International Journal of PharmTech Research
CODEN (USA): IJPRIF, ISSN: 0974-4304
Vol.8, No.1, pp 127-134, 2015

Deciphering relationships in disease networks using computational approaches: Fatty Liver, PCOD, Osteoarthritis, cholelithiasis & hyperlipdemia

Jaisri Jagannadham[1,2], Hitesh Kumar Jaiswal[1] and Kamal Rawal[1]*

[1]Department of Biotechnology, Jaypee Institute of Information Technology, Noida (UP) -201 307, India
[2]School of Life Sciences, Jaipur National University, Jagatpura, Jaipur (Rajasthan), India

Abstract: Most of the diseases are complex in nature, often influenced by genes, SNPs, proteins, pathways, environmental factors, and immune responses. Therefore, it is crucial to uncover their molecular pathways through network based approaches. Here, we elucidate the process of identification of disease genes and creating a disease network. We selected five diseases related to obesity which include fatty liver, hyperlipidemia, cholelithiasis, polycystic ovary disease and osteoarthritis. The genes associated with these diseases are retrieved using publically available databases and in-house developed literature-mining tool. A disease network was constructed using manually curated gene list. Finally, we predicted side effects of drugs based upon disease networks and literature mining.
Keywords: Disease network, genes, fatty liver, drug, target, hub, PCOD, Osteoarthritis, Obesity, hyperlipidemia, cholelithiasis.

Introduction

Genes, SNPs, proteins, metabolites and pathways play a major role in pathogenesis of multifactorial complex diseases such as obesity, type 2 diabetes, asthma and hypertension. Therefore, it is important to identify the genes associated with such diseases. Recently, several approaches have been reported to characterize candidate genes relevant to diseases. For example, functional genomics strategies have enabled the use of large-scale molecular and physiological data to help in discovery of gene modules that directly respond to genetic and environmental perturbations associated with the disease (1). High-throughput experimental methods such as DNA microarray (2), next-generation sequencing (3), and the two-hybrid screening system (4) as protein-protein interactions and gene expression profiles (5) have contributed tremendously in identification of candidate genes. In addition, several network based disease gene prioritization methods have been proposed (6), which include random walk method (7), CIPHER (8), PRINCE (9), MAXIF (10) and MINProp (11). Several properties of networks of disease genes have been reported in the past, for example protein products of disease gene interact with higher frequency. They also tend to get co-expressed in specific tissues(12).

Barabási et al, 2011 (13) proposed a hypothesis that a disease is rarely a consequence of an abnormality in a single effector gene product. Instead, the disease phenotype is a reflection of various patho-biological processes that interact in a complex network. Network medicine is based on a series of widely used hypotheses and organizing principles that link network structure to biological function and disease (13). In this work, we propose a new framework to understand disease networks.

Material & Method

A) Retrieval of disease genes

The genes associated with diseases in human are retrieved from databases such as the human malady compendium (MalaCards) (14), ingenuity pathway knowledge base (IPKB) (15) and gene disease database (DisGeNET) (16). In addition, our inbuilt literature mining tool in perl is used to retrieve the disease associated genes from PubMed abstracts. The step-wise process of the literature-mining tool is explained below:

1. Construction of the gene synonym list: A gene list was compiled from the HUGO Gene nomenclature committee website, www.genenames.org. A set of 35,960 genes, containing all coding and non-coding genes, was created by extracting the columns of approved names, approved symbols, previous names, previous symbols and synonyms of each gene.
2. Construction of the dataset: The abstracts were downloaded using RefNavigator (Version 2.5, ©Akossoft, 2008-2009). RefNavigator is a tool that searches for abstracts of research articles indexed in the MEDLINE database. It can also extract links to full text articles from PubMed.
3. Identification of the genes implicated in diseases: We built a text mining system to mine the genes implicated in diseases. To begin with set of 35,959 genes and their synonyms were searched in the abstracts dataset. If a synonym or a key word was found, then the preceding and successive words in the same sentence were screened for the presence of the approved name. If that was also found, the count for the particular gene was incremented by unity. The ones that revealed a positive count were compiled along with their synonyms, symbols and previous names into another file. Each name of the gene was separated by a semicolon in this file. For example, "INS; insulin" represents that, insulin is a gene that is known by two names INS (the symbol) and insulin (the approved name). A frequency of occurrence of gene symbol in a disease specific abstract dataset was also computed as a rough measure of their association.

B) Disease network & their analysis

The disease networks are constructed based on the curated gene sets. The networks are generated using GeneMania (17). It derives the knowledge for network generation from publicly available databases, which includes, Gene Expression Omnibus (GEO), BioGRID, I2D and Pathway Commons containing information from BioGRID, Memorial Sloan-Kettering Cancer Center, Human Protein Reference Database, HumanCyc, Systems Biology Center New York, IntAct, MINT, NCI-Nature Pathway Interaction Database and Reactome. Networks are generated from these data either directly or using the GeneMania in-house analysis pipeline. The functional analysis of the network genes is predicted using functional prediction algorithm. The hub object analyzer tool (Hubba) (18) is used in predicting the hubs in disease networks. The knowledge of the existing drugs and their associated targets were obtained through extensive literature survey. The side-effects associated with the drugs are retrieved from SIDER 2 database (19).

Results

Resources for disease-gene association

There is a limited availability of databases depicting the association of genes with diseases. Some of the examples are described as following : the human malady compendium (MalaCards) (14), ingenuity pathway knowledge base (IPKB) (15) and gene disease database (DisGeNET) (16) (Figure 1). MalaCards is an integrated database in biomedical research of human maladies and their annotations. Currently, MalaCards provides information on 16,919 disease entries compiled from 40 data resources. In addition, it provides the entire disease annotation such as disease name and their synonym, therapeutics, clinical tests and conditions, list of genes associated with the particular condition, network of related diseases and their relevant references (14). IPKB is a well-established database, for its rich source of information with data attained from experimental and clinical studies for all conditions. It also provides information on the expression level of a molecule implicated, such as increased, decreased, related or effected to a particular condition (15). DisGeNET is developed as a comprehensive database of human gene-disease associations by combining information from databases such as Online Mendelian Inheritance in Man (OMIM) and Pharmacogenomics Knowledge Base (PharmGKB) and from literature through literature-mining. It comprises the whole set of human diseases with genetic origin, including Mendelian, complex and environmental diseases (20).

Table 1 shows frequency of occurrences of terms related to 5 diseases in PubMed abstracts on dataset of obesity and human.

Diseases	Count in Abstracts (Obesity + human)
Diseases related to obesity	
Fatty liver	2142
Hyperlipidemia	1659
Osteoarthritis	814
Cholelithiasis	147
Polycystic Ovary disease	120
Diseases unrelated to obesity	
Lymphedema	81
Tuberculosis	80
Urolithiasis	62
Cystic fibrosis	59
Renal hypoplasia	2

Apart from above mentioned resources, we have developed an in-house automated literature-mining pipeline in Perl for extracting the genes associated with diseases through mining the PubMed abstracts. In this work, we are presenting five human diseases related to obesity to demonstrate applications of our approach. The primary data of this approach has already been used in biomedical screening of genes in obesity and its associated disorders (Jagannadham *et al* 2015. Manuscript submitted). This paper will discuss applications in fatty liver disease, PCOD, osteoarthritis, cholelithiasis and hyperlipidemia. The association of these diseases with obesity (Table 1) is proposed by their occurrences of disease term in PubMed abstracts on obesity and human as well as strongly estabilished clinical evidence. The diseases include fatty liver, hyperlipidemia, polycystic ovary disease, osteoarthritis and cholelithiasis. We found that 2.3% of total obesity abstracts have fatty liver as a key word (2142 PubMed abstracts), 1.8% abstract contain a key word hyperlipidemia, 0.9% abstracts contain osteoarthritis, 0.1% with polycystic ovary disease and cholelithiasis. As a control, we also computed occurrence of other unrelated diseases keywords in the obesity dataset and found to be comparatively less.

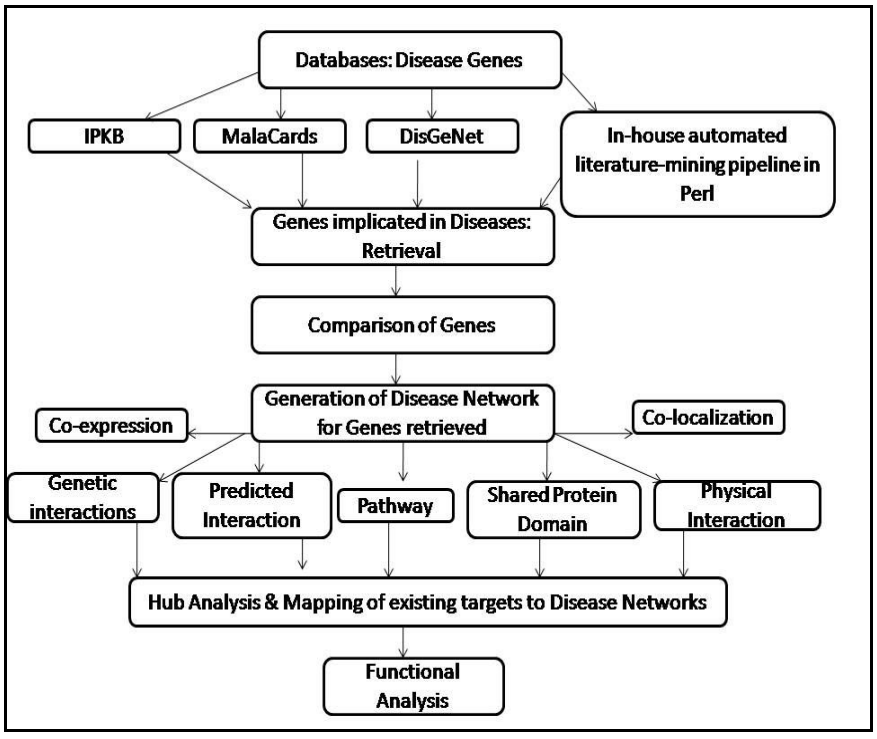

Figure1 shows the work-flow in deciphering the relationship in disease network.

Retrieval of genes associated with diseases

Understanding the mechanisms of human diseases is one of the most challenging problems in the biomedical research. The evidences suggest that diseases are not characterized by single genes, but emerge due to complex interactions between multiple genetic variants and environmental factors (21). The genes associated with the above mentioned five diseases are retrieved from MalaCards, IPKB, DisGeNET and our literature-mining tool. In our literature-mining system, we screened molecules implicated in each of the diseases by downloading the abstracts from PubMed having the disease term for instance, fatty liver and human. The genes implicated in each of these diseases are obtained by using the list of human genes (35,000) from a HUGO database (www.genenames.org). The list of genes obtained from these resources are given with their approved symbol (Supplementary Table 1). The frequency count of genes retrieved from each database is given in Table 2.

Table 2 shows the frequency of genes implicated in a 5 diseases from public resources.

Diseases	MalaCards	IPKB	DisGeNET	Inbuilt literature-mining tool
Fatty Liver	398	34	45	273
Hyperlipidemia	74	76	50	206
Osteoarthritis	630	157	149	238
Cholelithiasis	70	49	19	75
Polycystic Ovary disease	334	57	202	319

Comparison of candidate genes across databases

Candidate genes associated with diseases were retrieved from several databases and their frequency of occurrence was compared across databases (Figure 2). For example, in fatty liver, we found six genes that are reported to be common in all databases, namely MalaCards, IPKB, DisGeNET and inhouse literature-mining tool. These include leptin (LEP), perilipin 2 (PLIN2), tumor necrosis factor (TNF), peroxisome proliferator activator receptor gamma (PPARG), peroxisome proliferator activator receptor alpha (PPARA) and catalase (CAT). In cholelithiasis, 9 genes are found to reported in all databases. The example of these genes are apolipoprotein A-I (APOA1), apolipoprotein E (APOE), UDP glucuronosyltransferase 1 family, polypeptide A1 (UGT1A1), ATP-binding cassette, sub-family B member 4 (ABCB4), ATP-binding cassette, sub-family G member 8 (ABCG8), ATP-binding cassette, sub-family B member 11 (ABCB11), ATP-binding cassette, sub-family G member 5 (ABCG5), cholecystokinin A receptor (CCKAR). Likewise, we found several commom examples in hyperlipidemia, osteoarthritis and polycystic ovary disease (Supplementary Table 2).

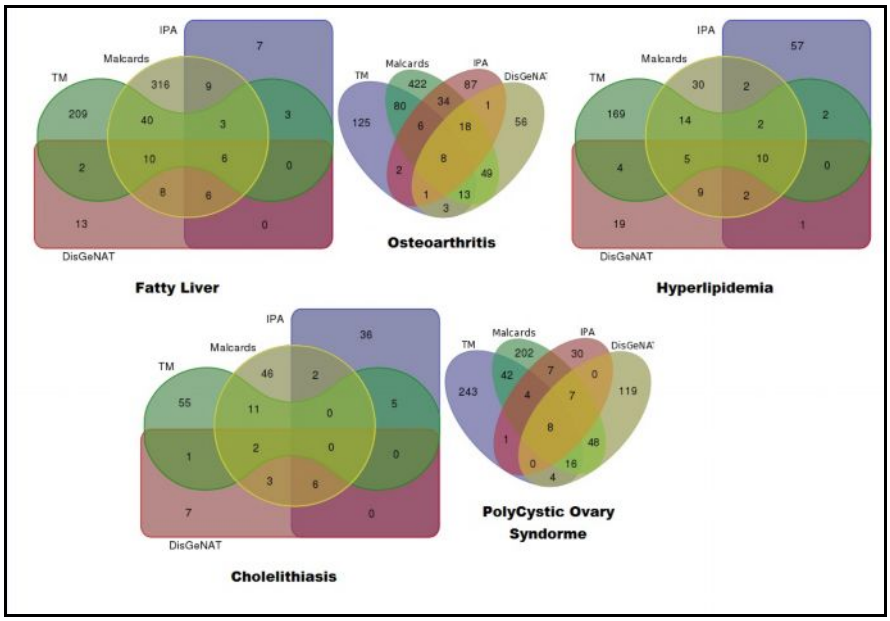

Figure 2 represents the venn diagrams to demonstrate common genes in fatty liver, osteoarthritis, hyperlipidemia, cholelithiasis and polycystic ovary disease obtained from sources such as MalaCards, IPA, literature-mining (TM) and DisGeNAT.

Creation of disease network

The genes retrieved from above mentioned databases are curated manually to understand their potential involment in a disease condition. These curated set of genes (Supplementary Table 3-7) form a set of core molecules for the disease network generation. Here, the functions of proteins encoded by these genes are computed through GeneMania (17). The networks on these disease conditions are predicted on the basis of co-expression, co-localization, physical interactions, predicted interactions, pathways, genetic interactions and their shared protein domains. Tables 3 elaborate the role of each of these in disease networks.

Table 3 shows fraction of genes categorized into different experimental resources such as co-expression, domain sharing etc.

	Fatty Liver	Polycystic Ovary disease	Hyperlipidemia	Osteoarthritis	Cholelithiasis
Co-expression	48.21%	53.33%	59.78%	63.88%	51.37%
Co-localization	20.98%	17.59%	19.10%	18.40%	14.66%
Physical interactions	12.16%	13.65%	9.46%	6.34%	11.15%
Predicted	8.41%	6.28%	3.44%	4.93%	6.56%
Pathway	5.06%	2.75%	4.18%	1.57%	9.11%
Genetic interactions	3.29%	5.05%	2.29%	3.53%	0.11%
Shared protein domains	1.87%	1.35%	1.74%	1.33%	7.05%

The disease network for fatty liver has 291 nodes (molecules) or species with 8635 interactions between them. These interactions are classified as: 4427 interactions from co-expression databases, 1753 from co-localization, 1587 genetic interactions, 285 from pathways, 301 from physical interactions, 85 predicted interactions and 197 from shared protein domains (Figure 3). The genes and their interaction evidences in fatty liver are provided in Supplementary Table 3. The polycystic ovary disease network has 331 nodes with 8022 interactions (Supplementary Table 4). Similarly, there are 6,314 interactions between 223 genes in hyperlipidemia network (Supplementary Table 5), the osteoarthritis network has 256 nodes with 7030 interactions between them (Supplementary Table 6) and in a cholelithiasis disease network there are 1049 interactions between 91 genes (Supplementary Table 7).

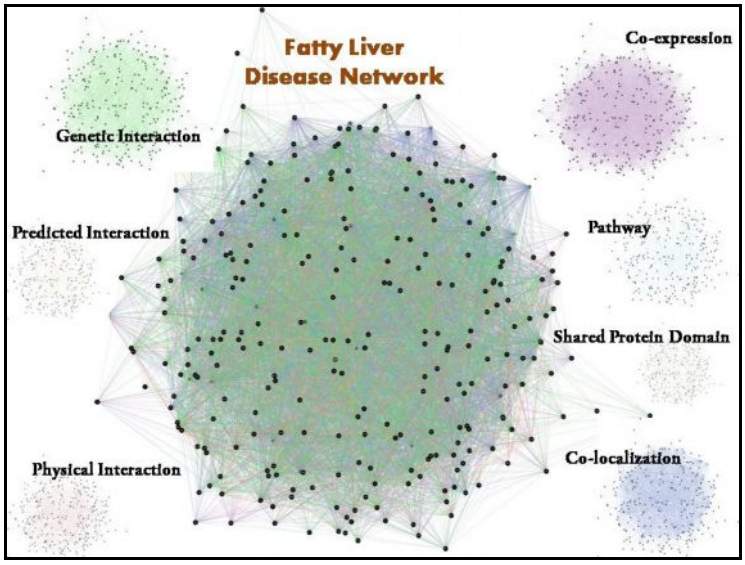

Figure 3 show the fatty liver disease network based upon various studies.

We identify *hubs* using hubs object analyzer software (Hubba) (18). A hub is a highly connected node of a network and may act as a target for a ligand/drug for therapeutic purpose. The top 10 hub molecules in all the networks are given in Table 4

Table 4 shows the highly connected nodes - "hubs" in the disease networks.

Fatty Liver	Poly Cystic Ovary disease	Hyperlipidemia	Osteoarthritis	Cholelithiasis
CPB2	JUN	SERPINA1	CXCL12	AFM
C5	PHOX2B	GC	SERPING1	GC
LIPC	PRL	C5	FN1	LIPC
GC	FMOD	FGA	BMP2	APOB
APCS	ESRRG	LIPC	CTSK	EGF
F9	JAG1	F9	VEGFC	GCG
PAH	EGFR	FGB	VCAM1	PLG
ARG1	F9	TTR	VCAN	AFP
AHSG	ESR1	F10	COL15A1	ADCYAP1
SERPINC1	VIM	PLG	CDH5	NPY

Comparison of drug targets from literature

We attempted to map the information on drugs and their targets on our networks. The drugs used in treatment of fatty liver includes: **gemfibrozil, atorvastatin and pravastin** (22). The targets associated with these drugs are cytochrome P450 2C8 (CYP2C8) for **gemfibrozil**, **and pravastin** and 3-hydroxy-3-methylglutaryl-coenzyme A reductase (HMGCR), dipeptidyl peptidase 4 (DPP4), aryl hydrocarbon receptor (AHR) for **atorvastatin**. The molecules associated with these targets were extracted and mapped to fatty liver disease network (Figure 4). We also mapped the side-effects associated with these drugs on to the target using SIDER 2 database (19). For instance, **gemfibrozil** is known to be associated with effects like gastrointestinal reaction, dyspepsia and diarrhea. The proteins targets predicted to be involved in diarrhea pathogenesis includes CYP2C8, CYP2C9 and PTGS1. These predictions can link causation of diarrhea by **gemfibrozil**, due to binding to its protein target CYP2C8.

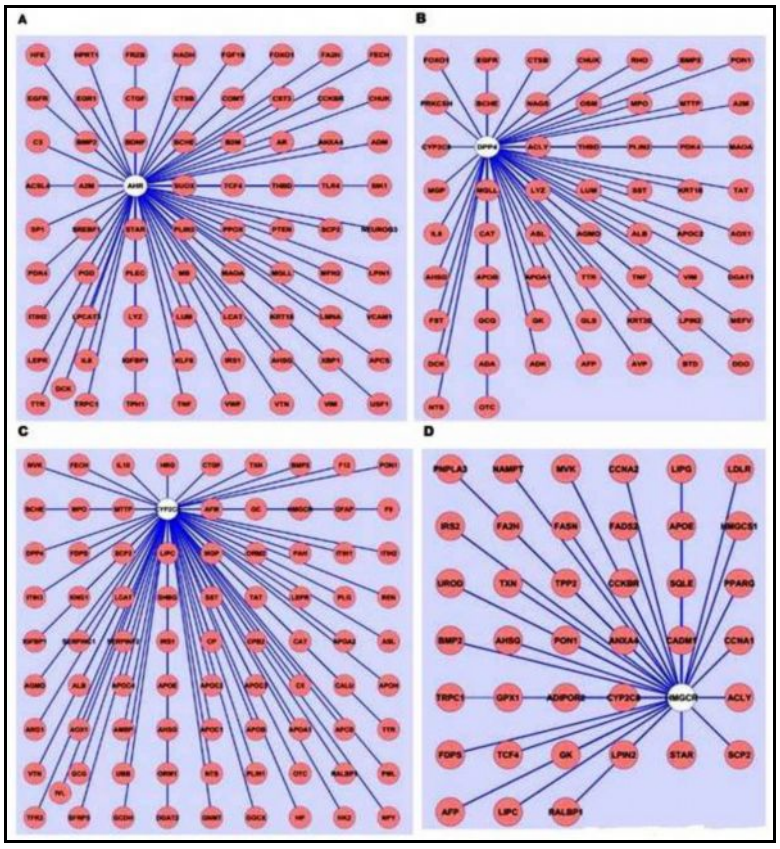

Figure 4 shows the sub-network of target signaling derived from fatty liver disease network: (A) Aryl hydrocarbon receptor- AHR (B) Dipeptidyl peptidase 4 - DPP4 (C) Cytochrome P450 2C8- CYP2C8 and (D) 3-hydroxy-3-methylglutaryl-coenzyme A reductase - HMGCR

Usage of metaformin leads to side effects such as hypoglycemia. 5'-AMP-activated protein kinase subunit beta-1 (PRKAB1), a known target for metformin, bind with a predicted protein DRD5 and involved in pathogenesis of hypoglycaemia. This interaction can explain the reason of hypoglycemia caused by metformin (23). U**rsodeoxycholic acid** binds to its target Cytochrome P450 2E1 (CYP2E1) and used for treatement of cholelithiasis (24). Due to the lack of detailed information on its target CYP2E1 involvment in cholelithiasis, no detailed information is inferred from disease network.

For hyperlipidemia, **clofibrate-** FDA approved antilipidemic agent for hyperlipidemia type III treatment exhibit its action as agonist on PPARA. PPARA is involved in hyperlipidemia network. For osteoarthritis, three drugs- **tenoxicam, piroxicam and oxaprozin** share common targets: Prostaglandin G/H synthase 1 (PTGS1) and Prostaglandin G/H synthase 2 (PTGS2). Due to lack of information in literature, we could not link the side effects with the molecular networks.

Functional analysis of disease network

It is essential to identify the functional role for a gene or a network. We assigned functional role to our networks using gene mania tool. This tool predicted functional roles for molecules listed in fatty liver network. For example, it predicted organic hydroxy compound metabolic process with a false discovery rate (FDR) of 1.10E-21, triglyceride metabolic process (3.22E-21), acylglycerol metabolic process (6.16E-21), neutral lipid metabolic process (6.16E-21), and lipid localization (1.07E-19) suggesting involvement of lipids and organic compounds metabolic pathways. This data indirectly support the role of such pathways in molecules implicated in fatty liver disease.

Similarly, in osteoarthritis network, we found involvement of molecules in inflammation, response to external stimuli and cell migration (See Supplementary Table 8).

Discussion

There are about 1733 diseases which affects humans (http://www.cdc.gov/diseasesconditions /az/a.html). Genes play a major role in pathogenesis of several of these diseases. Number of molecular and genetic studies of disease in last decades have produced an impressive list of gene–disease associations (12). Here, we used the concepts of network biology to integrate data from PubMed, SIDER, DrugBank and OMIM with information on side effects, gene expression and PPI. We extracted data from several databases such as MalaCards, IPKB, DisGeNET and our in-house literature-mining tool in perl to predict the genes association with a disease. We selected five diseases as a test case and used a computational approach developed inhouse (Jagannadham *et al* 2015, In Press). Different databases produce different set of genes for same clinical condition. For example, IPKB retrieves disease molecules based on knowledge of clinical and microarray studies. Our inbuilt literature mining approach in Perl screens the genes associated with a disease from PubMed abstracts.

Muhammed et al (2007) analyzed the relationships between drug targets and disease-gene products and observed a trend toward more rational drug design using network properties (25). With this lead, we attempted to map the existing targets for an FDA approved drug onto disease networks. Available drug targets were searched extensively in literature and mapped onto the disease networks. Next, we identified the cause/link associated with these drugs side-effects. We integrated information using networks to explain side effects such as diarrhea for gemfibrozil and hypoglycaemia with metformin. Using Gene Mania tool, a function for the particular group of genes or network was also predicted correctly. For example, in hyperlipidemia disease network, the majority of predictions were made for genes/protein molecules' role in lipid localization, regulation of plasma lipoprotein particle levels, protein-lipid complex, lipid homeostasis and transport. This allowed us to create an integrated framework based upon text mining tools, information gathered from publically available databases, side effects of drugs and network biology techniques to answer some of the fundamental questions on disease biology.

References

1. Schadt E.E., Molecular networks as sensors and drivers of common human diseases. Nature., 2009, 461, 218-223.
2. Heller M.J., DNA microarray technology: devices, systems, and applications., Annu Rev Biomed Eng, 2002, 4, 129-153.

3. Ansorge W.J., Next-generation DNA sequencing techniques, N Biotechnol, 2009, 25, 195-203.
4. Fields S. and Sternglanz R. The two-hybrid system: an assay for protein-protein interactions. Trends Genet.,1994, 10, 286-292.
5. Adie E.A., Adams R.R., Evans K.L., Porteous D.J. and Pickard B.S., Speeding disease gene discovery by sequence based candidate prioritization, BMC Bioinformatics, 2005, 6:55.
6. Suthram S., Dudley J.T., Chiang A.P., Chen R., Hastie T.J., Butte A.J., Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets, PLoS Comput Biol., 2010, 5,e1000662.
7. Köhler S, Bauer S, Horn D, Robinson PN. Walking the interactome for prioritization of candidate disease genes. Am J Hum Genet, 2008, 82, 949-958.
8. Wu X, Jiang R, Zhang MQ, Li S. Network-based global inference of human disease genes. Mol Syst Biol, 2008, 4, 189.
9. Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R. Associating genes and protein complexes with disease via network propagation. PLoS Comput Biol, 10, 6:e1000641.
10. Chen Y, Jiang T, Jiang R: Uncover disease genes by maximizing information flow in the phenome-interactome network, Bioinformatics, 2011, 27, i167-i176.
11. Hwang T, Kuang R. A heterogeneous label propagation algorithm for disease gene discovery. In: Proceeding of the SIAM International Conference on Data Mining, 2010, 583-594.
12. Goh, K.I. The human disease network. Proc. Natl. Acad. Sci. USA.,2007,104, 8685–8690
13. Barabási AL, Gulbahce N, Loscalzo J.Network medicine: a network-based approach to human disease. Nat Rev Genet., 2011,12, 56-68.
14. Rappaport N, Nativ N, Stelzer G, Twik M, Guan-Golan Y, Stein TI, Bahir I, Belinky F, Morrey CP, Safran M, Lancet D, . MalaCards: an integrated compendium for diseases and their annotation, Database, 2013
15. Viswanathan GA, Seto J, Patil S, Nudelman G, Sealfon SC, Getting Started in Biological Pathway Construction and Analysis,PLoS Comput Biol, 2008, 4, e16.
16. Bauer-Mehren A, Rautschka M, Sanz F, Furlong LI., DisGeNET: a Cytoscape plugin to visualize, integrate, search and analyze gene-disease networks., Bioinformatics, 2010, 26, 2924-2926
17. Warde-Farley D, Donaldson S. L., Comes O, Zuberi K, Badrawi R, Chao P, Franz M., Grouios C, Kazi F. Lopes C.T., Maitland A., Mostafavi. S., Montojo. J., Shao Q., Wright G., Bader G.D. and Morris. Q.The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function.Nucleic Acids Research, 2010,38, W214–W220.
18. . Lin CY, Chin CH, Wu HH, Chen SH, Ho CW, Ko MT. Hubba: hub objects analyzer--a framework of interactome hubs identification for network biology,Nucleic Acids Res., 2008,36,W438-443.
19. Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P.A side effect resource to capture phenotypic effects of drugs. Mol Syst Biol. 2010;6:343.
20. Bauer-Mehren A, Bundschus M, Rautschka M, Mayer MA, Sanz F, Furlong LI. Gene-disease network analysis reveals functional modules in mendelian, complex and environmental diseases. PLoS One, 2011,6:e20284.
21. Hirschhorn,J.N. and Daly,M.J., Genome-wide association studies for common diseases and complex traits, Nat. Rev. Genet., 2005, 6, 95–108.
22. Bayard M, Holt J, Boroughs E., Non alcoholic fatty liver disease.,Am Fam Physician., 2006,1,1961-1968.
23. Leeman L and Acharya U., The use of metformin in the management of polycystic ovary syndrome and associated anovulatory infertility: the current evidence.,J Obstet Gynaecol., 2009,29, 467-72.
24. Guarino MP, Cocca S, Altomare A, Emerenziani S, Cicala M., Ursodeoxycholic acid therapy in gallbladder disease, a story not yet completed.,World J Gastroenterol., 2013,21,5029-34.
25. Yildirim M A, Goh K, Cusick M E, Barabási A L and Vidal M., Drug—target network, Nature Biotechnology, 2007, 25, 1119–1126.

*****