Statistically significant subgraphs for genome-wide association study

Jun Sese[1], Aika Terada[1,2], Yuki Saito[3], and Koji Tsuda[4]

[1] Dept. of Computer Science, Ochanomizu University, Tokyo, 112-8610, Japan,
{sesejun@is.ocha.ac.jp, terada.aika@ocha.ac.jp}, http://seselab.org/
[2] Japan Society for the Promoting Science, Japan
[3] Dept. of Computer Science, Tokyo Institute of Technology, 152-8550, Japan
[4] Dept. of Computational Biology, University of Tokyo, Japan

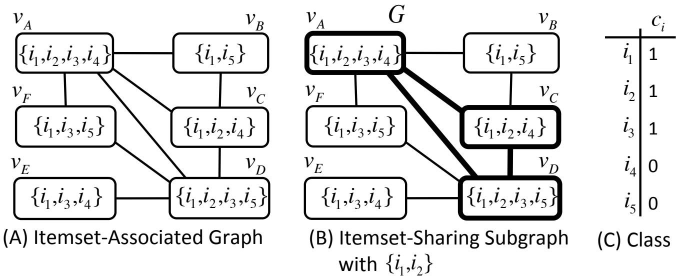
Abstract. Genome-wide association studies (GWAS) are widely applied for understanding the associations of single-nucleotide polymorphisms (SNPs) with a trait. GWAS data are often combined with known biological networks, and they have been analyzed using graph-mining techniques toward a systems understanding of the biological changes caused by the SNPs. To determine which subgraphs are associated with the trait, 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 trait. gLAMP integrates the limitless arity multiple-testing procedure (LAMP) with a graph-mining algorithm called COmmon Itemset Network mining (COIN). gLAMP controls the Bonferroni factor to the smallest possible value by showing that a larger subgraph tends to become untestable, which can be removed theoretically from the Bonferroni factor. The theoretical result shows that this combination has the potential to enumerate subgraphs statistically significantly associated with a trait.

Keywords: statistical significance, large graph, chi-squared test, GWAS

1 Introduction

Genome-wide association studies (GWAS) are a powerful analysis method of associating single-nucleotide polymorphism with a trait and has been widely used to understand both biology and disease analysis [2]. While many causal mutations of diseases have been uncovered using GWAS, diseases are regularly associated with multiple SNPs [6], and a systems understanding of why the SNPs cause these diseases is required to formulate new drugs and to develop new therapeutic methods. To this end, known biological networks are often integrated with GWAS data, and network analyses on the data have been widely performed [1].

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



(A) Itemset-Associated Graph (B) Itemset-Sharing Subgraph with $\{i_1,i_2\}$ (C) Class

Fig. 1. An example of GWAS and network data. v_n, i_m and c_m are associated with a SNP at position n, a patient m and a trait of m. Vertex v_A in Fig. 1 (A) indicates that SNP v_A was found in patients i_1, i_2, i_3 and i_4.

Table 1. Contingency Table of G.

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

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.

Statistically sound association discovery methods [10, 3, 11, 8] might provide us the statistical significance to the GWAS results. However, no existing methods have considered statistical significance of graph structures.

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. Performing a statistical test on each subgraph would require an enormous amounts of tests, and multiple testing correction would be required. 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 verified the statistical significance in graph mining.

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 correction (LAMP) [8] 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.

To describe the problem, we introduce graph structure whose vertex has an itemset label proposed by Sese *et al.* [5]

Definition 1 *(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 intersection of itemsets associated with the vertices in the subgraph is I. For an ISS G, we describe $V(G)$, $E(G)$ and $I(G)$ are the vertices, edges and sharing itemset in G.*

In the GWAS analysis, a vertex, an edge and an item are associated with a SNP, a connection between SNPs and a patient sample, respectively. Using a trait associated with a patient, we perform a statistical assessment of the subgraph.

Definition 2 *(P-value of an ISS) Suppose that item i is associated with 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 can be described as a contingency table in Table 1, where $x(G) = |\{i \mid i \in I(G)$ and $c_i = 1\}|$ and $C = |\{i \mid c_i = 1\}|$. We can perform a chi-squared test, Fisher's exact test, etc. on the contingency table. In GWAS analysis, the chi-squared test is widely used; therefore, we use a chi-squared test here. We define $P(G)$ as the P-value of the chi-squared test of G.*

A trait of a patient in the GWAS analysis is regarded as the class associated with each item. Figure 1 shows an example of an IA graph and the ISS. Table 2 shows a contingency table for G in Figure 1(C) with trait values in Figure 1(C). 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 1. (gLAMP problem) Suppose that we have an IA graph and class information for each item. 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) 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, the 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 [8].

Bonferroni correction is derived from the following inequality.

$$FWER = 1 - P(\cap_{i=1}^{M}\{P_i > \delta\}) = P(\cup_{i=1}^{M}\{P_i \leq \delta\}) \leq \sum_{i=1}^{M} P(\{P_i \leq \delta\}) \leq M\delta,$$

where P_i is a P-value of test i, and M is the number of tests. The problem that arises when the Bonferroni correction is applied to the graph-mining problem is that the corrected significance level might become extremely small because of the substantial number of subgraphs, and it may become impossible to find statistically significant results.

Limitless-arity multiple testing procedure (LAMP) [8] can enumerate statistically significant tests from multidimensional data. LAMP categorizes tests into testable and untestable. Let $f(n)$ be the minimum P-value of a statistical measure of tests having n or more objects using a fixed marginal distribution. In the contingency table in Table 1, when the marginal distribution is fixed, only $x(G)$ is variable. $f(n)$ is calculated when the values in the contingency table are the most biased, and the minimum P-value is achieved at $x(G) = \min\{n, C\}$. For Fisher's exact test, $f(n)$ is calculated as

$$f(n) = \begin{cases} \binom{C}{n} \Big/ \binom{N}{n} & \text{for } n \leq C, \\ 1 \Big/ \binom{N}{C} & \text{otherwise.} \end{cases} \tag{1}$$

Let n_i be the number of objects that satisfies i. A testable one satisfies $f(n_i) \geq \delta$ while an untestable one satisfies $f(n_i) < \delta$, where δ is a corrected significance level.

Untestable ones can be safely removed from Bonferroni factor. Let m_n be the number of testable ones that satisfy n or more objects. Tarone [7] showed that the untestable ones never cause false positives. With the property,

$$\begin{aligned} FWER = P(\cup_{i=1}^{M}\{P_i < \delta\}) \leq \sum_{i=1}^{M} P(\{P_i < \delta\}) \leq \sum_{i \in \{i | f(n_i) < \delta\}} P(\{P_i < \delta\}) \\ \leq |\{i | f(n_i) < \delta\}| \delta = m_n \delta. \end{aligned}$$

Hence, we can set δ to α / m_n unless $m_n \delta > \alpha$. 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 [9].

The pseudo-code of LAMP procedure is described in Algorithm 1. LAMP uses the property that $f(n)$ monotonically increases with decreasing n. n is initially set to the possible largest value, and subsequently decreases until $\delta > f(\lambda - 1)$.

An important point in the LAMP procedure is that the decrease in n increases m_n, which decreases FWER bound monotonically. In other words, any data structure can be used if it satisfies this property. 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 we need to count only maximal ISSes in Bonferroni factor.

Algorithm 1 LAMP (dataset $\mathcal{D}$, significance level α)

```
$\lambda \leftarrow$ the number of objects whose classes are 1, $\delta \leftarrow 1.0$
while $\lambda > 0$ do
    $\mathcal{I} \leftarrow$ itemsets that relate $\lambda$ or more objects in $\mathcal{D}$. (run the FIM algorithm)
    $m_\lambda \leftarrow |\mathcal{I}|$
    $\delta \leftarrow \alpha/m_\lambda$
    if $\delta < f(\lambda - 1)$ then
        break
    end if
    $\lambda \leftarrow \lambda - 1$
end while
Return the set of itemsets whose P-value $\leq \delta$ in $\mathcal{I}$
```

Definition 3 *(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.*

Property 1 *Only maximal ISSes should be counted in Bonferroni factor* ■

Proof. To proof the property, it is enough to show that non-maximal ISS is depend on a maximal ISS.

Suppose that G is not a maximal ISS. Let G' be a maximal ISS whose $V(G) \subseteq V(G')$, $E(G) \subseteq E(G')$ and $I(G) \subseteq I(G')$. When $I(G) = I(G')$, the test of G is identical to the test of G', and hence we can safely remove G from Bonferroni factor. When $I(G) \subset I(G')$, $V(G') = V(G)$ and $E(G') = E(G)$. Hence, $I(G)$ is subset of intersection of itemsets associated with $V(G)$, which indicates that G is not proper ISS. Then, any non-maximal ISS depends on a maximal ISS.

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

Property 2 *(Adding a vertex decreases the size of sharing 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 number of Bonferroni factor decreases according to the increase of λ, and hence the minimum P-value associated with the subgraphs increases.

Property 3 *Let $\mathcal{G}_n$ be a set of maximal ISSes that relate n or more items. Between $\mathcal{G}_\lambda$ and $\mathcal{G}_{\lambda+1}$, $\mathcal{G}_\lambda \supseteq \mathcal{G}_{\lambda+1}$ holds. Hence, $|\mathcal{G}_\lambda| \geq |\mathcal{G}_{\lambda+1}|$*

These properties allow us to replace the FIM algorithm with the graph-mining algorithm to find maximal ISSes called COmmon Itemset Network mining

Algorithm 2 gLAMP (IA graph G, class C, significance level α)

```
1: $\lambda \leftarrow |\{i|c_i = 1\}|$, $\delta \leftarrow 1.0$
2: while $\lambda > 0$ do
3:     $\mathcal{G}_\lambda \leftarrow$ run COIN to find maximal ISSes that relate $\lambda$ or more items in $G$
4:     $m_\lambda \leftarrow |\mathcal{G}_\lambda|$
5:     $\delta \leftarrow \alpha / m_\lambda$
6:     if $\delta < f(\lambda - 1)$ then
7:         break
8:     end if
9:     $\lambda \leftarrow \lambda - 1$
10: end while
11: Return the set of itemsets whose P-value $\leq \delta$ in $\mathcal{G}$
```

(COIN) [5] 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 in the LAMP with the COIN.

Minato *et al.* [4] 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 demonstrates 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

1. A.-L. Barabási, N. Gulbahce, and J. Loscalzo. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68, Jan. 2011.
2. M. Civelek and A. J. Lusis. Systems genetics approaches to understand complex traits. *Nature Reviews Genetics*, 15(1):34–48, Jan. 2014.
3. W. 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.
4. S.-i. Minato, T. Uno, K. Tsuda, A. Terada, and J. Sese. Fast statistical assessment for combinatorial hypotheses based on frequent itemset mining. In *Proc. of ECML/PKDD 2014*, 2014.
5. J. Sese, M. Seki, and M. 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.

6. R. Sladek, G. Rocheleau, J. Rung, C. Dina, L. Shen, D. Serre, P. Boutin, D. Vincent, A. Belisle, S. Hadjadj, B. Balkau, B. Heude, G. Charpentier, T. J. Hudson, A. Montpetit, A. V. Pshezhetsky, M. Prentki, B. I. Posner, D. J. Balding, D. Meyre, C. Polychronakos, and P. Froguel. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 445(7130):881–885, Feb. 2007.
7. R. E. Tarone. A modified Bonferroni method for discrete data. *Biometrics*, 46(2):515–522, June 1990.
8. A. Terada, M. Okada-Hatakeyama, K. Tsuda, and J. Sese. Statistical significance of combinatorial regulations. *PNAS*, 110(32):12996–13001, Aug. 2013.
9. T. Uno, T. Asai, Y. Uchida, and H. Arimura. LCM : An efficient algorithm for enumerating frequent closed item sets. In *Proceedings of Workshop on Frequent itemset Mining Implementations (FIMI'03)*, 2003.
10. G. Webb. Preliminary investigations into statistically valid exploratory rule discovery. In S. Simoff, G. Williams, and M. Hegland, editors, *Proceedings of the Second Australasian Data Mining Conference (AusDM03)*, pages 1–9, Sydney, 2003. University of Technology.
11. G. Webb and J. Vreeken. Efficient discovery of the most interesting associations. *Transactions on Knowledge Discovery from Data*, 8(3):15:1–15:31, 2014.