Interaction Variability of Human Protein Isoforms Identified through Biomedical Literature Mining

Şenay Kafkas

Submitted to the
Institute of Graduate Studies and Research
in partial fulfilment of the requirements for the Degree of

Doctor of Philosophy
in
Computer Engineering

Eastern Mediterranean University
February 2012
Gazimağusa, North Cyprus

Approval of the Institute of Graduate Studies and Research

Prof. Dr. Elvan Yılmaz
Director

I certify that this thesis satisfies the requirements as a thesis for the degree of Doctor of Philosophy in Computer Engineering.

Assoc. Prof. Dr. Muhammed Salamah
Chair, Department of Computer Engineering

We certify that we have read this thesis and that in our opinion it is fully adequate in scope and quality as a thesis for the degree of Doctor of Philosophy in Computer Engineering.

Assoc. Prof. Dr. Bahar Taneri
Co-Supervisor

Asst. Prof. Dr. Ekrem Varoğlu
Supervisor

Examining Committee

1. Assoc. Prof. Dr. Bahar Taneri
2. Assoc. Prof. Dr. Hakan Altınçay
3. Asst. Prof. Dr. Ahmet Ünveren
4. Asst. Prof. Dr. Ekrem Varoğlu
5. Asst. Prof. Dr. Özlem Uzuner

ABSTRACT

Over the last decade, advances achieved in genomic technologies have led to uncover vast amount of protein-protein interaction data. Nevertheless, the existing protein-protein interaction databases cover the interactions related only to a part of the proteome and protein isoform interaction databases are sparsely populated. Such isoforms are generated through transcript diversity mechanisms (e.g. alternative splicing) and could exhibit functional differences. Protein-protein interaction data on isoforms is necessary for analysing their functional similarities and understanding the effects of transcript diversity on protein-protein interaction networks. Biomedical literature is an invaluable complementary resource to experimental data. Automated tools are required to gather, view and analyse the isoform interactions from the biomedical literature.

This study presents a comprehensive automated text mining based analysis, which extracts protein interactions from the biomedical literature for human protein isoforms linked to the transcripts clustered in HumanSDB3 (an alternative splicing database of the human transcriptome). Extracted protein-protein interaction data is delivered to public through a new database called TBIID which stands for Transcript Based Isoform Interactions Database.

TBIID contains a total number of 31,819 interactions between 7,161 unique proteins. The interaction data is automatically gathered from a subset of 205,207 interaction abstracts, which are selected from about 4 million Medline abstracts belonging to the isoforms in HumanSDB3. The automatic extraction methods achieve state-of-the-art performance (53.22% precision, 68.94% recall, 60.07% F_1-score).

TBIID is utilised to quantify the variability in the isoform interactions based on their shared and unique interactions. Results reveal that almost all clusters analysed (99%) contain isoforms interacting with unique protein partners, with an average unique to shared interaction rate of ~5. Similar results are obtained by analysing the data from public protein-protein interaction databases. These findings are significant in that they demonstrate that isoforms tend to interact with unique partners, indicating that they could be involved in different interaction networks potentially for performing different functions. Hence, it can be concluded that transcript diversity has a potential to generate a significantly diverse interactome.

The literature analysis presented here gives access to protein interactions that are not yet contained in public resources and in particular, that are linked to transcript isoforms generated by alternative splicing and stored in HumanSDB3. TBIID is accessible at http://tbiid.emu.edu.tr serving as an up to date and comprehensive resource for future experiments on isoform interactions.

Keywords: alternative splicing, protein isoforms, biomedical text mining, abstract retrieval, interaction abstract selection, protein-protein interaction extraction, machine learning, interaction variability analysis.

ÖZ

Son on yılda, genomik teknolojilerde elde edilen gelişmeler, büyük miktarda protein-protein etkileşimi verisinin ortaya çıkarmasına yol açmıştır. Yine de, mevcut protein-protein etkileşimi veritabanları proteomun sadece bir kısmı ile ilgili etkileşim bilgisini kapsamakta ve protein izoformu etkileşimleri bilgisini de seyrek olarak içermektedirler. Bu izoformlar, transkript çeşitliliği mekanizmaları (örneğin alternatif sıplays) tarafından üretilirler ve işlevsel farklılıklar gösterebilirler. İzoformların protein-protein etkileşim verileri, fonksiyonel benzerliklerini analiz etmek ve transkript çeşitliliğinin, protein-protein etkileşim ağlarına etkilerini anlamak için gereklidir. Biyomedikal literatür, izoform etkileşim bilgisini, bilgisayara dayalı yöntemler ile toplamak, görüntülemek ve analiz etmek için deneysel yöntemlere paha biçilmez bir tamamlayıcı kaynak oluşturur.

Bu çalışmada, HumanSDB3'de (insan transkriptomu alternatif sıplays veritabanı) kümelenmiş transkriptler ile bağlantılı insan proteini izoformlarına ait protein etkileşimlerini biyomedikal literatürden çıkaran, kapsamlı bir otomatik metin madenciliği tabanlı analiz sunulmaktadır. Çıkarılan protein-protein etkileşimi verileri, transkript tabanlı izoform etkileşimleri veritabanı (ingilizce kısaltması TBIID) adı verilen yeni bir veritabanı üzerinden erişime ve kullanıma sunulmuştur.

TBIID 7,161 değişik proteine ait, toplam 31,819 etkileşim bilgisi içerir. Etkileşim verileri, otomatik olarak, HumanSDB3'deki izoformlara ait yaklaşık 4 milyon Medline kayıtından seçilen 205,207 etkileşim özetinden toplanmıştır. Kullanılan, otomatik ekstraksiyon yöntemleri, bu alanda ulaşılan en son gelişmeleri yansıtan yüksek bir performans sergilemektedir (53.22% Duyarlık, 68.94% Geri Çağırım, 60.07% F_1-skor).

TBIID, izoformların ortak ve özgün etkileşim ortaklarına dayalı olarak, izoform etkileşimleri değişkenliğini ölçmek için kullanılmıştır. Sonuçlar, hemen hemen tüm transkript kümelerinin (%99), özgün etkileşimin ortak etkileşime oranı ~5 olan izoformlar içerdiğini ortaya koymaktadır. Kamuya açık protein-protein etkileşimi veritabanlarının içeriği analiz edilerek benzer sonuçlar elde edilmiştir. Bu bulgular, izoformların, potansiyel farklı işlevleri yerine getirmek için, farklı etkileşim ağlarında görev alıp, farklı ortaklar ile etkileşim eğiliminde olabileceklerini gösterdiğinden önem taşımaktadır. Bu nedenle transcript çeşitliliğinin, önemli ölçüde çeşitlilik gösteren bir interaktom oluşturmak için potansiyele sahip olduğu söylenebilir.

Burada sunulan literatür analizi, var olan protein-protein etkileşimi veritabanlarında henüz bulunmayan ve özellikle HumanSDB3'de bulunan ve alternatif sıplays mekanizması ile ortaya çıkmış insan transkript izoformlarına ait proteinlerin etkileşimlerine erişim sağlamaktadır. TBIID, http://tbiid.emu.edu.tr adresinden erişilebilen, gelecekte yapılabilecek deneyler için güncel ve kapsamlı bir kaynak olarak hizmet vermektedir.

Anahtar Kelimeler: Alternatif Sıplays, protein izoformları, biyomedikal metin madenciliği, öz erişimi, etkileşim bilgisi içeren özlerin seçimi, proteinler arasındaki etkileşimlerin çıkarımı , otomatik öğrenme, etkileşim değişkenligi analizi.

To My Family

ACKNOWLEDGMENT

Firstly, I would like to thank my supervisors, Asst. Prof. Dr. Ekrem Varoğlu and Assoc. Prof. Dr. Bahar Taneri for their generous help and guidance throughout this study. They kindly devoted their time for helping me to develop my background in text mining and molecular biology and provided me with many stimulating ideas. This work would have encountered with many difficulties without their encouragement, patience and guidance.

I would like to thank Dr. Dietrich Rebholz-Schuhmann, head of the Rebholz Group, European Bioinformatics Institute, for his intense interest into this study and his invaluable suggestions which contributed to the quality of this work. I'm greateful for Dietrich's support and encouragement.

I would like to express my thanks to my monitoring jury members, Asst. Prof. Dr. Ahmet Ünveren and Assoc. Prof. Dr. Hakan Altınçay. They provided me with useful suggestions throughout this study and critically reviewed this thesis.

I'm thankful to Asst. Prof. Dr. Özlem Uzuner from the University of New York at Albany who kindly accepted to act as my external jury member, devoted her time to carefully review this study and provided me with useful suggestions.

I would like to thank all member of the Rebholz Group of European Bioinformatics Institute for their help and feedbacks on this study. My special thanks go to Christoph Grabmüller, Vivian Lee, Samuel Croset, Ian Lewin, Anika Oellrich and Ying (Yumi) Yan. Dear Christoph, thanks for your generous technical help. Dear

Vivian and Samuel, thanks for your critical review of this study as well as contribution to the manual analysis presented in this study. Dear Ian and Anika thanks for your support and feedbacks on this study. Dear Yumi, thanks from the bottom of my heart for your support, generously devoting your time to carefully review and give feedbacks on this study. Also, thanks for all of those thousands of steamy biological discussions which I believe they contribute my knowledge to a great extend. I'm grateful for always being there for me.

I would like to thank Sulan Yiğithan and Erdal Altun from the Computer Engineering Department of Eastern Mediterranean University, for their help during the publication of this thesis.

I owe a lot to my friends, Bahar Beydağlı, Ayça Kayıkçı, Ceren İnce, Laika Karşılı, Yöney Kırsal Ever, Enver Ever and Vincenzina Lo Re for their friendship, endless support, sharing all my happiness and sadness. Each of you has a special place in my heart and a big hug to all of you.

Words are not enough to express my gratefulness to my dear family. My parents, sister and brother supported, motivated and encouraged me throughout this study as well as my life. Even though I live thousands miles away from them, their endless love always shed light on my way. I dedicate this thesis to them as an indication of their significance in my life as well as in this study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS/ABBREVIATIONS

ACT	Article Classification Task
ASD	Alternative Splicing Database
ASTD	Alternative Splicing and Transcript Diversity
b	Threshold, offset from origin
BIND	the Biomolecular INteraction Database
BioCreative	Critical Assessment of Information Extraction Systems in Biology
BOW	Bag Of Words
BRF	Balanced Relative Frequency
CMT	Cluster with Multiple Defined Transcripts
CMT/MII	CMT with Multiple Interacting Isoforms
CMT/MII-eB	CMT/MII with isoforms having External Both types of interactions
CMT/MII-eS	CMT/MII with isoforms having External Shared interactions
CMT/MII-eU	CMT/MII with isoforms having External Unique interactions
CMT/MII-i	CMT/MII with isoforms having Internal interactions
CMT/NII	CMT with No Interacting Isoforms
CMT/SII	CMT with a Single Interacting Isoform
CRF	Conditional Random Field
CST	Cluster with Single defined Transcript
CUT	Cluster with Undefined Transcripts
DB	DataBase
DCS	Document Classification Score
DIP	Database of Interacting Proteins
DNA	DeoxyriboNucleic Acid
DT	Defined Transcript
DWL	Discriminating Word List
ESE	Exonic Splicing Enhancer
ESS	Exonic Splicing Silencer
EST	Expressed Sequence Tag
FN	False Negative
FP	False Positive
GN	Gene Normalisation
GO	Gene Ontology
HPRD	the Human Protein Reference Database
HumanSDB3	Human Splicing Database version 3
IAS	Interaction Article Subtask
IASEL	Interaction Abstract SELection
ID	Identifier
IDF	Inverse Document Frequency
iHOP	Information Hyperlinked Over Proteins
IR	Information Retrieval
ISE	Intronic Splicing Enhancer
ISS	Intronic Splicing Silencer
JNLPBA	Joint Workshop on Natural Language Processing in Biomedicine and its Applications
$K(\mathbf{x},\mathbf{y})$	Kernel function
L_d	Dual Lagrangian

L_p	Primary Lagrangian
MINT	Molecular INTeraction Database
MIPS	The MIPS Mammalian Protein-Protein Interaction Database
ML	Machine learning
MppDB	Mouse Protein-Protein interaction DataBase
mRNA	Messenger RNA
MUC	Message Understanding Conferences
MUTDB	MUTation Database
NCBI	National Center for Biotechnology Information
NE	Named Entity
NER	Named Entity Recognition
NLP	Natural Language Processing
NPM	Number of distinct Protein Mentions
OMIM	Online Mendelian Inheritance in Man
PAS	Predicate Argument Structure
PINA	Protein Interaction Network Analysis Platform
POS	Part of Speech
PPI	Protein-Protein Interaction
PPIN	Protein-Protein Interaction Network
Pre-mRNA	Precursor mRNA
RNA	Ribonucleic Acid
RF	Relative Frequency
S	Shared interaction
STS	Search Term Set
SVM	Support Vector Machine
SVM-TK	Support Vector Machine-Tree Kernels
TBIID	Transcript Based Isoform Interaction Database
TF	Term Frequency
TN	True Negative
TP	True Positive
TREC	Text REtrieval Conferences
U	Unique interaction
UKPMC	United Kingdom PubMed Central
UTR	UnTranslated Region
$ln()$	Natural Logarithm
$\mathbf{w}$	Weight vector
$\mathbf{x}$	Input vector
y_i	Class label, $1 \leq i \leq N$
$\mathbf{z}$	Input test sample
α	Precision/recall adjustment factor for F-score
ξ	Slack variable
λ_i	Lagrange multiplier with index *i*
χ^2	Chi-Square
$\Phi(x)$	Transformed space

Chapter 1

INTRODUCTION

1.1 Motivation

1.1.1 Biomedical Information Overload and the Need for Biomedical Text Mining

The biomedical science community has witnessed the completion of human genome sequence in this decade which still remains as one of the most important scientific events (International Human Genome Sequencing Consortium, 2004). Efforts in human genome sequencing have lead to significant developments in experimental techniques, which have tremendously accelerated high-throughput genome wide studies. Such studies reveal large amounts of experimental data which calls for automated methods from the field of bioinformatics. This field merges biology with computer science and enables scientists to discover new biological insights. Today, bioinformatics methods are being used to analyse experimental data and to deliver the results in a structured format through databases and ontologies. Such repositories often contain new observations such as protein interactions (e.g. Database of Interacting Proteins (DIP), http://dip.doe-mbi.ucla.edu/dip/Main.cgi), functions (e.g. Gene Ontology (GO), http://www.geneontology.org/), diseases and disorders (e.g. Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/omim), mutations (e.g. MUTation DataBase (MUTDB), http://mutdb.org/) and alternative splicing events (e.g. Alternative Splicing and Transcript Diversity (ASTD), http://www.ebi.ac.uk/astd/main.html). They serve as useful resources for assisting research in biomedical science. However, some of them such as GO and protein interaction databases are often sparsely populated (Cusick *et al.*, 2009). Furthermore, it is important to keep such repositories up to date. These problems call for complementary resources and require new methods for processing such resources.

Recent advances achieved in experimental techniques enable scientists to rapidly deliver new discoveries to the biomedical science community through publications. These new discoveries often cover invaluable findings about proteins, genes, pharmaceuticals and other compounds, protein interactions, and cellular as well as pathological processes. The amount of existing biomedical literature, which contains such knowledge, has already moved beyond the possibility of manual curation. Moreover, its volume expands at an increasing rate. For example, the National Center for Biotechnology Information's (NCBI) PubMed (http://www.ncbi.nlm.nih.gov/pubmed) which is the most comprehensive and widely used biomedical literature repository expands at a double exponential pace as shown in Figure 1.1. Its annual growth rate is estimated as ~4% (Lu, 2011) and as of December, 2011 PubMed includes over 21 million citations from Medline and other life sciences journals. Hence, the scientific literature remains an invaluable resource for complementing the existing Protein-Protein Interaction (PPI) repositories and keeping them up to date. However, this information rich resource is available only for those who can interpret the data represented in natural language form. This calls for techniques from text mining, a field in Natural Language Processing (NLP) aiming to discover new information from textual data, which is unstructured, highly variable and ambiguous. In particular, methods from the field of biomedical text mining are being used to process the scientific textual data. The aim is to find relevant pieces of information, which are represented in an unstructured form within scientific texts through computational methods. These methods are usually adopted from general artificial intelligence, statistics and data mining. They are applied in order to retrieve relevant documents, to classify them, to identify the biomedical entities and to extract relations between them by using various NLP methods such as part-of-speech (POS) taggers, stemmers, syntactic and semantic parsers as well as lexicons.

Biomedical text mining has increasingly attracted interest within the last decade and many applications have been developed. The field deals with the information overload problem by automatically “understanding” relevant pieces of data based on human-like comprehension of language from the scientific literature. However, this field is quite new and still remains challenging. The primary barrier lies in the nature of language ambiguity, i.e. multiple interpretations of the same string. Ambiguity complicates almost every single level of linguistic analysis. For example, protein/gene name recognition is a hard task given that protein/gene symbols are known to be ambiguous within and across different organisms and with other biomedical entities like chemicals. Hence, it is likely that many of the biomedical text mining tasks will remain unresolved in the near future.

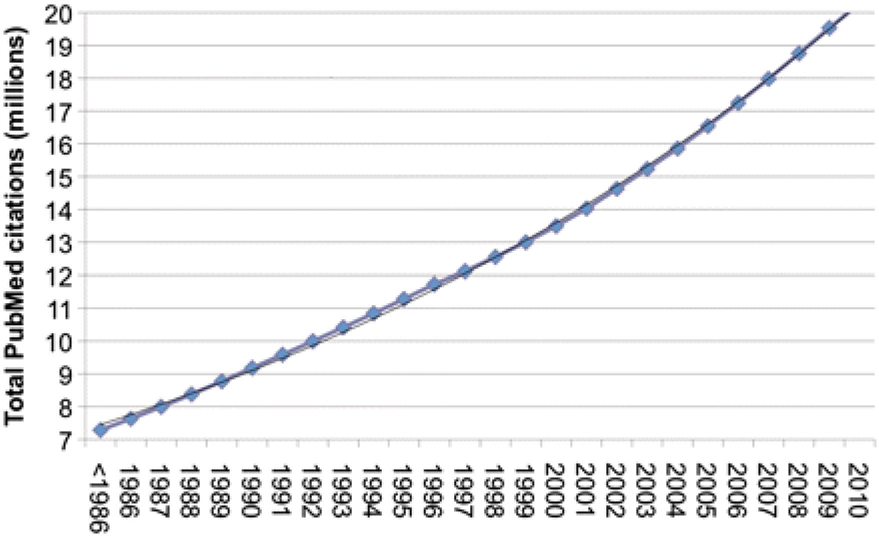


Figure 1.1: Growth of the PubMed Database between 1986 and 2010

(Source: (Lu, 2011))

1.1.2 Importance of Protein Isoform Interactions

Recent studies in the field of biomedicine have focused on PPIs to understand functions of organisms at the molecular level. Understanding PPIs is very important for understanding the signalling pathways involved in cellular activities and biochemical as well as disease processes. Some examples of cellular activities include Deoxyribonucleic Acid (DNA) replication, transcription and cell cycle control (Ono *et al*, 2001).

Scientific interest in PPIs led to generation of many PPI databases which deliver information in structured formats to the public through the internet. The PPI data archived in such databases can be gathered through experimental as well as literature mining methods (Zhou *et al.*, 2008). The experimental methods utilise either low-throughput (e.g. X-ray crystallography, fluorescence resonance energy transfer) or high-throughput (e.g. yeast two-hybrid, affinity purification) techniques to collect PPI data (Shoemaker and Panchenko, 2007). Low-throughput techniques provide accurate information on PPIs while high-throughput techniques provide less accurate information but allow performing proteome-wide experiments (Browne *et al*., 2010). On the other hand, literature mining methods rely on biomedical text mining approaches to gather PPI data from the scientific literature. Although there are many PPI databases such as DIP (Xenarios *et al*., 2000), the Molecular INTeraction Database (MINT) (Zanzoni *et al*., 2002) and IntAct (Hermjakob *et al*., 2004), they cover only a portion of the interactome and the overlap between them is small mainly due to different techniques and publications utilised to generate their contents (Mathivanan *et al*., 2006; Prieto and Rivas, 2006). Moreover, they contain a limited number of PPIs involving protein isoforms. For example, the Protein Interaction Network Analysis Platform (PINA) (Wu *et al*., 2009) which is a comprehensive PPI database contains only 772 interaction pairs (1.3% of all interactions in PINA) where at least one of them is identified to be an alternative splicing variant from the Uniprot Knowledge Base (http://www.uniprot.org/). Such variants are introduced by alternative splicing, an important cellular phenomenon, which significantly contributes to transcriptome diversity.

1.1.2.1 Alternative Splicing – The Main Source of Transcriptome Diversity

Alternative splicing is a cellular process which regulates gene expression and leads to production of multiple protein isoforms from a single gene by generating structurally different messenger ribonucleic acids (mRNAs) from the same precursor mRNA (pre-mRNA) sequence (Nilsen and Graveley, 2010; Taneri *et al*., 2011). A pre-mRNA molecule consists of protein coding regions (exons) and non-coding regions (introns) (House and Lynch, 2008). During the process of alternative splicing, different mRNAs are produced by making use of different splice sites of exon-intron junctions (Taneri, 2005). The splicing process is controlled by a mechanism called the spliceosome which is able to recognise specific sequences and signals marking the beginning and end of each intron within the pre-mRNA (House and Lynch, 2008).

Alternative splicing and other kinds of transcript diversity mechanisms, like gene duplication and allelism are the sources of transcript diversity leading to a diverse proteome in eukaryotic genomes (Chothia *et al.*, 2003; Graveley 2001). Gene duplication is the process by which a chromosome or a DNA portion is replicated resulting in two identical genes (Moleirinho *et al.*, 2011). Alleles are different forms of a given gene that occupy the same locus on a chromosome and control the same trait (Campbell and Heyer, 2004). Hence, in the case of gene duplication or allelism, protein isoforms are generated from related genes rather than a single gene which is the case in alternative splicing. Protein isoforms generated through these mechanisms are expected to exhibit differences such as gain, loss or divergence in their functions given that they have different structures (Modrek and Lee, 2002; Moleirinho *et al.*, 2011).

Analysing the sequence data across reference databases such as ASTD (Koscielny *et al.*, 2009), ProSAS (Birzele *et al.*, 2008) and ASAP II (Kim *et al.*, 2007), which gather transcript diversity and alternative splicing events have revealed that alternative splicing is widespread in various genomes. For example, in *homo sapiens* (human) up to 94% (Wang *et al.*, 2008), and in *mus musculus* (mouse) 79% (Taneri *et al.*, 2005) of the genes have been reported to exhibit alternative splicing. The proportion of the alternative splicing detected in a given genome depends on the number of transcript data available for that genome (Taneri, 2005). As the analysis techniques are improved and as a consequence the amount as well as quality of sequence data increase, the transcript diversity detected relevant to alternative splicing will also increase. Alternative splicing is known to be the major source of transcript diversity in eukaryotic genomes (Black, 2000) given that it is a widespread process leading to generation of multiple different mRNA transcripts from a given single pre-mRNA.

1.1.2.2 Effects of Alternative Splicing on Isoform Functions and Interactions

Protein isoforms produced from the same gene may share the same function, show minimal functional differences, or have entirely opposite functions (Stamm *et al.*, 2005). Isoforms having functional differences would be expected to show differences in selecting their interaction partners. For example, the human Slit receptor Robo3 has two isoforms, which are Robo3.1 and Robo3.2 differing in their carboxy terminal groups. Both isoforms interact with Slit ligands during neurogenesis regulation. However, they exhibit opposite functions due to their structural differences. Robo3.1 silences Slit repulsion while Robo3.2 favours Slit repulsion during the midline crossing events in the commissural axons (Chen *et al.*, 2008). Another example can be given from the *C. elegans* genome. The FGF receptor, EGL-15 has two isoforms namely EGL-15(5A) and EGL-15(5B). These isoforms differ in their extracellular domains leading to different functions in the gonadal chemoattraction of the migrating sex myoblasts (SMs). Isoform 5A plays a role in attraction of the migrating SMs to the gonad, while isoform 5B plays a role in repulsion of the migrating SMs from the gonad (Lo *et al.*, 2008). Although there are many studies in the literature which report on protein isoforms, their functions and interactions, they mainly focus on single genes. However, it is important to perform large-scale analyses in order to understand the effect of transcript diversity mechanisms on isoform interactions and to gain insight on their functions at a global level.

1.2 Overview of the Study

In this study, a large scale analysis is performed to measure the diversity in human protein isoform interactions based on the systematic analysis of the scientific literature by using biomedical text mining methods. The main goal of this study is to understand the effects of transcript diversity on the human protein interaction networks and to gain insights into functional similarities of protein isoforms. For this purpose, a comprehensive text mining pipeline utilising the content of the Human Splicing Database, version 3 (HumanSDB3) (Taneri *et al*., 2004; Taneri *et al*., 2005) is developed to gather interactions of the isoforms from the scientific literature. HumanSDB3 provides comprehensive genomic and transcriptomic data for human alternatively spliced genes. However, data regarding the protein isoform interactions is not included in the current version of the database (Taneri *et al*., 2004; Taneri *et al*., 2005).

In order to extract the interactions of the isoforms linked to the clustered transcripts from HumanSDB3, a total number of 4,083,094 Medline abstracts are analysed through an automated text mining pipeline (Kafkas *et al*., 2007; Kafkas *et al*., 2008; Kafkas *et al*., 2009a). For this purpose, a Support Vector Machine (SVM) (Vapnik, 1995) (here called the IASEL SVM classifier) trained on the BioCreative-II IAS corpus (Krallinger *et al*., 2008) with a novel and high performing feature set is used for selecting interaction abstracts (Kafkas *et al*., 2009b). Another SVM (here called the PPI SVM classifier) trained by utilising syntactic parsers information on the AIMed corpus (Bunescu *et al*., 2005) is utilised to extract the isoforms' interactions from the selected abstracts (Kafkas *et al*., 2010a). Manual analysis on a randomly selected set of findings reveals that overall the developed automated methods exhibit state of the art performance (53.22% precision, 68.94% recall, 60.07% F_1-score). The isoform interactions extracted from the scientific literature are archived in an interaction database called the Transcript Based Isoform Interaction Database (TBIID) (accessible via http://tbiid.emu.edu.tr). The database contains 31,819 distinct interactions belonging to 7,161 proteins. The content of TBIID is utilised to quantify the variability in isoform interactions. The variability analysis is based on the subset of interactions belonging to clusters having more than one distinct interacting protein isoform. During the variability analysis, differences in the number of interaction partners are quantified for a total number of 1,226 proteins and a total number of 1,540 interactions. The results reveal that almost all of the clusters analysed (99%) contain isoforms showing variation in their interactions. Similar results are obtained in comparison to the reference PPI databases. The results indicate that isoforms are characterised to interact with unique partners and hence they involve in different interaction networks for potentially exhibiting different biological functions (Kafkas *et al.*, 2010b). This study is important given that it demonstrates that alternative splicing and possibly other kinds of transcript diversity mechanisms lead proteome diversity and thus have a potential to generate a highly diverse interactome. The core of this study is published in (Kafkas *et al.*, 2011).

1.3 Hypotheses of the Study

The major hypotheses of this study: For the first time, this study bridges transcript diversity and protein interactions to analyse the effects of transcriptome and thus proteome diversity on the human interactome using data from the scientific literature at a large scale. Although, biomedical text mining methods have been used to tackle PPI extraction or other kinds of information extraction tasks (Albert *et al.*, 2003;

Donaldson *et al.*, 2003; Hakenberg *et al.*, 2010; Miwa *et al.*, 2009b; Waagmeester *et al.*, 2009), only a few studies have benefited from such methods to gather alternative splicing and transcript diversity relevant information from the scientific literature (Cheng *et al.*, 2008; Shah *et al.*, 2005). However, alternatively spliced variants as well as other kinds of isoforms are the main sources of proteome diversity which has a potential to lead to significant variation in protein interactions (Jaeger *et al.*, 2008). Therefore, it is important to have a global perspective on the variability in isoform interactions. For this purpose, an analysis is presented for the first time to quantify the isoform interaction variability.

TBIID is presented as a novel database which serves as a comprehensive resource on isoform interactions and supports further investigation on functional differences of isoforms based on the interaction variability presented.

Two designed PPI related tools, the IASEL SVM classifier and the PPI SVM classifier are presented as practical tools in the biomedical text mining domain. To the best of our knowledge, the IASEL SVM classifier used in this study has the second best performance reported in the literature on the BioCreative-II IAS test corpus. It is not possible to directly compare all the existing PPI extractors due to different pre-processing methods and/or evaluation metrics used to report their performances. However, the developed PPI extractor performs at the state-of-the art level on the AIMed corpus, given that its performance is within the performance range of the other PPI extractors reported in the literature which follows the same performance evaluation prodecure.

1.4 Organisation of the Thesis

This thesis is organised as follows: Chapter 1 presents the introduction. In Chapter 2, the biological background focusing on the major transcriptome diversity source, alternative splicing is presented. Chapter 3 provides background in PPI extraction related biomedical text mining tasks. Chapter 4 provides details of the developed text mining systems for extracting isoform interactions from the scientific literature. Chapter 5 describes the isoform interaction variability analysis in detail. Chapter 6 presents the generated database, TBIID. Comparison of the content of this database with publicly available PPI resources is discussed and the web interface of the database is described in this chapter. Lastly, in Chapter 7, the findings are summarised, discussed and future research directions are presented. The details of the evaluation metrics and SVM library which is used for designing the IASEL and PPI extraction classifiers are presented in the appendices.

Chapter 2

ALTERNATIVE SPLICING

2.1 Eukaryotic Gene Structure

All organisms, including the simplest uni-cellular and the most complex mammals consist of cells. Simple organisms (e.g. bacteria) have prokaryotic cells, while the more complex ones (e.g. vertebrates) have eukaryotic cells. One of the most fundamental differences between the two types of cells is that: a eukaryotic cell has a nucleus containing its DNA, while the genetic material is not membrane-bound in a prokaryotic cell. DNA is the heredity material nearly in every cell of almost all organisms (Miko and LeJeune, 2009). Structure of the DNA molecule is depicted in Figure 2.1. DNA stores the complete genetic information needed for building and maintaining an organism. This information is stored as a code made up of four nucleotide bases: Adenin (A), Guanine (G), Thymin (T), and Cytosine (C). DNA consists of two strands of nucleotides in the form of a double-helix. A nucleotide molecule is made up of one base, a sugar molecule (deoxyribose) and a phosphate molecule (Miko and LeJeune, 2009). Genes are made of DNA and are the basic functional units of heredity. They act as instructions to make protein molecules. In a eukaryotic cell, DNA is organised into structures called chromosomes. Eukaryotes have diverse number of genes and chromosomes (Miko and LeJeune, 2009). For example, it is estimated that humans have between 20,000 and 25,000 genes (International Human Genome Sequencing Consortium, 2004). It is known that a human cell contains 23 pairs of chromosomes, where 22 pairs are autosomes and the remaining pair is the sex chromosomes (Miko and LeJeune, 2009).

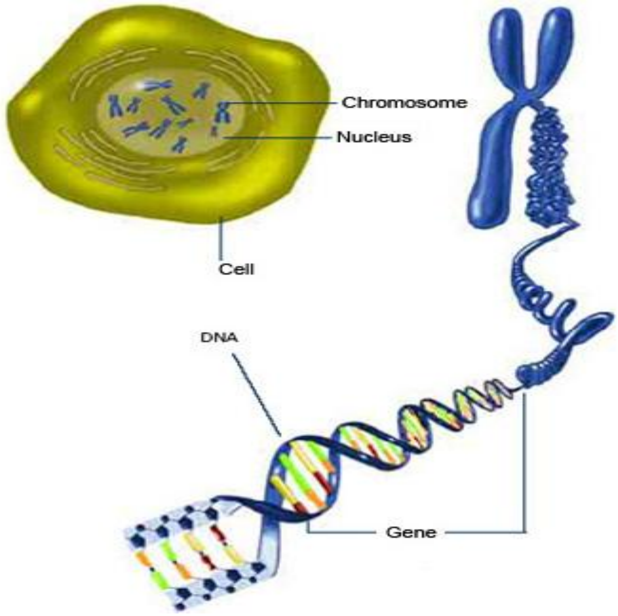


Figure 2.1: Structure of DNA

(Source: http://publications.nigms.nih.gov/thenewgenetics/chapter1.html)

Eukaryotic gene expression refers to the generation of protein or RNA from the information contained in the gene (Larson *et al.*, 2009). Structure of a eukaryotic gene is important for the regulation of gene expression. Eukaryotic genes often have regulatory regions that facilitate gene expression (Lynch, 2006). These regions are promoters, transcriptional start site, exonic, intronic regulatory motifs and transcriptional stop site (Figure 2.2). Transcription is the process of RNA synthesis from the DNA template during gene expression (see section 2.2 for details) (Moorhouse and Barry, 2004). Promoters which are also called transcription regulatory regions are located upsteam of genes. Promoters contain binding regions for the RNA polymerase enzyme and transcription factors, which are involved in the transcription regulation process. Transcription start sequences identify where DNA transcription starts. Exons are the coding regions and introns are non-coding regions of a gene. Transcription stop sequences are located downstream of a gene and specify where RNA transcription stops.

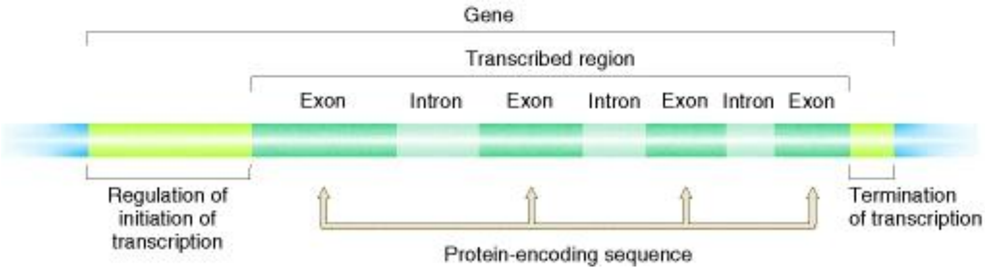

Gene
Transcribed region
Exon
Intron
Exon
Intron
Exon
Intron
Exon
Regulation of initiation of transcription
Protein-encoding sequence
Termination of transcription

Figure 2.2: Eukaryotic Gene Structure

(Source: http://www.ncbi.nlm.nih.gov/books/NBK22032/)

2.2 Eukaryotic Gene Expression and Cellular Mechanisms Increasing its Complexity

Each cell’s behaviour is controlled and determined by the functional molecules called proteins. Eukaryotic gene expression refers to the synthesis process of a functional product, which is often a protein as well as non-coding RNAs (Figure 2.3). It starts with the process of transcription, where messenger RNA (mRNA) is synthesized from the DNA template (Miko and LeJeune, 2009). Transcription starts with the binding of RNA polymerase enzyme to the promoter region of the DNA. This binding process is mediated by a protein complex formed from transcription factors. During the process of transcription, RNA polymerase reads through the DNA sequence and produces a complementary RNA sequence called precursor mRNA (pre-mRNA) (Miko and LeJeune, 2009). A modified guanine nucleotide is added to the 5'-end of the pre-mRNA, shortly after the start of transcription. This process is termed as capping. Capping is important for maintaining mRNA's stability and translation (Furuichi and Shatkin, 2007). Transcription is terminated by the cleavage of the produced transcript and followed by polyadenylation process, which involves the addition of A bases to the 3'-end of the pre-mRNA sequence. Polyadenylation is important for the translation and stability of mRNA molecule (Colgan and Manley, 1997).

A pre-mRNA molecule, which is transcribed from DNA consists of introns and exons. Introns are spliced out after transcription, while exons are retained in the final mRNA molecule. The process of intron removal and exon ligation is referred to as *splicing* (Nilsen, 2003). Splicing is performed by a ribonucleoprotein complex, called *spliceosome*. The *spliceosome* is composed of five small nuclear RNAs, termed as U1, U2, U4, U5, U6 and more than 300 distinct proteins (Nilsen, 2003). Briefly, the *spliceosome* works as follows: First it recognises the 3' and 5' splice sites of exon-intron junctions on the pre-mRNA. The 5' and 3' splice sites on the pre-mRNA are the conserved sequences that define starts and ends of introns. The recognition of 5' and 3' splice sites by the spliceosome is still not well understood. Nevertheless, it is known that the splice sites interact with some specific RNA and protein factors for engaging the spliceosome into splicing (Nilsen, 2003). Once the ends are recognised, the nucleotide sequence between these ends is removed and the exons are spliced together. In many cases, splice sites are joined together in different combinations resulting in different mRNAs and therefore generating RNA transcript diversity (Matlin *et al.*, 2005). This phenomenon is termed as *alternative splicing* which is discussed below.

Splicing is followed by *translation*, which is the protein synthesis from the mRNA template (Moorhouse and Barry, 2004). Translation starts with the binding of the ribosome small subunit to the 5' end of the mRNA sequence. This process is mediated by proteins called the initiation factors. During the translation process, the ribosome reads through the mRNA sequence and produces a protein. Termination of the process occurs when the ribosome reads a stop codon.

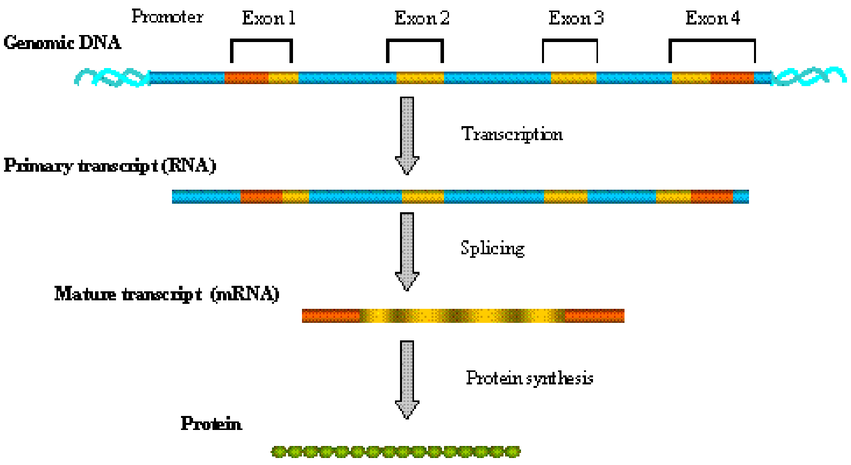


Figure 2.3: Eukaryotic Gene Expression

(Source: http://www.ncbi.nlm.nih.gov/projects/genome/probe/doc/ApplExpression.shtml)

Gene expression in eukaryotic cells is a complicated process which involves large number of protein-protein, protein-RNA and protein-DNA interactions (see section 2.3 for details). Several mechanisms including alternative splicing, alternative polyadenylation and RNA editing contribute to the complexity of gene expression. These mechanisms are discussed in detail in the following sections.

2.2.1 Alternative Splicing

Alternative splicing is the process producing different mRNAs from the same primary pre-mRNA by making use of different splice sites (Matlin *et al*, 2005). This process is depicted in Figure 2.4. As a result of this process, structurally and functionally different proteins can be produced from the same gene. Alternative splicing is widespread in different eukaryotic organisms (Kashyap and Sharma, 2007). Hence, this process contributes to transcriptome and proteome diversity. For example, alternative splicing leads to production of different *Neurospora crassa* Tob55 protein isoforms (a fungal protein), which differ in their ability to insert β-barrel proteins into the outer mitochondrial membrane (Hoppins *et al.*, 2007).

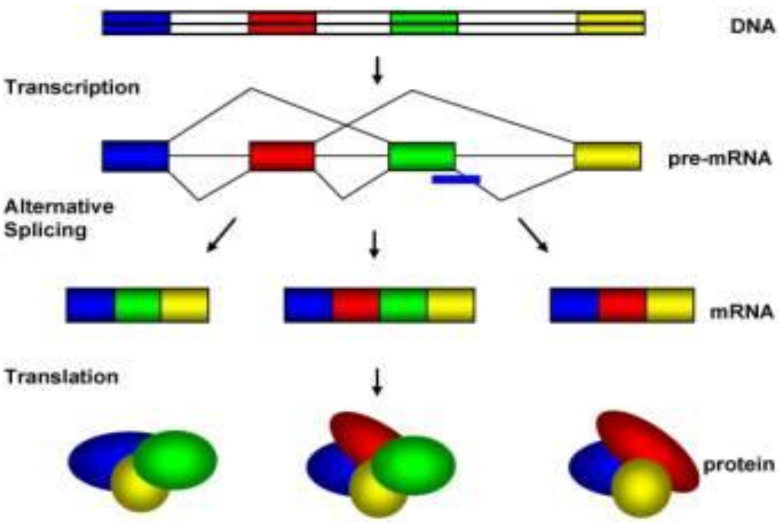


Figure 2.4: Alternative Splicing
Exons are shown in blue, red, green and yellow rectangles while introns are uncoloured. Protein domains are shown in circles.

(Source: http://designmatrix.wordpress.com/2010/03/30/introns-and-design-2/)

Effect of alternative splicing on a single gene as well as genome-wide level has been analysed in several studies. Its effect on a particular gene at the transcript level can range from a few transcript variants to a very large number of transcript variants. For example, it is possible to produce up to 38,016 different mRNAs from the *Drosophila* (fruit fly) DSCAM gene, which contains 95 alternatively spliced exons (Celotto and Graveley, 2001). The potential number of mRNAs that can be produced from the DSCAM gene is more than twice the number of genes in the entire *Drosophila* genome. Splice variants are generally termed as major and minor isoforms. Type of a particular isoform often highly depends on the tissue in which the gene is expressed (Trafton, 2008). Major isoforms more frequently appear than minor ones. In addition, it is possible to have multiple major isoforms for a given gene and they differ as the tissues differ. For example, in *M. galloprovincialis* (mussel), tropomyosin has 3 isoforms which are designated as TM1, TM2 and TM3. TM1 is identified as the major isoform, which appears in the various muscle tissues including adductor, cardiac, anterior pedal retractor, mantle and gills. On the other hand, TM2 and TM3 are identified as minor isoforms appearing only in the mantle and in both the mantle and gills, respectively (Itoh and Fujinoki, 2008).

Genome-wide analyses on sequence data have revealed that alternative splicing is widespread within and across different eukaryotic genomes. In human, 81-94% (Koscielny *et al.*, 2009; Taneri *et al.*, 2005; Wang *et al.*, 2008), in mouse 74-79% (Koscielny *et al.*, 2009; Taneri *et al.*, 2005), in rat 39-61% (Koscielny *et al.*, 2009; Lee *et al.*, 2007; Taneri *et al.*, 2005) and in rice 42% (Filichkin *et al.*, 2010) of genes have been found to exhibit alternative splicing. These findings indicate that alternative splicing is present across eukaryotic species and contributes to the transcript diversity to different degrees in various genomes.

2.2.2 Alternative Polyadenylation

Similar to alternative splicing, alternative polyadenylation is another widespread mechanism controlling the gene expression. This process leads to the generation of multiple mRNAs which differ in their 3' UTRs (untranslated regions) or coding regions from a single gene (Shen *et al.*, 2008). Polyadenylation starts with the interaction of the CPSF enzyme and CstF with specific sequences (termed as poly(A) signals) on the generated RNA for its cleavage from the 3' UTR after the transcription process (Giammartino *et al.*, 2011). Shortly after, the polyadenylation catalyser, polyadenylate polymerase creates the poly(A) tail (a sequence of adenine nucleotides). The protein PAB2, binds to the created poly(A) tail, for increasing the affinity of the catalyser. Polyadenylation stops whenever the poly(A) tail becomes long enough and the enzyme can no longer bind to CPSF. In the case of alternative polyadenylation, selection of the alternative sites depends on the expression of the proteins involving in polyadenylation as well as extracellular stimuli (Shell *et al*., 2005).

In some cases, the alternative polyadenylation sites are located in the 3' UTR only. This results in production of multiple mRNAs differing in their 3' UTRs but all of them code the same protein. In some other cases, alternative polyadenylation sites exist in internal introns/exons resulting in generation of different protein isoforms (Giammartino *et al*., 2011). Thus, similar to alternative splicing, alternative polyadenylation contributes to proteome diversity.

Genome-wide level bioinformatics studies have revealed that ~54% (Zhang, Lee and Tian, 2005; Tian, Zhang and Lutz, 2005) and ~60% (Shen *et al*., 2011) of genes in human and in *A. thaliana* respectively undergo alternative polyadenylation. These findings indicate that alternative polyadenylation plays a significant role in gene expression regulation across different species.

2.2.3 RNA Editing

RNA editing is a means for post-transcriptional alteration of RNA sequences, which can occur in different forms: nucleotide insertion, deletion and substitution (Farajollahi and Maas, 2010). RNA editing alters the basic coding sequence only. Thus, it does not include RNA processing events which are the cases in splicing and polyadenylation. RNA editing takes place prior to splicing of the pre-mRNA; where these modifications can also result in alternative splicing.

RNA editing can alter the genetic information content carried by an mRNA transcript, both by changing the coding sequence or by creating new splicing sites. Edited RNA transcripts exhibit different sequences compared to their unedited counterpart transcripts. Therefore, they may show different functional activities from that shown by the unedited transcripts. Hence, RNA editing can increase the proteome complexity of organisms.

Among the various types of RNA editing, the A-to-I base modification is the most widespread type in higher eukaryotes (Nishikura, 2010). During the A-to-I editing, adenosine (A) residues are deaminated and changed into inosine (I) residues. The editing of adenosines is catalysed by a small family of enzymes termed adenosine deaminases acting on RNA (ADARs) (Reenan, 2001). There are 3 different ADARs in human; ADAR1, ADAR2 and ADAR3 where the first two ones are responsible from most of the A-to-I editing process (Farajollahi and Maas, 2010). ADAR1 and ADAR2 are ubiquitously expressed in the brain, where 1/17,000 nucleotide is edited (Paul and Bass, 1998). Other known RNA editing types include C-to-U, G-to-A, and U-to-C. In C-to-U editing, a cytidine (C) nucletide is deaminated and changed into a uridine (U). The ApoBec-1 enzyme plays a catalyser role in the editing of cytidines (Chester *et al.*, 2000). For the case of G-to-A, U-to-C editing, neither the molecular mechanism(s) nor the involved enzymes are known to date.

2.3 Mechanism and Types of Alternative Splicing

The exons to be retained in the final mRNA molecule are determined by the process of splicing. Regulation and selection of splice sites is done by *trans*-acting protein (repressors and activators) networks which bind to *cis*-acting sites (silencers and enhancers) of the RNA forming splicing signals (Wang and Burge, 2008). In addition to these signals, there are other signals playing a role in alternative splice site selection. Exons and introns may have enhancer and silencer sites where proteins can bind and regulate alternative splice site selection. These sites are called Exonic Splicing Enhancer (ESE), Exonic Splicing Silencer (ESS), Intronic Splicing Enhancer (ISE) and Intronic Splicing Silencer (ISS) (Cartegni *et al.*, 2002).

Splicing activator proteins, which are generally members of the serine/arginine-rich (SR) protein family bind to ESEs and enable Exon Definition which is the process of recognition of a particular exon by the spliceosome (Cartegni *et al.*, 2002; Matlin *et al.*, 2005). They may also bind to ISEs and can promote Intron Definition which is the process of recognition of a particular intron by the spliceosome.

Splicing repressor proteins, such as heterogeneous nuclear ribonucleoproteins (hnRNPs) family including hnRNPA1 and polypyrimidine tract binding protein (PTB), bind to ESSs and silence splicing (Cartegni *et al.*, 2002). They may also bind to ISSs and result in skipping of alternative exons (Cartegni *et al.*, 2002; Matlin *et al.*, 2005)

There are five main types of alternative splicing (Figure 2.5). These are listed below:

(a) Cassette-exon inclusion or skipping: In this type, an exon may be retained or spliced out of the pre-mRNA.

(b) Alternative 3' splice-site selection: In this type, an alternative 3' splice site (acceptor) is used. This changes the 5' boundary of the downstream exon.

(c) Alternative 5' splice-site selection: In this type of splicing, an alternative 5' splice site (donor) is used. This changes the 3' boundary of the upstream exon.

(d) Mutually exclusive exons: In this type of splicing, only one of two exons is retained in mRNAs, but not both of them.

(e) Intron retention: In this type, a sequence may be retained to the final mRNA or spliced out as an intron. This type is different than exon skipping given that the sequence retained is not flanked by introns.

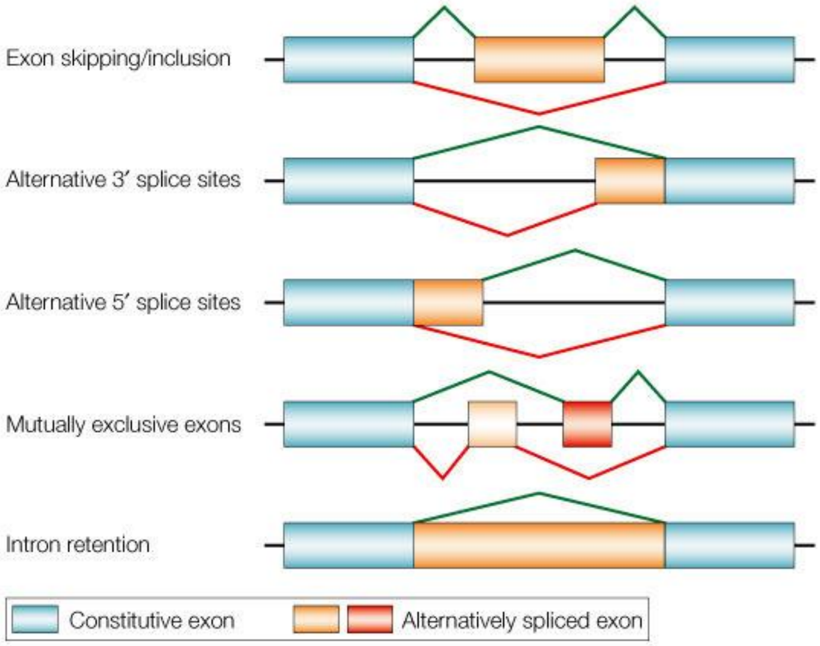

Exon skipping/inclusion
Alternative 3′ splice sites
Alternative 5′ splice sites
Mutually exclusive exons
Intron retention
Constitutive exon
Alternatively spliced exon

Figure 2.5: Types of Alternative Splicing
Constitutive exons are those which are not spliced out by a splicing reaction

(Source: (Cartegni *et al.*, 2002))

2.4 Regulation of Alternative Splicing

During the process of alternative splicing, different protein coding regions (exons) are combined in different ways. In this process, the choice of splice sites to be combined depends on interaction of the protein factors with specific sequences (signals) on the mRNA. Hence, alternative splicing is the result of a complex regulatory network depending on large number of sequences and factors. Specific sequences which take part in the process of alternative splicing are splicing enhancers and silencers (*cis*-acting elements) located on the pre-mRNA molecule which can lead to the selection or skipping of a particular splice site respectively. Protein factors (*trans*-acting elements) can affect the splicing process by binding to *cis*-acting elements or the spliceosome. Additionally, tissues, physiological conditions like stress, as well as developmental stages of organisms can play a role in the regulation of alternative splicing (Matlin *et al*., 2005; Woodley and Valcarcel 2002).

Recent large-scale studies have shown that alternative splicing is a tissue specific process (Castle *et al*., 2008; Wang *et al*., 2008). For example, Castle and colleagues have studied alternative splicing events in 48 different human tissues and have shown that 42% of the exons analysed are differently expressed in at least one of the tissues (Castle *et al*., 2008). Tissue specificity of alternative splicing is mainly driven by tissue-specific *trans*-acting factors targeting *cis*-acting RNA elements (Black 2003). For example, an RNA-binding protein in mouse called Fox-1 is expressed in three different tissues; brain, heart, and skeletal muscle. However, it regulates alternative splicing of the F1γ and α-actinin genes in muscle only (Jin *et al*., 2003).

Trans-acting factors also play a role in the regulation of alternative splicing events in specific developmental stages and/or psychological conditions. In *C. elegans*, the Fox-1 protein binds to the xol-1 gene and regulates its alternative splicing during the sex determination phase (Meyer 2000). In mouse, neuronal acetylcholinesterase (AChE) is alternatively spliced during the neuronal development phase. SC35 protein is one of the splicing factors regulating alternative splicing of this particular gene. Detailed analysis on AChE's alternative splicing events have revealed that under stress, increased SC35 leads to replacement of AChE-S by the AChE-R splicing variant (Meshorer *et al*., 2005).

Regulation of alternative splicing is very complicated and disturbance of this process such as mutations in *cis*-acting elements or *trans*-acting factors can cause diseases like cancer (Faustino and Cooper, 2003). For example, some variants of BRCA1 (Breast cancer gene 1) play a role in hereditary breast cancers. Such variants are generated due to an inherited nonsense mutation in a particular exon (exon 18) which causes disturbance of an exon silencing enhancer. This alters the binding of splicing factor, the SR protein SF2/ASF, leading to inappropriate skipping of the exon 18 (Millevoi *et al*., 2009). Therefore, it is important to analyse the role of alternative splicing in gene regulation at a large scale (Ramani *et al*., 2010) for understanding the diseases associated with its mechanism and enabling discovery of therapeutic drugs.

2.5 Conservation of Alternative Splicing

Splicing is a conserved mechanism in eukaryotes. The conserved splicing mechanism consists of the splicing signals that enableRNA recognition by the spliceosome (e.g. exon-intron junctions at the 5' and 3' ends of introns) and the core of the machinery which is formed by five spliceosomal small nuclear ribonucleoproteins and many protein factors (Keren *et al.*, 2010). Comparative genomic studies focusing on different eukaryotes, especially vertebrates have revealed that they have a high level of genetic similarity. For example, cats have 90% (Pontius *et al.*, 2007), cows have 80% (Elsik *et al.*, 2009) and mice have 80% (Mouse Genome Sequencing Consortium, 2002) genetic identity to humans. Such genomes share high number of genes, usually with conserved intron-exon structures. Scientists have studied evolutionarily conservation of alternative splicing by focusing on such conserved genes (orthologous) between different genomes. They mainly have analysed alternative splicing patterns (Baek and Green, 2005; Nurtdinov *et al.*, 2007; Sorek and Ast, 2003; Thanaraj *et al.*, 2003). For example, in a large scale study, 1,753 constitutive and 243 alternative exons (exons that are alternatively spliced across species) which are conserved between human and mouse genomes have been reported (Sorek and Ast, 2003). In another large scale study, it has been estimated that 7.2% (± 1.1%) of the human exons which are conserved in the mouse genome undergo alternative splicing in both genomes (Sorek *et al.*, 2006). Analyses on the conserved exons have shown that alternative exons are less conserved than constitutive exons in eukaryotes (Keren *et al.*, 2010; Nurtdinov *et al.*, 2007). In addition to exonic sequences, intronic sequences are also evolutionarily conserved. Thanaraj and colleagues have shown that 15% of the alternative and 67% of the constitutive human introns are conserved in mouse (Thanaraj *et al*, 2003).

Evolutionary conservation of splicing patterns provides insights into the functional significance of alternative splicing. For example, the sex determination pathway in *Drosophila melanogaster* (fruit fly) which is controlled by an alternative splicing cascade and is vital for the organism’s survival evolves rapidly (Sánchez, 2008). In the species that are closely related to *Drosophila*, the cascade exhibits slight differences. In other related insects, such as the *Musca domestica* (house fly) and the *Ceratitis capitata* (Mediterranean fruit fly), the pre-mRNA which corresponds to the first gene in the sex-determination pathway does not splice in a sex-specific manner (Sánchez, 2008). This shows that alternative splicing provides evolutionary plasticity given that splicing patterns constantly evolve.

2.6 Effect of Alternative Splicing on Protein Structures

Proteins vary in their biological activities which depend on their distinct three dimensional (3D) structures. The 3D structure of a protein is determined by its amino acid sequence which is coded by exons and translated from the mRNA sequence during the gene expression process. Alternative splicing leads to production of isoforms having differences in their amino acid sequences, and thus 3D protein structures by making use of different splice sites of the pre-mRNA (Möröy and Heyd, 2007). Since alternative splicing can insert, delete or modify functional protein domains (Taneri *et al.*, 2004), differences in protein structures potentially lead to differences in isoform function. Therefore, exons are crucial parameters playing a role in transcript and thus protein diversity in eukaryotes.

The effect of alternative splicing on a particular gene at protein function level can vary from the production of isoforms having the same function, to small functional differences to completely opposite functions. Functional differences of isoforms potentially lead to differences in their interaction partners. For example, the human gene Rab6A which plays an important role in eukaryotic cell membrane transport control has two isoforms, namely Rab6A and Rab6A'. Protein sequences of these isoforms differ in only three amino acid residues, which are located in regions flanking their PM3 GTP-binding domains. Analyses on these particular isoforms have revealed that both of them inhibit secretion in HeLa cells, but Rab6A stimulates the redistribution of Golgi proteins into the endoplasmic reticulum while Rab6A' does not. This shows that Rab6A can induce Golgi-to-endoplasmic reticulum retrograde transport whereas Rab6A cannot. Furthermore, analyses have shown that Rab6A' interacts with two Rab6A protein interaction partners, namely GAPCenA and clone 1, but not with the kinesin-like protein Rabkinesin-6 (a Golgi-associated Rab6A effector). These findings suggest that alternative splicing leads to production of Rab6A isoforms, which exhibit functional differences and interact with distinct sets of protein partners (Echard *et al.*, 2000).

Alternative splicing events are also prevalent in plant genomes such as *Arabidopsis thaliana*. Two alternatively spliced forms of the serine-arginine-rich (SR45) protein, which is a pre-mRNA splicing factor, have been studied in *A. thaliana* (Zang and Mount, 2009). Isoform 1 (SR45.1) differs by 8 amino acids from isoform 2 (SR45.2). A loss-of-function mutant plant that cannot make SR45 protein exhibits some developmental phenotypes affecting roots and flowers. When SR45.1 isoform is expressed in a mutant, the flower phenotype is restored but not the root phenotype. On the other hand, when SR45.2 isoform is expressed, the rooth growth is restored but not the floral morphology. Results show that two SR45 isoforms have distinct functions. Furthermore, this particular case shows that alternative splicing has an important role in the plant growth and development.

In order to gain a global insight, it is important to analyse the effect of alternative splicing on a large scale to understand the effects of this process on proteome and interactome. For example, Resch and colleagues have performed a comprehensive analysis on Alternative Splicing Database (ASD) to understand how alternative splicing affects protein domains. For this purpose, they have gathered 4422 major and 8962 minor isoforms from ASD and by using the PFAM and SMART domain databases they have identified 554 protein domains which have been modified by alternative splicing. In 92% of the cases (509), protein domains were partially or fully absent in the minor isoforms, while in 8% of the cases alternative splicing introduced new domains into minor isoforms. They have identified 50 different protein domains including some well-characterised interaction domains (like KRAB, Kelch) which have been preferentially removed by alternative splicing more frequently than average. Furthermore, they have shown on a number of selected examples like Kruppel transcription factors and Pbx2 that alternative splicing changes structure of the isoforms mainly by removing protein interaction domains which leads to redirection of protein interaction networks at key points (Resch *et al.*, 2003).

In another large-scale study, Fardilha and colleagues have reported on the high interaction diversity within the human testis protein phosphatase 1 (PP1) interactome (Fardilha *et al.*, 2011). PP1 is a serine/threonine-specific phosphatase, where its different forms form complexes with PP1 interacting proteins and affect functions of cell. PP1 is encoded by 3 different genes namely PP1-alpha, PP1-beta and PP1-gamma. There are two different forms of PP1-gamma: PP1-gamma1 and PP1-gamma2 generated through tissue-specific alternative splicing. PP1-gamma1 is expressed in many different tissues such as heart, brain and liver, while PP1-gamma2 is expressed in testis and is involved in the regulation of spermatozoa motility. The study has targeted to identify the PP1-gamma2 interacting proteins by using several experimental methods (such as yeast two hybrid and co-immunoprecipitation) in human testis. The formed interactome has been reported as the largest human testis PP1 interactome. By using this interactome, it has been shown that there is high diversity among the regulatory protein sets binding to PP1 isoforms in different tissues (77 interacting proteins in testis and 7 proteins in sperm). The reported PP1-binding proteins serve as potential targets for pharmacological interventions.

In short, the examples given above for large-scale analyses as well as the studies focusing on a single gene provide evidence that different isoforms of the same protein potentially are involved in different interaction networks.

Chapter 3

TEXT MINING FOR PROTEIN-PROTEIN INTERACTIONS

In response to the recent developments in the field of biomedicine, large amount of experimental and computational PPI data are gathered and accompanied with an exponential increase in the number of publications describing these findings. Hence, there is a great interest from scientific communities in automatically extracting PPI data from text which holds the promise of easily discovering biological knowledge.

Automated PPI extraction systems reported in the literature are based on a general text mining system architecture which is depicted in Figure 3.1.

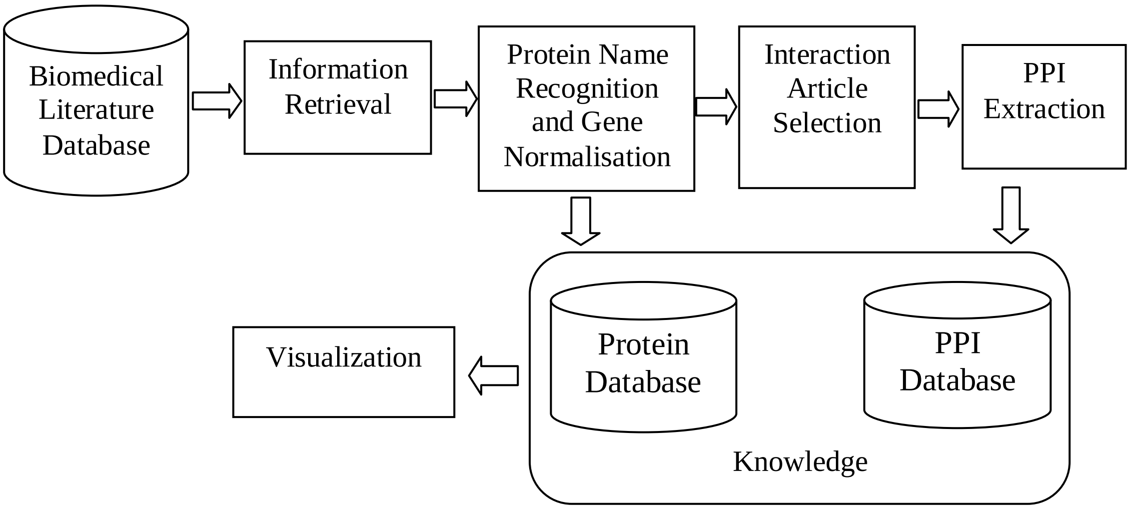

Biomedical Literature Database
Information Retrieval
Protein Name Recognition and Gene Normalisation
Interaction Article Selection
PPI Extraction
Protein Database
PPI Database
Knowledge
Visualization

Figure 3.1: General Text Mining System Architecture for PPI Extraction

The system consists of following main components:

1) Biomedical Literature Databases: Such databases are repositories keeping track the published biomedical text and serving as input sources for all typical text mining based systems. PubMed at NCBI which is the most widely used biomedical literature repository and United Kingdom PubMed Central (UKPMC) (http://ukpmc.ac.uk/) which gives access to full text articles can be listed as typical biomedical text resources.
2) Information Retrieval: This component focuses on searching and gathering relevant documents from large document collections based on user queries.
3) Protein Name Recognition and Gene Normalisation: The aim of this component is to identify protein mentions within text by their names and/or symbols. In some particular cases, the identified protein names are required to be linked to their protein database identifiers with an additional step called as normalisation. Identifying protein mentions (and normalising them) before attempting to extract interactions between them is strictly mandatory for this particular system. Therefore, success of this process in terms of recall and precision plays a crucial role in PPI extraction.
4) Interaction Article Selection: Initial large set of retrieved documents by component (1) can be further processed to sub-select documents which are likely to report on PPIs. Performing this process before PPI extraction is not as crucial as protein name recognition. Nevertheless, it is believed that this process increases success rate of PPI extraction (Krallinger *et al*., 2008).
5) PPI Extraction: PPI extraction is commonly addressed as identifying binary relationship between two proteins. This component remains as the core module in this particular text mining system.

6) Visualization: This component provides an interactive and user friendly way to end-users for utilising with the gathered knowledge. Although integration of this module in to the architecture is not mandatory, visualization helps to disclose the findings to the community working in the field of biomedicine and hence it can be considered to be very a rewarding component.

Following sub-sections provide background in detail on the system components presented above.

3.1 Information Retrieval

Information retrieval (IR) deals with the representation, organisation, storage and access to information items such as documents and web pages (Baeza-Yates and Ribeiro-Neto, 2011). Early IR systems required to submit requests that were run in batches and took hours or even days to return results given that computers had low computational power and capacity at that time (Hersh, 2009). However, advancements in computer technology led to the development of today's IR systems which can deal with exponentially growing massive amount of data. Google (http://www.google.com/) which is probably the leading Web search engine and Entrez, the PubMed IR system at NCBI (http://www.ncbi.nlm.nih.gov/pubmed) for biomedical domain could be listed as most widely used IR systems.

A typical IR system is composed of content, hardware to store the content and a software application enabling to access and retrieve the content based on user queries. There are two main processes in IR. These are indexing and retrieval. In indexing, metadata (i.e. meta-information about the information in the content) is assigned to items contained in the collection in order to retrieve them efficiently.

Content can be searched either based on user-queries or documents. In query-based search, queries are formed by using several search terms like protein or disease name of interest which are often connected by Boolean operators (AND, OR, NOT). Documents which match metadata are retrieved by interacting with the IR system. In the case of document-based search, rather than several keywords, the whole document is provided to the system and the retrieved set is a ranked list of documents similar to the provided one. Google Scholar (http://scholar.google.com/) which is dedicated to academic literature and Entrez IR systems support both types of searches.

During the last couple of decades, several challenges have been organised to evaluate IR systems in a community-wide manner. The largest IR challenge evaluation is the Text REtrieval Conferences (TREC) which started in 1992 (Harman, 1993). TREC is organised as an annual event where the tracks (tasks) are defined and queries as well as documents are provided to the challenge participants. TREC hosted Genomics track from 2003 to 2007 which was one of the largest as well as longest community-wide evaluation in biomedical domain. Several tasks such as ad hoc retrieval, document classification and extraction of GeneRIF were included in this track.

In PPI extraction, IR is utilised as a selection step which enables the retrieval of abstracts including proteins under investigation. Therefore, efficient selection of the input data to be processed by PPI extraction system is essential. Hence, search terms which are used to formulate queries for seeking the content should be carefully selected.

3.2 Protein Name Recognition and Gene Normalisation

Named Entity Recognition (NER) is the first step in many information extraction systems where the aim is to identify pre-defined objects called Named Entities (NE) in text. An NE is a word or group of words that denotes a specific object or group of objects. Such entities are location, person and organization as defined by 6th and 7th Message Understanding Conferences (MUC) in the newswire domain (Grishman and Sundheim, 1996; Kaufmann, 1998) and genes, chemicals, drugs, diseases, etc in the biomedical domain (Collier *et al*., 2001; Wilbur *et al*., 1999). In particular, NEs to be identified are genes/proteins in PPI extraction. This process is a crucial initial step affecting the overall PPI pipeline performance.

Several approaches have been proposed in the literature for protein name recognition. One such approach is the dictionary-based approach. Systems relying on this approach such as Whatizit Swissprot (Rebholz-Schuhmann *et al*., 2008) often exploit one or more terminological resources like Swissprot DB (http://www.expasy.ch/sprot) and Entrez Gene DB (http://www.ncbi.nlm.nih.gov/gene) to identify protein name locations in text. Such systems generally achieve high precision with the expense of recall given that they utilise well defined gene/protein names.

Another approach used to identify protein names is the rule-based. Systems based on this approach, such as YAPEX (Franzen *et al*., 2002) and KEX (Fukuda *et al*., 1998) generally utilise manually formed rules describing common naming structures for specific term groups by using orthographic or lexical evidences. Disadvantage of these systems is the generalisation problem of rules to apply new domains.

In recent years, machine learning (ML) based systems have gained popularity due to their robustness and high accuracy in NER. Such systems like Genia tagger (http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/), ABNER (Settles, 2005) and BioTagger-GM (Torii *et al.*, 2009) rely on a single classifier or a combination of multiple classifiers trained by using an ML algorithm such as Conditional Random Field (CRF), or SVM which employs a feature vector based on the training corpus. The feature vector often includes orthographic (e.g. digit, lowercase), morphological (e.g. prefix, suffix), and lexical features (e.g. part of speech tag). Bunescu and colleagues have shown in a comparative study that machine learning based protein name recognisers perform better than dictionary and rule based recognisers on the AIMed corpus (Bunescu *et al.*, 2005).

A number of systems reported in the literature have used ML approaches in combination with dictionary and/or ruled-based approaches to boost performance in protein name recognition. For example, NLProt (Mika and Rost, 2004) has combined a dictionary and a rule-based filtering module with several SVMs to tag protein names in text. Similarly, BioTagger-TM has utilised terminological information from Biothesarus (Liu *et al.*, 2006) and UMLS (Bodenreider, 2004) using machine learning frameworks and system combination.

Although there are many protein name recognisers proposed in the literature, it is difficult to compare their performances since they have been developed and evaluated based on different corpora including Genia Corpus (Ohta *et al.*, 2002), Yapex (Franzen *et al.*, 2002), BioCreative-I (Hirschman *et al.*, 2005a) and II (Smith *et al.*, 2008) datasets, and JNLPBA (Kim *et al.*, 2004) dataset. Nevertheless, several challenges have been organised to allow community-wide evaluations of the protein name recognisers. The shared task of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-2004) (Kim *et al.*, 2008) was one of such challenges. Others include, the Critical Assessment of Information Extraction systems in Biology (BioCreative) I and II challenges which were organised in 2004 and 2006 respectively (Hirschman *et al.*, 2005a; Smith *et al.*, 2008).

Despite all the efforts regarding protein name recognition, the overall performances of systems are still low compared to performances in the newswire domain. The highest scoring NER system has shown a human-curator comparable performance (F_1-score of 96%) at MUC (Kaufmann, 1998). On the other hand, in the biomedical domain, best performances have been reported as F_1-score values of 72.55% in JNLPBA-2004 (Kim *et al.*, 2008), 83% in BioCreative-I (Yeh *et al.*, 2005) and 87.21% in BioCreative-II (Smith *et al.*, 2008) which are significantly below the performance achieved in the newswire domain. The lower NER performances in biomedical domain can be attributed to several factors such as wide use of abbreviations, synonyms, homonyms, ambiguous names, inconsistent naming conventions, widespread and inconsistent use of white space and special characters such as '+', '-' and '/' (Dimililer *et al.*, 2009). Moreover, new molecular object names are introduced to the domain frequently and some of them are used for only a short time period. Hence, protein name recognition still remains as a challenging task in the biomedical domain and indeed, it is hard to apply in real use cases especially due to two factors: synonymy and ambiguity (Khalid, *et al.*, 2008). Synonymy occurs when one protein name is referred to by several different names. For example, the protein CD95 is also named as FAS and APT1. Ambiguity occurs when the same protein name can refer to more than one protein. For example, a search for the name CD95 in Entrez Gene DB returns 131 matches belonging to 19 different species.

In order to stimulate developments concerning the two aforementioned factors, BioCreative has organised several competitions for gene normalisation (GN) task (Hirschman *et al.*, 2005b; Morgan *et al.*, 2008; Lu *et al.*, 2011). This task involves recognition of the mentioned gene/protein names in text and linking them to database identifiers (IDs). The BioCreative challenges have introduced the first gold standard datasets for the GN task and have allowed implementation of a number of practical gene normalisers.

The focus of BioCreative-I has been on the normalisation of gene names in the Medline abstracts to their corresponding Entrez Gene DB IDs for different model organisms including fly, mouse and yeast. This challenge has attracted 8 participants. Highest performances have been reported as 92% for yeast, 82% for fly, 79% for mouse in F_1-score (Hirschman *et al.*, 2005b). In the BioCreative-II GN task, the genes and proteins have been associated with human only. Therefore, this task was easier compared to the GN task in BioCreative-I. In total, 20 participants have submitted results for the task and the best system has achieved an F_1-score value of 81% (Morgan *et al.*, 2008). In the BioCreative-III GN challenge, participants have been provided with full text articles instead of abstracts without any species information and asked to return a ranked list of gene IDs. These make the task harder than the GN task addressed in the two previous BioCreative challenges. The BioCreative-III GN challenge has attracted 14 different teams and performances have been reported in term of Threshold Average Precision (TAP-k), which is specifically used to measure the retrieval efficiency by taking ranking into consideration. The highest TAP-k scores have been reported as 0.3248 (k=5), 0.3469 (k=10), and 0.3466 (k=20) (Lu *et al*., 2011).

A typical gene normaliser integrates 3 main steps: (1) recognition of gene/protein mentions in the text, (2) gathering a list of candidate gene IDs by mapping the recognised genes to their corresponding DB IDs and (3) disambiguation. Various methods have been proposed in the literature for each of these steps. For (1), several state-of-the-art performing systems such as GNAT (F_1-score of 81.4% on the BioCreative data) (Hakenberg *et al*., 2008) and ProMiner (F_1-score of 80% on the BioCreative-II data) (Fluck *et al*., 2007) have utilised protein name dictionaries only while some of them such as GeNo (F_1-score of 86.4% on the BioCreative-II data) (Wermter *et al*., 2009) have employed both dictionaries as well as ML methods for detecting gene/protein names in text.

For (2) mentioned above, generally, similarity matching methods such as cosine similarity and exact matching have been applied to generate a candidate ID list for each recognised protein/gene name (Fundel *et al*., 2007; Hakenberg *et al*., 2008).

For (3) mentioned above, various similarity scores have been introduced or adopted from the existing solutions to eliminate false positive IDs (Dai *et al*., 2010; Fundel *et al*., 2007; Hakenberg *et al*., 2008, Wermter *et al*, 2009) given that multiple IDs can be gathered for a single gene. For example, in GNAT (Hakenberg *et al*., 2008) and in the system described by (Dai *et al*., 2010) external knowledge for each tagged gene, such as GO terms chromosome locations and alike, have been collected to calculate the likelihoods representing the similarity of the identified text with the knowledge to tackle the disambiguation problem.

Similarly, factors complicating the protein name recognition task play a role in gene normalisation. Especially ambiguity of gene/protein names with the common English words and with other organisms keeps this task challenging and leaves some room for performance improvement.

3.3 Interaction Article Selection

Interaction article detection can be considered as a binary text classification problem where the positive and negative classes correspond to relevant and irrelevant documents respectively. The aim of this step is to reduce the initial large number of retrieved documents to a manageable size by selecting those documents which are most likely to contain PPI.

Although this process is important for both reducing the workload significantly and increasing the success rate of PPI extraction, it has been neglected by many early PPI extraction systems (Krallinger *et al.*, 2008). Nevertheless, some studies carried out by (Marcotte *et al.*, 2001) and (Donaldson *et al.*, 2003) have reported the first such protein interaction document filtering systems by using Bayesian and machine learning approaches respectively. Recently, this task has been intensively studied in the BioCreative challenges enabling community-wide evaluations. In BioCreative-II, the task has been addressed as Interaction Article Subtask (IAS) revealing the first gold standard dataset in the domain (Krallinger *et al.*, 2008). This challenge has attracted 19 participants. Most successful systems have utilised machine learning approaches and adopted one or more forms of term weighting schemes from the standard text classification problem (Lan *et al.*, 2007, Abi-Haidar *et al.*, 2007) while some have used domain dependent features such as the number of protein mentions in text (Abi-Haidar *et al.*, 2007; Ehrler *et al.*, 2007) and mentioned named entities, and protein interaction verbs (Ehrler *et al.*, 2007). The highest F_1-score achieved in BioCreative-II IAS is 78% (Lan *et al.*, 2007). Their approach relied on an SVM trained using Bag-of-Words (BOW) features in combination with protein name entities. Lan and colleagues have improved their system after the challenge to an F_1-score of 80.25% by combining several features including BOW, trigger words describing PPIs and protein name entities (Lan *et al.*, 2009). Tsai and colleagues have reported an F_1-score 2.90% higher than that top-ranking system in BioCreative-II IAS (Tsai *et al.*, 2008). In their study, they have employed SVMs and exploited likely positive and unlabeled data to improve the classification performance. Another high performing system has been developed by (Wang *et al.*, 2008) who used the Adaboost (Freund and Schapire, 1997) method for feature combination and utilised the SVMs achieving an F_1-score of 84.38% on the BioCreative-II IAS dataset.

More recently, the task has been highlighted as Article Classification Task (ACT) in the BioCreative-II.5 and BioCreative-III challenges. The BioCreative-II.5 ACT was more challenging compared to the article selection tasks in BioCreative-II and III given that full texts have been provided instead of abstracts (Lietner *et al.*, 2010). In addition, in the BioCreative-II.5 and III challenges, participants have been asked to order articles by their likelihood to contain PPIs, in principle having the true hits in the top ranks. The evaluation scheme to be used had to make convenience for measuring performances of the systems by taking into account the produced ranked list of results that would match the gold standard. Therefore, an evaluation scheme called area under the interpolated precision/recall-curve (AUC iP/R) has been selected (Bradley, 1997). This scheme ideally measures precision and recall with respect to the ranked list of results generated by the systems.

In the BioCreative-II.5 challenge, the participants have been provided with training and test datasets each one containing 595 full text FEBS letters articles. 61 and 63 of these sets have been labelled as PPI relevant articles in the test and training sets respectively. This challenge has attracted 8 participants. Similar to the Biocreative-II IAS challenge, the best performing systems have utilised methods from machine learning and the best performance has been reported as an AUC iP/R value of 67.8% (Leitner *et al.*, 2010).

The BioCreative-III ACT challenge has attracted 10 participants. They have been provided with a balanced training set consisting of 2280 Medline abstracts, a development and a test set consisting of 4000 and 6000 abstracts respectively. 15% of these sets have been identified as PPI relevant abstracts. In this challenge, participants have relied on generally statistical methods like chi-square, mutual information and frequency cut-off for feature selection or term weighting. Some of them have also used dimensionality reduction methods on top of their features (Agarwal *et al.*, 2010; Doğan *et al.*, 2010; Fontaine *et al.*, 2010; Lourenco *et al.*, 2010). Half of the participants have used, either SVM or SVM in combination with another supervised machine learning approach for selecting interaction abstracts (Lourenco *et al.*, 2010; Agarwal *et al*, 2010; Wang *et al.*, 2010; Doğan *et al.*, 2010). For example, (Agarwal *et al.*, 2010) has used Naive Bayes and (Doğan *et al.*, 2010) has used Nearest Neighbour. The best performing system has used a Huber classifier (Zhang 2004) and achieved an F_1-score value of 61.42% with 67.98% AUC iP/R (Kim and Wilbur 2010). In this study, they have utilised grammatical relations extracted by C&C parser (Curran *et al.*, 2007) which indicates dependency relations between words to design their classifier. In addition to grammatical relations, they have used MeSH terms, unigrams, bigrams, trigrams, gene tagging (generated from the grammatical relations by replacing protein names with a special tag) and meta features (automatically induced meta bigram features).

3.4 Protein-Protein Interaction Extraction

PPI extraction is widely studied in the literature given that PPIs are crucial in analysing the cellular processes including signalling, regulation and metabolism at system biology level. Several approaches have been proposed to extract PPIs from the biomedical literature. Early studies have used the co-occurrence statistics of proteins (Shatkay *et al*., 2000; Stapley *et al*., 2000) and pre-defined patterns (templates) expressing rules (Blaschke *et al*., 1999; Hatzivassiloglou and Weng, 2002; Ono *et al*., 2001; Wong 2001) for PPI extraction. For example Ono and colleagues have used protein mentions, POS-tags of tokens constituting the sentences and a set of interaction keywords to form the patterns (Ono *et al*., 2001). Their system have been reported to achieve >80% recall and >90% precision.

Several studies have utilised parsers to parse sentences either fully (Friedman *et al*., 2001) or partially (shallow) (Thomas *et al*., 2000) to form more complicated templates with syntactic and semantic constraints. Shallow-parsing based systems decompose the sentences partially, identify certain phrasal components and extract local dependencies between them. (Yang *et al*., 2009) has employed a link grammar to analyse syntactic roles within sentences while (Fundel *et al*., 2007) has utilised a dependency parser for the PPI extraction which has been estimated to achieve 80% precision and 80% recall. Some rule-based PPI extraction systems have adopted dynamic programming techniques to discover patterns automatically and handle complex cases such as the system reported in (Huang *et al*., 2004). This system has achieved about 80.0% recall and about 80.5% precision. Co-occurrence and rule-based methods are known to have some drawbacks. Typically, co-occurrence based systems achieve high recall at the expense of low precision. The drawback associated with rule-based systems is the discovery of new patterns and applicability of rules to other data since usually they are generated by using a single training dataset.

In recent years, several publicly available PPI corpora such as AIMed (Bunescu *et al.*, 2005), BioInfer (Pyysalo *et al.*, 2007) and HPRD50 (Fundel *et al.*, 2007), IEPA (Ding *et al.*, 2002) and LLL (Nedellec 2005) as well as BioCreative datasets (Krallinger *et al.*, 2008; Leitner *et al.*, 2010) have been developed. These corpora make it convenient to implement and evaluate ML based PPI extraction systems. ML approaches such as SVM (Miwa *et al.*, 2009a; Miyao *et al.*, 2008; Yang *et al.*, 2010a), maximum entropy (Sun *et al.*, 2007) and bayesian network (Chowdhary *et al.*, 2009) have been extensively used to develop PPI extraction systems. Such systems can extract PPIs by learning rules automatically from a corpus based on a feature set as opposed to rule-based method that needs domain expert aid to define a set of rules. The feature set often includes, standard bag-of-word (Landeghem *et al.*, 2008; Mutsumori *et al.*, 2006), POS-tag and orthographic (Giuliano *et al.*, 2006) and syntactic features obtained through high-quality domain specific dependency and deep parsers such as Charniak-Lease (Andrew *et al.*, 2007), Ksdep (Sagae and Tsujii, 2007) and Enju (Miyao and Tsujii, 2008). In many recent studies, such parsers have been extensively used to extract different features based on the syntactic and semantic relations between words. (Airola *et al.*, 2008; Erkan *et al.*, 2007; Miwa *et al.*, 2009a; Miwa *et al.*, 2009b; Miyao *et al.*, 2008; Sætre *et al.*, 2007; Yang *et al.*, 2011). Such studies have proven that use of the syntactic features and combining different kernels (e.g. tree, linear) can boost performance in PPI extraction. For example, (Sætre *et al.*; 2007) has used an SVM with a tree kernel for syntactic shortest path features and a linear kernel for context features related to words between, before and after the target protein pair. This system has achieved an F_1-score of 37.80% on the AIMed corpus when only the linear kernel utilising BOW features is used. The performance has been increased to 52% when all the kernels are combined. Alternatively, (Airola *et al.*, 2008) has proposed an all-path graph kernel. In this approach, a given sentence is represented as a dependency graph and dependencies connecting two entities outside the shortest path as well as on the shortest path are considered. This method has been reported to achieve an F_1-score of 56.40% on AIMed. Miwa and colleagues have combined all the lexical and syntactic parsing features using multiple kernels to alleviate the limitations of each feature (Miwa *et al.*, 2008; Miwa *et al.*, 2009a; Miwa *et al.*, 2009b). Their systems have achieved at the state-of-the-art level on different benchmark data sets including AIMed (>60% F_1-score). (Yang *et al.*, 2011) has proposed to use weighted linear combination of the individual kernels proposed by Miwa and colleagues instead of assigning the same weight to each one. The system has achieved at the-state-of-the-art performance on different benchmark corpora (64.41% F_1-score on AIMed). In (Yang *et al.*, 2010b), the effect of different kernel combination strategies has been investigated. It has been reported that using ranking SVM for combining different kernels achieves the best performance among the methods used (64.88% F_1-score).

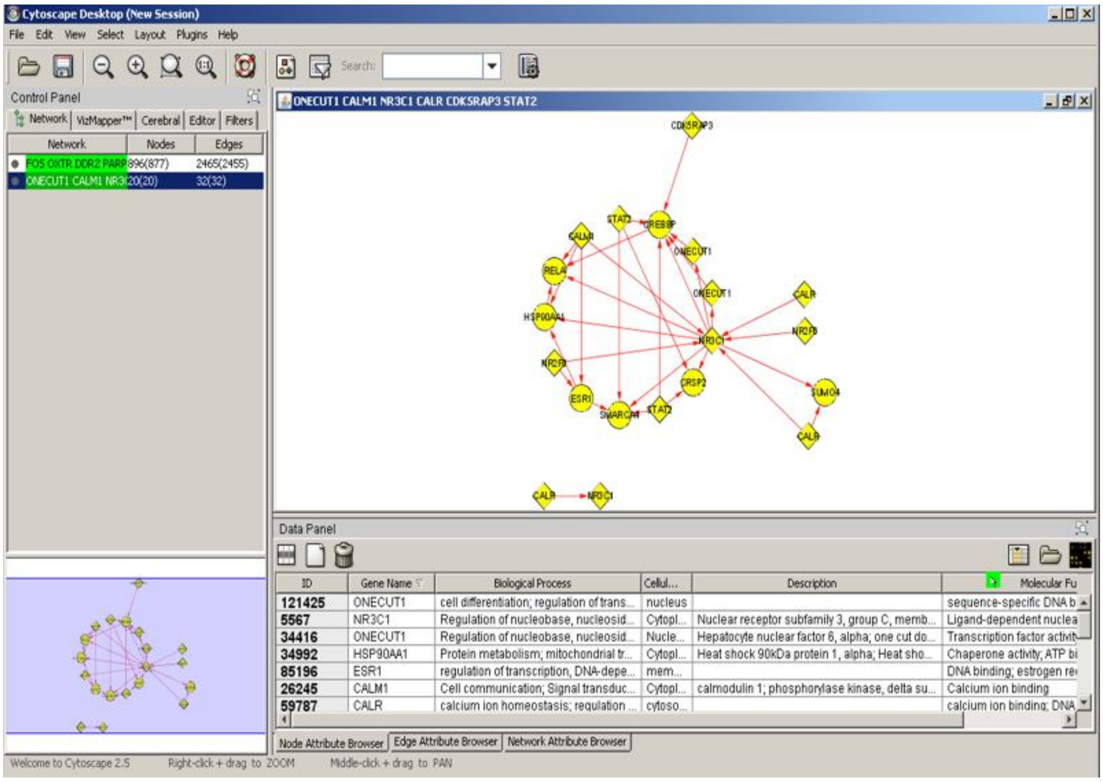
The benchmark datasets available for the PPI extraction task are small. Therefore, the supervised machine learning methods tend to suffer from the data sparseness problems given that they attempt to obtain knowledge from a limited amount of labelled data (Miwa *et al.*, 200b). Therefore, Li and colleagues have proposed to use unlabeled biomedical texts to enhance the performance of supervised learning for PPI extraction (Li *et al.*, 2010). Their semi-supervised learning algorithm trained by using local lexical features such as words and n-grams surrounding the protein pair of interest has achieved an F_1-score of 63.60% on AIMed.

Apart from the proposed methods, the impacts of individual parsers, features as well as kernels on the PPI extraction have been investigated. (Miyao *et al.* 2008) has shown that the accuracy of syntactic parsers play a role in the overall performance of the PPI systems. (Landeghem *et al*., 2008) has analysed the effect of syntactic and lexical features on different publicly available PPI datasets. Similarly, (Niu *et al*., 2010) has analysed various features including, lexical, interaction keyword, dependency, pattern and phrase. The effect of various kernels such as, tree, graph on PPI extraction has been analysed in (Tikk *et al*., 2010). In this study, the methods have been evaluated on different PPI datasets using cross-validation, cross-learning and cross-corpus evaluation. Their study has shown that the kernels utilising dependency trees generally perform better than kernels based on syntactic trees.

A wide range of results have been reported in the literature for the PPI extraction systems. Unfortunately, direct comparison of the systems is difficult due to differences in evaluation resources, metrics and strategies used to develop these systems (Aiorola *et al*., 2008; Sætre *et al*., 2007). While some systems have been reported to achieve 86-95% recall and precision (Ono *et al*., 2001), in the recent BioCreative-II and II.5 challenges the best systems have been reported to achieve 29% (Krallinger *et al*., 2008) and 30% (Leither *et al*., 2010) F_1-score. On the other hand, the reported results on the AIMed corpus are ranging from 33% (Yakushiji *et al*., 2005) to 65% (Miwa *et al*., 2009b) F_1-score. This difference is mainly due to the fact that in the BioCreative PPI tasks, gene normalisation is not separated from PPI identification. However, these results suggest that the PPI extraction problem is far from solved.

3.5 Protein-Protein Interaction Networks

In an organism, proteins systematically interact with each other creating dynamic interaction networks for regulating biological activities of cells. Hence, to fully observe the functional organization of the proteome efforts are directed to establish graph representations called protein-protein interaction networks (PPINs) (Cho *et al.*, 2004). In a PPIN, vertices represent proteins and edges represent protein interactions. Often, PPIN data is collected and stored in public databases, such as DIP (Xenarios *et al.*, 2000) and HPRD (Keshava *et al.*, 2009). Software platforms like Cytoscape (Shannon *et al.*, 2003) which specialised for graph representations of interactions are used to generate and analyse PPINs. Figure 3.1 is a screen shot representing the usage of Cytoscape for analysing a PPIN.



Network	Nodes	Edges
FOS OXTR DDR2 PARP	896(877)	2465(2455)
ONECUT1 CALM1 NR3(	20(20)	32(32)

ID	Gene Name	Biological Process	Cellul...	Description	Molecular Fu
121425	ONECUT1	cell differentiation; regulation of trans...	nucleus		sequence-specific DNA b
5567	NR3C1	Regulation of nucleobase, nucleosid...	Cytopl...	Nuclear receptor subfamily 3, group C, memb...	Ligand-dependent nuclea
34416	ONECUT1	Regulation of nucleobase, nucleosid...	Nucle...	Hepatocyte nuclear factor 6, alpha; one cut do...	Transcription factor activit
34992	HSP90AA1	Protein metabolism; mitochondrial tr...	Cytopl...	Heat shock 90kDa protein 1, alpha; Heat sho...	Chaperone activity; ATP bi
85196	ESR1	regulation of transcription, DNA-depe...	mem...		DNA binding; estrogen re
26245	CALM1	Cell communication; Signal transduc...	Cytopl...	calmodulin 1; phosphorylase kinase, delta su...	Calcium ion binding
59787	CALR	calcium ion homeostasis; regulation ...	cytoso...		calcium ion binding; DNA

Welcome to Cytoscape 2.5 Right-click + drag to ZOOM Middle-click + drag to PAN

Figure 3.2: Cytoscape Screenshot Showing Analysis of a PPIN

(Source: http://www.ncibi.org/gateway/mimiplugin.html)

Different methods including experimental, computational and automated literature mining have been used to generate PPINs for different organisms such as H. *pylori* (Rain *et al*., 2001), S. *cerevisiae* (Ito *et al*., 2001; Uetz *et al*., 2000), C. elegans (Li *et al*., 2004), D. *melanogaster* (Giot *et al*., 2001), A. *thaliana* (Li *et al*., 2011), mus musculus (Li *et al*., 2010) and Homo sapiens (Stelzl *et al*., 2005). Such PPINs are being utilised to understand evolution (Stumpf *et al*., 2007), mechanisms of diseases (Chen *et al*., 2009; Goñi *et al*., 2008; Hwang *et al*., 2008; Ideker *et al*., 2008; Zanzoni *et al*., 2009;) such as cancer (Gong *et al*., 2010) and Alzheimer's disease (Ofran *et al*., 2006) and drug targets (Ruffner *et al*., 2007).

There are several PPINs utilising text mining or text mining in combination with experimental methods in the literature. One such system is STRING (Szklarczyk *et* *al.*, 2011) which covers PPI data currently for more than 1100 different organisms. The PPI data stored in STRING is gathered based on experimental as well as text mining methods. Simple co-occurrence based approach is used to extract the PPIs from the scientific literature.

The Information Hyperlinked Over Proteins (iHOP) (Hoffman and Valencia, 2005) is the first open-access, large-scale biological literature navigation system for PPIs containing 23.4 million sentences and 110,000 different genes from more than 2,800 organisms. Sentences describing PPIs are extracted based on the tri-occurrence method which requires co-occurrence of two protein mentions and a verb describing protein interactions in the same sentence.

Info-Pubmed runs on PubMed database of NCBI aiding scientists to find protein-protein and gene-diseases associations which are extracted based on deep syntactic analysis of sentence structure (Ninomiya *et al.*, 2007).

CoPub (Fleuren *et al.*, 2011) is a text mining based system that detects co-occurring biomedical concepts in the Medline abstracts. The biomedical concepts covered by CoPub are human, mouse and rat genes, biological processes, molecular functions and cellular components from GO, liver pathologies, diseases, drugs and pathways. The retrieved relations between terms can be visualised using the Cytoscape web plug-in (Shannon *et al.*, 2003).

Mouse PPIN which runs on Mouse protein-protein interaction DataBase (MppDB) (Li *et al.*, 2010) can be given as an example for organism specific PPINs. MppDB contains PPI data from 6 different publicly available databases: HPRD (Keshava *et* *al*., 2009), IntAct (Hermjakob *et al*., 2004), the Biomolecular Interaction Database (BIND) (Bader *et al*., 2001), DIP (Xenarios *et al*., 2000), The MIPS Mammalian Protein-Protein Interaction Database (MIPS) (Pagel *et al*., 2005), MINT (Zanzoni *et al*., 2002) and also from the scientific literature. In order to gather the PPIs, first sentences containing co-occurrences of proteins are collected from the Medline abstracts. Second, a naïve Bayesian model is used on top of the sentences to filter false-positive interactions. Third, a SVM algorithm is further used to select protein pairs with physical interactions. Current version of the database includes more than 5,000 and 10,000 interactions of mouse proteins gathered from the reference PPI datasets and the literature respectively.

Another organism specific PPIN runs on the AtPID database (Li *et al*., 2011) which focuses on interactions between the Arabidopsis proteins. The PPI data covered by AtPID is collected through experiments and expanded by automated text mining methods. Currently, the database contains approximately 13,000 protein interactions.

Despite the efforts in generating extensive PPINs, for the most of the organisms, the PPI data gathered is far from complete. For example, the human interactome size is estimated as ~650,000 (Stumpf *et al*., 2008). However, PINA which is one of the most comprehensive PPI resources currently reports only 85,053 interactions for human. As the experimental and biomedical text mining methods improve, the PPINs will be enlarged in the future and will better assist research in the domain of biomedicine.

Chapter 4

MINING THE SCIENTIFIC LITERATURE FOR ISOFORM INTERACTIONS

4.1 Dataset Used - a Human Alternative Splicing Database

The transcript data coding protein isoforms are gathered from HumanSDB3. This database consists of clusters each of which containing a set of overlapping transcripts. Transcripts of a particular cluster are grouped according to their sequence similarities and are mapped to the same genomic region (locus). Methods developed to construct this database are described in (Taneri *et al.*, 2004, Taneri *et al.*, 2005). Briefly, a total of 4,635,471 transcript sequences are collected from UniGene human clusters (version no. 173) and are aligned to the genome (UCSC hg17) sequence by using blat (Kent, 2002). Transcripts are either full-length mRNAs or EST (Expressed Sequence Tag) sequences. Top 10% of the mapped transcripts are aligned to the genomic region by using SIM4 (Florea *et al.*, 1998). Top scoring matches are further screened for having at least 75% identity to the genome and containing at least two exons each of which having 95% or greater identity to the genome. Clusters having at least three transcripts are retained otherwise were discarded. Consequently, the database contains a total number of 1,459,966 transcripts from 20,707 different clusters. Each cluster has 70.5 transcripts on the average. 3,881 (18.69%) out of 20, 707 clusters are invariant while remaining 16,826 (81.31%) clusters are variant (Taneri, 2005). A variant cluster contains transcripts exhibiting alternative splicing.

On the other hand, an invariant cluster does not exhibit any alternative splicing (Taneri, 2005). Therefore, invariant clusters were excluded from this study. Table 4.1 shows two variant clusters of HumanSDB3 with a subset of their transcripts. The cluster with identifier (ID) Hs.3.chr6n.16927 contains alternatively spliced transcripts of gamma-aminobutyric acid (GABA) receptor, rho 1. NM_002042 and M62400 are GenBank IDs for two of the several overlapping transcripts clustered together in Hs.3.chr6n.16927. They are two different alternatively spliced transcripts of this gene. Cluster Hs.3.chr19n.10387 contains alternatively spliced transcripts of fibrillin 3. HumanSDB3 clusters are accessible through Scripps Genome Center web interface (http://emmy.ucsd.edu/).

Table 4.1 Two Variant Clusters of HumanSDB3 with Subsets of Their Transcripts

Cluster ID	GenBank Ids
Hs.3.chr6n.16927	NM_002042 M62400 CD672849 AW949752 AW949742
Hs.3.chr19n.10387	CD634027 CD634046 BI523489 NM_032447

4.2 Overview of the Text Mining Pipeline

The text mining pipeline developed to build TBIID is depicted in Figure 4.1. The pipeline consists of cluster based screening, abstract retrieval, abstract selection and PPI extraction units. The following sub-sections provide details on the units.

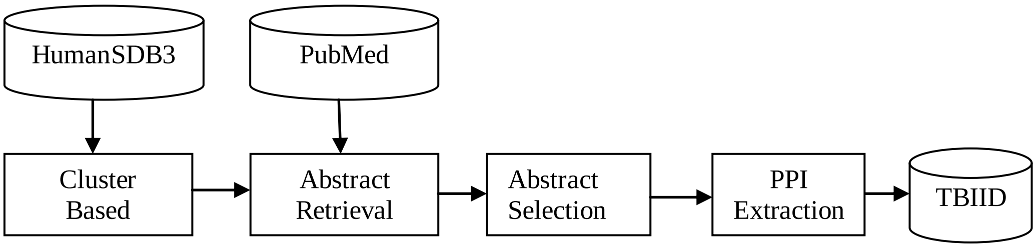


Figure 4.1: Text Mining Pipeline

4.3 Cluster Based Screening

4.3.1 Definitions

This section describes the definitions that are introduced and used during the screening process of the HumanSDB3 data.

Defined Transcript (DT): It is a transcript which is annotated in the Entrez Gene DB with at least one official symbol and name.

Following categorisation is used to distinguish the HumanSDB3 variant clusters according to the number of DTs in them:

(a) Cluster with Undefined Transcripts (CUT): Such clusters do not contain any DT.

(b) Cluster with a Single defined Transcript (CST): Such clusters contain only a single DT.

(c) Cluster with Multiple defined Transcripts (CMT): Such clusters contain multiple DTs.

4.3.2 Methods for Screening the Transcript Data

During the cluster-based screening depicted in Figure 4.2, each variant cluster from HumanSDB3 is screened for its annotated transcripts with its official symbol, name, aliases and designations in the Entrez Gene DB of NCBI (http://www.ncbi.nlm.nih.gov/gene). Annotated transcripts are termed in this work as *Defined Transcripts* (DTs). A DT has at least one official symbol and name from

Entrez Gene DB. Once the DTs are identified, each cluster is classified according to the number of DTs in them. A cluster which does not contain any DT is classified as *Cluster with Undefined Transcripts* (CUT). A cluster containing only a single DT is classified as *Cluster with a Single defined Transcript* (CST). A cluster containing multiple DTs is classified as *Cluster with Multiple Defined Transcripts* (CMT). Indeed, a very stringent procedure is applied to construct HumanSDB3. Transcripts of a given cluster are mapped to the same locus with high sequence similarities. Nevertheless, it is possible that DTs linked to the same CMT could also represent other kinds of isoforms, i.e. products of allelic or duplicated genes. Table 4.2 shows a sample CMT (cluster ID Hs.3.chr14p.5840) and a CST (cluster id Hs.3.chr15p.6725). The DTs with GeneBank IDs NM_000624 and CR601472 linked to the given CMT denote two different serpin isoforms, SERPINA5 (Entrez Gene ID:5104) and SERPINA3 (Entrez Gene ID:12), respectively. Analysis of this particular CMT based on the literature reveals that its DTs denote isoforms encoded by two different serpine genes located on the human chromosome 14q32 (Illingsley *et al.*, 1993). Previous studies have shown that these genes have a certain structural similarity and can be clustered together in the same serpin gene cluster. These evidences indicate that they evolved through gene duplication (Illingsley *et al.*, 1993; Pelissier *et al.*, 2008; Rollini and Fournier, 1997). The CST contains a single DT with Gene Bank ID X04665 encoding THBS1 (Entrez Gene ID:7057) protein.

Table 4.2: Example CST and CMT Clusters

Cluster Type	Cluster ID	Transcript ID	Gene ID	Official Symbol	Official Name
CMT	Hs.3.chr14p.5840	NM_000624	5104	SERPINA5	serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 5
CMT	Hs.3.chr14p.5840	CR601472	12	SERPINA3	serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 3
CST	Hs.3.chr15p.6725	X04665	7057	THBS1	thrombospondin 1

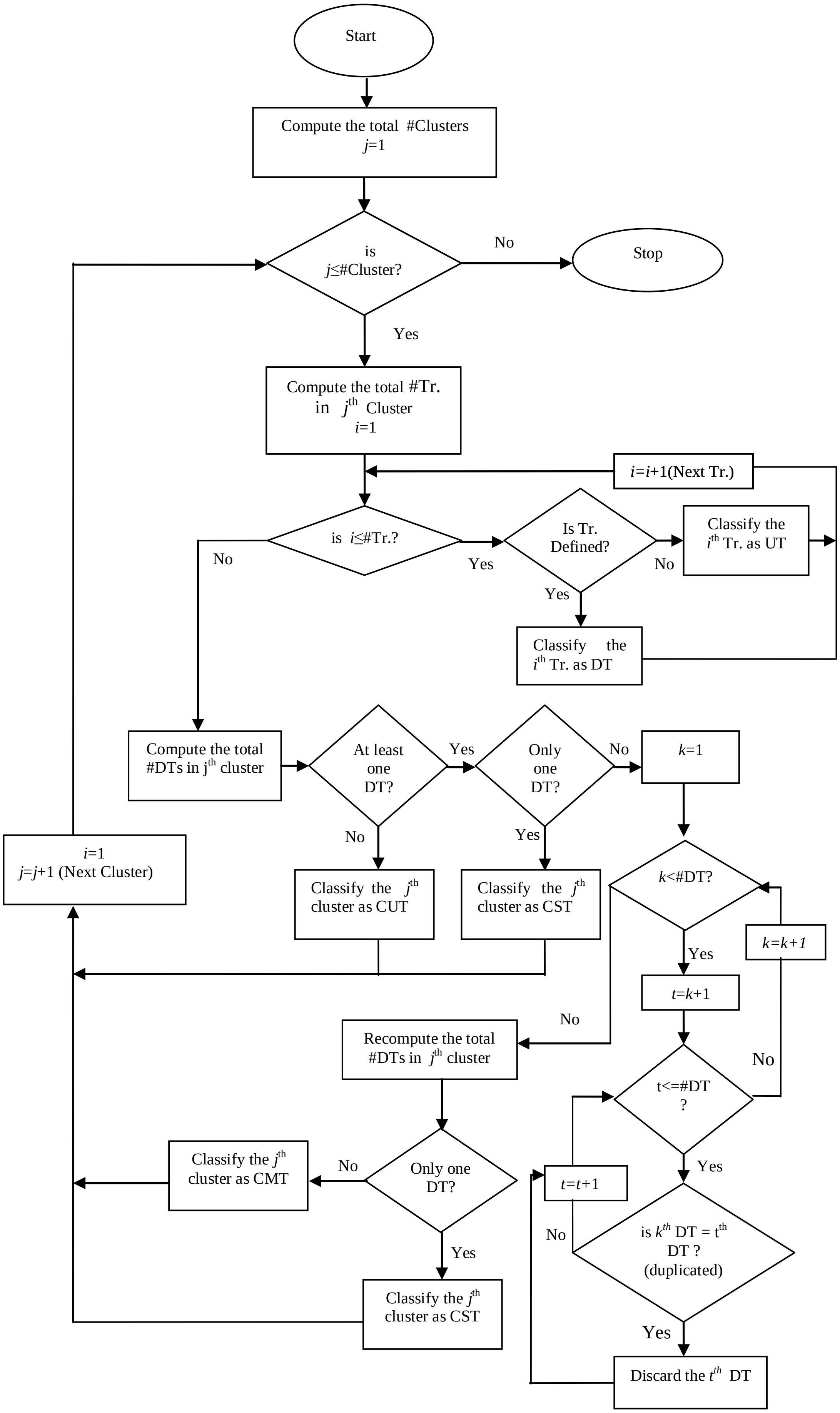


Figure 4.2: Cluster Based Screening

4.3.3 Analysis of Variant Clusters

A total number of 16,826 variant clusters from HumanSDB3 are analysed in the cluster-based screening phase. Results documented in Table 4.3 reveal that majority of the clusters corresponding to 72.50% (12,192) contain only one DT (i.e. they are CSTs). 21.21% (3,568) of clusters are identified as CUTs. Such clusters correspond to empty clusters since no annotation could be found in the Entrez Gene DB for any of their transcripts. In addition, 3.68% (620) clusters are identified as overlapping clusters. An overlapping cluster contains at least one DT which shares the same annotation with another DT belonging to a different cluster. CUTs and overlapping clusters are discarded from the study since they are not relevant for this work. A small portion corresponding to 2.65% (446) of the variant clusters are identified as CMTs. Analysis reveal that there are a total of 13,174 DTs contained in 12,638 CST and CMT clusters.

Table 4.3: Overview of the Distribution of HumanSDB3 Variant Clusters

Total	Variant Clusters	CUT	Overlapping Clusters	CST	CMT	CST+CMT
Numeric	16,826	3,568	620	12,192	446	12,638
Percentage (%)	100	21.21	3.68	72.46	2.65	75.11

4.4 Abstract Retrieval

All DTs from CSTs and CMTs are used during the abstract retrieval phase. In order to retrieve the relevant abstracts, first a rich Search Term Set (STS) is formed for each DT. Then, STSs are used to search and retrieve the abstracts relevant to the DTs from PubMed. For this purpose, the Entrez Programming Utilities (eUtils) toolkit which provides remote access to the NCBI's infrastructure is facilitated (Bathesda, 2010).

4.4.1 Methods Used for Retrieving Abstracts that Belong to Isoforms

4.4.1.1 Search Term Set Formation

Figure 4.3 illustrates the STS generation process. Each DT is screened in Entrez Gene DB of NCBI for its official symbol, name, aliases (other symbols) and designations (other names) by using Esearch utility of the eUtils toolkit. Relevant search fields for each transcript are gathered by using its GeneBank (transcript) ID as the search term. The general form of a query is as follows:

http://www.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=gene&term=GenBankID.

In order to increase recall, each DT is also screened in the Swissprot DB, with additional symbols and names. Further STS expansion is achieved by using synonym generation. Missing nomenclature rules in protein naming complicates the abstract retrieval process. Often, there are more than one representation forms for a single protein. For example; “OMA-1”, “OMA1” and “OMA 1” are all synonyms which can be used for the same protein. The Entrez abstract retrieval system automatically expands search term with a limited extent. For example; search for either “OMA-1” or “OMA 1” results in same set of abstracts. However, abstracts belonging to “OMA1” are missed. Therefore, in order to have a complete STS, “OMA1” is generated, where it includes a symbol like “OMA 1” or “OMA-1” in the synonym generation process. Similarly, “OMA 1” is generated, where the STS includes “OMA1”.

Another factor complicating the abstract retrieval is usage of common English words (e.g. “NOT”, “CELL”, “END”, “FISH”, “AGE”, “AIM”) and single letters (e.g. “P” and “H”) as symbols for proteins. Usage of such symbols results in retrieval of a very large set of abstracts, where many of them are expected to be irrelevant to the protein (i.e. false positive). For example, around 4,300,000 and 3,300,000 abstracts are retrieved from PubMed, when “CELL” and “P” are used as search terms, respectively. Hence, all English like words and single letters are removed from each final STS.

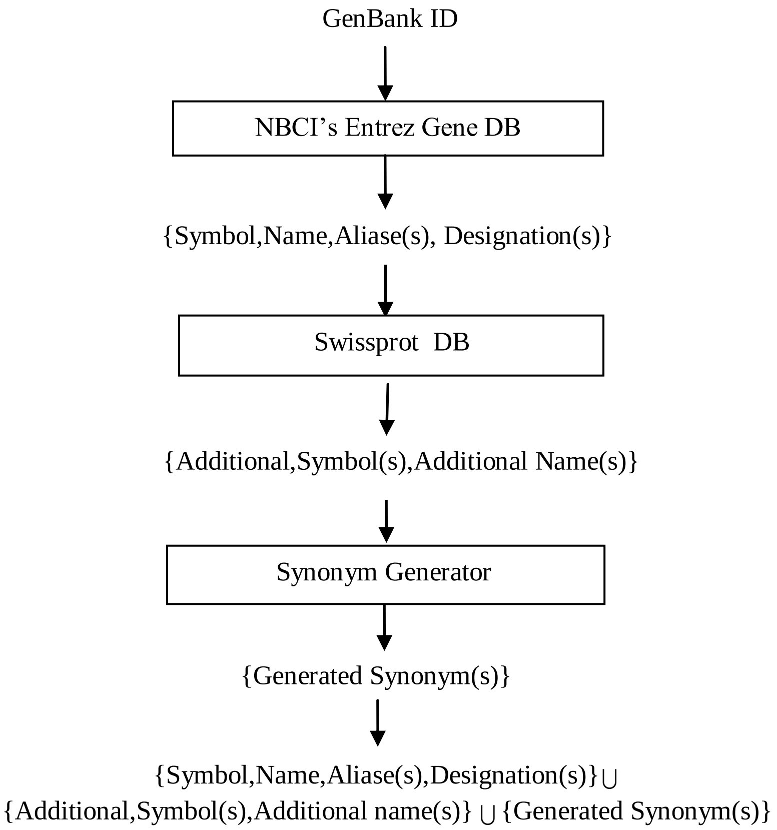

GenBank ID
NBCI’s Entrez Gene DB
{Symbol,Name,Aliase(s), Designation(s)}
Swissprot DB
{Additional,Symbol(s),Additional Name(s)}
Synonym Generator
{Generated Synonym(s)}
{Symbol,Name,Aliase(s),Designation(s)} ∪
{Additional,Symbol(s),Additional name(s)} ∪ {Generated Synonym(s)}

Figure 4.3: STS Generation Process

4.4.1.2 Retrieving the Relevant Abstracts from PubMed

In this phase, firstly relevant abstracts for each DT are searched in PubMed by using Esearch utility of the eUtils toolkit. For this purpose, a query is formed for each DT based on its STS. The search is restricted with documents in English language and the human organism. Following is a sample search query for human isoform CDH5 having a STS containing its symbol and name only:

http://www.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=PubMed&retmax=1&usehistory=y&term=(("CDH5"[Text Word]) OR ("CADHERIN 5, TYPE 2"[Text Word])) AND English [Lang] AND "humans"[MesH Terms]

Two parameters are returned as search result: QueryKey and WebEnv. Relevant abstracts of the isoform are retrieved by facilitating these parameters and Efetch utility of the eUtils. General form of a query used to retrieve abstracts is as follows:

http://www.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?retype=abstract&retmode=xml&retstart=$retstart&retmax=20db=PubMed&query_key=$QueryKey&WebEnv=$WebEnv

4.4.2 Results and Discussion on Abstract Retrieval

4.4.2.1 Improvement on Search Term Set Size

Total numbers of search terms gathered for different STSs is documented in Table 4.4. There are a total of 25,001 search terms corresponding to 1.90 search terms/DT on the average, when only official symbols and names are used. This number increases to 88,371, when aliases and designations from the Entrez Gene DB are added to the official symbols and names (6.72 search terms/DT). The search space is expanded with 26,220 additional search terms obtained from the Swissprot DB. Hence, a total of 117,852 search terms are gathered from both repositories, corresponding to 8.79 search terms/DT on the average. The search space is further expanded with additional 56,637 generated synonyms (4.30 search terms/DT). Consequently, a total number of 171,238 search terms corresponding to 13.00 search terms/DT on the average are used to retrieve the relevant abstracts. Hence, the final

STS is improved by a factor of 6.84 compared to the initial STS containing official symbols and names only.

Table 4.4: Number of Search Terms Corresponding to STSs

Search Term Set	Number of Search Terms	Number of Search Terms/DT
Symbols + Names (Entrez Gene DB)	25,001	1.90
Symbols + Names + Aliases + Designations (Entrez Gene DB)	88,381	6.71
Symbols + Definitions (Swissprot DB)	26,220	1.99
Total of Entrez Gene DB and Swissprot DB	114,601	8.70
Generated Synonyms	56,637	4.30
Total of Entrez Gene DB, Swissprot DB and Generated Synonyms	171,238	13.00

4.4.2.2 Effect of Search Term Set Expansion on Abstract Retrieval

The effect of STS expansion on abstract retrieval is analysed and depicted in Table 4.5. A total of 1,040,783 abstracts (79.00 abstracts/DT) are retrieved by using only symbol and name search fields from the Entrez Gene DB. This number dramatically increase to 4,002,003 (303.78 abstracts/DT) when all search fields from the Entrez Gene DB are used. 102,990 (7.81 abstracts/DT) additional abstracts are retrieved by using search terms gathered from the Swissprot DB. When search terms from both the Entrez Gene DB and the Swissprot DB are utilised, a total number of 4,104,993 abstracts are retrieved (311.60 abstracts/DT). Furthermore, 82,868 additional abstracts are retrieved by facilitating the synonym generation. Consequently, in total 4,187,861 relevant abstracts corresponding to 317.89 abstracts per DT on the average are retrieved for all DTs. The improvement is measured as a factor of 4.02 compared to the initial number of abstracts retrieved using STSs including official symbols and names only. Analysis results reveal that expansion of the STS leads to a significant increase in the total number of abstracts retrieved.

Table 4.5: Average Number of Abstracts Retrieved Corresponding to Search Term Sets

Search Term Set	**Number of Abstracts retrieved**	**Avg. Number of Abstracts retrieved/DT**
Symbols + Names (Entrez Gene DB)	1,040,783	79.00
Symbols + Names + Aliases + Designations (Entrez Gene DB)	4,002,003	303.78
Symbols + Definitions (Swissprot DB)	102,990	7.81
Total of Entrez Gene DB and Swissprot DB	4,104,993	311.60
Generated Synonyms	82,868	6.29
Total of Entrez Gene DB, Swissprot DB and Generated Synonyms	4,187,861	317.89

4.5 Abstract Selection

Identification of abstracts which are likely to contain PPI information from the retrieved abstract set is an important prior process which decreases the workload as well as affects the quality of the PPI extraction task. Therefore, an Interaction Abstract SELection (IASEL) system is developed. Different interactions of the isoforms are the main focus of this study. Therefore, each abstract is screened for its protein mentions using the Genia tagger (http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/) and number of distinct protein mentions is recorded. Abstracts containing at least two different protein names are selected for further investigation. Such abstracts are classified as interaction abstract or non-interaction abstract by using an SVM classifier (here called the IASEL SVM Classifier). Details of the developed classifier are provided in the following subsections.

4.5.1 IASEL Classifier System

The IASEL classifier system is depicted in Figure 4.4. The system is composed of a pre-processing unit, a feature extraction and combination unit and an SVM classifier.

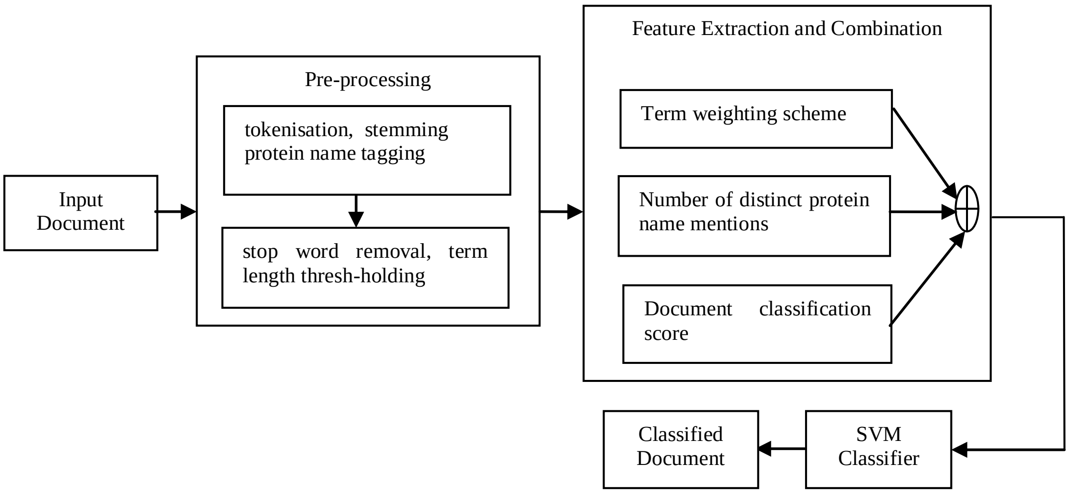

Input Document
Pre-processing
tokenisation, stemming protein name tagging
stop word removal, term length thresh-holding
Feature Extraction and Combination
Term weighting scheme
Number of distinct protein name mentions
Document classification score
SVM Classifier
Classified Document

Figure 4.4: IASEL System Overview

In the pre-processing unit, first each document is tokenised and tokens are converted to their stems. Next, each document is screened for its protein name mentions and all specific protein mentions are replaced with the word "PROT" to avoid the data sparseness problem. Then, all capitalisations, digits and special symbols are removed from the document. Genia Tagger is utilised for tokenisation, stemming and protein name tagging. Stop words are removed by using the list given in (www.pdg.cnb.uam.es/martink/LINKS/stop_word_list.txt). Tokens having length less than three characters are also removed, since many protein names contain more than two characters (Lan *et al*., 2007).

In the feature extraction and combination phase, performance contributions of 5 different term weighting schemes (TWSs) and domain specific features namely number of distinct protein mentions and document classification scores are investigated.

SVM which is one of the most widely used machine learning based classifiers in text mining applications is used to select the documents covering protein-protein interactions (Chen *et al.*, 2005). Robustness of these classifiers has been demonstrated in various text classification problems (Leopold and Kindermann, 2002; Burges, 1998) as well as in the recent BioCreative-II IAS challenge (Krallinger *et al.*, 2008). In this method, feature vectors are mapped into a higher dimensional space using a kernel function and then an optimal hyper-plane separating positive and negative data with the maximum margin is computed. Maximizing the margin improves generalisation ability of a given SVM classifier. The SVM^{Light} package (http://svmlight.joachims.org/) is utilised to implement the SVM classifier. The classifier is trained by using a linear kernel.

The BioCreative-II IAS dataset is used for training the SVM classifier and testing its performance. The training set consists of 3536 positive which are relevant (interaction) abstracts and 1959 negative which are irrelevant abstracts. The validation set includes 338 positive and 339 negative abstracts taken from the test data.

4.5.2 Features Used

4.5.2.1 Statistical Term Weighting Schemes

In text classification, TWSs are widely used to assign appropriate weights to the terms for improving the classification performance (Lan *et al.*, 2006). In order to design the IASEL SVM classifier, a set of TWSs is analysed to select the best performing one. The set consists of 5 different TWSs utilised in general text classification tasks. Standard BOW (Salton and Buckley, 1988) representation is used to represent each document as a vector of values, where each value represents a weight belonging to a term found in the document. Details on the TWSs used in the study are provided below.

Table 4.6 shows the numbers of documents which contain term t_i (i=1..M, M:total number of distinct terms in the training dataset) and do not contain term t_i in the positive and negative classes. These numbers play a role in calculating weights of the terms based on the TWSs listed below.

Table 4.6: Contingency Table for Document Frequency of Term t_i in Different Classes

Document Class	t_i	$\overline{t_i}$
Positive	a	b
Negative	c	d

(1) Normalised Term Frequency (NTF):

In this scheme, a weight is assigned to each term depending on the number of occurrences of the term in the document (Joachims, 2002). Term frequency (TF) is utilised in this task since it is a traditional scheme widely used in many text classification systems (Lan *et al.*, 2006). However, in order to eliminate the effect of the document's length, normalised version of the term frequency scheme is used. Normalised term frequency of each term is calculated by using the equation (4.1) (Joachims, 2002).

$$\text{NTF}=\frac{\text{number of occurence of term } t_i \text{ in the document}}{\text{length of the document}} \tag{4.1}$$

(2) NTF x Inverse Document Frequency (NTF.IDF):

The NTF scheme has the disadvantage of considering all terms as equally important while assessing the relevancy on a given document. This results in having important terms with little or no discriminating power in determining the relevance for a given document. For example, a set of documents on the protein families is likely to contain the term “protein” in every document leading to overestimate the term’s importance by assigning a high weight during the relevancy assessing process of a given document. The aim of NTF.IDF is to diminish the effect of terms which occur often in the set for making them meaningful in determining the relevant documents. Thus, high IDF scores are assigned to rare terms, whereas the IDF scores of frequent terms are likely to be low (Salton and Buckley, 1988).

This weighting scheme is selected to be used since it is popularly used in many text classification tasks (Lan *et al.,* 2006). NTF.IDF weight of each term is calculated by using the equation (4.2) (Joachims, 2002).

$$NTF \times \log_2 \frac{N}{(a+c)} \tag{4.2}$$

where,

N: Total number of documents in the corpus

(3) NTF x Relevance Frequency (NTF.RF):

The disadvantage associated with the IDF scheme lays in the absence of class information in calculation the term weights. More specifically, IDF scores of the terms are calculated based on only the sum of document frequencies of the terms where the terms appear, i.e. a and c values in Table 4.6. This leads to assign the same IDF value to the terms having different a/c ratios. In the cases of a>c or c>a, the traditional IDF can fail in improving the discriminating power of terms. Hence, (Lan *et al.*, 2006) has proposed the RF factor which takes a/c ratio into account.

NTF.RF is engaged into the analyses since its efficiency in text classification has been successfully demonstrated in the BioCreative-II IAS by the best performing system (Lan *et al.*, 2007). NTF.RF weight of each term is calculated by using the equation (4.3) (Lan *et al.*, 2006).

$$NTF \times \log_2(2 + \frac{a}{c}) \qquad (4.3)$$

(4) NTF x Balanced Relative Frequency (NTF.BRF):

Tsai and colleagues have proposed the BRF factor by mentioning that RF emphasises terms that are more frequently appear in positive rather than negative documents but ignore frequencies of terms other than target ones in documents (Tsai *et al.*, 2008). Therefore, RF works well when the portion of positive documents is smaller than the negatives one.

NTF.BRF is used in this study since it has been utilised in one of the previous IAS studies reporting high classification performance on the BioCreative-II IAS dataset (Tsai *et al.*, 2008). NTF.BRF weight of each term is calculated by using the equation (4.4).

$$NTF \times \log_2(2 + \frac{a \times b}{c \times d}) \qquad (4.4)$$

(5) NTF x Chi-square (NTF. χ^2):

Apart from the factors mentioned above, feature selection metrics, such as information gain and χ^2 can also be successfully used for term weighting (Lan *et al.*, 2006). For example, Deng and colleagues have reported that TF. χ^2 outperforms

TF.IDF in their study which they have used an SVM for text categorisation (Deng *et al.*, 2004). IDF and RF consider document frequencies of terms where the terms appear only (i.e. a and c numbers in Table 4.6). However, χ^2 considers document frequencies of terms where the terms do not appear (i.e. b and d numbers in Table 4.6) as well as appear (i.e. a and c numbers in Table 4.6).

NTF. χ^2 is selected to be used in this study since it is a promising term weighting scheme. NTF. χ^2 weight of each term is calculated by using the equation (4.5) (Yang and Pedersen 1997).

$$NTF \times \frac{N \times (ad - bc)^2}{(a+b) \times (c+d) \times (a+c) \times (b+d)} \tag{4.5}$$

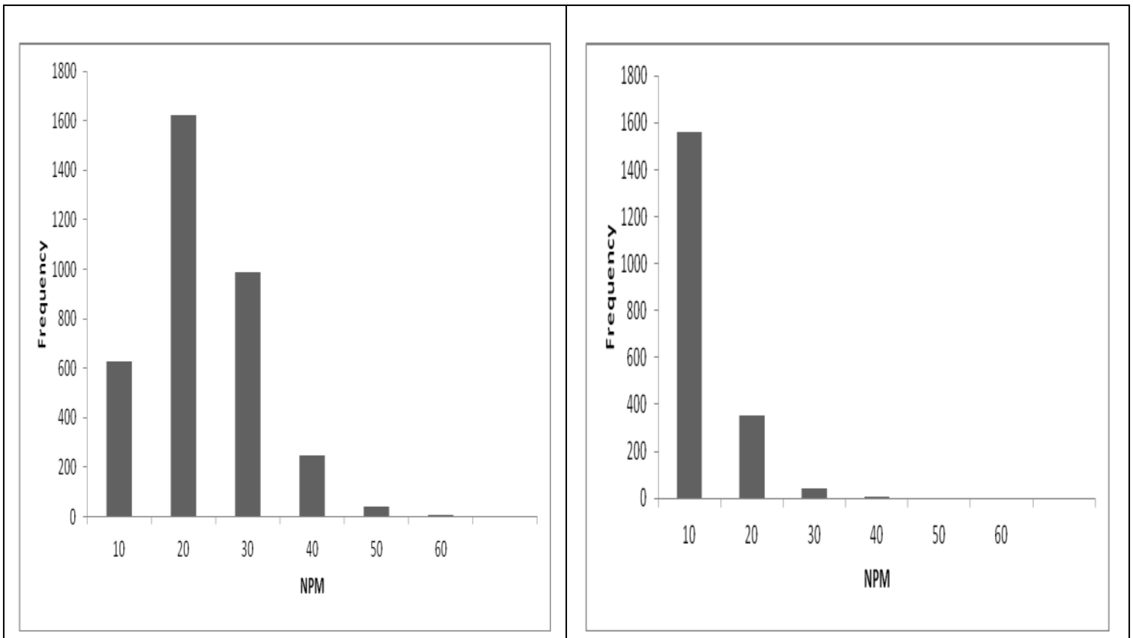
where,

N: Total number of documents in the corpus

4.5.2.2 Domain Specific Features

4.5.2.2.1 Number of Distinct Protein Mentions

It is demonstrated in (Abi-Haidar *et al.*, 2007) that for a randomly selected document, the probability of being a document discussing protein interactions increases with the number of distinct protein mentions in the document. The distribution of Number of distinct Protein Mentions (NPM) in interaction and non-interaction abstracts of the BioCreative-II IAS training set are documented (Figure 4.5). Results show that NPM values are higher in the majority of the interaction abstracts compared to the non-interaction abstracts. Therefore, NPM is believed to serve as a good domain specific feature for the IASEL task. In general, NPM value is obtained by dividing the number of distinct protein mentions by the maximum number of distinct protein mentions in the dataset used.



(a) Distribution of NPM values in the interaction abstract set

(b) Distribution of NPM values in the non-interaction abstract set

Figure 4.5: Histograms Showing the Distribution of NPM in the BioCreative-II IAS Training Dataset

4.5.2.2.2 Document Classification Scores

Marcotte and colleagues have proposed a Bayesian method to classify abstracts as interaction or non-interaction abstracts (Marcotte *et al.*, 2001). In this approach, first the “discriminating words” which are words describing PPIs are identified. Then, each document is assigned to a score calculated based on the frequencies of the discriminating words occurring in the document. A given score represents the likelihood of the document for being an interaction document. In this study, these scores are termed as Document Classification Scores (DCSs) and used as domain specific features to train the IASEL SVM classifier.

Discriminating word selection is based on occurrence statistics of the words from two different datasets: (1) A dataset consisted of interaction abstracts only and (2) A dataset consisted of a specific number of randomly selected abstracts. In the experimental set up, a total number of 250 interaction abstracts are randomly selected from the BioCreative-II IAS training set for (1). In order to form set (2), a total number of 61,777 abstracts are randomly selected from the PubMed DB. First, frequencies of all words in both dataset are calculated. Then, a “Dictionary” is formed from the set containing randomly selected abstracts’ words having frequencies greater than a pre-specified number (this number is selected as 3 as suggested in (Marcotte *et al.*, 2001). Each word is assigned to a score called the *ln(p-score)* (LNP-score) indicating its discriminative power by using equation (4.6)

$$ln\ p(n/\mathrm{N},f) \approx -Nf + n\ln(Nf) - \ln(n!) \tag{4.6}$$

where,

N: Total number of words in the dataset containing interaction abstracts

n: Total number of occurrences of a word in the dataset containing interaction abstracts

f: Dictionary frequency of a word

This approximation is valid when the total number of words used to generate the dictionary is much greater than *N* and when *f* is small.

Words having LNP-score smaller than a pre-specified number are selected as “Discriminating Words” and used to form a “Discriminating Word List” (DWL). The formed DWL includes the words within the top 40% range of the whole list. Any discriminating algorithm is biased by its training set. Therefore, gene and protein names as well as names of specific cellular systems and pathways are manually removed to alleviate this problem.

Document class scores are calculated by using the following equation:

$$S= \sum_i (n_i \ln \frac{f_{I,i}}{f_{N,i}} - N*(f_{I,i} - f_{N,i})) \tag{4.7}$$

where,

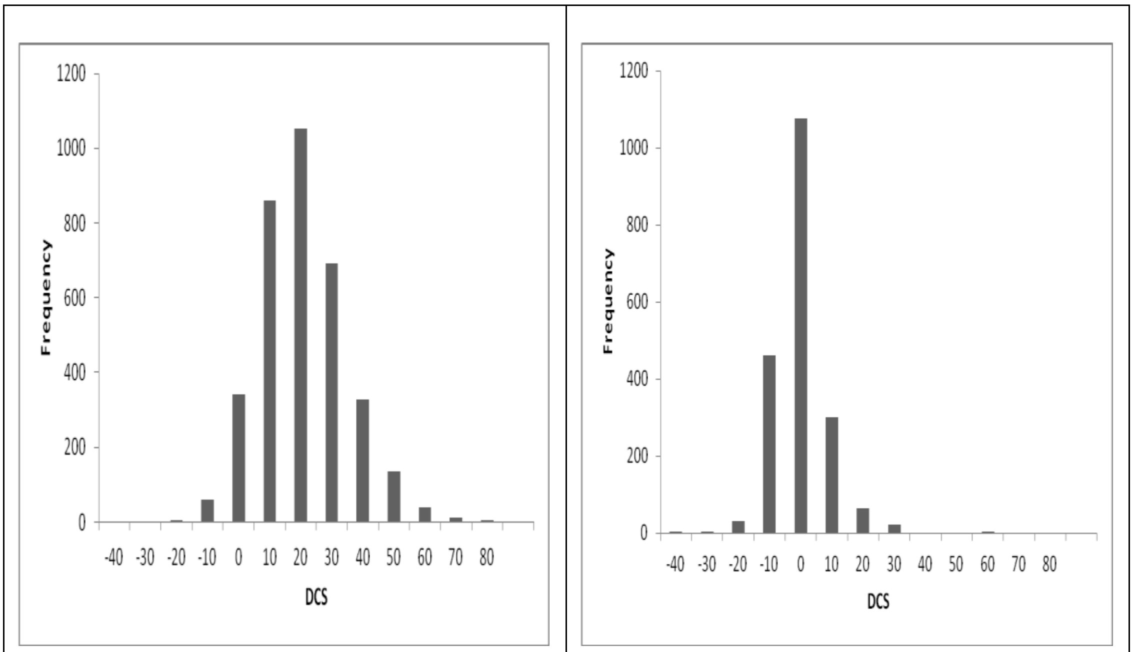
n_i: Number of occurrences of "Discriminating Word" *i* in the abstract

$f_{N,i}$: Dictionary frequency of "Discriminating Word" *i*

$f_{I,i}$: Frequency of "Discriminating Word" *i* in the dataset containing interaction abstracts

N: Total number of words in the abstract

Figure 4.6 shows the distribution of DCSs in the BioCreative-II IAS training dataset. Results demonstrate that majority of the interaction abstracts have positive scores while the majority of the non-interaction abstracts have scores around or less than zero. Hence, the usage of DCS could be extremely representative of the domain since it is generated from a Bayesian classifier trained on a set of biomedical documents.



(a) Distribution of DCS values in the interaction abstract set

(b) Distribution of DCS values in the non-interaction abstract set

Figure 4.6: Histograms showing the Distribution of DCSs in the BioCreative-II IAS Training Dataset

4.5.3 Results and Discussion on Interaction Abstract Selection

4.5.3.1 Effect of Feature Concatenation

The effect of concatenating different features is analysed using the BioCreative-II IAS train and test dataset. Firstly, the performance of the SVM classifier trained using domain specific features only is analysed and depicted in Table 4.7. Results show that, the classifier achieves an F_1-score value of 69.90% and 76.01% when only normalised NPM score and DCS is used, respectively. On the other hand, F_1-score increases to 79.37% when NPM is concatenated with DCS since these domain specific features exhibit complementary precision/recall behaviours.

Table 4.7: Classification Performance by Using Combination of Domain Specific Features Only

NPM	DCS	Precision(%)	Recall(%)	F_1-Score(%)
X		62.68	78.99	69.90
	X	76.58	75.44	76.01
X	X	76.44	82.54	79.37

NPM: distinct Number of Protein Mentions, DCS: Document Classification Scores, X sign indicates that the feature is used in the classifier design, Results are on the positive class

Secondly, both aforementioned pre-processing techniques namely stop word removal and term length thresholding are applied and the effect of combining TWSs with the domain specific features are analysed (Table 4.8). The total number of terms is calculated as 43,837 and the BOW approach is used to represent the documents as a vector of values to the SVM classifier. When, NPM is concatenated with the term weights, compared to the classifiers' performance utilising term weights only, the precision decreases for all of the TWSs except the ones utilising NTF.IDF and NTF.χ^2. On the other hand, the recall increases for all classifiers. This is reflected on to the F_1-score value where it decreases for all classifiers except the one using NTF.χ^2.

When, DCS is concatenated with the term weights, precision increases for all classifiers while recall decreases for all classifiers except the one utilising NTF.χ^2. This results in a decrease in F_1-score for all classifiers. On the other hand, concatenation of term weights with the two domain specific features results in an increase in the F_1-score values. This result is expected since combining features which have complementary precision/recall behaviour generally improves the classification performance (Zhang *et al.*, 2007). The best performing classifier uses NTF.χ^2 term weights concatenated with the two domain specific features and achieves an F_1-score of 81.31%.

Table 4.8: Classification Performance Using Different Feature Sets

TWS	NPM	DCS	Precision(%)	Recall(%)	F_1-Score(%)
NTF.IDF			73.04	82.54	77.50
NTF.IDF	X		73.20	84.02	78.24
NTF.IDF		X	74.58	78.99	76.72
NTF.IDF	**X**	**X**	**75.00**	**83.43**	**78.99**
NTF			73.35	82.25	77.55
NTF	X		71.19	86.96	78.29
NTF		X	77.74	77.51	77.62
NTF	**X**	**X**	**77.59**	**81.95**	**79.71**
NTF.χ^2			76.52	78.11	77.31
NTF.χ^2	X		75.27	83.73	79.27
NTF.χ^2		X	79.76	79.29	79.52
NTF.χ^2	**X**	**X**	**78.51**	**84.32**	**81.31**
NTF.BRF			76.23	77.81	77.01
NTF.BRF	X		75.55	81.36	78.35
NTF.BRF		X	77.95	76.33	77.13
NTF.BRF	**X**	**X**	**78.26**	**79.88**	**79.06**
NTF.RF			76.37	78.40	77.37
NTF.RF	X		76.11	81.07	78.51
NTF.RF		X	78.79	76.92	77.84
NTF.RF	**X**	**X**	**78.61**	**80.47**	**79.53**

TWS: Term Weighting Scheme, NPM: distinct Number of Protein Mentions, DCS: Document Classification Scores, X sign indicates that the domain specific feature is concatenated to the TWS, Results are on the positive class

In order to gain insights into the superior performance of NTF. χ^2 over the other TWSs when they are concatenated with NPM and DCS, dynamic ranges of the features used are analysed in Table 4.9. NTF. χ^2 has a larger range compared to the ranges of other TWSs. Therefore, its superior performance can be attributed to its large dynamic range of weights which fits with the dynamic ranges of the two domain specific features, especially with DCS.

Table 4.9: The Maximum and Minumim Values of the Features Analysed

Maximim/Minimum value for the classes	Features						
	NTF	NTF.IDF	NTF.RF	NTF.BRF	NTF.χ^2	NPM	DCS
maximim value (positive class)	1.40×10^{-1}	3.16	6.05×10^{-1}	6.03×10^{-1}	105.83	1	58.75
minimum value (positive class)	5.18×10^{-3}	6.34×10^{-2}	5.19×10^{-3}	5.18×10^{-3}	1.96×10^{-8}	4.88×10^{-2}	-23.22
maximim value (negative class)	1.50×10^{-1}	2.90	4.21×10^{-1}	4.19×10^{-1}	98.04	1	48.52
minimum value (negative class)	4.41×10^{-3}	5.66×10^{-2}	4.41×10^{-3}	4.40×10^{-3}	1.32×10^{-8}	0	-49.33

A Z-test analysis is appield on performances of the classifiers utilising TWSs in addition to the two domain specific features in order to understand if there is any statistically significant difference in the classifier performances. Z-score for each pair of classifier is calculated by using the following equation:

$$Z - score = \frac{|F_A - F_B|}{\sqrt{\frac{2F(1-F)}{N}}} \qquad (4.7)$$

where,

F_A and F_B are the F_1-scores of any two classifiers A and B respectively,

N is the number of training documents,

F is calculated by $(F_A+F_B)/2$.

Z-score for each pair of classifier, *p*(A,B) is documented in Table 4.10. The difference in the classifer performances is assumed to be statistically significant at a confidence level of 95% if Z-score>1.96. Based on this analysis, it can be concluded that the classifier trained by using NTF.χ^2 in concatenation with NPM and DCS performs better than the ones trained using term weights NTF.IDF and NTF.BRF in concatenation with NPM and DCS. The analysis shows that, there is no statistically significant difference in the performances for the remaining pairs of classifiers.

Table 4.10: Z-scores of classifier pairs

Feature Set Used to Train A	**Feature Set Used to Train B**	$\mathbf{F_A}$	$\mathbf{F_B}$	**Z-score**
{NTF, NPM, DCS}	{NTF.IDF, NPM, DCS}	0.7971	0.7899	0.75
{NTF, NPM, DCS}	{NTF. χ^2, NPM, DCS}	0.7971	0.8131	1.70
{NTF, NPM, DCS}	{NTF.BRF, NPM, DCS}	0.7971	0.7906	0.68
{NTF, NPM, DCS}	{NTF.RF, NPM, DCS}	0.7971	0.7953	0.19
{NTF. χ^2, NPM, DCS}	**{NTF.IDF, NPM, DCS}**	**0.8131**	**0.7899**	**2.45**
{NTF. χ^2, NPM, DCS}	**{NTF.BRF, NPM, DCS}**	**0.8131**	**0.7906**	**2.37**
{NTF. χ^2, NPM, DCS}	{NTF.RF, NPM, DCS}	0.8131	0.7953	1.89
{NTF.IDF, NPM, DCS}	{NTF.BRF, NPM, DCS}	0.7899	0.7906	0.07
{NTF.IDF, NPM, DCS}	{NTF.RF, NPM, DCS}	0.7899	0.7953	0.56
{NTF.BRF, NPM, DCS}	{NTF.RF, NPM, DCS}	0.7906	0.7953	0.49

The best classification performances from Table 4.7 and Table 4.8 are compared with other state-of-the-art performing systems reported in the literature in Table 4.11. The system trained by using NTF.χ^2, NPM and DCS is 3.31% better than the best system's performance of the BioCreative-II IAS challenge which achieves an F_1-score of 78.00% (Lan *et al*., 2007). Also, it is 1.06% better than the system reported in (Lan *et al*., 2009) utilising classifier combination for the IASEL task. Furthermore, the performance is 0.41% better than the system described in (Tsai *et al*., 2008) which reports an F_1-score of 80.90%. However, it falls behind the best reported system in (Wang *et al*., 2008) by 3.07% which uses the Adaboost method for feature concatenation on the same data set.

Using term weights as features in the BOW approach could suffer from large feature vector dimension. Therefore, the classifier's performance trained by using the two domain specific features only is also compared with the state-of-the-art systems. Its performance is competitive with the high performing systems listed in the table which use a feature set including term weights in addition to domain specific features. Furthermore, the performance of this system is 1.37% better than the top ranking system in the BioCreative-II IAS challenge which uses a large feature set based on the NTF.RF scheme (Lan *et al.*, 2007). However, it should be noted that, although, the feature vector dimension is reduced from thousands to only 2 by not using the BOW features in the feature set, computation of DCSs with the Naïve Bayes classifiers is still computationally expensive.

Table 4.11: Performances of High-Performing IASEL Systems on the BioCretive-II IAS Test Set

IASEL Study	**F_1-Score(%)**
(Wang *et al.*, 2008)	84.38
SVM trained on the set {NTF.χ^2, NPM and DSC}	**81.31**
(Tsai *et al.*,2008)	80.91
(Lan *et al.*, 2009)	80.25
SVM trained on the set {NPM and DCS}	**79.37**
(Lan *et al.*, 2007)	78.00

4.5.3.2 Training a Classifier for the Selection of Interaction Abstracts

The IASEL SVM classifier which is used for selecting the interaction abstracts from the set of retrieved abstracts is trained on the BioCreative-II IAS training and test datasets together. In order to design the classifier, performances of different feature sets including TWSs are compared when they are concatenated with NPM and DCS by using 10-fold cross validation (Table 4.12). The SVM trained with a feature set including term weights from NTF. χ^2 , NPM and DCS is selected as the IASEL SVM classifier given that it achieves the best F_1-score value (90.59%) on the BioCreative-II IAS dataset.

Table 4.12: Performances of Different Feature Sets on the BioCreative-II IAS Dataset

Feature Set	**10-Fold Cross Validation**		
	Precision(%)	**Recall(%)**	**F_1-score(%)**
NTF.χ^2 + NPM + DCS	91.28	89.92	90.59
NTF.IDF + NPM + DCS	89.97	88.77	89.36
NTF.RF + NPM + DCS	89.27	87.28	88.26
NTF.BRF + NPM + DCS	89.24	87.28	88.25
NTF + NPM + DCS	89.17	87.26	89.20

NPM: distinct Number of Protein Mentions, DCS: Document Classification Scores, + sign representes feature concatenation, Results are on the positive class

Figure 4.7 depicts the utilisation of the IASEL SVM classifier for selecting the abstracts which are likely to contain PPI information from the retrieved set of relevant abstracts.

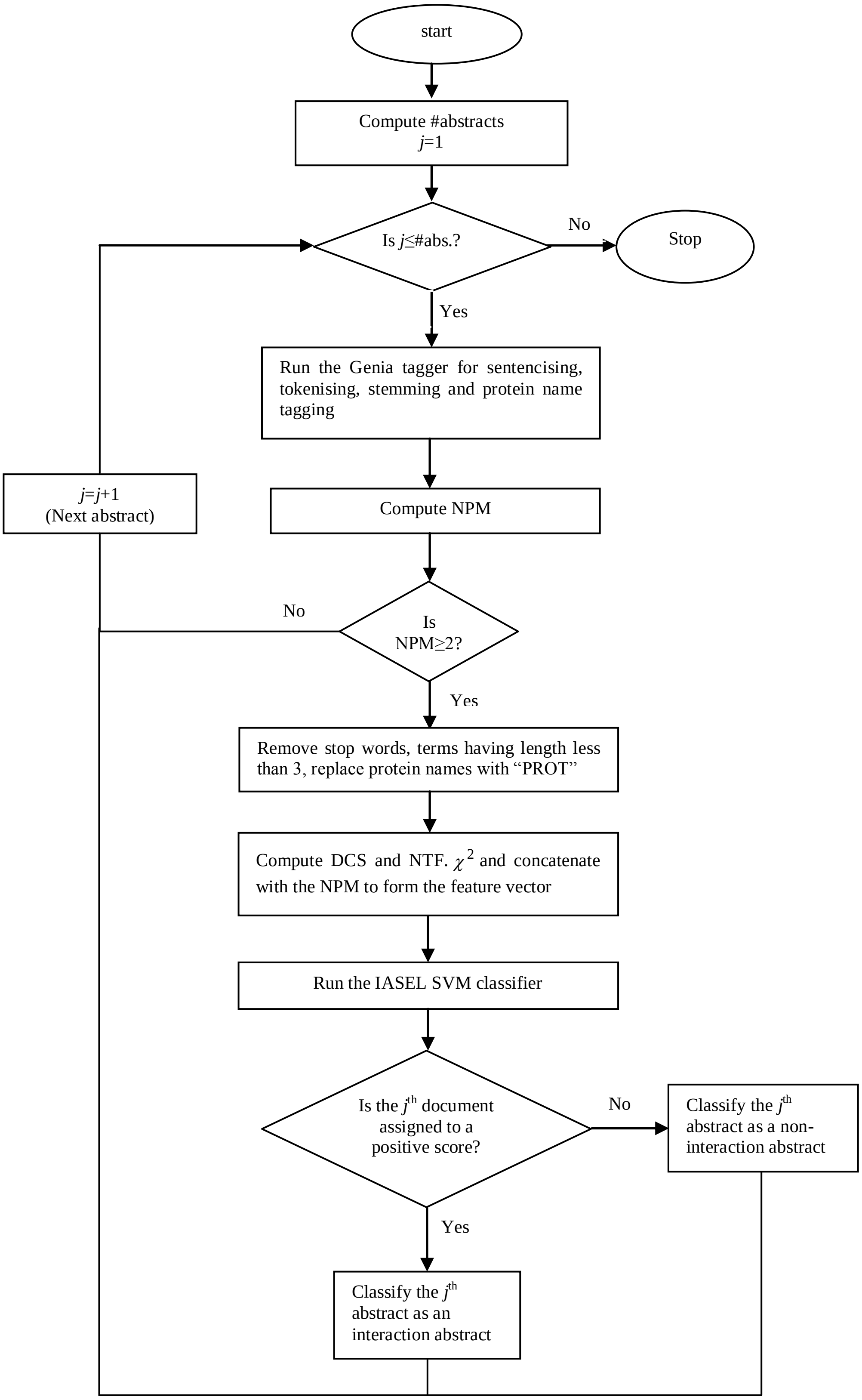

start
Compute #abstracts
j=1
Is j≤#abs.?
No
Stop
Yes
Run the Genia tagger for sentencising, tokenising, stemming and protein name tagging
j=j+1
(Next abstract)
Compute NPM
No
Is NPM≥2?
Yes
Remove stop words, terms having length less than 3, replace protein names with "PROT"
Compute DCS and NTF. χ2 and concatenate with the NPM to form the feature vector
Run the IASEL SVM classifier
Is the jth document assigned to a positive score?
No
Classify the jth abstract as a non-interaction abstract
Yes
Classify the jth abstract as an interaction abstract

Figure 4.7: Interaction Abstract Selection

4.6 PPI Extraction

PPI extraction is commonly formalised as a binary classification task where the system identifies the protein pairs having biological relationship in a given sentence. A sample interaction sentence and a pair of interest are shown in Figure 4.8.

LEC$_{[PROT1]}$ induces chemotaxis and adhesion by interacting with CCR1$_{[PROT2]}$ and CCR8$_{[PROT]}$

Figure 4.8: An Example Sentence Including PPI Data

4.6.1 PPI Extraction System Overview

The PPI extraction system developed for identifying the interaction pairs is depicted in Figure 4.9. The system consists of a pre-processing unit, a syntactic parsing unit and an SVM classifier.

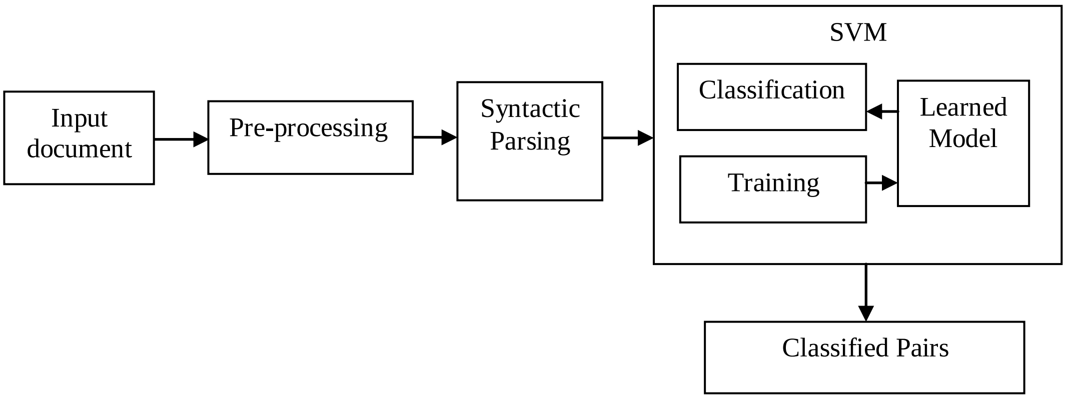


Figure 4.9: Overview of the PPI Extraction System

In the pre-processing unit, first, the tagged protein names in the selected interaction abstracts are normalised by using GNAT (Hakenberg *et al*., 2008). In this study, normalisation is particularly required in order to link the isoforms and their interactions to DTs from the HumanSDB3 clusters for further analysis. Next, sentences having *n* different proteins (*n*>2) are selected and replicated into C_2^n sentences. Each replicated sentence has exactly two of the protein names tagged and replaced with “PROT1” and “PROT2” while the rest of the protein tags are replaced with “PROT”. Figure 4.10 shows 3 replicated sentences for the sentence shown in Figure 4.8 which includes 3 protein mentions.

LEC$_{[\text{PROT1}]}$ induces chemotaxis and adhesion by interacting with CCR1$_{[\text{PROT2}]}$ and CCR8$_{[\text{PROT}]}$

LEC$_{[\text{PROT1}]}$ induces chemotaxis and adhesion by interacting with CCR1$_{[\text{PROT}]}$ and CCR8$_{[\text{PROT2}]}$

LEC$_{[\text{PROT}]}$ induces chemotaxis and adhesion by interacting with CCR1$_{[\text{PROT1}]}$ and CCR8$_{[\text{PROT2}]}$

Figure 4.10: Example Replicated Sentences

In the syntactic parsing module, two different syntactic parsers are employed to document dependency and deep relations between the words constituting the sentences respectively.

The SVM classifier implemented for identifying the interaction pairs is trained by using multiple kernels utilising the information from two different syntactic parsers. For this purpose, tree kernels in SVM$^{\text{Light}}$ (SVM-TK) (Moschitti, 2006) package is used. Tree kernels measure the similarity between two input trees by counting their common sub trees. For example, Figure 4.11 illustrates graphical representation of the parse trees of two noun phrases according to the syntactic tree kernels. The similarity between the two trees is calculated as 3 given that 3 out of 5 structures are identical.

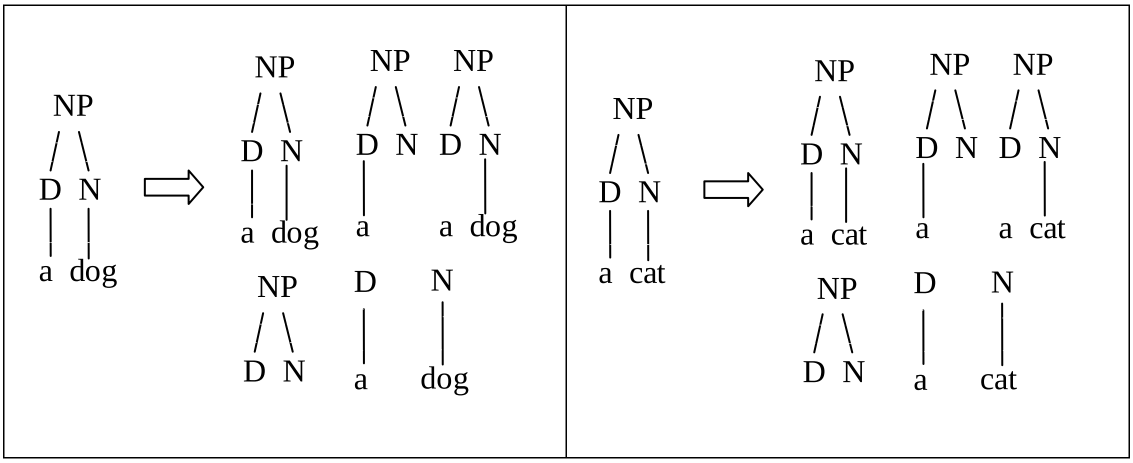


Figure 4.11: Example Parse Trees

(Source: http://disi.unitn.it/moschitti/Tree-Kernel.htm)

4.6.2 Features Used

Features used for PPI extraction are generated by utilising the standard BOW approach in addition to dependency and deep relations from the syntactic parsers used.

4.6.2.1 Bag-of-Words

It's widely accepted that words surrounding the candidate entities potentially carry evidence regarding their relationship (Phan *et al.*, 2007). Therefore, standard BOW representation is utilised as features. The set of features includes first three left stem words, first three right stem words and all the stem words between the two protein names. Figure 4.12 shows the BOW features for the sentence given in Figure 4.8.

Left: -
Between: induce, chemotaxis, and, adhesion, by, interact, with
Right: and, PROT

Figure 4.12: BOW Features

4.6.2.2 Syntactic Relations

Efficiency of syntactic parsing techniques in PPI extraction task have been demonstrated in several previous studies (Airola *et al.*, 2008; Miwa *et al.*, 2009a;

Miwa *et al.*, 200b; Miyao *et al.*, 2008; Sætre *et al.*, 2007). Such studies take into account the grammatical content of the sentences including dependencies between words and deep structures rather that the word itself. The designed PPI extraction system uses the dependency and deep relations generated using the dependency parser, Ksdep (Sagae and Tsujii, 2007) and the deep parser, Enju (Miyao and Tsujii, 2008) tuned for the biomedical domain.

A dependency parser takes sentences as input and produces a graph for each sentence where the nodes are the words and the arcs are dependency links between words. Figure 4.13 shows a parse tree produced by Ksdep for the interaction sentence shown in Figure 4.8. Ksdep generates binary relations between head and dependent nodes. For example, in the figure, the verb "induces" is the head node of "PROT1" as well as "by" which are the subject and a verb modifier in the sentence respectively. Existences of dependent nodes "PROT1" and "by" depend on the existence of their head node, "induce", in the sentence. The shortest path between the protein pairs of interest is used to extract dependency relations as features. Figure 4.14 shows a sample feature in tree format extracted from the dependency relations shown in Figure 4.8 where the prefix "r" refers the reverse relation. Reverse relation indicates that the direction of the arc between a given head-dependent pair is reverse.

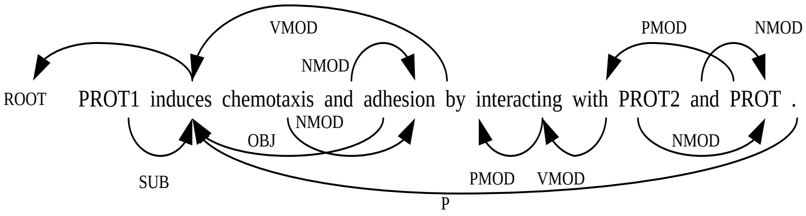

VMOD
PMOD
NMOD
NMOD
ROOT
PROT1 induces chemotaxis and adhesion by interacting with PROT2 and PROT .
NMOD
OBJ
NMOD
SUB
PMOD
VMOD
P

Figure 4.13: Dependency Relations Generated by Ksdep

(KSDEP (SUB (PROT1 induce))(rVMOD (induce by)) (rPMOD (by interact))(rVMOD (interact with))(rPMOD (with PROT))(rNMOD (PROT PROT2)))

Figure 4.14: Shortest Path Dependency Feature

A deep parser takes sentences as input and produces a graph for each sentence representing syntactic as well as semantic relations among the words. Figure 4.15 shows the parse graph generated for the interaction sentence shown in Figure 4.8 by Enju parser. This parser uses Predicate-Argument Structure (PAS) to represent the semantic relations between the words. For example, in the figure, the verb “induces” is the predicate and the subject “PROT1” is its first argument (arg1) while the object “chemotaxis” is its second argument (arg2). Figure 4.16 shows an example feature in the tree format extracted from the deep relations shown in Figure 4.15 where the prefix “r” refers the reverse relation.

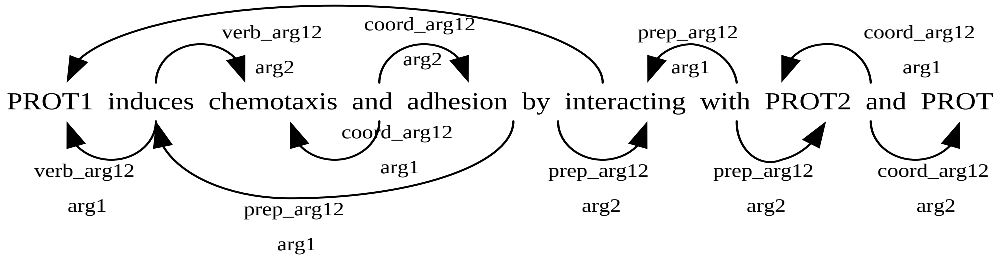


Figure 4.15: Deep Relations Generated by Enju

(ENJU (rverb_arg12_arg1 (PROT1 interact)) (rprep_arg12_arg1 (interact with))(prep_arg12_arg2 (with PROT2)))

Figure 4.16: Shortest Path Deep Feature

4.6.3 Results and Discussions on PPI Extraction

The SVM is trained on the AIMed corpus which is one of the largest PPI corpora consisting of 225 Medline abstracts belonging to human. In the corpus, protein names as well as the exact locations of statements expressing PPI information have been annotated providing an opportunity to develop PPI extraction systems using ML approaches. The corpus version and splitting procedure recommended in (Airola *et al.*, 2008) are used for conducting the experiments. The dataset includes 1000 positive and 4834 negative pairs.

Several experiments are conducted by using different combinations of the BOW and syntactic features described in section 4.6.2 on the AIMed corpus. Results from these experiments are reported in Table 4.13. The performance is measured in an abstract wise 10-fold cross validation (the corpus is splitted into 10 sets including equal number of abstracts) by using one-answer-per-occurrence criterion. In addition, the separating hyper-plane of the SVM is controlled by setting the regularization parameter C to 2. This value is chosen by cross-validation experiments.

Table 4.13: PPI Extraction Performance on AIMed Corpus by Using Different Feature Sets

Feature Set			**Kernel Type**		**Precision(%)**	**Recall(%)**	**F_1-Score(%)**
BOW	**K**	**E**	**T**	**L**			
X				X	51.69	37.36	42.82
	X		X		58.63	38.30	45.61
		X	X		58.88	33.92	42.40
	X	X	X		57.38	37.06	44.34
X	X		X	X	58.65	49.59	53.13
X		X	X	X	57.78	48.07	52.08
X	X	X	X	X	60.77	49.70	54.20

B: BOW features, K: Features extracted by using Ksdep, E: Features extracted by using Enju, T: Tree Kernel, L: Linear Kernel, Results are on the positive class

Different syntactic parsers can handle different layers of syntactic relations. The dependency parser misses some deep relations whereas the deep parser misses some shallow relations. In addition, the kernels used have some different advantages and disadvantages. For example, the linear kernel utilising BOW features can combine the words while ignoring the order of the words and their relations. The tree kernels can calculate the similarity between the shortest paths while ignoring the words and paths outside of the shortest paths (Miwa *et al.*, 2009b). Hence, different features capture different aspects from the sentences. In parallel to these, results show that the performance of the SVM classifier improves when different kernels are combined and it achieves the best performance (F_1-score value of 54.20%) when all features are concatenated. This classifier called the PPI SVM classifier.

Even though, many PPI extraction systems have been reported in the literature, each system has utilised different pre-processing methods resulting in different number of pairs from the AIMed corpus. Furthermore, different systems have used different evaluation procedures such as one-answer-per-occurrence (if the same protein is mentioned multiple times in the sentence, the interaction must be extracted for each occurrence) and one-answer- per-relation (multiple occurrences of the same protein interaction are considered one correct answer). Such factors make the comparison of different systems difficult (Pyysalo *et al.*, 2008). Therefore, the PPI SVM classifier is compared with other systems which have reported their performances on the AIMed corpus by using the same splitting method with (Airola *et al.*, 2008) only (Table 4.14). The state-of-the-art systems generally rely on an SVM classifier and combine multiple kernels to tackle the PPI extraction problem. The systems reported in (Miwa *et al.*, 2009a; Miwa *et al.*, 2009b; Miwa *et al.*, 2008; Miyao *et al.*, 2008) have trained SVM classifiers which combine linear kernels utilising BOW features, tree kernels utilising shortest path syntactic features generated through dependency and deep parsers and the all-path graph kernel proposed in (Airola *et al.*, 2008) which uses graph features. Miwa and colleagues have reported the highest F_1-score of 65.20% (Miwa *et al.*, 2009b) while the classifier using the graph kernel alone has been reported to achieve an F_1-score of 56.40% (Airola *et al.*, 2008). Yang and colleagues have expanded the feature set proposed by (Miwa *et al.*, 2009a) with additional

BOW features such as interaction keywords and protein names distance (Yang *et al.*, 2010a) and proposed to combine the kernels by using SVM ranking (Yang *et al.*, 2010b) or kernel weighting approaches (Yang *et al.*, 2011). These studies demonstrate that different kernel combination methods could also help to develop a high performing PPI classifier (F_1-score of 64.88% in (Yang *et al.*, 2010b) and F_1-score of 64.41% in (Yang *et al.*, 2011)). Li and colleagues have shown that semi-supervised learning is another method for designing a high performing classifier (Li *et al.*, 2010). Their system has achieved an F_1-score of 63.50%. On the other hand, lower performances have been obtained when a single kernel is used (F_1-score of 54.70% in (Liu *et al.*, 2010) and F_1-score of 53.50% in (Niu *et al.*, 2010)).

Table 4.14: Performances of Different PPI Extraction systems on AIMed

PPI Extraction Study	**#positive pairs**	**#pairs**	**Feature set**	**Kernel Type**	**P(%)**	**R(%)**	**F(%)**
(Miwa *et al.*, 2009b)	1000	5834	B+S	L+T+G	60.00	71.90	65.20
(Yang *et al.*, 2010b)	1000	5834	B+S	L+T+G	59.57	71.16	64.88
(Yang *et al.*, 2011)	1000	5834	B+S	L+T+G	57.72	71.07	64.41
(Miwa *et al.*, 2008)	1005	5648	B+S	L+T+G	60.40	69.30	64.30
(Li *et al.*, 2010)	1000	5834	B+S	L	60.47	68.31	63.50
(Miwa *et al.*, 2009a)	1000	5834	B+S	L+T+G	55.00	68.80	60.80
(Miyao *et al.*, 2008)	1059	5648	B+S	L+T	54.90	65.60	59.50
(Yang *et al.*, 2010a)	1000	5834	B+S	L+T+G	49.28	70.04	57.85
(Airola *et al.*,2008)	1000	5834	S	G	52.90	61.80	56.40
(Liu *et al.*, 2010)	1000	5834	B+S	L	63.40	48.80	54.70
Our system	**1000**	**5834**	**B+S**	**L+T**	**60.77**	**49.70**	**54.20**
(Niu *et al.*, 2010)	1000	5834	B+S	L	70.20	43.20	53.50

B: BOW, S: features generated by using syntactic parsers, L:Linear Kernel, T:Tree Kernel G: Graph kernel, P: Precision, R: Recall, F: F_1-Score, Results are on the positive class

The PPI SVM classifier developed in this study has an F_1-score of 54.20%. Syntactic features from Ksdep and Enju parsers as in (Miwa *et al.*, 2008) are utilised to train the SVM classifier. However, the BOW approach used is different from the approach used in their studies. The BOW approach used relies on the words surrounding the target protein pair while their approach relies on the top 1000 words with frequency information surrounding the target pair. Factors, which make the direct comparison between the PPI SVM classifier and the others difficult, are number of pairs reported for the AIMed corpus, parameter tuning and F_1-score calculation method. (Miyao *et al.*, 2008) has included self-interacting pairs and identified ~200 less negative pairs than other studies in AIMed. (Aiorola *et al.*, 2008) has applied leave-one-out principle to tune the parameters while Miwa and colleagues have controlled the position of the separating hyper-plane of the SVM by varying the threshold and calculating the average (Miwa *et al.*, 2008; Miwa *et al.*, 2009a). Furthermore, (Miwa *et al.*, 2009b) has reported their performances as macro-averaged F_1-score which is based on the calculation of F_1-score per document and then averaging it across documents. This factor makes the comparison difficult and unfair since other systems as well as the classifier developed in this work report their performance using the one-answer-per-occurrence criterion. The PPI SVM classifier achieves high precision which is desired for the interaction variation analysis described in Chapter 5. The system's performance is within the acceptable range of the state of the art considering the issues discussed above.

Figure 4.17 depicts the utilisation of the PPI SVM classifier for selecting the interaction protein pairs from the set of abstracts selected by the IASEL SVM classifier.

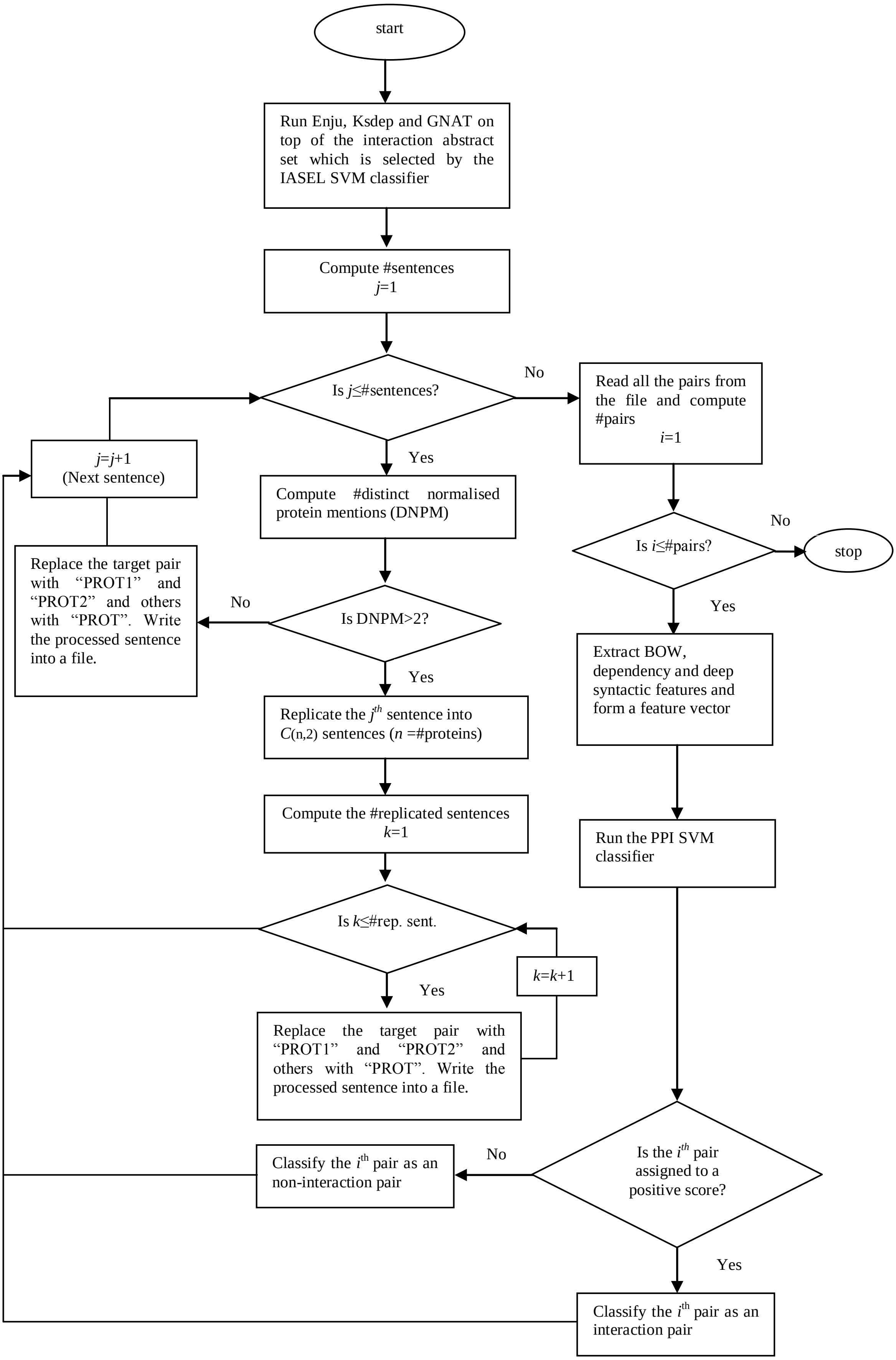

start
Run Enju, Ksdep and GNAT on top of the interaction abstract set which is selected by the IASEL SVM classifier
Compute #sentences
j=1
Is j≤#sentences?
No
Read all the pairs from the file and compute #pairs
i=1
Yes
j=j+1
(Next sentence)
Compute #distinct normalised protein mentions (DNPM)
No
Is i≤#pairs?
stop
Replace the target pair with "PROT1" and "PROT2" and others with "PROT". Write the processed sentence into a file.
No
Is DNPM>2?
Yes
Extract BOW, dependency and deep syntactic features and form a feature vector
Yes
Replicate the jth sentence into C(n,2) sentences (n =#proteins)
Compute the #replicated sentences
k=1
Run the PPI SVM classifier
Is k≤#rep. sent.
k=k+1
Yes
Replace the target pair with "PROT1" and "PROT2" and others with "PROT". Write the processed sentence into a file.
Classify the ith pair as an non-interaction pair
No
Is the ith pair assigned to a positive score?
Yes
Classify the ith pair as an interaction pair

Figure 4.17: Interaction Pair Selection

4.7 Literature Mining Results

Literature analysis results are shown in Table 4.15. A total number of 4,083,094 abstracts are retrieved from the PubMed DB for 12,638 different human alternatively spliced genes. Abstracts containing less than two different protein mentions are removed as mentioned earlier. From the remaining set of 2,465,692 abstracts, 205,270 abstracts are identified as containing PPI information according to the developed IASEL SVM classifier. From these abstracts, a total number of 267,718 sentences containing at least two different protein names are selected and a total number of 1,200,483 hypothetical protein interaction pairs are generated from them. Each pair is tested for interaction by using the developed PPI extraction classifier. Consequently, a total number of 33,158 distinct interactions are identified. Self-interacting proteins are excluded from the analysis since the interactions of different isoforms are the main focus of this study.

Table 4.15: Literature Analysis Results for Human Alternatively Spliced Genes

Phase		Total
Abstract Retrieval#		4,083,094
Abstract Selection	Abstract*	2,465,692
	Interaction abstracts	205,270
PPI Extraction	Sentence*	267,718
	Protein pairs generated	1,200,483
	Distinct Interaction protein pairs	33,158

#Although in total 4,187,861 records are retrieved, 4,083,094 of them have abstract text while the remaining 104,767 have title only

*Text containing at least two different protein mentions

4.8 Linking Literature Mining Results to HumanSDB3

4.8.1. Definitions

Following categorisation is used to distinguish between the extracted interacting pairs according to the cluster types of the protein partners constituting the pairs.

(a) CMT-CMT: Both interaction partners are from CMTs.

(b) CMT-CST: One of the interaction partners is from a CMT while the other is from a CST.

(c) CMT-NA: One of the interaction partners is from a CMT while the other one cannot be identified as an isoform linked to any variant cluster from HumanSDB3 (NA for Not Available).

(d) CST-CST: Both interaction partners are form CSTs

(e) CST-NA: One of the interaction partners is from a CST while the other one cannot be identified as an isoform linked to any variant cluster from HumanSDB3 (NA for Not Available).

4.8.2. Distribution of Isoform Interactions Based on the Validation against HumanSDB3

Each extracted protein interaction pair through the developed text mining pipeline is validated against HumanSDB3 (Table 4.16). For this purpose, Entrez Gene DB IDs of the isoforms are referenced to their corresponding DTs in HumanSDB3. Both protein partners are confirmed in the database for a total number of 22,018 (66.40%) interaction pairs which constitutes the majority of the extracted interactions. Only one of the interaction partners could be validated in the database for a total number of 9,801 (29.56%) pairs. There is no reference in the database for the remaining 1,339 (4.04%) interaction pairs. Interaction pairs having at least one protein referencing to a DT in HumanSDB3 are used to construct the TBIID. In total, 31,819 (96%) pairs of the extracted interactions are imported into TBIID.

Table 4.16: Distribution of Interaction Pairs Based on the Validation Against HumanSDB3

` **Reference in HumanSDB3**	**Number of Pairs**	**Percentage (%)**
Both proteins	22,018	66.40
Only one protein	9,801	29.56
None	1,339	4.04
Total	33,158	100

For the purpose of interaction variability analysis to be described in Chapter 5, all the 31,819 interaction pairs are further classified according to the cluster types of the protein partners as shown in Table 4.17. A total number of 102 (0.32%) pairs are identified as CMT-CMT type, where both protein partners belong to a CMT. For 2,548 (8.01%) pairs, one of the proteins belongs to a CMT while the other belongs to a CST. A total number of 697 (2.19%) pairs are identified as CMT-NA, where one of the proteins belongs to a CMT while there is no reference in HumanSDB3 for the other. Majority of the pairs (19,368, 60.87%) are CST-CST type. CST-NA type represents the pairs with only one of the partners belonging to a CST, while the other one has no reference in HumanSDB3. A total number of 9,104 (28.61%) of the interaction pairs are identified as such pairs.

Table 4.17: Distribution of Interaction Pairs According to Cluster Types of Protein Partners

Cluster Types of Protein Partners	**Number of Pairs**	**Percentage (%)**
CMT-CMT	102	0.32
CMT-CST	2,548	8.01
CMT-NA	697	2.19
CST-CST	19,368	60.87
CST-NA	9,104	28.61
Total	31,819	100

4.9 Manual Assessment of the Text Mining Pipeline

Text mining systems often generate systematic errors at the output level, mainly due to the limitations of the automated tools developed. As stated earlier, the text mining pipeline implemented consists of a protein name normalisation tool, an SVM classifier discriminating interaction abstracts and another SVM classifier for extracting the interaction protein pairs. An error introduced at the earlier stages of the pipeline can propagate causing a combined disturbance to the final output. The SVM classifier used for discriminating interaction abstracts is one of the best performing systems in the IASEL domain (F_1-score of 81.31% on BioCreative-II IAS test dataset, see section 4.5.2.1). Hence, the classifier performance is believed to achieve high enough to have sufficient recall for the analysis. However, several misclassification errors may araise during gene normalisation and PPI extraction tasks. For example, protein mentions in the abstracts are normalised to their corresponding Entrez Gene DB IDs by using GNAT. Although GNAT is one of the high performing normalisers in the domain, it comes with a limited recall (73.8%) due to missed protein names, their partial recognition or wrong assignments of protein IDs. For example the sentence *"Tudor domain missense mutations, including one found in an SMA patient, impair the interaction between SMN and fibrillarin (as well as the common snRNP protein SmB)*" states that "SMN" and "SmB" do interact.

However, the interaction information is missed (false negative) given that the GNAT does not normalise the protein “SmB”.

The SVM classifier utilised in the PPI extraction task uses dependency and semantic relations between the tokens constituting the sentence through several syntactic parsers. The efficiency of use of such parsing techniques in PPI extraction task has been demonstrated in previous studies (Airola *et al.*, 2008; Miwa *et al.*, 2009b; Miyao *et al.*, 2008). Nevertheless such systems often fail when the word describing the interaction does not occur in the shortest path between the interaction protein partners in the generated parse tree. The SVM classifier used in PPI extraction task is trained on the AIMed corpus which is a gold standard but a small corpus. It contains only 225 abstracts from DIP PPI database. Training the SVM classifier on such a small corpus limits its generalisation capabilities (the SVM classifier achieved an F_1-score of 54.20% by using 10-fold cross validation) (Miwa *et al.*, 2009b). For example the sentence *“CD26 mediates NH(2) terminus processing of CCL22, leading to the production of CCL22 (3-69) and CCL22 (5-69) that do not interact with CCR4”* from Table 6.3 contains a negation and a coordination leading to the extraction of an interaction between *“CCL22”* and *“CCR4”* which is a false positive. Several examples of false negatives and false positives introduced by the pipeline are shown in Table 4.18 and Table 4.19, respectively.

Table 4.18: Example False Negatives

PubMed ID	Sentence	Comments	Source of Error(s)
9878398	“In particular, p38 was found to associate with itself to form a dimer, but also with p43, with the class I tRNA synthetases ArgRS and GlnRS, with the class II synthetases AspRS and LysRS, and with the bifunctional GluProRS.”	Abstract is an interaction abstracts and GNAT detects and normalizes both p38 and p43	PPI Extraction Module
11509571	“Tudor domain missense mutations, including one found in an SMA patient, impair the interaction between SMN and fibrillarin (as well as the common snRNP protein SmB).”	Abstract is an interaction abstracts and GNAT misses the entire protein name SmB	GNAT
1646816	“Treatment of this PCI-binding material with chondroitinase ABC, but not with chondritinase AC or heparitinase, abolished binding to PCI-Sepharose, confirming the glycosaminoglycan nature of this material and suggesting the involvement of dermatan sulfate in binding.”	Abstract is an interaction abstract and GNAT misses the entire protein name Sepharose	GNAT

Table 4.19: Example False Positives

PubMed ID	Sentence	Comments	Source of Error(s)
11907260	“Palmitoylation of tetraspanin proteins: modulation of CD151 lateral interactions, subcellular distribution, and integrin-dependent cell morphology.”	Abstract is an IA but the sentence is not an interaction sentence	PPI Extraction Module
12618216	“The N-terminal non-RGS domain of human regulator of G-protein signalling 1 contributes to its ability to inhibit pheromone receptor signalling in yeast.”	Abstract is not an interaction abstract and sentence is not definitive	IASEL and PPI Extraction Module
14517274	“To analyze how M protein allows evasion of phagocytosis, we used the M22 protein, which has features typical of many M proteins and has two well-characterized regions binding human plasma proteins: the hypervariable NH2-terminal region binds C4b-binding protein (C4BP) , which inhibits the classical pathway of complement activation; and adjacent semivariable region binds IgA-Fc.”	Abstract is an interaction abstract but GNAT misses some protein name parts of cb4-binding protein and assigns another GeneID to its abbreviation C4BP	GNAT and PPI Extraction Module
15067078	“CD26 mediates NH(2) terminus processing of CCL22, leading to the production of CCL22 (3-69) and CCL22 (5-69) that do not interact with CCR4.”	Abstract is an interaction abstract and sentence contains negation of an interaction	PPI Extraction Module

Performance of the text mining system developed is evaluated manually. For this purpose, a total of 100 sentences are selected randomly and 212 protein interaction pairs belonging to these sentences are analysed. A total number of 91 out of 212 pairs are identified as true positives. A total number of 80 out of 212 pairs are identified as false positives and the remaining 41 pairs are identified as false negatives. Overall, the performance of the pipeline is estimated at an F_1-score of 60.07% with 68.94% recall and 53.22% precision.

Chapter 5

IDENTIFICATION OF THE VARIABILITY IN ISOFORM INTERACTIONS

5.1 Definitions

This section describes the definitions that are introduced and used during the process of variability analysis in isoform interactions.

Following categorisation is used to distinguish CMTs (Clusters with multiple defined transcripts) according to the number of interacting isoforms in them:

(a) CMTs with Multiple Interacting Isoforms (CMT/MII): Clusters with at least two isoforms with known interactions.

(b) CMTs with a Single Interacting Isoform (CMT/SII): Clusters with only one interacting isoform.

(c) CMT has No Interacting Isoforms (CMT/NII): Clusters which contain no interacting isoform.

For a given CMT/MII, different interaction types can be distinguished between the isoforms:

(a) Shared Interaction (S): In this type of interaction several isoforms have the same interaction partner.

(b) Unique Interaction (U): In this type of interaction only one isoform interacts with a distinct partner.

Following categorisation is used to distinguish CMT/MIIs according to the different combinations of interactions contained in them:

(a) CMT/MII having unique interactions and external interaction partners (CMT/MII-eU): The interaction partner is external to the CMT/MII and is unique for the given isoforms

(b) CMT/MII having shared interactions and external interaction partners (CMT/MII-eS): The interaction partner is external to the CMT/MII, and is shared between isoforms.

(c) CMT/MII having both types of interactions and external interaction partners (CMT/MII-eB): Isoforms have both external unique and external shared interaction partners.

(d) CMT/MII having internal interaction partners (CMT/MII-i): Isoforms from the same CMT/MII could interact with one another.

5.2 Interaction Types

Isoforms exhibiting variability in their structures could also exhibit variability in their functions and thus in their interactions. The variability in their interactions serves as a significant indicator for the functional variability of the isoforms. Therefore, the distribution of isoform interactions contained in CMT clusters is analysed in order to assess the variability within their interactions and gain insights into their functional variability. For this purpose, firstly CMTs are categorised based on the number of interacting isoforms contained in them. Figure 5.1 illustrates this categorisation. If a CMT cluster contains at least two isoforms having interactions as shown in Figure 5.1.a then it is categorised as a CMT with Multiple Interacting Isoforms (CMT/MII). There are CMTs, where only one of their isoforms has interactions (Figure 5.1.b). Such CMTs are categorised as CMT with only a Single Interacting Isoform (CMT/SII). A CMT can contain isoforms without any interaction information. Such a cluster is categorised as a CMT with No Interacting Isoforms (CMT/NII), as shown in Figure 5.1.c.

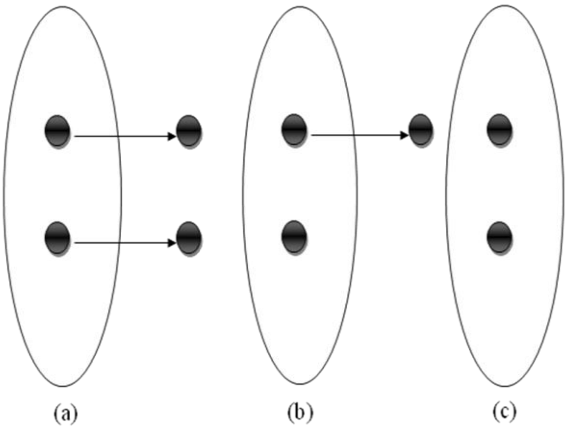


Figure 5.1: Categories of CMTs Based on the Number of Interacting Isoforms (a)CMT/MII (b)CMT/SII (c)CMT/NII (Definitions are provided in Secion 5.1)

The oval shapes represent CMTs. The black-filled circles in the CMTs represent isoforms while the ones outside the CMTs represent the interaction partners of the isoforms. Arrows represent interactions between the isoforms and their protein partners.

Secondly, in order to assess the variation in interactions, CMT/MIIs are categorised according to the interaction types of their isoforms (Figure 5.2). An isoform can have a *Shared Interaction* (S) if it has the same interaction partner with other isoforms. When only one isoform interacts with a distinct partner the interaction is termed as *Unique Interaction* (U). CMT/MIIs are grouped into four categories according to the different combinations of interactions. If the interaction partner is *external* to the CMT/MII and is *unique* for the given isoforms then the cluster is categorised as CMT/MII-eU (Figure 5.2.a). If the interaction partner is *external* to the CMT/MII, and it is *shared* between isoforms then the cluster is categorised as

CMT/MII-eS (Figure 5.2.a). Isoforms can have *both* external unique and *external* shared interaction partners. In such a case the cluster is categorised as CMT/MII-eB (Figure 5.2.c). Isoforms belonging to the same CMT/MII can have *internal* interactions. Such clusters are categorised as CMT/MII-i (Figure 5.2.d).

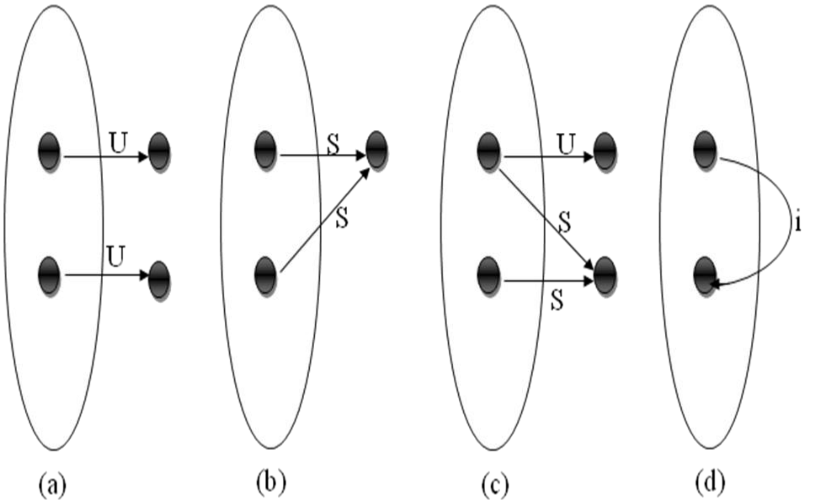


Figure 5.2: Categories of CMT/MIIs Based on the Interaction Types of Isoforms (a)CMT/MII-eU (b)CMT/MII-eS (c)CMT/MII-eB (d)CMT/MII-i (Definitions are provided in Secion 5.1)

The oval shapes represent CMTs. The black-filled circles in the CMTs represent isoforms while the ones outside the CMTs represent the interaction partners of the isoforms. Arrows represent interactions between the isoforms and their protein partners. U: Unique interaction, S: Shared interaction, i: Internal interaction.

5.3 Interaction Variability Analysis

Isoforms of a given gene can share the same function, they can show minimal functional differences, or they can have opposite functions (Stamm *et al.*, 2005). Changes in the functional behaviours of the isoforms could be expected to be reflected in their interactions and consequently in their interaction partners. More specifically, when isoforms have unique interactions then it is expected that they could exhibit different functional behaviours. On the other hand, when the isoforms interact with the same partners, it is expected to observe smaller functional variability. Therefore, distributions of shared and unique interactions of the isoforms are documented in order to gain insight into their functional variability.

Literature-based distribution of interaction pairs belonging to CMTs are depicted in Figure 5.3. No interaction information is found for a total of 164 CMTs (CMT/NII), while at least one interaction pair is found for 282 CMTs. A total of 194 out of 282 CMTs contain only one interacting isoform (CMT/SII), while the remaining 88 CMTs contain multiple interacting isoforms (CMT/MII). A total of 12 clusters include isoforms having intra-cluster interactions with possible external interactions (CMT/MII-i). Analysis on the clusters having external interactions reveals one CMT/MII-eS having only shared interactions, 70 CMT-MII-eU having only unique interactions and 5 CMT/MII-eB clusters having both shared and unique interactions. Hence, 87 out of 88 (99%) clusters analysed exhibit unique interactions. This is a significant finding in that it indicates that the human isoforms in the CMTs exhibit high variability in the selection of their interaction partners.

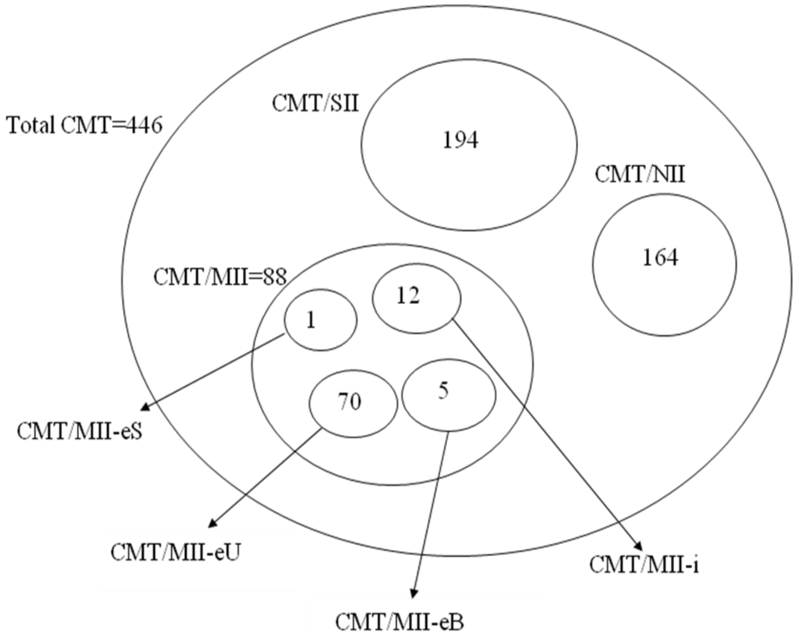


Figure 5.3: Interaction Variation Analysis of CMTs Based on Automated Literature Mining

(see section 5.1 for abbreviations)

The analysis shows that CMT/SIIs are more frequent than the other CMT types which could be due to for several reasons. Major isoforms are more frequently reported in the literature compared to the minor ones given that such isoforms are more commonly studied in experiments. In addition, depending on the tissues as well as developmental stage specificity of alternative splicing, mRNA or EST sequences of some isoforms may have not been available during the construction phase of the HumanSDB3. Hence such isoforms are not included in the study. In addition, some interaction data could be missed by the text mining system used.

Further investigation is carried out on 5 CMT/MII-eB and 10 out of 12 CMT/MII-i clusters in order to document the distribution of unique versus shared interactions (Figure 5.3, Table 5.1). The remaining two CMT/MII-i clusters contain only unique interactions. Hence, they are excluded from the analysis. All 15 clusters are categorised into two classes based on the total number of isoforms contained in them (>2 or 2 isoforms). The average ratio of unique versus shared interactions for each category is measured as above 5.60 for each category.

Table 5.1: Statistics on CMT/MII-eB and CMT/MII-i with Shared and Unique Interactions Based on the Literature Analysis

Iso/CMT*	HumanSDB3 Cluster ID	CMT/MII Type	Nof Iso	Nof S	Nof U	Nof S/Iso	Nof U/Iso	U/S	Avg U/S
2	Hs.3.chr6p.16643	CMT/MII-i	2	22	38	11	19	1.73	5.68
	Hs.3.chr17p.8013	CMT/MII-i	2	20	50	10	25	2.5	
	Hs.3.chr11p.3558	CMT/MII-eB	2	12	24	6	12	2	
	Hs.3.chr6n.17144	CMT/MII-i	2	10	46	5	23	4.6	
	Hs.3.chr1n.278	CMT/MII-i	2	8	35	4	17.5	4.38	
	Hs.3.chr5n.15390	CMT/MII-eB	2	8	52	4	26	6.5	
	Hs.3.chr14p.5840	CMT/MII-i	2	4	25	2	12.5	6.25	
	Hs.3.chr12p.4823	CMT/MII-i	2	2	40	1	20	20	
	Hs.3.chr17n.8529	CMT/MII-eB	2	2	6	1	3	3	
	Hs.3.chr19p.9432	CMT/MII-eB	2	2	19	1	9.5	9.5	
	Hs.3.chr22p.13094	CMT/MII-i	2	2	4	1	2	2	
>2	Hs.3.chr6p.16595	CMT/MII-i	3	14	38	4.67	12.67	2.71	5.62
	Hs.3.chr3p.13906	CMT/MII-eB	3	2	11	0.67	3.67	5.5	
	Hs.3.chr17n.8527	CMT/MII-i	4	5	15	1.25	3.75	3	
	Hs.3.chr17n.8355	CMT/MII-i	5	4	45	0.8	9	11.25	

*CMT is either a CMT/MII-eB or a CMT/MII-i, Iso:Isoforms, Nof:Number of, Avg:Average, S:Shared interactions, U:Unique interactions (see section 5.1 for abbreviations)

5.4 Validation of the Text Mining Results Against Public PPI DBs

The literature-based interaction variability analysis results are validated against the results obtained based on the publicly available PPI data from PINA (Wu *et al.*, 2009). PINA includes binary interactions from six major PPI DBs: IntAct (Hermjakob i., 2004), MINT (Zanzoni *et al.*, 2002), BioGRID (Stark *et al.*, 2006), DIP (Xenarios *et al.*, 2000), HPRD (Keshava *et al.*, 2009) and MIPS/MPact (Pagel *et al.*, 2005). In contrast to many other PPI DBs, PINA does not include either complex or genetic interactions. Therefore, it is suitable for the purpose of this study. All self and non-human interactions are removed from the PINA resulting in a comprehensive PPI set including 58,221 interactions between 11,856 different human proteins.

In order to validate the literature based findings, PINA is searched for the PPI data linked to the isoforms from all CMTs. For this purpose, Entrez Gene DB IDs of the isoforms are converted to their corresponding Uniprot accession numbers, given that protein partners in PINA are identified with their Uniprot accession numbers. Uniprot mapping system is utilised for the ID conversion process (http://www.uniprot.org/).

Distribution of the interaction pairs gathered from PINA in the CMTs is shown in Figure 5.4. A total of 101 out of 446 CMTs contain no interaction information for any of their isoforms (CMT/NII). A total number of 187 CMTs contain only one interacting isoform (CMT/SII), while the remaining 158 CMTs contain multiple interacting isoforms (CMT/MII). Further investigation based on the interaction types of isoforms from the CMT/MIIs shows that there are 4 CMT/MII-i, 9 CMT/MII-eS, 117 CMT-MII-eU and 28 CMT/MII-eB. The analysis reveals that majority of the CMT/MIIs (149 out of 158, 94.30%) exhibit variability in interactions. This finding serves as a significant indicator for the generation of interactome diversity due to transcript diversity.

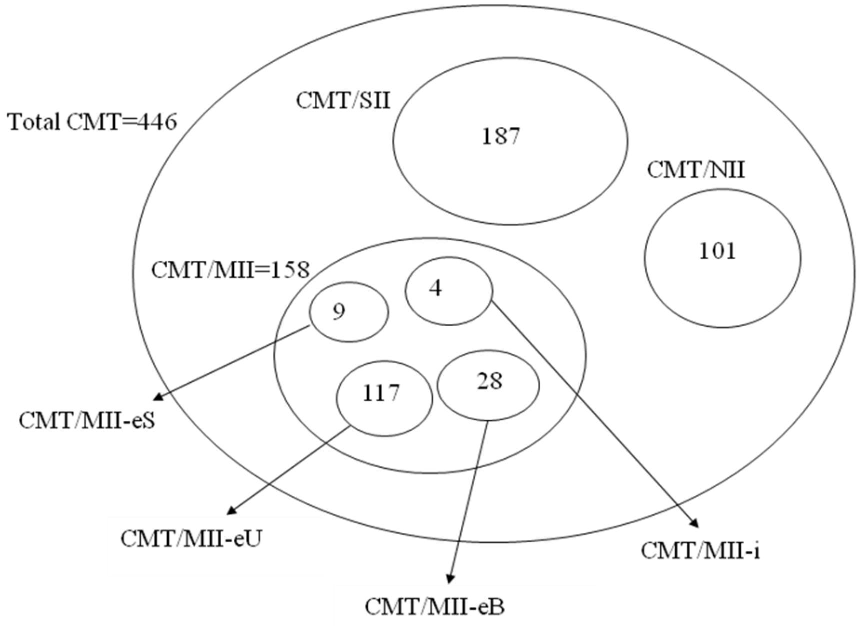


Figure 5.4: Interaction Variation Analysis of CMTs Based on PINA

(see section 5.1 for abbreviations)

Distributions of unique versus shared interactions, for a total of 30 clusters (28 CMT/MII-eB and 2 CMT/MII-i) which include both types of interactions, are documented in Table 5.2. The remaining two CMT/MII-i clusters are excluded, since they contain unique interactions only. The clusters are categorised according to the total number of interacting isoforms contained in them and the average ratio of unique versus shared interactions which are documented for each category. Clusters including two isoforms have slightly higher average ratio (8.82) than the average ratio obtained for the clusters including more than two isoforms (7.53). The average ratio of unique versus shared interactions calculated by using the gathered PPI data is slightly higher than the values calculated based on the literature analysis since the mapping procedure yields more a finely grained PPI dataset. It is worth to note that the distribution of the isoform interaction types in PINA is in agreement with the findings obtained through the literature analysis.

The interaction variability analysis results show that the isoforms are specialised towards selecting unique interaction partners and thus they are likely to be involved in different protein interaction networks. This indicates that the diversity introduced into isoform interactions by the transcript diversity mechanisms is potentially significant. In addition, given the different interactions of isoforms, it is expected that the isoforms exhibit different biological functions and thus they are likely to be involved in different molecular pathways.

Table 5.2: Statistics on CMT/MII-eB and CMT/MII-i with Shared and Unique Interactions Based on PINA

Iso/CMT*	**Cluster ID**	**CMT Type**	**Nof Iso**	**Nof S**	**Nof U**	**Nof S/Iso**	**Nof U/Iso**	**U/S**	**Avg U/S**
2	Hs.3.chr11p.3558	CMT/MII-eB	2	40	12	20	6	0.3	8.82
	Hs.3.chr17n.8529	CMT/MII-i	2	22	54	11	27	2.45	
	Hs.3.chr5n.15390	CMT/MII-eB	2	18	71	9	35.5	3.94	
	Hs.3.chr6p.16643	CMT/MII-eB	2	10	13	5	6.5	1.3	
	Hs.3.chr12n.4463	CMT/MII-eB	2	8	6	4	3	0.75	
	Hs.3.chr19n.10450	CMT/MII-eB	2	6	25	3	12.5	4.17	
	Hs.3.chr6n.17144	CMT/MII-eB	2	6	26	3	13	4.33	
	Hs.3.chr17n.8585	CMT/MII-eB	2	4	4	2	2	1	
	Hs.3.chr1n.278	CMT/MII-eB	2	4	10	2	5	2.5	
	Hs.3.chr6n.17040	CMT/MII-eB	2	4	13	2	6.5	3.25	
	Hs.3.chr11n.3142	CMT/MII-eB	2	2	48	1	24	24	
	Hs.3.chr17p.8013	CMT/MII-eB	2	2	9	1	4.5	4.5	
	Hs.3.chr17p.8043	CMT/MII-eB	2	2	43	1	21.5	21.5	
	Hs.3.chr1n.361	CMT/MII-eB	2	2	134	1	67	67	
	Hs.3.chr2p.10772	CMT/MII-eB	2	2	4	1	2	2	
	Hs.3.chr4p.14617	CMT/MII-eB	2	2	11	1	5.5	5.5	
	Hs.3.chr4p.14694	CMT/MII-eB	2	2	3	1	1.5	1.5	
>2	Hs.3.chr16p.7233	CMT/MII-eB	3	27	2	9	0.67	0.07	7.53
	Hs.3.chr15p.6760	CMT/MII-eB	3	8	18	2.67	6	2.25	
	Hs.3.chr3p.13906	CMT/MII-eB	3	8	57	2.67	19	7.13	
	Hs.3.chr17n.8437	CMT/MII-eB	3	6	8	2	2.67	1.33	
	Hs.3.chr11n.3383	CMT/MII-i	3	4	10	1.33	3.33	2.5	
	Hs.3.chr17n.8754	CMT/MII-eB	3	2	64	0.67	21.33	32	
	Hs.3.chr6p.16595	CMT/MII-eB	3	2	15	0.67	5	7.5	
	Hs.3.chr9n.19822	CMT/MII-eB	3	2	59	0.67	19.67	29.5	
	Hs.3.chr12n.4311	CMT/MII-eB	4	6	14	1.5	3.5	2.33	
	Hs.3.chr17n.8527	CMT/MII-eB	4	2	13	0.5	3.25	6.5	
	Hs.3.chr17n.8355	CMT/MII-eB	5	16	81	3.2	16.2	5.06	
	Hs.3.chr1p.1548	CMT/MII-eB	6	18	2	3	0.33	0.11	
	Hs.3.chr5p.15887	CMT/MII-eB	14	12	19	0.86	1.36	1.58	

*CMT is either a CMT/MII-eB or a CMT/MII-i cluster, Iso:Isoforms, Nof:Number of, Avg:Average, S:Shared interactions, U:Unique interactions (see section 5.1 for abbreviations)

Chapter 6

TBIID: TRANSCRIPT BASED ISOFORM INTERACTION DATABASE

6.1 TBIID in Comparison to Public PPI DBs

The PPI data gathered from the literature throughout this study is imported into a new database called TBIID. This databse contains the protein interactions of human protein isoforms. TBIID consists of 31,819 interactions between 7,161 distinct proteins out of which 5,615 are identified as being an isoform linked to either CSTs or CMTs of HumanSDB3. There are a total number of 1,540 interactions between 1,226 different proteins belonging to CMT/MII clusters exhibiting interaction variation (i.e. CMT/MII-eB, CMT/MII-eU and CMT/MII-i). 994 out of 1,226 proteins have a reference DT from HumanSBD3.

Here, TBIID content is compared against PINA (Wu *et al.*, 2009), in terms of the number of overlapping proteins and interactions are documented as shown in Figure 6.1. There are a total number of 4,944 (69.04%) overlapping proteins and 2,863 (9.00%) overlapping interactions between TBIID and PINA. The overlapping numbers are calculated as 927 (75.61%) and 141 (9.16%) for proteins and interactions respectively, when only the CMT/MIIs exhibiting interaction variation from TBIID are considered. These results show that the content of TBIID is complementary to the existing PPI DBs constituting PINA. Indeed, this result is expected given that the two PPI DBs exploits different resources and uses different methods to gather the PPI data.

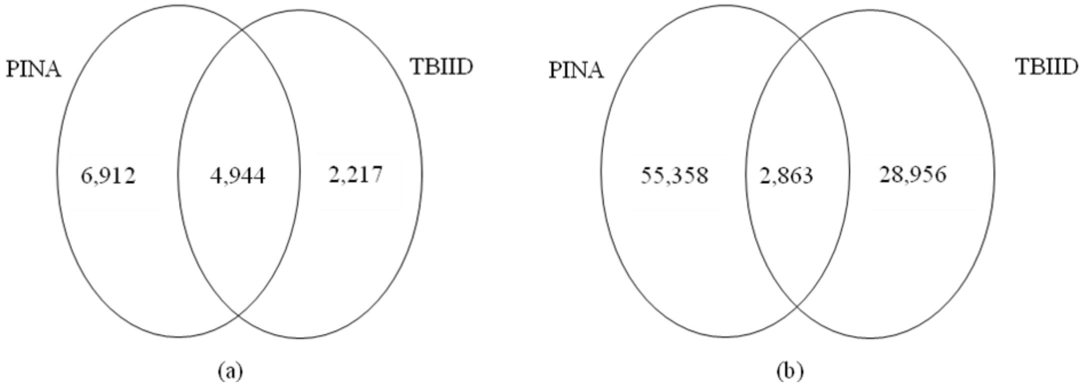


Figure 6.1: Venn Diagrams Showing Overlaps between PINA and TBIID
(a) Protein overlaps (b) Interaction overlaps

Analysis of the content of PINA (Table 6.1) shows that the majority of the interaction pairs (33,725 pairs) corresponding to 57.92% is reported by only one of the PPI DBs constituting PINA. Total numbers of 16,086 (27.63%), 5,182 (8.90%) and 3,102 (5.33%) interaction pairs are shared by two, three and four DBs, respectively. The remaining 126 (0.22%) interaction pairs are shared by five DBs while there is no interaction pair shared by all 6 DBs in PINA. Similarly, overlapping interaction pairs in TBIID are mainly reported in one (1,198 pairs, 41.84%) or two DBs (1,210 pairs, 42.26%). Total numbers of 338 (11.81%), 97 (3.39%) and 20 (0.70%) overlapping pairs are reported in three, four and five DBs, respectively. It is important to note that 85.55% of interactions in PINA are reported only in one or two DBs. This number is 84.1% when TBIID is considered.

Table 6.1: Interaction Pair Distribution in PINA and Overlapping Sets

Dataset	Number of PPI DBs containing the interaction pairs				
	1	2	3	4	5
PINA	33,725 (57.92%)	16,086 (27.63%)	5,182 (8.90%)	3,102 (5.33%)	126 (0.22%)
TBIID*	1,198 (41.84%)	1,210 (42.26%)	338 (11.81%)	97 (3.39%)	20 (0.70%)

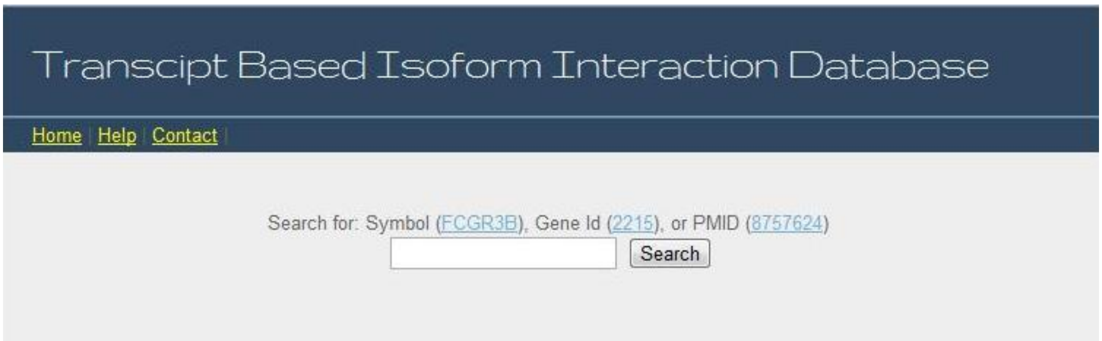
*Overlapping interactions with PINA only

These results indicate that the current major PPI DBs cover different sections of the interactome. Previous comprehensive analyses on the major PPI databases also report that overlapping portions between these DBs are very low given that often different extraction methods, curation methods and publication records are used to construct them (Cusick *et al*., 2009; Mathivanan *et al.*, 2006; Prieto and Rivas, 2006). These reasons could also lead to the low number of overlapping interaction pairs (2,863 pairs, 9.00%) between PINA and TBIID. Hence, publication records used to construct PINA and TBIID are analysed. Results show that there are a total number of 19,372 unique PubMed records referred by the interaction pairs in PINA. 8,326 (42.98%) of these records are found to be overlapping in the retrieved relevant abstract set of isoforms (4,083,094 abstracts) while 7,333 (37.85%) of them are found in the interaction abstract set (205,270 PubMed records). Results reveal that there is a high rate of discrepancy between the PubMed record sets used to generate TBIID and PINA. The portion of the source text used to gather the PPI data could be another important factor playing role in low percent of overlap. The PPI information contained in TBIID is extracted from the freely available Medline abstracts only. On the other hand, curators often facilitate full text articles leading to higher rates of PPIs for building literature-curated PPI databases, where some of them are included in PINA (Krallinger *et al*., 2008). In addition, some interactions could be missed during the building phase of the TBIID generation, given that it is developed by using automated text mining methods. Another factor is a 4% error rate introduced during the protein ID conversion process. Entrez Gene IDs of proteins from TBIID have to be converted to their corresponding Uniprot accession numbers in order to gather their interactions from PINA. Finally, research conducted in the scope of this study is based on the transcript data linked to the variant clusters from HumanSDB3 where a large portion (~81.00%) but not the complete genome is considered.

6.2 TBIID Web-Interface

A web-interface which enables users to analyse interaction data contained in TBIID is developed. This data in TBIID is linked to HumanSDB3. Hence, TBIID serves as a link between interactions of protein isoforms and their transcriptomic data. TBIID content is publicly accessible through http://tbiid.emu.edu.tr.

Search for interactions of a given protein can be done through a query system embedded into the web-interface (Figure 6.2). The query system is invoked by submitting the protein’s Entrez Gene ID or official symbol listed in Entrez Gene DB. In case of search based on a given symbol, the system allows several variations of the symbol. For example, it is not sensitive to upper/lower case letters, and spaces between digits and the letters are ignored. Users can also search the content of TBIID for interactions extracted from a particular PubMed record by submitting its PubMed ID to the system.


Transcipt Based Isoform Interaction Database

Search for: Symbol (FCGR3B), Gene Id (2215), or PMID (8757624)

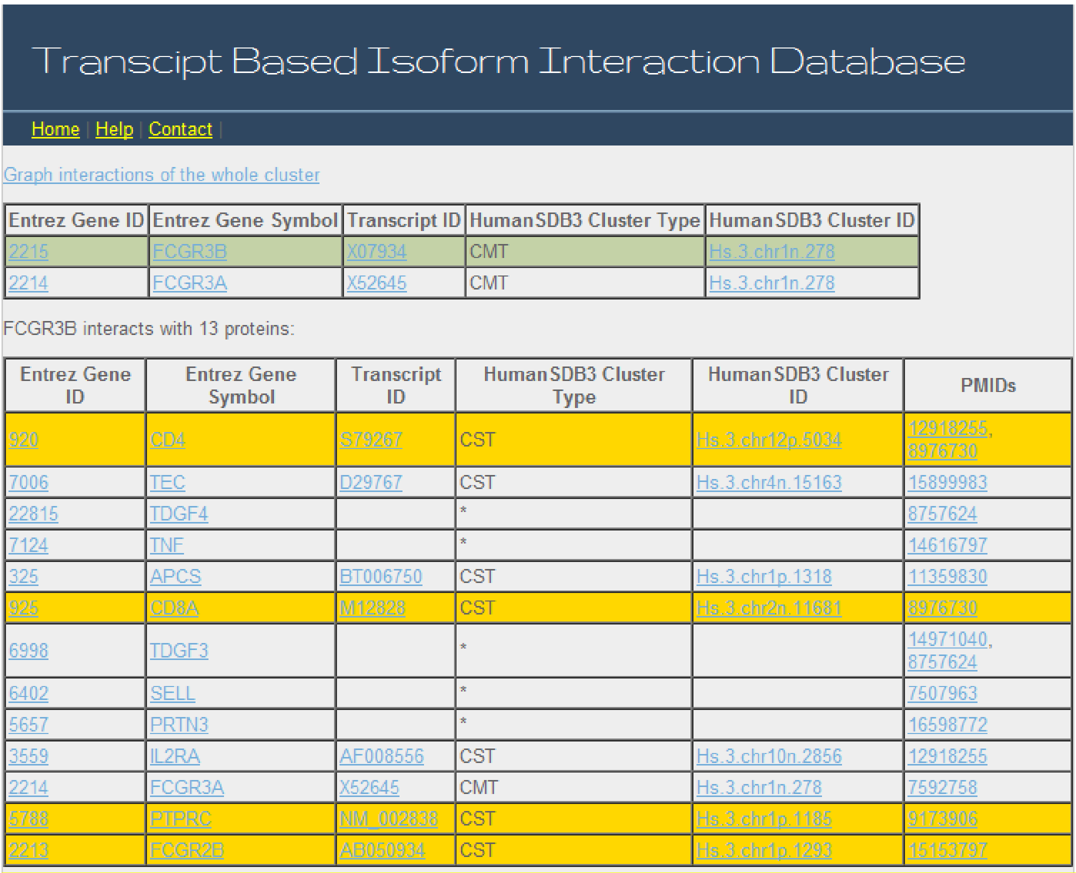
Figure 6.2: Query Interface of TBIID

Usage of the web-interface as well as the benefit of TBIID is demonstrated below by using an example CMT (HumanSDB3 Cluster ID: Hs.3.chr1n.278) from HumanSDB3. The CMT contains two DTs for human IgG Fc Receptor III (FCGR3). The DTs code two distinct but 97% identical allelic isoforms, namely FCGR3A and FCGR3B (Rogers and Scinicariello, 2006).

Figure 6.3 is a screenshot from TBIID illustrating the retrieval of the interactions of FCGR3B along with its isoform FCGR3A data by using its official symbol from TBIID. The table on top in the figure contains information on the isoforms linked to the CMT. The queried isoform is highlighted with green colour. In this table, the isoforms are linked to Nucleotide DB of NCBI (provides sequence information) and HumanSDB3 through their transcript IDs and HumanSDB3 cluster IDs respectively. They are also linked to the Entrez Gene DB through their Gene DB IDs. Entrez Gene DB provides some external information on proteins such as functions (based on GO concepts) and metabolic pathways that they involve in. Such information is important for understanding isoform interactions.

The lower table in the figure provides information on interaction partners of the isoform FCGR3B (Figure 6.3). Interaction data contained in TBIID is referenced to the source text from PubMed DB enabling manual analysis of the questionable records. Shared interactions of the isoforms are shown in yellow. Interactions of the protein partners can be retrieved from TBIID content by clicking on their symbols.

The utility of TBIID could be demonstrated by using FCGR3 isoforms. Examining the GO data provided in the Entrez Gene DB through the web-interface of TBIID reveals that FCGR3A and FCGR3B isoforms share several molecular functions (Ig binding and receptor activity) and are involved in immune response processes. Therefore it could be expected that those isoforms have shared interactions. Analysing the content of PINA reveals that there are 12 interaction partners for FCGR3A (APCS, CD247, CD38, CD4, FCER1G, GP6, FCGR1A, IGHG, LCK, PTPRC, SHC1, ZAP70) and only 4 partners for FCGR3B (APCS, IGHG1, M(2)21AB, Myb). According to PINA, these isoforms have two shared interaction partners (APRCS and IGHG1). On the other hand, by utilising TBIID, it is possible to expose other interesting interaction partners of the isoforms. For example, PTPRC (Entez Gene ID: 5788) is reported by TBIID as a shared interaction partner which is not reported by any of the major PPI DBs constituting PINA. This shared interaction is supported by evidences from the literature (see PubMed IDs: 8157290 and 9173906). In addition, TBIID reports TEC (see Entrez Gene ID:7006, PubMed ID: 15899983) as another unique interaction partner for FCGR3B isoform. These findings indicate that TBIID contains valuable data, which is complementary to the existing major PPI databases for analysing differential interactions of the isoforms. It is important to uncover differential interactions of isoforms for a good understanding of different biological processes that the isoforms are involved in.



Transcipt Based Isoform Interaction Database

Graph interactions of the whole cluster

Entrez Gene ID	Entrez Gene Symbol	Transcript ID	HumanSDB3 Cluster Type	HumanSDB3 Cluster ID
2215	FCGR3B	X07934	CMT	Hs.3.chr1n.278
2214	FCGR3A	X52645	CMT	Hs.3.chr1n.278

FCGR3B interacts with 13 proteins:

Entrez Gene ID	Entrez Gene Symbol	Transcript ID	HumanSDB3 Cluster Type	HumanSDB3 Cluster ID	PMIDs
920	CD4	S79267	CST	Hs.3.chr12p.5034	12918255, 8976730
7006	TEC	D29767	CST	Hs.3.chr4n.15163	15899983
22815	TDGF4		*		8757624
7124	TNF		*		14616797
325	APCS	BT006750	CST	Hs.3.chr1p.1318	11359830
925	CD8A	M12828	CST	Hs.3.chr2n.11681	8976730
6998	TDGF3		*		14971040, 8757624
6402	SELL		*		7507963
5657	PRTN3		*		16598772
3559	IL2RA	AF008556	CST	Hs.3.chr10n.2856	12918255
2214	FCGR3A	X52645	CMT	Hs.3.chr1n.278	7592758
5788	PTPRC	NM_002838	CST	Hs.3.chr1p.1185	9173906
2213	FCGR2B	AB050934	CST	Hs.3.chr1p.1293	15153797

Figure 6.3: Screenshot Representing the Content

ID: Identifier, PMID: PubMed Identifier

It is also possible to visualize interactions of the queried isoform graphically as shown in Figure 6.4. In this case, interactions of the isoform will be listed along with the interactions of other isoform(s) linked to the same cluster. In Figure 6.4, FCGR3B and its unique interaction partners are shown in green tones while FCGR3A and its unique interaction partners are shown in blue tones. Shared interaction partners are shown in yellow. This mode of visualization enables the TBIID users to simultaneously analyse the shared and unique interactions of the isoforms. This functional feature brings uniqueness to TBIID.

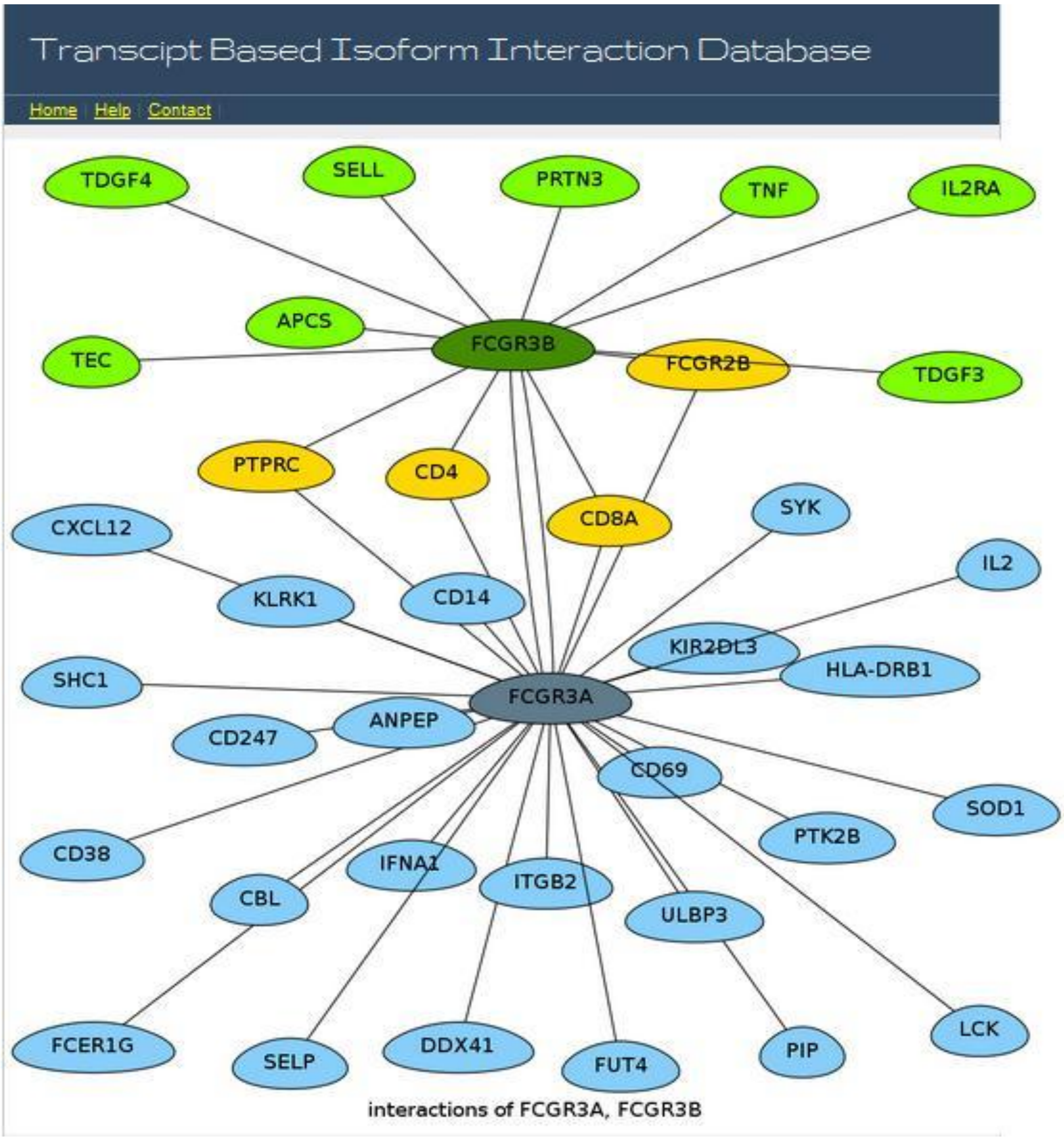


Figure 6.4: TBIID Screenshot Graphically Showing Visualization of Isoform Interactions

Chapter 7

DISCUSSION, CONCLUSION AND FUTURE WORK

7.1 Discussion and Concluding Remarks

In this thesis, (i) TBIID is presented as a new database covering PPI data on human protein isoforms and (ii) its content is utilised to investigate the variability in the isoform interactions. The biomedical literature is exploited automatically to gather protein interactions involving isoforms linked to clustered transcript data from HumanSDB3. For this purpose, the transcript data from clusters of HumanSDB3, which exhibit alternative splicing and which represent a significant portion of the human genome (~81%) are analysed. DTs within each cluster are identified and a rich STS by using Gene DB, Swissprot DB and synonym generation is compiled for each DT. Relevant abstracts are retrieved from the PubMed DB by using these STSs.

A state-of-the art performing SVM classifier is trained on the BioCreative-II IAS corpus with a novel set of features and used to select those abstracts which are likely to contain PPI information. This classifier achieves an F_1-score of 81.31% on the BioCreative-II IAS dataset, which is the second best performance reported in the literature to the best of our knowledge. Protein interactions involving isoforms from the selected abstracts are extracted by using another SVM trained by utilising features from syntactic parsers on the AIMed corpus. The performance of this classifier is measured at an F_1-score of 54.20% with a precision of 60.77% and a recall of 49.50% by using 10-fold cross validation. The classifier achieves at state-of-the-art level and has a high precision which is desirable for the purpose of the study. Nevertheless, the overall performance of the text mining pipeline is analysed manually on a random set of extracted interaction protein pairs. The performance is estimated at an F_1-score of 60.07% with 68.94% recall and 53.22% precision which is considered to be high enough to carry out further analysis on the extracted data.

The content of TBIID is compared against a comprehensive public PPI resource, PINA. A total number of 4,944 (69.04%) overlapping proteins and a total number of 2,863 (9.00%) overlapping interactions between the two resources are identified. Results are in parallel with the previous studies which highlight low overlap rates between the public PPI DBs due to the usage of different extraction and curation methods as well as publication records to generate them.

A large scale PPI analysis is applied on TBIID content to measure the variability in the isoform interactions. For this purpose, distributions of shared and unique interaction partners of the isoforms linked to CMT/MIIs are analysed. Results reveal that majority of the proteins coded by the transcripts isoforms in CMT/MIIs (99%) exhibit variation in their protein interactions. This is a significant finding in that it sheds light on how alternative splicing and possibly other transcript diversity mechanisms introduce variation in protein interactions. In addition, quantitative analysis on CMT/MIIs indicates that isoforms tend to interact with the same partners with a ratio of 1/5 only i.e isoforms are specialised towards forming unique interactions rather than shared ones. Importantly, with these results, it is quantitatively demonstrated that alternative splicing and other transcript diversity mechanisms generate transcript diversity, which generates proteome diversity, which leads to interactome diversity. Similar findings are obtained by using PINA which contains PPI data from major publicly available PPI DBs. These results indicate that the isoforms tend to form unique interactions, and are possibly involved in different interaction networks thus potentially achieving different biological functions. It is important to note that the obtained interaction variability and validation analyses results depend on the text mining approaches used in the study, available PPI data in the public resources as well as the cluster organisation in HumanSDB3. Use of a different cluster organisation, a set of public PPI data and text mining methods could yield different results complementing TBIID. Nevertheless, based on the methods used, current available data and the current results obtained in this study, it can be concluded that transcript diversity is a widespread process leading diverse proteomes and presents a potential to generate a significantly diverse interactomes.

TBIID is the first DB covering comprehensive PPI information linked to human protein isoforms. It serves as a bridge between isoform interactions and transcript diversity. Hence, documentation of and further analyses of potential differential interactions of protein isoforms from TBIID will help understand the effect of alternative splicing and also possibly other transcript diversity mechanisms on the human proteome and interactome at a large scale.

In this study, for the first time, a large scale analysis is applied on TBIID content to quantify the variability in the isoform interactions. Although, CMTs could also contain other kinds of isoforms in addition to alternative splicing variants, it is likely that source of the interaction variability is alternative splicing given that HumanSDB3 variant clusters contain transcripts exhibiting alternative splicing events. Nevertheless, further sequence-based detailed investigation on each CMT would identify the exact transcript diversity source and/or exact type of alternative splicing implicated in each differential isoform interaction. TBIID can be utilised towards such further experimental investigation on the CMTs and their unique as well as shared interactions. In addition to that, the developed text mining tools serves as practical tools in PPI related biomedical text mining tasks.

7.2 Future Work

In the future, the study may be extended to include further investigation on the functional variability of the protein isoforms. A possible variability analysis can be based on the distribution of functional annotations on the basis of Gene Ontology concepts, especially biological processes and molecular functions. Interaction partners of isoforms exhibiting functional diversity are known as good potential targets for pharmacological interventions (Da Cruz e Silva *et al.*, 2004). Therefore, data on interactions and functions of the isoforms play an important role in designing drugs specific to isoforms. Such drugs offer therapeutic advantages like preventing disease progress over their non-specific types given that specific isoforms could be involved in different biological pathways by playing different functional roles.

It is also possible to extend the study to gather disease-related information associated with the isofoms from CMTs by utilising the biomedical literature. This information is important for a good understanding of the transcript diversity mechanism, aberrant isoforms and their implications in abnormal protein functions. In addition, such information could serve as an important resource for molecular therapies.

Alternative splicing is a tissue specific cellular mechanism. Indeed, this information is important for a good understanding of alternative splicing events. Therefore, content of TBIID may be expanded to include tissue specificity information for the isoforms and link this information to their interaction specificity.

The transcript data provided in HumanSDB3 dates back to 2004 and was organised by using UCSC human genome version 17 (hg.17). Currently, Scripps Genome Center directs its efforts for generation of an updated version of HumanSDB by using recent transcript data and latest version of the human genome (hg.19). This will lead to a different cluster organization and clusters will contain more up to date transcript data compared to HumanSDB3. The developed text mining pipeline may be employed to analyse the new human alternative splicing database which would lead to generation of an updatedversion of TBIID.

TBIID is constructed based on an automated analysis of the biomedical literature indeed relying on existing knowledge from freely available PubMed abstracts. More PPI data can be extracted by searching full-length papers. However, this can be realised to a limited extend due to the copyright limitations of full-length articles. Nevertheless, this may be introduced to the database with forthcoming releases.

Furthermore, the study presented here may be expanded to several different organisms such as mouse and rat given that splicing DBs are available for several different transcriptomes (Taneri *et al.*, 2005; Taneri *et al.*, 2011). This would enable scientist to carry out species-specific as well as comparative studies on transcript diversity and isoform interactions.Such organisms have proximity since there are genes, which are evolutionarily conserved between them. For example, human shares at least 80% of genes with mouse (Mouse Genome Sequencing Consortium, 2002)

Therefore, coverage of each individual species specific database may be expanded based on the conserved genes and thus likely conserved interactions.

REFERENCES

Abi-Haidar, A., Kaur, J., Maguitman, A., Radivojac, P., Retchsteiner, A., Verspoor, K., Wang, Z., Rocha, L.M. (2007). Uncovering Protein-Protein Interactions in the Bibliome. In proceedings of the *Second BioCreative Challenge Workshop*, 247-256.

Agarwal, S., Liu, F., Li, Z., Yu, H. (2010). Machine learning-based approaches for BioCreative-III tasks. In proceedings of *BioCreative-III*, Maryland, USA, 42-47.

Airola, A., Pyysalo, S., Björne, J., Pahikkala, T., Ginter, F., Salakoski, T. (2008). All-path graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. *BCM Bioinformatics*, 9(Suppl II):S2.

Albert, S., Gaudan, S., Knigge, H., Raetsch, A., Delgado, A., Huhse, B., Kirsch, H., Albers, M., Rebholz-Schuhmann, D., Koegl, M. (2003). Computer assisted generation of a protein-interaction database for nuclear receptors. Molecular Endocrinology, 17, 1555-1567.

Andrew, B.C., Shepherd, A.J. (2007). Benchmarking natural-language parsers for biological applications using dependency graphs. *BCM Bioinformatics*, 8, 24-41.

Baek, D., Green, P. (2005). Sequence conservation, relative isoform frequencies, and nonsense-mediated decay in evolutionarily conserved alternative splicing. *PNAS*, 102(36), 12813-12818.

Baeza-Yates, R., Ribeiro-Neto, B. (2011). Modern Information Retrieval: The concepts and technology behind search. (Edt. Addison Wesley), Second Edition.

Bathesda, MD. (2010). Entrez Programming Utilities Help. National Center for Biotechnology Information (US).

Black, D.L. (2000). Protein Diversity from Alternative Splicing: A Challenge for Bioinformatics and Post-Genome Biology. *Cell*, 103, 367–370.

Black, D. L. (2003). Mechanisms of alternative pre-messenger RNA splicing. *Annual Review Biochemistry*, 72, 291–336.

Blaschke, C., Andrade, M.A., Ouzounis, C., Valencia, A. (1999). Automatic extraction of biological information from scientific text: Protein-protein interactions. In proceedings of the *7th Conference on Intelligent Systems in Molecular Biology*, 60-67.

Burges C.J.B. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining in Knowledge Discovery*, 2, 121-167.

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(Suppl 1), D267-D270.

Bradley, A.P (1997). The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms. *Pattern Recognition*, 30(7), 1145.

Browne, F., Zheng, H., Wang, H., Azuaje, F. (2010). From Experimental Approaches to Computational Techniques: A Review on the Prediction of Protein-Protein Interactions. *Advances in Artificial Intelligence*, doi:10.1155/2010/924529

Bunescu, R., Ge, R., Kate, R.J., Marcotte, E.M., Mooney, R.J., Ramani, A.K., Wong, Y.W. (2005). Comparative expeiments on learning information extractors for proteins and their interactions, *Artificial Intelligence in Medicine*, 33, 139-155.

Campbell, A.M., Heyer, L.J. (2004). Discovering Genomics, Proteomics, & Bioinformatics. *Pearson Education*, New Delhi.

Cartegni, L., Chew, S.L., Krainer, A.R (2002). Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nature Review Genetics*, 3(4), 285-98.

Castle, J.C., Zhang, C., Shah, J.K., Kulkarni, A.V., Kalsotra, A., Cooper, T.A., Johnson, J.M. (2008). Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. *Nature Genetics*, 40, 1416–1425.

Celotto, A.M., Graveley, B.R. (2001). Alternative splicing of the Drosophila Dscam pre-mRNA is both temporally and spatially regulated. *Genetics*, 159(2), 599-608.

Chen, H., Fuller, S.S., Friedman, C.P., Hersh, W. (2005). *Medical Informatics: Knowledge Management and Data Mining in Biomedicine Series: Integrated Series in Information Systems,* Springer-Verlag, New York.

Chen, J., Aronow, B.J., Jegga, A.G. (2009). Disease candidate gene identification and prioritization using protein interaction networks. *BMC Bioinformatics,* 10:73.

Chen, Z., Gore, B.B., Long, H., Ma, L., Tessier-Lavign, M. (2008). Alternative Splicing of the Robo3 Axon Guidance Receptor Governs the Midline Switch from Attraction to Repulsion. *Neuron*, 58, 325-332.

Cheng, C.Y., Hsu, F.R., Tang, C.Y. (2008) Extracting Alternative Splicing Information from Captions and Abstracts Using Natural Language Processing. In proceedings of the *IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing (SUTC2008),* Taichung.

Chester, A., Scott, J., Anant, S., and Navaratnam, N. (2000). RNA editing: cytidine to uridine conversion in apolipoprotein B mRNA. *Biochimica et Biophysica Acta-Gene Structure and Expression*, 1494(1-2), 1-13.

Cho, S., Sung, G.P., Lee, D.H. Park, B.C. (2004). Protein-protein Interaction Networks: from Interactions to Networks, *Journal of Biochemistry and Molecular Biology*, 37(1), 45-52.

Chothia, C., Gough, J., Vogel, C., Teichmann, S.A. (2003). Evolution of the protein repertoire. *Science*, 300, 1701–1703.

Chowdhary, R., Zhang, J., Liu, J.S. (2009). Bayesian inference of protein-protein interactions from biological literature. *Bioinformatics*, 25,1536–1542.

Colgan, D.F., Manley, J. (1997). Mechanism and regulation of mRNA polyadenylation. *Genes & Dev*elopment, 11, 2755-2766.

Collier, N., Nobata, C., Tsujii, J. (2001). Automatic acquisition and classification of terminology using a tagged corpus in the molecular biology domain. *Journal of Terminology*, 7(2), 239-257.

Curran, J.R., Clark, S., Bos, J. (2007). Linguistically motivated large-scale NLP with C&C and Boxer. In proceedings of the *ACL 2007 Demonstrations Session (ACL-07 demo)*, 33-36.

Cusick, M.E., Yu. H., Smolyar, A. , Venkatesan, K., Carvunis, A-R., Simonis, N., Rual, J-F., Borick, H., Braun, P., Dreze1, M., Vandenhaute, J., Galli, M., Yazaki, J., Hill, D.E., Ecker, J.R., Roth, F.P., Vidal, M. (2009). Literature-curated protein interaction datasets. *Nature Methods*, 6, 39-46.

Da Cruz e Silva, O.A., Fardilha, M., Henriques, A.G., Rebelo, S., Vieira, S., da Cruz e Silva, E.F. (2004). Signal transduction therapeutics: relevance for Alzheimer's disease. *Journal of Molecular Neuroscience*, 23(1-2), 123-42

Dai, H.J., Lai, P.T. and Tsai, R.T.H. (2010). Multistage Gene Normalization and SVM-Based Ranking for Protein Interactor Extraction in Full-Text Articles, *IEEE Transactions on Computational Biology and Bioinformatics*, 7, 412-420.

Deng, Z.-H., Tang, S-W., Yang, D-Q., Zhang, M., Li, L-Y, Xie, K.Q. (2004). A comparative study on feature weight in text categorization. Advanced web technologies and applications, Lecture Notes in Computer Science, (3007/2004), 588-597.

Dimililer, N., Varoğlu, E., Altınçay, H. (2009). Classifier subset selection for biomedical named entity recognition, *Applied Intelligence*, 31(3), 267-282.

Ding, J., Berleant, D., Nettleton, D., Wurtele, E. (2002). Mining Medline: abstracts, sentences, or phrases? In proceedings of the *Pacific Symposium on Biocomputing*, 326–337.

Divoli, A., Lopez, M.M., Mata, J., Wilbur, J.W. (2008). Overview of BioCreAtIvE-II gene mention recognition. *Genome Biology*, 9(2).

Doğan, R. I., Yang, Y., Névéol, A., Huang, M., Lu, Z. (2010). Identifying protein-protein interactions in biomedical text articles. In proceedings of *BioCreative-III*, Maryland, USA, 56-61.

Donaldson, I., Martin, J., Bruijin, B., Wolting, C., Lay, V., Tuekam, B., Zhang, S., Baskin, B., Bader, G.D., Michalickova, K., Pawson, T., Hogue, C.W.V. (2003). PreBIND and Textomy-mining the biomedical literature for protein-protein interactions using a support vector machine. *BCM Bioinformatics*, 4, 11-24.

Echard A., Opdam, F.J.M., Leeuw, H.J.P.C., Jollivet, F., Savelkoul, P., Hendriks, W., Voorberg, J., Goud, B., Fransen, J.A.M. (2000). Alternative Splicing of the

Human Rab6A Gene Generates Two Close but Functionally Different Isoforms. *Molecular Biology of the Cell*, 11, 3819–3833.

Ehrler, F., Gobeill, J., Tbahriti, I., Ruch, P. (2007). GeneTeam Site Report for BioCreative II: Customizing a Simple Toolkit for Text Mining in Molecular Biology. In proceedings of the *Second BioCreative Challenge Workshop*, 199-207.

Elsik, C.G., Tellam, R.L., Worley, K.C. (2009). The Genome Sequence of taurine Cattle: A Window to Ruminant Biology and Evolution. *Science*, 324(5926), 522-528.

Erkan, G., Özgür, A., Radev, D.R. (2007). Semi-supervised classification for extracting protein interaction sentences using dependency parsing. In proceedings of the *2007 joint conference on empirical methods in NLP and Computational Linguistics*, 228-237.

Farajollahi, S., Maas, S. (2010). Molecular diversity through RNA editing: a balancing act. *Trends in Genetics*, 26(5), 221-230.

Fardilha, M., Esteves, S.L., Korrodi-Gregório, L., Vintém, A.P., Domingues, S.C., Rebelo, S., Morrice, N., Cohen, P.T., da Cruz e Silva, O.A., da Cruz e Silva, E.F. (2011). Identification of the human testis protein phosphatise 1 intreactome. *Biochemical Pharmacology*, 82, 1403-1415.

Faustino, N.A., Cooper T.A. (2003). Pre-mRNA splicing and human disease. *Genes&Development,* 17, 419-437.

Filichkin, S.A., Priest, H.D., Givan, .SA., Shen, R., Bryant, D.W., Fox, S.E., Wong, W-K., Mockler, T.C. (2010). Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. *Genome Research*, 20, 45–58.

Fleuren, W.W. M., Verhoeven, S., Frijters, R. Heupers, B., Polman, J., van Schaik, R., de Vlieg, J., Alkema, W., (2011). CoPub update: CoPub 5.0 a text mining system to answer biological questions. *Nucleic Acids Research*, 9(Suppl 2), W450-W454.

Florea, L., Hartzell, G., Zhang, Z., Rubin, G.M., Miller, W. (1998). A computer program for aligning a cDNA sequence with genomic DNA sequence. *Genome Research,* 8, 967-974.

Fluck, J., Mevissen, H.T., Dach, H., Oster, M., Hofmann-Apitius, M. (2007). ProMiner: Recognition of human gene and protein names using regularly updated dictionaries", In proceedings of the *Second BioCreative Challenge Workshop*, 149-151.

Fontaine, J-F., Andrade-Navarro, M.A. (2010). Fast classification of scientific abstracts related to protein-protein interaction using a naïve Bayesian linear classifier. In proceedings of *BioCreative III*, Maryland, USA, 62-66.

Franzen, K., Eriksson, G., Olsson, F. (2002). Protein Names and How to Find Them, *International Journal of Medical Informatics, Special Issue on NLP in Biomedical Applications*, 67(1-3), 49-61.

Freund, Y., Schapire, R.E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting, *Journal of computer and system sciences*, 55, 119-139.

Friedman, C., Kra, P., Yu, H., Krauthammer, M., Rzhetsky, A. (2001). GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, 17(Suppl1), S74-S82.

Fukuda, K., Tsunoda, T., Tamura, A., Takagi, T. (1998).Toward Information Extraction: Identifying Protein Names from Biological Papers, In proceedings of the *Pacific Symposium on BioComputing*, Hawaii, 705-716.

Fundel, K., Küffner, R., Zimmer, R. (2007). RelEx: Relation extraction using dependency parse trees. *Bioinformatics*, 23, 365-371.

Furuichi, Y., Shatkin, A. (2007). Caps on Eukaryotic mRNAs. *In: Encyclopedia Of Life Sciences (ELS)***,** doi: 10.1002/9780470015902.a0000891.pub2.

Giammartino, D.C., Nishida, K., Manley, J.L. (2011). Mechanisms and Consequences of Alternative Polyadenylation. *Molecular Cell*, 43(6), 853-866.

Giot, L., Bader, J.S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y.L., Ooi, C.E., Godwin, B., Vitols, E., Vijayadamodar, G., Pochart, P., Machineni, H., Welsh, M., Kong, Y., Zerhusen, B., Malcolm, R., Varrone, Z., Collis, A., Minto, M., Burgess, S., McDaniel, L., Stimpson, E., Spriggs, F., Williams, J., Neurath, K., Ioime, N., Agee, M., Voss, E., Furtak, K., Renzulli, R., Aanensen, N., Carrolla, S., Bickelhaupt, E., Lazovatsky, Y., DaSilva, A., Zhong, J., Stanyon, C.A., Finley, R.L. Jr, White, K.P., Braverman, M., Jarvie, T., Gold, S., Leach, M., Knight, J., Shimkets, R.A., McKenna, M.P., Chant, J., Rothberg, J.M. (2003). A protein interaction map of Drosophila melanogaster. *Science*, 302, 1727-1736.

Giuliano, C., Lavelli, A., Romano, L. (2006). Exploiting shallow linguistic information for relation extraction from biomedical literature. In proceedings of the *EACL'06*, 401-408.

Gong, X., Wu, R., Zhang, Y., Zhao, W., Cheng, L., Gu, Y., Zhang, Y., Wang, J., Zhu, J., Guo, Z. (2010). Extracting consistent knowledge from highly inconsistent cancer gene data sources. *BMC Bioinformatics*, 11:76.

Goñi, J., Esteban, F.J., Mendizábal, N.V., Sepulcre, J., Ardanza-Trevijano, S., Agirrezabal, I., Villoslada, P. (2008). A computational analysis of protein-protein interaction networks in neurodegenerative diseases. *BMC Systems Biology*, 2:52.

Graveley, B.R. (2001). Alternative splicing: increasing diversity in the proteomic world. *Trends in Genetics*, 17(2), 100-107.

Grishman, R., Sundheim, B. (1996). Message Understanding Conference - 6: A Brief History. In proceedings of the *16th International Conference on Computational Linguistics (COLING), I,* Kopenhagen, 466–471.

Hakenberg J., Plake, C., Leaman, R., Schroeder, M., Gonzalez, G. (2008). Inter-species normalization of gene mentions with GNAT. *Bioinformatics*, 24, i126-132.

Hakenberg, J., Leaman, R., Vo, N.H., Jonnalagadda, S., Sullivan, R., Miller, C., Tari, L., Baral, C., Gonzalez, G. (2010). Efficient extraction of protein-protein interactions from full-text articles. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7(3), 481-94.

Harman D. (1993). The First Text REtrieval Conference (TREC-1). National Institute of Standards and Technology, Special Publication 500-207, Gaithersburg, Md. 20899.

Hatzivassiloglou, V., Weng, W. (2001). Learning anchor verbs for biological interaction patterns from published text articles. *International Journal of Medical Informatics*, 67, 19-32.

Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., Margalit, H., Armstrong, J., Bairoch, A., Cesareni, G., Sherman, D., Apweiler, R. (2004). IntAct: An open source molecular interaction database. *Nucleic Acids Research*, 32, D452-D455.

Hersh, W. (2009). Information Retrieval: A Health and Biomedical Perspective Series: Health Informatics, *Springer*, 3rd ed.

Hirschman, L., Yeh, A., Blaschke, C., Valencia, A. (2005a). Overview of BioCreatIve: Critical assessment of information extraction for biology. *BMC Bioinformatics*, 6(1):S1.

Hirschman, L., Colosimo, M., Morgan, A., Yeh, A. (2005b). Overview of BioCreAtIvE task 1B: Normalized Gene Lists, *BMC BioInformatics*, 6(Suppl 1):S11.

Hoffman, R., Valencia, A. (2005). Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics*, 21(Suppl 2), ii252-259.

House, A.E., Lynch, K.W. (2008). Regulation of Alternative Splicing: More than Just the ABCs. *The Journal of Biological Chemistry*, 283(3), 1217–1221.

Huang, M., Zhu, X., Hao, Y., Payan, D.G., Qu, K., Li, M. (2004). Discovering patterns to extract protein-protein interactions from full texts. *Bioinformatics*, 20, 1-9.

Hwang, S., Son, S., Kim, S.C., Kim, Y.J., Jeong, H., Lee, D. (2008). A protein interaction network associated with asthma. *Journal of Theoretical Biology*, 252,722–731.

Ideker, T., Sharan, R. (2008). Protein networks in disease. *Genome Research*, 18, 644-652.

Illingsley, G.D., Walter, M.A., Hammond, G.L., Cox, D.W. (1993). Physical mapping of four serpin genes: alpha 1-antitrypsin, alpha 1-antichymotrypsin, corticosteroid-binding globulin, and protein C inhibitor within a 280-kb region on chromosome 14q32.1. *American Journal of Human Genetics,* 52, 343-353.

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature*, 431, 931–945.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *PNAS,* 98(84), 4569–4574.

Itoh, A., Fujinoki, M. (2008). Tissue specificity of tropomyosin isoform in the mussel, Mytilus galloprovincialis. *Journal of Electrophoresis*, 52:(3), 47-52.

Jaeger, S., Gaudan, S., Leser, U., Rebholz-Schuhmann**,** D. (2008). Integrating Protein-Protein Interactions and Text Mining for Protein Function Prediction. *BMC Bioinformatics*, 9(Suppl 8):S2.

Jin, Y., Suzuki, H. Maegawa, S., Endo, H., Sugano, S., Hashimoto, K., Yasuda, K., Inoue, K. (2003). A vertebrate RNA-binding protein Fox-1 regulates tissue-specific splicing via the pentanucleotide GCAUG. *EMBO Journal*, 22(4), 905–912.

Joachims, T. (2002). Learning to Classify text using support vector machines. *Kluwer Academic Publishers*.

Kafkas, Ş., Varoğlu, E., Taneri, B. (2007). Building Interaction Networks of Proteins Coded by Alternatively Spliced Genes Using Text Mining Approaches. *Proccedings of the 2nd International Symposium on Health Informatics and Bioinformatics" (HIBIT 2007)*, Antalya, Turkey.

Kafkas, Ş., Varoğlu, E., Taneri, B. (2008). Methods for Abstract Retrieval from Pubmed Database for Alternatively Spliced Genes. *Proccedings of the* 3rd *International Symposium on Health Informatics and Bioinformatics (HIBIT 2008)*, Istanbul, Turkey.

Kafkas, Ş., Varoğlu, E., Taneri, B. (2009a). Interaction Networks for Proteins Coded by Alternatively Spliced Human Genes. *Proccedings of the 17th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) and 8th European Conference on Computational Biology (ECCB), (ISMB/ECCB 2009)*, Stockholm.

Kafkas, Ş., Varoğlu, E., Taneri, B. (2009b). Improving the Performance of Protein-Protein Interaction Article Selection Using Domain Specific Features. *Proccedings of the International Conference on Bioinformatics, Computational Biology, Genomics and Chemoinformatics (BCBGC 2009)*, Orlando, Florida, USA.

Kafkas, Ş., Varoğlu, E., Rebholz-Schuhmann, D., Taneri, B. (2010a). Identifying Variability of Human Splicing Forms in their Interaction by Using Literature Mining. *Proceedings of the 9th European Conference on Computational Biology (ECCB 2010)*, Ghent, Belgium.

Kafkas, Ş., Varoğlu, E., Rebholz-Schuhmann, D., Taneri, B. (2010b). Functional variation of alternative splice forms in their protein interaction networks: A literature mining approach. *BMC Bioinformatics*, 11(Suppl 5):P1 doi:10.1186/1471-2105-11-S5-P1. (Selected from the Workshop on Advances in Bio Text Mining (BioTM 2010), Ghent, Belgium.

Kafkas, Ş., Varoğlu, E., Rebholz-Schuhmann, D., Taneri, B. (2011). Diversity in the interactions of Isoforms Linked to Clustered Transcripts: A systematic Literature Analysis. *Journal of Proteomics and Bioinformatics*, 4:250-259. doi:10.4172/jpb.1000198.

Kashyap, L., Sharma, R.K. (2007). Alternative splicing: a paradoxical qudo in eukaryotic genomes, *Bioinformation*, 2(4), 155-156.

Kaufmann, M. (1998). *Proceedings of the seventh message understanding conference (MUC-7)*, Virginia.

Kent, W.J. (2002). BLAT-the BLAST-like alignment tool. *Genome Research*, 12, 656-664.

Keren, H., Lev-Maor, G., Ast, G. (2010). Alternative splicing and evolution: diversification, exon definition and function. *Nature Reviews Genetics*, doi:10.1038/nrg2776.

Keshava, P.T.S, Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D.S., Sebastian, A., Rani, S., Ray, S., Harrys, K.C.J., Kanth, S., Ahmed, M., Kashyap, M.K., Mohmood, R., Ramachandra, Y.L., Krishna, V., Rahiman, B.A., Mohan, S., Ranganathan, P., Ramabadran, S., Chaerkady, R., Pandey, A. (2009). Human Protein Reference Database - 2009 Update. *Nucleic Acids Research*, 37, D767-72.

Khalid, M., Jijkoun, V., De Rijke, M. (2008). The impact of named entity normalization on information retrieval for question answering, *Advances in Information Retrieval*, 705-710.

Kim, J., Ohta, T., Tsuruoka, Y., Tateisi, Y., Collier, N. (2004). Introduction to the Bio-Entity recognition task at JNLPBA. In proceedings of the *Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-2004)*, 70-75.

Kim, N., Alekseyenko, A.V., Roy, M., Lee, C. (2007). The ASAP II database: analysis and comparative genomics of alternative splicing in 15 animal species. *Nucleic Acids Research*,35(Database issue):D93-8.

Kim, S., Wilbur, W.J. (2010). Improving Protein-Protein Interaction Article Classification Performance by Utilizing Grammatical Relations. In proceedings of *BioCreative III*, Maryland, USA, 77-82.

Koscielny, G., Le Texier, V., Gopalakrishnan, C., Kumanduri, V., Riethoven, J.J., Nardone, F., Stanley, E., Fallsehr, C., Hofmann, O., Kull, M., Harrington, E., Boué, S., Eyras, E., Plass, M., Lopez, F., Ritchie, W., Moucadel, V., Ara, T., Pospisil, H., Herrmann, A., Reich, J., Guigó, R., Bork, P., Doeberitz, M.K., Vilo, J., Hide, W., Apweiler, R., Thanaraj, T.A., Gautheret, D. (2009). ASTD: The Alternative Splicing and Transcription Diversity database. *Genomics*, 93, 213-220.

Krallinger, M. Leitner, F., Rodriguez-Penagos, C., Valencia, A. (2008). Overview of the protein-protein interaction extraction task of BioCreative II. *Genome Biology,* 9(Suppl 2):S4.

Lan, M., Tan, C., Low, H. (2006). Proposing a new term weighting scheme for text categorization. In proceedings of the *21st AAAI*, 763-768.

Lan, M., Tan, C.L., Su, J. (2007). A Term Investigation and Majority Voting for Protein Interaction Article Sub-task 1 (IAS). In proceedings of the *Second BioCreative Challenge Workshop,* 183-185.

Lan, M., Tan, C.L., Su, J. (2009). Feature generation and representations for protein–protein interaction classification. *Journal of Biomedical Informatics*, 42, 866-872.

Landeghem, S.V., Saeys, Y., Peer, Y.V., Baets, B. (2008). Extracting protein-protein interactions from text using rich feature vectors and feature selection. In proceedings of the *2nd International workshop on Data and text mining in bioinformatics*, 61-68.

Larson, D.R., Singer, R.H., Zenklusen, D. (2009). A single molecule view of gene expression. *Trends in Cell Biology*, 19(11), 630-637.

Leitner, F., mardis, S.A., Krallinger, M., Cesareni, G., Hirschman, L.A., Valencia, A. (2010). An Overview of BioCreative II.5. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7(3), 385-399.

Leopold, E., Kindermann, J. (2002). Text categorization with support vector machines: How to represent texts in input space. *Machine Learning*, 46, 423-444.

Lee, Y., Lee, Y., Kim, B., Shin, Y., Nam, S., Kim, P., Kim, N., Chung, W.H., Kim, J., Lee, S. (2007). ECGene: an alternative splicing database update. *Nucleic Acids Research*, 35, D99-D103.

Li, S., Armstrong, C.M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P.O., Han, J.D., Chesneau, A., Hao, T., Goldberg, D.S., Li, N., Martinez, M., Rual, J.F., Lamesch, P., Xu, L., Tewari, M., Wong, S.L., Zhang, L.V., Berriz, G.F., Jacotot, L., Vaglio, P., Reboul, J., Hirozane-Kishikawa, T., Li, Q., Gabel, H.W., Elewa, A., Baumgartner, B., Rose, D.J., Yu, H., Bosak, S., Sequerra, R., Fraser, A., Mango, S.E., Saxton, W.M., Strome, S.,Van Den Heuvel, S., Piano, F., Vandenhaute , J., Sardet, C., Gerstein, M., Doucette-Stamm, L., Gunsalus, K.C., Harper, J.W., Cusick, M.E., Roth, F.P., Hill, D.E., Vidal, M. (2004). A map of the interactome network of the metazoan C. elegans. *Science,* 303, 540-543.

Li, P., Zang, W., Li, Y., Xu , F., Wang, J., Shi, T. (2011). AtPID: the overall hierarchical functional protein interaction network interface and analytic platform for *Arabidopsis*. *Nucleic Acids Research*, 39 (Suppl 1), D1130-D1133.

Li, X., Cai, H., Xu, J., Ying, S., Zhang, Y. (2010). A mouse protein interactome through combined literature mining with multiple sources of interaction evidence. *Amino Acids.* 38(4), 1237-52.

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

Liu, H., Hu, Z-Z., Zhang, J., Wu, C. (2006). BioThesaurus: a web-based thesaurus of protein and gene names. *Bioinformatics*, 22 (1), 103-105.

Liu, B. Qian, L., Wang, H., Zhou, G. (2010). Dependency-Driven Feature-based Learning for Extracting Protein-Protein Interactions from Biomedical Text. In proceedings of the *23rd International Conference on Computational Linguistics* (COLING), Poster volume, 757–765.

Lo, T.W., Branda, C.S., Huang, P., Sasson, I.E., Goodman, S.J., Stern, M.J. (2008). Different isoforms of the C. elegans FGF receptor are required for attraction and repulsion of the migrating sex myoblasts. Developmental Biology, 318, 268-275.

Lourenco, A., Conover, M., Wong, A., Pan, F., Abi-Haider, A., Nematzadeh, A., Shatkay, H., Rocha, L. (2010). Testing Extensive Use of NER Tools in Article Classification and a Statistical Approach for Method Interaction Extraction in the Protein-Protein Interaction Literature. In proceedings of *BioCreative III*, Maryland, USA, 105-109.

Lu, Z. (2011**).** PubMed and beyond: a survey of web tools for searching biomedical literature. *Database (Oxford)*, 2011: baq036.

Lu, Z. Kao, H-Y., Wei, C-H., Huang, M., Liu, J., Kuo, C-J. Hsu, C-N.**,** Tsai, R. T-H., Dai, H-J.**,** Okazaki, N. Cho, H-C., Gerner, M., Solt, I., Agarwal, S., Liu, F., Vishnyakova, D., Ruch, P., Romacker, M., Rinaldi, F., Bhattacharya, S., Srinivasan, P., Liu, H., Torii, M., Matos, S., Campos, D., Verspoor, K., Livingston, K.M., Wilbur, J.W. (2011). The gene normalization task in BioCreative III, *BMC Bioinformatics* 2011, 12(Suppl 8):S9.

Lynch, M. (2006).The Origins of Eukaryotic Gene Structure. *Molecular Biology and Evolution*, 23(2), 450-468.

Marcotte, E.M., Xenarios, L., Eisenberg, D. (2001). Mining literature for protein-protein interactions. *Bioinformatics*, 17(4), 359-363.

Mass, S. (2010). Gene regulation through RNA editing. *Discovery Medicine*, 10(54):379-86.

Mathivanan, S., Periaswamy, B., Gandhi, T.K., Kandasamy, K., Suresh, S., Mohmood, R., Ramachandra, Y.L., Pandey, A. (2006). An evaluation of human protein-protein interaction data in public domain. *BCM Bioinformatics*, 7(Suppl 5):S19.

Matlin, A.J., Clark, F., Smith, C.W.J. (2005). Understanding alternative splicing: towards a cellular code. *Nature Reviews*, 6 (5), 386–398.

Meshorer, E., Bryk, B., Toiber, D., Cohen, J., Podoly, E., Dori, A., Soreq, H. (2005). SC35 promotes sustainable stress-induced alternative splicing of neuronal acetylcholinesterase mRNA. *Molecular Psychiatry*, 10, 985–997.

Meyer, B.J. (2000). Sex in the wormcounting and compensating X-chromosome dose. *Trends Genet*ic, 16(6), 247-53.

Mika, S, Rost, B. (2004). Protein names precisely peeled off free text, *Bioinformatics*, 20(Suppl.1),1241-1247.

Miko, I., LeJeune, L. *Essentials of Genetics*. (Cambridge, MA: NPG Education, 2009).

Millevoi, S., bernat, S., Telly, D., Fouque, F., Gladieff, L., Favre, G., Vagner, S., Toulas, C. (2009). The c.5242C>A BRCA1 missense variant induces exon skipping by increasing splicing repressors binding. *Breast Cancer Research Treatment*, 120, 391

Miwa, M., Sætre, R., Miyao, Y., Ohta, T., Tsujii, J. (2008). Combining multiple layers of syntactic information for protein-protein interaction extraction. In proceedings of the *3rd International Symposium on Semantic Mining in Biomedicine*, 101-108.

Miwa, M., Sætre, R., Miyao, Y. (2009a). Protein-protein interaction extraction by Leveraging multiple kernels and parsers. *International Journal of Medical Informatics*, 78(12), e39-46.

Miwa, M., Sætre, R., Miyao, Y., Tsujii, J. (2009b). A Rich Feature Vector for ProteinProtein Interaction Extraction from Multiple Corpora. In proceedings of the *2009 Conference on Empirical Methods in Natural Language Processing*,1(1), 121-130.

Miyao, Y, Sætre R, Sagae K, Matsuzaki T, Tsujii J (2008). Task-oriented evaluation of syntactic parsers and their epresentation. In proceedings of the *ACL-08:HLT*, 46-54.

Miyao, Y., Tsujii, J. (2008). Feature forest models for probabilistic HPSG parsing. *Computational Linguistics* 2008, 34, 35-80.

Modrek, B., Lee, C. (2002). A genomic view of alternative splicing. *Nature Genetics*, 30:13-19.

Moleirinho, A., Carneiro, J., Matthiesen, R., Silva, R.M., Amorim, A., Azevedo, L. (2011). Gains, Losses and Changes of Function after Gene Duplication: Study of the Metallothionein Family. *PLoS ONE,* 6(4):e18487, doi:10.1371/journal.pone.001848

Moorhouse, M., Barry, P. (2004). Bioinformatics, Biocomputing and Perl. *John Willey and Sons (Edt.)*, 3-7.

Morgan, A.A., Lu, Z., Wang, X., Cohen, A.M., Fluck, j., Ruch, P., Divoli, A., Fundel, K., Leaman, R., Hakenberg, J., Sun, C., Liu, H., Torres, R., Krauthammer, m., Lau, W.W., Liu, H., Hsu, C-N., Schuemie, M., Cohen, K.B., Hirschman, L.(2008). Overview of BioCreative II gene normalization, *Genome Biology*, 9(Suppl 2):S3.

Moschitti, A. (2006). Making tree kernels practical for natural language learning. In proceedings of the *Eleventh International Conference on European Association for Computational Linguistics (EACL2006)*, Trento, Italy, 113-120.

Mouse Genome Sequencing Consortium (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature,* 420, 520-562

Möröy, T., Heyd, F. (2007). The impact of alternative splicing in vivo: Mouse models show the way. *RNA*, 13(8), 1155-1171.

Mutsimori, T., Murata, M., Fukuda, Y., Doi, K., Doi, H. (2006). Extracting protein-protein interaction information from biomedical text with SVM, *IEICE-Transactions on Information Systems*, E89, D2464-2466.

Nedellec, C. (2005). Learning language in logic-genic interaction extraction challenge. In proceedings of the *ICML05 workshop: Learning Language in Logic (LLL'05)*. Bonn, Germany, 97–99.

Nilsen, T.W. (2003). The spliceosome: the most complex macromolecular machine in the cell. *Bioesseys*, 25(12), 1147-1149.

Nilsen, T.W., Graveley, B.R. (2010). Expansion of the eukaryotic proteome by alternative splicing. *Nature*, 463(28).

Ninomiya, T., Matsuzaki, T., Miyao, Y., Tsujii, J. (2007). A log-linear model with an n-gram reference distribution for accurate HPSG parsing. In proceedings of *IWPT 2007*. Prague, Czech Republic.

Nishikura, K. (2010). Functions and regulation of RNA editing by ADAR deaminases, *Annual Reviews in Biochemistry*, 79:1–29.

Niu,Y., Otasek, D., Jurisica, I.(2010). Evaluation of linguistic features useful in extraction of interactions from PubMed; Application to annotating known, high-throughput and predicted interactions in I2D. *Bioinformatics*, 26, 111-119.

Nurtdinov, R.N., Neverov A.D., Favorov A.V., Mironov, A.A., Gelfand, M.S. (2007). Conserved and species-specific alternative splicing in mammalian genomes. *BMC Evolutionary Biology*, **7**:249.

Ofran, Y., Yachdav, G., Mozes, E., Soong, T., Nair, R., Rost, B. (2006). Create and assess protein networks through molecular characteristics of individual proteins. *Bioinformatics*, 22 (14):e402-e407.

Ohta, T., Tateisi, Y., Mima, H., Tsujii, J. (2002). The GENIA corpus: An annotated research abstract corpus in the molecular biology domain. In proceedings of the *2nd International Conference on Human Language Technology Research*, 82-86.

Ohta, T., Matsuzaki, T., Okazaki, N., Miwa, M., Sætre, R., Pyysalo, S., Tsujii, J. (2010). Medie and Info-pubmed: 2010 update. *BMC Bioinformatics*, 11(Suppl 5):P7.

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

Pagel, P., Kovac, S., Oesterheld, M., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Mark, P., Mewes, H-W., Ruepp, A., Frishman, D. (2005). The MIPS mammalian protein--protein interaction database Source. *Bioinformatics,* 21, 832-834.

Pelissier, P., Delourme, D., Germot, A., Blanchet, X., Becila, S., Maftah, A., Leveziel, H., Ouali, A., Bremaud, L. (2008). An original SERPINA3 gene cluster: elucidation of genomic organization and gene expression in the Bos taurus 21q24 region. *BMC Genomics*, 9:151.

Pontius, J.U., Mullikin, J.C., Smith, D.R., Agencourt Sequencing Team, Lindblad-Toh,K., Gnerre, S., Clamp, M., Chang, J., Stephens, R., Neelam, B., Volfovsky, N., Schäffer, A.A., Agarwala, R., Narfström, K., Murphy, W.J., Giger, U., Roca, A.L., Antunes, A., Menotti-Raymond, M., Yuhki, N., Pecon-Slattery, J., Johnson, W.E., Bourque, G. Tesler, G., NISC Comparative Sequencing Program, O'Brien, S.J. (2007). Initial sequence and comparative analysis of the cat genome. *Genome Research*, 17, 1675-1689.

Paul, M. S., Bass, B. L. (1998). Inosine exists in mRNA at tissue-specific levels and is most abundant in brain mRNA. *Embo Journal*, 17:1120-1127.

Phan, X, Horiguchi, S., Nguyen, L., Nguyen, C. (2007). Semantic Analysis of Entity Contexts towards Open Named Entity Classification on the Web. In proceedings of the *10th Conference of the Pacific Association for Computational Linguistics*, 137–144.

Prieto, C., Rivas, J. (2006). APID: Agile Protein Data Analyzer. *Nucleic Acid Research*, 34, W298-W302.

Pyysalo, S., Ginter, F., Heimonen, J., Bjorne, J., Boberg, J., Jarvinen, J., Salakoski, T. (2007). BioInfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8: 50.

Pyysalo, S., Sætre, R., Tsujii, J., and Salakoski, T. (2008). Why Biomedical Relation Extraction Results are Incomparable and What to do about it. In Proceedings of the *3rd International Symposium on Semantic Mining in Biomedicine (SMBM 2008)*, 149-152.

Rain, J-C., Selig, L., De Reuse, H., Battaglia, V., Reverdy, C., Simon, S., Lenzen, G., Petel, F., Wojcik, J., Schachter, V., Chemama, Y., Labigne, A., Legrain, P. (2001). The protein-protein interaction map of Helicobacter pylori. *Nature*, 409,211–215.

Ramani, A.K., Calarco1, J.A., Pan, Q., Mavandadi, S., Wang, Y., Nelson, A.C., Lee, L.J., Morris, Q., Blencowe, B.J., Zhen, M., Fraser, A.G. (2010). Genome-wide analysis of alternative splicing in *Caenorhabditis elegans. Genome Research*, 21, 342-348.

Reenan, R. A. (2001). The RNA word meets behavior: A to I pre-mRNA editing in animals, *Trends Genet*ic, 17, 53-56.

Rebholz-Schuhmann, D., Arregui, M., Kirsch, H., Jimeno, A. (2008). Text processing through Web services:calling Whatizit. *Bioinformatics*, 15:24(2), 298-298.

Resch, A., Xing, Y., Modrek, B., Gorlick, M., Riley, R., Lee, C. (2003). Assessing the Impact of Alternative Splicing on Domain Interactions in the Human Proteome. *Journal of Proteome Research*, 3(1), 76-83.

Rogers, K.A., Scinicariello, F., Attanasio, R. (2006). IgG Fc Receptor III Homologues in Nonhuman Primate Species: Genetic Characterization and Ligand Interactions. *Journal of Immunology*, 177, 3848-3856.

Rollini, P., Fournier, R.E. (1997). A 370-kb cosmid contig of the serpin gene cluster on human chromosome 14q32.1: molecular linkage of the genes encoding alpha 1-antichymotrypsin, protein C inhibitor, Kallistatin, alpha 1-antitrypsin and corticosteroid-binding globulin. *Genomics*, 46, 409-415.

Roy, S. W., Irimia, M. (2009). Splicing in the eukaryotic ancestor: form, function and dysfunction. *Trends in Ecology and Evolution*, 24(8), 447–455.

Ruffner, H., Bauer, A., Bouwmeester, T. (2007). Human protein-protein interaction networks and the value of drug discovery. *Drug Discovery Today*, 12(17-18), 706-716.

Sætre, R., Sagae, K., Tsujii, J. (2007). Syntactic features for protein-protein interaction extraction, In Proceeding of the *LBM'07*.

Sagae, K., Tsujii, J. (2007). Dependency parsing and domain adaptation with LR models and parser ensembles. In proceedings of the *EMNLP-CoNLL* 2007.

Salton, G., Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management,* 24(5), 513–523.

Sánchez, L. (2008). Sex-determining mechanisms in insects. *International Journal of Developmental Biology,* 52, 837–856

Saric, J., Jensen, L.J., Ouzounova, R., Rojas, I., Bork, P. (2006). Extraction of regulatory gene/protein networks from Medline. *Bioinformatics,* 22 (6), 645-650.

Settles, B. (2005). ABNER:an open source tool for automatically tagging genes, proteins, and other entity names in text, *Bioinformatics*, 21(14), 3191-3192.

Shah, P.K., Jensen, L.J., Boué, S., Bork, P. (2005). Extraction of Transcript Diversity from Scientific Literature. *PLoS Computational Biology,* 1:e10.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research,* 13(11), 2498-504.

Shatkay. H., Edwards, S., Wilbur,W.J., Boguski, M. (2000). Genes, themes and microarrays: using information retrieval for large-scale gene analysis. In proceedings of the *8th International Conference on Intelligent Systems for Molecular Biology (ISMB),*19-23.

Shell, S.A., Hesse, C., Morris Jr., S.M. Milcarek, C. (2005). Elevated levels of the 64-kDa cleavage stimulatory factor (CstF-64) in lipopolysaccharide-stimulated macrophages influence gene expression and induce alternative poly(A) site selection. *Journal of Biological Chemistry*, 280(48), 39950-39961.

Shen, Y., Ji, G., Haas, B.J., Wu, X., Zheng, J., Reese, G.J., Li, Q.Q. (2008). Genome level analysis of rice mRNa 3'-end processing signals and alternative polyadenilation. *Nucleic Acids Research*, 36(9):3150-3161.

Shen, Y., Venu, R.C., Nobuta, K., Wu, X., Notibala, V., Demirci, C., Meyers, B.C., Wang, G-L., Ji, G., Li, Q.Q. (2011). Transcriptome dynamics through alternative polyadenilation in developmental and environmental responses in plants revealed by deep sequencing. *Genome Research*, 21:1478-1486.

Shoemaker, B.A., Panchenko, A.R. (2007). Deciphering Protein–Protein Interactions. Part I. Experimental Techniques and Databases. *PLoS Computational Biology*, 3(3): e42. doi:10.1371/journal.pcbi.0030042.

Smith, L., Tanabe, L., Ando, R., Kuo, C.J., Chung, F.I., Hsu, C.N., Lin, Y.S., Klinger, R, Friedrich, C., Ganchev, K., Torii, M., Liu, H., Haddow, B., Struble, C., Povinelli, R., Vlachos, A., Baumgartner, W., Hunter, L., Carpenter, B., Tsai, R., Dai, H.J., Liu, F., Chen, Y., Sun, C., Katrenko, S., Adriaans, P., Blaschke, C., Torres, R., Neves, M., Nakov, P., Stapley, B.J., Benoit, G. (2000). Biobibliometrics: Information Retrieval and Visualizaiton from Co-occurrence of Gene Names in Medline Abstracts. In proceedings of the *5th Pacific Symposium on Biocomputing*, 529-549.

Sorek, R., Ast, G. (2003). Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Research*, 13, 1631–1637.

Sorek, R., Dror, G., Shamir, R. (2006). Assessing the number of ancestral alternatively spliced exons in the human genome. *BMC Genomics*, 7:273

Stamm, S., Ben-Ari, S., Rafalska, I., Tang, Y., Zhang, Z., Toiber, D., Thanaraj, T.A., Soreq, H. (2005). Function of alternative splicing. *Gene*, 34, 1-20.

Stapley, B., Benoit, G. (2000). Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. In Proceedings of the *Pacific Symposium on Biocomputing*, 529-540, Hawaii, U.S.A, 2000.

Stark, C., Breitkreutz, B-J., Reguly, T., Boucher, L., Breitkreutz, A., Tyers, M. (2006). BioGRID: a general repository for interaction datasets. *Nucleic Acids Research*, 34, D535-D539.

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F.H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., Timm, J., Mintzlaff, S., Abraham, C., Bock, N., Kietzmann, S., Goedde, A.,Toksöz, E., Droege, A., Krobitsch, S., Korn, B., Birchmeier, W., Lehrach, H., Wanker, E.E. (2005). A human protein protein interaction network: a resource for annotating the proteome. *Cell*, 122, 957-968.

Stumpf, M.P.H, Thorne, T., Silva, E., Stewart, R., An, H.J., Lappe, M., Wiuf, C. (2008). Estimating the size of the human interactome, *PNAS*, 105(19), 6959-6964.

Stumpf, M.P.H., Kelly, W.P., Thorne, T., Wiuf, C. (2007). Evolution at the system level: the natural history of protein interaction networks. *TRENDS in Ecology and Evolution*, 22(7), 366-373.

Sun, C., Lin, L., Wang, X., Guan, Y. (2007). Using maximum entropy model to extract protein-protein interaction information from biomedical literature. *Advanced Intelligent Computing Theories and Applications*, Springer, number 4681 in LNCS, 730–737.

Hoppins, S.C., Go, N.E., Klein, A., Schmitt, S., Neupert, W., Rapaport, D., Nargang, F.E. (2007). Alternative Splicing Gives Rise to Different Isoforms of the *Neurospora crassa* Tob55 Protein That Vary in Their Ability to Insert β-Barrel Proteins Into the Outer Mitochondrial Membrane. *Genetics*, 177(1), 137-149.

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P., Jensen, L.J., von Mering, C. (2011). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Research, 39, D561–D568.

Taneri, B., Snyder, B., Novoradovsky, A., Gaasterland, T. (2004). Alternative splicing of mouse transcription factors affect their DNA-binding domain architecture and is tissue specific. *Genome Biology*, 5(R75).

Taneri, B., Snyder, B., Novoradovsky, A., Synder, B., Gaasterland, T. (2005). Databases for comparative analysis of human-mouse orthologous alternative splicing. *Lecture Notes in Bioinformatics*, 3388, 123-131.

Taneri, B. (2005). Comparative Analysis of Alternative Splicing in Homo sapiens, Mus Musculus and Rattus norvegicus Transcriptomes. *Ph.D Thesis*, The Rockefeller University, USA.

Taneri, B., Snyder, B., Novoradovsky, A., Gaasterland, T. (2011). Alternative Splicing in the Fly and the Worm: Splicing Databases for Drosophila melanogaster and Caenorhabditis elegans, dexa. In proceedings of the *22nd International Workshop on Database and Expert Systems Applications*, 435-439.

Thanaraj, T.A., Clark, F., Muilu, J. (2003). Conservation of human alternative splice events in mouse. *Nucleic Acids Research*, 31(10), 2544-2552.

Thomas, J., Milward, D., Ouzounis, C., Pulman, S., Carrol, M. (2000). Automatic extraction of protein interactions from scientific abstracts. In proceedings of the *5th Pacific Symposium on Biocomputing*, 538-549.

Tian, B., Hu, J., Zhang, H., and Lutz, C.S. (2005).A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Research*. 33: 201–212.

Tikk, D., Thomas, P., Palaga, P., Hakenberg, J., Leser, U. (2010). A Comprehensive Benchmark of Kernel Methods to Extract Protein-Protein Interactions from Literature. *Plos Computational Biology,* 6(7)**:**e1000837. doi: 10.1371/journal.pcbi.1000837

Torii, M., Hu, Z., Wu, C.H., Liu, H. (2009). BioTagger-GM: A Gene/Protein Name Recognition System, *Journal of the American Medical Informatics Associations*,16(2), 247–255.

Trafton, A. (2008). Human genes sing different tunes in different tissues. *MIT, TechTalk*, 53(8).

Tsai, R.T., Hung, H., Dai, H., Lin, Y., Hsu, W. (2008). Exploiting likely-positive and unlabeled data to improve the identification of protein-protein interaction articles. *Bioinformatics*, 9(Suppl. I):S3.

Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Alia, Q-E., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., Rothberg, J.M. (2000). A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, 403, 623–627.

Vapnik, V. (1995). The nature of statistical learning theory. *Springer-Verlang*.

Waagmeester, A., Pezik, P., Coort, S., Tourniaire, F., Evelo, C., Rebholz-Schuhmann, D. (2009). Pathway enrichment based on text mining and its validation on carotenoid and vitamin A metabolism. *OMICS*, 13, 367-379.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456, 470–476.

Wang, H., Huang, M., Ding, S., Zhu, X. (2008). Exploiting and integrating rich features for biological literature classification. *BCM Bioinformatics*, 9(Suppl):S4.

Wang, Z., Burge, C. (2008). Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA*, 14 (5), 802–13.

Wang, X., Rak, Resti_car, A., Nobata, C., Rupp, C.J., Batista-Navarro, R.B., Nawaz, R., Ananiadou, S. (2010). NaCTeM Systems for BioCreative III PPI Tasks, In proceedings of *BioCreative III*, Maryland, USA, 142-147.

Wermter, J., Tomanek, K., Hahn, U. (2009). High-performance gene name normalization with GENO", *Bioinformatics*, 25(6), 815-821.

Wilbur, W.J., Hazard, G.F., Divita, G., Mork, J.G., Aronson, A.R., Browne A.C. (1999). Analysis of biomedical text for chemical names: A comparison of three methods. In proceedings of the *AMIA Annual Symposium*, 176-180.

Wong, L. (2001). PIES, a protein interaction extraction system. In proceedings of the *6th Pacific Symposium on Biocomputing*, 520-531.

Woodley, L.,Valcarcel, J. (2002). Regulation of alternative pre-mRNA splicing. *Briefings in Functional Genomics and Proteomics*, 1, 266-277.

Wu, J., Vallenius, T., Ovaska, K., Westermarck, J., Mäkelä, T.P., Hautaniemi, S. (2009). Integrated network analysis platform for protein-protein interactions. *Nature Methods*, 6, 75-77.

Xenarios, I., Rice, D.W., Salwinski, L., Baron, M.K., Marcotte, E.M., Eisenberg, D. (2000). DIP: the database of interacting proteins. *Nucleic Acids Research*, 28, 289-291.

Yakushiji, A., Miyao, Y., Tateisi, Y., Tsujii, J. (2005). Biomedical information extraction with predicate-argument structure patterns. In proceedings of the *11th annual meeting of the association for natural language processing*, 60-69.

Yang, Y., Pedersen, J.O. (1997). A comparative study on feature selection in text categorization. In proceedings of the *14th International Conference on Machine Learning*, 412-420.

Yang, Z.H., Lin, H.F., Wu, B.D. (2009). BioPPIExtractor: A protein-protein interaction extraction system for biomedical literature. *Expert Systems with Applications*, 36(2), 2228-2233.

Yang, Z., Lin, H., Li, Y. (2010a). BioPPISVMExtractor: A protein–protein interaction extractor for biomedical literature using SVM and rich feature sets, *Journal of Biomedical Informatics*, 43(1), 88-96.

Yang, Z., Lin, Y., Wu, J., Tang, N., Lin, H., Li, Y (2010b). Ranking SVM for multiple kernels output combination in protein-protein interaction extraction from biomedical literature. In proceedings of the 2010 *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 595-598.

Yang, Z., Tang, N., Zhang, X., Lin, H., Li, Y., Yang, Z. (2011). Multiple kernel learning in protein-protein interaction extraction from biomedical literature. *Artificial Intelligence in Medicine*,51(3),163-73.

Yeh, A., Morgan, A.,Colosimo, M., Hirschman, L. (2005). BioCreAtIvE Task 1A: gene mention finding evaluation, *BMC Bioinformatics*. 6(Suppl 1):S2.

Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., Ausiello, G., Helmer-Citterich, M., Cesareni, G. (2002). MINT: a Molecular INTeraction database. *FEBS Letters* 513, 135-140.

Zanzoni, A., Soler-López, M., Aloy, P. (2009). A network medicine approach to human disease. *FEBS Letters*, 583(11), 1759-1765.

Zhang, H., Lee, J.Y., Tian, B. (2005). Biased alternative polyadenilation in human tissues. *Genome Biology*, 6(12):R100.

Zhang, J., Marszalek, M., Lazebnik, S., Schmid, C. (2007). Local features and kernels for classification of texture and object categories: a comprehensive study. *International Journal of Computer Vision*, 73(2), 213-238.

Zhang, T. (2004). Solving large scale linear prediction problems using stochastic gradient descent algorithms. In proceedings of the *21st International Conference on Machine Learning*, Banff, 919-926.

Zhang, X-N., Mount, S.M. (2009). Two Alternatively Spliced Isoforms of the Arabidopsis SR45 Protein Have Distinct Roles during Normal Plant Development. *Plant Physiology*, 150(3), 1450-1458.

Zhou, D., He, Y., Kwoh, C.K. (2008). From biomedical literature to knowledge: mining protein-protein interactions. *Computational intelligence in Biomedicine and Bioinformatics: Current Trends and Applications. Studies in Computational Intelligence* (151), Springer, 397–421.

APPENDICES

Appendix A: Evaluation Metrics

Although there are a number of different metrics used to evaluate performances of the information extraction systems in the domain, the most popular metrics are Recall (R), Precision (P) and F-score. In this study, performances of the designed classifiers are evaluated based on these metrics. Following counts are involved in calculation of the R, P and F-score values:

True Positive (TP): Number of positive objects (correctly) classified as positive by the system.

False Negative (FN): Number of positive objects (incorrectly) classified as negative by the system.

False Positive (FP): Number of negative objects (incorrectly) classified as positive by the system.

True Negative (FN): Number of negative objects (correctly) classified as negative by the system.

Recall is defined as the ratio between the correctly identified objects and the total number of objects. It is calculated by using the equation (A.1).

$$R = \frac{TP}{TP + FN} \tag{A.1}$$

Precision is defined defined as the ratio between the correctly identified objects and the number of objects identified by the system. It is calculated by using the equation (A.2).

$$P = \frac{TP}{TP + FP} \tag{A.2}$$

F-score is defined as the harmonic mean between precision and recall. It is calculated by using the equation (A.3).

$$F_{\alpha} - score = \frac{1}{\alpha \frac{1}{P} + (1-\alpha) \frac{1}{R}} \tag{A.3}$$

where, α is a factor used for assigning weights to precision and recall. It is possible to adjust this factor according to the system requirements. Typically, equal weights are assigned to precision and recall ($\alpha = 0.5$). In this particular case, F-score is termed as the F_1-score and calculated by using the equation (A.4).

$$F_1 - score = \frac{2PR}{P+R} \tag{A.4}$$

Appendix B: Support Vector Machines

SVM (Vapnik, 1995) is a machine learning algorithm used for solving binary classification problem based on the recognised patterns from the analysis of a given set of training data. It has been successfully used in many classification problems including many text mining and document classification tasks (Joachims, 2002). An SVM training algorithm computes an optimal hyperplane separating positive and negative data by maximizing the margin between the hyperplane and nearest training data points (called the support vectors). In the classification phase, the margin is used to predict the class of new examples as positive or negative.

The concept can be expressed mathematically as follows:

A given set of labelled training data is denoted as, $L = \{(x_1, y_1), (x_2, y_2), ..., (x_l, y_l)\}$

where,

(x_i, y_i) (i=1..l) corresponds to (instance, class) pairs used to represent the i^{th} data point from the set,

$\mathbf{x_i} = \{x_{i1}, x_{i2}, ..., x_{id}\}^T$ denotes the the feature vector of the i^{th} data point in the d-dimensional space,

$y_i \in \{-1, +1\}$ denotes the class label of the i^{th} data point,

the SVM computes the maximum-margin hyperplane that divides the positive and negative data points by optimising the following problem:

In the case that data can be linearly separable, hyperplane which is used for the decision is defined as:

$$\mathbf{w} \cdot \mathbf{x} + b = 0 \qquad \text{(B.1)}$$

where, . denotes the dot product, $\mathbf{w} = [w_1, w_2, ..., w_N]^T$ is the weight vector and b (bias) is the threshold from the origin. The objective here is to choose the $\mathbf{w}$ and b values to maximize the distance (margin) between the two parallel hyperplanes which are represented by $\mathbf{w} \cdot \mathbf{x} + b = 1$ and $\mathbf{w} \cdot \mathbf{x} + b = -1$.

The classification task is depicted in Figure B.1. The straight line denotes the decision boundary that divides the feature space into two. Data points on the decision boundary must satisfy the equation (B.1).

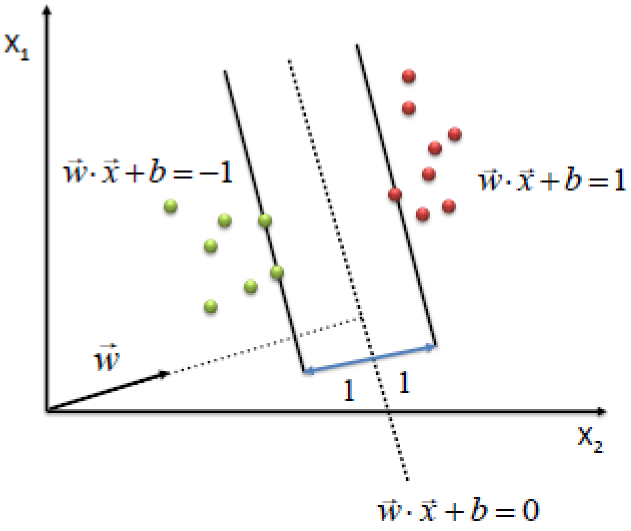

X_1
$\vec{w} \cdot \vec{x} + b = -1$
$\vec{w} \cdot \vec{x} + b = 1$
$\vec{w}$
1
1
X_2
$\vec{w} \cdot \vec{x} + b = 0$

Figure B.1. Illustration of a Linearly Separable Classification Problem

Data points on the margins are the support vectors.

The distance between the hyperplanes is calculated as:

$$d = \frac{2}{\|\mathbf{w}\|} \tag{B.2}$$

Hence the objective becomes minimizing $\|\mathbf{w}\|$. Following should be taken into account in order to prevent data points to fall into the margin:

$$\begin{cases} \mathbf{w}\cdot\mathbf{x_i}+b\geq 1 \text{ if } y_i=1 \\ \mathbf{w}\cdot\mathbf{x_i}+b\geq -1 \text{ if } y_i=-1 \end{cases} \tag{B.3}$$

where y_i denotes the target class label (1,-1) and i=1..N.

(B.2) can be substituted by:

$$\frac{\|\mathbf{w}\|^2}{2} \tag{B.4}$$

and (B.3) can be rewritten as:

$$\mathrm{y_i}(\mathbf{w}\cdot\mathbf{x_i}+b)\geq 1 \tag{B.5}$$

Therefore, the problem can be expressed as an optimization problem and formalised as shown below:

$$\min_{\mathbf{w}}\frac{\|\mathbf{w}\|^2}{2} \tag{B.6}$$

$$\mathrm{y_i}(\mathbf{w}\cdot\mathbf{x_i}+b)\geq 1 \;\;, \text{ for } i{=}1..N \tag{B.7}$$

Lagrange multiplier method is applied to solve the equation (B.6) which can be expressed as:

$$L_p=\frac{1}{2}\|\mathbf{w}\|^2 \text{ - } \sum_{i=1}^{N}\lambda_\mathrm{i}\left(\mathrm{y_i}(\mathbf{w}\cdot\mathbf{x_i}+b)-1\right), \tag{B.8}$$

where λ_i denotes the i^{th} lagrange multiplier.

Taking the derivative of L_p with respect to $\mathbf{w}$ gives:

$$\mathbf{w}=\sum_{i=1}^{N}\lambda_\mathrm{i}\mathrm{y_i}\mathbf{x_i} \tag{B.9}$$

Sum for $\mathbf{w}$ in (B.9) needs to be evaluated over the support vectors which are at the minimum distance away from the hyperplane. i.e. points where $\lambda_i \geq 0$.

Taking the derivative of L_p with respect to b gives.

$$\sum_{i=1}^{N} \lambda_{\mathrm{i}} \mathrm{y}_{\mathrm{i}} = 0 \tag{B.10}$$

(B.10) does not give b . Hence, by using the Karush-Kuhn-Tucker (KKT) conditions, (B.5) is represented as:

$$\lambda_{\mathrm{i}}\left[y_i\left(\mathbf{w} \cdot \mathbf{x_i} + b\right) - 1\right] = 0\,. \tag{B.11}$$

Substituting (B.9) and (B.10) into (B.8) gives the dual form of L_d:

$$L_D = \sum_{\mathrm{i}=1}^{\mathrm{N}} \lambda_i \text{ - } \frac{1}{2} \sum_{\mathrm{i,j}} \lambda_{\mathrm{i}} \lambda_{\mathrm{j}} y_i y_j \mathbf{x_i} \cdot \mathbf{x_j} \tag{B.12}$$

Hence, the function needs to be maximized in order to solve the optimization problem. Quadratic approaches can be applied to solve the problem and compute the values of $\lambda_{i,}$ b and **w.**

The decision boundary is represented as:

$$\left(\sum_{\mathrm{i}=1}^{\mathrm{N}} \lambda_i y_i \mathbf{x_i} \cdot \mathbf{x} \right) + b = 0 \tag{B.13}$$

and the target class of an test data point **z** is predicted as follows:

$$f(\mathbf{z}) = sign\left(\mathbf{w} \cdot \mathbf{z} + b\right) = sign\left(\sum_{i=1}^{N} \lambda_i y_i \mathbf{x_i} \cdot \mathbf{z} + b \right) \tag{B.14}$$

It is possible to extend the methodology of the SVM to data which is noisy. This introduces the idea of slack variable, ξ for allowing some data points to be misclassified and the trade-off (C) between maximizing the margin and minimizing the number of misclassified variables for penalizing the misclassification. The problem is represented as follows:

$$\min_{\mathbf{w}} \frac{\|\mathbf{w}\|^2}{2} + C\left(\sum_{i=1}^{N} \xi_i \right)^k \tag{B.15}$$

$$\text{subject to } \begin{cases} \mathbf{w}\cdot\mathbf{x}_i + b \geq 1\text{-}\xi_i \text{ if } y_i = 1 \\ \mathbf{w}\cdot\mathbf{x}_i + b \geq -1+\xi_i \text{ if } y_i = -1 \end{cases} \tag{B.16}$$

where, *i*=1..*N* and *C*&*k* are used to represent misclassification penalty.

In most of the real world applications the data is not linearly separable. In this case, a kernel function is used to transform the data into a higher dimensional space *Φ(x)* so that it becomes linearly separable. In such a case, the problem is represented as follows:

$$\min_{\mathbf{w}} \frac{\|\mathbf{w}\|^2}{2} \tag{B.17}$$

$$\text{subject to } \begin{cases} \mathbf{w}\cdot\Phi(\mathbf{x}_i) + b \geq 1 \text{ if } y_i = 1 \\ \mathbf{w}\cdot\Phi(\mathbf{x}_i) + b \geq -1 \text{ if } y_i = -1 \end{cases} \tag{B.18}$$

Based on the approach of the linear SVM and the quadratic programming, class of a given test data point **z**, can be predicted as:

$$f(\mathbf{z}) = sign(\mathbf{w}.\Phi(\mathbf{z}) + b) = sign\sum_{i=1}^{N}(\lambda_i y_i \Phi(\mathbf{x}_i).\Phi(\mathbf{z}) + b) \tag{B.19}$$

Kernel function is used to replace the dot product given that its computation in a high dimensional space is expensive. A kernel function is represented as $K(\mathbf{x},\mathbf{x}') = \Phi(\mathbf{x})^T\Phi(\mathbf{x}')$, which denotes the distance between $\mathbf{x}$ and $\mathbf{x}'$ transformed by Φ. Hence, a given test data point **z** is classified based on the following:

$$f(\mathbf{z}) = sign\left(\sum_{i=1}^{N}\lambda_i y_i K(\mathbf{x}_i,\mathbf{z}) + b\right) \tag{B.20}$$

Several examples to kernel functions are listed below (polynomial, radial basis function and sigmoid in order):

$$
\begin{aligned}
&K(\mathbf{x},\mathbf{y}) = (\mathbf{x}\cdot\mathbf{y}+1)^{q} \\
&K(\mathbf{x},\mathbf{y}) = e^{-\|\mathbf{x}-\mathbf{y}\|^{2}/\left(2\sigma^{2}\right)} \\
&K(\mathbf{x},\mathbf{y}) = \tanh(k\mathbf{x}\cdot\mathbf{y}-\delta)
\end{aligned}
\tag{D.21}
$$