Journal of Biomedical Informatics 54 (2015) 121–131

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

HPIminer: A text mining system for building and visualizing human protein interaction networks and pathways ☆

Suresh Subramani [a], Raja Kalpana [a], Pankaj Moses Monickaraj [b], Jeyakumar Natarajan [a,*]

[a] Data Mining and Text Mining Laboratory, Department of Bioinformatics, School of Life Sciences, Bharathiar University, Tamil Nadu, India
[b] Department of Computer Science, Bharathiar University, Coimbatore, Tamil Nadu, India

ARTICLE INFO

Article history:
Received 30 December 2013
Accepted 15 January 2015
Available online 4 February 2015

Keywords:
Protein–protein interactions
Network visualization
Pathway visualization
Information extraction
Text mining
Knowledge discovery
Biomedical informatics

ABSTRACT

The knowledge on protein–protein interactions (PPI) and their related pathways are equally important to understand the biological functions of the living cell. Such information on human proteins is highly desirable to understand the mechanism of several diseases such as cancer, diabetes, and Alzheimer's disease. Because much of that information is buried in biomedical literature, an automated text mining system for visualizing human PPI and pathways is highly desirable. In this paper, we present HPIminer, a text mining system for visualizing human protein interactions and pathways from biomedical literature. HPIminer extracts human PPI information and PPI pairs from biomedical literature, and visualize their associated interactions, networks and pathways using two curated databases HPRD and KEGG. To our knowledge, HPIminer is the first system to build interaction networks from literature as well as curated databases. Further, the new interactions mined only from literature and not reported earlier in databases are highlighted as new. A comparative study with other similar tools shows that the resultant network is more informative and provides additional information on interacting proteins and their associated networks.

1. Introduction

Protein–protein interactions (PPI) and their pathways are essential to understand almost all cellular functions and activities [1,2]. The biological functions of the living cell are results of many interacting molecules [3]. The knowledge about PPIs is of central importance to identify their related pathways in the biological systems. The amount of PPIs being published in the biomedical literature is huge with the results from high throughput experimental technologies [4,5]. Recently, significant amounts of work are carried out in building databases that store manually curated information on PPIs from the literature [6]. Some examples of these resources include HPRD [7], MINT [8], BioGRID [9], MIPS [10], PDZBase [11], IntAct [12], STITCH [13], and others. However, literature mining of PPIs is a time consuming task and is almost impractical for PPI extraction when compared to the rapid growth of biological publications. As a result, many PPI data are still available only in the literature [14]. The knowledge about metabolic pathways is equally important in organizing knowledge in system biology and often represented through collective interpretations of facts scattered throughout literature [15–18]. Pathways are very integrative in nature and require substantial human effort to construct. A huge number of PPI information still remains as hidden information in the biomedical literature and biologists have to read a large number of published papers to interpret and construct a pathway [19]. Further, the curation of a constructed pathway also requires monitoring of recent publications to extract relevant PPI information [20,21]. KEGG [3], Reactome [22] and BioCyc [23] are few best known primary pathway databases that are developed and maintained by a few dedicated research groups. According to Pathguide, there are 59 pathway related resources and 151,166 pathway entries within the list [24].

Among the PPI databases available, HPRD is specific for human with annotations pertaining to human PPIs and interactions of proteins with nucleic acids and other small molecules based on experimental evidence from the literature. Furthermore, the database includes information about posttranslational modifications, sub-cellular localization, protein domain architecture, tissue expression and association with human diseases [25]. Likewise, KEGG is a well-known and widely used pathway database among the scientific community [2]. The database links the genomic

☆ Availability and implementation: HPIminer is freely available on the web at http://www.biominingbu.org/HPIminer.

* Corresponding author at: Data Mining and Text Mining Laboratory, Department of Bioinformatics, School of Life Sciences, Bharathiar University, Coimbatore 641 046, Tamil Nadu, India. Fax: +91 422 2422387.

E-mail addresses: sureshsubramani@hotmail.com (S. Subramani), kalpana.rajaa@gmail.com (R. Kalpana), pankajmoses@hotmail.com (P.M. Monickaraj), n.jeyakumar@yahoo.co.in (J. Natarajan).

http://dx.doi.org/10.1016/j.jbi.2015.01.006
1532-0464/

information with higher order functional information by computerizing current knowledge on cellular processes and by standardizing gene annotations. KEGG incorporates a functional assignment process to link a set of genes in the genome with a network of interacting molecules in the cell, such as pathway to represent a higher order biological function [26].

In this paper, we introduce a new interaction and pathway text mining system called HPIminer with effective visualization support on retrieved information. HPIminer incorporates an interaction database curated from HPRD [7] and a pathway database collected from KEGG [26] for retrieving interaction and pathway information on a pair of interacting proteins that was previously extracted from the literature. The extraction of PPI pairs from the literature is carried out by incorporating our own tools, namely NAGGNER [27] for protein/gene name recognition, ProNormz [28] for normalization and PPInterFinder [29] for PPI extraction. All the three tools are highly specific to biomedical literature related to human and provide higher accuracy in the extraction of PPIs. HPIminer retrieves interaction and pathway information for each protein in the extracted PPI from the interaction and pathway databases respectively. Additionally, the tool builds an interaction network to visualize both protein–protein and protein–nonprotein interactions. To our knowledge HPIminer is the unique system that combines text mining PPI information with known interactions and associated pathways from curated databases related to human literature with a special emphasis on building protein interaction network with multiple visualization options.

2. Materials and methods

HPIminer consists of three separate parts. (i) Curated databases of protein interactions, pathways for retrieving, visualizing the protein interactions and pathways of a given query protein. (ii) A text mining engine for automatically retrieving protein interactions from the given sentence and then visualizing other protein interactions and pathways for the same protein from the curated database. (iii) A user friendly web interface to upload biomedical literature or query protein/gene name list directly for PPI extraction. The curated databases were imported and organized as relational database using MySQL. The text mining engine for text processing, and interaction extraction was implemented in Perl and Java. Cytoscape Web was used for PPI network visualization. The user friendly web interface was developed using Perl/CGI scripts. The overview of the system is illustrated in Fig. 1.

2.1. Database construction

The main idea behind HPIminer is to visualize all known protein interactions and pathways of a given protein or protein interactions directly retrieved from the literature. For this task, the system incorporates two specialized databases one for human protein interactions and other for pathways. Fig. 2 summarizes the major steps in the construction of both databases.

We constructed two separate databases for interactions and pathways. Interaction database is constructed from the HPRD database [7] containing 39,376 entries on human protein–protein interactions and 480 entries on human protein–nonprotein interactions. HPRD contains annotations pertaining to human proteins based on experimental evidence from the literature. Three datasets from HPRD database namely (i) human protein–protein interaction database, (ii) human protein–nonprotein interaction database and (iii) human protein pathway database are curated to form the interaction database. Likewise, entries related to human genes and pathways are collected from KEGG database [26] to form protein pathway database.

2.1.1. Human protein interaction database

We have chosen the Human Protein Reference Database (HPRD) for the construction of human protein interactions database as it contains a comprehensive collection of human PPIs with experimental evidence from the literature [7]. The database includes interactions of proteins with other proteins, nonproteins, nucleic acids and small molecules. HPRD is freely available in several different formats (http://www.hprd.org/). We downloaded binary protein–protein interactions, binary protein–nonprotein interactions and HPRD ID mappings files for constructing the human PPI database and protein–nonprotein interaction database. Table 1 lists the entries present in each file.

To map the HPRD PPI data with our PPI data, we used the Entrez Gene ID as the common identifier. However, HPRD database has its own HPRD ID for PPI mapping and does not have the Entrez Gene ID. So, we used our earlier protein normalization tool ProNormz and its specialized synonym dictionary [28] to find the official symbol, Entrez Gene ID, and known synonyms associated with each HPRD entries. Fig. 3 shows the mapping methodology of the HPRD PPI data with ProNormz's synonym dictionary. The final HPRD PPI database entries have the official gene names, Entrez Gene IDs, ref-seq ids, experimental id and reference id of both proteins.

In the mapping process, we removed a few PPIs which have no Entrez Gene ID and also filtered out the duplicate entries such as '*xy*' duplicate and '*yx*' duplicates using the symmetry property (i.e.) proteins in the intersection of the interactors of protein A and of protein B. The final curated human PPI database contains 39,376 entries on protein–protein interactions and 479 entries on protein–nonprotein interactions.

2.1.2. Human protein pathway database

We utilized the pathway data available at KEGG (Kyoto Encyclopedia of Genes and Genomes), which is one of the widely used resources for pathway information (http://www.kegg.jp). We have downloaded all KEGG pathways linked to each of the human genes and list of human pathways. Later, using a Perl script we constructed the human pathway database by mapping human synonyms dictionary. Human genes with their synonyms and associated pathways of the curated data are stored as human protein pathway database. Fig. 4 shows the mapping methodology for constructing our pathway database.

2.2. Text mining engine

The input text may be a PubMed abstract in Plain text/MEDLINE/XML format. We implemented the text mining engine by combining three of our earlier text mining systems viz NAGGNER [27] for named entity recognition of protein/gene names, ProNormz [28] for protein normalization and PPInterFinder [29] for extracting protein–protein interactions with two newly developed components Interaction Miner (Iminer) for mining interactions and Pathway Miner (Pminer) for mining pathways from interaction and pathway databases. The protein name recognition, normalization and PPI extraction tools are highly specific to human proteins (Fig. 5).

2.2.1. Gene/protein name recognition with NAGGNER

The gene/protein name recognition is achieved by our earlier developed tagger NAGGNER [27]. NAGGNER is a hybrid tagger that utilizes CRF tagging with additional rules and abbreviation identification modules specific to human literature for tagging of human proteins and genes. The system achieves a precision of 80.47%, recall of 71.60% and an overall *F*-score of 75.77% in tagging human proteins/genes on JNLPBA2004 corpus, which is comparable to the

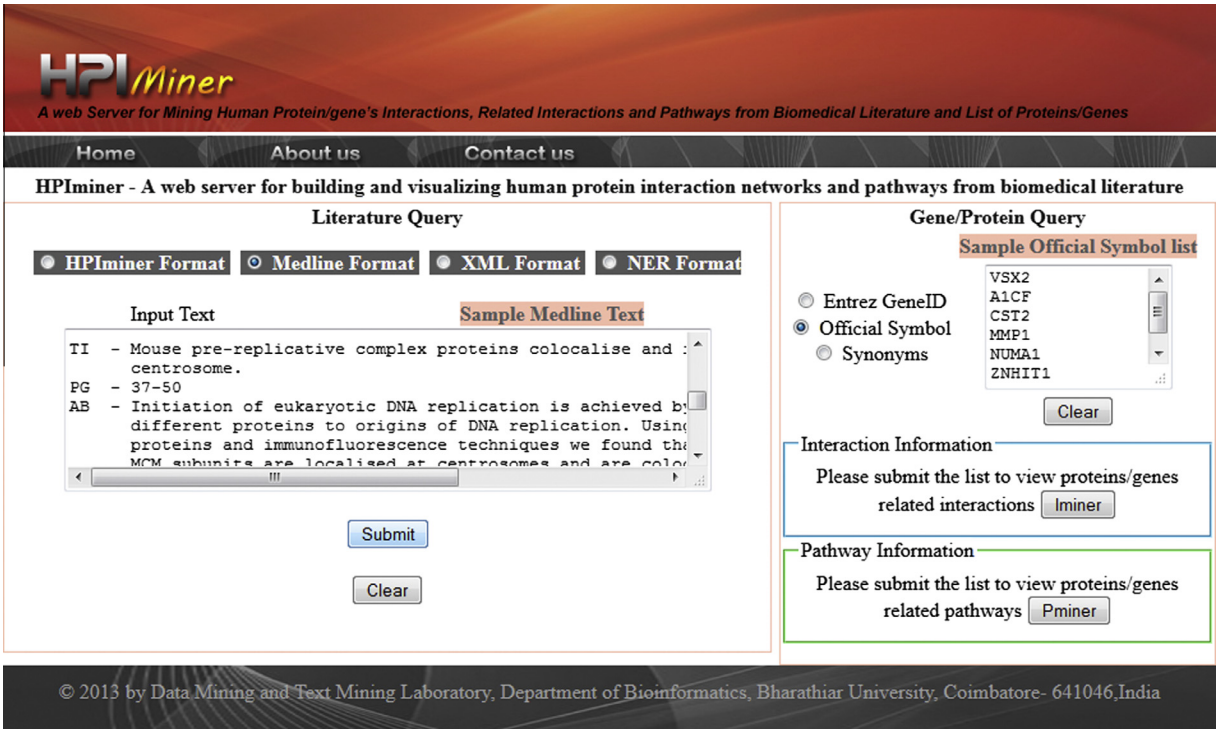

Fig. 1. Screenshot of HPIminer tool.

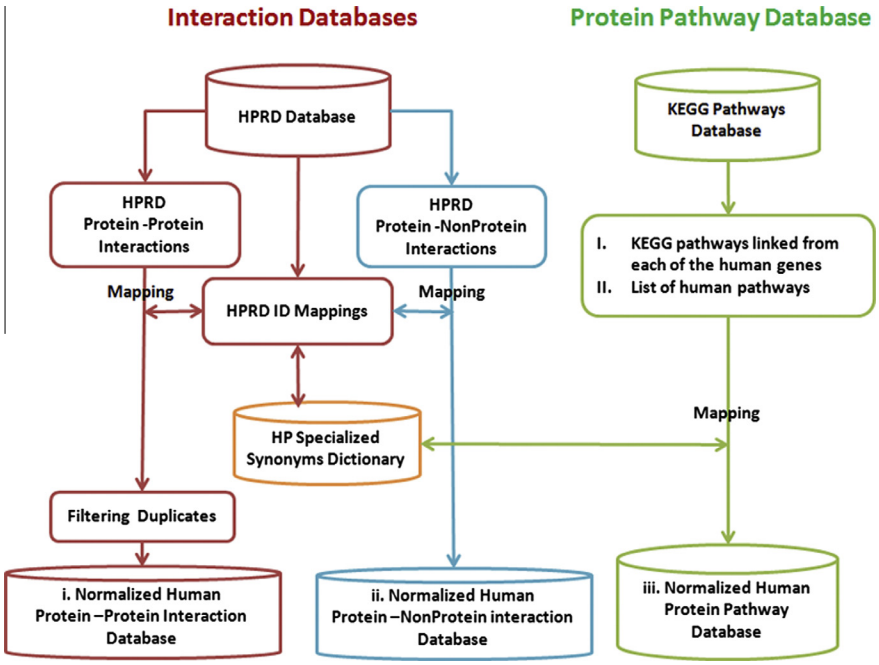

Fig. 2. Construction of HPIminer's curated databases.

other state-of-the-art systems. NAGGNER is freely available at http://biominingbu.org:8080/NAGGNER.

2.2.2. Gene/protein normalization with ProNormz

The tagged protein and gene names were normalized into Entrez Gene ID as the unique identifier using our earlier developed system called ProNormz [28]. ProNormz incorporates a specialized synonyms dictionary for human proteins with a set of 15 string matching rules classified into two main categories as dictionary rules and entity rules, and a disambiguation module to achieve the normalization. The system achieves a precision of 86.66%, recall of 80.25% and an overall *F*-score of 83.33% on BioCreative II test dataset available for normalization task. ProNormz is freely available at http://www.biominingbu.org/pronormz.

2.2.3. Protein–protein interaction extraction with PPInterFinder

Extraction of protein–protein interactions is achieved by our earlier developed web based tool called PPInterFinder [29].

Table 1
Benchmark results of the cascade oscillator's model.

Binary protein protein interactions	Binary protein nonprotein interactions	HPRD – ID mappings
interactor_1: geneSymbol hprd_id refseq_id	interactor: geneSymbol hprd_id refseq_id	hprd_id geneSymbol Accession: nucleotide protein
interactor_2: geneSymbol hprd_id refseq_id	non_protein_interactorname	entrezgene_id omim_id swissprot_id main_name
experiment_type reference_id	experiment_type reference_id	

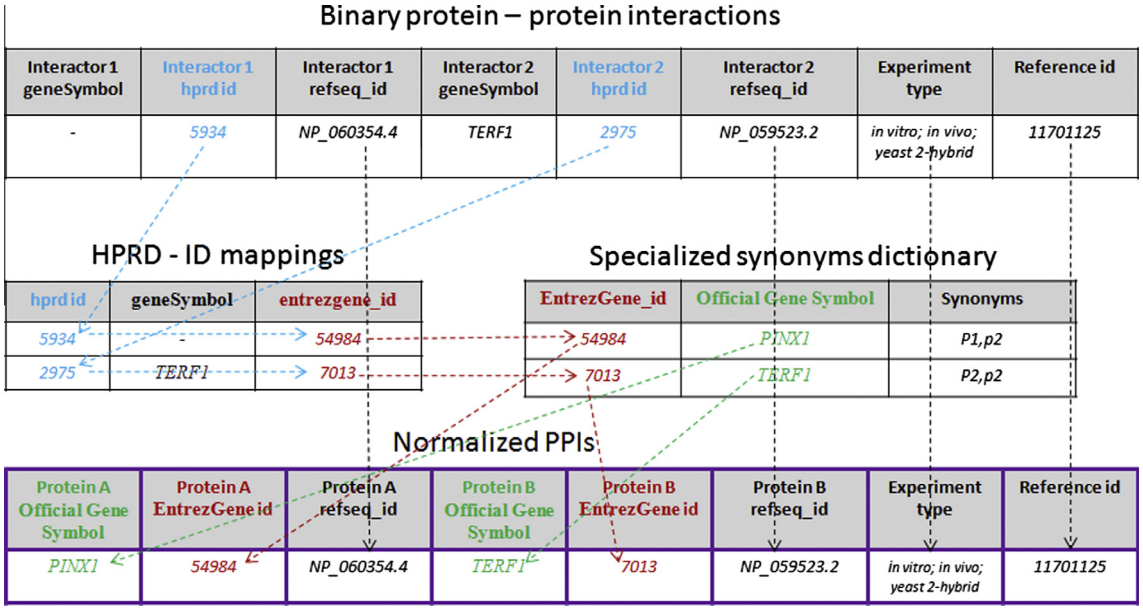

Fig. 3. Connectivity diagram for normalization of HPRD PPI database.

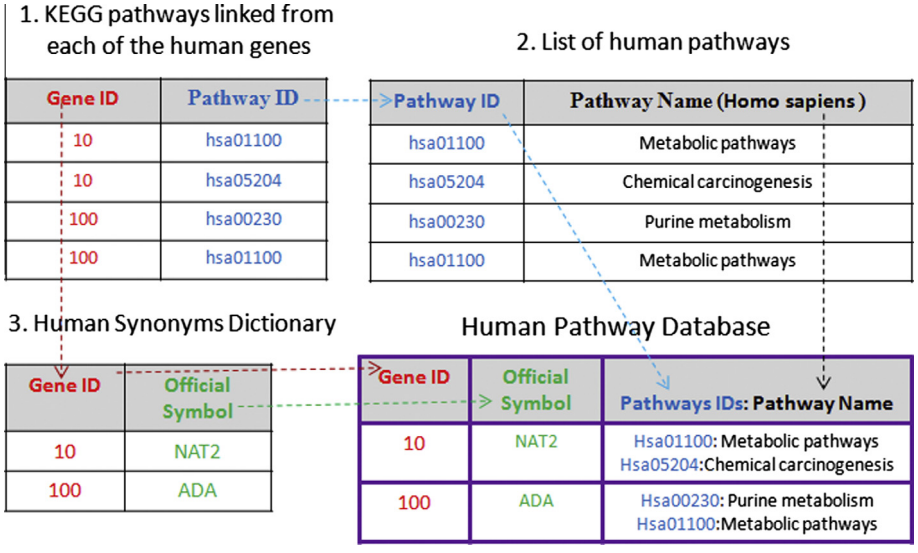

Fig. 4. Mapping methodology for pathway database construction.

PPInterFinder is an NLP tool applying a set of rules on grammatically parsed sentence to identify the candidate PPI pairs and matching the syntactic structure of the sentence with a dictionary of patterns. The reported accuracy of extraction of human PPIs by PPInterFinder was 66.05% for PPI detection alone and 57.15 for entity identification with PPI detection on AIMED corpus specific

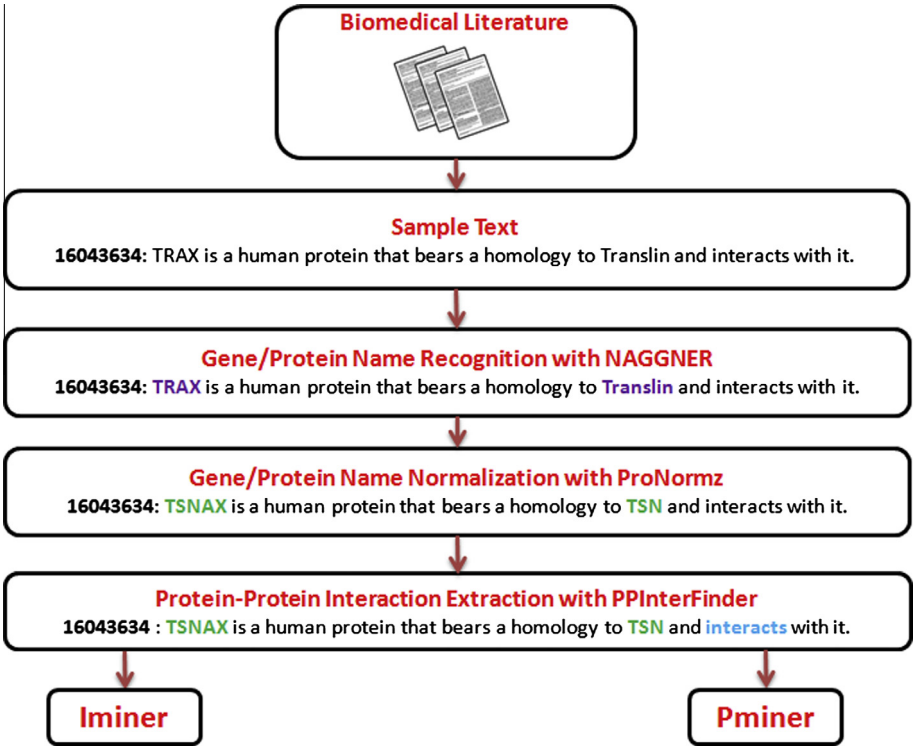

Fig. 5. Text preprocessing flowchart.

to humans. PPInterFinder is freely available at http://www.biominingbu.org/ppinterfinder.

2.2.4. Interaction Miner (Iminer)

The two proteins (protein A and protein B) from the extracted PPI pairs from PPInterFinder were matched with curated human protein interaction database to retrieve interactions of protein A and protein B with other proteins. We used a pair matching algorithm for retrieving the pairing proteins for the two proteins (protein A and protein B) of the PPI pair. The algorithm is implemented for retrieving both protein–protein interaction information as well as protein–nonprotein interaction information. Fig. 6 shows the algorithm for retrieving the interacted pair for each protein(x) in the database at location row(R), column(C) or row(R), column(C) + 1. Each of the obtained pairs from the processed text (protein A and protein B) are processed and checked individually in PPI database. The associated interacting protein is picked from the row when protein(x) matches with any database entry.

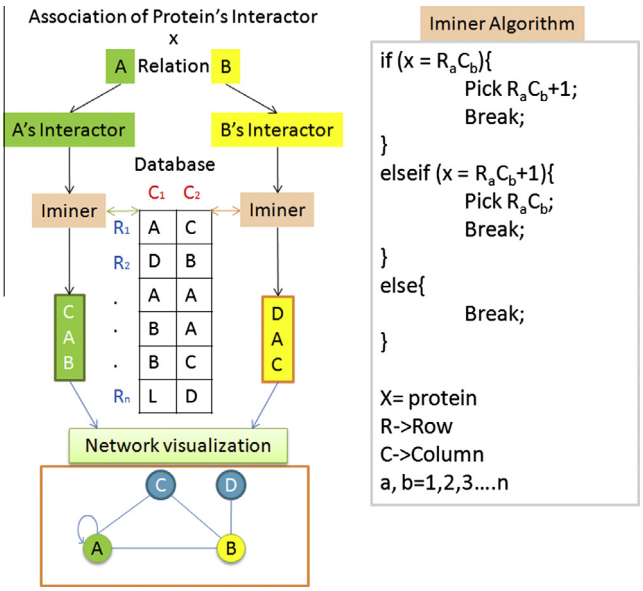

Fig. 6. Iminer algorithm and visualization for association of protein's interactor.

Next, we used Cytoscape Web [30] for constructing and visualizing interaction networks. For the given PPI pair, our network building process consists of three steps.

(i) Display the interactions of protein A,
(ii) Display the interactions of protein B,
(iii) Search for and display the possible connections between the two networks via the common interacting protein.

In addition to the individual PPI pair network visualization, we also provide few additional proteins options for network visualization. This includes

(iv) '*Combined network visualization option*' for visualizing all the PPI pairs mined from the literature with common interacting proteins between the networks.
(v) Visualize whether the extracted PPI pair is already known and exist in HPMiner's human protein interaction database or new and first mined from literature.
(vi) Input window to the user to input their known protein/gene list of interest directly and get their PPI interactions.

2.2.5. Pathway Miner (Pminer)

For the given two proteins (protein A and protein B) of the PPI pair, Pminer extracts and displays all the known pathways of protein A, protein B and their common pathways based on the intersection property.

Further, similar to Iminer, Pminer also displays known pathways and pathway map list of known proteins and genes directly mined from the pathway database.

3. Result and discussion

3.1. Human protein interaction database

We downloaded interactions from HPRD (http://www.hprd.org) and pathways from KEGG (http://www.kegg.jp). The downloaded interaction dataset from HPRD includes 39,376 binary protein protein interactions, 480 binary protein–nonprotein interactions and 19,701 HPRD ID mappings files. Similarly, the retrieved pathway dataset from KEGG includes 20,206 linking human genes to the related pathways and 257 pathways specifically belonging to human. Additionally, we utilized the curated human gene/protein names and synonyms dictionary constructed for normalization task which contains 33,580 human genes/proteins with their known synonyms [28]. All the three datasets were used for the construction of interaction database, and pathway database (Table 2).

The PPI database was generated by mapping binary PPIs, HPRD ID with the specialized gene/protein synonyms dictionary. We obtained 39,186 true interactions out of 39,376 interactions from HPRD after filtering out 190 interactions that are either incomplete or having protein/gene names not approved by HGNC (e.g. HPRD ID: 13664, 11791) or not having Entrez Gene ID (e.g. HPRD ID: 11786, 13632). Further, the normalized PPI database also had 36 *XY* duplicates (e.g. HPRD ID: 04992-15987, 01769-01080) and 9 *YX* duplicates (e.g. HPRD ID: 04274-07210, 01578-07211) that were removed. The final filtered and normalized human PPI database contains 9638 human proteins and 39,141 unique human PPI interactions (Supplementary file 1).

In addition, the HPRD database also contains 511 protein and nonprotein (e.g. small molecule) interactions. In our PPI database, we also included the protein and nonprotein interactions after curation. There are about 32 entries with missing gene id (e.g. HPRD ID: 14380, 03973) and one entry which is not approved by HGNC (HPRD ID: 19461). We removed 32 protein–nonprotein interactions and our final dataset contains 405 proteins and 151 nonproteins contributing for 479 protein–nonprotein interactions (Supplementary file 1).

3.2. Human protein pathway database

In a similar way, the retrieved 20,206 human genes were linked to the related pathways from KEGG database and the entries were mapped to gene/protein synonym dictionary. The final normalized dataset contains 6317 proteins and 257 associated pathways contributing to 20,206 entries on gene–pathway relationships.

3.3. Iminer/Pminer and retrieval of protein interactions, networks and pathways

Both Iminer and Pminer have several options to users for visualizing the protein interactions, networks and pathways. Fig. 7 displays the main output Iminer, which displays PPI pairs mined from literature with their associated sentences (Fig. 7A), individual interactions of each protein in the PPI pair (Fig. 7B), network view individual PPI pair (proteins A and protein B) (Fig. 7C), combined network view of all PPI pairs (Fig. 7D), Further, each of the PPI pairs mined from the literature were checked for their existence in the curated PPI database. If the PPI pair already exists in the PPI database, they were tagged as '*known*'. Otherwise if it was a new PPI pair not already present in PPI database they were tagged as '*New*' (Fig. 7E). In the above network, the central node of each network represents the protein in the each PPI pair and the surrounding nodes represent all interacting proteins with any possible common interactions among the networks.

Similarly, Fig. 8 displays the main output of Pminer, which displays all the known pathways of protein A, protein B and also any common pathways of both proteins (Fig. 8A) and pathway map view of each pathway with the positions of protein A and B highlighted (Fig. 8B).

Additionally, we also provide another input window to the user to input their known protein/gene list of interest. The user can input their query protein/gene names as Entrez Gene ID, or official symbol or known synonym name (Fig. 1). For the input protein/gene list, HPIminer directly extracts their known interactions and pathways from the database and visualize their associated networks and pathways (Fig. 9A–C).

3.4. Comparisons

HPIminer is an integrated text mining system with PPI identification combined with network and pathway visualization. Due to

Table 2
Source and derived databases.

Source databases		
HPRD	PPI entries: 39,376 interactions 19,701 HPRD – ID mappings Protein–nonprotein entries: 480 interactions 19,701 HPRD – ID mappings	
KEGG	20,206 entries on KEGG pathways linked from each of the human genes 257 entries on list of human pathways	
Derived databases		
	Filtered data	Final data
PPI database	190 incomplete entries 36 duplicates	39,141 interactions
Protein–nonprotein database	1 incomplete entries	479 interactions
Pathway database		6317 proteins 257 pathways

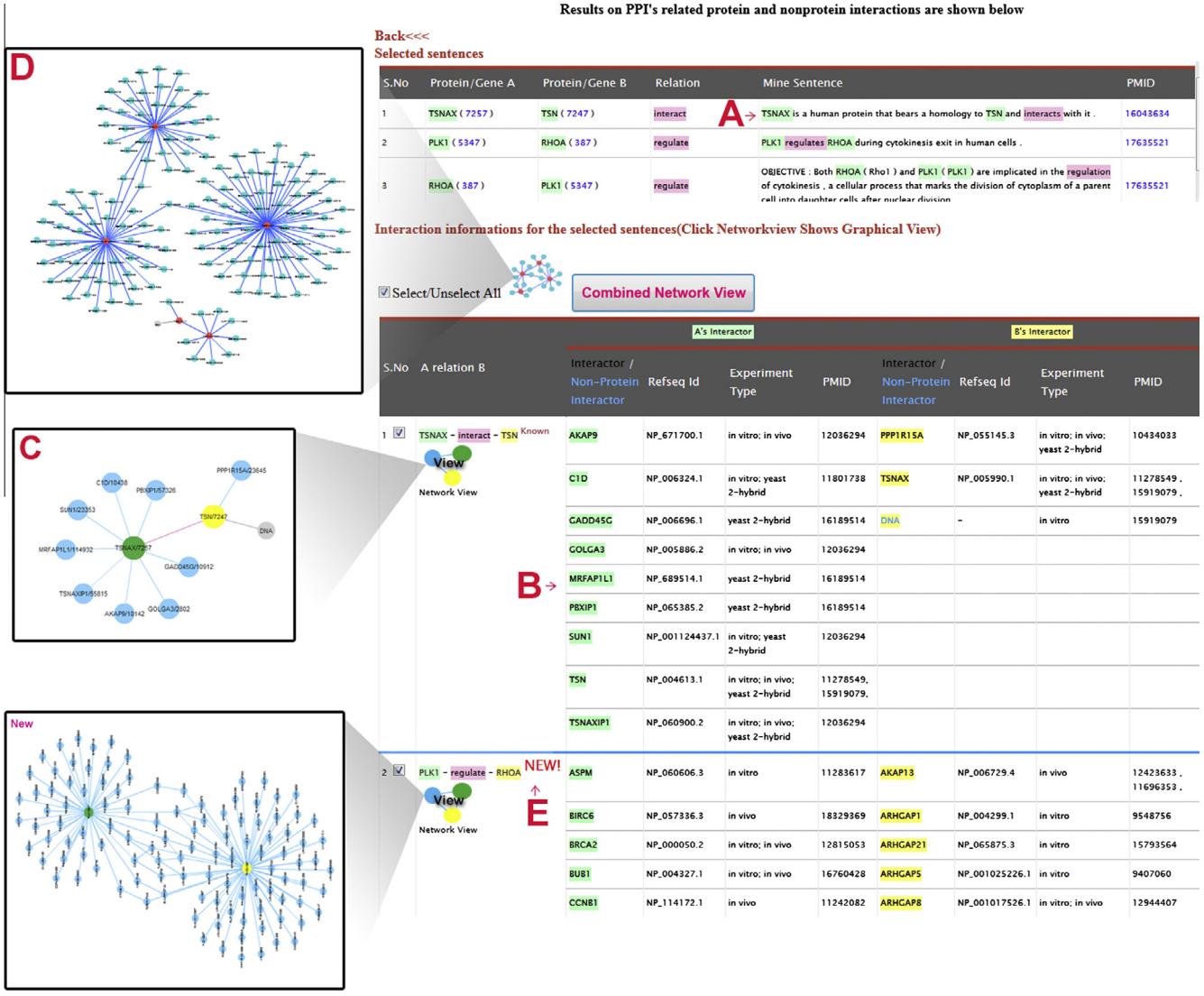

Fig. 7. Screenshot of HPIminer output on interaction information. (A) PPI pairs from literature, (B) interactor list of each protein in PPI pair, (C) individual PPI pair network, (D) combined network of all PPI pairs, and (E) new PPI pair with related interactions.

the lack of such an integrated system, direct comparison of HPIminer with other such systems is not possible. Further, three existing modules of the HPIminer namely, NAGGNER-named entity tagger module [27], ProNormz-protein normalization module [28] and PPInterFinder-PPI information extraction module [29] were already evaluated on standard corpora and published. For example, the PPInterFinder module for PPI information extraction from literature has been evaluated in the BioCreative Workshop 2012 Track III [31] using AIMED [32], HPRD50 [33], IntAct [34] Corpora [29]. The reported accuracy of PPInterFinder was 66.05% on AIMED, 68.24% on HPRD50 and 81.37% on IntAct corpora which is comparatively higher when compared to other best systems using these corpora [35].

Hence, to demonstrate and evaluate the performance of HPIminer's fourth text based novel PPI extraction and visualization module, we performed the following two case studies with a test dataset containing 32 genes related to Alzheimer's disease [36]. The first case study demonstrates how the TM module will help to extract novel PPI information from literature which was not included in the earlier PPI databases. The second study demonstrates the superiority of HPIminer over other network visualization system due to the inclusion of TM module.

3.5. Case study 1: text based novel PPI extraction

As reported earlier, we analyzed a dataset of 32 genes which were known to be involved in Alzheimer's disease [36]. First, we performed a literature search of all the 32 genes and retrieved 573 abstracts from PubMed database. The entity tagging and normalization modules of HPIminer determined 704 PPI sentences and filtered out 89 sentences for not having any Alzheimer's disease candidate genes. The PPI extraction module of HPIminer, resulted 89 candidate PPI pairs out of which 57 pairs were novel and retrieved only from literature and remaining 32 were already present in the HPRD database. Further, the manual curation of the above 57 novel text based PPI pairs results a total of 39 candidate PPI pairs as true PPI pairs (Supplementary file 2). The occurrence of false positives PPIs was due to the fact that certain proteins might be missed out during entity recognition and certain mismatches in pattern recognition of PPIs. The above 39 true and novel PPI pairs and not pre-reported in HPRD database.

In addition, we further validated these 39 true PPI pair by network analysis through other neighboring proteins in the network [37]. For this, network construction was performed for these 39 novel PPI pairs. It was already reported that in a protein network,

Selected sentences

S.No	Protein/Gene A	Protein/Gene B	Relation	Mine Sentence	PMID
1	PLK1 (5347)	ORC2 (4999)	phosphorylate	Mechanistically , we find that PLK1 phosphorylation of ORC2 maintains DNA replication on gemcitabine treatment .	23188630
2	PLK1 (5347)	MCM7 (4176)	interact	interaction of chromatin-associated PLK1 and MCM7 .	15654075

Pathway informations for the selected sentences (click pathway to view protein in pathway map)

S.No	A Relation B	Protein/Gene A's Pathway	Protein/Gene B's Pathway	Protein/Gene A & B 's Pathway
1	PLK1 - phosphorylate - ORC2	PLK1 1. Cell cycle 2. Oocyte meiosis 3. Progesterone-mediated oocyte maturation	ORC2 1. Cell cycle	PLK1 & ORC2 1. Cell cycle **A**
2	PLK1 - interact - MCM7	PLK1 1. Cell cycle 2. Oocyte meiosis 3. Progesterone-mediated oocyte maturation	MCM7 1. DNA replication 2. Cell cycle	PLK1 & MCM7 1. Cell cycle

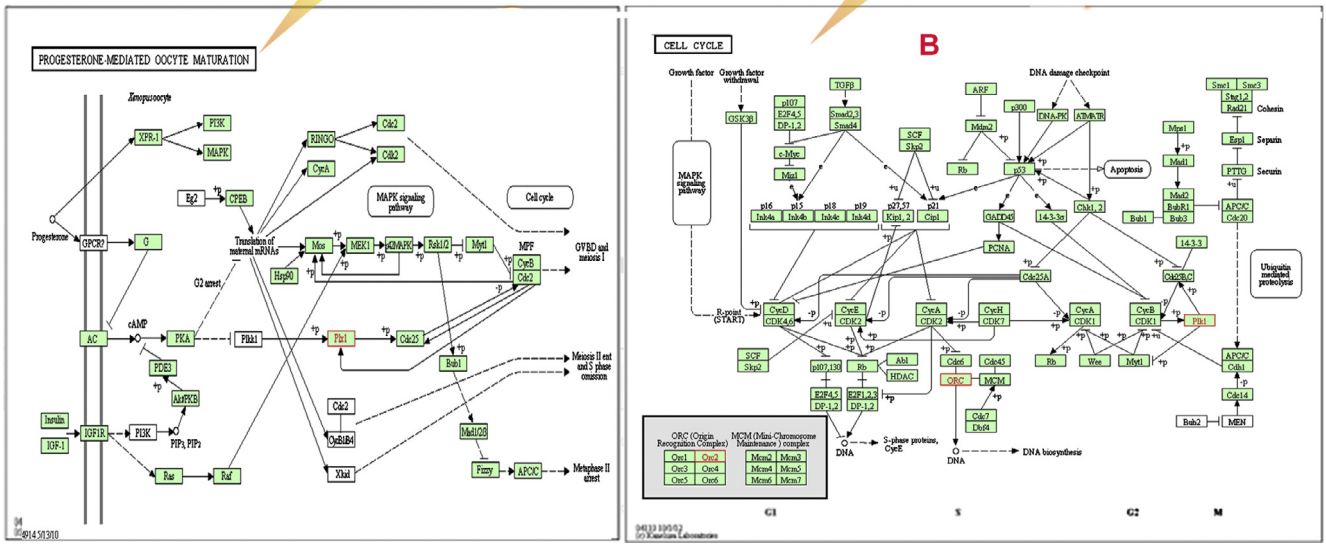

Fig. 8. Screenshot of HPIminer output on pathway information. (A) Common pathway of both proteins in PPI pair and (B) pathway with highlighted protein names.

first level neighbors were known to share similar functions as that of the target protein and hence functional similarity between the proteins are stronger in the network [38]. It is plausible that two proteins that interact with a common set of proteins have a good likelihood of sharing similar physical or biochemical characteristics, and thus exhibit a common function. Our network analysis result reveals that out of 39 novel PPI pairs, 25 PPI pairs had a relation through neighboring proteins (Supplementary file 3).

The above results demonstrate that HPIminer's network module not only identify novel PPIs based on literature mining but also validate them using other neighboring proteins in the network.

3.6. Case study 2: comparison with other network visualization systems

In case study 1, we successfully demonstrated how HPIminer can effectively extract novel PPI from literature which was not reported in the HPRD database, and its validation. In this case study, we further compare the performance of HPIminer with other similar network and pathway visualization systems. To our knowledge other three similar systems available for network construction and visualization were (i) Gene Interaction Miner (GIM) [39], (ii) STRING [40] and (iii) Pathway Studio [41]. GIM uses contextual information provided by iHOP (Information Hyperlinked over Proteins – http://www.ihop-net.org/) for the PPI information and pairs, whereas Pathway Studio and STRING uses their own curated databases. A comparative study of GIM, STRING and Pathway Studio was already reported and GIM outperformed the other two [39]. So, we directly compared HPIminer with GIM.

We use the same dataset of 32 genes which are known to be involved in Alzheimer's disease to compare the networks generated by both HPIminer and GIM. The resultant network is shown in Fig. 10. For the 32 genes related to Alzheimer's disease, HPIminer shows 17 connected edges whereas GIM's network has only 9 edges. Similarly, the number of disconnected genes/proteins in HPIminer network was only 15 proteins compared to GIM's 23 proteins. Further, the disconnected individual PPI pair network of HPIminer shows neighboring proteins for each protein in the PPI pair, however such information is missing in GIM as HPIminer includes novel text based PPI pairs in network construction. Some of the additional findings of HPIminer with its literature validation are discussed below.

For example, HPIminer results show that, the protein TAF9/6880 and YWHAE/7531 were common interacting partners to DRAP1/10589 – TAF7/6879 PPI pair and TAF7/6879 – YWHAH/7533 PPI pair. Such information provides new knowledge that

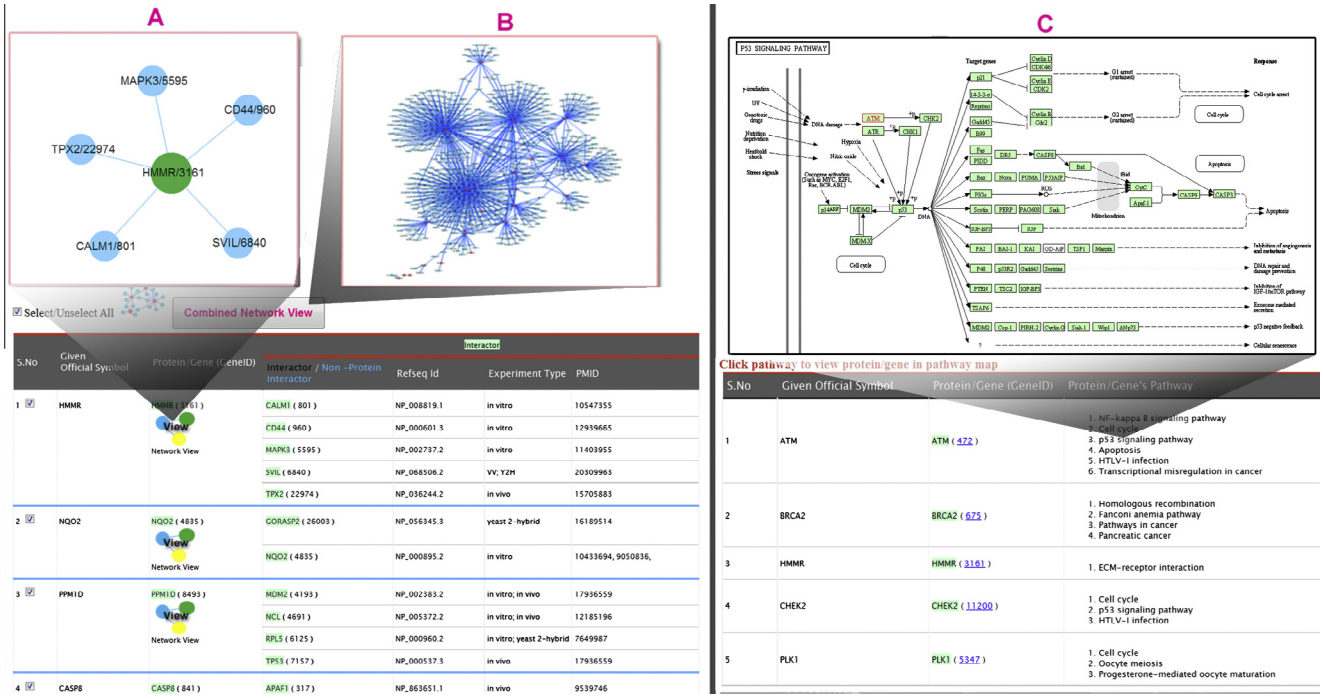

Fig. 9. Screenshot of HPIminer gene/protein list to view related interactions and pathways. (A) Interaction information of individual protein, (B) interaction information of list of all proteins and (C) pathway information of individual protein.

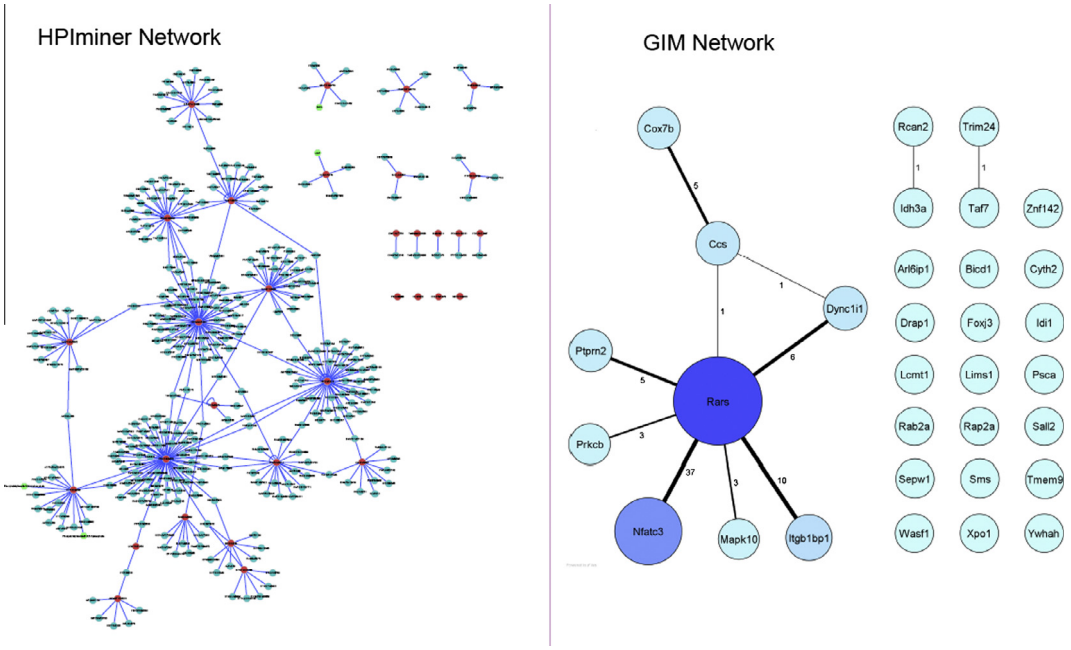

Fig. 10. Combined network generated by HPIminer and context-based network generated by GIM.

the common pairing partners may also be responsible for causing Alzheimer's disease. Similarly, the protein GSK3B/2932 is a common pairing partner for three target genes/proteins namely PRKCB/5579, BICD1/636 and DYNC1I1/1780. Therefore, protein GSK3B/2932 is more prone to cause Alzheimer's disease. Thus the network of HPIminer provides additional information on 30 neighboring proteins (Table 3). Further, the common interacting partners of these proteins predicted by HPIminer were found to be validated targets in several experimental studies as reported in the literature. For example, genes such as GSK3B, ESR1, ESR2 and FXR2 were known to increase the risk of Alzheimer's disease through several polymorphic studies [36,42–44].

Further, HPIminer provides two additional knowledge which includes (i) protein–nonprotein interactions and (ii) pathway related to proteins. For example, the combined network of HPIminer shows the interaction of disease causing genes NFATC3/4775 and IDH3A/3419 with nonprotein entities such as DNA and ADP. The overall combined network of HPIminer shows 370 nodes (proteins/genes) and 380 edges (shows interactions). Among them, 32 proteins are target proteins related to Alzheimer's disease, 4

Table 3
List of common interacting protein partners.

Target genes/proteins related to Alzheimer's disease		Common interacting protein partners
Protein A	Protein B	
DRAP1	TAF7	TAF9
TAF7	TRI124	TAF11
TRIM24	YWHAH	ESRI; ESR2; THRA; NR3C1
TAF7	YWHAH	YWHAE
TAF7	XPO1	AHR
YWHAH	WASF1	BAD; ABL1
YWHAH	XPO1	CDKN1B
WASF1	XPO1	CRK
WASF1	MAPK10	CDK5
MAPK10	CYTH2	ARRB2
YWHAH	PRKCB	YWHAG; PDPK1
CCS	YWHAH	YWHAG
CCS	PRKCB	YWHAG
YWHAH	RAP2A	RAF1
PRKCB	RAP2A	GRIN2D; GRIN1
XPO1	RAP2A	SMAD1
XPO1	LIMS1	SMURF1
RAP2A	LIMS1	TGFBR1
PRKCB	LCMT1	PPP2CB
LCMT1	ARL6IP1	FXR2
YWHAH	RAB2A	PRKCI
PRKCB	BICD1	GSK3A; GSK3B
PRKCB	DYNC1I1	GSK3B

are nonprotein entities and 334 are neighbor proteins (Supplementary file 4). Such additional information such as protein–nonprotein interaction, related pathways were not available in none of the present network visualization systems. These results imply that HPIminer is more versatile in extraction and visualization of PPI and pathways than other existing systems.

4. Conclusion

We have implemented an integrated text mining system HPIminer for extracting and visualizing human protein–protein interactions, interaction networks and pathways. HPIminer integrates three of our earlier developed text mining tools (i) NAGGNER for protein/gene name tagging (ii) ProNormz for protein/gene name normalization and (iii) PPInterFinder for extracting protein–protein interactions and two external knowledge sources (i) HPRD for existing protein interactions and (ii) KEGG for extracting pathway information. The two additional modules Iminer and Pminer of HPIminer combines resulting text mining PPI pairs with external knowledge resources for the visualization of interaction networks and pathways. To our knowledge HPIminer is first system which integrates text mining PPI information and pairs with curated databases to visualize the both interaction networks and pathways with many network visualization options.

Acknowledgments

The research has received funding from the Department of Biotechnology (DBT), Government of India, Grant No. BT/PR15378/BID/07/361/2011.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbi.2015.01.006.

References

[1] Mathivanan S, Periaswamy B, Gandhi TK, Kandasamy K, Suresh S, Mohmood R, et al. An evaluation of human protein–protein interaction data in the public domain. BMC Bioinform 2006;7(Suppl. 5):S19.
[2] Yu N, Seo J, Rho K, Jang Y, Park J, Kim WK, et al. hiPathDB: a human-integrated pathway database with facile visualization.. Nucleic Acids Res 2012;40(Database issue):D797–802.
[3] Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000;28(1):27–30.
[4] Kann MG. Protein interactions and disease: computational approaches to uncover the etiology of diseases. Brief Bioinform 2007;8(5):333–46.
[5] Huang M, Ding S, Wang H, Zhu X. Mining physical protein–protein interactions from the literature. Genome Biol 2008;9(Suppl 2):S12.
[6] Chowdhary R, Tan SL, Zhang J, Karnik S, Bajic VB, Liu JS. Context-specific protein network miner – an online system for exploring context-specific protein interaction networks from the literature. PLoS ONE 2012;7(4):e34480.
[7] Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, et al. Human Protein Reference Database – 2009 update. Nucleic Acids Res 2009;37(Database issue):D767–72.
[8] Ceol A, Chatr Aryamontri A, Licata L, Peluso D, Briganti L, et al. MINT, the molecular interaction database – 2009 update. Nucleic Acids Res 2010;38(Database issue):D532–9.
[9] Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucleic Acids Res 2006;34(Database issue):D535–9.
[10] Pagel P, Kovac S, Oesterheld M, Brauner B, Dunger-Kaltenbach I, Frishman G, et al. The MIPS mammalian protein–protein interaction database. Bioinformatics 2005;21(6):832–4.
[11] Beuming T, Skrabanek L, Niv MY, Mukherjee P, Weinstein H. PDZBase: a protein–protein interaction database for PDZ-domains. Bioinformatics 2005;21(6):827–8.
[12] Aranda B, Achuthan P, Alam-Faruque Y, Armean I, Bridge A, Derow C, et al. The IntAct molecular interaction database in 2010. Nucleic Acids Res 2010;38(Database issue):D525–31.
[13] Kuhn M, von Mering C, Campillos M, Jensen LJ, Bork P. STITCH: interaction networks of chemicals and proteins. Nucleic Acids Res 2008;36(Database issue):D684–8.
[14] Cusick ME, Yu H, Smolyar A, Venkatesan K, Carvunis AR, Simonis N, et al. Literature-curated protein interaction datasets. Nat Methods 2009;6(1):39–46.
[15] Kell DB, Oliver SG. Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. BioEssays 2004;26(1):99–105.
[16] Heiner M, Koch I, Will J. Model validation of biological pathways using Petri nets – demonstrated for apoptosis. Biosystems 2004;75(1–3):15–28.
[17] Luciano JS, Stevens RD. e-Science and biological pathway semantics. BMC Bioinform 2007;8(Suppl 3):S3.
[18] Ye Y, Doak TG. A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. PLoS Comput Biol 2009;5(8):e1000465.
[19] Ananiadou S, Kell DB, Tsujii J. Text mining and its potential applications in systems biology. Trends Biotechnol 2006;24(12):571–9.
[20] Kemper B, Matsuzaki T, Matsuoka Y, Tsuruoka Y, Kitano H, Ananiadou S, et al. PathText: a text mining integrator for biological pathway visualizations. Bioinformatics 2010;26(12):i374–81.
[21] Kell DB. Systems biology, metabolic modelling and metabolomics in drug discovery and development. Drug Discov Today 2006;11(23–24):1085–92.
[22] Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, et al. Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res 2011;39(Database issue):D691–7.
[23] Caspi R, Altman T, Dale JM, Dreher K, Fulcher CA, Gilham F, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res 2010;38(Database issue):D473–9.
[24] Bader GD, Cary MP, Sander C. Pathguide: a pathway resource list. Nucleic Acids Res 2006;34(Database issue):D504–6.
[25] Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, et al. Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res 2003;13(10):2363–71.
[26] Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res 2012;40(Database issue):D109–14.
[27] Raja K, Subramani S, Natarajan J. A hybrid named entity recognition for tagging human proteins/genes. Int J Data Min Bioinform 2014;10(3):315–28.
[28] Subramani S, Raja K, Natarajan J. ProNormz – an integrated approach for human proteins and protein kinases normalization. J Biomed Inform 2014;47:131–8.
[29] K. Raja, S. Subramani and J. Natarajan. PPInterFinder – a mining tool for extracting causal relations on human proteins from literature, Database (Oxford), 2013, bas052.

[30] Lopes CT, Franz M, Kazi F, Donaldson SL, Morris Q, Bader GD. Cytoscape Web: an interactive web-based network browser. Bioinformatics 2010;26(18):2347–8.
[31] C.N. Arighi, B. Carterette, K.B. Cohen, M. Krallinger, W.J. Wilbur, P. Fey, et al. An overview of the BioCreative 2012 Workshop Track III: interactive text mining task, Database (Oxford), 2013, bas056.
[32] Bunescu R, Ge R, Kate RJ, Marcotte EM, Mooney RJ, Ramani AK, et al. Comparative experiments on learning information extractors for proteins and their interactions. Artif Intell Med 2005;33(2):139–55.
[33] Fundel K, Kuffner R, Zimmer R. RelEx – relation extraction using dependency parse trees. Bioinformatics 2007;23(3):365–71.
[34] Aranda B, Achuthan P, Alam-Faruque Y, Armean I, Bridge A, Derow C, et al. IntAct Dataset, the IntAct molecular interaction database in 2010. Nucleic Acids Res 2010;38(Database issue):D525–D5231.
[35] Bui QC, Katrenko S, Sloot PM. A hybrid approach to extract protein–protein interactions. Bioinformatics 2011;27(2):259–65.
[36] Kwok JB, Loy CT, Hamilton G, Lau E, Hallupp M, Williams J, et al. Glycogen synthase kinase-3beta and tau genes interact in Alzheimer's disease. Ann Neurol 2008;64(4):446–54.
[37] Schwikowski B, Uetz P, Fields S. A network of protein–protein interactions in yeast. Nat Biotechnol 2000;18(12):1257–61.
[38] Chua HN, Sung WK, Wong L. Using indirect protein interactions for the prediction of Gene Ontology functions. BMC Bioinform 2007;8(Suppl 4):S8.
[39] Ikin A, Riveros C, Moscato P, Mendes A. The Gene Interaction Miner: a new tool for data mining contextual information for protein–protein interaction analysis. Bioinformatics 2010;26(2):283–4.
[40] Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, et al. STRING 8 – a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res 2009;37(Database issue):D412–6.
[41] Nikitin A, Egorov S, Daraselia N, Mazo I. Pathway studio – the analysis and navigation of molecular networks. Bioinformatics 2003;19(16):2155–7.
[42] den Heijer T, Schuit SC, Pols HA, van Meurs JB, Hofman A, Koudstaal PJ, et al. Variations in estrogen receptor alpha gene and risk of dementia, and brain volumes on MRI. Mol Psychiatry 2004;9(12):1129–35.
[43] Wang L, Andersson S, Warner M, Gustafsson JA. Morphological abnormalities in the brains of estrogen receptor beta knockout mice. Proc Natl Acad Sci USA 2001;98(5):2792–6.
[44] Malter JS, Ray BC, Westmark PR, Westmark CJ. Fragile X Syndrome and Alzheimer's Disease: Another Story About APP and β-amyloid. Curr Alzheimer Res 2010;7(3):200–6.