Niji: Analysis on the origin of biological systems using KEGG Pathways

Carlos Alberto Xavier Gonçalves[1], José Miguel Ortega[2],

1 UFMG
2 Universidade Federal de Minas Gerais. Laboratório de Biodados

Abstract

The Kyoto Encyclopedia of Genes and Genomes (Kegg) contains hundreds of pathways representing biological systems involved in metabolism, signaling, diseases and several other topics. These pathways are described in XML files and are graphically depicted in image files within Kegg's database. Kegg also contains data on clusters of orthologues, with which it is possible to obtain the taxonomic distribution of any given gene present in those pathways; by knowing all the organisms that contain a certain gene, it is possible do determine the lowest common ancestor (LCA) to those organisms, allowing us to infer the clade of origin of that gene. By using a local database containing LCA information for all genes on Kegg and also Kegg's automated programming interface (API), we generated colorized Kegg Pathways for the Homo sapiens in a way that each gene box's color is a representation of that gene's LCA; thus, genes that originate on the same clade are colorized with the same color. This allowed us to analyze how biological systems evolved over time. We also utilized Python scripts to recreate each pathway in graph objects, using the information contained in the XML files, and applied the LCA data to discover if the pathways were formed from a single connected component, or if they evolved from multiple subsystems that eventually coalesced. Of the 314 Kegg maps analyzed, 35 did not contain any edge information on Homo sapiens. We encountered 46 systems that reach full connectivity on the Homo sapiens, meaning no elements on those systems are disconnected at the most recent clade. Of these, 15 (32%) evolved on a single growing component, with new elements connecting directly to previously existing entities, whereas 31 (68%) evolved from multiple coalescing subsystems. Interestingly, six (13%) of the 46 fully-connected pathways are entirely ancient, with all elements dating back to the origin of eukaryotes, while there are seven (15%) maps containing up to early animals genes (from Metazoa through Bilateria). The remaining 33 (72%) maps have genes originated within the chordates. Most of these pathways reached full completeness within the Euteleostomi (modern fishes) clade, and some are as recent as the placental mammals (Theria and Eutheria clades). We created an online platform for consultation of these data, called Niji (the Japanese word for rainbow), that is available at: biodados.icb.ufmg.br/niji

Funding: CAPES, CNPq, FAPEMIG