GSEH: A novel approach to select prostate cancer-associated genes using gene expression heterogeneity

Hyunjin Kim, Sang-min Choi, Sanghyun Park*

Department of Computer Science, Yonsei University, 134 Sinchon-dong, Seodaemun-gu, Seoul 120-749, South Korea

* Corresponding author. Tel.:+82 2 2123 5714; fax:+82 2 365 2579; e-mail: sanghyun@yonsei.ac.kr

Abstract— When a gene shows varying levels of expression among normal people but similar levels in disease patients or shows similar levels of expression among normal people but different levels in disease patients, we can assume that the gene is associated with the disease. By utilizing this gene expression heterogeneity, we can obtain additional information that abets discovery of disease-associated genes. In this study, we used collaborative filtering to calculate the degree of gene expression heterogeneity between classes and then scored the genes on the basis of the degree of gene expression heterogeneity to find "differentially predicted" genes. Through the proposed method, we discovered more prostate cancer-associated genes than ten comparable methods. The genes prioritized by the proposed method are potentially significant to biological processes of a disease and can provide insight into them.

Index Terms— Gene selection, Gene prioritization, Disease-associated genes, Prostate cancer-associated genes, Gene expression heterogeneity

1 Introduction

The average life expectancy of human has increased throughout the world on account of advancements in medical science [1]. With large gains in life expectancy, a rising interest exists in disease management. If the diagnosis and prognosis are precisely predicted, the correct therapeutic methods can be used and significant disease damage can thereby be avoided. Indicators (biomarkers), such as genes or proteins, are typically used in predicting the diagnosis and prognosis of a disease [2-3]. Many biologists must choose which genes or proteins to investigate; therefore, gene prioritization has become increasingly important. Four computational strategies for gene prioritization exist [4]: filtering, text mining, similarity profiling and data fusion, and network-based. In the filtering strategy, filters are defined by properties of the ideal candidate gene. In the text-mining strategy, disease-relevant keywords are employed to retrieve disease-relevant literature, which is mined to identify candidate genes. In the similarity profiling and data fusion strategy, similarities between the candidate genes and known genes from various data sources are considered. In the network-based strategy, candidate genes in a gene network are selected based on the distance between the candidate genes and known disease genes. The proposed method is categorized as a filtering strategy because it employs a filter defined by heterogeneous gene expression characteristics.

Genes that are differentially expressed between two different conditions (i.e., malignant and benign) have received considerable attention because they are expected to predict the diagnosis and prognosis of the disease [5-6]. Feature selection methods can be used to identify genes that are differentially expressed between the two different conditions. In bioinfor-