Abstract

Uclust is currently the main algorithm for sequence clustering of large sets of RNA and DNA sequences, as its speed remains unsurpassed with the level of precision it has. With its high price, there is a demand for an alternative to Uclust which can compete in both speed and precision.

In this paper, I present two new greedy algorithms for competing with Uclust. As distance metric, one uses a sketch developed by the author of this paper which is a slight modification to MinHash, and the other uses a similarly developed modification of MinHash suggested in an earlier paper. We shall formulate the mathematics behind MinHash and the slight modifications made to it for the two algorithms. A detailed explanation of the two algorithms ensues, using pseudo-code. To increase the speed of both algorithms, they have been developed in MapReduce Frameworks, using Pig Latin. The Pig scripts will be examined to explain the MapReduce flow of both algorithms.

I have evaluated the speed and precision of both algorithms rigorously using real metagenome data. On this ground, I found that both algorithms are potentially quicker than Uclust, and one of them with a precision level that competes with Uclust.