Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2014, 6(2):359-367

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

An integrated web platform on disease-associated proteins

Chien-Hung Huang[a], Wen-Wei Zhuang[a] and Ka-Lok Ng[b,c*]

[a]*Department of Computer Science and Information Engineering, National Formosa University, Taiwan*
[b]*Department of Biomedical Informatics, Asia University, Taiwan*
[c]*School of Pharmacy, China Medical University, Taichung, Taiwan*

ABSTRACT

In human, a large portion of genes undergoes alternative splicing then translates into different protein isoforms. Translated proteins are activated or repressed through post-translational modification (PTM). Biological processes are mediated by protein-protein interaction (PPI). Many research studies suggested that disease formation involves differential expression of isoforms. Furthermore, both of PTM and PPI are essential for the signal transduction mechanism where defects in such process may lead to disease formation. In this work, disease-associated genes, proteins, alternative products, PTM, Gene Ontology (GO) annotations, subcellular localization and PPI information are integrated to provide a sophisticated platform for disease studies. A total of 39disease types and 47 subcellular localizations information are included in the platform. This platform also provided an index, Jaccard index; to quantify the portion of common proteins involved for any two of the diseases, which may be useful for comorbidity study. A few subcellular localization specific PPI information are available in Cytoscape display format. A web platform has been set up to display the results; it can be accessed at http://bioinfo.csie.nfu.edu.tw/Dis/index.php. Finally, using lung cancer associated genes as a case study, we demonstrate how to use the web platform resource to discover further disease information.

Keywords: Human disease, alternative splicing, post-translational modification, subcellular localization, protein-protein interaction, disease comorbidity.

INTRODUCTION

It is known that many genes or proteins are associated with human diseases. A strategy to gain a better understanding of disease formation is to integrate different levels of information from various biological molecular datasets. The availability of disease genes, proteins, alternative products, post-translational modification (PTM), Gene Ontology (GO) annotations, subcellular localization and protein-protein interaction (PPI) information; has made it possible to study human disease at a system level. In the present work, the above-mentioned datasets were used to construct a sophisticated platform for disease studies.

1.1. Diseases and Alternative Splicing

The protein-coding regions of most eukaryotic genes contain exons and introns, where introns are spliced, producing mRNA and translated into proteins. Differential expression of the p53 isoforms has been reported in renal cell carcinoma [1] and head and neck tumors [2].

When alternative splicing occurs in exons, part of the mRNA is removed, which results in altering protein domain composition and PPI. It is suggested that PPI, which is mediated by domain-domain interaction (DDI), may be affected by domain removal due to alternative splicing. In a previous work [3], we reported that the result effects can be divided into two categories: (i) the PPI due by differential domains interactions [4] and (ii) the PPI due to common domains interactions. The importance of alternative splicing in cancer formation can also be found in another

work [5], where the authors identified seven tumor-specific splice variants (ACTN1, CALD1, COL6A3, LRRFIP2, PIK4CB, TPM1, and VCL) in colon, bladder and prostate cancer. These results suggested that different isoforms may be involved in the pathogenesis of cancers and hence represent novel therapeutic targets.

1.2. Diseases and Post-translation Modification

Disease is a result of the emergence of problems in the communication system of the cells, that is, some problems are occurred in signal transduction pathways when the cells communicate with each other. The communications, either inter or intra cells are transmitted through the actions of bimolecular, such as growth factors, hormones, cytokines and the swapping of neurotransmitters.

Signal transduction pathways start with the binding of the cell membrane receptor protein with ligand, and then with PPI and PTM of the signaling molecule. For instance, extra-cellular signal, such as transcription factor, propagates and finally enters cell nucleus to activate or suppress genes.

One of the major purposes of this will be to study the relationship between signal transduction pathway and disease formation; such as, carcinogenesis. Three post-genomic era data; i.e. PPI, PTM of kinases and transcription factors, will be investigated in the near future.

Over the years, a growing number of experimental PPI data are available to researchers, such as DIP [6], IntAct [7] (http://mips.gsf.de/proj/ppi/), Mammalian Protein-Protein Interaction Database (MIPS) [8] and Human Protein Reference Data (HPRD) [9, 10]. These data are applied to cancer study [11], identifying cancer biomarkers [12, 13], and discover oncoprotein interaction network based on network topological structure analysis [14]. After protein translation synthesis, PTM is a rather common mechanism that activates or represses protein function. Phosphorylation, glycosylation and acetylation are three of the most popular PTM. Protein phosphorylation plays an important role in cellular regulation, if some problems are occurred in the PTM of the signal transduction pathway, such as the MAPK pathway, or cell proliferation pathway, it may results in cancer formation [15, 16]. However, the importance of protein post-modification and protein phosphor-proteome in cancer research did not pay much attention until recent years [17-19], due to the well developing of proteome technology, such as 2D gel and mass spectrometry. The research scope consists of identifying cancer biomarkers [19], how to conduct cancer treatment through protein drug targeting signal transduction pathway [20, 21], and designing drugs for regulating protein phosphorylation in signal transduction pathway [18-19, 22-23].

EXPERIMENTAL SECTION

1.3. Data Source

The data in our web platform are integrated and parsed from the following four databases: (1) Uniprot (http://www.uniprot.org/),(2) Gene Ontology (GO, http://www.geneontology.org/) (3) BioGrid (http://thebiogrid.org/), and OMIM (http://www.ncbi.nlm.nih.gov/omim/). The web platform obeys a multi-tiered architecture, and the data integration process is shown in Figure 1. Furthermore, Figures 2A and 2B represents the database E-R model and use case diagram of this system, respectively.

1.4. Protein Information from Disease and Subcellular Localization

OMIM is a database that catalogues the known diseases with a genetic component, while Uniprot is a database to provide relationship between genetic component, proteins and subcellular locations genomic analysis of a catalogued gene. By merging OMIM and Uniprot databases, we can derive the associated protein information according to the specified disease and subcellular localization.

Furthermore, Gene Ontology database stores molecular function information, which is also very useful in bioinformatics research. Our web platform also merged the GO database to provide biological annotation for disease-associated proteins.

A given protein may have distinct identifiers naming in different databases, therefore, Gene Name Service (http://bioagent.iis.sinica.edu.tw/GeneAlias/) is used to solve the issue of inconsistency. The protein related function can be accessed by Uniprot or Swissprot ID in our web platform.

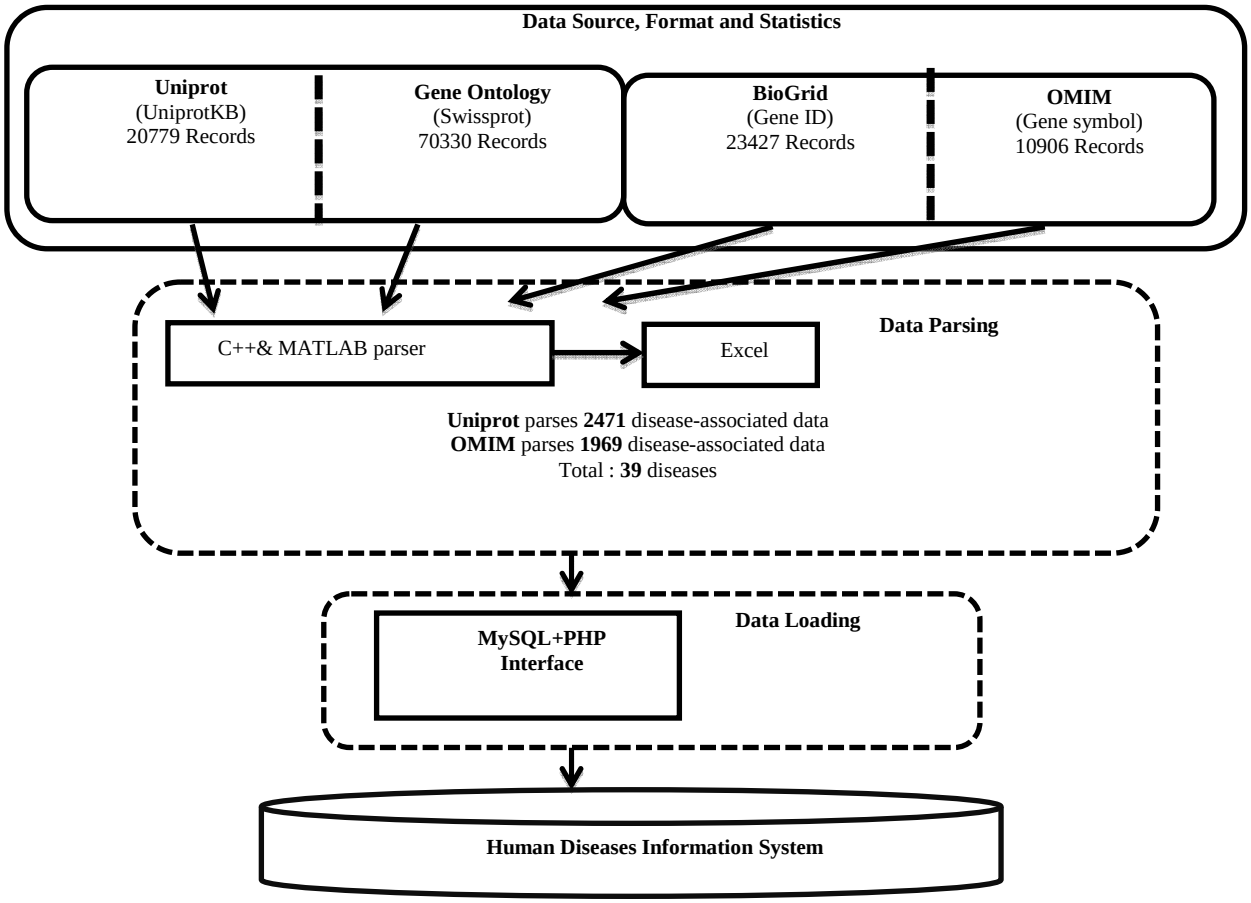


Fig. 1. An overview of the data integration process

1.5. *Comorbidity Study*

The present platform also provided an index, *Jaccard Index (JI)*; to quantify the number of common proteins involved in two diseases. It may be a useful starting point for disease comorbidity study. *JI* is a quantity which is used to quantify the similarity between two sets; hence, given two modules *A* and *B*, *JI* is given by:

$$JI(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

where $|A \cap B|$ and $|A \cup B|$ denote the cardinality of $A \cap B$ and $A \cup B$ respectively.

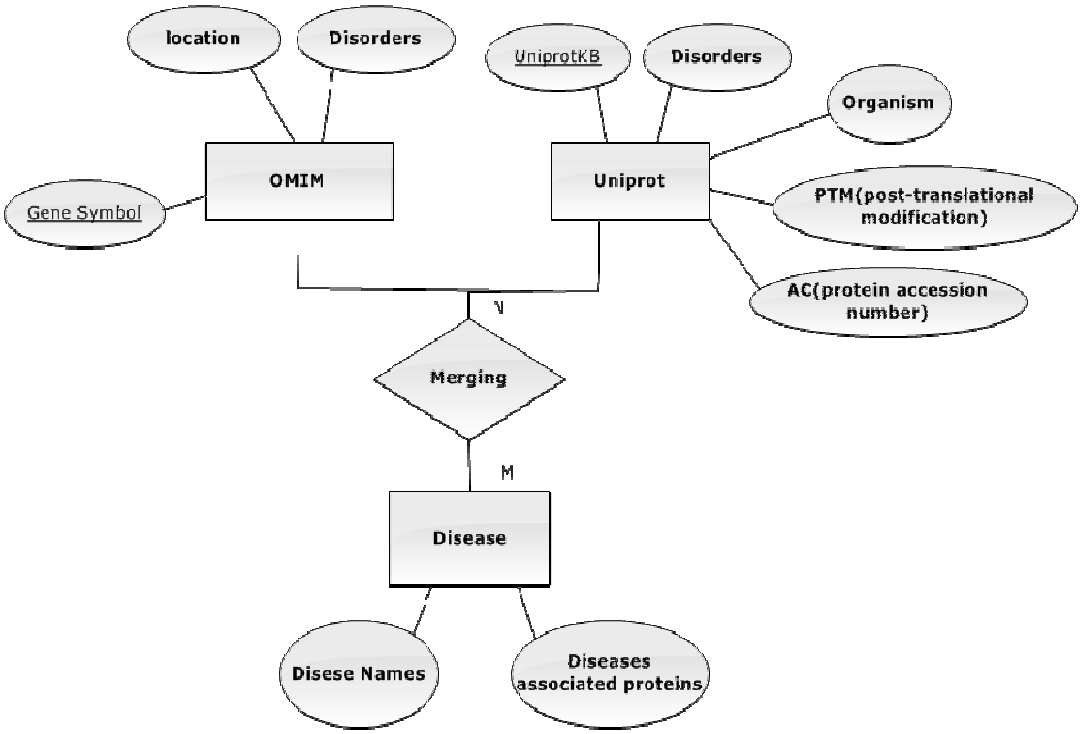


Fig. 2A. E-R model of the integrated database

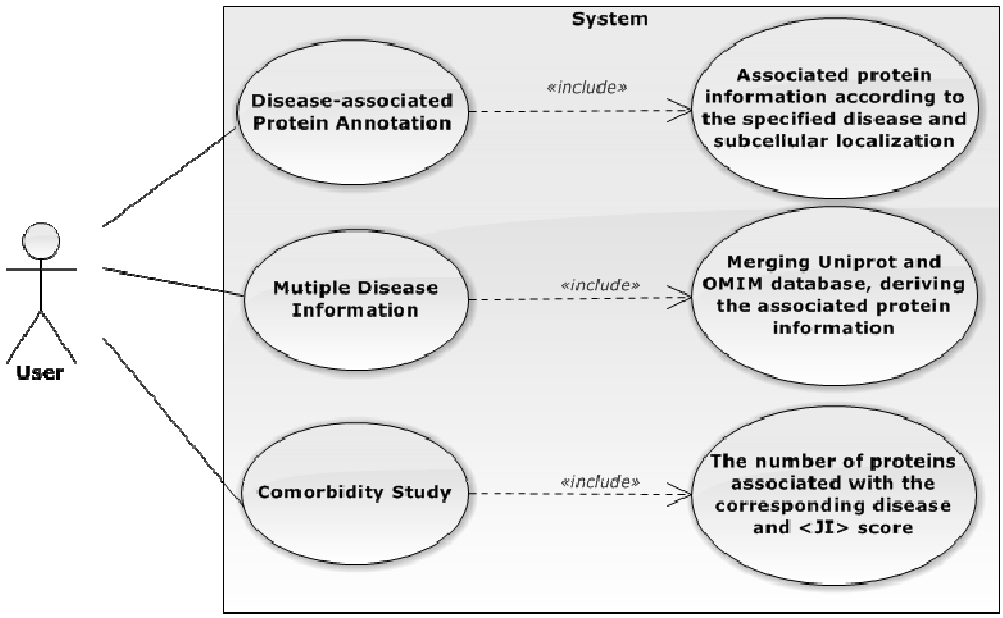


Fig. 2B. Use case diagram of the proposed web platform

1.6. Subcellular Localization Specific Protein-Protein Interactions

Protein-protein interaction network can be treated as a simple undirected graph, where each protein is mapped to a node and the interaction between two proteins is mapped to an edge. PPI data indicate relationships of each protein pair. However, protein can be found at different subcellular localization [24]; therefore, it is important to realize that two proteins can interact if they locate in the same localization.

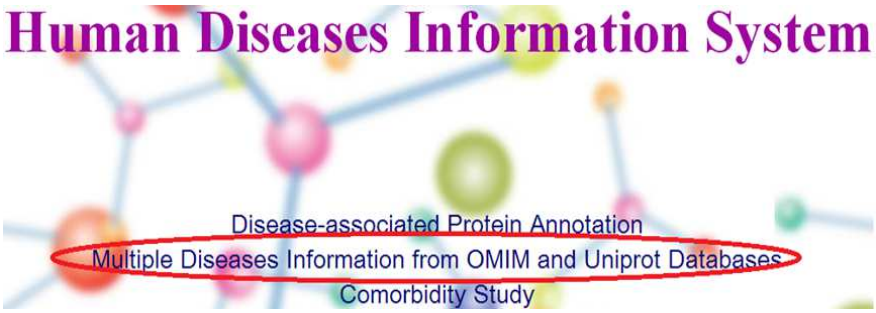
The BioGrid database provides PPI data for many species. By integrating human PPI data in BioGrid with the sub-cellular localization data, one can possibly recover more realistic biological information. This is because some of the PPI events are not allowed simply due to the fact that the two proteins are not resided in the same localization.

RESULTS

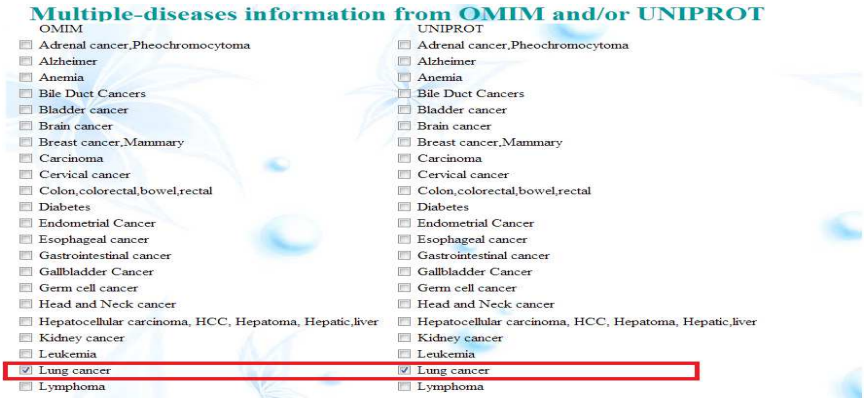
1.7. Associated Protein Information and Multiple Disease Information

Firstly, by selecting disease type and subcellular localization from the drop down menus in the web platform, users can retrieve the corresponding protein information, such as AC (protein accession number), OS (organism), PTM (post-translational modification) and alternate product. Moreover, the second major function of our web platform is to provide the associated information retrieved from OMIM and Uniprot databases for any disease combinations.

A



B



C

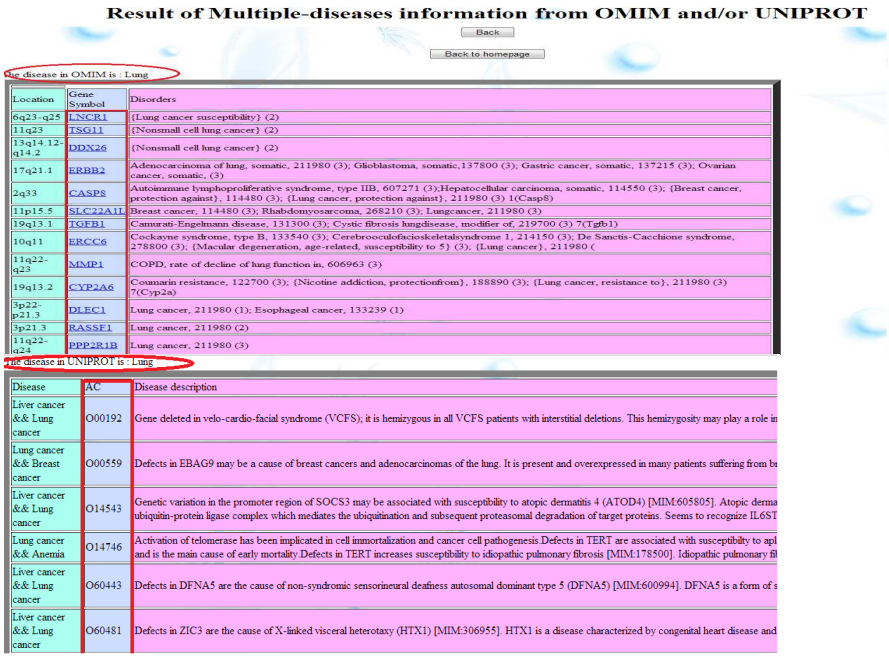
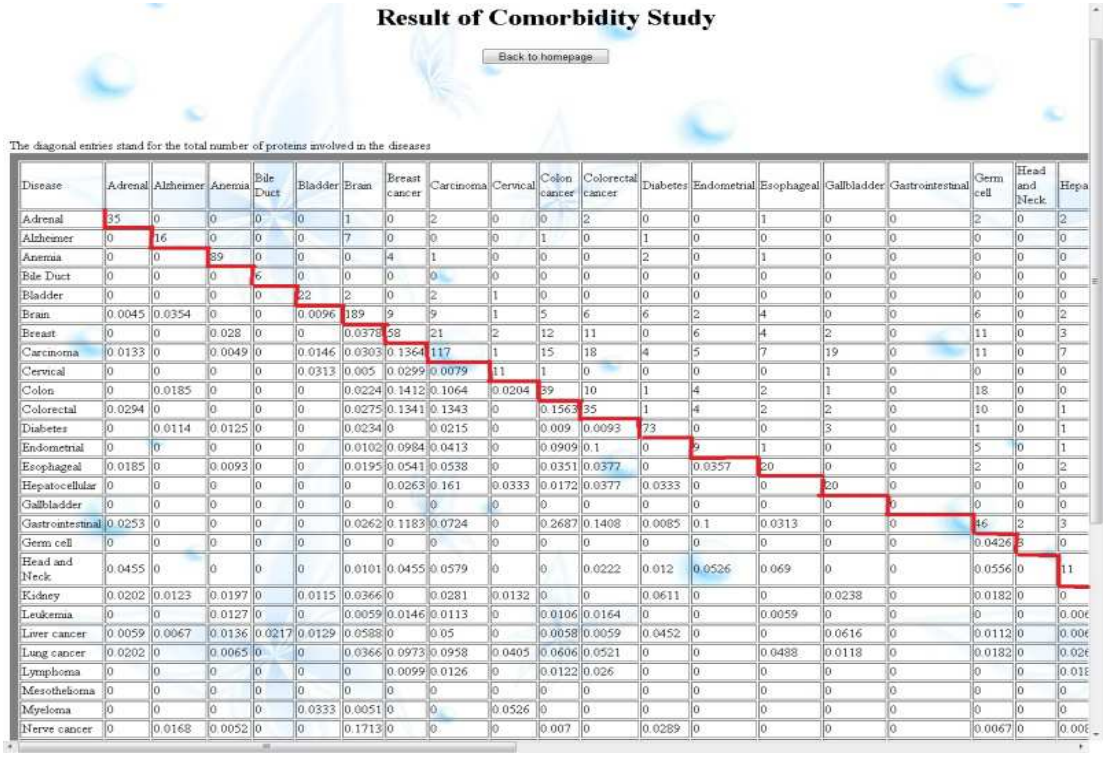


Figure 3. The steps for retrieving lung cancer associated genes/proteins in the proposed website

For example, following steps in Figure 3, users can extract lung cancer associated genes and proteins from OMIM and Uniprot respectively.

1.8. Comorbidity Study

The third major function of our web platform is to provide the comorbidity protein information and *JI* score for each disease pair. As shown in Figure 4, the first row and the first column list all disease type. The entries in the diagonal represent the number of proteins associated with the corresponding disease. Entries in upper-diagonal denote the number of common proteins for each disease pair, and in contrast, entries in lower-diagonal show its *JI* score. The top 10 disease pairs with the highest *JI* score are listed in descending order in Table 1. Through artificial mining in NCBI PubMed literatures, disease pairs with higher *JI* score truly tend to have closer relation.



Disease	Adrenal	Alzheimer	Anemia	Bile Duct	Bladder	Brain	Breast cancer	Carcinoma	Cervical	Colon cancer	Colorectal cancer	Diabetes	Endometrial	Esophageal	Gallbladder	Gastrointestinal	Germ cell	Head and Neck	Hepa
Adrenal	35	0	0	0	0	1	0	2	0	0	2	0	0	1	0	0	2	0	2
Alzheimer	0	16	0	0	0	7	0	0	0	1	0	1	0	0	0	0	0	0	0
Anemia	0	0	89	0	0	0	4	1	0	0	0	2	0	1	0	0	0	0	0
Bile Duct	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Bladder	0	0	0	0	22	2	0	2	1	0	0	0	0	0	0	0	0	0	0
Brain	0.0045	0.0354	0	0	0.0096	189	9	9	1	5	6	6	2	4	0	0	6	0	2
Breast	0	0	0.028	0	0	0.0378	58	21	2	12	11	0	6	4	2	0	11	0	3
Carcinoma	0.0133	0	0.0049	0	0.0146	0.0303	0.1364	117	1	15	18	4	5	7	19	0	11	0	7
Cervical	0	0	0	0	0.0313	0.005	0.0299	0.0079	11	1	0	0	0	0	1	0	0	0	0
Colon	0	0.0185	0	0	0	0.0224	0.1412	0.1064	0.0204	39	10	1	4	2	1	0	18	0	0
Colorectal	0.0294	0	0	0	0	0.0275	0.1341	0.1343	0	0.1563	35	1	4	2	2	0	10	0	1
Diabetes	0	0.0114	0.0125	0	0	0.0234	0	0.0215	0	0.009	0.0093	73	0	0	3	0	1	0	1
Endometrial	0	0	0	0	0	0.0102	0.0984	0.0413	0	0.0909	0.1	0	9	1	0	0	5	0	1
Esophageal	0.0185	0	0.0093	0	0	0.0195	0.0541	0.0538	0	0.0351	0.0377	0	0.0357	20	0	0	2	0	2
Hepatocellular	0	0	0	0	0	0	0.0263	0.161	0.0333	0.0172	0.0377	0.0333	0	0	20	0	0	0	0
Gallbladder	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Gastrointestinal	0.0253	0	0	0	0	0.0262	0.1183	0.0724	0	0.2687	0.1408	0.0085	0.1	0.0313	0	0	46	2	3
Germ cell	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0426	0	0
Head and Neck	0.0455	0	0	0	0	0.0101	0.0455	0.0579	0	0	0.0222	0.012	0.0526	0.069	0	0	0.0556	0	11
Kidney	0.0202	0.0123	0.0197	0	0.0115	0.0366	0	0.0281	0.0132	0	0	0.0611	0	0	0.0238	0	0.0182	0	0
Leukemia	0	0	0.0127	0	0	0.0059	0.0146	0.0113	0	0.0106	0.0164	0	0	0.0059	0	0	0	0	0.006
Liver cancer	0.0059	0.0067	0.0136	0.0217	0.0129	0.0588	0	0.05	0	0.0058	0.0059	0.0452	0	0	0.0616	0	0.0112	0	0.006
Lung cancer	0.0202	0	0.0065	0	0	0.0366	0.0973	0.0958	0.0405	0.0606	0.0521	0	0	0.0488	0.0118	0	0.0182	0	0.026
Lymphoma	0	0	0	0	0	0	0.0099	0.0126	0	0.0122	0.026	0	0	0	0	0	0	0	0.018
Mesothelioma	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Myeloma	0	0	0	0	0.0333	0.0051	0	0	0.0526	0	0	0	0	0	0	0	0	0	0
Nerve cancer	0	0.0168	0.0052	0	0	0.1713	0	0	0	0.007	0	0.0289	0	0	0	0	0.0067	0	0.008

Fig. 4. The number of common proteins and *JI* score for each disease pair

Table 1 Top 10 disease pairs with the highest *JI* score

Disease pair	*JI* score
(Gastrointestinal, Colon cancer)	0.2687
(Stomach, Colon cancer)	0.2162
(Stomach, Germ cell)	0.1915
(Stomach, Endometrial)	0.1875
(Renal, Endometrial)	0.1845
(Nerve, Brain)	0.1713
(Stomach, Breast cancer)	0.1525
(Colon cancer, Breast cancer)	0.1412
(Carcinoma, Breast cancer)	0.1364
(Colon cancer, Carcinoma)	0.1343

1.9. Subcellular Localization Specific Protein-Protein Interactions Study

Users are allowed to search for subcellular localization specific protein information by using the drop down menu. Subcellular localization specific PPI is available for cytoplasm, extracellular, membrane and nucleus. Our web platform provides PPI on specific subcellular localization, and the PPI data are ready for downloading for further study. In particular, the PPI data is prepared in Cytoscape http://www.cytoscape.org/) input format, which facilitate visualization purpose. The visualization graph proposed in this work can clearly show the interaction between each protein pair, and it can be extended to combine various graph clustering algorithms to predict protein attributes and protein complexes, which is critical for large-scale data analysis.

1.10. Case Study

The following case study demonstrates how to discover useful disease-related information by the proposed web platform.

In [25], by Robust Multi-array Average and Empirical Bayes(eBayes) statistics, the current authors extracted the most significant differentially expressed genes (DEGs) for lung cancer from the experimental verified microarray data, E-TABM-15, which consist of up and down regulated genes with *p*-values less than 0.0007. Among these DEGs, 952 and 1339 genes belong to the up and down group respectively. After querying the proposed web platform, a total of 127 lung cancer associated genes is retrieved, and among them, 25 genes are found to be the same with DEGs in [25]. The 25 genes are listed in Table 2, which can be provided for further drug-gene interaction and potential drugs investigation. By using the network visualization tool, Cytoscape, Figure 5 shows the PPI network of the 25 genes. In this network, proteins and interactions are drawn as nodes and edges respectively. Up-regulated genes, down-regulated genes and other adjacent genes are represented by different colors and shapes. The PPI information could possibly provide some crucial biological pathways for further investigation.

Table 2 The Gene symbols of overlapped lung cancer related genes by comparing the query results and DEGs data

UP regulated genes	Down regulated genes
ABCB4	ASPA
ATPAF2	CYBRD1
CP	DMPK
GALE	FBP1
HGD	GLUD1
HMBS	HADHA
MMACHC	HLA-DRA
SLC25A13	HPS5
TFR2	JAG1
UROS	LMNA
	MMRN1
	NDRG2
	PNPLA2
	RBP4
	SOCS3

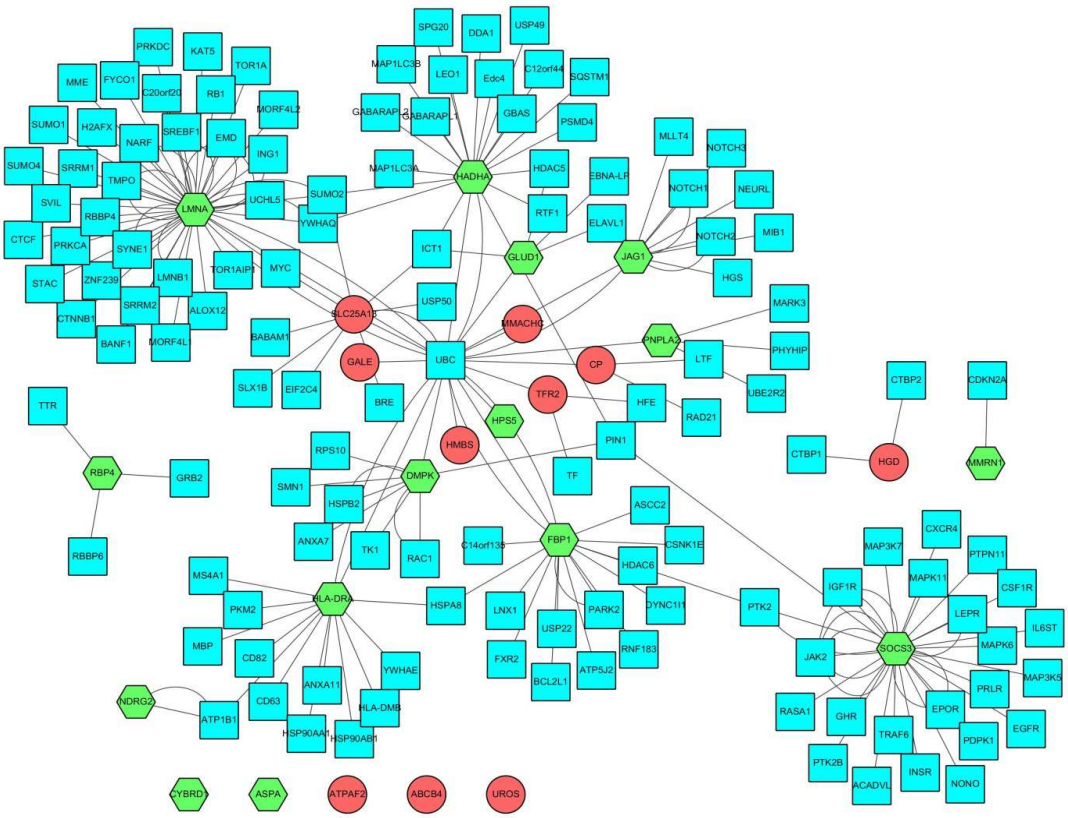


Fig. 5. PPI of the 25 genes in Table 2. Red color circles and green color hexagons denote up- and down- regulated genes respectively

DISCUSSION

In this study, disease-associated genes, proteins, alternative products, PTM, GO annotations, subcellular localization and PPI information are adopted to provide a sophisticated platform for biomedical researchers. A total of 39 disease types and 47 subcellular localizations are included in the platform. Subcellular localization specific PPI information is available in Cytoscape display format, which facilitate graph topological analysis. We also demonstrate by a case study that the proposed web platform can perform network inference for further disease information analysis.

It is expected that the platform should be of value for future studies into understanding molecular mechanism of disease formation and identify therapeutic drug targets. There are two tasks are undergoing or to be completed in the near future. The first one is to perform gene set enrichment analysis, i.e. GSEA, and pathway analysis, to identify enriched biological processes and pathways. The second task is to include lung cancer microarray data results for therapeutic target identification [25].

CONCLUSION

In conclusion, we have developed a pipeline to provide bio-molecular information for a rather comprehensive list of major diseases. The study starts from disease query, and then gene, protein, subcellular localization specific PPI, and GO annotation data can be retrieved. Although lung cancer is considered in our case study, the same analysis can be easily extended to other disease type as long as the microarray data is available. Furthermore, our platform provides the comorbidity protein information and *JI* score for each disease pair, which is useful for exploring the interplay between disease comorbidity [26].

Acknowledgments
The work of Chien-Hung Huang is supported by the National Science Council of Taiwan under grant NSC 101-2221-E-150-088-MY2. The work of Ka-Lok Ng is supported by the grants NSC 101-2221-E-468-027, NSC 102-2221-E-468-024, and NSC 102-2632-E-468-001-MY3.

REFERENCES

[1] W Song; SW Huo; JJ Lü; Z Liu; XL Fang; XB Jin; MZ Yuan. *Chin Med J*, **2009**, 122(8), 921-926.
[2] L Boldrup; JC Bourdon; PJ Coates; B Sjöström; K Nylander. *Eur J Cancer*, **2007**, 43(3), 617-623.
[3] YL Lee; JW Weng; WC Chiang; YW Lin; KL Ng; JJP Tsai; CY Huang. *IEEE International Conference on Bio-informatics and Bioengineering (BIBE)*, **2011**, 1(1), 33-38.
[4] S Amelie; C Arnaud; A Patrick. 3did: *Nucl Acids Res*, **2011**, 39 (Database issue):D718-23.
[5] EV Kriventseva; I Koch; R Apweiler; M Vingron; P Bork; MS Gelfand; S Sunyaev. *TRENDS in Genetics*, **2003**, 19(3), 124-128.
[6] L Salwinski; C S Miller; AJ Smith; FK Pettit; JU Bowie; D Eisenberg. *Nucl Acids Res*, **2004**, 32 (Database issue):D449-451.
[7] H Hermjakob et al. *Nucl Acids Res*, **2004**, 32 (Database issue):D452-D455.
[8] P Pagel et al. *Bioinfo*, **2005**, 21(6), 832-834.
[9] S Peri et al. *Genome Res*, **2003**, 13, 2363-2371.
[10] GR Mishra et al. *Nucl Acids Res*, **2006**, 34 (Database issue):D411-414.
[11] A Schramm; O Apostolov; B Sitek; K Pfeiffer; K Stühler; HE Meyer; W Havers; A Eggert. *Klin Padiatr*, **2003**, 215(6), 293-297.
[12] RJ Simpson; DS Dorow. *Trends Biotech*, **2001**, 19(10 Suppl), S40-48.
[13] C Mathelin; C Koehl; MC Rio. *Gynecol Obstet Fertil*, **2006**, 34(7-8), 638-646.
[14] P Hernandez; J Huerta-Cepas; D Montaner; F Al-Shahrour; J Valls; L Gomez; G Capella; J Dopazo; MA *BMC Genomics*, **2007**, 8, 185.
[15] AK Pullikuth; AD Catling. *Cell Signal*, **2007**, 19(8), 1621-1632.
[16] AS Dhillon; S Hagan; O Rath; W Kolch. *Oncogene*, **2007**, 26(22), 3279-3290.
[17] DF Stern. *Exp MolPathol*, **2001**, 70(3), 327-331.
[18] DF Stern. *Expert OpinTher Targets*, **2005**, 9(4), 851-860.
[19] YP Lim. *Clin Cancer Res*, **2005**, 11(9), 3163-3169.
[20] PJ Roberts; CJ Der. *Oncogene*, **2007**, 26(22), 3291-3310.
[21] DN Dhanasekaran; GL Johnson. *Oncogene*, **2007**, 26(22), 3097-30999.
[22] JA Kim. *Am J Surg*, **2003**, 186(3), 264-268.
[23] MF Moran; J Tong; P Taylor; RM Ewing. *Biochim Biophys Acta*, **2006**, 1766(2), 230-241.
[24] ZP Feng. *in Silico Biology*, **2002**, 2(3), 291-303.
[25] CH Huang; MY Wu; CY Huang; KL Ng. *International Conference on Bioinformatics (ICB'13)*, **2013**, 95-98.

[26] J Park; DS Lee; NA Christakis; AL Barab´asi. *Molecular Systems Biology*, **2009**, 5(262), 1-7.