The advent of information and combinatorial complexity: Understanding Darwinian evolution at the molecular level

Peter Schuster

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

ESF-COST Conference on Systems Chemistry

Acquafredda di Maratea, 03.– 08.10.2008

Web-Page for further information:

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

What is information ?

- *Information* is (only) what is understood.
- *Information* is (only) what creates information.

Carl Friedrich von Weizsäcker, 1912-2007, German physicist and philosopher.

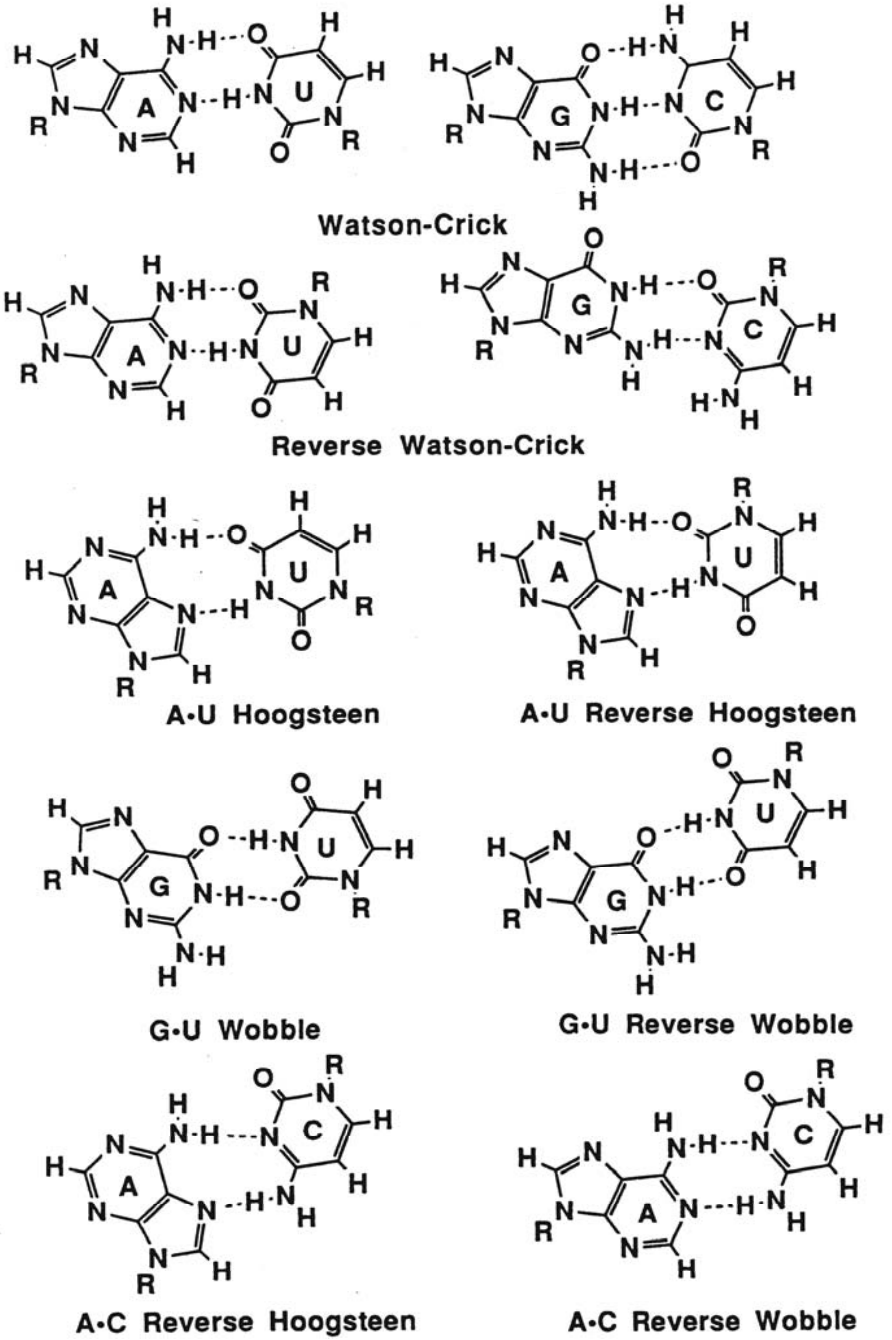
Information in biology

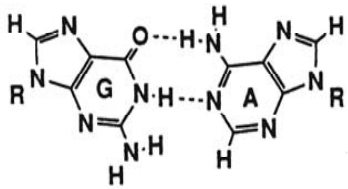
- ***Understanding of information*** is interpreted as decoding,
- ***maintenance of information*** requires reproduction, and
- ***creation of information*** occurs through adaptation to the environment by means of a Darwinian mechanism of variation and selection.

1. Requirements for information processing
2. The chemistry of Darwinian evolution
3. RNA sequences and structures
4. Consequences of neutrality
5. Evolutionary optimization of RNA structure

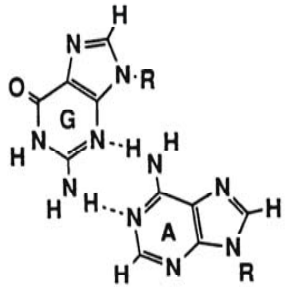
1. **Requirements for information processing**
2. The chemistry of Darwinian evolution
3. RNA sequences and structures
4. Consequences of neutrality
5. Evolutionary optimization of RNA structure

Classification of purine-pyrimidine base pairs

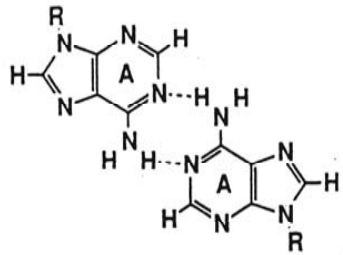




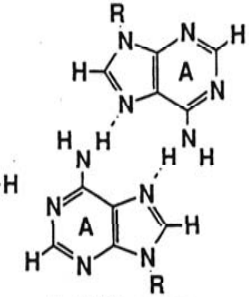
G•A N1-N1, carbonyl-amino



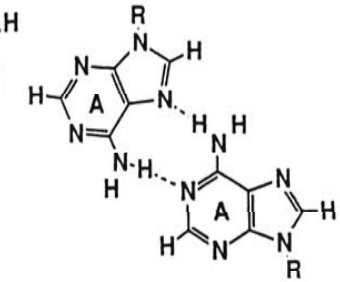
G•A N3-amino, amino-N1



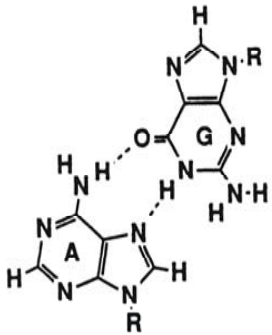
A•A N1-amino, symmetric



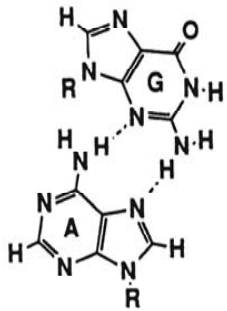
A•A N7-amino, symmetric



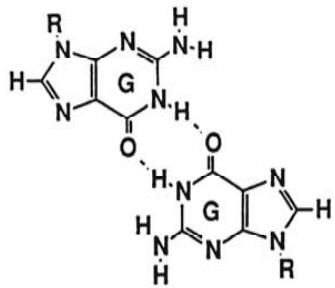
A•A N1-amino, N7-amino



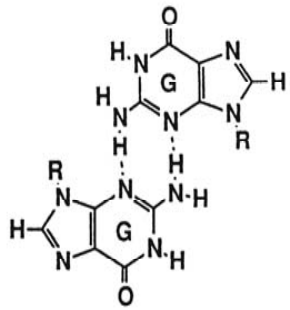
A•G N7-N1, amino-carbonyl



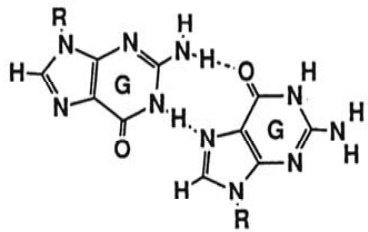
A•G N7-amino, amino-N3



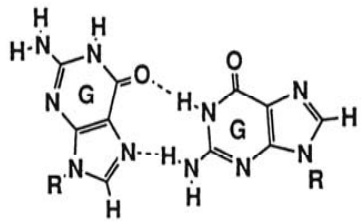
G•G N1-carbonyl, symmetric



G•G N3-amino, symmetric

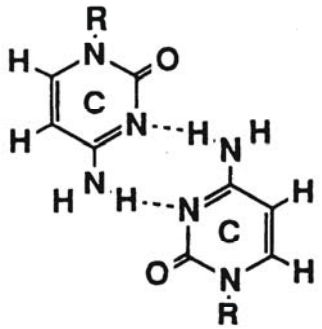


G•G N7-N1, carbonyl-amino

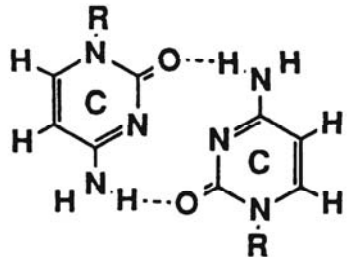


G•G N1-carbonyl, N7-amino

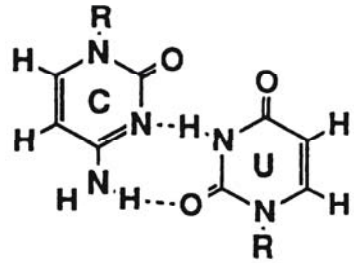
Classification of purine-purine base pairs



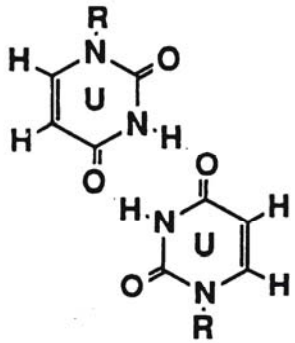
C•C N3-amino, symmetric



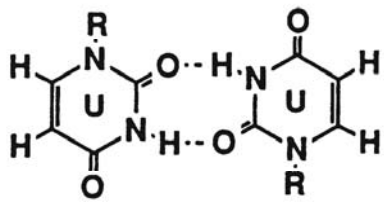
C•C carbonyl-amino, symmetric



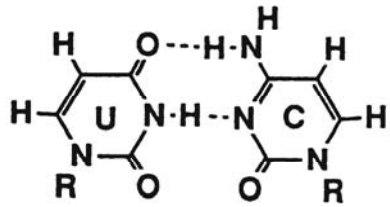
C•U N3-N3, 2-carbonyl-amino



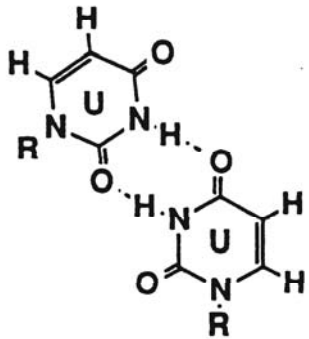
U•U 4-carbonyl-N3, symmetric



U•U 2-carbonyl-N3, symmetric



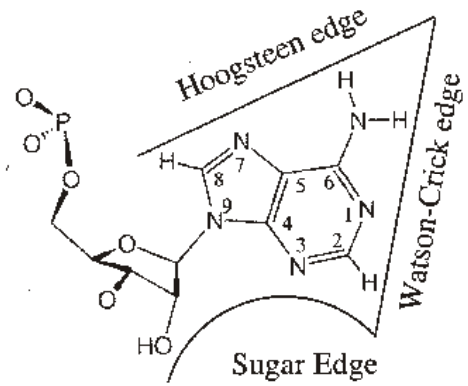
U•C N3-N3, 4-carbonyl-amino

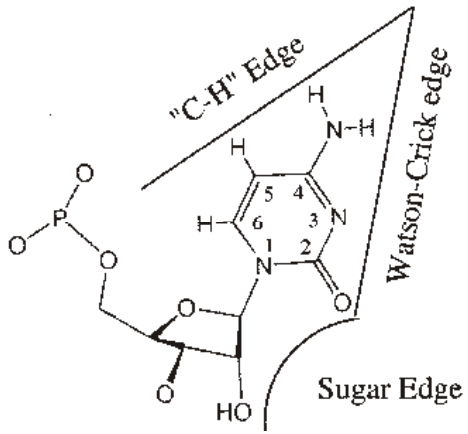


U•U 2-carbonyl-N3, 4-carbonyl-N3

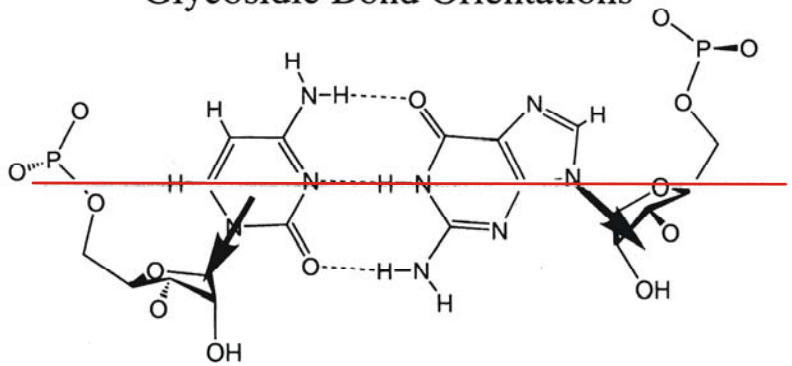
Classification of pyrimidine-pyrimidine base pairs

Interacting Edges

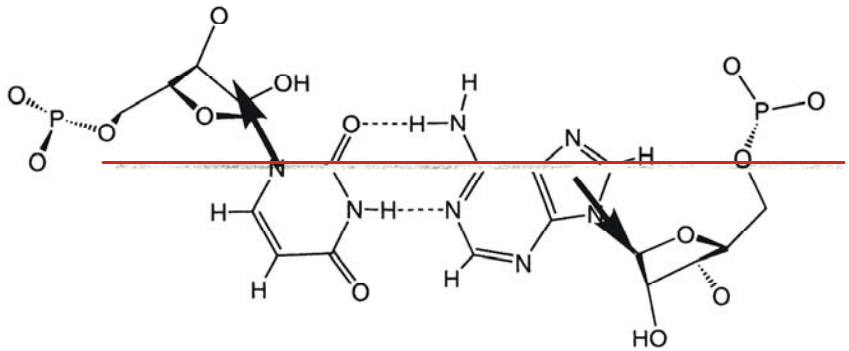




Glycosidic Bond Orientations



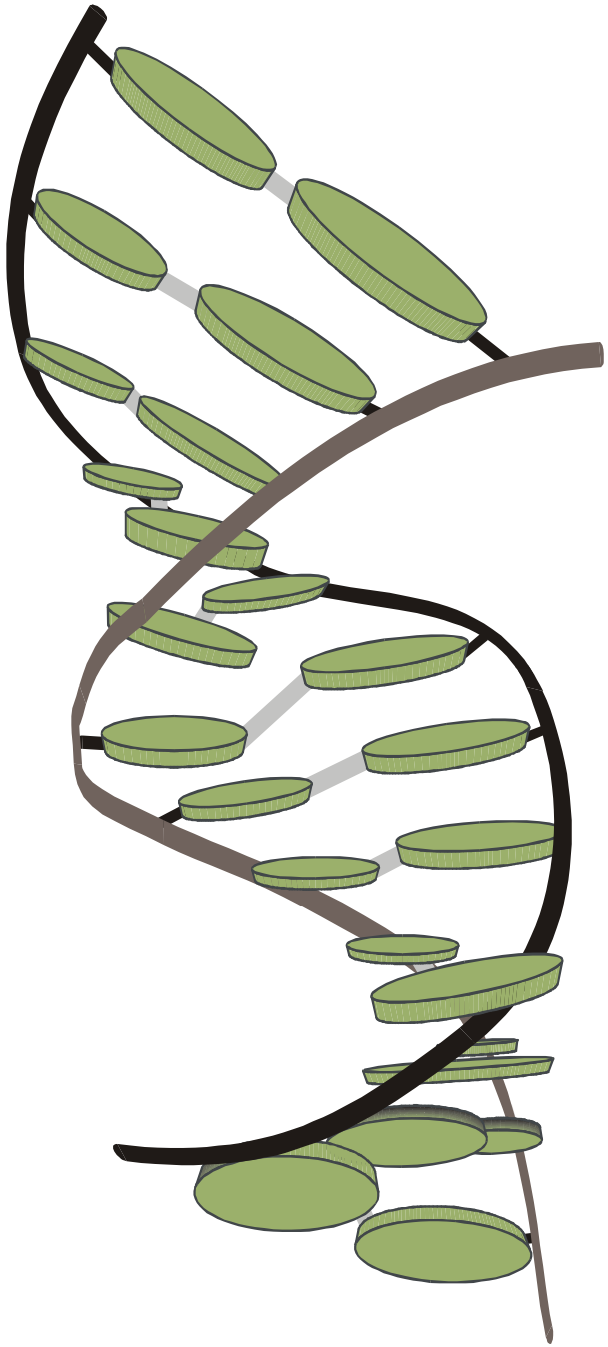
Cis orientation of the Glycosidic Bonds

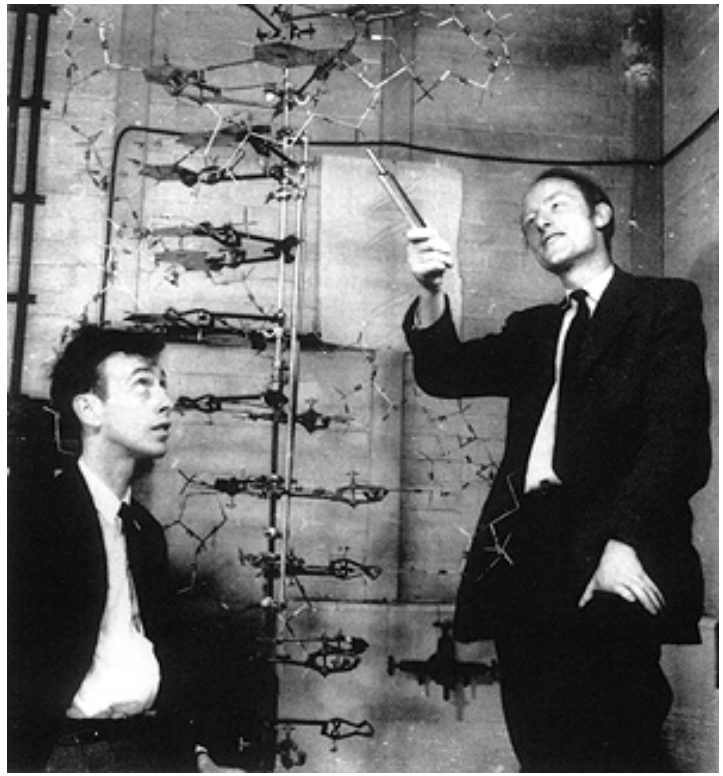


Trans orientation of the Glycosidic Bonds

General classification of base pairs

N.B. Leontis and E. Westhof, *RNA* **7**:499-512 (2001)





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

Nobel prize 1962

1953 – 2003 fifty years double helix

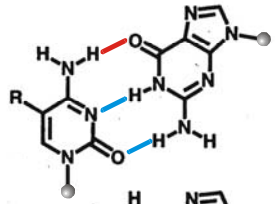
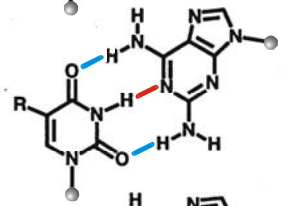
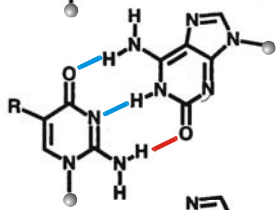
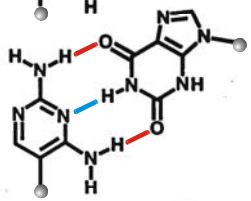
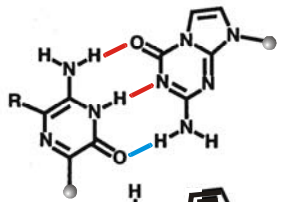
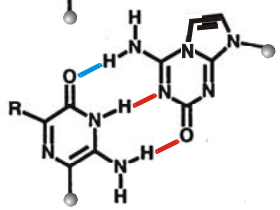
The three-dimensional structure of a short double helical stack of B-DNA

Color code:

Donor—Acceptor

Acceptor—Donor

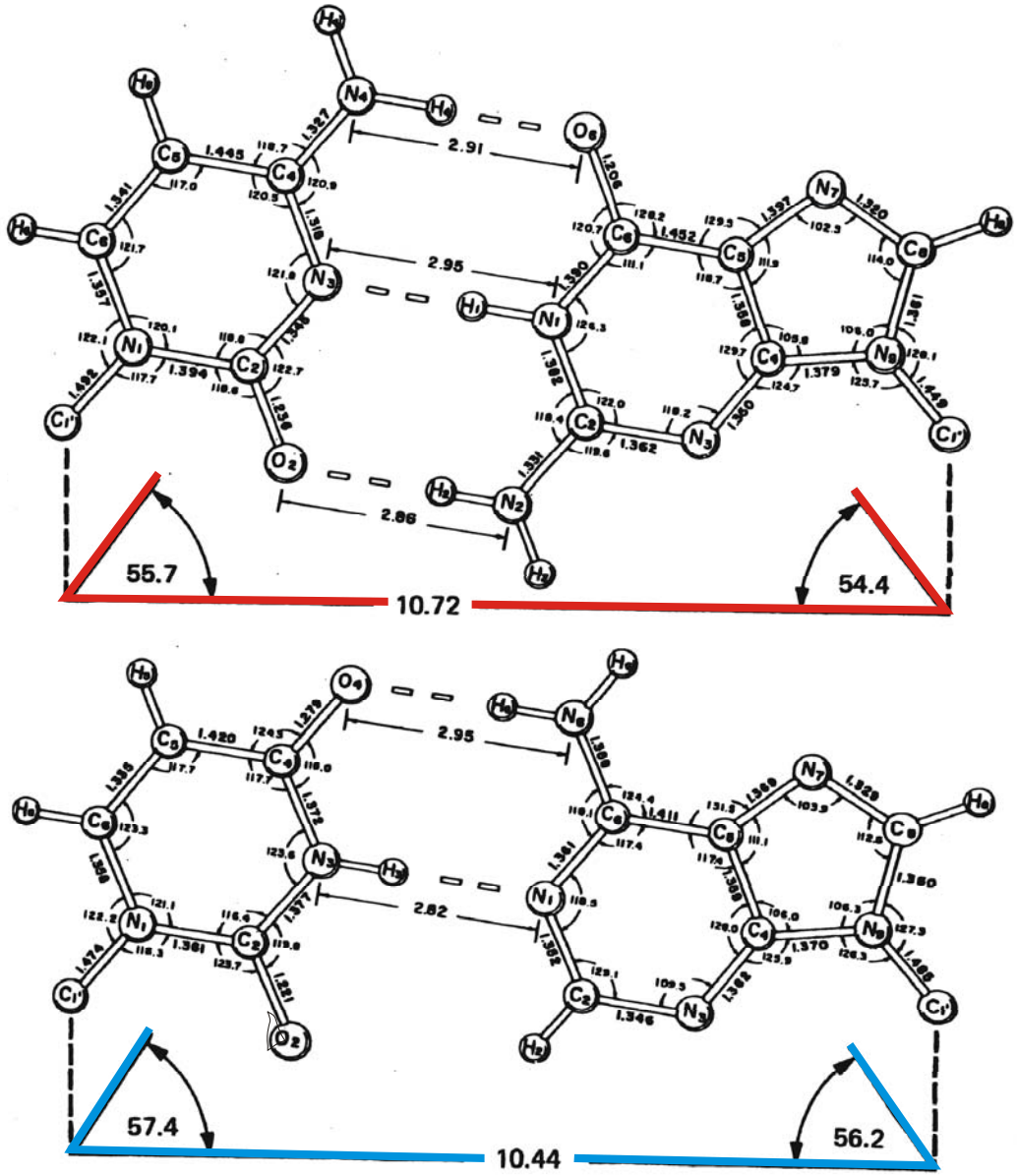
Hydrogen bonding patterns for Watson-Crick base pairs

2-keto, 4-amino pyrimidine	Donor Acceptor Acceptor	C		G	Acceptor Donor Donor	2-amino,6-keto purine
2,4-diketo pyrimidine	Acceptor Donor Acceptor	U		``A´´	Donor Acceptor Donor	2,6-diamino purine
2-amino, 4-keto pyrimidine	Acceptor Acceptor Donor				Donor Donor Acceptor	2-keto, 6-amino purine
2,6-diamino pyrimidine	Donor Acceptor Donor				Acceptor Donor Acceptor	2,6-diketo purine
2-amino, 6-keto pyrazine	Donor Donor Acceptor				Acceptor Acceptor Donor	5-keto, 7-amino, 1,6,8-triaza indolicine
2-keto, 6-amino pyrazine	Acceptor Donor Donor				Donor Acceptor Acceptor	5-amino, 7-keto, 1,6,8-triaza indolicine

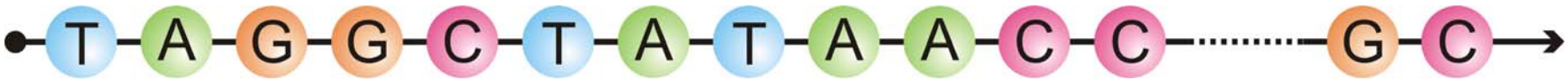
S.A. Benner *et al.*, Reading the palimpsest: Contemporary biochemical data and the RNA world. In: R.F.Gesteland and J.F.Atkins, eds. The RNA World, pp.27-70. CSHL Press, 1993

Canonical Watson-Crick base pairs:

cytosine – guanine
uracil – adenine



W.Saenger, Principles of Nucleic Acid Structure, Springer, Berlin 1984



4^n different sequences for chain length n

$n = 100$: $4^{100} = 1.6 \times 10^{60}$ sequences

Combinatorial complexity in biopolymer sequences

Information processing requires digitalization in the sense of „yes-or-no" decisions. Nature solves the problem through complementarity of nucleobases:

• *Biological information storage* in nucleic acids is extremely specific through applying the straightforward stereochemistry of the double helix.

• *Biological information processing* is overcoming thermodynamic restrictions without violating its rules.

• *Digitalization of biological information* is the key towards easily accessible combinatorial complexity and provides the basis for the inexhaustible reservoir of genotypes and shapes in nature.

1. Requirements for information processing
2. **The chemistry of Darwinian evolution**
3. RNA sequences and structures
4. Consequences of neutrality
5. Evolutionary optimization of RNA structure

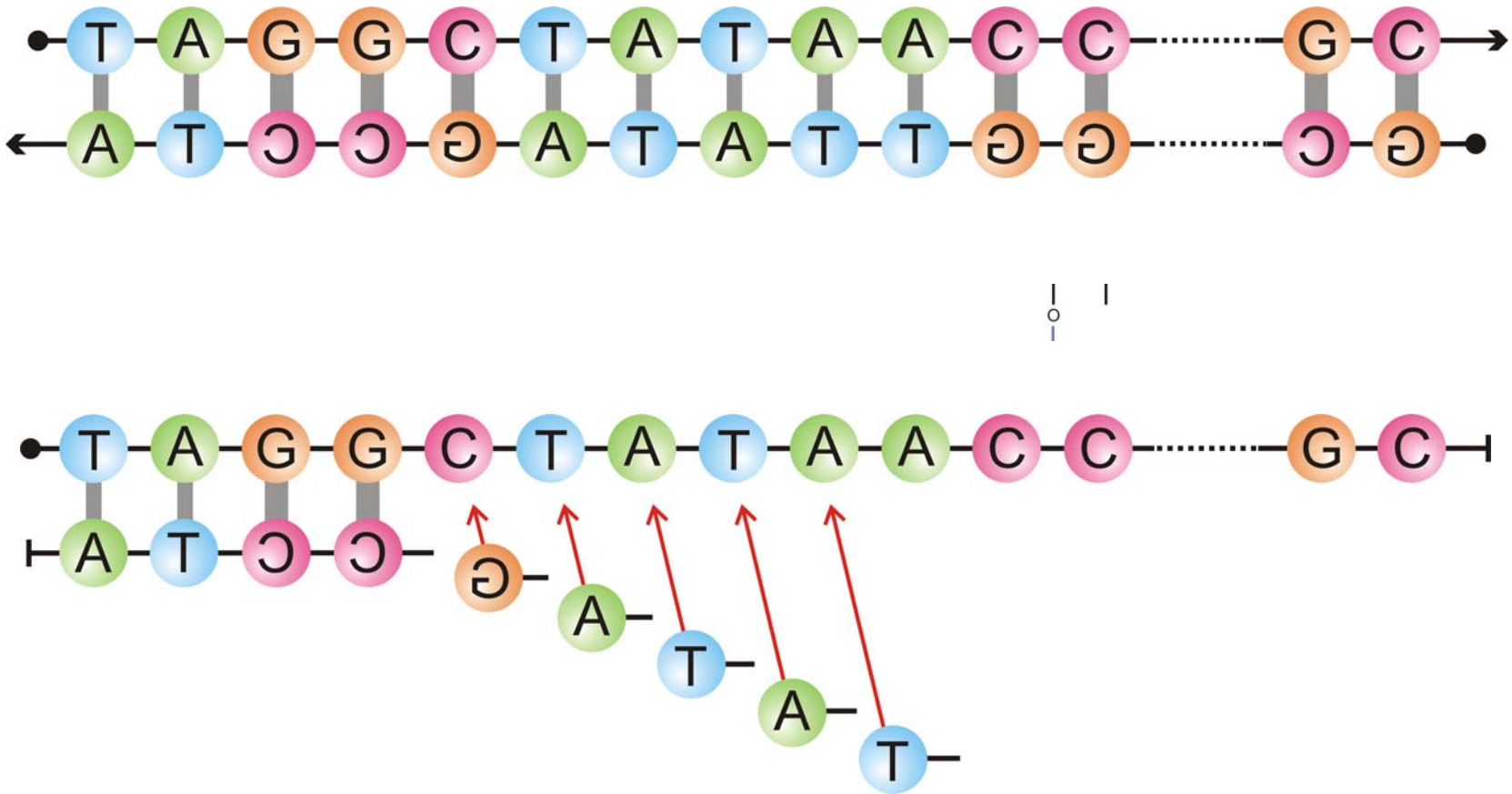
Three necessary conditions for Darwinian evolution are:

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

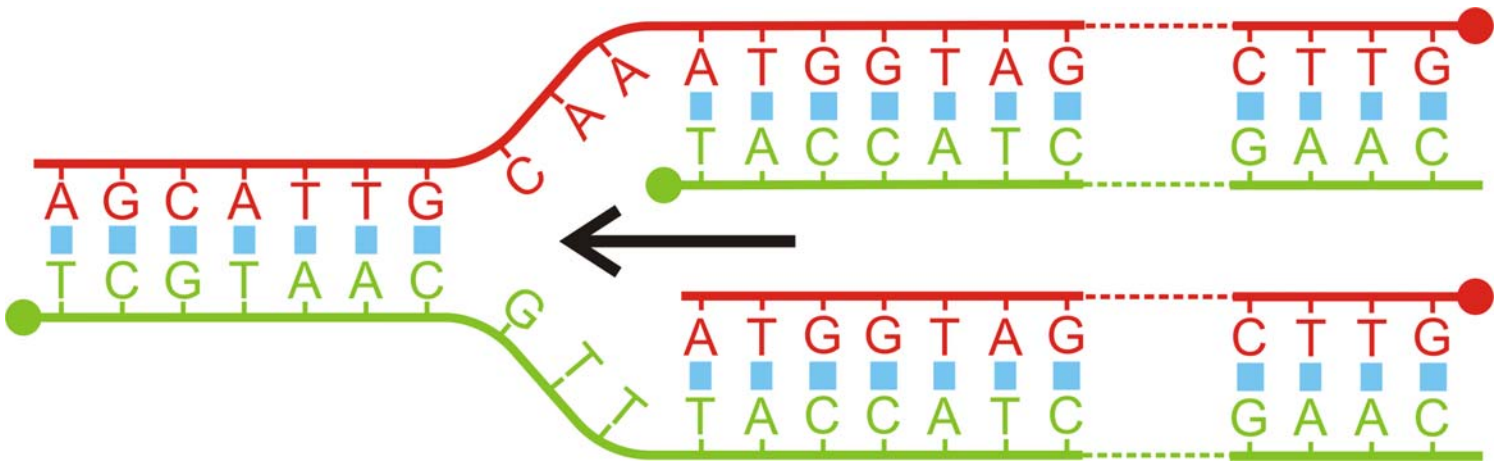
Variation through mutation and recombination operates on the **genotype** whereas the **phenotype** is the target of **selection**.

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**.

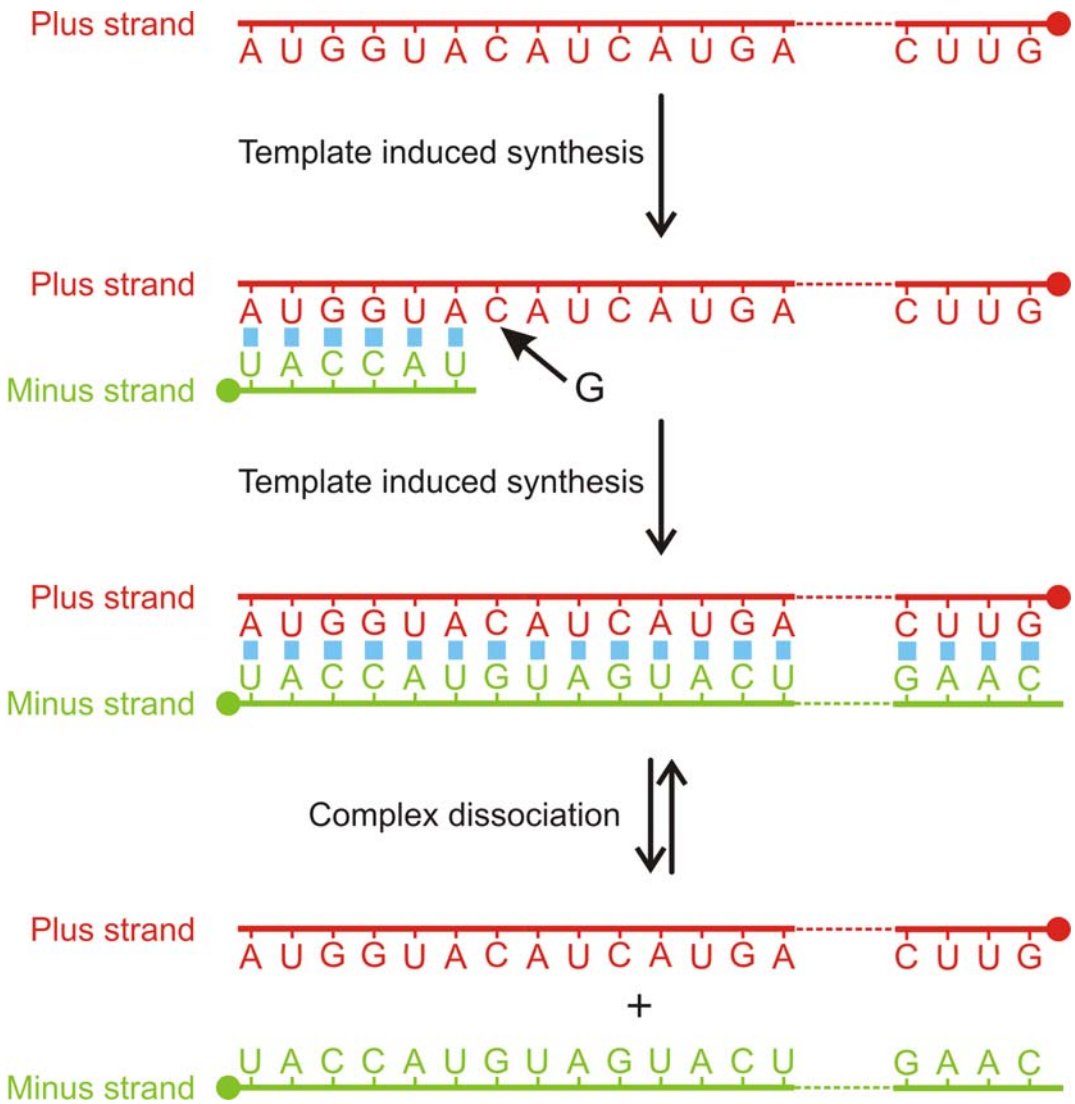


DNA structure and DNA replication



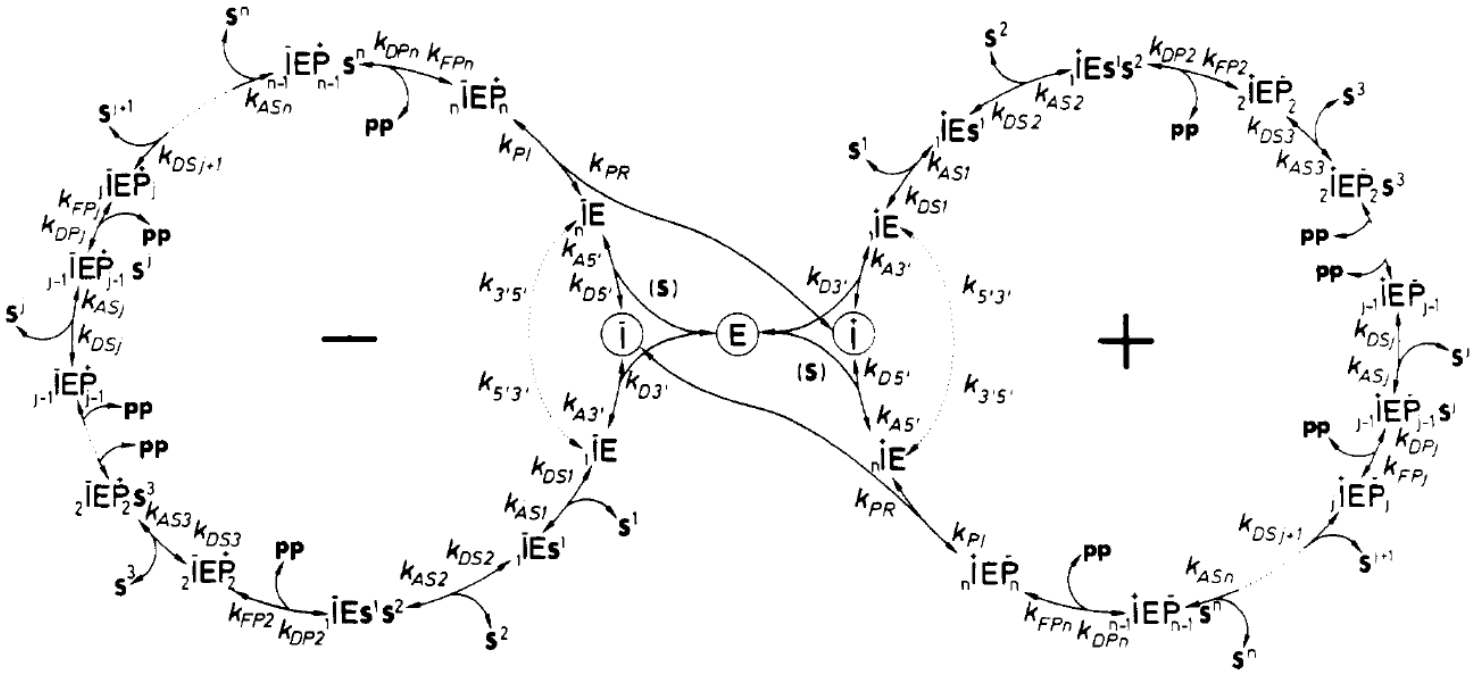
‚Replication fork' in DNA replication

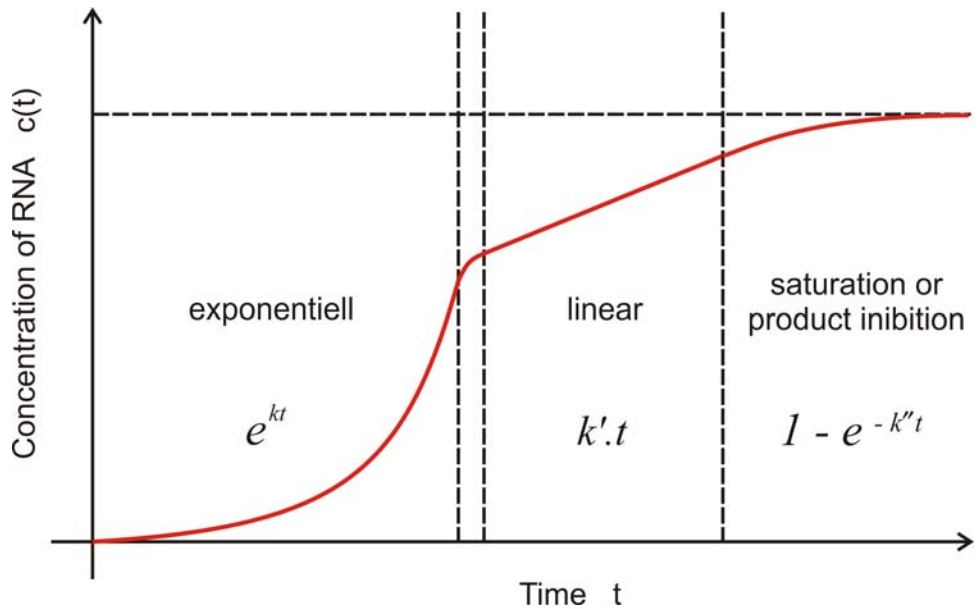
The mechanism of DNA replication is ‚semi-conservative'



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

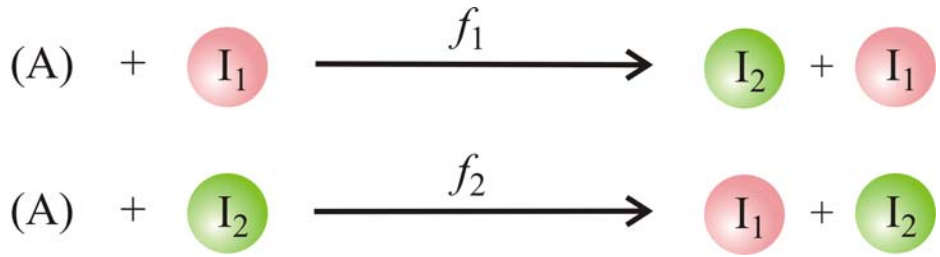
G≡C and **A=U**





Kinetics of RNA replication

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



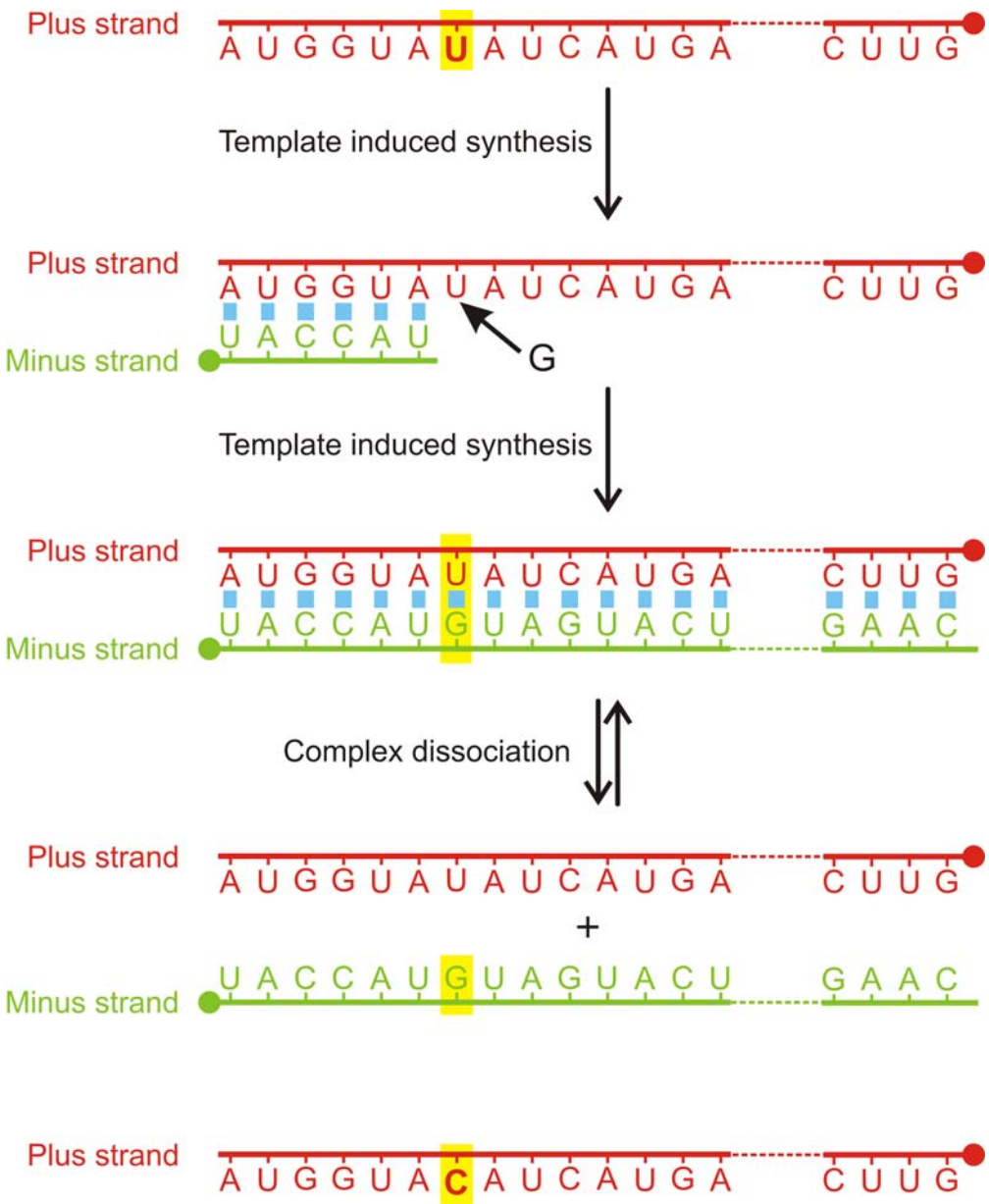
$$\frac{dx_1}{dt} = f_2\, x_2 \quad \text{and} \quad \frac{dx_2}{dt} = f_1\, x_1$$

$$x_1 = \sqrt{f_2}\;\xi_1\,, \quad x_2 = \sqrt{f_1}\;\xi_2\,, \quad \zeta = \xi_1 + \xi_2\,, \quad \eta = \xi_1 - \xi_2\,, \quad f = \sqrt{f_1 f_2}$$

$$\eta(t) = \eta(0)\, e^{-ft}$$

$$\zeta(t) = \zeta(0)\, e^{ft}$$

Complementary replication as the simplest molecular mechanism of reproduction



A point mutation is caused by an incorrect incorporation of a nucleobase into the growing chain during replication.

Replication and mutation are parallel chemical reactions.

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

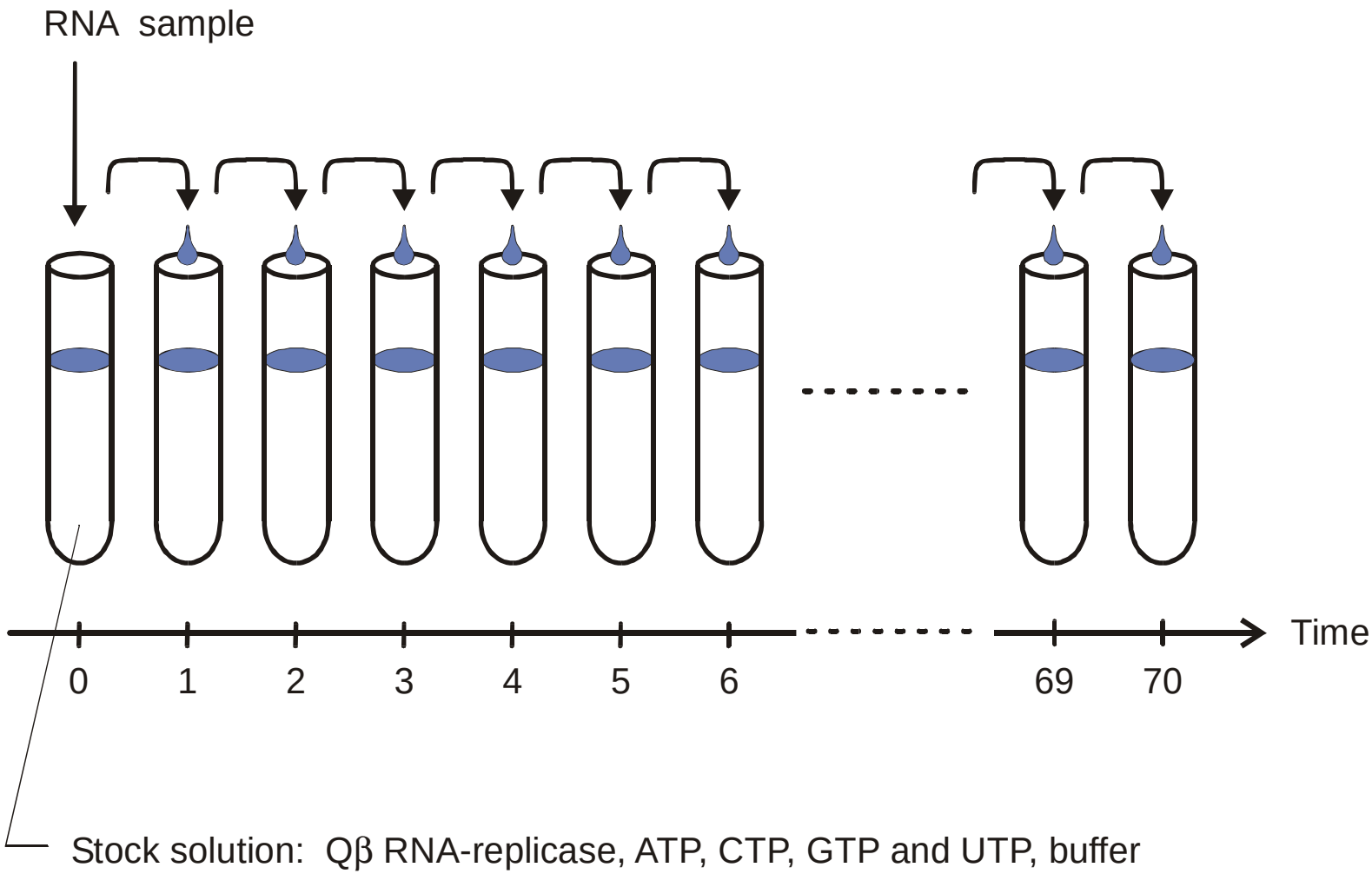
C.K.Biebricher, ***Darwinian selection of self-replicating RNA molecules***. Evolutionary Biology **16** (1983), 1-52

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



Application of serial transfer to RNA evolution in the test tube

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

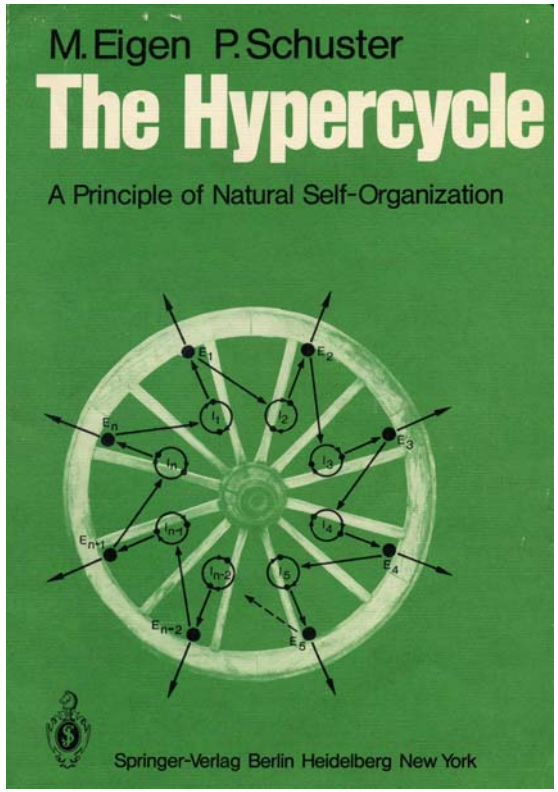
The mathematical analysis of dynamical systems using methods of differential topology, yields the result that there is only one type of mechanisms which fulfills the following requirements: The information stored in each single replicative unit (or reproductive cycle) must be maintained, i.e., the respective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

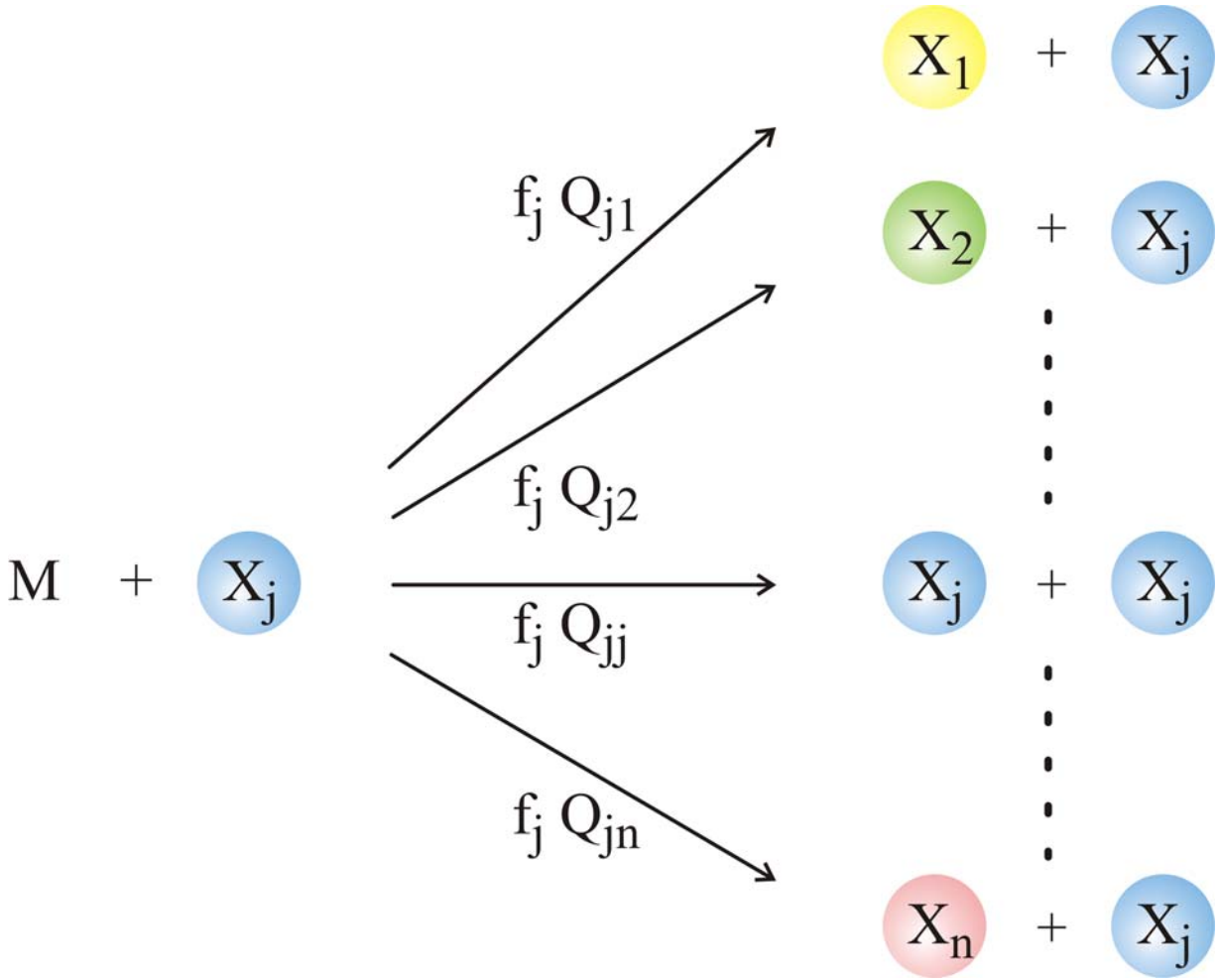
I. The Paradigm of Unity and Diversity in Evolution

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-

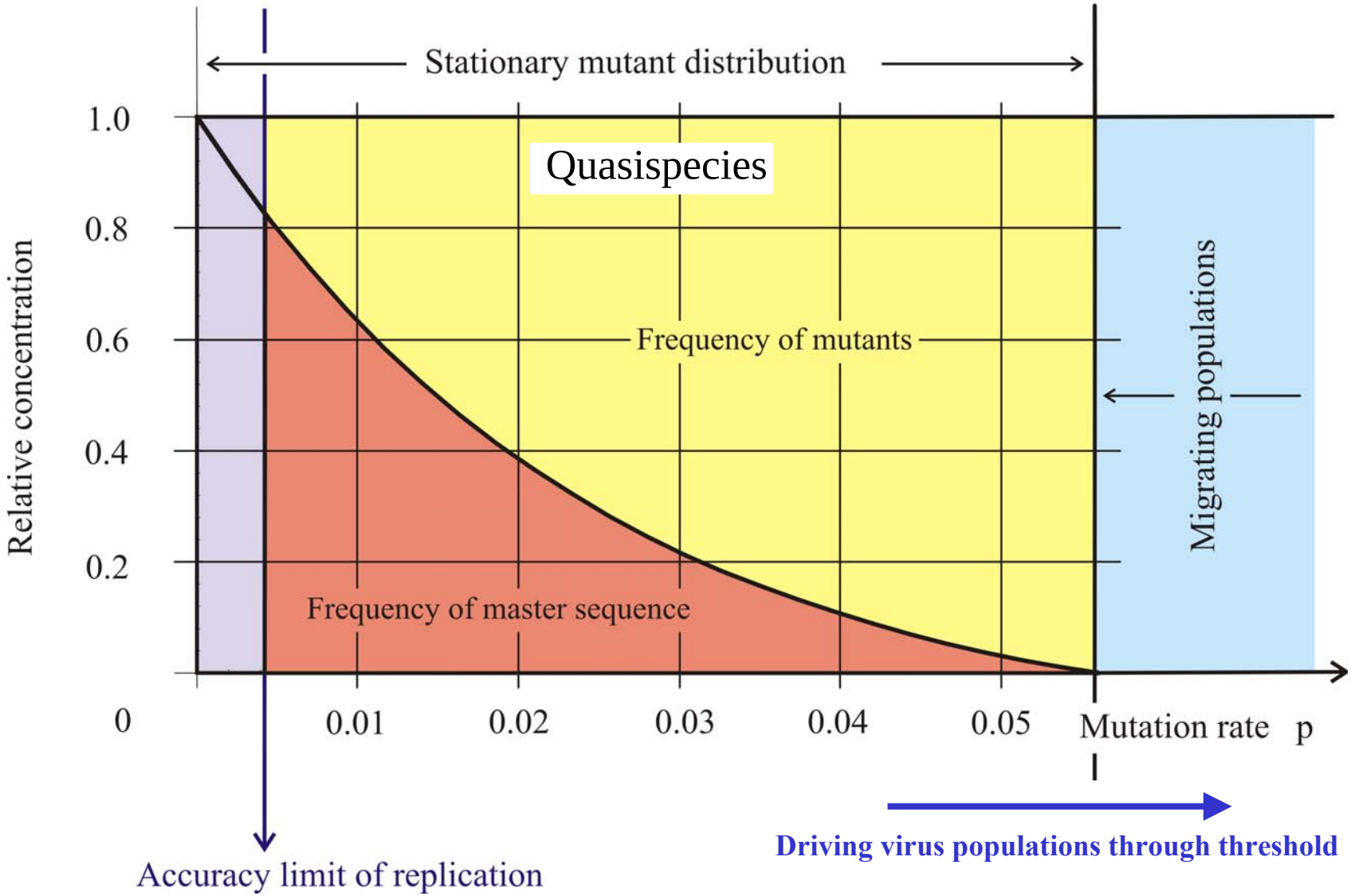


Chemical kinetics of molecular evolution

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



Chemical kinetics of replication and mutation as parallel reactions



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.

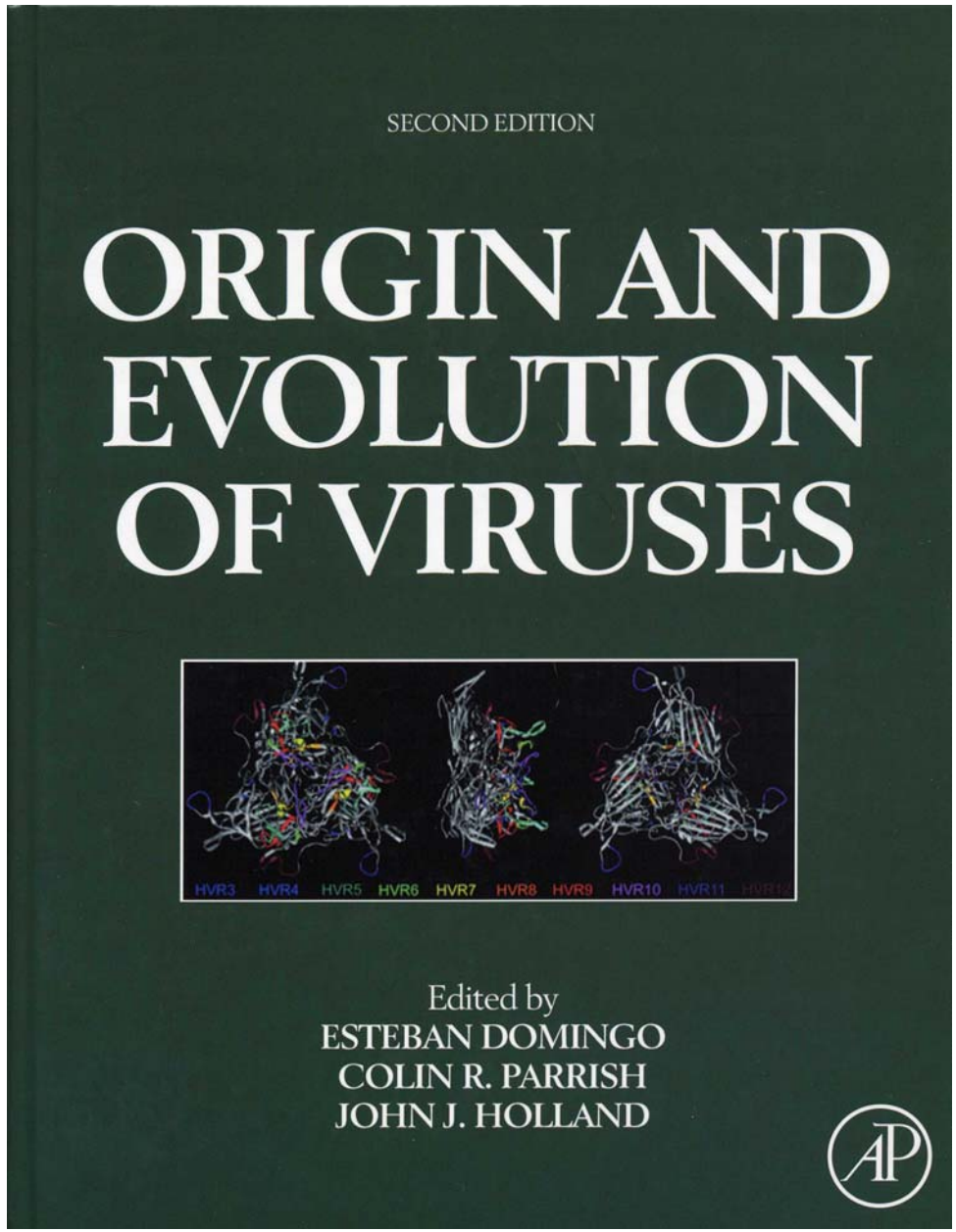
My thanks go also to the 19 authors who despite their busy schedules have taken time to prepare excellent manuscripts, to Elsevier staff for their prompt responses to my requests, and, last but not least, to Ms. Lucía Horrillo from Centro de Biología Molecular "Severo Ochoa" for her patient dealing with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

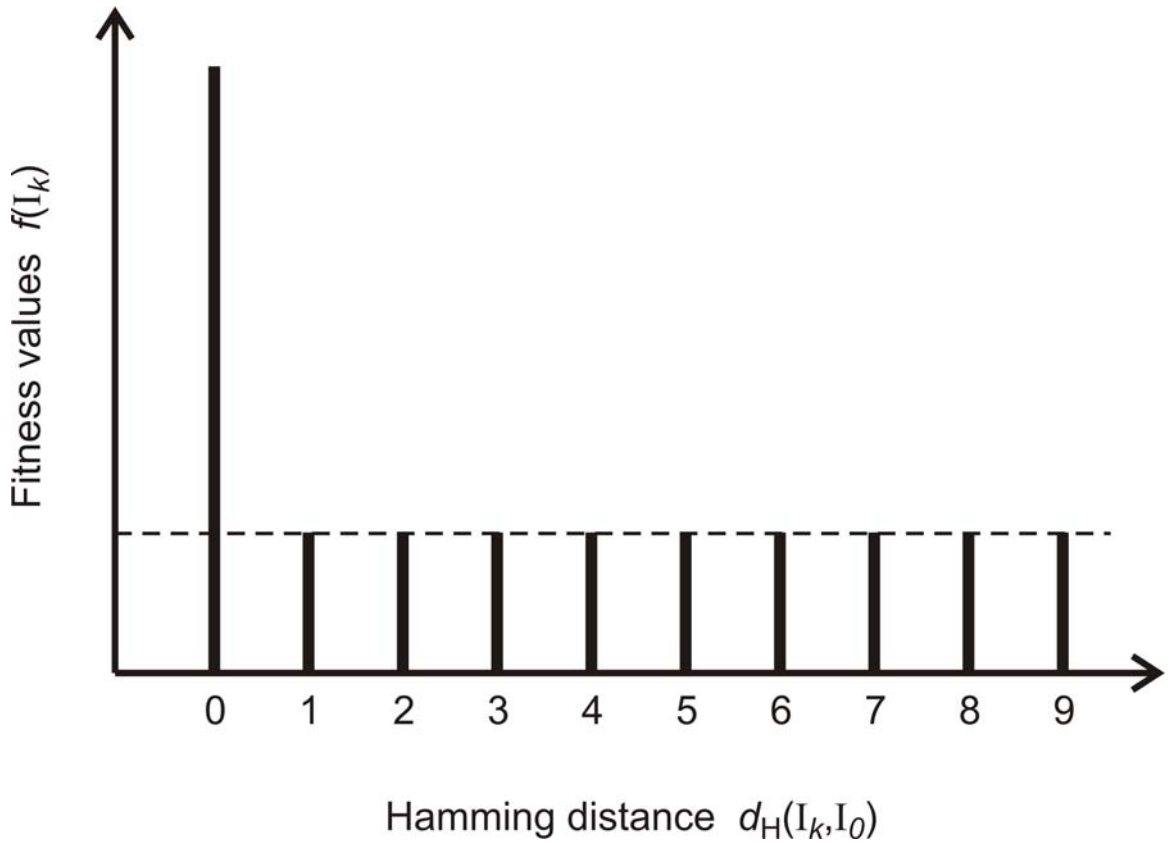
Esteban Domingo
Universidad Autónoma de Madrid
Centro de Biología Molecular "Severo Ochoa"
Consejo Superior de Investigaciones Científicas
Cantoblanco and Valdeolmos
Madrid, Spain
Tel.: +34 91 497 84858/9; fax: +34 91 497 4799
E-mail address: edomingo@cbm.uam.es

Available online 8 December 2004

doi:10.1016/j.virusres.2004.11.001

Molecular evolution of viruses





A fitness landscape showing an error threshold

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

329

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.

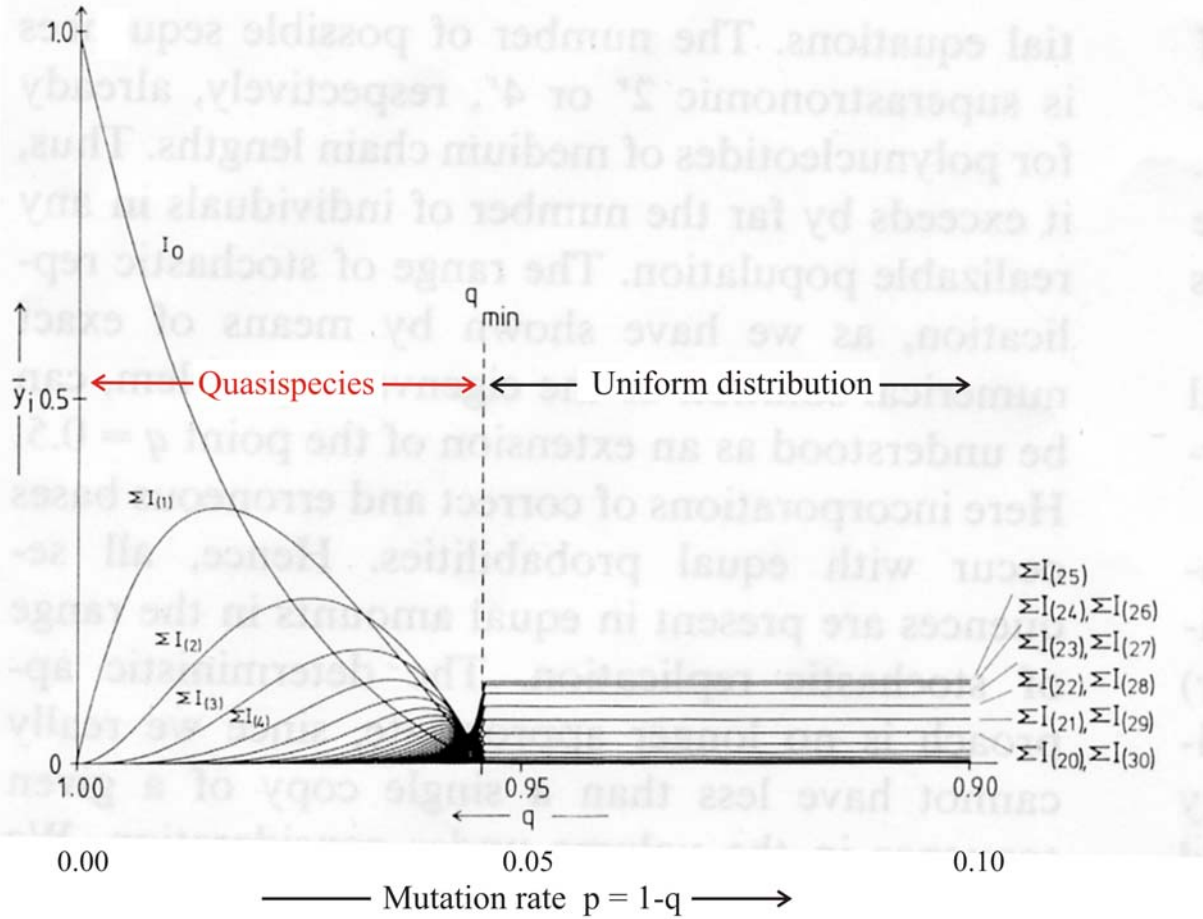
Rigorous mathematical analysis has been performed on eq. 1 [7,15,24,26]. In particular, it was shown that the non-linearity of eq. 1 can be removed by an appropriate transformation. The eigenvalue problem of the linear differential equation obtained thereby may be solved approximately by the conventional 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: $\Sigma_i = \Sigma_{i=1}^{n}$ and $\Sigma_{i,i\neq j} = \Sigma_{i=1}^{j-1} + \Sigma_{i=j+1}^{n}$, respectively.

0301-4622/82/0000-0000/$02.75



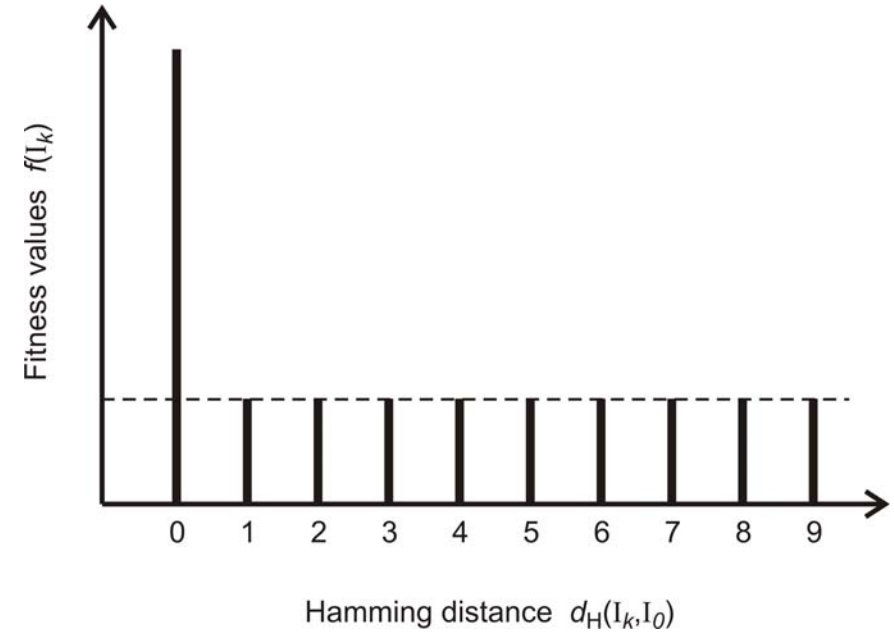
Single peak fitness landscape: $f_0 = f$ and $f_1 = f_2 = \ldots f_N = 1$

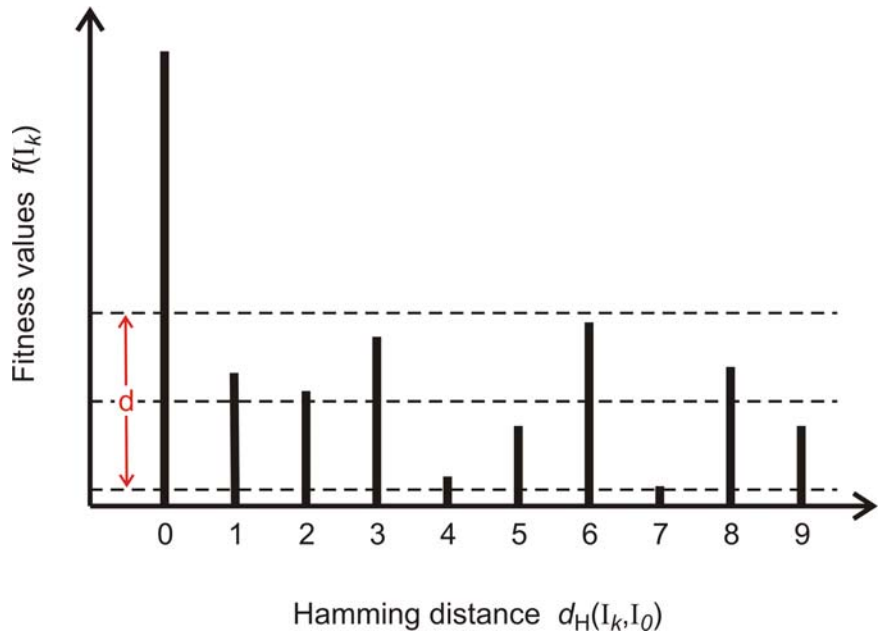
Quasispecies as a function of the mutation rate p

$$\sigma = \frac{f_0}{(1 - x_0)\sum_{i=1}^{N} f_i \, x_i}$$

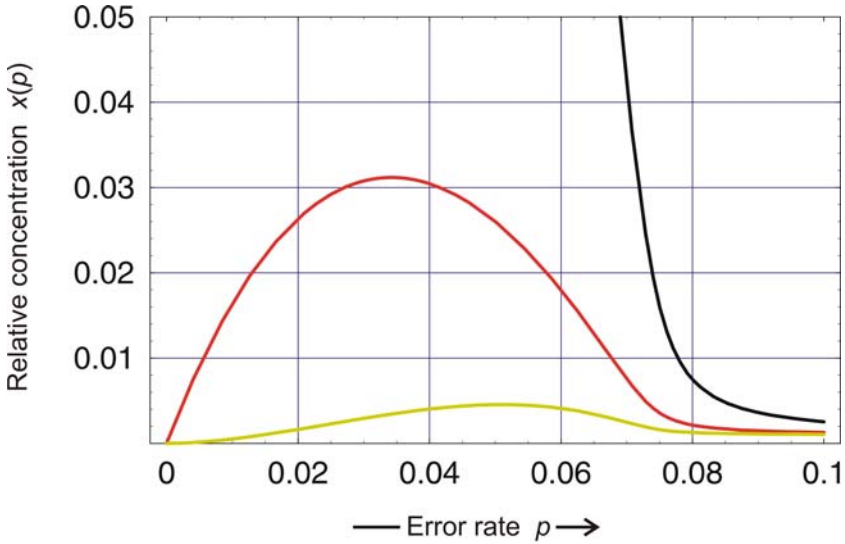
$f_0 = \sigma = 10$

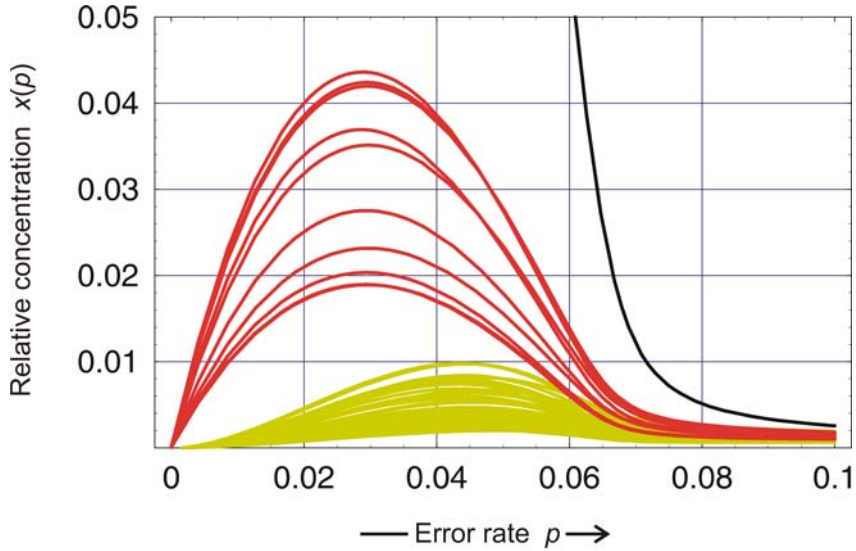
$I_0 \ldots$ master sequence; $N = \kappa^n$

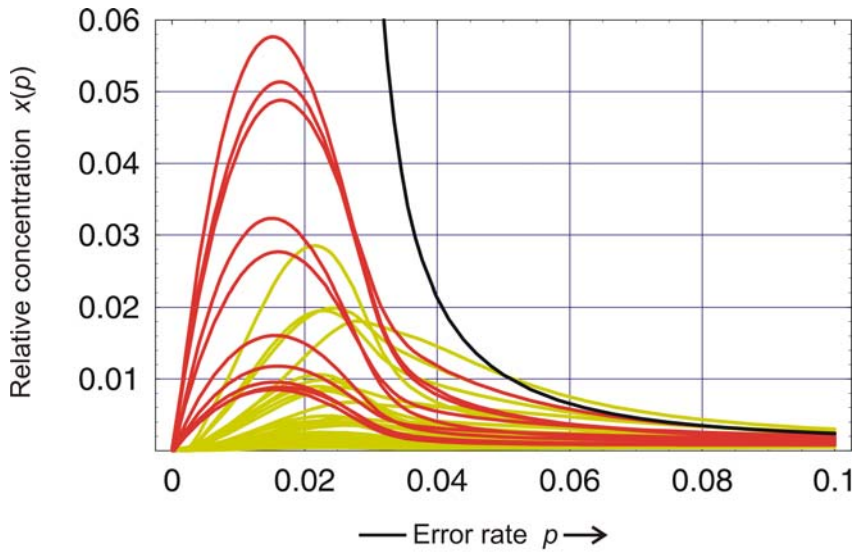




Fitness landscapes showing error thresholds







Error threshold: Individual sequences

n = 10, σ = 2 and d = 0, 1.0, 1.85

Evolutionary design of RNA molecules

A.D. Ellington, J.W. Szostak, **In vitro *selection of RNA molecules that bind specific ligands***. Nature **346** (1990), 818-822

C. Tuerk, L. Gold, **SELEX - *Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage* T4 *DNA polymerase***. Science **249** (1990), 505-510

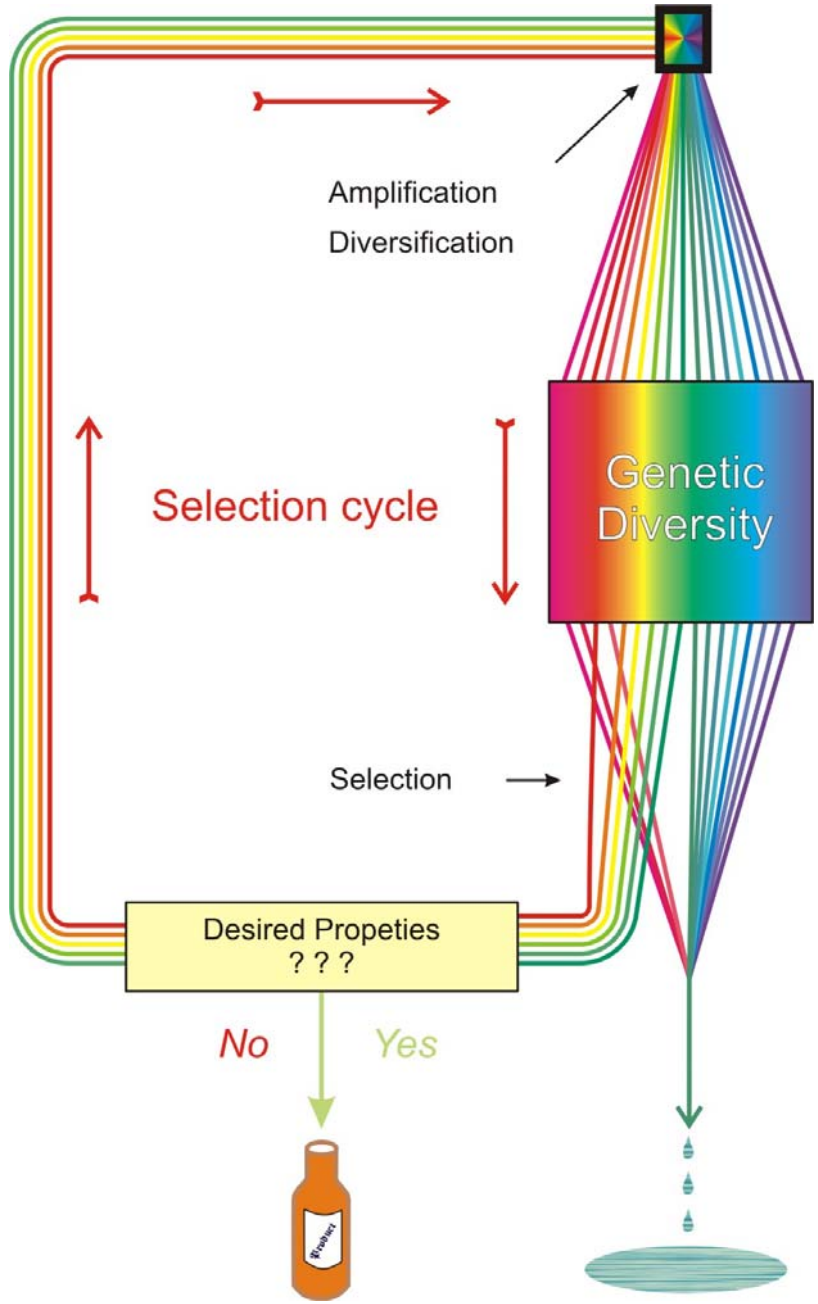
D.P. Bartel, J.W. Szostak, ***Isolation of new ribozymes from a large pool of random sequences***. Science **261** (1993), 1411-1418

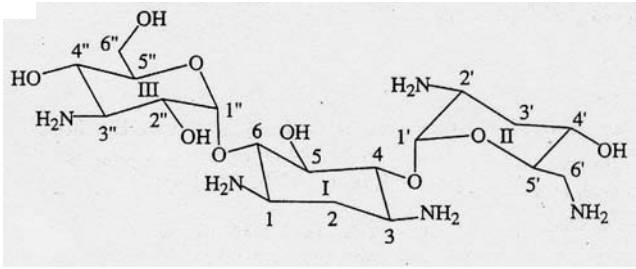
R.D. Jenison, S.C. Gill, A. Pardi, B. Poliski, ***High-resolution molecular discrimination by RNA***. Science **263** (1994), 1425-1429

Y. Wang, R.R. Rando, ***Specific binding of aminoglycoside antibiotics to RNA***. Chemistry & Biology **2** (1995), 281-290

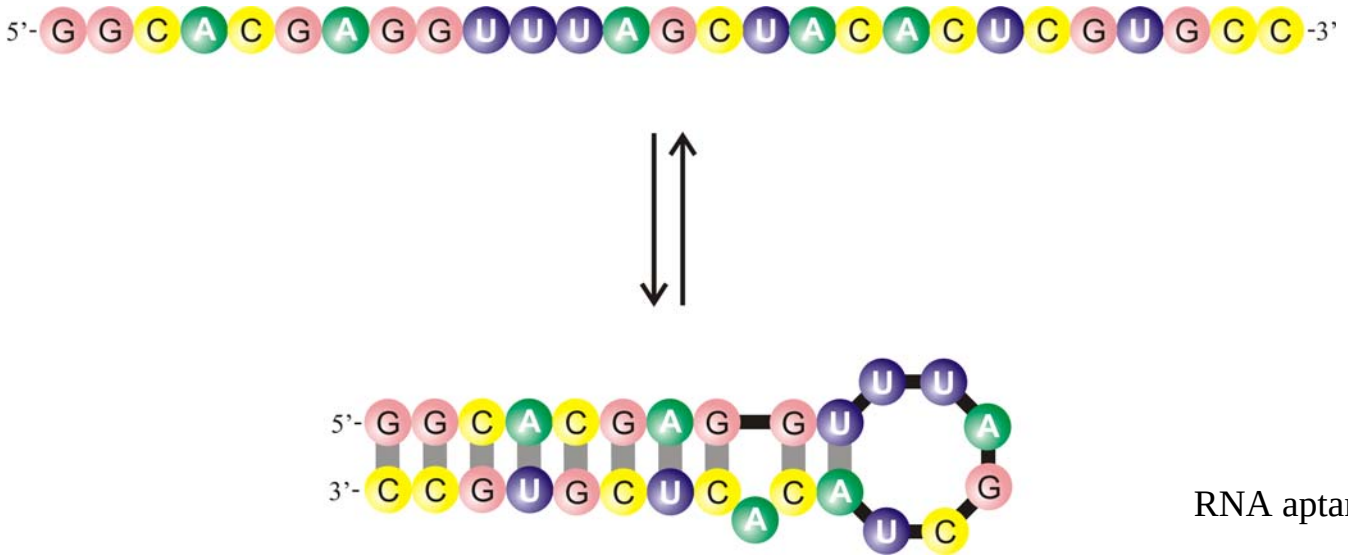
L. Jiang, A. K. Suri, R. Fiala, D. J. Patel, ***Saccharide-RNA recognition in an aminoglycoside antibiotic-RNA aptamer complex***. Chemistry & Biology **4** (1997), 35-50

An example of ‘artificial selection’ with RNA molecules or ‘breeding’ of biomolecules





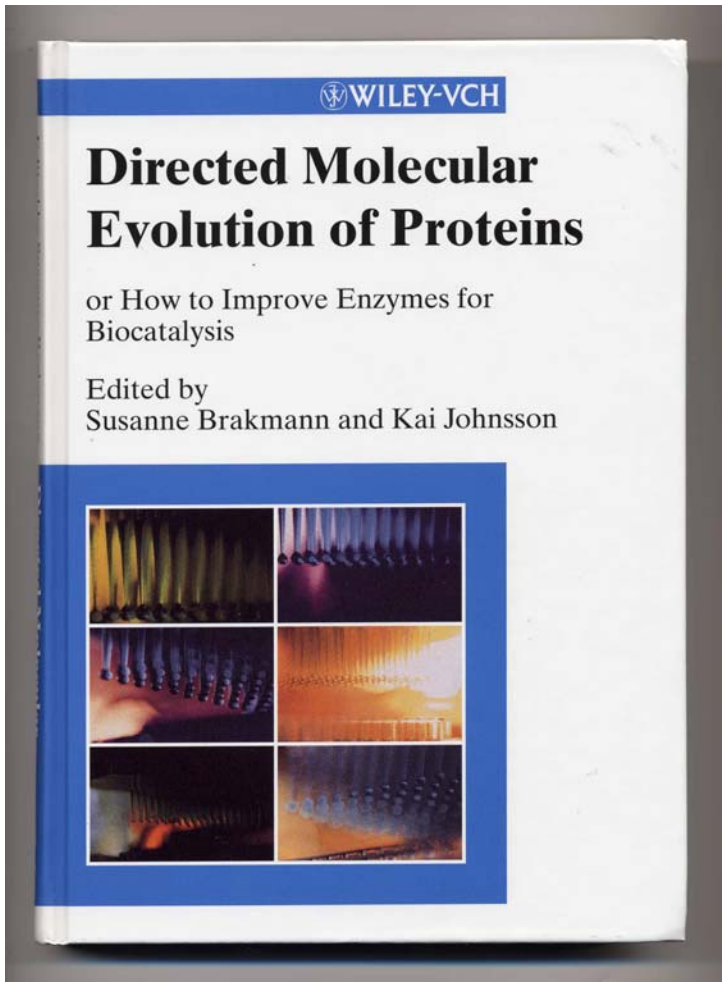
tobramycin

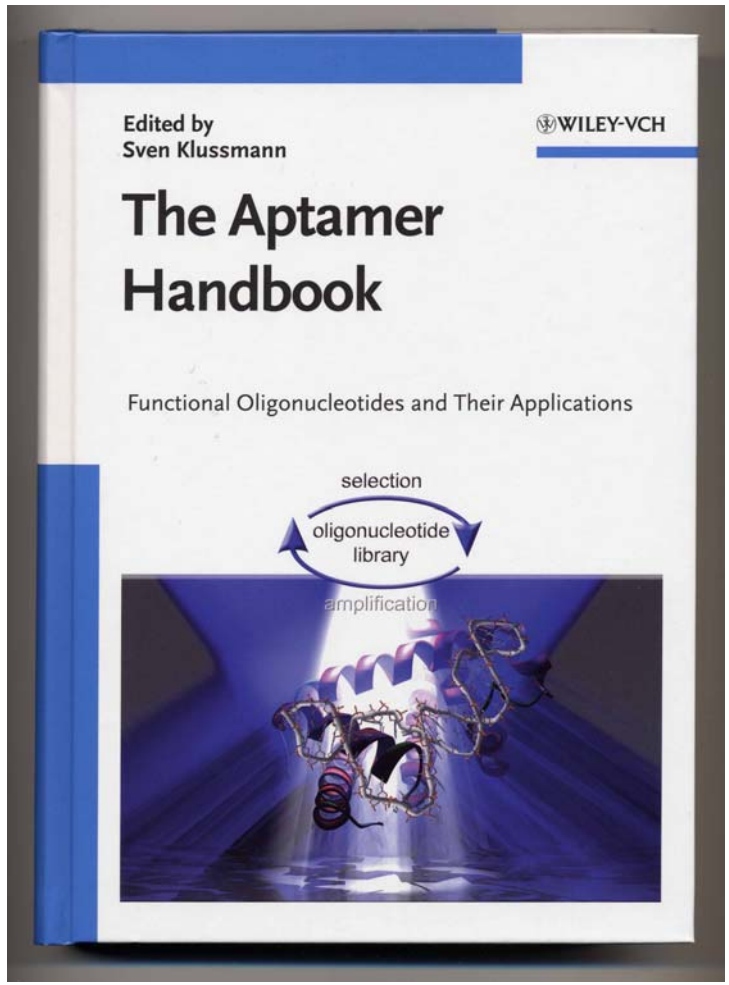


RNA aptamer, *n* = 27

Formation of secondary structure of the tobramycin binding RNA aptamer with **K_D = 9 nM**

L. Jiang, A. K. Suri, R. Fiala, D. J. Patel, ***Saccharide-RNA recognition in an aminoglycoside antibiotic-RNA aptamer complex.*** Chemistry & Biology **4**:35-50 (1997)





Application of molecular evolution to problems in biotechnology

Artificial evolution in biotechnology and pharmacology

G.F. Joyce. 2004. Directed evolution of nucleic acid enzymes. *Annu.Rev.Biochem*. **73**:791-836.

C. Jäckel, P. Kast, and D. Hilvert. 2008. Protein design by directed evolution. *Annu.Rev.Biophys*. **37**:153-173.

S.J. Wrenn and P.B. Harbury. 2007. Chemical evolution as a tool for molecular discovery. *Annu.Rev.Biochem*. **76**:331-349.

Results from kinetic theory of molecular evolution and evolution experiments:

• Evolutionary optimization does not require cells and occurs as well in cell-free molecular systems.

• Replicating ensembles of molecules form stationary populations called **quasispecies**, which represent the genetic reservoir of asexually reproducing species.

• For stable inheritance of genetic information mutation rates must not exceed a precisely defined and computable **error-threshold**.

•The error-threshold can be exploited for the development of novel antiviral strategies.

• *In vitro* evolution allows for production of molecules for predefined purposes and gave rise to a branch of biotechnology.

1. Requirements for information processing
2. The chemistry of Darwinian evolution
3. **RNA sequences and structures**
4. Consequences of neutrality
5. Evolutionary optimization of RNA structure

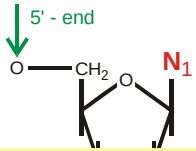
RNA folding **determination of RNA function**

molecular recognition catalysis

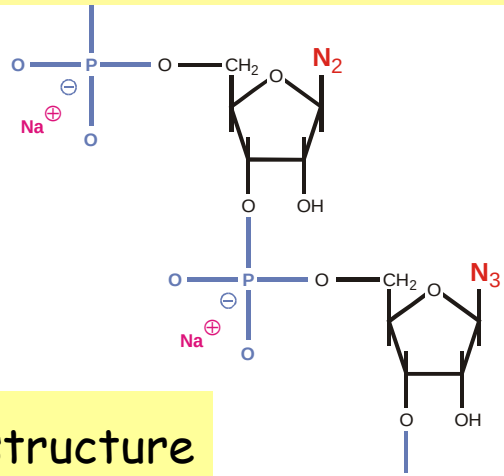
binding to: ground state transition state

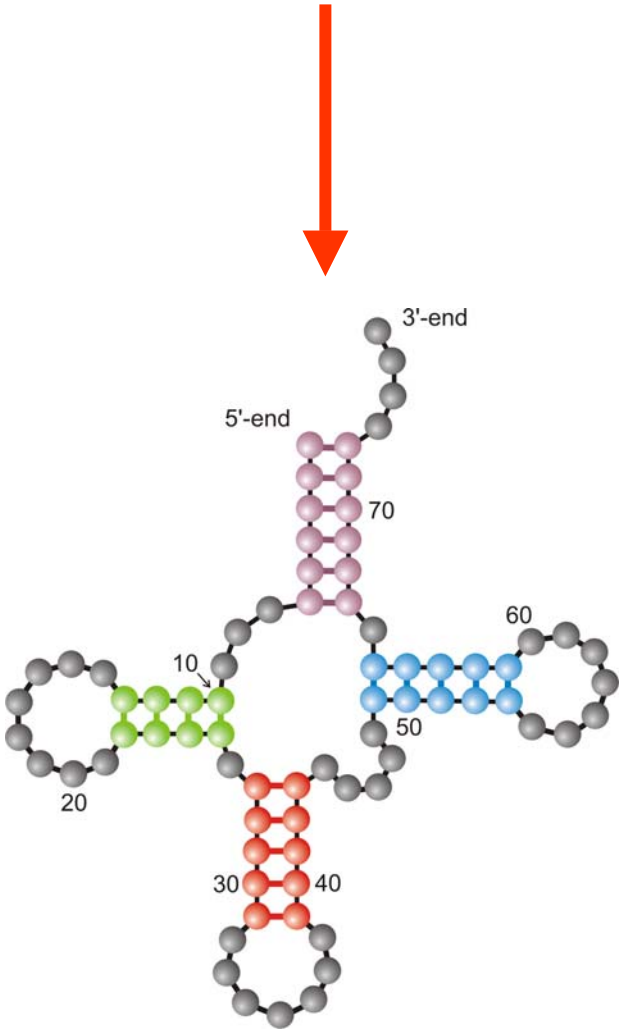
aptamers **ribozymes**

The paradigm of structural biology

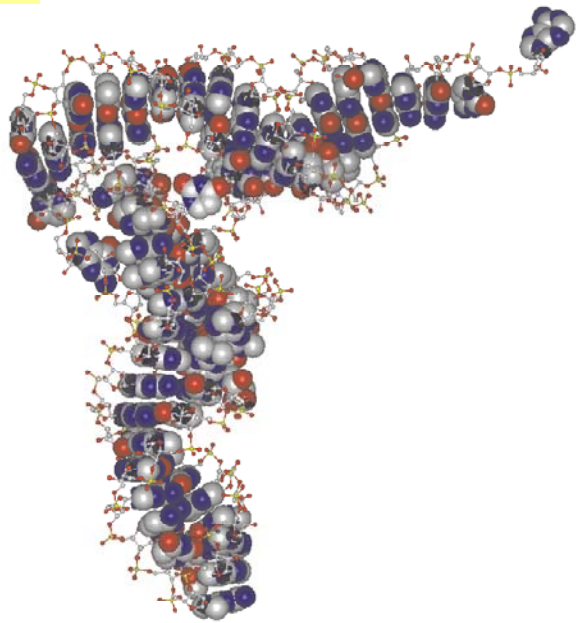


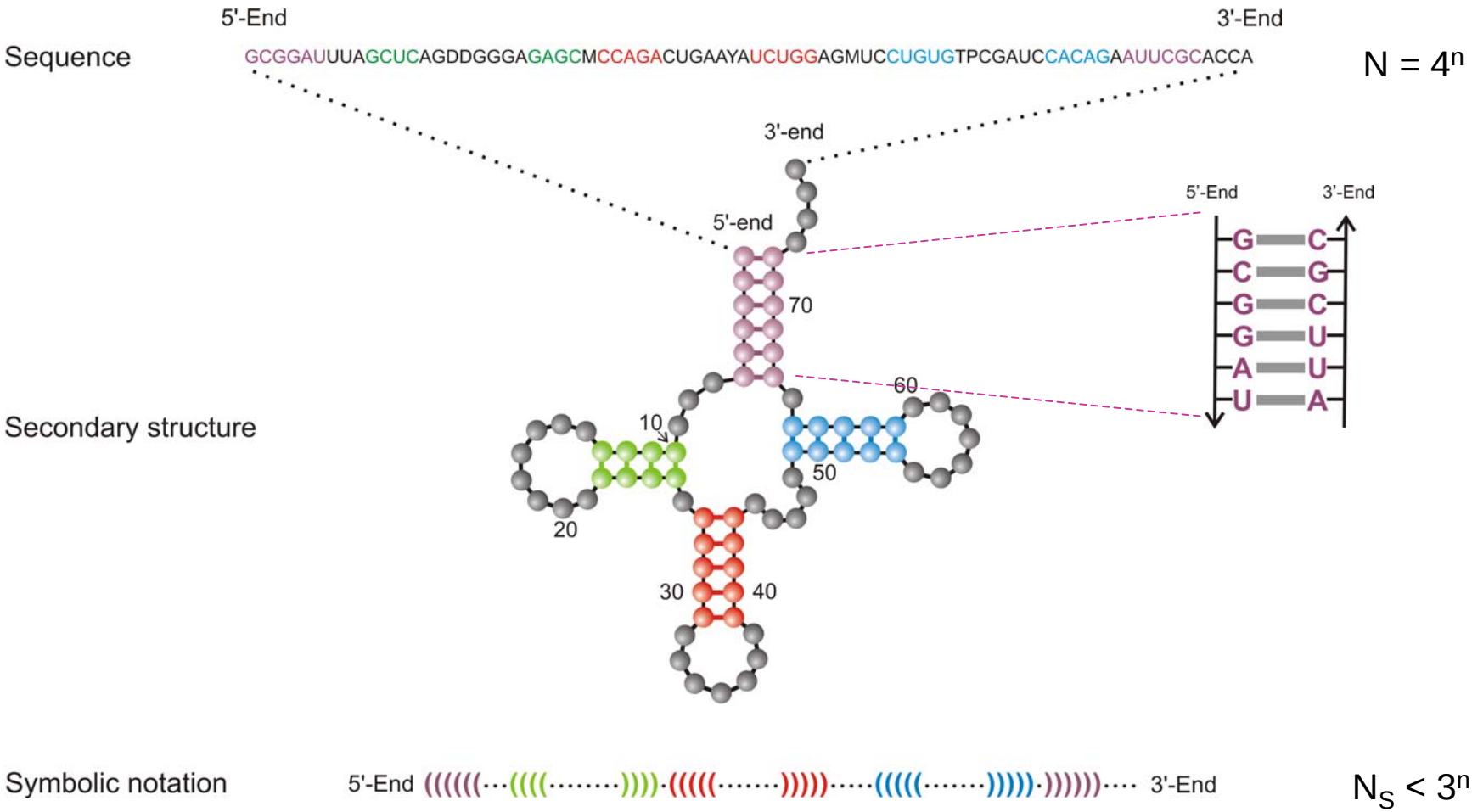
5'-end GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCACCA 3'-end





Definition of RNA structure





Criterion: Minimum free energy (mfe)

Rules: _ (_) _ ∈ {**AU**,**CG**,**GC**,**GU**,**UA**,**UG**}

A symbolic notation of RNA secondary structure that is equivalent to the conventional graphs

What is neutrality ?

Selective neutrality =
= several genotypes having the same fitness.

Structural neutrality =
= several genotypes forming molecules with the same structure.

From sequences to shapes and back: a case study in RNA secondary structures

PETER SCHUSTER[1,2,3], WALTER FONTANA[3], PETER F. STADLER[2,3] AND IVO L. HOFACKER[2]

[1] *Institut für Molekulare Biotechnologie, Beutenbergstrasse 11, PF 100813, D-07708 Jena, Germany*
[2] *Institut für Theoretische Chemie, Universität Wien, Austria*
[3] *Santa Fe Institute, Santa Fe, U.S.A.*

SUMMARY

RNA folding is viewed here as a map assigning secondary structures to sequences. At fixed chain length the number of sequences far exceeds the number of structures. Frequencies of structures are highly non-uniform and follow a generalized form of Zipf's law: we find relatively few common and many rare ones. By using an algorithm for inverse folding, we show that sequences sharing the same structure are distributed randomly over sequence space. All common structures can be accessed from an arbitrary sequence by a number of mutations much smaller than the chain length. The sequence space is percolated by extensive neutral networks connecting nearest neighbours folding into identical structures. Implications for evolutionary adaptation and for applied molecular evolution are evident: finding a particular structure by mutation and selection is much simpler than expected and, even if catalytic activity should turn out to be sparse in the space of RNA structures, it can hardly be missed by evolutionary processes.

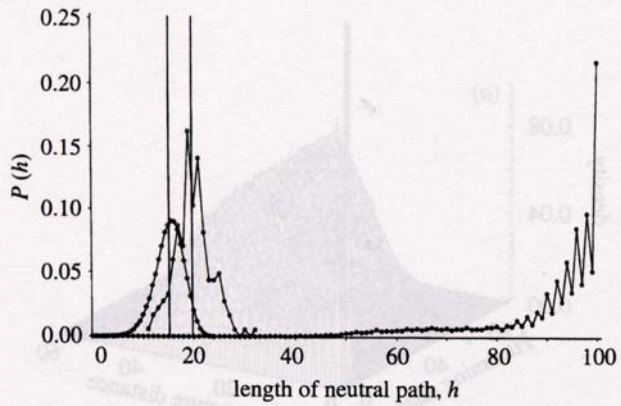
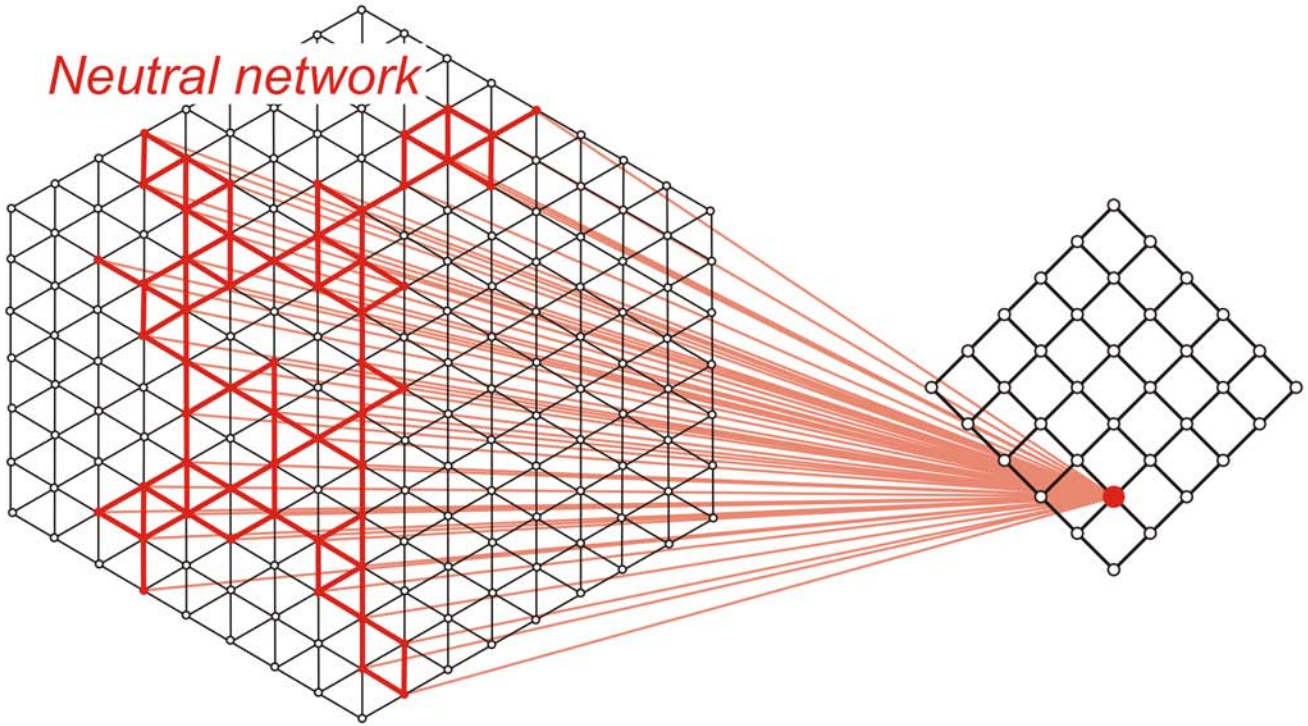


Figure 4. Neutral paths. A neutral path is defined by a series of nearest neighbour sequences that fold into identical structures. Two classes of nearest neighbours are admitted: neighbours of Hamming distance 1, which are obtained by single base exchanges in unpaired stretches of the structure, and neighbours of Hamming distance 2, resulting from base pair exchanges in stacks. Two probability densities of Hamming distances are shown that were obtained by searching for neutral paths in sequence space: (i) an upper bound for the closest approach of trial and target sequences (open circles) obtained as endpoints of neutral paths approaching the target from a random trial sequence (185 targets and 100 trials for each were used); (ii) a lower bound for the closest approach of trial and target sequences (open diamonds) derived from secondary structure statistics (Fontana *et al.* 1993*a*; see this paper, §4); and (iii) longest distances between the reference and the endpoints of monotonously diverging neutral paths (filled circles) (500 reference sequences were used).

Proc. R. Soc. Lond. B (1994) **255**, 279–284
Printed in Great Britain

Reference for postulation and ***in silico*** verification of ***neutral networks***

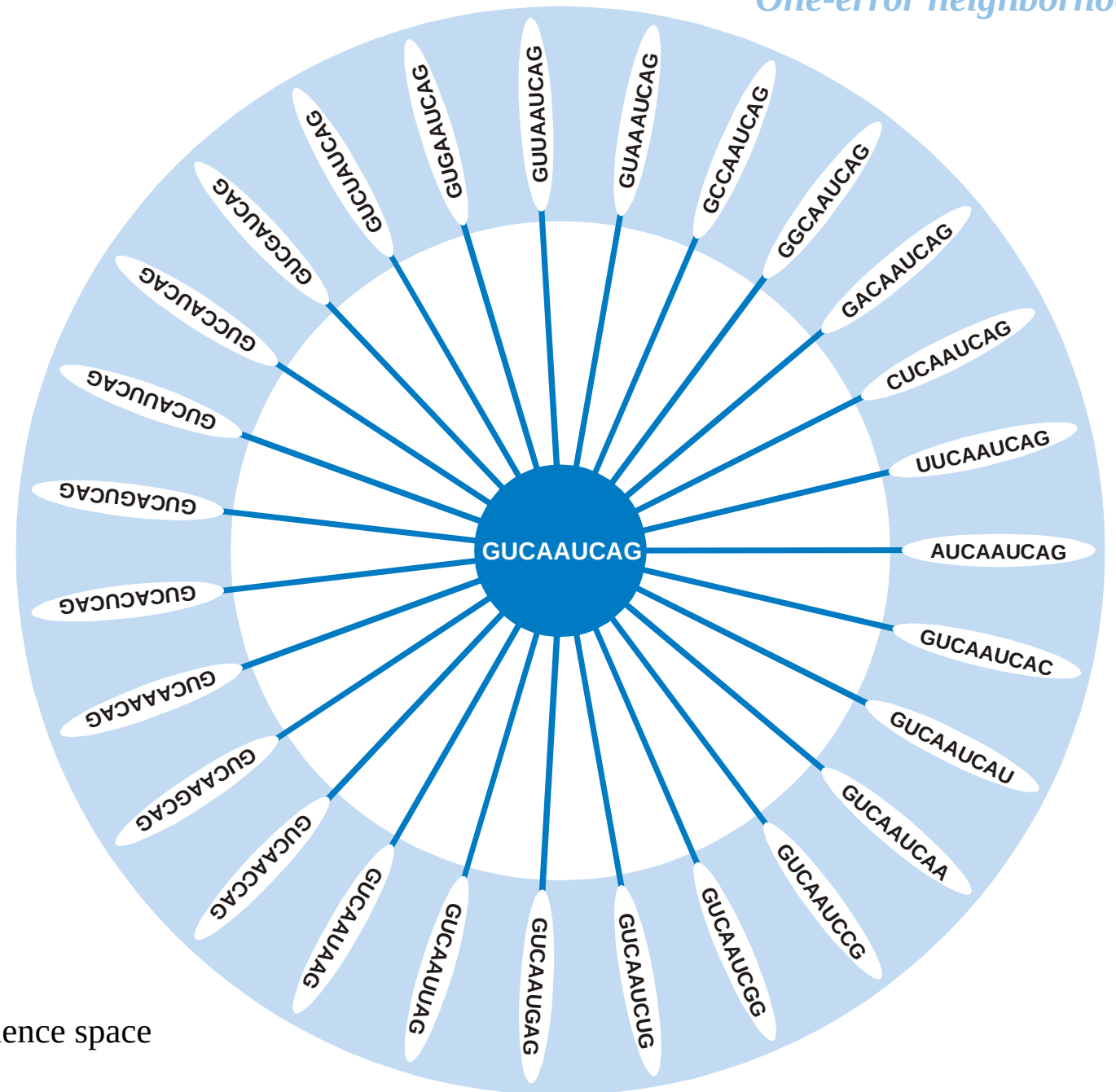


Sequence space

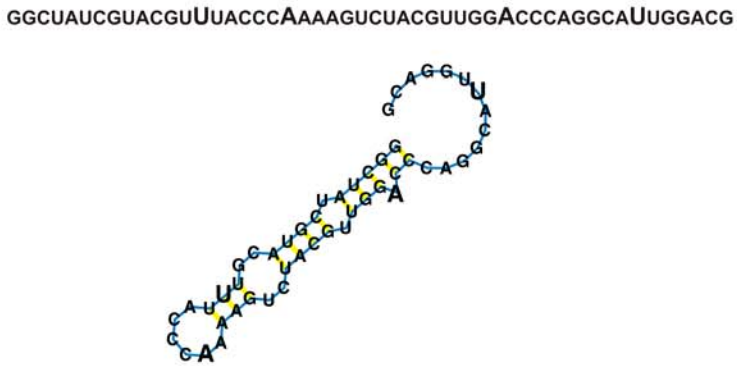
Structure space

many genotypes ⇒ one phenotype

One-error neighborhood



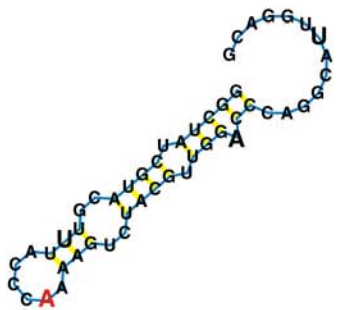
The surrounding of
GUCAAUCAG in sequence space



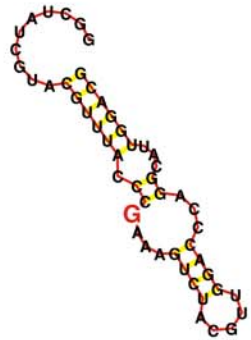
One error neighborhood – Surrounding of an RNA molecule of chain length n=50 in sequence and shape space

GGCUAUCGUACGUUUACCCGAAAGUCUACGUUGGACCCAGGCAUUGGACG

GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG

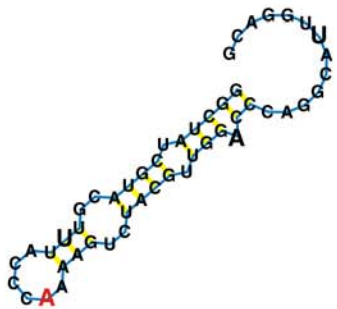


One error neighborhood – Surrounding of an RNA molecule of chain length n=50 in sequence and shape space

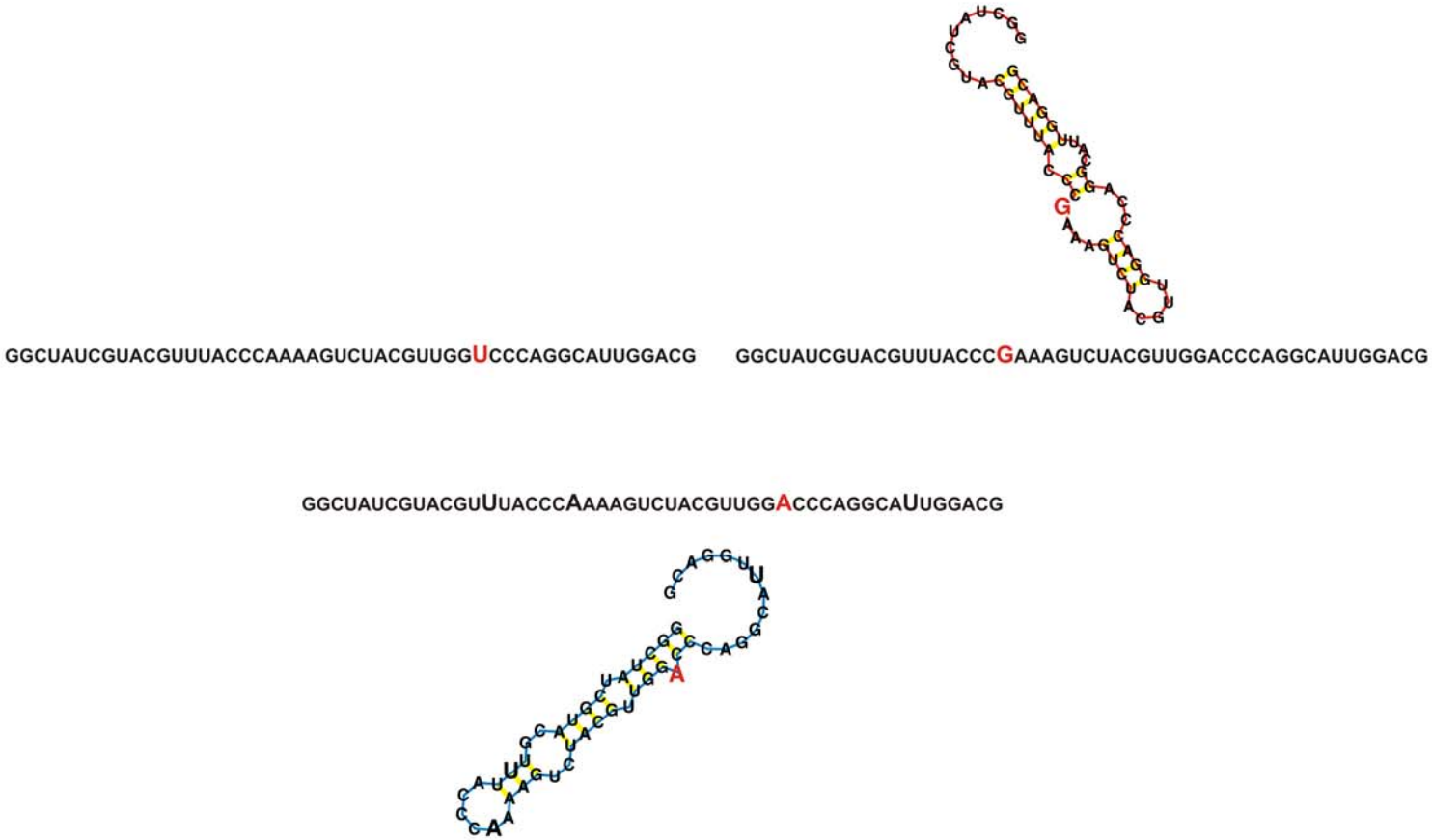


GGCUAUCGUACGUUUACCCGAAAGUCUACGUUGGACCCAGGCAUUGGACG

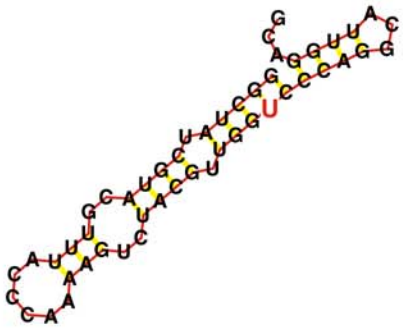
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG



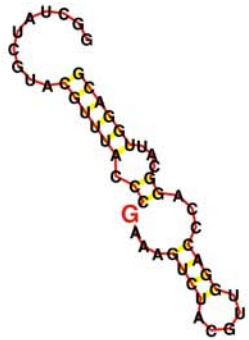
One error neighborhood – Surrounding of an RNA molecule of chain length n=50 in sequence and shape space



One error neighborhood – Surrounding of an RNA molecule of chain length n=50 in sequence and shape space

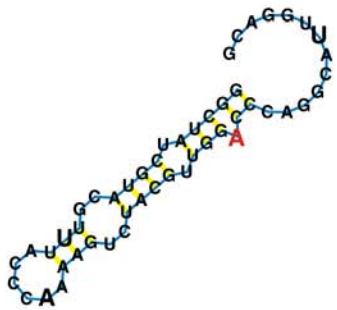


GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGUCCCAGGCAUUGGACG

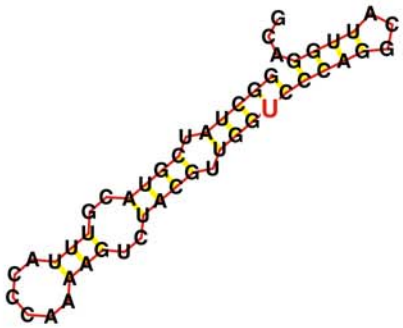


GGCUAUCGUACGUUUACCCGAAAGUCUACGUUGGACCCAGGCAUUGGACG

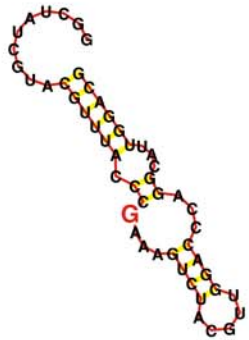
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG



One error neighborhood – Surrounding of an RNA molecule of chain length n=50 in sequence and shape space



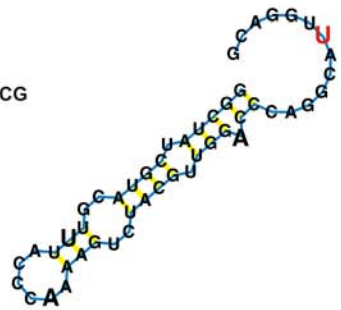
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGUCCCAGGCAUUGGACG



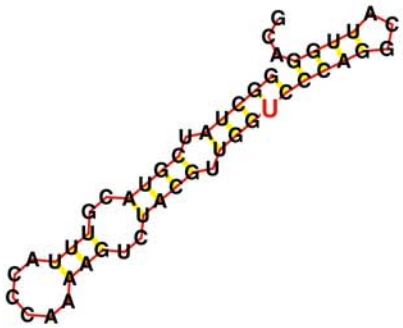
GGCUAUCGUACGUUUACCCGAAAGUCUACGUUGGACCCAGGCAUUGGACG

GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG

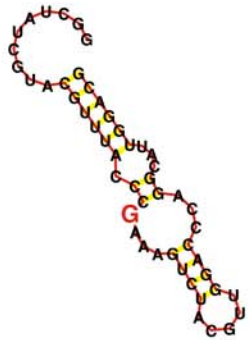
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCACUGGACG



One error neighborhood – Surrounding of an RNA molecule of chain length n=50 in sequence and shape space



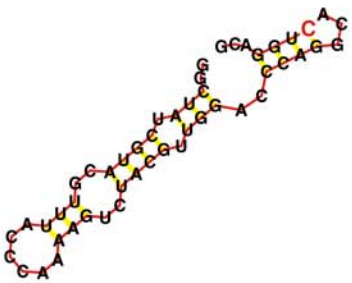
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGUCCCAGGCAUUGGACG

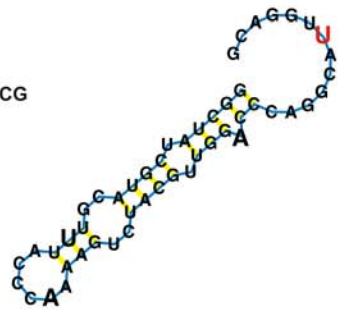


GGCUAUCGUACGUUUACCCGAAAGUCUACGUUGGACCCAGGCAUUGGACG

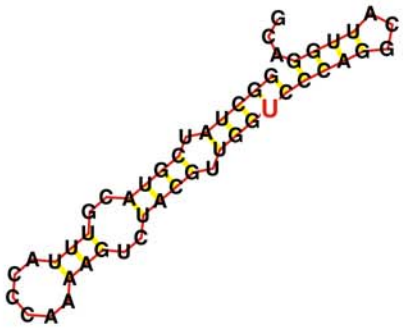
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG

GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCACUGGACG

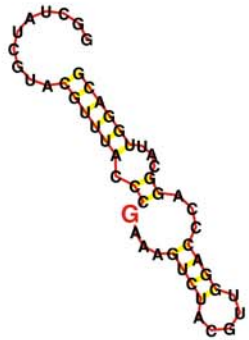




One error neighborhood – Surrounding of an RNA molecule of chain length n=50 in sequence and shape space



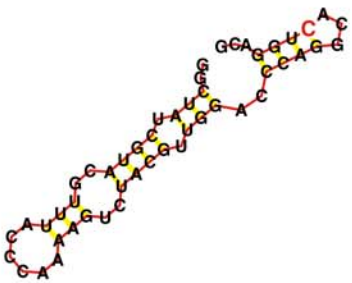
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGUCCCAGGCAUUGGACG

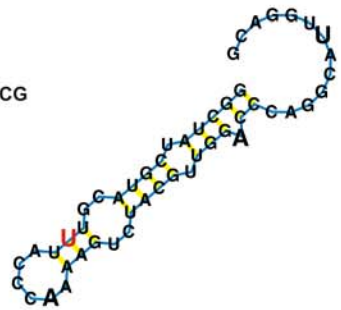


GGCUAUCGUACGUUUACCCGAAAGUCUACGUUGGACCCAGGCAUUGGACG

GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG

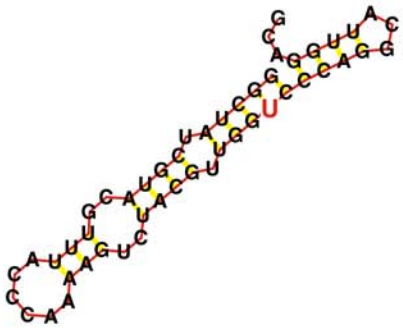
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCACUGGACG



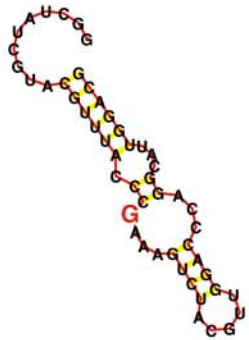


GGCUAUCGUACGUGUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG

One error neighborhood – Surrounding of an RNA molecule of chain length n=50 in sequence and shape space

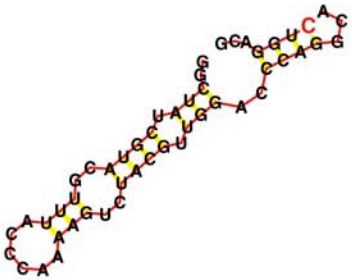


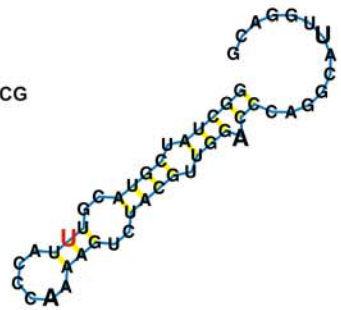
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGUCCCAGGCAUUGGACG



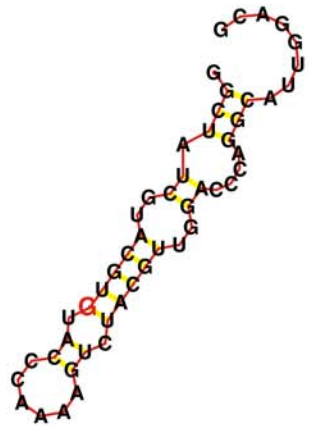
GGCUAUCGUACGUUUACCCGAAAGUCUACGUUGGACCCAGGCAUUGGACG

GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCACUGGACG





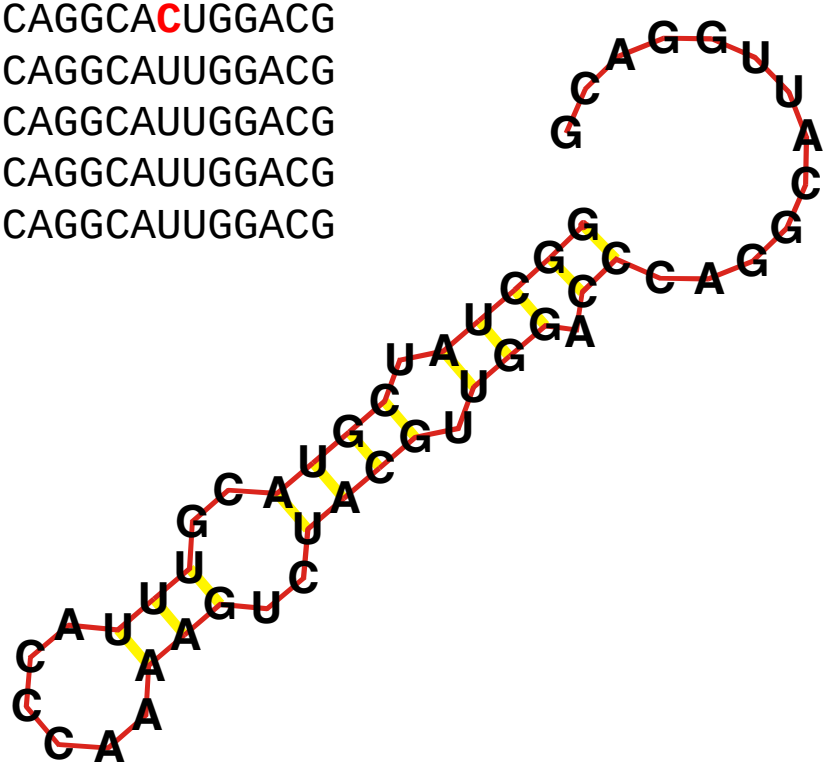
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG



GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG

One error neighborhood – Surrounding of an RNA molecule of chain length n=50 in sequence and shape space

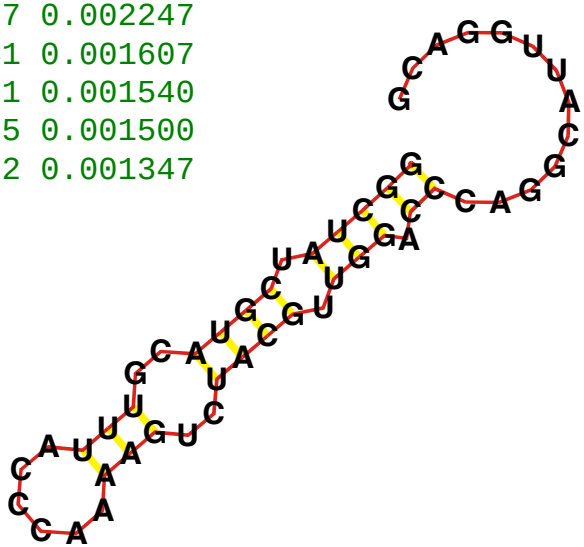
GGCUAUCGUA**U**GUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUU**A**GACG
GGCUAUCGUACGUUUAC**U**CAAAAGUCUACGUUGGACCCAGGCAUUGGACG
GGCUAUCGUACG**C**UUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG
GGC**C**AUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG
GGCUAUCGUACGU**G**UACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG
GGCUA**A**CGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCC**U**GGCAUUGGACG
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCA**C**UGGACG
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGG**U**CCCAGGCAUUGGACG
GGCUA**G**CGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG
GGCUAUCGUACGUUUACCC**G**AAAGUCUACGUUGGACCCAGGCAUUGGACG
GGCUAUCGUACGUUUACCCAAAAG**C**CUACGUUGGACCCAGGCAUUGGACG



One error neighborhood – Surrounding of an RNA molecule of chain length n=50 in sequence and shape space

	Number	Mean Value	Variance	Std.Dev.
Total Hamming Distance:	150000	11.647973	23.140715	4.810480
Nonzero Hamming Distance:	99875	16.949991	30.757651	5.545958
Degree of Neutrality:	50125	**0.334167**	0.006961	**0.083434**
Number of Structures:	**1000**	**52.31**	85.30	**9.24**

1	(((((.((((..(((.....))))..)))))..))).).....	**50125**	**0.334167**
2	..(((.((((..(((.....))))..)))))..))).....	**2856**	**0.019040**
3	((((((((((..(((.....))))..)))))))).)).....	**2799**	**0.018660**
4	(((((.((((..((((....))))..)))))..))).)).....	**2417**	**0.016113**
5	(((((.((((.((((.....))))).)))))..))).)).....	**2265**	**0.015100**
6	(((((.(((((.(((.....))).)))))..))).)).....	**2233**	**0.014887**
7	(((((..(((..(((.....)))..)))..)))..)).....	1442	0.009613
8	(((((.((((..((.......))..)))))..))).)).....	1081	0.007207
9	((((..((((..(((.....)))..)))))..)).)).....	1025	0.006833
10	(((((.((((..(((.....)))..)))))..))))).....	1003	0.006687
11	.((((.((((..(((.....)))..)))))..)))).....	963	0.006420
12	(((((.(((...(((.....)))...)))..))).)).....	860	0.005733
13	(((((.((((..(((.....)))..)))))..)).))).....	800	0.005333
14	(((((.((((...((.....))...)))))..))).)).....	548	0.003653
15	(((((.((((..............))))..))).)).....	362	0.002413
16	((.((.((((..(((.....)))..)))))..))..)).....	337	0.002247
17	(.(((.((((..(((.....)))..)))))..))).).....	241	0.001607
18	(((((.(((((((((.....)))))))))..))).)).....	231	0.001540
19	((((..((((..(((.....)))..)))))...)))).....	225	0.001500
20	((....((((..(((.....)))..))))).....)).....	202	0.001347



Shadow – Surrounding of an RNA structure in shape space:
AUGC alphabet, chain length n=50

1. Requirements for information processing
2. The chemistry of Darwinian evolution
3. RNA sequences and structures
4. **Consequences of neutrality**
5. Evolutionary optimization of RNA structure

ON

THE ORIGIN OF SPECIES

BY MEANS OF NATURAL SELECTION,

OR THE

PRESERVATION OF FAVOURED RACES IN THE STRUGGLE FOR LIFE.

By CHARLES DARWIN, M.A.,

FELLOW OF THE ROYAL, GEOLOGICAL, LINNÆAN, ETC., SOCIETIES;
AUTHOR OF 'JOURNAL OF RESEARCHES DURING H. M. S. BEAGLE'S VOYAGE ROUND THE WORLD.'

LONDON:
JOHN MURRAY, ALBEMARLE STREET.
1859.

This preservation of favourable individual differences and variations, and the destruction of those which are injurious, I have called Natural Selection, or the Survival of the Fittest. Variations neither useful nor injurious would not be affected by natural selection, and would be left either a fluctuating element, as perhaps we see in certain polymorphic species, or would ultimately become fixed, owing to the nature of the organism and the nature of the conditions.

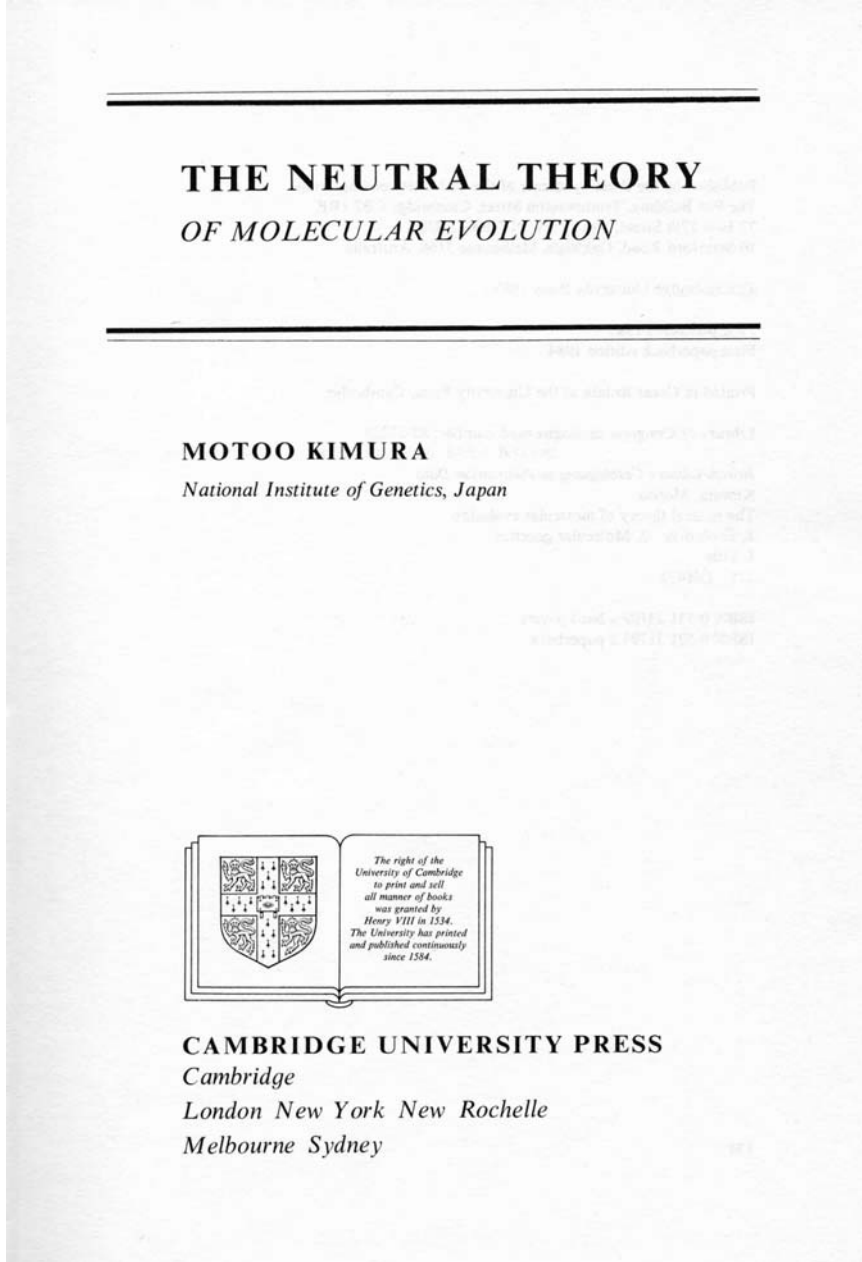
Charles Darwin. *The Origin of Species*. Sixth edition. John Murray. London: 1872

MOTOO KIMURA

Motoo Kimuras population genetics of neutral evolution.

Evolutionary rate at the molecular level. *Nature* **217**: 624-626, 1955.

The Neutral Theory of Molecular Evolution. Cambridge University Press. Cambridge, UK, 1983.



THE NEUTRAL THEORY

OF MOLECULAR EVOLUTION

MOTOO KIMURA

National Institute of Genetics, Japan

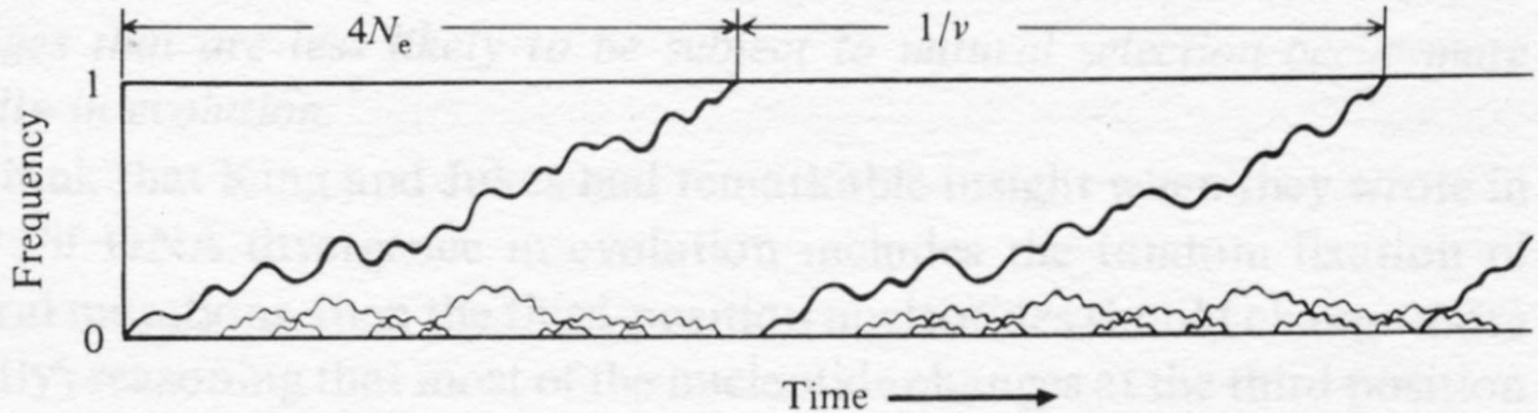
CAMBRIDGE UNIVERSITY PRESS

Cambridge

London New York New Rochelle

Melbourne Sydney

Fig. 3.1. Behavior of mutant genes following their appearance in a finite population. Courses of change in the frequencies of mutants destined to fixation are depicted by thick paths. N_e stands for the effective population size and v is the mutation rate.



The average time of replacement of a dominant genotype in a population is the reciprocal mutation rate, 1/ν, and therefore independent of population size.

Is the Kimura scenario correct for frequent mutations?

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

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

STATIONARY MUTANT DISTRIBUTIONS AND EVOLUTIONARY OPTIMIZATION

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

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

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

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

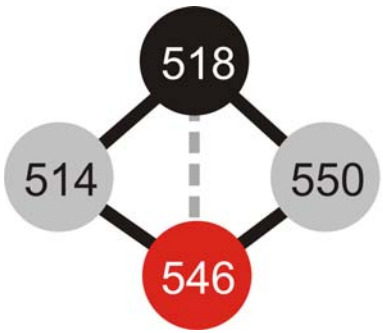
A kinetic theory of polynucleotide evolution which was developed during the past 15 years (Eigen, 1971; 1985; Eigen and Schuster, 1979; Eigen *et al.*, 1987; Schuster, 1986); Schuster and Sigmund, 1985) treats correct replication and mutation as parallel reactions within one and the same reaction network

Neutral network

$\lambda = 0.01,\ s = 367$

$\mathbf{d_H = 1}$

$$\lim_{p\to 0} x_1(p) = x_2(p) = 0.5$$



Neutral network

$\lambda = 0.01,\ s = 877$

$\mathbf{d_H = 2}$

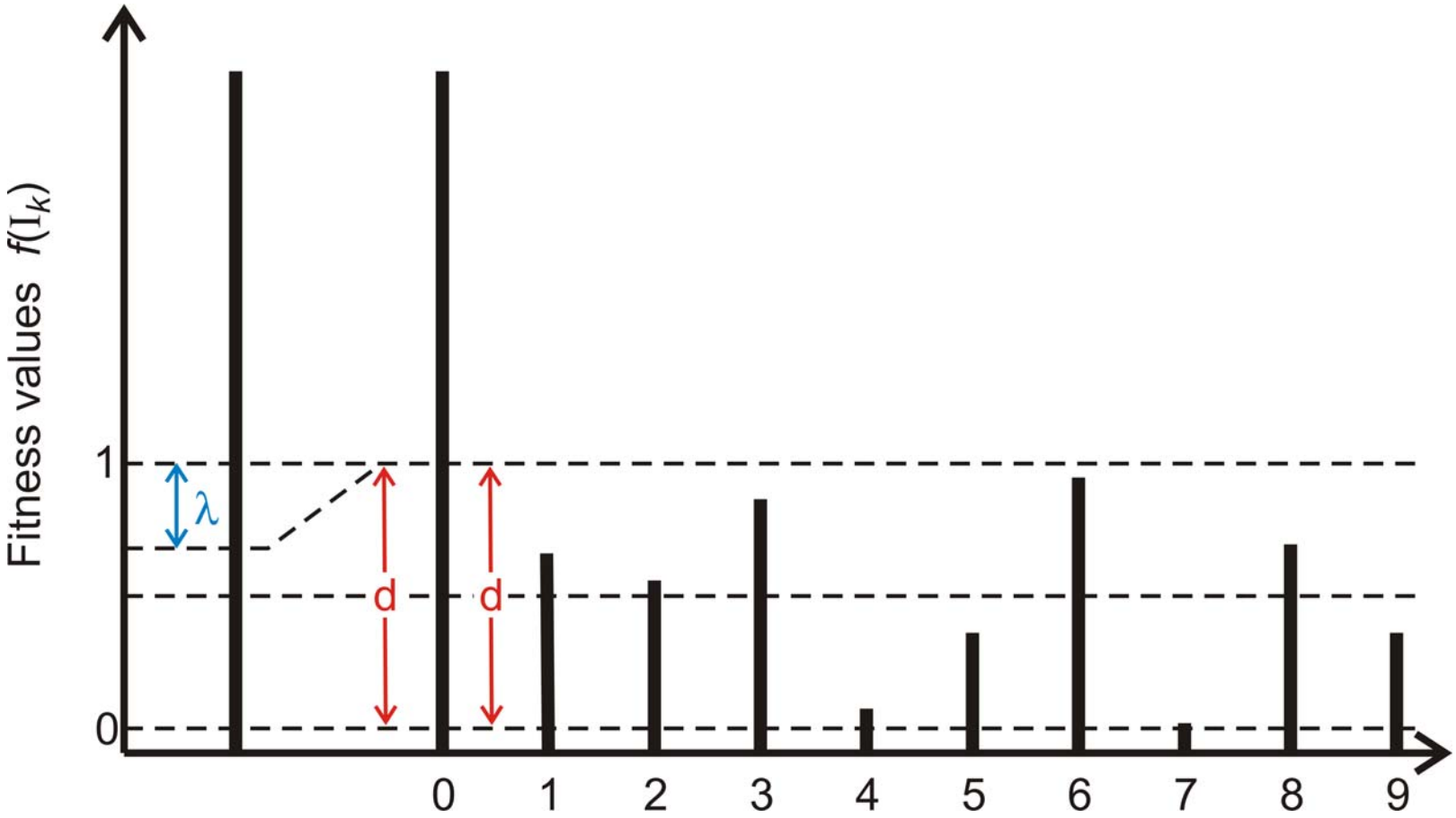
$$\lim_{p\to 0} x_1(p) = a$$

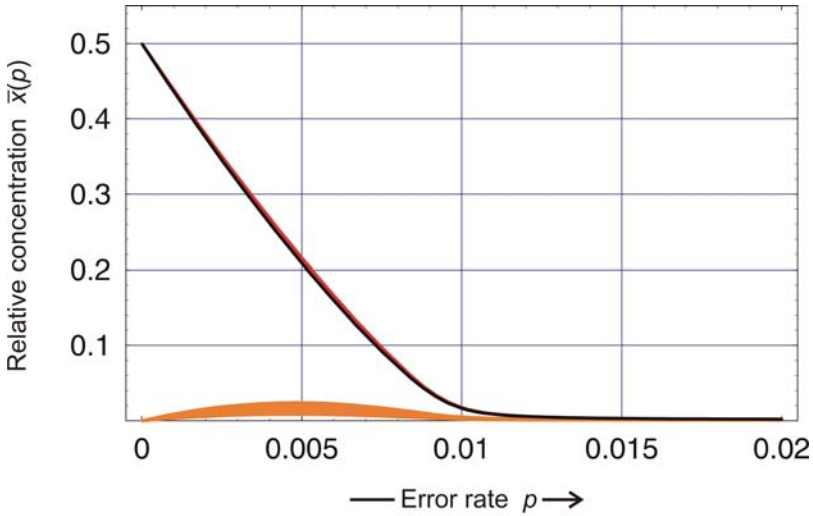
$$\lim_{p\to 0} x_2(p) = 1 - a$$

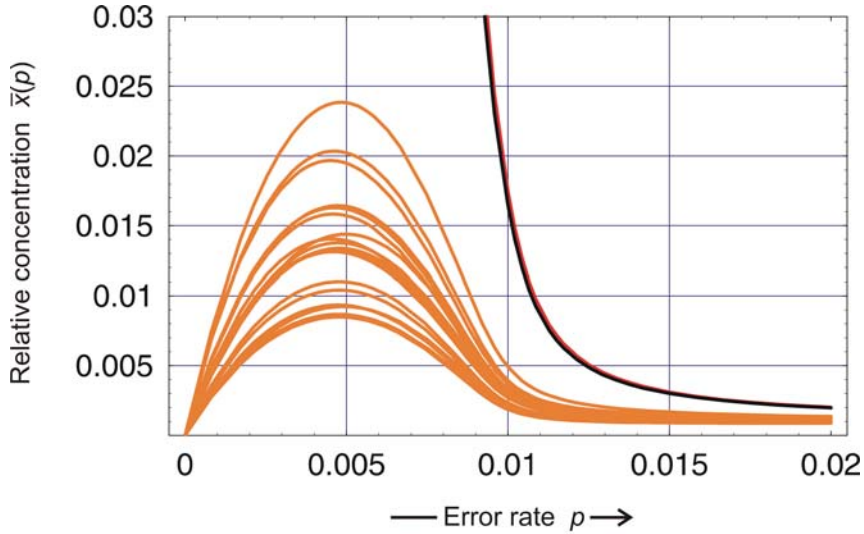
$\mathbf{d_H \geq 3}$

random fixation in the sense of Motoo Kimura

Pairs of genotypes in neutral replication networks


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

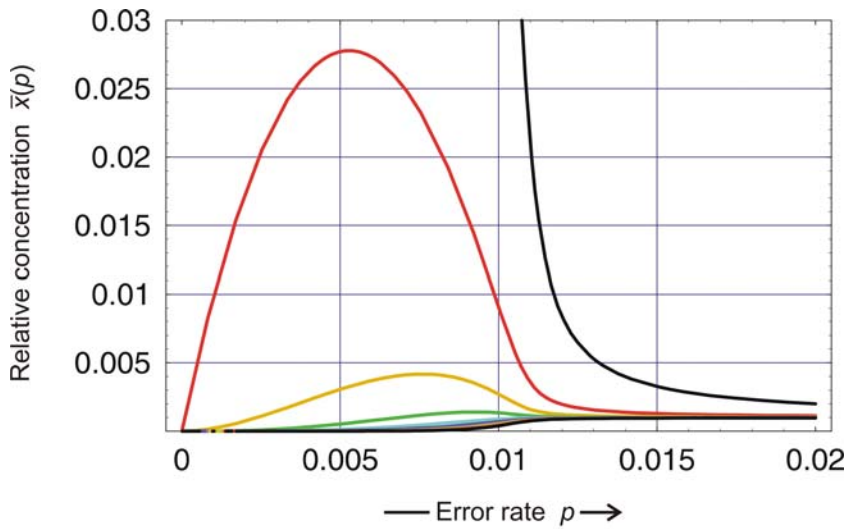




Neutral network

$\lambda = 0.01,\ s = 367$

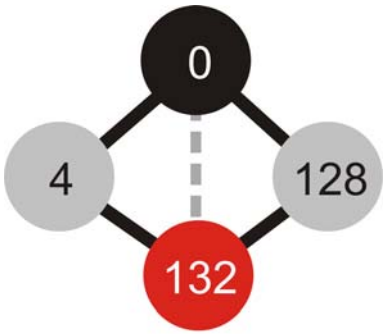
for comparison: $\lambda = 0$, $\sigma = 1.1$, d = 0



Neutral network: Individual sequences

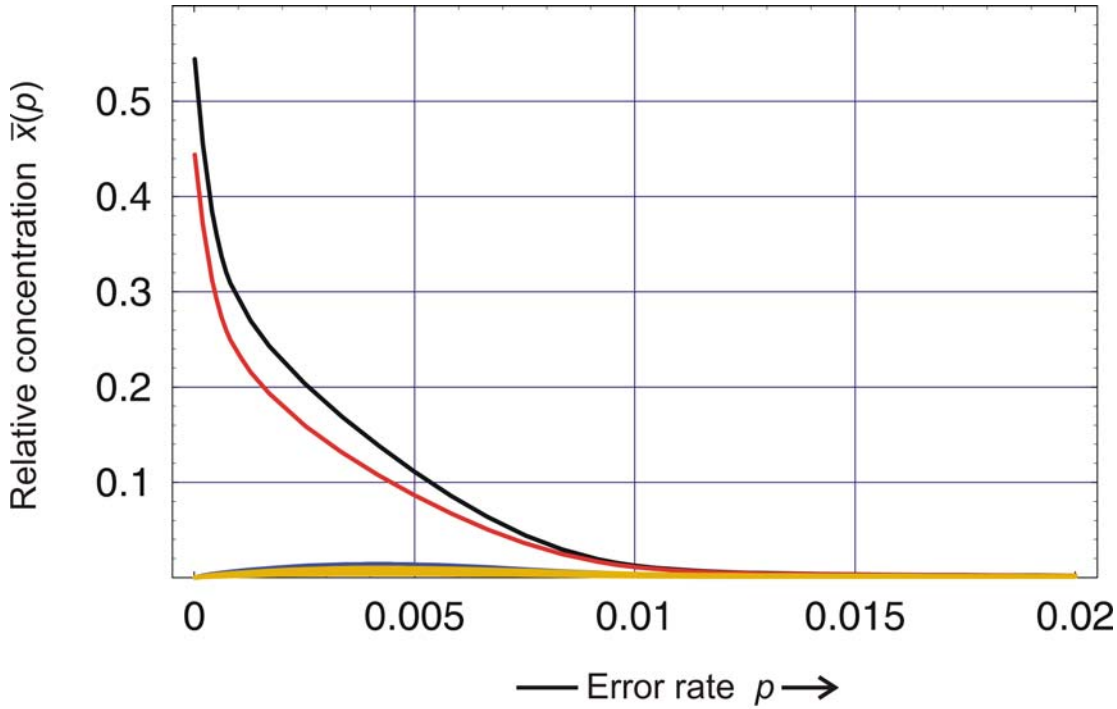
n = 10, $\sigma = 1.1$, d = 1.0

Consensus sequence of a quasispecies of two strongly coupled sequences of Hamming distance $d_H(X_{i,},X_j) = 1$.



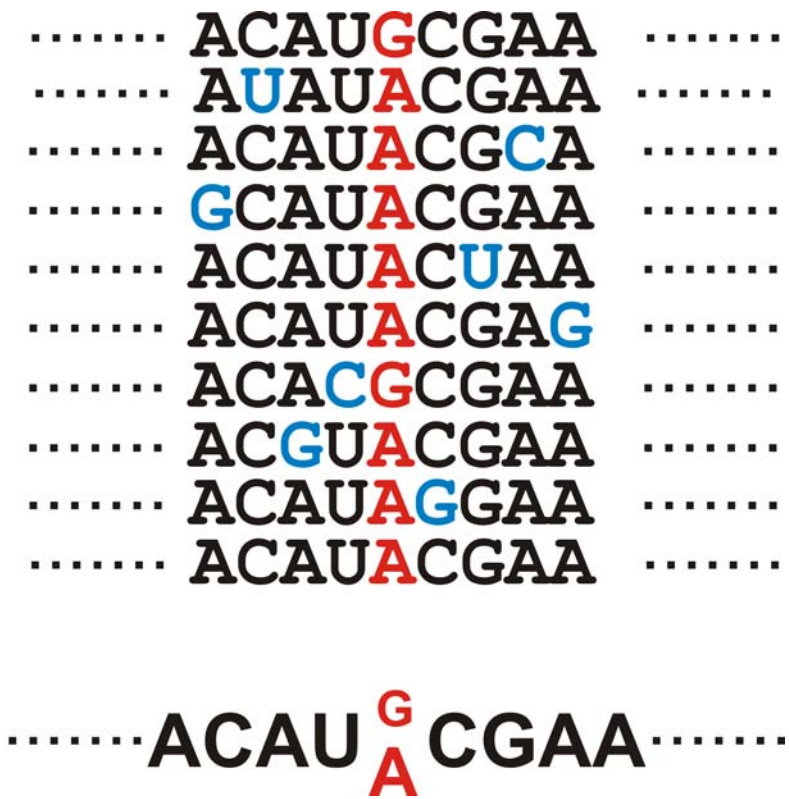
Neutral network

$\lambda = 0.01,\ s = 877$

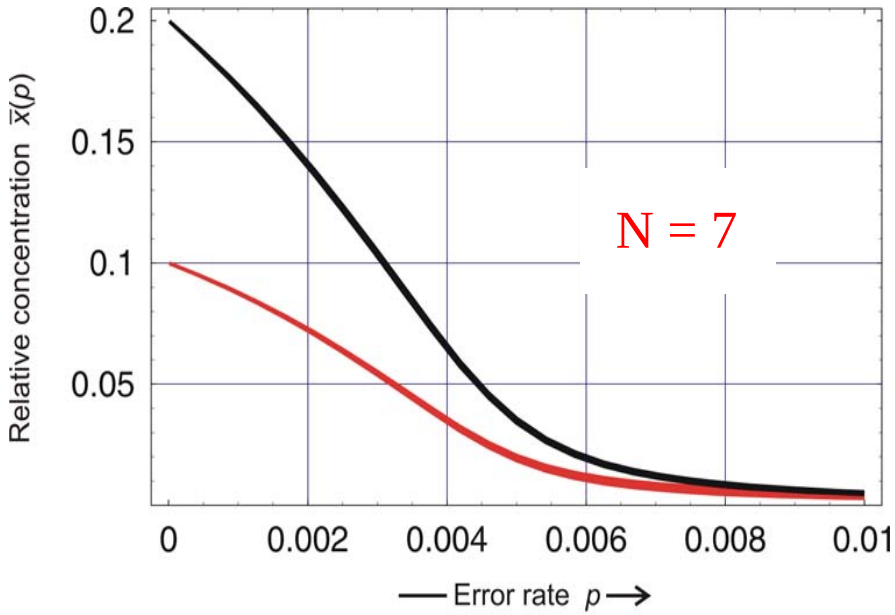


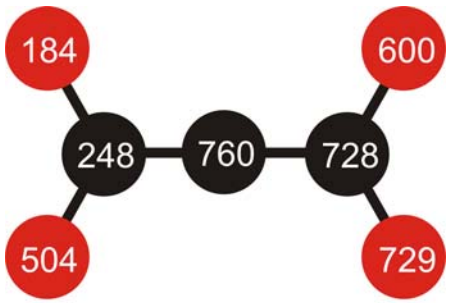
Neutral network: Individual sequences

n = 10, σ = 1.1, d = 1.0



Consensus sequence of a quasispecies of two strongly coupled sequences of Hamming distance $d_H(X_{i,},X_j) = 2$.





Neutral network

$\lambda = 0.10,\ s = 229$

Selection-mutation matrix W

$$W = \begin{pmatrix} f & O(\varepsilon^2) & \varepsilon & O(\varepsilon^2) & O(\varepsilon^2) & O(\varepsilon^2) & O(\varepsilon^2) \\ O(\varepsilon^2) & f & \varepsilon & O(\varepsilon^2) & O(\varepsilon^2) & O(\varepsilon^2) & O(\varepsilon^2) \\ \varepsilon & \varepsilon & f & \varepsilon & O(\varepsilon^2) & O(\varepsilon^2) & O(\varepsilon^2) \\ O(\varepsilon^2) & O(\varepsilon^2) & \varepsilon & f & \varepsilon & O(\varepsilon^2) & O(\varepsilon^2) \\ O(\varepsilon^2) & O(\varepsilon^2) & O(\varepsilon^2) & \varepsilon & f & \varepsilon & \varepsilon \\ O(\varepsilon^2) & O(\varepsilon^2) & O(\varepsilon^2) & O(\varepsilon^2) & \varepsilon & f & O(\varepsilon^2) \\ O(\varepsilon^2) & O(\varepsilon^2) & O(\varepsilon^2) & O(\varepsilon^2) & \varepsilon & O(\varepsilon^2) & f \end{pmatrix}$$

Adjacency matrix A

$$A = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

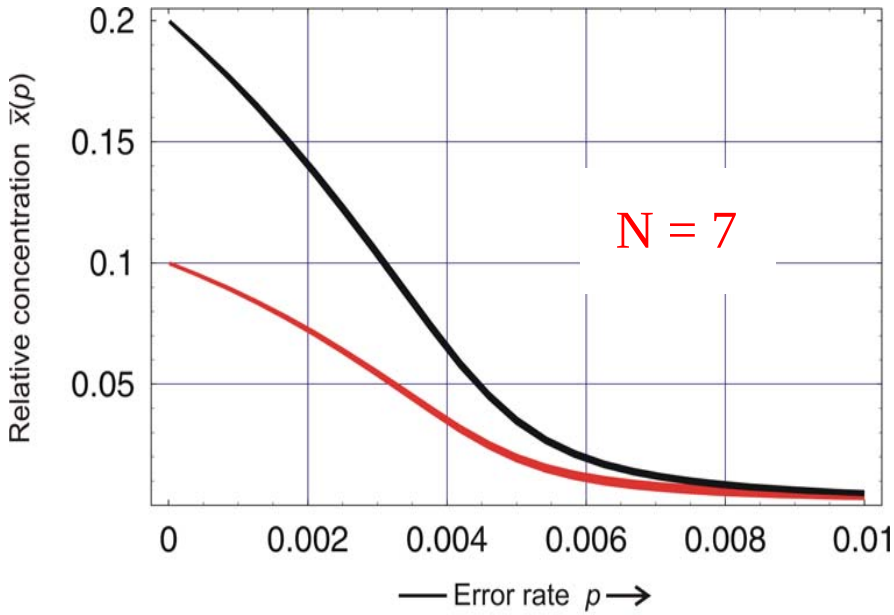
Eigenvalues of W and A

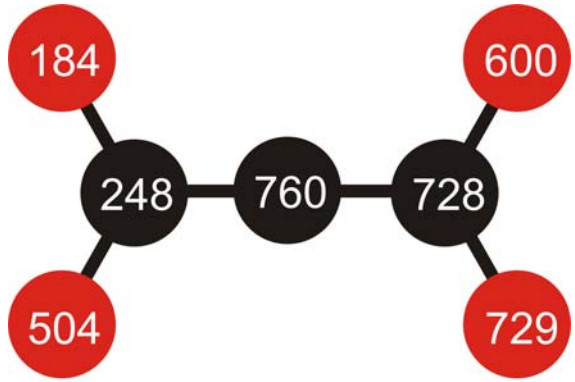
$$\begin{aligned} \lambda_0 &= f + 2\varepsilon\,, & \lambda_0 &= 2\,, \\ \lambda_1 &= f + \sqrt{2}\varepsilon\,, & \lambda_1 &= \sqrt{2}\,, \\ \lambda_{2,3,4} &= f\,, & \lambda_{2,3,4} &= 0\,, \\ \lambda_5 &= f - \sqrt{2}\varepsilon\,, & \lambda_5 &= -\sqrt{2}\,, \\ \lambda_6 &= f - 2\varepsilon\,, & \lambda_6 &= -2\,. \end{aligned}$$

Largest eigenvector of W and A

$$\xi_0 = (0.1, 0.1, 0.2, 0.2, 0.2, 0.1, 0.1)\,.$$

Computation of sequences in the core of a neutral network

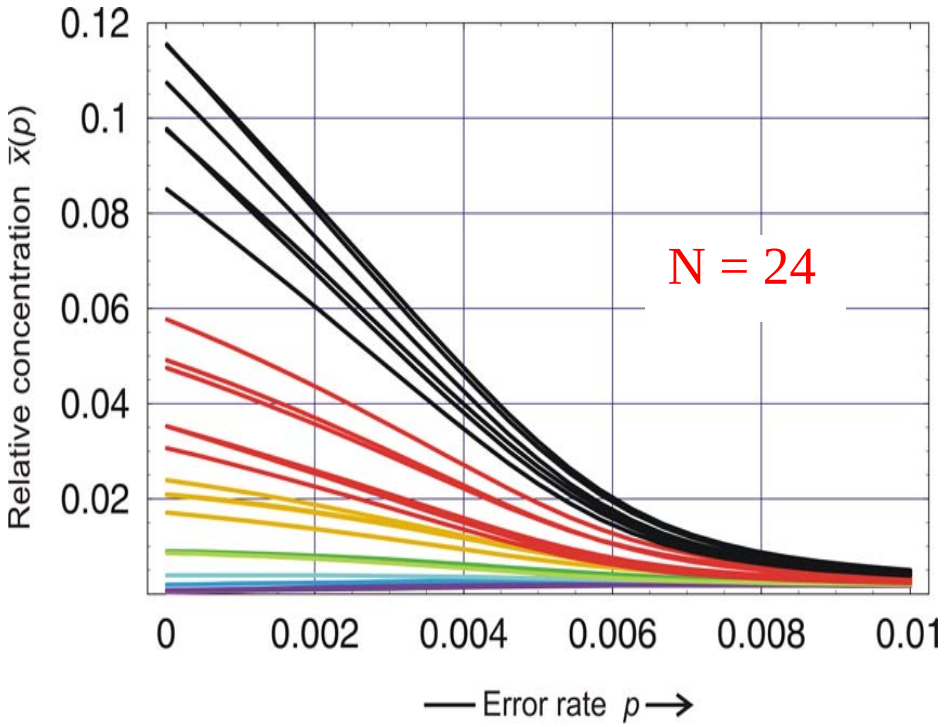


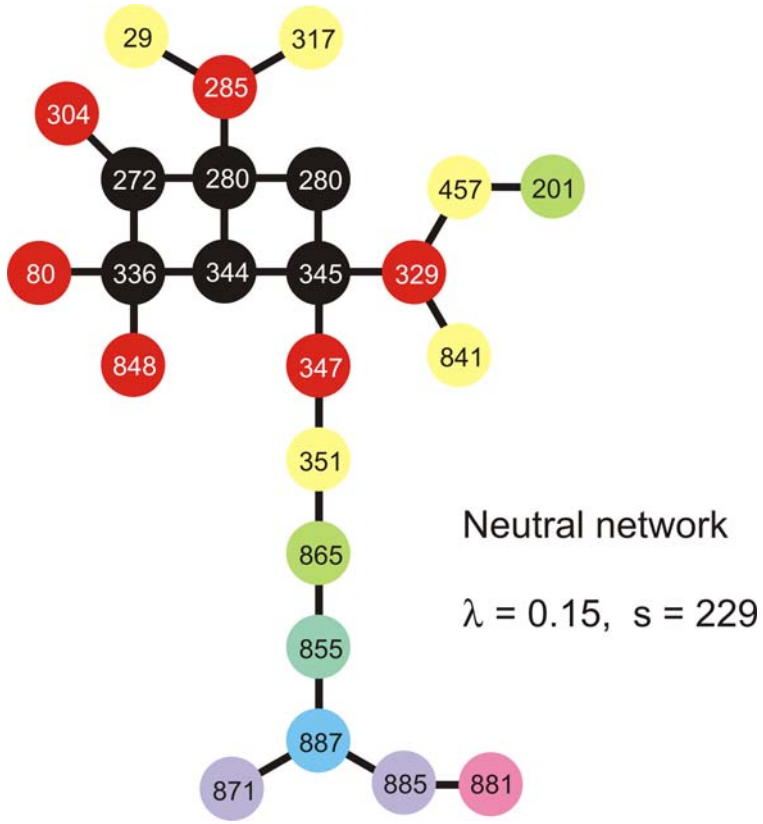


Neutral network

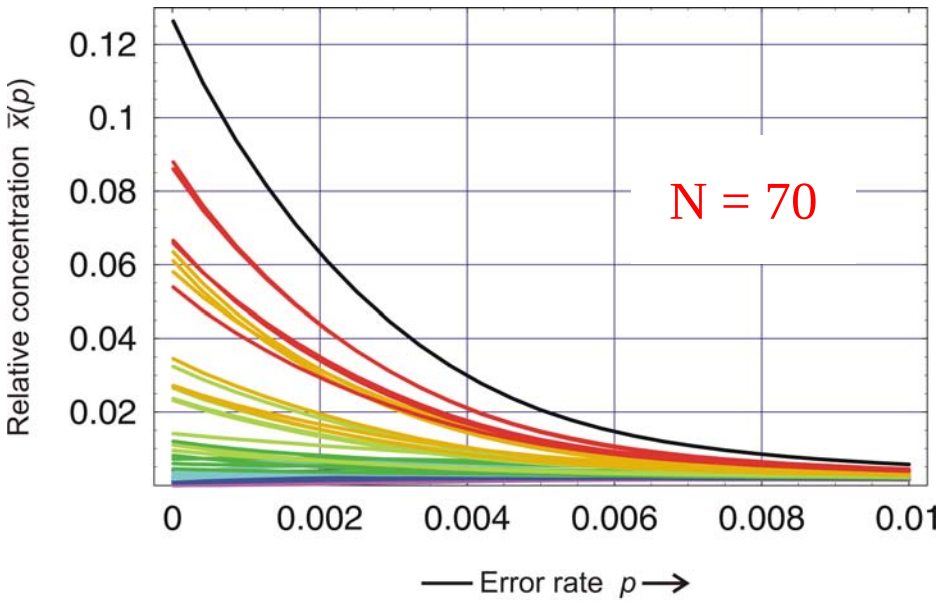
$\lambda = 0.10$, s = 229

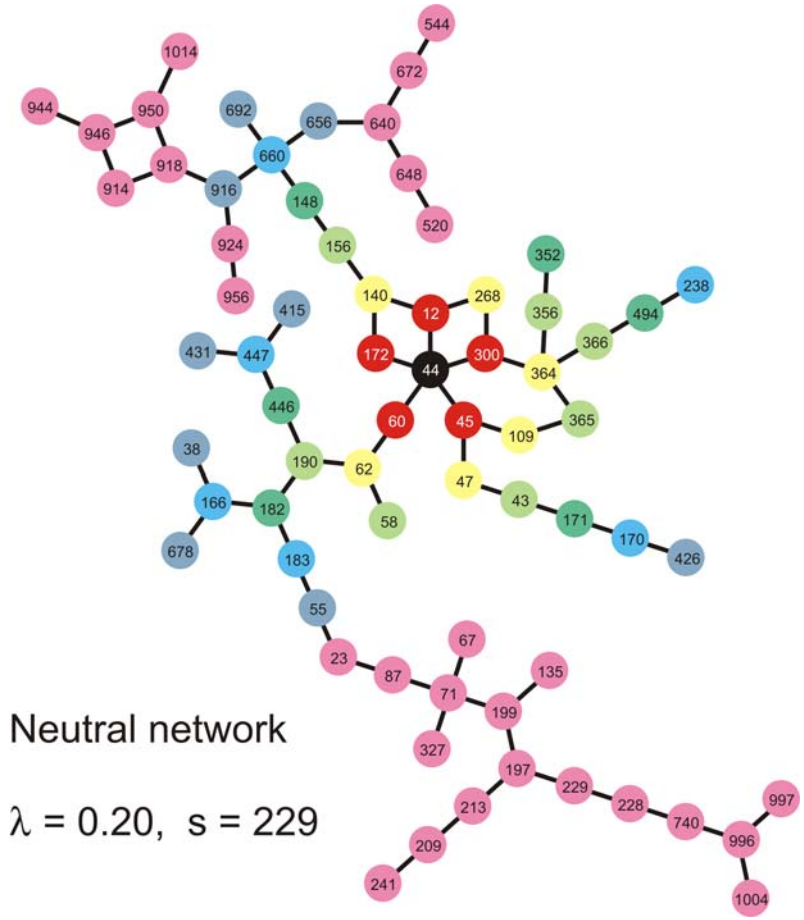
Neutral networks with increasing λ: $\lambda = 0.10$, s = 229



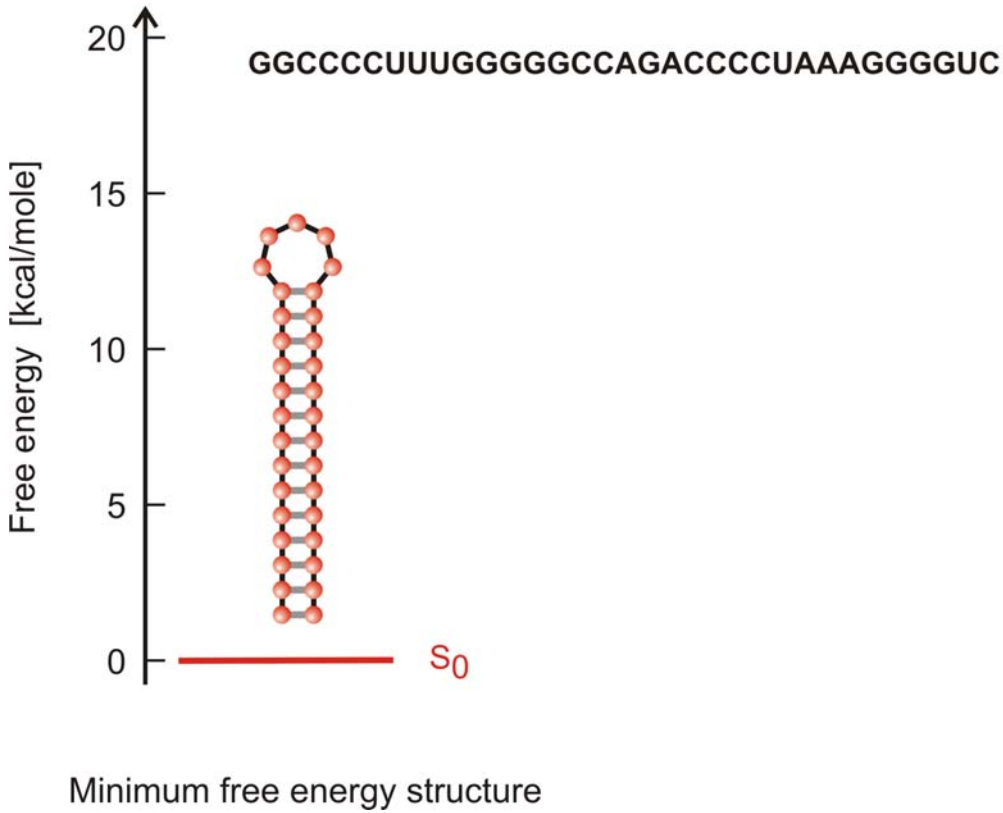


Neutral networks with increasing λ: λ = 0.15, s = 229

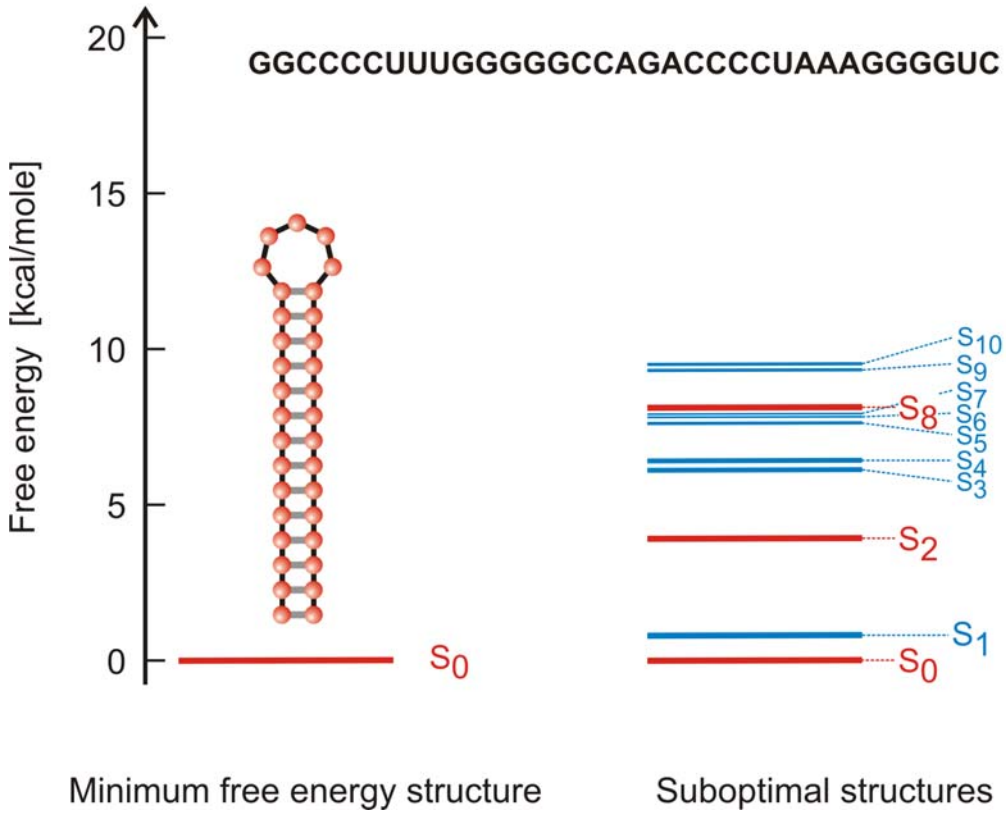




Neutral networks with increasing λ: λ = 0.20, s = 229



Extension of the notion of structure

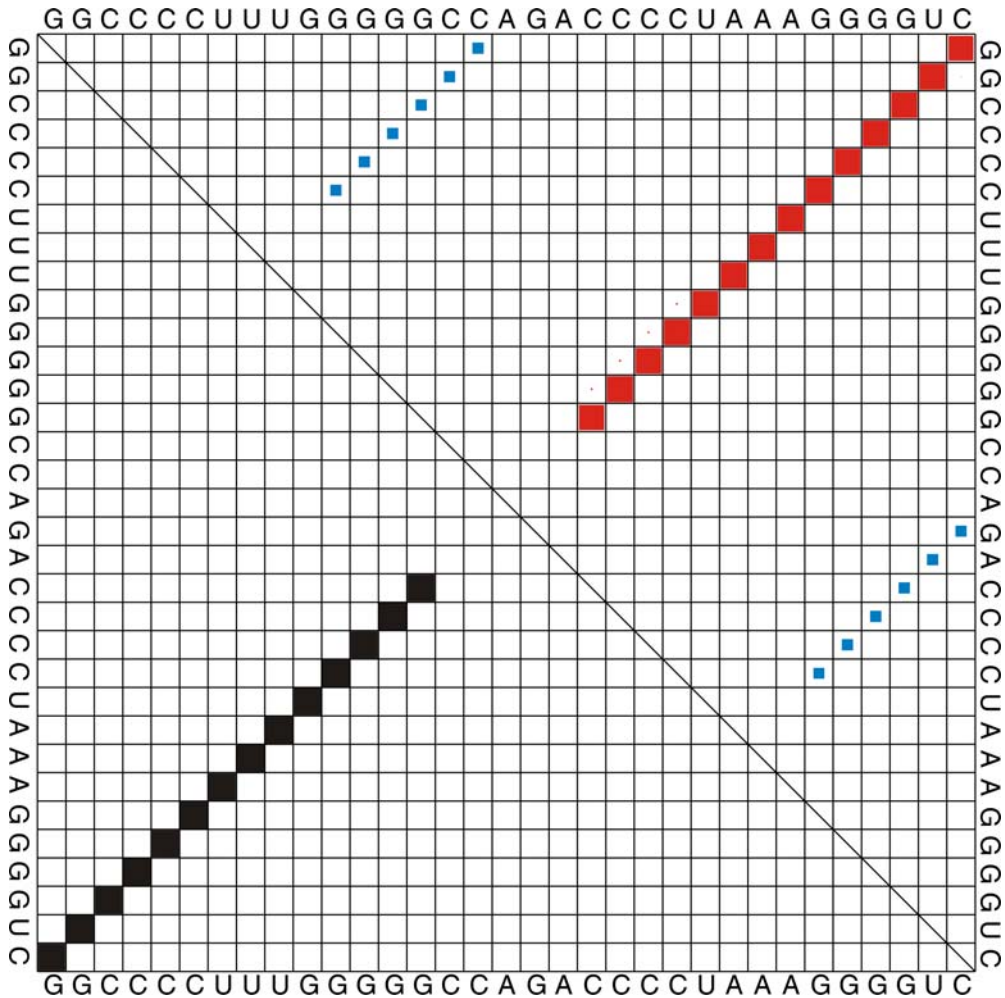


Extension of the notion of structure

GGCCCCUUUGGGGGCCAGACCCCUAAAGGGGUC

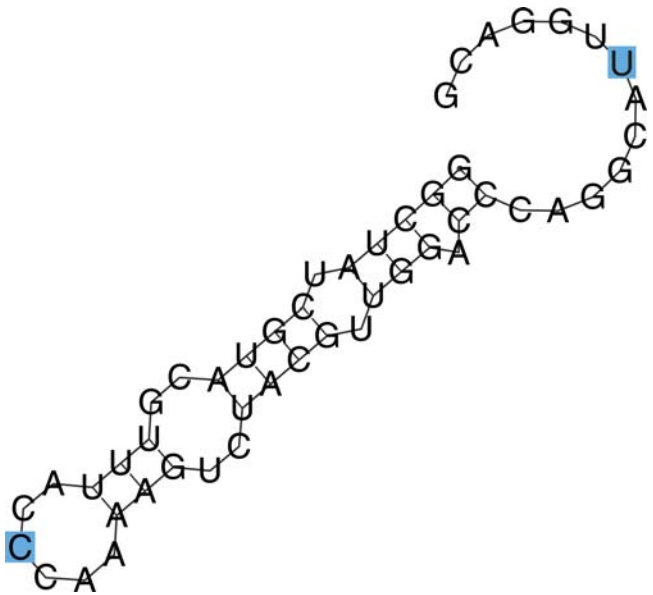
```
((((((((((((.....))))))))))))) -26.30

((((((....)))))).((((((....)))))) -25.30
.((((((((((((.....)))))))))))). -24.80
(((((((((((.......)))))))))))) -24.50
((((((....)))))).(((((......))))) -23.40
((((......)))))).((((((....)))))) -23.30
..((((((((((.....))))))))))).. -23.10
(((((((((((......)))).))))))) -23.00
.((((((((((.......))))))))))). -23.00
(((((((.((((((.....))))))))))))) -22.80
(((((((.(((((.....)))))).))))))) -22.70
((((((....))))))..((((....)))))). -22.70
((((((.(((((((.....))))))).)))))) -22.20
(((((((.((((.....)))).))))))) -22.10
(.((((((((((.....))))))))))).) -21.90
.((((((((((.....))))))))))).) -21.90
((((((....))))))...((((....)))).. -21.60
(((((((..(((((.....)))))..))))))) -21.50
.((((((((((......)))).))))))). -21.50
((((......)))))).((((......)))))) -21.40
.((((((.((((((.....))))))).))))). -21.30
..((((((((((.......))))))))))).. -21.30
```



mfe-weight: 0.7196

Suboptimal structures and partition function of a small RNA molecule: n = 33



GGCUAUCGUACGUUUAC**C**CAAAAGUCUACGUUGGACCCAGGCA**U**UGGACG

```
(((((.((((..(((......)))..)))).))).)).............  -7.30

..........((((((.((....((((.....))))...)).)))))).  -6.70
..........((((((.((....((((...))))).)).)))))).  -6.60
..(((.((((..(((......)))..)))).))).((((...)))).  -6.10
(((((.((((..(((......)))..)))).))).)).(........).  -6.00
(((((.((((..(((........)))..)))).))).)).........  -6.00
.(((.((..((((..(((......)))..)))).))....))).....  -6.00
```

GGCUAUCGUACGUUUAC**A**CAAAAGUCUACGUUGGACCCAGGCAUUGGACG

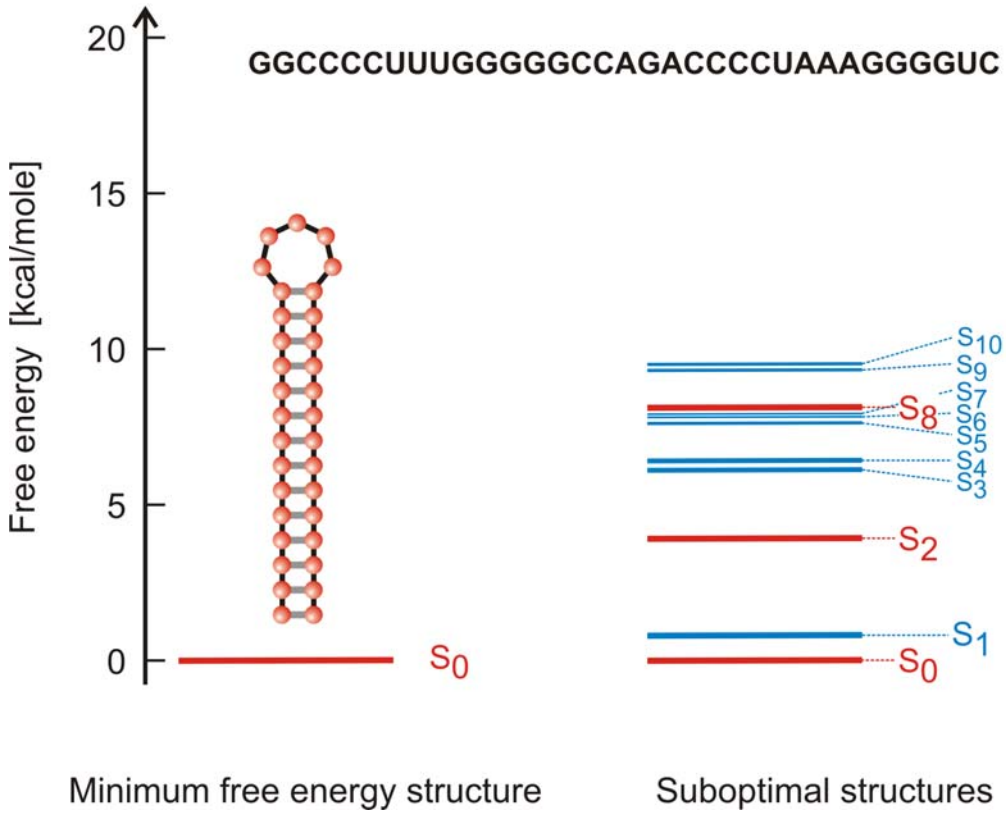
```
(((((.((((..(((......)))..)))).))).)).............  -7.30

.(((.((..((((..(((......)))..)))).))....))).....  -6.50
.(((.....((((..(((......)))..))))).....))).....  -6.30
..(((.((((..(((......)))..)))).))).((((...)))).  -6.10
(((((.((((..(((......)))..)))).))).)).(........).  -6.00
(((((.((((..(((........)))..)))).))).)).........  -6.00
.(((...((((((..(((......)))..)))))...))...))).....  -6.00
```

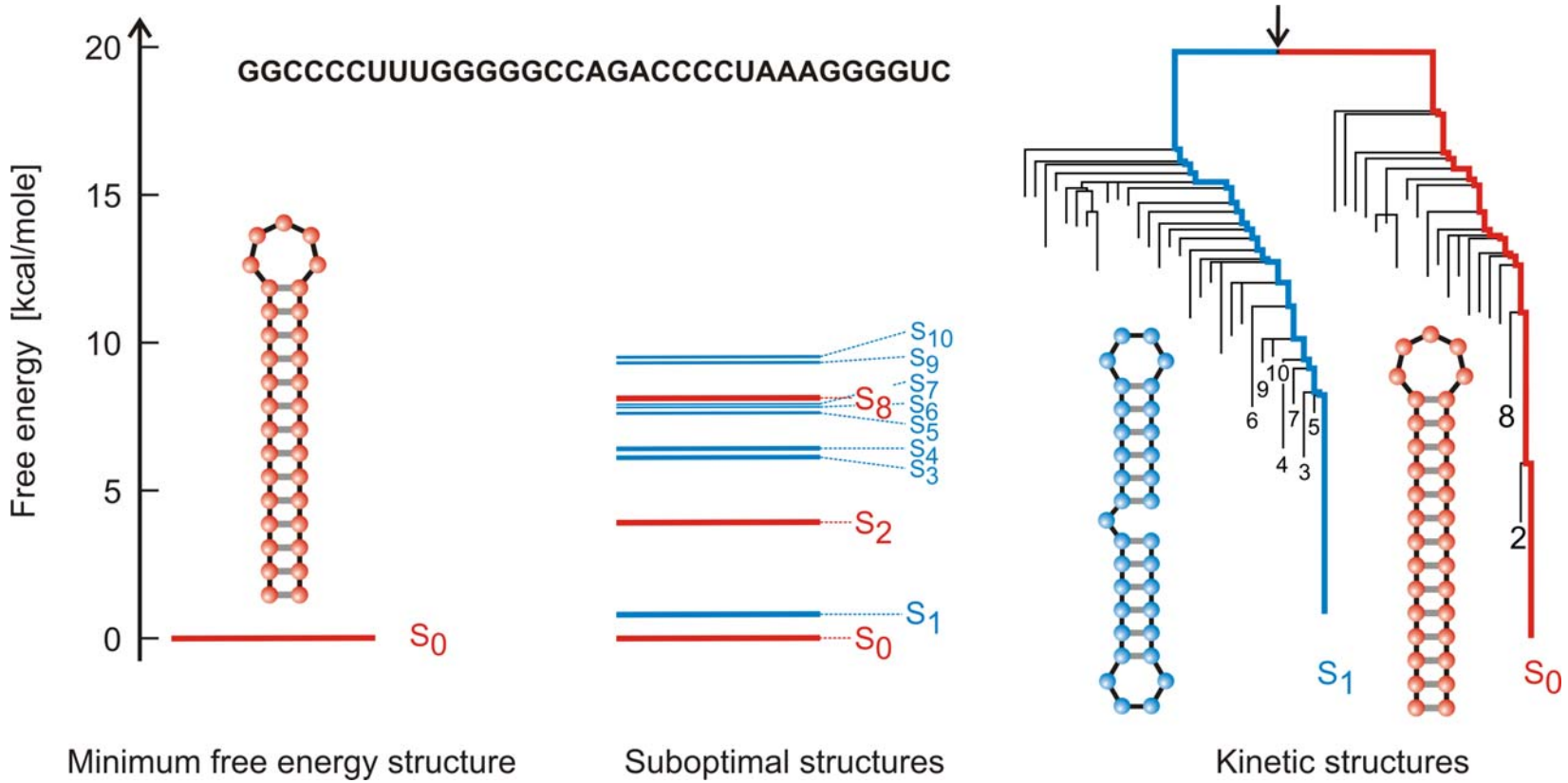
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCA**A**UGGACG

```
(((((.((((..(((......)))..)))).))).)).............  -7.30

..(((.((((..(((......)))..)))).))).((((.....)))...  -7.20
..........((((((.((....((((.....))))...)).)))))).  -6.70
..........((((((.((....((((...))))).)).)))))).  -6.60
(((((.((((..(((......)))..)))).))).)).)((.....)).  -6.50
(.(((.((((..(((......)))..)))).))).)((.....)))...  -6.30
.((((.((((..(((......)))..)))).))).)((.....)))...  -6.30
.....(((.(((..((......)).)))))).((.....)))...  -6.30
(.(((.((((..(((......)))..)))).))).((.....))).).  -6.10
.....((..((((..((......)).)))).)).((.....)))...  -6.10
......(((.((((...((....((((.....))))...)).)))).  -6.10
(((((.((((..(((......)))..)))).))).)).(........).  -6.00
(((((.((((..(((........)))..)))).))).)).........  -6.00
.(((.((..((((..(((......)))..)))).))....))).....  -6.00
......(((.((((...((....((((...))))).)).)))).))).  -6.00
```

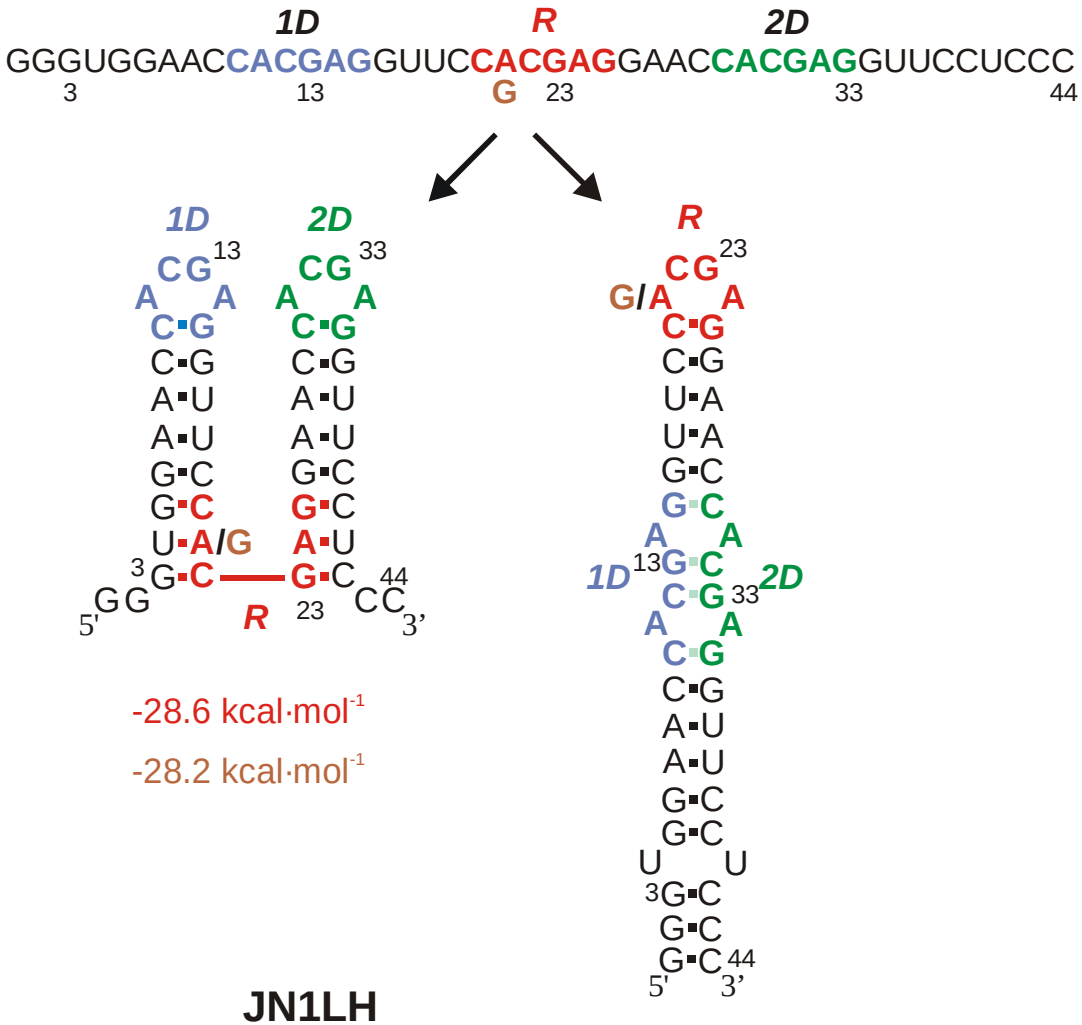


Extension of the notion of structure



Extension of the notion of structure

An RNA switch



JN1LH

J.H.A. Nagel, C. Flamm, I.L. Hofacker, K. Franke, M.H. de Smit, P. Schuster, and C.W.A. Pleij.

Structural parameters affecting the kinetic competition of RNA hairpin formation. *Nucleic Acids Res.* **34**:3568-3576, 2006.

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*).

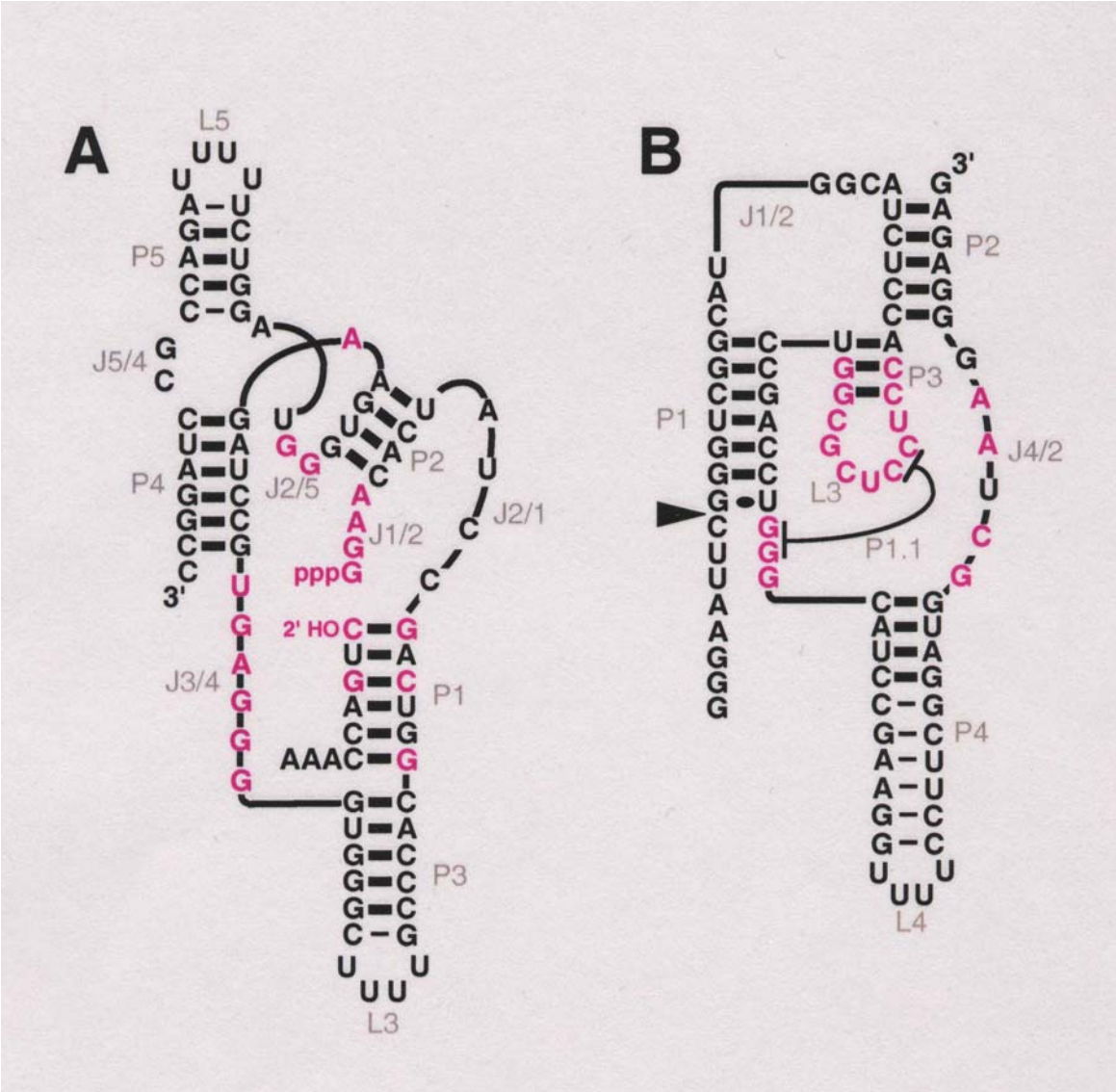
Theoretical analyses using algorithms for predicting RNA secondary structure have suggested that different neutral networks are interwoven and can approach each other very closely (*3*, *5–8*). Of particular interest is whether ribozyme neutral networks approach each other so closely that they intersect. 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.

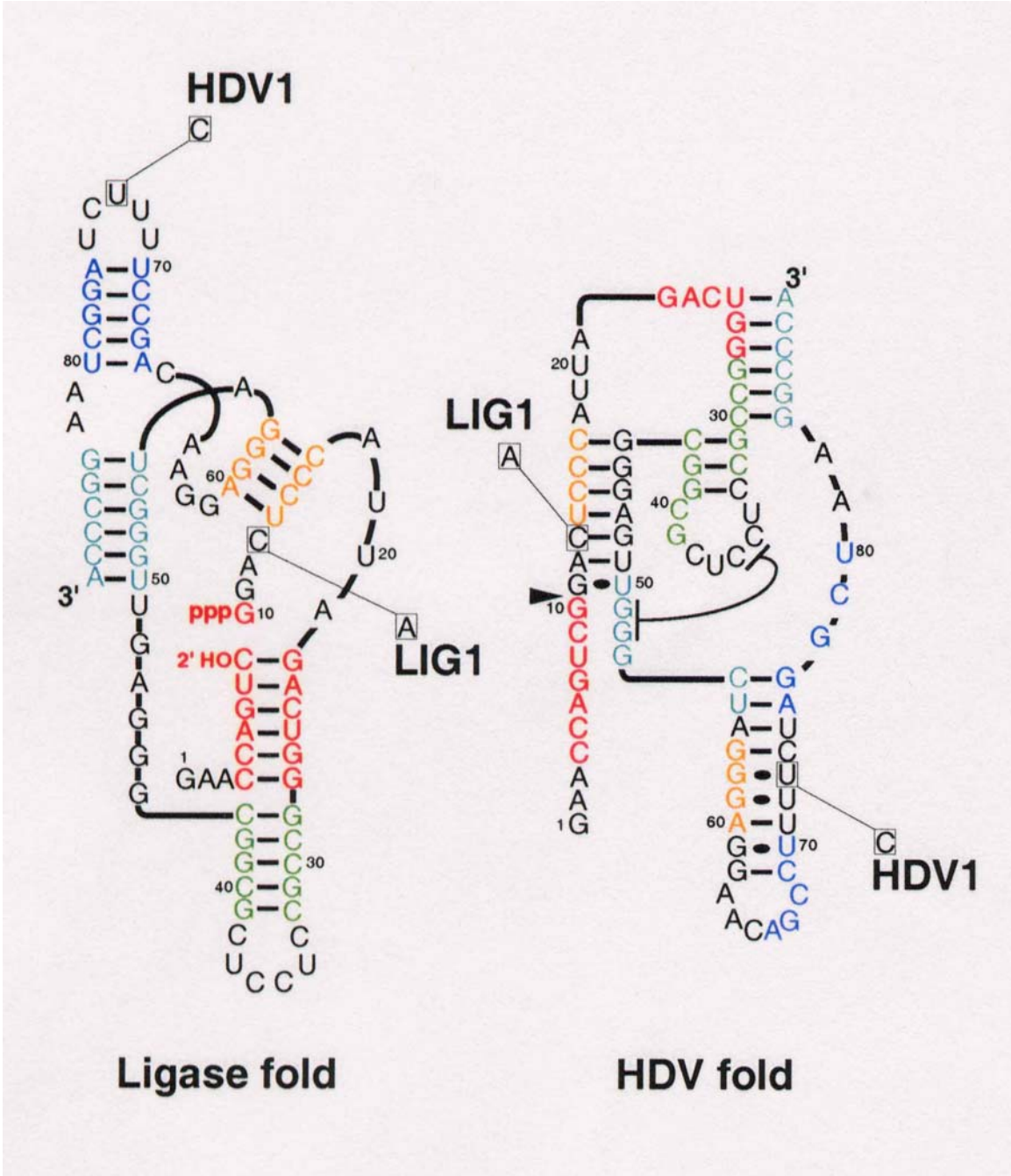
*To whom correspondence should be addressed. E-mail: dbartel@wi.mit.edu

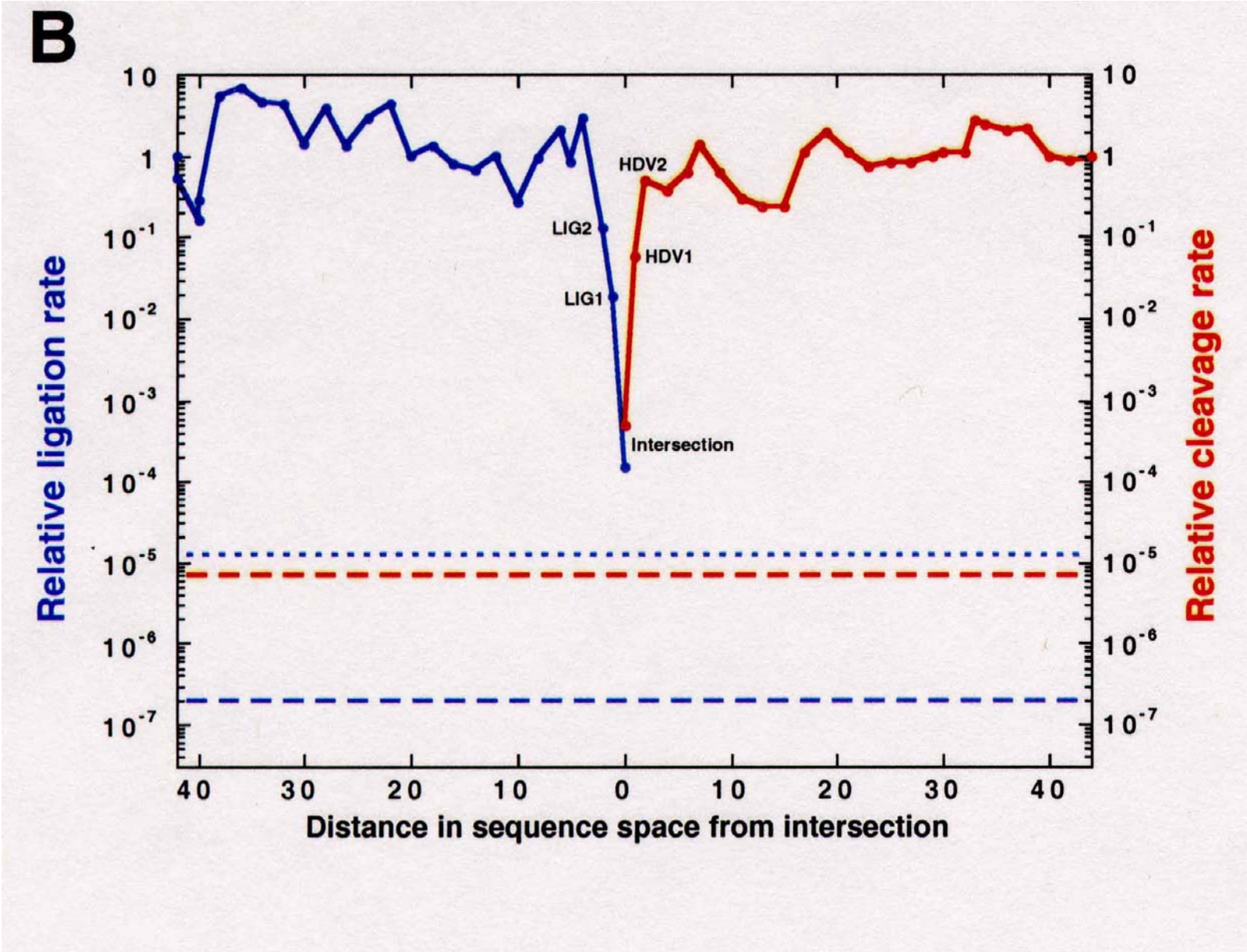


Two ribozymes of chain lengths n = 88 nucleotides: An artificial ligase (**A**) and a natural cleavage ribozyme of hepatitis-δ-virus (**B**)

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

Evolution of aptamers with a new specificity and new secondary structures from an ATP aptamer

ZHEN HUANG[1] and JACK W. SZOSTAK[2]

[1]Department of Chemistry, Brooklyn College, Ph.D. Programs of Chemistry and Biochemistry, The Graduate School of CUNY, Brooklyn, New York 11210, USA

[2]Howard Hughes Medical Institute, Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts 02114, USA

ABSTRACT

Small changes in target specificity can sometimes be achieved, without changing aptamer structure, through mutation of a few bases. Larger changes in target geometry or chemistry may require more radical changes in an aptamer. In the latter case, it is unknown whether structural and functional solutions can still be found in the region of sequence space close to the original aptamer. To investigate these questions, we designed an in vitro selection experiment aimed at evolving specificity of an ATP aptamer. The ATP aptamer makes contacts with both the nucleobase and the sugar. We used an affinity matrix in which GTP was immobilized through the sugar, thus requiring extensive changes in or loss of sugar contact, as well as changes in recognition of the nucleobase. After just five rounds of selection, the pool was dominated by new aptamers falling into three major classes, each with secondary structures distinct from that of the ATP aptamer. The average sequence identity between the original aptamer and new aptamers is 76%. Most of the mutations appear to play roles either in disrupting the original secondary structure or in forming the new secondary structure or the new recognition loops. Our results show that there are novel structures that recognize a significantly different ligand in the region of sequence space close to the ATP aptamer. These examples of the emergence of novel functions and structures from an RNA molecule with a defined specificity and fold provide a new perspective on the evolutionary flexibility and adaptability of RNA.

Keywords: Aptamer; specificity; fold; selection; RNA evolution

RNA **9**:1456-1463, 2003

Evidence for neutral networks and shape space covering

Evidence for neutral networks and intersection of apatamer functions

J Mol Evol (2003) 57:299–308
DOI: 10.1007/s00239-003-2481-y

JOURNAL OF MOLECULAR EVOLUTION

Evolutionary Landscapes for the Acquisition of New Ligand Recognition by RNA Aptamers

Daniel M. Held, S. Travis Greathouse, Amit Agrawal, Donald H. Burke

Department of Chemistry, Indiana University, Bloomington, IN 47405-7102, USA

Received: 15 November 2002 / Accepted: 8 April 2003

Abstract. The evolution of ligand specificity underlies many important problems in biology, from the appearance of drug resistant pathogens to the re-engineering of substrate specificity in enzymes. In studying biomolecules, however, the contributions of macromolecular sequence to binding specificity can be obscured by other selection pressures critical to bioactivity. Evolution of ligand specificity *in vitro*—unconstrained by confounding biological factors—is addressed here using variants of three flavin-binding RNA aptamers. Mutagenized pools based on the three aptamers were combined and allowed to compete during *in vitro* selection for GMP-binding activity. The sequences of the resulting selection isolates were diverse, even though most were derived from the same flavin-binding parent. Individual GMP aptamers differed from the parental flavin aptamers by 7 to 26 mutations (20 to 57% overall change). Acquisition of GMP recognition coincided with the loss of FAD (flavin-adenine dinucleotide) recognition in all isolates, despite the absence of a counter-selection to remove FAD-binding RNAs. To examine more precisely the proximity of these two activities within a defined sequence space, the complete set of all intermediate sequences between an FAD-binding aptamer and a GMP-binding aptamer were synthesized and assayed for activity. For this set of sequences, we observe a portion of a neutral network for FAD-binding function separated from GMP-binding function by a distance of three mutations. Furthermore, enzymatic probing of these aptamers revealed gross structural remodeling of the RNA coincident with the switch in ligand recognition. The capacity for neutral drift along an FAD-binding network in such close approach to RNAs with GMP-binding activity illustrates the degree of phenotypic buffering available to a set of closely related RNA sequences—defined as the set's functional tolerance for point mutations—and supports neutral evolutionary theory by demonstrating the facility with which a new phenotype becomes accessible as that buffering threshold is crossed.

Key words: Aptamers — RNA structure — Phenotypic buffering — Fitness landscapes — Neutral evolutionary theory — Flavin — GMP

Correspondence to: Donald H. Burke; *email:* dhburke@indiana.edu

Introduction

RNA aptamers targeting small molecules serve as useful model systems for the study of the evolution and biophysics of macromolecular binding interactions. Because of their small sizes, the structures of several such complexes have been determined to atomic resolution by NMR spectrometry or X-ray crystallography (reviewed by Herman and Patel 2000). Moreover, aptamers can be subjected to mutational and evolutionary pressures for which survival is based entirely on ligand binding, without the complicating effects of simultaneous selection pressures for bioactivity, thus allowing the relative contributions of each activity to be evaluated separately.

Neutrality in molecular structures and its role in evolution:

• Neutrality is an essential feature in biopolymer structures at the resolution that is relevant for function.

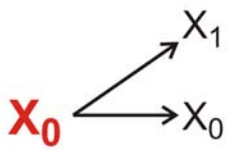
• Neutrality manifests itself in the search for minimum free energy structures.

• Diversity in function despite neutrality in structures results from differences in suboptimal conformations and folding kinetics.

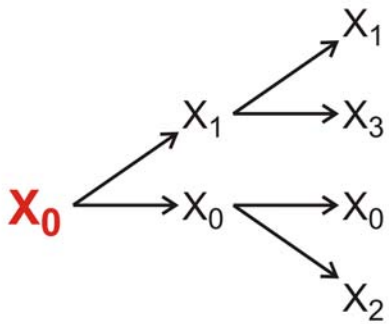
• Neutrality is indispensible for optimization and adaptation.

1. Requirements for information processing
2. The chemistry of Darwinian evolution
3. RNA sequences and structures
4. Consequences of neutrality
5. **Evolutionary optimization of RNA structure**

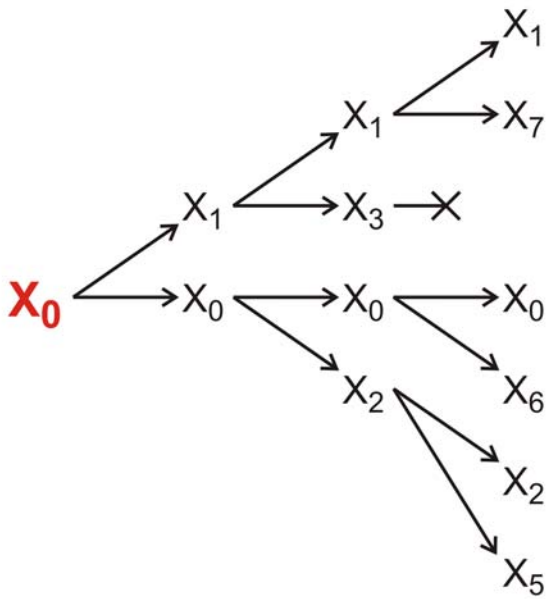
Evolution of RNA molecules as a Markow process and its analysis by means of the relay series



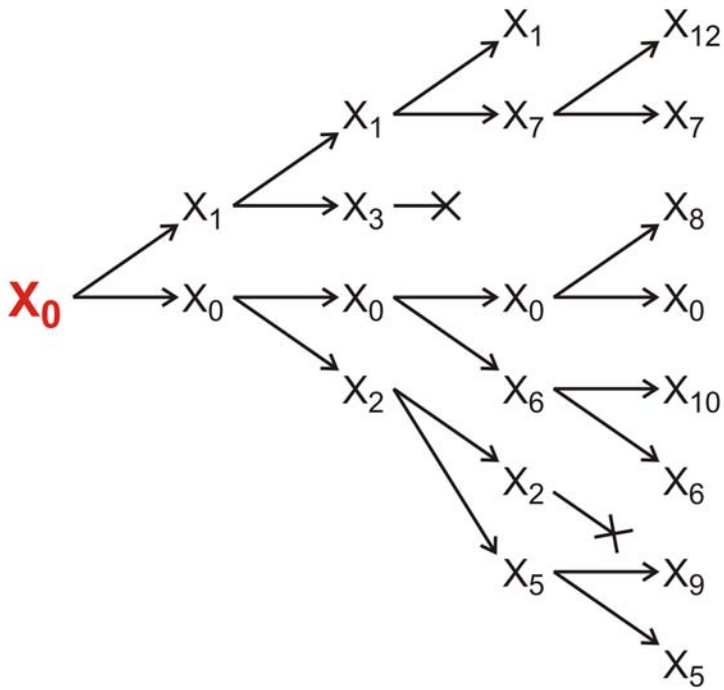
Evolution of RNA molecules as a Markow process and its analysis by means of the relay series



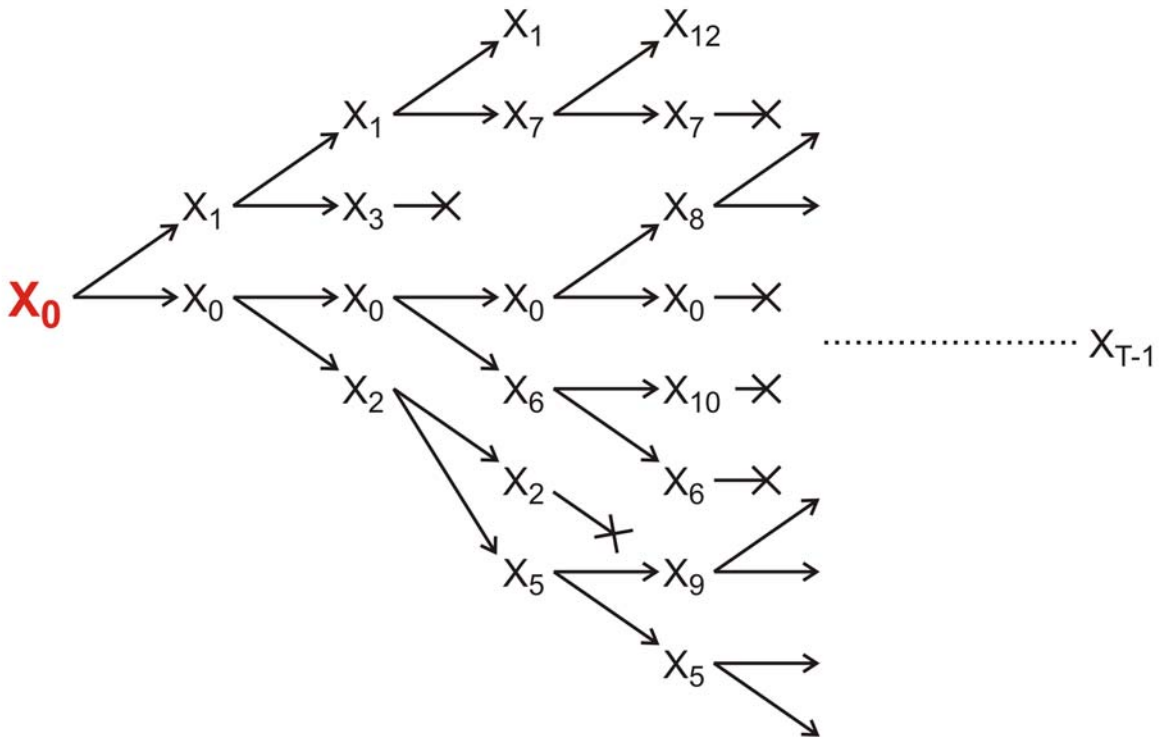
Evolution of RNA molecules as a Markow process and its analysis by means of the relay series



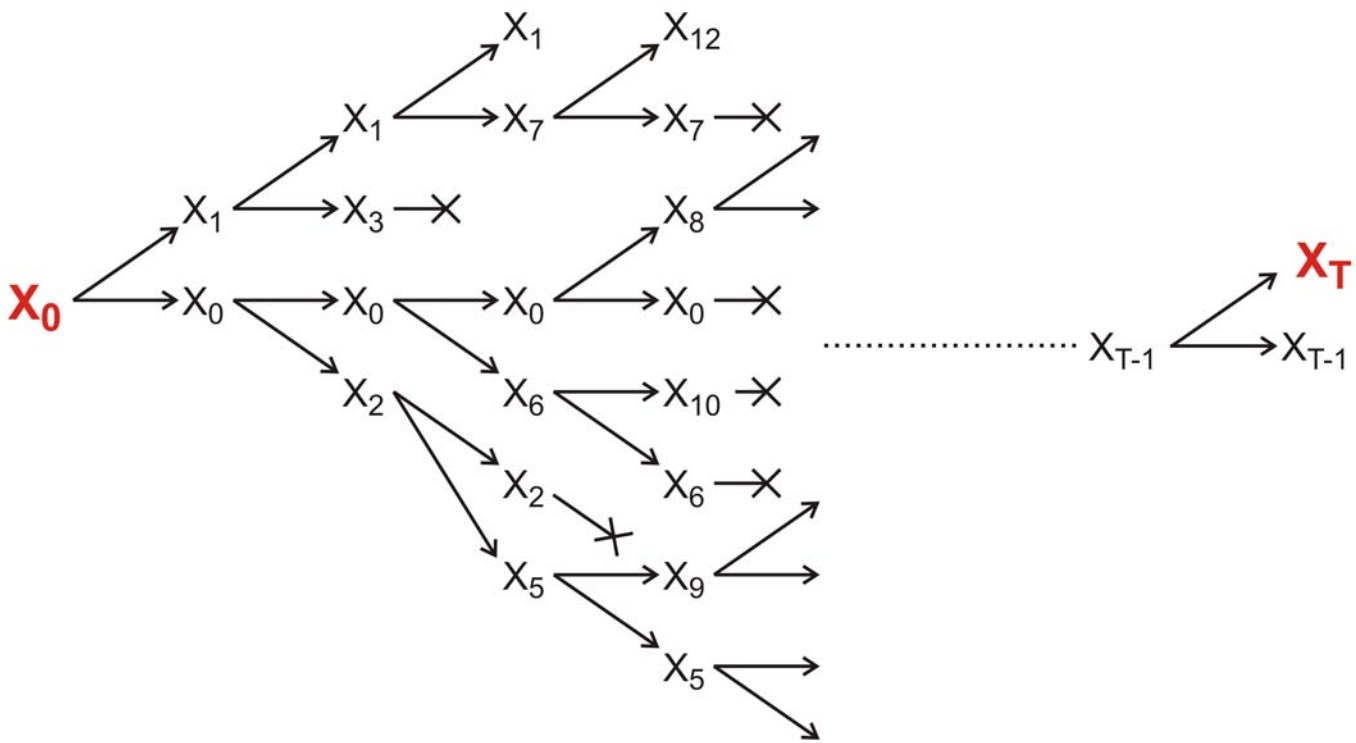
Evolution of RNA molecules as a Markow process and its analysis by means of the relay series



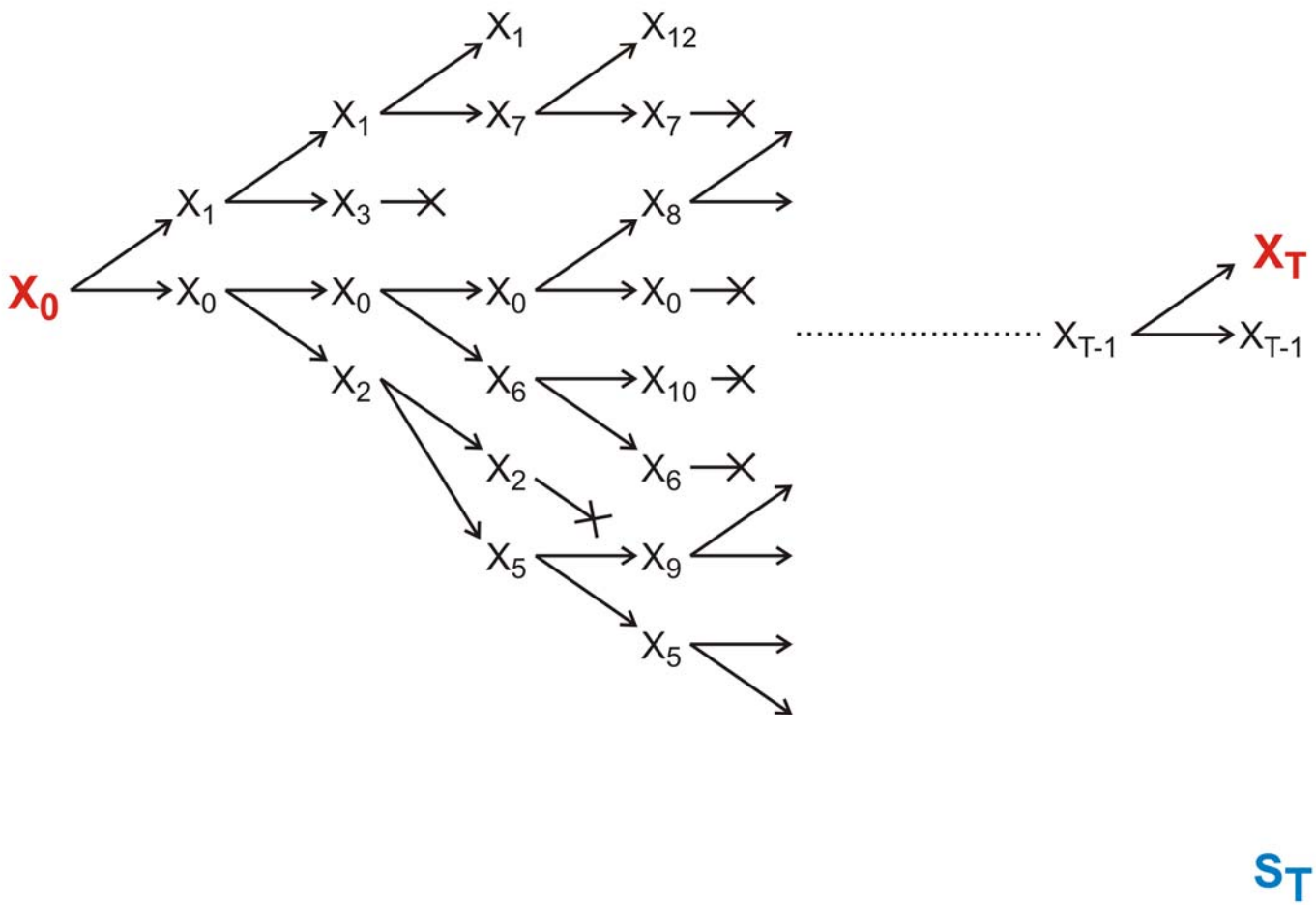
Evolution of RNA molecules as a Markow process and its analysis by means of the relay series



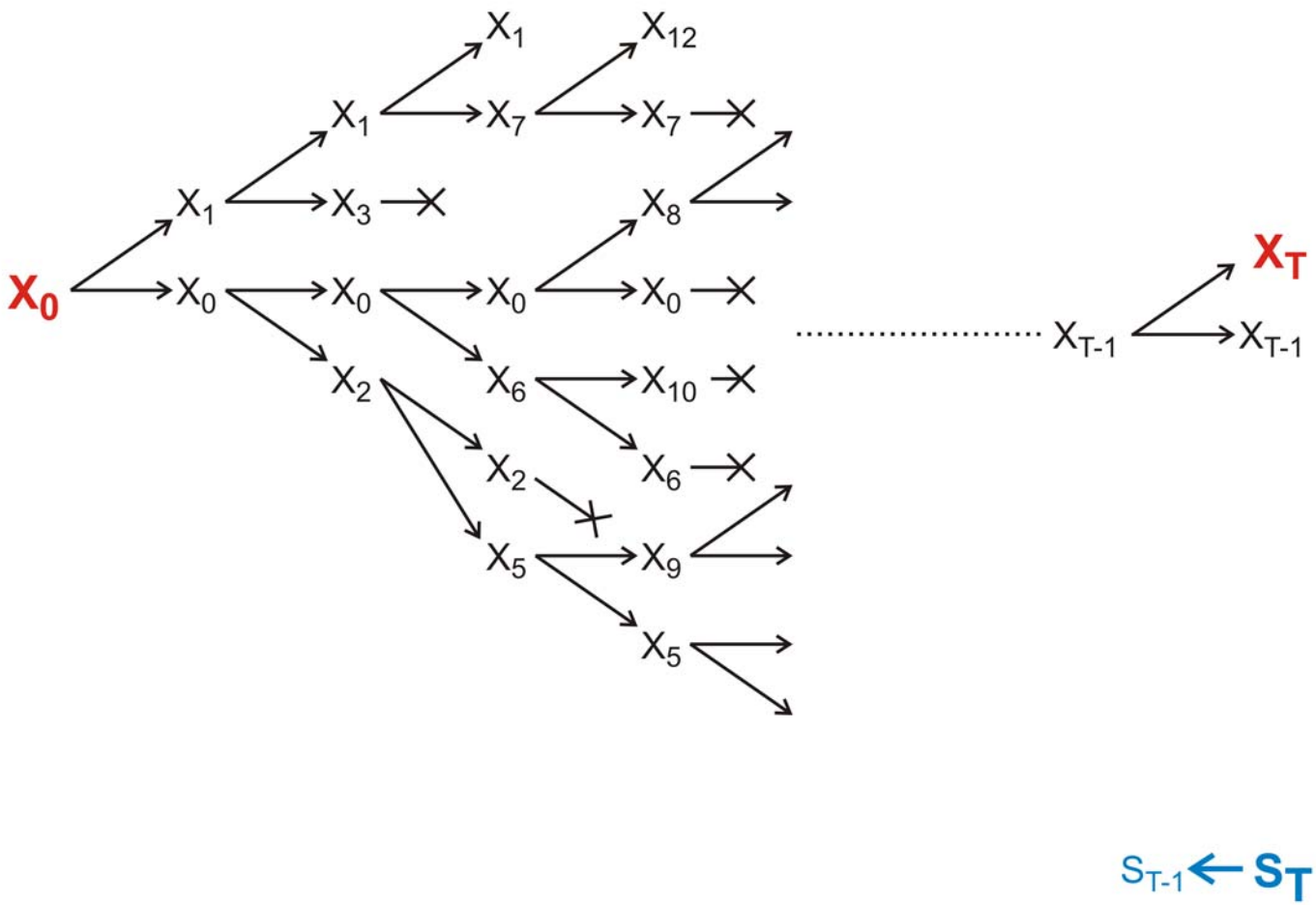
Evolution of RNA molecules as a Markow process and its analysis by means of the relay series



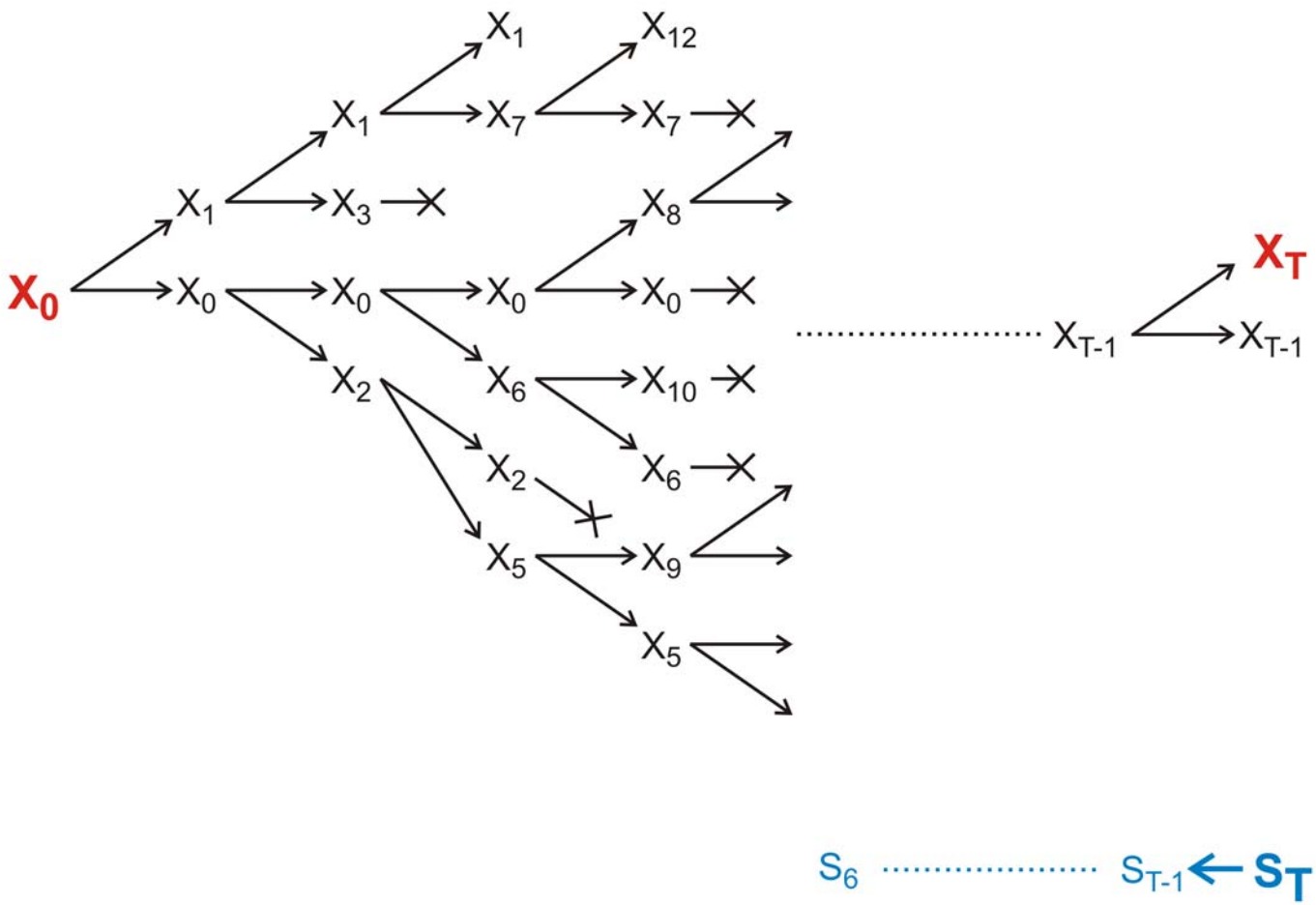
Evolution of RNA molecules as a Markow process and its analysis by means of the relay series



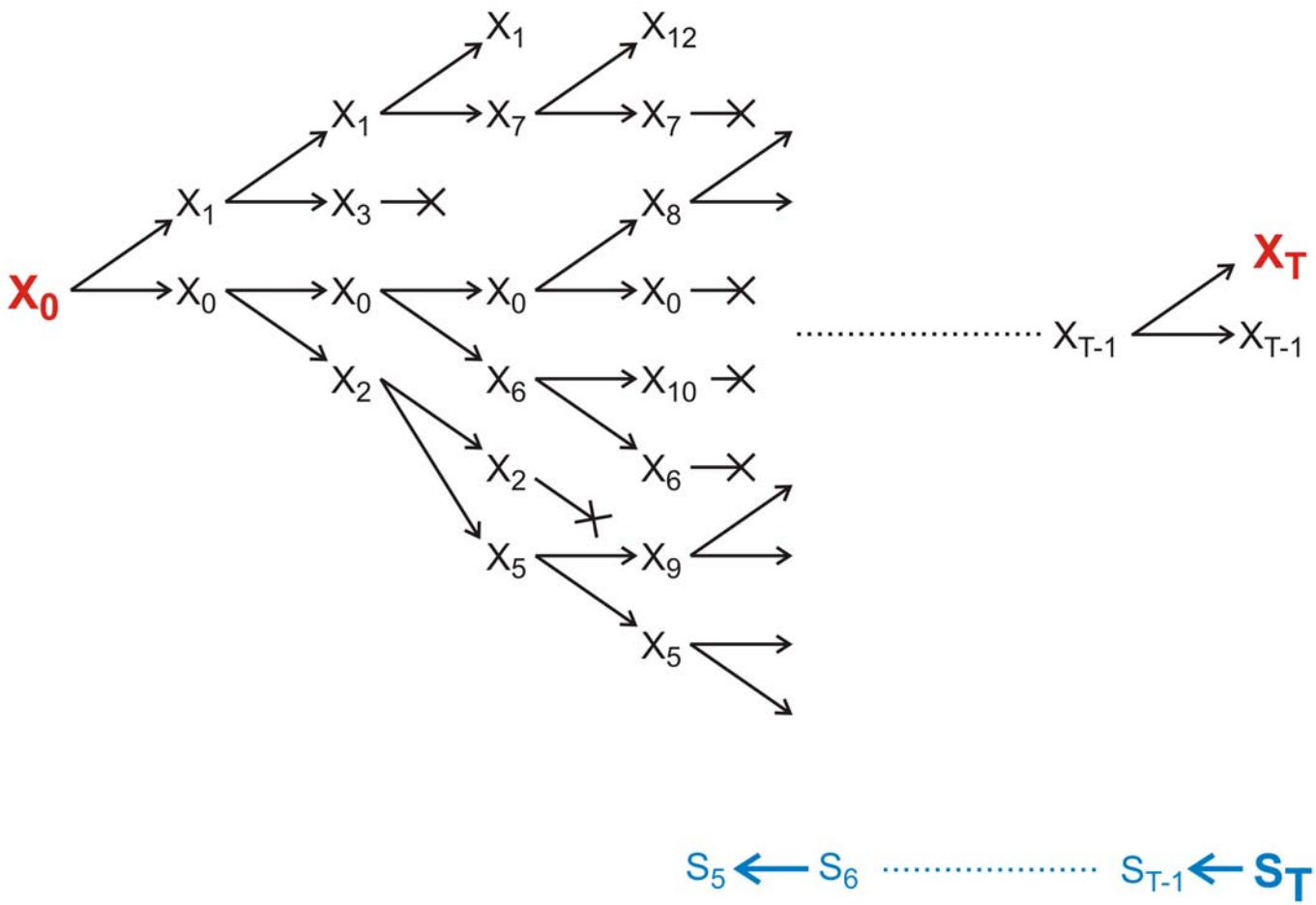
Evolution of RNA molecules as a Markow process and its analysis by means of the relay series



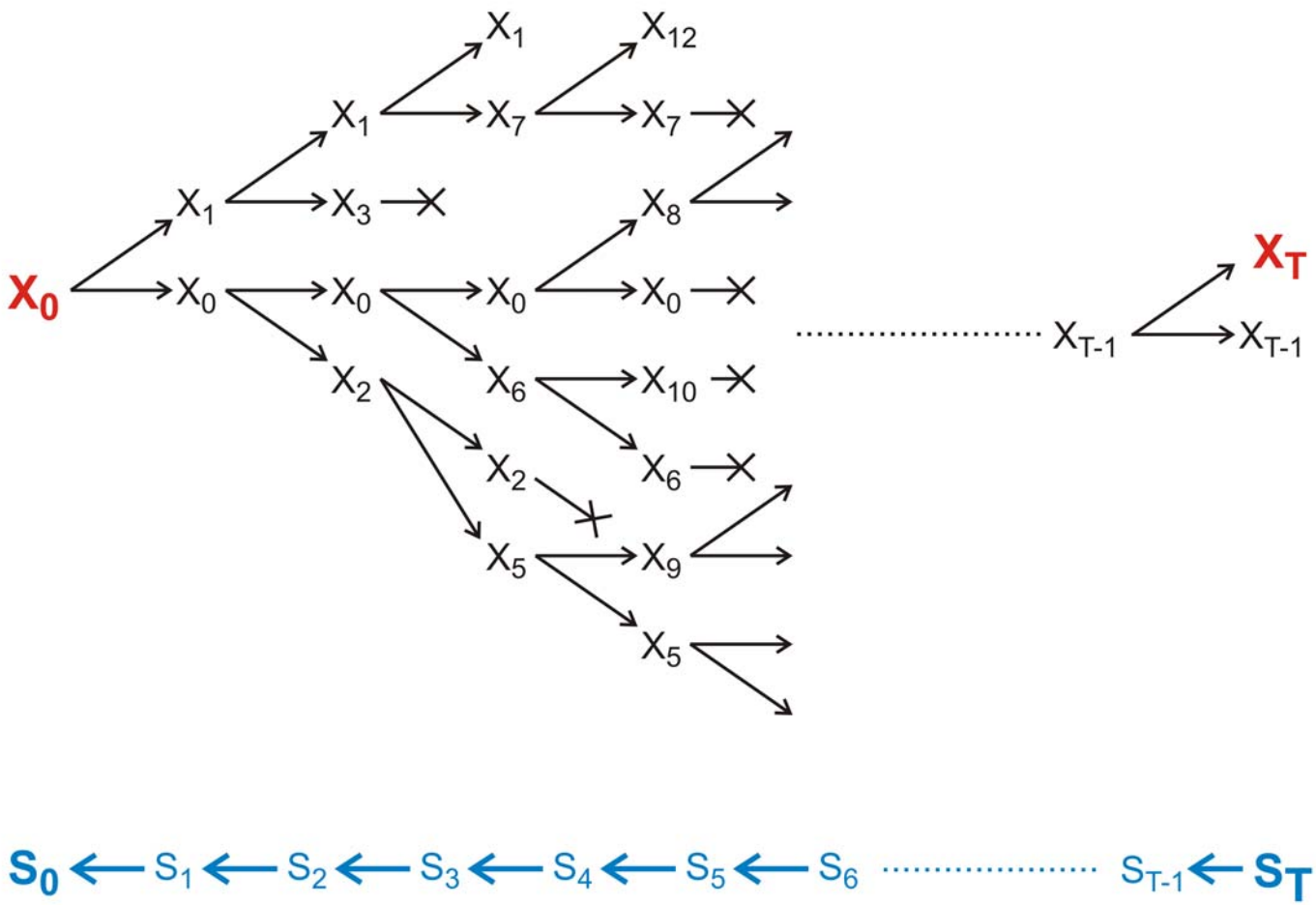
Evolution of RNA molecules as a Markow process and its analysis by means of the relay series



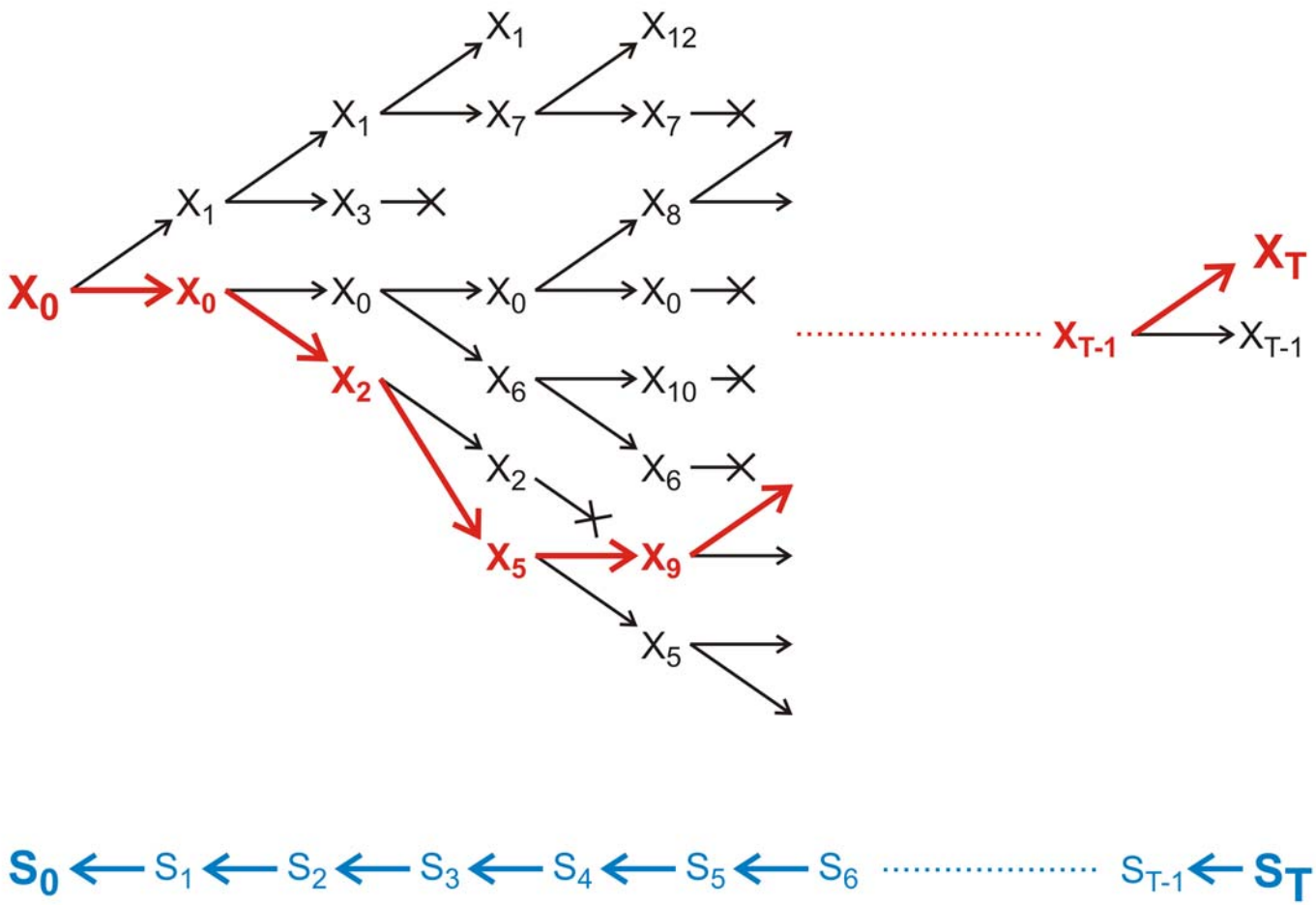
Evolution of RNA molecules as a Markow process and its analysis by means of the relay series



Evolution of RNA molecules as a Markow process and its analysis by means of the relay series



Evolution of RNA molecules as a Markow process and its analysis by means of the relay series



Evolution of RNA molecules as a Markow process and its analysis by means of the relay series

Evolution *in silico*

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

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 [(*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′-GCATGACCTGCCGGCTAAT-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 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.

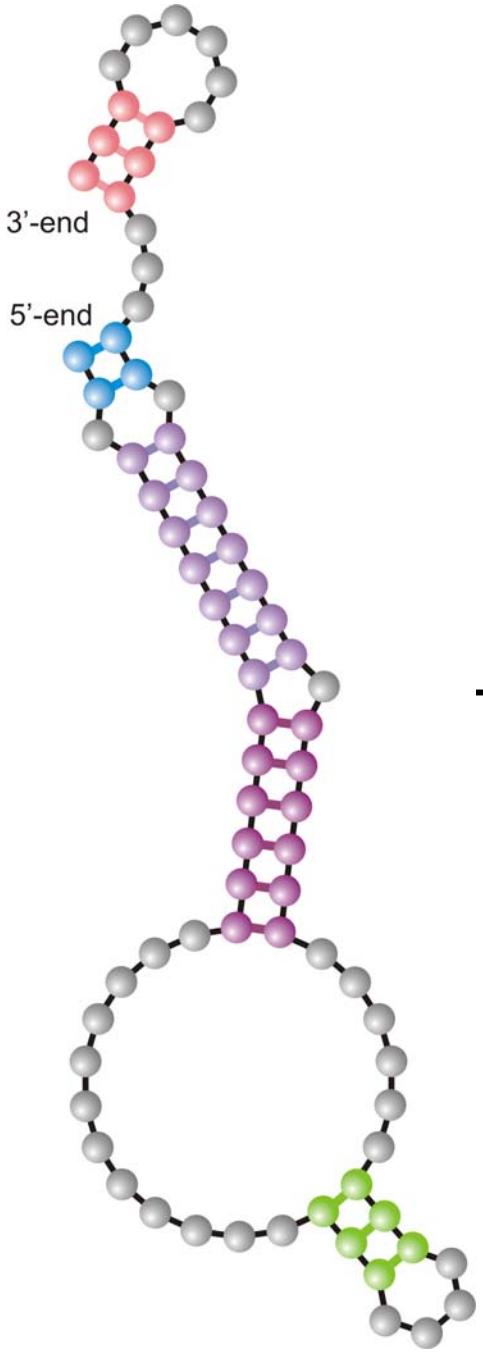
We focus on the narrow domain of RNA secondary structure, which is currently the simplest computationally tractable, yet realistic phenotype (*2*). This choice enables the definition and exploration of concepts that may prove useful in a wider context. RNA secondary structures represent a coarse level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

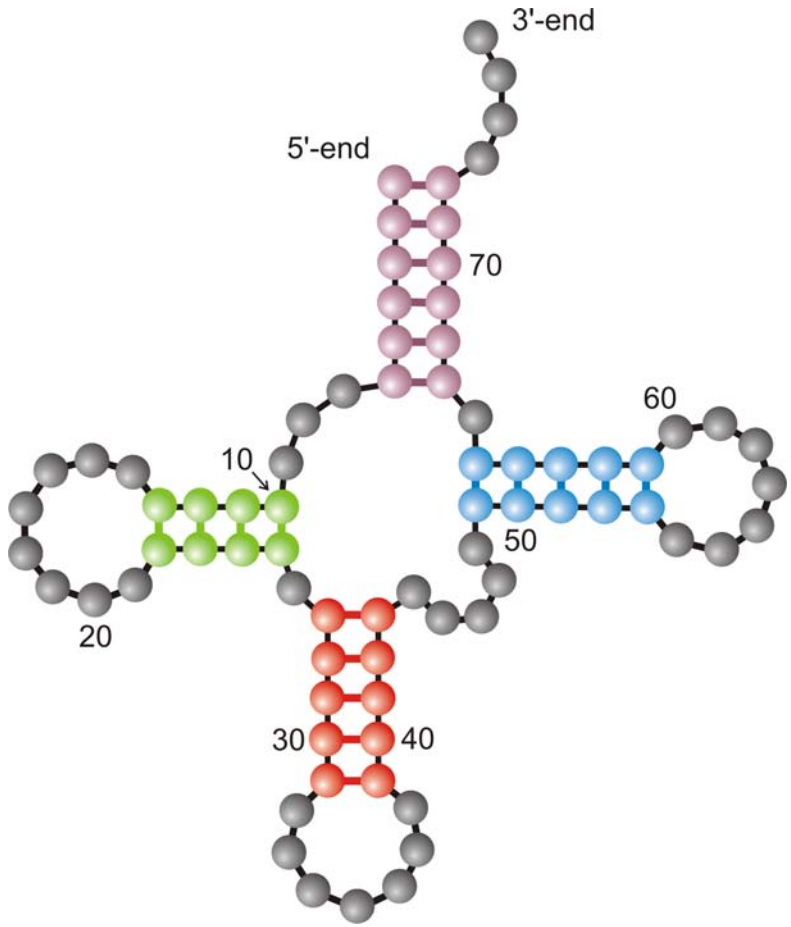
An instance representing in its qualitative features all the simulations we performed is shown in Fig. 1A. Starting with identical sequences folding into a random shape, the simulation was stopped when the population became dominated by the target, here a canonical tRNA shape. The black curve traces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

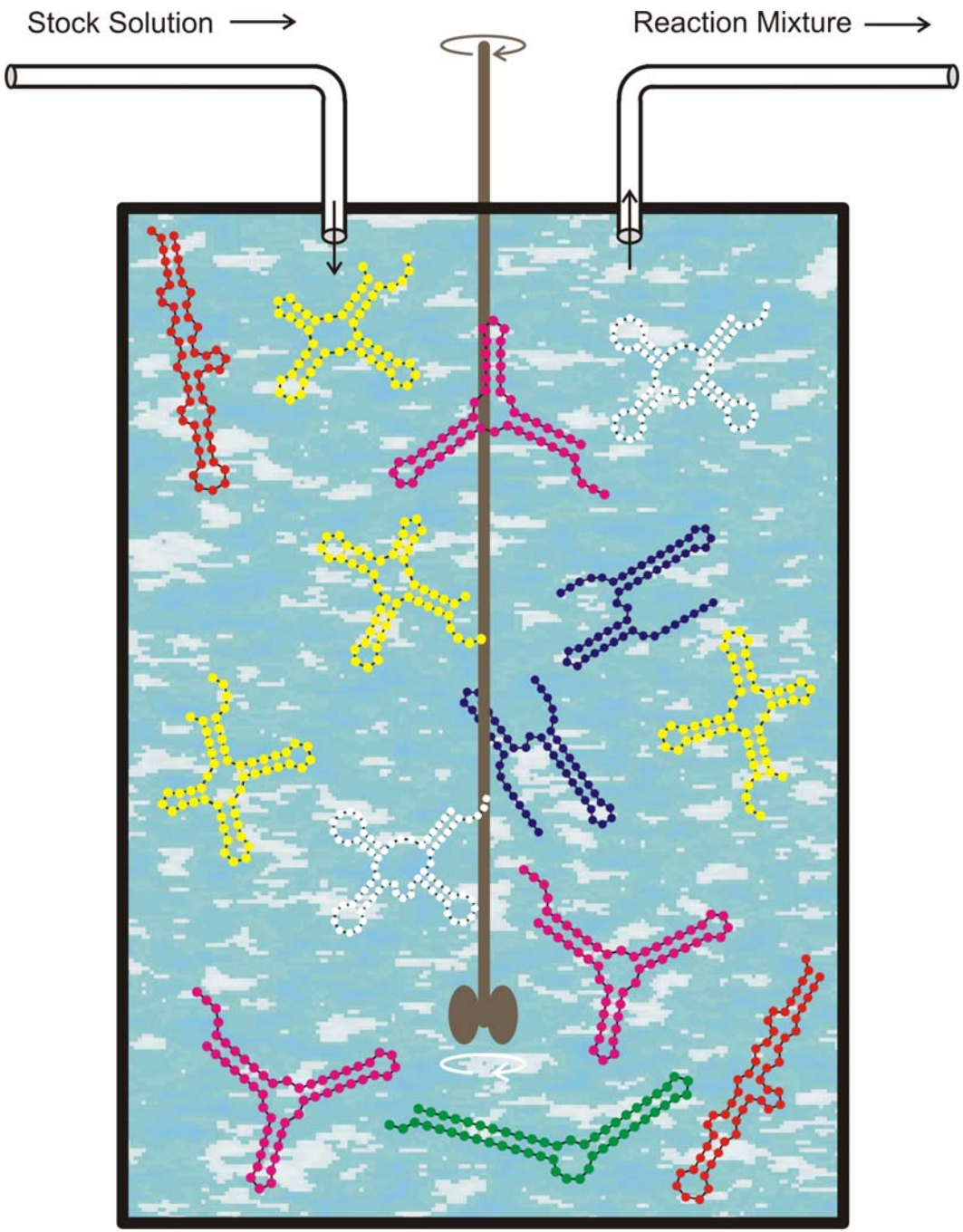
Institut für Theoretische Chemie, Universität Wien, Währingerstrasse 17, A-1090 Wien, Austria, Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA, and International Institute for Applied Systems Analysis (IIASA), A-2361 Laxenburg, Austria.



Structure of randomly chosen initial sequence



Phenylalanyl-tRNA as target structure



Replication rate constant

(Fitness):

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

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

Selection pressure:

The population size,

N = # RNA moleucles,

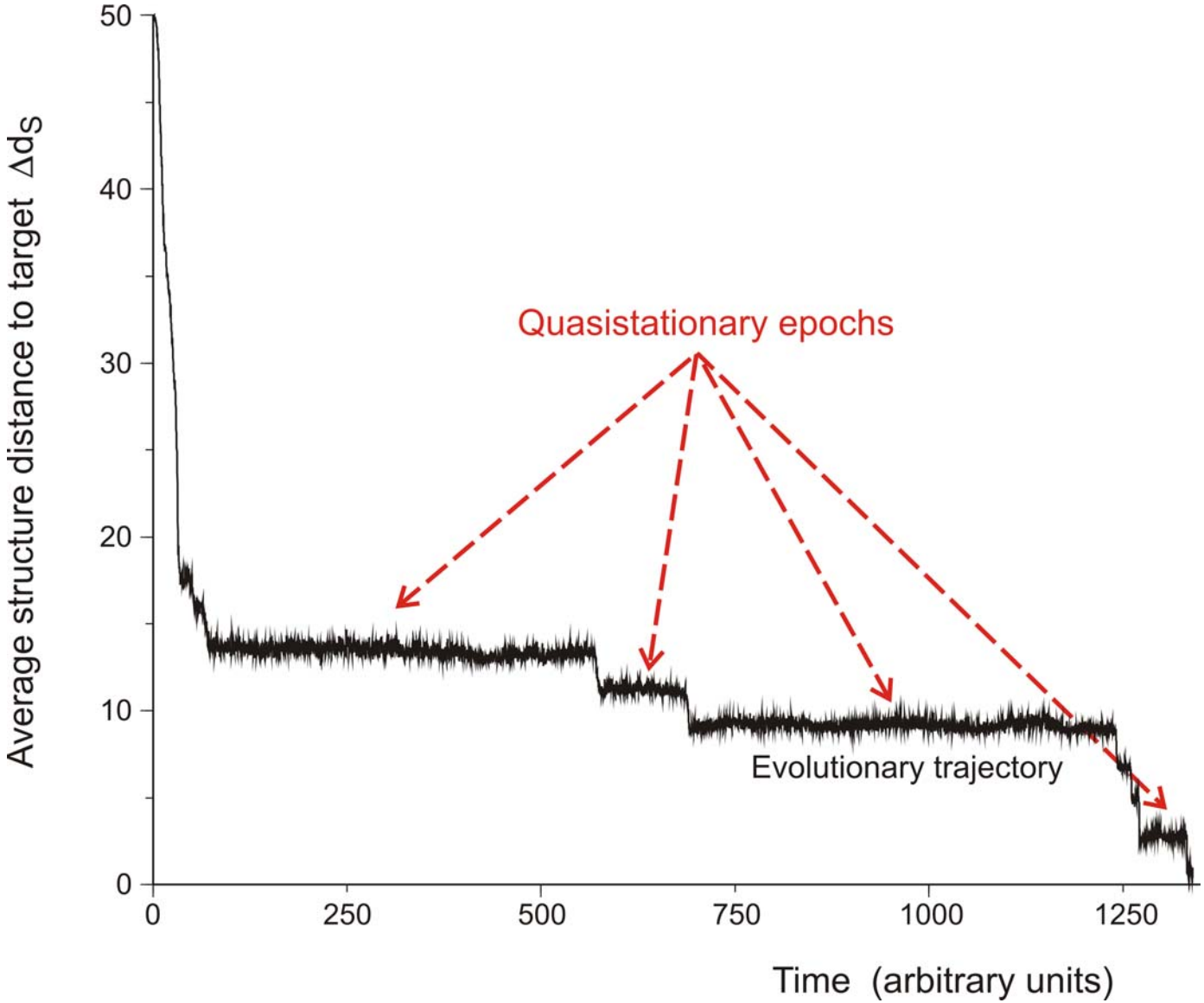
is determined by the flux:

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

Mutation rate:

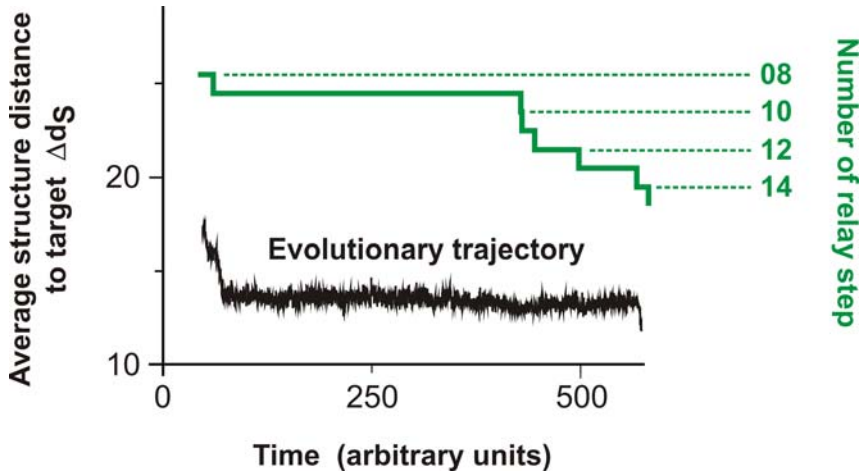
p = 0.001 / Nucleotide × Replication

The flow reactor as a device for studying the evolution of molecules ***in vitro*** and ***in silico***.



In silico optimization in the flow reactor: Evolutionary Trajectory

28 neutral point mutations during a long quasi-stationary epoch



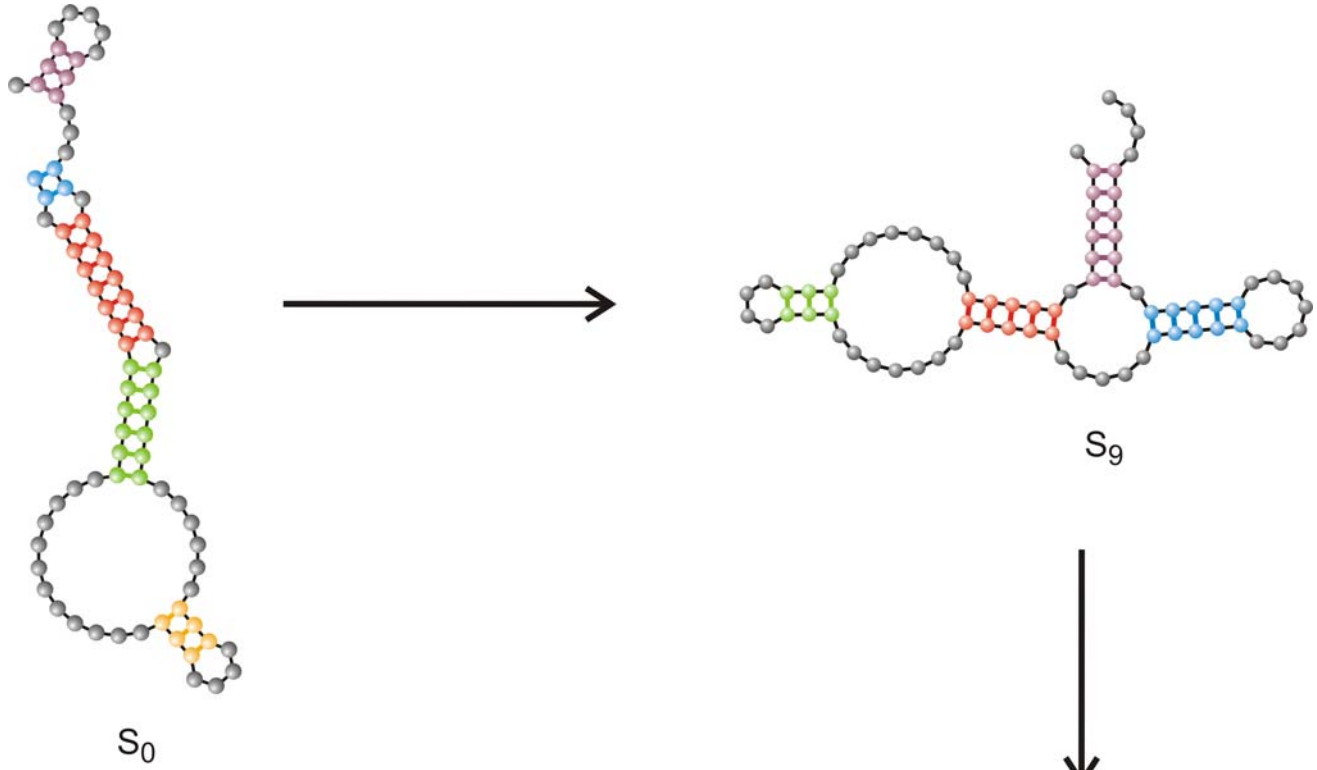

```
entry   GGUAUGGGCGUUGAAUAGUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUCCCGUACAGAA
8       .((((((((((((.........(((....)))).......)))))).....((((.......))))))))))....
exit    GGUAUGGGCGUUGAAUAAUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUCCCAUACAGAA
entry   GGUAUGGGCGUUGAAUAAUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUACCAUACAGAA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

Transition inducing point mutations change the molecular structure

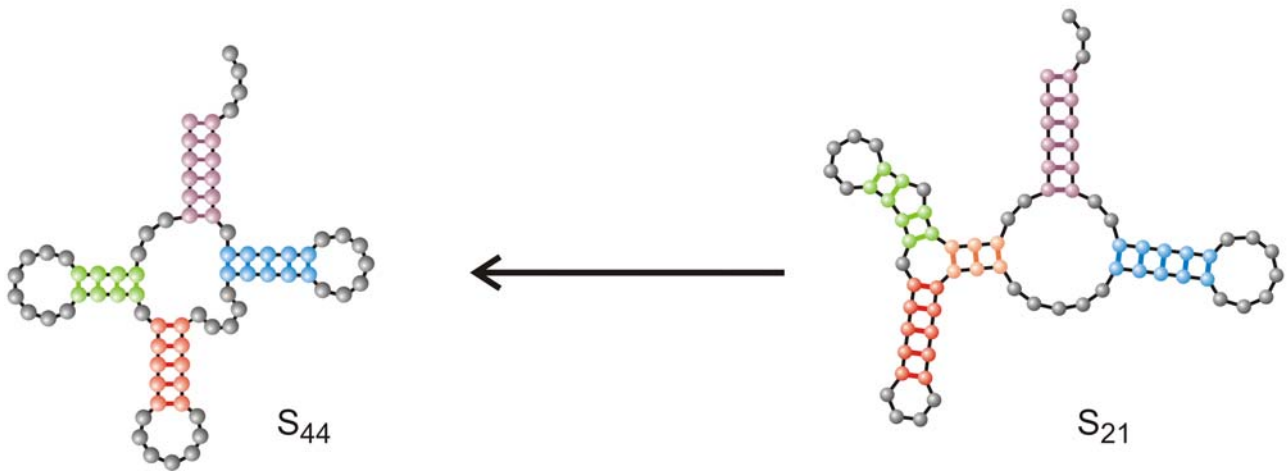
Neutral point mutations leave the molecular structure unchanged

Neutral genotype evolution during phenotypic stasis

Randomly chosen
initial structure



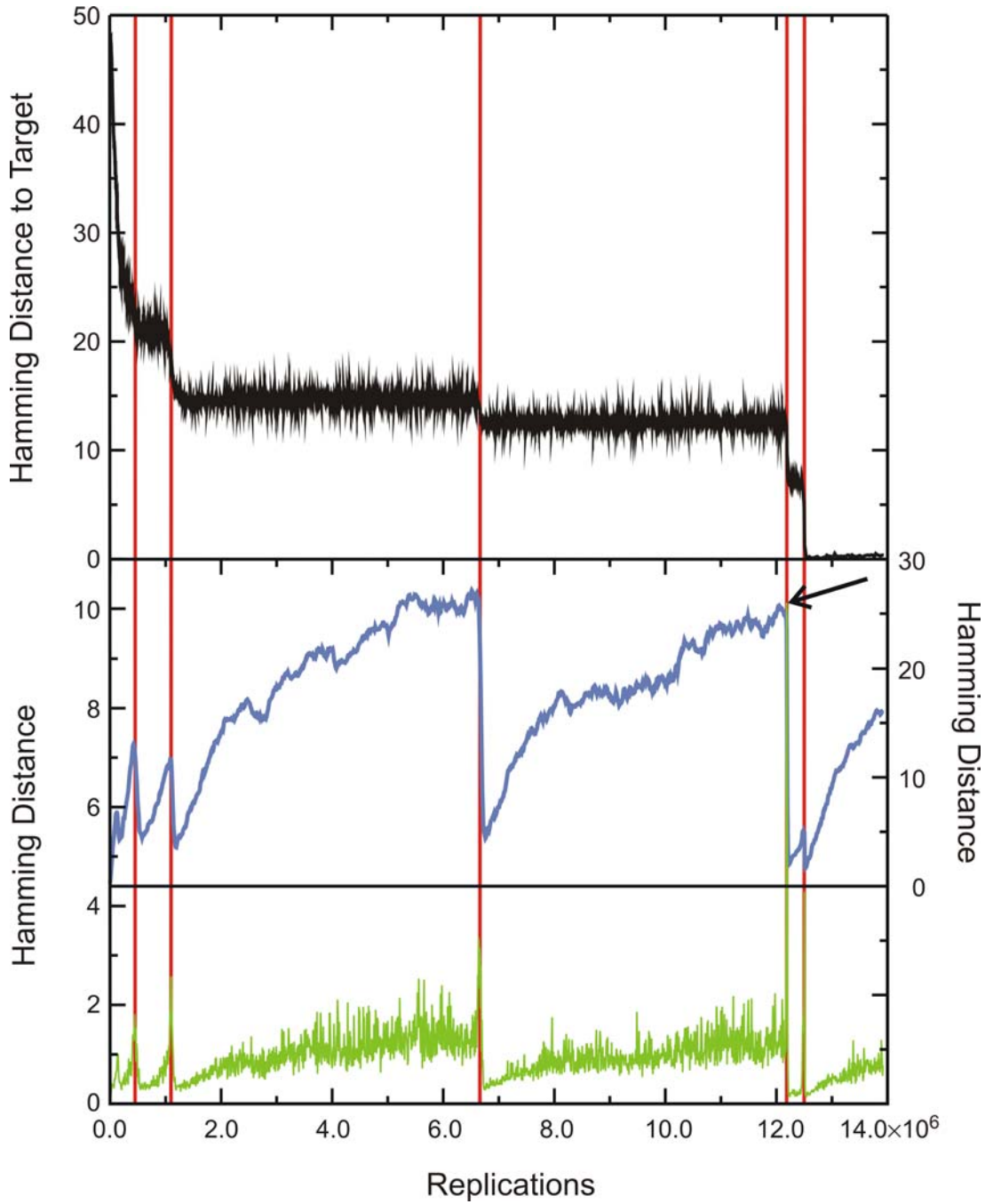
Phenylalanyl-tRNA
as target structure

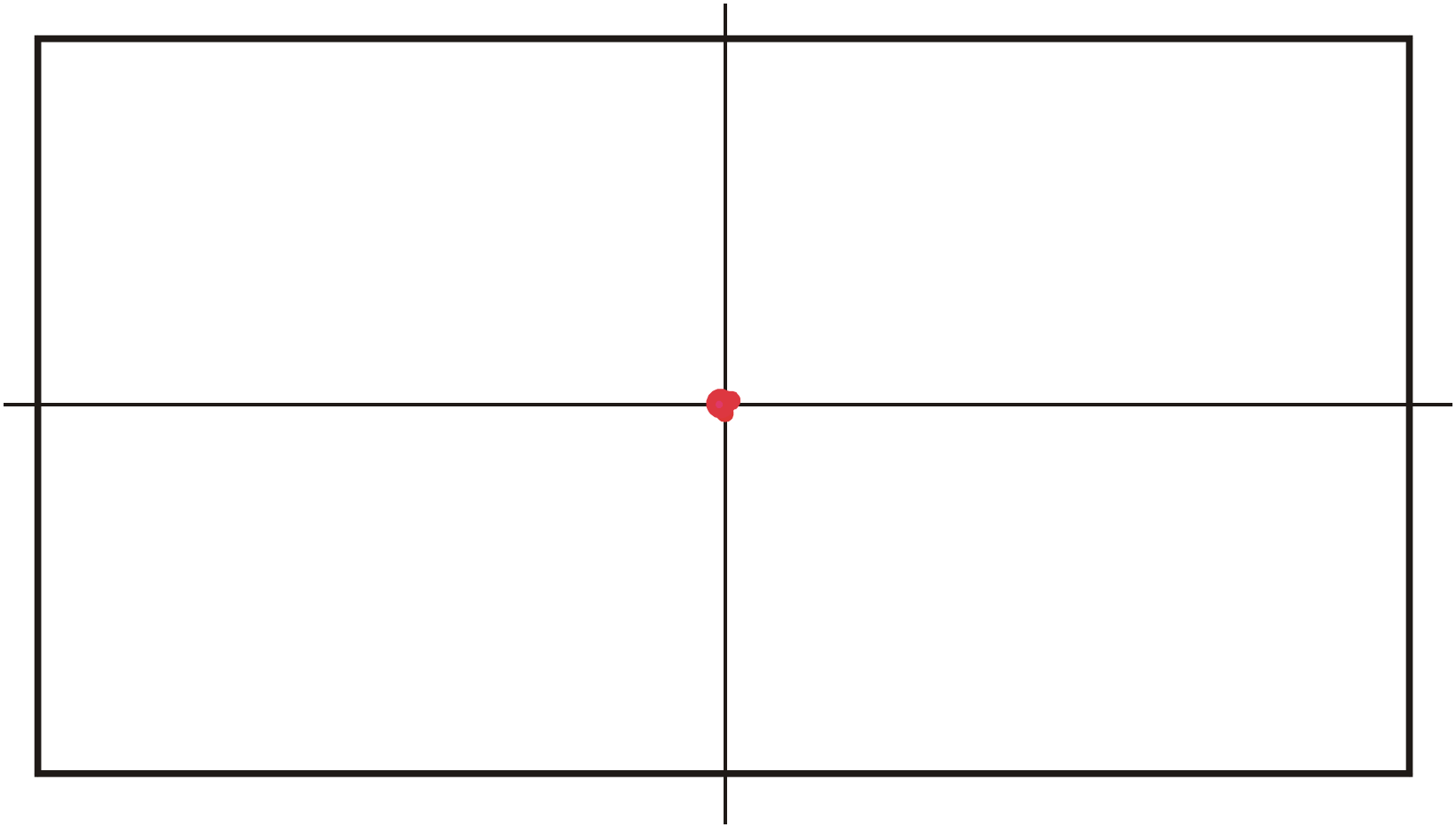


Evolutionary trajectory

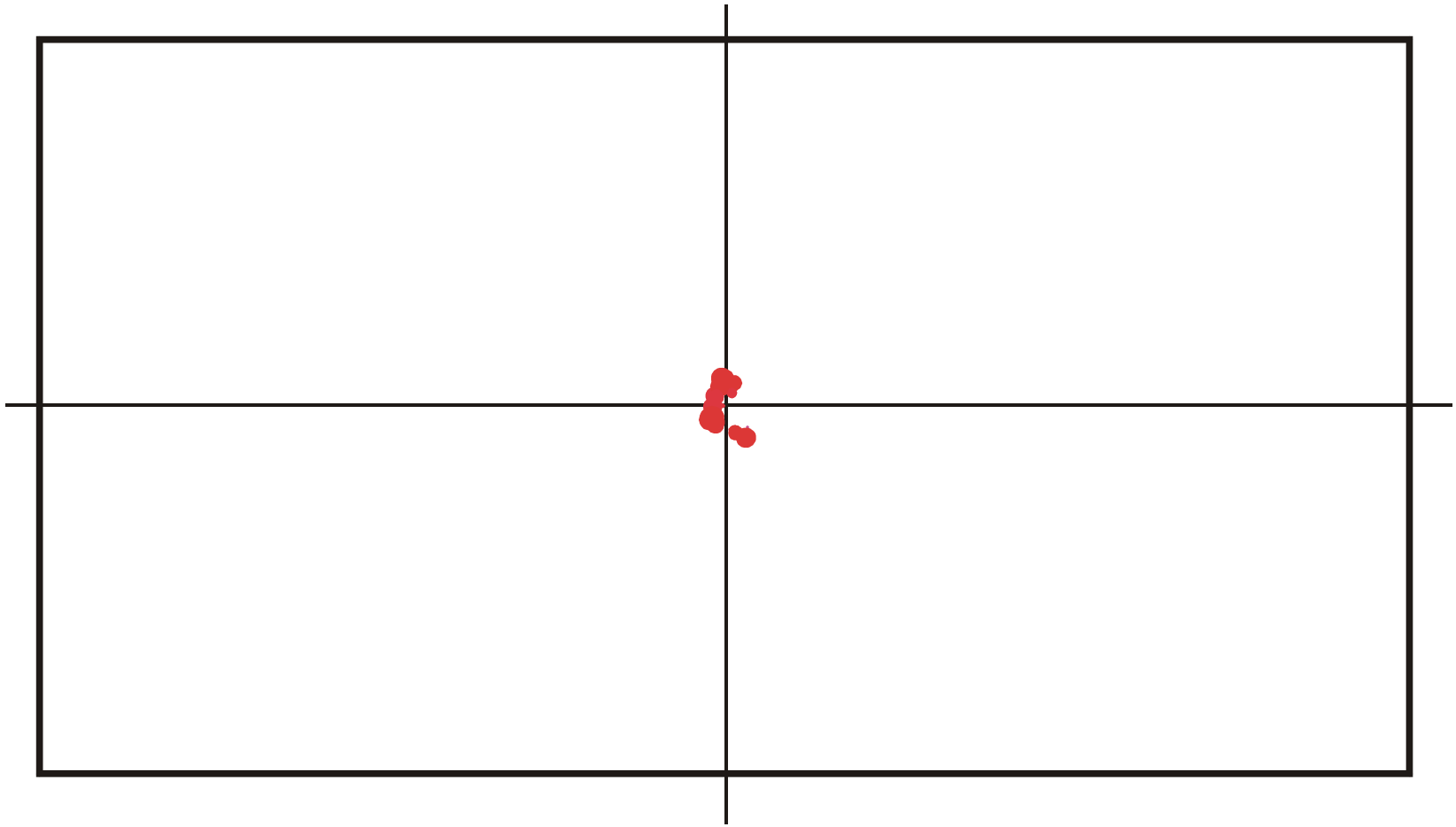
Spreading of the population on neutral networks

Drift of the population center in sequence space

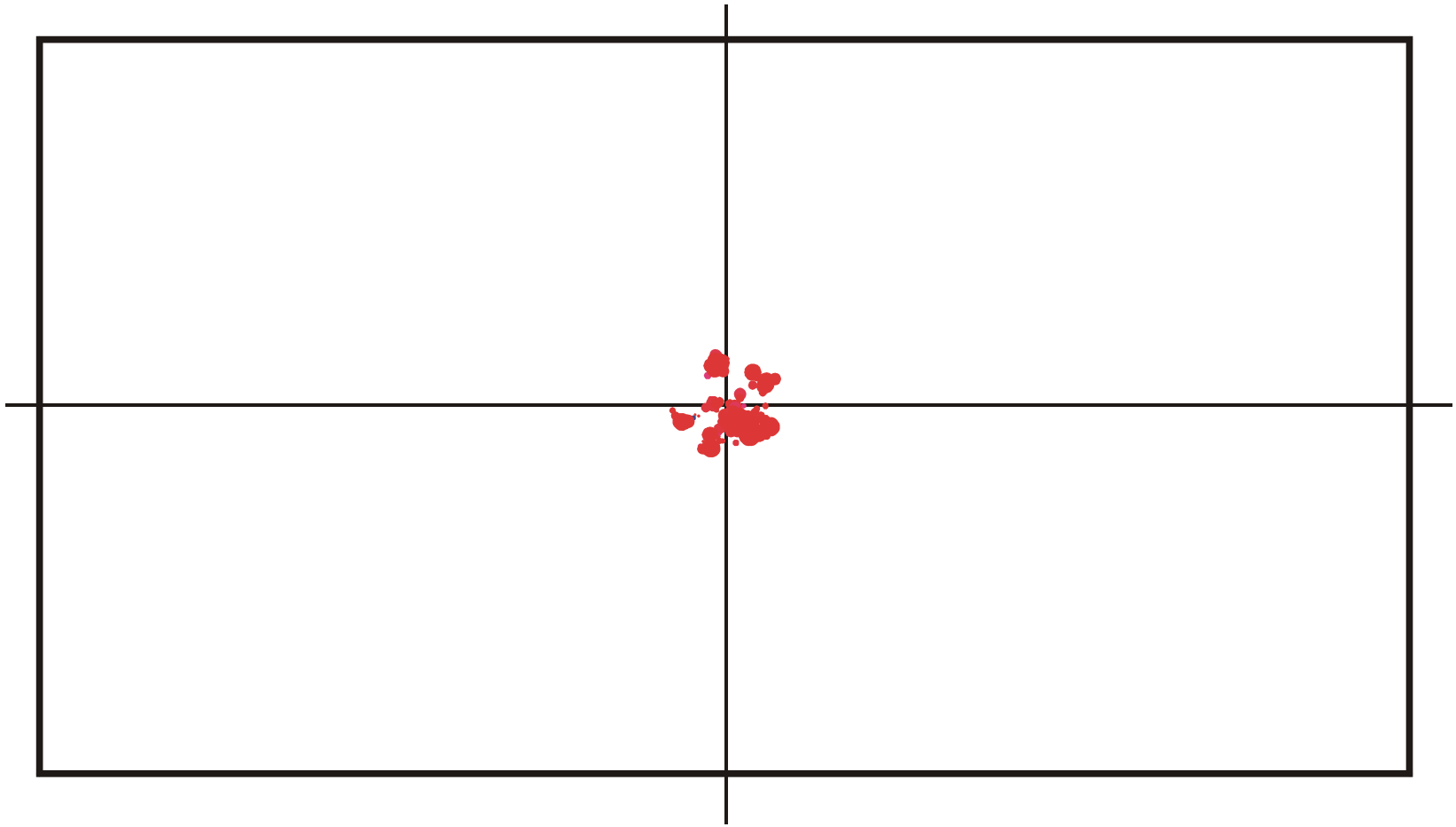




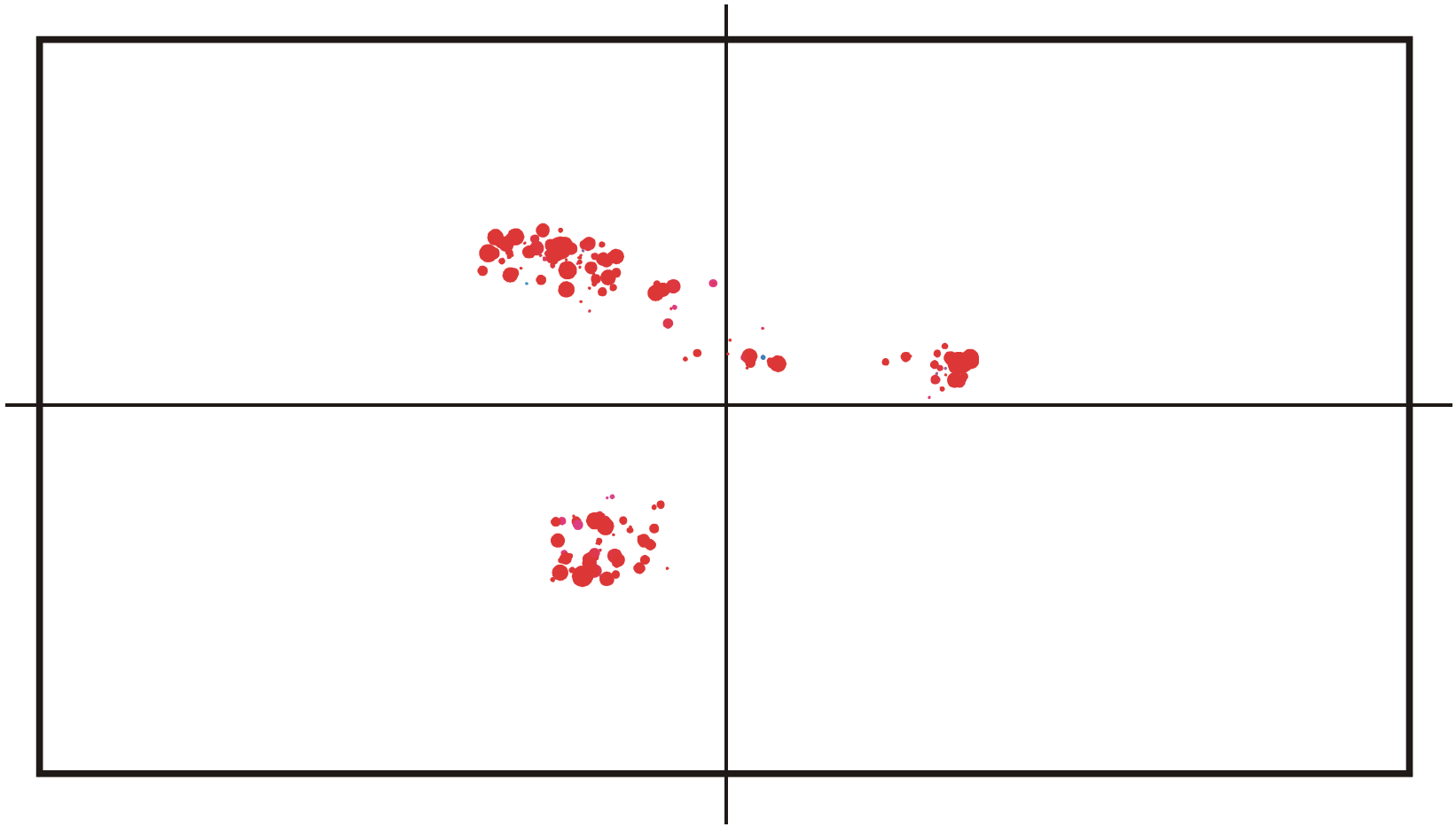
Spreading and evolution of a population on a neutral network: t = 150



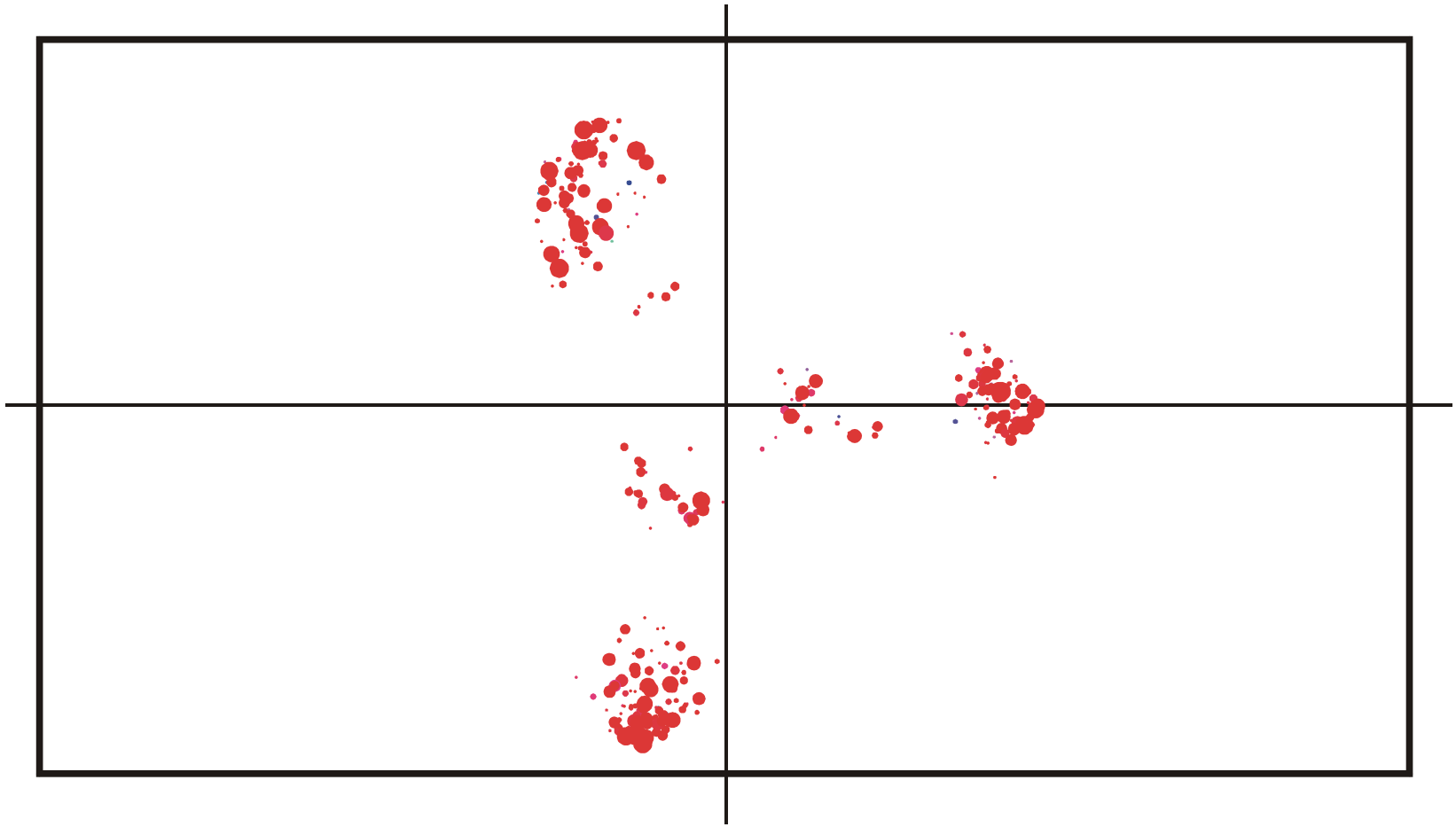
Spreading and evolution of a population on a neutral network : t = 170



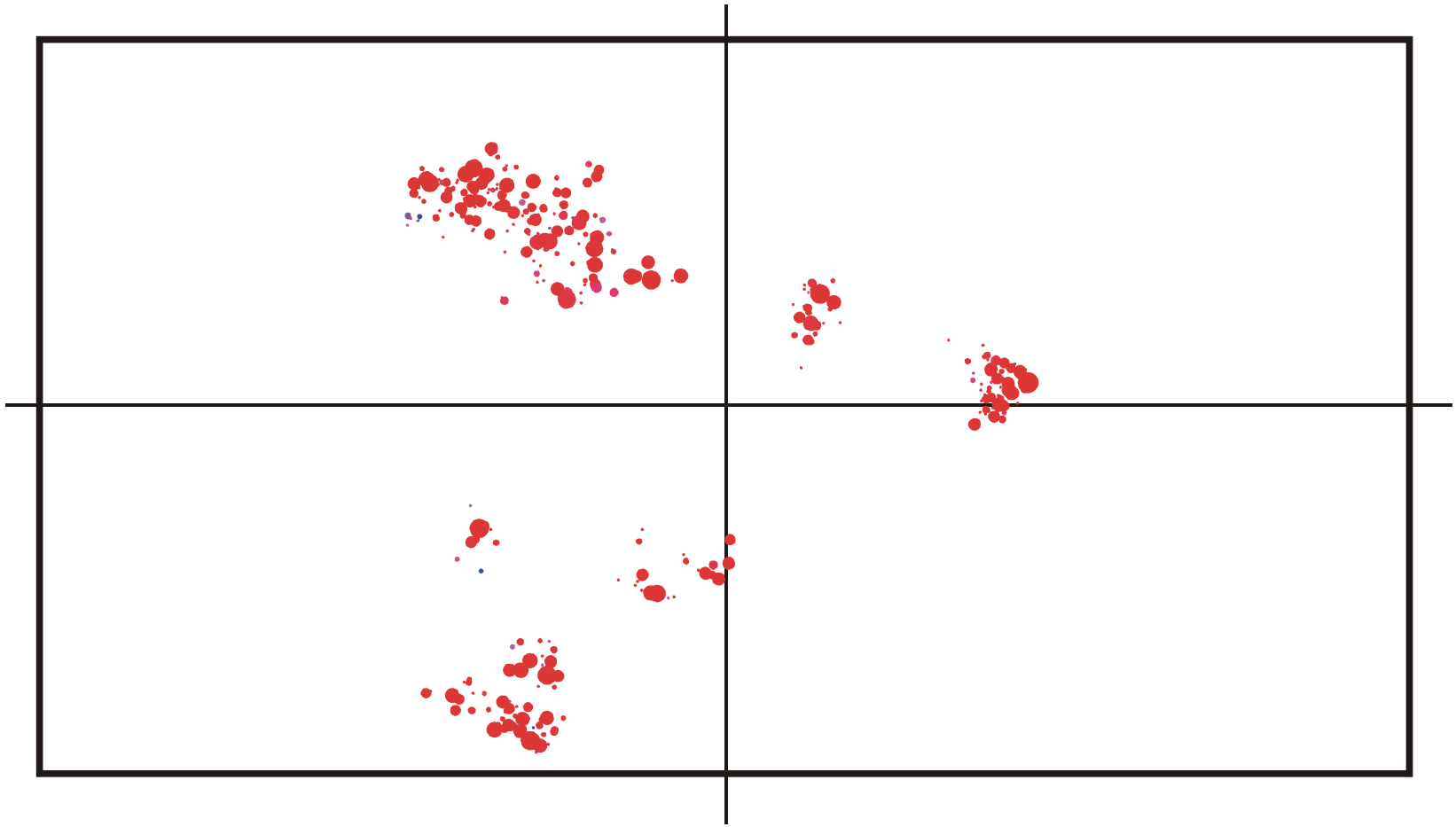
Spreading and evolution of a population on a neutral network : t = 200



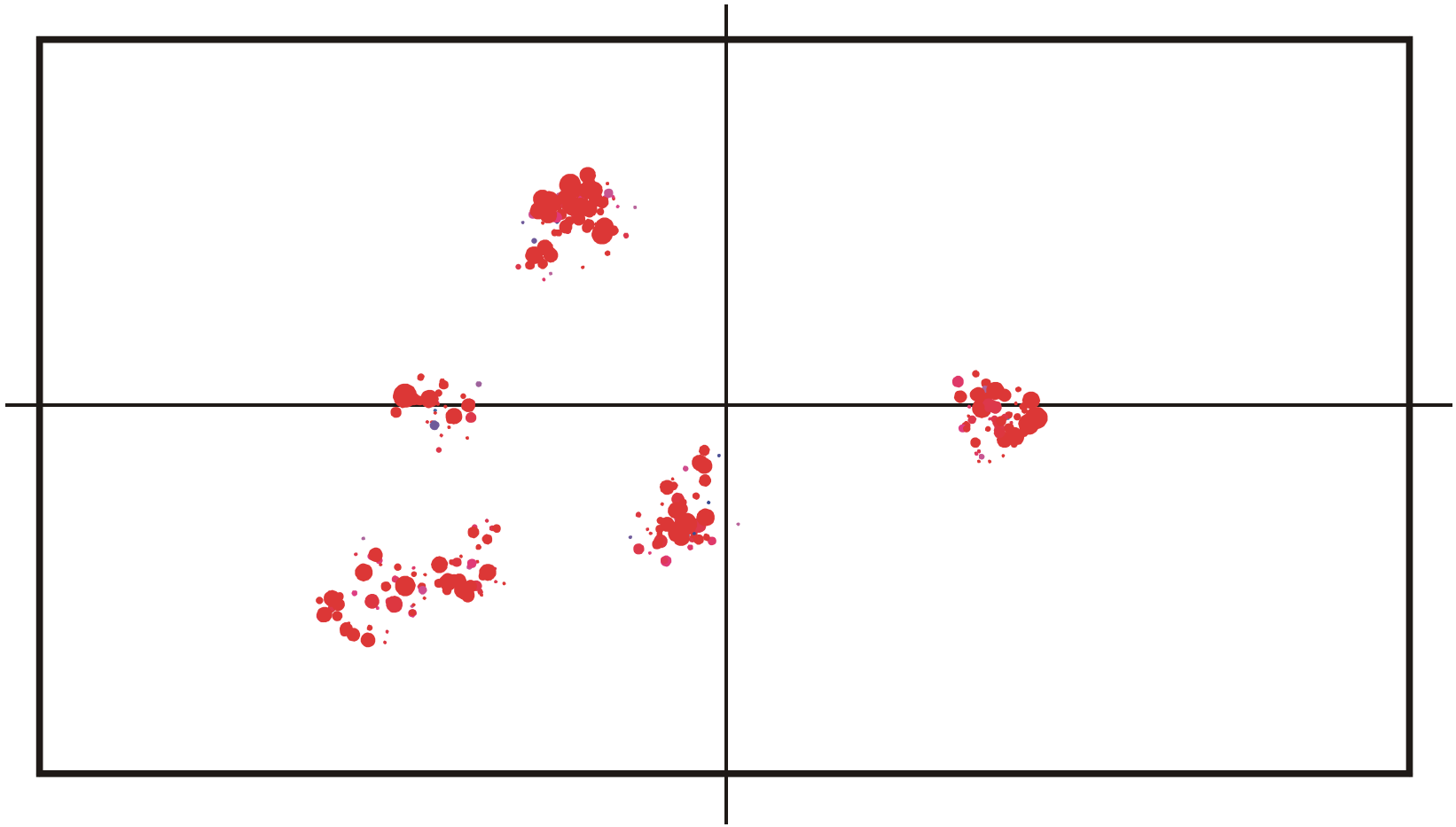
Spreading and evolution of a population on a neutral network : t = 350



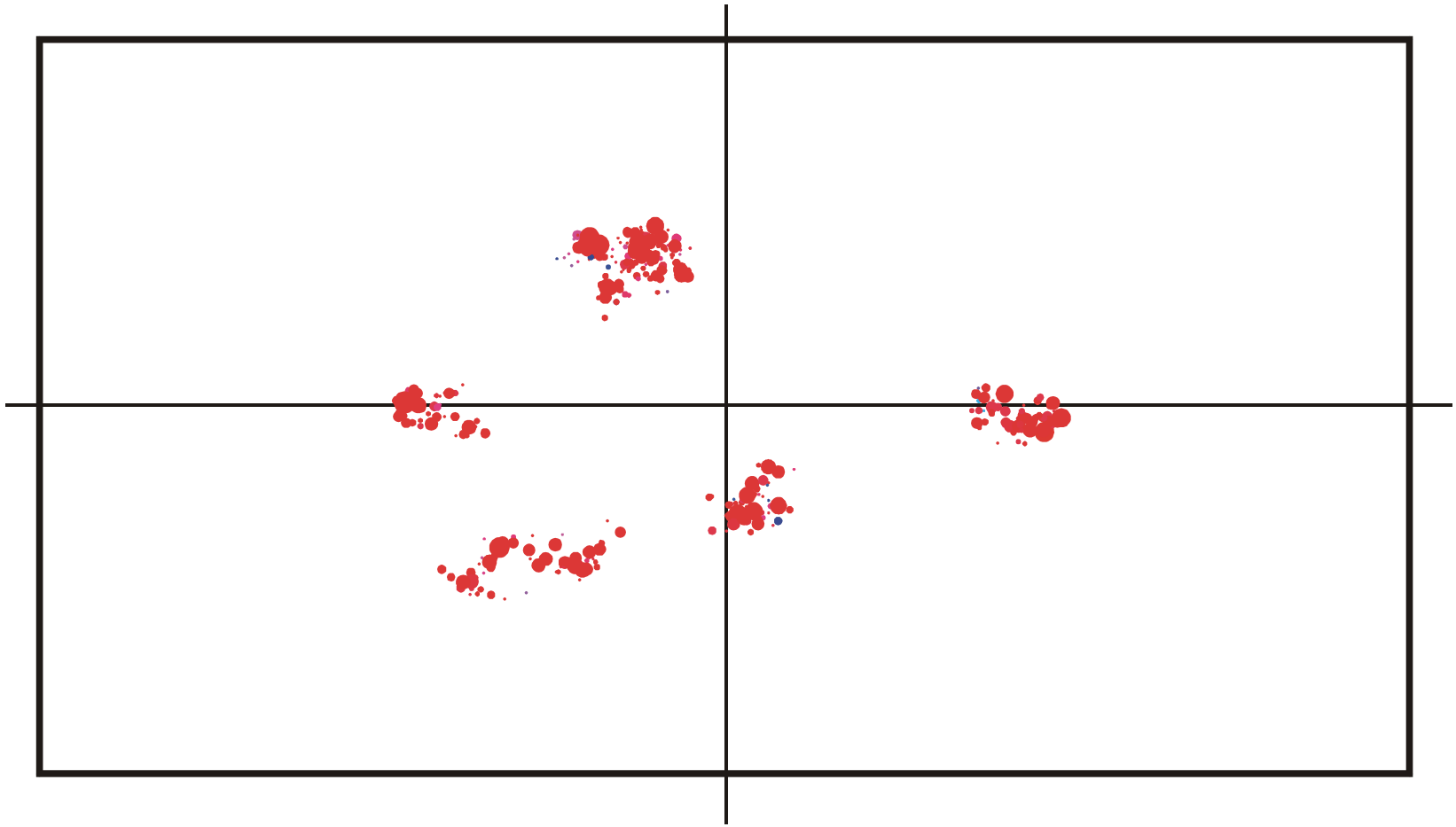
Spreading and evolution of a population on a neutral network : t = 500



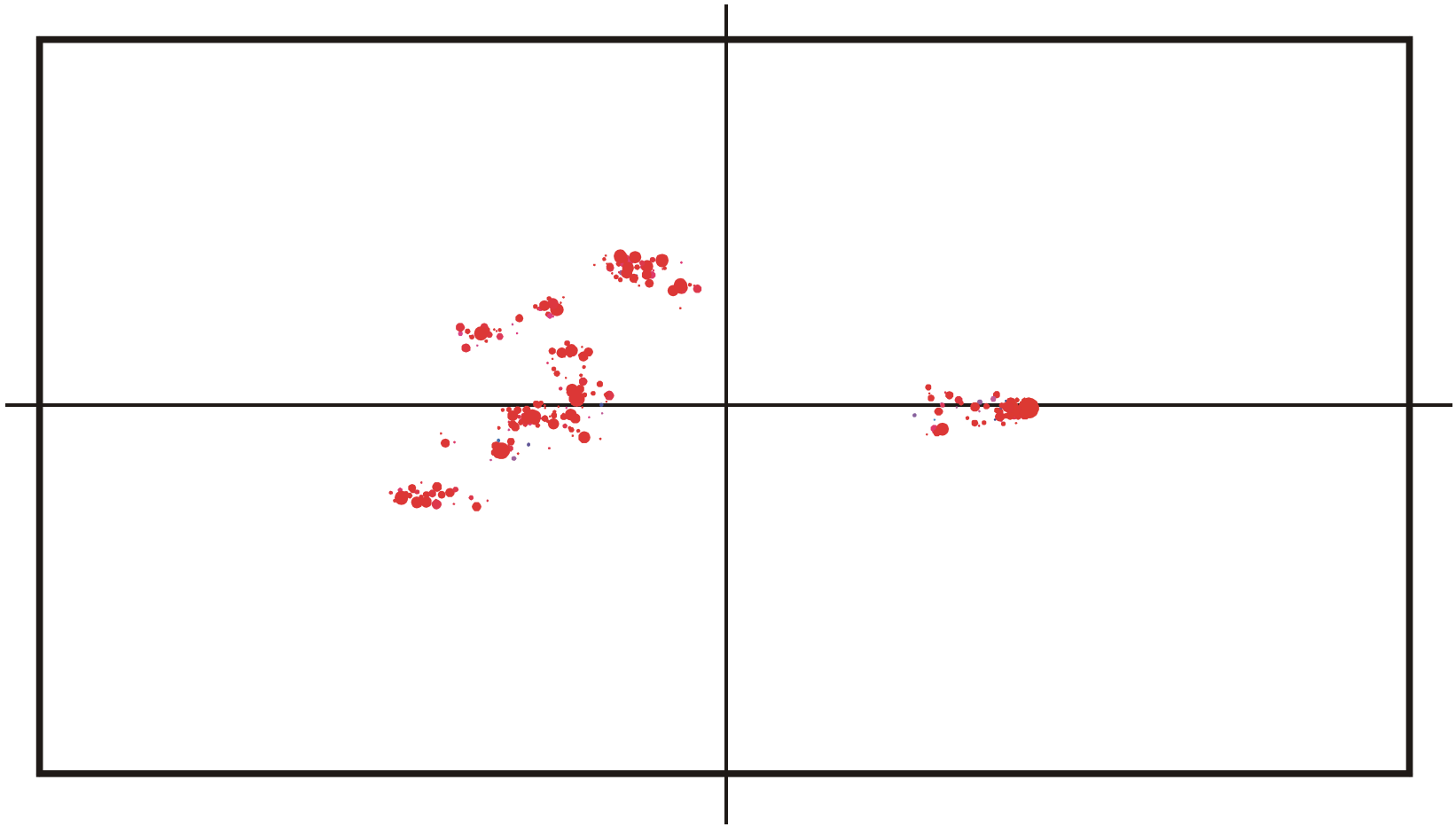
Spreading and evolution of a population on a neutral network : t = 650



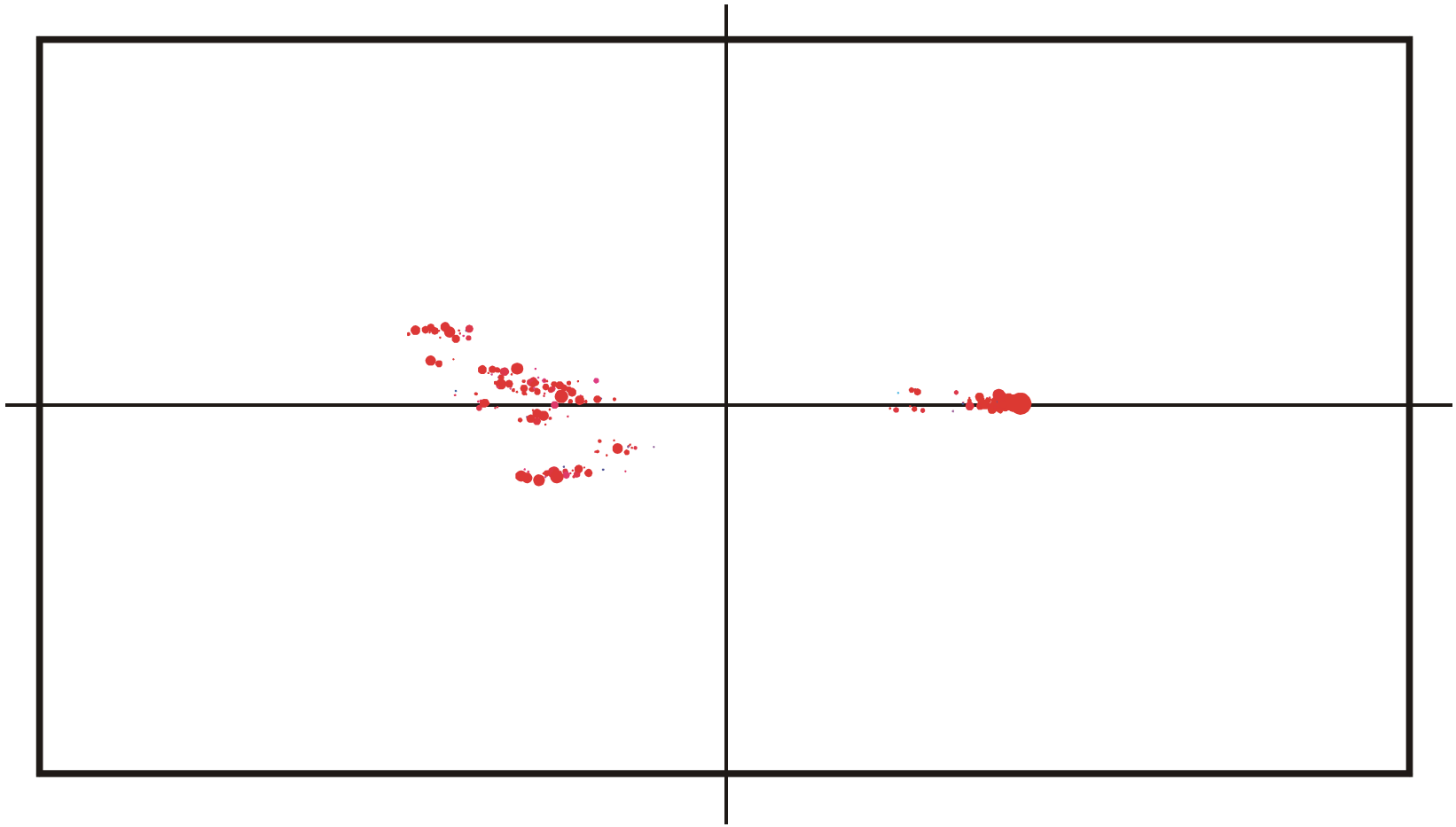
Spreading and evolution of a population on a neutral network : t = 820



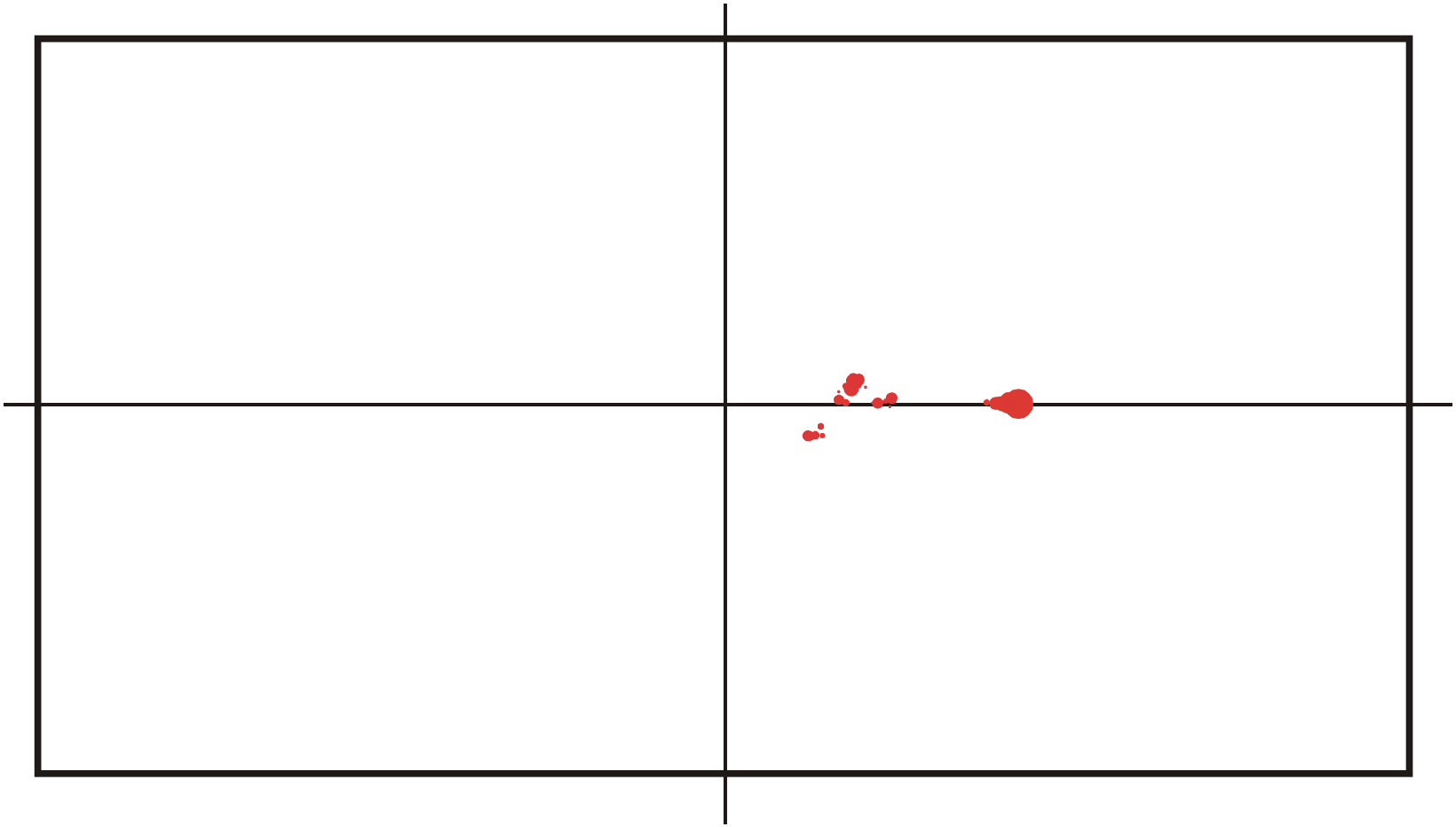
Spreading and evolution of a population on a neutral network : t = 825



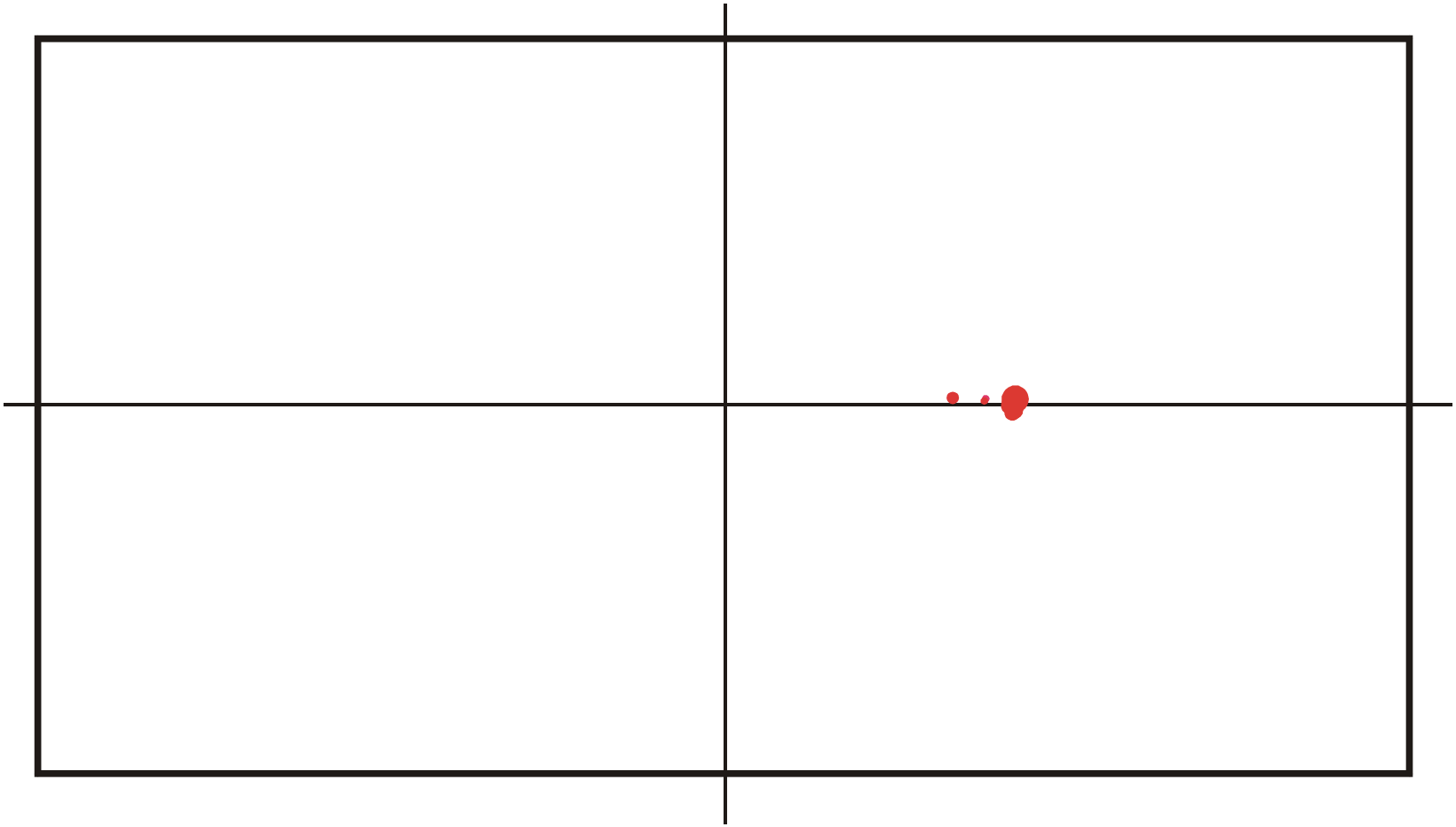
Spreading and evolution of a population on a neutral network : t = 830



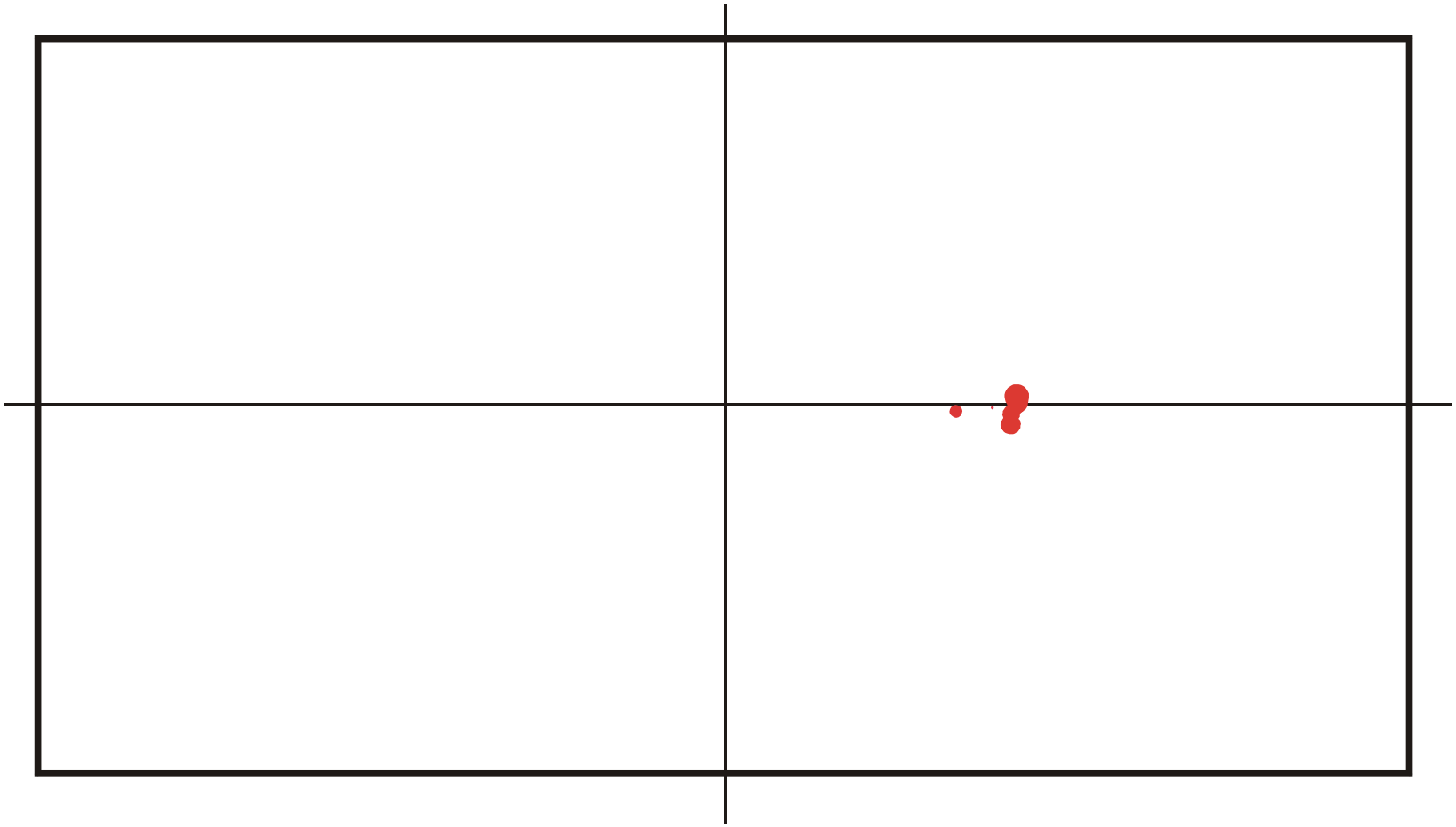
Spreading and evolution of a population on a neutral network : t = 835



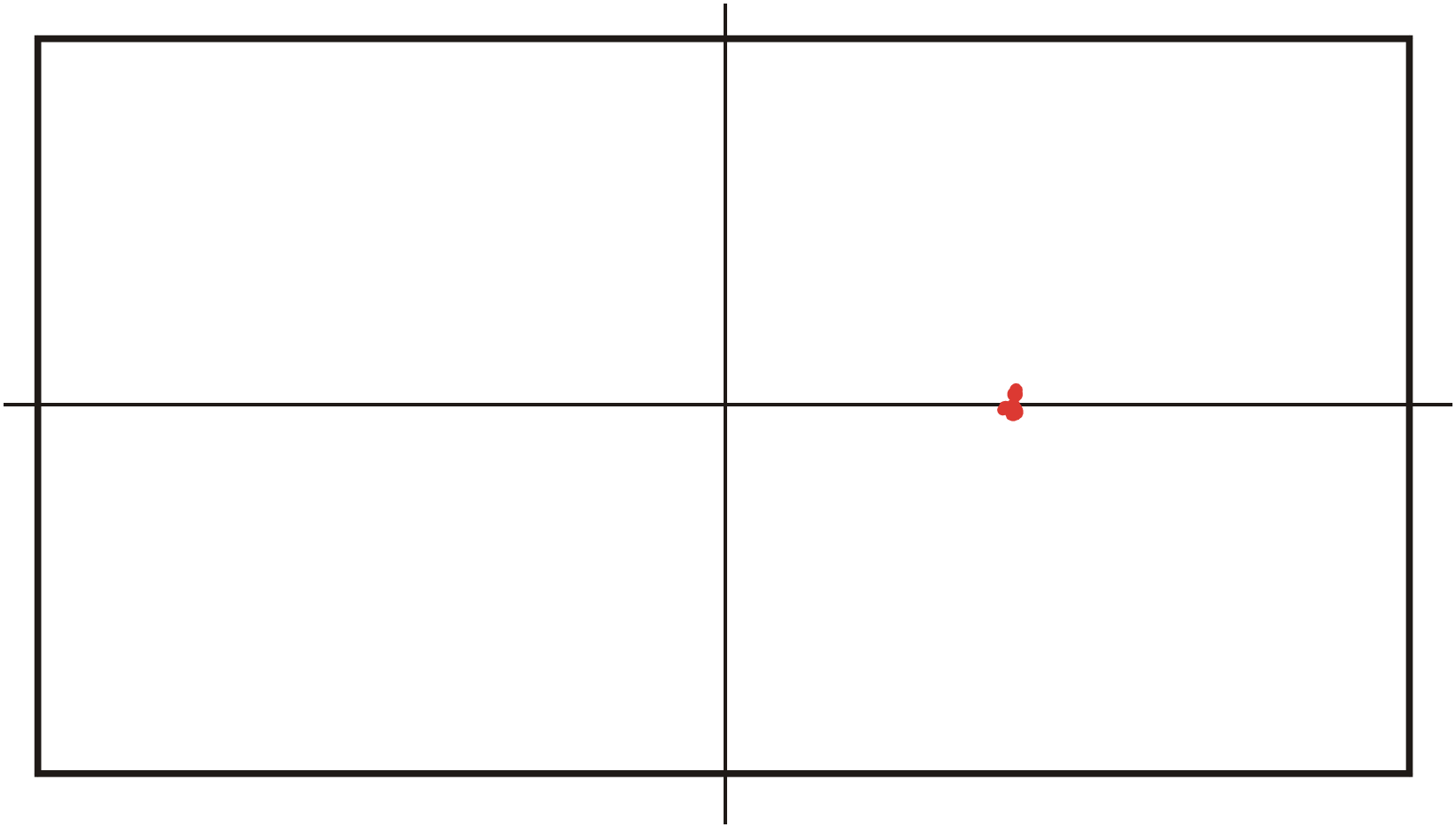
Spreading and evolution of a population on a neutral network : t = 840



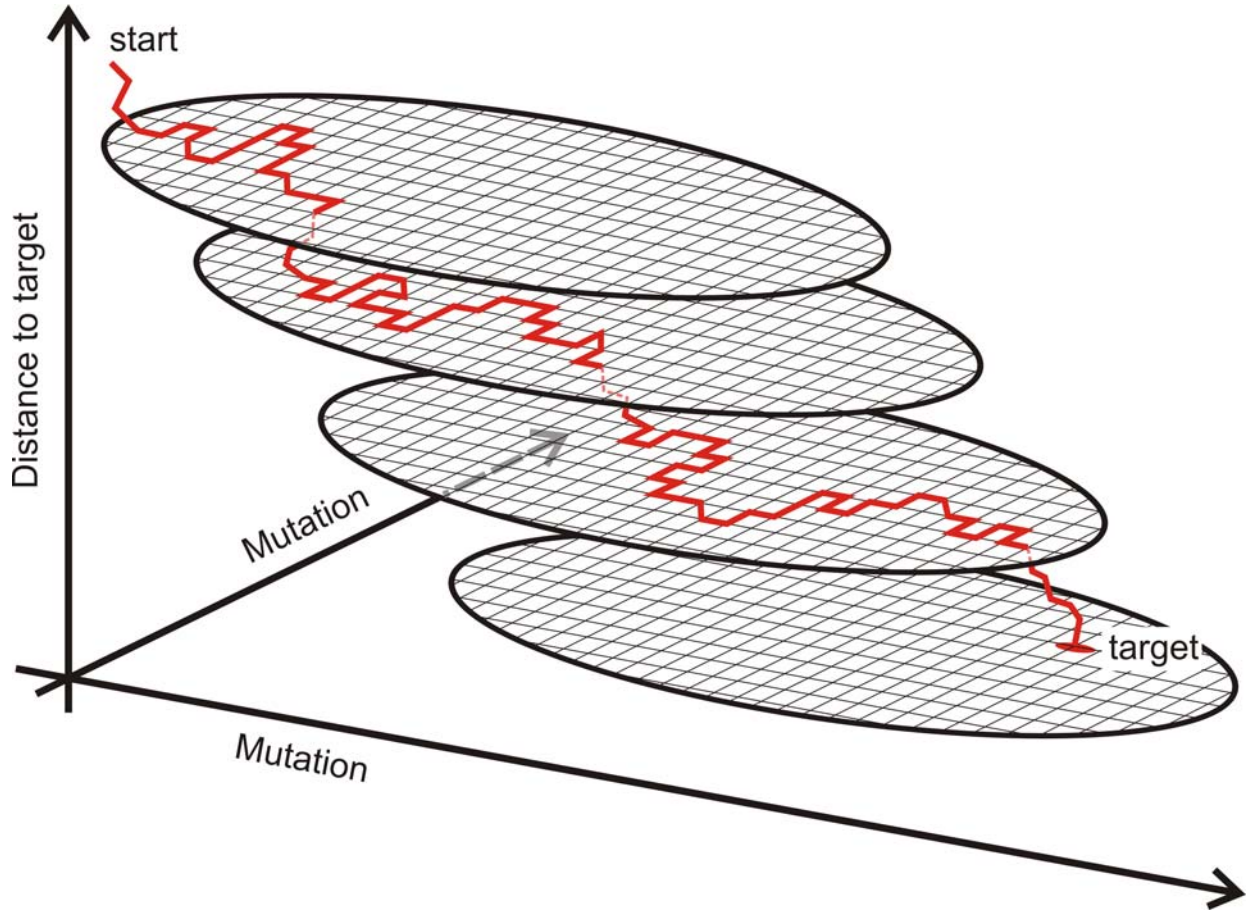
Spreading and evolution of a population on a neutral network : t = 845



Spreading and evolution of a population on a neutral network : t = 850



Spreading and evolution of a population on a neutral network : t = 855



A sketch of optimization on neutral networks

Table 8. Statistics of the optimization trajectories. The table shows the results of sampled evolutionary trajectories leading from a random initial structure, S_I, to the structure of $tRNA^{phe}$, S_T, as the target[a]. Simulations were performed with an algorithm introduced by Gillespie [55–57]. The time unit is here undefined. A mutation rate of $p = 0.001$ per site and replication were used. The mean and standard deviation were calculated under the assumption of a log-normal distribution that fits well the data of the simulations.

Alphabet	Population size, N	Number of runs, n_R	Real time from start to target		Number of replications [10^7]	
			Mean value	σ	Mean value	σ
AUGC	1 000	120	900	+1380 −542	1.2	+3.1 −0.9
	2 000	120	530	+880 −330	1.4	+3.6 −1.0
	3 000	1199	400	+670 −250	1.6	+4.4 −1.2
	10 000	120	190	+230 −100	2.3	+5.3 −1.6
	30 000	63	110	+97 −52	3.6	+6.7 −2.3
	100 000	18	62	+50 −28	–	–
GC	1 000	46	5160	+15700 −3890	–	–
	3 000	278	1910	+5180 −1460	7.4	+35.8 −6.1
	10 000	40	560	+1620 −420	–	–

[a] The structures S_I and S_T were used in the optimization:

S_I: ((.((((((((((((.....((((....)))).....))))))).)))))))..))...(((.....)))

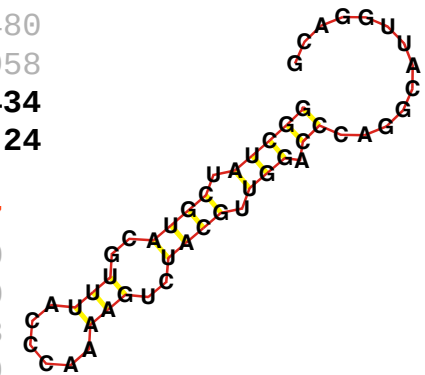
S_T: ((((((...((((........)))).((((.......)))))).....((((.......)))))).))))))....

Is the degree of neutrality in **GC** space much lower than in **AUGC** space ?

Statistics of RNA structure optimization: P. Schuster, Rep.Prog.Phys. 69:1419-1477, 2006

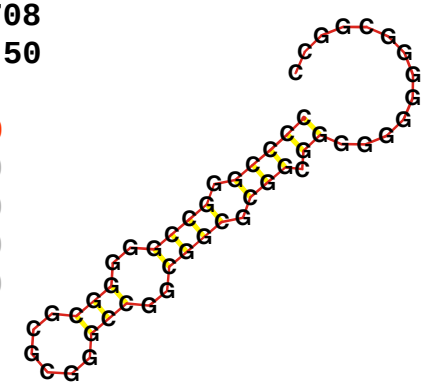
	Number	Mean Value	Variance	Std.Dev.
Total Hamming Distance:	150000	11.647973	23.140715	4.810480
Nonzero Hamming Distance:	99875	16.949991	30.757651	5.545958
Degree of Neutrality:	50125	**0.334167**	0.006961	**0.083434**
Number of Structures:	**1000**	**52.31**	85.30	**9.24**

1	(((((.((((..(((......))))..)))))..)))).....	**50125**	**0.334167**
2	..(((.((((..(((......))))..)))))..))).....	2856	0.019040
3	((((((((((..(((......))))..)))))))).)).....	2799	0.018660
4	(((((.((((..((((....))))..)))))..)))).....	2417	0.016113
5	(((((.((((.((((......))))..)))))..)))).....	2265	0.015100
6	(((((.(((((.(((......)))).)))))..)))).....	2233	0.014887



	Number	Mean Value	Variance	Std.Dev.
Total Hamming Distance:	50000	13.673580	10.795762	3.285691
Nonzero Hamming Distance:	45738	14.872054	10.821236	3.289565
Degree of Neutrality:	4262	**0.085240**	0.001824	**0.042708**
Number of Structures:	**1000**	**36.24**	6.27	**2.50**

1	(((((.((((..(((......))))..)))))..)))).....	**4262**	**0.085240**
2	((((((((((..(((......))))..)))))))).)).....	1940	0.038800
3	(((((.(((((.(((......)))).)))))..)))).....	1791	0.035820
4	(((((.((((.((((......))))..)))))..)))).....	1752	0.035040
5	(((((.((((..((((....))))..)))))..)))).....	1423	0.028460



Shadow – Surrounding of an RNA structure in shape space – **AUGC** and **GC** alphabet

Neutrality in evolution

Charles Darwin: *„ ... neutrality might exist ..."*

Motoo Kimura: *„ ... neutrality is unaviodable and represents the main reason for changes in genotypes and leads to molecular phylogeny ..."*

Current view: *„ ... neutrality is essential for successful optimization on rugged landscapes ..."*

Proposed view: *„ ... neutrality provides the genetic reservoir for functions in the rare and frequent mutation scenario ..."*

Outlook

Does understanding of life require more chemistry ?

Thinking in terms of processes rather than structures !

The difficulty to define the notion of „gene".

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

WHAT IS A GENE?

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

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

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

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

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

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

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

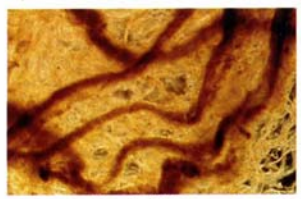
"All of that information seriously challenges our conventional definition of a gene," says molecular biologist Bing Ren at the University of California, San Diego. And the information challenge is about to get even tougher. Later this year, a glut of data will be released from the international Encyclopedia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

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

Under fire

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

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



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

P. PLAILLY/SPL

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

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

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

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

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

ENCODE stands for
ENCyclopedia Of DNA Elements.

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

Acknowledgement of support

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

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

Jubiläumsfonds der Österreichischen Nationalbank
Project No. Nat-7813

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

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

Österreichische Akademie der Wissenschaften

Siemens AG, Austria

Universität Wien and the Santa Fe Institute

Universität Wien

Coworkers

Universität Wien

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

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

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

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

Walter Fontana, Harvard Medical School, MA

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

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

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

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

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

INSTITUTE OF PHYSICS PUBLISHING REPORTS ON PROGRESS IN PHYSICS

Rep. Prog. Phys. **69** (2006) 1419–1477 doi:10.1088/0034-4885/69/5/R04

Prediction of RNA secondary structures: from theory to models and real molecules

Peter Schuster[1,2]

[1]Institut für Theoretische Chemie der Universität Wien, Währingerstraße 17, A-1090 Vienna, Austria
[2]The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA

E-mail: pks@tbi.univie.ac.at

Web-Page for further information:

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