The impact of selection at the amino acid level on the usage of synonymous codons

Paweł Błażej, Dorota Mackiewicz, Małgorzata WNętrzak, Paweł Mackiewicz*
Department of Genomics, Faculty of Biotechnology, University of Wrocław, ul. Joliot-Curie 14a, 50-383, Wrocław, Poland
*Corresponding author: pamac@smorfland.uni.wroc.pl

Abstract

There are two main forces that affect usage of synonymous codons: directional mutational pressure and selection. The effectiveness of protein translation is usually considered as the main selectional factor. However, the biased codon usage can be also a by-product of a general selection at the amino acid level interacting with nucleotide replacements. To evaluate the validity and strength of such effect, we superimposed more than 3.5 billion unrestricted mutational processes on the selection of non-synonymous substitutions based on the differences in physicochemical properties of the coded amino acids. Using a modified evolutionary optimization algorithm, we determined the conditions in which the effect on the relative codon usage is maximized. We found that the effect is enhanced by mutational processes generating more adenine and thymine than guanine and cytosine as well as more purines than pyrimidines. Interestingly, this effect is observed only under an unrestricted model of nucleotide substitution and disappears when the mutational process is time-reversible. The comparison of the simulation results with data for real protein coding sequences indicates that the impact of selection at the amino acid level on the synonymous codon usage cannot be neglected. Furthermore, it can considerably interfere, especially in AT-rich genomes, with other selections on the codon usage, e.g. translational efficiency. It may also lead to the difficulties in the recognition of other effects influencing the codon bias and an overestimation of the protein coding sequences whose codon usage is subjected to adaptational selection.

Introduction

Redundancy of the genetic code implies that there are more codons than amino acids. Consequently, many amino acids are encoded by more than one codon called synonymous. As a result, some substitutions between these codons are silent and do not change the coded amino acid. For example, in the case of the codons named four-fold degenerated (4FD), their third codon positions can be freely changed to any nucleotide, without consequences for the coded amino acid, and subsequently for protein composition and function. However, the synonymous codons are not uniformly used in real protein coding sequences, e.g. (COMERON 2004; GRANTHAM et al. 1980; IKEMURA 1985; PLOTKIN and KUDLA 2011; SHARP and LI 1986). Such preference of one synonymous codon over others is commonly known as codon usage bias (SHARP and LI 1986). The usage can be different for various genomes and genes within one genome and even within a single gene.
As far as the evolution of codon bias is concerned, two explanations, which are not mutually exclusive, have been proposed: directional mutations and specific selection (BULMER 1991; HERSBERG and PETROV 2008). From the mutational point of view, GC content is the strongest single determinant of codon usage in genomes (CHEN et al. 2004; ERMOLAeva 2001; KNIGHT et al. 2001; LI et al. 2015; MUTO and OSAWA 1987). Thus, in genomes with a high average GC content, the most frequent synonymous codons typically end with guanine or cytosine, whereas in genomes with a low average GC content, they usually have adenine or thymine in their silent positions. The GC content also fluctuates periodically along vertebrate chromosomes, creating an isochore structure and influencing the local codon usage of genes (CHEN et al. 2004; FEDOROV et al. 2002). In prokaryotic genomes, the chromosome-wide codon bias is related with various mutational pressures acting on differently replicating DNA strands, leading and lagging (e.g., FRANK and LOBRY 1999; LOBRY 1996; MACKIEWICZ et al. 1999b; MACKIEWICZ et al. 1999c; MORTON and MORTON 2007; MRAZEK and KARLIN 1998; ROCHA et al. 1999; TILLIER and COLLINS 2000). As a result, GT-rich codons are usually overrepresented in the leading strand genes, whereas AC-codons in the lagging strand. The codon biases characteristic of genomes enable the identification of potential genes that were horizontally transferred (GARCIA-VALLVE et al. 2003).

The bias resulting from the mutational effects can be modified by many selectional factors. The first reported influence of the selection on the codon bias was based on the observation that highly expressed sequences tend to use generally more frequent codons (e.g., AKASHI 2003; BENNETZEN and HALL 1982; CLARKE 1970; DURET and MOUCHIroud 1999; GHAEMMAGhamI et al. 2003; GOETZ and FUGLSANG 2005; IKemura 1981; IKemura 1985; KANAYA et al. 1999; MORTON 1998; ROCHA 2004). It was interpreted as an adaptation to the effectiveness of translation process and accuracy of protein synthesis, and is called as codon adaptation or translational selection on codon usage. In addition, a substantial coincidence between the gene copy number and the frequency of codons with a concentration of tRNA isoacceptors with their complementary anticodons was detected. The abundant tRNA isoacceptors, through their more fluent recognition of frequently used codons, enable the processivity of translational elongation (KANAYA et al. 1999; Xia 1998). Accordingly, there is a significant positive correlation between the gene expression level and the codon bias, and likewise, a negative correlation between the gene expression level and the rate of synonymous substitutions between the compared sequences (EYRE-WALKER and BULMER 1995; SHARP and Li 1987). The selection on synonymous codon usage can also result from selection for translational accuracy to reduce the costs of both missense and nonsense errors (STOLETZKI and EYRE-WALKER 2007). The effectiveness of translation is also enhanced by clustering some synonymous codons in the highly expressed genes, which is called a codon co-occurrence bias (CANNARROZZI et al. 2010; SHAO et al. 2012; ZHANG et al. 2013).

However, the analyses of many genomes show that there is a fraction of genes that show no evidence for the translational selection linked to codon usage (CARBONE and MADDEN 2005; DOS REIS et al. 2004; SHARP et al. 2005). This observation is not supported by the recent, multi-genome studies indicating that the translational selection for codon usage seems universal at least in prokaryotes (HERSBERG and PETROV 2008; SUPEK et al. 2010) and plastids (SUZUKI and MORTON 2016). On the other hand, recently developed
techniques measuring endogenous expression showed that the initiation rather than the elongation process limits the rate of protein production for most endogenous genes (INGOLIA 2014; INGOLIA et al. 2009; KERTESZ et al. 2010; TULLER et al. 2010).

Although the translational selection is thought to be the dominant explanation for systematic variation in the codon usage among genes (CHANEY and CLARK 2015; PLOTKIN and KUDLA 2011; QUAX et al. 2015), several other factors related to codon bias were put forward. One of them is the formation of the functional native structure of proteins, which is realised by the preference of common codons in regions critical for protein folding and structure (ORESIC and SHALLOWAY 1998; PECHMANN and FRYDMAN 2013; THANARAJ and ARGOS 1996; ZHOU et al. 2009). Furthermore, the bias in synonymous codon usage within coding sequence is also thought to be an additional layer of information influencing the stability of mRNA structure (BARTOSZEWSKI et al. 2010; LAZRAK et al. 2013), mRNA half-live (PRESNYAK et al. 2015) and the effectiveness of transcription (XIA 1996).

It was initially postulated that the enrichment of 5’ end of coding sequences in rare codons is intended to create a ramp at the 5’ end, which prevents ribosome traffic jams further down the length of the mRNA and increases translational efficiency (TULLER et al. 2010). Other authors proposed that the rare codons cause a translational pause, which helps targeting and export of secreted proteins (CLARKE and CLARK 2010; ZALUCKI et al. 2009). The ramp concept was revised in further studies showing that the reduced formation of stable mRNA structure is rather responsible for the higher translation rate (BENTELE et al. 2013; GOODMAN et al. 2013; KUDLA et al. 2009), whereas a computational model predicted that this ramp is caused by rapid initiation of short genes rather than rare codons at the 5’ end of transcripts (SHAH et al. 2013).

The presence of many selective constraints on the codon usage has consequences on the slower synonymous substitution rate of genes subjected to these selections, which is demonstrated by an inverse correlation between the rate and the degree of codon adaptation (MORTON et al. 2002; SHARP and Li 1986; SHARP et al. 1989; SHIELDS et al. 1988; SORHANNUS and FOX 1999). The understanding of the rules in codon usage is also important in order to better optimize heterologous gene expression (GUSTAFSSON et al. 2004), produce vaccines with attenuated viruses (COLEMAN et al. 2008) or find the association of diseases with synonymous single nucleotide polymorphism (DAIDONE et al. 2011; KIMCHI-SARFATY 2011; SAUNA and KIMCHI-SARFATY 2011). Therefore, it is still important to better recognize mechanisms that induce codon usage biases in nature, understand how the codon landscape evolves in time, and search for other factors affecting codon bias.

Morton postulated in his seminal work (MORTON 2001) another important factor influencing the synonymous codon usage. Interestingly, it is not related with the direct adaptational selection on the codon usage but results only from a general selection at the amino acid level of protein coding sequences. His analyses showed that the composition of their silent sites deviates from the composition of the noncoding (neutral) sites even in the absence of selective differences between synonymous codons. It results from various probabilities of fixation of codon replacements. He nicely demonstrated that, after considering this type of selection, there are far fewer genes with codon adaptation bias than it was previously thought. It
implicates that the selection on the codon usage associated with the translation efficiency may be overestimated. However, the study considered only four selected mutational processes generating equal frequencies of complementary nucleotides.

To further explore this subject, we created a mutation-selection model that includes the most general and unrestricted model of nucleotide substitutions and examines the large number of possible mutation processes generating almost 90,000 stationary distributions of codons. We applied an adapted version of the evolutionary optimization algorithm to find such conditions in which the mutation processes together with the selection at the amino acid level maximizes the degree of codon bias (deviation from the uniform codon usage). The results demonstrate that the effect under study cannot be neglected.

Methods

Overview

One of the best ways to assess the influence of selection, at the amino acid level, on synonymous codon usage is to compare the codon frequency resulting from the selection with the expected frequency without this constraint. To achieve that we constructed a mutation-selection model similar to (MORTON 2001). This model is based on the theory of homogeneous and continuous-time Markov processes. In contrast to the previous study, we examined the most general nucleotide substitution model. It was superimposed on the codon selection process associated with physicochemical properties of coded amino acids. Moreover, we tested almost 90,000 stationary distributions of codons which correspond to the mutational process. The effect of selection based on the differences between relative codon usage was measured before and after the applied selection. Since a given nucleotide stationary distribution can be realized by many Markov processes, we applied an evolutionary-based optimization algorithm in order to find conditions in which the differences between relative codon usage are maximized. As a result, we were able to determine the effect produced by the model with amino acid selection on synonymous sites. Below we presented in detail the stages of this approach presented in Supplemental Material Figure S1. Finally, the theoretical calculations were compared with the results provided by the bacterial genome analyses.

Mutation Process

To model the process of pure mutational pressure expressed by single nucleotide substitutions, we applied the homogeneous, stationary and continuous-time Markov process. The process is described by a substitution rate matrix $Q$ and stationary distribution of nucleotides $\pi$. This approach is most commonly used in the description of DNA sequence evolution (YANG 2006). Here, we used the most general unrestricted model of nucleotide substitution called UNREST (Table 1) (YANG 1994).
Table 1. Substitution rate matrix $Q$ for the unrestricted model of nucleotide substitutions (UNREST). The diagonals of $Q$ are determined by the requirement that each row sum to 0. The nucleotide stationary distribution $\pi = (\pi_A, \pi_T, \pi_G, \pi_C)$ is given by the set of equations $\pi Q = 0$ under the constraint $\sum_{i=A,T,G,C} \pi_i = 1$.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>T</th>
<th>G</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>$q_{AT}$</td>
<td>$q_{AG}$</td>
<td>$q_{AC}$</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>$q_{TA}$</td>
<td>$q_{TG}$</td>
<td>$q_{TC}$</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>$q_{GA}$</td>
<td>$q_{GT}$</td>
<td>$q_{GC}$</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>$q_{CA}$</td>
<td>$q_{CT}$</td>
<td>$q_{CG}$</td>
<td></td>
</tr>
</tbody>
</table>

The assumption about the stationarity of this process implies immediately the need to determine its exact stationary distribution, which should correspond in this case to the stationary frequency of nucleotides generated by the mutational process. In his work on this topic, Morton considered only four selected mutational processes with fixed nucleotide stationary distributions characterized by high A+T content (MORTON 2001). To formulate more general conclusions and assess the influence of the selection, at the amino acid level, on various mutational processes, we analysed 88,560 different nucleotide stationary distributions for the potential processes, which cover various mutational pressures. The frequencies of particular nucleotides range from 0.05 to 0.85 with the 0.01 increment. Therefore, they form the set:

$$\Pi = \{ \pi : \pi = (\pi_A, \pi_T, \pi_G, \pi_C), \sum_{i=A,T,G,C} \pi_i = 1 \}$$

where every $\pi \in \Pi$ is chosen according to the following assumptions:

1. $0.05 \leq \pi_i \leq 0.85$, $i \in \{A,T,G,C\}$;

2. for every $\pi \in \Pi$, the Euclidean distance to the nearest stationary distribution $\pi' \in \Pi$ is equal to 0.02 and the difference between $\pi'$ and $\pi$ in one coordinate is equal 0.01 or 0.

As a result, we obtained a dense subset of all possible nucleotide stationary distributions.

To find the rates of the matrix $Q$ for the particular distributions, we rested on the assumption that for the homogeneous, continuous-time and stationary Markov process the following set of equations holds:

$$\pi Q = 0.$$  \hfill (2)

These equations can be easily reformulated into the system of three equations:

$$V \beta^T = 0,$$  \hfill (3)

where:

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

and
\( \beta \in \mathbb{R}^{12} \) is composed of 12 substitution rates of matrix \( Q \):

\[
\beta = [q_{AT}, q_{AG}, q_{AC}, q_{TA}, q_{TG}, q_{TC}, q_{GA}, q_{GT}, q_{GC}, q_{CA}, q_{CT}, q_{CG}] \text{ under the constraint:}
\]

\[
\forall_{i \neq j} q_{ij} > 0, i, j \in \{A,T,G,C\},
\]

which is necessary to create homogeneous, continuous-time Markov processes with fixed stationary distribution.

The set of equations (3) has infinitely many nontrivial solutions. Moreover, each solution can be described by a linear combination of independent vectors \( v_1, v_2, \ldots, v_9 \in \mathbb{R}^{12} \) with coefficients \( \beta_i, i = 1, 2, \ldots, 9 \):

\[
\beta = \beta_1 v_1 + \beta_2 v_2 + \ldots + \beta_9 v_9.
\]

The \( \beta \) allows to create the matrix \( Q \), from which we derived a nucleotide transition probability matrix \( P \) by adopting the uniformization method (JENSEN 1953; TIJMS 2003). Generally, the uniformization procedure is used to transform the original continuous-time Markov process with non-identical leaving rates into an equivalent of stochastic process, in which the transition epoch is generated by a suitable Poisson process with a fixed rate. Following this method, for a given \( Q \) with stationary distribution \( \pi \), we could define a transition probability matrix \( P = (p_{ij}), i, j = A,T,G,C \) assuming that:

\[
p_{ij} = \begin{cases} 
    \frac{q_{ij}}{q}, & i \neq j; \\
    1 - \frac{|q_{ii}|}{q}, & i = j,
\end{cases}
\]

where \( q = \sum_{i=A,T,G,C} |q_{ii}| \). Clearly, \( P \) is the transition probability matrix describing Markov chain with the stationary distribution \( \pi \), which is the same for the continuous case. Moreover, the sum of all its off-diagonal elements is equal to 1. This representation turned out to be very useful in our mutation-selection model because it allowed the construction of quite a large set of possible nucleotide transition probability matrices \( P \) (Figure S1) under relatively weak mathematical assumptions.

Obviously, we are interested in a codon substitution process, and for this reason we calculated a codon \( k \) to codon \( l \) transition probability matrix \( P^* = (p^*_{k\rightarrow l}) \), using the nucleotide transition probability matrix \( P \) (Figure S1). In the matrix \( P^* \), we took into account all independent substitutions between codons resulting from a single nucleotide change. The Markov chain defined by \( P^* \) is also stationary with codon stationary distribution \( \pi^{\text{cod}} \) which is in accordance with the following system of equations:

\[
\pi^{\text{cod}} (I - P^*) = 0.
\]

Moreover, under assumptions presented above, the stationary relative frequencies of four-fold degenerated codons (4FD) are determined solely by the stationary distribution \( \pi \).
Process of selection

Similarly to (MORTON 2001), we were interested in a model of sequence evolution which combines mutation and selection at the amino acid level. At the selection stage, we introduced the acceptance matrix $D = (d_{m \to n})$, which contains probabilities that a change of amino acid $m$ to amino acid $n$ will be “accepted”. All diagonals of matrix $D$ are equal to 1, which means no selective costs of substitutions between synonymous codons. In this work, we employed the acceptance matrix presented in (MORTON 2001), which is based on Grantham’s chemical similarity matrix (GRANTHAM 1974) (Figure S1). Additionally, we took into account two cases involving substitutions to and from stop codons. In the first one, we assumed that such mutations are lethal (SL) and we set the probability of acceptance to 0. In the second case, the probability of acceptance of such substitutions was equal to the minimal probability in the matrix $D$ (SM).

As a consequence, we defined a general model, including the mutation and selection processes, in the same way Morton did (MORTON 2001). This model is expounded by a codon to codon transition probability matrix $C = (c_{k \to l})$. Furthermore, every codon to codon substitution $c_{k \to l}$ is defined by the following equation:

$$c_{k \to l} = p^*_{k \to l} \times d_{m \to n},$$

where $p^*_{k \to l}$ is the transition probability between codons $k$ and $l$, whereas $d_{m \to n}$ is the probability of accepting a change from amino acid $m$ to amino acid $n$ coded by codons $k$ and $l$, respectively. Obviously, all diagonals in matrix $C$ are set to make the rows sum to 1. In addition, the Markov chain described by $C$ has its own stationary distribution $\pi_{sel}$.

Measure of selection strength

The strength of selection at the amino acid level, which affects the composition in neutral sites of codons, was assessed for each stationary distribution $\pi$ by the normalized difference between the relative frequency of four-fold degenerated (4FD) codons after selection and their expected frequency resulting only from a mutation process:

$$F_{\pi(i)} = \sum_{i=A,T,G,C} \frac{\pi_{sel} - \pi_{i}^{\text{sel}}}{\pi_{i}},$$

where $s$ means a group of 4FD codons coding for one amino acid; $s_i$ is a codon from this group, in which a nucleotide $i$ occurs at the third position; $\pi_{sel}^{s_i} = \sum_{i=A,T,G,C} \pi_{sel}^{s_i}$ is the stationary frequency of this codon group after the selection (sel); $\pi_{s_i}^{\text{sel}}$ is the stationary frequency of the codon $s_i$ after the selection; $\pi$ is the relative stationary frequency of the codon $s_i$ obtained only from the mutation process. We analysed all five groups of 4FD codons and calculated the summarized effect of the selection at the amino acid level on these groups for each stationary distribution $\pi$. 7
\[
F_{\pi} = \sum_{s \in S} F_{\pi s},
\]

where \( S \) is the set of all groups of four-fold degenerated codons \( s \). Clearly, large values of \( F_{\pi} \) suggest a strong impact of selection on the usage of 4FD codons, whereas values equal to 0 indicate a lack of such effect on the relative frequencies of 4FD codons.

**Simulation procedure**

A nucleotide stationary distribution \( \pi \) can be realized by many Markov processes described by various substitution matrices \( P_{\pi} \), which can imply differences in stationary frequencies of codons after the selection \( \pi^{sel} \) and consequently different \( F_{\pi} \) values. To deal with this problem, we decided to find such probability \( F_{\pi}^{\text{max}} \) that maximize the \( F_{\pi} \) measure. The maximum value of \( F_{\pi} \) was denoted by \( F_{\pi}^{\text{max}} \). Thanks to that, we were able to assess the range of the selection strength at the amino acid level on synonymous codon usage for a given nucleotide stationary distribution \( \pi \).

The task of finding \( F_{\pi}^{\text{max}} \) is in fact an example of a single objective optimization problem, where \( F_{\pi} \) is a fitness function. Therefore, we decided to use Evolutionary Strategies (ES) approach (DE JONG et al. 1997), which is a commonly used technique in optimization problems when the solution is hard to find analytically. For each nucleotide distribution \( \pi \), we ran simulations with the population of 100 random candidate solutions according to ES principles. At the beginning of each simulation run, our candidate solutions were in fact substitution rate matrices \( Q \) selected at random according to the procedure described by equations (3) and the condition (4). In every simulation step, we applied mutation and selection operators. For a given rate matrix (individual), the process of mutation was realized by a random modification of its vector of coefficients \( \beta_i, i = 1, 2, \ldots, 9 \) according to the normal distribution \( N(0, \sigma) \) (Figure S1). The \( \sigma \) parameter was tuned during preliminary simulation tests to obtain a quick convergence to the satisfactory solution. The crossover operator used in this problem was a modified version of Linear Crossover LBGA (SCHLIERKAMP-VOOSEN and MUHLENBEIN 1994). It produced an offspring, which was a random linear combination of its parents in terms of equation (3) and (5). Understandably, at the end of these procedures, we checked the quality of newly produced offspring, to find out whether they possess a proper representation and fulfill the condition (4). In the next step, we made transformations of substitution nucleotide matrices \( Q \) to \( P \) and next to substitution codon matrices \( P^* \) and \( C \). This was done according to the procedure described in the previous sections. Therefore, we were able to calculate the codon stationary distribution after selection \( \pi^{sel} \) and values of the fitness function \( F_{\pi} \). Finally, we used the tournament selection as selection operator. Depending on the assumed fitness function, the algorithm selected individuals (rate matrices) that maximized the measure \( F_{\pi} \). The main program was developed by the author in C++ language. The stationary vectors \( \pi^{sel} \) were calculated using the Armadillo library (SANDERSON 2010).
Analysis of deviation in codon usage in protein coding sequences

The values determined for the $F_{\pi}^{\text{max}}$ measure were compared with an analogous parameter calculated for 4FD codons, using protein coding sequences from 4,879 fully sequenced bacterial genomes, whose sequences and annotations were downloaded from NCBI database (ftp://ftp.ncbi.nlm.nih.gov/genomes). We examined separately the genes located on the leading and the lagging DNA strands. The boundaries between the DNA strands were determined according to DNA walk methods, using DNA asymmetry parameters, i.e. the differences in complementary nucleotides: [G-C] and [A-T] (Kowalczik et al. 2001; Mackiewicz et al. 1999a). For these data, we calculated the summarized deviation from the expectation in the codon usage for all 4FD groups $S$. Clearly, it corresponds to the equations (9) and (10), i.e.:

$$F = \sum_{s \in S} f_s,$$

where:

$$f_s = \sum_{i \in \{A,T,G,C\}} \frac{e_i - o_s}{o_s} e_i$$

is the normalized difference between the relative frequencies of 4FD codons in one group and their expected frequencies. Therefore, $o_s$ is the observed frequency of a 4FD codon $s$ with a nucleotide $i$ at the third position, $o_s = \sum_{i \in \{A,T,G,C\}} o_s$ is the frequency of all codons in the group and $e_i$ is the expected frequency established as the average of relative frequencies of all 4FD codons with a nucleotide $i$ at the third codon position.

Data availability

The supplemental material includes two figures. Figure S1 illustrates the procedure leading to the assessment of the selection strength, at the amino acid level and Figure S2 compares nucleotide substitution probabilities for 5% of top matrices that maximized the values of $F_{\pi}$.

Results and discussion

The summarized effect of the selection on all codon groups

In total, we performed 88,560 simulations to find the maximum $F_{\pi}^{\text{max}}$ values for the normalized difference between the relative frequency of four-fold degenerated (4FD) codons after selection on amino acids and their expected frequency triggered only by a mutation process. This parameter expresses the most extreme impact of the selection at the amino acid level on the usage of 4FD codons. This impact can be found for a given mutation process with its specific nucleotide stationary distribution $\pi$. We studied
separately the selection with lethal stop codons’ substitutions (SL) and the variant with the minimal acceptance probability of such substitutions (SM).

The applied optimization algorithm enabled an effective maximization of the fitness function $F_{\pi}$ for all the nucleotide stationary distributions $\pi$ under study. Our results indicate that it is possible to find transition probability matrices $P_{\pi}^{\text{max}}$ for each nucleotide stationary distribution $\pi$, which maximize the impact of such a selection measured by $F_{\pi}^{\text{max}}$. Generally, depending on the applied stationary distribution $\pi$, $F_{\pi}^{\text{max}}$ varied from 0.25 to over 9.22 under SL variant and from 0.26 to over 9.1 under SM variant.

The nucleotide stationary distributions for which the extreme values of $F_{\pi}^{\text{max}}$ were found, are presented in Table 2. The findings indicate that the largest impact of selection at the amino acid level on deviations in the 4FD codons usage is for mutation processes that generate thymine with the high frequency at the expense of guanine and cytosine. The next frequent nucleotide is adenine. On the other hand, the smallest $F_{\pi}^{\text{max}}$ is for nucleotide distributions with high content of cytosine and next guanine. To systematically analyse the relationship of $F_{\pi}^{\text{max}}$ from the nucleotide stationary distributions, we carried out additional studies. Since results for models SL and SM were very similar, we focused on the latter.

Table 2. Nucleotide stationary distributions of mutation process for which the extreme $F_{\pi}^{\text{max}}$ values were found. Models with the selection assuming lethal substitutions involving stop codons (SL) and the variant with the minimal acceptance probability of such substitutions (SM) were considered, separately.

<table>
<thead>
<tr>
<th>Selection model</th>
<th>$F_{\pi}^{\text{max}}$</th>
<th>A</th>
<th>T</th>
<th>G</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>SL</td>
<td>9.22</td>
<td>0.15</td>
<td>0.74</td>
<td>0.05</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>0.25</td>
<td>0.05</td>
<td>0.07</td>
<td>0.13</td>
<td>0.75</td>
</tr>
<tr>
<td>SM</td>
<td>9.10</td>
<td>0.19</td>
<td>0.69</td>
<td>0.05</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>0.26</td>
<td>0.05</td>
<td>0.08</td>
<td>0.22</td>
<td>0.65</td>
</tr>
</tbody>
</table>

The results for the extreme values are supported by the radar chart, in which two sets of 100 stationary distributions responsible for the highest and the lowest values of $F_{\pi}^{\text{max}}$ are presented (Figure 1). It is visible that the $F_{\pi}^{\text{max}}$ value is clearly related to the frequency of nucleotides in the stationary distribution. The excess of thymine and next adenine leads to the highest values of $F_{\pi}^{\text{max}}$, whereas the lowest values of $F_{\pi}^{\text{max}}$ are observed for the domination of cytosine and next guanine.

In order to study how the possible maximum deviation in the usage of 4FD codons $F_{\pi}^{\text{max}}$ depends on the whole range of nucleotide stationary distributions in the combination of two nucleotides, we made the Wafer map, in which the gradient colouring corresponds to the $F_{\pi}^{\text{max}}$ value (Figure 2). Dark green denotes the lowest values and dark brown the highest values of $F_{\pi}^{\text{max}}$. The relationships are clearly non-linear. The highest values are observed for the distributions with the high frequency of thymine and substantially smaller contribution of other nucleotides, especially for $\pi_T > 0.6$ and $\pi_A$ in the range from 0 to 0.25 as well as for $\pi_T >$
0.7 and \( \pi_G < 0.2 \) or \( \pi_C < 0.2 \). The increase in \( F_{x}^{\text{max}} \) also correlates with the high frequency of adenine \( \pi_A > 0.7 \), but only together with the decline of guanine and cytosine frequencies to values lower than 0.2. However, there is a growth of \( F_{x}^{\text{max}} \) also for \( \pi_i \) from 0.2 to 0.4 with the excess of guanine in the range 0.5-0.7.

\[ \pi_A = 0.95 \quad \pi_A = \sum_{k \in K} \pi^A_k, \]

where \( \pi^A_k = \{ \pi : \pi \in \Pi \land \pi_A = k \} \) and \( k \in K = \{0.05, 0.06, \ldots, 0.84, 0.85\} \). For example, \( \Pi^A_{0.05} \) is the set of all stationary distributions \( \pi \) when the frequency of adenine is 0.05, i.e. \( \pi_A = 0.05 \), whereas frequencies of other nucleotides sum up to 0.95, i.e. \( \sum_{i \in T,G,C} \pi_i = 1 - \pi_A = 0.95 \). The sets \( \Pi^T, \Pi^G, \Pi^C \) were described in the same way.

**Figure 1.** Comparison of two sets of 100 stationary distributions for which \( F_{x}^{\text{max}} \) (the normalized difference between the relative frequency of four-fold degenerated codons after the selection on amino acids and their expected frequency resulting only from a mutation process) takes the highest (red) and the lowest values (green). The \( F_{x}^{\text{max}} \) is the highest for the distributions with the high frequency of thymine and adenine respectively, whereas the lowest for the distributions rich in cytosine and guanine, respectively.
Figure 2. Relationship between the $F_{\pi}^{\text{max}}$ value and combination of two nucleotides presented as coloured Wafer map. The colours correspond to the value of $F_{\pi}^{\text{max}}$, which depends on the frequency of the compared nucleotides. Dark green corresponds the lowest values and dark brown the highest values of $F_{\pi}^{\text{max}}$. Its highest values are for the high content of thymine and adenine, with simultaneous decrease in the guanine and cytosine frequency. The lowest values are for the low frequency of A and T as well as for moderate content of G and C.
Following this approach, we decided to calculate $\text{me}(F_{\pi}^{\max})$, i.e. the median value of $F_{\pi}^{\max}$ for every nucleotide $N = A, T, G, C$ and the stationary distribution $\pi \in \Pi^N_k$ separately. The main reason for using the median can be explained by the fact that it is an estimator of location parameter that is most resistant to outliers. Therefore, it is a useful and stable measure to detect general tendencies in large data sets. In addition, $\text{me}(F_{\pi}^{\max})$ for $\pi \in \Pi^N_k$ is a function of $k \in K$ for every $N = A, T, G, C$. In other words, the median was calculated from $F_{\pi}^{\max}$ values that were derived from substitution models generating nucleotide stationary distributions with the fixed frequency of one nucleotide and random frequencies of others.

**Figure 3.** Dependence of median value of $F_{\pi}^{\max}$, i.e. $\text{me}(F_{\pi}^{\max})$ on stationary frequencies of four nucleotides $\pi$. The median was calculated from $F_{\pi}^{\max}$ values that were derived from substitution models generating nucleotide stationary distributions with the given fixed frequency of one nucleotide $\pi$ and random frequencies of others. The dots represent exact values of $\text{me}(F_{\pi}^{\max})$, whereas lines are the best approximation based on generalized additive models with integrated smoothness estimation. The $\text{me}(F_{\pi}^{\max})$ depends non-linearly on the stationary distribution of particular nucleotides. Its strongest increase is for the growth of A and T.
In Figure 3, we illustrate the dependence of the median value of $F^{\text{max}}_\pi$ on the stationary frequencies of four nucleotides. Interestingly, the relationships are not linear and $me(F^{\text{max}}_\pi)$ shows a similar course for complementary nucleotides, especially for guanine and cytosine. The median value of $F^{\text{max}}_\pi$ starts from a relatively high value for small frequencies of G and C and decreases gradually with their growth, reaching a minimum for their frequencies around 0.36. After that, the median rises steadily, reaching its maximum for the highest frequencies of G and C. However, for the adenine frequency, $me(F^{\text{max}}_\pi)$ grows steadily to 0.6 and then increases rapidly for the highest frequencies. In the case of thymine, the median remains quite constant till it reaches 0.4 and then also quickly increases. The median values of $F^{\text{max}}_\pi$ for the fixed frequencies of G and C are generally lower than for A and T, with the exception of the frequency of G and C below 0.2.

Since the median value of $F^{\text{max}}_\pi$ depends on the complementary nucleotides in a similar way, we examined the dependence of $me(F^{\text{max}}_\pi)$ on the aggregated frequencies of the nucleotides, A+T and G+C, i.e. $\Pi^{A+T}$ and $\Pi^{G+C}$. They are both defined in the analogous way. For example, in the case of $\Pi^{A+T}$ we have:

$$\Pi^{A+T} = \bigcup_{k \in K} \{ \pi : \pi \in \Pi \land \pi_A + \pi_T = k \},$$

where $\pi_G + \pi_C = 1 - (\pi_A + \pi_T)$ and $k \in K = \{0.05, 0.06, ..., 0.84, 0.85\}$. In this case, we observed a sigmoidal increase of $me(F^{\text{max}}_\pi)$ with A+T content (Figure 4A), and the opposite trend for G+C (Figure 4B).

The similar dependence of $F^{\text{max}}_\pi$ on the frequencies of complementary nucleotides justifies considering also simpler models and stationary distributions, e.g. assuming equal frequencies of the complementary nucleotides: $\pi_A = \pi_T$ and $\pi_G = \pi_C$. This assumption was tested by Morton (MORTON 2001) on the example of four mutation-selection models. Here we included a generalization of this model, analysing a wider range of possible nucleotide frequencies $\pi \in \Pi^{A=T,G=C}$. The $F^{\text{max}}_\pi$ shows an exponential growth from 0.541 (for $\pi_A = \pi_T = 0.06, \pi_G = \pi_C = 0.41$) to 5.010 (for $\pi_A = \pi_T = 0.41, \pi_G = \pi_C = 0.09$) with an increase in the A and T frequency (Figure 4C, D).

The results show that there exists a strong relationship between $F^{\text{max}}_\pi$ and the frequencies of complementary nucleotides, regardless of the type of the assumed model. Therefore, we can infer that the impact of the selection, at amino acid level, on the usage of 4FD codons is connected with the structure of the stationary distribution generated by its mutation accumulation process. Generally, the selection is responsible for the high deviation in the synonymous codon usage when the nucleotide substitution process generates a high frequency of A and T nucleotides, while the processes with a high frequency of G+C in their stationary distributions reduces the impact of selection.
Figure 4. Dependence of median value of $F_{\pi}^{\max}$, i.e. $me(F_{\pi}^{\max})$ on stationary content of: adenine + thymine (A), guanine + cytosine (B), adenine and thymine (C) and guanine and cytosine (D) with equal frequencies, as well as purines (E) and pyrimidines (F). There is a clear non-linear relationship with the minimum for equal proportions of purines and pyrimidines.
Surprisingly, we observed a completely different dependence of $me(F_{\pi}^{\text{max}})$ on the total frequency of purines (A+G) and pyrimidines (C+T) in the assumed stationary distributions, i.e. $\pi \in \Pi^{A+G}$ and $\pi \in \Pi^{C+T}$. This dependence turned out to be non-linear and non-monotonic in contrast to the complementary nucleotides. In the case of purines, $me(F_{\pi}^{\text{max}})$ contains the local maximum at about $\pi_A + \pi_G = 0.13$ and then it drops below 1.5 at about $\pi_A + \pi_G = 0.5$, reaching the global minimum (Figure 4E, F). Next, it significantly increases to the global maximum at $\pi_A + \pi_G = 0.9$, with a value of around 3.9. The dependence of $me(F_{\pi}^{\text{max}})$ on pyrimidines shows a symmetrical course (Figure 4E, F), with the global maximum at $\pi_C + \pi_T = 0.11$, the global minimum at $\pi_C + \pi_T = 0.5$ and the local maximum at $\pi_C + \pi_T = 0.9$.

The effect of the selection on particular codon groups

The results presented above referred to the summarized effect of the selection at the amino acid level on all five groups of 4FD codons. However, it is interesting to analyse how the selection influences the deviation in the expected relative usage of particular groups of the codons $s$ for particular nucleotide stationary distributions $\pi$, i.e. $F_{\pi,s}^{\text{max}}$. The extreme values of $F_{\pi,s}^{\text{max}}$ found for the particular codon groups are shown in Table 3. The results demonstrate that the biggest deviation concerns the codons coding glycine, whereas the smallest is for valine codons. The other three groups of codons have comparable values. This effect could be explained by the differences in acceptance probabilities of the substitutions of amino acids coded by these codon blocks. However, the differences between the groups of codons disappear in the case of the lowest $F_{\pi,s}^{\text{max}}$ values, where we observed very similar values of $F_{\pi,s}^{\text{max}}$.

Table 3. The highest and the lowest values of $F_{\pi,s}^{\text{max}}$, which were found for five four-fold degenerated codon block denoted by coded amino acids. Models with the selection assuming lethal substitutions involving stop codons (SL) and the variant with the minimal acceptance probability of such substitutions (SM) were considered, separately.

<table>
<thead>
<tr>
<th>Selection model</th>
<th>Gly</th>
<th>Val</th>
<th>Thr</th>
<th>Ala</th>
<th>Pro</th>
</tr>
</thead>
<tbody>
<tr>
<td>SM</td>
<td>2.09</td>
<td>1.32</td>
<td>1.84</td>
<td>1.97</td>
<td>1.89</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.06</td>
<td>0.06</td>
<td>0.04</td>
<td>0.05</td>
</tr>
<tr>
<td>SL</td>
<td>2.09</td>
<td>1.33</td>
<td>1.89</td>
<td>1.99</td>
<td>1.92</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.08</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
</tr>
</tbody>
</table>

We additionally tested the deviation in the expected relative usage of particular codon groups $F_{\pi,s}^{\text{max}}$ as a function of stationary nucleotide distribution for the SM model (Figure 5). As in our analysis of the summarized effect on these groups, we likewise calculated the median of $F_{\pi,s}^{\text{max}}$ using the values that were obtained from the substitution matrices generating nucleotide stationary distributions with the fixed
frequency of one nucleotide and random frequencies of others. Similarly to the global effect, the same tendency in the case of adenine and thymine was noticed. The median of $F_{\pi_i}^{\text{max}}$ increases with a comparable intensity for all codon groups as a function of $\pi_i$, but in the case of $\pi_A$, $me(F_{\pi_i}^{\text{max}})$ for the glycine codons grow substantially faster than other codon blocks (Figure 5). The trends are different for the codon blocks depending on guanine and cytosine frequencies. In the case of guanine frequency, $me(F_{\pi_i}^{\text{max}})$ decreases substantially for the Gly codon group with the $\pi_G$ growth, in contrast to the other codon blocks, whose $me(F_{\pi_i}^{\text{max}})$ values begin to increase at $\pi_G = 0.35$. Two groups of codons can be distinguished when the relationship between $me(F_{\pi_i}^{\text{max}})$ and cytosine frequency is taken into account. One group, including the codons for Pro, Ala and Thr, shows a decreasing trend in their $me(F_{\pi_i}^{\text{max}})$, whereas the median values of $F_{\pi_i}^{\text{max}}$ of the other group, containing the Gly and Val codons, increase substantially with $\pi_C$.

**Figure 5.** Dependence of the median value of $F_{\pi_i}^{\text{max}}$, i.e. $me(F_{\pi_i}^{\text{max}})$ for four-fold degenerated codon groups (assigned by their coded amino acids) on the stationary frequencies of four nucleotides $\pi$ adenine (A), thymine (B), guanine (C) and cytosine (D). The dots represent exact values of $me(F_{\pi}^{\text{max}})$, whereas lines are the best approximation based on generalized additive models with integrated smoothness estimation. The median value depends differently on the codon groups and nucleotides.
The median of $F_{\pi}^{\text{max}}$ for all codon blocks shows a concordant increasing trend with A+T content (Figure 4A) and decreasing for G+C (Figure 4B) for all codons’ groups. The smallest deviation was observed in the codons for valine. As it was expected, the $F_{\pi}^{\text{max}}$ grows also for all codon groups with the A and T frequencies, with the assumption that $\pi_A = \pi_T$ and $\pi_G = \pi_C$ (data not shown).

![Figure 6](image.png)

**Figure 6.** Dependence of the median value of $F_{\pi}^{\text{max}}$ for four-fold degenerated codon groups (assigned by their coded amino acids) on the stationary frequencies of purines (A) and pyrimidines (B). The dots represent exact values of $me(F_{\pi}^{\text{max}})$, whereas lines are the best approximation based on generalized additive models with integrated smoothness estimation. The groups of codons respond differently to the frequencies.

The analysis of $me(F_{\pi}^{\text{max}})$ for particular codon groups well explains the non-linear relationship between the summarized effect of selection at the amino acid level on all 4FD codons and the purines and pyrimidines content (compare Figure 4E, F and Figure 6). The median of $F_{\pi}^{\text{max}}$ for the Thr, Pro and Ala codons increases with A+G frequency, whereas $me(F_{\pi}^{\text{max}})$ for the Gly codons decreases. This measure for Val codons also declines with the purines content but reaches its minimum at $\pi_A + \pi_G = 0.6$ and then goes up. The relationships between $me(F_{\pi}^{\text{max}})$ and C+T content show a mirror reflection. The superposition of these various trends for particular codon groups leads to the non-linear course of the relationship for the global measure $me(F_{\pi}^{\text{max}})$ for all synonymous codons (Figure 4E, F).
Characteristics of mutational probability matrices maximizing the selection effect

The present study revealed that the effect of selection, at the amino acid level, on the synonymous codon usage strongly depends on the nucleotide stationary distributions, which are the result of mutational processes described by the substitution probability matrices. Therefore, it is interesting to check what types of nucleotide substitutions are responsible for enhancing this effect. In Table 4 is presented the best matrix that together with selection on non-synonymous substitutions maximizes $F_\pi$. The matrix is characterized by the most frequent substitution $C \rightarrow G$ and next $G \rightarrow A$ and $C \rightarrow A$, whereas the rarest substitution in the matrix is $A \rightarrow G$ (Table 4).

Table 4. Transition probability matrix $P_\pi^{\text{max}}$ that maximizes the effect of selection on the relative usage of 4FD codons. It generates, together with the selection, the largest value of $F_\pi^{\text{max}} = 9.10$ under SM variant. The stationary distribution of the matrix is: $\pi_A = 0.19, \pi_T = 0.69, \pi_G = 0.05, \pi_C = 0.07$. A nucleotide in the column is substituted by a nucleotide in the row.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>T</th>
<th>G</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.8882</td>
<td>0.0013</td>
<td>0.0000</td>
<td>0.1104</td>
</tr>
<tr>
<td>T</td>
<td>0.0008</td>
<td>0.9882</td>
<td>0.0007</td>
<td>0.0103</td>
</tr>
<tr>
<td>G</td>
<td>0.1729</td>
<td>0.1549</td>
<td>0.5864</td>
<td>0.0858</td>
</tr>
<tr>
<td>C</td>
<td>0.1719</td>
<td>0.0025</td>
<td>0.2884</td>
<td>0.5372</td>
</tr>
</tbody>
</table>

To check if these properties are universal, we compared 5% (i.e. 4,428) of top matrices that generated the highest values of $F_\pi^{\text{max}}$ (shown in Supplemental Material Figure S2). The maximizing matrices are characterized by a higher probability of staying the same for adenine and thymine than for guanine and cytosine. Their most frequent substitutions are $C \rightarrow A$ and $C \rightarrow G$. It may result from the fact that these transitions belong to the most regular in all the 120 possible non-synonymous and single-nucleotide mutations of 4FD codons. Each of them occurs in 16 cases and they constitute in total 27% of the possible substitutions. On the other hand, the lowest probabilities show substitutions $A \rightarrow G$ and $A \rightarrow T$ (Figure S2). They are the least frequent mutations of all possible non-synonymous mutations involving 4FD codons. Each of them applies in only four cases.

Generally, the maximizing matrices have a tendency to generate more adenine and thymine at the expense of guanine and cytosine. These findings well correspond to the relationships observed between $F_\pi^{\text{max}}$ and nucleotide stationary distributions, indicating a greater deviation in the synonymous codon usage for the nucleotide distributions rich in A and T (Figure 3, 4).

We also noticed that the maximizing matrices are characterized by the preponderance of transversions over transitions, which enhances the impact of selection on the relative synonymous codon usage. The median and quartile range of transitions to transversions ratio is 0.207 [0.133-0.329]. It is less than 0.5 when there is no bias towards either transitions or transversions because there are twice as many possible transversions as transitions. It may be attributed to lower acceptance probabilities for amino acid substitutions in Grantham’s matrix employed in the research, which result rather from the transversions than
the transitions of the corresponding codons. Actually, the mean acceptance probability for transversions and transitions is 0.409 and 0.538, respectively. The difference is statistically significant in the Mann-Whitney test with \( p = 0.00001 \). A higher rate of transversions can, understandably, increase the rare substitutions of codons and lead to a marked bias in the relative usage of 4FD codons.

The maximizing matrices are also characterized by a significant deviation in the pairs of symmetric nucleotide substitutions, e.g. \( A \rightarrow C \) and \( C \rightarrow A \) expressed by:

\[
\text{Dev}_{\text{rev}} = \sum_{X,Y \in A,T,G,C} | p_{X \rightarrow Y} - p_{Y \rightarrow X} |.
\]  

(13)

Median and quartile range for these matrices were 0.623 [0.491-0.747]. The deviation in the pairs of such nucleotide substitutions can enhance the impact of selection on the relative 4FD codon usage through an unbalanced influx and outflow of these codons. For example, substitution \( C \rightarrow A \), which is more frequent than substitution \( A \rightarrow C \), can lower the content of alanine codon GCC at the expense of GAC coding for asparagine. In the case of comparable probabilities of these substitutions, the reversions could recover the number of the disappearing codon.

**Comparison of estimated deviations in codon usage with that observed in protein coding genes**

It would be desirable to assess the strength of the measured effect of selection, at the amino acid level, on the relative usage of 4FD codons in the context of empirical data. Therefore, we compared the difference observed between the relative frequencies of 4FD codons after the selection and their expected frequencies resulting only from the applied mutational process, with an analogous measure for such codons calculated in protein coding sequences present in almost 4,900 bacterial genomes. In an ideal situation, the expected occurrence of the observed relative frequencies of 4FD codons in protein coding sequences should be an aftermath of pure mutational pressures only. They are, however, not known. Therefore, we approximated the expected frequencies of 4FD codons by the average of the relative frequencies of 4FD codons in genes. Since bacterial genomes are characterized by a strong chromosome-wide compositional bias determined by two mutational pressures associated with differently replicated, leading and lagging, DNA strands, we examined the codon usage of genes separately from these DNA strands.

The distribution of the summarized deviation from the expectation in the codon usage for all 4FD groups in protein coding sequences was compared with the distributions of the calculated maximum deviation in the codon usage resulting from the selection at the amino acid level (the measure \( F_{\pi}^{\text{max}} \)). It was also compared with the starting values before the optimization procedure (Figure 7). The maximized values are clearly shifted from the initial distribution and overlap the distribution from genes. The average figure for the starting values is 0.3, for the maximized about 2, whereas for the genes 5.2. The figures for the maximized nucleotide substitution matrices constitute on average 37% of the deviation in codon usage found for protein coding sequences.
Figure 7. Distribution of the deviation from the expectation in the codon usage for all 4FD groups calculated for protein coding sequences, starting (randomly selected) nucleotide substitution matrices and matrices that maximized this measure. The maximized values are of the same order of magnitude as the deviation based on the empirical data.

The data seem to indicate that the estimated effect of the selection by the measure $F_{\pi}^{\text{max}}$ is not negligible. It is likely that the deviation calculated for the real genes is higher because of the additional factors influencing codon bias and linked, for example, with the effectiveness of translation, which appears universal in prokaryotic genomes and concerns substantial fraction of their genes (SUPEK et al. 2010). It is conceivable that the applied selection model based on the general Grantham’s amino acid matrix (GRANTHAM 1974) deviates the synonymous codon usage less than in real selections, which can be different for various genes and their products. Nevertheless, the comparison with the observed data demonstrates that the effect of selection at the amino acid level could help to explain a substantial proportion of the observed codon usage bias and as such cannot be disregarded.
Modelling of codon substitution process

In the paper, we used a mutational-selection model, the same as proposed by MORTON (2001). This approach has many advantages, which are relevant to our study. The model consists of separate mutation and selection components, which are easy to control. The mutation process can be simply defined by mutational matrices based on fixed nucleotide stationary distributions without any influence of selection. Apart from stationarity, the mutational matrix does not require the assumption on time-reversibility, which makes this model more general. The selection part is also easily expressed by an amino acid matrix, which does not require complicated transformations and implementations.

We decided to apply this model because others commonly used in the modelling of codon substitutions (HALPERN and BRUNO 1998; YANG and NIELSEN 2008) are not flexible enough to investigate the studied phenomenon. Particularly, they describe the mutation substitutions as a time-reversible stationary Markov process. They also introduce a selection mechanism in such way that the final mutation-selection process is also time-reversible and stationary. Thanks to that, the models are computationally effective tools. However, there are no biological reasons to expect that the substitution process proceeds in the reversible way (FELSENSTEIN 2004; SCHNEIDER and CANNAROZZI 2012; YANG 2006). This assumption is used only because of theoretical and practical computational benefits as well as mathematical convenience.

It is worth pointing out that when we implement in these models a selection based on amino acid frequencies determined by functional requirements in proteins then these models will produce the relative stationary frequencies of synonymous codons that will be the same as the stationary frequencies resulted from the strict mutational process (WALLACE et al. 2013; YANG and NIELSEN 2008). In consequence, under these assumptions, it is impossible to investigate the impact of selection at the amino acids level on the synonymous codon usage. Therefore, the model applied in this work appears to be more elastic and general because of less restrictive assumptions. In addition, it does not exclude “a priori” any possible additional factors which could influence the usage of synonymous codons.

The time-reversibility assumption is crucial in our consideration because if we assume that the nucleotide substitution process defined by the matrix $P^*$ is time-reversible and the acceptance matrix $D$ is symmetric then the impact of selection at the amino acid level on the usage of synonymous codons disappears. To prove this, it is enough to show that, under the above assumptions, the combined mutation-selection process defined by the probability matrix $C$ has the same stationary distribution as the mutational process alone, i.e. $\pi_{sel} = \pi_{cod}$. Thus, from the time-reversibility, we get:

$$p_{k\rightarrow l}^* \times \pi_{k}^{cod} = p_{l\rightarrow k}^* \times \pi_{l}^{cod}$$

and assuming the symmetry of the acceptance matrix $D$, i.e. $d_{m\rightarrow n} = d_{n\rightarrow m}$, we obtain the following equalities:

$$c_{k\rightarrow l} \times \pi_{k}^{cod} = p_{k\rightarrow l}^* \times d_{m\rightarrow n} \times \pi_{k}^{cod} = p_{l\rightarrow k}^* \times d_{n\rightarrow m} \times \pi_{l}^{cod} = c_{l\rightarrow k} \times \pi_{k}^{cod},$$

Clearly, the formula (15) is the detailed-balance equation of the process generated by the mutation-selection matrix $C$. As a result, this process is also time-reversible and consequently $\pi_{sel} = \pi_{cod}$. Thus, the studied influence of selection at the amino acid level on the synonymous codons usage cannot be demonstrated.
under the assumption of time-reversibility of the mutational process and the symmetry of the acceptance matrix $D$.

Nevertheless, this property (15) can be used in validation of the searching algorithm applied in our study. Since this algorithm considers the general class of nucleotide mutational matrices including also the time-reversible models as a subset, it should be possible to find by this algorithm such nucleotide transition probability matrices that would generate the exact equality between stationary codon distributions before and after selection, i.e. $\pi^{\text{rel}} = \pi^{\text{cod}}$. Such results would imply that the algorithm works efficiently. As expected, we received this equality for nucleotide substitution matrices that minimized the objective function $F_\pi$. The average values of $F_\pi$ were in these simulations almost equal to zero for all considered nucleotide stationary distributions.

It should be also added that if the deviation in synonymous codon usage is obtained under the time-reversible nucleotide substitution matrices then the acceptance probabilities matrix must be asymmetric. However, commonly used matrices describing physicochemical or biochemical differences/similarities between amino acids are symmetric. In our approach, we applied Grantham’s matrix of acceptance probabilities corresponding to chemical similarities between amino acids (GRANTHAM 1974). Since the matrix is symmetric, it favours no direction of amino acid replacement in contrast to a mutational matrix. Therefore, both mutation and selection are necessary to generate the bias in the usage of 4FD codons.

It is also possible to apply in our model other acceptance probability matrices based on various physicochemical or biochemical amino acid properties, e.g. hydropathy or polarity. Since such properties and resulting matrices are usually quite strong correlated, their use would not change the general conclusion about the influence of selection at the amino acid level on synonymous codon usage. The other matrices can slightly increase or decrease the codon bias observed for Grantham’s matrix in dependence of stationary distributions of mutational matrices but comprehensive and detailed studied are necessary to assess these relationships and intensity of this effect. However, commonly used PAM (Point Accepted Mutation) matrices are not appropriate because they are not free of a mutational influence, which is important in our considerations. It should be also noticed that Grantham’s matrix represents a mean field approximation of a model with fluctuating selection because this matrix describes only general similarities in chemical properties of amino acids but not specific selection for a particular protein. Various types of proteins can be characterized by different selection requirements because of their specific structure and function. The Grantham’s matrix is a general representation of constant selection but may not be a good approximation to the true evolutionary dynamics under time-varying selection. Such variable selections can produce their own distinctive pattern in codon usage bias in different types of sequences or regions (PLOTKIN and DUSHOFF 2003; PLOTKIN et al. 2006).

Properties of $F_\pi$ measure

To assess the strength of selection at the amino acid level on the 4FD codon usage, we applied $F_\pi$ measure being the normalized difference between the relative frequency of 4FD codons before and after
selection (9 and 10 equations). This measure was inspired by chi-squared statistics and has useful features like other standard measures, which can be also applied in the calculation of differences between the probability distributions. Most of all, $F_{\pi}$ is always non-negative and equals zero if and only if the relative stationary frequency of a codon subjected to a mutational process equals its frequency after selection, i.e. $\frac{\pi_i}{\pi_{i,sel}}$ for all $i \in \{A,T,G,C\}$. This property is called the identity condition and should be fulfilled by any measure, which is used to calculate the difference between two probability distributions. Thanks to that, the detection of any differences between stationary frequencies of $\pi$ and $\pi_{sel}$ is independent on the measure. In contrast to the standard measures like Kullback-Leibler divergence or total variation distance, $F_{\pi}$ includes information about the absolute value of the change of codon frequencies in relation to the expectation under the mutational process, which is useful to easy interpret the obtained results.

Nevertheless, our measure gives values compatible with Kullback-Leibler divergence and total variation distance. Spearman correlation coefficient between values of our measure with those of Kullback-Leibler divergence and total variation distance calculated for all 88,560 considered cases of nucleotide stationary distributions is very high and statistically significant (p-value < 2.2E-16), 0.921 and 0.915, respectively. Therefore, the application of these measures would change neither conclusions nor important results. With such great correlation, the trends presented in figures between the measure and nucleotide composition would be also the same.

**Concluding remarks**

The undertook study estimates the influence of selection, at the amino acid level, on the relative usage of four-fold degenerated codons. This impact is determined by different selection constraints on the non-synonymous replacement of these codons with others, which proceeds in a complex manner and depends on the probability of fixation of such substitutions as well as on the probability of particular nucleotide substitutions. We tested a wide range of conditions in which such influence can be valid, by the inclusion of nearly 90,000 stationary nucleotide distributions and associated unrestricted mutational processes. The selection was based on the differences in physicochemical properties of amino acids.

We noticed that mutational processes generating more adenine and thymine than guanine and cytosine enhance the influence of selection. The same is true for the processes yielding more purines than pyrimidines. It is noteworthy that the relationship between the effect under study and the content of these nucleotides is non-linear. On the other hand, the impact of selection at the amino acid level diminishes when the nucleotide processes generate 50% content of purines and pyrimidines as well as more guanine and cytosine than adenine and thymine. The nucleotide substitution matrices maximizing the consequence of the amino acid selection are characterized also by a greater probability of transversions outnumbering transitions, and a greater deviation in pairs of reversible nucleotide substitutions.

The influence of selection at the amino acid level was different for particular groups of 4FD codons. Generally, glycine codons show the strongest response to the selection impact under study, whereas codons
for valine the weakest. However, the deviation in the codon usage generated by the process with and without selection depends non-linearly from nucleotide stationary distribution. This effect could be explained by the discrepancies in the acceptance probabilities of substitutions of amino acids coded by these codon blocks.

The results indicate that the selection acting on non-synonymous substitutions, i.e. leading to amino acid replacements, can affect the usage of 4FD codons. This effect, however, is complex and depends on the properties of mutational pressure, which superimposes on the selection. Interestingly, we discovered that for each nucleotide distribution, it is possible to find such mutational probability matrices that will minimize and maximize the effect. It seems to suggest that the influence of selection, at the amino acid level, on the synonymous codon usage can vary in different organisms. Since the mutational pressure in genomes is not known and the selection at the amino acid level is also complicated, it is difficult to assess the exact contribution of this process in real protein coding sequences. The selection can both enhance and suppress the other effects on the codon usage, e.g. a selection related to translation efficiency. Nevertheless, the effect cannot be neglected because it correlates with the comparison between the calculated deviation in the codon usage subjected to this selection and an analogous measure estimated for protein coding sequences.

Our results show that the substitution matrices generating high A+T content affect the 4FD codon usage to the greatest extent. Assuming that the global genome content corresponds to a global mutational pressure (MUTO and OSAWA 1987), we can conclude that the effect of the selection under study would be most pronounced in AT-rich genomes. Consequently, the selection on non-synonymous substitutions can interfere in such genomes with other selections on codon usage, e.g. related with translational efficiency. In agreement with that we found that the difference between highly expressed genes coding for ribosomal proteins and the other genes, as far as the relative usage of particular 4FD codons is concerned, becomes smaller with A+T genomic content (Figure 8). There seems to be some evidence that it could be more difficult to maintain the appropriate codon bias in highly expressed genes in AT-rich genomes. Likewise, in genomes with more than 70% A+T, no influence of translational selection was reported, i.e. *Borrelia burgdorferi* (MCNERNEY 1998), *Buchnera* (RISPE et al. 2004), *Wigglesworthia* (HERBECK et al. 2003) and *Blochmannia floridanus* (BANERJEE et al. 2004). It may result from a greater difficulty in predicting genes with translational efficiency in AT-biased genomes using standard methods (e.g. CAI, codon adaptation index) because other methods based on random forest classifier revealed in these genomes the codon bias associated with the translational efficiency (SUPEK et al. 2010). Notwithstanding, our results show that AT-rich genomes have to cope with the greater influence of the selection, at the amino acid level, on the synonymous codon usage or adapt to it. It is possible that this type of selection can trigger codon bias in some genes, which can be misleading with regard to the selection on the translational effectiveness. Accordingly, Morton carried out an appropriate test decreasing the number of genes believed to have the codon usage associated with translational selection (MORTON 2001).

The aim of our study was to verify the effect of amino acid selection on the 4FD codon usage in global and general scales for the large number of possible mutational pressures and a fixed selection for amino acid replacements. Nevertheless, our results can be helpful to explain some effects related with codon bias also in local scales, i.e. codon usage variation across sites within a gene. Such variation was noticed by
AKASHI (1994) in orthologous genes from fruit flies. He found that the frequency of preferred codons is significantly higher at conserved amino acid positions than in non-conserved ones. This finding was further confirmed in bacteria (STOLETZKI and EYRE-WALKER 2007) as well as yeast, worm, mouse and human (DRUMMOND and WILKE 2008). This codon bias was interpreted as a result of selection for minimization of the chance for translation errors and protein misfolding during this process. On the other hand, PLOTKIN and DUSHOFF (2003) and PLOTKIN et al. (2004) found that some variable sequences coding for antigens and surface proteins or regions interacting with antibodies in pathogens *Mycobacterium tuberculosis*, *Plasmodium falciparum*, and influenza A virus are rich in “volatile” codons that can mutate with larger probability to codons encoding other amino acids. Such elevated volatility of these genes may be associated with a positive selection and greater pressure for amino-acid substitutions, which is favoured in order to avoid interactions with the host immune system. Although we showed that the synonymous codon bias can be generated by a general selection at the amino acid level, it cannot be excluded that more specific selections influencing particular sites in protein sequences with various intensity or pattern (BAZYKIN 2015) can also contribute with the other effects to the observed codon biases in the specific sequence sites.

**Figure 8.** Dependence of the difference in the relative usage of 4FD codons between genes coding for ribosomal and non-ribosomal proteins on the genomic A+T content. The difference was calculated based on 4802 genomes with annotated at least 30 genes for ribosomal proteins, separately for the leading and lagging strand. In total 5124 pairs of genes with at least 15 ribosomal genes on one strand were considered. The bars represent an average value for the given class of A+T content whereas whiskers standard deviation. The difference was calculated according to:

\[
\frac{\sum_{s \in S} \sum_{i \in A,T,G,C} o_{s,i}^{\text{rib}} \left( o_{s,i}^{\text{rib}} - o_{s}^{\text{nonrib}} \right)}{\sum_{s \in S} o_{s}^{\text{nonrib}}},
\]

where \( o_{s} \) is the observed frequency of a 4FD codon \( s \), with a nucleotide \( i \) at the third codon position, \( o_{s} = \sum_{i \in A,T,G,C} o_{s,i} \) is the frequency of all codons in the 4FD codon group \( S \). Indices *rib* and *nonrib* mean genes for ribosomal and non-ribosomal proteins, respectively. The calculated difference decreases with AT% and is the largest for the moderate AT content.
Acknowledgment

We are very grateful to the Reviewers for their valuable comments and insightful remarks that significantly improved the paper. The publication editing was funded by the KNOW Consortium, Faculty of Biotechnology, University of Wroclaw.

Supplemental Figures

Figure S1. The procedure leading to the assessment of the selection strength, at the amino acid level, on synonymous codons usage (the orange box). To achieve that, a mutation-selection model (the green boxes) was developed. It results from connecting the selection on amino acid substitutions (the blue box) with a codon substitution process derived from a nucleotide mutation process (the yellow boxes). To find the minimum and maximum of this selection strength, nucleotide rate matrices were optimized in Evolutionary Strategies approach (the pink box).

Figure S2. Comparison of nucleotide substitution probabilities for 5% of top matrices that maximized the values of $F_\pi$ (the normalized difference between the relative frequency of four-fold degenerated codons after the selection on amino acids and their expected frequency resulting only from a mutation process). The thick line indicates median, the grey box shows quartile range and the whiskers determine the range without outliers.

References


LaZrak, A., L. Fu, V. Bali, R. Bartoszewski, A. Rab et al., 2013 The silent codon change IS07-ATC->ATT contributes to the severity of the DeltaF508 CFTR channel dysfunction. FASEB J 27: 4630-4645.


XIA, X., 1998 How optimized is the translational machinery in Escherichia coli, Salmonella typhimurium and Saccharomyces cerevisiae? Genetics 149: 37-44.


