was originally supposed. RNA molecules are not just passive conduits through which the gene's message flows into the world but active regulators of cellular processes. In some cases, RNA may even pass information across generations — normally the sole preserve of DNA.

An eye-opening study last year raised the possibility that plants sometimes rewrite their DNA on the basis of RNA messages inherited from generations past[1]. A study on page 469 of this issue suggests that a comparable phenomenon might occur in mice, and by implication in other mammals[2]. If this type of phenomenon is indeed widespread, it "would have huge implications," says evolutionary geneticist Laurence Hurst at the University of Bath, UK.

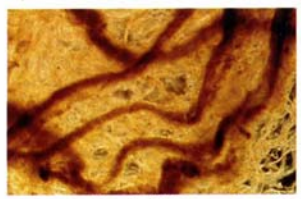
"All of that information seriously challenges our conventional definition of a gene," says molecular biologist Bing Ren at the University of California, San Diego. And the information challenge is about to get even tougher. Later this year, a glut of data will be released from the international Encyclopedia of DNA Elements (ENCODE) project. The pilot phase of ENCODE involves scrutinizing roughly 1% of the human genome in unprecedented detail; the aim is to find all the sequences that serve a useful purpose and explain what that purpose is. "When we started the ENCODE project I had a different view of what a gene was," says contributing researcher Roderic Guigo at the Center for Genomic Regulation in Barcelona. "The degree of complexity we've seen was not anticipated."

"We've come to the realization that the genome is full of overlapping transcripts."
— Phillip Kapranov

Under fire

The first of the complexities to challenge molecular biology's paradigm of a single DNA sequence encoding a single protein was alternative splicing, discovered in viruses in 1977 (see 'Hard to track', overleaf). Most of the DNA sequences describing proteins in humans have a modular arrangement in which exons, which carry the instructions for making proteins, are interspersed with non-coding introns. In alternative splicing, the cell snips out introns and sews together the exons in various different orders, creating messages that can code for different proteins. Over the years geneticists have also documented overlapping genes, genes within genes and countless other weird arrangements (see 'Muddling over genes', overleaf).

Alternative splicing, however, did not in itself require a drastic reappraisal of the notion of a gene; it just showed that some DNA sequences could describe more than one protein. Today's assault on the gene concept is more far reaching, fuelled largely by studies that show the previously unimagined scope of RNA.



Spools of DNA (above) still harbour surprises, with one protein-coding gene often overlapping the next.

P. PLAILLY/SPL

The one gene, one protein idea is coming under particular assault from researchers who are comprehensively extracting and analysing the RNA messages, or transcripts, manufactured by genomes, including the human and mouse genome. Researchers led by Thomas Gingeras at the company Affymetrix in Santa Clara, California, for example, recently studied all the transcripts from ten chromosomes across eight human cell lines and worked out precisely where on the chromosomes each of the transcripts came from[3].

The picture these studies paint is one of mind-boggling complexity. Instead of discrete genes dutifully mass-producing identical RNA transcripts, a teeming mass of transcription converts many segments of the genome into multiple RNA ribbons of differing lengths. These ribbons can be generated from both strands of DNA, rather than from just one as was conventionally thought. Some of these transcripts come from regions of DNA previously identified as holding protein-coding genes. But many do not. "It's somewhat revolutionary," says Gingeras's colleague Phillip Kapranov. "We've come to the realization that the genome is full of overlapping transcripts."

Other studies, one by Guigo's team[4], and one by geneticist Rotem Sorek[5], now at Tel Aviv University, Israel, and his colleagues, have hinted at the reasons behind the mass of transcription. The two teams investigated occasional reports that transcription can start at a DNA sequence associated with one protein and run straight through into the gene for a completely different protein, producing a fused transcript. By delving into databases of human RNA transcripts, Guigo's team estimate that 4–5% of the DNA in regions conventionally recognized as genes is transcribed in this way. Producing fused transcripts could be one way for a cell to generate a greater variety of proteins from a limited number of exons, the researchers say.

Many scientists are now starting to think that the descriptions of proteins encoded in DNA know no borders — that each sequence reaches into the next and beyond. This idea will be one of the central points to emerge from the ENCODE project when its results are published later this year.

Kapranov and others say that they have documented many examples of transcripts in which protein-coding exons from one part of the genome combine with exons from another

ENCODE stands for
ENCyclopedia Of DNA Elements.

ENCODE Project Consortium.
Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.
Nature **447**:799-816,2007

Acknowledgement of support

Fonds zur Förderung der wissenschaftlichen Forschung (FWF)
Projects No. 09942, 10578, 11065, 13093
13887, and 14898

Wiener Wissenschafts-, Forschungs- und Technologiefonds (WWTF)
Project No. Mat05

Jubiläumsfonds der Österreichischen Nationalbank
Project No. Nat-7813

European Commission: Contracts No. 98-0189, 12835 (NEST)

Austrian Genome Research Program – GEN-AU: Bioinformatics Network (BIN)

Österreichische Akademie der Wissenschaften

Siemens AG, Austria

Universität Wien and the Santa Fe Institute

Universität Wien

Coworkers

Universität Wien

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

Paul E. Phillipson, University of Colorado at Boulder, CO

Heinz Engl, Philipp Kügler, **James Lu**, **Stefan Müller**, RICAM Linz, AT

Jord Nagel, **Kees Pleij**, Universiteit Leiden, NL

Walter Fontana, Harvard Medical School, MA

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

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

Ivo L.Hofacker, **Christoph Flamm**, **Andreas Svrček-Seiler**, Universität Wien, AT

Kurt Grünberger, Michael Kospach , **Andreas Wernitznig**, **Stefanie Widder, Stefan Wuchty**, Universität Wien, AT

Jan Cupal, **Stefan Bernhart**, **Lukas Endler, Ulrike Langhammer**, **Rainer Machne, Ulrike Mückstein**, **Hakim Tafer, Thomas Taylor,** Universität Wien, AT

Web-Page for further information:

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