POSITIONAL INFORMATION STORAGE IN SEQUENCE PATTERNS

Dissertation
submitted to the
Department of Mathematics and Computer Science
of the Freie Universität Berlin
for the degree of
Doctor of Natural Sciences (Dr. rer. nat.)

by

Alexey Andreevich Shadrin

November 2013

First reviewer: Prof. Dr. Christof Schütte

Second reviewer: Prof. Dr. Andrey Grigoriev

Day of the disputation: September 24th, 2014

ACKNOWLEDGEMENTS

I would like to thank Prof. Dr. Hans Lehrach for giving me the great opportunity to conduct my dissertation in his department. I am grateful to Prof. Dr. Hans Lehrach and Dr. Christoph Wierling for allowing me the freedom in my research topics and for providing such promoting research environment at the MPI MG in Berlin.

This work was inspired and substantially influenced by the ideas of Dr. Dmitri Parkhomchuk, who also made a significant contribution to both papers, written during preparation of this thesis. I would like to express my sincere gratitude to Dmitri for his continuous comprehensive support and fruitful discussions.

I thank Prof. Dr. Christof Schütte for his kindness to serve as an academic advisor and reviewer of my thesis. I wish to thank Prof. Dr. Andrey Grigoriev for the assistance in writing the paper and for reviewing this thesis.

I would like to thank David DeMiglio for proofreading of my thesis.

Finally, I want to thank my parents Andrey and Svetlana and my sister Anna for their constant support throughout my thesis. And special thanks to my wife Evgenia for her unending encouragement, support and patience.

CONTENTS

ABSTRACT

Invariants (conservation laws) have served as the ultimate cornerstones of mathematical and physical theories from the early days of science to modern times. For example, the initial name of Einstein's theory was “Invariantentheorie”, and Klein in the “Erlanger Programm” saw geometry as the study of invariants under a group of transformations. However, in molecular evolution theories, the widely observed phenotype invariance, i.e. its preservation through generations, is not matched with any genomic sequence invariants. On the contrary, the genomic sequences are perceived to be quite fluid, evolving rapidly and opportunistically, frequently “neutrally”. The classical models of molecular evolution were elaborated more than 40 years ago with an extreme paucity of data. The consequent development of molecular evolution theory was primarily haphazard and superficial: minor ad hoc assumptions were introduced to fit newly obtained data but the core of these models remained unchanged. The concepts were expanded upon with more details and assumptions, becoming cumbersome and losing the ability of making verifiable predictions or explanations of observable phenomena. This lack of general fundamental principles has led to the crisis of molecular evolution theory. Current technologies supply us with an enormous amount of molecular data, allowing a deeper look into genome functionality, and demand a more profound understanding of genomic functionality.

This work introduces a novel paradigm into molecular evolution theory by proposing an invariant property of the genomic sequence, which does not vary at all or only slowly from generation to generation, while allowing the underlying sequences to change rapidly. The introduction of the invariant leads to more a “physical” and less opportunistic view on sequence evolution and provides testable predictions. The well-developed apparatus of Shannon’s informational theory is used as a mathematical framework of the model. A functional site is regarded as a positional probabilistic pattern, where each position of the pattern is a four-vector of nucleotide probabilities in the equilibrium population (i.e. abstract infinite population that has evolved for an infinite time without any disruptive events). Introducing the invariant allows us to simulate the genetic information dynamics and to apply basic physical principles such as the optimal efficiency and channel capacity. The model demonstrates a fundamental possibility of error-free information storage in sequences possessing arbitrarily low conservation. I show that the rate of beneficial mutations can be high in general—the lower the sequence conservation the higher the frequency of beneficial mutations. Experimental data demonstrates the tendency of real functional sites to optimization according to the proposed optimality criterion. The model allows

a fresh look at the well-known phenomena (e.g. it demonstrates that the "Molecular clock" and "Drake's rule" possibly emerge out of common underlying process). It is also able to provide reasonable explanations for some paradoxes (e.g. "Paradox of Variation") which are lacking a clear interpretation in the framework of classical theories. Therefore I believe that further development of the model will facilitate a deeper understanding of molecular evolution and population genetics processes.

OUTLINE

The thesis consists of three main sections: Introduction, Results and Discussion.

In the Introduction section I will survey basic theoretical concepts of molecular evolution, trace their historical development and discuss their current state. Here I will also briefly describe a number of noteworthy phenomena and paradoxes observed in molecular data. More attention will be paid to the "Quasispecies model", because it touches some aspects which are relative to the model presented in the Results section. Next I will outline some fundamental notions of Shannon's information theory, which will be hereinafter used in the model. Drawing a line under the Introduction, I will consider interrelations between information theory and molecular evolution.

In the first part of Results, the model of positional information storage in sequence patterns is introduced. First I discuss underlying assumptions and specific notions. Then the core principal of the model—the principle of sequence pattern conservation—is presented. Experimental evidence of pattern conservation is demonstrated in the example of human and mouse slice sites. After positing the conservation law, I suggest a criterion according to which the pattern can be optimized. The criterion is formulated in the form of a nonlinear constrained minimization problem which is then solved. The traces of expected optimal compositions of nucleotide frequencies are demonstrated in real binding sites. The second part of Results is entirely devoted to the "Drake's rule" phenomenon. I present its interpretation in the framework of the model. The proposed explanation is discussed in details and compared with conventional explanation.

The last section represents the overall discussion with some philosophical digressions. I speculate about the impact of the model on the status quo in molecular evolution theory, emphasizing the novelties introduced by this work.

1 INTRODUCTION

1.1 Concepts of molecular evolution

The literature of molecular biology contains a great variety of models of molecular evolution, however most of these models can be classified into two major types according to their theoretical foundations. The two most well-known concepts of molecular evolution are selectionism and neutralism. To some extent they represent the two extremes of an explanatory spectrum for understanding patterns of molecular evolution and the emergence of evolutionary innovation. The tension between these two concepts began when Motoo Kimura proposed his neutral theory of molecular evolution at the end of the 1960s, so it is without exaggeration almost as old as the field of molecular evolution itself. Historically, the most ferocious debates between selectionists and neutralists were focused around explanations for genetic variation either between populations of different species, divergence or in a population of single species, polymorphisms. In general, in explaining observed genetic variations, proponents both of neutralism and selectionism agree that deleterious mutations occur frequently in the course of evolution, but there is a deep discord between their positions on the relative importance of effectively neutral and beneficial mutations (Wagner 2008). Both mainstream concepts consider fixation of the mutation in a population as an elementary act of evolution. In this context, considering the observed spectrum of mutations, neutralists argue that beneficial mutations are rare and are fixed less frequently than neutral or slightly deleterious mutations (Kimura and Ohta 1974). Selectionists, by contrast, assert that beneficial mutations are abundant, so the majority of mutations that go to fixation in a population would be beneficial, or are at least linked to beneficial mutations. Considering polymorphic alleles, neutralists suggest their intermediate frequencies are simply a transient state in a continuously ongoing process of random genetic drift among neutral (i.e. functionally equivalent) alleles, while selectionism states that they provide selective advantage and thus are maintained by natural selection.

Here I need to digress briefly to explain the notion of random genetic drift, which is quite important for understanding both classical models described in this section, and the new model which is elaborated in the framework of this thesis. Furthermore, I will often refer to random genetic drift as genetic drift, or more simply as drift. However, randomness is a crucial feature of this process. According to the definition from the textbook, genetic drift is “random changes in allele frequency from one generation to the next in biological populations due to the finite samples of individuals, gametes, and ultimately alleles that contribute to the next generation" (Hamilton 2009). In contrast to natural selection, drift affects only alleles that have no effect on the fitness of organism (or whose effect is negligibly small). So the frequencies of such alleles in the population, due to their finite size, are determined by chance due to random sampling.

For a long time molecular evolution suffered from the a lack of real biological data; it was a field characterized by a wealth of theory and a paucity of data. However, over the last half-century rapid progress in nucleotide sequencing and other advanced investigation methods of molecular biology provided a vast amount of data and significantly changed our understanding of intercellular processes. These data made it possible at last to test theoretical predictions, to refine and supplement theoretical concepts. Under the influence of new experimental data, both selectionism and neutralism experienced a number of modifications, which made them closer in some aspects, but did not affect the core postulates of the theories. New knowledge also facilitated an emergence of a number of offshoot models. Primarily, they were generated simply by adding some minor assumptions to one of two major concepts and thus can scarcely be considered truly separate. However, there were a few (e.g. the near neutral theory (Ohta 1973, 1976)) that introduced new rather significant assumptions and thus can be treated as fully independent theories. There are also concepts, e.g. the so called quasispecies model (Eigen and Schuster 1977), which incorporate features of both selectionism and neutralism, thus attempting to reconcile them. In my opinion the latter models are the most interesting and therefore deserve special attention.

Although the last two decades demonstrate, somewhat, a decreasing tension about genetic variation between neutralists and selectionists and a convergence of their positions in some questions, the underlying discord persists. In fact, implications of this tension go far beyond explanations of genetic variation and envelope one of the most fundamental, however still unresolved, problems of evolutionary biology: the problem of the origin of evolutionary innovations (Wagner 2008).

1.1.1 Selectionism

Historically, selectionism was the first paradigm of molecular evolution. It appeared when molecular evolution came into its own in the late 1950's and early 1960's. At that time the majority of the molecular biologists saw the world through the lens of panselectionism (Neo-Darwinism) (Dietrich 1994), the broad evolutionary concept that grew in 1920's and 1930s from the synthesis of Mendalian genetics with the ideas of Darwin and Wallace, with Darwin’s notion of natural selection playing the dominant role. Panselectionism states that the strongest (perhaps even sole) driving force of evolution is natural selection, which acts as a purifying force removing deleterious alleles and promoting those that are beneficial. Natural selection thereby gradually improves the fitness of a population, and evolution represents a slow “directed” process consisting of a continuous sequence of minor variations.

Perhaps due to its earlier appearance and connection with Darwin’s ideas, the concept of selectionism is also sometimes called the classical theory. It is also worth noting that selectionism does not deny random drift and the formative influence of mutations, however, their role is considered to be minor. The ideas of selectionism were advocated by many outstanding scientists including R. A. Fisher, J. B. S. Haldane, G. G. Simpson , T. G. Dobzhansky, E. W. Mayr and others. These authors were aware that natural populations are far from genetic homogeneity and may potentially comprise a large variety of mutants. However, in general they assumed that there would be little genetic variation in populations. For this reason their work was primarily focused on the “wild-type” (i.e. the most prevalent allele) as a target of selection. Classical models are usually deterministic and consider infinite populations. These models usually contain mutation essentially as a 'perturbation term' in the differential equations describing selection. This term changes specific features of the solution, but the concept of the individual “wild-type” remains unchallenged (Eigen 1996). An assumption of infinite population size seems to be very strong and had been already questioned by Sewall Wright in early 1930s, before the molecular nature of DNA became known. He was the first who considered models with finite populations and proposed that fluctuations in allele frequencies due to stochastic effects (i.e. random genetic drift) can play an important role (Wright 1931). However, at that time there was very little knowledge regarding sizes of real populations or frequency of bottlenecks, and absolutely nothing was known about the physical nature of genes and DNA in general (Hughes 2007). Indeed, before the 1960s conventional models assumed that natural populations are large enough to ignore stochastic fluctuations of alleles. Due to this misconception, Wright’s models were found to be unrealistic and dismissed by the broad scientific community (Fisher and Ford 1950). The lack of knowledge gave rise to unrealistic views that were laid in the foundation of Neo-Darwinism and, despite the fact that most of postulates were later revised according to up-to-date experimental data, transferred (often implicitly) to selectionism concept when it was formed.

To summarize, we can say that from the perspectives of selectionism, evolutionary innovations emerge through both beneficial mutations, each changing the properties of a given phenotype very slightly, and natural selection which amplifies the frequency of beneficial alleles in the population.

1.1.2 Neutralism

Selectionism was a mainstream theory for less than a decade. During this period fragmentary information on molecular structure and mechanisms became available, changing insights into molecular evolution. In 1968 the Neutral theory of molecular evolution was proposed by Motoo Kimura (Kimura 1968), one year later essentially the same idea was proposed independently by King and Jukes (King and Jukes 1969). The idea quickly gained popularity among the scientific community partially due to strict and clear formulation carried out by Kimura, partially because it was able to explain numerous phenomena observed in experiments, which before had no clear interpretation in the framework of classical theory. Among these are such famous phenomena as Haldane’s dilemma (Haldane 1957)—an inconsistency of predicted (according to classical model) to empirically estimated rates of mutation accumulation, and unexpectedly high genetic variability of real populations. In contrast with typical models of the time, which were often based on the mathematically convenient, but obviously unrealistic, assumption of an infinite population size, Kimura examined the consequence of finite population on natural selection and genetic drift. The hypothesis proposed by Kimura was quite radical because it fundamentally contradicted the prevailing concept of molecular evolution. He assumed that evolutional innovations might be facilitated by mutations which, when they first arise, do not affect molecular functions (and hence are not targets for natural selection). In other words, he argued that genetic drift, rather than beneficial mutations, is the dominant process in evolution both within populations and over evolutionary time. This of course does not mean that all mutations are neutral, only that the vast majority of observed mutations are neutral. Initially his view was based mainly on the observation that the rate of amino acid substitutions among different groups of animals is roughly the same, and partially on the fact that genetic code is degenerate, which was already known at that time (Kimura 1968). It is also worth noting that Kimura did not deny the important role of natural selection. What he did is make a clear distinction between two types of selection: (1) purifying (or negative selection) which aims to remove deleterious alleles, and (2) positive (Darwinian selection) which favors advantageous mutants, causing rapid directional shift in alleles’ frequencies (Hughes 2008). Relying on existing data and knowledge he reasoned that because advantageous mutations are rare events, positive selection is a rare phenomenon. Deleterious mutations, in contrast, are common and, therefore, purifying selection is ubiquitous.

Declaring genetic drift to be the dominant process of molecular evolution, Kimura avoided making intricate assumptions about the potentially quite complex process of selection. This allowed him to reduce molecular evolution to a succinct, clear stochastic model with only two major parameters: population size and mutation rate. The neutral theory then provides a baseline hypothesis from which numerous testable predictions can easily be derived. Despite its simplicity, Kimura's was able to demonstrate explanatory and forecasting power from the very beginning.

Next I would like to discuss the history of two famous and popular concepts that are directly connected with the neutral theory. Preceding the invention of the neutral theory Zuckerkandl and Pauling (1962) noticed that the number of amino acid substitutions in α and β hemoglobin chains change roughly linearly in different lineages with time. The same observation was later made for cytochrome c (Margoliash 1963). These experiments led to the suggestion of a currently well-known technique called the molecular or evolution clock, which is now broadly used in phylogenetics for estimation of the speciation time, and building phylogenetic trees. The idea behind the concept of the molecular clock is that the evolution rate of DNA and protein sequences (i.e. the speed of accumulation of changes) is relatively constant over time and among different organisms. From this, it immediately follows that the genetic difference between any two species depends linearly on the time since their last shared common ancestor. Hence, if we accept this hypothesis, it can be used for estimation of evolutionary timescales. Initially the molecular clock was proposed based on purely empirical data estimated from fossil evidence. This phenomenon had no clear explanation in the framework of the classical concept of the molecular evolution, and to a great extent contradicted it. On the other hand, Kimura's theory states that the overwhelming majority of amino-acid substitutions are neutral, so if this holds true the constant rate of mutations' accumulation becomes an evident consequence. Thus, the neutral theory provides theoretical justification for the molecular clock. Another thing worth mentioning concerns the discrepancy of evolution between the functional and non-functional regions of a DNA sequence. Molecular biologists normally consider as self-evident the statement that functionally important sequences are conserved (i.e. evolve slowly). Many basic bioinformatics tools, such as homology searchers and sequence aligners are based on the fact that functionally essential sequences accumulate variations slower than sequences which have secondary importance or bear no function at all. However, this assumption is a prediction of the neutral theory that directly contradicts the prediction made by selectionists (Hughes 2007).

Soon after its presentation the neutral theory was brought to the center of evolutionary genetics, it showed forecasting power and provided a mechanism for the investigation of deep issues concerning molecular genetics. Very quickly, it gained strong influence in the scientific community and moved the classical theory from its dominant position. However, subsequent experimental evidence exposed phenomena contradicting to predictions of the neutral theory, which, however, could be better explained by the classical paradigm. A hot dispute between selectionists and neutralists continued. This has led to a revision of some key positions of neutralism and to the creation of the near neutral theory.

1.1.3 Near neutral theory

As was already mentioned, the core tenets of the neutral theory were established in a tight shortage of molecular data. When new experimental data became available, inconsistences between predictions of the neutral theory and observations emerged. E.g. the neutral theory (in its original form) predicts the so called generation-time effect, i.e. creatures with longer generation time will evolve slower in real time than those with shorter generation time. However, it was shown that proteins do not exhibit a strong generation-time effect, while noncoding DNA does (Kohne 1970). The reliance of the neutral theory on random genetic drift also fails to explain the "paradox of variation" (Lewontin 1974), where genetic diversity has not been found to depend strongly on the size of different populations. Although these observations do not contradict directly the possibility that many (or even most) substitutions are neutral it is thought that they can be better explained if allele frequency dynamics is driven by selection at the linked sites rather than by random genetic drift (Hahn 2008). The problems arising from the neutral theory promoted the development of new approaches in understanding the principles of molecular evolution. Taking key concepts of the neutral theory as a basis, Kimura’s associate Tomoko Ohta emphasized the role of a certain class of mutations with small selection coefficients (especially slightly deleterious). This class can be described as nearly neutral mutations. Her theoretical generalization of the neutral theory by including the concept of "near-neutrality" gave rise to the near neutral theory (Ohta 1973, 1976).

Here I digress slightly to clarify the notion of “effective population size”, because it can be found throughout theoretical papers on molecular genetics. This term is used in the original neutral theory; however, it is especially important for understanding the near neutral theory and will be often used further in this work. The concept of the effective size of a population was initially introduced by Sewall Wright (Wright 1931). Speaking simply, the effective population size indicates the number of individuals actually contributing alleles to the next generation. Due to various reasons (including sexual selection, sex ratio, inability of reproduction etc.) this number is usually smaller than the total number of individuals in the population (however, these two numbers usually correlate).

Returning to our topic, it should be noted that the fate of near neutral alleles, in contrast with strictly neutral, depends on effective population size. So if the effective population size becomes small enough, slightly deleterious mutations will behave as real neutral, i.e. they can drift to a higher frequency or even to fixation. But when the effective population size is large during a long enough time, these mutations will be eliminated from the population by selection. Slightly advantageous mutants are supposed to behave similarly to slightly deleterious, but they are expected to be less common than slightly deleterious.

Revisiting the generation-time effect we can now demonstrate how it can be potentially explained in the framework of the near neutral theory. Large organisms tend to have small population sizes and rather long generation times as compared to small organisms. Hence, according to the near neutral theory, the former will accumulate more slightly deleterious mutations. So the generation-time effect can be (at least partially) compensated by the difference of population sizes. So, according to the near neutral theory, population size plays a very important role in the formation of observed allele patterns. It is expected that when the population undergoes a bottleneck the power of natural selection is reduced, many slightly deleterious alleles become neutral, and their frequency in the population increases. However, if after the bottleneck effective population size increases, the efficiency of natural selection rises, these mutations again become slightly deleterious and are purged from the population.

The emergence of a bottleneck effect can be potentially demonstrated by the structure of single-nucleotide polymorphisms (SNPs) in the human population, which presumably underwent the last bottleneck nearly 70,000 years ago when, after the eruption of the Toba volcano in Indonesia, the total number of individuals was reduced to approximately 10,000 (Dawkins 2004). Zhao et al. (2003) has shown that the ratio of non-synonymous to synonymous mutations in human coding regions is less than half of that expected under the neutral mutation theory. This phenomenon is often interpreted as evidence of non-neutral evolution. However, there is at least one potential explanation in the framework of the near neutral theory. According to this explanation, the strength of the natural selection was reduced after the bottleneck and an abundance of slightly deleterious mutations spread over the population. The ratio of non-synonymous to synonymous mutations at that time was extremely high (as would be expected according to the strictly neutral theory). However, as the size of human population (as well as the effective population size) again started to grow, the pressure of purifying selection increased and began wiping out non-synonymous alleles (which are likely to be slightly deleterious), decreasing the ratio of non-synonymous to synonymous mutations. This process is ongoing, so currently we observe an intermediate state and this ratio will likely continue to decline as the size of human population is growing constantly. Bottlenecks are rather regular phenomena in different natural populations, so a correct understanding of how they affect the process of molecular evolution is very important.

Concluding this topic, I want to emphasize that perhaps it is not fully correct to consider the nearly neutral theory as an alternative theory competing with the neutral theory. Although the core assumption of the Kimura’s theory (namely that the majority of fixed mutations and polymorphisms within species are neutral) is modified in the near neutral theory, it may be more precise to say that the latter represents a corollary of the neutral theory that particularly focuses on the issues of slightly deleterious alleles.

1.1.4 Conclusion about neutralist-selectionist debates

The controversy between selectionists and neutralists was a main topic of population genetics and molecular evolution discourse from the late 1960s till the early 1990s. Starting with the severe lack of experimental data the debates were freshly innervated in the middle of 1970s, when DNA sequencing technology became available, which together with other technical innovations provided an exponential growth of molecular data. There was a hope that this large bulk of data would allow for the revelation of the real nature of molecular evolution mechanisms and facilitate a rapid resolution of the controversy. However, this expectation was ill-conceived. The abundance of DNA sequencing data shifted the main subject of the debates from proteins to DNA evolution, but it did not turn the balance towards neutralism or towards selectionism. In the late 1980s the debates quietly withered and came to an indeterminate demise (Hey 1999). Since the end of 1980s the situation can be characterized by a massive gathering of data and substantial lag of theoretical understanding (directly opposite to what we have seen in the beginning of debates). The state of affairs seemed so sad that proponents of both concepts proclaimed a looming crisis in the theory of molecular evolution, stating that all current theoretical models suffer either from unrealistic assumptions or are unable to describe known phenomena (Ohta and Gillespie 1996). So what are the current states of the debate and the problems of molecular evolution theory in general? I would say that this is largely a philosophical question. Exploring this question it is easy to dive deep into numerous arguments provided by supporters of both schools and get lost in them. As such, at the end of this section I will try to give a brief overview of current status quo.

Historically assuming that natural selection is the driving creative power in the processes of evolution, selectionists chose a more complicated position. Natural selection can take countless forms and it is not possible to elaborate a general testable mathematical model describing it. On the other hand the neutral theory (here I mean the original, i.e. strict neutral theory) provides a succinct and elegant model that can easily be tested. Until very recently most of noncoding DNA was considered non-functional. Non-functional sequence, by definition, avoids the pressure of natural selection, experiencing random genetic drift. That is why the discovery of the fact that a bulk of DNA in large genomes represents a noncoding sequence provided considerable vitality to the neutral theory.

As shown above, the neutral theory is able to explain much of observed phenomena. However, there are also a lot of cogent examples where its predictions are significantly violated. McDonald and Kreitman (1991) studied the mutational spectrum within coding regions of different Drosophila subgroups. They demonstrated that the ratio of fixed non-synonymous mutant alleles (i.e. alleles which differ between species but remain constant within them) to polymorphic non-synonymous alleles (i.e. alleles which show variation within species) often greatly exceeds the ratio between synonymous fixed and polymorphic mutations. This fact contradicts the neutral theory but can be easily interpreted as an emergence of positive selection. The inconsistency of neutral theory predictions to observed bias of alternative codon usage was described by Akashi (1995). There are also many lines of evidence against strictly neutral evolution of protein features (Kreitman and Akashi 1995), e.g. so called overdispersion of the molecular clock, i.e. an extremely high variance in the rate of protein evolution, which is much higher than predicted by the neutral theory (Takahata 1987). These are only some of the most notable examples where molecular data can be better explained assuming that natural selection rather than genetic drift plays the dominant role in molecular evolution. However, more can be easily found in the literature, for example see (Hahn 2008). Many of above mentioned phenomena can also be explained in the framework of Ohta’s near neutral theory, but the explanations and the theory behind this are “necessarily complex and cannot generate simple predictions in the way the strictly neutral model does” (Hey 1999). Despite uncertainties about the neutral theory it should be emphasized that all investigations regarding detection of the natural selection would have been impossible without the neutral model, which is always used as a viable and testable null alternative to selection (Kreitman 1996).

As can be seen, both theories are able to explain some spectrum of observable phenomena. Sometimes both models can provide a satisfactory explanation, in other cases one can be more convincing than the other, but often both concepts fail to give any reasonable interpretation. So, despite longstanding controversy, we still lack any general realistic model of molecular evolution. However, there is no doubt that desperate neutralist-selectionist debates were quite useful in vividly demonstrating that understanding mechanisms of molecular evolution is a fundamental problem of biology and that solution of this problem is quite complicated (Nei 2005). However, the truth comes only in discussion. The development of theoretical conceptions of molecular evolution resembles the classical reasoning triad: “Thesis, antithesis, synthesis”, most fully developed in Hegel’s philosophy, where two opposite extremes are logically followed by their fusion. By the early 1990s Li and Graur (1991) fairly noted in their book that any contemporary adequate theory of molecular evolution must be consistent with both natural selection and genetic drift.

From the very beginning of the controversy it is possible to trace convergence tendencies in both theories. These trends are sustained to this day. Recently, an interesting model combining both concepts was suggested by Wagner (2008). However, a good example of reconciliation of neutral and classical theories can be found in the much earlier work of Eigen (1971). In this seminal paper he speculated about the origin of life and proposed a strict mathematical model of evolution of self-replicating entities (potentially describing the evolution of early life). Later the model was named the “quasispecies model”. It intrinsically combines features of both concepts. However, for a long time this model did not receive proper attention from the community of evolutionary geneticists. Perhaps this can be partially explained by the fact that Manfred Eigen focused primarily on aspects and modes of molecular evolution which were not interesting for population and molecular geneticists at that time. I find this model quite appealing. For another thing, many of its aspects are similar to those of the model presented in the results section below. For this reason, I have decided to devote the whole next section to its detailed description, concentrating in particular on features that intersect with the features of the model developed in this work.

1.2 Quasispecies model

The quasispecies model will be described here in detail because some of its aspects are quite related to the model presented in the results section. However, this section (despite its name) is not only about the quasispecies model itself. Through the example of this model I will introduce some terms and describe some concepts important for the understanding of the results of this thesis.

The concept of quasispecies was introduced by Manfred Eigen in the early 1970s (Eigen 1971), but the term “quasispecies” was initially proposed later in his joint work with Peter Schuster devoted to the origin and evolution of life (Eigen and Schuster 1977). In that work they considered a process of Darwinian evolution in an infinite population consisting of asexual, self-reproducing, mutable entities (in particular DNA or RNA sequences) within the framework of physical chemistry. In the original work of Eigen and Schuster (Eigen and Schuster 1977) the term “quasispecies” is defined as an equilibrium distribution of mutants, which arises and is maintained under mutation-selection process. In a simple terms, quasispecies can be described as a large group, a cluster or a cloud of closely related molecular ‘organisms’ (nucleic acid sequences) which evolve asexually in the presence of a high mutation rate, so that each descendant is expected to contain several mutations relative to its parent. To avoid rather common misunderstanding it is also worth noting again that a quasispecies is not an arbitrary ‘swarm’ of mutants, but is a well-defined concept that requires specific conditions (presented in the next section) to be fulfilled. In addition to already mentioned prerequisites (namely: an infinite population of asexual replicators at high mutation rate), the model also needs time for ‘equilibration’ to occur, i.e. populations should evolve for a sufficient time without any disturbing events. In the framework of certain above mentioned assumptions the theory of quasispecies provides a strict mathematical model of molecular evolution described below.

1.2.1 Mathematical description

Several different variations of mathematical models of quasispecies exist, which contain different levels of details including different parameters such as death rate and concentration of energy-rich building material etc. (Eigen and Schuster 1977). Here I will not go into deep specific details and present only the essence, the most common model which can be found in (Nowak 1992).

Imagine that we have an infinite population of nucleotide sequences. Let’s assume that there are N different sequences s_1, s_2, …, s_N (usually sequences of the same length are considered, however, in general it is not necessary). Sequences are able to self-replicate with corresponding rates of replication denoted by r_i, $i \in [1,2, \ldots, N]$. These values can be considered as selective or fitness values of the certain alleles. The process of replication can be thought of as error-prone copying. So in general an offspring is not an exact copy of its parent, but has several mutations, thus parent sequence s_j spawns an offspring s_i, and in general i ≠ j. Let the mutation rate p_{ij} corresponds to the probability that parent s_j will produce a descendant s_i. Commonly, only point mutations where all substitutions are equiprobable are considered, so if we assume that the probability of each point substitution is equal to p (and all sequences have the same length) then $p_{ij} = p_{ji} = p^d$, where d is a hamming distance between s_i and s_j. The closer the sequences are, the higher the transition probability between them is. The quantities p_{ij}, i, j ∈ [1,2, …, N] form the so-called mutation matrix P, and $\sum_i p_{ij} = 1$. With this notation we can construct a system of ordinary differential equations, which describes the time evolution of the population of these sequences.

$$\frac{dx_i(t)}{dt} = \sum_{j=1}^{N} r_j p_{ij x_j} \quad i \in [1, \ldots, N] \tag{1}$$

Where $x_i(t)$ is a concentration (a fraction) of sequence s_i in the population (for the sake of brevity, I will often omit the time argument t). Let us denote the vector of concentrations as X = (x_1, x_2, …, x_N) and assume, that $\sum_i x_i = 1$ (we always can normalize the vector to achieve this), so at each moment of time X forms a distribution of alleles in the population. From this ODE system we can see that within the quasispecies model a frequency of a given sequence in the population does not depend on its replicative value alone, but also on the probability with which it is generated by erroneous replication of other sequences, their frequencies in the population and their replication rates. When time goes to infinity the distribution of sequences in the population eventually stabilizes. The resulting equilibrium distribution is called quasispecies. In mathematical terms this equilibrium is reached when the system is in steady state. Thus, it is possible to find an equilibrium distribution by solving the standard eigenvalue problem of linear algebra: WX = λX, where matrix W contains replication rates and probabilities of mutation:

$$W = \begin{pmatrix} r_1 p_{11} & r_2 p_{12} & \cdots & r_N p_{1N} \\ r_1 p_{21} & r_2 p_{22} & & r_N p_{2N} \\ \vdots & & \ddots & \vdots \\ r_1 p_{N1} & r_2 p_{N2} & \cdots & r_N p_{NN} \end{pmatrix} \tag{2}$$

So, in mathematical terms, quasispecies can be defined as the dominant eigenvector X_{max}, which belongs to the largest eigenvalue λ_{max} of the matrix W. The components of this eigenvector X_{max} completely determine the structure of population in the equilibrium state: the sequence s_i corresponds to the frequency $x_{max,i}$.

1.2.2 Survival of the flattest

One important outcome of the quasispecies model, which discriminates it from the standard population genetics, is a prediction (under certain conditions) of the equilibrium state of the population. This equilibrium population is a stable mixture of closely related organisms (defined by the dominant eigenvector of matrix W), which in general is not required to contain the fittest organism. Due to high rate of mutations and mutation coupling among variants, organisms in the model are not independent entities, and instead, the entire distribution of alleles acts coordinately and can be considered as a single organism. In this situation mutants with lower individual fitness can outperform organisms with higher fitness because they have a better support from their mutational neighbors (Wilke et al. 2001).

Schuster and Swetina (1988) were the first to describe this phenomenon in their theoretical work. They studied the dynamics of the population on a fitness landscape with two equally or almost equally fitted peaks. A fitness landscape is a metaphorical concept introduced by Wright (1932) used to illustrate the relationship between the fitness of the organism, which determines the height of the landscape, and its genotype. The distance between organisms on the surface of the fitness landscape is assumed to be proportional to the similarity of their genotypes, e.g. the hamming distance between their functional genomic sequences. They found that in the case when both peaks have the same height, the population always moves to the flatter peak (i.e. the peak with stronger mutational support). In the case when the flatter peak was slightly lower, the behavior of the population depended on the mutation rate: for lower rate of mutations the higher peak would be occupied by the population, but for higher mutation rates, a lower peak with higher mutational support would be favorable. Later on, it was demonstrated in the experiments with digital organisms that even in the case when organisms residing on the high narrow fitness peak replicate 10 times faster than organisms on the low flat peak, the latter can be preferable at a mutation rate of about 1.25 per genome per replication (Wilke et al. 2001), because mean fitness of this group will be higher. Mutation rates of this order can be observed in nature, e.g. they are common in viral populations (Drake and Holland 1999).

Thus we can summarize that under the high mutation pressure, the lower but flatter peak in the fitness landscape will be preferred over the higher (fitter) narrow peak, making the population more robust against mutations and allowing higher average fitness. In comparison to the Darwinian “survival of the fittest” this effect is usually called “the survival of the flattest” (Wilke et al. 2001). Figure 1 schematically shows how it works.

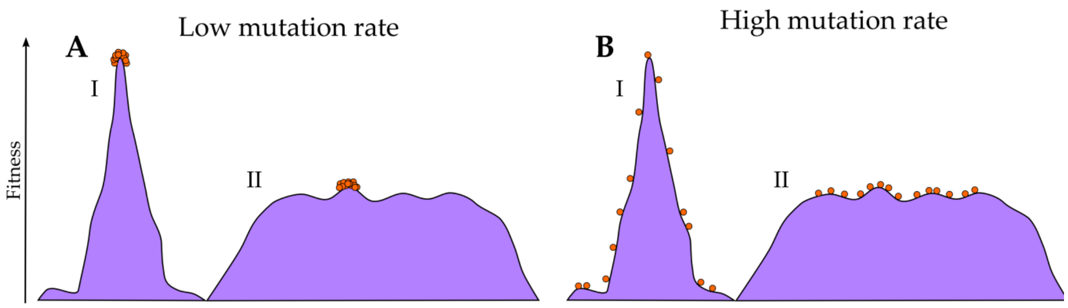

Figure 1. Schematic demonstration of "survival of the flattest" effect (Wilke and Adami 2003).
Left (A) and right (B) parts of the figure represent the same fitness landscape with different mutation rate. Under low mutation pressure the population occupying steep peak (A-I) outperform the population on the flat peak (A-II). On the other hand, when the rate of mutations is high, most organisms located on the steep peak (B-I) cannot hold on at the top and tumble down to the low fitness values. However, organisms of the population on the flat peak (B-II) are able to keep fitness values close to the local optimum. In the latter case the average fitness will be higher for the population on the flat peak.

1.2.3 Error threshold and error catastrophe

Another important inference of the quasispecies theory is the existence of an error threshold of replication leading to an error catastrophe. Eigen described it as a limit to the amount of information that can be stored in the population with some fixed rate of mutations (Eigen 1971). The amount of information in the population could be viewed in the framework of the quasispecies model as a degree of localization of the population in a space of all possible sequences. So it is inversely proportional to the variance of the distribution of alleles' number, contained in the population. Hence the maximum value of information is achieved only when all organisms in the equilibrium population are the same (only one allele is contained in the population, i.e. the eigenvector corresponding to the maximum eigenvalue has a single non zero component), thus the entire population is concentrated in a single point of sequence space. This, in turn, is possible only if the replication is error free (probability of mutation is equal to 0), so that when time goes to infinity, the fittest (the fastest replicating) organism will solely dominate in the population. However, in this case the evolution is impossible and the population will be static. On the other hand, if the probability of mutation is equal to 1, alleles in the population will be uniformly distributed in the sequence space and the population will contain no information. It is intuitively clear that for a probability of mutation between 0 and 1 some intermediate distributions will arise. The quasispecies model allows us to calculate this distribution for a given fitness landscape (i.e. for the fixed replication rates and probabilities of mutation).

Let’s consider a single-peak fitness landscape, so there is one allele that is the most fitted (it is usually called a master sequence). It can be shown (Nowak and Schuster 1989) that for low mutation rate values, the population consists of organisms concentrated in the neighborhood of the master sequence, and the master sequence itself dominates in the population. But when mutation exceeds the certain critical value, the distribution of sequences in the population spreads out rapidly to populate all possible sequences uniformly. Figure 2 shows the common shape of fraction of the master sequence in the population of binary sequences of length 50 as a function of the total mutation rate. There are 2^{50} different binary sequences of length 50, so as the mutation rate increases the fraction of master sequence (as well as the fraction any other sequence) approaches to 2^{-50} and the distribution becomes uniform.

This behavior resembles a phase transition which is quite common in different physical phenomena. Drawing a physical analogy, a mutation rate could be thought as a sort of temperature that "melts" the fidelity of the molecular sequences above the critical “temperature”.

The concepts of error threshold and error catastrophe may seem not intuitive, and probably they are the most misunderstood in the quasispecies theory. Thus I would

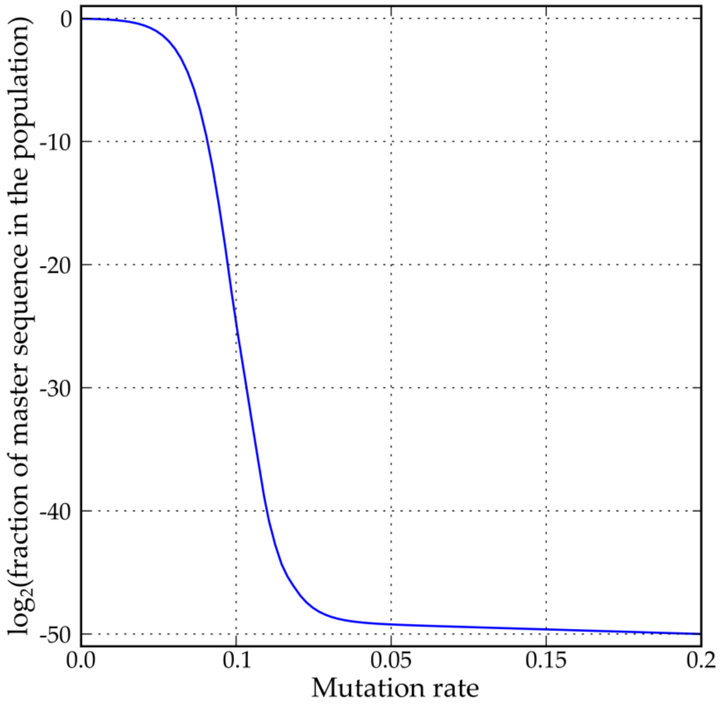

Figure 2. Diagram of the common shape of the logarithm of the fraction of the master sequence in the population as a function of the mutation rate.
A quick transition from the state where master sequence dominates in the population to the state where its fraction diminishes to virtually zero is observed near the critical value of per-base mutation rate equals to 0.1.

like to sum up once again the ideas of this paragraph. There is a value of mutation rate, beyond which the structure of a population breaks down and organisms rapidly disperse over the sequence space. This critical value of mutation rate is termed the error threshold. It depends on the length of genome, as well as on the shape of fitness landscape. For mutation rates below this value, the master sequence dominates in the population but when mutation rate is above this critical value, the fraction of master sequence rapidly drops down to practically zero and the population distributes uniformly in the sequence space. This phase-transition-like phenomenon, when information about the master sequence in the population is rapidly dissipated, is called the error catastrophe (Eigen and Schuster 1977).

A somewhat related phenomenon in finite populations is predicted by "Muller's ratchet" (Haigh 1978). "Muller's ratchet" describes an irreversible process of accumulation of deleterious mutations in the genome of organisms reproducing without recombination. In contrast with infinite population, stochastic effects of random genetic drift play a dominant role. The essence of the model relies on the fact that if back mutations are rare events and most mutations are detrimental, then after some finite time any finite asexual population will reach the state when each organism will carry at least one deleterious mutation so that the wild-type organism will be lost. Further accumulation of deleterious mutations will lead to a reduction in population fitness and eventually to random drift in the sequence space.

1.2.4 Viral Quasispecies

Because of its strong basic assumptions about the properties of the population under investigation, namely: asexuality, infinite population size, and high-mutability (high-mutability per se is not necessary, however, it is required to observe interesting effects predicted by the model), the quasispecies model is usually considered to be relevant to a limited number of real organisms. RNA viruses are among them (Domingo et al. 2002). They exhibit high mutation rates: around one mutation per genome per round of replication or even higher (Drake and Holland 1999) and viral populations, while not infinite, are extremely large.

The first attempt to demonstrate that populations of RNA viruses behave in accordance with the quasispecies model was made in the late 1970s by Domingo et al. (1978). In this classical in vitro experiment with Qβ phage it was shown that though the genome of the phage is statically defined (i.e. averaged "wild-type" remains the same from generation to generation) the structure of the equilibrium population is quite heterogeneous. Most of individuals differ in a few positions from the average sequence and the actual "wild-type" constitutes less than 15 %. Later, a number of studies for different viruses were presented which supported the importance of the quasispecies theory for viral populations: HIV (Hoffmann 1994), hepatitis C virus (Forns at al. 1999), vesicular stomatitis virus (Steinhauer et al. 1989), and foot-and-mouth disease virus (Domingo 1992) for example. An elegant mathematical basis and vivid experimental evidence have paved the way for the quasispecies concept to become the dominant paradigm for RNA virus evolution (Domingo et al. 1985).

However, several authors have reasonable doubt whether the quasispecies theory has any relevance for virus evolution (Jenkins et al. 2001; Holmes and Moya 2002). Others, while recognizing the importance of the quasispecies concept for RNA viruses evolution, argue that it contradicts conventional population genetics (Comas et al. 2005). The proponents of quasispecies, on the other hand, are persistent in defending their position, pointing out that the evidences against the theory are usually based on misinterpretation and incorrect application of the model (Eigen 1996; Wilke 2005). The model itself is quite appealing and arguments for it are strong enough to allow further development of the experimental aspects of the theory, with particular regard to RNA viruses.

1.2.5 Model limitations and discussion

For all its virtues the quasispecies model in practice faces difficulties which impede experimental validation of its predictions. Its rigid assumptions are quite convenient for building a mathematical model, but they are often incorrect for real biological populations. It is clear that all mathematical models assume some simplifications of real physical phenomenon they describe. After a simplifying assumption is made the question arises whether the model is still useful or if some crucial features of described phenomenon are violated and predictions of the model meaningless. Below, the most important shortcomings of quasispecies theory will be described and discussed.

The conventional model of quasispecies, as described above, deals with an infinite population. However, all real biological populations obviously have finite size. In particular, we have a situation when even for modest genome lengths, the number of possible genome sequences is much larger than the number of individuals in the population. For example, the largest observed populations of RNA viruses are on the order of 10^{12} viral particles within a single infected organism (Moya et al. 2000) while their genome contains 10^{3}-10^{4} nucleotides, and hence the total size of the sequence space is approximately $10^{6000} >> 10^{12}$. We can expect that even the region of sequence space with high fitness is typically much larger than the size of a real population in nature, so a population of realistic size will never cover it. In this situation we expect evolution to be governed mainly by random genetic drift which then renders the deterministic equations of the quasispecies model inapplicable (Jenkins et al. 2001). In order to get a better understanding of the limits of model's applicability, it should be tested whether phenomena observed in the conventional model persist in the case of finite population or if the stochastic effects of genetic drift become crucial and violate consistency of the model.

A lot of works were devoted to adopt the quasispecies theory to finite population sizes (Nowak and Schuster 1989; Campos and Fontanari 1999; Park et al. 2010). The authors of these studies usually take deterministic equations from the conventional model and then add a stochastic component to describe the influence of genetic drift in the finite population. These modifications can yield cumbersome results, and appealing analytical inferences of the original model transform into bulky expressions that are often difficult to interpret. It would be tempting to have a succinct model which at the same time would be able to estimate the main features of the finite population behavior.

It was shown, that at least in some cases, the information about the whole sequence space (which is available for infinite population) is not required and the behavior of the finite population residing in the vicinity of some local optimum in the sequence space can be described from deterministic equations (van Nimwegen et al. 1999b). I expect that it is possible to observe phenomena predicted by the quasispecies model in finite populations. First I would like to mention that the effect of the survival of the flattest was observed in extremely small digital populations, having a size lower by 50 orders of magnitude than the complete sequence space (Comas et al. 2005). However, the most noteworthy phenomenon is the emergence of quasispecies themselves, i.e. the formation of a population’s structure which facilitates the minimization of the mutational load via accumulating closely related sequences and thus reducing the risk to suffer from deleterious mutations (Bornberg-Bauer and Chan 1999; Wilke 2001). This phenomenon has been called the evolution of mutational robustness (van Nimwegen et al. 1999a). Van Nimwegen et al. (1999a) demonstrated the somewhat surprising fact that the population has the tendency to evolve toward highly connected regions of the sequence space as long as $\mu M >> 1$, where μ is mutation rate per genome per replication and M is the size of the population. If $\mu M >> 1$ this tendency is independent of evolutionary parameters such as mutation rate, selection advantage and population size. In many RNA viruses the number of mutations per genome per replication exceeds 1 (Drake and Holland 1999), thus, even relatively small populations of RNA viruses can be good candidates for experimental validation of the theory.

The core of the quasispecies model is quite general. However, most papers concerning the model focus on the narrow case of a single-peak fitness landscape in which a single fittest sequence (master sequence) has superior fitness 1+s while all other sequences have fitness 1 (Tannenbaum et al. 2004; Campos and Fontanari 1999; Swetina and Schuster 1982). As a result they lose much of generality of the original model and, what is even more unfortunate, move away from natural biological conditions. Another unrealistic feature contained in most existing models based on the quasispecies concept (which is, however, not inherent to the general quasispecies theory) is that they consider a type of selection where all sequences in the sequence space are assumed to have positive replication rates, that is, there are no lethal mutations and each genotype is able to produce a progeny. Such selection is termed "soft selection". Under soft selection only relative differences in fitness are important, the population size is held constant and hence the population cannot go extinct. The effect of the error threshold is usually studied in the framework of such special-case models (i.e. models with single-peak fitness landscape and absence of lethal mutations). Nonetheless, it is often presented as a general prediction of a quasispecies theory (Nowak 1992; Tannenbaum et al. 2004) and cited as an underlying theory for the concept of lethal mutagenesis for viruses, which has already proved its worth (Crotty et al. 2001; Pariente et al. 2005; Anderson et al. 2004). This, however, is not correct. Although, the concept of lethal mutagenesis seems to be similar to the concept of error catastrophe, these two concepts are not the same: an error catastrophe is a mere dissolution of population information in the sequence space, whereas extinction is a drop in the absolute abundance of individuals in the population which can lead to its complete extinguishing (Bull et al. 2007). It was demonstrated on a simple example (Summers and Litwin 2006) and strictly mathematically proven (Wagner and Krall 1993), that the complete absence of lethal mutations is the necessary condition for the existence of the error threshold and hence, when studied in the framework of the quasispecies model with soft selection, the error threshold per se makes no statements about population extinction.

On the other hand, the abundance of lethal mutations in biological organisms and in particular RNA viruses is now beyond doubt (Sanjuan et al. 2004). So we can expect that alternative models, based on a hard selection process in a finite population would be more adequate for description of lethal mutagenesis. If selection is hard, some organisms become unviable and do not participate in evolutionary competition. Therefore some fraction of a population will always remain in viable areas of the fitness landscape preventing uniform spread of the population in the sequence space and thus the complete loss of information. However, when hard selection is considered the population size will decline as the mutation rate increases and eventually the population can go extinct. Such extinction due to mutation pressure is usually called mutational meltdown (Lynch et al. 1993; Gabriel et al. 1993). Wilke proposed a simple graphical explanation for understanding the difference between the concepts of error threshold and lethal mutagenesis (Figure 3).

Summarizing these problems, I can say that nothing prevents the application of the quasispecies model to finite populations with hard selection. However, multiple additional, non-obvious assumptions are required to construct such models, leading to the difficulty of validation, and loss of generality and reliability.

To draw a line under this discussion on the quasispecies model I would like to say that I do not think the quasispecies theory should oppose classical theoretical population genetics, rather it should be viewed as a subset of it. Furthermore, it was shown that it is mathematically equivalent to the classical theory of mutation-selection balance (Wilke 2005). The main reason why these two concepts are usually considered as separate theories is that for a long time they were developed in parallel by more or

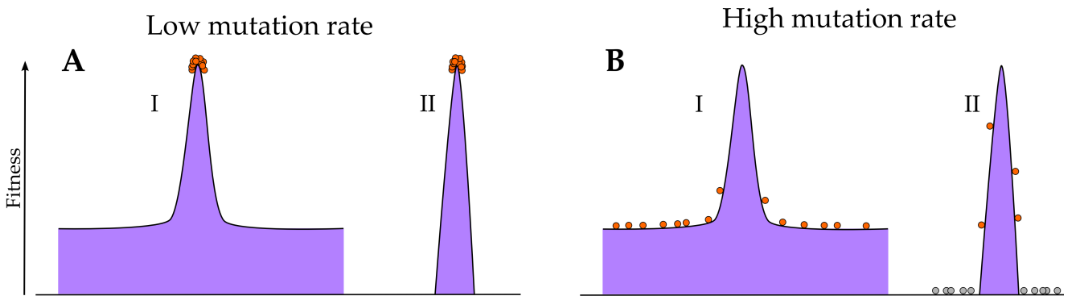

Figure 3. A diagram comparison of error threshold (A-I and B-I, soft selection) and lethal mutagenesis (A-II and B-II, hard selection) concepts under different mutational pressure (Wilke 2005).
If the mutation rate is low (A), both concepts give a similar picture: the population occupies the top of the fitness peak. However, when the rate of mutations becomes high enough (B), the selective strength becomes insufficient to hold back this pressure and the population, unable to retain its position on the peak, is scattered over the sequence space. If in the latter case a fitness landscape has a positive minimum (B-I) then the majority of organisms are pushed to this minimum level, while keeping their ability to reproduce. As a result they are still able to compete with individuals on the peak and win this competition by sheer abundance. On the other hand, if a fitness landscape has no positive minimum level (A-II and B-II), the high rate of mutation pushes a large fraction of the population to zero fitness (B-II). Organisms with zero fitness are unviable (gray dots in B-II), so they do not compete with organisms on the peak and, therefore, some fraction of the population will always remain there. It is worth noting, that all this reasoning is correct only if the population is infinite. Otherwise, stochastic pressure of random drift will move the population away from the peak (effect of Muller's ratchet) and the population will either drift across the sequence space (in case of soft selection, A-I and B-I) or go extinct (in the case of hard selection, A-II and B-II), demonstrating the mutational meltdown effect.

less independent groups of scientists representing different schools. That is why authors have been focused on different phenomena and their aspects, whereas the basic principles of both theories are the same. While the quasispecies model mainly focuses on high-dimensional sequence spaces, explicitly considering multiple loci under high mutational pressure, classical works in population genetics usually treat models with one or two alleles (Wilke and Adami 2003). Another common aspect which distinguishes quasispecies theory from the majority of works on conventional population genetics is the size of the population under investigation. While most current conventional models (based on Kimura's neutral theory) usually consider finite populations, taking into account different stochastic effects, the quasispecies model is originally a deterministic description of infinite populations. However, in recent years a lot of works on quasispecies model also incorporate stochastic components, so the boundaries between quasispecies model and models based on neutral theory became increasingly blurry.

1.3 Information theory. Basic concepts

In contrast with the overwhelming majority of scientific theories, the theory of communication or information theory (IT), as it is usually called, has an identifiable beginning—it is a paper from Claude Shannon called "A mathematical theory of communication", which was published in 1948 (Shannon 1948). One year later the theory appeared in a book by Shannon and Weaver (1949). Shannon was not the first who tried to create a mathematical framework for the data communication process. However, only a few isolated works, mainly focused on some specific applications, touched upon this topic. On the other hand, Shannon's work represents a unifying general idea that revolutionized the area of communication.

In his 1948 paper Shannon suggested innovative ideas and concepts which paved the way for the beginning of the digital age. First, Shannon demonstrated that the actual data content is irrelevant, any information can be represented as a sequence of 0's and 1's. He established two fundamental limits: the limit for possible degree of data compression (Shannon's source coding theorem) and the limit on a speed of the error-free data transmission for a given level of noise, measured in binary digits per second (Shannon's noisy-channel coding theorem). The former pertains to the concept of entropy, the latter relates to the notion of the mutual information and is also known as the channel capacity. Shannon showed that there are actually two alternative ways for reliable transmission of a given amount of information over a noisy channel: to use high power and low bandwidth, or high bandwidth and low power. While increasing the signal's power was at that time a common way to enhance the reliability of communication, the idea of using bandwidth extension as an alternative way to the same goal was groundbreaking.

Nonetheless, the fate of IT was not as easy as it may seem at first glance, and it took some years to fully appreciate the meaning of Shannon's work. In the beginning IT was primarily a theoretical study. Its practical applications were rare and had no essential demand. Indeed, at that time engineers preferred to simply use more bandwidth and power to achieve reliable communication in the presence of noise, instead of using complicated coding schemes for better compression and error correction. In addition, IT suffered heavily from hardware technological limitations: sophisticated coding algorithms require more processing power which was unavailable. These circumstances in turn contributed heavily to the poor funding of the information theoretical research. Eventually, lack of interest, technology and funding led the theorists to announce the "death" of IT in the mid-fifties.

However, the launch of the Sputnik in 1957 changed everything. The space race began and instantly IT received its first serious attention due to the tough efficiency demands in space flight communications. Indeed, signal transmission in deep space has features which make IT perfectly suited for it: foremost among these is the fact that power in space is quite expensive. Secondly, there is plenty of bandwidth which, as was shown by Shannon, is necessary for reliable energy-efficient communication. Extensive financial support from the military and rapid progress of computers and other hardware ensured swift flourishing of IT. Nowadays we use the achievements of Shannon's theory in almost all spheres of our lives often not even realizing it. Although the theoretical basis of IT was completely developed in the first years of its existence, algorithmic aspects continued to evolve dynamically over the last half-century. The rather recent (1993) invention of so called "turbo codes" provided doubling of the transmission efficiency, so significant breakthroughs are still possible and the progress is still far from complete. The story of the information theory is a unique possibility to trace the development of the idea initially expressed in a single theoretical paper that, during only few decades, evolved to a broad scientific field with widespread applications allowing for the beginning of current Digital Age (Aftab et al. 2001).

In the next few sections I will briefly describe some basic results of the information theory (mainly based on (Cover and Thomas 2006)) which are relevant to the framework of the model of molecular evolution developed below in the results section. A more detailed description of all these concepts and many others can be found in (Cover and Thomas 2006), for detailed introduction to information theory with clear examples and historical notes you can refer to (Pierce 1980).

1.3.1 Entropy, joint entropy, relative entropy and mutual information

Entropy and information are two basic concepts in Claude Shannon's information theory. However, in the literature they are often used in a conflicting manner. Below I will give mathematical descriptions of both notions, emphasizing their intuitive features, which, I hope, should facilitate precise understanding and the ability to distinguish between these two concepts.

First I would like to note that the notion of entropy is also present in thermodynamics and statistical mechanics. Although all three of these concepts were designed to describe phenomena in essentially different areas, they are somewhat similar. Sometimes this can lead to confusion. So to avoid misunderstanding, the entropy arising in information theory is usually called information entropy. Since in this work we are only interested in the specific aspects of this concept, for brevity I will use the term entropy as a synonym of information entropy.

Information entropy is usually defined as a measure of uncertainty (unpredictability or variability) of a random variable. Sometimes it is also called a self-information of a random variable, because, as shown below, it is equal to the amount of information required to describe the random variable. Let's consider discrete random variable X taking values from the set Ω_X and having probability mass function $p_X(x) = \Pr(X = x)$, $x \in \Omega_X$. Then the entropy of X is defined as:

$$H(X) = -\sum_{x \in \Omega_X} p_X(x) \log_2 p_X(x) \tag{3}$$

Further in this work I will always use the base of the logarithm equal to 2, which corresponds to the entropy measured in bits. Other typical units for entropy measurement are nats and bans, corresponding to logarithm bases e and 10. Also I will use the convention that $0 \log 0 = 0$. This can be easily verified using L'Hôpital's rule:

$$\begin{aligned}\lim_{x \to 0}(x \log x) &= \lim_{x \to 0}\left(\log x \middle/ \frac{1}{x}\right) = \lim_{x \to 0}\, (\log x)' \middle/ \left(\frac{1}{x}\right)' = \lim_{x \to 0}\left(\frac{1}{x}\right) \middle/ \left(-\frac{1}{x^2}\right) \\ &= \lim_{x \to 0} x = 0\end{aligned} \tag{4}$$

Thus, adding terms of zero probability does not change the entropy. It is also worth pointing out that entropy is a function only of the distribution of X, so it does not depend on the actual values taken by the random variable X, but on the probabilities. From logical considerations it is clear that the result of a random trial is most unpredictable when all possible outcomes are equally likely. Thus, the entropy of probability distribution is maximal if all outcomes are equiprobable. So for discrete random variable X with finite number of outcomes $|\Omega| = M$ the entropy can be bounded from above: $H(X) \leq -\sum_{i=1}^{M} \frac{1}{M} \log \frac{1}{M} = \log M$, with equality if and only if all outcomes have the same probability of $\frac{1}{M}$. For example the maximum entropy of binary random variable is 1 bit, corresponding to the Bernoulli distribution with probability of success $p = 0.5$. The above mentioned properties of the entropy show that the measure of uncertainty defined in this way naturally possesses many intuitive features.

In the same manner as we defined the entropy of a single random variable, we can define the entropy of a pair of random variables. The joint entropy $H(X,Y)$ of a pair of discrete random variables with a joint distribution $p_{XY}(x,y)$ is defined as:

$$H(X,Y) = -\sum_{x\in\Omega_X}\sum_{y\in\Omega_Y} p_{XY(x,y)} \log p_{XY}(x,y) \tag{5}$$

There is nothing actually new in this definition because we can consider a pair (X,Y) as a single vector-valued random variable. In the same way entropy of multiple random variables can be considered.

Also we can define conditional entropy of a random variable Y given another random variable X as an expected value of the entropies of the conditional distributions, averaged over the conditioning random variable:

$$\begin{aligned} H(Y|X) &= \sum_{x\in\Omega_X} p_X(x) H(Y|X=x) = \\ &= -\sum_{x\in\Omega_X} p_X(x) \sum_{y\in\Omega_Y} p_Y(y|x) \log p_Y(y|x) = \\ &= -\sum_{x\in\Omega_X}\sum_{y\in\Omega_Y} p_{XY}(x,y) \log p_Y(y|x) \end{aligned} \tag{6}$$

An intuitive property of conditional entropy is that it is always less than or equal to the unconditional: $H(Y|X) \leq H(Y)$, with equality if and only if X and Y are independent. Therefore, additional information cannot hurt.

The naturalness of the definition of joint entropy and conditional entropy is also manifested by the fact that the entropy of a pair of random variables is equal to the entropy of one plus the conditional entropy of the another: $H(X,Y) = H(X) + H(Y|X)$. So, entropy is an additive function if its arguments $X_1, X_2, \dots, X_N$ are independent random variables:

$$H(X_1, X_2, \dots, X_N) = \sum_{i=1}^{N} H(X_i) \tag{7}$$

For further discussion we will also need the notion called mutual information. Consider two random variables X and Y with a joint probability mass function $\mathrm{p}_{XY}(x,y)$ and marginal probability mass functions $p_X(x)$ and $p_Y(y)$. The mutual information $I(X,Y)$ is expressed as:

$$\begin{aligned} I(X,Y) &= \sum_{x\in\Omega_X} \sum_{y\in\Omega_Y} p_{XY(x,y)} \log \frac{p_{XY}(x,y)}{p_X(x)p_Y(y)} = H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned} \tag{8}$$

It is easy to see that mutual information is a symmetric function. It represents the reduction in the uncertainty (entropy) of X or Y due to the knowledge of another, Y or X correspondingly. For example, for independent X and Y mutual information $I(X,Y)$ is equal to zero. On the other hand, if $X = Y$ then $I(X,X) = H(X)$ (that is the reason why the entropy is sometimes called self-information). Speaking simply if $I(X,Y) > 0$, then the knowledge of one variable makes the predictions about the other more accurate, this is exactly what we mean by information in our common language. Summarizing, I want to point out that information can only be defined relative to two systems (random variables), it measures the correlation between these two systems and, therefore, in contrast with entropy, the information is always about something and never absolute.

1.3.2 Asymptotic equipartition property

Asymptotic equipartition property (AEP) is one of the fundamental results of the information theory. It is an analog of the law of large numbers in the IT and could be obtained as a direct consequence of the weak law of large numbers. The weak law of large numbers states that if we take a sum of N independent and identically distributed (i.i.d.) random variables (X_1, X_2, …, X_N), divided by N it will converge to the expected value of X as N approaches infinity: $\lim_{N\to\infty} \frac{1}{N} \sum_i X_i = E(X)$. In turn, AEP (in its original form) states that the value of $\frac{1}{N} \log \frac{1}{p_X(X_1, X_2, \dots, X_N)}$ converges in probability to the entropy $H(X)$, as $N \to \infty$, where $\mathrm{p}_X(X_1, X_2, \dots, X_N)$ is the probability of observing the sequence X_1, X_2, …, X_N. Speaking informally the probability of any actually observed sequence X_1, X_2, …, X_N will be close to $2^{-NH(X)}$ when the number of observations is large enough.

This fact become clear if we note that entropy can be also interpreted as the expected value of random variable $\log\frac{1}{p_X(X)}$, where X is drawn according to the probability mass function $p_X(x)$: $H(X) = E_{p_X}\left(\log\frac{1}{p_X(X)}\right)$. Now applying the weak law of large number to a sequence of i.i.d. random variables $\left(\log\frac{1}{p_X(X_1)}, \log\frac{1}{p_X(X_2)}, \dots, \log\frac{1}{p_X(X_N)}\right)$ we immediately obtain the statement of the AEP.

The requirement of independence and identical distribution are necessary for the proof on the basis of the weak law of large numbers. However, it is intuitively clear that the statement of the AEP is more general and the assumption of identical distribution is not essential for the AEP to hold. If identical distribution is not assumed, we need just a slight reformulation of the statement: $\frac{1}{N}\log\frac{1}{p_X(X_1, X_2, \dots, X_N)} \xrightarrow[N\to\infty]{} \bar{H}(X)$ in probability, where X_1, X_2,…, X_N are independent random variables and $\bar{H}(X) = \frac{1}{N}H(X_1, X_2, \dots, X_N) = \frac{1}{N}\sum H(X_i)$ is an average entropy. The case when the variables are independent but not identically distributed will be exploited in the model developed in the results section below. Additionally, it was proven that the requirement of independence can also be relaxed - for details see (Girardin 2005).

1.3.3 Typical and high-probability sets

For simplicity in this and the next section, I will consider the original form of AEP and thus use sequences of independent identically distributed random variables. However, all presented results can be easily generalized on the case of non-identical distribution.

AEP allows us to divide the set of all sequences $S_N = (X_1, X_2, \dots, X_N)$ into two sets. The first (typical) set contains elements with probabilities close to $2^{-NH(X)} = 2^{-H(S_N)}$, where $H(S_N) = H(X_1, X_2, \dots, X_N) = \sum H(X_i) = NH(X)$. The second (nontypical) set contains all other sequences. AEP states that as N grows the probability to get typical sequence converges to 1. Thus most of our attention will be concentrated on the typical sequences. Since any property that is true for the typical sequences will then be true with high probability and will determine the average behavior of the sequence, assuming that it is long enough.

To clarify this notion let's consider a simple example. Suppose we have a binary random variable $X \in \{0, 1\}$ with probability mass function $p(x)$, $x \in X$, defined as $p(1)$ = p and $p(0)$ = 1—p = q. If X_1, X_2, …, X_N are i.i.d. according to $p(x)$, the probability of sequence x_1, x_2, …, x_N is $\prod_{i=1}^{N} p(x_i)$. E.g., the probability of (0, 1, 0, 0, 1 0, 1) is $\mathrm{p}^{\sum X_i}\mathrm{q}^{N-\sum X_i} = \mathrm{p}^3\mathrm{q}^4$. There are in total 2^N sequences of length N, and it is clear that in general not all of them have the same probability. However, if we ask for the probability $p(X_1, X_2, \dots, X_N)$ of the outcomes $X_1, X_2, \dots, X_N$, where X_i, i = 1,2, .., N are i.i.d. distributed according to $p(x)$, what rational prediction can we make? It is reasonable to expect that the observed sequence will contain the number of 1's close to Np, thus the probability of such sequence will be close to $\mathrm{p}^{N\mathrm{p}}\mathrm{q}^{N\mathrm{q}}$. The entropy of this probability distribution is: $H(X) = -\mathrm{p}\ \log \mathrm{p} - \mathrm{q} \log \mathrm{q}$. Now let's note that $\mathrm{p}^{N\mathrm{p}}\mathrm{q}^{N\mathrm{q}}$ = $2^{-\ NH(X)}$. So expected sequence has a probability close to $2^{-\ NH(X)}$ and, according to AEP, the longer the sequence we observe (bigger N) is, the larger the fraction of sequences having probability close to $2^{-\ NH(X)}$ in the set of all possible sequences of length N.

In the two previous paragraphs I tried to give a somewhat informal explanation to the concept of typical set. In fact, in the framework of the thesis this explanation is sufficient. However, for correctness, I will also give the strict definition. The typical set $A_\epsilon^{(N)}$ with respect to the probability distribution with mass function $p_X(x)$ is the set of sequences (x_1, x_2,…, x_N) ∊ Ω_X^N with the property: $2^{-N(H(X)+\epsilon)} \leq p(x_1, x_2, \dots, x_N) \leq 2^{-N(H(X)-\epsilon)}$. Taking this, the simple corollary from the AEP statement is that for rather large values of N (for long enough sequences) the typical set has probability close to 1, all elements of the typical set are nearly equiprobable, and the number of elements in the typical set is nearly $2^{\ NH(X)}$. Considering the example with Bernoulli random variable X having probability of success equal to p, we can see that the size of its typical set is equal to the number of all possible sequences of length N if and only if $H(X)$ = 1. This, in turn, is the case only if p = 0.5. In all other cases the size of typical set is smaller than the number of all possible sequences.

Let's also introduce the notion of a high-probability set $B_\delta^{(N)}$. For each N = 1, 2, …, let $B_\delta^{(N)} \subset \Omega_X^N$ be the smallest set satisfying the condition: $Pr\left\{B_\delta^{(N)}\right\} \geq 1 - \delta$. To illustrate the difference between a typical set ($A_\epsilon^{(N)}$) and a high-probability set ($B_\delta^{(N)}$), let's once again consider a sequence of Bernoulli random variables X_1, X_2, …, X_N with parameter p = 0.8. The typical sequences in this case are the sequences with the fraction of 1's close to 0.8. However, this does not include the most likely single sequence, which is the sequence of all 1's. On the other hand the set $B_\delta^{(N)}$ by definition contains all the most probable sequences and therefore includes the sequence of all 1's.

From the definition of the typical set it is clear that it is a rather small set that concentrates most of the probability. It is not clear, however, whether this set is the smallest such set and how it compares to a high-probability set. From the definition of high-probability set it is clear, that for a given level of probability concentration δ and sequence length N, the number of elements in $B_\delta^{(N)}$ is less than or equal than in minimal typical set $A_{\epsilon_{min}}^{(N)}$ having the same probability, i.e. ∀ δ and N $\left|B_\delta^{(N)}\right| \le \left|A_{\epsilon_{min}}^{(N)}\right|$, where $A_{\epsilon_{min}}^{(N)} = \min_\epsilon \left\{A_\epsilon^{(N)} \middle| Pr\left\{A_\epsilon^{(N)}\right\} \ge 1-\delta\right\}$. However, it can be shown that $A_{\epsilon_{min}}^{(N)}$ and $B_\delta^{(N)}$ have significant intersection and, moreover, have about the same number of elements. Speaking more strictly, it is proved that the typical set has essentially the same number of elements as the high-probability set, to the first order in the exponent.

1.3.4 Channel capacity

Another important concept of the IT is the Channel Capacity. To define it we first need a definition of the communication channel itself. Generally speaking, a communication channel is a physical or logical transition medium which is used to convey information from the transmitter (T) to the receiver (R). Here only a case of the memory-less discrete channel will be considered, because it is exactly the case we need in this work. A discrete memory-less channel is a system comprising input (source) alphabet Ω_X, output alphabet Ω_Y and a transition probabilities $p(y|x)$ that defines the probability of observing the output symbol y given that the symbol x was sent, where the probability distribution of the output depends only on the current input and does not depend on previous input or output symbols. So T communicates with R means that T induces some state in R. In reality any communication is a physical action which causes transmitted information to be affected by the ambient noise as well as by distortions induced by sender or receiver. Communication is considered successful if the transmitter T and the receiver R agree on what was sent. The schematic diagram of physical communication system is presented in Figure 4.

An initially transmitted message (M) is encoded using source alphabet (Ω_X). Then the encoded sequence (X_1, X_2, …, X_N) is transmitted through communication channel, which maps source symbols from input sequence into the sequence (Y_1, Y_2, …, Y_N)

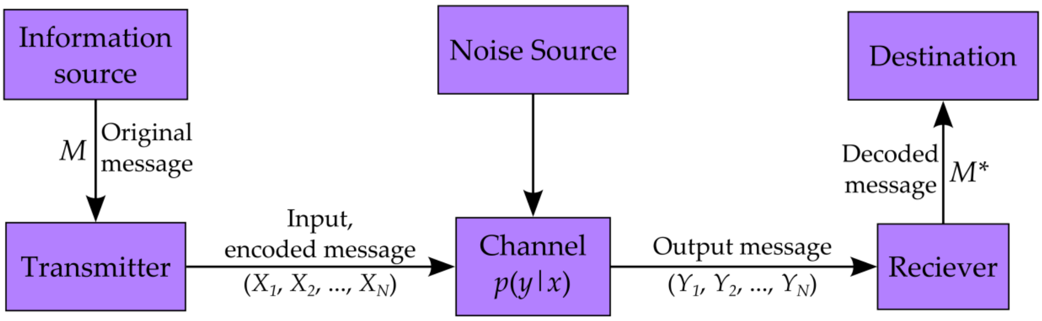

Figure 4. Schematic diagram of a general communication system.

of output symbols (Ω_Y). Usually, due to the presence of noise this correspondence is not exact, thus the output symbols are random but their distribution depends on input ($p(y|x)$). After receiving the output sequence, we attempt to decode the received (output) message and recover the original transmitted message. If the communication is error free, having an output sequence we can easily identify the required input sequence. In this case the reconstruction of the original message is a trivial task (because the procedure of message encoding is assumed to be a known deterministic process). However, the presence of noise in the system (which is always the case in real communication) can hinder the unambiguous reconstruction of the original sequence since each possible input sequence induces a probability distribution on the output sequences.

In general, the same input sequences can be mapped into different output sequences and different input sequences may give rise to the same output sequence. It is intuitively clear that higher levels of noise generate wider output distributions of input sequences with larger overlap with each other, which in turn makes recognition of the original message more difficult. To avoid this ambiguity we should add redundancy in the input coding procedure, i.e. expand spaces of potential input and output sequences (often, this is the same space, i.e. $\Omega_X = \Omega_Y$) and choose our input sequences to be rather dissimilar, so even after distortion, induced by different sources of noise, we can still identify with high confidence from what input sequence each particular output came. In his noisy-channel coding theorem Shannon showed (Shannon 1948) that by mapping the source into the appropriate "widely spaced" input sequences to the channel, it is possible to transmit a message with arbitrary low probability of error and reconstruct the source message at the output. However, this procedure usually requires increasing the length of code words, so the encoded (input) message becomes longer and the overall rate of transmission decreases. The theoretical maximum rate of reliable information transfer through the channel, for a particular noise level is called the capacity of the channel. The formal definition of the channel capacity (C) for discrete memory-less channel, defined above, can be formulated as follows:

$$C = \max_{p(x)} I(X, Y) \tag{9}$$

where the maximum is taken over all possible input distributions $p(x)$.

1.4 Information theory and molecular genetics

Application of IT approaches in molecular genetics has a long history. Although it is not directly related to the topic of this section, it is interesting and worth noting here that the PhD thesis of Claude Shannon (the father of IT) was devoted to the development of an algebra for studying dynamics of Mendelian populations. However, immediately after receiving his degree, Shannon went to work for Bell Laboratories, where he began his studies in the field of communications and never returned to population genetics (Crow 2001).

During the last 60 years a lot of work focused on various aspects of molecular genetics was done using information theory as a theoretical framework. For instance, Adami and Cerf (2000) studied information content of bacterial tRNA. IT-based tools were also suggested and effectively applied for determining secondary and tertiary structures of nucleic acid sequences (in particular tRNA and different subunits of rRNA) (Chiu and Kolodziejczak 1991; Gutell et al. 1992) and proteins (Gibrat et al. 1987; Thompson and Goldstein 1997). Another pressing concern where approaches utilizing concepts of information theory were successfully applied is the identification of DNA protein-binding regions (promoters, enhancers, splicing sites etc.) (Erill and O'Neill 2009) and protein-protein interactions. It was also demonstrated that IT methods can be useful in rational targeted drug design (Hamacher 2007).

Most of above mentioned works represent ad hock algorithms and therefore have purely practical orientation. However, it was shown that IT based approach gives possibility to reveal obscured general regularities of the genetic information structure. Gatlin (1966, 1968) suggested a coefficient based on the entropy of nucleotides distribution along the genome sequence (horizontal entropy). Gatlin's model actually represents a first order Markov chain. Given a probability distribution $p(x)$ and conditional probability distribution $p(y|x)$, where x is a nucleotide in a given genome and y is a nucleotide which follows x, he calculated the value $K = \mathrm{H}(X)—\mathrm{H}(Y|X)$, where X and Y are random variables distributed correspondingly as $p(x)$ and $p(y|x)$. So, the parameter K can be interpreted as a divergence from independence. As we see, the model is quite simple. However, Gatlin showed that broad classes of organisms (phage, bacteria and vertebrates) form clear clusters according to the value of this parameter, so, potentially it can reflect the emergence of deep forces acting behind the scene and shaping information structure of genomes. In more recent research Grosse et al. (2000) also studied horizontal correlations between nucleotides. He proposed a parameter which is based on mutual information and showed that its usage provides a possibility to discriminate protein coding regions from noncoding ones. In contrast with previous studies pursuing the same goal, here the authors were able to demonstrate that their approach is species independent.

Let's now move closer to the topic of this thesis. One of the key tasks in molecular genetics, which is also relevant to the material presented in the results section, is finding conserved regions in DNA, RNA or protein sequences. This problem is important because it is reasonable to suggest that a well-preserved sequence is (or was, not so long ago) under selective pressure and thus has (or had) a biological function (Fabris 2008). However, as is already known, nature has developed a variety of reliable methods for tolerating mutations in sequences without affecting their functionality: redundancy of amino acid protein sequences and functional DNA sites (in particular, the most well-known example is the redundancy of genetic code, however, noncoding functional sequences can possess much stronger redundancy), different intricate error-correction mechanisms etc. Thus, functional sequences are usually not absolutely conserved, but to some extent admit variability (sometimes rather strong). So the task of finding sequences which have the same function as the assigned probe becomes less trivial because we have to find not only completely identical sequences, but also sequences which are similar to our probe in some suitable sense. IT has proved to be a useful tool for the identification of these hypothetically meaningful sequence polymorphisms. However, in contrast with studies of Grosse (2000) (clustering of different groups of organisms) and Gatlin (1966, 1968) (discriminating between coding and noncoding regions of the genome) horizontal correlations alone cannot help in discovering functional polymorphic sequences, because the meaning of the sequence emerges only through the interaction with environment (i.e. the meaning is relative). Horizontal variability tells us nothing about actual information carried by the sequence, because a sequence per se does not bear any information about its function, rather this information is stored in correlations (interactions) between the sequence and environment which it describes or to which it corresponds (Adami 2004). The environment for the sequence can be given by the binding proteins and other intra-cellular components which somehow interplay with the sequence. In the light of the arguments expounded above, it seems apparent that we may never detect functional polymorphic sequences if we are only given a single sequence. However, it is possible to do so if we consider vertical entropies in the set of functionally equivalent sequences, i.e. aligning sequences from this set under each other and studying the patterns of nucleotide substitutions in each position. This strategy was initially introduced by Schneider et al. (1986) who investigated ribosome binding sites. His approach, though often criticized (e.g. see page 46 in (Yockey 2005)), proved its consistency.

Although, as we just saw, there are attempts to apply IT to a wide range of issues arising in molecular genetics, IT is not considered a conventional approach and is not widely used. Besides, almost all above mentioned works have an applied focus, whereas it is quite tempting to apply IT principals to elaborate a general theoretical model of information fluctuations in biological sequences.

2 RESULTS

2.1 The model of positional information storage in sequence patterns

> *Essentially, all models are wrong, but some are useful.*
>
> Box GEP and Draper NR (1987) Empirical Model Building and Response Surfaces, p. 424.

At the dawn of the era of molecular biology there was a strong held belief that the vast majority of any genome is occupied by genes. However, by the end of 1960s it became apparent that protein coding regions of genomes alternate with noncoding regions, and in large genomes the former usually constitutes only a minor fraction. The human genome was always of particular interest and at that time there were fierce debates over what fraction thereof is functional. Several different estimates were proposed, however, all agreed that the fraction of protein coding sequences within the human genome does not exceed 10%. In the beginning of 1970s Susumu Ohno published two landmark works in which he tried to explain the origin and role of noncoding DNA (Ohno 1970, 1972). According to his assumption, the majority of noncoding DNA originates from gene duplication followed by sequence degeneration and eventual loss of function by the duplicates. In line with this view he called noncoding sequence the "junk" DNA. Generally, this point of view was supported by the broad scientific community, and the insight into genome functionality was mainly restricted to protein-coding sequences. It is worth mentioning that the factors which impede our understanding of noncoding functionality lay mostly in the technical area. Genes are conserved, well-structured and have a universal mechanism of information encoding: the genetic code, which was discovered in the beginning of 1960s. Because of these properties, genes can be localized easily and investigated both experimentally and computationally. On the other hand, there is still no generic concept of informational storage for noncoding functional DNA, which looks quite variable, unstructured and thereby, difficult to recognize and study. In the light of all above mentioned it is not surprising that the majority of experimental research for decades was focused on the investigation of coding regions, while noncoding sequences were usually treated as nonfunctional garbage simply undergoing random drift. This approach created strong observational bias affecting theoretical studies and led to misconceptions about mechanisms of molecular evolution. In this regard classical models of molecular evolution are no exception as they were designed and verified according to contemporary data. They were tuned to describe the evolution of coding regions, where sequence conservation is reliable evidence of functionality, and noncoding DNA was primarily considered evolutionary junk.

Today our view of DNA functionality is gradually changing. Current technological advances provide an opportunity to look much deeper into genome functionality than was possible even a decade ago. The concept of “junk” DNA still holds true to some extent (the human genome is full of pseudogenes). However, it is apparent now that the picture is much more complex than was previously believed. While protein coding genes usually make up more than 90% of bacterial or archaeal genomes, their fraction in the human genome is only about 1.2% (The ENCODE Project Consortium 2012). There was never a doubt that at least some fraction of noncoding DNA is functional. Regions of noncoding sequence enriched with binding sites (promotors, enhancers, introns and etc.) and microRNA provides a mechanism for regulation of gene expression, ribosomal and transfer RNA mediates protein synthesis. These are just some of the more well-known examples of noncoding functionality. However, what fraction of noncoding DNA in human genome is functional remains largely unclear.

Recent extensive studies in the framework of The ENCODE (The Encyclopedia of DNA Elements) Project Consortium (2012) revealed that the proportion of the human genome potentially possessing some function is much higher than it was thought before, reaching 80.4%. This number is probably an overestimation. However, even for the most conservative estimates, the fraction of the human genome likely involved in direct gene regulation is about 8.5%, significantly exceeding the proportion ascribed to protein coding sequences (The ENCODE Project Consortium 2012). With these data, it is reasonable to suggest that more information in the human genome is responsible for regulation of gene expression than for biochemical properties of protein sequences. This fact can be also indirectly supported by the evidence that less than 10% of disease-associated single nucleotide polymorphisms (SNPs) are located in protein-coding exons, while the rest remains in intronic and intergenic regions (Kumar et al. 2013). So it is now beyond doubt that a significant fraction of noncoding DNA serves some purpose and thus is not “junk”. However, the majority of functional noncoding elements exhibit strong variability (of the sequence) not only between different species but also within single species, thus avoiding reliable detection by conventional conservation-based methods. On the other hand, there is evidence that deletion of some ultraconserved elements from the mouse genome does not have any apparent effect on the phenotype (Nóbrega et al. 2004; Ahituv et al.

2007). It is now becoming clear that conservation-based approaches in the investigation of functional elements within the genome have led to overemphasizing the importance of sequence conservation, and that sequence conservation, while being a useful indicator, is neither necessary nor a sufficient criterion of functionality (Elgar and Vavouri 2008). Despite preliminary efforts to characterize noncoding functional elements, understanding how information can be reliably stored in sequences possessing such a strong variability is still strongly lacking. However, at least one appropriate solution for this problem was already demonstrated by Claude Shannon more than 60 years ago in the framework of the information theory (IT).

As it was discussed in the introduction, Shannon considered an artificial communication system where a useful signal is clouded with noise. He showed that there are two alternatives to achieve reliable recognition of the message: increase the power of the signal or increase the redundancy of the coding scheme. While the former approach is straight and obvious (it was well-known long before Shannon's work and for a long time remained the only approach used by communication engineers) the latter is not so intuitive and it took significant amount of time to see its true value. Genetic information is encoded by nucleotide sequences. This information is transmitted from generation to generation, forming a kind of natural communication system. Drawing the analogy between artificial and natural communication systems further, we can suggest that different parts of genetic information are encoded using different coding systems. In particular "natural" codes can possess different degree of redundancy. It is worth noting that the only DNA code we truly understand now is the genetic code that provides correspondence between DNA sequence and amino acid sequence of proteins. As a rule, alterations in an amino acid sequence change the chemical properties of protein and disrupt functionality leading to nonviable phenotypes. Although the genetic code is redundant (each codon is a triplet of nucleotides, thus 4^3=64 codons are mapped to 20 "standard" amino acids and stop codons), 3 bases per codon is the most compressed configuration to encode a set of "standard" amino acids and a stop signal unambiguously, using 4 different coding symbols (nucleotides). Such conciseness of the genetic code is probably dictated by the tough demand for energy efficiency and other physical constraints in the early stages of the evolution of life, when the current variant of the genetic code was presumably shaped and stabilized at least since the last universal common ancestor of all modern (cellular) life forms (Koonin and Novozhilov 2009).

Once the genetic code becomes frozen, adding redundancy is quite cumbersome and inevitably leads to loss of viability. So, reasoning in the framework of IT, we can presume that it is very likely that nature, in its unconscious aspiration to efficiency, was compelled to enhance the reliability of transmission of protein coding information by “increasing the power of the signal”, which in biological terms corresponds to conservation of the sequence. That is exactly what we see—protein coding genes usually possess very high levels of conservation. On the other hand, as we already know, noncoding functional sequences can be extremely variable. This, however, does not mean that information contained in these sequences is not important. Presumably, regulatory mechanisms encoded in noncoding DNA are not subjected to such strong evolutionary constraints as are core protein coding sequences. So they are much more flexible, having the ability to evolve gradually while at each moment maintaining sufficient fitness of the population. According to ideas developed in IT, we can suggest that the nature was able to resort to alternative approach achieving reliability of noncoding genetic information transfer. Instead of maintaining high level of sequences conservation, which is apparently a very costly procedure, more redundant coding system is applied. As already known from the lesson of IT, a message, encoded with high redundancy can tolerate a lot of mutations without loss of meaning thus demonstrating high variability while preserving transmitted information. That is what we presumably observe in noncoding functional sequences.

In this section I try to adapt Shannon’s approach to molecular genetics and show the possible mechanism of reliable maintenance and evolution of genetic information in variable sequences. I suggest a model where genetic information represents well-defined values, expressed through the positional variability in the sequence pattern specified by the set of given (functional) sequences (e.g. nucleotide sequences of DNA/RNA or amino acid sequences of proteins). Looking ahead I would like to mention that in this framework it is convenient to consider a “sequence pattern” as a “fuzzy sequence”, where each position represents the probability distribution of nucleotides. When a pattern is defined, the effect of any mutation can be estimated. In the framework of the model it is possible to show that the frequency of beneficial mutations can be high in general and the same mutation, depending on the pattern’s context, can be either advantageous or deleterious. The model allows treating positional information (i.e. information related to a specific position in the corresponding sequence pattern) as a physical quantity, formulating its conservation law and modeling its continuous evolution in a whole genome. It provides the possibility for meaningful applications of basic physical principles such as optimal efficiency and channel capacity. I also suggest a plausible, according to the model, option for optimization of information storage, formulate it in strict mathematical form as a minimization problem, derive its solution and demonstrate experimental evidences of this phenomenon. The model shows that, in principle, it is possible to store error-free information in sequences with arbitrary low conservation. The suggested approach shares features both from neutralism and selectionism. However, it possesses considerable originality and thus cannot be directly attributed to any conventional theoretical schools. The described theoretical framework allows one to approach, from a novel general perspective, such long-standing paradoxes as excessive "junk" DNA in large genomes or the corresponding G- and C-values paradoxes. It can also shed light on some fundamental concepts of population genetics, including the cost-of-selection dilemma, error catastrophe and others. The model of positional information storage in sequence patterns presented in the next paragraphs was published in June 2013 (Shadrin et al. 2013).

2.1.1 Prior discussion and assumptions of the model

Information theory originally describes the process of sending discrete data over a noisy channel. It was already pointed out above that this process seems to be quite similar to transmitting DNA sequences through generations with mutational errors. Some applications of IT to molecular biology were attempted in order to exploit this similarity: for review see (Johnson 1970; Yockey 2005). However, despite the revolutionary role of the IT in communications engineering and the strong analogy between DNA sequences and discrete messages, the use of the IT in molecular genetics is disappointingly limited. In his seminal work Eigen (1971) indicated that the main challenge faced by researchers in adapting IT to the molecular genetics is how best to quantify the biological value of a sequence. The value that counts is the transmission of sequence (or a pattern, in the case of our model) to next generations. As was discussed in the introduction section, it is sensible to speak about information only in the context of system where there are more than one object interacting. To give a physical meaning for information we should provide a context, i.e. make information "relational" (information should be for something and about something). In the original IT problems the information is defined as a degree of correlation between sender and receiver in communication system. In the model presented I suggest considering an interaction of 3D molecular shapes between interacting molecules as an environment for information, where the degree of specificity of an interaction signifies corresponding amount of information.

All molecular interactions can be thought as being more or less a specific search ("homing") for an interacting partner with subsequent "docking" and energy dissipation. Perhaps the most remarkable and well-known examples of specific molecular interactions can be found in biological objects. For instance let's consider a "binding factor"—a protein (or protein complex) which seeks and binds to a specific spot (sequence) on DNA ("binding site") in order to regulate an expression of some gene. Usually binding sites are located in noncoding regions of DNA and can vary a lot while retaining their function, i.e. many binding factors recognize not a single specific sequence but a large set of sequences that have certain properties, forming the pattern for recognition. In other words a binding site for each binding factor is encoded with high redundancy. As mentioned above, the fact that a single binding factor is able to recognize a set of different binding sites could be explained by important IT-related reasons (later we will return to the discussion of this problem in more details). The theoretical model presented here describes the evolution of such sets and corresponding patterns. Using abundant and well-annotated splice sites of a several mammalian genomes I was able to provide support for my conclusions.

Previous applications of IT in molecular genetics were mainly focused on the problems of binding sites and factors operations in a genome. Combinatorial and thermodynamic properties of binding, such as their specific recognition mechanisms, were addressed by von Hippel and Berg (1986). Later these authors also attempted to adapt concepts of statistical-mechanics for descriptions of molecular affinity and binding dynamics (Berg and von Hippel 1987). Stormo (2000) focused on the purely practical problems of computational prediction and analysis of binding sites. Conversely, in this work I want to shift to the higher level of abstraction, presuming that the patterns for recognition are already established and known. Then, in the framework of this assumption, fundamental aspects of molecular evolution can be addressed, namely: how such patterns are maintained and evolve through generations, or how random mutations in patterns are redistributed between negative and advantageous. With enough data, the generality of this approach makes it possible to model information dynamics not of a single binding site but of the whole genome. Berg et al. (2004) studied properties of the molecular evolution of binding sites, however, in the investigation they used the classical model of adaptive evolution (i.e. evolution through variants amplification and fixation). Frank (2012) demonstrated another interesting approach to connect genetic information with environmental information but it was also based on the classical adaptive evolution formalism.

While the approaches for modeling adaptive selection are numerous and well-developed, the formalism for purifying (maintenance) selection are practically absent because this problem is usually considered merely a removal of negative mutations and is thus of no particular interest. This point of view likely holds true if sites under strong evolutionary constraints, where almost each mutation is strongly deleterious, are considered. Nonetheless, here I argue that in the case of functional sequences possessing sufficiently high variability (i.e. functional elements encoded with rather strong redundancy) the impact of maintenance evolution becomes significant, playing the dominant role in formation of the observed sequence compound. As far as I know the issue of pattern maintenance without progressive evolution, where a population acts as a "digital repeater" of stored information, has not been explored before. The preservation of genetic information through stability of the pattern, formed by the set of functionally allowed sequences, and achieved by keeping constant positional allele frequencies, has not been suggested before and represents a novel concept in population genetics. Another interesting and somewhat surprising thing I show here is that purifying selection without need of allele amplification (fixation) can potentially enjoy abundant beneficial mutations in functional sequences, allowing for high levels of variability. It will be shown how positive mutations can significantly ease the effort of removing negative alleles and that the pattern composition can be adjusted in order to optimize mutational load under different circumstances. Here I use the term "beneficial (positive) mutation" in a broader sense than it is commonly used. This extension of the concept is made in accordance with the proposed model and is discussed in details below. It is also important to mention that the problem of choice from a set (i.e. searching for a site in a genome) per se does not require a comprehensive application of IT and can be considered a simple combinatorial problem. Typical set, channel capacity and asymptotic equipartition property (in its basic version called Shannon-McMillan-Breiman theorem) are core and the most prominent concepts of IT and paved its way to the worldwide success. It is therefore quite tempting to adapt these notions to molecular and population genetics. However, neither of these concepts was applied for the positional information in molecular genetics. Therefore in the present work I make an attempt to integrate fundamental IT approaches into models of genetics.

2.1.2 Measure of genetic information

The genetic information (GI) can be viewed as positional in a very general sense. It defines the process of homing and consequent specific binding between molecules which also includes a binding of a molecule to itself, which is quite a common event for RNAs and proteins (i.e. secondary and tertiary structures). These processes transform sequential (one-dimensional) DNA information into 3D shapes and external energy inflow supplies binding/unbinding kinetics, unfolding the temporal dimension. An organized dynamic 3D structure with hereditary information stored in a molecular sequence provides all the basic "physical" properties of living system.

In this work I often use DNA binding sites to illustrate different phenomena. However, binding sites are merely convenient objects for visual demonstration of general phenomena. For example, the process of protein synthesis starts from transcription of DNA template; transcription in turn can be represented as a series of various homing, binding and unbinding events, therefore the notion of positional information is quite universal.

Now let's resort to an analogy again. Imagine we have an engineering problem: we have a set (population) of mutable self-replicating symbolic sequences (e.g. nucleotide sequences) and we want to maintain positional information during the course of iterative rounds of replication with mutagenesis and selection. The nature of sequences does not matter for a mathematical formulation of the problem. However, having our final goal in mind, for further simplicity I will consider and perform all calculations for nucleotide sequences. So each position in the sequence can take four different values corresponding to one of four nucleotides: A (adenine), G (guanine), C (cytosine), T (thymine). For our purposes we can design recognizers that respond to specific sub-sequences (e.g. binding factors and binding sites correspondingly). For instance if we want a sub-sequence (binding site) to determine a unique position in a (quasi-random) sequence of length L, we must use at least $\log_2 L$ bits of information, which requires half of this number of nucleotide positions ($\log_2 L/2$) to define (as far as each nucleotide position has 4 different states thus providing 2 bits). The total number of unique binding sites of this length obviously is $4^{\log_2 L/2} = L$. However, in this case maintaining of stability in the system will be quite challenging because any mutation in the site will break the recognition, thereby erasing the information. As such, information can be retained only if all mutations are avoided. This is possible if the mutation rate is sufficiently low while the rate of reproduction is high. In this case at least some of the progeny sequences will have no mutations, so the information can be maintained by discarding all mutated sequences. The latter case can be considered an extreme example of purifying selection. It is clear that this (trivial) mode of maintenance in reality can be potentially accomplished only in the very small genomes of microbes. However, usually mutations in binding sites cannot be avoided. What can be done in this situation? To resolve this problem we should deploy redundant coding in terms of IT, as it was discussed above. So, after adding redundancy to the code, a binding factor must recognize not a single sequence but a set of ("synonymous") sequences. When information is stored in redundant patterns there is a probability that even after mutation the sequence will be recognizable. After the mutagenesis round only recognizable sequences will be retained by selection in the population, keeping the ensemble of patterns unchanged as a whole.

Next I will define terms which will be frequently used in further discussion. Some of them were already mentioned. Here I use the term "site" (or "functional site") to define a specific site in a genome (usually bearing some specific function), "a (typical) set" means a set of functionally acceptable sequences for a site which keeps its functional performance (a phenotype) within acceptable limits, and "a pattern" (or "sequence pattern") brings a set together with its equilibrium frequencies because some sequences in a set might be more frequent than others.

I am not interested here in how selection chooses individual sequences or what mechanisms govern different schemes of binding factors functioning. For my purposes it is sufficient to know the final result—the shape of patterns and corresponding sets in the "homeostasis" (equilibrium state). Binding factors usually support general-purpose mechanisms of gene expression regulation (i.e. the same binding factor can facilitate regulation of many different genes). However, it is clear that sets of acceptable sequences and/or equilibrium distribution of sequences within the set for gene-specific binding sites of the same binding factor can differ significantly, depending on individual regulation requirements. Therefore corresponding patterns, although they are used by the same binding factor, should be considered different. However, currently in most real population the time since the last common ancestor of a site is insufficient to obtain site-specific patterns directly. Because of this, common computational methods involving GI formalism usually ignore discrepancies in site-specific patterns.

If the only purpose is reliable storage of hereditary information then the mutation rate can be simply pushed to a minimum. On the other hand, zero mutation rates will stop evolution as a species will not be able to resist the challenges of a changing environment and become extremely vulnerable on a geological time scale. So apparently the balance between information maintenance and evolvability is required. It is reasonable to suggest that the change of total genomic information is usually a very slow process and can be observed only on geologically large time scales (e.g. the human and chimpanzee genomes are ~ 99% identical). On the other hand, it is likely that maintenance evolution acts much more actively; its footprints are distinguishable on very short time scales and constitutes the majority of observed features in a genome sequence. This work therefore will be mainly devoted to the phenomenon of maintenance evolution. Once the maintenance mode is clarified transition to modeling of progressive evolution becomes relatively apparent.

For clarity and simplicity I will consider asexual population in equilibrium (i.e. a population that evolved a sufficiently long time without any disruptive events and progressive evolution), a constant population size and a genome with equal proportions of all four nucleotides. I also assume that positions of the pattern are independent, i.e. there are no correlations between different positions of the pattern. Under this assumption a sequence pattern can be viewed as a fuzzy sequence, where each position represents not a single nucleotide but a probability distribution of nucleotides (i.e. a four-component vector, with components equal to nucleotide probabilities (frequencies) in the position), derived from the underlying set of functional sequences. So a nucleotide sequence can be viewed as a pattern where at each position a single nucleotide has a frequency of 1 and all other have zero frequencies. Independence of pattern position could be a rather strong simplification and, of course, more sophisticated "encoding" schemes can be evaluated. However, at this stage, without loss of generality I prefer to keep things simple, because the main predictions and conclusions of the model are sufficiently interesting for the suggested simple encoding scheme. Another important assumption concerns the process of mutagenesis. In this work I will consider only single base substitutions, ignoring the role of indels and different mechanisms of genomic rearrangements. Epigenetic and ploidy effects as well as recombination and variability or evolution of recognizers are also out of scope of the thesis. So the concise IT "engineering" problem as defined above is considered. However, all mentioned phenomena can be added as interesting extensions to the model without interfering with general conclusions drawn from the basic model.

In accordance with the notion of a sequence pattern as presented above, the term "single position site" or simply "position" (P) will be used bearing in mind not a single specific nucleotide, but a four-component vector (f_A, f_G, f_C, f_T), where each of $f_N, N \in \{A, G, C, T\}$ is a frequency of the corresponding nucleotide in a given position of "population", as shown in Figure 5. "Population" denotes here a set of sufficiently diverged functional sites. It is important to note that in a validation experiment presented below, the sequences were actually taken from a single genome. However, in the idealized conditions of an engineering model of infinite asexual equilibrium population, the "population" can be taken literally. Moreover, in this model it is also possible to define sites' GIs precisely, without making the simplifying assumption "one binding factor to one pattern". However, in real populations this assumption is unavoidable due to insufficiency of data. In order to investigate sequence patterns in actual genomes (which is, however, not a goal of this work) one has to assume that position-specific patterns are sufficiently similar, i.e. properties of corresponding binding sites are sufficiently uniform. So the visualization of the pattern presented in Figure 5 represents essentially the average of exon-specific patterns (which are unobservable individually). Of course some exon-specific differences are expected in the set of human acceptor splice sites used to produce a pattern for Figure 5. However, Figure 5 is a vivid illustration for the definition of the sequence pattern.

If the composition of position does not affect the phenotype, selection ignores it. Such a position contains no information by definition (this definition is consistent with intuitive understanding of information). Due to random mutagenesis in an equilibrium infinite population, this position will be occupied by four nucleotides in equal proportions (i.e. the frequency of each of four nucleotides will be equal to 1/4). However, if the site is functional, its composition will be affected by selection and nucleotide frequencies will deviate from uniform distribution. The variability of a single position site (P) can be naturally quantified by the informational entropy ($H(P)$). Using formula (3) and the notation described above, it takes form of:

$$H(P) = -\sum_{N\in\{A,G,C,T\}} f_N \log_2 f_N \tag{10}$$

Non-functional positions with frequencies of 1/4 have, according to formula (10), the maximum variability of 2 bits. For a fully preserved site with a single acceptable nucleotide (i.e. one frequency equals 1, three are zero) the variability is zero. To obtain the measure of positional genetic information ($GI(P)$) we have to take the reciprocal value:

$$GI(P) = 2 - H(P) \tag{11}$$

So it is easy to see that for a fully conserved site, GI takes the maximum of 2 bits. For non-functional sites GI is equal to zero, while intermediate values quantify the degree of conservation. Hence GI determines the biological value of the position.

GI does not depend on permutations of components in the vector of nucleotide frequencies, and each value of GI, except for the degenerate cases of GI = 0 bit and GI = 2 bits, can be obtained with infinitely many variants of nucleotide frequency vectors.

Presented by formula (11), the definition of genetic information was proposed by Schneider et al. (1986) more than 25 years ago and soon became a standard tool for visualization of sequence pattern composition called "sequence logo" (Schneider and

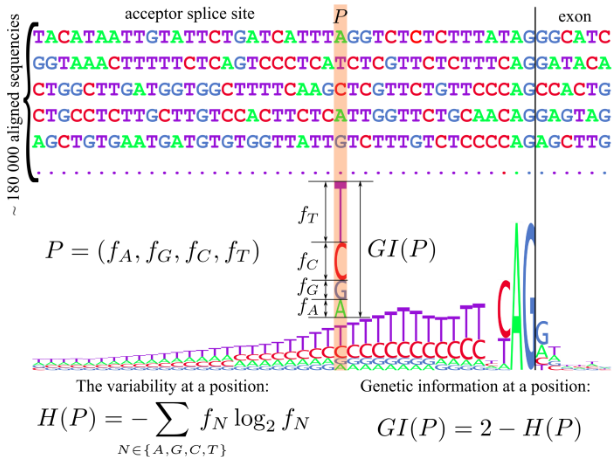
acceptor splice site
P
exon
~ 180 000 aligned sequencies
TACATAATTGTATTCTGATCATTTAGGTCTCTCTTTATAGGGCATC
GGTAAACTTTTTCTCAGTCCCTCATCTCGTTCTCTTTCAGGATACA
CTGGCTTGATGGTGGCTTTTCAAGCTCGTTCTGTTCCCAGCCACTG
CTGCCTCTTGCTTGTCCACTTCTCATTGGTTCTGCAACAGGAGTAG
AGCTGTGAATGATGTGTGGTTATTGTCTTTGTCTCCCCAGAGCTTG
$P = (f_A, f_G, f_C, f_T)$
f_T
f_C
f_G
f_A
$GI(P)$
The variability at a position:
$H(P) = -\sum_{N\in\{A,G,C,T\}} f_N \log_2 f_N$
Genetic information at a position:
$GI(P) = 2 - H(P)$

Figure 5. Illustration of sequence pattern and sequence logo; definition of genetic information.

Stephens 1990). The sequence logo of human acceptor splice site is shown in Figure 5. Acceptor splice site along with donor splice sites on pre-mRNA molecule are recognized by spliceosome—a protein complex, performing splicing of introns.

2.1.3 Typical sets and positional information

It was proposed (Stephens and Schneider 1992) and supported by simulations (Schneider 2000) that GI is additive and interpretable as localization (positional) information $GI_{binding} = \sum_i GI(P_i)$, i.e., the sum of GIs of individual positions in a binding site is equal to the information necessary to locate it in corresponding sequence context. It can then be hypothesized that apart from indicating the degree of conservation, $GI_{binding}$ has additional interpretations. Obviously the hypothesis is interesting and biologically important. However, its validation is not a trivial task because although both $GI_{binding}$ and localization information are measured in bits, their definitions are different and have no direct relationship. It is clear that the proof of this can significantly facilitate sensible GI-modeling applications.

In terms of IT, an abstract binding site can be described as a “source” which “generates” particular sequences (its realizations in a population, i.e. binding sites). In the framework of this schematic representation, $GI_{binding}$ and localization information can both be related by means of asymptotic equipartition property (AEP) (Girardin 2005). AEP applied to the described scheme effectively states that the majority of realizations of a binding site of length L mostly fall into a “typical set” and have similar values of $GI_{binding}$ (Cover and Thomas 2006). This means that while for non-degenerate GIs any sequence (out of possible 4^{L}) can be an outcome, the sequences actually observed, with high probability (tending to 1 when L tends to infinity) belong to the typical set, which has about $2^{2L-GI_{binding}}$ members distributed with approximately equal probabilities. The exponent value $2L - GI_{binding}$ reflects the entropy of a “source” or variability of a binding site. It is required $\log_2 N$ bits of information to select a single position site from a sequence of length N, which can be interpreted as a minimum number of yes/no questions required for the task. It is intuitively clear that a less specific search requires less information, e.g. selection of any item belonging to a set N_{set} requires $\log_2 N - \log_2 N_{set}$ bits. If we now return again to our scheme and recall that a binding factor defines a corresponding typical set of binding sites, recognizing sequences belonging to it and ignoring all others, it is then easy to see that corresponding localization information is equal to $GI_{binding}$.

This result provides a natural connection between continuous transversal (vertical) variability (i.e. across population, orthogonal to multiple sequences alignment) and discrete longitudinal (horizontal) localization on the sequence. It is also worth noting that due to its definition typical set of nucleotide sequences of length L represents a well-connected set (region) in the space of all possible L-nucleotide sequences with hamming distance. So a lot of mutations in typical sequences (i.e. sequences from typical set) will not throw them out of the typical set. Hence the structure of content of a typical set might provide a biological error protection mechanism, because if after mutation a sequence is still in the typical set, then this mutation is effectively “synonymous”. The “additivity” of GI, to our knowledge, was always used as an ad-hoc conjecture without a strict proof, since it is not possible to prove it without AEP. Here I want to focus the attention upon the “additivity”. It should not be confused with a simple additivity of information entropy, which obviously follows from formula (7) if we assume independency of positions in the pattern. This additivity concerns the statement that the sum of GIs for a functional site (or a whole genome) is linearly linked to the positional information (specificity of molecular interactions). Additivity in this sense is a critical property for whole-genome modeling of information dynamics and the form of function for measuring GI plays here very important role. One could use some other measure of frequencies biases (biases from the uniform distribution)—why is the defined GI function, based on information entropy “fundamental”? If we want the sum of GIs for the site to have informational meaning, the number of allowable functional sequences for the site (the size of its “typical set” in the context of above discussion) must depend exponentially on the defined site's variability (the value reciprocal to the sum of GIs). This exponential dependence is the nontrivial result of IT (AEP). The corresponding “natural choice” of the logarithmic function for information measure was discussed briefly in the introduction section and a detailed discussion can be found in the classical Shannon's paper (Shannon 1948). Having such a well-defined measure of positional information it is possible to build a formal (“mechanistic”) model of “molecular machines” evolution. The concept of sequence “typicality” (as an object for selection force) also may prove useful as it represents the collective property of a binding site, naturally accounting for a single position’s cumulative effects, as opposed to modeling the interaction of a large number of separate selection coefficients for each allele in each position. In the context of typicality the same mutation can either make a site more typical or less typical, depending on other positions of the site. Hence, selective values of the same mutation can be either positive or negative, depending on the background.

According to these ideas the conventional notion of “beneficial/positive mutation” (as well as reciprocal notion of “deleterious/negative mutation”) should be reformulated. Conventionally, a “beneficial mutation” is defined as a mutation that increases fitness (i.e. survival and reproductive success) of the carrier organism. I.e. conventional definition considers only contemporary effect of the mutation, on the other hand, the concepts of pattern and typicality allows us to naturally estimate the far-ranging effects of mutation (i.e. whether mutation increases or reduces the typicality of the carrier). In this work, therefore, all mutations that increase typicality of the sequence for a given pattern are considered beneficial and vice versa, while all mutations decreasing typicality are deleterious. In other words, beneficial mutation moves the sequence closer to the center of the functionally allowable set (where all sequences probably have insignificant differences in fitness and thus can be considered equal) providing more flexibility (more available mutations without loss of functionality) in the next generation. In the framework of classical concepts, such mutations are considered neutral, because they do not directly affect an organism's fitness. However, these variations increase tolerance to mutations in the next generation, making genetic information more robust in the course of maintenance evolution. The latter is assumed to be a ubiquitous process, occurring permanently in all loci of genome (in contrast with progressive evolution). Thus it seems reasonable to account these mutations to be beneficial.

2.1.4 Principle of conservation of sequence pattern

According to the conventional definition "population genetics is the study of allele frequency distribution and change" (Postlethwait 2009). If we consider a site with different alleles in the population, in the framework of classical models, the frequencies of these alleles will be not stable. If different alleles have different fitness, then such sites will evolve through fixation and carrying only the most advantageous allele, while all other (deleterious) alleles will be purged from the population by selection. If some alleles have equal or indistinguishably different fitness (e.g. if the site is nonfunctional, all alleles have the same fitness) their fate will be subject to genetic drift, due to which each allele can be either fixated or lost simply by chance. All these scenarios are transient. In spite of the "homeostasis" notion being ubiquitous in living systems, there are no valid terms or concepts for the description of evolution of weakly conserved sites in population genetics (e.g. the "tail" of the pattern in Figure 5), where what matters is the stable bias of frequencies rather than a neutrality, fixation or loss (which are the limiting cases for the model presented here, where GI is 0 or 2 bits correspondingly).

For brevity here I will not consider fitness of particular alleles altogether; the only thing I am formally concerned with is GI value. Traditionally the selective value for an allele is assigned somewhat ad hoc and its destiny in a population is traced using some mathematical model. In contrast with this conventional atomistic approach in this work more general and abstract concept of sequence pattern is suggested. Once a pattern (or a whole-genome pattern set) is defined, the value of any mutation is also defined (through its contribution to sequence typicality). Hence the pattern can be considered the lowest level phenotype, because higher level phenotypes are mechanistically derived from it. Having the pattern there is an opportunity to model whole-genome phenotype conservation without explicitly defining a high-level phenotype.

Classical models usually consider sites with only two different alleles since, due to common observations, the vast majority of observed variations (e.g. SNPs in the population) have two states because insufficient time has passed since the last common ancestor. However, here I suggest considering an idealized situation where time goes to infinity in a stable population without progressive evolution and other disruptive events. Understanding of underlying equilibrium dynamics in turn will pave the way for exploration of the evolution of variability “snapshots” created by recurring population bottlenecks, which are conspicuous but largely secondary companion processes.

The conservation laws of energy or momentum are the cornerstones of physical sciences. Here I suggest the law of conservation of sequence pattern in population genetics. It can be formulated as follows: a position with any intermediate value of GI can be at equilibrium, maintaining constant GI and nucleotide frequencies (hence the pattern and positional information of the corresponding binding site). So-called “balancing selection”, where the frequencies may be stable due to heterozygote advantage (Charlesworth 2006, Levene 1953) is apparently different from our generalization (the possibly interesting ploidy effects are not explored here for brevity).

2.1.5 Progressive and maintenance evolution

Progressive and maintenance evolution usually occur in the population simultaneously. Both represent important factors which affect allelic composition of the population. However, the effect of these phenomena is different and can be described in the framework of the model presented as follows: while progressive evolution reshapes the sequence pattern by changing the underlying set of functional sequences, maintenance evolution preserves the set of acceptable functional sequences, allowing only fluctuations inside this set. Thus the pattern remains unchanged. The model discussed focuses on characterizing a population in the static mode of evolution, which is essentially the maintenance evolution. Interestingly, a similar situation is classically described by the Hardy-Weinberg principle (Hardy 2003): constant allele frequencies in the absence of mutations and selection. However, we generalize this condition to functional sites explicitly including mutations and selection. The motivation behind the static mode of evolution is that the information already accumulated in a genome requires maintenance to prevent mutational degradation. When the maintenance is clarified, the next step can be exploration of how information in genome is redistributed or how novel information is added—progressive evolution. However, it is reasonable to suggest that increments of information are small in comparison with "what's already there", thus the majority of accumulated mutations (in functional sites) reflects the maintenance. Traditional models are often based on historical observational biases. Reasons for this are usually quite simple: for instance, it is easier to observe and study Mendelian traits compared to low penetrance effects (Houlston 2006). Other examples are the dramatic "selective sweeps" and "bottlenecks", which we believe are spectacular but special evolutionary events, scrambling the mundane maintenance phenomena as populations variability collapses. However, these events per se contribute negligibly to the bulk of genetic information, which is in the maintenance mode. Due to these collapses the observed variability of a site in a real population is typically much smaller than "potential" or "acceptable" variability which should be used to define the corresponding GI. Thus the actual sequence pattern for a specific splice site is usually unavailable (because the underlying sequence set is not available). However, this pattern exists in a platonic sense and if a site had a chance to diverge in a population without disturbances for sufficiently long time, functionally acceptable regions of the sequence space would be explored and actual patterns could be revealed, provided enough data.

Preemptively, it is possible to argue that mutational expansion into potential variability of functionally acceptable sites is commonly perceived as a "neutral evolution". However, in fact it is a "maintenance evolution" where observed deleterious (for GI value) mutations are compensated by approximately equal amount of beneficial mutations, preserving the sequence pattern. Thus, maintenance evolution suggests that while each individual mutation in general is not neutral, the cumulative effect of a large number of mutations can be considered neutral. The role of beneficial mutations is usually overlooked in classical models, as common wisdom dictates that they are rare, so that all the maintenance is carried out by purifying selection which is a special case in our model when GI is close to 2 bits.

2.1.6 Pattern conservation in the framework of the quasispecies model

As described in the introduction, an equilibrium population in the quasispecies model represents a “cloud” of organisms (sequences) with well-defined frequencies of alleles. Assuming for simplicity as before, that all sequences have the same length, by aligning them under one another we obtain a sequence pattern and corresponding GI profile. As quasispecies equilibrium is given by the steady state of the ODE system, equilibrium frequencies of alleles will hold static once the equilibrium state is reached. This, in turn, apparently leads to the conservation of the sequence pattern and GI profile, which is postulated in the model presented here. However, one should not be confused with the passing resemblance between these two models. The original model of quasispecies operates with infinite populations and obtains conservation of the sequence pattern as a corollary of stable alleles’ compound of the population at the equilibrium. In contrast, here the sequence pattern is implied as an inherent fundamental property of any functional site (or the whole genome, because the latter can be viewed as a composite functional site) which is predefined. It is suggested that the conservation of the pattern emerges naturally from the necessity of maintaining genetic information content, which ensures the viability of the population. Also I lay emphasis on the positional character of genetic information, stemming from the nature of molecular interactions. Besides, although the concept of pattern conservation is formulated in terms of infinite equilibrium population, I suppose that the phenomena of conservation might be also observed in finite natural populations (or their subsets) assuming that they are not extremely small and were not subjected to extensive positive selection, recent bottlenecks or other external influences abruptly distorting allelic composition of the population. Due to the finite sizes of real populations (leading to stochastic effects) and inevitability of disturbing impacts I expect that precise estimation of “ideal” sequences patterns will be impossible and the observed picture will represent a somewhat distorted reflection of the real underlying sequence pattern. However, for a population of rather large size that is stably evolving without drastic external pressure, the deviation of the observed sequence pattern from its “ideal” prototype will diminish over time.

2.1.7 Conservation of genetic information in splice sites

The constancy of sequence pattern postulated above can be exemplified by the divergence of splice site patterns between human and mouse—the difference between mouse and human splice site logos is quite small despite the large number of mutational and selective events that have happened since our divergence (Figure 6).

The maximum divergence of GI (less than 0.08 bit) can be observed in the fifth position of donor splice site. Notably, the number of splice sites is in the hundreds of thousands; the example in Figure 2 shows the phenomenon of constant pattern for the total of millions of nucleotide positions for a period of tens of millions of years. Obviously such an invariant, in such a fluid matter as genomic sequence, deserves close attention. As the average length of human exons is ~100 nucleotides, splice sites constitute a significant amount of the genomic sequence compared to coding sequences. Furthermore, splice sites are only one example of low-conservative functional noncoding sequences. So it is natural to assume that this mode of evolution affects a significant fraction of a genome besides splice sites. However, splice sites provide a unique opportunity for analysis because of their large number and well-defined locations in a genome. Other commonly known binding sites tend to be of sufficient length and high conservation (computational methods) and/or high binding affinity and specificity (experimental methods), creating observational biases with the preference for long sites with high GI per nucleotide.

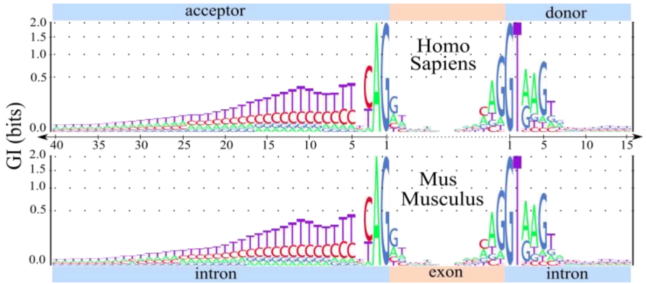

Figure 6. Comparison of splice sites sequence logos of Homo Sapiens and Mus Musculus.

2.1.8 Mutational load and its optimization

Optimality principles such as Maupertuis’ or Hamilton's and their different applications represent the foundations of physics, but they are applied moderately in life sciences. However, it is apparent that the drive to optimality is central in biological systems as well. All other things being equal, species with better energy efficiency have more available resources.

Examples of optimization are well-known in molecular genetics. E.g. the assignment of codons in the genetic code provides robustness due to the fact that a lot of mutations are synonymous (i.e. do not change the corresponding amino acid). It seems reasonable to suppose that before eventual stabilization, the genetic code was a subject to optimization by reassignment of codons in order to achieve better resistance to mutations. We can expect that tuning of noncoding functional elements, due to their high redundancy, is much more flexible. Thus many likely are (or recently were) affected by some kind optimization. However, in contrast with protein coding sequences, where the structure of information is known and provides clear understanding for potential mechanisms of optimization (which can be easily seen in the assignment of codons, as discussed above), the structure of information in noncoding functional elements is not so straight. However, our model suggests a universal approach for the representation of genetic information. On the basis of this representation we are able to propose a direct testable criterion, based on the notion of genetic load, according to which noncoding information can be optimized. However, it is not a unique option for optimization; other possibilities exists that are not considered in this work.

One classical definition states: “Genetic load is the reduction in selective value for a population compared to what the population would have if all individuals had the most favored genotype” (Crow 1958). If considered in the context of the proposed model, this definition describes a site with $GI = 2$ bits. However, if a site has a GI value lower than 2 bits, the load arising due to its maintenance should be defined in another way.

It is common practice to model equilibrium states through their stability to perturbations, i.e. if an equilibrium state is externally disturbed it should be restored by some stabilizing force. In the discussed case, the perturbations are random mutations, and counteracting stabilizing force is a (purifying) selection that tries to compensate for deviations from equilibrium. Therefore there is a straightforward way to model the maintenance of a pattern. Consider a single position site and assume that initially it is in equilibrium and has nucleotide frequencies (f_A, f_G, f_C, f_T), $\sum_{N\in\{A,G,C,T\}} f_N = 1$. Then mutagenesis pushes them into (f_a, f_g, f_c, f_t), after which these frequencies are corrected by reproduction and selection, returning nucleotides frequencies back to the initial values and preserving the initial value of GI:

$$\begin{pmatrix} f_A \\ f_G \\ f_C \\ f_T \end{pmatrix} \underset{\xleftarrow[selection]{}}{\xrightarrow{mutation}} \begin{pmatrix} f_a \\ f_g \\ f_c \\ f_t \end{pmatrix} \tag{12}$$

The changes in frequencies are assumed to be small.

Mutations can be divided into two different types: transitions change a purine to another purine or a pyrimidine to another pyrimidine: $ti = \{A \leftrightarrow G, C \leftrightarrow T\}$ and transversions represent all other mutations: $tv = \{AG \leftrightarrow CT\}$. Here, for brevity, I assume that all 4 transitions as well as all 8 transversions are equiprobable. The system for descendant nucleotide frequencies (f_a, f_g, f_c, f_t) can be written using the formula of total probability as follows:

$$\begin{cases} f_a = (1-p)f_A + p\left[kf_G + \dfrac{(1-k)}{2}(f_C + f_T)\right] \\ f_g = (1-p)f_G + p\left[kf_A + \dfrac{(1-k)}{2}(f_C + f_T)\right] \\ f_c = (1-p)f_C + p\left[kf_T + \dfrac{(1-k)}{2}(f_A + f_G)\right] \\ f_t = (1-p)f_T + p\left[kf_C + \dfrac{(1-k)}{2}(f_A + f_G)\right] \end{cases} \tag{13}$$

where p stands for probability of mutation and k is a probability of transition, upon condition that a mutation occurred. For mammals $k \approx 2/3$ (Collins and Jukes 1994) corresponding to the ratio of transversions to transitions $ti/tv = 1/2$. It is easy to see that for descendant nucleotide frequencies the property $\sum_{n\in\{a,g,c,t\}} f_n = 1$ also holds true. Hence the deviation of nucleotide frequencies from the equilibrium due to mutagenesis is:

$$\begin{cases} \Delta f_A = f_A - f_a = p\left[f_A - kf_G - \dfrac{(1-k)}{2}(f_C + f_T)\right] \\ \Delta f_G = f_G - f_g = p\left[f_G - kf_A - \dfrac{(1-k)}{2}(f_C + f_T)\right] \\ \Delta f_C = f_C - f_c = p\left[f_C - kf_T - \dfrac{(1-k)}{2}(f_A + f_G)\right] \\ \Delta f_T = f_T - f_t = p\left[f_T - kf_C - \dfrac{(1-k)}{2}(f_A + f_G)\right] \end{cases} \tag{14}$$

From equation (13) it is clear that mutagenesis will always push frequencies of nucleotides to uniform distribution, hence GI of the descendant frequencies vector is always less or equal to initial GI (equality happens only if initial GI = 0 or p = 0). As the target for potential optimization (among many alternatives) I suggest a variant of mutational load (ML), which is defined as Manhattan norm of frequencies deviation vector:

$$ML = |\Delta f_A| + |\Delta f_G| + |\Delta f_C| + |\Delta f_T| \tag{15}$$

Minimizing mutational load in the form as defined in equation (15) will minimize the number of mutations rejected by selection, thus minimizing the rate of “genetic deaths”, making this measure of ML biologically plausible. From the expression for optimal solution, presented in equation (17), it is easy to see that in this case: $-\Delta f_A =$

$\sum_{N\in\{G,C,T\}}\Delta f_N$, and $\Delta f_N > 0, \forall N \in \{G,C,T\}$, assuming A to be the highest frequency allele in the position. So after mutagenesis the selection can move frequencies back to the initial simply by removing variants whose frequencies have increased (i.e. G, C and T). For optimal frequencies, the number of individuals which must go extinct is proportional to the ML is defined by equation (15), and equal to $2\Delta f_A$.

In this work I consider the equilibrium state of a population, so in order not to overcomplicate matters, population size will be kept constant. In contrast with classical models, the size of a population does not really matter for maintenance evolution of GI. Population size matters for non-equilibrium phenomena such as selective sweeps, caused by spontaneous appearance and consequent fixation of a site with GI = 2 bits. Such events are out of scope of this model.

If the mutagenesis is biased ($k \neq 1/3$), different compositions (e.g. permutation of nucleotides) of nucleotide frequencies 4-vector with the same GI can produce different ML. A simple example demonstrating this fact is shown in Figure 7. When two major nucleotides (Figure 7 left) are connected by transition, the impact of mutagenesis is largely compensated as the most probable mutations are counteracting transitions. In the right half of Figure 7 the opposite effect can be seen—non-compensated composition—where the major nucleotides “leak” strongly into the minor ones, hence causing larger ML.

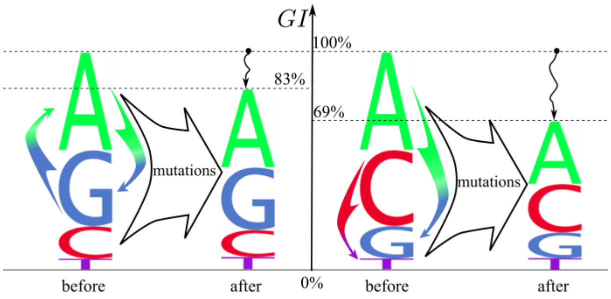

Figure 7. Variation of nucleotide frequencies and reduction of GI due to mutagenesis with transitions prevalence in positions with different nucleotide frequency vectors. The two most frequent mutations in a position are marked with colored arrows.

The task of finding nucleotide frequencies providing the lowest value to ML for a given GI can be formulated as the following optimization problem:

$$\begin{cases} ML(GI) \xrightarrow[f_A, f_G, f_C, f_T]{} min \\ \sum_{N\in\{A,G,C,T\}} f_N = 1 \\ 2 + \sum_{N\in\{A,G,C,T\}} f_N \log_2 f_N = GI = const \end{cases} \quad (16)$$

ML(GI) is a mutational load, which has to be minimized by adjusting the 4-vector of nucleotide frequencies for a given value of GI. It is easy to see that solution of problem (16) does not depend on the probability of mutation *p*. The solution was found numerically, using a homemade implementation of the evolutionary algorithm presented in (Runarsson and Yao 2000). However, its analytical representation, which in general case can be written in a parametric form, was also obtained and shown in formula (17).

$$\begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \end{pmatrix}_{opt} = \begin{cases} \begin{cases} f_2 = \frac{1}{4} + \left(f_1 - \frac{1}{4}\right)\frac{3k-1}{3-k} \\ f_3 = \frac{1}{2} - f_2 \\ f_4 = \frac{1}{2} - f_1 \end{cases} & if\ f_1 \in \left[\frac{1}{4}, \frac{1}{2}\right) \\ \begin{cases} f_2 = \frac{1}{4} + \left(f_1 - \frac{1}{4}\right)\frac{3k-1}{3-k} \\ f_3 = 1 - f_1 - f_2 \\ f_4 = 0 \end{cases} & if\ f_1 \in \left[\frac{1}{2}, \frac{1}{k+1}\right) \\ \begin{cases} f_2 = 1 - f_1 \\ f_3 = 0 \\ f_4 = 0 \end{cases} & if\ f_1 \in \left[\frac{1}{k+1}, 1\right] \\ GI = 2 + \sum_{i=1}^{4} f_i \log_2 f_i \end{cases} \quad (17)$$

where f_1 is the highest frequency, f_2 is the frequency connected to the f_1 by transition, f_3 is maximum of transversions to f_1, f_4 is transition to f_3, *k* is the probability of transition, upon condition that the mutation has occurred.

The vector of optimal nucleotide frequencies vs. GI, representing the solution of optimization problem (16), is demonstrated in Figure 8 A. It is noteworthy that assignment of certain nucleotides to the vector of optimal frequencies allows some permutations remaining in its optimality. In the presence of mutational bias ($k \neq 1/3$), top and bottom pairs of frequencies must be occupied by nucleotides connected through transition. In the case *k* = 1/3 (no mutational bias) all four frequencies are permutable with each other.

The solution shows a phenomenon resembling phase transitions: derivative discontinuities near 0.5 and 1 bits, with corresponding changes in the number of “degrees of freedom” and permutation symmetries. Phase transitions are generally assumed to be highly non-analytic as they stem from non-linear interactions of large numbers of objects. In this model the interactions in a population are effectively

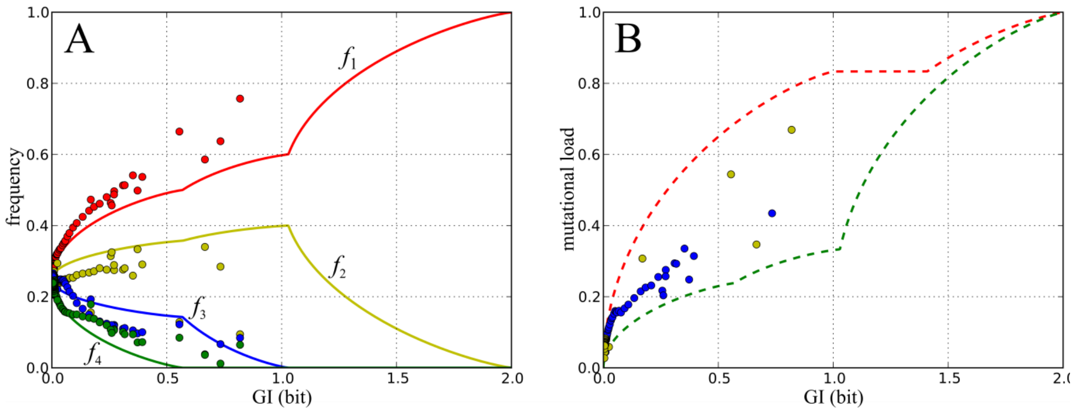

Figure 8. Minimization of mutational load. Comparison of nucleotide frequencies and mutational load observed in human splice sites (donor and acceptor) with optimal ones.
A (left): Nucleotide frequencies minimizing mutational load (k = 2/3) vs. GI and nucleotide frequencies of human splice sites. The red line (f_1) is the highest frequency nucleotide. The yellow line (f_2) is the frequency of nucleotide, connected to the f_1 by transition. The blue line (f_3) is maximum frequency of the remaining nucleotides coupled through transition. The green line (f_4) transition to f_3. Circles represent frequencies of Homo sapiens donor and acceptor sites (each circle represents single position of the site).
B (right): Minimum and maximum mutational load (k = 2/3) vs. GI and mutational load of human splice sites. Mutational load is normalized so that its maximum value equals 1. The green dashed line is the mutational load of optimal nucleotide frequencies (minimum mutational load). The red dashed line is the maximum mutational load. Blue and yellow circles are the mutational load of human acceptor and donor sites respectively.

hidden and are described by the optimal selection outcome. Detailed investigation of this phenomenon can potentially give an interesting result.

2.1.9 Experimental evidences of mutational load optimization

After the theory is presented, it is quite tempting to see its evidence in real data. The conservation of sequence pattern was already demonstrated by the comparison of human and mouse splice sites' sequence logos (Figure 6). In this section, verification of the theory will be continued. Above it was shown that in the framework of the model it is possible to minimize mutational load by arranging nucleotide frequencies in a way that results in specific compositions of the pattern. So if the proposed model adequately describes the process of maintenance evolution of genetic information and an assumption that the majority of this information is in maintenance mode holds true, we can hypothesize that the traces of optimization, in a form proposed above (i.e. specific distribution of nucleotide frequencies in the positions of sequence pattern) can be revealed in real sets of related functional sequences. To test this hypothesis we need to have sufficient reliable statistics of the site under investigation. For this reason

I decided to opt for human splice sites (because they are abundant and readily available) and some transcription factor binding sites from the open access database, for which sufficiently large samples of sequence variants are available.

However, even for these specially selected sites, due to a variety of reasons it is quite unlikely that experimental data will match this particular optimization precisely because on the one hand other optimization parameters are possible (for instance a total site length to optimize the transcription speed), and on the other hand, the pattern (i.e. the logo) itself was derived with the simplified assumptions outlined earlier (e.g. ignoring exons-specific individual patterns differences). Moreover, due to specific regulatory demands it is natural to expect that no optimal compositions exist.

2.1.9.1 Splice sites

Chromosome sequences and locations of human exons were retrieved from Ensembl database (Flicek et al. 2012) using the BioMart data mining extension (Kasprzyk 2011). Then, donor and acceptor splice sites were extracted. In order to avoid the influence of minor spliceosome, which may have a significantly different sequence pattern (Tarn and Steitz 1996), only the sites which conform to so-called “GT-AG” rule were kept. As a result more than 180 thousand of each donor and acceptor splice site sequences were obtained. For further analysis I took 70 acceptor and 50 donor positions adjacent to exon. Corresponding nucleotide frequency vectors together with optimal frequencies providing a minimum to ML are presented in Figure 8 A. It can be seen that the trajectories of nucleotide frequencies in splice site positions are fairly consistent with the optimal. Although some fine features of its behavior (e.g. collapsing of bottom pair of frequencies) are not reproduced by the model, the basic predicted trend, namely that the top and bottom pairs of nucleotide frequencies are connected through transition, is observed in 85% of positions with GI content higher than 0.05 bit (a total of 26 positions, with the arrangement of nucleotide frequencies in 22 corresponding to the model’s prediction).

It is also possible to find nucleotide frequencies providing a maximum ML for a given GI. It was done using the same evolutionary algorithm as I used for finding a minimum of the ML. Figure 8 B demonstrates the ML of human splice sites in comparison with maximum and minimum mutational load. According to Figure 8 B the strongest optimization is possible for position with a value of GI close to 1 bit where maximum and minimum ML differ most strongly. Such positions in principle would be favorable for the storage of genetic information. I also found that the higher mutational bias (lower *tv/ti*) provides an opportunity for better optimization in this region (i.e. reduces minimum ML), hence organisms with higher mutational bias are

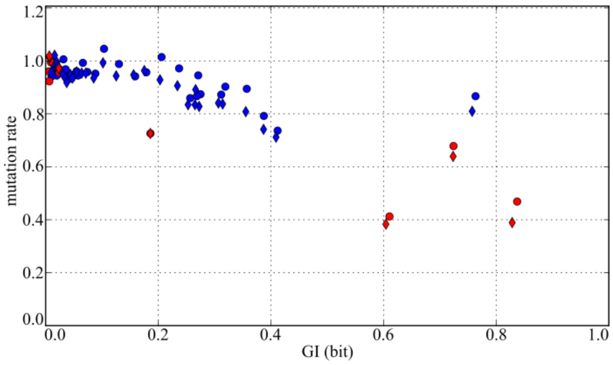

Figure 9. Normalized mutation rate of acceptor (blue) and donor (red) splice sites. Mutation rate was obtained from pairwise alignment of human vs. rhesus (diamonds) and human vs. chimp (circles) and then normalized to make mutation rate of positions with GI close to zero equal to 1.

able to achieve better optimization and we can expect that selection can somehow promote an increase of mutational bias.

Here I also want to note one interesting feature that I was able to observe in splice sites. Using pairwise alignments of human versus rhesus and chimpanzee, obtained from the UCSC Genome Browser database (Fujita et al. 2011), I compared the substitution rates for splice sites divergence between human and these two other primates (Figure 9).

I found that the conservation of the acceptor “tail” is quite weak: positions with GI < 0.4 bits have substitution rates higher than 80% of the neutral rate. However, the tail stores about 50% of positional information, i.e. approximately 5 bits as compared to ~10 bits of total acceptor information, 4 bits of which are provided by the “AG” site (Figure 5).

2.1.9.2 Transcription factor binding sites

Splice sites, as mentioned above, are rather unique genomic objects which may have special properties. It is therefore desirable to find evidences of mutational load optimization in other functional sites. For this purpose I selected 8 reliable transcription factor binding sites (TFBS) of vertebrates from the JASPAR open access database (Sandelin et al. 2004). In order to have enough statistics for building a reliable sequence pattern, only sites with average position coverage more than 1500 from JASPAR core database were selected. As a result 6 mouse sites (JASPAR IDs: MA0039.2, MA0035.2, MA0145.1, MA0141.1, MA0002.2, MA0140.1) and 2 human sites (IDs: MA0137.2, MA0138.2) were obtained. Their sequence logos are shown in Figure 10.

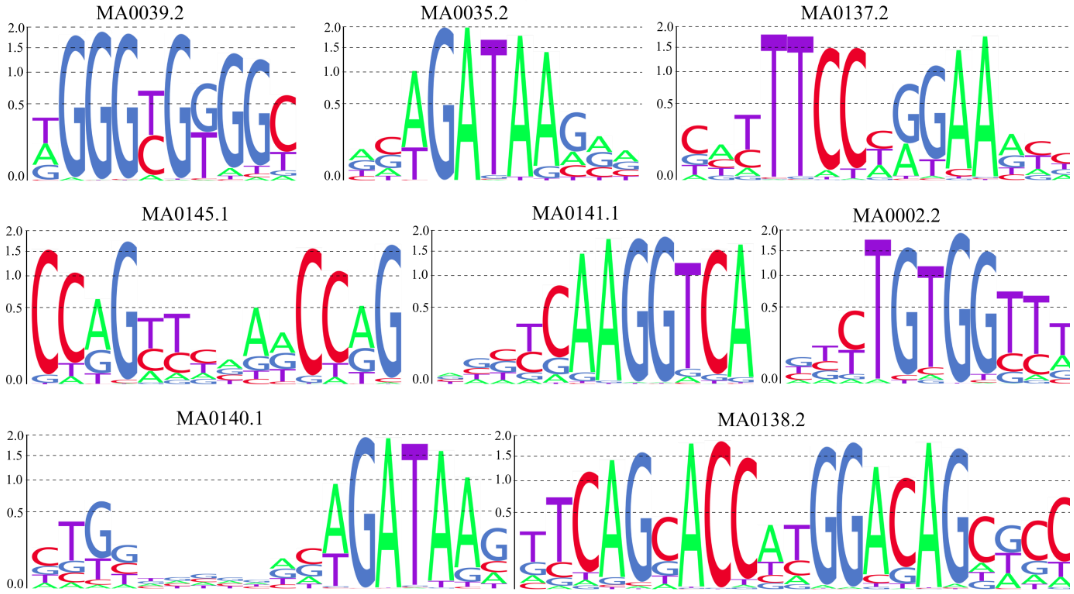

Figure 10. Sequence logos of 8 transcription factor binding sites. For each logo JASPAR ID of corresponding transcription factor binding site is indicated. Vertical axes represent log2-scaled GI measured in bits.

Vectors of nucleotide frequencies for the positions of these TFBS are shown in Figure 11 A. The optimization of nucleotide frequencies is not as clear as in the case of splice sites. This might be partially explained by observational biases towards highly conserved sites with underrepresented "tails" of patterns. The example of splice sites demonstrates that weakly conserved tails are well optimized while highly conserved positions are not.

Figure 11 A demonstrates rather a chaotic range of nucleotide frequencies. To suppress data noise and to reveal the obscured trend, I averaged nucleotide frequencies by kernel smoothing. Also in order to demonstrate non-random behavior of real sites, for each pattern of 8 selected TFBS, 10 000 random patterns were generated. Each position of random pattern has the same GI as the corresponding position of its real prototype, but its nucleotide frequencies were generated randomly. To get a single value of nucleotide frequency for random patterns in each position, nucleotide frequencies of corresponding nucleotide in corresponding positions were averaged over all 10 000 random patterns. The resulting (averaged) vectors of nucleotide frequencies of random patterns along with smoothed frequencies of sequence patterns of real sites are presented in Figure 11 B.

The most noticeable difference between random and real patterns in Figure 11 B is the mutual arrangement of the second and the third nucleotide frequencies. Real sites tend to pull up transition to nucleotide with the highest frequency (yellow), performing an optimization according to mutational load minimization. In contrast, random

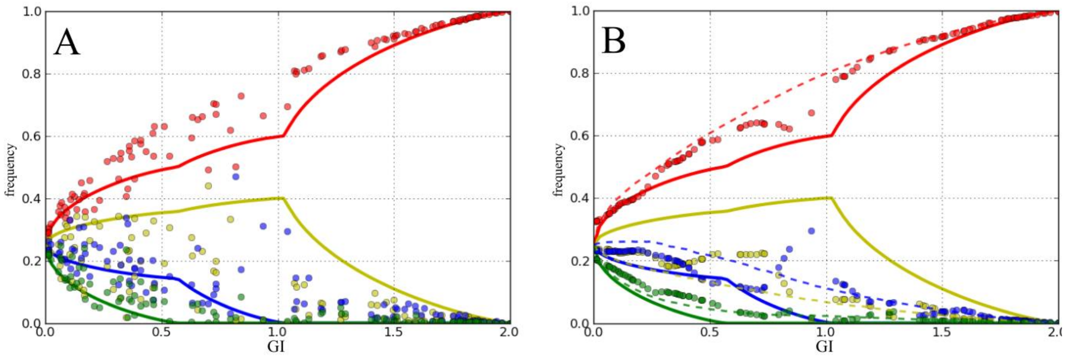

Figure 11. Optimization of nucleotide frequencies in TFBS.
A (left): Nucleotide frequencies vectors for positions of real TFBS patterns compared with optimal frequencies. Solid lines are optimal nucleotide frequencies. Circles are nucleotide frequencies of real TFBS patterns.
B (right): Averaged nucleotide frequencies of random generated patterns and real patterns of TFBS versus optimal frequencies. Solid lines are optimal nucleotide frequencies. Circles are smoothed frequencies of nucleotides in positions of real TFBS patterns. Dashed lines are averaged nucleotide frequencies in positions of random generated patterns.
Color scheme corresponds to one that was used in Figure 8.

generated patterns demonstrate behavior, where the second largest frequency nucleotide is transversion to one with the highest frequency. This could be easily explained from a mathematical point of view, as the probability of such outcome (second highest frequency nucleotide is transversion to the highest frequency nucleotide) is twice higher, because there are two transversions and only one transition to each nucleotide.

The tendency of TFBS towards minimization of mutational load becomes clearer if we directly compare the mutational load of real site patterns and random patterns, generated according to the procedure described above (Figure 12).

Here I also add human acceptor and donor sites to analysis. The sites under investigation have different lengths, therefore, in order to make possible a general comparison between them, the mutation load for each pattern was represented in terms of Z-score (distance from the mean in units of standard deviation).

Significant deviation from optimality observed in some sites can stem from the incompleteness of the data caused by the observational bias. However, it can be also an important signal, indicating that possibly some other mechanism of genetic information processing is involved in the evolution of the site. This site can be subject to ongoing progressive evolution. Another example could be so-called CpG sites. These sites are hyper-mutable and highly underrepresented in mammalian genomes. Thus they are costly in terms of maintenance, requiring either some special mechanisms of protection from mutations in functional regions or simple

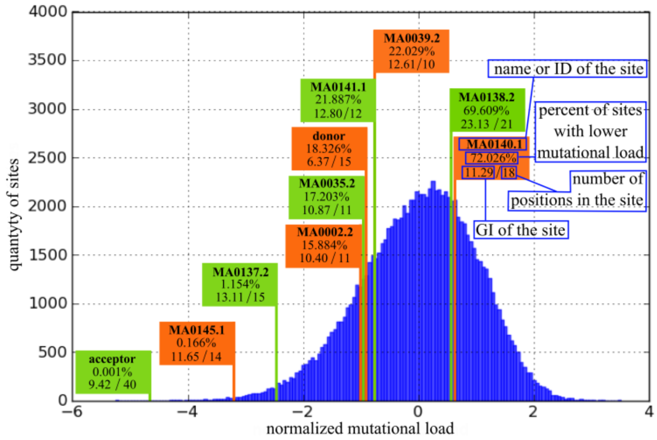

Figure 12. Normalized mutational load (Z-score, i.e. distance from the mean in units of standard deviation) of real binding site patterns (acceptor, donor and TFBS; for TFBS their JASPAR IDs are indicated) compared to distribution of 100 000 randomly generated patterns (10 000 for each real site). Z-scores of real patterns are indicated by colored flags with brief descriptions, which contain: the pattern's ID and the percentage of random patterns with lower mutational load, total GI and the length of corresponding site.

intensification of purifying selection. Also, as we observe CpG enrichment in functional promoters, we can suppose that they have additional informational value.

2.1.10 Simulation of single position evolution

In an attempt to get a deeper understanding of how the pattern is maintained and evolves in the framework of the proposed model, a simple simulation was constructed. The simulation describes the process of evolution of a single position site (i.e. individual organisms of our artificial population are simply single nucleotides). There are two processes counteracting each other in the course of simulation—mutation and selection. Mutation changes one nucleotide into another, with the probability of transition twice higher than probability of transversions (k = 2/3) and with equal probabilities of both transversions. The pattern of the site is not predefined, however, for each moment we define minimum level of GI, which must be maintained in the population (i.e. GI of single position site). According to simulation procedure, mutations, which reduce GI in a population, are called negative, and GI-rising mutations are positive. Selection in our simulation operates in a way similar to (back-) mutation. The act of selection replaces an individual with less frequent alleles by more frequent variants. Thus, after selection act the population GI increases. The most effective selection action replaces a nucleotide with the lowest frequency, by the nucleotide with the highest one. However, it is unlikely (similarly to Haldane's arguments) that selection on multiple sites in a genome has a chance to operate in the most efficient way for each of them. It is more realistic if selection increases GI in probabilistic manner—any increase of GI will do—not only through the most efficient replacements. It is noteworthy that neither mutation nor selection changes the size of population, thus it always remains constant.

Evolution of the population proceeds through stages simulating gradual increase of GI. All stages have a predefined minimum level of GI for the population (GImin), which increases from stage to stage. Each stage in turn is divided into two steps: slight increase of GImin level with consequent progressive evolution, during which frequencies of nucleotides adjust in order to reach the new level of GImin, and maintenance evolution when GImin is maintained in time. Initially, GImin is set equal to 0 and frequencies of all nucleotides are set to 0.25, (f_A, f_G, f_C, f_T) = (0.25, 0.25, 0.25, 0.25). During each stage the population is subjected to mutagenesis and selection: if after a mutation the level of GI falls below current GImin, selection starts to act. Acts of selection are repeated until the level of GI is restored to GImin. Thus, several acts of selection can occur after a single mutation. On the other hand, it is possible that selection is not required, because GI can rise due to positive mutations. As it was already noticed, most mutations lower the level of GI, and with increases of GImin, the fraction of such mutations increases as well, so selection has to generate more replacements. However, it is worth noting that the fraction of positive mutations remains considerably large until GI reaches the level of 1 bit. Figure 13 A demonstrates that approximately 25% of random mutations in a population with GI =

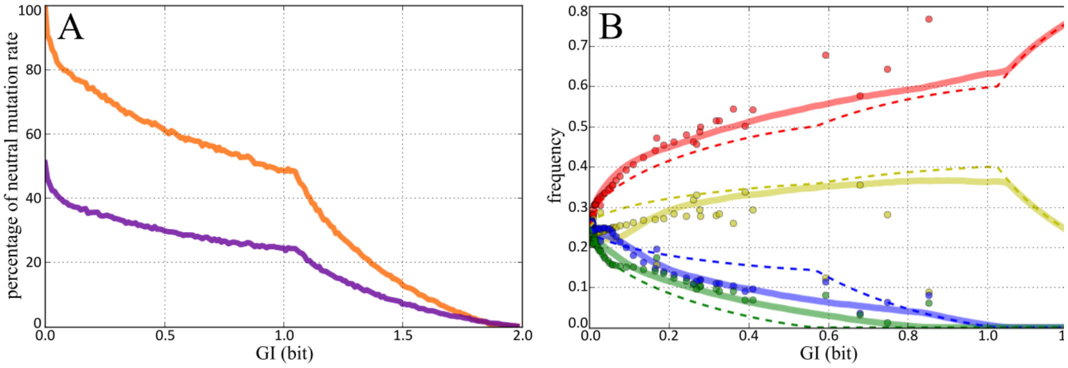

Figure 13. Simulation results.
A (left): All observed (not erased by selection) and positive mutations as a fraction of all mutations occurred in the simulation for a corresponding level of GImin.
The orange line represents the percentage of mutations which are observed in a population.
The purple line represents the percentage of GI -increasing mutations in a population.
B (right): Modeled nucleotide frequencies in comparison with real and optimal. Color denotes the same as in Figure 8.
Continuous transparent lines represent modeled nucleotide frequencies. Dashed lines represent optimal nucleotides frequencies. Circles represent frequencies of Homo sapiens donor and acceptor splice sites.

1 bit are positive.

The described simulation represents a simple way to model equilibrium and accumulation of genetic information in the population in the context of discussed pattern-centered concept. Its only principle is to maintain the required information content of the site. The trajectories of simulated nucleotide frequencies are presented in Figure 13 B. Their shape is very similar to the shapes of both frequencies observed in real splice sites and optimal nucleotide frequencies. Moreover, modeled frequencies even reproduce some subtle features of real splice sites' frequencies: the flipping of the second and the third frequencies near GI = 0 and collapse of the bottom pair of frequencies near GI = 0.1 bit. Hence, simulated trajectories of nucleotide frequencies in Figure 13 B demonstrate a somewhat surprising concordance between simple simulation and the perceivably complex behavior of millions of nucleotides involved in splice sites.

The simulation process is stochastic, so it is entirely possible that the observed picture occurred just by chance. In order to be sure that the behavior of the nucleotide frequencies is robust the same simulation was performed 20 times. Each time very similar trajectories of nucleotide frequencies were observed, so the result, presented in Figure 13, is highly-reproducible.

2.2 Drake's rule phenomenon

In the early 1990s John W. Drake (1991) examined the rate of spontaneous mutations in seven DNA-based haploid microbes, including two single stranded and two double stranded bacteriophages, a yeast, a bacterium and a filamentous fungus. Applying refined techniques for calculation of mutation rates he revealed a remarkable genomic regularity: the average mutation rate per base pair per replication is inversely proportional to the size of genome. Moreover, despite the significant differences between organisms, whose genome size varies 6500-fold and average mutation rates per base pair vary 16,000-fold, variation of their mutation rates per genome is only 2.5-fold with a mean value of 0.0033 per DNA replication, demonstrating a somewhat surprising invariance. Drake concluded that such constancy of genomic mutation rate indicates that this rate is highly evolved and "have been shaped in response to evolutionary forces of a very general nature, forces independent of kingdom and niche" (Drake 1991). Further he assumed that such regularity may be extrapolated over all microbial organisms. The pattern of mutation rate observed by Drake in microbes, i.e. the inverse relationship between mutation rate per nucleotide site per generation and total size of genome along with essential constancy of mutation rate per genome per generation has been called "Drake's rule" for its discoverer.

Initially Drake's conjectures faced reasonable skepticism because they were based on just seven species. However, during further investigations the set of species for which accurate estimates of the mutation rate is available was substantially extended, including not only DNA-based microbes but a broad variety of organisms from different taxa, from RNA and DNA viruses to plants and vertebrates (Lynch 2010a). These data provide strong support for "Drake's rule" with respect to viruses, bacteria and many of unicellular eukaryotes. However, there are some outliers among unicellular eukaryotes, which demonstrate the reverse trend, i.e. positive scaling between genome size and per base mutation rate. For instance, while documented microbes span four orders of magnitude in genome size, the range of their mutation rates per genome per replication (excluding above mentioned outliers) is remarkably narrow (less than ten folds) (Sniegowski and Raynes 2013). If instead of total size of genome the proteome size is regressed on the mutation rate however, then to a great extent the inverse relation between these two parameters also holds true for all documented microbes (including aforementioned outliers) (Massey 2013).

In striking contrast to the microbial pattern of mutation rate, multicellular eukaryotes demonstrate a strong positive relation between genome size and mutation rate per genome per generation (Sung et al. 2012). However, this last inconsistency can be at least partially explained by the drastic difference in the structure of genome between high eukaryotes and microbes. While the fraction of (conserved) protein coding sequence in microbial genomes usually exceeds 90%, in multicellular species its proportion is generally more than one order of magnitude less than that in the microbes and tends to decrease with the growth of total genome size (e.g. only ~1.5% of human genomic sequence encodes proteins). It should also be noted that estimated mutation rate per nucleotide site per generation for vertebrates (in particular for human) is approximately two orders of magnitude higher than for prokaryotes (Lynch 2010a). However, while for microbes the notions of "generation" and "replication" are synonymous, it is not so for higher eukaryotes. Thus for multicellular eukaryotes, and especially for vertebrates, the number of germ-line cell divisions can be large enough to play a substantial role. If this factor is taken into account, the mutation rate per cell replication in human germ-line is lower than that in any other species for which a reliable estimation of mutation rate is available (Lynch 2010b).

How mutations accumulate in genomes and what are the forces that shape genomic mutation rates are central questions of molecular evolution theories. Obtaining data to clarify this issue is laborious and technically challenging. Recent advances in highthroughput sequencing allowed rapid accumulation of data shedding light on mechanisms and rates of appearance of new mutations. However, the understanding of these processes is far from complete.

One potential and perhaps currently the most popular interpretation of observed mutation rate patterns, which is often used in particular to explain "Drake's rule", is given by the "Drift-barrier hypothesis" (Lynch 2010a; Sung et al. 2012). To put it in a nutshell the main idea of this hypothesis is that an observed pattern of mutation rates is a product of balance between two counteracting forces: natural selection and random genetic drift. More specifically, considering that the majority of mutations in functional sites are harmful, natural selection generally favors the alleles providing lower mutation rates. However, a reduction in the mutation rate can be achieved only by some physiological cost (e.g. slower replication speed, utilization of more complex schemes of preventing/correcting mutations etc., eventually leading to an increase in resources expenditure). Thus natural selection pushes mutation rate down, gradually increasing the fitness until any further advantages become smaller than the power of random drift. At this point selection is incapable of reducing the rate of mutations any further and the process freezes. The power of random genetic drift in turn is inversely proportional to the effective population size ($1/N_e$ and $1/(2N_e)$ for haploid and diploid organisms correspondingly, where N_e is the effective population size). So according to the "Drift-barrier hypothesis" species with higher N_e are expected to have lower rate of mutation. As a consequence, microbial populations that typically have large N_e, will demonstrate per a base mutation rate lower than multicellular eukaryotes (and especially vertebrates), whose effective population size is expected to be substantially lower.

Sung et al. (2012) regressed available estimations of per-base mutation rate in effective genome (which was estimated as a product of mutation rate/bp/generation and the size of protein-coding genome) on estimates of N_e (which, in turn, were obtained on the basis of average nucleotide heterozygosity estimated at silent sites). Consistency with predictions of the "Drift-barrier hypothesis" the regression demonstrates a strong negative relationship, largely independent of phylogenetic background (Figure 1 A in Sung et al. (2012)). However, Sung et al. acknowledge that estimation of N_e is "fraught with difficulties". This parameter cannot be estimated directly and its value can vary significantly, depending on estimations of other factors. Moreover, this study does not account for germ-line cell divisions, which, in my opinion, are important and if considered would substantially affect the observed picture. Although the "Drift-barrier hypothesis" suggests a possible explanation for observed patterns of mutation rate, the understanding of the mutation rate evolution and the phenomenon of "Drake's rule" is still far from being clear. However, it is beyond doubt that a clear conception of this process would shed the light on the evolutionary process.

Here I want to suggest an alternative interpretation of the "Drakes's rule" based on a formal model of storage of genetic information presented above. I will support a proposed hypothesis with numerical simulations, demonstrating that postulated by the "Drakes's rule" relation between mutation rate per genome per replication and genome size naturally emerges within the framework of the model of positional information storage. Random mutations deteriorate genomic information and must be compensated for by selection to maintain the total genomic information. Considering this process under informational equilibrium and suggesting that the genome of any species operates near its maximum informational storage capacity and the mutation rate is near its upper limit, a simple general explanation with minimal assumptions will be proposed for the "Drakes's rule". This part of results is based on (Shadrin and Parkhomchuk 2014).

2.2.1 Absolute and relative fitness

"Fitness" has been a key parameter in modeling Darwinian selection for almost a century. It determines which organisms are left to live and reproduce and which will be eliminated from the population. Generally, different alleles ("variants") are presumed to have different fitness. Mutations affect allele frequencies (density) in the population. The dynamics of the latter is traced with some mathematical model. Numerous models with different assumptions were proposed to simulate allele dynamics in real populations correctly. For instance the "Moran process" (Moran 1962) represents a model with "overlapping generations", defining an elementary time step as either death or reproduction of a random individual in the population, providing analytical solutions for this simple scenario. The "Wright-Fisher model" (Durrett 2008) represents an alternative, presuming non-overlapping generations, such as annual plants. There is no conventional approach of how the cumulative fitness for a few independent variants should be calculated, taking into account the effects of newly appearing variants, and many other subtleties. Traditional models usually consider the fitness as a relative value without any fixed baseline, so fitness value is distributed around the unit. Sometimes formulas analogous to what is introduced here are used for calculation of cumulative fitness for multiple alleles (Ofria et al. 2008; Strelioff et al. 2010; Frank 2012). The absolute value of the fitness in these cases is not interpreted so in fact it can also be normalized. For instance, in traditional approaches, there is no sense to take into account sites that are not variable in a population (i.e., sites where any variation is lethal) while calculating the fitness.

However, most of the genome usually is not variable in realistic modeling. In contrast for our formula it is essential to sum overall sites to obtain proper interpretation of fitness. Hence, for relative fitness, there is no fixed "baseline"—fitness can be assigned to an individual only in the context of the rest of the population. So a comparison of fitness between an elephant and a bacterium is impossible. Fitness defined in that way does not keep population history—a gain (or loss) of fitness for a whole population cannot be traced, after the fitness of the population increases the organisms are competing with each other in formally the same way. In this context the progressive evolution represents an opportunistic non-directional "Brownian" motion—fixation of accidental "positive" mutations. However, it is quite tempting to have a fitness measure that is "absolute", a baseline measure that reflects the organismal complexity: the total "genetic information" or "evolutionary progress". Such a measure can be naturally used as a fitness function within the population for modeling and also provides a possibility to compare different species. The model presented above is capable of recapitulating all traditional dynamics (e.g., "fixation", "drift", etc.); moreover, it quantifies an additional dimension—total genetic complexity. Modeling evolution without tracing this value can easily lead to "un-physical" outcomes—when the complexity is allowed to drift arbitrarily in the course of sequence evolution. From common observations provide natural expectation: genomic complexity of a given species is a sufficiently preserved on an evolutionary timescale, while underlying genomic sequence may undergo numerous ongoing changes. If the complexity has changed significantly in the course of simulation, the end product should be accounted as a different species. Modeling speciation events per se is a very different subject from modeling species preservation. Here, in essence, a very basic biological phenomenon is modeled: the preservation of form, while the matter (e.g., cells) in this form is continuously renewed. Instead of the matter, the model shows how functional genomics sequences can be continuously renewed while preserving species-specific phenotype, so the phenotype (typical set) and hence the total genomic information are invariants.

Besides introducing the total genomic information invariant, the "physical" property of the presented approach can be further illustrated by the notion of stability: in order for a system to be robust against external perturbations (e.g. random mutagenesis) it must reside in a "potential well". So perturbations are compensated for by forces returning the system to the initial state. In the absence of such compensations, the system would "smear out" in Brownian fashion. Random mutagenesis can both increase and decrease genetic information (GI), and in our case, these compensating forces are selection, which tries to increase GI, and the channel capacity limit, which makes impossible maintenance of GI level above a certain value and increasingly costly approaching the limit from below.

Despite the large number and diversity of existing models their explanatory power remains arguably limited. It is for this reason that Ohta and Gillespie (1996) declared the "looming crisis" referenced earlier, admitting that "all current theoretical models suffer either from assumptions that are not quite realistic or from an inability to account readily for all phenomena." The limitations of current models are likely rooted in the basic definition of fitness and/or the absence of suitable genomic information measure, because if similar in all models, their behavior and fundamental predictions will not change drastically after reshuffling other parameters. Proposed above information-theoretical model provides potential "absolute" measure (GI), estimating the total genomic information, which can be used for the fitness calculations, sensibly accounting for interactions of any number of variants in a genome. Technically, such fitness measure quantifies the degree of "typicality", the size of corresponding "typical set" is related to genomic information or complexity. Fitness connection with complexity is the most essential difference of our model from the traditional approaches, while the modes of reproduction and other parameters are of secondary importance. As the proposed fitness function is somewhat unique, possessing novel features, the properties of the model should be explored starting from the very basic considerations, omitting for the moment phenomena that are routinely considered in standard models such as the influences of recombination, linkage, sexual selection, fluctuating environment, etc. Nevertheless, it is clear that these phenomena are interesting and important, so it would be tempting to investigate them in the subsequent research and to compare the results with traditional approaches.

It is worth noting that technically any fitness expression including "relative" can be applied in the simulation used here to explain the Drake's rule. However, if "relative" fitness is used, additional care should be taken to monitor the equilibrium condition and the complexity dynamics, while the proposed "absolute" fitness expression automatically makes these tasks trivial.

Adami et al. (2000) proposed an interesting approach for quantifying complexity and modeling its increase for digital organisms. Approach considered here is different: information is quantified by a mechanistic model of molecular interactions and the main focus is on the preservation of such information with mutation/selection balance.

It is intuitively clear that the size of genome and total genomic information are somehow related. While in general this relation can be complex and multifactorial, it is clear that accounting for this information might shed light on the coevolution of the genome size and mutation rate, and thus on Drake's rule.

The concept of the "molecular clock" states that the rate of mutation accumulation is roughly proportional to the time of divergence from the last common ancestor. It is likely that both phenomena the "molecular clock" and the "Drake's rule" emerge from the same underlying principal. Roughly speaking, the idea is that the clock is a manifestation of the Drake's rule on the evolutionary timescale, because in spite of the variations in genome size and numerous properties of the population in the course of divergence (assuming the generation time changing slowly), the clock is sufficiently monotonic when comparing protein coding sequences (i.e., mutation rate and density are constant). Common explanation of the clock involves the neutral theory: the majority of mutations behave "as if" they were neutral (Kimura 1983). However, the proper application of the neutral theory to the molecular clock is not straightforward: the clock is ticking monotonously but at a different pace in strongly and weakly conserved genes. Should then some kind of differential "neutrality density" be introduced in order to accommodate "neutrality" for the explanation? A cleaner approach would be to treat mutations with a continuous spectrum of effects—from zero to lethal (fully conserved position). In latter case, it is obvious that zero point (pure neutrality) is not special, because the next infinitesimally close value (minuscule functional) will have indistinguishable properties. It seems that the object in focus of the neutral theory is not the point zero per se, but some loosely defined region in its vicinity.

The assumption of neutrality served as a necessary simplification at the time if its appearance explaining a number of phenomena. Its postulate that the majority of mutations avoids the pressure of selection and is driven by the stochastic process allows investigating evolutionary dynamics of populations without computationally expensive simulations. However, consider a mutation in a "non-functional" genomic region, first it will change replication dynamics due to different weights, shapes, and abundances of nucleotides, then it will affect the local chromosome or chromatin shape. Of course, these changes can be minuscule; however, in a strict mathematical sense the resulting organism is different. Thus, the strict neutrality is rather an exception than a rule. Now the computers are enough powerful to model all mutations as functional, with arbitrarily small effects if required.

Here, I show that while the majority of genomic dynamics can be attributed to the mutations with small effects (as expected), the less recognized aspect is that their cumulative contribution to genomic complexity evolution can be significant because of their abundance. There are also experimental indications of this phenomenon

(Yuan et al. 2013). Indeed, collectively such mutations can behave “as if” they were neutral. However, the reason such behavior is not their weak functionality per se (moreover it is shown below that a degree of functionality does not have a significant influence), but the “saturation” of genomes with information, the proximity to the “channel capacity”.

2.2.2 Prerequisites of the simulation

Let's recall, in our model of positional information storage a functional site (actually a group of functional sites or even the whole genome can be treated as a single “composite” site) is represented by the corresponding sequence pattern determining GI profile, so that for each position corresponding 4-vector of acceptable equilibrium frequencies and hence a value of GI (Formula 11) is defined. Then position-specific GIs can be used to calculate the total amount of information contained in functional site/genome as a simple sum over all positions:

$$GI_{total} = 2L + \sum_{j=1}^{L} \sum_{B \in \{A,G,C,T\}} f_{jB} \log_2 f_{jB} \qquad (18)$$

Where f_{jB}, $1 \leq j \leq L$, $B \in \{A, G, C, T\}$, is the frequency of nucleotide B at the position j and L is a length of the site. So it is obvious that the equilibrium population with allelic frequencies which are more biased from the uniform distribution contains more information. Let's also define an average density of genetic information in a population as $GI_\rho = GI_{total}/L$ bits per site. It is obvious that $0 \leq GI_\rho \leq 2$.

Here, for technical simplicity, individual positions in pattern are assumed to be independent (no epistasis effects), without this assumption it would be necessary to deal with general “typical sets” and the computation of GI would be much more complicated. However, it is reasonable to suggest that this assumption will not change the essential conclusions drastically. While covariable sites are well-known, the problem can be circumvented by grouping significantly correlated sites in “pseudo-sites” (now with more than four states) so that correlations can be broken up with a proper basis selection. However, it is clear that irrespectively of correlations between different sites in genome the latter still have some total genomic information (the formal calculation of GI in this case will be, however, very complicated). Simplified calculations described here merely illustrate the general principles: the interplay between the total genomic information, genome size, and mutation rate. While complexities of formal GI calculations in specific cases can be interesting to investigate, the basic principles of genomic complexity evolution, which are discussed here, are invariant.

It was already discussed that the equilibrium condition is very important for the correct definition and evaluation of GI. However, in general real populations are far from the equilibrium. It is important to understand that GI profile is a “prior”, inherent characteristic of molecular functionality. For instance, a protein domain can retain its functionality only within a certain set of sequences. In terms of GI it means that conserved protein domain has high GI value and hence small size of “typical set”. It is reasonable to suggest that functional genes are conserved in a similar manner (unless some novel mechanisms of molecular functioning are introduced). A simple corollary of this is that the average density GI_ρ cannot be significantly different in close species. The predefined GI profile and a population history (e.g. bottlenecks, selective sweeps and other disruptive events) determine an actual variability of alleles in a population.

Here I will model the dynamics of asexual population, which consists of sequences of same length, subjected to mutagenesis and selection. A series of such simulations for different lengths of sequences will be performed. By means of this numerical experiment I want to demonstrate that in the course of evolution with some predefined minimum level of total genetic information that must be maintained in order to keep the population viable, the pattern of observed mutations will follow the trend postulated by the “Drake’s rule”. In the simulation I assume that GI profile of the equilibrium population and thus the total GI is known (however, the underlying set of allowable sequences is unknown). Additionally, I will investigate observed mutational pattern considering the population already being in the equilibrium state, so that I am spared from having to take into account the population history. However, the equilibrium population is infinite and thus cannot be simulated directly. Due to this, the dynamics of its slice (small subset) will be simulated. This slice generally (except for the degenerate cases when extremely strong conservation is required) has smaller variability than the whole equilibrium population, but the properties of their mutational spectrums presumably will be the same. Realistic biological populations represent such a subset. It is often the case that they have recently (relatively to the mutation rate) undergone a bottleneck event and experienced the “founder effect”—all individuals are closely related through a few recent population founders. The variability in real populations is then very small and does not reflect correctly the underlying GI profile. However, this profile still “exists”, though more as an “ideal”, abstract object like “universals” in Plato’s philosophy. This profile could potentially be revealed if this subset was allowed to diverge for a sufficiently long time without disturbances (in reality only attainable for species with rather small genomes such as viruses and microbes). So the equilibrium population demonstrates the principle limit on the maintainable pattern (total GI, quantifying the total amount of allelic frequencies bias from the uniform distribution), which is defined solely by the mutation rate and reproduction/selection population properties, since the dynamical part (“history”) is excluded. However, it is clear that, with other things being equal, this limit plays the same limiting role for the “collapsed” population (after a bottleneck event).

Actually, the phenomenon to be demonstrated here is somewhat similar to the effect of “error threshold” in the quasispecies model. Also of significance is that modes of mutagenesis and maintenance of variability, which I want to investigate here, are similar to those in quasispecies theory: "The quasispecies concept becomes important whenever mutation rates are high. This is often the case in viral and bacterial populations" (Nowak 1992). In quasispecies theory a population represents a cluster of diverged genotypes. However, the distinction between “normal” species and quasispecies is blurred, and nothing prevents us from considering a “normal” population as the aforementioned subset of quasispecies (in the process of divergence). I presume that the mode of evolution when the mutation rate is high is of most interest and thus deserves careful examination. It is beyond doubt that viral and microbial populations have such mode of evolution. However, I suppose that it is also so for large genomes of higher organisms. What really matters is the mutation rate per genome per generation (this is actually the important corollary of the Drake’s rule) and it is known now that this parameter is quite large in mammals as well—about several hundred mutations, with few in coding regions. In order not to delve too deeply I presume that selection has an opportunity to act in a compensatory manner (to increase GI) in between generations only, however, as discussed above in metazoans germ-line selection issues may play a substantial role. Selection does not "see" a genome size or per-base mutation rate, what it does "see" is the effect of a number of functional mutations, for which it tries to compensate through genetic deaths—removal of the (most unfitted) genomes from a population. So, the natural “units” for selection actions are a genome and a bunch of mutations in it. Those are the reasons for focusing on mutation rates per genome per generation.

The theory of quasispecies was employed to characterize the evolution of HIV, which undergoes 1—10 mutations per replication, thus from the perspective of selection the functional impact (at least in terms of GI) is comparable with the evolution of mammals. These studies infer that HIV population “seems to operate very close to its error threshold” (Nowak 1992). In other words, even a slight increment of mutation rate will lead to a population’s inability to retain its spatial localization in the sequence space, genomic information will dissipate and the population will go extinct (however, it was discussed in the introduction that quasispecies theory says nothing about extinction because the model implies infinite population with soft selection). My main postulate here is the existence of this “threshold” for all species. With the provided IT framework, such a threshold seems to be well-defined and ready for modeling. The main differences between viral and mammalian populations seem to be the time of generation and the size of genome. The "cloud" of allowable viral genotypes can be potentially observed empirically. On the other hand, generation of the actual equilibrium "cloud" for a large, slowly replicating genome would take an astronomically long time and large population size. The latter fact, however, does not mean that it is impossible to explore features of this limit theoretically and then admit that these features will also reflect evolutionary characteristics of the aforementioned population slice. The mode of evolution where allelic variability is maintained in equilibrium is also considered in the quasispecies model. However, in the quasispecies model the fitness is a characteristic of the whole population, not of an individual organism (Nowak 1992). After the definition of a sequence pattern, genetic information and a fitness function based on them are introduced, we are able to measure absolute fitness for each individual in the population and arrive at the simulation procedure proposed here.

Here the analogy with quasispecies model should be taken very carefully since this model and the model presented here differ in some fundamental considerations. In contrast with the quasispecies model, where the "cloud" of genotypes arbitrarily depends on replication/mutation rates and other parameters, and has no deeper meaning, the "cloud" of genotypes in our model represents the typical set, defined as all genotypes producing a species-specific phenotype (defining species total GI, which is missing in quasispecies model). The threshold in presented model is based on the IT notion of channel capacity. Although it is conceptually similar to the "error threshold" from the quasispecies model (the central idea of both is an inability to maintain species-specific phenotype) (Eigen 1971), they have substantial differences. While quasispecies model considers dynamics of infinite populations with non-lethal mutagenesis, presented here model readily suits for modelling of finite populations with lethal mutagenesis. Moreover, Eigen's error threshold is not equivalent to channel capacity, in the sense that the latter does not necessarily entail "error catastrophe"—information can persist with arbitrarily low sequence conservation, where information density can be arbitrarily low. Similarly, in IT reliable transmission is possible with any noise level, but the rate will be lower for higher noise.

One relative issue concerns the fact that the "quasispecies dynamic" requires sufficiently high mutation rate in order to be invoked, otherwise the expediency of the quasispecies approach can be challenged (Holmes and Moya 2002; Wilke and Adami 2003; Wilke et al. 2001). However, some microbes (e.g., wild-type *Escherichia coli*) have mutation rates substantially lower than 1 per-genome per-generation, but still they acceptably fit into Drake's rule because it holds on a logarithmic scale. When mutation rate is so low, a simple simulation presented below will converge to monoclonal population, which will actually reflect *in vivo* situation for a separate bacterial colony. To regain Drake's rule as well as molecular clock phenomenon, some simplifying assumptions should be revised. Constant environment is a good candidate for this purpose: if the environment fluctuates, then we in fact have many different GI-profiles (or one multidimensional profile) where the global population is distributed and individuals are shifting from one environment to the other. In this case, after averaging over all sub-environments and transfers of the population between them, the effective (global) mutation rate must be higher and we will see a "cloud" instead of monoclonal population. Speaking simply, for a species in a fluctuating environment to have some "memory" about different and recurring sub-environments would be an advantage, so they do not need to adapt *de novo* when the environment changes. In that case, the decreased mutation rate can provide this improved memory. If considered in a single sub-environment, such species would look excessively complex and the mutation rate would be below the Drake's rule prediction. Lineages with higher mutation rate were erased by the environmental fluctuations.

These speculations lead to an interesting conclusion: microbes residing in more stable environments should have higher (properly normalized) mutation rates. This seems to be in line with observations: wild-living microbes usually have lower mutation rates in comparison with parasitic relatives who enjoy a host's homeostasis. Traditionally the increased mutation rate is interpreted in the context of the "arm-race" with an immune system of the host. However, the real story might be more complicated—the arm-race (an increased mutagenesis) could be restricted to a few specific genes (the phenomenon observed in some real cases, e.g., cell-surface proteins and so on) while elevation of the whole-genome mutation rate is costly and has no clear motivation. When wild-living microbes "compete" with a rapidly oscillating environment, it might impose genome-wide "racing" pressure, because of large differences in entire metabolism in different environments.

Presumably, in the case of equilibrium maintenance evolution (for $GI_\rho < 2$ bit) a large number of allowable sequences (constituting "typical set") are nearly synonymous and thus can coexist in a population. However, they are not completely synonymous and thus prevent stochastic effects from inducing significant variations in allelic frequencies. The process of selection is capable of maintaining the sequence pattern by discarding the most deviant ("atypical") sequences. The proposed model allows us to meaningfully quantify the information contained in the sequence pattern, given in addition a weight matrix of desired conservation profile, the model provides selective values of individuals accounting for all mutations, present and de novo. As shown above the substitution rate in functional sequences can be arbitrarily close to the neutral rate (see Figure 9), so in general the fraction of positive mutations can be substantial. A trivial requirement for the balance of GI is that about 50% of retained mutations must be "positive".

Feverati and Musso (2008) simulated evolution using the approach based on the formalism of Turing machines. The simulation procedure proposed here is somewhat similar and can be described as a "population of machines", which operate on symbolic sequences of limited length, reading out positional information and recognizing corresponding patterns (via typical sets) of molecular interactions and calculating a high-level phenotype. It should be noted that technically, for a general typical set the assumption of positional independence is not necessary. However, in comparison with the sequential algorithmic Turing machine, the approach presented has at least one clear advantage that makes it closer to real molecular mechanics; specification of the phenotype calculations per se is not required. Once sequence patterns and typical sets in genome are specified the problem of its maintenance or progressive evolution can be addressed (e.g. the cost or speed of the pattern's preservation or change). In this work I focus on the maintenance properties, treating such machines as genetic information storage device that must resist the random noise of mutagenesis. Selection uses a "typicality" of genome as a fitness measure, accounting for all variants (Equation 19). Actually, this fitness function possesses basic "common sense" features that are similar to traditional fitness functions—for instance a mutation in highly conserved site (i.e. site with high GI) will drop the fitness significantly. However, the described equilibrium mode of evolution, where mutation rate is close to its maximum value for which population remains viable, is robust to the changes of specific form of the fitness function. What really matters is the limited number of organisms (genomes) which can be eliminated by selection without leading to the extinction of the population. So it is important to understand that irrespective of the specific form of the fitness function, a limit to the amount of transferable genetic information will exist. Thereby results presented below are considerably general. Another thing worth mentioning is that in this model all sites and variants are functional. Thus there is no need to postulate "neutral" (Kimura 1983) or "near-neutral" (Ohta 1973) variants (to explain the high rates of sequence evolution)—in the described case, the equilibrium can be interpreted as the cumulative neutrality of all mutations remaining in a population, while assuming the individual neutrality of all or most mutations would be throwing the baby out with the bathwater.

2.2.3 Simulation terms

Describing a simulation process I will use some terms specifically, not generally, which will facilitate brevity and ease of understanding.

An "organism" is represented by the nucleotide sequence of given length (L), O = [B_1, B_2, …, B_L], where ∀ i ∈ [1,L], B_i ∈ {A,G,C,T}. A "population" is a set of organisms (sequences) of the same length. The parameters that govern the process of simulation are shown in Table 1.

Table 1. Simulation parameters.

Notation	Description
N	Number of organisms in the population (population size).
L	Length (number of bases) of genome of each organism in the population.
n_d	Number of descendants each organism produces in a single round of reproduction.
P_m	Probability of mutation per base.
P_{ti}	Probability that an occurring mutation will be a transition mutation.
$W = (W_j \mid j \in [1,L])$	Selection weights of nucleotides in each position. Where $W_j = (w_{jA}, w_{jG}, w_{jC}, w_{jT})$, $W_j(B) = w_{jB}$, $B \in \{A,G,C,T\}$ represents the selection weight of the corresponding nucleotide B in j-th position.

The mutational bias (P_{ti}) is included in the code for universality, but has no effect on the trends we investigate here. As discussed above species-specific biases can play an interesting role for GI storage optimization and may slightly affect species dispersion along the Drake's rule trend line. However, for brevity, here it is assumed to be constant. With this notation each organism (O) in a population can be associated with a fitness value (weight) specified by the weight matrix W:

$$W(O) = W([B_1, B_2, \dots, B_L]) = \sum_{i=1}^{L} W_i(B_i) \qquad (19)$$

A "typical" probability of the sequence/organism is its expected probability for a given sequence pattern. It is equal to the product over all positions of corresponding nucleotide probabilities. Ideally a "typical" probability of the organism should be used as its weight. Here, for computational convenience, we define the fitness (Equation 19) through summation of position-specific weights. These weights are assigned in accordance with a given sequence pattern so that a nucleotide with higher probability has larger weight. However, as discussed above, the specific form of the fitness function is not crucial for the investigation of equilibrium evolution if it induces the same topology as "typical" probability on fitness landscape of organisms of the given length. The form of fitness function can only affect the resulting GI_{total} and the magnitude of its fluctuations, not the existence of the limit to GI_{total} itself.

We do not know exactly the resulting GI-profile before the simulation is performed, because the weight matrix defines only a general direction of selection pressure and the final shape of pattern conservation, not the actual GI-profile per se. So the components of the weight matrix are used to determine preferences of selection which tries to maintain a pattern. In my experience, the particular recipes for selection actions (e.g. probabilistic/deterministic) and reproduction modes (overlapping/non-overlapping generations) play little role for the described trends as long as the main purpose of these actions is to maintain a pattern—a biased frequencies distribution, while the opposing force—random mutagenesis, tries to flatten the bias. Each round of mutagenesis decreases the genome's "typicality", on average. So a more "typical" genome has higher reproductive success because its progeny is more likely to stay typical and avoid elimination. As mentioned, GI can be viewed as a convenient measure of the functionally acceptable variant's frequencies biases. Such a fitness definition, in my opinion, is the key departure from traditional models. For example, it seems to be inherently difficult to approach the Drake's rule explanation with a fitness function that is relative—it reveals no information on an organisms' degree of complexity, hence, taken alone, it is "blind" to the size of a genome. In our case the total GI (reflecting organismal complexity) is measured by the amount of a pattern's (functionally acceptable) biases. It seems to be an intuitively appealing quantification—the larger the total amount of biases (further from the uniform distribution), the higher the information content and corresponding maintenance costs. However, such an approach is a necessary simplification—it works under the assumption that the rest ("higher order") information unfolding processes are approximately the same, which should at least work for similar species.

It is reasonable to suppose that sophisticated error correction mechanisms such as DNA repair constitute a biological burden. We can therefore question what value of mutation rate is the highest compatible to a given total GI. The differences of GI_{ρ} of functional sequences are assumed to be small for close species. Formally, for our phenotype-calculating machines, the conservation of GI is equivalent to the whole phenotype conservation, because as we reasoned describing the model of positional storage of genetic information, the GI conservation preserves positional information of molecular interactions so that a phenotype is mechanistically derived from the whole genome pattern.

2.2.4 Simulation process

The entire simulation process can be divided into three successive stages: initialization, spawning and selection. Initialization occurs only in the very beginning and then spawning and selection are repeated in a loop until the simulation process is stopped.

Initialization: the initial population consisting of N organisms (sequences) of length L is generated. All organisms in the initial population are identical and have the maximum possible weight according to matrix W, i.e. at each position j of each organism stands a nucleotide B_j: $B_j = \left[B | w_{jB} = \max_{B \in \{A,G,C,T\}} \left(w_{jB}\right)\right]$.

Spawning: the progeny is spawned. Each organism in the population produces n_d descendants (here we consider in detail only the case of binary fission, i.e. when n_d = 2). A descendant organism has the same length as its parent and is obtained by copying the parental sequence with a certain probability of per base mutation (P_m) and a bias of mutational spectrum (P_{ti}). The parental organism is excluded from the population after the reproduction, so that generations are non-overlapping, resulting in a population consisting of $n_{d*}N$ organisms.

Selection: selection reduces the number of organisms in the population back to the initial size. It acts deterministically, leaving *N* organisms, whose weight *W*(*O*) determined by formula (19) is larger.

The choice of procedure of the initial population generation does not affect the steady state of the simulation process, so we can simply generate a random initial population. However, generating the initial population as described above will provide the faster convergence to the steady state—the equilibrium condition that reveals the "error threshold"—the goal of our experiments. In order to keep things simple I decided to describe here the mode of reproduction with non-overlapping generations. However, I also experimented with overlapping generations (similarly to "Moran process") and found that the resulting trend is invariant.

2.2.4 GI behavior in the course of simulation

Since all organisms in the initial population are identical, GI_ρ of the initial population is equal to 2 bits. However, as I discussed earlier this is not the "correct" functional *GI* but a value formally computed in the course of simulation. If we start the simulation process as described above with the probability of mutation P_m high enough to allow occurring mutations to propagate in the population, diversity will emerge and GI_ρ will start to decrease. GI_ρ will finally reach the level where

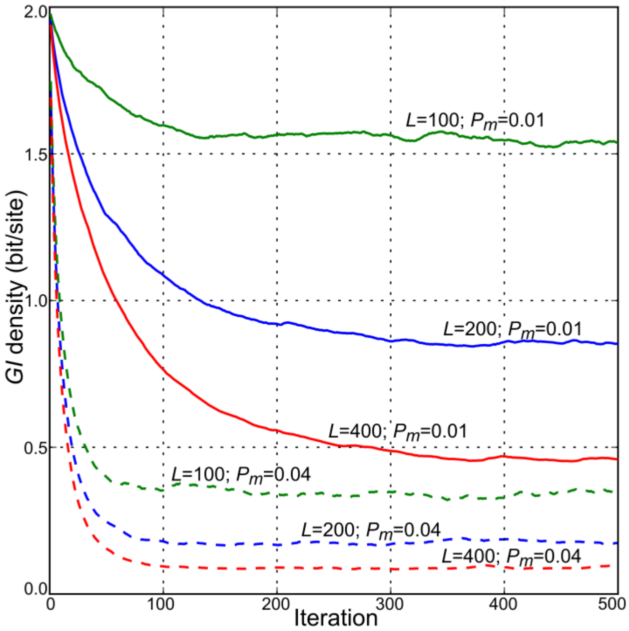

Figure 14. Convergence of GI_ρ for different parameters.
Common parameters for all demonstrated cases are: N = 1000; n_d = 2; P_{ti} = 2/3; W = (W_j = (0.8, 0.2, 0, 0) if j is even, else W_j = (0.5, 0.3, 0.1, 0.1)). Color determines organism length (L): green corresponds to L = 100, blue to L = 200 and red to L = 400. Line style determines probability of mutation per base (P_m): solid corresponds to P_m = 0.01 and dashed corresponds to P_m = 0.04.

mutagenesis is balanced by the force of selection and in consequent iterations will fluctuate in the vicinity of some value (referred to here as the "steady state" or "equilibrium state", though more akin to a "semi-steady state", strictly speaking). The existence of the balance (mean GI_ρ) is clear because the capacity (the averaged effect) of random mutagenesis to decrease GI monotonically drops from some value at GI_ρ = 2, to zero at GI_ρ = 0, while the corresponding selection capacity to increase GI behaves reciprocally—showing a non-zero value at GI_ρ = 0 and zero at GI_ρ = 2, thus these two functions intersect at some equilibrium point. In the numerical experiments I consider that the population has reached the equilibrium state if during the last T generations (T = 100 in our tests) two conditions are met: the sum of all GI_ρ changes between consequent generations is less than a specified threshold (1e-3 in our tests), and the maximum number of consequent generations increasing or decreasing GI_ρ is less than 0.1*T. The convergence of GI_ρ for different parameters is presented in Figure 14.

Fluctuations around the equilibrium depend on particular modeling features. However, in my experience, the fluctuations are smaller for larger population sizes but the mean value of GI_ρ does not vary significantly because he equilibrium state does not depend on the population size, which is natural to expect for the population maintaining constant allele frequencies. Even if we assume a more complicated scenario where fluctuations do not settle, the aforementioned capacities of mutagenesis and selection to change *GI* cannot depend significantly on the population size because they operate on the variant's frequencies, which are disentangled from the absolute population size. Hence the balance (even the dynamic balance) between these two forces is also free from the population size dependence.

Here I will call the state of the simulation when the population has already reached equilibrium as the GI-steady state and denote the mean value of GI_{ρ} in the equilibrium population as GI_{steady}. So GI_{steady} represents a maintainable level of genetic information for a given species. It can be also called a "mutation-selection balance". However, it is clearly different from Fisher's balance (Crow 1986), who considered a single site. In our case the balance is due to the compensatory effects of multiple positive and negative mutations. Other authors considered a balance similar to ours when the frequency of positive mutations is high so that they cannot be easily brought to fixation as in one-by-one case (Sniegowski and Gerrish 2010; Desai and Fisher 2007). This approach also differs from the approach described here. The most notable difference is that here we are not concerned with the fixations at all, and we quantify the limit on genomic complexity—as we discussed earlier, without considerations for this limit, a formal modeling might easily result in "un-physical" solutions. It should, however, be clearly understood that "steady" here connotes only the genetic information (and hence the phenotype). Individual genomes remain variable because new mutations still appear with a constant rate (Figure 15). The "molecular clock" is ticking and its empirical steadiness on the evolutionary scale is another indirect hint that the average *GI* density is a slowly varying parameter. For example, mutations are more frequent in a position with lower *GI* value, so if the density fluctuates strongly on the evolutionary scale, the clock will behave erratically. As already argued, *GI* increasing (positive) mutations constitute a significant fraction of random mutations (especially when *GI* in a position is low), thus allowing the same fraction (in the *GI* equivalent) of negative mutations to remain in the population. The monotonous molecular clock is traditionally explained by the neutrality assumption, which seems to be an oversimplification of reality. However, the proposed model provides the same prediction without resorting to implausible assumptions. Also the provided model shows that the steadiness of the clock is intimately connected with Drake's rule and the "error threshold", while the neutral theory is inherently unable to make such connections.

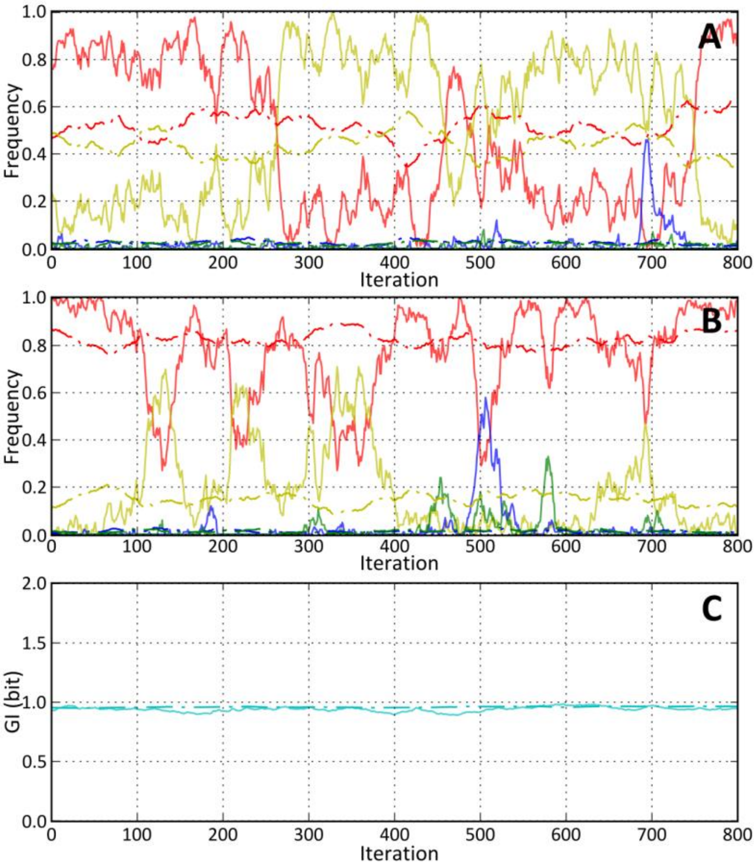

Figure 15. Fluctuation of positional nucleotide frequencies during *GI*-steady state for different selection weights (*W*) and population sizes (*N*). Common fixed parameters are: P_m=2^{-6}, P_{ti}=2/3, *L*=128, n_d=2. In all three subfigures (A, B, C) the line style defines population size: the dash and dot line correspond to *N*=10000, the solid line to *N*=100.

A: Fluctuations of nucleotide frequencies in a position (*P*) with selection weights W_P=(0.4, 0.38, 0.12, 0.1).

B: Fluctuations of nucleotide frequencies in a position (*P*) with selection weights W_P=(0.5, 0.3, 0.1, 0.1).

C: Dynamics of GI_{steady}.

2.2.5 Counting mutations

In contrast with *in vitro* experiments in the simulation, the number of fixed (observed) mutations per generation can be counted directly. Following the common notation I denote the number of observed mutations per base per generation as u_b, and observed mutation rate per generation per genome as u_g. Despite the fact that the values u_b and P_m are closely related, u_b is always less or equal than P_m because the organisms with more mutations are more likely to be eliminated at the selection stage.

In the first experiment described above, all parameters from Table 1 were fixed and the convergence of GI_ρ to its limiting (equilibrium) value (GI_{steady}) was traced (i.e. the value of GI_{steady} was not predefined, it was determined in the course of simulation).

Now let us look at the somewhat inverse experiment: we can fix the value of GI_{steady} and all parameters from Table 1 except P_m, and then numerically find the value of P_m which corresponds to the fixed parameters. This procedure was performed for all combinations of different organism lengths $L \in \{64, 128, 256, 512, 1024\}$, different values of $GI_{steady} \in \{1.2, 1.4, 1.6\}$ and different weights $W \in \{[W_j=(0.8, 0.2, 0, 0)$ if j is even, else $W_j=(0.5, 0.3, 0.1, 0.1)]$, $[W_j=(0.9, 0.1, 0, 0)$ if j is even, else $W_j=(0.4, 0.3, 0.2, 0.1)]\}$. Other parameters in all experiments were fixed: N=1000, n_d=2*;* P_{ti}=2/3. In the experiments we estimated the number of mutations observed in the GI-steady state and compared u_b parameters for different genome lengths. The results are summarized in Figure 16.

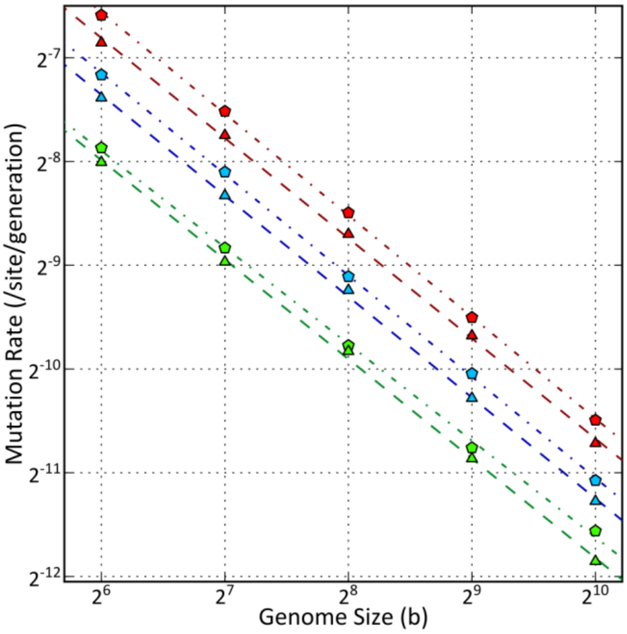

Figure 16. Relationship between the mutation rate per site per generation (u_b) and the genome size (L) observed in the simulation. Color determines density of genetic information in the steady state (GI_{steady}): red—GI_{steady}=1.2 bit/site, blue corresponds to GI_{steady}=1.4 bit/site, green to GI_{steady}=1.6 bit/site. The shape of the marker determines selection weights (W): the pentagon corresponds to $W_{pentagon}=(W_j=(0.8, 0.2, 0, 0)$ if j is even, else $W_j=(0.5, 0.3, 0.1, 0.1))$, the triangle corresponds to $W_{triangle}=(W_j=(0.9, 0.1, 0, 0)$ if j is even, else $W_j=(0.4, 0.3, 0.2, 0.1))$. Lines represent linear regression on a log-log scale. Dark red/blue/green lines correspond to light red/blue/green markers; dash and dot lines correspond to pentagons, dashed lines correspond to triangles.

Regression lines and corresponding correlation coefficients (r2):

GI_{steady}=1.2, $W_{pentagon}$ (red dash and dot line): $\log_2 u_b = -0.68 - 0.98 \log_2 L$, (r2=0.99)

GI_{steady}=1.2, $W_{triangle}$ (red dashed line): $\log_2 u_b = -1.02 - 0.97 \log_2 L$, (r2=0.99)

GI_{steady}=1.4, $W_{pentagon}$ (blue dash and dot line): $\log_2 u_b = -1.29 - 0.98 \log_2 L$, (r2=0.99)

GI_{steady}=1.4, $W_{triangle}$ (blue dashed line): $\log_2 u_b = -1.52 - 0.97 \log_2 L$, (r2=0.99)

GI_{steady}=1.6, $W_{pentagon}$ (green dash and dot line): $\log_2 u_b = -2.31 - 0.93 \log_2 L$, (r2=0.99).

GI_{steady}=1.6, $W_{triangle}$ (green dashed line): $\log_2 u_b = -2.23 - 0.96 \log_2 L$, (r2=0.99).

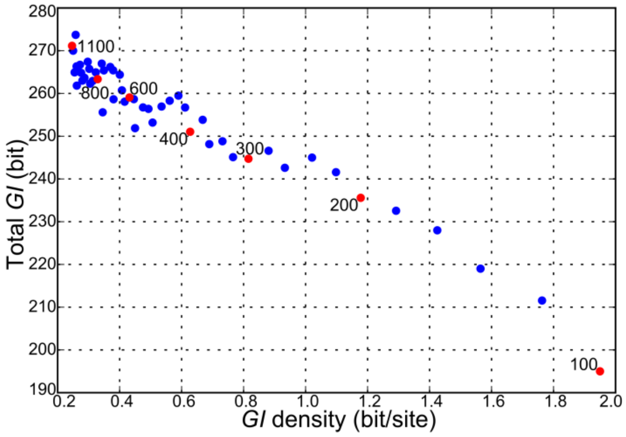

Figure 17. Dependence of total genetic information (GI_{total}) and density of genetic information (GI_{ρ}) on the length of genome (L) when the rate of mutations (P_m) is fixed. Each point represents a population with organisms having genome of size $L \in [100, 120, \ldots, 1080, 1100]$. For convenience of orientation some points are colored in red and genome size of corresponding population is labeled. Mutation rate (P_m) was fixed to *0.007*. Also, all other parameters were identical for all populations, namely: N=1000; n_d=2; P_{ti}=2/3; W=(W_j=(0.8, 0.2, 0, 0) if j is even, else W_j=(0.5, 0.3, 0.1, 0.1)).

Figure 17 demonstrates total genetic information (GI_{total}) and density of genetic information (GI_{ρ}) depending on the genome length (L) when the rate of mutations (*Pm*) is fixed. For each population displayed GI_{ρ} and GI_{total} were averaged over 1000 generations after the population reached *GI*-steady state.

Defining weight matrices in a different way, various scenarios of the GI density distribution were tested: homogeneous *GI* distribution in a genome (where all positions have the same distribution of weights, i.e $\forall\, B \in \{A, G, C, T\}$, $\forall$ i, j $\in [1, L]$ $W_j(B) = W_i(B)$) and bimodal. In the latter case one half of a genome consists of highly conserved ("lethal") sites to model the regions such as conserved protein domains, and the other half consists of weakly conserved sites to model the variable parts of proteins and weakly conserved non-coding regulatory DNA. I found that the meaning of the obtained results for the regression of the average GI density on the mutation rate at the equilibrium remains the same for different modes, so the observed trend is stable.

2.2.6 Manifestation of Drake's rule as a consequence of approaching channel capacity

As demonstrated in Figure 14 starting from some initial distribution of allelic frequencies, after a number of reproduction/selection rounds (during which all parameters from the Table 1 are fixed, simulating the evolution process without any disruptive events) the population comes to equilibrium state. This means that in the consequent iterations the density of GI (GI_ρ) fluctuates in the vicinity of the GI_{steady} value, preserving the total amount of information (GI_{total}) contained in the population. The value of GI_{steady} depends strongly only on the rate of mutagenesis (P_m) and intensity of reproduction (n_d, which determines selection potential). Changes of other parameters have minor effects. Since in the presented study the value of n_d is fixed in all experiments, GI_{steady} is determined by P_m. The value of GI_{steady} represents the maximum amount of information per position (averaged over all positions in genome) that is possible to maintain in the population from generation to generation with a given mutation rate (P_m). If we reduce mutation rate, GI_{steady} will expectedly increase, allowing for the transmission of more information *in toto* (GI_{total}). E.g. in the case of a mutation rate equal to zero ($P_m = 0$) GI_{steady} will be 2 bits. For another extreme case, when $P_m = 1$, GI_{steady} will obviously be zero.

Here I assume that the mutation rate in species is near the upper limit of tolerable values, so even a slight increase will make population unstable and eventually leads to extinction (the background of this assumption will be discussed in the next section). So from the viewpoint of IT, GI_{steady} can be naturally interpreted as a channel capacity that transmits the information about the underlying pattern from generation to generation, for a given level of noise (mutation rate).

Let's now take a look on channel capacity from the perspective of fitness. In the course of equilibrium evolution, genomes of species migrate within the set of allowable sequences (which determines corresponding sequence pattern), having for the most part very close fitness. It is reasonable to assume that allowable set forms a connected set (probably having rather simple shape at least locally) in the sequence space with Hamming distance. More conserved genomes form smaller allowable sets, e.g. in the limiting case when absolute conservation is required, the allowable set contains only one sequence. However, for all real species such sets are extremely large (probably much larger than their population sizes, especially for species with large genome). Positive mutation moves sequence "deeper" into the allowable set, increasing the distance from the exterior, negative conversely pushing the sequence towards the exterior. From these geometrical considerations it is clear that in general if an organism undergoes a positive mutation then its descendants will more likely experience further negative mutations. An assumption that mutation rate is at its upper tolerable limit suggests that most of the organisms are very close to the boundary of the allowable set (the neighborhood of the boundary contains most typical sequences). So from the viewpoint of fitness, the size of allowable sequence set of the population determines the channel capacity.

Remember now the second experiment, where for the given value of GI_{steady}, the corresponding mutation rate (P_m) was estimated, and after reaching the equilibrium state the number of fixed mutations was calculated for different lengths of genomes. In light of the above I can say that by fixing the value of GI_{steady} we determine the minimum value of average GI density required for viability of the population. Thus the second experiment allows us to estimate the upper limit of the mutational rate and the corresponding mutational spectrum for the equilibrium population having a predefined average density of GI. Let's recall the "Drake's rule" in its original form: there is an inverse relation between mutation rate per base per generation (u_b) and the size of genome (L) in microbes, so that mutation rate per genome per generation (u_g) is approximately constant. That is exactly what is demonstrated in Figure 16, which shows a clear inverse relation between genome size (L) and mutation rate per site per generation (u_b), while mutation rate per genome per generation (u_g) remains almost constant. For both values of GI_{steady} presented in Figure 16, the variation of u_g is less than 5% of the average u_g value among all experiments with corresponding GI_{steady}. The variations of u_g in real microbes are larger than in our *in silico* experiment. This may be due to the fact that while closely related species have s similar values of GI density (GI_ρ), these values are not identical (like in our simulation). The variations of GI_ρ among microbes probably are quite small, because in general more than 90% of a microbial genome constitutes protein coding sequence, a characteristic presumably conserved in different species to the same extent, so the general trend of relation between u_g and genome size in microbes remains clear. On the other hand, I expect that GI density in genomes of higher organisms is much lower than in microbes, because usually only a small fraction of their genomes encodes proteins, a significant part of functionality is borne by non-coding DNA possessing low conservation and thus low density of GI and potentially substantial part with no function. Figure 16 also demonstrates that species with lower density of GI can tolerate higher mutation rates (for a fixed length of genome). It is also clear from Figure 17 that for a fixed mutation rate longer genomes are capable of transmitting more information. These two facts possibly explain the reverse relation (direct proportionality) between u_b and L observed in multicellular eukaryotes. That said, however, if we consider only protein coding sequences, the relation between u_b and L observed in microbes mostly holds true for multicellular eukaryotes as well. The latter fact props up once again the presented explanation of the observed pattern of mutational rate, because as I mentioned above, what really matters is the GI density of the sequence, and the GI density of protein coding sequences is likely very similar among the vast majority of all biological species.

3 DISCUSSION

> *… every item of the physical world has at bottom—at a very deep bottom, in most instances—an immaterial source and explanation; that what we call reality arises in the last analysis from the posing of yes-no questions and the registering of equipment-evoked responses; in short, that all things physical are information-theoretic in origin and this in a participatory universe.*
>
> Wheeler JA (1989) "Information, physics, quantum: the search for links", Proceedings III International Symposium on Foundations of Quantum Mechanics, p. 354-368.

Only about 1.5% of human genome encodes protein sequences, while the functional significance of leftover noncoding DNA still remains largely unclear. Nonetheless, technological progress of the last decade gave us a substantial insight into noncoding functionality. Recent studies demonstrate that the manifestation of purifying selection in noncoding DNA is a widespread phenomenon (Kamal et al. 2006). According to a number of estimates the proportion of functional noncoding DNA in human genome may be more than 10-fold larger than the share of coding sequence (Smith et al. 2004; Ponting and Hardison 2011; The ENCODE Project Consortium 2012). Such estimates are often based on indirect evidence of functionality such as transcriptional activity (in case of *in vitro* experiments), or intricate techniques for detection of inter-species sequence conservation requiring precise calibration and nontrivial assumptions (*in silico* studies). Therefore the variance of these estimates is very high: while the most conservative suggest that the fraction of functional noncoding sequence is about 8% of human genome, the boldest estimates give a value which is one order of magnitude higher, reaching 80% of genome. However, whatever estimate is correct, it is clear now that a substantial part of genetic information in large genomes is concentrated in noncoding DNA. Functional noncoding elements are usually characterized by a high turnover rate, avoiding detection and investigation by conventional conservation-based methods. Up to now it remains largely unclear how genetic information can be reliably stored in such highly-variable sequences. So, despite the fact that the significance of noncoding functionality is currently apparent, we still lack an adequate model to describe the evolution of functional sequences experiencing a high mutation rate. The model of positional information storage in sequence patterns presented in this work is addressed to fill this gap. The model shows that, in principle, it is possible to store any amount of error-free genetic information with arbitrarily high substitution rates provided sufficiently long sequences. Considering the analogy between genetic and digital information transmission, this possibility is ensured by the noisy-channel coding theorem—one of the core results of Shannon's IT. Before Shannon's revelation, a high signal/noise ratio was accepted, in practice, with some errors during transmission to be unavoidable. However, the IT demonstrated that energy efficient error-free communication is possible given any level of noise. A similar situation is observed in genetics: the intuition that functional sites must possess high conservation (high signal/noise ratio) went as far as to call all weakly conserved sequences "junk DNA", while in this work I speculate (keeping faith in nature's thriftiness) that weakly conserved functional sequences (constituting an overwhelming majority in large genomes) represent evolutionary innovation for increasing efficiency. In spite of the strong analogy with Shannon's classical communication scheme, the situation considered here has a unique (somewhat counter-intuitive) feature: the proposed model suggests that a large fraction of random mutations is "positive" (i.e. is compensatory for GI storage as shown in Figure 7), while in the traditional IT model all noise is "bad". It was shown that the composition of positional nucleotide frequencies can be optimized to minimize the impact of GI-decaying mutations, taking advantage of compensatory (GI-increasing) mutations. However, for low GI values about 50% of mutations are "good", regardless of the nucleotide frequency's vector optimality.

It is clear that the applicability of the model to the evolution of real molecular machines should be thoroughly investigated since any formal model has its limits. For instance the interaction between particular alleles is apparently more complex than it is described by the model and as mentioned above, such interaction can be accounted for by constructing more complex typical sets. However, based on a few reasonable assumptions the model provides simple explanations for observable phenomena and operates similarly to our understanding of the machinery of molecular interactions, which could then be easily implemented in software. For these reasons I believe that this model deserves careful attention.

While classical models usually consider evolution of defined sequences, the paradigm in this work is shifted to model the evolution (and conservation as a special case of particular importance in this work) of probabilistic sequence patterns. In this framework, a mere shift of allele frequencies, rather than a fixation of allele is considered as an elementary act of evolution. This seems to have little sense for a single allele. However, applying this scheme to millions of alleles in a population and considering that frequency of beneficial mutations can be high we come to the mode of evolution which is in marked contrast to the traditionally considered mode. Instead of modeling fixation dynamics of single alleles with arbitrary assigned selective values, the framework, due to additivity of GI, allows us to model quantitatively the evolution of the total genomic information. A high frequency of beneficial mutations brings up the issue of determining the forces restricting the potential of progressive evolution and their limits, e.g. why some species are stable for millions of years.

Interestingly, we can consider the model as a simple generalization of the Hardy-Weinberg equilibrium (HWE) (Hardy 2003) where maintenance of functional sites is included explicitly. This may explain the persistent (half-century) illusion of the neutrality—mutations arising during maintenance evolution of equilibrium population will pretend to be neutral in the usual tests (e.g. Tajima's D test (Tajima 1989)) which actually assess the (local—in the case of recombining population) equilibrium condition, rather than the individual mutations neutrality. Below I discuss possible consequences for our understanding of evolution of real genomes, presuming that the model is sufficiently valid.

To prevent possible criticism, I want to emphasize that the model as presented above describes features of an idealized population. One interesting feature of the model is that it can potentially reconcile the "heated" debates between neutralists and selectionists, since it suggests that evolution is mostly neutral (in stasis) but this neutrality is maintained by the selection of the most typical individuals. I believe that the investigation of abstract models, regardless of their immediate relevance to "actual practice", is a normal epistemological practice since the corresponding applicability domains can be rather specific and are not yet well established and delimited. Naturally, this does not imply that the role of classical phenomena is ignored. For instance "selective sweeps" caused by "strong selection" are apparently non-equilibrium events and are out of scope of the equilibrium model. Roughly speaking, such an event provokes replacement of the whole population. However, in terms of GI this large-scale dramatic incident alone provides only 2 bits of GI for a given site. Generally any phenomena caused by a changing environment represent non-equilibrium events. The model assumes a constant environment and infinite time for equilibration, so all such events would occur with time for variability of the genome to settle around a new phenotype (allele composition). However, I propose that, even if a changing environment is considered, the model describes the "background" of such (presumably relatively rare) events. The frequency of such events should be limited by Haldane-type arguments (Haldane 1957), so it is reasonable to assume that the remaining mutational background can be better explained by the provided model than by the classical neutral approximation. Actually, according to the model a mutation (regardless of its selective value) per se, while changing the typicality or fitness of an individual organism, cannot affect the amount of total GI in the equilibrium population (this phenomenon will be discussed below). Therefore the model presented in this thesis has a well-defined, restricted applicability domain. However, through abrupt or gradual changes of a GI-profile, it is possible to extend the model to certain non-equilibrium scenarios of changing environment. Admittedly, the majority of variants in real populations have weak effects, however, their number can be quite large making their collective effects far from negligible (as in the simplistic interpretations of the neutral theory), here a consistent way to account for such effects is suggested.

As pointed out by Lynch (2010a): "mutation is the ultimate source of all variation, both adaptive and deleterious, a mechanistic understanding of the evolutionary process will be incomplete until a detailed account has been made of the rate of origin, molecular nature, and phenotypic consequences of spontaneous alterations for a diversity of organisms". So it is clear that understanding of how mutations accumulate in genomes is crucial for the comprehension of the evolutionary process. According to Drake (Drake et al. 1998) the genomic mutation rate "is likely to be determined by deep general forces, perhaps by a balance between the usually deleterious effects of mutation and the physiological costs of further reducing mutation rates". It is important to note that reflecting on the nature of genomic mutation rate, Drake did not include considerations for adaptive properties of evolution, practically solving the problem by hinting that it is rather a maintenance-related phenomenon. In the presented model maintenance is interpreted as the equilibrium in alleles' frequencies (and as a consequence, conservation of GI profile)—the main property of the model. In the framework of such interpretation, the size of population is obviously out of the equation (as in the case of HWE).

The key assumption of an explanation of the "Drake's rule" proposed here is that the total genomic information is saturated to its maximum maintainable level, or equivalently, that the mutation rate is near its upper limit for a given total GI of the species. I assume that the mutation rate and hence the total GI change slowly on an evolutionary timescale and hypothesize that the decrease of the rate is a basic prerequisite for progressive evolution. When after some spontaneous variation the mutation rate decreases, the total GI is gained promptly, reaching a new maximum maintainable level and restoring the equilibrium. Judging by the speed of convergence to the steady state demonstrated in Figure 14, equilibrium can be regained very quickly (~100 generations). The difficult question here is how the stability of the mutation rate for a given species can be motivated. So why does the rate of mutations not decrease or increase? Both an increase and decrease of the mutation rate have positive and negative impacts on evolution for different timescales.

On one hand, decreases in the mutation rate give an advantage that allows higher levels of total genetic information. However, this advantage is long-term because some generations must pass to fill newly accessible GI (if a niche requires it, which does not have to be the case in general). On the other hand it brings immediate disadvantages—“physiological costs”—since the lower mutation rate, in principle, must be associated with a slower replication rate and/or additional energy expenditures. Everything is exactly the opposite when we consider an increase of the mutation rate: it provides a long-term disadvantage, reducing the level of total GI and instant benefits, as well as “physiological costs”. So why does the mutation rate not degrade? For higher organisms, it is possible to speculate that an increased somatic mutagenesis might also cause short-term detriment, speeding up aging and promoting carcinogenesis. Besides somatic mutagenesis, there may be many other selectively important traits that are somehow linked to the changes of mutation rate.

Another idea asserts that while the decrease of the mutation rate must run into some “physiological costs”, the way back is not so easy—a mutation that degrades the rate does not necessary reduce the “physiological costs” back to the previous values. Such a mutation must be a rather specific “back-mutation” or, even more likely, a number of them. Thus it is practically almost impossible to achieve simultaneously both the increase of the mutation rate and a reduction of costs. Therefore, rates can only go down, locked from above by both short- and long-term disadvantages. The maintenance of mutation rate, in turn, might require a regular renewal of the population, described below. Of course there are plenty of examples demonstrating regressive evolution—such evolution can be easily caused, for instance, by moving to a simpler niche (habitat)—“use it or lose it”. It is woth noting that for the model presented here, regressive evolution is not a priory less frequent than progressive. If the niche is separated sufficiently, a subpopulation with increased mutation rate will degrade to a simpler species. The blind salamander living in caves is a potential example of such (organ-specific) decrease of complexity. This salamander has an atavism: rudimental eyes (sometimes the eyes are absent completely). If we had more data on such reversed “atavisms”, we could assess how popular “degrading” evolution is.Technically, it is not reasonable to reject the possibility that a number of simpler species might “devolve” from more complex ones. While the latter must have ascended from some simpler ancestors. In fact, it is likely that the topology of the evolutionary tree resemples a willow tree (i.e., numerous descending branches, from a few thick nearly horizontal branches—“living fossils”).

Hypothetically a slight change of the mutation rate is able to change species phenotype drastically and promptly (on the evolutionary timescale), since a minor relative change of mutation rate can cause a substantial modification of total accessible GI and a correspondingly significant change of the phenotype. Such argumentation can be used for an explanation of "punctuated equilibrium" phenomenon. In principal this hypothesis can be verified experimentally—the model predicts that population (for instance) of flies with a reduced mutation rate has an opportunity to give rise to an advanced species of flies. If we were able to select flies for lowered mutation rate, this prediction could be tested. However, a reduced rate of mutation is not enough to stimulate evolutionary progress. Another required prerequisite is a properly challenging environment which will provides an impetus to progressive evolution, thereby promoting an increase in complexity.

Due to the discreet nature of mutations modifying mutation rate and their (presumably) rather limited number, it is reasonable to expect that the mutation rate can take only a finite number of values (i.e. it is highly discrete). In this context it is possible to hypothesize about speciation scenarios. Consider a large population and suppose that the mutation rate is heterogeneous, so that the population has some average rate. Then imagine that a small group of organisms becomes separated from the main population. Suppose the group has much lower variance of mutation rates and the average mutation rate is far different from that of the origin population. These "founders" will produce a new population and the difference in mutation rate will lead to fast phenotype changes relative to the parent population.

It is presumed that the evolution of the size of a functional genome occurs by dint of gene duplications (Ohno 1970) so that sizes of "gene families" increase. This standpoint also promotes the postulate discussed above of slow changes of GI_ρ for functional sequences because of the likelihood that molecular functions of new sequences are in some way similar to those of the original. The theory presented predicts that partial or whole genome duplications accelerate the rate of sequence evolution and is followed by a shrinking back of the (functional) genome size and the loss of extra copies of genes due to an inability to maintain higher total *GI* without changes in mutation rates. Thus the reduction in mutation rate and/or the adjustment of lower GI_ρ functionality (Figure 17), rather than duplications per se, provide the basis for the evolutionary progress (an increase in complexity and total *GI*). It is also worth noting that duplications are relatively frequent events, whereas a slowdown of mutagenesis and involvement of functional sequences with lower density of GI are assumed to be "slow" processes. This hypothesis can be reinforced by recent experimental studies of RNA viruses. It is known that large- and intermediate-sized nidoviruses encode an enzyme implicated in controlling RNA replication fidelity, while other single stranded RNA viruses, with smaller genomes, do not encode the enzyme (Lauber et al. 2013). On one hand it is reasonable to argue that an acquisition of this enzyme have promoted genome extension (Nga et al. 2011). On the other hand, Eckerle et al. (2007) demonstrated that viruses containing a defective mutant of the enzyme-encoding gene possess an enhanced mutation accumulation rate. However, it is clear that progressive evolution is affected by external conditions (niche or habitat), which must be sufficiently complex to support the increase of species complexity. Duplications can lead to reproductive isolation. This fact, along with the hypothesis about the founder-specific mutation rate discussed above, provides a potential way to speciation and progressive evolution.

The IT notion of “channel capacity” is sufficiently weighty and general enough that I therefore suggest it can provide adequate comprehension of the “Drake’s rule”. Moreover, such an approach also readily allows numerical simulations of the process. Channel capacity is the tightest upper limit on the rate of error-free information transmission for a given level of noise. Real communication systems (currently) have a rate of transmission that is somewhat below this theoretical bound. Engineers make great efforts trying to approach this limit because the closer to it a communication system is, the more energy can be saved. Hence another principal consideration is that the nature is “thrifty” so it is not clear why it would not utilize the genomic informational capacity to its full extent, avoiding wasting resources on unused capacity. Selection should favor the thriftiness (though there are some opposing ideas of “selfish” or “parasitic” sequences). It is reasonable to presume that early genetic systems operated at the “error threshold”. If this holds true it is not clear at which point and wherefore the departure from this limit occurred. To stay always at the threshold seems to be the fastest and most energy-efficient way to progress, while the threshold itself (i.e. capacity) is pulled up by the enhancements of replication fidelity and possibly other mechanisms. If we consider the “costs”, it is difficult to come up with even a synthetic reason to push the fidelity beyond necessity. Thus, until we find at least some strong argument against, we have to admit (following “Occam's razor”) that contemporary species also operate at the “error threshold”. Taking this threshold into account can profoundly deepen our understanding of evolutionary processes, while ignoring it in evolutionary modeling is fraught with loss of adequacy.

Curiously, the problem of approaching channel capacity in the framework of IT has no general solution suitable for all practical situations, as it is related to the problem of achieving the best compression rate and, in practice, is limited by computational costs and memory. Following the reasoning of Chaitin (Chaitin 2012) it is possible to speculate that “molecular machines”, while attempting to approach the limit, have an infinite field for exercising mathematical creativity. The latter can be also used as an argument for the drive to complexity in living systems. Of course, the simple model presented here is able to capture only crude properties of the genetic information process as there are many features of real processes that are not included in the model—epigenetics, recombination, genomic rearrangements, the roles of transposable and repetitive elements, multiple ploidy, etc.

Large genomes contain a lot of repetitive and transposable elements (usually highly variable) at first glance this contradicts to the thriftiness notion. However, as shown in Figure 17 in case of informational saturation utilization of low-density GI sequences might be advantageous (e.g., the ENCODE project (The ENCODE Project Consortium 2012) seems to support the broad functionality of intergenic regions). From the IT point of view (semantically), the information content will not change if we will repeat the same message many times. Thus what really matters is not repetitive sequences per se but their structural properties.

Genome size of plants and animals can be both much larger and much smaller in comparison with mammalian genomes, while the number of genes is approximately the same. On one hand it is reasonable to suggest that the size of genome does not strongly affect organism performance. On the other hand, there should be a balance between the proliferation of these elements and some counteracting force (otherwise unconstrained multiplicative process would lead to an exponential growth). This balance represents independent from usual substitutions degree of freedom for phenotype tinkering, and its model can be developed similarly to the presented model. It is likely that the tinkering affects large-scale chromatin organization and is not much different from the usual mutagenesis. However, it could be useful that it is independent from it. Once the 3D nuclear organization became functionally important, some means of tinkering provided additional dimensions of variability. Presumably, substitutions and small indels cannot substantially affect nuclear organization (directly), so they are not well suited for this dimension of phenotype tinkering, in comparison with large-scale rearrangements and mobile elements. However, it is important to understand that proliferation of such elements does not immediately imply progressive evolution, an increase in complexity (which yet has to be defined formally for such elements). Similarly to the mutation-selection balance of normal mutations, the increase of complexity requires special "creative" events that affect the balance (mutation rate decrease etc.), and that balance is restored again quickly after such events.

Another thriftiness-based (posterior) "prediction" concerns CpG sites. Due to their hypermutability these sites are heavily under-represented in mammals and some other lineages. However,

it is known that highly conserved "CpG islands" exist in some functional regulatory regions. Either they are protected from mutagenesis by some special mechanisms or by simply purifying selection apparently they produce additional costs for their carriers. These costs must be balanced by some benefits, and indeed they have an additional informational capacity obtained due to the ability to be methylated. Similar reasoning may be applied to other over-conserved sequences (e.g. histones).

Also, the "silent" substitutions (which do not affect protein sequence) are unlikely to be strictly neutral, since organisms would capture unused informational capacity. Their applicability for calibration purposes should be carefully evaluated. Indeed, there are many reports which show their functionality potential.

Comparing the proposed above interpretation of the "Drake's rule" with another recently suggested (Sung et al. 2012), it is worth mentioning that the explanation presented here does not require additional difficult-to-define entities like "molecular refinements", "drift barrier" or "effective population size" -- the estimates of the latter are admitted by Sung et al. to be "fraught with difficulties". The verification of that evolutionary model *in silico* seems to be quite problematic, since genome-wide functionality and conservation are not defined. Hence there is no specific model for selection actions and many arbitrary parameters. As a consequence it is not clear how that model can be simulated. However, the ability to simulate and to assess the robustness in the parameter space is a very desirable feature of any "mechanistic" evolutionary model. Comparing Figure 1A in (Sung et al. 2012) with Figure 16, it is possible to hypothesize that eukaryotes have lower GI density on average, which is consistent with weaker genomic conservation observed. Figure 17 shows that storing genetic information in functional sequences with lower density can be advantageous. So perhaps eukaryotes resorted to this strategy. In general GI storage strategy can be affected by particular demands for optimization. Viruses and bacteria may prefer compact genomes with high density of GI, for example, utilizing the double stranded and overlapping coding and avoiding weakly conserved regulatory noncoding DNA because of their need to replicate fast and have small physical size.

One important conclusion that can be drawn from the above reasoning is that the molecular evolution on average is not about a continuous increase of total GI. This can shed light on the naive but still valid question of why we see "living fossils" (a species that "stopped evolving" and keeps its phenotype unchanged for millions of years), while on the other hand we can observe an amazing phenotypic plasticity (e.g. dogs pedigrees or Cetacean evolution). Despite being "adaptive" for a given environmental change, evolution is not "progressive" in terms of total GI, as I posit that each species already has the maximum total GI allowed by the mutation rate, which is assumed to vary slowly. Having this in mind it is also tempting to revisit the popular evolutionary concept stating that genes are near their best functional performance, balancing at the brink of "chaos and order". In terms of the proposed model we can say that performance is as good as allowed by the corresponding channel capacity, so that, in general, a random mutation has a high chance of being positive.

The postulated dependence of the "evolvability" on the size of the population is dogmatized in classical theories. This stems perhaps from the historically formed opportunistic "Brownian" concepts of the evolution. In my opinion, strong dependences on the population size in conventional models may have led to some contradictions with observed phenomena, such as Lewontin's "Paradox of Variation" (Lewontin 1974). Not to mention that the general tendency shows that on average more highly evolved species have smaller population sizes. However, the presented model allows us to reconsider the role of population size in the evolutionary process at least for the maintenance mode. The model suggests that if a population evolves at the limit of total GI, the gain of advance in one function entails losses for other. In this case, as I showed, the impact of population size may be diminished, at least for the maintenance mode of evolution. In this scenario, when an individual receives an advantageous mutation, its progeny will tolerate and keep more disadvantageous, new mutation-hitchhikers (and the outcomes of recombination), which will eventually nullify the effect of the initial mutation. Qualitatively, similar information "jamming" was also explored in the chapter "Conflict Resolution" in Forsdyke (2011). Moreover, the model presented here allows us to draw a scenario for evolvability vs. population size that is somewhat opposite to the conventional models: random mutations, if they have no instant deleterious effect, will on average increase the rate of mutations. So in the long-term a large population is able to accumulate many alleles, which enhances the average mutation rate, leading to degradation. Then the possible solution lies in a population bottleneck, i.e. the population must be periodically refreshed by establishing subpopulations, having decreased (below the average) mutation rate. Presumably, such subpopulation will quickly take advantage and overcome the parent population. This phenomenon can be called "genetic ageing of the population", in a sense, it resembles somatic ageing, however, occurs on the level of population. From this viewpoint bottlenecks, reproduction barriers and speciation events are necessary for progress and thus are inevitable companions of evolution, rather than peculiar accidental features. It is worth repeating that positive mutations are abundant in the framework of the model, therefore the role of population size and slow speed of evolution is not as significant as in classical models.

Breeders are well aware of the phenomenon that adaptation to new demands of selection occurs at a price to reduction of adaptation to other demands. In the framework of the presented model this is expressed in reshaping the genomic GI profile (which results in phenotype changes) while the total GI amount remains constant. From the biological point of view this phenomenon represents a directional decrease of variability (reflected in the model as the growth of corresponding GI) of one phenotypic characteristic, while the variability of the others increases (GI drops down). Except for some analogs of the “error threshold” (which are usually applicable only in some ad hoc cases), the notion of channel capacity is absent in conventional models. Therefore fitness functions traditionally used are relative and thus incapable of distinguishing among such “reshaping” and “progressive” modes of evolution. However, the suggested IT framework is sufficiently general allowing for direct modeling of such “reshaping selection”, evaluating its basic features and impacts of different evolutionary strategies. I expect that such evolutionary plasticity is to a greater extent inherent in functional sequences possessing low density of GI (e.g. most of eukaryotic noncoding functional sequences).

Despite its conceptual and practical importance, IT remains poorly known outside the communication engineering community. However, nature has to solve many engineering tasks in the course of evolution and many living processes (e.g. transmission of genetic information to next generation) seem to be exclusively relevant to literal communication, making an adaptation of IT to biology natural and highly promising (Battail 2013). Other things being equal, an optimization of GI storage makes species more efficient, providing higher informational capacity of the genome, increasing reliability of hereditary information transfer, and decreasing genetic load which eventually implies better survival rates. Natural selection leads to the “survival of the fittest”, which is equivalent to the survival of the most efficient, naturally including the efficiency of information processing. IT shows that in order to achieve better efficiency of information transmission we should apply more sophisticated algorithms of decoding and encoding. So in general, increasing memory and computational complexity is inevitable if we want to move closer to the channel capacity limit. Hence the IT provides a natural link between the drive to efficiency and the drive to complexity. While the former is usually considered as self-evident the comprehension of the latter presents considerable difficulties. Traditionally, biological complexity is assumed to passively emerge as simple rules (interactions), applied recursively. Such schemes are able to generate perceivably complex patterns, but these patterns remain simple algorithmically. In contrast, the lesson learned from IT is that an active drive to increase algorithmic complexity is necessary in order to be efficient. In this regard an instructive example is the “evolution” of IT itself. As it was mentioned in the introduction, the role of the IT among scientific and engineering communities was insignificant until the tough demand in energy-efficient communication appeared due to the beginning of the space race (Aftab et al. 2001). This demand boosted theoretical and practical developments, promoting the invention of complex hardware and algorithms allowing us to approach the channel capacity limit, and resulting in readily accessible various worldwide digital communication media. It is tempting to speculate that similar evolution with a drive to complexity occurs on the level of "molecular machines".

Invariants are the cornerstones of physical and mathematical theories. However, no invariants were proposed in the theory of evolution. Filling this gap, the invariance of a GI pattern in an equilibrium population is proposed in this work. Having this invariant and complementing it with a few reasonable assumptions I come up with a model of molecular evolution, which, to a wide extent, represents a completely novel conception. The use of Shannon's IT as a mathematical foundation for the model provides it with high level of abstraction and generality. The model allows to look at many evolutionary phenomena from a fresh perspective. New interpretations of some well-known phenomena have been already introduced in this thesis. Further theoretical elaboration and experimental verification of the model will promote a deeper understanding of molecular evolution and population genetics processes.

BIBLIOGRAPHY

Adami C (2004) Information Theory in Molecular Biology. arXiv:q-bio/0405004 [q-bio.BM].

Adami C and Cerf NJ (2000) Physical complexity of symbolic sequences. *Physica D. Nonlinear phenomena*, 137:62-69.

Adami C, Ofria C, Collier TC (2000) Evolution of biological complexity. *Proceedings of the National Academy of Sciences of the United States of America* 97(9):4463-4468.

Aftab O, Cheung P, Kim A, Thakkar S, Yeddanapudi N (2001) Information Theory and the Digital Age. 6.933—Final Paper, The Structure of Engineering Revolutions, Massachusetts Institute of Technology.

Ahituv N et al. (2007) Deletion of ultraconserved elements yields viable mice. *PLoS biology*, 5(9):e234.

Akashi H (1995) Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA. *Genetics*, 139(2):1067-76.

Anderson JP, Daifuku R, Loeb LA (2004) Viral error catastrophe by mutagenic nucleosides. *Annual review of microbiology*, 58:183-205.

Battail G (2013) Biology Needs Information Theory. *Biosemiotics*. 6(1):77-103.

Berg OG and von Hippel PH (1987) Selection of DNA Binding Sites by Regulatory Proteins. Statistical-Mechanical Theory and Application to Operators and Promoters. *Journal of molecular biology*, 193(4):723-750.

Berg J, Willmann S, Lässig M (2004) Adaptive Evolution of Transcription Factor Binding Sites. *BMC evolutionary biology*, 4:42.

Bornberg-Bauer E and Chan HS (1999) Modeling evolutionary landscapes: mutational stability, topology, and superfunnels in sequence space. *Proceedings of the National Academy of Sciences of the United States of America*, 96(19):10689-94.

Bull JJ, Sanjuán R, Wilke CO (2007) Theory of Lethal Mutagenesis for Viruses. *Journal of virology*, 81(6): 2930-2939.

Campos PRA and Fontanari JF (1999) Finite-size scaling of the error threshold transition in finite populations. *Journal of physics A: Mathematical and general*, 32 L1.

Chaitin G (2012) Proving Darwin: Making Biology Mathematical. Pantheon Books, New York.

Charlesworth D (2006) Balancing Selection and Its Effects on Sequences in Nearby Genome Regions. *PLoS Genetics,* 2(4):e64.

Chiu DK and Kolodziejczak T (1991) Inferring consensus structure from nucleic acid sequences. *Computer applications in the biosciences : CABIOS*, 7(3):347-52.

Collins DW and Jukes TH (1994) Rates of Transition and Transversion in Coding Sequences Since the Human-Rodent Divergence. *Genomics*, 20(3): 386-396.

Comas I, Moya A and Gonzalez-Candelas F (2005) Validating viral quasispecies with digital organisms: A re-examination of the critical mutation rate. *BMC evolutionary biology*, 5:5.

Cover TM and Thomas JA (2006) Elements of information theory, Second Edition. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Crotty S, Cameron CE and Andino R (2001) RNA virus error catastrophe: Direct molecular test by using ribavirin. *Proceedings of the National Academy of Sciences of the United States of America*, 98:6895-6900.

Crow JF (1958) Some Possibilities for Measuring Selection Intensities in Man. *Human Biology*, 30(1):1-13.

Crow JF (1986) Basic concepts in population, quantitative, and evolutionary genetics. W.H. Freeman, New York, p. 273.

Crow JF (2001) Shannon's brief foray into genetics. *Genetics*, 159(3): 915-917.

Dawkins R (2004) The Ancestor's Tale, A Pilgrimage to the Dawn of Life. Houghton Mifflin Company, Boston, p. 416.

Desai MM, Fisher DS (2007) Beneficial mutation-selection balance and the effect of linkage on positive selection. Genetics 176(3):1759-1798.

Dietrich MR (1994) The origins of the neutral theory of molecular evolution. *Journal of the history of biology*, 27(1):21-59.

Domingo E, Sabo D, Taniguchi T, Weissmann C (1978) Nucleotide sequence heterogeneity of an RNA phage population. *Cell*, 13(4):735-44.

Domingo E, Martínez-Salas E, Sobrino F, de la Torre JC, Portela A et al. (1985) The quasispecies (extremely heterogeneous) nature of viral RNA genome populations: biological relevance--a review. *Gene*, 40(1):1-8.

Domingo E (1992) Genetic variation and quasi-species. *Current opinion in genetics & development*, 2(1):61-3.

Domingo E, Biebricher CK, Eigen M, Holland JJ (2001) Quasispecies and RNA Virus Evolution: Principles and Consequences Georgetown. Landes Bioscience, Georgetown, TX, USA.

Drake JW (1991) A constant rate of spontaneous mutation in DNA-based microbes. *Proceedings of the National Academy of Sciences of the United States of America*, 88:7160-7164.

Drake JW, Charlesworth B, Charlesworth D and Crow JF (1998) Rates of Spontaneous Mutation. *Genetics*, 148:1667-1686.

Drake JW and Holland JJ (1999) Mutation rates among RNA viruses. *Proceedings of the National Academy of Sciences of the United States of America*, 96(24):13910-3.

Durrett R (2008) Probability Models for DNA Sequence Evolution. Springer, New York.

Eckerle LD, Lu X, Sperry SM, Choi L, Denison MR (2007) High fidelity of murine hepatitis virus replication is decreased in nsp14 exoribonuclease mutants. *Journal of Virology* 81(22):12135-12144.

Eigen M (1971) Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften*, 58:456-523.

Eigen M and Schuster P (1977) The hypercycle. A principle of natural self-organization. Part A: Emergence of the hypercycle. *Naturwissenschaften*, 64(11):541-65.

Eigen M (1996) On the nature of virus quasispecies. *Trends in microbiology*, 4:216-218.

Elgar G and Vavouri T (2008) Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. *Trends in genetics : TIG*, 24(7):344-52.

Erill I and O'Neill MC (2009) A reexamination of information theory-based methods for DNA-binding site identification. *BMC Bioinformatics*, 10:57.

Fabris F (2008) Shannon information theory and molecular biology. *Journal of Interdisciplinary Mathematics*, 12(1): 41-87.

Feverati G and Musso F (2008) Evolutionary model with Turing machines. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 77(6 Pt 1):061901.

Fisher RA and Ford EB (1950) The ‘Sewall Wright’ effect. *Heredity*, 4:117-119.

Flicek P, Amode MR, Barrell D, Beal K, Brent S et al. (2012) Ensembl 2012. *Nucleic Acids Research*, 40(Database issue):D84-90.

Forns X, Purcell RH, Bukh J (1999) Quasispecies in viral persistence and pathogenesis of hepatitis C virus. *Trends in microbiology*, 7(10):402-10.

Forsdyke DR (2011) Evolutionary bioinformatics. Springer, New York.

Frank SA (2012) Natural Selection. V. How to read the fundamental equations of evolutionary change in terms of information theory. *J Evol Biol* 25(12):2377-2396.

Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D et al. (2011) The UCSC Genome Browser database: update 2011. *Nucleic Acids Research*, 39(Database issue):D876-82.

Gabriel W, Lynch M, Burger R (1993) Muller's ratchet and mutational meltdowns. *Evolution*, 47:1744-1757.

Gatlin LL (1966) The information content of DNA. *Journal of theoretical biology*, 10:281-300.

Gatlin LL (1968) The information content of DNA, II. *Journal of theoretical biology*, 18:181-194.

Gibrat JF, Garnier J and Robson B (1987) Further developments of protein secondary structure prediction using information theory. New parameters and consideration of residue pairs. *Journal of molecular biology*, 198(3):425-43.

Girardin V (2005) On the different extensions of the ergodic theorem of information theory. In “Recent advances in applied probability”, eds Baeza-Yates R, Glaz J, Gzyl H, Hüsler J and Palacios JL. Springer, New York, pp. 163-179.

Grosse I, Herzel H, Buldyrev SV, Stanley HE (2000) Species independence of mutual information in coding and noncoding DNA. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 61(5 Pt B):5624-9.

Gutell RR, Power A, Hertz GZ, Putz EJ, Stormo GD (1992) Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nucleic Acids Research*, 20(21):5785-95.

Hahn MW (2008) Toward a selection theory of molecular evolution. *Evolution*, 62(2):255-65.

Haigh J (1978) Accumulation of deleterious mutations: Muller's ratchet. *Theoretical population biology*, 14:251-267.

Haldane JBS (1957) The Cost of Natural Selection. *Journal of Genetics*, 55:511-524.

Hamacher K (2007) Information theoretical measures to analyze trajectories in rational molecular design. *Journal of computational chemistry,* 28(16):2576-80.

Hamilton MB (2009) Population Genetics. Wiley-Blackwell, Hoboken, NJ, USA, p. 54-55.

Hardy GH (2003) Mendelian proportions in a mixed population. 1908. *Yale Journal of Biology and Medicine*, 76:79-80.

Hey J (1999) The neutralist, the fly and the selectionist. *Trends in ecology & evolution*, 14(1):35-38.

Hoffmann GW (1994) Co-selection in immune network theory and in AIDS pathogenesis. *Immunology and cell biology,* 72(4):338-46.

Holmes EC and Moya A (2002) Is the Quasispecies Concept Relevant to RNA Viruses? *Journal of virology*, 76:460-462.

Houlston R (2006) Mutations: Penetrance. General & Introductory Life Sciences. Online.

Hughes AL (2007) Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. *Heredity,* 99(4):364-73.

Hughes AL (2008) Near neutrality: leading edge of the neutral theory of molecular evolution. *Ann N Y Acad Sci*. 1133:162-79.

Jenkins GM, Worobey M, Woelk CH, Holmes EC (2001) Evidence for the non-quasispecies evolution of RNA viruses. *Molecular biology and evolution*, 18:987-994.

Johnson HA (1970) Information Theory in Biology after 18 Years. *Science,* 168(3939):1545-1550.

Kamal M, Xie X, Lander ES (2006) A Large Family of Ancient Repeat Elements in the Human Genome is under Strong Selection. *Proceedings of the National Academy of Sciences of the United States of America*, 103(8):2740-2745.

Kasprzyk A (2011) BioMart: driving a paradigm change in biological data management. *Database : the journal of biological databases and curation*, 2011:bar049.

Kimura M (1968) Evolutionary Rate at the Molecular Level. *Nature*. 217:624-6.

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, New York.

Kimura M and Ohta T (1974) On some principles governing molecular evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 71:2848-2852.

King JL and Jukes TH (1969) Non-Darwinian Evolution. *Science*, 164:788-97.

Kohne DE (1970) Evolution of higher-organism DNA. *Quarterly reviews of biophysics*, 3:327-375.

Koonin EV and Novozhilov AS (2009) Origin and evolution of the genetic code: the universal enigma. *IUBMB Life*, 61(2):99-111.

Kreitman M and Akashi H (1995) Molecular evidence for natural selection. *Annual Review of Ecology and Systematics*, 26:403-422.

Kreitman M (1996) The neutral theory is dead. Long live the neutral theory. *Bioessays*, 18(8):678-83.

Kumar V, Westra HJ, Karjalainen J, Zhernakova DV, Esko T et al. (2013) Human disease-associated genetic variation impacts large intergenic non-coding RNA expression. *PLoS Genetics*, 9(1):e1003201.

Lauber C, Goeman JJ, Parquet Mdel C, Nga PT, Snijder EJ, Morita K, Gorbalenya AE (2013) The footprint of genome architecture in the largest genome expansion in RNA viruses. *PLoS Pathogens* 9(7):e1003500.

Levene H (1953) Genetic Equilibrium When More than One Ecological Niche Is Available. *The American Naturalist*, 87(836):331-333.

Lewontin RC (1974) The genetic basis of evolutionary change. Columbia University Press, New York, USA.

Li W-H and Graur D (1991) Fundamentals of Molecular Evolution. Sinauer Associates, Sunderland, Massachusetts.

Lynch M, Burger R, Butcher D, Gabriel W (1993) The mutational meltdown in asexual populations. *The Journal of heredity*, 84:339-344.

Lynch M (2010a) Evolution of the mutation rate. *Trends in genetics: TIG*, 26(8):345-52.

Lynch M (2010b) Rate, molecular spectrum, and consequences of human mutation. *Proceedings of the National Academy of Sciences of the United States of America*, 107(3):961-8.

Margoliash E (1963) Primary structure and evolution of cytochrome c. *Proceedings of the National Academy of Sciences of the United States of America*, 50(4): 672-679.

Massey SE (2013) Proteome size as the major factor determining mutation rates. *Proceedings of the National Academy of Sciences of the United States of America*, 110(10):E858-9.

McDonald JH and Kreitman M (1991) Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, 351(6328):652-4.

Moran PAP (1962) The statistical processes of evolutionary theory. Clarendon Press, Oxford.

Moya A, Elena SF, Bracho A, Miralles R, Barrio E (2000) The evolution of RNA viruses: A population genetics view. *Proceedings of the National Academy of Sciences of the United States of America*, 97(13):6967–6973.

Nei M (2005) Selectionism and Neutralism in Molecular Evolution. *Molecular biology and evolution*, 22(12): 2318-2342.

Nga PT, Parquet Mdel C, Lauber C, Parida M, Nabeshima T, Yu F, Thuy NT, Inoue S, Ito T, Okamoto K, Ichinose A, Snijder EJ, Morita K, Gorbalenya AE (2011) Discovery of the first insect nidovirus, a missing evolutionary link in the emergence of the largest RNAvirus genomes. *PLoS Pathogens* 7(9):e1002215.

Nóbrega MA, Zhu Y, Plajzer-Frick I, Afzal V, Rubin EM (2004) Megabase deletions of gene deserts result in viable mice. *Nature*, 431(7011):988-93.

Nowak M and Schuster P (1989) Error Thresholds of Replication in Finite Populations–Mutation Frequencies and the Onset of Muller's Ratchet. *Journal of theoretical biology*, 137:375-395.

Nowak MA (1992) What is a quasispecies? *Trends in ecology & evolution*, 7(4):118-21.

Ofria C, Huang W, Torng E (2008) On the gradual evolution of complexity and the sudden emergence of complex features. *Artif Life* 14(3):255-263.

Ohno S (1970) Evolution by gene duplication. Springer, New York.

Ohno S (1972) So much ‘‘junk’’ DNA in our genome. *Brookhaven Symposia in Biology*, 23:366-370.

Ohta T (1973) Slightly Deleterious Mutant Substitutions in Evolution. *Nature*, 246:96-98.

Ohta T (1976) Role of very slightly deleterious mutations in molecular evolution and polymorphism. *Theoretical population biology*, 10(3):254-75.

Ohta T and Gillespie JH (1996) Development of Neutral and Nearly Neutral Theories. *Theoretical population biology*, 49(2):128-42.

Pariente N, Sierra S, Airaksinen A (2005) Action of mutagenic agents and antiviral inhibitors on foot-and-mouth disease virus. *Virus research*, 107(2):183-93.

Park JM, Muñoz E, Deem MW (2010) Quasispecies theory for finite populations. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 81(1 Pt 1):011902.

Pierce JR (1980) An Introduction to Information Theory: Symbols, Signals and Noise. Second revised edition. Dover Publications, New York, USA.

Ponting CP and Hardison RC (2011) What fraction of the human genome is functional? *Genome Research*, 21(11):1769-76.

Postlethwait JH (2009) Modern Biology. Holt, Rinehart and Winston.

Runarsson TP and Yao X (2000) Stochastic Ranking for Constrained Evolutionary Optimization. *IEEE Transactions on Evolutionary Computation*, 4(3):284-294.

Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research*, 32(Database issue):D91-4.

Sanjuan R, Moya A, Elena SF (2004) The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. *Proceedings of the National Academy of Sciences of the United States of America*, 101(22):8396-401.

Schneider TD, Stormo GD, Gold L, Ehrenfeucht A (1986) Information content of binding sites on nucleotide sequences. *Journal of molecular biology*, 188(3):415-431.

Schneider TD and Stephens RM (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research*, 18(20): 6097-6100.

Schneider TD (2000) Evolution of biological information. *Nucleic Acids Research*, 28(14):2794-9.

Schuster P and Swetina J (1988) Stationary mutant distributions and evolutionary optimization. *Bulletin of mathematical biology*, 50:635-660.

Shadrin AA, Grigoriev A, Parkhomchuk DV (2013) Positional information storage in sequence patterns. *Computational Molecular Bioscience*, 3(2):18-26.

Shadrin AA and Parkhomchuk DV (2014) Drake's rule as a consequence of approaching channel capacity. *Naturwissenschaften,* [Epub ahead of print].

Shannon CE (1948) A mathematical theory of communication. *Bell System Technical Journal*, 27:379-423, 623-656.

Shannon CE and Weaver W (1949) The Mathematical Theory of Communication. University of Illinois Press, Urbana, Illinois.

Smith NG, Brandström M, Ellegren H (2004) Evidence for turnover of functional noncoding DNA in mammalian genome evolution. *Genomics*, 84(5):806-13.

Sniegowski PD, Gerrish PJ (2010) Beneficial mutations and the dynamics of adaptation in asexual populations. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365(1544):1255-1263.

Sniegowski P and Raynes Y (2013) Mutation rates: how low can you go? *Current biology*, 23(4):R147-9.

Steinhauer DA, de la Torre JC, Meier E, Holland JJ (1989) Extreme heterogeneity in populations of vesicular stomatitis virus. *Journal of virology*, 63(5):2072-80.

Stephens RM and Schneider TD (1992) Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites. *Journal of molecular biology,* 228(4):1124-36.

Stormo GD (2000) DNA Binding Sites: Representation and Discovery. *Bioinformatics*, 16(1):16-23.

Strelioff CC, Lenski RE, Ofria C (2010) Evolutionary dynamics, epistatic interactions, and biological information. *Journal of Theoretical Biology* 266(4):584-594.

Summers J and Litwin S (2006) Examining the theory of error catastrophe. *Journal of virology*, 80(1):20-6.

Sung W, Ackerman MS, Miller SF, Doak TG, Lynch M (2012) Drift-barrier hypothesis and mutation-rate evolution. *Proceedings of the National Academy of Sciences of the United States of America,* 109(45):18488-92.

Swetina J and Schuster P (1982) Self-replication with errors. A model for polynucleotide replication. *Biophysical chemistry,* 16(4):329-45.

Takahata N (1987) On the overdispersed molecular clock. *Genetics*, 116(1):169-79.

Tannenbaum E, Deeds EJ, Shakhnovich EI (2003) Equilibrium distribution of mutators in the single fitness peak model. *Physical review letters,* 91(13):138105.

Tarn WY and Steitz JA (1996) A novel spliceosome containing U11, U12, and U5 snRNPs excises a minor class (AT-AC) intron in vitro. *Cell*, 84(5):801-11.

The ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414): 57-74.

Thompson MJ and Goldstein RA (1997) Predicting protein secondary structure with probabilistic schemata of evolutionarily derived information. *Protein science : a publication of the Protein Society*, 6(9):1963-75.

van Nimwegen E, Crutchfield JP, Huynen M (1999a) Neutral Evolution of Mutational Robustness. *Proceedings of the National Academy of Sciences of the United States of America* 96:9716-9720.

van Nimwegen E, Crutchfield JP, Mitchell M (1999b) Statistical Dynamics of the Royal Road Genetic Algorithm. *Theoretical Computer Science*. 229:41-102.

von Hippel PH and Berg OG (1986) On the Specificity of DNA-Protein Interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 83(6):1608-1612.

Wagner A (2008) Neutralism and selectionism: a network-based reconciliation. *Nature reviews. Genetics*, 9:965-974.

Wagner GP and Krall P (1993) What is the difference between models of error thresholds and Muller's ratchet? *Journal of Mathematical Biology*, 32(1):33-44.

Wilke CO (2001) Adaptive evolution on neutral networks. *Bulletin of mathematical biology*, 63(4):715-30.

Wilke CO, Wang JL, Ofria C, Lenski RE, Adami C (2001) Evolution of digital organisms at high mutation rate leads to survival of the flattest. *Nature*, 412:331-333.

Wilke CO and Adami C (2003) Evolution of mutational robustness. *Mutation research*, 522(1-2):3-11.

Wilke CO (2005) Quasispecies theory in the context of population genetics. *BMC evolutionary biology*, 5:44.

Wright S (1931) Evolution in Mendelian populations. *Genetics*, 16:97-159.

Wright S (1932) The roles of mutation, inbreeding, crossbreeding, and selection in evolution. *Proceedings of the Sixth International Congress of Genetics*, 1:356-366.

Yockey HP (2005) Information theory, evolution, and the origin of life. Cambridge University Press, New York.

Yuan D, Zhu Z, Tan X, Liang J, Zeng C, Zhang J, Chen J, Ma L, Dogan A, Brockmann G et al. (2013) Methods for scoring the collective effect of SNPs: Minor alleles of common SNPs quantitatively affect traits/diseases and are under both positive and negative selection. arXiv:1209.2911v2 [q-bio.GN]

Zhao Z, Fu YX, Hewett-Emmett D, Boerwinkle E (2003) Investigating single nucleotide polymorphism (SNP) density in the human genome and its implications for molecular evolution. *Gene*, 312:207-13.

Zuckerkandl E and Pauling LB (1962) Molecular disease, evolution, and genic heterogeneity. In "Horizons in Biochemistry", eds. Kasha M and Pullman B. Academic Press, New York, pp. 189-225.

APPENDIX

Zusammenfassung

Invarianten (Erhaltungssätze) dienten mathematischen und physikalischen Theorien als grundlegende Eckpfeiler, von der Frühzeit der Wissenschaft bis in die Neuzeit. So war beispielsweise die erste Bezeichnung für Einsteins Theorie „Invariantentheorie“ und Klein erachtete die Geometrie in seinem „Erlanger Programm“ als das Studium von Invarianten unter einer Transformationsgruppe. In den Theorien der molekularen Evolution hingegen wird die vielfach beobachtete Invarianz des Phänotyps, d. h. sein Erhalt über Generationen hinweg, nicht mit invarianten Genomsequenzen gleichgesetzt. Im Gegenteil, die Genomsequenzen werden als recht veränderlich betrachtet; sie entwickeln sich schnell und opportunistisch, oftmals „neutral“. Die klassischen Modelle der molekularen Evolution wurden vor mehr als 40 Jahren entwickelt, wobei damals keine umfassenden Datenmengen zur Verfügung standen. Die folgende Entwicklung der Theorie der molekularen Evolution war zunächst willkürlich und oberflächlich: unwesentliche Ad-hoc-Annahmen wurden eingeführt, um neu gewonnenen Daten zu entsprechen. Der Kern dieser Modelle blieb jedoch unverändert. Die Konzepte wurden mit mehr Details und Annahmen weiter ausgeführt, wodurch sie kompliziert wurden und die Fähigkeit verloren, nachweisbare Vorhersagen oder Erklärungen zu beobachtbaren Phänomenen abzugeben. Das Fehlen allgemeiner Grundprinzipien führte zur Krise der Theorie der molekularen Evolution. Heutige Technologien versorgen uns mit einer Unmenge an molekularen Daten, was einen tieferen Einblick in die Funktionsweise von Genomen ermöglicht und ein tiefgehenderes Verständnis der Funktionsweise von Genomen erfordert.

Diese Arbeit führt ein neues Paradigma in die Theorie der molekularen Evolution ein, indem eine invariante Eigenschaft der Genomsequenz eingebracht wird, die sich nicht oder nur langsam von Generation zu Generation ändert, während sich die Grundsequenzen schnell ändern können. Die Einführung der Invariante führt zu einer eher „physikalischen“ und weniger opportunistischen Sicht auf die Sequenzevolution und liefert prüfbare Vorhersagen. Das weit entwickelte System aus Shannons Informationstheorie wird als mathematischer Rahmen des Modells verwendet. Ein funktioneller Ort wird als ein positionell wahrscheinliches „Pattern“ betrachtet, wo jede Position des „Patterns“ einen Vierervektor von Nukleotidwahrscheinlichkeiten in der Gleichgewichtspopulation (d. h. abstrakte unendliche Population, die sich über einen unbegrenzten Zeitraum ohne störende Ereignisse entwickelt hat) darstellt. Die Einführung der Invariante ermöglicht uns die Simulation der Geninformationsdynamiken und die Anwendung grundlegender physikalischer

Prinzipien, wie die optimale Effizienz und Kanalkapazität. Das Modell beweist die grundsätzliche Möglichkeit einer fehlerfreien Informationsspeicherung in Sequenzen, deren Erhaltung willkürlich gering ist. Ich beweise, dass die Rate vorteilhafter Mutationen im Allgemeinen hoch sein kann. Je geringer die Sequenzerhaltung, desto höher die Frequenz der vorteilhaften Mutationen. Die Versuchsergebnisse zeigen die Tendenz wirklich funktioneller Orte zur Optimierung, in Übereinstimmung mit dem eingebrachten Optimalitätskriterium. Das Modell ermöglicht einen frischen Blick auf das wohlbekannte Phänomen (es zeigt beispielsweise, dass die „Molekulare Uhr" und „Drake's Rule" möglicherweise aus einem gemeinsamen Prozess heraus entstehen). Es kann ebenfalls sinnvolle Erklärungen für einige Paradoxa (z. B. „Paradox of Variation") liefern, denen es im Rahmen klassischer Theorien an einer eindeutigen Interpretation mangelt. Daher glaube ich, dass die Weiterentwicklung des Modells ein tieferes Verständnis der molekularen Evolution und populationsgenetischer Prozesse vermitteln wird.