Using Frequent Itemset Mining to Find Sets of Co-occurring Genomic Tracks
Boris Simovski, Geir Kjetil Sandve
University of Oslo

While immense amounts of genomic data are now publicly available, analyzing the data is a complicated and at times resource exhaustive task. A well established analysis is the computation of pairwise overlap between two genomic tracks. However, in certain situations it is valuable to consider a larger number of genomic tracks and e.g. discover subsets of the tracks that occur together at the same locations along the genome. An example of such a problem is to find combinations of transcription factor (TF) ChIP-seq tracks that occur at the same locations in the genome, either from a set of tracks for different TFs or from a set of tracks for the same TF in different cells/settings.

The problem at hand can be translated into a more general problem within the field of data mining, called frequent itemset mining. According to the itemset mining terminology, we take the genomic tracks to represent items and the base-pair positions of the genome to represent transactions.

Our Galaxy-based web tool at the Genomic HyperBrowser web server enables the user to run frequent itemset mining on large sets of genomic tracks. The result is a list of track combinations that occur together on at least a minimum number of base pairs along the genome. We present results for two different approaches, based on the breadth-first Apriori and the depth-first Eclat algorithm.

Additionally, we introduce another mining technique that can be of interest. We use the expected support of a given itemset, multiplied by a factor, as the decision threshold whether the itemset is frequent or not. The resulting itemsets are relatively frequent with regard to the chosen factor.