Discovering Disease-associated Drugs Using Web Crawl Data

Hyunjin Kim, Sanghyun Park*
Department of Computer Science, Yonsei University
50 Yonsei-ro, Seodaemun-gu, Seoul, South Korea
{chriskim, sanghyun}@cs.yonsei.ac.kr

ABSTRACT
The purpose of research on biomedical literature-based discovery is to bring out new knowledge from the existing biomedical information. Beginning with Dr. Swanson's ABC model, many studies extended or applied the ABC model to find new associations between biomedical entities. While the methods applied to data have advanced, in most cases biomedical literature has been used for the text data. Assuming that web crawl data is helpful in studying literature-based discovery as well as biomedical literature which is the existing but rather limited data source, we discovered new disease-drug associations using web crawl data in addition to biomedical literature. We also analyzed how helpful the additional use of web crawl data is for biomedical literature mining. Literature-based discovery using web crawl data has its significance as a pioneering work utilizing new data.

Categories and Subject Descriptors
J.3 [**Life and medical sciences**]: Biology and Genetics

General Terms
Experimentation, Measurement, Verification

Keywords
literature-based discovery; biomedical text mining; disease-associated genes; web crawl data; ABC model;

1. INTRODUCTION
Research on literature-based discovery has advanced a lot since Dr. Don R. Swanson's study on Raynaud's syndrome in 1986 [1]. Until recently text mining has been used to extract unpublished information from literature published in the field of biomedical research. The core concept is Dr. Swanson's ABC model. The ABC model is a rather simple concept that if A and B share a connection and B and C share also a connection, then A and C may have an implicit connection to each other. Since it was turned out that by using this concept new information can be obtained from literature mining without real biomedical tests [2], many researchers have joined literature-based discovery studies. Since then, most of subsequent methods either extended or applied the ABC model.

The final aim of the ABC model is to find new connections between biomedical entities using text data. Three steps are necessary to do this. First, names of bio-entities and text data should be obtained. Then, associations get extracted from the obtained text data using the names of the bio-entities and lastly, new association that has never existed should be found using the extracted associations. Names of bio-entities and text data are needed to extract a biomedical association from the text data. Bio-entities can be disease name, gene name, protein name, drug name, symptom name and miRNA name. And biomedical research papers are mostly used for text data. If text data and a name list of bio-entities are secured, it can be found from the text data whether the bio-entities have connections between them. The most accurate way is to read through and to check the connections in person but it is not feasible because there is too much biomedical literature which can be used for text data. Therefore, it is necessary to find a method of detecting whether a text describes a certain bio-entity association: if names of two bio-entities come up together in one literature or names of two bio-entities appear at the same time in one sentence of one literature, it is generally supposed that the two bio-entities have an association between them.

As mentioned earlier, most of studies on literature-based discovery have been on methods to extract new associations which have never existed. First of all, there is a conventional method, Dr. Swanson's ABC model. The ABC model searches for the associations between A and C which are not on the existing list when associations exist between A and B and between B and C. Petric [3] proposed an algorithm which advanced the ABC model: to use rare terms which come up infrequently in all the literature as B for a middle stage. This method is based on the assumption that if the rare terms appear together with A and C both, there is a high possibility of association between A and C. In case of Li's method [4], a set of MeSH (Medical Subject Headings: the National Library of Medicine's controlled vocabulary thesaurus, used for indexing articles as citing the content of the literature with 10-15 terms) terms are used to create a term document matrix. Mutual information value can be calculated based on the matrix, then the bigger the value, the more important the two terms become. The next step is to apply ABC model and to choose term B having the largest mutual information among associations between A and B. Then, to choose term C in order of mutual information among term C related to the chosen term B. Li's method has a significant point in creating term document matrix and using the mutual information based on the matrix. Tsuruoka [5] also proposed a method of expanded ABC model that utilizes ranking approach considering strength and reliability of direct association and indirect association between term A and term C to extract new associations that have never existed.

In literature-based discovery field, studies on discovering new biomedical information from biomedical literature have been broadly researched. However, it is hard to find approaches which use other text data sources. Finite and fixed data sources make finite and fixed results. If we use new data sources, we have a chance to identify new and undiscovered knowledge that have not been discovered from biomedical literature. Surely, there exist a few approaches which do not use other text data sources but utilize other data sources such as

* Corresponding author. Tel.: +82 2 2123 5714

SAC 2016, April 04-08, 2016, Pisa, Italy
© 2016 ACM. ISBN 978-1-4503-3739-7/16/04…$15.00
DOI: http://dx.doi.org/10.1145/2851613.2851615

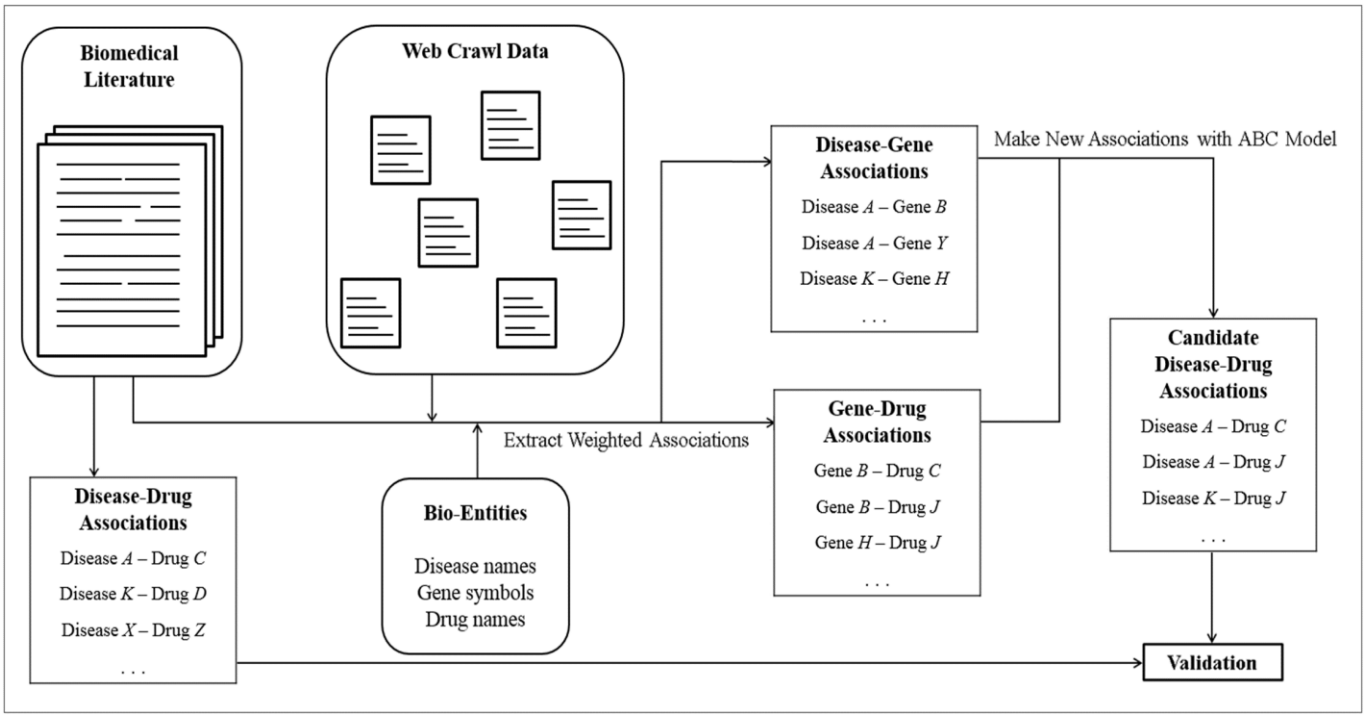

Figure 1: Overview of the proposed method

microarray data, protein-protein interactions, pathway data, and google search data.

Faro [6] proposed a method which exploits literature data and microarray data to find gene-disease associations. Bell [7] constructed integrated bio-entity network with literature, protein-protein interactions, Gene Ontology and KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway data. Eronen [8] predicted a link between bio-entities using Pubmed, Gene Ontology, KEGG pathway data, and OMIM (Online Mendelian Inheritance in Man) database. Kim [9] identified disease-associated genes using biomedical literature and google search results. The additional data sources mentioned above can help achieve a goal in literature-based discovery but they cannot play a major role as a biomedical literature. Currently, there is no related work which uses non-specific text data instead of biomedical literature in the literature-based discovery field. Web crawl data has extremely large texts, and therefore, we expected web crawl data can expand the limit of current literature-based discovery approaches. We proposed a novel approach which utilizes ABC model on biomedical literature and web crawl data. With the proposed method, we discovered disease-associated drugs which were not found from ABC model on biomedical literature. We also investigated the influence of the additional use of web crawl data and manually checked several disease-drug associations resulted from our method.

2. METHODS

The proposed method has four steps. It is similar to the Swanson's ABC model but we additionally utilize web crawl data (Figure 1). First, we prepare data sources required in the experiments. And then we extract disease-gene and gene-drug associations from biomedical literature and web crawl data. New candidate disease-drug associations are generated by using the ingredient associations. Lastly, we check whether the candidate associations are correct or not. We also decide the appropriate data amount ratio between biomedical literature and web crawl data.

2.1 Data Sets

The proposed method finds new associations through applying Swanson's ABC model to biomedical literature and web crawl data. Biomedical literature of 73.8GB size for the years 1980 to 2012 from MEDLINE (Medical Literature Analysis and Retrieval System Online) database is used for the biomedical literature. 81.7GB sized text data of Common Crawl which Amazon Web Services hosts is used for the web crawl data. Gene symbol, Disease name and Drug name which are necessary for the extraction from text data are found from PharmGKB (The Pharmacogenomics Knowledge Base) database.

2.2 Extracting Associations

Once the text data are gathered for the association extraction, bio-entity associations should be found from the text data. We extract disease-gene associations and gene-drug associations from the text data as bio-entity associations. For biomedical literature, the extraction is performed assuming that two entities have a relation when the two entities co-occurred in one sentence. However, the proposed method extracts association when the two entities simultaneously appear in one web page, not in one sentence. Since biomedical literature basically includes a lot of biological and medical information, we assume that there is a meaningful association when the two entities co-occurred in one sentence. Whereas web crawl data has relatively low probability of having biological and medical information, so we assume that there is a meaningful association when two entities co-occurred in one web page. There will be excessively large amount of associations if we make associations in one literature with biomedical literature data because the biomedical literature naturally contains a lot of biomedical-related concepts. On the other hand, web crawl texts are non-specific, so if we make

associations in one sentence, there will be few associations. Those are why we make associations when the two entities co-occur in one sentence in biomedical literature and when the two entities co-occur in one web page text in web crawl data. Disease-gene associations and gene-drug associations found in this step will be used to extract candidate disease-drug associations.

2.3 Generating Candidate Associations

Applying Swanson's ABC model to disease-gene associations and gene-drug associations extracted from biomedical literature and web crawl data, we build candidate disease-drug associations. As in ABC model, we assume that the disease and the drug related to the same gene possibly have relations to each other. The way to apply web crawl data to biomedical literature mining is a novel approach and it gets a different result depending on how much web crawl data is used to the existing biomedical literature, which is described in the next step.

2.4 Data Amount Ratio

Theoretically, the more information extracted from web crawl data is utilized in biomedical literature mining, the more candidate disease-drug associations and the better accuracy we get. However, it can be calculated how much web crawl data should be utilized for the limited size of biomedical literature to get efficient result. In our experiment, disease-gene associations and gene-drug associations extracted from 81.7 GB web crawl data are divided into ten equal parts and then cumulatively added to biomedical literature associations. Then we concluded that it is desirable to use 44% of biomedical literature size for web crawl data size in the experiment of this paper. There should be more studies on the data amount ratio between biomedical literature and web crawl data, but right now in this paper, 25:11 ratio between biomedical literature and web crawl data is recommended when building new candidate disease-drug associations. Details on this are described in Results section.

3. RESULTS

As experimental environments, we used Intel® Core™ i3 530 Dual 2.93 GHz, 8.00 GB RAM machine with Windows 7 operating system. We implemented our approach using JAVA programming language with JDK 7. The proposed method can find undiscovered disease-drug associations with biomedical literature and additional web crawl data. The main point of the proposed method is the utilization of the web crawl data in biomedical text mining. Therefore, we tried to identify that web crawl data is helpful for literature-based discovery, analyzed how efficient the additional use of web crawl data is, and investigated whether it can find disease-drug associations which are not found from biomedical literature data.

First, we implemented ABC model with only biomedical literature data to make candidate disease-drug associations. Then we compared the candidates with answer disease-drug associations to calculate the coverage. The disease-drug associations directly extracted from the MEDLINE database (This is not disease-drug associations from ABC model) is used as the answer disease-drug associations. We also did the same experiment only with web crawl data and with hybrid data (Biomedical literature + web crawl data) (Table 1). The ABC approach only with biomedical literature data found 23893 correct disease-drug associations out of 25721 answers and its coverage is 92.89%. The ABC approach only with web crawl data found 98.12% of the answer disease-drug associations which are 25238 disease-drug associations. Lastly, when we used both data sets, the result showed the largest coverage, 99.1%. Although the result of hybrid data showed the best coverage, it is identified that web crawl data can create more useful information than biomedical literature. It indicates that web crawl data can be an additional text source in literature-based discovery to support biomedical literature.

Table 1: Results of ABC model using 3 data types

	The number of Disease – Gene associations	**The number of Gene – Drug associations**	**The number of correct Disease – Drug associations**	**Coverage (%)**
Answer set	-	-	25,721	100
MEDLINE (73.8GB)	36,686	200,400	23,893	92.89
Web crawl data (81.7GB)	1,141,578	1,337,291	25,238	98.12
Hybrid [MEDLINE + Web crawl data (155.5GB)]	1,156,456	1,433,557	25,490	99.10

First, we implemented ABC model with only biomedical literature data to make candidate disease-drug associations. Then we compared the candidates with answer disease-drug associations to calculate the coverage. The disease-drug associations directly extracted from the MEDLINE database (This is not disease-drug associations from ABC model) is used as the answer disease-drug associations. We also did the same experiment only with web crawl data and with hybrid data (Biomedical literature + web crawl data) (Table 1). The ABC approach only with biomedical literature data found 23893 correct disease-drug associations out of 25721 answers and its coverage is 92.89%. The ABC approach only with web crawl data found 98.12% of the answer disease-drug associations which are 25238 disease-drug associations. Lastly, when we used both data sets, the result showed the largest coverage, 99.1%. Although the result of hybrid data showed the best coverage, it is identified that web crawl data can create more useful information than biomedical literature. It indicates that web crawl data can be an additional text source in literature-based discovery to support biomedical literature.

However, when using the web crawl data, the number of disease-gene associations and the number of gene-drug associations are larger than those from the biomedical literature data. The 1,141,578 disease-gene associations from 81.7GB web crawl data are about 30 times larger than the 36,686 disease-gene associations from 73.8GB MEDLINE data. In disease-gene association's case, associations from web crawl data are nearly 6.5 times larger than associations from MEDLINE. This is because we made associations when the two entities co-occurred in one sentence in biomedical literature and when the two entities co-occurred in one web page text in web crawl data. There will be excessively large amount of associations if we make associations in one literature with biomedical literature data because the biomedical literature naturally contains a lot of biomedical-related concepts. Similarly, web crawl texts are non-specific, so if we make associations in one sentence, there will be few associations. The disease-gene associations and the gene-drug associations are used to make candidate disease-drug associations, and therefore, the large number of the 'ingredient' associations can affect the final coverage. But our method is not intended to replace existing disease-associated

genes discovering methods; rather, it serves to provide additional information that, although there are a number of candidate associations, web crawl data has answer ingredients and it can play a key role in literature-based discovery.

We also analyzed how helpful the additional use of web crawl data is for literature-based discovery (Figure 2). The disease-gene associations and the gene-drug associations extracted from 81.7 GB web crawl data are divided into ten equal parts and then cumulatively evaluated. Those divided data also cumulatively added to biomedical literature associations to make hybrid results. The results of hybrid data showed better performance and less unstable than those of web crawl data. The results of hybrid data are more stable because they equally contain the result of biomedical literature. Although there are not many differences in the results of hybrid data, we investigated which data size induces the most efficient result (Figure 3). In Figure 3, the gradient of the tangent showed how much the number of correct associations changed in each data amount. It is obvious that the more the web crawl data is used, the more the correct associations can be found. But the more the data is used the less efficient it is. There is a certain data amount that satisfies both performance and efficiency. We found that using 40% of web crawl data is appropriate for the experiment of this paper.

There are significant changes using up to 40% of the web crawl data but when using 50% or more of the web crawl data, the results barely change. Based on that, we determined using 40% of the web crawl data (32.7GB) is efficient and shows considerable performance. 32.7GB size is 44% of the 73.8GB sized biological literature. In other words, if we use 25:11 ratio between biomedical literature and web crawl data, we may get good result with reasonable amount of data when making new candidate disease-drug associations.

We tried to suggest undiscovered disease-drug associations using both biomedical literature and web crawl data with the elimination of answer disease-drug associations but it is impossible to provide the list because there are an extremely large number of candidate disease-drug associations. Instead, we provide 1597 disease-drug associations gained only from the web crawl data (http://delab.yonsei.ac.kr/files/hjkim/SAC2016/Disease-Drug(1597).tsv). Moreover, we manually checked 10 disease-drug associations from the 1597 associations (Table 2). The diseases and the drugs of the manually checked 10 associations were proved to be actually related.

4. CONCLUSION

Our approach showed that by utilizing web crawl data in addition to biomedical literature, web crawl data may supplement biomedical literature. And it can find more diverse knowledge assisting biomedical literature out of those methods using mostly existing biomedical literature in searching for new knowledge. Since ABC model is a basic approach of literature-based discovery and those other methods also apply ABC model to theirs, we believe that these methods as well as ABC model will get good results by utilizing web crawl data. However, there are two considerations when utilizing web crawl data for literature-based discovery. First is what ways to extract bio-entity associations. New methods are required that are more suitable for the characteristics of web crawl data than the existing methods applied to biomedical literature. Secondly, large amounts of bio-entity associations are generated because of the large volume of web crawl data. While MEDLINE data is not over 100GB size that includes all the biomedical literature from year 1946 to date, web crawl data has much bigger size of over 500TB to beyond count. Though it has an advantage of having much information from the big data, it has so much noise by which too many bio-entity associations get generated. Therefore, methods are necessary to decrease the number of bio-entity associations accurately and effectively. These two points could be future works of literature-based discovery using web crawl data. The purpose of our approach is not to replace the existing methods but to verify that web crawl data can be assistive for literature-based discovery. In other words, it is to show the position of gold mine and that it is actually there. The methods of effective mining for gold should be developed to extract gold from stones in large quantities in the gold mine.

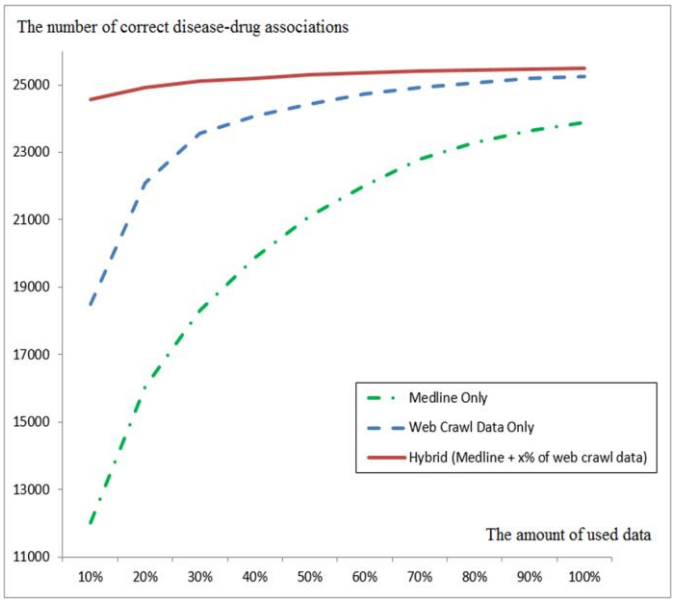

Figure 2: The influence of web data according to the data size

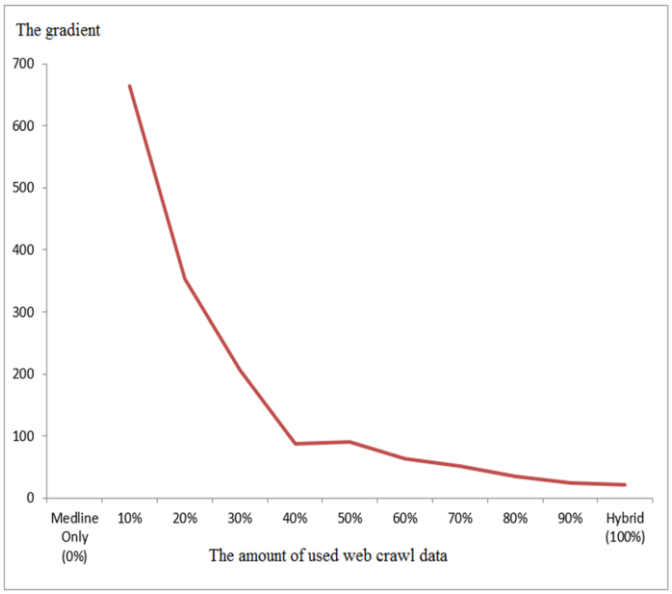

Figure 3: The gradient of the tangent of the influence graph

Table 2: Text summary of the manually checked associations

Disease	Drug	Text Summary
Breast Cancer	Tocopherol	Tocopherol is an apoptotic inducer for human breast cancer cells [10]. Tocopherol is effective in reducing tumor burden and metastasis of human breast cancer cells [11].
Gestational Diabetes	Metformin	Metformin is a logical treatment for gestational diabetes [12]. Metformin is a safe and effective alternative to insulin in the treatment of gestational diabetes especially suitable for women with mild gestational diabetes [13].
Menopause	Gabapentin	Gabapentin reduces frequency and severity of hot flashes in women with menopause [14].
Headaches	Cabergoline	Headache is one of the side effects of cabergoline [FDA (U.S. Food and Drug Administration) Prescribing information].
Hives	Zinc	Hives can be caused by zinc ["Zinc & Hives" from LIVESTRONG.COM (http://www.livestrong.com/article/548363-zinc-hives/)].
HIV	BMS-232632	BMS-232632 may be an effective HIV-1 inhibitor [15-16].
Apathy	Ropinirole	Ropinirole may be a treatment option for post-stroke apathy [17]. Ropinirole may be an effective treatment for patients who develop apathy after encephalitis [18].
Hypoglycemia	Liraglutide	Liraglutide can treat type 2 diabetes with low incidence of hypoglycemia [19-20].
Warts	Bleomycin	A bleomycin injection is effective and non-toxic treatment in periungual and palmo-plantar warts [21].
Hyperlipidemia	Retinoic Acid	Hyperlipidemia is often seen in patients treated with high-dose retinoic acid [22].

5. ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF-2015R1A2A1A05001845). We also appreciate Mr. Junsik Kim's proofreading efforts.

6. REFERENCES

[1] Don R. Swanson, "Fish oil, Raynaud's syndrome, and undiscovered public knowledge", Perspectives in Biology and Medicine, vol. 30(1), pp. 7-18, 1986.

[2] Don R. Swanson, "A second example of mutually isolated medical literatures related by implicit, unnoticed connections", Journal of the American Society for Information Science, vol. 40, no. 6, pp. 432-435, 1989.

[3] Ingrid Petric, T. Urbancic, B. Cestnik, and M. Macedoni-Luksic, "Literature mining method RaJoLink for uncovering relations between biomedical concepts", Journal of Biomedical Informatics, vol. 42, no. 2, pp. 219-227, 2009.

[4] Guangrong Li and Xiaodan Zhang, "Mining Biomedical Knowledge Using Mutual information ABC", IEEE International Conference on Granular Computing, 2011.

[5] Yoshimasa Tsuruoka, Makoto Miwa, Kaisei Hamamoto, Jun'ichi Tsujii, and Sophia Ananiadou, "Discovering and visualizing indirect associations between biomedical concepts", Bioinformatics, vol. 27, issue 13, pp. i111-i119, 2011.

[6] Alberto Faro, D. Giordano, and C. Spampinato, "Combining literature text mining with microarray data: advances for system biology modeling", Briefings in Bioinformatics, vol. 13, no 1, pp. 61-82, 2011.

[7] Lindsey Bell, Rajesh Chowdhary, Jun S. Liu, Xufeng Niu, and Jinfeng Zhang, "Integrated Bio-Entity Network: A System for Biological Knowledge Discovery", PLOS ONE, vol. 6, issue 6, e21474, 2011.

[8] Lauri Eronen and Hannu Toivonen, "Biomine: predicting links between biological entities using network models of heterogeneous databases", BMC Bioinformatics, vol. 13, 119, 2012.

[9] Jeongwoo Kim, Hyunjin Kim, Youngmi Yoon, and Sanghyun Park, "LGscore: A Method to Identify Disease-Related Genes using Biological Literature and Google Data", Journal of Biomedical Informatics, vol. 54, pp. 270-282, 2015.

[10] W. Yu, M. Simmons-Menchaca, A. Gapor, B. G. Sanders, and K. Kline, "Induction of apoptosis in human breast cancer cells by tocopherols and tocotrienols", Nutrition and cancer, vol. 33, issue 1, pp. 26-32, 1999.

[11] K. Kline, W. Yu, and B. G. Sanders, "Vitamin E and breast cancer", The Journal of nutrition, vol. 134, suppl. 12, pp. 3458-3462, 2004.

[12] Janet A. Rowan, William M. Hague, Wanzhen Gao, Malcolm R. Battin, and M. Peter Moore, "Metformin versus Insulin for the Treatment of Gestational Diabetes", The New England Journal of Medicine, vol. 358, pp. 2003-2015, 2008.

[13] H. Ijäs, M. Vääräsmäki, L. Morin-Papunen, R. Keravuo, T. Ebeling, T. Saarela, and T. Raudaskoski, "Metformin should be considered in the treatment of gestational diabetes: a prospective randomised study", An International Journal of Obstetrics and Gynaecology, vol. 118, issue 7, pp. 880-885, 2011.

[14] Konstantinos A. Toulis, T. Tzellos, D. Kouvelas, and D. G. Goulis, "Gabapentin for the treatment of hot flashes in women with natural or tamoxifen-induced menopause: A systematic review and meta-analysis", Clinical Therapeutics, vol. 31, issue 2, pp. 221-235, 2009.

[15] B. S. Robinson, K. A. Riccardi, Y. F. Gong, Q. Guo, D. A. Stock, W. S. Blair, B. J. Terry, C. A. Deminie, F. Djang, R. J. Colonno, and P. F. Lin, "BMS-232632, a highly potent human immunodeficiency virus protease inhibitor that can be used in combination with other available antiretroviral agents", Antimicrobial agents and chemotherapy, vol. 44, issue 8, pp. 2093-2099, 2000.

[16] A. Schuster, S. Burzawab, M. Jemala, E. Loizillonb, P. Couerbeb, and D. Whigan, "Quantitative determination of the HIV protease inhibitor atazanavir (BMS-232632) in human plasma by liquid chromatography–tandem mass spectrometry following automated solid-phase extraction", Journal of Chromatography B, vol. 788, issue 2, pp. 377-386, 2003.

[17] N. Kohno, S. Abe S, G. Toyoda, H. Oguro, H. Bokura, and S. Yamaguchi, "Successful treatment of post-stroke apathy by the

dopamine receptor agonist ropinirole", Journal of clinical neuroscience, vol. 17, issue 6, pp. 804-806, 2010.

[18] N. Kohno, N. Kohno N, Y. Nabika, G. Toyoda, H. Bokura, T. Nagata, and S. Yamaguchi, "The effect of ropinirole on apathy and depression after herpes encephalitis", Cognitive and behavioral neurology, vol. 25, issue 2, pp. 98-102, 2012.

[19] T. Vilsboll, "Liraglutide: a new treatment for type 2 diabetes", Drugs Today, vol. 45, issue 2, pp. 101-113, 2009.

[20] C. M. Mulligan, R. Harper, J. Harding, W. McIlwaine, A. Petruckevitch, and D. M. McLaughlin, "A Retrospective Audit of Type 2 Diabetes Patients Prescribed Liraglutide in Real-Life Clinical Practice", Diabetes Therapy, vol. 4, issue 1, pp. 147-151, 2013.

[21] Prasoon Soni, Kanika Khandelwal, Naushin Aara, Bhikam C. Ghiya, Rajesh D. Mehta, and Ram A. Bumb, "Efficacy of Intralesional Bleomycin in Palmo-plantar and Periungual Warts", Journal of Cutaneous and Aesthetic Surgery, vol. 4, issue 3, pp. 188-191, 2011.

[22] S. Y. Cai, H. He, T. Nguyen, A. Mennone, and J. L. Boyer, "Retinoic acid represses CYP7A1 expression in human hepatocytes and HepG2 cells by FXR/RXR-dependent and independent mechanisms", Journal of lipid research, vol. 51, issue 8, pp. 2265-2274, 2010.