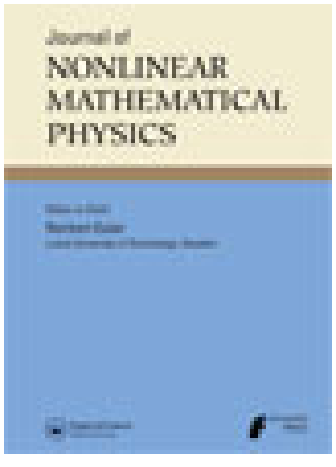


Journal of Nonlinear Mathematical Physics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tnmp20

HOW THE MUTATIONAL-SELECTION INTERPLAY ORGANIZES THE FITNESS LANDSCAPE

FRANCO BAGNOLI [a b] & PÍETRO LIÓ [c]

[a] Dept. Energy and CSDC, University of Florence, via S. Marta, 3 Firenze 50139 Firenze, Italy

[b] INFN, sez. Firenze, Italy

[c] Computer Laboratory, University of Cambridge, 15 JJ Thomson Avenue, Cambridge CB3 0FD, UK

Published online: 04 Mar 2013.

To cite this article: FRANCO BAGNOLI & PÍETRO LIÓ (2011) HOW THE MUTATIONAL-SELECTION INTERPLAY ORGANIZES THE FITNESS LANDSCAPE, Journal of Nonlinear Mathematical Physics, 18:sup2, 265-286, DOI: 10.1142/S1402925111001532

To link to this article: http://dx.doi.org/10.1142/S1402925111001532

ARTICLE

World Scientific
www.worldscientific.com

ATLANTIS PRESS

Journal of Nonlinear Mathematical Physics, Vol. 18, Suppl. 2 (2011) 265–286

DOI: 10.1142/S1402925111001532

HOW THE MUTATIONAL-SELECTION INTERPLAY ORGANIZES THE FITNESS LANDSCAPE

FRANCO BAGNOLI

Dept. Energy and CSDC, University of Florence
via S. Marta, 3 Firenze 50139 Firenze
Italy. Also INFN, sez. Firenze
franco.bagnoli@unifi.it

PÍETRO LIÓ

Computer Laboratory, University of Cambridge
15 JJ Thomson Avenue, Cambridge CB3 0FD, UK
pl219@cam.ac.uk

Received 14 November 2010
Revised 16 January 2011
Accepted 18 January 2011

Fundamental questions posed in classical genetics since early 20th century are still fundamental in today post genomic age. What has changed is the availability of huge amount of molecular genetics information on a broad spectrum of species and a more powerful and rich methodological approach, particularly that one based on statistical mechanics and dynamical system theory which is providing unprecedented prediction power. Here we focus on the behavior of basic life forms such as bacteria and viruses which have small genomes and short generation times. We show that central issues of the evolutionary theory, i.e. how genotype, phenotype and fitness are related, the effect of positive and negative natural selection, the specie formation could be described by simple models which allow predictions and validation using experimental data.

Keywords: Fitness; natural selection; mutation; quasi species.

1. Introduction

Quantitative methods have been applied to evolutionary biology for many years; statistical physics is increasingly applied to what is nowadays defined soft matter which includes complex ensembles such as colloids, membranes and biomolecules such as DNA and proteins. It seems reasonable to extend statistical physics even further to genes and proteins.

Biologists have introduced, using intuitive descriptions, important concepts such as phenotype, genotype and fitness which are fundamental to the understanding of heredity, development of organisms and species formation. The genotype of an organism is the class to which that organism belongs as determined by the DNA sequence passed to the organism by its parent(s). The phenotype of an organism is the class to which that organism belongs as determined by the behavioral characteristics of the organism, for example its metabolic

activities and its pattern of shape, development and movement. Natural selection acts on phenotypes; the propensity of an individual to survive and produce viable offsprings is termed fitness, and we assume that it is proportional to the average number of offsprings reaching the reproductive age for a given phenotype, and for a given time interval (one generation).

It is noteworthy that the spreading of infectious diseases, caused by bacteria, viruses or other pathogens, could be approached from an ecological perspective. Following this ecological framework, viruses and bacteria represent the most interesting laboratory of evolution to study the relationship between the genotype (molecular sequences which are affected by mutations) and phenotype (molecular structure, affected by selection). Viruses and bacteria have general very compact genomes, large population sizes, rapid reproductive rates, and, above all, high rates of mutation and recombination, so we do not need to wait very long time before observing the effects of selection pressures.

For simplicity, the phenotype of bacteria could be thought as an input-output devices where the "computational unit" is given by complex genetic and biochemical regulatory networks, i.e., the set of relationships that involve genes (which are portions of genotype) and proteins regulative mechanisms. The main purpose of such network is to process a subset of possible inputs and produce all outputs in the required ratios that, for instance, form the biomass of the cell.

Following a similar synthesis approach, the phenotype of viruses could be thought as generated by RNA or DNA strings which fold in space in a 3D structure. Of course these representations may not describe all possible evo-devo (evolution-development) dynamics in populating the phenotype space and particularly they do not describe completely the discontinuities of the relationship between genotype and phenotype mapping. Although we focus on bacteria and viruses, the complexity of sexual reproduction, which is characteristic of higher eukaryotes, could in principle be taken into account by simulating its effects on genotypes space, i.e., by considering the arising of large jumps (differences) in genotype space due to the recombination (which has a sort of combinatorial law) of the parental genotypes to generate the offspring genotypes.

The genetic and metabolic networks representation of bacteria and eukaryotes suggest the presence of epistasis, i.e., the nonlinear (nonadditive) interactions between genes. Note that the epistasis of genes affecting important fitness-related functions such as those involved in reproduction and survival, largely influences the evolutionary predictions. Recent results show that relatively simple fitness landscape models may be sufficient to quantitatively capture the complex nature of gene interactions and could represent a valid alternative for the more complex and specific metabolic network models.

Following this result, in this work we will develop models at the level of fitness landscape disregarding the regulatory networks, i.e., considering one gene at each time and the bulk of its relationship as a constant environment [1].

In the next section we introduce the necessary biology and the related mathematical description which will be used in the rest of the paper. In particular we introduce a measure of distance between genomes, and a mathematical description of genotype changes due to mutation and selection operators. We extend the discussion to include positive selection at different DNA sites as a meaningful statistical estimator of the mutation-selection coupled dynamics. We then show that the concept of fitness landscape naturally emerges from

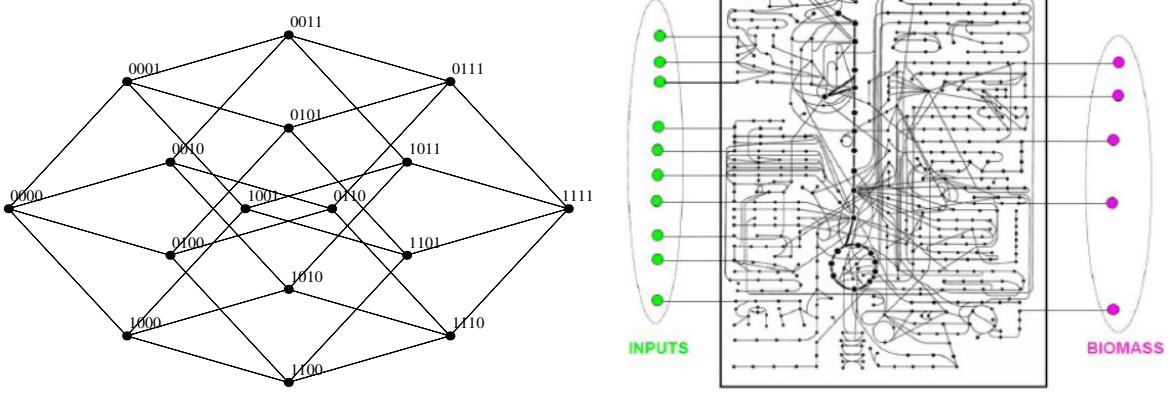
modeling evolution as a reaction-diffusion process. Section 6 makes use of Hiv biology to describe interesting quasi species theoretical issues such error threshold and Muller ratchet. Finally we highlight how the genomic available data could be used to fit even "light" statistical mechanics models and hopefully could result important in the analysis of pathogenic viruses and bacteria.

2. Modeling Blocks

From a mathematical point of view, the problem of modeling the evolution of haploid individuals is the following. Let us suppose that we can represent the genome g of an individual as a string of symbols. For simplicity, we use fixed-length strings of L Boolean symbols $g = (g_1, \ldots, g_L)$, $g_i \in \{0, 1\}$ or $g_i \in \{-1, 1\}$ All possible genomes can be mapped to the corners of a L-dimensional hypercube, see Fig. 1(a). Mutations corresponds to displacements on this hypercube.

Individuals are selected according to their phenotype u, which is a set of quantitative characters with which the individual interacts (and affects) the environment. A simplification that ease the treatment of the subject consists in assuming that the phenotype is just a function of the genotype: $u = u(g)$. This correspondence is generally quite complex, as indicated in Fig. 1(b). Each "gene" affects many phenotypic traits, and the genes that contribute to a phenotypic trait in general act in a nonlinear way. The phenotype of an individual can be thought as a relatively stable state of the metabolic network. For small variations around the stable state corresponding to an individual, we can assume that the variation of a phenotypic trait depends linearly by the variation of the activity level of genes. In this case, one can assume to perform a base change from genes to phenotypic traits and consider the evolution on the phenotypic space.

The survival and reproductive capacity of an individual depends on its phenotype. We introduce the fitness function A which gives the probability that a given phenotype is able to give rise to descendants. This function depends on the individual phenotype u, but also



(a) (b)

Fig. 1. (a) The two-dimensional projection of the Boolean hypercube for $L = 4$. (b) A schematic view of a metabolic network.

on the distribution of other phenotypes $p(u)$ in the environment (say: presence of preys, predators, parasites, etc.). Since A is essentially a probability, independent factors should affect it in a multiplicative way. So, it is convenient to write $A = \exp(H)$ (H is sometimes called the log-fitness), where H depends linearly on independent (nonepistatic) phenotypic traits.

A minimum example of the mapping between genotype and phenotype is given by a gene and the protein it codes; the phenotype is embedded in the correct folding of the amino acid chain into a 3-dimensional shape and its correct interactions with other proteins to form a metabolic network. Most of random mutations are simply unviable, so we can assume that most of genotypic space is "unobservable", in the sense that it corresponds to unviable phenotypes.

The evolution process occurs on the "backbone" that correspond to viable phenotypes, under the fitness selection which depends on the present population. In the case in which the rest of the population can be considered constant, the evolution from an individual's point of view corresponds to a stochastic walk (due to mutations) driven by selection, which is now just a function of the phenotype (or of the genotype, given our assumptions). The picture resembles that of a stochastic motion on a potential, which is termed "fitness landscape". Another view of the evolution is that of a reaction-diffusion pattern: reaction due to fitness and diffusion due to mutations.

One of the most prominent pattern is the concept of species (and that of quasi-species): a group of phenotypically and genotypically related individuals. It is possible to obtain, from very simple models some of which are presented in the following, that stable quasi-species occupy the maxima of the fitness landscape, and that their "width" and average fitness are related to the mutation rate and the curvature of the maximum. Another simple result concerns the coexistence of quasi-species: they can coexists only if their fitness is the same (Gause principle [2]). At first it would seems rather implausible that all coesisting species have exactly the same fitness by chance. Indeed, this is due to the interactions, i.e., epistasis, to the fact that the fitness A (and H) depends on the distribution of the population.

It is noteworthy that there are analogies between the dynamics between antibodies (immune response) and viruses and bacteria and the prey-predators dynamics. Predation or parasitism are able to "equalize" the fitness of their preys: the preys with higher fitness will grow at first, but this will stimulate their predator population to grow, and this lowers the preys' fitness. From the prey's point of view, predators and parasites are a variable load that can be exploited for competitive reasons. Since a predator targets (feeds on) a certain variety of phenotypes, an increase in the number of predators due to the presence of a certain phenotipe will also increase the predation on neighboring phenotypes. A similar effect is present among predators due to prey sharing.

This competition may promote speciation in an homogeneous environment. Even if there is no intersection in the prey preferences, the diffusion due to mutations makes parasites to attack also neighboring, i.e. similar, phenotypes, therefore acting as a competition term. In both cases, the competition strength (which could be considered a sort of load) can be lowered by broadening the phenotype distribution, even is this implies "occupying" phenotypes corresponding to lower fitness. An example is given by the thalassemia, in which slightly disadvantageous heterozigous are maintained in the population by malaria parasites.

Up to now, we have considered mutations as the biological equivalent of a mathematical constant diffusion term. However, for most of organisms, the "bare" (instantaneous) mutation rate is reduced by specialised protein complexes associated to the polymerase which act as correction molecular machineries. Therefore, the effective mutation rate is just a phenotype characteristics, which can therefore be selected in order to increase the variation in the population. An example is given by the hypermutation of antibodies during an immune response. A much more efficient mechanism developed to increase the variability of a population without incurring in the load of mutations is recombination.

In general, one observes only *neutral* variants, or the variations around a "wild type" phenotype with lower fitness, whose population is maintained only by the constant influx from the fitter strain. Indeed, the variation observed in the distribution of a given locus or portion of genome is a good indicator of the selective importance of such a portion. The term "negative selection" is used to characterize decrease of genetic variability, while "neutral selection" is used to characterize the variations that do not affect the phenotype. In recent years, signals of positive selection have been discovered in many genomes, and in particular in humans. The most powerful mechanism that promotes differentiation is the prey-predator (or better: parasite-host) interaction [3].

Positive selection can take two forms: directional, in which phenotypes (allele frequencies), are consistently driven (changed) in a given direction, or adirectional in which the natural selection pressure on the parasite genome simply broadens the allele distribution, for example to escape the host immune response. The first case is obviously a signature of a non-stationary function. We shall show in the following that a generic prey-predator dynamics promotes adirectional selection. It might happen that this pressure towards a broaden distribution may promote directional selection towards phenotypes with more variants.

Let us define the *distance* $d(x, y)$ between two genomes x and y as the minimum number of point mutations necessary to pass from genome y to genome x (and vice versa), $d(x, y) = L - \sum_{i=1}^{L} |x_i - y_i|$.

Note that the edit or Levenshtein distance between two strings, which is defined as the minimum number of edits needed to transform one string into the other, with the allowable edit operations being insertion, deletion, or substitution of a single character would be more appropriate but the cost to compute it, which is roughly proportional to the product of the two string lengths, makes this impractical for large genomes (the human genome is about 3 billion DNA bases).

For simplicity we assume that all single base mutations (base replacement, insertion, deletion) are equally likely, while in reality they depend on the identity of the symbol and on its positions on the genome [4–6]. Moreover, the manipulations of genetic material like DNA replication is subjected to proof-reading and error correction by several complexes, so that the probability of observing a mutation is quite small. The accuracy of this machinery is quite high for eukaryotic cells, less for bacteria and very low for the reverse transcription of RNA viruses.

As a consequence, RNA viruses, such as Hiv, have the highest mutation rates between 10^3 to 10^5 per base per generation; bacteria have rates of the order of 10^8 per base pair per generation (see also [7, 8] for statistical estimates of sequence variation across some virus and bacteria species); for human, the average mutation rate was estimated to be

2.5×10^{-8} mutations per nucleotide site based on a direct comparison of DNA sequences without function (pseudogenes) [9].

We assume that at most one mutation is possible per site per generation (replication). We denote with μ_s the probability of having one point mutation per generation.

The probability to have a point mutation from genotype y to genotype x is given by the short-range mutation matrix $\boldsymbol{M}_s(x, y)$ which is

$$\boldsymbol{M}_s(x|y) = \begin{cases} 1-\mu_s & \text{if x=y,} \\ \dfrac{\mu_s}{L} & \text{if d(x,y)=1,} \\ 0 & \text{otherwise.} \end{cases} \tag{2.1}$$

Other mutations correspond to long-range jumps in the genotypic space, such large scale genome segments duplications and recombinations. A very rough approximation consists in assuming all mutations are equally probable. Let us denote with μ_ℓ the probability per generation of this kind of mutations. The long-range mutation matrix, $\boldsymbol{M}_\ell$, is defined as

$$\boldsymbol{M}_\ell(x|y) = \begin{cases} 1-\mu_\ell & \text{if x=y,} \\ \dfrac{\mu_\ell}{2^L-1} & \text{otherwise.} \end{cases} \tag{2.2}$$

In the real world, certain types of mutations are more likely than others (for example the human protein ABOBEC3G causes G-to-A mutations), and in this case $\boldsymbol{M}_\ell$ becomes a sparse matrix $\widehat{\boldsymbol{M}}_\ell$. We introduce a sparseness index s which is the average number of nonzero off-diagonal elements of $\widehat{\boldsymbol{M}}_\ell$. The sum of these off-diagonal elements still gives μ_ℓ. In this case $\widehat{\boldsymbol{M}}_\ell$ is a quenched sparse matrix, and $\boldsymbol{M}_\ell$ can be considered the average of the annealed version.

As illustrated by the small-world effect, the combination of a vanishing long-range mutation rate and higher short-range mutation one may give origin to an effective fully-connected long-range mutation matrix [10]. Selection is modeled by a fitness function $A(u, \boldsymbol{p})$, which in general depends on the phenotype u of a given individual and on the whole distribution $p(u, t)$ of phenotypes (for example the presence of preys, predators, parasites, etc.).

In a mean-field approximation, in the limit of a vanishing mutation probability (per generation) and weak selection [11, 12] the evolution of the probability $p(x, t)$ of observing the genotype x at time t is given by

$$\dot{\boldsymbol{p}} = (\boldsymbol{A} - \langle A \rangle)\boldsymbol{M}\boldsymbol{p} \simeq (\boldsymbol{A} - \langle A \rangle)\boldsymbol{p} + \Delta\boldsymbol{p}, \tag{2.3}$$

where $\langle A \rangle = \int \mathrm{d}x\ A(x)p(x,t)$ is the average fitness, $\Delta = \boldsymbol{M} - \boldsymbol{1}$, and $\boldsymbol{1}$ is the identity. Considering only symmetric point mutations, Δ is the L-dimensional diffusion operator. One can recognize here the structure of a reaction-diffusion equation: evolution can be considered as a reaction-diffusion process in sequence space.

For discrete generations, the previous equation can be written as

$$p'(u) = \frac{1}{\langle A \rangle}\left(A(u, \boldsymbol{p})p(u) + \mu\frac{\partial^2 A(u, \boldsymbol{p})p(u)}{\partial u^2}\right), \tag{2.4}$$

with

$$\int_{-\infty}^{\infty} p(u)du = 1, \qquad \int_{-\infty}^{\infty} A(u, \boldsymbol{p})p(u)du = \langle A \rangle. \tag{2.5}$$

The analysis of the equation, Eq. (2.3) is much simpler if the fitness A does not depend on the population structure, i.e., $A = A(x)$, where x is the genome. The lineage of an individual corresponds to a walk on a static landscape, called the *fitness landscape*. A generic fitness function may be considered a landscape for short times, in which the species, other than the ones under investigations, can be considered constant.

In this case the evolution really corresponds to an optimization process: in the case of infinite population, and for mutations that are able to "connect" any two genomes (in an arbitrary number of passes — M is an irreducible matrix), the system may reach an equilibrium state (in the limit of infinite time).

There are a few landscapes that have been studied in details [13]. The simplest one is the flat landscape, where selection plays no role. It is connected to the concept of *neutral evolution*. In this landscape, evolution is just a random walk in sequence space, and one is interested in the probability of fixation of mutations in a finite populations, which corresponds to the divergence of an isolated bunch of individuals.

For very small probability of mutation, the asymptotic distribution is a *quasi-species* grouped around the peak. By increasing the mutation probability (or equivalently the genome length), this cloud spreads. It may happen that in finite populations no one has the right phenotype, so that the peak is lost, the so-called *error threshold* transition. Clearly, this transition poses the problem of how this peak has been populated for the first time. The idea is that the absence of population fitness space is actually similar to a Swiss cheese: paths of flat fitness and "holes" of unviable phenotypes (corresponding essentially to proteins that are unable to fold). The "roughness" of the fitness function is due to the presence of other species. So for instance a highly specialized predator, i.e. targeting a small ensemble of phenotypes, may become so tied to its specific prey, that its effective fitness landscape (for constant prey population) is extremely sharp.

Another consequences of an increased mutation rate, more effective for individuals with accurate replication machinery like multicellular ones, is the extinction of the species without losing their "shape", the so-called Muller's ratchet [14–16] or stochastic escape [17, 18], which, for finite populations, causes the loosing of the fitter strains by stochastic fluctuations.

Genes that have an additive effect are responsible for how much *quantitative traits* contribute to the phenotypes, and this results into a smooth landscape, shaped like the Fujiyama mount. It is possible to obtain a good approximation for the asymptotic distribution near such a maximum. Such results will be important in characterising the species competition. On a Fujiyama landscape, no error threshold transition is present [13]. Finally, one can study the problem of the evolution when the landscape has variable degree of roughness, a problem similar to that of disordered media in statistical mechanics [19, 20].

The evolution may appear as an optimization process. Indeed, if we neglect mutations, Eq. (2.3) becomes $\dot{\boldsymbol{p}} = (\boldsymbol{A} - \langle A \rangle)\boldsymbol{p}$, which is known as the *replicator equation*.

Assume that we start from a uniform distribution over the whole phenotypic space, and that the fitness shows a single, smooth maximum. The phenotypes u with $A(u) < \langle A \rangle$ tend to decrease in frequency, while those with $A(u) > \langle A \rangle$ tend to increase their frequency. Because of this, the average fitness $\langle A \rangle$ increases with time (Fisher theorem [21]).

The structure of Eq. (2.3) says that the fitness does not have absolute values. For a given genome, the survival of individuals depends on their *relative* fitness: those that

happen to have a fitness larger than average tend to survive and reproduce, the others tend to disappear. By doing so, the average fitness, in general, increases, so that a genome that is good at a certain time will become more common and the relative fitness more similar to the average one. The effects of this generic tendency depend on the form of the fitness.

3. How Fitness Shapes the Quasispecies

In the presence of a single maximum the asymptotic distribution is given by one quasispecies centers around the global maximum of the static landscape. The effect of a finite mutation rate is simply that of broadening the distribution from a delta peak to a bell-shaped curve.

We are interested in deriving the form of the asymptotic distribution near the maximum. We take a static fitness $A(u)$ with a smooth, isolated maximum for $u = 0$ (*smooth maximum* approximation). Numerical simulations show that for small mulation rates, the asymptotic quasispecies has a bell-shape form.

Let us assume that

$$A(u) \simeq A_0(1 - au^2), \tag{3.1}$$

where $A_0 = A(0)$.

We can try with a gaussian shape for the asymptotic distribution

$$p(u) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{u^2}{2\sigma^2}\right),$$

for which

$$\langle A\rangle = A_0 \int (1 - au^2)p(u)\, \mathrm{d}u = A_0(1 - a\sigma^2).$$

By substituting it into Eq. 2.4, by assuming $a\sigma^2 \ll 1$ and $au^2 \ll 1$, so that

$$\frac{1 - au^2}{1 - a\sigma^2} \simeq 1 - au^2 + a\sigma^2,$$

and equating the constant term and the term containing u^2, we get

$$\sigma^2 = \sqrt{\frac{\mu}{a}}.$$

The above approximations correspond to $\mu \ll 1/a$, i.e., the smoothness of the maximum is measured in terms of the mutation rate.

We shall see how this result can be used for determining the conditions of coexistence and of speciation induced by competition.

For completeness, we study here also the case of a *sharp maximum*, for which $A(u)$ varies considerably with u. In this case the growth rate of less fit strains has a large contribution from the mutations of fittest strains, while the reverse flow is negligible, thus

$$p(u-1)A(u-1) \gg p(u)A(u) \gg p(u+1)A(u+1)$$

neglecting last term, and substituting $q(u) = A(u)p(u)$ in Eq. (2.4) we get:

$$\frac{\langle A\rangle}{A_0} = 1 - 2\mu \qquad \text{for } u = 0 \tag{3.2}$$

and

$$q(u) = \frac{\mu}{(\langle A\rangle A(u) - 1 + 2\mu)} q(u-1) \qquad \text{for } u > 0 \tag{3.3}$$

Near $u = 0$, combining Eq. (3.2), Eq. (3.3) and Eq. (3.1)), we have

$$q(u) = \frac{\mu}{(1-2\mu)au^2} q(u-1).$$

In this approximation the solution is

$$q(u) = \left(\frac{\mu}{1-2\mu a}\right)^u \frac{1}{(u!)^2},$$

and

$$y(u) = A(u)q(u) \simeq \frac{1}{A_0}(1+au^2)\left(\frac{\mu A_0}{\langle A\rangle a}\right)^u \frac{1}{u!^2}.$$

We have checked the validity of these approximations by solving numerically Eq. (2.4); the comparisons are shown in Fig. (2). We observe that the *smooth maximum* approximation agrees with the numerics for small values of a, when $A(u)$ varies slowly with u, while the *sharp maximum* approximation agrees with the numerical results for large values of a, when small variations of u correspond to large variations of $A(u)$.

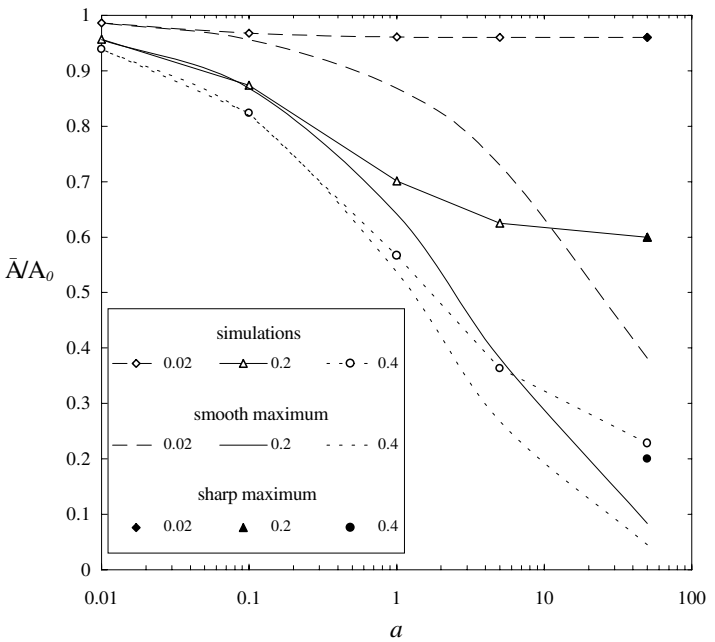


Fig. 2. Average fitness $\langle A\rangle/A_0$ versus the coefficient a, of the fitness function, Eq. (3.1), for some values of the mutation rate μ. Legend: *simulations* corresponds to the numerical solution of Eq. (2.4), *smooth maximum* refers to Eq. (3) and *sharp maximum* to Eq. (3.2).

4. Coexistence

We investigate here the conditions for which more than one quasi-species can coexist on a static fitness landscape without competition. Hiv quasi species have been found to change during anti viral therapies, in early and late stages of hiv infection and in different districts of the body [22]. Interestingly, we may imagine there are boundaries in which different quasi species may come in contact or nearby niches with different mutant frequencies. Wild type strains have the highest fitness in natural conditions but often much lower during different therapies. The need to minimize drug resistance and reduce treatment-related toxicities has engendered an interest in induction followed by maintenance regimen, in which a period of intensified antiretroviral therapy (induction phase) is followed by a lighter long-term regimen (maintenance phase). These alternance may generate oscillations in frequencies of quasi species which has not been fully investigated yet. Here we approach the problem as coexistence of quasi species. Let us assume that the fitness landscape has several distinct peaks, and that any peak can be approximated by a quadratic function near its maximum. For small but finite mutation rates, as shown by Eq. (3), the distribution around an isolated maxima is a bell shaped curve, whose width is given by Eq. (3) and average fitness by Eq. (3). Let us call thus distribution a quasi-species, and the peak a niche.

If the niches are separated by a distance greater than σ, a superposition of quasi-species (3) is a solution of Eq. (2.3). Let number the quasi-species with the index k, $p(u) = \sum_k p_k(u)$, where each $p_k(u)$ is centered around u_k and has average fitness $\langle A\rangle_k$. The condition for the coexistence of two quasi-species h and k is $\langle A\rangle_h = \langle A\rangle_k$ (this condition can be extended to any number of quasi-species). In other terms one can say that in a stable environment the fitness of all co-existing individuals is the same, independently on the species.

Since the average fitness (3) of a quasi-species depends on the height A_0 and the curvature a of the niche, one can have coexistence of a sharper niche with larger fitness together with a broader niche with lower fitness, as shown in Fig. 3. It is easy to derive the condition for coexistence. We assume that the conditions for smooth maximum hold, and that the maxima are so separated that the asymptotic distribution can be modeled as a sum of two

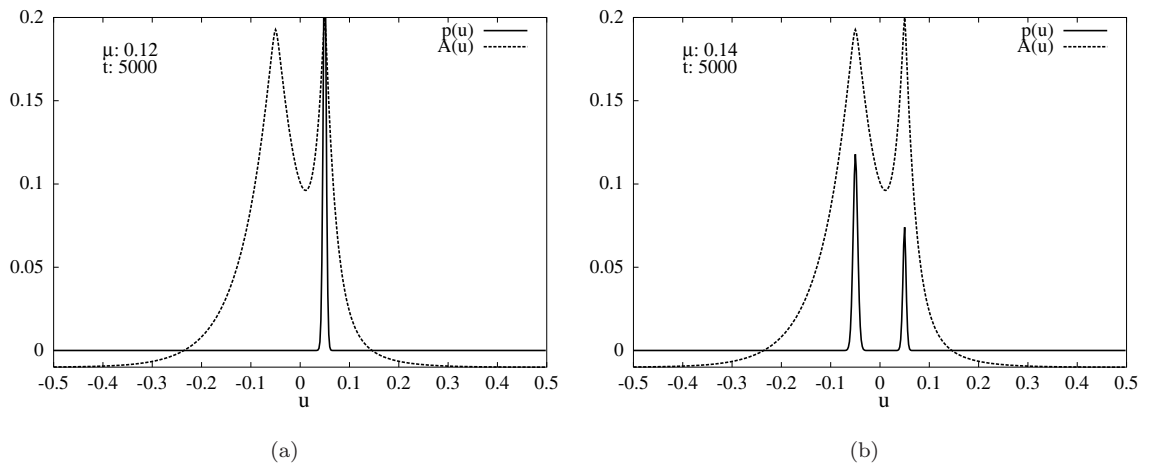


Fig. 3. Mutation-induced speciation. A two peaks static fitness landscape, increasing the mutation rate we pass from a single quasi-species population ((a), $\mu = 0.12$) to the coexistence of two quasi-species ((b), $\mu = 0.14$).

nonoverlapping Gaussians (labeled 1 and 2):

$$p(u) = \frac{\gamma_1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{u-\tilde{u}_1}{2\sigma_1^2}\right) + \frac{\gamma_2}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{u-\tilde{u}_2}{2\sigma_2^2}\right),$$

with $\gamma_1 + \gamma_2 = 1$.

Due to the nonoverlapping conditions, $\langle A\rangle = \gamma_1\langle A_1\rangle + \gamma_2\langle A_2\rangle \simeq \gamma_1 A_1(1-\sqrt{a_1\mu}) + \gamma_2 A_2(1-\sqrt{a_2\mu})$. By averaging Eq. 2.4 over distribution 1, we get

$$\gamma_1' = \frac{\langle A_1\rangle}{\gamma_1\langle A_1\rangle + \gamma_2\langle A_2\rangle}\gamma_1$$

and the condition for which distribution 2 will disappear is

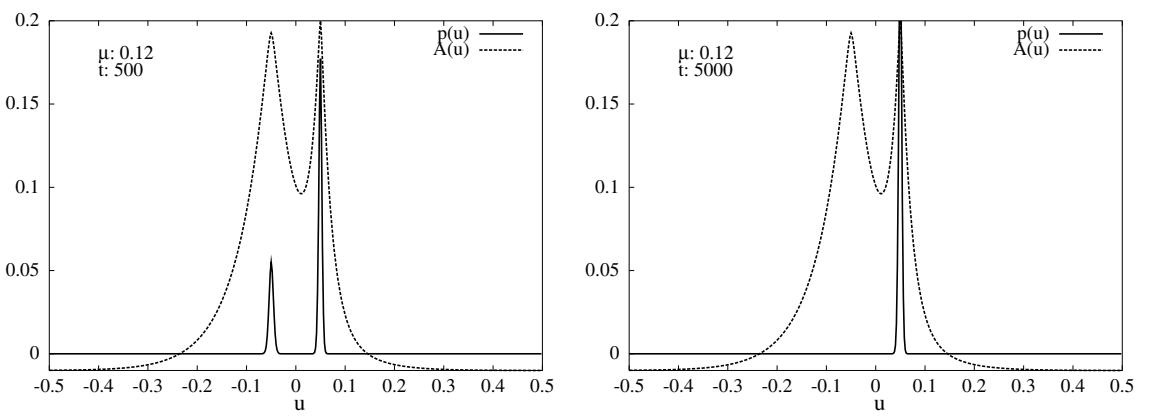
$$A_1 - A_2 > (A_1(1-\sqrt{a_1\mu}) - A_2(1-\sqrt{a_2\mu}))\sqrt{\mu}.$$

We clearly need $A_1 > A_2$ and $a_1 > a_2$. In this condition, it may happen that by increasing μ the previous condition will fail. The critical value μ_c of the mutation rate is

$$\mu_c = \left(\frac{A_2 - A_1}{A_2\sqrt{a_2} - A_1\sqrt{a_1}}\right)^2.$$

This coexistence depends crucially on the mutation rate μ. If μ is too small, the quasi-species occupying the broader niche disappears; if the mutation rate is too high the reverse happens. In this case, the difference of fitness establishes the time scale, which can be quite long. In presence of a fluctuating environment, these time scales can be long enough that the extinction due to global competition is not relevant. A transient coexistence is illustrated in Fig. 4. One can design a special form of the landscape that allows the coexistence for a finite interval of values of μ, but certainly this is not a generic case. This condition is known in biology under the name of Gause or *competitive exclusion* principle [2].

As shown in Reference [23], the existence of a degenerate effective fitness is a generic case in the presence of competition, if the two species can co-adapt before going extinct. The short-range competition lowers the fitness of the maximum, as shown in Fig. 5.



(a) (b)

Fig. 4. Evolution on a two-peaks static fitness landscape, after 500 (a) and 5000 (b) time steps. For a transient period of time the two species co-exist, but in the asymptotic limit only the fittest one survives.

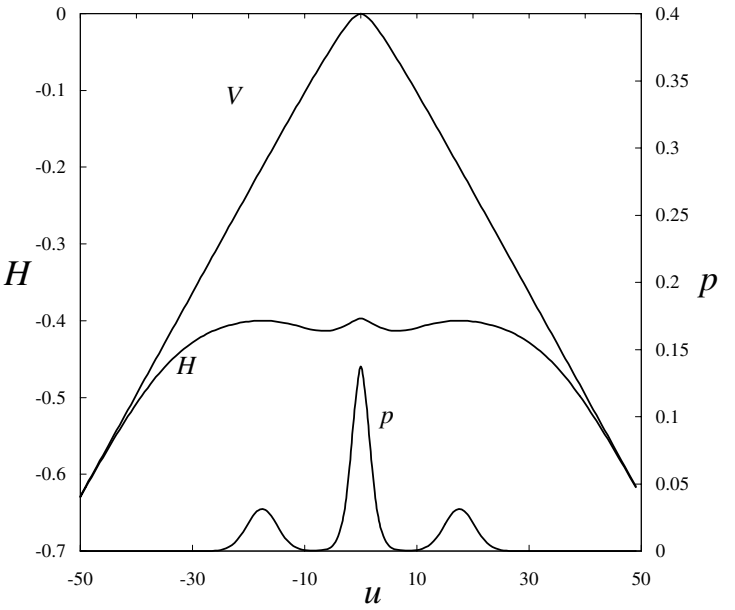


Fig. 5. Static fitness V, effective fitness H, and asymptotic distribution p numerically computed for the following values of parameters: $\alpha = 2$, $\mu = 0.01$, $V_0 = 1.0$, $b = 0.04$, $J = 0.6$, $R = 10$, $r = 3$ and $N = 100$.

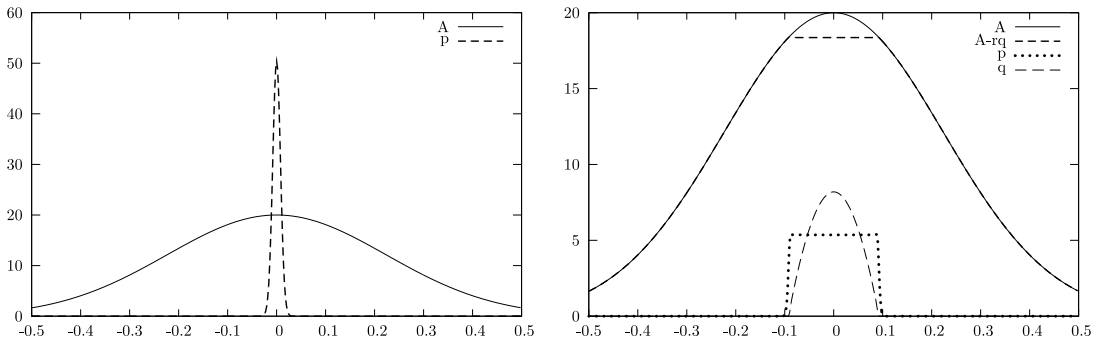
This short-range competition may originate by the presence of preys or predators. Let us we suppose there exists for instance predators or parasites that evolve so rapidly that they can be considered at equilibrium, and that each predator may exploit only the resources in a certain phenotypic range. Examples are for instances viruses that use some receptors, or the immune system that uses partial match to recognize invaders. The abundance of a certain phenotype triggers the increase of the related parasites and this is felt by nearby phenotypes as a short-range competition factor. Similar analysys apply competition due to the shortage of a given prey phenotype.

5. Positive Selection

We have considered up to now only the "bare" mutation rate, originated for instance by chemicals or cosmic rays. However, any replicating organism uses a correction mechanism for reducing this rate. This correction mechanism is coded into the genotype. In the case of RNA viruses, the main source of errors is the reverse transcription phase, which is influenced by the host environment. Therefore, in many cases the actual mutation rate is a function of the genotype, and under the influence of the selection mechanism [24]. In the presence of predation or parasitism of a faster-evolving species, it may become convenient to increase the variability of the phenotype.

This phenomenon may be illustrated by a simple model,

$$\begin{aligned} p'(x) &= \frac{A(x) - rq(x)}{\sum_{x'}(A(x') - rq(x'))p(x')}p(x) - \mu_p \frac{\partial \delta^2 p(x)}{\delta x^2}, \\ q'(x) &= \frac{p(x)}{\sum_{x'}(q(x'))p(x')}q(x) - \mu_q \frac{\partial \delta^2 q(x)}{\delta x^2}, \end{aligned} \tag{5.1}$$



(a) (b)

Fig. 6. (a) Static fitness A and asymptotic distribution of preys p in the absence of predation ($\mu_p = 0.01$); (b) Static fitness A, asymptotic distribution of preys p and of predators q, effective fitness of preys $A - rq$. Numerical values: $A_0 = 20$, $r = 0.2$, $\mu_q = 0.3$, $a = 10$.

where x is the phenotypic index ($x = 0$ corresponds to the wild type), $p(x)$ is the distribution of preys/hosts at time t and $q(x)$ that of predators/parasites. $A(x)$ is the fitness of preys in the absence of predators, and $rq(x)$ is the load due to the latters. Primes denote the quantities at the next time step and mutations are modeled as a simple diffusion process. We consider the case $A(x) = A_0 \exp(-ax^2) \simeq A_0(1 - ax^2)$ for small x.

The asymptotic state of the system $p' = p$ and $q' = q$ (see Fig. 6) in the limit of vanishing mutation rates is given by a flat distribution

$$p(x) = \begin{cases} \dfrac{1}{2L} & \text{for } -L < x < L, \\ 0 & \text{otherwise.} \end{cases}$$

Consisting of many phenotypes with the same fitness up to a width L (to be computed below). The distribution of $q(x)$ is linearly related to the fitness $A(x)$ (so that $A(x) - rq(x)$ is constant). By imposing the normalization condition $\int q(x)\, \mathrm{d}x = 1$ we get

$$L = \left(\frac{3r}{4aA_0} \right)^{\frac{1}{3}}.$$

In this way we obtain a broad distribution of prey phenotypes, whose abundance is unrelated to the fitness, while it is the predator distribution that is strongly correlated to the fitness of the prey.

In the real world, the effect of increasing mutation rate may be reached by using less-realible genetic manipulation as the reverse transcriptasi in RNA viruses. However, generic mutation lowers too much the viability, leading to the mutational meltdown (see Sec. 6). It is therefore expected the development of *positive* selection of viable variants by means of specific mechanisms that favors the mutability of selected portions of the phenotype, such as the hypermutation mechanism of the immune system.

The process of speciation requires the divergence of the genetic determinants of phenotypes; the selected characteristics may be identified by scanning genes for positive selected amino acids. While negative selection represents the conservation of functional and structural elements, positive or directional selection confers a fitness benefit to phenotype changes.

Positive selection is characterized by the quick emergence in a population of a species of a new allele with larger fitness advantage than the other alleles in the same population and has no advantage in the population of other species. Genes that contain positively selected mutations, yet remain constrained or selectively neutral in other closest species may offer insight into the significant genetic changes affecting phenotype difference. Positive selection is often limited to broadening frequencies in a small set of sites of a gene; their identification may highlight functionally important protein regions meaningful for the relationships between ecological and molecular change (for example viral proteins escaping immune system response). There are a variety of tests for detecting departures from neutral selection. One of the most used is Tajima's test which is based on the distribution of allele frequencies and/or segregating sites.

A second category of methods, based on Maximum Likelihood and Bayesian inference, identifies specific sites at which adaptive mutations occur by comparing nonsynonymous (i.e., amino acid change) to synonymous (i.e., nonchanging the amino acid) substitution rates with respect to a phylogenetic tree, i.e., taking into account the statistical relationships among sequences, in order to properly weight for multiple comparisons (). Recent methods consider as evidence for molecular adaptations how conserved, or radically different, nonsynonymous mutations are with respect to some important amino acid chemical physics properties such as charge and volume. Strongly positively selected genes are involved in a host's immune response to pathogens, or in a pathogen's evasion of this response (they include the human major histocompatibility complex (MHC)), vision and olfaction, neural development, and in male reproduction. Such genes are affected by sexual selection or sperm competition.

There is a strong link between positive selection, fitness and antiviral drug resistance arised during therapies for most viruses, particularly for Hiv. A selection pressure maps of HIV proteins involved in drug resistance provide clues on estimating probability of arising mutations contributing to drug resistance, distinguishing primary drug resistance mutations from accessory mutations, rate measurements of fast versus slow evolutionary pathways to multiple drug resistance, and the evolutionary dynamics of different types of mutations as the virus moves from untreated to drug-treated conditions and back during different regimen of therapy. The large increase of positive selection mutations occurring in hiv infected treated with drugs with respect the untreated is reported for example in [25].

6. Error Threshold and Muller's Ratchet

Viruses, and particularly those with small RNA genomes, often show a quasi species behavior. Important insights into the fitness landscape of viruses are coming from the analysis of Hiv variants in infected individuals; anti viral therapies influence the fitness landscape: usually drug resistance mutations reduce replication fitness and drug-sensitive viruses rapidly evolve after complete treatment interruption. The high genetic diversity in Hiv is caused by the high virion production rate (10^{10} daily), the short generation time and the error-prone reverse transcriptase, which generates an average of 10^5 mutations per site and generation. Time series analysis of Hiv strain abundances evolution in infected individuals shows that during the therapy treatment failure almost 100% of the virus population displays resistance mutations, while during treatment interruption there is the almost complete disappearance of resistance mutations from plasma virus. This suggests that the viral populations had

undergone genetic bottlenecks during the development and reversion of resistance and that treatment interruption results in reappearance of hidden strains in reservoirs, rare wild-type variants (which had very low fitness during therapy) or reversion of resistance through continued evolution. Interestingly, prolonged treatment failure causes a decrease of the number of viral variants suggesting the existence of an effective genetic bottle-neck [26].

The immune systems and pathogens have evolved with prey-predators exchange role dynamics. The human protein ABOBEC3G hypermutates the viral genome, via deamination, while the virus is reverse-transcribing its RNA genome to DNA. This hypermutational effect of ABOBEC3G on the virus genome could be though as leading to a mutational meltdown which produce mostly non viable viral strains. Note that during the viral life cycle, APOBEC3G is incorporated into the viral capsid. Typically, it is subsequently degraded by the virus's Vif protein, which is effectively the antidote to APOBEC3G. However, the imprecise inactivation of APOBEC3G by Vif causes a large spectrum of mutations to occur in Hiv, which allow the virus to finely explore the fitness landscape modified by the anti viral therapies. The analysis of strain abundances and variability during the antiviral drugs may provide an interesting mean of estimating the shape of the fitness landscape.

In order to present a modeling approach, let us consider a *sharp peak landscape*: the phenotype $u_0 = 0$, corresponding to the master sequence genotype $x = 0 \equiv (0, 0, \dots)$ has higher fitness $A_0 = A(0)$, and all other genotypes have the same, lower, fitness A_*. Due to the form of the fitness function, the dynamics of the population is fundamentally determined by the fittest strains.

Let us indicate with $n_0 = n(0)$ the number of individuals sharing the master sequence, with $n_1 = n(1)$ the number of individuals with phenotype $u = 1$ (only one bad gene, i.e., a binary string with all zero, except a single 1), and with n_* all other individuals. We assume also nonoverlapping generations,

During reproduction, individuals with phenotype u_0 can mutate, contributing to n_1, and those with phenotype u_1 can mutate, increasing n_*. We disregard the possibility of back mutations from u_* to u_1 and from u_1 to u_0. This last assumption is equivalent to the limit $L \to \infty$, which is the case for existing organisms. We consider only short-range mutation with probability μ_s. Due to the assumption of large L, the multiplicity factor of mutations from u_1 to u_* (i.e., $L - 1$) is almost the same of that from u_0 to u_1 (i.e., L).

The evolution equation (discrete time) of a population of N individuals with carrying capacity K [11, 27] is

$$\begin{aligned} n_0' &= \left(1 - \frac{N}{K}\right)(1 - \mu_s) A_0 n_0, \\ n_1' &= \left(1 - \frac{N}{K}\right)((1 - \mu_s) A_* n_1 + \mu_s A_0 n_0), \\ n_*' &= \left(1 - \frac{N}{K}\right) A_*(n_* + \mu_s n_1). \end{aligned} \tag{6.1}$$

and

$$\langle A \rangle = \frac{A_0 n_0 + A_*(n_1 + n_*)}{N}$$

is the average fitness of the population.

The steady state of Eq. (6.1) is given by $\boldsymbol{n}' = \boldsymbol{n}$. There are three possible fixed points $\boldsymbol{n}^{(i)} = \left(n_0^{(i)}, n_1^{(i)}, n_*^{(i)}\right)$: $\boldsymbol{n}^{(1)} = (0,0,0)$ $(N^{(1)} = 0)$, $\boldsymbol{n}_2 = (0,0,K(1-1/\langle A_*\rangle))$ $(N^{(2)} = n_*^{(2)})$ and

$$\boldsymbol{n}^{(3)} = \begin{cases} n_0^{(3)} &= N^{(3)} \dfrac{(1-\mu_s)A_0 - A_*}{A_0 - A_*}, \\ n_1^{(3)} &= N^{(3)} \dfrac{\mu_s}{1-\mu_s} \dfrac{A_0((1-\mu_s)A_0 - A_*)}{(A_0 - A_*)^2}, \\ n_*^{(3)} &= N^{(3)} \dfrac{\mu_s^2}{1-\mu_s} \dfrac{A_0 A_*}{(A_0 - A_*)^2}, \\ N^{(3)} &= 1 - \dfrac{1}{A_0(1-\mu_s)}. \end{cases}$$

The fixed point $\boldsymbol{n}^{(1)}$ corresponds to extinction of the whole population, i.e., to mutational meltdown (MM). It is trivially stable if $A_0 < 1$, but it can become stable also if $A_0 > 1$, $A_* < 1$ and

$$\mu_s > 1 - \frac{1}{A_0}. \tag{6.2}$$

The fixed point $\boldsymbol{n}^{(2)}$ corresponds to a distribution in which the master sequence has disappeared even if it has larger fitness than other phenotypes. This effect is usually called Muller's ratchet (MR). The point P_2 is stable for $A_0 > 1$, $A_* > 1$ and

$$\mu_s > \frac{A_0/A_* - 1}{A_0/A_*}. \tag{6.3}$$

The fixed point $\boldsymbol{n}^{(3)}$ corresponds to a coexistence of all phenotypes. It is stable in the rest of cases, with $A_0 > 1$. The asymptotic distribution, however, can assume two very different shapes. In the quasi-species (QS) distribution, the master sequence phenotype is more abundant than other phenotypes; after increasing the mutation rate, however, the numeric predominance of the master sequence is lost, an effect that can be denoted error threshold (ET). The transition between these two regimes is given by $n_0 = n_1$, i.e.,

$$\mu_s = \frac{A_0/A_* - 1}{2A_0/A_* - 1}. \tag{6.4}$$

Our definition of the error threshold transition needs some remarks: in Eigen's original work [28, 29] the error threshold is located at the maximum mean Hamming distance, which corresponds to the maximum spread of population. In the limit of very large genomes these two definitions agree, since the transition becomes very sharp [30]. See also [31, 32].

In Fig. 7(a) we reported the phase diagram of model (6.1) for $A_* > 1$ (the population always survives). There are three regions: for a low mutation probability μ_s and high selective advantage A_0/A_* of the master sequence, the distribution has the quasi-species form (QS); increasing μ_s the distribution undergoes the error threshold (ET) effect; finally, for very high mutation probabilities, the master sequence disappears and we enter the Muller's ratchet (MR) region [33, 34]. The error threshold phase transition is not present for smooth landscapes (for an example of a study of evolution on a smooth landscape, see [35]).

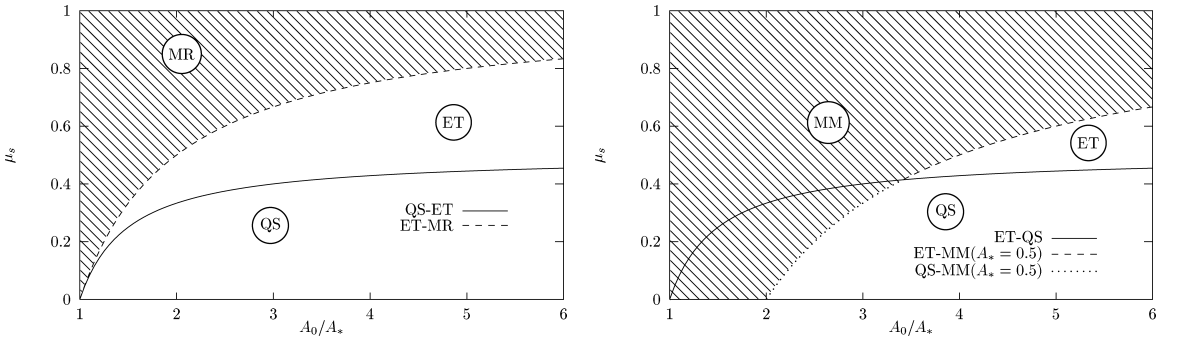


Fig. 7. (a) Phase diagram for the error threshold and Muller's ratchet transitions ($A_* > 1$). MR refers to the Muller's ratchet phase ET to the error threshold distribution and QS to quasi-species distribution. The phase boundary between the Muller's ratchet effect and the error threshold distribution (Eq. (6.2)) is marked ET-MR; the phase boundary between the error threshold and the quasi-species distribution (Eq. (6.4)) is marked QS-ET. (b) Phase diagram for the mutational meltdown extinction, the error threshold and the quasi-species distributions ($A_* < 1$). MM refers to the mutational meltdown phase, ET to the error threshold distribution and QS to quasi-species distribution. The phase boundary between the Mutational meltdown effect and the error threshold distribution (Eqs. (6.3) and (6.6)) is marked ET-MM; the phase boundary between the mutational meltdown and the quasi-species distribution (Eqs. (6.3) and (6.5)) is marked QS-MM.

In Fig. 7(b) we illustrate the phase diagram in the case $A_* = 0.5$. For a low mutation probability μ_s and high selective advantage A_0/A_* of the master sequence, again one observes a quasi-species distribution (QS), while for sufficiently large μ_s there is the extinction of the whole population due to the mutational meltdown (MM) effect. The transition between the QS and MM phases can occur directly, for

$$A_0/A_* < \frac{1-\sqrt{1-A_*}}{A_*} \tag{6.5}$$

(dotted QS-MM line in figure): during the transient before extinction the distribution keeps the QS form. For

$$A_0/A_* > \frac{1-\sqrt{1-A_*}}{A_*}, \tag{6.6}$$

one has first the error threshold transition (QS-ET line in figure), and then one observes extinction due to the mutational meltdown effect (dashed ET-MM line in figure). This mutation-induced extinction has been investigated numerically in [33]. The Error threshold for finite populations has been studied in [34, 36, 37].

7. Application to HIV Immunology

The theoretical immunology represents an important benchmark for modleing genotype-phenotype dynamics. Here we further discuss the phenotype — genotype mapping by framing it using a model which takes into account three components, the number of uninfected (T) and Hiv infected CD4+ T lymphocytes, and the number of free Hiv viruses (V). The immune system increases the fitness of the virus when the virus acts as a CD4 predator; it decreases the fitness when immune system cells act as Hiv predator (by killing infected cells,

or using specific antibodies). Given that the virus is quickly speciating in multi different quasi species, these will have slightly different fitness; the fitness advantage of a quasispecie will be affected by therapies and resources (CD4) competition.

A multi species modeling of Hiv virus could be described by the following set of equations, which represent an extension of a model by Perelson *el al.* [38].

$$\dot{T}_i = \left(\lambda_i + \sum_k \gamma_{ik}^{(T)} I_k T_i\right)\left(1 - \frac{1}{K}\sum_i T_i\right) - \left(\delta_T + \sum_k \beta_k V_k\right) T_i, \tag{7.1}$$

$$\dot{I}_k = \left(\sum_{k'} \mu_{kk'}\beta_{k'} V_{k'}\right)\left(\sum_i T_i\right) - \left(\delta_I + \sum_i \gamma_{ki}^{(I)} T_i\right) I_k, \tag{7.2}$$

$$\dot{V}_k = \pi I_k - \left(c + \sum_i \gamma_{ki}^{(V)} T_i\right) V_k. \tag{7.3}$$

The model considers the following cell types: T-helper (CD4+) cells responding to virus strain i, (T_i); T cells (any strain) infected by virus strain k, (I_k); abundance of viral strain k, (V_k). This means that viral strain k are identified by just one epitope, which is then displayed on the surface of the T cell of class k, and that a T cell of class i can be activated at least by one CD4+ T cell carrying the epitope k, which is specific of the viral strain k. The indices i (k) range from 1 to N_i (N_k), and in the following we have used $N_i = N_k = N$. This model allows to investigate time scale of quasi species evolution during superinfection (multiple infections at different times) or coinfection (simultaneous infection by different strains). For further details see [39].

8. Challenges in Fitting Nonlinear Models to Observed Data

Our approach has been to focus on building simple models which use as fewer parameters as possible that could be effective in describing the system under study; we should address the question: how simple models could still provide good fitting with real data? The large abundance of molecular data from many species (particularly bacteria and viruses) through next generation sequencing, provides the basis for evaluating mathematical models with real data. First, it is important to distinguish between "inverse" modeling, also called model calibration, in contrast to "forward" modeling approach which is used for forecasting or hypothesis testing. If there is scarcity of data, for example only parts of genetic and biochemical networks can be observed directly, not all the model parameters can be estimated uniquely.

Model calibration is the process of estimating these parameters by fitting the model to observed data as a nonlinear optimization problem. An important challenge is the estimation of nonidentifiable parameters which cannot be unambiguously determined with sufficient precision mainly due to correlations with other parameters. Note that, together with the best values estimated for a parameter we would like to estimate parameter uncertainties which lead to a measure of the robustness of model generated predictions.

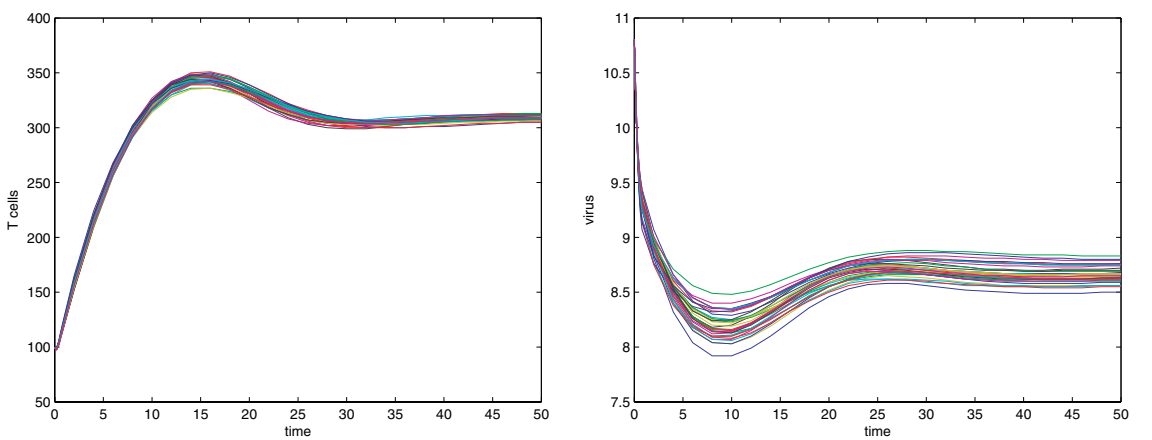
When implementing robust predictive analyses, the integrals over high-dimensional parameter spaces can neither be evaluated analytically, nor numerically in a straight-forward way. Although inference techniques, such as Maximum likelihood, are relatively easy to

implement, they suffer from quite few drawbacks, such as not been able to fully explore the entire parameter space. As partial solutions to this problem, solutions of differential equation could be approximated by nonparametric functions, which are estimated by penalized smoothing with penalties defined by the differential equation behaviors. The currently used inference methods have substantial limitations upon the form of models that can be fitted and, hence, upon the nature of the scientific hypotheses that can be made and the data that can be used to evaluate them. Instead, the so-called plug-and-play methods require only simulations from a model and are thus free from such restrictions.

Unfortunately the Bayesian methodology which provides a powerful inferential framework becomes computationally intensive when the amount of data is limited. Some useful computational tools are: Laplace's method of asymptotic approximation and Markov Chain Monte Carlo (MCMC) methods, including multi-level Metropolis-Hastings algorithms with tempering, the Gibbs sampler and the Hybrid Monte Carlo algorithm. Particle filter algorithms are useful for sequential Bayesian state estimation when the Kalman filter is not applicable because of nonlinear dynamics and/or nonGaussian probability models.

A partial availability of data would still allow to compute sensitivity, parameter identifiability, model fitting and estimate parameter confidence intervals through a Markov-chain based method. First, given the lack of quasi species data, we focus on single species, i.e. we drop the subscript of the variables V, I and T. We then use a MCMC procedure to estimate the effect of the parameter uncertainty on the model output (see Fig. 8). We take as input the sample of the parameter probability density function generated from the marginal distributions for each parameter (λ,β, δ, $\gamma^{(I)}$, $\gamma^{(T)}$).

Note that the arising of quasi species and the corresponding formation of immune cells specific for the different quasi species describe a simple mechanism for the generation of complexity. Ronald Fisher in early 1930 proposed that the more complex a plant or animal, the more difficulty it should have adapting to changes in the environment. This hypothesis seems strongly violated when observing how well-adapted, complex organisms — from orchids to bower birds to humans, are. Here we show that complexity can be easily



(a) (b)

Fig. 8. Sensitivity analysis of CD4+ T cells (a) and virus load (b) strain differences based on parameter distribution generated using a MCMC statistics. Data from [40].

generated through evolutionary processes and provide organisms to adapt to environmental change.

9. Conclusions and Potentialities

In this work we have stressed how statistical mechanics and dynamical system approaches allow evolutionary biology to develop more predictive and robust theory. The shift from allele data, the basis of classic evolutionary genetics, to sequence, gene expression and in the near future, biochemical and structure data, and hopefully dynamical aspects will allow evolutionary biology to develop more predictive and robust theory.

We have presented some simple models of evolution in phenotypic space. We start from the consideration that viruses represent the best opportunity to measure with high degree of accuracy, mutation rates, selection pressure and fitness, therefore most likely offering the "hopeful monsters", i.e., events of instantaneous speciation, theoretical biologists are searching for. In particular viruses provide meaningful examples of quasi species.

The intensive research effort over the past twenty years on the human immunodeficiency virus type 1 (HIV-1), which is the causative agent of the global AIDS pandemic, has led to the generation of a vast amount of clinical and sequence data on viral strains and on the action of several classes of effective anti-HIV-1 drugs that have significantly improved patient survival. One example is given by the Vif-APOBEC3G which could be modeled as a mutational meltdown system. Future direction could establish the fitness landscape generated by different timing and the kinds of antiviral drugs selected for inductionmaintenance and therapy-intensification strategies or the imperfect adherence to HIV induction therapy or to a personalised therapy.

The need for a nonlinear mathematical formulation of evolution is highlighted by recent experiments and theoretical advances showing that in some cases the mutation rate could be fitness-dependent [41]. Clearly, the ideal goals would be that of showing how macroevolutionary patterns may arise from a (not too over) simplified individual-based dynamics. However, evolutionary systems tend to develop highly correlated structure, so that it is difficult to operate the scale separation typical of simple physical system (say, gases). Nevertheless, here we show simple models could catch important aspects, in order to test the robustness of many hypotheses.

Acknowledgments

PL thanks the support of EC IST SOCIALNETS — Grant agreement number 217141.

References

[1] M. Guillaume, E. F. Santiago and L. Thomas, Distributions of epistasis in microbes fit predictions from a fitness landscape perelsonl, *Nature Genetics* **39** (2007) 555–560.
[2] G. F. Gause, *The Struggle for Existence* (Williams & Wilkins, Baltimore, MD 1934); G. Hardin, The competitive exclusion principle, *Science* **131** (1960) 1292–1297.
[3] E. J. Vallender and B. T. LahnHum, Positive selection on the human genome, *Hum. Mol. Genet.* **13** (2004) R245–R254.
[4] P. Lió and M. Bishop, Modeling sequence evolution, *Methods Mol Biol.* **452** (2008) 255–285.
[5] P. Lió, Dimensionality and dependence problems in statistical genomics, *Brief Bioinform.* **4** (2003) 168–177.

[6] S. Whelan, P. Lió and N. Goldman, Molecular phylogenetics: State-of-the-art methods for looking into the past, *Trends Genet.* **17** (2001) 262–272.
[7] P. Lió, Investigating the relationship between genome structure, composition, and ecology in prokaryotes, *Mol. Biol. Evol.* **19** (2002) 789–800.
[8] P. Lió and N. Goldman, Phylogenomics and bioinformatics of SARS-CoV, *Trends Microbiol.* **12** (2004) 106–111.
[9] M. W. Nachmana and S. L. Crowella, Estimate of the mutation rate per nucleotide in humans, Genetics **156** (2000) 297–304.
[10] F. Bagnoli and M. Bezzi, Small world effects in evolution, *Phys. Rev. E* **64** (2001) 021914.
[11] F. Bagnoli and M. Bezzi, An evolutionary model for simple ecosystems, in: *Annual Reviewies of Computational Physics VII*, ed. D. Stauffer (World Scientific, Singapore, 2000), pp. 265–310.
[12] F. Bagnoli, Evolutionary models for simple biosystems, in Handbook on Biological Networks, S. Boccaletti, V. Latora and Y. Moreno, eds., World Scientific Lecture Notes in Complex Systems, Vol. 10 (World Scientific, Singapore, 2010), pp. 329–372.
[13] L. Peliti, Introduction to the statistical theory of Darwinian evolution, Lectures at the Summer College on Frustrated System, Trieste, August 1997, cond-mat/9712027.
[14] M. Lynch and W. Gabriel, Mutation load and the survival of small populations, *Evolution* **44** (1990) 1725–1737.
[15] M. Lynch, R. Burger, D. Butcher and W. Gabriel, The mutational meltdown in asexual populations, *J. Hered.* **84** (1993) 339–344.
[16] A. T. Bernardes, Mutational meltdown in large sexual populations, *J. Physique* **I5** (1995) 1501–1515.
[17] P. G. Higgs and G. Woodcock, The accumulation of mutations in asexual populations, and the structure of genealogical trees in the presence of selection, *J. Math. Biol.* **33** (1995) 677–702.
[18] G. Woodcock and P. G. Higgs, Population evolution on a multiplicative single-peak fitness landscape, *J. Theor. Biol.* **179** (1996) 61–73.
[19] C. Amitrano, L. Peliti and M. Saber, Population dynamics in a spin-glass model of chemical evolution, *J. Mol. Evol.* **29** (1989) 513–525.
[20] L. Peliti, Fitness landscapes and evolution, in *Physics of Biomaterials: Fluctuations, Self-Assembly and Evolution*, eds. T. Riste and D. Sherrington (Kluwer, Dordrecht 1996), pp. 287–308.
[21] R. A. Fisher, *The Genetical Theory of Natural Selection* (Dover, New York, 1930).
[22] J. A. Anderson *et al.*, HIV-1 Populations in semen arise through multiple mechanisms, PLoS Pathog **6** (2010) e1001053; J. F. Salazar-Gonzalez *et al.*, Genetic identity, biological phenotype, and evolutionary pathways of transmitted/founder viruses in acute and early HIV-1 infection, *J. Exp .Med.* **206** (2009) 1273–1289.
[23] F. Bagnoli and M. Bezzi, Speciation as pattern formation by competition in a smooth fitness landscape, *Phys. Rev. Lett.* **79** (1997) 3302–3306.
[24] A. Sasaki and M. A. Nowak, Mutation landscapes, *J. Theor. Biol.* **224** (2003) 241–247.
[25] L. Chen and C. Lee, Distinguishing HIV-1 drug resistance, accessory, and viral fitness mutations using conditional selection pressure analysis of treated versus untreated patient samples, Biol Direct. **1** (2006) 14.
[26] C. Hedskog, *et al.*, Dynamics of HIV-1 quasispecies during antiviral treatment dissected using ultra-deep pyrosequencing. PLoS ONE **5** (2010) e11345.
[27] F. Bagnoli and M. Bezzi, Eigen's error threshold and mutational meltdown in a quasispecies model, *Internat. J. Modern Phys. C* **9** (1998) 555–562.
[28] W. Eigen, Selforganization of matter and evolution of biological Macromolecules, *Naturwissenshaften* **58** (1971) 465–523.
[29] W. Eigen and P. Schuster, The hypercycle: A principle of natural self-organization, *Naturwissenshaften* **64** (1977) 541–552.
[30] S. Galluccio, Exact solution of the quasispecies model in a sharply peaked fitness landscape, *Phys. Rev. E* **56** (1997) 4526–4539.

[31] E. Baake and T. Wiehe, Bifurcations in haploid and diploid sequence space models, *J. Math. Biol.* **35** (1997) 321–343.
[32] H. Wagner, E. Baake and T. Gerisch, Ising quantum chain and sequence evolution, *J. Stat. Phys* **92** (1998) 1017–1052.
[33] K. Malarz and D. Tiggemann, Dynamics in Eigen quasispecies model, *Internat. J. Modern Phys. C* **9** (1998) 481–490.
[34] T. Wiehe, E. Baake and P. Schuster, Error propagation in reproduction of diploid organisms. A case study on single peaked landscapes, *J. Theor. Biol.* **177** (1995) 1–15.
[35] L. S. Tsimring, H. Levine and D. A. Kessler, RNA virus evolution via a fitness-space model, *Phys. Rev. Lett.* **76** (1996) 4440–4443; D. A. Kessler, H. Levine, D. Ridgway and L. Tsmiring, Evolution on a Smooth Landscape, *J. Stat. Phys.* **87** (1997) 519–544.
[36] M. Nowak and P. Schuster, Error thresholds of replication in finite populations — Mutation frequencies and the onset of Muller's ratchet, *J. Theor. Biol.* **137** (1989) 375–395.
[37] D. Alves and J. F. Fontanari, Population genetics approach to the quasispecies model, *Phys. Rev. E* **54** (1996) 4048–4053.
[38] H. Wu, H. Zhu, H. Miao, A. Perelson, Parameter identiability and estimation of HIV/AIDS dynamic models, *Bull. Math. Biol.* **70** (2008) 785–799.
[39] F. Bagnoli, P. Lió and L. Sguanci, Modeling viral coevolution: HIV multi-clonal persistence and competition dynamics, *Physica A* **366** (2006) 333–346; L. Sguanci, F. Bagnoli and P. Lió, Modeling HIV quasispecies evolutionary dynamic, BMC Evolutionary Biology **7** (2007) S5.
[40] X. Xia, Estimation of HIV/AIDS parameters, *Automatica* **5** (2003) 1983–1988.
[41] C. F. Baer, Does mutation rate depend on itself? *PLoS Biol* **6** (2008) e52.