Using HipMCL, a high-performance parallel implementation of the Markov clustering algorithm, to understand microbial diversity

Georgios A. Pavlopoulos

Biomedical Sciences Research Center "Alexander Fleming", Athens, Greece

While various clustering algorithms have been proposed to find highly connected regions within a biological network, Markov Clustering (MCL) has been one of the most successful approaches. Despite its popularity, MCL's scalability suffers from high running times and memory demands. Here, we present the High-performance MCL (HipMCL), a parallel implementation of the original MCL algorithm that can run on distributed-memory computers. HipMCL can efficiently utilize 2000 compute nodes and cluster a network of ∼70 million nodes with ∼68 billion edges in ∼2.4 h. To demonstrate its capabilities, we show how we use HipMCL to understand the biological diversity and discover novel protein clusters.