University of Memphis
University of Memphis Digital Commons

Electronic Theses and Dissertations

4-19-2011

Knowledge Discovery Through Large-Scale Literature-Mining of Biological Text-Data

Vida Abedi

Follow this and additional works at: https://digitalcommons.memphis.edu/etd

Recommended Citation
Abedi, Vida, "Knowledge Discovery Through Large-Scale Literature-Mining of Biological Text-Data" (2011). *Electronic Theses and Dissertations*. 217.
https://digitalcommons.memphis.edu/etd/217

This Thesis is brought to you for free and open access by University of Memphis Digital Commons. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of University of Memphis Digital Commons. For more information, please contact khggerty@memphis.edu.

To the University Council:

The thesis committee for Vida Abedi certifies that this is the final approved version of the following electronic thesis: “Knowledge discovery through large-scale literature mining of biological text-data.”

Mohammed Yeasin, Ph.D.
Major Professor

We have read this thesis and recommend its acceptance:

Ramin Homayouni, Ph.D.

Ebenezer O. George, Ph.D.

Robert Williams, Ph.D.

Accepted for the Graduate Council

Karen D. Weddle-West, Ph.D.
Vice Provost for Graduate Programs

KNOWLEDGE DISCOVERY THROUGH LARGE-SCALE LITERATURE-MINING OF BIOLOGICAL TEXT-DATA

by

Vida Abedi

A Thesis

Submitted in Partial Fulfillment of the

Requirement for the Degree of

Masters of Science

Major: Bioinformatics

The University of Memphis

May, 2011

ACKNOWLEDGEMENTS

I would firstly thank my supervisor, Dr. Mohammed Yeasin, for taking me on as his student. His valuable knowledge and guidance has provided me with an immense foundation on which to continue my PhD and my career in research. I would also like to thanks Dr. Ramin Homayouni for his guidance. The guest speakers that Dr. Homayouni has scheduled over the years on Friday mornings on campus were and are still now an important part of my education. Through Dr. Homayouni's supportive actions and guidance, he has been an essential ambassador to my successes at The University of Memphis.

Also many thanks are extended to my teachers here at the University of Memphis and also special thanks to my committee members Dr. R. Williams, and Dr. E.O. George, for their guidance and continuous support. Special thanks are also extended to Mrs. Becky Ward and Mrs. Tallulah Campbell for their continuous help throughout the past two years.

Secondly, I would like to thank the members of the CVPIA lab for their friendship; but I would like to express a special thanks to Fazle E. Faisal, Geoffrey West, Pratiksha Subedi and Trinity Ownes for their help and complete support during the past two years.

And lastly, I would like to extend my thanks to my dear husband and colleague Ramin Zand for encouraging and supporting me throughout the years in graduate school, and also for his continuous love and friendship.

ABSTRACT

Abedi, Vida. M.Sc. The University of Memphis. 05/2011. Knowledge discovery through large-scale literature-mining of biological text-data. Major Professor: Dr. Mohammed Yeasin.

The aim of this study is to develop scalable and efficient literature-mining framework for knowledge discovery in the field of medical and biological sciences. Using this scalable framework, customized disease-disease interaction network can be constructed. Features of the proposed network that differentiate it from existing networks are its 1) flexibility in the level of abstraction, 2) broad coverage, and 3) domain specificity. Empirical results for two neurological diseases have shown the utility of the proposed framework. The second goal of this study is to design and implement a bottom-up information retrieval approach to facilitate literature-mining in the specialized field of medical genetics. Experimental results are being corroborated at the moment.

Table of Contents

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

LIST OF ABBREVIATIONS

ALS	Amyotrophic lateral sclerosis
BC	Breast cancer
CVA	cerebrovascular accident
IDF	Inverse Document Frequency
GO	Gene Ontology
IS	Ischemic stroke
LSA	Latent Semantic Analysis
MeSH	Medical Subject Headings
NCBI	National Center for Biotechnology Information
NLM	National Library of Medicine
PD	Parkinson's disease
POLSA	Parameter Optimized Latent Semantic Analysis
SVD	singular value decomposition
TF	Term Frequency
TIA	Transient ischemic attack

CHAPTER I

Introduction

In the post-genomic era and with the advances in high-throughput technologies new doors have been opened to study and map genetic networks including human diseases [1,2,3]. Most of the new research directions focused on the genetic causes of diseases looking only at one or few diseases at once. It was only in 2007 that Goh K.I. et al., [3] took a conceptually different approach; they proposed an interaction between two diseases when both diseases were associated with a common gene. This idea led to construction of *diseasome*, disease-disease interaction network [3]. This higher level of abstraction, moving from one disease and many genes or gene-by-products to many diseases and their respective genes or gene-by-products, provided a new outlook to view genetic networks. However, the disease-disease interaction network proposed by Goh K.I. et al., [3] relied only on gene-disease interaction data. It was only recently that a combination of disease-gene information and protein-protein information [4] was used to enhance the quality of such network. These types of high level analysis provide insights into topological features and functional properties of the disease interaction network. However, diseases can also be connected through non-genetic features such as risk factors, side-effect of drugs or treatments, or signs and symptoms. Therefore constructing a disease network based genetic and non-genetic factors can be a valuable reference for clinicians and medical researchers.

Constructing a disease interaction network at a higher level of abstraction, taking into account concepts such as risk factors, treatment options or symptoms, requires a dataset that is wide in its scope and can provide genetic, epigenetic as well as non-genetic

information. Literature data is the only source of knowledge with a wide extent of information from different sources. In addition to that literature-mining of biomedical text data can be implemented in a scalable framework for information retrieval purposes. The main goal of this study is to develop a scalable and effective literature-mining framework to model disease-disease interaction network through usage of associated factors. Distinctive features of the proposed network are the followings: 1) flexibility in the level of abstraction, 2) broad coverage, and 3) domain specificity.

Flexibility in the level of abstraction: A separation between a disease and a symptom is sometimes ambiguous. For instance, hypertension could be considered a disease or a symptom. In the proposed framework, customized disease networks can be constructed based on the required level of abstraction. Diseases, or factors are nodes of the network and literature-derived evidences are edges connecting the nodes. *Broad coverage:* The semantic model to build the disease-disease interaction network is constructed from biological text data obtained from titles and abstracts of journal publications. Hence, the primary collection of the data could be from a variety of sources such as high-throughput experimental evidences, clinical data, or empirical observations. Therefore, the resulting network incorporates a wide range of information providing robustness to noise and to technical variations. *Domain specificity:* Biological text data tend to be noisy and one of the main reasons for this noise is the specificity of the terminology and the large number of abbreviations in the field. To overcome some of the difficulties, additional measures were taken to customize this framework for biological sciences. In particular a multi-gram dictionary was constructed using the controlled vocabulary from the Medical Subject Headings (MeSH).

Empirical results from two neurological diseases, Ischemic stroke and Parkinson's disease have corroborated the efficacy of the proposed framework. Additional work is in progress to expand the dataset and to assess the performance of this framework using a more comprehensive benchmarking scheme. In addition, implementation of web service application is currently in process.

The second goal of this study is to develop a large scale literature-mining tool to facilitate information retrieval for the purpose of literature search in the field of medical genetics. The bottom-up approach provides an unbiased system to extract valuable information from biological text data, which will maximize the effectiveness of reviewing process for medical researchers. The key idea behind these two goals is to develop a large scale literature-mining tool for the purpose of knowledge discovery and information retrieval in the field of medical and biological sciences.

Goal and Objectives

The goal of this study is to develop large scale literature-mining framework for the purpose of knowledge discovery and information retrieval in the field of medical and biological sciences.

Goal I: Develop a scalable and effective literature-mining framework to model disease-disease interaction network through usage of associated factors.

Objectives:

I. Develop parameter optimized latent semantic analysis framework for associated factors
II. Develop a scalable approach for selection of associated factors
III. Design and implement a database system to store the literature for efficient mining and access of information
IV. Develop an optimized dictionary system for biological and medical literature data
V. Assess the performance of the framework

Goal II: Develop large scale literature-mining to facilitate information retrieval for the purpose of literature search in the field of medical genetics.

Objectives:

I. Identify potential text-mining tools in bioinformatics
II. Design an integrated system to fuse the existing bioinformatics tools
III. Assess the efficacy and benefits of the integrated system

Background

In the past decade, science has witnessed an explosion in *OMICS*-based technologies, such as genomics, proteomics, and pharmaco-genomics. These advances have generated a wave of analytical tools to tackle the complex biological networks and high dimensional datasets. A network in biology represents a set of nodes representing biochemical or chemical entities and a set of edges representing interaction between those entities. For instance, in a protein-protein interaction network, nodes represent proteins and edges could be evidence for physical interaction between proteins.

Some researchers have looked at the robustness and dynamical properties [5] of these networks while others have focused on the general characteristics [3,5] of these complex systems in order to gain a deeper understanding. Together the results from studying of these complex networks furthered our knowledge to a different level of understanding. For instance, network analysis of disease-disease interaction (where nodes are disease and edges are common genes between diseases) showed that the vast majority of genes associated with diseases are non-essential and do not tend to encode hub proteins; in addition to that, genes contributing to a common disorder i) have tendency for their by-products to interact with each other through protein-protein interactions, ii) have tendency to be co-expressed, and iii) tend to share Gene Ontology terms [3].

As a result, scientists no longer attempt to study one gene or one gene-by-product at a time; rather they plan to study a family of genes or even group of genes that respond to a given perturbation (using microarrays) at one time. This trend also promotes more interdisciplinary collaborations to unravel disease mechanisms and shed light on some of the fundamental biological questions.

Even though networks have been built and complex systems have been reverse-engineered to predict systems response to perturbations, no significant attempt has been made to create a high level view of such complex biological systems using sources of information other than genetics and other high-throughput experimental data. In a landmark study Kwang-Il Goh et al. [3] explored whether human genetic disorders are related to each other by analyzing their corresponding disease genes and gene-by-products. Interaction between genetic disorders provides a higher level of abstraction when compared to gene regulatory networks, or protein-protein interactomes. In the latter case, genes or proteins are the nodes in the network for a particular disease; in the former case, diseases themselves are nodes in the network. In essence, disease-disease interaction data provides a higher level view of the biological system. This higher level of abstraction facilitates translational research and is instrumental in clinical studies. This type of analysis can provide a valuable reference for clinicians and medical researchers. Disease networks constructed now are mostly based on genetic and proteomic data; but disease network could also be constructed based on literature data to incorporate a wider range of factors such as side effects and risk factors.

In fact, generating disease-models based on literature data is a very natural and efficient way to better understand and summarize the current knowledge about different high-level systems. Identifying connecting elements between diseases can provide a systematic approach to identify missing links and potential associations while presenting new opportunities for collaborations and interdisciplinary research. A connection between two diseases can be formalized as risk factor, symptom, treatment option, side-effect of drugs, or any other disease as compared to only common disease-genes. Figure 1

represent a pictograph representation of a disease interaction network, where connections are modeled based on different kinds of associated factors.

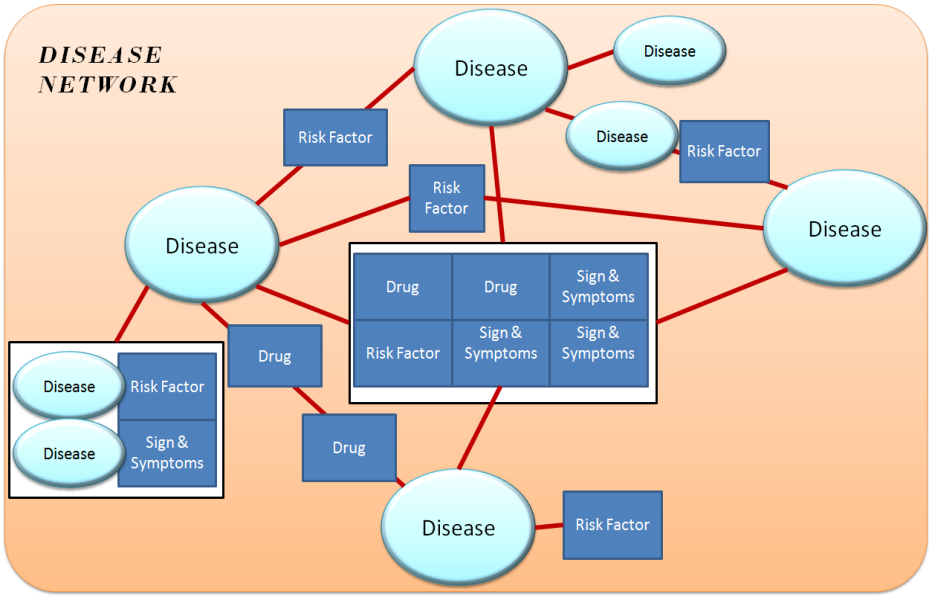

Figure 1. Pictograph representation of a disease interaction network using graph representation.
Light blue nodes represent diseases and dark blue nodes are associated factors. Note that because a disease can also be considered a risk factor to a second disease, therefore it is possible to have direct link between two or more diseases in the network.

Developing disease interaction network

In order to construct disease interaction networks, it is essential to identify factors associated with each disease independently. If factors are only genetic factors, then genomic data could be used; however, if factors are considered to cover a wider range then other type of data, such as text data, could be used. In essence, to build a high level view of the disease interaction network, it is essential to utilize factors at a higher level of granularity. For instance, instead of specific gene names it is more reasonable to know

that if the disease is associated with genetic factors. Therefore, "family history" could be used as a potential contributing factor. Similarly, instead of carefully analyzing the chemical structure of interacting compounds, it would be more appropriate to use groups of compounds such as "inorganic compounds", or "heterocyclic compounds". However, the system should be flexible enough to incorporate new factors when significant amount of information becomes available.

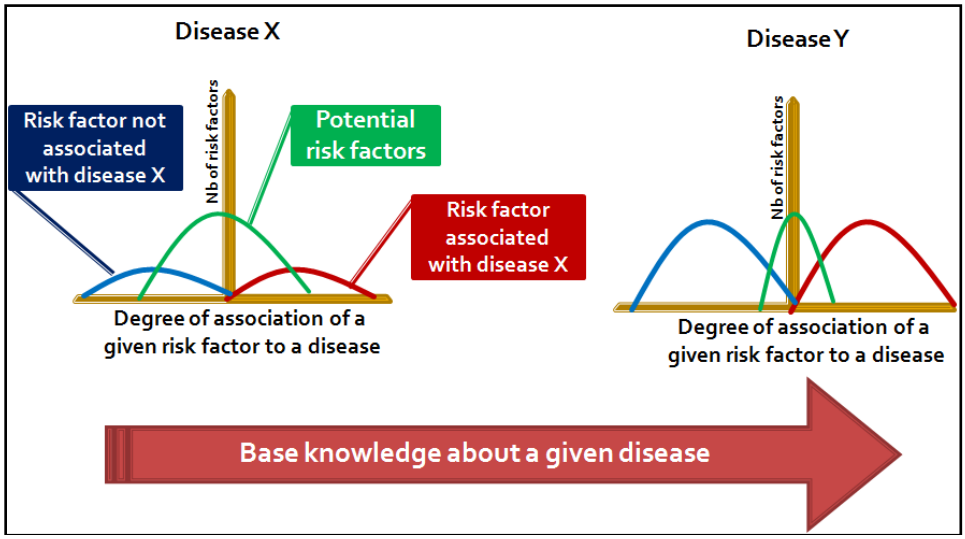

Figure 2. Model for the distribution of associated factors of a given disease.

The first step in the development of disease interaction network is the identification of factors that are associated to the disease. The next step is to evaluate the level of association of the factor to the disease. In order to accomplish the first task, literature-mining techniques could be used to mine the textual data in the PubMed[1] database and extract meaningful associations. If a disease is well documented, then it is

[1] http://www.ncbi.nlm.nih.gov/pubmed/

expected that a large body of literature will be available to corroborate the existence of such associations. If a disease is not well studied, then it is expected that most of the factors are weakly associated to the disease with only few factors displaying a high level of association (see figure 2, Disease X versus Disease Y). Hence the distribution of association level of factors (including risk factors) will be different in the two scenarios. In the first case (Disease X) the dominating distribution is that of potential factors; in the second case (Disease Y) the two dominating distributions are the factors that are associated and those that are not associated with the disease. Essentially, a large body of evidence is gathered to corroborate the existence of any association that may exist between the disease and the respective factor.

In order to evaluate the level of association of a factor to a disease, a more complex analysis is required. It is important to mention that factors are not independent of each other and if statistical tests are going to be used, it is imperative to account for that dependency whenever possible. Furthermore, if one could hypothesize that the distribution of associated factors follows a tri-modal distribution then it becomes simply more intuitive to measure the level of association for different factors.

The tri-modal distribution hypothesis proposed here is based on the fact that a disease mechanism is well understood if it is associated with a number of contributing factors at a relatively high score and vice-versa. Hence, as the figure 2 demonstrates, each distribution can be decomposed into three distributions, hence the tri-modal distribution model. The weights, or degree of association of a factor to a disease, are statistical measure of the association evaluated using literature-mining methodologies.

Knowledge discovery through literature-mining

The availability of huge textual resources provides the scientists with the chance to search for correlations or associations such as protein-protein interactions [6,7], and gene-disease associations [8,9]. However, biology and medicine are rich in terminology, for instance, in pathology reports and medical records, 12,000 medical abbreviations have been identified [10]. In addition, this large vocabulary is also dynamic and new terms emerge rapidly. For instance, the same object may have several names, or distinct objects can be identified with the same name; in the former case the names are synonyms and the latter case the objects are homonyms [11]. Consequently, literature-mining of biological and medical text becomes a very challenging task and the terms that suffer the most are gene and protein names [12,13]. However, even more challenging is the implementation of the information extraction, also known as deep parsing.

Deep parsing is built on formal mathematical models describing how text is generated in the human mind (i.e. formal grammar) [11]. The most popular formal grammars are deterministic or probabilistic context-free grammars [13]. Grammar-based information extraction techniques are computationally expensive because they require the evaluation and ranking of several alternative ways to generate the same sentence; it is therefore considerably slower but potentially much more precise [11]. An alternative to the grammar-based methods are vector-based methods such as Latent Semantic Analysis method. These alternative methods rely on bag-of-word concept, and have therefore reduced computational complexity. In addition to that, LSA technique has the added advantage of extracting direct and indirect association between entities.

In essence, since the traditional information retrieval framework, which relies on keyword-based approaches, cannot cope with the huge amount of information that is

being produced on a daily basis, scientists have focused on more sophisticated techniques such as text-mining [13] coupled with data-mining approaches. This shift has proved to be valuable in many instances. For example, titles from MEDLINE were used to make connections between disconnected arguments: 1) the connection between migraine and magnesium deficiency [14] which has been verified experimentally; 2) between indomethacin and Alzheimer's disease [14]; and finally 3) between *Curcuma longa* and retinal diseases, Crohn's disease and disorders related to the spinal cord [15]. Hypothesis generation in literature-mining relies on the fact that 'chance' connections can emerge to be meaningful [13].

Latent Semantic Analysis (LSA)

Latent Semantic Analysis (LSA) is a well known information retrieval technique which has been applied to many areas in bioinformatics. In LSA framework [16], a word-document matrix (also known as tf-idf matrix) is commonly used to represent a collection of text (corpus). LSA extracts relations based on second order co-occurrence from a corpus, and maps them on to K-dimensional vector space. The discrete indexed words are projected into an *eigen* space by applying singular value decomposition (SVD).

Arguably, LSA captures some semantic relations between various concepts based on their distance in the *eigen* space [17]. The most common similarity measure used to rank the vectors is the linear cosine similarity measure [17]. The three main steps of the LSA are outlined here and can be found in [16]:

I. **Creation of Term-Document matrix**: The text documents are represented using a bag-of-words model. This representation creates a term-document matrix in which the rows are the words (dictionary), the columns are the

documents, and the individual cell contains the frequency of the term appearance in the particular document. Term Frequency (TF) and Inverse Document Frequency (IDF) are used to create the TF-IDF matrix.

II. **Singular Value Decomposition (SVD)**: SVD or SparseSVD (approximation of SVD) is performed on the TF-IDF matrix and the *k* largest *eigen* vectors are retained. This k-dimensional matrix (encoding matrix) captures the relationship among words based on first and second order statistical co-occurrences.

III. **Information Retrieval**: Information related to a query can be retrieved by first folding-in the query into the LSA space and then performing a similarity measure between the documents and the query. Cosine similarity measure is usually used to rank and retrieve the documents.

Parameter Optimized Latent Semantic Analysis (POLSA)

Even though LSA has been has been applied to many areas in bioinformatics, the LSA models have been based on *adhoc* principles. In a recent work, a systematic study was performed on the parameters affecting the performance of LSA to develop a Parameter Optimized Latent Semantic Analysis (POLSA) [18]. The various parameters examined were corpus content, text preprocessing, sparseness of data vectors, feature selection, influence of the1st Eigen vector, and ranking of the encoding matrix. The optimized parameters should be chosen whenever possible.

Improving the semantic meaning of POLSA framework

Methods such as LSA have been successful in finding direct and indirect associations between various entities; however, these methods still use bag-of-words

concept; therefore, they do not take into account the order of words and hence the meaning of such words are often lost. Using multi keyword words would alleviate some of the problems of the bag-of-words model. In a multi keyword dictionary the word "vascular accident" (which is a synonym of "stroke") would be differentiated from "accident" which could also mean car accident in a different context.

However, it is challenging to generate such a dictionary. If all combinatorial words in the English dictionary are chosen, then the size of such dictionary would be considerably large even if one considers up-to three gram words. A larger dictionary implies also increased sparsity in the TF-IDF matrix. A possible solution is to construct the dictionary based on combinations of words that are biologically relevant for the case of biological text-mining. Identification of biologically relevant word combinations can be derived from biological ontology such as Gene Ontology[2] (GO) or Medical Subject Headings (MeSH)[3]. Using a multi-keyword dictionary could in principle improve the accuracy of the vector-based frameworks, such as the LSA, that rely only on bag-of-words models.

Benchmarks: knowledge-based systems

To evaluate any system, there is a need of ground truth or at least what is believed to be the current knowledge. There are a number of online resources that could be used to construct a ground truth in biology and medical sciences.

[2] www.geneontology.org

[3] http://www.ncbi.nlm.nih.gov/mesh

UpToDate: UpToDate[4] is a web accessible resource for clinicians. The system provides information for clinical decision making. It also answers clinical questions and summarizes sign, symptoms, side effect of drugs and treatments. UpToDate also provides a list of drugs or treatment options for a wide range of diseases. The system is not freely available to the public.

MedLink neurology: MedLink neurology[5] is similar to UpToDate in the sense it provides medical support to clinicians. However, MedLink neurology is especially targeted to neurologists; it has more in depth information about common and also rare neurological diseases. The system is not available to the public and requires registration. However, all the information is referenced and it is possible to extract original publications if needed. Medical summaries are usually done by few experts in the field, and only when sufficient amount of evidences become available, the experts will reference them in their clinical summaries.

PubMed: PubMed is a public database developed and maintained by the National Center for Biotechnology Information (NCBI[6]), and updated on a daily basis. Currently this database contains more 20 million citations for biomedical literature. Whenever an article has an abstract, that abstract is published through PubMed. One of the features of retrieving articles through PubMed is the fact that all entries are tagged using the Medical Subject Headings[7] controlled vocabulary. MeSH vocabularies are used to describe the subject of each journal article. MeSH contains approximately 26,000 terms and is

[4] http://www.uptodate.com/index

[5] http://www.medlink.com/medlinkcontent.asp

[6] http://www.ncbi.nlm.nih.gov/

[7] http://www.ncbi.nlm.nih.gov/mesh

updated annually to reflect changes in the medical field. MeSH terms are arranged hierarchically by subject categories and PubMed allows one to view this hierarchy and search the literature using the controlled vocabularies.

Constructing ground truth using PubMed MeSH controlled vocabulary is time consuming but feasible. However, the use of MeSH is extremely important because MeSH vocabulary ensures that articles are uniformly indexed by subject, whatever the author's words.

Bioinformatics tools

i. Text-mining methods in bioinformatics

Knowledge discovery is moving away from hypothesis-driven to data-driven approaches [19,20]; it is indeed increasingly difficult to sustain the deductive scientific method because these methods are not scalable for today's growing body of knowledge [21]. In fact, advances in technology and bioinformatics tools have significantly changed how we design and perform experiments and analyze the results. In the past decade, bioinformatics tools have also explored biological text-mining and today they are able to provide a faster approach to scan the literature by various means. One of the goals of this study is to perform an unbiased literature search using such tools as a first filtering step. It is clear that the expert knowledge will not be replaced by these bioinformatics tools; however, by using a text-mining approach, the researcher will only be required to read the relevant information and save a significant amount of time.

GeneIndexer: GeneIndexer[8] is an unsupervised approach that provides the user with a list of genes ranked for a given phenotype or disease. This tool is based on a

[8] http://www.computablegenomix.com/

Latent Semantic Analysis (LSA) approach, presenting thus direct and indirect associations between genes and phenotypes. The user may or may not enter a list of genes, which will be ranked based on the query phenotype. If a list of genes is not provided, the system will use all the known genes in human or mouse. This tool has great potential in microarray studies but also as we are using it here as a first step toward an unbiased literature review.

PubMatrix: PubMatrix[9] is a simple, yet powerful text-mining tool. Given two lists of terms (with a maximum of 100 for list 1 and 10 for list 2) this tool will count the number of co-occurrences and will generate a matrix. Note that the queries are sent to the PubMed exactly as they are listed (Gene AND Term). A link to the publications where this co-occurrence was observed is provided to the user as a link to PubMed. The advantage of using this high-throughput literature search is to speed up the search and also to provide an overview of the genes in the set and their relationship with respect to the keywords. The main weakness of this methodology is the selection of keywords. User-defined keyword introduces bias to the system and works against the data-driven knowledge discovery approach. On the other hand, integrating prior knowledge may be a good way to control noise and provide robustness to the system. Incorporation of prior knowledge has been of great value in many large scale integration studies over the past decade [22,23,24]. In a recent study, Steele *et al.*, [24] reported that massive incorporation of prior knowledge, from literature, can significantly improve reverse-engineering of biological systems and enhance the modeling process. The intelligent mining of biomedical literature, for decision making, hypothesis generation and knowledge discovery constitutes a formidable research challenge.

[9] http://pubmatrix.grc.nia.nih.gov/

Chilibot: Chilibot[10] is software designed to read the titles and abstracts from PubMed database in order to identify relationships between genes, proteins or any other entity. The results present key information that links the two entities. For instance, a sentence that contains a gene name and a disease name will be returned to the user with the reference to the publication. In addition to that, sentences are organized into different types of relationships based on linguistic analysis of the text. This tool is very useful to the researcher at the final stage of the analysis to speed up the review process of the PubMed articles.

ii. Enrichment analysis tools

Enrichment analysis tools[11] are comprehensive global analysis toolboxes. They systematically translate a set of gene list into functional profiles. Intuitively, enrichment analysis tools have been mostly used for study of microarray data. An array of such tools have been implemented and improved over the years [review in ref. 25]. One example is the collection of Onto-Tools. The latter is web-accessible tool which is based on Gene Ontology (GO) [26]. These tools attempt to translate a list of differentially regulated genes into an understandable set of biological phenomena. Most of these tools, including Onto-Express, have one common feature: they share the same infrastructure in the sense that they all use Gene Ontology (GO) to group genes into categories. The main difference among these tools are the following: 1) statistical model; 2) visualization capabilities; 3) level of abstraction (some of the tools only present the lowest level of GO categories, when other are dynamic or present a tree view representation); 4) supported microarrays; 5) supported input IDs; and 6) user interface.

[10] http://www.chilibot.net/

[11] http://vortex.cs.wayne.edu/projects.htm#Onto-Express and http://biit.cs.ut.ee/gprofiler/

In order to make best use of these tools and to be able to obtain meaningful insight from them, it is imperative to understand the structure of Gene Ontology that all these tools are based upon (review in ref. 27). Gene Ontology has Directed Acyclic Graph structure within each of the three main domains (biological process, molecular function, and cellular component). GO structure is not uniform in its information content; therefore using a simple level-based analysis would not be adequate. Different approaches have been taken to correct for this problem [28,29]. A second major issue, which is also ignored often, is the fact that over 95% of the GO annotations are computationally derived and are not manually curated (these are indicated as IEP evidence code, or inferred from expression pattern) [27]; these automations often time use high-throughput experiments (co-expression data can be used for this purpose). This fact causes a circular problem: for instance consider using gene expression data to predict gene function while including annotations that are derived from gene expression datasets. In order to avoid this problem, one should consider filtering annotations based on the evidence code [27]. This becomes particularly important when one wishes to experimentally validate a hypothesis generated by these models, especially when the required experiment is expensive and time consuming (for instance generating a knock out mouse). Finally, a tool such as Onto-Express can automatically translate a list of genes into functional profiles using information derived from gene ontology.

CHAPTER II

Proposed Methodology

To achieve the main goals of this study, which is to develop a scalable and effective literature-mining framework to model disease-disease interaction network through usage of associated factors, Parameter Optimized LSA (POLSA) was used. In the first section (section A), the methodologies as well as the parameters of the system are described in detail. These include the dataset, parameters of POLSA, disease-disease interaction network model, evaluation, and web-service design and implementation. In order to achieve the second goal of this study, to develop a large scale literature-mining tool to facilitate information retrieval for the purpose of literature search in the field of medical genetics, integration of a number of text-mining tools was done. Additionally, in the second section of this chapter (section B), a detailed description of the tool integration strategy for the development of large scale literature-mining is presented.

Finally, the key idea behind these two goals is to develop a large scale literature-mining tool for the purpose of knowledge discovery and information retrieval in medical and biological sciences. The ultimate aim is to examine if literature-mining techniques or a fusion of these techniques are capable of providing a reliable source of evidences beyond just a source of prior information.

Literature-mining framework to model disease-disease interaction network

In order to develop a scalable and effective literature-mining framework to model disease-disease interaction network, through usage of associated factors, five different objectives have to be fulfilled. These objectives are detailed in the introduction section but can be summarized as follows: 1) *Framework design:* develop parameter optimized LSA framework for associated factors; 2) *Selection step:* develop a scalable approach for selection of associated factors; 3) *Database system:* design and implement a database system to store the literature for efficient mining and access of information; 4) *Optimization of the system:* develop an optimized dictionary system for biological and medical literature data; 5) *Assessment of the framework:* assess the performance of the framework by comparing the output of the system with the ground truth for two disease.

Framework design: disease modeling using tri-modal distribution scheme

A set of associated factors, including risk factors, side effects of drugs or treatments, are selected to generate a set of documents, where each factor is represented by one document. Using the LSA technique or the improved POLSA technique, it is possible to rank the documents based on their association to a given query, which in this case is a disease such as “stroke”. The highly ranked documents with respect to a disease are considered factors associated with that disease.

Furthermore, the distribution of a set of associated factor with respect to a disease can be modeled as a tri-modal distribution: sum of three normal distributions. This is due to the fact that, some factors are known to be associated to the disease and these have

high scores; similarly some factors are known to not be associated to the diseases and these have negative scores; in addition, some factors may or may not be associated to the disease and these have low similarity scores. Using a tri-modal model it is possible to separate the three distributions.

To obtain the three separate distributions from the ranked documents, Matlab's curve fitting toolbox is used. The input to this toolbox is a one-dimensional distribution and the resulting output is the set of parameters of three normal distributions. In fact, the cosine scores, representing the association of a disease with all the factors, are used to generate a histogram. This histogram is imported in the curve fitting toolbox where a general Gaussian model is used to generate the tri-modal distribution. Goodness of fit is measured using an R-square score. R-square is expected to be higher for well studied diseases such as Ischemic Stroke as compared to Parkinson's disease because the well studied diseases tend to have a better separation between the associated and non-associated factors.

The parameters of the POLSA model are detailed in the following section: To increase the performance of the LSA, two parameters were pre-set based on a previous work done in our lab: 97% criteria for semantic space reduction and exclusion of the 1st *eigen* value component in the projected space.

Dataset: *The set of 97 factors*. PubMed database was used to download titles and abstracts of articles from the past twenty years using risk factors as queries to PubMed. Based on the results obtained by the first analysis, it was clear that only twenty years of literature may not be enough to capture our existing knowledge. Therefore, the time frame was expanded to fifty years of publications for the set of 276 factors. *The set of 276*

factors derived from MeSH. PubMed database was used to download titles and abstracts of articles from the past fifty years following an integrated strategy (see figure 12) to generate the queries that were sent to PubMed. In order to implement this framework, NCBI Eutilies, in particular Esearch and Efetch were utilized. Detailed about the selection process of these factors is presented in the following section.

Weighting scheme: Term Frequency-Inverse Document Frequency (TF-IDF) is used as a weighting scheme in the LSA model. No pre-processing is performed for the initial analysis where the set of risk factors is limited to 97 factors. Stop word removal is used as a pre-processing step when the set of risk factors is 276. No stemming is performed in the two experimental set up.

Query to the system: Any Medical Subject Heading could be used as query to the system. This will allow users to search and construct content-rich relationship network between diseases, risk factors, chemical compounds and more. There are over 38,000 possible queries to the system based on the multi-gram based dictionary (as described in section 4) and over 19,000 keywords based on simple MeSH dictionary.

Similarity measure: Cosine similarity measure is used as the similarity measure in this study. The latter is defined as the score given to a gene-document d_j with respect to a query q by the measure of cosine of the angle between the corresponding vectors in the latent semantic space. This measure is calculated as follows:

$$\cos\theta_j = \frac{d_j^T (U_k^T q)}{||d_j||^2 ||q||^2}; \quad \text{j=1, 2,..,n.}$$

Selection Step: risk factor identification using expert knowledge and MeSH hierarchy

A set of *97 associated factors* (see table 1 in the supplemental material) were identified through a systematic review of medical articles and case reports; this list was the result of a careful revision by an expert in the medical field.

Table 1: Potential risk factors / contributing factors selected by medical expert

Tobacco smoking, alcohol consumption, health education and health promotion, pregnancy outcome, heterosexual, homosexual, maternal influenza, chemical agents, wood dust (exposure), silica dust (exposure), postmenopause, lifestyle intervention, diet nutrition, stress, age gender, oral contraceptive (OC), depression, breast-feeding, head trauma, abdominal adiposity, bisphenol-A (PBA), diethylstibestrol (DES), estradiol (E2), outdoor worker, indoor worker, fracture, bone mineral density (BMD), body mass index (BMI), Hormone, exposure polycyclic aromatic hydrocarbons, hepatitis B virus, Aflatoxin, night shift work, cholesterol level, family history, morning cortisol level, Mood, hypothyroidism, hyperthyroidism, insomnia, retrovirus, enterovirus, volume of cerebrum, volume of hippocampus, volume of lateral ventricle, caffeine, motor activity assessment, Cannabis, cocaine, viral infection, bacterial infection, addiction, air pollutants, volatile organic compounds, Pesticide, calcium deficiency or calcium overdose, phosphorus deficiency or phosphorus overdose, magnesium deficiency or magnesium overdose, sodium deficiency or sodium overdose, potassium deficiency or potassium overdose, sulphur deficiency or sulphur overdose, chloride deficiency or chloride overdose, chromium deficiency or chromium overdose, copper deficiency or copper overdose, fluoride deficiency or fluoride overdose, iodine deficiency or iodine overdose, iron deficiency or iron overdose, manganese deficiency or manganese overdose, molybdenum deficiency or molybdenum overdose, selenium deficiency or selenium overdose, zinc deficiency or zinc overdose, vitamin A or Retinol, vitamin B1 or Thiamine, vitamin B2 or Riboflavin, vitamin B3 or Niacin, vitamin B5 or Pantothenic acid, vitamin B6 or Pyridoxine, vitamin B7 or Biotin, vitamin B9 or Folic acid, vitamin B12 or Cyanocobalamin, vitamin C or Ascorbic acid, vitamin D or Calciferol, vitamin E or Tocopherol, vitamin K or Phylloquinone, Asthma, Autism, Schizophrenia, HIV, immunological disorder, Bipolar, Hypertension, Osteoporosis, coronary heart disease (CHD), Diabetes, Allergy, Herpes, leukemia, Breast cancer, Lymphoma.

Subsequently, a set of *276 MeSH* were selected to construct a list of potential risk factors and other associated factors (see table 6). The advantages of using MeSH-based factors are: 1) the terms are controlled vocabulary, constituting more than 25,000 subject headings in an eleven-level hierarchy and 83 subheadings in the 2010 edition; 2) using an

up-to-dated hierarchical subject heading provides common language for communication and information exchange in a hierarchical manner. The list of 276 factors can be expanded to eventually include all the headings and sub-headings. In order to do that, it is essential to overcome limitations of the existing text mining tools as well as statistical tools at hand. It is essential to address the systematic bias that exists for factors with limited number of documents; this concept is further explored in the discussion section. The list of 276 factors was manually revised by a medical expert. The MeSH headings for the construction of the corpus utilized in this study constitute two levels of hierarchy, giving the user different degree of granularity to analyze the results.

Using the set of 97 factors was important to assess the system before moving to a large scale study in which the large amount of literature required design and construction of a database. However, the set of 97 factors included generic (such as "stress") and also specific factors (such as "maternal influenza") which caused an imbalance in the dataset. Data imbalance can be a significant source of bias in the data. Therefore, by choosing the factors from the MeSH hierarchy, this problem could be partially alleviated. In addition to that, current work is going in our lab to implement a local POLSA model and alleviate this problem systematically.

Database system: design and implementation of a database system to store and mine the literature

The dataset for the 276 factors is downloaded from PubMed and stored in MySQL database. The databse construction is based on the following design (see Figure 3). The advantage of using a database to store the data has one key advantage. Since the relationship between abstract and factors is many-to-one, by saving the data into a database each abstract will only be downloaded once, saving significant amount of storage.

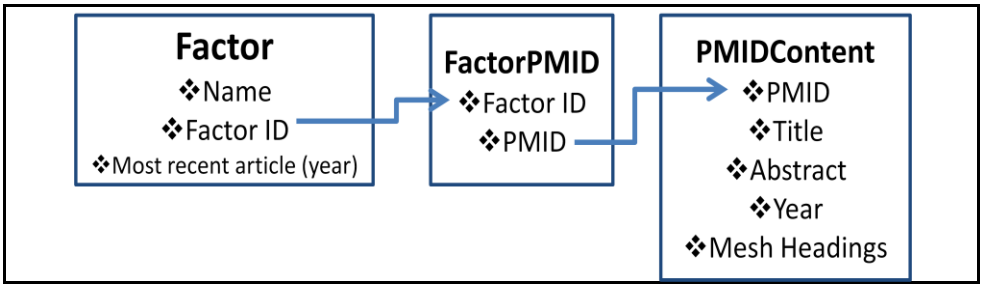

Figure 3: MySQL database design.
Three tables are used to construct the database for the MeSH-based factors. **Factor** table contains information regarding the 276 MeSH factors, "most recent article (year)" is used to update the entry in the database; **FactorPMID** contains information need to link the factor to PubMed abstracts using PMIDs (unique identifies of PubMed abstracts); **PMIDContent** contains information about each abstract. In PMIDContent Mesh Headings are separated by ";".

Optimization of the system: development of an optimized dictionary system for biological and medical literature data

For the set of 97 factors: A simple dictionary is constructed using MeSH. All the headings are reduced to single words. For example "Reproductive and Urinary Physiological Phenomena" is reduced to five words in the dictionary (Reproductive, and,

Urinary, Physiological, Phenomena); duplicates as well as stop words such as "and" are removed from the dictionary. The size of this dictionary is a slightly over 19,000.

For the set of 276 factors derived from MeSH: A refined multi-gram dictionary is constructed to improve the quality of LSA model (see figure 13). The main advantage of using a multi-gram dictionary with LSA is to overcome some of the limitations of the bag-of-word model. LSA model does not take into account the order of the words and therefore semantic meaning is sometimes lost, especially when two or three words have a completely different meaning when expressed in a specific order. For instance, "vascular accident" is a synonym for "ischemic stroke"; however, if vascular and accident are considered independently then the true meaning is lost. The size of the multi-gram dictionary is a little over 38,000 words.

Assessment of the framework: assess the performance of the framework

Ground Truth for evaluation process: *for the set of 97 factors.* MedLink Neurology is one of the most comprehensive resources, used by neurologists, for neurological diseases. MedLink Neurology is used as a ground truth for two of the diseases studied here: Parkinson disease and Ischemic stroke.

For the set of 276 factors derived from MeSH. A systematic literature review, using PubMed, is done to evaluate the existence of any association between a given disease and the 276 factors. Ground truth is constructed based on the information from PubMed using MeSH search with etiology OR complications. The information is meticulously read by expert and the ground truth is built accordingly. An example of PubMed search for breast cancer (disease) and Age Group (associated factor) is as follows:

(("Breast Neoplasms/complications"[Mesh] OR "Breast Neoplasms/etiology"[Mesh]) AND ("Age Groups"[Mesh]))

Based on the literature, each associated factor is categorized as "Established Factor", "Possible Factor" or "Unknown". The ground truth is built for two diseases: breast cancer and Ischemic Stroke. Breast cancer is considered a systemic disease and therefore associated with a larger set of factors. Ischemic stroke on the other hand is a relatively well studied disease with known risk factors and complications. The latter would provide us with a higher quality ground truth. It is imperative to mention that our understanding of causes and side effects of diseases are not completely accurate and complete, therefore it is impossible to expect the ground truth to be perfect. Improvement should be made on regular basis when new information becomes available.

Validation strategy: *for the set of 97 factors.* The results are validated with the ground truth when the ground truth is based on the information from MedLink Neurology. Distribution of the similarity measure obtained by the LSA model is plotted and a gradient level of association is determined based on the shape of the distribution. For instance the peak representing at cosine similarity of zero represents the factors with potential association and the peak in the negative similarity measure is representative of no-association. Similarly the peak with a mean in the higher cosine measure represents factors associated with the disease at high, medium and low degree. This method is not statistically derived, yet it provides a first overview of the results. Since the number of

factors was limited, a statistical model or a machine learning approach would have been unrealistic, or would have only identified few factors as being slightly significant.

However, using a disease model (by tri-modal distribution) allows a better identification of the three sets of factors: unknown association, potential association and established association. The latter method is empirical and provides an intuitive approach to evaluate the results, especially when the number of factors is limited. A more sophisticated methodology is proposed when the number of factors is larger. Figure 4 demonstrate an example of such analysis.

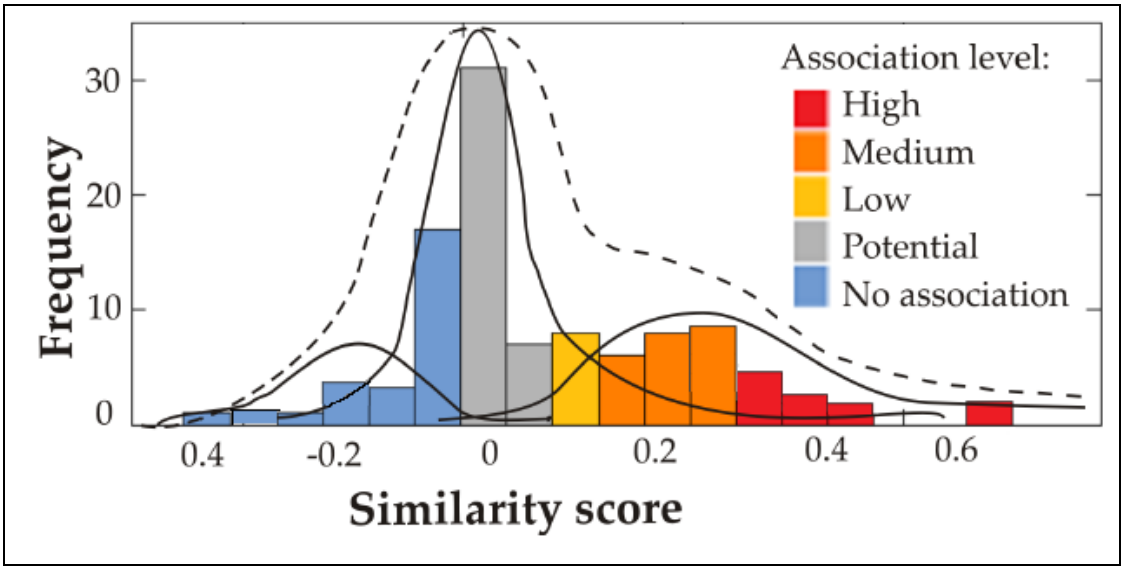

Figure 4: Illustrative example of the validation strategy for a small set of risk factor. The histogram represents number of factors at each similarity level when the query is one of the keywords in the dictionary (ex: "stroke"). Cosine similarity measure is used in this study, the latter ranges from -1 to +1; however similarity measure is more concentrated around zero especially when the dataset is large. When the set of risk factors is limited, a disease-model strategy is used to group the risk factors into three main categories: orange for high to low association level between the disease and the associated factor, gray for potential association and blue for no association detected. If a tri-model distribution is assumed for a relatively well studied disease, then this model can be used to group the risk factors.

For the set of 276 factors derived from MeSH. Validation Strategy (work in progress) – Comparing ground truth to the results obtained by the system using three different strategies: 1) empirical method: Ground truth is compared to the results obtained by the system. A similarity score above 0.3 is considered strong; a similarity score above 0 is considered possible association and a negative similarity score is indicative of unknown association. 2) Method based on experimental observation: Ground truth is compared to the results obtained by the system. Similarity scores are sorted and divided into four bins. The highest score will be compared to the established factors, the second highest scores are compared to the possible factors and the remaining two sets are compared to the unknown factors. This subdivision is based on the fact that given the two ground truths extracted from PubMed, approximately 50% of the factors were marked as unknown and the remaining were roughly divided into two sets, one being established factors and one being possible factors (see the results section). 3) Unsupervised machine learning method: Ground truth is compared to the results obtained by the system. Similarity scores are clustered into three bins using K-mean clustering algorithm (with K=3). The cluster with highest values is compared to the cluster with established factors, the cluster with average values is compared to the factors with possible associations; similarly, the cluster with the lowest scores is compared to factors with unknown associations.

Develop large scale literature-mining to facilitate information retrieval for the purpose of literature search in the field of medical genetics

In order to develop a large scale, scalable and effective literature-mining framework to facilitate information retrieval for the purpose of literature search in the field of medical genetics, three different objectives are outlined in the introduction section of this report and also summarized here. 1) *Tool identification:* Identification of a number of text-mining tools in bioinformatics; 2) *System design:* Design an integrated system to fuse the existing bioinformatics tools; 3) *Assessment of the framework:* Assess the integrated framework.

Tool identification: Identification of a number of text-mining tools in bioinformatics

In conventional top-down approach, manual literature review is the only tool used to search the literature in order to identify genes associated with a given disease state. In bottom-up approach, all the genes have initially the same probability of being among the associated genes. A number of tools have been selected for the bottom-up approach of searching the literature. These tools are presented in the order that they will be integrated starting with GeneIndexer.

GeneIndexer: Input to GeneIndexer is the set of all human genes and the keyword used is "stroke". The top 800 genes (cosine score > 0.1) have been selected for further analysis.

PubMatrix: Input to PubMatix is top 800 genes, selected form GeneIndexer, and seven keywords: stroke, cerebral ischemia, brain infarction, CVA, cerebrovascular accident, TIA, and transient ischemic attack. The average score is calculated for each gene and used a measure to select genes important for further analysis. Since PubMatrix

allows a maximum of one-hundred terms and ten modifiers for each run, eight different requests were made and results were combined for the analysis.

Literature search, Chilibot and Enrichment analysis: Literature review of the top selected genes is performed by reading articles, abstracts, case reports as well as using *Chilibot* to direct the literature search. This process was performed partly by the author but also by an expert in the field.

In addition, enrichment analysis is performed on the set of genes to identify major biological processes and molecular functions that are shared among the genes. Enrichment analysis was performed using *Onto-Express* toolbox. This technique, as also described in introduction section, relies on the Gene Ontology information of the genes.

System design: Design an integrated system to fuse the existing bioinformatics tools

Figure 5 outlines the design of the integrated system. At the first stage, GeneIndexer is used; the latter ranks the genes and approximately the top 5% can be used for further investigation (top N genes). The identified genes can then be forwarded to a second text-mining tool such as PubMatrix along with a set of related keywords or synonyms. Using a cut-off, a smaller set of genes can be selected for further investigation by the researcher. The top M genes are limited and therefore possible to analyze them manually or with the help of additional tools such as Chilibot or enrichment analysis tools as also described in the introduction.

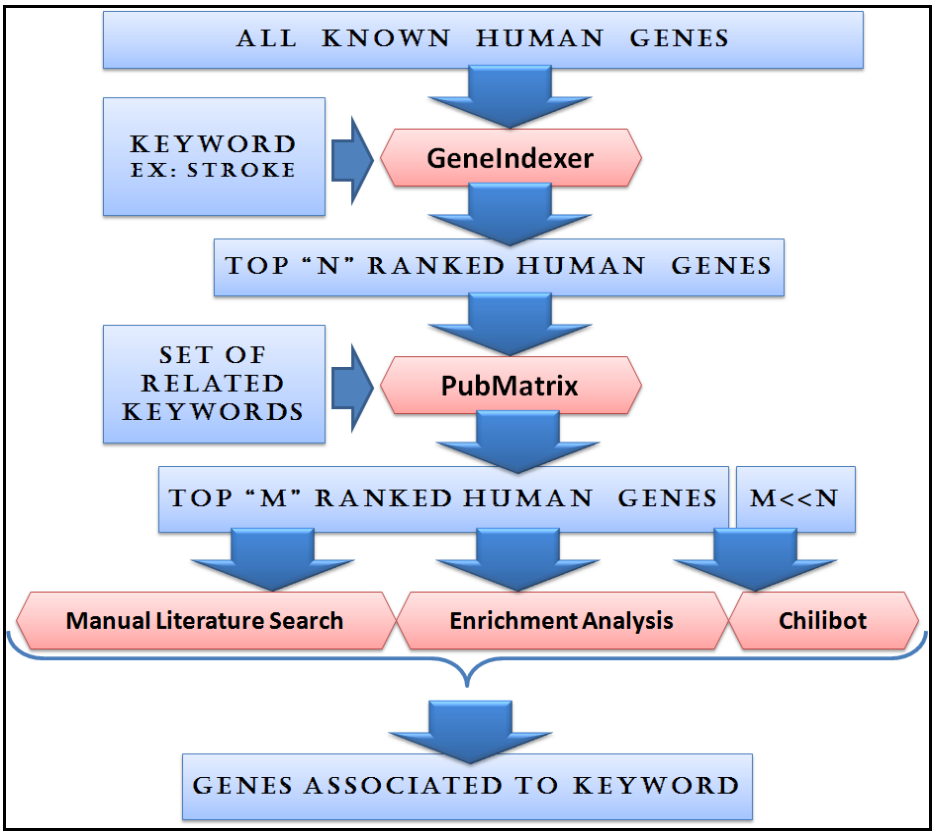

Figure 5: A flow chart describing the steps taken to perform a bottom-up literature mining of ischemic stroke.

Assessment of the framework: Assess the integrated framework:

A literature review of the genetics of stroke is in progress. The presented framework was used to generate 90 genes that could potentially be directly or indirectly associated with stroke. A manual review of the gene list is in progress by two neurologists at UTHSC. This work will be presented at the AAN2011 conference in Hawaii this April.

CHAPTER III

RESULTS

Literature-mining framework to model disease-disease interaction network

Histogram representation of factors at different level of association with respect to a given disease was used to generate three Gaussian distributions. Figure 6 represents the data for Ischemic stroke while figure 7 represents Parkinson's disease. Risk factors and treatment options, as well as side-effects of therapies for Ischemic stroke are better understood than those of Parkinson's disease. The previous is reflected in three distributions: 1) the tri-modal distribution for Parkinson's disease has greater variance as compared to Ischemic stroke; 2) mean value of the rightmost distribution is higher in the case of Ischemic stroke. The higher variance in the case of Parkinson's disease is an indication of greater uncertainty in terms of the level of association of various factors to the disease.

Additionally, the higher mean value for the right-most curve indicates a higher confidence; this is scored by cosine similarity measure on the factors identified to be linked to the disease.

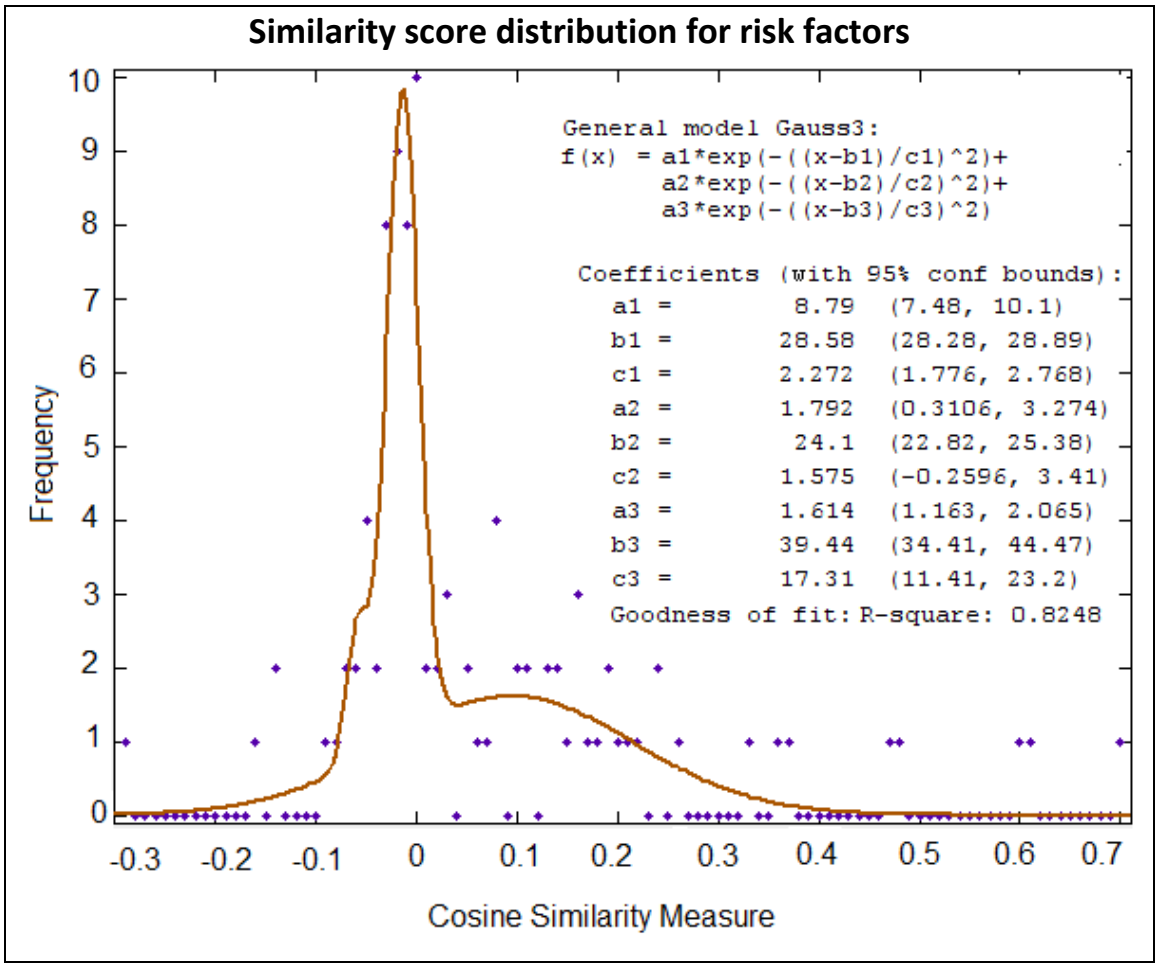

Figure 6: Distribution of factors for Ischemic stroke.
Data and the tri-modal distribution fitted to the data using Matlab's curve fitting toolbox. The frequency represents number of factors at each similarity level when the query is "stroke".

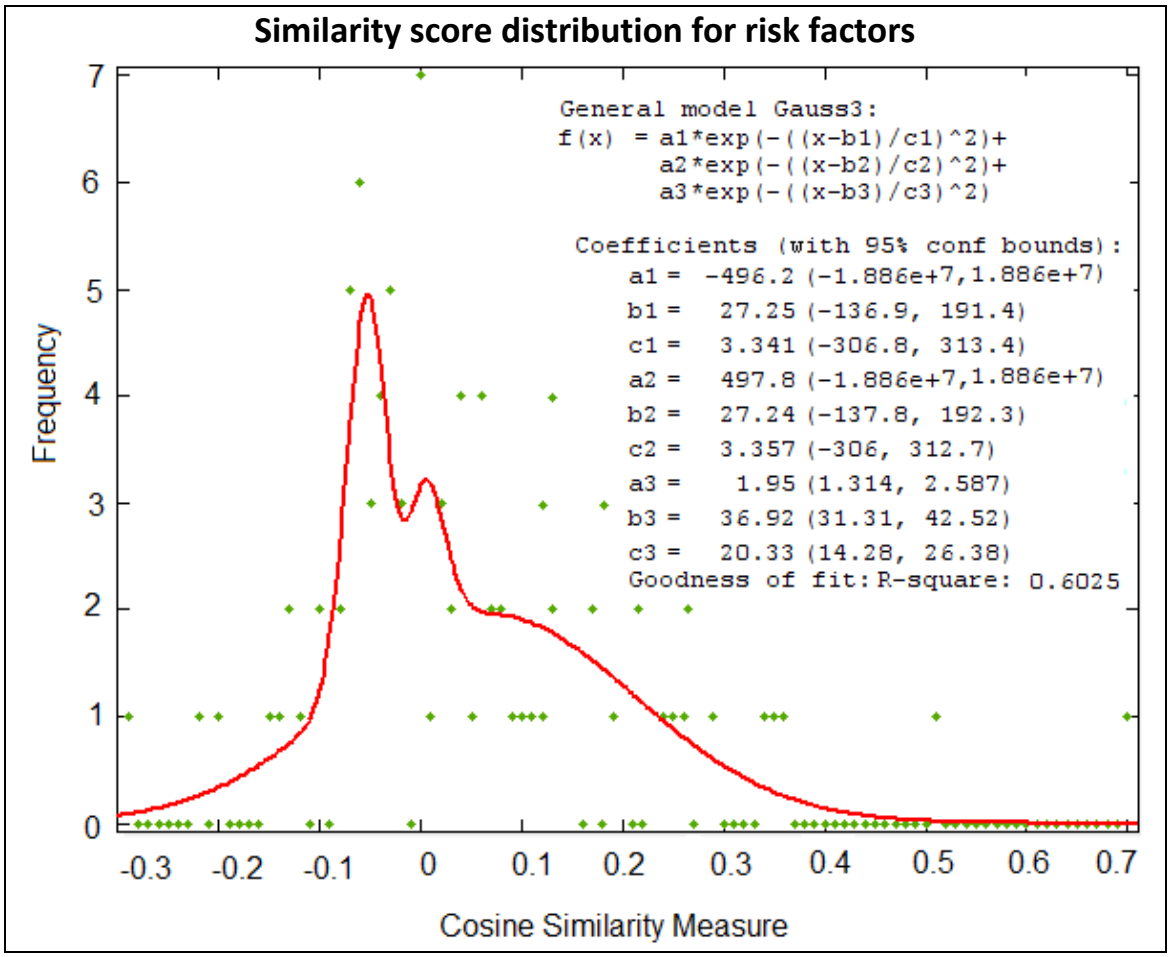

Figure 7: Distribution of factors for Parkinson's Disease.
Data and the tri-modal distribution fitted to the data using Matlab's curve fitting toolbox. The frequency represents number of factors at each similarity level when the query is "parkinson".

The goal of this analysis is to identify and score association level of different factors with respect to a given phenotype state. Using a larger bin size, the histogram representation from figure 6 and 7 are plotted again in figure 8. Higher cosine values of factors correspond to higher level of association to the disease state. Negative or zero cosine values indicate that no association was found between the factors and the disease. Since the number of factors is limited to 97, it will be difficult to use p-value to select the association factors. Statistical methods used to generate p-value will only identify a very

small set of factors (i.e. 3-4) to be associated to the disease with high level of confidence; however this number is expected to be much larger. Using a larger set of factors will alleviate this limitation and p-value can be calculated to evaluate the level of association.

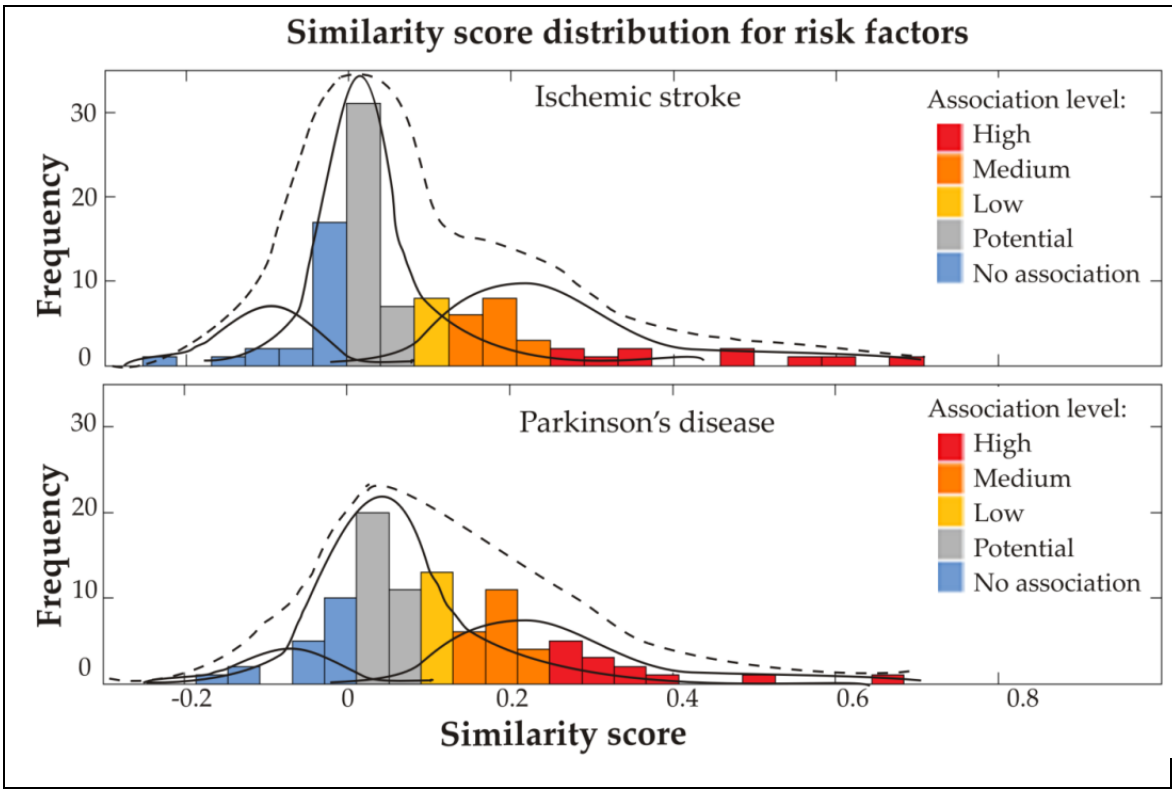

Figure 8: Similarity score distribution (dashed line) for risk factors associated with Ischemic Stroke and Parkinson's disease.
Tri-modal distribution models are represented by solid lines (approximated from figure 6 and R2).

Comparison between the retrieved factors at varying levels of association and the common knowledge in medicine was made possible by reviewing the literature through MedLink. Medlink is a web resource for neurologist to access up-to-date information for

clinical use. The following figure (figure 9) demonstrates the level of cohesion and overlap between the identified factors and those listed in the MedLink neurology.

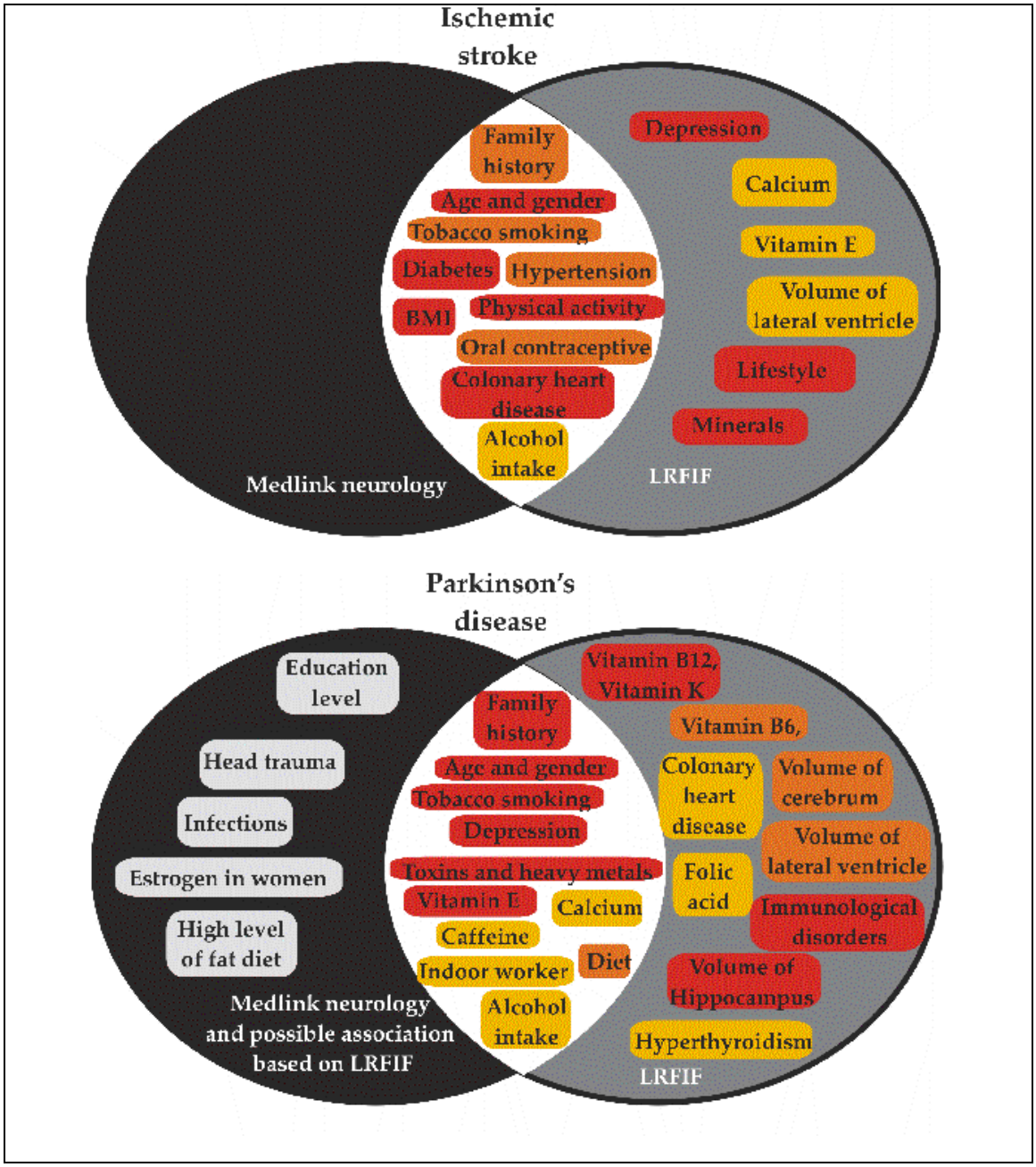

Figure 9: Venn diagram of potential risk factors identified by MedLink Neurology and the proposed literature-based risk factor identification framework (LRFIF) for Ischemic stroke and Parkinson's disease
Association level; red: high; orange: medium; yellow: low; gray: potential; blue: none.

In the case of Ischemic stroke, most of the factors are identified by both systems; however there is a set of factors that have only been identified by the proposed approach. A systematic review of this set is included in the discussion section of this thesis. In the case of Parkinson's disease, a large number of factors have been identified by both systems; there are however a number of factors that are only been identified by one of the methods. A review of each of these factors is presented in the discussion section. Furthermore, a small scale disease network is constructed based on the 97 factors. To create this network, four diseases are queried, ALS, Breast Cancer, Ischemic stroke and Parkinson's disease (see figure 10); the resulting network demonstrate the association that exists between these diseases using the different factors.

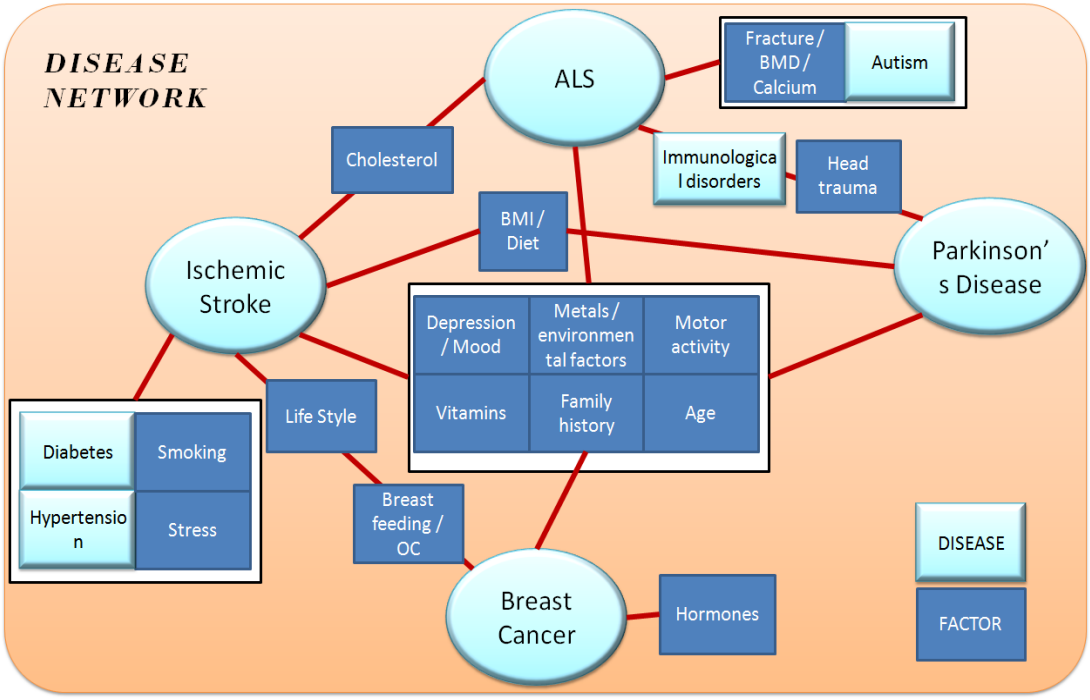

Figure 10: Disease network for ALS, IS, PD and BC.
Factors include, risk factors other diseases, side effects, and signs and symptoms.

Increasing the number of risk factors by MeSH hierarchy provides a robust and scalable approach. At a first level, 276 factors are selected to construct the set of factors and a ground truth is constructed based on the information from PubMed (see Material and Methods). Two diseases are considered for this analysis: 1) breast cancer a systemic disease, associated with a larger number of factors; 2) Ischemic stroke, a well understood and well controlled disease with better defined factors and treatment options. Table 2 presents summery of the ground truth analysis.

Table 2: Number of identified factors for Breast Cancer and Ischemic Stroke.

	Breast Cancer			Ischemic Stroke		
Association level	Established factor	Possible factor	Unknown	Established factor	Possible factor	Unknown
Number of factors	**40**	**101**	**135**	**58**	**63**	**155**
Percentage	**14%**	**37%**	**49%**	**21%**	**23%**	**56%**

The number of established factors is higher for stroke when compared to breast cancer; contrarily the number of possible factors is higher in the case of breast cancer. This observation is in accordance with the fact that Ischemic stroke is a better understood disease. Factors which have been associated with the disease based on a limited number of evidences are grouped in the category of "possible factors". Factors that have accepted by the community to be associated to the disease are grouped in the "established factor" group, and the remaining of the factors are in the group of "unknown" factors. Simulations are still in progress; however, the ground truth which is the most time-consuming step of this process is completed and analysis of the results will follow shortly.

Develop large scale literature-mining to facilitate information retrieval for the purpose of literature search in the field of medical genetics

Using a cut-off value of 0.1 cosine similarity score and the keyword "stroke", 800 human genes were selected using GeneIndexer software. These 800 genes were imported in PubMatrix with seven keywords as representative of the keyword stroke. PubMatrix returned the number of abstract for each search combination and the resulting matrix was analyzed. Table 3 demonstrates a small set of this analysis for illustrative purposes.

Table 3: Partial results obtained from GeneIndexer and PubMatrix.

Gene indexer		PubMatrix - number of articles							
Gene name	Cosine Similariry Score	1. stroke	2. Cerebral ischemia	3. Brain infarction	4. CVA	5. Cerebrovascular accident	6. TIA	7. Transient Ischemic Attack	*Average*
mthfr	0.576	310	9	11	311	3	8	18	*95.7*
f5	0.403	9	0	0	9	0	1	0	*2.7*
nbpf3	0.396	1	0	0	1	0	0	0	*0.3*
ggt2	0.38	0	0	0	0	0	0	0	*0.0*
kif6	0.37	2	0	0	2	0	0	0	*0.6*
cpb2	0.335	1	0	0	1	0	0	0	*0.3*
tmtc2	0.33	0	0	0	0	0	0	0	*0.0*
lpa	0.33	8	0	0	8	0	1	1	*2.6*
caq14	0.326	0	0	0	0	0	0	0	*0.0*

To select the best threshold for PubMatix, different cut-offs were considered (see figure 11). A threshold of 1 indicates that on average, the gene had one retrieved abstract from PubMed when searched in combination with the seven keywords. A threshold of 1, resulted in 173 genes from the 800 genes, and a threshold of 5 reduces the gene list to 113. In this analysis, a threshold of 10 was considered, resulting in 90 genes from the set of 800.

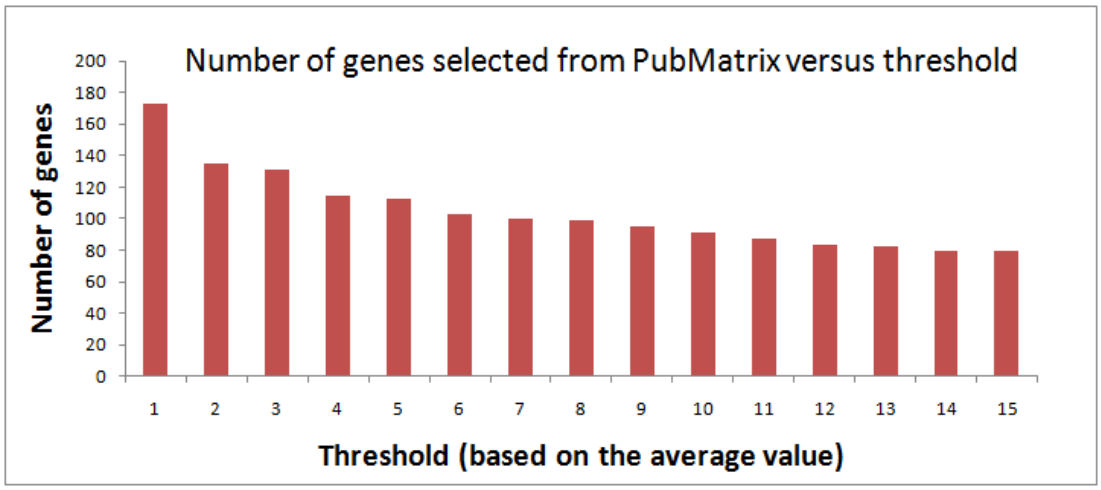

Figure 11: Number of genes when different thresholds were considered in PubMatrix analysis.

The list of 90 genes and their respective description is provided in the supplemental results section (see table 7). Further analysis of the reduced gene-set using enrichment analysis tools allows further classification of the genes using their respective Gene Ontology (GO) information. Top ten GO categories from the second, third, and fourth levels of GO hierarchy are listed in table 4. At level 2, the categories are more abstracts. The top ten categories in the fourth level of GO are analyzed further and 62 genes were identified to be in at least one of the ten categories (see table 5). Analysis of 62 genes or even 90 genes for a comprehensive literature review by an expert is more realistic than analysis all possible genes that could be implicated in Stroke.

Table 4: Top 10 categories at level 2, 3 and 4 of Gene Ontology. Ten other interesting categories with respect to stroke are also listed.

Number of genes	GO category
At level 2 of GO	
40	organelle
41	extracellular region
25	extracellular region part
54	metabolic proecss
39	response to stimulus
24	organelle part
45	regulation of biological process
11	locomotion
50	biological regulation
42	multicullular organismal process
At level 3 of GO	
19	coagulation
35	response to external stimulus
39	extracellular region
40	intracellular organelle
33	membrane-bounded organelle
19	regulation of body fluid levels
13	negative regulation of multicellular organismal process
24	extracellular space
3	nucleic acid binding
8	cell surface binding
At level 4 of GO	
19	blood coagulation
39	extracellular region
10	platelet activation
33	intracellular membrane-bound organelle
19	regulation of body fluid levels
14	regulation of response to external stimulus
20	extracellular space
9	extracellular coagulation
6	eukaryotic cell surface binding
8	negative regulation of coagulation
Other interesting categories (with fdr corrected p-value <0.05)	
5	fibrinolysis
9	inflamatory response
11	bood circulation
3	regulation of colagen
4	regulation of systemic arterial blood pressure by hormone
6	regulation of blood vessel size
6	vascular process in circulatory system
6	regulation of angiogenesis
7	lipid metabolic process
5	regulation of tissue remodeling

Table 5: Genes in the top ten categories at the fourth level of Gene Ontology

ABO	APOH	F2	GP1BA	LPA	PLAT	TCN2
ACE	CDC24	F5	HGF	MGP	PON2	TNF
ACE2	CHAD	F7	HP	MMP9	PSMA6	
ADD1	CRP	FGA	IL6	NOS3	PTX3	
AGTR1	CS	FGB	ITGA2	NOTCH3	RETN	
ALBP	CST3	FGG	ITGA2B	NPC1	SELP	
ANXA5	CYP7A1	FLG	ITGB3	OLR1	SERPINB1	
APCS	EEF1A2	GARS	KL	P11	SERPINE1	
APLN	EPHX2	GK	LBP	PC	SERPINE2	
APOE	F11	GNB3	LCN2	PF4	SORL1	

CHAPTER IV

DISCUSSION

The implementation of heterogeneous data integration for scientific and medical decision support systems, in addition to knowledge discovery processes is an ambitious project. Using the presented framework for the identification of associated factors, it is possible to build disease models and disease networks at different levels of granularity. The fact that many risk factors are diseases and can contribute to the dysfunction of important physiological pathways creates a circular dilemma but also a great opportunity. We have shown for instance, that Parkinson's disease is highly associated to a number of factors including immunological disorders, which itself is a disease. If one can corroborate that hypothesis, then PD can significantly benefit from innovative new treatment options that are currently being used for immunological disorders.

These interactions can be used to build disease networks and promote interdisciplinary research in different fields. The idea of using risk and other associated factors derived from MeSH hierarchy as a feature space is unique in the sense that it provides potential for hypothesis generation and advances our understanding of complex diseases. In addition, this framework provides prior knowledge based on the current literature and can be used for a multimodal data integration strategy. More specifically, factors can be used to measure the degree of association between disease groups and generate new hypotheses for future research directions. *De novo* hypothesis generation can provide a novel approach on how we design experiments and select the parameters for the study.

Finally, in the second part of this study, a fusion on different text-mining tools was performed to provide a customized framework for the identification of genes related to a disease. The aim of this exercise was to assist medical researchers perform a review of the literature in a timely manner by using more than just one text-mining tool and exploit benefits of different techniques at once.

Literature-mining framework to model disease-disease interaction network

Modeling associated factors of phenotypic states (such as diseases) using a tri-modal distribution can be beneficial in grouping factors and score their association levels. In a tri-modal distribution model, the rightmost distribution represents factors with the highest degree of association to the phenotypic state while the leftmost distribution represents the lowest degree of association. In the middle distribution, factors are slightly associated to the phenotypic state or disease; future research directions can be helpful in corroborating the existence or absence of such interactions. This type of analysis facilitates interdisciplinary research and brings together scientists from an array of fields. The main characteristic of the tri-modal distribution is the following: the greater the separation between the leftmost and rightmost distributions the more a disease is studied. Similarly, the more a disease is studied, the more bias there is in the distribution (mean values are farther apart from zero, or from each other) which translates into a smaller variance.

These general features are observed in the comparative analysis of Ischemic stroke (IS) and Parkinson's disease (PD). Ischemic stroke is relatively better understood and researched disease with many of its risk factors and treatment outcomes documented.

This fact translates as expected in the disease model (see figure 6). In fact, the difference of mean values of the rightmost and leftmost distributions is higher in the case of IS than in the case of PD. Likewise, variance of all the three distributions in IS are lower than the variance of all the three distributions in PD. This is also clear from the shape of the two tri-modal distributions.

Comparing the results obtained by the POLSA-association-study in Ischemic stroke (IS) and Parkinson's disease (PD) a number of factors were identified to be associated to the disease in either the proposed method or by the MedLink web-resources only. In the case of IS, six factors were only identified by the proposed method: lifestyle, calcium, minerals, depression, vitamin E and volume of lateral ventricle. A manual review of the literature is performed to find evidences for these associations.

Lifestyle: In a comprehensive literature review by Galimanis et al., [30], important articles were analyzed to highlight recent advances in stroke prevention by identifying old and new lifestyle factors. Many lifestyle elements such as behavior, heavy alcohol consumption, active or passive smoking, physical activity, education and diet have been linked to stroke. It is possible that in each individual paper there are no direct references made to lifestyle choices; however, LSA-based technique captures direct and indirect associations. Based on the vast amount of literature reviewed on individual lifestyle factors, lifestyle was found to be highly linked to stroke.

Calcium / minerals: Calcium was found to fairly associated to Ischemic stroke in our analysis (cosine score of 0.13). In a systematic review of 2906 papers, Catling LA *et al.,* [31] showed a statistically significant inverse association between magnesium and cardiovascular mortality (OR 0.75 (95% CI 0.68, 0.82), P<0.001); however evidence for

calcium remained unclear. Further research is warranted to corroborate presence or absence of such association. An advantage of the proposed method is to expose the researcher to potential factors that might have been overlooked in the manual literature review. In addition, this analysis provides a faster and more efficient method to identify new research directions in the field. Moreover, POLSA-based association identification framework identified other minerals to be fairly to highly associated with stroke. Zinc, selenium, chloride, copper and phosphorus were fairly associated to stroke; while magnesium was moderately and chromium highly associated to stroke. Note that chromium is used in cobalt-chromium stent platforms for treatment of patients with coronary artery disease [32] therefore this association is naturally expected.

Depression: To assess the association of depression to ischemic stroke, three keywords were selected: morning cortisol level, mood, and stress. This is important since depression can manifest itself through mood change, stress level or biochemical alteration of hormones such as cortisol. For instance, it is known that in "non-depressed people" the level of cortisol peaks in the morning and decreases as the day progresses. In depressed people, however, cortisol peaks earlier in the morning but does not decrease as is the case for normal people [33]. In a detailed review of the literature in 2010 [34] cortisol was found to significantly contribute to the incidence and intensity of anxiety and depression as well as stroke. In the POLSA framework, morning cortisol level, mood and stress had a cosine score of 0.48, 0.18 and 0.12 respectively. This level of association provides evidence that depression and stroke are likely to be associated through various elements such as hormone level, mood, or stress level. Such analysis can help direct further

research to better understand stroke and depression as diseases and plan preventive measures accordingly.

Vitamin E: Vitamin E was also found to be fairly associated to stroke (cosine score of 0.12) in POLSA-based approach. In a systematic review and meta-analysis [35] of randomized, placebo controlled trials (published until January 2010) nine trials were included that investigated the effect of vitamin E on incident of stroke for a total of 118,765 participants. The results corroborate that vitamin E increased the risk for hemorrhagic stroke by 22% and reduced the risk of ischemic stroke by 10%. This is a very important observation because our analysis was based on the literature prior to this review article; and based on our prediction such association was probable.

Volume of lateral ventricle: Finally volume of lateral ventricle is also found to be fairly associated with stroke in our analysis (cosine score of 0.13). No strong evidence was found in the literature for this association; however, this may be due to the fact that "lateral ventricle" is directly linked to stroke [36] and articles retrieved to generate the document set for this factor were mostly related to "lateral ventricle" rather than volume of lateral ventricle. This is due to how the PubMed search returned the articles. Using MeSH-base factors, these types of noises are reduced to a minimum, because, if an article is tagged to a given MeSH category, it will most likely contain relevant information. Note that, some of the MeSH tags are also computationally derived; therefore this problem may not be completely solved.

Similarly, in the case of PD, nine factors were identified by the proposed method and not by MedLink web-resource. These included the followings: immunological disorders; vitamin B6, B12 and K; folic acid; volume of lateral ventricle, cerebrum and

hippocampus; coronary heart disease and hyperthyroidism. In addition to that, five factors were identified by MedLink but these were only potentially link to PD in our analysis: Head trauma, education level, infections, estrogen in women and high level of fat diet. A manual review of the literature is performed to investigate the presence or lack of evidences for these associations.

Immunological disorders: Immunological disorder is highly linked with PD in our analysis (cosine score of 0.29) but not mentioned in MedLink neurology. Decreased olfaction is one of the earliest non-motor signs of PD [37,38]. Additionally, the sense of smell has a key impact in the physiology of the immune system. Specifically, the oral-gastrointestinal tract is the major source of immune education [39] and changes in the diet can alter the immunoreactivity. In fact it was shown that immunologically manipulated mice have changes in the diet selection [40]. Additionally immunological diseases are accompanied by a decrease in the sense of smell (HIV, cancer) [41,42,43]. These studies point to a probable link between impaired immune system and PD [38]. Nonetheless more directed research is strongly recommended because understanding the nature of PD may lay the ground for successes in the quest of delaying the progression of this degenerative disease.

Vitamin B6, B12 and K: Vitamin B6, B12 and vitamin K are moderately to highly associated with PD in our study (cosine of 0.2, 0.21 and 0.36 respectively). Reviewing the literature, it was found that retrospective studies in healthy subjects showed no connection between intake of vitamin B12 and the risk of developing Parkinson's disease [44], although one study reported an association of low levels of vitamin B6 and Parkinson's disease [45]. However, elevated homocysteine levels in

patients with Parkinson's disease may accelerate neurodegeneration. But elevated levels of homocysteine can be treated with folate, vitamin B6, or B12 since deficiencies of these vitamins can lead to high homocysteine levels [46]. In addition, vitamin B6 is necessary for homocysteine trans-sulfuration supporting the synthesis of glutathione. The latter is used for the defense against oxidative stress which is also involved in PD [review in ref. 47]. Finally, even though a lack of folate and vitamin B12, as well as elevated levels of homocysteine have been shown to promote neurodegeneration, there is no evidence at the clinical level that illustrates these compounds are involved in the specific etiology or pathogenesis of Parkinson's disease [47]. Therefore, more direct clinical studies are warranted in this field. In addition to vitamin B, vitamin K has also been shown to be associated with PD in a small scale meta-analysis study [48].

Folic acid: Folate is an essential vitamin (also known as vitamin B9) and its deficiency is related with an increased level of homocysteinemia. It has been shown that in some, an antibody is produced against folate that prevents it from properly entering the brain. This process has been shown to worsen certain cognitive functions of the patients with PD [49]. Studies are in progress to assess the impact of folate on the progression of PD [50]. The association of folic acid and PD in our analysis was existent but relatively low (cosine score of 0.1).

Volume of lateral ventricle, cerebrum and hippocampus: PD and brain volume seems to be correlated [51,52,53,54]. In fact, in a recent study [51] there was evidence of an association between asymmetrical lateral ventricular enlargement with PD motor asymmetry and progression ($r=0.96$, $P<0.001$). Further studies are needed to investigate the underlying mechanisms and potential for using the lateral ventricle volume

measurement as a marker for the progression of PD. In a different study patients with PD showed atrophy in the hippocampus and the prefrontal cortex [52]. Similarly, in another study it was found that hippocampal atrophy, prominent in PD with dementia, is primarily centered in the head of the hippocampus [53]. Even though it is known that atrophy in the hippocampus (or amygdala) is not specific to PD, [53,55] more directed research can facilitate differentiating patients suffering from dementia in PD and Alzheimer's disease based on non-invasive, imaging-based diagnostics. Volume of lateral ventricle, cerebrum and hippocampus were moderately-to-highly linked to PD in our study (cosine score of 0.17, 0.24 and 0.27 respectively).

Coronary heart disease: Coronary heart disease (CHD) may be indirectly associated with PD through vascular dementia. Vascular dementia results from interrupted blood flow to the brain, often after one or a series of stroke attacks. In our analysis CHD was linked to PD by a cosine score of 0.12.

Hyperthyroidism: in a retrospective study although hypothyroidism was not found to be more common in patients with PD, yet two major observations were made: men with Parkinson's were more likely to have abnormal thyroid laboratory tests as compared with controls; and '*subclinical*' hyperthyroidism was found to be more prevalent in patients with PD [56]. Furthermore, possible association of PD and hypothyroidism could be through pathological intraneuronal accumulation of neurofilaments [57] or based on the fact that many of the symptoms of the two disorders are similar [58,59]. In a recent review article, where authors compared different types of cancer in patients with PD, it was found that thyroid cancer was reported at a higher than expected rate ($P<0.0001$) [60]. Therefore more studies are needed to shed light on the

underlying mechanism for this association. Furthermore it is important to note that, depression, an associated factor to PD (based on our analysis as well as MedLink data), is well known to be highly linked to thyroid dysfunction. Our analysis assigned a score of 0.1 to the association between hyperthyroidism and PD.

Missed associations: The relationship between **head trauma** and the development of Parkinson's disease has been an issue in neurology since James Parkinson's initial 1817 essay. [61,62]. As we have only reviewed publications from the past 20 years, it is very likely that we have overlooked the wave of publications that associated head trauma as a risk factor for PD (cosine score between head trauma and PD was 0.07). **Education level** was not one of our factors, but "health education and health promotion" was; the latter had a low but positive cosine score (0.04). **Infection** was divided into two categories: bacterial and viral infection. The score for bacterial infection was positive, and the score for viral infection was negative. However, since bacterial and viral infections are very broad categories, the amount of literature was significantly biased causing the cosine score to be an underestimate of the true value. This is when systematic bias in the dataset could cause bias in the final analysis. To alleviate this problem, we are working on a local-based-LSA technique, meanwhile the factors selected using MeSH hierarchy are more specific and therefore the bias in the dataset is expected to reduce. **High-level of fat in the diet** was not specifically selected as a factor; instead cholesterol was a factor in our dataset. Cholesterol was slightly linked to PD in our analysis (cosine score of 0.069). **Estrogen** in women was not specifically used in our list of 97 factors; instead estradiol was one of the factors and its level of association with PD was very low (cosine value of 0.001).

Parkinson's disease (PD) is less studied than ischemic stroke (IS) and it is clear from this analysis that the proportion of factors detected by both systems (POLSA-based analysis versus MedLink web-resource) is higher in the case of IS. Interestingly, associations detected only by our framework can facilitate extraction of interesting observations and possible trends in the field. For instance, the fact that PD could possibly be associated with immunological disorders is intriguing and could provide new prospects to prevention and possibly treatment of the disease. This also facilitates interdisciplinary research and enhances interaction between scientists from sub-specialized fields.

Improvements to the literature-mining framework to model disease-disease interaction network

The limitations caused by manually selecting the set of factors proved to be one of the main drawbacks from the previous implementation of the factor identification network. The set of 97 factors had no hierarchical structure and was not designed to systematically cover a wide range of topics. In addition, it was observed that some of the factors were very specific (such as volume of lateral ventricle) and some very general (bacterial infection). Furthermore, some of these terms were not exactly mentioned in the ground truth due to the level of their abstraction; hence making the comparison difficult and not accurate.

To alleviate these problems, and make the system scalable to a larger set of factors, MeSH-based hierarchy was used and a set of 276 MeSH terms were selected to represent the set of factors. This selection was made by an expert in the medical field. The set of 276 terms were selected from the eight main categories. Selecting a larger set

of factors also allows factor identification in a more systematic manner; in addition, use of statistics and p-values for selection of relevant factors becomes realistic and feasible.

To further improve the LSA technique, multi-gram MeSH dictionary (mono, bi and tri-gram dictionary) was designed and constructed; the main objective was to incorporate added semantic meaning and enhance the query-based selection process of the LSA. For example, the term "vascular accident" is equivalent to stroke, yet if the two keywords ("vascular" and "accident") are considered independently the semantic meaning is lost in this case. We have made this observation when exploring risk factors of "stroke" and compared them with risk factors for "vascular accident" and other synonym words (Cerebral ischemia, Brain infarction, CVA, Cerebrovascular accident, TIA, Transient Ischemic Attack). By implementing an improved method based on multi-gram dictionary, quality of the results is expected to be enhanced and be robust to multi-keyword queries. However, due to the complexity of the implementation and additional features, simulation results are not yet completed. Nevertheless, ground truth is constructed for two diseases, breast cancer and ischemic stroke, based on the 276 MeSH terms.

Develop large scale literature-mining to facilitate information retrieval for the purpose of literature search in the field of medical genetics

Typically a literature review is written by an expert in the field and requires a great amount of time for the expert to browse and gather relevant information from the large amount of textual data. Ultimately, the final product is extremely useful for new researchers; it helps them gain a deeper understanding of the topic in a timely manner. However, due to the huge amount of textual data, this task is very time consuming and

becomes more difficult as the number of publications increases. In this study we proposed a semi-supervised literature review technique. The methodology has the goal of identifying genes that have been associated to stroke in the literature and utilizes existing bioinformatics tools. The result is unbiased toward the year of publication, impact factor of the journal, ranking of the establishments reporting the results. In addition, the system is designed to gather direct and indirect association using textual data as well as ontological resources (Gene Ontology). Using this approach, the focus will not only be on the genes but also on the biological processes and molecular functions that are associated with queried disease, in this case Stroke.

As also described in the results section, using GeneIndexer, top 800 genes were selected. The threshold was not to discriminate genes based on low cosine score. The goal was to have a large pool of genes to input to a second tool (PubMatrix). Note that only one keyword was used for GeneIndexer, because the latter is capable of extracting indirect associations as well as direct associations. The large pool of genes was reduced to 90 genes using PubMatrix. PubMatrix relies only on direct associations, co-occurrence of words, to score the genes. This implies that many of the genes that were among the 800, and were filtered out could still be important for stroke; however, the goal of this exercise was not to propose new hypotheses, but to gather evidences for the association between a given gene and the disease in a systematic manner. Hence, the methodology could be altered to allow gene discovery or hypothesis generation.

The reduced gene set was analyzed using Onto-Express, an enrichment analysis toolbox. At the third and fourth level of GO tree, interesting categories were observed. For instance, coagulation, response to stimulus, and regulation of body fluid levels were

among the highly significant categories. In addition, other categories were identified to be significantly enriched such as fibrinolysis, inflammatory response, regulation of blood vessel size and lipid metabolic process. Identification of these categories and the genes in these categories facilitate the literature review process greatly. It is time efficient and allows researchers to only review the relevant information. In addition, by having a tolerant threshold more autonomy can be given to the researcher. Finally from the 90 genes, 62 genes were in at least one of the highly significant GO categories. Many of these genes are well known to be linked to stroke. For instance there is a large amount of literature associating F5, Notch3, Ace, and Nos3 to ischemic stroke as described below.

F5: This gene encodes an essential cofactor (coagulation factor V) of the blood coagulation cascade. Several studies including a meta-analysis identified an association between this gene and stroke [see review in ref. 63,64].

Notch3: The mutation of this gene causes Cerebral autosomal dominant arteriopathy with subcortical infarcts and leucoencephalopathy (CADASIL). The clinical symptoms include recurrent strokes and transient ischemic attacks, as well as progressive cognitive impairment [see review in ref. 65].

Ace: ACE produces angiotensin II and catabolises bradykinin which affect vascular tone, endothelial function, and smooth-muscle-cell proliferation. Several studies have indicated that genetic variation in this protein contributes to the risk of IS [see review in ref. 63,64].

Nos3: The glu298asp polymorphism in NOS3 gene was reported to be associated with IS [see review in ref. 64].

A careful review of the literature based on the 90 genes, with more emphasis on the 62 genes, is being performed by two neurologists.

CHAPTER V

Conclusion

Finally, in the post-genomic era, much focus was aimed towards utilization of high-throughput technologies, such as microarrays, to generate large amount of data at once. However, analyzing large datasets proved to be challenging and part of the research was focus to better understand the global view of the biological systems. In order to achieve that goal and shed light on some of the fundamental biological questions, researchers have used the large datasets to build networks; these networks helped scientists understand the relationship among entities and gain a deeper insight of the system topologies and natural laws governing the biological systems. For instance, it is well known today that biological networks are scale free. This has major implications; for example, in a protein-protein interaction network many proteins are associated with few proteins and only a handful of proteins are connected to a large set of proteins. This observation is also valid for gene-regulatory networks, in which nodes are transcription factors and edges are protein-DNA interactions [5].

In essence developing a disease-disease interaction network proposed in this study is not a novel idea; however, the concept of using the literature to model association between nodes and using concepts (such as risk factors, side effects or treatment options) as nodes of the network is novel and can be very insightful. In a conventional network analysis, nodes are usually one type of entity such as genes, proteins, or even diseases, and edges are associates between the two entities which can be modelled differently. For instance, in a disease-disease interaction network, nodes are diseases and edges are

common genes among the disease; in a protein-protein interaction network, nodes are proteins and edges are genetic and/or physical association between two proteins. In the proposed disease-disease interaction network, nodes are factors, which could be diseases, side-effects of drugs, risk factors or any other biological entity (from the MeSH hierarchy), and edges are literature-derived associations. The flexibility of the system makes it practical and valuable for medical scientists. The system can be queried for any factor and the output returns a set of associated factors and their relative level of association. By querying a number of diseases it is possible to build a customized disease network and investigate potential relationships among different entities. Furthermore, since associations are modeled using literature data, it is possible to update the system regularly, and as the time progresses and our knowledge expands this framework becomes more valuable and more reliable.

Furthermore, in the second section of this thesis we demonstrated that bioinformatics tools can be of great value in knowledge discovery. It is clear that the expert knowledge cannot be replaced by bioinformatics tools; however, by using for instance, a literature-mining approach, the researcher will only be required to read the relevant information and save a significant amount of time. For example, a biologist can then look for logical inconsistencies in the text. In addition, a reasonable assumption would be to give more weight to the latest findings, there are exceptions including errors or fabricated facts and an expert in the field is better at identifying these discrepancies. It will be unproductive not to take advantage of the array of text-mining tools that are being developed and offered freely to the community. However, one should always be aware of the limitations of these technologies. For instance as a community we agree that scientists

should share all useful findings. Yet some datasets are too large to include in the text-based articles format; some facts are simply too trivial to merit a paper; isolated finding or negative results are often withheld from the published record. Therefore, missing elements in the text data are not necessarily a weakness but characteristic of the data, and these should not be a reason for not taking advantage of such tools.

References

1. Pasternak JJ: *An Introduction to Human Molecular Genetics.* 2nd Ed. Hoboken, NJ: Wiley; 2005.

2. Broeckel U, Schork NJ. **Identifying genes and genetic variation underlying human diseases and complex phenotypes via recombination mapping**: *J Physiol* 2004, **554(Pt 1):**40-45.

3. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL: **The human disease network**. *Proc Natl Acad Sci U S A* 2007, **104(21)**:8685-8690.

4. Xuehong Z, Ruijie Z, Yongshuai J, Peng S, Guoping T, Xing W, Hongchao L, Xia L: **The expanded human disease network combining protein–protein interaction information.** *European Journal of Human Genetics* 2011, doi:10.1038/ejhg.2011.30

5. Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L: **Integrated genomic and proteomic analyses of a systematically perturbed metabolic network**. *Science* 2001, **292(5518)**:929-934.

6. Ono T, Hishigaki H, Tanigami A, Takagi T: **Automated extraction of information on protein–protein interactions from the biological literature**. *Bioinformatics* 2001, **12**:155–161

7. Blaschke C, Hirschman L, Valencia A: **Information extraction in molecular biology**. *Brief Bioinform* 2002, **3(2)**:154-165.

8. Hao, Y. Zhu X, Huang M, Li M: **Discovering patterns to extract protein–protein interactions from the literature. Part II.** *Bioinformatics* 2005, **21**:3294–3300

9. Chun HW, Tsuruoka Y, Kim JD, Shiba R, Nagata N, Hishiki T, Tsuji J: **Extraction of gene-disease relations from Medline using domain dictionaries and machine learning**. *Pac Symp Biocomput* 2006:4-15.

10. Berman JJ: **Pathology abbreviated - A long review of short terms.** *Archives of Pathology & Laboratory Medicine* 2004,**128(3)**:347-352.

11. Rzhetsky A, Seringhaus M, Gerstein M: **Seeking a new biology through text mining**. *Cell* 2008, **134(1)**:9-13.

12. Hirschman L, Morgan AA, Yeh AS: **Rutabaga by any other name: extracting biological names**. *J Biomed Inform* 2002, **35(4)**:247-259.

13. Wilbur WJ, Hazard GF, Divita G, Mork JG, Aronson AR, Browne AC: **Analysis of biomedical text for chemical names: a comparison of three methods.** *Proc AMIA Symp* 1999:176-180.

14. Swanson, D, Smalheiser, N: **Assessing a gap in the biomedical literature: magnesium deficiency and neurologic disease**. *Neurosci Res Commun* 1994, **15**: 1–9

15. Srinivasan P, Libbus B: **Mining MEDLINE for implicit links between dietary substances and diseases**. *Bioinformatics* 2004, **20 Suppl 1**:i290-296.

16. Landauer, TK, Dumais, ST**: A solution to plato's problem: the latent semantic analysis theory of the acquisition, induction, and representation of knowledge**. *Psychological Review* 1997, **104**:211-240.

17. Berry MW, Browne M: *Understanding Search Engines: Mathematical Modeling and Text Retrieval*. SIAM, Philadelphia 1990.

18. Yeasin M, Malempati H, Homayouni R, Sorower MS: **A systematic study on latent semantic analysis model parameters for mining biomedical literature**. *Conference Proceedings: BMC Bioinformatics2009,* **10**(Suppl. 7):A6

19. Kell DB: **Metabolomics and systems biology: making sense of the soup**. *Curr Opin Microbiol* 2004, **7(3)**:296-307.

20. Kell DB: Theodor Bücher Lecture. **Metabolomics, modelling and machine learning in systems biology - towards an understanding of the languages of cells**. Delivered on 3 July 2005 at the 30th FEBS Congress and the 9th IUBMB conference in Budapest. *FEBS J* 2006, **273(5)**:873-894.

21. Brent R, Lok L: Cell biology. **A fishing buddy for hypothesis generators**. *Science* 2005, **308(5721)**:504-506.

22. Castelo R, Siebes A: **Priors on network structures: biasing the search for Bayesian networks**. *Int J Approx Reason* 2000, **24(1)**: 39-57.

23. Imoto S, Higuchi T, Goto T, Tashiro K, Kuhara S, Miyano S: **Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks**. *Proc IEEE Comput Soc Bioinform Conf* 2003, **2**:104-113.

24. Steele E, Tucker A, 't Hoen PA, Schuemie MJ: **Literature-based priors for gene regulatory networks**. *Bioinformatics.* 2009, **25(14)**:1768-1774.

25. Khatri P, Drăghici S: **Ontological analysis of gene expression data: current tools, limitations, and open problems**. *Bioinformatics* 2005, **21(18)**:3587-3595.

26. Khatri P, Voichita C, Kattan K, Ansari N, Khatri A, Georgescu C, Tarca AL,

Drăghici S: **Onto-Tools: new additions and improvements in 2006**. *Nucleic Acids Res* 2007, **35(Web Server issue)**:W206-211.

27. Rhee SY, Wood V, Dolinski K, Draghici S: **Use and misuse of the gene ontology annotations**. *Nat Rev Genet* 2008, **9(7)**:509-515.

28. Wong SL, Zhang LV, Roth FP: **Discovering functional relationships: biochemistry versus genetics.** *Trends Genet* 2005, **21(8)**:424-427.

29. Myers CL, Barrett DR, Hibbs MA, Huttenhower C, Troyanskaya OG: **Finding function: evaluation methods for functional genomic data**. *BMC Genomics* 2006, **7**:187.

30. Galimanis A, Mono ML, Arnold M, Nedeltchev K, Mattle HP: **Lifestyle and stroke risk: a review.***Curr Opin Neurol* 2009, **22(1)**:60-8.

31. Catling LA, Abubakar I, Lake IR, Swift L, Hunter PR: **A systematic review of analytical observational studies investigating the association between cardiovascular disease and drinking water hardness.** *J Water Health* 2008, **6(4)**:433-42.

32. Menown IA, Shand JA: **Recent advances in cardiology**. *Future Cardiol* 2010, **6(1)**:11-7.

33. Tafet GE, Idoyaga-Vargas VP, Abulafia DP, Calandria JM, Roffman SS, Chiovetta A, Shinitzky M: **Correlation between cortisol level and serotonin uptake in patients with chronic stress and depression**. *Cogn Affect Behav Neurosci* 2001, **1(4)**:388-93.

34. Williams GP: **The role of oestrogen in the pathogenesis of obesity, type 2 diabetes, breast cancer and prostate disease**. *Eur J Cancer Prev* 2010, **19(4)**:256-71.

35. Schürks M, Glynn RJ, Rist PM, Tzourio C, Kurth T: **Effects of vitamin E on stroke subtypes: meta-analysis of randomised controlled trials.** *BMJ* 2010, **341**:c5702. doi: 10.1136/bmj.c5702.

36. Zhang RL, Zhang ZG, Chopp M: **Ischemic stroke and neurogenesis in the subventricular zone.** *Neuropharmacology* 2008, **55(3)**:345-52.

37. Benkler M, Agmon-Levin N, Shoenfeld Y: **Parkinson's disease, autoimmunity, and olfaction.** *Int J Neurosci* 2009, **119(12)**:2133-2143.

38. Moscavitch SD, Szyper-Kravitz M, Shoenfeld Y: **Autoimmune pathology accounts for common manifestations in a wide range of neuro-psychiatric disorders: the olfactory and immune system interrelationship.** *Clin Immunol* 2009, **130(3)**:235-243.

39. Faria AM, Weiner HL: **Oral tolerance**. *Immunol Rev* 2005, **206**:232-259.

40. Teixeira G, Paschoal PO, de Oliveira VL, Pedruzzi MM, Campos SM, Andrade L, Nobrega A: **Diet selection in immunologically manipulated mice**. *Immunobiology* 2008, **213(1)**:1-12.

41. Schiffman SS, Sattely-Miller EA, Taylor EL, Graham BG, Landerman LR, Zervakis J, Campagna LK, Cohen HJ, Blackwell S, Garst JL: **Combination of flavor enhancement and chemosensory education improves nutritional status in older cancer patients.** *J Nutr Health Aging* 2007, **11(5)**:439-454.

42. Murphy C, Davidson TM, Jellison W, Austin S, Mathews WC, Ellison DW, Schlotfeldt C: **Sinonasal disease and olfactory impairment in HIV disease: endoscopic sinus surgery and outcome measures**. *Laryngoscope.* 2000, **110(10 Pt 1)**:1707-1710.

43. Zucco GM, Ingegneri G: **Olfactory deficits in HIV-infected patients with and without AIDS dementia complex.** *Physiol Behav* 2004, **80(5)**:669-674.

44. Chen H, Zhang SM, Schwarzschild MA, Hernan MA, Logroscino G, Willett WC, Ascherio A: **Folate intake and risk of Parkinson's disease**. *Am J Epidemiol* 2004, **160(4)**:368-375.

45. de Lau LM, Koudstaal PJ, Witteman JC, Hofman A, Breteler MM: **Dietary folate, vitamin B12, and vitamin B6 and the risk of Parkinson disease.** *Neurology* 2006, **67(2)**:315-318.

46. Miller JW, Nadeau MR, Smith D, Selhub J: **Vitamin B-6 deficiency vs folate deficiency: comparison of responses to methionine loading in rats**. *Am J Clin Nutr* 1994, **59(5)**:1033-1039.

47. Stanger O, Fowler B, Piertzik K, Huemer M, Haschke-Becher E, Semmler A, Lorenzi S, Linnebank M: **Homocysteine, folate and vitamin B12 in neuropsychiatric diseases: review and treatment recommendations.** *Expert Rev Neurother* 2009, **9(9)**:1393-1412.

48. Iwamoto J, Matsumoto H, Takeda T: **Efficacy of menatetrenone (vitamin K2) against non-vertebral and hip fractures in patients with neurological diseases: meta-analysis of three randomized, controlled trials.** *Clin Drug Investig* 2009, **29(7)**:471-479.

49. Miller JW, Selhub J, Nadeau MR, Thomas CA, Feldman RG, Wolf PA. **Effect of L-dopa on plasma homocysteine in PD patients: relationship to B-vitamin status.** *Neurology* 2003, **60(7)**:1125-1129.

50. Klivenyi P, Vecsei L. **Novel therapeutic strategies in Parkinson's disease.** *Eur J*

Clin Pharmacol 2010, **66(2)**:119-125.

51. Lewis MM, Smith AB, Styner M, Gu H, Poole R, Zhu H, Li Y, Barbero X, Gouttard S, McKeown MJ, Mailman RB, Huang X: **Asymmetrical lateral ventricular enlargement in Parkinson's disease.** *Eur J Neurol* 2009, **16(4)**:475-481.

52. Jokinen P, Brück A, Aalto S, Forsback S, Parkkola R, Rinne JO: **Impaired cognitive performance in Parkinson's disease is related to caudate dopaminergic hypofunction and hippocampal atrophy.** *Parkinsonism Relat Disord* 2009, **15(2)**:88-93.

53. Bouchard TP, Malykhin N, Martin WR, Hanstock CC, Emery DJ, Fisher NJ, Camicioli RM: **Age and dementia-associated atrophy predominates in the hippocampal head and amygdala in Parkinson's disease**. *Neurobiol Aging* 2008, **29(7)**:1027-1039.

54. Acharya HJ, Bouchard TP, Emery DJ, Camicioli RM: **Axial signs and magnetic resonance imaging correlates in Parkinson's disease.** *Can J Neurol Sci* 2007, **34(1)**:56-61.

55. Laakso MP, Partanen K, Lehtovirta M, Hallikainen M, Hänninen T, Vainio P, Riekkinen P Sr, Soininen H: **MRI of amygdala fails to diagnose early Alzheimer's disease.** *Neuroreport* 1995, **6(17)**:2414-2418.

56. Tandeter H, Levy A, Gutman G, Shvartzman P: **Subclinical thyroid disease in patients with Parkinson's disease.** *Arch Gerontol Geriatr* 2001, **33(3)**:295-300.

57. Chinnakkaruppan A, Das S, Sarkar PK: **Age related and hypothyroidism related changes on the stoichiometry of neurofilament subunits in the developing rat brain.** *Int J Dev Neurosci* 2009, **27(3)**:257-261.

58. García-Moreno JM, Chacón-Peña J: **Hypothyroidism and Parkinson's disease and the issue of diagnostic confusion.** *Mov Disord* 2003,**18(9)**:1058-1059.

59. Munhoz RP, Teive HA, Troiano AR, Hauck PR, Herdoiza Leiva MH, Graff H, Werneck LC: **Parkinson's disease and thyroid dysfunction.** *Parkinsonism Relat Disord* 2004, **10(6)**:381-383.

60. Ferreira JJ, Neutel D, Mestre T, Coelho M, Rosa MM, Rascol O, Sampaio C: **Skin cancer and Parkinson's disease**. *Mov Disord* 2010, **25(2)**:139-148.

61. Parkinson, J: Essay on the shaking palsy. London: Whittingham and Rowland, for Sherwood,Neely and Jones, 1817.

62. Factor SA, Sanchez-Ramos J, Weiner WJ: **Trauma as an etiology of**

parkinsonism: a historical review of the concept. *Mov Disord* 1988, **3(1)**:30-36.

63. Casas JP, Hingorani AD, Bautista LE, Sharma P: **Meta-analysis of genetic studies in ischemic stroke: thirty-two genes involving approximately 18,000 cases and 58,000 controls.** *Arch Neurol* 2004, **61(11)**:1652-1661.

64. Bersano A, Ballabio E, Bresolin N, Candelise L: **Genetic polymorphisms for the study of multifactorial stroke.** *Hum Mutat* 2008, **29(6)**:776-795.

65. Hervé D, Chabriat H: **CADASIL.** *J Geriatr Psychiatry Neurol* 2010, **23(4)**:269-276.

Appendices

Supplemental Material

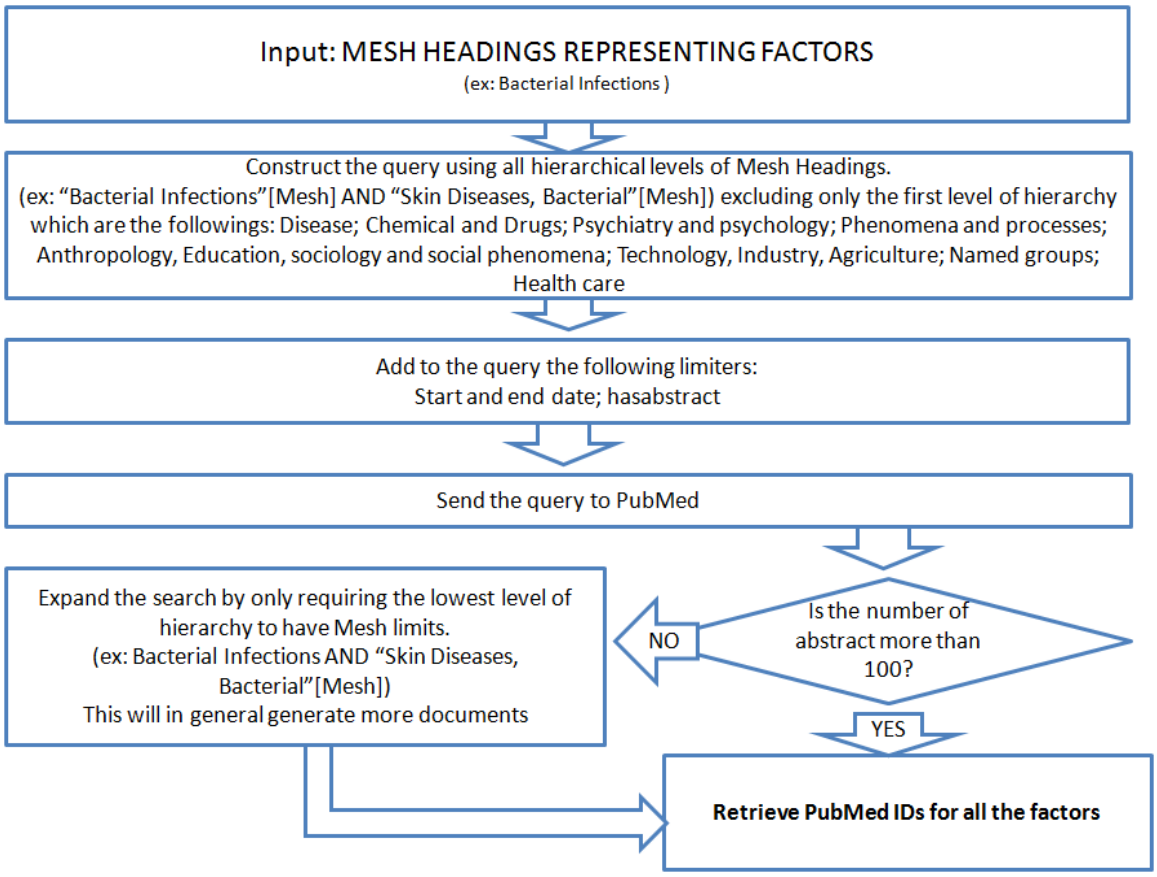

Figure 12: PubMed search strategy for the 276 risk and associated factors. The queries to PubMed are constructed using all levels of hierarchy except the highest level, which is very generic. Limiters, such as start and end date are used to control the retrieval and also to make the system more efficient for regular updates. If the number of retrieved abstract is low (one-hundred) the query will expand to include more documents, this is done by removing the quotation marks of the higher level of the MeSH hierarchy (note: The number of abstract for each of the 276 factor was analyzed and based on that distribution a threshold of 100 was selected (data not shown)).The advantage of this is to remove some degree of bias (the goal is reduce the number of risk factors having fewer than one-hundred documents in the database). PubMed IDs are retrieved and data is downloaded to the MySQL database using these identifications numbers.

Table 6: The 276 factors derived from MeSH.
NO- indicates that those factors are too general and therefore not considered; however, a selected number of subheadings are considered and listed. In brackets: MeSH identification number.

NO- Diseases [C]
NO- Bacterial Infections and Mycoses [C01] +
Bacterial Infections [C01.252] +
Bacteremia [C01.252.100] +
Central Nervous System Bacterial Infections [C01.252.200] +
Endocarditis, Bacterial [C01.252.300] +
Eye Infections, Bacterial [C01.252.354] +
Fournier Gangrene [C01.252.377]
Gram-Negative Bacterial Infections [C01.252.400] +
Gram-Positive Bacterial Infections [C01.252.410] +
Pneumonia, Bacterial [C01.252.620] +
Sexually Transmitted Diseases, Bacterial [C01.252.810] +
Skin Diseases, Bacterial [C01.252.825] +
Spirochaetales Infections [C01.252.847] +
Mycoses [C01.703] +
Zoonoses [C01.908]
NO- Virus Diseases [C02] +
Arbovirus Infections [C02.081] +
Bronchiolitis, Viral [C02.109]
Central Nervous System Viral Diseases [C02.182] +
DNA Virus Infections [C02.256] +
Eye Infections, Viral [C02.325] +
Fatigue Syndrome, Chronic [C02.330]
NO- Hepatitis, Viral, Human [C02.440] +
Hepatitis A [C02.440.420]
Hepatitis B [C02.440.435] +
Hepatitis C [C02.440.440] +
Hepatitis D [C02.440.450] +
Hepatitis E [C02.440.470]
Opportunistic Infections [C02.597] +
Pneumonia, Viral [C02.705]
RNA Virus Infections [C02.782] +
Sexually Transmitted Diseases [C02.800] +
Skin Diseases, Viral [C02.825] +
Slow Virus Diseases [C02.839] +
Tumor Virus Infections [C02.928] +
Viremia [C02.937]
Zoonoses [C02.968]
NO- Parasitic Diseases [C03] +
Central Nervous System Parasitic Infections [C03.105] +
Eye Infections, Parasitic [C03.300] +
Helminthiasis [C03.335] +
Intestinal Diseases, Parasitic [C03.432] +
Liver Diseases, Parasitic [C03.518] +
Lung Diseases, Parasitic [C03.582] +
Mesomycetozoea Infections [C03.600] +
Parasitemia [C03.695]
Protozoan Infections [C03.752] +
Skin Diseases, Parasitic [C03.858] +
Zoonoses [C03.908]
Neoplasms [C04] +

NO- Musculoskeletal Diseases [C05] +
Bone Diseases [C05.116] +
Cartilage Diseases [C05.182] +
Fasciitis [C05.321] +
Foot Deformities [C05.330] +
Hand Deformities [C05.390] +
Jaw Diseases [C05.500] +
Joint Diseases [C05.550] +
Muscular Diseases [C05.651] +
Musculoskeletal Abnormalities [C05.660] +
Rheumatic Diseases [C05.799] +
Digestive System Diseases [C06] +
Stomatognathic Diseases [C07] +
Respiratory Tract Diseases [C08] +
Otorhinolaryngologic Diseases [C09] +
NO- Nervous System Diseases [C10] +
Autoimmune Diseases of the Nervous System [C10.114] +
Autonomic Nervous System Diseases [C10.177] +
NO- Central Nervous System Diseases [C10.228] +
Encephalomyelitis [C10.228.440] +
High Pressure Neurological Syndrome [C10.228.470]
Movement Disorders [C10.228.662] +
Spinal Cord Diseases [C10.228.854] +
Chronobiology Disorders [C10.281] +
Cranial Nerve Diseases [C10.292] +
Demyelinating Diseases [C10.314] +
Nervous System Malformations [C10.500] +
Nervous System Neoplasms [C10.551] +
Neurocutaneous Syndromes [C10.562] +
Neurodegenerative Diseases [C10.574] +
Neuromuscular Diseases [C10.668] +
NO- Neurotoxicity Syndromes [C10.720] +
Botulism [C10.720.150]
Heavy Metal Poisoning, Nervous System [C10.720.475] +
MPTP Poisoning [C10.720.606]
Neuroleptic Malignant Syndrome [C10.720.737]
Sleep Disorders [C10.886] +
Trauma, Nervous System [C10.900] +
NO- Male Urogenital Diseases [C12] +
Genital Diseases, Male [C12.294] +
Urogenital Abnormalities [C12.706] +
Urogenital Neoplasms [C12.758] +
Urologic Diseases [C12.777] +
Kidney Diseases [C12.777.419] +
Urinary Bladder Diseases [C12.777.829] +
Urinary Tract Infections [C12.777.892] +
Urolithiasis [C12.777.967] +
NO- Female Urogenital Diseases and Pregnancy Complications [C13] +
Female Urogenital Diseases [C13.351] +
Pregnancy Complications [C13.703] +
NO- Cardiovascular Diseases [C14] +
Cardiovascular Abnormalities [C14.240] +
Cardiovascular Infections [C14.260] +
Vascular Diseases [C14.907] +
Aortic Diseases [C14.907.109] +
Arterial Occlusive Diseases [C14.907.137] +

Arteriovenous Malformations [C14.907.150] +
Arteritis [C14.907.184] +
Cerebrovascular Disorders [C14.907.253] +
Diabetic Angiopathies [C14.907.320] +
Hyperemia [C14.907.474]
Hypertension [C14.907.489] +
Hypotension [C14.907.514] +
Myocardial Ischemia [C14.907.585] +
Peripheral Vascular Diseases [C14.907.617] +
Vasculitis [C14.907.940] +
Venous Insufficiency [C14.907.952] +
NO- Hemic and Lymphatic Diseases [C15] +
Hematologic Diseases [C15.378] +
Lymphatic Diseases [C15.604] +
NO- Skin and Connective Tissue Diseases [C17] +
Connective Tissue Diseases [C17.300] +
NO- Nutritional and Metabolic Diseases [C18] +
NO- Metabolic Diseases [C18.452] +
Acid-Base Imbalance [C18.452.076] +
Calcium Metabolism Disorders [C18.452.174] +
DNA Repair-Deficiency Disorders [C18.452.284] +
Glucose Metabolism Disorders [C18.452.394] +
Iron Metabolism Disorders [C18.452.565] +
Lipid Metabolism Disorders [C18.452.584] +
Malabsorption Syndromes [C18.452.603] +
Metabolic Syndrome X [C18.452.625]
Metabolism, Inborn Errors [C18.452.648] +
Mitochondrial Diseases [C18.452.660] +
Phosphorus Metabolism Disorders [C18.452.750] +
Porphyrias [C18.452.811] +
Proteostasis Deficiencies [C18.452.845] +
Wasting Syndrome [C18.452.915] +
Water-Electrolyte Imbalance [C18.452.950] +
NO- Nutrition Disorders [C18.654] +
Hypervitaminosis A [C18.654.301]
Infant Nutrition Disorders [C18.654.422] +
Malnutrition [C18.654.521] +
Overnutrition [C18.654.726] +
Wasting Syndrome [C18.654.940] +
NO- Endocrine System Diseases [C19] +
Adrenal Gland Diseases [C19.053] +
Bone Diseases, Endocrine [C19.149]
Diabetes Mellitus [C19.246] +
Dwarfism [C19.297] +
Gonadal Disorders [C19.391] +
Parathyroid Diseases [C19.642] +
Pituitary Diseases [C19.700] +
Thyroid Diseases [C19.874] +
NO- Immune System Diseases [C20] +
Autoimmune Diseases [C20.111] +
Addison Disease [C20.111.163]
Antiphospholipid Syndrome [C20.111.197]
Arthritis, Rheumatoid [C20.111.199] +
Glomerulonephritis, IGA [C20.111.525]
Hepatitis, Autoimmune [C20.111.567]
Lupus Erythematosus, Systemic [C20.111.590] +

Purpura, Thrombocytopenic, Idiopathic [C20.111.759]
Thyroiditis, Autoimmune [C20.111.809]
Hypersensitivity [C20.543] +
NO- Disorders of Environmental Origin [C21] +
DNA Damage [C21.111] +
Occupational Diseases [C21.447] +
Agricultural Workers Diseases [C21.447.080] +
Dermatitis, Occupational [C21.447.270]
Inert Gas Narcosis [C21.447.426]
Persian Gulf Syndrome [C21.447.653]
Pneumoconiosis [C21.447.800] +
Poisoning [C21.613] +
Argyria [C21.613.068]
Arsenic Poisoning [C21.613.097]
Bites and Stings [C21.613.127] +
Cadmium Poisoning [C21.613.165]
Carbon Tetrachloride Poisoning [C21.613.177]
Fluoride Poisoning [C21.613.380]
Gas Poisoning [C21.613.455] +
Lead Poisoning [C21.613.589] +
Manganese Poisoning [C21.613.618]
Mercury Poisoning [C21.613.647] +
Mycotoxicosis [C21.613.680] +
Neurotoxicity Syndromes [C21.613.705] +
Plant Poisoning [C21.613.756] +
Psychoses, Substance-Induced [C21.613.809] +
Water Intoxication [C21.613.932]
Preconception Injuries [C21.676]
NO- Substance-Related Disorders [C21.739] +
Alcohol-Related Disorders [C21.739.100] +
Amphetamine-Related Disorders [C21.739.225] OR Cocaine-Related Disorders [C21.739.300] OR Marijuana Abuse [C21.739.635] OR Tobacco Use Disorder [C21.739.912]
Wounds and Injuries [C21.866] +
NO- Pathological Conditions, Signs and Symptoms [C23] +
NO- Pathologic Processes [C23.550] +
Arrhythmias, Cardiac [C23.550.073] +
Ascites [C23.550.081]
Azotemia [C23.550.145]
Dehydration [C23.550.274]
Emphysema [C23.550.325] +
Hemorrhage [C23.550.414] +
Hyperammonemia [C23.550.421]
Hyperbilirubinemia [C23.550.429] +
Hyperuricemia [C23.550.449]
Hypovolemia [C23.550.455]
Leukocytosis [C23.550.526]
Menstruation Disturbances [C23.550.568] +
Muscle Weakness [C23.550.695]
Nerve Degeneration [C23.550.737] +
NO- Signs and Symptoms [C23.888] +
Body Temperature Changes [C23.888.119] +
Body Weight [C23.888.144] +
Cardiac Output, High [C23.888.176]
Cardiac Output, Low [C23.888.192]
Chills [C23.888.208]

Cyanosis [C23.888.248]
Eye Manifestations [C23.888.307] +
Fatigue [C23.888.369] +
Flushing [C23.888.388]
Heart Murmurs [C23.888.447] +
Hot Flashes [C23.888.475]
Hypergammaglobulinemia [C23.888.512]
Intermittent Claudication [C23.888.531]
Mobility Limitation [C23.888.550]
Pain [C23.888.646] +

NO- Chemicals and Drugs [D]
Inorganic Chemicals [D01] +
Organic Chemicals [D02] +
Heterocyclic Compounds [D03] +
Polycyclic Compounds [D04] +
Macromolecular Substances [D05] +
Complex Mixtures [D20] +
Biomedical and Dental Materials [D25] +

NO- Psychiatry and Psychology [F]
NO- Behavior and Behavior Mechanisms [F01] +
Defense Mechanisms [F01.393] +
Human Development [F01.525] +
Personality [F01.752] +
NO- Psychological Phenomena and Processes [F02] +
NO- Psychophysiology [F02.830] +
Appetite [F02.830.071]
Sleep [F02.830.855] +
Stress, Psychological [F02.830.900] +
Religion and Psychology [F02.880] +
Resilience, Psychological [F02.940]

NO- Phenomena and Processes [G]
NO- Metabolic Phenomena [G03] +
NO- Body Composition [G03.180] +
Body Fat Distribution [G03.180.134]
NO- Immune System Phenomena [G12] +
CD4-CD8 Ratio [G12.248]
Immunocompetence [G12.460]
Immunocompromised Host [G12.470] +
NO- Integumentary System Physiological Phenomena [G13] +
NO- Skin Physiological Phenomena [G13.750]
NO- Skin Physiological Processes [G13.750.829]
Sweating [G13.750.829.855]
Skin Temperature [G13.750.844]
NO- Ocular Physiological Phenomena [G14] +
Refraction, Ocular [G14.760]
Vision Disparity [G14.930]
Visual Acuity [G14.940] +

NO- Anthropology, Education, Sociology and Social Phenomena [I]
NO- Social Sciences [I01] +
Quality of Life [I01.800]
NO- Sociology [I01.880] +
Culture [I01.880.143] +
Hierarchy, Social [I01.880.298]
Minority Groups [I01.880.371]
Social Class [I01.880.552] +
Social Welfare [I01.880.787] +

Socialization [I01.880.813]
Socioeconomic Factors [I01.880.840] +
Education [I02] +
Human Activities [I03] +
Exercise [I03.350] +
Leisure Activities [I03.450] +
Physical Fitness [I03.621]
Travel [I03.883] +
NO- Technology, Industry, Agriculture [J]
NO- Technology, Industry, and Agriculture [J01] +
Household Products [J01.516] +
NO- Food and Beverages [J02] +
NO- Beverages [J02.200] +
Alcoholic Beverages [J02.200.100] +
Carbonated Beverages [J02.200.300]
Coffee [J02.200.325]
Milk [J02.200.700] +
Milk Substitutes [J02.200.712] +
Mineral Waters [J02.200.806]
Tea [J02.200.900]
Food [J02.500] +
NO- Named Groups [M]
NO- Persons [M01]
Age Groups [M01.060] +
Alcoholics [M01.066]
Athletes [M01.072]
Caregivers [M01.085]
Child, Abandoned [M01.097]
Child, Exceptional [M01.102] +
Child of Impaired Parents [M01.106]
Child, Orphaned [M01.108]
Child, Unwanted [M01.111]
Consultants [M01.120]
Crime Victims [M01.135] +
Criminals [M01.142]
Disabled Persons [M01.150] +
Drug Users [M01.169]
Emigrants and Immigrants [M01.189]
Homebound Persons [M01.276]
Homeless Persons [M01.325] +
Medically Uninsured [M01.385]
Prisoners [M01.729]
Refugees [M01.755]
Single Person [M01.785]
Students [M01.848] +
Terminally Ill [M01.873]
NO- Health Care [N]
NO- Population Characteristics [N01] +
Socioeconomic Factors [N01.824]
NO- Environment and Public Health [N06] +
Environment [N06.230] +

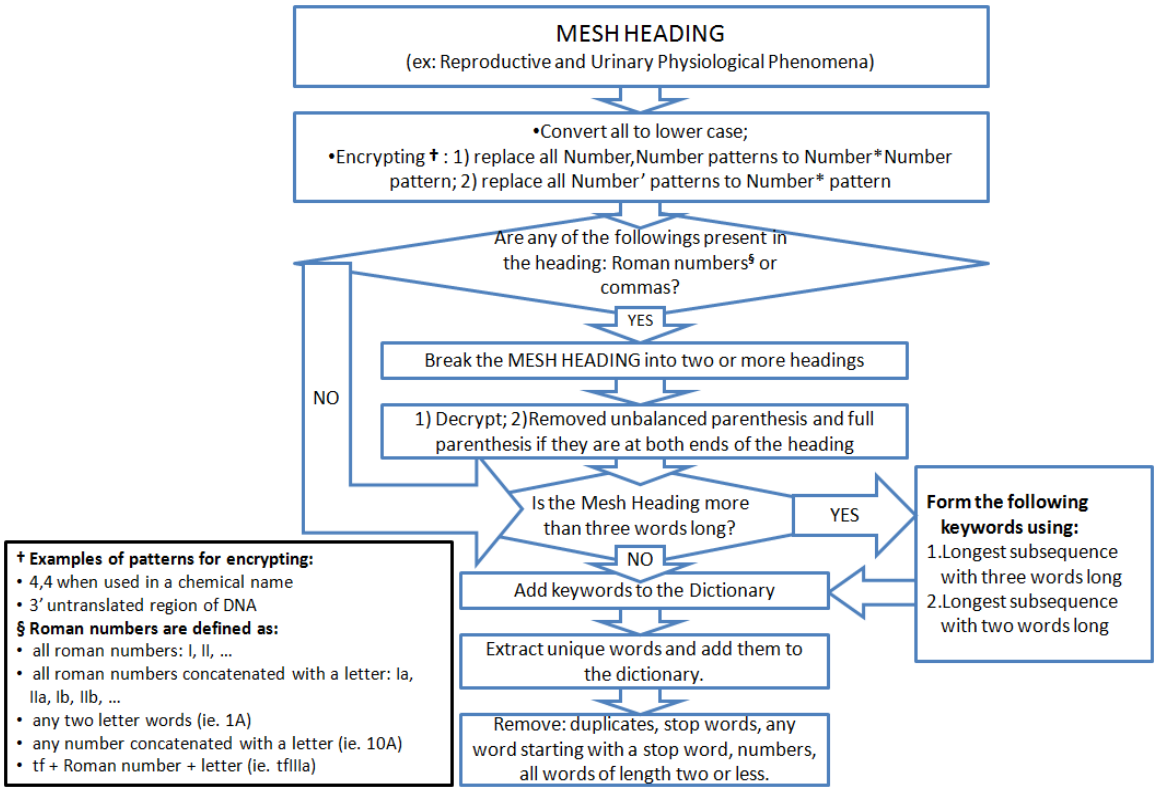

Figure 13: Multi-gram dictionary construction scheme based on MeSH.

This scheme would guarantee construction of all possible 1, 2 and 3 keyword terms for the dictionary. The encryption methodology designed here would also provide a way to ensure the keywords are coherent and logical. For instance, the comma is used a separator of different terms, however, if it separates two numbers such as the case for many chemical names then in those cases it will not be used as a term separator. In addition, the implementation is such that, it is easy to increase the multi-gram dictionary to more than 3 terms per keyword. Increasing the number of terms per keyword would increase the complexity of the search and also the dictionary size; however, with technological advances and increased in computational power it will be possible in near future to include more terms in the dictionary using this approach.

Supplemental results

Table 7: List of 90 genes and their respective description.

Gene Name	Description
APOH	Beta-2-glycoprotein 1 precursor (Beta-2-glycoprotein I) (Apolipoprotein H) (Apo-H) (B2GPI) (Beta(2)GPI) (Activated protein C- binding protein) (APC inhibitor) (Anticardiolipin cofactor); Binds to various kinds of negatively charged substances such as heparin, phospholipids, and dextran sulfate. May prevent activation of the intrinsic blood coagulation cascade by binding to phospholipids on the surface of damaged cells (345 aa)
TCN2	Transcobalamin-2 precursor (Transcobalamin II) (TCII) (TC II); Primary vitamin B12-binding and transport protein. Delivers cobalamin to cells (427 aa)
TUBB1	Tubulin beta-1 chain; Tubulin is the major constituent of microtubules. It binds two moles of GTP, one at an exchangeable site on the beta chain and one at a non-exchangeable site on the alpha-chain (By similarity) (451 aa)
LBP	Lipopolysaccharide-binding protein precursor (LBP); Binds to the lipid A moiety of bacterial lipopolysaccharides (LPS), a glycolipid present in the outer membrane of all Gram-negative bacteria. The LBP/LPS complex seems to interact with the CD14 receptor (481 aa)
PLAT	Tissue-type plasminogen activator precursor (EC 3.4.21.68) (tPA) (t- PA) (t-plasminogen activator) (Alteplase) (Reteplase) [Contains- Tissue-type plasminogen activator chain A; Tissue-type plasminogen activator chain B]; Converts the abundant, but inactive, zymogen plasminogen to plasmin by hydrolyzing a single Arg-Val bond in plasminogen. By controlling plasmin-mediated proteolysis, it plays an important role in tissue remodeling and degradation, in cell migration and many other physiopathological events. Play a direct role in facilitating neuronal migration (562 aa)
RETN	Resistin precursor (Cysteine-rich secreted protein FIZZ3) (Adipose tissue-specific secretory factor) (ADSF) (C/EBP-epsilon-regulated myeloid-specific secreted cysteine-rich protein) (Cysteine-rich secreted protein A12-alpha-like 2); Hormone that seems to suppress insulin ability to stimulate glucose uptake into adipose cells. Potentially links obesity to diabetes (108 aa)
SULT2A1	Bile salt sulfotransferase (EC 2.8.2.14) (Hydroxysteroid Sulfotransferase) (HST) (Dehydroepiandrosterone sulfotransferase) (DHEA-ST) (ST2) (ST2A3); Catalyzes the sulfation of steroids and bile acids in the liver and adrenal glands (285 aa)
HGF	Hepatocyte growth factor precursor (Scatter factor) (SF) (Hepatopoeitin-A) [Contains- Hepatocyte growth factor alpha chain; Hepatocyte growth factor beta chain]; HGF is a potent mitogen for mature parenchymal hepatocyte cells, seems to be an hepatotrophic factor, and acts as growth factor for a broad spectrum of tissues and cell types. It has no detectable protease activity (728 aa)
PON2	Serum paraoxonase/arylesterase 2 (EC 3.1.1.2) (EC 3.1.8.1) (PON 2) (Serum aryldialkylphosphatase 2) (A-esterase 2) (Aromatic esterase 2); Hydrolyzes the toxic metabolites of a variety of organophosphorus insecticides. Capable of hydrolyzing a broad spectrum of organophosphate substrates and a number of aromatic carboxylic acid esters (By similarity). Has antioxidant activity. Is not associated with high density lipoprotein. Prevents LDL lipid peroxidation, reverses the oxidation of mildly oxidized LDL, and inhibits the ability of MM-LDL to induce monocyte chemotaxis (354 aa)
SERPINE1	Plasminogen activator inhibitor 1 precursor (PAI-1) (Endothelial plasminogen activator inhibitor) (PAI); This inhibitor acts as 'bait' for tissue plasminogen activator, urokinase, and protein C. Its rapid interaction with TPA may function as a major control point in the regulation of fibrinolysis (402 aa)

MGP	Matrix Gla-protein precursor (MGP) (Cell growth-inhibiting gene 36 protein); Associates with the organic matrix of bone and cartilage. Thought to act as an inhibitor of bone formation (103 aa)
P11	Placental protein 11 precursor (EC 3.4.21.-) (PP11); Probable serine protease (369 aa)
GNB3	Guanine nucleotide-binding protein G(I)/G(S)/G(T) subunit beta 3 (Transducin beta chain 3); Guanine nucleotide-binding proteins (G proteins) are involved as a modulator or transducer in various transmembrane signaling systems. The beta and gamma chains are required for the GTPase activity, for replacement of GDP by GTP, and for G protein- effector interaction (340 aa)
ENSG00000111980	NG,NG-dimethylarginine dimethylaminohydrolase 2 (EC 3.5.3.18) (Dimethylargininase-2) (Dimethylarginine dimethylaminohydrolase 2) (DDAHII) (DDAH-2) (S-phase protein) (Protein G6a); Hydrolyzes N(G),N(G)-dimethyl-L-arginine (ADMA) and N(G)-monomethyl-L-arginine (MMA) which act as inhibitors of NOS. Has therefore a role in nitric oxide generation (285 aa)
PPP3CC	Serine/threonine-protein phosphatase 2B catalytic subunit gamma isoform (EC 3.1.3.16) (Calmodulin-dependent calcineurin A subunit gamma isoform) (Calcineurin, testis-specific catalytic subunit) (CAM- PRP catalytic subunit); Calcium-dependent, calmodulin-stimulated protein phosphatase. This subunit may have a role in the calmodulin activation of calcineurin (512 aa)
APOE	Apolipoprotein E precursor (Apo-E); Mediates the binding, internalization, and catabolism of lipoprotein particles. It can serve as a ligand for the LDL (apo B/E) receptor and for the specific apo-E receptor (chylomicron remnant) of hepatic tissues (317 aa)
ACE2	Angiotensin-converting enzyme 2 precursor (EC 3.4.17.-) (ACE-related carboxypeptidase) (Angiotensin-converting enzyme homolog) (ACEH); Carboxypeptidase which converts angiotensin I to angiotensin 1-9, a peptide of unknown function, and angiotensin II to angiotensin 1-7, a vasodilator. Also able to hydrolyze apelin- 13 and dynorphin-13 with high efficiency. May be an important regulator of heart function. In case of human coronaviruses SARS and HCoV-NL63 infections, serve as functional receptor for the spike glycoprotein of both coronaviruses (805 aa)
CRP	C-reactive protein precursor [Contains- C-reactive protein(1-205)]; Displays several functions associated with host defense- it promotes agglutination, bacterial capsular swelling, phagocytosis and complement fixation through its calcium-dependent binding to phosphorylcholine. Can interact with DNA and histones and may scavenge nuclear material released from damaged circulating cells (224 aa)
APCS	Serum amyloid P-component precursor (SAP) (9.5S alpha-1-glycoprotein) [Contains- Serum amyloid P-component(1-203)]; Can interact with DNA and histones and may scavenge nuclear material released from damaged circulating cells. May also function as a calcium-dependent lectin (223 aa)
IL6	Interleukin-6 precursor (IL-6) (B-cell stimulatory factor 2) (BSF-2) (Interferon beta-2) (Hybridoma growth factor) (CTL differentiation factor) (CDF); IL-6 is a cytokine with a wide variety of biological functions- it plays an essential role in the final differentiation of B-cells into Ig-secreting cells, it induces myeloma and plasmacytoma growth, it induces nerve cells differentiation, in hepatocytes it induces acute phase reactants (212 aa)
CHAD	Chondroadherin precursor (Cartilage leucine-rich protein); Promotes attachment of chondrocytes, fibroblasts, and osteoblasts. This binding is mediated (at least for chondrocytes and fibroblasts) by the integrin alpha(2)beta(1). May play an important role in the regulation of chondrocyte growth and proliferation (By similarity) (359 aa)
TTPA	Alpha-tocopherol transfer protein (Alpha-TTP); Binds alpha-tocopherol and enhances its transfer between separate membranes (278 aa)

SORL1	Sortilin-related receptor precursor (Sorting protein-related receptor containing LDLR class A repeats) (SorLA) (SorLA-1) (Low-density lipoprotein receptor relative with 11 ligand-binding repeats) (LDLR relative with 11 ligand-binding repeats) (LR11); Likely to be a multifunctional endocytic receptor, that may be implicated in the uptake of lipoproteins and of proteases. Binds LDL, the major cholesterol-carrying lipoprotein of plasma, and transports it into cells by endocytosis. Binds the receptor- associated protein (RAP). Could play a role in cell-cell interaction (2214 aa)
PSMA6	Proteasome subunit alpha type 6 (EC 3.4.25.1) (Proteasome iota chain) (Macropain iota chain) (Multicatalytic endopeptidase complex iota chain) (27 kDa prosomal protein) (PROS-27) (p27K); The proteasome is a multicatalytic proteinase complex which is characterized by its ability to cleave peptides with Arg, Phe, Tyr, Leu, and Glu adjacent to the leaving group at neutral or slightly basic pH. The proteasome has an ATP-dependent proteolytic activity (246 aa)
ITGB3	Integrin beta-3 precursor (Platelet membrane glycoprotein IIIa) (GPIIIa) (CD61 antigen); Integrin alpha-V/beta-3 is a receptor for cytotactin, fibronectin, laminin, matrix metalloproteinase-2, osteopontin, osteomodulin, prothrombin, thrombospondin, vitronectin and von Willebrand factor. Integrin alpha-IIb/beta-3 is a receptor for fibronectin, fibrinogen, plasminogen, prothrombin, thrombospondin and vitronectin. Integrins alpha-IIb/beta-3 and alpha-V/beta-3 recognize the sequence R-G-D in a wide array of ligands. Integrin alpha-IIb/beta-3 recognizes the sequence H-H-L-G-G-G-A-K-Q-A-G-D- [...] (788 aa)
MARS	Methionyl-tRNA synthetase, cytoplasmic (EC 6.1.1.10) (Methionine--tRNA ligase) (MetRS) (900 aa)
NOTCH3	Neurogenic locus notch homolog protein 3 precursor (Notch 3) [Contains- Notch 3 extracellular truncation; Notch 3 intracellular domain]; Functions as a receptor for membrane-bound ligands Jagged1, Jagged2 and Delta1 to regulate cell-fate determination. Upon ligand activation through the released notch intracellular domain (NICD) it forms a transcriptional activator complex with RBP-J kappa and activates genes of the enhancer of split locus. Affects the implementation of differentiation, proliferation and apoptotic programs (By similarity) (2321 aa)
CNTN3	Contactin-3 precursor (Brain-derived immunoglobulin superfamily protein 1) (BIG-1) (Plasmacytoma-associated neuronal glycoprotein); Contactins mediate cell surface interactions during nervous system development. Has some neurite outgrowth-promoting activity (By similarity) (1028 aa)
SELP	P-selectin precursor (Granule membrane protein 140) (GMP-140) (PADGEM) (Leukocyte-endothelial cell adhesion molecule 3) (LECAM3) (CD62P antigen); Ca(2+)-dependent receptor for myeloid cells that binds to carbohydrates on neutrophils and monocytes. Mediates the interaction of activated endothelial cells or platelets with leukocytes. The ligand recognized is sialyl-Lewis X. Mediates rapid rolling of leukocyte rolling over vascular surfaces during the initial steps in inflammation through interaction with PSGL1 (830 aa)
LCT	Lactase-phlorizin hydrolase precursor (Lactase-glycosylceramidase) [Includes- Lactase (EC 3.2.1.108); Phlorizin hydrolase (EC 3.2.1.62)]; LPH splits lactose in the small intestine (1927 aa)
F11	Coagulation factor XI precursor (EC 3.4.21.27) (Plasma thromboplastin antecedent) (PTA) (FXI) [Contains- Coagulation factor XIa heavy chain; Coagulation factor XIa light chain]; Factor XI triggers the middle phase of the intrinsic pathway of blood coagulation by activating factor IX (625 aa)
ADD1	adducin 1 (alpha) isoform b; Membrane-cytoskeleton-associated protein that promotes the assembly of the spectrin-actin network. Binds to calmodulin (768 aa)
GARS	Glycyl-tRNA synthetase (EC 6.1.1.14) (Glycine--tRNA ligase) (GlyRS) (749 aa)
NPC1	Niemann-Pick C1 protein precursor; Involved in the intracellular trafficking of cholesterol. May play a role in vesicular trafficking in glia, a process that may be crucial for maintaining the structural and functional integrity of nerve terminals

	(1278 aa)
AGTR1	Type-1 angiotensin II receptor (AT1) (AT1AR) (AT1BR); Receptor for angiotensin II. Mediates its action by association with G proteins that activate a phosphatidylinositol- calcium second messenger system (359 aa)
ACE	Angiotensin-converting enzyme, somatic isoform precursor (EC 3.4.15.1) (Dipeptidyl carboxypeptidase I) (Kininase II) (CD143 antigen) [Contains- Angiotensin-converting enzyme, somatic isoform, soluble form]; Converts angiotensin I to angiotensin II by release of the terminal His-Leu, this results in an increase of the vasoconstrictor activity of angiotensin. Also able to inactivate bradykinin, a potent vasodilator. Has also a glycosidase activity which releases GPI-anchored proteins from the membrane by cleaving the mannose linkage in the GPI moiety (1306 aa)
PTX3	Pentraxin-related protein PTX3 precursor (Pentaxin-related protein PTX3) (Tumor necrosis factor-inducible protein TSG-14); Plays a role in the regulation of innate resistance to pathogens, inflammatory reactions, possibly clearance of self- components and female fertility (By similarity) (381 aa)
PF4	Platelet factor 4 precursor (PF-4) (CXCL4) (Oncostatin A) (Iroplact); Released during platelet aggregation. Neutralizes the anticoagulant effect of heparin because it binds more strongly to heparin than to the chondroitin-4-sulfate chains of the carrier molecule. Chemotactic for neutrophils and monocytes. Inhibits endothelial cell proliferation, the short form is a more potent inhibitor than the longer form (101 aa)
ANXA5	Annexin A5 (Annexin V) (Lipocortin V) (Endonexin II) (Calphobindin I) (CBP-I) (Placental anticoagulant protein I) (PAP-I) (PP4) (Thromboplastin inhibitor) (Vascular anticoagulant-alpha) (VAC-alpha) (Anchorin CII); This protein is an anticoagulant protein that acts as an indirect inhibitor of the thromboplastin-specific complex, which is involved in the blood coagulation cascade (320 aa)
ITGA2	Integrin alpha-2 precursor (Platelet membrane glycoprotein Ia) (GPIa) (Collagen receptor) (VLA-2 alpha chain) (CD49b antigen); Integrin alpha-2/beta-1 is a receptor for laminin, collagen, collagen C-propeptides, fibronectin and E-cadherin. It recognizes the proline-hydroxylated sequence G-F-P-G-E-R in collagen. It is responsible for adhesion of platelets and other cells to collagens, modulation of collagen and collagenase gene expression, force generation and organization of newly synthesized extracellular matrix (1181 aa)
NOS3	Nitric-oxide synthase, endothelial (EC 1.14.13.39) (EC-NOS) (NOS type III) (NOSIII) (Endothelial NOS) (eNOS) (Constitutive NOS) (cNOS); Produces nitric oxide (NO) which is implicated in vascular smooth muscle relaxation through a cGMP-mediated signal transduction pathway. NO mediates vascular endothelial growth factor (VEGF)-induced angiogenesis in coronary vessels and promotes blood clotting through the activation of platelets (1203 aa)
EEF1A2	Elongation factor 1-alpha 2 (EF-1-alpha-2) (Elongation factor 1 A-2) (eEF1A-2) (Statin S1); This protein promotes the GTP-dependent binding of aminoacyl-tRNA to the A-site of ribosomes during protein biosynthesis (463 aa)
SERPINB2	Plasminogen activator inhibitor 2 precursor (PAI-2) (Placental plasminogen activator inhibitor) (Monocyte Arg-serpin) (Urokinase inhibitor); Inhibits urokinase-type plasminogen activator. The monocyte derived PAI-2 is distinct from the endothelial cell- derived PAI-1 (415 aa)
CYP7A1	Cytochrome P450 7A1 (Cholesterol 7-alpha-monooxygenase) (CYPVII) (EC 1.14.13.17) (Cholesterol 7-alpha-hydroxylase) (504 aa)
GPR25	Probable G-protein coupled receptor 25; Orphan receptor (361 aa)
APLN	apelin, AGTRL1 ligand (APLN), mRNA (122 aa)
FGB	Fibrinogen beta chain precursor [Contains- Fibrinopeptide B]; Fibrinogen has a double function- yielding monomers that polymerize into fibrin and acting as a

	cofactor in platelet aggregation (491 aa)
FGA	Fibrinogen alpha chain precursor [Contains- Fibrinopeptide A]; Fibrinogen has a double function- yielding monomers that polymerize into fibrin and acting as a cofactor in platelet aggregation (866 aa)
OLR1	Oxidized low-density lipoprotein receptor 1 (Ox-LDL receptor 1) (Lectin-type oxidized LDL receptor 1) (Lectin-like oxidized LDL receptor 1) (Lectin-like oxLDL receptor 1) (LOX-1) (hLOX-1) [Contains- Oxidized low-density lipoprotein receptor 1, soluble for; Receptor that mediates the recognition, internalization and degradation of oxidatively modified low density lipoprotein (oxLDL) by vascular endothelial cells. OxLDL is a marker of atherosclerosis that induces vascular endothelial cell activation and dysfunction, resulting in pro-inflammatory responses, pro- oxidative conditions and a [...] (273 aa)
TPMT	Thiopurine S-methyltransferase (EC 2.1.1.67) (Thiopurine methyltransferase); Catalyzes the S-methylation of thiopurine drugs such as 6-mercaptopurine (245 aa)
STAB1	Stabilin-1 precursor (Fasciclin, EGF-like, laminin-type EGF-like and link domain-containing scavenger receptor 1) (FEEL-1) (MS-1 antigen); Acts as a scavenger receptor for acetylated low density lipoprotein. Binds to both Gram-positive and Gram-negative bacteria and may play a role in defense against bacterial infection. When inhibited in endothelial tube formation assays, there is a marked decrease in cell-cell interactions, suggesting a role in angiogenesis. Involved in the delivery of newly synthesized CHID1/SI-CLP from the biosynthetic compartment to the endosomal/lysosomal system (2570 aa)
ABO	Histo-blood group ABO system transferase (NAGAT) [Includes- Glycoprotein-fucosylgalactoside alpha-N- acetylgalactosaminyltransferase (EC 2.4.1.40) (Fucosylglycoprotein alpha-N-acetylgalactosaminyltransferase) (Histo-blood group A transferase) (A transfera; This protein is the basis of the ABO blood group system. The histo-blood group ABO involves three carbohydrate antigens-A, B, and H. A, B, and AB individuals express a glycosyltransferase activity that converts the H antigen to the A antigen (by addition of UDP-GalNAc) or to the B antigen (by addition of UDP-Gal), whereas O individu [...] (353 aa)
LPA	Apolipoprotein(a) precursor (EC 3.4.21.-) (Apo(a)) (Lp(a)); Apo(a) is the main constituent of lipoprotein(a) (Lp(a)). It has serine proteinase activity and is able of autoproteolysis. Inhibits tissue-type plasminogen activator 1. Lp(a) may be a ligand for megalin/Gp 330 (2040 aa)
PGA3	pepsinogen 3, group I (388 aa)
GP1BA	Platelet glycoprotein Ib alpha chain precursor (Glycoprotein Ibalpha) (GP-Ib alpha) (GPIbA) (GPIb-alpha) (Antigen CD42b-alpha) (CD42b antigen) [Contains- Glycocalicin]; GP-Ib, a surface membrane protein of platelets, participates in the formation of platelet plugs by binding to the A1 domain of vWF, which is already bound to the subendothelium (627 aa)
CD24	Signal transducer CD24 precursor; Modulates B-cell activation responses. Signaling could be triggered by the binding of a lectin-like ligand to the CD24 carbohydrates, and transduced by the release of second messengers derived from the GPI-anchor. Promotes AG-dependent proliferation of B-cells, and prevents their terminal differentiation into antibody-forming cells (80 aa)
SERPINA5	Plasma serine protease inhibitor precursor (PCI) (Protein C inhibitor) (Serpin A5) (Plasminogen activator inhibitor 3) (PAI-3) (PAI3) (Acrosomal serine protease inhibitor); Inhibits activated protein C as well as plasminogen activators (406 aa)
TBCD	Tubulin-specific chaperone D (Tubulin-folding cofactor D) (Beta- tubulin cofactor D) (tfcD) (SSD-1); Tubulin-folding protein; involved in the first step of the tubulin folding pathway. Modulates microtubule dynamics by capturing GTP-bound beta-tubulin (TUBB) (999 aa)

KRTAP17-1	Keratin-associated protein 17-1 (Keratin-associated protein 16.1); In the hair cortex, hair keratin intermediate filaments are embedded in an interfilamentous matrix, consisting of hair keratin-associated proteins (KRTAP), which are essential for the formation of a rigid and resistant hair shaft through their extensive disulfide bond cross-linking with abundant cysteine residues of hair keratins. The matrix proteins include the high- sulfur and high-glycine-tyrosine keratins (105 aa)
FGG	Fibrinogen gamma chain precursor; Fibrinogen has a double function- yielding monomers that polymerize into fibrin and acting as a cofactor in platelet aggregation (453 aa)
CS	Citrate synthase, mitochondrial precursor (EC 2.3.3.1) (466 aa)
APOD	Apolipoprotein D precursor (Apo-D) (ApoD); APOD occurs in the macromolecular complex with lecithin- cholesterol acyltransferase. It is probably involved in the transport and binding of bilin. Appears to be able to transport a variety of ligands in a number of different contexts (189 aa)
PDE4D	cAMP-specific 3',5'-cyclic phosphodiesterase 4D (EC 3.1.4.17) (DPDE3) (PDE43); Regulates the levels of cAMP in the cell (809 aa)
GPT	Alanine aminotransferase 1 (EC 2.6.1.2) (ALT1) (Glutamic--pyruvic transaminase 1) (GPT 1) (Glutamic--alanine transaminase 1); Participates in cellular nitrogen metabolism and also in liver gluconeogenesis starting with precursors transported from skeletal muscles (496 aa)
PDE5A	cGMP-specific 3',5'-cyclic phosphodiesterase (EC 3.1.4.35) (CGB-PDE) (cGMP-binding cGMP-specific phosphodiesterase); Plays a role in signal transduction by regulating the intracellular concentration of cyclic nucleotides. This phosphodiesterase catalyzes the specific hydrolysis of cGMP to 5'- GMP (875 aa)
PC	Pyruvate carboxylase, mitochondrial precursor (EC 6.4.1.1) (Pyruvic carboxylase) (PCB); Pyruvate carboxylase catalyzes a 2-step reaction, involving the ATP-dependent carboxylation of the covalently attached biotin in the first step and the transfer of the carboxyl group to pyruvate in the second. Catalyzes in a tissue specific manner, the initial reactions of glucose (liver, kidney) and lipid (adipose tissue, liver, brain) synthesis from pyruvate (1178 aa)
HP	Haptoglobin precursor [Contains- Haptoglobin alpha chain; Haptoglobin beta chain]; Haptoglobin combines with free plasma hemoglobin, preventing loss of iron through the kidneys and protecting the kidneys from damage by hemoglobin, while making the hemoglobin accessible to degradative enzymes (405 aa)
ADD3	Gamma-adducin (Adducin-like protein 70); Membrane-cytoskeleton-associated protein that promotes the assembly of the spectrin-actin network. Binds to calmodulin (706 aa)
F5	Coagulation factor V precursor (Activated protein C cofactor) [Contains- Coagulation factor V heavy chain; Coagulation factor V light chain]; Central regulator of hemostasis. It serves as a critical cofactor for the prothrombinase activity of factor Xa that results in the activation of prothrombin to thrombin (2230 aa)
MT-ND1	NADH-ubiquinone oxidoreductase chain 1 (EC 1.6.5.3) (NADH dehydrogenase subunit 1); Core subunit of the mitochondrial membrane respiratory chain NADH dehydrogenase (Complex I) that is believed to belong to the minimal assembly required for catalysis. Complex I functions in the transfer of electrons from NADH to the respiratory chain. The immediate electron acceptor for the enzyme is believed to be ubiquinone (By similarity) (318 aa)
MT-ND4	NADH-ubiquinone oxidoreductase chain 4 (EC 1.6.5.3) (NADH dehydrogenase subunit 4); Core subunit of the mitochondrial membrane respiratory chain NADH dehydrogenase (Complex I) that is believed to belong to the minimal assembly required for catalysis. Complex I functions in the transfer of electrons from NADH to the respiratory chain. The immediate electron acceptor for the

	enzyme is believed to be ubiquinone (By similarity) (459 aa)
FLG	Filaggrin; Aggregates keratin intermediate filaments and promotes disulfide-bond formation among the intermediate filaments during terminal differentiation of mammalian epidermis (4061 aa)
MMP9	Matrix metalloproteinase-9 precursor (EC 3.4.24.35) (MMP-9) (92 kDa type IV collagenase) (92 kDa gelatinase) (Gelatinase B) (GELB) [Contains- 67 kDa matrix metalloproteinase-9; 82 kDa matrix metalloproteinase-9]; May play an essential role in local proteolysis of the extracellular matrix and in leukocyte migration. Could play a role in bone osteoclastic resorption. Cleaves KiSS1 at a Gly-|-Leu bond (707 aa)
LCN2	Neutrophil gelatinase-associated lipocalin precursor (NGAL) (p25) (25 kDa alpha-2-microglobulin-related subunit of MMP-9) (Lipocalin-2) (Oncogene 24p3); Transport of small lipophilic substances (Potential) (200 aa)
ALOX5	Arachidonate 5-lipoxygenase (EC 1.13.11.34) (5-lipoxygenase) (5-LO) (674 aa)
F7	Coagulation factor VII precursor (EC 3.4.21.21) (Serum prothrombin conversion accelerator) (SPCA) (Proconvertin) (Eptacog alfa) [Contains- Factor VII light chain; Factor VII heavy chain]; Initiates the extrinsic pathway of blood coagulation. Serine protease that circulates in the blood in a zymogen form. Factor VII is converted to factor VIIa by factor Xa, factor XIIa, factor IXa, or thrombin by minor proteolysis. In the presence of tissue factor and calcium ions, factor VIIa then converts factor X to factor Xa by limited proteolysis. Factor VIIa will also convert factor IX to factor I [...] (466 aa)
TNF	Tumor necrosis factor precursor (TNF-alpha) (Tumor necrosis factor ligand superfamily member 2) (TNF-a) (Cachectin) [Contains- Tumor necrosis factor, membrane form; Tumor necrosis factor, soluble form]; Cytokine that binds to TNFRSF1A/TNFR1 and TNFRSF1B/TNFBR. It is mainly secreted by macrophages and can induce cell death of certain tumor cell lines. It is potent pyrogen causing fever by direct action or by stimulation of interleukin-1 secretion and is implicated in the induction of cachexia, Under certain conditions it can stimulate cell proliferation and induce cell differentiation (233 aa)
MTHFR	Methylenetetrahydrofolate reductase (EC 1.5.1.20); Catalyzes the conversion of 5,10- methylenetetrahydrofolate to 5-methyltetrahydrofolate, a co- substrate for homocysteine remethylation to methionine (697 aa)
CST3	Cystatin-C precursor (Cystatin-3) (Neuroendocrine basic polypeptide) (Gamma-trace) (Post-gamma-globulin); As an inhibitor of cysteine proteinases, this protein is thought to serve an important physiological role as a local regulator of this enzyme activity (146 aa)
GK	Glycerol kinase (EC 2.7.1.30) (ATP-glycerol 3-phosphotransferase) (Glycerokinase) (GK); Key enzyme in the regulation of glycerol uptake and metabolism (553 aa)
KL	Klotho precursor (EC 3.2.1.31) [Contains- Klotho peptide]; May have weak glycosidase activity towards glucuronylated steroids. However, it lacks essential active site Glu residues at positions 239 and 872, suggesting it may be inactive as a glycosidase in vivo. May be involved in the regulation of calcium and phosphorus homeostasis by inhibiting the synthesis of active vitamin D (By similarity) (1012 aa)
EPHX2	Epoxide hydrolase 2 (EC 3.3.2.10) (Soluble epoxide hydrolase) (SEH) (Epoxide hydratase) (Cytosolic epoxide hydrolase) (CEH); Acts on epoxides (alkene oxides, oxiranes) and arene oxides. Plays a role in xenobiotic metabolism by degrading potentially toxic epoxides. Also determines steady-state levels of physiological mediators. Has low phosphatase activity (555 aa)

ALOX5AP	Arachidonate 5-lipoxygenase-activating protein (FLAP) (MK-886-binding protein); Seems to be required for the activation of 5-LO (5- lipoxygenase). Could play an essential role in the transfer of arachidonic acid to 5-LO. Binds to MK-886, a compound that blocks the biosynthesis of leukotrienes (161 aa)
TARS	Threonyl-tRNA synthetase, cytoplasmic (EC 6.1.1.3) (Threonine--tRNA ligase) (ThrRS) (723 aa)
ITGA2B	Integrin alpha-IIb precursor (Platelet membrane glycoprotein IIb) (GPalpha IIb) (GPIIb) (CD41 antigen) [Contains- Integrin alpha-IIb heavy chain; Integrin alpha-IIb light chain]; Integrin alpha-IIb/beta-3 is a receptor for fibronectin, fibrinogen, plasminogen, prothrombin, thrombospondin and vitronectin. It recognizes the sequence R-G-D in a wide array of ligands. It recognizes the sequence H-H-L-G-G-G-A-K-Q-A-G-D-V in fibrinogen gamma chain. Following activation integrin alpha- IIb/beta-3 brings about platelet/platelet interaction through binding of soluble fibrinogen. This step leads [...] (1039 aa)
F2	Prothrombin precursor (EC 3.4.21.5) (Coagulation factor II) [Contains- Activation peptide fragment 1; Activation peptide fragment 2; Thrombin light chain; Thrombin heavy chain]; Thrombin, which cleaves bonds after Arg and Lys, converts fibrinogen to fibrin and activates factors V, VII, VIII, XIII, and, in complex with thrombomodulin, protein C. Functions in blood homeostasis, inflammation and wound healing (622 aa)