BiDIP: a Biological Data Integration Platform for Transcriptome Analysis

Junho Park[1], Min-Ji Kim[2], Eung-Hee Kim[1], Sungkwon Yang[2], Sungin Lee[2], Jin-Muk Lim[2], Hyunwhan Joe[1], Kyung-Sik Ha[1], and Hong-Gee Kim[1]

Biomedical Knowledge Engineering Lab. (BiKE),
Seoul National University, Seoul, Korea
[1]{naon,eungheekim,hyunwhanjoe,hakyung,hgkim}@snu.ac.kr
[2]{mingmaroo,syang0531,sunginlee,bikeljm}@gmail.com

Abstract. Many studies aimed to construct an automated gene expression analysis platform for researchers. However, they lack an integrated data model for analyzing heterogeneous data. In order to address this issue, we created a biological data integration platform for transcriptome analysis (BiDIP) for managing various kinds of databases. As part of this platform, we developed a biological interaction data model (BIM). Also we provide a Web application and OpenAPIs that allow users to search for connections among multiple databases conformant to the proposed data model. In this paper, we will present the current status of the platform as well as its future research venues.

Keywords: Biological data integration, Biological ontology, Interaction data model, Gene expression analysis, Transcriptomics

1 Introduction

Gene expression analysis has been a valuable research venue that provides for explanatory power to specific biological phenomena, given a biological context for a patient's histology or medicinal intervention. With the advent of microarray technology, transcriptome analysis has greatly overcome the limitations resident in studying individual genes, which offers fragmented and insufficient explanation for biological phenomena. Microarray technology enables us to measure and identify compositions of transcriptome under certain biological conditions. It, however, poses a significant challenge to biologists, especially those who are not well-versed in using the technology. Seemingly small, but still challenging to biologists, was the necessity for adequate amount of knowledge in coding required for data processing and statistical analysis of the data gained from microarray experiments. Almost 10 years after the introduction of the technology, we have acquired standard protocols for microarray analysis, though biologists still quite often need helping hands of computer experts or of companies that provide experiment tool kits.

We aimed to develop a platform that facilitates transcriptome analysis called Online Transcriptome Analysis Pipeline (OnTAP). Two major components constitute OnTAP: a Biological Data Integration Platform (BiDIP) and analysis modules including data pre-processing. There are four sets of analysis modules: gene set enrichment analysis, gene network analysis, drug target prioritization analysis, and miRNA target prioritization analysis.

BiDIP consists of four main components: 1) a comprehensive database model, BIM (Biological Interaction data Model) that encompasses 4 types of biological databases; 2) 4 integrated databases, which deal with PPI (Protein-Protein Interaction) databases, DGI (Drug-Gene Interaction) databases, microRNA databases, and pathway databases; 3) BiDIP browser, and 4) OpenAPI. BiDIP provides a unified view on various biological databases facilitating and streamlining transcriptome analysis, alleviating some burden off biology researchers. This paper presents OnTAP in general, and BiDIP in particular.

2 Related Work

Studies abound that aimed to provide streamlined services for microarray analysis: ArrayPipe [1], Expression Profiler [2], L2L microarray analysis tool [3], RACE [4], WebArray [5], MIDAW [6], CARMAweb [7], Asterias [8], MAGMA [9], GEPAS [10], EzArray [11], ArrayMining [12], MAGNET [13] and GALAHAD [14]. The most common services of these studies are data pre-processing, data normalization, gene name conversion, gene or sample clustering, gene set enrichment analysis, and data visualization. It should be noted that more recent studies used a higher volume of data, offered more analysis services, such as network-based analysis [13] and drug-focused analysis [14].

The notable drawbacks of the previous studies are: 1) limited coverage/volume of data - it is arguably true that the higher volume of, and wider coverage of, data will plug the gaps in our knowledge for biological phenomena; 2) absence of a comprehensive and inclusive data model by which to view and analyze heterogeneous databases.

3 BiDIP

This section presents detailed explanation on BiDIP, with focus on Biological Interaction data Model (BIM), integrated databases, Web-based browser, and Open API.

3.1 Snapshot of BiDIP

Fig. 1 shows the overall picture of OnTAP. In order to enable analysis on gene set enrichment, gene network, drug target prioritization, and miRNA prioritization, four types of databases are required: biological pathway, PPI (Protein-Protein interaction), DGI (Drug-gene Interaction), and MGI (miRNA-gene interaction). BiDIP provides in the main data groundwork for such analyses.

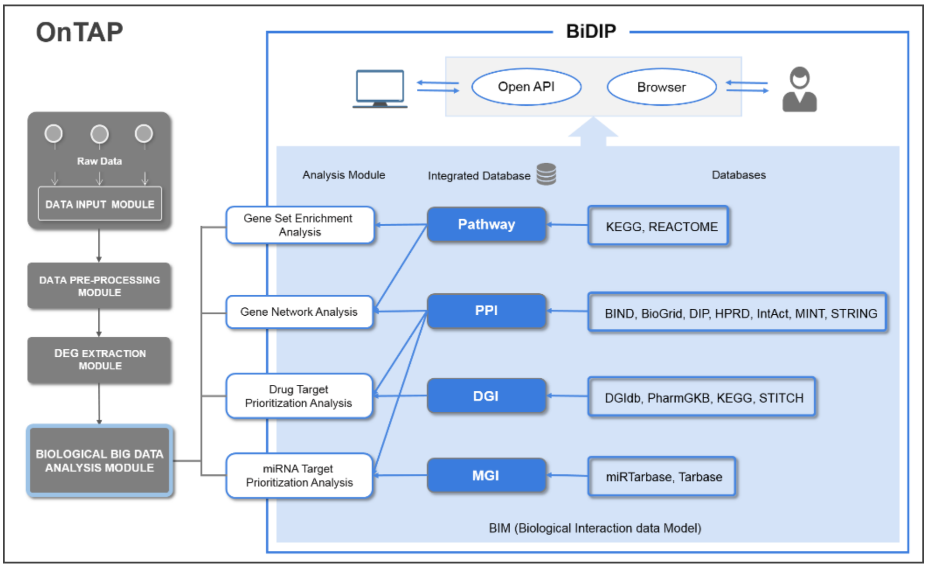

OnTAP
BiDIP
Open API
Browser
Raw Data
DATA INPUT MODULE
DATA PRE-PROCESSING MODULE
DEG EXTRACTION MODULE
BIOLOGICAL BIG DATA ANALYSIS MODULE
Analysis Module
Integrated Database
Databases
Gene Set Enrichment Analysis
Gene Network Analysis
Drug Target Prioritization Analysis
miRNA Target Prioritization Analysis
Pathway
PPI
DGI
MGI
KEGG, REACTOME
BIND, BioGrid, DIP, HPRD, IntAct, MINT, STRING
DGIdb, PharmGKB, KEGG, STITCH
miRTarbase, Tarbase
BIM (Biological Interaction data Model)

Fig. 1. The overall picture of OnTAP

3.2 Biological Interaction Data Model (BIM)

The core relations that BIM is designed to express are gene-gene, gene-drug, gene-miRNA relations, and pathway data. As shown in Fig. 2, the top classes in BIM are *process*, *interaction* and *biochemical entity*. *Biochemical entity* represents objects involved in a biological interaction. An *interaction* connects two biochemical entities. A sequence of interactions makes a *process*. In a *process*, the order of the interactions carries significant meaning. Fig. 3 shows a more detailed overview of BIM.

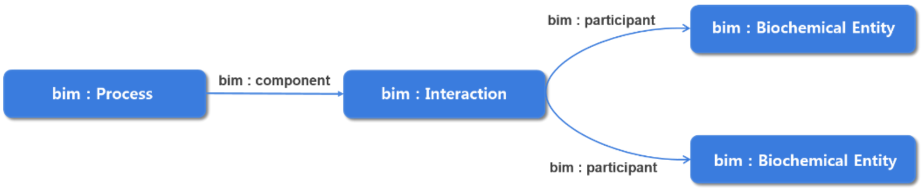


Fig. 2. Top-level Classes of BIM

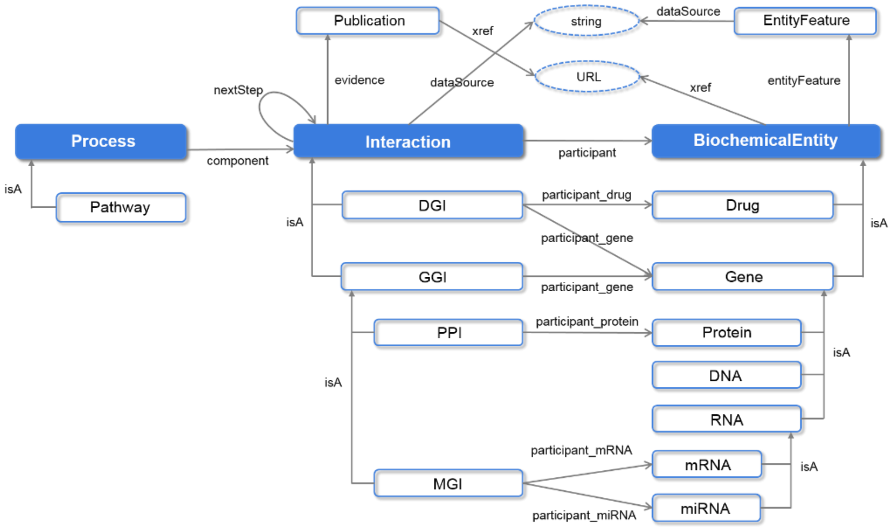


Fig. 3. Detailed View on BIM

3.3 BIM classes

Biochemical entity contains two main subclasses: *gene* and *drug*. *Gene* is defined as a biological region with a specific function. *Gene* has three subclasses: 1) *DNA*-specific region that will become a protein or functional RNA, 2) transcribed *RNA* from DNA-specific region, 3) translated *protein* from transcribed RNA. *RNA* has two subclasses: *miRNA* and *mRNA*. *Drug* represents any material used to treat or prevent diseases. In transcriptome analysis, the distinction between gene and its subclasses has not been used in general. However, we have made that distinction to reflect biological reality, which allows BIM to link with other ontologies that require such distinction. *Biochemical entity* is linked to *EntityFeatures* class with properties expressing full-name, genomic region, function annotation, and molecular weight.

Interaction has subclasses representing various unions of biochemical entities such as Gene to Gene Interaction (*GGI*) and Drug to Gene Interaction (*DGI*). *GGI* has miRNA to Gene Interaction (*MGI*) subclass, and Protein-Protein Interaction (*PPI*) subclass. *GGI* has two *gene* participants, *MGI* has *mRNA* and *miRNA* participants, *PPI* has two *protein* participants and *DGI* has *gene* and *drug* participants. *Interaction* also has a link to *Publication* class, each object of which is identified by PubMed id (PMID). *Process* has a subclass *pathway* representing a sequence of *GGI*.

3.4 BIM properties

A *process* has a 1: N relationship with its interactions and each *interaction* has 2 biochemical entities using a participant property. *Participant* property has two sub properties: *participant gene* and *participant drug*. *participant gene* has three sub properties:

participant_mRNA, *participant_miRNA*, and *participant_ptn*. Each *interaction* has a *datasource* property to represent the origin of database. The *xref* (external reference) property is used in two cases 1) to link *publication* objects to publication url 2) to link a biochemical object to an external site for more information. To express a sequence in a *pathway* we used the *nextstep* property, as shown in Fig. 5.

3.5 Integrated Databases

Biological databases are integrated based on BIM as shown in Table 1. We extracted human species (*Homo sapiens*) records from the databases and then converted Gene ID to official Gene Symbol by the HGNC standard.

Table 1. Integrated Databases

Database Type	Data Source	# of instances
Pathway	KEGG [15], Reactome [16]	833
Drug-Gene Interaction	KEGG, DGIdb [17], PharmGKB [18], STITCH [19]	793,405
Protein-Protein Interaction	BIND [20], BioGrid[21], DIP [22], HPRD [23], IntAct [24], MINT [25], STRING [26]	4,912,839
miRNA-Gene Interaction	miRTarBase [27], TarBase [28]	1,059,998

3.6 BIM-based Representation Example

This section presents an example of how a gene is expressed in BIM. An ion channel gene TRPC1 is used as an example in Fig. 4. TRPC1 is linked to Protein STIM1 with PPI (PPI 307419), to miR-124-3p with MGI (MGI 235456), and to Dantrolene with DGI (DGI 51715). Each interaction instance has a *datasource* property and an *xref* property. Fig. 5 shows a serotonergic synapse pathway that includes TRPC1. A *pathway* instance has more than one *GGI* component and a *GGI* instance has 2 participating genes. The order of each *GGI* instance is determined by using the *nextstep* property to represent the pathway flow.

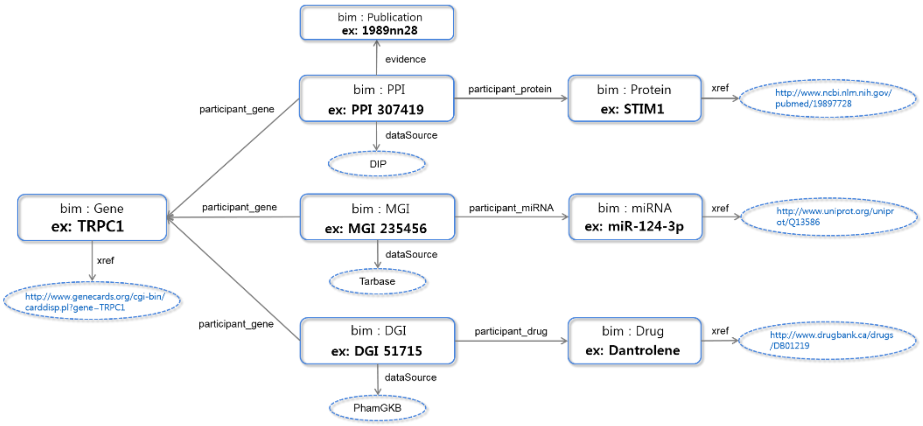


Fig. 4. BIM Representation of Entities Participating in PPI, MGI, and DGI of TRPC1

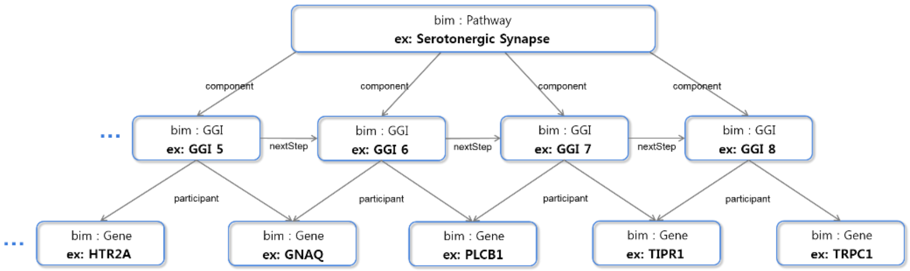


Fig. 5. BIM Representation of Serotonergic Synapse Pathway including TRPC1

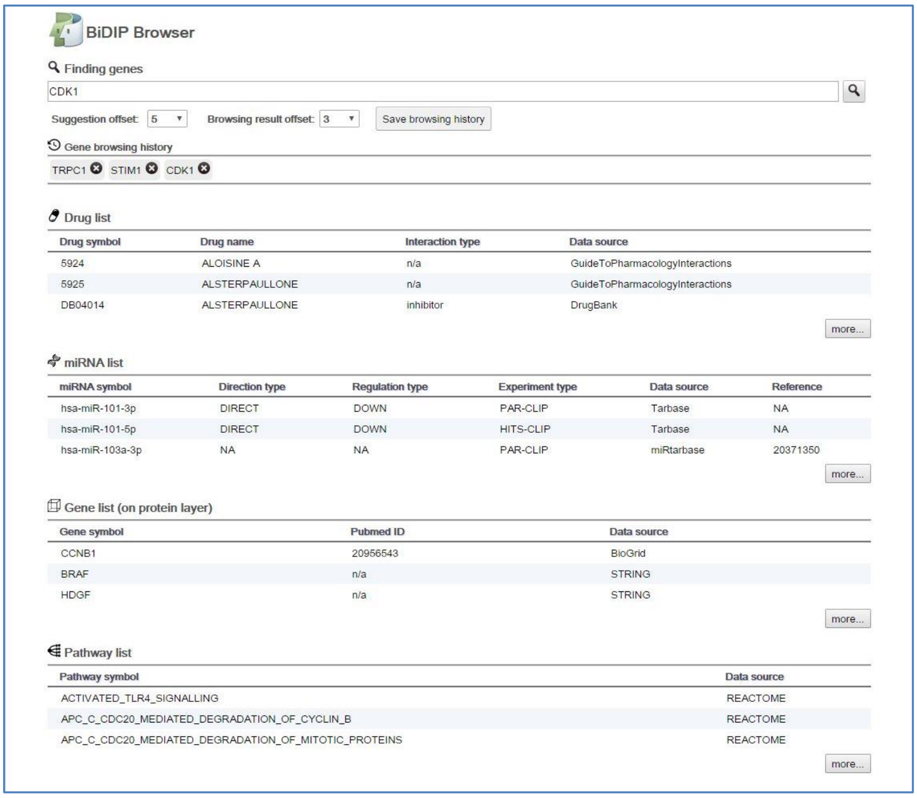
3.7 BiDIP Browser and OpenAPI

A Web-based application, BiDIP Browser, and open-access APIs were developed. Both services were designed and implemented to enable efficient retrieval of gene-centric interaction information on drugs, miRNAs, pathways and proteins based on BIM. As shown in Fig. 6, the browser has 5 panels: 1) Finding genes & Gene browsing history, 2) drug list, 3) miRNA list, 4) gene (on protein layer) list, and 5) pathway list. Given a keyword entered in the Finding Genes box, a list of candidate genes is listed whose symbols start with the keyword. Once a gene is selected from the search results, it is first recorded and listed as a previously-selected gene in the Gene browsing history list. The drug, miRNA, gene and pathway panels show elements interacting with the selected gene. Clicking the ‘Save browsing history’ button will enable the user to save his/her browsing history as an XML file. The browser is available at http://147.47.41.58:8080/BiDIP_Browser/. Most of the information the user can access from the browser can be obtained by using our RESTful web services, BiDIP OpenAPI; and third parties can use the services to develop their own systems and applications that make use of the BIM-based gene-centric interaction datasets. Table 2 gives a list of the web services with examples.

Table 2. OpenAPI Web Services

Service	Parameter syntax	Example
getGenes	keyword:KEY: from:START to:OFFSET	*baseURL*/getGenes/keyword:AA:from:5to:10
getDrugs	geneIndex:GINDEX: from:START to:OFFSET	*baseURL*/getDrugs/geneIndex:139:from:0to:5
getMiRnas	geneIndex:GINDEX: from:START to:OFFSET	*baseURL*/getMiRnas/geneIndex:139:from:0to:5
getPpis	geneIndex:GINDEX: from:START to:OFFSET	*baseURL*/getPpis/geneIndex:139:from:0to:5
getPathways	geneIndex:GINDEX: from:START to:OFFSET	*baseURL*/getPathways/geneIndex:139:from:0to:5

*** *baseURL*: http://147.47.41.58:8080/BiDIP_OpenAPI/openApi**



CDK1

Suggestion offset: 5

Browsing result offset: 3

TRPC1 STIM1 CDK1

Drug list

Drug symbol	Drug name	Interaction type	Data source
5924	ALOISINE A	n/a	GuideToPharmacologyInteractions
5925	ALSTERPAULLONE	n/a	GuideToPharmacologyInteractions
DB04014	ALSTERPAULLONE	inhibitor	DrugBank

miRNA list

miRNA symbol	Direction type	Regulation type	Experiment type	Data source	Reference
hsa-miR-101-3p	DIRECT	DOWN	PAR-CLIP	Tarbase	NA
hsa-miR-101-5p	DIRECT	DOWN	HITS-CLIP	Tarbase	NA
hsa-miR-103a-3p	NA	NA	PAR-CLIP	miRtarbase	20371350

Gene list (on protein layer)

Gene symbol	Pubmed ID	Data source
CCNB1	20956543	BioGrid
BRAF	n/a	STRING
HDGF	n/a	STRING

Pathway list

Pathway symbol	Data source
ACTIVATED_TLR4_SIGNALLING	REACTOME
APC_C_CDC20_MEDIATED_DEGRADATION_OF_CYCLIN_B	REACTOME
APC_C_CDC20_MEDIATED_DEGRADATION_OF_MITOTIC_PROTEINS	REACTOME

Fig. 6. Screenshot of BiDIP Browser

4 Conclusion and Future Work

In this paper, a biological data integration platform, BiDIP, is presented. BiDIP integrates four sets of interaction databases into a common data model. The differentiating factors of this study that distinguish it from the extant similar gene expression analysis studies are as follows: 1) development of comprehensive data model, BIM, for interaction databases; 2) creation of consolidated databases for four types of heterogeneous interaction databases; 3) two points of access, BiDIP Browser and open APIs, to the database. BiDIP enables researchers to gain access to integrative biological information about genes of interest.

The focus of this study has been to develop a solid basis for interaction-focused databases. Hence, there are certain details left out for now such as epigenetic information, DNA variant, and etc. These data will be progressively added to BIM.

Finally, BiDIP is part of a larger project called OnTAP. We are currently in the process of developing data input modules, data pre-processing modules, and data analysis modules for full deployment of OnTAP for general use.

Acknowledgements.
This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and future Planning (NRF-2014R1A2A1A11049728)

References

1. Hokamp, K., Roche, F., Acab, M., Rousseau, M., Kuo, B., Goode, D., Aeschliman, D., Bryan, J., Babiuk, L., Hancock, R., Brinkman, F.S.: ArrayPipe: a flexible processing pipeline for microarray data. Nucleic Acids Res. 32 (Web Server), W457 (2004)
2. Kapushesky, M., Kemmeren, P., Culhane, A., Durinck, S., Ihmels, J., Korner, C., Kull, M., Torrente, A., Sarkans, U., Vilo, J., Brazma, A.: Expression Profiler: next generation-an online platform for analysis of microarray data. Nucleic Acids Res. 32 (Web Server), W465 (2004)
3. Newman, J.C., Weiner, A.M.: L2L: a simple tool for discovering the hidden significance in microarray expression data. Genome Biol. 6(9), R81 (2005)
4. Psarros, M., Heber, S., Sick, M., Thoppae, G., Harshman, K., Sick, B.: RACE: remote analysis computation for gene expression data. Nucleic Acids Res. 33 (Web Server), W638 (2005)
5. Xia, X., McClelland, M., Wang, Y.: WebArray: an online platform for microarray data analysis. BMC Bioinformatics. 6, 306 (2005)
6. Romualdi, C., Vitulo, N., Favero, M., Lanfranchi, G.: MIDAW: a web tool for statistical analysis of microarray data. Nucleic Acids Res. 33 (Web Server), W644 (2005)
7. Rainer, J., Sanchez-Cabo, F., Stocker, G., Sturn, A., Trajanoski, Z.: CARMAweb: comprehensive R-and bioconductor-based web service for microarray data analysis. Nucleic Acids Res. 34 (Web Server), W498 (2006)

8. Diaz-Uriarte, R., Alibes, A., Morrissey, E., Canada, A., Rueda, O.M., Neves, M.L.: Asterias: integrated analysis of expression and aCGH data using an open-source, web-based, parallelized software suite. Nucleic Acids Res. 35 (Web Server), W75 (2007)
9. Rehrauer, H., Zoller, S., Schlapbach, R.: MAGMA: analysis of two-channel microarrays made easy. Nucleic Acids Research. 35 (Web Server), W86 (2007)
10. Tárraga, J., Medina, I., Carbonell, J., Huerta-Cepas, J., Minguez, P., Alloza, E., Al-Shahrour, F., Vegas-Azcárate, S., Goetz, S., Escobar, P., Garcia-Garcia, F., Conesa, A., Montaner, D., Dopazo, J.: GEPAS, a web-based tool for microarray data analysis and interpretation. Nucleic Acids Res. 31(13), 3461-3467 (2008)
11. Zhu, Y., Zhu, Y., Xu, W.: EzArray: A web-based highly automated Affymetrix expression array data management and analysis system. BMC Bioinformatics. 9(46) (2008)
12. Glaab, E., Garibaldi, J.M., Krasnogor, N.: ArrayMining: a modular web-application for microarray analysis combining ensemble and consensus methods with cross-study normalization. BMC Bioinformatics. 10(358) (2009)
13. Linderman, G.C., Chance, M.R., Bebek, G.: MAGNET: Microarray gene expression and network evaluation toolkit. Nucleic Acids Res. 40 (Web server), W152-156 (2012)
14. Laenen, G., Ardeshirdavani, A., Moreau, Y., Thorrez, L.: Galahad: a web server for drug effect analysis from gene expression. Nucleic Acids Res. 43 (Web server), W208-212 (2015)
15. KEGG: Kyoto Encyclopedia of Genes and Genomes, http://www.genome.jp/kegg
16. Reactome, http://www.reactome.org
17. DGIdb: The Drug Gene Interaction Database, http://dgidb.genome.wustl.edu
18. PharmGKB: The Pharmacogenomic Knowledgebase, https://www.pharmgkb.org
19. STITCH: Chemical-Protein Interactions, http://stitch.embl.de
20. BINDTranslation, http://baderlab.org/BINDTranslation
21. BioGRID: The Biological General Repository for Interaction Datasets, http://thebiogrid.org
22. DIP: Database of Interacting Proteins, http://dip.doe-mbi.ucla.edu/dip/Main.cgi
23. HPRD: Human Protein Reference Databse, http://www.hprd.org
24. IntAct, http://www.ebi.ac.uk/intact
25. MINT: The Molecular INTeraction database, http://mint.bio.uniroma2.it
26. STIRNG: functional protein association networks, http://www.string-db.org
27. miRTarBase: the experimentally validated microRNA-target interactions database, http://mirtarbase.mbc.nctu.edu.tw
28. TarBase, http://diana.imis.athena-innovation.gr/DianaTools/index.php?r=tarbase/index