MEMNAR: Finding Mutually Exclusive Mutation Sets through Negative Association Rule Mining

Iman Deznabi, Ahmet Alparslan Celik, Oznur Tastan

Abstract—It has been reported in multiple cancers that certain set of gene mutations tend not to occur concurrently in the same patient. This mutual exclusivity pattern hints at a functional relation and can help uncover cancer-driver alterations. We address the problem of discovering mutually exclusive mutation gene sets through mining negative association rules. Our proposed algorithm, MEMNAR, efficiently mines for negative association rules in patient mutation data and constructs mutually exclusive gene sets based on these extracted rules with high accuracy. We also define and detect more complex mutual exclusivity patterns that have not been addressed in earlier approaches. Evaluations on simulated data sets demonstrate that MEMNAR can discover mutual exclusive gene sets faster with improved accuracy compared to the state-of-the-art methods. When we apply MEMNAR on breast cancer, we identify several mutually exclusive gene sets that are biologically relevant and some of which have not been reported in the literature.

1 Introduction

Cancer genomes harbor many genomic alterations. However, only a small portion of these act as ”drivers” while the rest are ”passengers” with no significant effect on cancer. Distinguishing driver and passenger mutations is critical for understanding mechanisms underlying cancer. Candidate driver genes are typically determined by conducting statistical tests of mutational frequency [8]. However, individual tumors exhibit a high level of diversity with different combinations of mutations, limiting the utility of statistical tests that detect drivers based on recurrence. The combinatorial patterns of mutations can help understanding the functional relations of genes in cancer. One method of identifying driver genes is to investigate the frequently observed patterns of mutations. One such interesting pattern is mutual exclusivity, where a set of mutated genes rarely co-occurs in the same tumor.

The computational problem of discovering mutually exclusive sets are addressed in the literature with different approaches [11], [1], [2], [10], [7], [15], [9]. Some of these methods are limited by computational efficiency and some suffer from high false discovery rate. In this work, we propose a new approach wherein we formalize the problem of finding mutually exclusive gene sets as a negative association rule mining problem. We present an algorithm, MEMNAR that exploits efficient data structures and pruning strategies developed for frequent item set mining.

As a second contribution, we define a novel mutual exclusivity pattern of the following form: a mutation in a gene, say X, induces that a set of genes, say Y and Z, are not simultaneously mutated with gene X. A patient that is mutated in X and Y but not Z, or a patient mutated in X and Z but not Y supports this pattern; while a patient that contains all three X, Y and Z mutations contradicts with the pattern. MEMNAR discovered such novel patterns in breast cancer (BRCA) mutation data.

Our results on simulated data demonstrate that MEMNAR can outperform the compared state-of-the-art approaches in terms of accuracy and runtime. Applying MEMNAR on BRCA somatic mutation data verifies known mutually exclusive sets as well as discovers several other novel mutually exclusive sets.

2 Methods

2.1 Association Rules

Positive and negative association rules are used in market basket analysis, where the aim is to detect interesting relations in large customer transactions [14]. Positive association rules aim at finding products that are often bought along with other products. In contrast, a negative association (or ”disassociation”) rule looks for item sets that are rarely bought with other items, i.e. ”Customers that buy Coke are unlikely to buy Pepsi as well”. Below, we formally define these concepts in the context of cancer mutation data.

Let $M = \{m_1, m_2, ..., m_N\}$ be the set of N distinct genes that are found to be mutated in a patient cohort, D. Let $T_i \subset M$ be the set of mutated genes in patient i. *Support* of a mutation set, $X \in M$, $supp(X)$, is defined as the fraction of patients with mutations in X. A negation of a mutation set is denoted as $\neg X$ and indicates the absence of all the mutations in X.

A positive association rule is of the form $X \rightarrow Y$, where $X, Y \in M$ and $X \cap Y = \phi$. X is the *antecedent* of the rule and Y is the *consequent* of the rule. The association's strength is measured with its *support* and *confidence* and with high statistical significance. The support of a rule $X \rightarrow Y$ is defined as the percentage of patients that harbor mutations in genes of X and Y : $supp(X \rightarrow Y) = supp(X \cup Y)$. On the other hand, confidence represents the fraction of patients with mutations in Y among those that have mutations in the genes of X; it is the conditional probability of observing Y given X: Confidence, $c(X \rightarrow Y) = supp(X \cup Y)/supp(X)$.

A negative association rule includes at least one negative mutation set either in the antecedent or the consequent of the rule [16]; it can take one of these forms: $X \rightarrow \neg Y$,

- *All authors are with the Computer Engineering Department, Bilkent University, Ankara, Turkey.*
  E-mail: oznur.tastan@cs.bilkent.edu.tr

$\neg X \rightarrow Y$ or $\neg X \rightarrow \neg Y$. Among these rules we are only interested in finding rules of the form $X \rightarrow \neg Y$ due to its relation to mutual exclusive sets as described in the next section. The support of this negative association rule can be calculated from the positive sets: $supp(X \rightarrow \neg Y) = supp(X) - supp(X \cup Y)$.

2.2 Mutual Exclusivity and Negative Associations

Consider a mutual exclusive mutation set with three genes m_i, m_j and m_k. If a tumor harbors one of these genes mutations, it does not contain the other two mutations. Such a mutual exclusivity pattern can be represented with the following three negative association rules:

$$\{m_i\} \rightarrow \{\neg m_j, \neg m_k\}$$
$$\{m_j\} \rightarrow \{\neg m_i, \neg m_k\}$$
$$\{m_k\} \rightarrow \{\neg m_i, \neg m_j\}$$

The first rule states that if m_i is present in a patient, neither m_j nor m_k is likely to be present in the patient. If dataset includes these three negative rules with high support and confidence, then we can conclude that there is a mutual exclusivity pattern among the three genes. We term such negative rules as complimentary rules. In general, if there are k complimentary rules for k different mutated genes, the k genes form a mutually exclusive gene set.

We also introduce the following more complex mutual exclusivity pattern that has not been studied in the literature: if a specific mutation is present in a patient tumor, certain mutations can still occur but not simultaneously. For example, a patient with mutation m_i and m_j but not with m_k fits this pattern, but a patient with all three mutations does not. For a three gene set, this mutual exclusivity pattern corresponds to a single negative association rule of the form: $\{m_i\} \rightarrow \neg\{m_j, m_k\}$.

2.3 MEMNAR Algorithm

Mining for negative association rules is a challenging task. The absence of mutations in a patient is frequently encountered and their combinations lead to an exponential number of negative associations rules, most of which are indeed uninteresting. There are a few methods developed for mining negative association rules [3], [16], [13], [18]. Our method is closest to PNAR [3] in the way it generates bigger items; however, PNAR does not discover the first type of rules discussed in the previous section. The steps of MEMNAR are as follows and it is illustrated in Figure 1:

1) Generate all positive frequent mutation sets, insert them in P. A frequent mutation set is considered frequent if its support is above a minimum support threshold. The subset of P with set size 1 will be referred as P_1.
2) Negate all mutation sets in P_1 and put the frequent negated items in N_1 which is set of frequent negative item sets with only one item.
3) Combine all pairs in P_1 and N_1 to generate mutation sets that contain one positive and one negative mutation. Insert the frequent ones in $PN_{1,1}$.
4) Join the generated frequent item sets in part 3 to get bigger item sets; filter them to get $PN_{1,2}$ and $PN_{2,1}$ which include frequent item sets with 1 positive and 2 negative mutations and frequent item sets with 2 positive and 1 negative mutations, respectively. This joining and filtering operation is defined in the next section.
5) Keep joining and filtering item sets inside each set to generate bigger item sets called $PN_{p,n}$ where p is the number of positive mutations in the item set and n is the number of negative mutations in it.
6) Generate all valid positive and negative association rules. A valid rule should have support, confidence and significance bigger than preset thresholds.
7) Generate mutual exclusive gene sets by combining complimentary valid rules.

We generate the positive frequent mutation sets in Step 1 using Frequent-Pattern Tree algorithm due to its efficiency [6]. In Steps 3 and 4, the downward-closure (*anti-monotonicity*) property of support is used, namely, if a set is frequent all of its subsets should be frequent. Then after generating 2-length mutation set, we form bigger item sets by joining smaller item sets, as discussed in the next section.

2.3.1 Positive and Negative Join

In Steps 4 and 5 of the algorithm we merge smaller mutation sets to get larger frequent mutation sets, we adapt the methodology introduced in Cornelis et al. [3]. There are two kinds of joining operations, *positive join* increases the positive subset of the mutation set, we call this the positive part, and *negative join* that increases the size of the negative part. For the positive join operation, the following criteria should be satisfied: i) the length of both mutation sets should be the same, ii) the negative parts should be identical, iii) their positive parts should differ in only one item and iv) the union set of the positive parts should also be frequent. Negative join has a similar criteria: i) for two sets to be merged with the negative join operation, the length of the mutation sets should be the same, ii) they should differ in only one item in their negative part, iii) their sets in the positive parts should be identical. In negative join, in addition to these conditions, we also check if the mutations that exists in one set but not in the other set could form a mutual exclusivity pattern as well by checking if they exist in $PN_{1,1}$. This additional criterion ensures that all the mutations will eventually form a mutual exclusive set. For example suppose that we want to negative join $\{m_1, \neg m_2\}$ with $\{m_1, \neg m_3\}$, their positive parts are identical and they only differ on one item in their negative part. Thus, we check if $\{m_2, \neg m_3\}$ and $\{\neg m_2, m_3\}$ are in $PN_{1,1}$ and if they are frequent, we join these two item sets to get $\{m_1, \neg m_2, \neg m_3\}$.

2.3.2 Filtering Mutation Sets

In step 5 of the algorithm we filter generated mutation similar to Apriori algorithm and PNAR algorithm based on support. However, to be able to find rare mutations, we set the support threshold very low and introduce additional filtering strategies. Firstly, we require the rule generated from mutation set to satisfy a significance lower bound. Secondly, if the confidence of the rule corresponding to the mutation set is not above a preset confidence threshold this set will not contribute in negative join function because confidence is monotonic in negative join. Additionally, we filter the mutation sets, which cannot lead to a mutually

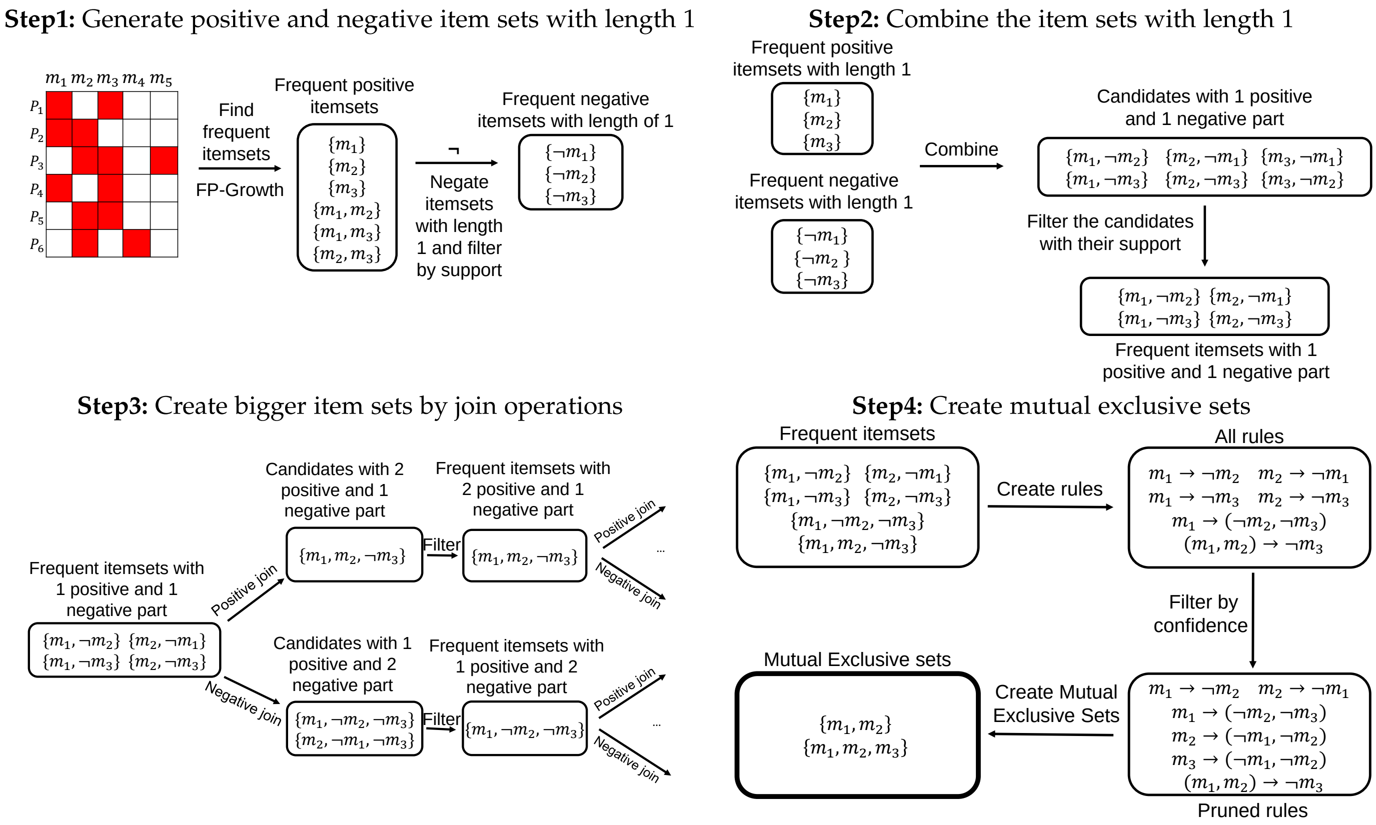
Step1: Generate positive and negative item sets with length 1
$m_1\ m_2\ m_3\ m_4\ m_5$
P_1
P_2
P_3
P_4
P_5
P_6
Find frequent itemsets
FP-Growth
Frequent positive itemsets
$\{m_1\}$ $\{m_2\}$ $\{m_3\}$ $\{m_1, m_2\}$ $\{m_1, m_3\}$ $\{m_2, m_3\}$
¬
Negate itemsets with length 1 and filter by support
Frequent negative itemsets with length of 1
$\{\neg m_1\}$ $\{\neg m_2\}$ $\{\neg m_3\}$
Step2: Combine the item sets with length 1
Frequent positive itemsets with length 1
$\{m_1\}$ $\{m_2\}$ $\{m_3\}$
Frequent negative itemsets with length 1
$\{\neg m_1\}$ $\{\neg m_2\}$ $\{\neg m_3\}$
Combine
Candidates with 1 positive and 1 negative part
$\{m_1, \neg m_2\}$ $\{m_2, \neg m_1\}$ $\{m_3, \neg m_1\}$ $\{m_1, \neg m_3\}$ $\{m_2, \neg m_3\}$ $\{m_3, \neg m_2\}$
Filter the candidates with their support
$\{m_1, \neg m_2\}$ $\{m_2, \neg m_1\}$ $\{m_1, \neg m_3\}$ $\{m_2, \neg m_3\}$
Frequent itemsets with 1 positive and 1 negative part
Step3: Create bigger item sets by join operations
Frequent itemsets with 1 positive and 1 negative part
$\{m_1, \neg m_2\}$ $\{m_2, \neg m_1\}$ $\{m_1, \neg m_3\}$ $\{m_2, \neg m_3\}$
Positive join
Candidates with 2 positive and 1 negative part
$\{m_1, m_2, \neg m_3\}$
Filter
Frequent itemsets with 2 positive and 1 negative part
$\{m_1, m_2, \neg m_3\}$
Positive join
...
Negative join
Negative join
Candidates with 1 positive and 2 negative part
$\{m_1, \neg m_2, \neg m_3\}$ $\{m_2, \neg m_1, \neg m_3\}$
Filter
Frequent itemsets with 1 positive and 2 negative part
$\{m_1, \neg m_2, \neg m_3\}$
Positive join
...
Negative join
Step4: Create mutual exclusive sets
Frequent itemsets
$\{m_1, \neg m_2\}$ $\{m_2, \neg m_1\}$ $\{m_1, \neg m_3\}$ $\{m_2, \neg m_3\}$ $\{m_1, \neg m_2, \neg m_3\}$ $\{m_1, m_2, \neg m_3\}$
Create rules
All rules
$m_1 \rightarrow \neg m_2$ $m_2 \rightarrow \neg m_1$ $m_1 \rightarrow \neg m_3$ $m_2 \rightarrow \neg m_3$ $m_1 \rightarrow (\neg m_2, \neg m_3)$ $(m_1, m_2) \rightarrow \neg m_3$
Filter by confidence
Pruned rules
$m_1 \rightarrow \neg m_2$ $m_2 \rightarrow \neg m_1$ $m_1 \rightarrow (\neg m_2, \neg m_3)$ $m_2 \rightarrow (\neg m_1, \neg m_2)$ $m_3 \rightarrow (\neg m_1, \neg m_2)$ $(m_1, m_2) \rightarrow \neg m_3$
Create Mutual Exclusive Sets
Mutual Exclusive sets
$\{m_1, m_2\}$ $\{m_1, m_2, m_3\}$

Fig. 1: Overview of MEMNAR.

exclusive set. For this we check if the mutation set leads to complimentary rules as defined in Section 2.2. For example, for item set $\{m_1, \neg m_2, \neg m_3\}$, we require if $\{m_2, \neg m_1, \neg m_3\}$ and $\{m_3, \neg m_1, \neg m_2\}$ to be frequent, and if they are not, they are pruned.

2.3.3 Generating rules

The negative association rules of the form, $X \rightarrow \neg Y$, are created such that the positive set constitutes the antecedent and the negative set constitutes the consequent of the rule. For example, from mutation set $\{m_1, m_2, \neg m_3, \neg m_4\}$, the following rule will be generated: $\{m_1, m_2\} \rightarrow \{\neg m_3, \neg m_4\}$.

2.3.4 Assessing statistical significance

We assess the significance of a rule by calculating the z-score. For this we calculate the χ^2 statistic as described in [5] and calculate the z-score based on χ^2 statistic:

$$z(X \rightarrow \neg Y) = \frac{\sqrt{n}\delta(X \rightarrow \neg Y)}{\sqrt{2supp(X)supp(\neg X)supp(Y)supp(\neg Y)}}, \quad (1)$$

where δ is the *leverage* for a rule and it is defined as:

$$\delta(X \rightarrow \neg Y) = supp(X \cup \neg Y) - supp(X)supp(\neg Y) \quad (2)$$

2.3.5 Generating mutual exclusive sets

To find the first type of mutual exclusivity sets as discussed in Section 2.2, we look for k valid complementary negative association rules among k genes. If they all exist, we form a mutual exclusive set. The z-score and confidence for the set are calculated by averaging the z-scores and confidences of the rules forming it. For the second type of mutual exclusivity pattern, the negative association rules with high z-scores are found that fits the pattern. The rule's z-score is considered as the mutual exclusive set's score.

3 Results

3.1 Results on simulated data

We repeated the simulated experiments in MEGSA [7] and compared MEMNAR with three algorithms: MEGSA [7], Mutex [1], and Multi-Dendrix [9]. A mutation matrix of 54 genes and 500 patients is generated. A mutual exclusive set of 4 genes is implanted in the patient mutation data; the coverage of these mutual exclusive sets is varied in assessing the performance. The background mutation rate for simulated genes is set to 1%. Two settings is used: i) in the balanced setting mutual exclusive mutations are equally distributed among patients, ii) in the unbalanced setting, one mutation covers more patient than the remaining three with the proportion 3:1:1:1. The mutations in the remaining 50 genes were randomly distributed. These genes are divided into 5 groups with frequencies 1%, 5%, 10%, 20% and 30%. Each algorithm is evaluated based on whether it can find the simulated mutual exclusive gene set as the top ranked mutual exclusive set. The simulation is repeated 100 times in each case.

Figures 2 A and B display simulation results for balanced and unbalanced cases respectively. MEMNAR outperforms all the methods in both settings. We also compared runtimes (Figure 2 C). Here Multi-Dendrix algorithm was too slow to complete the analysis, so it is not shown. MEMNAR is much faster than both MEGSA and Mutex.

3.2 Mutual Exclusive Gene Sets in Breast Cancer

We applied MEMNAR on breast cancer (BRCA) somatic mutation data obtained from TCGA and post-processed by [9]. MEMNAR found 21 significant (p-value < 0.005)

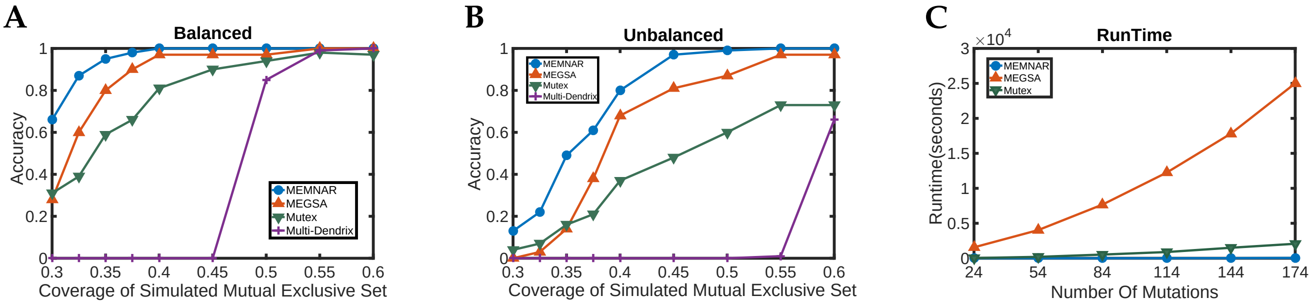

Fig. 2: Comparison of algorithms on simulated data for A) the balanced and B) imblanced cases. C) Runtimes of the algorithms as the number of total mutations are varied.

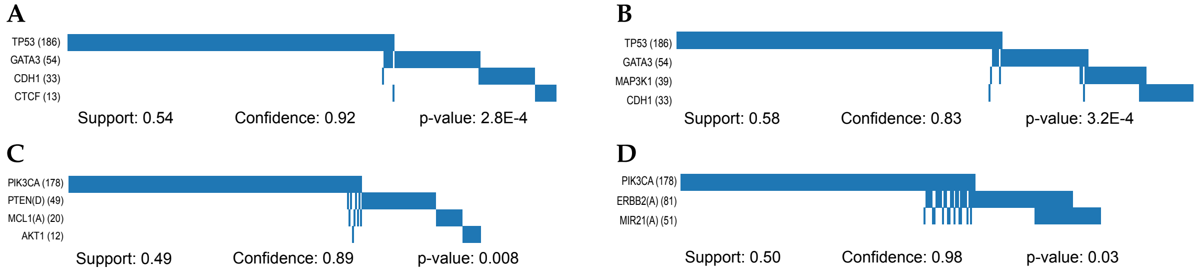

Fig. 3: Example mutual exclusive sets discovered in BRCA. Rows are genes, columns are samples, blue color indicates the gene is mutated in the patient. The numbers next to genes report the number of patients the gene is mutated. D or A next to gene name indicates that it is a copy number variation with deletion or amplification type.

mutually exclusive mutation sets, which includes all the significant mutual exclusive sets found previously by other methods. The top set includes TP53, GATA3, CDH1, and CTCF (Figure 3 A), which is altered in 54.83% of the BRCA samples. All of these genes have been reported as driver mutations for breast cancer [17], [12], [4]. The set with TP53, CDH1, GATA3 and MAP3K1/MAP2K4 (Figure 3 B), belongs to MAPK/ERK pathway, which is known to be driver pathway[17]. The set that comprise of PIK3CA, PTEN(D), MCL1(A), AKT1 (Figure 3 C) was not found before by any *de novo* methods and is interesting as the genes participate in to Jak-STAT signaling pathway. As an example of complex mutual exclusive rule, that states that when PIK3CA is mutated in a patient, ERBB2(A) and MIR21(A) show a mutual exclusive relation (Figure 3 D).

References

[1] Ö. Babur, M. Gönen, B. A. Aksoy, N. Schultz, G. Ciriello, C. Sander, and E. Demir. Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations. *Genome Biology*, 16(1):1–10, 2015.

[2] G. Ciriello, E. Cerami, C. Sander, and N. Schultz. Mutual exclusivity analysis identifies oncogenic network modules. *Genome research*, 22(2):398–406, 2012.

[3] C. Cornelis, P. Yan, X. Zhang, and G. Chen. Mining positive and negative association rules from large databases. In *Cybernetics and Intelligent Systems, 2006 IEEE Conference on*, pages 1–6. IEEE, 2006.

[4] F. Graziano, B. Humar, and P. Guilford. The role of the e-cadherin gene (cdh1) in diffuse gastric cancer susceptibility: from the laboratory to clinical practice. *Annals of oncology*, 14(12):1705–1713, 2003.

[5] W. Hämäläinen. Kingfisher: an efficient algorithm for searching for both positive and negative dependency rules with statistical significance measures. *Knowledge and information systems*, 32(2):383–414, 2012.

[6] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. In *ACM Sigmod Record*, volume 29, pages 1–12. ACM, 2000.

[7] X. Hua, P. L. Hyland, J. Huang, L. Song, B. Zhu, N. E. Caporaso, M. T. Landi, N. Chatterjee, and J. Shi. Megsa: A powerful and flexible framework for analyzing mutual exclusivity of tumor mutations. *The American Journal of Human Genetics*, 98(3):442–455, 2016/03/12.

[8] M. S. Lawrence, P. Stojanov, P. Polak, G. V. Kryukov, K. Cibulskis, A. Sivachenko, S. L. Carter, C. Stewart, C. H. Mermel, S. A. Roberts, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–218, 2013.

[9] M. D. Leiserson, D. Blokh, R. Sharan, and B. J. Raphael. Simultaneous identification of multiple driver pathways in cancer. *PLoS Comput Biol*, 9, 2013.

[10] M. D. Leiserson, H.-T. Wu, F. Vandin, and B. J. Raphael. Comet: a statistical approach to identify combinations of mutually exclusive alterations in cancer. *Genome biology*, 16(1):1–20, 2015.

[11] C. A. Miller, S. H. Settle, E. P. Sulman, K. D. Aldape, and A. Milosavljevic. Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors. *BMC medical genomics*, 4(1):34, 2011.

[12] E. J. Peterson, O. Bögler, and S. M. Taylor. p53-mediated repression of dna methyltransferase 1 expression by specific dna binding. *Cancer research*, 63(20):6579–6582, 2003.

[13] A. Savasere, E. Omiecinski, and S. Navathe. Mining for strong negative associations in a large database of customer transactions. In *Data Engineering, 1998. Proceedings., 14th International Conference on*, pages 494–502. IEEE, 1998.

[14] P.-N. Tan, S. Michael, and V. Kumar. *Association Analysis: Basic Concepts and Algorithms*, chapter 6, pages 327–404. Addison-Wesley, 2005.

[15] F. Vandin, E. Upfal, and B. J. Raphael. Algorithms for detecting significantly mutated pathways in cancer. *Journal of Computational Biology*, 18(3):507–522, 2011.

[16] X. Wu, C. Zhang, and S. Zhang. Efficient mining of both positive and negative association rules. *ACM Transactions on Information Systems (TOIS)*, 22(3):381–405, 2004.

[17] W. Yan, Q. J. Cao, R. B. Arenas, B. Bentley, and R. Shao. Gata3 inhibits breast cancer metastasis through the reversal of epithelial-mesenchymal transition. *Journal of Biological Chemistry*, 285(18):14042–14051, 2010.

[18] X. Yuan, B. P. Buckles, Z. Yuan, and J. Zhang. Mining negative association rules. In *Computers and Communications, 2002. Proceedings. ISCC 2002. Seventh International Symposium on*, pages 623–628. IEEE, 2002.