A GO BASED REPRESENTATION FOR PROGNOSIS AND INFERENCE FROM MICROARRAY DATA

Shashank Srivastava, Snigdha Chaturvedi, Arnab Bhattacharya

Dept of Computer Science, Indian Institute of Technology, Kanpur

ssriva@iitk.ac.in

Study of microarrays can assist mining of valuable biological information, building predictive models, and unraveling latent correlations signifying biological pathways. Several techniques have focused on identifying differentially expressed genes, and use of dimensionality reduction techniques to overcome the "curse of dimensionality"; including the Gene-set approach to evaluate expression patterns of gene groups instead of individual genes. While existing methods have been useful in determining differentially regulated genes, and building predictive models from gene-expression values, they do not allow direct inference of relations between gene-expression values, or insight into higher level biological concepts.

We focus on analyzing microarray for interpretability of various pathological conditions. This study provides a concise quantitative representation of microarray data in terms of biological concepts, using a knowledge infusion from the structure of the Gene Ontology database. We exploit the hierarchical structure in the Gene Ontology to quantitatively express a medical condition in terms of gene concepts. All genes reachable from a GO term are used to estimate the relative expression of the GO term, taking into account the number of paths. The representation then leads to identification of differentiating gene concepts between two conditions to draw precise biological inferences and differences; directly yielding biological understanding and interpretability.

The proposed method is tested on two standard datasets (DLBCL and Leukemia datasets). Prognostic predictions in both cases are seen to be corroborated by existing biological literature. Classification accuracies from the representation scheme using decision trees compare with advanced statistical methods, suggesting stability of the approach. Inferences additionally suggest a few pointers indicating directions for future exploration for biochemists. The approach can be especially useful in case of complex diseases, where underlying causes and etiology are still unknown.

PAPER ID: 213