Magdalena Karolina Biesiadecka

Experimental test of hypotheses explaining the negative correlation between molecular evolution rate and gene expression level

Summary:

The rate of molecular evolution of can differ by several orders of magnitude between genes. For over a decade, it has been firmly established that genes which have the highest level of expression accumulate the least mutations regardless whether being essential for life or not. Several hypotheses attempting to explain this unexpected discovery have been developed. Some hypotheses focus on expression costs, especially on the metabolism of RNA or protein. These hypotheses assume that selection has conserved highly expressed proteins against mutations as a consequence of avoiding the synthesis of abundant but ineffective proteins which would be a waste of resources. Other hypotheses are focused on the risk appearing when polypeptides adopt improper structure. The basic assumption is that highly expressed proteins evolve slowly in order to maintain a sequence that helps to keep the rate of mistranslation error slow and thus avoid the cost of protein toxicity. Paralogous genes appear especially suitable to evaluate and compare the two kinds of costs. Two paralogs often differ in the rate of evolution and the level of expression. They involve comparable metabolic costs as they are generally similar but the already accumulated differences in sequence may result in unequal costs of toxicity.

The aim of the study was to test the mentioned hypotheses by comparing directly the fitness of expressing paralogous genes. A collection of *Saccharomyces cerevisiae* strains were used in which every strain contained a plasmid with one overexpressed gene. In total, more than six hundred strains grouped in three hundred pairs were analyzed. The paralogous genes were expressed by using a strong inducible promoter. It was expected that a high level

of unnecessary protein would affect fitness of the cell and help to expose differences between pairs of strains with paralogous genes. The concentration of overexpressed protein was measured and found to be very high and equal on average for the fast and slow evolving paralogs, ensuring that the experimental system worked correctly.

Results of two independent experiments did not confirm the expectation that the cost of protein overexpression should be higher for faster evolving paralogs. One experiment was based on measurements of the maximum growth rate done separately for every overexpressing strain. The other one, involved competition within pairs of paralogs. The expected differences did not appear even under experimental environments known to reduce general protein stability. Thus, our results did not support the hypothesis that the natively low expressed proteins evolve under reduced purifying selection because they are rare in cells and therefore less toxic after misfolding. The genes and proteins used in these experiments were of similar properties when compared in paralogous pairs but substantially diverse when compared between pairs. Therefore, a series of analyses was added in which the entire gene collection was used to test the impact of specific features of proteins and mRNAs on the cell fitness. None of the features potentially affecting the translation rate or destabilization risk of tertiary protein structure were found to be a significant determinant of fitness. Neither the number of genetic and physical interactions of a protein nor its functional role as categorized by Gene Ontology were found to have an impact on fitness. Only one basic protein feature, its size, was found to determine the fitness effect. It was also critical for the realized level of protein overexpression. On the one hand, this result is not surprising. On the other hand, it demonstrates that the fitness cost depends mostly, if not entirely, on the quantity of an overproduced polypeptide and not on its quality as determined by its various properties.

The present results were compared with those recently published. Both the results from this PhD thesis and from other new studies appear to contradict the hypotheses linking the rate of molecular evolution with the risk of toxicity stemming from polypeptide destabilization. Furthermore, the idea that the functional importance of a gene may affect the rate of its evolution is only weakly supported by both this and other recent research. The present work points to the metabolic costs of expression as potentially the most important determinant of the fitness cost but this conclusion is not clearly supported by other recent results. In sum, a satisfactory explanation of the negative correlation between the rate of sequence evolution and the level of expression appears still missing.