Секция «Биоинформатика»

## Phylogenetic inference using protein evolutionary domains

### Научный руководитель – Спирин Сергей Александрович

***Ноздрин Владимир Александрович***
*Student (specialist)*
Московский государственный университет имени М.В.Ломоносова, Факультет биоинженерии и биоинформатики, Москва, Россия
*E-mail: nozdrin01vl@gmail.com*

**Introduction.** Most studies about inferring of mammalian phylogeny used full-length coding sequences. For example, in paper [1] sequences of APOB gene exons and in paper [4] sequences of AT-rich genes were used. However sequences of full-length proteins are subjects of domain shuffling that often makes it difficult to use them in phylogenetics. Maybe there is a point in using orthologous parts of protein sequences. The aim of this study is to test a suitability of evolutionary domains for solving mammalian phylogeny.

**Methods.** There are 64 species of mammals whose evolutionary domains sequences are present in Pfam release 33.1 (2020) [2]. We selected 39 of them, one from each mammalian family. Those Pfam families that are common for those 39 species were selected. We fetched sequences of evolutionary domains from those families and obtained orthologous groups of them. Sequences of each orthologous group were aligned, and the obtained alignments were used to infer three phylogenetic trees. The first tree is based on concatenation of all alignments. The second tree is the extended majority rule consensus of trees inferred from the alignments. The third tree is ASTRAL-III [5] consensus. All trees were inferred with FastME 2.0 [3].

**Results.** The gained trees have much in common with generally accepted phylogeny. Although, all of gained trees have one common error, namely Afrotheria (Loxodonta africana and Trichechus manatus) is inside Boreoeutheria though it should be outgroup for Boreoeutheria. The second significant error is present only in concatenation tree and Extended Majority rule tree, namely Erinaceus europaeus (order Eulipotyphla) is outside Boreoeutheria but it is supposed to be inside Laurasiatheria. The latter error may be caused by long branch attraction.

**Conclusions.** If the causes of these errors will be found and fixed, sequences of evolutionary protein domains have a potential to be used in phylogenetic inference. Maybe such trees can be more accurate if there are more data.

**Acknowledgments.** Author would like to thank the study's scientific supervisor S.A.Spirin.

### References

1) Amrine-Madse et al. A new phylogenetic marker, apolipoprotein B, provides compelling evidence for eutherian relationships. Molecular Phylogenetics and Evolution (2003) doi: 10.1016/S1055-7903(03)00118-0
2) J. Mistry et al. Pfam: The protein families database in 2021. Nucleic Acids Research (2020) doi: 10.1093/nar/gkaa913
3) Lefort et al. FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program. Molecular Biology and Evolution (2015) doi: 10.1093/molbev/msv150

4) Romiguier et al. Less Is More in Mammalian Phylogenomics: AT-Rich Genes Minimize Tree Conflicts and Unravel the Root of Placental Mammals. Molecular Biology and Evolution (2013) doi: 10.1093/molbev/mst116

5) Zhang et al. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics (2018) doi: 10.1186/s12859-018-2129-y