299
Current Topics in Microbiology and Immunology

Editors

R.W. Compans, Atlanta/Georgia
M.D. Cooper, Birmingham/Alabama
T. Honjo, Kyoto · H. Koprowski, Philadelphia/Pennsylvania
F. Melchers, Basel · M.B.A. Oldstone, La Jolla/California
S. Olsnes, Oslo · M. Potter, Bethesda/Maryland
P.K. Vogt, La Jolla/California · H. Wagner, Munich

E. Domingo (Ed.)

Quasispecies: Concept and Implications for Virology

With 44 Figures and 7 Tables

Springer

Esteban Domingo
Centro de Biología Molecular "Servero Ochoa" (CSIC-UAM)
Universidad Autónoma de Madrid
Cantoblanco
28049 Madrid
Spain
e-mail: edomingo@cbm.uam.es

Cover illustration: Amino acid residues in the parvovirus MVMi capsid selected during evolutionary processes in mice. Highlighted in colours are residues of the spike whose replacement conferred the MAR phenotype (red), residues involved in primary receptor recognition (green), and residues conferring hematotropism (yellow). All other capsid residues are colored blue and shown as wireframe, except for the white-space-filling model of the residues defining the spike at the three-fold axis of symmetry.The structure of MVMi viral capsid is shown by the program RasMol (Sayle and Milner-White 1995) and the MVMi coordinates (1MVM) deposited in the PDB (Agbandje-McKenna et al. 1998).

Library of Congress Catalog Number 72-152360

ISSN 0070-217X
ISBN-10 3-540-26395-0 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-26395-1 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September, 9, 1965, in its current version, and permission for use must always be obtained from Springer-Verlag. Violations are liable for prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media
springeronline.com
© Springer-Verlag Berlin Heidelberg 2006
Printed in Germany

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
Product liability: The publisher cannot guarantee the accuracy of any information about dosage and application contained in this book. In every individual case the user must check such information by consulting the relevant literature.

Editor: Simon Rallison, Heidelberg
Desk editor: Anne Clauss, Heidelberg
Production editor: Nadja Kroke, Leipzig
Cover design: design & production GmbH, Heidelberg
Typesetting: LE-TEX Jelonek, Schmidt & Vöckler GbR, Leipzig
Printed on acid-free paper SPIN 11501312 27/3150/YL – 5 4 3 2 1 0

Dedication

Many observations on the great potential for phenotypic change of RNA viruses were made during the twentieth century, and John Holland reviews them in the closing chapter of this volume. There is a very remarkable precedent that concerns the noted Catalan virologist Jordi Casals, a pioneer of virology research who sadly passed away last year. Born Jordi Casals i Ariet in Viladrau, Girona on May 15, 1911 he died in New York City on February 10, 2004 after a productive and distinguished career in the United States. Jordi Casals left Spain in 1936 at the onset of the civil war and occupied positions at Rockefeller Institute for Medical Research in New York, Cornell University Medical College, Rockefeller Foundation, Yale University, and Mount Sinai School of Medicine (see a bibliographical note by Charles Calisher (2004) Arch Virol 149:1264-1266). A survivor of Lassa fever, he was a devoted, meticulous and observant scientist interested in virus classification. He was particularly concerned by confounding antigenic cross-reactions among viruses and viral diversity within a virus group. In a letter addressed to Charles Calisher in 1968 Jordi Casals wrote "A virus or species is a cluster of different individualities grouped around and resembling a prototype or model, rather than a number of strains all identical with a prototype." (I am indebted to Charles Calisher for sharing this information). Jordi Casals was honored at the recent annual meeting of the Virology Group of the Catalan Society of Biology (Societat Catalana de Biologia) held in Barcelona on 25 October 2004, with a biographical note read by Dr. Albert Bosch. This volume on quasispecies is dedicated to his memory.

Esteban Domingo

Preface

High mutation rates and quasispecies dynamics are essential features of RNA viruses. Continuous genetic variation and selection of virus subpopulations in the course of RNA virus replication are intimately related to viral disease mechanisms. Experience has taught that the adaptive potential of viruses must be taken into consideration in designing preventive and therapeutic antiviral strategies. The central topics of this volume are the origins of the quasispecies concept, and the implications of quasispecies dynamics for viral populations. It includes chapters that emphasize general concepts (quasispecies, sequence space, fitness, error catastrophe, lethal defection, adaptive responses, population bottlenecks, etc.) and chapters that deal with population dynamics of specific viruses such as Picornaviruses, Pestiviruses, Arenaviruses, Arboviruses, human immunodeficiency virus, plant viruses and some DNA viruses that display features of RNA genetics. In particular, implications of quasispecies dynamics in vivo are dealt with in several chapters. I thank all authors who have contributed their time and expertise to produce a markedly transdisciplinary volume that should provide a stimulus for future research. Hopefully, as sometimes happens with mutant swarms, the entire volume will be more "fit" than the sum of its chapters! I am also deeply indebted to Dr. Michael B.A. Oldstone for his generous invitation to serve as guest editor for a volume of this prestigious CTMI series. The Springer-Verlag staff and Lucia Horrillo of Centro de Biología Molecular "Severo Ochoa" were decisive to successfully completing this volume.

Madrid, April 2005 *Esteban Domingo*

List of Contents

List of Contributors

(Addresses stated at the beginning of respective chapters)

CTMI (2006) 299:1–31
© Springer-Verlag Berlin Heidelberg 2006

What Is a Quasispecies?

C. K. Biebricher (✉) · M. Eigen

Max Planck Institute for Biophysical Chemistry, Am Fassberg,
37077 Göttingen, Germany
cbiebri@gwdg.de

Abstract The concept of the quasispecies as a society formed from a clone of an asexually reproducing organism is reviewed. A broad spectrum of mutants is generated that compete one with another. Eventually a steady state is formed where each mutant type is represented according to its fitness and its formation by mutation. This quasispecies has a defined wild type sequence, which is the weighted average of all genotypes present. The quasispecies concept has been shown to affect the pathway of evolution and has been studied on RNA viruses which have a particularly high mutation rate. They (and possibly the majority of other species) operate close to the error threshold that allows maximum exploration of sequence space while conserving the information content of the genotype. The consequences of the quasispecies concept for the new 'evolutionary technology' are discussed.

1
Species and Quasispecies

The taxonomic classification of living organisms by Carl von Linné (1707–1778) was a milestone in biology. According to the view of his time, the living world was divided into different species that preexisted since the creation of the world and would persist until its end. Linné determined the systematic kinship of species by similarities in their anatomical build. A species was considered a society of individuals that are able to generate fertile offspring. This definition remained valid after the view of invariable species was shattered: Since Darwin it has been accepted that species can go extinct and new ones are formed by diversification.

In the first half of the twentieth century, this view took on a theoretical fundament with the Mendelian view of biological species and the Neodarwinistic synthetic theory. With the advent of the molecular biology in the second half of the twentieth century, interest focussed on organisms that have a much simpler, 'vegetative' propagation mechanism, like Prokarya. Obviously, the biological concept of species stated above does not apply to them, even less to forms of life such as viruses that have no cellular organisms. There is no gene pool to select from and no continuous shuffling of gene alleles. Nevertheless, these forms of life have been classified with reasonable results. Obviously, the reproducing individuals in prokaryotic taxons and virus populations must lead to a society that shares many properties with the 'classical' species; it is called a 'quasispecies'. Its emergence and its properties are described in this article.

2
Growth

The purpose of a theory is not to describe *how* processes occur in nature in detail, but rather to understand *why* certain regularities can be observed. Theory is no alternative to experiment, but theory and experiment support one another, theory by interpreting experimental results and suggesting further meaningful experiments, which then can be used to refine the theory. However, theory needs reduction to the most important parameters while neglecting less important contributions. It is necessary to find suitable experimental conditions where complexity is low enough to verify theoretical predictions.

The different reproduction mechanisms among the superkingdoms in nature cause difficulties; hence one must find the fundamental property of reproduction. Fisher [85], in deriving the laws of population dynamics of

a Mendelian population, was fully aware of this difficulty. He assumed that the population contains a sufficient number of males to fertilize all females, regarded only the female part of the population and obtained Malthusian, i.e. exponential, growth. The potential for exponential growth is an inherent property of life, even though it is rarely observed in nature because growth is rapidly limited by a shortage of resources. Indeed, processes such as leavening, where exponential spread can be observed, have always served as a metaphor for life. Exponential growth can be easily observed in the superkingdom of Prokarya. For further considerations, horizontal gene transfer phenomena such as transformation, infection, transfection, and conjugation should be disregarded.

An experimental system that allows the study of exponential growth is the growth of a bacterial culture in a defined nutrient medium under controlled conditions. Bacteria grow by metabolizing nutrients from the medium, and divide after a certain time τ to two daughter cells. While the τ values vary from one bacterium to another, the average τ values for large populations of a species can be determined with high precision; the population density can be preferentially derived from macroscopic properties of the population, e.g. by turbidity measurements. After the time t, the initial concentration of bacteria has grown to

$$c(t) = c(t = 0) \cdot e^{At} \tag{1}$$

where the growth rate $A = \ln 2t/\tau$ (Table 1 lists the symbols for the parameters used in this article). Dynamical processes are best described by differential equations. The rate of population increase is directly proportional to the number of parental cells and their specific growth rate parameter:

$$dc/dt = Ac\,. \tag{2}$$

In experimental measurements, this equation is only strictly observed at large dilution. At higher concentrations, the consumption of resources, in particular of oxygen, slow down bacterial growth until the growth curve eventually levels in to a maximal concentration. While this behaviour can be described by the logistical equation, we shall use for our arguments the simple form. A specially designed apparatus, the chemostat, was invented to avoid large concentration increases by pumping in fresh medium at the constant rate Φ and removing a balancing volume of bacterial culture [93]. The dynamical equation then becomes

$$dc/dt = (A - \Phi)c\,. \tag{3}$$

If $A - \Phi > 0$, the population increases, if $A - \Phi < 0$, the population decreases. When $A - \Phi = 0$, the bacterial concentration should be constant,

Table 1 Parameters

A_i	Growth rate parameter of type i
$\bar{A}$	Average growth rate parameter of the population ($\bar{A} = \sum_k A_k x_k$)
D_i	Mortality or decomposition rate parameter of type i
E_i	Excess reproduction rate of type i ($E_i = A_i - D_i$)
$\bar{E}$	Average excess reproduction rate of the population ($\bar{E} = \sum_k E_k x_k$)
N	Number of generations
Q_{ii}	Probability of precise reproduction of sequence i ($Q_{ii} = \bar{q}^{\nu_i}$)
Q_{ik}	Probability of producing type i while replicating sequence k
W_{ii}	Rate parameter of precise (excess) production of sequence i: ($W_{ii} = Q_{ii}A_i - D_i$)
W_{ik}	Rate parameter of production of sequence i by erroneous copying of sequence k ($W_{ik} = Q_{ik}A_k$)
c_i	Population density of type i
d_{ik}	The Hamming distance between two sequences i and k, i.e., the number of positions at which both genomes differ.
$\bar{q}$	Average fidelity of single digit reproduction (insertion of correct nucleotide)
$1 - \bar{q}$	Probability of single error production
x_i	Fraction of type i in the total population (type frequency) ($x_i = c_i / \sum_k c_k$)
$\tilde{x}_i$	type frequency of i in the evolution steady state
$\mathcal{F}_i$	Fitness value of species i; $\mathcal{F}_i = A_i/A_m$
ε_i	Average error rate per sequence i ($\varepsilon_i = \nu_i(1 - \bar{q})$)
ν_i	Genome length of type i
$\bar{\sigma}_m$	Average superiority of master sequence over competitor mutants ($\bar{\sigma}_m = A_m / (D_m + \bar{E}_{k \neq m})$, requiring $\sigma_m > 1$)
τ	Duplication time of a bacterium
Φ	Flux rate of growth medium in reactor

but measurements show irreproducible changes in the bacterial concentration. Experiments with finite populations never observe deterministic laws exactly: inevitable statistical fluctuations of the A values lead to concentration fluctuations that are not compensated because A is independent of the concentration. Another experimental device, the turbidostat, keeps the turbidity of the bacterial culture and thus the bacterial concentration constant.

A nonbiotic system of studying exponential growth is RNA replication [111] in a cell-free system, which is even capable of Darwinian evolution [86]. An RNA strand serving as template replicates, producing a complementary replica or minus strand, which can itself replicate again to produce plus strands. This cross-catalysis leads to exponential growth of the RNA [8, 9]. Direct measurement of the exponential growth is difficult,

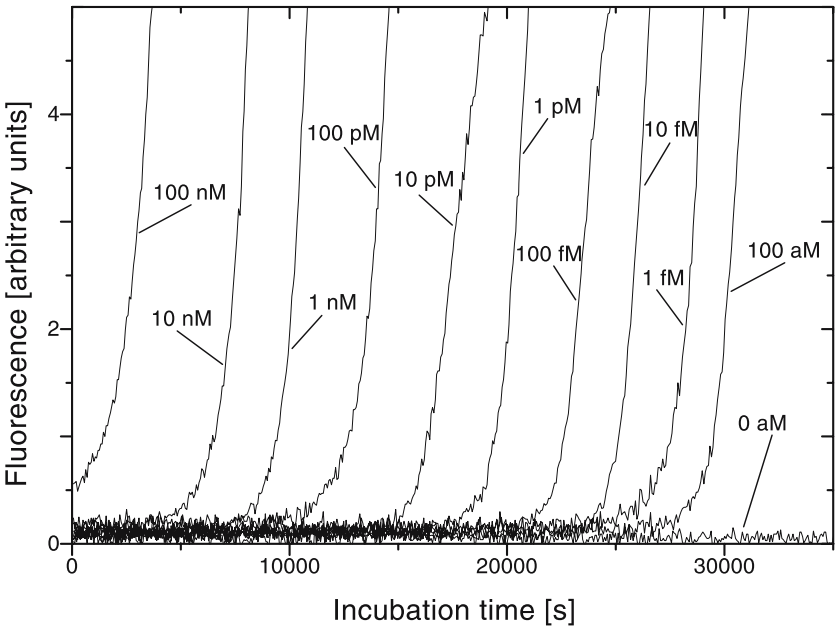


Fig. 1 Growth profile of replicating RNA [121]. A mixture containing 0.75 mM ATP, CTP, GTP and UTP, 1 μM RNA polymerase from *Escherichia coli* and 3 μM thiazole orange was inoculated with RNA species EcorpG at different concentrations and the fluorescence of the probes was measured in a multichannel fluorimeter at an excitation wavelength of 488 nm and an emission wavelength of 515 mm.

because exponential growth requires that the replicase enzyme that catalyses the replication is present in large excess of the growing RNA. As in measuring the bacterial population with turbidity, only populations above a boundary give a measurable signal. The solution is that one measures growth at different initial concentrations (Fig. 1). If the population is diluted by the factor F_{dil}, the growth curve appears to be shifted on the time axis by the difference Δt and one calculates the replication rate to be $A = \ln F_{\mathrm{dil}}/\Delta t$ [9, 121] (Fig. 1).

3
Selection of the Fittest

Populations are rarely composed of absolutely identical individuals, but contain usually *types* that can be distinguished by certain criteria. In the past, the criteria were obtained by visual inspection, e.g. Mendel distinguished the types of his pea population by the color of the flowers and the shape

of the seeds. In the first half of the twentieth century, a large number of different variants of the fruit fly *Drosophila melanogaster* were investigated that were distinguished by the visual differences in the appearance at the larval or imago stages. Bacterial types can be distinguished by inspection of the colonies they form on solid nutrient media. Types classified by such criteria are called *phenotypes*.

In contrast to our lack of quantitative predictions of evolutionary events, the molecular basis of transmitting phenotypic traits to offspring is well understood: The information needed for morphogenesis and function is encoded into the *genotype*, the nucleotide sequence in each organism's genome. The key step in reproduction is replication of the genome. The enormously complex process of decoding the genotypic program resulting in a phenotype with certain properties is called expression.

Let us assume that the different types in a population ignore one another and that each individuum grows as if it were alone, by observing the growth of different types, e.g. species, of bacteria in a medium that contains nutrients in vast excess. Each of them is present at a relative concentration $x_i = c_i / \sum_k c_k$ and grows with the specific fecundity A_i. As in the previous examples, the total concentration of bacteria shall be kept constant by removing the excess production of offspring caused by the average growth of the types $\bar{A} = \sum_k A_k x_k$ by appropriately diluting the medium. The relative population of each type then changes with the rate

$$dx_i(t)/dt = \{A_i - \bar{A}(t)\}x_i(t) \,. \tag{4}$$

Note that equation (4) is nonlinear because $\bar{A}(t)$ is time-dependent. Types whose growth rate exceeds the average are enriched while types with $A_i < \bar{A}(t)$ are depleted in the population. It is obvious that the value for A_i of a type i is a good measure of its 'fitness' in Darwin's sense. The enrichment of fitter types in the population raises the average growth rate, and more and more types fall below the average, until eventually the population reaches a maximum growth rate where only the *master* type m, the type with the maximal fitness, survives:

$$\bar{A}(t) \to A_{\max} \qquad x_m \to 1 \qquad x_{i \neq m} \to 0 \,. \tag{5}$$

Natural selection is thus indeed an immediate consequence of autocatalytic reproduction [36]. The following conditions must be fulfilled:

1. One type cannot be converted into another, e.g. because the types belong to different species.
2. The growth of one species is not influenced by the presence of another.
3. The selection proceeds far from chemical equilibrium.

In reality, populations do not change solely by differences in fecundity. Mortality rates D_i also influence the composition since types with higher mortality also get depleted in the population. In this case, fitness is correlated with the excess reproduction rate $E_i = A_i - D_i$. Not only maximizing the reproduction rate, but also minimizing the mortality rate is then important for selection. A practical application of this selection process is the preparation of pure bacterial cultures from a natural specimen by the use of selective media.

In nature, the selection process we described is much more complicated. Fecundity and mortality rates are often not time-independent. Under realistic conditions, say in an ecosystem, the types compete for nutrients, or interact one with another by mutually influencing their fecundities and mortalities, most conspicuously in a predator–prey relationship. The fecundity and mortality rates of one species depend on the presence of other types and change with time. The situation is further complicated by variations of the environmental conditions. Calculation of selection values or fitness values is then very difficult.

When working under defined conditions with the RNA replication system described above, it was possible to predict precise selection values. At low RNA concentrations, with nucleoside triphosphate substrates and replicase in vast excess, the RNA species with the highest growth rate is selected as described above. When the RNA concentration reaches the replicase concentration, however, the growth characteristic as well as the selection drastically changes. Since an RNA strand must bind a replicase molecule for replication, exponential increase of the RNA concentration is suddenly replaced by linear growth when the template concentration reaches the enzyme concentration. Free RNA strands, both of plus and minus polarity, accumulate, which can react one with the other by formation of a double helix that is unable to replicate. With the onset of the linear phase it is no longer the RNA species with the highest growth rate that wins, but rather the one with the highest rate of replicase binding. At still higher RNA concentrations, minimizing the loss rate by double strand formation also becomes important. Eventually, often a stable ecosystem is formed where different RNA species occupy constant parts of the population. The rather complicated but quite instructive calculation of the selection values can be found in the literature [12, 13].

4
Mutation

Selection among preformed and invariable types alone does not suffice for Darwinian evolution. Evolution needs the formation of new types by a process

called mutation. While in Mendelian populations the offspring usually is of a different type than the parents because the different parental gene alleles are reshuffled, 'vegetatively' growing populations such as bacteria and viruses have been assumed to produce identical offspring. New types are created by mutation, i.e. by misincorporation of nucleotides during the replication process. Evolutionary innovation is driven in all biological species by mutation.

It is straightforward to introduce *mutation rates* W_{ji} describing the rate of producing type j during reproduction of type i. However, in many cases mutation rates are too low to be realistically described by deterministic rate coefficients, even in large populations. When writing down an equation with rate coefficients, one has to be aware that the parameters are merely averaged probabilities. Before further conclusions can be drawn, one has to estimate whether there is a reasonable probability that the pertaining mutation will take place at all.

How do we modify the equation to take mutation into account? Reproduction can proceed with fidelity, producing offspring of the parental type (W_{ii}) or it can erroneously produce offspring of another type ($W_{ji}, i \neq j$). The production rates of each type i are composed by the rates of fidelity reproduction of the same type and the rates of erroneous reproduction from other types [32]. The equation considering selection and mutation then is

$$dx_i/dt = \{W_{ii} - \bar{E}(t)\}x_i(t) + \sum_{k \neq i} W_{ik}x_k(t) \tag{6}$$

where $\bar{E}(t)$ is the average excess growth rate of the total population. If the mutation terms become negligible, i.e. at high fidelity reproduction, equation (6) is converted to the selection equation (5). The reproduction rate with fidelity can also be written as $W_{ii} = Q_{ii}A_i - D_i$, where Q_{ik}, the *mutation probability* is the probability of producing type i in a reproduction round of type k. When D_i can be neglected, which can often be obtained by providing suitable environmental conditions, then equation (6) converts to

$$dx_i/dt = \{Q_{ii}A_i - \bar{A}(t)\}x_i(t) + \sum_{k \neq i} Q_{ik}A_kx_k(t)\,. \tag{7}$$

There we have two mutation terms: (i) the *mutational gain* $\sum_{k \neq i} Q_{ik}A_kx_k$, producing type i by replication of other types, and (ii) the *mutational loss* $(1 - Q_{ii})A_ix_i = \sum_{i \neq k} Q_{ki}A_ix_i$, producing mutants in reproduction.

How can Q_{ik} values be measured? With the sequencing machines available today, it is in principle possible to screen the genotypes of the offspring of a single reproduction cycle, but, since mutations are very rare, millions of sequence determinations would be needed to get values with some statistical significance. In the past, one had to work with phenotypic markers.

The challenge is determining a small number of mutants in a vast excess of wild type.

Luria and Delbrück [84] have invented an ingenious method to measure mutations by observing the formation of mutants resistant to phage T1 (type r) in a sensitive wild type (type s) population. Since the wild type population has to be created, one needs a conditional lethal mutation, i.e. under one set of conditions, the 'permissive' one, the bacteria grow normally while under 'nonpermissive' conditions growth is inhibited. Mutation can be observed only under permissive conditions. Since the initial population already contains mutants and since more than a single reproduction round takes place, the accumulation of mutants with time has to be observed.

When performing several independent experiments, the absolute number of resistant bacteria in the offspring was found to scatter widely, depending on whether the mutation was formed early or late during growth, showing the stochastic and undirected nature of the mutation. From the kinetic profiles of mutant accumulation after further growth, one can directly calculate Q_{rs} values, where r is the viable, resistant and s the lethal, sensitive phenotype. The above equation can be simplified. Mutations are rare ($Q_{ss} \approx 1$), wild type bacteria are in large excess ($x_s \approx 1$) and the growth of the total population depends only on the wild type ($\bar{A} \approx A_s$), and we obtain

$$dx_r/dt = \{A_r - A_s\}x_r(t) + Q_{rs}A_r \,. \tag{8}$$

When the mutation is entirely neutral, the first term cancels and the linear increase of the type frequency gives the *mutation rate* $W_{rs} = Q_{rs}A_r$ [93]. Otherwise one observes an exponential and a linear growth component and one can determine the mutation probability and, from the exponential part, the selection rate $A_r - A_s$. Generally, we can introduce classical fitness values $\mathcal{F}_i$ for a type i by dividing its growth rate by the rate of the predominant wild type w and divide the time by the growth rate to obtain the generation number N. The equation, which is valid only for a population dominated by the wild type where the contributions from other mutants to the mutational gain are negligible, then reads

$$dx_i/dN = \{\mathcal{F}_i - 1\}x_i + Q_{iw} \,. \tag{9}$$

Simplifications introduced by experimental conditions always limit the validity of the interpretation. When the time periods for observing the accumulation of phenotypic mutations was extended to several weeks by working in the turbidostat [31], non-reproducible, erratic patterns appeared, comprising periods where the relative mutant increased linearly as described above, but also periods where the relative mutant concentration decreased. How can

the periods of decrease be understood? Obviously, the model chosen for the analysis was too simplistic: there were not only two types in the solution. Other mutations not changing the phenotype also occurred and, if one happened to be advantageous, it was enriched in the population and eventually replaced the old wild type together with its mutant spectrum. A new mutant spectrum was then built around the new wild type.

Furthermore, the measured mutation probabilities were rather high since a degenerative mutation has been chosen. Resistance to phage T1 is simply achieved by abolishing the phage receptor, which is not needed by the host cell under these conditions. Hundreds of different mutational events may lead to destruction of a gene, causing a specific phenotypic change while only one (or a very few if pseudorevertants are possible) leads to the restoration of a lost gene. The found mutation probability was thus the sum of many specific mutation probabilities that led to the same mutant phenotype.

Benzer [5, 6] succeeded in adapting this technique to map mutations within a gene of bacteriophage T4 by measuring the probabilities to restore wild type function by recombination after double infections. A surprisingly precise map of the gene was obtained. Two mutants not able to recombine to wild type were judged to be identical, and Benzer thus also obtained type frequencies. Benzer noted that type frequencies scattered for different loci, and called the loci with particularly high type frequencies 'hot spots'. This name suggests that type frequencies are particularly enhanced at loci with high mutation rates, but this conclusion is not justified because Benzer's measurements were taken from a snapshot of the mutant distribution at a certain time. Other factors also contribute to the type frequency, most notably error propagation by replication of the mutant genomes. As previously seen, only a kinetic study of the mutant spectrum can clarify the contributions of mutation and error propagation.

Measurements of *error rates* of replicating enzymes by determining the increase of the number of revertants from a lethal phage mutant in vitro [42, 43, 45, 79, 78] restricted the growth to one replication round to avoid error propagation. Strictly seen, these error rates only apply to the specific mutation tested in the experiment, because error rates are not uniformly distributed within the sequence. Average enzyme fidelities $\bar{q}$, i.e. the probability of inserting the correct nucleotide in the incorporation of a single nucleotide, and average error rates $1 - \bar{q}$, i.e. the probability of producing a mismatch in incorporating a nucleotide can be estimated only by statistical analysis of several such experiments. With an average fidelity of $\bar{q}$, the probability of obtaining a precise copy of a sequence with the chain length ν is calculated to be $Q_{ii} = \bar{q}^{\nu}$ [37]. This suggests that at sufficiently large chain lengths, the probability of making a precise copy can become small. While the probabilities of

making specific mutations are always small in comparison to the probability of making correct offspring, the summation of the huge number of mutants may add up to a majority, causing $Q_{ii} \ll 1$. This will cause a fundamental difference to the notion of a population dominated by a single type.

5
Sequence Heterogeneity in Virus Populations

The experimental observation of mutant spectra with normal organisms is difficult. Their genome sizes are huge, their obtainable population sizes are small and DNA replication is highly accurate. With RNA viruses, which have small genomes, large population numbers and especially high mutation rates [27, 28, 26], observation becomes feasible. Still easier is the observation of mutant spectra in replicating RNA populations, since the phenotypic expression of the RNA type is reduced to directing the replicase to replicate it as efficiently as possible [8, 9, 101].

The first method of sequencing RNA involved digesting radioactively labelled RNA with specific RNases and separating the resulting oligonucleotides by two-dimensional electrophoresis [103]. A highly reproducible spot pattern called fingerprint was obtained, where the radioactivity of single spots related one to another as integer multiples, clearly indicating a defined sequence of the population. This is in agreement with a population dominated by the wild type.

As the sequence analysis of RNA phage Qβ ($\nu = 4216$ bases) was progressing in the laboratory of Charles Weissmann [14], the phage was recloned by making a population from a single plaque. Surprisingly, the fingerprint pattern of the resulting clone deviated from the former wild type [25]. Was it a rare mutant that was picked by chance? Obviously not: fingerprinting several clones derived from single plaques revealed frequent deviations from the wild type fingerprint, but the deviations were in different places. Passaging the clones resulted in reappearance of the 'wild type' fingerprint. Apparently, the phage RNA population is highly heterogeneous, but has a defined *average* of the sequence, the wild type sequence [25]. Subsequent studies with eukaryotic viruses [96, 26] confirmed this view.

An average mutation rate $\bar{q} = 3 \times 10^{-4}$ was estimated [2] by calculating the frequency of mutant and revertant clones in reversion and competition experiments. This rate is quite high and suggests that a large proportion of the virus progeny contains errors. Indeed, the specific infectivity of most RNA viruses was found to be quite low: only a fraction of virions produce plaques. The majority of them may have lethal mutations, but this is not yet proven.

Reverse transcription of phage RNA and cloning at the DNA level revealed that the phage population does contain many lethal or nearly lethal phages that fail to produce a plaque (Biebricher, unpublished observations). While each mutation may be rather rare and cannot be propagated, the chance of picking a lethal mutant is substantial because there are so many of them.

When viruses from a single clone are amplified, they produce mutants. The mutants compete with each other, until eventually a stable steady state mutant spectrum is obtained, where the contributions from mutation and selection balance each other for every mutant. When starting from a different mutant of the same population, the same steady state is obtained. When determining a virus sequence, it is thus instrumental to work with a reproducible, equilibrated mutant spectrum. It is formed whenever a large population grows for sufficient time. On the other hand, when picking small parts of the population, in the most extreme part a single individuum as in the plaque test, one obtains irreproducible results: different mutants are picked, often with substantially lower fitness values than the wild type [29] (Muller's ratchet [87]; see the chapter by C. Escarmís et al., this volume).

6 The Sequence Space

Mutants are ordered by aligning their sequences with the wild type sequence (Fig. 2). A quantitative kinship relation can be given with the Hamming distance [54], i.e. the number of bases different in their sequences. When doing so, one often observes that certain mutants share several base exchanges and form a clan within the population. The clan resulted from error propagation derived from a common ancestor, because it is highly unlikely that several identical base exchanges take place in a single reproduction cycle. One-dimensional alignment cannot adequately display this kinship of one mutant to another.

The adequate topography of mutant distribution is the *sequence space*, a hypercube [33, 107] with at least as many dimensions as the chain length of the genome, ν. The Hamming distance between two mutants is then the minimum number of steps to go from one sequence to the other. For nucleic acids, which are composed of four different nucleotides, one would require a 2ν-dimensional space. In the full sequence space, each mutant would occupy two positions because the replication mechanism produces the complementary sequence, occupying the antipodal corner of the hypercube and having the maximal Hamming distance of 2ν. If one wants to avoid two mirror landscapes, one can neglect the complement by assuming a $(2\nu - 1)$-dimensional

sequence space. A hypercube can of course not be imagined geometrically and projections on paper are very difficult to survey for larger ν, but the mathematical relations can be used for understanding evolutionary pathways [33]:

1. The storage capacity is very high. For a viral RNA with a genome length of 10^4 nucleotides, it would be 10^{6000}, an astronomically large number.
2. Distances between corners remain small, the maximal Hamming distance is 2ν.
3. The connectivity is very high: at each corner one has 2ν directions to chose.

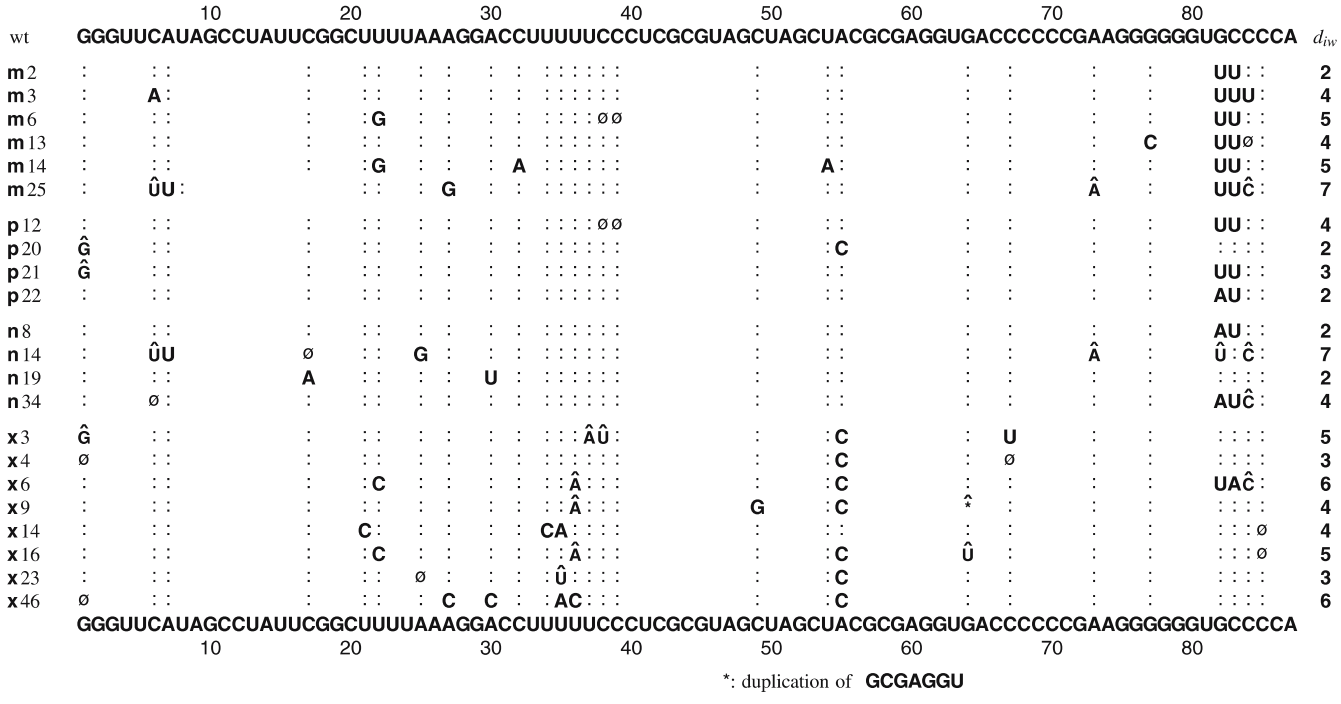
It is obvious that the small distances between any pair of corners does not bring any advantage if one tries to go to a certain target corner by a *random* walk. Without guidance in choosing the right direction, one gets lost in the vast space.

Guidance is provided by fitness differences. Instead of representing the sequence space by a cube, we span a 2ν-dimensional hyperplane and add as 'vertical' coordinate the selection values for each locus. We obtain a 'fitness' landscape that has some properties in common with landscapes on Earth. Selection values are not distributed at random in the sequence space but form coherent mountains with several peaks. Gradients provide guidance: while gravity directs to lower heights, selection drives uphill.

How are the landscapes formed? Genotypes express into their corresponding phenotypes, and a global evaluation of their reproduction and survival successes determines their selection values. Even though expression of viruses is simpler than expression of multicellular organisms, its complexity is still much too high for making accurate calculations, but experimental data give us valuable insight.

Probably the best understood virus is the levivirus Qβ [11]. It contains three major cistrons and at least one minor one, caused by the occasional read through a stop codon. The viral RNA folds into a highly compact structure. Much is known about the secondary structure of the RNA [4]. Folding and refolding of the secondary and tertiary structure provides a highly sophisticated regulation of the gene translation resulting in a 'phage clock': the linear nucleotide sequence precisely determines a cascade of chemical reactions leading eventually to reproduction. The sophistication is so high that hardly any nucleotide position can be replaced without consequences for the selection value. The proof is the experiment described above: even mutants that appear to be neutral since they form normal plaques eventually return to a reproducible wild type after sufficient time for selection to do its work.

Arguments that much of the sequence is not important for fitness because they neither change the amino acid sequence nor the secondary structure [62]


10 20 30 40 50 60 70 80
wt GGGUUCAUAGCCUAUUCGGCUUUUAAAGGACCUUUUUCCCUCGCGUAGCUAGCUACGCGAGGUGACCCCCCGAAGGGGGGUGCCCCA d_{iw}
m2 UU 2
m3 A UUU 4
m6 G øø UU 5
m13 C UUø 4
m14 G A A UU 5
m25 ÛU G Â UUĈ 7
p12 øø UU 4
p20 Ĝ C 2
p21 Ĝ UU 3
p22 AU 2
n8 AU 2
n14 ÛU ø G Â Û Ĉ 7
n19 A U Û 2
n34 ø AUĈ 4
x3 Ĝ ÂÛ C U Û 5
x4 ø C ø 3
x6 C Â C UAĈ 6
x9 Â G C * 4
x14 C CA C ø 4
x16 C Â C Û ø 5
x23 ø Û C 3
x46 C C AC C 6
GGGUUCAUAGCCUAUUCGGCUUUUAAAGGACCUUUUUCCCUCGCGUAGCUAGCUACGCGAGGUGACCCCCCGAAGGGGGGUGCCCCA
10 20 30 40 50 60 70 80
*: duplication of GCGAGGU

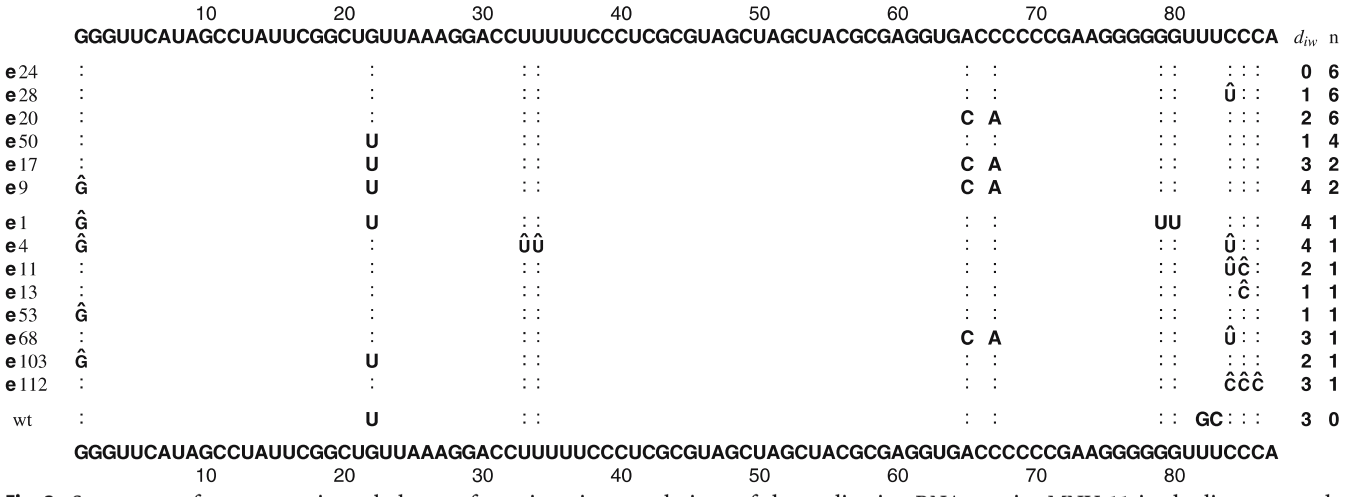


Fig. 2 Sequences of representative subclones of quasispecies populations of the replicating RNA species MNV-11 in the linear growth phase (top) and the exponential growth phase (bottom) [101]. Â: A inserted, ø: nucleotide deleted, n: number of clones.

are thus not in agreement with the experimental data. If these arguments were true, passaging of the virus would lead to random drift of the sequence [66] and not to the reselection of a wild type mutant distribution. Expression involves much more than translation and secondary structure:

1. The tertiary structure includes a large repertoire of intramolecular nucleotide interactions in addition to those directing the secondary structure [19];
2. Site-directed mutagenesis experiments [95, 94, 67, 118] showed that structural dynamics of the RNA folding is of high importance for the expression of the virus: during replication and translation, extensive refolding of the genome is crucial;
3. Plus and minus strands have different roles in expression; nevertheless, there are structural constraints for both strands (See the chapter by E. Domingo et al., this volume, for further discussions of possible multiple cis-acting and trans-acting interactions among components of the mutant spectra of RNA viruses).

Work with replicating RNA has corroborated this argument. As mentioned, expression of short replicating RNA species is reduced to a fraction of the expression of a viral RNA: translation and its regulation is irrelevant and only replication by the extraneously added replicase matters. Nevertheless, defined sequences of the various replicating RNA species are obtained. Even in this model system, extensive amplification of a mutant, no matter whether isolated from a natural population or generated artificially, results in the ultimate reselection of the optimal wild type mutant distribution.

The mutant spectrum of the replicating species MNV-11 has been studied in greater detail after extensive amplification to obtain mutant distributions in the quasispecies equilibrium [101] by retrotranscribing and cloning on the DNA level RNA strands picked at random from the RNA population. When the RNA population was grown in the exponential growth phase (where replicase is present in excess of RNA), the overall replication rate decided the selection and the population was dominated by three mutants which were equally represented. When RNA template was in excess of replicase, the mutant spectrum was dominated by a single master sequence which was identical to the sequence of the average population (Fig. 2). Under these conditions, not only the replication rate, but also the competition of RNA for enzyme decides the selection success. Even small changes of the external conditions altered the spectrum [101].

7
The Quasispecies

We can also build a population landscape by assigning to the vertical dimension the type frequency for each possible genotype in sequence space. As discussed earlier, type frequency values can be easily measured and mutation probabilities and selection values are normally calculated from measuring the temporal change of the population landscape. Relative population values can be determined by picking members of the population at random and ordering them by sequence, as in Benzer's experiments. We find that they are closely grouped together; as in a landscape on earth, they form more or less coherent mountain regions containing several peaks. How do the population landscape and the fitness landscape compare?

- Obviously, the two landscapes are not congruent! Therefore, Darwin's theory is not the trivial tautology 'survival of the survivor'. In model systems like RNA replication, it was not necessary to derive fitness values by studying the dynamics of the evolutionary process itself, but they could be calculated from physical and chemical parameters.
- The population landscape is poorly correlated with mutation rates, and it is not possible to derive mutation rates from type frequencies. The population dynamics is controlled by selection *and* mutation and two of the parameters must be known to calculate the third.
- The shape and extension of the population landscape is strongly dependent on population size, while the fitness landscape is independent of population size.
- Fitness landscapes are 'rugged': since many single mutations are lethal, many steps lead precipitously to zero fitness. This does not inhibit evolution severely: the more dimensions there are, the higher the probability of finding one with substantial fitness is.
- Population landscapes are less rugged because the mutation term smoothes them; there is no nonpopulated locus in the immediate neighborhood of a highly populated one.
- The population space does not monotonously decrease with increasing Hamming distance to the master sequence. As in a natural landscape on earth, hills with smaller elevations group around the highest peak. The fitness landscape is also not monotonous: two deleterious mutations can compensate each other, e.g. by restoring a base pair in a stem region or by correcting a frame shift.

– Hills in the population landscape comprise individuals of close kinship. The population is thus divided into subpopulations that compete one with another.

– The highest mountain peak in the fitness landscape is the master *m*. It is not necessarily the highest peak in the population landscape because of the contribution of the mutational gain term.

If two or more species are present, a corresponding number of isolated mountains are found. This property allows us to extend the notion of the species barrier, defined in Mendelian populations with unsuccessful interbreeding, to non-Mendelian populations: coherent regions in the population landscape, sufficiently far apart one from another in sequence space to make mutational jumps from one region to the other extremely unlikely, define different species.

Another suitable metaphor for the population landscape is a cloud, denser in the interior, fuzzy and fluctuating at the periphery. The periphery is the arena of evolutionary progress: a newly found fitness maximum can lead to a shift in the cloud to the new optimum or to a bifocal cloud. When the centres of the latter float apart in sequence space to their new fitness optima, the cloud eventually is separated by a new species barrier: a new species has been formed. Peaks in the fitness landscape far outside the species barrier are without consequences: they cannot be reached by a mutational event.

This formation mechanism for new species, first proposed by Darwin, makes it inevitable that the fitness landscape remains largely unexplored. This is illustrated by the fact that the sequences of ribosomal [124, 70] and transfer RNA [38] still show evidence that they departed from the progenote, the ancestor cell. Crick [22] argued that in such central features of the expression as translation, a once designed optimum could not be changed without destroying most of the accumulated information. It cannot be known whether the solutions found were the only or at least the optimal ones, but from our biological knowledge this seems unlikely: the fixed solution is probably merely a 'frozen accident'. The opposite phenomenon is also observed. The coliphages of the genus Leviviridae can be grouped into two main groups. Their structure and organization are so similar that it seems inconceivable that they did not have a common phylogenetic origin, but their sequences do not reflect any relationship. Apparently, a wide-spread flat fitness landscape allowed the sequences to drift widely apart. It is quite clear that a particularly high heterogeneity does not allow the conclusion on a particularly high mutation rate; often it is rather a measure of how much variability the expression can tolerate.

Under realistic conditions, the population cloud in the sequence space is in constant movement due to changes in the environmental conditions (See the chapter by C.O. Wilke et al., this volume). This rapid adaptation does not have to wait for a rare advantageous mutation to happen. Instead, the mutants best adapted to the environment are rapidly enriched by selection and shift the center of the cloud to another position. Instead of creating advantageous mutants from the master sequence, adaptation starts with the selection of hills that are closest to the new optimum.

When conditions can be kept constant, the population eventually reaches its optimal composition: a steady state is reached ($d\tilde{x}_i = 0$) where all mutants retain a fixed proportion of the population $\tilde{x}_i$ because the selection term and the mutational gain balance one another. As seen with the sequence determination of viruses, this steady state is highly reproducible. No matter at which point of the sequence space (within the species barrier) the evolution process starts, one ends up with the same steady state distribution. This 'reference' population has reproducible properties and a defined sequence and therefore resembles a species. It is called quasispecies. Its wild type is not a predominant type, but rather the gravity centre of the steady state mountain in the population landscape. It may, but does not need to, coincide with the master type.

The steady state concentration of a type can be approximated by neglecting the mutational gain term [32] to

$$\tilde{x}_i = \frac{W_{ii} - \tilde{E}_{k\neq i}}{\tilde{E}_i - \tilde{E}_{k\neq i}} \tag{10}$$

This approximation holds only when type i belongs to a master-dominated population, where the fitness of the master far exceeds any other fitness value. However, this conditions is not fulfilled in real mutant distributions, which contains mutants in a continuous fitness range from lethal to almost neutral mutations.

Analysis requires an exact solution of the system of differential equations [Eq. (6)], which despite their inherent nonlinearity is possible [115, 64, 39]. Its maximum eigenvalue (which is the only stable one [39]) describes the quasispecies distribution.

Of course, evolution does not know a permanent steady state. A change of external conditions changes the wild type. Furthermore, a new fitness hill may be hit by chance in the still unexplored part of the sequence space. The steady state is then replaced by a new rapid evolution process until a new steady state is found. Evolution thus proceeds via 'punctuated equilibria' [49].

What is the target of 'Selection of the fittest'? The type, the clan or the (quasi)species? This depends on what is investigated. If selection among

species is investigated, not the fittest type, but the whole wide-spread quasispecies distribution is the target of selection. Within the quasispecies distribution, the clans represented by hills in the population space compete one with another. Within the clan, the types compete for reaching their maximum representation. (Measurements and implications of fitness variations of viruses are covered in the chapter by M.E. Quiñones-Mateu, E.J. Arts, this volume.)

8 The Error Threshold

Balancing of the focussing force of selection and the diversifying force of mutation is only possible within certain bounds: the latter must not exceed a certain threshold for maintaining the information content of the quasispecies. According to Eq. (10), a master sequence should disappear whenever its selection rate value W_{mm} becomes equal to or smaller than $\bar{E}_{k\neq m}$. Under constant external conditions, W_{mm} is a constant, but in nature external conditions usually vary. Moreover, in addition to the kinetic parameters A_m and E_m, W_{mm} depends on the fidelity probability Q_{mm}, which itself depends on ν_m, the chain length of the sequence to be reproduced. The larger ν_m, at given reproduction mechanisms, the smaller $Q_{mm} = \bar{q}^{\nu_m}$ (with $\bar{q} < 1$). If the average superiority of the master over its competitors is introduced as $\bar{\sigma}_m = A_m/(D_m + \bar{E}_{k\neq m})$, Eq. (10) can be written as $\tilde{x}_m = (\bar{\sigma}_m \bar{q}^{\nu_m} - 1)/(\bar{\sigma}_m - 1)$, and the error threshold relation is obtained [32]:

$$1 - \bar{q}_m \leq \frac{\ln \bar{\sigma}_m}{\nu_m} \,. \tag{11}$$

Above the critical error rate given by Eq. (11), the master type and with it the information contained in the sequence will be lost. To maintain the information, the mutation loss $1 - Q_{mm}$ must be compensated by at least an average superiority $\bar{\sigma}_m > 1$. Though Eq. (10) is an approximation, relation (11) generally holds surprisingly well, even for wide-spread quasispecies distributions.

In all studied cases with single-stranded RNA viruses, the critical product $\nu_m(1 - q_m)$ turned out to be in the vicinity of 1. Therefore, $\bar{\sigma}_m$ must clearly exceed 1 to obtain a natural logarithm $\ln \bar{\sigma}_m$ that is not too far from 1. In contrast, $\bar{\sigma}_m$-values in Mendelian species are near unity, because large parts of the genome do not contribute to the immediate survival under the pertaining conditions. For these species, the average error rate per sequence i, ε_i, is very small and the term $\ln \bar{\sigma}_m$ can be approximated by ε. For the cases studied, ε turned out to be between 10^{-2} and 10^{-3}.

The disintegration of information at the error threshold behaves like a first-order phase transition [113, 114, 34]. In a numerical simulation carried out by Swetina and Schuster [113], where all mutants were assumed to have a uniform fitness of one tenth of the master, the relative population of the master $\tilde{x}_m$ dropped with increasing error rate to quite small values, until, at the error threshold, an instability occurred and all types in the population, *including the master*, became equally populated. While a defined wild type sequence could be determined below the error threshold, it suddenly disappeared at the error threshold.

The chosen example of a single master sequence surrounded by uniformly less fit mutants is quite instructive, but unrealistic in nature. In reality, much of the sequence space cannot be accessed because nonviable mutants do not produce progeny. Hence, at the error threshold, the mutant distribution cannot evaporate into the whole sequence space. Instead, the cloud spreads out so much into the lethal area that fewer and fewer progeny are produced, resulting in the eventual annihilation of the population. This has been shown to happen with viral populations after the error rate was artificially raised by the addition of mutagenic drugs [80, 109, 23, 35, 50] (see E. Domingo et al, this volume, for discussion of error catastrophe as an antiviral strategy).

9
Evolutionary Biotechnology

9.1
Principles

A new research field of increasing impact is evolutionary biotechnology. It is an application of the Darwinian principles of generation of diversity and selection with the aim of deriving novel molecules with desired properties [65]. The latter are improved by further cycles of amplification and selection (Fig. 3), whereby natural selection is replaced by artificial selection.

Two general strategies can be chosen: rational and irrational design [15, 116, 92]. Rational design is a well-tested method that is suitable for well-understood systems. It is frequently used to modify protein properties [75, 44, 102] by site-directed, not random mutations. It is evolutionary in the sense that it involves evaluation and selection of successful mutants. Exploration of fitness is restricted to the mountain itself and its close surroundings.

The irrational design is a gunshot method: we try to obtain success by playing a lottery game. The advantage of this method is that it does not depend on prior knowledge of how to achieve a desired function. The whole

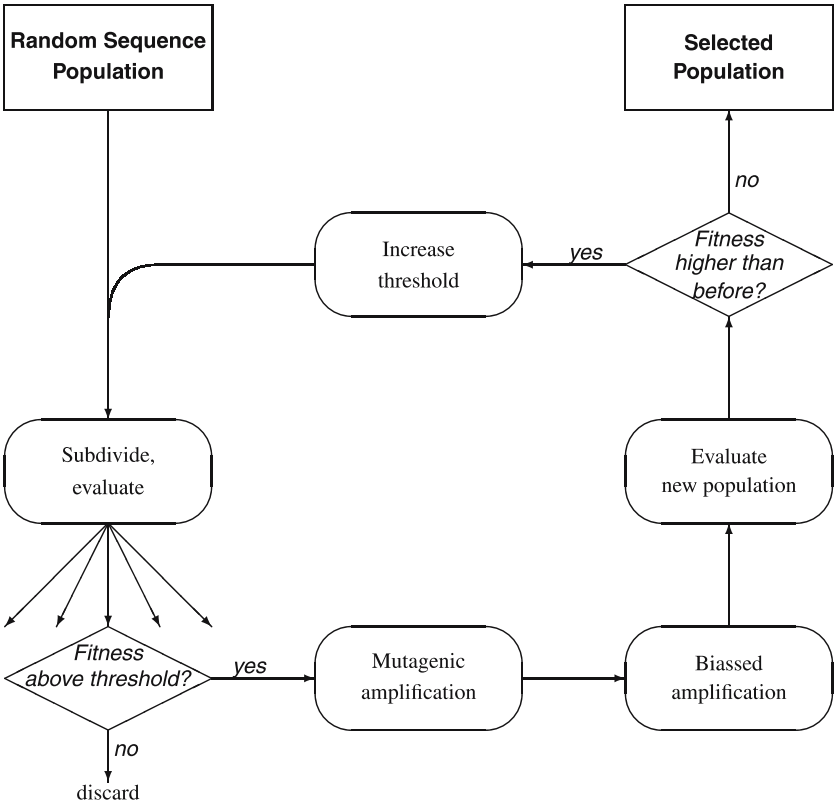


Fig. 3 General flow diagram for evolution experiments using artificial selection by a biassed process.

sequence space is used for exploration. This approach is very old: nearly all information on food and drugs accumulated in the history of mankind has been gained using this method, even though essentially by serendipity. A systematic exploration requires high-throughput methods and the invention of efficient screening methods. The extension to chemically synthesized polymers composed by suitable nonbiotic chemical residues (combinatorial chemistry [18, 77, 83]) is also new, as is the application of evolution strategies to technical processes [99]. Unfortunately, the expectation of success drops sharply with increasing complexity. Once a positive result has been obtained, the neighbourhood of its locus in sequence space is explored.

For most applications, both strategies have severe limitations. A mixed approach is better: shotgunning is restricted to areas in sequence space with