Bioinformatics Advance Access published April 8, 2014

HyperModules: identifying clinically and phenotypically significant network modules with disease mutations for biomarker discovery

Alvin Leung [1], Gary D. Bader [1,*] and Jüri Reimand [1,*]

[1]The Donnelly Centre, University of Toronto

Associate Editor: Prof. Alfonso Valencia

ABSTRACT

Summary: Correlating disease mutations with clinical and phenotypic information such as drug response or patient survival is an important goal of personalised cancer genomics and an important first step in biomarker discovery. HyperModules is a network search algorithm that finds frequently mutated gene modules with significant clinical or phenotypic signatures from biomolecular interaction networks.

Availability: HyperModules is available in Cytoscape App Store and as a command line tool at www.baderlab.org/Sofware/HyperModules.

Contact: Juri.Reimand@utoronto.ca, Gary.Bader@utoronto.ca

1 INTRODUCTION

Establishing functional links between genetic variation and human disease is a key goal of cancer genome sequencing (Gonzalez-Perez et al., 2013) and genome-wide association studies (Hardy and Singleton, 2009). Complex diseases like cancer are often driven by infrequent changes in multiple genes in pathways (Vogelstein et al., 2013). Network analysis helps interpret mutations in systems context and find disease genes, pathways, and biomarkers for precision medicine (Barabasi et al., 2011).

Discovery of modules (sub-networks) in biological networks helps isolate systems with disease-related properties and reduces interactome complexity. A growing number of methods is available for this purpose. A landmark paper combines gene expression signatures with protein-protein interactions (PPI) to find predictive modules of cancer outcome (Chuang et al., 2007). The NETBAG method studies genetic associations and copy number variants to find autism-related modules (Gilman et al., 2011). HotNet detects frequently mutated pathways in networks (Vandin et al., 2011, 2012). Net-Cox builds prognostic cancer signatures in network analysis of gene expression data (Zhang et al., 2013). The Reactome FI Cytoscape plugin uncovers prognostic gene modules from networks and gene expression data (Wu and Stein, 2012). Network based stratification predicts tumour subtypes from mutations in network regions (Hofree et al., 2013). Such modules maximize a feature of genes such as differential expression, disease mutation frequency, or enrichment of interactions.

As clinical profiles of patients are increasingly available in cancer genomics efforts such as the TCGA pan-cancer project (Weinstein et al., 2013), new methods are needed to discover multivariate biomarkers in networks. We recently analyzed cancer mutations in phosphorylation signalling and found that kinase-substrate networks are informative of patient survival and therapy response (Reimand and Bader, 2013; Reimand et al., 2013). In particular, we found network modules with rare mutations in ovarian cancer patients with improved prognosis. We created the HyperModules method to systematically discover clinically correlated modules from gene and protein networks (Reimand and Bader, 2013) based on our earlier work on functional sub-network discovery (Reimand et al., 2008; Altmae et al., 2012). Here we present the previously unavailable software in open-source Java as a command line tool for automated work and a Cytoscape app for interactive graphical analysis.

2 SOFTWARE

HyperModules assumes that clinically informative mutations of complex disease occur in systems of closely interacting genes. The greedy network search algorithm focuses on a local network area, defined by a central seed node (a mutated gene) and its surrounding subnetwork. All mutated genes are sequentially considered as seeds in module discovery. Search starts from the seed and grows the module towards increased benefit by adding connected genes that best improve clinical significance. This objective is driven by statistical tests where patients defined by the module are compared to other patients. Categorical clinical variables are studied with Fisher's exact test and survival times with log-rank test. Cox regression is currently not supported, however we plan to add this feature in the future. To establish statistical significance of detected modules, we build a null distribution by searching networks with permuted gene names. Each module of the true network is quantified with an empirical p-value reflecting the fraction of seed-specific modules from shuffled networks exceeding the significance of the true module. This removes artefacts of the greedy strategy and corrects for topological features such as highly connected nodes.

The analysis pipeline is outlined in Figure 1. Interaction networks are loaded into Cytoscape using standard features (Shannon et al., 2003). HyperModules requires gene mutations and patient clinical info in two tables. The user selects type of clinical analysis and columns in data table (survival time or variable such as tumor relapse). Survival analysis requires follow-up time and vital status of patients. Detected modules are studied further with

*to whom correspondence should be addressed

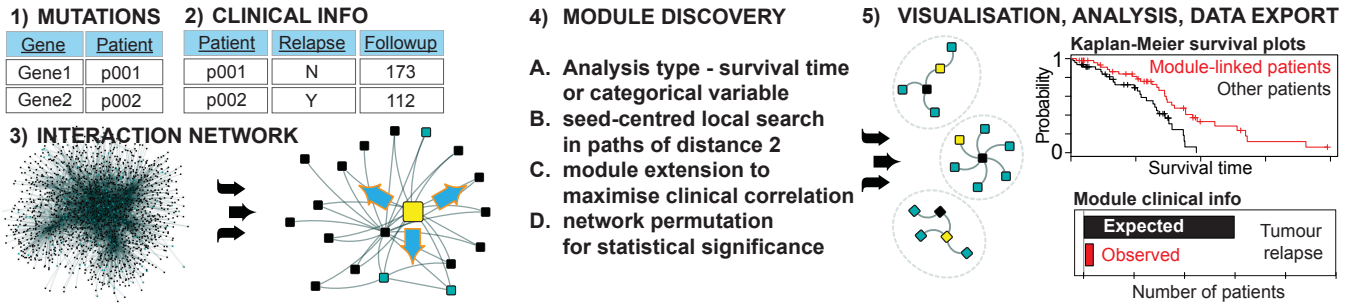

Fig. 1. HyperModules requires three inputs - (1) mutated genes in patients, (2) patient clinical information, and (3) protein or gene network. Search is performed for all mutated genes as seeds (4). Network visualisation, clinical variable statistics and data export facilitate further analysis (5).

network visualisation, survival curves, and data export. We tested HyperModules on protein networks to find survival modules with cancer mutations. We extracted three human PPI networks of variable size from iRefWeb (Turner et al., 2010), and five cancer mutation datasets from ICGC portal v.12 (Supplementary Figure 1). For example, network analysis of 30,000 interactions with 121 liver cancer patients, 686 mutated genes, and 10,000 permutations takes 10 min on a 8-core computer with 16 GB RAM. HyperModules is thus applicable to a range of networks and mutation datasets.

3 EXAMPLE ANALYSIS

An example dataset is provided in Supplementary File 1. It comprises 183 ovarian cancer patients from the TCGA study (Cancer Genome Atlas Research Network, 2008) and the network of 4,823 kinase-substrate interactions from our earlier study (Reimand and Bader, 2013). The ovarian cancer mutations are restricted to 163 proteins with single nucleotide variants affecting protein phosphorylation sites or kinase domains. Two sets of modules were computed with 10,000 network permutations and are shown in Supplementary Figures 2-3. First, the search for survival correlations in the kinase-substrate network with log-rank test identified 19 modules where associated patients have significantly different survival rates compared to other patients in the cohort (empirical $p \leq 0.05$). Second, the categorical variable search with Fisher's exact test revealed 5 modules with significant enrichment of alive patients (empirical $p \leq 0.05$). The modules are also summarised in Supplementary Table 1.

4 DISCUSSION

HyperModules is a biological network-mining algorithm that reveals modules of interacting genes with clinically informative disease mutations. Diverse biomolecular interaction networks can be analysed, including PPI networks, gene regulatory networks, and curated biological pathways. Disease mutations are also broadly defined. While we initially studied cancer point mutations, other types of alterations such as copy number and gene expression changes can be used. HyperModules finds correlations with groups of genes where mutations may be infrequent but the signature strengthens through network integration. Such modules are not often directly usable as biomarkers due to small sample size, however we believe that our approach helps discover genes and pathways as potential multivariate biomarkers for further experiments.

ACKNOWLEDGEMENTS

We thank Jason Montojo and Harold Rodriguez for Cytoscape help. This work was supported by NRNB (US NIH NCRR grant P41 GM103504) and Google (Google Summer of Code 2013).

REFERENCES

Altmae,S, Reimand,J, Hovatta,O, Zhang,P, Kere,J, et al. (2012) Research resource: interactome of human embryo implantation: identification of gene expression pathways, regulation, and integrated regulatory networks. Mol. Endocrinol., **26** (1), 203–217.

Barabasi,AL, Gulbahce,N and Loscalzo,J (2011) Network medicine: a network-based approach to human disease. Nat. Rev. Genet., **12** (1), 56–68.

Cancer Genome Atlas Research Network (2008) Integrated genomic analyses of ovarian carcinoma. Nature, **474**, 609-15.

Chuang,HY, Lee,E, Liu,YT, Lee,D and Ideker,T (2007) Network-based classification of breast cancer metastasis. Mol. Syst. Biol., **3**, 140.

Gilman,SR, Iossifov,I, Levy,D, Ronemus,M, Wigler,M et al. (2011) Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. Neuron, **70** (5), 898–907.

Gonzalez-Perez,A, Mustonen,V, Reva,B, Ritchie,GR, Creixell,P., et al. (2013) Computational approaches to identify functional genetic variants in cancer genomes. Nat. Methods, **10** (8), 723–729.

Hardy,J and Singleton,A (2009) Genomewide association studies and human disease. N. Engl. J. Med., **360** (17), 1759–1768.

Hofree,M, Shen,JP, Carter,H, Gross,A and Ideker,T (2013) Network-based stratification of tumor mutations. Nat. Methods, **10** (11), 1108–1115.

Reimand,J and Bader,GD (2013) Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. Mol. Syst. Biol., **9**, 637.

Reimand,J, Tooming,L, Peterson,H, Adler,P and Vilo,J (2008) GraphWeb: mining heterogeneous biological networks for gene

modules with functional significance. Nucleic Acids Res., **36** (Web Server issue), W452–459.

Reimand,J, Wagih,O and Bader,GD (2013) The mutational landscape of phosphorylation signaling in cancer. Sci Rep, **3**, 2651.

Shannon,P, Markiel,A, Ozier,O, Baliga,NS, Wang,JT, et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res., **13** (11), 2498–2504.

Turner,B, Razick,S, Turinsky,AL, Vlasblom,J, Crowdy,EK, et al. (2010) iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. Database, **2010**, baq023.

Vandin,F, Clay,P, Upfal,E and Raphael,BJ (2012) Discovery of mutated subnetworks associated with clinical data in cancer. Pac Symp Biocomput, **1**, 55–66.

Vandin,F, Upfal,E and Raphael,BJ (2011) Algorithms for detecting significantly mutated pathways in cancer. J. Comput. Biol., **18** (3), 507–522.

Vogelstein,B, Papadopoulos,N, Velculescu,VE, Zhou,S, Diaz,LA et al. (2013) Cancer genome landscapes. Science, **339** (6127), 1546–1558.

Weinstein,JN, Collisson,EA, Mills,GB, Shaw,KR, Ozenberger,BA, et al. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet., **45** (10), 1113–1120.

Wu,G, and Stein,L (2013) A network module-based method for identifying cancer prognostic signatures. Genome Biol., **13**, R112.

Zhang,W, Ota,T, Shridhar,V, Chien,J, Wu,B and Kuang,R (2013) Network-based survival analysis reveals subnetwork signatures for predicting outcomes of ovarian cancer treatment. PLoS Comput. Biol., **9** (3), e1002975.