Approximate Subspace Pattern Mining for Mapping Copy-Number Variations

Nicholas Johnson, Gang Fang, and Rui Kuang

Department of Computer Science and Engineering
University of Minnesota
Minneapolis, MN

Abstract

DNA copy-number variations (CNVs) are genome aberrations that could disrupt normal biological functions and lead to tumor genesis. Identifying causal copy-number variations is an important step in understanding the molecular mechanisms of cancer. In this paper, we introduce an open-source tool for mining CNV subspace patterns (SubPatCNV). SubPatCNV is an approximate association pattern mining algorithm under a spatial constraint on the positional CNV probes. In the experiments on a bladder cancer dataset, SubPatCNV discovered many large aberrant CNV events in patient subgroups and reported CNV regions highly specific to clinical variables and enriched with more known oncogenes than other existing CNV discovery methods.

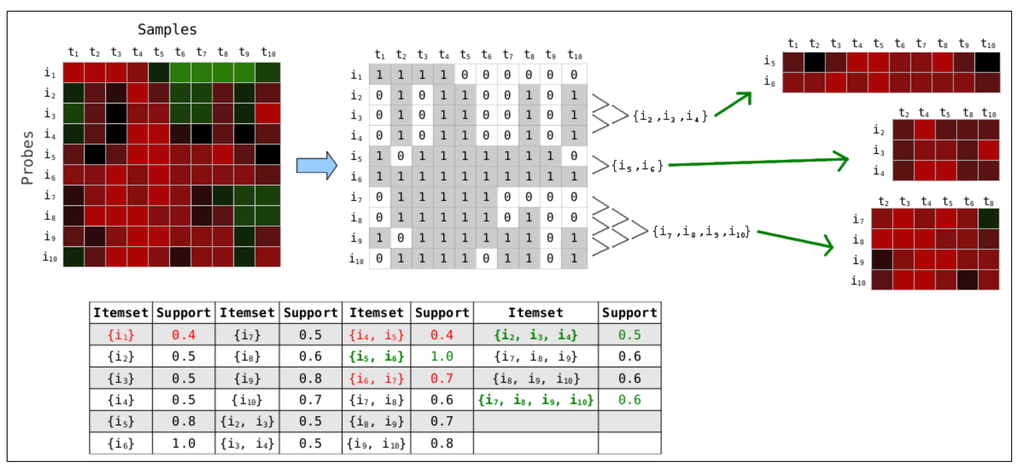
Background

Cancers are thought to develop through the accumulation of genome aberrations such as mutations and structural variations. DNA copy-number variations are one type of structural aberrations that could disrupt normal biological functions. One recent tool for CNV identification called "Genomic Identification of Significant Targets in Cancer" (GISTIC) [5] takes a statistical approach by calculating a probe summary statistic (G-score) for the positional CNV across all patients in the dataset. Using this G-score and a calculated q-value GISTIC uses a "peel-off" heuristic procedure that iteratively selects CNV regions with the greatest G-score within each continuous region of significant CNVs in order to grow the positions into a “peak region”. Another tool called JISTIC [6] improves upon GISTIC by introducing a variation of the "peel-off" search heuristic called "limited peel-off“ which works by only peeling off the G-score portions that contribute to the current peak region. Two limitations of these methods are: (1) calculations of G-scores do not utilize the positional dependence between probes and (2) they do not analyze all potential aberrant CNV regions due to greedy-like search heuristics. These limitations prevent them from discovering potential candidate aberrant regions of interest to patient subgroups.

Methods

Approximate Frequent Patterns: Given a MxN matrix of items and transactions where transactions contain a subset of items, the number of transactions that an item is in, $|T' : I' \subset T'|$, is called the item's support. An itemset I' is considered a frequent itemset if its support is greater than a user-specified threshold $minsup$. Frequent patterns are very useful, but they are not effective in discovering associations within noisy data. As such, we allow a specified percentage of items to be missing from an itemset which are called error-tolerant itemsets (ETIs). An ETI I' is said to be weak with tolerance ε if $\exists T' \in T : |T'| \geq minsup$ and $\frac{\sum_{i \in I'} \sum_{t \in T'} D_{i,t}}{|I'| * |T'|} \geq 1 - \varepsilon$. However, weak ETI’s suffer from spurious items. The algorithm in [7] was developed to overcome spurious items by adding a recursive constraint to the weak ETI definition. In this algorithm an itemset I' is a recursive weak ETI if I' and all subsets of I' are weak ETIs.

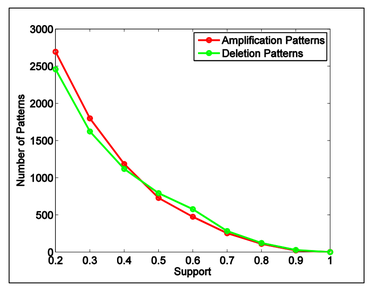
Subspace Patterns for Copy-Number Variation Discovery: Our algorithm has two changes catered to CNV data. First, SubPatCNV implements a CNV specific pruning heuristic to take advantage of natural correlations within CNV data (nearby probes are positively correlated) in order to work on large scale CNV datasets. As such, our pruning strategy only considers sets of items (probes) that are continuous. Second, SubPatCNV limits itemset merging by requiring the two itemsets to be similar in support. SubPatCNV proceeds in rounds by considering each individual frequent itemset from the previous round and tries to merge them together. Each itemset will be merged if two criteria are passed: (1) itemsets are neighbors going down the chromosome and (2) the two frequent itemsets to be merged need to be similar enough in support. When both criteria are passed the two itemsets are merged together into a candidate frequent itemset which is then subjected to the recursive weak ETI criteria. If an itemset passes this criteria it is deemed a frequent itemset. This is illustrated in the figure below.

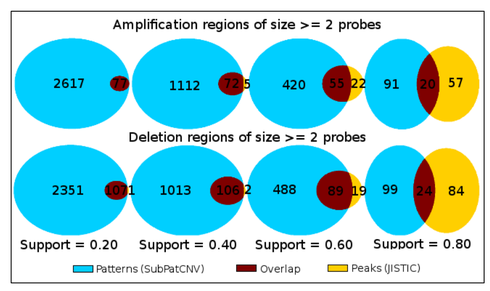


Itemset	Support	Itemset	Support	Itemset	Support	Itemset	Support
{i1}	0.4	{i7}	0.5	{i4, i5}	0.4	{i2, i3, i4}	0.5
{i2}	0.5	{i8}	0.6	{i5, i6}	1.0	{i7, i8, i9}	0.6
{i3}	0.5	{i9}	0.8	{i6, i7}	0.7	{i8, i9, i10}	0.6
{i4}	0.5	{i10}	0.7	{i7, i8}	0.6	{i7, i8, i9, i10}	0.6
{i5}	0.8	{i2, i3}	0.5	{i8, i9}	0.7		
{i6}	1.0	{i3, i4}	0.5	{i9, i10}	0.8		

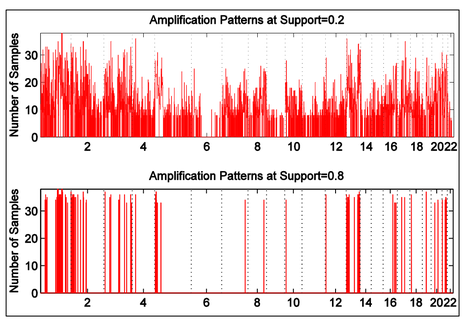
Results and Discussion

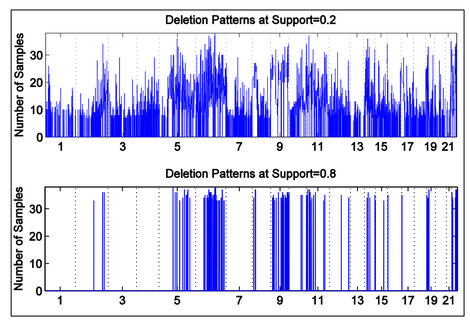
The most important user-defined parameter for SubPatCNV is support. This value is likened to the statistical significance of a pattern, that is, if a pattern has a $minsup = 1.0$ then all patients in the dataset have a log intensity value outside the range of normal CNVs. By varying the parameter $minsup$, a user can control the type of the patterns to be discovered. We show in the next figures how the value of $minsup$ affects the number of discovered patterns and how many of these overlap with peaks discovered by JISTIC.



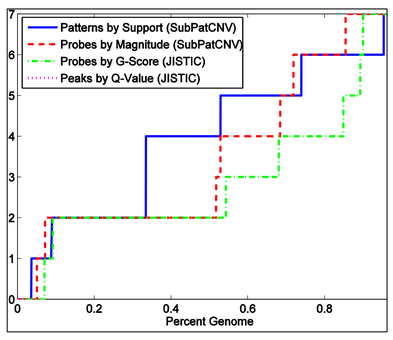


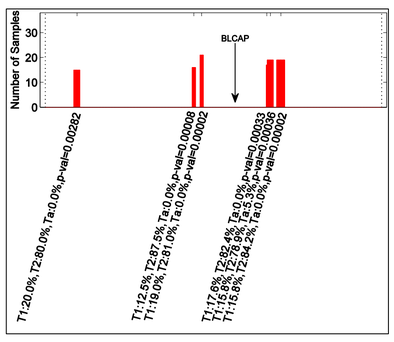
To get a better understanding of the aberrant region discovery we illustrate the positions of each pattern discovered with $minsup = 0.20$ and $minsup = 0.80$ in a genome-wide scale.



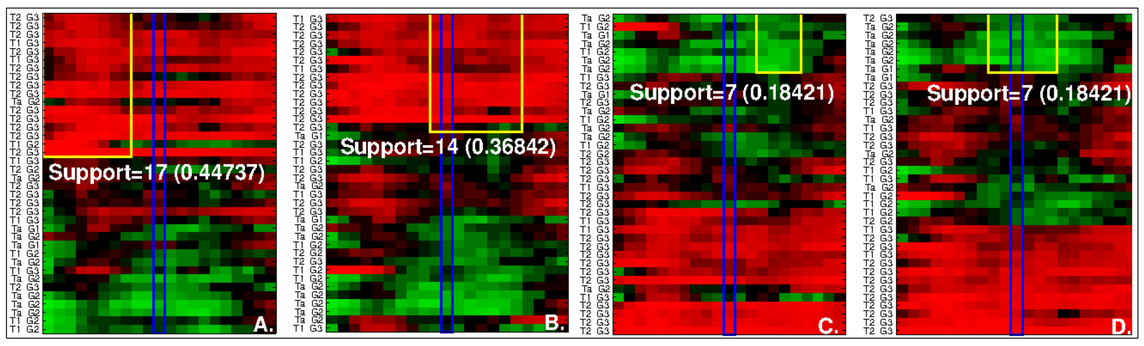


To better measure the cancer relevance of the discovered aberrant CNV regions, we observed how many known oncogenes are overlap with the candidate aberrant CNV regions. We ran SubPatCNV with $minsup = 0.20$ and JISTIC with q-value threshold of 0.15. We then sorted the patterns discovered by SubPatCNV by support or average log-intensity value within the pattern in ascending order and compared with the sorted JISTIC peaks by the G-score or q-value in ascending order. We then plotted the overlap between the patterns/peaks and the oncogenes in the figure below. In both ways of sorting the patterns discovered by SubPatCNV overlap with more oncogenes than JISTIC. SubPatCNV patterns are naturally associated with patient subsets in the support group thus it is able to discover patient subgroup specific aberrant regions and provide information about which patients are associated with the pattern while JISTIC does not explicitly provide this information. The figure below shows a select few amplification patterns discovered on chromosome 20 with $minsup = 0.40$.





Finally, we visualize the original CNV probe log-intensities within several patterns on chromosome 20. The pattern in the Figure (A) below contains shows probes that are highly amplified for a subgroup of 17 patients with 88.2% of the patient samples annotated with tumor grade “G3". Figure (B) shows another highly amplified pattern with 85.7% of the patients in tumor stage "T2" and 100% of the patients in tumor grade "G3". This observation suggests that this pattern may be used as a biomarker to predict patient tumor stage and grade. In Figure (C) a low supported but highly deleted pattern with 85.7% of the patients in tumor grade "G2" is shown. In Figure (D) a low supported but highly deleted pattern with 85.7% of the patients having tumor stage "Ta" is shown.



Conclusion

Identifying causal CNVs driving cancer development is a difficult problem. SubPatCNV is an easy to use, open-source software tool that provides the flexibility of identifying aberrant CNV regions specific to patient subgroups of different sizes. SubPatCNV is freely available at http://compbio.cs.umn.edu/SubPatCNV/

References and Funding

All references can be found in the paper version of this poster.

This work is supported by NSF grant # IIS1117153, NSF grant # IIS0916439 and a University of Minnesota Rochester Biomedical Informatics and Computational Biology Program Traineeship Award.