International Letters of Natural Sciences *Online: 2015-07-31*
ISSN: 2300-9675, Vol. 44, pp 1-9
doi:10.56431/p-u2hm84

Text Mining Approach to analyse the relation between obesity and breast cancer data

Dr. Ashok Kumar[1,a], Priyanka Thakur[1,b], Kanika Gupta[1,c] and Dr. Amit Pal[2,d]*

Centre for Systems Biology and Bioinformatics, Panjab University, Chandigarh -160014, India[1]
Deptt. of Biochemistry, PGIMER, Chandigarh-160012, India[2]

E-mail address: [a]ashokbiotech@gmail.com[a];
[b]ptpriyanka010@gmail.com[b]
[c]guptakainka18@gmail.com[c]
[d]maximus1134@gmail.com*; [d]apal1134@gmail.com*
Corresponding Author: Dr. Amit Pal[d]*

Keywords: BMI, Text mining, Obesity, Breast cancer

ABSTRACT. Biomedical research needs to leverage and exploit large amount of information reported in scientific publication. Literature data collected from publications has to be managed to extract information, transforms into an understandable structure using text mining approaches. Text mining refers to the process of deriving high-quality information from text by finding relationships between entities which do not show direct associations. Therefore, as an example of this approach, we present the link between two diseases i.e. breast cancer and obesity.Obesity is known to be associated with cancer mortality, but little is known about the link between lifetime changes in BMI of obese person and cancer mortality in both males and females. In this article, literature data for obesity and breast cancer was obtained using PubMed database and then methodologies which employs groups of common genes and keywords with their frequency of occurrence in the data were used, aimed to establish relation between obesity and breast cancer visualized using Pi-charts and bar graphs. From the data analysis, we obtained 1 gene which showed the link between both the diseases and validated using statistical analysis and disease-connect web server. We also proposed 8 common higher frequency keywords which could be used for indexing while searching the literature for obesity and breast cancer in combination.

1. INTRODUCTION

The scientific literature provides a wealth of information to researchers. It may serve as a source of information that can be used for building research hypotheses that subsequently can be experimentally validated. This knowledgebase may serve as a source for interpretation of experimental results [1]. Current biomedical research needs to leverage and exploit the enormous amount of information reported in scientific publications. One of the most important entry points to scientific literature sources for biomedical research is PubMed. Now-a-days, text is the most common vehicle for the formal exchange of information. Text mining refers to the process of deriving high-quality information from text by finding relationships between entities which do not show direct associations. This approach uses the automated methods for exploiting the enormous amount of knowledge available in text documents [2]. It is used to identify different relations like gene-disease, drug-disease [3] and drug-target associations etc. [4]. Large-scale extraction and analysis of gene-disease associations, and integration with current biomedical knowledge, provided insights of information, found in the literature, and raised challenges regarding data prioritization and curation [5]. We will be using text data from obesity and breast cancer for this current project. Obesity defined as it is an abnormal accumulation of body fat. Obesity is major risk factors for a number of chronic diseases, cardiovascular diseases and cancer. Breast cancer is a kind of cancer that develops from breast cells. Obesity is known to be associated with cancer mortality, but little is known about the link between lifetime changes in BMI of obese person and cancer mortality in both males and females. Numerous epidemiological studies have reported a possible differential impact of BMI on breast cancer risk in women of various life stages[6].Conversely, there is statistically

significant positive association between body weight and breast cancer risk among postmenopausal women[7].Our approach, particular aimed to establish relationships between entities. Software's have been used to execute relation between obesity and breast cancer on the basis of co-occurrence of common keywords and genes. We have taken obesity (group I), breast cancer (group II), obesity and breast cancer (group III) as our query. Literature data will be obtained from PubMed database. Keywords and genes with their frequency of occurrence will be extracted from the data using RapidMiner [8], Coremine and Pubmed.mineR [9]. Keywords and genes corresponds to higher frequency of occurrence will be selected and visualized in WordCloud [13]. Relative frequencies of co-occurrence of selected keywords will be obtained using PubMatrix [10].Gene Ontology of selected genes will be analyzed using Gene Ontology Consortium [11]. In the analysis of selected data for three groups, common genes and keywords will be selected. Co-occurrence of both the genes and keywords would be analyzed using PubMatrix. Validation of the final result will be done using Disease-connect [12].

2. METHODOLOGY

Dataselection. Literature data is obtained using PubMed database for 3 main groups that is group I as Obesity, group II as Breast Cancer, and group III as obesity and breast cancer in combination. Data was obtained by applying filter of text availability as 'Abstracts' and publication date as '5years'.

Mining of data. Three softwares are used to extract keywords and genes from the data. Keywords are extracted using RapidMiner and Coremine. RapidMiner is stand-alone software which is used to retrieve keywords with their frequency of occurrence in the literature data for all the three groups. Coremine is a web based search engine. Keywords are extracted from disease literature. Coremine gave keywords with their frequency of occurrence in NCBI for both the queries obesity and breast cancer but not for group III. From literature data, genes are extracted using Pubmed.mineR. It uses R package. List of number of genes with their frequency of occurrence in the data is obtained

Data analysis. Data obtained from 3 softwares then analyzed using spreadsheets in Ms-Excel and visualized in the form of Pi charts and Bar graphs. Highly frequent keywords and genes then selected from the data for further procedure. WordCloud: The WordCloud is a Cytoscape plugin generates a visual summary of these annotations by displaying them as a tag cloud, where more frequent words are displayed using a larger font size (5). Highly frequent keywords are visualized using WordCloud which shows the words connected with their frequency of occurrence in the data.PubMatrix: PubMatrix is a web-based tool that allows simple text based mining of the NCBI literature search service PubMed using any two lists of keywords terms, resulting in a frequency matrix of term co-occurrence (2). It is used to relate the keywords of 1 group with another. Co-occurrence of keywords in two groups is obtained.

From the results of PubMatrix 8 common highly frequent keywords then selected for further procedure.

Data enrichment. Data enrichment of selected genes with high frequency then done by using Gene Ontology Consortium. By comparing the gene ontologies of selected genes of all the three groups, 3 common genes then selected.

Co-occurrence of keywords and genes. PubMatrix is used to obtain co-occurrence between the selected 3 genes and 8 keywords. It is used to relate genes with keywords and their frequency of occurrence together

Validation. Final results obtained then validated using Disease-Connect web server which is the first public web server integrates comprehensive-omics and literature data, including a large amount of gene expression data, GenomeWide Association Studies catalog, and text-mined knowledge, to discover disease–disease connectivity via common molecular mechanisms (3).

3. RESULTS AND DISCUSSION

Literature data from pubmed. Literature data is retrieved using PubMed database for group I(obesity),group II(breast cancer),group III(obesity and breast cancer) with applying filter of Text Availability as 'Abstracts' and Publication dates as '5 years'. Total 76776, 76487, 1212 results were obtained for group I, II and III, respectively.

Rapidminer to extract keywords. RapidMiner converts the literature data into the list of keywords corresponds to their frequency of occurrence in the data. Results obtained as attribute name,total occurrences, document occurrences and occurrence of words. List of total 43763, 107913, 33206 keywords were obtained for groups I, II and III respectively.

Coremine to extract keywords. Coremine extract keywords from medical literature. Coremine was used to decrease the redundancy of results of RapidMiner and to avoid the biasness. Total 9 and 11 keywords were obtained with their frequency of occurrence for group I and II respectively.

From the results of RapidMiner and Coremine 26,25and 8 keywords were selected on the basis of their higher frequency for group I, group II and Group III data, respectively.

Table1. Result of three softwares for all the groups

Groups	PubMed	RapidMiner	Coremine
Group I	76776 papers	43763 keywords	9 keywords
Group II	76487 papers	107913 keywords	11 keywords
Group III	1212 papers	33206 keywords	No result

Data visualization using WordCloud. Selected data for all the three groups then visualized using WordCloud plugin of Cytoscape. Figure 1, 2 and 3 shows network view, preferred layout view and in dock window, respectively.

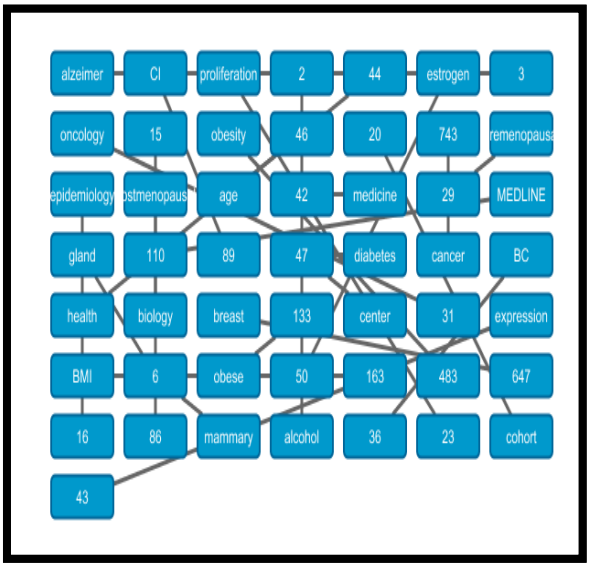

Figure 1: Network view

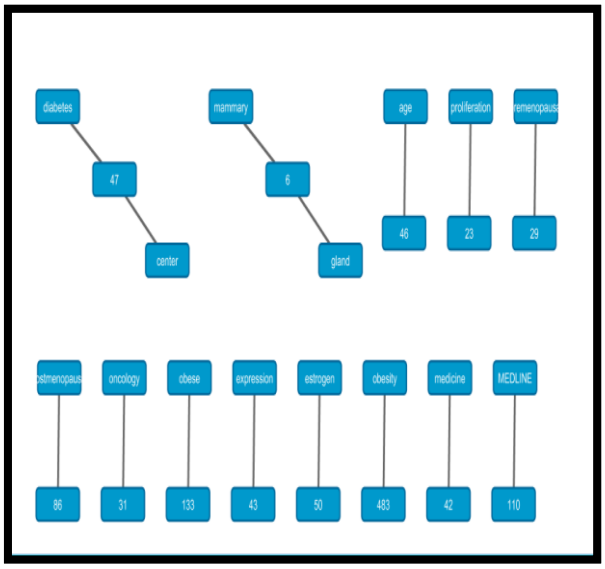

Figure 2: Preferred layout

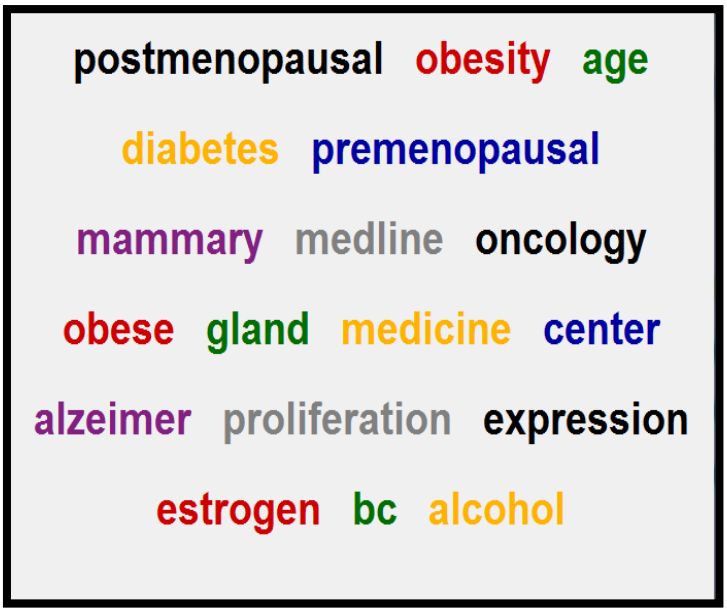

Figure 3: Visualization of selected keywords in dock window

Co-occurrence of keywords using pubmatrix. Selected keywords from RapidMiner and Coremine, considered as input for PubMatrix which analyze the co-occurrence of keywords in the form of matrix. Matrix is formed by pairwise comparison of each SEARCH TERM against each MODIFIER TERM. Frequency number in the table demonstrates the co-occurrence of both the keywords together.

Table 2: Matrix table shows frequency of co-occurrence of keywords of group I data with group III data

Breast cancer	Obesity and breast cancer							
	Age	BMI	BRCA	Estrogen	Post menopausal	Adult	Proliferation	Prognostic
Alcohol	70760	4556	28	6918	212	207622	10231	2290
BC	5035	292	83	987	61	11261	1367	960
BMI	49827	85896	24	1768	392	58490	446	1224
Biology	49194	1571	386	16810	253	104133	47365	8064
Breast	54055	2520	1934	46307	1768	142981	23752	18026
CI	162882	13437	240	19150	840	402730	12847	17078
Cancer	274134	9030	2476	71026	2626	1215148	166242	116523
Center	263481	15637	706	24474	904	657064	63541	33401
Cohort	148772	12898	265	5242	462	239601	2661	17686
Epidemiology	507857	29677	513	15890	1347	936656	8373	64412
Gland	25181	386	20	12533	134	87313	11078	2098
Health	457089	33092	561	24402	1814	949263	29757	18542
MEDLINE	8385	445	29	820	83	13649	321	1352
Mammary	57090	2552	1937	49494	1778	149507	26616	18245
Medicine	413007	26643	593	37362	1679	1131726	99657	46567
Obesity	58126	38366	13	3470	476	96256	2209	1291
Estrogen	29278	1768	229	224654	2925	72015	14169	6399
Expression	78282	2566	441	42629	526	257065	172203	37972
Obese	63683	41690	14	3865	523	104400	2619	1460
Oncology	39569	1199	625	7867	324	118503	18813	25548
Postmenopausal	15670	2162	32	20200	2074	33833	1166	800
Premenopausal	6149	1059	36	4784	704	11479	362	518
Proliferation	18191	446	106	14169	229	68513	406446	8278
Age	1883725	49827	615	29278	2685	1116442	18191	47808
Diabetes	110707	20120	9	3833	359	217891	8190	4899

Table 3: Matrix table shows frequency of co-occurrence of keywords of group II data with group III data

Obesity	Breast cancer and obesity							
	Age	BMI	BRCA	Estrogen	Post menopausal	Adult	Proliferation	Prognostic
AR	13179	827	15	3556	96	29011	7264	853
Abdominal	47590	5764	30	2252	269	176376	3530	3852
Active	58938	2907	52	9104	307	165774	25586	4088
Adult	1116442	58490	1081	72015	5206	6047184	68513	112164
Age	1883725	49827	615	29278	2685	1116442	18191	47808
Alcohol	70760	4556	28	6918	212	207622	10231	2290
Alzeimer	0	0	0	0	0	0	0	0
Appetite	3717	1337	1	296	17	10604	198	151
Asthma	21206	887	0	316	16	53976	2116	389
BC	5035	292	83	987	61	11261	1367	960
BMI	49827	85896	24	1768	392	58490	446	1224
BRCA	615	24	2590	229	10	1081	106	93
Baseline	90612	10867	50	6127	699	222804	2826	8304
Biological	135112	7544	571	36263	922	347902	98893	38993
Blood	401828	33848	184	58599	2309	1315430	102927	43961
Brain	152043	2130	23	22210	270	440339	26222	12290
Breast	54055	2520	1934	46307	1768	142981	23752	18026
Cancer	274134	9030	2476	71026	2626	1215148	166242	116523
CI	162882	13437	240	19150	840	402730	12847	17078
Cohort	148772	12898	265	5242	462	239601	2661	17686
Agriculture	10116	296	1	1377	14	11587	1993	46
Obesity	58126	38366	13	3470	476	96256	2209	1291
Oncology	39569	1199	625	7867	324	118503	18813	25548
Estrogen	29278	1768	229	224654	2925	72015	14169	6399
Postmenopausal	15670	2162	32	20200	2074	33833	1166	800
Diabetes	110707	20120	9	3833	359	217891	8190	4899
Women	277764	25806	1014	48734	5798	487368	5798	11007

By comparing the data of group I, group II and group III using outputs of RapidMiner, Coremine and PubMtarix, 8 common keywords were extracted. These 8 common were selected on the basis of their higher frequency of occurrence and co-occurrence in the data.

Pubmed.miner to extract genes. Pubmed.mineR extract genes form the literature data. Pubmed.mineR gave result as total number of genes with their frequency of occurrence in the literature data. Total 979, 4449, 1849 genes with their frequency of occurrence is obtained for group I,II and III.

Data enrichment using gene ontology consortium. Gene ontology for selected genes of all the three data was obtained.

SR. NO	Gene ID	Mapped ID's	Gene name	Gene symbol	PANTHER protein class	Species
1	HGNC=9582	PTEN	Dual-specificity protein phosphatase PTEN	PTEN	protein phosphatase	Homo Sapiens
2	HGNC=1100	BRCA1	Breast cancer type 1 susceptibility protein	BRCA1	ubiquitin protein ligase	Homo Sapiens
3	HGNC=11089	SLN	Sarcolipin	SLN	-	Homo Sapiens
4	HGNC=3236	EGFR	Epidermal growth factor receptor	EGFR	-	Homo Sapiens
5	HGNC=11515	T	Brachyury protein	T	transcription factor nucleic acid binding	Homo Sapiens
6	HGNC=1101	BRCA2	Breast cancer type 2 susceptibility protein	BRCA2	damaged DNA binding protein	Homo Sapiens
7	HGNC=644	AR	Androgen receptor	AR	nuclear hormone receptor, nucleic acid binding	Homo Sapiens
8	HGNC=5172	HR	protein hairless	HR	Zinc finger transcription factor , nucleic acid binding	Homo Sapiens

Figure 3: Gene Ontology of selected genes of higher frequency from group I data

SR. NO	Gene ID	Mapped ID's	Gene Name	Gene Symbol	PANTHER Protein Class	Species
1	HGNC=1424	CAD	CAD protein	CAD	Transferase , ligase	Homo Sapiens
2	HGNC=3678	FGF21	Fibroblast growth factor 21	FGF21	Growth factor	Homo Sapiens
3	HGNC=24678	F10	Alpha-ketoglutarate dependent dioxygenase FTO	F10	-	Homo Sapiens
4	HGNC=11764	TG	Thyroglobulin	TG	Esterase , lipase	Homo Sapiens
5	HGNC=11316	SSB	Lupus la protein	SSB	Ribonucleoprotein	Homo Sapiens
6	HGNC=11515	T	Brachyury protein	T	Transcription factor, nucleic acid binding	Homo Sapiens
7	HGNC=2367	CRP	C-reactive protein	CRP	Antibacterial response protein	Homo Sapiens
8	HGNC=6932	MC4R	Melanocortin receptor 4	MC4R	G-protein coupled receptor	Homo Sapiens
9	HGNC=644	AR	Androgen receptor	AR	Nuclear hormone receptor , nucleic acid binding	Homo Sapiens
10	HGNC=5172	HR	Protein hairless	HR	Zinc finger transcription factor , nucleic acid binding	Homo Sapiens

Figure 4: Gene Ontology of selected genes of higher frequency from group II data

Correlation between genes and keywords using pubmtarix. 3 common genes and 8 common keywords were selected from previous results. Correlation between genes and keywords was checked using Pubmatrix.

Table 4: Matrix table showing co-occurrence of genes with keywords

Keywords	Genes		
	HR	AR	T
Age	29649	13183	1
BMI	2275	827	0
Estrogen	3375	3556	0
Postmenopausal	1046	997	0
Proliferation	4498	7266	0
Cancer	32963	15289	0
Diabetes	6981	2728	0
Breast	5256	2045	0

Table 4 shows the probability of occurrence of keywords and genes together. Genes AR and HR show higher frequency of co-occurrence with all the keywords but Gene T shows its frequency of co-occurrence only with the keyword Age. This shows that genes AR and HR relate the diseases, obesity and breast cancer with each other.

Validation using disease-connect web server.

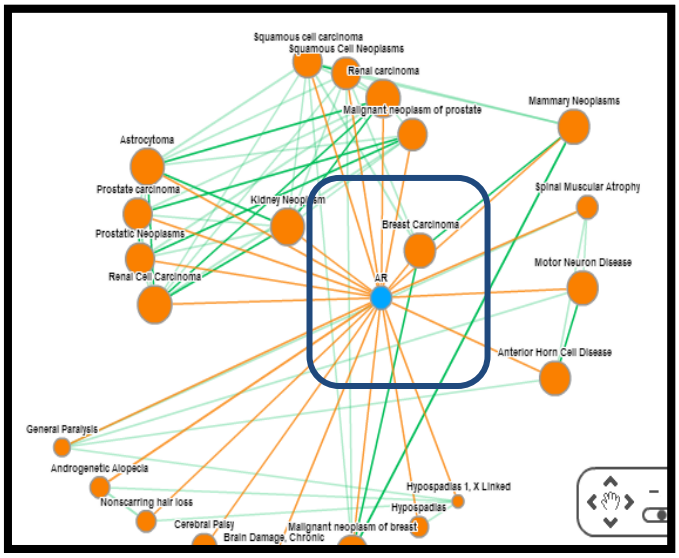

Figure 5: Network of Disease associated with Gene AR

Gene AR when searched for associated disease network, then Breast Carcinoma was associated with the gene. Network shows the diseases associated with **AR** based on the GWAS/OMIM/DEG records.

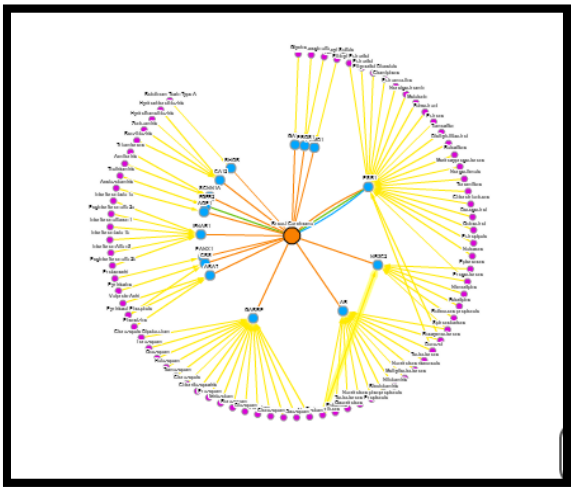

Figure 6: Disease-gene-drug network

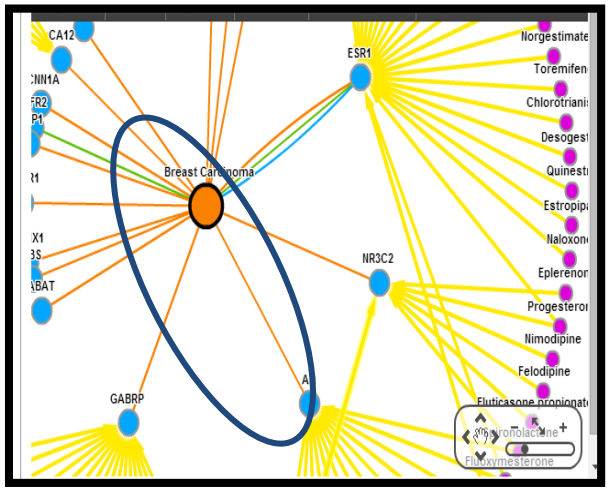

Figure 7: Disease-gene network

Figure 6 shows the disease-gene-drug network contains genes related to Breast Carcinoma based on the GWAS/OMIM/DEG records. Figure 7 shows the presence of AR gene in the disease gene network of breast Carcinoma.

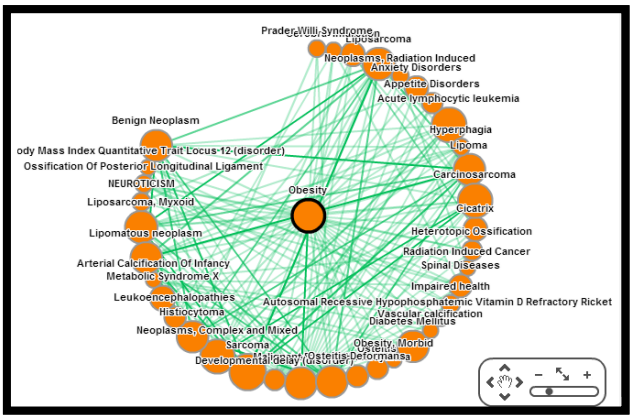

Figure 8: Network shows the disease

AQP9	GeneRIF	18401671 19615702 22425521
AR	GeneRIF	18805913 20083725 21940984
ARNTL	GeneRIF	23003921
ARRDC3	GeneRIF	21982743

Figure 9: Genes Associated with Obesity

Figure 8 shows disease associated with obesity. Figure 9 shows the Genes that are relevant to Obesity based on the GeneRIF (Gene Reference into Function) and GeneWays records. GeneWays provides disease-gene relations extracted from full-text articles and abstracts in PubMed.

4. CONCLUSION

Obesity has differential effect on breast cancer risk in women of various life stages. Text mining approach was used to reveal the relationship between both the diseases on the basis of co-occurrence of molecular markers and keywords. On the basis of text mining procedure in present study it is concluded that AR gene could be potentially linked between obesity and breast cancer. The frequency of selective keywords demonstrates the linking between both the diseases and these words could be used for indexing while searching the literature for obesity and breast cancer in combination.

Acknowledgement

Authors are thankful to Vice Chancellor of Panjab University, Chandigarh and coordinator of Centre for Systems biology and bioinformatics for providing all the facilities to carry out this work. We would also like to acknowledge the support provided by online tools, servers, databases and softwares for successful accomplishment of our work that is PubMed Database (http://www.ncbi.nlm.nih.gov/pubmed), Rapid Miner (https://rapidminer.com/products/studio/), Coremine (http://www.coremine.com/medical/), Pubmed.mineR (https://cran.r-project.org/web/packages/pubmed.mineR/index.html), WordCloud (http://baderlab.org/Software/WordCloudPlugin), Gene Ontology Consortium (http://geneontology.org/), DiseaseConnect (http://disease-connect.org/).

References

[1] Funk, C. S., I. Kahanda, A. Ben-Hur and K. M. Verspoor (2015). "Evaluating a variety of text-mined features for automatic protein function prediction with GOstruct." J Biomed Semantics6: 9.

[2] Preiss, J., M. Stevenson and R. Gaizauskas (2015). "Exploring Relation Types for Literature-based Discovery." J Am Med Inform Assoc.

[3] Ramezankhani, A., O. Pournik, J. Shahrabi, F. Azizi and F. Hadaegh (2015). "An application of association rule mining to extract risk pattern for type 2 diabetes using tehran lipid and glucose study database." Int J Endocrinol Metab13(2): e25389.

[4] Burkhart, K. K., D. Abernethy and D. Jackson (2015). "Data Mining FAERS to Analyze Molecular Targets of Drugs Highly Associated with Stevens-Johnson Syndrome." J Med Toxicol.

[5] Bravo, A., J. Pinero, N. Queralt-Rosinach, M. Rautschka and L. I. Furlong (2015). "Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research." BMC Bioinformatics16(1): 55.

[6] Taghizadeh, N., H. M. Boezen, J. P. Schouten, C. P. Schroder, E. G. Vries and J. M. Vonk (2015). "BMI and Lifetime Changes in BMI and Cancer Mortality Risk." PLoS One10(4): e0125261.

[7] Scholz, C., U. Andergassen, P. Hepp, C. Schindlbeck, T. W. Friedl, N. Harbeck, M. Kiechle, H. Sommer, H. Hauner, K. Friese, B. Rack and W. Janni (2015). "Obesity as an independent risk factor for decreased survival in node-positive high-risk breast cancer." Breast Cancer Res Treat.

[8] "Rapid-I: Rapid Miner." *Rapid - I*. Rapid - I, n.d. Web. 10 Nov. 2012.

[9] Jyoti Rani, S.Ramachandan and Ab. Rauf Shah (2014). Text mining of PubMed abstracts. R package version 1.0.4.

[10] Kevin Becker, Douglas Hosack, Glynn Dennis, Richard A Lempicki, Tiffani J Bright ,Chris Cheadle and Jim Engel (10 December 2003), PubMatrix: atool for multiplex literature mining BMC Bioinformatics, Vol. 4, No. 161.

[11] Ashburner et al, Gene ontology: tool for the unification of biology (2000) Nat Genet 25(1):25-9

[12] Liu CC, Tseng YT, Li W, Wu CY, Mayzus I, Rzhetsky A, Sun F, Waterman M, Chen JJ, Chaudhary PM, Loscalzo J, Crandall E, Zhou XJ. (2014) DiseaseConnect: a comprehensive web server for mechanism-based disease-disease connections. Nucleic Acids Research.

[13] Layla Oesper, Daniele Merico, Ruth Isserlin and Gary D Bader (2011). WordCloud: a cytoscape plugin to create a visual semantic summary of networks. Source Code for Biology and Medicine, 6:7.