Network-based Metric space for Phenotypic Stratification of Samples Using Transcriptome Profiles

Inyoung Sung[1], Dohoon Lee[1†], Sangseon Lee[2†] and Sun Kim[1,3,4,*]

[1]Interdisciplinary Program in Bioinformatics, Seoul National University
[2]BK21 FOUR Intelligence Computing, Seoul National University
[3]Department of Computer Science and Engineering, Seoul National University
[4]Institute of Engineering Research, Seoul National University

Abstract

- Although there are recent studies using transcriptome profiles to stratify samples in a clinically or phenotypic meaningful way, it is difficult to deal with the high-dimension genetic space while considering the complex interactions between genes.
- To reduce high-dimension genetic space to a lower dimension space and address the complex in the gene-gene interactions, we proposed a network-based two-step computational framework using transcriptome profiles and biological network.
- The proposed method successfully stratified the samples.

Motivation

- The development of high-throughput sequencing technology has made it possible to obtain a large amount of transcriptome data of biological samples, and it will be able to measure the dysfunction of regulatory mechanisms according to the abnormality using the transcriptome data.
- However, it is difficult to analyze transcriptome data considering more than 20,000 genes and their interactions through a relatively small sample.
- To overcome these difficulties, we proposed a computational method that can measure phenotypic changes in samples using transcriptome data and biological network.

Methods

- As illustrated in Figure 1, the proposed network-based computational method consists of two steps:
  1. **Construction of condition specific subnetworks**
     : in a public biological network, we used a network propagation algorithm [1] to redefine gene interactions and a community detection algorithm [2] to cluster genes in the network.
  2. **Map into a metric space**
     : in each subnetwork constructed in the step 1, measure the phenotypic changes of each sample compare to normal state sample in two perspectives – network-level Jensen-Shannon divergence JSD (motivated by Scientific Reports 2016 [3]) and network structure-level importance CC, eventually we defied $NetJSD$

$$\frac{1}{|\mathrm{N}|}\sum_{v\in \mathrm{N}} CC_v \times JSD_v(P^N||P^Q)$$

where N is set of nodes in network, |N| is number of nodes, v is node in N, N is normal state sample and Q is query sample. As a result single sample mapped into a metric space generated by subnetworks in the form of vector consisted $NetJSD$s.

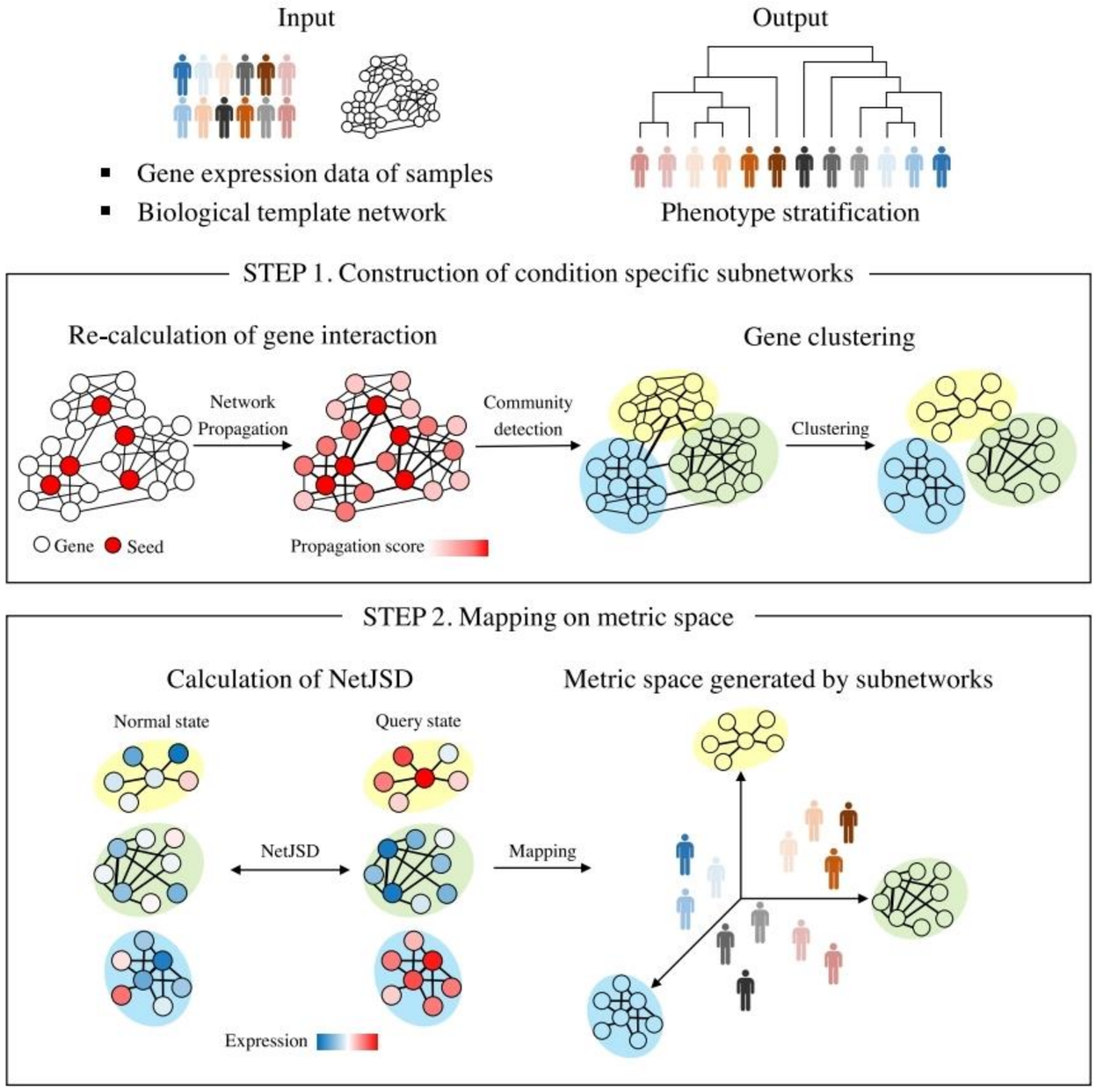
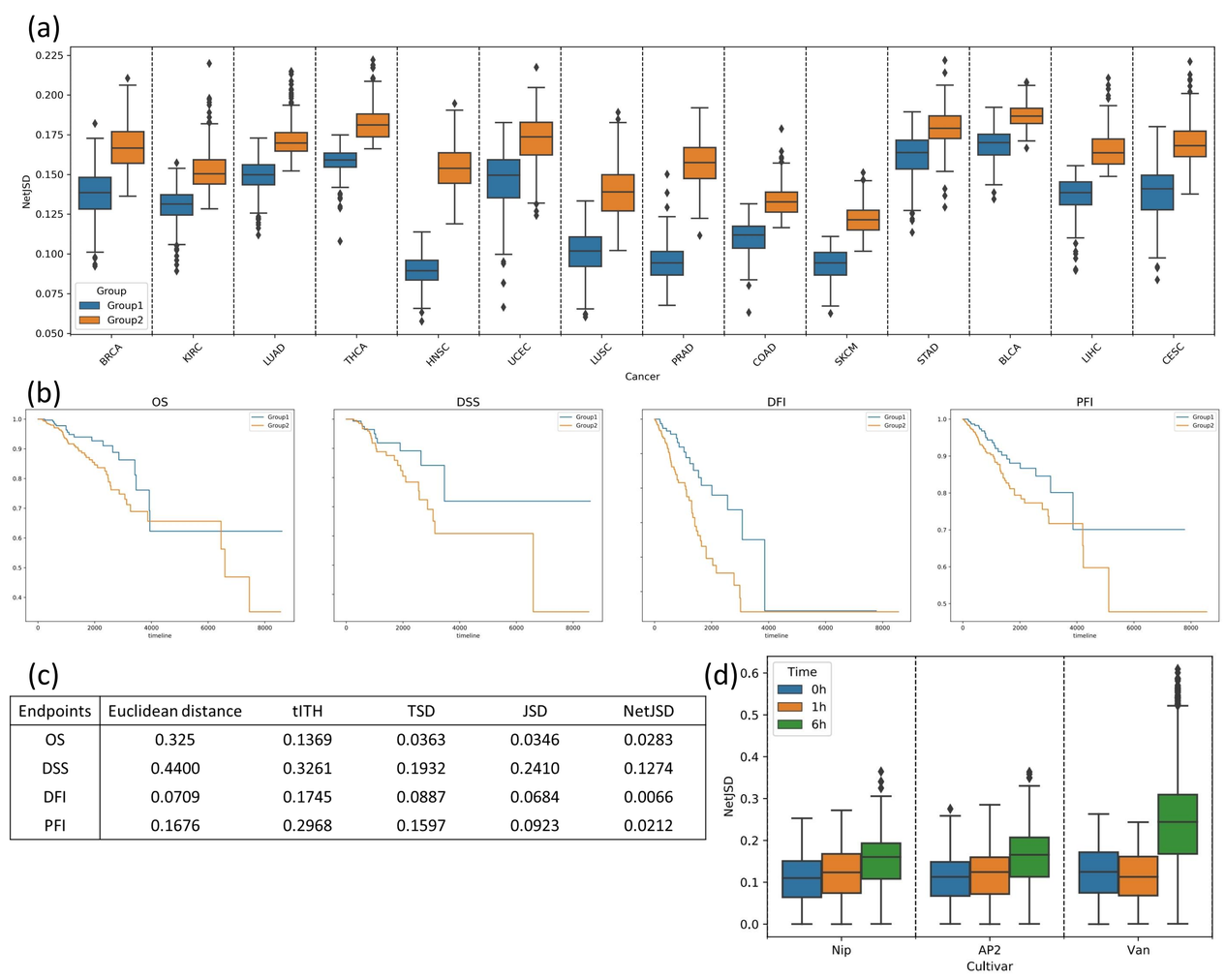


Figure 1. The overview of the proposed method. We generated metric spaces with a two-step computational framework using transcriptome profiles and biological networks.

Results

- Dataset – transcriptome data and public biological network
  - 14 cancer patients from Pan-Cancer Atlas and three Oryza sativa cultivars from GEO
  - Homo sapiens and Oryza sativa protein-protein interaction network from SRING
- **Pan-cancer dataset**: for each cancer, the proposed method constructed subnetworks generating a metric space, in which the samples are divided into two survival groups: high- and low-risk. As a result, the two risk groups had different distributions of average $NetJSD$s (Figure 2 (a)).
  - For breast cancer patients, it showed significant results in the survival analysis for the four clinical endpoints: OS, DSS, DFI and PFI (Figure 2 (b)). In addition, the proposed method outperformed the existing three methods in survival analysis for all endpoints (Figure 2 (c)).

  Therefore, these result showed that our method stratified cancer patients as clinical meaningful way in the metric space while reducing the gene space.
- **Oryza sativa dataset**: in Figure 2 (d), we showed how expression differed between cultivars or over time in three Oryza sativa cultivars as drought stress persisted. From this result, it can be argued that the $NetJSD$ calculated in each subnetwork measured the difference between cultivars and the passage of time in the same way as the biological perspective.



(c)

Endpoints	Euclidean distance	tITH	TSD	JSD	NetJSD
OS	0.325	0.1369	0.0363	0.0346	0.0283
DSS	0.4400	0.3261	0.1932	0.2410	0.1274
DFI	0.0709	0.1745	0.0887	0.0684	0.0066
PFI	0.1676	0.2968	0.1597	0.0923	0.0212

Figure 2. (a) Boxplot of $NetJSD$ on 14 cancers from Pan-caner Atlas. (b) Survival analysis results on four clinical endpoints of the breast cancer. (c) Performance comparison results on the breast cancer. (d) Boxplot of $NetJSD$ on three Oryza sativa cultivars.

Conclusion

- We proposed a **novel network-based sample stratification method** using transcriptome data and biological networks to generate a metric space for quantitatively measuring phenotypic changes.
- Through gene clustering and network propagation, this method not only reduced the genomics space from high to low dimensions, but also constructed condition specific subnetwork reflecting the properties of the given data.
- Furthermore, $NetJSD$ measured from two perspectives – network-level JSD and network structure-level importance in the network was able to capture the phenotypic change of the sample.
- The proposed computational method has the potential to measure the difference between samples with any kind of data.

References

[1] Can, Tolga, Orhan Çamoğlu, and Ambuj K. Singh. "Analysis of protein-protein interaction networks using random walks." Proceedings of the 5th international workshop on Bioinformatics. 2005.
[2] Blondel, Vincent D., et al. "Fast unfolding of communities in large networks." Journal of statistical mechanics: theory and experiment 2008.10 (2008): P10008.
[3] Park, Youngjune, et al. "Measuring intratumor heterogeneity by network entropy using RNA-seq data." Scientific reports 6.1 (2016): 1-12.