JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 24, Number 11, 2017

Pp. 1089–1098
DOI: 10.1089/cmb.2017.0017

Representations of Search Spaces in the Problem of Mutational Pressure Optimization According to Protein-Coding Sequences

PAWEŁ BŁAŻEJ, MAŁGORZATA WNĘTRZAK,
MAŁGORZATA GRABIŃSKA, and PAWEŁ MACKIEWICZ

ABSTRACT

The proper representation of the search space is the fundamental step in every optimization task, because it has a decisive impact on the quality of potential solutions. In particular, this problem appears when the search spaces are nonstandard and complex, with the large number of candidate solutions that differ from classical forms usually investigated. One of such spaces is the set of continuous-time, homogenous, and stationary Markov processes. They are commonly used to describe biological phenomena, for example, mutations in DNA sequences and their evolution. Because of the complexity of these processes, the representation of their search space is not an easy task but it is important for effective solving of the biological problems. One of them is optimality of mutational pressure acting on protein-coding sequences. Therefore, we described three representations of the search spaces and proposed several specific evolutionary operators that are used in evolutionary-based optimization algorithms to solve the biological problem of mutational pressure optimality. In addition, we gave a general formula for the fitness function, which can be used to measure the quality of potential solutions. The structures of these solutions are based on two models of DNA evolution described by substitution-rate matrices, which are commonly used in phylogenetic analyzes. The proposed representations have been successfully utilized in various issues, and the obtained results are very interesting from a biological point of view. For example, they show that mutational pressures are, to some extent, optimized to minimize cost of amino acid substitutions in proteins.

Keywords: algorithms, DNA, evolutionary optimization, Markov processes, mutation, substitution rate matrix.

1. INTRODUCTION

THE PROBLEM OF NON-STANDARD SEARCH SPACES with numerous, complex, or multidimensional solutions occurs frequently in practice. In such cases, the optimal solutions can be extremely hard to find by analytical and exhaustive search methods. In this context, evolutionary-based algorithms appear to be a

Department of Genomics, Faculty of Biotechnology, University of Wrocław, Wrocław, Poland.

promising approach because of their elasticity and relatively weak restrictions (De Jong et al., 1997). However, these procedures require a well-defined potential candidate solution to describe evolutionary operators effectively. They also need a suitable representation of the fitness function, which is responsible for the evaluation of a given solution.

Such complex search spaces characterize many biological processes, which are inherently nonlinear and complicated. These phenomena are usually modeled by some realizations of stochastic processes. Therefore, the problem of finding an appropriate solution is, in fact, a question about properties of a particular stochastic process.

In this article, we show three representations of continuous-time, homogeneous, and stationary Markov processes, which are commonly used in description of mutations and evolution of DNA sequences. They constitute specific classes of stochastic processes, which were tested in empirical examples as search spaces in studies on the potential optimality of the mutation accumulation process (Błażej et al., 2013, 2015, 2017). We described in detail the technical difficulties that can appear in solving these problems. It is clear that the way of representation of the candidate solution or, more precisely, the definition of its structure and properties has a strong impact on the shape of the evolutionary operators and, in consequence, on the quality of potential solutions. Therefore, we described evolutionary operators, that is, mutation and crossover, which were adopted to the specificity of the proposed search spaces in an evolutionary-based algorithm to find theoretically optimal solutions. Finally, we gave a detailed overview of the structure and the properties of the fitness function, which can be used in the mutational pressure optimization problems. The presented methods can be useful in any optimization processes in which the solutions are described by homogeneous and stationary Markov processes with finite state space.

1.1. Mutations in DNA sequences

The mutational pressure is a process of introducing spontaneous changes, that is, mutations, into DNA sequences. It is an indispensable component of biological evolution, because together with selection it is responsible for genetic variation observed in all living organisms. These changes arise as a result of many factors such as radiation or chemicals and can be also introduced during the synthesis and repair of the DNA molecule.

The most deleterious are nonsense mutations, that is, changes of sense codons into stop codons, which lead to shortening the length of a gene product (protein). Consequences of mutations of sense codons into other sense codons depend on differences in the physicochemical properties (e.g., size, charge, hydrophobicity) of replaced amino acids. The more different these amino acids are, the more harmful their replacement is for the coded protein.

It is supposed that the mutations observed in real genomes are not only a result of a strict random process but also the coevolution between the mutational pressure and additional constraints that are imposed on gene expression and products by the process of natural selection and properties of the genetic code (Freeland and Hurst, 1998; Freeland et al., 2003; Archetti, 2004; Dudkiewicz et al., 2005; Najafabadi et al., 2005; Itzkovitz and Alon, 2007; Mackiewicz et al., 2008; Massey, 2015). The consequences of mutational pressure are ambiguous. From one point of view, most changes are deleterious and generate unwanted costs of their repairing (Kimura, 1967; Drake, 1991). Therefore, a tendency to decrease the mutation rate should exist in organisms. On the other hand, mutational pressure is crucial to generate a genetic variability that is necessary for quick adaptation of a given organism to a changing environment (Travis and Travis, 2002; Bedau and Packard, 2003; Denamur and Matic, 2006). Therefore, it seems reasonable to assume that mutational pressure evolves as a result of a specific trade-off between the accuracy to preserve genetic information in protein-coding genes and the requirements for adaptational flexibility (Radman et al., 1999; Sniegowski et al., 2000).

It is worth mentioning that the potential optimization of mutational pressure may be associated not only with the global mutational rate but also with the pattern of nucleotide substitutions, that is, the rates of change between one type of nucleotide and another. Therefore, it is interesting to find sets of these rates that are optimized to these two extremes, that is, minimizing changes in genes or maximizing their variation and comparing them with the empirical mutational matrices (Błażej et al., 2013, 2015). However, it is not an easy task. Although the process of DNA mutation is commonly described by a class of transition probability matrices, which represent stationary, homogenous, and continuous-time stochastic processes, they can be realized in a huge number of possibilities, even for a fixed nucleotide composition of DNA

sequence. That is why it is challenging to construct the search space of these solutions, especially for the optimality problem presented earlier.

1.2. Models of nucleotide substitutions

Mutations introducing spontaneous changes into DNA sequences are usually described by models that are based on the theory of continuous-time, homogeneous, and stationary Markov processes. It is assumed that mutational pressure is a realization of a four-state, continuous-time, homogeneous, and stationary Markov process. Such a process is uniquely defined by a substitution-rate matrix

$$Q = \{q_{ij}\}, \quad i, j \in \{A, T, G, C\}$$

and stationary distribution of nucleotides $\pi = (\pi_A, \pi_T, \pi_G, \pi_C)$. This representation is commonly used in the description of DNA sequence evolution. We applied two popular models describing such evolution (Tables 1 and 2). The first one was the generalized time-reversible model (GTR) (Lanave et al., 1984; Tavare, 1986), in which time reversibility of the Markov process means that the *detailed-balance* condition is fulfilled:

$$\pi_i q_{ij} = \pi_j q_{ji}, \ i \neq j.$$

In this case, π_i is the stationary probability of being in state i, that is, one of the four possible nucleotides, whereas q_{ij} denotes the rate of substitution from nucleotide i to nucleotide j. It should be noted that there is no biological reason to expect that the substitution process is reversible and the GTR model is used because of mathematical convenience and easiness of application (Felsenstein, 2004; Yang, 2006). Therefore, we also considered the general unrestricted (UNREST) model (Yang, 1994), which includes 12 different parameters (rates) and is the most general representation of the process of nucleotide substitutions with only one restriction on the stationary distribution π (Table 2).

These two models were used to describe three special classes of potential solutions that were denoted as: M_{GTRs}, M_{GTR}, and M_{UN}. The M_{GTRs} class is composed of the stochastic processes that are time-reversible with a fixed stationary distribution π and have a similar speed of convergence to the stationarity as the corresponding empirical processes evaluated from real data. M_{GTR} is a generalization of the M_{GTRs} class, because it contains all GTR processes with a fixed π and without any additional restrictions. Finally, the M_{UN} is composed of all UNREST type Markov processes with a fixed π. It is clear that the class of Markov processes generated according to the UNREST model is more general than the one generated by the GTR assumptions; therefore, the following property is fulfilled:

$$M_{GTRs} \subset M_{GTR} \subset M_{UN}.$$

We used M_{GTRs} and M_{GTR} classes as potential search spaces because of their mathematical convenience. Moreover, M_{GTRs} enabled us to compare the empirical data with a special class of GTR models assuming the same speed of convergence to the stationarity. Although these search spaces are sets of continuous-time, homogeneous, and stationary Markov processes, they require different methods to generate potential solutions, which is interesting from a computational point of view. They also need appropriate evolutionary operators, which have to take into account the specificity of the selected search spaces.

Table 1. The Nucleotide Substitution-Rate Matrix Q of the Generalized Time-Reversible Model

	A	T	G	C
A	—	$a\pi_T$	$b\pi_G$	$c\pi_C$
T	$a\pi_A$	—	$d\pi_G$	$e\pi_C$
G	$b\pi_A$	$d\pi_T$	—	$f\pi_C$
C	$c\pi_A$	$e\pi_T$	$f\pi_G$	—

Nucleotides in rows are substituted by nucleotides in columns. π_i is the stationary frequency of i nucleotide, for $i \in \{A, T, G, C\}$, whereas a to f are rate parameters.

Table 2. The Nucleotide Substitution-Rate Matrix Q of the UNREST Model

	A	T	G	C
A	—	q_{AT}	q_{AG}	q_{AC}
T	q_{TA}	—	q_{TG}	q_{TC}
G	q_{GA}	q_{GT}	—	q_{GC}
C	q_{CA}	q_{CT}	q_{CG}	—

Nucleotides in rows are substituted by nucleotides in columns.

q_{ij} is a substitution rate from nucleotide i to j. For fixed stationary distribution π, the equations $\pi Q = 0$ under the constraints $\sum_i \pi_i = 1$ are fulfilled.

UNREST, unrestricted.

2. METHODS OF GENERATING CANDIDATE SOLUTIONS

2.1. Generation of substitution-rate matrices in the M_{GTR} class

The GTR approach is commonly used to model the processes of nucleotide substitutions and assumes the time reversibility of Markov processes. The basic GTR model is based directly on the GTR substitution-rate matrix presented in Table 1. According to the theory of continuous-time Markov processes, we obtained that

$$M_{GTR} = \{\vec{s}_1 : \ \vec{s}_1 = (a, b, c, d, e, f), \ a, b, c, d, e, f > 0\} \tag{1}$$

is a class of six-dimensional vectors of parameters, which uniquely define a time-reversible process of nucleotide substitutions with a fixed stationary distribution π. It is evident that in this simple case, we can incorporate evolutionary operators, that is, mutation and crossover, which are commonly used in optimization problems in which the potential search space is a subset of n-dimensional Euclidean space. The mutation operator can be realized by a random shift of a vector $\vec{s}_1$ according to the normal distribution $N(0, \sigma)$. Moreover, M_{GTR} is a convex cone and therefore, we are able to adopt each crossover operator, which produces an offspring as a linear combination of its parents, with positive coefficients. For example, it is possible to use here a modified version of the Linear Crossover (Schlierkamp-Voosen and Muhlenbein, 1994).

2.2. Generation of substitution-rate matrices in the M_{GTRs} class

The M_{GTR} class contains many potential solutions that are represented by substitution-rate matrices with a fixed stationary distribution π and without any additional assumptions on eigenvalues. However, there may be a need to restrict the search space to a subset of matrices that are characterized by special properties, for example, a fixed speed of convergence to the stationary distribution. It may be useful to compare the reference empirical matrices characteristic for real genomes with the optimal matrices found in the search space (Błażej et al., 2015). In this case, the theoretical alternatives should possess not only the same nucleotide stationary distribution but also the same speed of convergence.

To solve this problem, it is necessary to construct a new representation of a candidate solution, in which we include some additional assumptions from the theory of Markov processes. First of all, we transform the substitution-rate matrix Q into a transition probability matrix $P = \{p_{ij}\}$ by adopting the uniformization method (Tijms, 2003). As a result, we obtain a matrix P that is defined in the following way:

$$p_{ij} = \begin{cases} \frac{q_{ij}}{q}, & i \neq j \\ 1 - \frac{|q_{ii}|}{q}, & i = j \end{cases} \tag{2}$$

where $q = \sum_{i \in A, T, G, C} |q_{ii}|$. In general, the uniformization procedure is used to transform the original continuous-time Markov process with nonidentical leaving rates into an equivalent of a stochastic process, in which the transition epoch is generated by a suitable Poisson process with a fixed rate. In consequence, we were able to apply the following representation of a discrete time-reversible Markov chain (Brémaud, 1998):

$$P = A \Lambda A^{-1}. \tag{3}$$

This is a unique spectral decomposition of the transition probability matrix P, where A is an orthogonal matrix, in which the rows consist of right eigenvectors, whereas A^{-1} is the transpose of A ($A^{-1}=A^T$) and its columns are composed of left eigenvectors; Λ is a diagonal matrix with real eigenvalues on its diagonal. It is clear that all eigenvalues are the solutions of the characteristic equation and have many interesting probabilistic interpretations. The maximum of the eigenvalues is equal to 1 and corresponds to the stationary distribution π, that is, the left eigenvector. Furthermore, the second largest eigenvalue gives an upper bound on the speed of convergence of the Markov process to the stationary distribution, generated by the transition probability matrix P, which is a direct consequence of the Perron-Frobenius theorem. These properties are sufficient to define a convenient representation of a candidate solution from the subclass of M_{GTR}. Therefore, we assume that every individual, that is, the transition probability matrix P is expressed by an equation:

$$P = A\Lambda A^T \Pi. \tag{4}$$

In this representation, we have that the matrix

$$A = \begin{bmatrix} 1 & x_1 & y_1 & z_1 \\ 1 & x_2 & y_2 & z_2 \\ 1 & x_3 & y_3 & z_4 \\ 1 & x_4 & y_4 & z_4 \end{bmatrix}. \tag{5}$$

It is a real-valued and an orthogonal matrix with three column vectors $\vec{x}=(x_1, x_2, x_3)$, $\vec{y}=(y_1, y_2, y_3)$, $\vec{z}=(z_1, z_2, z_3)$, whereas

$$\Lambda = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \lambda_3 & 0 \\ 0 & 0 & 0 & \lambda_4 \end{bmatrix} \tag{6}$$

is the matrix with eigenvalues on its diagonal, where the first and the second row are fixed. Π is a diagonal matrix with $\pi=(\pi_A, \pi_T, \pi_G, \pi_C)$ on its diagonal. Moreover, A is orthogonal in terms of the stationary distribution, that is, $A^{-1}=A^T\Pi$. It is evident that the Equation (4) is a special case of Equation (3) and is very useful in generating at random a sample of matrices with desired properties. As a consequence, it is possible to explore a specific subclass of time-reversible Markov processes with a fixed stationary distribution π, because each candidate solution could be expressed as a set of vectors:

$$M_{GTRs} = \{\{\vec{x}, \vec{y}, \vec{z}, (\lambda_2, \lambda_2, \lambda_3)\}\}, \tag{7}$$

where $(\lambda_2, \lambda_2, \lambda_3)$ is a vector of eigenvalues. The following assumptions on eigenvalues $(\lambda_2, \lambda_2, \lambda_3)$ were considered (Błażej et al., 2015): (a) all eigenvalues are exactly the same as in the reference process; (b) only the second eigenvalue is the same as in the reference process. These two constraints guarantee that the generated stochastic processes converge to the fixed stationary distribution with the same speed.

To generate at random a candidate solution, we need to create three independent eigenvectors $\vec{x}$, $\vec{y}$, $\vec{z}$. It can be realized by drawing three points from the unit sphere. Next, they are orthogonalized according to the Gramm-Schmidt orthogonalization procedure, whereas the three eigenvalues λ_1, λ_2, and λ_3 are generated at random according to one of the assumptions on the eigenvalues.

As a mutation operator, a random shift of eigenvectors $\vec{x}$,$\vec{y}$, and $\vec{z}$ can be applied and/or eigenvalues can be drawn from the normal distribution $N(0, \sigma)$. Every individual that is modified by the mutation operator must be orthogonalized again and it should be checked whether the structure of the candidate-solution [Eq. (3)] is preserved. However, in this case, it is not clear as to how to modify the representation [Eq. (3)] to effectively describe a crossover operator in such a way that selected individuals can exchange partial information.

2.3. Generation of substitution-rate matrices in the M_{UN} class

The time-reversibility assumption in the process of nucleotide substitution is generally accepted in many phylogenetic studies. It is assumed that this property is a good representation of the real substitution process and facilitates mathematical operations on matrices describing this process, although no biological

justification for the time-reversibility models was proposed (Felsenstein, 2004; Yang, 2006). However, some studies show that the time-reversibility causes a problem in modeling synonymous codon usage (Błażej et al., 2017). It implies that this assumption is not always a good approximation of biological processes. Therefore, it seems reasonable to develop a new procedure under more general assumptions, which will be an extension of methods presented in the previous sections.

The UNREST model is the most general, because it does not assume time reversibility in the process of nucleotide substitution, which is imposed on others. The representation of each candidate solution that belongs to the M_{UN} class is based on the structure of the substitution-rate matrix defined in Table 2. In addition, every stationary and continuous-time Markov process fulfills the following *balance equation* (Brémaud, 1998):

$$\pi Q = 0 \tag{8}$$

under the constraint

$$\sum_{i \in \{A, T, G, C\}} \pi_i = 1.$$

The fundamental step in this investigation is to reformulate [Eq. (8)] into the system of three equations:

$$V\beta^T = 0, \tag{9}$$

where

$$V^T = \begin{bmatrix} -\pi_A & \pi_A & 0 \\ -\pi_A & 0 & \pi_A \\ -\pi_A & 0 & 0 \\ \pi_T & -\pi_T & 0 \\ 0 & -\pi_T & \pi_T \\ 0 & -\pi_T & 0 \\ \pi_G & 0 & -\pi_G \\ 0 & \pi_G & -\pi_G \\ 0 & 0 & -\pi_G \\ \pi_C & 0 & 0 \\ 0 & \pi_C & 0 \\ 0 & 0 & -\pi_C \end{bmatrix}$$

and $\beta \in \mathbf{R}^{12}$ is composed of 12 substitution rates of the matrix Q, that is

$$\beta = [q_{AT}, q_{AG}, q_{AC}, q_{TA}, q_{TG}, q_{TC}, q_{GA}, q_{GT}, q_{GC}, q_{CA}, q_{CT}, q_{CG}].$$

The set of Equations (9) was obtained by reformulating (8) so that the variables from Equation (8), that is, π became factors in Equation (9). As a result, we got a homogeneous set of algebraic equations with infinitely many nontrivial solutions. Moreover, each potential solution is a linear combination of independent vectors (base) $v_1, v_2, \ldots, v_8, v_9 \in \mathbf{R}^{12}$ with coefficients $\beta_i,\ i = 1, 2, \ldots, 9$. In consequence, every solution of Equation (9) is of the form:

$$\beta = \beta_1 v_1 + \beta_2 v_2 + \ldots + \beta_9 v_9. \tag{10}$$

Therefore, β allows to create the matrix Q and thus, each potential candidate solution could be described in the following way:

$$M_{UN} = \{\vec{s} :\ \vec{s} = (\beta_1, \beta_2, \ \ldots, \beta_9), \}, \tag{11}$$

where the coefficients $\beta_1, \beta_2, \ldots, \beta_9$ guarantee a proper representation of the vector β. In other words, the condition:

$$q_{ij} > 0, i \neq j \tag{12}$$

must be fulfilled. Clearly, Equation (11) is a space of vector coefficients related to the solutions [Eq. (10)] of the Equation (9).

This general model of nucleotide substitutions needs specific evolutionary operators. From linear algebra, we get immediately that the solutions of Equation (9) constitute a vector space. Threrefore, the base coefficients also create a vector space. However, in this case, we have to check the condition [Eq. (12)], which implies that M_{UN} is just a subset of the whole set of potential coefficients. As a result, we could define the mutation operator as random changes in the vector of coefficients $\beta_i,\ i = 1, 2, \ldots, 9$. They are generated according to the normal distribution $N(0, \sigma)$. The crossover generator adopted to this problem can be a modified version of the Linear Crossover LBGA (Schlierkamp-Voosen and Muhlenbein, 1994). In short, it produces an offspring, which is a linear combination of its parents in terms of Equations (9) and (10). Clearly, it is necessary to check the quality of the newly produced offspring at the end of these procedures, particularly whether the rates [Eq. (12)] are positive.

2.4. The fitness function

The representations of search spaces described earlier can be used for finding a proper solution in various optimality problems. These problems require to define an appropriate fitness function F, which is necessary to compare the quality of potential solutions. In the case of the study on the optimality of mutational pressure, this function should combine several features of protein-coding sequences with properties of the process of nucleotide substitution. The cost of mutations in these sequences should take into account the potential differences between amino acids that are coded by mutated codons. The general form of the fitness function F can be given by the following formula:

$$F = \sum_{<k,\, l> \in C} p(k) p_{k \to l} g(k, l), \tag{13}$$

where C is the set of pairs of codons $< k, l >$, which differ in one codon position and $p(k)$ is a probability of selecting a given codon k. Moreover, $p_{k \to l}$ is a probability of transition from the codon k to l in one nucleotide substitution. This single change is generated by a uniformized transition probability matrix [Eq. (2)] calculated for fixed candidate solutions. Finally, $g(k, l)$ is a measure of differences between the properties of the amino acids coded by the codons k and l, respectively (which is called a selection factor). Various representations of the function g can be applied (Błażej et al., 2013, 2015), for example:

$$g(k, l) = \begin{cases} 0 & \textit{if } k \textit{ and } l \textit{ code the same amino acid} \\ 1 & \textit{if } k \textit{ and } l \textit{ code various amino acids.} \end{cases} \tag{14}$$

The case Equation (13) is, in fact, the sum of probabilities of nonsynonymous substitutions, which are calculated from the codon frequencies and a given mutational matrix. It is also possible to use:

$$g(k, l) = \begin{cases} 0 & \textit{if } k \textit{ and } l \textit{ code the same amino acid} \\ pr_{kl} & \textit{if } pr_{kl} \in [0, 1] \textit{ and } k \textit{ and } l \textit{ code various amino acids,} \end{cases} \tag{15}$$

where pr_{kl} is a probability that the transition from the codon k to l will be accepted by a selection process. Moreover, using a numerical description of amino acid properties, the following selection factor can be proposed:

$$g(k, l) = [A(k) - A(l)]^2, \tag{16}$$

which is a squared difference between the properties of the amino acids described by a function A and coded by the codons k and l, respectively. In consequence, under the assumption Equation (16), the function F has an interesting interpretation, because it is a mean value of amino acids substitution costs.

3. APPLICABILITY

The methods presented in this work can be successfully applied to many biologically inspired problems. The general representation M_{GTR} of a candidate solution can be used to solve the problem of finding stochastic mutational processes, which together with selection minimize or maximize the evolutionary cost of generated changes in protein-coding sequences (Błażej et al., 2013). The impact of these processes was investigated on real genes present in the genome of bacteria *Borrelia burgdorferii*. The optimal mutational

Table 3. The Transition Probability Matrix P, Which Is a Representation of the Optimal Mutational Pressure in Terms of Minimizing the Probability of Nonsynonymous Nucleotide Substitutions

	A	T	G	C
A	0.9978	0.0001	0.0005	0.0017
T	0.0000	0.8806	0.0000	0.1193
G	0.0011	0.0002	0.9982	0.0006
C	0.0091	0.9896	0.0013	0.0000

The selection strength was calculated according to the formula [Eq. (13)], in which codon frequencies p were evaluated from bacteria *Borrelia burgdorferii* genes and $g(k, l)$ function was of the form [Eq. (14)].

pressures were expressed as transition probability matrices (see Table 3 for example). They were found by a searching algorithm, which was based on the Evolutionary Strategies (ES) technique. Each candidate solution belonged to the class [Eq. (1)], whereas the mutation and crossover operators were fitted to the M_{GTR} representation.

In contrast to the general results presented (Błażej et al., 2013), the representation M_{GTRs} was used in more detailed comparisons between empirical processes and their corresponding optimized alternatives (Błażej et al., 2015). Seven different mutational pressures deduced for different bacterial genomes were studied (see Table 4 for example). Particularly, the representation [Eq. (7)] appeared very useful in the problem of searching for an optimal solution under additional assumptions on the speed of convergence to the stationarity. What is more, to evaluate the quality of a given solution, Błażej et al. (2015) applied different measures of amino acids properties. They were used to establish selected fitness functions in which the function $g(k, l)$ was of the form Equation (16). All the optimal solutions were found by using a searching algorithm based on the ES approach. This method worked on M_{GTRs} search space representation under the conditions defined in the previous sections. In Table 3, we presented an example of an optimal stochastic process, which was established by the searching algorithm. The described studies using M_{GTR} and M_{GTRs} representations showed that empirical mutational pressures in bacterial genomes are rather optimized to minimize cost of amino acid replacements, simultaneously allowing for some variation in the protein-coding sequences (Table 5).

The M_{UN} model is the most general representation of a stochastic process considered in this work. Therefore, it could be applied in wider aspects of possible optimization problems in comparison to the M_{GTR} and M_{GTRs} models. The M_{UN} representation was used by Błażej et al. (2017) to find nucleotide substitution matrices that maximized the differences in usage of synonymous codons together with the selection described by Equation (15). They showed that the M_{UN} model allows to include all possible mutation-selection effects acting on the synonymous codons usage in the protein-coding sequence, which is impossible under the M_{GTR} assumption.

Table 4. The Empirical Transition Probability Matrix Calculated for the Leading DNA Strand from *Borrelia burgdorferii* Genome (Kowalczuk et al., 2001)

	A	T	G	C
A	0.4924	0.2713	0.1762	0.0601
T	0.1731	0.6428	0.0918	0.0924
G	0.4323	0.3058	0.2231	0.0388
C	0.1855	0.6904	0.1241	0.0000

It describes the stationary stochastic process with stationary distribution: $\pi_A = 0.3167$, $\pi_T = 0.4876$, $\pi_G = 0.1370$, $\pi_C = 0.0588$.

Table 5. The Transition Probability Matrix Describing the Optimal Mutational Process, Acting on *Borrelia burgdorferii* Genome, in Which the Mean Cost of Amino Acid Substitutions Is Minimized

	A	*T*	*G*	*C*
A	0.77	0.15	0.08	0.00
T	0.09	0.87	0.01	0.02
G	0.19	0.05	0.72	0.03
C	0.00	0.15	0.08	0.77

This process is stationary with the same stationary distribution as in the empirical case. All eigenvalues are exactly the same as in the empirical case. The cost of amino acid substitution was calculated according to the polar property (Woese, 1973).

The derived representations of search spaces can be applied in problems, which are described by continuous-time, homogeneous, and stationary Markov processes. In particular, they can be used in finding nucleotide substitution matrices that fulfill appropriate properties, for example, minimize or maximize consequences of substitution or compositional differences at the nucleotide, codon, and amino acid levels.

AUTHOR DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Archetti, M. 2004. Codon usage bias and mutation constraints reduce the level of error minimization of the genetic code. *J. Mol. Evol*. 59, 258–266.

Bedau, M.A., and Packard, N.H. 2003. Evolution of evolvability via adaptation of mutation rates. *Biosystems* 69, 143–162.

Błażej, P., Mackiewicz, D., Wnętrzak, M., et al. 2017. The impact of selection at the amino acid level on the usage of synonymous codons. *G3* 7, 967–981.

Błażej, P., Mackiewicz, P., and Wańczyk, M. 2013. Using evolutionary algorithms in finding of optimized nucleotide substitution matrices, 41–42. In *Proceedings of the Genetic and Evolutionary Computation Conference, GECCO'13, Companion* Association for Computing Machinery (ACM), New York, NY. *978-1-4503-1964-5/13/07.*

Błażej, P., Miasojedow, B., Grabińska, M., et al. 2015. Optimization of mutation pressure in relation to properties of protein-coding sequences in bacterial genomes. *PLoS One*. 10, e0130411.

Brémaud, P. 1998. *Markov Chains Gibbs Fields, Monte Carlo Simulation and Queues*. Springer Verlag, New York, NY.

De Jong, K., David, D., Fogel, B., et al. 1997. A history of evolutionary computation, A2.3:1–12. *In* Back, T., Fogel, D., Michalewicz, Z., et al., eds. *Handbook of Evolutionary Computation*. Oxford University Press, Bristol, UK.

Denamur, E., and Matic, I. 2006. Evolution of mutation rates in bacteria. *Mol. Microbiol*. 60, 820–827.

Drake, J.W. 1991. A constant rate of spontaneous mutation in dna-based microbes. *Proc. Natl. Acad. Sci. U. S. A.* 88, 7160–7164.

Dudkiewicz, M., Mackiewicz, P., Nowicka, A., et al. 2005. Correspondence between mutation and selection pressure and the genetic code degeneracy in the gene evolution. *Future Gener. Comput. Syst*. 21, 1033–1039.

Felsenstein, J. 2004. *Inferring Phylogenies*. Sinauer Associates, Inc., Sunderland, MA.

Freeland, S.J., and Hurst, L.D. 1998. The genetic code is one in a million. *J. Mol. Evol.* 47, 238–248.

Freeland, S.J., Wu, T., and Keulmann, N. 2003. The case for an error minimizing standard genetic code. *Orig. Life Evol. Biosph.* 33, 457–477.

Itzkovitz, S., and Alon, U. 2007. The genetic code is nearly optimal for allowing additional information within protein-coding sequences. *Genome Res*. 17, 405–412.

Kimura, M. 1967. On evolutionary adjustment of spontaneous mutation rates. *Genet. Res*. 9:23–34.

Kowalczuk, M., Mackiewicz, P., Mackiewicz, D., et al. 2001. High correlation between the turnover of nucleotides under mutational pressure and the DNA composition. *BMC Evol. Biol.* 1:13.

Lanave, C., Preparata, G., Saccone, C., et al. 1984. A new method for calculating evolutionary substitution rates. *J. Mol. Evol*. 20, 86–93.

Mackiewicz, P., Biecek, P., Mackiewicz, D., et al. 2008. Optimisation of asymmetric mutational pressure and selection pressure around the universal genetic code, 100–109. *In* Bubak, M., Dongarra, J., VanAlbada, G.D., and Sloot, P.M.A., eds. *Computational Science—ICCS 2008, PT 3*, volume 5103 of *Lecture Notes in Computer Science*. Elsevier; Springer. Berlin, Heidelberg.

Massey, S.E. 2015. Genetic code evolution reveals the neutral emergence of mutational robustness, and information as an evolutionary constraint. *Life*. 5, 1301–1332.

Najafabadi, H.S., Goodarzi, H., and Torabi, N. 2005. Optimality of codon usage in escherichia coli due to load minimization. *J. Theor. Biol*. 237, 203–209.

Radman, M., Matic, I., and Taddei, F. 1999. Evolution of evolvability, 146–155. *In* Caporale, L.H., ed. *Molecular Strategies in Biological Evolution*, volume 870 of Annals of the New York Academy of Sciences, Whiley, New York, NY.

Schlierkamp-Voosen, D., and Muhlenbein, H. 1994. Strategy adaptation by competing subpopulations, 199–208. In *Proceedings Parallel Problem Solving from Nature III*. Springer; Berlin, Heidelberg.

Sniegowski, P., Gerrish, P., Johnson, T., et al. 2000. The evolution of mutation rates: Separating causes from consequences. *Bioessays*. 22, 1057–1066.

Tavare, S. 1986. Some probabilistic and statistical problems of the analysis of DNA sequences. *Lect. Math. Life Sci.* 17, 57–86.

Tijms, H. 2003. *A First Course in Stochastic Processes*. John Wiley & Sons LTD, England.

Travis, J., and Travis, E. 2002. Mutator dynamics in fluctuating environments. *Proc. R. Soc. Lond. B Biol. Sci.* 269, 591–597.

Woese, C.R. 1973. Evolution of the genetic code. *Naturwissenschaften* 60, 447–459.

Yang, Z. 1994. Estimating the pattern of nucleotide substitution. *J. Mol. Evol*. 39, 105–111.

Yang, Z. 2006. *Computational Molecular Evolution*. Oxford University Press, New York, NY.

Address correspondence to:
Dr. Paweł Błażej
Department of Genomics
Faculty of Biotechnology
University of Wrocław
Ul. F. Joliot-Curie 14a
Wrocław 50-383
Poland

E-mail: blazej@smorfland.uni.wroc.pl