INTEGRATION OF TEXT MINING WITH SYSTEMS BIOLOGY PROVIDES NEW INSIGHT INTO THE PATHOGENESIS OF DIABETIC NEUROPATHY

by

Junguk Hur

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2010

Doctoral Committee:

Professor Eva L. Feldman, Co-Chair
Professor Hosagrahar V. Jagadish, Co-Chair
Professor Matthias Kretzler
Research Assistant Professor Maureen Sartor
Professor David J. States, University of Texas

DEDICATION

To my family

ACKNOWLEDGMENTS

Over the past few years, I have been tremendously fortunate to have the company and mentorship of the most wonderful and smartest scientists I know. My advisor, Prof. Eva Feldman, guided me through my graduate studies with constant support, encouragement, enthusiasm and infinite patience. I would like to thank her for being a mom in my academic life and raising me to become a better scientist. I would also like to thank my co-advisors Prof. H. V. Jagadish and Prof. David States. They provided sound advice and inspiration that have been instrumental in my Ph.D. study. I am also very grateful to Prof. Matthias Kretzler and Prof. Maureen Sartor for being an active part of my committee as well as for their continuous encouragement and guidance of my work.

I would also like to thank Prof. Daniel Burns, Prof. Margit Burmeister, Prof. Gil Omenn and Prof. Brain Athey from the Bioinformatics Program, who have been very generous with their support and advice about academic life. I am also very grateful to Sherry, Julia and Judy, who helped me through the various administrative processes with cheerful and encouraging dispositions. I would also like to thank the Rackham Graduate School for supporting my studies through a prestigious Rackham Pre-doctoral Fellowship.

I was very fortunate to have the unique opportunity of interacting with different research groups on campus spanning biology and bioinformatics. I have learnt a lot from my interactions with my fellow group members, who have provided me with valuable experiences both within and outside of work. I would like to particularly thank Dr. Kelli Sullivan for her continuous encouragement and invaluable guidance, which has been a great asset to my graduate life. I would also like to thank the Geek bay members, Dr.

Adam Schuyler, Manjusha Pande, Tim Wiggin and George Zhao in the Feldman lab for their support and invaluable discussions. I am also very grateful to my fellow researchers in the Feldman lab: Bhumsoo Kim, Claudia Figueroa-Romero, Stacey Sakowski, Simon Lunn, Andrea Vincent, Yu Hong, Carey Backus, Lisa McLean, Lisa Baiz, Alex Oh and all other lab members for their biological insights and continuous support. I would also like to thank Ann Randolph, Viji Nair and Eichinger Felix from the Kretzler lab for their invaluable collaboration and discussion.

Finally, my special thanks go to my family and friends who continuously supported me during my studies. Without their support and inspiration, I would not be who I am now. Especially, I would like to thank my wife, Mira Park, for fully supporting me throughout my educational journey, and my daughters, Yoon and Naye (Katelyn), whose existence has been a blessing to me.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Diabetic neuropathy (DN) is the most common complication of diabetes affecting approximately 60% of all diabetic patients leading to significant mortality, morbidity, and poor quality of life. Though more than 50% of patients with DN develop substantial nerve damage prior to noticeable symptoms, no biomarkers for predicting the onset or progression of DN are currently available. Here we present a biomarker discovery platform integrating literature mining and a systems biology approach to identify potential DN biomarkers. A web-based target identification and functional analysis tool, SciMiner (http://jdrf.neurology.med.umich.edu/SciMiner), was developed that identifies targets using a context specific analysis of MEDLINE abstracts and full texts. A comprehensive list of 1,026 targets from diabetes and reactive oxygen species (ROS) related literature was compiled by SciMiner. The expression levels of nine genes, selected from the over-represented ROS-diabetes targets, were measured in the dorsal root ganglia (DRG) of diabetic and non-diabetic DBA/2J mice. Eight genes exhibited significant differential expression and the directions of expression change in six of those genes paralleled enhanced oxidative stress in the DRG, suggesting the involvement of ROS related targets in DN. A microarray analysis was also performed on sural nerve biopsies from two DN patient groups with fast or slow DN progression to identify gene expression profiles related to DN progression. In the fast progressing DN, defense response and inflammatory response related genes were up-regulated, while lipid metabolic process and peroxisome proliferator-activated receptor (PPAR) signaling pathway related genes were down-regulated. We also developed mRNA expression signatures that predict DN progression in humans with a high prediction accuracy. Ridge-regression based predictive models with 14 genes achieved a prediction accuracy of 92% (correct prediction of 11 out of 12 patients). Our results identifying the unique gene signatures of progressive DN and compiling ROS-diabetes targets can facilitate the development of new mechanism-based therapies and predictive biomarkers of DN.

CHAPTER 1

INTRODUCTION

1.1 DIABETIC NEUROPATHY

1.1.1 DIABETES MELLITUS

Diabetes is a metabolic disease in which the body does not produce or properly use insulin, a hormone required to convert sugar into energy for daily life. Approximately twenty-three million children and adults, 7.8% of the population in the United States, have diabetes and the incidence is increasing by 5% per year [1]. An additional fifty-seven million Americans have impaired glucose tolerance, or pre-diabetes, in which blood glucose levels are higher than normal but not high enough to be diagnosed as diabetes [1]. The American Diabetes Association estimated the total annual economic cost of diabetes in 2007 to be $174 billion ($116 billion in medical expenditures, $27 billion in direct diabetes care, $58 billion in treating diabetes-related chronic complications, and $31 billion in excess general medical costs) [1].

Diabetes is typically characterized by elevated blood glucose levels (known as hyperglycemia) with symptoms such as increased urination, increased thirst, unexpected weight change, fatigue and blurred vision. There are two main types of diabetes: Type 1 and Type 2 diabetes. In Type 1 diabetes, often referred to as juvenile diabetes or insulin dependent diabetes mellitus (IDDM), the body cannot produce enough insulin due to a destruction of insulin-creating beta-cells in the pancreas by the body's immune system. Thus, patients with Type 1 diabetes need to be supplied with exogenous insulin. In Type 2 diabetes, often called adult-onset diabetes, patients usually begin by developing insulin resistance, in which cells in the body do not properly respond to insulin. Both types of diabetes will result in collateral damage to nerves and blood vessels that accumulates over years of hyperglycemic conditions, leading to the spectrum of conditions referred to as diabetic complications.

1.1.2 DIABETIC NEUROPATHY

Diabetic complications are the most common cause of renal failure, blindness and amputations, and lead to significant mortality, morbidity, and poor quality of life. The most common complication of diabetes is diabetic neuropathy (DN) affecting approximately 60% of all diabetic patients [2, 3]. Diabetic neuropathy (DN) is characterized by a progressive loss of peripheral nerve axons, resulting in decreased sensation, pain, and eventually complete loss of sensation. Twenty-five years following diagnosis of diabetes, patients have a cumulative risk of 22% for a lower extremity amputation [4], which makes DN the leading cause of non-traumatic amputation in the United States [5].

Diagnosis of DN is based on symptoms and a physical examination, which often reveals a characteristic "stocking and glove" loss of sensation. Patients lose the function of their longest sensory axons first, causing loss of sensation in the feet and hands that progressively moves more proximally along the limbs. Qualitative assessments of sensory loss can be quantified by electrophysiological measures of sensory and motor function, including nerve conduction studies. Changes in such studies are a direct reflection of nerve fiber loss and routinely reveal low or absent sensory responses and slowed sensory and motor nerve conduction velocities (NCV) [6, 7]. These parameters therefore closely correlate with anatomical evidence of decreased myelinated fiber density (MFD) in the sural nerve and decreased intra-epidermal nerve fiber density (IENF) measured from skin biopsies in the lower leg [7-9]. Treatment for DN is currently limited to glycemic control, good foot hygiene (as insensate feet are often injured without the patient's knowledge, leading to chronic wound ulceration) and symptomatic care for pain [10]. While tight glycemic control may slow peripheral nerve deterioration, no treatment is currently available to reverse nerve fiber loss and restore function in DN.

1.1.3 CELLULAR PATHWAYS IMPLICATED IN DN

The current understanding of the underlying mechanisms of DN pathogenesis is far from complete, which hinders not only the development of mechanism-based therapies targeting the genes, proteins and signaling cascades underlying DN but also the identification of potential DN biomarkers. Potential risk factors for DN such as hyperglycemia, hypertension, duration of diabetes, and hyperlipidemia have been studied for their roles in the pathogenesis of DN [10, 11]. Among these factors, hyperglycemia has long been considered the primary risk factor for DN and extensively investigated to elucidate its potential downstream mechanisms. Numerous mechanisms downstream of hyperglycemia have been implicated in the pathogenesis of DN, including increased polyol pathway activity with NAD(P)-redox imbalance [2, 12], oxidative stress [2, 12, 13], mitochondrial dysfunction [14, 15], inflammation [16, 17] and the accumulation of advanced glycosylation endproducts (AGEs) [18].

There is also increasing evidence suggesting that hyperlipidemia substantially contributes to the development of DN [19-21]. In the Eurodiab Trial, a longitudinal study of over 3,000 individuals with type 1 diabetes, dyslipidemia was closely associated with the onset of DN [19, 20]. In our recent retrospective analysis of fast and slow progressing DN, elevated triglycerides was the only clinical parameter that correlated with decreased myelinated fiber density in the sural nerves [21]. Oxidized low density lipoprotein (oxLDL) also strongly correlates with nervous system injury in animal models of high fat induced diabetes [22]. Therefore, lipid-lowering drugs may have potential in the treatment of DN. One such drug, the PPARα agonist fenofibrate demonstrated improved symptoms and slowed the progression of DN as well as other diabetic complications (retinopathy and nephropathy) [23-25].

1.1.4 OXIDATIVE STRESS

Despite this progress, the exact mechanisms underlying the pathogenesis of DN are still not fully understood. Only recently has a link among the various implicated pathways been established that provides a unified mechanism of tissue damage. Each of these pathways directly and indirectly leads to overproduction of reactive oxygen species (ROS) [2, 12]. ROS are highly reactive ions or small molecules including oxygen ions, free radicals and peroxides, and are formed as natural byproducts of cellular energy metabolism. Due to the highly reactive properties of ROS, excessive ROS may cause significant damage to proteins, DNA, RNA and lipids.

All cells have protective mechanisms against ROS; however, under diabetic conditions, these protective mechanisms are overwhelmed due to the substantial increase in ROS, leading to cellular damage and dysfunction [26] . The idea that increased ROS and oxidative stress contribute to the pathogenesis of diabetic complications has led scientists to investigate different oxidative stress pathways [27, 28]. Inhibition of ROS or maintenance of euglycemia restores metabolic and vascular imbalances and blocks both the initiation and progression of complications [29, 30]. Despite the significant implications and extensive research into the role of ROS in diabetes, no comprehensive database regarding ROS-related genes or proteins is currently available.

1.1.5 BIOMARKERS FOR DN PREDICTION

In greater than 50% of patients with DN, there is substantial, irreparable nerve damage prior to the development of noticeable symptoms. The ability to measure specific biomarkers of DN prior to the onset of permanent damage would permit the initiation of aggressive therapy in time to preserve nerve function. Biomarkers that are highly predictive of the development and worsening of diabetic complications are available for diabetic nephropathy [31]. Currently, no biomarkers exist for DN, making it impossible to detect until clinically obvious symptoms appear, at which point irreparable damage has occurred. Our overall goal for this thesis is to develop a biomarker discovery system for DN by integrating literature mining and systems biology.

1.2 BIOMEDICAL LITERATURE MINING

Knowledge obtained through scientific discoveries in biomedical disciplines has been accumulated in the biomedical literature. One key resource for the biomedical literature, the PubMed database maintained by the National Center for Biotechnology Information (NCBI), comprises a vast amount of biomedical articles. Approximately 20 million records from over 25,000 journals are indexed by PubMed today and the volume is rapidly expanding [32]. Due to this rapid growth of published information, it is no longer feasible to keep up to date with all of the new literature manually, even within one’s own research area. The field of text mining, a computer-assisted information extraction from text data, is becoming increasingly important to cope with the increasing volume of electronically available biomedical literature. The goal of biomedical text mining is to aid researchers by having computers read and summarize the literature and efficiently identify publications that are most relevant to the researchers’ interests.

1.2.1 DISCIPLINES OF BIOMEDICAL LITERATURE MINING

There are three main areas of biomedical literature mining [33, 34]: (1) information retrieval (IR), which is the process of identifying relevant papers; (2) entity recognition (ER) or named entity recognition (NER), which is the recognition of biological entities such as genes, proteins, metabolites and chemical compounds within papers; and (3) information extraction (IE), which is the extraction of specific facts from papers. Each of these three main areas constitutes its own research topic and application, but they are also closely related to each other.

1.2.1.1 Information Retrieval

IR, one fundamental discipline of biomedical literature mining, aims to properly identify and retrieve text data relevant to a certain topic from document repositories. The topic could be a pre-defined set of papers or a user-provided query. IR technologies have been widely adopted in the biomedical community and the best-known system is the Entrez query and retrieval system available through PubMed [33, 35]. This system employs two IR methodologies: simple Boolean query searches (logical combinations of terms connected by operators such as ‘AND’, ‘OR’ and ‘NOT’) on indexed documents, and a vector model, which incorporates a frequency-based scoring scheme to calculate document similarity and identify related articles [36]. PubMed is indexed with Medical Subject Headings (MeSH), a controlled vocabulary of over 25,000 terms organized in a hierarchical fashion with 15 top-level categories [37].

1.2.1.2 Entity Recognition

Named Entity Recognition (NER) or simply Entity Recognition (ER) is the most fundamental discipline of biomedical literature mining. ER aims to identify all the instances of named entities such as genes, proteins, chemical compounds and diseases in the biomedical literature. Correct identification of such named entities provides a solid ground for facilitating the retrieval of relevant literature (IR) and the identification of relationships among these entities (IE) [38].

Various methods and algorithms have been applied to ER, which can be categorized into two groups: rule-based approaches and machine learning (ML)-based approaches. The first approach relies on manually devised rules and patterns of systematic variations in names, such as identifying gene symbols with letters followed by numbers, or identifying protein names as those ending with ‘-ase’ or followed by terms like ‘receptor’ and ‘protein’ [33, 39-42]. The ML-based approaches use statistical and probabilistic models that estimate the degree of confidence in making identifications of terms. Various ML approaches, such as the hidden Markov model (HMM) [43, 44], the support vector machine (SVM) [45, 46], and the naïve Bayesian learning (along with decision tree and inductive rule learning) [47] have been applied to biomedical literature mining.

Though the task of ER has the simple goal of identifying biological entities in text, it has been challenging for several reasons. Firstly, there is no complete dictionary for most types of biological named entities. The lack of a complete dictionary makes it impossible for relatively simple text matching algorithms to identify all the entities in the literature. Secondly, different expressions such as synonymous protein names, gene symbols (acronyms) and typographical variants can refer to the same entity (protein or gene) [34]. For example, the terminology referring to 'tumor necrosis factor alpha' can include 'TNFA', 'Tnfa', 'TNF-A', 'TNF-alpha', 'cachectin', 'APC1 protein', 'TNF, monocyte-derived', 'TNF superfamily, member 2' and many other different forms. There are ongoing efforts to standardize and maintain gene symbols and protein names by biomedical science communities and annotation databases such as the HGNC (HUGO Gene Nomenclature Committee) [48] and the National Center for Biotechnology Information (NCBI) Entrez Gene database [49]. Such standardization will substantially improve the accuracy of the literature mining techniques.

Lastly, abbreviations and acronyms of genes are often ambiguous and can refer to different genes, or even non-gene symbols such as experimental methods. For instance, the abbreviation 'PSA' refers to five different human genes in the NCBI Gene database: 'pleiomorphic adenoma gene 1 (PLAG1)', 'kallikrein-related peptidase 3 (KLK3)', 'aminopeptidase puromycin sensitive (NPEPPS)', 'protein S alpha (PROS1)', and 'phosphoserine aminotransferase 1 (PSAT1)'. PSA refers to KLK3 in the context of prostate cancer, while it refers to PSAT1 in the context of amino acid metabolism. Various methods have been introduced to resolve this ambiguity issue in ER using rule and ML-based approaches [43-46, 50, 51]. Chapter 2 of this thesis will introduce a conflict resolving scheme based on the co-occurrence of gene symbols and longer descriptions (names).

1.2.1.3 Information Extraction

IE is an approach to identify pre-defined types of facts such as the relationship between biological entities within the text. Particularly, the identification of physical protein-protein interactions (PPI) has been an active application of IE in the molecular biology domain [46, 52-54]. The Michigan Molecular Interactions database (MiMI, http://mimi.ncibi.org/), the most comprehensive PPI database, incorporates literature-derived PPI information to augment the coverage of PPI in the database [55]. The scope of IE can be extended to 'Knowledge Discovery', which aims at discovering hidden information in the literature and providing new insights for future research [34].

1.2.2 APPLICATIONS OF BIOMEDICAL LITERTURE MINING

Biomedical literature mining has already had a critical impact on several biomedical research areas, and as discussed already, will increase in value for biomedical research in general as the body of published information continues to expand beyond the point of manageability. Some of the areas already impacted by literature mining are summarized below.

1.2.2.1 Functional Annotation of Genes

With the advance of genome sequencing technology, we are now seeing a dramatic increase in the number of available complete genomic sequences. Over 1,000 species have been completely sequenced to date, and more than 5,500 species are currently being sequenced [56]. Traditional functional analysis approaches cannot keep up with the speed of newly sequenced genomes. Therefore, computational approaches relying on protein or DNA sequence similarities with known species have been extensively used to annotate newly sequenced genomes [57-59]. Text mining technologies have also been employed to assist gene annotation. Keywords found in the biomedical literature that are highly associated with protein families, were assigned to the newly identified genes belonging to the same protein families [60]. Gene Ontology (GO) terms identified in the literature were also used in annotating newly identified genes [61, 62].

1.2.2.2 High-throughput Expression Analysis

High-throughput gene expression assay platforms like Microarray and RNA-Seq are extensively used to survey the expression profiles of tens of thousands of genes. Typically, sets of genes (clusters) with similar expression patterns are identified and studied for their coherence in biological functions or common regulatory mechanisms. Text mining approaches are also used to assist array analyses. One such approach is assigning highly relevant terms to clusters based on significant association of the genes and terms within the literature [63, 64]. Another approach is using gene-gene co-citation networks in the literature to explore the possible connections among the genes of interest. PubGene (http://pubgene.org) [65] and BiblioSphere PathwayEdition (Genomatix Software GmbH, Munich, Germany) [66, 67] allows researchers to navigate and visualize the potential functional connections among the genes from microarray results using the co-citation networks constructed from the entire PubMed abstracts.

1.2.2.3 Extracting Protein-Protein Interactions (PPI)

As introduced in Section 1.2.1.3, identifying relationships between biological entities within the text is one of the major disciplines of biomedical literature mining. PPI has been of particular interest to researchers since it forms the basis for the majority of cellular events such as transcriptional regulation and signal transduction [68]. The current PPI networks collectively in all PPI databases are estimated to represent only 10~50% of the complete PPI networks [69, 70]. Thus, identifying potential PPIs from the literature to augment the current PPI networks is an active research area. Various methods based on semantic proposition [71], rules and patterns [72, 73], or support vector machine (SVM) [46, 74] have been developed. The current release of the MiMI database, which provides a comprehensive literature derived PPI information [55], is a good example of this application.

1.2.2.4 Application in this Thesis

Chapter 3 of this thesis will describe how SciMiner, a literature mining tool (Chapter 2), is employed to identify targets related to the reactive oxygen species (ROS) and diabetes context. A subset of the ROS-diabetes targets is evaluated for their biological relevance in diabetic mice.

1.3 SYSTEMS BIOLOGY APPROACH

A significant challenge in the biomedical sciences is to decipher the biological functions of individual genes, pathways, and networks that drive complex phenotypes. Over the last decade, we have witnessed the paradigm shift from trying to understand the isolated functions of individual genes and pathways to trying to piece together the complex interactions of regulatory networks in biological systems. Systems biology is an interdisciplinary research field that focuses on how these complex networks function to control cell physiology and the biology of diseases by integrating multi-level data, such as gene expression, protein expression and metabolite profiles [75, 76]. Analyses of these data via '-omics' (transcriptomics for gene expression, proteomics for protein expression, and metabolomics for metabolite profiles) must be integrated in the setting of a strong bioinformatics infrastructure. This thesis focuses on how bioinformatics systems and transcriptomics can be used to analyze systems-level changes in gene expression that correlate to a disease state.

1.3.1 MICROARRAY AS TRANSCRIPTOMICS TOOL

In the 1990s, DNA microarray technology was introduced to the field of molecular biology, allowing for measurement of the expression of tens of thousands of genes simultaneously on a single glass slide [77, 78]. DNA microarray technology is based on the specific hybridization of cRNA to probes imprinted on the array. Probes, which are oligonucleotides, cDNA, or small fragments of PCR products, are either spotted or synthesized right on the array using photolithographic technology (and available as commercial products from Agilent and Affymetrix). After labeled complementary cRNAs obtained from samples are hybridized to microarrays under high-stringency conditions, laser scanners are used to detect and quantify the intensity of fluorescence of spots on a microarray. With the advance of microarray technology, there has been an explosion of studies using this technology to examine whole genome expression patterns. Transcriptomics, the analysis of the set of all RNA molecules (especially mRNAs) produced in an individual cell or population of a particular cell type [79], has become one of the most prominent fields of study in biomedical science. In Chapter 4 of this thesis, we will use microarrays to survey differences in the transcriptional profiles from patients with different degrees of DN progression.

1.3.2 MICROARRAY ANALYSIS PLATFORM (GENEPATTERN AND CHIPINSPECTOR)

1.3.2.1 GenePattern Analysis Platform

Computational methods for microarray analysis are typically developed as either stand-alone applications or as libraries for a mathematical toolkit such as R (http://www.r-project.org/) or MATLAB (http://www.mathworks.com/). This approach scales poorly with increases in the number of tools and the complexity of the desired analyses. Users must cope with numerous technology maintenance tasks, including locating programs and keeping them up to date, converting between file formats, retaining old versions of analysis applications and entering parameters for each analysis step, as well as self-correction when errors occur. GenePattern (http://www.broadinstitute.org/cancer/software/genepattern/) is an analysis environment that automates some aspects of data analysis, making the process less cumbersome for the user [80]. GenePattern provides the ability to use programs written in Java, Perl, R and MATLAB in an environment with standardized file formats, a common user interface, integrated versioning, and mechanisms for documentation. These programs, called modules, can run automatically as part of analysis pipelines, which reduce analysis time and may be distributed to other users to enable reproduction of the precise computational methods performed [80]. In Chapter 4 of this thesis, GenePattern is used as the primary analysis platform for microarray data using Robust Multi-array Average (RMA), which is a typical probe set-based microarray analysis approach using a quantile normalization method [81].

1.3.2.2 ChipInspector

ChipInspector (CI; version 2.1; Genomatix Software GmbH, Munich, Germany) provides a unique opportunity to investigate the gene expression microarray data at the level of an individual probe, instead of at the level of the Affymetrix probe set [82]. Recent studies estimated that approximately 50~75% of multi-exon genes in human and mouse genomes generate multiple mRNA isoforms through alternative splicing [83-85]. A probe set-based approach that averages signals in a probe set may obscure the resolution of alternative transcripts, leading to increased noise in the analysis [76]. Therefore, keeping the mRNA signal on the transcript level allows identification of genes that change only in the relative abundance of alternative transcripts but without a net overall change in gene expression [76]. This ChipInspector-based approach is expected to provide a reduced level of experimental background noise with enhanced sensitivity. In this thesis, we take an integrative approach by combining RMA- and ChipInspector-based results.

1.3.3 FUNCTIONAL ENRICHMENT ANALYSES

The typical initial results of microarray experiments are lists of genes that are differentially expressed between the sample groups being compared. These lists are then subjected to a functional enrichment analysis that attempts to identify over-represented biological functions or pathways among these genes. Various tools for this purpose have been developed, including the Gene Set Enrichment Analysis (GSEA) (http://www.broadinstitute.org/gsea/) [86, 87], the Database for Annotation, Visualization and Integrated Discovery (DAVID) (http://david.abcc.ncifcrf.gov/) [88, 89], ConceptGen (http://conceptgen.ncibi.org) [90], and LRpath [91]. Using publicly available sources, such as GO [92] and pathways databases including Kyoto Encyclopedia of Genes and

Genomes (KEGG, http://www.genome.jp/kegg/) [93] and Reactome (http://www.reactome.org/) [94], these tools provide statistical significance of enrichment in diverse types of biological categories. Typically, these tools require users to provide a predefined set of genes as input; however, LRpath accepts p-values of all genes in microarrays tested by a statistical significance test (e.g., t-test) and calculates the odds of gene set membership with the significance of differential expression [91].

1.3.4 PREDICTION MODELING USING GENE EXPRESSION PROFILES

Prediction modeling is a process of finding characteristics of an object that predicts its class (discrete) or values (continuous) based on a set of training data [95]. When performing microarray experiments, the object is usually a tissue sample and its characteristics are the observed levels of gene expression. Biomedical applications of prediction modeling include both diagnostic and prognostic uses, such as distinguishing types of cancer or predicting the outcome of a specific cancer [96, 97]. There are many ML- or regression-based classification algorithms available, such as Naïve Bayse, neural network, support vector machine, decision tree, and logistic regression [98, 99]. Their performance varies depending on the data set, but ridge regression-based prediction has demonstrated superior performance in predicting the survival of cancer patients [100, 101] and chronic kidney diseases [102] using microarray expression data. Ridge regression is particularly useful for modeling gene expression data from microarrays, which often have genes with high correlations with each other, referred to as collinearity. Ridge regression handles this collinearity issue by reducing the number of dimensions, which imposes some bias on the regression coefficients. In Chapter 4, we will employ ridge regression-based modeling to classify (predict) the class of DN patients based on the gene expression profiles obtained from microarray analyses of sural nerves.

1.4 SPECIFIC AIMS AND APPROACHES OF THIS THESIS

Our overall goal for this thesis is to develop a biomarker discovery system for DN by integrating literature mining and systems biology. These two approaches will provide unprecedented insight into the ROS-diabetes genes and the large-scale changes in gene expression that occur during DN. This systems-level knowledge will serve as the first step in our effort to identify biomarkers of DN. We have the following three specific aims in this thesis.

1.4.1 AIM1: DEVELOP A LITERATURE MINING TOOL

We hypothesized that an easy-to-use web-based literature mining tool that correctly identifies targets from the vast amount of biomedical literature for a given topic would substantially enhance researchers' understanding of the topic by providing functional summaries of the related targets. We developed SciMiner, a web-based literature mining tool [42]. The target identification is accomplished using comprehensive dictionaries of gene symbols and names along with term expansion rules. Ambiguous symbols are resolved by a unique confidence-scoring scheme based on the co-occurrence of abbreviated symbols and longer descriptions in the same document. SciMiner provides a convenient web-based platform for mining targets (genes and proteins) from the biomedical literature with the capacity for functional enrichment analyses. SciMiner performs well compared to other methods [103-105], but is unique in that it (i) searches full text documents (not just abstracts), (ii) allows users to directly edit the mining results, and (iii) allows comparisons to be made between search results of multiple queries.

1.4.2 AIM2: IDENTIFY REACTIVE OXYGEN SPECIES (ROS) - AND DIABETES-RELATED TARGETS USING SCIMINER

We hypothesized that SciMiner would be able to compile a comprehensive list of ROS-diabetes targets from the biomedical literature. To identify ROS-diabetes targets, SciMiner was applied to ROS-diabetes related literature indexed by ("Reactive Oxygen Species"[MeSH] AND "Diabetes Mellitus"[MeSH]) in PubMed. The identification accuracy was improved by reviewing the sentences in which each target was identified. The collected ROS-diabetes targets were further tested against randomly selected non-ROS-diabetes literature to identify targets significantly over-represented in the ROS-diabetes literature. Functional enrichment analyses were performed on these targets to summarize the biological functions of the ROS-diabetes targets in terms of Gene Ontology (GO) terms and pathways. In order to confirm the biological relevance of the over-represented ROS-diabetes targets, the gene expression levels of nine selected targets were measured in dorsal root ganglia (DRG) from mice with and without diabetes.

1.4.3 AIM3: IDENTIFY GENE EXPRESSION SIGNATURES PREDICTIVE OF DIABETIC NEUROPATHY PROGRESSION

We hypothesized that diabetes directly affects gene expression in peripheral nerves that could be detected by microarray analyses. To identify the gene expression signatures correlated with DN progression, we performed the first high-throughput genome-wide expression study of human sural nerve biopsies obtained from patients with DN. Our laboratory is in possession of a unique repository of human sural nerve biopsies from participants in a large randomized placebo-controlled clinical DN trial testing acetyl-L-carnitine (ALC), which improved neuropathic pain but did not affect nerve conduction velocities and amplitudes in the sural nerves [21, 106]. Sural nerve samples from two groups of patients with either fast progressing or slow/non-progressing DN were surveyed with high-throughput Affymetrix Microarray for their gene expression profiles. A series of bioinformatics analyses were performed to analyze differential gene expression profiles between the two groups and revealed gene networks and pathways that are potentially responsible for the progression of DN. Ridge regression-based computational predictive models using the expression profiles of these genes were then developed to accurately predict the class of DN progression (fast or slow progression).

1.5 THESIS OUTLINE

Chapter 2 presents the development of SciMiner, a web-based literature mining tool for target identification and functional enrichment analysis, and demonstrates the superior performance of SciMiner using the BioCreAtIvE (Critical Assessment of Information Extraction systems in Biology) evaluation system. The basic architecture and implemented text mining techniques are described.

Chapter 3 demonstrates the application of SciMiner to the ROS-diabetes related literature to compile a comprehensive list of ROS-diabetes targets. Identification of the key ROS-diabetes targets and a preliminary evaluation of these targets as potential DN biomarkers are presented.

Chapter 4 presents the first high-throughput genome-wide expression analysis of human sural nerve biopsies obtained from patients with either fast or slow DN progression. Differentially regulated genes and cellular pathways between different rates of DN progression are presented. Computational prediction models for DN progression class (progressing or non-progressing) based on the gene expression profiles of differentially expressed genes are also presented.

Chapter 5 summarizes overall conclusions from the study and presents possible future work.

CHAPTER 2

SCIMINER: WEB-BASED LITERATURE MINING TOOL FOR TARGET IDENTIFICATION AND FUNCTIONAL ENRICHMENT ANALYSIS

2.1 INTRODUCTION

The PubMed database maintained by the National Center for Biotechnology Information (NCBI) is a key resource for biomedical science. It is a large and rapidly expanding data set; more than 20 million records from over 25 thousands journals are indexed by PubMed today. With the increasing volume of the published biomedical literature, text mining has emerged as an increasingly important technology. The goal of biomedical text mining is to aid researchers in identifying relevant information more efficiently by having computers read the literature.

Currently available web-based biomedical text mining tools include EBIMed [104]. ALI BABA [103], and PolySearch [105]. EBIMed provides a simple interface for identifying associations among named entities (genes/proteins, gene ontologies, drug names, and species). Ali Baba visualizes associations in a graphical way. PolySearch provides more than 50 different classes of queries against various types of text, scientific abstract or biomedical databases. Though these tools provide valuable resources, they are limited in that (i) they only access MEDLINE abstracts as their literature data source, (ii) they do not allow users to edit the mining results, and (iii) they are unable to perform comparisons between search results of multiple queries.

Here we present SciMiner, a web-based literature mining tool, which automatically collects MEDLINE records and available full text. Targets (genes and proteins) are extracted and ranked by the number of documents in which they appear. To provide an overall summary of their biological functions, these targets are further analyzed for their enrichments in Gene Ontology terms, pathways, MeSH terms, and protein-protein interaction networks based on external annotation resources. Table 2-1 compares the features of SciMiner and other existing tools.

The performance of target symbol and name identification is assessed using the BioCreAtIvE II (Critical Assessment of Information Extraction systems in Biology) Gene Normalization Task [107]. SciMiner achieved 87.1% recall, 71.3% precision, and 75.8% F-measure. SciMiner's literature mining performance coupled with functional enrichment analyses provides an efficient platform for retrieval and summary of rich biological information from corpora of users' interest.

2.2 IMPLEMENTATION

2.2.1 DOCUMENT RETRIEVAL

Figure 2-1 illustrates the overall workflow how SciMiner processes users' queries to retrieve literature data. Once a query is submitted to SciMiner, it is sent to the NCBI PubMed server (http://www.ncbi.nlm.nih.gov/pubmed/) via Entrez E-utilities (http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html) to retrieve all the resulting PMIDs (PubMed Unique Identifiers). The retrieved PMID list checked against the SciMiner database to determine which documents need to be retrieved and processed. Only the PMIDs that were not previously processed are retrieved and processed.

Table 2-1. Comparison of SciMiner Features with Other Web-based Literature Mining Tools

	Features	SciMiner	EBIMed	Ali Baba	PolySearch
Text Data Source	MEDLINE Abstracts[1]	O	O	O	O
	Full text HTMLs[2]	O	X	X	X
Query Input	Search terms as in PubMed[3]	O	O	O	X
	Structured query[4]	X	X	X	O
	Accepting PMID list as input[5]	O	O	O	X
	Limit of document number[6]	500-Unlimited[a]	10,000	10,000	500-Unlimited
Document processing	Genes and protein recognition[7]	O	O	O	O
	Ambiguity (conflict) resolution[8]	O	O	X	?
	Other named entity recognition[9]	X [b]	drug, species, cells, diseases	GO, drug, species	Drug, disease, metabolite, tissue, cell
	User editability to increase accuracy[10]	O	X	X	X
	Document processed on the fly[11]	O	O	O	O
Filtering	Minimum number of citations per target[12]	O	X	X	O
	Minimum score[13]	O	X	X	O
Enrichment Analysis	Comparison among search results[14]	O	X	X	X
	Functional enrichment (GO, Pathway, MeSH) of queries[15]	O	X	X	Δ
Result	Result notification by email[16]	O	X	X	O
	Downloadable result[17]	O	Graph	X	X
	Results are cached[18]	O	X	X	O
	Highlighted target in abstract[19]	O	O	O	O
	Visualization of interacting targets[20]	Via Cytoscape	O	X	X
	Links to PubMed[21]	O	O	X	O
	Links to journal HTML (publishers) [22]	O	X	X	X
	Links to other databases[23]	NCBI Gene, HGNC, MiMI, QuickGO, KEGG, Reactome, NCBI MeSH	PubMed, MeSH, DrugBank, UniProt	QuickGO, NCBI Taxonomy viewer, DrugBank	PubMed, OMIM, DrugBank, UniProt, HMDB, HPRD, GAD
	Document based summary[24] (Highlight in abstract)	O	O	O	X
Other features	Execution time estimation[25]	O	O	O	X
	Standalone version[26]	O	X	X	X
	Bulk data set[27]	O	X	X	O
	EndNote citation export[28]	O	X	X	X

O: available or supported, X: unavailable or unsupported. Compared features are as follows: **1) MEDLINE Abstracts (**MEDLINE abstracts are used as the source of the text data); **2) Full Text HTMLs** (full text HTML documents are used as the source of the text data if available); **3) Search terms as in PubMed** (tool supports search terms as being used in PubMed query); **4) Structured Query** (available in PolySearch such as "Given X condition, find every Y"); **5) Accepting PMID list as input** (a list of PMIDs can be used as a search query.); **6) Limit of document number** (the maximum number of documents per query); **7) Genes and protein recognition** (tool identifies genes and proteins names and symbols); **8) Ambiguity (conflict) resolution** (conflicting symbols are resolved); **9) Other named entity recognition** (other named entities like GO terms and drug names are identified by text-mining); **10) User editability to increase accuracy** (users can manually edit individual target identification results and additional filters of IGNORE, EXCLUDE, and INCLUDE may be used); **11) Documents processed on the fly** (documents are processed depending on users' queries. Previously processed documents by any previous queries would not need to be reprocessed); **12) Minimum number of citations per target** (the number of associated documents per target in users' search can be specified to show such targets with a specified number of documents associated); **13) Minimum score** (a score thresholds can be specified as an additional filter); **14) Comparison among search results** (comparisons can be performed among different search results (or different queries) in terms of target lists, GO terms, MeSH terms and pathways); **15) Functional enrichment (GO, Pathway, MeSH) of queries** (enriched biological features can be identified by Fisher's exact test in comparisons among different search results); **16) Result notification by email** (an email notice will be sent out to users when the results are ready); **17) Downloadable results** (mining and analysis results are available for download); **18) Results are cached** (processed documents will be kept in database to provide a quicker result in later queries. Users can maintain their search and analysis results in users' account); **19) Highlighted target in abstract** (identified targets are color highlighted. In SciMiner, this is limited to those targets from abstracts); **20) Visualization of interacting targets** (Protein-Protein Interaction (PPI) are visualized in Cytoscape); **21) Links to PubMed** (tool provides a URL link to the PubMed AbstractPlus page for each document); **22) Links to journal HTML** (tool provides a URL link to the full text page for each document); **23) Links to other databases** (tool provides links to external databases including NCBI Entrez Gene, HGNC (HUGO Nomenclature, MiMI (Michigan Molecular Interactions), QuickGO, KEGG (Kyoto Encyclopedia of Genes and Genomes), Reactome, DrugBank, UniProt, HMDB (Human Metabolome Database), HPRD (Human Protein Reference Database), and GAD (Genetic Association Databsae)); **24) Document based summary** (tool provides a document centric summary page for each processed document. SciMiner provides detailed list targets and related information, while others only highlight identified entities in the abstract); **25) Execution time estimation** (tool provides an estimated time for each query and analysis job); **26) Standalone version** (a standalone package is available for download); **27) Bulk data set** (tool provides annotation and gene / protein details in a downloadable format. SciMiner provides some of these data on the public web version. All of the other data are available in the standalone package); **28) User account** (tool allows users to manage their previous search and analysis results. User account is essential since SciMiner allows user to merge and compare multiple search results); **29) EndNote citation export** (tool provides EndNote citation data for every processed document. Users can import these citation data directly from the SciMiner webpage without visiting PubMed or publishers' websites to download such citation files). **a)** the public version of SciMiner has a default limit of 500 documents per query, which can be increased upon request. For standalone version, 20,000 documents per 1GB of RAM are recommended. **b)** GO terms, MeSH terms, pathways, protein-protein interactions are not directly identified from the text data. Instead, SciMiner uses external annotation resources to associate identified targets to these entities.

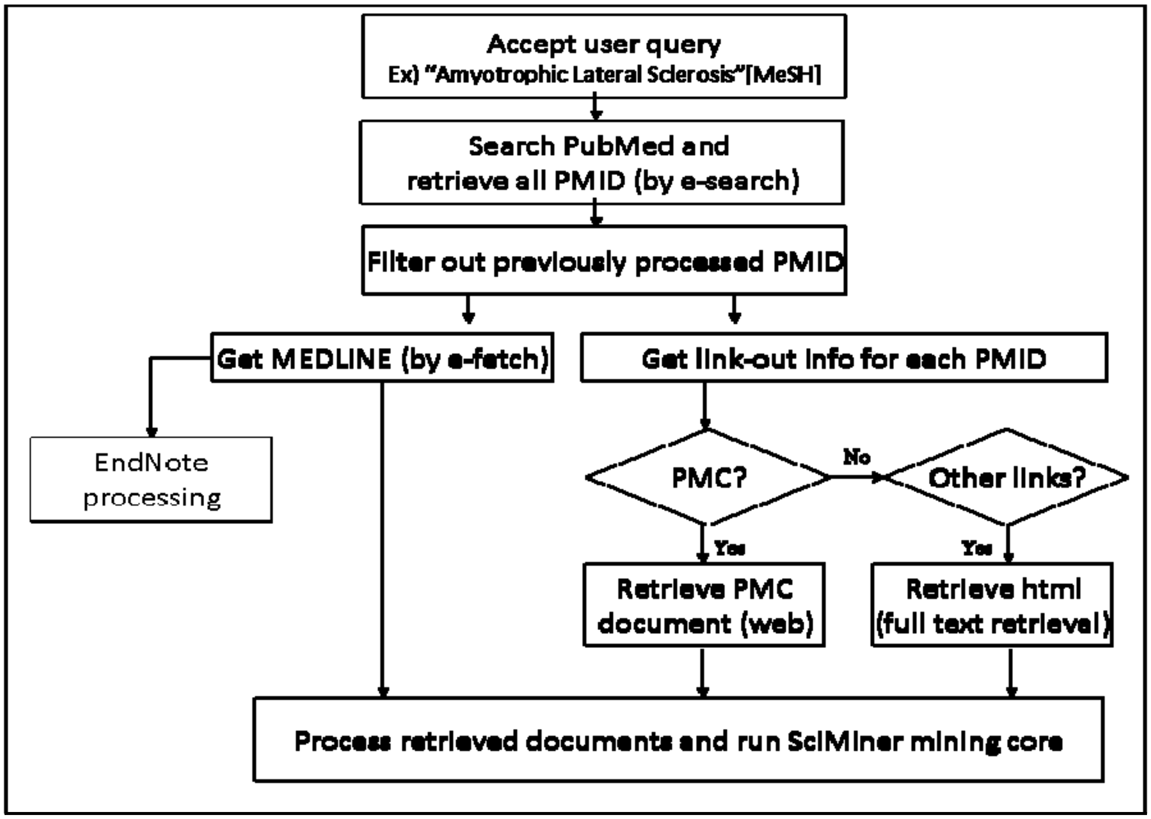
Accept user query
Ex) "Amyotrophic Lateral Sclerosis"[MeSH]
Search PubMed and
retrieve all PMID (by e-search)
Filter out previously processed PMID
Get MEDLINE (by e-fetch)
Get link-out info for each PMID
EndNote
processing
PMC?
No
Other links?
Yes
Yes
Retrieve PMC
document (web)
Retrieve html
(full text retrieval)
Process retrieved documents and run SciMiner mining core

Figure 2-1. Schematic Diagram of SciMiner Query Process and Document Retrieval

Each document's MEDLINE record is retrieved by NCBI e-fetch utility and processed by SciMiner document processing pipeline. In order to retrieve available full text HTML, 'NCBI PubMed's link-out to journal' information is fetched for each document to acquire the corresponding journal's URL. If multiple links are available, the PubMed Central (PMC) gets the highest priority since PMC maintains a consistent layout for html document, which makes the parsing process smooth. For URLs directing to an abstract page or a service provider selection page, SciMiner automatically tries to locate possible full text URLs and retrieve the full text HTML. Such publishers are noted as 'Multi' in the 'Retrieval Steps' column in. Depending on the subscription status of users' institutional library, full text availability may be variable for the standalone local installation of SciMiner. shows the list of journal publishers that are currently supported by SciMiner. PDF documents are not supported by SciMiner.

Table 2-2. Full Text Sources (Journal Publishers)

Publishers	Full Text			MEDLINE Abstract	Note
	Availability	Supported by SciMiner	Retrieval Steps$^{1)}$		
PubMed Central	O	O	Single	O	
Nature	O	O	Multi	O	
Science	O	O	Single	O	
Elsevier (ScienceDirect)	O	O	Single	O	
Highwire	O	O	Single	O	
Blackwell synergy	O	O	Multi	O	
Informaworld	O	O	Multi	O	
Springer link	O	O	Multi	O	
Ovid	O	O	Multi	O	
Karger	O	O	Multi	O	
PortlandPress	O	O	Single	O	
Generic$^{2)}$	O, X	O, X	Single	O	
Wiley Science	O	X		O	PDF
Libert online	O	X		O	PDF
Ingenta	O	X		O	

O: available or supported, X: unavailable or unsupported. 1) There are two types of retrieval steps: **Single** where the NCBI Link-out URL directly points to the full text html and **Multi** where the NCBI Link-out URL should be processed in multiple steps to get the correct full text html. 2) Generic indicates all other publishers that are not specified in this table.

Table 2-3 shows the statistics in the SciMiner database as of 11/10/2008 with 92,089 documents processed. These statistics do not necessarily represent the whole literature available in PubMed due to its very small number of documents.

Table 2-3. Proportion of Publishers in the SciMiner Database Sorted by Percentage (a Snapshot at 11/10/2008 with 92,089 Documents)

Document Source Type	Publisher/Journal Type	count	percentage
Full text (if available) AND MEDLINE Abstract	Elsevier (Science Direct)	19654	21.3%
	HighWire	9672	10.5%
	Other (Generic)	6333	6.9%
	PubMed Central	6218	6.8%
	Wiley Science	4874	5.3%
	Springer link	2287	2.5%
	Ingenta	1156	1.3%
	Nature	1122	1.2%
	Ovid	1066	1.2%
	Informaworld	898	1.0%
	Libert online	595	0.6%
	Karger	559	0.6%
	Science	159	0.2%
	Portland Press	120	0.1%
	BlackWell Synergy	78	0.1%
MEDLINE Abstract Only	NCBI (MEDLINE only)	25312	27.5%
	PDFONLY	4942	5.4%
	Full Text Retrieval Failed	444	0.5%

2.2.2 ANNOTATION RESOURCES

Targets (genes/proteins) used in the SciMiner system are based on the HUGO (Gene Nomenclature Committee, http://www.genenames.org/) gene set [48, 108]; thus, SciMiner reports are based on human targets. SciMiner does not distinguish between human genes from other species genes. Even though the actual biological functions can vary from species to species, SciMiner assumes that overall biological functions are relatively well conserved among species if genes are using the same symbol or name. For example, if a paper mentions superoxide dismutase 1 (Sod1) in mice, SciMiner assigns this occurrence to SOD1 (as human gene) and disregards taxonomy information. This approach is expected to get an overview of biological functions.

Annotation information for each HUGO entry is collected from the following resources. It should be noted that from the literature text data, SciMiner tries to identify only targets (genes and proteins). Other data types (pathway, Gene Ontology terms, MeSH, and Protein-Protein interactions) for identified targets are collected through external annotation resources listed below in Table 2-4. Mapping among database entries have been made available by in-house Perl scripts and the Clone/Gene ID converter [109].

Table 2-4. External Annotation Databases

Type	Data Source Name	URL
Gene Protein	HGNC (HUGO Gene Nomenclature)	http://www.genenames.org/
	NCBI Entrez Gene	http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene
Pathway	KEGG Pathway	http://www.genome.ad.jp/kegg/pathway.html
	Reactome Pathway	http://www.reactome.org/
	NCBI Entrez Gene	http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene
PPI	MiMI (Michigan Molecular Interactions)	http://mimi.ncibi.org/
MeSH	PubMed	http://www.ncbi.nlm.nih.gov/pubmed/
Gene Ontology	Gene Ontology Consortium	http://www.geneontology.org
	NCBI Entrez Gene	http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene

2.2.3 SCIMINER DICTIONARIES

2.2.3.1 Dictionary Compilation and Expansion

SciMiner uses two dictionaries, referred to as ‘Symbol’ and ‘Name’, compiled from the HGNC (HUGO Gene Nomenclature) database (http://www.genenames.org) and the NCBI Entrez Gene database (http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene) previously known as LocusLink. The Symbol dictionary holds single word acronyms, while the Name dictionary contains longer descriptions (at least two words) of targets. In the current version (SciMiner 2.2), the Symbol dictionary has 83,735 unique entries and the Name dictionary has 136,827 unique entries. These dictionaries are extended to 87,014 and 263,304 entries, respectively, via the SciMiner dictionary expansion rules, which include relaxed special character handling and Greek character conversions such as TNF-alpha to TNF-A and TNFA.

2.2.3.1.1 Source of Symbol and Name

As the primary source for gene symbols and names, the following data from the HGNC database for each gene are retrieved: ‘Approved Symbols’, ‘Approved Names’, ‘Previous Symbols’, and ‘Previous Names’. As the secondary source, the following data from the NCBI Entrez Gene database for each gene are retrieved: ‘OFFICIAL_SYMBOL’, ‘ALIAS_SYMBOL’, ‘OFFICIAL_GENE_NAME’, ‘ALIAS_PRODUCT’, and ‘ALIAS_PROT’. For any conflicting symbols and names, the ‘Symbol’ and ‘Name’ dictionaries maintain a default assignment. Full conflict information is kept in separate files and used during the SciMiner mining process. The following order of preference is used for assigning a default HUGO ID to conflicting symbols or names: for the Symbol dictionary (HUGO Approved Symbol > NCBI Official Symbol > HUGO Previous Symbols > NCBI Alias Symbol) and for the Name dictionary (HUGO Approved Name > NCBI Official Gene Name > HUGO Previous Names > NCBI Alias Product > NCBI Alias Prot).

2.2.3.1.2 Dictionary Expansion Rules

The Symbol and Name dictionaries are extended to include possible variations of the terms to increase the overall sensitivity of SciMiner target identification system.

1) For entries in the Symbol dictionary
   a) Special characters are removed for entries in the symbol dictionary, Greek words frequently used in gene symbols such as alpha, beta, gamma and kappa are replaced by their corresponding single characters if no such symbol already exists in the dictionary. (e.g., TNF-alpha (HGNC ID: 11892) ➔ 'TNF-A' is added to the Symbol dictionary for HGNC ID: 11892)
   b) Dashes '-' are removed and the resulting symbol is added as a new entry unless there is already such symbol in the dictionary. (e.g., TNF-A (HGNC ID: 11892) ➔ TNFA is added to the dictionary terms for HGNC ID:11892)
2) For entries in the Name dictionary
   a) Special characters are removed and resulting names are added to the dictionary as explained.
   b) Phrases in parentheses are removed. (e.g., AFG3 ATPase family gene 3-like 1 (yeast) ➔ AFG3 ATPase family gene 3-like 1)
   c) Phrases before the first comma, if any, are added as a new term. (e.g., 39S ribosomal protein L10, mitochondrial ➔ 39S ribosomal protein L10)

d) Commas are removed. (e.g., ANKRD26-like family C, member 1A → ANKRD26-like family C member 1A)

2.2.3.1.3 English Dictionary

General English words are excluded in the SciMiner mining process. A dictionary containing 135,000 words was obtained from http://vburton.ncsa.uiuc.edu/wordlist.txt. Any word ending with '~ase' or those frequently used for gene symbols such as ski and Jun were manually removed from this dictionary. SciMiner checked any identification of a symbol against this English dictionary to filter out possible false positive findings. For example, SciMiner does not accept the gene 'CLOCK' if it is identified by a relaxed match as 'Clock' in text and without any relevant longer description form of 'CLOCK'. For any symbols having the same alphabets with the symbol dictionary, case should be completely matched.

2.2.4 TARGET IDENTIFICATION

2.2.4.1 Overall Target Recognition (Mining) Process

Retrieved documents (both MEDLINE records and full text HTMLs) are pre-processed for removal of unnecessary hyperlinks and UTF-8 characters in text, and then split by an in-house sentence-splitter into individual sentences. Sentences undergo the target recognition process via 'Symbol search' and 'Gene name search' depending on the base dictionary as introduced in the Section 2.2.3. In the current version of SciMiner v2.2, targets identified by Symbol search are subjected to a name-resolving process, while those targets identified by Name search are not. This is because single-word symbols or acronyms can have many different meanings, while multi-word names are usually specific. Figure 2-2 illustrates the overall mining process.

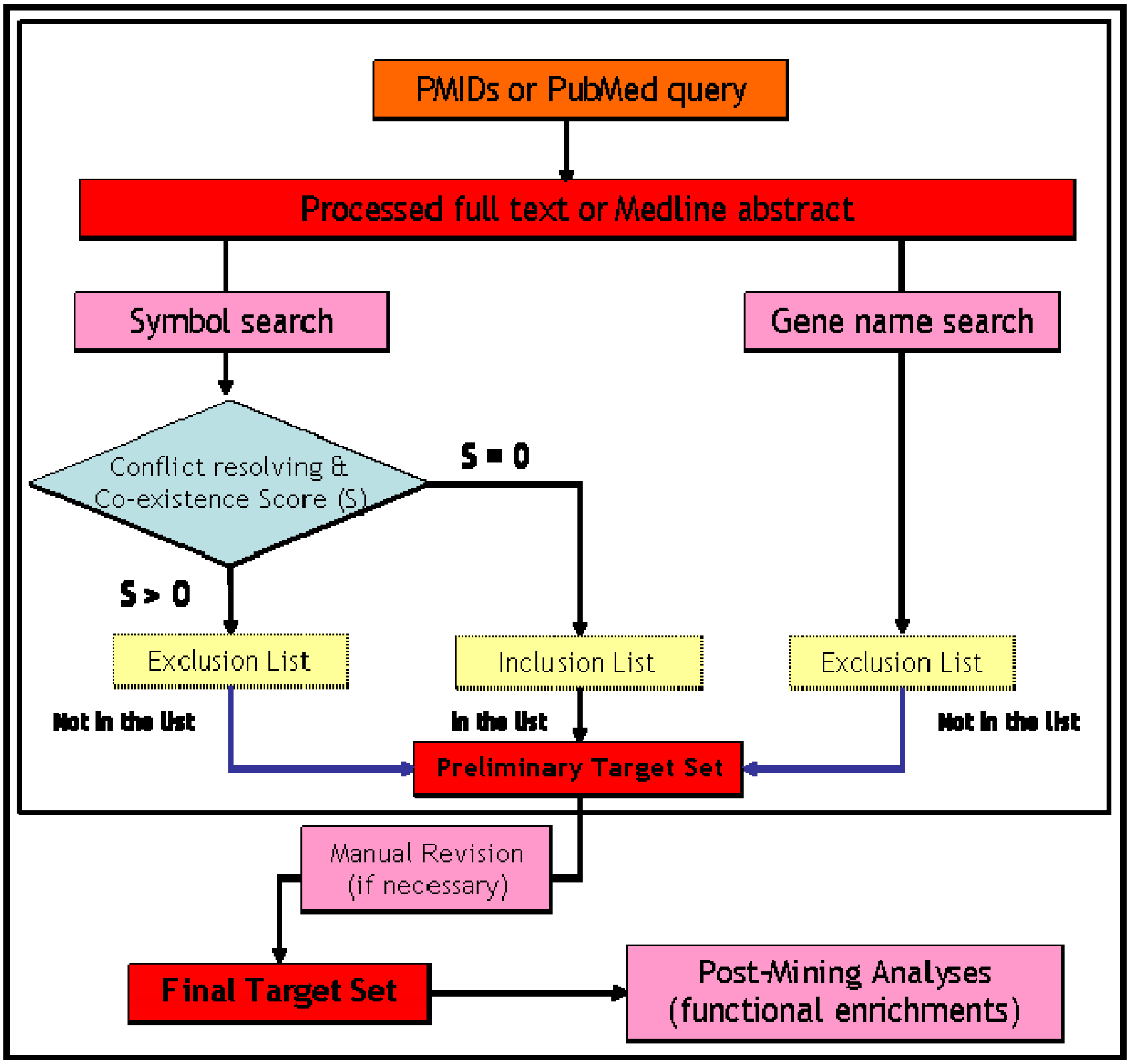

Figure 2-2. Schematic Diagram of SciMiner Target Recognition Process

Retrieved documents are subjected to 'Symbol search' and 'Gene name search'. In Symbol search, conflict-resolving step is additionally employed to correctly assign HGNC ID to identified target (see Section 2.2.4.2 Confidence Scoring Scheme for more details). All identified targets are further filtered by user-provided 'EXCLUDE' and 'INCLUDE' lists to improve the overall accuracy.

2.2.4.1.1 Symbol search

The following steps summarize the Symbol search process.

1) Symbol dictionary loading: The pre-compiled symbol dictionary is loaded into two hash tables; one with keeping the case and another with making all in lower case from the second character. For example, SOD1 is hashed as ‘SOD1’ in the first hash and ‘Sod1’ in the second hash. The following rules are employed to create possible alternative symbols as an extension.
   a) Anything ending with –ALPHA, -BETA, -GAMMA, and –KAPPA to –A or A, -B or B, -G or G, and –K or K. (e.g., PI4K-BETA ➔ PI4K-B and PI4KB)
   b) [All non-numeric characters]-[Numbers] to [All non-numeric characters][Numbers] without ‘-dash’ in-between. (e.g., SOD-1 ➔ SOD1)
   c) Any symbol entry is filtered by the IGNORE list as well as the general English dictionary.
2) Sentences in the retrieved documents are further split by a space ‘ ‘ and each word is checked against the hash tables of symbol dictionary. A word containing special characters such slash ‘/’ or parentis ‘(‘ are checked as a whole as well as split forms by such special characters. (e.g., Smad3/Smad4 ➔ Smad3 Smad4)
3) Checking against dictionary hash tables and further rules:
   a) Check the word keeping the case.
   b) If no match has been identified above, try to convert characters from the second position to the end into a lower case, while keeping the first character as it is and check hash table if there is a match. (e.g., SOD1 -> Sod1)

c) If no match has been identified above and the word has the following pattern (^[h|m]([A-Z].*)), then [h|m] is removed and the remaining is checked. (e.g., hPop1 ➔ POP1, or mSin3a => SIN3A)

d) If no match has been identified above and the word contains a dash ‘-‘,

   i. A starting or ending dash is removed. (e.g., SOD1- ➔ SOD1)

   ii. Ending ‘–receptor ‘ is converted to –R or R. (e.g., INS-receptor ➔ INS-R and INSR)

   iii. Ending ‘–(alpha|beta|gamma|kappa)’ is converted to–A or A and etc. (e.g., TNF-alpha ➔ TNF-A)

      (1) Anything ending with the followings are truncated: like|dependent|specific|receptor|staining|induced|inducible|activated|repressed|stimulated|controlled|enhanced|mediated (e.g., SOD1-induced ➔ SOD1)

   iv. If no match has been identified above and the word is in the following pattern /^(alpha|beta|gamma)(\d+)-(\S+)&/, the order of words are changed. (e.g., beta1-syntrophin ➔ syntrophin beta 1)

   v. Dashes are simply removed unless it is in a (number)-dash-(number) pattern.

e) If no match has been identified above and the word ends with (alpha|beta|gamma|kappa), single greek characters are used with and without a dash. (e.g., TNFgamma ➔ TNF-G and TNFG)

f) If no match has been identified above and the word ends with a ‘s’ which is a probable plural form.

i. If the word is in all lower case, then it is ignored.

ii. Otherwise, the word is checked against the dictionary alone.

iii. Accompanying words are also checked for any possible expansion forms. (e.g., SMADs 3 and 4 ➔ SMAD3 and SMAD4)

g) If no match has been identified above and the word contains at least one upper case character, then the word is converted to all upper case and checked. Discard any word in all lower case.

h) Further expansions rules are applied if a match has been identified from above (at ith position) and accompanying word (i+1th position) is either ‘and’ or ‘to’ (e.g., SMAD3 and 4 ➔ SMAD3 and SMAD4).

4) Confidence scoring calculation: Positive score is assigned if there exist longer descriptions of the word (acronym) being tested in the same document (not only in the same sentence) (see Section 2.2.4.2 for more details). Longer descriptions refer to the entries in the name dictionary and expanded names.

5) Conflict resolution: If the identified symbol has a conflict (belonging to a precompiled symbol-conflict set), confidence scores is calculated for all of the possible candidates specified by the symbol-conflict set. Only the top scoring match is selected and reported.

6) Acronyms that are less likely to be gene symbols are further checked and filtered out.

a) Anything followed by (et al, buffer, score, version, medium, media, cell, software, program, algorithm, system, test, company, agent) and their plural forms if available. (e.g., SPSS program or MES buffer)

b) Anything that has ‘acknowledge’ or ‘thank’ in its flanking text. Usually identified words with these words are author names matched usually by the 5)B) step above. Sentences from the acknowledgement section are excluded in SciMiner process.

c) Anything that has a pattern of multiple /[AGCTU] [AGCTU] [AGCTU]$/i is filtered out as RNA codons such as AGT, AUG and etc.

7) Score boosting: Partial positive scores is given if zero-scored match meet any of the following criteria. This is based on the assumption that a legitimate symbol can have a zero-confidence score if there is no supporting longer form of gene names. This happens quite frequently, where only the abstract is available.

a) If there is any positive score target within the same block, a partial score of 0.2 is assigned to the zero-scored target. (e.g., ‘SOD1/NOS1’ where scores were 0 for SOD1 and 0.5 for NOS1. In this case, SOD1 gets a partial score of 0.2. Our assumption is that if part of the word contains a positive scored match, the remaining is also likely to be a legitimate gene symbol (as long as it is found as a match from the mining steps above).

b) If there are one or more positive scored neighbors within two-word distance, a partial score of 0.2 is assigned to the zero-scored target (e.g., ‘SOD1 TP53 NOX3 were up-regulated’ where scores were 0 for SOD1, 2.5 for TP53, and 0.9 for NOX3. Our algorithm assigned 0.2 to SOD1).

8) Any remaining matches go through EXCLUDE/INCLUDE filtering process.

a) Any positive scored matches is checked against EXCLUDE list. If it belongs to EXCLUDE list and a corresponding condition is found in the same document, then this positive scored match is marked as 'EXCLUDED'.

b) Any zero scored matches are also checked against INCLUDE list. If it belongs to the INCLUDE list unless it also belongs to EXCLUDE list and its corresponding condition is met.

2.2.4.1.2 Gene name search

Compared to the Symbol search above, Gene name search is much simpler, as summarized below.

1) Dictionary loading: The pre-compiled Gene name dictionary is loaded into an array (@UNIQNAME) after the following processing.

   A. Replacement of any special characters with a blank space.

   B. Removal any multiple consecutive spaces into a single space.

   C. Conversion into lower case.

   D. Removal of any gene name entry with less than 4 characters.

   E. Additional hash of partial gene names: The official HUGO gene names are processed and used for confidence-scoring purpose only. These are not used as individual gene name identification but only used during confidence score calculation for Symbol search above. Anything that has a match in IGNORE list or English dictionary are not included. Minimum word length is 3 to be included. FAS: Fas (TNF receptor superfamily, member 6) is given as an example.

      i. Fas: Not to be included since it is identical to the main entry.

ii. TNF: Included with a partial score of 0.3

iii. receptor: Not included since it is filtered by English dictionary.

iv. superfamily: Not included since it is filtered by English dictionary.

v. member: Not included since it is filtered by English dictionary.

vi. Fas TNF: Included with a partial score of 0.3

vii. Fas TNF receptor: Included with a partial score of 0.3

viii. Fas TNF receptor superfamily: Included with a partial score of 0.3

ix. Fas TNF receptor superfamily member: Included with a partial score of 0.3

F. The main gene name array is sorted in an alphabetical order.

G. First four characters of each gene name are collected and two hash tables are generated. These hash tables are employed as indexes to reduce the search space during gene name searching. Four-character threshold has been empirically chosen.

i. %first4codeStart: This hash contains the starting index of @UNIQNAME for the entries starting with the given 4-character.

ii. %first4codeEnd: This hash contains the last index of @UNIQNAME for the entries starting with the given 4-character.

2) Sentences are further split by a space ‘ ‘ and any special characters are removed.

3) The first four characters of every single word in the document are collected.

4) For each four-character code, candidate gene names are obtained from the @UNIQNAME array using %first4codeStart (as starting index)

and %first4codeEnd (as ending index). These names are used in Perl regular expression to identify any occurrence in the full document.

   A. Since gene names are so variable in length, we cannot use hash-table based approach as in the Symbol search.

5) Any identification made above is checked against the EXCLUDE filter.
6) Remaining identifications is reported and the confidence score of 1 is assigned.

In the final step, symbol based identification result and gene name based identification result are merged into a single result.

2.2.4.2 Confidence Scoring Scheme

The same acronym can be shared by multiple distinct targets, which becomes a major obstacle in correctly recognizing abbreviated forms of target names. This ambiguity is resolved with a confidence scoring scheme based on the co-occurrence of abbreviated symbols and longer descriptions in the same document. Compared to other systems employing co-occurrence based approaches (e.g. ProMiner [50]), SciMiner extends the co-occurrence search scope to the MEDLINE MeSH records and further allows partial name matches. This becomes particularly useful when the full text bodies are not available. This scoring scheme is used to 1) resolve the name conflict and 2) increase the precision.

Targets identified through the name dictionary are given a score of '1' and do not go through the name resolution process. Targets identified through the symbol dictionary are subjected to confidence score calculation based on co-occurrence of longer descriptions (entries in the name dictionary and expanded forms). Scores are assigned to each identification (match) according to the following rules: 1) A score of ‘0.5’ is given to perfect matches with a longer description from the unique name dictionary; 2) A score of ‘0.3’ is given to a partial match to the approved name of the corresponding HUGO symbol or full match to the expanded names.

The following rules are further employed to increase the overall accuracy of the mining result by minimizing possible false positives:

1) A score of 0.5 is given to a match preceding or followed by the following terms; gene(s), protein(s), mRNA(s);
2) A score of 0.3 is given to a match within in a same block of positively scored target (a single word in the original text but only separated by special characters) like ‘Bcl-xL/Bad’;
3) A score of 0.2 is given to a match with one word apart from other positive matches.
4) A score of 0.1 is given to a match with two words apart from other positive matches.
5) Any sentence from Acknowledgement and author lines are excluded since matches from these sections are very prone to have false positive matching of symbols to author names.

2.2.4.3 User-Provided Filters and Manual Correction

SciMiner accuracy is increased by allowing users to provide their own filters. The IGNORE list may contain entities to be ignored. The INCLUDE and EXCLUDE lists of acronyms (or symbols) are included or excluded when conditions are met. For example, the default SciMiner EXCLUDE list has ‘SDS’ and ‘sodium dodecyl sulfate’ as its condition. Identification of ‘SDS’ in a text as ‘serine dehydratase’ will be excluded if there is an occurrence of ‘sodium dodecyl sulfate’ in the same document. In order to further improve the accuracy of mined targets, SciMiner allows users to manually edit identified targets on the mining result pages.

2.2.5 POST-MINING ANALYSIS

Functional enrichment analyses are performed by comparing the identified targets of one search to those of other search results. Fisher’s exact test [110] is used to identify statistically significant over-representations of target list entries, Gene Ontology terms, MeSH terms, and pathways. This post-mining analysis step provides a simple but intuitive way to understand over-represented biological functions.

2.2.6 VISUALIZATION

A summary is provided for the Target Recognition and Post-Mining Analysis results and the full results are available as a web-page, a simple text file, and an Excel file. The molecular interaction networks of the targets can be visualized in Cytoscape [111] by following links from the Target Recognition result page. This functionality is enabled by the Cytoscape MiMI plug-in [112] and Java Web Start (http://java.sun.com/).

2.2.7 DATA MANAGEMENT

SciMiner is implemented in Perl and uses a MySQL database to store compiled dictionaries and identified targets. Server-client communication is handled by CGI (Common Gateway Interface) scripts.

2.3 RESULTS AND DISCUSSION

2.3.1 PERFORMANCE EVALUATION ON BIOCREATIVE CORPUS

The performance of target identification was evaluated using the BioCreAtIvE (Critical Assessment of Information Extraction systems in Biology) II (Year 2006) Gene Normalization (GN) Task as a gold standard [107]. The Gene Normalization task aims at correctly identifying the unique identifiers of genes and proteins mentioned in literature data and linking them to the NCBI Entrez Gene database. The gold standard set contains 785 human gene identifiers in a corpus of 262 abstracts.

With the scoring scheme disabled, SciMiner identified 1,114 human gene identifiers of which 677 identifiers were matched to the gold standard set. This corresponds to 86.2% recall, 60.8% precision, and 71.3% F-measure. Utilizing the SciMiner scoring scheme and optimally tuning the score threshold parameter for each of the evaluation measures resulted in maximum values of 87.1% recall (at score threshold of zero), 71.3% precision (at score threshold of 0.7), and 75.8% F-measure (at score threshold of 0.3). This result suggests that using scoring scheme based improves the precision of the target identification.

Without scoring scheme, an acronym conflict is not resolved and the default HUGO ID for the acronym in SciMiner dictionary was reported. Unresolved conflicts have contributed to a slight increase in the total number of identifications (1,114 vs 1,092). For example, the document of PMID 10072587 (titled as “Cloning of a novel gene (ING1L) homologous to ING1, a candidate tumor suppressor”) includes the following sentence, “The ING1 gene encodes p33 (ING1), a putative tumor suppressor for neuroblastomas and breast cancers, which has been shown to cooperate with p53 in controlling cell proliferation.”. With the scoring scheme on, SciMiner correctly identified p33 as ING1, but without the scoring scheme, SciMiner incorrectly identified p33 as LTB (lymphotoxin beta TNF superfamily, member 3), which had ‘p33’ as one of its synonyms.

Table 2-5 shows the performance summary of SciMiner with various scoring thresholds applied and Figure 2-3 illustrates how these measures vary with different score thresholds. Recall decreases as score threshold increases. Precision improves as the score threshold increases to a value of 0.7, and then decreases slightly. This indicates that utilizing the scoring scheme increases the overall precision, but further optimization is required.

Table 2-5. Recall, Precision, and F-score for Multiple SciMiner Confidence Score Thresholds.

Score Threshold	Total Identification	True Positive	False Positive	False Negative	Recall	Precision	F-measure
No Scoring Scheme	1114	677	437	108	0.862	0.608	0.713
0	1092	684	408	101	**0.871**	0.626	0.729
0.1	956	659	297	126	0.839	0.689	0.757
0.2	937	652	285	133	0.831	0.696	0.757
0.3	929	650	279	135	0.828	0.700	**0.758**
0.4	842	597	245	188	0.761	0.709	0.734
0.5	838	595	243	190	0.758	0.710	0.733
0.6	816	581	235	204	0.740	0.712	0.726
0.7	764	545	219	240	0.694	**0.713**	0.704
0.8	764	545	219	240	0.694	0.713	0.704
0.9	738	521	217	264	0.664	0.706	0.684
1	699	490	209	295	0.624	0.701	0.660
1.1	689	484	205	301	0.617	0.702	0.657
1.2	670	468	202	317	0.596	0.699	0.643
1.3	643	446	197	339	0.568	0.694	0.625
1.4	637	440	197	345	0.561	0.691	0.619
1.5	610	417	193	368	0.531	0.684	0.598

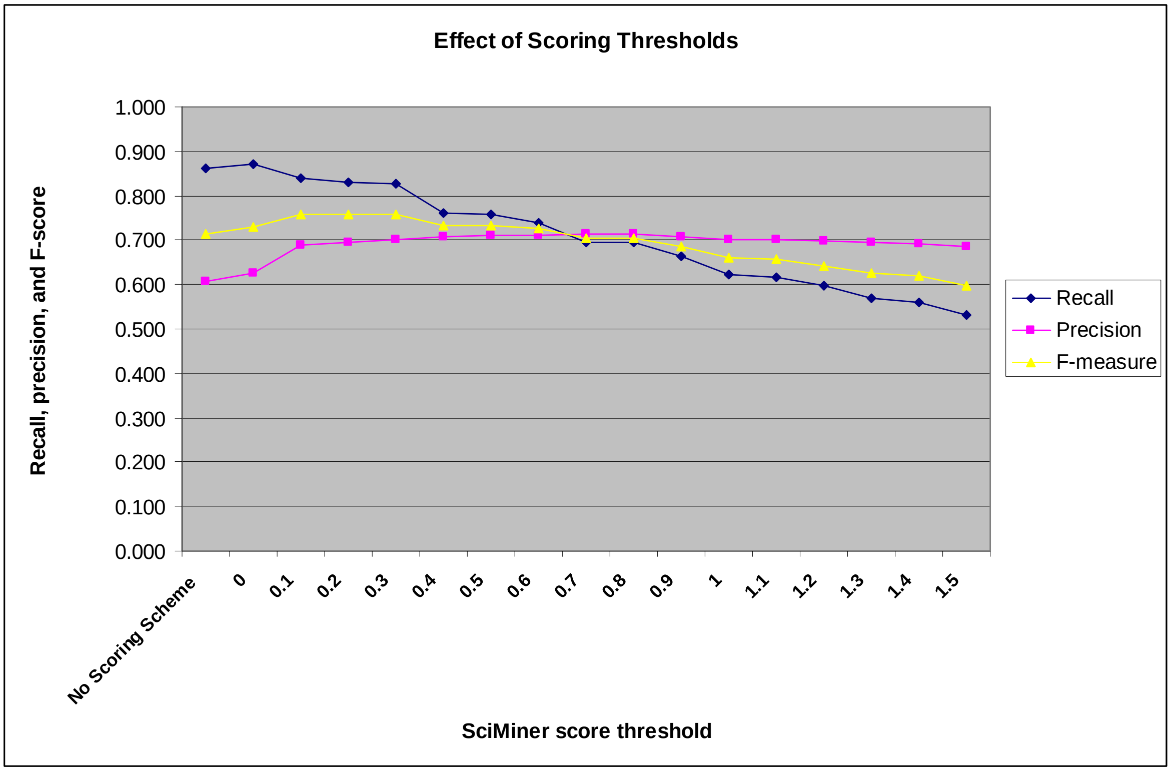

Figure 2-3 SciMiner Recall, Precision, and F-measure by Different Score Thresholds

At low or zero confidence score thresholds, SciMiner shows very high recall rates, but also registers high numbers of false positive identifications leading to relatively low precision. However, it should be noted that SciMiner provides users with opportunities to improve the overall accuracy. Users are allowed to edit the identification results if they find any misidentified targets and can also use custom filters (IGNORE, INCLUDE, and EXCLUDE) to improve accuracy.

Compared to the 54 BioCreAtIvE II Gene Normalization Task results posted by 20 groups [107], SciMiner's recall, precision and F-measure rank 2nd, 34th, and 19th, respectively.

2.3.2 APPLICATION

SciMiner was run on a query of “Amyotrophic Lateral Sclerosis” and found 3,226 targets from 10,625 documents as of 08/31/2008. The most frequently found target was superoxide dismutase 1 (SOD1) from 2,198 papers, followed by amyloid beta (APP), ubiquitin (RPS27A), microtubule-associated protein tau (MAPT). Post-Mining Analysis identified 183 enriched pathways in these targets (p<0.001). They include KEGG pathways of amyotrophic lateral sclerosis, apoptosis, and signaling pathways (e.g. MAPK and JAK-STAT).

Post-Mining Analysis was performed between two subsets of the above corpus; Query1 (“Amyotrophic Lateral Sclerosis” AND “Reactive Oxygen Species”) and Query2 (“Amyotrophic Lateral Sclerosis” AND “Inflammation”). This comparison identifies targets that are over-represented in either “Reactive Oxygen Species” or “Inflammation” in the domain of “Amyotrophic Lateral Sclerosis”. Query1 found 401 targets from 172 documents, while Query2 found 561 from 168 documents. Catalase (CAT) and SOD1 were highly over-represented in the Query1 result, while tumor necrosis factor (TNF) and interleukin-6 (IL-6) were highly over-represented in the Query2 result. The pathway enrichment analysis further found that “DNA repair” and “cytokine-cytokine receptor interactions” were the most significantly enriched pathways from the targets of Query1 and Query2, respectively.

2.3.3 CONCLUSION

SciMiner provides a convenient web-based platform for mining targets (genes and proteins) from the biomedical literature with the capacity for functional enrichment analyses. SciMiner performs well compared to other methods, but is unique in that it (i) searches full text documents (not just abstracts), (ii) allows users to directly edit the mining results, and (iii) allows comparisons to be made between search results of multiple queries.

CHAPTER 3

LITERATURE-BASED DISCOVERY OF DIABETES- AND ROS-RELATED TARGETS

3.1 INTRODUCTION

Diabetes is a metabolic disease in which the body does not produce or properly respond to insulin, a hormone required to convert carbohydrates into energy for daily life. According to the American Diabetes Association, twenty-three million children and adults, approximately 7.8% of the population in the United States, have diabetes [1]. The cost of diabetes in 2007 was estimated to be $174 billion [1]. The micro- and macro-vascular complications of diabetes are the most common causes of renal failure, blindness and amputations leading to significant mortality, morbidity and poor quality of life; however, incomplete understanding of the causes of diabetic complications hinders the development of mechanism-based therapies.

In vivo and *in vitro* experiments implicate a number of enzymatic and non-enzymatic metabolic pathways in the initiation and progression of diabetic complications [2] including: (1) increased polyol pathway activity leading to sorbitol and fructose accumulation, NAD(P)-redox imbalances and changes in signal transduction; (2) non-enzymatic glycation of proteins yielding "advanced glycation end-products" (AGEs); (3) activation of protein kinase C (PKC), initiating a cascade of intracellular stress responses;

and (4) increased hexosamine pathway flux [2, 12]. Only recently has a link among these pathways been established that provides a unified mechanism of tissue damage. Each of these pathways directly and indirectly leads to overproduction of reactive oxygen species (ROS) [2, 12].

ROS are highly reactive ions or small molecules including oxygen ions, free radicals and peroxides, formed as natural byproducts of cellular energy metabolism. ROS are implicated in multiple cellular pathways such as mitogen-activated protein kinase (MAPK) signaling, c-Jun amino-terminal kinase (JNK), cell proliferation and apoptosis [113-115]. Due to the highly reactive properties of ROS, excessive ROS may cause significant damage to proteins, DNA, RNA and lipids. All cells express enzymes capable of neutralizing ROS. In addition to the maintenance of antioxidant systems such as glutathione and thioredoxins, primary sensory neurons express two main detoxifying enzymes: superoxide dismutase (SOD) [28] and catalase [116]. SOD converts superoxide (O_2^-) to H_2O_2, which is reduced to H_2O by glutathione and catalase [116]. SOD1 is the main form of SOD in the cytoplasm; SOD2 is located within the mitochondria. In neurons, SOD1 activity represents approximately 90% of total SOD activity and SOD2 approximately 10% [117]. Under diabetic conditions, this protective mechanism is overwhelmed due to the substantial increase in ROS, leading to cellular damage and dysfunction [26] .

The idea that increased ROS and oxidative stress contribute to the pathogenesis of diabetic complications has led scientists to investigate different oxidative stress pathways [27, 28]. Inhibition of ROS or maintenance of euglycemia restores metabolic and vascular imbalances and blocks both the initiation and progression of complications [29, 30]. Despite the significant implications and extensive research into the role of ROS in diabetes, no comprehensive database regarding ROS-related genes or proteins is currently available.

In the present study, a comprehensive list of ROS- and diabetes-related targets (genes/proteins) was compiled from the biomedical literature through text mining technology. SciMiner, a web-based literature mining tool [42], was used to retrieve and process documents and identify targets from the text. The collected ROS-diabetes targets were further tested against randomly selected non-ROS-diabetes literature to identify targets that are significantly over-represented in the ROS-diabetes literature. Functional enrichment analyses were performed on these targets to identify significantly over-represented biological functions in terms of Gene Ontology (GO) terms and pathways.

In order to confirm the biological relevance of the over-represented ROS-diabetes targets, the gene expression levels of 9 selected targets were measured in dorsal root ganglia (DRG) from mice with and without diabetes. DRG contain primary sensory neurons that relay information from the periphery to the central nervous system (CNS) [26, 28, 118]. Unlike the CNS, DRG are not protected by a blood-nerve barrier, and are consequently vulnerable to metabolic and toxic injury [119]. We hypothesize that differential expression of identified targets in DRG would confirm their involvement in the pathogenesis of diabetic neuropathy.

3.2 METHODS

3.2.1 DEFINING ROS-DIABETES LITERATURE

To retrieve the list of biomedical literature associated with ROS and diabetes, PubMed was queried using (“Reactive Oxygen Species”[MeSH] AND “Diabetes Mellitus”[MeSH]). This query yielded 1,154 articles as of April 27, 2009. SciMiner, a web-based literature mining tool [42], was used to retrieve and process the abstracts and available full text documents to identify targets (full text documents were available for approximately 40% of the 1,154 articles). SciMiner-identified targets, reported in the form of HGNC [HUGO (Human Genome Organization) Gene Nomenclature Committee] genes, were confirmed by manual review of the text.

3.2.2 COMPARISON WITH HUMAN CURATED DATA (NCBI GENE2PUBMED)

The NCBI Gene database provides links between Gene and PubMed. The links are the result of (1) manual curation within the NCBI via literature analysis as part of generating a Gene record, (2) integration of information from other public databases, and (3) GeneRIF (Gene Reference Into Function) in which human experts provide a brief summary of gene functions and make the connections between citation (PubMed) and Gene databases. For the 1,154 ROS-diabetes articles, gene-paper associations were retrieved from the NCBI Gene database. Non-human genes were mapped to homologous human genes through the NCBI HomoloGene database. The retrieved genes were compared against the SciMiner derived targets. Any genes missed by SciMiner were added to the ROS-diabetes target set.

3.2.3 PROTEIN-PROTEIN INTERACTIONS AMONG ROS-DIABETES TARGETS

To indirectly evaluate the association of literature derived targets (by SciMiner and NCBI Gene2PubMed) with ROS and diabetes, protein-protein interactions among the targets were surveyed. This was based on an assumption that targets are more likely to have protein-protein interactions if they are truly associated within the same biological functions/pathways. A Protein-Protein Interaction (PPI) network of the ROS-diabetes targets was retrieved from the Michigan Molecular Interactions (MiMI, http://mimi.ncibi.org/) database and compared against the networks of 100 randomly drawn sets from HUGO. A standard Z-test and one sample T-test were used to calculate the statistical significance between the ROS-diabetes PPI network and the random PPI networks.

3.2.4 FUNCTIONAL ENRICHMENT ANALYSIS

Literature derived ROS-diabetes targets (by SciMiner and NCBI Gene2PubMed) were subject to functional enrichment analyses to identify significantly over-represented biological functions in terms of Gene Ontology [92], pathways (Kyoto Encyclopedia of Genes and Genomes (KEGG, http://www.genome.jp/kegg/) [93] and Reactome (http://www.reactome.org/) [94]). Fisher's exact test [110] was used to calculate statistical significance with Benjamini-Hochberg adjusted p-value < 0.05 [120] as the cut-off.

3.2.5 OVER-REPRESENTED ROS-DIABETES TARGETS

3.2.5.1 Defining Background Corpora

To identify a subset of targets that are highly over-represented within the ROS-diabetes targets, the frequency of each target (defined as the number of documents in which the target was identified divided by the number of total documents in the query) was compared against the frequencies in randomly selected background corpora. Depending on how the background set is defined, over-represented targets may vary widely; therefore, to maintain the background corpora close to the ROS and diabetes context, documents were selected from the same journal, volume, and issue of the 1,154 ROS-diabetes documents, but were NOT indexed with "Reactive Oxygen Species"[MeSH] nor "Diabetes Mellitus"[MeSH]. For example, one of the ROS-diabetes articles (PMID:18227068), was published in the Journal of Biological Chemistry, Volume 283, Issue 16. This issue contained 85 papers, 78 of which were not indexed with either "Reactive Oxygen Species"[MeSH] or "Diabetes Mellitus"[MeSH] indexed. One of these 78 papers was randomly selected as a background document. Three sets of 1,154 documents were selected using this approach and processed using SciMiner. Identified targets were confirmed by manual review for accuracy.

3.2.5.2 Identifying Significantly Over-represented Targets

ROS-diabetes targets were tested for over-representation against targets identified from the three background sets. Fisher's exact test was used to determine if the frequency of each target in the ROS-diabetes target set was significantly different from that of the background sets. Any targets with a p-value < 0.05 after Benjamini-Hochberg multiple testing corrections in at least two of the three comparisons were deemed to be an over-represented ROS-diabetes target. Functional enrichment analyses were performed on these over-represented ROS-diabetes targets as described above in Section 3.2.4.

3.2.5.3 Selecting Targets for Real-time RT-PCR

A subset of targets were selected for real-time RT-PCR from the top 10 over-represented ROS-diabetes targets excluding insulin and NADPH oxidase 5 (NOX5), which does not have a mouse ortholog. Nitric oxide synthase 1 (NOS1), the main generator of nitric oxide, ranked at the 15th position and was additionally selected for inclusion in the test set.

3.2.6 DIFFERENTIAL GENE EXPRESSION IN DIABETIC DRG

3.2.6.1 Mice

DBA/2J mice were purchased from the Jackson Laboratory (Bar Harbor, ME). Mice were housed in a pathogen-free environment and cared for following the University of Michigan Committee on the Care and Use of Animals guidelines. Mice were fed AIN76A chow (Research Diets, New Brunswick, NJ). Male mice were used for this study.

3.2.6.2 Induction of Diabetes

Two treatment groups were defined: control (n=4) and diabetic (n=4). Diabetes was induced at 13 weeks of age by low-dose streptozotocin (STZ) injections, 50 mg/kg/day for 5 consecutive days. All diabetic mice received LinBit sustained release insulin implants (LinShin, Toronto, Canada) at 8 weeks post-STZ treatment. Insulin implants were replaced every 4 weeks, at 12 and 16 weeks post-STZ treatment. At 20 weeks post-STZ treatment, mice were euthanized by sodium pentobarbital overdose and DRG were harvested as previously described [121].

3.2.6.3 Real-time RT-PCR

The gene expression of the selected 9 literature-derived ROS-diabetes targets in DRG was measured using real-time RT-PCR. The amount of mRNA isolated from each DRG was normalized to an endogenous reference [Tbp: TATA box binding protein; Δ cycle threshold (CT)] [122].

3.3 RESULTS

3.3.1 IDENTIFICATION OF ROS-DIABETES TARGETS

A total of 1,021 unique targets were identified by SciMiner from the 1,154 ROS-diabetes papers defined by the query of (“Reactive Oxygen Species”[MeSH] AND “Diabetes Mellitus”[MeSH]) and confirmed by manual review. Table 3-1 contains the top 10 most frequently mentioned targets in the ROS-diabetes papers. Insulin was the most frequently mentioned target, followed by SOD1 and catalase.

Table 3-1. Top 10 Most Frequent ROS-diabetes Targets

Symbol	Name	#Paper	Match Strings
INS	insulin	503	INS | insulin | proinsulin |
SOD1	superoxide dismutase 1	368	Sod1 | SOD1 | SOD-1| *
CAT	catalase	241	CAT | catalase | CAT-reversible | *
PRKCA	protein kinase C, alpha	194	PKCA | PKC-alpha | PKC-A *
ALB	albumin	179	albumin | serum albumin |
NOX5	NADPH oxidase 5	177	NOX5 | nadph oxidase |
NOS2A	nitric oxide synthase 2A	144	NOS | iNOS | Nos2 |*
XDH	xanthine dehydrogenase	133	XOR |xanthine dehydrogenase| *
AGT	angiotensinogen	131	Ang-II | ANG | AGT | AngI | *
TNF	tumor necrosis factor	120	TNFA | TNF | TNF-alpha | *

* Matching strings were truncated to fit in the table. The full contents are available at the ROS-diabetes webpage (http://jdrf.neurology.med.umich.edu/ROSDiabetes/) [123]. ‘#Paper’ refers to the number of documents in which each target was mentioned at least once.

The NCBI Gene2PubMed database, containing expert-curated associations between the NCBI Gene and PubMed databases, revealed 90 unique genes associated with the 1,154 ROS-diabetes papers. SciMiner identified 85 out of these 90 targets, indicating a 94% recall rate. These Gene2PubMed associations were integrated with the

SciMiner results to augment the ROS-diabetes target list to result in 1,026 unique ROS-diabetes targets (see the ROS-diabetes webpage for more details [123]).

3.3.2 PROTEIN-PROTEIN INTERACTION NETWORK OF THE ROS-DIABETES TARGETS

The PPI network among the ROS-diabetes targets was evaluated using MiMI interaction data. This was based on the assumption that targets commonly related to a certain topic are more likely to have frequent interactions with each other. One hundred PPI networks were generated for comparison using the same number of genes (1,026) randomly selected from the complete HUGO gene set (25,254). The PPI network of the ROS-diabetes targets was significantly different from the randomly generated networks indicating their strong association with the topic “ROS and Diabetes”.

Table 3-2 demonstrates that the mean number of targets with any PPI interaction in the randomly generated target sets was 528.9 (approximately 52% of 1,026 targets), while the number of targets with any PPI interaction in the ROS-diabetes target was 983 (96%). The number of targets interacting with each other was also significantly different between the random networks (mean=155.4) and the ROS-diabetes network (mean=879). Figure 3-1 illustrates the distributions of these measurements from the 100 random networks with the ROS-diabetes set depicted as a red vertical line. It is obvious that the PPI network of the ROS-diabetes targets is significantly different from the random networks.

Table 3-2. Summary of 100 Randomly Generated PPI Networks

	# of targets with any interaction	# of targets interacting with each other	# of direct interactions among targets	Max degree
ROS-diabetes Targets	983	879	5002	173
Mean (100 networks)	528.9	155.4	165.4	25.0
STDEV (100 networks)	16.0	36.2	54.2	39.7
Z-Score	28.5	20.0	89.2	3.7
P-value(Z)	0	0	0	9.6E-05
T-Statistics	-284.8	-200	-891.9	-37.3
P-value(T)	4.6E-146	6.7E-131	4.0E-195	4.2E-60

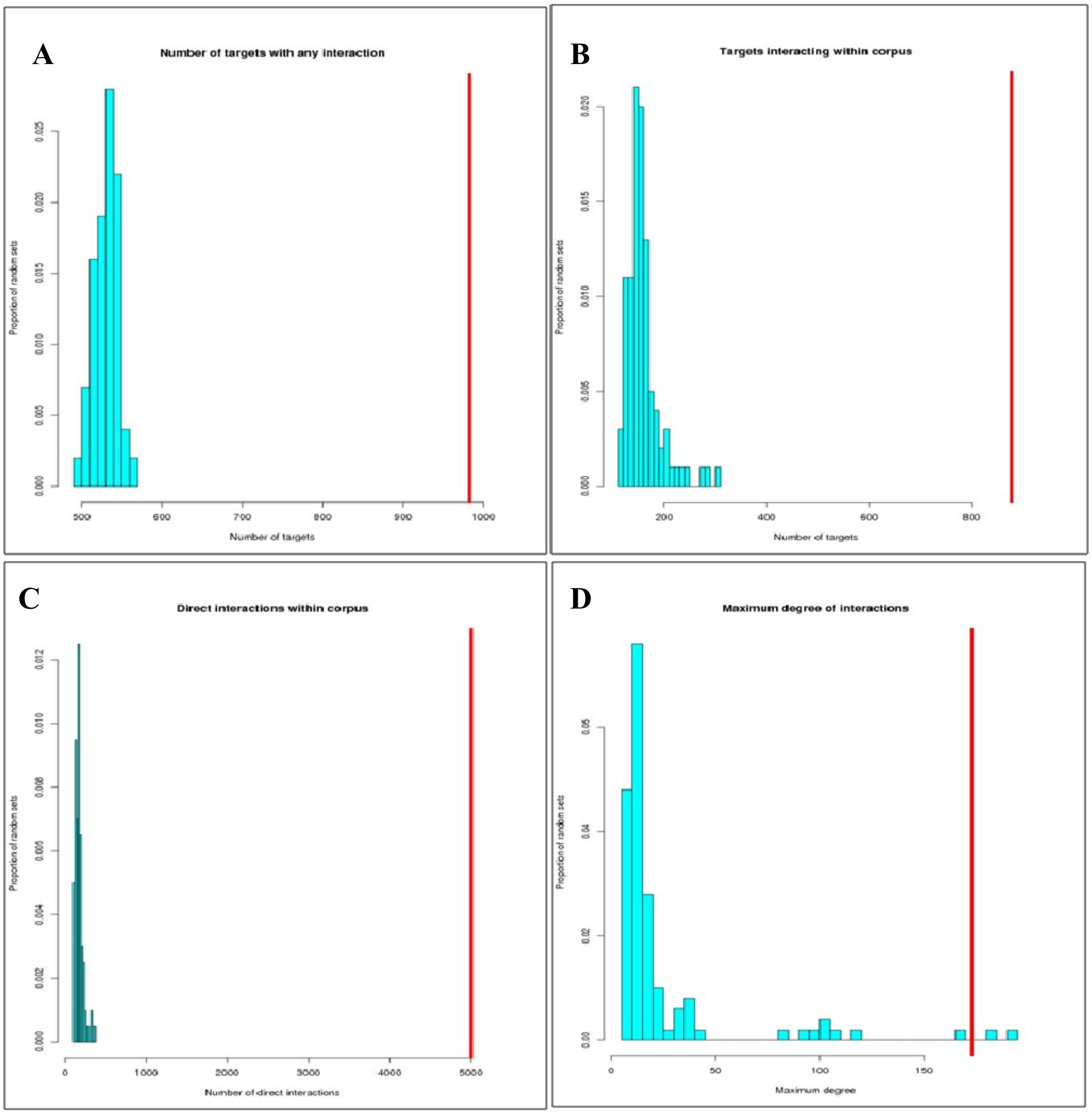

Figure 3-1. Histograms of Randomly Generated PPI Networks

The blue histograms illustrate the distributions of 100 randomly generated networks, while the red line indicates the ROS-diabetes targets. The network of the ROS-diabetes targets is significantly different from the 100 randomly generated networks, indicating the overlap of ROS-diabetes targets with respect to the topic “Reactive Oxygen Species and Diabetes”

3.3.3 FUNCTIONAL ENRICHMENT ANALYSES OF THE ROS-DIABETES TARGETS

Functional enrichment analyses of the 1,026 ROS-diabetes targets were performed to identify over-represented biological functions of the ROS-diabetes targets. After Benjamini-Hochberg multiple testing corrections, a total of 189 molecular functions, 450 biological processes, 73 cellular components and 341 pathways were significantly enriched in the ROS-diabetes targets (see the ROS-diabetes webpage for more details [123]). Table 3-3 lists the top 3 most over-represented GO terms and pathways ranked by p-values of Fisher's exact test: e.g., apoptosis, oxidoreductase activity and insulin signaling pathway.

Table 3-3. Enriched Functions of 1,026 ROS-diabetes Targets

Category	Term	#target	p-value	Fold
Biological Processes GO	metabolic process	113	3.40E-26	3.3
	protein amino acid phosphorylation	98	2.90E-24	3.5
	response to hypoxia	36	8.80E-24	12
Molecular Functions GO	protein binding	514	2.80E-71	2.1
	oxidoreductase activity	103	1.50E-31	4.2
	transferase activity	148	1.70E-26	2.7
Cellular Components GO	cytoplasm	381	1.50E-57	2.3
	extracellular region	220	9.10E-44	2.9
	mitochondrion	154	6.30E-43	3.9
Pathway	Focal adhesion	75	2.40E-42	9.4
	Apoptosis	49	6.70E-35	14.5
	MAPK signaling pathway	73	4.30E-34	6.9

'#target' refers to the number of ROS-diabetes targets with each biological function with Benjamini-Hochberg adjusted p-values. Fold is the ratio of targets from the ROS-diabetes set to the complete HUGO gene set.

3.3.4 IDENTIFICATION OF OVER-REPRESENTED ROS-DIABETES TARGETS

To identify the ROS-diabetes targets highly over-represented in ROS-diabetes literature, three sets of background corpora of the same size (n=1,154 documents) were generated using the same journal, volume and issue approach. The overlap among the three background sets in terms of documents and identified targets are illustrated in Figure 3-2. Approximately 90% of the selected background documents were unique to the individual set, while 50% of the identified targets were identified in at least one of the three background document sets. The frequencies of the identified targets were compared among the background sets for significant differences. None of the targets had a p-value < 0.05 after Benjamini-Hochberg corrections, indicating no significant difference among the targets from the three different background sets (Table 3-4).

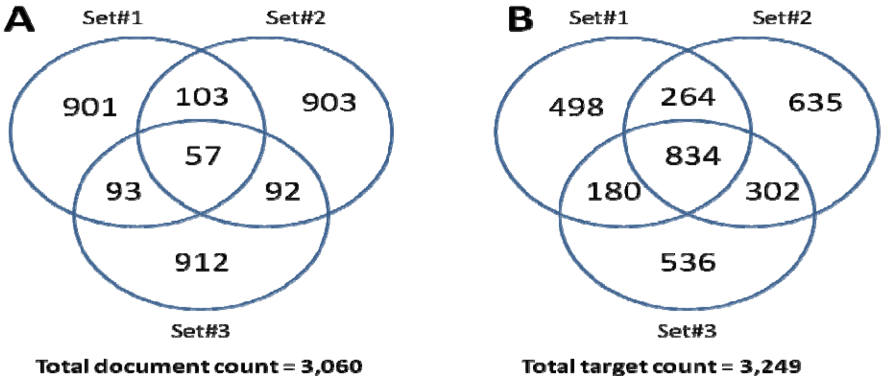

Figure 3-2. Venn Diagrams of Document Compositions and Identified Targets of the Randomly Generated Background Sets.

Approximately 90% of the selected background documents were unique to individual set (A), while 50% of the identified targets were identified in at least two of the three background document sets (B).

Table 3-4. Comparison of Target Frequencies among Three Background Sets.

	Symbol	Name	p-value	BH adjusted p-value
BG Set#1 Vs BG Set#2	PDHX	pyruvate dehydrogenase complex, component X	3.82E-02	1
	IL2	interleukin 2	4.53E-02	1
	MPO	myeloperoxidase	4.83E-02	1
	PLAU	plasminogen activator, urokinase	6.22E-02	1
	SLC12A1	solute carrier family 12 (sodium/potassium/chloride transporters), member 1	6.22E-02	1
	JAK2	Janus kinase 2 (a protein tyrosine kinase)	6.49E-02	1
	ACACA	acetyl-Coenzyme A carboxylase alpha	6.49E-02	1
	MET	met proto-oncogene (hepatocyte growth factor receptor)	6.49E-02	1
	PTEN	phosphatase and tensin homolog (mutated in multiple advanced cancers 1)	6.49E-02	1
	ACTC1	actin, alpha, cardiac muscle 1	6.90E-02	1
BG Set#1 Vs BG Set#3	GFAP	glial fibrillary acidic protein	2.21E-02	1
	CD44	CD44 molecule (Indian blood group)	3.11E-02	1
	MYST2	MYST histone acetyltransferase 2	3.11E-02	1
	FCGR3A	Fc fragment of IgG, low affinity IIIa, receptor (CD16a)	3.11E-02	1
	IGF1	insulin-like growth factor 1 (somatomedin C)	3.44E-02	1
	IL6	interleukin 6 (interferon, beta 2)	3.62E-02	1
	AGT	angiotensinogen (serpin peptidase inhibitor, clade A, member 8)	4.36E-02	1
	PLCG1	phospholipase C, gamma 1	5.14E-02	1
	CBL	Cas-Br-M (murine) ecotropic retroviral transforming sequence	6.22E-02	1
	NES	nestin	6.22E-02	1
BG Set#2 Vs BG Set#3	ACTC1	actin, alpha, cardiac muscle 1	1.46E-03	1
	IGF1	insulin-like growth factor 1 (somatomedin C)	6.06E-03	1
	VWF	von Willebrand factor	1.87E-02	1
	PLCG1	phospholipase C, gamma 1	2.20E-02	1
	PLCB1	phospholipase C, beta 1 (phosphoinositide-specific)	2.33E-02	1
	CS	citrate synthase	3.11E-02	1
	ALPP	alkaline phosphatase, placental (Regan isozyme)	3.82E-02	1
	NPY	neuropeptide Y	5.66E-02	1
	CD28	CD28 molecule	6.22E-02	1
	MYST2	MYST histone acetyltransferase 2	6.22E-02	1

Comparisons of the ROS-diabetes targets against these background sets revealed 53 highly over-represented ROS-diabetes targets as listed in Table 3-5. These 53 targets were significant ($p\text{-value} < 0.05$) against all three background sets and significant following Benjamini-Hochberg corrections (BH adjusted $p\text{-value} < 0.05$) against at least two of the three background sets. SOD1 was the most over-represented in the ROS-diabetes targets.

3.3.5 FUNCTIONAL ENRICHMENT ANALYSES OF THE OVER-REPRESENTED ROS-DIABETES TARGETS

Functional enrichment analyses of the 53 ROS-diabetes targets were performed to identify over-represented biological functions. Following Benjamini-Hochberg correction, a total of 65 molecular functions, 209 biological processes, 26 cellular components and 108 pathways were significantly over-represented when compared against all the HUGO genes (see the ROS-diabetes webpage for more details [123]). Table 3-6 shows the top three most significantly over-represented GO terms and pathways ranked by p-values of Fisher's exact test. GO terms related to oxidative stress such as "superoxide metabolic process", "superoxide release", "electron carrier activity" and "mitochondrion" were highly over-represented in the 53 ROS-diabetes targets.

Table 3-5. Fifty-three Targets Over-represented in ROS-diabetes Literature

Rank	Symbol	HUGO_ID	Name	#Paper	BG #1	BG #2	BG #3
1	SOD1	11179	superoxide dismutase 1, soluble (amyotrophic lateral sclerosis 1 (adult))	368	3.1E-84	2.0E-78	2.0E-78
2	CAT	1516	catalase	241	2.1E-50	3.9E-44	3.9E-44
3	NOX5	14874	NADPH oxidase, EF-hand calcium binding domain 5	177	3.1E-42	3.6E-39	2.1E-37
4	INS	6081	insulin	503	5.9E-41	2.0E-43	2.3E-39
5	XDH	12805	xanthine dehydrogenase	133	1.5E-30	1.2E-28	8.8E-28
6	PRKCA	9393	protein kinase C, alpha	194	7.1E-23	6.4E-26	8.9E-24
7	NCF1	7660	neutrophil cytosolic factor 1, (chronic granulomatous disease, autosomal 1)	72	7.6E-19	7.7E-16	8.7E-16
8	NOS3	7876	nitric oxide synthase 3 (endothelial cell)	115	1.6E-18	3.9E-16	7.6E-18
9	SOD2	11180	superoxide dismutase 2, mitochondrial	85	2.1E-18	7.7E-16	3.8E-15
10	CYBA	2577	cytochrome b-245, alpha polypeptide	69	4.2E-17	5.0E-13	6.9E-14
11	NOS2A	7873	nitric oxide synthase 2A (inducible, hepatocytes)	144	3.9E-16	5.2E-12	4.5E-14
12	AGT	333	angiotensinogen (serpin peptidase inhibitor, clade A, member 8)	131	1.8E-14	1.4E-09	3.5E-08
13	AKR1B1	381	aldo-keto reductase family 1, member B1 (aldose reductase)	61	8.0E-13	9.5E-13	3.6E-11
14	CYBB	2578	cytochrome b-245, beta polypeptide (chronic granulomatous disease)	49	4.0E-12	2.6E-09	5.8E-11
15	NOS1	7872	nitric oxide synthase 1 (neuronal)	82	4.9E-12	3.7E-10	4.7E-09
16	NCF2	7661	neutrophil cytosolic factor 2 (65kDa, chronic granulomatous disease, autosomal 2)	50	2.4E-11	1.5E-09	3.8E-08
17	CYCS	19986	cytochrome c, somatic	81	8.7E-10	2.2E-10	2.1E-10
18	HBB	4827	hemoglobin, beta	101	1.4E-08	5.9E-10	2.2E-08
19	GSR	4623	glutathione reductase	61	1.4E-08	4.8E-08	4.8E-08
20	UCP1	12517	uncoupling protein 1 (mitochondrial, proton carrier)	38	4.1E-07	2.1E-06	9.7E-06
21	NOX4	7891	NADPH oxidase 4	31	6.2E-07	2.3E-04	2.7E-05
22	PARP1	270	poly (ADP-ribose) polymerase family, member 1	37	7.1E-07	1.1E-07	5.3E-05
23	UCP2	12518	uncoupling protein 2 (mitochondrial, proton carrier)	34	7.0E-07	4.5E-06	2.1E-05
24	HBA1	4823	hemoglobin, alpha 1	30	1.1E-06	1.2E-06	9.3E-06
25	ALB	399	albumin	179	7.0E-06	4.9E-06	1.7E-06
26	NOX1	7889	NADPH oxidase 1	30	8.2E-06	8.6E-06	9.7E-06
27	NFKB1	7794	nuclear factor of kappa light polypeptide gene enhancer in B-cells 1 (p105)	90	9.4E-06	1.2E-04	4.5E-04
28	VEGFA	12680	vascular endothelial growth factor A	57	2.6E-04	1.9E-04	4.1E-03
29	SOD3	11181	superoxide dismutase 3, extracellular	18	2.5E-04	8.1E-02	3.4E-02
30	REN	9958	renin	51	3.6E-04	2.2E-02	7.2E-02

Rank	Symbol	HUGO_ID	Name	#Paper	BG #1	BG #2	BG #3
31	MPO	7218	myeloperoxidase	28	5.7E-04	2.4E-01	5.1E-02
32	SORD	11184	sorbitol dehydrogenase	15	1.8E-03	1.9E-03	1.8E-03
33	COL4A1	2202	collagen, type IV, alpha 1	15	1.8E-03	1.3E-02	1.8E-03
34	TGFA	11765	transforming growth factor, alpha	46	2.1E-03	3.5E-02	3.5E-04
35	ACE	2707	angiotensin I converting enzyme (peptidyl-dipeptidase A) 1	69	3.8E-03	1.1E-02	1.1E-02
36	AGTR1	336	angiotensin II receptor, type 1	36	3.7E-03	4.9E-02	1.8E-03
37	G6PD	4057	glucose-6-phosphate dehydrogenase	19	5.6E-03	3.7E-01	2.1E-01
38	CP	2295	ceruloplasmin (ferroxidase)	13	6.2E-03	3.1E-01	2.9E-01
39	NCF4	7662	neutrophil cytosolic factor 4, 40kDa	16	6.7E-03	9.9E-04	9.9E-04
40	MT-CYB	7427	mitochondrially encoded cytochrome b	15	1.3E-02	1.3E-02	1.3E-01
41	DUOX1	3062	dual oxidase 1	11	2.2E-02	2.9E-01	1.1E-01
42	SERPINE1	8583	serpin peptidase inhibitor, clade E (nexin, plasminogen activator inhibitor type 1),	37	2.4E-02	2.5E-02	1.1E-03
43	GSTCD	25806	glutathione S-transferase, C-terminal domain containing	37	2.4E-02	3.8E-01	9.1E-02
44	COQ7	2244	coenzyme Q7 homolog, ubiquinone (yeast)	16	2.8E-02	1.9E-01	3.1E-02
45	RAC1	9801	ras-related C3 botulinum toxin substrate 1 (rho family, small GTP binding protein	18	3.0E-02	4.3E-01	7.8E-02
46	MAOB	6834	monoamine oxidase B	10	3.9E-02	4.1E-01	4.4E-01
47	UCP3	12519	uncoupling protein 3 (mitochondrial, proton carrier)	17	4.7E-02	1.7E-02	1.8E-02
48	VCAM1	12663	vascular cell adhesion molecule 1	29	5.4E-02	6.3E-02	3.5E-02
49	AKT1	391	v-akt murine thymoma viral oncogene homolog 1	75	5.5E-02	4.9E-02	6.4E-02
50	LEPR	6554	leptin receptor	21	8.7E-02	3.1E-01	1.4E-02
51	EDN1	3176	endothelin 1	38	8.8E-02	3.8E-01	2.6E-02
52	COL1A1	2197	collagen, type I, alpha 1	84	8.7E-02	2.6E-02	1.7E-01
53	CCL2	10618	chemokine (C-C motif) ligand 2	38	2.0E-01	4.9E-02	1.0E-02

‘ #Paper’ is the number of papers in ROS-diabetes corpus; BG#1, BG#2 and BG#3 are Benjamini-Hochberg adjusted p-values between ROS-diabetes targets and background sets.

Table 3-6. Enriched Functions of the 53 Over-represented Targets in ROS-diabetes Literature

Category	Term	# target	p-value	Fold
Biological Processes GO	superoxide metabolic process	7	3.70E-15	303
	electron transport	13	1.50E-12	16
	superoxide release	5	4.20E-11	298
Molecular Functions GO	electron carrier activity	15	1.80E-17	27
	oxidoreductase activity	18	2.20E-16	14
	iron ion binding	15	4.20E-16	21
Cellular Components GO	mitochondrion	13	9.90E-08	6
	extracellular space	10	6.60E-07	8
	soluble fraction	7	3.20E-06	11
Pathway	Leukocyte transendothelial migration	9	6.40E-12	36
	Small cell lung cancer	7	1.00E-09	38
	Formation of Platelet plug	6	1.10E-08	41

3.3.6 GENE EXPRESSION CHANGE IN DIABETES

Two groups of DBA/2J mice exhibited significantly different levels of glycosylated hemoglobin (%GHb). The mean ± SEM were 6.2 ± 0.3 for the non-diabetic control group and for 14.0 ± 0.8 for the diabetic group ($p<0.001$), indicative of prolonged hyperglycemia in the diabetic group [121]. DRG were harvested from these animals for gene expression assays. Nine genes were selected from the top ranked ROS-diabetes targets: superoxide dismutase 1 (Sod1), catalase (Cat), xanthine dehydrogenase (Xdh), protein kinase C alpha (Prkca), neutrophil cytosolic factor 1 (Ncf1), nitric oxide synthase 3 (Nos3), superoxide dismutase 2 (Sod2), cytochrome b-245 alpha (Cyba), and nitric oxide synthase 1 (Nos1). Eight genes exhibited differential expression between diabetic and non-diabetic mice (P < 0.05) as shown in Figure 3-3. Cat, Sod1, Sod2, Prkca, and Nos1 expression levels were decreased, while Ncf1, Xdh, Cyba expression levels were increased in diabetes.

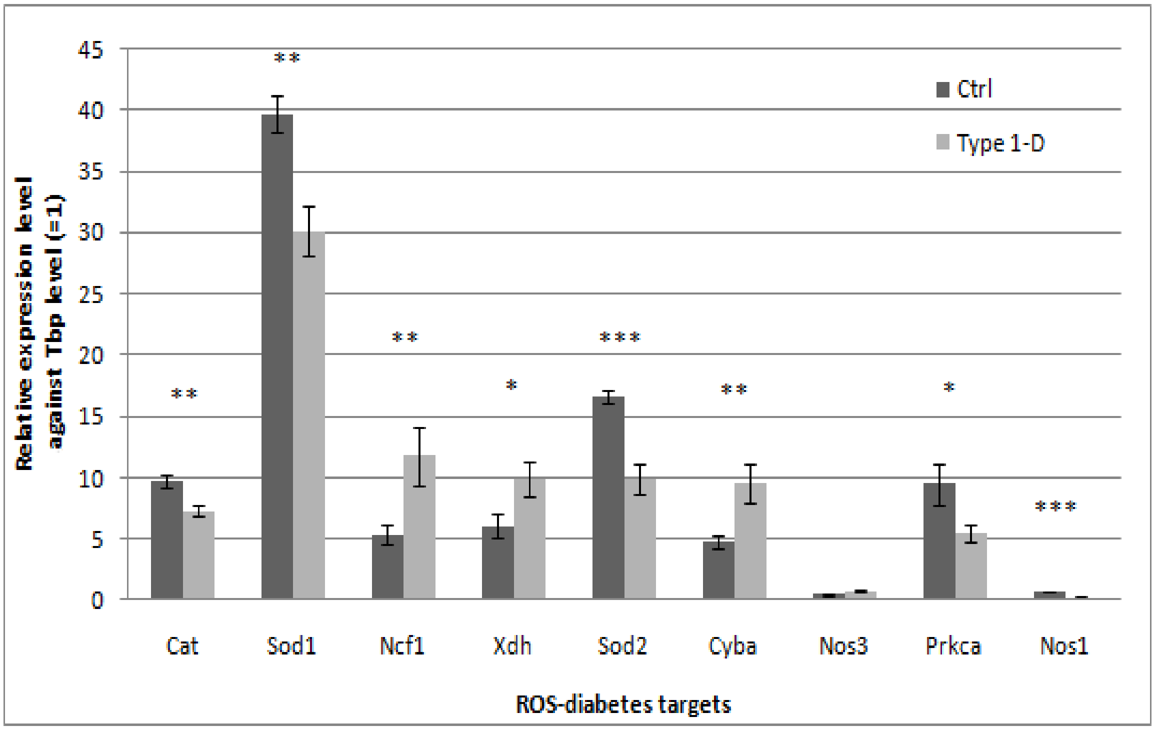

Figure 3-3. Gene Expression Levels of 9 ROS-diabetes Targets in DRG Examined by Real-time RT-PCR

Expression levels are relative to Tbp, an internal control (error bar = Standard Error Mean) (*, P < 0.05; **, P < 0.01; and ***, P < 0.001). Eight (Cat, Sod1, Ncf1, Xdh, Sod2, Cyba, Prkca, and Nos1) out of the 9 ROS-diabetes genes were significantly regulated by diabetes.

3.4 DISCUSSION AND CONCLUSIONS

Reactive oxygen species (ROS) are products of normal energy metabolism and play important roles in many other biological processes such as the immune response and signaling cascades [113-115]. As mediators of cellular damage, ROS are implicated in pathogenesis of multiple diseases including diabetic complications [124-127]. With the aid of literature mining technology, we collected 1,026 possible ROS-related targets from a set of biomedical literature indexed with both ROS and diabetes.

Fifty-three targets were significantly over-represented in the ROS-diabetes papers when compared against three background sets. Depending on how the background set is defined, the over-represented targets may vary widely. An ideal background set would be the entire PubMed set; however, this is not possible due to limited access to full texts and computational resources. An alternative method would be to use only abstracts in PubMed, but this may not fully represent the literature. In the present study, background documents were randomly selected from the same journal, volume, and issue of the 1,154 ROS-diabetes documents, which were not indexed with "Reactive Oxygen Species"[MeSH] nor "Diabetes Mellitus"[MeSH]. This approach maintained the background corpora not far from the ROS and diabetes context.

The gene expression levels of nine targets selected from the 53 over-represented ROS-diabetes targets were measured in diabetic and non-diabetic DRG. Our laboratory is particularly interested in deciphering the underlying mechanisms of diabetic neuropathy, a major complication of diabetes. Data published by our laboratory both *in vitro* and *in vivo* confirm the negative impact of oxidative stress in complication-prone neuronal tissues like DRG [22, 26, 28, 118]. In an effort to obtain diabetic neuropathy specific targets, SciMiner was employed to further analyze a subset of the ROS-diabetes papers (data not shown). Nerve growth factor (NGF) was identified as the most over-represented target in this subset when compared to the full ROS-diabetes set; however, the Benjamini-Hochberg adjusted p-value of NGF was not statistically significant (P = 0.06). The relatively small numbers of papers and associated targets may have contributed to this non-significance. Therefore, the candidate targets for gene expression validation were selected from among the 53 over-represented ROS-diabetes targets derived from the full ROS-diabetes corpus.

Among the tested genes, the expression levels of Cat, Sod1, Sod2, Prkca, and Nos1 were decreased, while the expression levels of Ncf1, Xdh, and Cyba were increased under diabetic conditions. Cat, Sod1, and Sod2 are responsible for protecting cells from oxidative stress by destroying superoxides and hydrogen peroxides [26, 27, 116, 117]. Decreased expression of these genes may result in oxidative stress [128]. Increased expression of Cyba and Ncf1, subunits of superoxide-generating nicotinamide adenine dinucleotide phosphate (NADPH) oxidase complex [127], also supports enhanced oxidative stress. Xdh and its inter-convertible form, Xanthine oxidase (Xod), show increased activity in various rat tissues under oxidative stress conditions with diabetes [129], and also showed increased expression in diabetic DRG in the current study.

Unlike the above concordant genes, protein kinase C and nitric oxide synthases did not exhibit predicted expression changes in diabetes. Protein kinase C activates NADPH oxidase, further promoting oxidative stress in the cell [130, 131]. Decreased expression of Prkca in our diabetic DRG is not parallel with expression levels of other enzymes expected to increase oxidative stress. Between the two nitric oxide synthases tested in the present study, Nos1 (neuronal) expression was significantly decreased ($P<0.001$) in diabetes, while Nos3 (endothelial) expression was not significant ($P = 0.06$). The neuronal Nos1 is expected to play a major role in producing nitric oxide, another type of highly reactive free radical. Thus, with some exceptions, the majority of the differentially expressed genes in DRG show parallel results to the known activities of these targets in diabetes, suggesting enhanced oxidative stress in the diabetic DRG.

Among the 53 over-represented ROS-diabetes targets, SOD1 was the most over-represented and was differentially expressed under diabetic and non-diabetic conditions. To the best of our knowledge, no published study has investigated the role of SOD1 in the onset and/or progression of diabetic neuropathy. Mutations of SOD1 have long been associated with the inherited form of amyotrophic lateral sclerosis (ALS) [132] and the theory of oxidative stress-based aging [133]. Early reports indicate that knockout of the SOD1 gene does not affect nervous system development [134], although recovery following injury is slow and incomplete [135, 136]. With respect to diabetes, SOD1 KO accelerates the development of diabetic nephropathy [137] and cataract formation [138]. Thus, examining the SOD1 KO mouse as a model of diabetic neuropathy would be a reasonable follow-up study.

One limitation of the current approach using literature mining technology is incorrect or missed identification of the mentioned targets within the literature. Based on a performance evaluation using a standard text set BioCreAtIvE (Critical Assessment of Information Extraction systems in Biology) version 2 [139], SciMiner achieved 87.1% recall (percentage identification of targets in the given text), 71.3% precision (percentage accuracy of identified target) and 75.8% *F*-measure (harmonious average of recall and precision = (2 x recall x precision) / (recall + precision)) before manual revision [42]. In order to improve the accuracy of SciMiner's results, each target was manually reviewed and corrected by checking the sentences in which each target was identified. Approximately, 120 targets (~10% of the initially identified targets from the ROS-diabetes papers) were removed during the manual review process. The overall accuracy is expected to improve through the review process; however, the review process did not address targets missed by SciMiner, since we did not thoroughly review individual papers. Instead, five missed targets, whose associations with ROS-diabetes literature were available in the NCBI Gene2PubMed database, were added to the final ROS-diabetes target list.

Even with these limitations, the present approach enabled us to collect a comprehensive list of ROS and diabetes related targets and led us to confirm the biological relevance to diabetic neuropathy of the selected ROS-diabetes targets. Using SciMiner to identify significantly enriched targets is applicable to any disease or topic of interest and will shorten the time needed to assess the literature for relevant and potential biological markers of disease.

CHAPTER 4

GENE EXPRESSION PROFILES PREDICTIVE OF DIABETIC NEUROPATHY PROGRESSION

4.1 INTRODUCTION

Twenty-three million Americans have diabetes and the incidence is increasing by 5% per year [1]. The most common complication of diabetes is peripheral neuropathy occurring in approximately 60% of all diabetic patients [2, 3]. An additional fifty-seven million Americans have impaired glucose tolerance, or pre-diabetes, and up to 30% of these patients will exhibit peripheral neuropathy at diagnosis [1, 140, 141]. Diabetic neuropathy (DN) is characterized by progressive loss of peripheral nerve axons, resulting in decreased sensation, pain, and eventually complete loss of sensation.

In greater than 50% of patients with DN, there is substantial and irreversible nerve damage prior to the development of noticeable symptoms. The discovery of DN biomarkers measurable prior to the onset of permanent damage would permit aggressive early therapy of DN to preserve nerve function. Biomarkers that are highly predictive of the development and worsening of diabetic complications are only available for diabetic nephropathy [142-145]. Currently, no biomarkers exist for DN, making it impossible to detect until clinically obvious symptoms and signs appear, at which point irreparable damage has occurred.

Our goal is to develop rational treatment paradigms to halt or reverse DN progression and to identify individuals at high risk of developing DN. We are in possession of a unique repository of human sural nerve biopsies with matched blood chemistries, electrophysiology, and nerve function tests from participants in a large randomized placebo-controlled clinical trial [21, 106]. Our initial analyses revealed that after correcting for baseline DN severity, the strongest factor correlating with loss of sural myelinated fiber density (MFD) was serum triglycerides [21]. These results indicate a role for dyslipidemia in the progression of DN.

In the current study, we use a bioinformatics approach to identify genes and pathways altered by DN over the course of one year. We report the first high-throughput genome-wide expression study of human sural nerve biopsies obtained from patients with DN. Gene expression profiles were examined in sural nerve samples from two groups of patients with either fast progressing or slow/non-progressing DN by high-throughput Affymetrix microarray. A series of bioinformatics analyses were employed to analyze differential gene expression profiles between the two groups and revealed gene networks and pathways linked to the progression of DN. Computational predictive models, based on the expression profiles of selected genes, were developed and correctly classified patients as exhibiting either progressing or non-progressing DN. Our best predictive models included 14 genes and demonstrated a prediction accuracy of 92% in a separate test cohort of patients. To our knowledge, these gene sets provide the first predictive measures of human DN progression and may be used to explore new pathways underlying disease pathogenesis. In addition, it provides a unique starting point for targeted serum biomarker development to identify patients at risk for DN prior to the onset of irreversible peripheral nerve damage.

4.2 RESEARCH DESIGN AND METHODS

4.2.1 HUMAN SURAL NERVE SAMPLES

Human sural nerve biopsies were obtained as part of a double-blind, placebo-controlled, 52-week clinical trial of acetyl-L-carnitine (ALC) for DN treatment [21, 106]. The inclusion and exclusion criteria were described previously [21, 106]. In brief, both type 1 and 2 diabetic patients were included, all with existing, mild neuropathy. Measures of nerve conduction velocity and sensory function were measured prior to the collection of a sural nerve biopsy (week 0 – denoted as the primary sample). Following 52 weeks of treatment, measures of DN were re-assessed and a second sural nerve biopsy was harvested (week 52 – denoted as the secondary sample).

By comparing changes in sural nerve myelinated fiber density (MFD) across the course of the study, our post-hoc analysis classified the patient samples into two groups: progressors and non-progressors. Patient samples in the progressor group lost ≥ 500 fibers/mm^2 between the primary and secondary biopsies, while patient samples in the non-progressor group lost ≤ 100 fibers/mm^2 over 52 weeks [21]. Primary and secondary biopsies from 36 patients (18 progressors and 18 non-progressors) were used in this study. The selection of patient samples from each group was adjusted for MFD at trial onset, insulin treatment, gender and type of diabetes. The use of the human sural nerve samples was approved by the Institutional Review Board for Human Subject Research at the University of Michigan.

4.2.2 RNA PREPARATION

Total RNA was isolated from a 1 cm segment of each sural nerve biopsy using a commercially available kit (RNeasy Mini Kit; QIAGEN, Inc., Valencia, CA), including an on-column deoxyribonuclease digestion and following the manufacturer's protocol. RNA quality and quantity were assessed by microfluid electrophoresis using an RNA 6000 Pico LabChip on a 2100 Bioanalyzer (Agilent Technologies, Inc., Santa Clara, CA). Samples with a minimum RNA Integrity Number (RIN) of 6.5 were used for Microarray hybridization [146].

4.2.3 AFFYMETRIX MICROARRAYS

Samples meeting the RNA quality criteria were analyzed by microarray. Total RNA (75 ng) was amplified and biotin labeled using the Ovation Biotin-RNA Amplification and Labeling System (NuGEN Technologies, Inc., San Carlos, CA) according to the manufacturer's protocol. The University of Michigan Comprehensive Cancer Center Affymetrix and Microarray Core Facility (University of Michigan, Ann Arbor, MI) performed the amplification and hybridization using the Affymetrix GeneChip Human Genome U133 Plus 2.0 Array. Intensities of target hybridization to respective probe features were detected by laser scan of the array. Image files were generated by Affymetrix GeneChip software (MAS5).

4.2.4 DATA ANALYSIS

4.2.4.1 Quality Assessment and Data Preprocessing

The Affymetrix CEL files were initially analyzed using a local version of the GenePattern genomic analysis platform from the Broad Institute [80]. The samples were Robust Multi-array Average (RMA) normalized using the BrainArray Custom CDF HGU133Plus2_Hs_ENTREZG version 12 [147]. Microarray quality was assessed using the probe-level modeling (PLM) and quality metrics provided by the *affy* package of BioConductor [148-150]. Outlier arrays that did not cluster with other arrays in Principal Component Analysis (PCA) results were excluded from further analyses [151].

4.2.4.2 Identification of differentially expressed genes (DEGs)

Two independent analysis platforms were employed to identify DEGs between different biological groups (secondary biopsies; progressors and non-progressors): the GenePattern platform using the standard RMA based probe-set approach and ChipInspector (CI; version 2.1; Genomatix Software GmbH, Munich, Germany). The RMA approach averages normalized expression levels across all probes for the gene (probe set level analysis) whereas Genomatix ChipInspector calculates the change in each probe (probe level analysis) [82]. Genes were deemed as DEGs using Cyber-T [152], based on a Bayesian regularized t-test, p-value < 0.05 in the RMA approach and a False Discovery Rate (FDR) < 0.1% using ChipInspector [82] with a minimum of 4 probes per transcript.

4.2.4.3 Functional Enrichment Analyses

The Database for Annotation, Visualization and Integrated Discovery (DAVID) (http://david.abcc.ncifcrf.gov/) [88, 89], ConceptGen (http://conceptgen.ncibi.org) [90], and LRpath [91] were used to identify over-represented biological functions and pathways among the DEGs.

4.2.4.4 Network Analysis

A gene co-citation network of the DEGs was generated by Genomatix BiblioSphere (Genomatix Software GmbH, Munich, Germany) using a sentence level co-citation filter. This network allows us to visualize an entire network of DEGs and their biological connections identified in the literature. The topology of the network was further analyzed by the Fast-Greedy community-structure identification algorithm, implemented in Cytoscape plug-in GLay (http://brainarray.mbni.med.umich.edu/sugang/glay/) to identify coherent sub-networks. Identified sub-networks were subjected to functional enrichment analyses by DAVID to reveal the over-represented biological functions within each sub-network.

4.2.5 PREDICTIVE MODELING USING GENE EXPRESSION PROFILES

The expression profiles of the DEGs were evaluated for their ability to predict progressors versus non-progressors using ridge regression modeling [100-102]. The gene expression profiles of the secondary samples, excluding those samples with paired primary samples, were used as the training set (13 progressors and 11 non-progressors). The expression profiles of the primary samples were used as the testing set (5 progressors and 7 non-progressors). Three different sets of DEGs were used as predictors; set 1 included all 532 DEGs, set 2 contained 63 DEGs with a minimum fold-change of 1.5, and set 3 included 10 DEGs with a minimum fold-change of 2. To identify a set of genes with the least number of genes but with high prediction accuracy, the genes from set 2 were added to set 3 one at a time until the prediction accuracy of the expanded set reached the maximum level of accuracy.

4.2.6 REAL-TIME RT-PCR

The gene expression of eight DEGs identified by microarray was confirmed by real-time RT-PCR performed on five independent samples from each secondary group (progressor and non-progressor). Reverse transcription was performed using iScript cDNA Synthesis kit (Bio-Rad, Hercules, CA). Real-time PCR amplification and SYBR Green fluorescence detection were performed using iCycler iQ Real-time Detection System (Bio-Rad Laboratories, Hercules, CA). The fluorescence threshold value (C_T) was calculated using iCycler iQ system software and the levels were normalized to an endogenous reference gene TATA box binding protein (TBP) [122]. A Pearson correlation coefficient was calculated for each gene between the $\log_2$-transformed expression values as measured by microarray and the negative of the C_T by RT-PCR [153].

4.3 RESULTS

4.3.1 SAMPLE INFORMATION

Patient information regarding type and duration of diabetes, gender, body mass index and circulating lipids is provided in Table 4-1. The O'Brien score for neuropathy and baseline and final MFD are also listed. The only significant difference between the progressor and non-progressor groups was the change in MFD over 52 weeks. Eighty percent of the study participants had type 2 diabetes and 61% were treated with insulin.

Table 4-1. Patient Characteristics (n=36)

		Non-Progressor	Progressor	P-value
Gender	Male	10	11	1
	Female	8	7	
Diabetes Types	Type 1	3	4	1
	Type 2	15	14	
Insulin Treatment	Yes	11	11	1
	No	7	7	
Age (years)		54.7 ± 12.9	52.2 ± 10.3	0.524
Diabetes Duration (years)		10.8 ± 7.2	12.0 ± 7.3	0.622
Body Mass Index (kg/cm2)		30.0 ± 5.7	31.6 ± 10.3	0.568
Homoglobin A1C (%)		8.9 ± 1.6	9.2 ± 1.4	0.449
Triglyceride (mmol/L)		1.8 ± 0.8	2.7 ± 1.9	0.088
Cholesterol (mmol/L)		5.5 ± 0.8	5.5 ± 0.9	0.938
O'Brein Score		4179.7 ± 772.2	3854.5 ± 860.1	0.241
MFD_Base (fibers/mm2)		5133.2 ± 1139.2	5132.8 ± 1450.8	0.999
MFD_52Wks (fibers/mm2)		5256.8 ± 1200.0	4066.6 ± 1538.7	0.014 *
MFD_Change (fibers/mm2)		123.6 ± 209.2	-1066.2 ± 391.1	3.84E-13 ***

Continuous variables are reported as mean ± standard deviation. P-values were calculated by two-sample t-test for continuous variable and Fisher's exact test for categorical variables (*: $p < 0.05$, ***: $p < 0.001$) .

4.3.2 IDENTIFICATION OF DIFFERENTIALLY EXPRESSED GENES AND MICROARRAY QUALITY ASSESSMENT

Figure 4-1 illustrates how many sural nerve biopsies are used in this study. Fifty samples (14 primary (7 progressors and 7 non-progressors) and 36 secondary (18 progressors and 18 non-progressors) samples) met the RNA quality criteria and were examined by global gene expression profiling using the Affymetrix Human Genome U133 Plus 2.0 platform. Two outlier arrays from the primary progressor group were excluded from further analyses (data not shown) and one array from the secondary non-progressor group was also excluded due to a labeling error.

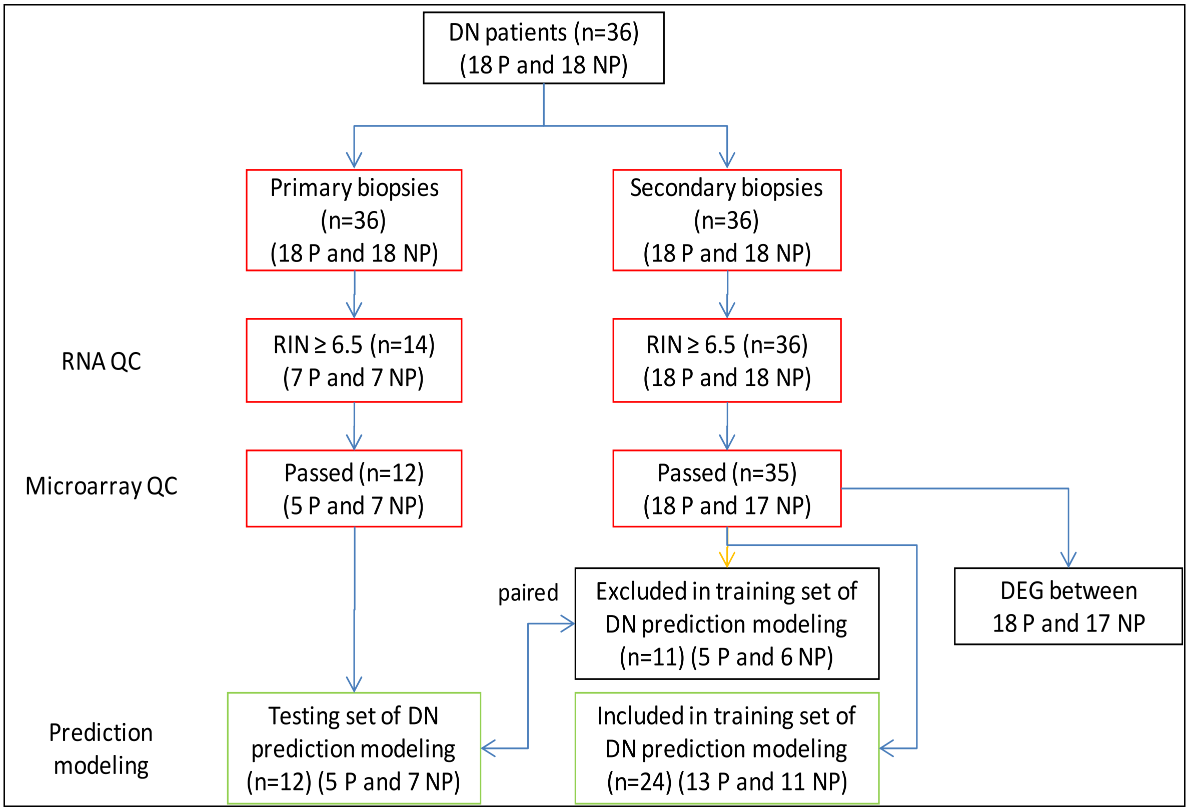

Figure 4-1. Primary and Secondary Biopsies Selection in the Present Study

Primary and secondary biopsies of 36 DN patients were included in this study. Samples with a minimum RIN of 6.5 were used for microarray hybridization. Two outlier arrays (primary) and one array with a mislabeling error (secondary) were excluded from further analyses. The secondary samples, excluding those samples with paired primary samples, were used as the training set (13 progressors and 11 non-progressors) for DN prediction modeling, and the primary samples were used as the testing set (5 progressors and 7 non-progressors). P denotes progressor and NP denotes non-progressor.

The changes in gene expression described below represent changes between secondary biopsies from the progressor (n=18) and non-progressor (n=17) groups tested by Cyber-T and ChipInspector. Of the 22,288 Entrez genes available on the array, 14,885 genes were expressed above background in at least one of the 47 samples. Genes with a Cyber-T Bayesian p-value of less than 0.05 and a ChipInspector FDR < 0.1% were considered DEGs. A total of 558 genes had a Bayesian p-value of less than 0.05, while 4,899 genes had a ChipInspector FDR < 0.1%. Only 532 genes deemed as a DEG by both methods were included for further analyses.

Technical validation of the microarray data was performed by real-time RT-PCR of 8 DEGs with a minimum fold-change of 1.5. A subset of 4 DEGs including cystatin SN, hepcidin antimicrobial peptide, MLX interacting peptide and beta A2 crystallin demonstrated strong positive correlations with the microarray data. Eosinophil-derived neurotoxin (RNASE2) and "family with sequence similarity 43, member B" (FAM43B) were negatively correlated with the microarray data. Real-time RT-PCR compared to microarray data revealed 63% of the DEGs were regulated in parallel (Table 4-2).

Table 4-2. Genes Tested by Real-time RT-PCR

Gene ID	Symbol	Description	Correlation
1469	CST1	cystatin SN	0.9864
57817	HAMP	hepcidin antimicrobial peptide	0.9373
51085	MLXIPL	MLX interacting protein-like	0.8397
1412	CRYBA2	crystallin, beta A2	0.7377
56605	ERO1LB	ERO1-like beta (S. cerevisiae)	0.4713
10804	GJB6	gap junction protein, beta 6, 30kDa	0.1152
6036	RNASE2	ribonuclease, RNase A family, 2 (liver, eosinophil-derived neurotoxin)	-0.0535
163933	FAM43B	family with sequence similarity 43, member B	-0.7099

4.3.3 FUNCTIONAL ENRICHMENT ANALYSES

Functional enrichment analyses of the 532 DEGs discovered in the secondary biopsies (progressor versus non-progressor) were performed to identify over-represented biological functions in Gene Ontology (GO) terms and pathways. DAVID identified 31 and 168 over-represented biological functions among the up- and down-regulated DEGs in the progressor group, respectively (DAVID p-value < 0.05). Table 4-3 lists a selected subset of the over-represented biological functions; the up-regulated genes in progressors (i.e., down-regulated in non-progressors) were enriched in "extracellular region", "defense response" and "inflammatory response" (Table 4-4), while down-regulated genes in progressors (i.e., up-regulated genes in non-progressors) were enriched in energy metabolism related functions such as "glucose metabolic process", "PPAR signalling pathway" and "regulation of lipid metabolic process" (Table 4-5).

Table 4-3. Over-represented Biological Functions in DEGs

	Biological function	Gene Count	p-value	Enrichment fold
Up-regulated in progressors	extracellular region	58	6.86E-07	1.9
	prostanoid metabolic process	4	2.24E-03	14.9
	defense response	18	5.99E-03	2.1
	inflammatory response	12	6.24E-03	2.6
	regulation of axonogenesis	5	8.34E-03	6.2
	response to wounding	15	1.75E-02	2.0
Down-regulated in progressors	chemical homeostasis	20	8.25E-06	3.3
	glucose metabolic process	11	1.21E-05	6.1
	glycerolipid metabolic process	11	2.01E-05	5.8
	PPAR signaling pathway	8	7.04E-05	7.6
	regulation of lipid metabolic process	9	4.93E-05	6.8
	response to insulin stimulus	8	1.67E-04	6.8

Table 4-4. DEGs Related to Defense Response and Inflammatory Response (Up-regulated Genes in Progressors)

		Entrez ID	Symbol	Description	P-value	Fold-change
defense response		136	ADORA2B	adenosine A2b receptor	0.0133	1.4
		2788	GNG7	guanine nucleotide binding protein (G protein), gamma 7	0.0336	1.2
		7033	TFF3	trefoil factor 3 (intestinal)	0.0127	1.5
		23601	CLEC5A	C-type lectin domain family 5, member A	0.0207	1.7
		57817	HAMP	hepcidin antimicrobial peptide	0.0025	2.4
		81035	COLEC12	collectin sub-family member 12	0.0152	1.2
	inflammatory response	140	ADORA3	adenosine A3 receptor	0.0327	1.3
		624	BDKRB2	bradykinin receptor B2	0.0273	1.2
		3075	CFH	complement factor H	0.0010	1.2
		4282	MIF	macrophage migration inhibitory factor (glycosylation-inhibiting factor)	0.0282	1.1
		4973	OLR1	oxidized low density lipoprotein (lectin-like) receptor 1	0.0003	1.6
		7852	CXCR4	chemokine (C-X-C motif) receptor 4	0.0388	1.3
		10344	CCL26	chemokine (C-C motif) ligand 26	0.0004	3.2
		10630	PDPN	podoplanin	0.0202	1.2
		25824	PRDX5	peroxiredoxin 5	0.0200	1.1
		53833	IL20RB	interleukin 20 receptor beta	0.0185	1.3
		57834	CYP4F11	cytochrome P450, family 4, subfamily F, polypeptide 11	0.0372	1.4
		148022	TICAM1	toll-like receptor adaptor molecule 1	0.0379	1.2

Table 4-5. DEGs Related to Lipid Metabolism and PPAR Signaling Pathway (Down-regulated Genes in Progressors)

	Entrez ID	Symbol	Description	P-value	Fold-change
regulation of lipid metabolic process	348	APOE	apolipoprotein E	0.0266	0.8
	2180	ACSL1	acyl-CoA synthetase long-chain family member 1	0.0379	0.8
	3952	LEP	leptin	0.0099	0.7
	5140	PDE3B	phosphodiesterase 3B, cGMP-inhibited	0.0126	0.7
	5468	PPARG	peroxisome proliferator-activated receptor gamma	0.0426	0.7
	8660	IRS2	insulin receptor substrate 2	0.0118	0.9
	9370	ADIPOQ	adiponectin, C1Q and collagen domain containing	0.0094	0.7
	51085	MLXIPL	MLX interacting protein-like	0.0083	0.7
	57104	PNPLA2	patatin-like phospholipase domain containing 2	0.0206	0.8
	100129500	LOC100129500	hypothetical LOC100129500	0.0485	0.7
PPAR signaling pathway	948	CD36	CD36 molecule (thrombospondin receptor)	0.0373	0.7
	2180	ACSL1	acyl-CoA synthetase long-chain family member 1	0.0379	0.8
	4199	ME1	malic enzyme 1, NADP(+)-dependent, cytosolic	0.0181	0.8
	5105	PCK1	phosphoenolpyruvate carboxykinase 1 (soluble)	0.0497	0.8
	5346	PLIN	perilipin	0.0092	0.7
	5468	PPARG	peroxisome proliferator-activated receptor gamma	0.0426	0.7
	6319	SCD	stearoyl-CoA desaturase (delta-9-desaturase)	0.0241	0.7
	9370	ADIPOQ	adiponectin, C1Q and collagen domain containing	0.0094	0.7

The DEGs were further analyzed by ConceptGen, another gene set enrichment analysis tool equipped with a network visualization capability. ConceptGen also applies additional biological information such as MeSH terms and differential expression profiles from the Gene Expression Omnibus (GEO). Figure 4-2 illustrates the over-represented concepts (sets of genes associated with common biological functions) of the up-regulated (1A) and down-regulated (1B) DEGs. Over-represented GO terms and pathways are similar to those identified by DAVID, for example, the concepts of "defense response" and "inflammatory response" in the up-regulated DEGs and the energy metabolism related concepts in the down-regulated DEGs. ConceptGen also identified MeSH terms that are highly associated with the DEGs. Interestingly, the down-regulated DEGs were enriched with MeSH terms such as lipids, fatty acids, triglycerides, cholesterol, and insulin suggesting decreased energy metabolism in fast progressing DN.

LRpath identified over-represented biological functions regarding GO terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. Unlike DAVID or ConceptGen, LRpath does not require users to provide a predefined set of DEGs. Instead, LRpath analyzes statistical significance (Cyber-T p-value in the current study) of all genes expressed above background in the microarray. LRpath identified 606 GO terms and 31 KEGG pathways, which were differentially expressed (LRpath p-value < 0.05). The most over-represented terms included metabolic process related terms such as "regulation of lipid metabolic process", "monocarboxylic acid metabolic process" and "PPAR signaling pathway".

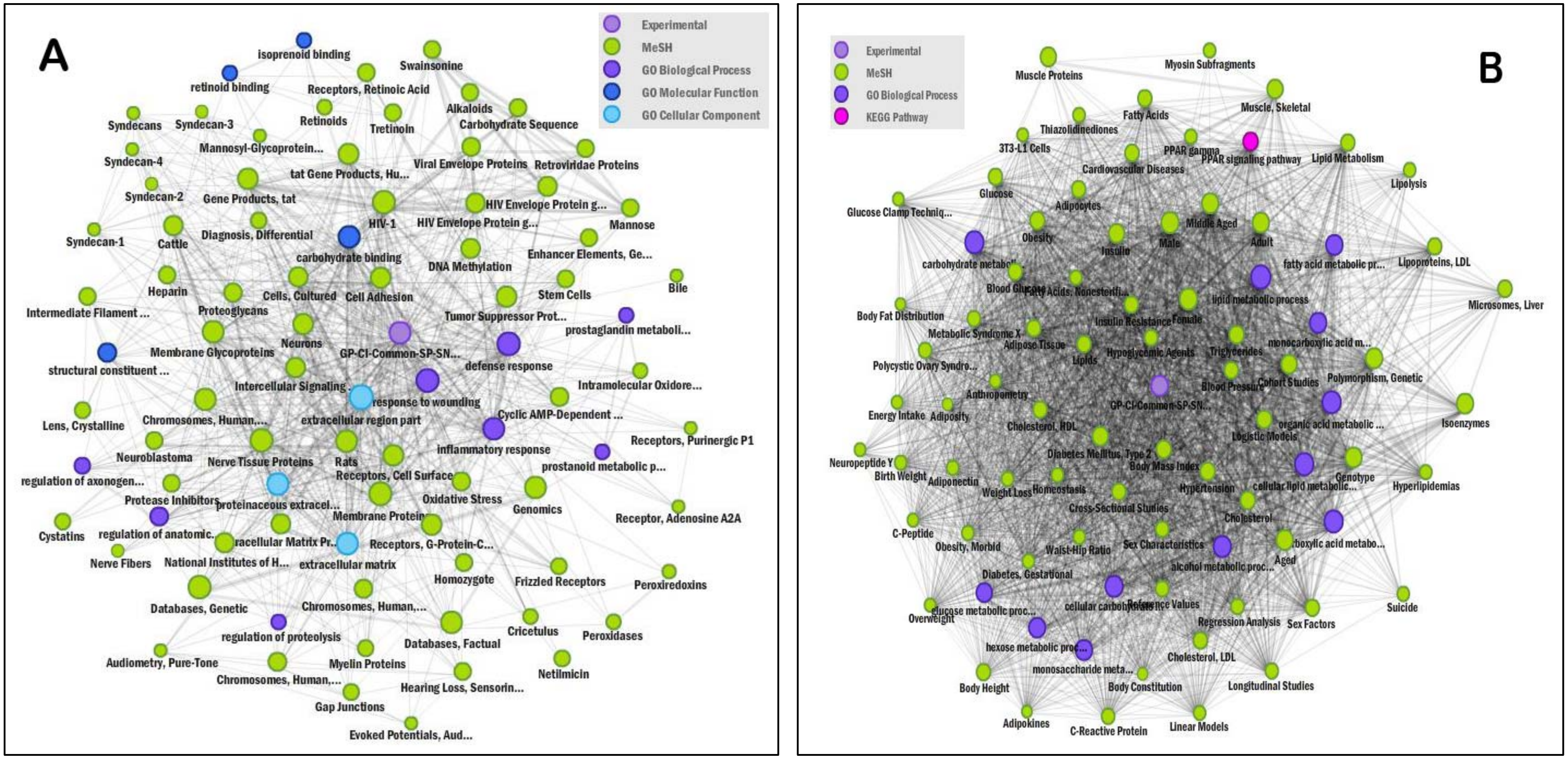

Figure 4-2. A Network of Over-represented Biological Concepts by ConceptGen

The concepts (gene sets) over-represented in the up-regulated genes (A) and down-regulated genes (B) in progressors. The center nodes in violet, titled as "GP-CI-Common-SP-SN…", refers to the DEGs

4.3.4 NETWORK ANALYSIS

Once over-represented biological functions of DEGs were identified, we examined potential relationships among DEGs. Figure 4-3 illustrates a literature derived gene network of the DEGs created by BiblioSphere based on sentence level co-citations of DEGs. The network is composed of five sub-networks centered on the 5 most connected genes: jun oncogene (JUN), leptin (LEP), serpin peptidase inhibitor E Type 1 (SERPINE1), apolipoprotein E (APOE) and PPARG. The complete network was further analyzed by a Cytoscape plug-in GLay to cluster the genes into subgroups based on the network structure. As depicted in Figure 4-4, 6 clusters with a minimum of 8 genes were identified by the Fast-Greedy algorithm [154] implemented in the GLay plug-in. Functional enrichment analyses of these sub-networks using DAVID identified representative biological functions within each cluster: cell death and inflammatory response for cluster 1, glucose and lipid metabolism for cluster 2, cell projection and axonogenesis for cluster 3, cellular homeostasis and cofactor metabolic process for cluster 4, cytoskeletal protein binding for cluster 5, and Wnt receptor signaling pathway for cluster 6.

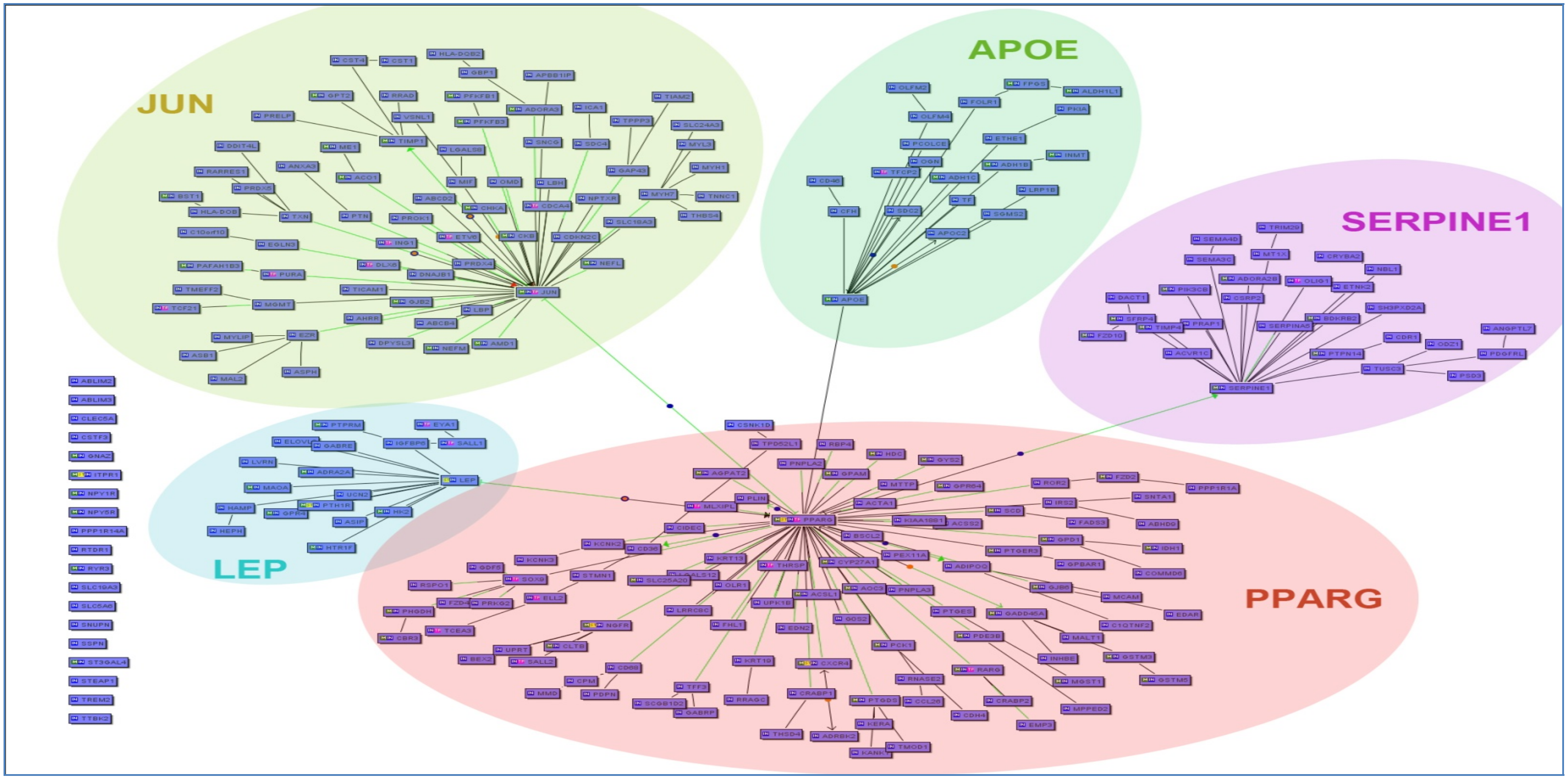

Figure 4-3. Gene Co-citation Network of DEGs by BiblioSphere PathwayEdition

A literature derived gene network of the DEGs was created by BiblioSphere PathwayEdition using sentence level co-citations of DEGs. The network is composed of five sub-networks centered on the 5 most connected genes: JUN, PPARG, LEP, SERPINE1 and APOE.

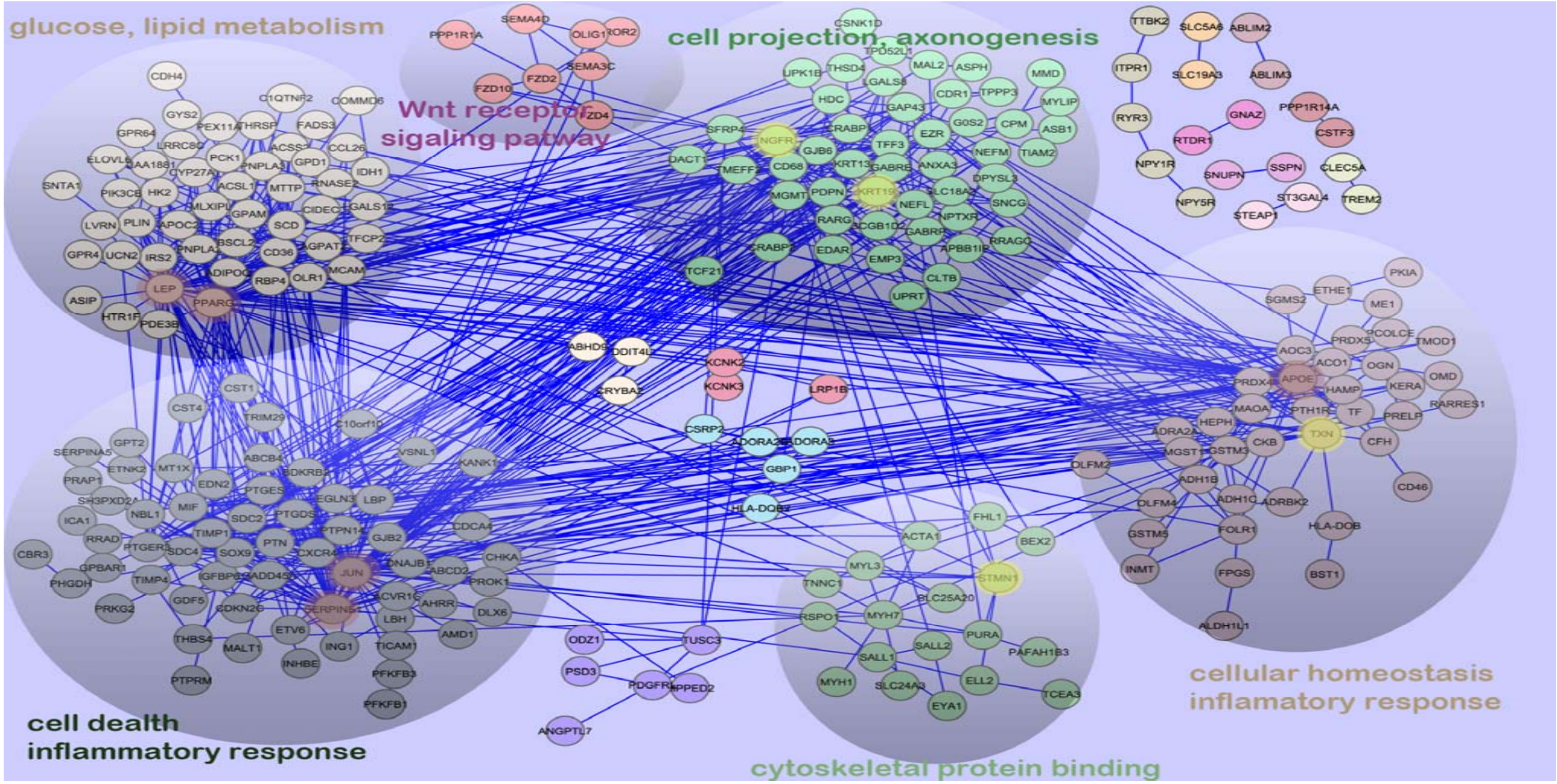

Figure 4-4. Gene Co-citation Network Clustered by Fast-Greedy Community Structuring Algorithm

The complete co-citation network of the DEGs was clustered based on network topology by Fast-Greedy algorithm implemented in the Cytoscape GLay plug-in. Nodes (genes) highlighted in red or yellow refer to the highly connected genes: nodes in red refer to the core genes in Figure 2, while nodes in yellow refer to highly connected genes that were not core genes in Figure 2.

4.3.5 PREDICTIVE MODELING OF DN PROGRESSION BASED ON GENE EXPRESSION PROFILES

Ridge regression models based on a subset of DEGs were used to predict the status of patient samples as progressors/non-progressors. Three different sets of DEGs were used in our initial models; set 1 included all 532 DEGs, set 2 contained 63 DEGs with a minimum fold-change of 1.5, and set 3 included 10 DEGs with a minimum fold-change of 2. The regression models were trained using the gene expression profiles of the 24 secondary samples obtained at the 52-week time point and then tested on 12 primary samples obtained at study initiation. Table 4-6 summarizes the performance of these models. Briefly, both models using DEG sets 1 and 2 achieved a prediction accuracy of 92% (11 correct predictions of 12), while the smallest model based on set 3 demonstrated a prediction accuracy of 50%.

Table 4-6. Performance of Predictive Models

Model	# of genes	classification accuracy	sensitivity (true progressor prediction)	specificity (true non-progressor prediction)
M1	529	0.92 (11/12)	0.80 (4/5)	1.00 (7/7)
M2	63	0.92 (11/12)	0.80 (4/5)	1.00 (7/7)
M3	10	0.5 (6/12)	0.40 (2/5)	0.57 (4/7)

In order to create a model using the smallest set of DEGs but with the same prediction accuracy, we began with set 3 and individual DEGs from set 2 were sequentially added. The predictive power was assessed following each addition until the new set achieved the original prediction accuracy. The result was 4 sets of 14 DEGs, beyond which the addition of new DEGs did not increase accuracy. Table 4-7 lists these 4 models, all of which achieved a prediction accuracy of 92% when tested on the primary samples. Each model includes 10 base DEGs from set 3 and 4 combinations of 11 DEGs from set 2 (Table 4-8).

Table 4-7. Predictive Models with 14 Genes Achieving a Prediction Accuracy of 92% on the Primary Sample

Model	classification accuracy	sensitivity (true progressor prediction)	specificity (true non-progressor prediction)
1	0.92 (11/12)	0.80 (4/5)	1.00 (7/7)
2	0.92 (11/12)	1.00 (5/5)	0.86 (6/7)
3	0.92 (11/12)	0.80 (4/5)	1.00 (7/7)
4	0.92 (11/12)	1.00 (5/5)	0.86 (6/7)

Iterative search identified 4 models composed of 14 achieving the highest prediction accuracy on the primary samples.

Table 4-8. The Gene Content of the 4 Models with 14 Genes

	Model#				EntrezID	Symbol	Description	Cyber-T P-value	Fold-Change
Base					1469	CST1	cystatin SN	0.0349	10.0
					10804	GJB6	gap junction protein, beta 6, 30kDa	0.0011	5.5
					10344	CCL26	chemokine (C-C motif) ligand 26	0.0004	3.2
					10647	SCGB1D2	secretoglobin, family 1D, member 2	0.0033	2.9
					57817	HAMP	hepcidin antimicrobial peptide	0.0025	2.4
					6036	RNASE2	ribonuclease, RNase A family, 2 (liver, eosinophil-derived neurotoxin)	0.0088	2.4
					3860	KRT13	keratin 13	0.0315	2.4
					4741	NEFM	neurofilament, medium polypeptide	0.0036	2.1
					1412	CRYBA2	crystallin, beta A2	0.0065	2.1
					80763	C12orf39	chromosome 12 open reading frame 39	0.0057	0.4
Additional		2			1381	CRABP1	cellular retinoic acid binding protein 1	0.0361	1.8
	1				11341	SCRG1	stimulator of chondrogenesis 1	0.0096	1.7
	1			4	163933	FAM43B	family with sequence similarity 43, member B	0.0268	1.7
				4	11081	KERA	keratocan	0.0010	1.6
				4	10562	OLFM4	olfactomedin 4	0.0013	1.6
			3		55825	PECR	peroxisomal trans-2-enoyl-CoA reductase	0.0043	0.7
	1				51085	MLXIPL	MLX interacting protein-like	0.0083	0.7
		2	3		153918	FAM164B	family with sequence similarity 164, member B	0.0052	0.6
			3		3112	HLA-DOB	major histocompatibility complex, class II, DO beta	0.0356	0.6
	1	2	3	4	56605	ERO1LB	ERO1-like beta (S. cerevisiae)	0.0014	0.6
		2			225	ABCD2	ATP-binding cassette, sub-family D (ALD), member 2	0.0054	0.6

4.4 DISCUSSION

DN is the most common diabetic complication, affecting up to 60% of all diabetic patients and is a major factor contributing to injury, poor wound healing and lower extremity amputation [155, 156]. The pathogenesis of DN includes hyperglycemia-induced oxidative stress and deranged polyol metabolism, changes in nerve microvasculature, decreased growth factor support and dysregulated lipid metabolism [2, 124]. Any one of these factors is enough to severely impair nerve function and all are likely to contribute to DN. Addressing these deficits alone or in combination has yet to result in effective DN treatment, confirming that an increased understanding of the mechanisms underlying the onset and progression of DN is of prime importance. The current study takes an important first step towards this goal by identifying a specific set of genes whose expression is predictive of human DN progression and analyzing their interactions within known cellular pathways. Identifying common elements in these complex networks will yield novel insights into disease pathogenesis, provide new therapeutic targets and identify potential DN biomarkers.

Our initial analyses of this data set classified the patient samples based on MFD and found that two large groups emerged; those with a loss of MFD $\geq$ 500 fibers/mm^2 over 52 weeks (progressors) and those whose MFD was relatively stable (MFD loss $\leq$ 100 fibers/mm^2 over 52 weeks, non-progressors) [21]. We examined sural nerve biopsies from these two patient groups to discover differences in gene expression that could account for the differences in the clinical course of human DN. We employed statistical methods to refine our data analyses [82, 122, 152] and narrowed our results from 14,885 expressed genes to 532 DEGs that were differentially expressed between the secondary nerve biopsies from the progressors and non-progressors.

Functional enrichment analyses identified two biological functions, “defense response” and “inflammatory response”, containing 12 or more up-regulated genes in patients classified as progressors. Increasing evidence implicates these same two processes in the development and progression of diabetic nephropathy [157-159]. For example, chemokines, toll-like receptors, adhesion molecules, and cytokines, which are all involved in inflammation and identified as up-regulated DEGs in progressors, are instrumental in the pathogenesis of diabetic nephropathy [160]. One specific gene from the inflammatory group, bradykinin receptor B2 (BDKRB2), is of particular interest for two reasons. First, BDKRB2 regulates the expression of genes involved in progressive glomerulosclerosis such as tumor growth factor beta 1 (TGF-β1) and p53 [161] and, second, we recently reported that type 1 diabetic mice with dysregulated BDKRB2 developed enhanced nephropathy and neuropathy [162]. Another specific gene of interest, membrane-associated adenosine A3 receptor (ADORA3), comes from the family of “defense” genes. It is also implicated in the pathogenesis of diabetic nephropathy as ADORA3 along with other adenosine receptors exhibit differential gene expression and cellular and tissue distribution in diabetic kidney [163]. Thus, the up-regulation of cytokines, chemokines and genes such as DBKRB2 and ADORA3 in our study (Table 4-4) suggests enhanced inflammation and dysregulated defense responses in the sural nerves of patients with progressive DN. Further exploration of these pathways in experimental models of DN could yield sufficiently new insight into the human disease process to allow for the development of mechanism-based therapies.

The down-regulated DEGs in the progressors were enriched with biological functions related to energy metabolism including “glucose metabolic process” and “PPAR signaling pathway”. Among these DEGs, PPARG, encoding a nuclear receptor for glitazone, plays a key role in regulating glucose and lipid metabolism [164-166]. Agonists of PPARG are effective in ameliorating DN and nephropathy in animal models [122, 167, 168]. Another key gene is APOE, encoding an apolipoprotein class, which regulates the normal catabolism of triglycerides and cholesterol [169, 170]. A polymorphism of this gene is linked to the progression of diabetic nephropathy [171, 172]. Decreased levels of PPARG and APOE as well as other lipid metabolism related DEGs correlates well with the increased levels of lipids in these patients and our recent finding that altered lipid metabolism may play a role in the progression of DN [21].

Although the functional enrichment analyses identify over-represented biological functions in general, they do not reveal how the DEGs are potentially interacting with each other in a network. To obtain a global view of the network, we examined gene interaction networks based on literature derived co-citation information (Figure 4-3 and Figure 4-4). Although co-citation of two genes in a single sentence does not necessarily indicate there is a direct interaction, this process may reveal novel associations and lend new insights into function [65, 173, 174]. In the current study, the BiblioSphere co-citation network demonstrated potential interactions among DEGs and identified 5 major sub-networks centered on the following genes; PPARG, APOE, SERPINE1, JUN and LEP. Figure 4-3 is a snapshot of a BiblioSphere Pathway View of this network with PPARG used as the seed node; however, this view does not show all of the identified co-citations.

The majority of the key genes identified in our network analyses are implicated in the pathogenesis of diabetes and diabetic complications (mainly diabetic nephropathy) as summarized in Table 4-9. PPARG and APOE, of particular interest in diabetic nephropathy, regulate fatty acid storage and glucose metabolism [165], and are down-regulated in progressors. Systemically, down-regulation of either gene in adipocytes leads to a decrease in serum lipid uptake with subsequent hyperlipidemia [166] and a predisposition towards developing DN [21]. Thiazolidinediones, powerful PPARG agonists, have proven effective in ameliorating DN in experimental models of diabetes [122, 167, 168]. These effects may result from direct action on neurons [175] or Schwann cells [122, 167] or by indirectly inhibiting macrophage infiltration [167] and local cytokine regulation [168]. SERPINE1, encoding plasminogen activator inhibitor 1 (PAI-1), regulates fibrinolysis [176]. Elevated levels of plasma PAI-1 are associated with higher incidences of diabetes [177-180] and knocking out PAI-1 ameliorated diabetic nephropathy in mice [181, 182]. Mutations in the appetite controlling LEP gene (ob/ob) and its receptor LEPR (db/db) have been extensively studied as animal models of type 2 diabetes with hyperglycemia and hyperlipidemia [121, 183]. A recent study suggested leptin's therapeutic effect in a combinatorial treatment with insulin in type 1 mice [184]. JUN oncogene, forming the AP-1 transcription factor, is involved in cellular processes of cell cycle control and death [185]. JUN's associated protein family, c-Jun N-terminal kinases (JNK), key signaling molecules linking inflammation and insulin resistance, are significantly activated in multiple tissues including sural nerve of type 1 and 2 diabetic patients [186-188]. Thus, the enriched biological functions and the networks of the

DEGs reflect current theories with regard to metabolic dysregulation in diabetes and its complications [189].

To fully incorporate all of the co-citation connections among the DEGs, we applied the Fast-Greedy algorithm, a community structure identification algorithm, to the entire co-citation network. Fast-Greedy, based on a hierarchical agglomeration algorithm, outperformed other methods in detecting community structure or sub-networks [190-192], the gathering of nodes into groups such that nodes are more densely connected within groups than between groups [154]. Analysis by Fast-Greedy grouped LEP and PPARG together within the context of glucose and lipid metabolism and JUN and SERPINE1 within the context of cell death and inflammation. Three other sub-networks were identified with noteworthy key genes: "cell projection and axonogenesis" with nerve growth factor receptor (NGFR), "cellular homeostasis and inflammatory response" with thioredoxin (TXN) and "cytoskeletal protein binding" with stathmin 1 (STMN1). NGFR exerts protection against nerve damage [193, 194] and the expression of NGFR protein in plasma correlates with DN progression in diabetic rats [195]. Thioredoxin, which regulates cellular oxidative stress, is also implicated in diabetes. Thioredoxin's anti-oxidant activity was significantly inhibited by hyperglycemia, which suggests its important role in vascular oxidative stress and inflammation in diabetes [196]. No direct implication of STMN1, a major regulator of microtubule dynamics, in diabetes is currently known; however, microtubule-stabilizing agents (MSTA) including taxanes and epothilones induce severe peripheral neuropathy in over 30% of MSTA-treated cancer patients [197]. This finding may suggest STMN1's possible involvement in neuropathy pathogenesis, and is worth further investigation.

Our next goal was to use observed DEG expression to predict and/or classify the separate subset of biopsies. Ridge regression models based on subsets of DEGs were evaluated for their accuracy in identifying samples as progressors and non-progressors. Regression modeling is extremely useful in predicting the progression of cancer and diabetic nephropathy [100-102]. In the current study, gene expression profiles from secondary biopsy samples of known MFD (progressors or non-progressors) were compared and used in training the models. The models were then used to classify the expression profiles of a set of primary biopsies for the progression endpoint 12 months later. The best predictive models included 14 genes and correctly classified 11 out of 12 test samples.

Naïve Bayes classification algorithm based on physiologic and demographic data of these patients demonstrated a classification accuracy of 63% in our previous study [21]. The most influential factors in this model were triglycerides, cholesterol, and a clinical symptom score. The present study demonstrates that gene expression profiles achieve much higher prediction accuracy (92%) than the clinical parameters and are better predictors of future DN progression. As shown in Table 1, the study population included both type 1 and 2 diabetes patients with a wide range of clinical parameters. Nevertheless, our expression-based predictive models correctly predicted DN progression regardless of diabetes type and other clinical characteristics.

We hypothesize that the genes identified in our best predictive models (Table 4-8) represent products or "genetic biomarkers" of the biological networks involved in DN onset and progression. This idea is reinforced by the fact that several of the genes have known associations with diabetes or diabetic complications. We are particularly interested in CST1, whose expression was increased by 10 fold in progressors. CST1, encoding a cysteine protease inhibitor, was initially implicated in gastric and colorectal tumorigenesis [198, 199]. Another member of this protein family, cystatin C (CST3), has been identified as a prime predictor of diabetic nephropathy progression [200, 201]. Although the CST1 gene product has not been investigated in the context of diabetic complications, it is detectable in saliva, tears and urine [199]. To date, there are no definitive biomarkers of DN progression easily accessed from body fluids and we speculate that CST1 could prove to be a novel and easily measureable biomarker for DN. Other identified genes with measurable endproducts and a previous association with diabetes and its complications include CRYBA2, encoding beta-crystallin A2. Beta-crystallin A2 is enhanced in both the lens and serum of STZ-induced diabetic rats during the course of cataractogenesis [202], but has not been measured in either experimental animals or humans with DN. Neurofilament medium peptide (NEFM) along with neurofilament heavy peptides (NEFH) are decreased in experimental DN and significantly correlate with lowering of both motor and sensory nerve conduction velocities in DN [203]. Serum antibodies to NEFM have been reported in an experimental model of toxic neuropathy, but have not been studied in DN [204]. HLA-DOB, encoding the histocompatibility complex class II DO beta, is identified as a type 1 diabetes susceptibility gene [205], but there are no reports how this HLA antigen is associated with DN.

As we hypothesized, other genes in the predictive models are not directly known to be involved in diabetes or its complications but are downstream markers of the biological networks we identified as being activated in DN progression. Of these, the most noteworthy are the gene products associated with inflammation or oxidative stress that may be measured in serum. These include CCL26 encoding eotaxin-3, an immune-regulatory cytokine [206]; HAMP encoding hepcidin antimicrobial peptide, whose expression is increased by inflammation [207]; and olfactomedin 4 (OLFM4), a robust marker for stem cells, is regulated by NF-kappaB, which modulates inflammation and oxidative stress [208]. Two genes are associated with another identified biological network, lipid metabolism. Peroxisomal trans-2-enoyl-CoA reductase (PECR) is involved in fatty acid metabolic processes [209] and ABCD2, encoding ATP-binding cassette subfamily D (ALD) member 2, regulates the peroxisomal import of fatty acids and fatty acyl-CoAs [210]. Like the markers of inflammation, these products may be measured in serum and could prove to be novel biomarkers.

In summary, we report for the first time differential gene expression of human sural nerves from patients with progressive and non-progressive DN. Biological enrichment and network analyses identified several novel areas of biological importance, yielding new insight into disease pathogenesis and opening up new areas of potential investigation for the discovery of mechanism-based therapies. While translating gene expression to predictive biomarkers measurable in the clinic remains a challenge, we report several novel potential biomarker candidates for DN. Collectively, our results represent the first exploration of gene expression arrays from human sural nerves of patients with varying degrees of DN and provide new insight into disease pathogenesis and biomarker identification.

Table 4-9. Key Genes Identified in the Co-citation Network Analyses

PPARG. PPARG gene encodes a type 2 nuclear receptor, the glitazone receptor or nuclear receptor subfamily 1 group C member (NR1C3) [164]. PPARG plays a crucial role in regulating glucose metabolism and fatty acid storage [165], where the genes activated by PPARG stimulate uptake of serum lipid and glucose. Down-regulation of PPARG increases serum lipid with subsequence hyperlipidemia [166], a predisposition towards developing DN [21]. Agonists of PPARG have proven effective in ameliorating DN in experimental models of diabetes [122, 167, 168].

APOE. APOE gene encoding a class of apolipoprotein, a major component of the chylomicron, plays an essential role for the normal catabolism of triglyceride and cholesterol [169]. APOE was mainly studied in the context of dyslipoproteinemia [170] and cardiovascular diseases [211]; however, APOE has been implicated other biological processes such as Alzheimer's disease [212] and immunoregulation [213]. With respect to diabetes, polymorphism of APOE has also been implicated in the progression of diabetic nephropathy [171, 172].

SERPINE1. SERPINE1 gene encodes plasminogen activator inhibitor 1 (PAI-1), a member of the serine protease inhibitor superfamily. PAI1 inhibits fibrinolysis (the physiological breakdown of blood clots) by suppressing tissue plasminogen activator (tPA) and urokinase (uPA) [176]. Elevated levels of plasma PAI-1are associated with higher incidence of both type 1 and 2 diabetes [177-180]. Kidney of human with diabetic nephropathy and animal models exhibited enhanced levels of PAI-1 [214, 215]. Diabetic nephropathy was ameliorated by knockout of PAI-1 in a mouse model [181, 182].

JUN. JUN gene encodes c-Jun forming the AP-1 transcription factor in combination with c-Fos, c-Maf, and ATF [185]. The transcription factor AP-1 is involved in cellular processes of cell cycle control, proliferation, transformation and death [185]. Mutations in JUN have been implicated in various cancers including hepatocellular carcinomas [216] and intestinal cancer [217]. The associated protein family, c-Jun N-terminal kinases (JNK), considered as key signaling molecules linking inflammation and insulin resistance, are significantly activated in various tissues of type 2 diabetes [186, 187]. Increased activation and total level of JNK was reported in sural nerve of type 1 and 2 diabetic patients [188].

LEP. LEP gene encodes an appetite controlling hormone leptin, whose increased level inhibit food intake and regulate energy expenditure. Mutations in leptin (ob/ob) or in leptin receptor (db/db) mice, exhibiting hyperglycemia and hyperlipidemia, have been extensively studied as animal models of type 2 diabetes [121, 183]. Leptin has been demonstrated its therapeutic potential in multiple studies: over-expression of leptin completely prevented the development of hyperglycemia and nephropathy in a genetic model of lipoatrophic diabetes (A-ZIP/F-1 mice) [218], and combinatorial treatment with insulin demonstrated a better efficiency in glycemic and lipidemic control in type 1 mice than a treatment of insulin alone [184].

NGFR. NGFR gene encodes a receptor for the neurotrophins, growth factors stimulating survival, differentiation, or growth of neurons as neurotrophic factors. NGFR exerts protection against nerve damage in human [193, 194]. With respect to diabetes, diabetic rats exhibited the expression of NGFR protein in plasma, possibly indicating diabetic neuropathy [195]; however, protein expression of NGFR in serum was not observed in type 2 diabetes patients [219].

KRT19. KRT19 gene encodes keratin type 1 cytoskeletal 19 proteins, which are intermediate filament proteins that maintain the structural integrity of epithelial cells. Keratin 19 is biomarkers in various cancers [220, 221].

TXN. TXN gene encodes a 12-kD oxidoreductase enzyme, thioredoxin regulating oxidative stress status in cell. Hyperglycemia inhibits thioredoxin activity through thioredoxin-interacting protein thus resulting in increased vascular oxidative stress in diabetes mellitus [196].

STMN1. STMN1 gene encodes stathmin 1, a cytosolic phosphorprotein regulating microtubule dynamics [222]. Mutation in STMN1 is implicated in development of various types of cancers [223-225].

CHAPTER 5

CONCLUSION AND FUTURE STUDY

5.1 CONCLUSION

DN is the most common complication of diabetes affecting approximately 60% of all diabetic patients and leading to significant mortality, morbidity, and poor quality of life [2, 3]. More than 50% of patients with DN will develop substantial nerve damage prior to noticeable symptoms. No treatments are currently available to reverse nerve damage in DN; therefore, early identification of DN prior to the onset of symptoms indicative of nerve damage is extremely important to allow for early intervention. To date, no biomarkers are available to identify or predict the development and progression of DN in patients with diabetes. To discover potential DN biomarkers and to better understand the pathogenesis of DN, this thesis took two distinctive approaches, employing both literature mining technology and genome-wide gene expression studies in nerve tissues with DN.

In Chapter 2, we described the development of a new literature mining tool, SciMiner [42]; a web-based target identification and functional analysis tool that identifies targets (genes and proteins) using a context specific analysis of MEDLINE abstracts and full texts. SciMiner uses both regular expression patterns and dictionaries of gene symbols and names compiled from multiple sources. Ambiguous acronyms, a major challenge in the literature mining community, are resolved by a scoring scheme based on the co-occurrence of acronyms and corresponding description terms that incorporates optional user-defined filters. In a performance evaluation using BioCreAtIvE II Gene Normalization Task, SciMiner demonstrated 86.2% recall, 60.8% precision, and 71.3% F-measure and ranked 2nd, 34th, and 19th, respectively among the 54 results published by other groups [107]. In addition to comparable performance with other methods [103-105, 107], SciMiner has unique features including the capability to examine full text documents, user feedback-based improvement of target identification, and comparisons between search results among multiple queries.

In Chapter 3, we applied SciMiner to the identification of ROS-diabetes targets from the biomedical literature and evaluated the biological relevance of selected targets in the pathogenesis of DN. A total of 1,026 ROS-diabetes targets were identified from 1,154 papers indexed with diabetes and ROS by PubMed. Fifty-three targets were significantly over-represented in the ROS-diabetes literature compared to a randomly selected set of papers. These over-represented targets included well-known members of the oxidative stress response including catalase, the NADPH oxidase, and superoxide dismutase families of proteins. The expression levels of nine genes, selected from the top ranked ROS-diabetes set, were measured in the dorsal root ganglia (DRG) of diabetic and non-diabetic DBA/2J mice. Eight genes exhibited significant differential expression between diabetic and non-diabetic mice and the directions of expression change in diabetes of 6 genes paralleled enhanced oxidative stress in the DRG.

In Chapter 4, microarray analysis was performed on sural nerve biopsies from two patient groups with fast or slow DN progression to identify gene expression profiles related to DN progression. In progressors, defense response and inflammatory response-related genes were up-regulated, while lipid metabolic process and PPAR pathway-related genes were down-regulated. Analysis of literature-derived co-citation network of the DEGs revealed gene networks centered on APOE, JUN, LEP, SERPINE1 and PPARG. We also developed mRNA expression signatures that predict DN progression in humans with high accuracy. Ridge-regression based models with 14 genes achieved a prediction accuracy of 92% (correct prediction of 11 out of 12 patients).

In summary, we report for the first time differential gene expression in human sural nerves from patients with progressive and non-progressive DN. Biological enrichment and network analyses identified several novel areas of biological importance, yielding new insight into disease pathogenesis and opening up new areas of potential investigation for the discovery of mechanism-based therapies. While translating gene expression to predictive biomarkers measurable in the clinic remains a challenge, we report several novel potential biomarker candidates for DN.

Collectively, our results represent the first exploration of gene expression in human sural nerves with varying degrees of DN. Along with the compiled ROS-diabetes targets, our results provide new insight into disease pathogenesis and biomarker identification.

5.2 FUTURE STUDY

As demonstrated in Chapter 3, SOD1 was the most over-represented ROS-diabetes targets in literature and was differentially expressed in diabetic and non-diabetic DRG. Mutations of SOD1 have long been associated with the inherited form of amyotrophic lateral sclerosis (ALS) [132] and the theory of oxidative stress-based aging [133]. Early reports indicate that knockout of the SOD1 gene does not affect nervous system development [134], although recovery following injury is slow and incomplete [135, 136]. With respect to diabetes, SOD1 KO accelerates the development of diabetic nephropathy [137] and cataract formation [138]. To the best of our knowledge, however, no published study has investigated the role of SOD1 in the onset and/or progression of diabetic neuropathy. Thus, examining the SOD1 KO mouse as a model of DN would be a reasonable next step.

The computational predictive models in Chapter 4 require validation in a larger independent patient cohort. Our laboratory currently maintains a repository of over a thousand of human sural nerve biopsies. We will retrieve and examine additional biopsies to further evaluate our predictive models. Since we have a limited set of genes in our models, a middle-throughput expression measurement system such as Applied Biosystems® TaqMan Low Density Arrays (TLDA, Carlsbad, California) could be used instead of microarray analyses.

For practical use in clinical care, these predictive biomarkers must be easily measured in readily accessible body fluids such as urine and plasma or skin biopsies. To apply the predictive models to the expression profiles in these non-invasive tissues, the correlation of gene expression profiles between sural nerve and non-invasive tissues must also be examined. Using only genes with high correlation, the current models may be applied to these non-invasive tissues, which make the genes in the models excellent biomarker candidates.

We still do not entirely understand the underlying mechanisms of DN pathogenesis, so critical to the development of mechanism specific therapeutic strategies. The level of complexity of the potentially dysregulated biological pathways in DN is too high to be easily understood without intelligent integration of multiple levels of biological information. Transcriptomics only represents one aspect of gene regulation, and changes in the transcriptome do not always correlate with protein expression or activity [226, 227]. Therefore, integrating gene expression data (transcriptome) with protein (proteome) and metabolites (metabolome) will provide a more complete picture of DN pathogenesis.

In order to identify essential cellular responses in humans that lead to DN, a careful comparison with other complications, for example, diabetic nephropathy would be necessary. Comparison of transcriptional networks between DN and diabetic nephropathy will identify cellular responses shared by these complications. From this cross-tissue comparison, those responses amenable to conventional and novel therapies will be identified and validated in murine models of diabetic complications. Following verification in animal models, these responses would be evaluated as potential biomarkers in non-invasive plasma and urine samples. These future studies have the power to discover relevant regulatory networks in DN and diabetic nephropathy and to identify candidate pathways and molecules whose regulation alters disease progression.

BIBLIOGRAPHY

1. American-Diabetes-Association: **Economic costs of diabetes in the U.S. In 2007**. *Diabetes Care* 2008, **31**(3):596-615.

2. Edwards JL, Vincent AM, Cheng HT, Feldman EL: **Diabetic neuropathy: mechanisms to management**. *Pharmacol Ther* 2008, **120**(1):1-34.

3. Feldman EL: **Diabetic neuropathy**. *Curr Drug Targets* 2008, **9**(1):1-2.

4. Otiniano ME, Du X, Ottenbacher K, Black SA, Markides KS: **Lower extremity amputations in diabetic Mexican American elders: incidence, prevalence and correlates**. *J Diabetes Complications* 2003, **17**(2):59-65.

5. Feldman EL, Stevens MJ, Russell JW, Greene DA, Becker KL: **Diabetic neuropathy**. In: *Principles and Practice of Endocrinology and Metabolism.* vol. 3rd. Philadelphia: Lippincott Williams & Wilkins; 2001: 1391-1399.

6. Bril V, Perkins BA: **Validation of the Toronto Clinical Scoring System for diabetic polyneuropathy**. *Diabetes Care* 2002, **25**(11):2048-2052.

7. Perkins BA, Greene DA, Bril V: **Glycemic control is related to the morphological severity of diabetic sensorimotor polyneuropathy**. *Diabetes Care* 2001, **24**(4):748-752.

8. Lauria G, Lombardi R: **Skin biopsy: a new tool for diagnosing peripheral neuropathy**. *BMJ* 2007, **334**(7604):1159-1162.

9. Polydefkis M, Hauer P, Sheth S, Sirdofsky M, Griffin JW, McArthur JC: **The time course of epidermal nerve fibre regeneration: studies in normal controls and in people with diabetes, with and without neuropathy**. *Brain* 2004, **127**(Pt 7):1606-1615.

10. Habib A, Brannagan T: **Therapeutic Strategies for Diabetic Neuropathy**. *Current Neurology and Neuroscience Reports* 2010, **10**(2):92-100.

11. Tesfaye S, Selvarajah D: **The Eurodiab study: what has this taught us about diabetic peripheral neuropathy?** *Curr Diab Rep* 2009, **9**(6):432-434.

12. Folli F, Guzzi V, Perego L, Coletta DK, Finzi G, Placidi C, La Rosa S, Capella C, Socci C, Lauro D *et al*: **Proteomics reveals novel oxidative and glycolytic mechanisms in type 1 diabetic patients' skin which are normalized by kidney-pancreas transplantation**. *PLoS One* 2010, **5**(3):e9923.

13. Brownlee M: **The pathobiology of diabetic complications: a unifying mechanism**. *Diabetes* 2005, **54**(6):1615-1625.

14. Chowdhury SK, Zherebitskaya E, Smith DR, Akude E, Chattopadhyay S, Jolivalt CG, Calcutt NA, Fernyhough P: **Mitochondrial respiratory chain dysfunction in dorsal root ganglia of streptozotocin-induced diabetic rats and its correction by insulin treatment**. *Diabetes* 2010, **59**(4):1082-1091.

15. Knott AB, Bossy-Wetzel E: **Impairing the mitochondrial fission and fusion balance: a new mechanism of neurodegeneration**. *Ann N Y Acad Sci* 2008, **1147**:283-292.

16. Schaper NC, Huijberts M, Pickwell K: **Neurovascular control and neurogenic inflammation in diabetes**. *Diabetes Metab Res Rev* 2008, **24 Suppl 1**:S40-44.

17. Goldberg RB: **Cytokine and cytokine-like inflammation markers, endothelial dysfunction, and imbalanced coagulation in development of diabetes and its complications**. *J Clin Endocrinol Metab* 2009, **94**(9):3171-3182.

18. Barbosa JH, Oliveira SL, Seara LT: **[The role of advanced glycation end-products (AGEs) in the development of vascular diabetic complications]**. *Arq Bras Endocrinol Metabol* 2008, **52**(6):940-950.

19. Tesfaye S: **Advances in the management of painful diabetic neuropathy**. *Clin Med* 2007, **7**(2):113-114.

20. Tesfaye S, Chaturvedi N, Eaton SE, Ward JD, Manes C, Ionescu-Tirgoviste C, Witte DR, Fuller JH: **Vascular risk factors and diabetic neuropathy**. *N Engl J Med* 2005, **352**(4):341-350.

21. Wiggin TD, Sullivan KA, Pop-Busui R, Amato A, Sima AA, Feldman EL: **Elevated triglycerides correlate with progression of diabetic neuropathy**. *Diabetes* 2009, **58**(7):1634-1640.

22. Vincent AM, Hayes JM, McLean LL, Vivekanandan-Giri A, Pennathur S, Feldman EL: **Dyslipidemia-induced neuropathy in mice: the role of oxLDL/LOX-1**. *Diabetes* 2009, **58**(10):2376-2385.

23. Davis TM, Yeap BB, Davis WA, Bruce DG: **Lipid-lowering therapy and peripheral sensory neuropathy in type 2 diabetes: the Fremantle Diabetes Study**. *Diabetologia* 2008, **51**(4):562-566.

24. Firth J: **Fenofibrate and diabetic retinopathy**. *Lancet* 2008, **371**(9614):722; author reply 722.

25. Sacks FM: **After the Fenofibrate Intervention and Event Lowering in Diabetes (FIELD) study: implications for fenofibrate**. *Am J Cardiol* 2008, **102**(12A):34L-40L.

26. Vincent AM, Kato K, McLean LL, Soules ME, Feldman EL: **Sensory neurons and schwann cells respond to oxidative stress by increasing antioxidant defense mechanisms**. *Antioxid Redox Signal* 2009, **11**(3):425-438.

27. Vincent AM, Feldman EL: **New insights into the mechanisms of diabetic neuropathy**. *Rev Endocr Metab Disord* 2004, **5**(3):227-236.

28. Vincent AM, Russell JW, Low P, Feldman EL: **Oxidative stress in the pathogenesis of diabetic neuropathy**. *Endocr Rev* 2004, **25**(4):612-628.

29. Brownlee M: **Biochemistry and molecular cell biology of diabetic complications**. *Nature* 2001, **414**(6865):813-820.

30. Feldman EL: **Oxidative stress and diabetic neuropathy: a new understanding of an old problem**. *J Clin Invest* 2003, **111**(4):431-433.

31. Ozer JS: **A guidance for renal biomarker lead optimization and use in translational pharmacodynamics**. *Drug Discov Today* 2010, **15**(3-4):142-147.

32. **PubMed** [http://www.ncbi.nlm.nih.gov/pubmed/]

33. Jensen LJ, Saric J, Bork P: **Literature mining for the biologist: from information retrieval to biological discovery**. *Nat Rev Genet* 2006, **7**(2):119-129.

34. Krallinger M, Erhardt RA, Valencia A: **Text-mining approaches in molecular biology and biomedicine**. *Drug Discov Today* 2005, **10**(6):439-445.

35. Baxevanis AD: **Searching NCBI databases using Entrez**. *Curr Protoc Bioinformatics* 2008, **Chapter 1**:Unit 1 3.

36. Wilbur WJ, Coffee L: **The effectiveness of document neighboring in search enhancement**. *Inf Process Manage* 1994, **30**(2):253-266.

37. **Medical Subject Headings - Fact Sheet** [http://www.nlm.nih.gov/pubs/factsheets/mesh.html]

38. Cohen AM, Hersh WR: **A survey of current work in biomedical text mining**. *Brief Bioinform* 2005, **6**(1):57-71.

39. Fukuda K, Tamura A, Tsunoda T, Takagi T: **Toward information extraction: identifying protein names from biological papers**. *Pac Symp Biocomput* 1998:707-718.

40. Tanabe L, Wilbur WJ: **Tagging gene and protein names in biomedical text**. *Bioinformatics* 2002, **18**(8):1124-1132.

41. Krauthammer M, Rzhetsky A, Morozov P, Friedman C: **Using BLAST for identifying gene and protein names in journal articles**. *Gene* 2000, **259**(1-2):245-252.

42. Hur J, Schuyler AD, States DJ, Feldman EL: **SciMiner: web-based literature mining tool for target identification and functional enrichment analysis**. *Bioinformatics* 2009, **25**(6):838-840.

43. Nobata C, Collier N, Tsujii J: **Automatic term identification and classification in biology texts**. In: *Proc Natural Language Pacific Rim Symposium.* 1999.

44. Zhou G, Zhang J, Su J, Shen D, Tan C: **Recognizing names in biomedical texts: a machine learning approach**. *Bioinformatics* 2004(20):1190.

45. Kazama Ji, Makino T, Ohta Y, Tsujii Ji: **Tuning support vector machines for biomedical named entity recognition**. In: *Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain - Volume 3.* Phildadelphia, Pennsylvania: Association for Computational Linguistics; 2002: 1-8.

46. Donaldson I, Martin J, de Bruijn B, Wolting C, Lay V, Tuekam B, Zhang S, Baskin B, Bader GD, Michalickova K *et al*: **PreBIND and Textomy--mining the biomedical literature for protein-protein interactions using a support vector machine**. *BMC Bioinformatics* 2003, **4**:11.

47. Hatzivassiloglou V, Duboue PA, Rzhetsky A: **Disambiguating proteins, genes, and RNA in text: a machine learning approach**. *Bioinformatics* 2001, **17 Suppl 1**:S97-106.

48. Povey S, Lovering R, Bruford E, Wright M, Lush M, Wain H: **The HUGO Gene Nomenclature Committee (HGNC)**. *Hum Genet* 2001, **109**(6):678-680.

49. **NCBI Entrez Gene Database** [http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene]

50. Hanisch D, Fundel K, Mevissen HT, Zimmer R, Fluck J: **ProMiner: rule-based protein and gene entity recognition**. *BMC Bioinformatics* 2005, **6 Suppl 1**:S14.

51. Chang JT, Schutze H, Altman RB: **GAPSCORE: finding gene and protein names one word at a time**. *Bioinformatics* 2004, **20**(2):216-225.

52. Ramani AK, Bunescu RC, Mooney RJ, Marcotte EM: **Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome**. *Genome Biol* 2005, **6**(5):R40.

53. Cooper JW, Kershenbaum A: **Discovery of protein-protein interactions using a combination of linguistic, statistical and graphical information**. *BMC Bioinformatics* 2005, **6**:143.

54. Li Y, Hu X, Lin H, Yang Z: **Learning an enriched representation from unlabeled data for protein-protein interaction extraction**. *BMC Bioinformatics* 2010, **11 Suppl 2**:S7.

55. Tarcea VG, Weymouth T, Ade A, Bookvich A, Gao J, Mahavisno V, Wright Z, Chapman A, Jayapandian M, Ozgur A *et al*: **Michigan molecular interactions r2: from interacting proteins to pathways**. *Nucleic Acids Res* 2009, **37**(Database issue):D642-646.

56. **Entrez Genome** [http://www.ncbi.nlm.nih.gov/sites/genome]

57. Abascal F, Valencia A: **Automatic annotation of protein function based on family identification**. *Proteins* 2003, **53**(3):683-692.

58. Tamames J, Ouzounis C, Casari G, Sander C, Valencia A: **EUCLID: automatic classification of proteins in functional classes by their database annotations**. *Bioinformatics* 1998, **14**(6):542-543.

59. Ashurst JL, Collins JE: **Gene annotation: prediction and testing**. *Annu Rev Genomics Hum Genet* 2003, **4**:69-88.

60. Andrade MA, Valencia A: **Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families**. *Bioinformatics* 1998, **14**(7):600-607.

61. Lovering RC, Dimmer EC, Talmud PJ: **Improvements to cardiovascular gene ontology**. *Atherosclerosis* 2009, **205**(1):9-14.

62. Arnaud MB, Costanzo MC, Shah P, Skrzypek MS, Sherlock G: **Gene Ontology and the annotation of pathogen genomes: the case of Candida albicans**. *Trends Microbiol* 2009, **17**(7):295-303.

63. Blaschke C, Oliveros JC, Valencia A: **Mining functional information associated with expression arrays**. *Funct Integr Genomics* 2001, **1**(4):256-268.

64. Shatkay H, Edwards S, Wilbur WJ, Boguski M: **Genes, themes and microarrays: using information retrieval for large-scale gene analysis**. *Proc Int Conf Intell Syst Mol Biol* 2000, **8**:317-328.

65. Jenssen TK, Laegreid A, Komorowski J, Hovig E: **A literature network of human genes for high-throughput analysis of gene expression**. *Nat Genet* 2001, **28**(1):21-28.

66. Vock C, Gleissner M, Klapper M, Doring F: **Oleate regulates genes controlled by signaling pathways of mitogen-activated protein kinase, insulin, and hypoxia**. *Nutr Res* 2008, **28**(10):681-689.

67. Lim CK, Hwang WY, Aw SE, Sun L: **Study of gene expression profile during cord blood-associated megakaryopoiesis**. *Eur J Haematol* 2008, **81**(3):196-208.

68. Raman K: **Construction and analysis of protein-protein interaction networks**. *Autom Exp* 2010, **2**(1):2.

69. Hart GT, Ramani AK, Marcotte EM: **How complete are current yeast and human protein-interaction networks?** *Genome Biol* 2006, **7**(11):120.

70. Stumpf MP, Thorne T, de Silva E, Stewart R, An HJ, Lappe M, Wiuf C: **Estimating the size of the human interactome**. *Proc Natl Acad Sci U S A* 2008, **105**(19):6959-6964.

71. Rindflesch TC, Hunter L, Aronson AR: **Mining molecular binding terminology from biomedical text**. *Proc AMIA Symp* 1999:127-131.

72. Ng SK, Wong M: **Toward Routine Automatic Pathway Discovery from On-line Scientific Text Abstracts**. *Genome Inform Ser Workshop Genome Inform* 1999, **10**:104-112.

73. Ono T, Hishigaki H, Tanigami A, Takagi T: **Automated extraction of information on protein-protein interactions from the biological literature**. *Bioinformatics* 2001, **17**(2):155-161.

74. Yang Z, Lin H, Li Y: **BioPPISVMExtractor: a protein-protein interaction extractor for biomedical literature using SVM and rich feature sets**. *J Biomed Inform* 2010, **43**(1):88-96.

75. Li Z, Chan C: **Systems biology for identifying liver toxicity pathways**. *BMC Proc* 2009, **3 Suppl 2**:S2.

76. Werner T: **Regulatory networks: linking microarray data to systems biology**. *Mech Ageing Dev* 2007, **128**(1):168-172.

77. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray**. *Science* 1995, **270**(5235):467-470.

78. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H *et al*: **Expression monitoring by hybridization to high-density oligonucleotide arrays**. *Nat Biotechnol* 1996, **14**(13):1675-1680.

79. Gomase VS, Tagore S: **Transcriptomics**. *Curr Drug Metab* 2008, **9**(3):245-249.

80. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP: **GenePattern 2.0**. *Nat Genet* 2006, **38**(5):500-501.

81. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data**. *Biostatistics* 2003, **4**(2):249-264.

82. Cohen CD, Lindenmeyer MT, Eichinger F, Hahn A, Seifert M, Moll AG, Schmid H, Kiss E, Grone E, Grone HJ *et al*: **Improved elucidation of biological processes linked to diabetic nephropathy by single probe-based microarray data analysis**. *PLoS One* 2008, **3**(8):e2937.

83. Forrest AR, Taylor DF, Crowe ML, Chalk AM, Waddell NJ, Kolle G, Faulkner GJ, Kodzius R, Katayama S, Wells C *et al*: **Genome-wide review of transcriptional complexity in mouse protein kinases and phosphatases**. *Genome Biol* 2006, **7**(1):R5.

84. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W *et al*: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**(6822):860-921.

85. Modrek B, Resch A, Grasso C, Lee C: **Genome-wide detection of alternative splicing in expressed sequences of human genes**. *Nucleic Acids Res* 2001, **29**(13):2850-2859.

86. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES *et al*: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles**. *Proc Natl Acad Sci U S A* 2005, **102**(43):15545-15550.

87. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E *et al*: **PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes**. *Nat Genet* 2003, **34**(3):267-273.

88. Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA: **DAVID: Database for Annotation, Visualization, and Integrated Discovery**. *Genome Biol* 2003, **4**(5):P3.

89. Huang da W, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources**. *Nat Protoc* 2009, **4**(1):44-57.

90. Sartor MA, Mahavisno V, Keshamouni VG, Cavalcoli J, Wright Z, Karnovsky A, Kuick R, Jagadish HV, Mirel B, Weymouth T *et al*: **ConceptGen: a gene set enrichment and gene set relation mapping tool**. *Bioinformatics* 2010, **26**(4):456-463.

91. Sartor MA, Leikauf GD, Medvedovic M: **LRpath: a logistic regression approach for identifying enriched biological groups in gene expression data**. *Bioinformatics* 2009, **25**(2):211-217.

92. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet* 2000, **25**(1):25-29.

93. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes**. *Nucleic Acids Res* 2000, **28**(1):27-30.

94. Vastrik I, D'Eustachio P, Schmidt E, Joshi-Tope G, Gopinath G, Croft D, de Bono B, Gillespie M, Jassal B, Lewis S *et al*: **Reactome: a knowledge base of biologic pathways and processes**. *Genome Biol* 2007, **8**(3):R39.

95. Hanai T, Hamada H, Okamoto M: **Application of bioinformatics for DNA microarray data to bioscience, bioengineering and medical fields**. *J Biosci Bioeng* 2006, **101**(5):377-384.

96. Catalano A, Iland H: **Molecular biology of lymphoma in the microarray era**. *Pathology* 2005, **37**(6):508-522.

97. Quackenbush J: **Microarray analysis and tumor classification**. *N Engl J Med* 2006, **354**(23):2463-2472.

98. Ressom HW, Varghese RS, Zhang Z, Xuan J, Clarke R: **Classification algorithms for phenotype prediction in genomics and proteomics**. *Front Biosci* 2008, **13**:691-708.

99. Dong G, Li J, Wong L: **The Use of Emerging Patterns in the Analysis of Gene Expression Profiles for the Diagnosis and Understanding of Diseases**: IEE Press/Wiley; 2005.

100. Yoshihara K, Tajima A, Yahata T, Kodama S, Fujiwara H, Suzuki M, Onishi Y, Hatae M, Sueyoshi K, Kudo Y *et al*: **Gene expression profile for predicting survival in advanced-stage serous ovarian cancer across two independent datasets**. *PLoS One* 2010, **5**(3):e9615.

101. Shedden K, Taylor JM, Enkemann SA, Tsao MS, Yeatman TJ, Gerald WL, Eschrich S, Jurisica I, Giordano TJ, Misek DE *et al*: **Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study**. *Nat Med* 2008, **14**(8):822-827.

102. Ju W, Eichinger F, Bitzer M, Oh J, McWeeney S, Berthier CC, Shedden K, Cohen CD, Henger A, Krick S *et al*: **Renal gene and protein expression signatures for prediction of kidney disease progression**. *Am J Pathol* 2009, **174**(6):2073-2085.

103. Plake C, Schiemann T, Pankalla M, Hakenberg J, Leser U: **AliBaba: PubMed as a graph**. *Bioinformatics* 2006, **22**(19):2444-2445.

104. Rebholz-Schuhmann D, Kirsch H, Arregui M, Gaudan S, Riethoven M, Stoehr P: **EBIMed--text crunching to gather facts for proteins from Medline**. *Bioinformatics* 2007, **23**(2):e237-244.

105. Cheng D, Knox C, Young N, Stothard P, Damaraju S, Wishart DS: **PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites**. *Nucleic Acids Res* 2008, **36**(Web Server issue):W399-405.

106. Sima AA, Calvani M, Mehra M, Amato A: **Acetyl-L-carnitine improves pain, nerve regeneration, and vibratory perception in patients with chronic diabetic neuropathy: an analysis of two randomized placebo-controlled trials**. *Diabetes Care* 2005, **28**(1):89-94.

107. Morgan A, Lu Z, Wang X, Cohen A, Fluck J, Ruch P, Divoli A, Fundel K, Leaman R, Hakenberg J *et al*: **Overview of BioCreative II gene normalization**. *Genome Biology* 2008, **9**(Suppl 2):S3.

108. Bruford EA, Lush MJ, Wright MW, Sneddon TP, Povey S, Birney E: **The HGNC Database in 2008: a resource for the human genome**. *Nucleic Acids Res* 2008, **36**(Database issue):D445-448.

109. Alibes A, Yankilevich P, Canada A, Diaz-Uriarte R: **IDconverter and IDClight: conversion and annotation of gene and protein IDs**. *BMC Bioinformatics* 2007, **8**:9.

110. Fisher RA: **On the interpretation of χ2 from contingency tables, and the calculation of P**. *Journal of the Royal Statistical Society* 1922, **85 (1)**:87-94.

111. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks**. *Genome Res* 2003, **13**(11):2498-2504.

112. Gao J, Ade AS, Tarcea VG, Weymouth TE, Mirel BR, Jagadish HV, States DJ: **Integrating and Annotating the Interactome using the MiMI plugin for Cytoscape**. *Bioinformatics* 2008:btn501.

113. Erol A: **Insulin resistance is an evolutionarily conserved physiological mechanism at the cellular level for protection against increased oxidative stress**. *Bioessays* 2007, **29**(8):811-818.

114. Pan JS, Hong MZ, Ren JL: **Reactive oxygen species: a double-edged sword in oncogenesis**. *World J Gastroenterol* 2009, **15**(14):1702-1707.

115. Sarsour EH, Kumar MG, Chaudhuri L, Kalen AL, Goswami PC: **Redox control of the cell cycle in health and disease**. *Antioxid Redox Signal* 2009, **11**(12):2985-3011.

116. Schwedhelm E, Maas R, Troost R, Boger RH: **Clinical pharmacokinetics of antioxidants and their impact on systemic oxidative stress**. *Clin Pharmacokinet* 2003, **42**(5):437-459.

117. Shaw PJ: **Molecular and cellular pathways of neurodegeneration in motor neurone disease**. *J Neurol Neurosurg Psychiatry* 2005, **76**(8):1046-1057.

118. Russell JW, Berent-Spillson A, Vincent AM, Freimann CL, Sullivan KA, Feldman EL: **Oxidative injury and neuropathy in diabetes and impaired glucose tolerance**. *Neurobiol Dis* 2008, **30**(3):420-429.

119. Jimenez-Andrade JM, Herrera MB, Ghilardi JR, Vardanyan M, Melemedjian OK, Mantyh PW: **Vascularization of the dorsal root ganglia and peripheral nerve of the mouse: implications for chemical-induced peripheral sensory neuropathies**. *Mol Pain* 2008, **4**:10.

120. Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing**. *Journal of the Royal Statistical Society Series B (Methodological)* 1995, **57**(1):289-300.

121. Sullivan KA, Hayes JM, Wiggin TD, Backus C, Su Oh S, Lentz SI, Brosius F, 3rd, Feldman EL: **Mouse models of diabetic neuropathy**. *Neurobiol Dis* 2007, **28**(3):276-285.

122. Wiggin TD, Kretzler M, Pennathur S, Sullivan KA, Brosius FC, Feldman EL: **Rosiglitazone treatment reduces diabetic neuropathy in streptozotocin-treated DBA/2J mice**. *Endocrinology* 2008, **149**(10):4928-4937.

123. **The ROS-diabetes webpage** [http://jdrf.neurology.med.umich.edu/ROSDiabetes/]

124. Figueroa-Romero C, Sadidi M, Feldman EL: **Mechanisms of disease: the oxidative stress theory of diabetic neuropathy**. *Rev Endocr Metab Disord* 2008, **9**(4):301-314.

125. Hakim FA, Pflueger A: **Role of oxidative stress in diabetic kidney disease**. *Med Sci Monit* 2010, **16**(2):RA37-48.

126. Ilieva H, Polymenidou M, Cleveland DW: **Non-cell autonomous toxicity in neurodegenerative disorders: ALS and beyond**. *J Cell Biol* 2009, **187**(6):761-772.

127. Lee HB, Yu MR, Yang Y, Jiang Z, Ha H: **Reactive oxygen species-regulated signaling pathways in diabetic nephropathy**. *J Am Soc Nephrol* 2003, **14**(8 Suppl 3):S241-245.

128. Cetinkalp S, Delen Y, Karadeniz M, Yuce G, Yilmaz C: **The effect of 1alpha,25(OH)2D3 vitamin over oxidative stress and biochemical parameters in rats where Type 1 diabetes is formed by streptozotocin**. *J Diabetes Complications* 2009, **23**(6):401-408.

129. Aliciguzel Y, Ozen I, Aslan M, Karayalcin U: **Activities of xanthine oxidoreductase and antioxidant enzymes in different tissues of diabetic rats**. *J Lab Clin Med* 2003, **142**(3):172-177.

130. Iaccio A, Collinet C, Gesualdi NM, Ammendola R: **Protein kinase C-alpha and -delta are required for NADPH oxidase activation in WKYMVm-stimulated IMR90 human fibroblasts**. *Arch Biochem Biophys* 2007, **459**(2):288-294.

131. Fontayne A, Dang PM, Gougerot-Pocidalo MA, El-Benna J: **Phosphorylation of p47phox sites by PKC alpha, beta II, delta, and zeta: effect on binding to p22phox and on NADPH oxidase activation**. *Biochemistry* 2002, **41**(24):7743-7750.

132. Turner BJ, Talbot K: **Transgenics, toxicity and therapeutics in rodent models of mutant SOD1-mediated familial ALS**. *Prog Neurobiol* 2008, **85**(1):94-134.

133. Perez VI, Bokov A, Van Remmen H, Mele J, Ran Q, Ikeno Y, Richardson A: **Is the oxidative stress theory of aging dead?** *Biochim Biophys Acta* 2009, **1790**(10):1005-1014.

134. Ohlemiller KK, McFadden SL, Ding DL, Flood DG, Reaume AG, Hoffman EK, Scott RW, Wright JS, Putcha GV, Salvi RJ: **Targeted deletion of the cytosolic Cu/Zn-superoxide dismutase gene (Sod1) increases susceptibility to noise-induced hearing loss**. *Audiol Neurootol* 1999, **4**(5):237-246.

135. Kawase M, Murakami K, Fujimura M, Morita-Fujimura Y, Gasche Y, Kondo T, Scott RW, Chan PH: **Exacerbation of delayed cell injury after transient global ischemia in mutant mice with CuZn superoxide dismutase deficiency**. *Stroke* 1999, **30**(9):1962-1968.

136. Reaume AG, Elliott JL, Hoffman EK, Kowall NW, Ferrante RJ, Siwek DF, Wilcox HM, Flood DG, Beal MF, Brown RH, Jr. *et al*: **Motor neurons in Cu/Zn superoxide dismutase-deficient mice develop normally but exhibit enhanced cell death after axonal injury**. *Nat Genet* 1996, **13**(1):43-47.

137. DeRubertis FR, Craven PA, Melhem MF: **Acceleration of diabetic renal injury in the superoxide dismutase knockout mouse: effects of tempol**. *Metabolism* 2007, **56**(9):1256-1264.

138. Olofsson EM, Marklund SL, Behndig A: **Enhanced diabetes-induced cataract in copper-zinc superoxide dismutase-null mice**. *Invest Ophthalmol Vis Sci* 2009, **50**(6):2913-2918.

139. Hirschman L, Yeh A, Blaschke C, Valencia A: **Overview of BioCreAtIvE: critical assessment of information extraction for biology**. *BMC Bioinformatics* 2005, **6 Suppl 1**:S1.

140. Smith AG, Russell J, Feldman EL, Goldstein J, Peltier A, Smith S, Hamwi J, Pollari D, Bixby B, Howard J *et al*: **Lifestyle intervention for pre-diabetic neuropathy**. *Diabetes care* 2006, **29**(6):1294-1299.

141. Smith AG, Rose K, Singleton JR: **Idiopathic neuropathy patients are at high risk for metabolic syndrome**. *Journal of the neurological sciences* 2008, **273**(1-2):25-28.

142. Keller B, Martini S, Sedor J, Kretzler M: **Linking variants from genome-wide association analysis to function via transcriptional network analysis**. *Semin Nephrol* 2010, **30**(2):177-184.

143. Bhavnani SK, Eichinger F, Martini S, Saxman P, Jagadish HV, Kretzler M: **Network analysis of genes regulated in renal diseases: implications for a molecular-based classification**. *BMC Bioinformatics* 2009, **10 Suppl 9**:S3.

144. Schmid H, Henger A, Kretzler M: **Molecular approaches to chronic kidney disease**. *Curr Opin Nephrol Hypertens* 2006, **15**(2):123-129.

145. Yasuda Y, Cohen CD, Henger A, Kretzler M: **Gene expression profiling analysis in nephrology: towards molecular definition of renal disease**. *Clin Exp Nephrol* 2006, **10**(2):91-98.

146. Schroeder A, Mueller O, Stocker S, Salowsky R, Leiber M, Gassmann M, Lightfoot S, Menzel W, Granzow M, Ragg T: **The RIN: an RNA integrity number for assigning integrity values to RNA measurements**. *BMC Mol Biol* 2006, **7**:3.

147. Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, Bunney WE, Myers RM, Speed TP, Akil H *et al*: **Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data**. *Nucleic Acids Res* 2005, **33**(20):e175.

148. Gautier L, Cope L, Bolstad BM, Irizarry RA: **affy--analysis of Affymetrix GeneChip data at the probe level**. *Bioinformatics* 2004, **20**(3):307-315.

149. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias**. *Bioinformatics* 2003, **19**(2):185-193.

150. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP: **Summaries of Affymetrix GeneChip probe level data**. *Nucleic Acids Res* 2003, **31**(4):e15.

151. Raychaudhuri S, Stuart JM, Altman RB: **Principal components analysis to summarize microarray experiments: application to sporulation time series**. *Pac Symp Biocomput* 2000:455-466.

152. Baldi P, Long AD: **A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes**. *Bioinformatics* 2001, **17**(6):509-519.

153. Gyorffy B, Molnar B, Lage H, Szallasi Z, Eklund AC: **Evaluation of microarray preprocessing algorithms based on concordance with RT-PCR in clinical samples**. *PLoS One* 2009, **4**(5):e5645.

154. Clauset A, Newman ME, Moore C: **Finding community structure in very large networks**. *Phys Rev E Stat Nonlin Soft Matter Phys* 2004, **70**(6 Pt 2):066111.

155. Feldman EL, Stevens MJ, Russell JW, Greene DA, Porte D, Jr., Sherwin RS, Baron A: **Somatosensory neuropathy**. In: *Ellenberg and Rifkin's Diabetes Mellitus.* vol. 6th McGraw Hill; 2003: 771-788.

156. Feldman EL, Stevens MJ, Russell JW, Peltier A, Inzucchi S, Porte JD, Sherwin RS, Baron A: **Somatosensory neuropathy**. In: *The Diabetes Mellitus Manual.* vol. 6th: McGraw-Hill; 2005: 366-384.

157. Tuttle KR: **Linking metabolism and immunology: diabetic nephropathy is an inflammatory disease**. *J Am Soc Nephrol* 2005, **16**(6):1537-1538.

158. Navarro-Gonzalez JF, Mora-Fernandez C: **The role of inflammatory cytokines in diabetic nephropathy**. *J Am Soc Nephrol* 2008, **19**(3):433-442.

159. Mora C, Navarro J: **Inflammation and diabetic nephropathy**. *Current Diabetes Reports* 2006, **6**(6):463-468.

160. Rivero A, Mora C, Muros M, Garcia J, Herrera H, Navarro-Gonzalez JF: **Pathogenic perspectives for the role of inflammation in diabetic nephropathy**. *Clin Sci (Lond)* 2009, **116**(6):479-492.

161. Kakoki M, Kizer CM, Yi X, Takahashi N, Kim HS, Bagnell CR, Edgell CJ, Maeda N, Jennette JC, Smithies O: **Senescence-associated phenotypes in Akita diabetic mice are enhanced by absence of bradykinin B2 receptors**. *J Clin Invest* 2006, **116**(5):1302-1309.

162. Kakoki M, Sullivan KA, Backus C, Hayes JM, Oh SS, Hua K, Gasim AM, Tomita H, Grant R, Nossov SB *et al*: **Lack of both bradykinin B1 and B2 receptors enhances nephropathy, neuropathy, and bone mineral loss in Akita diabetic mice**. *Proc Natl Acad Sci U S A* 2010.

163. Pawelczyk T, Grden M, Rzepko R, Sakowicz M, Szutowicz A: **Region-specific alterations of adenosine receptors expression level in kidney of diabetic rat**. *Am J Pathol* 2005, **167**(2):315-325.

164. Elbrecht A, Chen Y, Cullinan CA, Hayes N, Leibowitz M, Moller DE, Berger J: **Molecular cloning, expression and characterization of human peroxisome proliferator activated receptors gamma 1 and gamma 2**. *Biochem Biophys Res Commun* 1996, **224**(2):431-437.

165. Robinson E, Grieve DJ: **Significance of peroxisome proliferator-activated receptors in the cardiovascular system in health and disease**. *Pharmacol Ther* 2009, **122**(3):246-263.

166. Duan SZ, Usher MG, Mortensen RM: **PPARs: the vasculature, inflammation and hypertension**. *Curr Opin Nephrol Hypertens* 2009, **18**(2):128-133.

167. Yamagishi S, Ogasawara S, Mizukami H, Yajima N, Wada R, Sugawara A, Yagihashi S: **Correction of protein kinase C activity and macrophage migration in peripheral nerve by pioglitazone, peroxisome proliferator activated-gamma-ligand, in insulin-deficient diabetic rats**. *J Neurochem* 2008, **104**(2):491-499.

168. Maeda T, Kiguchi N, Kobayashi Y, Ozaki M, Kishioka S: **Pioglitazone attenuates tactile allodynia and thermal hyperalgesia in mice subjected to peripheral nerve injury**. *J Pharmacol Sci* 2008, **108**(3):341-347.

169. Ordovas JM, Mooser V: **The APOE locus and the pharmacogenetics of lipid response**. *Curr Opin Lipidol* 2002, **13**(2):113-117.

170. Smit M, de Knijff P, van der Kooij-Meijs E, Groenendijk C, van den Maagdenberg AM, Gevers Leuven JA, Stalenhoef AF, Stuyt PM, Frants RR, Havekes LM: **Genetic heterogeneity in familial dysbetalipoproteinemia. The E2(lys146----gln) variant results in a dominant mode of inheritance**. *J Lipid Res* 1990, **31**(1):45-53.

171. Li Y, Tang K, Zhang Z, Zhang M, Zeng Z, He Z, He L, Wan C: **Genetic diversity of the apolipoprotein E gene and diabetic nephropathy: a meta-analysis**. *Mol Biol Rep* 2010.

172. Rosario RF, Prabhakar S: **Lipids and diabetic nephropathy**. *Curr Diab Rep* 2006, **6**(6):455-462.

173. Kimpel MW, Strother WN, McClintick JN, Carr LG, Liang T, Edenberg HJ, McBride WJ: **Functional gene expression differences between inbred alcohol-preferring and -non-preferring rats in five brain regions**. *Alcohol* 2007, **41**(2):95-132.

174. Schmelzer C, Lindner I, Rimbach G, Niklowitz P, Menke T, Doring F: **Functions of coenzyme Q10 in inflammation and gene expression**. *Biofactors* 2008, **32**(1-4):179-183.

175. Churi SB, Abdel-Aleem OS, Tumber KK, Scuderi-Porter H, Taylor BK: **Intrathecal rosiglitazone acts at peroxisome proliferator-activated receptor-gamma to rapidly inhibit neuropathic pain in rats**. *J Pain* 2008, **9**(7):639-649.

176. Binder BR, Christ G, Gruber F, Grubic N, Hufnagl P, Krebs M, Mihaly J, Prager GW: **Plasminogen activator inhibitor 1: physiological and pathophysiological roles**. *News Physiol Sci* 2002, **17**:56-61.

177. Cesari M, Sartori MT, Patrassi GM, Vettore S, Rossi GP: **Determinants of plasma levels of plasminogen activator inhibitor-1 : A study of normotensive twins**. *Arterioscler Thromb Vasc Biol* 1999, **19**(2):316-320.

178. Festa A, D'Agostino R, Jr., Tracy RP, Haffner SM: **Elevated levels of acute-phase proteins and plasminogen activator inhibitor-1 predict the development of type 2 diabetes: the insulin resistance atherosclerosis study**. *Diabetes* 2002, **51**(4):1131-1137.

179. Bosnyak Z, Forrest KY, Maser RE, Becker D, Orchard TJ: **Do plasminogen activator inhibitor (PAI-1) or tissue plasminogen activator PAI-1 complexes predict complications in Type 1 diabetes: the Pittsburgh Epidemiology of Diabetes Complications Study**. *Diabet Med* 2003, **20**(2):147-151.

180. Erem C, Hacihasanoglu A, Celik S, Ovali E, Ersoz HO, Ukinc K, Deger O, Telatar M: **Coagulation and fibrinolysis parameters in type 2 diabetic patients with and without diabetic vascular complications**. *Med Princ Pract* 2005, **14**(1):22-30.

181. Nicholas SB, Aguiniga E, Ren Y, Kim J, Wong J, Govindarajan N, Noda M, Wang W, Kawano Y, Collins A *et al*: **Plasminogen activator inhibitor-1 deficiency retards diabetic nephropathy**. *Kidney Int* 2005, **67**(4):1297-1307.

182. Brosius FC, 3rd: **New insights into the mechanisms of fibrosis and sclerosis in diabetic nephropathy**. *Rev Endocr Metab Disord* 2008, **9**(4):245-254.

183. Wolf G, Chen S, Han DC, Ziyadeh FN: **Leptin and renal disease**. *Am J Kidney Dis* 2002, **39**(1):1-11.

184. Wang MY, Chen L, Clark GO, Lee Y, Stevens RD, Ilkayeva OR, Wenner BR, Bain JR, Charron MJ, Newgard CB *et al*: **Leptin therapy in insulin-deficient type I diabetes**. *Proc Natl Acad Sci U S A* 2010, **107**(11):4813-4819.

185. Shaulian E, Karin M: **AP-1 as a regulator of cell life and death**. *Nat Cell Biol* 2002, **4**(5):E131-136.

186. Yang R, Trevillyan JM: **c-Jun N-terminal kinase pathways in diabetes**. *Int J Biochem Cell Biol* 2008, **40**(12):2702-2706.

187. Wellen KE, Hotamisligil GS: **Inflammation, stress, and diabetes**. *J Clin Invest* 2005, **115**(5):1111-1119.

188. Purves T, Middlemas A, Agthong S, Jude EB, Boulton AJ, Fernyhough P, Tomlinson DR: **A role for mitogen-activated protein kinases in the etiology of diabetic neuropathy**. *FASEB J* 2001, **15**(13):2508-2514.

189. Obrosova IG: **Diabetic painful and insensate neuropathy: pathogenesis and potential treatments**. *Neurotherapeutics* 2009, **6**(4):638-647.

190. Radicchi F, Castellano C, Cecconi F, Loreto V, Parisi D: **Defining and identifying communities in networks**. *Proc Natl Acad Sci U S A* 2004, **101**(9):2658-2663.

191. Newman ME: **Fast algorithm for detecting community structure in networks**. *Phys Rev E Stat Nonlin Soft Matter Phys* 2004, **69**(6 Pt 2):066133.

192. Wu F, Huberman BA: **Finding communities in linear time: a physics approach**. *Eur Phys J B* 2004, **38**(2):331-338.

193. Kocsis JD, Lankford KL, Sasaki M, Radtke C: **Unique in vivo properties of olfactory ensheathing cells that may contribute to neural repair and protection following spinal cord injury**. *Neurosci Lett* 2009, **456**(3):137-142.

194. Chen LW, Yung KK, Chan YS, Shum DK, Bolam JP: **The proNGF-p75NTR-sortilin signalling complex as new target for the therapeutic treatment of Parkinson's disease**. *CNS Neurol Disord Drug Targets* 2008, **7**(6):512-523.

195. Chilton L, Middlemas A, Gardiner N, Tomlinson DR: **The p75 neurotrophin receptor appears in plasma in diabetic rats-characterisation of a potential early test for neuropathy**. *Diabetologia* 2004, **47**(11):1924-1930.

196. Schulze PC, Yoshioka J, Takahashi T, He Z, King GL, Lee RT: **Hyperglycemia promotes oxidative stress through inhibition of thioredoxin function by thioredoxin-interacting protein**. *J Biol Chem* 2004, **279**(29):30369-30374.

197. Lee JJ, Swain SM: **Peripheral neuropathy induced by microtubule-stabilizing agents**. *J Clin Oncol* 2006, **24**(10):1633-1642.

198. Yoneda K, Iida H, Endo H, Hosono K, Akiyama T, Takahashi H, Inamori M, Abe Y, Yoneda M, Fujita K *et al*: **Identification of Cystatin SN as a novel tumor marker for colorectal cancer**. *Int J Oncol* 2009, **35**(1):33-40.

199. Choi EH, Kim JT, Kim JH, Kim SY, Song EY, Kim JW, Yeom YI, Kim IH, Lee HG: **Upregulation of the cysteine protease inhibitor, cystatin SN, contributes to cell proliferation and cathepsin inhibition in gastric cancer**. *Clin Chim Acta* 2009, **406**(1-2):45-51.

200. Shimizu A, Horikoshi S, Rinnno H, Kobata M, Saito K, Tomino Y: **Serum cystatin C may predict the early prognostic stages of patients with type 2 diabetic nephropathy**. *J Clin Lab Anal* 2003, **17**(5):164-167.

201. Taglieri N, Koenig W, Kaski JC: **Cystatin C and cardiovascular risk**. *Clin Chem* 2009, **55**(11):1932-1943.

202. Ranjan M, Nayak S, Rao BS: **Immunochemical detection of glycated beta- and gamma-crystallins in lens and their circulating autoantibodies (IgG) in streptozocin induced diabetic rat**. *Mol Vis* 2006, **12**:1077-1085.

203. Sayers NM, Beswick LJ, Middlemas A, Calcutt NA, Mizisin AP, Tomlinson DR, Fernyhough P: **Neurotrophin-3 prevents the proximal accumulation of neurofilament proteins in sensory neurons of streptozocin-induced diabetic rats**. *Diabetes* 2003, **52**(9):2372-2380.

204. El-Fawal HA, McCain WC: **Antibodies to neural proteins in organophosphorus-induced delayed neuropathy (OPIDN) and its amelioration**. *Neurotoxicol Teratol* 2008, **30**(3):161-166.

205. Johansson S, Lie BA, Todd JA, Pociot F, Nerup J, Cambon-Thomsen A, Kockum I, Akselsen HE, Thorsby E, Undlien DE: **Evidence of at least two type 1 diabetes susceptibility genes in the HLA complex distinct from HLA-DQB1, -DQA1 and -DRB1**. *Genes Immun* 2003, **4**(1):46-53.

206. Blanchard C, Rothenberg ME: **Basic pathogenesis of eosinophilic esophagitis**. *Gastrointest Endosc Clin N Am* 2008, **18**(1):133-143; x.

207. Chung B, Matak P, McKie AT, Sharp P: **Leptin increases the expression of the iron regulatory hormone hepcidin in HuH7 human hepatoma cells**. *J Nutr* 2007, **137**(11):2366-2370.

208. Chin KL, Aerbajinai W, Zhu J, Drew L, Chen L, Liu W, Rodgers GP: **The regulation of OLFM4 expression in myeloid precursor cells relies on NF-kappaB transcription factor**. *Br J Haematol* 2008, **143**(3):421-432.

209. Gloerich J, Ruiter JP, van den Brink DM, Ofman R, Ferdinandusse S, Wanders RJ: **Peroxisomal trans-2-enoyl-CoA reductase is involved in phytol degradation**. *FEBS Lett* 2006, **580**(8):2092-2096.

210. Fourcade S, Ruiz M, Camps C, Schluter A, Houten SM, Mooyer PA, Pampols T, Dacremont G, Wanders RJ, Giros M *et al*: **A key role for the peroxisomal ABCD2 transporter in fatty acid homeostasis**. *Am J Physiol Endocrinol Metab* 2009, **296**(1):E211-221.

211. Bocksch L, Stephens T, Lucas A, Singh B: **Apolipoprotein E: possible therapeutic target for atherosclerosis**. *Curr Drug Targets Cardiovasc Haematol Disord* 2001, **1**(2):93-106.

212. Bookheimer S, Burggren A: **APOE-4 genotype and neurophysiological vulnerability to Alzheimer's and cognitive aging**. *Annu Rev Clin Psychol* 2009, **5**:343-362.

213. Zhang HL, Wu J, Zhu J: **The role of apolipoprotein E in Guillain-Barre syndrome and experimental autoimmune neuritis**. *J Biomed Biotechnol* 2010, **2010**:357412.

214. Eddy AA, Giachelli CM: **Renal expression of genes that promote interstitial inflammation and fibrosis in rats with protein-overload proteinuria**. *Kidney Int* 1995, **47**(6):1546-1557.

215. Hirano T, Kashiwazaki K, Moritomo Y, Nagano S, Adachi M: **Albuminuria is directly associated with increased plasma PAI-1 and factor VII levels in NIDDM patients**. *Diabetes Res Clin Pract* 1997, **36**(1):11-18.

216. Eferl R, Ricci R, Kenner L, Zenz R, David JP, Rath M, Wagner EF: **Liver tumor development. c-Jun antagonizes the proapoptotic activity of p53**. *Cell* 2003, **112**(2):181-192.

217. Nateri AS, Spencer-Dene B, Behrens A: **Interaction of phosphorylated c-Jun with TCF4 regulates intestinal cancer development**. *Nature* 2005, **437**(7056):281-285.

218. Suganami T, Mukoyama M, Mori K, Yokoi H, Koshikawa M, Sawai K, Hidaka S, Ebihara K, Tanaka T, Sugawara A *et al*: **Prevention and reversal of renal injury by leptin in a new mouse model of diabetic nephropathy**. *FASEB J* 2005, **19**(1):127-129.

219. Humpert PM, Kopf S, Djuric Z, Laine K, Korosoglou G, Rudofsky G, Jr., Hamann A, Morcos M, von Eynatten M, Nawroth PP *et al*: **Levels of three distinct p75 neurotrophin receptor forms found in human plasma are altered in type 2 diabetic patients**. *Diabetologia* 2007, **50**(7):1517-1522.

220. Saloustros E, Mavroudis D: **Cytokeratin 19-positive circulating tumor cells in early breast cancer prognosis**. *Future Oncol* 2010, **6**(2):209-219.

221. Gama A, Alves A, Schmitt F: **Expression and prognostic significance of CK19 in canine malignant mammary tumours**. *Vet J* 2010, **184**(1):45-51.

222. Jourdain L, Curmi P, Sobel A, Pantaloni D, Carlier MF: **Stathmin: a tubulin-sequestering protein which forms a ternary T2S complex with two tubulin molecules**. *Biochemistry* 1997, **36**(36):10817-10821.

223. Bjorklund P, Cupisti K, Fryknas M, Isaksson A, Willenberg HS, Akerstrom G, Hellman P, Westin G: **Stathmin as a marker for malignancy in pheochromocytomas**. *Exp Clin Endocrinol Diabetes* 2010, **118**(1):27-30.

224. Jeon TY, Han ME, Lee YW, Lee YS, Kim GH, Song GA, Hur GY, Kim JY, Kim HJ, Yoon S *et al*: **Overexpression of stathmin1 in the diffuse type of gastric cancer and its roles in proliferation and migration of gastric cancer cells**. *Br J Cancer* 2010, **102**(4):710-718.

225. Kuo MF, Wang HS, Kuo QT, Shun CT, Hsu HC, Yang SH, Yuan RH: **High expression of stathmin protein predicts a fulminant course in medulloblastoma**. *J Neurosurg Pediatr* 2009, **4**(1):74-80.

226. Gygi SP, Rochon Y, Franza BR, Aebersold R: **Correlation between protein and mRNA abundance in yeast**. *Mol Cell Biol* 1999, **19**(3):1720-1730.

227. Cravatt BF, Simon GM, Yates JR, 3rd: **The biological impact of mass-spectrometry-based proteomics**. *Nature* 2007, **450**(7172):991-1000.