

Evolutionary dynamics and statistical physics

Daniel Fisher¹, Michael Lässig² and Boris Shraiman³

¹ McCullough Laboratory, Department of Applied Physics,
Stanford University, CA 94305, USA

² Institut für theoretische Physik, University of Cologne, D-50937 Cologne,
Germany

³ Kavli Institute for Theoretical Physics and Department of Physics,
University of California, Santa Barbara, CA 91306, USA

E-mail: dsfisher@stanford.edu, mlassig@uni-koeln.de
and shraiman@kitp.ucsb.edu

Received 18 January 2013

Accepted 18 January 2013

Published 12 February 2013

Online at stacks.iop.org/JSTAT/2013/N01001

[doi:10.1088/1742-5468/2013/01/N01001](https://doi.org/10.1088/1742-5468/2013/01/N01001)

Abstract. This introductory article provides the background to and motivation for this special issue and the relationship between evolutionary dynamics and statistical physics.

Contents

1. Statistics of evolution: why now, why here?	2
1.1. From experiments to fitness landscapes	4
1.2. Statistics of adaptive processes	5
1.3. Sexual and spatial effects	6
References	8

1. Statistics of evolution: why now, why here?

This special issue is focused on quantitative approaches to the dynamics of biological evolution. This field—population genetics and its relatives—is currently undergoing a burst in its own evolution. It has always been focused not on the natural history aspects of evolution, but on the dynamics of genetic change in populations that is driven by natural selection and chance. There is a century-long mathematical tradition that started with the integration of Darwinian evolution and Mendelian genetics. Its lineage includes Fisher, Haldane, Wright, and Kimura, who mapped out many of the conceptual and statistical issues, without close contact with the parallel development of statistical physics in the 20th century.

The current resurgence of quantitative evolutionary theory is driven by the tremendous progress of DNA sequencing technology. At first, sequencing of individual genes and then whole genomes provided data on the genetic differences between species that had accumulated on geological time scales. This supplemented, and to a remarkable extent agreed with, the phylogenetic trees inferred from properties of the organisms in the fossil record and extant species. But the incredible drop in costs of sequencing—by more than six orders of magnitude even since the human genome was sequenced—has made possible far more. It is now routine to sequence many members of the same population revealing the genetic variation that distinguishes these individuals from each other. Moreover, genetic changes can be tracked in real time: sequencing is providing the first direct observations of the dynamical heart of evolution.

For example, sequencing from the population of HIV viruses within a single patient at multiple times provides a direct glimpse of the dynamics of viral evolution, which the high mutation rate of HIV causes to take place on time scales of weeks to years. The evolution of the viral population is countered by the host's adaptive immune system which itself works by the rapid evolution of a small portion of the DNA of white blood cells to produce new antibody proteins and receptors for detection of invaders. The outcome of this evolutionary battle determines the progression of AIDS. More generally, acquisition of drug resistance by pathogens, global spread of diseases, and the development of cancers are all evolutionary processes. Thus, a better understanding of evolutionary dynamics is crucial for medicine. And it is also for responding to the effects of climate change on the biosphere.

On the fundamental science side, we are faced with a great opportunity: to take the study of evolution from its largely historical and qualitative state to a field with

understanding built up from observations, laboratory experiments, phenomenology, and—crucially—quantitative theory. Many of the aspects of evolution that are least understood are intrinsically quantitative and statistical. The basic processes of evolution are sufficient to evolve anything given enough time. But very little is understood about what sets the time scales of evolution, whether for antibiotic resistance, emergence of new animal species, or the diversification of life from the first cells. One is usually told that puzzlement about how the complexities of life arose can be swept away by the sheer magnitude of geologic time. If selection on existing variation is the limiting process, as Darwin emphasized, then this might be reasonable. But selection can act only on differences within a population, which are continuously generated by mutations and recombination. If that process is the limiting factor for evolution, then other huge and tiny parameters, such as population sizes, lengths of genomes, mutation and recombination rates, and the size of the Earth, are surely important. But, in contrast to the cases for most fields of physics, it is not even known which combinations of these or other less obvious parameters determine evolutionary time scales. One of the valuable roles for physicists is in framing such quantitative questions which are often given short shrift in contemporary evolutionary literature. For, as stated clearly by Haldane in 1924 [1]: ‘A satisfactory theory of natural selection must be quantitative. In order to establish the view that natural selection is capable of accounting for the known facts of evolution, we must show not only that it can cause a species to change, but that it can cause it to change at a rate which will account for present and past transmutations.’

While the deluge of sequencing data may answer the ‘why now’ question, why is this journal the right venue for work on evolutionary genetics? We argue that the new data necessitates re-examination of some key assumptions that have dominated population genetics throughout its history. Until recently, most biologists agreed on a rather simple picture of molecular evolution, based on Kimura’s neutral theory [2]. This view holds that beneficial mutations are generated sufficiently rarely that they spread in populations one at a time. Deleterious mutations are generated frequently, but are rapidly eliminated by ‘purifying’ selection. Therefore, most of the large amount of genetic variation observed within and between species was assumed selectively neutral. The dynamics of neutral variation is loosely analogous to that of an ideal gas near equilibrium. But it has become increasingly clear that many genetic variants that affect fitness arise and coexist in populations all the time. This has profound consequences for the dynamics of genomes. The emerging picture of genetic evolution is that of a strongly interacting stochastic system with large numbers of interacting components far from equilibrium. It is close to the heart of modern statistical physics.

At a qualitative level, it is clear why the dynamics of concurrent beneficial mutations are coupled. Mutations on different genomes interfere with each other’s spread through a population, while those on the same genome can enhance or impede each other by fitness interactions between them. While traditional genetics has focused on identifying specific mutations with large effects, statistical approaches address a complementary component of heritable variation: populations harbor a multitude of genetic differences, called *polymorphisms*, which have small individual effects. This line of inquiry goes back to the foundational work of Fisher in 1914 on quantitative traits that are affected by multiple genes [3]. Fisher reconciled the inheritance statistics of quantitative traits with that of discrete Mendelian traits, observing that the analog properties of the former arise

from a superposition of large numbers of the latter. An important aspect of statistical genetics is the study of heritability and evolution of such extensively polymorphic traits. As demonstrated by several articles in this issue, statistical physics has an important role to play: building a coherent theory of non-neutral diversity, a problem which is just beginning to come into focus. Statistical methods are also critical for understanding the trajectories of populations over extended periods of evolutionary time. Interactions between genomic changes have again an important effect: they make the effects of genetic changes at one point in time dependent on the genetic changes in the population's past. It is these complex interactions that ultimately build biological functions on macro-evolutionary time scales. But only now can one begin to explore the statistical aspects of such interactions and their dynamical consequences via experiments, genomic data, and new theory.

The purpose of the present collection of articles is to illustrate the many connections between statistical evolutionary theory and statistical physics, to review recent progress, and to raise questions for future research. We have grouped these articles into three sections.

1.1. From experiments to fitness landscapes

The first steps in understanding the dynamics of natural selection are based on the simplifying assumption that the fitness of an organism depends only on its genome. This presumes the existence of a map from genotype to fitness, defined as the difference between the birth and death rates. Such maps are known as *fitness landscapes* [4]. This paradigm has been a traditional playground for theorists, but we have still limited empirical knowledge about the structure of fitness landscapes, whether in a general statistical sense or for specific examples.

Recent laboratory evolution experiments, starting with seminal work by Lenski's group on bacterial populations [5, 6], have produced systematic data on adaptive processes and the fitness effects of mutations that arise. A few years ago, Weinreich and colleagues produced the first concrete experimental fitness landscape, which describes the evolution of antibiotic resistance of a single bacterial protein [7]. These experiments have prompted a burst of theoretical efforts towards the more realistic description of adaptive processes. Several articles in this issue are testimony to these developments. They describe new ways of inferring fitness from data, including effects of spatial structure and of sexual recombination.

The article by Desai provides a broad survey of exciting new evolution experiments and of their implications for the inference of fitness landscapes. He discusses the challenges of linking experiments to population genetics models and pinpoints the spectrum of mutational trajectories as a crucial empirical input for dynamical models.

Illingworth and Mustonen take a somewhat complementary view, focusing on the inference of beneficial fitness effects from time series of genomic data. In particular, they address the challenging problem of telling apart the *driver mutations* of an adaptive process, which carry a large positive fitness effect, from the *passenger* mutations that hitchhike on successful genomes.

Nourmohammad, Schiffels, and Lässig look at the inference of selection acting on *quantitative traits*, which can be regarded as 'mesoscopic' observables that are controlled by many genomic sites. Binding affinities of interaction domains, folding stabilities,

and expression levels of proteins are examples of such molecular traits, quantitative information on which is now accumulating from high-throughput experiments.

Szendro *et al* analyze the adaptive dynamics on explicit fitness landscapes inferred from evolution experiments and show that the ‘Rough Mount Fuji’ model provides a useful framework for characterizing and comparing such landscapes. Their analysis applies if a limited set of mutations is the dominant target of selection, as is often the case for evolution of antibiotic resistance in bacteria. The speed of evolution in such landscapes is often limited by the small number of routes to higher fitness.

1.2. Statistics of adaptive processes

How fast are adaptive processes? This is one of the simplest quantitative questions about evolution. But in a strongly interacting system involving many beneficial mutations, answering this question turns out to be surprisingly difficult even in simple fitness landscapes. Recent research has revealed several connections to deep problems in statistical physics. Fitness is the most important quantitative trait of an organism, and its rate of change is given by Fisher’s so-called Fundamental Theorem [3]. The theorem relates, straightforwardly, the rate of increase of the average fitness of a large population to the variance of fitness within the population in situations in which new mutations can be neglected. But this only gives the initial increase and is not sufficient to determine the longer term dynamics of the moving fitness distribution: the speed of this traveling fitness wave is set by stochastic effects at its leading edge, as for various front propagation problems in statistical physics [8]. This is an example of a general rule: evolution depends in an essential way on rare mutation or recombination events—in marked distinction to most familiar problems in statistical physics. Nevertheless, in the last few years, studies of traveling fitness waves have produced fundamental new insights into the dynamics of adaptation in large and rapidly mutating populations, with and without recombination.

The new traveling-wave theory of adaptation is discussed at length in the article by Fisher. In particular, he discusses stochastic effects at the tip of the wave, which are generated by the fluctuations of the number of fittest individuals in a population. These fluctuations translate into universal properties of the adaptive process, which turn out to be quite independent of details of the local statistical properties of the fitness landscape that determine the supply of potential beneficial mutations.

The paper by Ganusov, Neher and Perelson re-examines the problem of HIV escape from virus specific cytotoxic lymphocytes (CTL), pointing out the importance of understanding the dynamics of escape from multiple simultaneous CTLs. This process requires accumulation of multiple mutations and is thus directly related to the traveling-wave theory of adaptation. The paper is an interesting link between theory and biological reality of evolutionary dynamics. The authors specifically address issues of inferring model parameters and of designing data acquisition, so as to facilitate comparisons between data and models.

Rouzine and Weinberger discuss HIV evolution within individual patients and review the relevant evolutionary models of rapid adaptive processes in asexual populations, which are characterized by the competition between clones with different beneficial mutations. They argue that treated and untreated patients show characteristic differences in the mode of HIV evolution, and discuss the consequences for antiviral drug treatment.

Another fundamental problem concerns the genealogy of a population—the phylogenetic trees of evolution writ small. Understanding genetic relatedness of individuals is absolutely essential for correct interpretation of empirical data. At present, all common approaches are based on the Kingman *coalescent* [9]. This branching process completely describes the statistical properties of the genealogy of a population—provided all the variation is neutral. As evidence against neutrality has accumulated, alternatives and generalizations of Kingman’s coalescent have emerged [10, 11]. Recent work of statistical physicists has provided a better description of genealogies in continually evolving asexual populations: the Bolthausen–Sznitman coalescent which, in contrast to the binary trees of Kingman, includes multiple lineages merging at the same time in the past with a particular set of probabilities [12]. Related features are, at least qualitatively, observed in genealogical trees of actual populations [13]–[15]. Amusingly, the Bolthausen–Sznitman process first emerged in a completely different context as a relation between different equilibrium states of mean-field spin glasses [16].

Genealogies and coalescents are the focus of the article by Brunet and Derrida. They discuss, in particular, which statistical features of the Bolthausen–Sznitman coalescent are universal properties of asexual adaptive processes with a large number of beneficial mutations. It is exactly these properties that we can hope to identify in real populations.

1.3. Sexual and spatial effects

The articles in this section address a challenging issue: extending the statistical description of evolutionary processes beyond the mean-field and other decoupling assumptions of traditional population genetics. Specifically, they discuss the effects of sexual reproduction with a finite recombination rate and of spatial population structure. Both extensions generate dynamical field theories, which exhibit nontrivial correlation functions depending on genomic distance and on spatial distance, respectively. This field is still at its beginning. But as shown by the contributions to this issue, its development will profit from links to broader topics in statistical mechanics, including systems with quenched disorder and reaction–diffusion dynamics.

Neher, Vucelja, Mezard and Shraiman relate the dynamics of selection on standing variation in a population undergoing a combination of sexual and asexual evolution to the statistical mechanics of Derrida’s random energy model [17]. This establishes a connection between population genetics and spin glass theory: the selection-driven collapse of population diversity in the absence of new mutations is mapped onto the glass transition, and methods developed for dealing with quenched randomness are put to use in calculating the statistics of relatedness in the population.

Weinreich, Sindi, and Watson study an adaptive processes driven by selection and recombination. They show that recombination affects the fixed points of this dynamics and can even jeopardize the very concept of uphill motion in a fitness landscape, because the velocity field of evolution may acquire a curl.

As discussed above, the simplest situations are populations without spatial structure. As long as the environment is constant in space and time, and the interactions between individuals only act to constrain the local population density, the dynamics with spatial motion is still a physics-like problem. As discussed in the article of Etheridge and Barton, the dynamics and statistics of genealogies change with spatial structure even for neutral evolution.

Geyrhofer and Hallatschek consider the effects of selection in spatially uniform environments. They show, in particular, that a dynamics of selection, migration, and genetic drift can lead to fluctuation-dominated delocalization phenomena, which have close analogies to problems in nonequilibrium statistical physics.

Where to?

Recent progress notwithstanding, many key problems in statistical evolutionary genetics remain open. As discussed above and in several articles of this issue, we still do not have a satisfactory understanding of either local or global structure of fitness landscapes. Locally, we would like to characterize the neighborhood of a given point in genome space: What is the spectrum of effect sizes of mutations? What is the fraction of mutations that are beneficial? It is clear that the rate of beneficial mutations is higher in poorly adapted than in well-adapted states [18, 19]. This is one example of a feedback of earlier evolution on later stages. Making such feedbacks quantitative is important for more realistic models of non-neutrally diverse populations that need to replace the neutral model as the default for interpreting genomic data from populations.

The global structure of the landscape becomes particularly important for sexual populations, because recombination cuts and pastes parental genotypes, so lineages do not move through landscapes in small steps. It is important whether effects of mutations are additive or contingent on genetic background, that is, whether mutations at different loci have fitness interactions: *epistasis* in population genetics parlance. Everything we know about organismic biology suggests a great deal of genetic interdependence, which implies extensive epistasis. But epistasis has proved rather difficult to pin down in genetic data (except in the case of outcrossing, which brings together divergent genotypes). Epistasis has large effects on the dynamics of selection. Neher *et al.*, in the present issue, decompose the dependence of fitness on multilocus genotypes into an additive and a (quenched) random part. The latter is motivated by the assumption that interactions are complicated enough to make the non-additive portion of the offspring fitness essentially independent of the parental fitnesses. The resulting model is similar to the random energy model of statistical physics [17]. Yet considerably more work is needed to develop a more realistic model of epistasis and recombination, especially with chromosomal structure which makes the separations between loci of sets of genetic differences important. More generally, better biologically motivated models of epistasis are needed to describe the effect of new mutations in sexually reproducing populations: there is no reason to believe that a primarily additive assumption is a good starting point for studying sustained evolution.

Another major step will be to move beyond evolution in static fitness landscapes. While it is a reasonable starting point for modeling of evolutionary processes under selection, the landscape approach contains several drastic simplifications. It neglects the dependence of fitness on the environment and, therefore, on time and space. Time dependent environments are ubiquitous and essential for evolution. If time dependence arises primarily from external changes, the metaphor of a dynamic *fitness seascape* is instructive [20]. For modest changes, or with little diversity in a population, random models of the seascape can provide useful starting points. But often the most important changes are intrinsic to the evolving population: evolution changes the interactions between organisms, directly or by modifications of their environment. Responses to this can include differentiation within a population: genetically identical organisms that do

not all behave the same way. Hence, the challenge for the future is to develop models that include some caricature of both the organismic biology and the ecology, yet are still simple enough for developing an understanding beyond the particular details of the model. For already well-adapted microbial organisms, almost all beneficial mutations are likely to be conditionally beneficial, depending both on the genome on which they occur and on the environment. Similarly, feedbacks on an organism's fitness resulting from changes in the environment that their or their relatives' evolution induces are likely to result from sums of positive and negative effects. Random models of these interactions may again be a good starting point, but an important challenge is to disentangle the ecological stochasticity from stochastic effects in the response of organisms at the genetic or phenotypic level.

These examples of unsolved problems show that statistical physicists now have many great opportunities to frame new questions for evolutionary biology—and to answer some of them.

References

- [1] Haldane J B S, *A mathematical theory of natural and artificial selection part I*, 1924 *Trans. Camb. Phil. Soc.* **23** 19–41
- [2] Kimura M, 1983 *The Neutral Theory of Molecular Evolution* (Cambridge: Cambridge University Press)
- [3] Fisher R A, 2000 *The Genetical Theory of Natural Selection: A Complete Variorum Edition* ed H Bennet (Oxford: Oxford University Press)
- [4] Wright S, *The roles of mutation, inbreeding crossbreeding and selection in evolution*, 1932 *Proc. of the 6th International Congress on Genetics* pp 356–66
- [5] Lenski Richard E, Rose M R, Simpson S C and Tadler S C, *Long-term experimental evolution in Escherichia coli. I. Adaptation and divergence during 2000 generations*, 1991 *Am. Naturalist* **138** 1315
- [6] Gerrish P J and Lenski R E, *The fate of competing beneficial mutations in an asexual population*, 1998 *Genetica* **102/103** 127
- [7] Weinreich D M, Delaney N, DePristo M A and Hartl D L, *Darwinian evolution can follow only very few mutational paths to fitter proteins*, 2006 *Science* **312** 111
- [8] Cross M C and Hohenberg P C, *Pattern formation outside equilibrium*, 1993 *Rev. Mod. Phys.* **65** 851
- [9] Kingman J, *On the genealogy of large populations*, 1982 *J. Appl. Probab.* **19A** 27
- [10] Krone S M and Neuhauser C, *Ancestral processes with selection*, 1997 *Theor. Popul. Biol.* **51** 210
- [11] Durrett R and Schweinsberg J, *A coalescent model for the effect of advantageous mutations on the genealogy of a population*, 2005 *Stochast. Process Appl.* **115** 1628
- [12] Brunet E, Derrida B and Simon D, *Universal tree structures in directed polymers and models of evolving populations*, 2008 *Phys. Rev. E* **78** 061102
- [13] Strelkova N and Lässig M, *Clonal interference in the evolution of influenza*, 2012 *Genetics* **192** 671
- [14] Neher R and Hallatschek O, *Genealogies of rapidly adapting populations*, 2013 *Proc. Natl Acad. Sci. USA* **110** 437
- [15] Desai M M, Walczak A M and Fisher D S, *Genetic diversity and the structure of genealogies in rapidly adapting populations*, 2012 *Genetics* **112.147157** (Epub ahead of print)
- [16] Bolthausen E and Sznitman A S, *On Ruelles probability cascades and an abstract cavity method*, 1998 *Commun. Math. Phys.* **197** 247
- [17] Derrida B, *Random-energy model: an exactly solvable model of disordered systems*, 1981 *Phys. Rev. B* **24** 2613
- [18] Silander O K, Tenailon O and Chao L, *Understanding the evolutionary fate of finite populations: the dynamics of mutational effects*, 2007 *PLoS Biol.* **5** e94
- [19] Goyal S, Balick D J, Jerison E R, Neher R A, Shraiman B I and Desai M M, *Dynamic mutation–selection balance as an evolutionary attractor*, 2012 *Genetics* **191** 1309
- [20] Mustonen V and Lässig M, *From fitness landscapes to seascaapes: non-equilibrium dynamics of selection and adaptation*, 2009 *Trends Genet.* **25** 111