**Creating an efficient compression based DNA sequence clustering algorithm for the analysis and comparison of metagenomes**

Abstract

DNA clustering algorithms allow the comparison and analysis of metagenomes (environmental genomes) that contain unknown and uncultivated microorganisms otherwise difficult to analyze using traditional genomic methods. Although successful, these algorithms lack the efficiency and accuracy needed to push Metagenomics as a common and accessible lab procedure. The compression based clustering approach seeks to address these gaps. By applying a lossless compression algorithm to the DNA sequences the size of these sequences can be reduced allowing increased efficiency. By creating a similarity method attuned to the compressed sequences, these compressed sequences can be clustered using any current clustering method. We demonstrate the accuracy and diversity of compression based clustering by testing six sample algorithms created by this approach, extensively over a wide range of data sets, concluding that compression based clustering has the potential to bridge the gaps in current clustering algorithms.