Gene Expression Classification Based on Metabolic Modeling

Livnat Jerby

Introduction

Transcriptomic Classification

Advances in DNA microarray technologies have made it feasible to obtain genome-scale gene expression profiles of different tissue samples and cell-lines under various conditions. Expression data enables us to gain insight on biological mechanisms via the comparison of gene profiles in normal and pathological conditions. It aids the identification of interrelationships among genes, the understanding of progressive mechanisms of diseases at a molecular level, as well as the identification and validation of novel drug targets. Frequently, patients that are diagnosed with the same disease vary in the underlying mechanism of their pathology, in their response to treatment, and in their prognosis. The basic idea is that the more accurate the diagnosis is, the more effective the treatment would be. This observation and the increasing availability of DNA microarray data have led the utilization of a plethora of classifier design approaches (e.g., Classification and Regression Trees (CART), Support Vector Machines (SVM), Artificial Neural Networks (ANN)), in the attempt to classify expression profiles.

To design a classifier we utilize available training samples (e.g., gene expression profiles) from different classes. Each sample is composed of features (e.g. genes). During the training phase the classifier estimates or adjusts the parameters for classification according to its particular underlying paradigm and the information extracted from the training set to optimize the classification. In the next stage the predictive performances of the classifier are tested on unseen samples. The accuracy of the classifiers performances depends strongly on the representation of the samples. In a way, the classifier can be considered as an artificial intelligence device that could be used as a computational oracle. Support vector machines (SVMs) [1] are supervised classifiers. They often outperform other learning algorithms and are fairly insensitive to the phenomena known as the curse of dimensionality (i.e., deteriorate classification performances caused by a large number of features),

which is relevant for gene expression profiles, that consist of several thousand genes with only a few dozen of samples available as training data.

Constraint-Based Modeling (CBM)

Deducing a phenotype based on expression profiles has been addressed by diverse computational methodologies, outside the realm of machine learning. One of these methods is metabolic modeling. Mathematical modeling of cellular metabolism has traditionally been performed through kinetic modeling techniques that require detailed information on kinetic constants and on enzyme and metabolite levels [2]. However, the lack of accurate cellular information of enzymes kinetics and levels currently limits the applicability of such methods to small-scale systems. An alternative computational approach, constraint-based modeling (CBM), bypasses these hurdles as it does not depend on such detailed information. CBM assumes a metabolic steady state under which feasible flux distributions satisfy a stoichiometric mass-balance requirement, thermodynamic constraints and constraints on enzymes capacities that are based on experimental observations of flux rates. This modeling paradigm has been extensively applied with considerable success to study microbial physiology [3–10]. Though still less developed, large-scale modeling of human metabolism is constantly progressing [11]; earlier study has focused on characterizing distinct human metabolic pathways [12, 13], and modeling specific cell types and organelles [15–17]. In 2007, two generic human metabolic models were presented on the basis of an extensive evaluation of genomic and bibliomic data [8, 14]. The potential clinical utility of the generic model was previously demonstrated by its ability to identify functionally related sets of reactions that are causally related to hemolytic anemia, and potential drug targets for treating hypercholesterolemia [8].

Research Objectives

In this study we set out to integrate between CBM and transcriptomic classification by exploiting the metabolic model to aid and improve the classification. The objective is to generate a paradigm that would enable the classifier to account for the metabolic aspects that are concealed in the expression data. To do so, we applied different methods of CBM to preprocess the samples for classification. To experiment and validate the different possibilities, we used gene expression in brain and in liver of laboratory mice, fed one of four different diets.

The work was done with the guidance of Dr. Lior Wolf and Prof. Eytan Ruppin.

References

[1] Vapnik V. (1998) Statistical Learning Theory, Chichester, Wiley, UK

[2] Garfinkel D, Hess B (1964) Metabolic control mechanisms. VII.A detailed computer model of the glycolytic pathway in ascites cells. J Biol Chem **239**: 971983

[3] Edwards JS, Ibarra RU, Palsson BO (2001) In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. Nat Biotechnol **19**: 125130

[4] Segre D,VitkupD, ChurchGM (2002) Analysis of optimality in natural and perturbedmetabolic networks. Proc Natl Acad Sci USA 99: 1511215117

[5] Shlomi T, Berkman O, Ruppin E (2005) Regulatory on/off minimization of metabolic flux changes after genetic perturbations. Proc Natl Acad Sci USA 102: 76957700

[6] Feist AM, Palsson BO (2008) The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli. Nat Biotechnol 26: 659667

[7] AbuOun M, Suthers PF, Jones GI, Carter BR, Saunders MP, Maranas CD,Woodward MJ, Anjum MF (2009) Genome scale reconstruction of a Salmonella metabolic model: comparison of similarity and differences with a commensal Escherichia coli strain. J Biol Chem 284: 2948029488

[8] Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, Srivas R, Palsson BO (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc Natl Acad Sci USA 104: 17771782

[9] Kumar VS, Maranas CD (2009) GrowMatch: an automated method for reconciling in silico/in vivo growth predictions. PLoS Comput Biol 5: e1000308

[10] Oberhardt MA, Palsson BO, Papin JA (2009) Applications of genomescale metabolic reconstructions. Mol Syst Biol 5: 320

[11] Goodacre R, Vaidyanathan S, Dunn WB, Harrigan GG, Kell DB (2004) Metabolomics by numbers: acquiring and understanding global metabolite data. Trends Biotechnol 22: 245252

[12] Kanehisa M, Goto S (2000) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28: 2730

[13] Romero P, Wagg J, Green M, Kaiser D, Krummenacker M, Karp P (2004) Computational prediction of human metabolic pathways from the complete human genome. Genome Biol 6: R2

[14] Ma H, Sorokin A, Mazein A, Selkov A, Selkov E, Demin O, Goryanin I (2007) The Edinburgh human metabolic network reconstruction and its functional analysis. Mol Syst Biol 3: 135

[15] Wiback SJ, Palsson BO (2002) Extreme pathway analysis of human red blood cell metabolism. Biophys J 83: 808818

[16] Chatziioannou A, Palaiologos G, Kolisis FN (2003) Metabolic flux analysis as a tool for the elucidation of the metabolism of neurotransmitter glutamate. Metab Eng 5: 201210

[17] Vo TD, Greenberg HJ, Palsson BO (2004) Reconstruction and functional characterization of the human mitochondrial metabolic network based on proteomic and biochemical data. J Biol Chem 279: 3953239540