UNIVERSITY OF THE WITWATERSRAND

Abstract

School of Electrical and Information Engineering
Engineering and the Build Environment

Master of Science in Engineering

A Machine Learning Approach to DNA Shotgun Sequence Assembly

by Radu-Ionut Constantinescu

DNA sequencing and assembly is becoming increasingly prevalent in the field of bioinformatics. It is used in a variety of fields such as forensics and genetic engineering in order to sequence DNA of a species or specific individuals. The high computational complexity associated with DNA sequencing and assembly makes the process expensive to implement. In order to help reduce this complexity, a read grouping machine learning approach, which breaks the problem of assembly into multiple smaller sub-problems, is proposed. The shotgun sequencing process was performed on a 50456 base pair portion of the Drosophila Melanogaster (fruit fly) genome. The sequencing and assembly process was simulated under varying conditions of read size, coverage depth and sequencing error rates. The greedy and de Bruijn algorithms were first implemented as stand-alone assemblers and their performance was compared. Thereafter, a neural network system was implemented together with each of the two assemblers in order to investigate the effects a read grouping approach has on assembly performance. The performance of each of the four assemblers was then compared in terms of computational complexity and assembly accuracy using information theoretic principles along with a proposed coverage metric. It was found that the simulation time of the stand-alone greedy assembler was significantly improved when combined with the neural network read grouping approach. However, due to the higher relative complexity associated with the neural network training and grouping process, the same can not be said about the de Bruijn assembler. In order for the de Bruin assembler to benefit from this "divide and conquer" approach, faster grouping techniques need to be implemented.