Protein–protein interaction and pathway databases, a graphical review

Tomas Klingström and Dariusz Plewczynski

Submitted: 13th July 2010; Received (in revised form): 14th August 2010

Abstract
The amount of information regarding protein–protein interactions (PPI) at a proteomic scale is constantly increasing. This is paralleled with an increase of databases making information available. Consequently there are diverse ways of delivering information about not only PPIs but also regarding the databases themselves. This creates a time consuming obstacle for many researchers working in the field. Our survey provides a valuable tool for researchers to reduce the time necessary to gain a broad overview of PPI-databases and is supported by 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 local copy of Cytoscape cys file can be downloaded from http://bio.icm.edu.pl/~darman/ppi web page.

Keywords: proteins; interactome; pathways; signalling; metamining; literature curation; protein–protein interaction; binary-interactions

INTRODUCTION
Understanding the organization of genetic networks and protein pathways to establish how they contribute to cellular and organism phenotypes is one of the major challenges in the post-genomic era [1]. The protein–protein interaction (PPI) community has been characterized by a wide and open distribution of proteomic data [2] through the collection of PPI and pathway databases. The ability to distribute and share data between various research groups has resulted in a large number of different source databases. However, the general overlap between those databases is very limited [3, 4], which means that a common procedure for researchers is to unify these diverse data sets to support their own work [5–8]. Several metamining databases have been created that perform such unification. This has lead to the spontaneous development of a network of data exchange between literature curated databases, metamining databases and databases generating predicted PPIs. The exchange of information is supported by three major data exchange formats: BioPAX [9], PSI-MI [10] and SBML [11]. An extensive review of these data formats has already been published by others [12], and a more detailed description has therefore been omitted from this article. Previous reviews regarding PPI and pathway databases have largely focused on the potential of using databases to support drug development [13, 14] or cover the early development of literature curated and pathway databases [15, 16].

A single illustration containing all databases would create an extremely complicated image clouded by a large amount of superfluous information. We have...
therefore created a set of separate visual networks defined by our own classification scheme. PPI source databases are databases that conduct in-house curation of data, this class includes both pure source databases and databases combining in-house curation with metamining of other databases. Pathway source databases are databases creating networks, or pathways, giving information not only regarding the PPI but also its biological context. Metamining databases are databases exclusively relying on source databases and high-throughput experiments published as finished data sets. In our classification we have decided to combine both PPI metamining databases and metamining pathway databases in order to allow readers to easily compare their coverage of available source databases. A separate category is predictive databases that use interaction data to computationally predict previously unknown interactions. In this class, we have also decided to incorporate databases (HAPPI and STITCH) that mainly rely on predicted data but do not make predictions on their own. Furthermore a visualization of support for BioPAX, PSI-MI and SBML is included in the online visualization available at http://pathguide.org. A complete list of databases covered in this article and their classification can be found in Supplementary Data 1. A brief summary of each database can be found in Supplementary Data 2.

CREATION OF THE NETWORK
Selection of data and databases
The initial selection of investigated databases was made from the index of databases kept at Pathguide (http://www.pathguide.org/) and based on their popularity ranking. Most databases have at least one article describing them in the annual database issue of Nucleic Acids Research (NAR) journal. Articles published in this annual issue are obligated to contain a section comparing it to other similar databases. Relevant databases mentioned in those articles have also been added to complement the initial selection from Pathguide.

Visualization of network
Visualization of databases interactions was carried out by the open source visualization program Cytoscape [17]. Each database is represented as a node and interactions with other nodes are visualized as an edge. Nodes are also connected to Pathguide IDs which provide linkage to a Pathguide summary of the functionality of the database and a direct link to the database. Collected data is included in a single CYS file (the file format used by Cytoscape) that is freely available upon request from authors for further development. Databases with the default (beige) colour have not been specifically investigated, blue nodes are interaction databases, yellow nodes are predictive interaction databases, green nodes are pathway databases, red nodes represent metamining databases and grey nodes are databases or standards, allowing the user to access several different databases through a standardized entrance point.

We have three different shapes for edges. An arrow means that the database pointed to is mining data from the node connected by the edge. A diamond means that the database pointed to is being mapped from the other database, i.e. a database entry in the database pointed ‘to’ can be found by searches in the database pointed ‘from’. A straight line shows that there is an exchange agreement between the databases, meaning that data can go between the databases in both directions.

Future development
All data is stored in a single CYS library. By using a controlled vocabulary, we are able to easily incorporate new data to the already existing maps. A web implementation of Cytoscape network is available Pathguide (http://www.pathguide.org/interactions). A simple html submission system will allow users to submit updates to the network.

Interaction databases
The topical databases such as DroID (PPIs in Drosophila melanogaster), MatrixDB (extracellular PPIs), InnateDB (PPIs in the immune system) and MPIDB (PPIs in microbes) combine datamining from other source databases with their own curation efforts (Figure 1). The combination of extensive datamining and in house curation make interaction databases an appealing choice for researchers interested in the topical area of the database.

BioGRID imported the HPRD and Flybase databases in 2006 [18], but have not added any more data from other databases since then. BIND, DIP, HPRD, IntAct and MINT do not incorporate data from other databases. Overlap between most source databases is currently small [3, 4]. Each literature curating database extracts its data from a set of selected scientific journals, meaning that shared interactions can only occur if two databases curate...
the same journal or the same PPI is mentioned in two different journals. Searches in several databases are therefore complementary in the sense that each new search provides access to a large set of interactions that were not previously searched.

All the major literature curation databases except HPRD are affiliated with the IMEx consortium (http://imex.sourceforge.net/). The first step of this cooperation is the joint effort of literature curation, leading to continuous increase of shared data in the databases. Literature sources for curation are divided among the IMEx members and curated in an IMEx standardized format, curated data is then made available to all IMEx members.

Metamining databases
APID, MiMI and UniHI are metamining databases with the mission to unify source databases into a single comprehensive source meta-database (Figure 2). APID is dedicated completely to experimental data from interaction databases, and incorporates data from BioGRID, BIND, DIP, HPRD, IntAct and MINT.

UniHI incorporates experimental and computationally predicted human PPI data. Apart from the literature curated databases BioGRID, BIND, DIP, HPRD and IntAct they also include data from Reactome, high-throughput Y2H experiments (MDC-Y2H [19] and CCSB [20]), searches for proteins mentioned together in Medline abstracts (COCIT [21]), and predictions of orthologous proteins in model animals predicted in humans (OPHID [22], HomoMint [23] and ORTHO [24]).

MiMI integrates data from BioGRID, BIND, DIP, HPRD, IntAct and MINT, and it also adds interactions from KEGG and Reactome pathway

Figure 1: Interaction databases. Among PPI source databases we see three distinct patterns of data exchange. Stand alone database who are not actively sharing data (BIND, HPRD and to a certain extent BioGRID), generalist databases participating in the IMEx agreement (DIP, IntAct and MINT) and topical databases who combine their own in-house curation with metamining of other databases (DroID, InnateDB, MatrixDB and MPIDB). The combination of in-house curation and extensive metamining make the topical databases an attractive one stop choice for researchers focused on *D. melanogaster* (DroID), the immune system (InnateDB), extracellular proteins (MatrixDB) or physical microbial interactions (MPIDB). IMEx databases only share data curated in accordance with the IMEx curation rules. This mean that both non-sharing and IMEx databases are likely to contain a large proportion of entries unique to the database unless the IMEx exchange is combined with a metamining process (like in the case with the topical databases). Researchers unwilling to spend a significant amount of time on querying generalist source databases may therefore find it convenient to use metamining databases despite the issues described in the metamining databases section.
Figure 2: Metamining databases. The view of metamining databases is obscured by the fact metamining databases in several cases not only extract data from PPI databases but also extract data from other kinds of databases to provide supporting information. The view does however clearly show three distinctive approaches to metamining. APID, MIMI, PINE and UniHI are traditional PPI metamining databases in the sense that they periodically download data from source PPI databases to create a centralized repository. ConsensusPathDB and Pathways Commons provide a similar service aimed at integrating different pathway databases. ConsensusPathDB here provide access to a far larger number of source databases and also makes use of PPI databases to create its own PPI based pathways. DASMI is unique in that it maps to several databases rather than datamining them. This mean that DASMI makes it own queries to the databases mapped to each time a user enters a query into DASMI. Therefore DASMI is always up to date with the most recent version of the source databases being mapped. But also less suited for large queries where bandwidth limitations may be an issue.
that each meta-database has its own scope and therefore often only contain parts of the source databases.

A search in for example APID is not equal to querying BIND, BioGRID, DIP, IntAct and MINT directly. This can cause serious issues with reproducibility, if the usage of metamining resources is not clearly mentioned in the report. Researchers conducting experiments similar to the original experiments might find themselves with different sets of data compared to the original author, thereby causing serious doubts about the integrity of the results presented by the original author.

Each metamining database contains its own unique algorithm for unification, with some strengths and weaknesses. The most easily measured factor is the coverage of the different source databases. Only APID and ConsensusPathDB provide a web site access to the number of interactions which are taken from their sources. MiMI and UniHI provide similar data, but in their most recent publication [34, 35]. PathwayCommons does not provide public access to this kind of data and is therefore omitted from Table 1.

Table 1 shows the current coverage of the metamining databases. The coverage of a database

<table>
<thead>
<tr>
<th>Database</th>
<th>BIND</th>
<th>BioGRID</th>
<th>DIP</th>
<th>HPRD</th>
<th>IntAct</th>
<th>MINTa</th>
</tr>
</thead>
<tbody>
<tr>
<td>APID</td>
<td>45 276</td>
<td>28.6%</td>
<td>94 197</td>
<td>89.2%</td>
<td>34 235</td>
<td>59.5%</td>
</tr>
<tr>
<td>MiMI</td>
<td>233 201</td>
<td>147.5%</td>
<td>167 330</td>
<td>159%</td>
<td>49 677</td>
<td>86.1%</td>
</tr>
<tr>
<td>UniHI</td>
<td>7394</td>
<td>4.7%</td>
<td>24 624</td>
<td>23.3%</td>
<td>1397</td>
<td>2.42%</td>
</tr>
<tr>
<td>ConsensusPathDB</td>
<td>0</td>
<td>0%</td>
<td>27 044</td>
<td>22.8%</td>
<td>1459</td>
<td>2.53%</td>
</tr>
<tr>
<td>APID</td>
<td>45 276</td>
<td>28.6%</td>
<td>94 197</td>
<td>89.2%</td>
<td>34 235</td>
<td>59.5%</td>
</tr>
<tr>
<td>MiMI</td>
<td>233 201</td>
<td>147.5%</td>
<td>167 330</td>
<td>159%</td>
<td>49 677</td>
<td>86.1%</td>
</tr>
<tr>
<td>UniHI</td>
<td>7394</td>
<td>4.7%</td>
<td>24 624</td>
<td>23.3%</td>
<td>1397</td>
<td>2.42%</td>
</tr>
<tr>
<td>ConsensusPathDB</td>
<td>0</td>
<td>0%</td>
<td>27 044</td>
<td>22.8%</td>
<td>1459</td>
<td>2.53%</td>
</tr>
</tbody>
</table>

Each source database is divided into two columns. The first column show how many protein-protein interactions the metamining database has extracted. The second column describes the coverage of the database, i.e. the number of interactions extracted divided by the number of interactions reported on the source database webpage (03 August 2009). Values >100% where initially thought to be caused by the expansion of complex interactions according to either the matrix model or the spoke model [32]. According to the spoke model a bait-protein interacting with a complex consisting of three proteins is expanded into three interactions, one interaction with each one of the proteins in the complex. Expansion by the matrix model is similar to the spoke model, but also adds interactions between the proteins participating in the protein complex. Communication with developers of the metamining databases have, however, shown this theory to be incorrect since MiMI, UniHI and ConsensusPathDB does not expand protein complexes imported from other databases (Glenn Tarcea, Matthias Futschik and Atanas Kamburov, personal communication).

APID uses the spoke model to expand protein complexes from source databases when applicable (Carlos Prieto, personal communication). But entries in MINT are already expanded according to the matrix model which mean that any expansion would be redundant and add 0 interactions. Further investigation by Glenn Tarcea at MiMI solved the issues with the MiMI statistics but could not provide further insight in the issues among the other databases. In the case of MiMI, it turns out that the raw identifier counter providing data for their article [31] was faulty and will in the future be replaced by a ‘gene – gene’ interaction counter (Tarcea, G. personal communication). Please note that this table is not a quality assessment of databases, most meta-databases only extract data fulfilling certain criteria and are therefore expected to have <100% coverage. aMINT previously expanded interactions according to the matrix model. On 01 August 2009 MINT contained 11 518 interactions. All databases use data downloads from before this time. We therefore use this older number in the table. bDoes not use MINT, but does use HomoMINT.

A databases of human PPIs predicted from orthology with proteins in other species. UniHI includes 10 174 interactions from this database and thus a coverage of 41.63%. The number of non-redundant protein-protein interactions in each database is (webpages accessed 18 November 2009): BIND 158 149, BioGRID 105 593, DIP 57 683, HPRD 38 806, IntAct 201 652, MINT 83 323 (please see footnote ‘a’) and HomoMINT 24 439 interactions (only used by UniHI). BioGRID also contain genetic interactions. MiMI is the only database in the table that also incorporates this kind of data which mean that the total number of interactions in BioGrid (169 723) is a more fitting comparasion.
is calculated as the proportion of PPIs in a source database that have been imported into the metamin-
ing database. We would like to stress that coverage is not a quality estimation of databases since each meta-databases has its own scope which does not always correspond to the scope of the source databases. Of the meta-databases, MiMI is the most non-discriminatory database, aiming to include all interaction data in one single comprehensive database and therefore incorporating all interactions they can find. APID is focused on experimentally verified proteins and relies on a unifying algorithm that uses a protein sequence as an identifier, meaning that interactions not supported by an experiment or containing a protein with an unknown sequence are omitted. UniHI focuses on human PPIs extracted from databases, thereby giving it a low coverage of databases hosting data from other species. ConsensusPathDB is a pathway database requiring more contextual biological information surrounding an interaction, thus greatly limiting the number of interactions it can integrate from each source database.

More worrying is that there are several instances of meta-databases reporting that they have extracted >100% of the source interactions. Of the databases in the table only APID expands imported complexes (Carlos Prieto, Glenn Tarcea, Matthias Futschik and Atanas Kamburov, personal communication) and we have so far been unable to find other satisfactory explanations.

PINA [36] is a network analysis platform with a built in data set of data from MINT, IntAct, DIP, BioGRID, HPRD and MIPS/MPACT. PINA also allow users to upload their own private PPI data sets to create custom meta-databases. This combined with the advanced query and analysis tools make PINA suitable for researchers who wish to perform PPI network analysis on specific pathways or to create novel pathways based on known PPIs.

DASMI [37] is an alternative approach to the integration of protein and domain interaction data. Rather than collecting all data into a single database, this extension of the Distributed Annotation System (DAS) [38] accesses all included databases and then collects the data upon request from the user. Decentralized approaches are therefore less suited than metaminining databases for large-scale analysis, such as genome-wide interaction studies, as huge amounts of data have to be transferred for each query. This approach heavily depends on the participation of the interaction database providers and thus far has been only partially successful. Several databases have been integrated in this way, yet in other cases the DASMI developers have been forced to perform static integration of data and host those resources on a central server. Temporary caches are updated when the databases are updated (like HPRD and DIP), or final data sets of finished projects (like Sanger and CCSB-HI1) are used. DASMI data sources are programmatically accessible, which has led to the development of different user clients that present the interactions in different ways [37, 39]. It remains to be seen, whether a growing support from client developers will convince more database providers to offer a DAS-based interface to their data, as this would increase their visibility.

Predictive interactions databases

There are four major predictive PPI databases: HAPPI, STRING, STITCH and Scansite. Both HAPPI and STITCH are directly reliant on STRING meta-database to find their predicted interactions (Figure 3). STITCH novel features come from the fact that it also provides predictions regarding protein–chemical interactions.

STRING combines known interaction data from interaction databases BIND, BioGRID, DIP, IntAct MINT and HPRD with interactions from the pathway databases PID, Reactome, KEGG and EcoCyc. To further supplement this database it also includes interactions predicted by algorithms specifically made for STRING [40, 41].

HAPPI extracts the human interactions from STRING and makes a comparison with data found among main sources: BIND, HPRD, KEGG, MINT and OPHID. Each source has its own confidence value and the reporting of an interaction in several databases increases the confidence score. This yields a database of more than 138 000 non-redundant high-quality PPIs, in HAPPI terminology referred to as 3, 4 or 5 star-quality interactions. Lower quality predictions are also kept in the database marked as 1 or 2 star interactions bringing the total number to >1.2-million ‘predicted’ interactions.

The Scansite database is based on a 1D sequence comparison with known motifs. Scansite was created to identify short protein sequences likely to be recognized by modular signaling domains,
phosphorylated by protein Ser/Thr- or Tyr-kinases or mediate specific interactions with protein or phospholipid ligands [43].

Pathway databases
Pathway databases put protein interactions into a biological context by creating pathways to describe biological processes (Figure 4). Some databases such as KEGG track a direct relationship of protein interactions while others such as Reactome base their pathways on biochemical transition and transportation rather than the protein interactions.

Integration of PPI data from PPI databases is uncommon, of the investigated databases only SPIKE reports such integration into their database to create their pathways [44]. But ConsusPathDB and recent work by Wu et al. [45] prove that despite the differences it is possible to combine data sets of both kinds to increase proteome coverage but still maintain the quality of functional PPIs. It should however be noted that much useful information, such as semantic relationships between proteins, is lost during the merging (Guanming Wu, personal communications).

Exchange of data has been hindered by the diverse ways of defining pathway data but is increasing. There are currently two major data standards for pathway data exchange: BioPAX [9] and SBML [11]. Each one having its own distinct advantages, SBML provides support for modelling and simulations while BioPAX creates a rich hierarchy with a strong degree of flexibility [12]. Many databases therefore support both formats with BioPAX being the dominant one regarding data exchange due to its capability of exchanging many semantic contents.

The BioPAX data exchange format is currently used by several databases for data exchange among pathway databases. NCI-PID mines BioCarta and Reactome through BioPAX level 2 transfers. Reactome have extracted data through BioPAX level 2 from NCI-PID, the Cancer Cell Map and is working on integrating RiceCyc [46] (Guanming Wu, personal communication). KEGG data has
instead been imported in KEGG's own format, KGML and NCI-PID have been integrated through SBML with Cell-Designer additions for PANTHER. These efforts show that the ability to exchange data between pathway databases have been greatly enhanced through the introduction of the BioPAX and SBML formats.

Unifying efforts
Apart from DASMI (covered in metamining databases) there are two other services that can provide a single point of access to a wide variety of databases, InterPro and BioMart (Figure 5). InterPro has recently launched their own BioMart service enabling users to access InterPro through the BioMart program. BioMart is similar to DASMI, in that both aim to provide a single portal that allows access to all decentralized source databases. BioMart is aimed at general description of proteins rather than PPIs, but Reactome includes its own BioMart service. Therefore using BioMart one can access InterPro and further get access to PANTHER and mapping to IntAct, as a part of the services offered by InterPro.

DISCUSSION
The lack of a commonly agreed upon definition of binary interactions is a source of major confusion in research regarding PPIs [47, 48]. Cusick et al. [47] attribute these issues to a lack of transparency on the behalf of the source databases who do not clearly enough describe the nature of their source data. We believe that the problem rather is caused by each database having its own unique way of delivering this information. Some databases put all the necessary information in an article submitted to Nucleic Acids Research (NAR) database issue, some present their information online and others require the inquirer to E-mail the database team for answers. The result is that each individual database is easy to understand in purpose and content. But gaining an overview of several databases is almost impossible to do within a reasonable time frame.
Figure 5: Unifying efforts. Both BioMart and DASMI aim to create a single point of access to a multitude of databases. The systems rely on mapping methods as the client sends queries to multiple servers and assemble the returned information into a comprehensible view. The BioMart system does not support exceptions to this method while DASMI often download databases of interest and makes the data available from servers maintained by the DASMI team. InterPro provides a one-stop shop for protein classification by creating protein sequence signatures from its member databases and store them in its own database. This protein classification scheme is also available through BioMart since InterPro now supports queries made from BioMart clients.
The effect of these inconsistencies can also be seen in the confusion regarding the size of the human interactome [49–51]. Venkatesan et al. [51] estimates the size to $\sim130\,000$ binary interactions, Hart et al. [49] to $154\,000–369\,000$ interactions and Stumpf et al. [50] to $\sim650\,000$ interactions. Closer inspection reveals that each team has defined its own search space as the human interactome, Venkatesan et al. [51] use the most restrictive definition and only include binary physical interactions. Hart et al. [49] use in-house experimental data obtained by TAP-MS to create its source networks which means that proteins belonging to the same protein complex are also considered to be interacting, thus increasing the size of their defined interactome. Stumpf et al. [50] rely on a combination of yeast two hybrid (Y2H) derived data sets and literature curated data from DIP [52] and IntAct [53]. Data obtained by Y2H experiments will, due to experimental constraints, follow the definition used by Rual et al. [20] while literature curated databases uses a more flexible definition of ‘interaction’ [48] that also considers non-physical functional interactions to be a form of interaction.

This situation is not likely to improve unless a conscious effort is made to solve these issues. One of our initial theories was that it is likely that the PPI community would undergo a consolidation similar to how almost all protein information now can be accessed through Uniprot.org. A quick glance at the IMEx consortium also supports this idea but non-IMEx compliant legacy records mean that each database will contain a mixture of unique and shared entries (Arnaud Ceol, Lukasz Salwinski and Sandra Orchard, personal communication). The PSIQUIC standard will make it possible to make a single search query to access each database (Aranda et al., manuscript in preparation) but each database will still host its own unique legacy records. These legacy records combined with the slightly different scope of each database mean that it is unlikely that we will see a quick reduction of the number of available databases. Among pathway databases this consolidation is even further away since each single database has a unique way of creating and defining its pathways. Therefore it is unlikely that transparency is going to increase by a reduction of the number of sources. This means that we should focus on increasing transparency by presenting the available information in a more uniform way.

At the moment there are two major sources of database information, the annual NAR database issue and the webpage of each database. NAR articles are indispensable as a tool when assessing a database but the lack of article standardization makes it highly time consuming to carefully read articles and still need to contact the authors with further questions. Therefore it would greatly increase the perceived transparency of databases if a more consistent article layout is required in the database issue. Likewise a voluntary effort by database teams to present a webpage which clearly describes the scope and content of the database would make it easier for users to properly assess the database. A more clear presentation of database methodology may also provide a major advantage in attracting other research groups to conduct experimental or bioinformatical research based on the database data. By clearly presenting data the database becomes a more attractive source for information to use in projects like the prediction of new protein interactions [54], cross-validation to improve the quality of currently available data sets [42] or the creation of predictor models used to increase the economics of high throughput experiments [55]. Making the presentation of different databases more uniform is likely to prove a cheap and efficient way to make PPI databases more attractive partners for researchers active in other fields.

**SUPPLEMENTARY DATA**

Supplementary data are available online at http://bib.oxfordjournals.org/.

**Key points**

- The Cytoweb visualization of PPI databases is a convenient to identify independent PPI resources (available at: http://www.pathguide.org/interactions.php).
- Querying metamining databases is not equivalent to independently querying the source databases independently.
- The term ‘protein–protein interaction’ is ambiguous and can refer to direct physical interactions, membership of the same protein complex or functional (indirect) interactions.

**Acknowledgements**

We would like to thank all the researchers who have taken the time to personally answer our questions. We would also like to thank Hagen Blankenburg for taking the time to extensively explain DAS to us and Jonas Klingström for his linguistic advice.

**FUNDING**

Linnéska stipendiestiftelsen, EC OxyGreen (KBBE-2007-212281) 6FP projects as well as the Polish
Ministry of Education and Science (N301 159735, N518 409238 and others).

References


