PLOS ONE

RESEARCH ARTICLE

Optimization of Mutation Pressure in Relation to Properties of Protein-Coding Sequences in Bacterial Genomes

Paweł Błażej[1], Błażej Miasojedow[2], Małgorzata Grabińska[1], Paweł Mackiewicz[1]*

1 Department of Genomics, Faculty of Biotechnology, University of Wrocław, Wrocław, Poland, **2** Section of Mathematical Statistics, The Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Warszawa, Poland

* pamac@smorfland.uni.wroc.pl

Citation: Błażej P, Miasojedow B, Grabińska M, Mackiewicz P (2015) Optimization of Mutation Pressure in Relation to Properties of Protein-Coding Sequences in Bacterial Genomes. PLoS ONE 10(6): e0130411. doi:10.1371/journal.pone.0130411

Editor: Igor B. Rogozin, National Center for Biotechnology Information, UNITED STATES

Received: October 21, 2014

Accepted: May 19, 2015

Published: June 29, 2015

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The publication fee was funded by the KNOW Consortium.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Most mutations are deleterious and require energetically costly repairs. Therefore, it seems that any minimization of mutation rate is beneficial. On the other hand, mutations generate genetic diversity indispensable for evolution and adaptation of organisms to changing environmental conditions. Thus, it is expected that a spontaneous mutational pressure should be an optimal compromise between these two extremes. In order to study the optimization of the pressure, we compared mutational transition probability matrices from bacterial genomes with artificial matrices fulfilling the same general features as the real ones, e.g., the stationary distribution and the speed of convergence to the stationarity. The artificial matrices were optimized on real protein-coding sequences based on Evolutionary Strategies approach to minimize or maximize the probability of non-synonymous substitutions and costs of amino acid replacements depending on their physicochemical properties. The results show that the empirical matrices have a tendency to minimize the effects of mutations rather than maximize their costs on the amino acid level. They were also similar to the optimized artificial matrices in the nucleotide substitution pattern, especially the high transitions/transversions ratio. We observed no substantial differences between the effects of mutational matrices on protein-coding sequences in genomes under study in respect of differently replicated DNA strands, mutational cost types and properties of the referenced artificial matrices. The findings indicate that the empirical mutational matrices are rather adapted to minimize mutational costs in the studied organisms in comparison to other matrices with similar mathematical constraints.

Introduction

Mutations occurring in DNA sequences are an inherent component of biological evolution. They, together with recombinations, generate a genetic variation which is subsequently subjected to selection. The most frequent mutations are substitutions, i.e., single nucleotide changes, which may be spontaneous, induced by radiation or chemicals, or introduced during

the replication and repair of DNA. One of the most evident effects of mutations that arise during replication is DNA asymmetry. It manifests itself in different nucleotide and codon compositions of the diversely replicated DNA strands, called leading and lagging. This effect comes from various synthesis mechanisms of the DNA strands [1, 2], and is observed in most bacterial genomes, [3–11]. It has also significant consequences on various divergence rates of the genes located on the differently replicated DNA strands [12–17] as well as stability of their positions [18–20] and distribution on chromosomes [21, 22].

In practice, however, it is very hard to detect the effect of spontaneous mutations because many of them are eliminated by selection, especially those that happen in protein-coding sequences. Deleterious changes, such as nonsense mutations, which generate premature stop codons can lead to truncation of protein sequences. The harmfulness of other mutations, called missense, which change one coded amino acid to another depend on the differences in physicochemical properties of the substituted amino acids. The more these amino acids differ, e.g., in size, charge or hydrophobicity, the more harmful their replacement. Since many amino acids are encoded by two, three, four or six different codons, called synonymous, there are some mutations, called silent or synonymous, that do not change encoded amino acids and, consequently, protein composition and structure.

Mutations occurring in biological DNA sequences are the result of coevolution between mutational pressure and selection constraints around the genetic code [23–25], and can be optimized to some extent during evolution; see for review: [26, 27]. On the one hand, we should expect a tendency "for selection" to decrease the mutation rate because most mutations are deleterious and generate energetically costly repairing [28, 29]. On the other hand, mutations are responsible for genetic diversity, which is necessary for the adaptation of organisms to changing environments on the evolutionary scale. Therefore, an elevated level of mutation rate should be also expected in these cases [30–32]. This trade-off between the necessity to preserve accurate genetic information and requirements for adaptational flexibility indicates that some optimal mutation pressure can evolve [27, 33, 34]; however, this may depend on fitness landscape [35] and population structure [36].

The selection can operate to refine DNA replication and repair [37–43], which can influence the global mutation rate in organisms. An improvement of fidelity of DNA polymerases in DNA synthesis as well as effectiveness of their proofreading properties and post-replicative DNA mismatch repair mechanisms would decrease the general mutation rate. Otherwise, the rate would increase. The pattern of nucleotide substitution, i.e. relative rates of change from one type of nucleotide to another, can be also subjected to the optimization. For example, transitions, i.e., substitutions for the same chemical type of nucleotides, purine for purine or pyrimidine for pyrimidine, often cause fewer changes in coded amino acids or their properties than transversions, i.e., substitutions for different type of nucleotides, purine for pyrimidine and *vice versa*. Therefore, we can expect that a higher transitions/transversions ratio will be favored in mutational pressures.

The problem of optimization related to mutations has been studied in the context of the genetic code origin and its evolution [24, 25, 44–47]. It was postulated that many more assignments of codons to amino acids existed at the dawn of life on Earth; however, they were lost because they did not effectively minimize harmful effects of mutations on protein-coding sequences and translation errors. Optimization of codon usage has also been analyzed in terms of the reduction of deleterious mutational effects [46, 48, 49].

Since a significant fraction, usually more than 90% of bacterial genomes constitute protein-coding sequences [50], it is worthwhile to study the optimization of mutation pressure in respect to proteins. In this approach, we studied whether empirical mutational pressures expressed by transition probability matrices for particular bacterial genomes are better

optimized to protein-coding sequences than other such types of matrices characterized by the same stationary distribution and the convergence speed to the stationarity. The optimization was considered according to probability of non-synonymous substitutions and different costs of amino acid substitutions occurring in products of protein-coding sequences.

Materials and Methods

Mutational transition probability matrices

We tested empirical mutational pressures described by transition probability matrices which were detected in six bacterial genomes by [15, 51, 52]. It is important in our studies to the matrices reflect neutral mutations in the absence of selection. To achieve that as much as possible, the authors in the inferring these matrices made a big effort to eliminate the potential influence of selection. The matrix for *Borrelia burgdorferi* was obtained by comparison of gene sequences with their potential pseudogenes found in intergenic regions [51], whereas matrices for *Escherichia coli*, *Chlamydia muridarum*, *Chlamydia trachomatis*, *Rickettsia*, *Staphylococcus aureus* and *Streptococcus pyogenes* genomes were inferred from comparison of synonymous sites in orthologous genes from closely related species or strains [15, 52]. However, there is a subset of highly expressed genes in which also synonymous substitutions are subjected to selection because of some preferences in codon usage, which is positively correlated with tRNA content in cells and the rate of translation [53–58]. Therefore, the authors removed the top 10% genes with most biased codon usage, expected to be the most highly expressed, to obtain the sites subjected to neutral substitutions. In our studies, we considered matrices for differently replicated DNA strand, leading and lagging separately, because they are subjected to various mutational patterns.

The matrices determine a unique homogeneous Markov chain, which characterizes the process of nucleotide substitutions and converges to the stationary distribution (Table 1). We used basic concepts of linear algebra and theory of Markov processes to investigate the properties of empirical matrices and define a class of artificial transition probability matrices M with similar properties. The artificial matrices were used as a reference to the empirical ones. The mathematical properties of the matrices are related to the stationary distribution and spectral decomposition of transition probability matrix [59–61].

It is well known from linear algebra and theory of Markov processes that every finite positive transition probability matrix P has a unique spectral decomposition

$$P = \mathrm{A}\Lambda\mathrm{A}^{-1}, \tag{1}$$

where: A and A^{-1} are matrices whose rows/columns consist of right/left eigenvectors, respectively; Λ is a diagonal matrix with eigenvalues on its diagonal. The eigenvalues of the matrices are the solution of the characteristic equation. Therefore, there are four eigenvalues and four right/left eigenvectors for the matrix P. In general, some of the eigenvalues could be complex, not real numbers. The stationary distribution of the Markov process π is the left eigenvector (i.e., $\pi P = \pi$) and corresponds to the maximum of eigenvalues, which is always equal to 1 in the case of transition probability matrix. The second largest eigenvalue is responsible for the speed of convergence of Markov process to the stationary distribution, generated by P, which is a direct consequence of the Perron-Frobenius theorem [59–61]) Using these properties, we took into account the class M of transition probability matrices $P = (p_{ij})$, $i, j \in \{A,T,G,C\}$, where p_{ij} denotes the probability of substitution from a nucleotide i to a nucleotide j, and A, T, G, C are nucleotides: adenine, thymine, guanine and cytosine, respectively. Each matrix $P \in M$ is

Table 1. Nucleotide stationary distribution of leading and lagging strand matrices for studied genomes.

Genome	Leading strand				Lagging strand			
	A	T	G	C	A	T	G	C
Borrelia burgdorferi	0.32	0.49	0.14	0.06	0.49	0.32	0.06	0.14
Chlamydia muridarum	0.24	0.25	0.28	0.22	0.22	0.23	0.29	0.26
Chlamydia trachomatis	0.23	0.21	0.29	0.26	0.25	0.25	0.25	0.24
Escherichia coli	0.25	0.33	0.25	0.18	0.27	0.31	0.21	0.22
Rickettsia species	0.30	0.31	0.21	0.19	0.33	0.27	0.24	0.16
Staphylococcus aureus	0.41	0.39	0.12	0.08	0.35	0.45	0.09	0.11
Streptococcus pyogenes	0.33	0.42	0.12	0.13	0.30	0.40	0.09	0.20

doi:10.1371/journal.pone.0130411.t001

expressed by an equation:

$$P = \mathrm{A}\Lambda\mathrm{A}^{T}\Pi, \tag{2}$$

where A is a real valued orthogonal matrix. Λ is a diagonal matrix with fixed the first and the second eigenvalue. Π is a diagonal matrix with the empirical stationary distribution $\pi = \{\pi_A, \pi_T, \pi_G, \pi_C\}$ on its diagonal. $\mathrm{A}^{-1} = \mathrm{A}^{T}\Pi$, which means that A is orthogonal in terms of stationary distribution. The Eq (2) is the special case of the Eq (1) and a general representation of the probability matrix *P* for a time-reversible Markov process [60]. Thanks to the Eq (2) it is very convenient to easily generate at random a sample of matrices from the class *M*, which is crucial from the computational point of view. Moreover, it is generally accepted in phylogenetic studies [62, 63] that the time reversible matrix is a very good description of the real substitution process and it is not necessary to apply more general unrestricted models with larger number of parameters, which could cause over-parameterization.

To search a wide class of possible alternatives to the empirical matrices in the class of time reversible Markov processes we applied the same stationary distribution and the same restrictions on some eigenvalues. We assumed that the second eigenvalues of artificial matrices are the same as in empirical matrices, which corresponds to the same time scale for stochastic processes generated by these matrices. In other words, it means that their stationary distributions converge with the same speed. In the representation (2), the third and the fourth eigenvalues are real variables. We tested three constraints on these eigenvalues to check a possible influence of these assumptions on obtained results: (i) all generated matrices had a constant probability of nucleotide substitutions under the stationary state (i.e., $\sum_{i\in\{A,T,G,C\}} \pi_i(1-p_{ii}) = const.$) as a corresponding empirical matrix (*constant* assumption), (ii) all generated matrices had the same eigenvalues as a corresponding empirical matrix (*equal* assumption), (iii) all generated matrices had the same sum of their eigenvalues as a corresponding empirical matrix (*trace* assumption).

Fitness function

We considered several fitness functions *F* to investigate the influence of mutational matrices found for particular bacterial genomes on protein-coding sequences lying on the leading and lagging DNA strands in the corresponding genome (S1 Table). Sequences of these genes were downloaded from the GenBank database [64] and a decision about the location of these genes on the DNA strands was deduced according to the DNA asymmetry calculated in the Oriloc software [65]. Sequences from closely related genomes for which one mutational matrix was deduced were considered as one set.

In the functions F, we took into account the probability of non-synonymous substitutions and the mean value of amino acid substitution cost with and without nonsense mutations. Therefore, the optimizing matrices were tested on sites different from those used in inferring the empirical mutational matrices, which were derived from synonymous sites. The probability of non-synonymous substitutions was calculated based on the empirical codon frequency and the probability of change of one codon to another, coding different amino acids. The probability of the codon change was realized by a single nucleotide mutation based on the appropriate nucleotide substitution matrix. In the calculation of the mean value of the amino acid substitution costs, we additionally multiplied the probability of the codon change by a value reflecting differences between the amino acids. These differences were based on several amino acid scoring matrices and indices describing various physicochemical and biochemical properties of amino acids: chemical distance [66], hydropathy index [67], amino acid pair distance [68], EMPAR matrix [69] and polar requirement matrix [70]. All the matrices and indices were downloaded from the AAindex database [71]. The matrices and indices include only knowledge about properties of amino acids without any influence of underlying nucleotide mutations. When substitutions of stop codon were considered, we assumed their cost as the highest value of all amino acid substitution costs in the given measure.

To assess the optimality of empirical mutational matrix and easily compare the results for different genomes, we normalized the obtained values of fitness function according to the formula:

$$F_{norm} = (F_{emp} - F_{min})/(F_{max} - F_{min}), \quad (3)$$

where: F_{emp} is the value of fitness function for the empirical matrix, F_{min} is the smallest and F_{max} the largest value of fitness function found for the artificial matrices under given conditions. Clearly, $F_{\mathrm{norm}} = 0$ indicates that the empirical matrix minimizes costs of mutation just as the best artificial matrix, whereas $F_{\mathrm{norm}} = 1$ means that the empirical matrix maximizes the effect of mutations just as the best generated matrix.

Searching for the extreme values of fitness function

To find the maximum or minimum value of a given function F for artificial mutational matrices, we used the Evolutionary Strategies (ES) approach [72, 73]. This technique is an adequate tool in the finding solution of optimization problems, where the search space is not exactly defined and the solution is hard to find analytically. Similarly to the classical ES procedure, we started with the population of 100 random candidate solutions, i.e., transition probability matrices and carried out simulations according to the ES principles. The matrices at the initialization stage were computed according to the Eq (2). Three eigenvectors, which constitute matrix A, were selected at random and orthogonalized according to the Gramm-Schmidt orthogonalization procedure. The third and the fourth eigenvalue, which are necessary to well define the matrix Λ were selected at random based on one of three assumptions (see above). Unfortunately, this method does not guarantee to obtain always a probability transition matrix because negative elements can appear. Therefore, we had to repeat this procedure up to the valid probability matrix was generated.

At each generation step, we applied a mutation and selection procedure. The process of mutation introduction was realized by a modification of a matrix (an individual), which was done by a random shift of eigenvectors and/or two eigenvalues according to the normal distribution $N(0,\sigma)$. The scale parameter $\sigma = 0.3$ was tuned in initial simulation tests to obtain a quick convergence to the optimal solution. We run the Gramm-Schmidt orthogonalization procedure so many times to obtain a probability transition matrix.

In the selection stage, the half of the worst candidate solutions (according to their *F* values) were deleted and replaced by survived individuals. Simulations were run over 10,000 steps. The length of the simulations proved sufficient because all important parameters stabilized till this time (S1 Fig). Finally, the best matrices under a given criterion were extracted from the population.

In S1 Fig, we presented a typical simulation run, which was described by the average, minimum and maximum value of the fitness function *F*, calculated from the population of 100 individuals in every generation. It is clear that all considered statistics increased sharply about the 500th simulation steps and then remained stable until the end of the simulation despite small fluctuations of the minimum. To check the stability of the algorithm, we compared simulations with the same parameters but under different random seeds (S2 Fig). These simulations converged to the same value of fitness function.

Comparisons of empirical and optimized matrices by Principal Component Analysis (PCA) and Kruskal-Wallis test were carried out in Statistica (StatSoft Inc. 2011, version 10, www.statsoft.com). In PCA, we assumed a covariance matrix in the calculation of principal components.

Results

The aim of our study was to test to what extent spontaneous mutational pressures described by transition probability matrices are optimized according to harmful effects on protein-coding sequences. We considered six mutational matrices inferred from various bacterial genomes (S1 Table). Their effects were measured by probability of non-synonymous substitutions and costs of amino acid substitutions with and without nonsense mutations. The amino acid costs were described by different matrices and indices, which characterized various physicochemical and biochemical properties of amino acids. As a reference set to the empirical mutational matrices, we used optimized matrices that were initially randomly generated. For their generation, we applied spectral representation of time reversible Markov processes, which fulfilled the assumed conditions. Our model allowed us to control several parameters of the Markov process such as the speed of convergence to the stationary distribution and probability of substitutions in the stationary state. The generated matrices had the same stationary distribution as well as the first and the second eigenvalues as the empirical matrices. However, the third and the fourth eigenvalue were chosen according to three additional claims. The constant assumption meant that all generated matrices had the constant probability of nucleotide substitutions under stationarity as a corresponding empirical matrix. The equal assumption indicated that all generated matrices had all the same eigenvalues as the empirical matrix. Finally, the trace assumption meant that all generated matrices had the sum of their eigenvalues as the empirical matrix. Using the modified Evolutionary Strategy approach, we searched the space of randomly generated matrices to find the best optimized matrices, i.e., maximizing or minimizing harmful effects on protein-coding sequences expressed by a fitness function. Finally, the empirical mutational matrices were compared with the solutions found.

General comparison of empirical and optimized matrices

In the example shown in Fig 1, we presented values of a given fitness function *F* for three types of assumptions on eigenvalues. It is visible that the value for empirical leading strand matrix from *Borrelia burgdorferi* is located closer to the smallest fitness function than to the largest value. It indicates that the empirical matrix is to some extent optimized to minimize the probability of non-synonymous substitution. To easily compare results between different genomes, measures of mutation effect and assumptions for the third and fourth eigenvalues, we

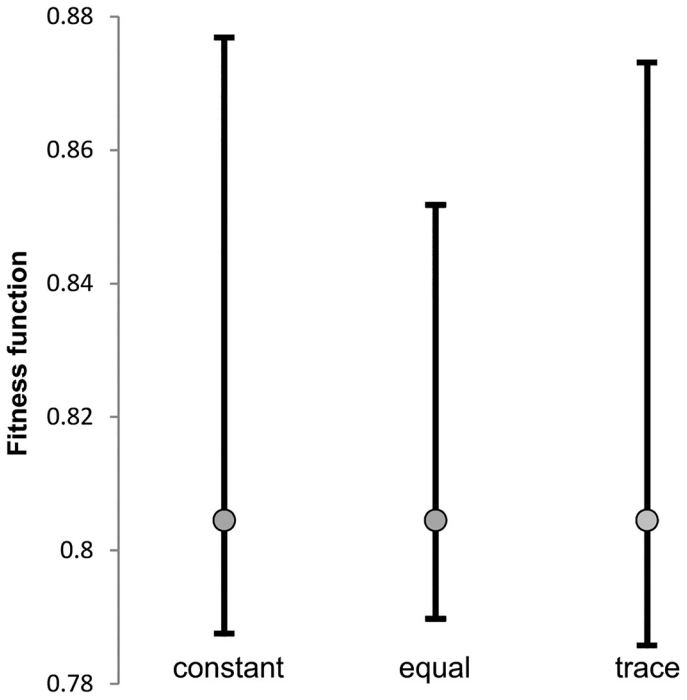

Fig 1. Fitness function for three assumptions on eigenvalues. Comparison of the fitness function (measured by the probability of non-synonymous substitutions) for the *B*. *burgdorferi* mutational matrix from the leading strand (grey dot) with the largest (upper whiskers) and the smallest (lower whiskers) values found for artificial matrices considering three types of assumptions on eigenvalues.

doi:10.1371/journal.pone.0130411.g001

normalized the value of fitness function comparing empirical matrix value with values for optimized matrices (Eq 3). Briefly, the normalized fitness function (F_{norm}) approaching 0 indicates that the empirical matrix has a tendency to minimize costs of mutations, whereas F_{norm} approaching 1 means that this matrix maximizes the effect of mutation on protein-coding sequences. Sample results for selected conditions were shown in Tables 2 and 3.

Fig 2 presents a distribution of F_{norm} values for all 231 combinations of genomes, different measures of mutational effect on protein-coding genes and three assumptions on eigenvalues of the generated matrices for two DNA strand separately. For all these combinations, the

Table 2. The normalized fitness function for non-synonymous substitutions.

Genome	Assumption on eigenvalues		
	constant	equal	trace
Borrelia burgdorferi	0.190	0.238	0.215
Chlamydia muridarum	0.170	-0.004	0.161
Chlamydia trachomatis	0.161	0.042	0.155
Escherichia coli	0.128	-0.010	0.124
Rickettsia species	0.249	0.131	0.254
Staphylococcus aureus	0.270	0.206	0.287
Streptococcus pyogenes	0.238	0.163	0.284

The normalized fitness function was measured by the probability of non-synonymous substitutions for leading strand matrices from seven genomes and three types of assumptions on eigenvalues.

doi:10.1371/journal.pone.0130411.t002

Table 3. The normalized fitness function for costs of amino acid substitutions.

Genome	Assumption on eigenvalues					
	constant		equal		trace	
	AA	AA+Stp	AA	AA+Stp	AA	AA+Stp
Borrelia burgdorferi	0.266	0.209	0.015	0.010	0.144	0.121
Chlamydia muridarum	0.182	0.193	0.246	0.265	0.221	0.221
Chlamydia trachomatis	0.216	0.231	0.320	0.360	0.217	0.234
Escherichia coli	0.125	0.165	0.074	0.147	0.058	0.145
Rickettsia species	0.225	0.218	0.176	0.171	0.198	0.201
Staphylococcus aureus	0.245	0.279	0.019	0.017	0.142	0.109
Streptococcus pyogenes	0.289	0.210	-0.001	0.048	0.115	0.114

The normalized fitness function measured by the mean cost of amino acid substitutions without (AA) and with (AA+Stp) stop codons using polar requirement for the leading strand matrices from seven genomes and three types of assumptions on eigenvalues.

doi:10.1371/journal.pone.0130411.t003

empirical mutational matrices were closer (F_{norm} was lower than 0.5) to the artificial matrices minimizing costs of mutations than to the matrices maximizing them. The leading strand matrices were slightly better optimized (the mean F_{norm}: 0.202, the range: -0.010 to 0.487) than the lagging strand matrices (the mean F_{norm}: 0.212, the range: -0.006 to 0.471). The empirical leading strand matrices showed $F_{norm} < 0.25$ in more than 73% of tested conditions, whereas the lagging strand matrices in more than 67% of conditions. It indicates slightly better minimization of the leading strand matrices. The largest F_{norm} value (0.487) was for the *Chlamydia*

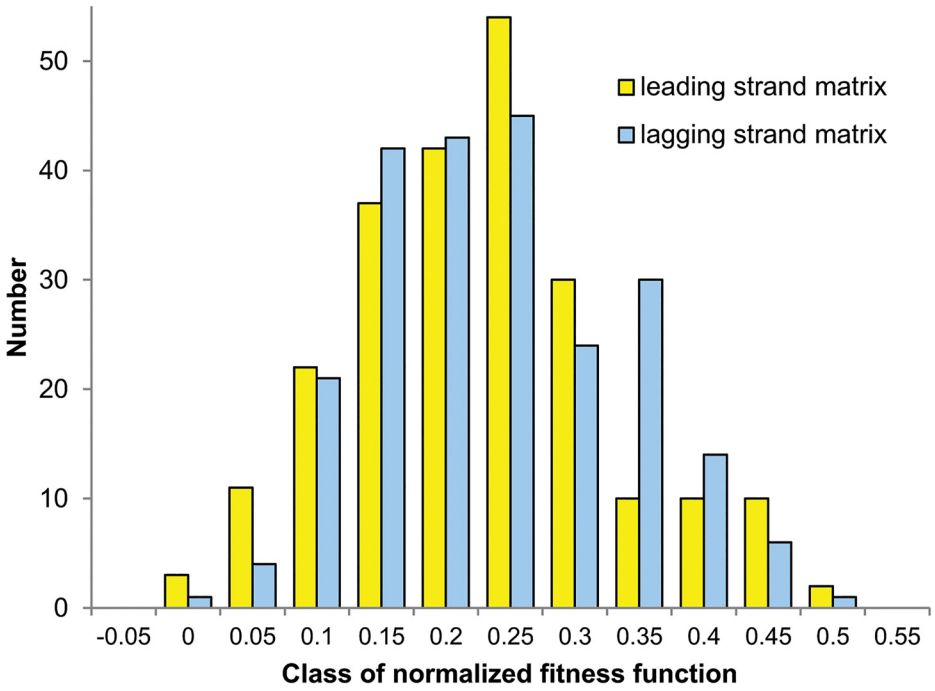

Fig 2. Distribution of normalized fitness function for empirical matrices. The distribution of normalized fitness function F_{norm} for all 231 combinations of genomes, assuming different measures of mutational effect on protein-coding genes and three assumptions on eigenvalues of generated matrices for two differently replicated DNA strands.

doi:10.1371/journal.pone.0130411.g002

trachomatis leading strand matrix compared with the best artificial matrices optimized according to the equal assumption on eigenvalues and the costs of amino acids substitutions considering their hydrophobic properties. What is more, in four cases, the applied algorithm found no better artificial matrix than the empirical one, which performed slightly better than the best artificial matrix (F_{norm} obtained negative values from -0.001 to -0.01). Such empirical matrices were the leading strand matrix for *Streptococccus pyogenes* tested according to the costs of amino acids substitutions under polar requirement, as well as the lagging strand matrix for *C. trachomatis* and the leading strand matrices for *C. muridarum* and *Escherichia coli* tested according to the probability of non-synonymous substitution. All these instances fulfilled the equal assumption on eigenvalues, i.e., the most restrictive one, which can explain the close similarity between the empirical and found optimized matrices.

To check if in these four cases the algorithm got stuck in a local minimum and was unable to find better solution, we carried out 100 additional simulations with different seeds. However, none of them produced matrices that were better optimized than the empirical ones. Moreover, the algorithm always converged almost to the same solution. The percentage difference between extreme values of fitness function of these solutions was extremely small, from 0.002% (for *S. pyogenes*) to 0.026% (for *E. coli*). We also checked if initial matrices were, in fact, randomly generated and represented a wide range of starting points for the algorithm to search a vast space of potential solutions. In fact, the range of their fitness function was from 46 (for *E. coli*) to 1326 (for *S. pyogenes*) times larger than the found solutions (Fig 3). In Fig 4, we also visualized the matrices by Principal Component Analysis (PCA), which showed that the initial matrices represented a wide spectrum of starting points, whereas the minimizing matrices were restricted to a very small region. Interestingly, the empirical matrix was placed in the middle of the solutions found.

We observed a small difference in the value of normalized fitness function F_{norm} between genomes. Considering three types of assumptions on eigenvalues, *E. coli* matrices from two DNA strands were on average better optimized according to the probability of non-synonymous substitutions (0.06) than *Chlamydia* matrices (0.11) and other genomes (0.21 – 0.28). However, considering costs of amino acids substitutions, *Chlamydia* matrices were on average slightly worse (0.27 and 0.28 for *C. muridarum* and *C. trachomatis*, respectively) than other genomes (0.17 – 0.19). Nevertheless, these results indicate that there are no significant differences between optimization of matrices coming from different genomes and tested on different fitness functions.

A small difference in F_{norm} was found considering assumptions on the third and fourth eigenvalues of the generated matrices. For the probability of non-synonymous substitutions and the equal assumption, F_{norm} was nearly on average two times smaller (0.13) than for constant (0.20) and trace assumptions (0.23) for all considered matrices. For different costs of amino acid substitutions, the normalized fitness function was on average larger for the constant assumption (0.26) than for the trace and the equal assumptions (0.18 for these two cases). Small values for the equal assumption can result from the most restrictive conditions on eigenvalues.

Considering different measures of mutation effects on protein-coding sequences, we found that the empirical matrices were generally the best optimized according to EMPAR matrix [69] (F_{norm} = 0.16), next to polar requirement [70] (0.17) and non-synonymous substitutions (0.18). A slightly higher value of F_{norm} = 0.24 was obtained for chemical distance [66], hydropathy index [67] and amino acid pair distance by [68]. We did not find a significant difference between F_{norm} calculated for amino acid costs with or without nonsense mutations, with the exception of the equal assumption (0.20 vs 0.17, respectively).

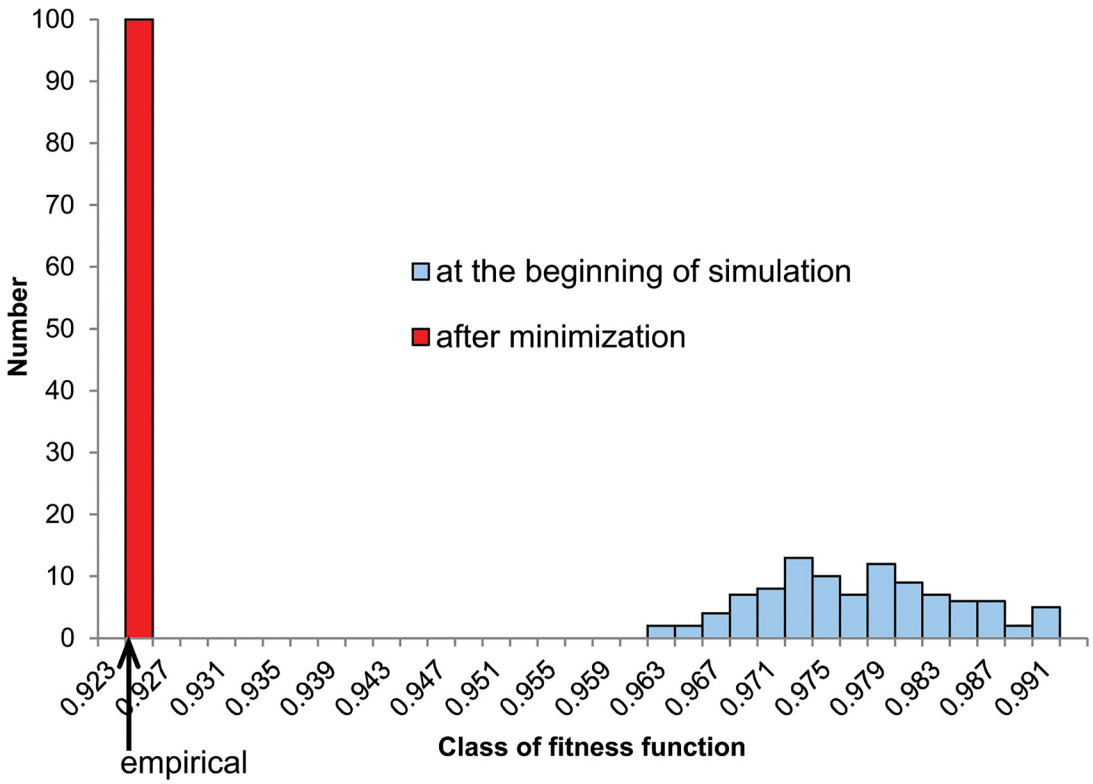

Fig 3. Distribution of fitness function for artificial matrices for *S. pyogenes*. The distribution of fitness function *F* for 100 artificial starting matrices (with equal assumption) at the beginning of simulation and after minimization (1000 steps) according to the costs of amino acids substitutions under polar requirement. The value for the empirical leading strand substitution matrix from *S. pyogenes* (0.923161) was indicated by the arrow.

doi:10.1371/journal.pone.0130411.g003

Properties of empirical and optimized matrices

One of parameters that was used to describe spontaneous mutational pressure is transitions to transversions ratio. The expected ratio should be 1:2, if all nucleotide substitutions happen with the same probability. However, transitions are usually observed several times more often in real sequences than transversions [74]. The observed bias results from higher rate of chemical changes between structurally similar nucleotides and more probable transition substitutions introduced during DNA replication. In addition to that, transitions are less harmful than transversions in terms of changing coded amino acids or their properties and, therefore, more often accepted. Thus, we should observe that matrices minimizing costs of amino acid and non-synonymous substitutions are characterized by the high transitions/transversions ratio compared to maximizing matrices. Actually, the ratio for the minimizing matrices was on average about nine to ten times larger than for maximizing ones and did not differ significantly between matrices for two DNA strands (Table 4). Interestingly, the average ratio for empirical matrices was even bigger than that for the minimizing matrices. Although the distribution of the ratio for the minimizing matrices was quite wide, almost 72% of the values for the leading strand matrix were larger than one and, likewise, 80% of the values for the lagging strand matrix were also greater than one (Fig 5). In contrast to that, none of the maximizing matrices exceeded this value and all but one were smaller than 0.5. The empirical matrices well corresponded to the distribution of the minimizing matrices.

Examples of optimized matrices in comparison to corresponding empirical matrix were presented in S2 Table. To easily visualize matrices according to all possible 12 nucleotide

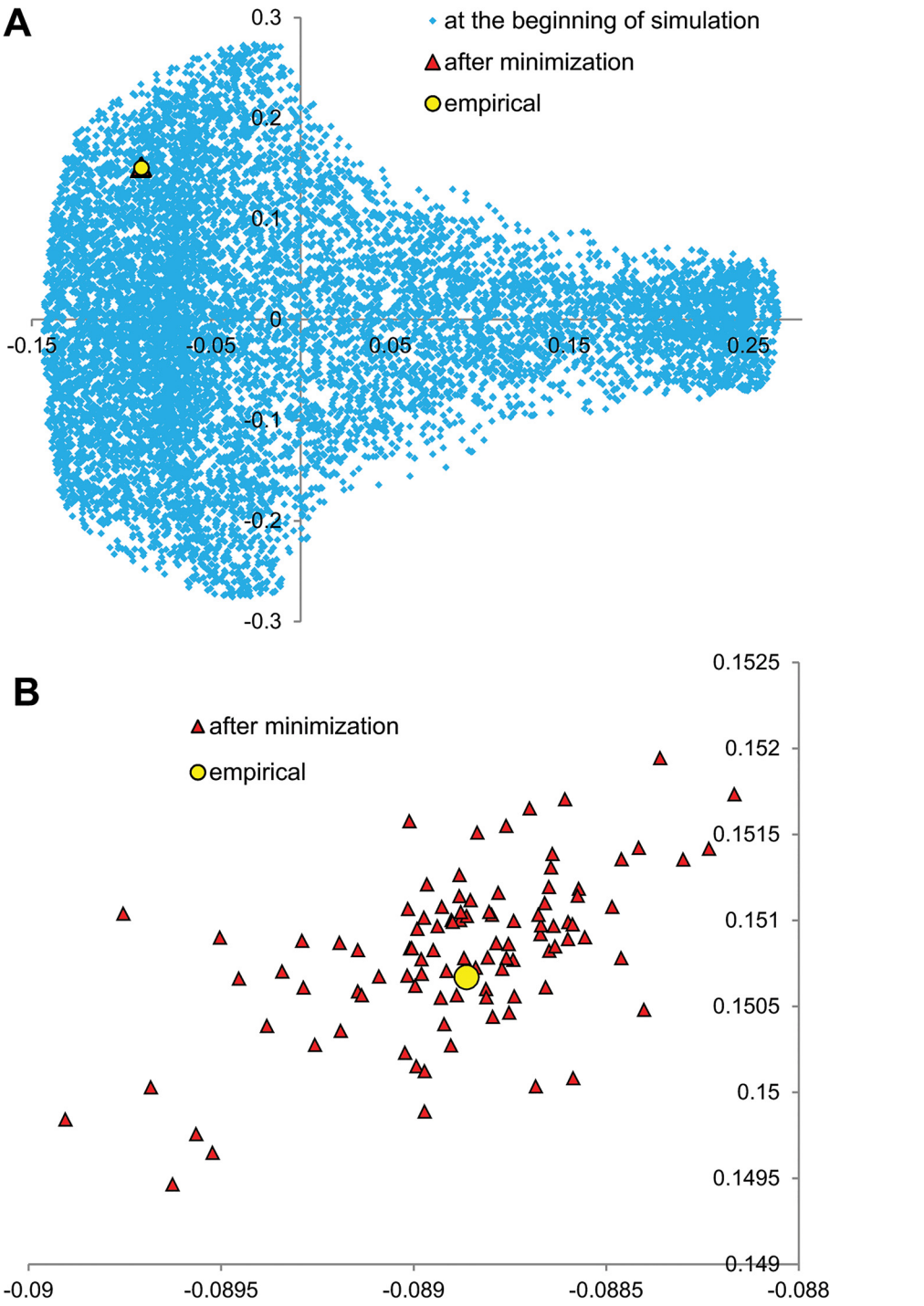

Fig 4. PCA of *S. pyogenes* empirical matrix and artificial matrices. (A) The Principal Component Analysis of artificial starting matrices (with equal assumption) at the beginning of simulation and after the minimization according to the costs of amino acids substitutions under polar requirement. The empirical leading strand substitution matrix from *S. pyogenes* was indicated by the open circle. (B) The enlarged part of A focused on the region occupied by the minimizing and empirical matrices.

doi:10.1371/journal.pone.0130411.g004

Table 4. The ratio of transitions to transversions for different types of matrices.

	Empirical matrices		Minimizing matrices		Maximizing matrices	
	leading	lagging	leading	lagging	leading	lagging
Mean	1.88	1.80	1.38	1.49	0.15	0.14
Minimum	1.11	1.11	0.39	0.38	0.00	0.00
Maximum	2.18	2.14	2.99	2.75	0.53	0.46

The transitions/transversions ratio was calculated for seven empirical matrices and, in the case of optimizing matrices, for all 231 combinations of genomes, assuming different measures of mutational effect on protein-coding genes and three assumptions on eigenvalues.

doi:10.1371/journal.pone.0130411.t004

substitutions (corresponding to elements of these matrices), we carried out Principal Component Analysis to reduce the number of dimensions from 12 matrix elements to two main variables (Fig 6). Maximizing and minimizing matrices were clearly separated into two non-overlapping groups by the first component, which indicates that they differ in their elements. All empirical matrices were placed among the set of minimizing matrices, which indicates a similar pattern (rate) of nucleotide substitutions described by the empirical and the optimized matrices. There was also no difference in the distribution of matrices in respect to differently

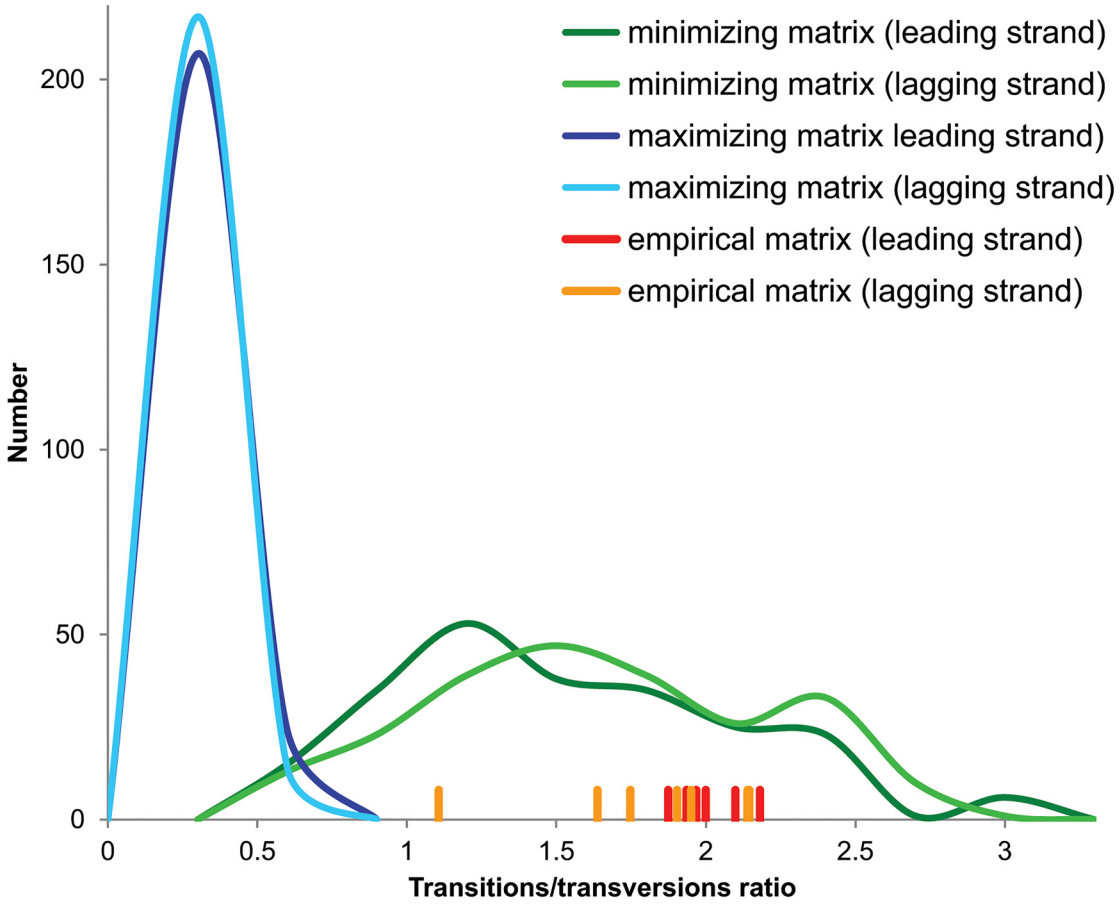

Fig 5. Transitions/transversions distribution. The distribution of transitions to transversions ratio for empirical matrices as well as matrices maximizing and minimizing costs of amino acid and non-synonymous substitutions. Data for two differently replicated DNA strands (leading and lagging) were considered, separately.

doi:10.1371/journal.pone.0130411.g005

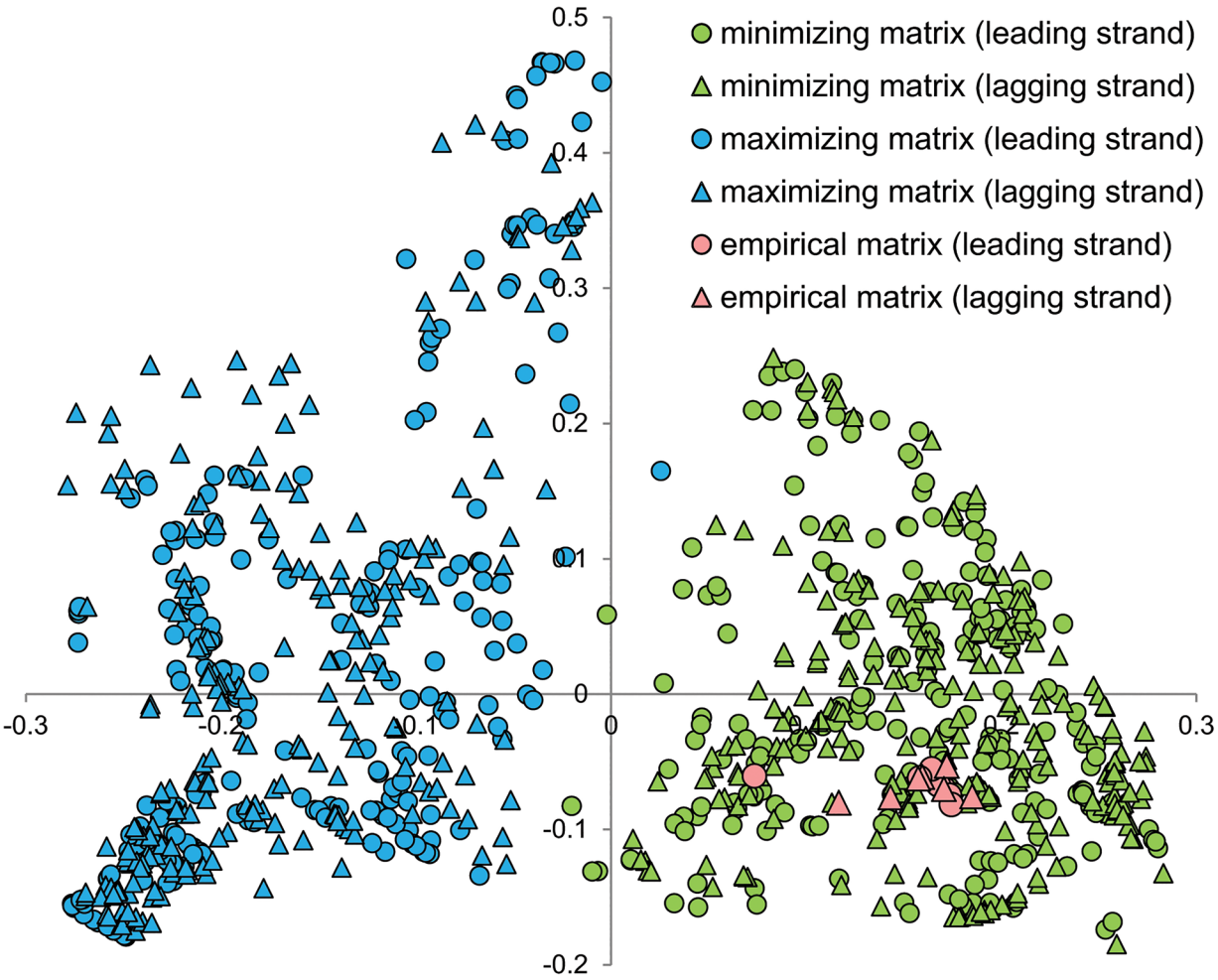

Fig 6. PCA of empirical and optimized artificial matrices. The Principal Component Analysis of the empirical substitution matrices as well as 231 matrices maximizing and minimizing costs of amino acids and non-synonymous substitutions. Data for two differently replicated DNA strands (leading and lagging) were considered, separately.

doi:10.1371/journal.pone.0130411.g006

replicated DNA strands. The largest correlation with the first (most discriminative) component showed transversions: A→T (-0.73), T→A (-0.72), G→T (-0.69) and T→G (-0.59), as well as transitions A→G (0.75) and G→A (0.80). The opposite signs at the correlation coefficients are connected with a different influence of these variables on the first component and the separation of sets.

The transversions A→T and T→A had statistically significantly (Kruskal-Wallis test, p-value < 0.001) smaller rates in the analyzed empirical and minimizing matrices than in the maximizing matrices, whereas the transitions A→G, G→A, T→C and C→T were significantly larger in the first-mentioned matrices than in the latter (p-value < 0.0032) (Fig 7). It is worth emphasizing that we did not find significant differences between the empirical and minimizing matrices for any of these substitutions. Transitions T→G and G→T also showed smaller values for the empirical and minimizing matrices than the maximizing ones. Differences between the minimizing and the maximizing matrices for these two substitutions and between the empirical and maximizing matrix for the lagging strand in the case of G→T were statistically significant (p-value = 0.043). The other substitutions did not show significant differences between the

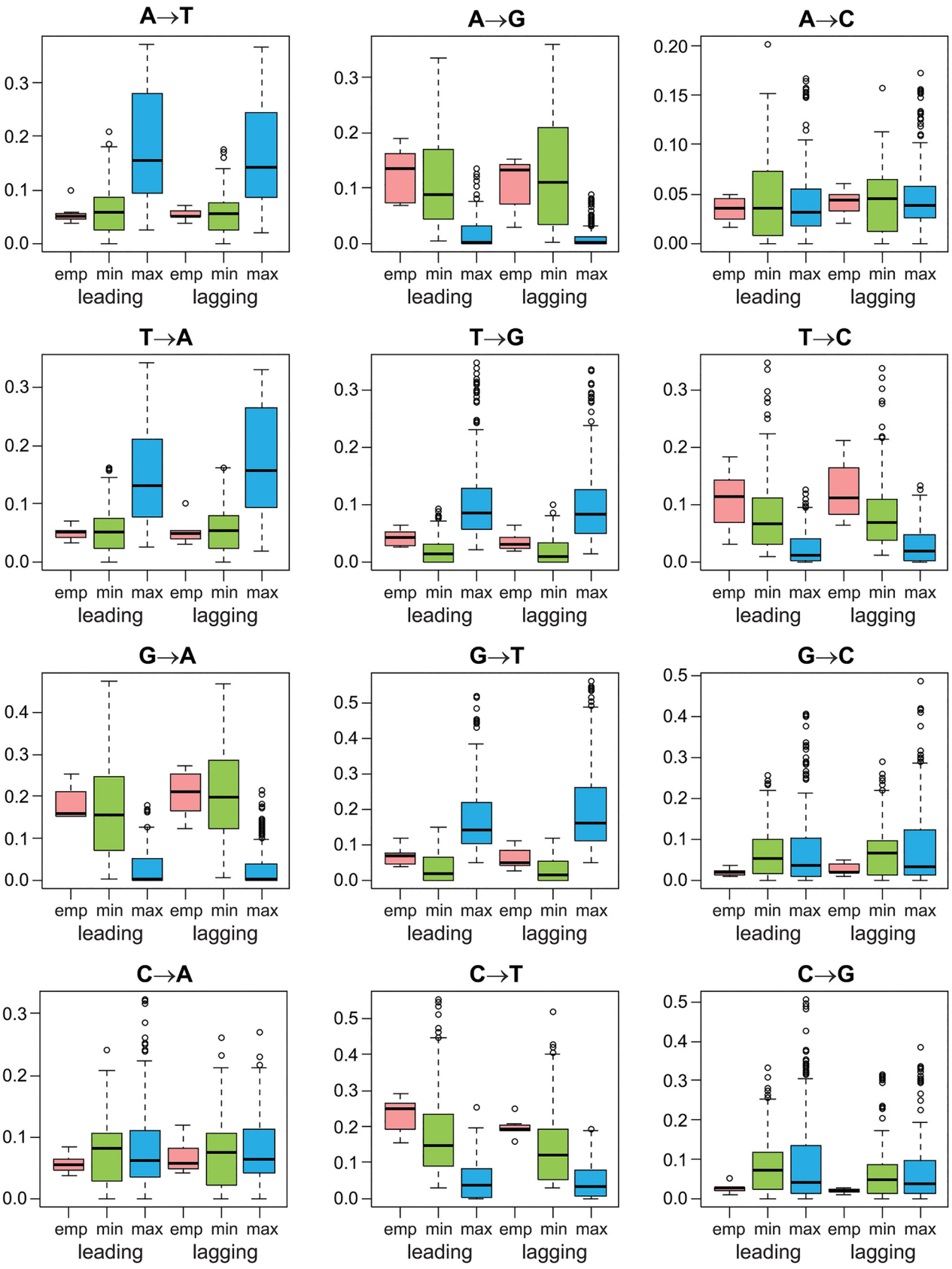
A→T
A→G
A→C
T→A
T→G
T→C
G→A
G→T
G→C
C→A
C→T
C→G
emp min max emp min max
leading lagging

Fig 7. Box-plots of nucleotide substitutions' rates for empirical and optimized artificial matrices. The thick horizontal lines indicate median, the colored boxes show quartile range and the whiskers determine the range without outliers. Data for two differently replicated DNA strands (leading and lagging) were considered, separately.

doi:10.1371/journal.pone.0130411.g007

matrices in almost all cases, although the empirical matrices were characterized by the smallest rate of transversions G→C and C→G.

Three most frequent substitutions of the minimizing leading and lagging strand matrices were the same but differed in proportions. These were transitions: C→T, G→A and A→G, with percentages 34%, 33% and 18% for the leading strand and 26%, 52% and 11% for the lagging strand (S3 Table). The same substitutions dominated also in four, one and three of seven empirical matrices from the leading strand, respectively. For the lagging strand matrices, C→T and G→A dominated in two and four cases, respectively. In contrast to that, these three substitutions were selected as the rarest substitution in 53% (for the leading strand) and 61% (for the lagging strand) of maximizing matrices. In turn, three the least frequent substitutions were T→G (32%), A→C (20%) and G→C (14%) in the minimizing leading strand matrices, whereas in the minimizing lagging strand matrices, T→G (39%), C→G (17%) and A→T (11%). The T→G and A→C transversions had also the smallest rate in two empirical leading strand matrices and G→C substitution in five of them. In the case of the empirical lagging strand matrices, four had C→G and three G→C as the rarest substitution but none T→G or A→T. However, three the least frequent substitutions in 66% of the minimizing leading strand matrices were selected as the most dominant mutation in only 13% of maximizing matrices (A→C in none of them). The proportions for the lagging strand matrices are 57% and 37%, respectively. The lack of the full correspondence between the rarest substitutions in the minimizing matrices and the most common substitutions in the maximizing ones, as could be expected, may result from the same restrictions imposed on the generated matrices, e.g., stationary distribution and the same speed of convergence to the distributions.

Discussion

The minimization of harmful effects of mutations can be achieved by a decrease in the global mutation rate by evolution of high-fidelity polymerases, which select and incorporate nucleotides into newly synthesized DNA strands [38, 39, 75]. The other adaptations can be more effective mechanisms of mutation correction: exonucleolytic proofreading [43] and post-replicative DNA mismatch repair [40, 41], which excise and replace incorrectly inserted bases. Here we focused on more sophisticated aspects of this optimization, namely on the pattern of nucleotide substitutions, i.e., relative rates between changes of particular nucleotides. Such optimization can be connected with changes in the quantity of nucleotides in the cellular dNTP pools [76, 77] as well as preferences of polymerases and correction mechanism to particular nucleotides [78–84], which may favor introduction of one nucleotide over another to DNA.

The nucleotide substitution patterns are usually described by transition probability matrices. To assess the optimization of the empirical mutational matrices, we compared them with the reference set of optimized artificial matrices. In contrast to the method used by Błażej et al. [23], in which the class of General Time Reversible matrices with six parameters was considered as the reference set, here we analyzed a more convenient class of transition probability matrices. In both cases, the generated matrices had the stationary distribution (the left eigenvector corresponding to the first eigenvalue) as the empirical matrix. However, in this approach we also assumed the second eigenvalue (corresponding to the speed of convergence to the stationarity) as in the empirical matrices, while the third and fourth eigenvalues could

vary. Thus, the empirical mutational matrices were compared with other matrices in their class with similar mathematical properties.

The obtained results indicate that spontaneous mutational pressures in bacterial genomes, described by transition probability matrices, are optimized to minimize rather than maximize the frequency of non-synonymous substitutions and costs of amino acid replacements according to their different physicochemical and biochemical properties. In all 231 analyzed cases, the influence of mutational pressure on protein-coding sequences in comparison to the best optimized artificial matrices, measured by the normalized fitness function, never exceeded 0.5 (in this scale, 0 indicates that the empirical matrix minimizes costs of mutations as the best artificial matrix, whereas 1 means that this matrix maximizes costs of amino acid substitutions as the best artificial matrix). Interestingly, in four cases the empirical matrix minimized the influence of mutation on protein-coding sequences slightly better than any optimized matrices. We thoroughly tested if the applied algorithm got stuck in a local minimum during searches of the best solution. However, 100 simulations with different seeds always converged to the same or very similar solution.

It is interesting that empirical nucleotide mutation matrices showed a similar trend in the minimization of harmful substitutions, despite generating different stationary distributions and testing on different protein-coding sequences. This trend did not depend, or depended very weakly, on the analyzed genomes, effects of mutations on protein-coding sequences and assumptions on eigenvalues of artificial optimized matrices, to which the empirical ones were compared. It indicates that various mutational pressures are similarly optimized to minimize costs of mutations in different biological systems. We also did not find significant differences between matrices from two differently replicated DNA strands (leading and lagging), although the leading strand matrices were slightly better minimized than matrices from the lagging strand. It may be related with the preferred location of genes that are essential for cell functioning (e.g., coding for ribosomal proteins) in the leading strand [21, 22]. The genes for ribosomal proteins maintain conserved positions on bacterial chromosomes with the phylogenetic distance of compared genomes [18]. Moreover, it was observed a higher frequency of gene translocations from the lagging to the leading strand rather than in the opposite direction [19] and smaller rate of substitutions' accumulation in the leading than lagging strand genes [12–15].

Both the empirical matrices and matrices minimizing mutational effects on protein-coding sequences demonstrated the excess of transitions over transversions as it would be expected because the latter have more harmful impact on proteins. Although the leading and lagging strand matrices for the same genome are characterized by different nucleotide stationary distributions (Table 1), they also showed similar patterns of nucleotide substitutions with the large rate of transitions A→G, G→A, T→C and C→T as the minimizing matrices. It indicates that the minimization of mutational costs is realized by the same relative rates.

Although most mutations are deleterious (especially those replacing amino acids with different properties), we should not expect the perfect minimization of mutational effects by empirical matrices. The mutational pressure can approach two extremes, the minimization and maximization [26, 27]. The pressure is minimized to decrease the number of harmful mutations and cost of DNA repair. On the other hand, the pressure can be maximized to increase a genetic variation and the number of profitable substitutions driven by positive selection, which are necessary in changing environmental conditions and strong competition between organisms. The mutations balance between these two extremes and obtain some optimal values specific for a given biological system (genome) [23]. In agreement with that, the analyzed empirical mutational matrices, when compared with the optimized matrices, locate between these two extremes showing, however, closer similarity to matrices that minimize costs of mutations.

Supporting Information

S1 Fig. Variation of fitness function. The minimum, maximum and mean of fitness function *F* for the rate matrix maximizing amino acid costs and characterizing by a constant probability of nucleotide substitutions as the empirical leading strand matrix from *B. burgdorferi* (the constant assumption).
(PDF)

S2 Fig. Fitness function for 10 runs. The mean value of fitness function *F* for ten runs of the algorithm with different seeds aimed to find the rate matrix maximizing amino acid cost and characterizing by a constant probability of nucleotide substitutions as the empirical leading strand matrix from *B. burgdorferi* (the constant assumption).
(PDF)

S1 Table. Genomes and their protein-coding genes used in the study.
(DOCX)

S2 Table. *S. pyogenes* empirical matrix and optimized artificial matrices. The comparison of the empirical matrix from *S. pyogenes* leading strand with matrices optimized according to the costs of amino acids substitutions under polar requirement and equal assumption on eigenvalues. The substitutions were sorted in descending order according to the empirical values.
(DOCX)

S3 Table. Number of matrices in which a given substitution showed the largest or the smallest rate for the leading (before slash) and lagging (after slash) DNA strands.
(DOCX)

Acknowledgments

We are very grateful to the Reviewers and Editors for their excellent comments and insightful remarks that significantly improved the paper. We would like also to thank Prof. Stanisław Cebrat for inspiring and helpful discussions as well as to Przemysław Gagat for English editing. The publication fee was funded by the KNOW Consortium.

Author Contributions

Conceived and designed the experiments: PB PM BM. Performed the experiments: PB BM PM MG. Analyzed the data: PM PB BM MG. Contributed reagents/materials/analysis tools: PB BM MG PM. Wrote the paper: PM PB.

References

1. Frank AC, Lobry JR. Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. Gene. 1999; 238(1):65–77. Epub 1999/11/26. PMID: 10570985.
2. Kowalczuk M, Mackiewicz P, Mackiewicz D, Nowicka A, Dudkiewicz M, Dudek MR, et al. DNA asymmetry and the replicational mutational pressure. J Appl Genet. 2001; 42(4):553–77. Epub 2003/10/18. PMID: 14564030.
3. Lobry JR. Asymmetric substitution patterns in the two DNA strands of bacteria. Mol Biol Evol. 1996; 13 (5):660–5. Epub 1996/05/01. PMID: 8676740.
4. McLean MJ, Wolfe KH, Devine KM. Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. J Mol Evol. 1998; 47(6):691–6. Epub 1998/12/16. PMID: 9847411.
5. Mrazek J, Karlin S. Strand compositional asymmetry in bacterial and large viral genomes. Proc Natl Acad Sci U S A. 1998; 95(7):3720–5. Epub 1998/05/09. PMID: 9520433; PubMed Central PMCID: PMC19903.

6. Rocha EP, Danchin A, Viari A. Universal replication biases in bacteria. Mol Microbiol. 1999; 32(1):11–6. Epub 1999/04/27. PMID: 10216855.
7. Tillier ER, Collins RA. The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. J Mol Evol. 2000; 50(3):249–57. Epub 2001/02/07. PMID: 10754068.
8. Lobry JR, Sueoka N. Asymmetric directional mutation pressures in bacteria. Genome Biol. 2002; 3(10): RESEARCH0058. Epub 2002/10/10. PMID: 12372146; PubMed Central PMCID: PMC134625.
9. McInerney JO. Replicational and transcriptional selection on codon usage in Borrelia burgdorferi. Proc Natl Acad Sci U S A. 1998; 95(18):10698–703. Epub 1998/09/02. PMID: 9724767; PubMed Central PMCID: PMC27958.
10. Mackiewicz P, Gierlik A, Kowalczuk M, Szczepanik D, Dudek MR, Cebrat S. Mechanisms generating long-range correlation in nucleotide composition of the Borrelia burgdorferi genome. Physica A. 1999; 273(1–2):103–15. doi: 10.1016/s0378-4371(99)00345-3 PMID: WOS:000084121100010.
11. Mackiewicz P, Gierlik A, Kowalczuk M, Dudek MR, Cebrat S. How does replication-associated mutational pressure influence amino acid composition of proteins? Genome Res. 1999; 9(5):409–16. Epub 1999/05/18. PMID: 10330120; PubMed Central PMCID: PMC310782.
12. Mackiewicz P, Mackiewicz D, Kowalczuk M, Dudkiewicz M, Dudek MR, Cebrat S. High divergence rate of sequences located on different DNA strands in closely related bacterial genomes. J Appl Genet. 2003; 44(4):561–84. Epub 2003/11/18. PMID: 14617839.
13. Szczepanik D, Mackiewicz P, Kowalczuk M, Gierlik A, Nowicka A, Dudek MR, et al. Evolution rates of genes on leading and lagging DNA strands. J Mol Evol. 2001; 52(5):426–33. Epub 2001/07/10. doi: 10.1007/s002390010172 PMID: 11443346.
14. Tillier ER, Collins RA. Replication orientation affects the rate and direction of bacterial gene evolution. J Mol Evol. 2000; 51(5):459–63. Epub 2000/11/18. PMID: 11080368.
15. Rocha EP, Danchin A. Ongoing evolution of strand composition in bacterial genomes. Mol Biol Evol. 2001; 18(9):1789–99. Epub 2001/08/16. PMID: 11504858.
16. Lin CH, Lian CY, Hsiung CA, Chen FC. Changes in transcriptional orientation are associated with increases in evolutionary rates of enterobacterial genes. BMC Bioinformatics. 2011; 12 Suppl 9:S19. Epub 2011/12/22. doi: 10.1186/1471-2105-12-S9-S19 PMID: 22152004; PubMed Central PMCID: PMC3283321.
17. Paul S, Million-Weaver S, Chattopadhyay S, Sokurenko E, Merrikh H. Accelerated gene evolution through replication-transcription conflicts. Nature. 2013; 495(7442):512–5. Epub 2013/03/30. doi: 10.1038/nature11989 PMID: 23538833; PubMed Central PMCID: PMC3807732.
18. Mackiewicz D, Mackiewicz P, Kowalczuk M, Dudkiewicz M, Dudek MR, Cebrat S. Rearrangements between differently replicating DNA strands in asymmetric bacterial genomes. Acta Microbiol Pol. 2003; 52(3):245–60. Epub 2004/01/28. PMID: 14743977.
19. Mackiewicz P, Mackiewicz D, Gierlik A, Kowalczuk M, Nowicka A, Dudkiewicz M, et al. The differential killing of genes by inversions in prokaryotic genomes. J Mol Evol. 2001; 53(6):615–21. Epub 2001/10/26. doi: 10.1007/s002390010248 PMID: 11677621.
20. Mackiewicz P, Mackiewicz D, Kowalczuk M, Cebrat S. Flip-flop around the origin and terminus of replication in prokaryotic genomes. Genome Biol. 2001; 2(12):INTERACTIONS1004. Epub 2002/01/16. PMID: 11790247; PubMed Central PMCID: PMC138987.
21. Rocha EP, Danchin A. Essentiality, not expressiveness, drives gene-strand bias in bacteria. Nat Genet. 2003; 34(4):377–8. Epub 2003/07/09. doi: 10.1038/ng1209 PMID: 12847524.
22. Rocha EP, Danchin A. Gene essentiality determines chromosome organisation in bacteria. Nucleic Acids Res. 2003; 31(22):6570–7. Epub 2003/11/07. PMID: 14602916; PubMed Central PMCID: PMC275555.
23. Błażej P, Mackiewicz P, Cebrat S, Wańczyk M, editors. Using Evolutionary Algorithms in Finding of Optimized Nucleotide Substitution Matrices. Genetic and Evolutionary Computation Conference, GECCO'13; 2013; Amsterdam, The Netherlands: Companion ACM; 2013.
24. Dudkiewicz M, Mackiewicz P, Nowicka A, Kowalczuk M, Mackiewicz D, Polak N, et al. Correspondence between mutation and selection pressure and the genetic code degeneracy in the gene evolution. Future Generation Computer Systems. 2005; 21(7):1033–9.
25. Mackiewicz P, Biecek P, Mackiewicz D, Kiraga J, Bączkowski K, Sobczyński M, et al. Optimisation of asymmetric mutational pressure and selection pressure around the universal genetic code. In: Bubak M, Dongarra J, VanAlbada GD, Sloot PMA, editors. Computational Science—ICCS 2008, PT 3. Lecture Notes in Computer Science. 5103: Elsevier, Springer; 2008. p. 100–9.
26. Radman M, Matic I, Taddei F. Evolution of evolvability. Ann N Y Acad Sci. 1999; 870:146–55. Epub 1999/07/23. PMID: 10415480.

27. Sniegowski PD, Gerrish PJ, Johnson T, Shaver A. The evolution of mutation rates: separating causes from consequences. Bioessays. 2000; 22(12):1057–66. Epub 2000/11/21. doi: 10.1002/1521-1878(200012)22:12<1057::AID-BIES3>3.0.CO;2-W PMID: 11084621.

28. Drake JW. A constant rate of spontaneous mutation in DNA-based microbes. Proc Natl Acad Sci U S A. 1991; 88(16):7160–4. Epub 1991/08/15. PMID: 1831267; PubMed Central PMCID: PMC52253.

29. Kimura M. On evolutionary adjustment of spontaneous mutation rates. Genetical Research. 1967; 9 (1):23–34.

30. Travis JM, Travis ER. Mutator dynamics in fluctuating environments. Proc Biol Sci. 2002; 269 (1491):591–7. Epub 2002/03/28. doi: 10.1098/rspb.2001.1902 PMID: 11916475; PubMed Central PMCID: PMC1690933.

31. de Visser JA. The fate of microbial mutators. Microbiology. 2002; 148(Pt 5):1247–52. Epub 2002/05/04. PMID: 11988499.

32. Denamur E, Matic I. Evolution of mutation rates in bacteria. Mol Microbiol. 2006; 60(4):820–7. Epub 2006/05/09. doi: 10.1111/j.1365-2958.2006.05150.x PMID: 16677295.

33. Johnson T, Barton NH. The effect of deleterious alleles on adaptation in asexual populations. Genetics. 2002; 162(1):395–411. Epub 2002/09/21. PMID: 12242249; PubMed Central PMCID: PMC1462245.

34. Orr HA. The rate of adaptation in asexuals. Genetics. 2000; 155(2):961–8. Epub 2000/06/03. PMID: 10835413; PubMed Central PMCID: PMC1461099.

35. Clune J, Misevic D, Ofria C, Lenski RE, Elena SF, Sanjuan R. Natural selection fails to optimize mutation rates for long-term adaptation on rugged fitness landscapes. PLoS Comput Biol. 2008; 4(9): e1000187. Epub 2008/09/27. doi: 10.1371/journal.pcbi.1000187 PMID: 18818724; PubMed Central PMCID: PMC2527516.

36. Stich M, Manrubia SC, Lazaro E. Variable mutation rates as an adaptive strategy in replicator populations. PLoS One. 2010; 5(6):e11186. Epub 2010/06/23. doi: 10.1371/journal.pone.0011186 PMID: 20567506; PubMed Central PMCID: PMC2887357.

37. Kunkel TA. Evolving views of DNA replication (in)fidelity. Cold Spring Harb Symp Quant Biol. 2009; 74:91–101. Epub 2009/11/12. doi: 10.1101/sqb.2009.74.027 PMID: 19903750; PubMed Central PMCID: PMC3628614.

38. Kunkel TA, Bebenek K. DNA replication fidelity. Annu Rev Biochem. 2000; 69:497–529. Epub 2000/08/31. doi: 10.1146/annurev.biochem.69.1.497 PMID: 10966467.

39. Loh E, Salk JJ, Loeb LA. Optimization of DNA polymerase mutation rates during bacterial evolution. Proc Natl Acad Sci U S A. 2010; 107(3):1154–9. Epub 2010/01/19. doi: 10.1073/pnas.0912451107 PMID: 20080608; PubMed Central PMCID: PMC2824296.

40. Schofield MJ, Hsieh P. DNA mismatch repair: molecular mechanisms and biological function. Annu Rev Microbiol. 2003; 57:579–608. Epub 2003/10/07. doi: 10.1146/annurev.micro.57.030502.090847 PMID: 14527292.

41. Kunkel TA, Erie DA. DNA mismatch repair. Annu Rev Biochem. 2005; 74:681–710. Epub 2005/06/15. doi: 10.1146/annurev.biochem.74.082803.133243 PMID: 15952900.

42. Pavlov YI, Shcherbakova PV, Rogozin IB. Roles of DNA polymerases in replication, repair, and recombination in eukaryotes. Int Rev Cytol. 2006; 255:41–132. Epub 2006/12/21. doi: 10.1016/S0074-7696(06)55002-8 PMID: 17178465.

43. Reha-Krantz LJ. DNA polymerase proofreading: Multiple roles maintain genome stability. Biochim Biophys Acta. 2010; 1804(5):1049–63. Epub 2009/06/24. doi: 10.1016/j.bbapap.2009.06.012 PMID: 19545649.

44. Freeland SJ, Hurst LD. The genetic code is one in a million. J Mol Evol. 1998; 47(3):238–48. Epub 1998/09/11. PMID: 9732450.

45. Freeland SJ, Wu T, Keulmann N. The case for an error minimizing standard genetic code. Orig Life Evol Biosph. 2003; 33(4–5):457–77. Epub 2003/11/08. PMID: 14604186.

46. Marquez R, Smit S, Knight R. Do universal codon-usage patterns minimize the effects of mutation and translation error? Genome Biol. 2005; 6(11):R91. Epub 2005/11/10. doi: 10.1186/gb-2005-6-11-r91 PMID: 16277746; PubMed Central PMCID: PMC1297647.

47. Sella G, Ardell DH. The impact of message mutation on the fitness of a genetic code. J Mol Evol. 2002; 54(5):638–51. Epub 2002/04/20. doi: 10.1007/s00239-001-0060-7 PMID: 11965436.

48. Zhu CT, Zeng XB, Huang WD. Codon usage decreases the error minimization within the genetic code. J Mol Evol. 2003; 57(5):533–7. Epub 2004/01/24. doi: 10.1007/s00239-003-2505-7 PMID: 14738311.

49. Archetti M. Codon usage bias and mutation constraints reduce the level of error minimization of the genetic code. J Mol Evol. 2004; 59(2):258–66. Epub 2004/10/16. doi: 10.1007/s00239-004-2620-0 PMID: 15486699.

50. Rogozin IB, Makarova KS, Natale DA, Spiridonov AN, Tatusov RL, Wolf YI, et al. Congruent evolution of different classes of non-coding DNA in prokaryotic genomes. Nucleic Acids Res. 2002; 30(19):4264–71. Epub 2002/10/05. PMID: 12364605; PubMed Central PMCID: PMC140549.

51. Kowalczuk M, Mackiewicz P, Mackiewicz D, Nowicka A, Dudkiewicz M, Dudek MR, et al. High correlation between the turnover of nucleotides under mutational pressure and the DNA composition. BMC Evol Biol. 2001; 1:13. Epub 2002/01/22. PMID: 11801180; PubMed Central PMCID: PMC64649.

52. Rocha EP, Touchon M, Feil EJ. Similar compositional biases are caused by very different mutational effects. Genome Res. 2006; 16(12):1537–47. Epub 2006/10/28. doi: 10.1101/gr.5525106 PMID: 17068325; PubMed Central PMCID: PMC1665637.

53. Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R. Codon Catalog Usage Is a Genome Strategy Modulated for Gene Expressivity. Nucleic Acids Res. 1981; 9(1):R43–R74. PMID: ISI:A1981KZ75700019.

54. Supek F, Skunca N, Repar J, Vlahovicek K, Smuc T. Translational Selection Is Ubiquitous in Prokaryotes. PLoS Genet. 2010; 6(6). ARTN e1001004 doi: 10.1371/journal.pgen.1001004 PMID: ISI:000279805200033.

55. Stoletzki N, Eyre-Walker A. Synonymous codon usage in Escherichia coli: selection for translational accuracy. Mol Biol Evol. 2007; 24(2):374–81. Epub 2006/11/15. doi: 10.1093/molbev/msl166 PMID: 17101719.

56. Kanaya S, Yamada Y, Kudo Y, Ikemura T. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene. 1999; 238(1):143–55. Epub 1999/11/26. PMID: 10570992.

57. Ikemura T. Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol. 1985; 2(1):13–34. Epub 1985/01/01. PMID: 3916708.

58. Ikemura T. Correlation between the Abundance of Escherichia-Coli Transfer-Rnas and the Occurrence of the Respective Codons in Its Protein Genes—a Proposal for a Synonymous Codon Choice That Is Optimal for the Escherichia-Coli Translational System. J Mol Biol. 1981; 151(3):389–409. doi: 10.1016/0022-2836(81)90003-6 PMID: ISI:A1981MJ92600003.

59. Cinlar E. Introduction to Stochastic Processes. New York, USA: Springer-Varlag; 1975.

60. Brémaud P. Markov Chains Gibbs Fields, Monte Carlo Simulation and Queues: Springer Verlag; 1998.

61. Norris J. Markov chains. Cambridge, United Kingdom: Cambridge University Press; 1998.

62. Felsenstein J. Inferring Phylogenies. Sunderland, MA: Sinauer Associates, Inc.; 2004.

63. Yang Z. Computational Molecular Evolution. New York: Oxford University Press; 2006.

64. Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res. 2014; 42(Database issue):D32–7. Epub 2013/11/13. doi: 10.1093/nar/gkt1030 PMID: 24217914; PubMed Central PMCID: PMC3965104.

65. Frank AC, Lobry JR. Oriloc: prediction of replication boundaries in unannotated bacterial chromosomes. Bioinformatics. 2000; 16(6):560–1. Epub 2000/09/12. PMID: 10980154.

66. Grantham R. Amino acid difference formula to help explain protein evolution. Science. 1974; 185 (4154):862–4. Epub 1974/09/06. PMID: 4843792.

67. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol. 1982; 157(1):105–32. Epub 1982/05/05. PMID: 7108955.

68. Miyata T, Miyazawa S, Yasunaga T. Two types of amino acid substitutions in protein evolution. J Mol Evol. 1979; 12(3):219–36. Epub 1979/03/15. PMID: 439147.

69. Mohana Rao JK. New scoring matrix for amino acid residue exchanges based on residue characteristic physical parameters. Int J Pept Protein Res. 1987; 29(2):276–81. Epub 1987/02/01. PMID: 3570667.

70. Woese CR. Evolution of the genetic code. Naturwissenschaften. 1973; 60(10):447–59. Epub 1973/10/01. PMID: 4588588.

71. Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M. AAindex: amino acid index database, progress report 2008. Nucleic Acids Res. 2008; 36(Database issue):D202–5. Epub 2007/11/14. doi: 10.1093/nar/gkm998 PMID: 17998252; PubMed Central PMCID: PMC2238890.

72. De Jong K, Fogel DB, Schwefel H- P. A history of evolutionary computation. In: Back T, Fogel D, Michalewicz Z, editors. Handbook of Evolutionary Computation: IOP Publishing Ltd and Oxford University Press; 1997. p. A2.3:1–12.

73. Michalewicz Z. Genetic algorithms + data structures = evolution programs. London, United Kingdom: Springer-Verlag; 1996.

74. Wakeley J. The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance. Trends Ecol Evol. 1996; 11(4):158–62. Epub 1996/04/01. PMID: 21237791.

75. Johnson KA. The kinetic and chemical mechanism of high-fidelity DNA polymerases. Biochim Biophys Acta. 2010; 1804(5):1041–8. Epub 2010/01/19. doi: 10.1016/j.bbapap.2010.01.006 PMID: 20079883; PubMed Central PMCID: PMC3047511.

76. Kumar D, Abdulovic AL, Viberg J, Nilsson AK, Kunkel TA, Chabes A. Mechanisms of mutagenesis in vivo due to imbalanced dNTP pools. Nucleic Acids Res. 2011; 39(4):1360–71. Epub 2010/10/22. doi: 10.1093/nar/gkq829 PMID: 20961955; PubMed Central PMCID: PMC3045583.

77. Waisertreiger IS, Liston VG, Menezes MR, Kim HM, Lobachev KS, Stepchenkova EI, et al. Modulation of mutagenesis in eukaryotes by DNA replication fork dynamics and quality of nucleotide pools. Environ Mol Mutagen. 2012; 53(9):699–724. Epub 2012/10/12. doi: 10.1002/em.21735 PMID: 23055184; PubMed Central PMCID: PMC3893020.

78. Deschavanne P, Filipski J. Correlation of GC content with replication timing and repair mechanisms in weakly expressed E.coli genes. Nucleic Acids Res. 1995; 23(8):1350–3. Epub 1995/04/25. PMID: 7753625; PubMed Central PMCID: PMC306860.

79. Strauss BS. The 'A rule' of mutagen specificity: a consequence of DNA polymerase bypass of non-instructional lesions? Bioessays. 1991; 13(2):79–84. Epub 1991/02/01. doi: 10.1002/bies.950130206 PMID: 2029269.

80. Ide H, Murayama H, Sakamoto S, Makino K, Honda K, Nakamuta H, et al. On the mechanism of preferential incorporation of dAMP at abasic sites in translesional DNA synthesis. Role of proofreading activity of DNA polymerase and thermodynamic characterization of model template-primers containing an abasic site. Nucleic Acids Res. 1995; 23(1):123–9. Epub 1995/01/11. PMID: 7870577; PubMed Central PMCID: PMC306639.

81. Pavlov YI, Rogozin IB, Galkin AP, Aksenova AY, Hanaoka F, Rada C, et al. Correlation of somatic hypermutation specificity and A-T base pair substitution errors by DNA polymerase eta during copying of a mouse immunoglobulin kappa light chain transgene. Proc Natl Acad Sci U S A. 2002; 99 (15):9954–9. Epub 2002/07/18. doi: 10.1073/pnas.152126799 PMID: 12119399; PubMed Central PMCID: PMC126606.

82. Choi JY, Lim S, Eoff RL, Guengerich FP. Kinetic analysis of base-pairing preference for nucleotide incorporation opposite template pyrimidines by human DNA polymerase iota. J Mol Biol. 2009; 389 (2):264–74. Epub 2009/04/21. doi: 10.1016/j.jmb.2009.04.023 PMID: 19376129; PubMed Central PMCID: PMC4010588.

83. Suzuki M, Yoshida S, Adman ET, Blank A, Loeb LA. Thermus aquaticus DNA polymerase I mutants with altered fidelity. Interacting mutations in the O-helix. J Biol Chem. 2000; 275(42):32728–35. Epub 2000/07/25. doi: 10.1074/jbc.M000097200 PMID: 10906120.

84. Pursell ZF, Isoz I, Lundstrom EB, Johansson E, Kunkel TA. Yeast DNA polymerase epsilon participates in leading-strand DNA replication. Science. 2007; 317(5834):127–30. Epub 2007/07/07. doi: 10.1126/science.1144067 PMID: 17615360; PubMed Central PMCID: PMC2233713.