EMBnet.journal
Bioinformatics in Action

Theofilatos et al. (2016) EMBnet.journal 22, e871
http://dx.doi.org/10.14806/ej.22.0.871

Reports

InSyBio BioNets, an efficient tool for network-based biomarker discovery

Konstantinos Theofilatos[1,✉], Christos Dimitrakopoulos[1,✉], Christos Alexakos[1], Aigli Korfiati[1], Spiros Likothanassis[1], Seferina Mavroudi[1]

[1]InSyBio Ltd, London, UK

Abstract

Biological networks have been widely used in systems biology in order to model the complex interactions of molecular players such as proteins, genes, mRNAs, non-coding RNAs and others. However, most of the current methods for biomarker discovery do not use biological networks, but just deploy simple statistical methods to identify differentially expressed genes and gene products. In the present paper, we present InSyBio BioNets, which is a cloud-based web platform offering a unique biomarker discovery pipeline, which combines differential expression analysis and a method for comparing biological networks to identify biomarkers efficiently. As a case study, InSyBio BioNets was applied to a Parkinson dataset of gene expression measurements and outperformed a standard statistical approach by recovering a more compact and informative set of biomarkers.

Introduction

The execution of complex biological processes requires the precise interaction and regulation of thousands of molecules. These interactions can be modeled as networks, which typically consider molecular components within a cell as nodes and their direct or indirect interactions as edges. Network representation enables data integration from a wide range of studies, including protein-protein interaction (PPI) and gene expression measurements, into a single framework. The analysis of these networks can aid in understanding the disease mechanisms, but it has not been successfully linked to clinical applications until now.

Biomarker discovery is a field currently dominated by statistical analysis on the actual expression values of genes or quantitative values of proteins to identify the cellular molecules, which differ significantly in experiments between biological or clinical conditions (*i.e.* disease vs control samples). The results of this approach present certain drawbacks including the high number of discovered biomarkers that require experimental validation as well as the high number of false positives. Moreover, standard statistical approaches are prone to identify biomarkers descriptive of the disease's outcome and not of its cause. For this reason, the current trend in biomarker discovery is to detect biomarkers by comparing biological networks. Biological network metrics are more stable to changes between biological conditions compared to absolute gene expression differences and can be associated with the causes of disease mechanisms.

InSyBio BioNets is a tool providing a unique biomarker discovery pipeline that overcomes the aforementioned constraints of existing biomarker discovery methods capitalizing on biological networks' comparison. In specific, InSyBio BioNets offers a novel systems medicine approach, which provides biomarker sets with increased predictive accuracy. The proposed pipeline is offered through a flexible semi-automated web-based analysis enabling the users to easily navigate through its different steps, while also being able to use their own algorithms and methods. This can be accomplished either by selecting among a variety of algorithms offered through InSyBio BioNets web interface or by downloading intermediate results, processing them locally and uploading results to continue the analysis through the web interface.

In addition to the biomarker discovery pipeline, InSyBio BioNets provides a set of tools for the construction, preprocessing, meta-analysis and visualisation of biological networks and it supports tools for parsing and creating gene expression files to enable the construction and analysis of gene co-expression networks. Moreover, users can fast analyse large biological networks and gene expression files using the tool's user-friendly job management mechanism. Regarding the uncovered biomarkers, users have access to informative biomarker reports, which provide information from publicly available databases and from InSyBio Interact tool's PPI repository. These reports also include information about the prior knowledge linking biomarkers to diseases.

Article history
Received: 25 November 2016
Published: 1 December 2016

Theofilatos et al. (2016) EMBnet.journal 22, e871
http://dx.doi.org/10.14806/ej.22.0.871

Methodologies

Gene expression data parsing, preprocessing and analysis

InSyBio BioNets offers a set of tools for handling, preprocessing and analysing gene expression data in order to construct gene co-expression networks. It supports the universally accepted format for gene expression data named SOFT (Simple Omnibus Format in Text) supported also by Gene Expression Omnibus. InSyBio BioNets SOFT parsing includes the following preprocessing steps a) logarithmic normalisation, b) missing values estimation and c) filtering based on average expression or expression variance. The different experimental states (conditions) defined in the SOFT file are automatically recognised and a gene expression tab delimited file is constructed for each state. Gene expression files can also be used to generate weighted gene co-expression networks. A weighted edge is added to the network if the metric among the expression profiles of the two nodes adjacent to this edge exceeds a predefined threshold (which is automatically derived for each node from the dataset). InSyBio BioNets also provides a network-clustering tool that supports state of the art clustering algorithms, a network analysis tool to compute network metrics and components as well as various Cytoscape-based network visualisation tools. Most algorithms included in the InSyBio BioNets tool were implemented using Python programming language and standard python libraries such as Biopython.

Biomarker discovery pipeline based on network comparison

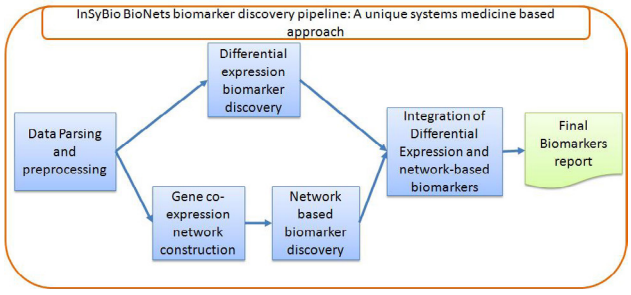

Figure 1. InSyBio BioNets biomarker discovery pipeline.

The proposed pipeline (Figure 1) is a five step procedure: I) parse and preprocess transcriptomics experiments, II) perform differential expression analysis to uncover differentially expressed biomarkers, III) construct gene-co expression networks for each one of the biological conditions examined in the transcriptomics experiments, IV) compare biological networks to uncover network-based biomarkers and V) combine differentially expressed and network-based biomarkers by a confidence score. For the networks comparison (step IV), users are able to select one of the offered network metrics (degree centrality, clustering coefficient or PageRank centrality with the PageRank centrality being the default option) and the tool detects the network's nodes for which the selected metric is significantly altered. Network-based biomarkers are merged with the differentially expressed ones (step V) by intersecting the two sets and by computing a combined confidence score. The final biomarker list is annotated with information from public available repositories including OMIM, DisGeNet, Genecards and other InSyBio Suite Tools (InSyBio Interact and InSyBio ncRNAseq). In step (II), differential expression is performed by using the Wilcoxon rank sum test and users can state a P-value threshold. Bonferroni corrections are applied to reduce the number of false positive predictions.

Results and conclusions

As a case study, InSyBio BioNets was used to uncover Parkinson Disease (PD) biomarkers from gene expression measurements in blood samples. PD remains a disease whose diagnosis is based on clinically detectable symptoms. When these symptoms arise, it is quite late for the effective patients' treatment. PD has been attributed to genetic and environmental causes in the relevant scientific literature. However, until now, there exist only a few early stage diagnostic tests of limited predictive accuracy. We analysed a Gene Expression Omnibus dataset (GDS2519) which has been constructed by microarray experiments on blood samples of 50 early stage Parkinson Disease patients, 22 control patients and 33 other neurodegenerative disease control patients. We used InSyBio BioNets to detect biomarkers for PD and we compared our results with the Wilcoxon rank sum test (used by Scherzer *et al.*, 2016) which detects biomarkers based on differential expression and ranks them based on their P-values. InSyBio BioNets uncovered a more specific set of biomarkers with increased predictive accuracy for PD-related genes (Table 1). These biomarkers were further reduced to five (HNRNPA3, ZFC3H1, SSR1, ATRX, SNCA) without the loss of classification accuracy when an ensemble genetic algorithm/SVM method was applied for feature selection.

Table 1. InSyBio BioNets vs. a standard differential expression analysis for identifying PD biomarkers from transcriptomics experiments.

Method	#PD Biomarkers	Precision using genes associated with PD from DisGeneNet DB	Classifiers Accuracy (SVM classifiers used) with 10-fold cross validation
Standard Approach (Wilcoxon Rank Sum method)	834	7.47%	96.4%
InSyBio BioNets	24	12.5%	100%

Availability

InSyBio BioNets is one of the tools included in the integrated bioinformatics web platform of InSyBio named InSyBio Suite. A Demo version of InSyBio BioNets is freely available at http://demo.insybio.com. A free evaluation version includes a one-month free license and it can be purchased by sending an email at info@insybio.com. To purchase the commercial version of InSyBio BioNets users can contact sales@insybio.com for the detailed quota and information. InSyBio is registered with the Information Commissioner's Office under registration reference number ZA182885 to provide data security.

Acknowledgements

InSyBio participates in the NBG Business Seeds program by the National Bank of Greece.

References

1. Theofilatos, KA, Likothanassis S, Mavroudi S (2015) Quo vadis computational analysis of PPI data or why the future isn't here yet. *Frontiers in genetics* **6**, 289. http://dx.doi.org/10.3389/fgene.2015.00289
2. Tong H, Faloutsos C, Pan JY (2006) Fast random walk with restart and its applications. *Sixth International Conference on Data Mining (ICDM'06)*, Hong Kong, pp. 613-622 https://doi.org/10.1109/ICDM.2006.70
3. Adler CH, Beach TG, Hentz JG, Shill HA, Caviness JN, Driver-Dunckley E, Dugger BN *et al.* (2014) Low clinical diagnostic accuracy of early vs advanced Parkinson disease Clinicopathologic study. *Neurology* **83**(5), 406-412. https://dx.doi.org/10.1212/WNL.0000000000000641
4. Scherzer CR, Eklund AC, Morse LJ, Liao Z *et al.* (2007) Molecular markers of early Parkinson's disease based on gene expression in blood. *Proc Natl Acad Sci USA*, **104**(3), 955-960. https://dx.doi.org/10.1073/pnas.0610204104