Robust Prediction-Based Analysis for Genome-Wide Association and Expression Studies

Skanda K.Koppula[14], Amin Zollanvari[123], Ning An[5], Gil Alterovitz[1234*]

[1] Center for Biomedical Informatics, Harvard Medical School [Boston, MA 02115].

[2] Children's Hospital Informatics Program at Harvard-MIT Division of Health Science [Boston, MA 02115].

[3] Partners Healthcare Center for Personalized Genetic Medicine [Boston, MA 02115].

[4] Department of Electrical Engineering and Computer Science at Massachussetts Institute of Technology [Cambridge, MA 02139].

[5] Gerontechnology Lab, Hefei University of Technology [Hefei, China 230009]

* Corresponding author. Contact: gil@mit.edu

Abstract: Here we describe a prediction-based framework to analyze omic data and generate models for both disease diagnosis and identification of cellular pathways which are significant in complex diseases. Our framework differs from previous analysis in its use of underlying biology (cellular pathways/gene-sets) to produce predictive feature-disease models. In our study of alcoholism, lung cancer, and schizophrenia, we demonstrate the framework's ability to *robustly* analyze omic data of *multiple* types and sources, identify significant features sets, and produce accurate predictive models.

We found novel significant pathways for each disease and developed models with predictive powers of 83%, 81%, and 76%, respectively for the three diseases. This is thus the first 'good' predictor of alcohol dependence [1]. Our higher consistency between analyses of multiple datasets demonstrates that our method is robust. This enables intervention when the onset of alcoholism or schizophrenia appears likely, and can help us better understand underlying biological mechanisms and risk factors.

Introduction

The rapid accumulation of high-throughput genomic and proteomic data has offered unprecedented opportunities for biologists to gain insights into disease. Here we demonstrate a robust, prognostic framework to identify biologically meaningful feature sets and develop predictive models for disease.

Method

We first develop Bayesian models linking each feature in a pathway to the disease phenotype, via a curated SNP-gene-pathway mapping. Then, our method sorts the models based on accuracy as a phenotypic classifier, identifying models and corresponding feature sets with significant predictive accuracy. From this, we identified disease-significant pathways and eliminated the necessity for background gene selection. Using the best performing feature-sets in a final model, we created an aggregate prognostic model for each disease. Three multifactorial diseases were studied: alcoholism (SNP data), lung cancer (gene expression data), and schizophrenia (SNP data from a differing chip platform). For each disease, we independently analyzed two data sets (n = 1395and 1367) to understand the method's robustness.

Results and Discussion

In our study of alcoholism, we identified four clinically-ascertainable pathways that have not been previously tied to the behavior. These pathways were common to our analysis results of our two alcoholism data-sets. The aggregate model of these gene-sets yielded the first 'good' predictive model of alcoholism (AUC = .83) when testing on a merged dataset. In our second study, studying lung cancer, we found seven pathways common significant pathways and produced an aggregate model (AUC of .81). Our significant pathway results from lung cancer exhibited higher percentages of than cross-data overlap than a corresponding one done in the seminal 2005 GSEA paper. In our final study, for schizophrenia, we identified three novel gene sets significant to the disease and developed a fair aggregate predictive model (AUC = .76).

The power of our predictive enrichment framework lies in its biological meaning (bases models on the studied pathway biology in order to draw further insight), robustness, and flexibility – the choice of data modality, predictive structures, choice of feature sets, and the number of central phenotypic outcomes.

[1] Simon. 2004. *Evaluating the AUC for an ROC curve***.** Children's Mercy Hospital Report.