Boltzmann and Evolution

Basic Questions of Biology seen with Atomistic Glasses

Peter Schuster

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

Boltzmann‘s Legacy

Erwin Schrödinger Institute, Wien, 07.– 09.06.2006

Web-Page for further information:

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

… Wenn Sie mich nach meiner innersten Überzeugung fragen ob man es (das 19. Jahrhundert) einmal das eiserne Jahrhundert oder das Jahrhundert des Dampfes oder der Elektrizität nennen wird, so antworte ich ohne Bedenken, das Jahrhundert der ***mechanischen Naturauffassung****, das* ***Jahrhundert Darwins*** *wird es heißen.*

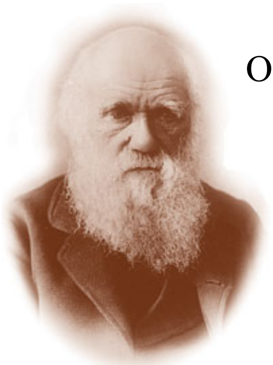
Ludwig Boltzmann, *Der zweite Hauptsatz der mechanischen Wärmetheorie.* Vortrag, gehalten in feierlichen Sitzung der Kaiserlichen Akademie der Wissenschaften am 29. Mai 1886.

Three necessary conditions for Darwinian evolution are:

1. **Multiplication,**
2. **Variation**, and
3. **Selection**.

Variation through mutation and recombination operates on the **genotype** whereas the **phenotype** is the target of **selection**.

One important property of the Darwinian scenario is that **variations** in the form of mutations or recombination events occur **uncorrelated** with their **effects on the selection process**.

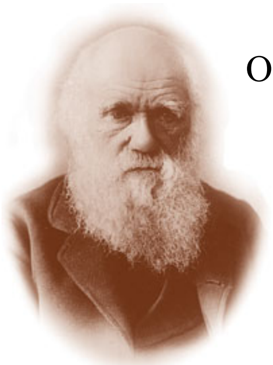


Origin of evolutionary biology
1859

Origin of genetics
1865

Charles Darwin

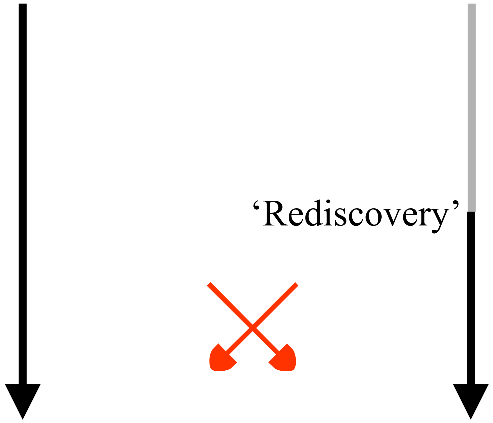
Gregor Mendel



Charles Darwin

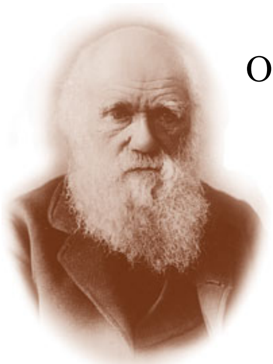
Origin of evolutionary biology
1859

Origin of genetics
1865



‘Rediscovery’ 1900

Gregor Mendel

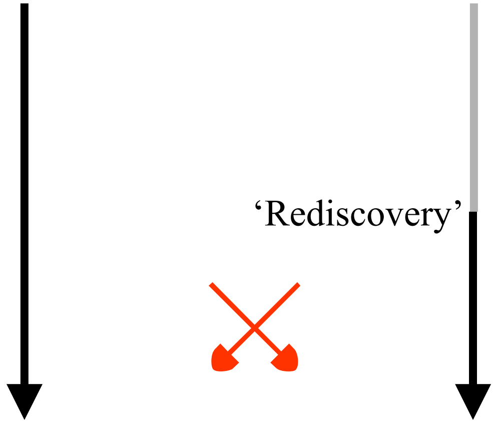


Charles Darwin

Origin of evolutionary biology
1859

Origin of genetics
1865

Gregor Mendel



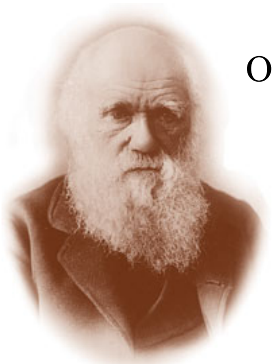
‘Rediscovery’ 1900

First unification: Population genetics 1930

Ronald Fisher

JSB Haldane

Sewall Wright



Charles Darwin

Origin of evolutionary biology
1859

Origin of genetics
1865

Gregor Mendel

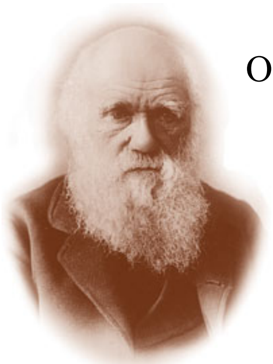
‘Rediscovery’ 1900

First unification: Population genetics 1930

Ernst Mayr

Synthetic or
Neo-Darwinian theory
1940 - 1950

Theodosius
Dobzhansky



Charles Darwin

Origin of evolutionary biology
1859

Origin of genetics
1865

‘Rediscovery’ 1900

Gregor Mendel

Friedrich Woehler

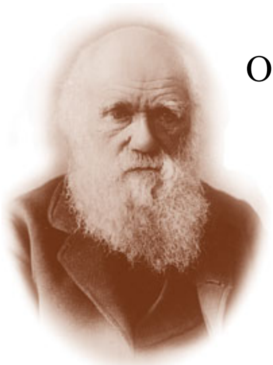
Origin of
biochemistry
1828

First unification: Population genetics 1930

Ernst Mayr

Synthetic or
Neo-Darwinian theory
1940 - 1950

Theodosius
Dobzhansky



Charles Darwin

Origin of evolutionary biology
1859

Origin of genetics
1865

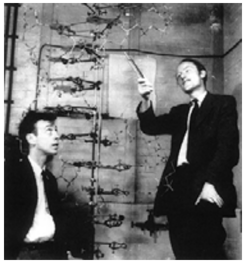
‘Rediscovery’ 1900

Gregor Mendel

Friedrich Woehler

Origin of biochemistry
1828

Origin of molecular biology 1953



James Watson and Francis Crick

First unification: Population genetics 1930

Ernst Mayr

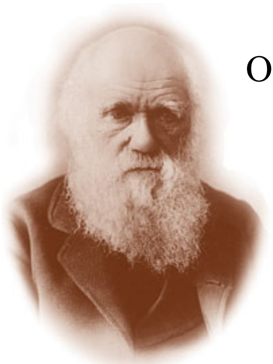
Synthetic or
Neo-Darwinian theory
1940 - 1950

Theodosius Dobzhansky

Max Perutz

John Kendrew

Biology of the 21st century



Charles Darwin

Origin of evolutionary biology
1859

Origin of genetics
1865

‘Rediscovery’ 1900

Gregor Mendel

Friedrich Woehler

Origin of
biochemistry
1828

Origin of molecular
biology 1953

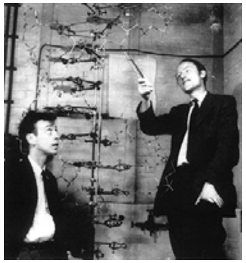
First unification: Population genetics 1930

Ernst Mayr

Synthetic or
Neo-Darwinian theory
1940 - 1950

Theodosius
Dobzhansky

Jacques Monod



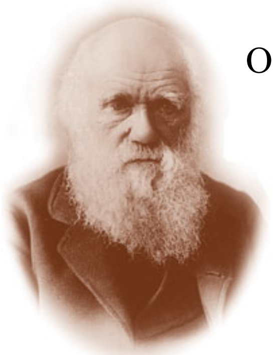
James Watson and
Francis Crick

François
Jacob

Max Perutz

John Kendrew

Biology of the 21st century



Charles Darwin

Origin of evolutionary biology
1859

Origin of genetics
1865

‘Rediscovery’ 1900

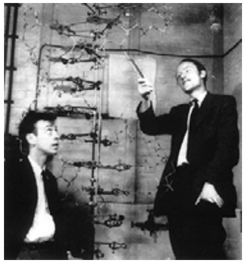
Gregor Mendel

Friedrich Woehler

Origin of biochemistry
1828

First unification: Population genetics 1930

Origin of molecular biology 1953



James Watson and Francis Crick

Jacques Monod

Ernst Mayr

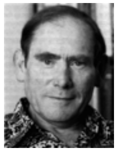
Synthetic or
Neo-Darwinian theory
1940 - 1950

Theodosius Dobzhansky

François Jacob

Max Perutz

Manfred Eigen



Sydney Brenner

John Kendrew

Biology of the 21st century

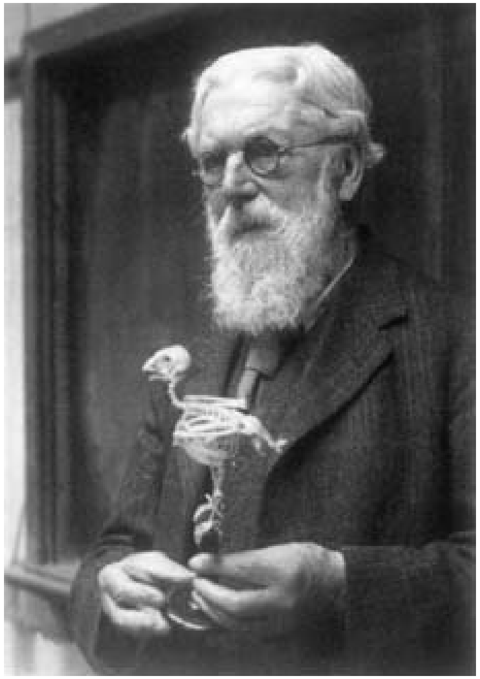
Biomathematics, bioinformatics, … , biophysics, biochemistry, … , molecular genetics, … , systems biology, biomedicine, macroscopic biology, evolutionary biology, sociobiology, anthropology, …

… Der allgemeine Daseinskampf der Lebewesen ist daher nicht ein Kampf um die Grundstoffe – die Grundstoffe aller Organisman sind in Luft, Wasser und Erdboden im Überflusse vorhanden – auch nicht um **Energie**, welche **in Form von Wärme** leider **unverwandelbar** in jedem Körper reichlich vorhanden **ist**, sondern ein **Kampf um die Entropie**, **welche durch den Übergang der Energie von der heißen Sonne zur kalten Erde disponibel wird**. Diesen Übergang möglichst auszunutzen, breiten die Pflanzen die unermeßliche Fläche ihrer Blätter aus und **zwingen die Sonnenenergie** in noch unerforschter Weise, ehe sie auf das Temperaturniveau der Erdoberfläche herabsinkt, **chemische Synthesen auszuführen**, von denen man in unseren Laboratorien noch keine Ahnung hat. …

Ludwig Boltzmann, *Der zweite Hauptsatz der mechanischen Wärmetheorie.* Vortrag, gehalten in feierlichen Sitzung der Kaiserlichen Akademie der Wissenschaften am 29.Mai 1886.

Available energy (free energy) is the main object at stake in the struggle for existence and the evolution of the world.

Quoted in D'Arcy W. Thompson. *On Growth and Form***,** Cambridge (UK), 1917.



Nothing in biology makes sense except in the light of evolution.

Theodosius Dobzhansky, 1973.

Genotype, Genome

Collection of genes

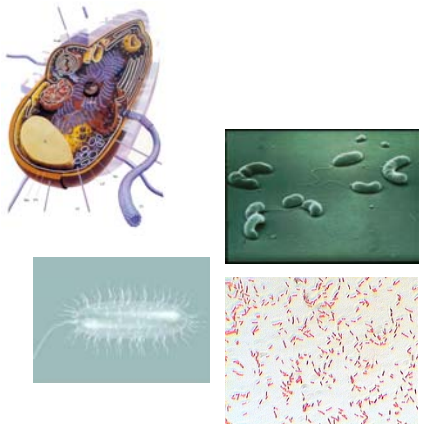
Developmental program

Unfolding of the genotype

Highly specific environmental conditions

Phenotype

Evolution explains the origin of species and their interactions



The holism versus reductionism debate

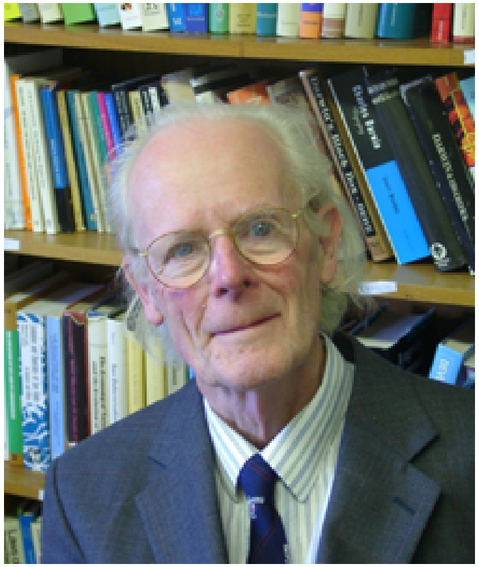
The holistic approach

Macroscopic biologists aim at a top-down approach to describe the phenomena observed in biology.

↔

The reductionists' program

Molecular biologist perform a bottom-up approach to interpret biological phenomena by the methods of chemistry and physics.



As I happens, I do not understand how modern sewing-machines work, but this does not lead me suppose that the laws of topology have been broken: Indeed, I feel confident I could find out if someone would let me take one to pieces.

Molecular biologists are quite right to disbelieve in (any kind of) *elán vital*.

John Maynard Smith, The problems of biology.
Oxford University Press, 1986.

What should be the attitude of a biologist working on whole organisms to molecular biology? It is, I think, foolish to argue that we (the macroscopic biologists) are discovering things that disprove molecular biology. It would be more sensible to say to molecular biologists that **there are phenomena that they will one day have to interpret in their terms**.

John Maynard Smith, The problems of biology.
Oxford University Press, 1986.

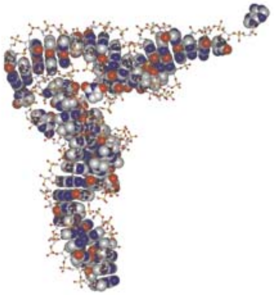
Genotype, Genome

Genetic information

GCGGATTTAGCTCAGTTGGGAGAGCGCCAGACTGAAGATCTGGAGGTCCTGTGTTCGATCCACAGAATTCGCACCA

Omics

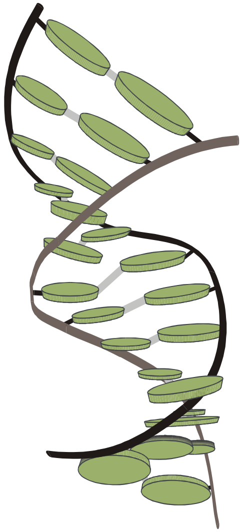
'The new biology is the chemistry of living matter'



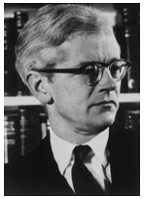
Biochemistry
molecular biology
structural biology
molecular evolution
molecular genetics
systems biology
bioinfomatics

Unfolding of the genotype

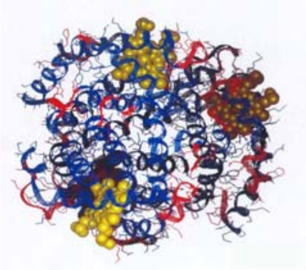
Highly specific environmental conditions



Phenotype



John Kendrew



Evolution of RNA molecules,
ribozymes and splicing,
the idea of an RNA world,
selection of RNA molecules,
RNA editing,
the ribosome is a ribozyme,
small RNAs and RNA switches.

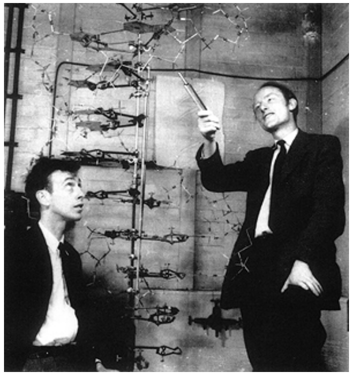
The exciting RNA story

Manfred Eigen

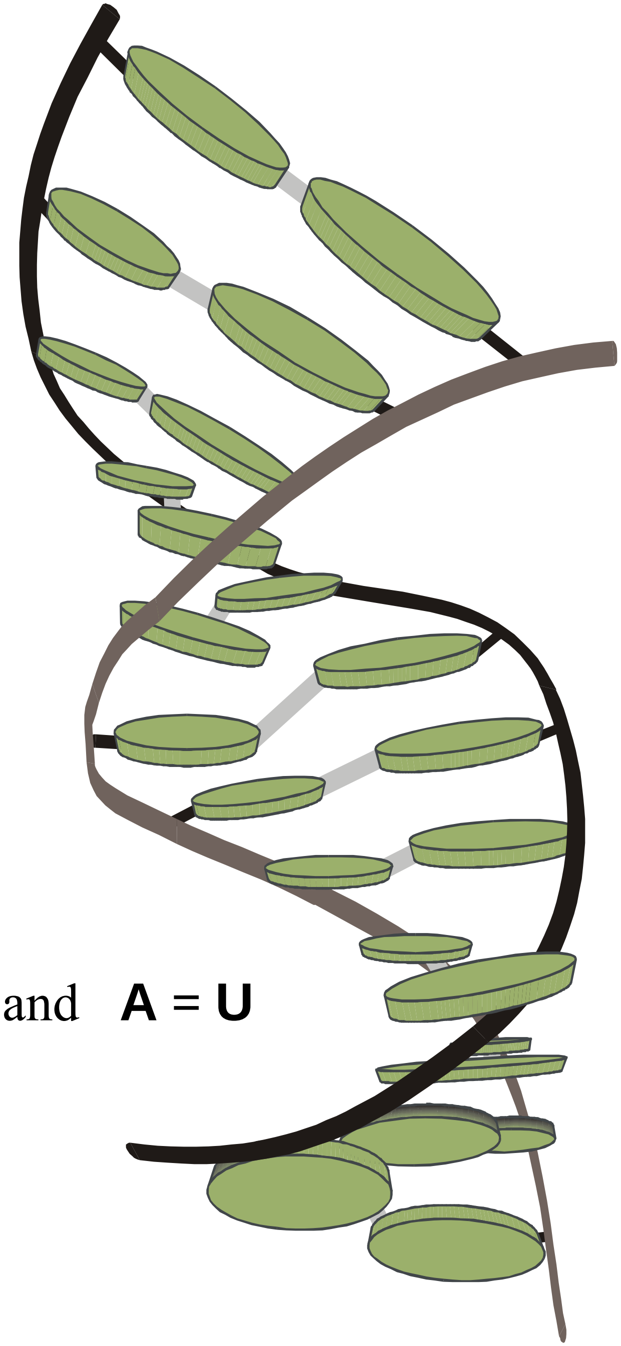
Molecular evolution
Linus Pauling and
Emile Zuckerkandl

Hemoglobin sequence
Gerhard Braunitzer

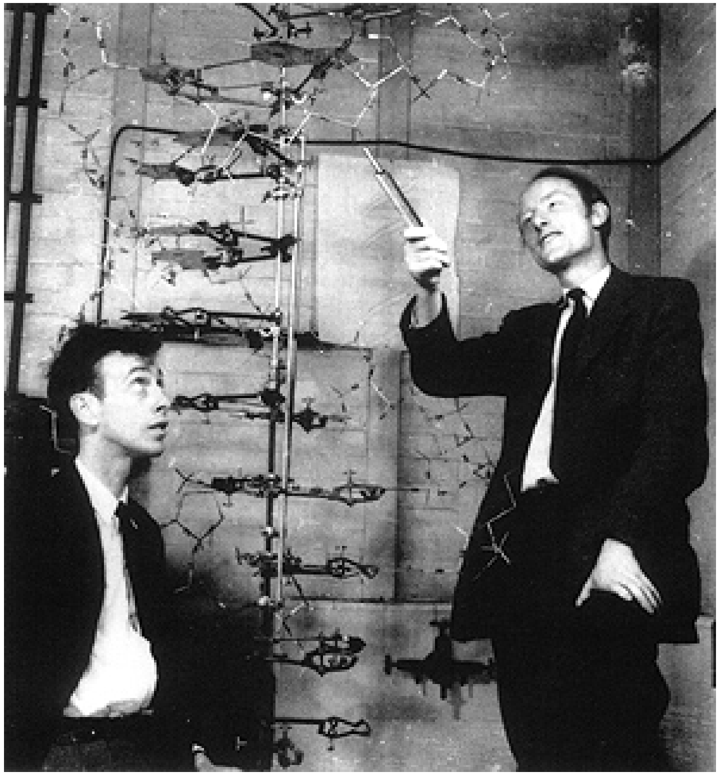
Max Perutz



James D. Watson und
Francis H.C. Crick

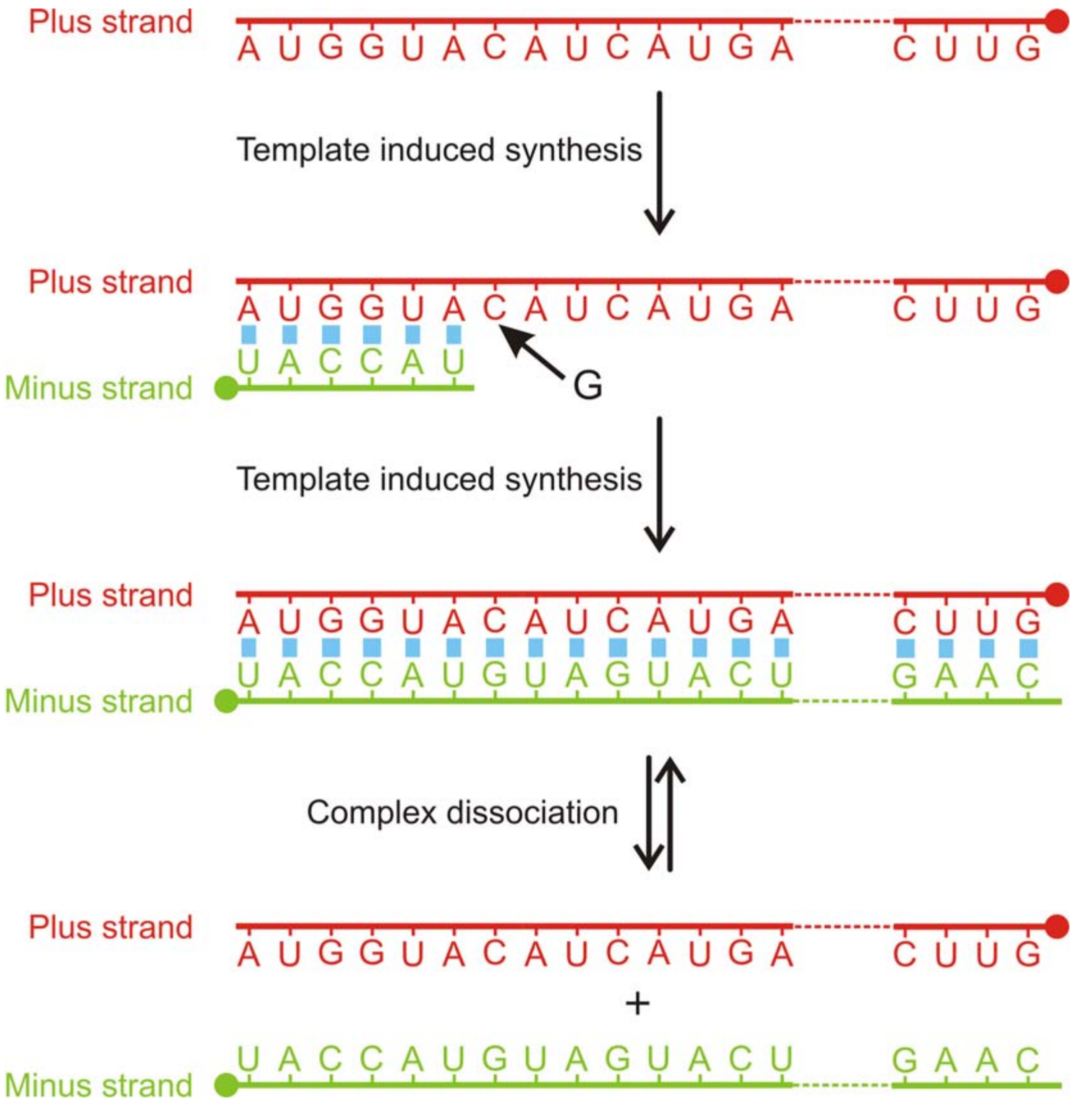


G ≡ C and **A = U**



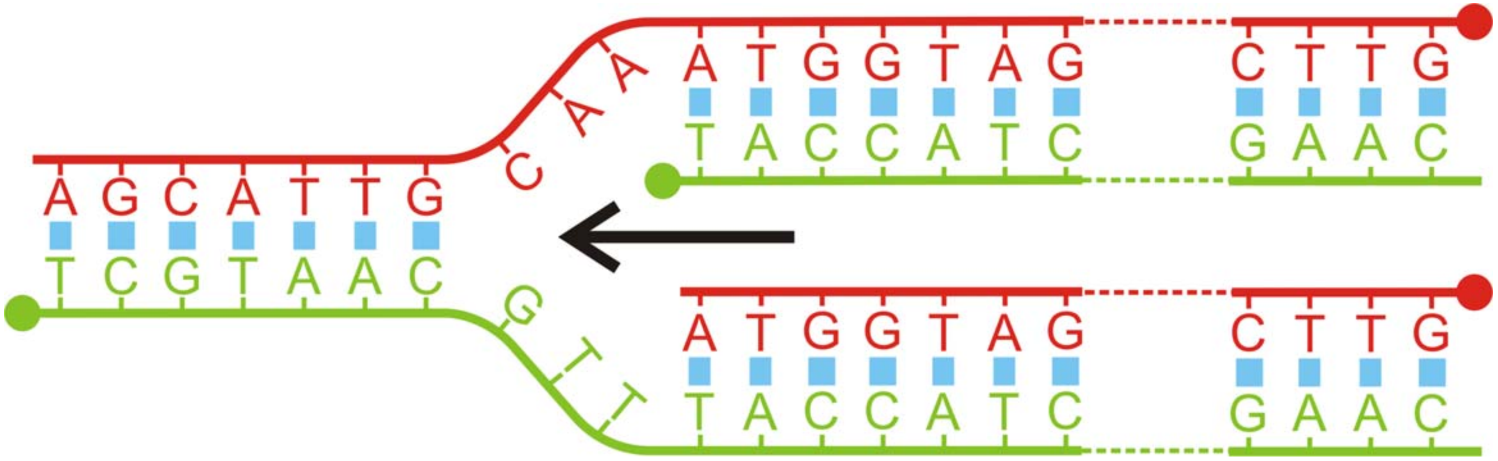
James D. Watson, 1928- , and Francis Crick, 1916-2004,
Nobel Prize 1962

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


Plus strand
A U G G U A C A U C A U G A C U U G
Template induced synthesis
Plus strand
A U G G U A C A U C A U G A C U U G
U A C C A U
G
Minus strand
Template induced synthesis
Plus strand
A U G G U A C A U C A U G A C U U G
U A C C A U G U A G U A C U G A A C
Minus strand
Complex dissociation
Plus strand
A U G G U A C A U C A U G A C U U G
+
U A C C A U G U A G U A C U G A A C
Minus strand

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

G≡C and **A=U**



‚Replication fork' in DNA replication

The mechanism of DNA replication is ‚semi-conservative'

Three necessary conditions for Darwinian evolution are:

1. **Multiplication,**
2. **Variation**, and
3. **Selection**.

Variation through mutation and recombination operates on the **genotype** whereas the **phenotype** is the target of **selection**.

One important property of the Darwinian scenario is that **variations** in the form of mutations or recombination events occur **uncorrelated** with their **effects on the selection process**.

All conditions can be fulfilled not only by cellular organisms but also by **nucleic acid molecules** in suitable **cell-free experimental assays**.

Evolution of RNA molecules based on Qβ phage

D.R.Mills, R.L.Peterson, S.Spiegelman, ***An extracellular Darwinian experiment with a self-duplicating nucleic acid molecule***. Proc.Natl.Acad.Sci.USA **58** (1967), 217-224

S.Spiegelman, ***An approach to the experimental analysis of precellular evolution***. Quart.Rev.Biophys. **4** (1971), 213-253

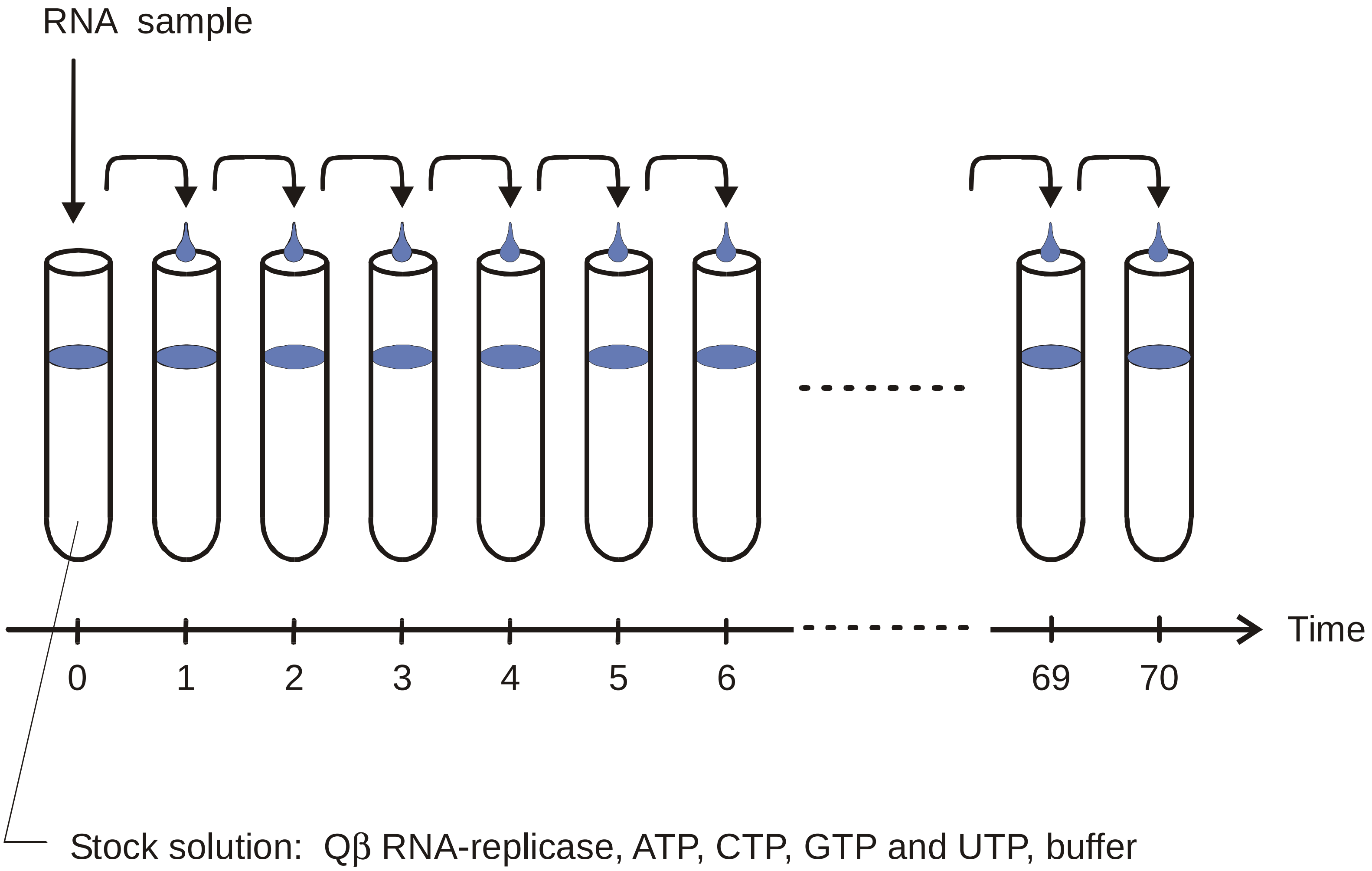
C.K.Biebricher, ***Darwinian selection of self-replicating RNA molecules***. Evolutionary Biology **16 (**1983), 1-52

G.Bauer, H.Otten, J.S.McCaskill, ***Travelling waves of*** **in vitro** ***evolving RNA.*** *Proc.Natl.Acad.Sci.USA* **86** (1989), 7937-7941

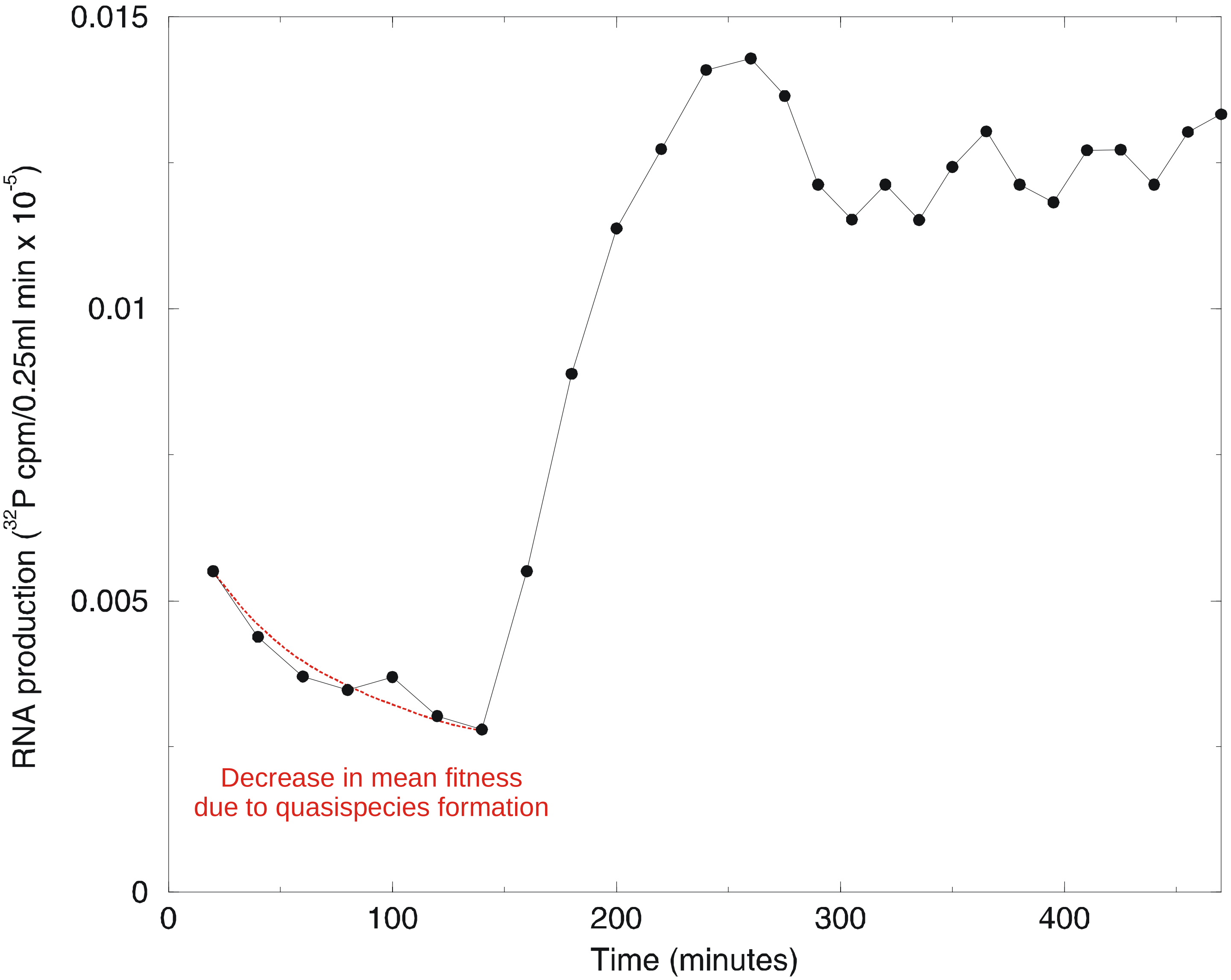
C.K.Biebricher, W.C.Gardiner, ***Molecular evolution of RNA*** **in vitro**. Biophysical Chemistry **66** (1997), 179-192

G.Strunk, T.Ederhof, ***Machines for automated evolution experiments*** **in vitro** ***based on the serial transfer concept***. Biophysical Chemistry **66** (1997), 193-202

F.Öhlenschlager, M.Eigen, ***30 years later – A new approach to Sol Spiegelman‘s and Leslie Orgel‘s*** **in vitro** ***evolutionary studies***. Orig.Life Evol.Biosph. **27** (1997), 437-457


RNA sample
Time
0
1
2
3
4
5
6
69
70
Stock solution: Qβ RNA-replicase, ATP, CTP, GTP and UTP, buffer

The serial transfer technique applied to RNA evolution ***in vitro***



The increase in RNA production rate during a serial transfer experiment

DIE NATURWISSENSCHAFTEN

58. Jahrgang, 1971 Heft 10 Oktober

Selforganization of Matter and the Evolution of Biological Macromolecules

Manfred Eigen*

Max-Planck-Institut für Biophysikalische Chemie,
Karl-Friedrich-Bonhoeffer-Institut, Göttingen-Nikolausberg

I. Introduction

I.1. „Cause and Effect"

The question about the origin of life often appears as a question about "cause and effect". Physical theories of macroscopic processes usually involve answers to such questions, even if a statistical interpretation is given to the relation between "cause" and "effect". It is mainly due to the nature of this question that many scientists believe that our present physics does not offer any obvious explanation for the existence of life, which even in its simplest forms always appears to be associated with complex macroscopic (i.e. multimolecular) systems, such as the living cell.

As a consequence of the exciting discoveries of "molecular biology", a common version of the above question is: *Which came first, the protein or the nucleic acid?* – a modern variant of the old "chicken-and-the-egg" problem. The term "first" is usually meant to define a causal rather than a temporal relationship, and the words "protein" and "nucleic acid" may be substituted by "function" and "information". The question in this form, when applied to the interplay of nucleic acids and proteins as presently encountered in the living cell, leads ad absurdum, because "function"

[*] Partly presented as the "Robbins Lectures" at Pomona College, California, in spring 1970.

Die Naturwissenschaften 64. Jahrgang Heft 11 November 1977

The Hypercycle

A Principle of Natural Self-Organization

Part A: Emergence of the Hypercycle

Manfred Eigen
Max-Planck-Institut für biophysikalische Chemie, D-3400 Göttingen

Peter Schuster
Institut für theoretische Chemie und Strahlenchemie der Universität, A-1090 Wien

This paper is the first part of a trilogy, which comprises a detailed study of a special type of functional organization and demonstrates its relevance with respect to the origin and evolution of life.
Self-replicative macromolecules, such as RNA or DNA in a suitable environment exhibit a behavior, which we may call Darwinian and which can be formally represented by the concept of the quasi-species. A quasi-species is defined as a *given* distribution of macromolecular species with closely interrelated sequences, dominated by one or several (degenerate) master copies. External constraints enforce the selection of the best adapted distribution, commonly referred to as the wild-type. Most important for Darwinian behavior are the criteria for internal stability of the quasi-species. If these criteria are violated, the information stored in the nucleotide sequence of the master copy will disintegrate irreversibly leading to an error catastrophy. As a consequence, selection and evolution of RNA or DNA molecules is limited with respect to the amount of information that can be stored in a single replicative unit. An analysis of experimental data regarding RNA and DNA replication at various levels of organization reveals, that a sufficient amount of information for the build up of a translation machinery can be gained only via integration of several different replicative units (or reproductive cycles) through *functional* linkages. A stable functional integration then will raise the system to a new level of organization and thereby enlarge its information capacity considerably. The hypercycle appears to be such a form of organization.

Preview on Part B: The Abstract Hypercycle

The mathematical analysis of dynamical systems using methods of differential topology, yields the result that there is only *one* type of mechanisms which fulfills the following requirements:
The information stored in each single replicative unit (or reproductive cycle) must be maintained, i.e., the respective master copies must compete favorably with their *error* distributions. Despite their competitive behavior these units must establish a cooperation which includes all functionally integrated species. On the other hand, the cycle as a whole must continue to compete strongly with any other single entity or linked ensemble which does not contribute to its integrated function.
These requirements are crucial for a selection of the best adapted functionally linked ensemble and its evolutive optimization. Only hypercyclic organizations are able to fulfil these requirements. Non-cyclic linkages among the autonomous reproduction cycles, such as chains or branched, tree-like networks are devoid of such properties.
The mathematical methods used for proving these assertions are fixed-point, Lyapunow- and trajectorial analysis in higher-dimensional phase spaces, spanned by the concentration coordinates of the cooperating partners. The self-organizing properties of hypercycles are elucidated, using analytical as well as numerical techniques.

Preview on Part C: The Realistic Hypercycle

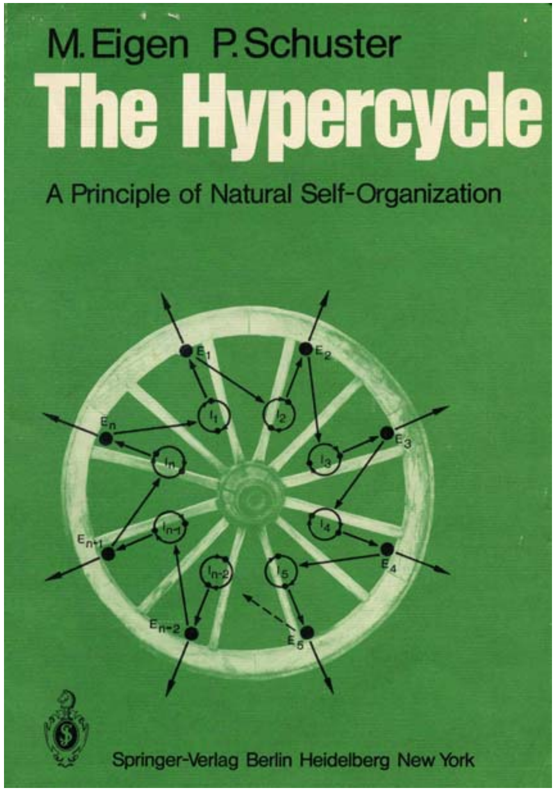
A realistic model of a hypercycle relevant with respect to the origin of the genetic code and the translation machinery is presented. It includes the following features referring to natural systems:
1) The hypercycle has a sufficiently simple structure to admit an origination with finite probability under prebiotic conditions.
2) It permits a continuous emergence from closely interrelated (t-RNA-like) precursors, originally being members of a stable RNA quasi-species and having been amplified to a level of higher abundance.
3) The organizational structure and the properties of single functional units of this hypercycle are still reflected in the present genetic code in the translation apparatus of the prokaryotic cell, as well as in certain bacterial viruses.

I. The Paradigm of Unity and Diversity in Evolution

Why do millions of species, plants and animals, exist, while there is only one basic molecular machinery of the cell: one universal genetic code and unique chiralities of the macromolecules?

The geneticists of our day would not hesitate to give an immediate answere to the first part of this question. Diversity of species is the outcome of the tremendous branching process of evolution with its myriads of single steps of reproduction and mutation. It in-

Naturwissenschaften 64, 541–565 (1977) © by Springer-Verlag 1977



M. Eigen P. Schuster

The Hypercycle

A Principle of Natural Self-Organization

Springer-Verlag Berlin Heidelberg New York

Chemical kinetics of molecular evolution

M. Eigen, P. Schuster, `The Hypercycle´, Springer-Verlag, Berlin 1979

$$(A) + I_1 \xrightarrow{f_1} I_1 + I_1$$

$$(A) + I_2 \xrightarrow{f_2} I_2 + I_2$$

$$(A) + I_i \xrightarrow{f_i} I_i + I_i$$

$$(A) + I_m \xrightarrow{f_m} I_m + I_m$$

$$(A) + I_n \xrightarrow{f_n} I_n + I_n$$

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

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

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

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

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

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

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

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

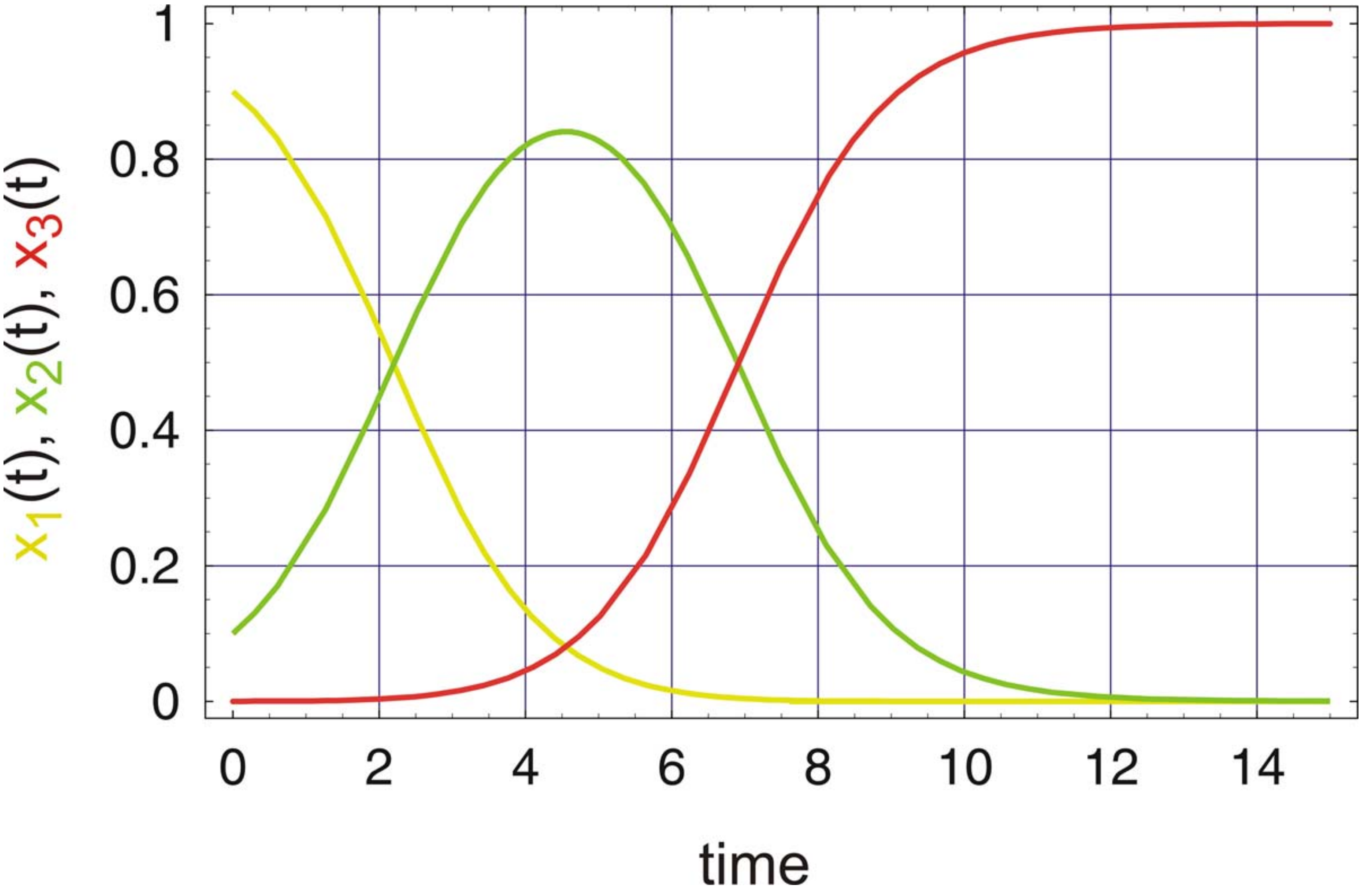
$$\frac{dx_i}{dt} = x_i\left(f_i - \phi\right), \quad i = 1,2,\cdots,n; \quad \sum\nolimits_{i=1}^{n} x_i = 1; \quad \phi = \sum\nolimits_{j=1}^{n} f_j\, x_j = \overline{f}$$

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

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

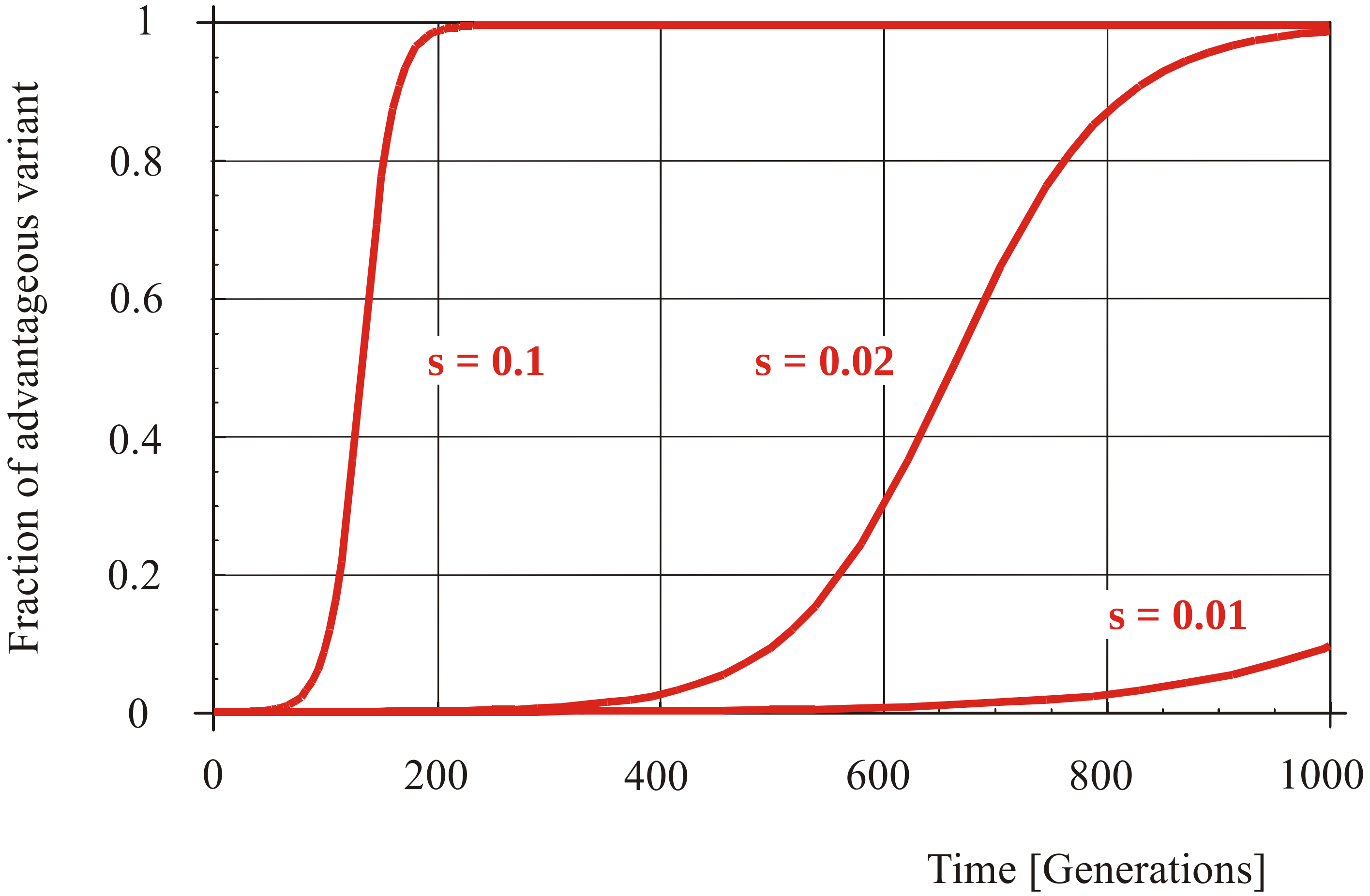
Solutions are obtained by integrating factor transformation

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

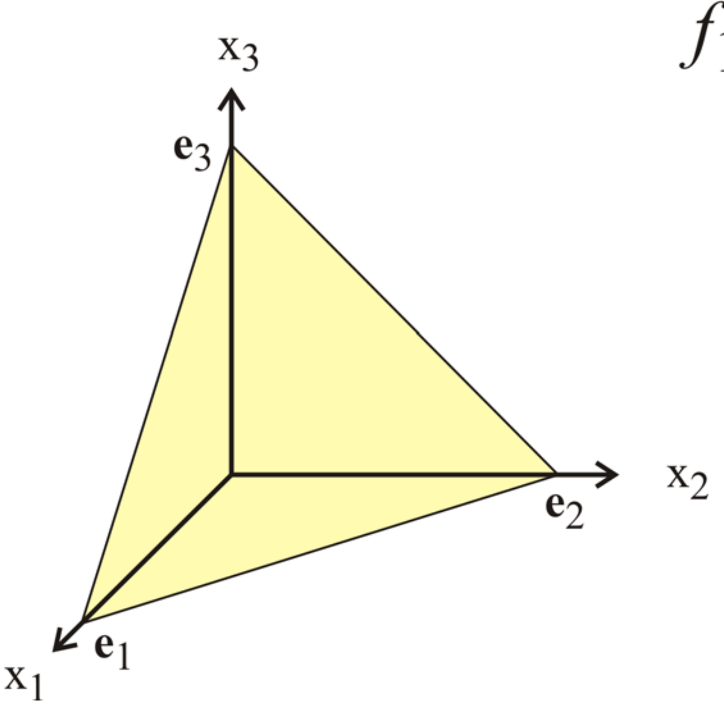


Selection between three species with $f_1 = 1$, $f_2 = 2$, and $f_3 = 3$

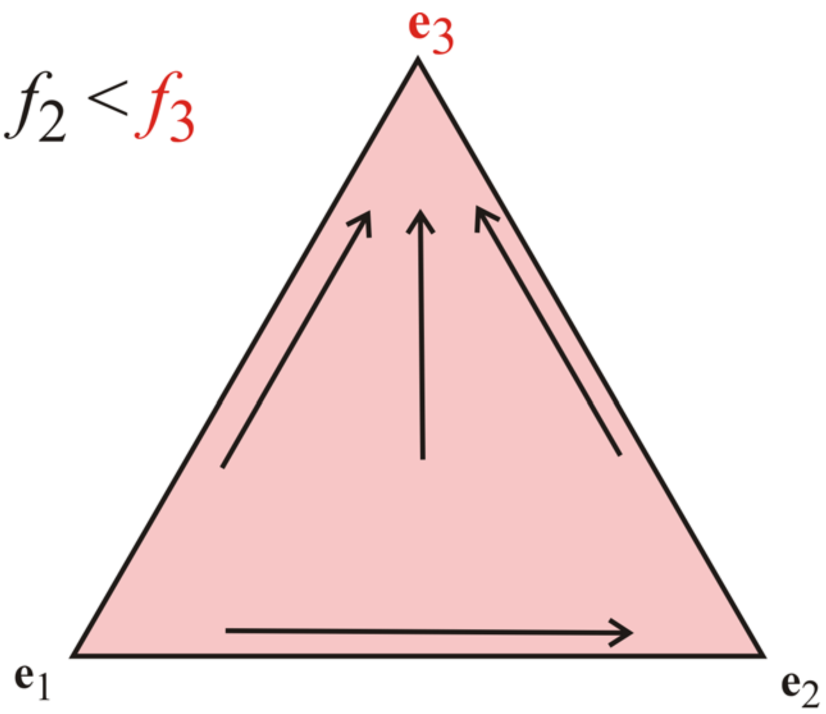
$\mathbf{s} = (f_2\text{-}f_1) \,/\, f_1;\ f_2 > f_1\ ;\ x_1(0) = 1 - 1/\mathrm{N}\ ;\ x_2(0) = 1/\mathrm{N}$



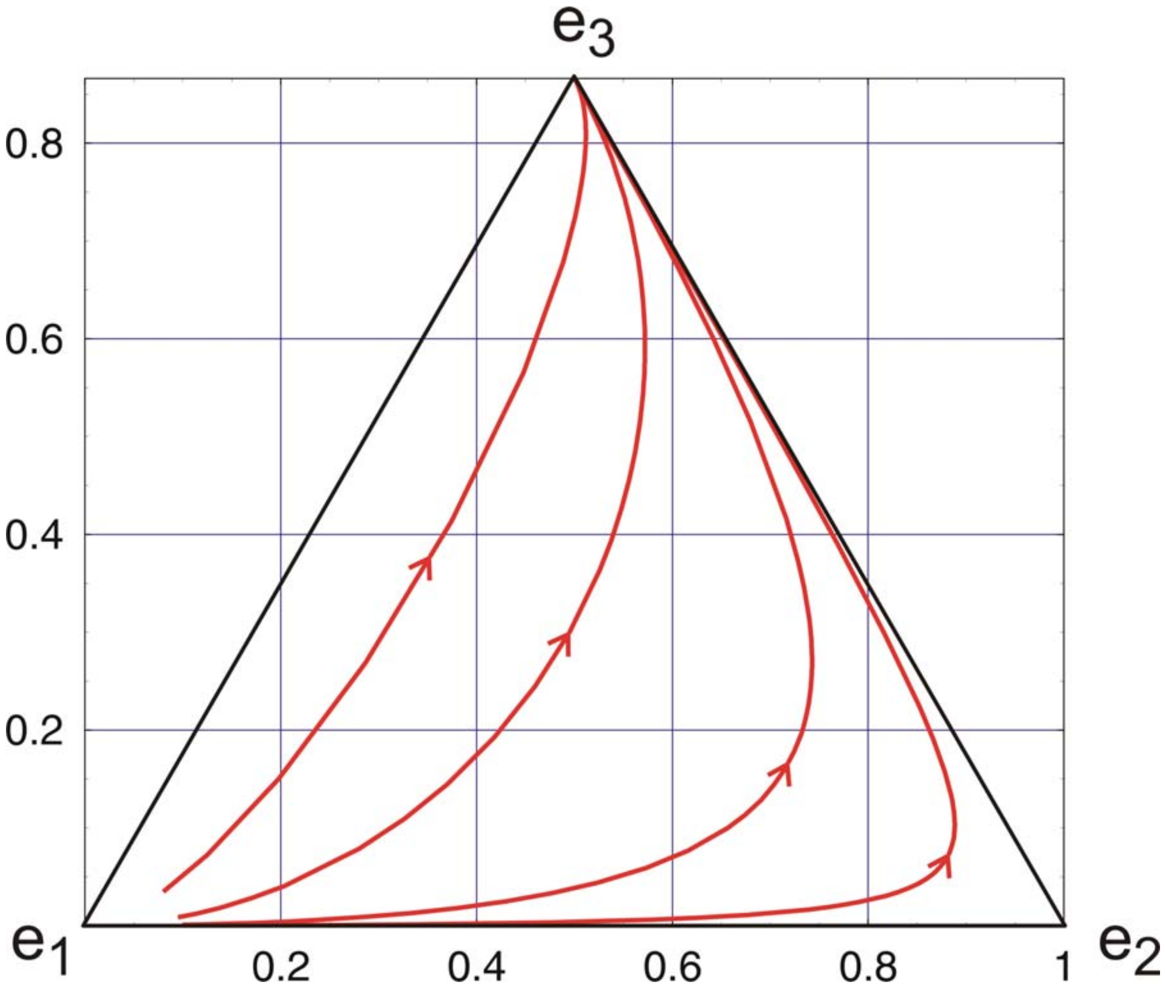
Selection of advantageous mutants in populations of N = 10 000 individuals



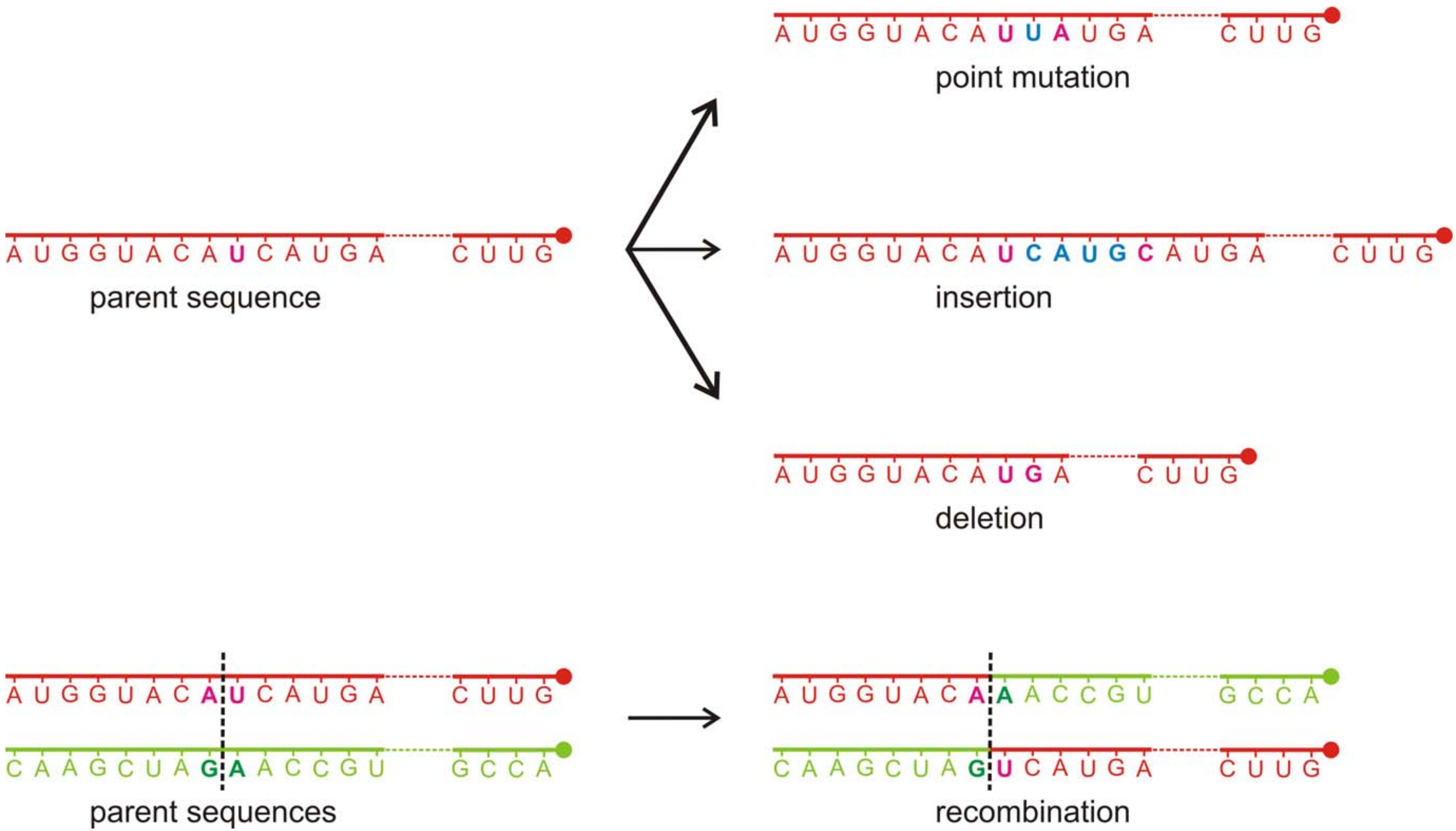
$f_1 < f_2 < f_3$



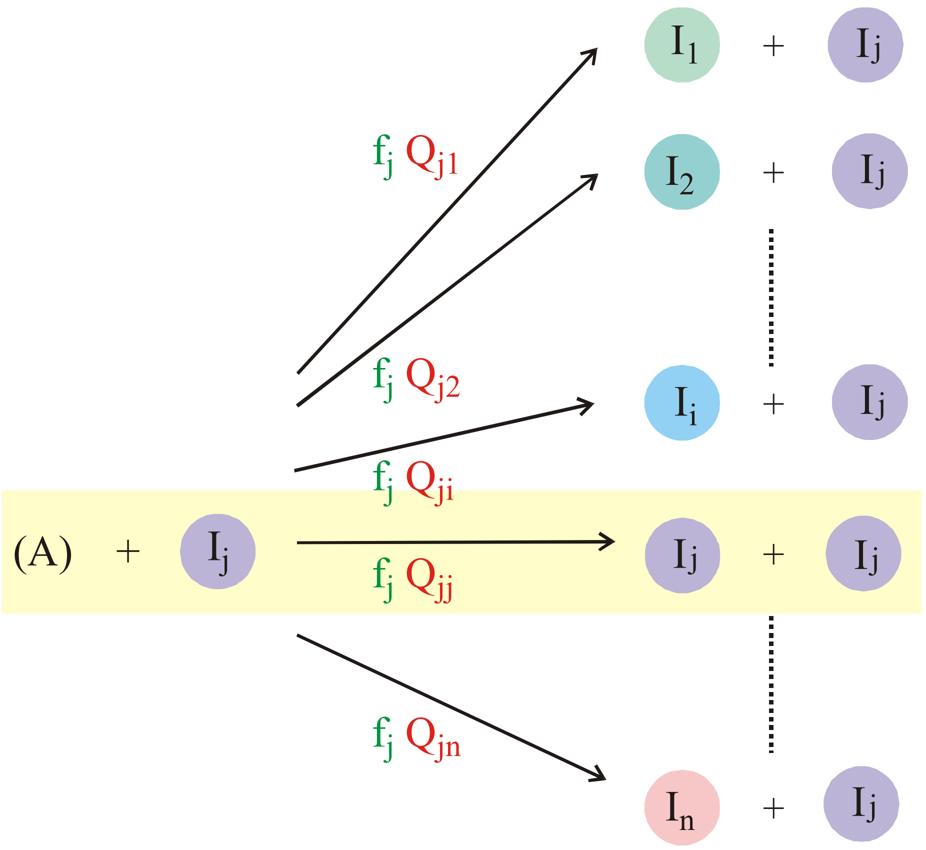
Selection on the concentration simplex: $S_3 = \left\{ x_i \geq 0,\ i = 1,2,3;\ \sum_{i=1}^{3} x_i = 1 \right\}$



Selection between three species with $f_1 = 1, f_2 = 2,$ and $f_3 = 3$, parametric plot on S_3


A U G G U A C A U U A U G A C U U G
point mutation
A U G G U A C A U C A U G A C U U G
parent sequence
A U G G U A C A U C A U G C A U G A C U U G
insertion
A U G G U A C A U G A C U U G
deletion
A U G G U A C A U C A U G A C U U G
C A A G C U A G A A C C G U G C C A
parent sequences
A U G G U A C A A A C C G U G C C A
C A A G C U A G U C A U G A C U U G
recombination

Variation of genotypes through mutation and recombination



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

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

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

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

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

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

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

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

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

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

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

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

Perron-Frobenius theorem applied to the value matrix W

W is primitive: (i) λ_0 is real and strictly positive

(ii) $\lambda_0 > |\lambda_k| \text{ for all } k \neq 0$

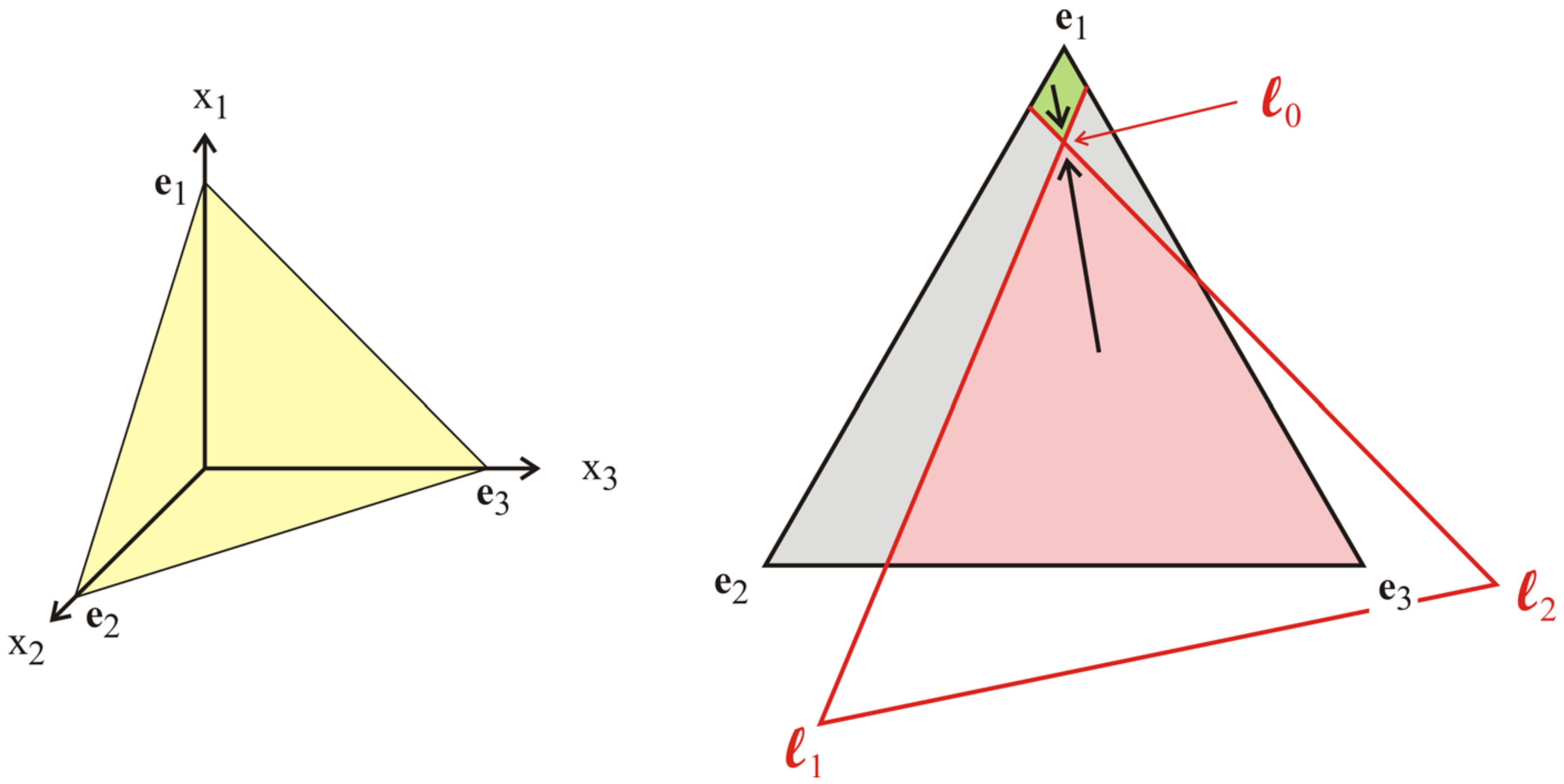
(iii) λ_0 is associated with strictly positive eigenvectors

(iv) λ_0 is a simple root of the characteristic equation of W

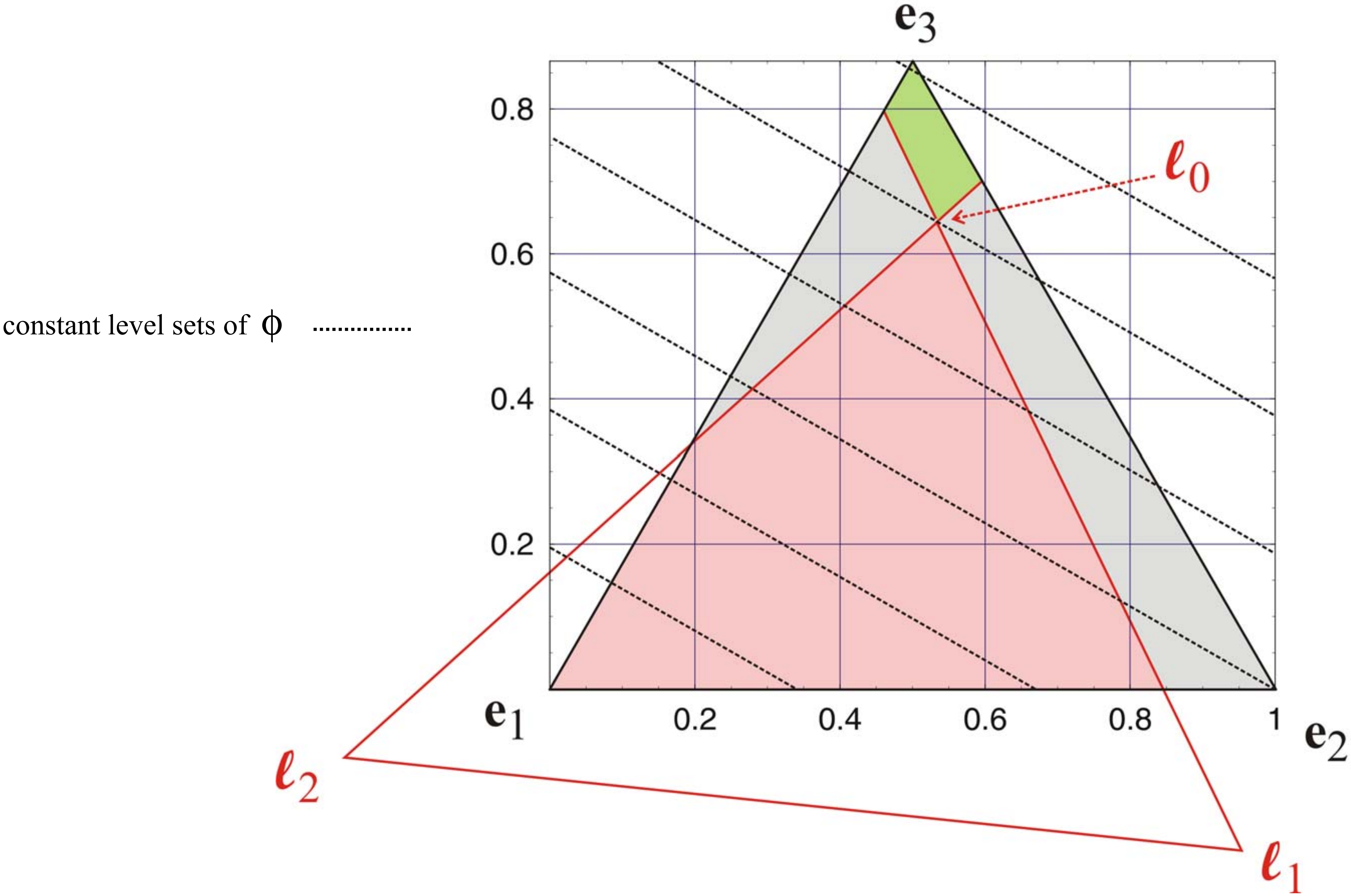
(v-vi) etc.

W is irreducible: (i), (iii), (iv), etc. as above

(ii) $\lambda_0 \geq |\lambda_k| \text{ for all } k \neq 0$

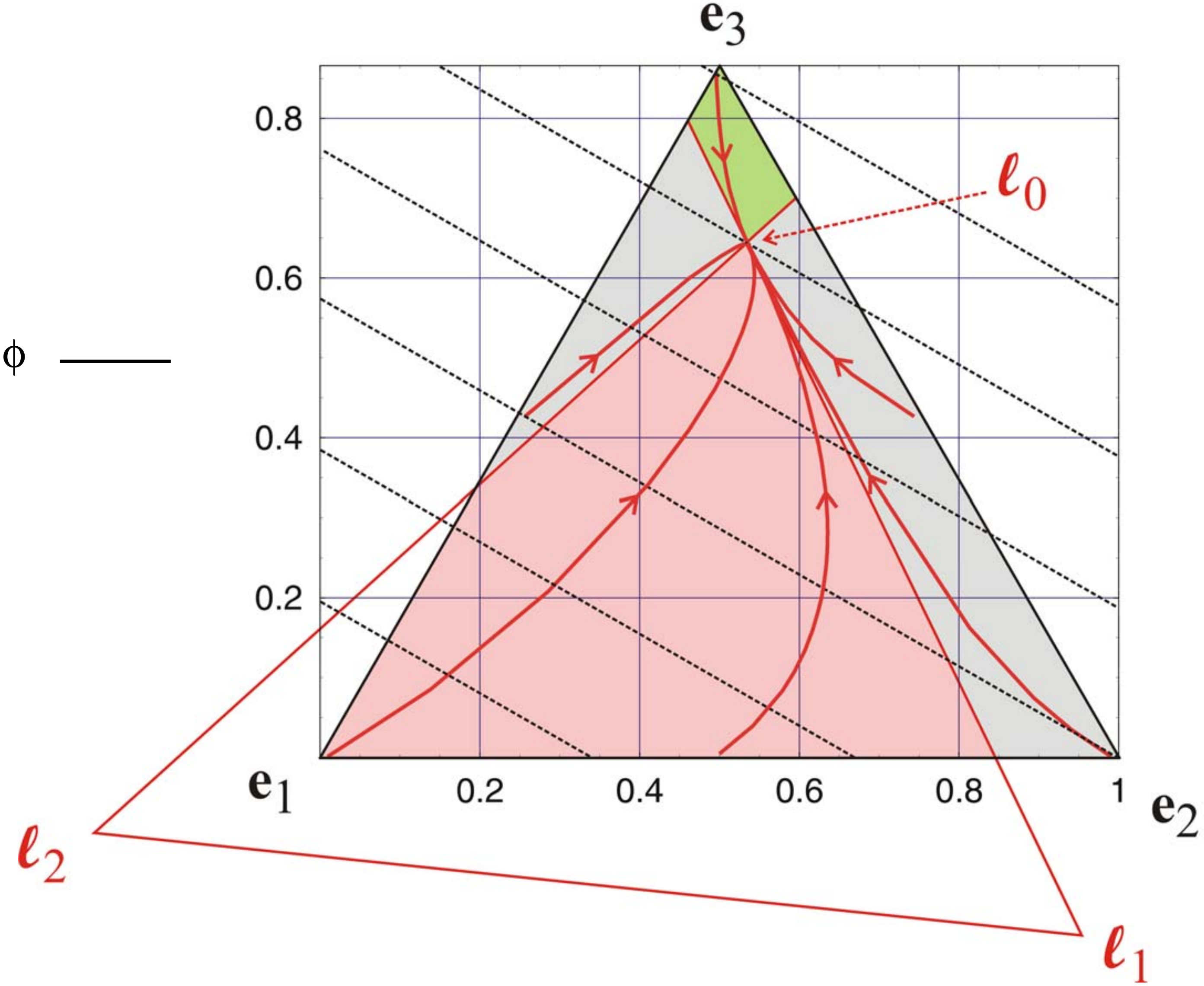


The quasispecies on the concentration simplex: $S_3 = \left\{ x_i \geq 0,\ i = 1, 2, 3;\ \sum_{i=1}^{3} x_i = 1 \right\}$

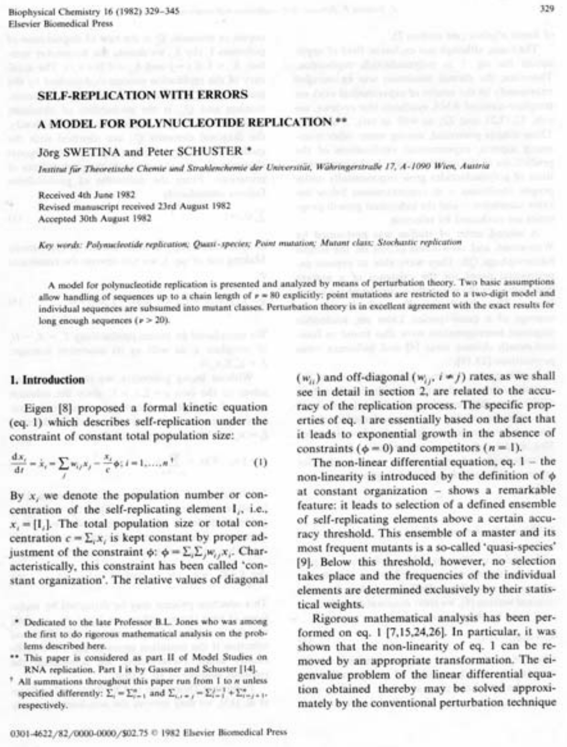


Selection of quasispecies with $f_1 = 1.9$, $f_2 = 2.0$, $f_3 = 2.1$, and $p = 0.01$

constant level sets of ϕ ——



Selection of quasispecies with $f_1 = 1.9$, $f_2 = 2.0$, $f_3 = 2.1$, and $p = 0.01$, parametric plot on S_3



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 $\nu = 80$ 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 $c = \Sigma_i x_i$ is kept constant by proper adjustment of the constraint ϕ: $\phi = \Sigma_i \Sigma_j w_{ij} x_j$. 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.

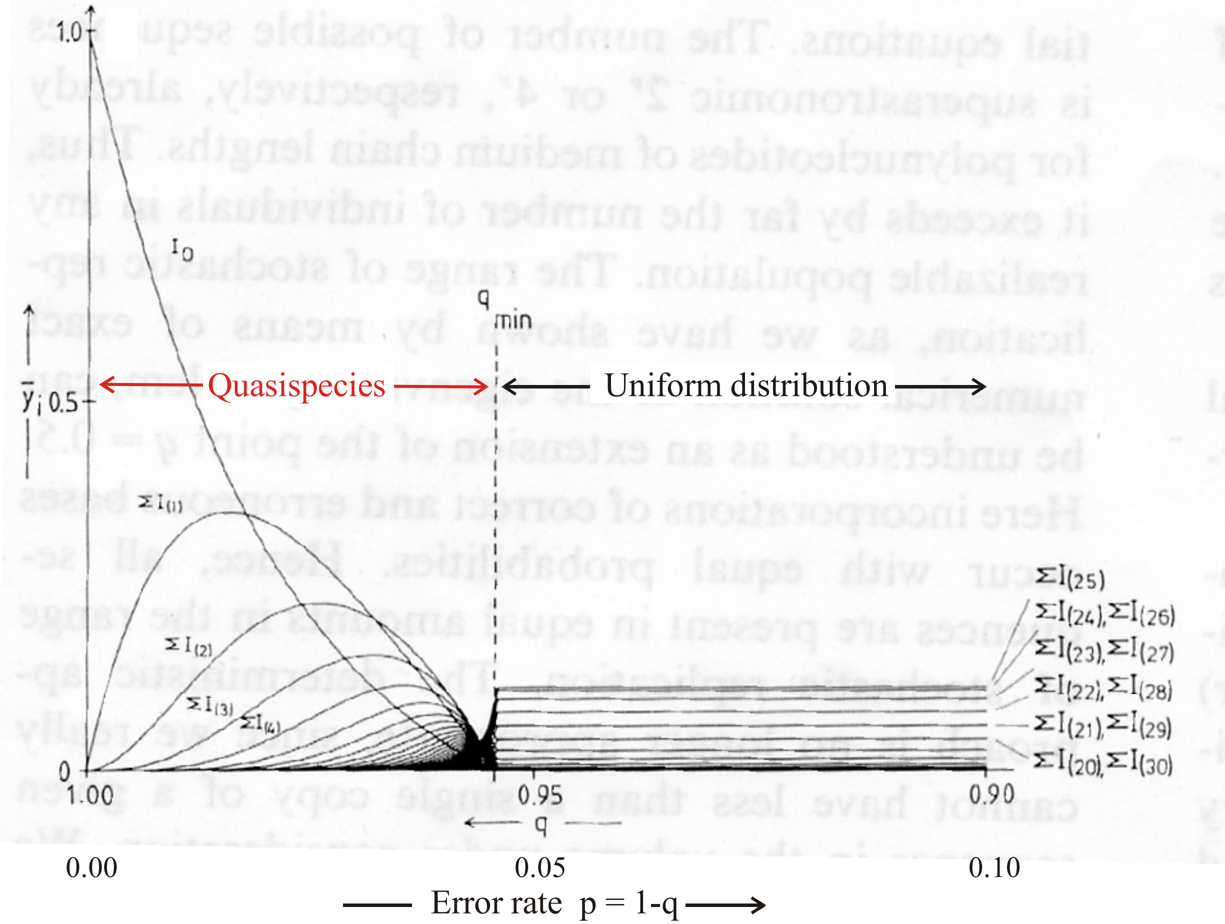
Rigorous mathematical analysis has been performed on eq. 1 [7,15,24,26]. In particular, it was shown that the non-linearity of eq. 1 can be removed by an appropriate transformation. The eigenvalue problem of the linear differential equation obtained thereby may be solved approximately by the conventional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

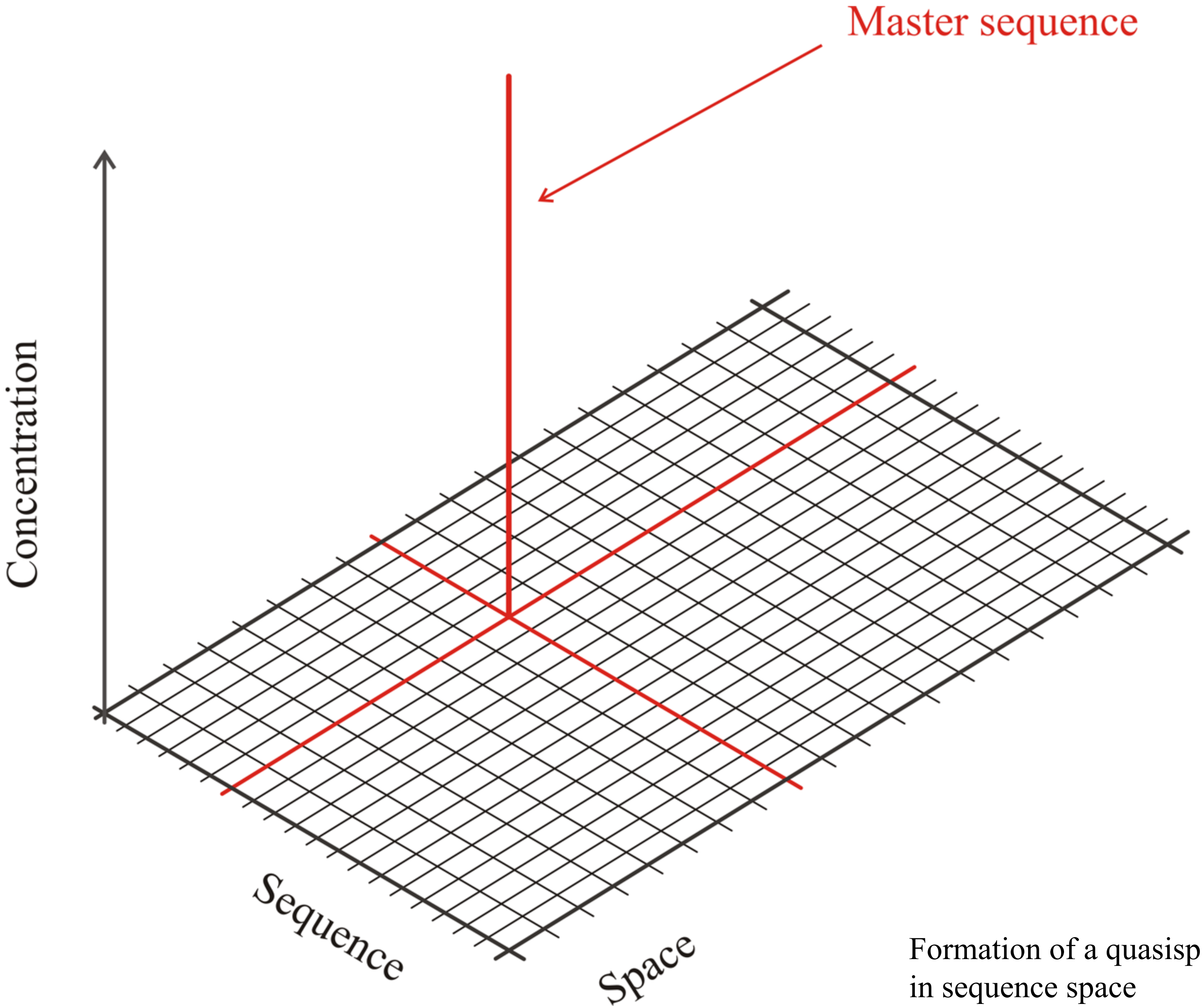
[**] This paper is considered as part II of Model Studies on RNA replication. Part I is by Gassner and Schuster [14].

[†] All summations throughout this paper run from 1 to n unless specified differently: $\Sigma_i = \Sigma_{i=1}^{n}$ and $\Sigma_{i,i\neq j} = \Sigma_{i=1}^{j-1} + \Sigma_{i=j+1}^{n}$, respectively.

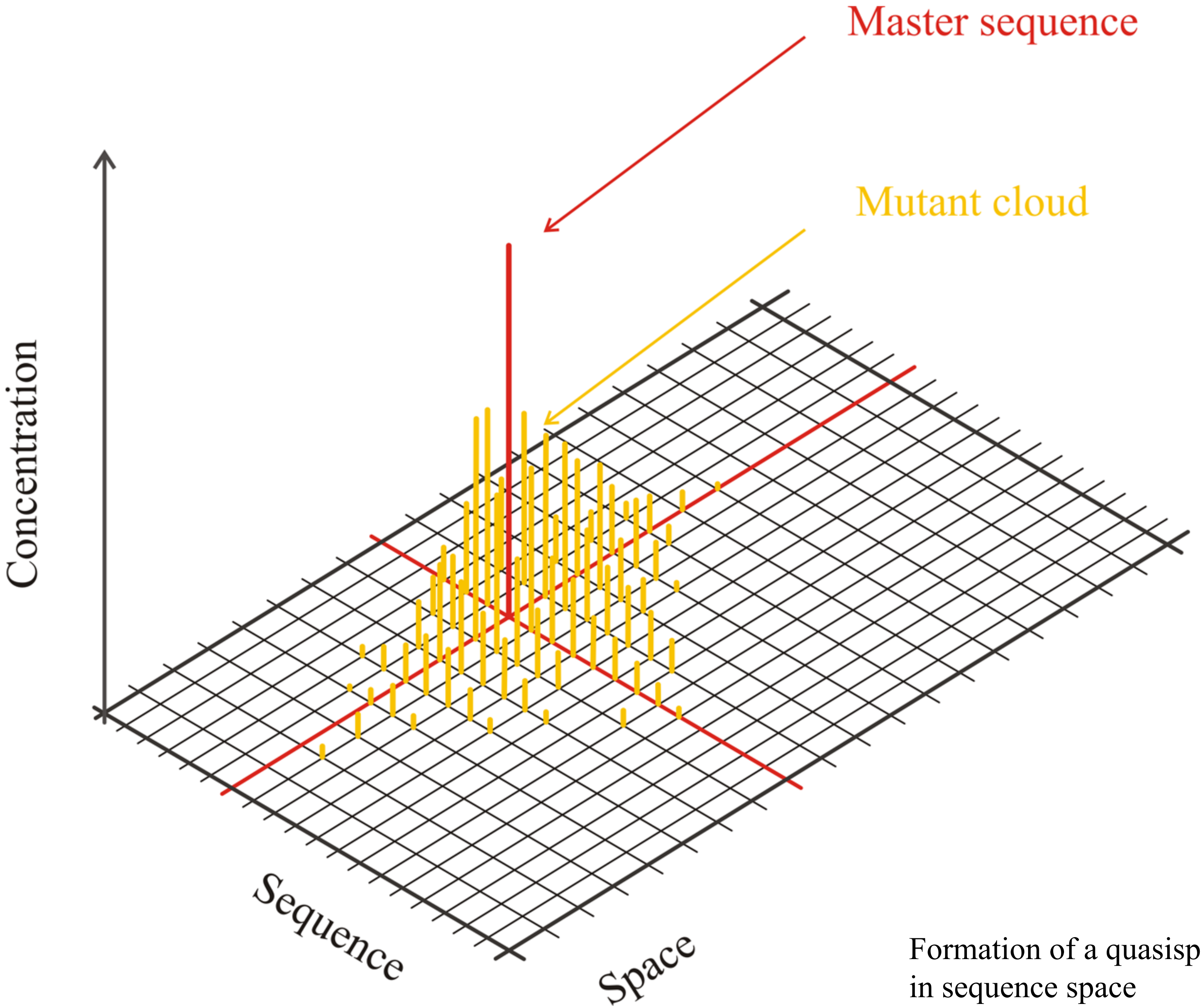
0301-4622/82/0000-0000/$02.75 © 1982 Elsevier Biomedical Press



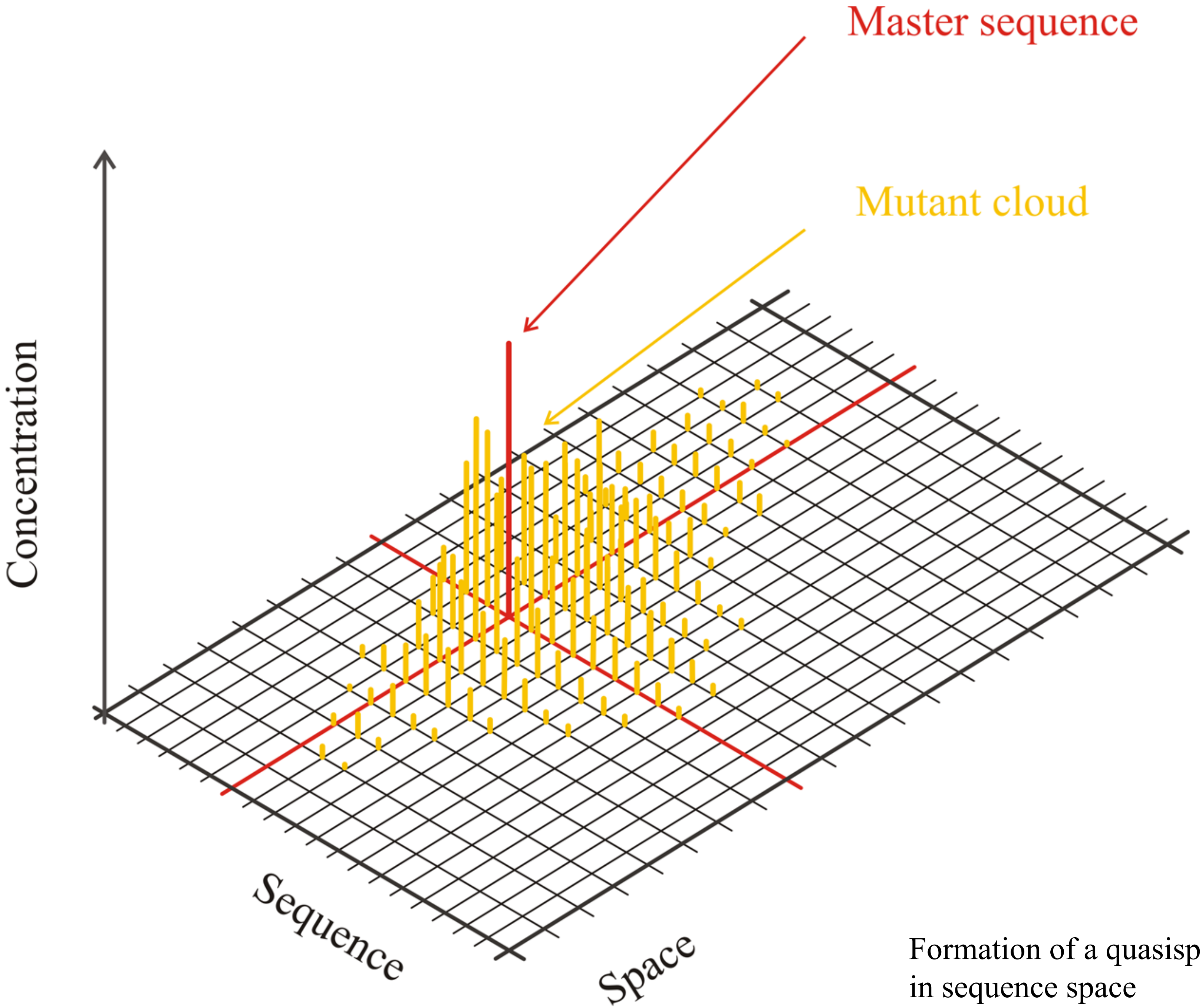
Quasispecies as a function of the replication accuracy q



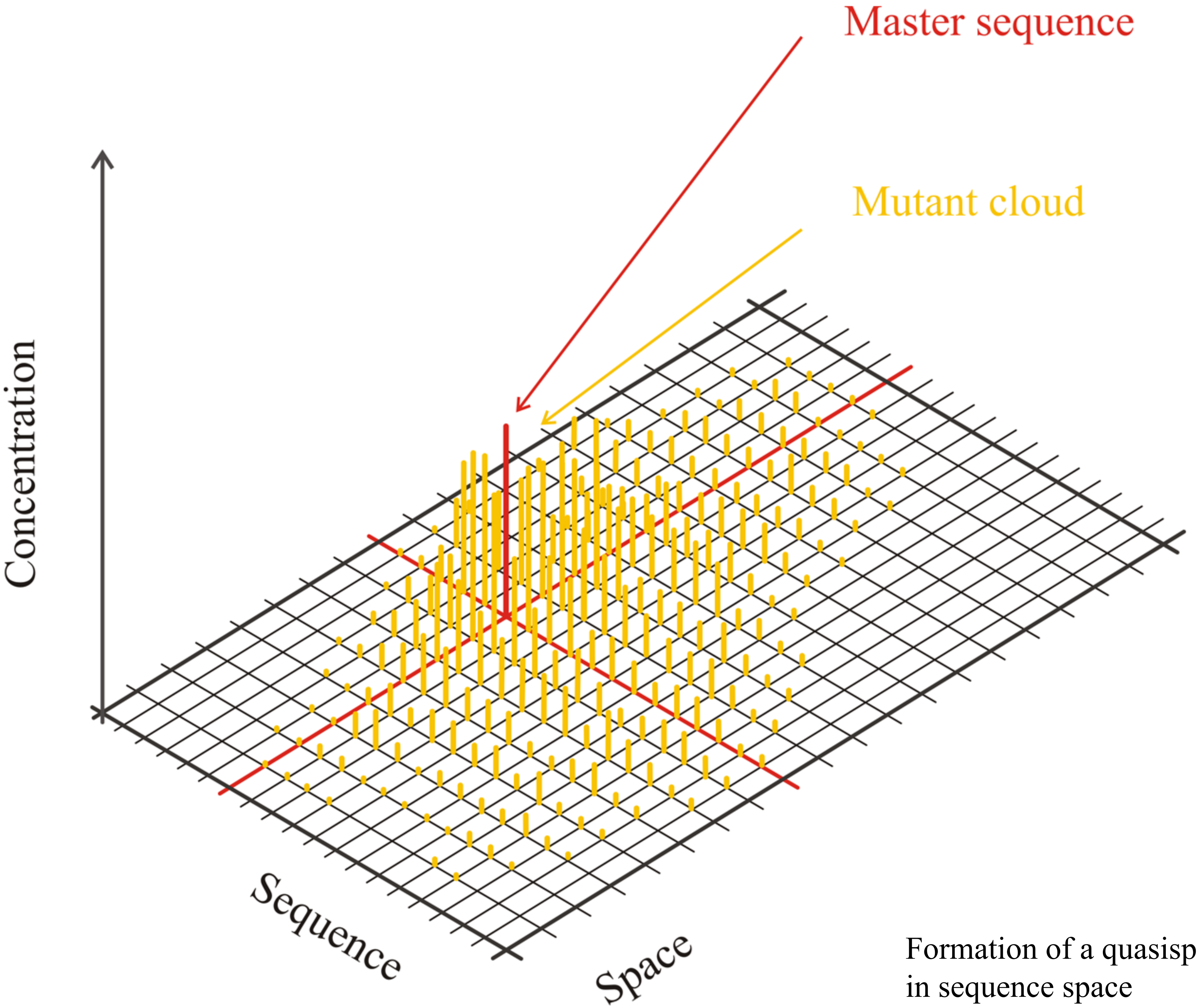
Formation of a quasispecies
in sequence space



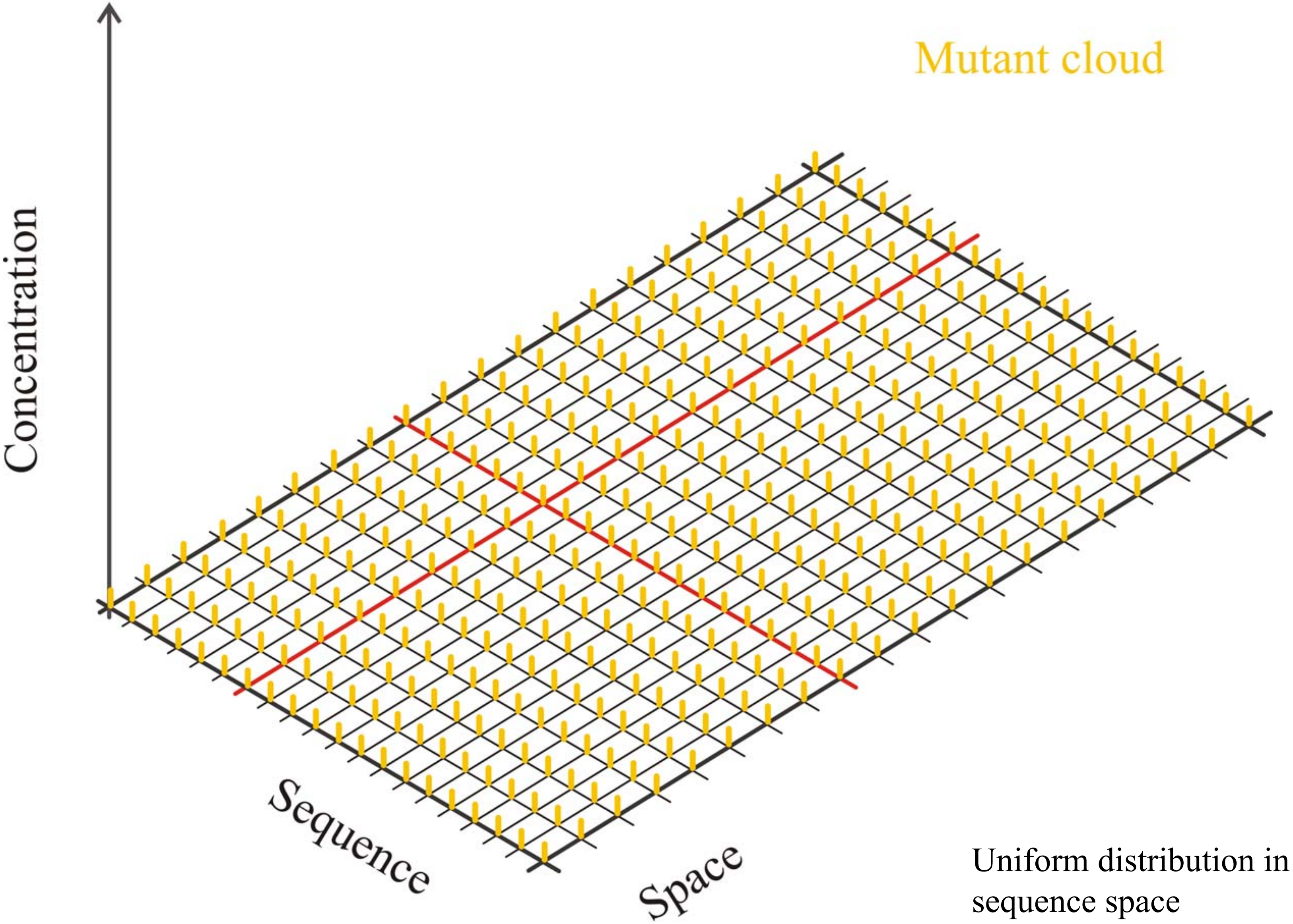
Formation of a quasispecies
in sequence space



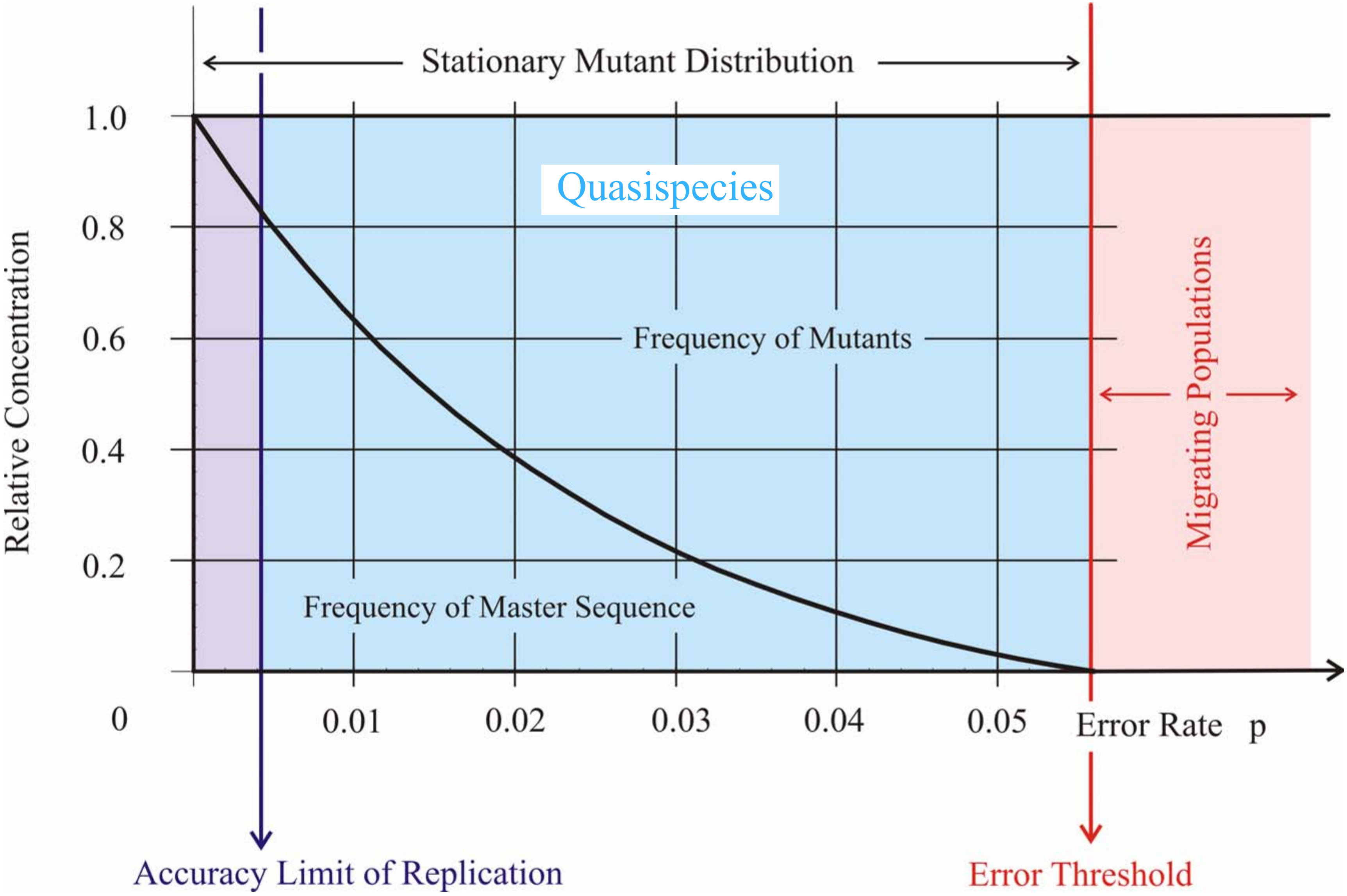
Formation of a quasispecies
in sequence space



Formation of a quasispecies
in sequence space



Uniform distribution in
sequence space



The error threshold in replication

Chain length and error threshold

$$Q \cdot \sigma \;=\; (1-p)^n \cdot \sigma \;\geq\; 1 \quad\Rightarrow\quad n \cdot \ln(1-p) \geq -\ln\sigma$$

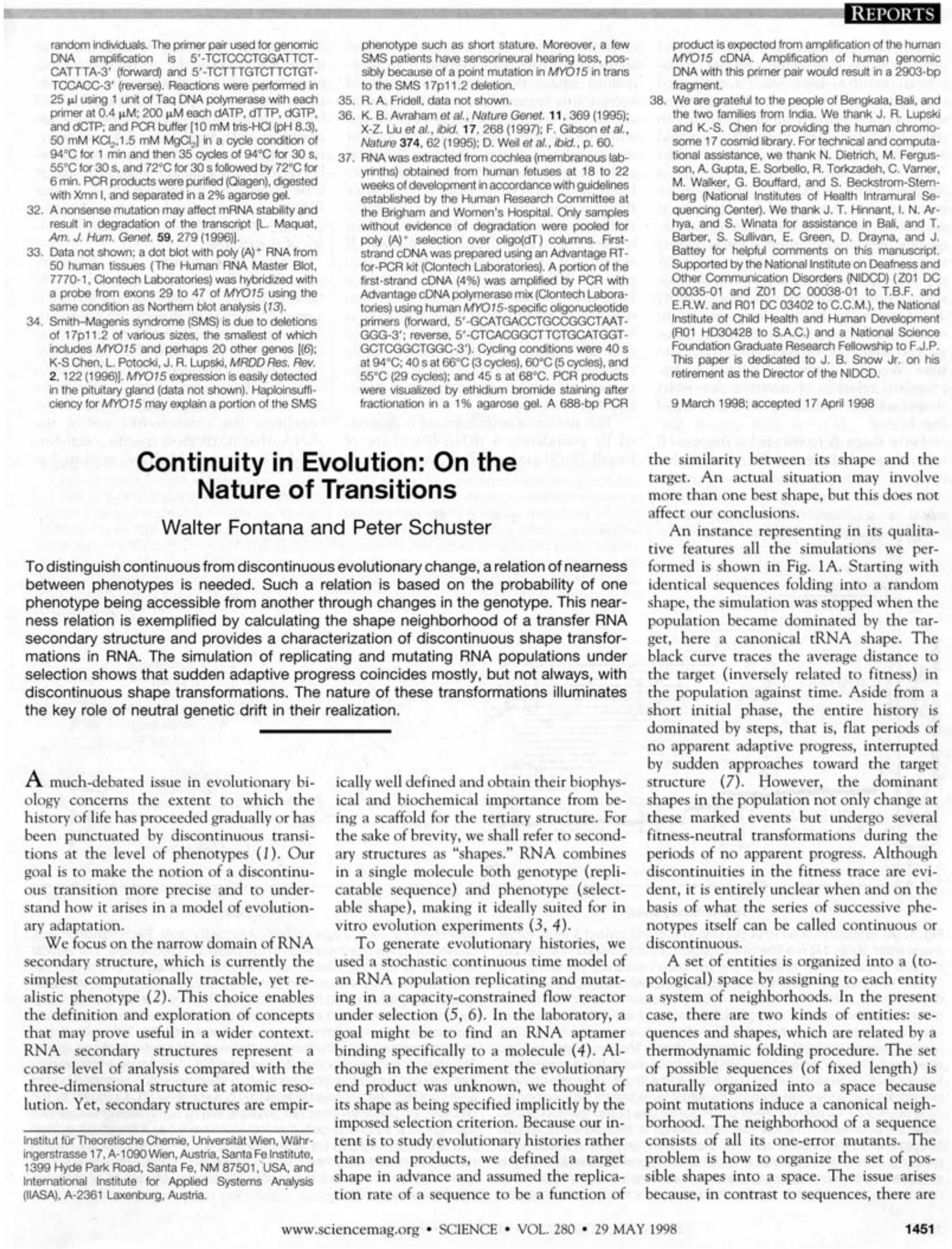
$$p \;\ldots\; \text{constant}: \quad n_{\max} \approx \frac{\ln\sigma}{p}$$

$$n \;\ldots\; \text{constant}: \quad p_{\max} \approx \frac{\ln\sigma}{n}$$

$Q=(1-p)^n$	…	replication accuracy
p	…	error rate
n	…	chain length
$\sigma = \dfrac{f_m}{\sum_{j\neq m} f_j}$	…	superiority of master sequence

Evolution *in silico*

W. Fontana, P. Schuster,
Science **280** (1998), 1451-1455



random individuals. The primer pair used for genomic DNA amplification is 5′-TCTCCCTGGATTCTCATTTA-3′ (forward) and 5′-TCTTTGTCTTCTGTTCCACC-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 [(*6*); K-S Chen, L. Potocki, J. R. Lupski, *MRDD 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′-GCATGACCTGCCGGCTAATGGG-3′; reverse, 5′-CTCACGGCTTCTGCATGGTGCTCGGCTGGC-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 2903-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. Fergusson, A. Gupta, E. Sorbello, R. Torkzadeh, C. Varner, M. Walker, G. Bouffard, and S. Beckstrom-Sternberg (National Institutes of Health Intramural Sequencing Center). We thank J. T. Hinnant, 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 00038-01 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.

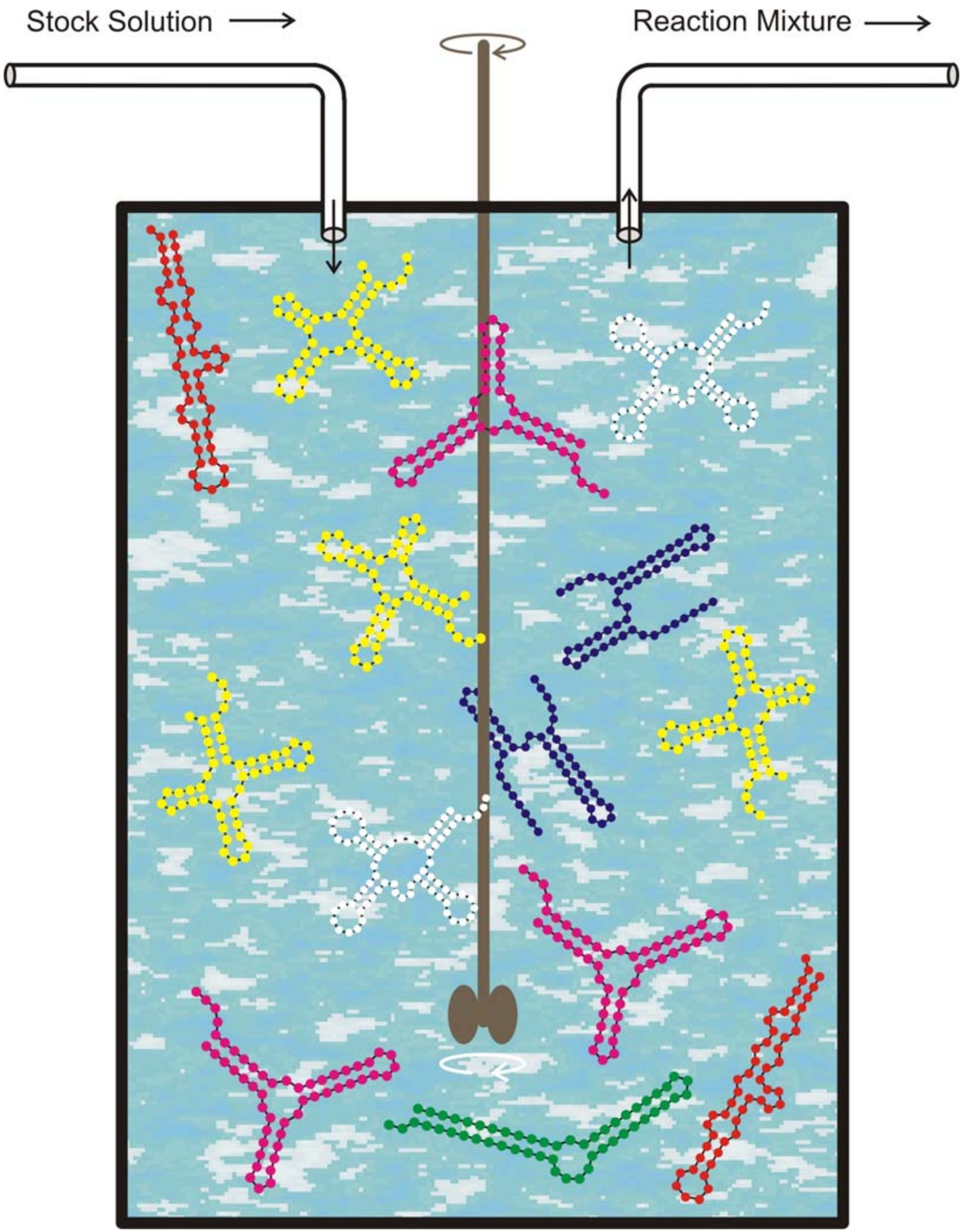
We focus on the narrow domain of RNA secondary structure, which is currently the simplest computationally tractable, yet realistic phenotype (*2*). This choice enables the definition and exploration of concepts that may prove useful in a wider context. RNA secondary structures represent a coarse level 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.



Replication rate constant:

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

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

Selection constraint:

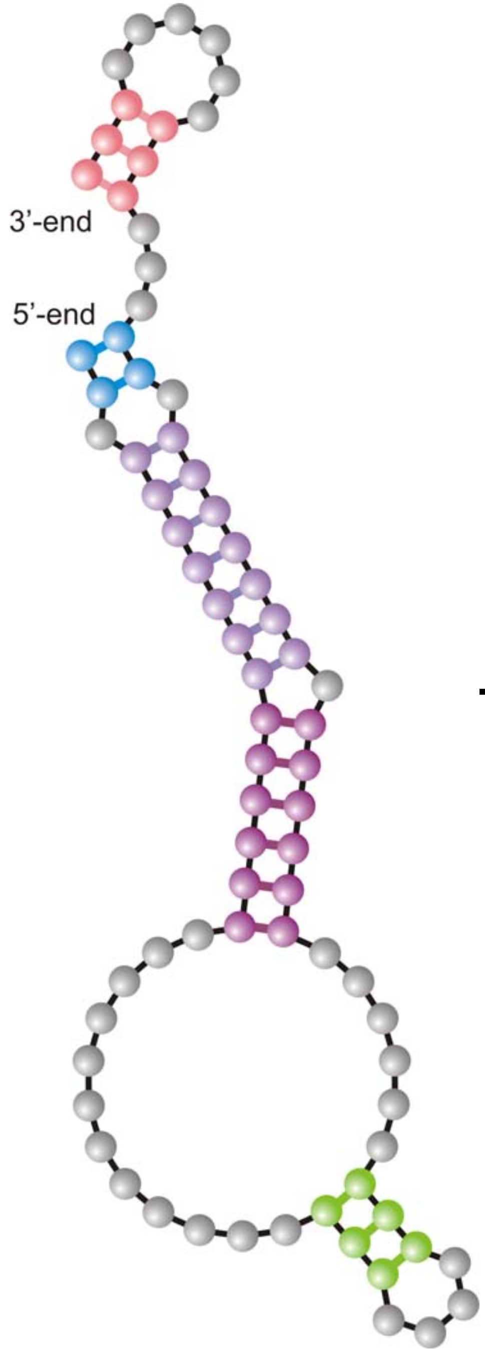
Population size, N = # RNA molecules, is controlled by the flow

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

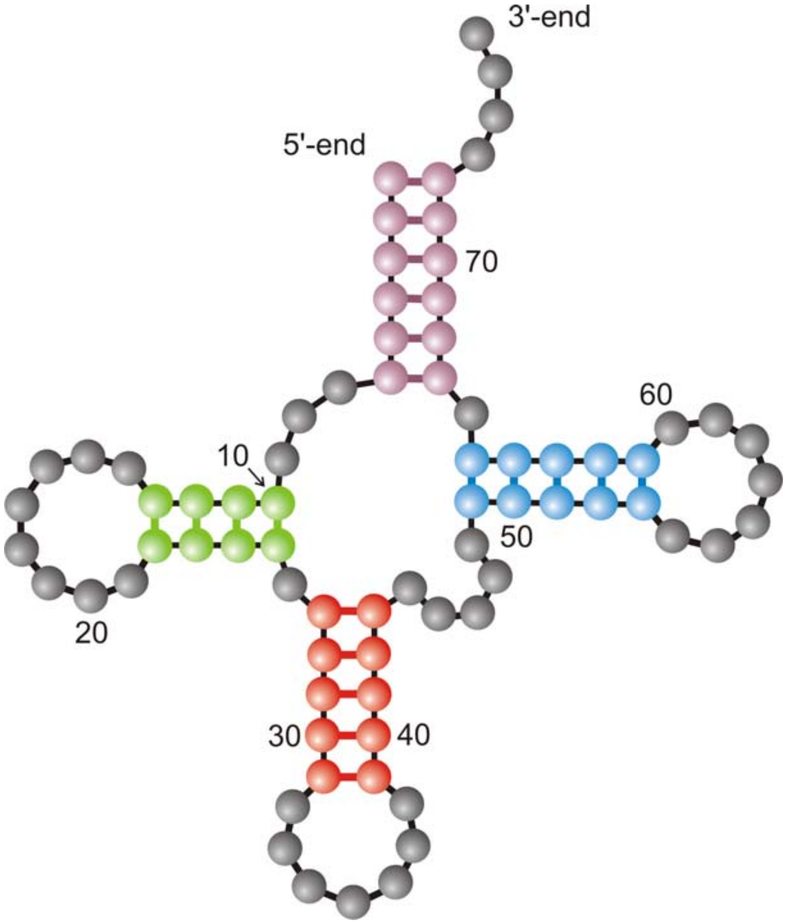
Mutation rate:

p = 0.001 / site × replication

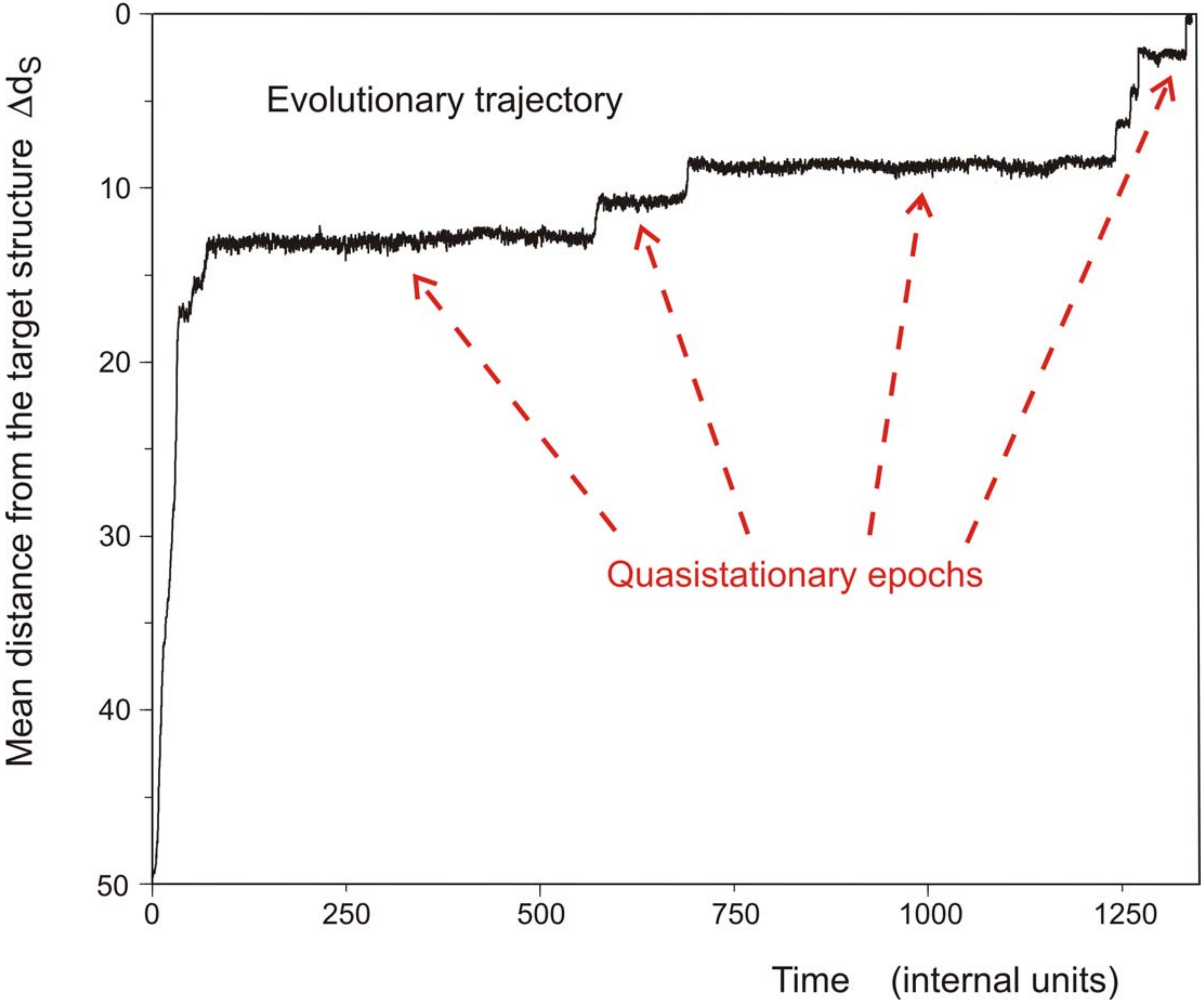
The flowreactor as a device for **studies** of evolution ***in vitro*** and ***in silico***



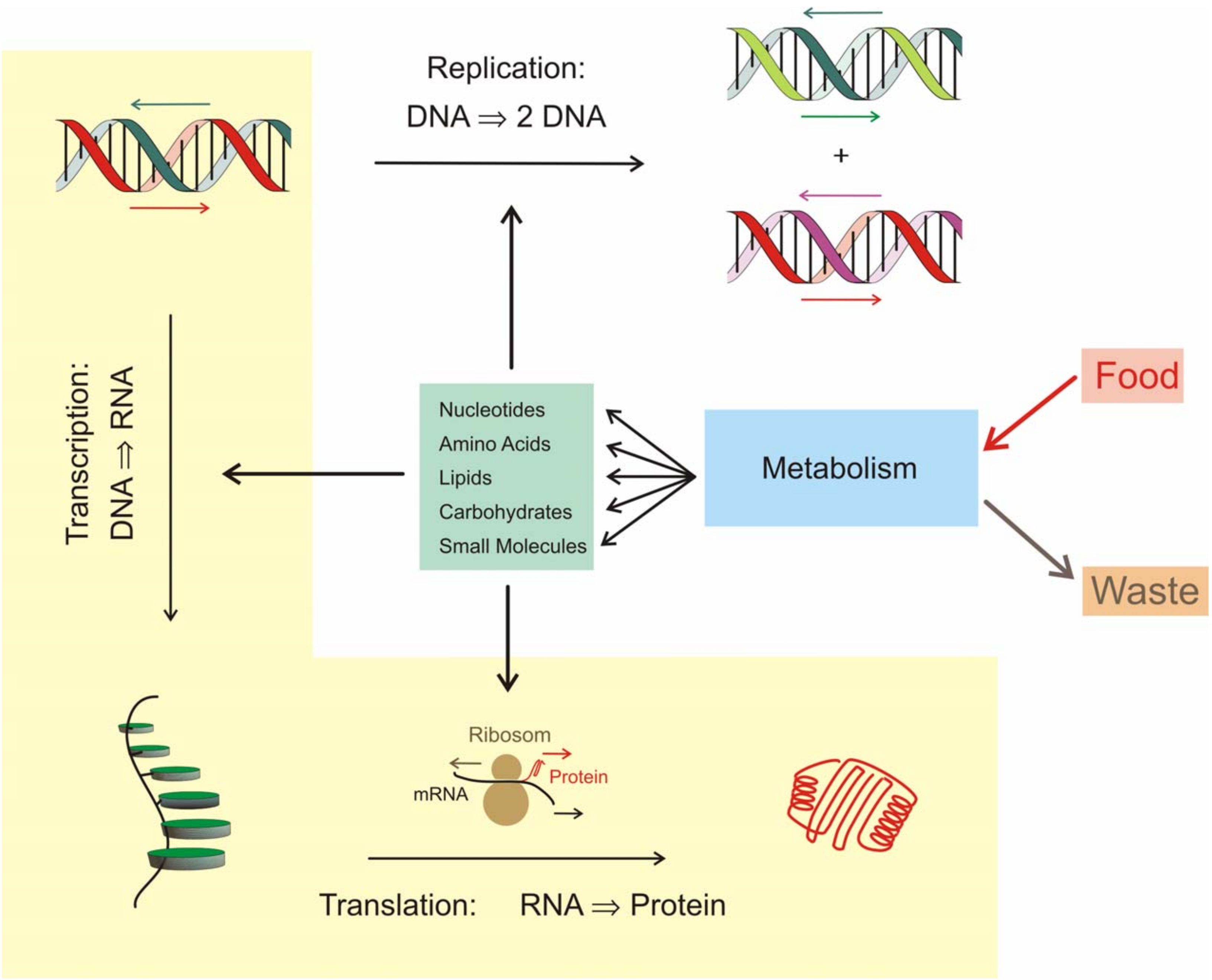
Randomly chosen initial structure



Phenylalanyl-tRNA as target structure



In silico optimization in the flow reactor: Evolutionary Trajectory



Replication:
DNA ⇒ 2 DNA

Transcription:
DNA ⇒ RNA

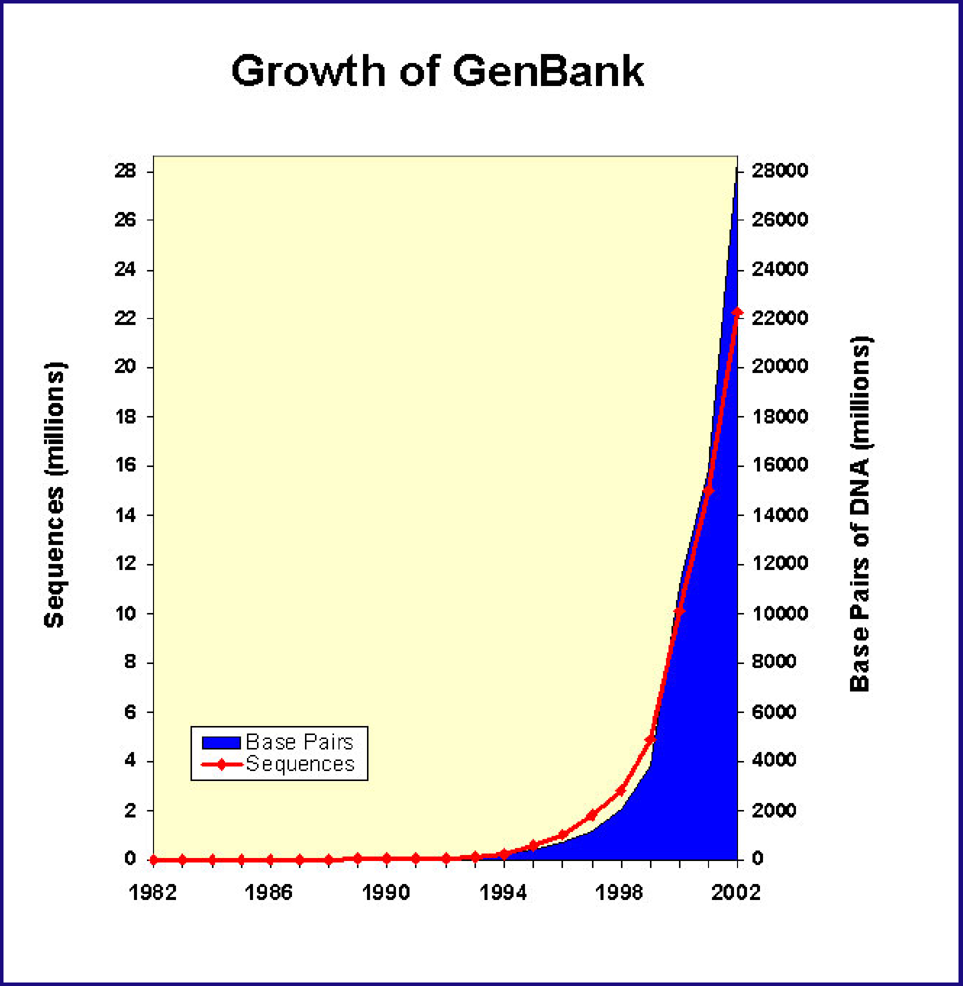
Nucleotides
Amino Acids
Lipids
Carbohydrates
Small Molecules

Metabolism

Food

Waste

Translation: RNA ⇒ Protein



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*

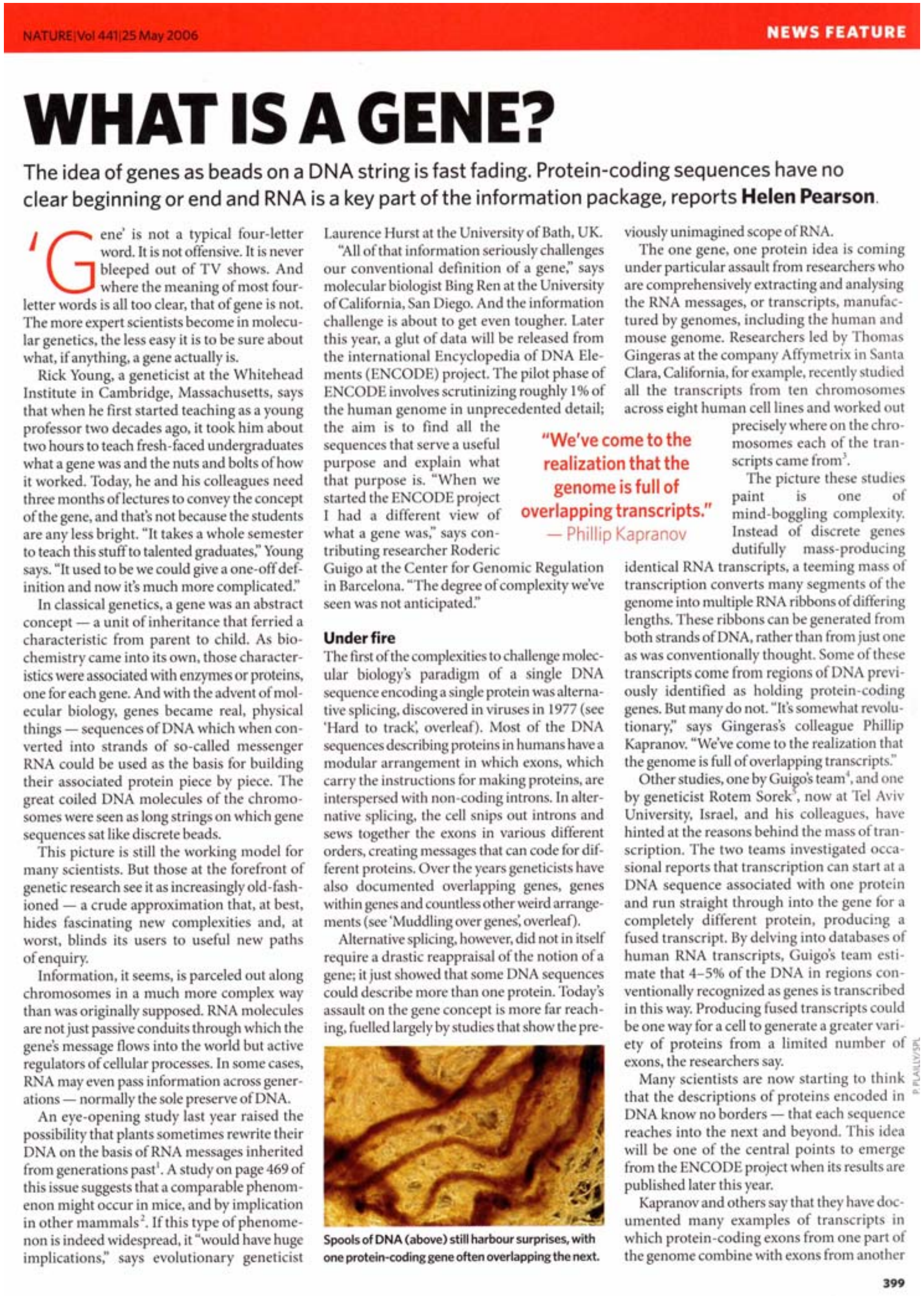
68 phages
1328 viruses
35 viroids
472 organelles (423 mitochondria, 32 plastids, 14 plasmids, 3 nucleomorphs)

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

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 - 60 000

The difficulty defining the gene

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



NATURE|Vol 441|25 May 2006

NEWS FEATURE

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.

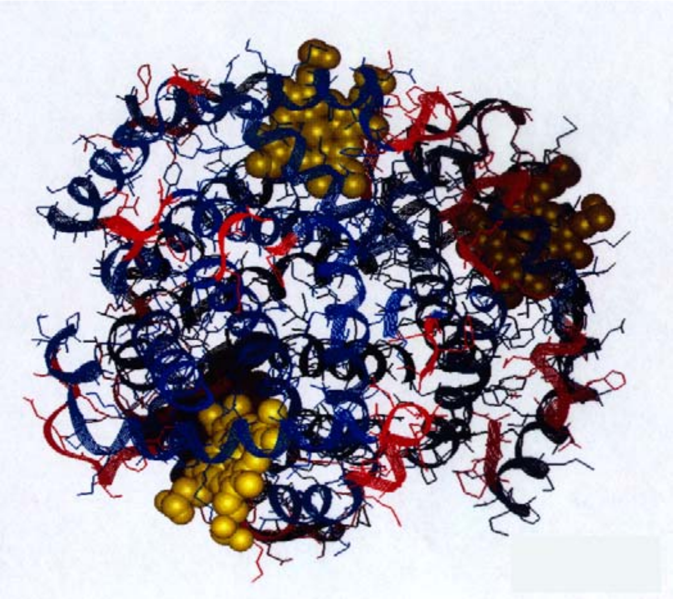
Kapranov and others say that they have documented many examples of transcripts in which protein-coding exons from one part of the genome combine with exons from another

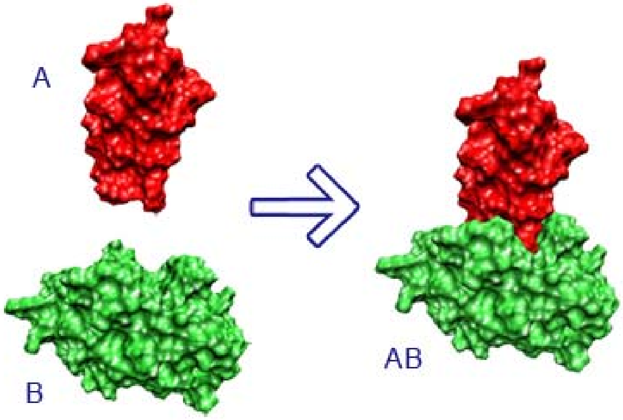
399

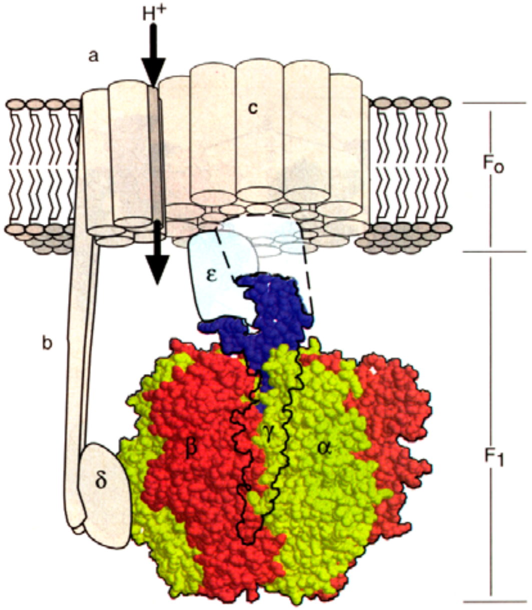
Structural biology

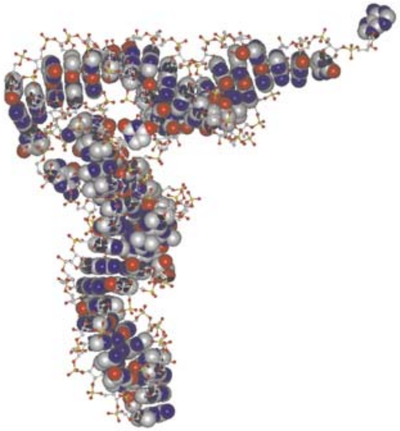
Proteins, nucleic acids, supramolecular complexes,

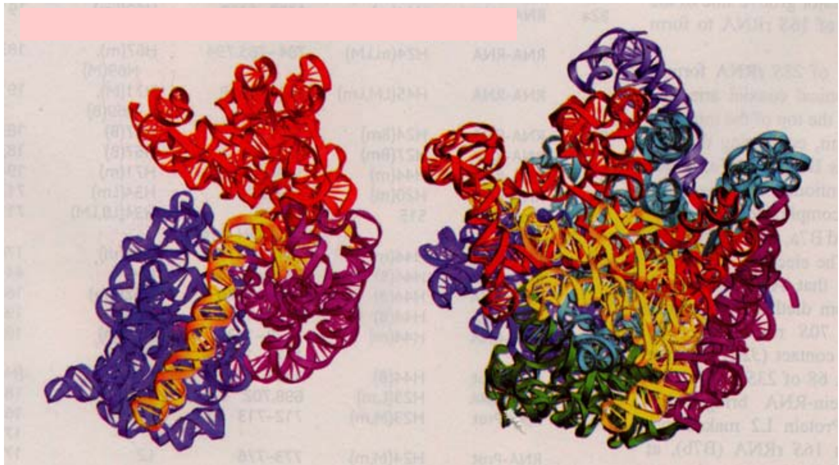
molecular machines

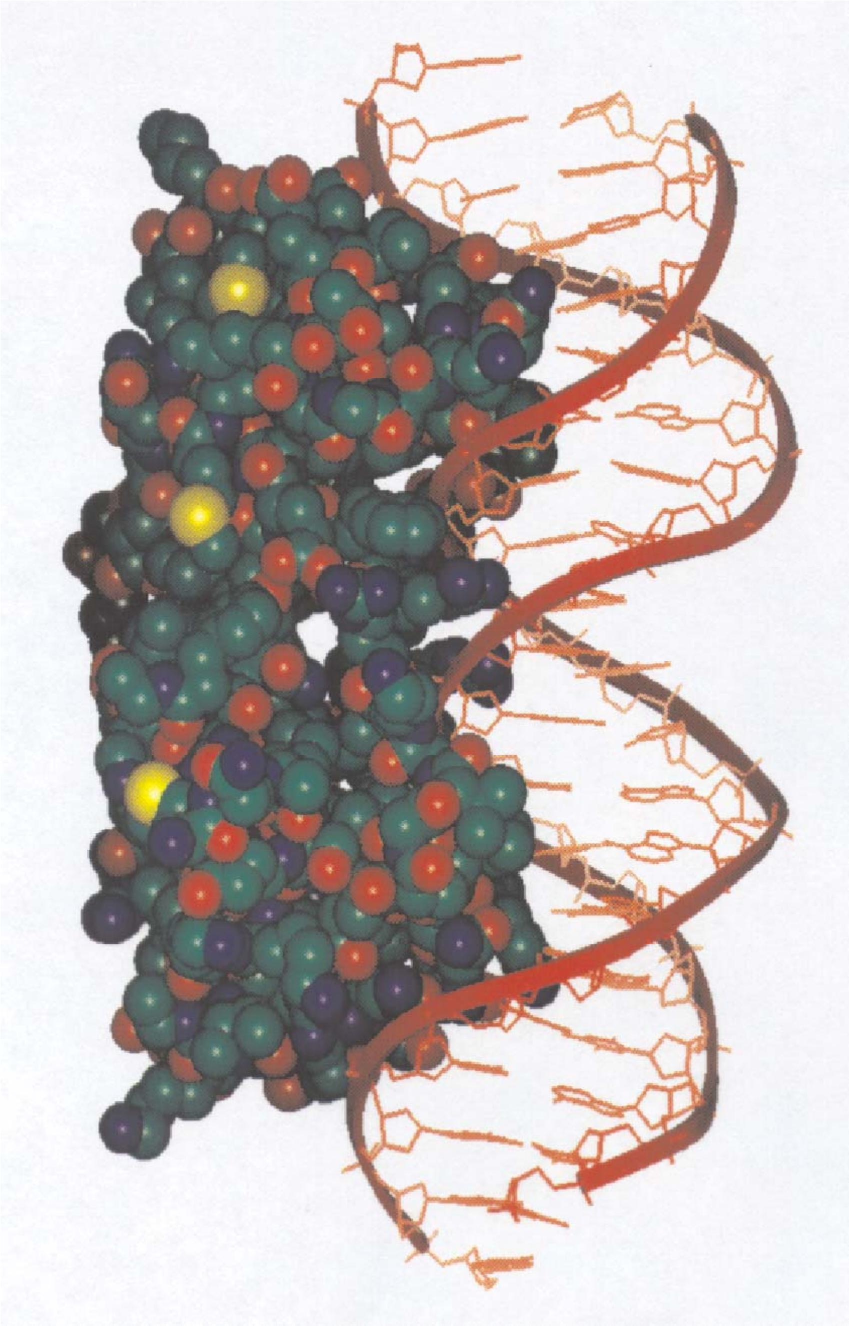




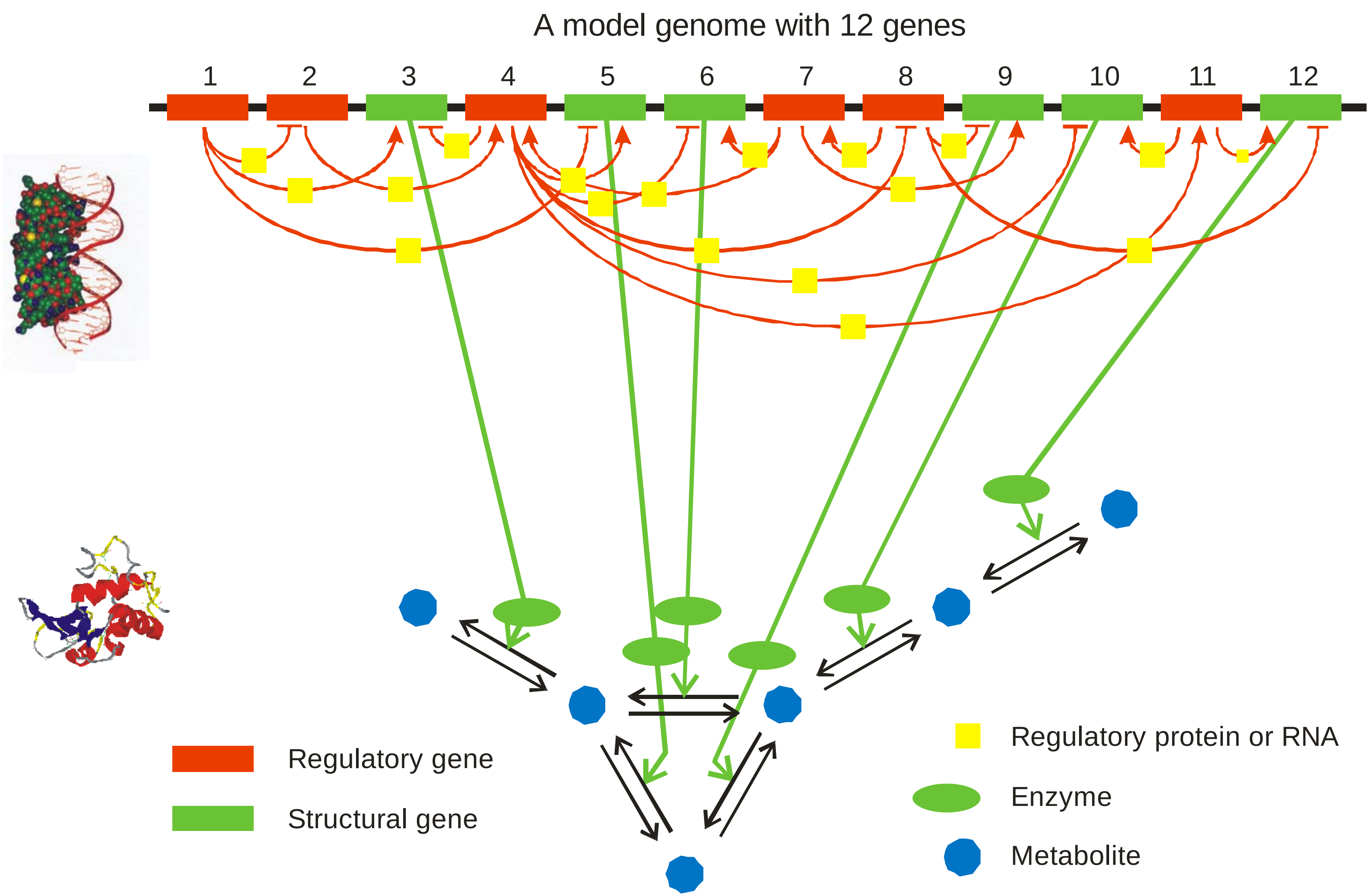








Three-dimensional structure of the complex between the regulatory protein **cro-repressor** and the binding site on λ-phage **B-DNA**



A model genome with 12 genes

🟥 Regulatory gene

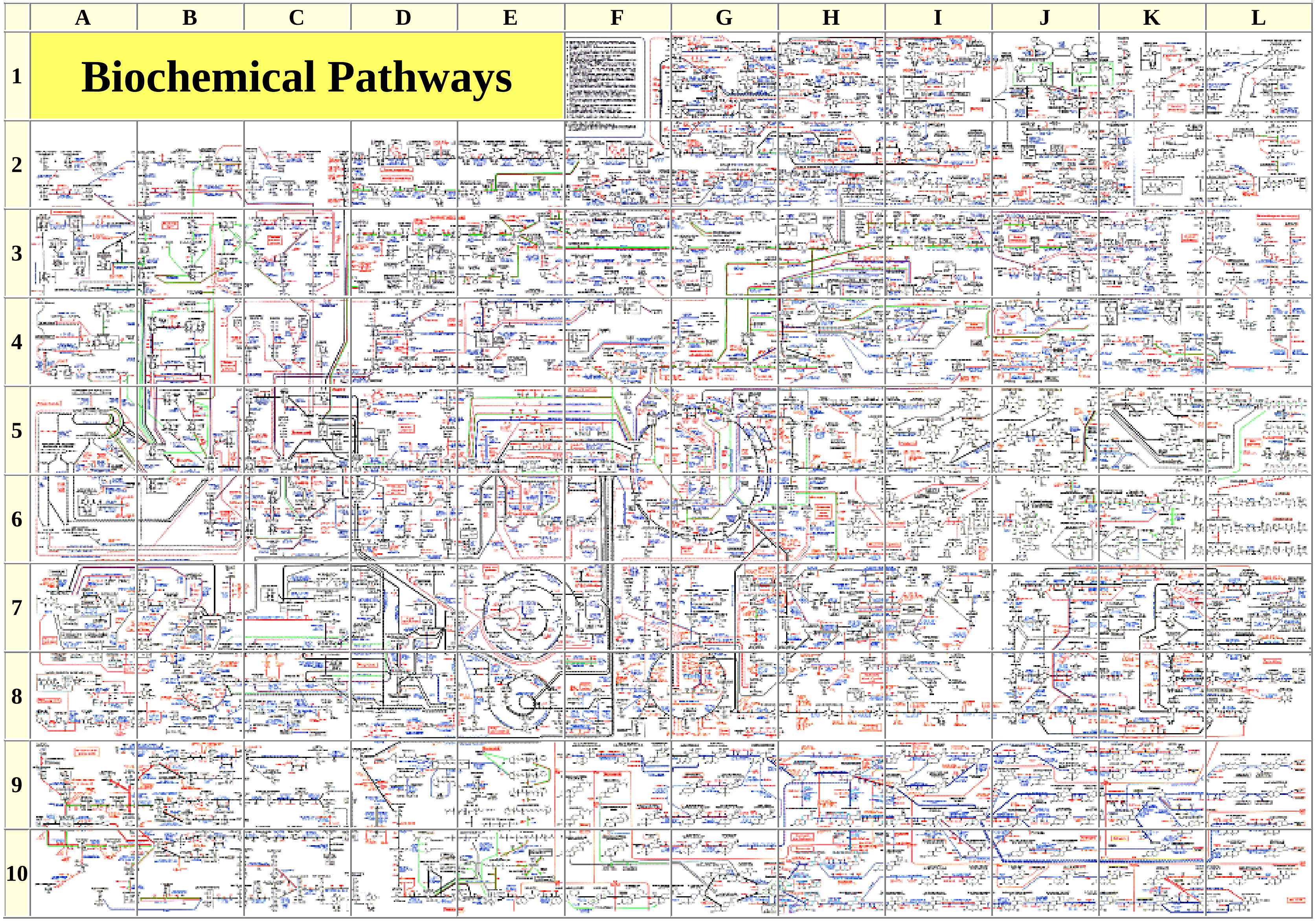
🟩 Structural gene

🟨 Regulatory protein or RNA

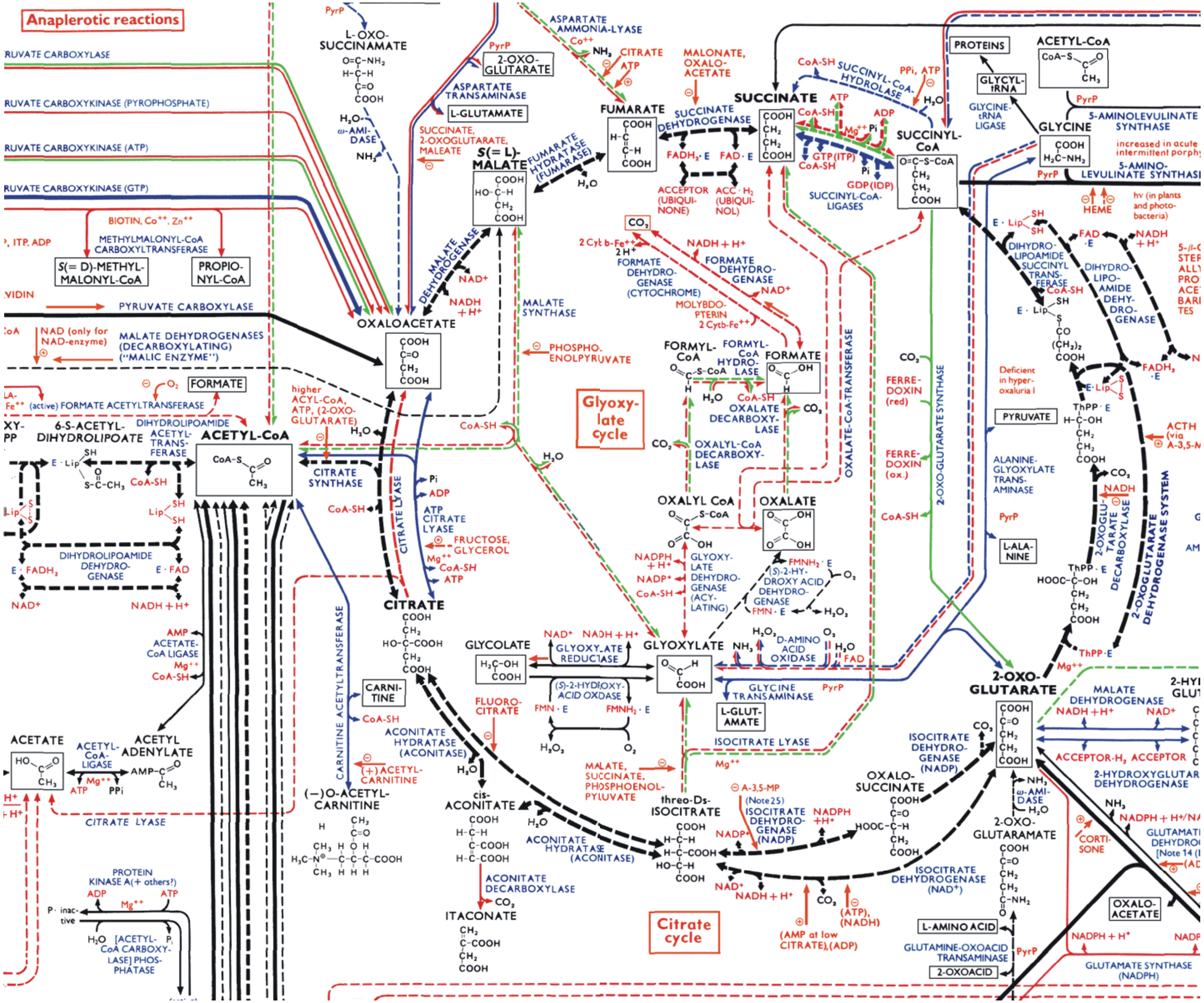
⬬ Enzyme

● Metabolite

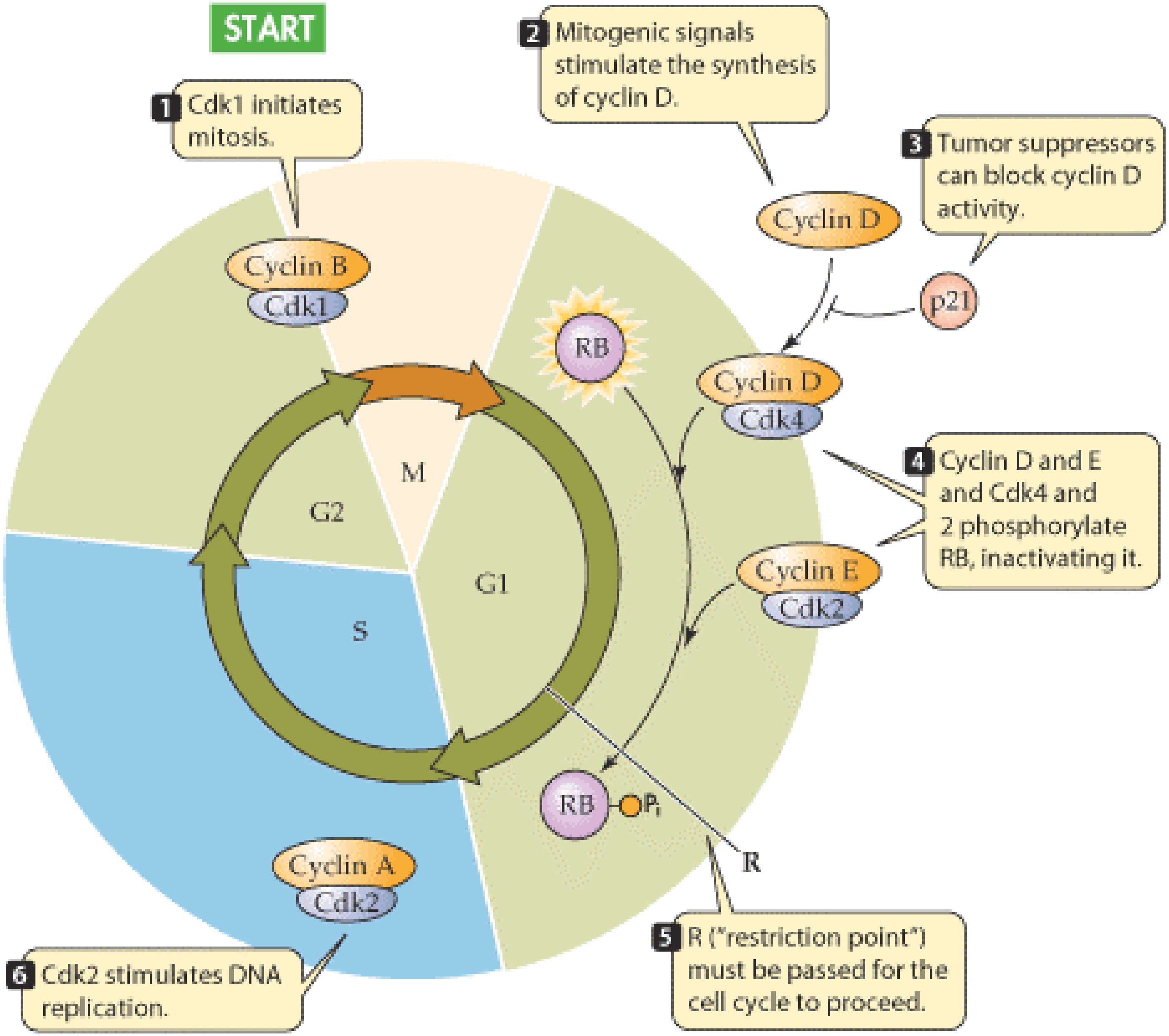
Sketch of a genetic and metabolic network



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



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



1 Cdk1 initiates mitosis.

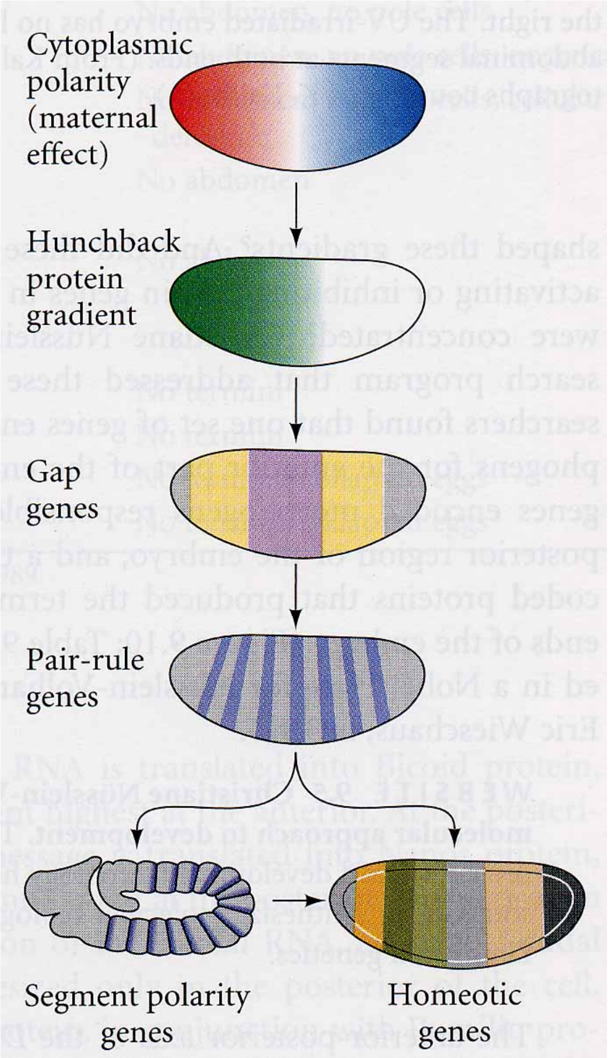
2 Mitogenic signals stimulate the synthesis of cyclin D.

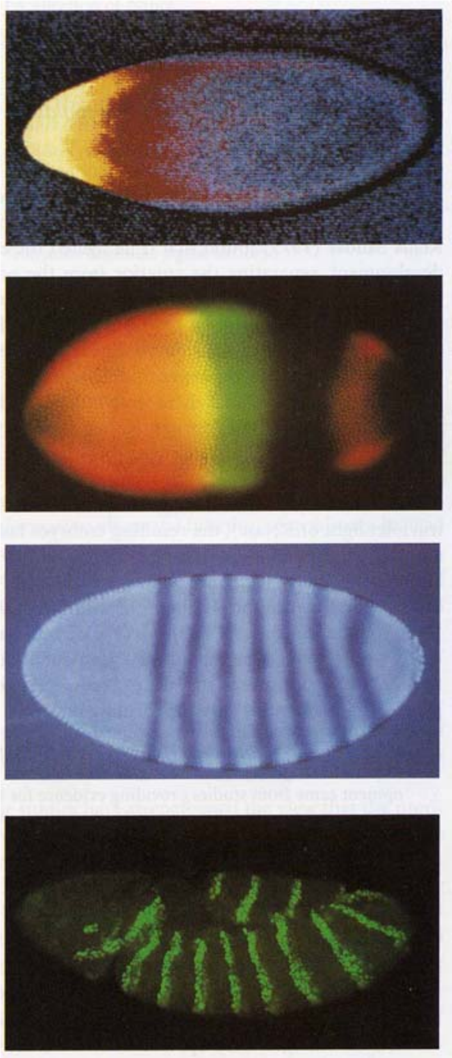
3 Tumor suppressors can block cyclin D activity.

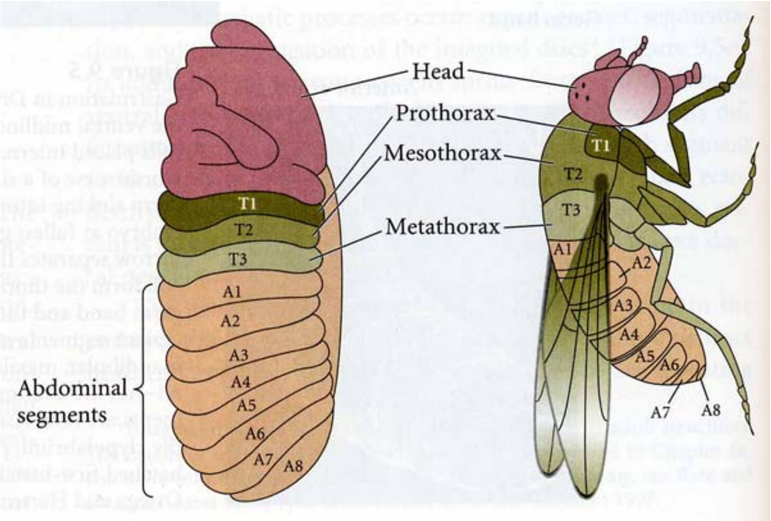
4 Cyclin D and E and Cdk4 and 2 phosphorylate RB, inactivating it.

5 R ("restriction point") must be passed for the cell cycle to proceed.

6 Cdk2 stimulates DNA replication.


Cytoplasmic polarity (maternal effect)
Hunchback protein gradient
Gap genes
Pair-rule genes
Segment polarity genes
Homeotic genes





Cascades, A ⇒ B ⇒ C ⇒ ... , and networks of genetic control

Turing pattern resulting from reaction-diffusion equation ?

Intercelluar communication creating positional information

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

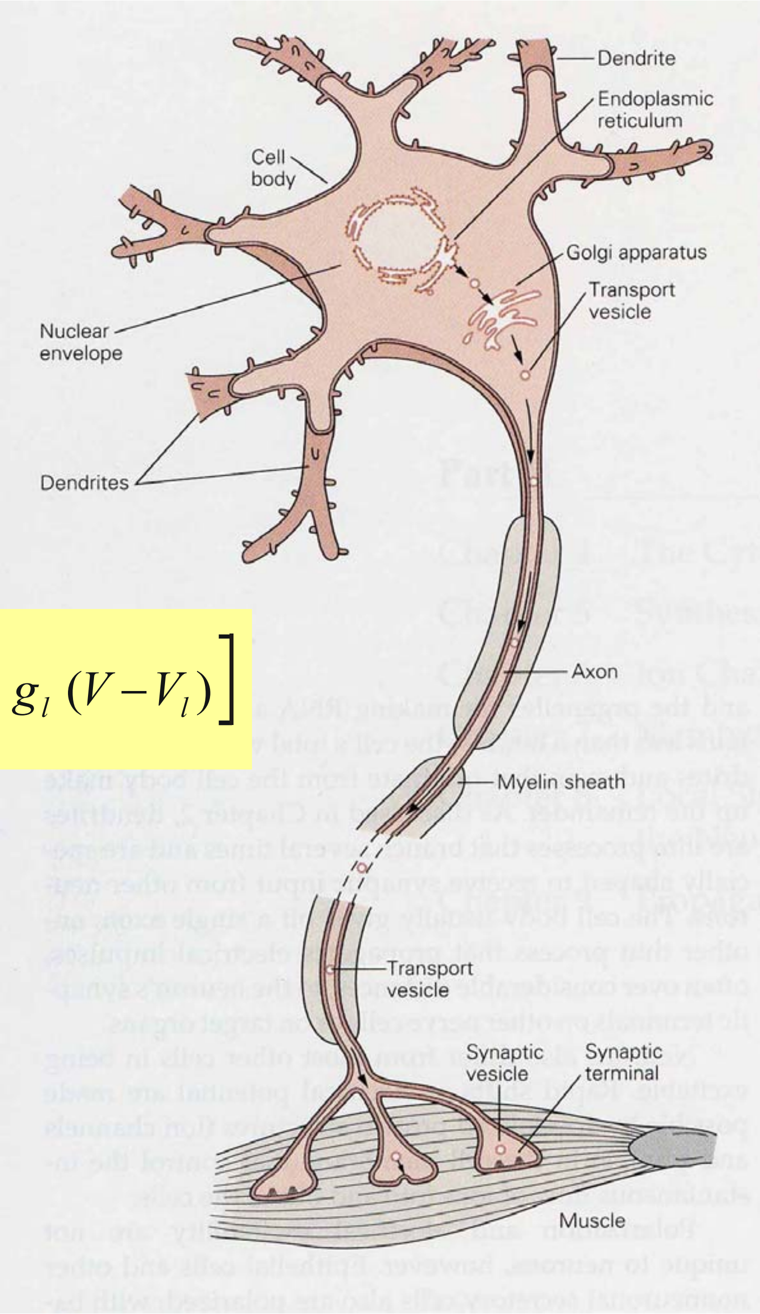
$$\frac{dV}{dt} = \frac{1}{C_M}\Big[I - g_{Na}\, m^3\, h\,(V - V_{Na}) - g_K\, n^4\,(V - V_K) - g_l\,(V - V_l)\Big]$$

$$\frac{dm}{dt} = \alpha_m\,(1-m) - \beta_m\, m$$

$$\frac{dh}{dt} = \alpha_h\,(1-h) - \beta_h\, h$$

$$\frac{dn}{dt} = \alpha_n\,(1-n) - \beta_n\, n$$

Hogdkin-Huxley OD equations



A single neuron signaling to a muscle fiber

$$\frac{1}{R}\frac{\partial^2 V}{\partial x^2} = C\frac{\partial V}{\partial t} + [g_{Na}\, m^3\, h\,(V - V_{Na}) + g_K\, n^4\,(V - V_K) + g_l\,(V - V_l)]\, 2\pi\, r\, L$$

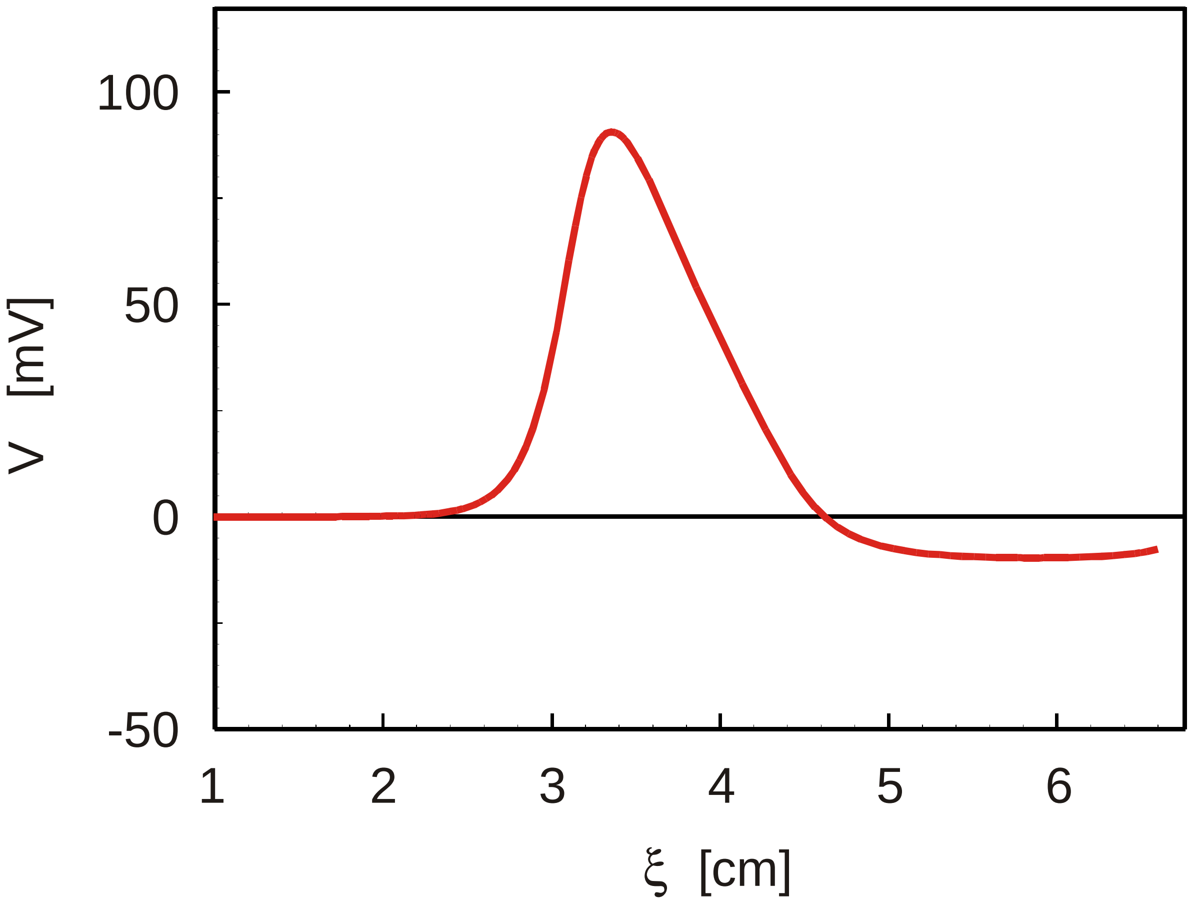
$$\frac{\partial m}{\partial t} = \alpha_m (1-m) - \beta_m m$$

$$\frac{\partial h}{\partial t} = \alpha_h (1-h) - \beta_h h$$

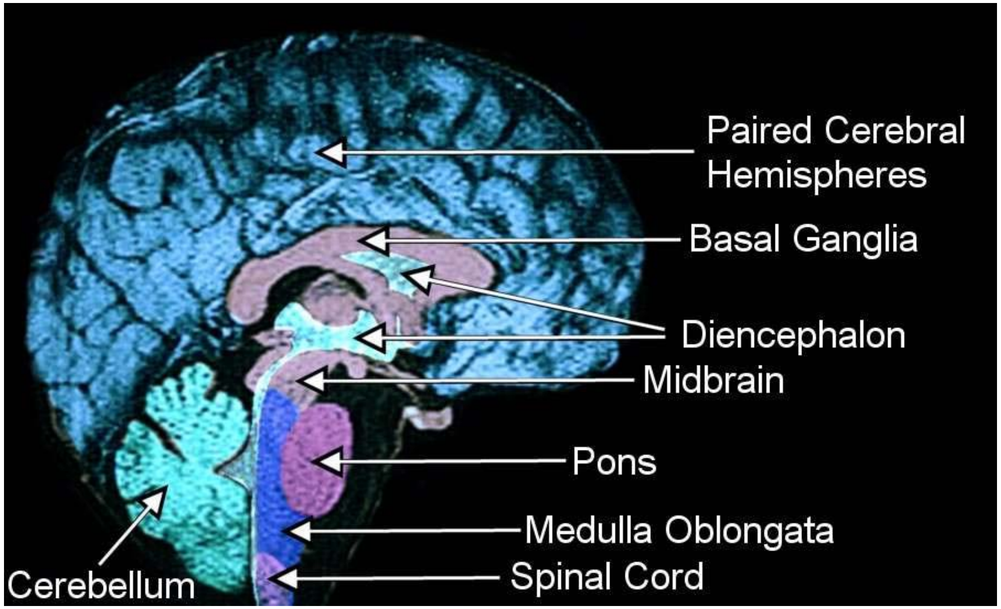
$$\frac{\partial n}{\partial t} = \alpha_n (1-n) - \beta_n n$$

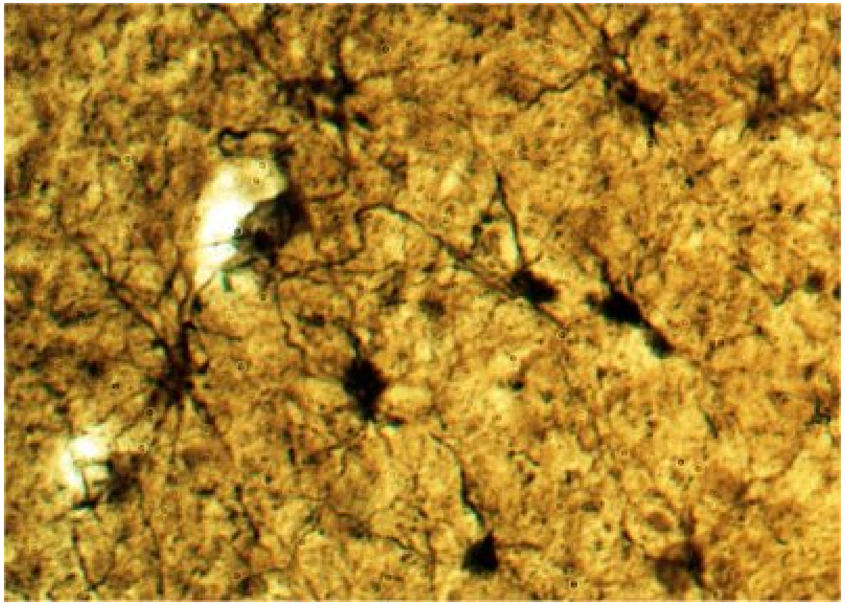
Hodgkin-Huxley partial differential equations (PDE)

Hodgkin-Huxley equations describing pulse propagation along nerve fibers



T = 18.5 C; θ = 1873.33 cm / sec





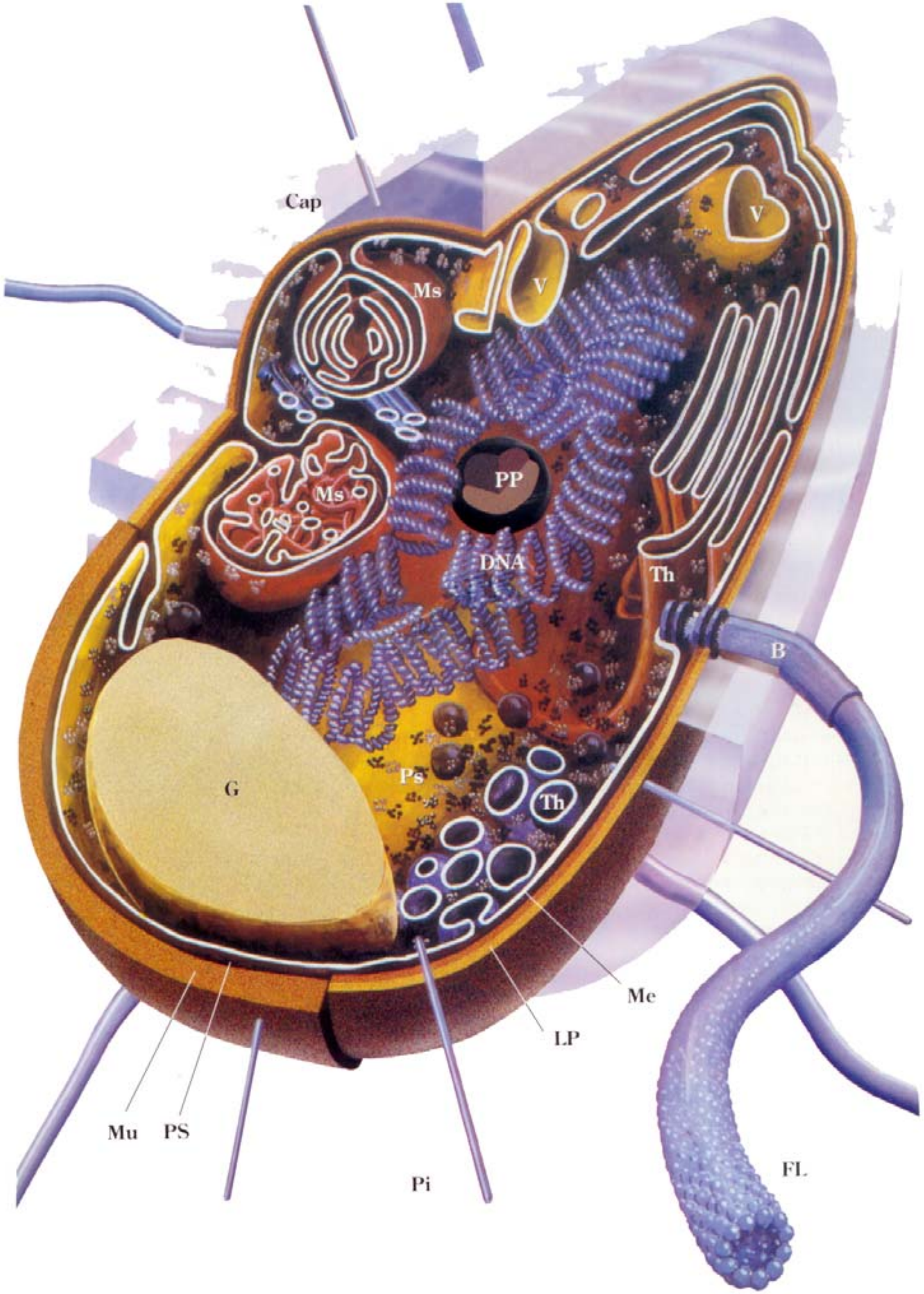
The human brain

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

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

The human body:

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



The spatial structure of the bacterium *Escherichia coli*

Im Restaurant des Nordwestbahnhofs verzehrte ich noch in aller Gemütlichkeit Jungschweinsbraten mit Kraut und Erdäpfel und trank einige Gläser Bier dazu. **Mein Zahlengedächtnis**, sonst erträglich fix, **behält die Zahl der Biergläser stets schlecht**.

Ludwig Boltzmann und die diskrete Beschreibung der Natur.

Ludwig Boltzmann, *Reise eines deutschen Professors ins Eldorado*. Aus *Ludwig Boltzmann, Populäre Schriften*. Eingeleitet und herausgegeben von Engelbert Broda. Friedrich Vieweg & Sohn, Braunschweig 1979, p.258.

Web-Page for further information:

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