Neutrality in Structural Bioinformatics and Molecular Evolution

Peter Schuster

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

Bioinformatics Research and Development 2008

Technische Universität Wien, 07.07.2008

Web-Page for further information:

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

ON

THE ORIGIN OF SPECIES

BY MEANS OF NATURAL SELECTION,

OR THE

PRESERVATION OF FAVOURED RACES IN THE STRUGGLE FOR LIFE.

By CHARLES DARWIN, M.A.,

FELLOW OF THE ROYAL, GEOLOGICAL, LINNÆAN, ETC., SOCIETIES;

AUTHOR OF 'JOURNAL OF RESEARCHES DURING H. M. S. BEAGLE'S VOYAGE ROUND THE WORLD.'

LONDON:

JOHN MURRAY, ALBEMARLE STREET.

1859.

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

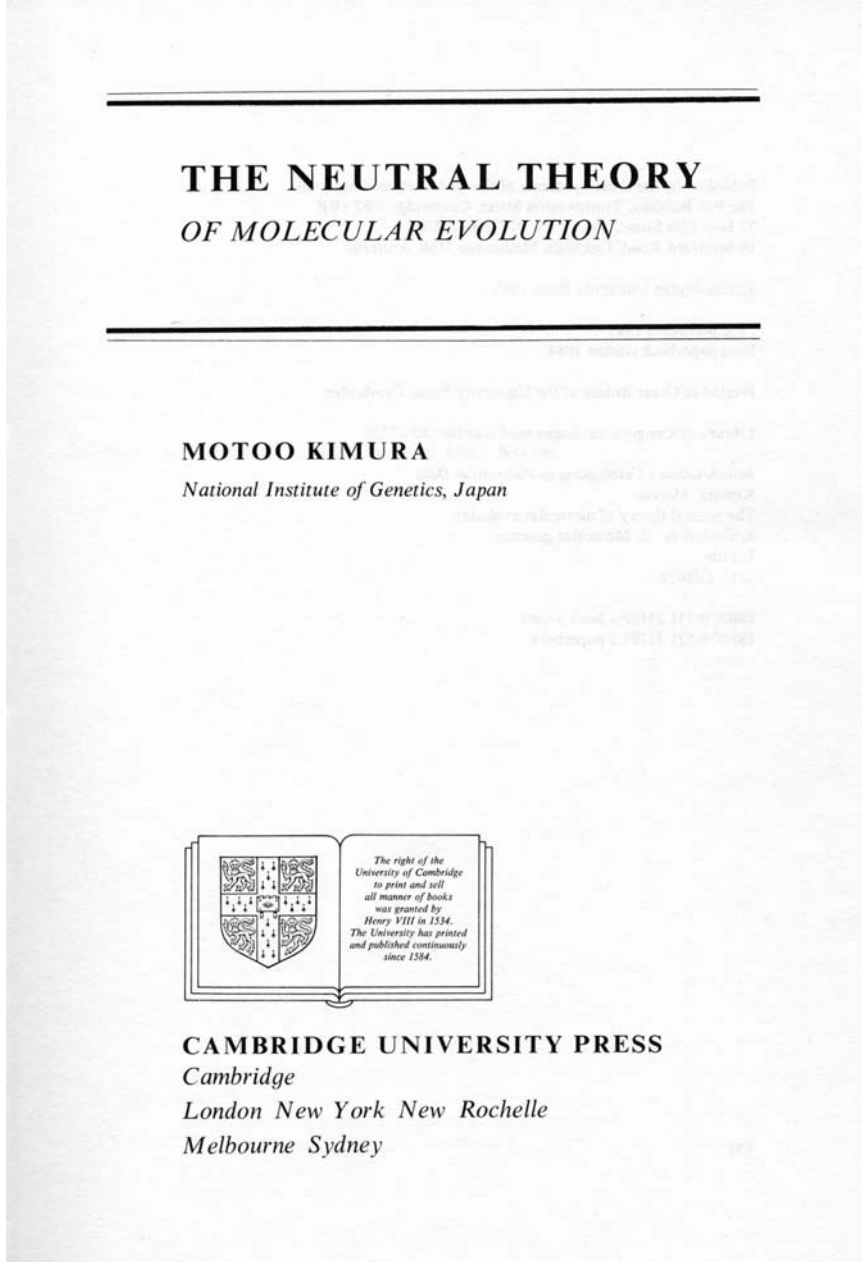
Charles Darwin. *The Origin of Species*. Sixth edition. John Murray. London: 1872

MOTOO KIMURA

Motoo Kimuras population genetics of neutral evolution.

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

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



THE NEUTRAL THEORY

OF MOLECULAR EVOLUTION

MOTOO KIMURA

National Institute of Genetics, Japan

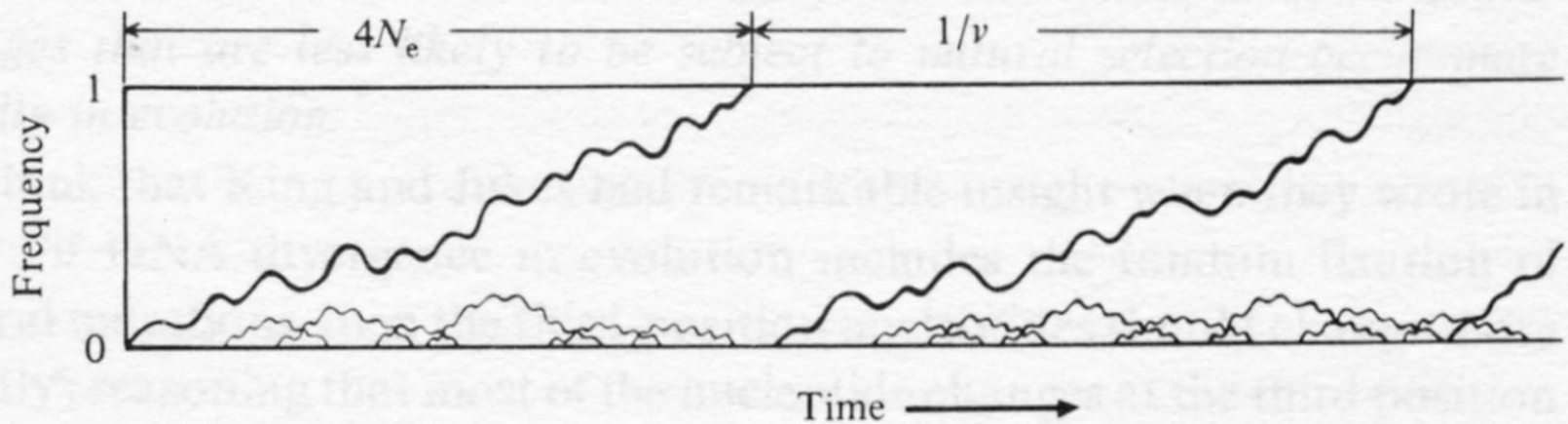
CAMBRIDGE UNIVERSITY PRESS

Cambridge

London New York New Rochelle

Melbourne Sydney

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

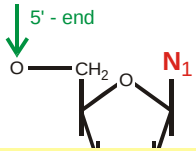


The average time of replacement of a dominant genotype in a population is the reciprocal mutation rate, 1/ν, and therefore independent of population size.

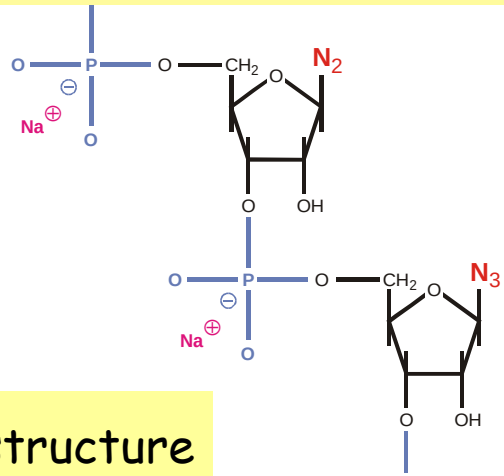
Fixation of mutants in neutral evolution (Motoo Kimura, 1955)

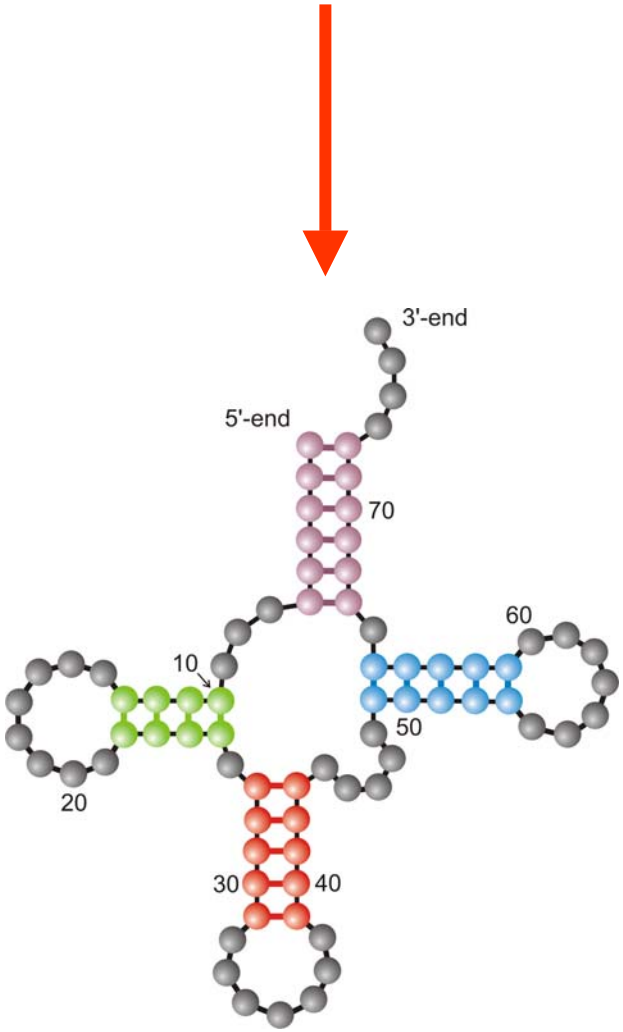
1. Ruggedness of molecular landscapes
2. Replication-mutation dynamics
3. Models of fitness landscapes
4. Ruggedness and error thresholds
5. Stochasticity of replication and mutation
6. Population dynamics on neutral networks

1. **Ruggedness of molecular landscapes**
2. Replication-mutation dynamics
3. Models of fitness landscapes
4. Ruggedness and error thresholds
5. Stochasticity of replication and mutation
6. Population dynamics on neutral networks

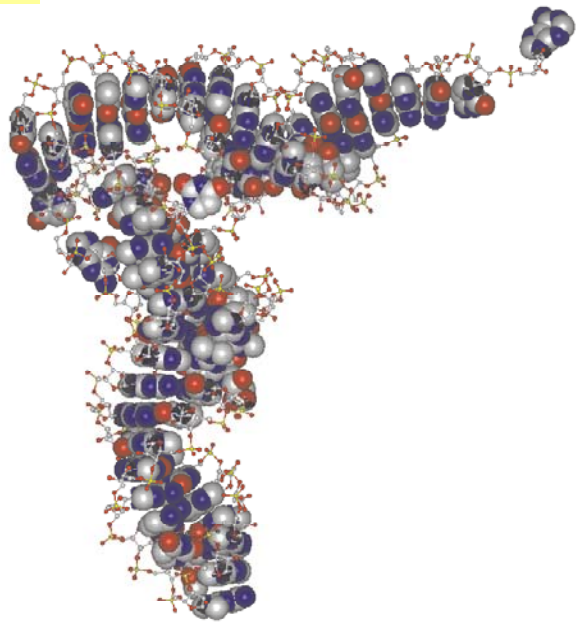


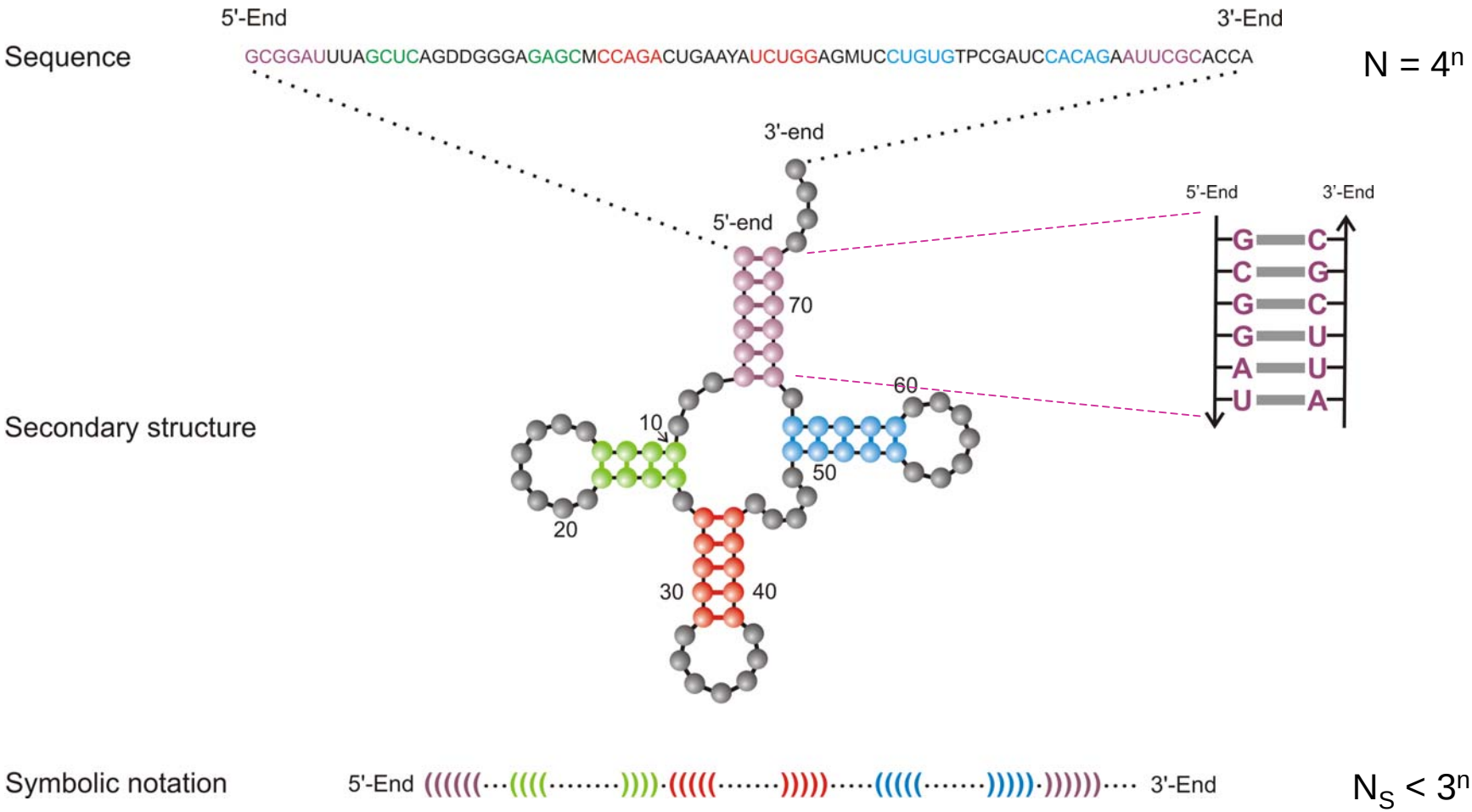
5'-end GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCACCA 3'-end





Definition of RNA structure

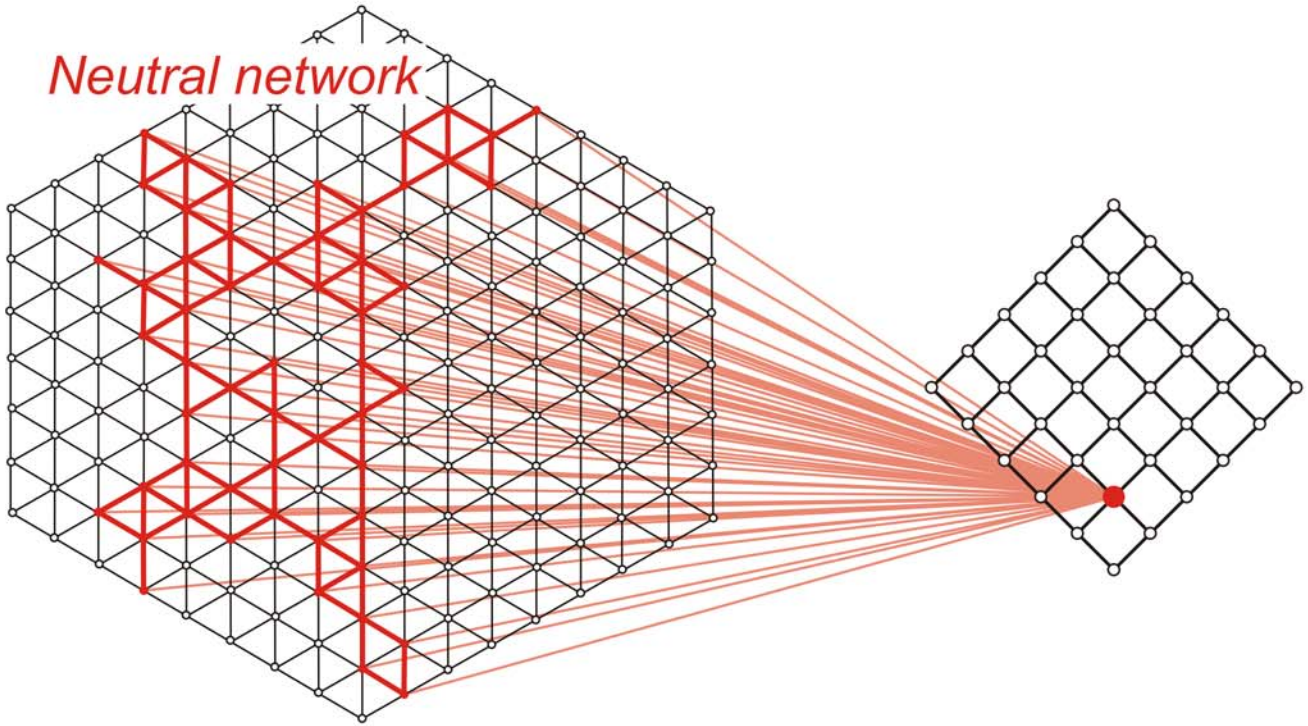




Criterion: Minimum free energy (mfe)

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

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

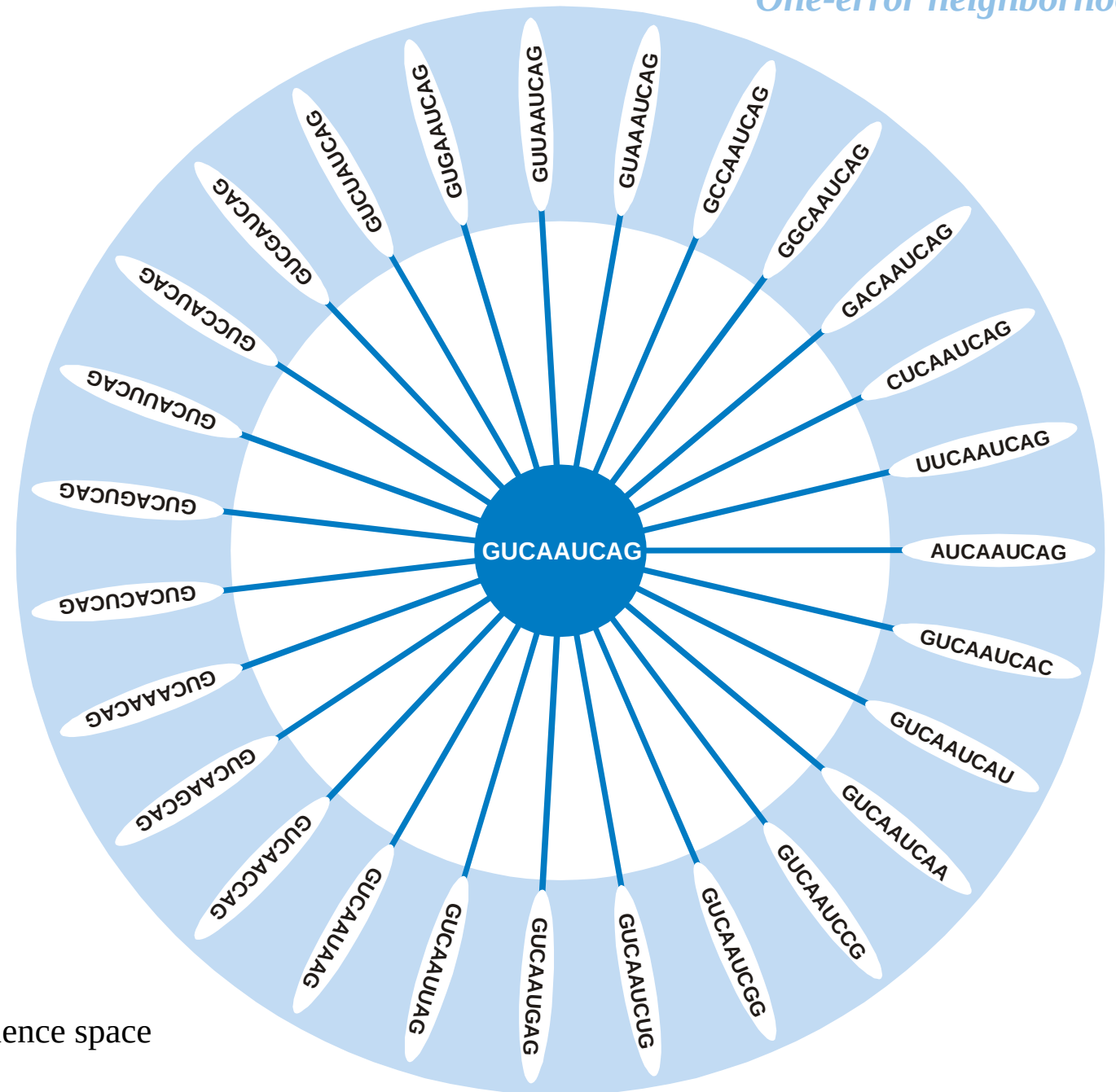


Sequence space

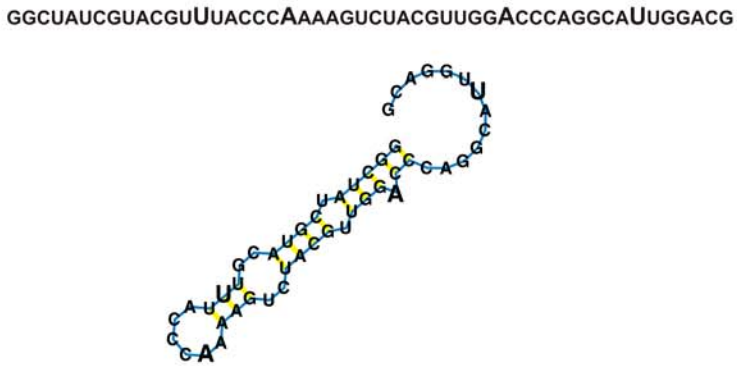
Structure space

many genotypes ⇒ one phenotype

One-error neighborhood



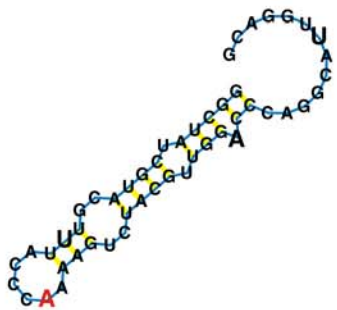
The surrounding of
GUCAAUCAG in sequence space



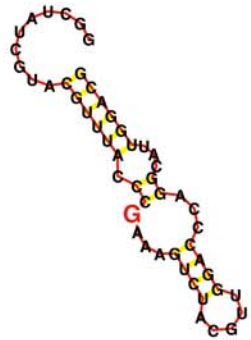
One error neighborhood – Surrounding of an RNA molecule of chain length n=50 in sequence and shape space

GGCUAUCGUACGUUUACCCGAAAGUCUACGUUGGACCCAGGCAUUGGACG

GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG

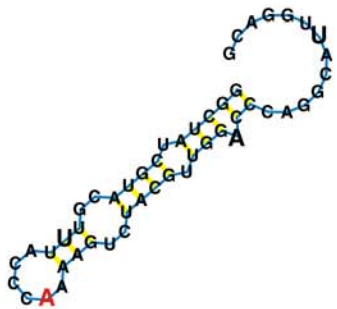


One error neighborhood – Surrounding of an RNA molecule of chain length n=50 in sequence and shape space

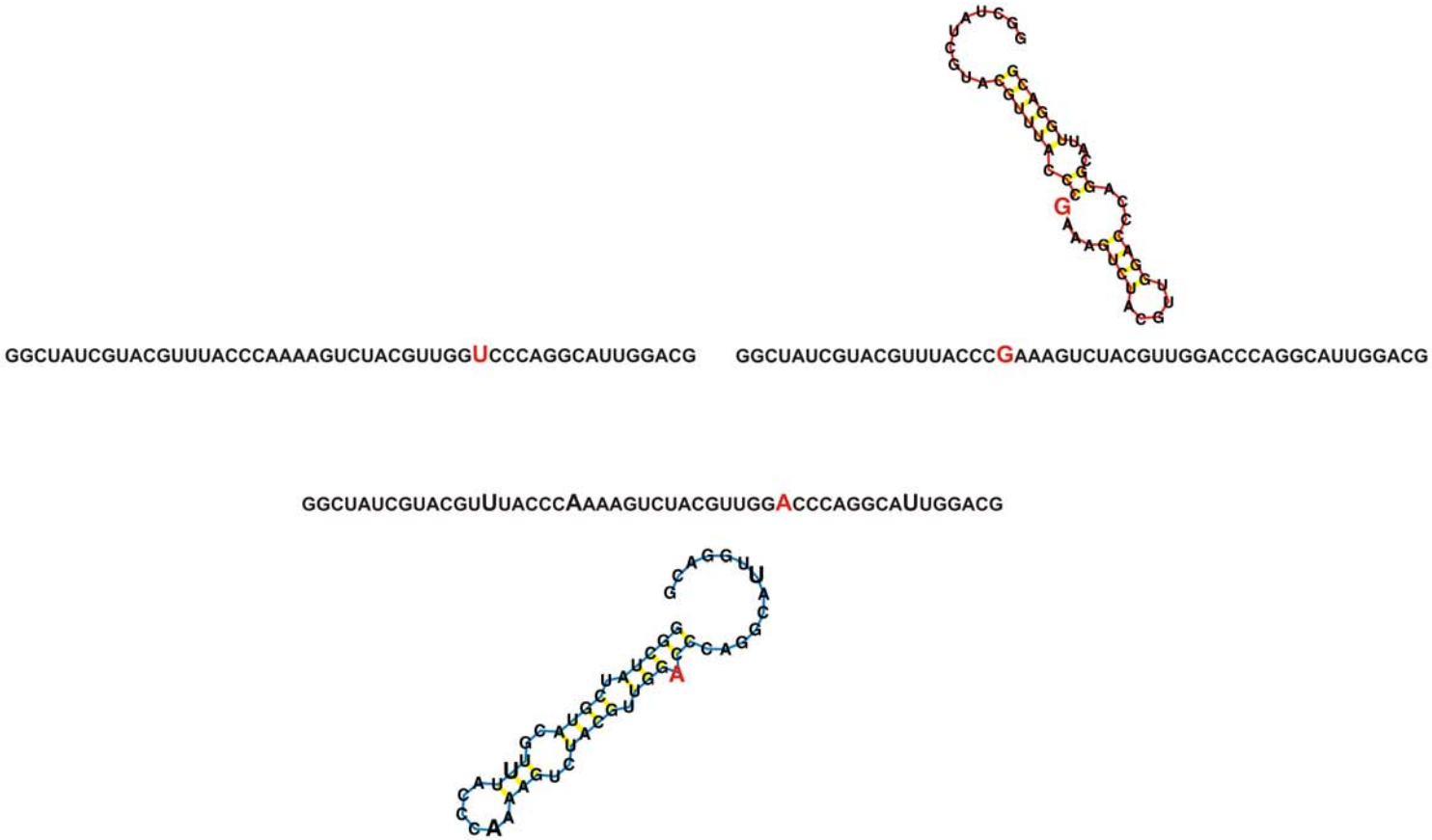


GGCUAUCGUACGUUUACCCGAAAGUCUACGUUGGACCCAGGCAUUGGACG

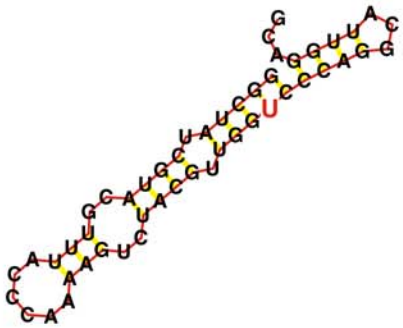
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG



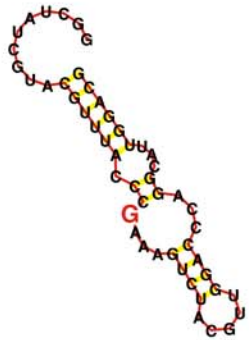
One error neighborhood – Surrounding of an RNA molecule of chain length n=50 in sequence and shape space



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

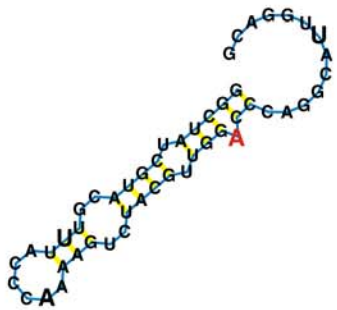


GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGUCCCAGGCAUUGGACG

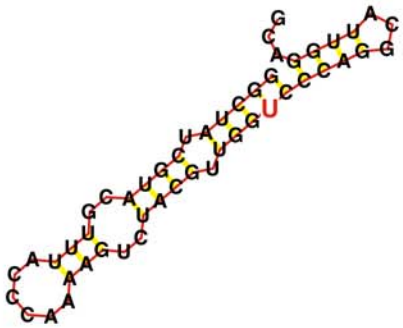


GGCUAUCGUACGUUUACCCGAAAGUCUACGUUGGACCCAGGCAUUGGACG

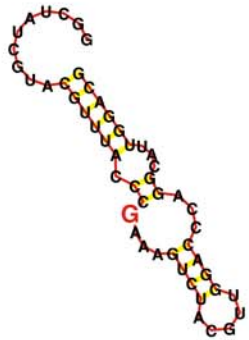
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG



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



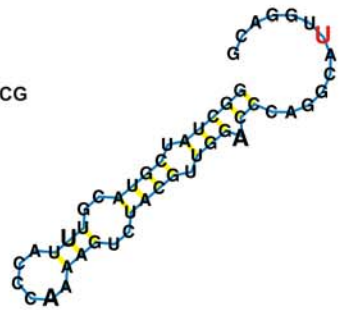
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGUCCCAGGCAUUGGACG



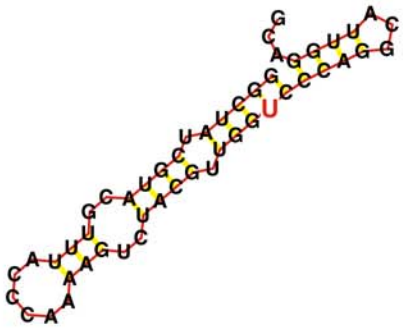
GGCUAUCGUACGUUUACCCGAAAGUCUACGUUGGACCCAGGCAUUGGACG

GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG

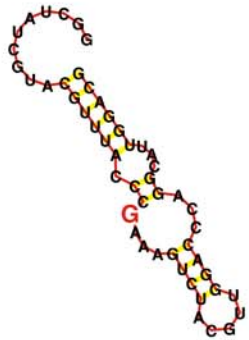
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCACUGGACG



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



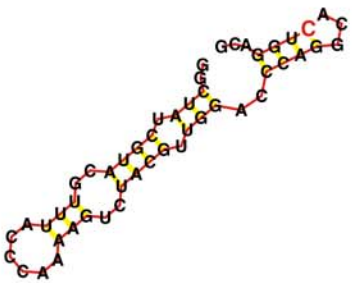
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGUCCCAGGCAUUGGACG

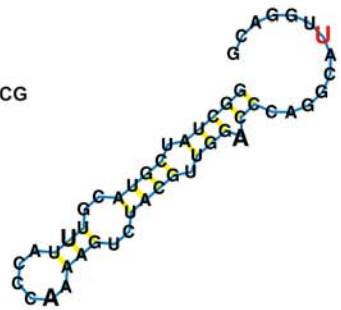


GGCUAUCGUACGUUUACCCGAAAGUCUACGUUGGACCCAGGCAUUGGACG

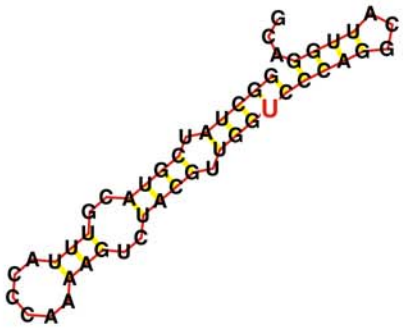
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG

GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCACUGGACG

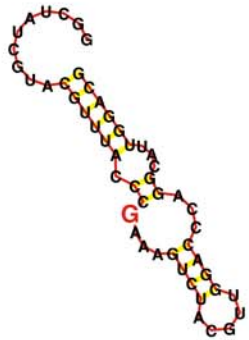




One error neighborhood – Surrounding of an RNA molecule of chain length n=50 in sequence and shape space



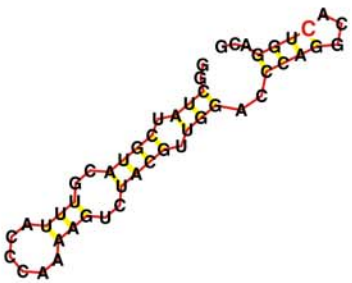
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGUCCCAGGCAUUGGACG

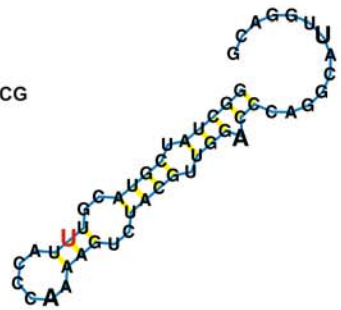


GGCUAUCGUACGUUUACCCGAAAGUCUACGUUGGACCCAGGCAUUGGACG

GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG

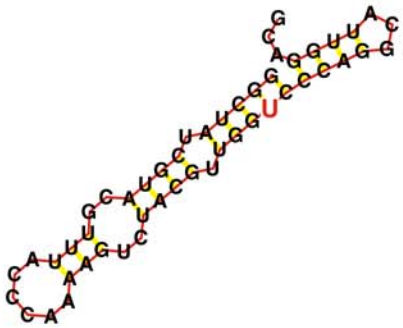
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCACUGGACG



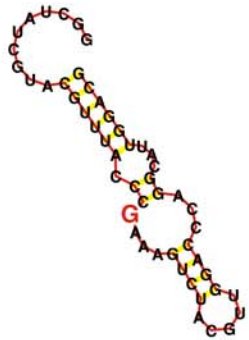


GGCUAUCGUACGUGUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG

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

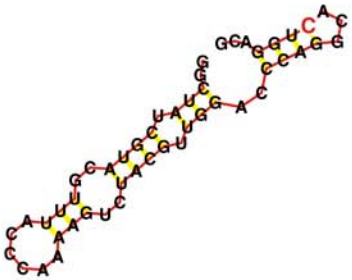


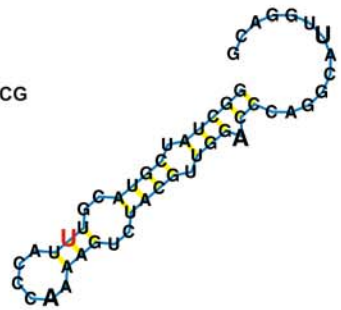
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGUCCCAGGCAUUGGACG



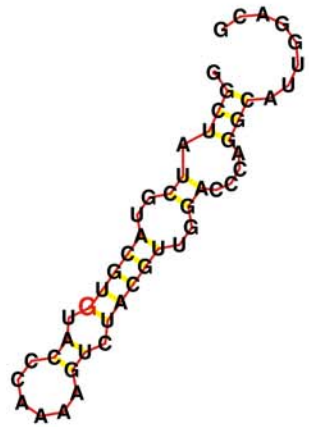
GGCUAUCGUACGUUUACCCGAAAGUCUACGUUGGACCCAGGCAUUGGACG

GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCACUGGACG





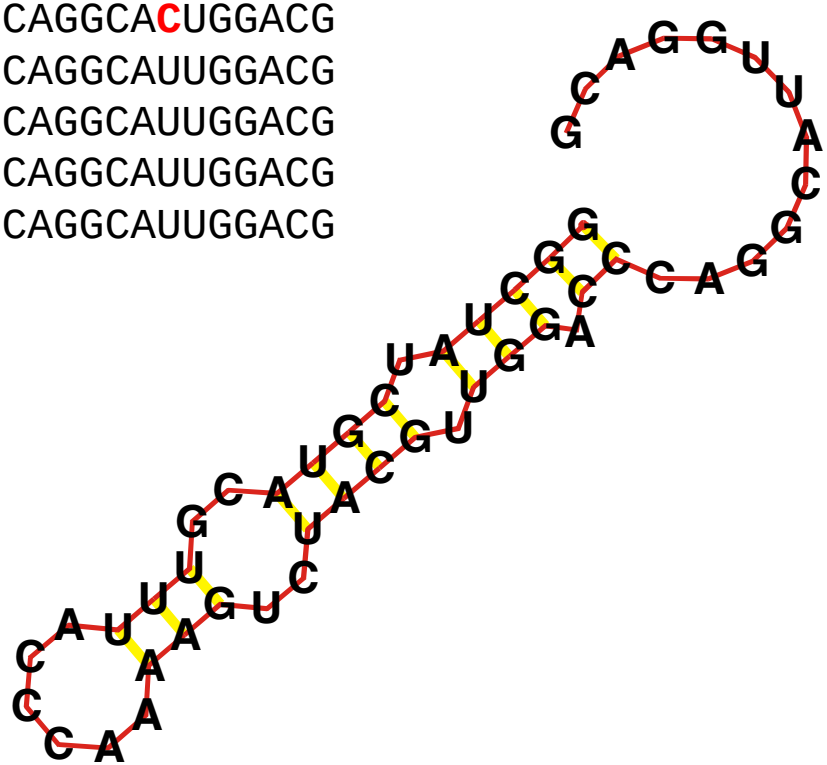
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG



GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG

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

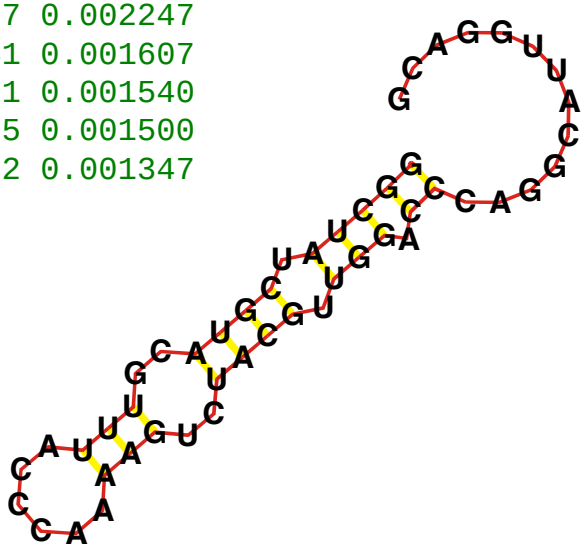
GGCUAUCGUA**U**GUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUU**A**GACG
GGCUAUCGUACGUUUAC**U**CAAAAGUCUACGUUGGACCCAGGCAUUGGACG
GGCUAUCGUACG**C**UUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG
GGC**C**AUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG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



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

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

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



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

1. Ruggedness of molecular landscapes
2. **Replication-mutation dynamics**
3. Models of fitness landscapes
4. Ruggedness and error thresholds
5. Stochasticity of replication and mutation
6. Population dynamics on neutral networks

DIE NATURWISSENSCHAFTEN

58. Jahrgang, 1971 — Heft 10 Oktober

Selforganization of Matter and the Evolution of Biological Macromolecules

MANFRED EIGEN*

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

I. Introduction

I.1. „Cause and Effect"

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

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

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

Die Naturwissenschaften 64. Jahrgang Heft 11 November 1977

The Hypercycle

A Principle of Natural Self-Organization

Part A: Emergence of the Hypercycle

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

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

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

Preview on Part B: The Abstract Hypercycle

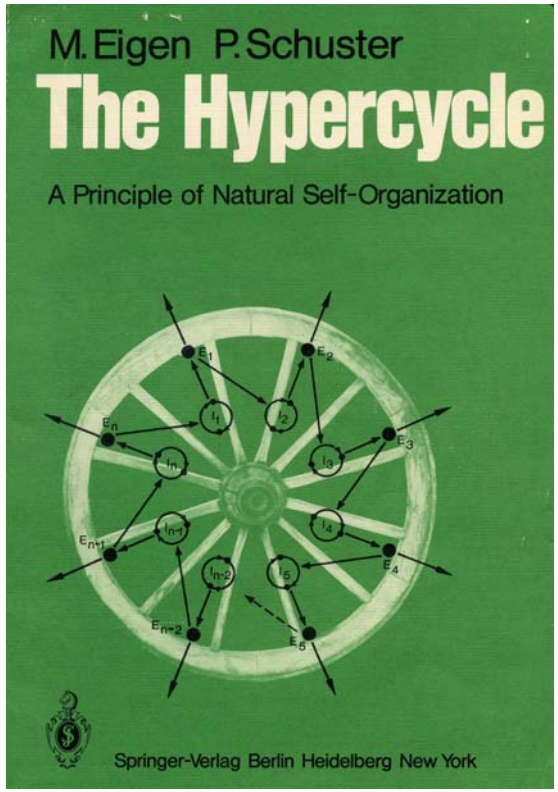
The mathematical analysis of dynamical systems using methods of differential topology, yields the result that there is only one type of mechanisms which fulfills the following requirements: The information stored in each single replicative unit (or reproductive cycle) must be maintained, i.e., the respective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

Preview on Part C: The Realistic Hypercycle

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

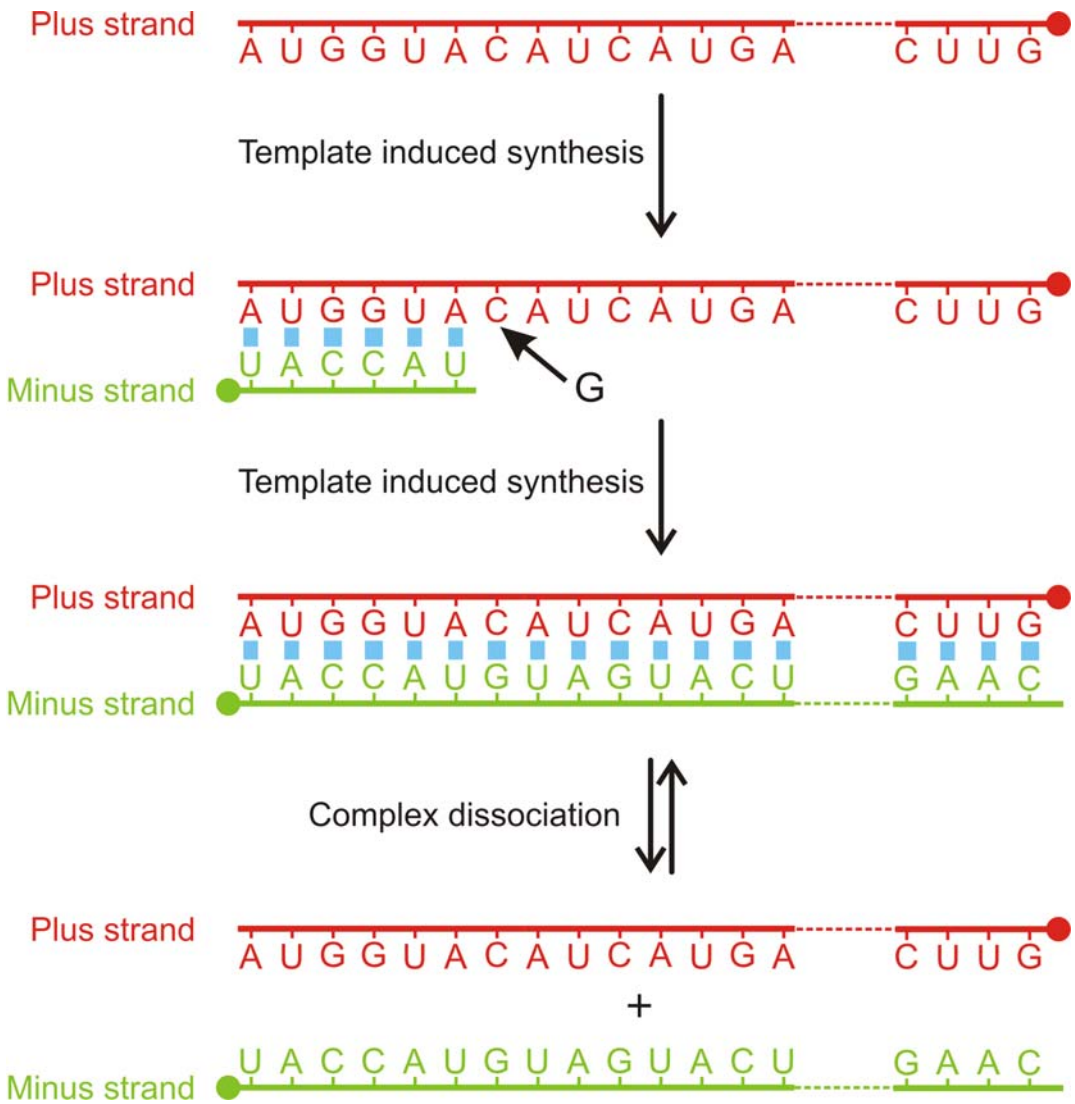
I. The Paradigm of Unity and Diversity in Evolution

Why do millions of species, plants and animals, exist, while there is only one basic molecular machinery of the cell: one universal genetic code and unique chiralities of the macromolecules?
The geneticists of our day would not hesitate to give an immediate answere to the first part of this question. Diversity of species is the outcome of the tremendous branching process of evolution with its myriads of single steps of reproduction and mutation. It in-



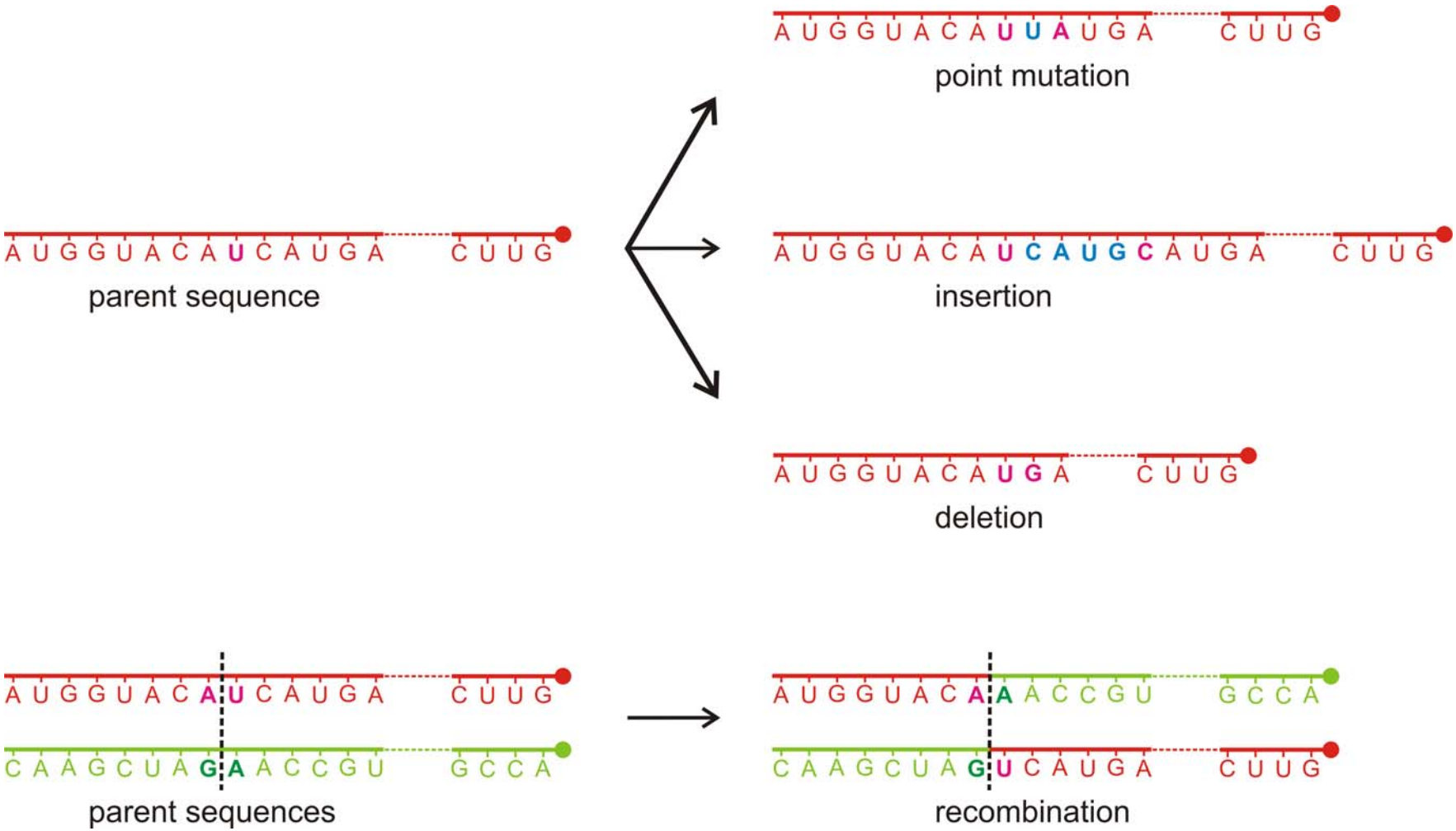
Chemical kinetics of molecular evolution

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

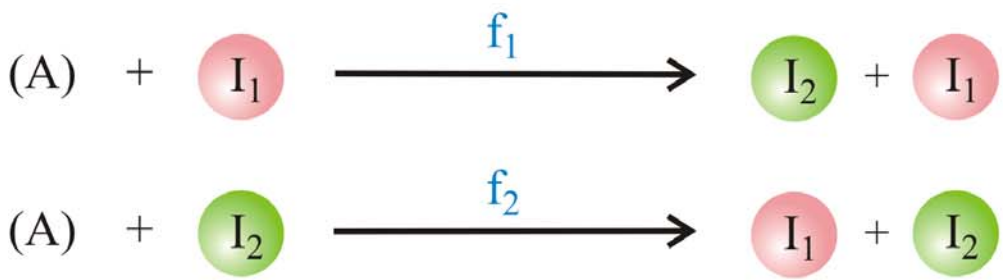


Complementary replication is the simplest copying mechanism of RNA.
Complementarity is determined by Watson-Crick base pairs:

G≡C and **A=U**



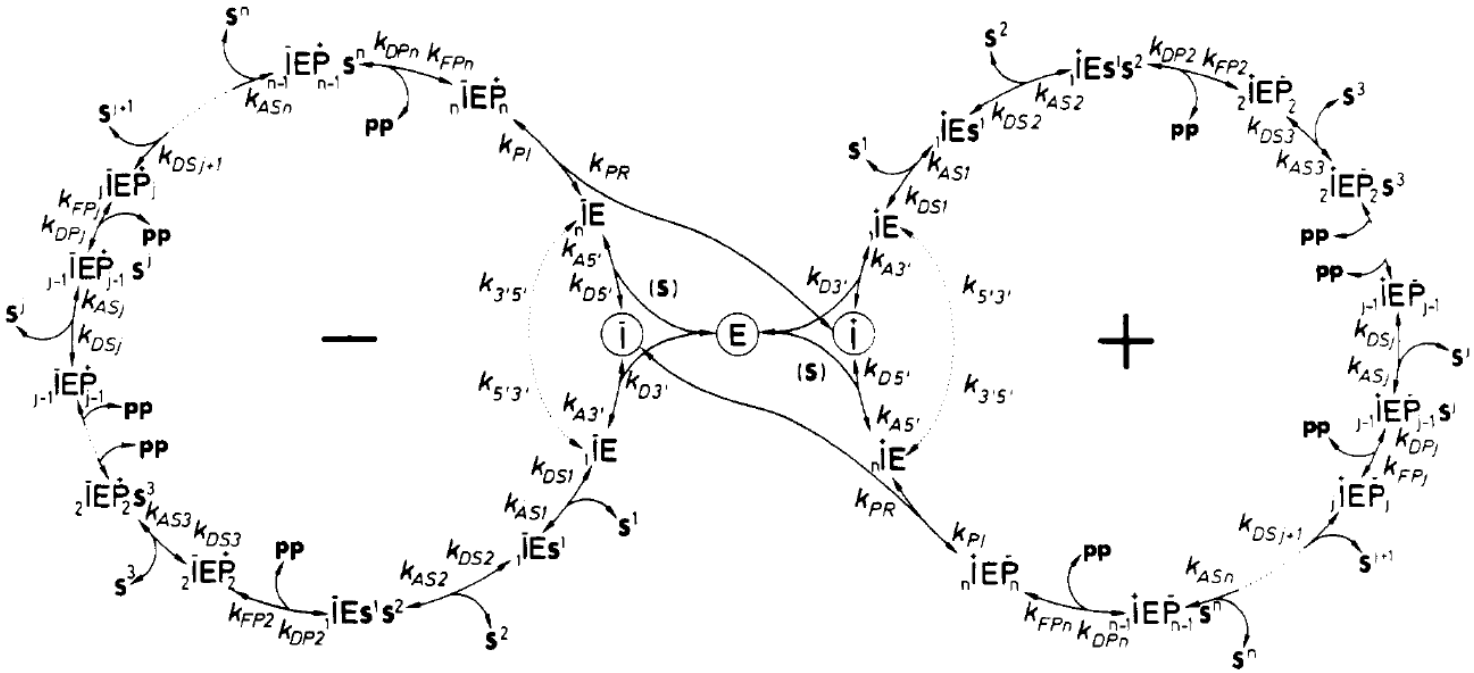
Variation of genotypes through mutation and recombination

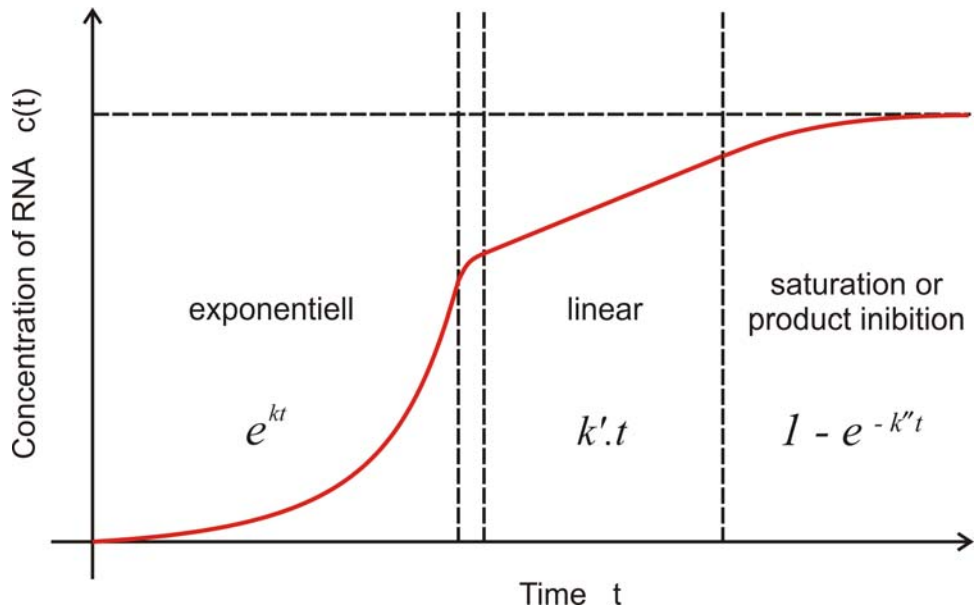


$$dx_1 / dt = f_2 x_2 - x_1 \Phi$$
$$dx_2 / dt = f_1 x_1 - x_2 \Phi$$

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

Complementary replication as the simplest molecular mechanism of reproduction





Kinetics of RNA replication

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

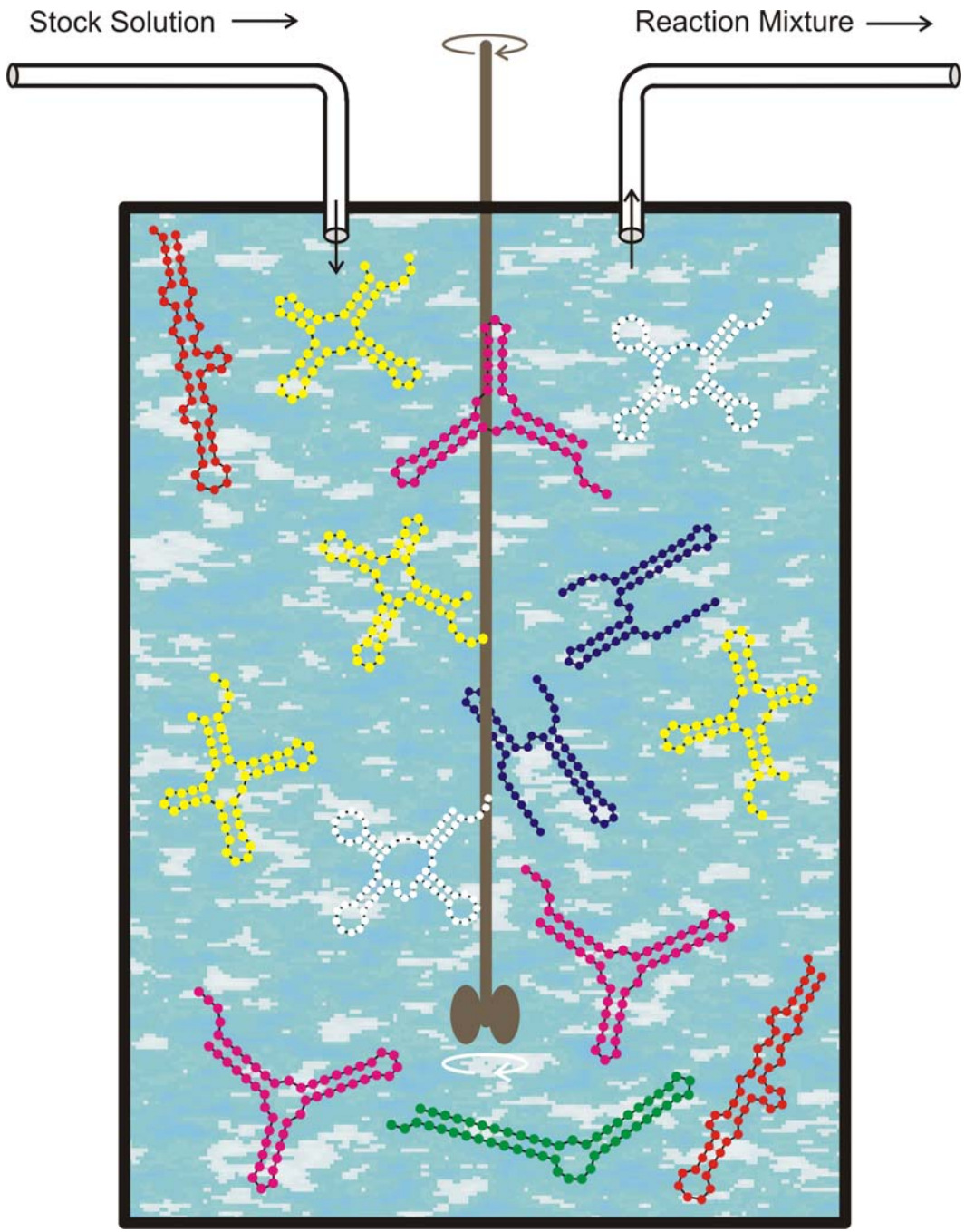
Stock solution:

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

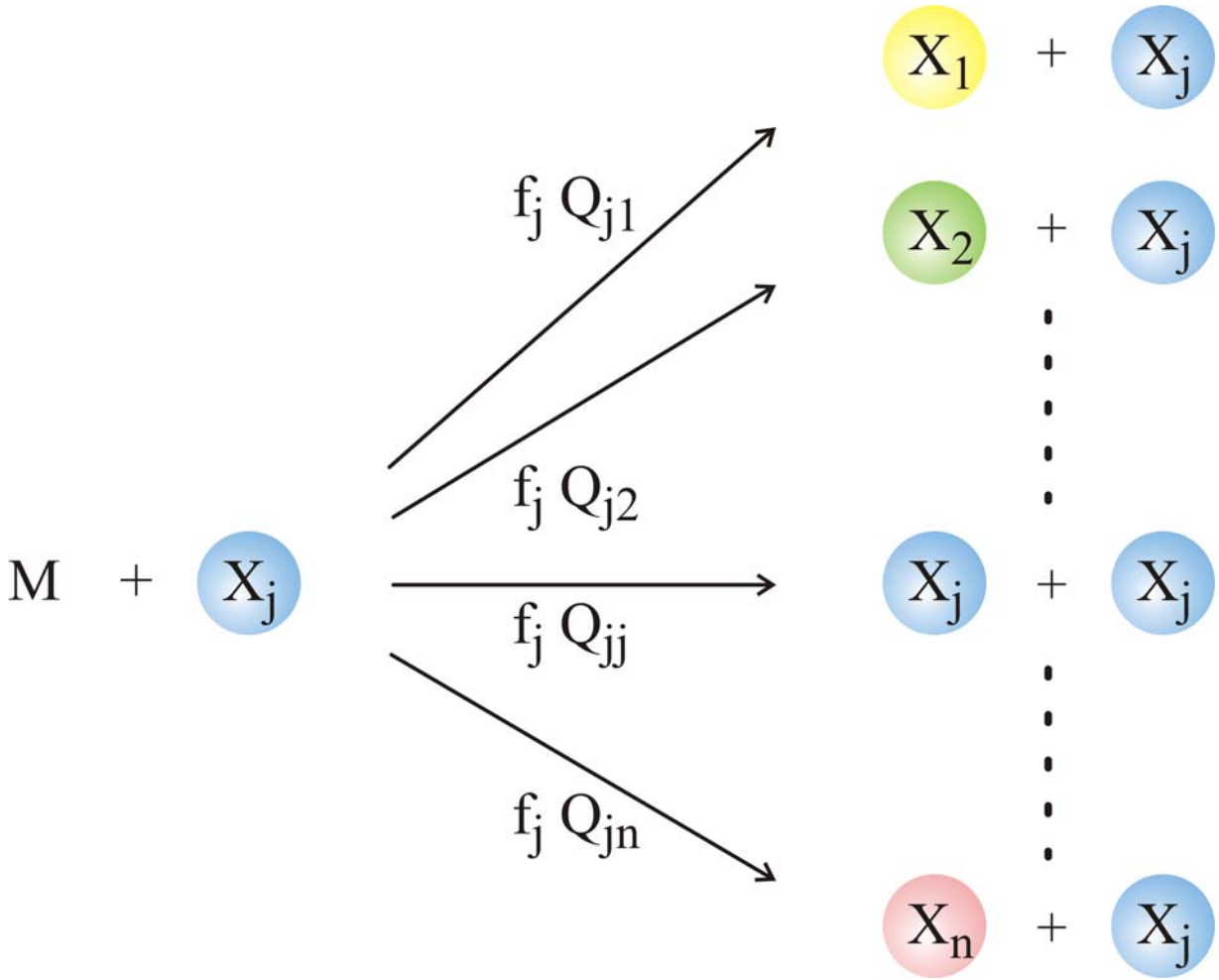
Flow rate: $r = \tau_R^{-1}$

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

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



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



Chemical kinetics of replication and mutation as parallel reactions

$$\frac{dx_j}{dt} = \sum_{i=1}^{n} Q_{ji}\, f_i\, x_i - x_j\, \Phi \quad \text{with} \quad \Phi = \sum_{i=i}^{n} f_i\, x_i$$

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

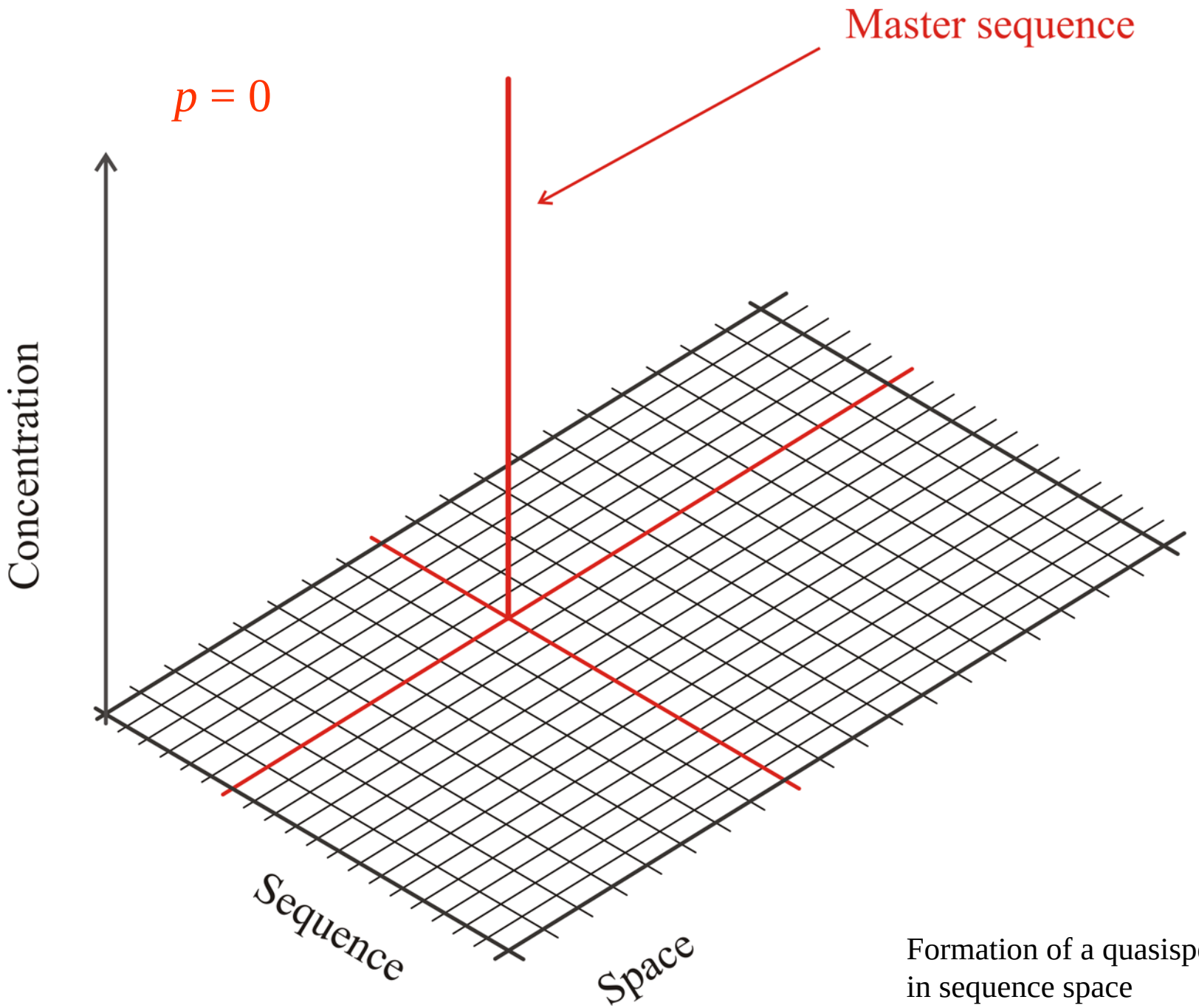
Uniform error rate model

$$Q_{ij} = (1-p)^{n-d_H(X_i,X_j)}\; p^{d_H(X_i,X_j)},\; p \ldots \text{error rate per digit}$$

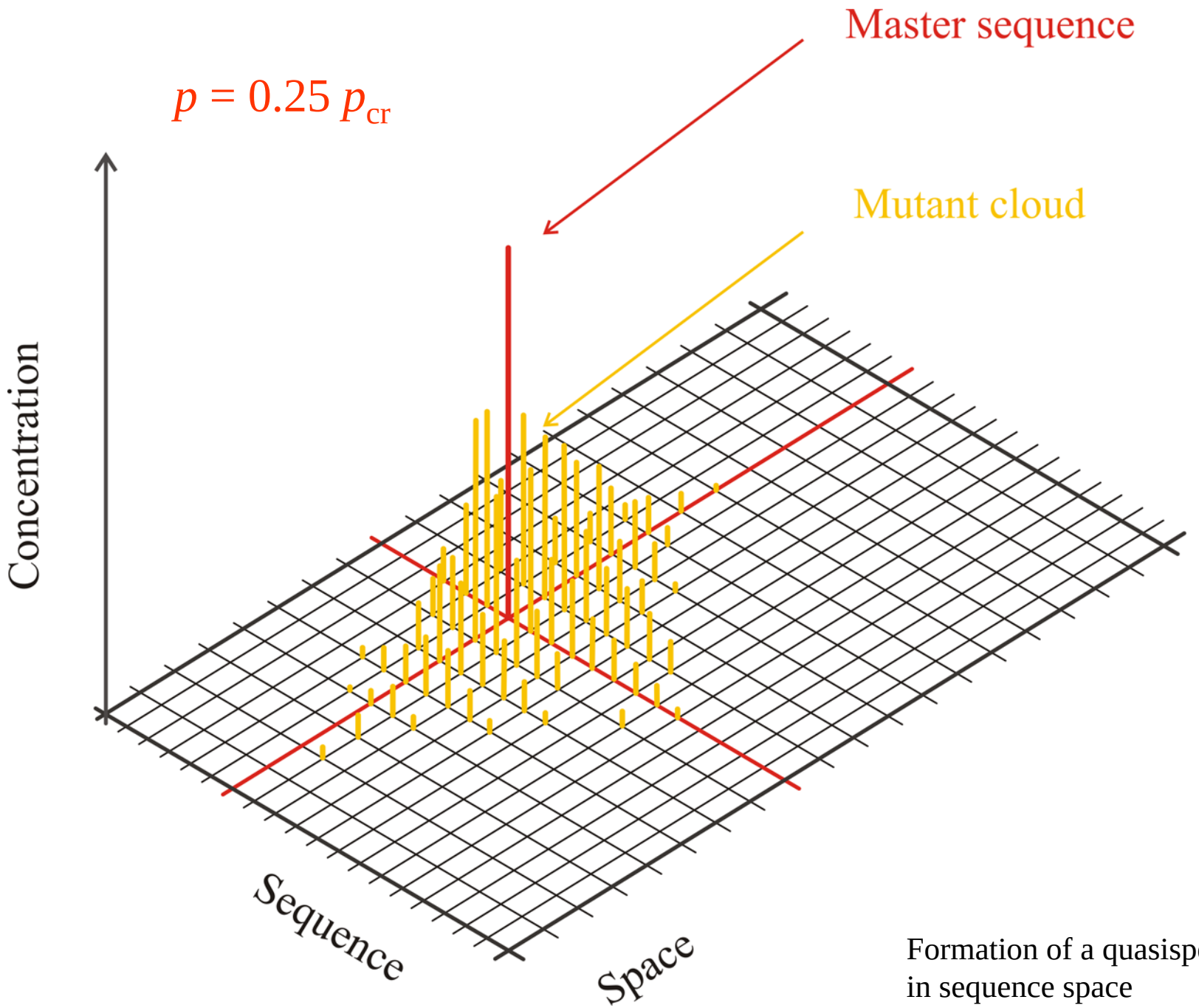
$$d_H(X_i,X_j) \ldots \text{Hamming distance between } X_i \text{ and } X_j$$

$$\sum_{j=1}^{n} Q_{ji} = 1$$

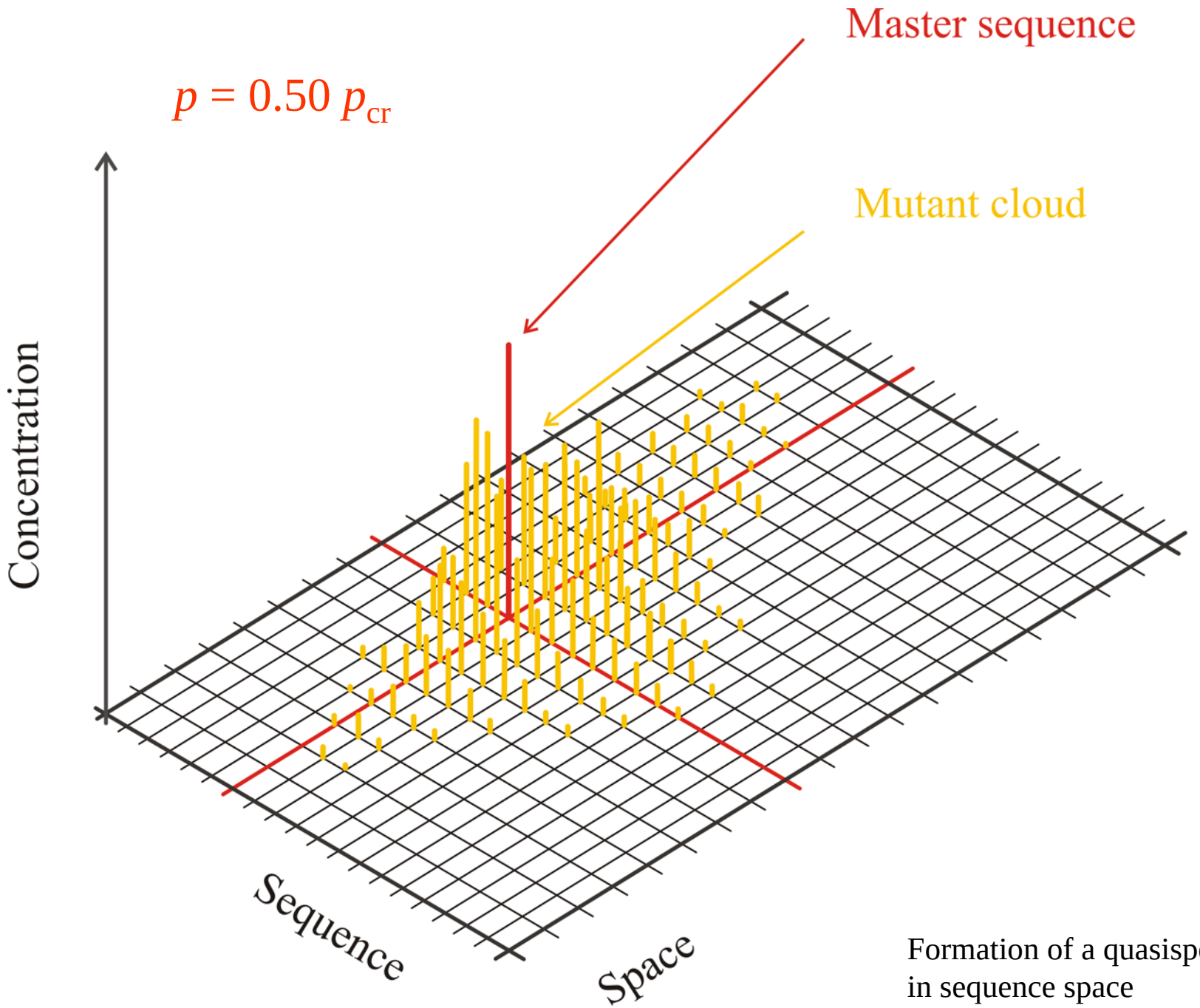
The replication-mutation equation



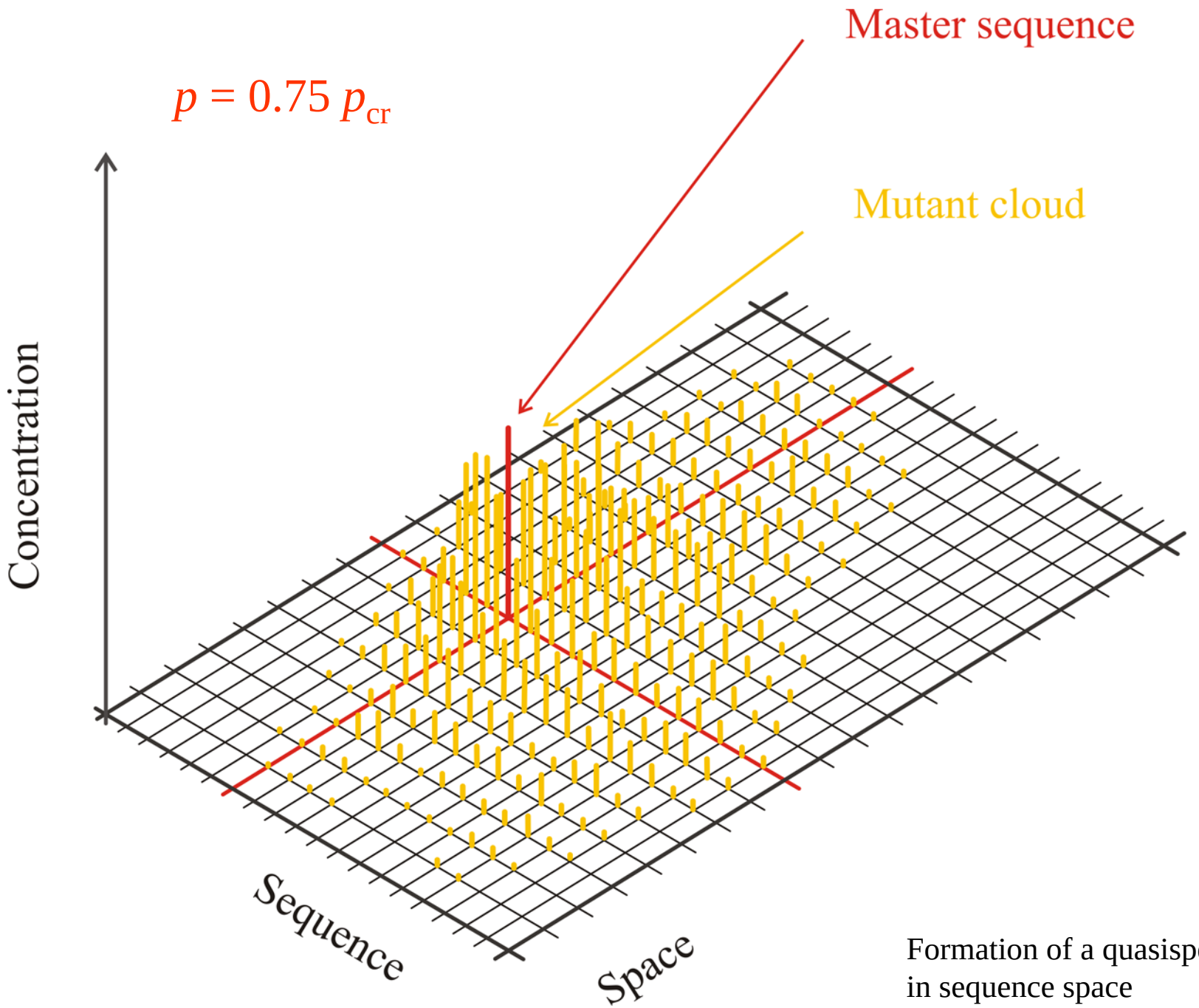
Formation of a quasispecies in sequence space



Formation of a quasispecies in sequence space

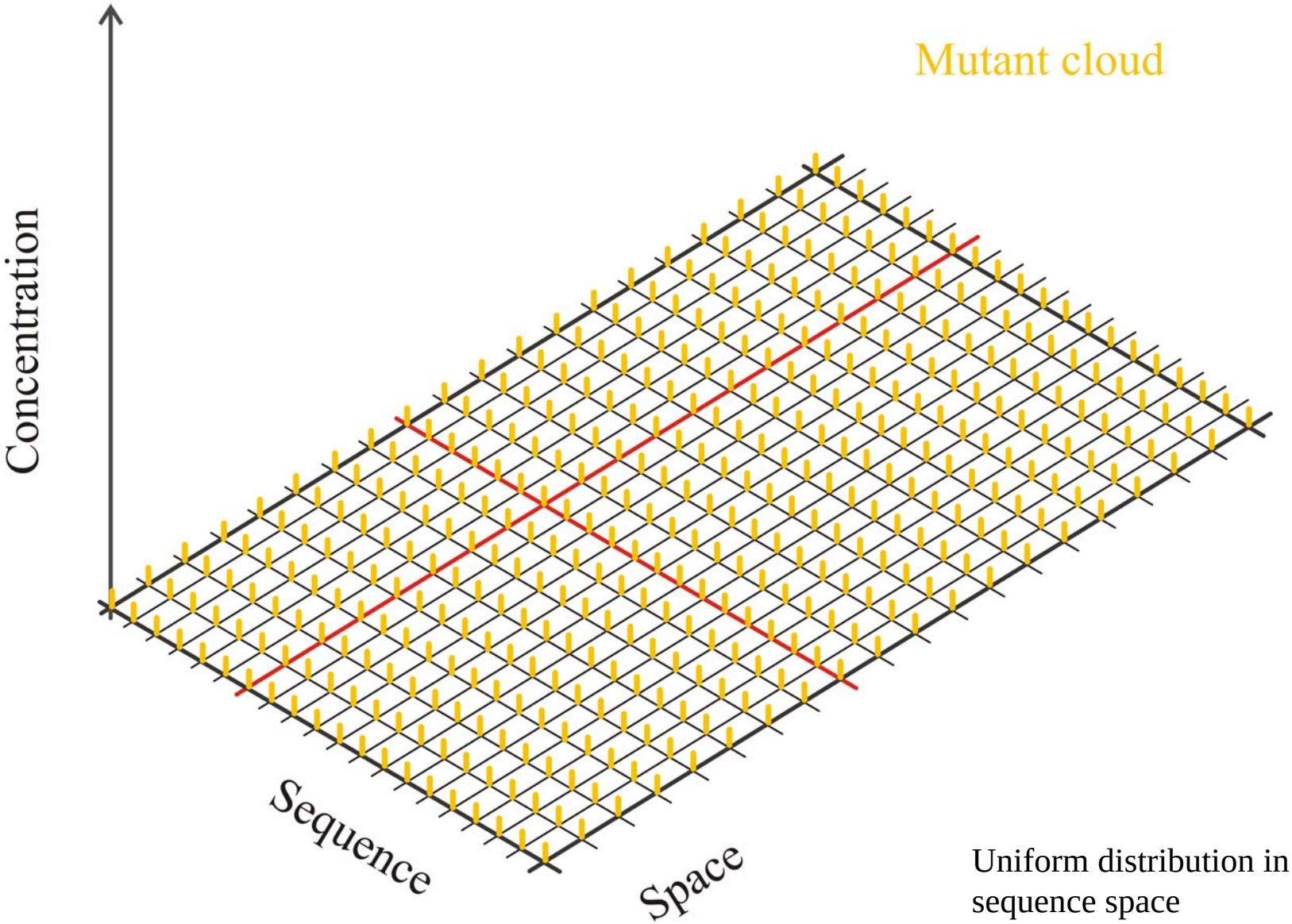


Formation of a quasispecies in sequence space



Formation of a quasispecies in sequence space

$p \geq p_{cr}$



Uniform distribution in sequence space

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.

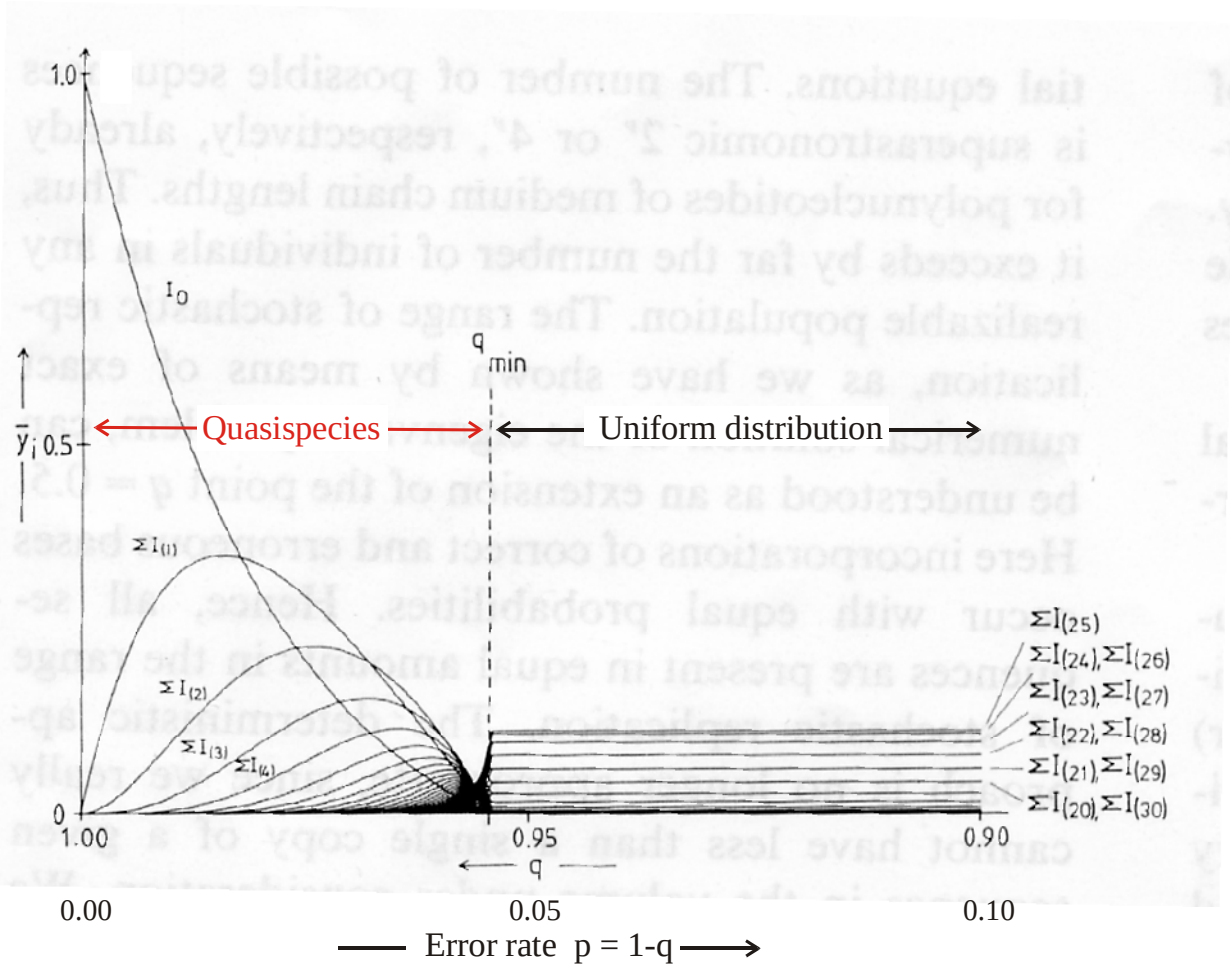
Rigorous mathematical analysis has been performed on eq. 1 [7,15,24,26]. In particular, it was shown that the non-linearity of eq. 1 can be removed by an appropriate transformation. The eigenvalue problem of the linear differential equation obtained thereby may be solved approximately by the conventional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

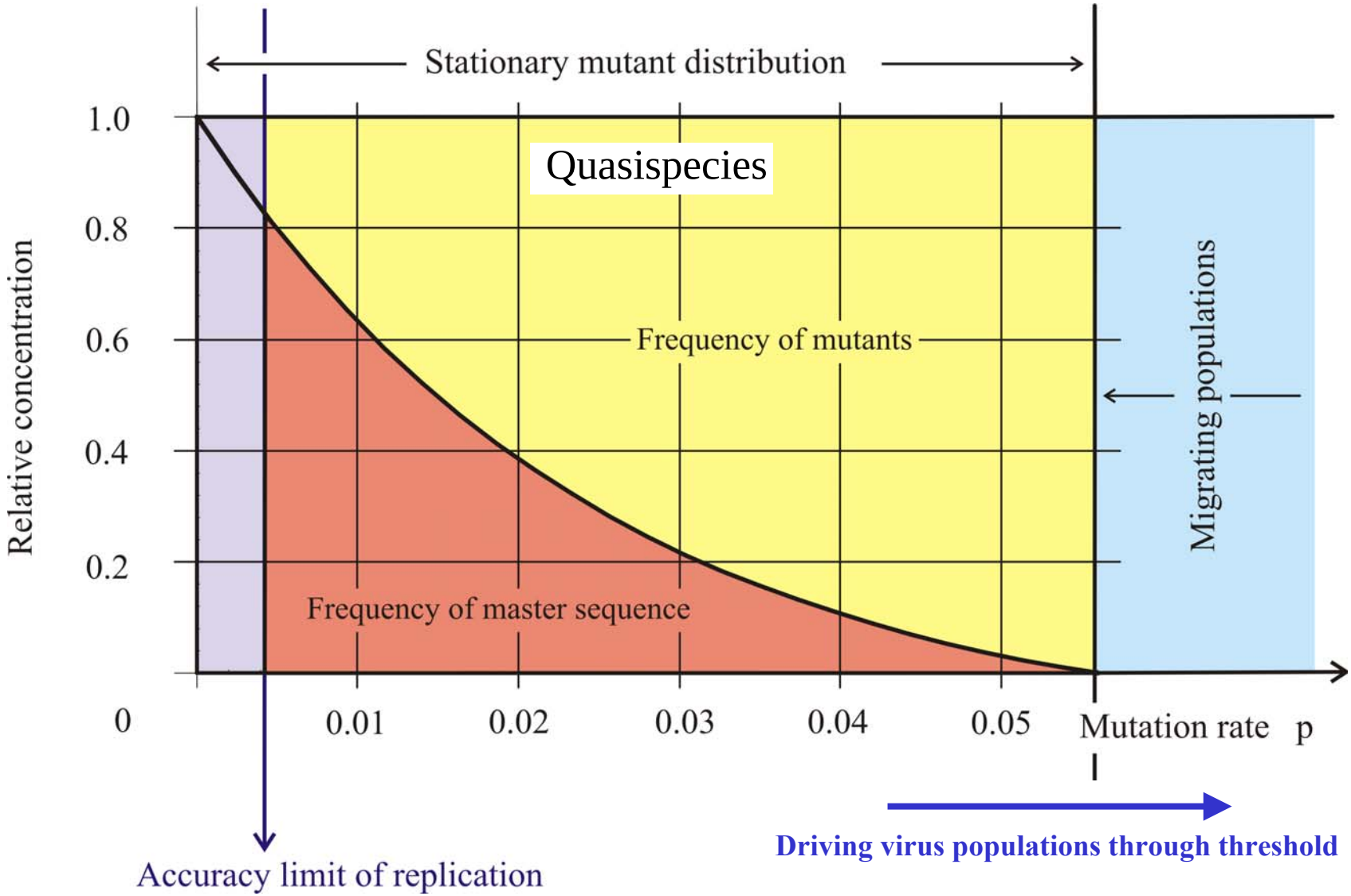
** This paper is considered as part II of Model Studies on RNA replication. Part I is by Gassner and Schuster [14].

† 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

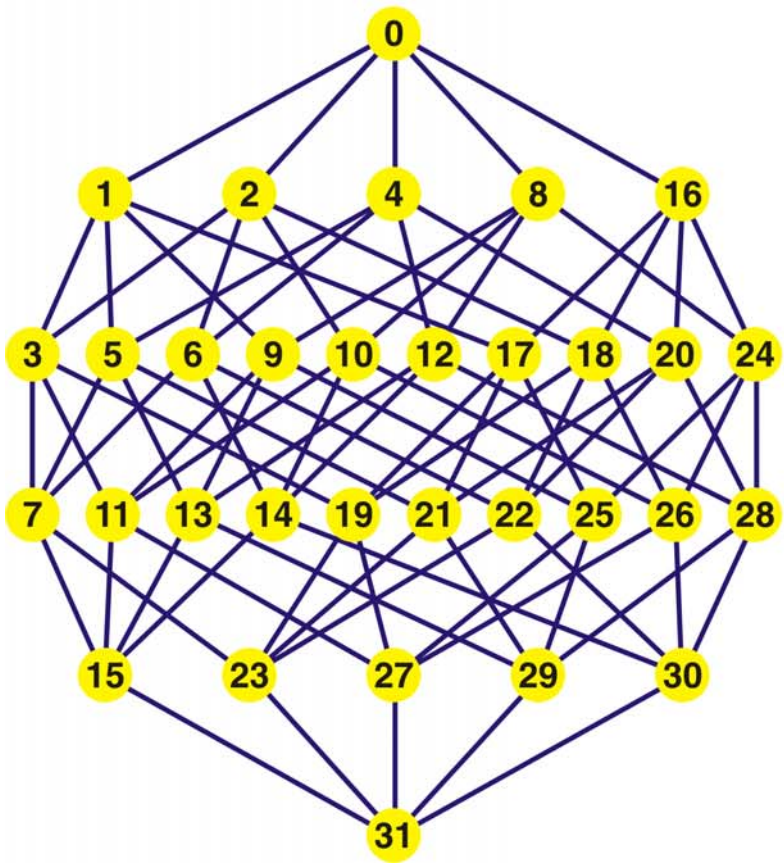


Stationary population or quasispecies as a function of the mutation or error rate *p*



The error threshold in replication

1. Ruggedness of molecular landscapes
2. Replication-mutation dynamics
3. **Models of fitness landscapes**
4. Ruggedness and error thresholds
5. Stochasticity of replication and mutation
6. Population dynamics on neutral networks



Mutant class

0

1

2

3

4

5

Binary sequences can be encoded by their decimal equivalents:

C = 0 and G = 1, for example,

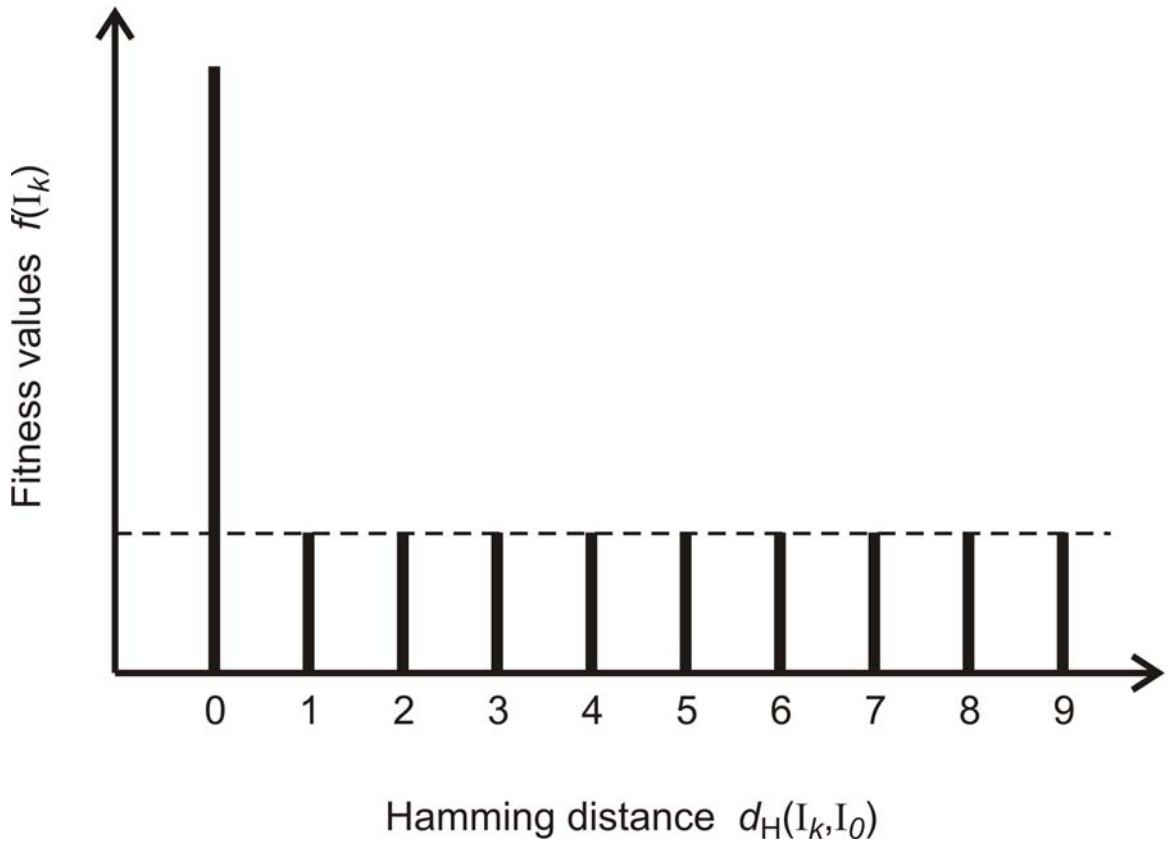
"0" ≡ 00000 = CCCCC,

"14" ≡ 01110 = CGGGC,

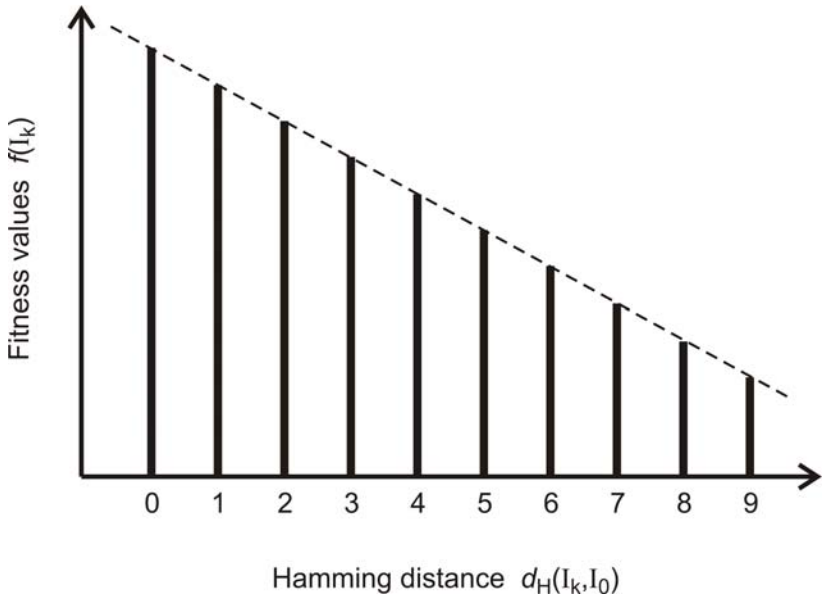
"29" ≡ 11101 = GGGCG, etc.

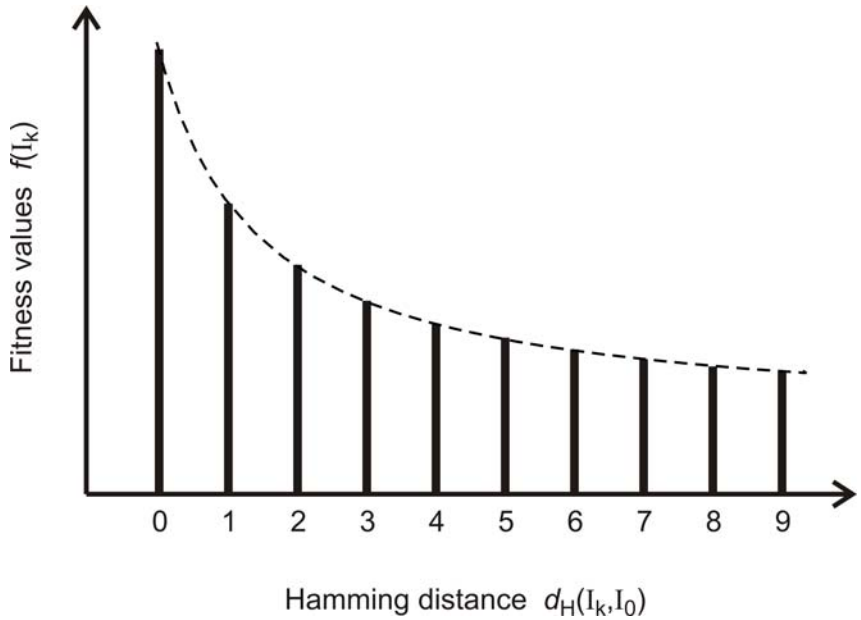
Every point in sequence space is equivalent

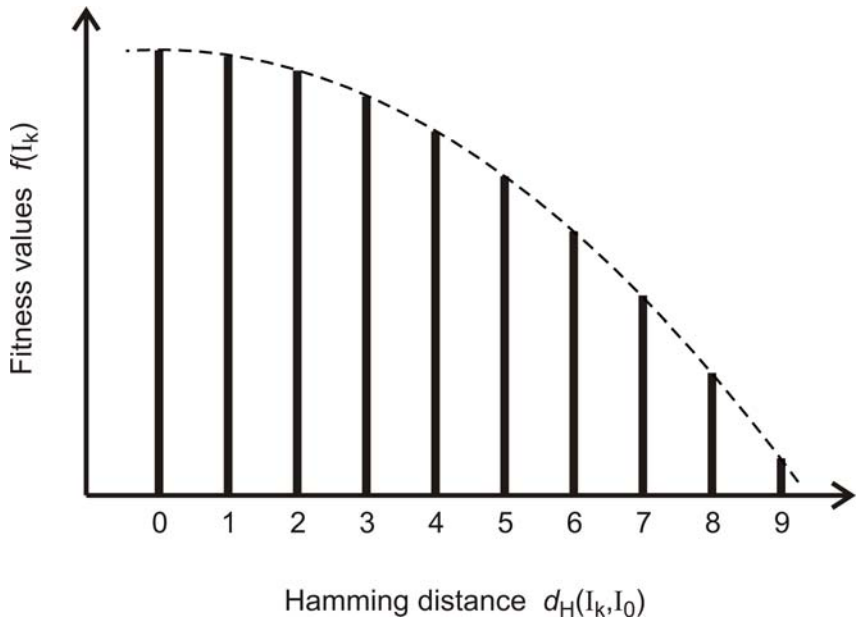
Sequence space of binary sequences with chain length $n = 5$



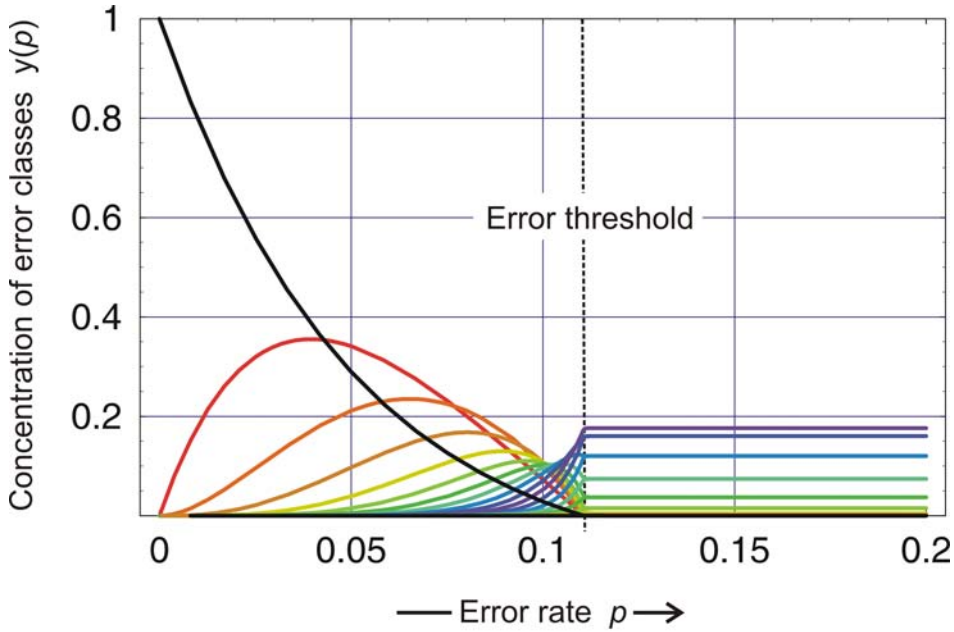
A fitness landscape showing an error threshold

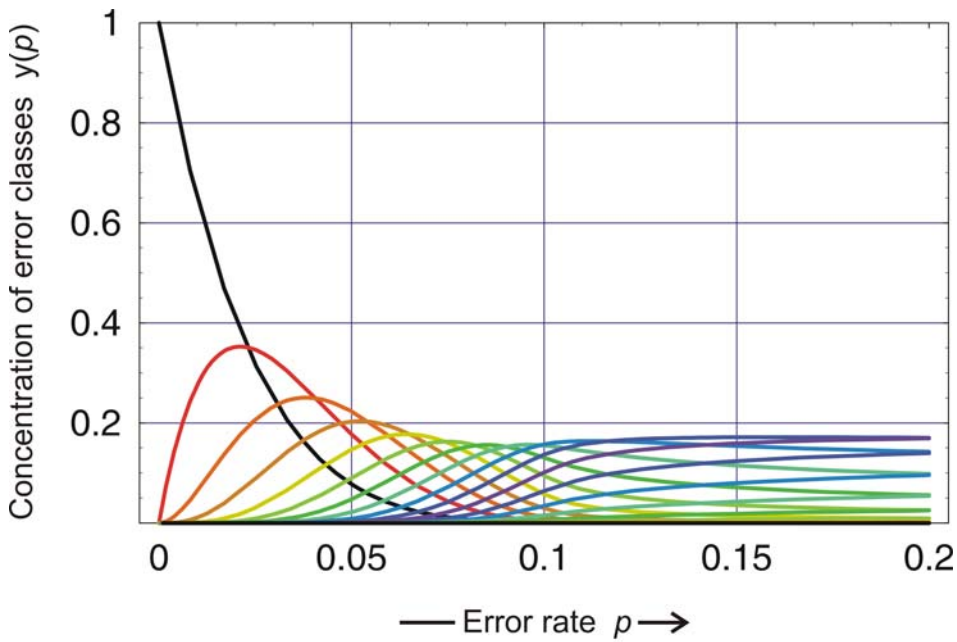






Fitness landscapes **not** showing error thresholds





Error thresholds and gradual transitions

n = 20 and $\sigma = 10$

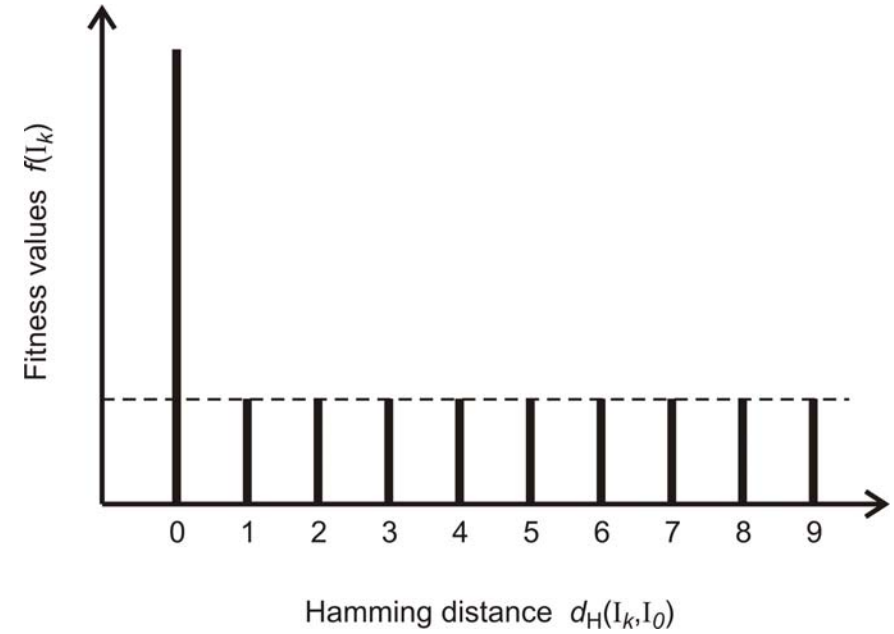
1. Ruggedness of molecular landscapes
2. Replication-mutation dynamics
3. Models of fitness landscapes
4. **Ruggedness and error thresholds**
5. Stochasticity of replication and mutation
6. Population dynamics on neutral networks

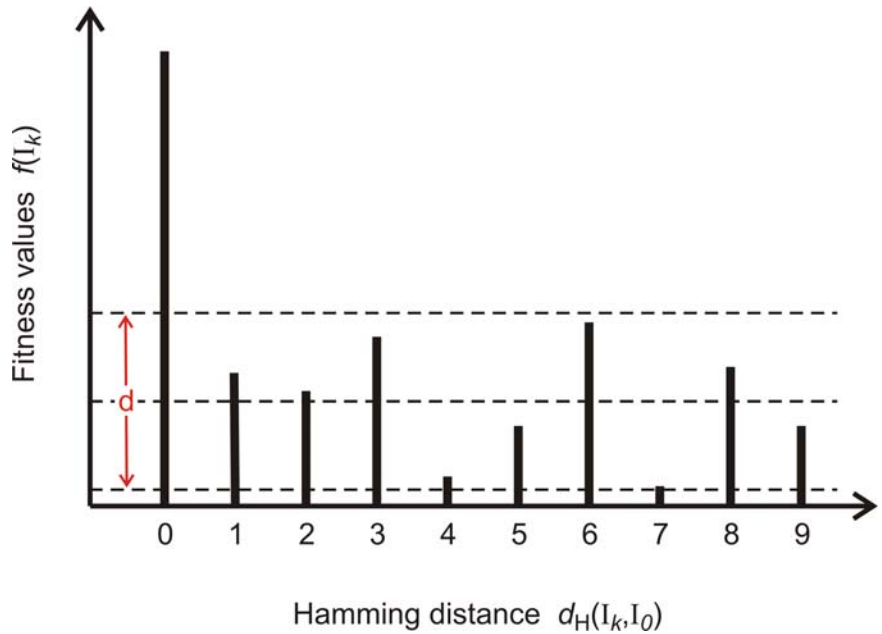
Sources of ruggedness:

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

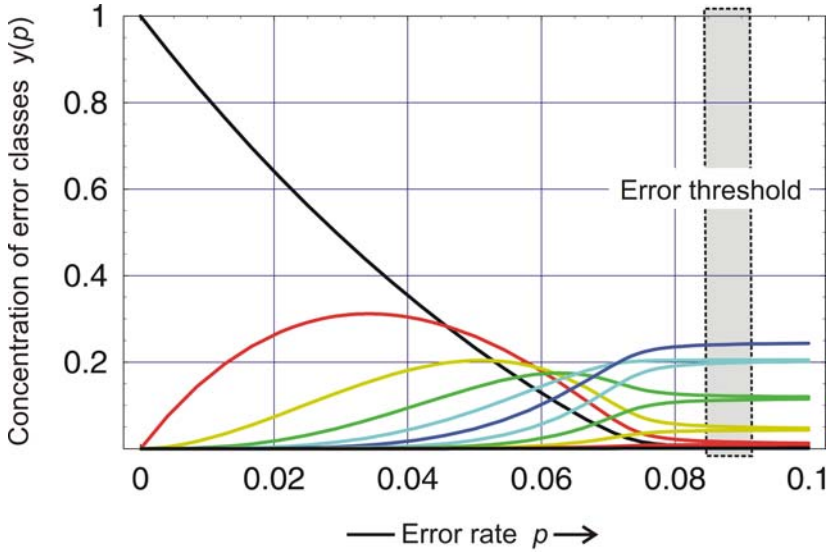
Three sources of ruggedness:

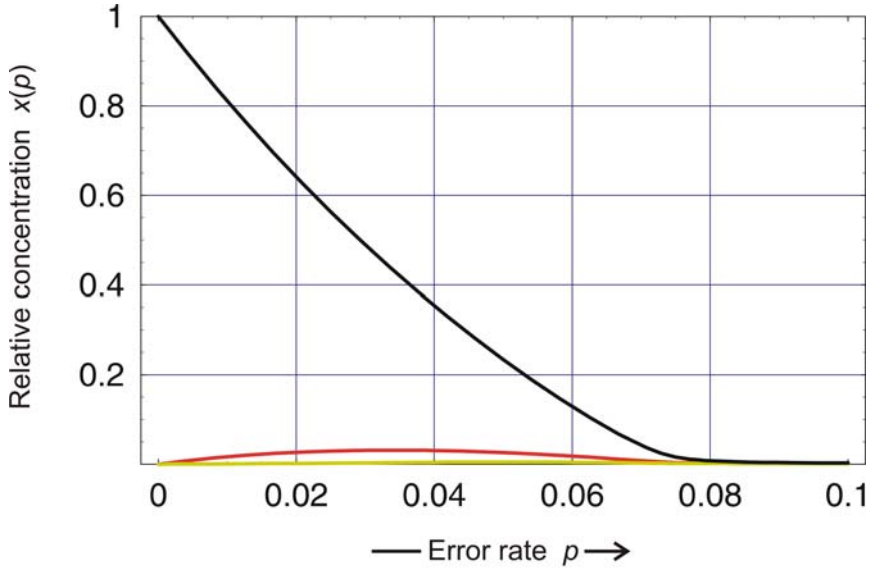
1. **Variation in fitness values**
2. Deviations from uniform error rates
3. Neutrality

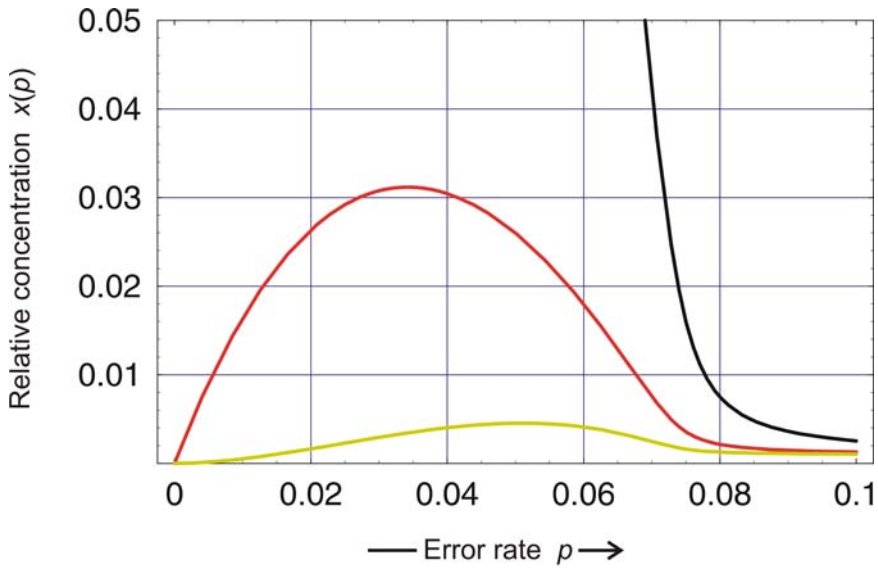




Fitness landscapes showing error thresholds

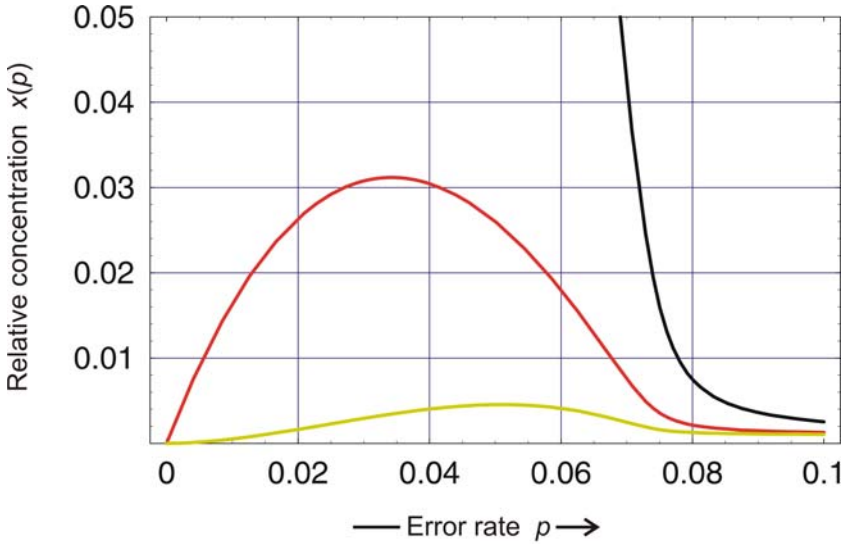


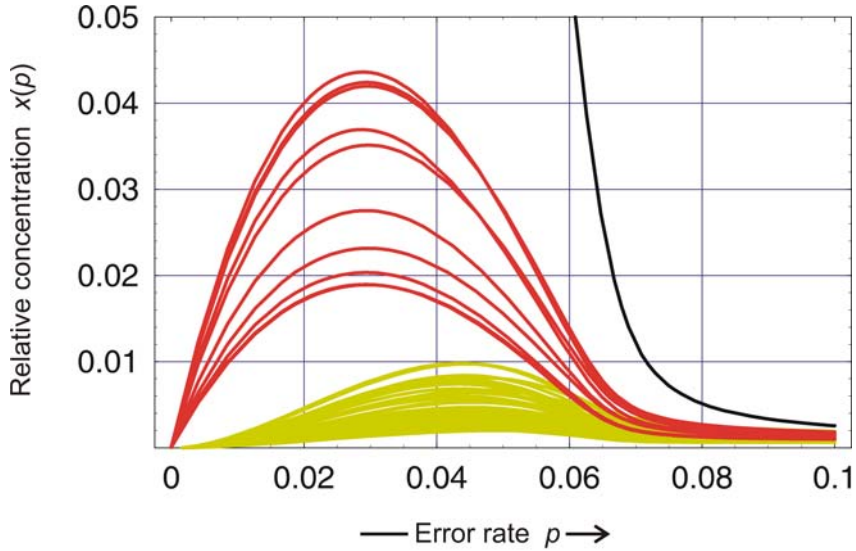


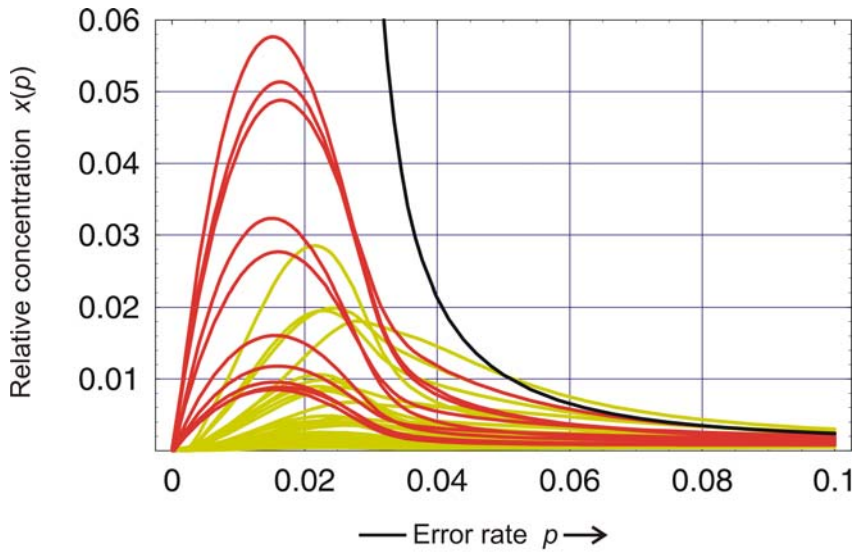


Error threshold: Error classes and individual sequences

n = 10 and $\sigma = 2$

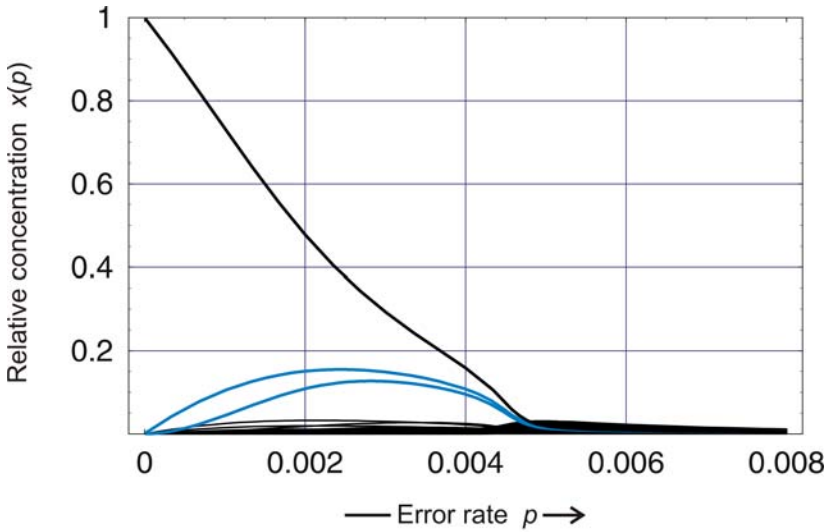


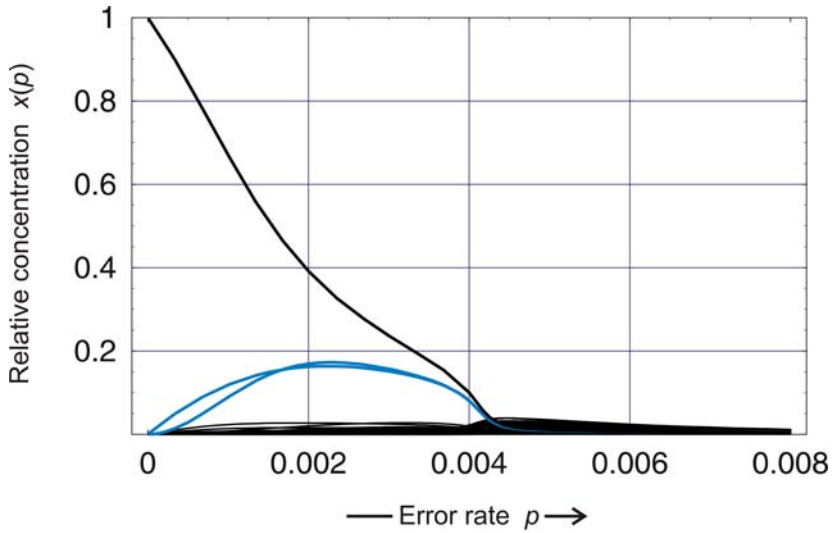


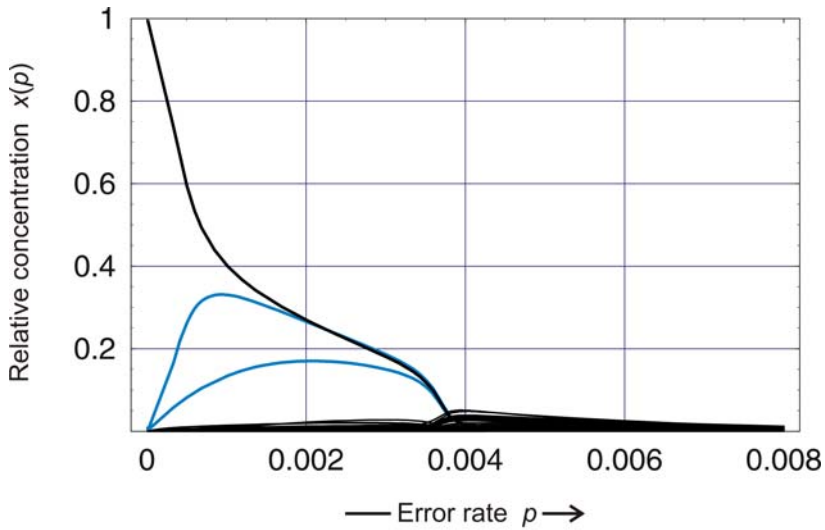


Error threshold: Individual sequences

n = 10, σ = 2 and d = 0, 1.0, 1.85





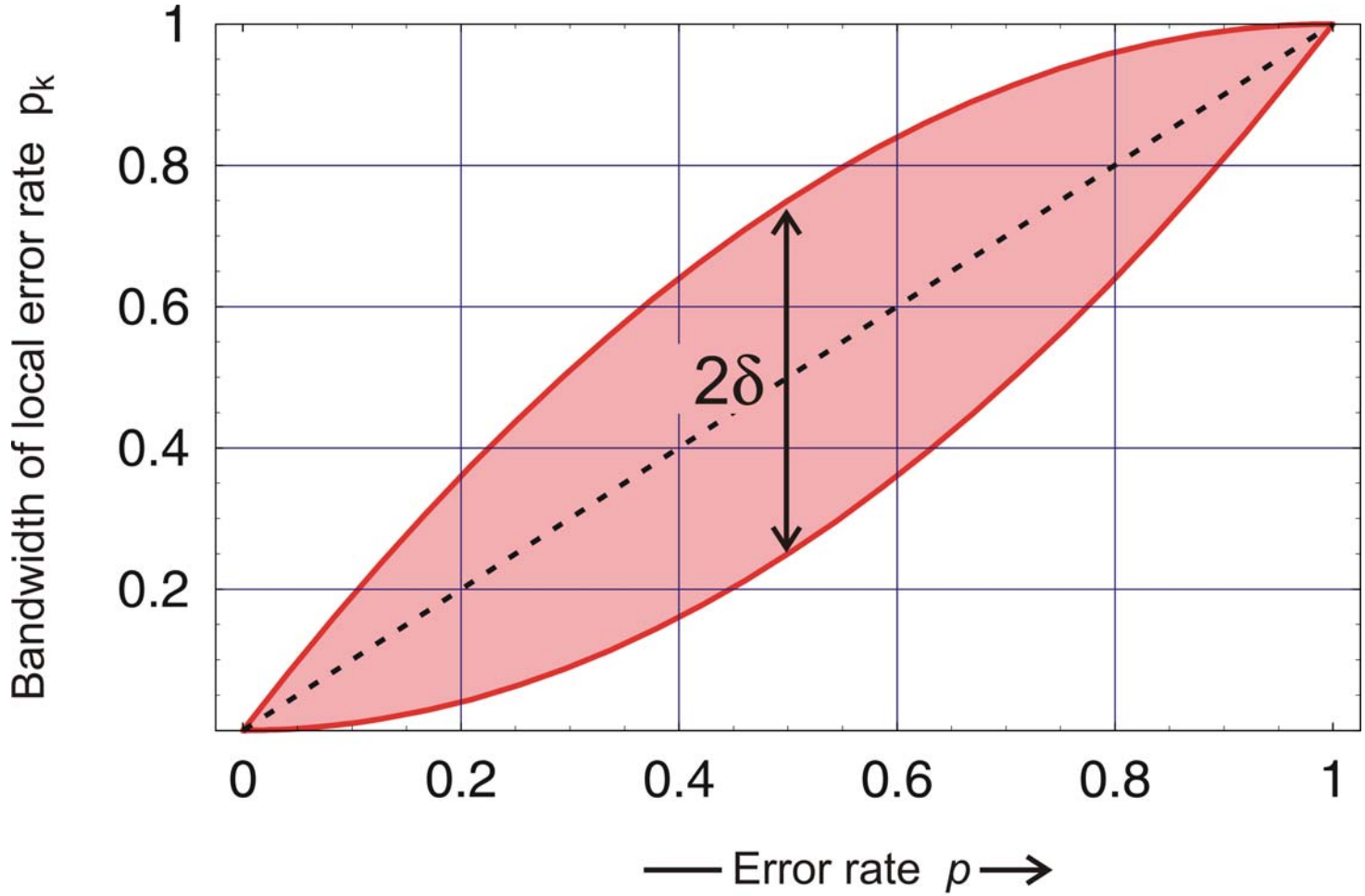


Error threshold: Individual sequences

n = 10, σ = 1.1, d = 1.95, 1.975, 2.00 and seed = 877

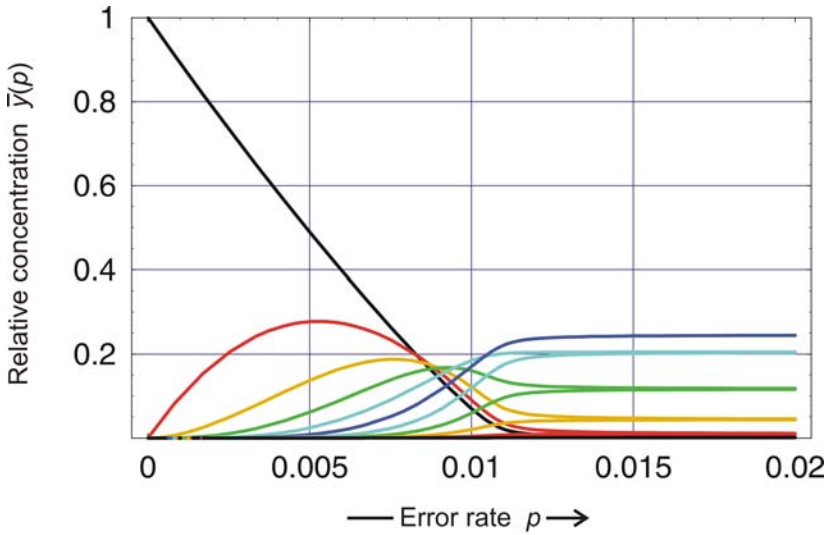
Three sources of ruggedness:

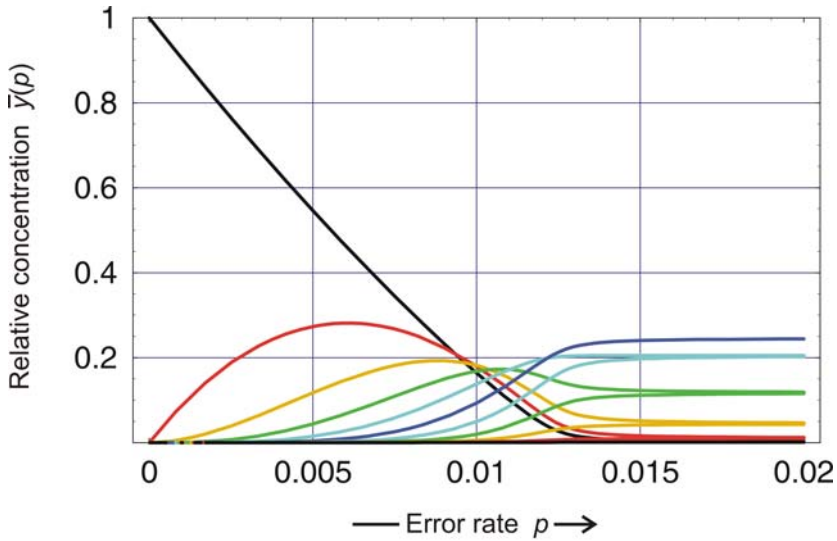
1. Variation in fitness values
2. **Deviations from uniform error rates**
3. Neutrality

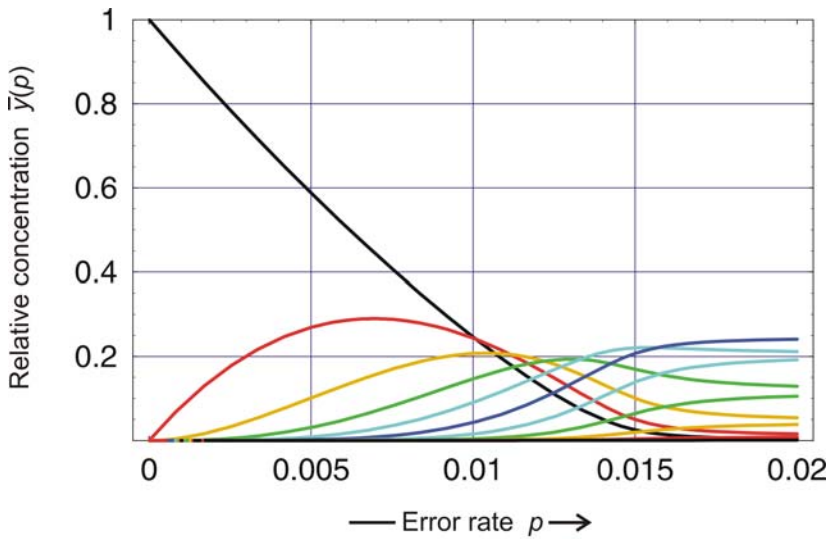


Local replication accuracy p_k:

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





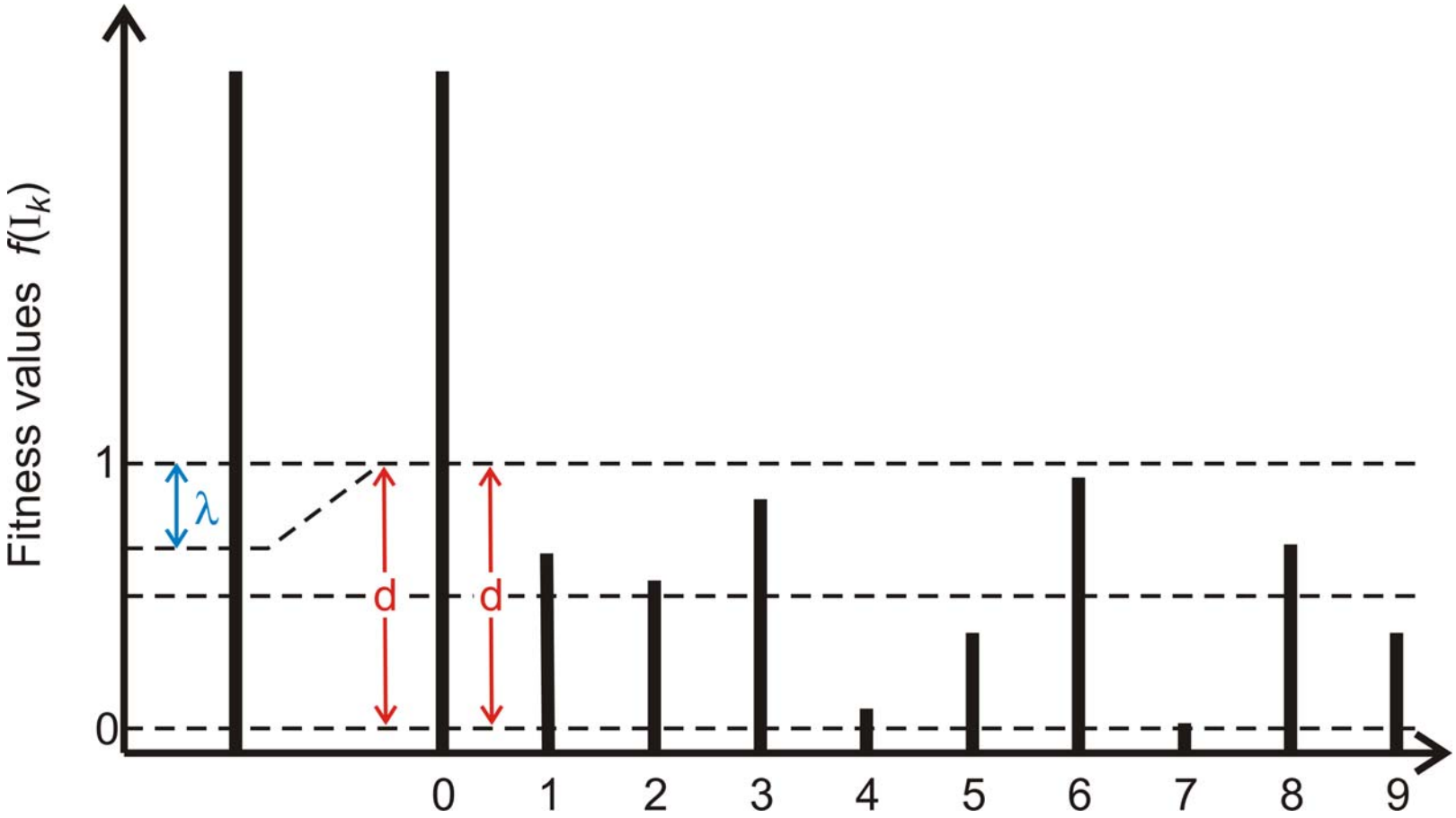


Error threshold: Classes

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

Three sources of ruggedness:

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


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

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.

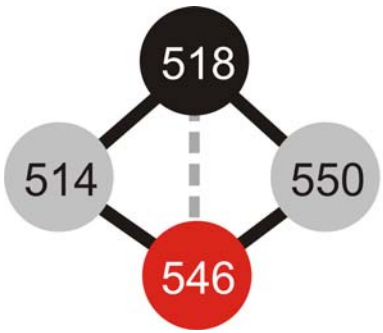
1. Introduction. Recent data from populations of RNA viruses provided direct evidence for vast sequence heterogeneity (Domingo *et al.*, 1987). The origin of this diversity is not yet completely known. It may be caused by the low replication accuracy of the polymerizing enzyme, commonly a virus specific, 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

λ = 0.01, s = 367

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



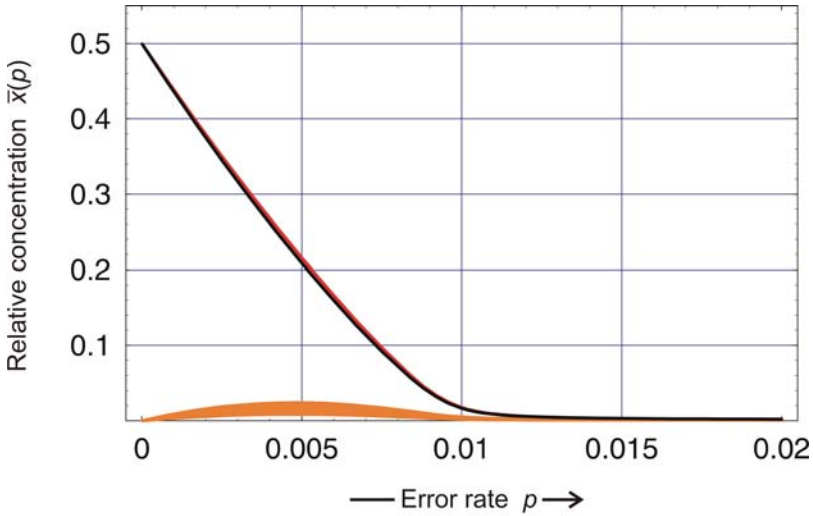
Neutral network

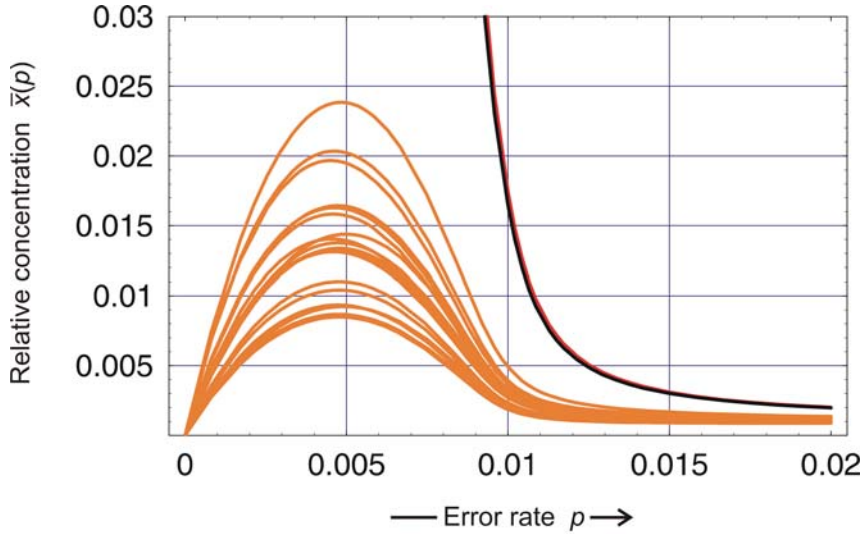
λ = 0.01, s = 877

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

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

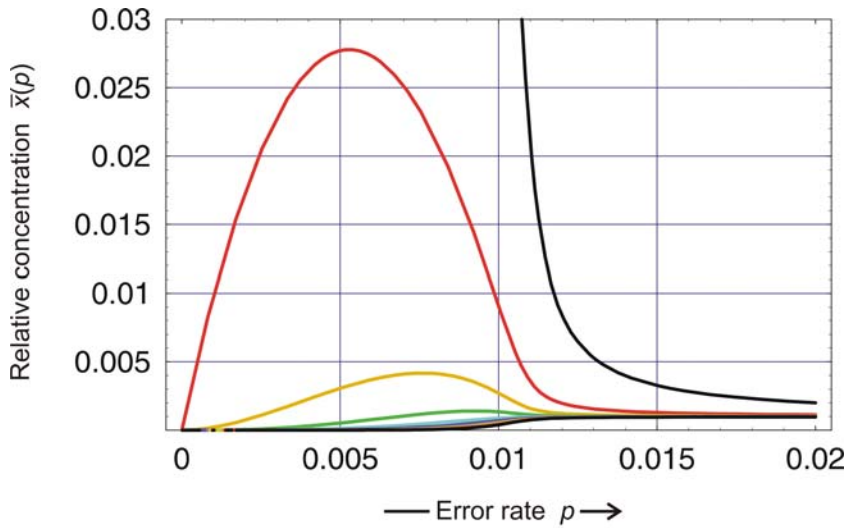
Elements of neutral replication networks





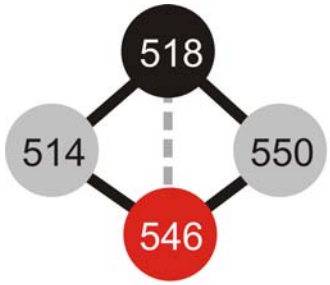
Neutral network

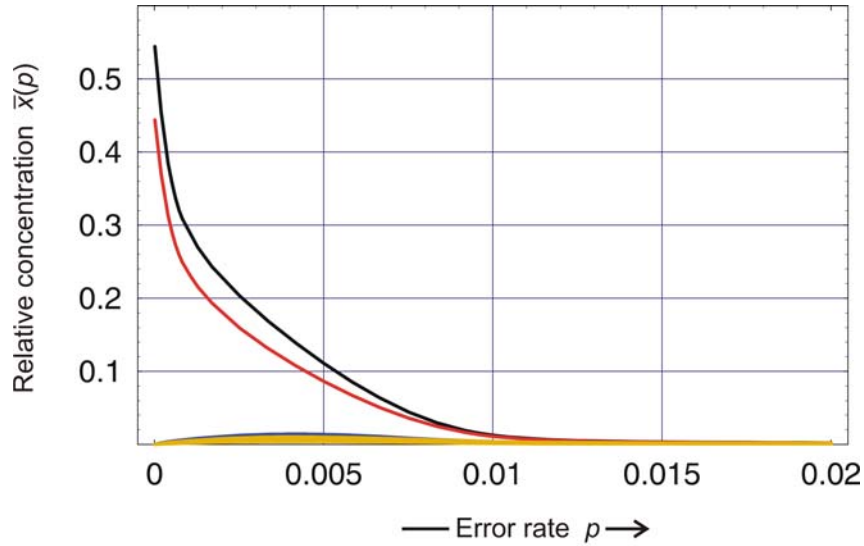
$\lambda = 0.01,\ s = 367$



Error threshold: Individual sequences

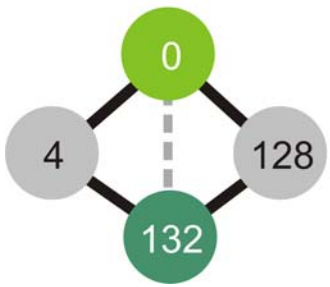
$n = 10,\ \sigma = 1.1,\ d = 1.0$

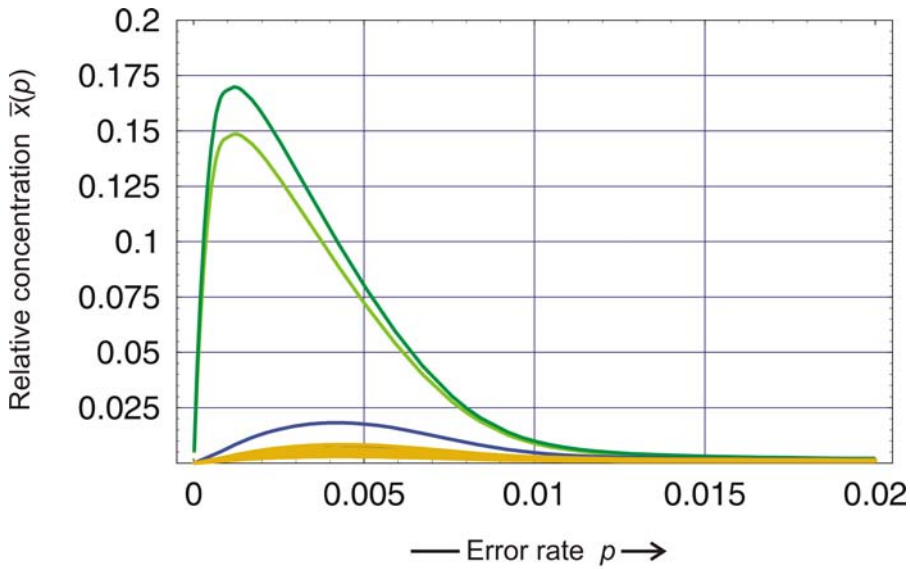




Neutral networks

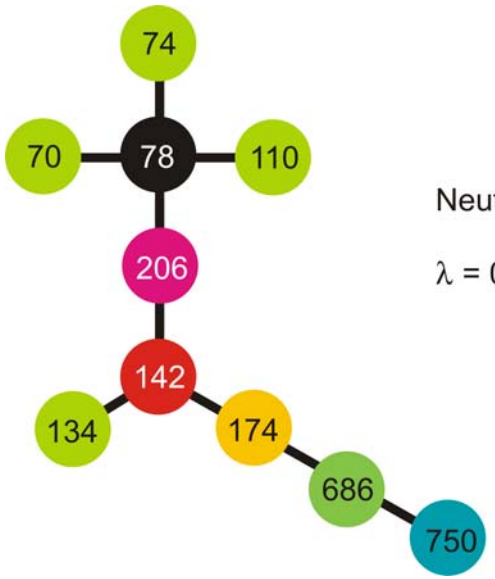
$\lambda = 0.01,\ s = 877$





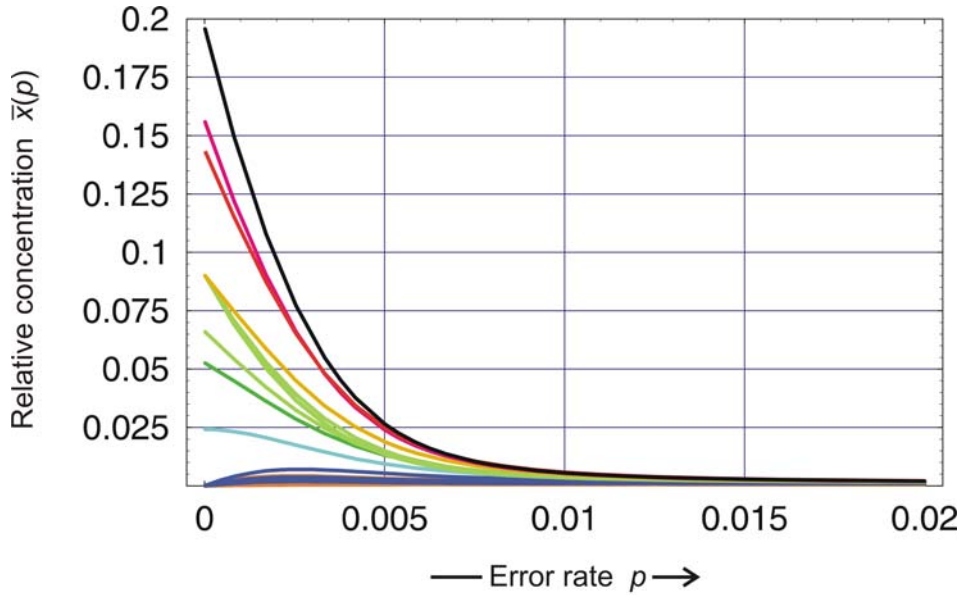
Error threshold: Individual sequences

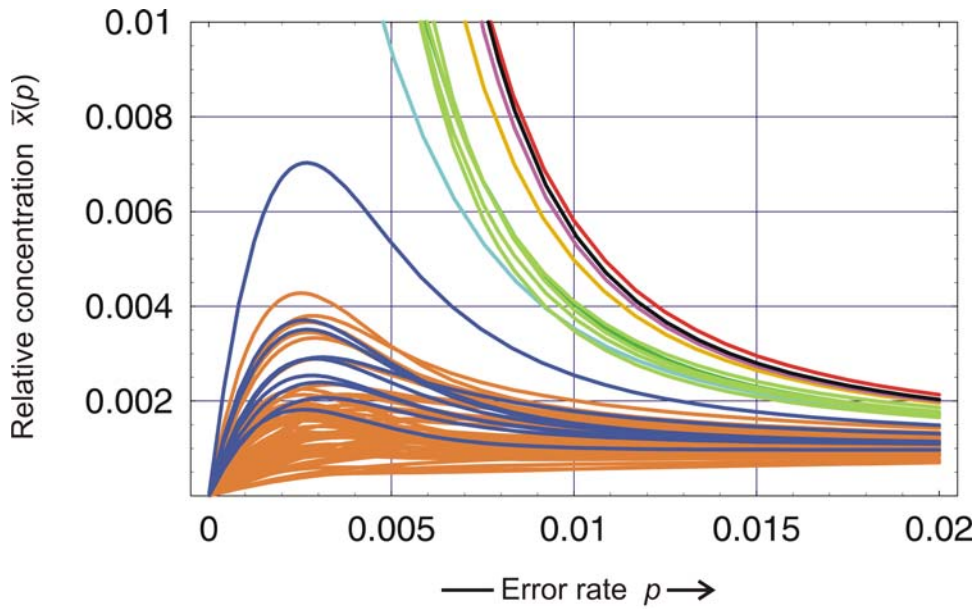
n = 10, σ = 1.1, d = 1.0



Neutral network

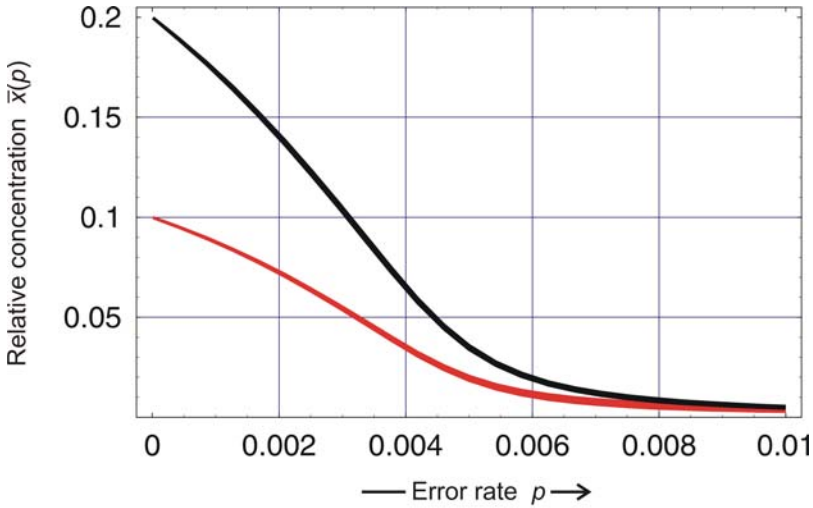
$\lambda = 0.10,\ s = 367$

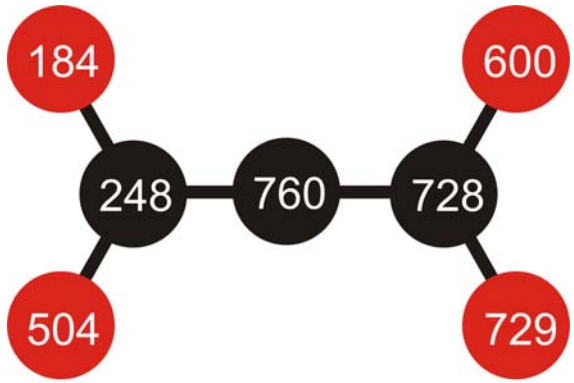




Error threshold: Individual sequences

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





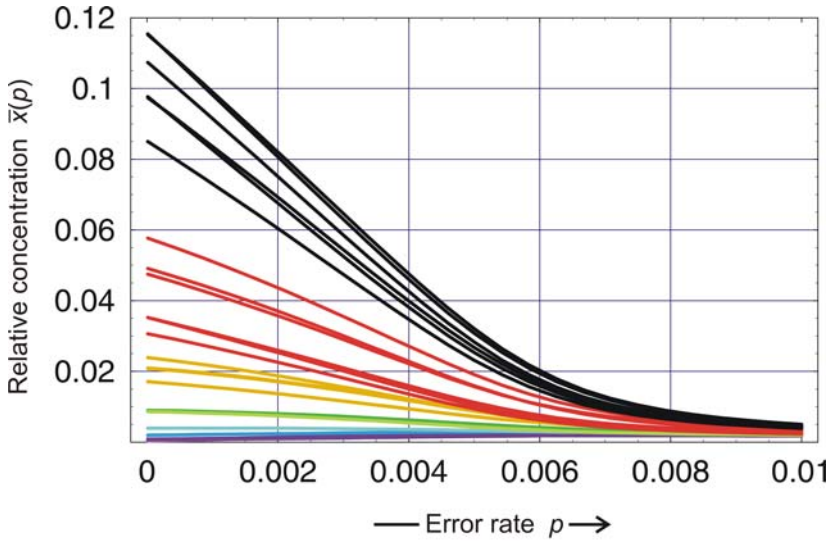
Neutral network

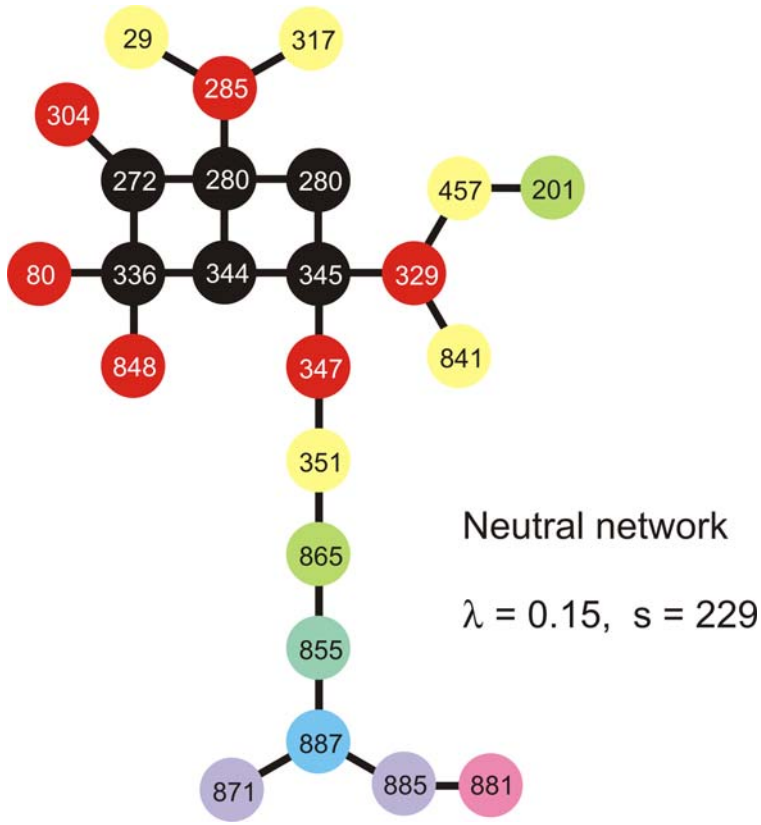
$\lambda = 0.10,\ s = 229$

$\lambda = 0.10$

N = 7

Neutral networks with increasing λ

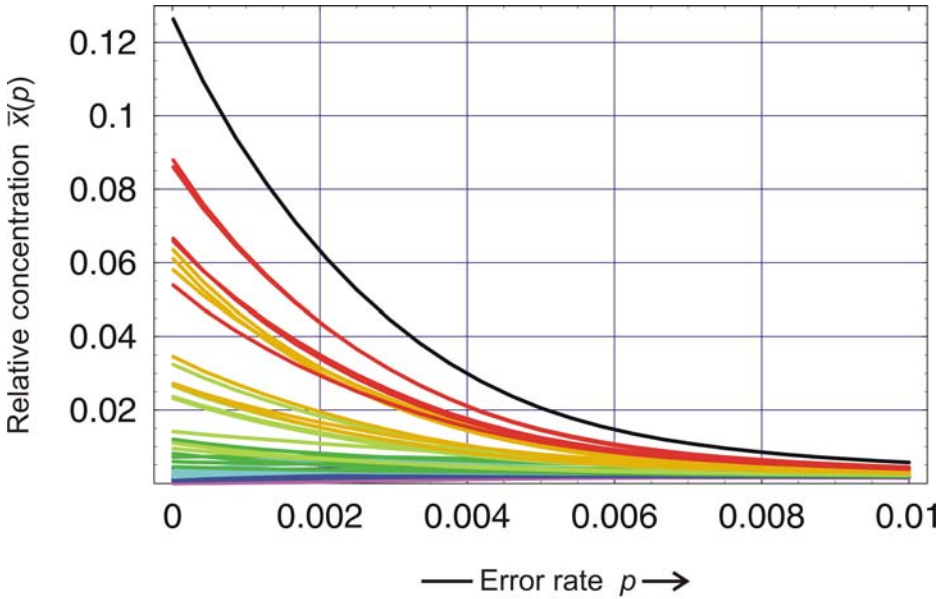


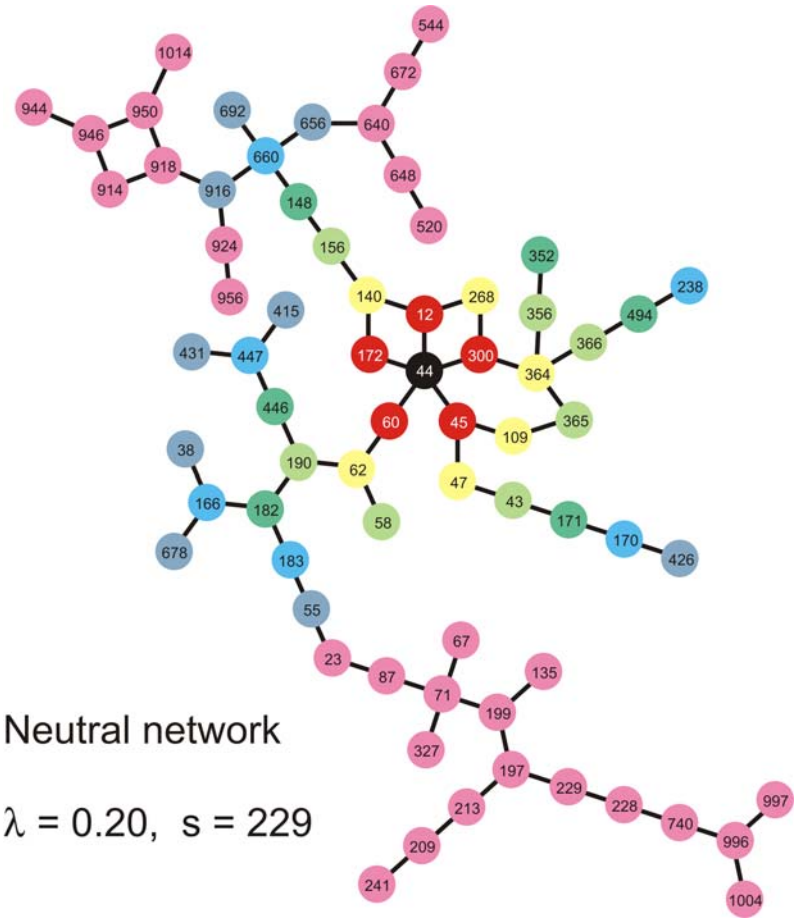


$\lambda = 0.15$

N = 24

Neutral networks with increasing λ





Neutral network

$\lambda = 0.20$, s = 229

$\lambda = 0.20$

N = 70

Neutral networks with increasing λ

random number seed σ

λ	229	367	491	673	877
0.005	1	1	1|1	1	1|1
0.01	**2**	**2**	**2**	1	1|1
0.015	**2**	**2**	**2**	**2**	1|1
0.02	**3**	**2**	**2**	2|2	1|1|1|1
0.025	**3**	**2**	**2**	**3**	1|1|1|1
0.03	**3**	**3**	**2**	**3**	**3**
0.035	**3**	**3**	**2**	**3**	**3**
0.04	**3**	**3|3**	**2**	**3**	**3**
0.045	**3**	**5**	**3**	**3**	**4**
0.05	**3**	**5**	**3**	**5**	**7**
0.06	**6**	**5**	**3**	**7**	**7**
0.07	**6**	**8**	**5**	**7**	**7**
0.08	**7**	**8**	**5**	**4**	**8**
0.09	**7**	**8**	**10**	**5**	**9**
0.10	**7**	**10**	**9**	**5**	**9**
0.11	**8**	**14**	**22**	**6**	**9**
0.12	**10**	**17**	**44**	**14**	**9**
0.13	**11**	**40**	**49**	**43**	**9**
0.14	**16**	**52**	**70**	**84**	**28**
0.15	**24**	**72**	**71**	**95**	**12**
0.20	**70** (69)	180	152	181	151

Size of selected neutral networks in the limit $p \rightarrow 0$ as a function of the degree of neutrality λ

1. Ruggedness of molecular landscapes
2. Replication-mutation dynamics
3. Models of fitness landscapes
4. Ruggedness and error thresholds
5. **Stochasticity of replication and mutation**
6. Population dynamics on neutral networks

Evolution *in silico*

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

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

32. A nonsense mutation may affect mRNA stability and result in degradation of the transcript [L. Maquat, *Am. J. Hum. Genet.* **59**, 279 (1996)].
33. Data not shown; a dot blot with poly (A)$^+$ RNA from 50 human tissues (The Human RNA Master Blot, 7770-1, Clontech Laboratories) was hybridized with a probe from exons 29 to 47 of *MYO15* using the same condition as Northern blot analysis (*13*).
34. Smith–Magenis syndrome (SMS) is due to deletions of 17p11.2 of various sizes, the smallest of which includes *MYO15* and perhaps 20 other genes [(*6*); K-S Chen, L. Potocki, J. R. Lupski, *MRDD Res. Rev.* **2**, 122 (1996)]. *MYO15* expression is easily detected in the pituitary gland (data not shown). Haploinsufficiency for *MYO15* may explain a portion of the SMS phenotype such as short stature. Moreover, a few SMS patients have sensorineural hearing loss, possibly because of a point mutation in *MYO15* in trans to the SMS 17p11.2 deletion.
35. R. A. Fridell, data not shown.
36. K. B. Avraham *et al.*, *Nature Genet.* **11**, 369 (1995); X-Z. Liu *et al.*, *ibid.* **17**, 268 (1997); F. Gibson *et al.*, *Nature* **374**, 62 (1995); D. Weil *et al.*, *ibid.*, p. 60.
37. RNA was extracted from cochlea (membranous labyrinths) obtained from human fetuses at 18 to 22 weeks of development in accordance with guidelines established by the Human Research Committee at the Brigham and Women's Hospital. Only samples without evidence of degradation were pooled for poly (A)$^+$ selection over oligo(dT) columns. First-strand cDNA was prepared using an Advantage RT-for-PCR kit (Clontech Laboratories). A portion of the first-strand cDNA (4%) was amplified by PCR with Advantage cDNA polymerase mix (Clontech Laboratories) using human *MYO15*-specific oligonucleotide primers (forward, 5′-GCATGACCTGCCGGCTAAT-GGG-3′; reverse, 5′-CTCACGGCTTCTGCATGGT-GCTCGGCTGGC-3′). Cycling conditions were 40 s at 94°C; 40 s at 66°C (3 cycles), 60°C (5 cycles), and 55°C (29 cycles); and 45 s at 68°C. PCR products were visualized by ethidium bromide staining after fractionation in a 1% agarose gel. A 688-bp PCR product is expected from amplification of the human *MYO15* cDNA. Amplification of human genomic DNA with this primer pair would result in a 2903-bp fragment.

38. We are grateful to the people of Bengkala, Bali, and the two families from India. We thank J. R. Lupski and K.-S. Chen for providing the human chromosome 17 cosmid library. For technical and computational assistance, we thank N. Dietrich, M. Fergusson, A. Gupta, E. Sorbello, R. Torkzadeh, C. Varner, M. Walker, G. Bouffard, and S. Beckstrom-Sternberg (National Institutes of Health Intramural Sequencing Center). We thank J. T. Hinnant, I. N. Arhya, and S. Winata for assistance in Bali, and T. Barber, S. Sullivan, E. Green, D. Drayna, and J. Battey for helpful comments on this manuscript. Supported by the National Institute on Deafness and Other Communication Disorders (NIDCD) (Z01 DC 00035-01 and Z01 DC 00038-01 to T.B.F. and E.R.W. and R01 DC 03402 to C.C.M.), the National Institute of Child Health and Human Development (R01 HD30428 to S.A.C.) and a National Science Foundation Graduate Research Fellowship to F.J.P. This paper is dedicated to J. B. Snow Jr. on his retirement as the Director of the NIDCD.

9 March 1998; accepted 17 April 1998

Continuity in Evolution: On the Nature of Transitions

Walter Fontana and Peter Schuster

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

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

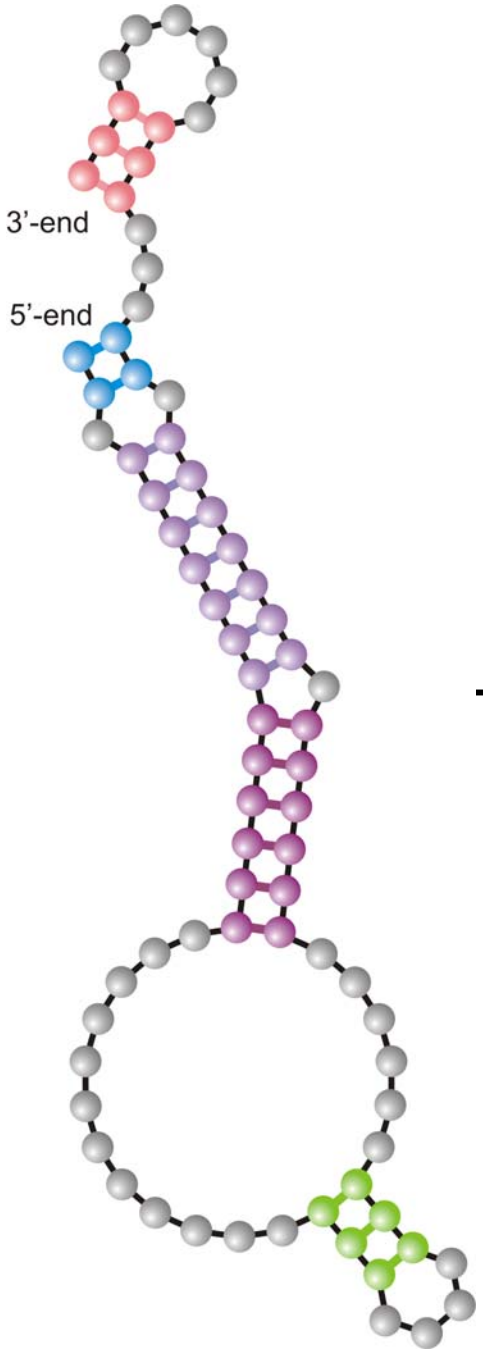
We focus on the narrow domain of RNA secondary structure, which is currently the simplest computationally tractable, yet realistic phenotype (*2*). This choice enables the definition and exploration of concepts that may prove useful in a wider context. RNA secondary structures represent a coarse level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

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

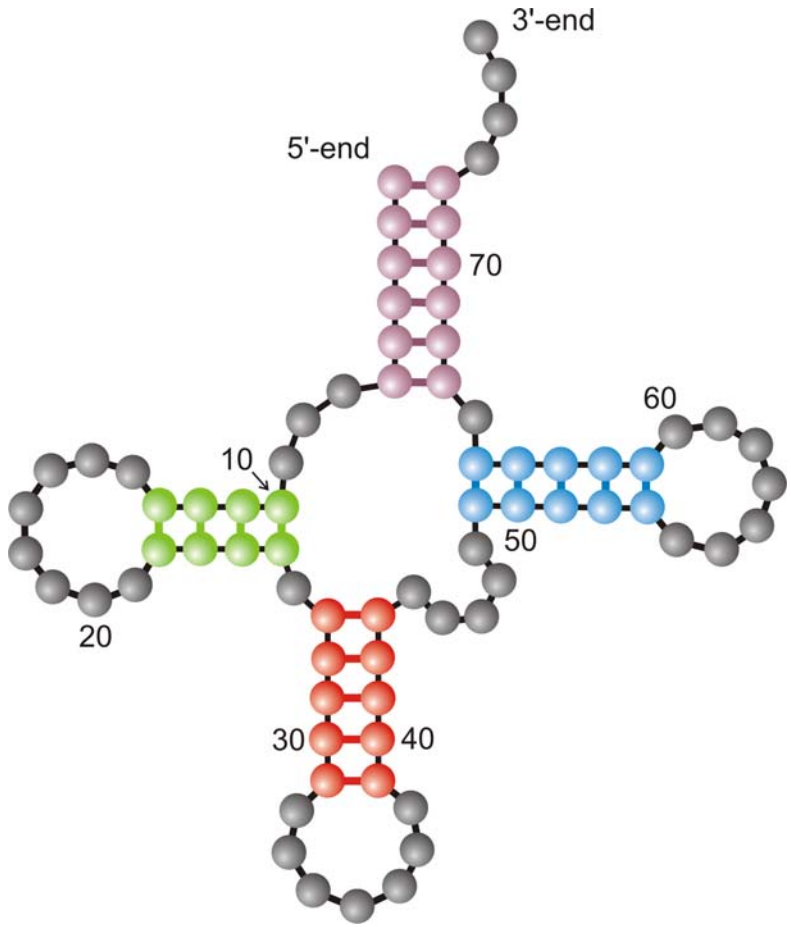
An instance representing in its qualitative features all the simulations we performed is shown in Fig. 1A. Starting with identical sequences folding into a random shape, the simulation was stopped when the population became dominated by the target, here a canonical tRNA shape. The black curve traces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

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

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

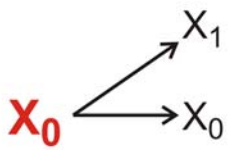


Structure of randomly chosen initial sequence

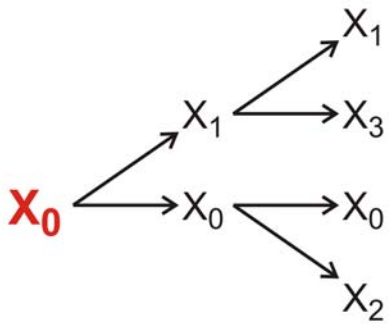


Phenylalanyl-tRNA as target structure

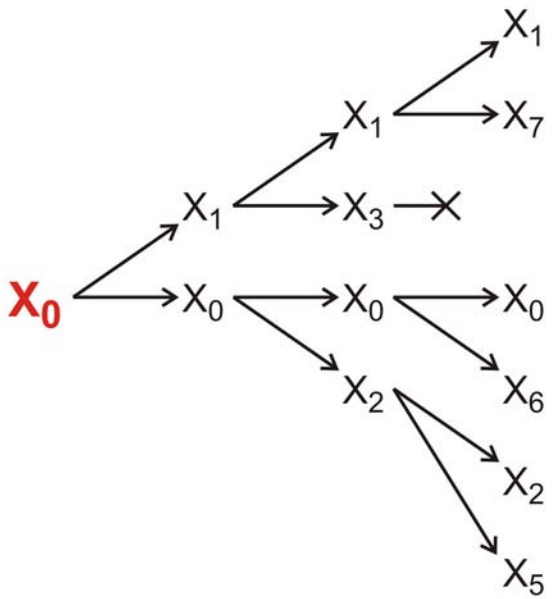
Evolution of RNA molecules as a Markow process and its analysis by means of the relay series



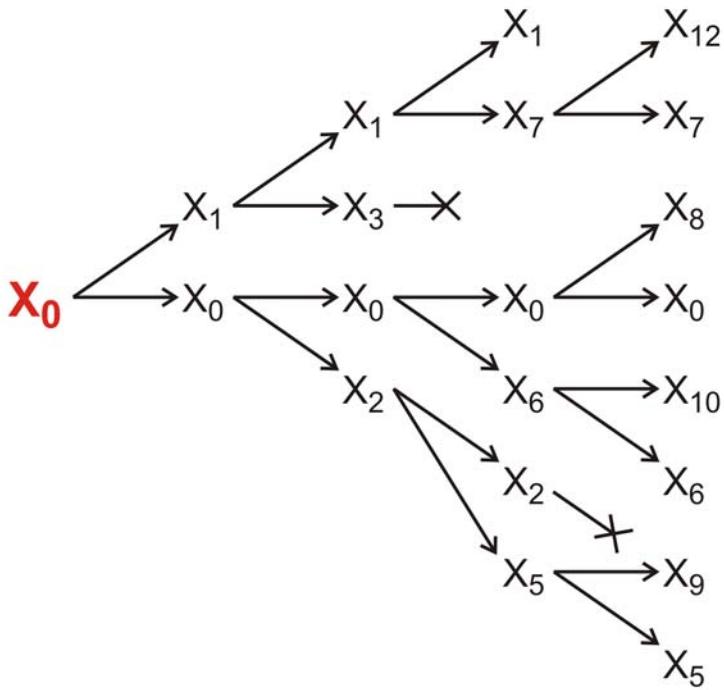
Evolution of RNA molecules as a Markow process and its analysis by means of the relay series



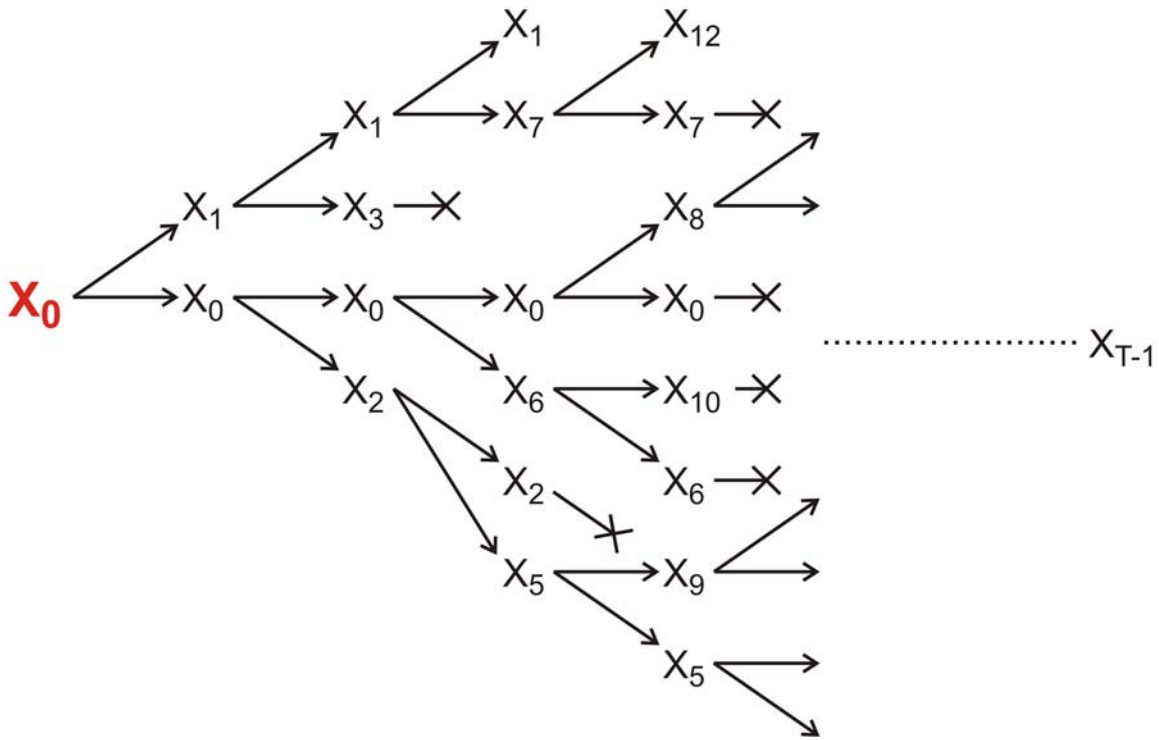
Evolution of RNA molecules as a Markow process and its analysis by means of the relay series



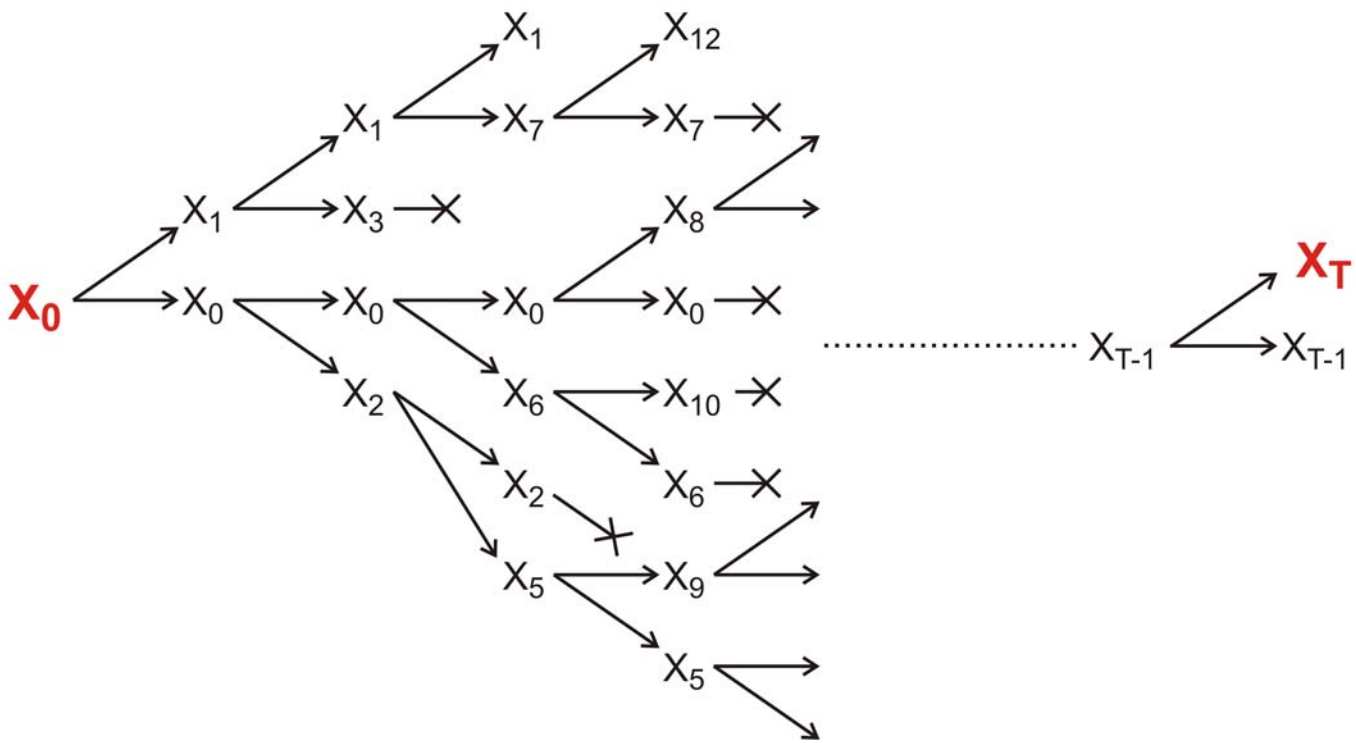
Evolution of RNA molecules as a Markow process and its analysis by means of the relay series



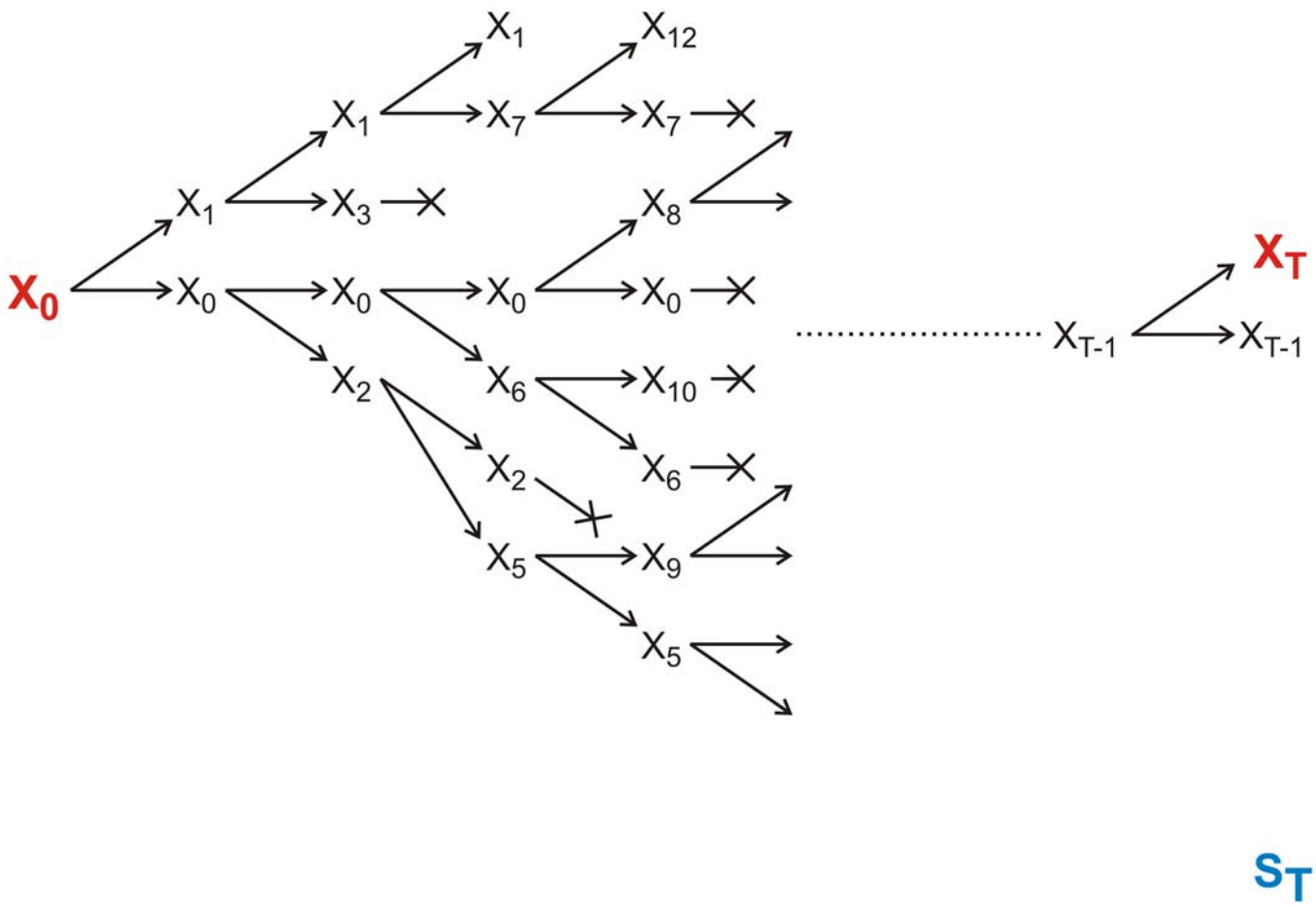
Evolution of RNA molecules as a Markow process and its analysis by means of the relay series



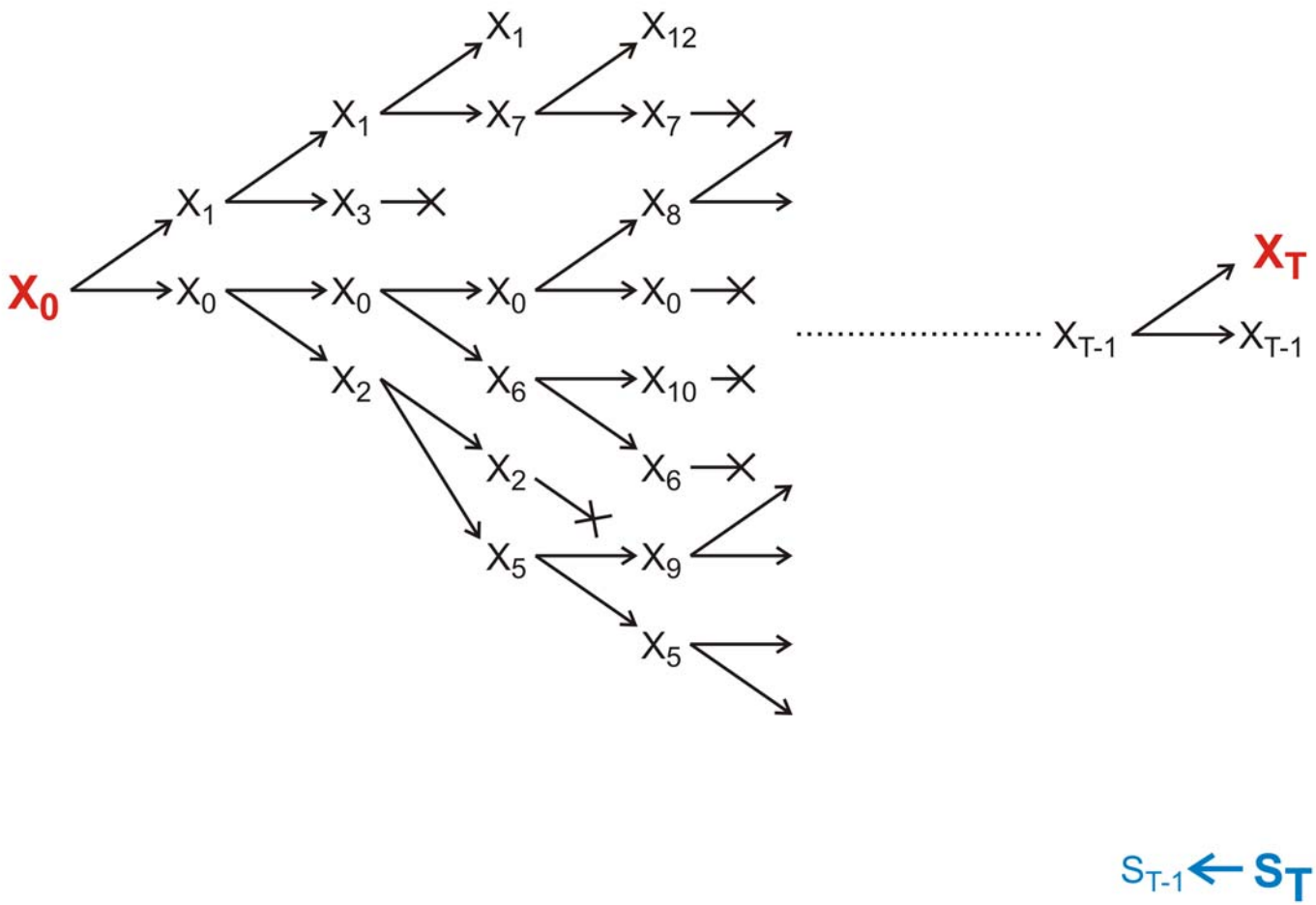
Evolution of RNA molecules as a Markow process and its analysis by means of the relay series



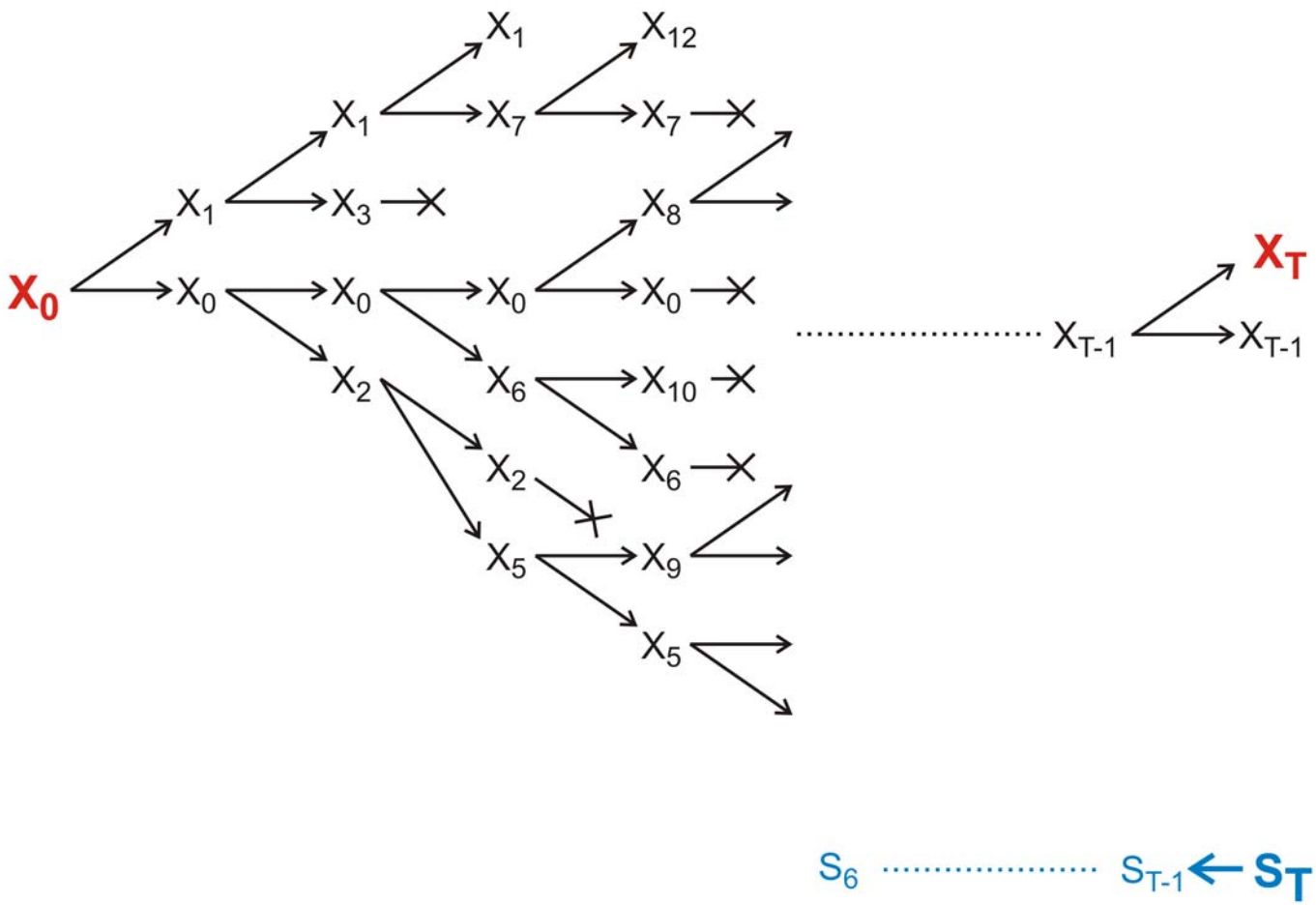
Evolution of RNA molecules as a Markow process and its analysis by means of the relay series



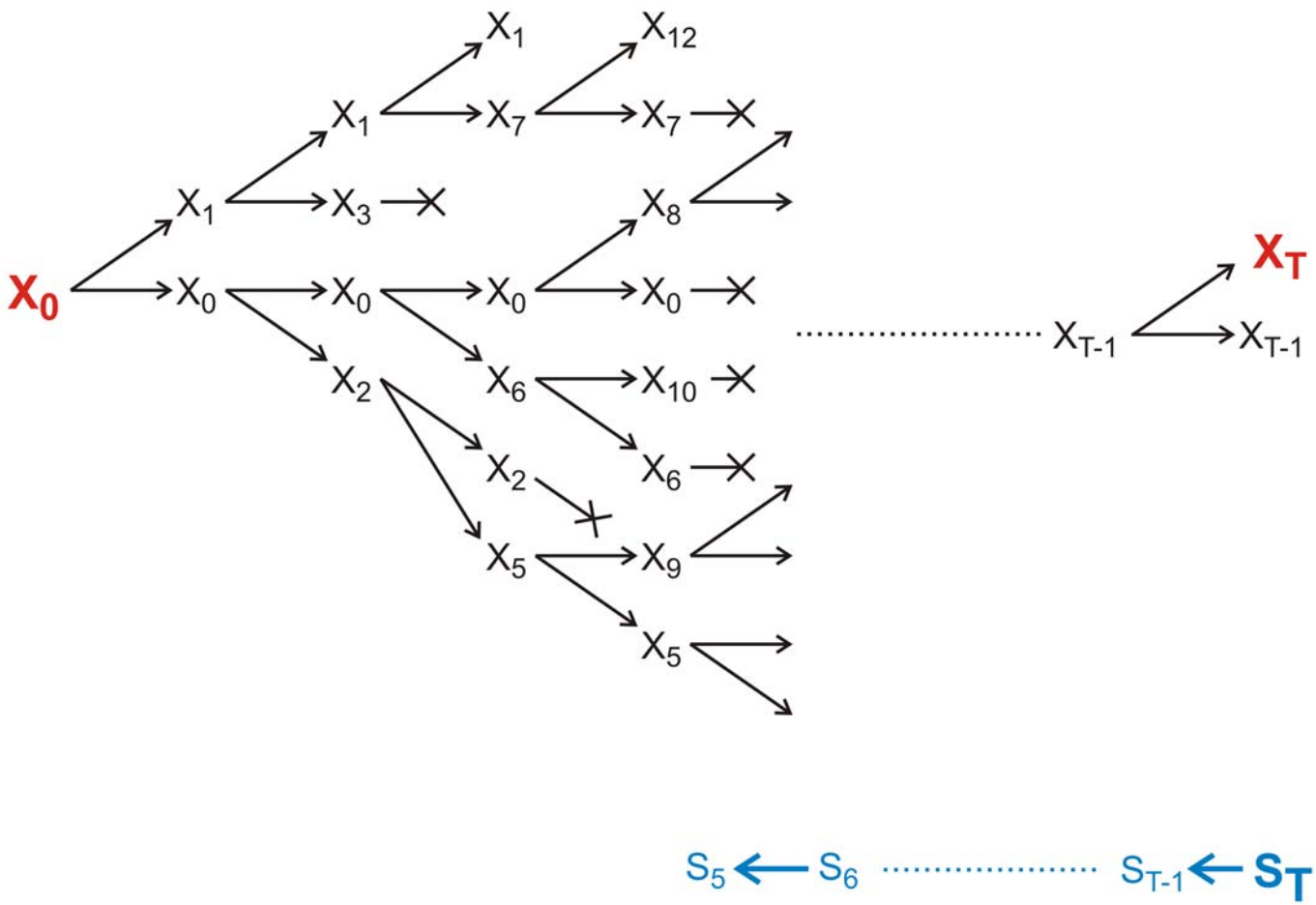
Evolution of RNA molecules as a Markow process and its analysis by means of the relay series



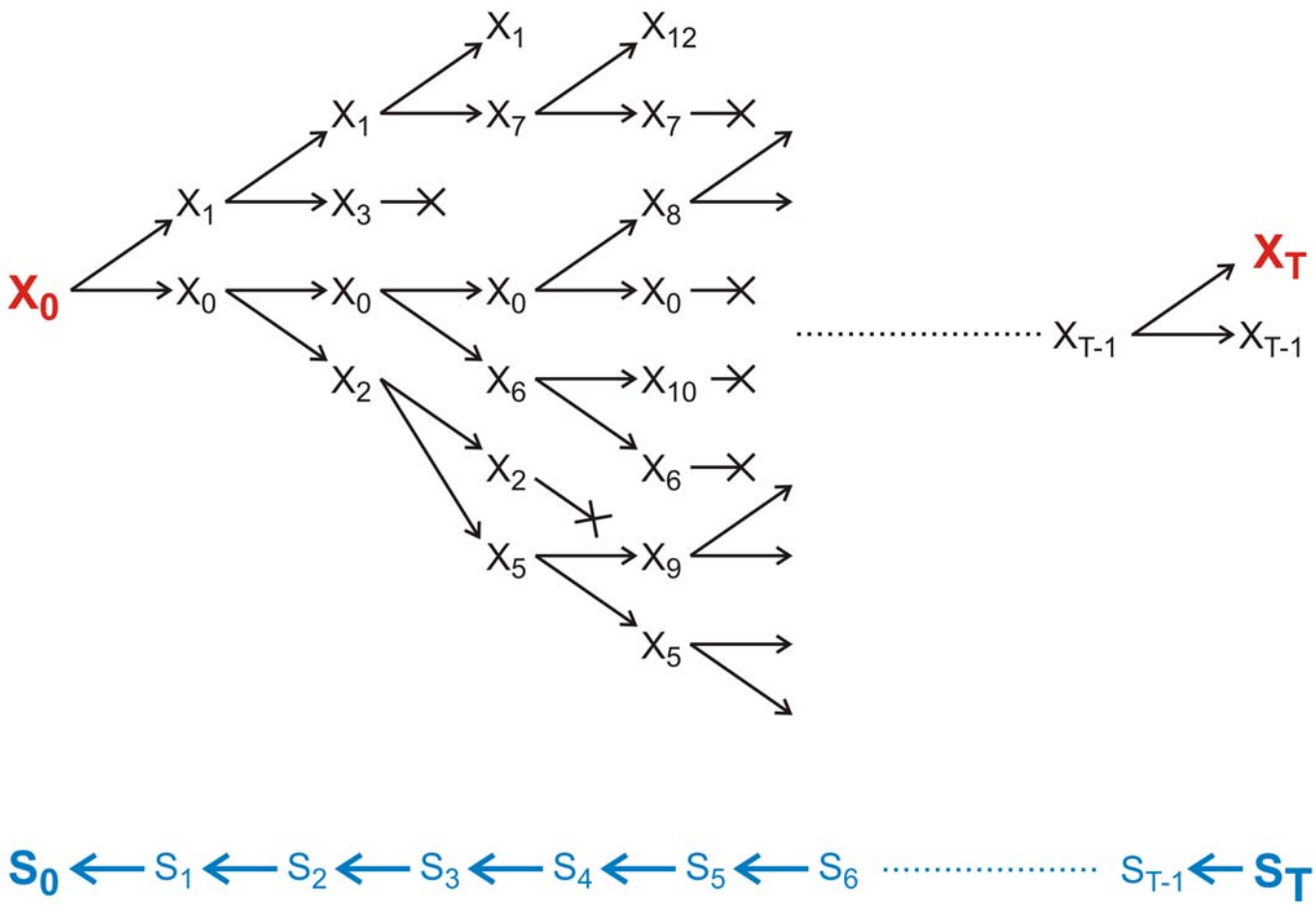
Evolution of RNA molecules as a Markow process and its analysis by means of the relay series



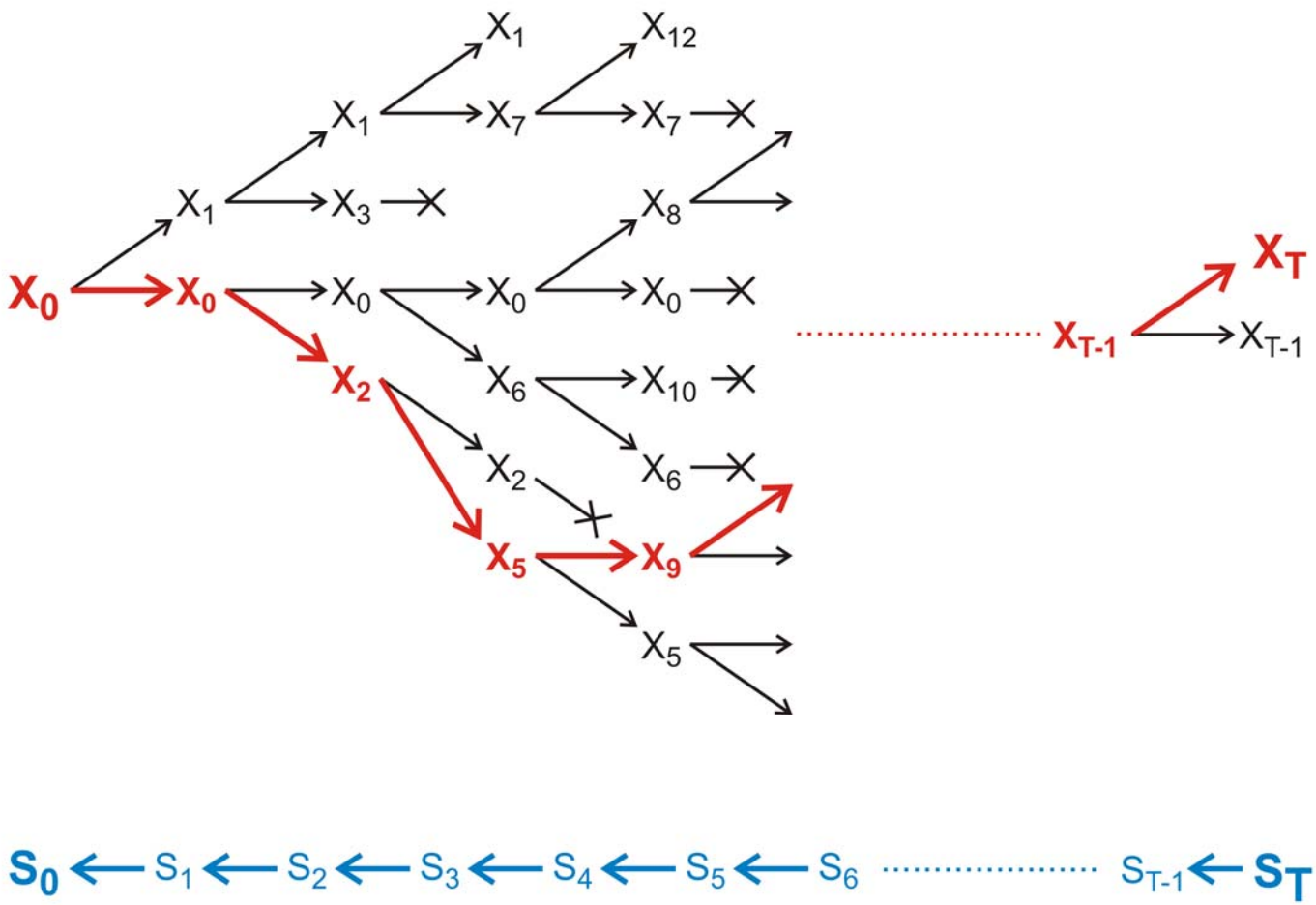
Evolution of RNA molecules as a Markow process and its analysis by means of the relay series



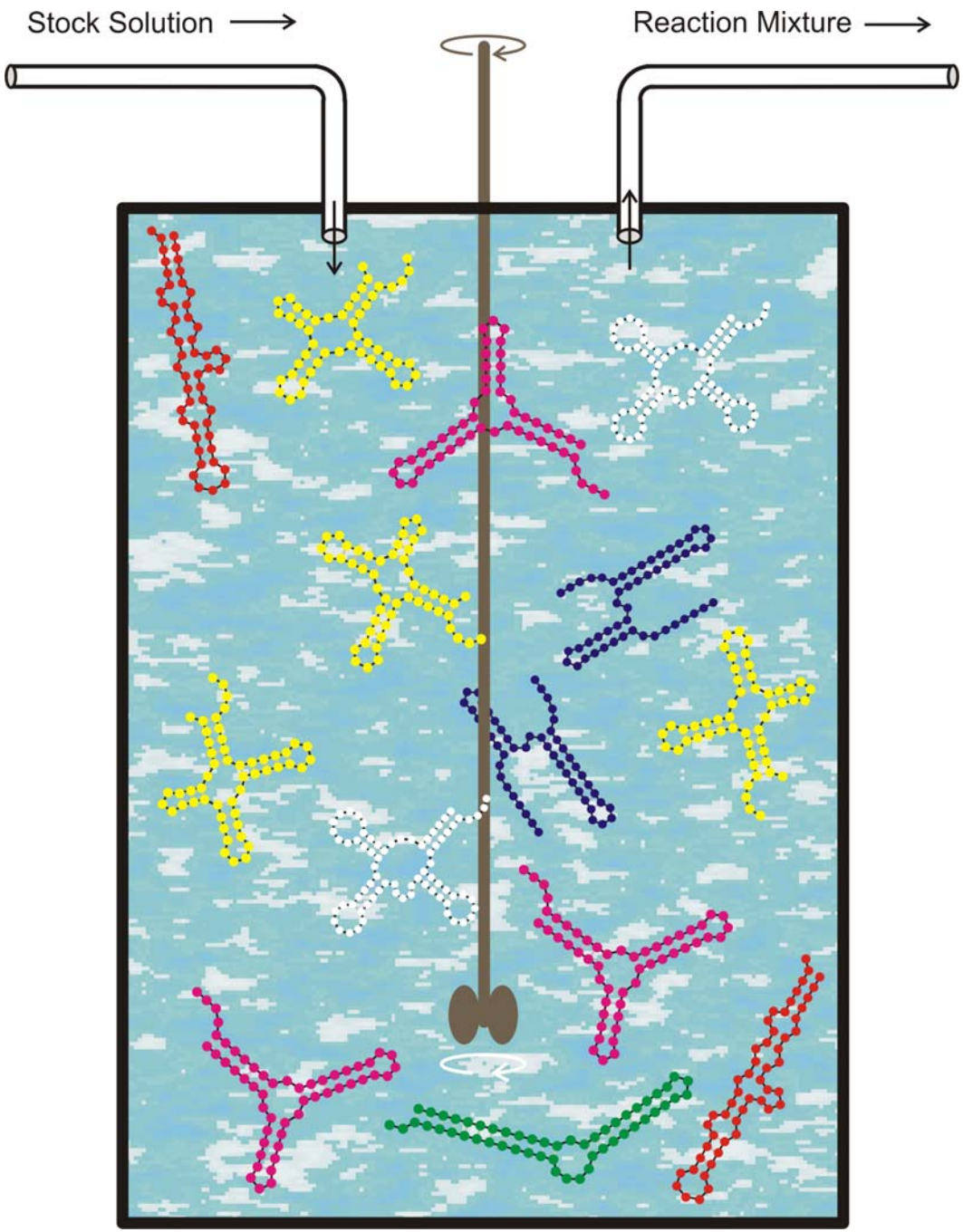
Evolution of RNA molecules as a Markow process and its analysis by means of the relay series



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



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



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,

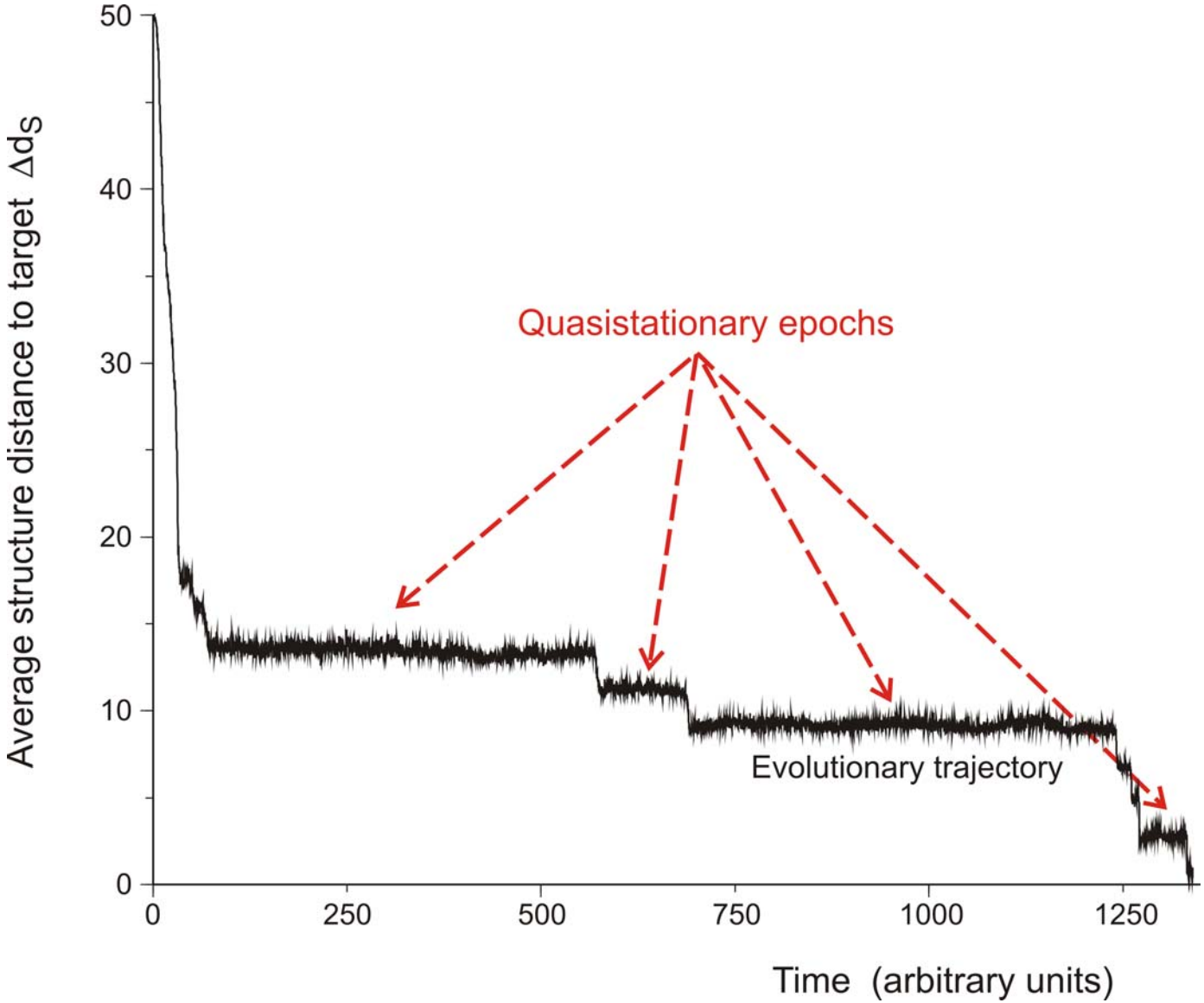
is determined by the flux:

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

Mutation rate:

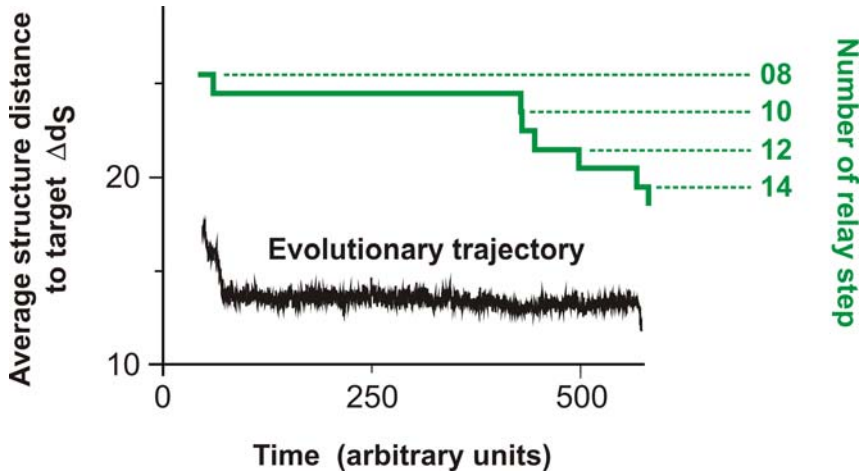
p = 0.001 / Nucleotide × Replication

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



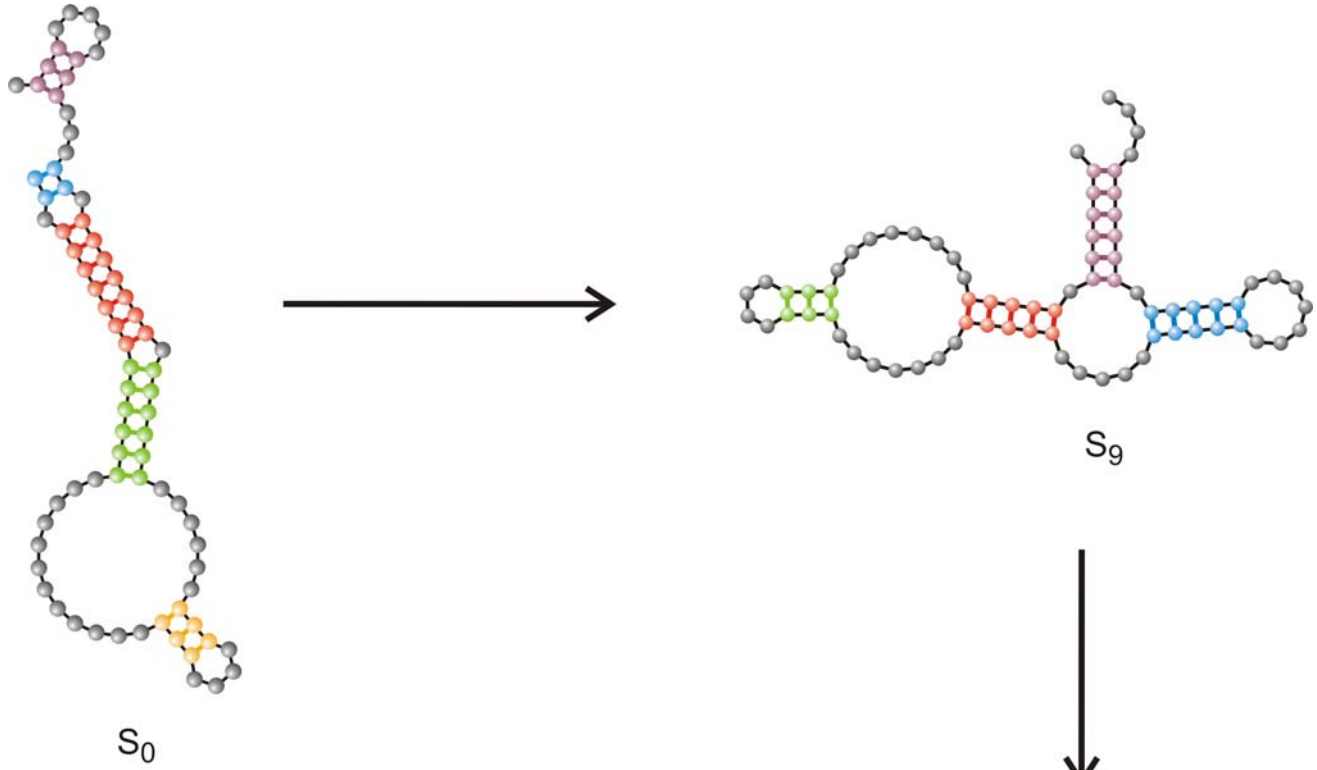

```
entry   GGUAUGGGCGUUGAAUAGUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUCCCGUACAGAA
8       .((((((((((((.........(((....)))).......)))))).....((((.......))))))))))....
exit    GGUAUGGGCGUUGAAUAAUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUCCCAUACAGAA
entry   GGUAUGGGCGUUGAAUAAUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUACCAUACAGAA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

Transition inducing point mutations change the molecular structure

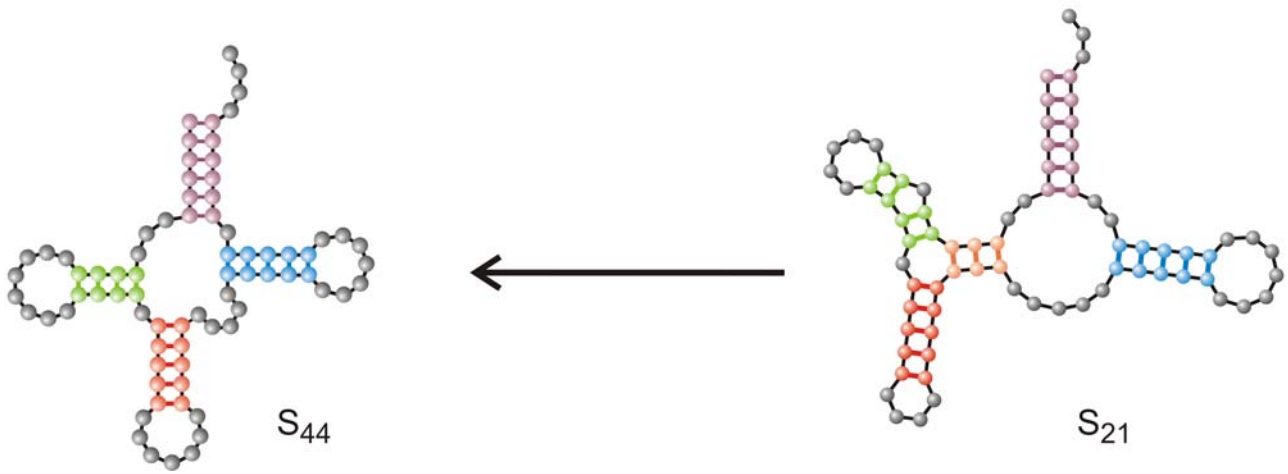
Neutral point mutations leave the molecular structure unchanged

Neutral genotype evolution during phenotypic stasis

Randomly chosen
initial structure



Phenylalanyl-tRNA
as target structure

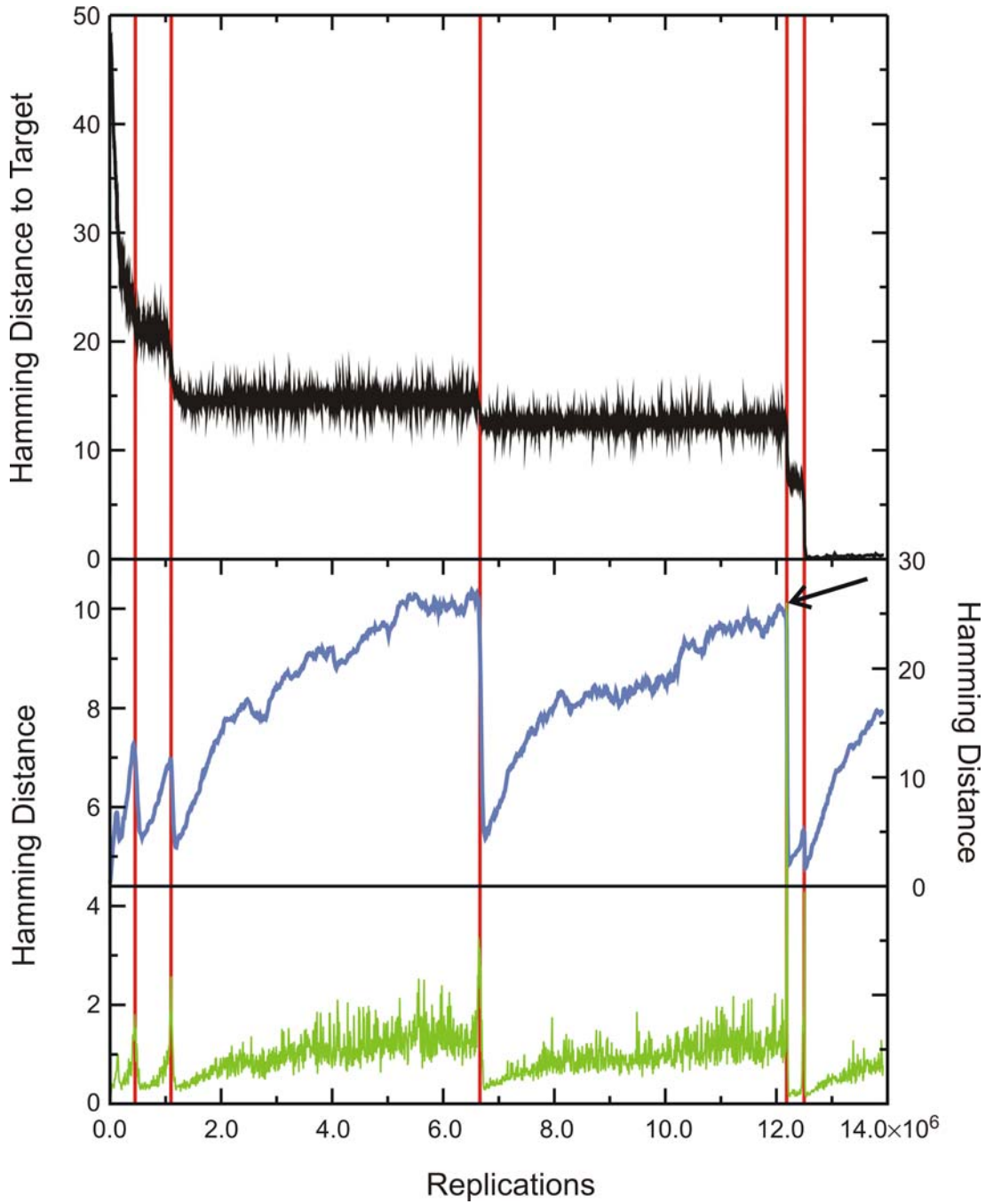


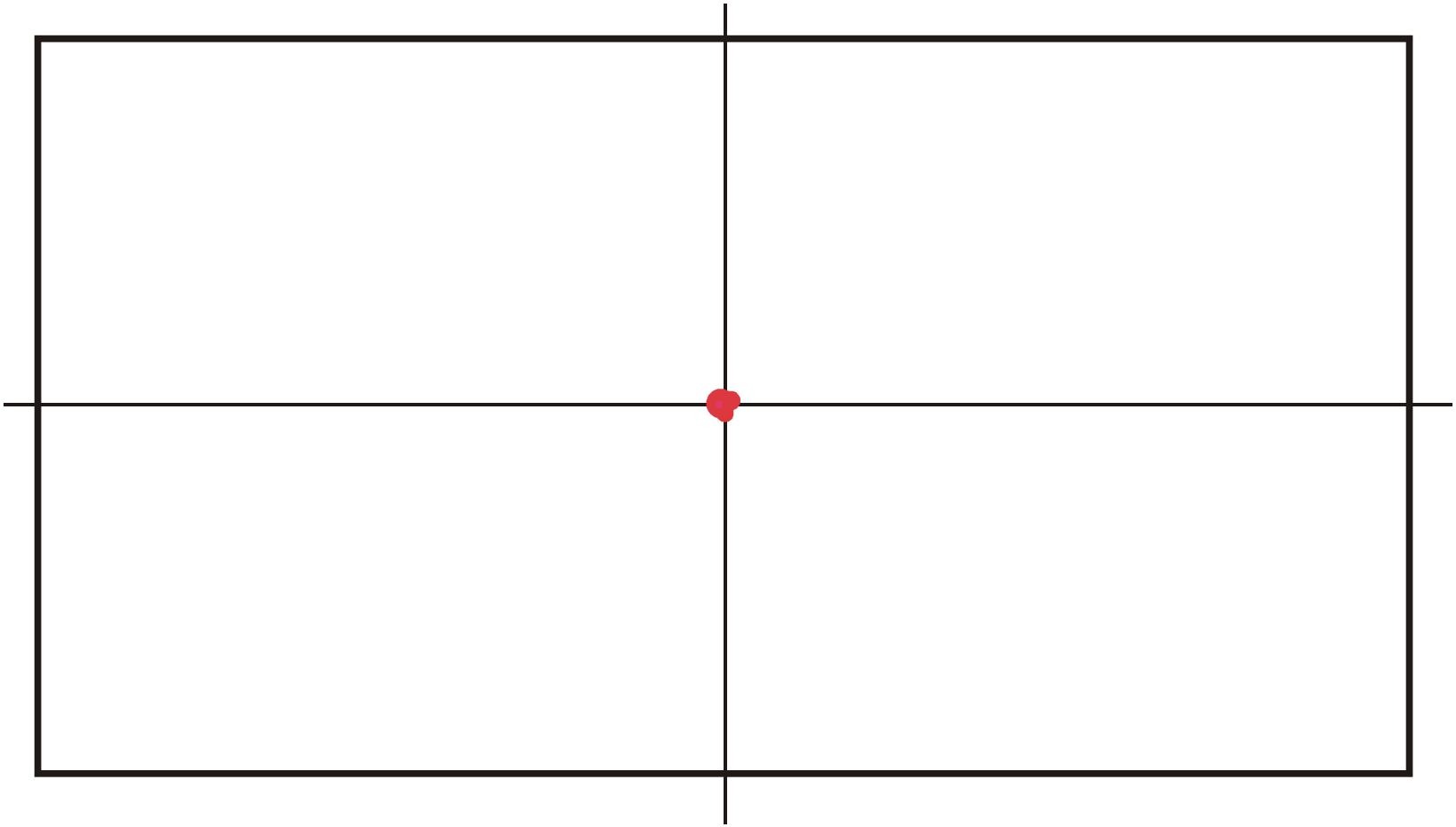
1. Ruggedness of molecular landscapes
2. Replication-mutation dynamics
3. Models of fitness landscapes
4. Ruggedness and error thresholds
5. Stochasticity of replication and mutation
6. **Population dynamics on neutral networks**

Evolutionary trajectory

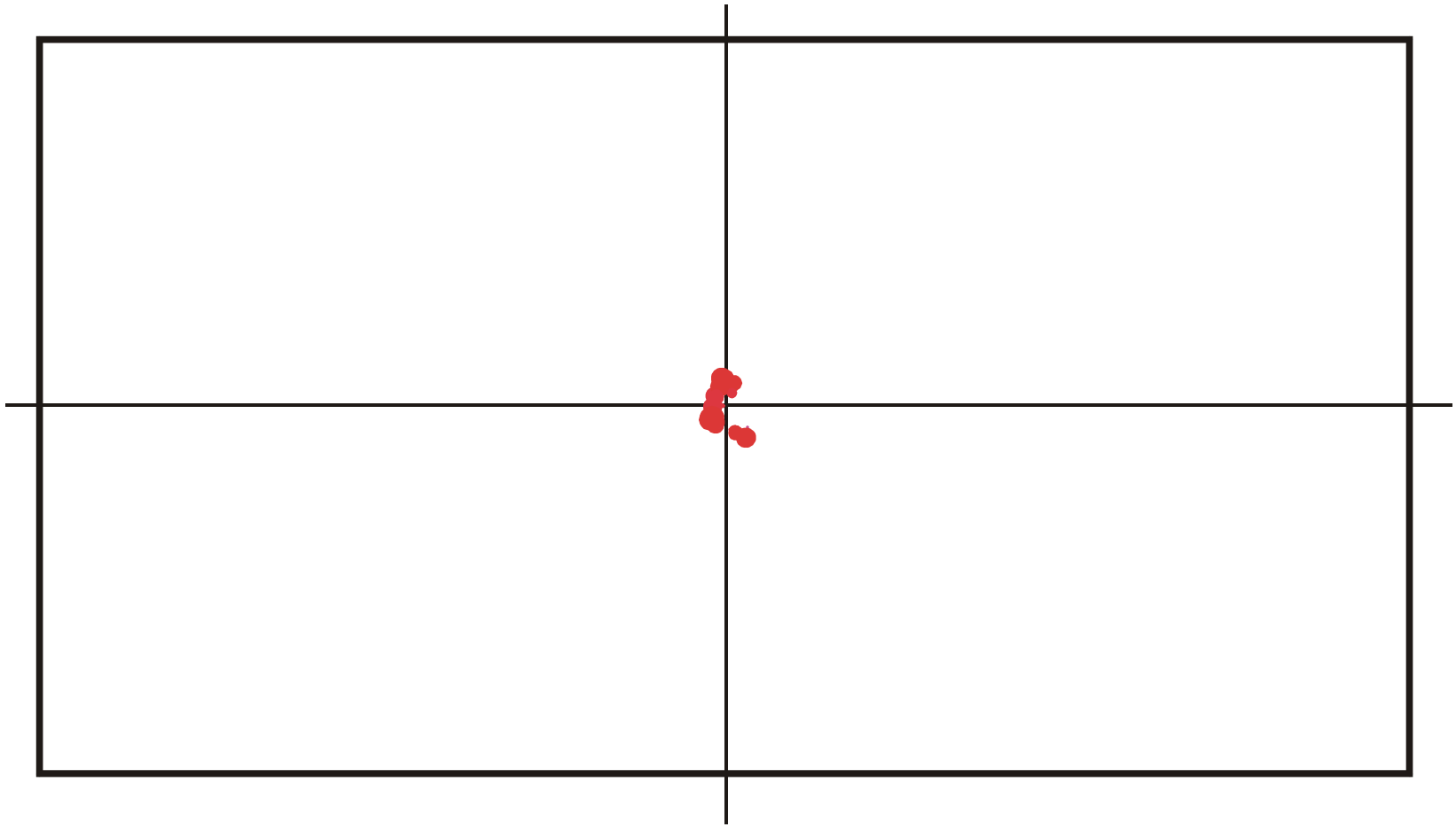
Spreading of the population on neutral networks

Drift of the population center in sequence space

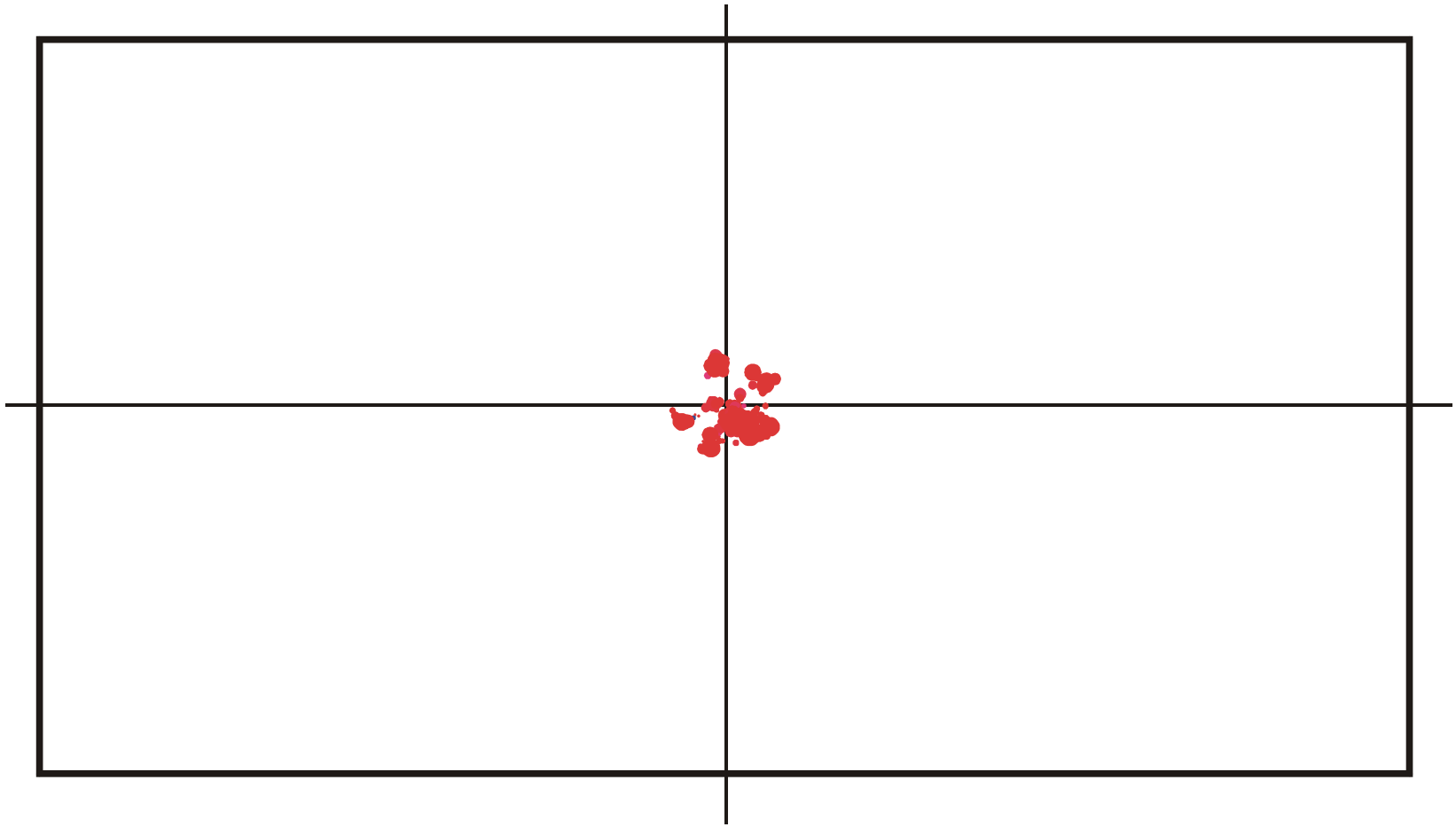




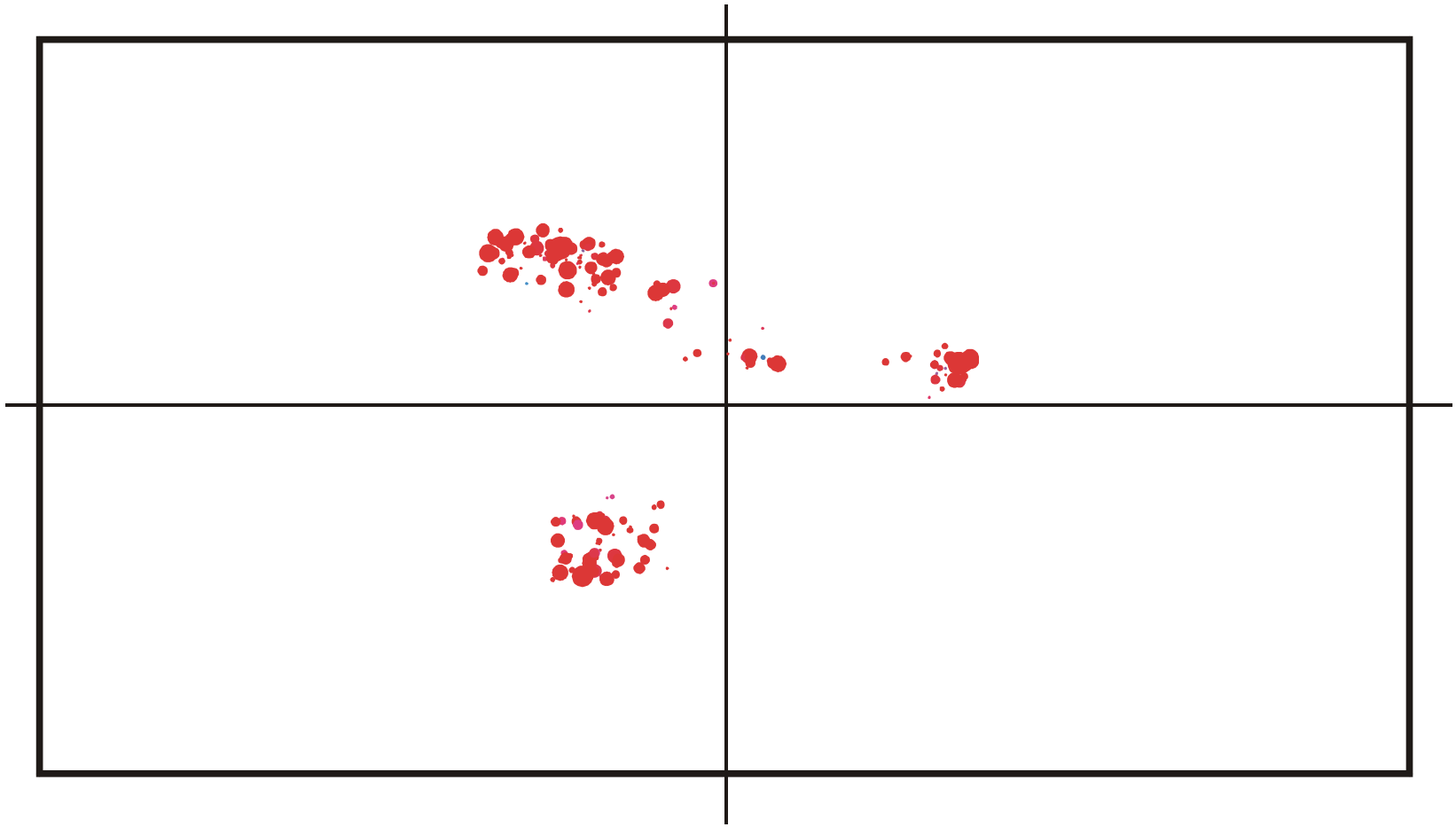
Spreading and evolution of a population on a neutral network: t = 150



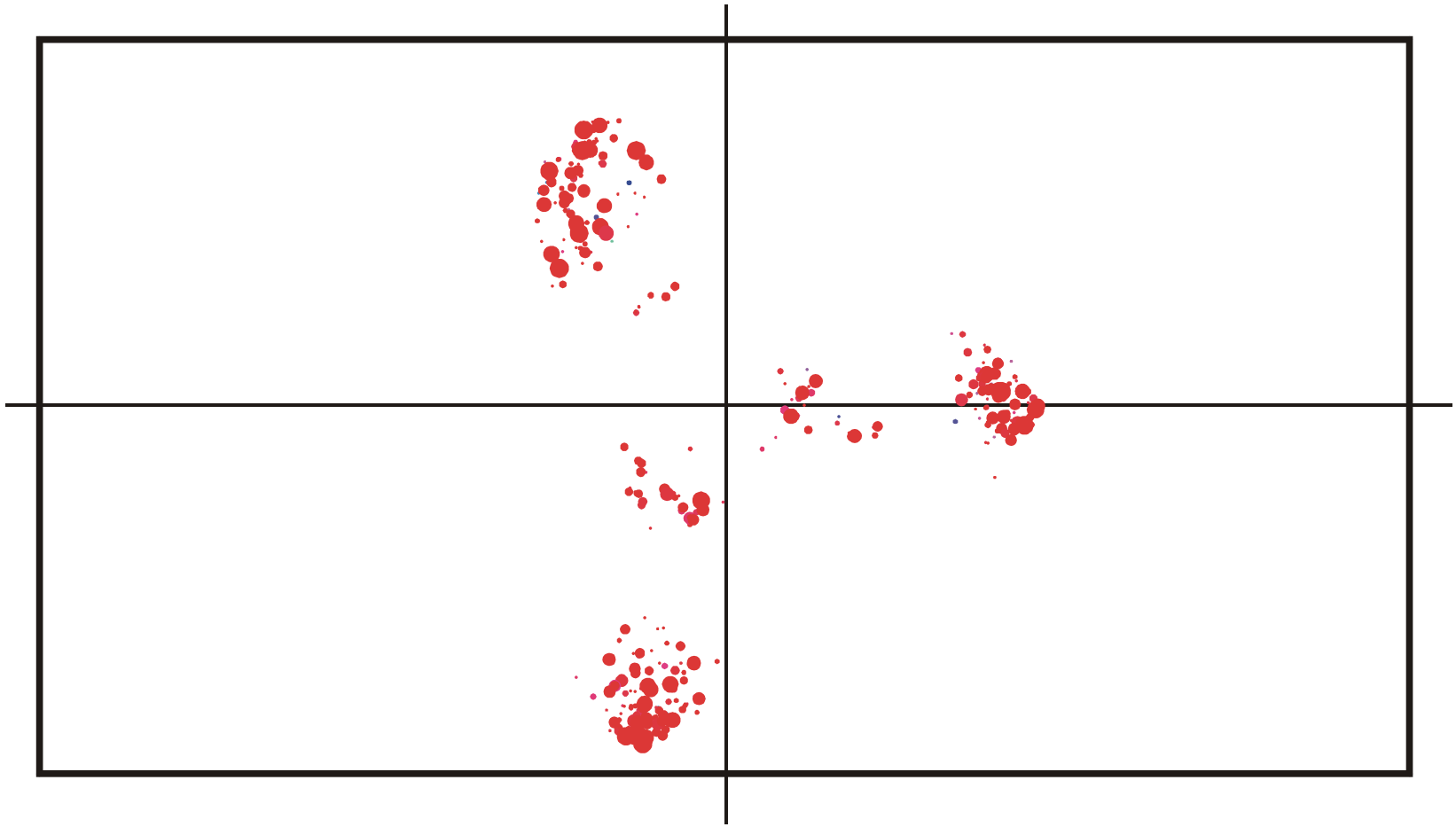
Spreading and evolution of a population on a neutral network : t = 170



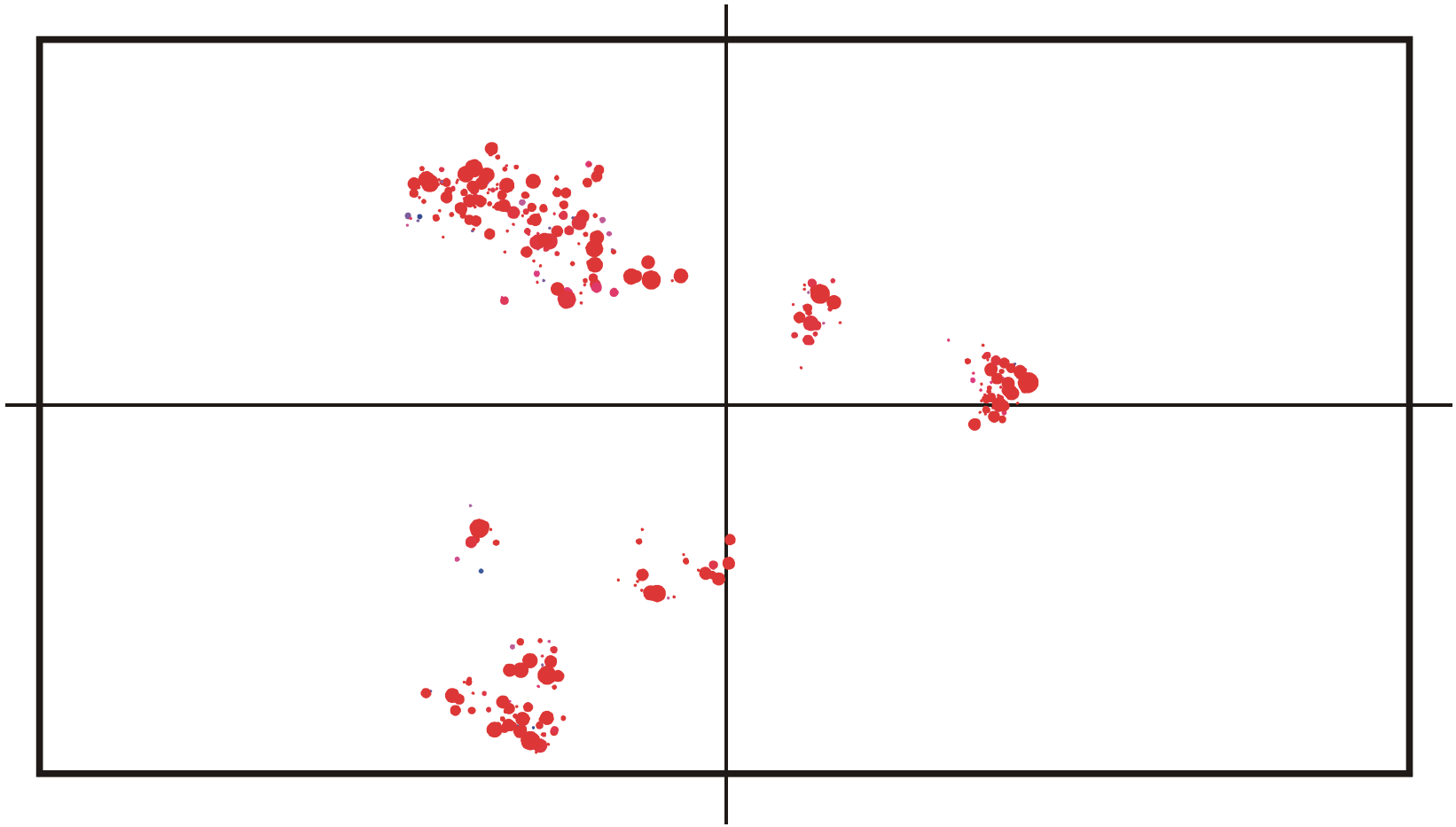
Spreading and evolution of a population on a neutral network : t = 200



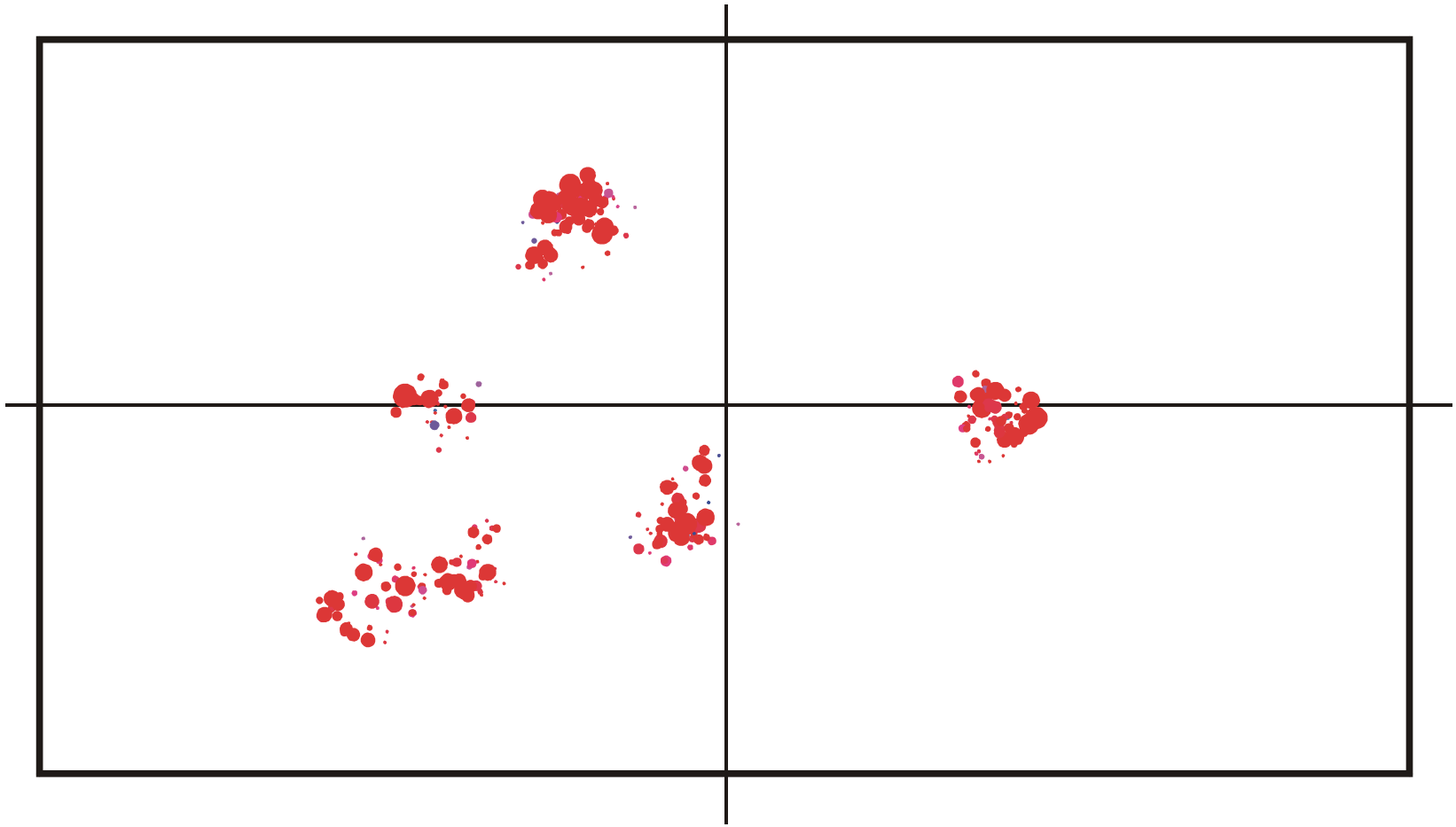
Spreading and evolution of a population on a neutral network : t = 350



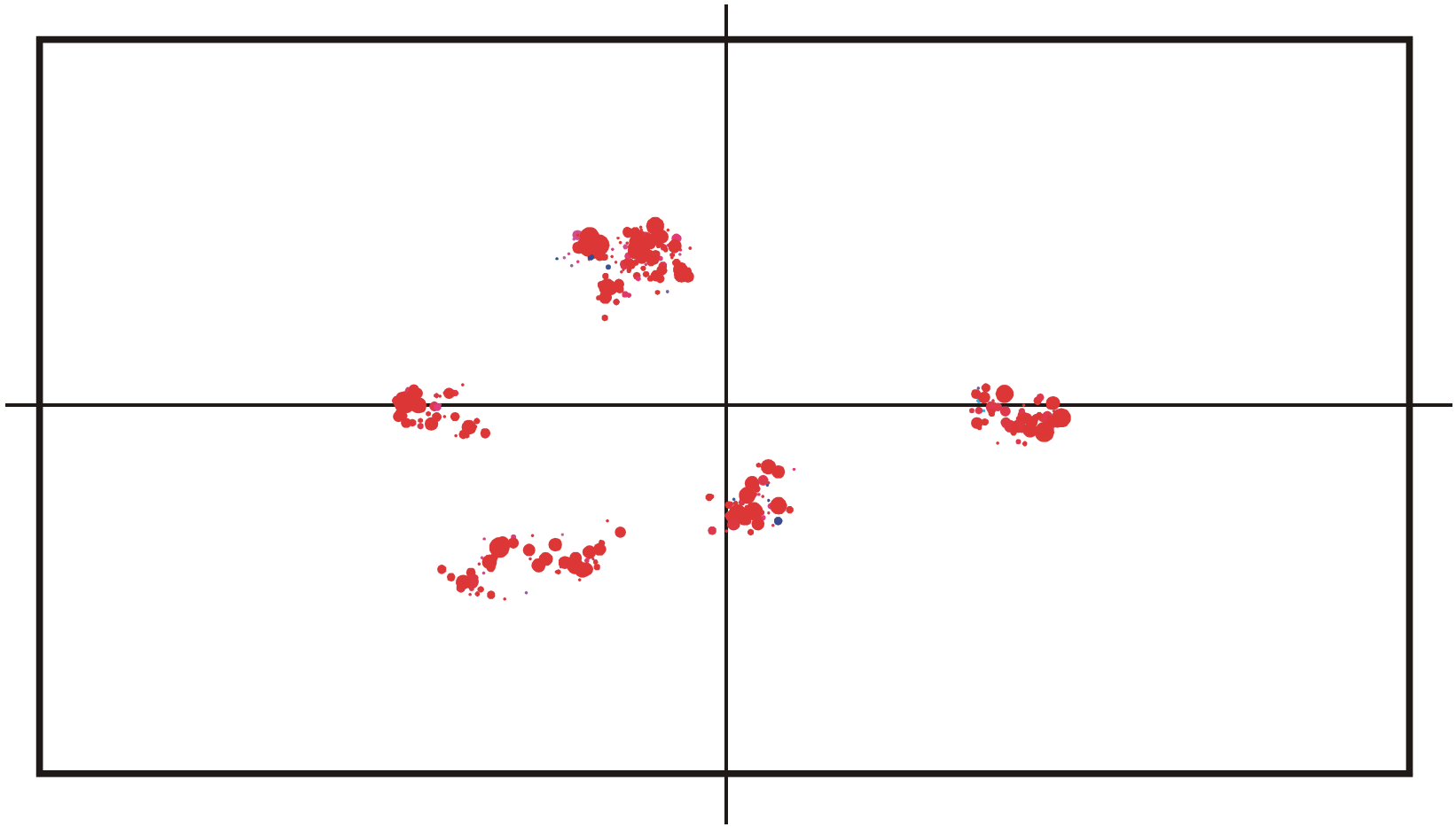
Spreading and evolution of a population on a neutral network : t = 500



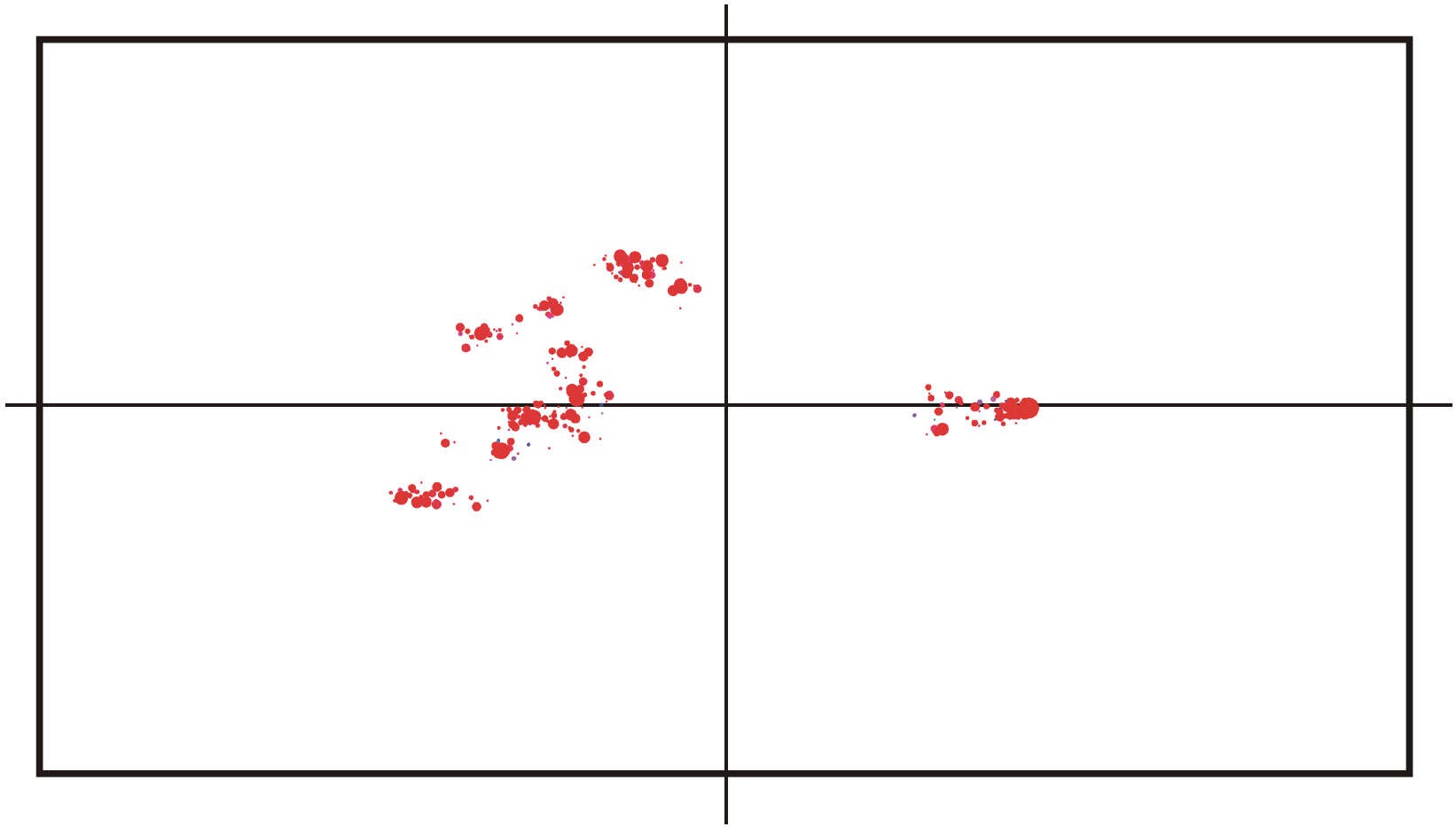
Spreading and evolution of a population on a neutral network : t = 650



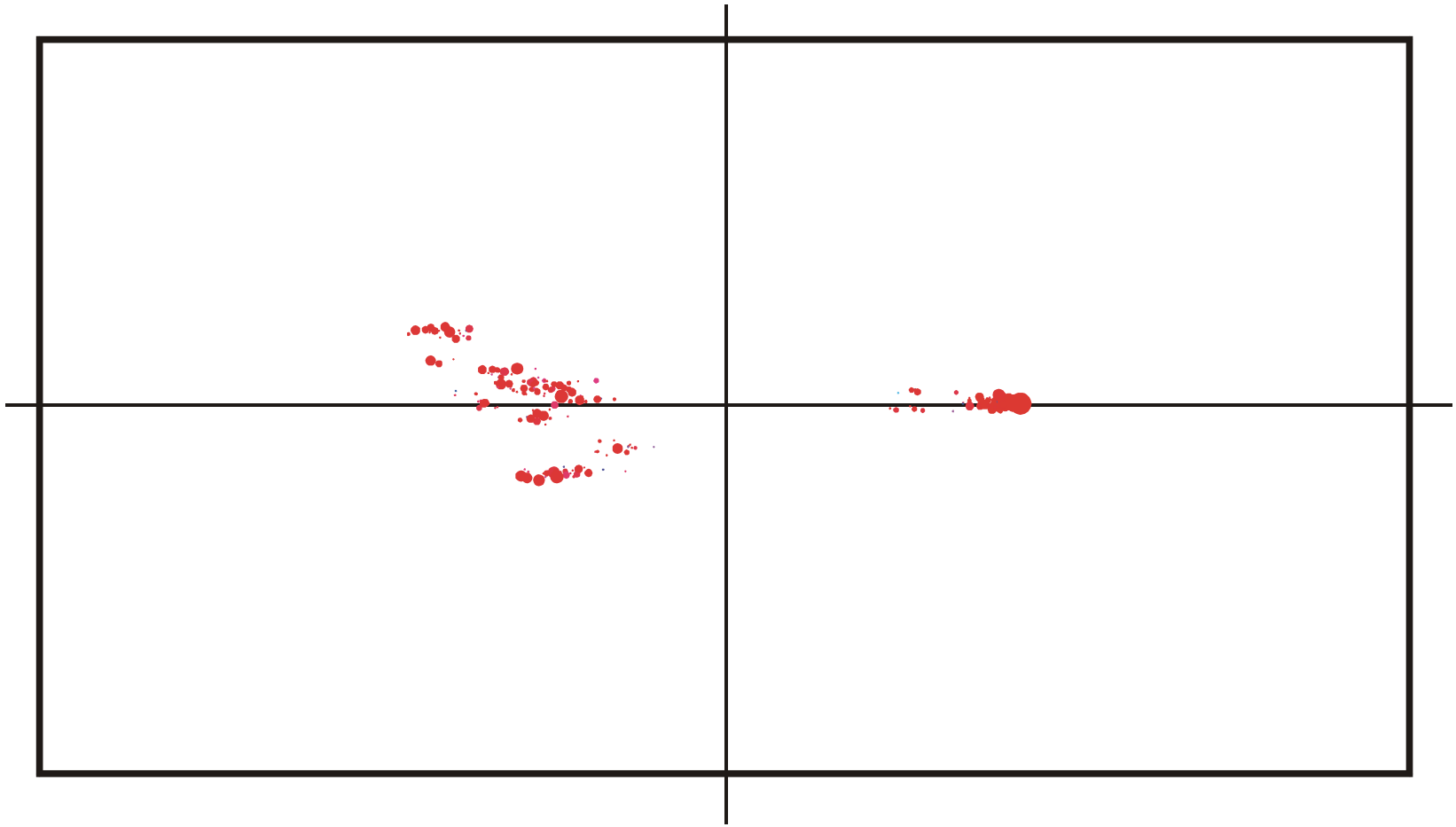
Spreading and evolution of a population on a neutral network : t = 820



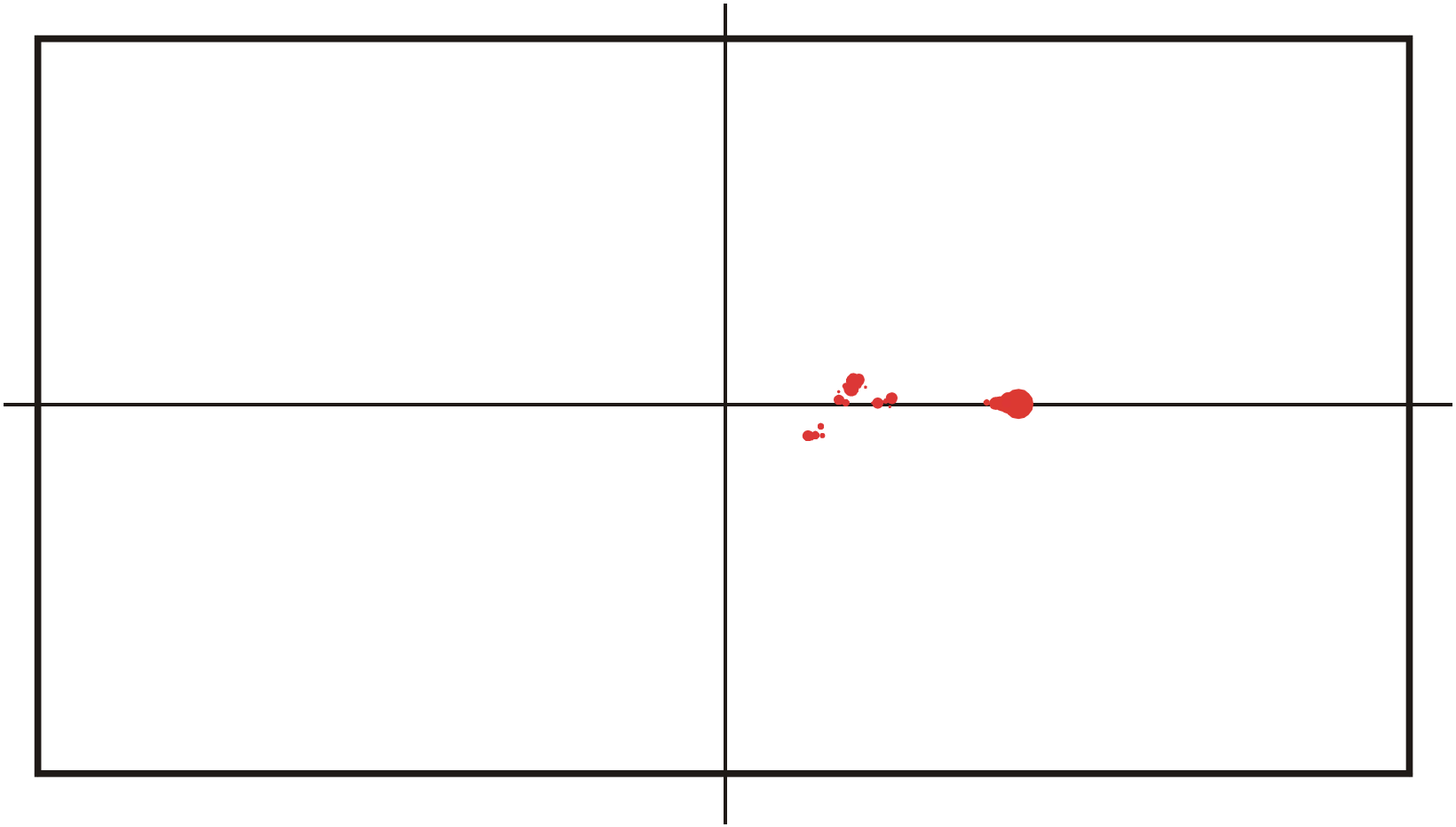
Spreading and evolution of a population on a neutral network : t = 825



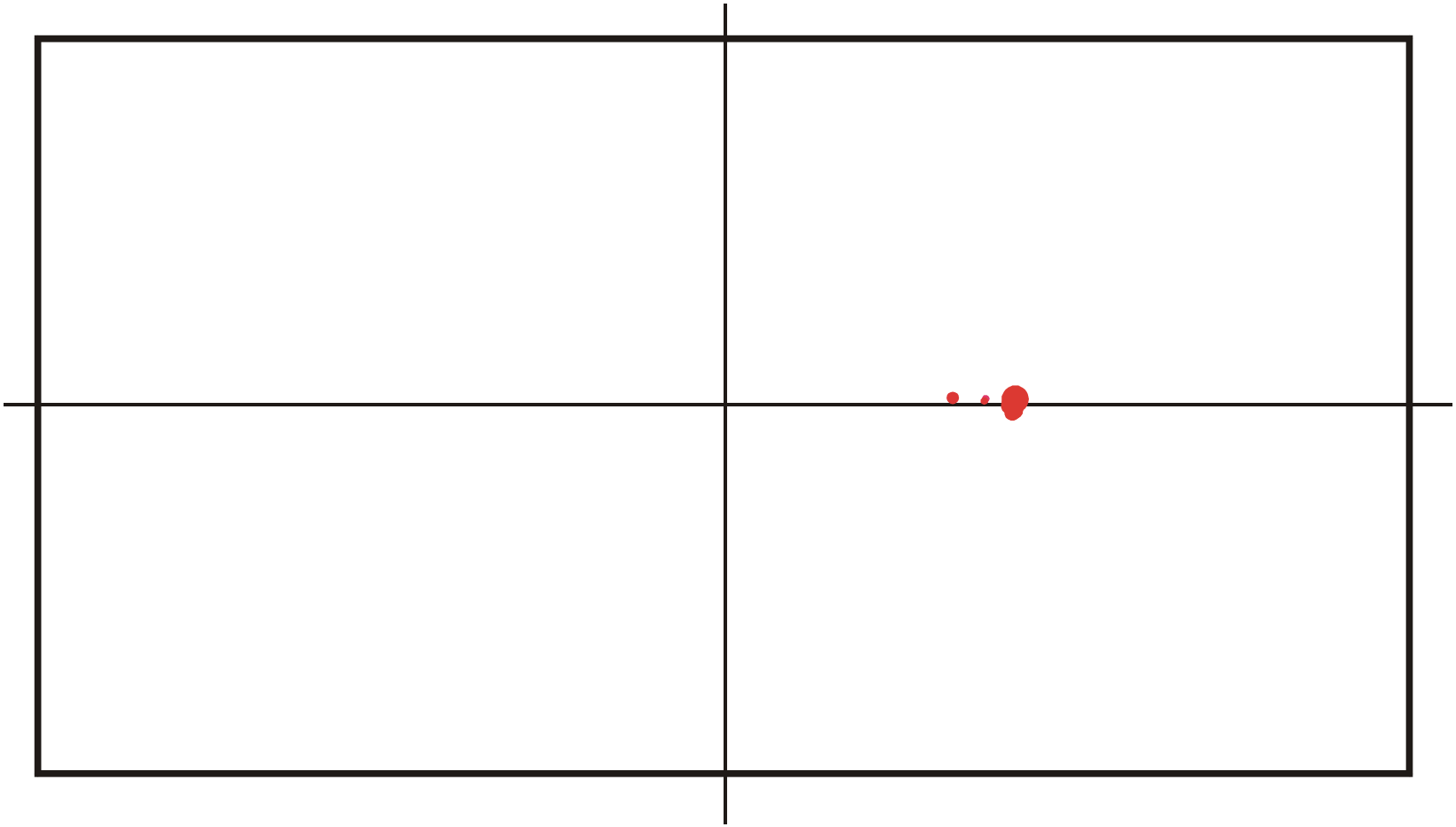
Spreading and evolution of a population on a neutral network : t = 830



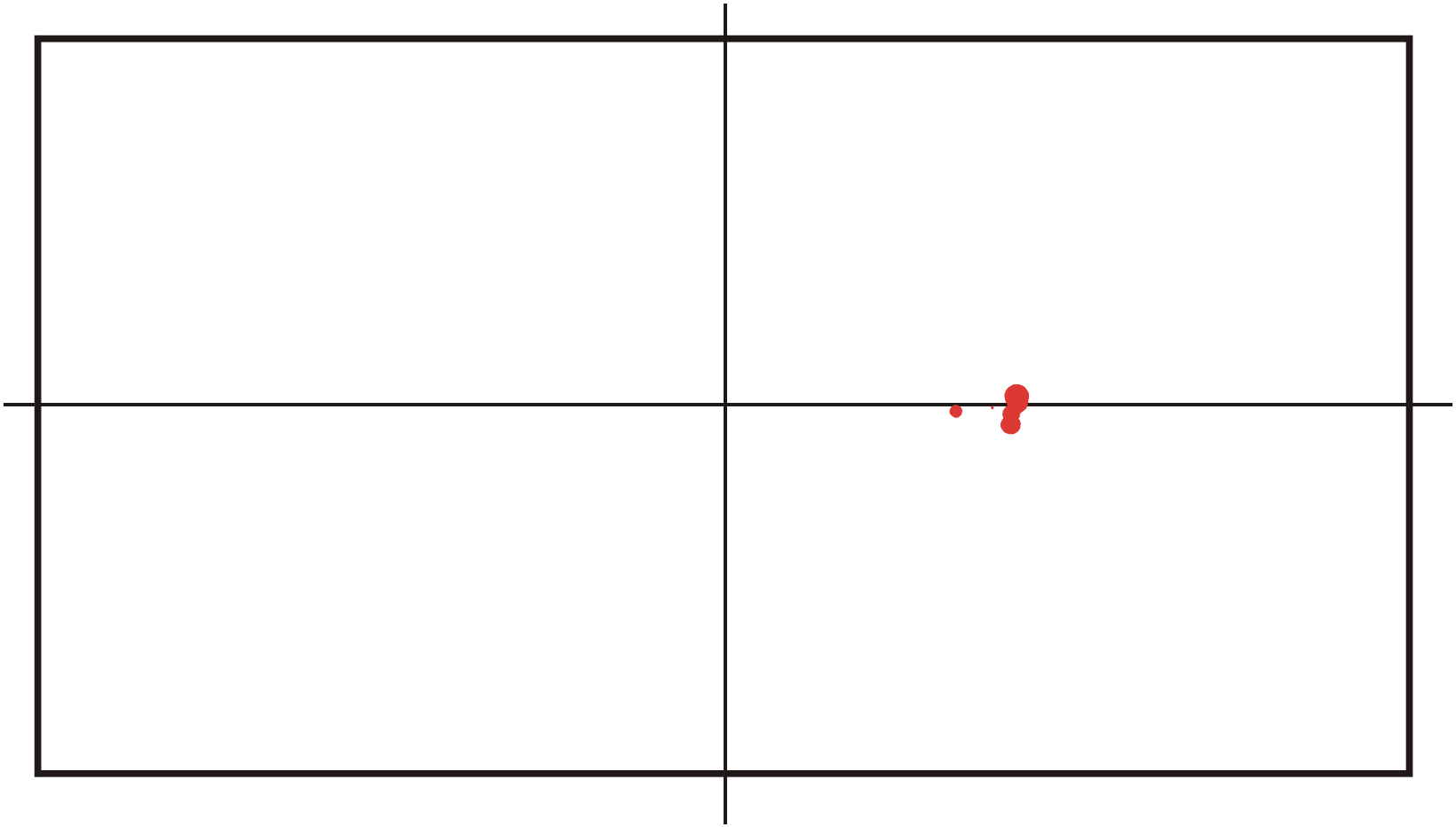
Spreading and evolution of a population on a neutral network : t = 835



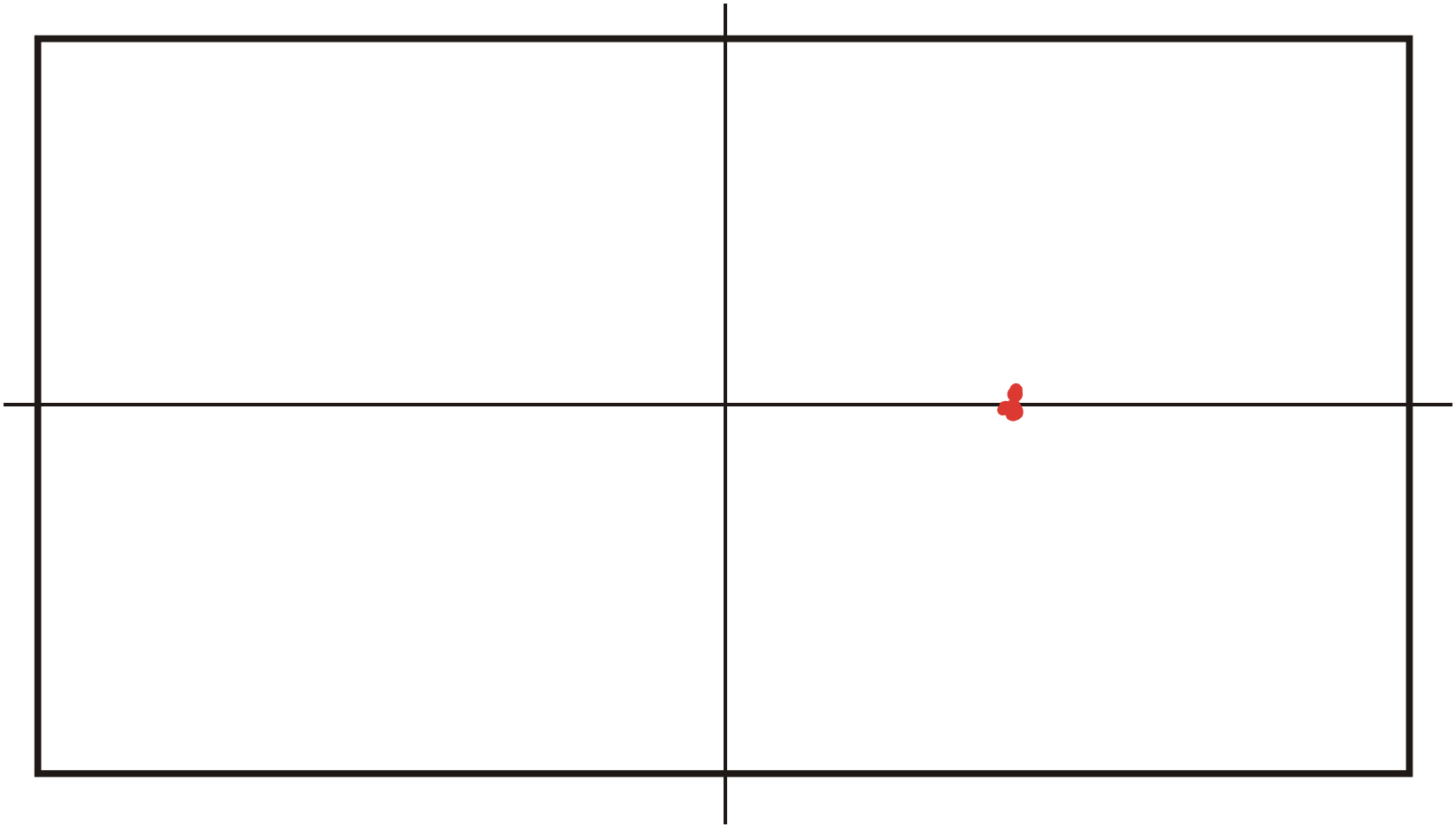
Spreading and evolution of a population on a neutral network : t = 840



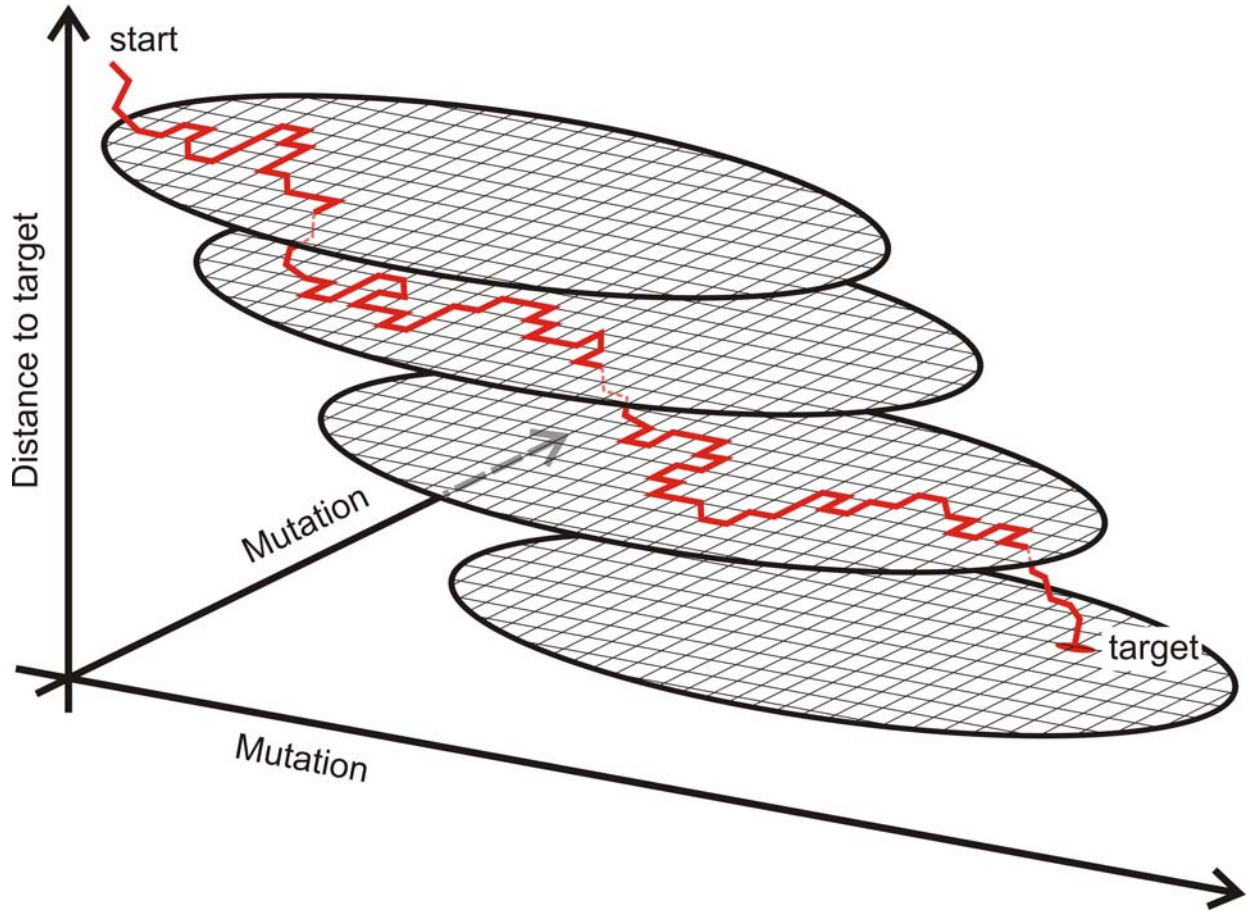
Spreading and evolution of a population on a neutral network : t = 845



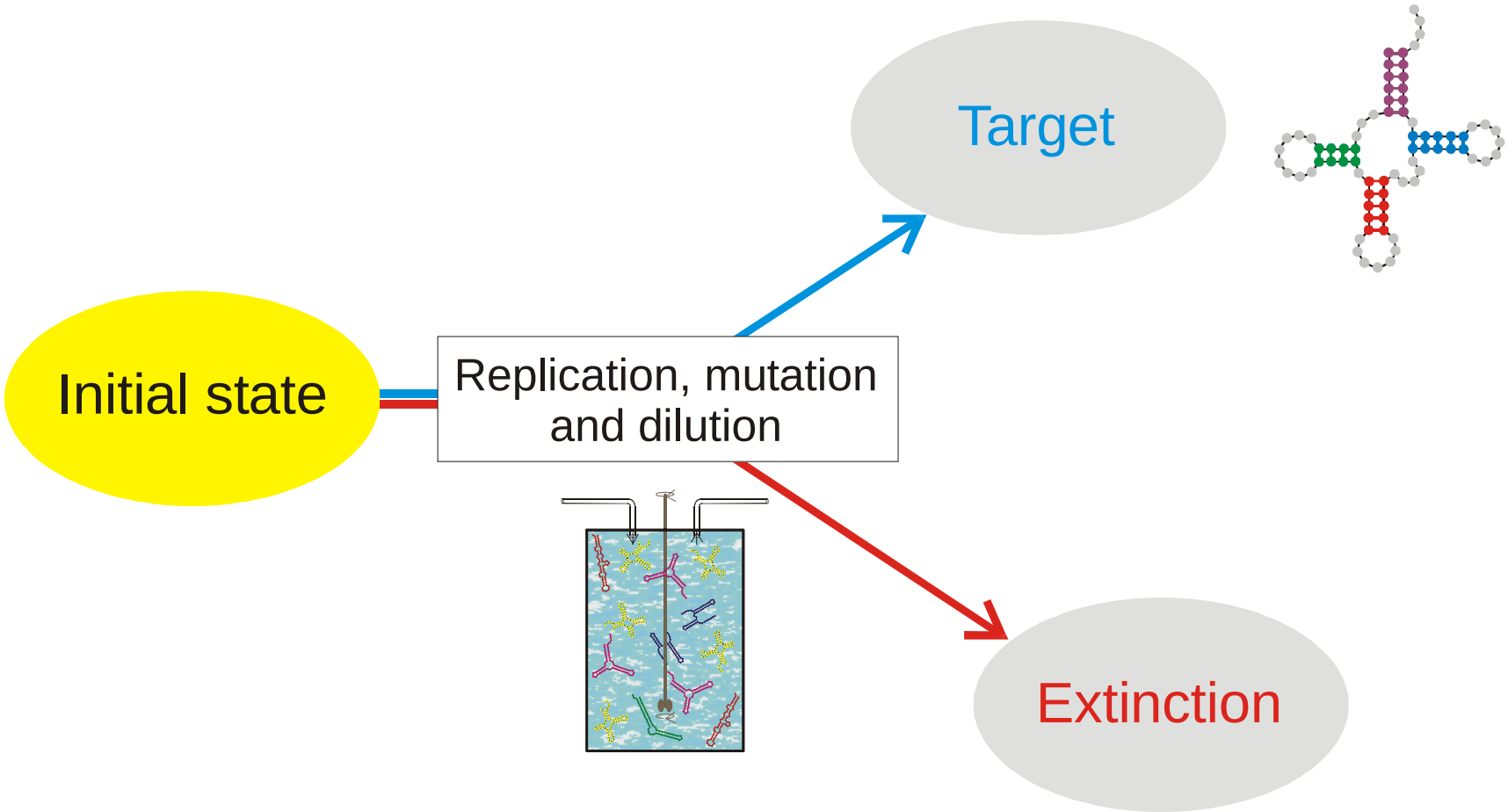
Spreading and evolution of a population on a neutral network : t = 850



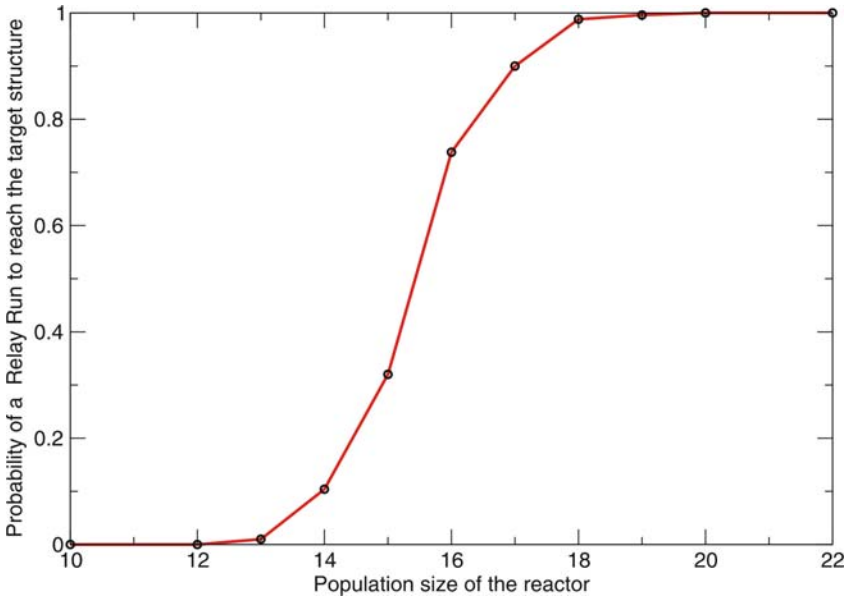
Spreading and evolution of a population on a neutral network : t = 855

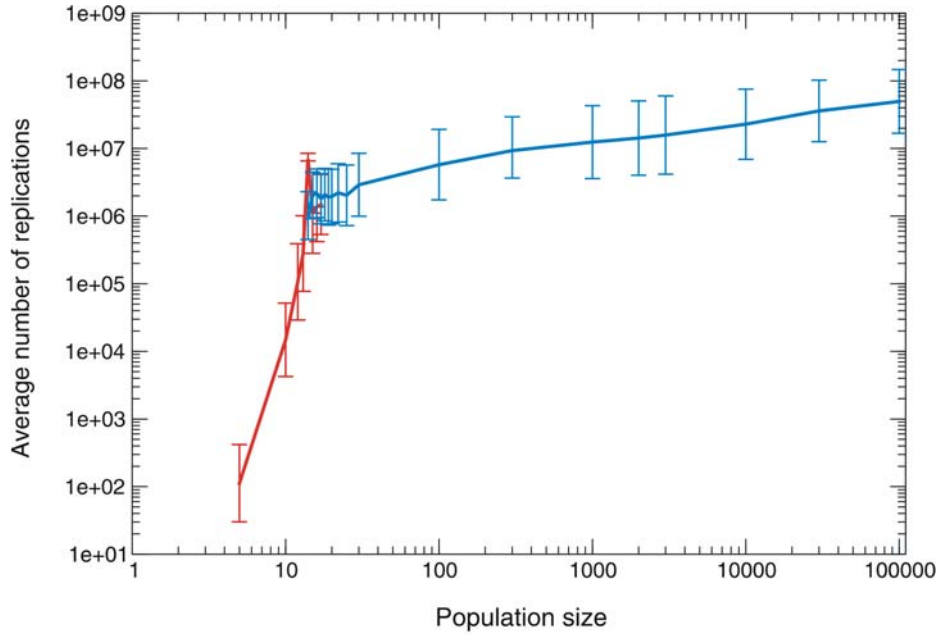


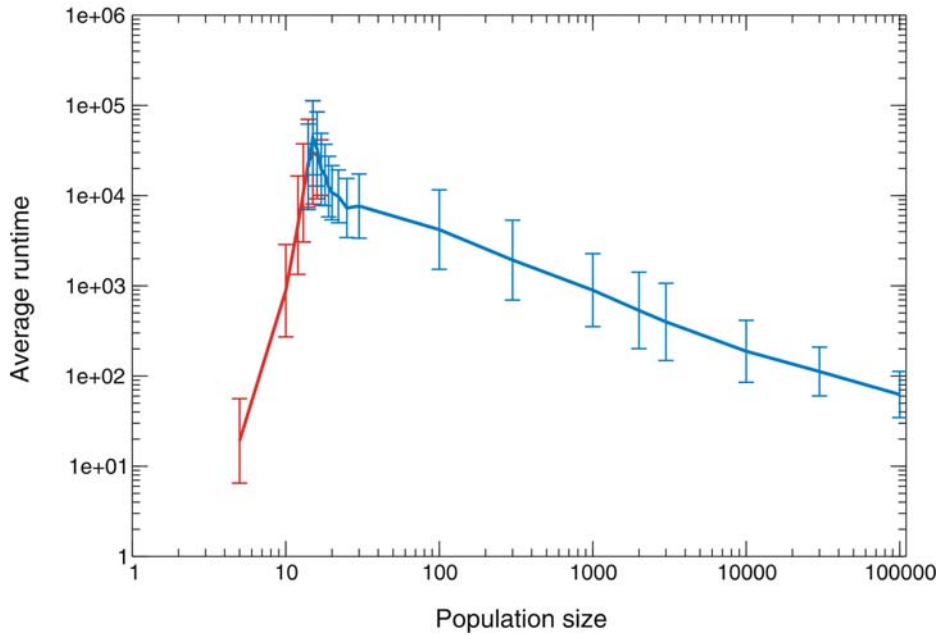
A sketch of optimization on neutral networks



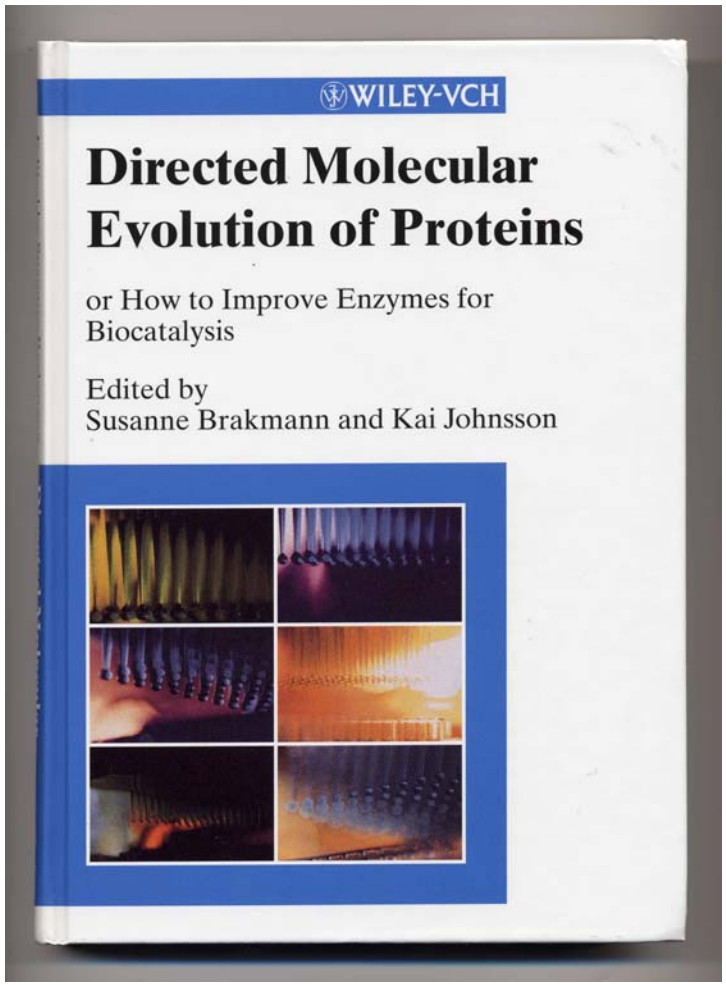
Replication and mutation as a stochastic process

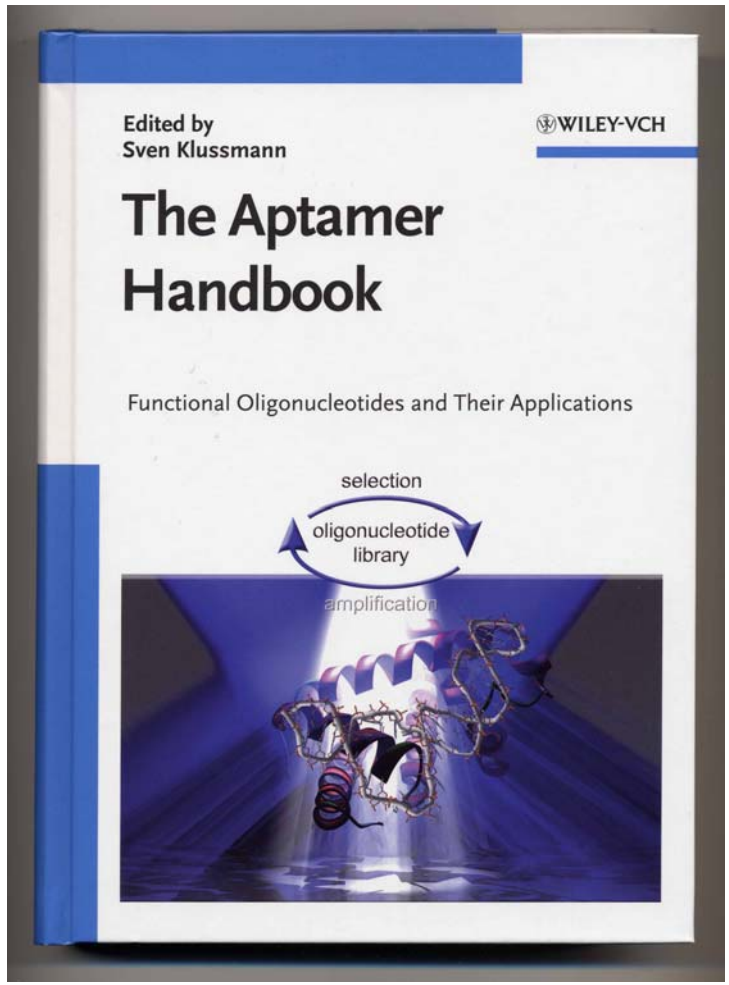






Expectation values as functions of population size:
Extinction probability, average number of replications and run time





Application of molecular evolution to problems in biotechnology

Acknowledgement of support

Universität Wien

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

Siemens AG, Austria

Universität Wien and the Santa Fe Institute

Coworkers

Universität Wien

Walter Fontana, Harvard Medical School, MA

Christian Forst, Los Alamos National Laboratory, NM

Christian Reidys, Nankai University, Tientsin, China

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

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

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

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

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

Web-Page for further information:

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