Research Article

# Biomedical Text Mining of Obesity, Diabetes and Hypertension Genes

**Jaisri Jagannadham[1,2], Hitesh Kumar Jaiswal[1], Stuti Agrawal[1], Kamal Rawal[1*]**
[1]Department of Biotechnology, Jaypee Institute of Information Technology, Noida, Uttar Pradesh, India.
[2]School of Life Sciences, Jaipur National University, Jaipur, India.
***Corresponding author's E-mail:** kamalrawal.jnu@gmail.com

**Accepted on:** 23-06-2015; **Finalized on:** 31-07-2015.

**ABSTRACT**

The major goal of the study is to screen the diseases that are associated with obesity as well as the genes implicated in obesity and its related diseases. To facilitate the screening of thousands of literature data on obesity as well as its associated diseases, an in-house text mining system is developed in Perl. The abstracts of about 90,000 from 1960-2013 till September on obesity were screened by the text-mining system for about 1700 diseases. It is predicted that they are about 526 diseases that are associated or occurred with the term "*Obesity*". In which seven of the well-studied diseases such as Type II Diabetes Mellitus, Hypertension, Hyperlipidemia, Fatty Liver, Cholelithiasis, Osteoarthritis, Polycystic ovary disease including obesity are screened at the gene level in the abstracts obtained for the respective diseases. The screened genes implicated in each of the disease are compared and identified 26 genes including Leptin, Insulin etc, as crucial as they are common in all diseases. The gene ontology and signaling pathway analysis is done for these crucial genes. This work has led to the identification of diseases and genes in obesity and its associated disorders. Further, the list of diseases that are less-studied or less-frequently associated with obesity is predicted apart from the strongly related diseases. The signaling pathway analysis of the 26 crucial genes has led to identification that some of these are associated with cancer pathways. This provides an important link between cancer and these diseases which has to be investigated further.

**Keywords:** Obesity, genes, text-mining, diseases, algorithm

## INTRODUCTION

Obesity is a multi-factorial epidemic affecting millions across the world.[1] The imbalance between energy intake and expenditure is presumed to be the cause for the deposition of fat cells as energy storage. Abdominal obesity is defined as fat located in subcutaneous and visceral compartment around the abdominal region.[2] The visceral fat includes the fat deposited around omentum and mesentery in the abdominal cavity. Excess visceral fat is primarily associated with metabolic risk factors.[3] Increase in number of fat cells leads to several health problems which includes type 2 diabetes and hypertension. The subcutaneous fat includes gluteofemoral and truncal subcutaneous adipose tissue. Out of which, truncal fat is associated with metabolic risk factors. The frequently co-occurring diseases with obesity are collectively referred to 'metabolic syndrome X', 'insulin resistance syndrome', or 'Reaven syndrome' includes dyslipidemia, non insulin-dependent diabetes mellitus (NIDDM) and heart disease.[4-6]

Large numbers of research studies are being conducted in the obesity field. These studies include clinical, genetic, mutational and meta-studies. These studies contain a wealth of information which can be used to find an association of a variety of clinical conditions relevant to obesity. Further, we can also identify overlap of genes, proteins and pathways using such information. A major issue is high manpower cost to create and maintain such kind of resource. Use of the text-mining algorithm seems to offer a solution to this problem. It allows the experimenter to scan millions of published articles on a specific research area. These articles serve as a rich source of information covering a diverse range of research areas such as biochemistry, biology, pharmacology, etc. Text-mining has the potential to analyze results derived from different types of experiments such as microarray techniques[7], drug-drug interaction[8], and metabolomics[9] as well as to integrate results published in these articles. In this study, we find the correlation of obesity and other associated diseases through screening millions of published research records. Further, we also find candidate genes for obesity as well as associated clinical conditions to find overlap amongst them.

## MATERIALS AND METHODS

### Text-Mining Process

The rapid growth in the number and size of literature data needs efficient intensive search algorithms that can retrieve biological data depending on the study type. Text mining is applied in biomedical study for both hypothetical generation and biological discovery. Here initially the literature on obesity and its 7 associated diseases obtained from PubMed are categorised as "*Information retrieval system" (IRS)*.

From the IRS, the algorithm looks for the "*Entity recognition (ER)*" followed by "*Information extraction (IE)*". The step-wise process of the Text-mining algorithm is illustrated below:

***Compilation & Processing of Abstracts***

Abstracts having the term "obesity" and "human" were downloaded from PubMed using RefNavigator (version 2.0). RefNavigator is a tool, for retrieving the literature data form PubMed, Google Scholar, IEEE, ACM, Science Direct, Arxiv and Scirus. There are about 97,345 abstracts on obesity and human. These downloaded abstracts consisted of names of authors and their corresponding affiliations, journal of publication, year of publication and the abstract itself. The abstracts are downloaded as follows:

Open RefNavigator- Click PubMed in the menu bar Type "*Obesity*" and select "*human*" and check the box "*has abstracts*"- Click Search which will retrieve the documents- In the menu bar: click Import and Export- Export the references.

The compiled abstracts are processed using algorithms in perl.

First, they are filtered of from any redundancy. Followed by, aligning each abstract per line and removing any blank space. Further the abstracts are segregated into two parts that is the references are filtered separately from the main content and is given a specific identification number for each abstract and its corresponding reference.

***Compilation of disease/gene list***

(i) The list of diseases is obtained from centre of disease control and prevention website http://www.cdc.gov/diseasesconditions/az/a.html. (ii) The complete set of 35,959 genes in humans along with the symbol, approved name, previous names and synonyms was compiled from the HUGO database (www.genenames.org)

***Screening of diseases/genes using in-house text-mining system***

The approved disease/gene names for the complete set of disease/genes in humans were searched in the above set of abstracts using a Perl script. Here we describe the text-mining algorithm process;

Let X be the dictionary of disease/gene list to be searched in the abstracts Y. The X is considered as a vector with row representing gene and its synonym and column representing gene list. If any of gene/disease were found, the count for that particular disease/gene is incremented by unity.

Let X = Dictionary Set;

Let Y=Abstracts;

Let Z=disease/gene output;

Let count= 0;

Let a is a disease/gene in X

for a =1 to n do

if a ϵ *Y* then

count= a++ ;

write a & count to Z;

Next;

**Gene Ontology & Signalling Analysis**

The gene ontology (GO) analysis for the common genes in obesity and its related diseases is done on the basis of functional categories such as molecular function, biological process and cellular component using Network Ontology Analysis (NOA).[10] To know the signaling associated with these common molecules, these genes are searched against the gene sets involved in signaling pathways such as KEGG, Biocarta, Reactome, other canonical pathways using Molecular Signature Database (MSigDB).[11] MSigDB computes the significant value using the probability of a match between a test and a database gene set.

## RESULTS AND DISCUSSION

**Mining association of obesity with other diseases from literature**

We extracted 97,345 abstracts published from 1960 till September 2013 on obesity in human using RefNavigator. We prepared a list of 1733 disease terms (http://www.cdc.gov/diseasesconditions/az/a.html). These abstracts were searched for the occurrences of 1733 disease terms. We found 527 disease terms occurred in abstracts relevant to obesity. Most frequent terms (diseases) associated with obesity include insulin resistance (found in 47363 abstracts), hypertension (13,293), atherosclerosis (3224), coronary heart disease (2232), fatty liver (2142), hyperlipidemia (1659), breast cancer (1520), hypertriglyceridemia (1141), prader-willi syndrome (1093), prostate cancer (590), colorectal cancer(487), colon cancer (327), diabetic nephropathy (241), and lipodystrophy (220). Several clinical conditions which include apnea (2005 abstracts), stroke (1927), asthma (1251), cirrhosis (810), and osteoporosis (806) were also discovered. Several genetic syndromes that have a strong association with obesity include Prader-willi syndrome, Alstrom syndrome (75), Barder-Biedl syndrome (199), Cohen syndrome (32), and Simposn-Golabu-Behmel syndrome (17) (Supplementary Table 1: http://tinyurl.com/qxgb5co).

**Mining molecules relevant to obesity and other diseases**

Some of the obesity associated diseases such as type 2 diabetes mellitus, hypertension, fatty liver, polycystic ovarian disease (PCOD), hyperlipidemia, cholelithiasis, osteoarthritis are analyzed for their association at a molecular level. With our text-mining process, the molecules implicated in each of the diseases are screened from PubMed abstracts using the list of about 35,000 genes in human from HUGO database (www.genenames.org). The number of molecules implicated in obesity and its associated disease screened by our system is given in Table 1. The detailed molecule

list on all diseases is provided in Supplementary Table 2: http://tinyurl.com/oygh6uw.

**Table 1:** Shows the number of molecules screened by our text mining process from PubMed abstracts on obesity and its related diseases.

| Diseases associated with Obesity | Implicated genes |
|---|---|
| Hypertension | 495 |
| Fatty Liver | 274 |
| Diabetes Mellitus, Type II | 494 |
| Polycystic Ovary Syndrome | 319 |
| Hyperlipidemia | 207 |
| Osteoarthritis | 239 |
| Cholelithiasis | 75 |
| Obesity | 2809 |

After the screening of molecules from abstracts, the common genes between obesity and the related diseases are filtered implementing an identity algorithm for pairwise comparison in Perl. Let *O* be the obesity gene list, and let *OR* be the obesity-related disease gene list. Let *i* denote a gene in *O* to be searched in *OR*, the presence of *i* in both set is possible if and only if $i \epsilon O, OR$. From the results it is predicted that the maximum overlap of molecules is between type II diabetes mellitus and obesity showing 345 molecules in common. Some of the well-known molecules are insulin, insulin receptor, adiponectin, leptin, lipoprotein lipase and so on. Similarly is with hypertension and obesity showing a common genes of 345 (Figure 1). The next highest common genes with Obesity are of 292 with PCOD. The common genes between Obesity and Fatty liver are 209, Obesity and Hyperlipidemia are 181, Obesity and Osteoarthritis are 168 and Obesity and Cholelithiasis are 59.

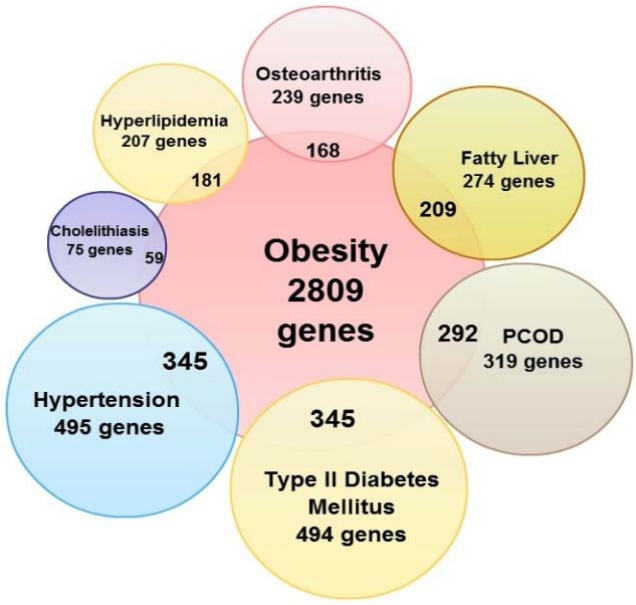


**Figure 1:** Obesity and its associated diseases with the molecules screened and the number of overlap molecules with obesity.

To know the crucial molecules that are common to all that is obesity and its 7 other related diseases, the above identity algorithm is used for screening the gene *i* in multiple disease gene set. Such mining resulted in a set of 26 genes that are common in all the above diseases. Some of the molecules are leptin and its receptor (LEP), glucagon (GCG), peroxisome proliferator-activated receptor gamma (PPARG), tumor necrosis factor (TNF) etc (See Table 2).

**Gene Ontology & Signaling of the crucial molecules in obesity and its 7 associated diseases**

The gene ontology (GO) describes the functional association of the molecules. In biology, GO has three functional categories: biological process, cellular component and molecular function. The Network Ontology Analysis (NOA) is used for the analysis of the GO associated with the molecules that are common in all obesity and its 7 related diseases. The detailed results on the functional activity at all the 3 terms along with the p-value are given in Supplementary Table 3: http://tinyurl.com/ofcsena. The biological process associated with these molecules includes regulation of biological quality, regulation of signal transduction, signaling process and cell communication which is further categorized with positive regulation in all the above processes. On the cellular component basis these molecules are associated with extracellular region, protein-lipid complex such as very-low-density lipoprotein and triglyceride rich lipoprotein particle. At molecular function, they are associated with DNA regulatory region binding such as promoter and protein binding such as enzyme, hormone such as NADP or NADPH binding, receptor such as insulin-like growth factor receptor binding, parathyroid hormone receptor binding. The Figure 2 explains the activity and the number of these molecules at all three GO terms.

Here the molecules those are common in all obesity related diseases are predicted for their signaling in databases such as KEGG, Reactome, BioCarta using Molecular Signature Database (MSigDB). There are about 1320 gene sets in collection for the above mentioned signaling databases. These 26 genes are searched against these gene sets to know their associated signaling pathways. It is predicted that these 26 genes are involved in 96 signaling events.

The corresponding overlap significance is also predicted through p-value. The maximum overlap is with 6 genes in signaling such as genes involved in developmental biology (396) (the number indicate the number of genes involved in this particular signaling), cytokine-cytokine receptor interaction (267), signaling events mediated by protein tyrosine phosphatase (PTP1B) (52).

**Table 2:** Shows the list of 26 genes (approved name, HUGO ID, chromosome location) common in Obesity, Hypertension, Hyperlipidemia, Osteoarthritis, Cholelithiasis, Type 2 DM, Fatty liver, and PCOD.

| Gene Symbol | Gene Name | HGNC ID | Location |
|---|---|---|---|
| LEP | leptin | HGNC:6553 | 7q31 |
| TF | transferrin | HGNC:11740 | 3q21 |
| HP | haptoglobin | HGNC:5141 | 16q22.2 |
| TNF | Tumor necrosis factor | HGNC:11892 | 6p21.3 |
| EGF | Epidermal growth factor | HGNC:3229 | 4q25 |
| CCND1 | Cyclin D1 | HGNC:1582 | 11q13 |
| SST | somatostatin | HGNC:11329 | 3q28 |
| GSR | glutathione reductase | HGNC:4623 | 8p21.1 |
| INS | Insulin | HGNC:6081 | 11p15.5 |
| EGFR | Epidermal growth factor receptor | HGNC:3236 | 7p12 |
| G6PD | Glucose-6-phosphate dehydrogenase | HGNC:4057 | Xq28 |
| PTH | Parathyroid hormone | HGNC:9606 | 11p15.3-p15. 1 |
| LPL | Lipoprotein lipase | HGNC:6677 | 8p22 |
| CAT | catalase | HGNC:1516 | 11p13 |
| REN | renin | HGNC:9958 | 1q32 |
| PRL | Prolactin | HGNC:9445 | 6p22.3 |
| MS | Multiple sclerosis | HGNC:7314 | |
| PPARG | Peroxisome proliferator-activated receptor gamma | HGNC:9236 | 3p25 |
| LEPR | leptin receptor | HGNC:6554 | 1p31 |
| VIM | vimentin | HGNC:12692 | 10p13 |
| CCK | cholecystokinin | HGNC:1569 | 3p22.1 |
| HFE | hemochromatosis | HGNC:4886 | 6p21.3 |
| DIANPH | Diabetic nephropathy | HGNC:2875 | 18q22.3-q23 |
| GCG | glucagon | HGNC:4191 | 2q36-q37 |
| APOA1 | Apolipoprotein A-I | HGNC:600 | 11q23-q24 |
| AR | Androgen receptor | HGNC:644 | Xq12 |

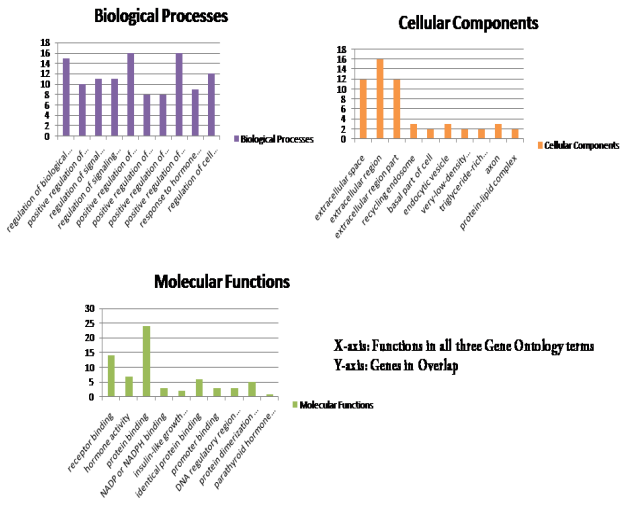


**Figure 2:** Describes the gene ontology prediction in all three terms for the common molecules in obesity and its 7 associated diseases.

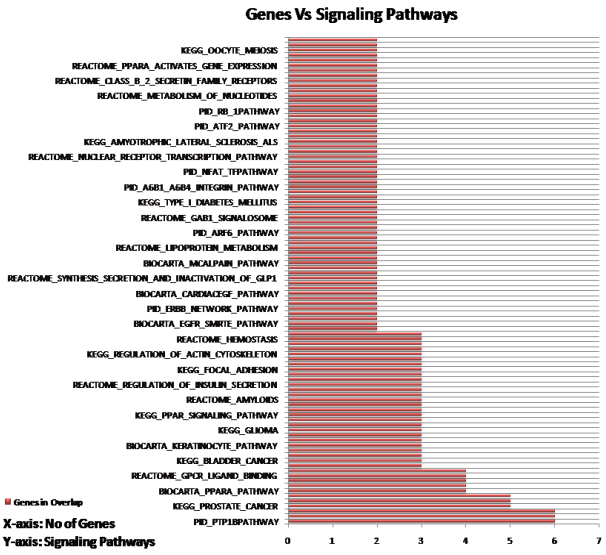


**Figure 3:** Shows the gene overlap in signaling pathways such as KEGG, Biocarta, Reactome.

The 5 of these molecules are involved in pathways of cancer (328) and prostate cancer (89). They are androgen receptor (AR), cyclin D1 (CCND1), epidermal growth factor and its receptor (EGF, EGFR), peroxisome proliferator activated receptor gamma (PPARG) and insulin (INS). These genes overlap distribution with the signaling pathways are provided in Figure 3. The least overlap is of 2 genes in 58 signaling pathways provided in Supplementary Table 4: http://tinyurl.com/pp24wbr.

## CONCLUSION

This work enhances our understanding of obesity and its associated diseases/disorder through our in-house text-mining system in Perl. Our text-mining system process help in searching the huge set of literature data for various diseases and genes. The list of about 527 diseases that may be associated with obesity with the frequency of occurrences in the obesity abstracts is identified by our method. From this finding the diseases that are strongly as well as weekly associated with obesity including the syndromes is predicted. Further the molecular screening is carried out with the text-mining system in obesity and 7 of its associated diseases such as type II diabetes mellitus, hypertension, hyperlipidemia, cholelithiasis, fatty liver, polycystic ovary syndrome, osteoarthritis. The pairwise overlap of the molecules in obesity and 7 of its diseases resulted in the maximum overlap of obesity genes with hypertension and type II diabetes mellitus. This explains that our study is explaining an in-depth understanding of obesity with its associated diseases at molecular level. The list of 26 genes such as leptin and its receptor, insulin, glucagon etc are identified as common in all obesity and its 7 associated diseases through our identity algorithm in Perl. The study of these common molecules on gene ontology terms has given a detailed depiction on the functional annotation at molecular, cellular and biological processes. The signaling level understanding of these common molecules has resulted in some important links. It is observed that insulin, peroxisome proliferator activator receptor gamma, cyclin D 1, androgen receptor epidermal growth factor and its receptor are among the common molecules that are involved in pathways associated with cancer. Further from the screening of diseases related to obesity, it is evident that obesity may be the cause for cancer involving 20 different types such as prostrate cancer, breast cancer etc. This provides a key evidence for the association of cancer with obesity and its related diseases but in-depth further investigation is needed.

**Acknowledgement:** We thank Jaypee Institute of Information Technology and Jaipur National University for their constant support and encouragement in carrying out the research. We are thankful to Indian Institute of Technology, New Delhi for providing the access to their super-computing facility for the execution of our programs.

## REFERENCES

1. Popkin BM, Drewnowksi A, Dietary fats and the nutrition transition: new trends in the global diet, Nutr Rev, 55, 1997, 31-43.
2. GRUNDY SM, Obesity, Metabolic Syndrome, and Cardiovascular Disease. The Journal of Clinical Endocrinology & Metabolism, 89, 2004, 2595–2600.
3. Bosello O, Zamboni M, Visceral obesity and metabolic syndrome, Obes Rev, 1, 200, 47-56.
4. Petrie JR, Cleland SJ, Small M, The metabolic syndrome: overeating, inactivity, poor compliance or 'dud'advice? Diabet. Med., 15, 1998, 529–53.
5. Reaven GM, Role of insulin resistance in human disease (syndrome X): an expanded definition, Annu. Rev. Med., 44, 1993, 121–131.
6. Reaven GM, Pathophysiology of insulin resistance in human disease, Physiol. Rev., 75, 1995, 473–486.
7. Fleuren WW, Verhoeven S, Frijters R, Heupers B, Polman J, van Schaik R, de Vlieg J, Alkema W, CoPub update: CoPub 5.0 a text mining system to answer biological questions, Nucleic Acids Res, 39, 2011, W450-W454.
8. Tari L, Anwar S, Liang S, Cai J, Baral C, Discovering drug-drug interactions: a text-mining and reasoning approach based on properties of drug metabolism, Bioinformatics, 26, 2010, 547-553.
9. Nobata C, Dobson PD, Iqbal SA, Mendes P, Tsujii J, Kell DB, Ananiadou S, Mining metabolites: extracting the yeast metabolome from the literature, Metabolomics, 7, 2011, 94-101.
10. Wang J, Huang Q, Liu ZP, Wang Y, Wu LY, Chen L, Zhang XS, NOA: a novel Network Ontology Analysis method, Nucleic Acids Res., 39, 2011, e87.
11. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, Proc Natl Acad Sci U S A, 102, 2005, 15545-15550.

**Source of Support:** Nil, **Conflict of Interest:** None.