Nucleic Acids Research, 2017 **1**
doi: 10.1093/nar/gkx930

The online Tabloid Proteome: an annotated database of protein associations

Surya Gupta[1,2,3,†], Demet Turan[1,2,3,†], Jan Tavernier[1,2] and Lennart Martens[1,2,3,*]

[1]VIB-UGent Center for Medical Biotechnology, VIB, Ghent 9000, Belgium, [2]Department of Biochemistry, Ghent University, Ghent 9000, Belgium and [3]Bioinformatics Institute Ghent, Ghent University, Ghent 9000, Belgium

Received August 04, 2017; Revised September 11, 2017; Editorial Decision October 02, 2017; Accepted October 10, 2017

ABSTRACT

A complete knowledge of the proteome can only be attained by determining the associations between proteins, along with the nature of these associations (e.g. physical contact in protein–protein interactions, participation in complex formation or different roles in the same pathway). Despite extensive efforts in elucidating direct protein interactions, our knowledge on the complete spectrum of protein associations remains limited. We therefore developed a new approach that detects protein associations from identifications obtained after re-processing of large-scale, public mass spectrometry-based proteomics data. Our approach infers protein association based on the co-occurrence of proteins across many different proteomics experiments, and provides information that is almost completely complementary to traditional direct protein interaction studies. We here present a web interface to query and explore the associations derived from this method, called the online Tabloid Proteome. The online Tabloid Proteome also integrates biological knowledge from several existing resources to annotate our derived protein associations. The online Tabloid Proteome is freely available through a user-friendly web interface, which provides intuitive navigation and data exploration options for the user at http://iomics.ugent.be/tabloidproteome.

INTRODUCTION

Protein–protein interactions are well known to play key roles in living cells (1,2). These protein interactions can be studied through a variety of approaches, including circuit-completion assays such as yeast-two-hybrid and MAPPIT (3), affinity purification and cross-linking mass spectrometry analyses (4) and computational analyses of protein structure (5). Yet despite the immense popularity and remarkable successes of these methods, these remain focused on direct protein interactions through actual contact between the molecules, a focus that is also found in the often complex algorithms that re-use such specialized data to determine direct protein interactions (6). As such, these approaches do not currently detect other types of biologically meaningful protein associations, for instance between proteins that are active in the same pathway. Such indirect associations are, however, likely to be as relevant as direct protein interactions for the understanding of cells and tissues in health and disease (7). Indeed, the disruption of signaling or biochemical pathways through small molecule toxins, pathogen-produced macromolecules, or cancer mutations can have profound health consequences (8). We have therefore developed a novel approach to detect novel and biologically associated protein pairs, which is based on the orthogonal re-use of publicly available data (9).

In this approach, we have used publicly available mass spectrometry-based human proteomics experiments from the PRIDE database (10). These experiments are re-processed using a pipeline build from our pride-asap (11), SearchGUI (12) and PeptideShaker (13) tools, integrated through our Pladipus (14) platform. Co-occurring protein pairs across many experiments are detected using a Jaccard Similarity metric based on the number of distinct peptides identified for each protein. Validation of the detected protein associations was performed by cross-referencing the obtained protein pairs with biological knowledge from various existing resources: pathways from Reactome (15), protein–protein interactions from IntAct (16) and BioGRID (17), protein complexes from CORUM (18), and paralog information from Ensembl (19). The majority of detected protein associations also had strong existing biological annotations, and very significantly more so than random, showing that the approach was sound (9). Further details on methodology and validation are provided in Supplementary Information S1.

Interestingly, however, only very few of the obtained associations are explained through direct protein interactions (<3% of obtained associations are known as protein interactions), which indicates that our method is very comple-

*To whom correspondence should be addressed. Tel: +32 92 649 358; Fax: +32 9 264 94 84; Email: lennart.martens@vib-ugent.be
†These authors contributed equally to the paper as first authors.

mentary to traditional protein-protein interaction analyses. Moreover, protein associations with little or no annotation were also observed, typically because one or both proteins in the pair are very poorly annotated in existing databases.

To make the results of our analysis easily available to the research community, we here present the online Tabloid Proteome as a convenient and user-friendly means to access and query all obtained protein associations, along with extensive annotation based on existing knowledgeb bases (http://iomics.ugent.be/tabloidproteome).

DATA COLLECTION AND INTEGRATION

Data sources

Mass spectrometry-based human proteomics experiments were retrieved from the PRIDE database (release May_2015). The biological annotation included in online Tabloid Proteome is collected from six resources (Figure 1A): biological pathways are derived from Reactome (15) (V56); binary protein–protein interaction data were obtained from IntAct (16) (release 2016_01) and BioGRID (17) (version 3.4.145); protein complexes were obtained from CORUM (18) (release 2012_02), and paralog information was obtained from Ensembl (19) (version 83). To annotate protein pairs with little or no biological information, Gene Ontology (GO) annotation from UniProtKB/Swiss-Prot (20) (release 2016_06) was obtained for biological process, molecular function and cellular component annotation. Disease related information was obtained from DisGeNET (21) (version 4.0), and tissue annotation was added from The Human Protein Atlas (22) (version 17). We have used the UniProt and DAVID (23) accession number conversion.

Database architecture and database content

All data underlying the Tabloid Proteome are stored in the Neo4j graph database (https://neo4j.com). The Tabloid Proteome currently stores 551 420 distinct associations for 4562 unique proteins pairs, all with a minimum Jaccard similarity of 0.1. These associations map against 1231 Reactome leaf pathways, 1104 CORUM complexes, 13 593 protein–protein interactions from BioGRID and IntAct, 53 tissue annotations from The Human Protein Atlas and 9697 DisGeNET diseases. These associations are currently derived from 99 distinct PRIDE projects comprising a total of 1063 assays.

Web development

The web interface was developed using the Java Server Faces (JSF) framework, with JQuery (https://jquery.com), Bootstrap (http://getbootstrap.com) and Primefaces (https://www.primefaces.org). Visualization of the protein association graphs is performed by the Cytoscape JavaScript library (24).

TABLOID PROTEOME FEATURES AND APPLICATION

User interface

The online Tabloid Proteome can be queried by protein, gene, pathway, tissue or disease (Figure 1C), and results can be filtered by a Jaccard Similarity threshold score. The user can search by an entity's name or its accession number (UniProt accession, Entrez gene ID, Reactome accession or DisGeNET ID). It is also possible to search for two or more proteins simultaneously, and for up to two genes simultaneously.

The web interface displays the results in a data table (Figure 2A). Initially, the table shows summary information in each row, which the user can then expand to see interactions, common PRIDE projects, pathways, complexes, gene ontology terms and common diseases between the two proteins in that pair (Figure 2C). Furthermore, results can also be visualized through a graph view. The user can choose to see genes or proteins, which will be rendered as nodes, while different types of associations are shown as color-coded edges (Figure 2B). All retrieved data can also be downloaded in tab-delimited or CSV format (for the tabular view), or as a PNG image (for the graph view).

A RESTful Web Service is also provided for other applications to access the Tabloid Proteome. The JSON structure and detailed information about the web service is available in the API section of the Tabloid Proteome website (http://iomics.ugent.be/tabloidproteome/tabloidApi.html).

Example use cases

The online Tabloid Proteome can be queried in six different ways, each of which supports different use cases. These five use cases, along with some examples, are detailed below.

The first way to access the online Tabloid Proteome is via a single protein or gene search. In this case, the user provides either a single protein (e.g. Hemopexin) or gene name (e.g. *HPX)*, or a single UniProt accession number (e.g. P02790) or Entrez gene ID (e.g. 3263) in the corresponding text field. When a protein or gene name, or an Entrez gene ID is used for the query, a dialog will pop up to show the matching proteins for confirmation and/or disambiguation (e.g. searching MTA1 provides two possible matching proteins). Clicking the magnifying glass icon in front of the desired protein reveals all known association partners for that protein.

Based on the query protein or gene, the resulting set of associations can consist of already well-known associations, as-yet unknown associations, or a combination of these. When querying the online Tabloid Proteome with protein MTA1 (Q13330), for instance, four associated proteins are found, all four of which are already well known to directly interact with MTA1. On the other hand, querying by protein Hemopexin (P02790), yields seven associated proteins, none of which have any prior knowledge linking it to Hemopexin. The system does indicate that many of these have a shared role in diseases. A literature search for the first associated protein, Vitamin-D binding protein (P02774, *GC*) with a Jaccard similarity score of 0.58, revealed that it has been found to be co-overexpressed with Hemopexin in breast tumor (25), which illustrates that de-

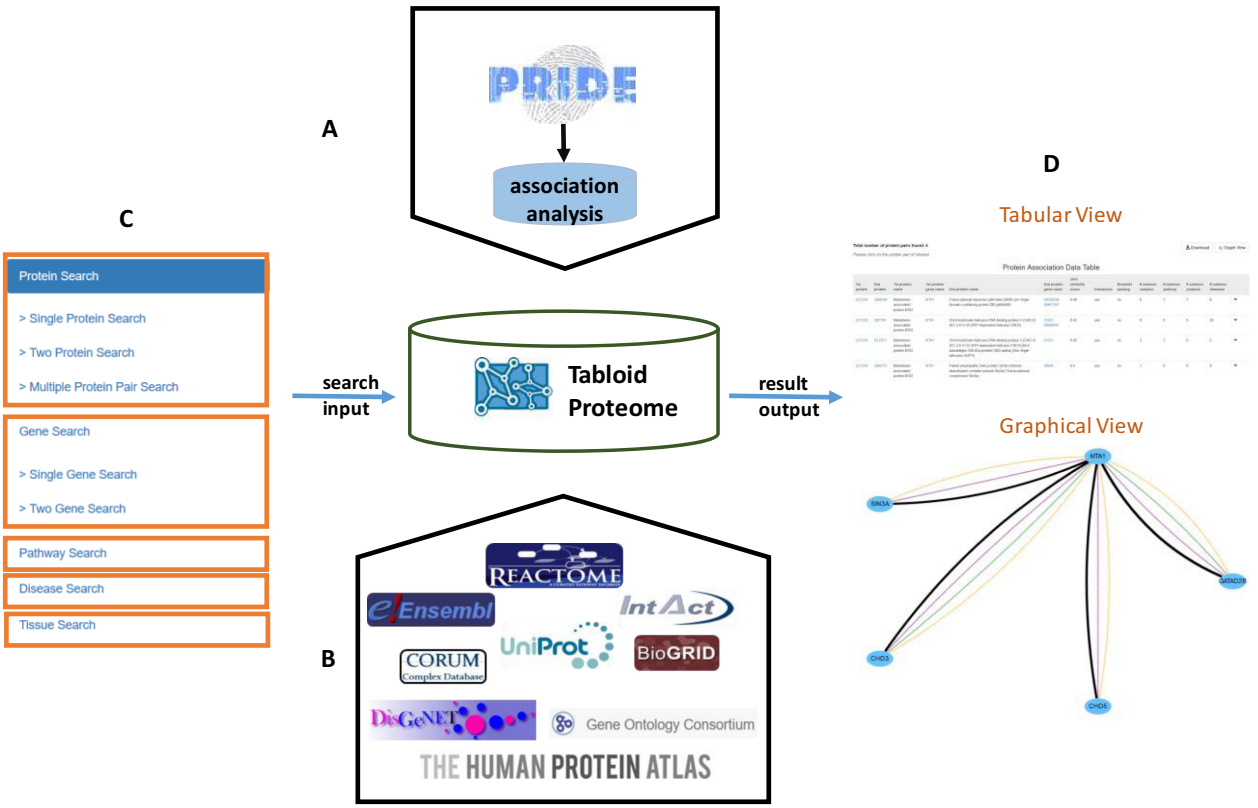

Figure 1. Overall view of the online Tabloid Proteome. (**A**) Raw data are derived from PRIDE database, which is further reprocessed and analyzed with a Jaccard similarity score. (**B**) The derived protein association pairs are then annotated with data from nine different resources. (**C**) The online Tabloid Proteome interface allows searches by proteins, genes, pathway, tissue and disease. (**D**) The results are depicted in two forms: a tabular view and an interactive graphical view.

spite a lack of existing knowledge, the association is likely biologically meaningful.

The user is thus able to consider all retrieved associations, to choose only well-known associations, or to focus exclusively on novel associations; the online Tabloid Proteome provides ample information to make such a triage easily possible.

A second use case is to focus on a protein or gene pair, by using the two protein (or two gene) search option. The supported input types are identical to the single protein or gene search. Also, as for the single search, when using anything but a UniProt accession number, a dialog pops up listing all possible protein associations for confirmation and/or disambiguation.

Importantly, for a two protein query, the online Tabloid Proteome also automatically retrieves indirect associations if the searched pair does not have a direct association. For instance, when queried with the protein accession numbers Q96ST3 and Q8TDI0, no direct associations are found but the system presents the user with a possible indirect relation between both proteins, as both are associated to *MTA1*.

A third use case is to query the online Tabloid Proteome for multiple protein association pairs, which is supported by the multiple pair protein search option. Here, the user can upload a file with protein pairs, which are then queried. Interestingly, the multiple pair protein search also allows users to provide their own links for these pairs in the uploaded file, which are then visualized in the graphical view alongside the other edges.

The online Tabloid Proteome search is not limited to only protein or gene-based searches.

Indeed, the fourth use case allows the system to be queried by a Reactome pathway, using either a pathway name or Reactome accession number. The input text can be generic, such as 'metabolism', or can be specific, for instance 'Pyruvate metabolism' or Reactome accession number 'R-HSA-70268'. The result is a list of all known associated protein pairs in which both members are involved in the provided pathway. Each individual pair can then be explored in more detail by clicking the magnifying glass icon for that pair in the overview table.

The fifth use case then, consists of retrieving all associated protein pairs that are detected for a particular disease. This can be achieved by querying the online Tabloid Proteome with either a disease name or a DisGeNET ID. As for the pathways, the query can be generic (e.g. 'cancer', which will result in 139 matched diseases), or selective (e.g. 'Cancer of Nasopharynx' or DisGeNET ID: 'C0238301').

The sixth case allows the user to restrict their search for associated protein pairs by tissue annotation, as provided by The Human Protein Atlas. This search can be performed by selecting one of the 53 tissue names, for example, 'lung', resulting in 653 protein pairs. Similar to the disease and pathway search, the results are shown in an overview table.

Figure 2. Results in the online Tabloid Proteome are displayed in two different views: a tabular and a graphical view. (**A**) In the tabular view, each row provides an overall summary of the association, while a dropdown section per row provides further information (as shown in (**C**)). (**B**) The interactive graphical view shows proteins as nodes and associations as color-coded edges. (C) Detailed information about each type of association is revealed using dropdowns the tabular view or by clicking the corresponding edge in the graphical view.

DISCUSSION AND OUTLOOK

The online Tabloid Proteome is an easily searchable website that presents protein associations derived from public mass spectrometry-based proteomics datasets, along with annotation of these associations using information from a variety of existing knowledge bases. Due to the strong complementarity with existing approaches (most notably with direct protein-protein interaction studies) the online Tabloid Proteome provides unique added value for researchers that are interested in understanding the relations between proteins.

Future plans for the online Tabloid Proteome focus on the inclusion of association data obtained from public pro-

teomics data from other model organisms (notably mouse, yeast and Arabidopsis). This will also allow orthologous association pairs that have been conserved across evolution to be derived. Additionally, we also compared the protein pairs to information from the STRING database (26) and found 296 protein pairs that had text mining and/or co-evolution annotation in STRING, but that had no annotation in any of the other annotation databases. We will therefore also add STRING database annotations to the online Tabloid Proteome.

AVAILABILITY

The online Tabloid Proteome is freely available via http://iomics.ugent.be/tabloidproteome. The documentation for database usage is available via http://genesis.ugent.be/uvpublicdata/Tabloid_Proteome/TabloidProteome1.2_documentation.pdf and the documentation for API REST is available via http://iomics.ugent.be/tabloidproteome/tabloidApi.html.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Ghent University Concerted Research Action [BOF12/GOA/014]; ERC Advanced Grant [N° 340941, CYRE]; Research Foundation - Flanders (FWO) and Hermesfonds for ELIXIR Belgium [AH.2017.051, IO 17001306]. Funding for open access charge: ERC Advanced Grant [N° 340941, CYRE].
Conflict of interest statement. None declared.

REFERENCES

1. Lievens,S., Eyckerman,S., Lemmens,I. and Tavernier,J. (2010) Large-scale protein interactome mapping: strategies and opportunities. *Expert Rev. Proteomics*, **7**, 679–690.
2. Luck,K., Sheynkman,G.M., Zhang,I. and Vidal,M. (2017) Proteome-scale human interactomics. *Trends Biochem. Sci.*, **42**, 342–354.
3. Lemmens,I., Lievens,S. and Tavernier,J. (2010) Strategies towards high-quality binary protein interactome maps. *J. Proteomics*, **73**, 1415–1420.
4. Mehta,V. and Trinkle-Mulcahy,L. (2016) Recent advances in large-scale protein interactome mapping. *F1000Research*, **5**, 782.
5. Gromiha,M.M., Yugandhar,K. and Jemimah,S. (2017) Protein-protein interactions: scoring schemes and binding affinity. *Curr. Opin. Struct. Biol.*, **44**, 31–38.
6. Drew,K., Lee,C., Huizar,R.L., Tu,F., Borgeson,B., McWhite,C.D., Ma,Y., Wallingford,J.B. and Marcotte,E.M. (2017) Integration of over 9, 000 mass spectrometry experiments builds a global map of human protein complexes. *Mol. Syst. Biol.*, **13**, 932.
7. Wilson,J.L., Hemann,M.T., Fraenkel,E. and Lauffenburger,D.A. (2013) Integrated network analyses for functional genomic studies in cancer. *Semin. Cancer Biol.*, **23**, 213–218.
8. Wang,E. (2013) Understanding genomic alterations in cancer genomes using an integrative network approach. *Cancer Lett.*, **340**, 261–269.
9. Gupta,S., Verheggen,K., Tavernier,J. and Martens,L. (2017) Unbiased protein association study on the public human proteome reveals biological connections between co-occurring protein pairs. *J. Proteome Res.*, **16**, 2204–2212.
10. Vizcaíno,J.A., Csordas,A., del-Toro,N., Dianes,J.A., Griss,J., Lavidas,I., Mayer,G., Perez-Riverol,Y., Reisinger,F., Ternent,T. *et al.* (2016) 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.*, **44**, D447–D456.
11. Hulstaert,N., Reisinger,F., Rameseder,J., Barsnes,H., Vizcaíno,J.A. and Martens,L. (2013) Pride-asap: automatic fragment ion annotation of identified PRIDE spectra. *J. Proteomics*, **95**, 89–92.
12. Vaudel,M., Barsnes,H., Berven,F.S., Sickmann,A. and Martens,L. (2011) SearchGUI: an open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. *Proteomics*, **11**, 996–999.
13. Vaudel,M., Burkhart,J.M., Zahedi,R.P., Oveland,E., Berven,F.S., Sickmann,A., Martens,L. and Barsnes,H. (2015) PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nat. Biotechnol.*, **33**, 22–24.
14. Verheggen,K., Maddelein,D., Hulstaert,N., Martens,L., Barsnes,H. and Vaudel,M. (2016) Pladipus enables universal distributed computing in proteomics bioinformatics. *J. Proteome Res.*, **15**, 707–712.
15. Fabregat,A., Sidiropoulos,K., Garapati,P., Gillespie,M., Hausmann,K., Haw,R., Jassal,B., Jupe,S., Korninger,F., McKay,S. *et al.* (2016) The reactome pathway knowledgebase. *Nucleic Acids Res.*, **44**, D481–D487.
16. Orchard,S., Ammari,M., Aranda,B., Breuza,L., Briganti,L., Broackes-Carter,F., Campbell,N.H., Chavali,G., Chen,C., del-Toro,N. *et al.* (2014) The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.*, **42**, D358–D363.
17. Chatr-Aryamontri,A., Oughtred,R., Boucher,L., Rust,J., Chang,C., Kolas,N.K., O'Donnell,L., Oster,S., Theesfeld,C., Sellam,A. *et al.* (2017) The BioGRID interaction database: 2017 update. *Nucleic Acids Res.*, **45**, D369–D379.
18. Ruepp,A., Waegele,B., Lechner,M., Brauner,B., Dunger-Kaltenbach,I., Fobo,G., Frishman,G., Montrone,C. and Mewes,H.-W. (2010) CORUM: the comprehensive resource of mammalian protein complexes–2009. *Nucleic Acids Res.*, **38**, D497–D501.
19. Yates,A., Akanni,W., Amode,M.R., Barrell,D., Billis,K., Carvalho-Silva,D., Cummins,C., Clapham,P., Fitzgerald,S. and Gil,L. (2016) Ensembl 2016. *Nucleic Acids Res.*, **44**, D710–D716.
20. Consortium,U. (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **45**, D158–D169.
21. Piñero,J., Bravo,À., Queralt-Rosinach,N., Gutiérrez-Sacristán,A., Deu-Pons,J., Centeno,E., García-García,J., Sanz,F. and Furlong,L.I. (2017) DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.*, **45**, D833–D839.
22. Uhlén,M., Fagerberg,L., Hallström,B.M., Lindskog,C., Oksvold,P., Mardinoglu,A., Sivertsson,Å., Kampf,C., Sjöstedt,E., Asplund,A. *et al.* (2015) Tissue-based map of the human proteome. *Science*, **347**, 1260419.
23. Huang,D.W., Sherman,B.T. and Lempicki,R.A. (2008) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
24. Franz,M., Lopes,C.T., Huck,G., Dong,Y., Sumer,O. and Bader,G.D. (2015) Cytoscape.js: a graph theory library for visualisation and analysis. *Bioinformatics*, **38**, 309–311.
25. Pawlik,T.M., Hawke,D.H., Liu,Y., Krishnamurthy,S., Fritsche,H., Hunt,K.K. and Kuerer,H.M. (2006) Proteomic analysis of nipple aspirate fluid from women with early-stage breast cancer using isotope-coded affinity tags and tandem mass spectrometry reveals differential expression of vitamin D binding protein. *BMC Cancer*, **6**, 68.
26. Szklarczyk,D., Franceschini,A., Wyder,S., Forslund,K., Heller,D., Huerta-Cepas,J., Simonovic,M., Roth,A., Santos,A., Tsafou,K.P. *et al.* (2015) STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.*, **43**, D447–D452.