

Correlation between the variability of mammalian molecular genetic sequences and the geoclimatic characteristics of their habitats

Efimov V.M.^{1,2,3,5*}, Polunin D.A.², Kovaleva V.Y.³, Efimov K.V.⁴

¹ *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*

² *Novosibirsk State University, Novosibirsk, Russia*

³ *Institute of Systematics and Ecology of Animals, SB RAS, Novosibirsk, Russia*

⁴ *Higher School of Economics, Moscow, Russia*

⁵ *Tomsk State University, Tomsk, Russia*

* efimov@bionet.nsc.ru

Key words: *Cytb*, amino acid sequences, Glires, climatic factors, correlation

Motivation and Aim: The mitochondrial *Cytb* gene is currently the basis of the molecular taxonomy of organisms, as it is considered to be selectively neutral. Recently, however, this assertion has been increasingly questioned. Therefore, we performed a multivariate analysis of the covariance of amino acid sequences of the *Cytb* gene with environmental factors for a set of different rodent mammalian species. This analysis made it possible to reveal the adaptive components of the variability of the amino acid sequences of the *Cytb* gene, which significantly correlate with the geoclimatic characteristics of species habitats.

Materials and Methods: Amino acid sequences of the mitochondrial gene *Cytb* (378 aa) of 72 species of Old World rodent mammals Glires, which list the latitude and longitude of sampling sites, were taken from GenBank (NCBI). Based on these geographical coordinates, for each sequence, using the AQUASTAT service, 13 monthly climatic characteristics averaged over 30 years (156 variables in total), as well as the height of the terrain, were obtained.

Based on these data, two matrices of interspecies Euclidean distances of size 72x72 were calculated: by primary amino acid sequences and by a set of standardized climatic variables. For amino acid sequences, the square root of p-distance was used. Principal component matrices (PC) were calculated for each distance matrix by the method of principal coordinates [1]. Between them, the matrix of Pearson's correlation coefficients is calculated (Table 1). The Figure shows scatterplots corresponding to significant correlation coefficients. The calculations were carried out using the Jacobi4 package [2].

Results: Based on the Table, the first amino acid PC (22.6 % of total variance) does not correlate with any climatic PC. However, the next two amino acid PCs significantly ($p < 0.001$ according to Bonferroni criterion) correlate with the first two PCs of climatic variability. Judging by the correlations of AaPC1, it reflects the mutual substitutions of amino acids with similar physicochemical properties, which is consistent with the evolution neutrality hypothesis. In turn, ClimPC1 correlates with the temperature gradient associated with the latitude of habitats, while ClimPC2 reflects precipitation and humidification.

Table 1. Pearson correlation coefficient matrix (×1000) between amino acid and climatic principal components

PC	AaPC1	AaPC2	AaPC3	AaPC4	AaPC5	AaPC6	AaPC7	λ, %	S, %
ClimPC1	197	472***	66	43	-60	-16	59	41.6	41.6
ClimPC2	-72	133	479***	-163	-164	-86	-244	26.9	68.5
ClimPC3	-94	169	-47	-218	253	28	239	11.3	79.8
ClimPC4	126	43	104	-56	185	156	-274	5.2	85.0
ClimPC5	-184	-82	-33	35	-46	54	68	3.4	88.5
λ, %	22.6	10.3	8.7	6.5	5.3	4.1	2.7		
S, %	22.6	32.9	41.6	48.1	53.4	57.5	60.2		

Note: *** means $p < 0.001$ according to Bonferroni criterion

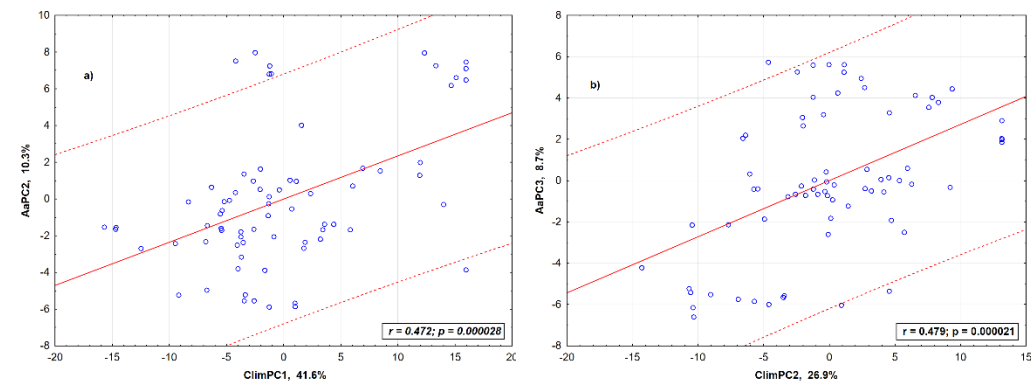


Fig. 1. Configuration of 72 species of Old World rodent mammals Glires on the plane of climatic and amino acid principal components: a) ClimPC1 – AaPC2; b) ClimPC2 – AaPC3

Conclusion: The results obtained support both hypotheses. The first amino acid PC, in all likelihood, corresponds to the hypothesis of neutral evolution, the next two PCs are adaptations to environmental factors. The next task is to identify the positions of substitutions and their biological meaning in both cases.

Acknowledgements: The study is supported by the Kurchatov Genomic Center of the Institute of Cytology and Genetics, SB RAS (075-15-2019-1662).

References

1. Gower J.C. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*. 1966;53(3-4):325-338.
2. Polunin D. et al. JACOBI4 software for multivariate analysis of biological data. *bioRxiv*. 2019:803684.5.