Ignet: A centrality and INO-based web system for analyzing and visualizing literature-mined networks

Arzucan Özgür[1*], Junguk Hur[2*], Zuoshuang Xiang[3*], Edison Ong[3], Dragomir R. Radev[3], Yongqun He[3§]

[1] Bogazici University, 34342 Istanbul, Turkey; [2] University of North Dakota, US; [3] University of Michigan, Ann Arbor, USA.

ABSTRACT

Ignet (Integrative Gene Network) is a web-based system for dynamically updating and analyzing gene interaction networks mined using all PubMed abstracts. Four centrality metrics, namely degree, eigenvector, betweenness, and closeness are used to determine the importance of genes in the networks. Different gene interaction types between genes are classified using the Interaction Network Ontology (INO) that classifies interaction types in an ontological hierarchy along with individual keywords listed for each interaction type. An interactive user interface is designed to explore the interaction network as well as the centrality and ontology based network analysis. Availability: http://ignet.hegroup.org.

1 INTRODUCTION

Many web systems exist for literature mining of gene interactions, e.g., Chilibot (http://www.chilibot.net/) and iHOP (http://www.ihop-net.org/UniPub/iHOP/). Some of these tools mark the interaction keywords in the sentences. One common obstacle is that these interaction keywords are not classified; so detailed interaction types cannot be studied.

Ontology-based literature mining is an emerging research field that applies ontology to support literature mining. The Interaction Network Ontology (INO) is a newly developed interaction ontology that supports biomedical literature mining (Hur et al., 2015). INO was initially developed to represent over 800 interaction keywords (Ozgur et al., 2011), and their hierarchical structure using ontological format, and more interaction terms were later added to INO with well-defined axioms (Hur et al., 2015).

In our previous studies, we also ranked the genes in the literature-mined gene networks using different types of centralities: *degree centrality*, *eigenvector centrality*, *closeness centrality*, and *betweenness centrality* (Ozgur et al., 2011). Theses centralities measure different levels of importance. For example, in betweenness centrality a node is considered important if it occurs on many shortest paths between other nodes, whereas in degree centrality a node is considered important if it is connected to many other nodes.

We have named our literature mining strategy Centrality and Ontology-based Network Discovery using Literature data (CONDL) (Ozgur et al., 2011). CONDL was successfully applied to extract and analyze IFN-γ and vaccine-related gene interaction network as well as vaccine and fever-related gene interaction network (Hur et al., 2012).

Based on the CONDL strategy, we have developed Ignet (http://ignet.hegroup.org), a web-based literature mining database system that stores gene-gene interactions extracted from PubMed abstracts. A gene–gene interaction in this study corresponds to an interaction between genes and/or gene products such as proteins.

2 FEATURES AND USAGE

Briefly, all article abstracts available in PubMed are retrieved. The sentences, split by Java's internal splitter (BreakIterator), were examined using SciMiner (Hur et al., 2009) to identify gene names and interaction keyword(s) (e.g., interacts, binds, activates) represented in INO. We obtained the dependency parse trees of the sentences using the Stanford Parser (http://nlp.stanford.edu/software/lex-parser.shtml) and extracted the shortest dependency path between each pair of genes in a sentence. Our assumption is that the shortest path between two gene names in a dependency tree is a good description of the semantic relation between them in the corresponding sentence. We defined an edit distance-based kernel function among these dependency paths and used support vector machines (SVM) in the SVM^{light} package (Joachims, 1999) to classify each path as describing an interaction between the gene pair or not (Erkan et al., 2007). The value output by the decision function of the SVM classifier (*i.e.*, the score field in Fig. 1B) can be used as a confidence score to measure the confidence of association between two genes in a sentence. Positive score means that the SVM classifier predicts an "interaction", whereas negative score corresponds to a prediction of "not interaction". The larger the absolute value of a score, the more confidant the classifier is in the classification decision. The higher the score of a sentence is, the more likely it is that the sentence describes an interaction between the pair of genes. The current database contains only those interactions with a positive SVM score.

3 IGNET USE CASE DEMONSTRATION

Ignet contains user-friendly web query interface (Fig. 1). A user can query one gene or two genes. Each gene has its own centrality scores, which indicate its degree of importance in a network. All sentences associated with the queries are obtained, with gene name and INO interaction verbs highlighted. A click to a specific INO interaction verb

* These authors contributed equally.

§ To whom correspondence should be addressed: yongqunh@umich.edu

links to a page that shows the hierarchy of the INO verbs (Fig. 1C).

In addition, Ignet also includes a subprogram called Dignet (http://ignet.hegroup.org/dignet), which applies PubMed search to define the scope of papers for generating the network, and use the Ignet execution pipeline to generate gene-gene interactions and networks and calculate centrality scores for genes in the networks.

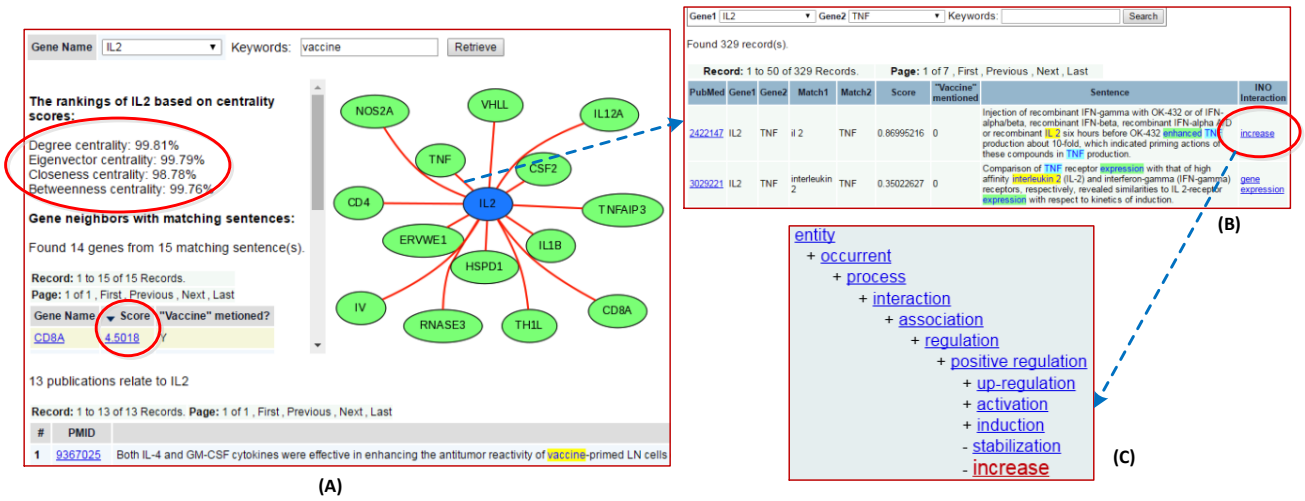

Fig. 1. Ignet web query of literature mined gene interaction network. (A) The list of genes as the neighbors associated with IL2 in the vaccine context, and the automatically generated visualization graphic displaying the interactions among IL2 and its associated genes. Red circled are centrality scores and a confidence score (e.g., 4.5018) for ranking gene-gene interactions. (B) Once the edge between IL2 and TNF is clicked, the publication records to support the interaction are shown in another page. In addition to the two gene/protein names, the interaction keywords (e.g., increase) are also shown. (C) Once the interaction word "increase" is clicked, the ontology hierarchy of this term in INO is displayed in an Ontobee web page (Xiang et al., 2011).

4 SUMMARY

Ignet is a web-based literature mining system that integrates the centrality-based literature mining approach with INO-based ontology analysis of interaction types. The gene-gene relationships are extracted using machine learning methods with the syntactic and semantic structures of the sentences. To the best of our knowledge, Ignet is the first web system that provides centrality analysis for literature-mined gene interaction networks and ontology representation of interaction types. Ignet not only provides access to automatically extracted gene interactions, but it also enables generations of new hypotheses (Özgür *et al.,* 2010; Özgür *et al.,* 2011; Hur *et al.*, 2012).

ACKNOWLEDGEMENTS

This work was supported by grant R01AI081062 from the US NIH National Institute of Allergy and Infectious Diseases and by Marie Curie FP7-Reintegration-Grants within the 7th European Community Framework Programme.

REFERENCES

Erkan G, Ozgur A, Radev D: Semi-Supervised Classification for Extracting Protein Interaction Sentences using Dependency Parsing. In Proceedings EMNLP-CoNLL. 2007: 228-237.

Hur, J., Ozgur, A., Xiang, Z., and He, Y. (2012). Identification of fever and vaccine-associated gene interaction networks using ontology-based literature mining. *J Biomed Semantics* 3**,** 18.

Hur, J., Ozgur, A., Xiang, Z., and He, Y. (2015). Development and application of an interaction network ontology for literature mining of vaccine-associated gene-gene interactions. *J Biomed Semantics* 6**,** 2.

Hur, J., Schuyler, A.D., States, D.J., and Feldman, E.L. (2009). SciMiner: web-based literature mining tool for target identification and functional enrichment analysis. *Bioinformatics* 25**,** 838-840.

Joachims, T. (1999). "Making large-scale support vector machine learning practical," in *Advances in Kernel Methods: Support Vector Learning,* ed. C.J.B. B. Schölkopf, And A. J. Smola, Eds. (Cambridge, MA.: MIT Press), 169-184.

Ozgur, A., Xiang, Z., Radev, D.R., and He, Y. (2011). Mining of vaccine-associated IFN-gamma gene interaction networks using the Vaccine Ontology. *J Biomed Semantics* 2 Suppl 2**,** S8.

Xiang, Z., Mungall, C., Ruttenberg, A., and He, Y. (Year). "Ontobee: A linked data server and browser for ontology terms", in: *The 2nd International Conference on Biomedical Ontologies (ICBO)*: CEUR Workshop Proceedings), Pages 279-281.