Bioinformatics Advance Access published September 16, 2010

Evidence mining and novelty assessment of protein-protein interactions with the ConsensusPathDB plugin for Cytoscape

Konstantin Pentchev[1†], Keiichiro Ono[2†], Ralf Herwig[1], Trey Ideker[2] and Atanas Kamburov[1,2,*]

[1]Department of Vertebrate Genomics, Max Planck Institute for Molecular Genetics, Ihnestr. 63-73, 14195 Berlin, Germany

[2]Departments of Medicine and Bioengineering, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093

Associate Editor: Dr. Jonathan Wren

ABSTRACT

Summary: Protein-protein interaction detection methods are applied on a daily basis by molecular biologists worldwide. After generating a set of potential interactions, biologists face the problem of highlighting the ones that are novel and collecting evidence with respect to literature and annotation. This task can be as tedious as searching for every predicted interaction in several interaction data repositories, or manually screening the scientific literature. To facilitate the task of evidence mining and novelty assessment of protein-protein interactions, we have developed a Cytoscape plugin that automatically mines publication references, database references, interaction detection method descriptions and pathway annotation for a user-supplied network of interactions. The basis for the annotation is ConsensusPathDB – a meta-database that integrates numerous protein-protein, signaling, metabolic and gene regulatory interaction repositories for currently three species: *Homo sapiens*, *Saccharomyces cerevisiae* and *Mus musculus*.

Availability: The ConsensusPathDB plugin for Cytoscape (version 2.7.0 or later) can be installed within Cytoscape on a major operating system (Windows, Mac OS, Unix/Linux) with Sun Java 1.5 or later installed through Cytoscape's Plugin manager (category 'Network and Attribute I/O'). The plugin is freely available for download on the ConsensusPathDB web site (http://cpdb.molgen.mpg.de).

Contact: kamburov@molgen.mpg.de

1 INTRODUCTION

Due to the high explanatory power of protein-protein interactions for biological processes in health and disease (Ideker and Sharan, 2008), dedicated interaction detection methods like yeast-two-hybrid (Y2H) screening (Fields, 2005) and co-purification (Aebersold and Mann, 2003) are applied on a daily basis by molecular biologists worldwide and contribute to the completion of the map of protein-protein interactions for human and other species. An immediate task after generating a network of predicted interactions is to identify the ones that have not been published previously and to collect evidence for every single interaction from literature and annotation. This information is useful in order to estimate the performance of the interaction screen and to assess the contribution to the protein-protein interaction map of the species in question. To accomplish this task, biologists typically search their new data against every single protein-protein interaction repository like IntAct (Huntley *et al.*, 2007) or MINT (Chatr-aryamontri *et al.*, 2007). Even more tedious is the manual mining for interactions in scientific literature to collect the publication references and detection methods for the novel interaction list.

Cytoscape (Shannon *et al.*, 2003) is a widely used, freely available software tool for visualization, manipulation and analysis of biomolecular interaction networks. To aid the process of interaction evidence mining, we have developed a plugin for Cytoscape that searches all interactions from the network of interest in the interaction space stored in ConsensusPathDB. ConsensusPathDB (Kamburov *et al.*, 2009) is an interaction meta-database that integrates functional interaction repositories forming a heterogeneous interaction network which comprises protein-protein interactions, as well as signaling, metabolic and gene regulatory interactions. Currently, the database integrates 18 open-access repositories on human interactions and 8 repositories for both yeast and mouse interactions and contains around 150,000 human, 195,000 yeast and 13,000 mouse distinct interactions (many of which are of non-binary nature, i.e. contain more than two interaction partners). In this paper, we describe the functionality of the ConsensusPathDB plugin for Cytoscape and demonstrate its usage and performance.

2 DESCRIPTION

After installing the plugin, the user starts by loading the network of interest (denoted query network) represented by binary interactions in Cytoscape and launching the ConsensusPathDB plugin through Cytoscape's 'Plugins' menu (Figure 1 A). After setting a few parameters which we describe below, the user starts the evidence mining process. The plugin then communicates with the repository of ConsensusPathDB through a web service. Once the plugin sends the query network to the server, a search is executed on the server-side for all (or, optionally, just the selected) proteins and interactions from the query network in ConsensusPathDB through SQL queries. Proteins from the query network are matched to the data repository on the basis of accession numbers such as UniProt (The UniProt Consortium, 2010) or Ensembl (Flicek *et al.*, 2010). Inter-

*To whom correspondence should be addressed.

†These authors contributed equally to this work.

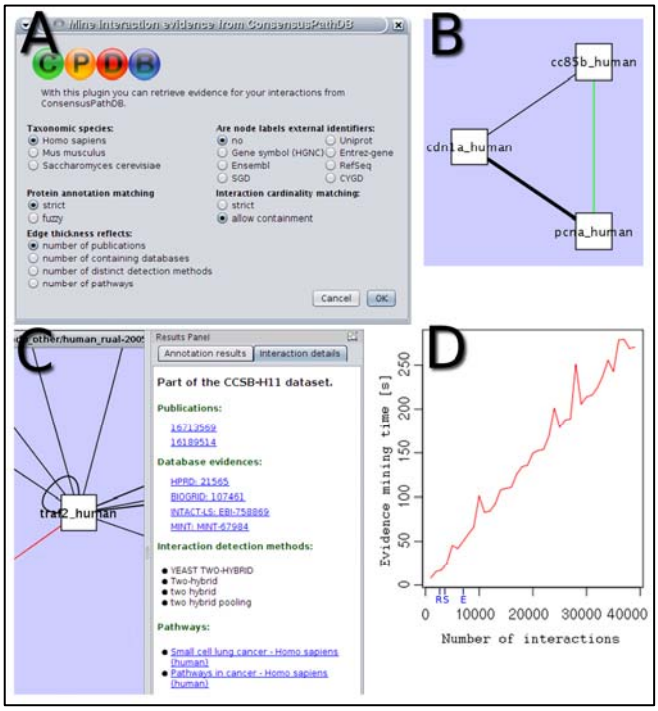

Fig. 1. **(A)** The splash screen of the plugin showing the different parameters; **(B)** the ConsensusPathDB visual style where reproduced interactions are weighted by evidence and novel interactions are highlighted in green; **(C)** newly imported attributes of a selected interaction are shown in the 'Interaction details' tab of Cytoscape's results panel; **(D)** evidence mining time plot for networks of different size with default parameters (here, all query interactions were present in ConsensusPathDB such that the mining process took maximal time). The sizes of the networks predicted using large-scale interaction screening by Rual et *al.*,2005 (R), Stelzl *et al.*,2005 (S) and Ewing *et al.*,2006 (E) are marked on the x-axis.

actions from the query network are matched to the repository based on their participants.

The performance of the interaction matching depends critically on how well proteins in the query network are annotated with accession numbers. In the case that accession numbers are not available, the user is prompted to specify whether the node labels represent accession numbers of a certain type. The interaction matching performance is influenced by two parameters, 'protein annotation matching' (strict / fuzzy) and 'interaction cardinality matching' (strict / allow containment). Strict protein annotation matching denotes that a protein from the query network and a protein from the database are considered identical only if all identifiers of a type match. Fuzzy matching means that the identifiers of the query protein may form a sub-set of the identifiers of the database counterpart or vice versa. Fuzzy matching is useful e.g. when proteins on the one side are compared with protein families on the other side. The 'interaction cardinality matching' parameter specifies whether the binary interactions from the query network should be matched only with binary interactions from the database network (strict matching) or whether they may be matched to complex interactions, i.e. interactions of more than two proteins that contain the binary interactions. More details about protein and interaction mapping can be found in the Supplementary text to his paper.

After matching proteins and interactions, the web service server sends annotation attributes for matched query interactions in the form of publication references (Pubmed identifiers), interaction detection methods, database references (such as IntAct and MINT) and pathway annotations (i.e. pathways that contain both participants of a protein-protein interaction) to the client plugin. The plugin creates a custom visual style in Cytoscape where the thickness of interaction edges reflects (optionally) the number of publications, number of containing interaction databases, number of distinct detection methods, or number of containing pathways for the protein interaction (Figure 1 B). Interactions that are not found in the repository, and thus represent potential novel interactions, are highlighted in green. In the results tab of Cytoscape, an interaction mapping summary is displayed together with a legend. The interaction attributes that have been retrieved from ConsensusPathDB can be viewed for selected interactions under the 'Interaction details' tab of the results panel (Figure 1 C). If applicable, this information is provided as web links to the primary data and can be viewed in a web browser.

Figure 1 D shows the performance of the plugin implementation with respect to the mining of interaction annotation for different network sizes. Results show that even for large networks evidence mining executes in minutes, for example ~2 minutes for a network with 20,000 nodes. It should be noted, however, that the Internet connection speed of the client influences the overall speed of interaction matching.

ACKNOWLEDGEMENTS

Funding: Max Planck Society (IMPRS-CBSC); Cytoscape project (GM070743); European Union's project APO-SYS (HEALTH-F4-2007-200767); BMBF MedSys project PREDICT (0315428A).
Conflict of interest: none declared.

REFERENCES

Aebersold,R. and Mann,M. (2003) Mass spectrometry-based proteomics. *Nature*, **422**, 198-207.
Chatr-aryamontri,A. *et al.* (2007) MINT: the Molecular INTeraction database. *Nucleic Acids Res.*, **35**, D572-574.
Ewing,R.M. *et al.* (2006) Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol. Syst. Biol.*, **3**, 89.
Fields,S. (2005) High-throughput two-hybrid analysis. The promise and the peril. *FEBS J*, **272**, 5391-5399.
Flicek,P. *et al.* (2010) Ensembl's 10th year. *Nucleic Acids Res.*, **38**, D557-562.
Huntley,R. *et al.* (2007) IntAct--open source resource for molecular interaction data. *Nucleic Acids Res*, **35**, D561-565.
Ideker,T. and Sharan,R. (2008) Protein networks and disease. *Genome Res.*, **18**, 644-652.
Kamburov,A. *et al.* (2009) ConsensusPathDB--a database for integrating human functional interaction networks. *Nucleic Acids Res.*, **37**, D623-628.
Rual,J.F. *et al.* (2005) Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, **437,** 1173-1178.
Shannon,P. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498-2504.
Stelzl,U. *et al.* (2005) A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, **122**, 957-968.
The UniProt Consortium (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.*, **38**, D142-148.