Homology detection using multilayer maximum clustering coefficient

Caio Rafael do Nascimento Santiago[1], Luciano Antonio Digiampietri[1],

1 Universidade de São Paulo

Abstract

Sequence clustering is an important tool for helping the understanding of homology relations in protein sets, by grouping them into related families. The identification of these families is not a trivial task, and there are many studies dedicated to solving it, the majority of them are graph-based ones. In the graph-model, nodes represent sequences and edges represent homology relations, in general, defined considering metrics obtained from local alignments.

Some studies use the concept of transitivity to explain the homology, i.e., if two proteins are homologous and a third is homologous to one of the two, then the three are considered a family of homologous proteins. The main concern about this approach is the establishment of how much this proposition could be extended based only on transitivity.

This work presents a graph-based clustering method that maximizes the clustering coefficient based entirely on the transitivity of the homology relationship. It creates a undirected graph considering the e-value metric (or any other alignment metric chosen by the user) and produces a multilayer list of gene families according to thresholds progressively more restrictive. This allows to the user to work with genes families composed of genes with greater distances (first layers) and more restricted families (with more similar genes) in the last layers.

Some advantages of this approach are: it preserves the graph constructed considering the local alignments and, therefore, it is easy to understand the process that generated a certain family; it is possible to analyze the topology of the graph, in order to, for example, find multi-domain proteins or find the proteins that phylogenetically separate two families.

This approach was tested in two phylogenetically closely related sets of genomes, the first contains 69 strains from Xanthomonadaceae family and the second contains 55 Streptococcus pyogenes strains. The results from our approach were compared with the TribeMCL results. In the case studies, our solution identified a bigger core genome, considering the number of homologous families (4% bigger than the one identified by TribeMCL), and, when ignoring the paralogs genes, our approach identified a core with 42% more homologous families than the one identified by TribeMCL. When analyzing the gene annotation/products in the homologous families, our solution was 6% better in grouping genes with the same annotations in the same family when compared with the families produced by TribeMCL and using the annotations provided by Patric. Finally, the biggest families produced by our approach are smaller than the ones produced by TribeMCL (from 9% to 15% smaller), being able to not group together genes that are not very similar and have different annotations.

Funding: Capes