The Advantage of Using Mathematics in Biology

Peter Schuster

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

Erwin Schrödinger-Institut

Wien, 15.04.2008

Web-Page for further information:

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

1. Fibonacci – Rabbits, plants, and the golden ratio
2. Mendel – Colors, genes, and inheritance
3. Fisher – Synthesis of genetics and Darwinian evolution
4. Turing – The origin of patterns
5. Hodgkin and Huxley – Neurons and PDEs
6. Error thresholds and antiviral strategies
7. Neutral networks – How evolution works

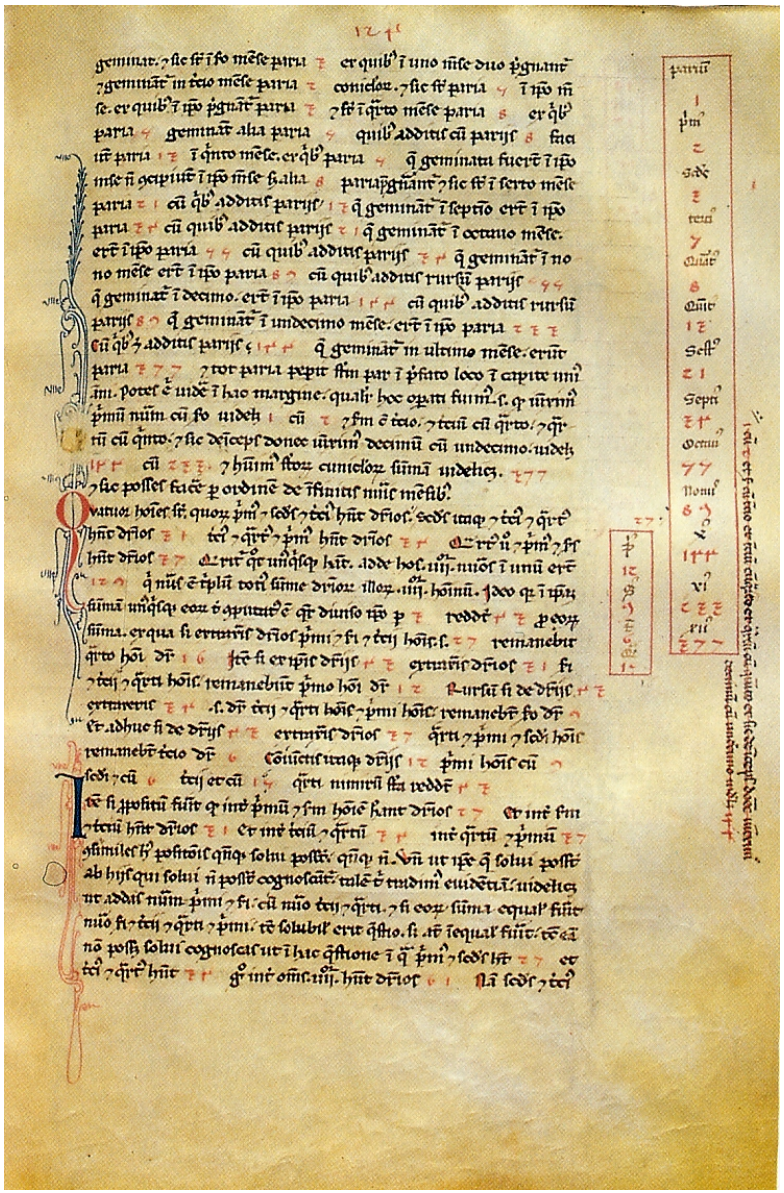
1. **Fibonacci – Rabbits, plants, and the golden ratio**
2. Mendel – Colors, genes, and inheritance
3. Fisher – Synthesis of genetics and Darwinian evolution
4. Turing – The origin of patterns
5. Hodgkin and Huxley – Neurons and PDEs
6. Error thresholds and antiviral strategies
7. Neutral networks – How evolution works

$$F_n := \begin{cases} 0 & \text{if } n = 0\,; \\ 1 & \text{if } n = 1\,; \\ F_{n-1} + F_{n-2} & \text{if } n > 1\,. \end{cases}$$

$$F_n = 0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, \ldots, \text{ for}$$
$$n = 0, 1, 2, 3, 4, 5, 6, \ 7, \ 8, \ 9, 10, \ldots\ .$$

Leonardo da Pisa
„Fibonacci“ – Filius Bonacci
~1180 – ~1240

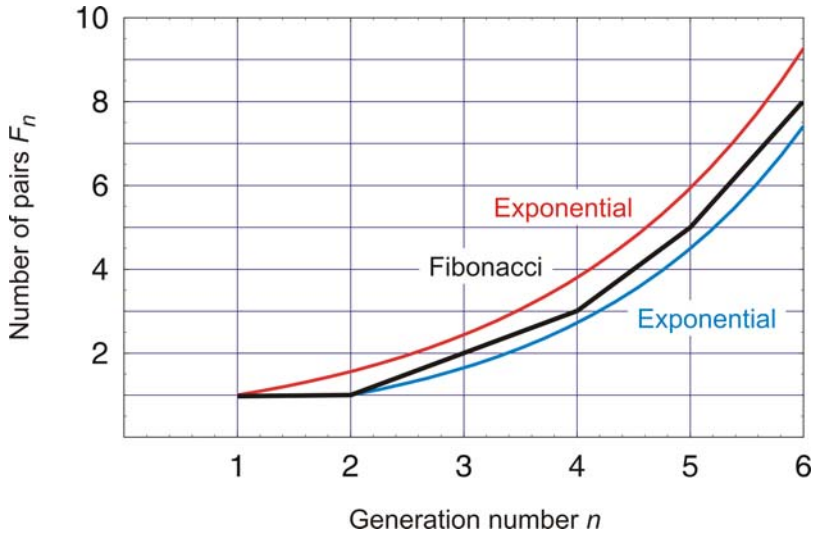
The Fibonacci numbers

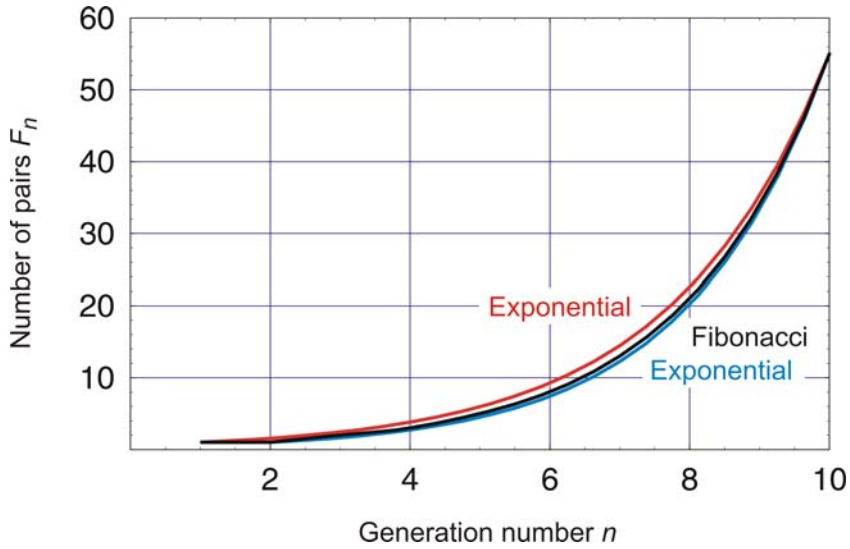


generation	1	2	3	4	5	6
# pairs	1	1	2	3	5	8

The Fibonacci numbers

Bodo Werner, Universität Hamburg, 2006





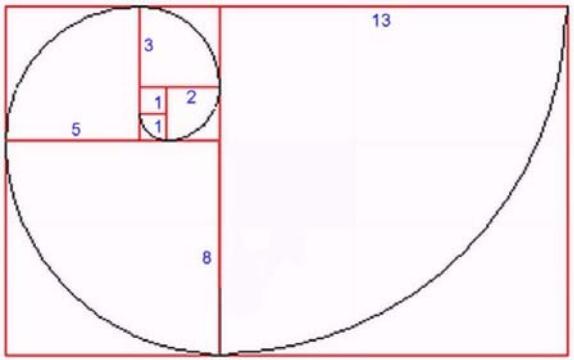
$$f_{\text{upper}}(n) = \exp\Big(0.445259 \cdot (n-1)\Big)$$

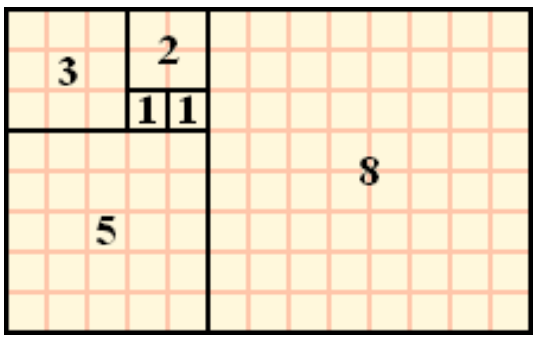
$$f_{\text{lower}}(n) = \exp\Big(0.500917 \cdot (n-2)\Big)$$

$$\lim_{n\to\infty} \frac{F_{n+1}}{F_n} = = \frac{1}{2}(1+\sqrt{5}) = 1.61803\ldots$$

Johannes Kepler (1571-1630)

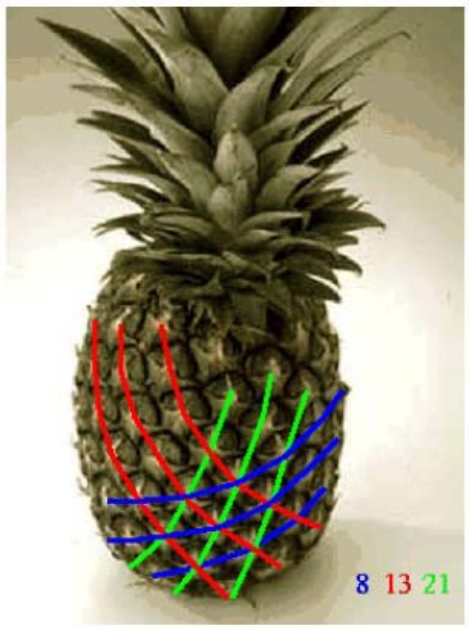
The Fibonacci numbers

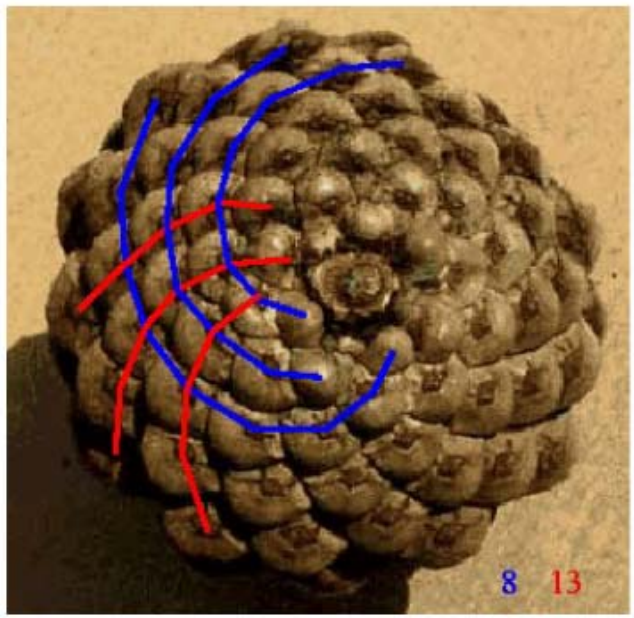




Space filling squares

The Fibonacci spirals



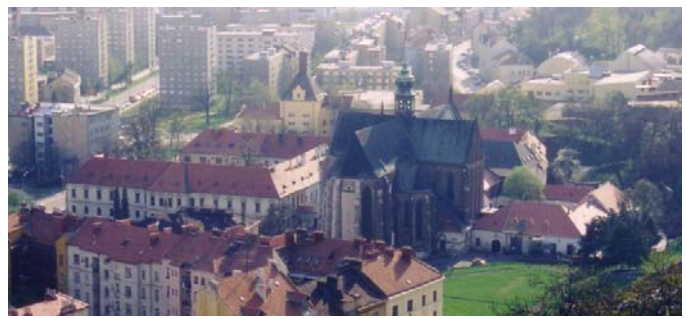


1. Fibonacci – Rabbits, plants, and the golden ratio
2. **Mendel – Colors, genes, and inheritance**
3. Fisher – Synthesis of genetics and Darwinian evolution
4. Turing – The origin of patterns
5. Hodgkin and Huxley – Neurons and PDEs
6. Error thresholds and antiviral strategies
7. Neutral networks – How evolution works

1. flower color is purple or white
2. flower position is axil or terminal
3. stem length is long or short
4. seed shape is round or wrinkled
5. seed color is yellow or green
6. pod shape is inflated or constricted
7. pod color is yellow or green

Gregor Mendel (1882-1884)

```
1st experiment  ⇒  60 fertilizations on 15 plants
2nd experiment  ⇒  58 fertilizations on 10 plants
3rd experiment  ⇒  35 fertilizations on 10 plants
4th experiment  ⇒  40 fertilizations on 10 plants
5th experiment  ⇒  23 fertilizations on  5 plants
6th experiment  ⇒  34 fertilizations on 10 plants
7th experiment  ⇒  37 fertilizations on 10 plants
```



Gregor Mendel‘s experiments on plant genetics

Versuche über Pflanzen-Hybriden. *Verhandlungen des naturforschenden Vereines in Brünn* **4**: 3–47, 1866.
Über einige aus künstlicher Befruchtung gewonnenen Hieracium-Bastarde. *Verhandlungen des naturforschenden Vereines in Brünn* **8**: 26–31, 1870.

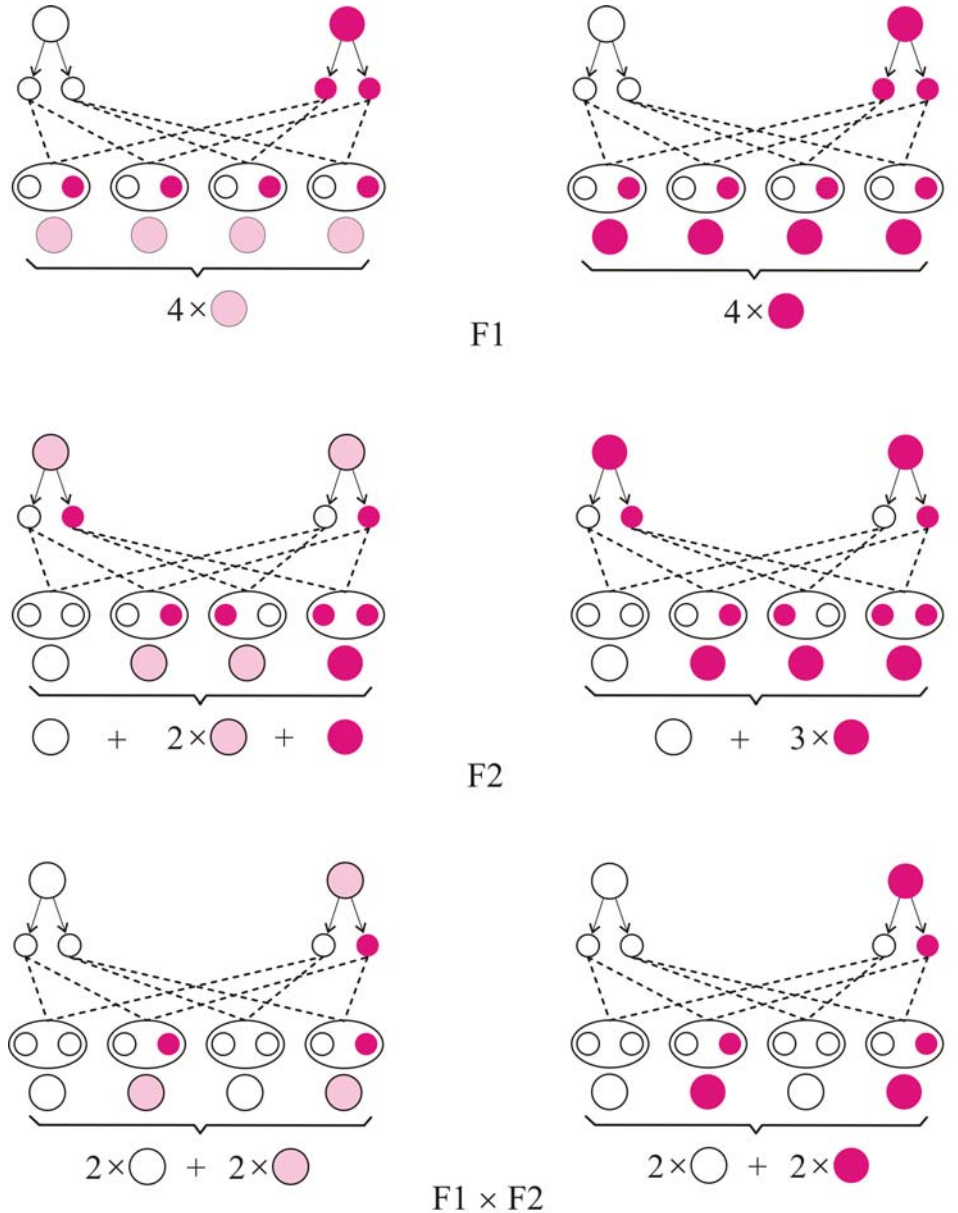
	Experiment 1		Experiment 2	
	Form of Seed		Color of Albumen	
Plants	Round	Angular	Yellow	Green
1	45	12	25	11
2	27	8	32	7
3	24	7	14	5
4	19	10	70	27
5	32	11	24	13
6	26	6	20	6
7	88	24	32	13
8	22	10	44	9
9	28	6	50	14
10	25	7	44	18

- Expt. 1: Form of seed. From 253 hybrids 7324 seeds were obtained in the second trial year. Among them were 5474 round or roundish ones and 1850 angular wrinkled ones. Therefrom the ratio **2.96:1** is deduced.
- Expt. 2: Color of albumen.. 258 plants yielded 8023 seeds, 6022 yellow, and 2001 green; their ratio, therefore, is as **3.01:1**.

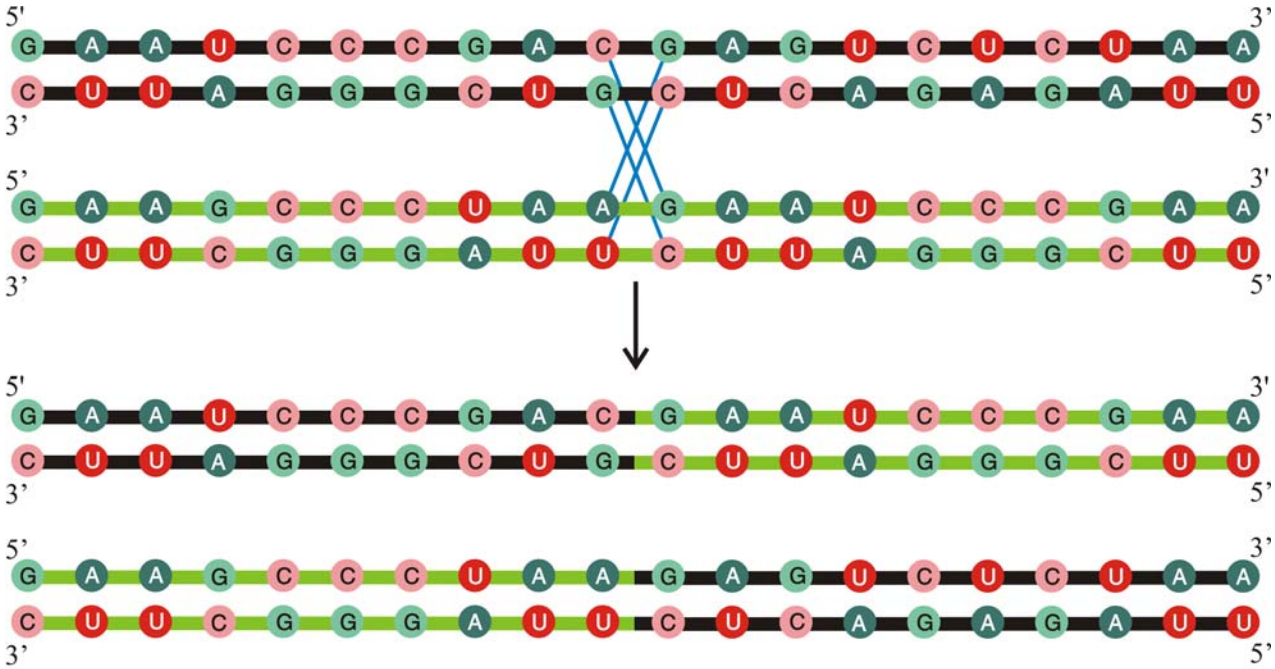
Gregor Mendel concluded correctly from his experiments:

1. that the inheritance of each trait is determined by "units" or "factors" that are passed on to descendents unchanged (these units are now called genes)
2. that an individual inherits one such unit from each parent for each trait
3. that a trait may not show up in an individual but can still be passed on to the next generation.

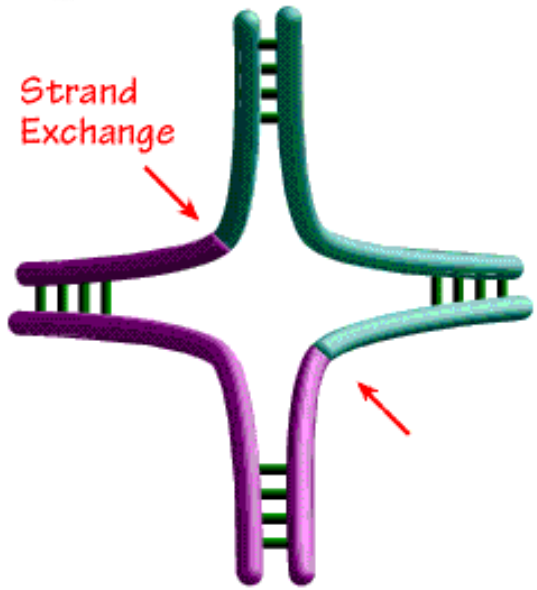
Gregor Mendel‘s experiments on plant genetics



Gregor Mendel‘s experiments
on plant genetics



genetic recombination



Molecular explanation of Mendel‘s expriments – recombination

1. Fibonacci – Rabbits, plants, and the golden ratio
2. Mendel – Colors, genes, and inheritance
3. **Fisher – Synthesis of genetics and Darwinian evolution**
4. Turing – The origin of patterns
5. Hodgkin and Huxley – Neurons and PDEs
6. Error thresholds and antiviral strategies
7. Neutral networks – How evolution works

Ronald Fisher (1890-1962)

alleles: $A_1, A_2, \ldots, A_n$

frequencies: $x_i = [A_i]$; genotype: $A_i \cdot A_k$

Fitness values: $a_{ik} = f(A_i \cdot A_k)$, $a_{ik} = a_{ki}$

$$\frac{dx_i}{dt} = \sum_{j=1}^{n} a_{ij} x_i x_j - x_i \sum_{j=1}^{n} \sum_{k=1}^{n} a_{jk} x_j x_k = x_i \left(\sum_{j=1}^{n} a_{ij} x_j - \Phi \right)$$

$$\text{with} \quad \Phi = \sum_{j=1}^{n} \sum_{k=1}^{n} a_{jk} x_j x_k \quad \text{and} \quad \sum_{j=1}^{n} x_j = 1$$

$$\frac{d\Phi}{dt} = 2\left(< \bar{a}^2 > - < \bar{a} >^2 \right) = 2 \,\mathrm{var}\{\bar{a}\} \geq 0$$

Ronald Fisher‘s selection equation

1. Fibonacci – Rabbits, plants, and the golden ratio
2. Mendel – Colors, genes, and inheritance
3. Fisher – Synthesis of genetics and Darwinian evolution
4. **Turing – The origin of patterns**
5. Hodgkin and Huxley – Neurons and PDEs
6. Error thresholds and antiviral strategies
7. Neutral networks – How evolution works

A. M. Turing. The chemical basis of morphogenesis.
Phil.Trans.Roy.Soc. London B **327**:37, 1952

Alan Turing (1912-1954)

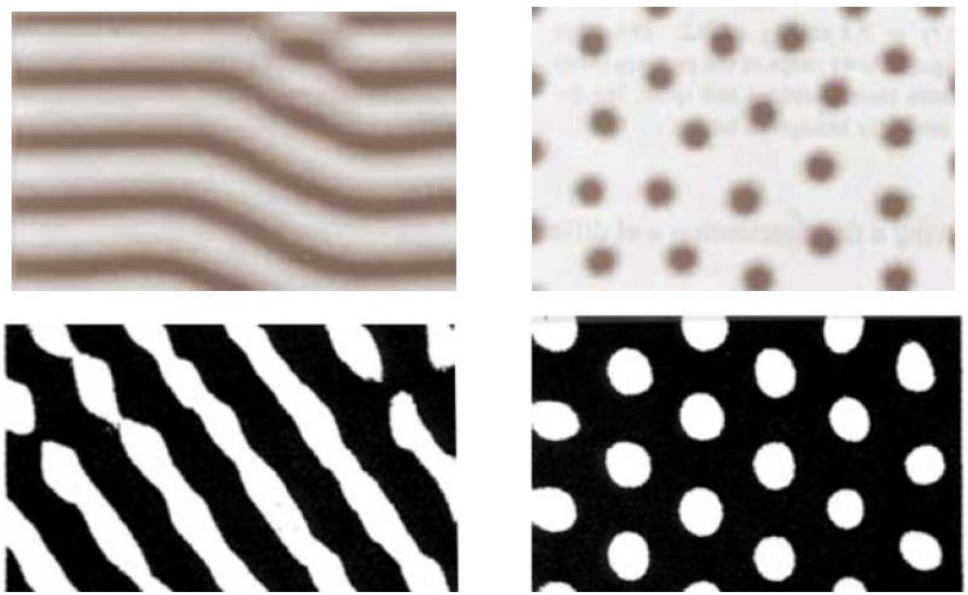
$$\frac{\partial c_i}{\partial t} = D_i \triangle c_i + F_i(c_1, c_2, \ldots, c_n)\,;\; i = 1, 2, \ldots, n$$

D_i ... diffusion coefficient of substance "i"

Spontaneous pattern formation in reaction diffusion equations

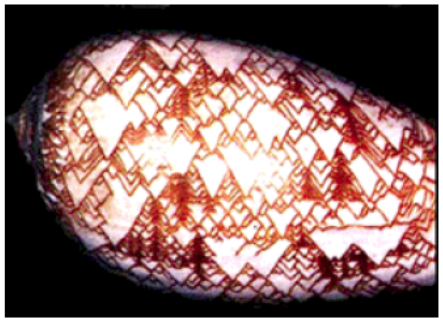
Boris Belousov and Anatol Zhabotinskii

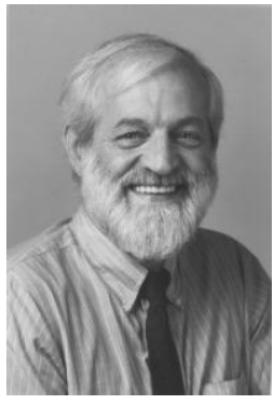
Boris Belousov



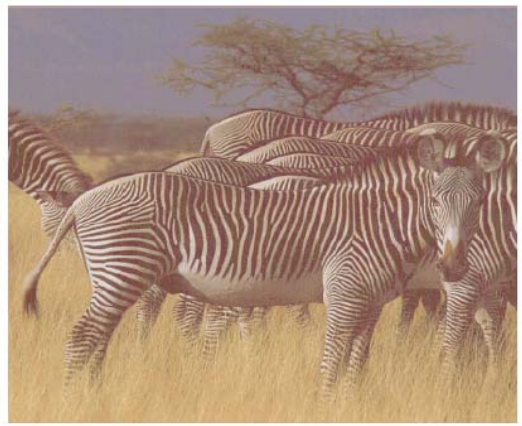
Vincent Castets, Jacques Boissonade, Etiennette Dulos and Patrick DeKepper, Phys.Rev. Letters 64:2953, 1990

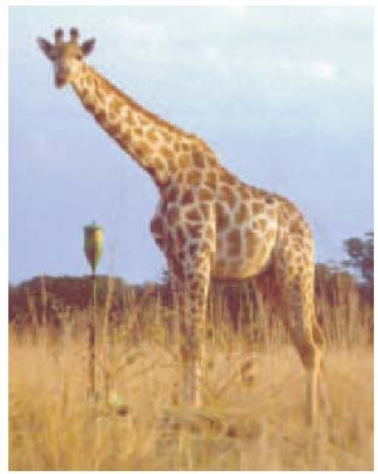
Experimental verification of Turing patterns in chemical reactions

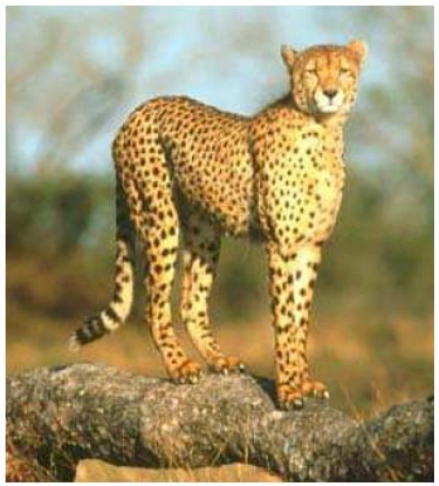


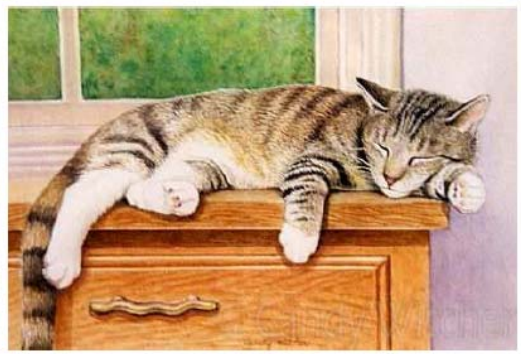


James D. Murray Hans Meinhardt Alfred Gierer



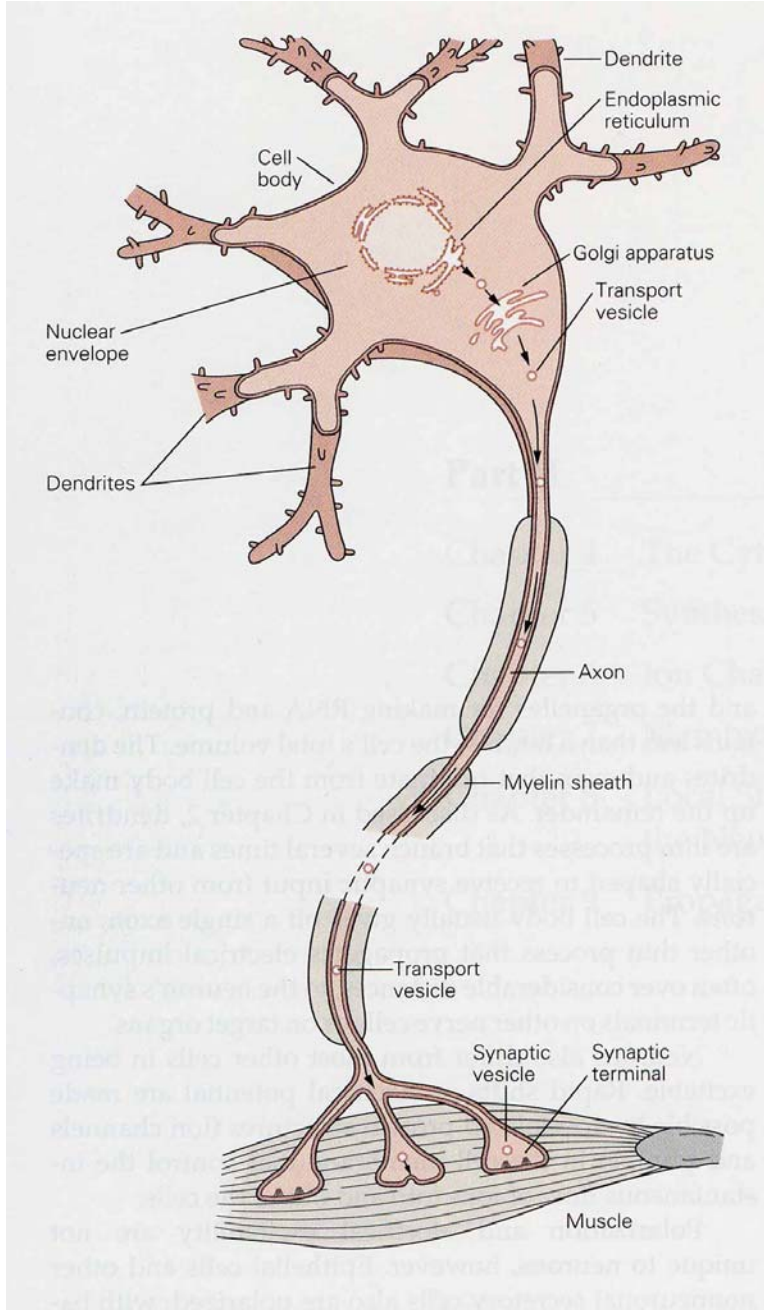




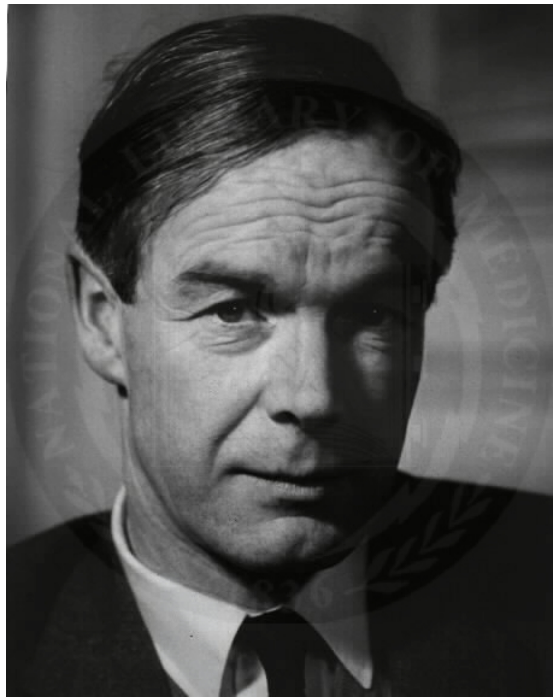


Turing patterns
on animal skins
and shells

1. Fibonacci – Rabbits, plants, and the golden ratio
2. Mendel – Colors, genes, and inheritance
3. Fisher – Synthesis of genetics and Darwinian evolution
4. Turing – The origin of patterns
5. **Hodgkin and Huxley – Neurons and PDEs**
6. Error thresholds and antiviral strategies
7. Neutral networks – How evolution works



A single neuron signaling to a muscle fiber



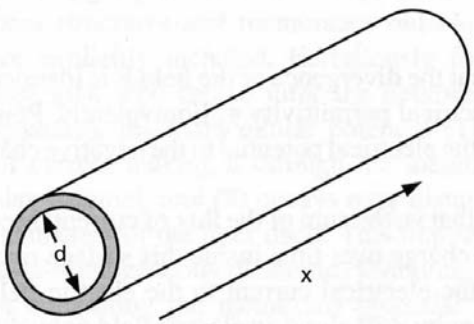
Alan Hodgkin

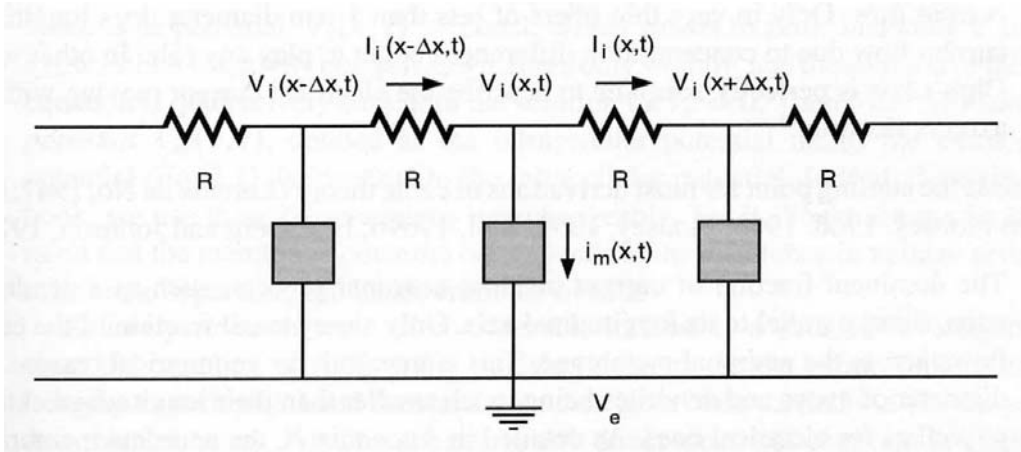
Andrew Huxley

A. L. Hodgkin and A. F. Huxley. *A Quantitative Description of Membrane Current and its Application to Conduction and Excitation in Nerve*.

Journal of Physiology **117**: 500-544, 1952

The Hodgkin-Huxley equation





A **B**

Fig. 2.2 ELECTRICAL STRUCTURE OF A CABLE **(A)** Idealized cylindrical axon or dendrite at the heart of one-dimensional cable theory. Almost all of the current inside the cylinder is longitudional due to geometrical (the radius is much smaller than the length of the cable) and electrical factors (the membrane covering the axon or dendrite possesses a very high resistivity compared to the intracellular cytoplasm). As a consequence, the radial and angular components of the current can be neglected, and the problem of determining the potential in these structures can be reduced from three spatial dimensions to a single one. On the basis of the bidomain approximation, gradients in the extracellular potentials are neglected and the cable problem is expressed in terms of the transmembrane potential $V_m(x,t) = V_i(x,t) - V_e$. **(B)** Equivalent electrical structure of an arbitrary neuronal process. The intracellular cytoplasm is modeled by the purely ohmic resistance R. This tacitly assumes that movement of carriers is exclusively due to drift along the voltage gradient and not to diffusion. Here and in the following the extracellular resistance is assumed to be negligible and V_e is set to zero. The current per unit length across the membrane, whether it is passive or contains voltage-dependent elements, is described by i_m and the system is characterized by the second-order differential equation, Eq. 2.5.

Christof Koch, Biophysics of Computation. Information Processing in single neurons.
Oxford University Press, New York 1999.

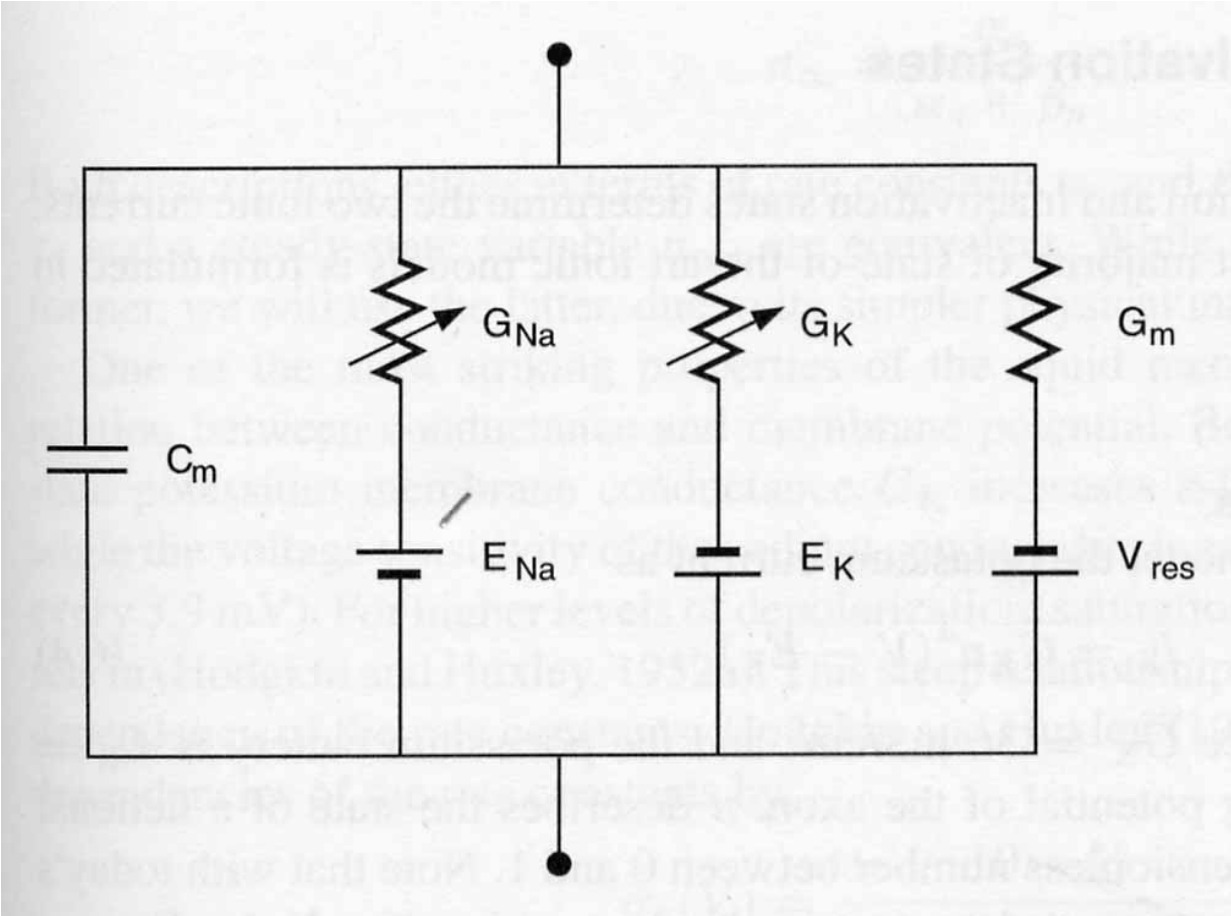


Fig. 6.2 Electrical Circuit for a Patch of Squid Axon Hodgkin and Huxley modeled the membrane of the squid axon using four parallel branches: two passive ones (membrane capacitance C_m and the leak conductance $G_m = 1/R_m$) and two time- and voltage-dependent ones representing the sodium and potassium conductances.

Christof Koch, Biophysics of Computation. Information Processing in single neurons. Oxford University Press, New York 1999.

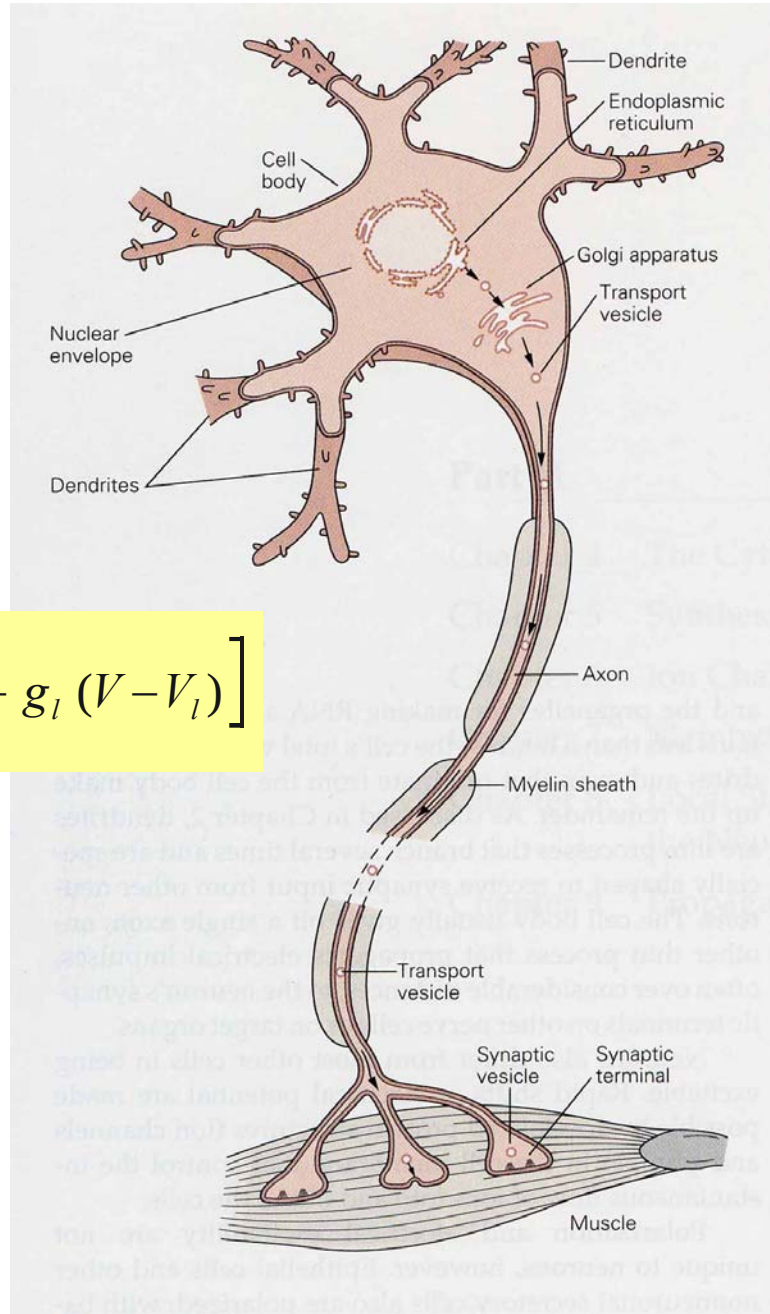
$$\frac{dV}{dt} = \frac{1}{C_M}\left[I - g_{Na}\, m^3\, h\,(V - V_{Na}) - g_K\, n^4\,(V - V_K) - g_l\,(V - V_l)\right]$$

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

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

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

Hogdkin-Huxley OD equations



A single neuron signaling to a muscle fiber

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

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

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

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

Hodgkin-Huxley PDEquations

Travelling pulse solution: $V(x,t) = V(\xi)$ with

$$\xi = x + \theta\, t$$

Hodgkin-Huxley equations describing pulse propagation along nerve fibers

$$\frac{1}{R}\frac{d^2V}{d\xi^2} = C_M\,\theta\frac{dV}{d\xi} + \left[\,\bar{g}_{Na}\,m^3\,h(V-V_{Na}) + \bar{g}_K\,n^4\,(V-V_K) + \bar{g}_l\,(V-V_l)\right]\,2\pi r L$$

$$\theta\,\frac{dm}{d\xi} = \alpha_m\,(1-m) - \beta_m\,m$$

$$\theta\,\frac{dh}{d\xi} = \alpha_h\,(1-h) - \beta_h\,h$$

$$\theta\,\frac{dn}{d\xi} = \alpha_n\,(1-n) - \beta_n\,n$$

Hodgkin-Huxley PDEquations

Travelling pulse solution: $V(x,t) = V(\xi)$ with

$$\xi = x + \theta\,t$$

Hodgkin-Huxley equations describing pulse propagation along nerve fibers

$$\frac{\partial m}{\partial t} = \Theta(T)\big[\alpha_m(1-m) - \beta_m m\big]$$

$$\frac{\partial h}{\partial t} = \Theta(T)\big[\alpha_h(1-h) - \beta_h h\big]$$

$$\frac{\partial n}{\partial t} = \Theta(T)\big[\alpha_n(1-n) - \beta_n n\big] \ ,$$

$$\text{where } \ \Theta(T) = 3^{(T-6.3)/10}$$

Temperature dependence of the Hodgkin-Huxley equations

Systematic investigation of pulse behavior

International Journal of Bifurcation and Chaos, Vol. 16, No. 12 (2006) 3605–3616

ANALYTICAL DYNAMICS OF NEURON PULSE PROPAGATION

PAUL E. PHILLIPSON* and PETER SCHUSTER
Institut für Theoretische Chemie, Universität Wien, Währingerstraße 17, A-1090 Wien, Austria
**Austrian Academy of Sciences, Dr. Ignaz Seipel-Platz 2, A-1010 Wien, Austria*

Received July 12, 2005; Revised January 11, 2006

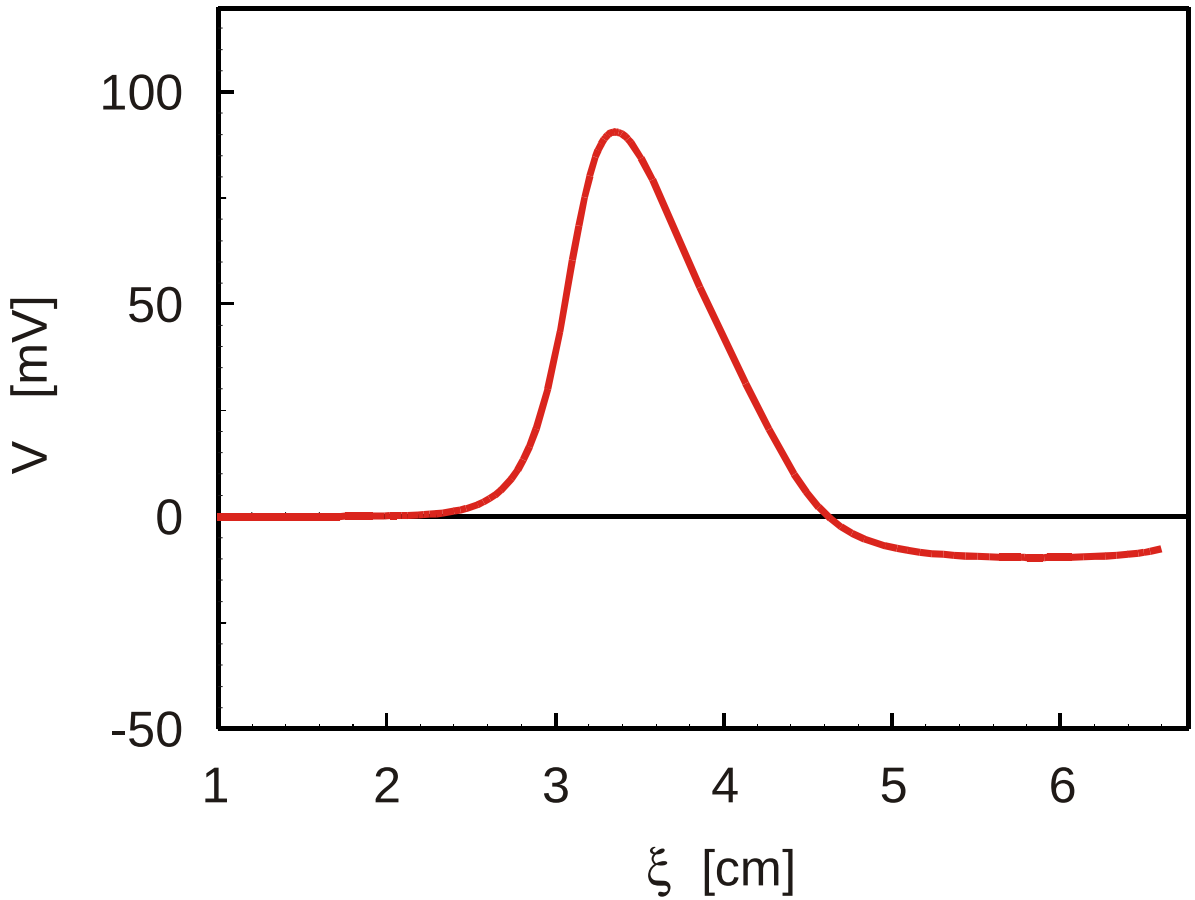
The four-dimensional Hodgkin–Huxley equations describe the propagation in space and time of the action potential $v(z)$ along a neural axon with $z = x + ct$ and c being the pulse speed. The potential $v(z)$, which is parameterized by the temperature, is driven by three gating functions, $m(z)$, $n(z)$ and $h(z)$, each of which obeys formal first order kinetics with rate constants that are represented as nonlinear functions of the potential v. It is shown that this system can be analytically simplified (i) in the number of gating functions and (ii) in the form of associated rate functions while retaining to close approximation quantitative fidelity to computer solutions of the exact equations over the complete temperature range for which stable pulses exist. At a given temperature we record two solutions ($T < T_{max}$) corresponding to a high-speed and a low-speed branch in speed-temperature plots, $c(T)$, or no solution ($T > T_{max}$). The pulse is considered as composed of two contiguous parts: (i) a pulse front extending from $v(0) = 0$ to a pulse maximum $v = V_{max}$, and (ii) a pulse back extending from V_{max} through a pulse minimum V_{min} to a final regression back to $v(z \to \infty) = 0$. An approximate analytic solution is derived for the pulse front, which is predicted to propagate at a speed $c(T) = 1203\,\Theta^{\frac{1}{2}}$ (T°C) cm/sec, $\Theta = 3^{\frac{T-6.3}{10}}$ in close agreement with computer solution of the exact Hodgkin–Huxley equations for the entire pulse. These results provide the basis for a derivation of two-dimensional differential equation systems for the pulse front and pulse back, which predict the pulse maximum and minimum over the operational temperature range $0 \le T \le 25$°C, in close agreement with the exact equations. Most neuron dynamics studies have been based on voltage clamp experiments featuring external current injection in place of self-generating pulse propagation. Since the behaviors of the gating functions are similar, it is suggested that the present approximations might be applicable to such situations as well as to the dynamics of myelinated fibers.

Keywords: Hodgkin–Huxley equations; action potentials; neuron models; nonlinear dynamics; neuron pulse propagation.

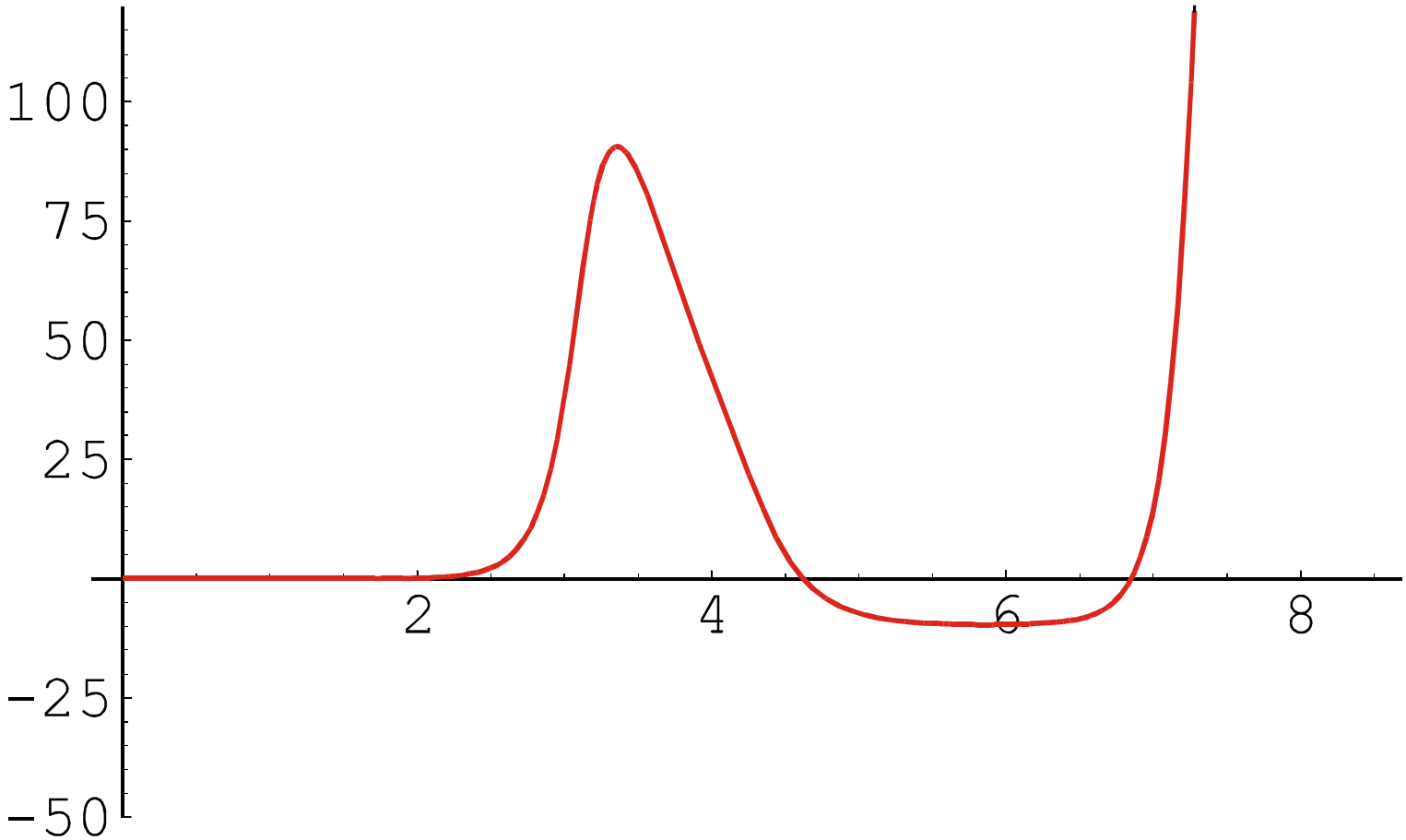
1. Neuron Pulse Propagation and the Hodgkin–Huxley Equations

Conductance mechanisms for the propagation of a pulse along an unmyelinated neural axon were encapsulated within a predictive theory by the equations of Hodgkin and Huxley [1952]. These equations became the prototype for description of neural pulse propagation and provide the basis for all subsequent conduction models of neural behavior. The Hodgkin–Huxley equations relate the propagating action potential v to sodium, potassium and leak conductances I_{Na}, I_K, I_{leak} causing the

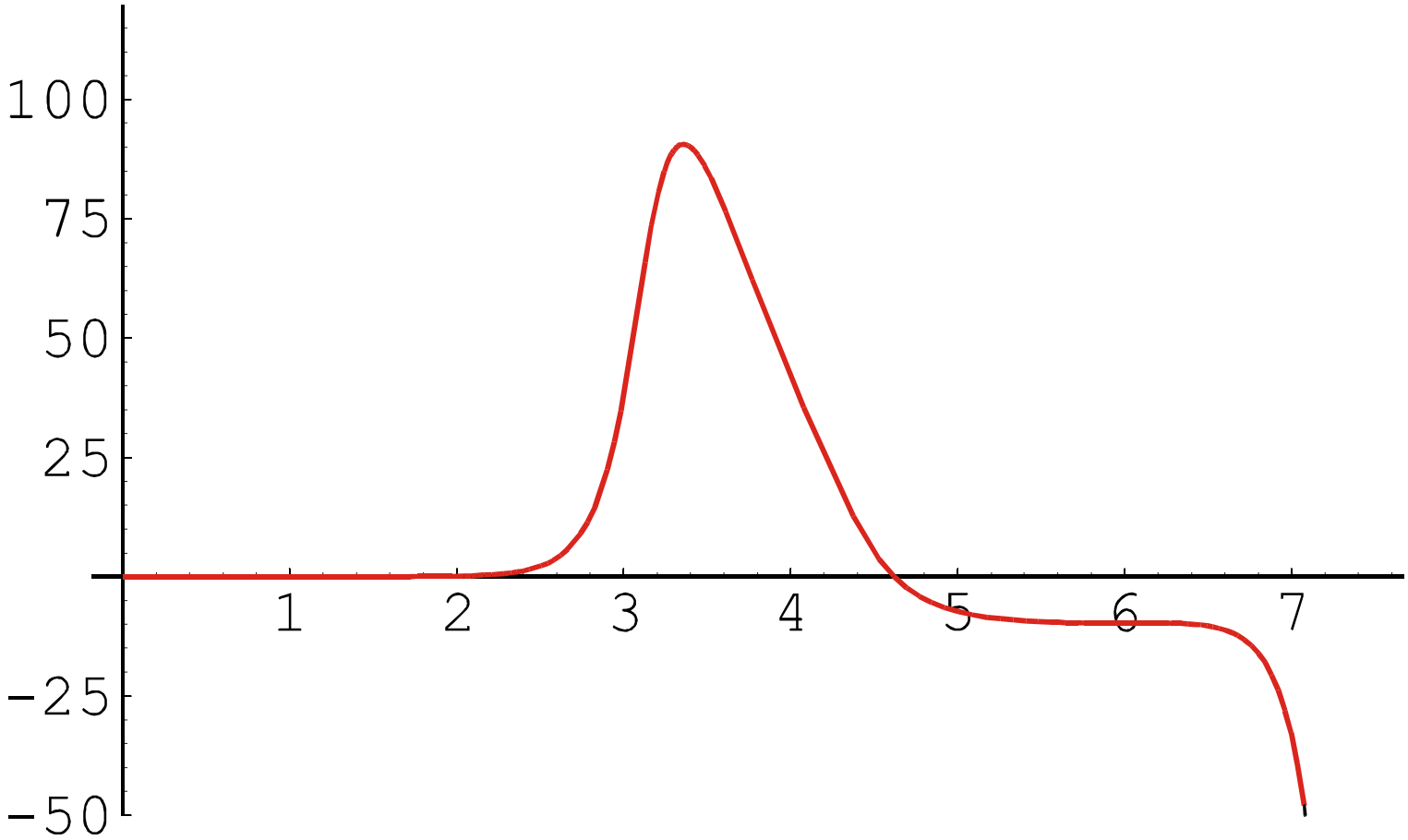
*Permanent Address: Department of Physics, Box 390, University of Colorado, Boulder, CO 80309, USA E-mail: phillips@colorado.edu



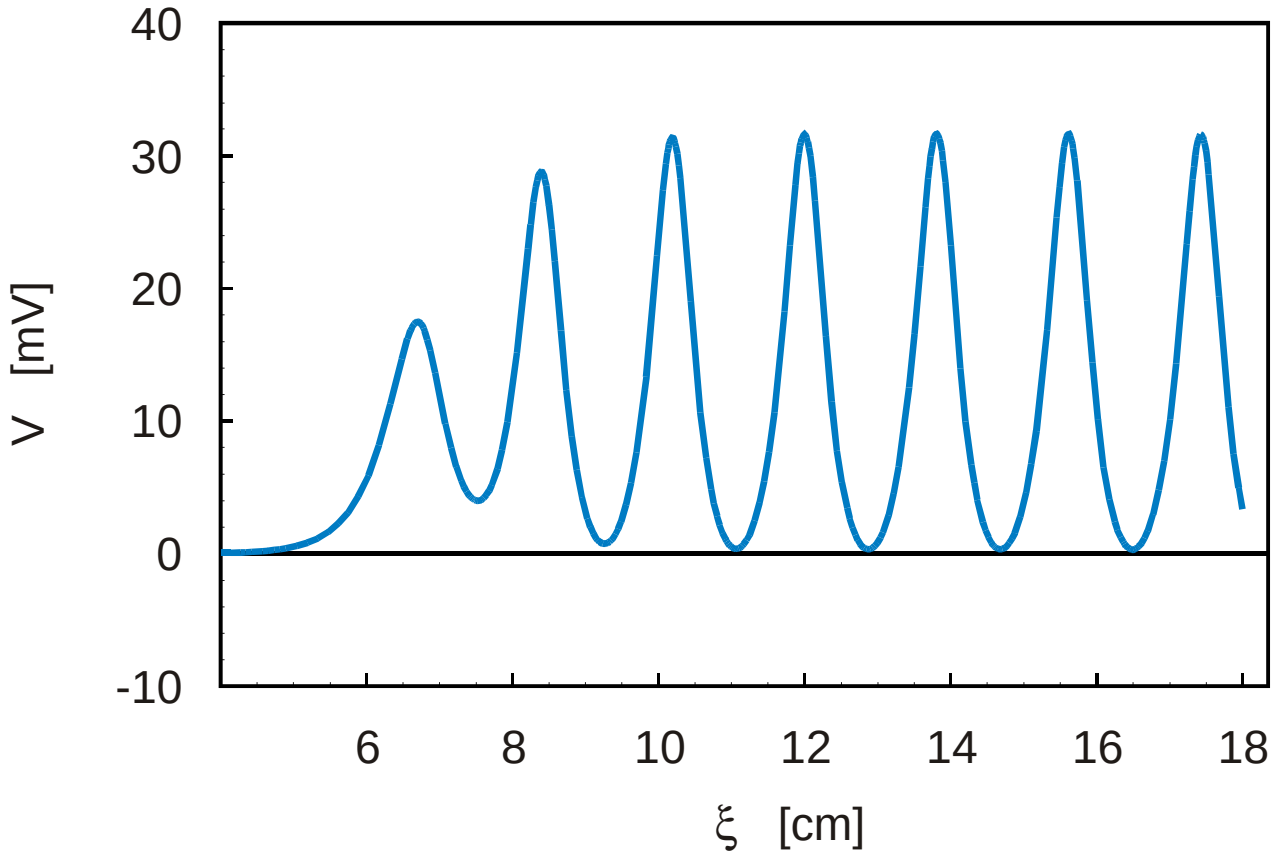
T = 18.5 C; θ = 1873.33 cm / sec



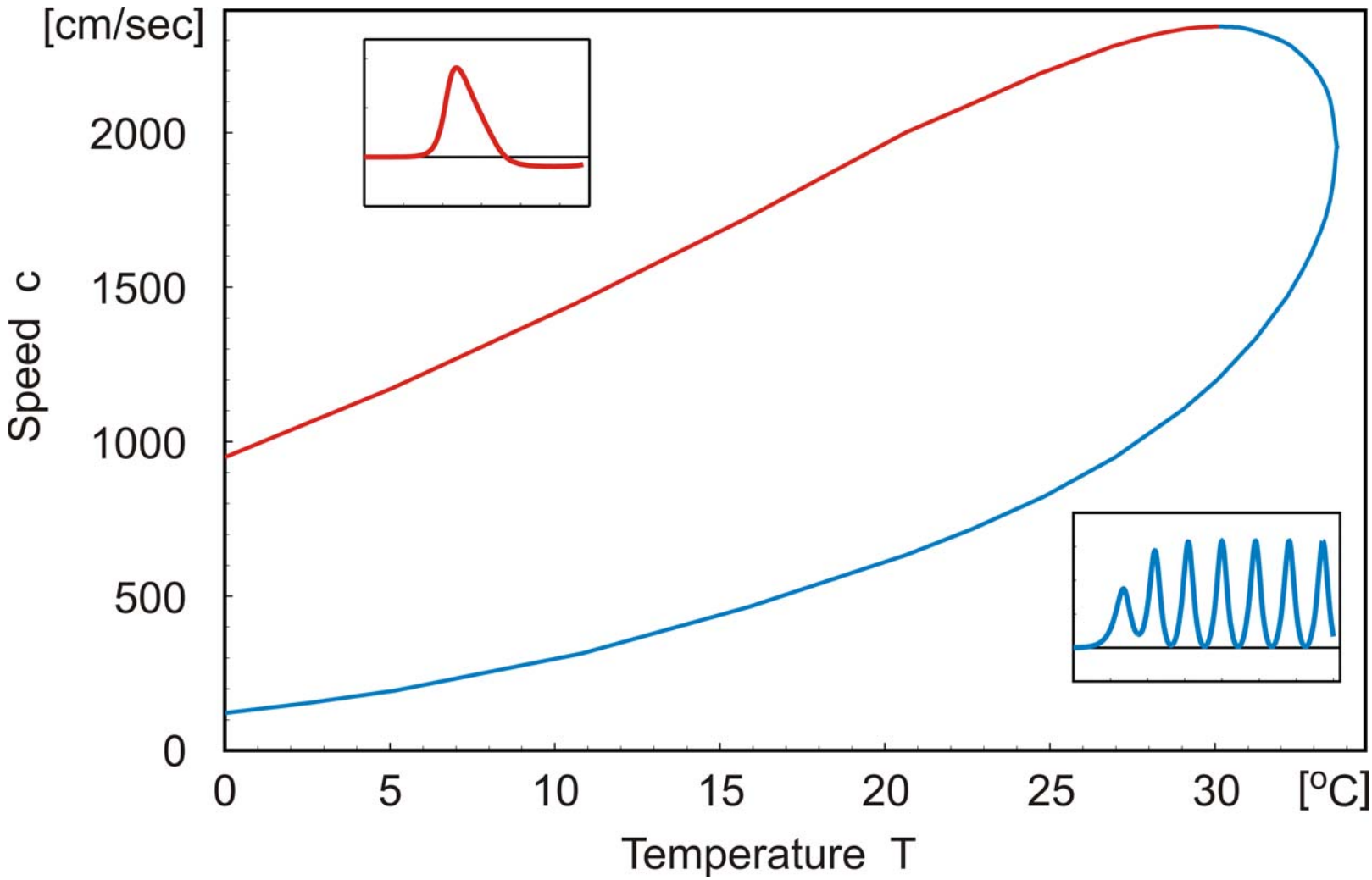
T = 18.5 C; θ = 1873.3324514717698 cm / sec



T = 18.5 C; θ = 1873.3324514717697 cm / sec



T = 18.5 C; θ = 544.070 cm / sec



Propagating wave solutions of the Hodgkin-Huxley equations

1. Fibonacci – Rabbits, plants, and the golden ratio
2. Mendel – Colors, genes, and inheritance
3. Fisher – Synthesis of genetics and Darwinian evolution
4. Turing – The origin of patterns
5. Hodgkin and Huxley – Neurons and PDEs
6. **Error thresholds and antiviral strategies**
7. Neutral networks – How evolution works

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 deteriorate 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

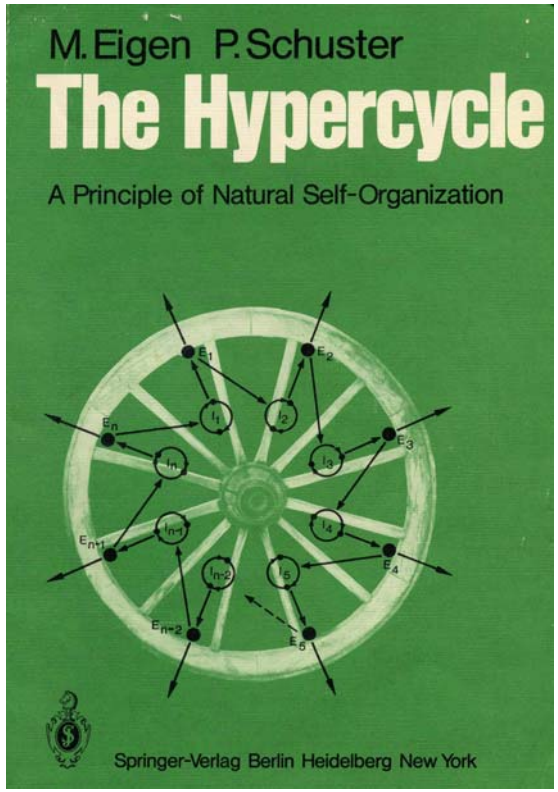
The mathematical analysis of dynamical systems using methods of differential topology, yields the result that there is only one type of mechanisms which fulfills the following requirements: The information stored in each single replicative unit (or reproductive cycle) must be maintained, i.e., the respective 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 functions.
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, Lyapunov- 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-



Chemical kinetics of molecular evolution

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

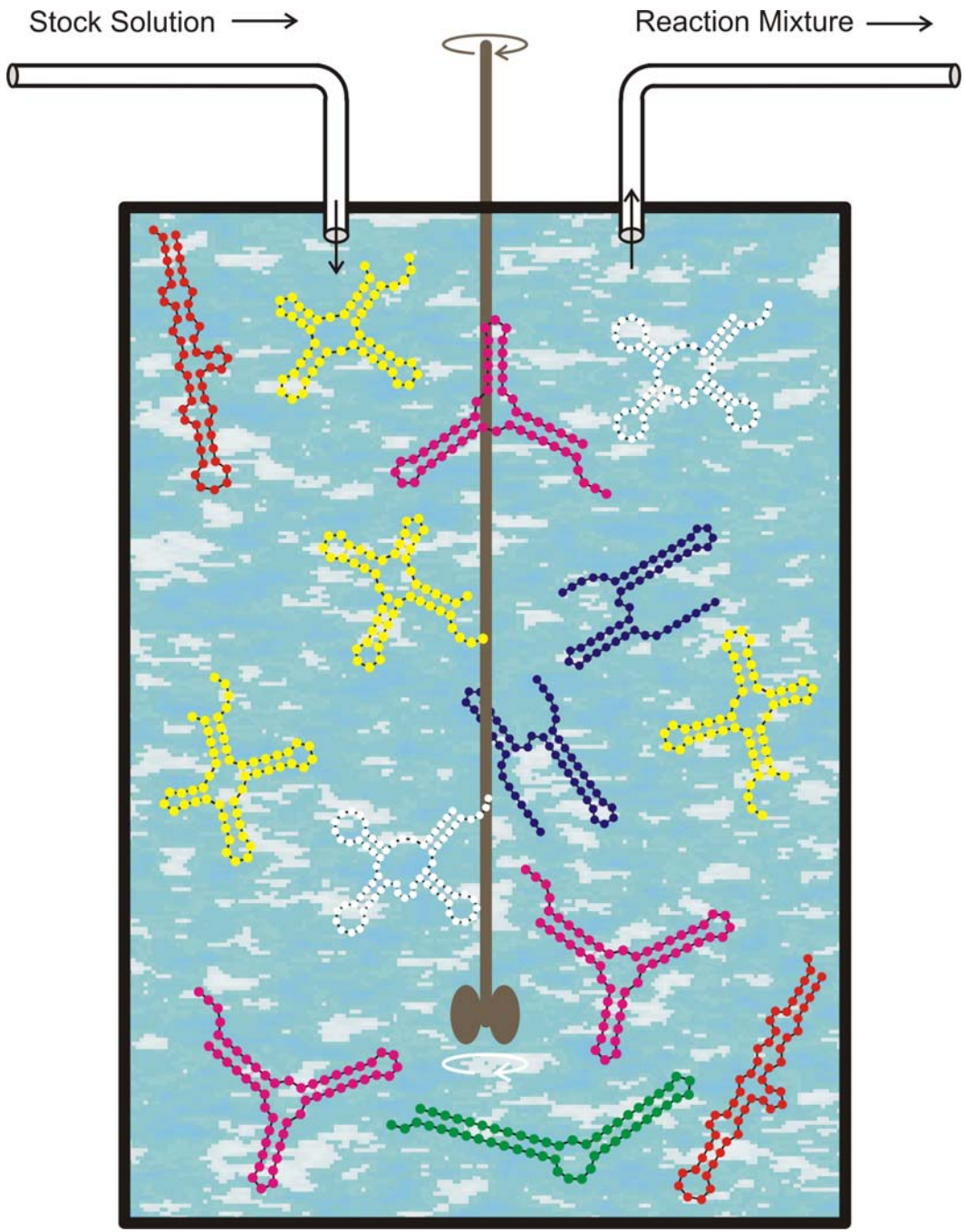
Stock solution:

activated monomers, **ATP, CTP, GTP, UTP (TTP);**
a replicase, an enzyme that performs complemantary replication;
buffer solution

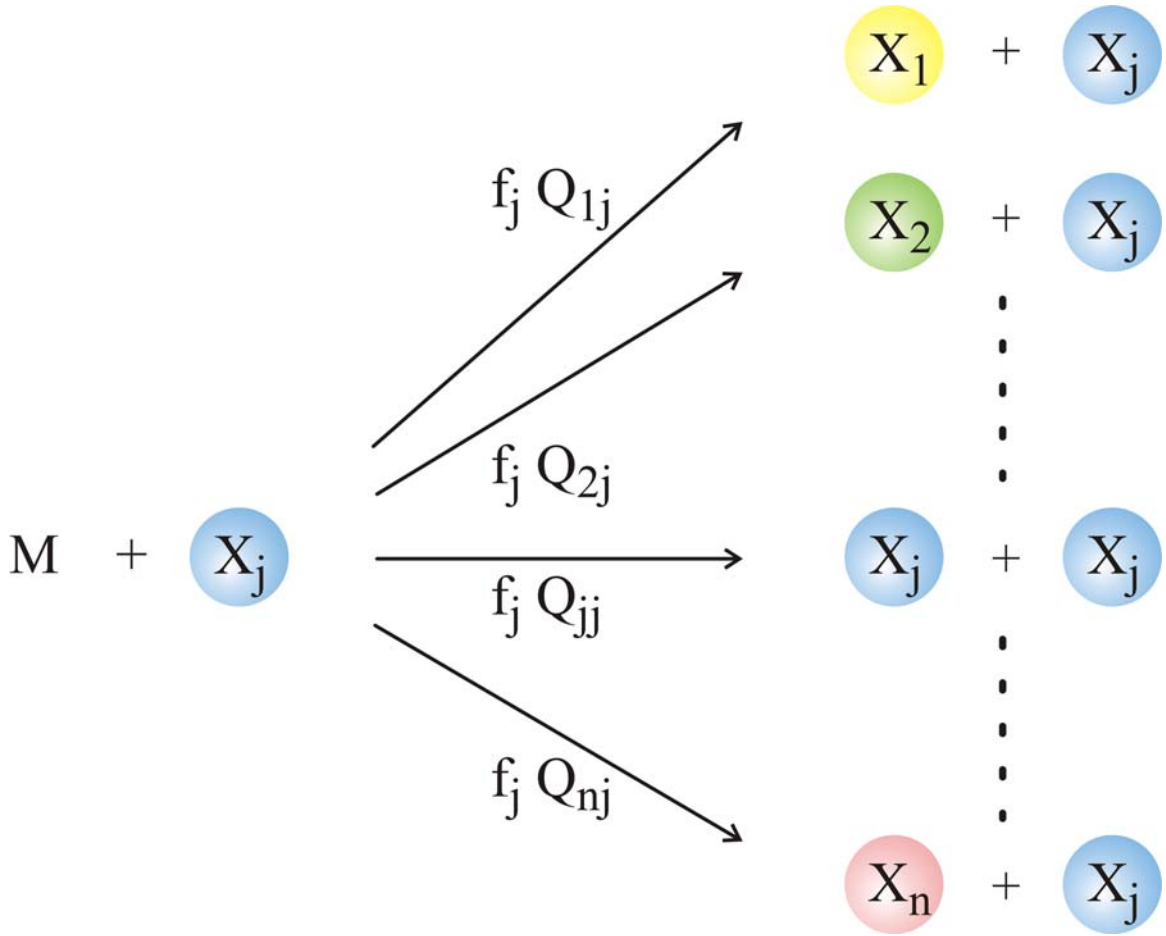
Flow rate: $r = \tau_R^{-1}$

The population size N, the number of polynucleotide molecules, is controlled by the flow r

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



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



Chemical kinetics of replication and mutation as parallel reactions

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

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

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

Manfred Eigen‘s replication-mutation equation

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

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

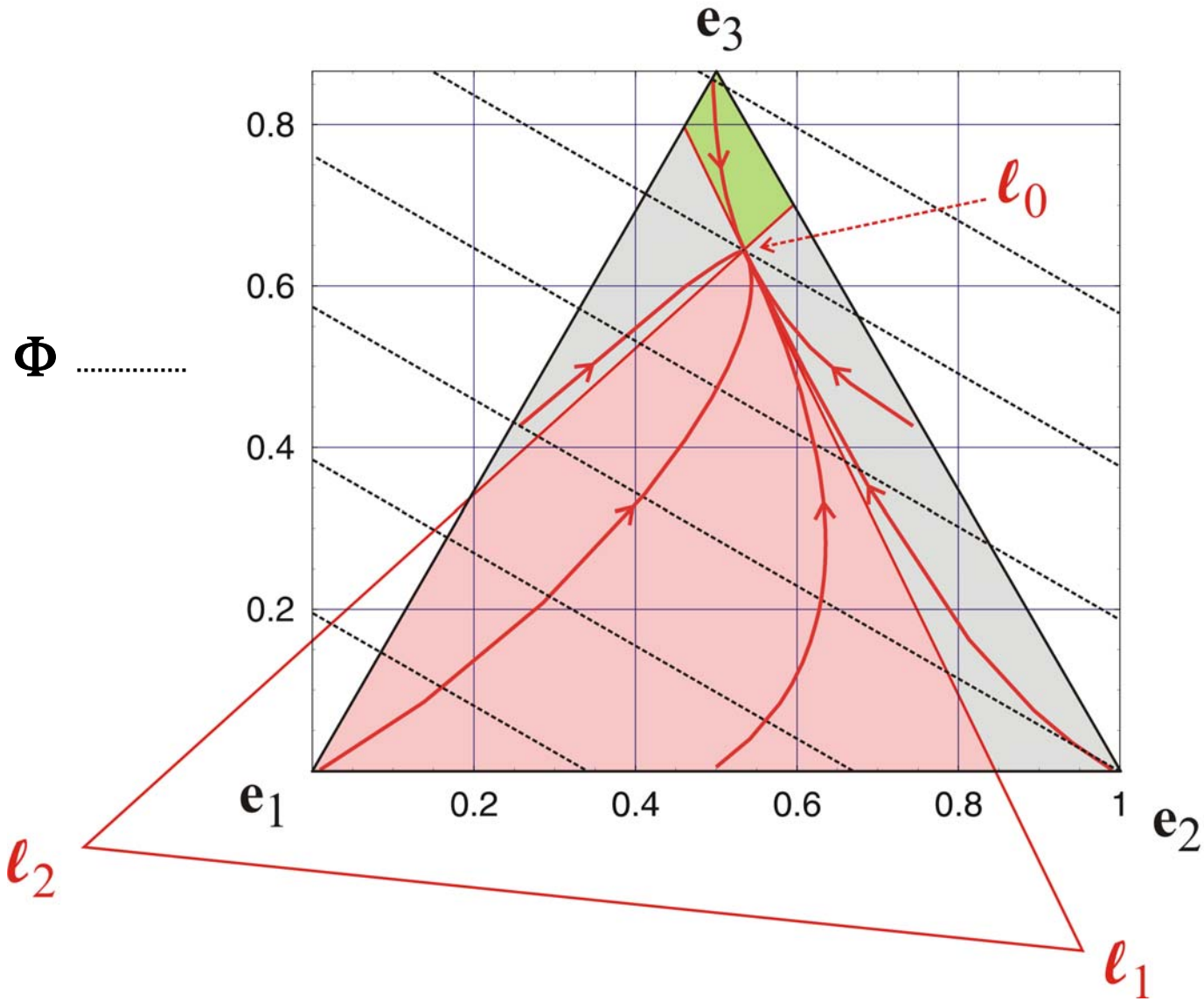
Solutions are obtained after integrating factor transformation by means of an eigenvalue problem

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

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

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

constant level sets of Φ .....

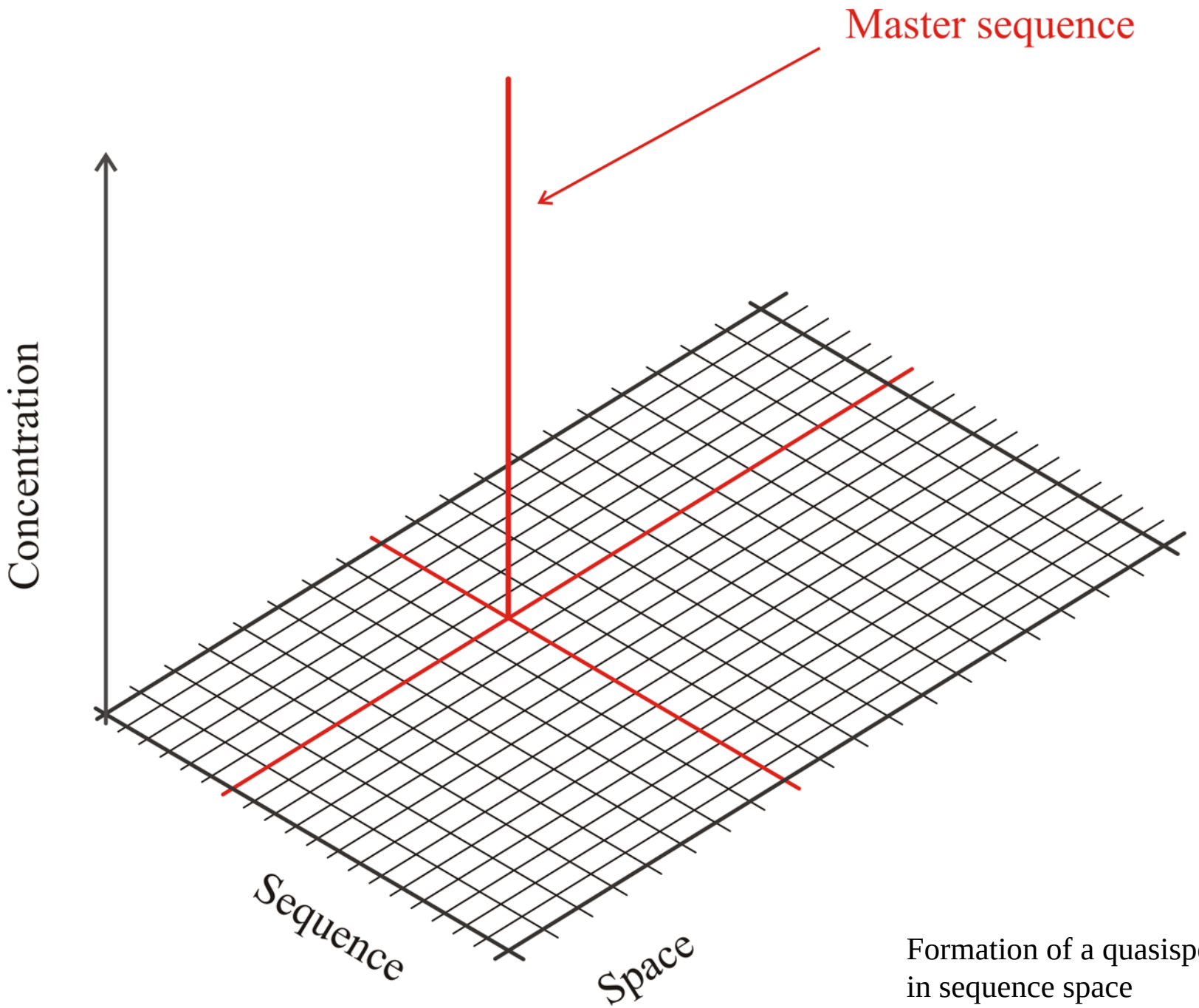


Selection of quasispecies with $f_1 = 1.9$, $f_2 = 2.0$, $f_3 = 2.1$, and $p = 0.01$
parametric plot on S_3

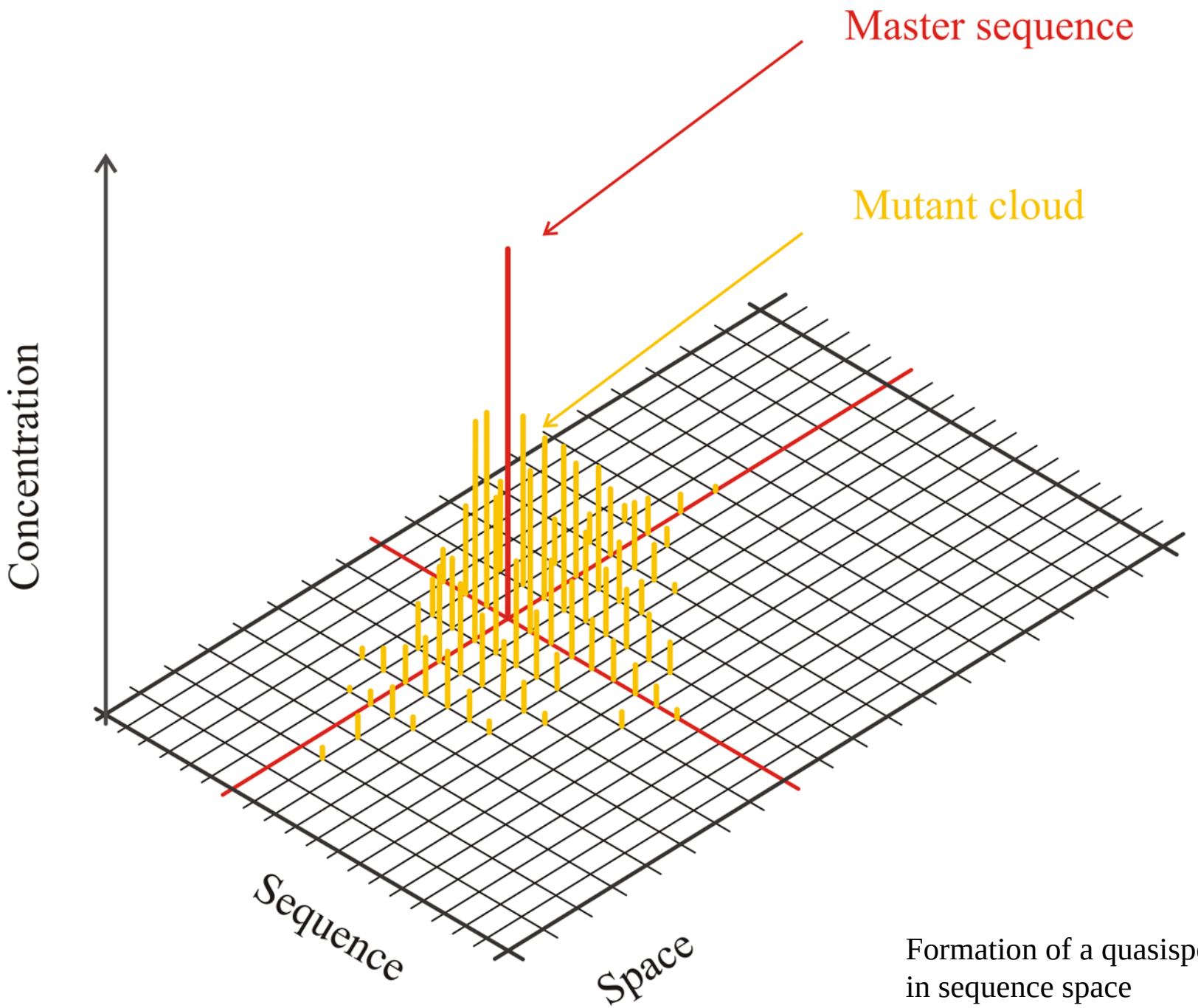
Uniform error rate model:

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

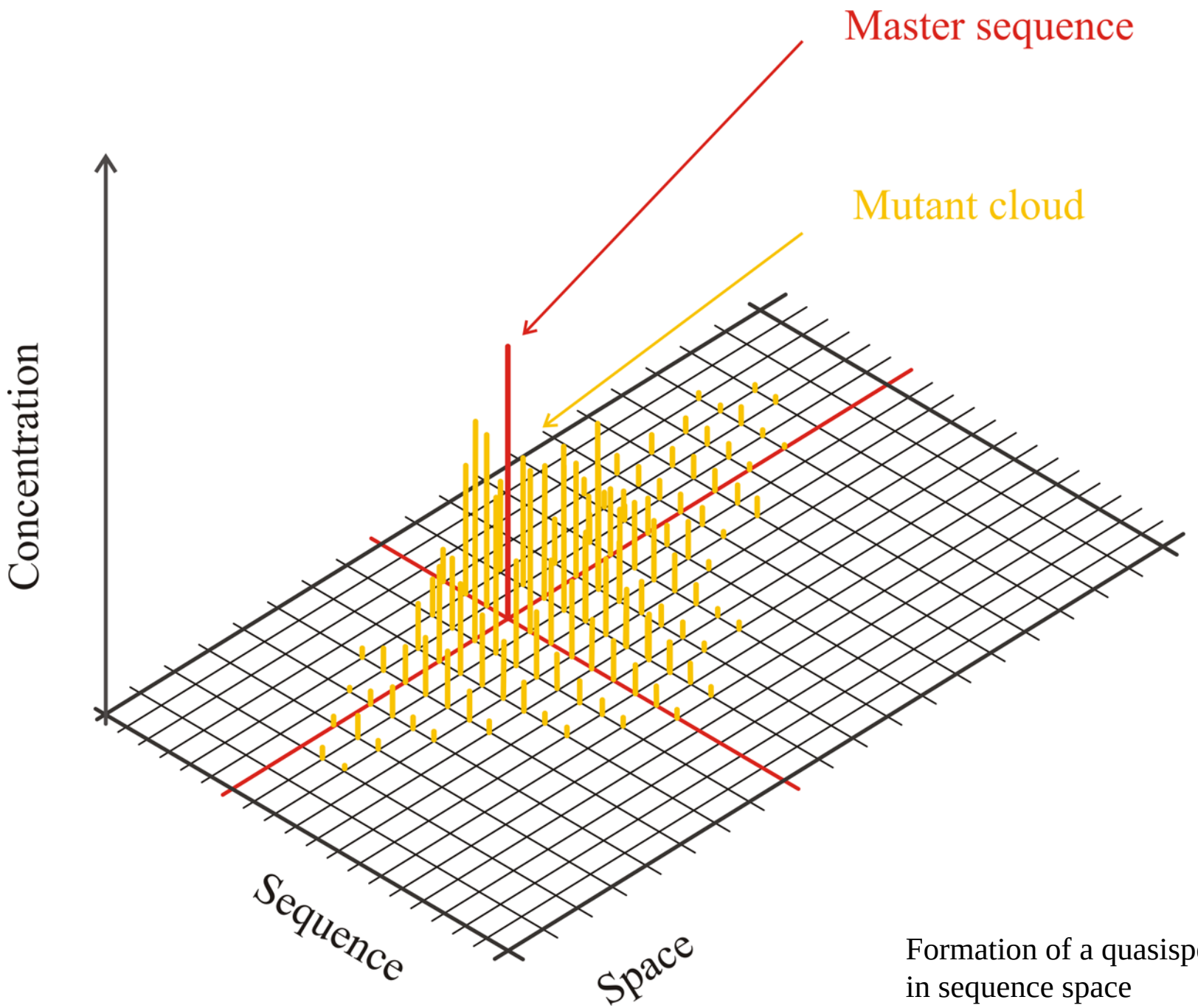
$d_H(\mathbf{X}_i, \mathbf{X}_j)$ ... Hamming distance



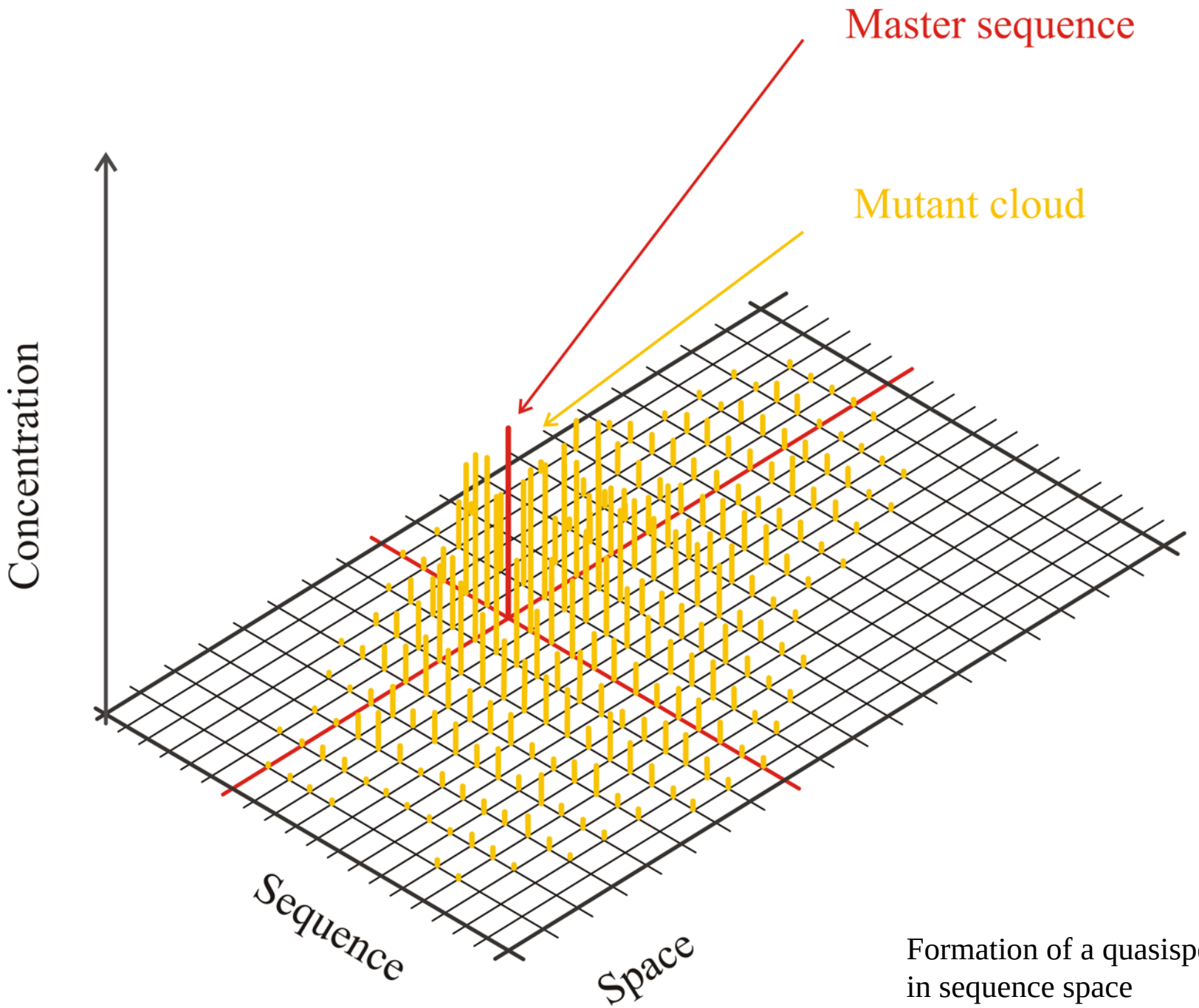
Formation of a quasispecies in sequence space



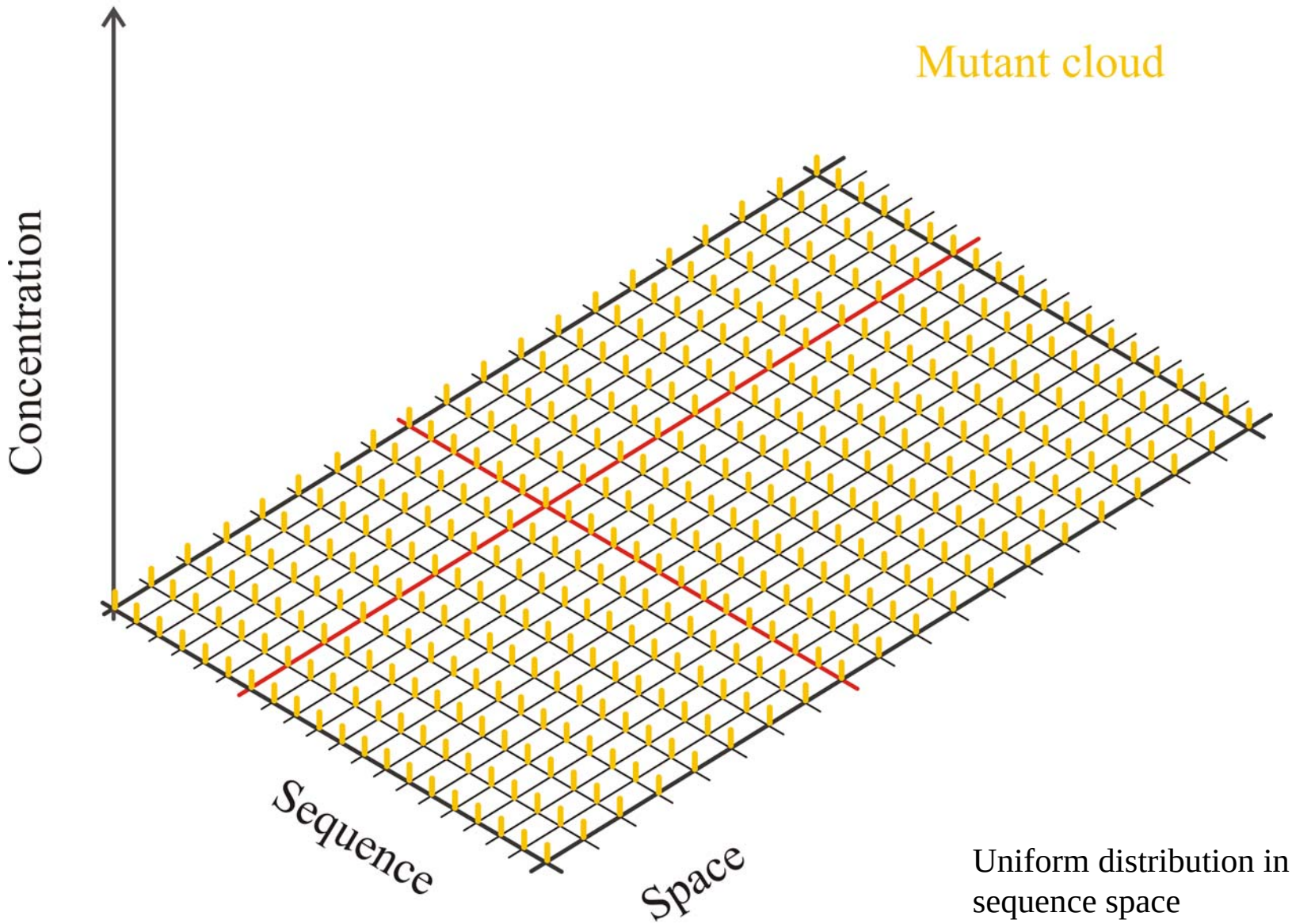
Formation of a quasispecies in sequence space



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:

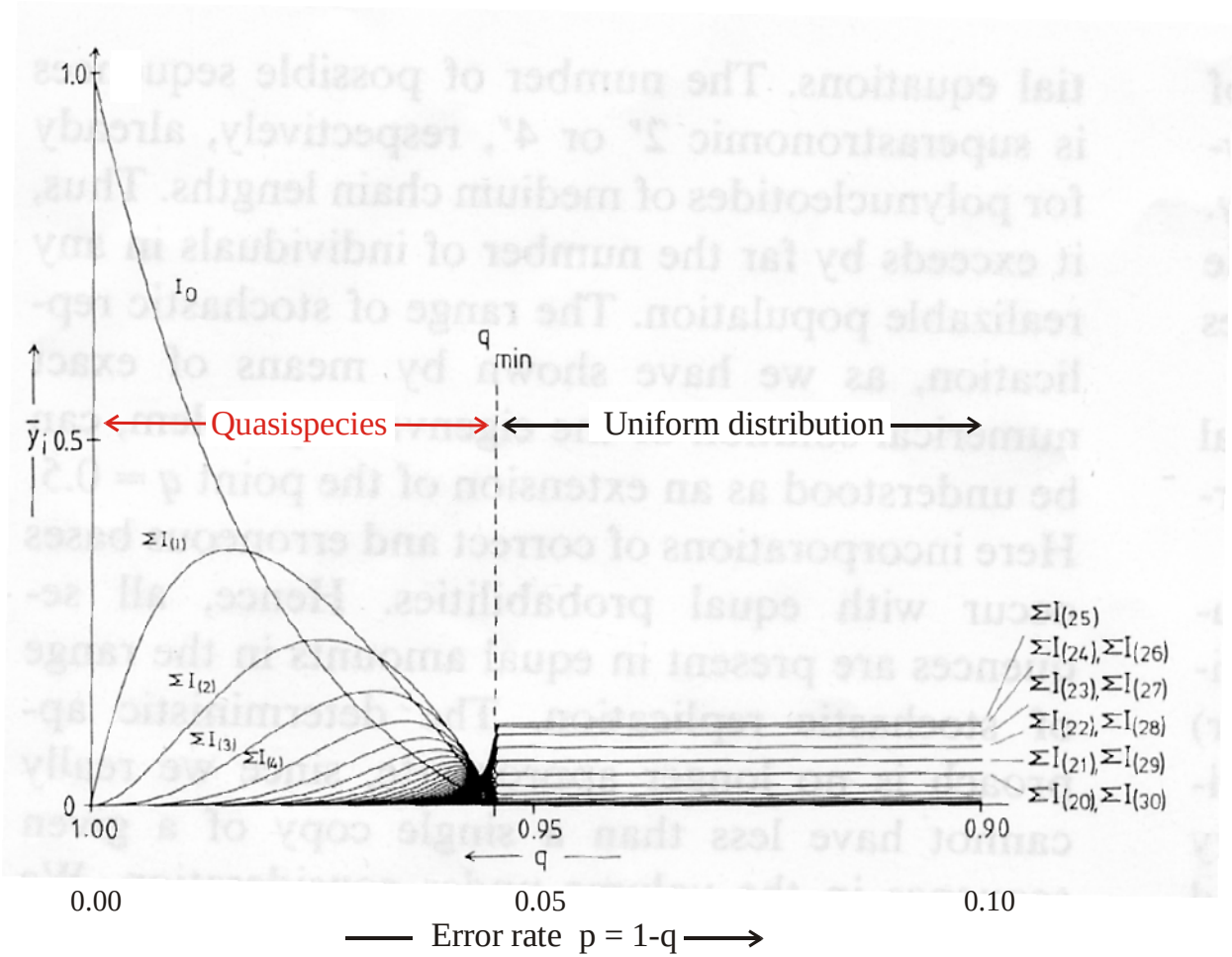
$$\frac{dx_i}{dt} = \dot{x}_i = \sum_j w_{ij} x_j - \frac{x_i}{c}\phi; \; i = 1, \ldots, n \; ^{\dagger} \qquad (1)$$

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

Chain length and error threshold

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

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

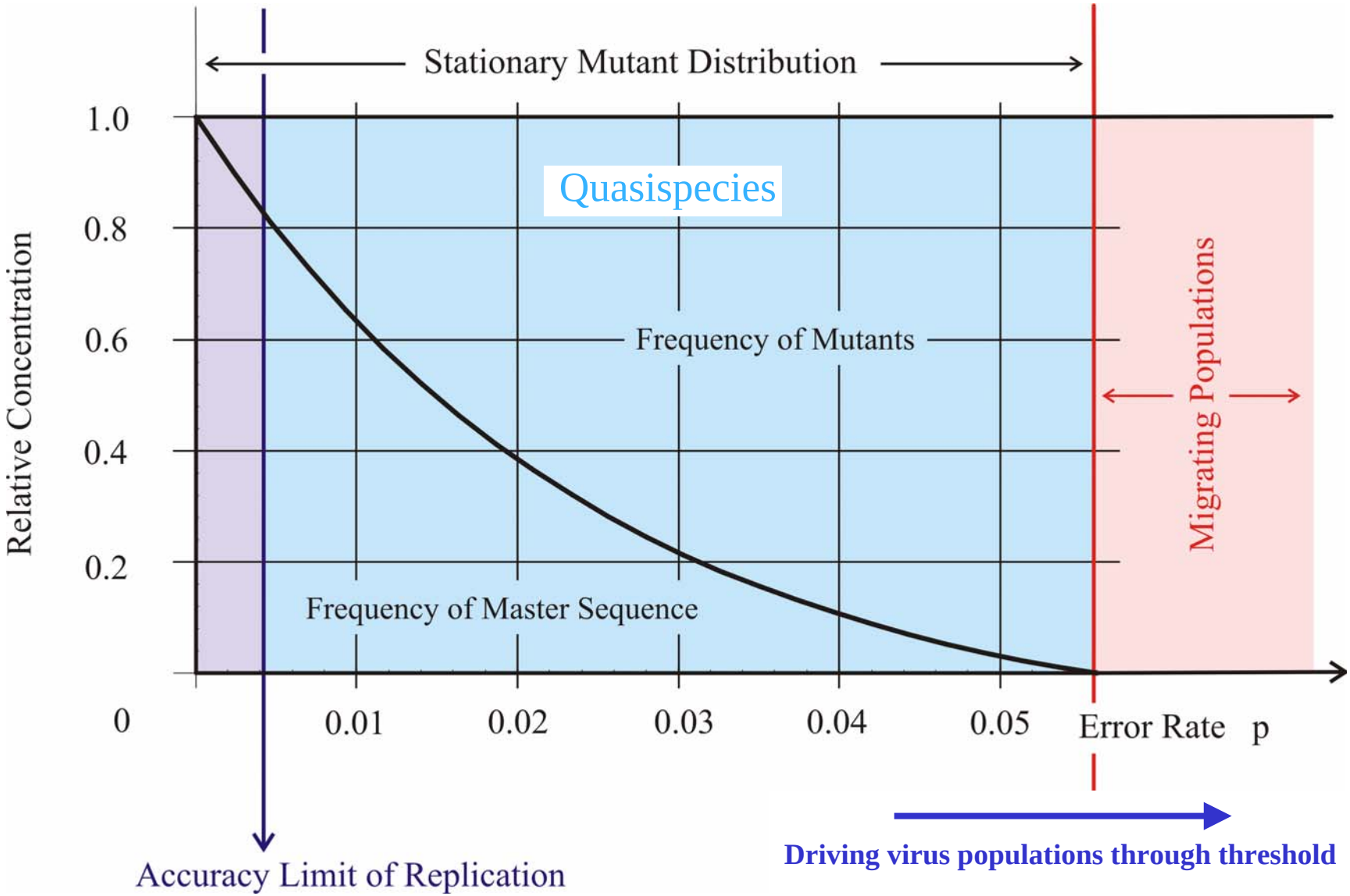
$$p \ \ldots \ \text{constant} : \quad n_{\max} \approx \frac{\ln\sigma}{p}$$

$Q = (1-p)^n$ … replication accuracy

p … error rate

n … chain length

$\sigma = \dfrac{f_m}{(1-x_m)\sum_{j \neq m} f_j}$ … superiority of master sequence



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

1. Fibonacci – Rabbits, plants, and the golden ratio
2. Mendel – Colors, genes, and inheritance
3. Fisher – Synthesis of genetics and Darwinian evolution
4. Turing – The origin of patterns
5. Hodgkin and Huxley – Neurons and PDEs
6. Error thresholds and antiviral strategies
7. **Neutral networks – How evolution works**

RNA sequence: GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA

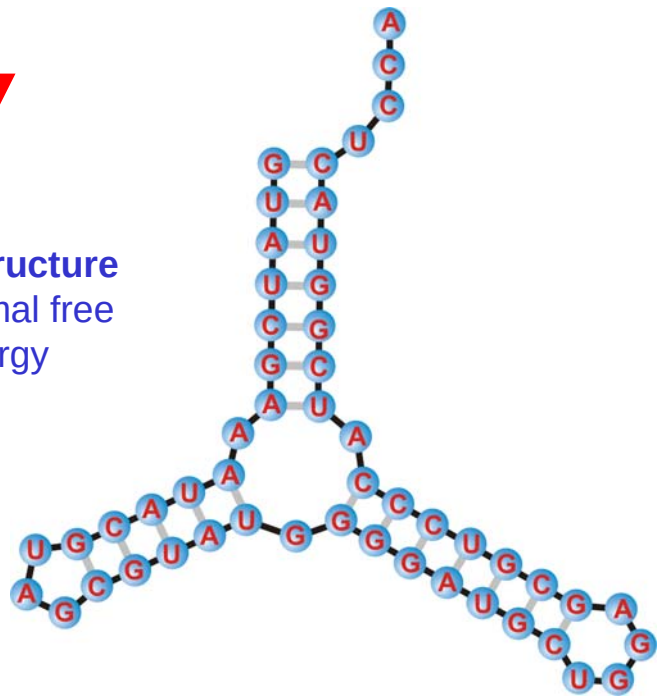
RNA folding:

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

Biophysical chemistry: thermodynamics and kinetics

Empirical parameters

RNA structure of minimal free energy



Sequence, structure, and design

The **Vienna RNA-Package**:

A library of routines for folding, inverse folding, sequence and structure alignment, kinetic folding, cofolding, …

Monatshefte für Chemie 125, 167–188 (1994)

Monatshefte für Chemie
Chemical Monthly

Printed in Austria

Fast Folding and Comparison of RNA Secondary Structures

I. L. Hofacker[1,*], W. Fontana[3], P. F. Stadler[1,3], L. S. Bonhoeffer[4], M. Tacker[1] and P. Schuster[1,2,3]

[1] Institut für Theoretische Chemie, Universität Wien, A-1090 Wien, Austria
[2] Institut für Molekulare Biotechnologie, D-07745 Jena, Federal Republic of Germany
[3] Santa Fe Institute, Santa Fe, NM 87501, U.S.A.
[4] Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, U.K.

Summary. Computer codes for computation and comparison of RNA secondary structures, the Vienna RNA package, are presented, that are based on dynamic programming algorithms and aim at predictions of structures with minimum free energies as well as at computations of the equilibrium partition functions and base pairing probabilities.

An efficient heuristic for the inverse folding problem of RNA is introduced. In addition we present compact and efficient programs for the comparison of RNA secondary structures based on tree editing and alignment.

All computer codes are written in ANSI C. They include implementations of modified algorithms on parallel computers with distributed memory. Performance analysis carried out on an Intel Hypercube shows that parallel computing becomes gradually more and more efficient the longer the sequences are.

Keywords. Inverse folding; parallel computing; public domain software; RNA folding; RNA secondary structures; tree editing.

Schnelle Faltung und Vergleich von Sekundärstrukturen von RNA

Zusammenfassung. Die im Vienna RNA package enthaltenen Computer Programme für die Berechnung und den Vergleich von RNA Sekundärstrukturen werden präsentiert. Ihren Kern bilden Algorithmen zur Vorhersage von Strukturen minimaler Energie sowie zur Berechnung von Zustandssumme und Basenpaarungswahrscheinlichkeiten mittels dynamischer Programmierung.

Ein effizienter heuristischer Algorithmus für das inverse Faltungsproblem wird vorgestellt. Darüberhinaus präsentieren wir kompakte und effiziente Programme zum Vergleich von RNA Sekundärstrukturen durch Baum-Editierung und Alignierung.

Alle Programme sind in ANSI C geschrieben, darunter auch eine Implementation des Faltungsalgorithmus für Parallelrechner mit verteiltem Speicher. Wie Tests auf einem Intel Hypercube zeigen, wird das Parallelrechnen umso effizienter je länger die Sequenzen sind.

1. Introduction

Recent interest in RNA structures and functions was caused by their catalytic capacities [1, 2] as well as by the success of selection methods in producing RNA

RNA sequence: GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA

RNA folding:

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

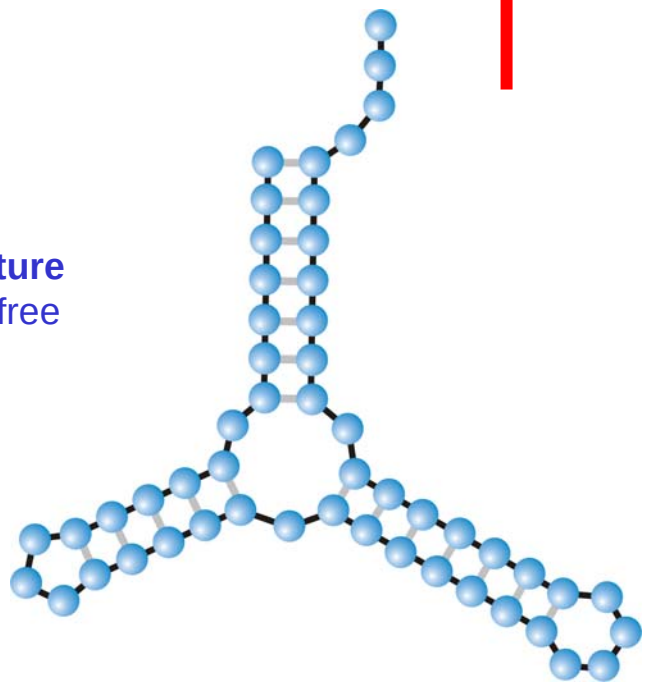
Iterative determination of a sequence for the given secondary structure

Inverse Folding Algorithm

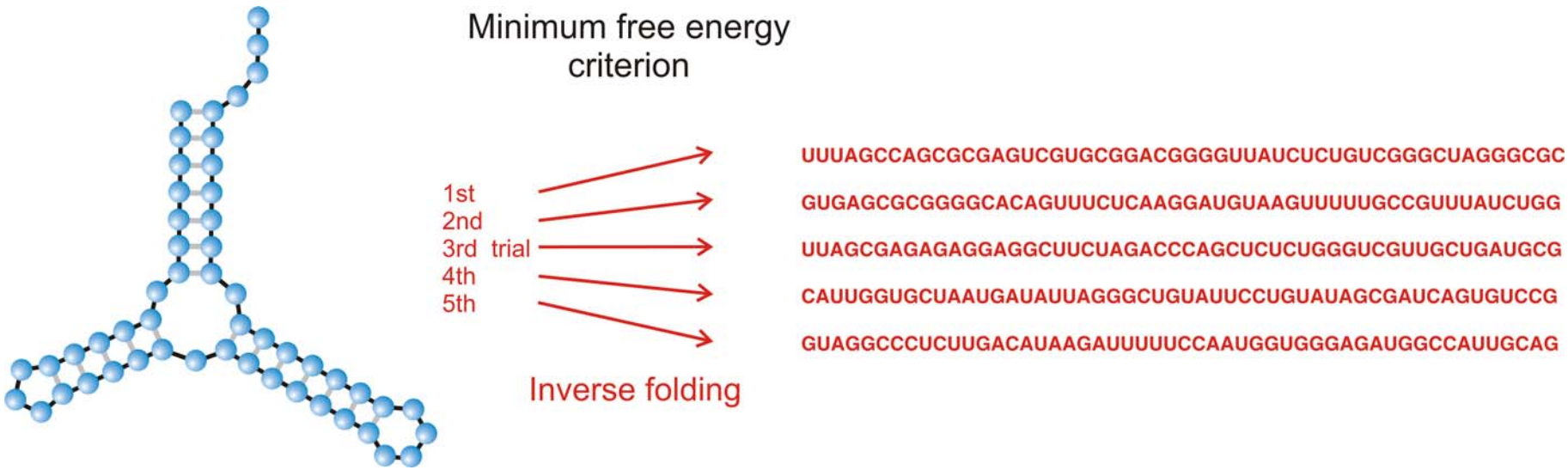
Inverse folding of RNA:

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

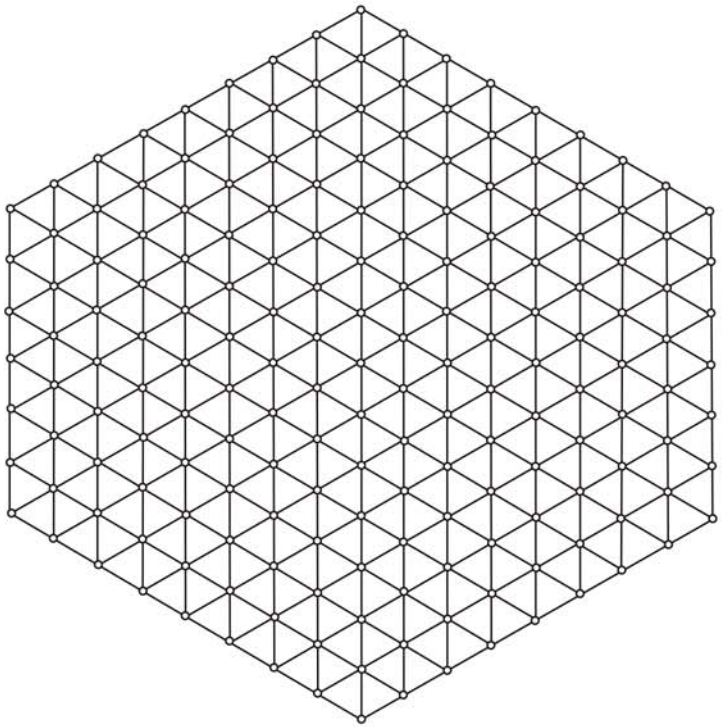
RNA structure of minimal free energy



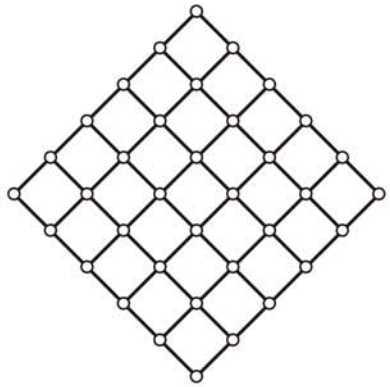
Sequence, structure, and design



The inverse folding algorithm searches for sequences that form a given RNA secondary structure under the minimum free energy criterion.

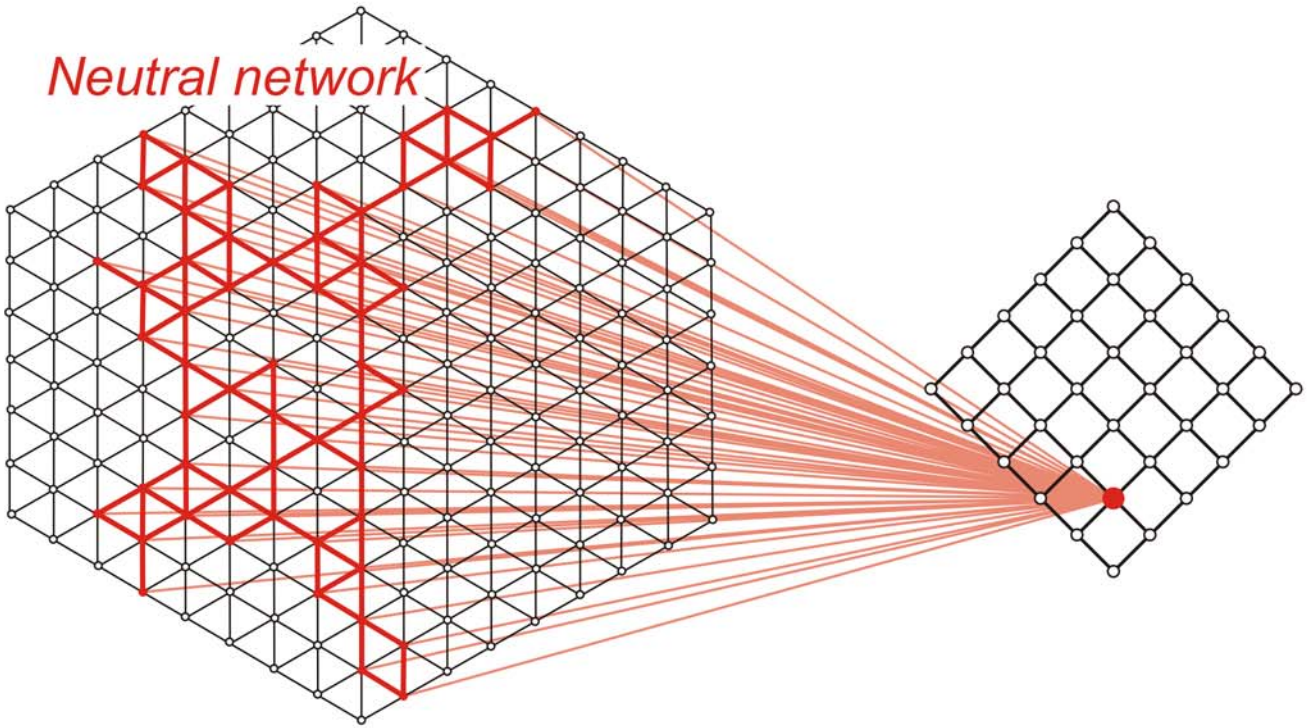


Sequence space



Structure space

Sequence space and structure space


Neutral network
Sequence space
Structure space

Space of genotypes: $\mathcal{I} = \{I_1, I_2, I_3, I_4, \ldots, I_N\}$; Hamming metric

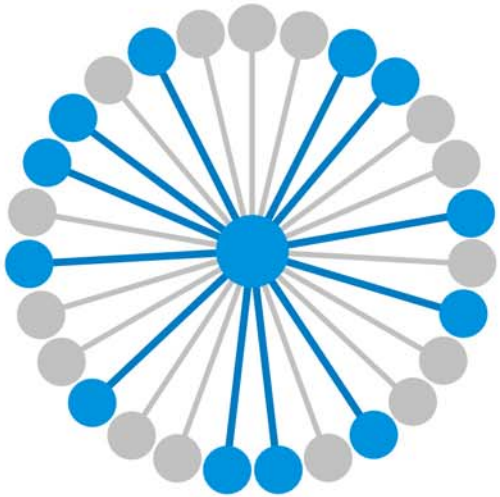
Space of phenotypes: $S = \{S_1, S_2, S_3, S_4, \ldots, S_M\}$; metric (not required)

$$N >> M$$

$$\psi(I_j) = S_k$$

$$\mathbf{G}_k = \psi^{-1}(S_k) \cup \{ I_j \mid \psi(I_j) = S_k \}$$

A mapping ψ and its inversion



$$G_k = \psi^{-1}(S_k) \doteq \{ I_j \mid \psi(I_j) = S_k \}$$

$$\bar{\lambda}_k = \frac{\sum_{j \in |G_k|} \lambda_j(k)}{|G_k|}$$

$\lambda_j = 12 / 27 = 0.444$

Alphabet size κ :

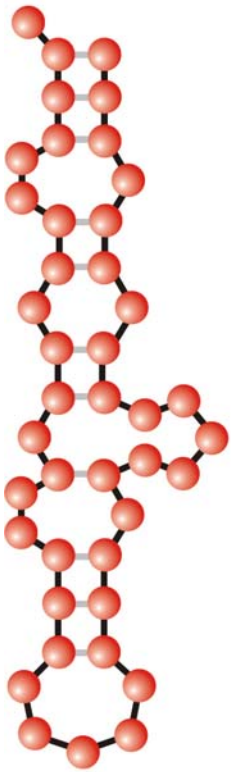
κ	λ_{cr}	
2	0.5	**AU,GC,DU**
3	0.423	**AUG , UGC**
4	0.370	**AUGC**

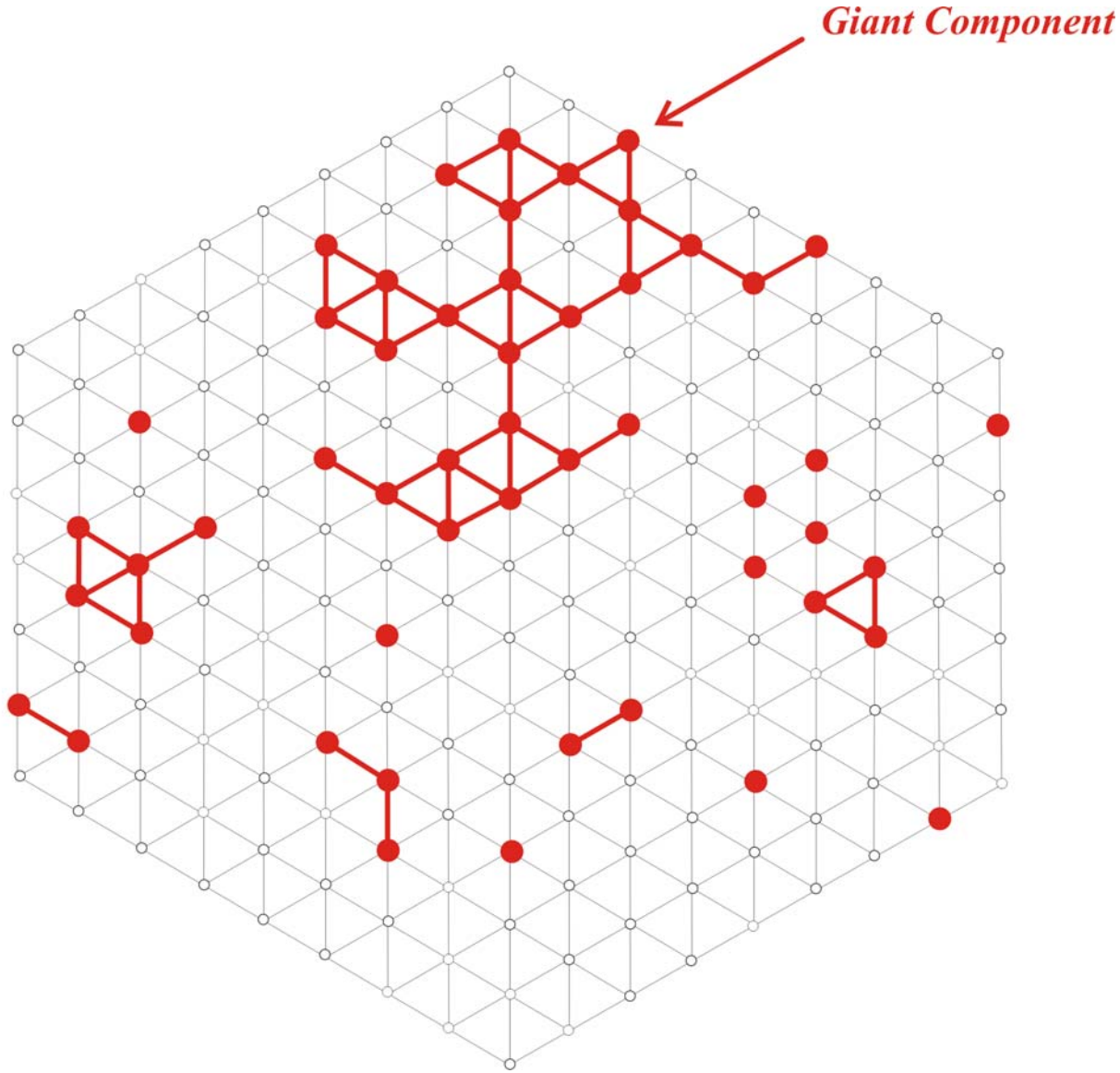
$\bar{\lambda}_k > \lambda_{cr}$ .... network G_k is connected

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

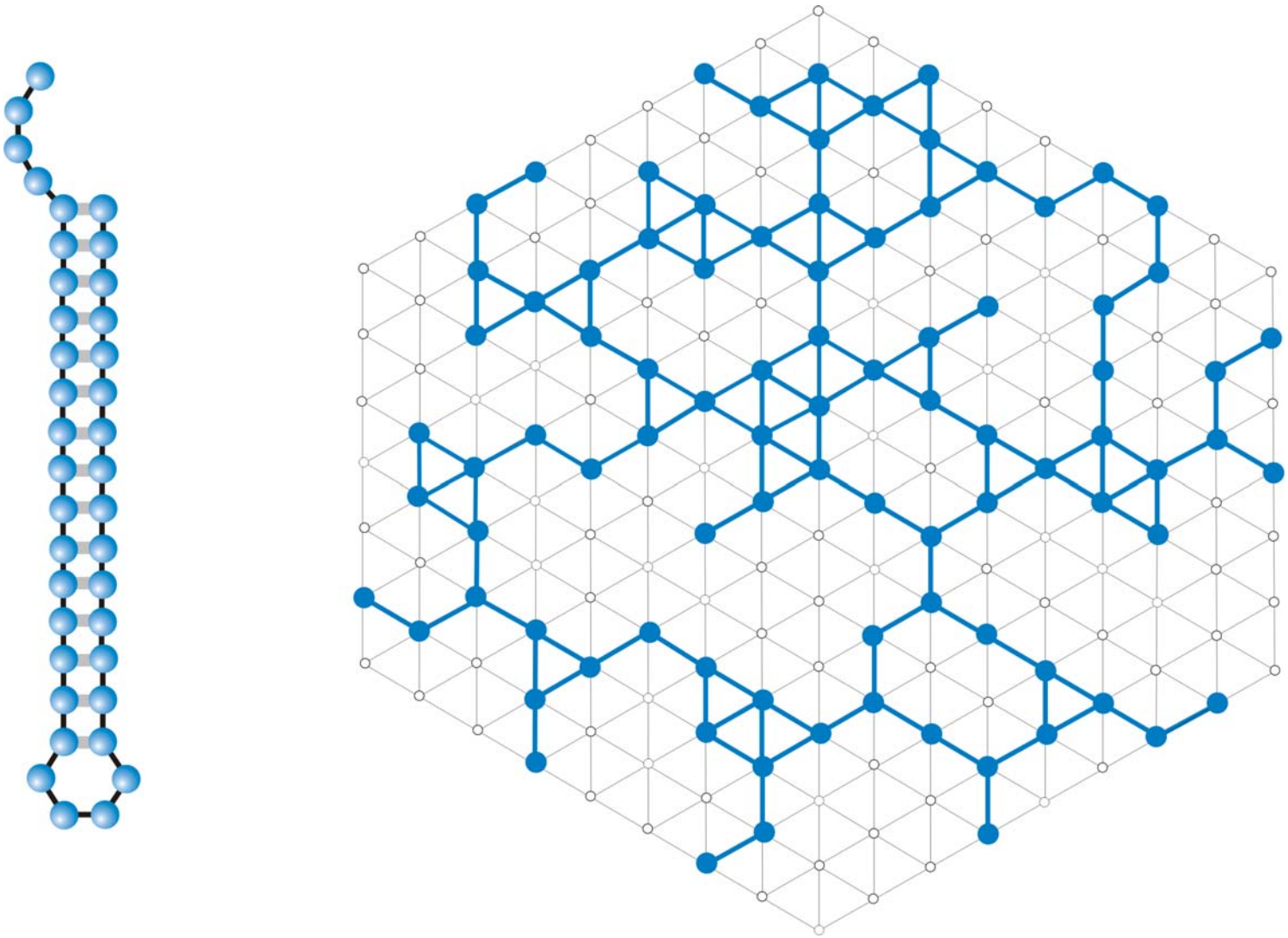
Connectivity threshold: $\lambda_{cr} = 1 - \kappa^{-1/(\kappa-1)}$

Degree of neutrality of neutral networks and the connectivity threshold





A multi-component neutral network formed by a rare structure: $\lambda < \lambda_{cr}$



A connected neutral network formed by a common structure: $\lambda > \lambda_{cr}$

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.

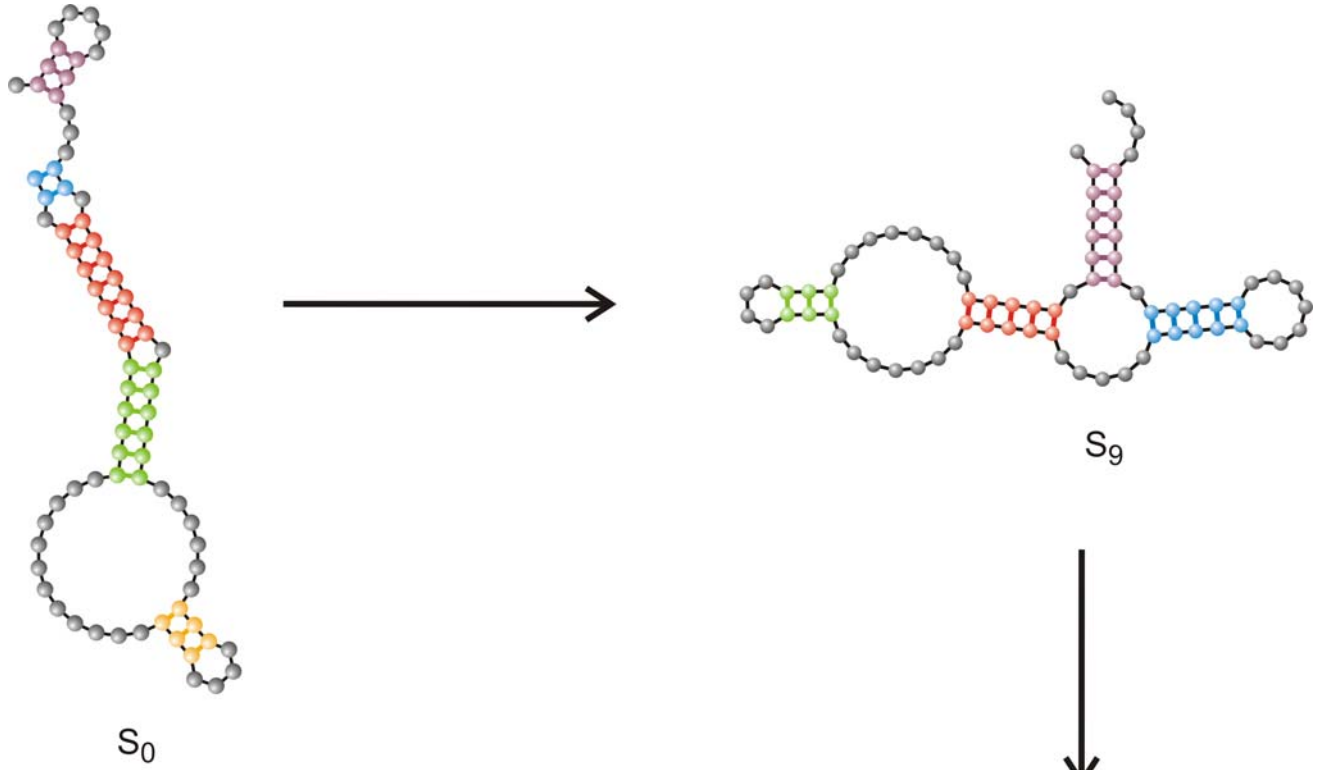
An instance representing in its qualitative features all the simulations we performed is shown in Fig. 1A. Starting with identical sequences folding into a random shape, the simulation was stopped when the population became dominated by the target, here a canonical tRNA shape. The black curve traces 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

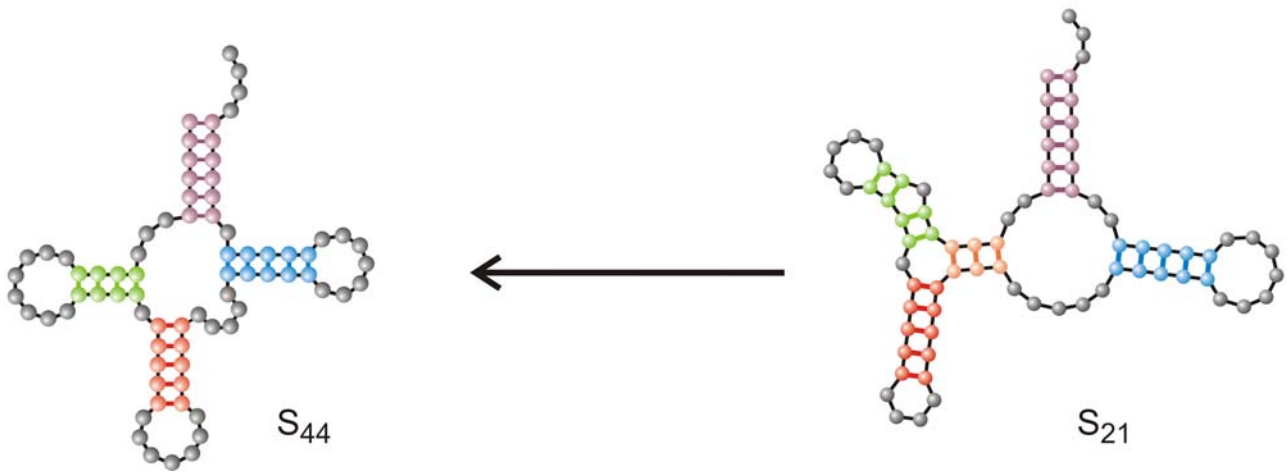
Institut für Theoretische Chemie, Universität Wien, Währingerstrasse 17, A-1090 Wien, Austria, Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA, and International Institute for Applied Systems Analysis (IIASA), A-2361 Laxenburg, Austria.

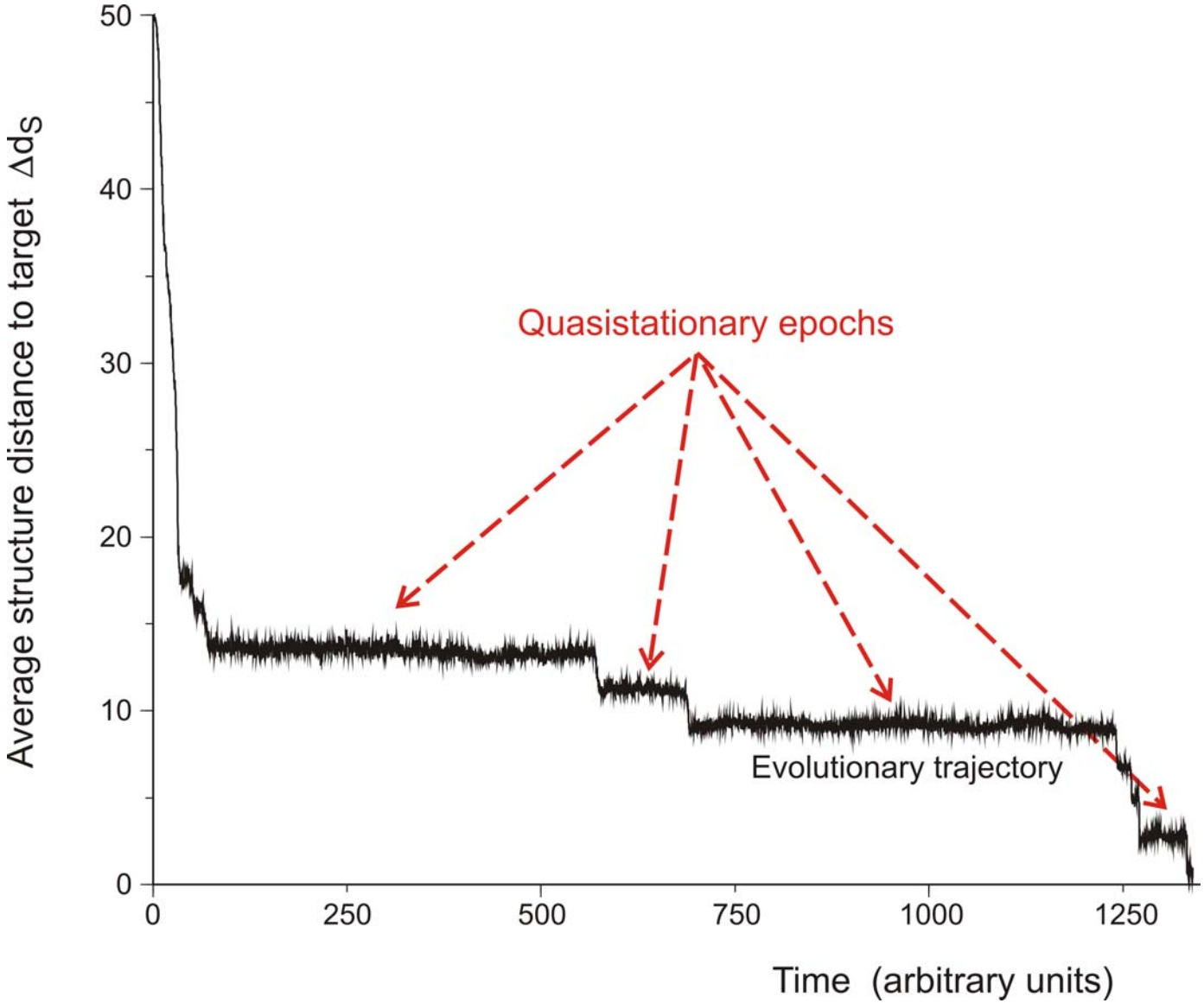
Stochastic simulation of evolution of RNA molecules

Randomly chosen
initial structure

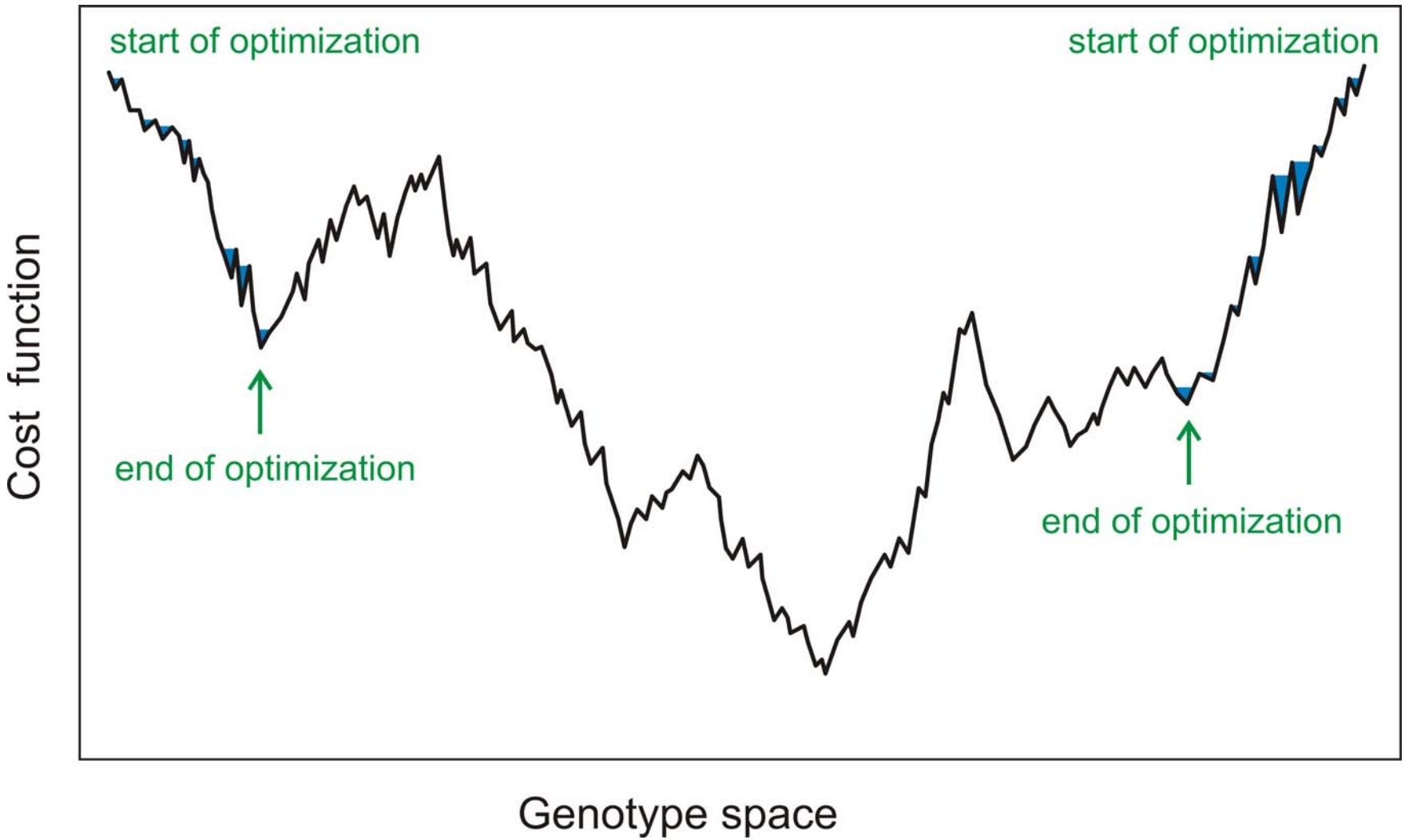


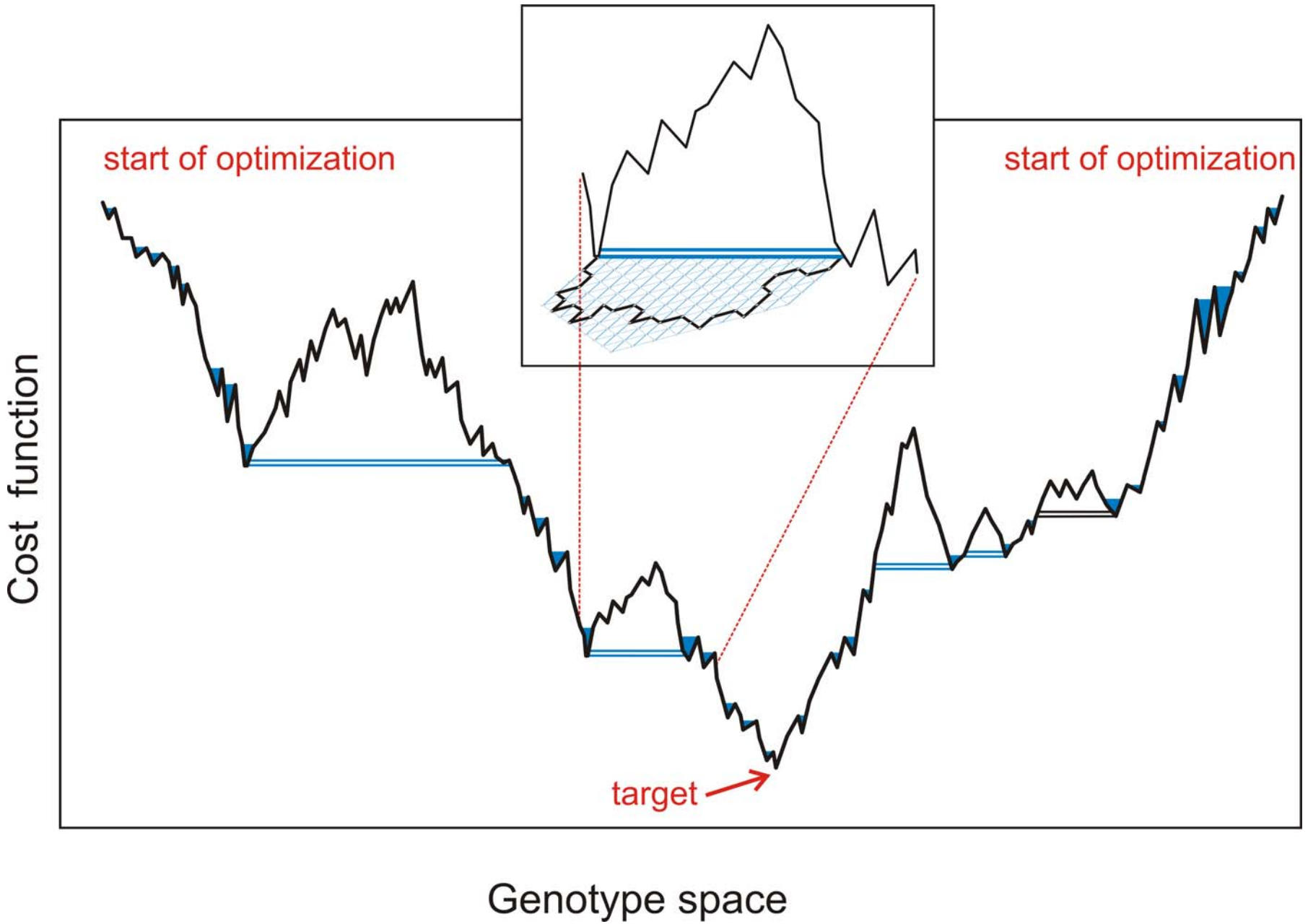
Phenylalanyl-tRNA
as target structure

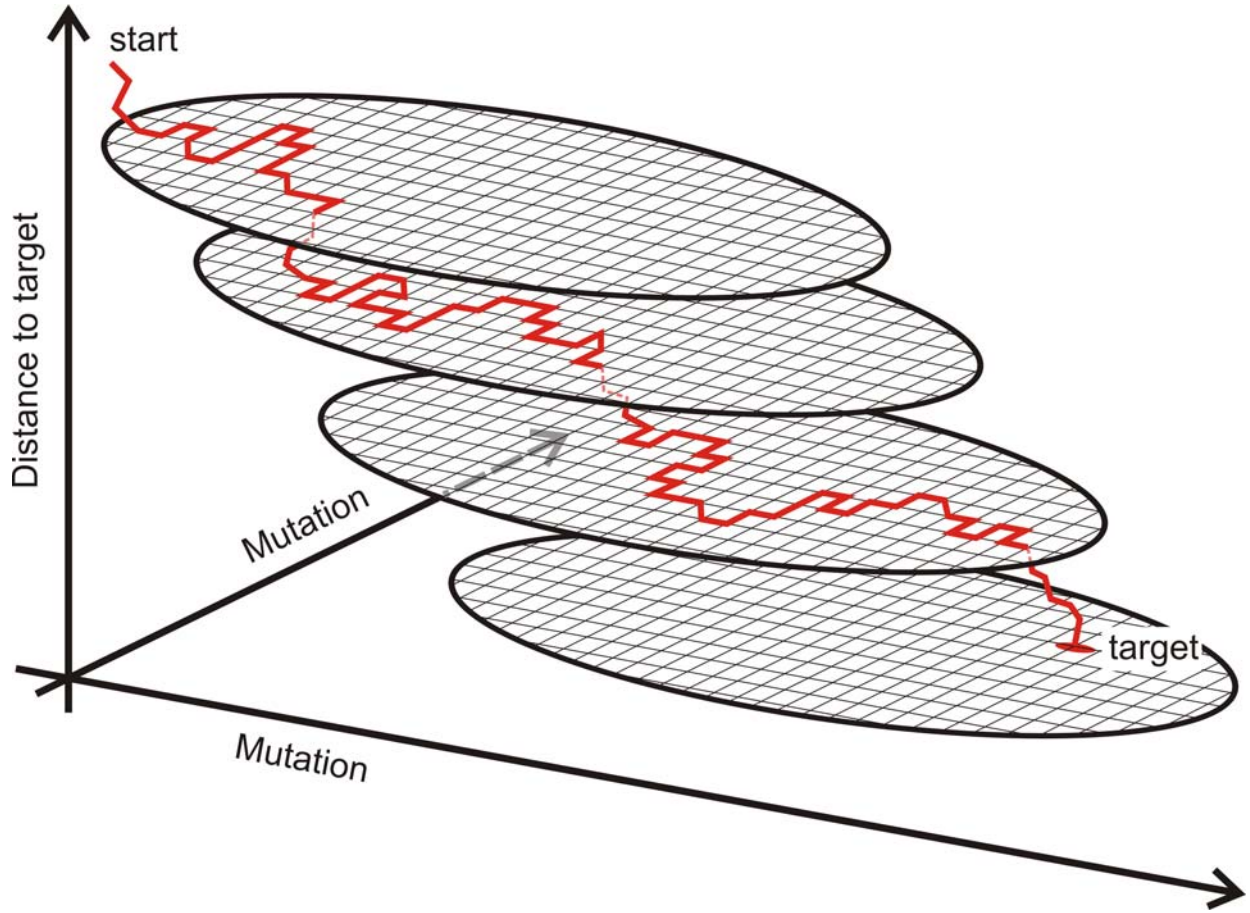




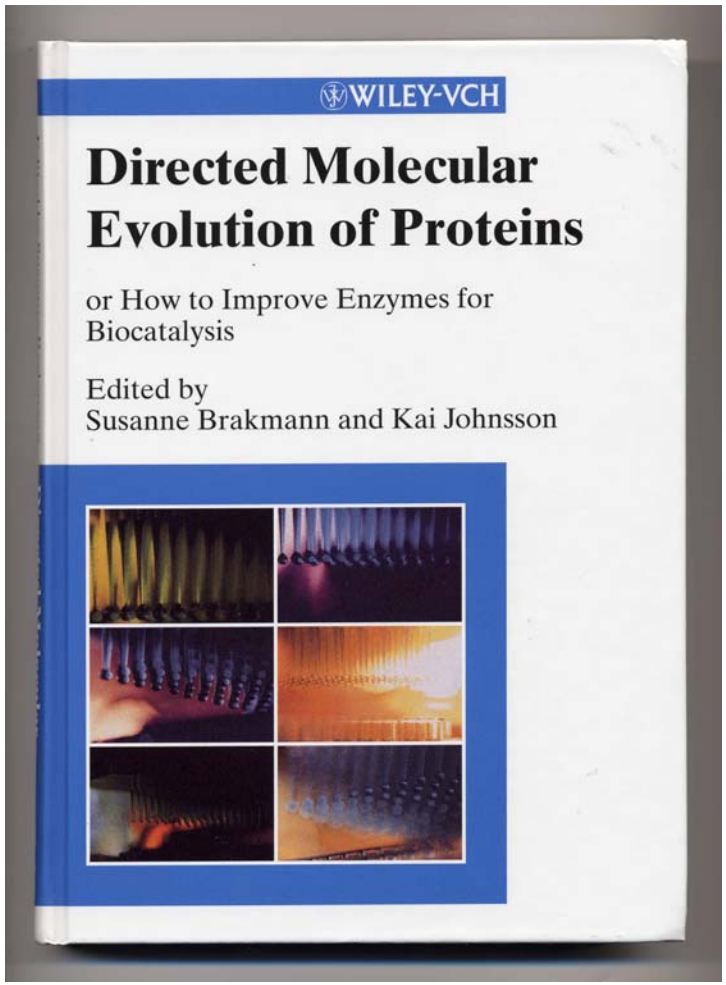
In silico optimization in the flow reactor: Evolutionary Trajectory

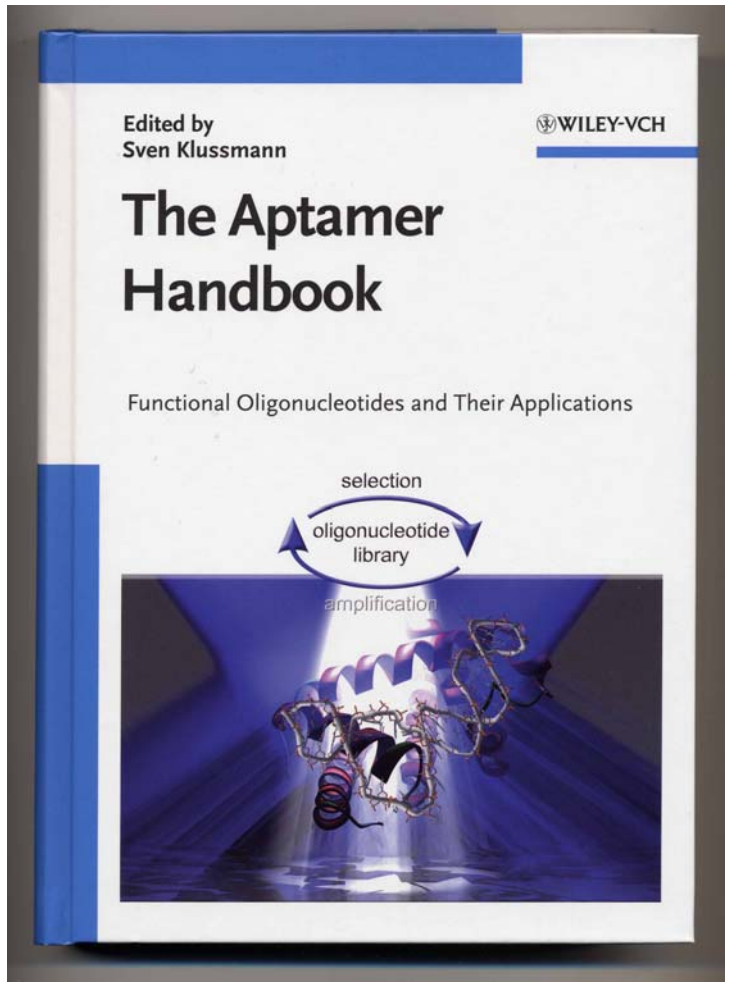

start of optimization
start of optimization
end of optimization
end of optimization
Cost function
Genotype space


start of optimization
start of optimization
target
Cost function
Genotype space

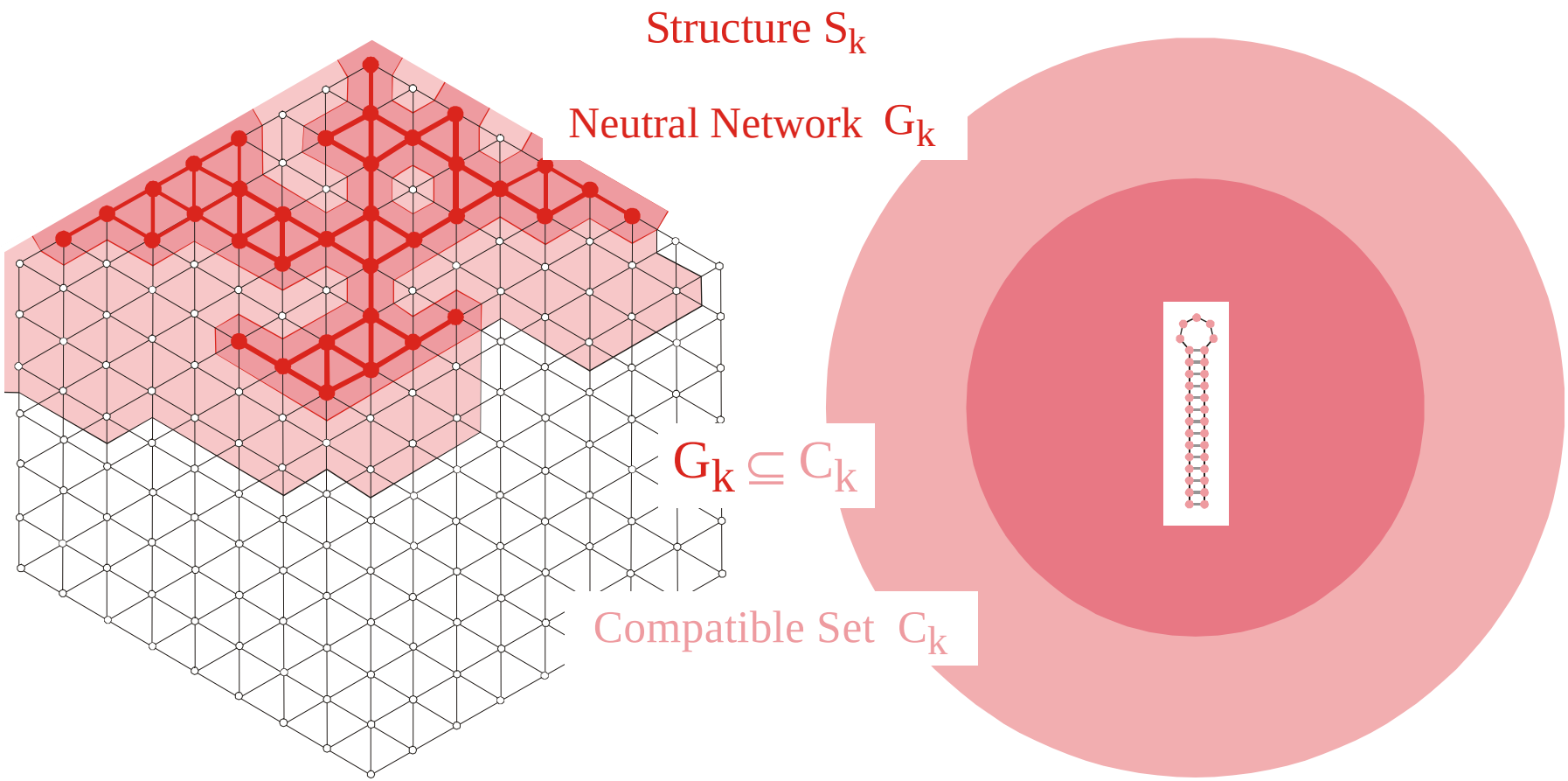


A sketch of optimization on neutral networks

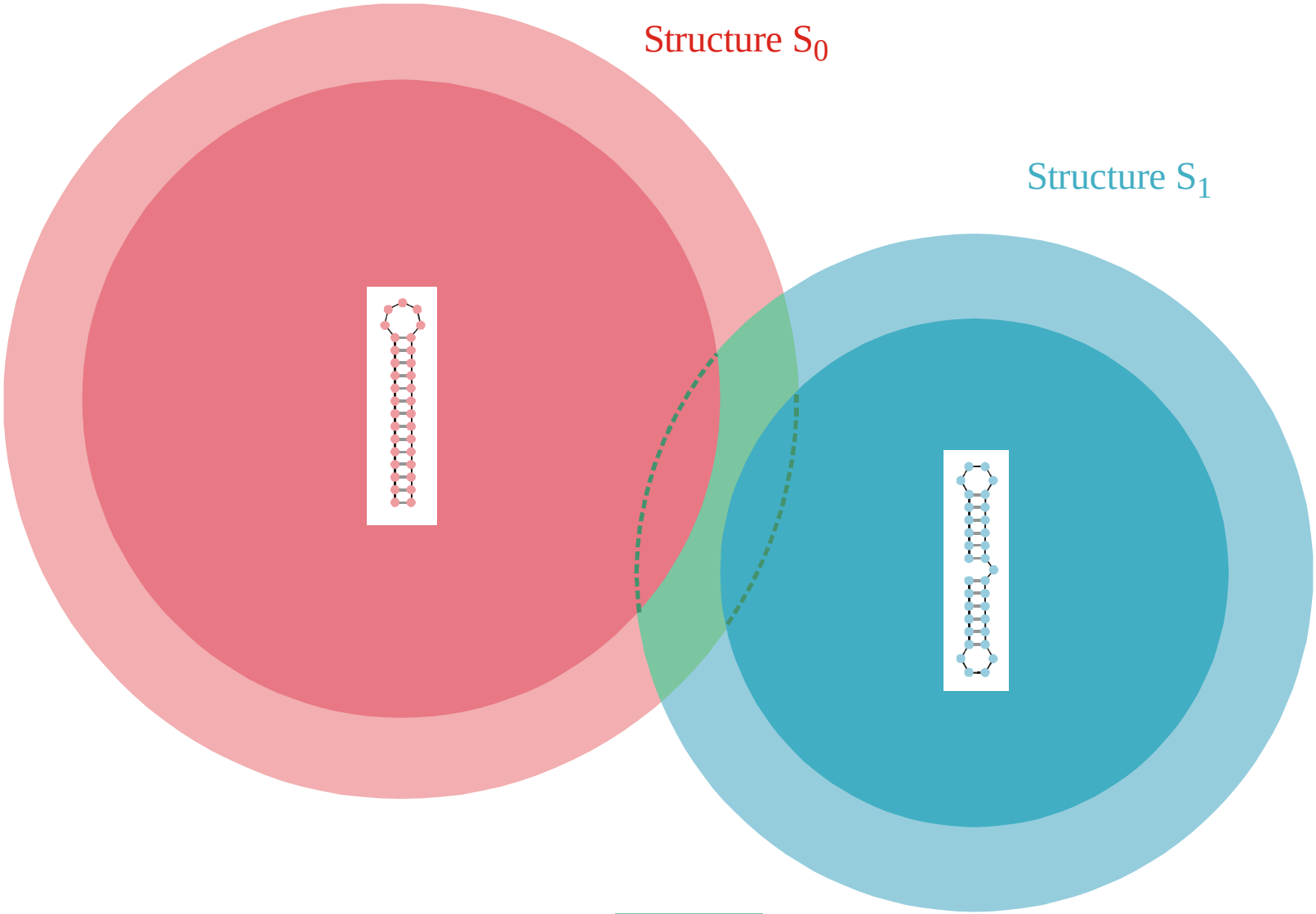




Application of molecular evolution to problems in biotechnology



The **compatible set C_k** of a structure **S_k** consists of all sequences which form **S_k** as its minimum free energy structure (the neutral network **G_k**) or one of its suboptimal structures.



Intersection of two compatible sets: $C_0 \cap C_1$

The intersection of two compatible sets is always non empty: $C_0 \cap C_1 \notin \varnothing$

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

0092-8240/97 $17.00 + 0.00

S0092-8240(96)00089-4

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

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

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

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

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

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

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

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

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

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

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

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

A ribozyme switch

E.A.Schultes, D.B.Bartel, Science **289** (2000), 448-452

minus the background levels observed in the HSP in the control (Sar1-GDP–containing) incubation that prevents COPII vesicle formation. In the microsome control, the level of p115-SNARE associations was less than 0.1%.
46. C. M. Carr, E. Grote, M. Munson, F. M. Hughson, P. J. Novick, *J. Cell Biol.* **146**, 333 (1999).
47. C. Ungermann, B. J. Nichols, H. R. Pelham, W. Wickner, *J. Cell Biol.* **140**, 61 (1998).
48. E. Grote and P. J. Novick, *Mol. Biol. Cell* **10**, 4149 (1999).
49. P. Uetz *et al.*, *Nature* **403**, 623 (2000).
50. GST-SNARE proteins were expressed in bacteria and purified on glutathione-Sepharose beads using standard methods. Immobilized GST-SNARE protein (0.5 μM) was incubated with rat liver cytosol (20 mg) or purified recombinant p115 (0.5 μM) in 1 ml of NS buffer containing 1% BSA for 2 hours at 4°C with rotation. Beads were briefly spun (3000 rpm for 10 s) and sequentially washed three times with NS buffer and three times with NS buffer supplemented with 150 mM NaCl. Bound proteins were eluted three times in 50 μl of 50 mM tris-HCl (pH 8.5), 50 mM reduced glutathione, 150 mM NaCl, and 0.1% Triton X-100 for 15 min at 4°C with intermittent mixing, and elutes were pooled. Proteins were precipitated by MeOH/CH_3Cl and separated by SDS–polyacrylamide gel electrophoresis (PAGE) followed by immunoblotting using p115 mAb 13F12.
51. V. Rybin *et al.*, *Nature* **383**, 266 (1996).
52. K. G. Hardwick and H. R. Pelham, *J. Cell Biol.* **119**, 513 (1992).
53. A. P. Newman, M. E. Groesch, S. Ferro-Novick, *EMBO J.* **11**, 3609 (1992).
54. A. Spang and R. Schekman, *J. Cell Biol.* **143**, 589 (1998).
55. M. F. Rexach, M. Latterich, R. W. Schekman, *J. Cell Biol.* **126**, 1133 (1994).
56. A. Mayer and W. Wickner, *J. Cell Biol.* **136**, 307 (1997).
57. M. D. Turner, H. Plutner, W. E. Balch, *J. Biol. Chem.* **272**, 13479 (1997).
58. A. Price, D. Seals, W. Wickner, C. Ungermann, *J. Cell Biol.* **148**, 1231 (2000).
59. X. Cao and C. Barlowe, *J. Cell Biol.* **149**, 55 (2000).
60. G. G. Tall, H. Hama, D. B. DeWald, B. F. Horazdovsky, *Mol. Biol. Cell* **10**, 1873 (1999).
61. C. G. Burd, M. Peterson, C. R. Cowles, S. D. Emr, *Mol. Biol. Cell* **8**, 1089 (1997).
62. M. R. Peterson, C. G. Burd, S. D. Emr, *Curr. Biol.* **9**, 159 (1999).
63. M. G. Waters, D. O. Clary, J. E. Rothman, *J. Cell Biol.* **118**, 1015 (1992).
64. D. M. Walter, K. S. Paul, M. G. Waters, *J. Biol. Chem.* **273**, 29565 (1998).
65. N. Hui *et al.*, *Mol. Biol. Cell* **8**, 1777 (1997).
66. T. E. Kreis, *EMBO J.* **5**, 931 (1986).
67. H. Plutner, H. W. Davidson, J. Saraste, W. E. Balch, *J. Cell Biol.* **119**, 1097 (1992).
68. D. S. Nelson *et al.*, *J. Cell Biol.* **143**, 319 (1998).
69. We thank G. Waters for p115 cDNA and p115 mAbs; G. Warren for p97 and p47 antibodies; R. Scheller for rbet1, membrin, and sec22 cDNAs; H. Plutner for excellent technical assistance; and P. Tan for help during the initial phase of this work. Supported by NIH grants GM 33301 and GM42336 and National Cancer Institute grant CA58689 (W.E.B.), a NIH National Research Service Award (B.D.M.), and a Wellcome Trust International Traveling Fellowship (B.B.A.).

20 March 2000; accepted 22 May 2000

One Sequence, Two Ribozymes: Implications for the Emergence of New Ribozyme Folds

Erik A. Schultes and David P. Bartel*

We describe a single RNA sequence that can assume either of two ribozyme folds and catalyze the two respective reactions. The two ribozyme folds share no evolutionary history and are completely different, with no base pairs (and probably no hydrogen bonds) in common. Minor variants of this sequence are highly active for one or the other reaction, and can be accessed from prototype ribozymes through a series of neutral mutations. Thus, in the course of evolution, new RNA folds could arise from preexisting folds, without the need to carry inactive intermediate sequences. This raises the possibility that biological RNAs having no structural or functional similarity might share a common ancestry. Furthermore, functional and structural divergence might, in some cases, precede rather than follow gene duplication.

Related protein or RNA sequences with the same folded conformation can often perform very different biochemical functions, indicating that new biochemical functions can arise from preexisting folds. But what evolutionary mechanisms give rise to sequences with new macromolecular folds? When considering the origin of new folds, it is useful to picture, among all sequence possibilities, the distribution of sequences with a particular fold and function. This distribution can range very far in sequence space (*1*). For example, only seven nucleotides are strictly conserved among the group I self-splicing introns, yet secondary (and presumably tertiary) structure within the core of the ribozyme is preserved (*2*). Because these disparate isolates have the same fold and function, it is thought that they descended from a common ancestor through a series of mutational variants that were each functional. Hence, sequence heterogeneity among divergent isolates implies the existence of paths through sequence space that have allowed neutral drift from the ancestral sequence to each isolate. The set of all possible neutral paths composes a "neutral network," connecting in sequence space those widely dispersed sequences sharing a particular fold and activity, such that any sequence on the network can potentially access very distant sequences by neutral mutations (*3–5*).

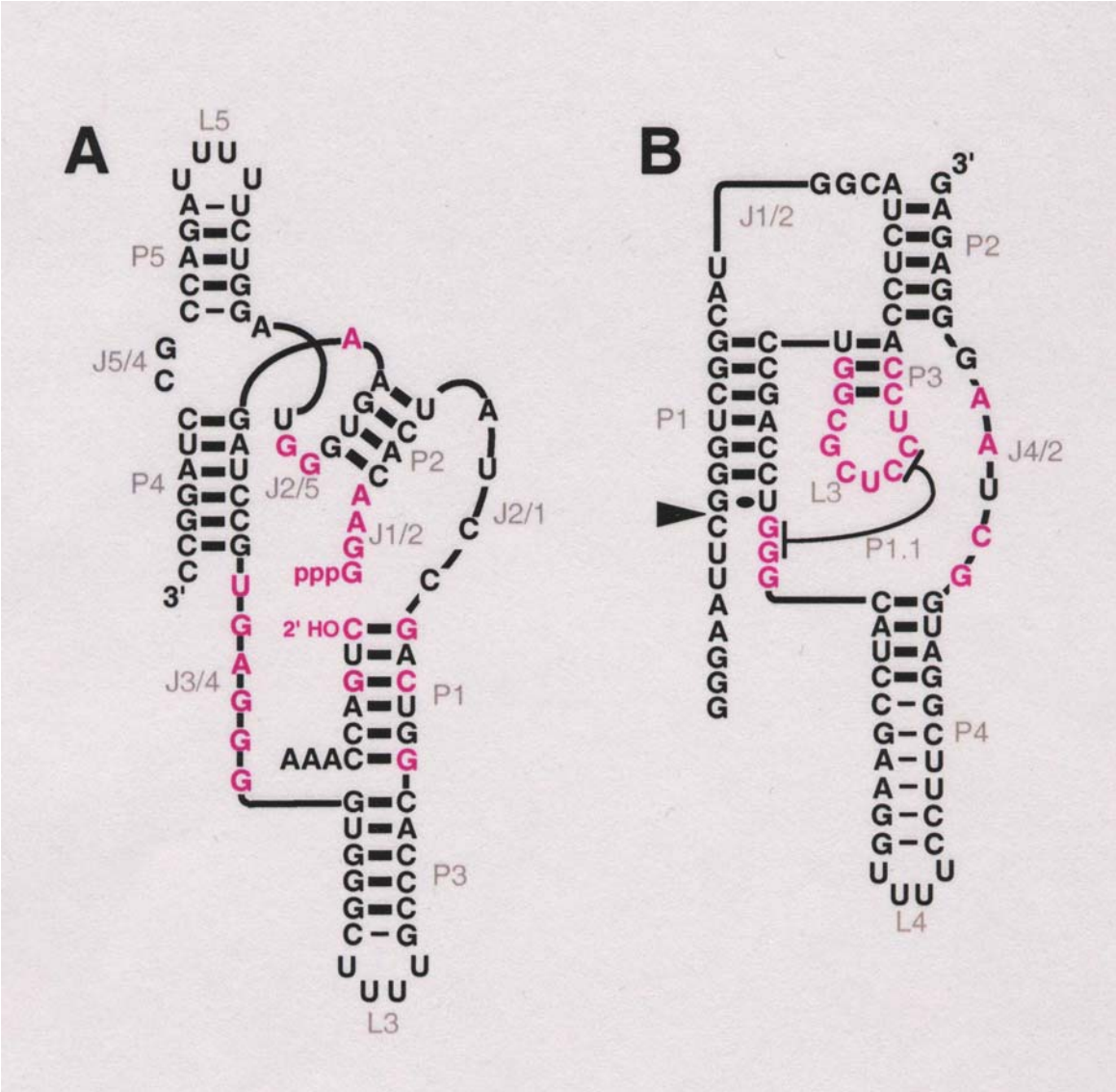
Theoretical analyses using algorithms for predicting RNA secondary structure have suggested that different neutral networks are interwoven and can approach each other very closely (*3*, *5–8*). Of particular interest is whether ribozyme neutral networks approach each other so closely that they intersect. If so, a single sequence would be capable of folding into two different conformations, would have two different catalytic activities, and could access by neutral drift every sequence on both networks. With intersecting networks, RNAs with novel structures and activities could arise from previously existing ribozymes, without the need to carry nonfunctional sequences as evolutionary intermediates. Here, we explore the proximity of neutral networks experimentally, at the level of RNA function. We describe a close apposition of the neutral networks for the hepatitis delta virus (HDV) self-cleaving ribozyme and the class III self-ligating ribozyme.

In choosing the two ribozymes for this investigation, an important criterion was that they share no evolutionary history that might confound the evolutionary interpretations of our results. Choosing at least one artificial ribozyme ensured independent evolutionary histories. The class III ligase is a synthetic ribozyme isolated previously from a pool of random RNA sequences (*9*). It joins an oligonucleotide substrate to its 5′ terminus. The prototype ligase sequence (Fig. 1A) is a shortened version of the most active class III variant isolated after 10 cycles of in vitro selection and evolution. This minimal construct retains the activity of the full-length isolate (*10*). The HDV ribozyme carries out the site-specific self-cleavage reactions needed during the life cycle of HDV, a satellite virus of hepatitis B with a circular, single-stranded RNA genome (*11*). The prototype HDV construct for our study (Fig. 1B) is a shortened version of the antigenomic HDV ribozyme (*12*), which undergoes self-cleavage at a rate similar to that reported for other antigenomic constructs (*13*, *14*).

The prototype class III and HDV ribozymes have no more than the 25% sequence identity expected by chance and no fortuitous structural similarities that might favor an intersection of their two neutral networks. Nevertheless, sequences can be designed that simultaneously satisfy the base-pairing requirements

Whitehead Institute for Biomedical Research and Department of Biology, Massachusetts Institute of Technology, 9 Cambridge Center, Cambridge, MA 02142, USA.

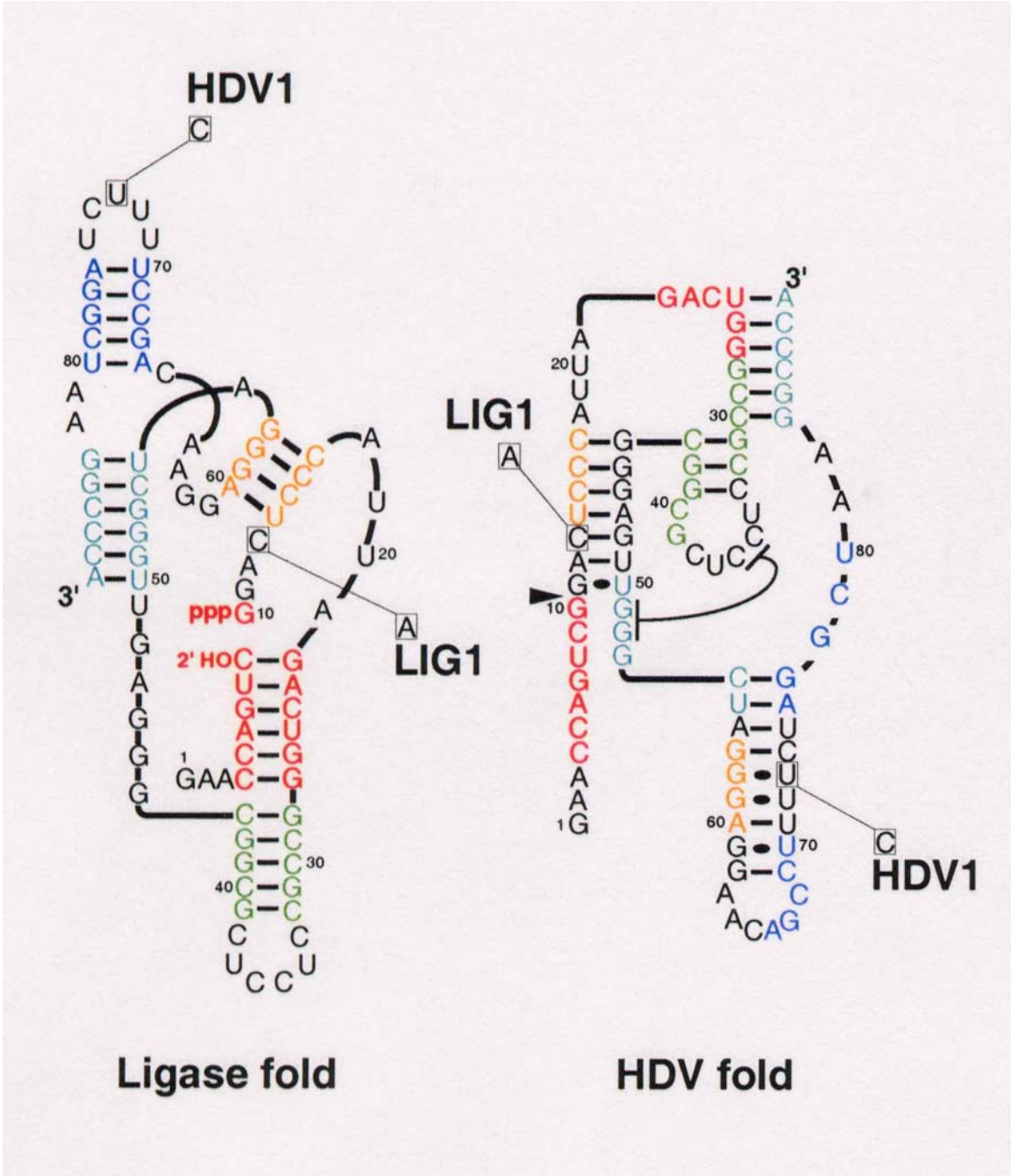
*To whom correspondence should be addressed. E-mail: dbartel@wi.mit.edu

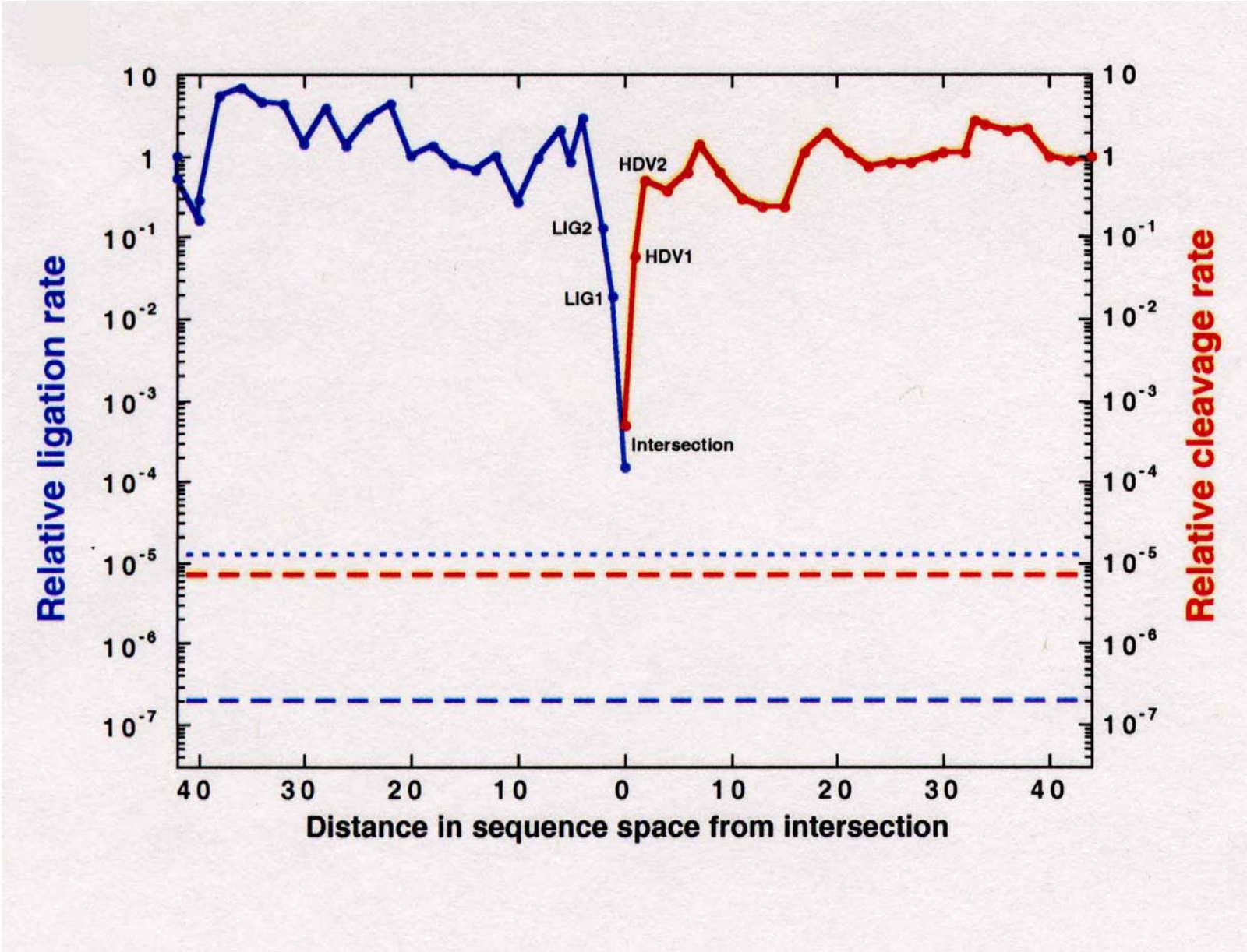


Two ribozymes of chain lengths n = 88 nucleotides: An artificial ligase (**A**) and a natural cleavage ribozyme of hepatitis-δ-virus (**B**)

The sequence at the ***intersection***:

An RNA molecules which is 88 nucleotides long and can form both structures





Two neutral walks through sequence space with conservation of structure and catalytic activity

Web-Page for further information:

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