AN INTEGRATIVE CANCER CLASSIFICATION BASED ON GENE EXPRESSION DATA

H.F. Ong[1], *N. Mustapha*[*, 2]

[1] School of Information Technology, Faculty of Business and Information Science, UCSI University, Cheras , Kuala Lumpur, Malaysia

[2] Department of Computer Science, Faculty of Computer Science and Information Technology, University Putra Malaysia, Serdang, Selangor Darul Ehsan, Malaysia

* corresponding author : norwati@fsktm.upm.edu.my

The advent of integrative approach has shifted cancer classification task from purely data-centric to incorporate prior biological knowledge. Integrative analysis of gene expression data with multiple biological sources is viewed as a promising approach to classify and to reveal relevant cancer-specific biomarker genes. The identification of biomarker genes can be used as a powerful tool for understanding the complex biological mechanisms, and also for diagnosing and treatment of cancer diseases. However, most integrative-based classifiers only incorporate a single type of biological knowledge with gene expression data within the same analysis. For instance, gene expression data is normally integrated with functional ontology, metabolic pathways, or protein-protein interaction networks, where they are then analysed separately and not simultaneously. Apart from that, current methods generates a large number of candidate genes, which still require further experiments and testing to identify the potential biomarker genes. Hence, this study aims to resolve the problems by proposing a systematic integrative framework for cancer gene expression analysis to the classification task. The association based framework is capable to integrate and analyse multiple prior biological sources simultaneously. Set of biomarker genes that are relevant to the cancer diseases of interest are identified in order to improve classification performance and its interpretability. In this paper, the proposed approach is tested on a breast cancer microarray dataset and integrated with protein interaction and metabolic pathway data. The results shows that the classification accuracy improved if both protein and pathways information are integrated into the microarray data analysis.