The role of neutrality in molecular evolution

Novel variations of an old theme

Peter Schuster

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

Evolutionary Dynamics Program

Cambridge, MA, 07.11.2008

Web-Page for further information:

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

Prologue

The work on a molecular theory of evolution started 40 years ago .....

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.

1971

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 favorably with any other single entity or linked ensemble which does not contribute to its integrated function.
These requirements are crucial for a selection of the best adapted functionally linked ensemble and its evolutive optimization. Only hypercyclic organizations are able to fulfil these requirements. Non-cyclic linkages among the autonomous reproduction cycles, such as chains or branched, tree-like networks are devoid of such properties.
The mathematical methods used for proving these assertions are fixed-point, 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-

1977

Molecular Quasi-Species†

Manfred Eigen,* John McCaskill,
Max Planck Institut für biophysikalische Chemie, Am Fassberg, D 3400 Göttingen-Nikolausberg, BRD
and Peter Schuster*
Institut für theoretische Chemie und Strahlenchemie, der Universität Wien, Währinger Strasse 17, A-1090 Wien, Austria (Received: June 9, 1988)

The molecular quasi-species model describes the physicochemical organization of monomers into an ensemble of heteropolymers with combinatorial complexity by ongoing template polymerization. Polynucleotides belong to the simplest class of such molecules. The quasi-species itself represents the stationary distribution of macromolecular sequences maintained by chemical reactions effecting error-prone replication and by transport processes. It is obtained deterministically, by mass-action kinetics, as the dominant eigenvalue of a *value* matrix, **W**, which is derived directly from chemical rate coefficients, but it also exhibits stochastic features, being composed to a significant fraction of unique individual macromolecular sequences. The quasi-species model demonstrates how macromolecular information originates through specific nonequilibrium autocatalytic reactions and thus forms a bridge between reaction kinetics and molecular evolution. Selection and evolutionary optimization appear as new features in physical chemistry. Concentration bias in the production of mutants is a new concept in population genetics, relevant to frequently mutating populations, which is shown to greatly enhance the optimization properties. The present theory relates to asexually replicating ensembles, but this restriction is not essential. A sharp transition is exhibited between a drifting population of essentially random macromolecular sequences and a localized population of close relatives. This transition at a threshold error rate was found to depend on sequence lengths, distributions of selective values, and population sizes. It has been determined generically for complex landscapes and for special cases, and, it was shown to persist generically in the presence of nearly neutral mutants. Replication dynamics has much in common with the equilibrium statistics of complex spin systems: the error threshold is equivalent to a magnetic order–disorder transition. A rational function of the replication accuracy plays the role of temperature. Experimental data obtained from test-tube evolution of polynucleotides and from studies of natural virus populations support the quasi-species model. The error threshold seems to set a limit to the genome lengths of several classes of RNA viruses. In addition, the results are relevant even in eucaryotes where they contribute to the exon–intron debate.

1. Molecular Selection

Our knowledge of physical and chemical systems is, in a final analysis, based on models derived from repeatable experiments. While none of the classic and rather besieged list of properties rounded up to support the intuition of a distinction between the living and nonliving—metabolism, self-reproduction, irritability, and adaptability, for example—intrinsically limit the application of the scientific method, a determining role by unique or *individual* entities comes into conflict with the requirement of repeatability. Combinatorial variety, such as that in heteropolymers based on even very small numbers of different bases, even just two, readily provides numbers of different entities so enormous that neither consecutive nor parallel physical realization is possible. The physical chemistry of finite systems of such macromolecules must deal with both known regularities and the advent of unique copolymeric sequences. Normally this would present no difficulty in a statistical mechanical analysis of typical behavior, where rare events play no significant role, but with autocatalytic polymerization processes even unique single molecules may be amplified to determine the fate of the entire system. Potentially creative, self-organizing around unique events, the dynamics of this simplest living chemical system is invested with regularities that both allow and limit efficient adaptation. The quasi-species model is a study of these regularities.

The fundamental regularity in living organisms that has invited explanation is adaptation. Why are organisms so well fitted to their environments? At a more chemical level, why are enzymes optimal catalysts? Darwin's theory of natural selection has provided biologists with a framework for the answer to this question. The present model is constructed along Darwinian lines but in terms of specific macromolecules, chemical reactions, and physical processes that make the notion of survival of the fittest precise. Not only does the model give an understanding of the physical limitations of adaptation, but also it provides new insight into the role of chance in the process. For an understanding of the structure of this minimal chemical model it is first necessary to recall the conceptual basis of Darwin's theory.

Darwin recognized that new inheritable adaptive properties were not induced by the environment but arose independently in the production of offspring. Lasting adaptive changes in a population could only come about by natural selection of the heritable traits or *genotype* based on the full characteristics or *phenotype* relevant for producing offspring. A process of chance, i.e., uncorrelated with the developed phenotype, controls changes in the genotype from one generation to the next and generates the diversity necessary for selection. Three factors have probably prevented chemists from gaining a clear insight into these phenomena in the past, despite the discovery of the polymeric nature of the genotype (DNA): the complexity of a minimum replication phenotype, the problem of dealing with a huge number of variants, and the nonequilibrium nature of these ongoing processes.

The formulation of a tractable chemical model based on Darwin's principle may be understood in several steps:

1. The major constituents of the system have to be inherently self-reproductive. Only two classes of molecules are presently

† This is an abridged account of the quasi-species theory that has been submitted in comprehensive form to *Advances in Chemical Physics*.[1]

(1) Eigen, M.; McCaskill, J. S.; Schuster, P. *Adv. Chem. Phys.*, in press.

0022-3654/88/2092-6881$01.50/0

1988

Chemical kinetics of molecular evolution

Error Thresholds for Quasispecies on Dynamic Fitness Landscapes

Martin Nilsson
Institute of Theoretical Physics, Chalmers University of Technology and University of Gothenburg, Gothenburg, Sweden

Nigel Snoad
Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501
and The Australian National University, ACT 0200, Australia†
(Received 29 March 1999)

In this paper we investigate error thresholds on dynamic fitness landscapes. We show that there exists both a lower and an upper threshold, representing limits to the copying fidelity of simple replicators. The lower bound can be expressed as a correction term to the error threshold present on a static landscape. The upper error threshold is a new limit that only exists on dynamic fitness landscapes. We also show that for long genomes and/or highly dynamic fitness landscapes there exists a lower bound on the selection pressure required for the effective selection of genomes with superior fitness independent of mutation rates, i.e. there are distinct nontrivial limits to evolutionary parameters in dynamic environments.

PACS numbers: 87.23.Kg, 87.10.+e, 87.15.Aa

Maternal Effects in Molecular Evolution

Claus O. Wilke*
Digital Life Laboratory, Mail Code 136-93, Pasadena, California 91125
(Received 27 June 2001; published 31 January 2002)

We introduce a model of molecular evolution in which the fitness of an individual depends both on its own and on the parent's genotype. The model can be solved by means of a nonlinear mapping onto the standard quasispecies model. The dependency on the parental genotypes cancels from the mean fitness, but not from the individual sequence concentrations. For finite populations, the position of the error threshold is very sensitive to the influence from parent genotypes. In addition to biological applications, our model is important for understanding the dynamics of self-replicating computer programs.

DOI: 10.1103/PhysRevLett.88.078101 PACS numbers: 87.23.Kg

Quasispecies theory for multiple-peak fitness landscapes

David B. Saakian,[1,2] E. Muñoz,[3] Chin-Kun Hu,[1] and M. W. Deem[3]
[1]*Institute of Physics, Academia Sinica, Nankang, Taipei 11529, Taiwan*
[2]*Yerevan Physics Institute, Alikhanian Brothers St. 2, Yerevan 375036, Armenia*
[3]*Department of Physics and Astronomy, Rice University, Houston, Texas 77005-1892, USA*
(Received 15 September 2005; revised manuscript received 13 December 2005; published 11 April 2006)

We use a path integral representation to solve the Eigen and Crow-Kimura molecular evolution models for the case of multiple fitness peaks with arbitrary fitness and degradation functions. In the general case, we find that the solution to these molecular evolution models can be written as the optimum of a fitness function, with constraints enforced by Lagrange multipliers and with a term accounting for the entropy of the spreading population in sequence space. The results for the Eigen model are applied to consider virus or cancer proliferation under the control of drugs or the immune system.

DOI: 10.1103/PhysRevE.73.041913 PACS number(s): 87.23.Kg, 02.50.−r, 87.10.+e, 87.15.Aa

Phase Diagrams of Quasispecies Theory with Recombination and Horizontal Gene Transfer

J.-M. Park[1,2] and M. W. Deem[1]
[1]*Department of Physics & Astronomy and Department of Bioengineering, Rice University, Houston, Texas 77005-1892, USA*
[2]*Department of Physics, The Catholic University of Korea, Bucheon, 420-743, Korea*
(Received 9 October 2006; published 29 January 2007)

We consider how transfer of genetic information between individuals influences the phase diagram and mean fitness of both the Eigen and the parallel, or Crow-Kimura, models of evolution. In the absence of genetic transfer, these physical models of evolution consider the replication and point mutation of the genomes of independent individuals in a large population. A phase transition occurs, such that below a critical mutation rate an identifiable quasispecies forms. We show how transfer of genetic information changes the phase diagram and mean fitness and introduces metastability in quasispecies theory, via an analytic field theoretic mapping.

DOI: 10.1103/PhysRevLett.98.058101 PACS numbers: 87.23.Kg, 87.15.Aa

PHYSICAL REVIEW E **75**, 061109 (2007)

Emergence of order in selection-mutation dynamics

Christoph Marx, Harald A. Posch,* and Walter Thirring†
Faculty of Physics, Universität Wien, Boltzmanngasse 5, A-1090 Wien, Austria
(Received 7 March 2007; published 8 June 2007)

We characterize the time evolution of a d-dimensional probability distribution by the value of its final entropy. If it is near the maximally possible value we call the evolution mixing, if it is near zero we say it is purifying. The evolution is determined by the simplest nonlinear equation and contains a $d \times d$ matrix as input. Since we are not interested in a particular evolution but in the general features of evolutions of this type, we take the matrix elements as uniformly distributed random numbers between zero and some specified upper bound. Computer simulations show how the final entropies are distributed over this field of random numbers. The result is that the distribution crowds at the maximum entropy, if the upper bound is unity. If we restrict the dynamical matrices to certain regions in matrix space, to diagonal or triangular matrices, for instance, then the entropy distribution is maximal near zero, and the dynamics typically becomes purifying.

DOI: 10.1103/PhysRevE.75.061109 PACS number(s): 05.20.−y, 87.23.Kg, 05.45.Pq, 87.10.+e

PHYSICAL REVIEW E **76**, 041133 (2007)

Emergence of order in quantum extensions of the classical quasispecies evolution

Heide Narnhofer,* Harald A. Posch,† and Walter Thirring‡
Faculty of Physics, Universität Wien, Boltzmanngasse 5, A-1090 Wien, Austria
(Received 12 June 2007; published 24 October 2007)

We study evolution equations which model selection and mutation within the framework of quantum mechanics. The main question is to what extent order is achieved for an ensemble of typical systems. As an indicator for mixing or purification, a quadratic entropy is used which assumes values between zero for pure states and $(d-1)/d$ for fully mixed states. Here, d is the dimension. Whereas the classical counterpart, the quasispecies dynamics, has previously been found to be predominantly mixing, the quantum quasispecies (QS) evolution surprisingly is found to be strictly purifying for all dimensions. This is also typically true for an alternative formulation (AQS) of this quantum mechanical flow. We compare this also to analogous results for the Lindblad evolution. Although the latter may be viewed as a simple linear superposition of the purifying QS and AQS evolutions, it is found to be predominantly mixing. The reason for this behavior may be explained by the fact that the two subprocesses by themselves converge to different pure states, such that the combined process is mixing. These results also apply to high-dimensional systems.

DOI: 10.1103/PhysRevE.76.041133 PACS number(s): 05.30.−d, 87.23.Kg, 04.20.Ha, 87.10.+e

What is neutrality ?

Selective neutrality =
= several genotypes having the same fitness.

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

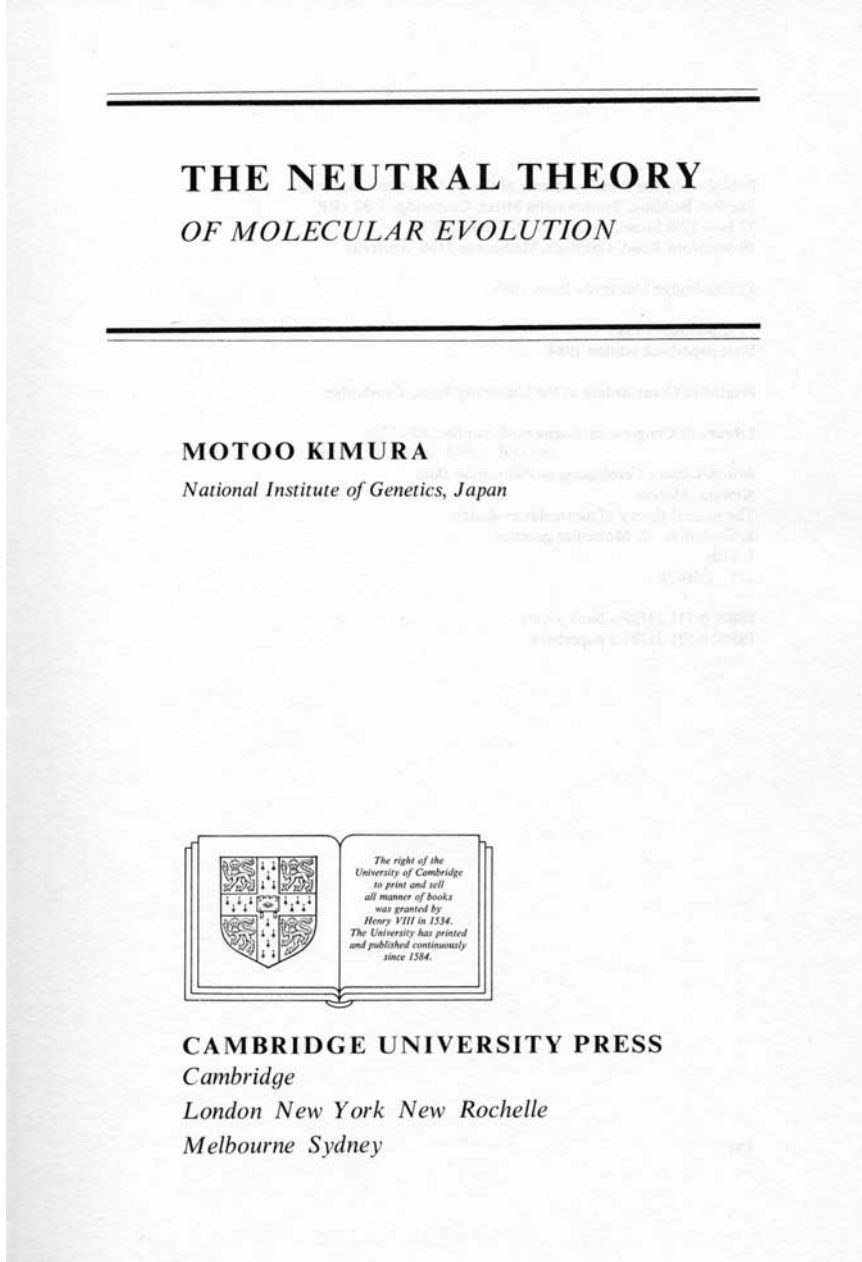
Charles Darwin. *The Origin of Species*. Sixth edition. John Murray. London: 1872

MOTOO KIMURA

Motoo Kimura's 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

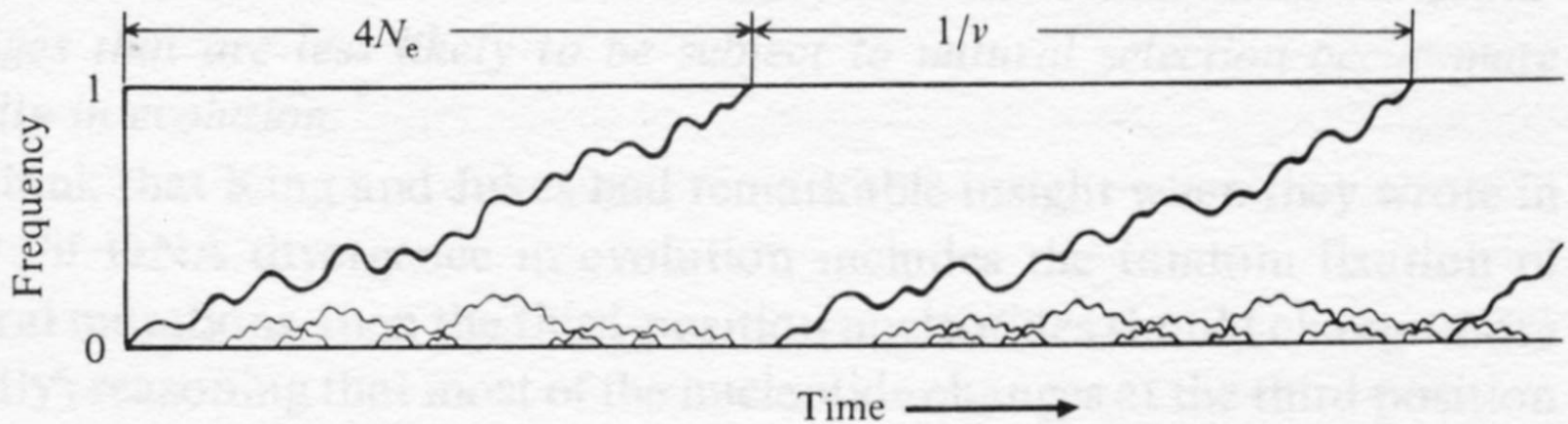
OF MOLECULAR EVOLUTION

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.

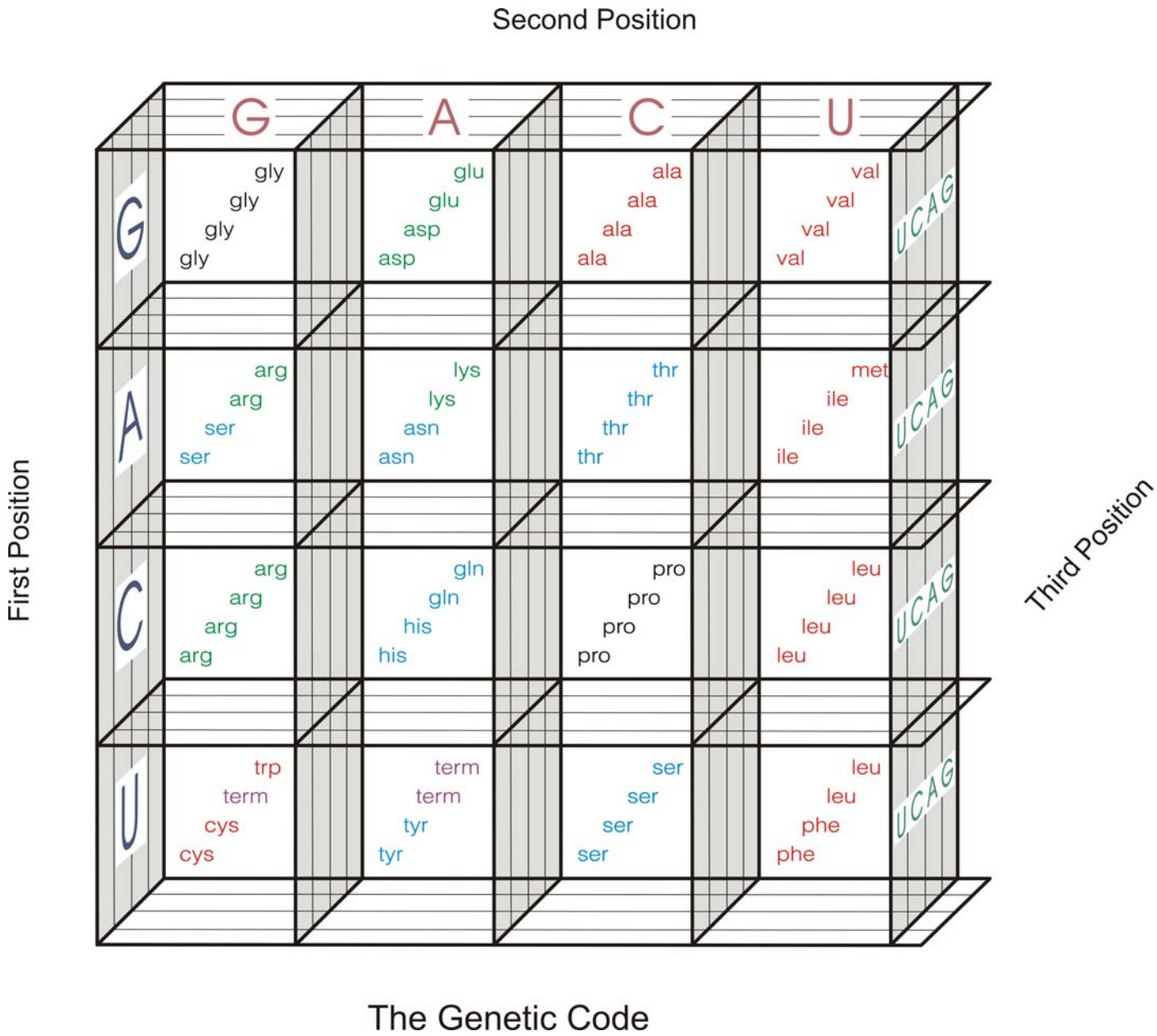


The average time of replacement of a dominant genotype in a population is the reciprocal mutation rate, 1/ν, and therefore independent of population size.

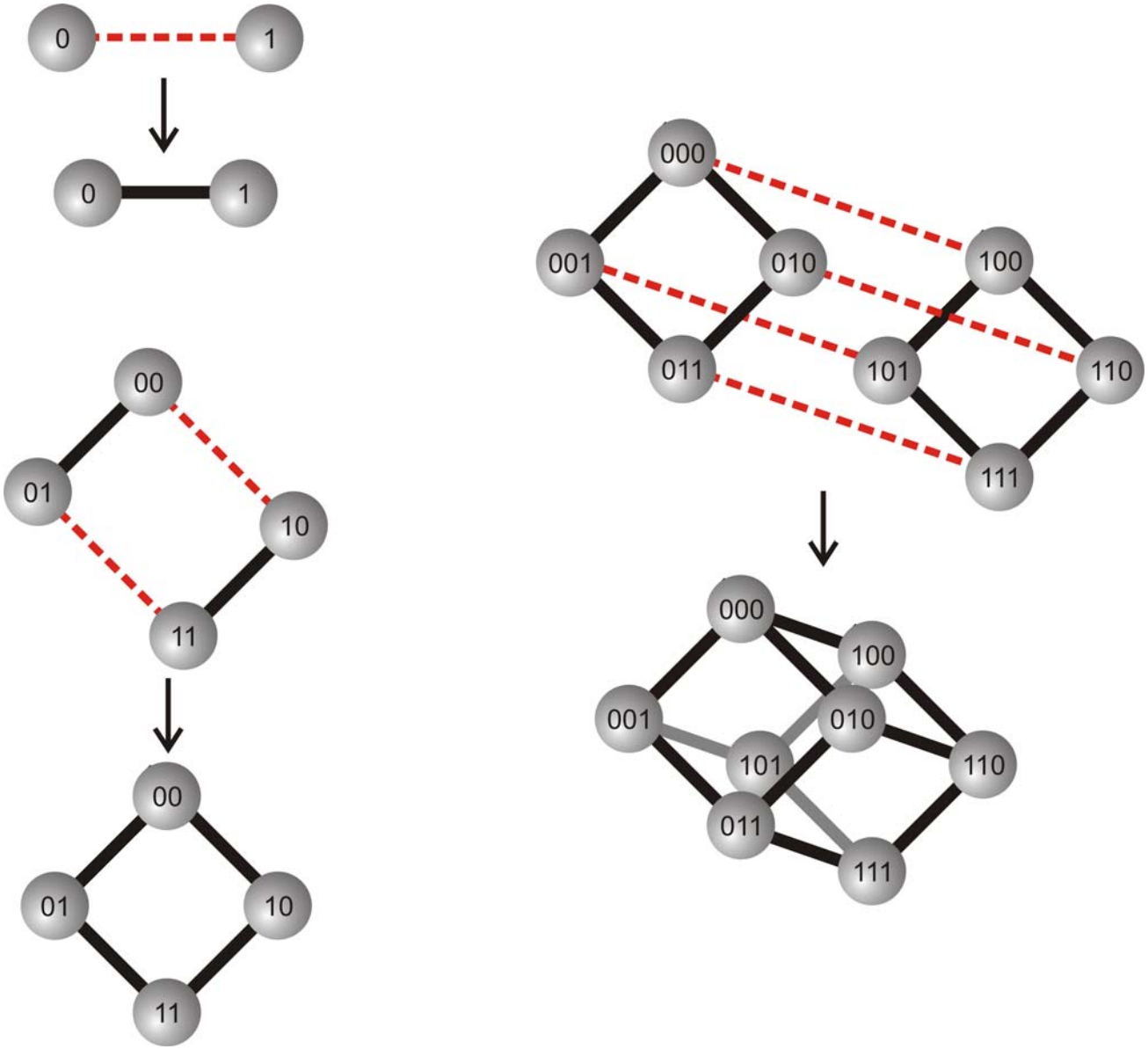
Fixation of mutants in neutral evolution (Motoo Kimura, 1955)

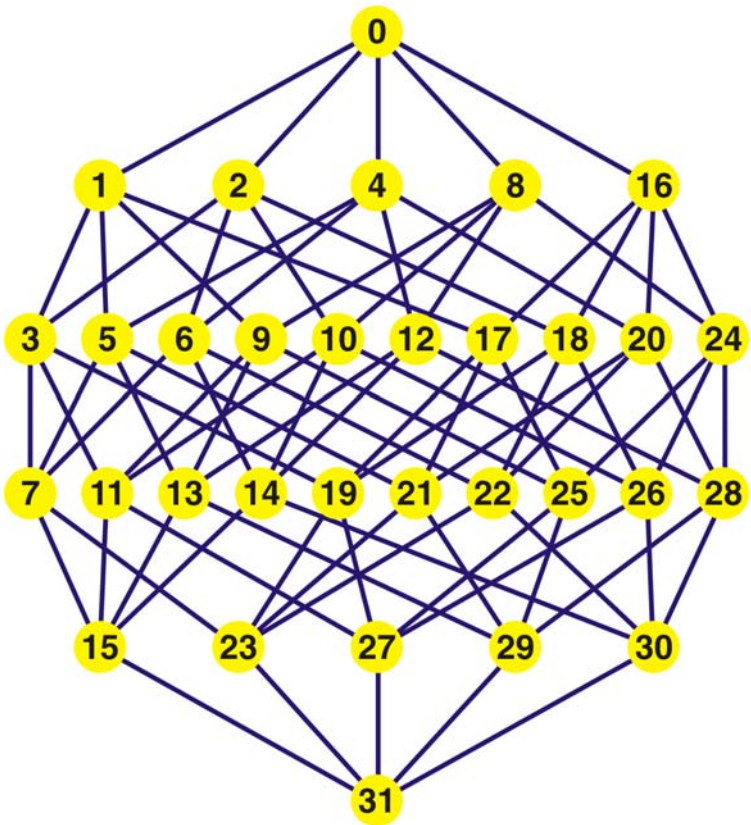
1. The origin of neutrality
2. RNA structures as a useful model
3. RNA replication and quasispecies
4. Selection on realistic landscapes
5. Consequences of neutrality
6. Evolutionary optimization of structure
7. The richness of conformational space

1. **The origin of neutrality**
2. RNA structures as a useful model
3. RNA replication and quasispecies
4. Selection on realistic landscapes
5. Consequences of neutrality
6. Evolutionary optimization of structure
7. The richness of conformational space



Redundancy of the genetic code as a source of neutrality


0
1
0
1
00
01
10
11
00
01
10
11
000
001
010
100
011
101
110
111
000
001
100
010
101
110
011
111



Mutant class

0

1

2

3

4

5

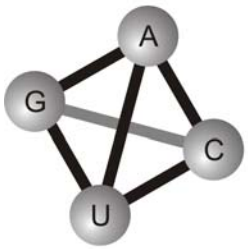
Binary sequences can be encoded by their decimal equivalents:

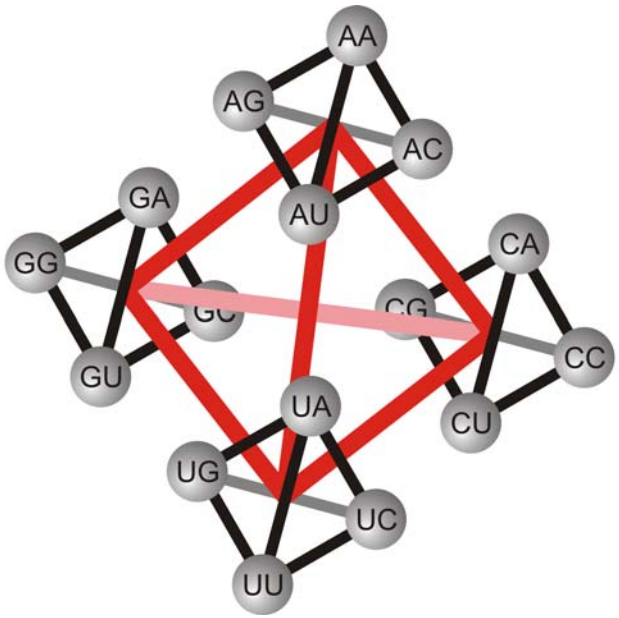
C = 0 and **G = 1**, for example,

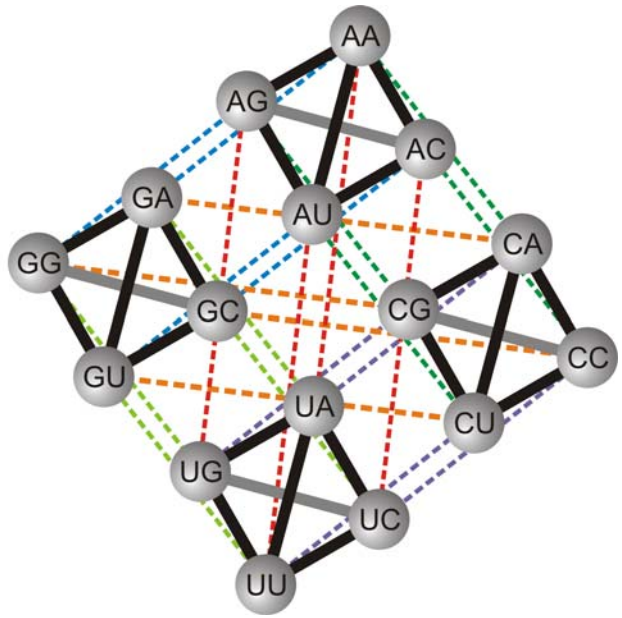
"0" ≡ 00000 = CCCCC,

"14" ≡ 01110 = CGGGC,

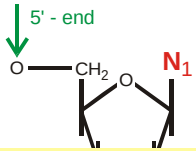
"29" ≡ 11101 = GGGCG, etc.


A
G
C
U

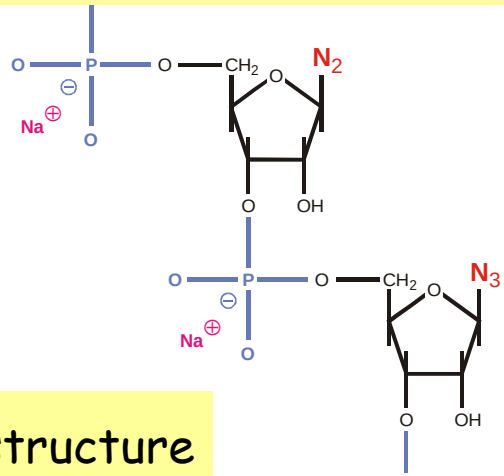

AA
AG
AC
GA
AU
GG
CA
GC
CG
CC
GU
UA
CU
UG
UC
UU

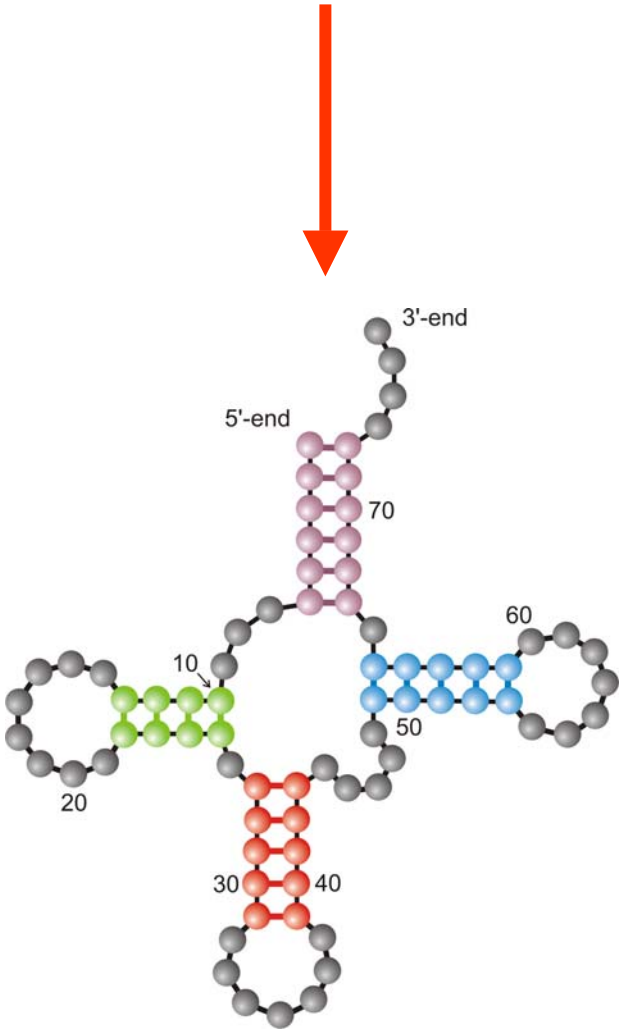

AA
AG
AC
GA
AU
GG
CA
GC
CG
CC
GU
UA
CU
UG
UC
UU

1. The origin of neutrality
2. **RNA structures as a useful model**
3. RNA replication and quasispecies
4. Selection on realistic landscapes
5. Consequences of neutrality
6. Evolutionary optimization of structure
7. The richness of conformational space

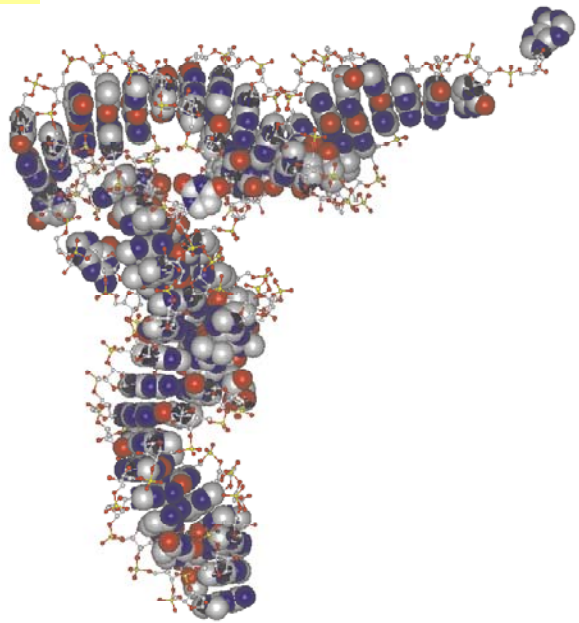


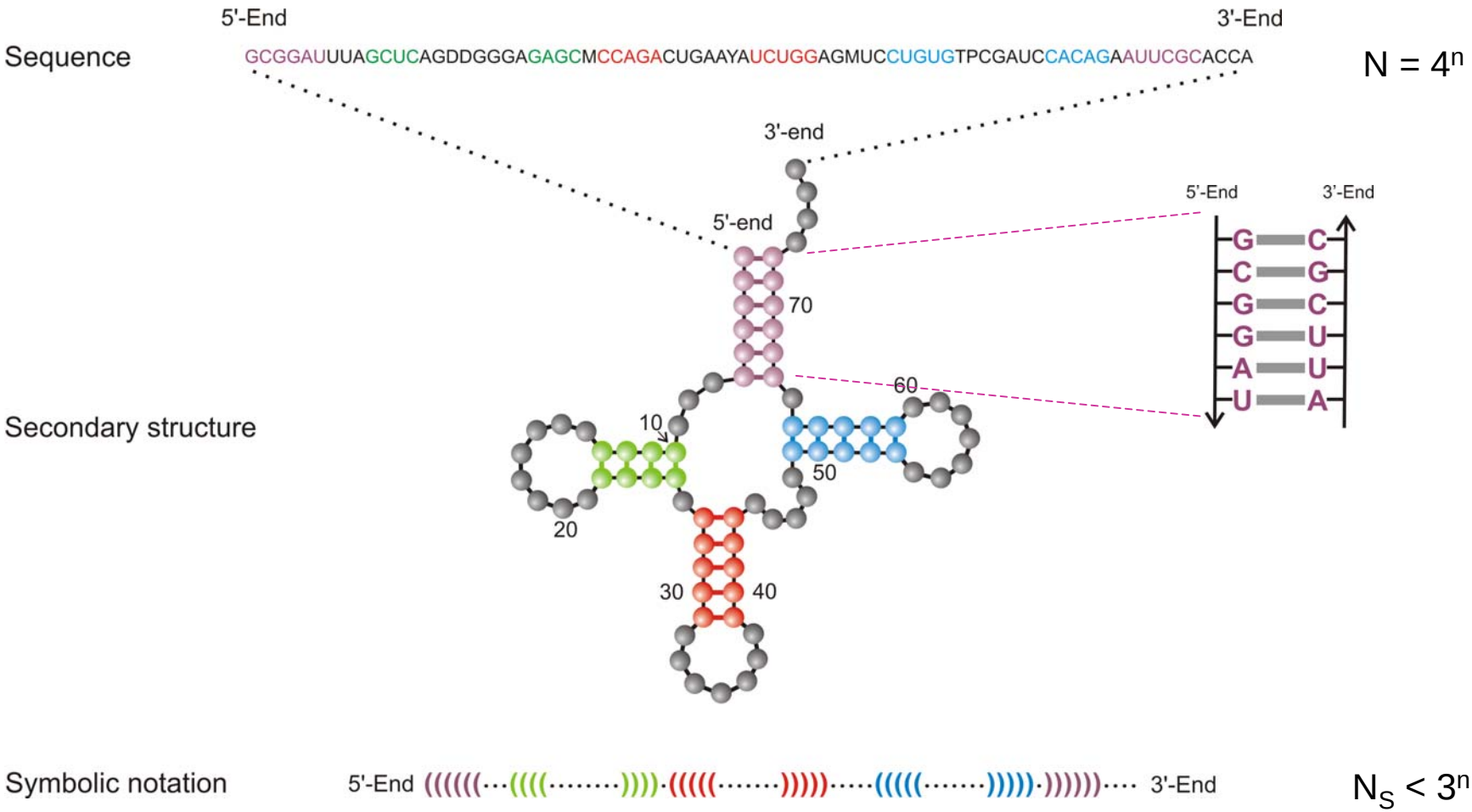
5'-end GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCACCA 3'-end





Definition of RNA structure

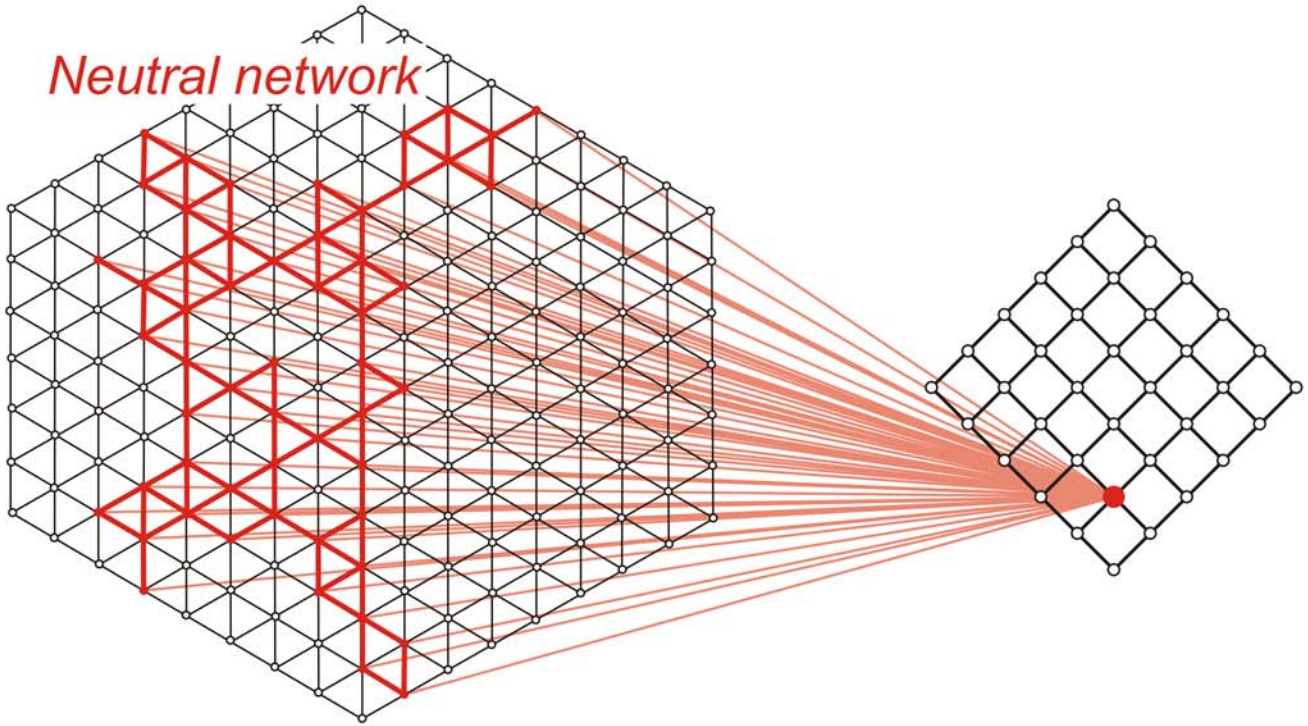




Criterion: Minimum free energy (mfe)

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

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

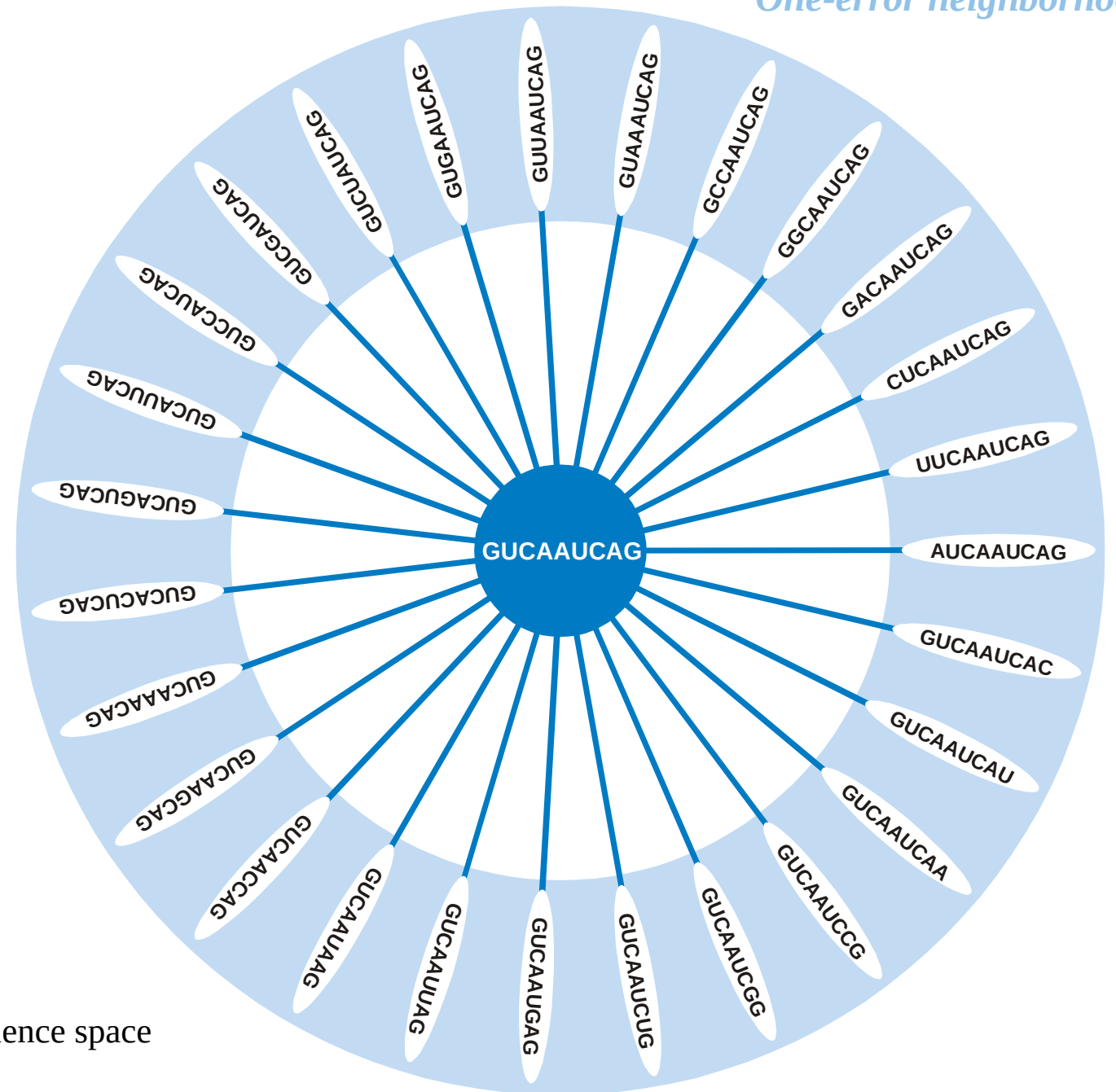


Sequence space

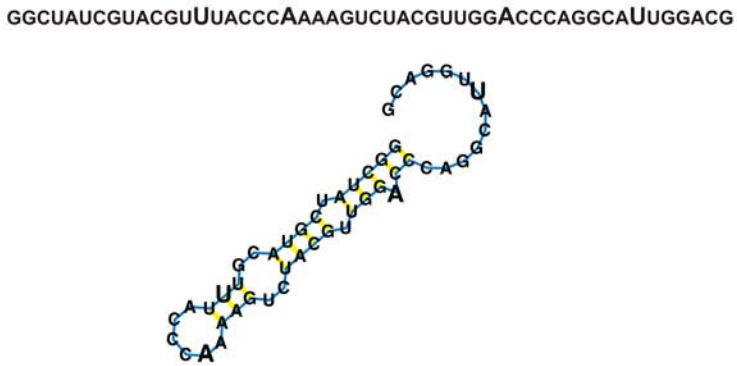
Structure space

many genotypes ⇒ one phenotype

One-error neighborhood



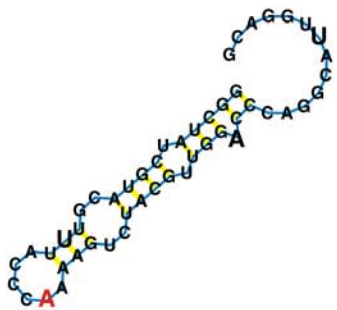
The surrounding of
GUCAAUCAG in sequence space



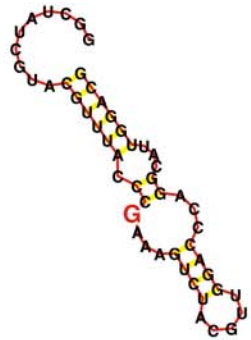
One error neighborhood – Surrounding of an RNA molecule of chain length n=50 in sequence and shape space

GGCUAUCGUACGUUUACCCGAAAGUCUACGUUGGACCCAGGCAUUGGACG

GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG

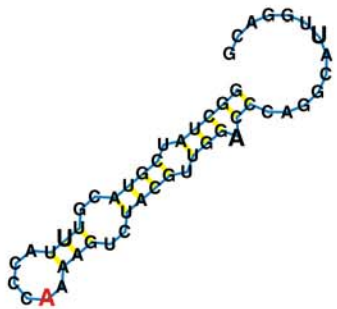


One error neighborhood – Surrounding of an RNA molecule of chain length n=50 in sequence and shape space

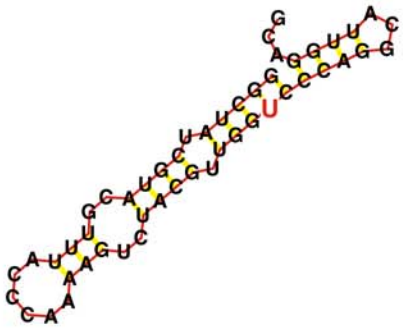


GGCUAUCGUACGUUUACCCGAAAGUCUACGUUGGACCCAGGCAUUGGACG

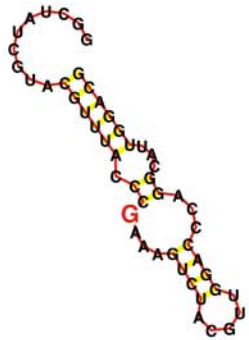
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG



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

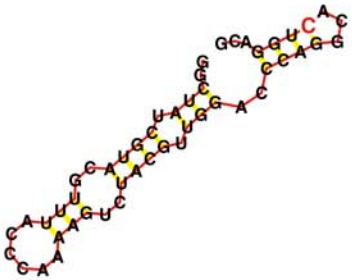


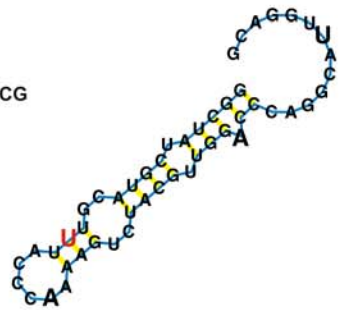
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGUCCCAGGCAUUGGACG



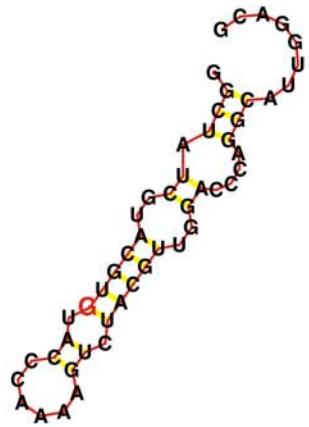
GGCUAUCGUACGUUUACCCGAAAGUCUACGUUGGACCCAGGCAUUGGACG

GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCACUGGACG





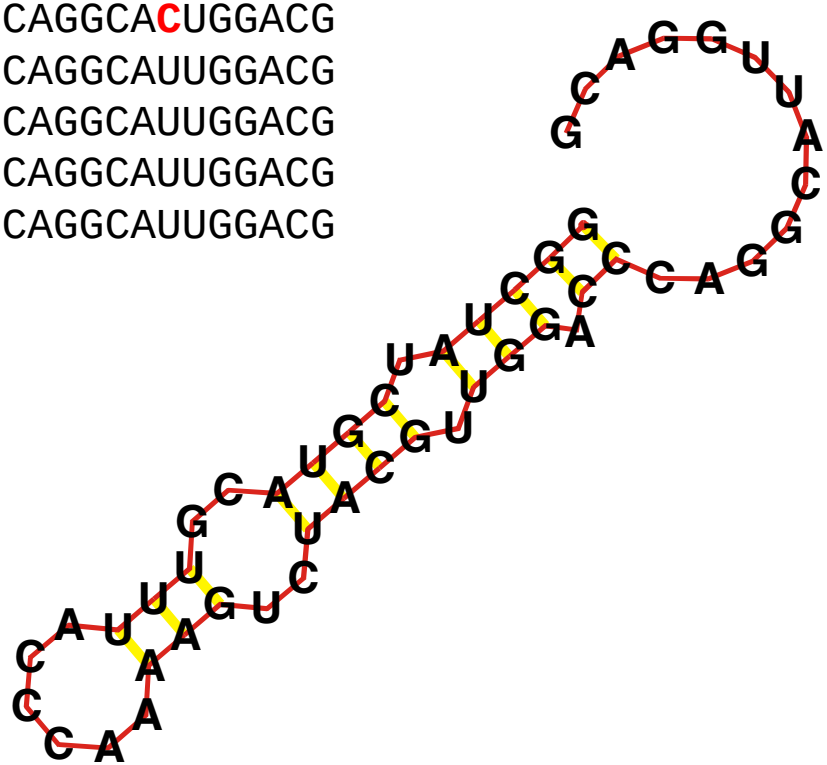
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG



GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG

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

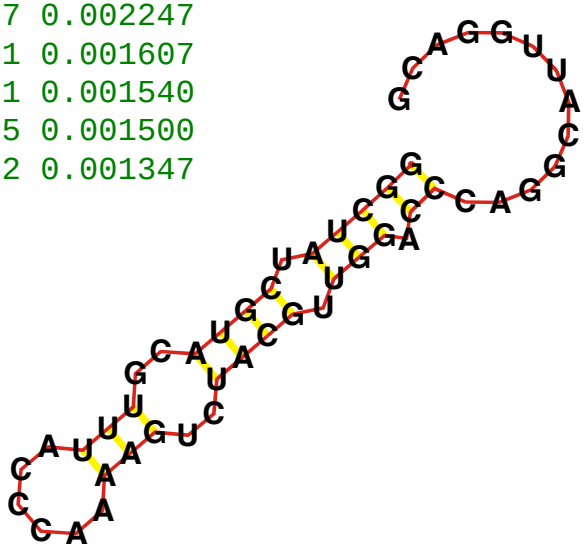
GGCUAUCGUA**U**GUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUU**A**GACG
GGCUAUCGUACGUUUAC**U**CAAAAGUCUACGUUGGACCCAGGCAUUGGACG
GGCUAUCGUACG**C**UUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG
GGC**C**AUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG
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

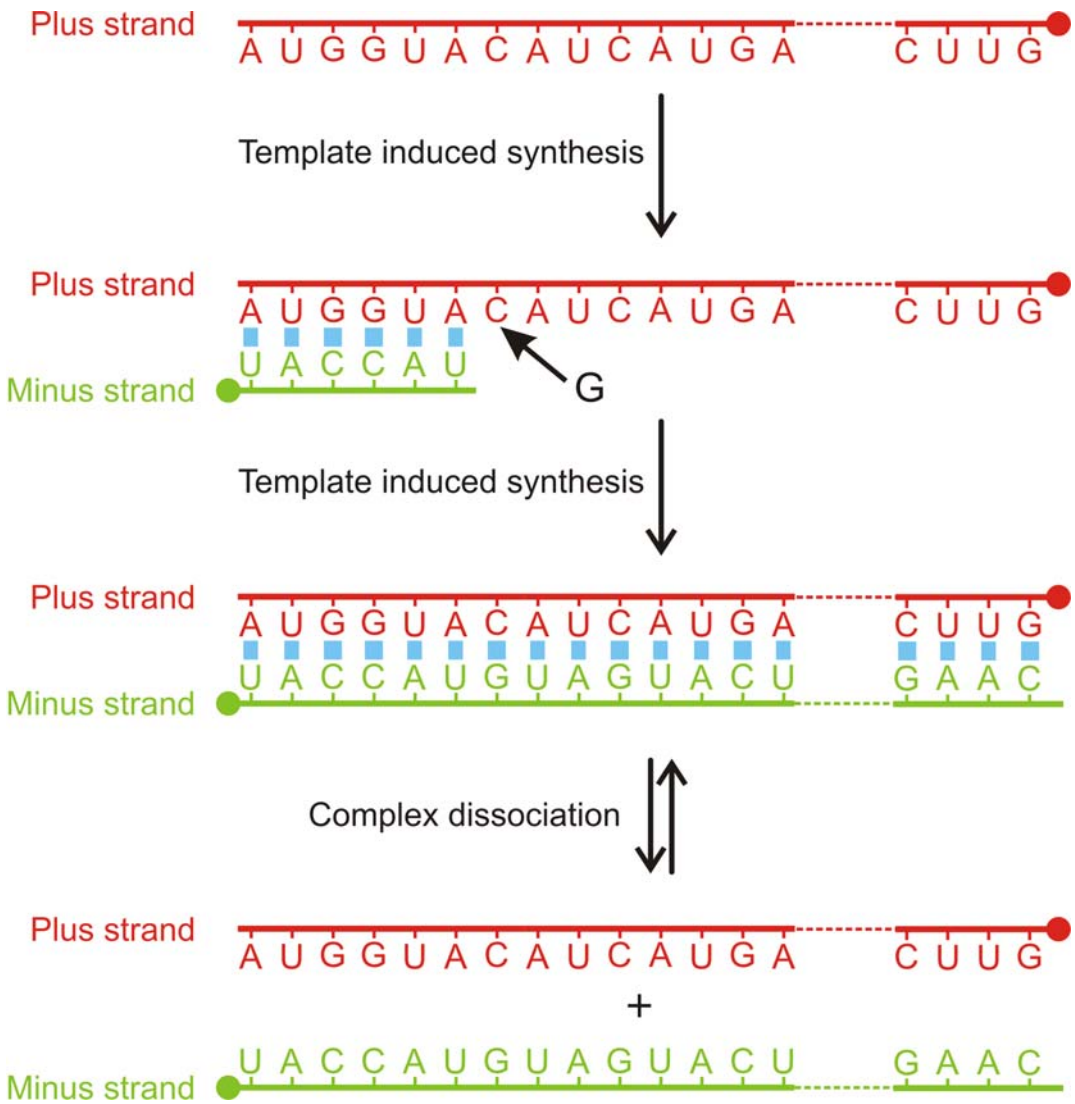
	Number	Mean Value	Variance	Std.Dev.
Total Hamming Distance:	150000	11.647973	23.140715	4.810480
Nonzero Hamming Distance:	99875	16.949991	30.757651	5.545958
Degree of Neutrality:	50125	**0.334167**	0.006961	**0.083434**
Number of Structures:	**1000**	**52.31**	85.30	**9.24**

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



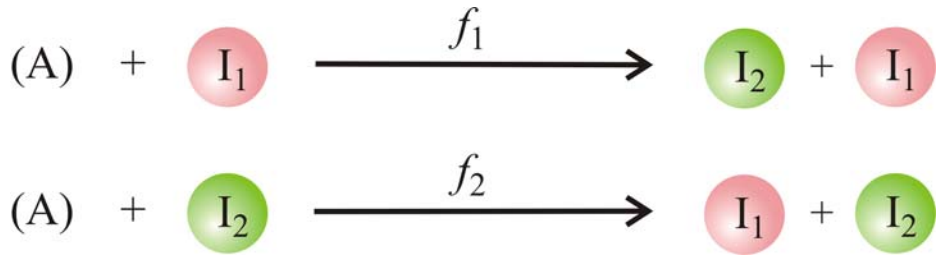
Shadow – Surrounding of an RNA structure in shape space:
AUGC alphabet, chain length n=50

1. The origin of neutrality
2. RNA structures as a useful model
3. **RNA replication and quasispecies**
4. Selection on realistic landscapes
5. Consequences of neutrality
6. Evolutionary optimization of structure
7. The richness of conformational space



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

G≡C and **A=U**



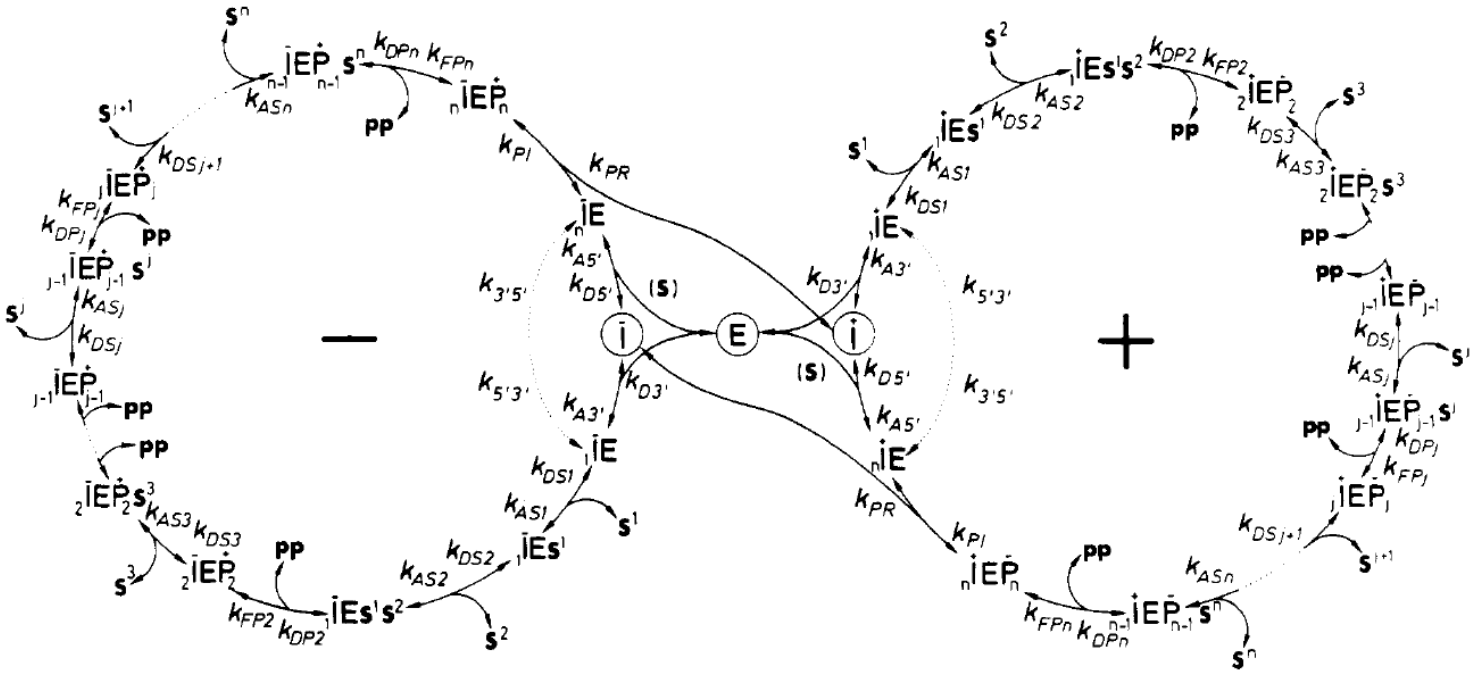
$$\frac{dx_1}{dt} = f_2\, x_2 \quad \text{and} \quad \frac{dx_2}{dt} = f_1\, x_1$$

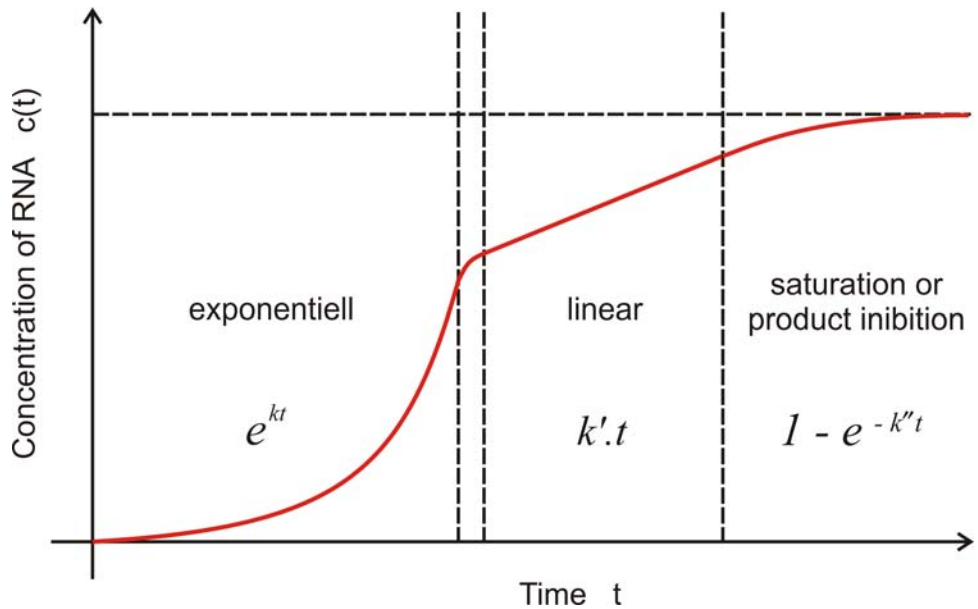
$$x_1 = \sqrt{f_2}\;\xi_1\,, \quad x_2 = \sqrt{f_1}\;\xi_2\,, \quad \zeta = \xi_1 + \xi_2\,, \quad \eta = \xi_1 - \xi_2\,, \quad f = \sqrt{f_1 f_2}$$

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

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

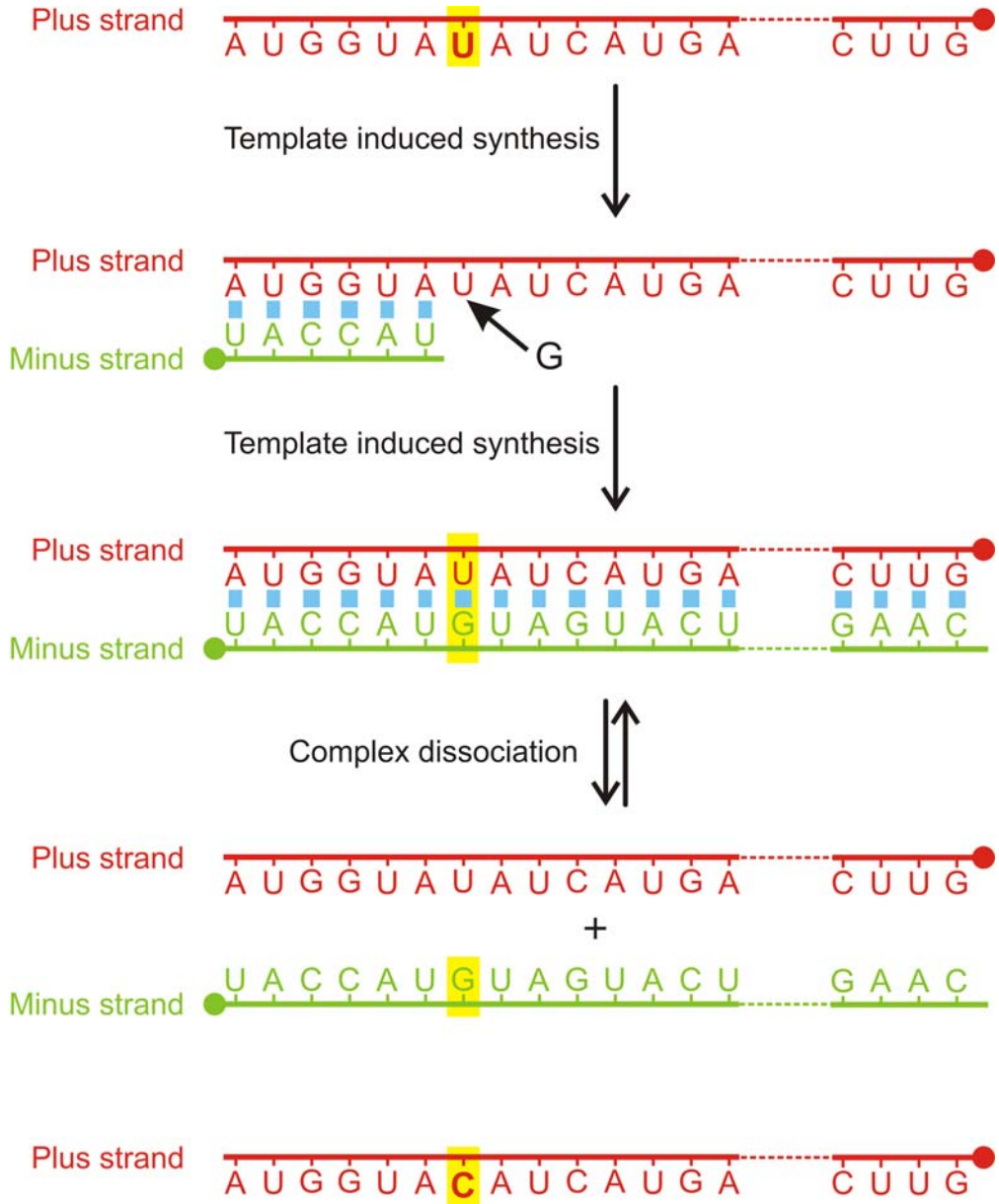
Complementary replication as the simplest molecular mechanism of reproduction

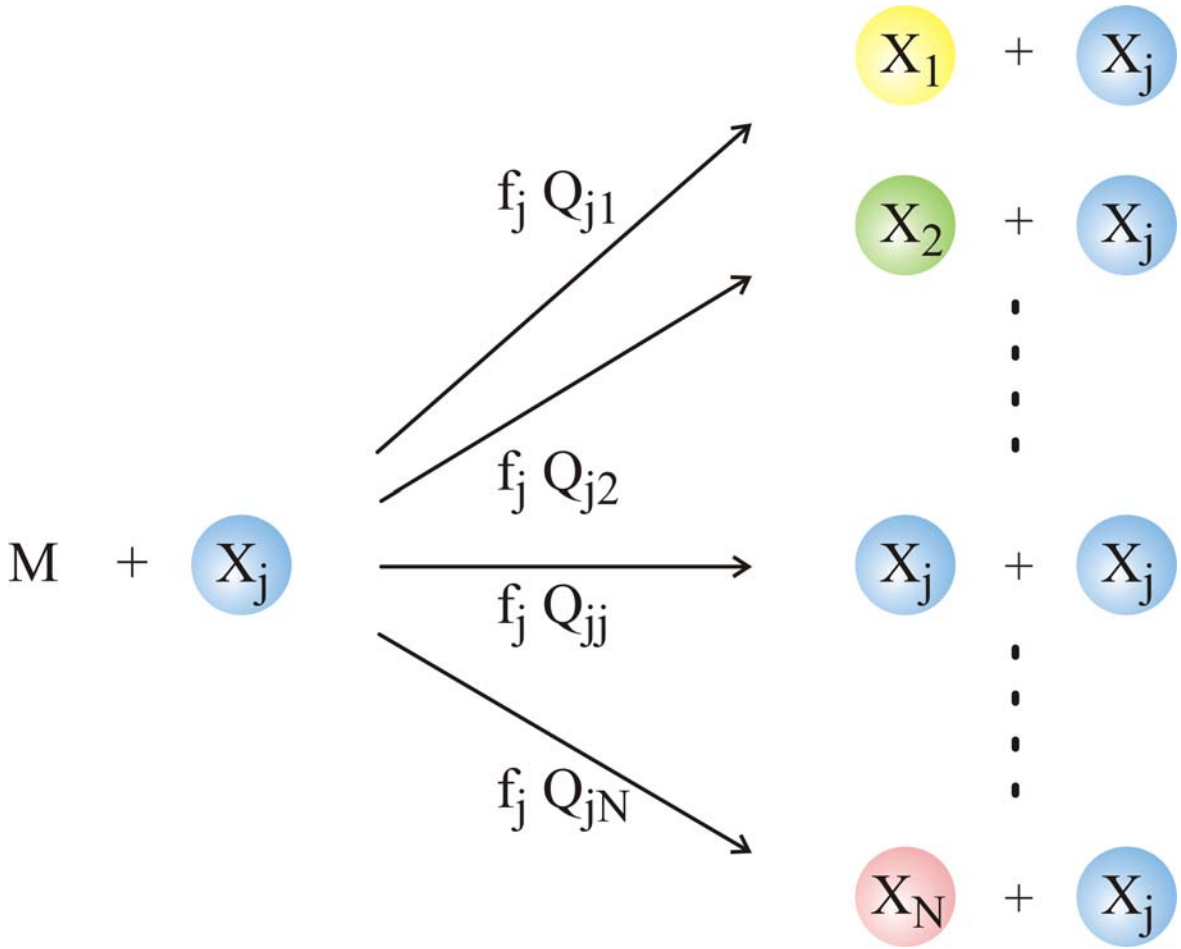




Kinetics of RNA replication

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


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



Chemical kinetics of replication and mutation as parallel reactions

$$\frac{dc_i}{dt} \;=\; \sum_{j=1}^{N} Q_{ij}\, f_j\, c_j\,; \quad i = 1, 2, \ldots, N$$

$$\frac{d\mathbf{c}}{dt} \;=\; \mathrm{W} \cdot \mathbf{c}\,; \quad \sum_{1=1}^{N} c_i(t) \;=\; c(t)\,; \quad \mathrm{W} \;=\; \{W_{ij} \doteq Q_{ij}\, f_j\}$$

Normalization

$$x_i \;=\; c_i/c\,; \quad \sum_{i_1}^{n} x_i \;=\; 1$$

$$\frac{d\mathbf{x}}{dt} \;=\; \mathrm{W} \cdot \mathbf{x} \;-\; \bar{f}\,\mathbf{x} \;=\; (\mathrm{G} \cdot \mathrm{F} - \bar{f}\,\mathbb{E}) \cdot \mathbf{x}\,; \quad \bar{f} \;=\; \sum_{i=1}^{N} x_i\, f_i$$

Matrix W and Frobenius theorem:

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

Primitive matrix W:

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

Perron-Frobenius theorem applied to the value matrix W

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

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

(iii) λ_0 is associated with strictly positive eigenvectors

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

(v-vi) etc.

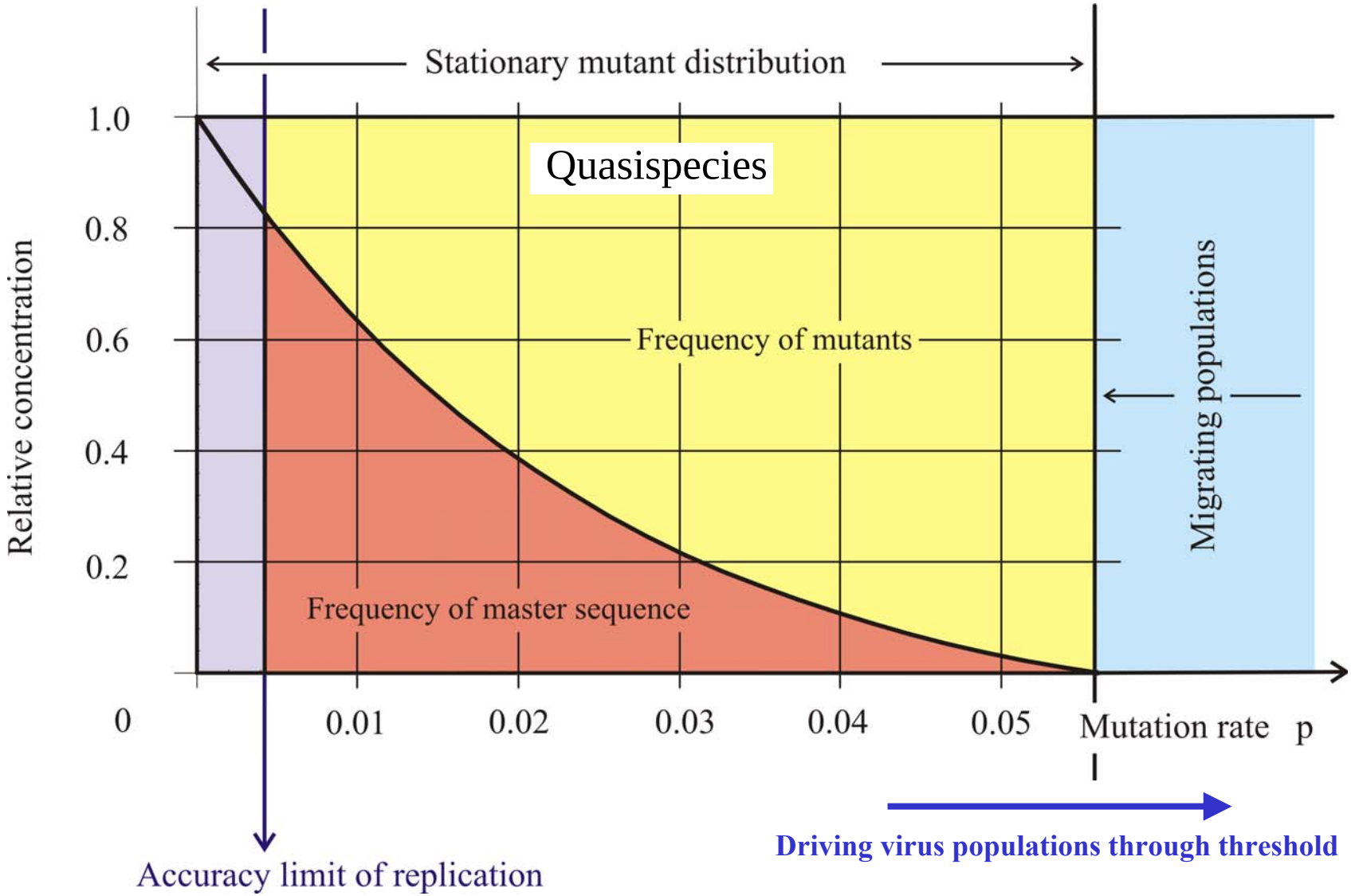
W is irreducible: (i), (iii), (iv), etc. as above

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

Decomposition of matrix W

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

$$\mathrm{Q} = \begin{pmatrix} Q_{11} & Q_{12} & \dots & Q_{1n} \\ Q_{21} & Q_{22} & \dots & Q_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ Q_{n1} & Q_{n2} & \dots & Q_{nn} \end{pmatrix} \quad \text{and} \quad \mathrm{F} = \begin{pmatrix} f_1 & 0 & \dots & 0 \\ 0 & f_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & f_n \end{pmatrix}$$



The error threshold in replication

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

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.

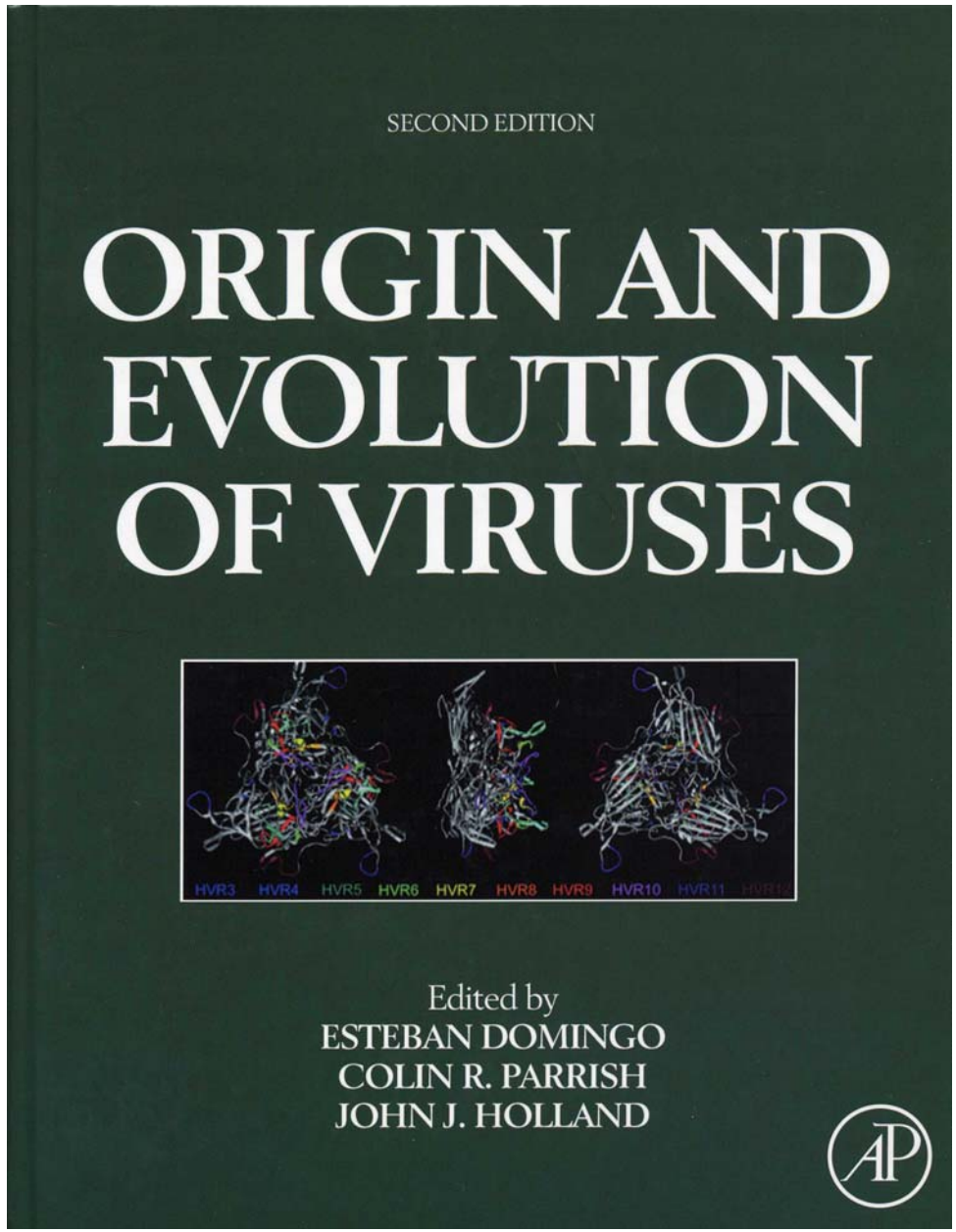
My thanks go also to the 19 authors who despite their busy schedules have taken time to prepare excellent manuscripts, to Elsevier staff for their prompt responses to my requests, and, last but not least, to Ms. Lucía Horrillo from Centro de Biología Molecular "Severo Ochoa" for her patient dealing with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



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

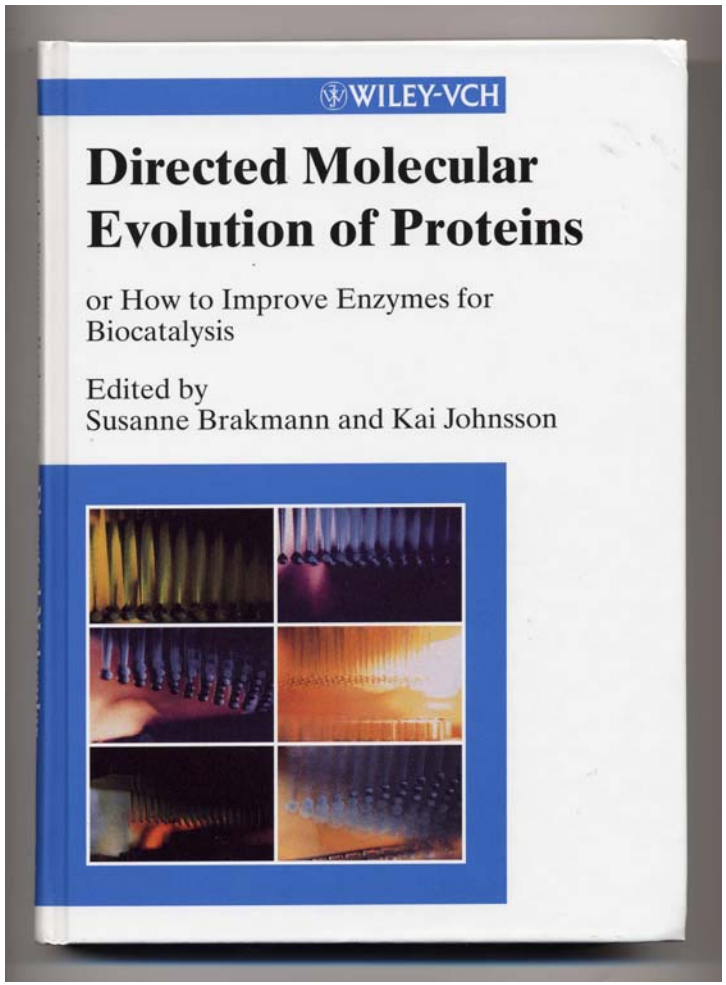
C. Tuerk, L. Gold, **SELEX - *Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage* T4 *DNA polymerase***. Science **249** (1990), 505-510

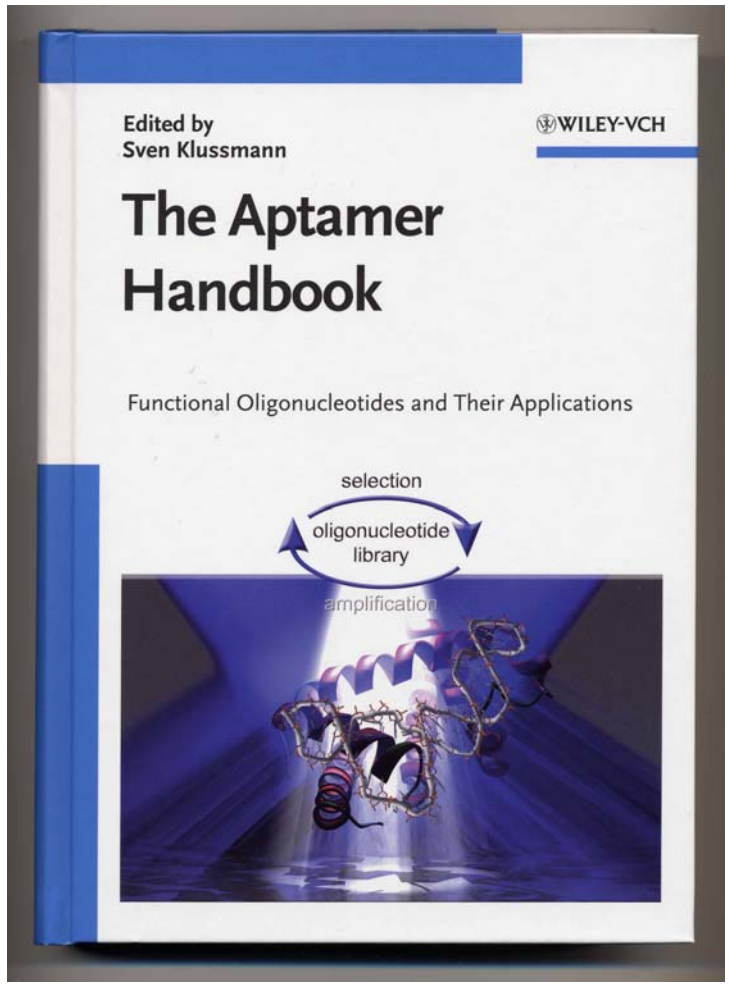
D.P. Bartel, J.W. Szostak, ***Isolation of new ribozymes from a large pool of random sequences***. Science **261** (1993), 1411-1418

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





Application of molecular evolution to problems in biotechnology

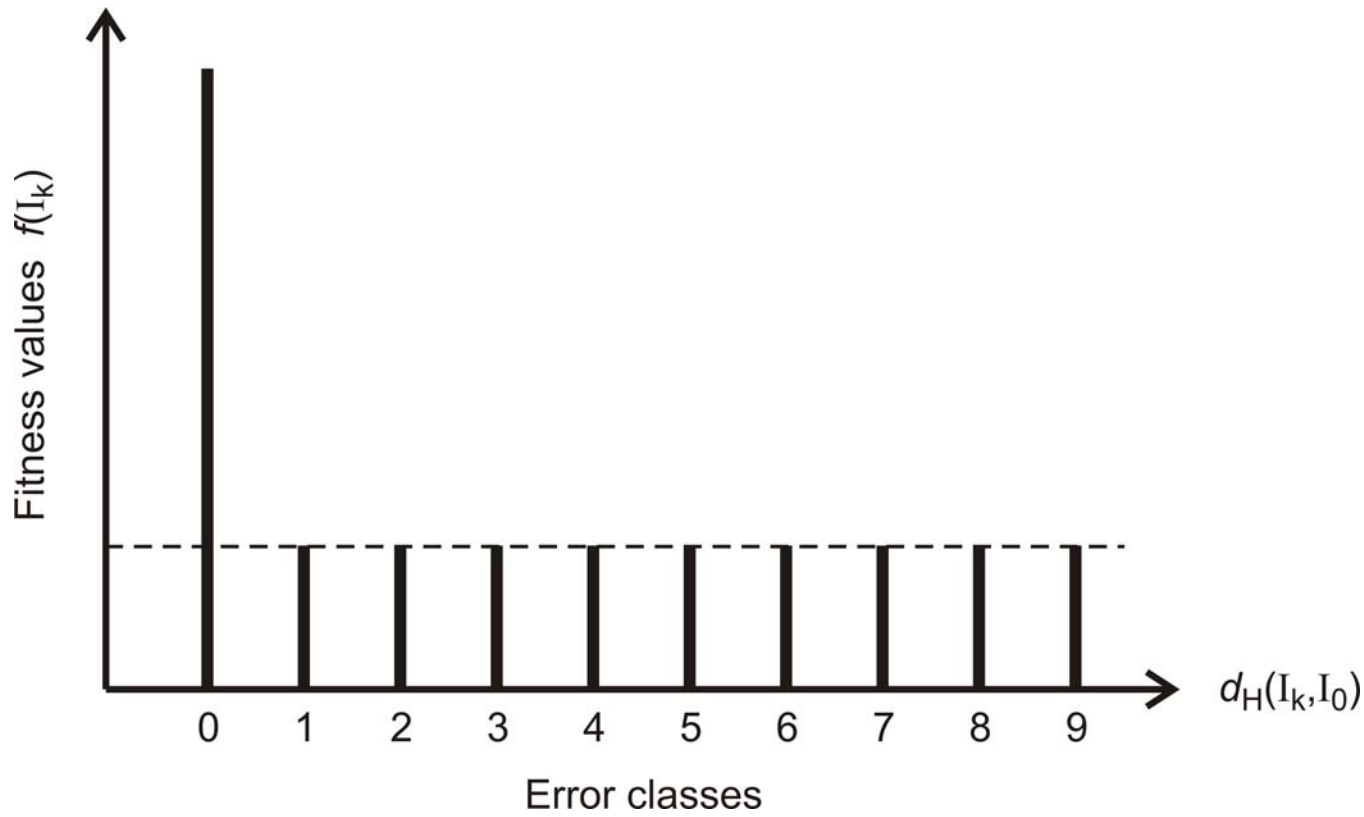
Artificial evolution in biotechnology and pharmacology

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.

1. The origin of neutrality
2. RNA structures as a useful model
3. RNA replication and quasispecies
4. **Selection on realistic landscapes**
5. Consequences of neutrality
6. Evolutionary optimization of structure
7. The richness of conformational space



A fitness landscape showing an error threshold:

The single-peak landscape

Uniform error rate model:

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

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

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.

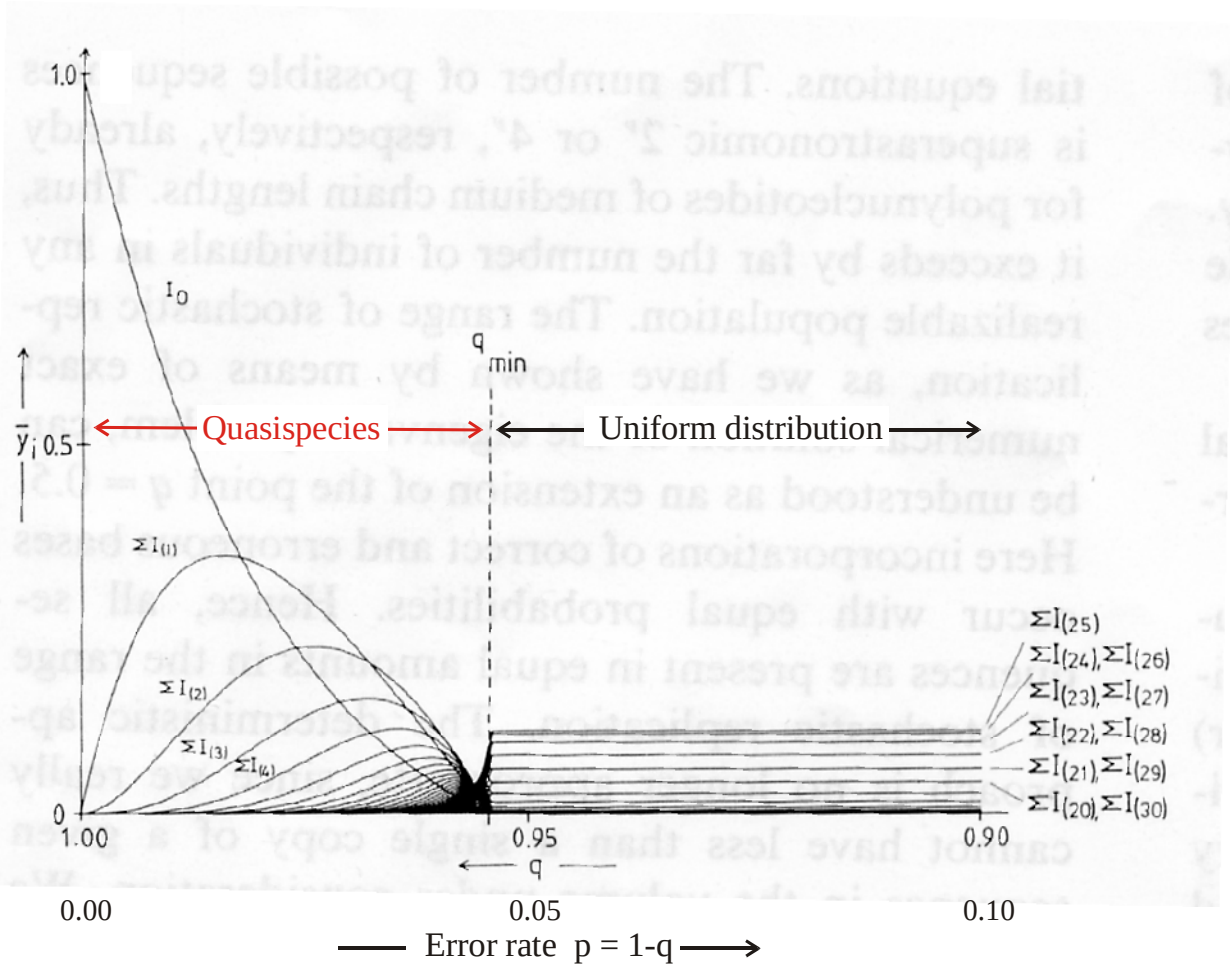
Rigorous mathematical analysis has been performed on eq. 1 [7,15,24,26]. In particular, it was shown that the non-linearity of eq. 1 can be removed by an appropriate transformation. The eigenvalue problem of the linear differential equation obtained thereby may be solved approximately by the conventional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

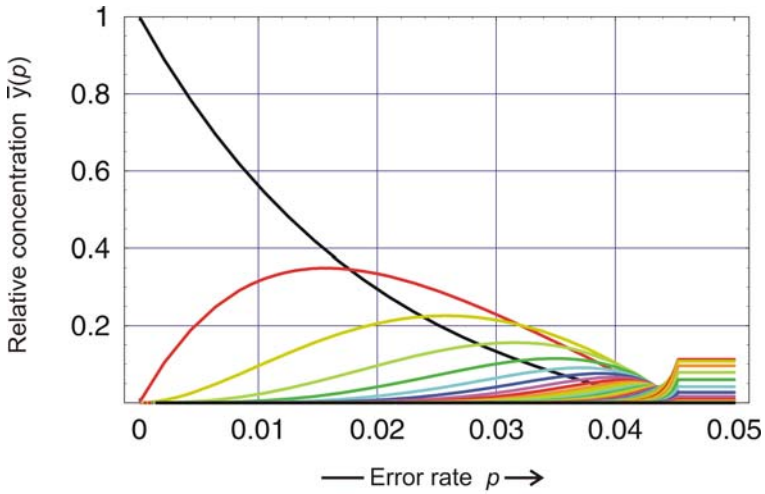
** This paper is considered as part II of Model Studies on RNA replication. Part I is by Gassner and Schuster [14].

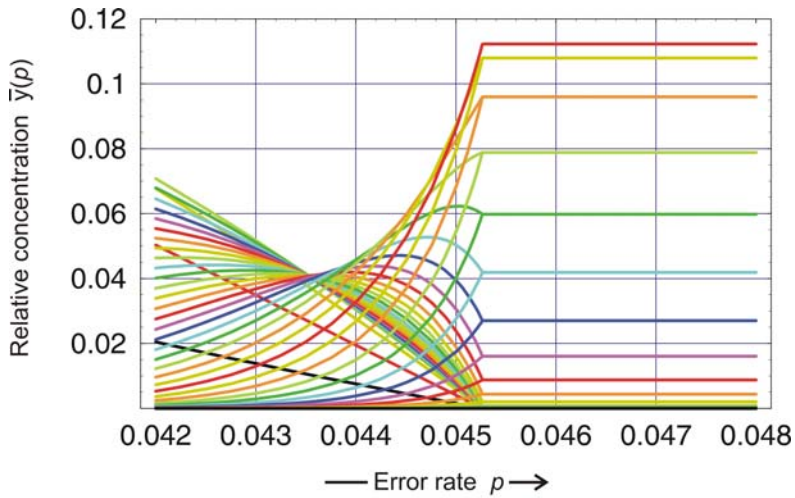
† All summations throughout this paper run from 1 to n unless specified differently: $\Sigma_i = \Sigma_{i=1}^{n}$ and $\Sigma_{i,i\neq j} = \Sigma_{i=1}^{j-1} + \Sigma_{i=j+1}^{n}$, respectively.

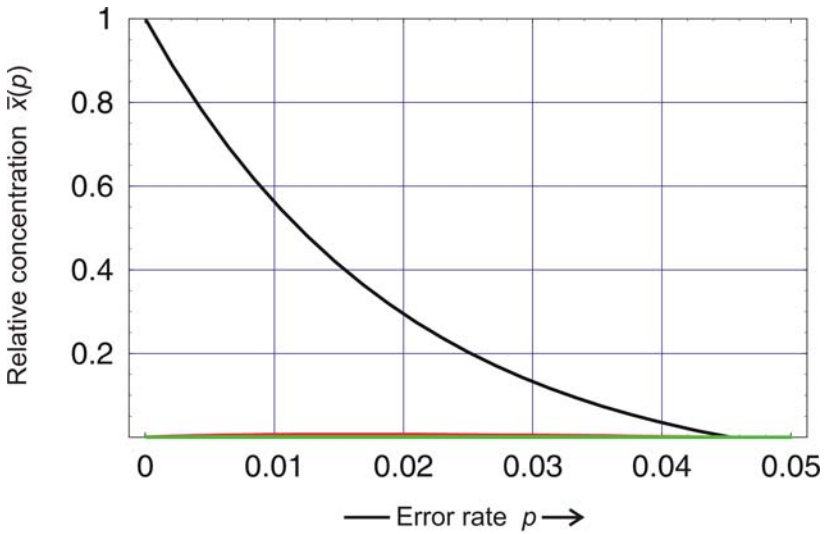
0301-4622/82/0000-0000/$02.75

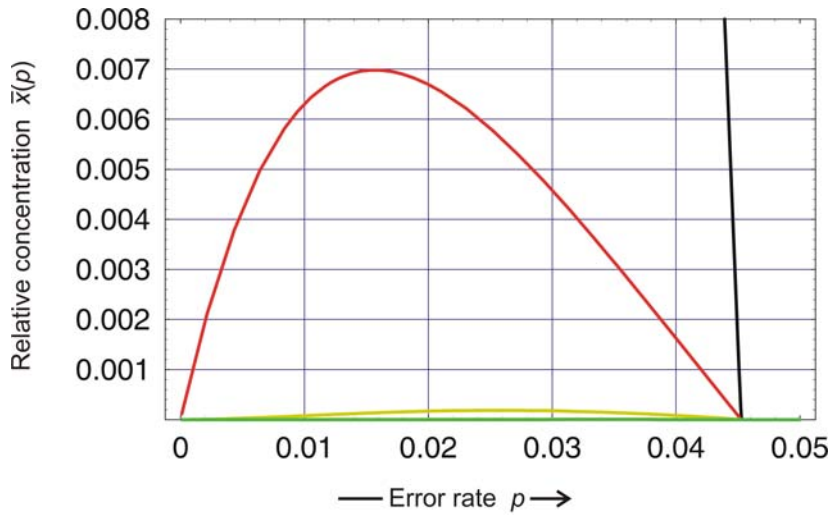


Stationary population or quasispecies as a function of the mutation or error rate *p*









Error threshold on a single peak fitness landscape with n = 50 and σ = 10

PHYSICAL REVIEW A VOLUME 45, NUMBER 8 15 APRIL 1992

Error thresholds for molecular quasispecies as phase transitions: From simple landscapes to spin-glass models

P. Tarazona
Institut für Theoretische Chemie der Universität Wien, A-1090 Wien, Austria
*and Departamento de Fisica de la Materia Condensada, Universidad Autonoma de Madrid, E-28049, Madrid, Spain**
(Received 19 June 1991)

The correspondence between Eigen's model [Naturwissenschaften **58**, 465 (1971)] for molecular quasispecies and the equilibrium properties of a lattice system proposed by Leuthäusser [J. Chem. Phys. **84**, 1884 (1986); J. Stat. Phys. **48**, 343 (1987)] is used to characterize the error thresholds for the existence of quasispecies as phase transitions. For simple replication landscapes the error threshold is related to a first-order phase transition smoothed by the complete wetting of the time surface. Replication landscapes based on the Hopfield Hamiltonian for neural networks allow for the tuning of the landscape complexity and reveal the existence of two error thresholds, bracketing a region of spin-glass quasispecies between the simple quasispecies and the fully disordered mixture of sequences.

PACS number(s): 87.10.+e, 64.60.Cn, 05.50.+q

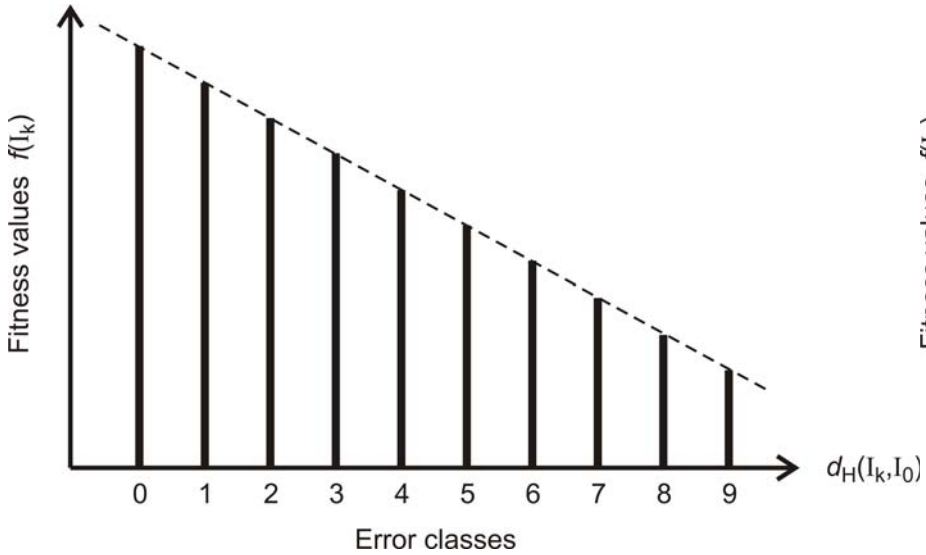
VOLUME 91, NUMBER 13 PHYSICAL REVIEW LETTERS week ending 26 SEPTEMBER 2003

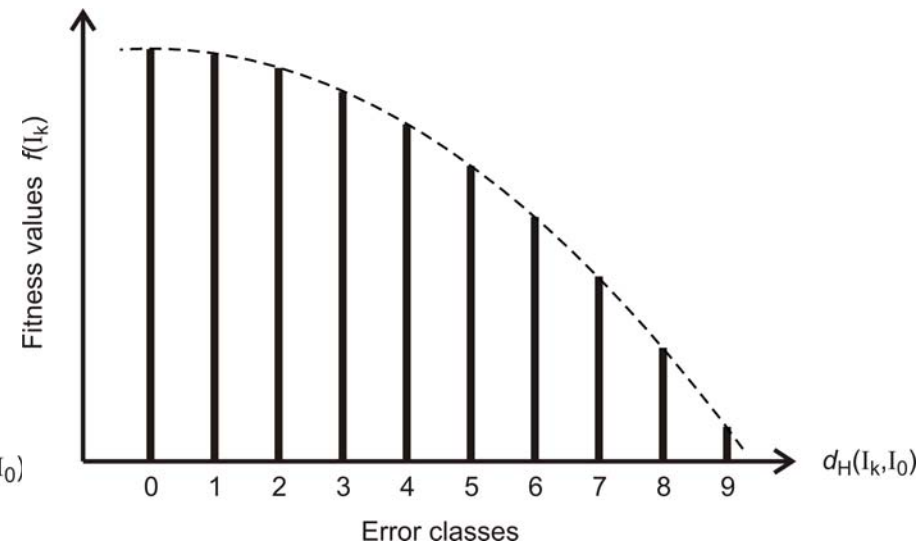
Equilibrium Distribution of Mutators in the Single Fitness Peak Model

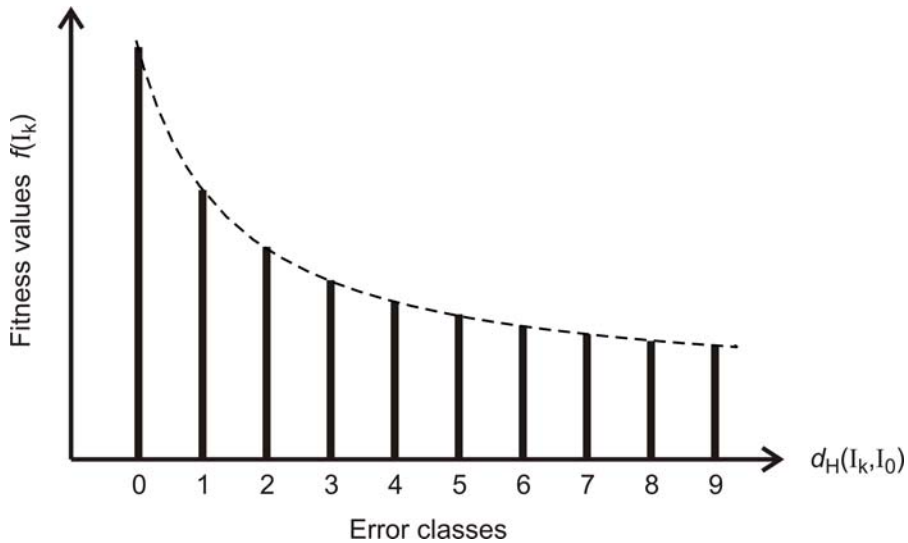
Emmanuel Tannenbaum,* Eric J. Deeds, and Eugene I. Shakhnovich
Harvard University, Cambridge, Massachusetts 02138, USA
(Received 25 April 2003; published 26 September 2003)

This Letter develops an analytically tractable model for determining the equilibrium distribution of mismatch repair deficient strains in unicellular populations. The approach is based on the single fitness peak model, which has been used in Eigen's quasispecies equations in order to understand various aspects of evolutionary dynamics. As with the quasispecies model, our model for mutator-nonmutator equilibrium undergoes a phase transition in the limit of infinite sequence length. This "repair catastrophe" occurs at a critical repair error probability of $\epsilon_r = L_{\text{via}}/L$, where L_{via} denotes the length of the genome controlling viability, while L denotes the overall length of the genome. The repair catastrophe therefore occurs when the repair error probability exceeds the fraction of deleterious mutations. Our model also gives a quantitative estimate for the equilibrium fraction of mutators in *Escherichia coli*.

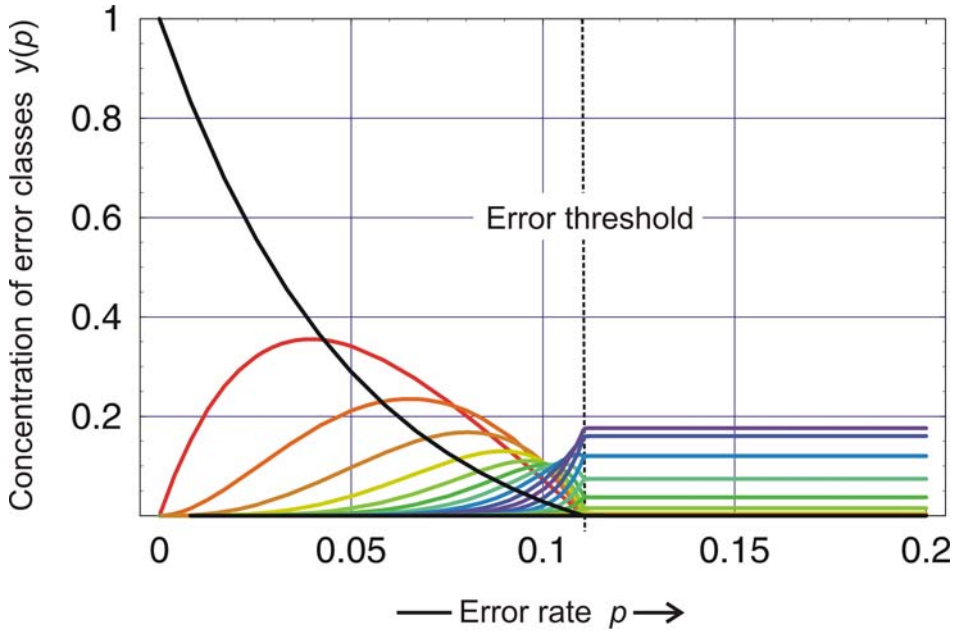
DOI: 10.1103/PhysRevLett.91.138105 PACS numbers: 87.23.Kg, 64.60.–i, 87.16.Ac

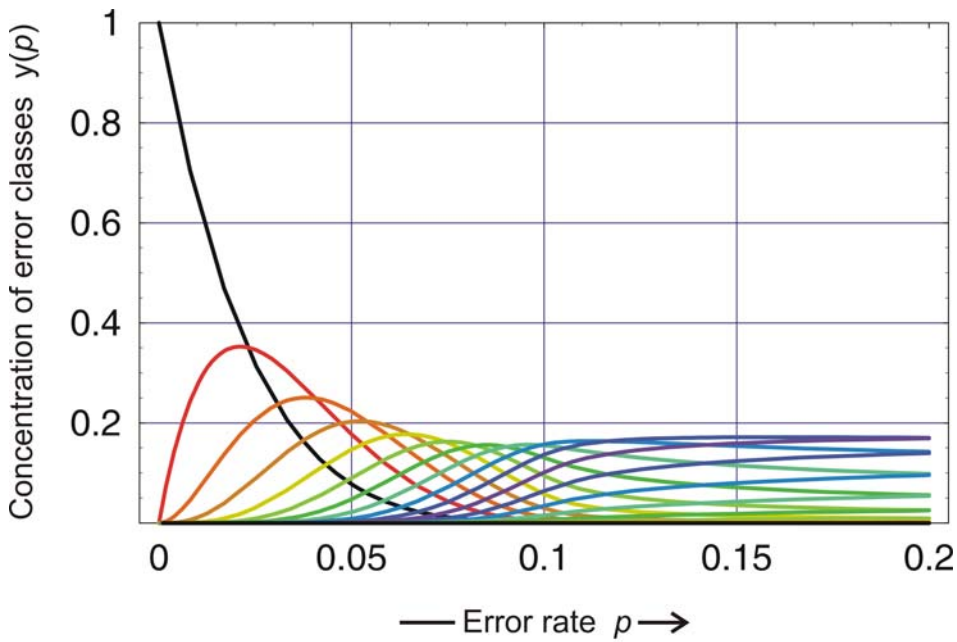






Fitness landscapes **not** showing error thresholds





Error thresholds and gradual transitions

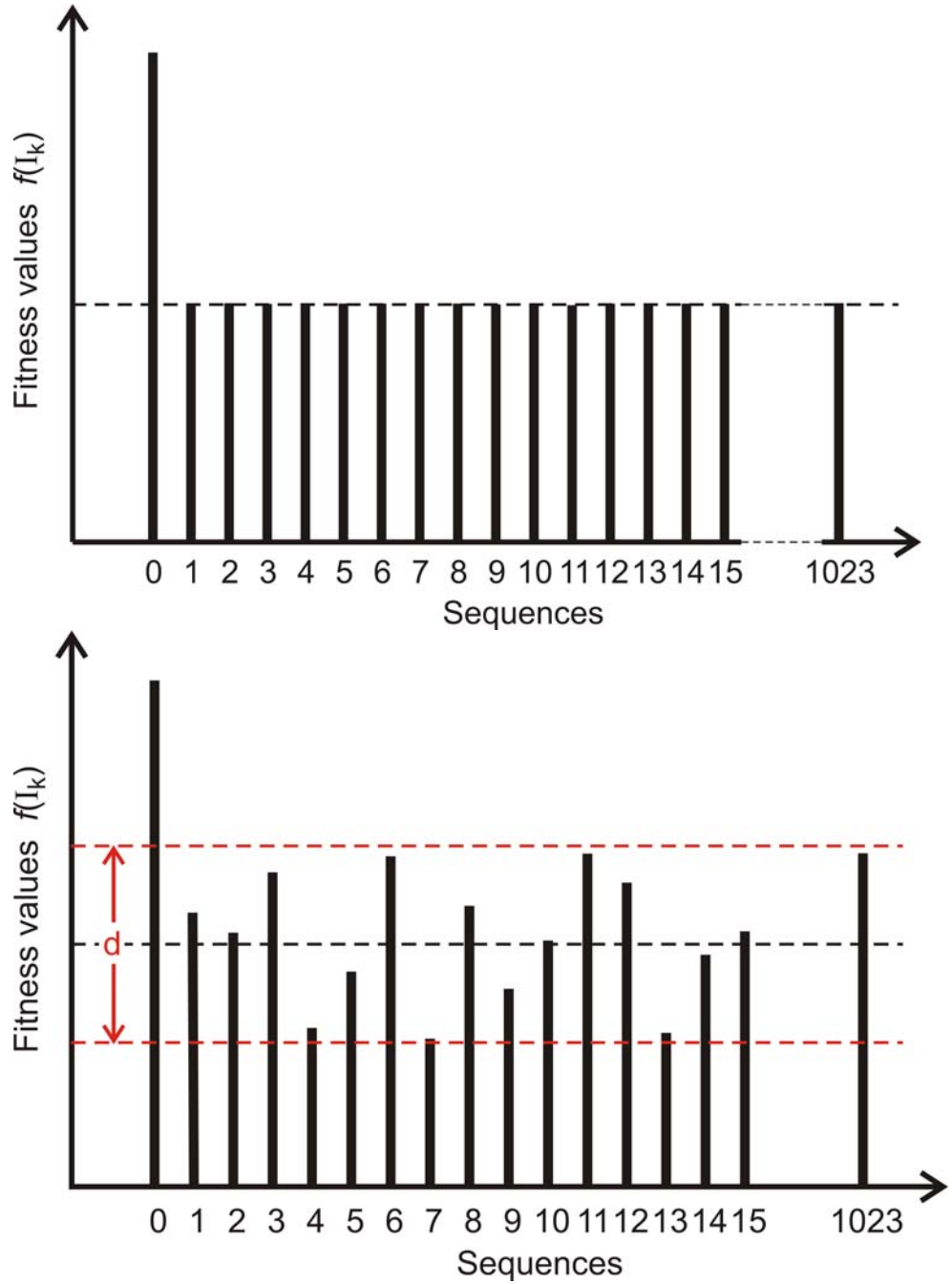
n = 20 and $\sigma = 10$

Features of realistic landscapes:

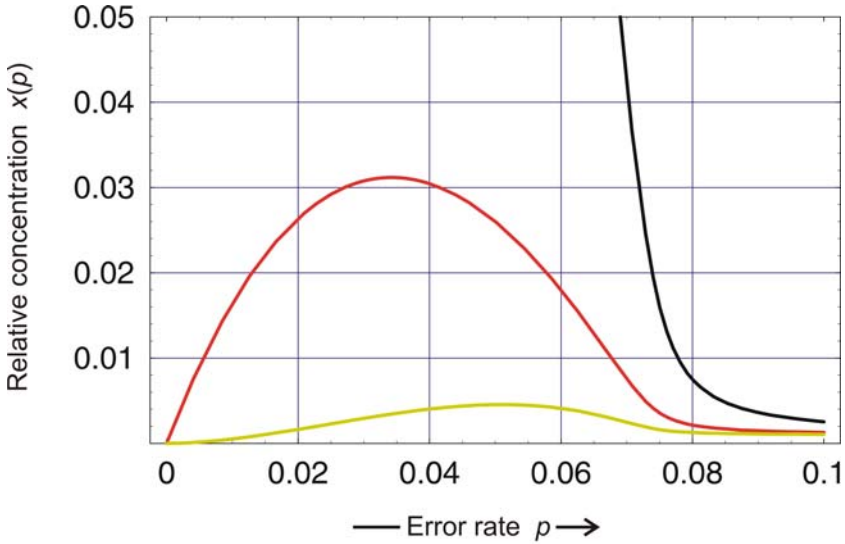
1. Variation in fitness values
2. Deviations from uniform error rates
3. Neutrality

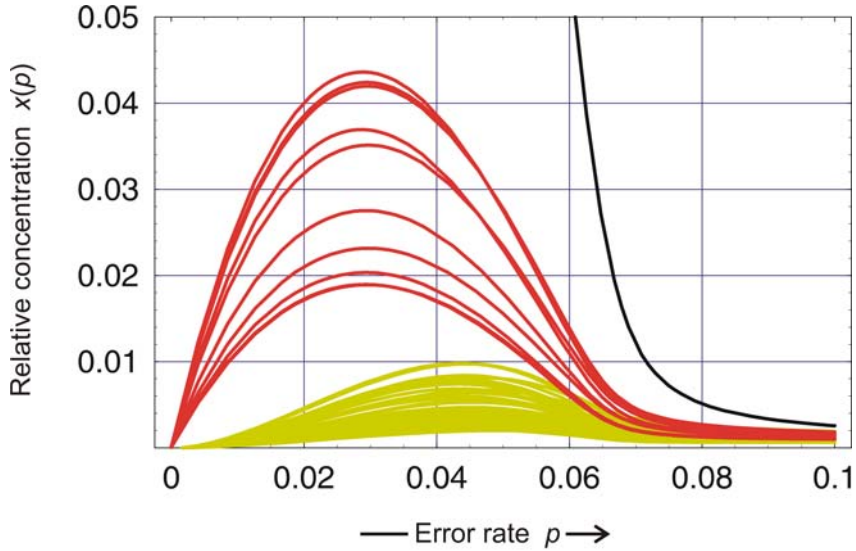
Features of realistic landscapes:

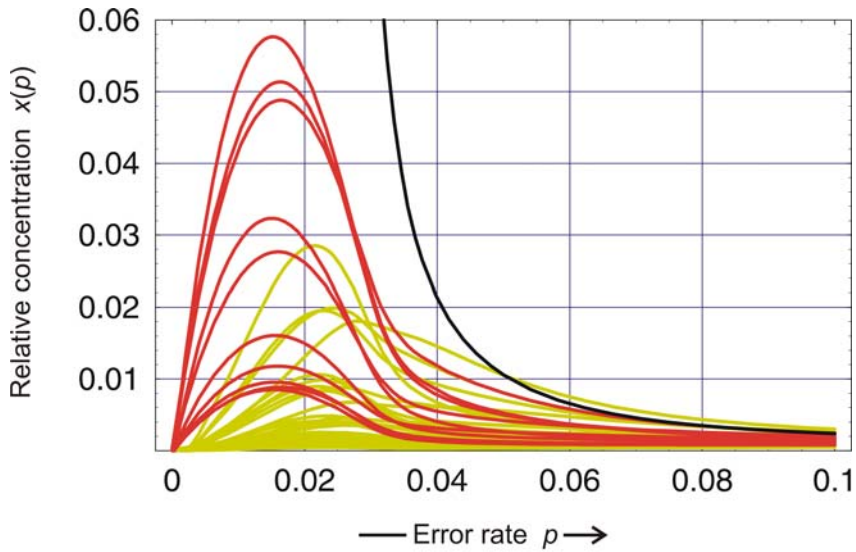
1. **Variation in fitness values**
2. Deviations from uniform error rates
3. Neutrality



Fitness landscapes showing error thresholds





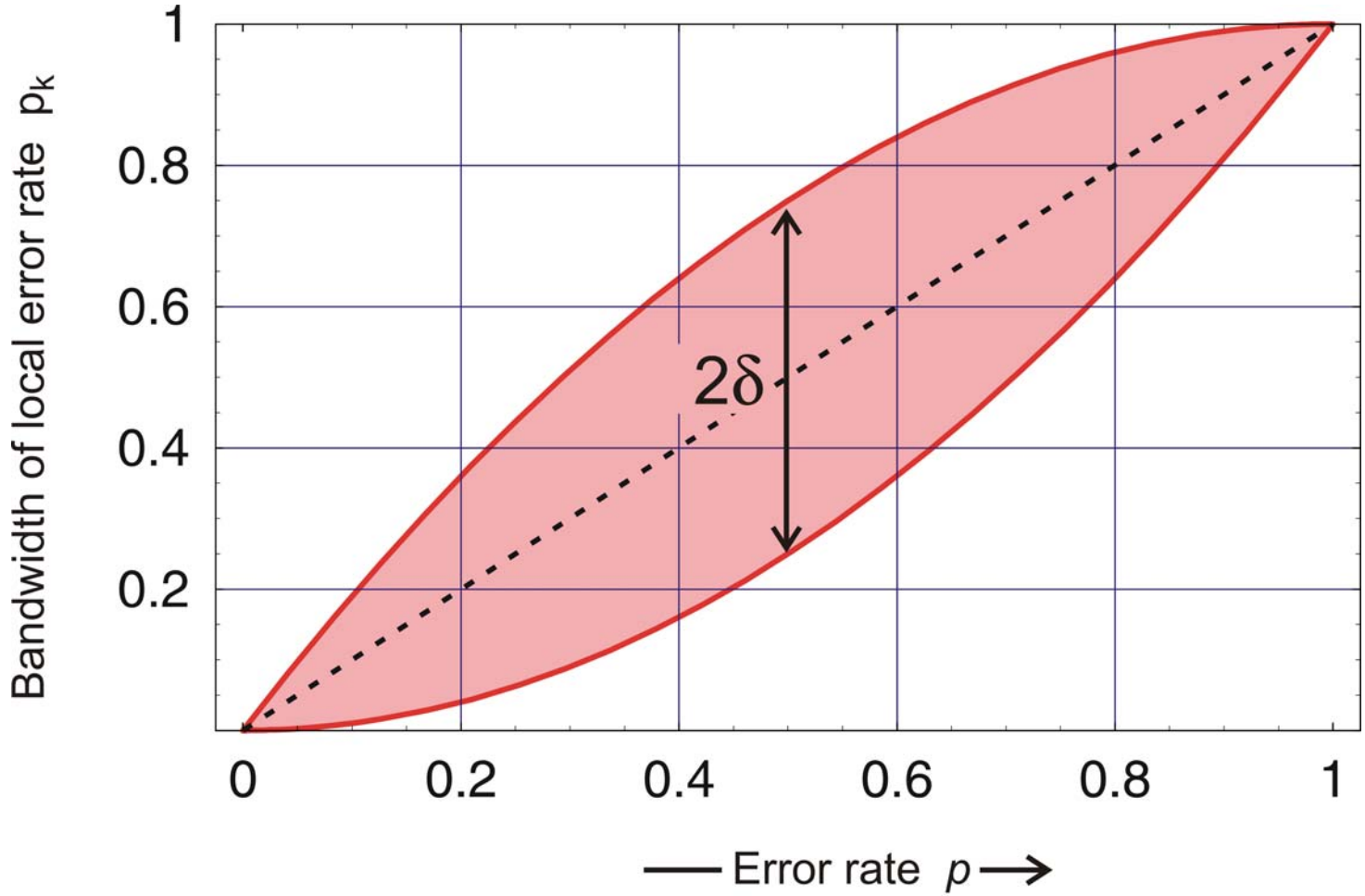


Error threshold: Individual sequences

n = 10, σ = 2 and d = 0, 1.0, 1.85

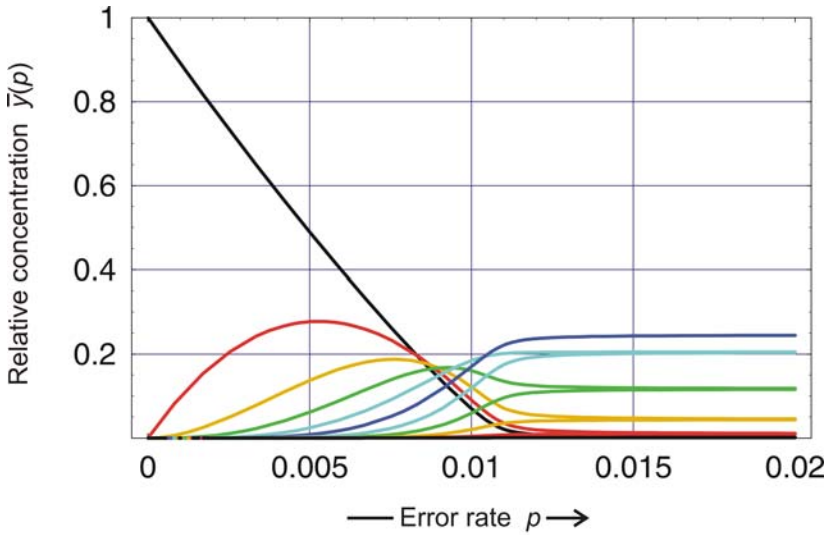
Features of realistic landscapes:

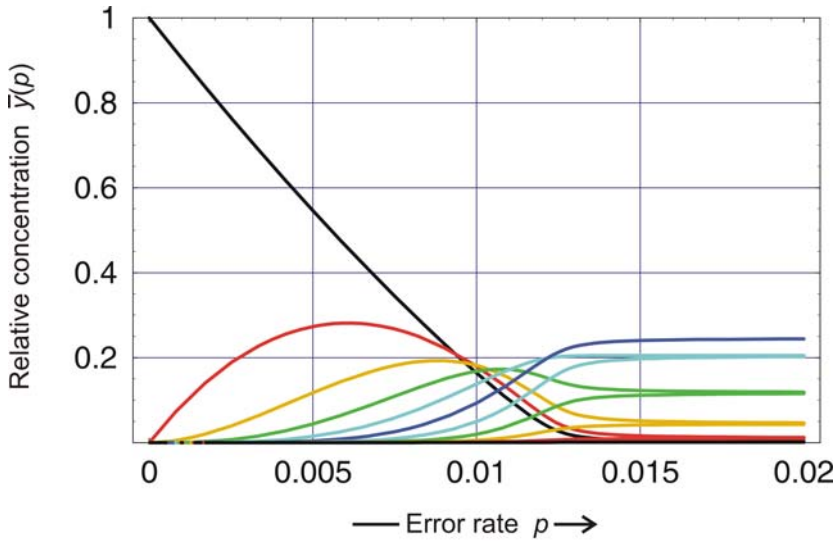
1. Variation in fitness values
2. **Deviations from uniform error rates**
3. Neutrality

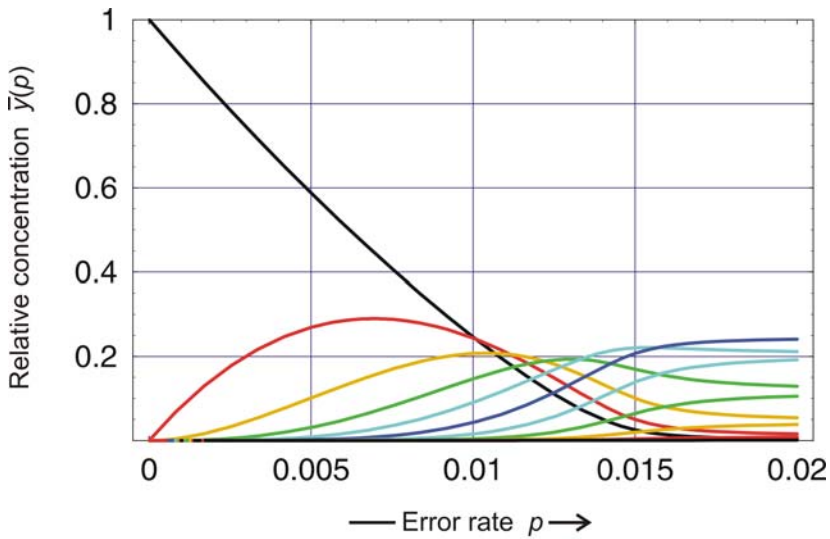


Local replication accuracy p_k:

$$p_k = p + 4\,\delta\,p(1-p)\,(X_{rnd}-0.5)\,,\quad k = 1,2,\ldots,2^{\nu}$$



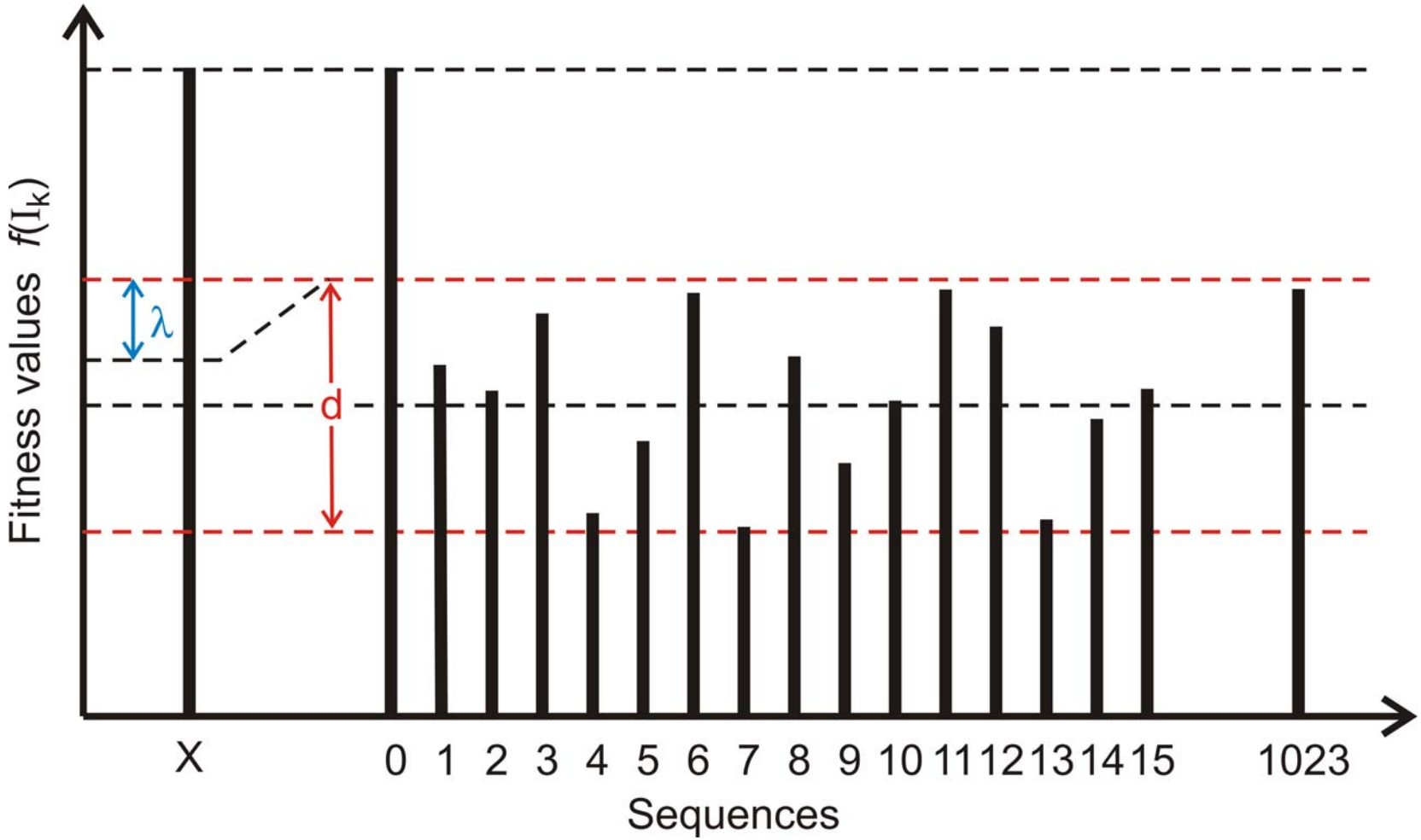




Error threshold: Classes

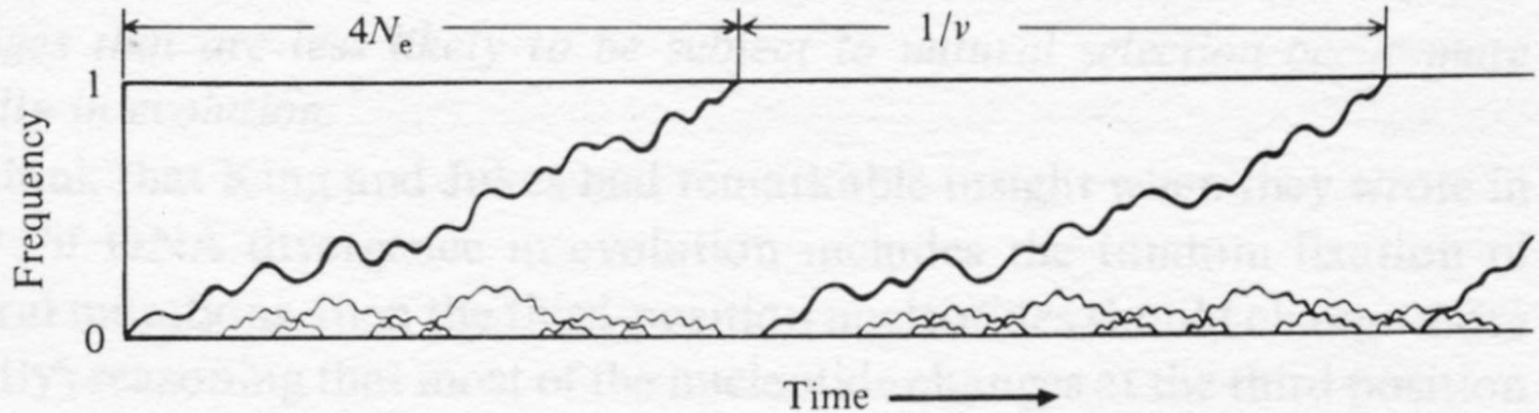
n = 10, σ = 1.1, δ = 0, 0.3, 0.5, and seed = 877

1. The origin of neutrality
2. RNA structures as a useful model
3. RNA replication and quasispecies
4. Selection on realistic landscapes
5. **Consequences of neutrality**
6. Evolutionary optimization of structure
7. The richness of conformational space



A fitness landscape including neutrality

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.



Motoo Kimura

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.

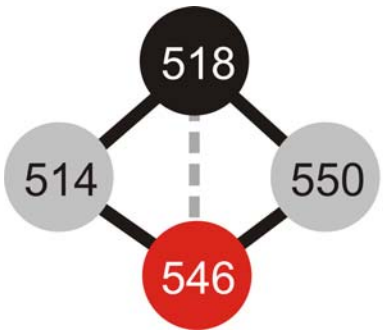
A kinetic theory of polynucleotide evolution which was developed during the past 15 years (Eigen, 1971; 1985; Eigen and Schuster, 1979; Eigen *et al.*, 1987; Schuster, 1986); Schuster and Sigmund, 1985) treats correct replication and mutation as parallel reactions within one and the same reaction network

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

$\lambda = 0.01,\ s = 877$

$\mathbf{d_H = 2}$

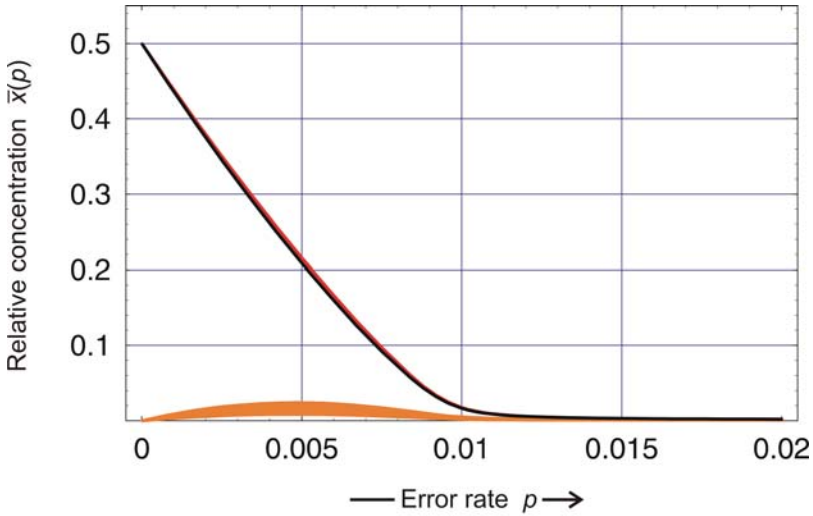
$$\lim_{p\to 0} x_1(p) = a$$

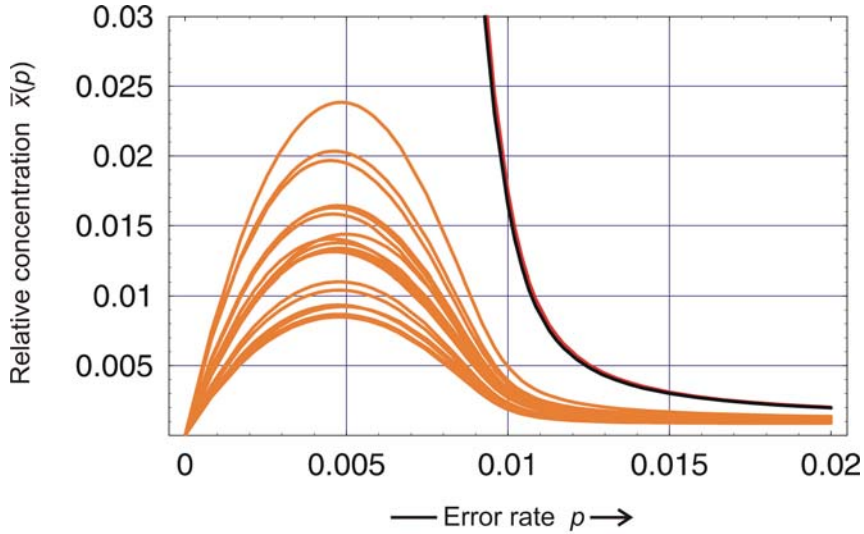
$$\lim_{p\to 0} x_2(p) = 1 - a$$

$\mathbf{d_H \geq 3}$

random fixation in the sense of Motoo Kimura

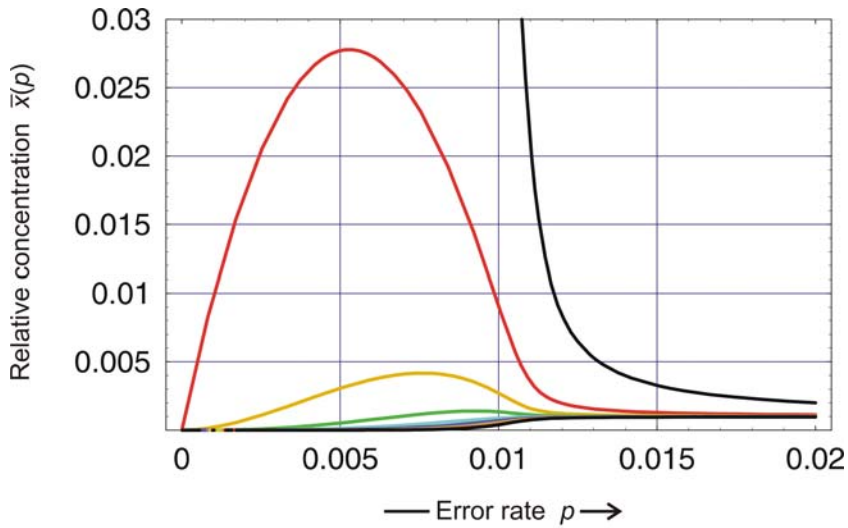
Pairs of genotypes in neutral replication networks





Neutral network

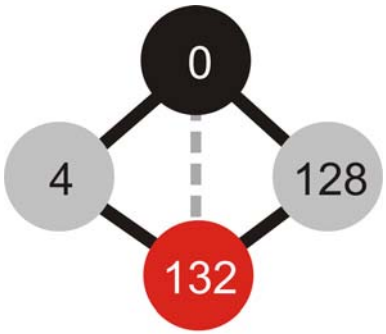
$\lambda = 0.01,\ s = 367$



Neutral network: Individual sequences

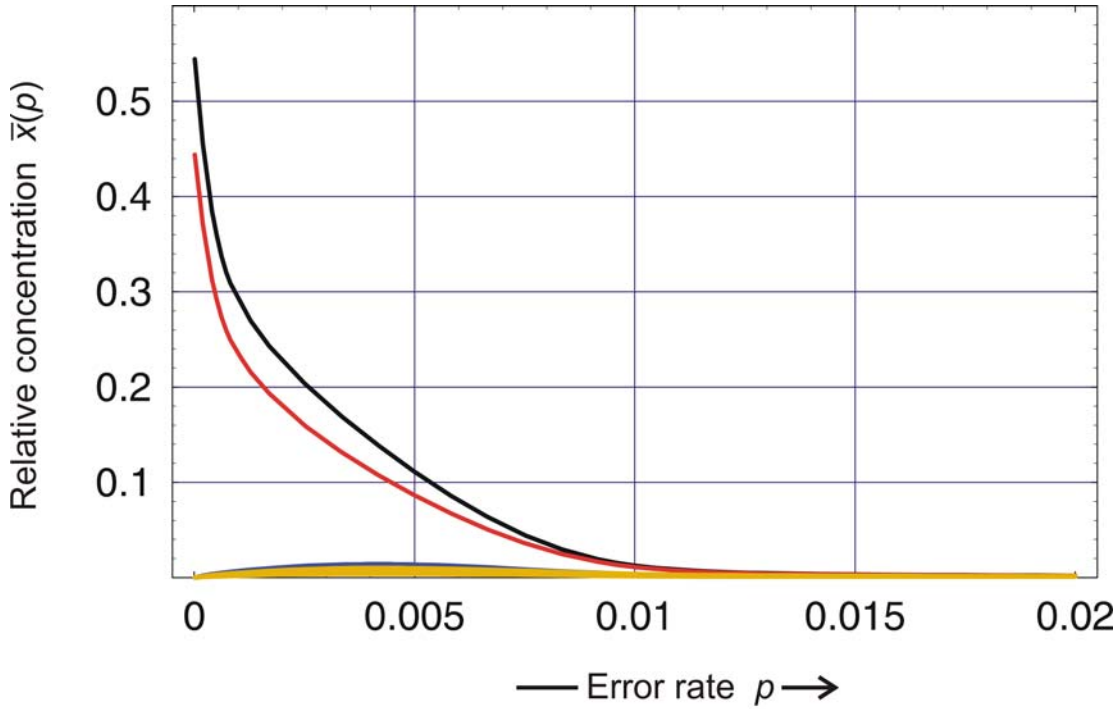
$n = 10,\ \sigma = 1.1,\ d = 1.0$

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



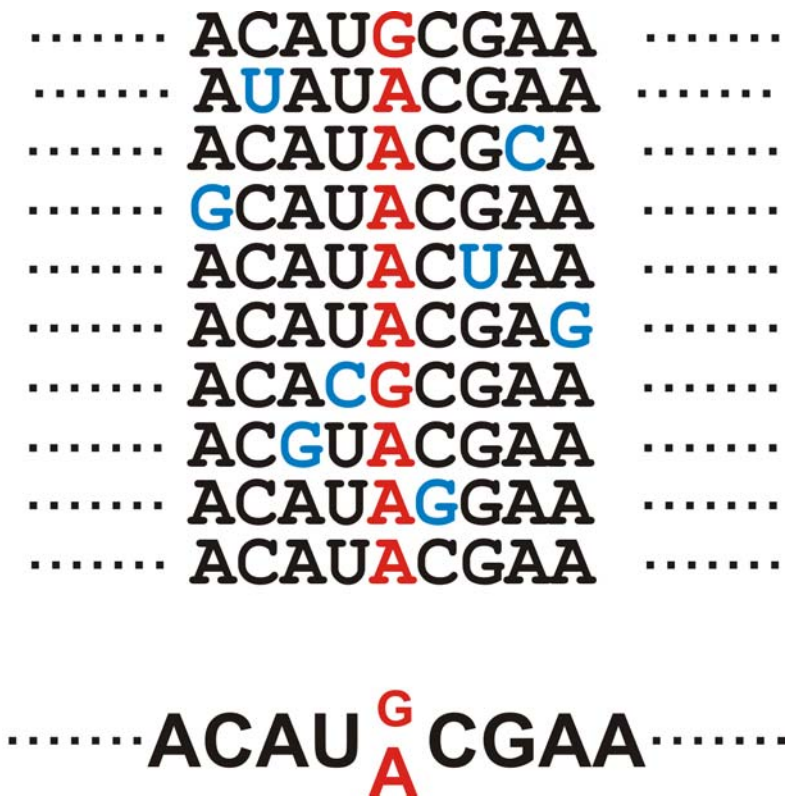
Neutral network

$\lambda = 0.01,\ s = 877$

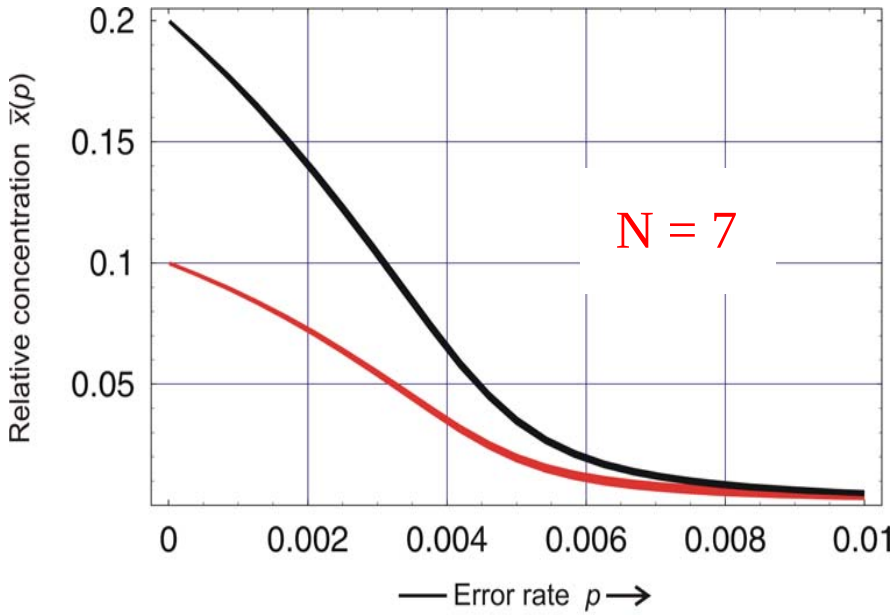


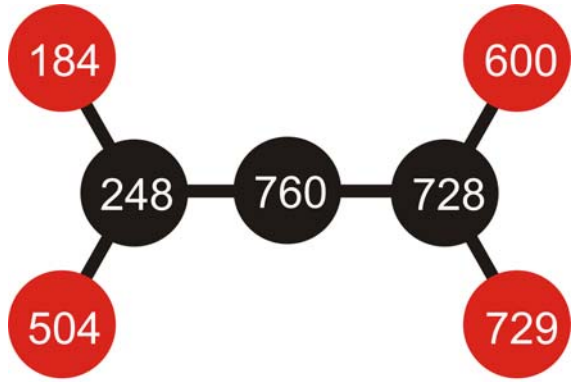
Neutral network: Individual sequences

n = 10, σ = 1.1, d = 1.0



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

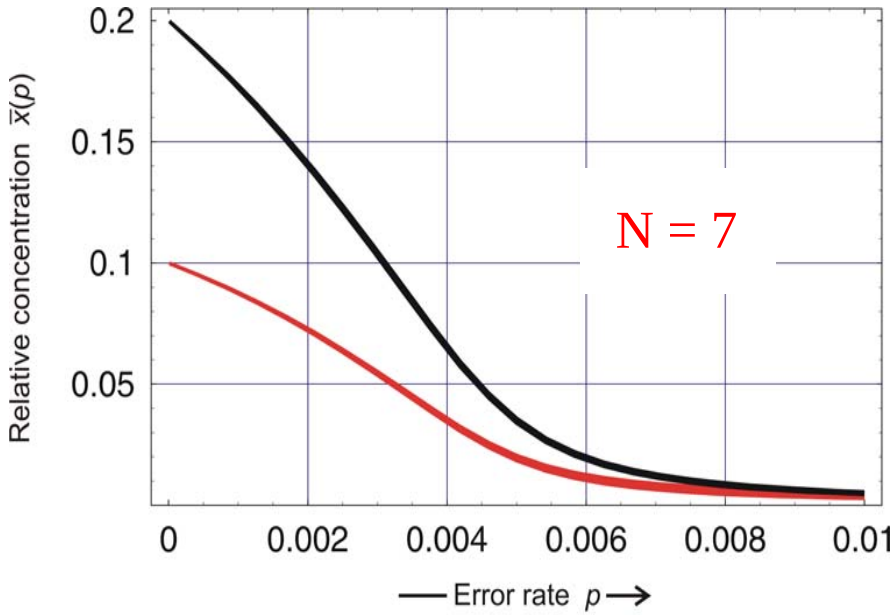


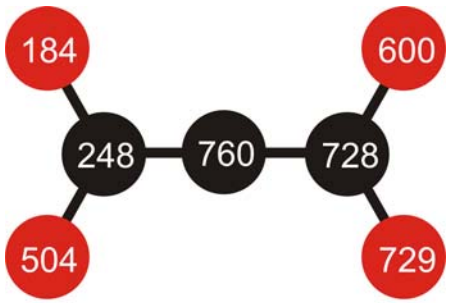


Neutral network

$\lambda = 0.10$, s = 229

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





Neutral network

$\lambda = 0.10,\ s = 229$

Perturbation matrix W

$$W = \begin{pmatrix} f & 0 & \varepsilon & 0 & 0 & 0 & 0 \\ 0 & f & \varepsilon & 0 & 0 & 0 & 0 \\ \varepsilon & \varepsilon & f & \varepsilon & 0 & 0 & 0 \\ 0 & 0 & \varepsilon & f & \varepsilon & 0 & 0 \\ 0 & 0 & 0 & \varepsilon & f & \varepsilon & \varepsilon \\ 0 & 0 & 0 & 0 & \varepsilon & f & 0 \\ 0 & 0 & 0 & 0 & \varepsilon & 0 & f \end{pmatrix}$$

Eigenvalues of W

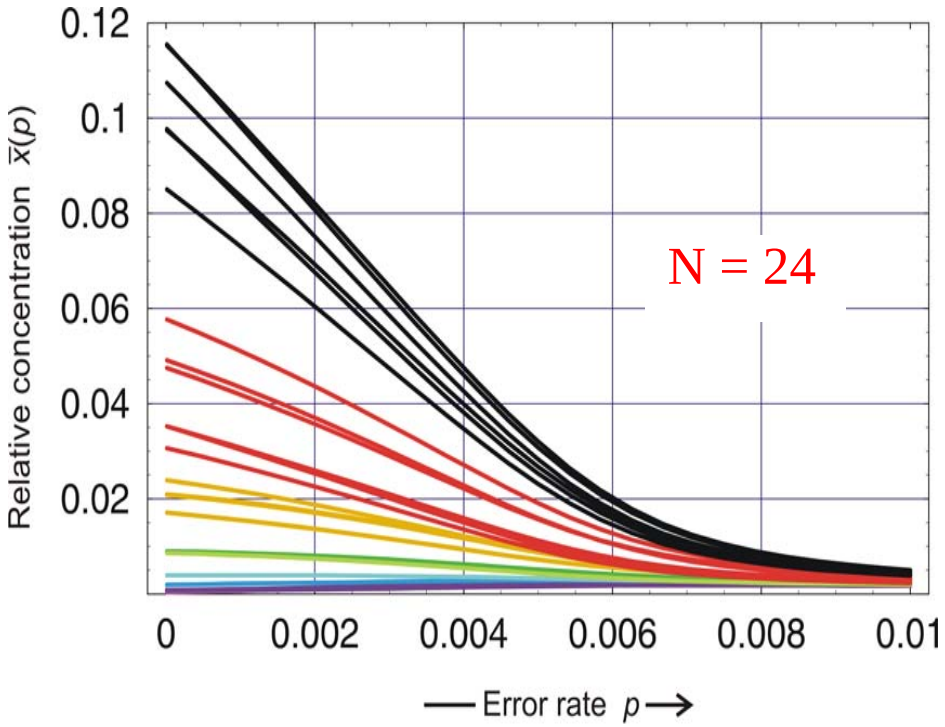
$$\lambda_0 = f + 2\varepsilon ,$$

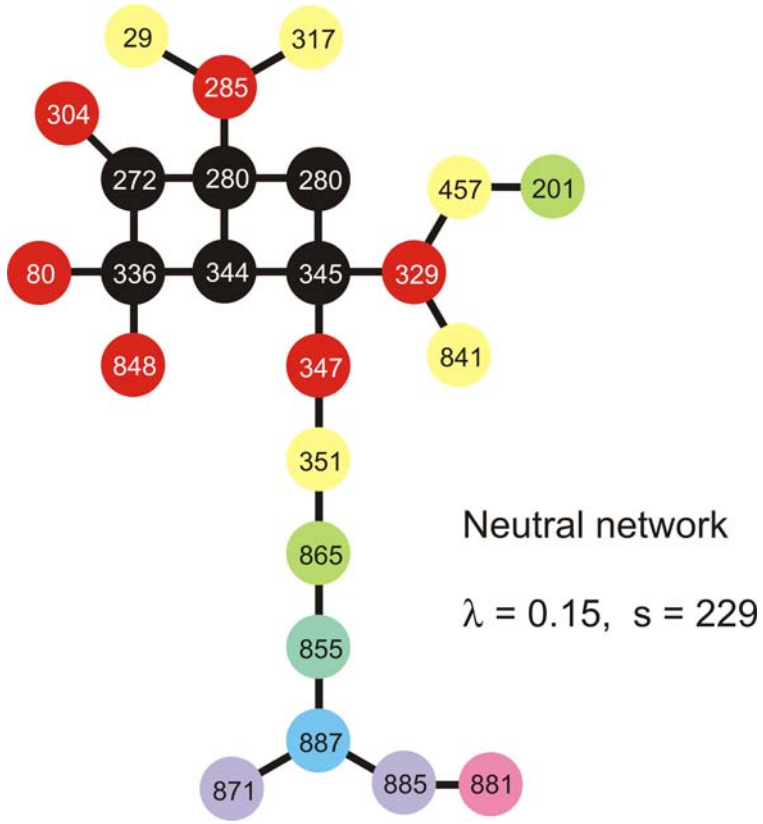
$$\lambda_1 = f + \sqrt{2}\varepsilon ,$$
$$\lambda_{2,3,4} = f ,$$
$$\lambda_5 = f - \sqrt{2}\varepsilon ,$$
$$\lambda_6 = f - 2\varepsilon .$$

Largest eigenvector of W

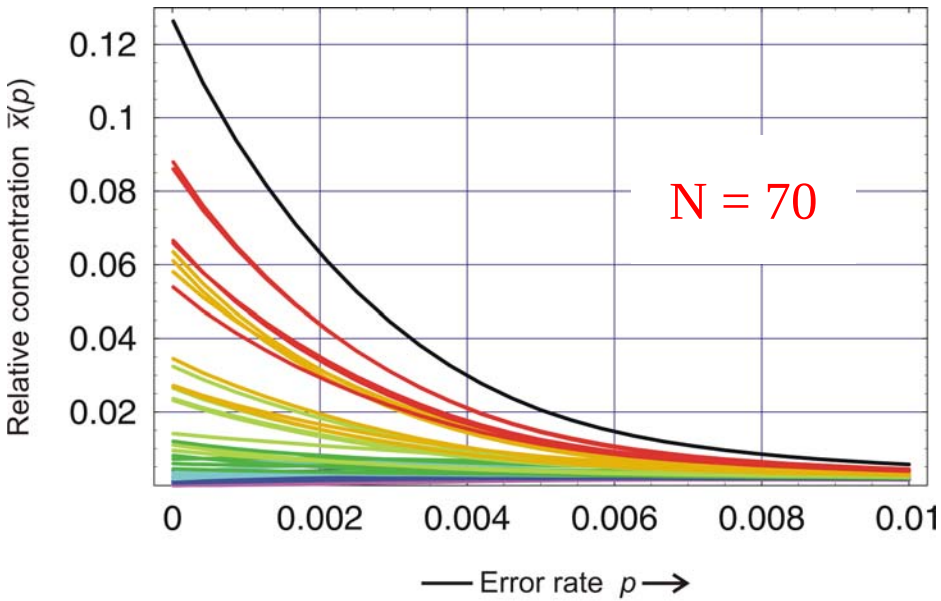
$$\xi_0 = (0.1, 0.1, 0.2, 0.2, 0.2, 0.1, 0.1) .$$

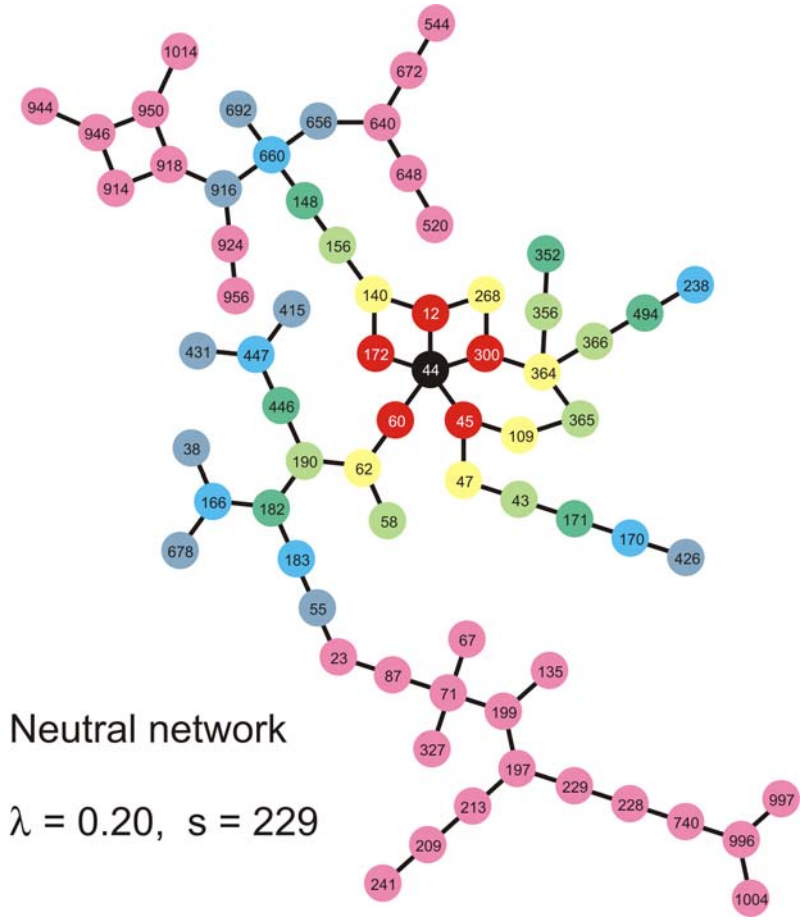
Neutral networks with increasing λ: λ = 0.10, s = 229





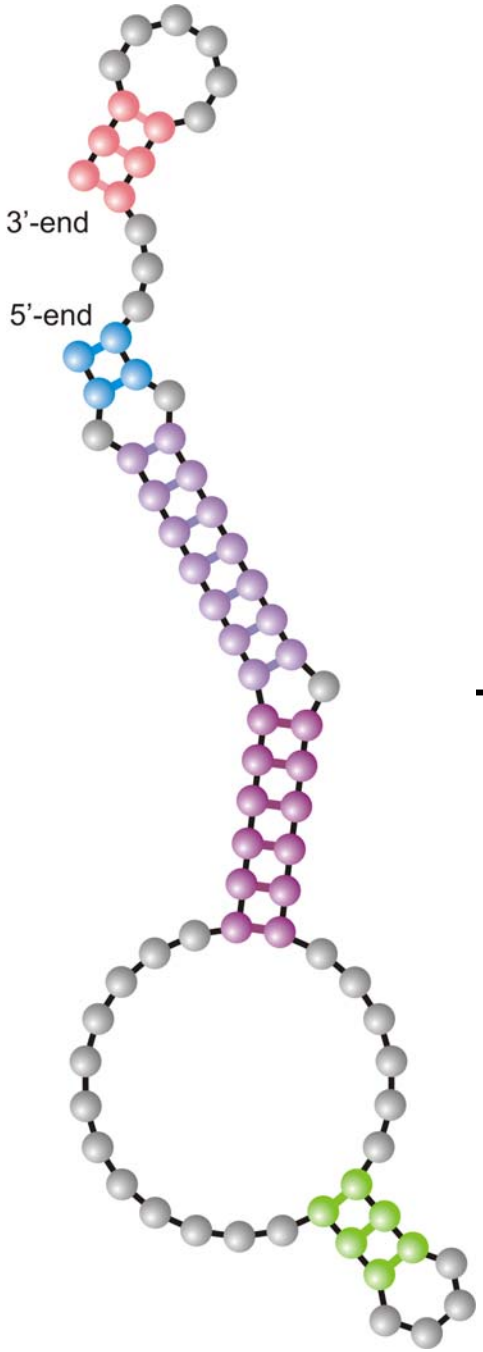
Neutral networks with increasing λ: λ = 0.15, s = 229



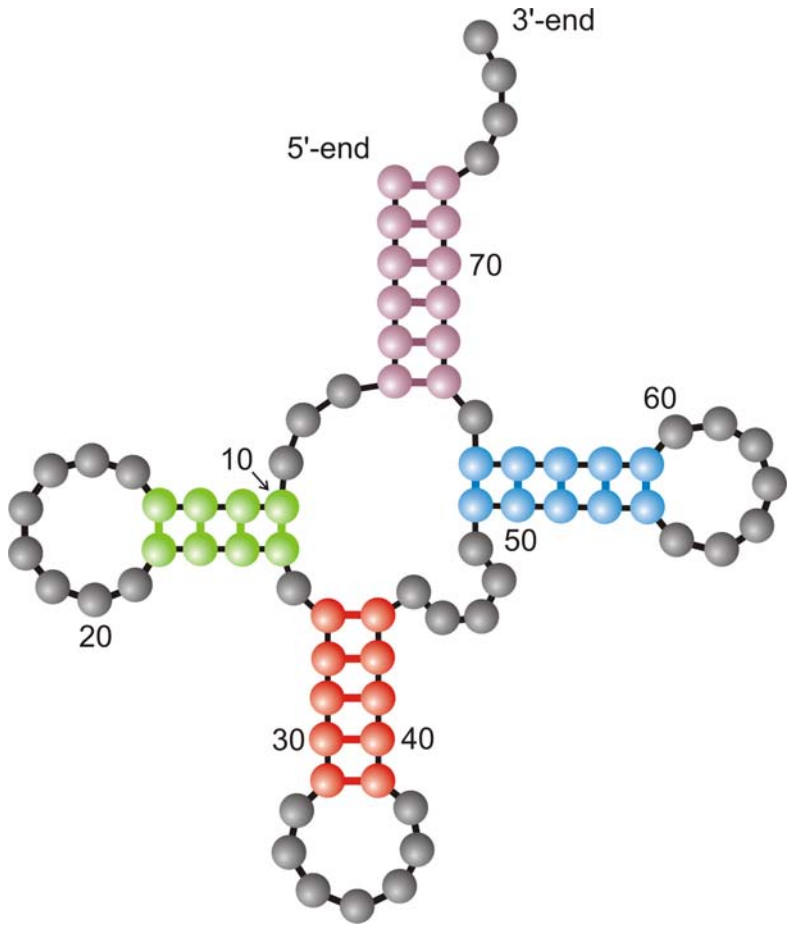


Neutral networks with increasing λ: λ = 0.20, s = 229

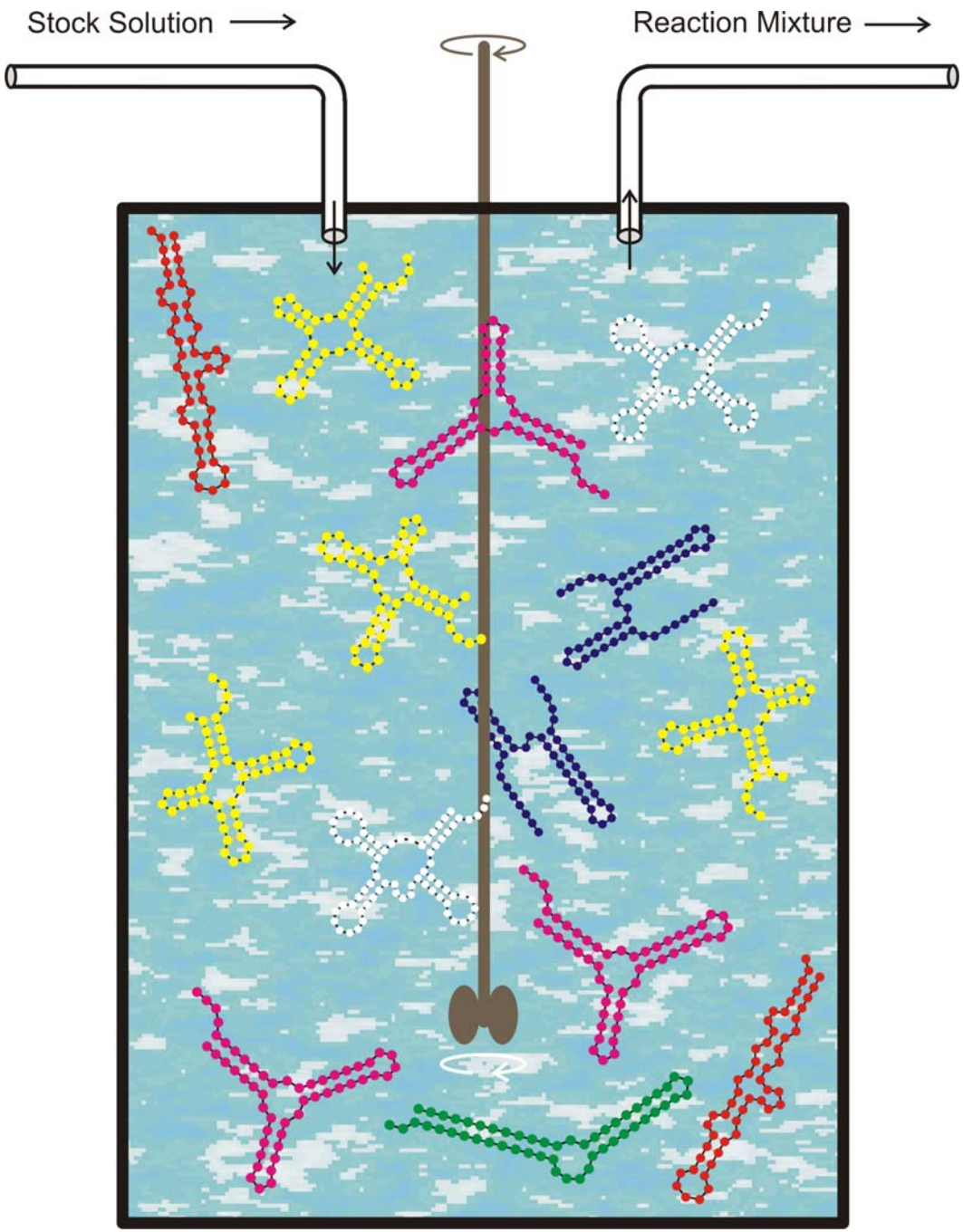
1. The origin of neutrality
2. RNA structures as a useful model
3. RNA replication and quasispecies
4. Selection on realistic landscapes
5. Consequences of neutrality
6. **Evolutionary optimization of structure**
7. The richness of conformational space



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,

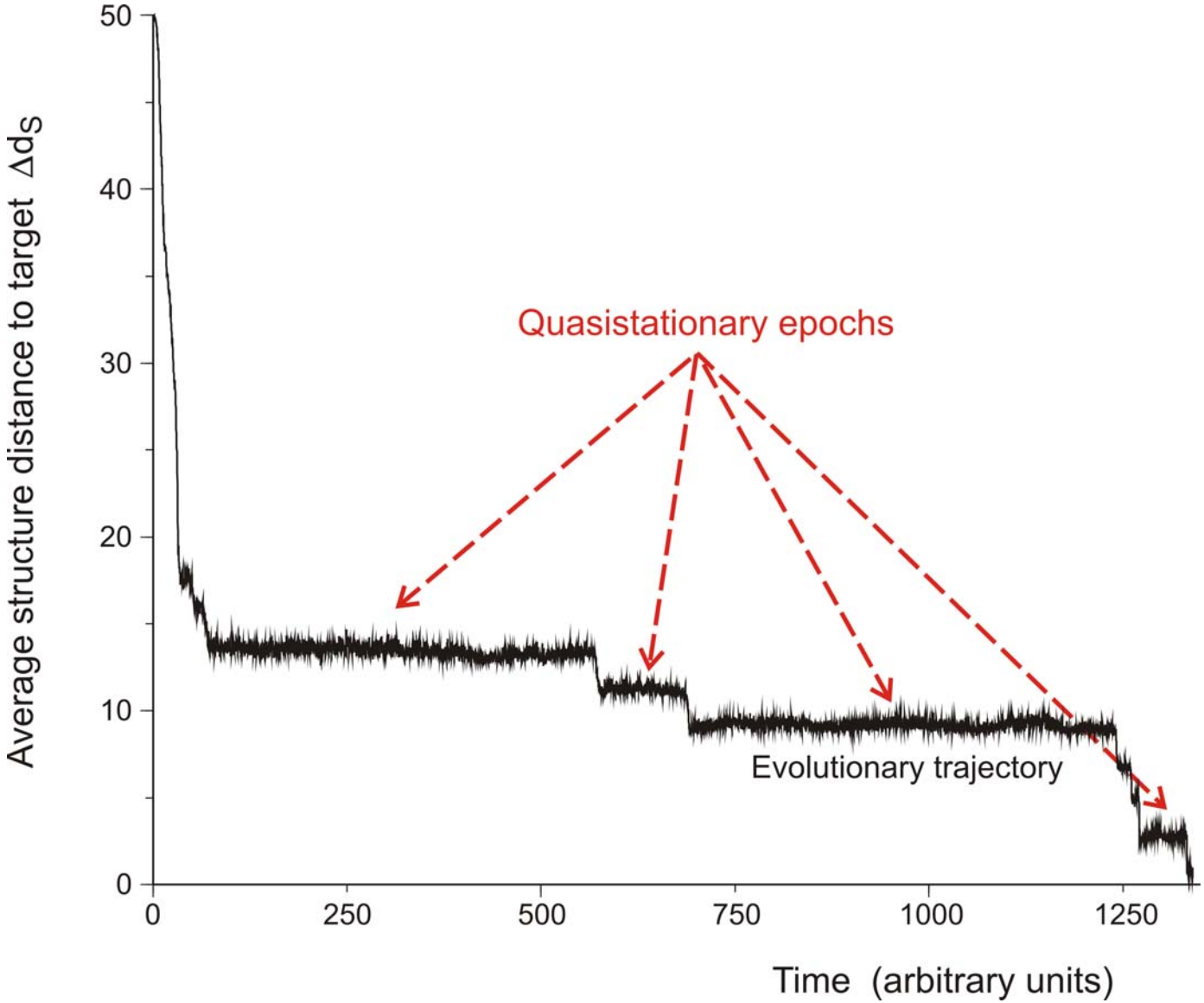
is determined by the flux:

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

Mutation rate:

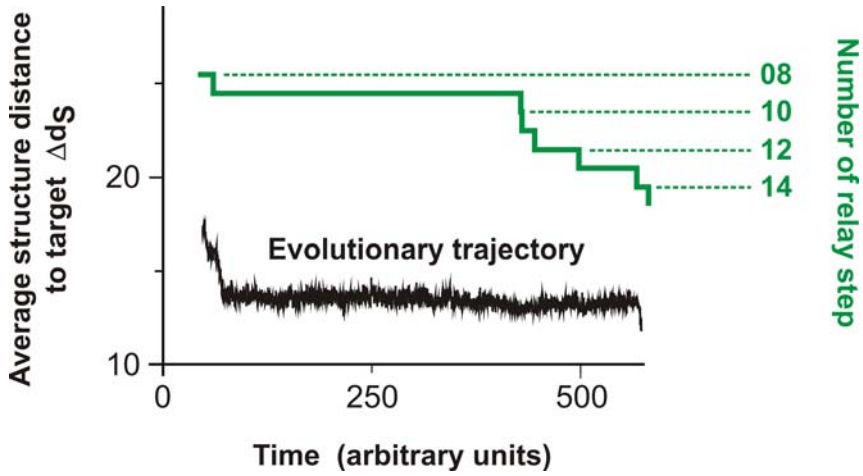
p = 0.001 / Nucleotide × Replication

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



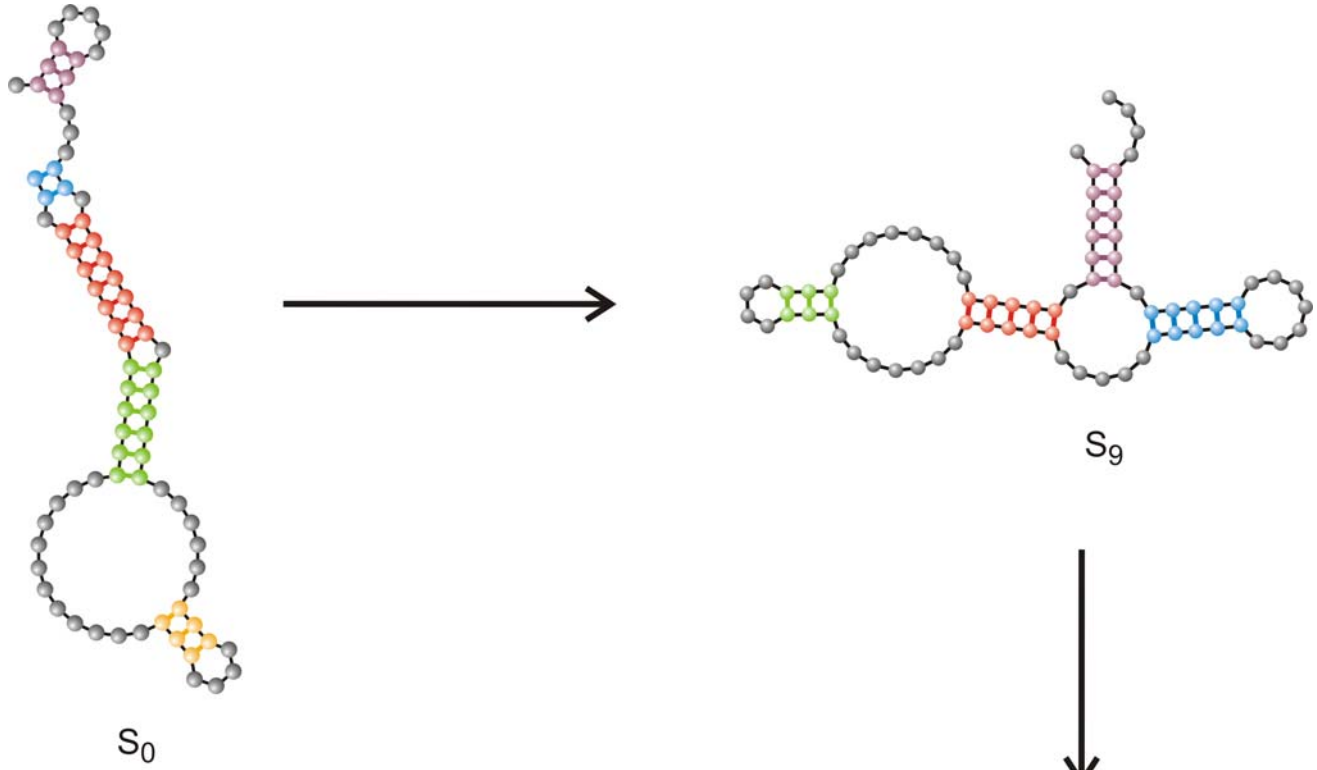

```
entry   GGUAUGGGCGUUGAAUAGUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUCCCGUACAGAA
8       .((((((((((((.........(((....)))).......)))))).....((((.......))))))))))....
exit    GGUAUGGGCGUUGAAUAAUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUCCCAUACAGAA
entry   GGUAUGGGCGUUGAAUAAUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUACCAUACAGAA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

Transition inducing point mutations change the molecular structure

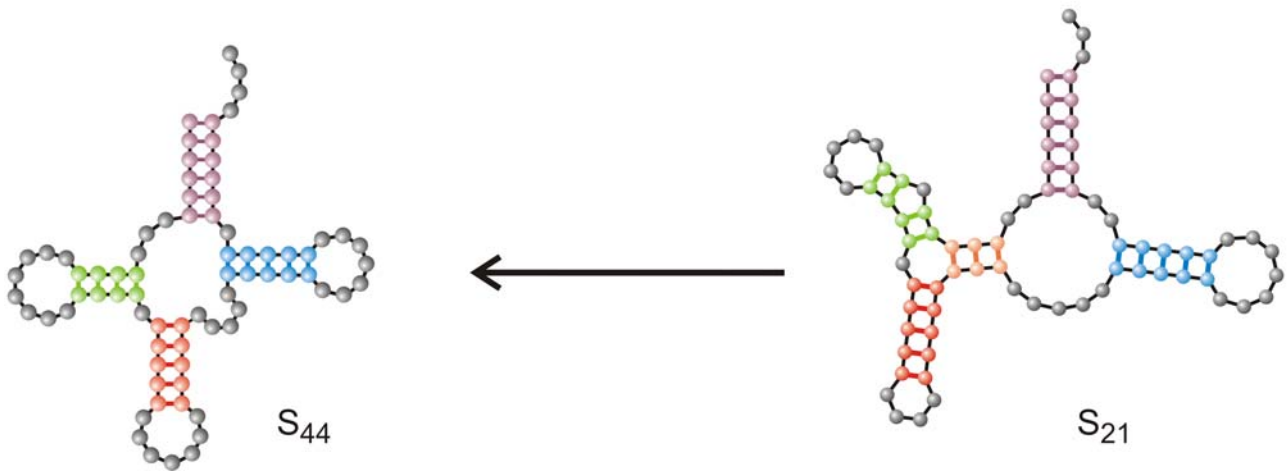
Neutral point mutations leave the molecular structure unchanged

Neutral genotype evolution during phenotypic stasis

Randomly chosen
initial structure



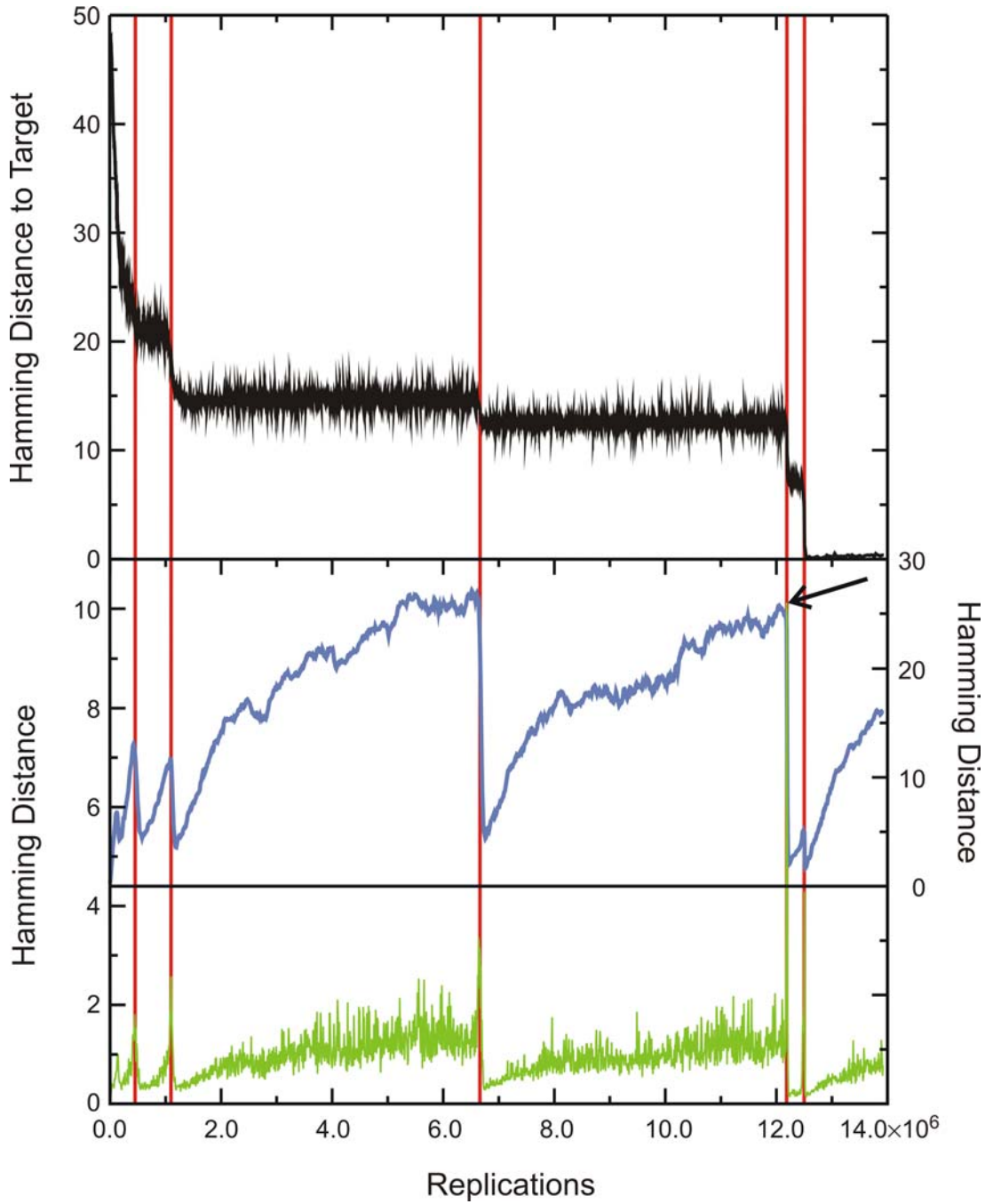
Phenylalanyl-tRNA
as target structure



Evolutionary trajectory

Spreading of the population on neutral networks

Drift of the population center in sequence space



Proc. Natl. Acad. Sci. USA
Vol. 93, pp. 397–401, January 1996
Evolution

Smoothness within ruggedness: The role of neutrality in adaptation

MARTIJN A. HUYNEN*†, PETER F. STADLER†‡, AND WALTER FONTANA†‡§

*Los Alamos National Laboratory, Theoretical Biology and Biophysics and Center for Nonlinear Studies, MS-B258, Los Alamos, NM 87545; †Institut für Theoretische Chemie, Universität Wien, Währingerstrasse 17, A-1090 Vienna, Austria; and ‡Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501

Communicated by Hans Frauenfelder, Los Alamos National Laboratory, Los Alamos, NM, September 20, 1995 (received for review June 29, 1995)

ABSTRACT **RNA secondary structure folding algorithms predict the existence of connected networks of RNA sequences with identical structure. On such networks, evolving populations split into subpopulations, which diffuse independently in sequence space. This demands a distinction between two mutation thresholds: one at which genotypic information is lost and one at which phenotypic information is lost. In between, diffusion enables the search of vast areas in genotype space while still preserving the dominant phenotype. By this dynamic the success of phenotypic adaptation becomes much less sensitive to the initial conditions in genotype space.**

To explain the high fixation rate of nucleotide substitutions in a population, Kimura (1) argued that the vast majority of genetic change at the level of a population must be neutral rather than adaptive. Sewall Wright's reaction to Kimura's point was politely neutral (ref. 2, p. 474): "Changes in wholly nonfunctional parts of the molecule would be the most frequent ones but would be unimportant, unless they occasionally give a basis for later changes which improve function in the species in question which would then become established by selection." Today, in view of the data generated by comparative sequence analysis, the surprise is no longer over the existence of neutrality but over how little conservation there is at the sequence level (3–6). This makes Wright's point even more pertinent. How are we to imagine the relation between neutral evolution and adaptation? An answer to this question requires a model of the relationship between genotype and phenotype. Such a model is available for RNA secondary structure. The latter can be computed from the sequence by means of procedures based on thermodynamic data which have become standard in the past 15 years (7, 8). Secondary structure covers the major share of the free energy of tertiary structure formation and is frequently used to interpret RNA function and evolutionary data. As such, the case is a qualitatively important one.

Robust Properties of RNA Folding

The mapping from sequences to secondary structures is many to one for two reasons: (*i*) there are many more sequences than secondary structures, and (*ii*) some structures are realized much more frequently than others (9). Call two sequences connected if they differ by one or at most two point mutations. A neutral network, then, is a set of sequences with identical structure so that each sequence is connected to at least one other sequence. The crucial point for our discussion comes from a recent study of the standard secondary structure prediction algorithm (9), which showed that such networks exist and that for frequent structures these networks percolate through sequence space. For example, starting at a sequence that folds into a tRNA structure, it is possible to traverse sequence space along a connected path, thus changing every nucleotide position without ever changing the structure. Moreover, due to the high-dimensionality of sequence space, networks of frequent structures penetrate each other so that each frequent structure is almost always realized within a small distance of any random sequence. These features seem to be intrinsic to RNA folding, since they are insensitive to whether the folding algorithm is thermodynamic, kinetic, or maximum matching (E. Bornberg-Bauer, M. Tacker, and P. Schuster, personal communication) or whether one considers one minimum free energy structure or the entire Boltzmann ensemble (10).

A Simple Model for Test Tube Evolution

To assess the consequences of these properties for molecular evolution, we study a model in which the replication rate (fitness) of an RNA sequence depends on its secondary structure. Our folding procedure¶ is a speed-tuned implementation of the Zuker–Stiegler algorithm (8). The model consists of a population of RNA sequences of fixed length ν, which replicate and mutate in a stirred flow reactor. RNA populations manageable in the computer or in the laboratory are tiny compared to the size of the sequence space (4^ν), and a correct simulation must, therefore, resort to stochastic chemical reaction kinetics (11, 12). A selection pressure is induced by a dilution flow, which adjusts over time to keep the total RNA population fluctuating around a constant capacity N (11, 13). This setup mimics Spiegelman's serial transfer technique (14), where sequences with a replication rate above (below) the average increase (decrease) in concentration.

When a sequence undergoes a replication, each base is copied with fidelity $1 - p$. The overall replication rate of an individual sequence is defined to be a function of the distance (9, 30) between its secondary structure and a predefined target structure. Here the target structure is the $tRNA^{Phe}$ cloverleaf, but the structure of any randomly chosen sequence would do as well. This corresponds to the artificial *in vitro* selection of a structure with some desired function or affinity to a target (14–21). A similar situation, though with proteins and not RNA, occurs in the affinity maturation of the immune response (22). In both artificial and natural selection there are two sources of neutrality: one is the sequence (genotype) to structure (phenotype) mapping, and the other is the structure to replication rate (fitness) mapping. It is the former source that is central to this discussion. Notice, thus, that in the present model the second source of neutrality arises only for sequences whose structures differ from the target.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

§To whom reprint requests should be addressed at Institut für Theoretische Chemie, Währinger Strasse 17, A-1090 Vienna, Austria.

¶Hofacker, I. L., Fontana, W., Stadler, P. F., and Schuster, P. RNA folding package available by anonymous ftp from ftp.itc.univie.ac.at in/pub/RNA.

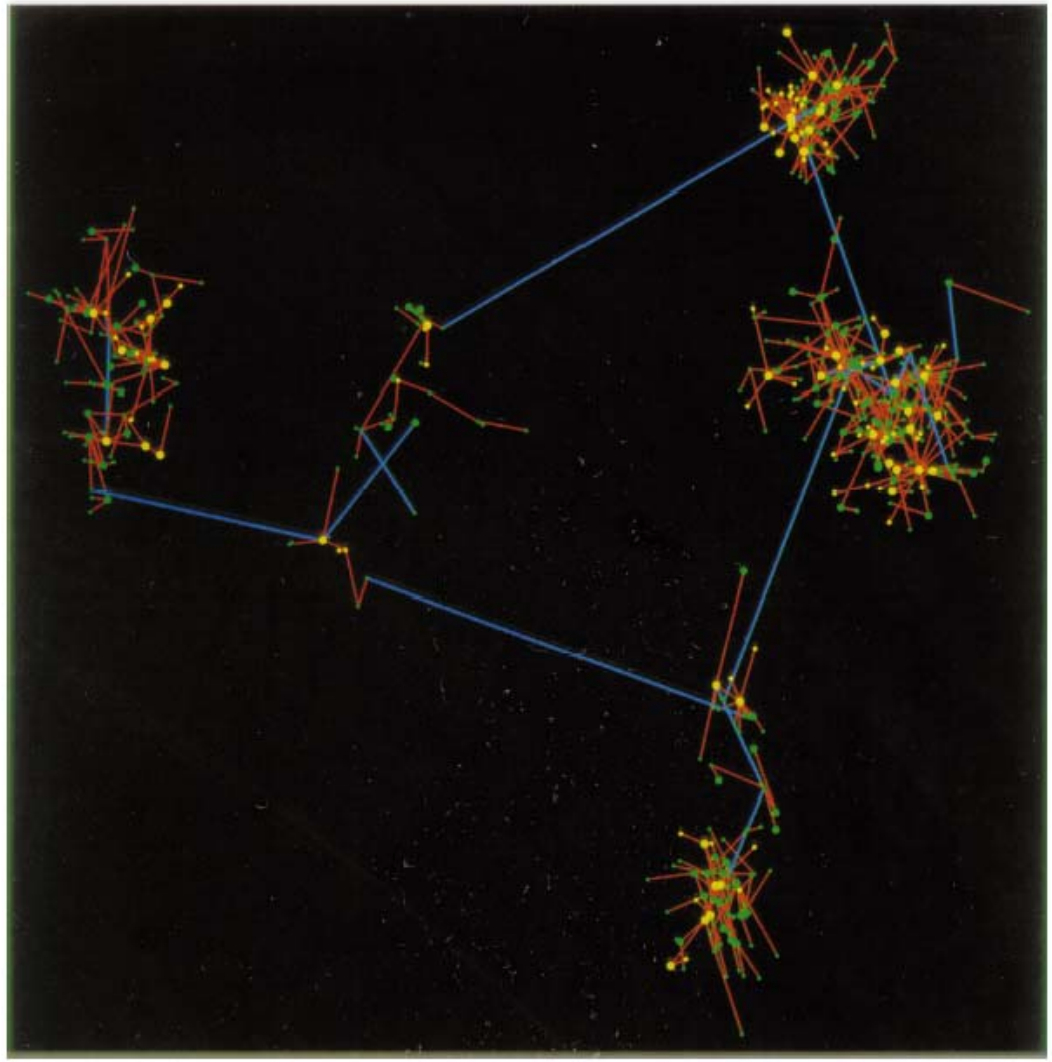
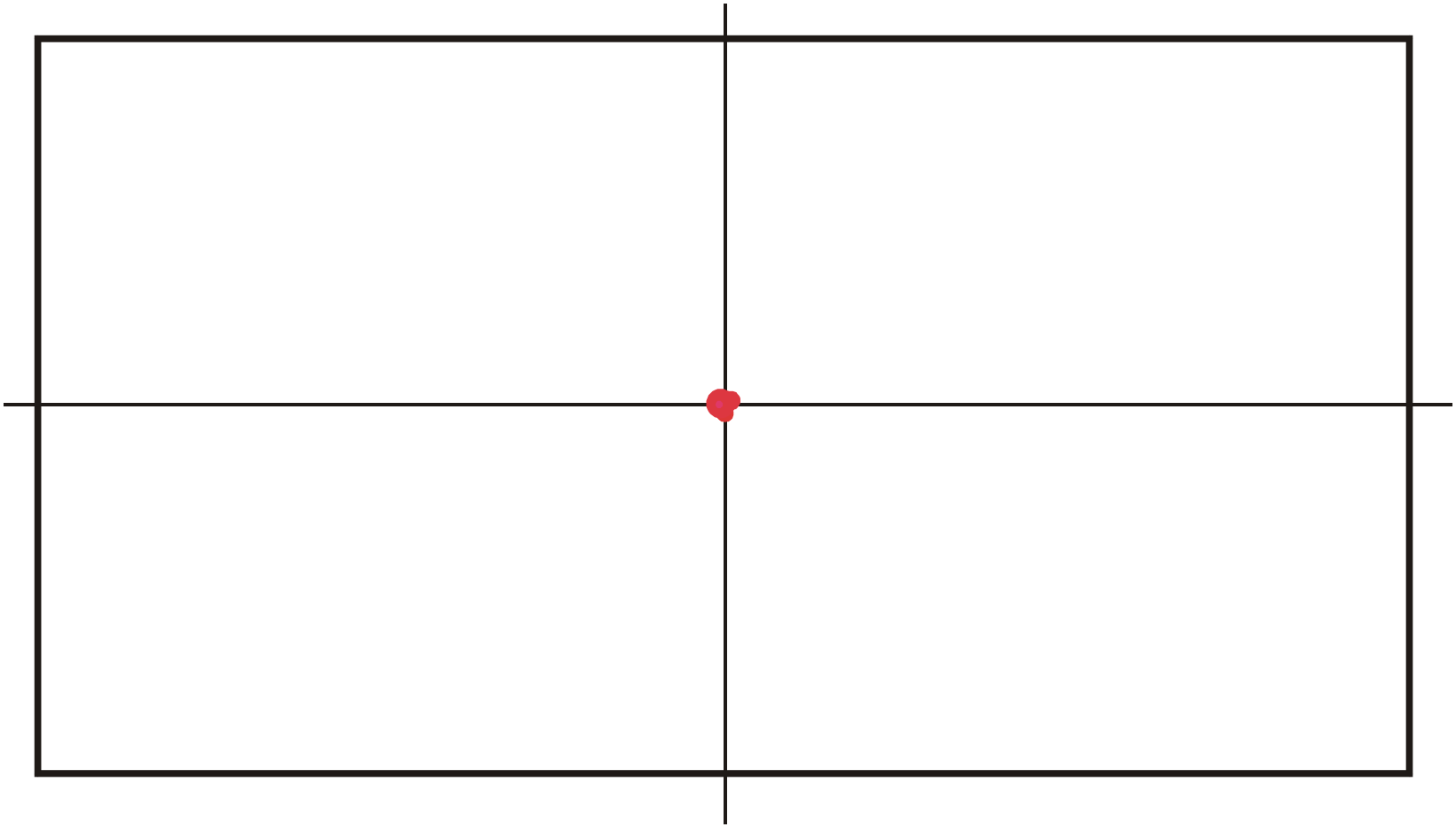
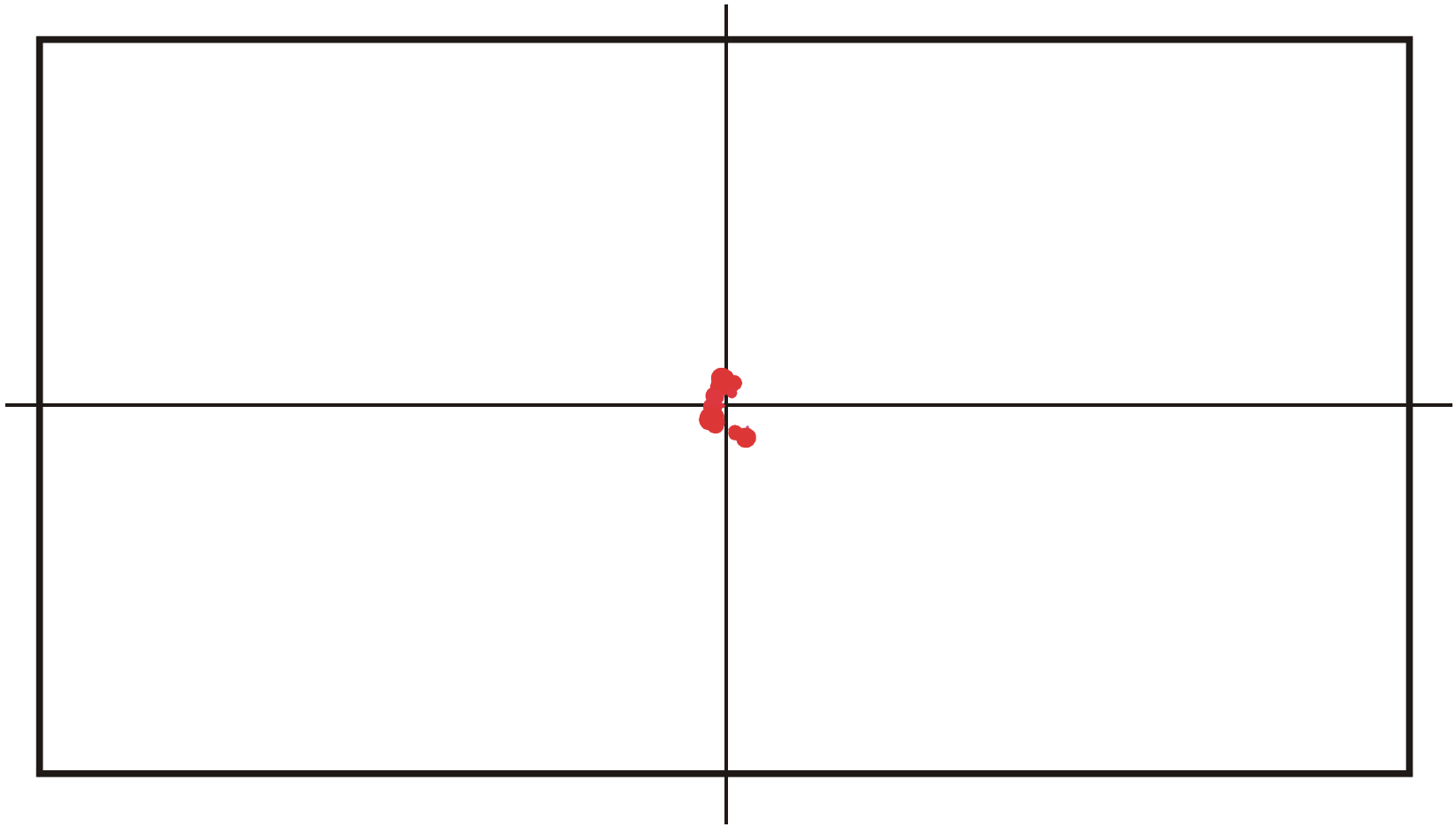


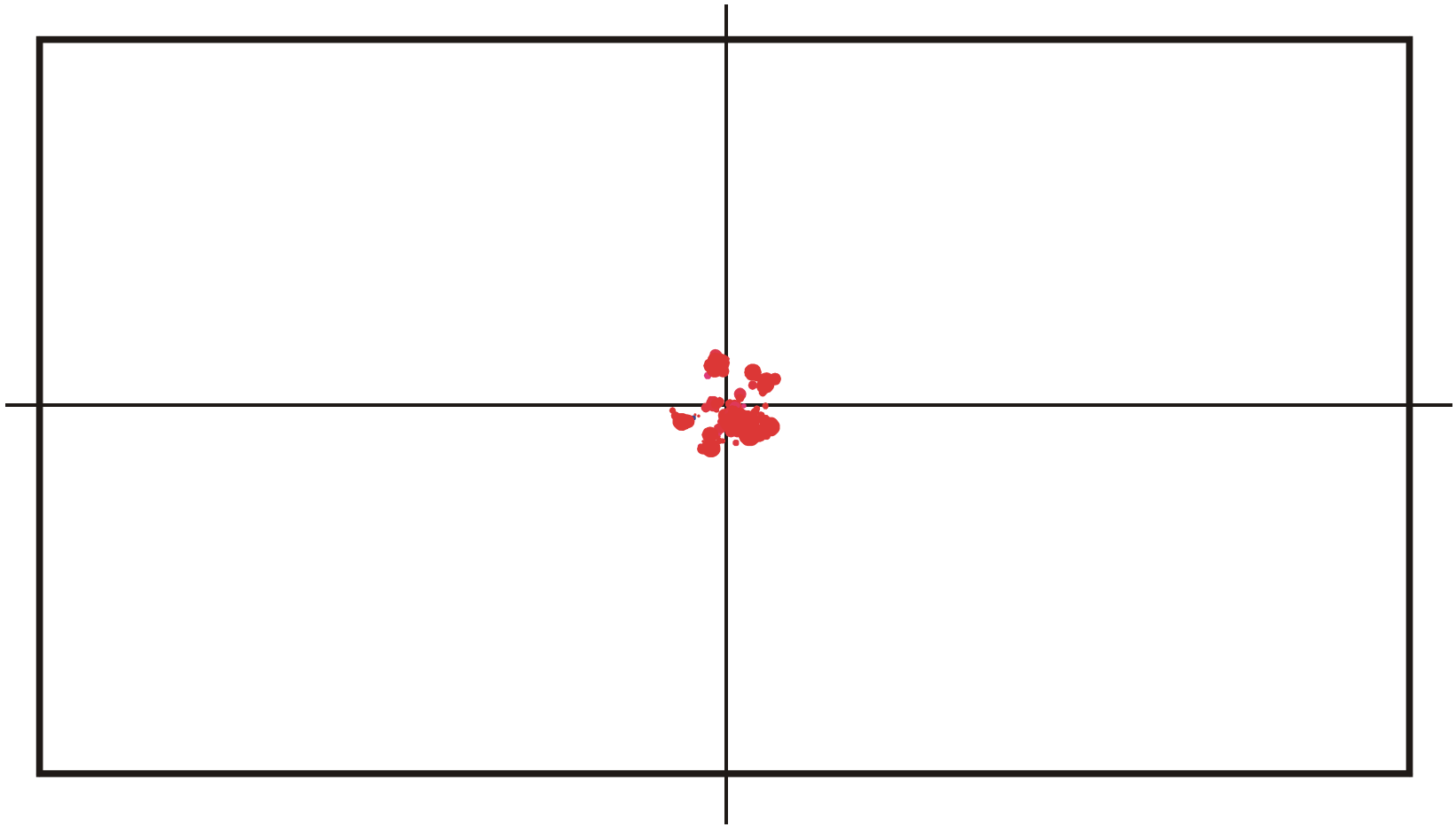
FIG. 2. Population structure in sequence space. The support of a population in sequence space is the set of sequences present in at least one copy. The population support can be pictured in two dimensions using some theorems from distance geometry (27). We compute the metric matrix M with entries $m_{ij} = (d_{0i}^2 + d_{0j}^2 - d_{ij}^2)/2$, where d_{ij} is the Hamming distance between sequences i and j and 0 is the center of mass of the support. Sequences are expressed in principal axes coordinates by diagonalizing M. Only the components corresponding to the largest two eigenvalues are kept, yielding a projection onto the plane that captures most of the variation. Dots represent a static snapshot of $N = 2000$ individuals after 135 time units replicating with $p = 0.002$. Among the 2000 individuals, 631 are different and among them 301 fold into different structures. To help correct for the distortions of the projection, the dots are connected by the edges of the minimum spanning tree. Edges connect closest points. Red (blue), Hamming distance less (more) than 6; dot size large (small), more (less) than four copies in the population; yellow (green), sequences that do (do not) fold into the tRNA target structure.



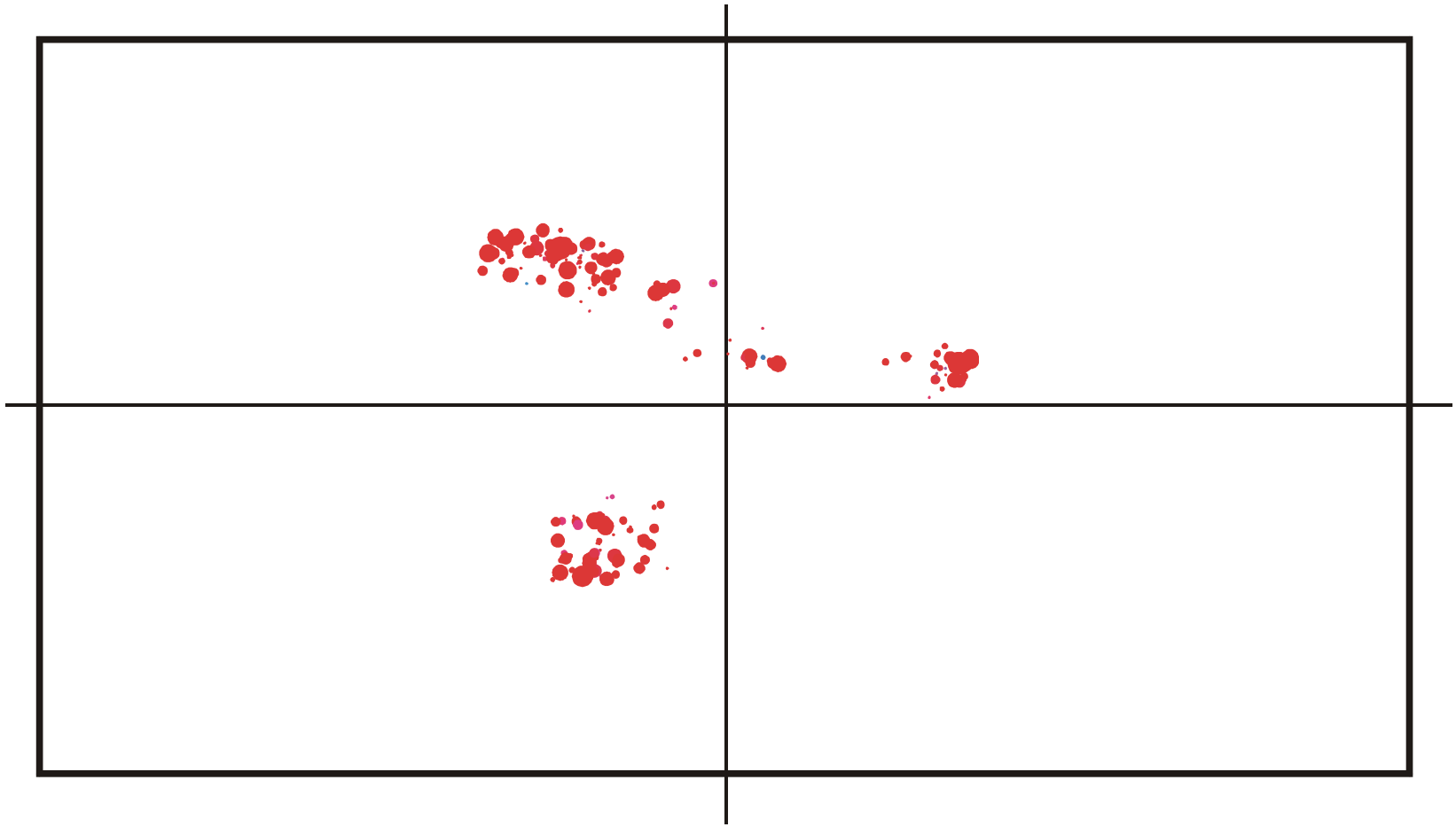
Spreading and evolution of a population on a neutral network: t = 150



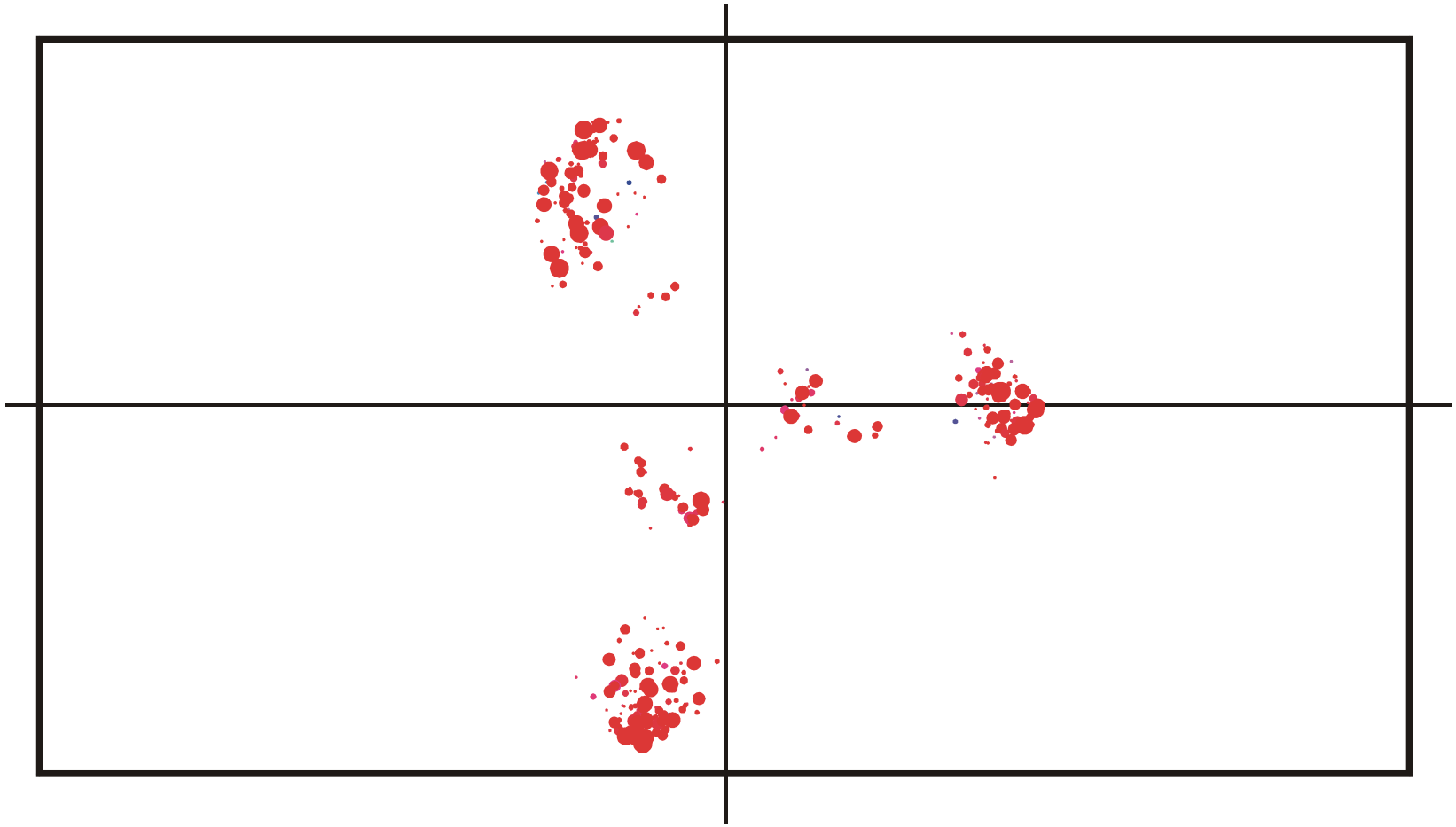
Spreading and evolution of a population on a neutral network : t = 170



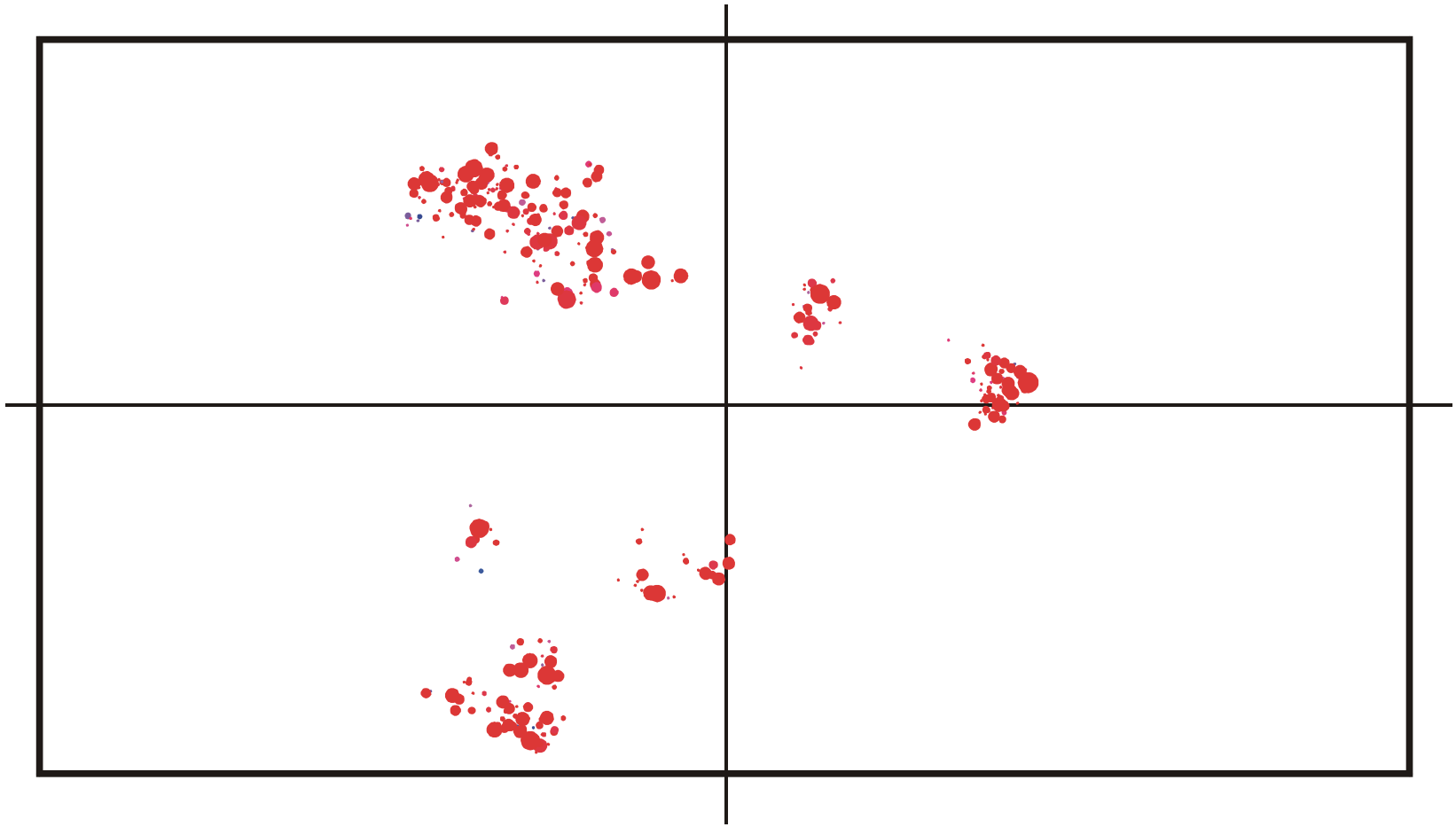
Spreading and evolution of a population on a neutral network : t = 200



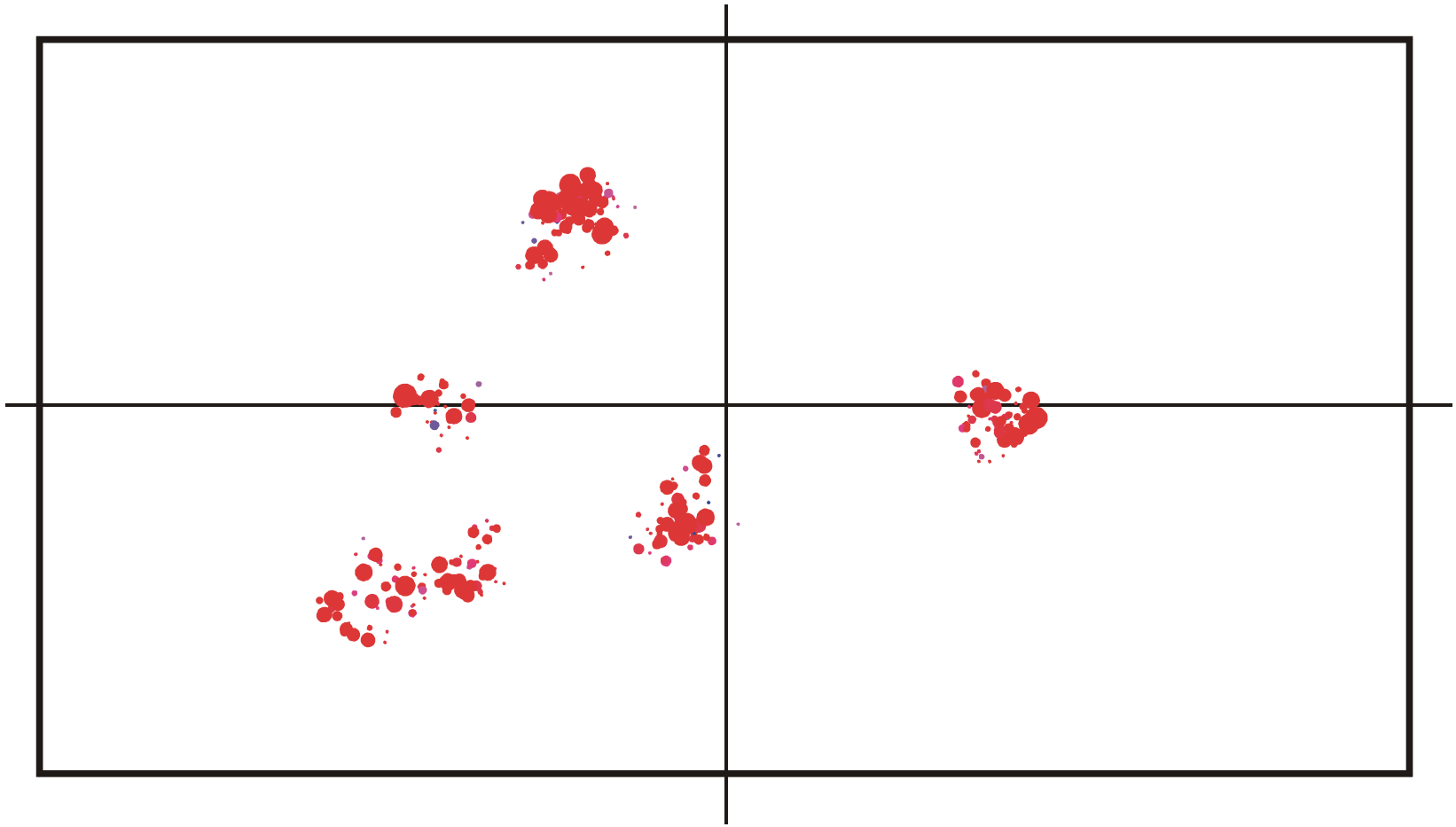
Spreading and evolution of a population on a neutral network : t = 350



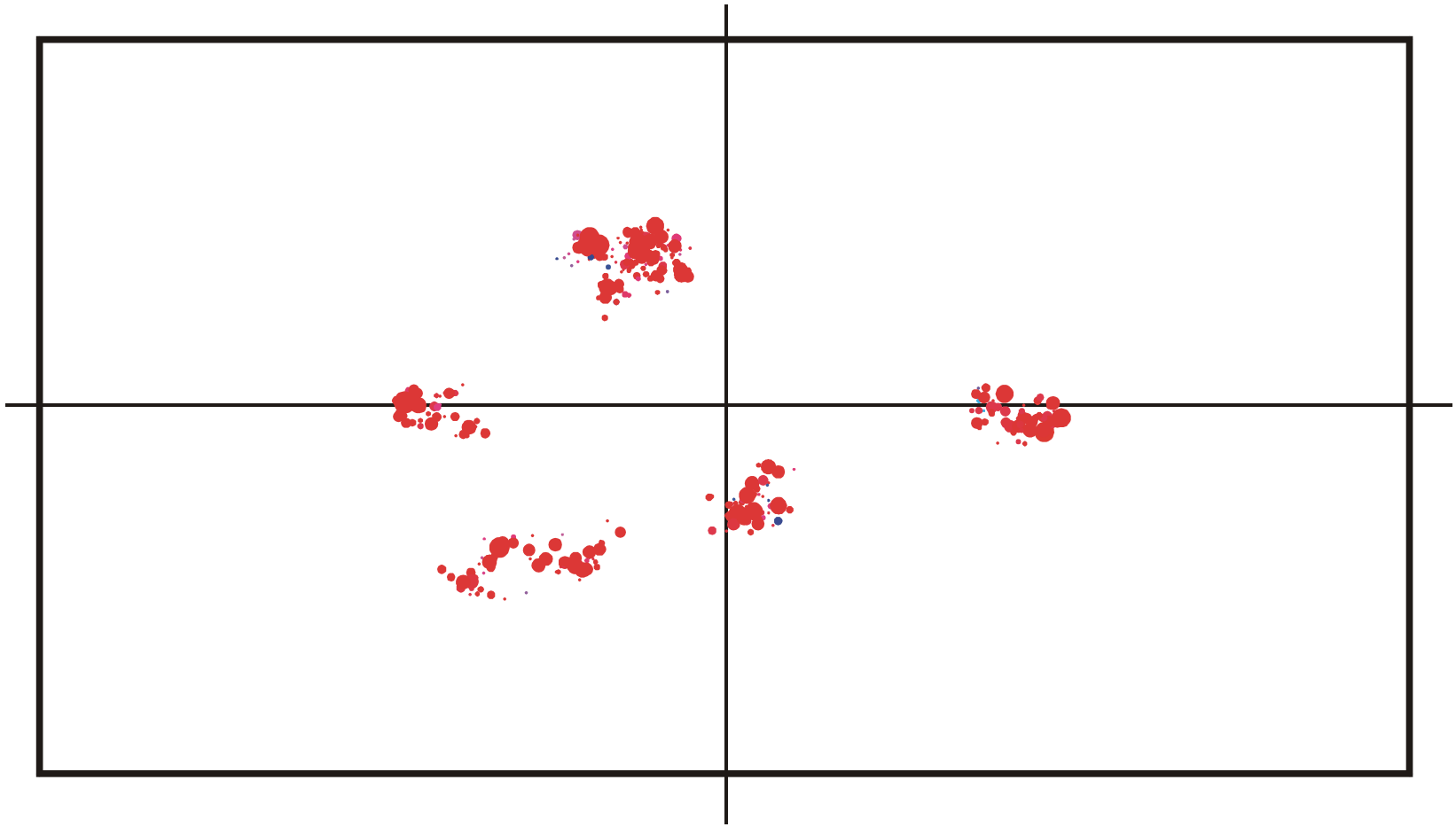
Spreading and evolution of a population on a neutral network : t = 500



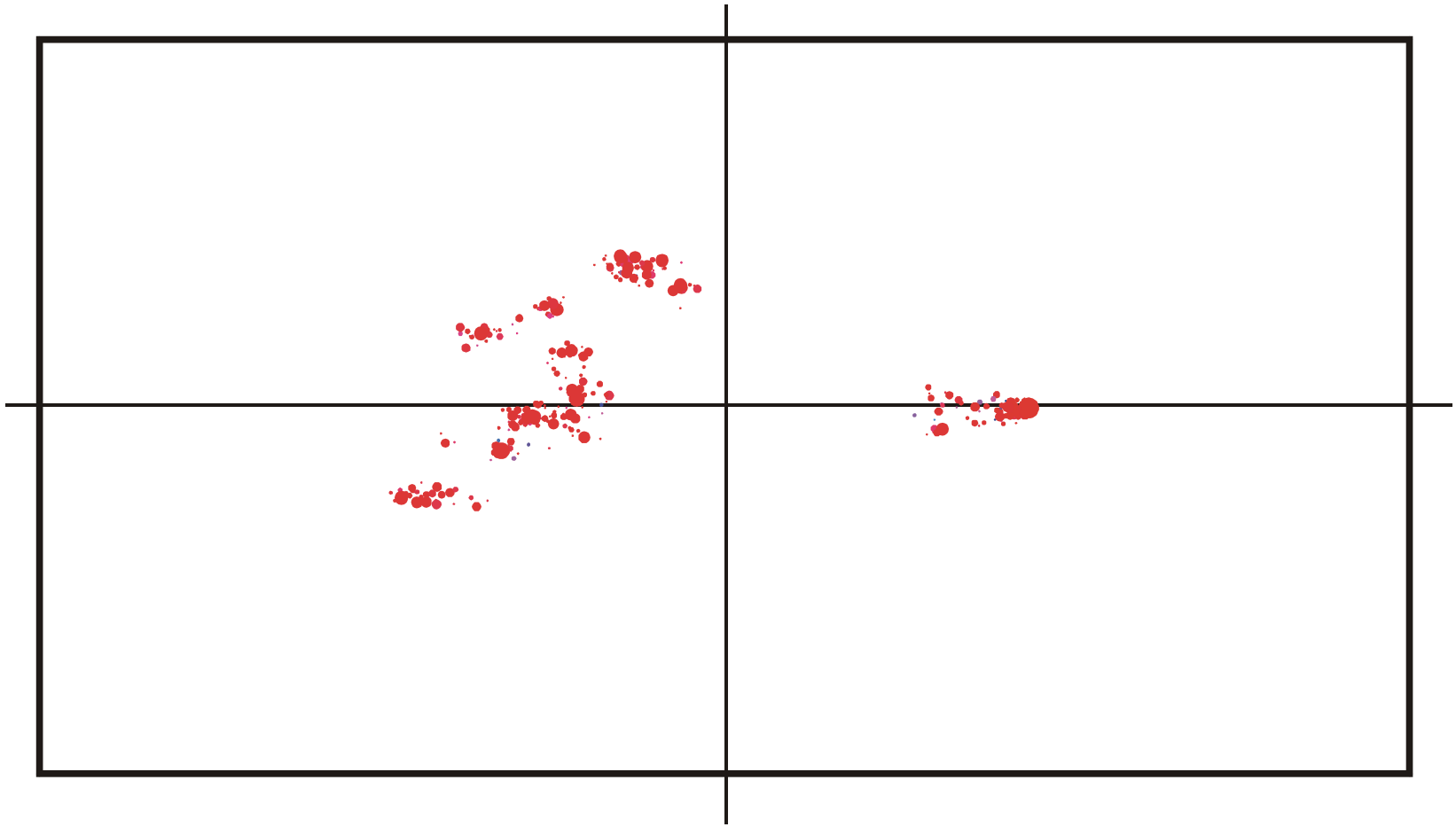
Spreading and evolution of a population on a neutral network : t = 650



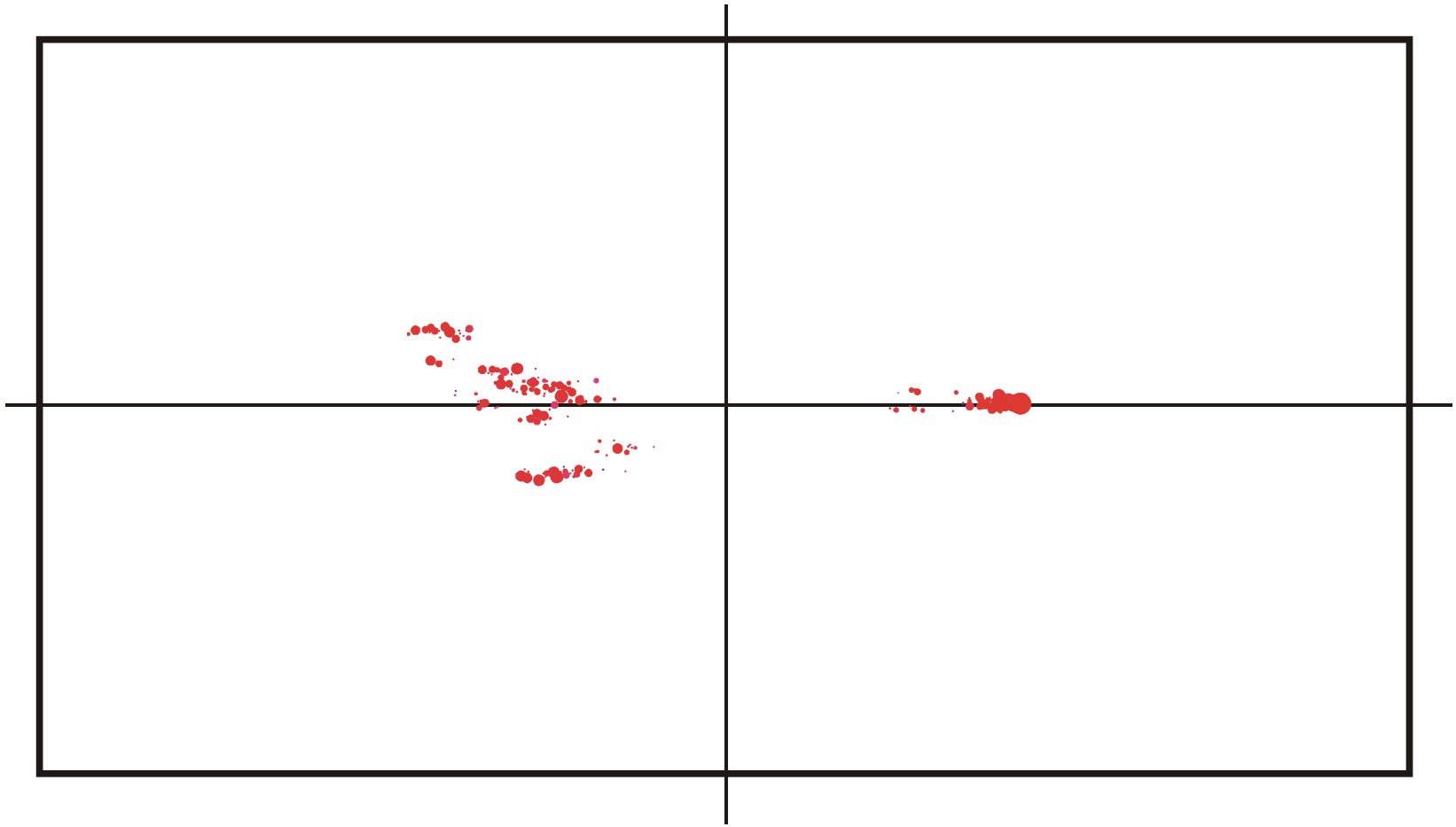
Spreading and evolution of a population on a neutral network : t = 820



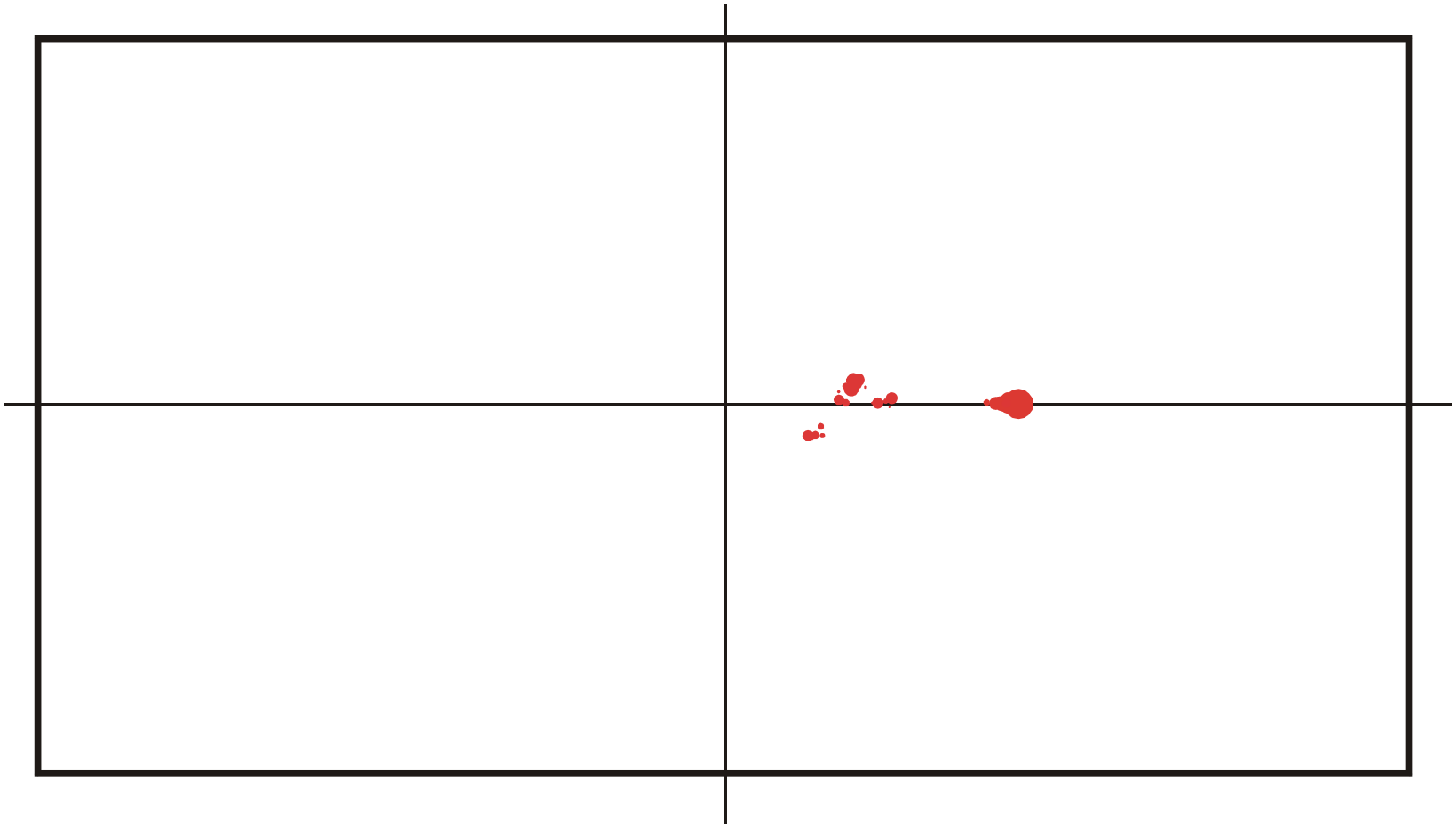
Spreading and evolution of a population on a neutral network : t = 825



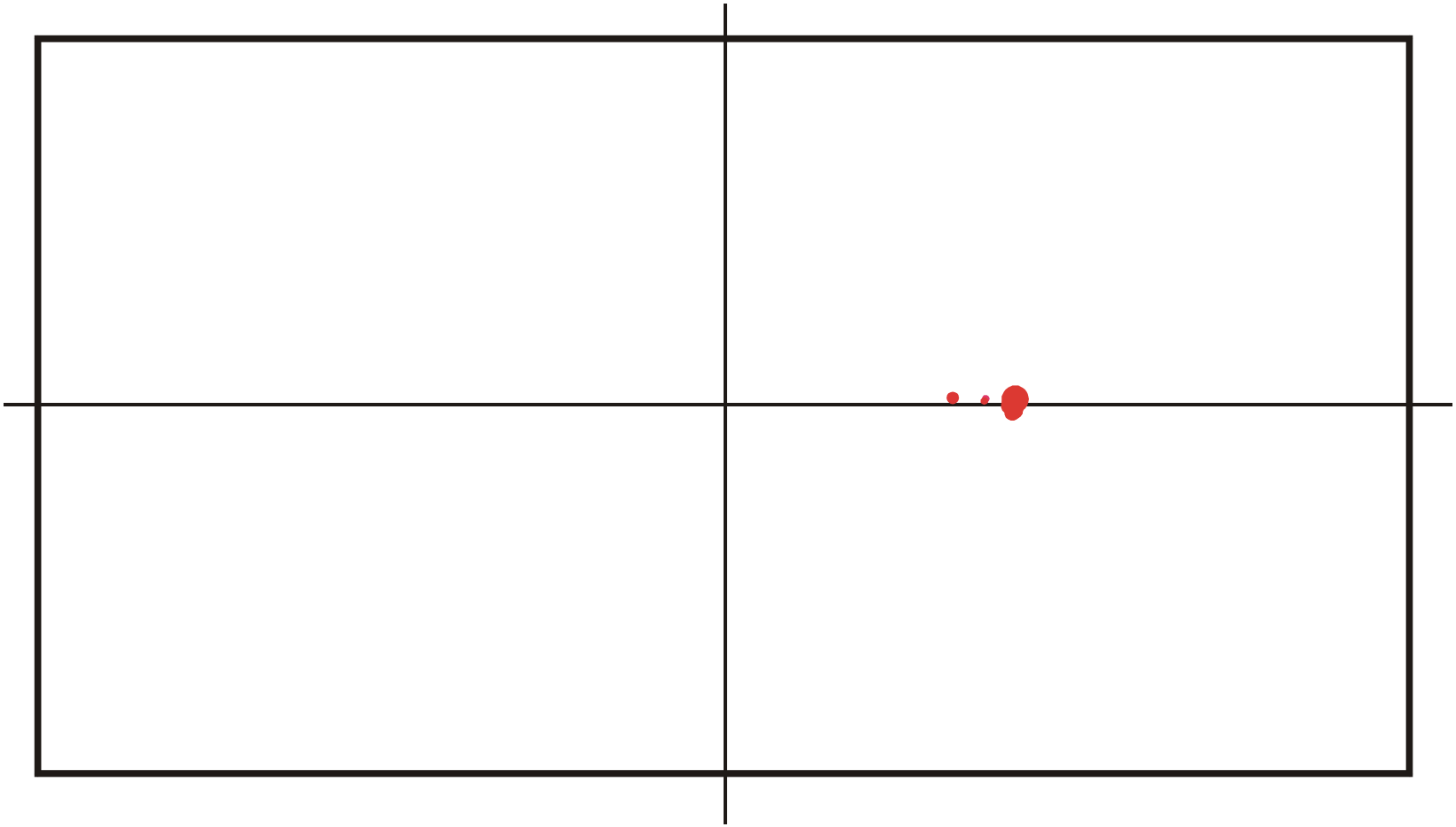
Spreading and evolution of a population on a neutral network : t = 830



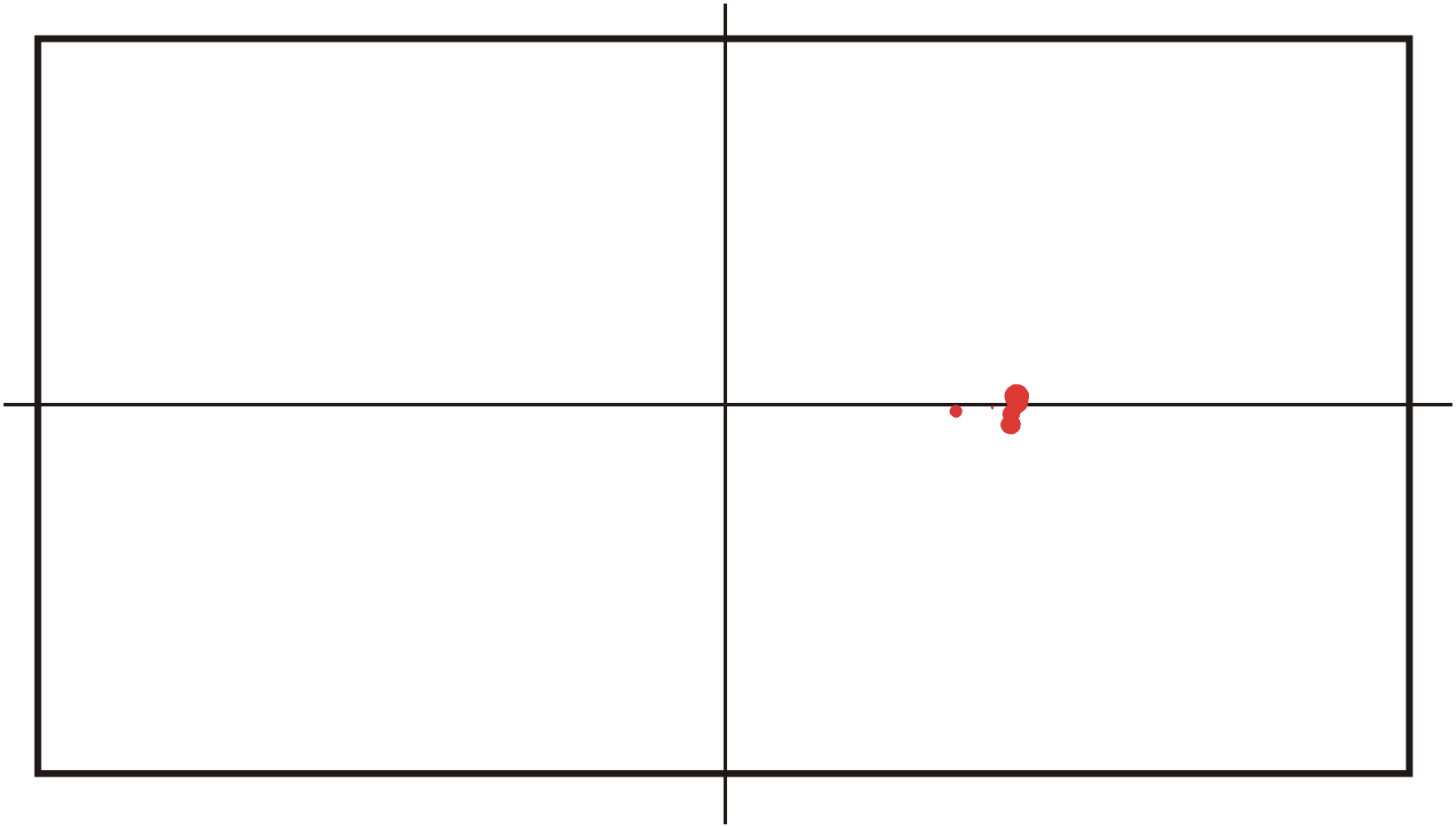
Spreading and evolution of a population on a neutral network : t = 835



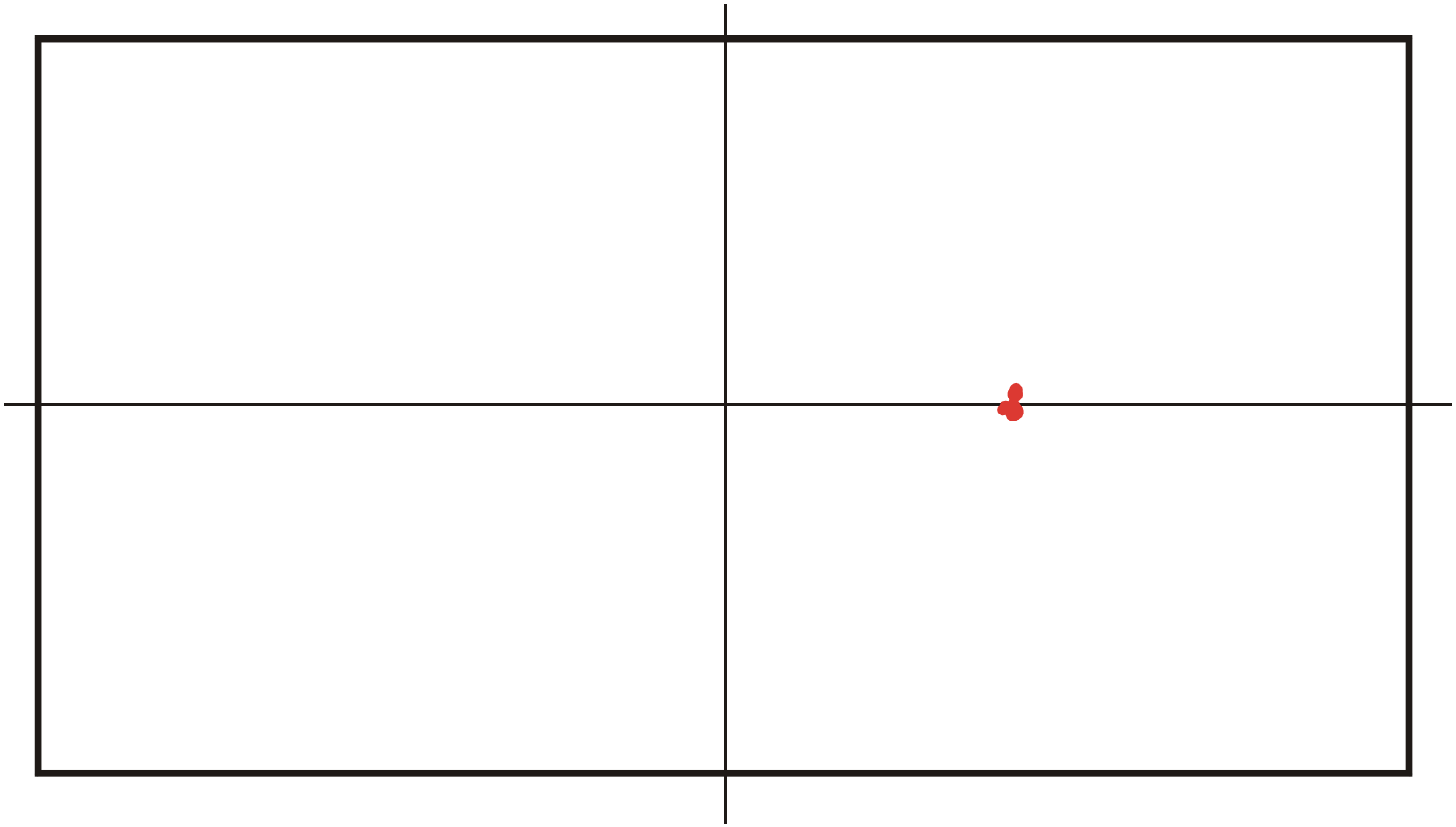
Spreading and evolution of a population on a neutral network : t = 840



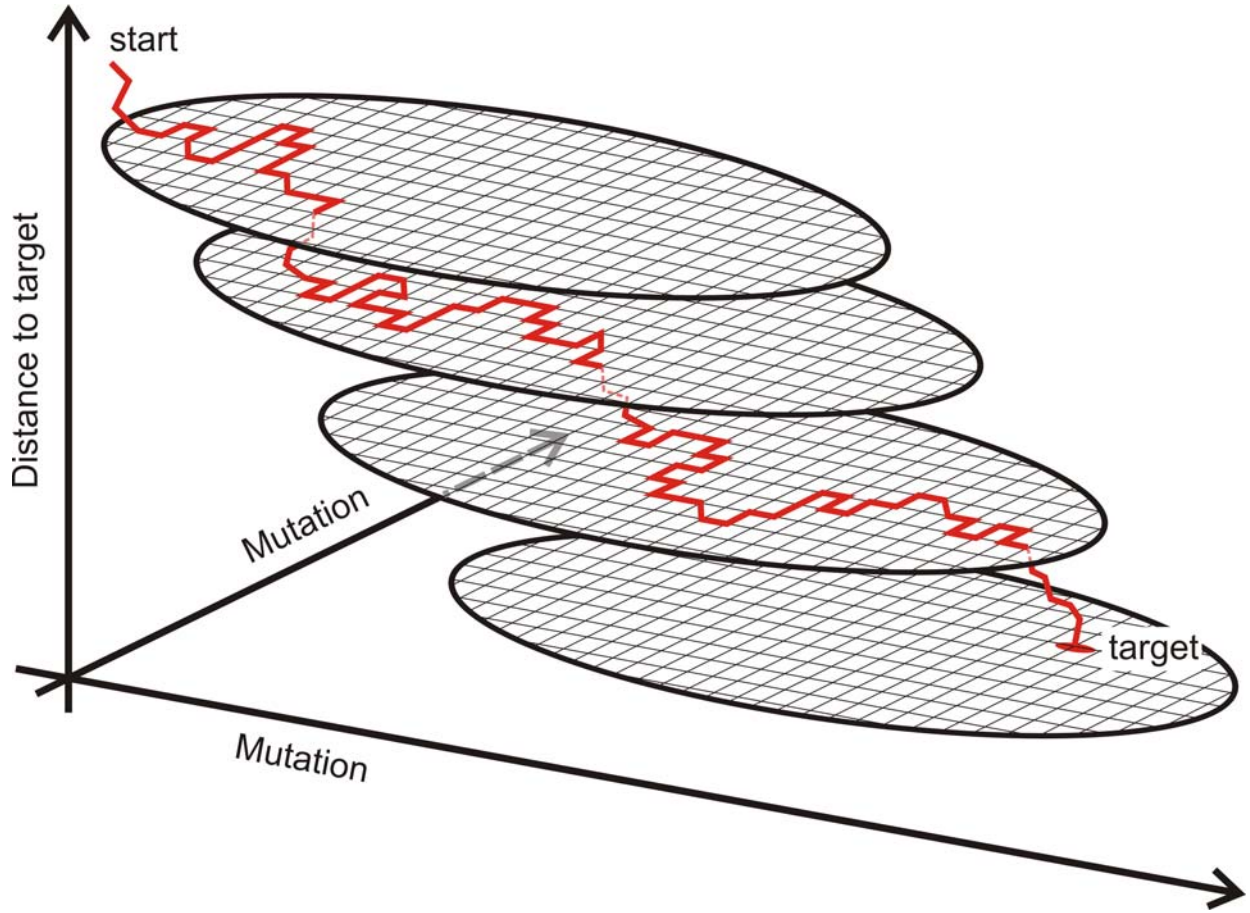
Spreading and evolution of a population on a neutral network : t = 845



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

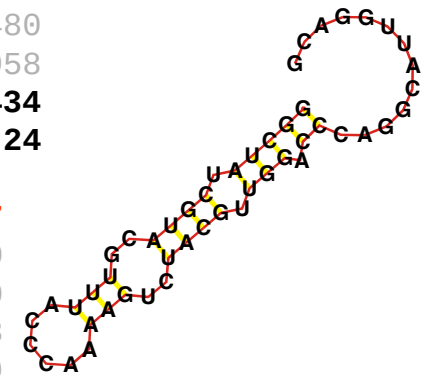
S_T: ((((((...((((........)))).((((.......)))))).....((((.......)))))).))))))....

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

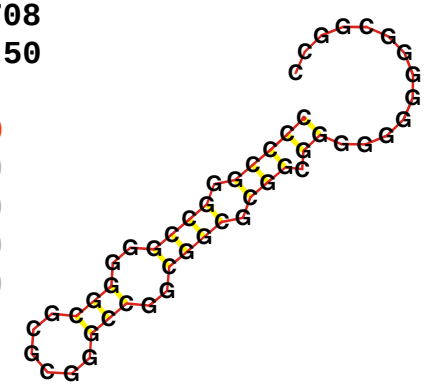
	Number	Mean Value	Variance	Std.Dev.
Total Hamming Distance:	150000	11.647973	23.140715	4.810480
Nonzero Hamming Distance:	99875	16.949991	30.757651	5.545958
Degree of Neutrality:	50125	**0.334167**	0.006961	**0.083434**
Number of Structures:	**1000**	**52.31**	85.30	**9.24**

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



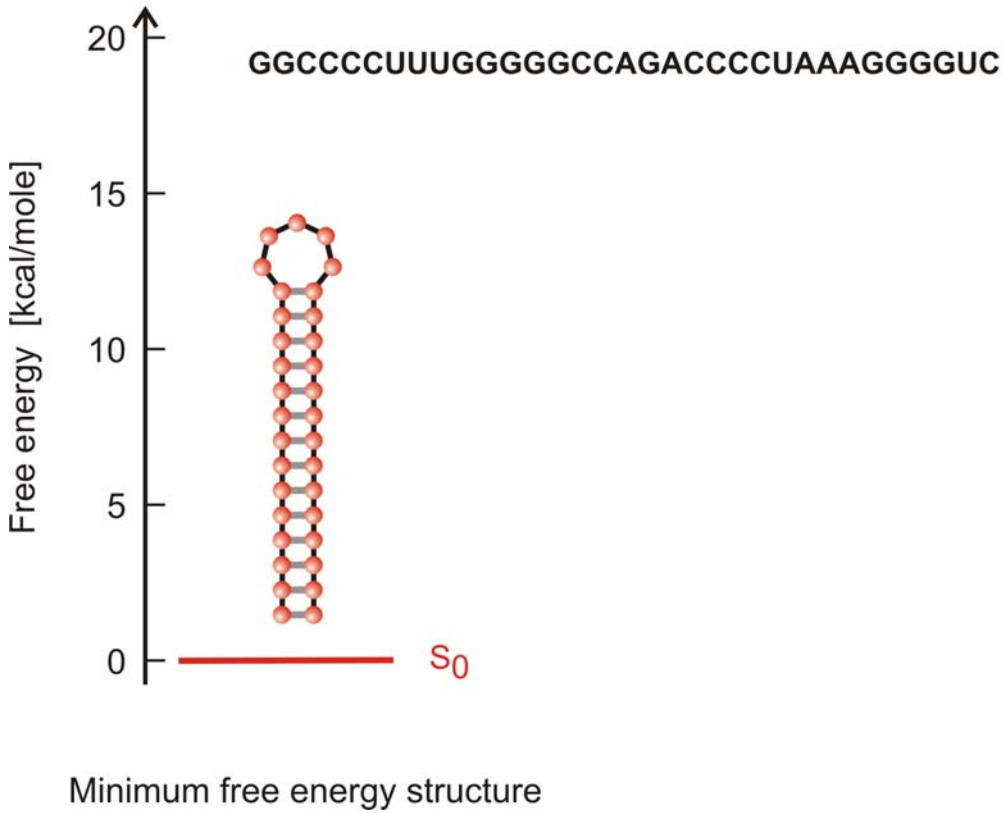
	Number	Mean Value	Variance	Std.Dev.
Total Hamming Distance:	50000	13.673580	10.795762	3.285691
Nonzero Hamming Distance:	45738	14.872054	10.821236	3.289565
Degree of Neutrality:	4262	**0.085240**	0.001824	**0.042708**
Number of Structures:	**1000**	**36.24**	6.27	**2.50**

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

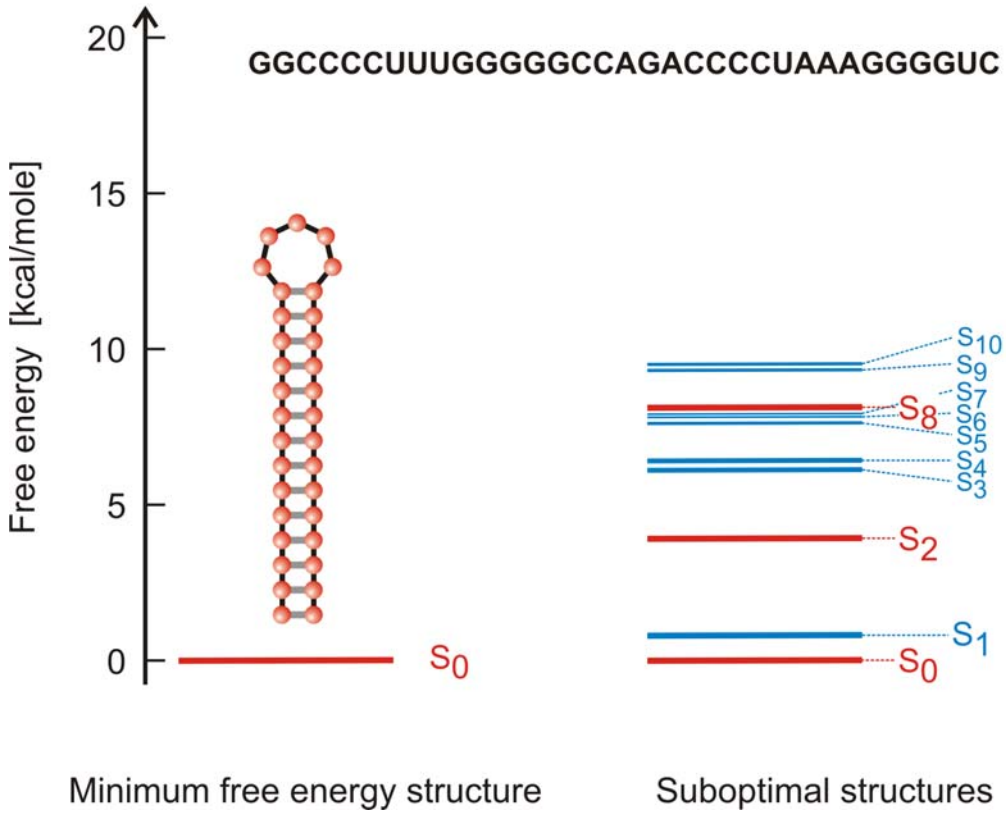


Shadow – Surrounding of an RNA structure in shape space – **AUGC** and **GC** alphabet

1. The origin of neutrality
2. RNA structures as a useful model
3. RNA replication and quasispecies
4. Selection on realistic landscapes
5. Consequences of neutrality
6. Evolutionary optimization of structure
7. **The richness of conformational space**



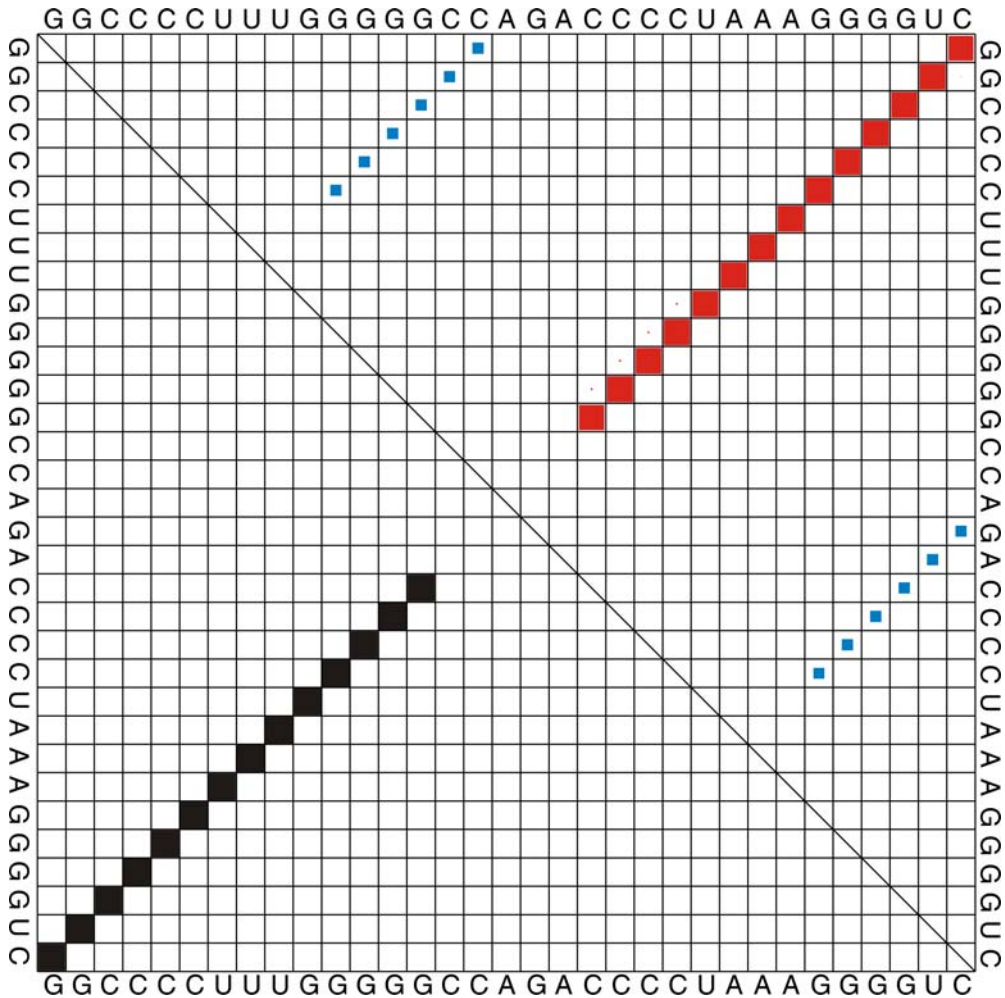
Extension of the notion of structure



Extension of the notion of structure

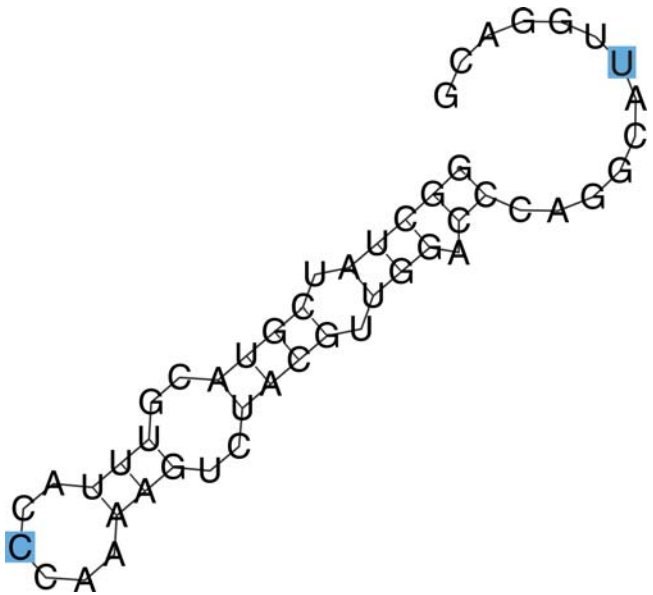
GGCCCCUUUGGGGGCCAGACCCCUAAAGGGGUC

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

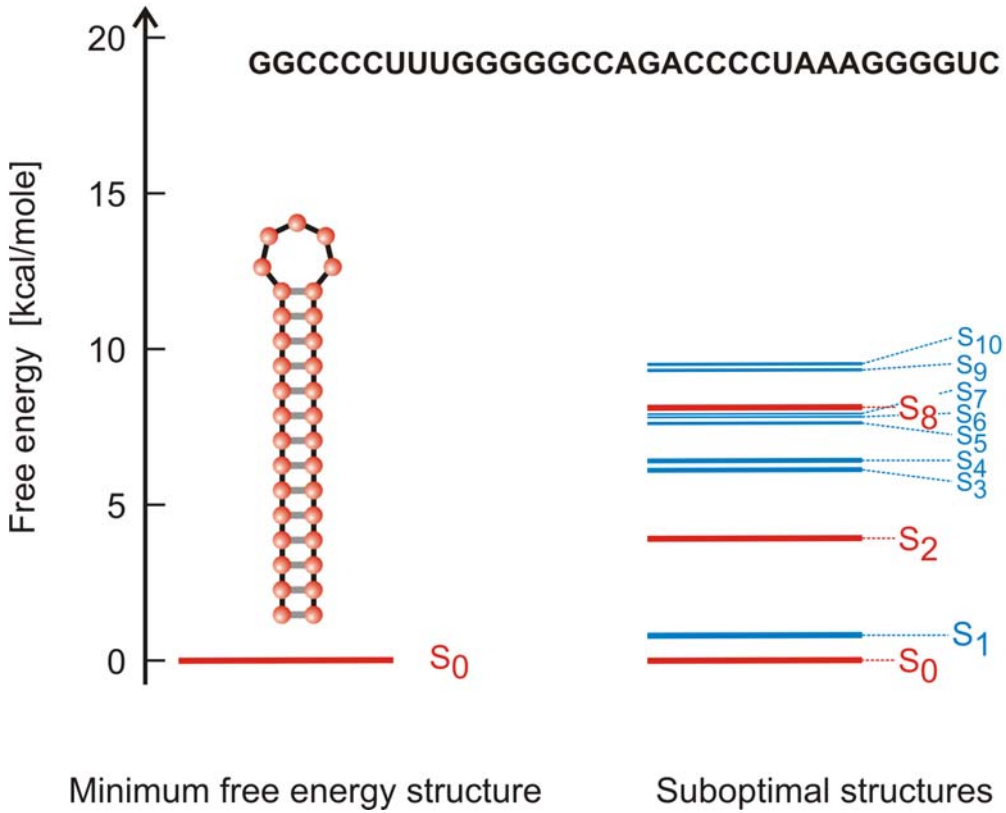
```
(((((.((((..(((......)))..)))).))).)).............  -7.30

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

GGCUAUCGUACGUUUAC**A**CAAAAGUCUACGUUGGACCCAGGCAUUGGACG

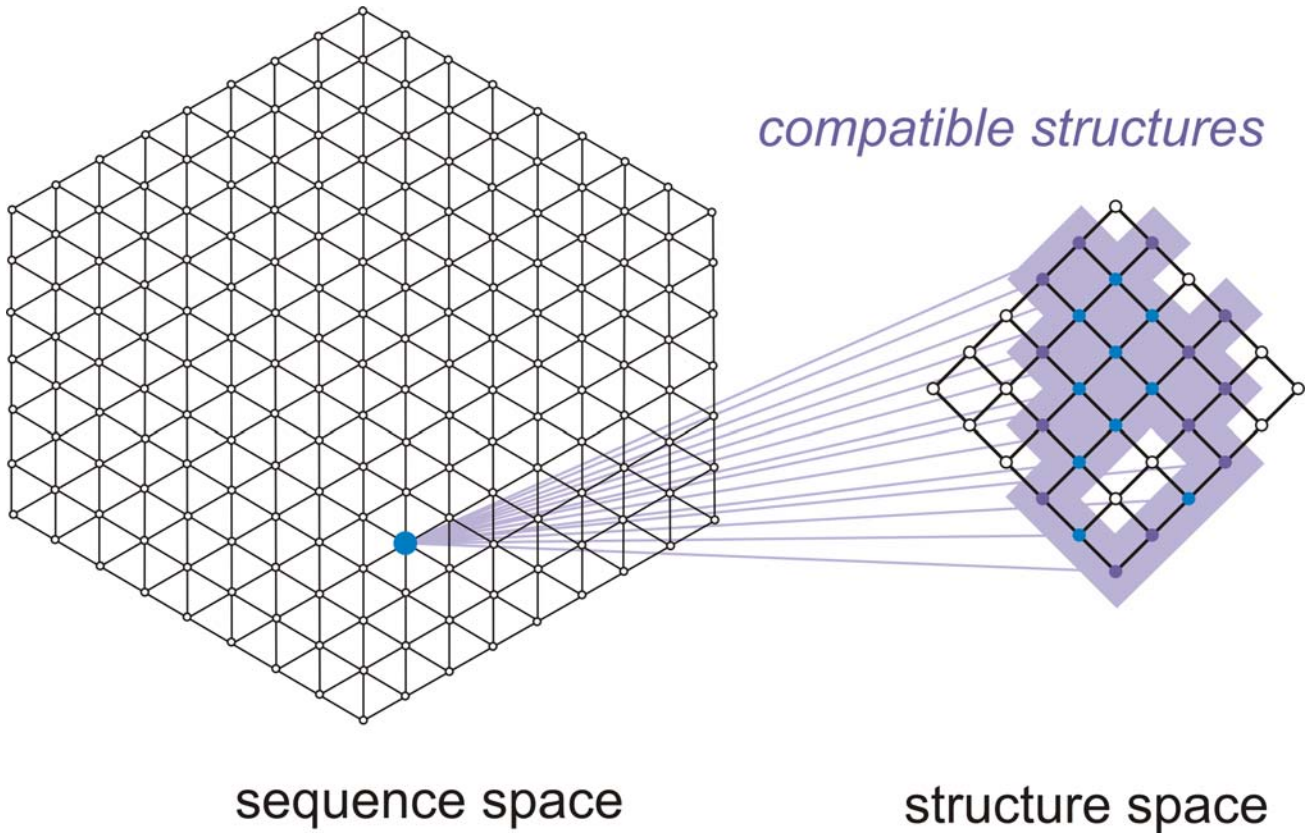
```
(((((.((((..(((......)))..)))).))).)).............  -7.30

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

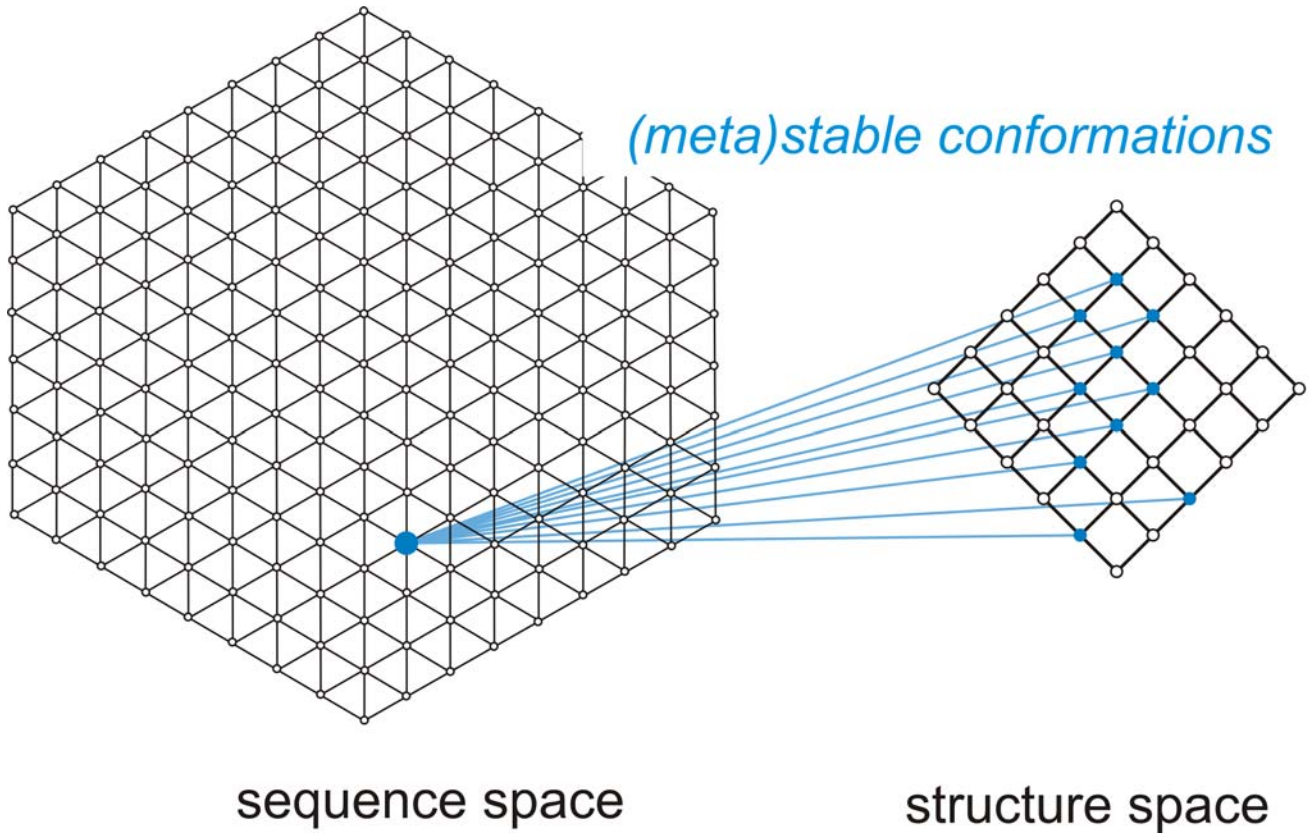
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCA**A**UGGACG

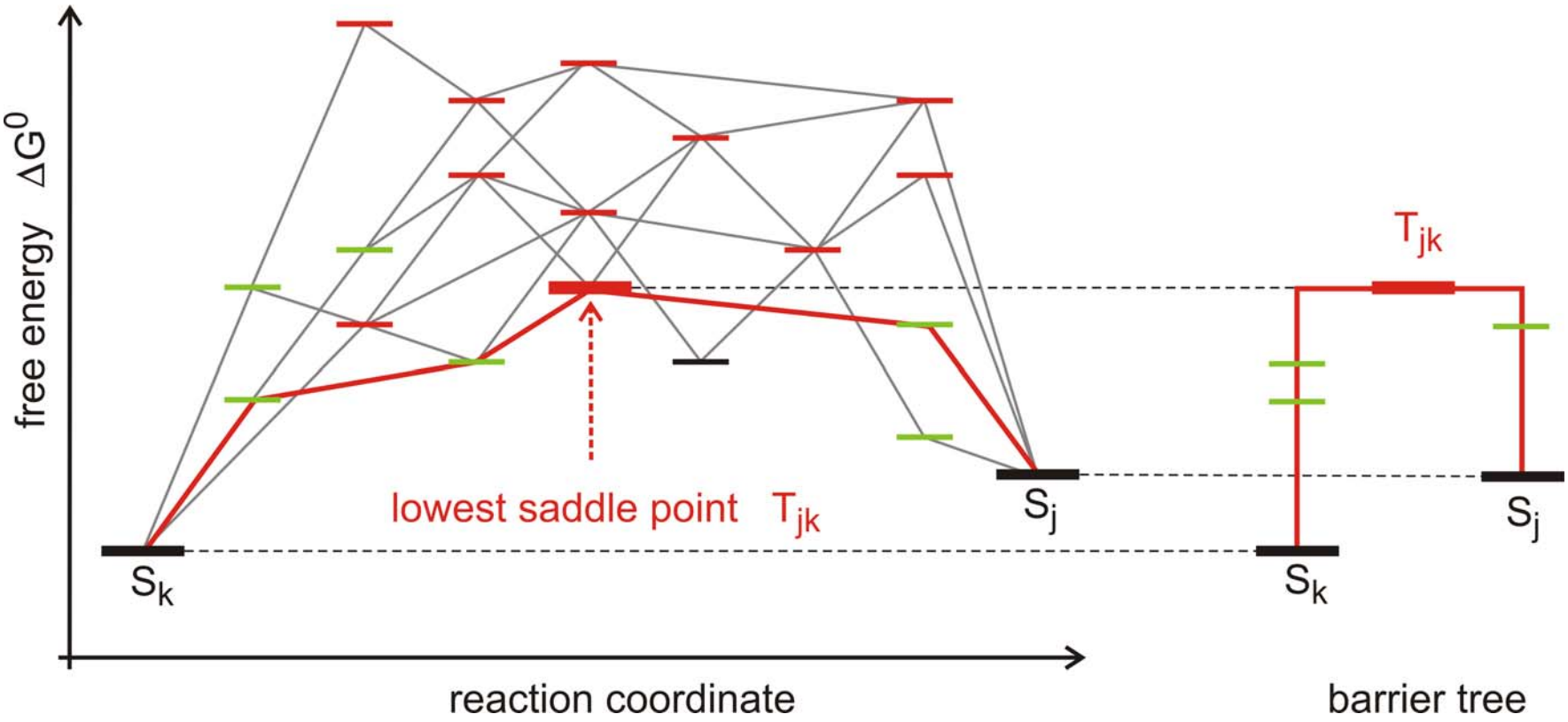
```
(((((.((((..(((......)))..)))).))).)).............  -7.30

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

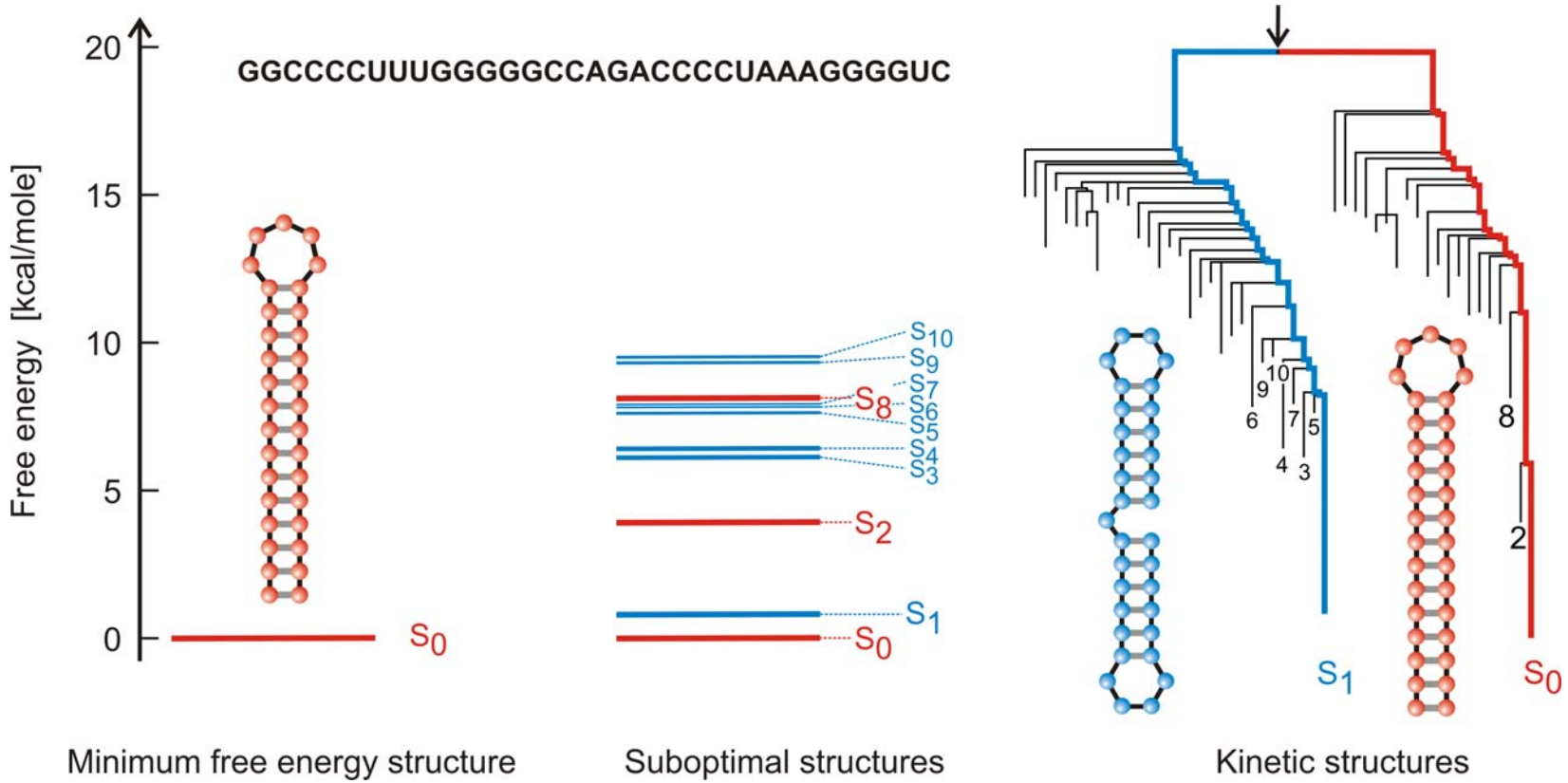


Extension of the notion of structure


compatible structures
sequence space
structure space

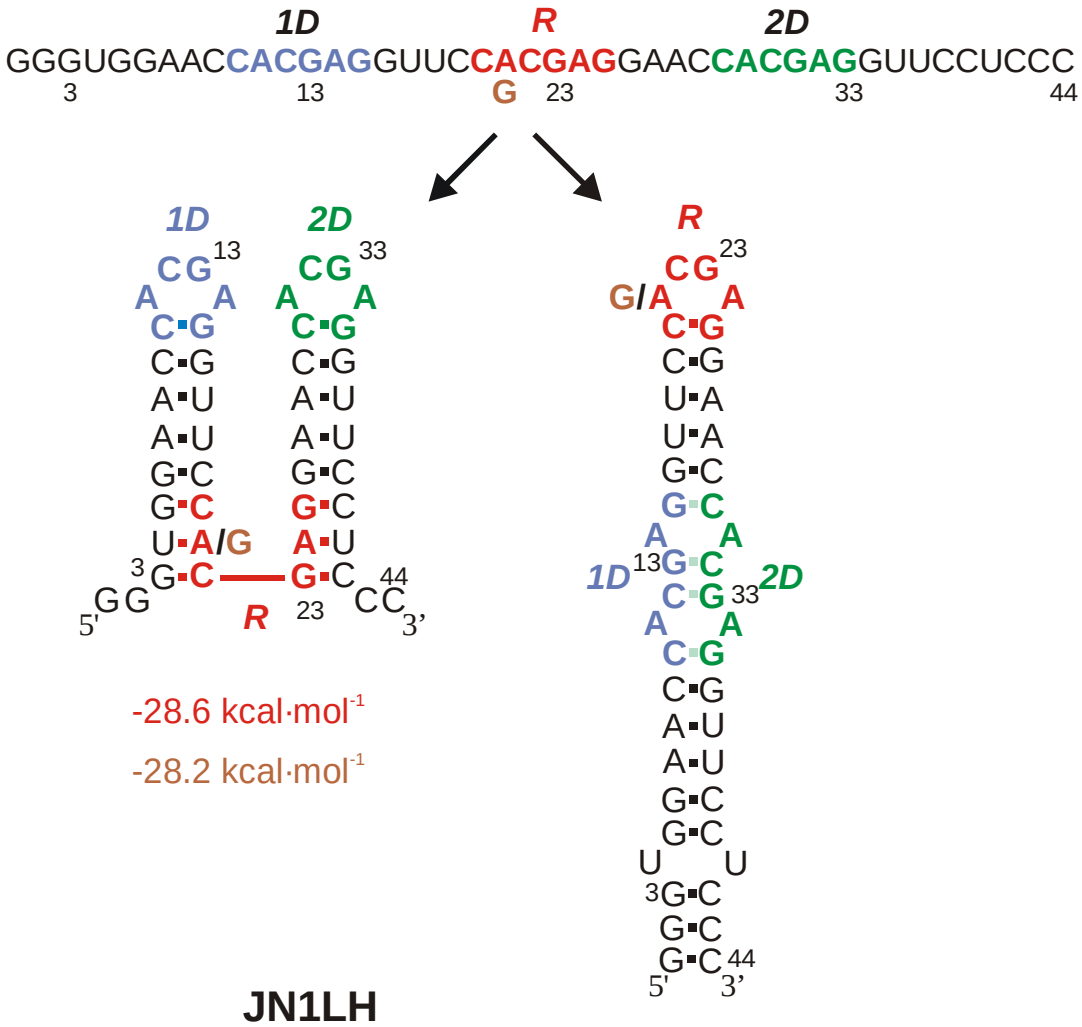

(meta)stable conformations
sequence space
structure space


free energy ΔG^0
lowest saddle point T_{jk}
S_k
S_j
reaction coordinate
T_{jk}
S_k
S_j
barrier tree



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*).

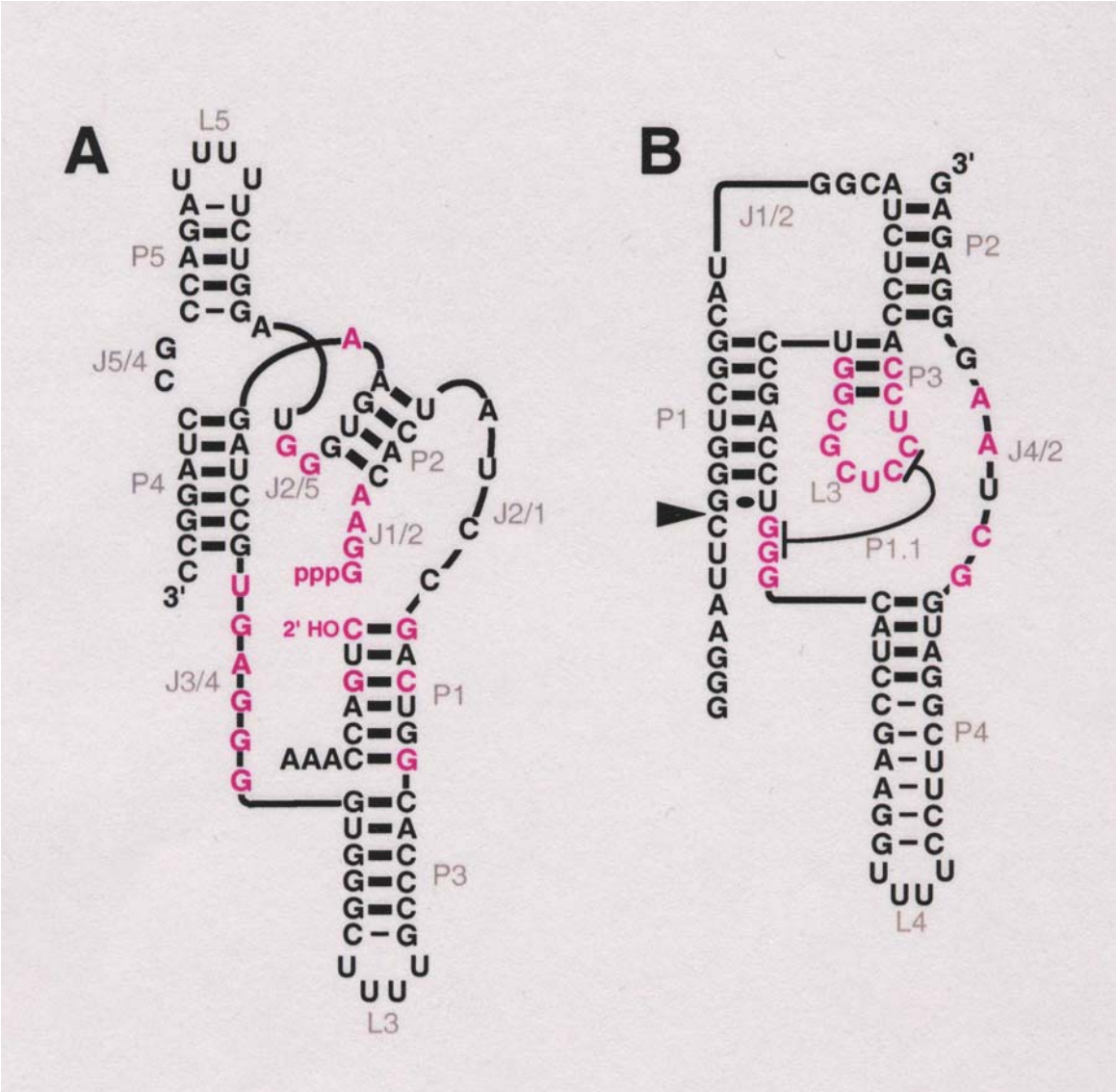
Theoretical analyses using algorithms for predicting RNA secondary structure have suggested that different neutral networks are interwoven and can approach each other very closely (*3*, *5–8*). Of particular interest is whether ribozyme neutral networks approach each other so closely that they intersect.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

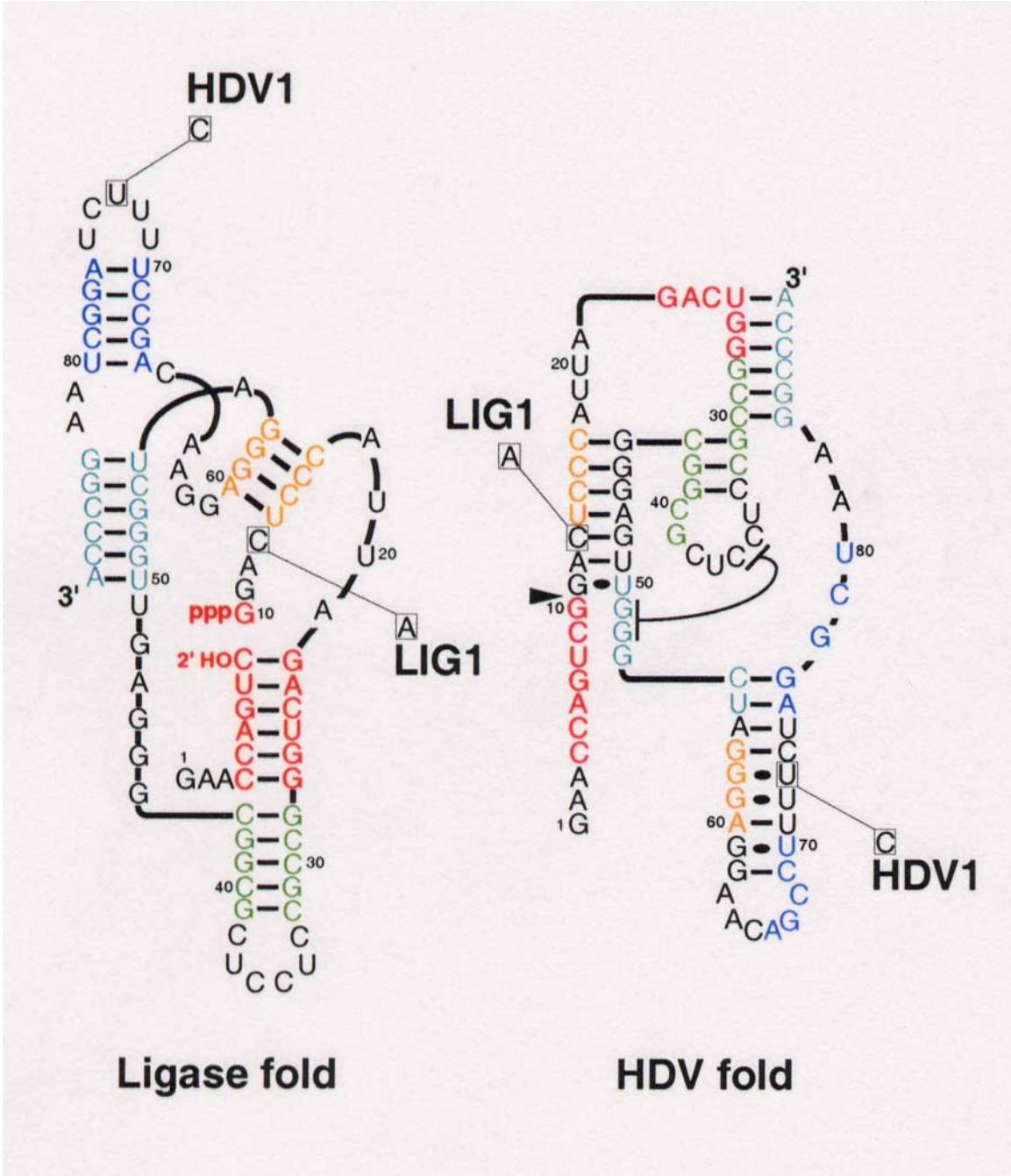
*To whom correspondence should be addressed. E-mail: dbartel@wi.mit.edu

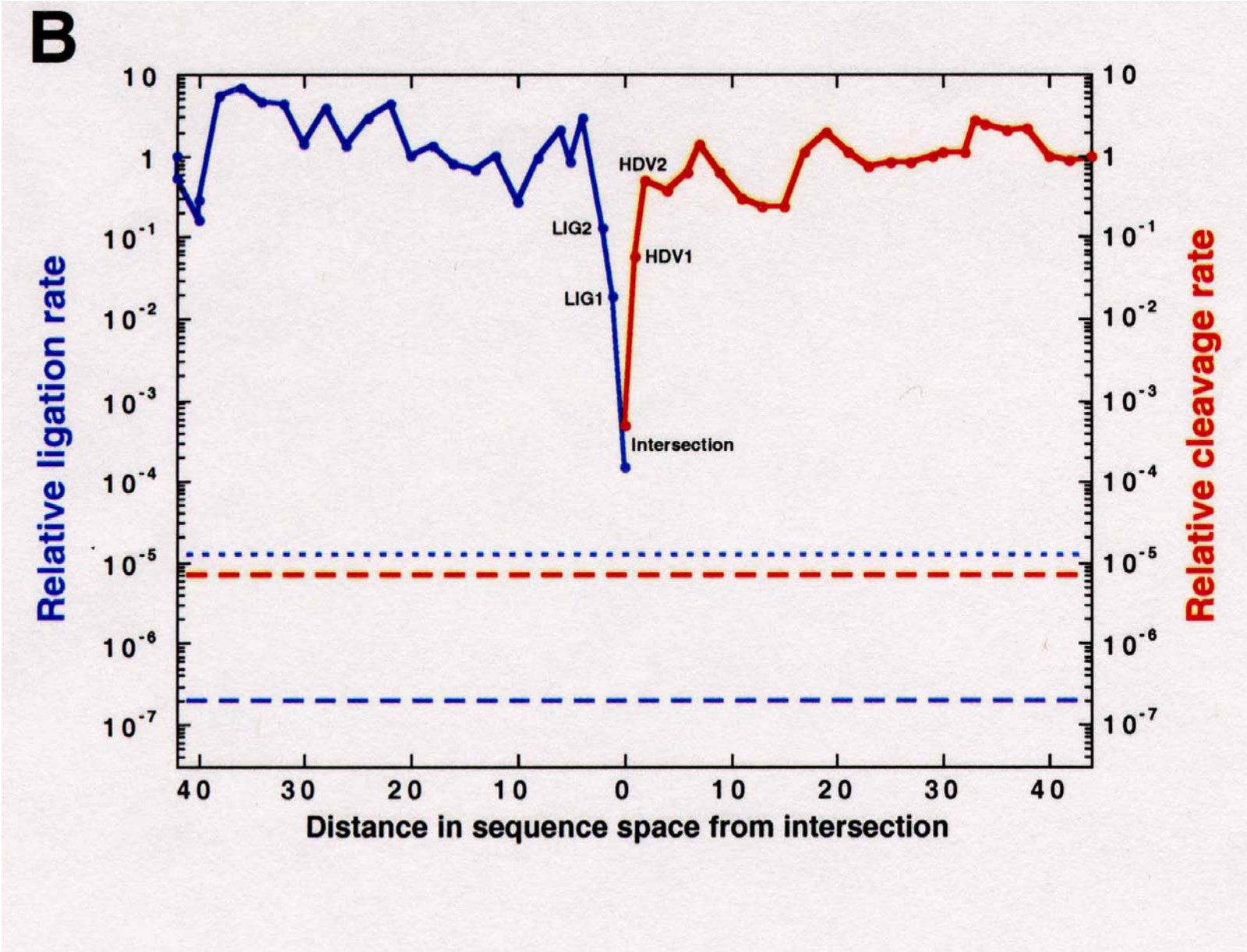


Two ribozymes of chain lengths n = 88 nucleotides: An artificial ligase (**A**) and a natural cleavage ribozyme of hepatitis-δ-virus (**B**)

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

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