Data Mining in MEDLINE for Disease-Disease Associations via Second Order Co-Occurrence

Modest von Korff, Bernard Deffarge and Thomas Sander
Research Information Management
Actelion Pharmaceuticals Ltd., Allschwil, Switzerland
Email: modest.korff@actelion.com

Abstract—**DDMiner, a new method for mining disease-disease associations in MEDLINE, is presented together with its first results. DDMiner searches for co-occurrences of gene names and disease terms, and finds relationships between diseases by word vector-similarity calculations. All records in PubMed were labeled with around 40,000 gene and protein names, and around 4,000 disease terms. Each disease term was described by a word vector from which the length equals the number of gene names. Each field in the vector represented a gene or a protein. The value in the field was derived from the number of publications in which this gene occurred together with the disease term. Disease-disease associations were calculated by vector-similarity calculation. Five diseases were examined together with their closest neighbor diseases to show the validity of our approach. All five examples showed only disease-disease associations that could be validated by medical literature. These results show that mining for disease-disease associations by second order co-occurrence is a powerful tool for medical science.**

I. Introduction

Many diseases, if not most, are associated with other diseases. These disease–disease associations are often considered by physicians while examining their patients. Elderly or very sick patients rarely suffer from a single disease alone, and ignoring the relationships between diseases might impair the treatment of patients. Studying disease–disease associations is also of paramount importance for drug discovery. Patients may be receiving multiple drugs for various diseases, which could result in undesired drug–drug interactions. Thus, searching for disease–disease associations is crucial, however not a new topic in medical science. Physicians have observed for decades that some diseases were often accompanied by others [1, 2]. The observations made were sometimes collected in reviews [3]. Due to the high interest in disease–disease associations, several data mining techniques were already applied to support the research on this topic. The initiatives were triggered in particular by new databases such as the Online Mendelian Inheritance in Man (OMIM) database, which relates genetic disorders to diseases [4]. OMIM was used by several research groups to derive disease–disease associations; for example, a network of human diseases was created by Goh et al. [5]. They used the gene sets from OMIM to build their disease network. In addition, disease similarities for 54 diseases were calculated by Suthram et al. [6], and an enrichment of disease candidate genes via text mining of OMIM descriptions was implemented by van Driel et al. [7]. MeSH annotations of MEDLINE articles were analyzed by Liu et al. to derive genetic and environmental factors for diseases [8]. Regardless of all the approaches already developed for detecting disease–disease associations, there is still room to include more information than what was previously done.

Our presented approach, DDMiner, differs from the previous methods by annotating all records in MEDLINE with gene names and MeSH terms. Disease–disease associations were derived by comparing stochastic gene patterns. These gene patterns were extracted from the records in MEDLINE by mining for co-occurrences of gene names and disease terms. This technique is known as second order co-occurrence and was already applied by Schütze for word sense discrimination [9]. He discovered that similar words often appear with groups of identical words. Second-order co-occurrence is able to calculate the similarity between two words that do not co-occur frequently, because they co-occur with the same neighboring words. Consequently, second order co-occurrence can be regarded as a small world phenomenon [10].

In DDMiner, we have limited the considered co-occurring expressions to gene-name synonyms. Many diseases are related to some change in the expression of proteins, and usually, a protein expression change is caused by a change in gene expression. In cancer related diseases, the changes seem to be tremendous, when compared to healthy control groups. Inflammation-related diseases also tend to cause large changes in the protein expression pattern. A prominent example is rheumatoid arthritis, which affects millions of people worldwide. Neurodegenerative diseases such as Alzheimer, Parkinson and Huntington also show different protein expression patterns when compared to controls. Searching disease–disease associations by gene patterns decouples the results from the necessity that the co-occurrence of two diseases was already published. The DDMiner approach searches for gene–disease associations and compares the gene

978-1-4799-7560-0/15 $31.00 © 2015 IEEE
DOI 10.1109/SSCI.2015.54

patterns to detect associated diseases. Gene–disease associations are an important field of research in medicine and drug discovery [11]. A query on PubMed [12] with the expression “gene–disease associations” resulted in about 278,000 hits. Additionally, the vast majority of publications on gene–disease associations do not mention these key words. Unfortunately, most medical information is not published in open-access journals and is therefore not freely available. However, almost all relevant biomedical literature is indexed in MEDLINE [13]. With more than 22 million bibliographic records, MEDLINE is the largest repository for biomedical literature. Interfaces such as PubMed enable interactive and programmatic access. Many software tools that exploit gene–disease associations already exist. A comprehensive overview has been recently published by Piro et al. [14]. They described 36 tools for disease–gene extraction, and additional tools were recently described by: Rappaport et al. [15], Peng et al. [16], and Pletscher-Frankild et al. [17]. The most recent one provides also a webpage with the results for numerous genes [18]. Nevertheless, no standard algorithm exists for searching gene-name synonyms in PubMed records.

Section II describes the DDMiner algorithms while section III summarizes the mining of the PubMed database and discusses the disease-disease associations that were found for five diseases. The results are summarized and some conclusions are given in Section IV.

II. Methods

A. Gene names and synonyms

A gene name is the starting point for any gene–disease association. DDMiner takes a gene name and derives a list of synonyms from two sources. A table with Human Genome Organization (HUGO) identifiers, gene names, approved symbols and synonyms is automatically compiled from HGNC (HUGO Gene Nomenclature Committee) [19]. The HUGO Gene Nomenclature Committee is part of the European Bioinformatics Institute and works under the supervision of the Human Genome Organization. The second source is the MEDLINE database EntrezGene, which also provides HUGO ids, gene names, and synonyms [20, 21]. Both these databases are used because they do not completely overlap. Not all gene-name synonyms are useful for searching PubMed. Three-letter synonyms often have more than one meaning. Using a three-letter word as a query would retrieve all records containing this three-letter word either in the title or in the abstract. Therefore, these synonyms are excluded from the database queries. Synonyms containing two letters and at least one digit are allowed in database queries.

B. Programmatic access to PubMed

The MEDLINE databases can be accessed programmatically via the Entrez tools [22]. A query, containing a search term, submitted to MEDLINE via the PubMed interface returns a list of identifiers (PMID) used to obtain the publication records (**R**). These records contain bibliographic information, often an abstract and the MeSH term headings, which are used to index these articles. These MeSH term headings are not used by DDMiner. DDMiner relies on its own indexing of the retrieved PubMed records.

C. Querying PubMed Central with gene name synonyms

With the synonyms retrieved from HGNC and EntrezGene, the queries are generated to search the PubMed Central database. One PubMed query is created for every single synonym. Without any further specification, all fields in the PubMed database are searched. Depending on the synonym, no records at all, or up to several tens of thousands, are retrieved. The result is a dataset ($\mathbf{R}_{\mathrm{Gene}}$) for each gene, containing the retrieved records. PubMed queries with the Entrez tool do not distinguish between lower-case and upper-case letters. Unfortunately, many letter combinations exist, which differ in capitalization and are shared by different terms. Consequently, up to ten thousands of false-positive records can be retrieved for a single gene.

D. Text normalization

The sentences are extracted from the text with the Apache library OpenNLP 1.5.3 [23]. All words that contain only one capital letter are de-capitalized. Greek letters are out-written as full text. A further complication for the Greek letter beta is caused by the misuse of the German letter 'sharp s' as beta. German sharp s without surrounding letters is changed into 'beta'. The Greek letter 'µ' is also available in the enhanced ASCII code. This needs an extra detection and 'µ' is retyped as 'mu'. Text that contains text and numbers like '12hydroxy' is split up into '12 hydroxy'. Gene name synonyms containing numbers like BACE1 remain untouched. Finally, all punctuation is removed.

E. Filtering of PubMed records

Two post-processing steps can be added to get rid of the false-positive records:

1) White list filtering of PubMed records

If a synonym consists of less than six characters and does not contain a space, the retrieved PubMed records are filtered for the exact upper- and lower-case pattern of the synonym. However, after this filtering process, many false-positive records still remain. These records contain terms with an identical synonym to the gene under consideration. False-positive records that contain the exact synonym can be detected by analyzing the context of the synonym. It has to be related to the concept of a 'gene'. For the record filter in DDMiner, a gene context list of 25 terms has been defined: activation, activator, allosteric, chromatin, chromosome, codon, exon, expression, gene, genome, genotype, histone, homolog, inhibitor, inhibition, intron, modulator, mutant, nucleosome, peptide, phenotype, phenotypic, polymerase, protein, target, transcript, and transposon. If a PubMed record does not contain any of these words, it is very unlikely that the record is related to a gene. Consequently, a PubMed record is only accepted if it contains at least one of the words from the context list.

2) Filtering for disease MeSH terms

All records are skipped that do not contain a disease MeSH term.

3) Filtering for gene name synonyms

After normalizing the PubMed records and the gene-name synonyms, the records are searched with the synonyms. For matching gene-name synonyms the following algorithm has been developed: for a PubMed record to be considered as a match for a given gene under consideration it has to fulfill the following criteria: Each word of a gene-name synonym has to match with at least two–third of its words, the words in a text string that were delivered by a sliding window algorithm. For a gene-name synonym with *n* words all *n* word n-grams were word wise compared. Two words are considered as a match if their similarity is equal or higher than 0.45. Word–word similarities are calculated using the Damerau–Levenshtein algorithm as implemented by Kevin Stern [24]. The sliding window algorithm provides strings with a length of two more words than the gene-name synonym term. This made the search equivalent to Lucene Fuzzy Query with a sloppiness value of 2 [25].

F. Stoplists

Stoplists are lists of node names that are used to activate or inactivate branches in the MeSH tree [26]. Only active MeSH terms are used for searching corresponding expressions in the data. Branch 'C' in the MeSH tree is already dedicated to diseases. For the disease stoplist used in this examination, the sub-branch C22, which contains MeSH terms related to animal diseases, has been excluded. Also, the disease-relevant sub-branches from branch F containing terms for psychiatry and psychology has been included. In total, the stoplist disease is described by 11,177 nodes with 4606 unique MeSH descriptors.

G. Indexing with Lucene

The Apache library Lucene is a powerful text indexing and search tool written in Java [27]. It is an open-source tool and is well documented. Lucene is widespread and powers the indexing and the search of many websites. Lucene provides a lot of powerful features such as in-memory indexing, ranked searching, sorting by field, combined fields search, and different types of text analyzer for different languages. In this study, a distinct in-memory index has been created for every set of PubMed records ($\mathbf{R}_{\text{Gene}}$) Lucene is configured to use the standard analyzer from version 4.6. Title, MeSH descriptors, and summary of the PubMed records are indexed by Lucene. In this way, an index ($\mathbf{I}_{\text{Gene, R}}$) is generated for every ($\mathbf{R}_{\text{Gene}}$).

H. Searching Lucene indices with disease MeSH terms

The indices ($\mathbf{I}_{\text{R}}$) is searched for matching disease MeSH terms by using the standard analyzer from Lucene. A multi-field query analyzer is initialized to query all fields in the indices. A list of query terms has been compiled from the disease stoplist. For example, an entry in this list can contain a descriptor like 'Alzheimer Disease'. A list of entry points is connected to this descriptor. Each entry point is a synonym or a closely related term used in literature (e.g. a list of 32 entry points is given for 'Alzheimer Disease', containing the terms 'Disease, Alzheimer', 'Alzheimer Syndrome', 'Senile Dementia' or 'Presenile Dementia'). To query an index for a disease, a dedicated Lucene query is created for the descriptor and for every entry point. If either the descriptor or at least one of the entry points is found by the Lucene index searcher and the similarity score is ≥0.5, the disease is added to a result list. It is also recorded how often a disease term is found by the index searcher. For a single record, an occurring disease MeSH term is only counted once. The algorithm produces for every gene a list of diseases together with their frequency ($\mathbf{LD}_{\text{Gene}}$). This list is sorted according to the frequency of occurrence with the most frequent disease at the top.

I. Disease by gene descriptor-vector

A number of 39,410 approved gene and protein symbols has been retrieved from the HUGO table. This number defines the size of the descriptor-vector that is used to describe a disease from its genes. Each field in the vector corresponds to an approved gene or protein symbol from the HUGO table. To create a descriptor-vector ($\mathbf{DD}_{\text{Disease}}$) for a disease from its genes, all tables ($\mathbf{LD}_{\text{Gene}}$) are searched for the disease and the frequencies of occurrence of the genes are kept in the corresponding fields of the descriptor-vector. One field in the disease descriptor-vector represents the number of publications in which the corresponding gene was mentioned together with the disease. Some genes are of higher interest in the medical literature than others. They are mentioned much more often than genes of lower interest. For this reason each field in the descriptor-vector is scaled between 0 and 100. The scaled descriptor-vectors ($\mathbf{DD}_{\text{Disease scaled}}$) are used for the similarity calculations. A similarity between two vectors ($\mathbf{DD}_{\text{Disease scaled}}$) is calculated by the normalized Manhattan Block distance.

III. Results

The synonyms for the 39,410 approved symbols from the HUGO table, proteins and gene names, were extracted from PubMed and HUGO. One to twenty synonyms were retrieved, including the approved symbol itself. These synonyms were used to search the PubMed database for publication records. A total number of about 10 million unique PubMed records were retrieved. Every retrieved record went through the text normalization procedure. The gene name synonyms were searched in the normalized text. The indexing with the disease MeSH terms was done on the original text. All PubMed records containing a gene name synonym and a MeSH disease term were used to create the disease descriptor-vectors ($\mathbf{DD}_{\text{Disease}}$). For 14,256 genes at least one disease term was found. A disease descriptor-vector could be calculated for 1,790 diseases. Five diseases were chosen to exemplify the results of the similarity calculations: Alzheimer's Disease, Dermatitis, Glioblastoma, Hepatitis C, and Hypertension. The similarity values between the five examples and all other diseases were calculated. Similarities below a threshold of 0.85 were not considered. Tables 1-5 show the disease similarity and the number of shared genes. A shared gene is a gene that was mentioned together with both diseases, but not necessarily in the same publication. In Figures 1-5 the similarity relations were sketched as a graph. Every node represents a disease. An edge between two nodes represents a similarity above or equal to the threshold. The graph was calculated up to a depth of two nodes.

Alzheimer's Disease is a neurodegenerative disease [28]. Five MeSH terms with a similarity above the threshold were found: Neurodegenerative Diseases, Dementia, Nerve Degeneration, Brain Ischemia and Neuroblastoma (Table I). Neurodegenerative Diseases and Alzheimer's Disease share 1260 genes. This means that 1260 different genes were mentioned together with Alzheimer's Disease or together with Neurodegenerative Diseases in at least one publication. The Neurodegenerative Diseases node is an ancestor node in the branch of the MeSH tree, where the Alzheimer's Disease node is located. Dementia is located in another branch than Neurodegenerative Diseases, however Alzheimer's Disease is also a leaf in this branch. Brain Ischemia, a condition where restricted blood flow leads to low oxygen levels in the brain, influences the development of Alzheimer's Disease [29]. Figure 1 shows the association of Alzheimer's Disease with other diseases. On the graph, it can be seen that Alzheimer's Disease and Stroke share a common node: Brain Ischemia. This indirect link has been validated by literature; Honig et al. examined the relation between these two diseases [30].

TABLE I. ALZHEIMER AND SIMILAR DISEASES

Disease	***Similarity***	***Shared gene set***
Neurodegenerative Diseases	0.897	1260
Dementia	0.890	970
Nerve Degeneration	0.872	991
Brain Ischemia	0.860	940
Neuroblastoma	0.859	1180

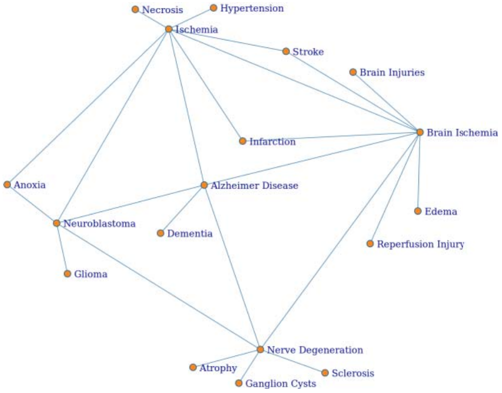

Fig. 1. Disease similarity graph for Alzheimer disease.

Dermatitis is a generic term and means any inflammatory disease of the skin. A number of ten similar disease MeSH nodes were found (Table II). These nodes was sub-divided into three groups: 1) the generic terms Hypersensitivity, Skin Diseases and Lung Diseases; 2) sub-types of dermatitis, and 3) similar diseases. Dermatitis sub-types included: Contact Dermatitis and Atopic Dermatitis. Similar diseases included: Psoriasis, Asthma, Colitis, Pneumonia, and Bacterial Infections. Psoriasis, Asthma and Colitis are auto immune diseases of the skin and are related to Dermatitis by their inflammatory nature [31]. Pneumonia is an inflammatory disease and co-occurs with dermatitis. The disease graph in Figure 2 shows relationships between Dermatitis and Arthritis via two nodes: Hypersensitivity and Asthma. All four diseases have an inflammatory component, but the involved biochemical pathways differ for Dermatitis and Arthritis [32].

TABLE II. DERMATITIS AND SIMILAR DISEASES

Disease	***Similarity***	***Shared gene set***
Dermatitis, Atopic	0.970	722
Psoriasis	0.907	746
Skin Diseases	0.903	737
Hypersensitivity	0.870	889
Asthma	0.867	784
Colitis	0.865	734
Dermatitis, Contact	0.862	508
Pneumonia	0.859	771
Bacterial Infections	0.855	676
Lung Diseases	0.852	687

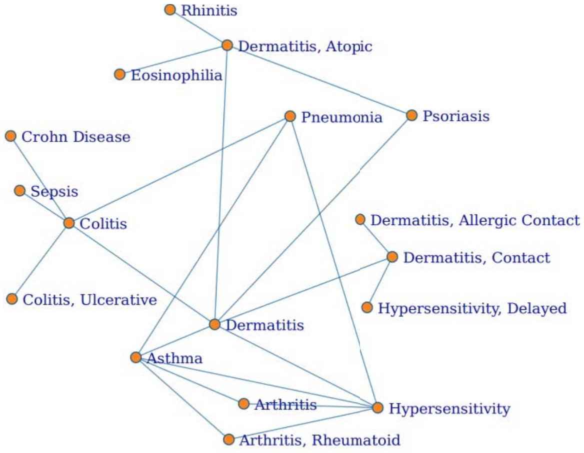

Fig. 2. Disease similarity graph for Dermatitis.

Glioblastoma is a malignant brain tumor. All ten similar disease MeSH terms were tumor related (Table III). The three most similar diseases, Brain Neoplasms, Glioma and Astrocytoma, are also tumors of the brain. Pathologic Neovascularization is the formation of blood micro vessels, which are needed by the tumor for further growth. From the disease graph in Figure 3, it can be seen that Glioblastoma is related to other cancer types above and over the first layer of similar disease MeSH nodes.

TABLE III. GLIOBLASTOMA AND SIMILAR DISEASES

Disease	***Similarity***	***Shared gene set***

Disease	***Similarity***	***Shared gene set***
Brain Neoplasms	0.958	1468
Glioma	0.949	1582
Astrocytoma	0.886	1014
Neovascularization, Pathologic	0.861	985
Neoplasm Invasiveness	0.860	1268
Cell Transformation, Neoplastic	0.855	1312
Pancreatic Neoplasms	0.854	1038
Colonic Neoplasms	0.854	1136
Carcinoma, Squamous Cell	0.853	1176
Carcinoma, Non-Small-Cell Lung	0.851	943

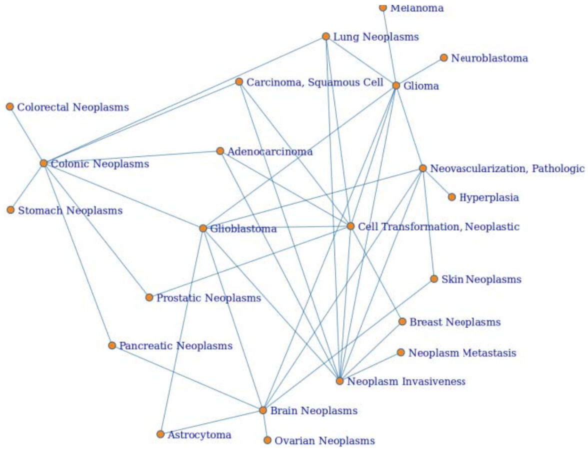

Fig. 3. Disease similarity graph for Glioblastoma.

Hepatitis C is an infectious disease caused by a virus. The four similar disease MeSH terms were all Hepatitis related (Table IV). In Figure 4, indirect connections from Hepatitis C to Liver Cirrhosis, Fibrosis and Liver Neoplasms are shown. Hepatitis and Liver Cirrhosis are associated diseases, as described by Bruix et al. [33] and Fibrosis is often observed in Hepatitis C patients [34]. Moreover, Hepatitis C is regarded as a risk factor for Liver Neoplasms [35].

TABLE IV. HEPATITIS C AND SIMILAR DISEASES

Disease	***Similarity***	***Shared gene set***
Hepatitis	0.956	1384
Hepatitis C, Chronic	0.911	706
Hepatitis, Chronic	0.910	746
Hepatitis B	0.874	799

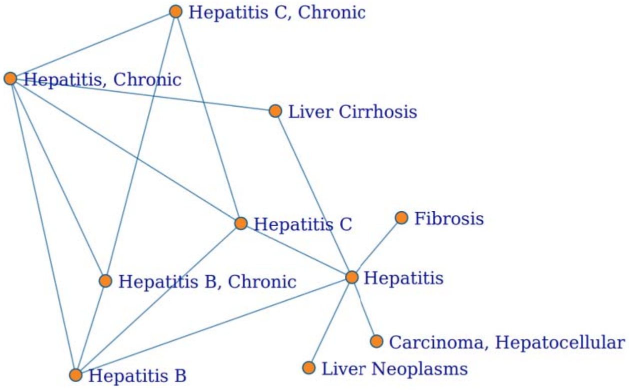

Fig. 4. Disease similarity graph for Hepatitis C.

Hypertension is a disease of the vascular system where high blood pressure persists in arteries. Hypertension is also known to be associated with multiple diseases. Consequently, our network included a node with many connections: Hypertrophy, Stroke, Cardiomyopathies, Hypertrophy, Infarction, Fibrosis, Body Weight are only some of them. Hypertrophy is the most similar disease MeSH term to hypertension (Table V). Hypertrophy of the heart means an increase of the left or right ventricle volume and is a frequent consequence of hypertension. Fibrosis is also a known complication in hypertension [36]. Interestingly, a relation between body weight and hypertension was already published fifty years ago [37]. Preventing stroke and infarction by treating hypertension is common practice in medicine [38] [39]. Another interesting MeSH disease node is Smoking, with a similarity of 0.86. Smoking is not a disease in itself, but it is known as a risk factor for developing hypertension. The disease graph in Figure 5 shows many known influence factors for hypertension. Some cancer types can be found in the second level around Hypertension. It is known that some cancer types influence the blood pressure. Additionally, an increase in blood pressure is a possible side effect of cancer treatment.

TABLE V. HYPERTENSION AND SIMILAR DISEASES

Disease	***Similarity***	***Shared gene set***
Hypertrophy	0.910	1462
Ischemia	0.906	1499
Diabetes Mellitus	0.905	1667
Infarction	0.905	1315
Cardiovascular Diseases	0.903	1242
Fibrosis	0.900	1688
Body Weight	0.897	1383
Atherosclerosis	0.896	1346
Myocardial Infarction	0.895	1201
Stroke	0.893	1325
Anoxia	0.890	1574
Chronic Disease	0.881	1244

Disease	*Similarity*	*Shared gene set*
Pain	0.871	1249
Obesity	0.871	1513
Diabetes Mellitus, Experimental	0.864	1083
Depression	0.864	1262
Hemorrhage	0.862	1119
Necrosis	0.862	1720
Disease Progression	0.861	1662
Hyperplasia	0.861	1316
Smoking	0.858	1200
Diabetes Mellitus, Type 2	0.858	1231
Insulin Resistance	0.857	1285
Edema	0.852	1075
Shock	0.851	1533

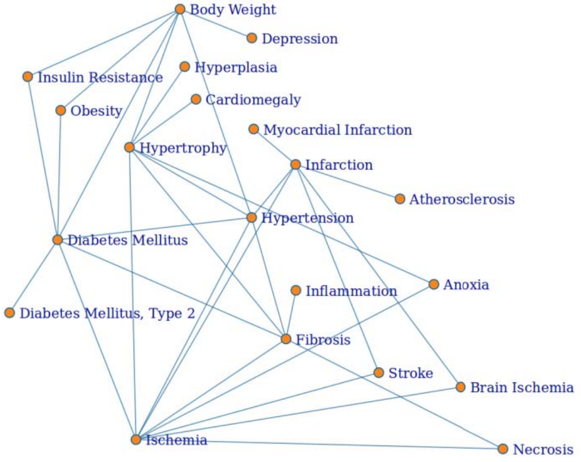

Fig. 5. Disease similarity graph for Hypertension.

Results summary: Five diseases and their most similar disease MeSH terms were analyzed. The smallest set of shared genes was found for Dermatitis and Contact Dermatitis with 508 shared genes. Hypertension and Necrosis shared the largest set of genes; with 1720 genes it is more than three times the size of the smallest set. Similar disease MeSH terms contained generic disease terms like Neurodegenerative Diseases for Alzheimer's Disease or Hepatitis for Hepatitis C. DDMiner found these relations with similarity comparisons of the shared gene sets. These similarity relations were verified with the MeSH tree. In the MeSH tree the more generic terms Neurodegenerative Diseases and Hepatitis are ancestor nodes for the specific terms Alzheimer's Disease and Hepatitis C. Many of the similar disease MeSH nodes are already obviously related. Examples are the different types of Hepatitis to Hepatitis C and the different types of brain tumors to Glioblastoma. Nevertheless, all similarity connections were validated by searching the medical literature. For most of the similarity connections a literature example was given. The disease graphs in Figures 1-5 visualize the similarity relations beyond the direct neighbour nodes. Some of these neighbour relations were checked for plausibility. For example, Hypertension is related to different cancer types. And Hepatitis C has a strong association to Fibrosis.

IV. Summary and Conclusions

A total number of around 10 million PubMed records were annotated with almost 40,000 genes and 4,000 disease MeSH terms. At least one gene-name synonym and one disease MeSH term were found in 2.5 million PubMed records. These labeled records were used to create around 4000 word vectors. One vector for each disease and each field in the vector corresponded to a gene. A similarity matrix was calculated from the 4000 vectors. From this similarity matrix the similarity tables for five diseases were analyzed in detail. These five diseases related to 56 direct neighbors. All relations were validated by medical literature. No incorrect disease–disease associations were found in the analysis. Additionally, the neighbor relations were visualized by a graph. Graph visualization extends the similarity relationships over and above the first layer of similar diseases. Also for these indirect connections evidence was found in the medicinal literature. The results demonstrated that diseases can be related by co-occurrences of genes. Second–order co–occurrence for disease–disease associations is a powerful tool. This approach can be used in drug discovery and medicinal sciences. Further analysis will show the validity of disease associations that connected at lower thresholds. The shared gene sets between two connected diseases are also of high interest and will be the subject of further studies.

References

[1] D. A. Bennett, J. A. Schneider, J. L. Bienias, D. A. Evans, and RS Wilson. "Mild cognitive impairment is related to Alzheimer disease pathology and cerebral infarctions"; Neurology, 64, 5, 834-841, 2005.

[2] R. Kalaria. "Similarities between Alzheimer's disease and vascular dementia"; J Neurol Sci, 203-204, 29-34, Nov 15, 2002.

[3] John Stuart Garrow. "Obesity and related diseases". (Churchill Livingstone, 1988. 1988)

[4] A. Hamosh, A. F. Scott, J. Amberger, C. Bocchini, D. Valle, and V. A. McKusick. "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders"; Nucleic Acids Res, 30, 1, 52-55, Jan 1, 2002.

[5] K. I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A. L. Barabasi. "The human disease network"; Proc Natl Acad Sci USA, 104, 21, 8685-8690, May 22, 2007.

[6] S. Suthram, J. T. Dudley, A. P. Chiang, R. Chen, T. J. Hastie, and A. J. Butte. "Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets"; PLoS Comput Biol, 6, 2, e1000662, Feb, 2010.

[7] M. A. van Driel, and H. G. Brunner. "Bioinformatics methods for identifying candidate disease genes"; Hum Genomics, 2, 6, 429-432, Jun, 2006.

[8] Y. I. Liu, P. H. Wise, and A. J. Butte. "The "etiome": identification and clustering of human disease etiological factors"; BMC Bioinformatics, 10 Suppl 2, S14, 2009.

[9] Hinrich Schütze. "Automatic word sense discrimination"; Computational linguistics, 24, 1, 97-123, 1998.

[10] Stanley Milgram. "The small world problem"; Psychology today, 2, 1, 60-67, 1967.

[11] D. Botstein, and N. Risch. "Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease"; Nat Genet, 33 Suppl, 228-237, Mar, 2003.

[12] http://www.ncbi.nlm.nih.gov/pmc/, accessed 03/01/2014 2014.

[13] http://www.nlm.nih.gov/pubs/factsheets/medline.html, accessed Feb 8 2014.

[14] R. M. Piro, and F. Di Cunto. "Computational approaches to disease-gene prediction: rationale, classification and successes"; FEBS Journal, 279, 5, 678-696, Mar, 2012.

[15] N. Rappaport, N. Nativ, G. Stelzer, M. Twik, Y. Guan-Golan, T. I. Stein, I. Bahir, F. Belinky, C. P. Morrey, M. Safran, and D. Lancet. "MalaCards: an integrated compendium for diseases and their annotation"; Database (Oxford), 2013, bat018, 2013.

[16] K. Peng, W. Xu, J. Zheng, K. Huang, H. Wang, J. Tong, Z. Lin, J. Liu, W. Cheng, D. Fu, P. Du, W. A. Kibbe, S. M. Lin, and T. Xia. "The Disease and Gene Annotations (DGA): an annotation resource for human disease"; Nucleic Acids Res., 41, Database issue, D553-560, Jan, 2013.

[17] Sune Pletscher-Frankild, Albert Pallejà, Kalliopi Tsafou, Janos X. Binder, and Lars Juhl Jensen. "DISEASES: Text mining and data integration of disease–gene associations"; Methods, 12, 1046, 2015.

[18] http://www.hugenavigator.net/HuGENavigator/startPagePedia.do, accessed 19.09.2014 2014.

[19] http://www.genenames.org, accessed Jul. 2015 2015.

[20] D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova. "Entrez Gene: gene-centered information at NCBI"; Nucleic Acids Res, 33, Database issue, D54-58, Jan. 2004, 2005.

[21] http://www.ncbi.nlm.nih.gov/gene, accessed Feb 10 2011.

[22] D. L. Wheeler, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, L. Y. Geer, W. Helmberg, Y. Kapustin, D. L. Kenton, O. Khovayko, D. J. Lipman, T. L. Madden, D. R. Maglott, J. Ostell, K. D. Pruitt, G. D. Schuler, L. M. Schriml, E. Sequeira, S. T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, T. O. Suzek, R. Tatusov, T. A. Tatusova, L. Wagner, and E. Yaschenko. "Database resources of the National Center for Biotechnology Information"; Nucleic Acids Res., 34, Database issue, D173-180, Jan, 2006.

[23] https://opennlp.apache.org/

[24] https://github.com/KevinStern/software-and-algorithms/blob/master/src/main/java/blogspot/software_and_algorithms/stern_library/string/DamerauLevenshteinAlgorithm.java

[25] Michael McCandless, Erik Hatcher, and Otis Gospodnetic: 'FuzzyLikeThisQuery': 'Lucene in Action: Covers Apache Lucene 3.0' (Manning Publications Co., 2010), pp. 284

[26] Don R. Swanson, Neil R. Smalheiser, and Vetle I. Torvik. "Ranking indirect connections in literature-based discovery: The role of medical subject headings"; Journal of the American Society for Information Science and Technology 57, 11, 1427-1439, 2006.

[27] Andrzej Białecki, Robert Muir, and Grant Ingersoll: 'Apache Lucene 4', in Editor (Ed.)^(Eds.): 'Book Apache Lucene 4' (Department of Computer Science, University of Otago, Dunedin, New Zealand, 2012, edn.), pp.

[28] A. Burns, and S. Iliffe. "Alzheimer's disease"; BMJ, 338, b158, 2009.

[29] J. C. de la Torre, and T. Mussivand. "Can disturbed brain microcirculation cause Alzheimer's disease?"; Neurol Res, 15, 3, 146-153, Jun, 1993.

[30] L. S. Honig, M. X. Tang, S. Albert, R. Costa, J. Luchsinger, J. Manly, Y. Stern, and R. Mayeux. "Stroke and the risk of Alzheimer disease"; Arch Neurol, 60, 12, 1707-1712, Dec, 2003.

[31] I. Nomura, E. Goleva, M. D. Howell, Q. A. Hamid, P. Y. Ong, C. F. Hall, M. A. Darst, B. Gao, M. Boguniewicz, J. B. Travers, and D. Y. Leung. "Cytokine milieu of atopic dermatitis, as compared to psoriasis, skin prevents induction of innate immune response genes"; J Immunol, 171, 6, 3262-3269, Sep 15, 2003.

[32] A. J. Schuerwegh, L. S. De Clerck, L. De Schutter, C. H. Bridts, A. Verbruggen, and W. J. Stevens. "Flow cytometric detection of type 1 (IL-2, IFN-gamma) and type 2 (IL-4, IL-5) cytokines in T-helper and T-suppressor/cytotoxic cells in rheumatoid arthritis, allergic asthma and atopic dermatitis"; Cytokine, 11, 10, 783-788, Oct, 1999.

[33] J. Bruix, J. M. Barrera, X. Calvet, G. Ercilla, J. Costa, J. M. Sanchez-Tapias, M. Ventura, M. Vall, M. Bruguera, C. Bru, and et al. "Prevalence of antibodies to hepatitis C virus in Spanish patients with hepatocellular carcinoma and hepatic cirrhosis"; Lancet, 2, 8670, 1004-1006, Oct 28, 1989.

[34] C. T. Wai, J. K. Greenson, R. J. Fontana, J. D. Kalbfleisch, J. A. Marrero, H. S. Conjeevaram, and A. S. Lok. "A simple noninvasive index can predict both significant fibrosis and cirrhosis in patients with chronic hepatitis C"; Hepatology, 38, 2, 518-526, Aug, 2003.

[35] Z. D. Goodman, and K. G. Ishak. "Histopathology of hepatitis C virus infection"; Semin Liver Dis, 15, 1, 70-81, Feb, 1995.

[36] H. D. Intengan, and E. L. Schiffrin. "Vascular remodeling in hypertension: roles of apoptosis, inflammation, and fibrosis"; Hypertension, 38, 3 Pt 2, 581-587, Sep, 2001.

[37] J. M. Chapman, and F. J. Massey, Jr. "The Interrelationship of Serum Cholesterol, Hypertension, Body Weight, and Risk of Coronary Disease. Results of

the First Ten Years' Follow-up in the Los Angeles Heart Study"; J Chronic Dis, 17, 933-949, Oct, 1964.

[38] "Prevention of stroke by antihypertensive drug treatment in older persons with isolated systolic hypertension. Final results of the Systolic Hypertension in the Elderly Program (SHEP). SHEP Cooperative Research Group"; JAMA, 265, 24, 3255-3264, Jun 26, 1991.

[39] W. B. Kannel, A. L. Dannenberg, and R. D. Abbott. "Unrecognized myocardial infarction and hypertension: the Framingham Study"; Am Heart J, 109, 3 Pt 1, 581-585, Mar, 1985.