


Gene set analysis with graph-embedded kernel association test

Jialin Qu and Yuehua Cui  *

Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824, USA

*To whom correspondence should be addressed.

Associate Editor: Teresa Przytycka

Received on September 24, 2021; revised on November 20, 2021; editorial decision on December 13, 2021; accepted on December 16, 2021

Abstract

Motivation: Kernel-based association test (KAT) has been a popular approach to evaluate the association of expressions of a gene set (e.g. pathway) with a phenotypic trait. KATs rely on kernel functions which capture the sample similarity across multiple features, to capture potential linear or non-linear relationship among features in a gene set. When calculating the kernel functions, no network graphical information about the features is considered. While genes in a functional group (e.g. a pathway) are not independent in general due to regulatory interactions, incorporating regulatory network (or graph) information can potentially increase the power of KAT. In this work, we propose a graph-embedded kernel association test, termed gKAT. gKAT incorporates prior pathway knowledge when constructing a kernel function into hypothesis testing.

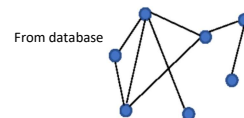
Results: We apply a diffusion kernel to capture any graph structures in a gene set, then incorporate such information to build a kernel function for further association test. We illustrate the geometric meaning of the approach. Through extensive simulation studies, we show that the proposed gKAT algorithm can improve testing power compared to the one without considering graph structures. Application to a real dataset further demonstrate the utility of the method.

Availability and implementation: The R code used for the analysis can be accessed at <https://github.com/JialinQu/gKAT>.

Objective:

1. A statistical test of group effect (e.g. a set of genes $\mathbf{X}_i = [x_{ij}, x_{i2}, \dots, x_{iP}]$ in a pathway) on response (e.g., phenotype, y_i)

Interaction network: G



Novelty:

1. Adding **gene interaction structure** into \mathbf{K} :
 $\mathbf{K} = \mathbf{X}\mathbf{S}\mathbf{X}^T$, $\mathbf{S} = \exp(\delta\mathbf{H})$ is diffusion kernel based on structure of gene set, \mathbf{H} is negative Laplacian matrix of gene structure graph \mathbf{G} .
2. Ranging δ for various \mathbf{S} , combining **p-values** via **Cauchy transformation**

Existing Test Framework:

1. Semiparametric model:
 $y_i = \alpha_0 + \alpha^T \mathbf{Z}_i + \beta(\mathbf{X}_i) + \epsilon_i$
 if $\beta(\mathbf{X}_i)$ is linear in \mathbf{X}_i :
 - regression:
 $y_i = \alpha_0 + \alpha^T \mathbf{Z}_i + \beta^T \mathbf{X}_i + \epsilon_i$
 - Classification:
 $\text{Logit}(P(y_i = 1)) = \alpha_0 + \alpha^T \mathbf{Z}_i + \beta^T \mathbf{X}_i$
2. To inhibit large degrees of freedom --> Kernel-based association test (kernel machine regression):
 Let $\beta \sim \text{Normal}(\mathbf{0}, \tau^2 \mathbf{K})$, $\mathbf{K}_{ij} = \mathbf{K}(\mathbf{X}_i, \mathbf{X}_j)$
 Test statistic Q under null hypothesis $H_0: \tau^2 = 0$ is:
 $Q = (\mathbf{y} - \hat{\mu})^T \mathbf{K} (\mathbf{y} - \hat{\mu})$,
 where $\hat{\mu} = \alpha_j + \mathbf{Z} \hat{\alpha}$, is the prediction under null hypothesis

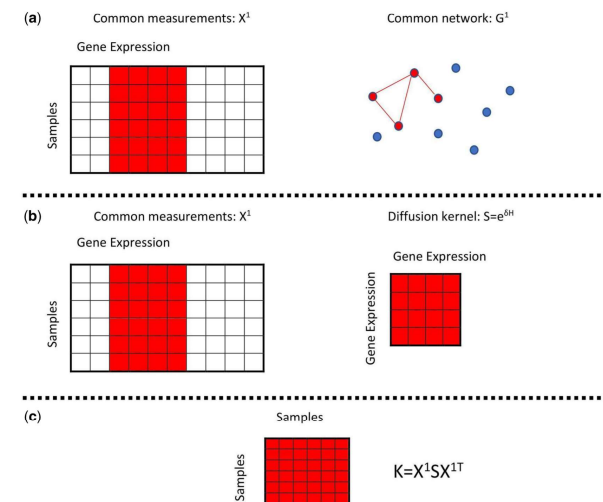


Fig. 1. Step (a): a subset of gene variables (red color) is extracted to form a sub-network based on some prior knowledge (e.g., a KEGG pathway or a GO term); (b) compute a diffusion kernel based on the sub-network graph information extracted from the database; and (c) calculate the final kernel that contains the graph information for further association test