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
The work on a molecular theory of evolution started 40 years ago ......
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 Snow
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 pressures required for the effective selection of genotypes with superior fitness independent of mutation rates, i.e., there are distinct nonlinear 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 135–93, 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. Mitroz,3 Chin-Kun Hu,1 and M. W. Deem3
1Institute of Physics, Academia Sinica, Nankang, Taipei 11529, Taiwan
2Yerevan Physics Institute, Alikhanian Brothers St. 2, Yerevan 375036, Armenia
3Department 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-peak fitness landscapes 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 numbers: 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. Park1,2 and M. W. Deem1
1Department of Physics & Astronomy and Department of Bioengineering, Rice University, Houston, Texas 77005-1892, USA
2Department 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 form. 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
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

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, LINNEAN, ETC., SOCIETIES;
AUTHOR OF ‘JOURNAL OF RESEARCHES DURING H. M. S. BEAGLE’S VOYAGE
ROUND THE WORLD’

LONDON:
JOHN MURRAY, ALBEMARLE STREET.
1859.

The right of Translation is reserved.
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.
Motoo Kimura's population genetics of neutral evolution.


The average time of replacement of a dominant genotype in a population is the reciprocal mutation rate, $1/v$, 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
Binary sequences can be encoded by their decimal equivalents:

\begin{align*}
C = 0 \quad \text{and} \quad G = 1, \quad \text{for example,}\\
"0" & \equiv 00000 = \text{CCCCC}, \\
"14" & \equiv 01110 = \text{CGGGC}, \\
"29" & \equiv 11101 = \text{GGGC\textsuperscript{G}}, \text{ etc.}
\end{align*}
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
Definition of RNA structure

5'-end GCGGAUUAGCUCAGUUGGAGAGGCGCCAGACUGAAGAUCCUGGAGGUGUCUGUCGUCCAUCCACAGAAUUCGACC 3’-end
Criterion: Minimum free energy (mfe)

Rules: \_ (\_ ) \_ \in \{AU, CG, GC, GU, UA, UG\}

A symbolic notation of RNA secondary structure that is equivalent to the conventional graphs
many genotypes \Rightarrow one phenotype
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
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
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
<table>
<thead>
<tr>
<th>Number</th>
<th>Mean Value</th>
<th>Variance</th>
<th>Std.Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Hamming Distance:</td>
<td>150000</td>
<td>11.647973</td>
<td>23.140715</td>
</tr>
<tr>
<td>Nonzero Hamming Distance:</td>
<td>99875</td>
<td>16.949991</td>
<td>30.757651</td>
</tr>
<tr>
<td>Degree of Neutrality:</td>
<td>50125</td>
<td>0.334167</td>
<td>0.006961</td>
</tr>
<tr>
<td>Number of Structures:</td>
<td>1000</td>
<td>52.31</td>
<td>85.30</td>
</tr>
</tbody>
</table>

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:

\[ G \equiv C \text{ and } A = U \]
\[
\frac{dx_1}{dt} = f_2 x_2 \quad \text{and} \quad \frac{dx_2}{dt} = f_1 x_1
\]

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

\[
\eta(t) = \eta(0) e^{-ft}
\]

\[
\zeta(t) = \zeta(0) e^{ft}
\]

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
Chemical kinetics of replication and mutation as parallel reactions
\[
\frac{dc_i}{dt} = \sum_{j=1}^{N} Q_{ij} f_j c_j; \quad i = 1, 2, \ldots, N
\]

\[
\frac{dc}{dt} = W \cdot c; \quad \sum_{i=1}^{N} c_i(t) = c(t); \quad W = \{W_{ij} \doteq Q_{ij} f_j\}
\]

Normalization

\[
x_i = \frac{c_i}{c}; \quad \sum_{i_1}^{n} x_i = 1
\]

\[
\frac{dx}{dt} = W \cdot x - \bar{f} x = (G \cdot F - \bar{f} E) \cdot x; \quad \bar{f} = \sum_{i=1}^{N} x_i f_i
\]
Matrix $W$ and Frobenius theorem:

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

Primitive matrix $W$:

A nonnegative square matrix $W = \{w_{ij}\}$ is said to be a primitive matrix if there exists $k$ such that $W^k \gg 0$, 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) $\lambda_0$ is real and strictly positive

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

(iii) $\lambda_0$ is associated with strictly positive eigenvectors

(iv) $\lambda_0$ 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 $W$

$$W = \begin{pmatrix}
w_{11} & w_{12} & \cdots & w_{1n} \\
\vdots & \vdots & & \vdots \\
w_{n1} & w_{n2} & \cdots & w_{nn}
\end{pmatrix} = Q \cdot F$$

with

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

Driving virus populations through threshold

The error threshold in replication
Evolution of RNA molecules based on Qβ phage


Antiviral strategy on the horizon

Error catastrophe laid the conceptual origins at the middle of the 20th century, when the consequences of mutations on enzymes involved in protein synthesis, as a theory of aging. In these times, biological processes were generally perceived differently from today. Infections 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. Variations in differentiated organisms were seen as residing essentially from exchanges of genetic material associated with sexual reproduction.

The problem was to unravel the mechanisms of inheritance, expression of genetic information and metabolism. Few saw that genetic change is occurring at variance in all organisms, and still fewer recognized Darwinian principles as essential to the biology of pathogens versus, and cells. Population genetics rarely used bacteria or viruses as experimental systems to define concepts of biological evolution. The extent of genetic polymorphism among isolates of the same biological species came as a surprise when the first results on comparative electrophoretic mobility of enzymes were observed. The advent of in vitro DNA, RNA, and protein analysis sequencing techniques, molecular analyses of enzymes reinforced the conclusion of extreme inter-bacterial genetic variation within one species. Now, due largely to parallel progress in comparative genetics, we see cellular RNAs, both prokaryotic and eukaryotic, as highly dynamic. Most cellular processes, including such essential functions as transcription, translation, and growth, are increasingly perceived in inter-subcellular, and intracellular interactions.

The contributions to this volume have been chosen to reflect current issues of evidence (both theoretical and experimental) on which antiviral strategies based on genetic determination, rather than on viruses are being constructed. Two central themes have been underlined: the copying theory, which accounts for the inter-redundancy, and the copying fidelity that must be satisfied by any information-bearing replication system for the essential genetic information to be transmitted to progeny. Clearly related to the theoretical developments have been numerous experimental studies on quantispecies dynamics and their multiple biological manifestations. The latter can be summarized by saying that RNA viruses, like viruses of mammalian or bacterial nature, rather than defined genetic entities, are able to engage their potential to overcome selective pressures imposed by their environment. Instead of the virus, a complex of proteins in clinical practice and the design of vaccines for a number of diseases, are currently presented by a large number of strains, another line of growing interest is the enzymology of viral replication, mainly of RNA viruses, named at understanding the molecular basis for antiviral activities. Error catastrophe as a potential new concept strategy, is the importance of understanding the response to infections when such a host-virus interaction model may be inserted in virology. The challenge is to develop antiviral strategies through enhanced understanding of the field, but also a stimulating opportunity to the major problems to be attacked.

The idea of preparing this special issue arose as a brainstorming initiative of Ulrich Daleser, former Editor of Virus Research, and the idea was enthusiastically adopted by Luis Espuelas, recently appointed as Editor of Virus Research. I take this opportunity to thank Luis and the Editor-in-Chief of Virus Research, BrianMadey, for their continued interest and support to the research on viral evolution over the years.

My thanks go also to the 10 authors who despite their busy schedules have taken time to prepare excellent manuscripts, to Elsevier staff for their prompt responses to any requests, and, last but not least, to Mrs. Lourdes Orona at Centro de Biología Molecular "Severo Ochoa" for her patient dealing with the correspondence with authors and the final organization of the issue.

Enrique Domínguez
Centro de Biología Molecular "Severo Ochoa"
Consejo Superior de Investigaciones Científicas
E28049, Madrid, Spain
Tel.: +34 91 407 4542/0; fax: +34 91 407 4700
E-mail: edominge@cnb.csic.es
Available online 3 December 2004
Molecular evolution of viruses
Evolutionary design of RNA molecules


Application of molecular evolution to problems in biotechnology
Artificial evolution in biotechnology and pharmacology


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:

\[ Q_{ij} = p^{d_H(x_i, x_j)} (1 - p)^{(n - d_H(x_i, x_j))} \]

\[ d_H(x_i, x_j) \ldots \text{Hamming distance} \]
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 $\sigma = 10$
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 Física de la Materia Condensada, Universidad Autónoma 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

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 \( e_c = L_{\text{via}}/L \), where \( L_{\text{via}} \) 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

\[ n = 20 \text{ and } \sigma = 10 \]
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

\( n = 10, \sigma = 2 \) 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 \( p_k \):

\[
p_k = p + 4 \delta p(1-p) (X_{\text{rnd}}-0.5), \quad k = 1,2,\ldots,2^v
\]
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
Is the Kimura scenario correct for frequent mutations?
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", i.e. 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 wild 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.
Pairs of genotypes in neutral replication networks

Neutral network
$\lambda = 0.01, \ s = 367$

$d_H = 1$
\[
\lim_{p \to 0} x_1(p) = x_2(p) = 0.5
\]

$d_H = 2$
\[
\lim_{p \to 0} x_1(p) = a
\]
\[
\lim_{p \to 0} x_2(p) = 1 - a
\]

$d_H \hat{O} 3$
random fixation in the sense of Motoo Kimura
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 $d_H(X_i,X_j) = 1$. 

\[ \text{\ldots
dot\ldots ACAU GCGAA \dot\ldots
dot\ldots AUAUACGAA \dot\ldots
dot\ldots ACAUGCGCA \dot\ldots
dot\ldots GCAUAACGAA \dot\ldots
dot\ldots ACAUGCUAA \dot\ldots
dot\ldots ACAUGCGAG \dot\ldots
dot\ldots ACAUGCAGAA \dot\ldots
dot\ldots ACACGCAGAA \dot\ldots
dot\ldots ACGUACGAA \dot\ldots
dot\ldots ACAUAGGAA \dot\ldots
dot\ldots ACAUACGAA \dot\ldots
dot\ldots ACAUAGCAGAA \dot\ldots\]
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 $d_H(X_i,X_j) = 2$.
Neutral networks with increasing $\lambda$: $\lambda = 0.10, s = 229$
Neutral networks with increasing $\lambda$: $\lambda = 0.10$, $s = 229$
Neutral networks with increasing $\lambda$: $\lambda = 0.15$, $s = 229$
Neutral networks with increasing $\lambda$: $\lambda = 0.20$, $s = 229$
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
Phenylalanyl-tRNA as target structure

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,
\[ N = \# \text{ RNA molecules}, \]
is determined by the flux:
\[ N(t) \approx \bar{N} \pm \sqrt{N} \]

Mutation rate:
\[ p = 0.001 / \text{Nucleotide} \times \text{Replication} \]

The flow reactor as a device for studying the evolution of molecules \textit{in vitro} and \textit{in silico}. 
In silico optimization in the flow reactor: Evolutionary Trajectory

Quasistationary epochs

Evolutionary trajectory
28 neutral point mutations during a long quasi-stationary epoch

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
Smoothness within ruggedness: The role of neutrality in adaptation

MARTIN A. HOVMOELLER, JR., PETER F. STADLER, and WALTER FORTANA

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

To explain the high fraction rate of nucleotide substitutions in a population, Kimura (1) argued that the vast majority of genetic change as the level of a population must be neutral rather than adaptive. Sewall Wright's reaction to Kimura's point was probably 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 gave 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 conservatism there is at the sequence level (3). This makes Wright's point even more pertinent. How are we to imagine the relation between neutral evolution and adaptation? One aims 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 compared from the sequence by means of procedures based on thermodynamic data which have become standard in the past 15 years (4, 5). 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, it is a qualitatively important one.

Robust Properties of RNA Folding

The mapping from sequences to secondary structures is many to one (two exceptions: (a) there are many more sequences than secondary structures, and (b) some structures are realized much more frequently than others (6)). 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 RNA 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-dimensional character of sequence space, network of frequent structures percolates almost in such a way that each frequent structure is almost always realized within a small distance of an arbitrary sequence. These strong features seem to be intrinsic to RNA folding, since they are insensitive to whether the folding algorithm is thermodynamic, kinetic, or minimum free energy structure (9) or whether one considers one minimum free energy structure or the entire Boltzmann ensemble (10).

A Simple Model for In Vivo Evolution

To assess the consequences of these models 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-up implementation of the Zhang-Frederickson algorithm (8). The model consists of a population of RNA sequences of fixed length L, which replicate and mutate in a stirred flow reactor. RNA populations imaginable in the computer are in the laboratory are too large to be compared to the size of the sequence space (L), and a correct simulation must, therefore, resort to stochastic chemical reactions kinetics (11, 12). A selection pressure is induced by a dilution flow, which adjusts one time to keep the total RNA population fluctuating around a constant capacity N (11, 12). This setup mimics large-scale serial transfer technique (13), where sequences with a replication rate above (below) the average become (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, 10) between its secondary structure and a predefined target structure. How the target structure is the RNA secondary structure, 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, then, that in the present model the second source of neutrality arises only for sequences whose structures differ from the target.
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 $m_{ij} = \frac{d_i + d_j - d_{ij}}{2}$, 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 neural networks
Is the degree of neutrality in $\text{GC}$ space much lower than in $\text{AUGC}$ space?

<table>
<thead>
<tr>
<th>Number</th>
<th>Mean Value</th>
<th>Variance</th>
<th>Std.Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>52.31</td>
<td>85.30</td>
<td>9.24</td>
</tr>
<tr>
<td>2</td>
<td>36.24</td>
<td>6.27</td>
<td>2.50</td>
</tr>
</tbody>
</table>

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
Suboptimal structures and partition function of a small RNA molecule: \( n = 33 \)
Extension of the notion of structure

![Diagram showing RNA structure and free energy levels.](image-url)
Extension of the notion of structure
An RNA switch

J.N1LH

-28.6 kcal·mol⁻¹

-31.8 kcal·mol⁻¹


A ribozyme switch

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
Coworkers

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
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