BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 14 2010, pages 1802–1803
doi:10.1093/bioinformatics/btq277

Data and text mining

Advance Access publication May 27, 2010

Integrative platform to translate gene sets to networks

Marko Laakso and Sampsa Hautaniemi*

Computational Systems Biology Laboratory, Institute of Biomedicine and Genome-Scale Biology Program, University of Helsinki, PO Box 63, 00014 University of Helsinki, Finland

Associate Editor: John Quackenbush

ABSTRACT

Summary: We have implemented a computational platform (Moksiskaan) that integrates pathway, protein–protein interaction, genome and literature mining data to result in comprehensive networks for a list of genes or proteins. Moksiskaan is able to generate hypothetical pathways for these genes or proteins as well as estimate their activation statuses using regulation information in pathway repositories. An automatically generated result document provides a detailed description of the query genes, biological processes and drug targets. Moksiskaan networks can be downloaded to Cytoscape for further analysis. To demonstrate the utility of Moksiskaan, we use gene microarray and clinical data from >200 glioblastoma multiforme primary tumor samples and translate the resulting set of 124 survival-associated genes to a network.

Availability and Implementation: Moksiskaan and user guide are freely available under GNU General Public License at http://csbi.ltdk.helsinki.fi/moksiskaan/

Contact: Sampsa.Hautaniemi@Helsinki.FI

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on January 28, 2010; revised on April 29, 2010; accepted on May 24, 2010

1 INTRODUCTION

High-throughput technologies, such as gene microarrays, are effective in profiling gene expressions genome-wide and have become standard means in biomedicine. Consequently, a multitude of computational methods have been successfully applied to high-throughput data resulting in the discovery of a large number of disease-associated genes. An unresolved issue to date, however, is how to interpret these results, such as gene lists, in biological context. A popular solution is to apply statistical methods to identify Gene Ontology categories or pathways that are associated with the differentially expressed genes (DEGs), such as SPIA (Tarca *et al.*, 2008), SubpathwayMiner (Li *et al.*, 2009) and KOBAS (Wu *et al.*, 2006). These methods typically use a pathway repository, such as KEGG (Kanehisa *et al.*, 2009), to assess whether a pathway is significantly influenced by the DEGs and produce a *P*-value or a score for each pathway.

Many pathways crosstalk, i.e. two or more pathways influence each other. Thus, a gene may belong to a number of canonical pathways or be influenced by many canonical pathways. Pathway analysis methods that do not take crosstalking into account may not be able to produce a comprehensive view on the impact of the DEGs to pathways. In order to address the crosstalk issue, we have implemented a computational platform called Moksiskaan that integrates data from KEGG pathway and drug databases (Kanehisa *et al.*, 2009), cPath (Cerami *et al.*, 2006)-based Pathway Commons repository of pathways (http://www.pathwaycommons.org/pc/), Ensembl genome database (Hubbard *et al.*, 2009), PINA protein–protein interaction database (Wu *et al.*, 2009) and SNPs3D literature mining database (Yue *et al.*, 2006) to translate gene or protein lists to networks. Moksiskaan is able to (i) rapidly give an overview of the biological functions of a list of genes, such as DEGs; and (ii) form hypothetical pathways consisting of interconnections between the genes. It can also be used to suggest drugs to target proteins. Moksiskaan-generated networks can be viewed as PDFs or downloaded to Cytoscape (Cline *et al.*, 2007) for further analysis.

*To whom correspondence should be addressed.

2 APPROACH

Moksiskaan provides a flexible and powerful schema to store and query elements of an abstract and extensible data type structure *bioentity* that currently represents genes, transcripts, proteins, drugs and pathways as bioentity types. The connections between bioentity types are strongly typed. For instance, a *protein coding gene* link is used to connect proteins to their source genes only.

Moksiskaan can be used in four different modes. The first mode ('connected') constructs a comprehensive network for a list of genes (or other bioentities) by using all imported pathways to identify known regulations between the genes in the list. Connections between these genes are visualized along intermediate genes controlled with a user-defined parameter n, which describes the maximum number of genes between two genes in the list. For example, if $n=2$ and the gene list contains *Sos* and *MEK* as DEGs in the ErbB signaling pathway, also their intermediate genes *Ras* and *Raf* are visualized in the network. The coloring of *Sos* and *MEK* in the resulting network corresponds to their observed expression value, whereas *Ras* and *Raf* expression values are estimated from the network using Boolean logic. In this example, if *Sos* is upregulated, also *Ras* and *Raf* are predicted to be upregulated. Finally, topology of the network is pruned using expression data on all DEGs so that only edges that do not conflict with the observed states are retained. Orphan genes with no neighbors are removed from the network. The other three modes are: 'up' (only genes that are in upstream are searched), 'down' (only genes that are in downstream are searched) and 'both' (genes are searched using the modes 'up' and 'down').

3 METHODS

Moksiskaan runs on the Anduril workflow framework (freely available at http://csbi.ltdk.helsinki.fi/anduril/) and has been implemented using

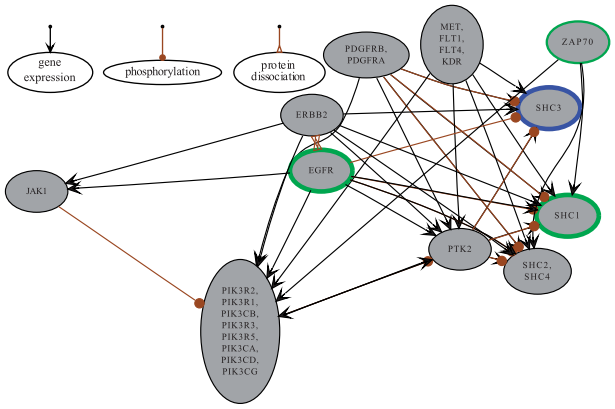

Fig. 1. A snapshot of a Moksiskaan network constructed with 124 survival-associated genes. Edge colors and end point shapes distinguish between the connection types as described in the upper left corner. Green and blue borders refer to up- and downregulated genes (absolute fold-change >2), respectively. Measured gene expressions are shown with bold borders while predictions made by Moksiskaan are thin.

the Hibernate framework, which simplifies the data accession (Fevre *et al.*, 2007). Hibernate enables a separation between the database engine (PostgreSQL) and the code, and provides a single point of maintenance for the schema that is then propagated to:

- SQL statements responsible for the database construction,
- Entity-relationship (ER) diagrams,
- HTML-based descriptions of the relations and their attributes, and
- Java classes that are used to represent the data.

To demonstrate Moksiskaan, we used The Cancer Genome Atlas dataset on glioblastoma multiforme, which is the most aggressive form of brain cancers (McLendon *et al.*, 2008). We downloaded the RMA normalized gene expression (Affymetrix) data for 228 tumor samples belonging to 198 patients along with clinical data, as well as expression profiles for 10 normal tissue samples that were used as controls (McLendon *et al.*, 2008). We then used Kaplan–Meier method with log-rank test to identify survival-associated genes. This resulted in 124 genes having a P-value <0.01, which were used as an input to Moksiskaan. We used 'down' mode ($n=1$) to identify downstream genes that the survival-associated genes regulate. Protein–protein interactions were disabled to reduce network complexity for visualization. The resulting network figure is presented in Supplementary Material, and a zoom-in to the network in Figure 1.

4 DISCUSSION

In the network for 124 survival-associated genes, the genes *EGFR* and *ErbB2* are linked with a brown line with a triangle, which denotes a protein complex (Fig. 1). Indeed, the genes *EGFR* and *ErbB2* encode receptors that can form a heterodimer, which activates downstream signaling affecting various phenotypes, such as cell growth and cell motility (Yarden and Shilo, 2007). The Moksiskaan network further suggests that the *EGFR* and *ErbB2* complex activates the PI3K pathway directly (a node with eight PIK3 genes) and via JAK1. Activation of the PI3K pathway directly by EGFR and ErbB2 is well-known (Liu *et al.*, 2009; Yarden and Shilo, 2007). However, JAK1 has been mainly associated with the immune system (Igaz *et al.*, 2001) and is not considered as a part of the canonical EGFR pathway (Yarden and Shilo, 2007). Thus, the latter connection between EGFR/ErbB2 and PI3K pathway could have remained unnoticed without a comprehensive network approach. The drug data integration together with Gene Ontology analysis, protein interactions and gene annotations are provided in Supplementary Material.

In summary, we have designed and implemented a platform that allows translation of gene lists to comprehensive networks by integrating pathway, protein–protein interaction, literature mining and genome data. This integrated set of regulatory connections allows identification of novel and experimentally testable hypotheses. The networks generated by the Moksiskaan platform can be downloaded to Cytoscape for editing and advanced analysis.

ACKNOWLEDGEMENT

We are grateful to Riku Louhimo for his help with the Moksiskaan user guide and web site, Minna Miettinen for implementing GSEA, Kristian Ovaska for suggestions to the database schema, Jukka Westermarck and Henk Stunnenberg for helpful discussions, and Tiia Pelkonen for proofreading.

Funding: Academy of Finland (projects 125826 and 128416); The Sigrid Jusélius Foundation, Biocentrum Helsinki and Finnish Graduate School in Computational Sciences (FICS).

Conflict of Interest: none declared.

REFERENCES

Cerami,E. *et al.* (2006) cPath: open source software for collecting, storing, and querying biological pathways. *BMC bioinformatics*, **7**, 497.

Cline,M. *et al.* (2007) Integration of biological networks and gene expression data using cytoscape. *Nat. Protoc.*, **2**, 2366.

Fevre,F. *et al.* (2007) Cyclone: java-based querying and computing with Pathway/Genome databases. *Bioinformatics*, **23**, 1299–1300.

Hubbard,T.J.P. *et al.* (2009) Ensembl 2009. *Nucleic Acids Res.*, **37**, D690–D697.

Igaz,P. *et al.* (2001) Biological and clinical significance of the JAK-STAT pathway; lessons from knockout mice. *Inflamm. Res.*, **50**, 435–441.

Kanehisa,M. *et al.* (2009) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.*, **38**, D355–D360.

Li,C. *et al.* (2009) SubpathwayMiner: a software package for flexible identification of pathways. *Nucleic Acids Res.*, **37**, e131.

Liu,P. *et al.* (2009) Targeting the phosphoinositide 3-kinase pathway in cancer. *Nat. Rev. Drug Discov.*, **8**, 627–644.

McLendon,R. *et al.* (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061–1068.

Tarca,A.L. *et al.* (2008) A novel signaling pathway impact analysis (SPIA). *Bioinformatics*, **25**, 75–82.

Wu,J. *et al.* (2006) KOBAS server: a web-based platform for automated annotation and pathway identification. *Nucleic Acids Res.*, **34**, W720–W724.

Wu,J. *et al.* (2009) Integrated network analysis platform for protein-protein interactions. *Nature Met.*, **6**, 75–77.

Yarden,Y. and Shilo,B. (2007) Snapshot: EGFR signaling pathway. *Cell*, **131**, 1018.

Yue,P. *et al.* (2006) SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics*, **7**, 166.