Spark-Based Feature Linking for Scalable High-Throughput Metabolomics Data Processing

Introduction
Metabolomics is the study of the small molecules involved in metabolic pathways, referred to as metabolites, through their identification and quantification [1]. Liquid chromatography-mass spectrometry (LC-MS) gained momentum in metabolomics, as it provides good metabolites coverage along with its high-throughput nature. In fact, a modern mass spectrometer is able to produce 35K spectra per hour [2]. Therefore, automated approaches need to be adopted in order to process the raw data produced by the machines.

In the coming decade a significant part of the EU-citizens will have their genome determined routinely, and linking this to metabolomics data, from biofluid samples, will enable personalized and evidence-based medicine. This will imply an increase in the size of the datasets, hence parallelized data processing solutions in metabolomics need to be developed. While most of the tasks in MS data processing are trivially parallelizable, feature linking, which needs to consider multiple spectra in a non-independent fashion, is challenging.

Project description
Unsupervised learning has been used in order to implement feature linking in some popular software tools. For instance, OpenMS uses a modified version of the Quality-Threshold (QT) algorithm [2]. Apache Spark (http://spark.apache.org/) is a general cluster-computing engine for large-scale data processing, which includes parallel implementations of many commonly used supervised/unsupervised learning algorithms. The aim of this project is to implement a scalable tool for MS data feature linking, using Spark-based unsupervised learning.

Spark machine learning library (MLlib) offers a distributed implementation of K-means. In K-means the number of clusters $K$, in the the dataset, is assumed to be known. This contrasts with the QT algorithm, where a threshold is used to tentatively add points to a number of clusters that is not known a priori. In fact, we do not know the number features groups in the mass spectrometry data. Hence, tuning the $K$ parameter will be crucial. Various approaches are reported in the literature e.g. $K=\sqrt{n/2}$ rule of thumb [3], gap statistic [4]. Choosing the right approach for metabolomics data will have major role in this project. An alternative, more advanced, approach would be to hack the K-means implementation in Spark, in order to make it behave like the QT algorithm.

The applicability of the tool will be tested on metabolomics data available in the Kultima lab.

Supervisor:
Ola Spjuth, Docent
Department of Pharmaceutical Biosciences and Science for Life Laboratory
Uppsala University
ola.spjuth@farmbio.uu.se
References