Master Project Proposal

Implementing a Statistical Model for Clustering Metagenomic Contigs, Using Sequence Signatures and Correlation Over Multiple Samples, by Efficient Parallel Computing and Test Driven Development

Brynjar Smári Bjarnason 840824-4690
Computational and Systems Biology
School of Computer Science and Communication
KTH Royal Institute of Technology

CSC Supervisor: Jens Lagergren
School of Computer Science and Communication
KTH Royal Institute of Technology

Supervisor: Anders Andersson
School of Biotechnology
KTH Royal Institute of Technology

Co-supervisor: Christopher Quince
School of Engineering
University of Glasgow

January 31, 2013

The Problem

In metagenomic research, which is an analysis of microbial communities from environmental samples, we want to know which organisms are found in samples such as water, soil and human gut. Most of these organisms have not been sequenced and many can not be cultured in laboratory. With massively parallel sequencing (MPS) we get short reads of all genetic material in these samples which are assembled into continuous genetic strands called contigs. Each contig represents a fragment of an organisms genome, and to reconstruct the genomes in the sample the contigs need to be clustered. This is referred to as the binning process and after, each bin should contain only contigs from one organism or very related organisms. This allows us to examine which organisms are in the sample, in what abundance and mutations within organism.

While assembly has evolved and is quite established, the binning process has lagged behind.

Description

The main goal of the two master projects is to create an efficient and statistically modelled pipeline for using correlation between multiple samples and sequence signatures to cluster contigs into bins representing the organisms found in these samples.

This master project will focus on implementing the statistical models and binning algorithms. The models and algorithms will be analysed and optimized for efficient parallel computing on a cluster such as the Uppmax's Kalkyl cluster using test driven development. To evaluate quality and performance, both simulated and real metagenomic datasets will be used.