Current Topics in Microbiology and Immunology

Esteban Domingo
Peter Schuster
Santiago F. Elena
Celia Perales *Editors*

Viral Fitness and Evolution

Population Dynamics and Adaptive Mechanisms

Springer

Current Topics in Microbiology and Immunology

Volume 439

Series Editors

Rafi Ahmed, School of Medicine, Rollins Research Center, Emory University, Atlanta, GA, USA

Shizuo Akira, Immunology Frontier Research Center, Osaka University, Suita, Osaka, Japan

Arturo Casadevall, W. Harry Feinstone Department of Molecular Microbiology and Immunology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

Jorge E. Galan, Boyer Center for Molecular Medicine, School of Medicine, Yale University, New Haven, CT, USA

Adolfo Garcia-Sastre, Department of Microbiology, Icahn School of Medicine at Mount Sinai, New York, NY, USA

Bernard Malissen, Parc Scientifique de Luminy, Centre d'Immunologie de Marseille-Luminy, Marseille, France

Rino Rappuoli, GSK Vaccines, Siena, Italy

The reviews series *Current Topics in Microbiology and Immunology* publishes cutting-edge syntheses of the latest advances in molecular immunology, medical microbiology, virology and biotechnology. Each volume of the series highlights a selected timely topic, is curated by a dedicated expert in the respective field, and contains a wealth of information on the featured subject by combining fundamental knowledge with latest research results in a unique manner.

2020 Impact Factor: 4.291, 5-Year Impact Factor: 5.110
2020 Eigenfactor Score: 0.00667, Article Influence Score: 1.480
2020 Cite Score: 7.7, h5-Index: 38

Esteban Domingo · Peter Schuster ·
Santiago F. Elena · Celia Perales
Editors

Viral Fitness and Evolution

Population Dynamics and Adaptive Mechanisms

Springer

Editors
Esteban Domingo
(CSIC-UAM)
Centro de Biología Molecular Severo Ochoa
Madrid, Spain

Santiago F. Elena
(CSIC-UV)
Instituto de Biología Integrativa de Sistemas
Valencia, Spain

Peter Schuster
Institut für Theoretische Chemie
Universität Wien
Vienna, Austria

Celia Perales
(CSIC)
Centro Nacional de Biotecnología
Madrid, Spain

ISSN 0070-217X ISSN 2196-9965 (electronic)
Current Topics in Microbiology and Immunology
ISBN 978-3-031-15639-7 ISBN 978-3-031-15640-3 (eBook)
https://doi.org/10.1007/978-3-031-15640-3

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, 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.
The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Foreword

Over the last years, I used to deliver a lecture on viral transmission within a course on viral dynamics at Universitat Pompeu Fabra in Barcelona. It was a review of standard approaches to virus complexity with a special focus on the modelling of how epidemic thresholds emerge from simple toy descriptions of interactions between viruses and their hosts. As with other standard lectures on the topic, some relevant examples were discussed. Measles, hepatitis C virus, HIV or polio and their particular idiosyncrasies and historical impact on our societies, as well as the major success of vaccination with the eradication of smallpox as an outstanding victory of science. I also discussed the concept of emerging viruses and the threat posed by novel infectious diseases associated with them. Along with Hanta, Lassa or the more recent Zika virus, the SARS-related coronavirus case study came easily as a reminder that new viruses are expected to emerge any time. I used to discuss the 2002 outbreak that took place in Southern China which caused alarm due to its rapid spread (cases were reported on all continents) and high mortality. The virus was eventually contained thanks to a rapid public health response and its zoonotic origins established. We could say that we were safe: It seemed just another event taking place somewhere, far away and just scratching our doors. At the end of the lecture, I also cited several authors warning about the risks associated to our globalised world where transport networks have virtually eliminated geographic boundaries and where humans are putting an enormous pressure on nature. The last slide was a cover of TIMES magazine: "Warning: we are not ready for the next pandemic".

As I write this foreword, the world emerges from 2 years of a devastating pandemic caused by another SARS-CoV strain that emerged at the end of 2019. In a few months, what appeared to be just another controllable outbreak led to a global spread. Millions of people have died from the COVID-19 pandemic and the confinement measures required to control further spread had to be taken at a big economic and social cost. It has been estimated that the total mass of viruses involved could fit within a soda can. A reminder of the power of these tiny molecular parasites, able to damage our civilisation despite their apparent simplicity.

Along with the political and public health measures used to limit the spread, a historical research effort was deployed towards understanding the virus, from its

genome sequence and molecular structure to the underlying strategies to find its way into tissues and organs. As a result, a whole picture of this new threat emerged in a record time, along with the first vaccines that marked the start of its control (but perhaps not its demise). The success of this scientific challenge was largely due to a combination of economic expenditure and the accumulated knowledge of past studies in virology. Despite the uncertainties created by the new virus, it was only one member of the virosphere, i.e. the huge and always expanding universe of viruses. It was known that they were equipped with RNA genomes, exhibiting high mutation rates (due to their error-prone polymerases) and able to generate a highly heterogeneous population. Moreover, theoretical works initiated in the 1970s on the Darwinian evolution of viruses also offered a rationale to understand the impact of their heterogeneity on adaptation. As discussed by different contributors of this book, viral populations (including both animal and plant viruses) are *quasispecies*: there is no single genome responsible for adaptation, but a whole cloud of sequences connected via mutations. And these populations evolve and spread within space and time but also across their astronomically vast fitness landscapes. Viruses are complex adaptive systems, always evolving and adapting to their hosts and sometimes causing major trouble.

Over the last decades, the combination of novel techniques of deep sequencing along with improved models that can be compared with massive amounts of data is sharpening our picture of virus dynamics. These advances have shown that it is possible to describe their landscapes and the development of new antiviral agents. This is not an easy task, since coevolutionary forces are always at work as mutation (but also recombination, as illustrated by Geminiviruses) can produce novelties responsible for new epidemic events. As pointed out by Domingo and co-workers, viruses typically live in a non-equilibrium state. This makes difficult (but not impossible) to search and develop new resistance alleles to help creating reliable crops or predict clinical outcomes. The COVID-19 pandemic has been the most recent scenario to use the many and complementary approximations covered here by means of both general principles and well-defined case studies.

The emergent picture that results from the interdisciplinary nature of virology is a promising one, where ecological and evolutionary components of adaptation (displaying similar time scales when dealing with RNA viruses) are both integrated into a *phylodynamics* perspective. Similarly, the study of plant-virus adaptation suggests that a proper integration will be achieved by means of a multilayer network, where multiple scales (from cells and tissues to individuals and communities) would connect key concepts associated to replication modes, bottlenecks and path-dependent propagation events. Understanding viruses, as illustrated by the contributions collected here, will always need to combine the idiosyncrasies of each case study (there is plenty of room for variation, and no other entity illustrates this

so well) and general principles. There is a long and winding road to map the virosphere, from its molecular possibilities and evolutionary trees to its mathematical properties. Much work is still needed, but progress is fast and the whole picture is rapidly improving. Hopefully, we might not yet be able to predict the next pandemic, but we are getting more aware and well prepared.

Barcelona, Spain Ricard Solé

Contents

Virus Evolution on Fitness Landscapes

Peter Schuster and Peter F. Stadler

Abstract The landscape paradigm is revisited in the light of evolution in simple systems. A brief overview of different classes of fitness landscapes is followed by a more detailed discussion of the RNA model, which is currently the only evolutionary model that allows for a comprehensive molecular analysis of a fitness landscape. Neutral networks of genotypes are indispensable for the success of evolution. Important insights into the evolutionary mechanism are gained by considering the topology of sequence and shape spaces. The dynamic concept of molecular quasispecies is viewed in the light of the landscape paradigm. The distribution of fitness values in state space is mirrored by the population structures of mutant distributions. Two classes of thresholds for replication error or mutations are important: (i) the—conventional—genotypic error threshold, which separates ordered replication from random drift on neutral networks, and (ii) a phenotypic error threshold above which the molecular phenotype is lost. Empirical landscapes are reviewed and finally, the implications of the landscape concept for virus evolution are discussed.

In: **Viral Fitness and Evolution—Population Dynamics and Adaptive Mechanisms**
Esteban Domingo, Peter Schuster, Santiago F. Elena and Celia Perales, Eds.
Current Topics in Microbiology and Immunology, Volume xxx
Springer Nature Switzerland AG, Cham, CH 2021

P. Schuster (✉)
Institut für Theoretische Chemie der Universität Wien, Währingerstraße 17, 1090 Wien, Austria
e-mail: pks@tbi.univie.ac.at

P. F. Stadler
Institut für Informatik der Universität Leipzig, Härtelstraße 16-18, 04107 Leipzig, Germany
e-mail: stadler@bioinf.uni-leipzig.de; pfs@santafe.edu

The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe NM 87501, USA

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

E. Domingo et al. (eds.), *Viral Fitness and Evoluation*, Current Topics in Microbiology and Immunology 439, https://doi.org/10.1007/978-3-031-15640-3_1

1 Introduction

Evolutionary genetics in the twentieth century was shaped by the life-long controversy between Ronald Aylmer Fisher and Sewall Wright on the mechanism of evolution (Provine 1992) that did not end before both opponents had died. After Fisher's death in 1962, Wright continued to put forward his view of evolution in many papers and in his extensive four-volume treatise of population genetics (Wright 1968, 1969, 1977, 1978). In a nutshell:

(i) Fisher wanted to construct a theory of evolution similar to thermodynamics (Crow 2002; Ewens and Lessard 2015), whereas Wright favored the uniqueness of biology. Central to Fisher's thinking was a reference state with independent genes, which were optimized each for themselves. Epistasis and pleiotropy, he thought, play the role of a perturbation (see Sect. 2.3).
(ii) Wright's view was the opposite in the sense that he emphasized the importance of gene interactions and thought that the successes of natural selection as well as animal and plant breeding result from optimizing the entire system of interacting genes rather than single genes.

A consequence of the different views is that Fisher considered evolution to be the most effective in large populations, whereas Wright was thinking about partially isolated subgroups, sufficiently small to make random drift efficient and to allow for fixation but large enough to prevent degeneration and extinction. In addition, Fisher like most of his contemporaries involved in the development of the synthetic theory of evolution was an extreme *gradualist* in the sense that he insisted that all evolutionary change is brought about by the succession of a large number of very small steps (see Fig. 22 in Sect. 4.1).

The concept of an *adaptive fitness landscape* is commonly attributed to Wright (1931, 1932, 1988) who introduced it as a metaphor underlying the illustration of adaptive evolution as a hill-climb or *adaptive walk* on a multi-peak fitness (hyper) surface (Fig. 1).[1] According to McCoy (1979) the idea of evolution as an adaptive process on a fitness landscape has been used for the first time by Janet (1895) in order to provide an explanation for the lack of intermediate forms of species in the fossil record much earlier than Wright's seminal publication (Wright 1932). Although there is a certain similarity between the thoughts of Janet and Wright there are also fundamental differences. Janet, for example, when searching for an explanation of the occurrence of gaps in the fossil record, assumed that species located in valleys are pulled down by selection similar to gravitation acting on bodies in reality. He thought

[1] The expression *hypersurface* points at the fact that fitness landscapes are surfaces in high-dimensional space. The notion *hypersurface* was introduced first in quantum molecular sciences where the motions of a molecule have commonly more than three degrees of freedom. Since we shall be dealing here almost exclusively with such high-dimensional objects we drop from now on the prefix 'hyper'.

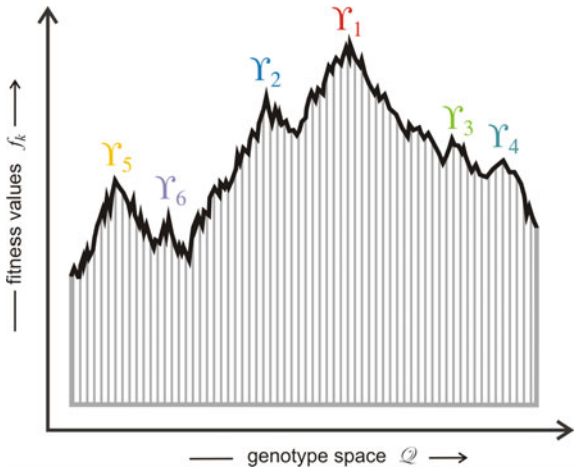

Fig. 1 Sketch of Sewall Wright's fitness landscape. The figure sketches one-dimensional fitness landscapes where individual species Υ_j occupy local fitness maxima. Species are ordered by fitness: $f_1 \geq f_2 \geq f_3 \ldots$. Under constant environmental conditions evolution approaches the global optimum corresponding to the fittest species Υ_1. In low-dimensional spaces substantial fitness drawdowns are required in order to cross valleys. Wright (1932) himself points out that real fitness landscapes are high-dimensional and sketches with low-dimensional supports, in particular one-dimensional ones, may be misleading

that populations occupy areas corresponding to the variability so that strong selection and small areas correspond to narrow valleys and populations with little variability whereas more variable species would be represented by broader valleys and larger areas. Dietrich and Skipper (2012) argue that '*Wright's discovery of the adaptive landscape was independent of Janet's, and certainly more influential historically.*'

The fitness landscape in the sense of Sewall Wright is a mapping of the space of genotypes called *genotype* or *sequence space*[2] onto non-negative real numbers representing fitness parameters, which enter the deterministic or stochastic dynamical systems describing the evolution of populations (Fig. 1). Wright's illustration visualizes a genotype space with several alleles per locus. Genotype space then comprises all possible allele combinations and for a few hundred genes with two alleles each, the number of combinations is already hyper-astronomical—four hundred genes lead to $2^{400} = 2.58 \times 10^{120}$ different genotypes. As sketched in Fig. 1 species are situated on local maxima or peaks of the landscape. Evolution progresses through migration from peak to peak, Wright's original fitness landscape contains many local fitness maxima and minima, and migrating populations have to pass lower terrain crossing intermediate saddle points or valleys (Wright 1932, pp. 358 and 361). Optimization through natural selection on fitness landscapes implies stepwise or steady increase and encounters a problem, since naïve adaptive walks in the Darwinian sense cannot

[2] The genotype space $\mathcal{Q}$ will be called *sequence space* when we understand genotypes as RNA or DNA sequences here.

take downward steps. Wright accounts for this obvious problem with his *shifting balance* model of evolution, which combines Darwinian selection and genetic drift (Masel 2011). The model consists of three phases: (i) random genetic drift splitting the global population into subpopulations, (ii) selection within subpopulations, and (iii) selection between subpopulations. The mean fitness of the population is assumed to decrease during phase (i) and to increase during the phases (ii) and (iii). Wright's original fitness landscape is mapped upon a 2D model support of genotype space that contains several local maxima, saddle points, and valleys. Landscapes on low-dimensional supports are misleading as a simple example demonstrates: On 1D-landscapes all valleys have to be crossed at the bottom, because there are no saddle points, which exist already on 2D supports and it becomes possible to migrate from one peak to another without passing the lowest points. Generic landscapes on high-dimensional supports provide a multiplicity of paths leading from peak to peak along ridges and passing high-dimensional saddle points, crossing valleys at the bottom is no longer necessary (Fragata et al. 2019). Apart from high-dimensionality it is the existence of neutral networks (Grüner et al. 1996a, b; Reidys et al. 1997; Schuster et al. 1994), that enables adaptive walks on fitness landscapes (see also sections 2.5 and 3.6). Despite the apparent success in popularity for the illustration of Darwinian selection Wright's metaphor was facing objection and has been heavily debated in the following years (see, e.g., Provine 1986; Ruse 1996) and for a more recent well-founded analysis of Wright's landscape concept we recommend (Skipper 2004, 2012)). Several authors find it questionable whether or not Wright's fitness landscape is more than a didactically useful metaphor (Ruse 1996) and indeed, Wright developed his adaptive landscape metaphor and its diagrams as a way to translate his shifting balance theory from the mathematics of population genetics to a more accessible idiom for general biologists.

The majority of studies dealing with fitness landscapes assume time-independent fitness values. This may be questionable because evolution takes place in ecosystems and the adaptation of one species provides changes in the environment for all other species. Stronger coupling and consequential time dependence are caused by the coevolution of species. In mathematical terms time dependent landscapes can be described in different ways: (i) by static landscapes, which are externally driven, (ii) by spatially extended dynamical systems, or (iii) by coevolutionary models (Richter and Engelbrecht 2014). Here we shall be concerned with static landscapes only.

2 Landscape Models and Landscape Features

A complete and comprehensive representation of the fitness landscape with all important properties is out of reach at present, even for the smallest organisms and for viruses, and even in case of evolution in vitro (Joyce 2007). The maximal achievable population sizes are in the range of $N = 10^{15}$ molecules and they are much too small to cover all possible variants. A practicable escape of this dilemma is the concep-

tion of models, which focus on a given aspect of evolution on landscapes. Therefore this section is dealing with different models and the properties they can describe or explain.

2.1 Classes of Landscapes

Adaptive landscapes for evolution are easier to discuss and analyze when three classes are distinguished: (i) genetic, (ii) phenotypic, and (iii) molecular landscapes (Dietrich and Skipper 2012). Sewall Wrights landscape is a *genetic landscape*: fitness is assigned to genotypes, the collection of all possible genotypes together with a distance measure between pairs of genotypes constitutes the genotype space, and the fitness landscape is the result from plotting fitness on genotype space. Already in his first paper Wright (1932) makes clear that one fundamental problem is the dimensionality of the carrier upon which the landscape is built. Wright illustrates the problem by means of a fictive mini-genome of five loci with two alleles each giving rise to $2^5 = 32$ different haploid combinations and $3^5 = 243$ diploid genotypes. Viruses encode for one up to 2500 protein genes, prokaryotes from 182 up to 9400, and eukaryotes from 20 000 up to about 100 000 (Milo and Phillips 2016, p. 286). A second complication is the complexity of the relations between genotypes, phenotypes, and fitness, which involve a variety of still unsolved problems like protein folding, the development of multicellular organisms as well as the prediction of biological function from molecular structure.

The existence of a *genetic landscape* as a useful tool in modeling evolution is based on a number of assumptions, for example

(i) constant environments are required, since the unfolding of the phenotype as well as its fitness depends on environmental factors,
(ii) fitness is a property of the carrier of the genotype alone and does not depend on other individuals in the population,
(iii) random events during unfolding of the genotype, which exert influence on phenotypes and/or fitness values, are excluded, and
(iv) epigenetic effects are negligible.

In reality assumptions (i) to (iv) are restrictive and not fulfilled in general.

In *phenotypic landscapes* the first part of the highly complex unfolding of genotypes is skipped and evolutionary relevant properties like fitness are assigned directly to phenotypes. The assignment of fitness to phenotypes is certainly less difficult and easier to interpret than the relation to genotypes. The first definition and application of phenotypic landscapes is due to the paleontologist Simpson Simpson (1944, 1953). Among other things the phenotypic landscape has been applied to explain the evolution of horses. The phenotypic landscape became popular among paleontologists and received a mathematical model through the works of Lande (1976, 1979). A detailed description of phenotypic landscapes is found in Rice (2012).

The fast development of molecular biology in the second half of the twentieth century opened a new avenue to understanding, analyzing, and handling of fitness landscapes. The resulting model is often characterized as the *molecular landscapes* paradigm: Fitness is a function of genotypes through the expression in molecular structures and according to structural biology this allocation is based on two mappings

$$\begin{array}{ccccc} \text{genotype} & \Longrightarrow & \text{phenotype} & \Longrightarrow & \text{fitness}\,, \\ X_k & \overset{\Psi}{\Longrightarrow} & S_k & \overset{\Phi}{\Longrightarrow} & f_k\,, \\ \text{sequence} & \Longrightarrow & \text{structure} & \Longrightarrow & \text{function}\,, \end{array} \qquad (1)$$

$$S_k = \Psi(X_k) \qquad f_k = \Phi(S_k)\,.$$

In other words, the structure S_k—understood as the phenotype—is a function of the genotype X_k, which is the nucleotide sequence of a DNA or an RNA molecule, and fitness f_k is a function of the molecular structure. In strict mathematical sense both mappings are considered to be unique in the forward direction: A given genotype X_k determines the structure S_k and the structure in turn determines the fitness f_k of the genotype's carrier. The first mapping, $S_k = \Psi(X_k)$, is the genotype-phenotype (GP) map and the second map, $f_k = \Phi(S_k)$, is called the structure-function relation.

The obvious question for the introduction of a molecular fitness landscape concerns the choice of appropriate supports for the functions $\Psi(X)$ and $\Phi(S)$. Maynard Smith (1970) suggested a *protein space* as the basis for modeling protein evolution. The protein space is a point space: Every point represents one particular amino acid sequence and accordingly, for a twenty-letter alphabet, $\kappa = 20$, the number of points in protein space is *hyper-astronomically* large already for small oligopeptides. To give an example: The space of proteins of chain length ℓ is denoted by $\mathcal{Q}_\ell^{(20)}$ and has the *cardinality* $|\mathcal{Q}_\ell^{(20)}| = 20^\ell$. Thus the number of individual sequences of chain length $\ell = 18$ over the amino acid alphabet amounts to $|\mathcal{Q}_{18}^{(20)}| = 20^{18} = 2.62 \times 10^{23}$ and this is more than one-third of Avogadro's number or approximately one hundred times the number of protein molecules per liter at cellular concentration.[3] The protein space is not only huge it is also quite complex as far as the manifestations of evolutionary moves are concerned (Dayhoff et al. 1978).[4] The primary moves occur at the polynucleotide level—DNA or RNA—and because of translation amino acid replacements involve the genetic code and its redundancies. Empirical determination of the probabilities of particular amino acid replacements yields the so-called *point accepted mutation matrices* (PAM_n) (Böckenhauer and Bongartz 2007, pp. 95–97), where n counts the numbers of exchanged amino acid per one hundred amino acid residues—common are PAM_1, PAM_{100}, and PAM_{250}. Margaret Dayhoff's path-breaking contribution to computational biology was to develop methods for the calculation of the PAM-matrices and to relate them to evolutionary time scales.

[3] Milo (2013) provides a value of $2 - 4 \times 10^6$ protein molecules per μm^3 cell volume.

[4] An evolutionary move is the change of a biopolymer sequence as a consequence of a mutation.

Considering evolution at the level of DNA or RNA has the advantage that mutations, in particular point mutations, i.e., exchanges of single nucleotides, are straightforward to handle. The space of polynucleotide sequences is completely determined by the length of the sequences[5] ℓ and the size of the alphabet κ (Fig. 2). For binary sequences, $\kappa = 2$, of constant chain length ℓ the sequence space is a hypercube of dimension $n = \ell$. The binary sequence space, $\mathcal{Q}_\ell^{01}$ of sequences built from the two digits 0 and 1, was introduced by Richard Hamming at Bell Labs (Hamming 1950, 1986) in the context of error handling in communication theory. The binary sequence space is adequate for dealing with all sequences built from two nucleotides, $\mathcal{Q}_\ell^{GC}$ and $\mathcal{Q}_\ell^{AU}$, respectively. The extension to the natural four-letter alphabet, $\mathcal{Q}_\ell^{AUGC}$ with the four nucleotides (A, U (T), G, C), is straightforward although hard to illustrate properly, because projections of the high-dimensional objects onto a 2D plane commonly look rather confusing. In Fig. 2 we sketch the buildup of binary and four-letter sequence spaces: The sequence spaces for longer sequences can be derived by means of a recursive construction: $\mathcal{Q}_\ell^{AUGC} \rightarrow \mathcal{Q}_{\ell+1}^{AUGC}$ (Swetina and Schuster 1982). A sequence space is visualized as a graph where the nodes correspond to individual sequences and the edges connect all pairs of sequences with Hamming distance $d_H = 1$. The Hamming distance counts the number of positions in which two end-to-end aligned sequences of equal length ℓ differ. It induces a metric on sequence space and provides the basis for a formal mathematical analysis of sequence space and shape space topology (see Sect. 4). In Fig. 3 the sequence space is resolved into mutant classes. The class Γ_k comprises all sequences, which are at Hamming distance $d_H = k$ from the reference sequence X_0: $\Gamma_k = \{X_j | d_H(X_0, X_j) = k\}$. In particular, Γ_0 contains only the reference sequence X_0 and $\nu_0 = 1$, Γ_1 all $\nu_1 = (\kappa - 1)\ell$ one-error mutants of X_0, Γ_2 all $\nu_2 = (\kappa - 1)^2\ell(\ell - 1)/2$ two-error mutants, and Γ_k all $\nu_k = (\kappa - 1)^k \binom{\ell}{k}$ mutants with k errors. Binary sequences follow a binomial distribution: $|\Gamma_k| = \nu_k = \binom{\ell}{k}$. There is also a rich literature on this topic (Eigen 1985; Feistel and Ebeling 1982; Fontana and Schuster 1998b; Rechenberg 1973; Reidys et al. 1997; Reidys 1997; Reidys and Stadler 2002; Stadler et al. 2001; Strasser 2010). At the same time the Hamming metric represents the natural distance between genotypes in evolution, because it is well defined as the minimal number of point mutations converting two sequences of equal lengths into each other, and in addition, point mutations are most common in evolution.

Based on the concept of sequence space, mutations at the molecular level can be easily classified and analyzed. The binary sequence space has a symmetry that is missing in the sequence spaces over all other alphabets: The numbers of sequences in class Γ_k, $\nu_k = \binom{\ell}{k}$, are the same as the numbers in class $\Gamma_{\ell-k}$ since $\nu_{\ell-k} = \binom{\ell}{\ell-k} = \binom{\ell}{k}$. In particular, there is only one sequence in class Γ_ℓ and this is the sequence that is complementary to the reference, whereas class Γ_ℓ contains the complementary sequence and $\nu_\ell = (\kappa - 1)^\ell - 1$ other sequences in all sequence spaces with $\kappa \neq 2$. This symmetry is nicely reflected by the appearance of probability density surfaces for structure distances (Fontana et al. 1993a, Fig. 12).

[5] For the sake of simplicity we consider here sequences of the same length ℓ.

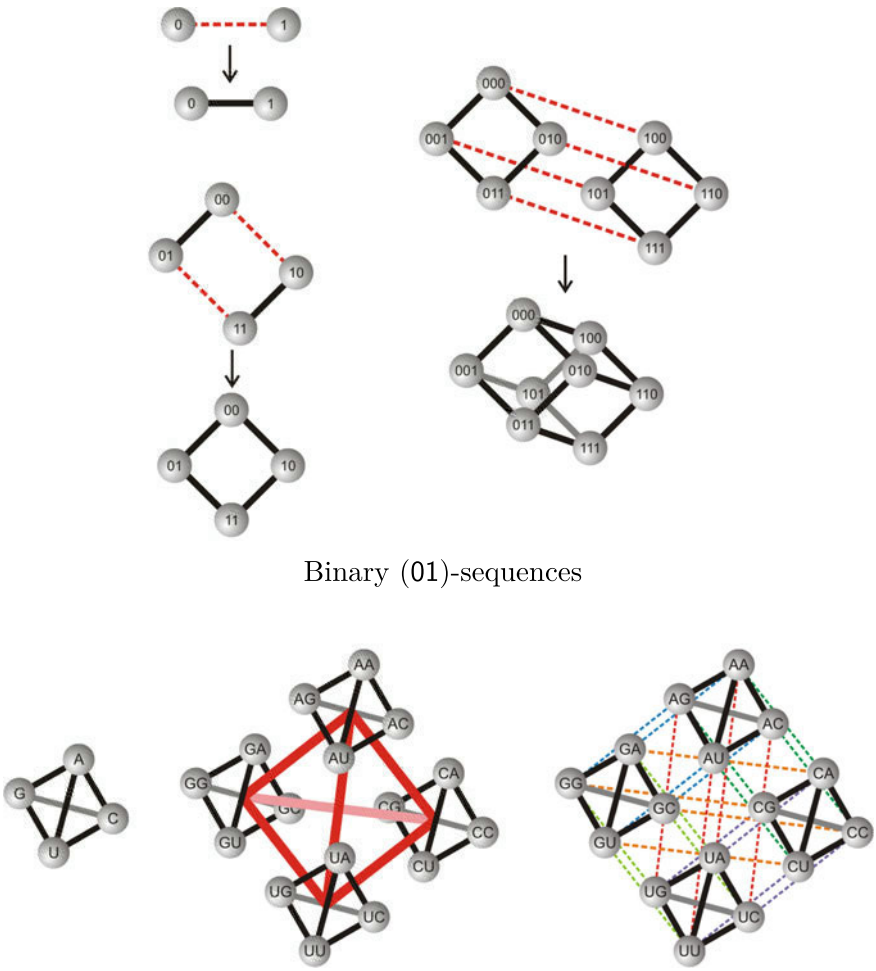

Fig. 2 Buildup of binary (01) and four-letter (AUGC) sequence spaces. The sequence spaces $\mathcal{Q}_{\ell+1}^{\mathcal{A}}$ for strings with chain length $\ell+1$ over an alphabet $\mathcal{A}$ of size $|\mathcal{A}|=\kappa$ is constructed from sequence spaces $\mathcal{Q}_{\ell}^{\mathcal{A}}$ through adding one symbol, either 0 or 1 for binary sequences or A, or U(T) or G or C for natural sequences, on the l.h.s. to the string (see, e.g., Schuster 2009). Joining all pairs of sequences with $d_{\mathrm{H}}=1$ by straight lines yields the sequence space $\mathcal{Q}_{\ell+1}^{\mathcal{A}}$. The upper part of the figure deals with binary sequences, $\mathcal{A}_2=$ 01: The sequence space $\mathcal{Q}_{\ell}^{01}$ is a hypercube of dimension ℓ. The lower part of the figure presents the same construction for the natural nucleotide alphabet, $\mathcal{A}_4=$ AU(T)GC. The single digit element, $\mathcal{Q}_1^{01}$ or $\mathcal{Q}_1^{\mathsf{AUGC}}$, is a straight line and one dimensional for binary sequences or a tetrahedron and three-dimensional for the four-digit alphabet, respectively. The binary sequence space for two digits $\mathcal{Q}_2^{01}$ is a square and two dimensional. For natural (AUGC)-strings of two letters ($\ell=2$) the sequence space, $\mathcal{Q}_2^{\mathsf{AUGC}}$, is a tetrahedron of tetrahedra (middle drawing). This is an object in six-dimensional space that looks quite complicated in the projection onto a (two dimensional) plane (drawing on the r.h.s.)

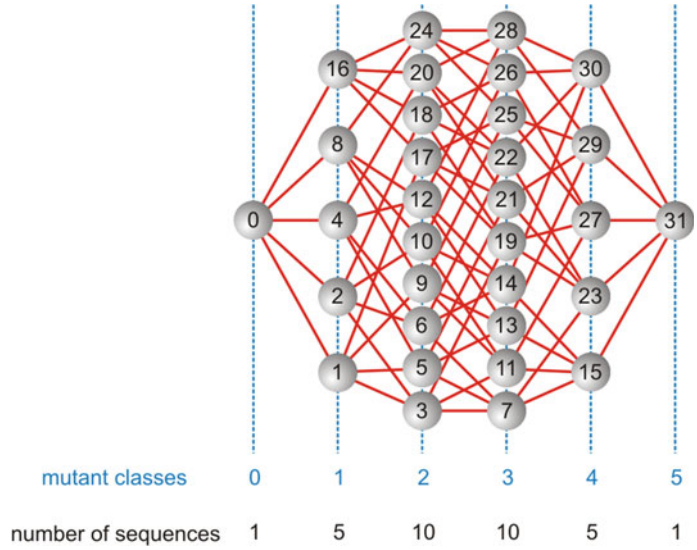

Fig. 3 Sketch of the binary sequence space with $\ell = 5$. The sequence space $\mathcal{Q}_5^{(2)} \equiv \mathcal{Q}_5^{\mathsf{GC}}$ contains 32 sequences, which are indicated here by their equivalent decadic numbers. Nucleotides are assigned, for example $'0' \equiv \mathsf{C}$ and $'1' \equiv \mathsf{G}$: $0 =$ '00000' $\equiv$ 'CCCCC', $1 =$ '00001' $\equiv$ 'CCCCG', $2 =$ '00010' $\equiv$ 'CCCGC', ..., $31 =$ '11111' $\equiv$ 'GGGGG'. Individual sequences are grouped in mutant classes Γ_k that are defined by their Hamming distance to the reference sequence 'CCCCC', $d_{\mathrm{H}}(X_j, X_1) = k$. The numbers of binary sequences in each Γ_k are given by the binomial distribution: $|\Gamma_k| = \nu_k = \binom{\ell}{k}$

The concept of evolution at the polynucleotide level was implemented by Eigen (1971) in his kinetic theory of self-organization of biological macromolecules. Implicitly in this theory it was assumed that evolution takes place by means of moves in a nucleic acid sequence space. These moves are mutations, most frequently point mutations—especially in in vitro and virus evolution—but other changes in the genetic message like deletions, insertions, and genome rearrangements occur as well although with lower frequency. In case the mutation rates are sufficiently high, continued reproduction and mutation lead to a distribution of sequences related by mutation, which may approach stationarity depending on population size, reproduction parameters, and environmental conditions. For now we mention three different scenarios: (i) at small mutation rates and not too large population sizes the populations are homogeneous and consist of single genotypes only, (ii) in the intermediate range mutations occur sufficiently frequently and mutants show up regularly in the genotype distributions, and (iii) at high mutation rates the supply of new mutations is sufficiently large in order to prevent the population from becoming stationary and hence it drifts randomly through sequence space. Evolution in scenario (i) is confined to successive fixation of (advantageous) mutants in the population and all previously selected sequences are lost in the transition to the next variant. Scenario (ii) allows for the simultaneous existence of several genotypes in the population and the genetic reservoir of the population comprises

the whole ensemble. The notion *quasispecies* has been coined for such an ensemble that has reached stationarity (Eigen and Schuster 1977). In Sect. 5 we shall derive an approximate quantitative expression for the mutation rates at which the first transition between the scenarios, (i)→(ii), occurs. The background picture of Eigen's theory is a fitness landscape built upon a genotype space (Fig. 1), which consists of a complete set of polynucleotide sequences. There are many alternative concepts; as an example we mention the fitness-space model (Tsimring et al. 1996) that has been successfully applied to explain the experimental observation of two distinct stages in the growth of viral fitness (Novella et al. 1995, 1999).

2.2 Abstract Rugged Landscapes

The dynamics of evolution, and its efficiency as a search process on the genotype or phenotype space, depends crucially on the structure of landscapes. It is instructive, therefore, to study simple mathematical models of landscapes before delving in more detail into the molecular and empirical landscapes on which viruses evolve. Abstracting equation (1) further, a fitness landscape is simply a function $f(\mathsf{X})$ that assigns a (fitness) value to each object or *configuration* in a (finite) set $\mathcal{Q}$, the *configuration space*. This set $\mathcal{Q}$ is endowed with some structure that expresses *nearness* or *accessibility*. For the moment, this is captured by a *move set*, that defines for $\mathsf{X} \in \mathcal{Q}$ the set $\mathcal{N}(\mathsf{X})$ of neighbors of X. Assuming symmetry of the move set, i.e., we obtain, as above, an undirected graph as our configuration space. A detailed account of the formal aspects can be found in Reidys and Stadler (2002). Fitness landscapes are thus simply functions on the vertex set of a graph.

At this level of abstraction, landscapes are a quite common model in many fields of science:

(i) In physics, spin glasses (Garstecki et al. 1999) comprise sets of N spins $s_i = \pm 1/2$ to which an energy $f(\mathbf{s}) = \sum_{i<j} J_{ij} s_i s_j$ is assigned that models their pairwise interaction. The interaction strength J_{ij} encapsulates details of the model. For instance, if the spins are arranged on a regular lattice, $J_{ij} = 0$ unless i and j located at adjacent lattice positions. The neighbors of spin configuration $\mathbf{s} = (s_1, s_2, \ldots, s_N)$ are obtained by flipping a single spin $s_i \to -s_i$.

(ii) The folding of a given biopolymer (Onuchic et al. 1997; Mallamace et al. 2016) may be understood in terms of an energy landscape $f(\mathbf{x})$ where $\mathbf{x}$ is a (discretized) spatial conformation (Flamm et al. 2000; Wolfinger et al. 2006), e.g., a particular contact matrix of the polypeptide chain or a list of nucleotide base pairs in a nucleic acid. A natural notion of neighborhood in this case involves the open or closing of a base pair or a contact between amino acids. Landscape of this type is used to model the dynamics of biopolymer folding.

(iii) In the setting of computer-aided drug design, $\mathcal{Q}$ is e.g., a set of small molecular ligands and their fitness $f(\mathsf{X})$ models their binding affinity to a given target

protein of interest (Vogt 2018). Neighborhood $\mathcal{N}$ on $\mathcal{Q}$ is defined by chemical similarity or similarity in syntheses of combinatorial and make-on-demand libraries (Saldívar-González et al. 2020).

(iv) Landscapes also serve as abstract models of combinatorial optimization problems with the aim to design general purpose optimization schemes (Gendreau and Potvin 2010) such as genetic algorithms or simulated annealing. Here, the aim is to find neighborhood definitions that make the problem easier to solve.

The neighborhood structure in the configuration space provides us with a notion of locality. A *local optimum*, then, is simply a configuration $\widehat{\mathsf{X}}$ for which $f(\widehat{\mathsf{X}}) \geq f(\mathsf{Y})$ for all $\mathsf{Y} \in \mathcal{N}(\mathsf{X})$, i.e., a configuration with a fitness larger than that of all its neighbors. In the same vain, walks can be used to explore the landscape. An *adaptive walk* (Weinberger 1991b; Jain 2011; Neidhart and Krug 2011) is a sequence of configurations such that fitness increases in each step, ending when no further improvement can be found, i.e., when a local optimum is reached. A *gradient walk* is an adaptive walk in which in each step the neighbor $\mathsf{Y} \in \mathcal{N}(\mathsf{X})$ is chosen that maximizes the fitness among all neighbors of X.

The definition of locality makes it possible to render the intuition of *ruggedness* precise. A landscape is *smooth* if it has few local optima and (on average) long adaptive and gradient walks. In such a landscape, it suffices to follow an up-hill direction to eventually reach one of the few optima. A landscape is *rugged*, on the other hand, if it has many local optima and commensurably short adaptive walks that usually get stuck quickly in suboptimal configurations.

A random walk on $\mathcal{Q}$ is simply a sequence $\mathcal{R} = (\mathsf{X}_0, \mathsf{X}_1, \ldots, \mathsf{X}_t, \ldots)$ of configurations $\mathsf{X}_i \in \mathcal{Q}$ such that $\mathsf{X}_{i+1} \in \mathcal{N}(\mathsf{X}_i)$ is a randomly chosen neighbor of its predecessor X_i. Random walks provide a one-dimensional slice through the landscapes if one considers the corresponding sequence of fitness values $f(\mathcal{R}) = \big(f(\mathsf{X}_0), f(\mathsf{X}_1), \ldots, f(\mathsf{X}_t), \ldots\big)$. In fact, in this manner one obtains a picture that very much looks like Fig. 1, except the horizontal axis now represents *time*, i.e., the number of steps along the random walk (Weinberger 1990). Standard methods from time-series analysis can be applied to the pseudo-time-series $t \mapsto f(\mathsf{X}_t)$ sampled along random walks. Of particular interest is the autocorrelation function

$$\rho(\tau) = \lim_{t\to\infty} \frac{1}{t-\tau} \sum_{i=\tau}^{t} \widetilde{f}(\mathsf{X}_t)\, \widetilde{f}(\mathsf{X}_{t-\tau}) \tag{2}$$

where $\widetilde{f}(\mathsf{X}) = (f(\mathsf{X}) - \bar{f})/s_f$ is the landscape normalized by subtracting the mean fitness value $\bar{f} = (1/|X|)\sum_{x\in X} f(x)$ and dividing by the square root of the variance $s_f^2 = (1/|X|)\sum_{x\in X}(f(x) - \bar{f})^2$. The autocorrelation functions satisfies $\rho(0) = 1$ and eventually approaches $\lim_{\tau\infty} \rho(\tau) = 0$ as the random walk moves to unrelated regions of configuration space. The speed at which it drops measures how quickly the information about the local fitness is lost, which is fast in rugged landscapes and slow in smooth landscapes. The *correlation length* $\ell = \sum_{\tau=0}^{\infty} \rho(\tau)$ thus serves as a convenient measure of ruggedness that can easily and efficiently be estimated.

2.3 Additivity, Epistasis, and Elementary Landscapes

The sequence space $\mathcal{Q}_\ell^\kappa$ with $\kappa = |\mathcal{A}| = 2$ is a Boolean hypercube. Instead of the digits 0 and 1 or the nucleotides G and C we consider here the directions of *spins*, $x_k = \pm 1$ or in other words, a configuration X is a sequence of spins: $\mathsf{X} = (x_1, \ldots, x_n)$ where the spin at the position is either $+1$ or -1. Then, every function of X can be written in a power series expansion[6]:

$$f(\mathsf{X}) = a_0 + \sum_i a_i x_i + \sum_{i<j} a_{ij} x_i x_j + \sum_{i<j<k} a_{ijk} x_i x_j x_k + \ldots + \sum_{i_1<i_2<\cdots<i_p} a_{i_1 i_2 \ldots i_p} x_{i_1} x_{i_2} \ldots x_{i_p} + \ldots \,. \tag{3}$$

The coefficients, a_i, a_{ij}, a_{ijk}, etc., or in multi-index notation a_I, quantify the contribution $\varphi_I(x) = \prod_{j\in I} x_j$ of the interaction of the subset I of spins. The multi-index $I \subseteq \{1, 2, \ldots, N\}$ specifies a set of interacting positions. The cardinality $|I|$ stands for the number of interacting positions. The expansion above, therefore, gives a decomposition of the landscape into contributions of different interaction orders. The first term, a_0 is not associated with any position since $I = \varnothing$ and captures simply the average fitness $a_0 = \bar{f}$. The next term, $\sum_i a_i x_i$, specifies the *additive* contributions, each of which depends on exactly the spin of one sequence position i corresponding to $|I| = 1$. The additive terms correspond to the association of an *effect* in form of a fitness difference between the two spin states $x_i = \pm 1$. The quadratic terms with $|I| = 2$, $\sum_{i<j} a_{ij} x_i x_j$, correspond to pairwise contribution of spins like in the simple spin glass models mentioned in the previous Sect. 2.2. They describe the interactions between exactly two sequence positions. In the language of quantitative genetics, for example in genome-wide association studies, the additive terms with $|I| = 1$ measure the first order effects of independent genes.[7] The higher-order terms, $|I| \geq 2$, therefore constitute a measure for the contributions that are subsumed under the notion of *epistasis*. Based on an extensive data collection the fitness landscape of HIV-I has been modeled and analyzed by means of the series expansion (3) (Kouyos et al. 2012) and additive and epistatic effects were discussed.

From a mathematical point of view, Eq. (3) is the discrete analog of a Fourier expansion of the landscape. It can be written in much more compact form as $f(\mathsf{X}) = \sum_I a_I \varphi_I(x)$. The functions $\varphi_I(x)$ are the so-called *Walsh functions* (Walsh 1923). The relative importance of interactions of order p is quantified by the *amplitude spectrum* of landscapes (Hordijk and Stadler 1998),

[6] This expansion reminds of the *cluster expansion* for the partition function used in statistical physics (Mayer and Montrol 1941).

[7] According to Ronald Fisher's assumption these were the dominant contributions to fitness $f(\mathsf{X})$ (see Sect. 1).

$$A_p = \sum_{I:|I|=p} a_I^2 \Big/ \sum_{I\neq\varnothing} a_I^2 \tag{4}$$

By construction $A_p \geq 0$ for all $p \geq 1$ and $\sum_{p=1}^{N} A_p = 1$ since there is no contribution of $I = \varnothing$, i.e., of the landscape average a_0. The values of A_p quantify the (relative) contribution of interactions to the total variability in the landscape in each order p. Thus a landscape with $A_1 = 1$ is additive, and the spin glass mentioned earlier is characterized by $A_2 = 1$. The amplitude spectrum A_p is also closely related to the autocorrelation length (Reidys and Stadler 2002). In general, the neighborhood structure of a graph is expressed by the adjacency matrix A with entries $A_{ij} = 1$ is $x_i \in \mathcal{N}(x_j)$ and $A_{ij} = 0$ otherwise. The so-called Laplacian matrix is $\Delta = \mathrm{A} - \mathrm{d}\,\mathrm{I}$ for regular graphs, these are graphs where all vertices have the same number of neighbors: $\mathrm{d} = |\mathcal{N}(x)|\,\forall\, x \in X$. For the Boolean hypercube, the Walsh functions are the eigenvectors of Δ, satisfying $\Delta\varphi_I = \lambda_p\varphi_I$ with $p = |I|$ and $\lambda_p = 2p$. For regular graphs, the autocorrelation function of a landscape can be written as

$$\rho(\tau) = \sum_{p\geq 1} A_p(1 - \lambda_p/\mathrm{d})^{\tau} \tag{5}$$

in terms of the eigenvalue λ_p of the Laplacian, the size of the neighborhoods D, and the amplitude spectrum. The Laplacian spectrum $\{\lambda_p\}$ and the corresponding eigenvectors φ_I of course depend on the configuration space and take on different value for graphs different from the Boolean hypercube. The simple form of the autocorrelation function ρ yields a simple expression for the correlation length:

$$\ell_{\mathrm{c}} \;=\; \mathrm{d} \sum_{p\geq 1} A_p/\lambda_p \tag{6}$$

Thus landscapes become more rugged, in the sense of shorter correlation length, the more the amplitude spectrum shifts toward large values of p.

Landscapes with $A_p = 1$ for some p, and thus $A_q = 0\,\forall\, q \neq p$, are called *elementary* (Stadler 1996). By definition, their autocorrelation functions are simple exponentials, $\rho(\tau) = (1 - \lambda_p/\mathrm{d})^{\tau}$, and thus the pseudo-time-series sampled along random walks resemble so-called AR(1) processes (Stadler 1996; Dimova et al. 2005). Elementary landscapes also have some other useful properties (Stadler 1996; Davies et al. 2001; Whitley et al. 2008; Dimova et al. 2009; Klemm and Stadler 2014). All local maxima, for example, have an above-average fitness value or in other words, if $\widehat{\mathsf{X}}$ is a local maximum, then $f(\widehat{\mathsf{X}}) \geq \overline{f}$ (Grover 1992).

2.4 Kauffman's NK Model

The *NK model* (Kauffman and Weinberger 1989) serves as a general model of fitness landscapes with tunable ruggedness introduced by epistatic interactions among N genes, each with two alleles denoted by 0 and 1, or $x_i \in \{0, 1\}$. *Neighboring genomes* are reached by changing the state of one allele, i.e., by flipping a single x_1 from 0 to 1 or *vice verse*, similar to spin flips. Therefore the NK-landscape 'lives' on a Boolean hypercube. Its fitness function $f(\mathsf{X})$ is the sum of N contributions $f_i(\mathsf{X})$ $(i = 1, \ldots, N)$, one for each gene or locus, which in turn depends on the i-th bit of X as well as K other bits of X:

$$f(\mathsf{X}) = \frac{1}{N}\sum_{i=1}^{N} f_i(\mathsf{X}) \tag{7}$$

This construction explicitly models epistatic interactions. For $K = 0$ there is no epistasis and $f_i(\mathsf{X})$ depends only on the value of x_i, i.e., the landscape is additive. A particular instance of the model is thus specified in terms of a table of $N = 2^{K+1}$ fitness values, one for each configuration of the $K + 1$ bits contributing to $f_i(\mathsf{X})$, together with a rule that determines which K bits influence $f_i(\mathsf{X})$. Common choices are the nearest neighbors of the locus 'i', or a set of K randomly assigned sites. The fitness values themselves are assigned from a uniform distribution. An example for $N = 3$ and $K = 2$ is shown in Fig. 4. Further details can be found in Kauffman (1993).

The ruggedness of the landscape increases with K. For $K = 0$ there are independent, additive contributions from each bit. For $K = N - 1$, on the other hand, there is a different random contribution $f_i(\mathsf{X})$ for each value of the bit string X, hence fitness values are uncorrelated random numbers. The landscapes resulting from the

value of bit			fitness contribution			total fitness
b_1	b_2	b_3	f_1	f_2	f_3	$f = \frac{1}{N}\sum_{i=1}^{N} f_i$
0	0	0	0.6	0.3	0.5	0.47
0	0	1	0.1	0.5	0.9	0.50
0	1	0	0.4	0.8	0.1	0.43
0	1	1	0.3	0.5	0.8	0.53
1	0	0	0.9	0.9	0.7	0.83
1	0	1	0.7	0.2	0.3	0.40
1	1	0	0.6	0.7	0.6	0.63
1	1	1	0.7	0.9	0.5	0.70

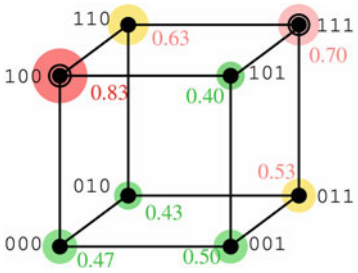

Fig. 4 Example of an NK-landscape. A simple example of an instance of the NK model for $N = 3$ and $K = 2$. Left: Table of the (randomly assigned) fitness contributions for each of the $2^{K+1} = 2^3 = 8$ possible neighborhood configurations. The fitness of each string is the average of the individual fitness contributions. Right: Fitness values assigned to the vertices of the 3-dimensional cube. The landscape has two local optima, 100 and 111 indicated in red by ⊙

NK model have been investigated in quite some detail in the literature (Weinberger 1991b; Kaul and Jacobson 2006; Neidhart et al. 2013; Nowak and Krug 2015) (see also Campos et al. 2002; Ochoa 2006 in the context of an error threshold).

Except for the additive case $K = 0$, the NK models are not additive. Their amplitude spectra, however, take a simple form (Neidhart et al. 2013)

$$A_p = \frac{1}{2^{K+1}} \binom{K+1}{p} \text{ if } p \leq K+1 \qquad A_p = 0 \text{ if } p > K+1 \tag{8}$$

We see that the landscapes thus become more rugged with increasing K. For large values of K, neither the maximal interaction order $p = K + 1$ nor the additive mode $p = 1$ contribute much. The dominating contribution comes from interaction orders around $(K + 1)/2$, which yields a correlation length of approximately $\ell = N/\big(2(K + 1)\big)$ (Hordijk et al. 2020), again showing that K tunes the ruggedness. Closed-form expressions are also available for the length of adaptive walks and gradient walks (Weinberger 1991b; Fontana et al. 1993b), resulting in walk lengths proportional to $N \ln(K + 1)/(K + 1)$.

2.5 Neutrality

Kimura's *neutral theory* of biological evolution focuses exclusively on the aspects of neutrality (Kimura 1983), in particular on the behavior of populations in the absence of selective differences. This corresponds to the extreme case of a *flat* landscape, $f(\mathsf{X}_k) = \overline{f}\ \forall\ \mathsf{X}_k \in \mathcal{Q}$. Neutrality, however, plays its important role in evolution through its interactions with selection. Two configurations $\mathsf{X}_1, \mathsf{X}_2 \in \mathcal{Q}$ are said to be *neutral* if $f(\mathsf{X}_1) = f(\mathsf{X}_2)$. More colloquially, one refers to a landscape as *neutral* if a substantial fraction of adjacent pairs of configurations are neutral or, in other words, if the existence of neutral neighbors is a common phenomenon.

One source of neutrality are many-to-one genotype-phenotype maps (see also Sect. 3.5). Here, neutral neighbors arise from genotypic changes that are phenotypically silent. Such models have been studied quite extensively just before the turn on the millennium in the context of cellular automata (Hordijk 1997) and sequential dynamical systems (Barrett and Reidys 1999). In case of biomolecules the classical source of neutrality are the synonymous codons in protein synthesis through translation. We shall also see below that neutrality plays also a key role for RNA molecules (see Sect. 3). The second source of neutrality comes from the second map Φ in Eq. (1): Different phenotypes may have the same fitness, $f_j = \Phi(S_j) = \Phi(S_k) = f_k$. It is important to make clear the meaning of *the same* in the context of both mappings. Clearly it cannot be mathematically or exactly the same, because there will be always some, eventually a tiny difference. In the case of the synonymous codons differences are introduced through processing of the messenger RNAs on the ribosome (Brule and Grayhack 2017). Sufficiently small differences, however, are effectively 'invisible' to natural selection and will be indistinguishable by a given measurement procedure.

Neutrality more realistically means 'not distinguishable' for selection within the bandwidth of natural randomness in the sense of the *nearly neutral theory* developed by Ohta (1992).

At the first glance, neutrality and ruggedness appear to be two sides of the same (molecular) coin. Surprisingly, however, ruggedness (as measured by the correlation length of the landscapes) and neutrality (as measured by the number of neutral neighbors) can be tuned independently of each other. A described in detail in Reidys and Stadler (2001), it is possible to introduce neutrality by setting a (large) fraction μ of the coefficients a_I in the Fourier expansion to 0. The expected number of neutral neighbors then increases with μ. If this is done in an elementary landscape, i.e., assuming $a_I \neq 0$ only if $|I| = p$, we have already seen above, that $A_p = 1$, irrespective of the distribution of zero and non-zero values among the a_I. Thus autocorrelation function and correlation length remain unchanged, while neutrality is tuned by μ. Conversely, one can fix a desired level of neutrality and then choose μ accordingly for series of different interactions orders p. For the p-spin models—these are elementary landscapes with a fixed value of p—the average fraction of neighbors can be computed as $\overline{\nu} = \mu^{\binom{N-1}{p-1}}$, and in these settings most coefficients a_I must vanish. This is very natural in some important physical models, however. In short range spin glass all but a linear number of coefficient vanishes, thus $\mu = 1 - z/N^{p-1}$, and one obtains $\overline{\nu} \approx e^{-z}$, where z measures the connectivity of the lattice, i.e., the effective number of spins that are potentially available as interaction partners. In summary, therefore, ruggedness and neutrality are independently tunable properties of landscapes (Reidys and Stadler 2001).

2.6 Holey Landscapes

Models focused on neutrality often disregard all details of selection and only distinguish viable, $f = 1$, from inviable, $f = 0$, configurations.[8] Considering the viable set in a sequence space $\mathcal{Q}_\ell^{\mathcal{A}}$ is tantamount to analyzing an induced subgraph of configuration space. The *holey landscape* model introduced by Gavrilets (1997a) is an example of such a simple mapping. The model borrows the basic idea from Dobzhansky (1937): Fitness values one and zero are assigned at random to genotypes or to alleles in sequence space. The result is a flat surface with holes (Gavrilets and Gravner 1997). The so-called Bateson-Dobzhansky-Muller model (Dobzhansky 1937; Bateson 2009; Muller 1942) does not deal with viruses but shows beautifully how complex observations can be explained by making use of simple model assumptions. It studies reproductive isolation through the incompatibility of alleles, in particular in the context of speciation. As shown in Fig. 5 incompatibility of two alleles on different loci leads to a zone of unviable genotypes—a hole, which is surrounded by a connected ridge of viable genotypes.

[8] For evolutionary purposes it makes little difference whether a genotype is inviable or unfit for reproduction.

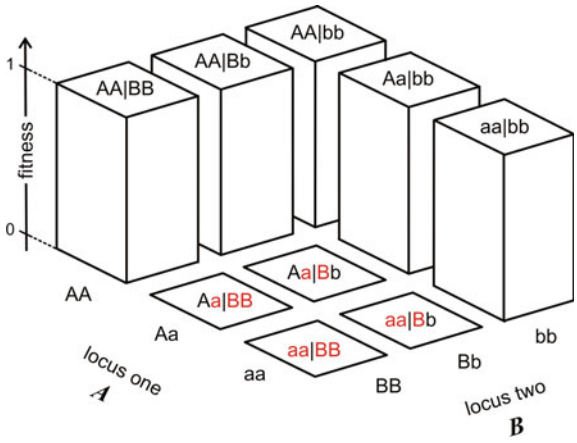

Fig. 5 Bateson-Dobzhansky-Muller landscape. The fitness landscape for two locus, ***A*** and ***B***, two alleles, (A, a) and (B, b), model is shown in Fig. 5. It consists of nine genotypes, which are classified as viable (1) and inviable (0). Genotypes that contain the two incompatible alleles a and B (red), do not survive and have fitness $f = 0$. Nevertheless it is possible to migrate from AA|BB to aa|bb along a ridge of viable genotypes with $f = 1$ (Gavrilets 1997b)

The holey landscape model is based, in essence, on three assumptions (Gavrilets 1997a; Gavrilets and Gravner 1997):

(i) fitness values are assigned at random to individual points in sequence or genotype space,
(ii) fitness values are generated independently and accordingly, the resulting landscape is uncorrelated, and
(iii) fitness is a function with a Boolean output in the sense that only two fitness values (1, 0) are allowed and the probability of a given genotype to be viable is a parameter, $0 \leq p \leq 1$.

The viable genotypes are considered as the nodes of a graph in genotype space, whose edges are all connections of pairs of nodes with Hamming distance $d_{\mathrm{H}} = 1$. Viable genotypes form clusters in genotype space, which are connected subgraphs commonly characterized as *connected components*.

Despite its simplicity the holey landscape model makes a number of important predictions, which follow directly from the theory of random graphs (Erdős and Rényi 1959; Erdős 1959; Gilbert 1959) (for a recent comprehensive presentation see also Bollobás 2001): The structure of connected components and their distribution in genotype space is determined by the probability of viability p. In the subcritical regime p is below a critical value, $p < p_{\mathrm{cr}}$, the graph consists of many small components and evolution confines the population to a small part of genotype space. The supercritical regime, $p > p_{\mathrm{cr}}$, is characterized by the existence of a single *giant component* that commonly spans all genotype space. Evolution allows the migration

of populations through the whole sequence space. The critical value of $p = p_{cr}$ is tantamount to a *percolation threshold*, which separates the regime of graphs spanning sequence space from the regime of dominant local networks.

Another example considering random graphs as reference state are *neutral networks* of RNA structures (Schuster et al. 1994; Reidys et al. 1997; Reidys 1997) (see Sect. 3.5). This line of investigation focuses in particular on percolation-like phenomena (Reidys et al. 1997; Reidys 2009) and a key result of this line of reasoning is the existence of a threshold for the fraction $\overline{\lambda}$ of neutral neighbors above which neutral conformations form a single, essentially connected *neutral network*. Such large degrees of neutrality support drift in a manner very similar Kimura's neutral theory on flat landscapes (Huynen et al. 1996).

3 The RNA Model

The molecular approach to global fitness landscapes requires a tractable model of the genotype-phenotype mapping ψ, the first mapping in Eq. (1), which relates biopolymer sequences and structures. Molecular structures of biopolymers—addressed as phenotypes in the RNA model or in virology—are obtained through folding sequences of polynucleotides or polypeptides into 3D-structures. In general sequence-structure mappings are highly complex and—at the present state of knowledge—can be approached reliably only by combined experimental and theoretical approaches, which are very demanding in general. Despite substantial progress in computational structure predictions within the last two decades, calculation of biopolymer structures from sequence data alone is still in its infancy. Therefore only two approaches seem to be promising: (i) studies of large parts of or entire sequence spaces with approximations that dispense from the claim of being quantitative, or (ii) investigations with accurate methods of small local regions of sequence space—consisting, for example, of the wild type and a few selected mutants. We mention examples for both approaches: (i) analysis of fitness landscapes through applying the RNA model in this section and (ii) empirical fitness landscapes, which necessarily are confined to small parts of sequence space, and will be discussed in Sect. 6. The RNA model is presented here in some detail, because it comes closest to simple evolving systems like virus evolution and evolution in vitro and several results are directly transferable to the analysis of experimental data.

3.1 RNA Secondary Structures

In the RNA model (Schuster 2003, 2006) structures are represented by simplified molecular structures, in particular *RNA secondary structures* of minimal free energies (Fig. 6), which will be characterized by *shapes*. RNA sequences, which are often addressed as *primary structures*, fold into complex 3D-structures (Leontis and Westhof 2012; Miao and Westhof 2017; Li et al. 2020). The primary structure or

sequence of an RNA molecule defines the configuration, i.e., the position of each nucleotide along the one-dimensional or unbranched backbone chain. The folding process in aqueous media can be visualized easier when separated into two steps: (i) the primary structure folds into a secondary structure by forming six types of *legal base pairs*[9]—the four Watson-Crick base pairs, and the two wobble pairs subsumed in a paring alphabet $\mathcal{B}(\mathcal{A}) = \{\mathsf{A} = \mathsf{U}, \mathsf{U} = \mathsf{A}, \mathsf{G} \equiv \mathsf{C}, \mathsf{C} \equiv \mathsf{G}, \mathsf{G} - \mathsf{U}, \mathsf{U} - \mathsf{G}\}$ (Fig. 6),[10] which have similar planar geometries and fit perfectly into the RNA double helix, and (ii) further folding of the secondary structure into the 3D, spatial or tertiary structure through tertiary interactions and formation of 3D-motifs. One reason for the partitioning of the folding process into these two steps concerns the free energy of folding[11]: (i) the major part of the (negative) free energy of secondary structure formation results from base pair stacking, and (ii) the energies of additional interactions in the RNA tertiary structures—hydrogen bonding and other intermolecular interactions—are commonly weaker than the stacking energies. As a rule the secondary structure remains unchanged when the 2D shape is folded into the 3D structure.

An RNA secondary structure is equivalent to a listing of the base pairs in the structure belonging to the six legal conformations ($\mathcal{B}$). In addition, three conditions or *rules* have to be fulfilled for valid structures:

(i) *binary interaction rule*: An individual nucleotide 'i' is either involved in a single base pair 'i-k' or is unpaired,
(ii) *no-pairs-between-neighbors rule*: Base pairs to nearest neighbors, 'i-j' with $j = i \pm 1$, are forbidden, and
(iii) *no-pseudoknot rule*: Two base pairs 'i-j' and 'k-l' with $i < j$, $i < k$ and $k < l$ are only accepted if either $i < k < l < j$ or $i < j < k < l$ is fulfilled or, in other words, the second base pair is either enclosed by the first pair or lies completely outside.

The binary interaction rule (rule (i)) prohibits the participation of one and the same nucleotide in two or more base pairs. The no-pairs-between-neighbors rule (rule (ii)) implies the exclusion of empty parentheses: () is not allowed. Conventional RNA structures are free of pseudoknots (rule (iii); Fig. 7). Then, an RNA secondary structure can be encoded uniquely by strings of length ℓ over an alphabet with three symbols, $\mathcal{H} = \{'\bullet','(', ')'\}$ for unpaired nucleotides, nucleotides opening and nucleotides closing base pairs in the direction from the 5'- to the 3'-end. In other words, '(' denotes a paired nucleotide with the pairing partner toward the 3'-end and ')' one with the partner toward the 5'-end, respectively.[12] For example, '(((•(((•••)))•)))'

[9] The notion legal indicates the existence of other stable base pairs, which do not fit into the structure of the double helix.

[10] The number of lines connecting the two nucleotides indicates the strength of the interaction.

[11] Throughout this chapter we mean *Gibbs free energy* when we write *free energy*.

[12] Here the exclusion of pseudoknotted structures is essential, because a unique coding of pseudoknotted structures would require colored parenthesis as illustrated, for example, by the pseudoknot in the shape shown in Fig. 7.

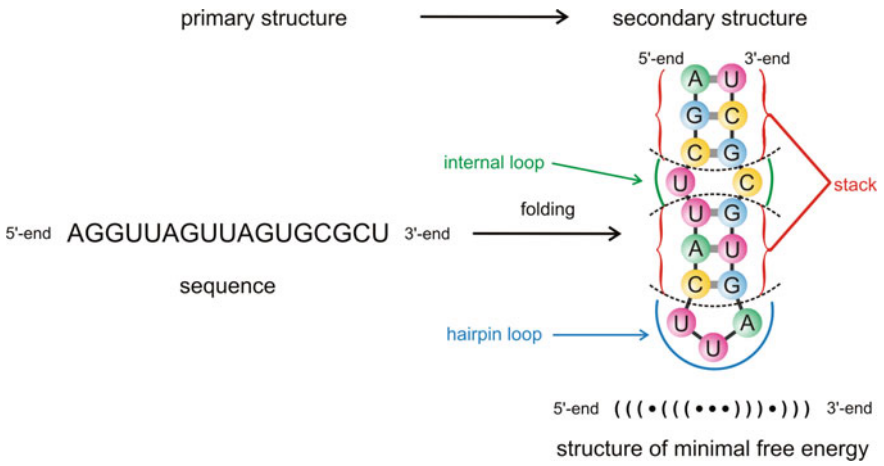

Fig. 6 RNA secondary structures of minimal free energy. RNA sequences or primary structures X_k fold into secondary structures also called *shapes* S_k through formation of base pairs that fit into the double helix. Six base pairs, the four Watson-Crick pairs and the two wobble pairs, $\mathcal{B} = \{\mathsf{A}=\mathsf{U}, \mathsf{U}=\mathsf{A}, \mathsf{G}\equiv\mathsf{C}, \mathsf{C}\equiv\mathsf{G}, \mathsf{G}-\mathsf{U}, \mathsf{U}-\mathsf{G}\}$, are allowed. Only the structure with the lowest free energy—*minimum free energy* (mfe) *structure*—is considered. The two ends of polynucleotides are chemically different and denoted as 5'- and 3'-ends indicating the position of the terminal phosphate moiety on the ribose ring. A structure consists of a backbone (black) and base pairs (gray) and can be decomposed into modules—free ends, bulges, hairpin loops (blue), internal loops (green), multiloops, and stacks (red), which are assumed to contribute additively to the free energy of the molecule (see also Fig. 10). A general convention for sequences and structures writes the 5'-end on the l.h.s. and the 3'-end on the r.h.s. of the diagram. Conventional secondary structures are shapes, which can be drawn on the plane without intersections of the backbone or the base pairs. Then the *structure string* defines the base paring pattern uniquely, because the mathematical parentheses convention is valid. Exceptions are structures with *pseudoknots* that are conventionally understood as motifs of tertiary structures (Fig. 7)

is the secondary structure of the RNA molecule in Fig. 6 in encoded form. Further restrictions concerning physically improbable structural elements will be discussed in Sect. 3.3.

3.2 Counting Structures

The RNA model is the only currently known example based on real molecules where exact values for the cardinalities of structure spaces are available. The three-symbol encoding of RNA secondary structures defines a *shape space* $\mathcal{S}_\ell^{\mathcal{H}}$ where ℓ is the chain length of the polynucleotide and $\mathcal{H} = \{\bullet, (,)\}$ the alphabet used to encode the shapes. The genotype-phenotype relation in the RNA model,

$$\Psi : \left\{\mathcal{Q}_\ell^{\mathcal{A}}; d_{\mathrm{H}}(\mathsf{X}_i, \mathsf{X}_j)\right\} \quad \Longrightarrow \quad \left\{\mathcal{S}_\ell^{\mathcal{H}}; d_{\mathrm{S}}(\mathsf{S}_i, \mathsf{S}_j)\right\} \tag{9}$$