GO-Genesis: finding the origin of biological processes and molecular functions from Gene Ontology

Carlos Gonçalves[1], J.M. Ortega[1]

[1]*Laboratório de Biodados, Departamento de Bioquímica e Imunologia, ICB, UFMG*

Gene Ontology (GO) is a database comprised of terms that can be annotated to proteins to describe biological information. These terms are hierarchically organized in three distinct ontologies (Biological Process, Molecular Function and Cellular Component), with more generic terms being parents of more specific ones. Biological Processes are complex systems that require the participation of several gene products, such as "photosynthesis" or "innate immune system"; Molecular Functions are sequence-related properties of the genes, such as being able to catalyze a reaction (i.e., "phosphatase activity") or being able to interact with another substance (like "calcium ion binding"); Cellular Components describe regions within the cell or its periphery where the gene product can be located, such as "cell membrane", "nucleus" or "synapse". Several of these biological processes and molecular functions are clearly very ancient, existing since the origin of life, whereas others have to be more recent innovations on the course of evolution, so we decided to estimate the moment of origin for all of them. For each GO term, we determined all organisms that had at least one protein annotated to that particular term; with that, we calculated the lowest common ancestor (LCA) for those organisms – that is, the clade on which the function or the process itself originated, at least according to the existing Gene Ontology annotations. Given the fact that each individual term had its origin independently determined, there were some cases on which a given term could be dated more recently than one of its children; since, by logic, this is an invalid result, we corrected the LCA of those ancestor terms to match the origin of the oldest of their children terms. Overall, most of the terms (over 53%) of molecular functions existing on the *Homo sapiens* are very ancient, dating back to the origin of the cellular organisms, with an expressive number also appearing on the ancient clades of Eukaryota (13%) and Opisthokonta (4%). More recently, peaks of origin of terms are observed on the clades of Bilateria (7%), Euteleostomi (4%) and Amniota (10%). As for the biological processes, about 23% of them originated on the cellular organisms, with several others also appearing on the early clades of Eukaryota (13%) and Opisthokonta (6%). Surprisingly, the clade which had most processes being originated was Amniota (26%), a rather recent one. The origin of the GO terms is available for consultation at http://biodados.icb.ufmg.br/GO-Genesis.

Financial support: CAPES