# Using Evolutionary Algorithms in Finding of Optimized Nucleotide Substitution Matrices

Paweł Błażej, Paweł Mackiewicz, Stanisław Cebrat, Małgorzata Wańczyk
Department of Genomics, Faculty of Biotechnology, University of Wrocław
ul. Przybyszewskiego 63/77, 51-148 Wrocław, Poland
blazej@smorfland.uni.wroc.pl, pamac@smorfland.uni.wroc.pl,
cebrat@smorfland.uni.wroc.pl, Malgorzata.Wanczyk@gmail.com

## Categories and Subject Descriptors

J.3.1 [**Computer Applications**]: Life and Medical Sciences—*Biology and Genetics*; G.3 [**Mathematics of Computing**]: Probability and Statistics—*Markov processes, probabilistic algorithms (including Monte Carlo)*; I.2.8 [**Computing Methodologies**]: Problem solving, Control Methods and Search—*Heuristics methods*

## Keywords

DNA, evolutionary algorithm, genome, General Time Reversible model, Markov processes, mutational pressure, nucleotide substitution matrix, sequence evolution

## 1. INTRODUCTION

Mutations occurring in biological DNA sequences are not completely random but are a result of coevolution between mutational pressure with selection constraints around the genetic code [2, 5] and can be optimized to some extent during evolution ([6]). On one hand, most mutations are deleterious and generate energetic costs of their repairing, therefore a tendency to decrease the mutation rate should exist. On the other hand, mutations are responsible for genetic diversity, which is necessary for adaptation in changing environment. Therefore, an elevated level of mutation rate should be also expected in these cases. It indicates that the real mutational pressure should be subjected to some optimization in biological systems.

## 2. METHODS

To find the optimized mutational pressures expressed by nucleotide substitution matrices we combined classic methods from probability theory with basic principles of evolutionary strategies [1]. The matrices were described by General Time Reversible (GTR) model that requires six different rate parameters and the stationary distribution of four nucleotides [4, 7]. The matrices were optimized for four objective functions: the minimum and maximum number of all mutations, and the minimum and maximum number of non-synonymous substitutions, i.e. mutations not changing encoded amino acid in protein coding sequences. The best solutions were searched among the population of 100 individuals (rate matrices), which were initially randomly generated. In every simulation run, an individual was selected with probability 0.5 and mutated by random modifications of its rate parameters. The found solutions obtained very good convergence and small variation. The matrices were compared with the empirical replication-associated substitution rate matrix and assumed the fixed stationary distribution of nucleotides as in the real mutational pressure found for bacteria *Borrelia burgdorferi* genome [3]. The matrices were used to generate the process of nucleotide substitutions in protein coding sequences from this genome. Besides the original protein coding sequences (the gene set), we also searched the best matrices for two other cases: (i) randomly generated sequences with the same length and global nucleotide composition as the original protein coding sequences (the random set) and (ii) sequences with the same length as the original protein coding sequences but assuming uniform composition of four nucleotides (the uniform set).


*GECCO'13 Companion,* July 6–10, 2013, Amsterdam, The Netherlands.
ACM 978-1-4503-1964-5/13/07.

## 3. RESULTS AND DISCUSSION

In Fig. 1 we compared, as an example, matrices according to the number of all mutations that they introduced into gene sequences. As expected, the distribution of these numbers for minimizing and maximizing matrices are very narrow and are located at two extremes in the plot whereas the distribution for initial randomly generated matrices is very wide and located between the former ones. The number of substitutions introduced by minimizing matrices are about 5.5 times lower than by maximizing one. These results indicate that the applied algorithm efficiency optimized the objective function. Interestingly, the expected number of mutations introduced by the empirical matrix describing mutational pressure in the bacteria *B. burgdorferi* genome is located between the extreme values for optimized matrices with a slight shift toward the value characteristic of minimizing matrices.

To easy compare the elements of all obtained matrices we carried out PCA analysis to reduce the number of dimensions from 16 matrix elements to two main variables (Fig. 2). Based on the PCA visualization we can recognize two main groups of matrices. The one cluster includes relatively very similar all matrices that minimized the number of all or non-synonymous mutations for all types of sequences. These matrices clearly separate from the second group also relatively very similar matrices that maximized these numbers of mutations. This group does not comprise the matrix that maximized the number of non-synonymous substitutions for sequences with the uniform distribution of four nucleotides. However, this matrix is separated only by the second principal component that explains only about 7% of variance.

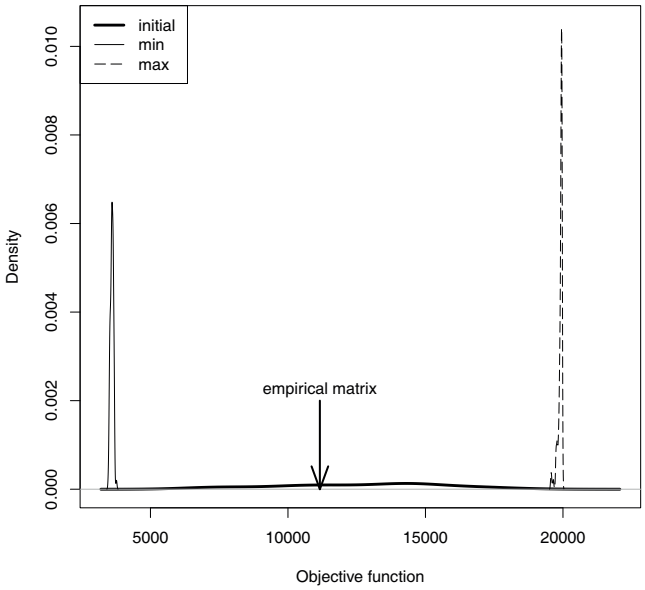


**Figure 1: Empirical density function of the number of all mutations introduced into real protein coding sequences sequences by hundred best minimizing (min) and maximizing (max) matrices as well as hundred initial randomly generated matrices (initial). The arrows indicates the expected number of mutations introduced by the empirical mutational matrix found for** *B. burgdorferi***.**

Interestingly, the empirical matrix is situated between the clusters for matrices minimizing and maximizing mutations with slightly smaller distance to the former group.

Analyses showed significant differences between rate matrices minimizing and maximizing mutations but very close similarity between matrices of the same optimization type and tested on different kinds of sequences. As it should be expected in biological systems, the empirical matrix located between these two extremes with some affinity to minimizing matrices. The applied approach turned out to work effectively and could be used in searching for the best rate matrices under other objective functions.

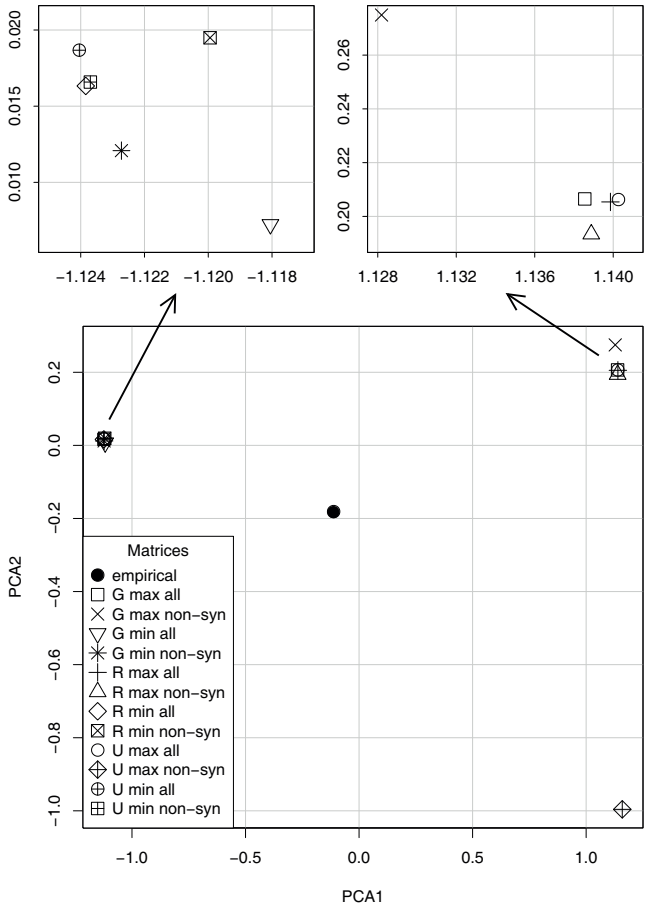


**Figure 2: PCA analysis of the best uniformized rate matrices optimized according to the minimum (min) or maximum (max) number of all mutations (all) or non-synonymous (non-syn) substitutions for real protein gene sequences (G), the same sequences but with randomized nucleotide positions (R) and sequences with the uniform distribution of four nucleotides (U). The empirical mutational matrix (empirical) found for** *B. burgdorferi* **genome was included for comparison. Two main principal components explain more than 97% of variance.**

# 4. REFERENCES


[1] K. De Jong, D. David, B. Fogel, H.-P. Schwefel, et al. A history of evolutionary computation. In T. Back, D. Fogel, Z. Michalewicz, et al., editors, *Handbook of Evolutionary Computation*, pages A2.3:1–12. Oxford University Press, 1997.

[2] M. Dudkiewicz, P. Mackiewicz, A. Nowicka, M. Kowalczuk, D. Mackiewicz, N. Polak, K. Smolarczyk, J. Banaszak, M. Dudek, and S. Cebrat. Correspondence between mutation and selection pressure and the genetic code degeneracy in the gene evolution. *Future Generation Computer Systems*, 21(7):1033–1039, 2005.

[3] M. Kowalczuk, P. Mackiewicz, D. Mackiewicz, A. Nowicka, M. Dudkiewicz, M. Dudek, and S. Cebrat. High correlation between the turnover of nucleotides under mutational pressure and the DNA composition. *BMC Evol. Biol.*, 1:13, 2001b.

[4] C. Lanave, G. Preparata, C. Saccone, and G. Serio. A new method for calculating evolutionary substitution rates. *J. Mol. Evol*, 20:86–93, 1984.

[5] P. Mackiewicz, P. Biecek, D. Mackiewicz, J. Kiraga, K. Baczkowski, M. Sobczynski, and S. Cebrat. Optimisation of asymmetric mutational pressure and selection pressure around the universal genetic code. In Bubak, M and Dongarra, J and VanAlbada, GD and Sloot, PMA, editor, *Computational Science - ICCS 2008, PT 3*, volume 5103 of *Lecture Notes in Computer Science*, pages 100–109. Elsevier; Springer, 2008.

[6] P. Sniegowski, P. Gerrish, T. Johnson, and A. Shaver. The evolution of mutation rates: separating causes from consequences. *Bioessays*, 22(12):1057–1066, 2000.

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