CyEVEX: Literature-scale network integration and visualization through Cytoscape

Kai Hakala[1], **Sofie Van Landeghem**[2,3], **Suwisa Kaewphan**[1,4], **Tapio Salakoski**[1,4], **Yves Van de Peer**[2,3], **Filip Ginter**[1]
[1]Department of Information Technology, 20014 University of Turku, Finland
[2]Department of Plant Systems Biology, VIB, Technologiepark 927, 9052 Gent, Belgium
[3]Department of Plant Biotechnology and Bioinformatics, Ghent University, Gent, Belgium
[4]Turku Centre for Computer Science (TUCS), Joukahaisenkatu 3-5, 20520 Turku, Finland

Abstract

EVEX is a literature-scale event extraction resource, publicly available via a web application and as a relational database. In this paper we present CyEVEX, a plug-in which integrates EVEX with the widely used Cytoscape network analysis platform, making the text mining data readily available for integration with experimental data sources and subsequent biological analysis. CyEVEX can populate existing networks with edges corresponding to EVEX events, as well as add new nodes to the network, revealing novel interesting genes and proteins and their relationships within the existing network.

1 Introduction

Information extraction is one of the major research tasks within the BioNLP community, aiming to automatically extract bio-entity associations stated in text and present them to life science researchers. *Event extraction*, a particular formalization of the task, has rapidly gained interest, owing to the resources provided within the BioNLP'09 and '11 Shared Tasks on Event Extraction (Kim et al., 2009; Kim et al., 2011). In the wake of the Shared Tasks, top-ranking systems (Björne et al., 2009; Miwa et al., 2012; McClosky et al., 2011) as well as large-scale datasets obtained by processing PubMed and PubMed Central texts with these systems have been publicly released. In particular, the *EVEX* dataset of Van Landeghem et al. (2011), comprises in its current state of over 34 million biomolecular events among more than 67 million gene/protein name occurrences, extracted from all articles available in the 2011 distribution of PubMed and PubMed Central Open Access section.

The EVEX dataset is stored in a complex relational database which is not intended to be directly queried by end users. Therefore, Van Landeghem et al. (2012) have developed a web interface for the dataset, available at `www.evexdb.org`. This interface is intended for fast intuitive search of events and access to relevant literature. However, it is not suitable for automated text-mining analysis of large gene/protein networks or integration of text-mining data with experimental results.

A particularly wide-spread tool among life science researchers is *Cytoscape* (Smoot et al., 2011), a platform providing functionality for network analysis and visualization supported by a large set of plug-ins for numerous specialized tasks including data integration, clustering, ontology enrichment, filtering, and others. Providing access to EVEX within Cytoscape would thus allow straightforward integration of textual information with other databases and experimental data, analysing the resulting networks using the tools readily available in Cytoscape.

In this paper, we introduce *CyEVEX*, a Cytoscape plug-in which enables the population of large networks with EVEX events in what is essentially a one-click process, not requiring any background in relational databases or text mining, nor local installation of the EVEX database. The Cytoscape plug-in complements and links to the existing web application, the two together providing a comprehensive interface to the EVEX event data specifically targeting end-users. The plug-in is freely available at `www.evexdb.org`.

2 Overview of CyEVEX

The CyEVEX plug-in has been developed to enable straightforward integration of text mining information from EVEX into various networks analyses. Within this context, the nodes of the biomolecular networks are genes, identified through their unique Entrez Gene identifiers[1]. Between these nodes, various types of relations (edges) may exist, forming gene regulatory networks, protein-protein interaction networks, metabolic networks or any other type of molecular network.

The only prior requirement for the application of CyEVEX is that the nodes in the network have Entrez Gene identifiers available, which is typically the case when analysing large experimental networks. Cytoscape also provides various plug-ins that add these identifiers to genes/proteins based on other information, such as their symbol or UniProt ID. As CyEVEX communicates directly with the EVEX web application, the text mining data does not have to be stored locally.

Currently, the plug-in offers two main functions: Populating an existing network with edges corresponding to interactions derived from EVEX, and expanding an existing network with the interaction partners of its genes. Both of the CyEVEX functionalities are available as menu options within Cytoscape and are briefly discussed in the next sections, as well as illustrated in more detail in Sections 3 and 4.

To search for pairwise interactions given an existing network of input genes, the EVEX resource is queried to find biomolecular events between any two of the genes in the network, and subsequently generates pairwise interactions that are translated to new edges. The coarse type of such an interaction is stored in the conventional *interaction* attribute of the Cytoscape edge and is specified as either *regulation*, *indirect regulation* or *binding* (Van Landeghem et al., 2012).

In addition, the original events may contain complex regulatory chains as well as other physical event types. Consequently, the attribute *evex.subtype* provides a more precise classification of the original events, such as *negative regulation of transcription* or *positive regulation of phosphorylation*. Furthermore, details on the affirmative or negative context (e.g. does not regulate) and the speculative context (e.g. may regulate) of the original event are also stored as attributes (*evex.negation* and *evex.speculation*).

Finally, to judge the reliability of an interaction extracted from text, confidence values are automatically derived from the Turku Event Extraction System and represented in the attribute *evex.confidence*. These confidence values are normalized classification scores, ranking events from least to most reliable, allowing for selection of high-precision events (Van Landeghem et al., 2012). They enable various filtering and visualization possibilities through built-in functionality of Cytoscape (cf. Section 3).

The second functionality offers node expansion of a certain gene in a given network by searching all pairwise interactions for this particular gene. If the retrieved interaction partners are not present in the network, new nodes are created accordingly and connected to the original network using similar edges as described above.

By design, CyEVEX works with genes and proteins as nodes, given by Entrez Gene identifiers. When a certain use-case requires analysis of interologs or regulogs (i.e. interactions and regulatory relationships derived through homology), a family-based generalization of the nodes in the network is possible. This is supported by the existing integration of event data with gene families from HomoloGene, Ensembl, and Ensembl Genomes (Van Landeghem et al., 2011; Van Landeghem et al., 2012). When this functionality is selected, edges reflect associations that are observed among any genes belonging to the same families as the genes in the input network. The family-based edges are categorized by the resource used to define the families, thus making them easily distinguishable from gene-specific interactions. When expanding a node over the family-based generalizations, only interacting families with a gene from the same organism are included in the results. The newly created nodes are then identified by the Entrez Gene ID of this organism-specific gene from the resulting gene family. This allows for species-specific filtering of results while still incorporating relevant information from closely related

[1]The data of linking events from EVEX to gene normalization results, is currently under review.

species.

While the networks generated through CyEVEX are already extensively annotated with interaction data and contextual information, a menu option also allows linking out to the EVEX website. This functionality opens a page with detailed information on the selected event, showing its explicit (formal) structure as well as the source texts supporting the statement, linking to the original PubMed abstract or PubMed Central full-text article. Furthermore, the site allows exploratory browsing of related genes and events. When this functionality is selected for edges in the network that were not created by CyEVEX, but originate from another data source, the website displays the main search page which lists all textual interactions for the given gene or gene pair, allowing further validation of the external data through EVEX.

Various data sources, notably PPI databases such as STRING (Szklarczyk et al., 2011) and BioGRID (Winter et al., 2011) provide similar tools for populating Cytoscape networks. However, they sometimes lack pointers to experimental evidence, or merely link to full-text articles, preventing a quick manual evaluation. The fine-grained interaction descriptions provided by CyEVEX can thus provide additional information on top of these existing efforts, as the event-based visualizations in CyEVEX are fully compatible with the PPI-based view.

3 Use case: Constructing networks from seed genes

Previously, the EVEX resource has already been proven useful for hypothesis generation. A recent study involved finding candidate regulators for a set of *Escherichia coli* genes (Kaewphan et al., 2012), accomplished by integrating text mining data with microarray-based co-expression networks. The text mining data was sourced from EVEX, starting with a list of 14 key genes (or 'seed genes') that are known to influence NADP(H)-metabolism, and subsequently retrieving their candidate regulators and binding partners. Through the final integrated network, a set of interesting candidate regulators which were involved in specific triangular patterns could be identified.

In this section, we demonstrate how CyEVEX offers similar functionality by enabling the construction of a network, starting from a small set of seed genes and expanding it to a larger network. This is applied to the original set of 14 seed genes, for which interactions between them are sought in the first step (Figure 1). Most genes are interconnected, but a few seed genes can be considered as outliers in this restrictive network. To connect these isolated nodes, the node expansion functionality of CyEVEX is applied, resulting in the discovery of indirect connections via common interaction partners.

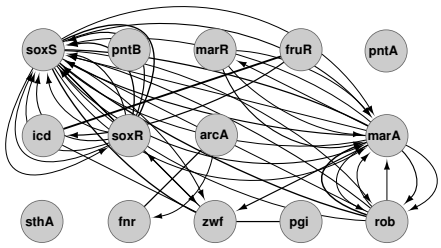

Figure 1: Direct interactions found for the 14 seed genes using gene families from Ensembl Genomes. Plain lines represent binding events, and arrows denote regulatory control. Note that multiple regulation events can be seen between two genes, as their subtypes may differ, e.g. *positive regulation* and *regulation of transcription*. Line widths reflect the confidence values.

Applying this method to all seed genes, a network of 155 nodes and 347 edges is constructed, in about 10 seconds of run-time. In contrast to the original study where manual evaluation was applied to confirm the retrieved edges (Kaewphan et al., 2012), with CyEVEX we can automatically apply certain filters to prune the network. To this end, excessive edges are deleted according to the event confidence with threshold values of -2.0 for regulation and -1.0 for binding events, and regulations are limited to those in which the seed genes appear as targets, as we are only interested in upstream regulation of the seed genes.

The resulting network can be used as a hypothetical gene regulatory network or can be integrated with other data sources, in this case with an *E. coli* gene expression network derived from microarray data. Triangular patterns such as those in the study of Kaewphan et al. (2012) can be found, retrieving 35 out of the 41 previously identified candidate genes (Figure 2). Further, the network also

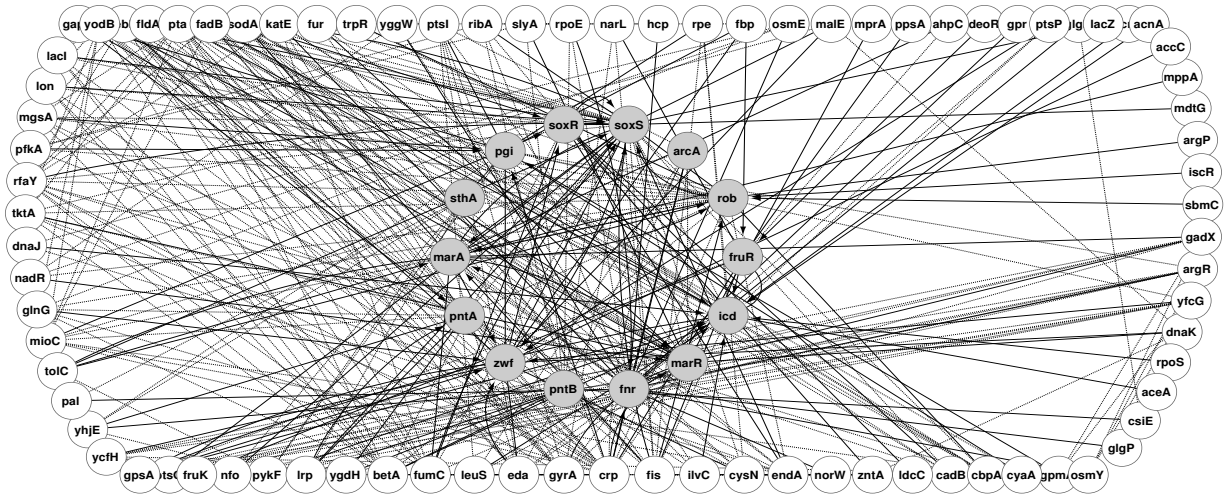

Figure 2: Overlayed network containing both co-expression and EVEX edges.

contains nodes which were previously annotated to be false positives. These findings suggest that caution must be taken when filtering out edges by their confidence, and that finding a suitable confidence threshold is crucial. However, the ideal threshold will vary between different use cases and thus needs some project-specific tuning. For this purpose, limited manual evaluation efforts may be conducted by using the link out functionality to the textual details of the extracted events.

4 Use case: Analysing large-scale networks

The small-scale networks described in the previous section can still be analysed and interpreted manually. However, this manual analysis becomes infeasible for larger networks that include a few thousand genes or represent the whole interactome of a genome. There is thus a need for a platform that can analyse large-scale networks and present a smaller set of meaningful results that can be readily interpreted by human. For this task, Cytoscape is a widely used tool as many excellent plug-ins have specifically been developed for such network analysis and visualization.

In this section, we demonstrate the compatibility of CyEVEX with other publicly available Cytoscape plug-ins to perform large-scale network analysis. Specifically, we focus on motif clustering and functional annotation by using two external plug-ins: CyClus3D and ClueGO. The CyClus3D plug-in identifies motif clusters in integrated networks from different data sources by using a 3-dimensional spectral clustering algorithm (Audenaert et al., 2011), while ClueGO provides functional enrichment of gene clusters by extracting non-redundant biological information directly from multiple ontology resources such as Gene Ontology and KEGG (Bindea et al., 2009).

Building upon the previously described use case of NADP(H) metabolism in *E. coli*, we now retrieve the full text-mining network of *E. coli* genes based on the Ensembl Genomes family generalization via CyEVEX, rather than restricting the search to the 14 seed genes. The retrieval of 13393 edges among the 3312 genes took approximately one minute of runtime. We further integrate this text-mining data with co-expression data relevant to the *E. coli* genome. Next, the integrated network is clustered to illustrate the network motifs of binding and gene expression events under NADP(H) perturbation by CyClus3D. Such network motifs consist of undirected associations, with binding interactions originating from text-mining and co-expression correlations from the microarray data. The resolution and cluster size parameters are set to their recommended default values: 0.5 and 4 respectively. The clustering algorithm identifies 391 clusters including 1672 associated genes. In a final step, these clusters are enriched

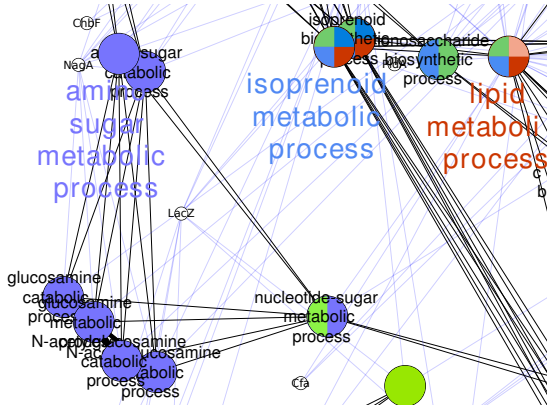

Figure 3: A snapshot of the integrated network in which genes are grouped and enriched with biological process terms from Gene Ontology. The functionally grouped network represents links (edges) of biological process terms (nodes) based on their kappa score levels (0.3). The terms assigned to the groups are partially overlapped and presented with multiple corresponding colors. Only most significant terms are shown with labels of corresponding colors and their node sizes represent the term enrichment significance. The terms, which are not grouped, are shown in white.

with their corresponding Gene Ontology terms using ClueGO (Figure 3). The resulting network represents 1111 nodes of terms in biological process of *E. coli* linking through edges with a default kappa score level (0.3) which measures the association strength between the terms. The functional group of associated genes, annotated with their biological processes, are readily available for biological interpretation.

The seamless integration illustrated above is due to the well-defined functionality of CyEVEX, which includes all relevant information of the genes in the network, allowing straightforward integration of the textual data with external resources. The wealth of publicly available Cytoscape plug-ins creates the opportunity for many more similar use cases, offering a powerful tool for analysing molecular networks and integrating textual information. Such analyses are the foundation for subsequent biological interpretations of experimental results.

5 Conclusions and Future Work

We presented CyEVEX, a tool for integrating the EVEX text mining resource with Cytoscape, a widely used network analysis framework. CyEVEX provides straightforward access to the 34 million EVEX biomolecular events as well as all the additional functionality implemented in EVEX, such as the scoring mechanism and gene family based event generalizations. CyEVEX complements the existing EVEX web application, the two comprising a comprehensive interface focused on end-users.

In addition to soliciting user feedback and implementing the resulting feature requests, possible future work includes the enhancement of functionality provided by CyEVEX to ensure closer integration with the EVEX web application.

Acknowledgments

We would like to thank Sanna Kreula and Patrik R. Jones for providing the analysed microarray expression data for integration, the Academy of Finland, the Research Foundation Flanders and Turku Centre for Computer Science (TUCS) for funding the study, the Ghent University Multidisciplinary Research Partnership 'Bioinformatics: from nucleotides to networks' for additional support, and CSC – IT Center for Science Ltd for computational resources.

References

Pieter Audenaert, Thomas Van Parys, Florian Brondel, Mario Pickavet, Piet Demeester, Yves Van de Peer, and Tom Michoel. 2011. CyClus3D: a Cytoscape plugin for clustering network motifs in integrated networks. *Bioinformatics*, 27(11):1587–1588.

Gabriela Bindea, Bernhard Mlecnik, Hubert Hackl, Pornpimol Charoentong, Marie Tosolini, Amos Kirilovsky, Wolf-Herman Fridman, Franck Pagés, Zlatko Trajanoski, and Jérôme Galon. 2009. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*, 25(8):1091–1093.

Jari Björne, Juho Heimonen, Filip Ginter, Antti Airola, Tapio Pahikkala, and Tapio Salakoski. 2009. Extracting complex biological events with rich graph-based feature sets. In *Proceedings of the BioNLP'09 Shared Task on Event Extraction*, pages 10–18.

Suwisa Kaewphan, Sanna Kreula, Sofie Van Landeghem, Yves Van de Peer, Patrik R. Jones, and Filip Ginter. 2012. Integrating large-scale text mining and co-expression networks: Targeting NADP(H) metabolism in E. coli with event extraction. In *Proceedings of*

the Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2012), pages 8–15.

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of BioNLP'09 Shared Task on event extraction. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pages 1–9. Association for Computational Linguistics.

Jin-Dong Kim, Sampo Pyysalo, Tomoko Ohta, Robert Bossy, Ngan Nguyen, and Jun'ichi Tsujii. 2011. Overview of BioNLP Shared Task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 1–6. Association for Computational Linguistics.

David McClosky, Mihai Surdeanu, and Christopher Manning. 2011. Event extraction as dependency parsing. In *Proceedings of ACL-HLT*, pages 1626–1635. Association for Computational Linguistics.

Makoto Miwa, Paul Thompson, John McNaught, Douglas Kell, and Sophia Ananiadou. 2012. Extracting semantically enriched events from biomedical literature. *BMC Bioinformatics*, 13(1):108.

Michael E. Smoot, Keiichiro Ono, Johannes Ruscheinski, Peng-Liang Wang, and Trey Ideker. 2011. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, 27(3):431–432.

Damian Szklarczyk, Andrea Franceschini, Michael Kuhn, Milan Simonovic, Alexander Roth, Pablo Minguez, Tobias Doerks, Manuel Stark, Jean Muller, Peer Bork, Lars J. Jensen, and Christian von Mering. 2011. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.*, 39(Database issue):D561–568.

Sofie Van Landeghem, Filip Ginter, Yves Van de Peer, and Tapio Salakoski. 2011. EVEX: A PubMed-scale resource for homology-based generalization of text mining predictions. In *Proceedings of BioNLP'11 Workshop*, pages 28–37. Association for Computational Linguistics.

Sofie Van Landeghem, Kai Hakala, Samuel Rönnqvist, Tapio Salakoski, Yves Van de Peer, and Filip Ginter. 2012. Exploring biomolecular literature with EVEX: Connecting genes through events, homology and indirect associations. *Advances in Bioinformatics, special issue Literature-Mining Solutions for Life Science Research*, (582765).

Andrew G. Winter, Jan Wildenhain, and Mike Tyers. 2011. BioGRID REST Service, BiogridPlugin2 and BioGRID WebGraph: new tools for access to interaction data at BioGRID. *Bioinformatics*, 27(7):1043–1044.