Using Frequent Itemset Mining to Find Sets of Co-occurring Genomic Tracks

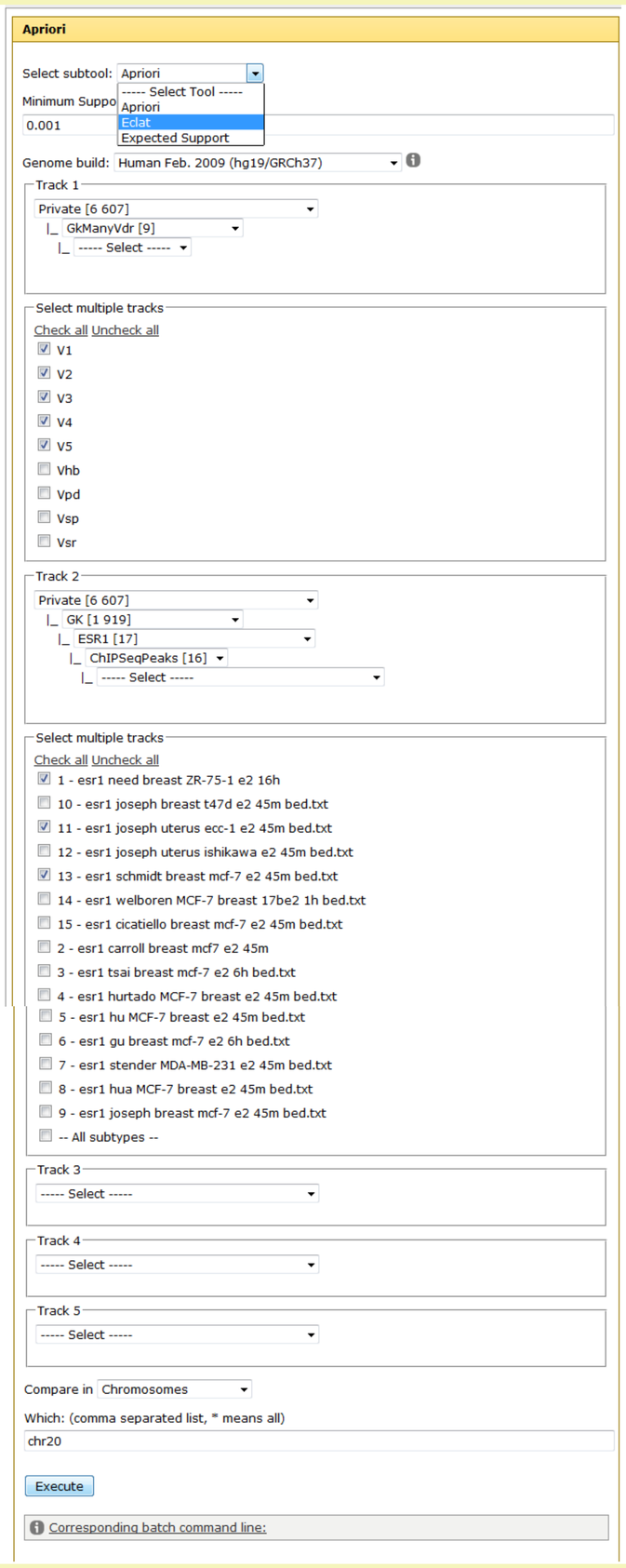
Boris Simovski, Geir Kjetil Sandve
University of Oslo

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.

Aditionally, 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 wether the itemset is frequent or not. The resulting itemsets are relatively frequent with regard to the chosen factor.



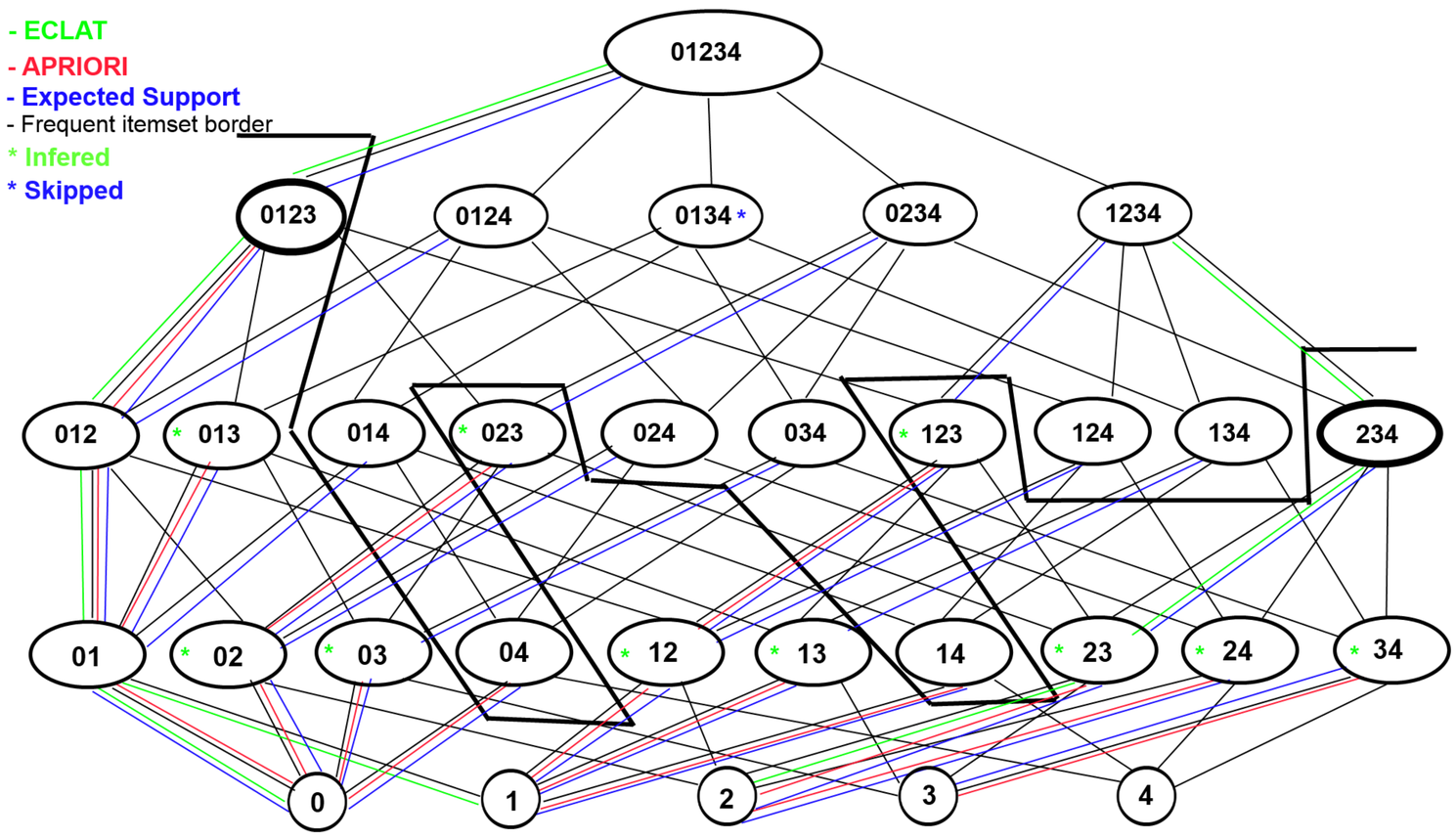


Fig 1. Latice representation of the search space; Explore routes of the algorithms.

Tracks

Track index	Track name
0	VDR 1
1	VDR 2
2	VDR 3
3	VDR 4

Frequent itemsets

Level	Itemset	Support ▾
1	[0]	0.00121895067268
1	[1]	0.00054639771318
1	[3]	0.000524009956602
1	[2]	0.000481725497862
2	[3, 0]	0.000239553755368
2	[2, 0]	0.000189732667021
2	[1, 0]	0.000164282658834
2	[2, 3]	0.000157174427121
3	[2, 0, 3]	0.000110939187808
2	[1, 2]	7.67466892776e-05
2	[1, 3]	7.10664505426e-05
3	[1, 0, 3]	6.20383616034e-05
3	[1, 0, 2]	5.44223990536e-05
3	[1, 2, 3]	3.07177156174e-05
4	[1, 2, 3, 0]	2.99561193624e-05

Table 1. Frequent itemsets and their corresponding supports.

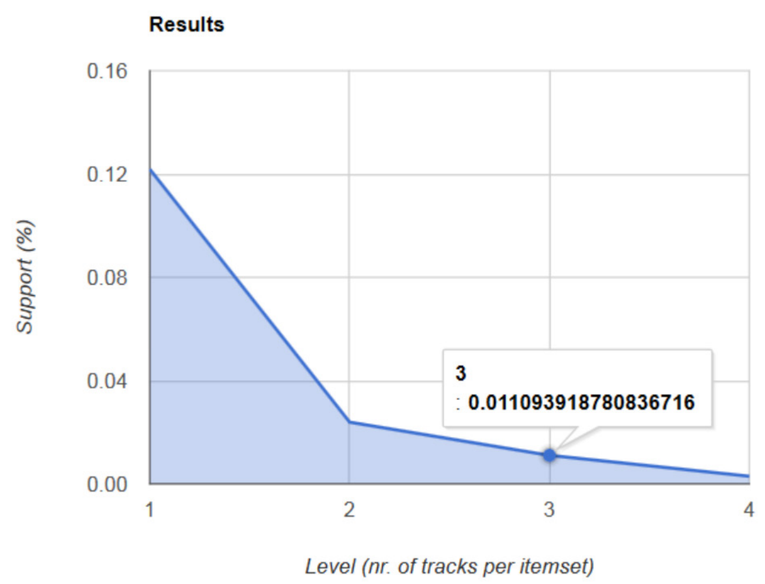


Chart 1. Pareto front of the maximum frequent itemset supports per level.

= The Apriori algorithm uses a breadth-first strategy.
+ Will calculate the support of all frequent itemsets
- Will perform more poorly than ECLAT when larger size frequent itemsets are expected

= The ECLAT algorithm uses a depth-first strategy with the intent to discover **maximal frequent itemsets,** from which all of its subsets can be infered as frequent.
+ Will execute significantly faster than other algorithms (when larger size frequent itemsets are expected, i.e. the minimum support is small)
- If the frequent itemsets are of small size it will perform more poorly than Apriori
- Support of infered frequent itemsets is not calculated (inherited from maximal)

= The Expected Support mining algorithm defines the frequent itemset differently. It takes into account the "frequency" of all the single items in the itemset and an itemset is designated frequent when its support is larger than its expected support by a selected factor.
+ Gives better insight into the interconnection of the items (genomic tracks)
- Performance is poor in regard to time of execution, almost all subsets will be tested for frequency

The results of the frequent itemset mining can reveal interesting combinations of genomic tracks, which can be followed up with more detailed analyses into how they are related.

Agrawal, Rakesh, et al. "Fast discovery of association rules." Advances in knowledge discovery and data mining 12 (1996): 307-328.

Zaki, Mohammed J., et al. "New algorithms for fast discovery of association rules." 3rd Intl. Conf. on Knowledge Discovery and Data Mining. Vol. 20. 1997.

Goethals, Bart. "Survey on frequent pattern mining." Univ. of Helsinki (2003).

Gundersen, Sveinung, et al. "Identifying elemental genomic track types and representing them uniformly." BMC bioinformatics 12.1 (2011): 494.