Network−Based Metric Space for Phenotypic Stratification of Samples Using Transcriptome Profiles

Inyoung Sung[1], Dohoon Lee[1+], Sangseon Lee[2+] and Sun Kim[134*]

[1] Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Republic of Korea.

[2] BK21 FOUR Intelligence Computing, Seoul National University, Seoul, Republic of Korea.

[3] Department of Computer Science and Engineering Seoul National University, Seoul, Republic of Korea.

[4] Institute of Engineering Research Seoul National University, Seoul, Republic of Korea.

[+] Equal contributor

* Correspondence: sunkim.bioinfo@snu.ac.kr

Clinical or phenotypic stratification of samples through transcriptome data is of great interest in the era of high−throughput sequencing. However, existing stratification methods lack efficient utilization of gene interaction information, and furthermore, handling more than 20,000 genes causes the curse of high dimensionality that hinders elucidating the linkage between genetic profiles and clinical or phenotypic differences. To reduce high dimensional genetic space to a lower dimension space, we propose a network−based two−step computational framework. We first reduce dimensions of transcriptome to a few tens of dimensions by mapping transcriptome to protein interaction network followed by performing network propagation algorithm and clustering analysis. Then, each network is converted into a single numeric metric by utilizing information theoretic quantification of gene expression abnormality, which results in a single sample mapping to a metric space generated by each subnetwork in the form of vectors. The proposed network−based stratification method was used to analysis two different public datasets: 14 cancer types patients from the Pan−Cancer Atlas and drought response over time for three Oryza sativa cultivars from GEO. Extensive experiments showed that our method generates a metric space that captures data−specific biological functions and improves the stratification performance compared to existing methods. Therefore, the metric space generated by the proposed method efficiently captured clinical or phenotype information of samples and successfully stratified the samples, addressing the problem in the complex gene space. The proposed method is implemented in Python and available at https://github.com/Sunginyoung/net_stratification.

Keywords: Information theory, Network propagation, Gene clustering, Pan−Cancer, Oryza sativa, Sample stratification