ProteinWorldDB in a multidimensional representation of the Network of Life

Edson Machado[1], Marcos Catanho[1], Paulo Carvalho[2] and Wim Degrave[1]

[1]*Oswaldo Cruz Institute, FIOCRUZ, Rio de Janeiro, Brazil*
[2]*Carlos Chagas Institute, FIOCRUZ, Curitiba, Brazil*

Inter-genomic distance estimations reflect genome complexities and divergence, but are not yet well understood. Calculations of such distances with the purpose of constructing evolutionary relationships are based on the selection of a subset of homologous sequences (orthologs) and provide understanding of evolutionary relationships between genes, proteins and species, with the phylogenetic tree being a primary tool in analysis and visualization. However, inferring the "true tree" is fundamentally a difficult problem, and the traditional two-dimensional visualization is hard to interpret when a larger amount of organisms is involved. In addition, this approach ignores most of the unique traits and dissimilar aspects of the genome organization and coding/non-coding potential of organisms, paralogs, putative proteins of unknown function, etc. One can thus express inter-genomic distances comparing overall nucleotide sequences, which is not uniform due to a high variability in genetic/genomic complexities and GC bias, or by comparing the complete predicted protein set for each genome. We adapted a genomic distance method, originally based on protein similarities scores measured in bidirectional comparisons with BLAST, to use protein similarities scores measured in unidirectional comparisons using Smith-Waterman algorithm, with SSEARCH program, to infer distances between genomes and construct a distance matrix. As an example, we used data of ProteinWorldDB to apply our genomic distance method to infer distances between 210 species (117 bacterias, 49 eukaryotes and 44 archaeas) in order to construct an initial representation of the Network of Life. ProteinWorldDB stores the results of the "Uncovering Genome Mysteries" project, which examined close to 200 million predicted protein sequences from a wide variety of life forms. Those protein sequences were compared against each other through the IBM World Community Grid with the SSEARCH program to assess their similarity. This represents about 20 quadrillion ($2x10^{16}$) comparisons and the total computation time is projected to take the equivalent of one computer running continuously for 40,000 years. In addition, the results of the thus measured distances between the 210 species under analysis were used to construct a three-dimensional representation of the Network of Life. Support: Fiocruz, CNPq, IBM.