Rare amino acid changes fixation drives divergence in Metazoa evolution

K.V. Gunbin, V.V. Suslov, Y.L. Orlov

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia, genkvg@gmail.com

Novosibirsk State University, Novosibirsk, Russia

One of the features of the molecular evolution of any amino acid sequence is an unevenness of the rates of mutation accumulation. Fixation bursts were observed during analysis of the divergence of mouse and rat [1]. Mutation bursts usually are accompanied by the fixation of statistically rare amino acid substitutions [2, 3]. These changes can be observed, for example, as a "Stokes shifts" in protein evolution [3]. It is likely that such bursts of mutations can be the result of "reusing" of existing protein domains in the context of various functions [4]. Therefore, it is of interest to detect such changes in protein families over the course of the Metazoan evolution maximally using all the available data.

A comprehensive study of fixation of atypical amino acid substitutions during Metazoan protein evolution was made based on over one million proteins maintaining orthologous status among one hundred completely sequenced organisms. In order to identify the atypical amino acid substitutions on the each branch of the Metazoa phylogenetic tree, the previously developed approach [5-7] was used. The search for atypical amino acid substitutions was confined to all the internal branches of the phylogenetic trees. It is of importance that the empirical model of amino acid replacement derived for each analyzed protein family. This allowed us to study protein evolution more accurately (taking into account functional and structural protein features) comparing to protein studies based on standard general models. Using the orthology relationships between proteins described in OrthoDB v. 6 we for the first time selected phylogenetically connected orthologous protein groups (or PCOPGs) and grouped these PCOPGs in clusters on the basis of protein sharing feature. This procedure allowed us to study orthologous protein groups with evolutionarily important functions rarely lost in the evolution of large taxa.

To test whether there is an increased selection pressure on the sites with atypical amino acid substitutions we analyzed the fraction of descendant branches with (secondary) amino acid

substitutions occurred after the (primary) amino acid substitution on ancestral branch. The analysis clearly demonstrated that after the primary atypical amino acid substitution on ancestral branch the frequency of occurrence of secondary substitutions on descendant branches three to five times lower comparing to the number of secondary substitutions after any types of substitutions ($p<1*10^{-16}$). Therefore, a low number of secondary substitutions occurred on descendant branches after the atypical amino acid replacements may indicates disruptive selection, under which the original ancestral state switches to other alternative stable state. To check this in depth, we performed a statistical regression between the rates of taxon formation in the Vertebrata fossil records and the fraction of PCOPGs clusters that fix atypical amino acid substitutions on the internal nodes of the phylogenetic tree of multicellular animals with good paleontological descriptions in PaleoDB [8]. A clear-cut positive statistically significant association ($p<0.01$) were observed between increases in the frequencies of fixation of rare amino acid substitutions and the genus birth rate in the paleontological history of taxa when maximum number of rare amino acid substitutions per PCOPG cluster under consideration. This implies that the rate of divergent morphological evolution in vertebrates is associated with the rate of fixation of atypical amino acid substitutions in proteins.

To identify the functional role of atypical substitutions, we analyzed the frequencies of their occurrence in different types of functional domains of proteins based on annotations for domains in Pfam / SUPERFAMILY / SMART / PANTHER and annotations only for domains in Pfam / SUPERFAMILY databases. For each divergence (internal node) of 25 major taxonomic groups of multicellular animals, a search was made for statistically overrepresented InterPro and Gene Ontology (GO) terms that characterized PCOPG clusters with statistically rare amino acid substitutions. To answer the question as to what kind of protein domains are characterized by the most intensive fixation of atypical substitutions in animals, we divided all the protein domains into two large groups, the evolutionarily old domains, which are possessed by both animals and their distantly related relatives, fungi and amoeba, and the evolutionarily young domains, which are only possessed by animals. After that, we performed functional enrichment tests using the randomization procedures. Protein domains associated with the eggNOG functional categories of the basal biochemical cell machinery, the replication, recombination and repair, the energy production and conversion, the extracellular structures

maintenance, the translation processes, and the nuclear and chromatin structure maintenance are totally enriched (p=1*10^{-6}) with fixation of atypical amino acid substitutions. This fact, clearly demonstrates the functional importance of atypical amino acid substitutions over the evolution, because this relation is very clear only on evolutionary old protein domains. Another interesting observation obtaining using Gene Ontology is the high enrichment (p<1*10^{-6}) with atypical amino acid substitutions of evolutionarily old proteins which are characterized by cell-oriented functions related to development and differentiation. Interestingly, this shift is so strong that the GO category 'cell differentiation' most frequently characterizes protein domains that fix atypical amino acid substitutions (p=1*10^{-10}). By contrast, the enrichment with atypical amino acid substitutions of GO categories characterizing evolutionary young protein domains is not outstanding, if exists.

Taking into account that atypical amino acid fixations in the protein related with the cell differentiation in the context of a multicellular organism, a detailed analysis of the protein molecular evolution via searching for atypical (statistically rare) amino acid substitutions was performed on proteins known to be related to cellular differentiation in mammals. It was shown that several embryonic stem cell specific proteins containing fixed rare amino acid substitutions (FOXD3, ESSRB, GABRB3, HCK, NR5A2, SALL4, TFCP2L1) had fixed rare amino acid substitutions at the divergence of Placentalia. Other phylogenetic hotspots of rare amino acid fixations are related to the respective divergences of Chiroptera, Eulipotyphla and Boreoeutheria. It was clearly demonstrated that the burst of atypical amino acid substitutions in proteins with pluripotency functions occurred at the time when Placentalia was forming.

One of the most interesting results for broad biologists' community is the relation of acceleration/deceleration in the fixation of atypical amino acid substitutions with the type of observed macroevolutionary changes. In particular all general (or aromorphic) morphological transformations in the paleontological record coincide with reduced frequencies of fixation of atypical amino acid substitutions in orthologous protein groups. This fact can be explained by the Haldane's dilemma that is in agreement with the scarcity of paleontological findings of intermediate taxa and a poor representation of ancestral genera of the taxon evolving by aromorphic morphological transformation in the paleontological record. On the contrary, when

taxa evolves in an explosion-like manner the fixation of atypical amino acid substitutions significantly accelerates.

Thus, our results clearly show the importance of rare amino acid substitutions as divergence markers leading to new phylogenetic branches in the course of Metazoa evolution.

The work was supported in part by RSF grant 14-24-00123 and by ICG SB RAS budget project VI.61.1. Computing was done at Siberian Supercomputer center SB RAS and at Novosibirsk State University Supercomputer Center.

References:

1. G.A. Bazykin, et al. (2004) Positive selection at sites of multiple amino acid replacements since rat-mouse divergence, *Nature,* **429**:558-562.

2. M.S. Breen, et al. (2012) Epistasis as the primary factor in molecular evolution, *Nature,* **490**:535-538.

3. D.D. Pollock, et al. (2012) Amino acid coevolution induces an evolutionary Stokes shift, *Proc Natl Acad Sci U S A,* **109**:E1352-1359.

4. A.J. Sardar, et al. (2014) The evolution of human cells in terms of protein innovation. *Mol Biol Evol*., **31**:1364-1374.

5. K.V. Gunbin, et al. (2011) Molecular evolution of cyclin proteins in animals and fungi, *BMC Evol Biol.,* **11**:224.

6. K.V. Gunbin, et al. (2012) Computer System for Analysis of Molecular Evolution Modes (SAMEM): analysis of molecular evolution modes at deep inner branches of the phylogenetic tree, *In Silico Biol.,* **11**:109-123.

7. K.V. Gunbin, A. Ruvinsky (2013) Evolution of general transcription factors, *J. Mol Evol.*, **76**:28-47.

8. A. K. Behrensmeyer, and A. Turner (2013) Taxonomic occurrences of Suidae recorded in the Paleobiology Database. Fossilworks. http://fossilworks.org.