CHAPTER 11

Correlations between Quantitative Measures of Genome Evolution, Expression and Function

Yuri I. Wolf, Liran Carmel and Eugene V. Koonin*

Abstract

In addition to multiple, complete genome sequences, genome-wide data on biological properties of genes, such as knockout effect, expression levels, protein-protein interactions, and others, are rapidly accumulating. Numerous attempts were made by many groups to examine connections between these properties and quantitative measures of gene evolution. The questions addressed pertain to the most fundamental aspects of biology: what determines the effect of the knockout of a given gene on the phenotype (in particular, is it essential or not) and the rate of a gene's evolution and how are the phenotypic properties and evolution connected? Many significant correlations were detected, e.g., positive correlation between the tendency of a gene to be lost during evolution and sequence evolution rate, and negative correlations between each of the above measures of evolutionary variability and expression level or the phenotypic effect of gene knockout. However, most of these correlations are relatively weak and explain a small fraction of the variation present in the data. We propose that the majority of the relationships between the phenotypic ("input") and evolutionary ("output") variables can be described with a single, composite variable, the gene's "social status in the genomic community", which reflects the biological role of the gene and its mode of evolution. "High-status" genes, involved in house-keeping processes, are more likely to be higher and broader expressed, to have more interaction partners, and to produce lethal or severely impaired knockout mutants. These genes also tend to evolve slower and are less prone to gene loss across various taxonomic groups. "Low-status" genes are expected to be weakly expressed, have fewer interaction partners, and exhibit narrower (and less coherent) phyletic distribution. On average, these genes evolve faster and are more often lost during evolution than high-status genes. The "gene status" notion may serve as a generator of null hypotheses regarding the connections between phenotypic and evolutionary parameters associated with genes. Any deviation from the expected pattern calls for attention—to the quality of the data, the nature of the analyzed relationship, or both.

Introduction

Quantitative genomics involves numerous measures reflecting different aspects of the evolutionary history and the physiological role of a given gene (protein). One can estimate the evolution rate of a gene, measured in different organisms; its expression level in different tissues

*Corresponding Author: Eugene V. Koonin—National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, U.S.A. Email: koonin@ncbi.nlm.nih.gov.

Discovering Biomolecular Mechanisms with Computational Biology,
edited by Frank Eisenhaber.

Table 1. Connections between various measures of sequence evolution rate, gene loss, expression, and fitness effect[a]

		K_{aa}	K_N	K_S	K_5	K_3	PGL	E_H	B_H	E_C	E_Y	E_Y
1	protein evolution rate (K_{aa})	x										
2	CDS non-synonymous evolution rate (K_N)	+	x									
3	CDS synonymous evolution rate (K_S)		+	X								
4	5'-UTR evolution rate (K_5)		+	+	x							
5	3'-UTR evolution rate (K_3)		+	+	+	x						
6	propensity for gene loss (PGL)	+					x					
7	expression level in human (E_H)	-	-	-	0	-	-	x				
8	expression breadth in human (B_H)	-	-	-	0	-		+	x			
9	expression level in *C. elegans* (E_C)	-					-	+		x		
10	expression level in *S. cereviseae* (E_Y)	-					-	+		+	x	
11	viability of gene disruption in *S. cereviseae* (E_Y)	+					+	-	-	-	-	x

[a] The data was from references15, 16.

and in different taxonomic groups; the tendency of a gene to be lost during evolution of different lineages of organisms or its tendency to produce paralogous copies via duplication; its position in the metabolic, signaling and protein interaction networks; and a variety of other quantities (e.g., refs. 1-4). Not unexpectedly, many of such measures are not independent. The literature on the subject (see specific references below) reports numerous positive and negative correlations: between the synonymous and nonsynonymous evolution rates within a gene; between evolution rate and expression level; between propensity of gene loss and fitness effect; and many more (Table 1). Some of these correlations are very strong for quite obvious reasons, such as evolution rates in different lineages or expression levels of orthologous genes; others are less trivial, e.g., the correlation between the degree of conservation of a gene's presence in different lineages and the degree of conservation of its sequence; yet others are remarkably low or absent, sometimes running contrary to expectations (evolution rate vs. number of protein-protein interactions or conservation of gene sequence and that of expression profiles).

Diverse as they are, all these purported correlations, except for the most obvious ones, share one somewhat disturbing feature: although they may be highly statistically significant due to the large number of data points, they typically explain only a small fraction of the variance of the analyzed quantities. Hence considerable debate around many of these observations, which is further compounded by problems with the completeness and quality of much of the data involved, particularly that coming from genome-scale analyses of gene expression, protein-protein interactions, and other aspects of gene functioning. For example, the argument about the link—or lack thereof—between the connectivity of a protein in protein-protein interaction networks