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 context. A popular solution is to apply statistical methods to identify known regulations between the genes in the list. 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’).

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
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 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


