JMLR: Workshop and Conference Proceedings 47:29–36, 2014 Statistically Sound Data Mining

Statistically significant subgraphs for genome-wide association study

Jun Sese SESE.JUN@AIST.GO.JP
Computational Biology Research Center, AIST
2-4-7 Aomi, Koto, Tokyo, 135-0064, Japan
Aika Terada TERADA.AIKA@AIST.GO.JP
Research Fellow of Japan Society for the Promotion of Science
and *Computational Biology Research Center, AIST*
2-4-7 Aomi, Koto, Tokyo, 135-0064, Japan
Yuki Saito SAITO@SS.CS.TITECH.AC.JP
Dept. of Computer Science, Tokyo Institute of Technology
2-12-1 Ookayama, Meguro, Tokyo, 152-8550, Japan
Koji Tsuda TSUDA@K.U-TOKYO.AC.JP
Dept. of Computational Biology, The University of Tokyo
5-1-5 Kashiwanoha, Kashiwa, Chiba, 277-8568, Japan

Editors: Wilhelmiina Hämäläinen, Francois Petitjean, and Geoff Webb

Abstract

Genome-wide association studies (GWAS) have been widely used for understanding the associations of single-nucleotide polymorphisms (SNPs) with a disease. GWAS data are often combined with known biological networks, and they have been analyzed using graph-mining techniques toward a systems understanding of the biological changes caused by the SNPs. To determine which subgraphs are associated with the disease, a statistical test on each subgraph needs to be conducted. However, no statistically significant results were found because multiple testing correction causes an extremely small corrected significance level.

We introduce a method called gLAMP to enumerate subgraphs having statistically significant associations with a diagnosis. gLAMP integrates the Limitless Arity Multiple-testing Procedure (LAMP) with a graph-mining algorithm called COmmon Itemset Network mining (COIN). LAMP gives us the smallest possible Bonferroni factor, and COIN provides us with efficient enumeration of testable subgraphs. Theoretical results of their combination show the potential to enumerate subgraphs statistically significantly associated with a disease.

Keywords: statistical significance, subgraph enumeration, chi-squared test, GWAS

1. Introduction

Genome-wide association study (GWAS) is a powerful analysis method of associating single-nucleotide polymorphism with a trait and has been widely used to understand both biology and disease analysis (Civelek and Lusis, 2014). While causal mutations of diseases have been uncovered using GWAS, two problems remain. One is that the most of GWAS analysis focused on the associations between single SNP and a disease while diseases are regularly

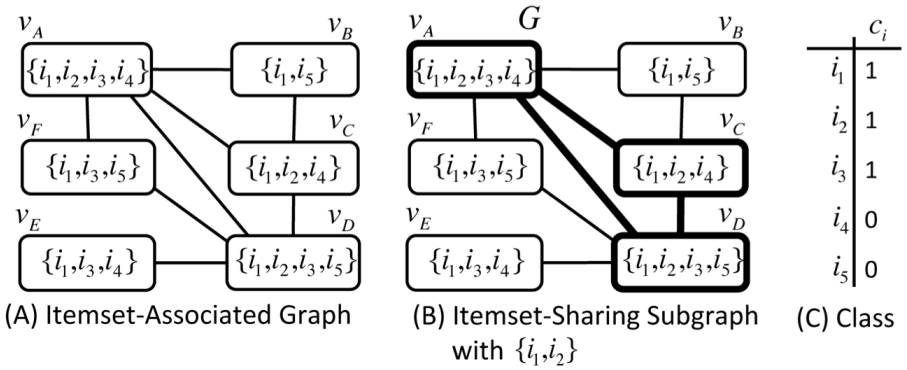

Figure 1: (A) An example of IA-graph. In GWAS data, v_n and i_m indicate a SNP at position n and a patient having the SNP. Edges mean the relationships between SNPs. (B) ISS with common itemset $\{i_1, i_2\}$. Subgraph G (bold lines) indicate the ISS having three vertices, three edges. (C) Class of each patient. $c_m = 1$ means that patient i_m has a target trait/disease. The contingency table associated with G is shown in Table 2.

associated with multiple SNPs (Sladek et al., 2007). The requirements of large computational time and statistical assessment of the results prohibited the genome-wide analysis of the combinations of SNPs. The other is the difficulty to find the connection of the mutations with the understanding of why the SNPs cause these diseases, which is required to formulate new drugs and to develop new therapeutic methods. To connect the SNPs with the systems understandings, known biological networks such as protein-protein interactions (Chatr-aryamontri et al., 2013) and metabolic pathways (Kanehisa et al., 2014) are often integrated with GWAS data. Network analyses on the data have been widely performed (Barabási et al., 2011). However, only a few analysis results have been confirmed biologically because of the lack of statistical assessment of the results. In biology and the medical science, the statistical significance of the results of an analysis is an important criterion of whether they are confirmed experimentally. Computational results without statistical assessments cannot be confirmed and thus will never be published in any biological or medical journals.

Statistical assessment of the graph-mining result may lead to no significant results because of multiple testing correction. Most graph-mining algorithms check the importance on every subgraph, which causes an enormous amount of tests and requires multiple testing correction. When we use Bonferroni correction on the situation, the corrected significance level would be extremely small, and no significant result might be found. This is one reason why few studies in graph mining verified the statistical significance.

Statistically sound association discovery methods(Webb, 2003; Hamalainen, 2010; Terada et al., 2013; Webb and Vreeken, 2014) might provide us the statistical significance to the GWAS results. However, the calculation of the statistical significance may take large computational time because of the large size of GWAS data, and even if the calculation

Table 1: Contingency Table of G_m.

	$c_i = 1$	$c_i = 0$	
G_m	$x(G_m)$	$n_m - x(G_m)$	$n_m = \lvert I(G_m)\rvert$
$\bar{G_m}$	$C - x(G_m)$	$N - n_m - C + x(G_m)$	$N - n_m$
	C	$N - C$	N

Table 2: Contingency Table of the ISS in Fig. 1(B)

	$c_i = 1$	$c_i = 0$	
G	2	0	2
$\bar{G}$	1	2	3
	3	2	5

is finished, the corrected significance level might become too small to discover statistically significant associations. Furthermore, no existing methods have considered statistical significance of graph structures.

In this paper, we formalize a statistical graph-mining problem for the GWAS using graph data, and introduce a method to solve the problem. Our solution uses the advantage of Limitless Arity Multiple-testing Procedure (LAMP) (Terada et al., 2013) to calibrate the Bonferroni factor to the smallest possible value, and tries to efficiently find a statistically significant result even after multiple testing correction is performed.

We introduce a graph whose vertex has an itemset label proposed by Sese *et al.* (Sese et al., 2010)

Definition:
(Itemset-associated graph and itemset-sharing subgraph) An itemset-associated (IA) graph is an undirected graph whose vertex contains a set of items (an itemset). An itemset-sharing subgraph (ISS) with an itemset I means a connected subgraph of a given IA graph whose all vertices contain I. For an ISS G, we describe $V(G)$, $E(G)$ and $I(G)$ are the vertices, edges and common (largest) itemset in G.

Figure 1(A) shows an example of the IA-graph with six vertices, eight edges and five items. Vertex v_B has an itemset $\{i_1, i_5\}$. A subgraph indicated by bold lines in Figure 1(B) is an ISS with vertices v_A, v_C and v_D and common itemset $\{i_1, i_2\}$.

In the GWAS analysis, a vertex, an edge and an item represent a SNP, a connection between SNPs and a patient sample, respectively. An ISS in Figure 1(B) shows that patients i_1 and i_2 (common items of the graph) have SNPs v_A, v_C and v_D (vertices in the graph).

Definition:
(P-value of an ISS) Suppose that item i is related to a class $c_i \in \{0, 1\}$. With ISS G, items are divided into two groups. One is in $I(G)$, and the other is not. The status is described as a contingency table in Table 1, where $x(G) = |\{i \mid i \in I(G) \text{ and } c_i = 1\}|$ and $C = |\{i \mid c_i = 1\}|$. On the contingency table, we can calculate a P-value of G using chi-squared test and define it as $P(G)$.

We can perform a chi-squared test, Fisher's exact test, etc. on the contingency table, but the chi-squared test is widely used in GWAS analysis, and hence we used a chi-squared test here.

A trait of a patient in the GWAS analysis is regarded as the class associated with each item. An ISS G described in Figure 1(B) have common itemset $\{i_1, i_2\}$, and both of which have class 1. Table 2 shows a contingency table for G. Its chi-squared value and P-value are 2.22 and 0.137, respectively.

With these definitions, we introduce a statistical graph mining problem.

Problem:
(gLAMP problem) Suppose that we have an IA graph and a class table. Given the data and significance level α, enumerate statistically significant ISSes $\mathcal{G}$ in the IA graph where $P(G) \leq \delta$ for $G \in \mathcal{G}$, and δ is a corrected significance level to control family-wise error rate (FWER), the probability of at least one false discoveries, below α. ■

The results of the problem are related to the combinations of SNPs having statistically significant associations with the target trait.

2. Limitless Arity Multiple-testing Procedure (LAMP)

Bonferroni correction has been used in almost all GWAS analyses to control FWER below the significance level α. However, Bonferroni correction is too conservative to control the FWER in practice because it assumes that any tests can cause false positive. To avoid the problem, we here introduce LAMP (Terada et al., 2013).

Given IA-graph contains M subgraphs $G_1, G_2, \ldots, G_M$, a statistical test is performed for M subgraphs to find statistically significantly associated subgraphs with a trait. Fixing corrected significance level δ, Bonferroni correction calculates FWER as $M\delta$ from the following inequality.

$$\begin{aligned} FWER &= 1 - \Pr(\{m \mid P(G_m) > \delta \text{ for } m \in \{1, \ldots, M\}\} \neq \phi) \\ &= \Pr(\{m \mid P(G_m) \leq \delta \text{ for } m \in \{1, \ldots, M\}\} = \phi) \\ &\leq \sum_{i=1}^{M} \Pr(P(G_m) \leq \delta) \leq M\delta \end{aligned}$$

This value should be less than significance level α, and hence δ is set to α/M in Bonferroni correction. Generally, M is substantial number in the graph-mining problem, causing an extremely small corrected significance level. It may become impossible to find statistically significant results.

LAMP (Terada et al., 2013) achieves higher sensitivity than Bonferroni correction by rigorously calculating FWER and can enumerate statistically significant tests from multidimensional data. LAMP categorizes tests into testable and untestable since untestable ones are safely removed from the Bonferroni factor. Untestable ones are defined as the tests that never cause significant results under corrected significance level δ. When we have subgraph G_m, its marginal distribution (N, C and n_m in Table 1) of the contingency table can be calculated without performing the statistical test. From contingency tables satisfying the distribution, P-values can be calculated. The smallest P-value among them is the possible minimum P-value of G_m. If the value is larger than δ, G_m is untestable and can be removed from Bonferroni factor because $\Pr(P(G_m) \leq \delta) = 0$. The minimum P-value can be calculated on the discrete statistics such as Fisher's exact test, chi-squared test and Mann-Whitney test.

In the GWAS problem, the minimum P-value depends only on n_m because N and C are fixed (details are in Terada et al. (2013)). Therefore, we use $f(n_m)$ to describe the minimum P-value of G_m, and testable G_m satisfy $f(n_m) \leq \delta$ while untestable ones satisfy

Data: associations between SNPs and patients, traits of patients, and significance level α
Result: the set of itemsets whose P-value $\leq \delta$ in $\mathcal{I}$
$n \leftarrow$ the number of patients whose classes are 1
$\delta \leftarrow 1.0$
while $n > 0$ **do**
  $\mathcal{I} \leftarrow$ itemsets (combinations of SNPs) that relate n or more patients.
  (run the FIM algorithm)
  $m_n \leftarrow |\mathcal{I}|$
  $\delta \leftarrow \alpha / m_n$
  if $\delta < f(n-1)$ **then**
    break
  end
end
Algorithm 1: LAMP

$f(n_m) > \delta$. $f(n_m)$ is calculated when the values in the contingency table are the most biased, and the minimum P-value is achieved at $x(G_m) = \min\{n_m, C\}$. Note that $f(x)$ monotonically increases with decreasing x(Terada et al., 2013).

We need to solve an optimization problem to find the largest δ so that FWER keeps below α. With the property,

$$\begin{aligned} FWER &= \Pr(\{m \mid P(G_m) \leq \delta \text{ for } m \in \{1, \ldots, M\}\} = \phi) \\ &\leq \sum_{m=1}^{M} \Pr(P(G_m) \leq \delta) \leq \sum_{m \in \{m | f(n_m) \leq \delta\}} \Pr(P(G_m) \leq \delta) \\ &\leq |\{m \mid f(n_m) \leq \delta \text{ for } m \in \{1, \ldots, M\}\}| \delta = M' \delta, \end{aligned}$$

where $M' = |\{m | f(n) \leq \delta\}|$ and n is the largest value that satisfy $f(n) \leq \delta$. In other words, M' is the number of testable tests. Hence, we can set δ to α / M' unless $M' \delta \leq \alpha$. Because δ depends on n, LAMP determines the largest n to set FWER bound δm_n below α. Calculating m_n from high-dimensional data can be performed using a frequent pattern mining (FIM) algorithm (Uno et al., 2003).

The pseudo-code of LAMP procedure is described in Algorithm 1. LAMP uses the property that $f(n)$ monotonically increases with decreasing n when $n \leq C$. n is initially set to the possible largest value, and subsequently decreases until $\delta > f(n-1)$. In the next section, we use the property to address the graph mining setting.

3. Enumerating testable itemset-associated subgraphs

We here introduce the testable subgraphs that are associated with the maximal itemset-sharing subgraphs and show that LAMP can address subgraphs using the replacement of the FIM algorithm with a graph-mining algorithm.

We here show that the number of maximal ISSes is used as the Bonferroni factor.

Definition:
(Maximal ISS) For ISSes G, when no ISS G' whose $V(G) \subseteq V(G')$, $E(G) \subseteq E(G')$ and $I(G) \subseteq I(G')$ exists, G is defined as the maximal ISS.

Data: IA graph G, class C, significance level α
Result: the set of itemsets whose P-value $\leq \delta$ in $\mathcal{G}$
$n \leftarrow |\{i|c_i = 1\}|$
$\delta \leftarrow 1.0$
while $n > 0$ **do**
  $\mathcal{G}_n \leftarrow$ run COIN to find maximal ISSes that relate n or more items in G
  $m_n \leftarrow |\mathcal{G}_n|$
  $\delta \leftarrow \alpha / m_n$
  if $\delta < f(n-1)$ **then**
    break
  end
  $n \leftarrow n - 1$
end

Algorithm 2: gLAMP

Property:
Only maximal ISSes should be counted in Bonferroni factor. ■

Proof Suppose that G is not a maximal ISS. In this case, a maximal ISS G' whose $I(G') = I(G)$ exists from the definition. When $I(G) = I(G')$, the contingency table of G is identical to the contingency table of G', and hence we can safely remove G from Bonferroni factor. Therefore, we need to count only maximal ISSes in Bonferroni factor. ■

The following property guarantees that we use the ISS enumeration technique instead of FIM algorithm in LAMP.

Property:
(Adding a vertex decreases the size of common itemset) Let G be an ISS. Let G' be an ISS generated by adding node $v \notin V(G)$. $I(G') \subseteq I(G)$ for any v.

By adding node v to G, For a maximal graph G' having vertices $V \cup \{v\}$ where $v \notin V$, $I(G') \subset I(G)$.

From the property, we can conclude the following property. The property shows that the Bonferroni factor decreases according to the increase of n, and hence the minimum P-value associated with the subgraphs increases.

Property:
Let $\mathcal{G}_n$ be a set of maximal ISSes that relate n or more items. Between $\mathcal{G}_n$ and $\mathcal{G}_{n+1}$, $\mathcal{G}_n \supseteq \mathcal{G}_{n+1}$ holds. Hence, $|\mathcal{G}_n| \geq |\mathcal{G}_{n+1}|$

These properties allow us to replace the FIM algorithm with the graph-mining algorithm to find maximal ISSes called COmmon Itemset Network mining (COIN) (Sese et al., 2010) in LAMP to enumerate statistically significant subgraphs (Algorithm 2). The difference between LAMP in Algorithm 1 and gLAMP in Algorithm 2 is only at line 3, in which the FIM algorithm is replaced with COIN.

4. Summary and Future Work

We introduced an algorithm to a multiple testing procedure algorithm for subgraphs in a large complex graph. The procedure uses the main framework of LAMP and replaces the FIM algorithm in LAMP with COIN.

Minato *et al.* (Minato et al., 2014) introduced an efficient algorithm for LAMP, which uses depth-first traversal instead of LAMP's breadth-first traversal. gLAMP inherits the LAMP's breadth-first traversal, and the dept-first traversal would be applicable to the proposed problem.

This paper only demonstrated the theoretical points of the statistically sound graph mining problem. We plan on implementing this procedure, and evaluating the efficiency and usefulness of this algorithm in the future.

References

Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68, January 2011.

Andrew Chatr-aryamontri, Bobby-Joe Breitkreutz, Sven Heinicke, Lorrie Boucher, Andrew Winter, Chris Stark, Julie Nixon, Lindsay Ramage, Nadine Kolas, Lara O'Donnell, Teresa Reguly, Ashton Breitkreutz, Adnane Sellam, Daici Chen, Christie Chang, Jennifer Rust, Michael Livstone, Rose Oughtred, Kara Dolinski, and Mike Tyers. The BioGRID interaction database: 2013 update. *Nucleic Acids Research*, 41(Database issue):D816–23, January 2013.

Mete Civelek and Aldons J Lusis. Systems genetics approaches to understand complex traits. *Nature Reviews Genetics*, 15(1):34–48, January 2014.

Wilhelmiina Hamalainen. Efficient Discovery of the Top-K Optimal Dependency Rules with Fisher's Exact Test of Significance. In *IEEE 10th International Conference on Data Mining (ICDM 2010)*, pages 196–205. IEEE, 2010.

Minoru Kanehisa, Susumu Goto, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Research*, 42(Database issue):D199–205, January 2014.

Shin-ichi Minato, Takeaki Uno, Koji Tsuda, Aika Terada, and Jun Sese. Fast statistical assessment for combinatorial hypotheses based on frequent itemset mining. In *Proc. of ECML/PKDD 2014*, 2014.

Jun Sese, Mio Seki, and Mutsumi Fukuzaki. Mining Networks with Shared Items. In *Proc. of the 19th ACM international conference on Information and knowledge management*, pages 1681–1684, New York, New York, USA, 2010. ACM Press.

Robert Sladek, Ghislain Rocheleau, Johan Rung, Christian Dina, Lishuang Shen, David Serre, Philippe Boutin, Daniel Vincent, Alexandre Belisle, Samy Hadjadj, Beverley

Balkau, Barbara Heude, Guillaume Charpentier, Thomas J Hudson, Alexandre Montpetit, Alexey V Pshezhetsky, Marc Prentki, Barry I Posner, David J Balding, David Meyre, Constantin Polychronakos, and Philippe Froguel. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 445(7130):881–885, February 2007.

Aika Terada, Mariko Okada-Hatakeyama, Koji Tsuda, and Jun Sese. Statistical significance of combinatorial regulations. *PNAS*, 110(32):12996–13001, August 2013.

Takeaki Uno, Tatsuya Asai, Yuzo Uchida, and Hiroki Arimura. LCM : An efficient algorithm for enumerating frequent closed item sets. In *Proceedings of Workshop on Frequent itemset Mining Implementations (FIMI'03)*, 2003.

G.I. Webb. Preliminary investigations into statistically valid exploratory rule discovery. In S.J. Simoff, G.J. Williams, and M. Hegland, editors, *Proceedings of the Second Australasian Data Mining Conference (AusDM03)*, pages 1–9, Sydney, 2003. University of Technology.

G.I. Webb and J. Vreeken. Efficient discovery of the most interesting associations. *Transactions on Knowledge Discovery from Data*, 8(3):15:1–15:31, 2014.