**ExPlain: Causal Analysis of Gene Expression Data from Promoter Models to Signaling Pathways**

Alexander Kel\(^1\)  
Holger Karas\(^1\)  
Nico Voss\(^1\)  
Tagir Waleev\(^2\)

ake@biobase.de  
hka@biobase.de  
nvo@biobase.de  
amaembo@gmail.com

Edgar Wingender\(^1,3\)  
ewi@biobase.de

\(^1\) BIOBASE GmbH, Halchtersche Str. 33, D-38304 Wolfenbüttel, Germany  
\(^2\) A.P. Ershov's Institute of Informatics Systems, 6, Lavrentiev ave., 630090 Novosibirsk, Russia  
\(^3\) Dept. Bioinformatics, UKG/Univ. Göttingen, Goldschmidstr. 1, 37077 Göttingen, Germany

Keywords: gene expression, microarray analysis, promoter models, transcription factors, binding sites, genetic algorithm

1 Introduction

Cellular signal transduction networks of multicellular organisms are enormously complex though very robust in providing fast and appropriate response to any extracellular signal. This is achieved through combinatorial usage of a rather limited set of signaling molecules and pathways. These combinatorics must be mirrored by the structure of gene promoters as combinations of transcription factor binding sites (composite modules). Different signal transduction pathways leading to the activation of transcription factors converge at key molecules that master the regulation of certain cellular processes. Such crossroads of signaling networks often appear as “Achilles Heels” causing a disease when not functioning properly.

Several methods were developed for the analysis of signal transduction and gene regulatory networks associated with gene expression data. However, these approaches often underestimate the role of molecular processes that occur in the cell on the post- (or pre-) transcriptional level – the “ iceberg” of signal transduction that cannot be seen on the level of gene expression changes. In addition, the promoter structure which is the key component linking the gene regulation with signal transduction network through the multiple interaction of transcription factors to their DNA binding sites are purely understood. Eventually, all the afore mentioned approaches to network analysis operate with the final products of differentially expressed genes as well as their effects on the physiology of the cell, but are not focused on the molecular mechanisms that cause the observed changes in gene expression.

2 Method and Results

We developed an integrated computational tool, ExPlain for causal interpretation of gene expression data. It analyzes microarray data and proposes complexes of transcription factors as well as “upstream” key signaling molecules that master the observed gene expression profile. The method utilizes data from three databases (TRANSFAC® (Matys et al., 2006), TRANSPATH® (Krull et al., 2006) and HumanPSD http://www.biobase-international.com/) and integrates two programs: 1) Composite Module Analyst (CMA) analyzes 5'-upstream regions of co-expressed genes and applies a genetic algorithm to reveal composite modules (CMs) consisting of co-occurring single TF binding sites and composite elements (Waleev, at al., 2006; Kel et al., 2006a;); 2) ArrayAnalyzer (Kel et al., 2006b) is a fast network search engine that analyzes signal transduction networks controlling the activities of the corresponding TFs and seeks key molecules responsible for the observed concerted gene activation.

In the Figure 1 we show the user interface of ExPlain system and present the results of applying the system to a set of microarray data on a skin disease. A set of 150 promoters of differentially expressed genes in human fibroblasts of the patients having the skin disease has been compared to the set of 300 promoters of genes that did not showed any significant change of expression. Site frequency analysis showed that promoters of differentially expressed genes have significantly higher frequency of sites for such transcription factors as NF-kappaB, IRF-1, EGR-2 and some others (Figure 1a). Analysis of composite modules has
revealed a highly significant combination of single matrices and matrix pairs that include also matrixes for such factors as AML, OCT and pairs: SP-1/ERG-1 and AP-1/OCT (Figure 1b). This composite promoter model was able to discriminate more than 60% of the differentially expressed promoters from the background promoters (Figure 1c). Finally, the analysis of the signal transduction pathways upstream of these transcription factors helps to identify several potential key molecules such as the ActR-II, which is an important factor of the Atrophin-1 (DRPLA) pathway (Figure 1d).

Figure 1: UI of ExPlain with the results of analysis of site overrepresentation in promoters of genes differentially regulated in a human skin disease: a) promoter model; b) the histogram of the corresponding promoter composite score for differentially expressed genes (red) versus non-changed genes (blue) and c) the identified key-node molecule, ActR-II, with the corresponding signaling network leading to the regulation of activity of the transcription factors found in the promoter model.

References


