# Biomarker Discovery Using Bioinformatics Methods

**by**

**Md Tawhidul Islam**
**Bachelor of Science (Computing)**
**Macquarie University, Sydney, Australia**

**A thesis submitted in fulfilment of the requirements for the degree of**
**Master of Philosophy**

**Department of Chemistry and Biomolecular Sciences**
**Macquarie University**
**Sydney, Australia**

**March 2010**

# Table of contents

## Abstract

A biomarker is a biochemical indicator of a biologic state that may serve as an indicator or predictor of a disease. Biomarker is used to measure presence, risk, progress or the effect of treatment of a disease rather than measuring the disease itself. Biomarkers act as a basis for the selection of lead candidates for clinical trials. Scientists have been searching for biomarkers for decades. Methods of discovery have developed as the technology emerges. Advances in genomics and proteomics have made it easier to interrogate hundreds or thousands of potential markers at a time and produced an unprecedented growth in the volume of new data in the field of biomarker, drug discovery and patient care. However success and progress of such work is very much dependent on prior knowledge and experience with the potential markers of interest. The diverse data generated by high-throughput biotechnology is an ideal starting point for gaining knowledge in system bioinformatics. This information is only useful if it is easily accessible. However, majority of them are presented in free-text format that are not readily available for automatic computerized analysis.

In this thesis we present a novel knowledge aggregation approach based on statistical, user-defined structural rules, machine learning, text mining and Natural Language Processing (NLP) techniques to automatically extract biomarker related information from scientific literatures. Our knowledge aggregation approach combines of two major tasks namely, Information Extraction and Relationship Extraction. Therefore the thesis first presents an automatic information retrieval, summarization and extraction (mExtract) tool. Built on statistical and pattern matching NLP technique our intelligent agent system, mExtract is capable of retrieving most relevant documents from the web based on user queries. Once the documents are retrieved, system then uses its underlying techniques to extract biomarker specific information (i.e. protein, gene, genome, disease) from the text by finding out the focal topic of the document and extracting the most relevant properties of that topic and also generates a summary of the topic. Secondly, we present our extended system namely Biomarker Information Extraction Tool (BIET), that is capable of extracting biomarker relationship within disease, gene and protein. For a given set of oncology related texts (i.e., Abstract), BIET extracts biomarker relationship namely, *is biomarker of* (disease, gene/protein) from the

texts. Built on state-of-the-art statistical models and machine learning techniques BIET consists of three major components; Semantic Category Recognition to identify the evaluative sentences among other sentences by recognizing words and phrases in the text belonging to semantic categories of interest to bio-medical entities, Assertion Classification to determine whether the statement refers to biomarker entity (protein, gene and disease) relationship and Semantic Relationship Classification to identify the biomarker relationship among the bio-medical entities.

The diverse applications presented in this thesis demonstrate that our new knowledge aggregation approach is practical , effective in the sense it utilizes a series of statistical models that are heavily reliant on local lexical and syntactic context and achieve competitive results compared to more complex NLP solutions; versatile as it is easily extendable to similar or more complex relation extraction task and represents an important contribution to bioinformatics and to the fields of biomedical research in which it is applied.

## Declaration

This thesis contains no material that has been accepted for the award of any higher degree or diploma at any University or Institution, and to the best of my knowledge, contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

## Acknowledgement

At first I would like to record my gratitude to Professor Shoba Ranganathan for her supervision, advice, and guidance from the very early stage of this research as well as giving me extraordinary experiences throughout the work. She went above and beyond to assist me with my work and given me time in both business hours, after hours including weekends. I am indebted to her more than I can express. I gratefully thank and acknowledge Dr Abhaya Nayak for his co-supervision, crucial contribution, and guidance throughout my candidature.

Many thanks go in particular to Dr Mostafa Al Masum for his valuable advice, collaboration and furthermore, using his precious times to read this thesis to provide critical comments about it. Dr Mostafa always granted me his time for answering any question relating to my work. I have also benefited by advice and guidance from former PHD students of our research group, Dr Durgaprashad Bolina. I am also grateful to my current lab member Gaurav Kumar for his support.

Many thanks to MICROS-FIDELIO Australia management and my colleagues, with special thanks to Chris Gribble (former country Manager) for his ongoing inspiration, Marcus Crowley (Managing Director), Sue Savage (General Manager) for allowing me to take time off from work (as required) and for their inspirations.

My parents deserve special mention for their mental support and prayers. My Father, Late Md Abdullah Ansary, whom I am indebted for rest of life for inspiring me to do higher education, for showing me the dream since my childhood and giving me all his savings to come to Australia. He always had great confidence on me. While my father's inspiration keeps me going, it is my mother, Achia Khatun, who sincerely raised me with great care and gentle love. Anwar, Tariq thanks for being supportive and caring siblings.

Words cannot express my appreciation to my wife Saima whose dedication, love and persistent confidence in me, has greatly helped me to take some of the load off my shoulder. I owe her for all her supports and inspirations.

## Dedication

MST. ACHIA KHATUN (MOTHER), LATE MD. ABDULLAH ANSARY (FATHER), SAIMA JESMIN (WIFE), MOHAMMAD IMDADUL ISLAM (BROTHER), MD ANWARDUL ISLAM (BROTHER), MD TARIQUL ISLAM(BROTHER).

## List of Abbreviations

| | |
|---|---|
| CLL | Chronic Lymphocytic Leukaemia |
| CRF | Conditional Random Fields |
| DNA | Deoxyribonucleic acid |
| IE | Information Extraction |
| IR | Information Retrieval |
| KNN | K-Nearest Neighbourhood |
| LSA | Latent Semantic Analysis |
| MEMM | Maximum Entropy Markov Model |
| MeSH | Medical Subject Heading |
| ML | Machine learning |
| NER | Named Entity Recognizer |
| NLP | Natural Language Processing |
| PCA | Principal Component Analysis |
| SVD | Singular Value Decomposition |
| SVM | Support Vector Machine |
| TF-IDF | Term Frequency – Inverse Document Frequency |
| UMLS | Unified Medical Language System |

## Contribution to thesis as publications

1. M.T. Islam, D. Bollina, A. Nayak, S. Ranganathan : Towards an Agent-based Information Retrieval System for Computational Biomarker Discovery, in the International Conference on Information and Communication Technology (ICICT 2007), March 2007, Dhaka, Bangladesh, pp. 57-63.

2. M.T. Islam, M. Shaikh, A. Nayak, S. Ranganathan, Biomarker Information Extraction Tool (BIET) Development using Natural Language Processing and Machine Learning, Proc.2010 IEEE/ACM Int'l Conference and Workshop on Emerging Trends in Technology (ICWET 2010), February 2010, Mumbai, India, pp. 121-126.

3. M.T. Islam, M. Shaikh, A. Nayak, S. Ranganathan, Extracting Biomarker Information applying Natural Language Processing and Machine Learning, Proc.2010 IEEE 4th Int'l Conference on Bioinformatics and Biomedical Engineering (iCBBE 2010), Chengdu, China. (accepted for publication and presentation in June 2010).

## List of Tables

## List of Figures

# Chapter 1: Introduction and Background

## 1.1 Introduction

Bioinformatics aims to provide a better understanding of biological processes using various computational methods to analyse and integrate complex genomic information that are applied to gene-based drug discovery and development. Research and technological advances such as DNA microarray, serial analysis of gene expression and mass spectrometry proteomics, genome-scale sequencing and microarray analysis produced substantial quantities of data about genes and their products [1-3]. These high-throughput technologies allow us to analyse hundreds and thousands of molecular components at a time that are making a huge contribution to the online literature repositories. Biomedical literatures are great source of information as they are experimentally tested and validated before publication. Success to any new research development is very much dependant on existing and historical data on the same domain. In a discovery process, computational biology techniques are applied to provide data analysis and statistical support for a research hypothesis using most relevant biological data [4]. This phenomenon is referred as *in silico* assays. This *in silico* or computational pre-screening is essential to narrow the scope of *in vitro* and *in vivo* research of the discovery to investigate complex biological problem, and it leads to more coherent and accurate excremental model [5]. As such, there is an increasing need to combine the analysis of data from multiple experiments with knowledge accumulated from the other kinds of analysis in system biology [6]. This information can contribute to human welfare if they are easily accessible to researchers and scientists. The underlying biomedical knowledge base is expanding at an increasing rate. For example, at present (March, 2010) PubMed and MEDLINE database contain more than 19 and 17 million records respectively [7]. Figure 1.1 shows the exploding number of articles available from MEDLINE since 1978 [8]. However majority of this information is presented in a free text format that relies heavily on manual intervention and curation tasks to extract necessary information. Therefore this enormous growth in online literature is the catalyst for automated knowledge aggregation systems. Text mining or text data mining is the process of extracting high-quality information from unstructured textual sources (articles, textual databases).

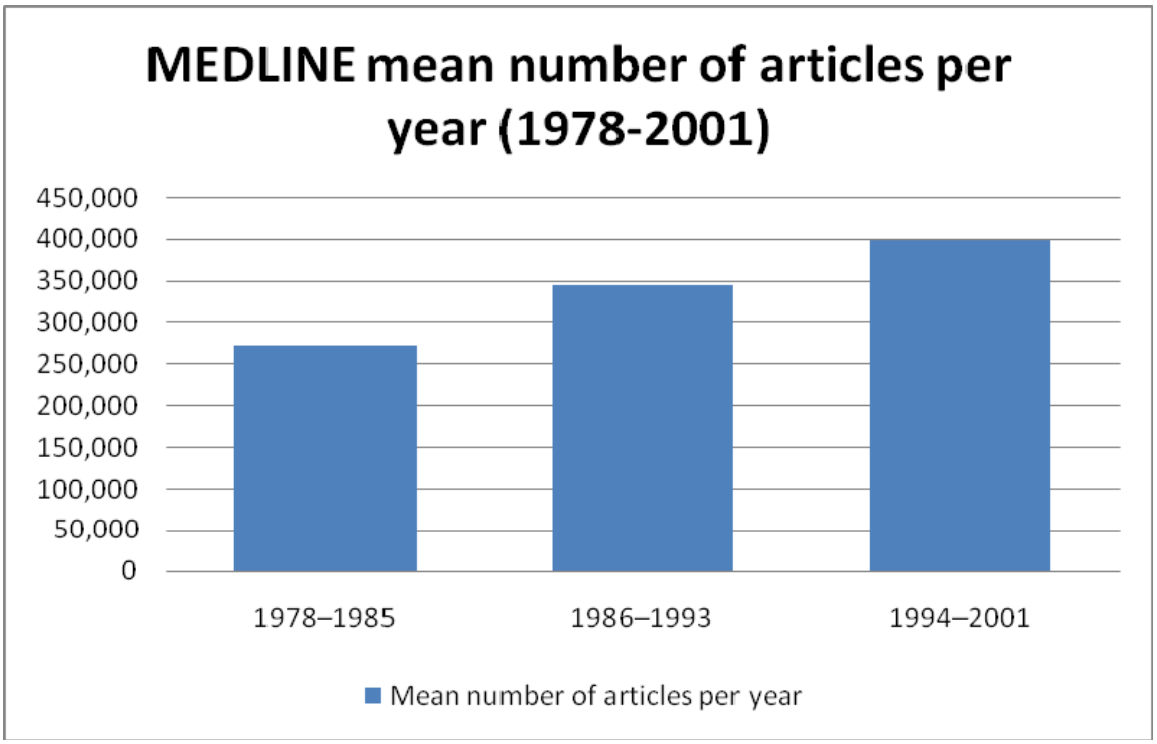


**Figure 1.1 Growth trends of MEDLINE journal articles**

Biomedical text mining helps researchers to obtain necessary information more efficiently and uncover relationships obscured by the sheer volume of available information. Text mining applies various algorithmic, statistical and data management methods to the vast amount of biomedical knowledge that are available in both structured and free text format. The advances in text mining in the biomedical domain includes but not limited to Named Entity Recognition (NER), text classification, micro-array analysis, gene expression and annotation, synonym and abbreviation extraction, relationship extractions.

Abgene [9] uses rule-based techniques to extract gene and protein name. A hybrid approach using Conditional Random Fields (CRF) and a normalizing tagger was used by Roman et al. [10] for NER. Latent Semantic Analysis (LSA) was used for acronym disambiguation. GAPSCORE [11] uses trained classifies where the features are identified by assigning numerical scores to each token within a sentence based on morphological and contextual analysis. MERGE [12] uses Maximum Entropy Markov Model (MEMM). It is capable of defining domain specific feature functions for NER. Text classification is the task of assigning one or more categories to documents based on its contents. PreBIND and Textomy [13] use Support Vector Machine (SVM) that is trained on the words from

MEDLINE abstracts to distinguish abstracts containing protein-protein interaction. It is designed to locate protein-protein interaction in the literature for classification task. Wallace et al. [14] uses support vector machine built over different feature spaces to automatically classify “relevant” and “irrelevant” citations. FigSearch [15] classifies figures from any corpus of full-text biological papers based on schematic illustrations of protein interactions and signalling events. Waree et al. [16] performs correlation analysis and Principal Component Analysis (PCA) for microarray experiment data. It uses Singular Value Decomposition (SVD) to solve the PCA problem and showed superior results compared to traditional systems. MedSummarizer [17] uses summarization algorithm on biomedical literatures to assign semantic biological meaning to gene cluster. Seki et al. [18] uses K-Nearest Neighbourhood (KNN) algorithm together with supervised term weighting schemes for gene ontology annotation task. GeneWebEx [19] uses user defined templates to identify and extract gene annotation from web-based databanks. iProLINK [20] presents a framework that links text mining with ontology in systems biology. It includes a user interface for text mining and a text mining module to create, evaluate, rank text mining outputs. Yu et al. [21] automatically extracts gene name, synonyms from full texts and Liu et al. [22] detects abbreviation and phrases from MEDLINE abstracts. While most of these works are superior in their individual perspective, they do not indicate any correlations between entities. Increasingly scientists are more interested on how these entities and facts are related to each other. As such relationship extraction tools are getting more popular. Relationship extraction tool aims to detect occurrences of a specified type of relationship between a pair of entities defined in the system. Several approaches to extracting relations of interest have been applied by researchers in different areas of interest in the biomedical domain. Ono et al. [23] and Ramani et al. [24] uses rule based techniques to detect protein-protein interactions. Glenisson et al. [25] uses vector space and k-medoids algorithm with a cosine similarity metric for text based gene clustering. Protein Active Site Template Acquisition (PASTA) [26] detects relationships between amino acid residues and their function within a protein using manually created templates, type , POS tagging and lexicons. Albert et al. [27] detects tri-occurrences of two proteins and one interaction within a sentence using dictionaries of protein and interaction terms. Automatic extraction system for biological entities (i.e. gene, protein) and diseases has been less prominent; this area is getting increasingly popular to

improve human life from the burden of diseases. Chun et al. [28] detects disease-gene relation based on dictionary lookup method that uses six public databases for disease and gene names and machine learning based NER system to extract relation. Robert et al. [29] uses machine learning algorithm for similar task form structured patients narratives. Rindflesch et al. [30, 31] extracts relation between biological entities or genetic phenomena and disease using NLP methods.

Although there has been a significant improvement in the field of Natural Language Processing (NLP) specially, text mining and data mining but scientists still lack from appropriate tools to extract or infer answers to specific queries where information is usually linked up to another information. This thesis addresses the two main phenomena of scaling up the computational knowledge aggregation (i.e. information retrieval and relationship extraction) on drug discovery process to make the best use of the growing quantities of published biological texts. Firstly, we have developed an automatic information retrieval, summarization and extraction tool that is capable of retrieving relevant document from the online repository, extract scientifically important sentence and summarize the document. We then go further by developing a sophisticated tool to extract relationship among the biomedical entities within the extracted text from the first system. Both systems are built on advanced computing technologies such as natural language processing, statistical analysis and machine learning algorithm for the respective aggregation tasks. The information obtained by our method can be extended to interpret and validate new research hypothesis and high-throughput experimental results.

Our knowledge aggregation approach enables biomedical researchers to obtain domain specific information (i.e. disease, gene, protein and biomarker) from a large-scale of datasets comprising thousands of records without needing to have computational knowledge. The results of these applications demonstrate the effectiveness of our approach, and its applicability to the biomedical drug discovery process and similar tasks. The applications presented are relevant to the biomedical fields of biomarker and drug discovery. These applications were chosen as proof of concepts, and the applications of the approach presented here are not limited to these fields.

## 1.2 Objectives

In view of the emerging research growth in biomedical information in the World Wide Web, this thesis attempts to comprehend the information flow to develop knowledge extraction and discovery software framework that can be useful for the researchers for future data analysis without having much domain knowledge or computation skills. This thesis presents two novel tasks of knowledge aggregation tasks; an information retrieval tool to retrieve the information from the web and a relationship extraction tool to extract relationship from the retrieved text.

A biomarker is a biochemical indicator of a biologic state that may serve as an indicator or predictor of a disease. Biomarker is used to measure presence, risk, progress or the effect of treatment of a disease rather than measuring the disease itself. Biomarkers are used for disease characterization and diagnosis and remains very popular and active research domain. Majority of these researches aim to find effective response to dose and early detection of adverse events in the patient population.

According to Cancer Australia [32], in 2004 more than 98,000 non-melanoma related new cases were diagnosed in Australia and an estimated 382 000 were treated in the previous year. In 2009, the estimated new figures for new cases were 110,000 and estimated 42,000 were expected to die in the same year as cancer continues to be considered as one of the leading cause of death in Australia. One in 2 men and 1 in 3 women are expected to be diagnosed with cancer before the age of 85. The direct cost of cancer to Australian health system is $3.8 billion a year and approximately $378 million was spent in cancer research just in 2000-01.These figures are alarming as Australian Institute of Health and Welfare reports, the new cases figures are expected to increase 29% for women and 32% for men in 2011 [33].

Evidently the current drug discovery tools are considered to be insufficient in catering the high performance needs in the current process as well as there is a growing need to fast track the discovery process to keep up with these increasing threat to human life. Hence our aim is to avail oneself of the multitude of analytical tools that can assess new biomarkers as well as the existing data to validate new

claims. In this thesis we choose to deal with Cancer Biomarkers as the proof of concepts.

In the next section, we describe our problem statement with example and chapter 2 of this thesis contains the detail description of our proposed method and solution.

### 1.2.1 Problem Statement

Here we first describe a problem scenario. For example, a researcher wants to know the biomarkers of *Lymphocytic Leukemia.* In a typical scenario he or she will need to go to existing online journals or databases and use different search strings to get relevant documents. A search string "Biomarker of Lymphocytic Leukemia" was used by the user on the PubMed repository. The search returned over 7200 literatures. Now he or she needs to read through all these literatures and documents get the information on biomarker. By the time he or she finishes reading the 7200 literatures, more literatures are added to this repository and the

**FCRL2 expression predicts IGHV mutation status and clinical progression in chronic lymphocytic leukemia.**

Li FJ, Ding S, Pan J, Shakhmatov MA, Kashentseva E, Wu J, Li Y, Soong SJ, Chiorazzi N, Davis RS.

Division of Hematology/Oncology, Department of Medicine, University of Alabama at Birmingham, AL 35294-2182, USA.

Comment in: Blood. 2008 Jul 1;112(1):2-3.

CD38 and ZAP-70 are both useful prognostic markers for B-cell chronic lymphocytic leukemia (CLL), but are variably discordant with IGHV mutation status. A total of 5 human Fc receptor-like molecules (FCRL1-5) have tyrosine-based immunoregulatory potential and are expressed by B-lineage subpopulations. To determine their prognostic potential in CLL, FCRL expression was compared with IGHV mutation status, CD38 and ZAP-70 expression, and clinical features from 107 patients. FCRL1, FCRL2, FCRL3, and FCRL5 were found at markedly higher levels on CLL cells bearing mutated IGHV genes than on unmutated CLL cells or CD19(+) polyclonal B lymphocytes. Univariate comparisons found that similar to CD38 and ZAP-70, FCRL expression was strongly associated with IGHV mutation status; however, only FCRL2 maintained independent predictive value by multivariate logistic analysis. Strikingly, FCRL2 demonstrated 94.4% concordance with IGHV mutation compared with 76.6% for CD38 and 80.4% for ZAP-70. Compared with other indicators, FCRL2 was also superior at predicting the time to first therapy; the median treatment-free interval was 15.5 years for patients with high FCRL2 expression compared with 3.75 years for FCRL2-low patients. Our studies indicate that FCRL2 has robust predictive value for determining IGHV gene mutation status and clinical progression and thus may further improve prognostic definition in CLL.

PMID: 18314442 [PubMed - indexed for MEDLINE]

**Figure 1.2 Sample Text – proof of concept**

list goes on. In layman's term, our researcher is only interested to know the biomarkers of X disease; he or she is not concerned about all other information that is available in these literatures at this stage. So the intention is to get the most updated data at any given time without having to go back to the repository everyday to look for updates. This sounds like a tedious and rather impossible task in the absence of automated software tools.

If we take a closer look at this problem and consider the sample text shown in figure 1.2; from systematic perspective, following needs to happen

- Automated software needs to be able to search online repositories to collect all the papers that are related to *Lymphocytic Leukemia*. The system should also be able to consider synonyms; acronyms etc., of the search term and retrieve all papers (i.e. Lymphocytic Leukemia VS CLL).
- Once the documents are retrieved, the system then needs to analyse the information and suggest the user about the biomarkers of a given disease. In this example (figure 1.3), *FCRL2, CD38* and *ZAP-70* are biomarkers of *Lymphocytic Leukemia*.

In order to do this, the system needs to solve the following problems:

- The system needs to identify target entities like gene, protein and diseases (highlighted texts)
- The system needs to understand syntactically and semantically important sentences (underlined sentences).
- The system needs to understand the features or phenomenon to decide the relationship.

- For example in the first underlined sentence (title) - *"FCRL2 expression predicts IGHV mutation status and clinical progression in chronic lymphocytic leukemia",* in this sentence, the author of the paper indicates that *FCRL2* is a biomarker of *chronic lymphocytic leukemia*. So the system needs to be able to identify this relationship by understanding important words and features i.e. the words, *predict, clinical progression*.
- In the 2nd underlined sentence – '*CD38 and ZAP-70 are both useful prognostic markers for B-cell chronic lymphocytic leukemia (CLL), but are variably discordant with IGHV mutation status' –* the author mentions that CD38 and ZAP-70 are biomarkers of *chronic lymphocytic leukemia (CLL).* So the systems needs to

identify this relationship by understanding the key features or clues left by the author (i.e. the phrase *useful prognostic markers.*)

- In the third underlined sentence – '*Our studies indicate that FCRL2 has robust predictive value for determining IGHV gene mutation status and clinical progression and thus may further improve prognostic definition in CLL'* – the author again mentions *FCRL2* to be the biomarker of *CLL* or *chronic lymphocytic leukemia* and this time author backs up the idea with clinical findings and leaves some important clues like, *studies indicate, predictive value, clinical progression, improve prognostic* for the system. So our system needs be able to identify these features to suggest the user about the biomarker of a given disease.

### 1.3 Biomedical Text Mining

Biomedical text mining is a computation process of analysing textual data by automatic or semi automatic means to discover new, previously unknown information or rediscovering existing information. The volume of scientific literature has created an increased interest in linking the entities and concepts in unstructured texts. Several systems have been built to accomplish specific mining tasks. iHop [34] extracts annotation and detects interactions, PreBind [13] uses machine learning approach to extract protein-protein interactions. Biomedical discovery support system (BITOLA) [35] uses association rules from MeSH [36] descriptors to detect candidate genes for diseases and indirect relationships, EUCLID [37] classifies proteins into functional groups based on SwissProt keywords, CaRE [38] detects semantic categories and associated relationships in medical discharge papers. SemGen [30] characterizes the semantics relationship among the entities within the text. EDGAR [31] extracts gene, drugs and cell line for cancers. Chilibot [39] is a web-based text mining application that extracts term-term relationships. Some other similar systems are text mining tool for microarrays microGENIE [40], PubGene [41], Medgene [42] and Geisha [43].

A typical text mining process flow consists of two major tasks. First task is to extract important information from relevant documents. This task is called Information Retrieval and Information Extraction. Second task is to discover specific knowledge from the relevant articles and extracted texts. This is called knowledge discovery. As depicted in Figure 1.1, generally a text mining process begins with collections of unannotated, untagged raw articles. These articles are

then automatically ranked by their relevance to the system goal. Articles with higher ranks are then tagged by categories; terms and evaluative information are extracted directly from individual documents. Finally the extracted categories, entities and evaluative information are used to support a range of data mining operations on the articles.

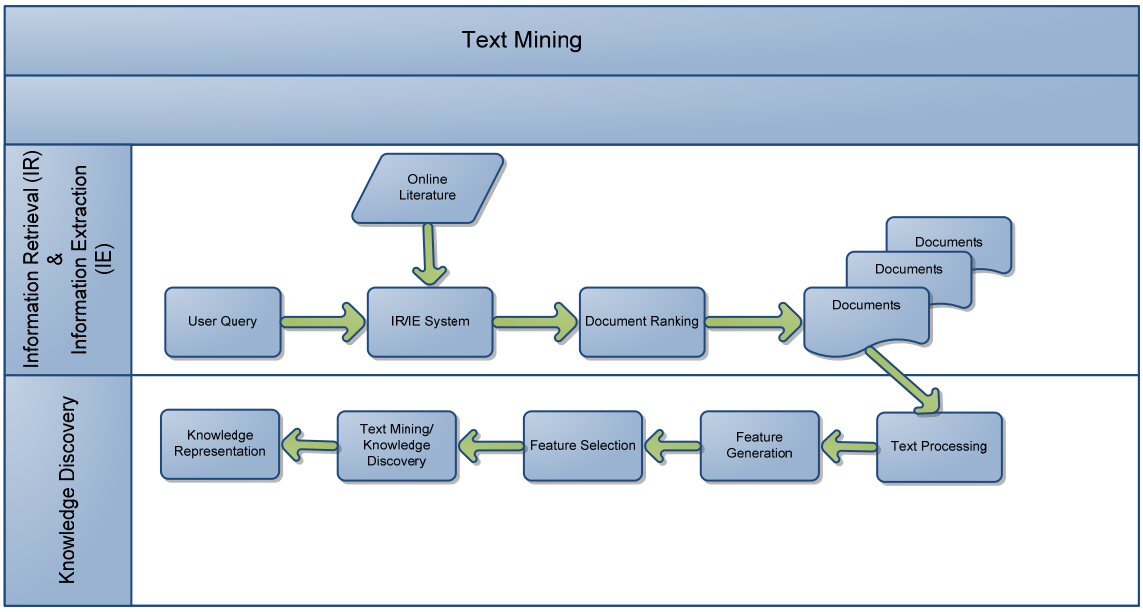


**Figure 1.3 Typical Text Mining Process Flow**

In the next two subsections we will briefly discuss the main components of these tasks

### 1.3.1 Information Retrieval & Information Extraction

The parallel advances in information technology and biomedical research has embraced researchers with thousands of useful documents in their finger tips. Information sources are no longer limited to printed books; internet based technologies now allow us access to almost every information that is available in different journals and online databases. The major drawback of this information flood is, most of them are presented in unstructured texts and it is impossible for any individuals to go through all of them to find all relevant information. This initiates the immediate need of automated complex systems that are capable of extracting the scientifically important information for researcher. Natural Language Processing (NLP) is one such technical advance that provides techniques and methodology to derive such information from free texts. NLP mostly uses different

rule based techniques to determine the syntactic and semantic associations of biomedical entities to retrieve texts.

The first and foremost task of a text mining process is to locate the relevant articles. Given a set of source of articles and a user query, Information Retrieval (IR) system aims to find a set of articles that are relevant to the query and Information Extraction (IE) system aims to find the sentences with relevant information and extracts relevant information and presents the information in a predefined format. Commonly used automatic IR techniques are rule-based, linguistic, statistical, machine-learning, and hybrid approaches. Two major supervised tasks of IE are ranking and classification. The task of ranking is to measure how relevant documents or texts based on a given user query. Keyword based Term Frequency – Inverse Document Frequency (TF-IDF) [44] is one of the common, simple and effective ranking techniques that aim to rank document or texts based on the keyword statistics in a given corpus. Classification aims to categorize documents or texts to one or more category. Documents are treated as a bag of strings and strings are considered as bag of characters. Semantically or syntactically distinct characters, words correspond to a feature that is used for statistical and Machine Learning (ML) based system for classification. Again, TF-IDF [44] technique can be used to reduce the magnitude of the feature vectors. We will discuss more about supervised and unsupervised techniques in the next sub section.

### 1.3.2 Knowledge Discovery & Knowledge Interpretation

To date almost every IR tools returns large number of documents in response to user queries. This phenomenon is truly driven by the vast growth of online literature due to research development and not because of the lack of advance IR techniques. Although IR tools limit the number of retrieved documents to a theoretically manageable figure; in practical the human driven manual analysis of such information are still very labour intensive and costly. Hence the computational biologists need to come up with automatic knowledge interpretation, discovery and decision support system that allows users to get the exact information in their finger tips.

Machine learning (ML) techniques are proven to be very effective for such tasks. Machine learning is a process that facilities a computer program to improve its performance based on previous experience. Machine learning techniques are divided in to two types, supervised and unsupervised techniques. Supervised technique uses training data and feature to learn the agent for classification. Training data includes input objects and desired outputs that are based on observations. Supervised system then predicts output of a given input after analysing the training data. These features, written by the developers most often focus on the syntactic, semantic and lexical knowledge of input and output texts. On the other hand unsupervised techniques do not rely on predefined input rules. The focus is on how the data is organized in the unlabeled inputs. The key differences to these approaches are supervised techniques requires sufficient number of well defined target variables and their values whereas in unsupervised learning target variables are recorded in small number of cases or some cases they are unknown. In general unsupervised methods are more preferable as they are less labour intensive but it requires large set of training data. Despite the recent advances in biomedical research, this domain still lacks on appropriate training data. Hence supervised models are considered to be much more superior to the unsupervised models for biomedical text mining. In our approach we used supervised Support Vector Machine (SVM) algorithm [45-49].

As shown in figure 1.3, generally a knowledge discovery process consists of five major tasks; text processing, feature generation, feature selection, knowledge discovery and knowledge representation. Text processing is the syntactic and semantic analysis of text that combines different NLP tasks such as parts-of-speech tagging, word sense disambiguation, parsing, named entity recognition etc. In Feature generation texts are considered as bag of words and syntactic, semantic issues including their lexical information are considered. These features are used during the learning phase. In order to reduce the dimensionality, stemming techniques are used to take the root form the words. All learners require reducing dimensionality to improve the quality and efficiency. Statistical and mathematical techniques are used to select the important features and ignore any unnecessary features. In the knowledge discovery phase, a collection of labelled records that contains its feature attribute from the previous phase are used to assign a class to input texts. The text data set is divided into training and test sets.

The training set is used to build the model and test set used to validate the same. Final task of the process is to represent the gathered knowledge or idea to desired form which can range from visual representation, decision tree, database or even plain text format.

### 1.3 Key Contributions

The original work presented in this thesis makes several important contributions to the fields of bioinformatics, biomedical knowledge mining and drug discovery, which are summarized here:

- A new biological knowledge mining framework for modelling “second generation” biological discovery processes, in which knowledge flows through analysis pipelines consisting of multiple cascaded tasks.
- A novel user-driven intelligent agent based text mining method for information retrieval and summarization. The method is customizable by end users without the need for programming. Knowledge acquired through this process is injected in to the next text mining process that is used for other sophisticated knowledge discovery tasks.
- A novel knowledge aggregation method to extract biomedical entity relationship based on structural rules for extracting, aggregating and reconciling information from multiple heterogeneous biological data sources, regardless of their native data structures.
- A proof-of-concept demonstration that aggregated biological knowledge, expressed using complex NLP, statistical and machine learning techniques can be processed by software tool that does not require users to have domain knowledge.
- A biomarker discovery tool to discover biomarker from scientific articles that can act as the basis of new research hypothesis or validating newly discovered biomarkers. This work can be extended to discover new markers by analysing the discovered relationships.

## 1.4 Structure of the thesis

This thesis introduces the field of Biological Knowledge Aggregation and Biological Knowledge Mining, analysing the current problems inherent to the evolution of bioinformatics from small-scale entity-based discovery leading to large-scale systemic discovery.

- Chapter 1 provides an introduction, background, aims, contributions and structure of the works are presented.
- Chapter 2 describes the methods and applications for the presented work.
- Chapter 3 discusses presented information retrieval and summarization tool that selects relevant abstracts and full particles from online document repositories (i.e. PubMed) based on user queries and summarizes them.
- Chapter 4 presents a text mining method which selects relevant abstracts from our previous system (chapter 3) to discover the entity relationships from the extracted abstracts. It uses machine learning techniques to identify biomedical relationship from the free texts. We show that this method can substantially reduce curation workload.
- Chapter 5 contains the overall conclusions. The contributions of this thesis are summarized and reviewed, and future research directions are discussed here.

## Chapter 2: Methods and Applications

Figure 2.1 shows the top level architecture of the system that incorporates two systems; mExtract & BIET. We first attempted to solve the information retrieval problem and presented a system (mExtract) that will retrieve topic specific literature from the web, statistically rank the documents and produce a summary of the topic which is detailed in chapter 3.

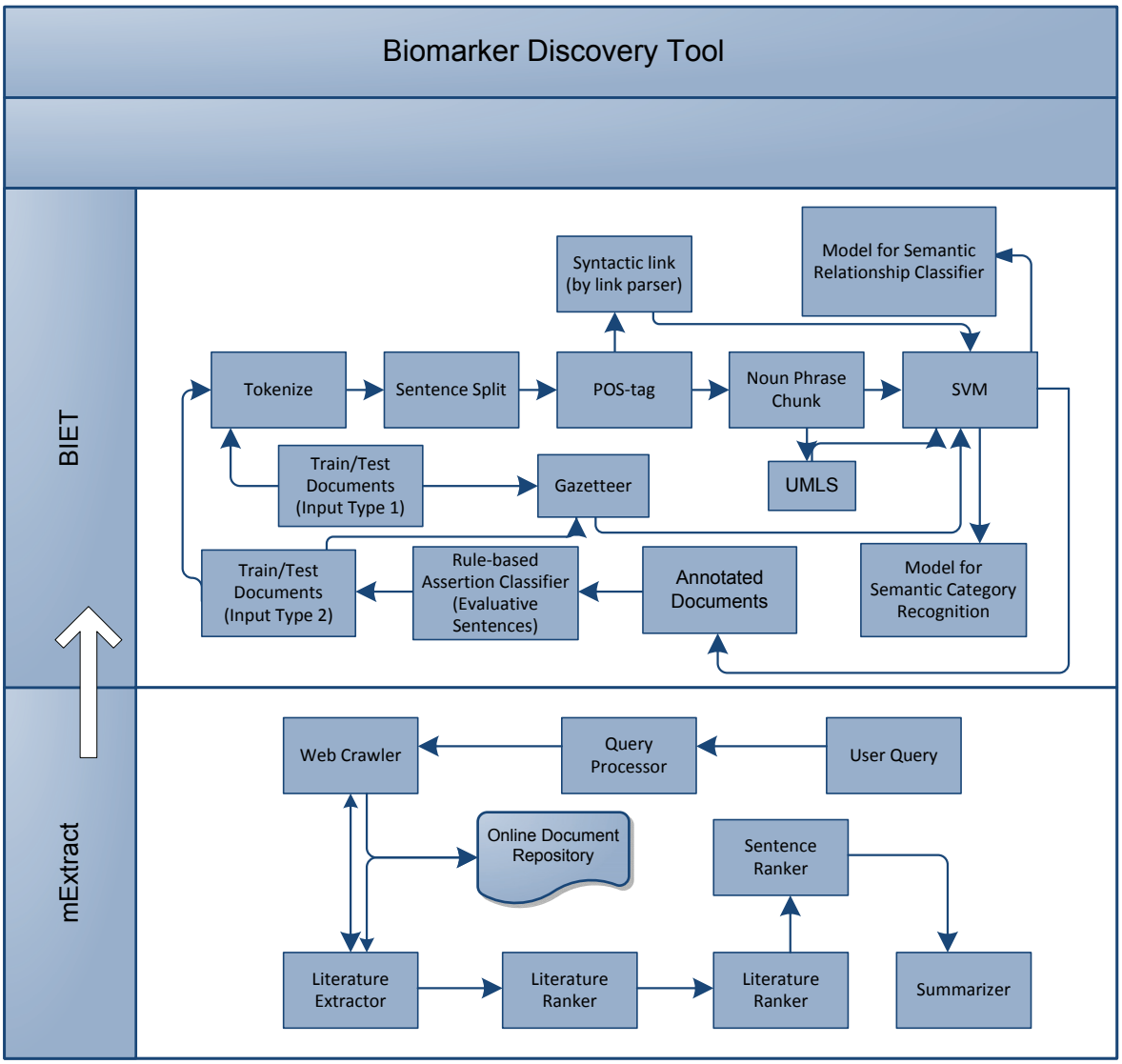


**Figure 2.1 System Architecture**

In chapter 4, we detailed our approach to address the second problem and present a relationship extraction tool (BIET) that is capable of extracting biomarker relationship from the scientific literatures. This tool utilizes our first solution for information retrieval tasks then it goes through a series of complex NLP and learning pipeline that ranges from named entity recognition to defining relationship

and stores the information in a database. System defines the output as *is_biomarker_of (g/p,d)* where '*g/p'* represents gene/protein name respectively and '*d'* represents disease name. In other words, based on our sample text in figure 1.3; a sample output would be is_biomarker_of (*FCRL2, lymphocytic leukemia).*

**Table 2.1. A list of the methods and applications developed during this work**

| **Method/Application** | **Chapter** | **Refer to publication number(s)** |
|---|---|---|
| Information Retrieval and Summarization | 3 | 1 |
| Biomarker and Relationship Extraction | 4 | 2,3 |

The methods and applications developed in this work are summarised in Table 2.1. Most of the work presented in this thesis along with its methods and key results has been published in international peer-reviewed conferences during the course of the candidature period. Within the scope of this work, the present author published three papers as first author. Detailed descriptions of each item can be found in the corresponding publications or section.

# Chapter 3: Design and Development of Information Retrieval, Extraction and Summarization Tool

## 3.1 Introduction and significance

Automated information retrieval (IR) and information extraction (IE) tools are widely used as the initial source of accessing the knowledge from online document repositories. Over the past few decades a large number of such systems were developed for both commercial and non-commercial purposes. A number of these tools also provide auto summarization features. However most of these systems including the commercial ones use keyword based techniques to retrieve documents and extraction. Evidently information overload is the main driven factor to these developments, but the key challenge to this development is how the information is presented in the web. Most of these documents are not only in free text format but also contains conceptual information, especially biomedical literatures. It is important that IE systems are capable to deal with conceptual or semantic meaning of terms for its tasks.

In this paper we present an intelligent agent based searching phase and summarization phase for biomedical literature using hybrid NLP techniques. Our developed system first uses conceptual techniques to expand user query then uses traditional statistical and pattern matching techniques to retrieve the relevant documents. System then extracts the most important scientific information and biological phenomenon from the retrieved documents. In addition to this; based on user's selection, our system generates an extractive or abstractive summary of the retrieved document.

## 3.2 Results presented as publication articles

Although traditional keyword and index based commercial tools seem to suffice the need for most individuals, but this doesn't suffice the need of research communities. Understanding the underlying concept remains the key element for automatic knowledge discovery. Traditional approaches to IR system are often centralized and hierarchical that is not feasible for large scale computing tasks.

Multi-agent systems consist of multiple autonomous interacting agents to complete complex tasks that has a number of advantages over traditional approaches.

- A multi-agent system efficiently retrieves, filters, and globally coordinates information from spatially distributed sources.
- It also enhances overall system performance, specifically along the dimensions of computational efficiency, extensibility, robustness, maintainability, reliability, responsiveness and reusability.

Considering the aforementioned advantages, we presented a multi-agent framework for IR and IE system that is more favourable to academic and scientific research needs. Here we have integrated a number of agents to perform specific tasks to accomplish the IR and IE tasks that is capable of retrieving information from multiple online repositories and capable of processing huge textual data simultaneously. Extracted information is stored in machine readable formats that can be used for further data analysis and more complex knowledge aggregation tasks. In this project we used automatic extraction of knowledge from different information sources to create a conceptual structure to enhance system performance. Information sources used in this project are utilized as proof of concept, more advanced information source like Unified Medical Language System (UMLS), Medical Subject Heading (MeSH) etc. can be used to obtain domain specific concept to improve performance.

Pages 18-24 of this thesis have been removed as they contain published material. Please refer to the following citation for details of the article contained in these pages.

Islam, M. T., Bollina, D., Nayak, A., & Ranganathan, S. (2007, March). Intelligent agent system for bio-medical literature mining. In *2007 International Conference on Information and Communication Technology* (pp. 57-63). IEEE.

# Chapter 4: Biomarker Extraction Tool Development

### 4.1 Introduction and significance

Biomarkers play a vital in role in drug discovery process. Biomarker discovery is one of the popular and active research fields. Scientists need to have access to existing research advances and knowledge to validate new findings. In order to combat the enormous growth of scientific document, automatic knowledge discovery and decision support systems are necessary. The rationale is not just to find existing relationships between entities, such systems needs to be able to do complex analysis to create new research hypothesis. In Paper 2, we focus on a specific task of finding biomarkers of diseases from online literatures by finding the relationship between the entities (i.e. gene, protein and disease).

Our approach employed Support Vector Machine learning classifier in conjunction with other complex rule based techniques to accomplish this task. The classifier is trained with manually-annotated sets of oncology related documents that contain documents of both interest i.e. positives and negatives. The best results to date have been obtained as a result of laborious choices of algorithms and document features, to suit the specifics of this particular problem (Paper 3).

### 4.2 Results presented as publication articles

In this project, we have reviewed semantic learning technologies as a knowledge representation layer for biological knowledge mining. We have shown that the combination of machine learning and rule based NLP techniques allows flexible and extensible encoding of knowledge, and therefore supports the flow and augmentation of knowledge necessary for biological knowledge mining. The study presented in this paper showed that carefully selected semantic features are a powerful addition to semantic knowledge representation, and are capable of restructuring and extending existing knowledge through the developed application that can be used to build much more complex system for system biology.

# Biomarker Information Extraction Tool (BIET) Development using Natural Language Processing and Machine Learning

Md Tawhidul Islam
Department of Chemistry and Biomolecular Science, Biotechnology Research Institute, Macquarie University, Balaclava Rd, NSW 2109, Australia
+61 2 9485 1247
md.islam@students.mq.edu.au

M Shaikh
Dept. of Information and Comm. Engineering University of Tokyo 7-3-1 Hongo, Bunkyo Ku Tokyo 113-8656, Japan
+81-3-5841-6767
almasum@gmail.com

A Nayak
Department of Computing Macquarie University, Balaclava Rd, NSW 2109, Australia
+61 2 9850 9565
abhaya@science.mq.edu.au

S Ranganathan
Department of Chemistry and Biomolecular Science, Biotechnology Research Institute, Macquarie University, Balaclava Rd, NSW 2109, Australia
+61 2 9850 6262
shoba.ranganathan@mq.edu.au

## ABSTRACT

In recent years, there has been a rising interest in extracting entities and relations from biomedical literatures. A vast number of systems and approaches have been proposed to extract biological relations but none of them achieves satisfactory results due to the failure of handling the grammatical complexities and subtle features of biomedical texts. In this paper, we detail an approach to a very specific task of information extraction namely, extracting biomarker information in biomedical literature. Starting with the abstract of a given publication, we first identify the evaluative sentence(s) among other sentences by recognizing words and phrases in the text belonging to semantic categories of interest to bio-medical entities (semantic category recognition). For the entities like, protein, gene and disease, we determine whether the statement refers to biomarker relationship (assertion classification). Finally, we identify the biomarker relationship among the bio-medical entities (semantic relationship classification). Our approach utilizes a series of statistical models that rely heavily on local lexical and syntactic context and achieve competitive results compared to more complex NLP solutions. We conclude the paper by presenting the design of a system namely, the Biomarker Information Extraction Tool (BIET). BIET combines our solutions to semantic category recognition, assertion classification and semantic relationship classification into a single application that facilitates the easy extraction of semantic information from medical text. We designed and implemented ML-based BIET system for biomarker extraction, using support vector machines and trained and tested it on a corpus of oncology related PubMed/MEDLINE literatures hand-annotated with biomarker information. Several tests are performed to assess the performance of the system's component namely semantic category recognizer, assertion classifier and semantic relationship classifier and the system achieves an average F-score of 86% for the task of biomarker information extraction comparing to the human annotated dataset (i.e. gold standard) scores.



ICWET'10, February 26–27, 2010, Mumbai, Maharashtra, India.



## Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data Mining; I.2.4 [**Knowledge Representation Formalisms and Methods**]; I.5.4 [**Applications**]: Text Processing; J.3 [**LIFE AND MEDICAL SCIENCES**]: Medical information systems;

## General Terms

Algorithms, Documentation, Design, Experimentation, Security, Human Factors, Standardization, Languages, Theory, Verification

## Keywords

Biomarker, relationship extraction, text mining, literature mining.

## 1. INTRODUCTION

Advances in biomedical research have produced an unprecedented growth in the volume and diversity of biological data. Majority of this research is focused on furthering our understanding of cellular and molecular mechanisms that are critical to the design of targeted therapies and preventions. PubMed/MEDLINE repository is growing exponentially with new publications [15] which makes it almost impossible for researchers to keep up with the relevant publications in this domain. Vast amounts of this knowledge are only presented in free-text format which are not readily available for automatic computerized analysis. To answer complex research questions, bioinformatics analysis needs to aggregate increasing quantities of information from mounting number of diverse sources, combining multiple tasks into analysis pipelines.

The information found in these research papers is important for biologists, chemists, pharmacist doctors and researchers. Researchers trying to find statistics on patients with X and Y disease or symptoms linked to gene or protein related cell damage records. The problem is that this information is not readily available. The project described here takes another step forward towards accomplishing relationship extraction task in the direction of identifying which is the biomarker of which cancerous disease. Many programs have been developed to help with similar problem. Much of the work done so far has focused on discovering gene-gene relations, protein-protein relations or protein- protein interactions [1][2][9][10][13][14] from biomedical documents. Category and Relationship Extractor

(CaRE) [16] system finds semantic categories and their relationships in medical discharge papers. Other methods include, gene-to-gene co-citation network [7][17], co-occurrence based [10], rule-based [15], kernel-based [3, 18] relationship extraction. EDGAR [11] extracts gene, drugs and cell line for cancers. Other approaches attempts to extract and characterize the type of relation between entities, SemGen [12] which attempts to characterize the semantics of the relations based on whether a gene causes, predisposes, or is simply associated with a disease. NLP methods are applied to generate a set of candidate relationship features, which are evaluated by biological experts to generate a final set of relationship features [12].

Our approach uses Support Vector Machine (SVM) classifiers to learn these relationships. The classifiers are trained and evaluated using novel data: a gold standard corpus of oncology narratives, hand-annotated with semantic entities and relationships. We describe a range of experiments that were done to aid development of the approach and to test its applicability to the biomarker studies. We train classifiers using a number of different features sets and investigate their contribution to system performance. These sets include some comparatively simple text-based features and others based on a linguistic analysis, including some derived from a full syntactic analysis of sentences. Clinically interesting relationships may span several sentences, and so we compare classifiers trained for both intra- and inter-sentential relationships (spanning one or more sentence boundaries).

The rest of this paper is organized as follows: Section 2 discusses the primary task of BIET which is semantic category recognition and assertion classification. Section 3 describes the secondary task of the tool which is semantic relationship classification. Since we are utilizing General Architecture for Text Engineering (GATE) [5] to implement this tool in the subsequent section the tool is explained from GATE's perspective. Therefore, section 4 discusses GATE and why it was chosen as a framework for our research. Section 5 discusses the Link Grammar parser and its usage inside GATE. Section 6 describes the modified GATE plugin used to preprocess the documents. Section 7 illustrates the SVM capabilities of GATE. Sections 8 and 9 cover the results and conclusions.

## 2. SEMANTIC CATEGORY RECOGNITION AND ASSERTION CLASSIFICATION

There are two parts to this task. The first part, referred to as semantic category recognition, is to identify the semantic category of each sentence in a given abstract. We defined two semantic categories of interest: evaluative sentence, and non-evaluative sentence. Each sentence in the corpus must be classified with one of these two categories. For example, given the following three sentences the system would classify the first sentence as non-evaluative and the other two as evaluative sentences based on the medical entities described therein.

*"In this study we examined the prognostic value of SR-A1 gene expression using a semi-quantitative RT-PCR method. High SR-A1 expression was observed in 31/81 (38.3%) breast cancer tissues and was found to be more frequent in patients with tumors of large size (p=0.027), as well as in lymph node-positive patients (p=0.035). Our results suggest that SR-A1 may possibly be characterized as a new marker of unfavorable prognosis for breast cancer."*

To fully understand the implication of medical findings, we must be able to distinguish between positive, negative, and uncertain assertions of these problems, as the following examples illustrate:

• *SR-A1 expression was observed in ... breast cancer tissues and was found to be more frequent in patients with tumors of large size ....*

• *… SR-A1 may possibly be characterized as a new marker of .... breast cancer*

• *… clinical impact of TRAG-3 in ovarian carcinoma has not been demonstrated previously.*

In the first sentence, the cancer diseases is asserted as being identified by the gene, in the second case the disease possibly is identified by gene, and in the final case the disease is asserted as not being associated with the gene. For each problem the second part of our task is to distinguish between the three possible scenarios highlighted as: the entity (i.e., gene or protein) states the diseases; the entity may have reference to the diseases; and the entity does not have the reference to the diseases. We refer to this as assertion classification.

The statistical semantic category (SC) recognizer is our solution for semantic category recognition. We frame the problem as a binary classification task. Given a sentence, the statistical SC recognizer considers each word in isolation and uses SVMs with linear kernel to classify the sentence as belonging to either evaluative text or non-evaluative text category. The statistical SC recognizer incorporates the following features that capture the contextual, ontological and surface cues that human annotators use in determining semantic categories of the sentences:

- The targets: disease (i.e., name of the cancer), gene and protein.
- Left and right lexical bigrams of the targets
- The heading of the section that the targets appear in.
- Syntactic bigrams of the targets.
- The head of the noun phrase that the targets is part of and the syntactic bigrams of the head.
- The part of speech of the targets and the words within a +/- 2 context window.
- The UMLS semantic types of the noun phrase containing the target.
- Whether or not the target is a diseases name.
- Whether or not the target is a gene name.
- Whether or not the target is a protein name.

Using these features, the statistical SC recognizer obtains F-measures above 90% for most categories. These results are significantly better than the performance of a baseline, which simply maps phrases in the text to UMLS semantic types.

To distinguish between positive, uncertain and negative assertions of each medical evidence identified by semantic category recognition, we employ a regular-expression-based algorithm referred to as the rule-based assertion classifier. Initially, we used a development corpus to create the following dictionaries:

- Common phrases that precede or succeed gene or protein

entity and imply that the disease is not associated (later referred to as negation phrases).
- Common phrases that precede or succeed gene or protein entity and imply that the disease shows uncertain relationship (later referred to as uncertainty phrases).

Common phrases that precede or succeed a gene or protein entity and imply that the disease is associated with entity (later referred to as positive phrases).

To classify a problem, the rule-based assertion classifier determines which phrases occur within a four-word window of the problem. The classifier is greedy, first checking for positive phrases, followed by negation phrases and finally uncertainty phrases. If at any stage a match is found, the assertion classifier labels the problem according to the assertion implied by the matching phrase. If no match is found we label the statement as unknown and discard to process. The rule-based assertion classifier achieves F-measures above 90%.

## 3. SEMANTIC RELATIONSHIP CLASSIFICATION

Given two concepts (i.e. in our case gene/protein and disease) in a sentence, the final task is to define the relationship between them. Hence, we focused on interactions involving gene/protein and diseases; for simplicity we defined one type of binary relationships that encapsulate most of the information pertaining to medical entities: relationships between diseases and genes/proteins. In this case, we define the following possible relationships between disease and research evidence related to gene or protein: The gene/protein is a biomarker of the disease (i.e. explicit biomarker) or the gene/protein maybe a biomarker of the disease (i.e. implicit biomarker).

Our statistical semantic relationship (SR) recognizer consists of SVM classifier corresponding to the aforementioned binary relationship types. Thus, there is an SVM classifier for relationships between explicit or implicit biomarkers. Unmarked input text is passed through our statistical semantic category recognizer and the rule-based assertion classifier that mark semantic categories of evaluative text and problem assertions respectively. For each sentence in the text and for each candidate pair of entities covered by one of our relationship types (for example, the gene is a certain biomarker of a disease relationship), the statistical SR recognizer uses the SVM classifier to determine which specific relationship exists between the entities of the evaluative texts.

We list the features for our SVM classifier as follows:

- The number of words between the candidate entities.
- Whether or not the disease precedes the gene/protein entity.
- Whether or not other entities (e.g. medical tests) occur between the disease and gene/protein.
- The verbs between the disease and the gene/protein entity.
- The two verbs before and after the disease and the two verbs before and after the gene/protein.
- The headwords of the disease and the gene/protein related noun-phrases.
- The right and left lexical bigrams of the disease and the gene/protein.
- The right and left syntactic bigrams of the disease and the gene/protein.
- The words between the disease and the gene/protein.
- The path of syntactic links (as found by the Link Grammar Parser) between the disease and the gene/protein.
- The path of syntactically connected words between the disease and the gene/protein.

The modularized design of the system is shown in Figure 1.
We have two SVM models to deal with two types of inputs. The first model (i.e. model for semantic category recognition) deal with the abstract text documents and outputs the evaluative and non-evaluative sentences. The second model (i.e. model for semantic relationship classifier) deals with the evaluative sentences to output the explicit and non-explicit biomarker relationships of a disease with respect to gene or protein. The modular application also allows us to customize and extend the program with the interchangeable components.

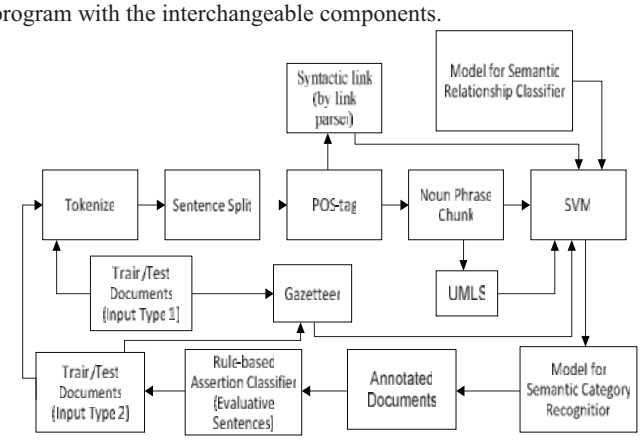


**Figure 1: Flow diagram of BAIT**

## 4. GATE

The General Architecture for Text Engineering (GATE) is a free Java-based software available online, widely used for creating text mining projects and for natural language processing (NLP). GATE also has a built-in "Annotation Diff" tool to measure precision and recall. GATE is versatile; it works on many different operating systems and can process several other languages besides English [5].

According to [8], GATE is made up of many plugins that can be put together in a pipeline to form an application. This plugins can be used as is or modified to fit particular needs of a project. The tokeniser module splits text into simple tokens, such as numbers, punctuation, symbols, and words of different types. The sentence splitter is a cascade of finite state transducers which segments the text into sentences. This module is required for the tagger.

The gazetteer consists of lists such as cities, organizations, days of the week etc. It not only consists of entities but also names of useful indicators, such as typical company designators that are compiled into finite state machines, which annotate the occurrence of the list items in the given document. Gazetteer lists can be extended or modified to fit particular needs of a project. We have supplied an extensive list of gene, protein and disease names to the gazetteer and also matched the record of these entities in UMLS database for further information.

The *JAPE transducer* is the module that runs *JAPE* grammars, which could be doing tasks like chunking, named entity recognition and so on. GATE is supplied with an *NE transducer,* which performs named entity recognition for English. The *orthomatcher* performs co-reference or entity tracking by recognizing relations between entities based on orthographically matched names. It also improves named entity recognition by assigning annotations to previously unclassified names, based on relations with existing entities.

## 5. LINK GRAMMAR PARSER GATE INTERFACE

The Link Grammar Parser [6] is an open source tool that parses English text and extracts syntactic dependencies by labeling the relationship between pairs of words. The interface to the Link Grammar parser is found in a program called findlink2. To improve the output capability of the program we added "panic mode" in our modified findlink2 program into Java. The "panic mode" is used to parse long sentences when a valid parse is not found within a certain time. GATE's bootstrap wizard was used to create a plugin of the Link Grammar Interface, called FindLink. The FindLink plugin takes as input, a tokenized file and produces a file with the link structure for each sentence. The input file is tokenized using GATE's tokenizer in order to bind the syntactic bigrams to each token. FindLink extracts the left and right syntactic bigrams for each token and adds them as features to GATE's token annotation set, to be used later by a support vector machine. A syntactic bigram is "the right-hand links originating from the target; the words linked to the target through single right-hand links (call this set R1); the right-hand links originating from the words in R1; the words connected to the target through two right-hand links; the left-hand links originating from the target; the words linked to the target though single left-hand links (call this set L1); the left-hand links originating from the words in L1; and the words linked to the target through two left-hand links."

## 6. GATES PREPROCESSING RESOURCES

The first step in preprocessing the documents is tokenization. GATE's tokenizer is more complicated because not only does it split the document into tokens, it also extracts a multitude of information for each token. It places the tokens in different categories such as punctuation or word and it stores the length of the token and orthographical information such a capitalization. Having all this information already stored in each token eliminated the need to create a separate regular expression program to extract these features. All this information was later used to build the SVM model.

Many tools work on one sentence at a time therefore it is important to split the input into individual sentences. It does not suffice to simply split the text after a period or line break as GATE's ANNIE Sentence Splitter does. For example, sentences that use a period in abbreviations, personal titles or numbers can be incorrectly split. GATE's RegEx Sentence Splitter handles these cases correctly. RegEx Sentence Splitter's default configuration file splits sentence after two or more new line characters and was modified to split after one or more new line characters. One of the tools that require the sentence splitter is the part-of-speech (POS) tagger. Like the other tools, the tagger stores its result as a feature of each token.

GATE's Gazetteer tool is a kind of dictionary lookup. We added list of genes, proteins, and diseases to GATE's extensive list of dictionaries. When an entry in the input document is found in one of the dictionaries, an annotation type of "Lookup" is added to the document. Each Lookup annotation also holds a "majorType" feature, which equals the entry type such as name or location.

Another step required to preprocess a document is to mark the noun phrases. GATE already has a Noun Phrase Chunker plugin, but it also had problems with dates and numerals. We used Java Annotation Patterns Engine (JAPE) rules to make the plug-in handle such situations correctly. JAPE is a language developed to recognize patterns within the annotations of a document and produce new annotations out of the patterns. For example, below is an example of noun phrase chunk, where the words enclosed in brackets are noun phrases.

[The hypothetical protein C7orf24] has been implicated as [a cancer marker] with [a potential role] in [cell proliferation].

## 7. SUPPORT VECTOR MACHINE

We used Support Vector Machines [4] to perform semantic category recognition and relationship extraction. The task of the semantic category recognition SVM is to find and label the evaluative sentence in a document. The steps required to create an SVM in GATE are described below:

1. Manually annotate a document with the classes you want the SVM to learn. This file will be used to train the SVM. This can be achieved from end also by loading the document onto GATE, highlighting a text to be annotated and entering the annotation details in the popup box.
2. Pick the features that are needed to add to the SVM and create a configuration file based on these features..
3. Run the necessary tools on the manually annotated document to extract the features from step 2.
4. Train the SVM with the resulting file from step 3.

To apply the learned SVM model to a new document we simply run the same tools from step 3 above on the new document and run the SVM on the resulting document, with the learning mode set to application. The SVM uses features of each word to try and figure out how to categorize the word.

The features used for this case are discussed in section 2. Inside BIET, the assertion classifier is a rule based (i.e. JAPE rules are created) component and semantic relationship classifier (SC) component is another SVM model which is created by following the similar steps described above. One of the big advantages of GATE's SVM plugin is that, if one is not satisfied with the results of the SVM, one can manually delete the incorrect annotations and insert new annotations. Then the corrected output can be saved and added to the training corpus.

## 8. RESULTS

We have collected 100 abstract texts related to oncology research domain from the PubMed/MEDLINE. Two domain experts manually annotated these 100 abstracts in three aspects: selecting the evaluative sentences that refer to information related to gene, protein, disease etc.; further mark each evaluative sentence whether its assertion type is either positive, uncertain or negative type and then identifying the biomarker relationship (i.e. explicit and implicit types) from the evaluative sentences. Only those

annotations are considered for which both of the experts reached to a consensus regarding those three aspects of annotation. This annotated data serves as the gold standard of our experiment and is depicted in Table 1.

The BIET tool is consisting of three components namely, semantic category recognizer (SR) to recognize evaluative sentences, assertion classifier (AC) to group three types of assertions: positive, uncertain, and negative statement of gene/protein with respect to disease within the evaluative texts and finally the semantic relationship classifier (SC) does classify explicit and implicit biomarker relationships among the entities. Therefore we have tested the performance of the three components and evaluated precision, recall and F-measure with respect to the gold-standard scores.

The statistical SR component achieves a micro F-measure of 93.6%, and a macro F-measure of 92.1% for evaluative sentence recognition. Rule-based AC achieves average F-measure of 96.8% considering the true positive outputs (compared to the gold standard) made by SR component and the statistical SC component achieves a micro F-measure of 87.4% and a macro F-measure of 86.7% for both explicit and implicit biomarker relationship extraction.

**Table 1: The training and test corpus (Gold Standard)**

| | |
|---|---|
| Total Number of abstracts related to Oncology | 100 |
| Average Length of an Abstract | 8 sentences |
| Total Number of Sentences in the Corpus | 807 |
| Number of Evaluative Sentences | 233 |
| Number of Sentences with positive assertion | 112 |
| Number of Sentences with uncertain assertion | 93 |
| Number of Sentences with negative assertion | 28 |
| Explicit Biomarker relationships | 119 |
| Implicit Biomarker relationships | 97 |

Regarding experiments we followed two strategies, 8-fold cross validation and overall validation strategy. In 8-fold cross validation strategy, we took 80 abstracts (i.e. 645 sentences) to be used in eight fold cross-validation. These 80 abstracts are further sub-divided into eight sub-corpora having 10 abstracts in each fold. For each of the eight testing folds, the corresponding seven folds of gold standard data were used to train both SR and SC components and the leftover eighth one is used to test the performance of that fold. In this manner the 8-fold cross-validation allowed us to check the performance of the two SVM models in for each sub-corpora of the given dataset. On the contrary, the overall validation strategy utilized the 80 abstracts to train the SVMs and the rest 20 abstracts were used to test the components. The performance measures of the system components are summarized in Table 2. Since the output made by SR is used by the subsequent components, the performance of SR affects the performace of AC and SC compoents. Therefore we performed exhaustive experiments with SR components considering different featuresets and the best output achieved with the best set of features are described here and others are not mentioned due to space limitation. In the same manner the experiment result of SC is given for the best set of features and it is calculated with respect to the gold-standard score. Figure 2 shows the F-measure of the different experiments.

Chi-Square test revealed that the performance obtained for SR and SC from the different folds doesn't vary significantly, which is probably an indication that the training features are optimal.

**Table 2: Performance evaluation**

| Experiment | Semantic Category Recognizer SVM | | Assertion Classifier (Rule-Based) | | Semantic Relationship Classifier SVM | |
|---|---|---|---|---|---|---|
| Fold 1 | P | 96.9 | P | 96.6 | P | 87.4 |
| | R | 95.3 | R | 97.2 | R | 87.3 |
| | F | 96.1 | F | 96.9 | F | 87.3 |
| Fold 2 | P | 97.6 | P | 96.6 | P | 88.2 |
| | R | 95.4 | R | 97.2 | R | 86.9 |
| | F | 96.5 | F | 96.9 | F | 87.5 |
| Fold 3 | P | 96.8 | P | 96.6 | P | 87.7 |
| | R | 95.9 | R | 97.2 | R | 87.0 |
| | F | 96.4 | F | 96.9 | F | 87.4 |
| Fold 4 | P | 95.8 | P | 96.6 | P | 87.9 |
| | R | 96.9 | R | 97.2 | R | 87.0 |
| | F | 96.4 | F | 96.9 | F | 87.5 |
| Fold 5 | P | 95.4 | P | 96.6 | P | 88.0 |
| | R | 97.1 | R | 97.2 | R | 87.1 |
| | F | 96.2 | F | 96.9 | F | 87.6 |
| Fold 6 | P | 97.1 | P | 96.6 | P | 87.7 |
| | R | 95.8 | R | 97.2 | R | 86.8 |
| | F | 96.4 | F | 96.9 | F | 87.3 |
| Fold 7 | P | 97.1 | P | 96.6 | P | 87.7 |
| | R | 95.3 | R | 97.2 | R | 86.9 |
| | F | 96.2 | F | 96.9 | F | 87.3 |
| Fold 8 | P | 97.2 | P | 96.6 | P | 87.7 |
| | R | 95.4 | R | 97.2 | R | 87.0 |
| | F | 96.3 | F | 96.9 | F | 87.3 |
| 8-Fold Average | P | 96.7 | P | 96.6 | P | 87.8 |
| | R | 95.9 | R | 97.2 | R | 87.0 |
| | F | 96.3 | F | 96.9 | F | 87.4 |
| Overall Strategy | P | 92.0 | P | 96.8 | P | 86.9 |
| | R | 92.2 | R | 96.8 | R | 86.5 |
| | F | 92.1 | F | 96.8 | F | 86.7 |

The performance for AC remains unchanged in the 8-folds because this is a rule-based component and for each fold the same set of data is given to this component.The performance of AC component is measured by counting the true positive outputs made by SR components. We notice that if SR varies the performance of SC also varies. Therefore the better SR performs the better SC is likely to perform.

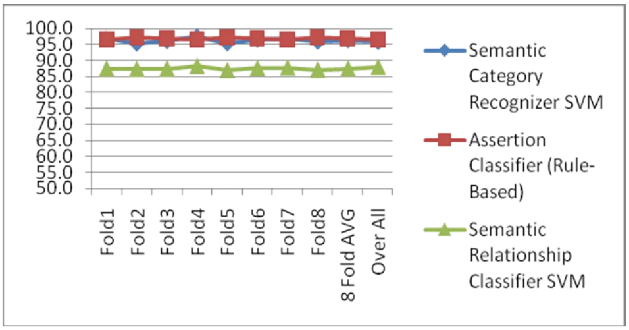


**Figure 2: Performance evaluation (F-Scores)**

## 9. CONCLUSION

In this paper we explored the machine learning approach along with semantic category recognition, assertion classification, and systematic relationship classification to extract biomarker information from PubMed abstracts. This method is designed to extract biomarker and disease based relationship from text using supervised machine learning techniques. Our result based on aggregation and classification rules indicates an average F- score 92.1 for category recognizer, 96.8 for assertion classifier and 86.7 for relationship classifier module.

The relationship extraction methods and techniques proposed in this paper are applicable to a broad range of applications and not limited to the biomarker identifications. Biological and genomic text mining is currently an active research area, in particular biomarker and disease analysis. Our results demonstrated that generic, mainstream machine learning software can produce substantial curation effort savings, when expert knowledge is channelled into the analysis task. Further, we like extend our relationship model with ontology and other clinically important information to BIET and develop it as a web-based application.

## 10. REFERENCES

[1] Blaschke C, Andrade MA, Ouzounis C, Valencia A. Automatic extraction of biological information from scientific text: protein-protein interactions. Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology. 1999; 60-7.

[2] Bunescu R, Ge R, Kate RJ, Marcotte EM, Mooney RJ, Ramani AK, Wong YW: Comparative experiments on learning information extractors for proteins and their interactions. Artificial Intelligence in Medicine 2005, 33(2):139-155.

[3] Bunescu RC, Mooney RJ: Subsequence Kernels for Relation Extraction. Proceedings of the 19th Conference on Neural Information Processing Systems 2005.

[4] Chih-Chung Chang and Chih-Jen Lin, LIBSVM: a library for support vector machines, 2001.

[5] Cunningham H, Maynard D, Bontcheva K, Tablan V: GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA 2002:168-175.

[6] D. Sleator and D. Temperley. Parsing English with a link grammar. Technical Report CMU-CS-91-196, Carnegie Mellon University, 1991.

[7] Jenssen TK, Læ´ greid A, Komorowski J, Hovig E.A literature network of human genes for highthroughput analysis of gene expression. Nature Genetics 2001; 28: 21-8.

[8] K. Bontcheva, H. Cunningham, V. Tablan, D.Maynard, O.Hamza. Using GATE as an Environment for Teaching NLP. Proceedings of the ACL Workshop on Effective Tools and Methodologies in Teaching NLP, 2002.

[9] Ono T, Hishigaki H, Tanigami A, Takagi T. Automated extraction of information on protein-protein interactions from the biological literature.Bioinformatics2001;17:155-61.

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

[11] Rindflesch TC, Tanabe L, Weinstein JN, Hunter L: EDGAR: Extraction of Drugs, Genes And Relations from the Biomedical Literature. Proceedings of Pacific Symposium on Biocomputing 2000:517-528.

[12] Rindflesch TC, Libbus B, Hristovski D, Aronson AR, Kilicoglu H: Semantic relations asserting the etiology of genetic diseases. AMIA Annu Symp Proc, National Library of Medicine, National Institutes of Health, Department of Health and Human Services, Bethesda, Maryland20894, USA 2003:554-558.

[13] Rosario B, Hearst A: Multi-way Relation Classification: Application to Protein-Protein Interaction. Human Language Technology Conference on Empirical Methods in Natural Language Processing 2005.

[14] Rzhetsky A, Iossifov I, Koike T, Krauthammer M, Kra P, Morris M, Yu H, Duboué PA, Weng W, Wilbur WJ, Hatzivassiloglou V, Friedman C:GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. J Biomed Inform 2004, 37:43-53.

[15] S Hunter L, Cohen KB. Biomedical language processing: what's beyond PubMed? Mol Cell 2006; 21:589–94.

[16] Sibanda, Tawanda. Was the Patient Cured? Understanding Semantic Categories and Their Relationships in Patient Records. Massachusetts Institute of Technology, June 2006.

[17] Yanhui Hu, Lisa M. Hines, Haifeng Weng, Dongmei Zuo, Miguel Rivera, Andrea Richardson, and Joshua LaBaer: Analysis of Genomic and Proteomic Data Using Advanced Literature Mining. Journal of Proteome Research 2003, 2, 405-412

[18] Zelenko D, Aone C, Richardella A: Kernel Methods for Relation Extraction. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, PA, USA 2002:71-78.

Pages 32-35 of this thesis have been removed as they contain published material. Please refer to the following citation for details of the article contained in these pages.

Islam, M. T., Shaikh, M., Nayak, A., & Ranganathan, S. (2010, June). Extracting biomarker information applying natural language processing and machine learning. In *2010 4th International Conference on Bioinformatics and Biomedical Engineering* (pp. 1-4). IEEE.

# Chapter 5: General discussion

## 5.1. Information extraction and summarization tool

The main objective of this development is to demonstrate that effective aggregations in domain specific scientific literatures might augment knowledge discovery. mExtract is built on multi-techniques that combines three major IE techniques; rule based, statistical and conceptual mapping that distinguishes it from other IE tools. Our system has proven to be effective in retrieving and extracting information. The proposed approach has the advantages of not being limited to a certain medical domain and making the use of some already existing components. The semantic representations can be used for different purposes. During the extraction process, information is collected from the written text itself and from its semantic representation. Given a user query our system first looks for information that are available in the conceptual data sources to automatically enhance and expand the user query and target words, then based on the extraction rules looks for target words in the input documents.

This system was enhanced recently to lookup concepts from UMLS and MeSH terms. The evaluation of the extraction process has been performed with a data set of 200 articles. For these articles the recall and precession values for retrieving the correct document are 92.15% and 93.24% respectively. For the evaluation of extraction techniques a data set of 50 articles are used and the recall and precession values are 91.29% and 92.65% respectively. Our preliminary experimental results demonstrate that information extraction and knowledge mining can be integrated to ameliorate both of the tasks.

## 5.2 Biomarker discovery Tool

In Chapter 4, we have described a novel conceptual biological knowledge-mining framework for describing multi-stage bioinformatics pipelines, and introduced a notation that simply but effectively captures biomedical entity relationship through the analysis process based on the features. We have done extensive performance testing using additive strategy measure to validate the performance of each feature. We started with a baseline feature and added one feature at a time to get

the most effective features. Features that had minimum or no positive impact on the performance were excluded. We started with an initial set of features and conducted the experiment, and only the best features are selected for each module. The key results are presented in chapter 4, and selected features for the two main modules are summarised in Table 5.1 of this subsection.

**Table 5.1 Feature Set for Biomarker Relationship extraction**

| Semantic Category Recognizer | Semantic Relationship Classifier |
|---|---|
| The targets: disease (i.e., name of the cancer), gene and protein. | The number of words between the candidate entities. |
| Left and right lexical bigrams of the targets | Whether or not the disease precedes the gene or protein entity. |
| The heading of the section that the targets appear in. | Whether or not other entities (e.g. medical tests) occur between the disease and gene/protein. |
| Syntactic bigrams of the targets. | The verbs between the disease and the gene/protein entity. |
| The head of the noun phrase that the targets is part of and the syntactic bigrams of the head | The two verbs before and after the disease and the two verbs before and after the gene/protein. |
| The part of speech of the targets and the words within a +/- 2 context window. | The headwords of the disease and the gene/protein related noun-phrases. |
| The UMLS semantic types of the noun phrase containing the target. | The right and left lexical bigrams of the disease and the gene/protein |
| Whether or not the target is a diseases name. | The right and left syntactic bigrams of the disease and the gene/protein. |
| Whether or not the target is a gene name. | The words between the disease and the gene or protein. |
| Whether or not the target is a protein name. | The path of syntactic links (as found by the Link Grammar Parser) between the disease and the gene or protein. |
| | The path of syntactically connected words between the disease and the gene/protein. |

Performance of relationship depends on the performance of entity recognition. The limitations of entity recognition will impact the performance of relation extraction. To get a measure of this effect, we evaluated the system and applied test data containing imperfectly extracted entities. These results are shown in Table 5.2. Our experiment with imperfect entity shows significantly lower overall F-score of 72.6 for semantic category recognizer and overall F-score of 64.6 for semantic relationship classifier. After a careful consideration, to improve the quality of the Named Entity Recognizer (NER), we have created an extensive list of all known gene, protein and disease names including their synonyms to the system and matched the record of these entities in UMLS database for further information. We then trained the system over texts containing gold standard entities. We have not trained our model with data containing imperfect entities; this can be taken as a future work to see if it makes any difference to the overall performance.

For the relationship extraction tasks the experiments described in Chapter 4 assumes perfect entity recognition. We used the entities of the gold standard as input to the relation extraction process for both training and testing purposes. During the feature set development phase, this measure was taken to separate the complexities of imperfect entity recognition in relation extraction process. Our system achieves an average F-score 92.1 for semantic category recognizer, 96.8 for assertion classifier and 86.7 for relationship classifier module.

In this framework, knowledge is not assumed to emerge from data alone, but as a result of combining data with other knowledge (such as descriptive metadata, or analysis results from analysis tasks). Since knowledge flow and task composition are key aspects of multi-stage analysis, our framework is a contribution towards the design and formalization of large-scale bioinformatics projects.

25. Glenisson, P., et al., *Evaluation of the vector space representation in text-based gene clustering.* Pac Symp Biocomput, 2003: p. 391-402.
26. Gaizauskas, R., et al., *Protein structures and information extraction from biological texts: the PASTA system.* Bioinformatics, 2003. **19**(1): p. 135-43.
27. Albert, S., et al., *Computer-assisted generation of a protein-interaction database for nuclear receptors.* Mol Endocrinol, 2003. **17**(8): p. 1555-67.
28. Chun HW, Tsuruoka Y, Kim JD, Shiba R, Nagata N, Hishiki T, Tsujii J. Maui, *Extraction of Gene-Disease Relations from Medline Using Domain Dictionaries and Machine Learning*. in *Pacific Symposium on Biocomputing*, Hawaii, USA.
29. Roberts, A., et al., *Mining clinical relationships from patient narratives.* BMC Bioinformatics, 2008. **9 Suppl 11**: p. S3.
30. Rindflesch, T.C., et al., *Semantic relations asserting the etiology of genetic diseases.* AMIA Annu Symp Proc, 2003: p. 554-8.
31. Rindflesch, T.C., et al., *EDGAR: extraction of drugs, genes and relations from the biomedical literature.* Pac Symp Biocomput, 2000: p. 517-28.
32. *Cancer in Australia*. [cited 25/03/2010]; Available from: http://www.canceraustralia.gov.au/about-cancer/cancer-information/cancer-in-australia.aspx.
33. *Cancer incidence projections Australia 2002 to 2011*. [cited 15/03/10]; Available from: http://www.aihw.gov.au/publications/can/cipa02-11/.
34. Hoffmann, R. and A. Valencia, *A gene network for navigating the literature.* Nat Genet, 2004. **36**(7): p. 664.
35. Dimitar Hristovskia and B. Peterlinb. *Literature-Based Disease Candidate Gene Discovery*. in *Proceedings of Medinfo*. 2004: American Medical Informatics Association.
36. *Medical Subject Headings*. 14/10/2010 [cited 02/03/2010]; Available from: http://www.nlm.nih.gov/mesh/.
37. Tamames, J., et al., *EUCLID: automatic classification of proteins in functional classes by their database annotations.* Bioinformatics, 1998. **14**(6): p. 542-3.
38. Sibanda, T., *Was the patient cured? Understanding semantic categories and their relationships in patient records*. 2006, Massachusetts Institute of Technology.

## 5.3 Future Work

The work presented in this thesis has a great potential for future research development. This thesis has identified key objectives and directions of relationship extraction and knowledge mining task, but the technologies and methods proposed here needs to be considered for further research for greater contribution to the biomedical community. In the near future, we like to extend our relationship model with ontology and other clinically significant concepts (not limited to disease→gene/protein) and develop web interface to make it available in the web.

In future we like to do further analysis on extracted relations to help scientists to develop different hypothesis to associate other possible biomarkers, having an established relationship with other diseases, as a potential biomarker of a given diseases for which the inferred biomarker relationship has not yet discovered. In other words we like to uncover previously unrecognised relationships worthy of further investigation. We want to find all the relations Y related to the X (i.e. starting point). For example if X is a disease then Y can be the causes, symptoms etc of the disease X. Similarly from other set of literatures we may obtain Z as the causes, symptoms etc of the disease X. Thus we get all the possible biomarkers of X. If there is no previously identified relation between X and Z, we may infer a new relation between X and Z. This new relation can then generate new research hypothesis and scientist can verify their suitability by further literature review, lab experiments and analysis.

## 5.4. Concluding comments

Biomedical knowledge mining and biomarker discovery is an emerging and active research domain as scientists, pharmacists and physicians continue to fight with diseases to improve the stability of human life. Our research has developed methods and analytical tools for systematic analysis of biomedical data for comparative research analysis. We have applied these approaches to the oncology research data to demonstrate and evaluate our system based on one relationship (i.e., bio-marker relationship). We have also provided some future directions to our system. However, in order to understand the interactions and

relations between biomedical entities, this work will need to be extended to several other entities and relationships. Future directions will focus on automation of updating different relationship, in view of the constant updating of relationship network databases, and more complex analysis of extracted information to build a decision support system.

## References


1. Bernstein, B. and M. Kellis, *Large-scale discovery and validation of functional elements in the human genome.* Genome Biology, 2005. **6**(3): p. 312.
2. Chagoyen, M., et al., *Discovering semantic features in the literature: a foundation for building functional associations.* BMC Bioinformatics, 2006. **7**: p. 41.
3. Hall, N., *Advanced sequencing technologies and their wider impact in microbiology.* J Exp Biol, 2007. **210**(9): p. 1518-1525.
4. Brazma, A., et al., *Approaches to the automatic discovery of patterns in biosequences.* J Comput Biol, 1998. **5**(2): p. 279-305.
5. Yu, U., et al., *Bioinformatics in the post-genome era.* J Biochem Mol Biol, 2004. **37**(1): p. 75-82.
6. Kanehisa, M. and P. Bork, *Bioinformatics in the post-sequence era.* Nat Genet, 2003. **33 Suppl**: p. 305-10.
7. *NLM System: Data, News and Update Information*. [cited 19/03/2010]; Available from: http://www.nlm.nih.gov/bsd/revup/revup_pub.html#med_update.
8. Druss, B.G. and S.C. Marcus, *Growth and decentralization of the medical literature: implications for evidence-based medicine.* J Med Libr Assoc, 2005. **93**(4): p. 499-501.
9. Tanabe, L. and W.J. Wilbur, *Tagging gene and protein names in biomedical text.* Bioinformatics, 2002. **18**(8): p. 1124-32.
10. Roman Klinger, C.M.F., Juliane Fluck, Martin Hofmann-Apitius. *Named Entity Recognition with Combinations of Conditional Random Fields*. in *Second BioCreative Challenge Evaluation Workshop*. 2007. Centro Nacional de Investigaciones Oncologicas, CNIO, Madrid, Spain.
11. Chang, J.T., H. Schutze, and R.B. Altman, *GAPSCORE: finding gene and protein names one word at a time.* Bioinformatics, 2004. **20**(2): p. 216-25.
12. Moshe, F., R. Binyamin, and F. Ronen, *A hybrid approach to NER by MEMM and manual rules*, in *Proceedings of the 14th ACM international conference on Information and knowledge management*. 2005, ACM: Bremen, Germany.

13. Donaldson, I., et al., *PreBIND and Textomy--mining the biomedical literature for protein-protein interactions using a support vector machine.* BMC Bioinformatics, 2003. **4**: p. 11.
14. Wallace, B.C., et al., *Semi-automated screening of biomedical citations for systematic reviews.* BMC Bioinformatics. **11**: p. 55.
15. Liu, F., et al., *FigSearch: a figure legend indexing and classification system.* Bioinformatics, 2004. **20**(16): p. 2880-2.
16. Waree, R. and O. Carlos, *Microarray data analysis with PCA in a DBMS*, in *Proceeding of the 2nd international workshop on Data and text mining in bioinformatics*. 2008, ACM: Napa Valley, California, USA.
17. Pankaj, K. and M. Sougata, *Text-based summarization and visualization of gene clusters*, in *Proceedings of the 2005 ACM symposium on Applied computing*. 2005, ACM: Santa Fe, New Mexico.
18. Seki, K. and J. Mostafa, *Gene ontology annotation as text categorization: An empirical study.* Information Processing & Management, 2008. **44**(5): p. 1754-1770.
19. *GeneWebEx: Gene Annotation Web Extraction, Aggregation, and Updating from Web-Based Biomolecular Databanks*, in *Proceedings of the 4th IEEE Symposium on Bioinformatics and Bioengineering*. 2004, IEEE Computer Society.
20. Zhang-Zhi, H., et al., *iProLINK: A Framework for Linking Text Mining with Ontology and Systems Biology*, in *Proceedings of the 2008 IEEE International Conference on Bioinformatics and Biomedicine*. 2008, IEEE Computer Society.
21. Yu, H. and E. Agichtein, *Extracting synonymous gene and protein terms from biological literature.* Bioinformatics, 2003. **19 Suppl 1**: p. i340-9.
22. Liu, H. and C. Friedman, *Mining terminological knowledge in large biomedical corpora.* Pac Symp Biocomput, 2003: p. 415-26.
23. Ono, T., et al., *Automated extraction of information on protein-protein interactions from the biological literature.* Bioinformatics, 2001. **17**(2): p. 155-61.
24. Ramani, A.K., et al., *Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome.* Genome Biol, 2005. **6**(5): p. R40.

25. Glenisson, P., et al., *Evaluation of the vector space representation in text-based gene clustering.* Pac Symp Biocomput, 2003: p. 391-402.
26. Gaizauskas, R., et al., *Protein structures and information extraction from biological texts: the PASTA system.* Bioinformatics, 2003. **19**(1): p. 135-43.
27. Albert, S., et al., *Computer-assisted generation of a protein-interaction database for nuclear receptors.* Mol Endocrinol, 2003. **17**(8): p. 1555-67.
28. Chun HW, Tsuruoka Y, Kim JD, Shiba R, Nagata N, Hishiki T, Tsujii J. Maui, *Extraction of Gene-Disease Relations from Medline Using Domain Dictionaries and Machine Learning*. in *Pacific Symposium on Biocomputing*, Hawaii, USA.
29. Roberts, A., et al., *Mining clinical relationships from patient narratives.* BMC Bioinformatics, 2008. **9 Suppl 11**: p. S3.
30. Rindflesch, T.C., et al., *Semantic relations asserting the etiology of genetic diseases.* AMIA Annu Symp Proc, 2003: p. 554-8.
31. Rindflesch, T.C., et al., *EDGAR: extraction of drugs, genes and relations from the biomedical literature.* Pac Symp Biocomput, 2000: p. 517-28.
32. *Cancer in Australia*. [cited 25/03/2010]; Available from: http://www.canceraustralia.gov.au/about-cancer/cancer-information/cancer-in-australia.aspx.
33. *Cancer incidence projections Australia 2002 to 2011*. [cited 15/03/10]; Available from: http://www.aihw.gov.au/publications/can/cipa02-11/.
34. Hoffmann, R. and A. Valencia, *A gene network for navigating the literature.* Nat Genet, 2004. **36**(7): p. 664.
35. Dimitar Hristovskia and B. Peterlinb. *Literature-Based Disease Candidate Gene Discovery*. in *Proceedings of Medinfo*. 2004: American Medical Informatics Association.
36. *Medical Subject Headings*. 14/10/2010 [cited 02/03/2010]; Available from: http://www.nlm.nih.gov/mesh/.
37. Tamames, J., et al., *EUCLID: automatic classification of proteins in functional classes by their database annotations.* Bioinformatics, 1998. **14**(6): p. 542-3.
38. Sibanda, T., *Was the patient cured? Understanding semantic categories and their relationships in patient records*. 2006, Massachusetts Institute of Technology.

39. Chen, H. and B.M. Sharp, *Content-rich biological network constructed by mining PubMed abstracts.* BMC Bioinformatics, 2004. **5**: p. 147.
40. Maksym, K., et al., *A tool for gene expression based PubMed search through combining data sources.* Bioinformatics, 2004. **20**(12): p. 1980-1982.
41. Jenssen, T.K., et al., *A literature network of human genes for high-throughput analysis of gene expression.* Nat Genet, 2001. **28**(1): p. 21-8.
42. Hu, Y., et al., *Analysis of Genomic and Proteomic Data Using Advanced Literature Mining.* Journal of Proteome Research, 2003. **2**(4): p. 405-412.
43. Oliveros, J.C., et al., *Expression profiles and biological function.* Genome Inform Ser Workshop Genome Inform, 2000. **11**: p. 106-17.
44. Gerard, S. and B. Chris, *Term Weighting Approaches in Automatic Text Retrieval*. 1987, Cornell University.
45. N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*. 2000: Cambridge University Press.
46. Vapnik, V., *Statistical Learning Theory.* . 1998, New York: John Wiley & Sons.
47. Osuna, E.F., Robert; Girosi, Federico, *Support Vector Machines: Training and Applications*. AI Memo 1602. 1997.
48. Christopher, J.C.B., *A Tutorial on Support Vector Machines for Pattern Recognition.* Data Min. Knowl. Discov., 1998. **2**(2): p. 121-167.
49. Thorsten, J., *Making large-scale support vector machine learning practical*, in *Advances in kernel methods: support vector learning*. 1999, MIT Press. p. 169-184.

# Appendix A:

### Bioinformatics resources for Text Mining

| Resource Name | Type | URL |
|---|---|---|
| Disease Database | Database with names of diseases | http://www.diseasesdatabase.com |
| Entrez Gene | Searchable Gene Database | http://www.ncbi.nlm.nih.gov/gene |
| GATE | General Architecture for Text Engineering, computer architecture for a broad range of Natural Language Processing tasks. | http://gate.ac.uk/ |
| GeneCards | Searchable, integrated, database of human genes | http://www.genecards.org |
| GO | Gene Ontology Database | http://www.geneontology.org |
| GPSDB | Gene and Protein Synonym Database,organized by species | http://biomint.cs.kuleuven.be/protop/bin/bmsynonyms.pl?s=&userType=guest |
| MeSH | Medical Subject Headings Database | http://www.nlm.nih.gov/mesh/meshhome.html |
| PubMed | Database of citations and abstracts for biomedical literature from MEDLINE and additional life science journals | http://www.ncbi.nlm.nih.gov/pubmed |
| PubMed Central | Digital archive of full-text biomedical and life sciences journal literature | http://www.ncbi.nlm.nih.gov/pmc/ |
| UMLS | Unified Medical Language Metathesaurus | http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/index.html |

## Appendix B:

### Peer-reviewed Conference papers

1. M.T. Islam, D. Bollina, A. Nayak, S. Ranganathan (2007): Towards an Agent-based Information Retrieval System for Computational Biomarker Discovery, in the International Conference on Information and Communication Technology (ICICT 2007), March 7-9, 2007, Dhaka, Bangladesh, pp. 57-63.
2. M.T.Islam, M. Shaikh, A. Nayak, S. Ranganathan, Biomarker Information Extraction Tool (BIET) Development using Natural Language Processing and Machine Learning, Proc.2010 IEEE/ACM Int'l Conference and Workshop on Emerging Trends in Technology(ICWET 2010), February 2010, Mumbai, India, pp. 121-126.
3. M.T.Islam, M. Shaikh, A. Nayak, S. Ranganathan, Extracting Biomarker Information applying Natural Language Processing and Machine Learning, Proc.2010 IEEE 4th Int'l Conference on Bioinformatics and Biomedical Engineering (iCBBE 2010), Chengdu, China. (To be published and presented, June 2010).

### Oral presentations

4. M.T. Islam, D. Bollina, A. Nayak, S. Ranganathan (2007): Towards an Agent-based Information Retrieval System for Computational Biomarker Discovery, in the International Conference on Information and Communication Technology (ICICT 2007), March 7-9, 2007, Dhaka, Bangladesh, pp. 57-63.
5. M.T.Islam, M. Shaikh, A. Nayak, S. Ranganathan, Biomarker Information Extraction Tool (BIET) Development using Natural Language Processing and Machine Learning, Proc.2010 IEEE/ACM Int'l Conference and Workshop on Emerging Trends in Technology (ICWET 2010), February 2010, Mumbai, India, pp. 121-126.
6. M.T. Islam, M. Shaikh, A. Nayak, S. Ranganathan, Extracting Biomarker Information applying Natural Language Processing and Machine Learning, Proc.2010 IEEE 4th Int'l Conference on Bioinformatics and Biomedical Engineering (iCBBE 2010), Chengdu, China. (To be published and presented, June 2010).