VERSITA

CELLULAR & MOLECULAR BIOLOGY LETTERS
http://www.cmbl.org.pl

Received: 06 October 2010
Final form accepted: 28 February 2011
Published online: 12 March 2011

Volume 16 (2011) pp 258-263
DOI: 10.2478/s11658-011-0004-1

Short communication

GIDMP: GOOD PROTEIN-PROTEIN INTERACTION DATA METAMINING PRACTICE

DARIUSZ PLEWCZYNSKI[1]* and TOMAS KLINGSTRÖM[1,2]
[1]Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw, ul. Pawinskiego 5a, 02-106 Warsaw, Poland, [2]Master of Science Programme in Molecular Biotechnology Engineering at Uppsala University, Norbyvägen 14, 752 36 Uppsala, Sweden

Abstract: Studying the interactome is one of the exciting frontiers of proteomics, as shown lately at the recent bioinformatics conferences (for example ISMB 2010, or ECCB 2010). Distribution of data is facilitated by a large number of databases. Metamining databases have been created in order to allow researchers access to several databases in one search, but there are serious difficulties for end users to evaluate the metamining effort. Therefore we suggest a new standard, "Good Interaction Data Metamining Practice" (GIDMP), which could be easily automated and requires only very minor inclusion of statistical data on each database homepage. Widespread adoption of the GIDMP standard would provide users with:

- a standardized way to evaluate the statistics provided by each metamining database, thus enhancing the end-user experience;
- a stable contact point for each database, allowing the smooth transition of statistics;
- a fully automated system, enhancing time- and cost-effectiveness.

The proposed information can be presented as a few hidden lines of text on the source database www page, and a constantly updated table for a metamining database included in the source/credits web page.

Key words: Proteins, Interactome, Pathways, Signaling, Metamining, Literature curation, Protein-protein interaction, Bioinformatics, Systems biology

* Author for correspondence. e-mail: darman@icm.edu.pl

INTRODUCTION

The importance of giving proper credit to the source databases is recognized by all major data mining databases (APID [1], ConsensusPathDB [2], DASMI [3], MiMI [4], PathwayCommons [5] and UniHI [6]). APID, DASMI, MiMI and UniHI are focused only on protein-protein interactions (PPI), while ConsensusPathDB and PathwayCommons strive to create biological pathways containing PPIs. Any result found in these databases provides information about the source database and the link to the corresponding data web resource. This means that the credit system works perfectly when the query finds something. It is however possible that a query to a metamining database fails despite the interaction being available in one of the source databases mined by the metamining database. Unifying diverse sets of data with different accession numbers and IDs is a major challenge here. Therefore, each metamining database contains its own unique solution to this issue. One of the most important features of these solutions is the coverage of the unification method. The purpose of metamining databases is to minimize the time spent by researchers and bioinformaticians trying to find relevant data. A database with a low coverage means that there is a significant risk that 'no results' simply means that the metamining database failed to extract the data, or that the interaction does not fit within the scope of the metamining database. For a biologist researching a single protein, this means that they must still have to access all the databases to find non-integrated interactions, and a bioinformatician performing automated proteome-scale research is almost certainly going to miss a significant number of interactions. Knowing the coverage and the scope of the different databases is therefore of great importance when selecting a datamining service. Several databases (APID, ConsensusPathDB and UniHI) provide data on their homepage regarding the number of interactions they import from source databases. However, none of the databases provide data about how many interactions they did not manage to import successfully and how many they did not want to import,. This makes it unnecessarily difficult for the careful scientist to evaluate the reliability of their findings.

MATERIALS AND METHODS

Coverage varies widely between the databases and is decided by a combination of the topical requirements of the metamining database and its ability to accurately incorporate protein-protein interactions. Separating these two factors is currently a massive undertaking. By using the statistics on the respective homepage of APID, ConsensusPathDB and UniHI (all accessed 09-07-21) and the data provided by MiMI in their latest relevant article [4] we estimated their coverage of the IntAct [7] source database. According to the IntAct web site, IntAct currently includes 195,553 protein interactions (09-07-21). APID contains 155,746 interactions from IntAct, giving a coverage of 79.6%, ConsensusPathDB

contains 22,049 interactions and has a coverage of 11.2%, MiMI contains 77,780 interactions and has a coverage of 39.8%, and UniHI contains 19,404 interactions and has a coverage of 9.9%. The numbers can however vary drastically between the databases. MiMI finds 167,330 interactions in the BioGRID [8] database containing 166,002 non-redundant interactions, giving it a somewhat confusing coverage of 100.7%. APID, which contained the highest coverage in IntAct, this time, only reached coverage of 56.7% by containing 94,197 interactions from the interactions included in BioGRID. Obviously, finding a suitable metamining database currently requires a significant work effort and cannot be done in the cases of DASMI and PathwayCommons.

A closely related problem is the issue of synchronization. All databases except DASMI rely on centralized storage of extracted data (DASMI also does this for non-DAS compliant databases). Most metamining databases are not continuously synchronized. We will illustrate this with another example using BioGRID (current version 2.0.54, released 1-07-09). APID contains version 2.0.47, downloaded 08-12-10, ConsensusPathDB 2.0.54. DASMI searches BioGRID when the query is made. MiMI does not state this information on its homepage. PathwayCommons contains 2.0.49, downloaded 09-01-28. UniHI downloaded its version of BioGRID on 08-08-22 but does not state the version number. This should be compared with the fact that BioGRID on average adds 1620 new non-redundant interactions every month (based on the latest news period 09-03-01 to 09-07-01). Therefore, for every update not covered by the metamining database, the coverage has gone down by approximately 1% since BioGRID started carrying out monthly updates.

RESULTS AND DISCUSSION

To raise awareness of these issues and also more easily provide researchers with a way to comprehensively evaluate the quality of their data sources, we propose a simple standard called “Good Interaction Data Metamining Practice” (GIDMP). It requires a small degree of cooperation between databases, but we estimate the effort of following GIDMP to be roughly equivalent to answering five e-mails from concerned scientists regarding the quality of data. Each database following the GIDMP standard should include one page with the title Source Databases or SourceDB including a table containing the relevant statistical information (see Tab. 1 for an example of this) and other relevant information as deemed necessary by the independent databases.

The new standard is supported by two recent publications where we have encountered unnecessary difficulties due to the issues with database coverage and scope. The first one, from Briefings in Bioinformatics**,** “Protein-protein interaction and pathway databases, a graphical review” 2010 by T. Klingström and D. Plewczynski [11], is an extended review of the available protein-protein databases. It provides a valuable tool for researchers to reduce the time necessary to gain a broad overview of PPI databases and is supported by

Tab. 1. Status of APID with the assumption that all source databases are in their most recent versions extracted 2009-07-21.

The seven defined columns are divided into two groups. The first five are static data entered by the metamining database author. The following two rely on the corresponding output from the source databases and those are the two most important ones. The list of columns includes:
Name: Official name of the database in abbreviated form. **Link to resource**: Link to portal page in the case of database, link to describing article if the data are taken from a specific low-throughput project. **Description of resource**: Article describing the database or project from where data are taken from. **Notes**: Free text area available to explain any data- specific comments. **Version**: Tells the user when the extraction was made. Continuously updated databases only give the date of extraction, while databases such as Reactome or BioGRID provide both the version and the date. **Latest version**: Each source database should provide a code denoting when the last update was made. We suggest that the code should take the form of the date of the update, e.g. 20090721. In the source database this code should be updated whenever a new entry is made in the database. A simple script comparing the code on the metamining site with the code in the source database can then be run at certain time intervals. A mismatch in code causes "Latest version" to switch to "No". **Coverage**: Below the update code appears the number of non-redundant physical PPIs for each species covered by the source database. The metamining database can run a script which uses this number to calculate the coverage of the metamining database at the same time interval as "Latest version" is checked. Coverage is calculated of the coverage is done by the metamining database by dividing the number of integrated PPIs in the metamining database with the number of PPIs in the source database from the relevant species.

Name	Link to resource	Description of resource	Notes	Version	Latest version	Coverage
BIND	http://bond.unleashedinformatics.com	Bader *et al.* BIND the biomolecular interaction network database. **Nucleic Acids Res.** 31 (2003) 248-250		N/A	Yes	28.6%
BioGRID	http://www.thebiogrid.org/	Breitkreutz *et al.* The BioGRID interaction database: 2008 update. **Nucleic Acids Res.** 6 (Database issue) (2008) D637-D640		Version 2.0.54 2009-07-01	Yes	56.3%
DIP	http://dip.doe-mbi.ucla.edu/	Salwinski *et al.* The database of interacting proteins: 2004 update. **Nucleic Acids Res**. 32 (2004) D449-D451	Is updated without new versions	2009-07-21	Yes	59.5%
HPRD	http://www.hprd.org	Keshava Prasad *et al.* Human protein reference database: 2009 update. **Nucleic Acids Res.** 37 (Database issue) (2009) D767-D772	MiMI also extracts interactions from pathway data available in HPRD. Is updated without new versions	2009-07-21	Yes	76.8%
IntAct	http://www.ebi.ac.uk/intact/site/index.jsf	Kerrien *et al.* IntAct-open source resource for molecular interaction data. **Nucleic Acids Res.** 35 (Database issue) (2007) D561-D565	Is updated without new versions	2009-07-21	Yes	79.6%
MINT	http://mint.bio.uniroma2.it/mint	Andrew Chatr-aryamontri *et al.* MINT: the Molecular INTeraction database **Nucleic Acids Res.** 35 (Database issue) (2007) D572-574	Is updated without new versions	2009-07-21	Yes	117.8%

a graphical representation of data exchange. The graphical representation is made available in cooperation with the team maintaining www.pathguide.org and can be accessed at http://www.pathguide.org/interactions.php in a new Cytoscape web implementation. The second supporting publication is our recent article in Cellular & Molecular Biology Letters, "The interactome: predicting the protein-protein interactions in cells" 2009 by D. Plewczynski and K. Ginalski [9], where we reviewed computational methods for prediction of protein-protein interactions including consensus machine learning approaches [10], and compared them for their effectiveness in predicting protein-protein interactions using sequence and structure information.

Here, we encourage biologists to consider the issues of synchronization and coverage. Interaction metamining databases are an excellent tool for proteomics research but require a more careful evaluation than commonly thought. Therefore we also strongly encourage developers of metamining databases to produce a GIDMP table to support biologists in their evaluations. We believe that the GIDMP standard is a good compromise that takes into consideration both the need for transparency in the production of data, and the limited resources available for the metamining database teams.

Acknowledgments. This work was supported by Linnéska stipendiestiftelsen and the Polish Ministry of Education and Science (N301 159735).

REFERENCES

1. Prieto, C. and De Las Rivas, J. APID: Agile Protein Interaction DataAnalyzer. **Nucleic Acids Res.** 34 (2006) W298-302.
2. Kamburov, A., Wierling, C., Lehrach, H. and Herwig, R. ConsensusPathDB-a database for integrating human functional interaction networks. **Nucleic Acids Res.** 37 (2009) D623-D628.
3. Blankenburg, H., Finn, R.D., Prlić, A., Jenkinson, A.M., Ramírez, F., Emig, D., Schelhorn, S.E., Büch, J., Lengauer, T. and Albrecht, M. DASMI: exchanging, annotating and assessing molecular interaction data. **Bioinformatics** 25 (2009) 1321-1328.
4. Jayapandian, M., Chapman, A., Tarcea, V.G., Yu, C., Elkiss, A., Ianni, A., Liu, B., Nandi, A., Santos, C., Andrews, P., Athey, B., States, D. and Jagadish, H.V. Michigan Molecular Interactions (MiMI): putting the jigsaw puzzle together. **Nucleic Acids Res.** 35 (2007) D566-D571.
5. http://www.pathwaycommons.org/
6. Chaurasia, G., Iqbal, Y., Hanig, C., Herzel, H., Wanker, E.E. and Futschik, M.E. UniHI: an entry gateway to the human protein interactome. **Nucleic Acids Res.** 35 (2007) D590-D594.
7. Kerrien, S., Alam-Faruque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C., Dimmer, E., Feuermann, M., Friedrichsen, A., Huntley, R., Kohler, C., Khadake, J., Leroy, C., Liban, A., Lieftink, C., Montecchi-Palazzi, L., Orchard, S., Risse, J., Robbe, K., Roechert, B., Thorneycroft, D., Zhang, Y.,

Apweiler, R. and Hermjakob, H. IntAct-open source resource for molecular interaction data. **Nucleic Acids Res.** 35 (2007) D561-D565.
8. Breitkreutz, B.J., Stark, C., Reguly, T., Boucher, L., Breitkreutz, A., Livstone, M., Oughtred, R., Lackner, D.H., Bahler, J., Wood, V., Dolinski, K. and Tyers, M. The BioGRID Interaction Database: 2008 update. **Nucleic Acids Res.** 36 (2008) D637-D640.
9. Plewczynski, D. and Ginalski, K. The interactome: Predicting the protein-protein interactions in cells. **Cell. Mol. Biol. Lett.** 14 (2009) 1-22.
10. Plewczynski, D. Brainstorming: weighted voting prediction of inhibitors for protein targets. **J. Mol. Model.** (2010) in press.
11. Klingström, T. and Plewczynski, D. Protein-protein interaction and pathway databases, a graphical review. **Brief. Bioinform.** (2010) in press, DOI: 10.1093/bib/bbq064.