SPADIS: Selecting Predictive and Diverse SNPs in GWAS

Serhan Yılmaz, Öznur Taştan, A. Ercüment Çiçek

Contact: cicek@cs.bilkent.edu.tr or otastan@sabanciuniv.edu

TALK: Saturday, July 7, 2018 11:30 AM-11:40 AM @Grand Ballroom C-F

Introduction

Genome-Wide Association Studies (GWAS) have led to a wide range of discoveries over the last decade where individual variations in DNA sequences, usually single nucleotide polymorphisms (SNPs), have been associated with phenotypic differences. However, individual variants often fail to explain the heritability of complex traits and diseases as large number of variants contribute to these phenotypes and each variant has a small overall effect. Thus, evaluating and associating multiple loci with a given phenotype is critical. An approach to achieve this is to utilize a SNP-SNP interaction network to guide the SNP selection process. An efficient method called SConES follow such an approach to select predictive SNPs over a SNP-SNP network by encouraging the selection of connected SNPs (Azencott et al., 2013). However, we argue that enforcing the selected features to be in close proximity encourages the algorithm to pick features that are in linkage disequilibrium or that have similar functional consequences. Hence, it may lead to the selection of functionally redundant SNPs and the loss of variants that cover different processes. This is illustrated in Figure 1.

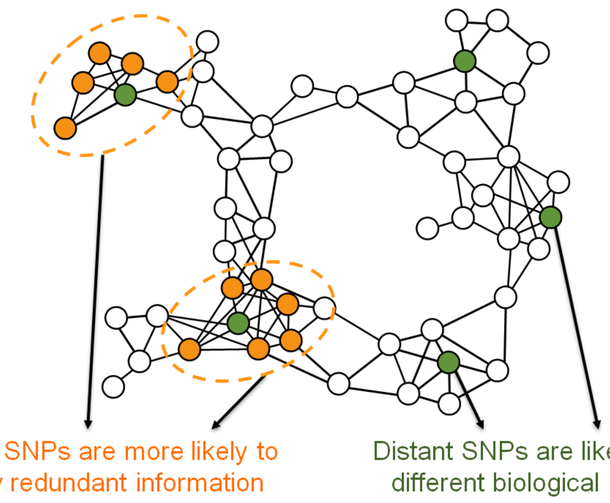

Fig. 1. A toy example that illustrates the intuition behind SPADIS.

Our Method — SPADIS

We hypothesize that diversifying the SNPs in terms of location would result in covering complementary modules in the underlying network that lead to the phenotype. Based on this rationale, we present SPADIS: A novel SNP selection algorithm over a SNP-SNP network that favors;

- loci with high univariate associations with the phenotype
- loci that are diverse in the sense that they are far apart on a loci interaction network.

SPADIS is formalized as a feature selection problem over a network of SNPs. The problem is to find a SNP subset S with cardinality at most $k \ll n$ (the number of SNPs) that explains the phenotype. To this end, we utilize a two-step approach. In the first step, we assess the relation of each SNP to the phenotype individually using the Sequence Kernel Association Test (SKAT) (Wu et al., 2011). In the second step, our goal is to maximize the total score of SNP set while ensuring the selected set consists of SNPs that are remotely located on the network. To encode this intuition, we define the submodular set function shown in Figure 2.

SNP association score — Proximity penalization function

$$F(S) = \sum_{i \in S} \left(c_i + \beta \left(1 - \sum_{j \in S} \frac{K(i,j)}{2k} \right) \right)$$

F is submodular!

β determines the magnitude of penalization

Shortest path distance

$$K(i,j) = \begin{cases} 1 - d(i,j)/D & d(i,j) \le D, \quad i \neq j \\ 0 & otherwise \end{cases}$$

D is distance threshold parameter

Fig. 2. Submodular set function F that is maximized by SPADIS.

How is submodularity useful?

Subset selection problem with cardinality constraint is NP-hard. Thus, exhaustive search is infeasible when k or n is not small. For this reason, we make use of the fact that the function defined in Figure 2 is submodular. Although submodular optimization itself is NP-hard as well, the greedy algorithm given in **Algorithm 1,** proposed by Nemhauser et al. (1978), guarantees a (1-1/e)-factor approximation to the optimal solution under cardinality constraint for monotonically non-decreasing and non-negative submodular functions.

Algorithm 1 Greedy Algorithm

Input: Set function F, ground set V, cardinality constraint $k \le |V|$.
Output: Set $S \subset V$ such that $|S| = k$.

1: $S \leftarrow \emptyset$
2: **while** $|S| < k$ **do**
3: $\quad S \leftarrow S \cup \operatorname{argmax}_{x \in V \setminus S} F(S \cup x)$
4: **end while**

SNP-SNP Networks

We construct four undirected SNP-SNP networks, three of which are defined in Azencott et al. (2013): GS (gene sequence) network, GM (gene membership) network and GI (gene interaction) network. Additionally, we introduce a new network (GS-HICN) to investigate the usefulness of the 3D conformation of the genome in the context of SNP selection problem. GS-HICN connects loci that are significantly close in 3D (p-value < 0.05) in addition to adjacent loci on the DNA sequence (GS). To assess thxe statistical significance for close loci, we process the intrachromosomal contact matrices using Fit-Hi-C method (Ay et al.,2014). All four networks are illustrated in Figure 3.

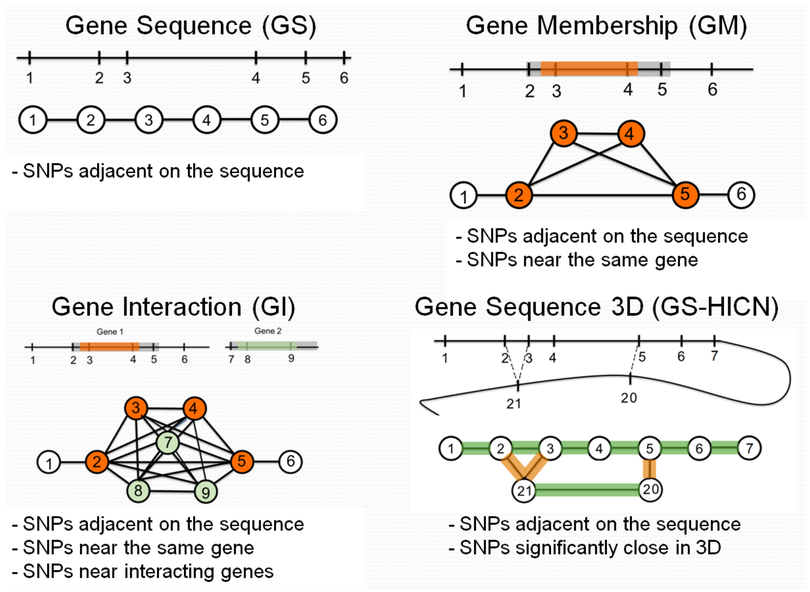

Fig. 3. Visualization of the SNP-SNP networks that are constructed.

Experimental Setting

We evaluate performance of the compared SNP selection methods on the 17 flowering time phenotype dataset of *Arabidopsis Thaliana (AT)* containing up to 180 samples and 173 219 SNPs with minor allele frequency (MAF) ≥ 10%.

We compare SPADIS with the following methods:

- *SConES*: A network-constrained SNP selection method with a max flow based solution (Azencott et al., 2013).
- *Univariate*: We run univariate linear regression and select SNPs that are found to be significantly associated with the phenotype.
- *Lasso*: The Lasso regression that minimizes the prediction error with the L1-regularizer of the coefficient vectors.

We perform the parameter selection of the compared methods using two metrics separately *stability*, denoted with (S) and measured using the consistency index, and *regression performance*, denoted with (R) measured using Pearson's squared correlation coefficient.

Phenotype Prediction Performance

First, we compare the methods with maximum cardinality constaint where the number of SNPs selected is upper bounded by 1733 i.e. 1% of the number of all SNPs. When regression performances (R^2) are averaged over all phenotypes and all networks, SPADIS outperforms all other methods by a fair margin followed by SConES(R) —see Figure 4 (Left). Next, we perform additional experiments for the best two performing methods (SPADIS and SConES) by constainting them to select close to the same number of SNPs of k. We experiment with 4 different constraints and show that SPADIS outperforms SConES in all cases, on average —see Figure 4 (Right).

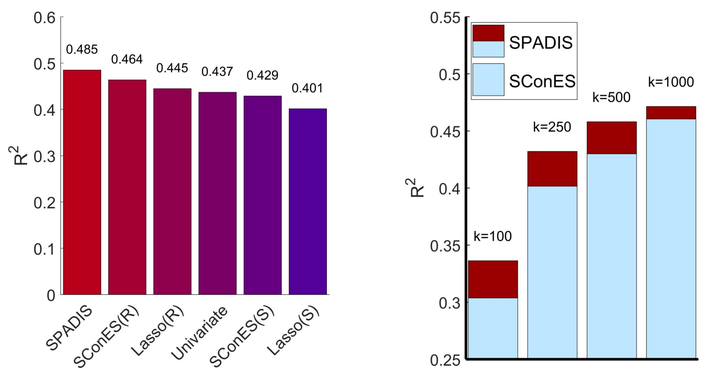

Fig. 4. (Left) Pearson's squared correlation coefficient (R^2) obtained for maximum cardinality constraint of 1733. Methods are ordered in descending order of R^2. (Right) The improvement of SPADIS over SConES in terms of R^2 for different tight cardinality constraints k. Blue bar indicates the maximum of SConES(S) and SConES(R), red bar indicates the amount of improvement of SPADIS over SConES. (Left & Right) All values shown are averages over 17 phenotypes and 4 networks.

References

Ay, F. et al. (2014). *Genome research*, 24(6), 999–1011.
Azencott, C. A. et al. (2013). *Bioinformatics*, 29(13), i171–i179.
Nemhauser, G. L. et al. (1978). *Mathematical Programming*, 14(1), 265–294.
Wu, T. T. et al. (2009). *Bioinformatics*, 25(6), 714–721.

Funding: This work is supported by TUBITAK via Career Grant #116E148 to AEC

Time Performance

We report the CPU runtime of all methods on all four networks. The runtime tests are conducted for one cross-validation fold with preset parameters on a single phenotype of FT Field with the most number of samples available. Results show that SPADIS is more efficient than all other methods except Univariate (baseline method) —see Figure 5.

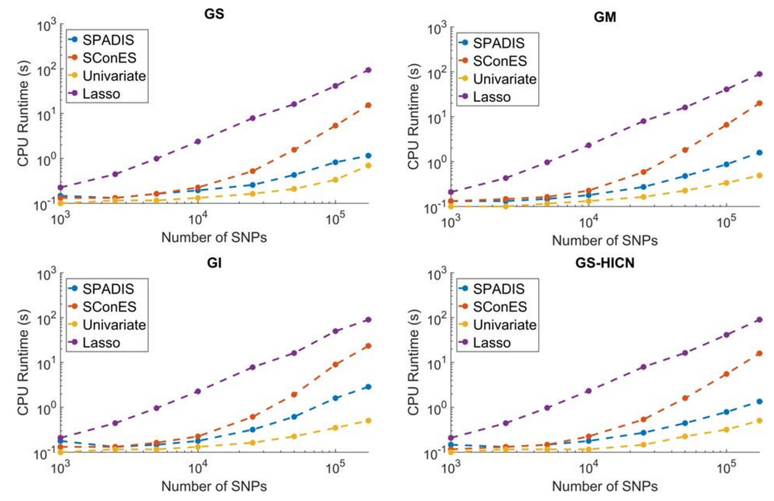

Fig. 5. CPU time measurements of SPADIS, SConES, Univariate and Lasso from 1.000 to 173.219 SNPs on all four networks.

Diverse Selection of SNPs

The goal of SPADIS is to select a diverse set of SNPs over the SNP-SNP network. We hypothesize that SNPs selected with SPADIS overlap with more diverse biological processes and that the prediction performance is reinforced by this effect. We investigate whether this hypothesis is supported by empirical values by utilizing two metrics:

(1) Genes-Hit:

- There are some genes that are known *a priori* to be associated with the phenotype, denote them as **target genes**
- A gene is considered **hit**, if a SNP near that gene is selected
- *Genes-Hit* is the number of **target genes** that are **hit**

(2) GO-Hit:

- Each gene are associated with a number of Gene Ontology (GO) annotated biological processes
- A GO term is considered **hit**, if a gene associated with that term is **hit**
- GO-Hit is the number of GO-terms that are **hit**.

As shown in Table 1, SPADIS hits 7% to 46% more distinct candidate genes and 5% to 17% more GO annotated biological processes compared to the next best performing method on average.

Table 1. Statistics about the genes and biological processes hit by the selected SNPs sets of all SNP selection methods when tight cardinality constraint of k is applied. The reported results are averages over all 17 phenotypes and 4 networks. The best result for each cardinality constraint k is marked as bold.

Metric	k	SPADIS	SConES(S)	SConES(R)	Univariate	Lasso
Genes-Hit	100	**5.9**	4.4	4.5	3.8	5.5
	250	**12.9**	8.7	9.0	7.6	10.9
	500	**23.4**	14.3	15.0	13.8	18.3
	1000	**40.8**	24.7	23.6	24.2	27.9
GO-Hit	100	**151**	114	117	137	144
	250	**306**	230	236	266	280
	500	**491**	373	382	424	441
	1000	**747**	597	581	659	636

Contribution of Hi-C data

We assess the contribution of using the Hi-C data by comparing the regression performances of the models built on GS-HICN to the models built on other three networks (GS, GM, GI). We perform five experiments with different cardinality constaints. As shown in Table 2, Hi-C data provides improvements in regression performance on average: 1.4% higher than GS and GM and 1.9% higher than GI. Moreover, the improvement can be considered consistent since GS-HICN performs better than other networks on average in 4 out of 5 experiments. Furthermore, GS-HICN hits 3.0% to 6.6% more genes and 2.7% to 21.9% more biological processes compared to other networks, on average (result not shown).

Table 2. The average Pearson's squared correlation coefficient averaged over all 17 phenotypes and all methods (SPADIS, SConES(S) and SConES(R)). The best result for each experiment is marked as bold.

Experiment Constraint	k	Network GS	GM	GI	GS-HICN
Tight	100	0.310	0.311	0.309	**0.314**
Tight	250	0.403	0.406	0.398	**0.415**
Tight	500	0.438	0.438	0.432	**0.445**
Tight	1000	0.461	0.461	0.459	**0.467**
Maximum	1733	0.457	0.456	**0.462**	0.461
Average		0.414	0.414	0.412	**0.420**