Master Project Proposal

Implementing and Evaluating a Statistical Model for Clustering Metagenomic Contigs

Brynjar Smári Bjarnason 840824-4690
Computational and Systems Biology
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 11, 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 into bins. This is referred to as the binning process. After this process, each bin should contain only contigs from one organism. This allows us to examine which organisms are in the sample, in what abundance and mutations within organism.

The goal with this project (and a project run in parallel) is to develop a rigorous statistical model and efficient implementation for a clustering method using multiple samples from the same environment taken over period of time or from close regions. Using multiple samples will hopefully give us better view of the organisms in that environment since microbial communities vary over time and place and contigs from the same organism should co-vary in abundance across samples (as opposed to contigs from different organisms).

Description

The focus of this project is implementation and optimization of a statistical model for clustering contigs from multiple samples. The statistical model will be designed in parallel so many models need to be compared for quality and efficiency, also comparison to current binning projects should be done. Therefore a framework for comparing different models and implementations will be developed.