KeyPathwayMiner - Combining OMICS data and biological networks

Josch Pauling, Nicolas Alcaraz, Alexander Junge, Jan Baumbach
Max Planck Institute for Informatics, Saarbrücken, Germany

jpauling@mpi-inf.mpg.de

Abstract: KeyPathwayMiner is a method for extracting and visualizing disease-specific key pathways. We identify sub-graphs, where most genes are dysregulated in a typical case-control study. Therefore, we extract all maximal connected sub-networks where all but K genes are differentially expressed/methylated/etc. in all but L cases. This model yields a very high interpretability of the results since K and L have real-world implications. We will exemplarily demonstrate KeyPathwayMiner's flexibility by analyzing promoter methylation as well as gene expression assays of complex diseases: Huntington's disease and colorectal cancer, respectively. Here, we identify biologically sound key pathways that highly overlap with known disease-related genes (literature research). Our KeyPathwayMiner implementation uses a combination of fixed-parameter, approximation and heuristic algorithms for tackling the underlying NP-hard problem. It is available as a Cytoscape plugin and has been downloaded and installed ~900 times since its first release in Oct. 2011 (~5x per day). Availability: http://keypathwayminer.mpi-inf.mpg.de

1 Introduction and Overview

While combining networks with OMICS data (known as network enrichment, for instance) is a long-standing problem in computational biology, little attention has been paid to interpretability of the results. We usually seek to identify a densely connected sub-graph in a given PPI network that is highly expressed in a given OMICS data set (typically a transcriptomics study). For complex diseases, such as cancer, gold standard data doesn't exist, i.e. known key pathways with many relevant genes, such that setting the parameters, thresholds, etc. for the underlying combined statistics is tricky and still unsolved. When computing such statistics, we need at least one such parameter that balances network density and correlation in the expression data, even when we neglect modeling the noise

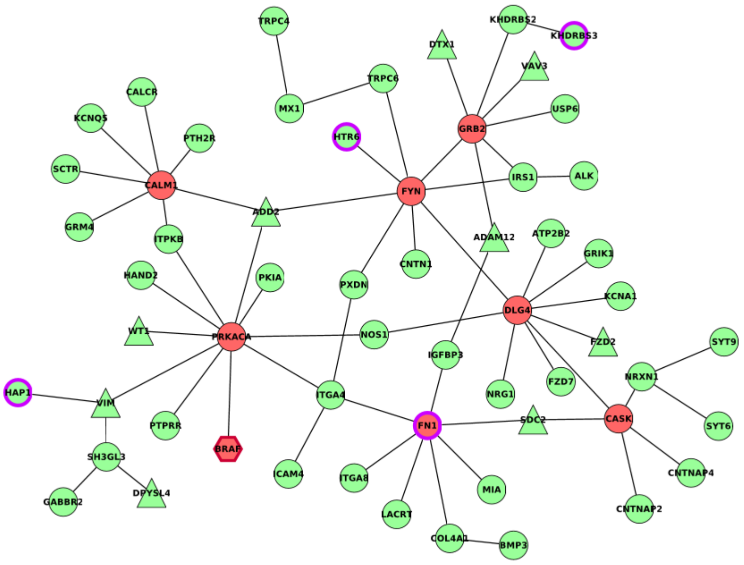

Figure 1: Largest subnetwork found containing the *BRAF* gene for K=8 and L=25. Red nodes represent exception nodes, triangle nodes are hypermethylated genes that also show significant decrease in gene expression levels, and nodes with a purple border are genes with promoters classified as CIMP.

levels in the two data types. We circumvent this problem by providing the end user with an easy-to-interpret model that asks for two parameters with a strong real-world meaning: K and L. KeyPathwayMiner computes all maximal connected sub-networks where all genes but K are expressed/differentially expressed/methylated/active/etc. in all patients but at most L. For the colorectal cancer data set, for instance, we find a 58-genes-key pathway (Figure 1) in the human interactome (approx. 10k proteins, 40k interactions) where all genes but K=8 have a hypermethylation event in the promoter in all 128 patients but at most L=25. In another example we studied Huntington's disease with gene expression data (see Figure 2 for the corresponding key pathway). We applied KeyPathwayMiner to many more data sets and compared it to similar tools obtaining equal or better results (see [AFK$^+$12, BFK$^+$12]). Since our first publication in Oct. 2011, the community downloaded and installed the Cytoscape plugin ~900x (~5x per day).

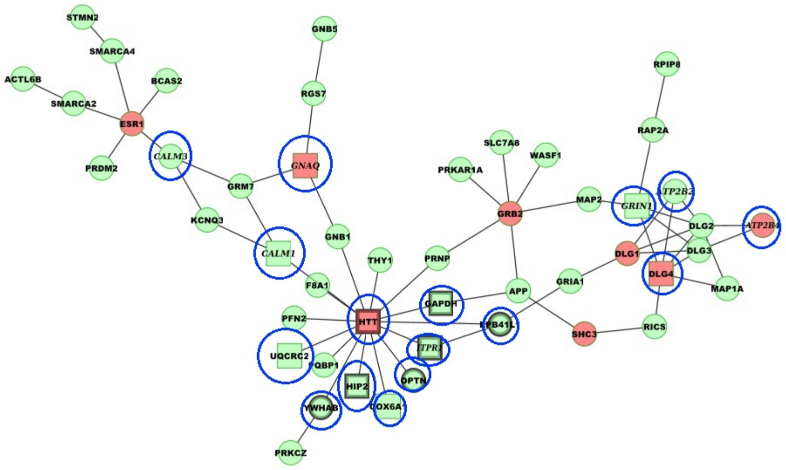

Figure 2: Huntington's disease (HD) key pathway. Here, our KeyPathwayMiner Cytoscape plug-in also used the human interactome network and genome-wide gene expression studies for 38 HD patients (and 32 healthy persons in control group) as input. The illustrated network is the maximal connected sub-network where all genes/proteins but K=8 are differentially expressed in all 38 HD patients but L=6. Red nodes represent exception genes, nodes with blue circles are genes known to be HD-related (from literature).

2 Methods and Model Summary

We provide two slightly varying models for the above introduced problem of finding key pathways:

1. INES: For all genes that have been measured in the case-control study, the profile over all cases is attached. All genes that are not dysregulated in all cases but L are considered ”exception genes”. We find all maximal, connected sub-networks containing at most K such ”exception-genes”.

2. GLONE: This is a slightly modified, alternative model. Now we identify all maximal, connected sub-networks where all but at most K nodes are expressed in all cases but in total (!) at most L, i.e. accumulated over all cases and all nodes in a solution. While INES tends to prefer solutions with many hub nodes as exception genes, GLONE circumvents this potential drawback (see [BFK$^+$12]).

Since the underlying optimization problems are computationally hard, we developed a set of three different algorithmic strategies: an exact fixed-parameter algorithm (INES only, fast for $K < 3$), a greedy approximation (INES only, fast but less accurate for higher values of K and L), as well as two Ant Colony Optimization schemes (INES and GLONE, fast and accurate for medium to high values of K, generally accurate for all tested K and L values).

3 Conclusion

Overall, KeyPathwayMiner tackles the problem of finding biomedically relevant pathways by directly combining biological networks with different types of OMICS data. In contrast to existing methods, we ensure interpretability and usability while still being robust, accurate and fast on real world application cases. For details, please refer to the three corresponding papers: [AKW$^+$11, AFK$^+$12, BFK$^+$12]

References

[AFK$^+$12] Nicolas Alcaraz, Tobias Friedrich, Timo Kötzing, Anton Krohmer, Joachim Mueller, Josch Pauling, and Jan Baumbach. Efficient key pathway mining - Combining networks and OMICS data. *Integr Biol*, 4(7):756–764, 2012.

[AKW$^+$11] Nicolas Alcaraz, Hande Kucuk, Jochen Weile, Anil Wipat, and Jan Baumbach. KeyPathwayMiner - Detecting case-specific biological pathways using expression data. *Internet Mathematics*, 7(4):299–313, 2011.

[BFK$^+$12] Jan Baumbach, Tobias Friedrich, Timo Kötzing, Anton Krohmer, Joachim Müller, and Josch Pauling. Efficient algorithms for extracting biological key pathways with global constraints. In *Proceedings of the fourteenth international conference on Genetic and evolutionary computation conference*, GECCO '12, pages 169–176, New York, NY, USA, 2012. ACM.