University of Crete
School of Sciences and Engineering
Computer Science Department

Mining the Biomedical Literature – The MineBioText system: Discovery of Gene, Protein and Disease Correlations

Despoina Antonakaki

Master Science Thesis

Heraklion, Crete, Greece
February 2006

University of Crete
School of Sciences and Engineering
Computer Science Department

Mining the Biomedical Literature – The MineBioText system: Discovery of Gene, Protein and Disease Correlations

by

Despoina Antonakaki

MASTER OF SCIENCE

Author:

Despoina Antonakaki
Computer Science Department

Supervisory Committee:

Vassilis Christophides
Supervisor
Associate Professor, Computer Science Department

Yannis Tollis
Professor, Computer Science Department

George Potamias
Researcher B, ICS-FORTH

Dimitris Kafetzopoulos
Researcher B, IMBB-FORTH

Researcher B, ICS-FORTH

Approved by:

Dimitris Plexousakis
Associate Professor, Computer Science Department
Chairman of the Graduate Studies Committee

Heraklion, Crete, Greece
February 2006

Mining the Biomedical Literature – The MineBioText system: Discovery of Gene, Protein and Disease Correlations

DESPOINA ANTONAKAKI

MASTER OF SCIENCE
THESIS

ABSTRACT

Automatic knowledge discovery from biomedical free-texts appears as a necessity considering the growing of the massive amounts of biomedical scientific literature. A special problem that makes this task more challenging, and difficult as well, is the over-abundance and diversity of the related genomic/proteomic ontologies and the respective gene and protein terminologies. Specifically, a genomic/proteomic term, e.g., gene, protein and their functional descriptions, as well as the diseases, are referred with many different ways in scientific documents regarding the organization, research context and the naming conventions that the authors are adherent to. The work reported in this thesis presents methods and tools for the efficient and reliable mining of biomedical literature, based on advanced text-mining techniques. Specifically it covers the following R&D challenges: (a) Identification of gene/protein--gene/protein and gene/protein--disease correlations following a text mining approach. The approach utilizes data-mining and statistical techniques, algorithms and metrics to deal with the following problems: (i) identification and recognition of terms in text-references – based on an appropriately devised and implemented algorithmic process that utilises the Trie data-structure; and (ii) ranking of terms and their (potential) relations or, links – based on the MIM entropic metric (Mutual Information Metric) to measure the respective terms' association strength. (b) Construction of a genes association network – based on the assessed terms' (genes, proteins, diseases) association strengths. (c) Categorization / Classification of text-references (mainly from the PubMed abstracts repository) into class categories utilizing an appropriately devised classification metric and procedure, and using the most descriptive (i.e, strong) associations between terms. Pre-assignment of text-references (i.e., PubMed abstract) to categories is performed by posting respective queries to PubMed, i.e., querying PubMed with "breast cancer" the retrieved documents are considered to belong to the "breast cancer" category. (d) Assessment on the texts' categorization / classification results – based on respective PubMed abstract collections, their pre-categorization and careful experimental set-up to measure prediction results, i.e., accuracy and precision. (e) Design and development of a tool – the ***MineBioText*** (Mining Biomedical Texts), that encompasses all of the aforementioned operations with extra functionalities for setting-up the domain of reference and study, e.g., gene/protein and disease names, their synonyms and free-text descriptions, text collections, parameterization of build-in algorithmic processes etc.

Supervisor: **Vassilis Christophides**
Associate Professor
Computer Science Department
University of Crete

ΕΞΟΡΥΞΗ ΓΝΩΣΕΩΝ ΑΠΟ ΒΙΟΙΑΤΡΙΚΗ ΒΙΒΛΙΟΓΡΑΦΙΑ – ΤΟ ΣΥΣΤΗΜΑ MINEBIOTEXT: ΑΝΑΚΑΛΥΨΗ ΣΥΣΧΕΤΙΣΕΩΝ ΜΕΤΑΞΥ ΓΟΝΙΔΙΩΝ, ΠΡΩΤΕΪΝΩΝ ΚΑΙ ΑΣΘΕΝΕΙΩΝ

ΔΕΣΠΟΙΝΑ ΑΝΤΩΝΑΚΑΚΗ

ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΠΕΡΙΛΗΨΗ

Η αυτόματη ανακάλυψη γνώσεων από έγγραφα βιοϊατρικού περιεχομένου ελεύθερης γραφής (free-texts) αποτελεί μια αναγκαιότητα κυρίως λόγω του τεράστιου, και συνεχώς αυξανόμενου, πλήθους σχετικών επιστημονικών αναφορών. Το βασικό πρόβλημα που κάνει αυτόν τον στόχο περισσότερο προκλητικό και δύσκολο είναι η υπεραφθονία καθώς και η ποικιλομορφία σχετικών γονιδιωματικών ορολογιών και των εμπλεκόμενων γονιδιακών/πρωτεϊνικών ορολογιών. Συγκεκριμένα, ένας γονιδιωματικός όρος, π.χ., γονίδιο ή πρωτεΐνη και η περιγραφή της λειτουργία, αλλά και σχετιζόμενες ασθένειες, αναφέρονται με πολλούς διαφορετικούς τρόπους σε σχετικά επιστημονικά έγγραφα ανάλογα με το ερευνητικό πλαίσιο και τις συμβάσεις ονοματολογίας που ο συντάκτης του εγγράφου αποδέχεται και ακολουθεί. Η εργασία που αναφέρεται σε αυτήν την μεταπτυχιακή διατριβή παρουσιάζει μεθόδους και τα εργαλεία για την αποδοτική και αξιόπιστη ανακάλυψη γνώσεων από τη σχετική βιοϊατρική βιβλιογραφία και αναφορές, και βασίζεται σε προηγμένες τεχνικές εξόρυξης γνώσης από κείμενα (text-mining). Συγκεκριμένα, συνδιαλέγεται και προσφέρει λύσεις στις παρακάτω ερευνητικές και αναπτυξιακές (Ε&Α) προκλήσεις: (α) Αυτόματη ανακάλυψη συσχετίσεων μεταξύ γονιδίων/πρωτεϊνών και μεταξύ γονιδίων/πρωτεϊνών και ασθενειών. Το θέμα προσεγγίζεται με τεχνικές και αλγοριθμικές διαδικασίες text-mining καθώς και τη δημιουργία και χρήση σχετικών στατιστικών μετρικών: (i) Προσδιορισμός, αναγνώριση και διαχείριση όρων σε βιοϊατρικά έγγραφα – για το σκοπό αυτό επινοήθηκε και προσαρμόστηκε κατάλληλα μια αλγοριθμική διαδικασία που χρησιμοποιεί την ευέλικτη και αποδοτική δομή δεδομένων Trie, και (ii) ταξινόμηση των όρων και (των πιθανών) σχέσεών τους ή, συνδέσεων – για το σκοπό αυτό η εντροπική μετρική υπολογισμού της αμοιβαίας πληροφορίας έχει κατάλληλα προσαρμοστεί και χρησιμοποιηθεί. (β) Κατασκευή δικτύου συσχέτισης γονιδίων/πρωτεϊνών (gene correlation network) – βασίζεται στην αξιολόγηση της δύναμης συσχέτισης (correlation strength) των προσδιορισμένων και αναγνωρισμένων γονιδιωματικών όρων στα διαθέσιμα έγγραφα. (γ) Κατηγοριοποίηση/Ταξινόμηση εγγράφων (κυρίως από την αποθήκη περιλήψεων PubMed) η οποία βασίζεται στην επινόηση και χρήση μιας μετρικής ταξινόμησης και την εισαγωγή σχετικής αλγοριθμικής διαδικασίας ταξινόμησης εγγράφων (texts classification) – η μετρική χρησιμοποιεί τη δύναμη συσχέτισης μεταξύ όρων που εμφανίζονται στα διαθέσιμα έγγραφα. Η αλγοριθμική διαδικασία στηρίζεται στην εκπαίδευση (training) του ταξινομητή εγγράφων με βάση έγγραφα-εκπαίδευσης από τη βάση/αποθήκη περιλήψεων PubMed και την εκ' των προτέρων ταξινομησή τους (pre-assignment to classes) από σχετικά ερωτήματα στο PubMed, δηλ., θέτοντας το ερώτημα στο PubMed "καρκίνος του μαστού" τα ανακτημένα έγγραφα θεωρούνται ότι ανήκουν στην κατηγορία "καρκίνος-μαστού". (δ) Εκτεταμένα πειράματα για την επικύρωση (validation) και αξιολόγηση (evaluation) αποτελεσμάτων σε σχέση με την αξιοπιστία και 'χρησιμότητα' των συσχετίσεων που ανακαλύπτονται, καθώς και σε σχέση με την αξιοπιστία (ακρίβεια) κατάταξης και ταξινόμησης εγγράφων. (ε) Σχεδίαση και ανάπτυξη ενός εργαλείου – το σύστημα ***MineBioText***, το οποίο ενσωματώνει όλες τις προαναφερθείσες τεχνικές και διαδικασίες με τις πρόσθετες λειτουργίες για τη δημιουργία του πεδίου-αναφοράς (domain of reference) σε ολοκληρωμένες διαδικασίες εξόρυξης γνώσης από βιβλιογραφικές αναφορές, π.χ., εκμετάλλευση πολλαπλών ονοματολογιών γονιδίων/πρωτεϊνών και ασθενειών, των συνωνυμών τους και των αντίστοιχων ελεύθερου-κειμένου περιγραφών τους, συλλογές εγγράφων, παραμετροποίηση διαδικασιών, οπτικοποίηση (visualization) αποτελεσμάτων κ.λπ.

Επόπτης: **Βασίλης Χριστοφίδης**
Αναπληρωτής Καθηγητής
Τμήμα Επιστήμης Υπολογιστών
Πανεπιστήμιο Κρήτης

Ευχαριστίες

Ξεκινώντας πριν δύο περίπου χρόνια την εργασία αυτή, δεν είχα συνειδητοποιήσει τις δυνατότητες αλλα και την γοητεία που θα μπορούσε να μου παρέχει η μελέτη της εφαρμογής τεχνικών Μηχανική Μάθησης στο πεδίο των βιοϊατρικών δεδομένων αλλα και τον τρόπο που διευρύνουν την έρευνα στην βιολογία.

Θα ήθελα να ευχαριστήσω όλους όσους βοηθήσαν στην εκπόνηση αυτής της εργασίας. Καταρχήν τον επιβλέποντα καθηγητή μου Γιώργο Ποταμιά όχι μόνο για την καθοδήγηση και για την ευκαιρία που μου προσέφερε να γνωρίσω ένα δυναμικό χώρο έρευνας και σκέψης αλλα και την ευχάριστη συνεργασία που είχαμε όλον αυτόν τον καιρό.

Νιώθω την ανάγκη να ευχαριστήσω τον αείμνηστο καθηγητή Στέλιο Ορφανουδάκη για τα βασικά θεωρητικά θεμέλια αυτής της εργασίας και φυσικά τον μετέπειτα επόπτη αυτής της εργασίας Βασίλη Χριστοφίδη. Σε αυτό το σημείο θα ήθελα επίσης να ευχαριστήσω τον Αλέξανδρο Καντεράκη για την ουσιαστική βοήθεια που μου παρείχε για την διεκπεραίωση αυτης της εργασίας.

Ιδιαίτερα ευχαριστώ τους Δημήτρη Καφετζόπουλο και Γιάννη Τόλλη για την συμμετοχή τους στην περάτωση αυτής της εργασίας καθώς επίσης το Ινστιτούτο Πληροφορικής και το Πανεπιστήμιο Κρήτης για την οικονομική και υλικοτεχνική υποστήριξη όλα αυτά τα χρόνια.

θα ήθελα επίσης να ευχαριστήσω μια ξεχωριστή παρέα -έτοιμη να αντιμετωπίσει τις προκλήσεις του χρόνου με ενθουσιασμό, ζωντάνια, ομορφία, δύναμη και ευφυία- για την συντροφιά, την ηθική υποστήριξη τους αλλά και το πλήθος των εμπειριών που μοιραστήκαμε .

Τέλος θα ήθελα να ευχαριστήσω τα αδέλφια μου Ζαχαρία, Κωνσταντίνο και Μάριο, και τους γονείς μου Δημήτριο και Σταυρούλα στους οποίους αφιερώνω αυτήν την εργασία.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. Introduction

1.1 Motivation

After completion of the human genome sequencing (http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml) we are now entering the *post-genomic* age. The main focus in genomics research is switching from sequencing to using the genome sequences in order to understand how genomes are functioning (i.e., *functional genomics*). Approximately a decade ago, the concept of being able to simultaneously measure the concentrations of every transcript in the cell in a single experiment seemed impossible to most researchers. A new demanding need is raising namely, the linkage between the clinical and the 'genomics' worlds. Identification of genes and proteins that affect biological function in humans and other organisms is a critical step towards the discovery of new medical therapies. In this context the study of *relations* (*correlations* and/or *interactions*) between genes/proteins, and between genes/proteins and diseases is vital, especially with the need to approach and realise the vision of *genomic and individualised medicine*. This is mainly because an individual's (potential) genetic predisposition is more likely to be *polygenic*, i.e., depending from multiple genes and their interrelations.

The target and main contribution of this thesis is the automated **discovery** of relationships among genes - over hundreds or thousands of them.

The quest relies on the fact that many individual genes and their function are already discussed in the literature. The main assumption and the rational behind this approach is that common (between genes or, proteins) function related relevant literature is a strong indicator of common function among genes. Subsequently, it is a strong indicator for a ***correlation*** (or, association) between the referred genes/proteins.

The vehicle towards the objective of the thesis relies on information retrieval and **text mining** approaches. In this setting novel document indexing, term (gene, protein, diseases) identification; and document categorisation methods and techniques are introduced.

In this context an integrated biomedical literature mining system was designed and implemented – the **MineBioText** system.

In a relatively recent report (Committee on Quality of Health Care in America, 2001) it is stated that: *"What is perhaps most disturbing is the absence of real progress toward restructuring health care systems to address both quality and cost concerns, or toward applying advances in information technology to improve administrative and clinical processes"*. Our position is that, difficulties and failures of medical decision-making in everyday practice are largely failures in *knowledge coupling*, due to the over-reliance on the unaided human mind to recall and organize all the relevant details. They are not, specifically and essentially, failures to reason logically with the medical knowledge once it is presented completely and in a highly organized form within the framework of the patient's total and unique situation (Weed, 1991). If we are to reduce errors and provide quality of care, we must transform the current healthcare enterprise to one in which caregivers exercise their unique human capacities within supportive systems that compensate for their inevitable human limitations. Achieving that vision, however, requires that we first build the appropriate technology and enable clinicians to integrate it into their practices by adopting system-oriented values.

Systems Biology – *a 'holistic' approach*. Correlations between genotypes, gene regulatory networks and biochemical pathways allow the intervention and metabolic readjustments for combating complex diseases (Evans and Relling, 2004; Bonetta, 2004). An ambitious direction is to attempt to model and infer *gene regulatory networks* on a global scale, or along more specific subcomponents such as a pathway or a set of co-regulated genes. A major obstacle is that our knowledge of transcription and other critical molecular level mechanisms remains incomplete, especially as refers to in-vivo perturbations or "noise" at various stages of regulation in molecular processes which could mark the difference between changes, often epigenetic, which may significantly affect other processes, versus those which do not. Furthermore, there are very few examples of regulatory circuits for which detailed information is available, and they all appear to be very complex. On the theoretical side, several mathematical formalisms have been applied to model genetic networks. These range from discrete models, such as Boolean networks, as in the pioneering work of Kauffman, to continuous models based on differential equations, such as continuous recurrent neural networks or power-law formalism, probabilistic graphical models and Bayesian networks. None of these formalisms appears to capture all the dimensions of gene regulation and most of the work in this field is still very preliminary. The manual inference of pathway information as it occurs e.g. in the interpretation of gene expression data (Apica *et al*. 2005) is assisted with the use of pre-complied protein interaction databases, like those available from ingenuity, Transfac, GeneGo (Nikolsky *et al*. 2005), Ariadne. A review of most of these tools can be found in (Bonetta 2004). Understanding biology at the system level - not only gene networks, but also protein networks, signalling networks, metabolic networks, and specific systems, such as the immune system or neuronal networks - is likely to remain at the center of the bioinformatics efforts of the next few decades.

Interdisciplinary Research. With the introduction of sophisticated laboratory instrumentation, robotics and large, complex data sets, biomedical research is increasingly becoming a *cross-disciplinary* effort requiring the collaboration of biologists, engineers, software and database designers, physicists etc. Techniques and technological infrastructure comes mainly from *Bioinformatics* (BI) and *Medical Informatics* (MI) – two disciplines that up to now have followed separate development with few contacts and synergies between them. The publication of the human genome has evidenced the need and the possibilities for a strong synergy between these two disciplines. The *integration* and exploitation of the data and information generated at all levels in both fields requires a new approach that enables a two-way dialogue between them that comprises data, methods, technologies, tools and applications.

- ***Biomedical Informatics*** (**BMI**) is the emerging area that aims to put these two worlds together. The mission of BMI is to provide the technical and scientific infrastructure and knowledge to allow evidence-based, individualised healthcare using all relevant sources of information. These sources include the "classical" information as currently maintained in the health record, as well as new genomic, proteomic and other molecular-level information. Aiming at a change from late stage diagnosis towards early detection or even prediction of disease, BMI bears the potential to foster discovery and creation of novel diagnostic and therapeutic methods, in order to improve the health and quality of life of the individual, as well as the efficiency of expenditure in healthcare systems (Martin-Sanchez *et al.*, 2004; Diaz, 2005).

1.2 Literature Data Mining

Almost every known or postulated piece of information pertaining to genes, proteins, and their role in biological processes is reported somewhere in the vast amount of published biomedical literature. However, the advancement of genome sequencing techniques is accompanied by an overwhelming increase in the literature discussing the discovered genes. This combined abundance of genes and literature produces a major bottleneck for interpreting and planning genome-wide experiments. Thus, the ability to rapidly survey this literature constitutes a necessary step toward both the design and the interpretation of any large-scale experiment. Moreover, automated literature mining offers a yet untapped opportunity to integrate many fragments of information gathered by researchers from multiple fields of expertise into a complete picture exposing the interrelated roles of various genes, proteins, and chemical reactions in cells and organisms (Shatkay and Feldman, 2003).

In a committee report for the US National Academy of Science Harold J.Morowitz argued that biological research had reached a point where "*new generalizations and higher order biological laws are being approached, but may be obscured by the simple mass of data*". Now, twenty years later, those words seem to us a cruel reality as well as an impulsion for the development of improved computer-aided tools to aid the human experts. Referring to the US National Library of Medicine (NLM – *PubMed*, www.pubmed.com), the citation database published from the mid 1960s more that 4700 biomedical journals published in over 70 countries. In 1985, the total amount of sequence entries found in EBI nucleotide database reached almost the size of 5000. Four years ago, in 2001 that number was increased about five times. The expansion of wider applications of intensive data technologies includes DNA and protein chips, high-throughput protein three-dimensional structure determination and real time molecular and cellular imaging. This swamp of digital data seems to have a dramatic view.

The scientific community has to deal with the handling of growing of the massive amounts of scientific literature, which is as well impressive. The number of review articles on gene technology probably exceeds the number of primary research publications in this field. According to NLM and the web database system there is an amount of metadata for more than 11 million articles (MEDLINE, 2005). There are a number of efficient, publicly available tools for data processing, storing and retrieving the information and analyzing the results in the context of existing knowledge. The NCBI's web-based search system allows searching MEDLINE according to the journal and date of original publication, retrieval of full text of the publication.

The increased complexity and the importance of searching the vast amount of bibliographic information makes the developing of improved computer-aided tools a necessity. The above is reinforced if we consider that this information is scattered throughout the published literature. Although the availability of the articles in different form that can be viewed from the scientists, there is a great need for a transformation in a computer friendly form. This is because of the limited ability to search for the computer the full texts such as PDF, html or text forms. It is difficult for a computer-based algorithm to retrieve, analyze and combine the data if there are plenty of sources in paper-based form.

The earliest respective works in the biomedical domain focused on tasks needing linguistic context and processing at level of words like identifying protein names [Fukuda *et al.*, 1998] or on tasks relying on word co-occurrence [Stapley and Benoit, 2000] and pattern matching [Ng and Wong, 1999]. During the last few years, there was a surge of interest in using the biomedical literature, (e.g., Andrade and Valencia, 1997; Craven and Kumlien, 1999; Friedman *et al.*, 2001; Fukuda *et al.*, 1998; Hanisch *et*

al., 2003; Jenssen *et al.*, 2001; Leek, 1997; Rindflesch *et al.*, 2000; Shatkay *et al.*, 2000; Yandell and Majoros, 2002), ranging from relatively modest tasks such as finding reported gene location on chromosomes [Leek, 1997] to more ambitious attempts to construct putative gene networks based on gene-name co-occurrence within articles [Jenssen *et al.*, 2001]. Since the literature covers all aspects of biology, chemistry, and medicine, there is almost no limit to the types of information that may be recovered through careful and exhaustive mining. Some possible applications for such efforts include the reconstruction and prediction of pathways, establishing connections between genes and disease, finding the relationships between genes and specific biological functions, and much more. It is important to note that *a single mining strategy is unlikely to address this wide spectrum of goals and needs* [Shatkay and Feldman, 2003].

Literature data mining is the process of identifying and extracting valid, novel and useful nuggets of information and patterns from scientific literature. It comprises two technologies; text mining and information extraction (Figure 1). Literature data mining has progressed from simple recognition of terms to extraction of interaction relationships from complex sentences.

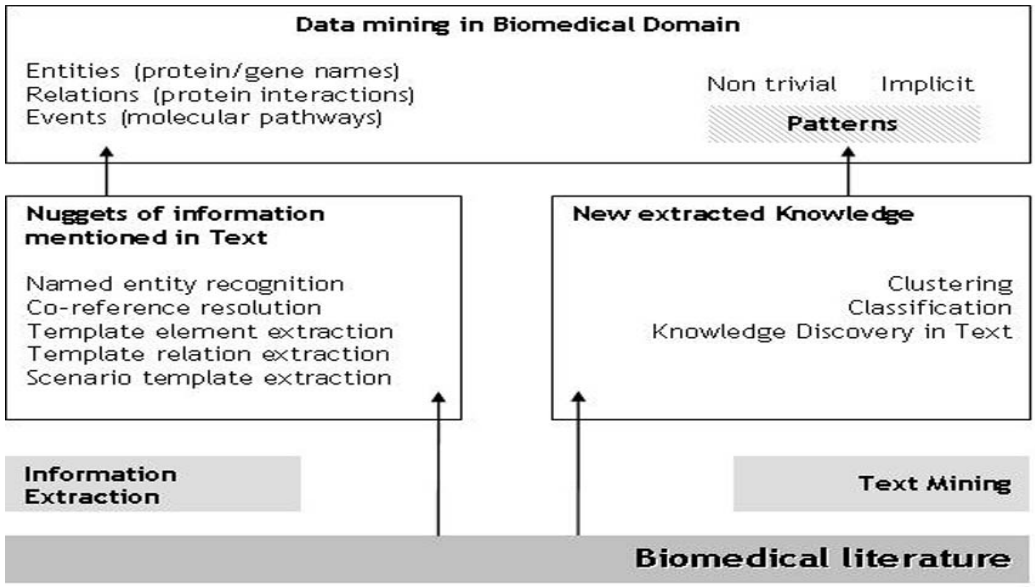


Fig. 1: Data mining in the Biomedical Domain. *Literature data mining in the biomedical domain tries aims to identify and extract valid, novel, potentially useful and ultimate understandable novel nuggets of information and patters in scientific literature. It combines two technologies: Information Extraction (IE) & Text Mining (TM). IE identifies predefined classes of entities, relations and events that are explicitly mentioned in the literature. TM identifies non-trivial, implicit, previously unknown and useful patters in text which are not explicitly mentioned in the text.*

The automated handling of text is an active research area, spanning several disciplines. These include the following: *information retrieval*, which mostly deals with finding documents that satisfy a particular information need within a large database of documents (for an introduction see, for instance, Sahami [1998], Salton [1989], Witten *et al.* [1999]); *natural language processing (NLP)*, a broad discipline concerned with all aspects of automatically processing both written and spoken language (Allen, [1995],

Charniak [1993], and Russell and Norvig [1995] are some introductory references); *information extraction (IE)*, a subfield of NLP, centered around finding explicit entities and facts in unstructured text (e.g., Cardie, 1997; Cowie and Lehnert, 1996). For instance, identifying all the positions in the text that mention a protein or a kinase (entity extraction), or finding all phosphorylation relationships to populate a table of phospohrylated proteins along with the responsible kinase (relationship extraction) are both IE tasks. Finally, *text mining* (Hearst, 1999), the combined, automated process of analyzing unstructured, natural language text in order to discover information and knowledge that are typically difficult to retrieve.

Text mining – **TM** refers to the emerging research area that can be roughly characterized as *knowledge discovery from large text collections*, thus combining knowledge discovery and text processing methods. It uses techniques from the general field of data mining (Frawley *et al.*, 1991), but since it handles unstructured data, a major part of the process deals with the crucial stage of pre-processing the document collections; term extraction (Daille *et al.*, 1994; Frantzi, 1997), and information extraction.

It is concerned mainly with the discovery of interesting *patterns* such as clusters, associations, deviations, similarities, and differences between terms, between documents, and between terms and documents (Feldman, 1999; Mladenic, 2000; Ciravegna *et al.*, 2001).

The current thesis reports on work done into five – (5) R&D directions:

1. Identification of *gene/protein--gene/protein* and *gene/protein--disease* ***associations*** following a text mining approach. The approach utilizes data-mining and statistical techniques, algorithms and metrics to tackle the problems of: (i) *identification and recognition of terms* in text-references – based on an appropriately devised and implemented algorithmic process; and (ii) *ranking* of terms and their (potential) *relations* or, *links* – based on the *MIM* entropic metric (Mutual Information Metric) to measure the respective terms' *association strength*.
2. Construction of a genes ***association network*** – based on the assessed terms (i.e., genes, proteins, diseases) association strengths.
3. ***Categorization*** / ***Classification*** of text-references (mainly from the PubMed abstracts repository) into class categories utilizing an appropriately devised classification metric and procedure, and using the most descriptive (i.e, strong) associations between terms. Pre-assignment of text-references (i.e., PubMed abstract) to categories is performed by posting respective queries to PubMed, i.e., querying PubMed with "breast cancer" the retrieved documents are considered to belong to the "breast cancer" category.
4. Assessment on the texts' categorization / classification results – based on respective PubMed abstract collections, their pre-categorization and careful experimental set-up to measure ***prediction results***, i.e., accuracy and precision.
5. Design and development of a tool – the ***MineBioText*** (Mining Biomedical Texts), that encompasses all of the aforementioned operations with extra functionalities for setting-up the domain of reference and study, e.g., gene/protein and disease names, their synonyms and free-text descriptions, text collections, parameterization of build-in algorithmic processes etc.

1.3 Organization of the Thesis

In the previous chapters we focused on the main problems met in the Biomedical domain which however impulse researchers from different but eventually assembled sections of science. The main contributions of the work are mentioned, and why literature data mining is the approach we followed to accomplish them.

In chapter (2) Background work on Text-Mining and Biomedical Literature Mining Tasks is mentioned. How Text mining and Information retrieval are combined and the fundamental tasks employed by IR approaches which are also utilised in the context of text mining. Chapter (2.3) focuses mainly on the basic tasks and approaches of Text Mining in the Biomedical Domain. We initially refer to the main problems in the biomedical literature that have to be tackled; Term Identification and Term recognition Works are mentioned in chapter (2.3.1), which as well contains the ontologies in biomedical domain. Machine Learning Approaches, including Supervised, Unsupervised learning, Support Vector Machines, and hybrid approaches are reported in the next section. Chapter (2.3.2) deals with background work referring to Interactions Discovery (Inducing Gene/Protein Correlations), and the next chapters with Classification and Clustering Approaches (2.3.3) and Biomedical Information Extraction via Natural Language Processing approaches (2.3.4).

Chapter (3) presents the main data structures, algorithms and statistical metrics used in this work. Chapter (3.1) deals with the Biomedical Texts Collections including Literature and Gene Terms that are collected; Gene and Proteins are explained as well as Parsing and Term Identification. In the next chapters (3.1) the TRIE data structure is presented, how data is structured in it, the main definitions and relations that will be used later are explained in (3.1.5) and (3.1.6) The next chapters (3.1.7, to 3.1.9) contain the main algorithms and metrics used in order to extract the Gene Association's Networks. After the weight assignment, the term frequency is measured according to the well-established formula of MIM and a gene association's network is extracted in chapter (3.1.9). In chapter (3.1.10) the classification method is described, including the 'strength' assignment and the similarity-scoring scheme applied to each gene term located in the set of abstracts; finally, the prediction formulas used for validation are mentioned.

In chapter (4), the architecture of the system is decomposed and the parts of MIM computation and gene association's Network are analyzed. The Graphical User Interface of the Application is decomposed in the next chapter (4.3); basic input and dialog Windows are explained.

The next chapter comprises the process we followed in order to validate the approach we propose. We describe the sources from which the input was retrieved, the reason why we focused on abstracts, and the whole process we followed to extract the gene network. The extracted visualized graph and the classification results for several experiments are shown in chapter (5.2), (5.3). The next part of this chapter explains the Evaluation of the *MineBioText* on a Trec Genomics Task, using an evaluation scheme provided by National Institute of Standards and Technology (http://www.nist.gov/) and the evaluation results in chapter (5.3.2).

2. Background to Text-Mining & Biomedical Literature Mining Tasks

2.1 Text-Mining: An Outline

The great amount of information referring to gene and protein related biological functions raised a great interest for automating the techniques of identification, extraction, management integration and exploitation of relevant knowledge. In order to accomplish the task of literature data mining the primary tasks of *Information Extraction* and *Text Mining* are combined. Information Extraction, as it will be detailed in the sequel, constitutes the pre-processing phase. Information extraction includes the tasks of *term and relation extraction* as well as the *co-occurrence resolution*. In the next step, where text analysis takes place, techniques from machine learning, statistics and data mining are utilized.

Text mining - TM is defined as the process of *discovering and extracting knowledge from unstructured data*, contrasting it with data mining, which discovers knowledge from structured data (Hearst, 1999). Instead of leaving the user with the problem of having to read several tens of thousands of documents, text mining gives the possibility of extracting precise facts, and finding interesting associations among disparate facts, leading to the discovery of new or unsuspected and hidden knowledge in text references. Normally TM comprises three steps (see figure 2, below):

1. In the first step includes *relevant* text-references are *collected*, mainly based on *Information Retrieval* approaches.
2. In the next step, known as ***Information Extraction***, identification and extraction of the information pieces (mainly terms or, small-phrases) from the (retrieved) texts is performed – this is done in accordance to user's requests; in principal it is based on ***Information Retrieval*** techniques, and mainly on text *parsing* operations.
3. In the last step, mainly data-mining, machine learning and statistical techniques are employed in order to induce and identify associations among the pieces of the extracted information.

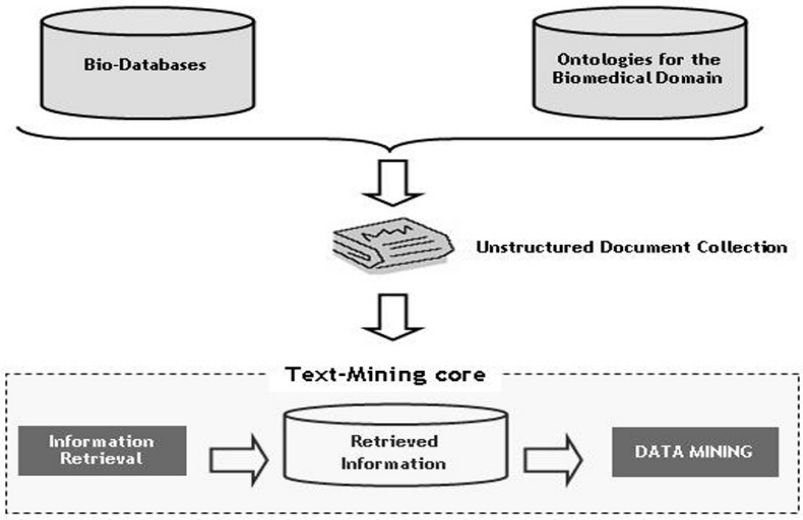


Fig. 2: *A potential view of Text Mining components and operations*

In the biomedical domain, the input ***corpus*** provides *implicit* information about the correlation between different biological objects, but the structure of biomedical literature is not always adaptable to parsing operations. Some steps are necessary in order to prepare the text for processing and *transforming* it into a representation suitable for the data-mining tasks.
A common type of text representation and processing ideal, followed by information retrieval researchers, is concerned with the contained *word*s as the basic representation unit - called '*bag of words*'. In this approach, we may distinguish two steps:

- First, an *attribute-value* representation of texts is followed, where each text document is represented as a *vector* in the lexicon space;
- Second, based on the formed vectors, a '*term-frequency*' process is initiated. The later includes operations for the removal of *inflexion* information based on *stemming* where, stop-words (or, other words from a pre-specified repository of words to disregard) are also removed. Moreover, Instead of treating each different word as a different instance of a term, a *mapping* or, *projection* operation is applied on terms that refer to grammatically or syntactically related words.

2.2 Text Mining and Information Retrieval: A Synergistic Endeavour

Information Retrieval (IR) is concerned with locating information according to user's needs. The primal task in IR is the retrieval task in which several (available) documents referring to a specific domain are used for searching specific terms (as an analogue imagine a researcher doing a literature search in a library). In this environment the retrieval system knows the set of documents to be searched, but cannot anticipate the particular topic that will be investigated. We call this an *ad-hoc* retrieval task, reflecting the *arbitrary* subject of the search and its short duration. This may be contrasted with the text mining approach and underlying philosophy: where *'long-term' are posted and tackled in the sense that 'life-long', universal and 'more' stable findings are inquired, i.e.,* ***knowledge*** *hidden in document references*. But, IR techniques present the (basic) infrastructure (approaches, techniques, systems and tools) for the text mining machinery.

2.2.1 Enabling Text Mining via Information Retrieval

Information retrieval is concerned with identifying documents that are *most relevant to a user's need* within a very large set of documents. More precisely, given a large database of documents, and a specific information need - usually expressed as a *query* by the user, the goal of information retrieval methods is to find the documents in the database that satisfy the information need. Naturally, the task has to be performed accurately and efficiently (Shatkay and Feldman, 2003). There are three fundamental tasks employed by IR approaches, which are also utilised in the context of text mining.

① ***Boolean queries and index structures***. A simple and common way for a user to express her need is through a *Boolean* query. Under this setting, the user provides a term (e.g., *OLE1*), or a Boolean term-combination (e.g., *OLE1* and *lipid*). The result is the set of *all* the documents in the database satisfying the query constraints, e.g., containing both the query terms *OLE1* and *lipid*. This query paradigm is used by the biomedical literature database PubMed (www.pubmed.org) and by many other text databases and search engines over the Web. It is supported by an index covering all the terms in the whole database of documents. Each *term* may be a single word (e.g., *blood*) or a phrase (e.g., *blood pressure*). It is common practice

to omit from the index terms that are frequent and non–content–bearing, such as prepositions. These terms are usually referred to as *stop words* and are viewed as delimiters when processing text. The *index structure* contains all the terms, typically sorted alphabetically for quick access, and holds for each term a reference to all the documents in the database that contain it. When a user poses a query, the index structure is efficiently searched for the query terms occurring in it, and all the documents found to contain the terms (or the Boolean combination of the terms) are retrieved. There are various methods to create indices and use them. The simple form of Boolean query, which has the advantage of efficient implementation over large databases, suffers several limitations: (a) the number of documents typically retrieved is *prohibitively large*; (b) a substantial part of the retrieved documents are *irrelevant* to the user's information need; and (c) many relevant documents *may not be retrieved*. For instance, if we were retrieving from PubMed, using the query '*OLE*', abstracts discussing *OLE1* under any other of its aliases (e.g., "*DNA repair protein* or *fatty-acid desaturase 1"*) would not be retrieved. Problem b above stems from the well-known *polysemy* phenomenon: a word may have multiple meanings in different contexts. For instance, when looking in PubMed for the term '*Cytosine Deaminase'* under its acronym '*CD'*, we may retrieve all abstracts referring to *Cytosine Deaminase* in which we are actually interested, but also all those discussing *Crohn's Disease* (also *CD*) which are completely unrelated. On the other hand, limitation c, stems from *synonymy*: a single concept is discussed in various abstracts under different names.

② ***Similarity queries and the vector model***. A broadly used alternative to the Boolean query is the *similarity query*, which is typically based on the *vector space* model. Under this setting, documents are viewed as *vectors over terms*. A query, *q*, consisting of many terms (it may even comprise a complete document), is in-and-of-itself viewed as a body of text, rather than merely as a search-terms combination. Thus, it too is represented as a vector. The retrieval task reduces to searching the database for document vectors that are *most similar (most distant,* as a dual process to similarity) to the query vector. Various similarity measures over documents have been devised and used (Salton, 1989).

To explicitly define the vector model, we refer to the large set of documents from which retrieval is conducted as the *database* and denote it as *DB*. The *controlled vocabulary* of the database is the set of all the terms occurring within *DB*'s documents. Let *M* be the number of distinct terms $\{t_1, \ldots, t_M\}$ in this vocabulary. A term, t_i, may be a single word or a longer phrase such as "*blood pressure*" or "*acquired immunodeficiency syndrome*" - stop-words are typically disregarded. Some systems may also *stem* words, removing common suffixes such as '*ing'* or '*e's* - as for instance with known and widely utilised Porter's stemming algorithm (Porter, 1980; 1997, Jones and Willet, 1997). A *document*, *d*, in the database is represented as an *M*-dimensional vector: $<w_1, w_2, \ldots, w_M>$, where w_i is a ***weight*** representing the *occurrence* or the *significance* of term t_i within the document. The particular choice of term-weights can significantly influence the results of a similarity search, and there are several schemes for calculating the weights, e.g., based on vector *similarity/distance metrics* like the Euclidean distance, 'cos'ine, and correlation coefficient similarity and others (Salton, 1989).

- ***Latent Semantics Indexing*** & ***Latent Similarity***. A more flexible approach that depends less on the explicit query terms and, to some extent, accommodates synonyms and polysyms is *latent semantics analysis* (Deerwester *et al.*, 1990; Dumais et al., 1988; Dumais, 1990; Furnas *et al.*, 1988). Two main ideas

underlying the method are: (i) there are abstract concepts that the explicit words in the documents are trying to convey. *Different word combinations may be used to identify the same concept* (*synonymy*), while the *same word may denote different concepts under different contexts* (*polysemy*). The *semantics* of words is the concept they are conveying. While the words are overtly present in the document, the semantics is not explicitly stated and is therefore *latent*; and (ii) a collection of documents, each represented by an *M*-dimensional vector, can be viewed as a matrix. As such, algebraic operators can be applied to it. One particular operator, namely *singular value decomposition*, can be used to identify and extract the "significant components" of the matrix. These are its largest *k* singular values, where *k* is much smaller than the original number of terms *M*. Each document in the matrix can thus be approximately represented as a linear combination of these *k* singular values, or equivalently, as a *k*-dimensional vector, rather than as the original *M*-dimensional vector. By combining these two ideas, each of the *k* large singular values of a document collection is viewed as a surrogate for a class of terms with a common *hidden* semantics. Both queries and documents are transformed and expressed as vectors over these singular values rather than as vectors over *M* terms, and the similarity measure is applied to these transformed vectors, whose dimensionality is lower than that of the original term-space. The method has shown a lot of promise, but suffers from two main drawbacks: - It was so far only shown effective on small collections of documents; and - The algebraic transform to singular-values space overrides the actual words in the documents. Thus, the method does not provide the intuition or the ability to observe the terms responsible for the document similarity.

For the shake of the completeness of the presentation, we also mention the *probabilistic models* for indexing documents and assessing the similarity between them (van Rijsbergen, 1979; Ponte and Croft, 1998; Hofmann 1999).

③ ***Text categorization***. This is the *labelling* of text references with *thematic categories* from a predefined set of category tags. There are two main approaches to categorization. One is the *knowledge engineering* approach (Hayes, 1992; Hayes and Weinstein, 1990) where the user manually defines a set of rules to encode expert knowledge regarding the correct classification of documents into given categories. The other approach is based on *machine learning* (Yang and Chute, 1994; Lewis and Ringuette, 1994; Lewis and Hayes, 1994; Lewis, 1995; Lewis et al., 1996; Larkey and Croft, 1996; Dumais et al., 1998; Joachims, 1998; Yang and Liu, 1999; Cohen and Singer, 1999; Potamias et al., 2001; Potamias, 2001; Sebastiani, 2002) where a general inductive process automatically builds a ***text classifier*** *by training over a set of pre-classified documents*.

Some indicative IR works in the biomedical domain, also inspired the work reported in this thesis, follows.

- In the context of TREC-2003, a relevant ad-hoc retrieval work is reported for the biomedical domain (Hull and Waldman, 2003). In this work, the authors target the recognition of gene and protein functions in MEDLINE/PubMed abstracts. They suggested an approach that uses simple syntax and domain semantics to identify sentences in the abstracts that suggest/point to (pre-specified) gene functions.

They ranked these abstracts by a metric that assess the number of appropriate function instances they contain. They achieved ~0.32 average precision and approximately ~0.30 R-precision[1] figures.

- In addition, TREC-2003 is referred to a task that incorporates two retrieval tasks into a set of experiments for the retrieval of known-items. They hypothesized than not all retrieval tasks should be approached by the same retrieval approach when a single search entry point is used. They applied task-classifiers on a top of traditional web-retrieval approaches. The traditional IR approach was based on the fusion of result sets generated by query runs over independent parts of the document structure. The task classifier combined query term analysis with known information resources and URL depth (Beitzel *et al.*, 2003).
- In the BioCreAtIvE competition (http://www.pdg.cnb.uam.es/BioLINK/BioCreative.eval.html) a similar to ad-hoc retrieval task is reported for task 1B and the provided text references (Hachey *et al.*, 2004). The difference with the previous works is found in the limited number of the used documents. In this work the authors, identify gene references in the texts with respect to (accessible) databases of organism gene identifiers, and presented several approaches for gene identification, lookup, and disambiguation. The system provided an organism database containing unique gene identifiers with lists of synonyms and an accompanying abstract. They created a list of unique gene identifiers for genes that were mentioned in the abstract, including explicit mentions as well as those implicitly mentioned in gene mutants, alleles, and products. Results were presented with two possible baseline systems and a discussion on the source of precision and recall errors, as well as an estimate of precision and recall for organism-specific gene synonym lists.
- Another kind of conventional ad-hoc retrieval task, where the system is expected to retrieve relevant documents in response to a user's query is reported in (Song *et al.*, 2003). The authors tested several techniques such as a 'phrase indexing strategy', two query-weighting methods, and two post-processing methods. Documents and queries in this task were limited to the biomedical domain. According to the experimental results, query weighting methods and document filtering methods can improve the performance of the retrieval system, but there still remain a room for improvement. This task had some significant differences to previous ad-hoc tasks, because of its environment.

The work presented in this thesis utilises and introduces novel IR approaches and techniques – for document indexing, term identification and text categorization, in the context of an integrated biomedical text mining system *MineBiotext*.

2.3 Text-Mining in the Biomedical Domain: Basic Tasks & Approaches

Despite the great need that emerges from the large amount of bibliography concerning biomedicine, several problems pop out transmuting our task into a challenge.

i. A main problem that appears concerns the *interoperability* of the available biomedical resources especially with respect to the non-unified format they seems

[1]R-precision is the precision after R-docs retrieved (http://www.scils.rutgers.edu/~muresan/IR/TREC/Proceedings/t8_proceedings/appendices/A/appendixa.cover.pdf).

to appear. The most important problem is the variety and multiplicity of utilized *terminologies* as well as the *lexical coverage*. The problem arises from the fact that there is not a standard adopted vocabulary.

The problem of gene or protein name identification in the free-text publications is emerging and not adequately tackled. In some organisms scientists have enjoyed applying gene names with primary meaning outside the biological domain[2]. Names such as 'vamp', 'eve', 'disco', 'boss', 'zip' or 'ogre' are therefore not easily recognized as representative gene references [Proux *et al.*, 1998]. Other major problems that are considered refer to the existence of *many different names* for the same entity and cause problems to literature (free-text) searching algorithms. In particular *synonymy* reduces the number of recalls of a given object (gene, protein, molecular pathway and function, disease etc).

According to Ensemble,[3] several naming conventions are entry points into the ensemble database (http://www.ensembl.org/index.html). Identification for a gene can be the Ensemble Gene ID, the ensemble identifier, known gene name, OMIM diseases and free text search of OMIM[4], SWISSPROT (http://ca.expasy.org/sprot/) and InterPro annotation (http://www.ebi.ac.uk/interpro/). There seem to be a great need for the organization and centralization of terminologies in the biological domain. This calls for experts from different but eventually assembled sections of science.

ii. Another problem that lies over the working out of the literature is the fact of *multiple meanings* for a given term. The effect is the reduction in precision, and ambiguities of a term's sense. The term 'insulin' for example can refer to a gene, a protein, a hormone or a therapeutic agent depending on the context.

Additionally the use of *pronouns* and *definite articles*, *long*, *complex* or *negative sentences* or, those in which information is implicit can be also a speculative situation for a searching algorithm. The term ambiguity can also arise from the fact of identification with common English words or bad encoding of human genes; for example the 'BCL-2' family of proteins. The problem becomes worse because the existing biology terminological resources lack of information that can support *term disambiguation*.

2.3.1 The 'Curse of Genes Naming': Term Identification & Recognition

The basic tasks in text-mining include: the *classification* or, *categorization* of texts to specified *classes* (or, *categories*) according to their *content*. In an approximate setting, the content of a text reference may be approached by the identification of the *terms occurring in a text and their potential interdependence*.

The basic step towards *term identification* is the detection of given terms in the corpus. It consists of three steps: (i) *term recognition*; (ii) *term classification*; and (iii) *term mapping*. For each step several approaches and trends have been proposed.

Term recognition refers to the '*marking*' of the words being (pre)specified to belong to the domain. The occurrence of a single term has such significance as well as the co-

[2] A representative example is the fruit fly *Drosophila*.
[3] Ensemble is a joint project between EMBL - EBI and the Sanger Institute to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes. Ensemble is primarily funded by Wellcome Trust.
[4] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM

occurrence with other terms. Potential consideration that must be consulted is the differentiation between terms and non-terms and the variation of a specific one. Here comes the need for *ontological* and *terminological* support.

Ontology Utilization & Engineering. The biological domain is supplied with several *ontologies*. For the genomics domain the most known and widely utilized ontologies include the "Gene Ontology" (GO, www.geneontology.org), and the "HUGO Gene Nomenclature" (http://www.gene.ucl.ac.uk/nomenclature/ - from the Human genome Organization – HUGO, http://www.hugo-international.org/). For the medical domain known and widely utilized ontologies include: the Unified Medical Language System (UMLS; http://www.nlm.nih.gov/research/umls/), and the Medical Subject Headings (MeSH, http://www.nlm.nih.gov/ mesh/meshhome.html).

The critical step in *ontology engineering* is the identification of relations between terms based on the mapping of the terminologically valid concepts to the terms. A basic problem in the mapping of terms (genes, proteins, pathways, diseases etc) to biomedical ontologies (and related databases) refers to the fact that a term can receive *multiple semantic tags*. Unfortunately, not all ontologies are devised in a consistent way, following best practice design approaches. Subsequently, a re-*engineering* process is needed to render them more useful for applications such as text mining. However, even with a well designed ontology and appropriate lexicons there is an extra need to establish the missing link, i.e., to provide the mappings from terms in lexicons into corresponding ontology concepts.

A common ontology engineering approach is the combination of different ***dictionaries*** with utilization of various *distance-measures* (e.g., the *edit distance*) in order to achieve flexible string matching. In (Krauthammer *et al.*, 2000) a method is presented based on 'approximate string comparison' for the recognition of genes' and proteins' names and their variations. In the reported approach, both protein dictionaries and target text were encoded using the "nucleotide" code (a four-letter encoding over the {A, C, G, T} alphabet). Then, BLAST-like techniques (originally used for alignment of DNA and protein sequences) are applied to the converted text in order to identify character sequences that are similar (i.e. may be aligned) to existing gene and protein names (also encoded by the corresponding nucleotide codes). In the reported experiments the approach achieved ~0.79 recall, and ~0.72 (overall) precision figures.

Recent techniques for term identification and recognition in biomedical are based on '*episodes/episode rules*' and '*rule based information systems*'. These approaches rely on finding information in text references partly through *dictionary look-up* by looking-up to resources which are divorces from the reality of the textual term. They utilize dictionaries as term resources, as well as *term formation patterns*. With a (mainly manual) technique respective mapping and recognition rules are formed. In a recent workshop (Ananiadou *et al.*, 2005) a general grammar based technique was suggested that utilizes a 'morphological unification grammar's and a lexicon with instances of specific affixes, roots, and Greek/Latin neoclassical combining forms. An *episode* in data mining is the assignment of temporal values to items of data. It is actually a sequence of *tuples* consisting of a *feature vector* and an index describing its temporal location. Specifically in a text reference the tuple representation is used to represent each word occurrence in the document and its location [Alohen *et al.*] However a feature is not restricted to just a word (gene or protein), it can also be a phrase, punctuation mark or a mark-up tag. Episodes are designed to look for patterns such as co-occurring terms or phrases which might be used in constructing *'concordance' lists* or, learning grammatical rules in a particular type of text.

- ***Machine learning*** - **ML** approaches as well as statistical techniques are also utilized. While statistical approaches mainly address the recognition of general terms, ML-systems are usually designed for the integration of both term recognition and term classification tasks. An issue that is handled by ML methods is the selection of *representative* features (i.e., features with highly accurate recognition and classification figures), as well as the detection of term boundaries of *multi-word* terms. However, the success of the machine learning approach is bounded by appropriate training sets – for both term and text collections, at least for supervised ML tasks. This limits its applicability to biomedical literature mining tasks because of the lack of representative and reliable training resources (i.e., benchmark term and text sets). In general, ***supervised*** ML algorithms include: *decision tree* learners, *neural networks*, *support vector machines* (SVM), (Naïve) *Bayesian* approaches, *linear function* learners, and *instance-based* classification (Mitchell, 1997). Respective ***unsupervised*** ML algorithms include: *hierarchical*.
  - In (Collier *et al.*, 2000) a Hidden Markov Models (HMM; [Rabiner & Juang, 1986]) and specific *orthographic* features (e.g., "*consisting of letter and digits*" and "*having initial capital letter*"), are utilized for the discovery of terms pre-assigned to ten classes. The GENIA corpus is used (http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/topics/Corpus/). The reported results depend on the quality of training resources; an example was that of the protein class which was the most frequent in the training sets; so, influencing the final classification. On the other hand, instances of the RNA class were very rare, so it was difficult to learn descriptive and highly accurate features. Even though, the results were encouraging with an F-score of ~0.76 achieved. A similar approach is reported in (Morgan *et al.*, 2003) for the recognition of Drosophila gene names. Besides to orthographic features, *prefix*/*suffix* information, *part-of-speech* (POS) tags, and *noun heads* may be also utilized, as in (Shen *et al.*, 2003). They reported that POS-tags proved to be among the most useful features. They achieved F-scores of ~0.17 to 0.80 depending on the class. The overall F-score was 66.1% and the protein class F-score was ~0.71.
  - In (Stapley *et al.*, 2002) a SVM approach is used to classify terms derived by standard *term weighting* techniques and predict the cellular location of proteins from abstracts. The accuracy of the classifier on a benchmark of proteins with known cellular locations was better than that of a support vector machine trained on amino acid composition and was comparable to a hand-crafted rule-based classifier (Eisenhaber and Bork, 1999).
  - In (Finkel *et al.*, 2004) an ML system is presented for the recognition of names in biomedical texts. The system could make extensive use of local and syntactic features within the text, as well as external resources including the web; it achieved an F-score of 0.70 on the Coling 2004 NLPBA/BioNLP shared task of identifying five biomedical named entities in the GENIA corpus.
  - Also in (Dingare *et al.*, 2004) a maximum-entropy based system is presented aiming towards the identification of Named Entities (NEs) in biomedical abstracts. Its performance is also presented, on the only two biomedical Named Entity Recognition (NER) comparative evaluations that have been held to date, namely BioCreative and Coling BioNLP. The system obtained an exact match F-score of 0.83 in the BioCreative evaluation and 0.70 in the BioNLP evaluation.
- Another trend used towards term identification is based on ***hybrid*** approaches and systems. They combine rule and statistically based techniques, as well as linguistic

and contextual processing in order to rank candidate terms. Hybridization is performed with an 'amalgama' of machine learning, dictionary-based and probabilistic approaches.

- A specific protein and gene name tagger, "ABGene", is presented in (Tanabe and Wilbur, 2002), with training performed with MEDLINE abstracts and adaptation of Brill's POS tagger (Brill, 1992). In order to improve precision and recall they performed transformation rules for filtering and "recovering" of results. False positive term names were filtered-out by a list of precompiled biological and non-biological terms. False negatives were recovered by a list of terms that were compiled from the Locus-Link database (http://www.ncbi.nlm.nih.gov/LocusLink/, Please Note: As of July 1st, 2005 the NCBI is no longer updating Locuslink: http://www.ncbi.nih.gov/entrez/query.fcgi?db=gene.) and Gene Ontology (www.geneontology.org). Context words were also consulted. If a word was surrounded by "good" context words, it was tagged as a protein or gene. "Good" context words had been generated by a probabilistic algorithm by assigning Bayesian weights to all non-gene names that co-occurred with known names in the training set. Compound names were also extracted by relying on the combination of frequently occurring components in known multi-word gene names and a set of regular expressions. Overall, ABGene achieved precision figures in the range of 0.60 to 0.90.
- A remarkable hybrid method called *C/NC-value* is reported in (Frantzi *et al.*, 2000) for the task terminology recognition in many biomedical sub-domains. A set of *morpho-syntactic filters* suggested the term candidates, and statistical measures on corpus estimated the term-hoods. More specifically the frequency of occurrence of a term, the frequency of occurrence as a substring of other candidate terms, the number of candidate terms containing the given candidate term as a substring, and the number of words contained in the candidate term were the main factors that were assigned to each candidate according to their co-occurrence with top-ranked context words. Reported results (with wxperiments performed on 2,082 MEDLINE abstracts) exhibit precision figures of 0.91-0.98 for top ranked terms.
- In (Bodenreider *et al.*, 2002) a corpus and both lexical and terminological knowledge are utilised in order to extend an existing biomedical terminology. The adjectival modifiers were removed from terms extracted from the corpus (three million noun phrases extracted from MEDLINE), and de-modified terms were searched in the terminology. A phrase from MEDLINE became a candidate term in the medical UMLS *Metathesaurus*[5] if the following two requirements were met: 1) a de-modified term created from this phrase was found in the terminology, and 2) the modifiers removed to create the de-modified term also modify existing terms from the terminology, for a given semantic category. A manual review of a sample of candidate terms was performed. The results showed that out of the 3 million simple phrases randomly extracted from MEDLINE, 125,000 new terms were identified for inclusion in the UMLS – 0.83 of the 1000 terms reviewed manually were associated with a relevant UMLS concept.

[5] http://www.nlm.nih.gov/research/umls/

- In (Zhou *et al.*, 2005) a protein/gene name recognition system is presented in which three classifiers are combined: Support Vector Machines (SVM) and two discriminative HMMs Hidden Markov Models, with three post processing modules including an abbreviation resolution module, a proteins/gene refinement and a simple dictionary matching module. The experiments achieved an F-score of ~0.83.
- In (Ibushi *et al.*, 1999) two classification methods based in SVM and Adaboost are combined on a set of MEDLINE abstracts, achieved the best results at 0.48 precision, 0.49 recall, and 0.48 F-value levels.
- A named entity recognition system called PowerBioNE is introduced in (Zhou, 2004) where, HMM, SVM and sigmoid approaches are combined. Evaluation showed F-measure of 0.69, 0.71 and 0.78 on different classes of the GENIA corpus.
- In (Dingare *et al.*, 2005) a maximum-entropy based approach is followed that incorporates a diverse set of features for identifying genes and proteins in biomedical abstracts. The system was entered in the BioCreative comparative evaluation and achieved a precision of 0.83 and recall of 0.84 in the "open" evaluation and a precision of 0.78 and recall of 0.85 in the "closed" evaluation.

Although it is not possible to thoroughly compare different systems with different targets and test collections for the task of term identification and recognition - TAR in the biomedical domain, there are attempts aiming to organize different evaluation schemes such as the BioCreative[6] endeavor. In general, and with an approximate summation, we may report that TAR precision and recall figures range from 0.70-0.90, and (around) 0.70-0.85, respectively.

2.3.2 Interactions Discovery: Inducing Gene/Protein Correlations

The following works indicate the fact of identifying biomedical terms through *Machine learning approaches* focused on *rule induction* and *Support Vector Machines (SVM)*.

- (Bunescu et. Al., 2002; 2004) utilize machine-learning approaches based on SVM techniques to identify human proteins with higher accuracy than several previous approaches. In these references, it is also demonstrated that various rule induction methods are able to identify *protein interactions* with higher precision than manually developed rule.
- In (Bunescu *et al.*, 2005) an information extraction system is presented for identifying human protein names in MEDLINE abstracts, and subsequently, to extract information on interactions between the proteins. They demonstrated that ML approaches using SVMs and maximum entropy are able to identify human proteins with higher accuracy than several previous approaches. They also demonstrated that various rule induction methods are able to identify protein interactions with higher precision than manually developed rules.

---

[6] BIOCREATIVE (Critical Assessment of Information Extraction systems in Biology) was organized for the first time as a challenge cup in 2003, in which one of the subtasks was related to protein/gene name recognition and identification (in the same, shared set of documents). The evaluation showed that the best methods achieved F-scores of 80%, with both the best precision and recall values of around 80%. For details see http://www.mitre.org/public/biocreative/.

- In (Dehoney *et al.*, 2003) the present NLP approach is presented where, with the aid and utilization of Gene Ontology (GO) collects and translates unstructured text data into structured interaction data. NLP is realized by a rule induction program, RAPIER (http://www.cs.utexas.edu/users/ml/rapier.html). RAPIER was modified to learn rules from tagged documents, and then it is trained on a corpus tagged by expert curators. The resulting rules are used to extract information from a test corpus automatically. Extracted genes and proteins are mapped onto Locuslink[7], and extracted interactions are mapped onto GO. Once information is structured it is stored in a molecular-pathway database and this formal structure allows to perform advanced data mining and visualization.

2.3.3 Biomedical Text Categorization: Classification & Clustering Approaches

Biomedical Text Categorization – BTC aims to the better retrieval of relevant text references, and improves the potential of knowledge discovery (i.e., gene/protein correlations) from the retrieved texts. In general we may refer to two main methods of text categorization: (i) *unsupervised* method – the task here is the induce clusters of texts with high *intra* (within cluster) similarity, and with high *inter* (between clusters) dissimilarity; and (ii) supervised method – the task here is to devise a feature-based prediction model (procedure, metric or both) in order to predict the category in which a text belongs (a training phase is required, based on collections of pre-categorized text references). Below we present some of the basic R&D developments for this task – in most of the presented methods both supervised and unsupervised text categorization approaches are utilized.

Classification approaches in Text-Categorization

- In (Sathiya *et al.*, 2002) a BTC approach is presented for the Curation of Biomedical Literature[8], realize by an automated text classification system for the classification of biomedical papers. Text classification is based on the existence of experimental evidence for the expression of molecular gene products for specified genes within a given paper. The system performs pre-processing and data cleaning, followed by feature extraction from the raw text. It subsequently classifies texts using the extracted features with a Naïve Bayes Classifier.
- In (Mullen *et al.*, 2005) a sentences classification system is presented which uses Naïve Bayes and SVM methods. The method was tested on ZAISA-1 Dataset[9], a set of 20 full-annotated journal articles. On full articles, the highest overall F-score was 0.70, obtained by the SVM model.
- In (Yildiz and Pratt, 2005) the presented text classification system extracts medical phrases from text by incorporating a medical knowledge base and natural language processing techniques. Experiments were made on MEDLINE documents from the OHSUMED dataset. They achieved the best results with the hybrid method of combining bag of words and bag of phrases: F-score = 0.60, precision 0.87 and recall = 0.46.
- In (Craven and Kumlien, 1999) a ML system is presented to induce routines for extracting facts from the text. It applies a statistical text classification method

[7] http://www.ncbi.nih.gov/entrez/query.fcgi?db=gene
[8] It was presented in KDD Cup 2002 as task 1.
[9] http://research.nii.ac.jp/~collier/projects/ZAISA/index.htm

and a relational learning method in a corpus of 2,889 MEDLINE abstracts by querying on the names of six proteins. The most significant results on several datasets achieved 0.77 precision at 0.30 Recall. The same authors have also presented an approach to decrease the cost of learning information-extraction routines by learning from "weakly" labeled training data (Craven and Kumlien, 1999).

- In (Eskin and Agichtein, 2004) the presented approach aims to the discovery of protein functional regions and combines text mining and DNA sequence analysis. It is based on the creation of a seed dataset to train a text classifier and uses the classifier to predict additional sequences, which correspond to a class. It combines SVMs and the Hamming distance (for sequences similarity), achieving approximately a precision level of 0.80.

Clustering approaches in Text-Categorization

- In (Chang *et al.,* 2001) a modified PSI-BLAST similarity-based process is utilized. They showed that supplementing sequence similarity with information from biomedical literature search could increase the accuracy of homology search result.
- In (Iliopoulos *et al.,* 2001) a method for clustering MEDLINE abstracts - based on a statistical treatment of terms, together with stemming, a `go-list', and unsupervised machine learning, is given.
- In (Harte *et al.,* 2003) procedures and tools are presented for the pre-qualification of documents for further analysis. A corpus of documents for proteins was initially built from MEDLINE search terms. The documents space was examined using a strategy employing *Latent Semantic Indexing* (LSI; http://www.cs.utk.edu/~lsi/), which uses Entrez's "related papers" utility for MEDLINE. Document's relationships were visualized using an undirected graph and scored by their relatedness. Distinct document clusters, formed by the most highly connected related papers, are mostly composed of abstracts relating to ones aspect of research. This feature was used to filter irrelevant abstracts, which resulted in a reduction in corpus size of 0.10 to 0.30 depending on the domain. The excluded documents were examined to confirm their lack of relevance. Corpora consisted of the most relevant documents thus reducing the number of false positives and irrelevant examples in the training set for pathway mapping. Documents were tagged, using modified version of GATE2, with term based on GO for rule induction using RAPIER (http://www.cs.utexas.edu/users/ml/rapier.html).
- In (Uramoto et al., 2004) a text-mining system for knowledge discovery from biomedical documents is presented. It was the application of "IBM TAKMI for Biomedical Documents" to facilitate knowledge discovery from very large text collections and database. The respective set of tools, designated MedTAKMI, was an extension of the TAKMI (Text Analysis and Knowledge MIning) - a system originally developed for text mining in customer relationship management applications. MedTAKMI dynamically and interactively mines a collection of documents to obtain characteristic features within them. It also utilizes natural language techniques to extract deeper relationships among biomedical concepts.

2.3.4 Natural Language Approaches in Biomedical Information Extraction

Natural Language Processing - NLP has been applied to a broad range of information extraction problems in biology such as the recognition of protein interactions from scientific text.

- In (Set *et al.*, 1999) the presented NLP-based system is used to discover knowledge from GeneBank DNA sequences databases. It utilises a grammatical model of gene structure to create a parse tree. The parse tree was transformed into an augmented feature table that represented the gene structure, and through it a classification hierarchy that reflected the evolutionary relationships between genes is built.
- In (Dickerson and Berleant, 2003) a project is presented for the development of a publicly available software suite called the Gene Expression Toolkit (GET). A Java™-based tool helps to dynamically find and visualize metabolic networks. The overall system is quite complex. A text-mining tool pulls out potential metabolic relationships from the PubMed database. These relationships are then reviewed by a domain expert and added to an existing network model. The results are visualized using an interactive graph display module. The basic metabolic or regulatory flow in the network is modeled using fuzzy cognitive maps. Causal connections are pulled out from sequence data using a genetic algorithm-based logical proposition generator that searches for temporal patterns in microarray data.
- In (Gondy *et al.*, 2003) the Genescene system is presented. A tool for biomedical researchers where, research findings and background relations are automatically extracted from text and experimental data. The extracted relations were evaluated by qualified researchers and are precise. A qualitative ongoing evaluation of the current online interface indicated that this method was more useful and efficient than keyword based searching.
- Finally REGEN - Retrieval and Extraction of GENomics Data (Tasmin, 2003) is an NLP-based system that retrieves and extracts information from genomic data. The retrieval task is based on the combination of exact- and partial-match searching approaches using syntactic and semantic cues as patterns.

2.4 MineBioText: Contributions on Text Mining

2.4.1 External Static Term Repository vs. Term Identification

Most of the current research efforts are primarily focusing in *Term Identification*. Namely, they parse the abstracts to identify ***indicative*** terms, or else terms that are not common English words, but they reveal a potential significant role context-specific. By introducing these methods there is no need to provide an external repository of ***Gene Terms***. The repository is self-generated via various methods. One of the most known methods is the *TF * IDF* metric. For each term, we measure the number of documents that do contain (at least once) a specific word that is a potential term.

$$TF = \frac{n_i}{\sum_k n_k}$$

With n_i being the number of occurrences of the considered term, and the denominator is the number of occurrences of all terms.

$$IDF = \log\left(\frac{|D|}{|d_j \supset t_i|}\right)$$

- $|D|$ *is the number of documents in the corpus.*
- $|d_j \supset t_j|$ *is the number of documents where the term t_j appears (that is $n_j \neq 0$).*

$$TF\_IDF = TF \cdot IDF$$

The term frequency in the given document gives a measure of the importance of the term within the particular document.

A term that scores high in this formula is considered an ***indicative*** term and subsequently it is stored. The main disadvantage of this method is that it cannot reveal the potential *semantics* of the terms discovered. For example if we would have to visualize a graph where the nodes where terms identified by this procedure, each node could be a gene, protein, disease name or something completely irrelevant.

Our approach considers a fixed, stable set of terms. These terms are already stored in a public available database named ***Ensembl*** (http://www.ensembl. org/index. html)and can be browsed via a special tool called Entrez(http://www.ncbi.nlm.nih.gov/entrez/ query.fcgi). Moreover, this set of terms is constantly curated by a set of experts, where new genes/proteins discovered, or a new nomenclature is introduced. By using an external provided set of nomenclatures, we introduce a novel area for bioinformatics where the term identification process is neither essential nor necessary.

2.4.2 Semantic Related Terms vs. Simple set of Terms.

According to related works, the set of terms discovered or supplied is a simple vector of unprocessed terms. Although this collection may be huge covering all the aspects of gene-naming it cannot hold the semantic relations that may exist between different

nomenclatures. Especially in the gene-naming domain we have identify three types of semantic relations:

- ***Significant Term/Relation***. This name stands for a unique identifier among ***Gene Terms***. Like the primary key in relational database notation, this relation stands for a nomenclature that all the other nomenclatures should be assigned to. Each gene name has one or more primary identifier. For our implementation, we have chosen the ***Ensembl*** identifier as the *Significant Term*.
- ***Synonym Terms/Relations***. As we have seen one of the major drawbacks in biomedical literature text mining is the existence of many heterogeneous naming conventions, or else nomenclatures. Fortunately, the service Entrez of ***Ensembl*** has a state of the art database with all naming convention with respect their Ensembl nomenclature. By querying this database we were able to extract a numerous of nomenclatures and build a vigorous set of gene names under almost all possible nomenclatures.
- ***Free text description Words/Relation***. This contribution will be discussed in the next chapter.

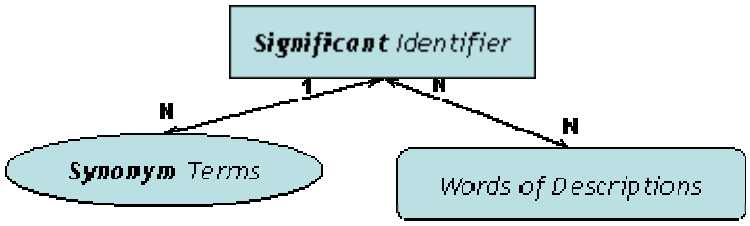


Fig. 3: Semantic Relations of Gene Terms. For each *Synonym Term* the corresponding *significant identifier (Ensembl id)* is unique, but for each *significant identifier* the corresponding synonym terms can be more than one. As we will explain in the next subchapter, for each word belonging to a *free text description* the corresponding *significant identifiers* can be more than one, because of the fact that the *free text descriptions* can contain common words.

2.4.3 Free Text Description vs. Plain List of Terms.

All existing works in B*iomedical Text Mining* consider a plain set of gene terms that should be located in the document corpus. Meanwhile we have noticed that the most genes are referred indirectly by domain specific words. These words can be found in a free text description form provided by many sources such as the ***Gene Ontology Consortium***. These descriptions are also curated and managed officially by a group of experts. In this Thesis we propose a general schema and algorithm to mine their semantics, store the underlying information and to retrieve it during document parsing. Although many references in genes are made through words presented only in gene descriptions, the presence of such word in an abstract does not ensures the direct reference of a gene. We resolve this conflict by introducing the *Float Term Vector* presented in the next section.

2.4.4 Float Term Vector vs. Binary Term Vector

By introducing *free text description* as a possible form of input, we have to express the presence of a gene term in relative, not binary way. That is, if a gene term from a

specific nomenclature is located in an abstract then we are certain that the specific gene is referred. On the other hand, if we locate a word that is used in a description of more than one genes then we are not certain to which gene we are referred. This is why we introduce a metric to indicate how strong a gene is referred in an abstract. Other works often use a binary vector where a gene is located or absent in an abstract. With *float term vectors*, we enrich the expressing ability of our mechanism.

2.4.5 Implementation Contributions.

- ***'Yes Word' List.*** In contrary with the ***Stop Word List***, we can limit the inserted set of genes to these exported by an external study. Namely, the *Term Vector* can contain a customized set of genes. The inserted genes can also be part of any naming convention. This is useful if we want to extract the underlying relations between genes that have discovered to exhibit some form of common behavior different that the coexistence in biomedical documents. These external discoveries could be, sequence similarity (i.e. through a BLAST algorithm), expression similarity (i.e. through microarray data mining operations) or functional similarity (i.e. through biomedical research).
- ***Arbitrary Term Hierarchy.*** According to our implementation, we can predefine the kind of interactions that we want to be revealed. For example, we can insert different nomenclatures for gene terms, diseases and proteins. Then we can state that we want only interactions that include genes-genes, gene-diseases and proteins-diseases interactions. By that mean, we exclude gene-protein interactions that are very common and may be subject to redundant information.

2.5 Other system approaches vs. MineBioText

At present, most literature analysis tools use some form of text mining and focus on interactions between proteins. An example is ***InterWeaver***, which automatically extracts interactions from sources that include niche databases of curated protein interactions, and scientific abstracts [Zhang Z., 2004]. In addition, such tools tend to focus on particular domains, such as cancer and neuroscience, lacking of scalability. A good example of a text-mining tool that specifically looks at neuroscience articles is ***NeuroText*** [Crasto CJ., 2003]. Another well-known example of a literature analysis tool that has been used to automatically extract gene-gene associations from Medline abstracts and assist with microarray data analysis is ***PubGene*** [Jenssen T.K., 2001, *PubGene Gene Database and Tools: http://www.pubgene.org/].* It is important to note that although these tools perform important tasks, they are not integrated text-mining environments because they do not integrate information from multiple research options. They perform standard gene-gene relation mining through a unique abstract repository. Critical to the successful application of text-mining tools is the integration within the same computer environment of multiple research parameters. At present, most systems will integrate data between two parameters, usually gene expression and phenotype responses. An example is the combination of genomic information with clinical data for personalized medicine [Pittman J., 2004]. This integration is currently being mainly applied to human cancers and has the potential to also evolve into a significant predictive capability [Nervins JR., 2003]. Another partial integrating text-mining tool is the *Dragon TF Association Miner* (DTFAM), which carries out text mining of abstracts from scientific papers and focuses on integrating links between transcriptions factors with disease and terms from the Gene Ontology database.

2.5.1 Comparison with Commercial approaches

Several commercial biomedical-text analysis platforms are currently available. Some of them have been developed directly by pharmaceutical companies, such as the *Novartis Knowledge Space Portal* [*http://www.novartis.com/*]. Also, bioinformatics companies have constructed biomedical-text-mining applications such as the Alma Knowledge Discovery system [*http://www.almabioinfo.com/*], which incorporates powerful database systems, version control, security systems and integrated representation mechanisms. There are also other commercial text-mining and knowledge-discovery applications, including *Biovista* [*http://www.biovista.com/*], *BioWisdom* [*http://www.biowisdom.com/*], *SAS Text Miner* [*http://www.sas.com/technologies/analytics/datamining/textminer/*] and *TextSense* [*http://www.inforsense.com/products/textsense.html*]. *BioWisdom*, for example uses an extensive ontology of pharmaceutically relevant concepts within its knowledge-discovery platform. *Biovista* exploits the use of different views or representations of biological knowledge, taking into account context information, and can extract interactions between genes and proteins from free text. Any interaction identified by the system is subject to manual verification, so the correct identification of these interactions is performed by a human user. This manual curation is not only time and resource consuming but is reflective of any bias the human expert will have and does not lend itself to convenient and frequent updating.

2.6 In depth differences between MineBioText and TREC-genomics assignments

The intrinsic challenge of Genomics Track The goal of *Trec 2004 Genomics Track* was to create test collection for evaluation of information retrieval and related tasks in the genomics domain focused on the biomedical scientists need in order to gather biomedical literature. The considerable challenge for the whole attempt was not only better Information Systems and Management in the Biomedical Domain but also the improvement of information extraction and text mining. The submitted works include Term identification using several biomedical ontologies, *Term Tagger, Probabilistic Model for Stemming, Synonymy Management,* ***GO/MESH*** terms retrieval, traditional Machine Learning techniques (*Bayes Classifier, Support Vector Machines (SVM), Decision Trees* e.t.c.) as well as novel ML-techniques , novel classification techniques, statistical similarity schemes, *Inverse Document Frequency* (IDF) and *TF-IDF* retrieval schemes.

The value of co-occurrence Many efforts use various grammatical and *natural language processing* techniques (NLP) to extract genes/proteins associations.

Several issues must be considered such as the *sentence* and *word-level tokenization, stemming, entity identification, part of speech tagging, stemming, and abbreviation expansion.*

We considered the **efficiency** of a system in which the *Biomedical Annotation* was retrieved and recognized in the set of abstracts, instead of using Natural *Language Processing* techniques to extract the terms.

The difficulty is that current technology is not at the level where it can correctly ***identify the relationships*** from a sentence and accurately link them to the genes or other biomedical objects with an ***acceptable accuracy and recall*** across all domains.

The second problem is the ***disambiguation*** of gene names resulting from ***polysemous terms***, while ***synonyms*** create difficulties when attempting to present information in a qualified way. (Persidis A., 2004)

MineBioText expresses the fact of co-occurring of terms by means of statistic measurements. ***Mutual Information Measure***, a well-established formula, used to assign associations between biomedical objects. For each pair of Gene Terms located in the literature, the estimated MIM value indicates the '***strengthens'*** of their correlation. The task of database population by the discovered relationships between genes/genes-proteins and gene-diseases remains as future work.

In contrast with NLP works, in this work entity identification is performed by *string matching* in the bibliography given the set of gene/protein/disease Terms from the ***Ensembl Genome Browser***, the biological annotation repository.

In contrast with NLP-techniques, the ***DB-population*** by sets of identified *Gene Terms* cannot be achieved, since biomedical terms are not discovered in text but retrieved from ***Ensembl***.

A gene/protein/disease ***associations' network*** is constructed through the discretization of the computed values of MIM, emphasizing the most important relations discovered in the set of abstracts.

3. Mining the Biomedical Literature with MineBioText

The MineBioText system encompasses a set of operations summarised below.

MineBioText:
Tasks/Sub-Tasks tackled, Operations & Services

1. Literature **collection** – abstracts from ***PubMed*** (www.pubmed.com); and Genes/Proteins **Terminology** utilization from the ***Ensembl*** (http://www.ensembl.org/index.html) and ***GO*** (www.geneontology.org) resources.
   - Special operations and services are provided for the device of the **domain-of-reference**, i.e., genes/proteins and diseases, their synonyms, and their ***GO descriptions***. The provided operations are adaptable to different domains of reference. Therefore, ***MineBioText*** is easily **adaptable** to other, except the biology, domains of reference.
2. **Terms identification** – implemented by specially devised **parsing** operations:
   - Elaboration of an efficient tree-based data-structure to parse documents for **terms** (genes/proteins, diseases) **identification**;
   - Stemming and removal of common used words and patterns; a special exclusion-words lexicon is also provided accompanied with special operations to edit and revise it.
   - Formation of a special **vector-based** data-structure to hold the terms, and their frequency of occurrence, in the given text references.
3. Extraction of **Gene/Protein Associations** & **Genes/Protein Correlations Network Construction**:
   - Computation of ***weight values*** to the extracted terms according to frequency statistics.
   - Estimation of ***term-hoods*** including occurrences as well as co-occurrences of gene/protein terms in the corpus through; the basic correlation measure is based on the Mutual Information Metric – MIM entropic, which is applied on the terms weight values.
   - Discretization of the extracted MIM genes/proteins correlation –MIM values into three levels of ***association-strengths***: Strong, Medium and Weak.
   - Construction of ***Genes/Proteins Associations Network*** - based on the computed genes/proteins association strengths.
   - ***Visualization*** of the genes/proteins associations' network based on the utilization of the tulip graph tool – based on appropriate formatting of the output genes/proteins associations network (http://www.tulip-software.org/).
4. **Texts Categorization**:
   - Selection of ***highly discriminant terms*** – based on the terms included in high ***weight*** values; user specified or, automatic thresholding capabilities are provided.
   - Elaboration of a novel ***texts classification metric*** and process; the texts classification metric is based and utilizes the highly discriminate terms determined by 4.i.

The theoretical basis of the thesis is mainly implemented by operations 2, 3, and 4, which are covered and presented in this section. Operations in 1 are covered and presented in the next section (section 5).

3.1 Biomedical Texts Collections, Parsing & Gene/Protein Identification

3.1.1 Biomedical Literature & Texts Collection

In order to achieve a large-scale experiment in biomedicine we must first collect text and documents containing the genes we are interested in. PubMed a service of the National Library of Medicine - NLM (http://www.nlm.gov/) includes over 15 million citations for biomedical articles back to the 1950's (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi). These citations are from MEDLINE and additional life science journals. PubMed includes links to many sites providing full text articles and other related resources. Also it provides access to bibliographic information that includes: MEDLINE, old MEDLINE, as well as the out-of-scope citations (e.g., articles on plate tectonics or astrophysics) from certain MEDLINE journals; primarily general science and chemistry journals - for which the life sciences articles are indexed for MEDLINE; citations that precede the date that a journal was selected for MEDLINE indexing; and some additional life science journals that submit full text to PubMed Central and receive a qualitative review by NLM. We found PubMed and MEDLINE as the most reliable sources of biomedical literature, also utilized by most researched in the fields of medicine and biology.

The respective biomedical texts collections of interest, to focus the MineBioText inquiries and respective knowledge discovery operations, comes from queries post on PubMed. For example with the query:

```
"breast cancer" AND ("gene expression" OR "microarray")
```

we may focus on references dealing with *breast cancer from a gene-expression profiling perspective*, with "microarray" used and interpreted as a synonym of gene-expression.

MineBioText offers services for storing the retrieved biomedical text references - actually PubMed abstracts accompanied with their full citation details (i.e., title, authors, affilication etc). Here we have to note that, at the current MineBioText implementation, the utilization of PubMed, and the formation of respective biomedical texts collections is performed off-line. It is in our future R&D plans to offer the respective functionality from within MineBioText (see last section of this thesis).

3.1.2 Genes/Proteins Terminology

As mentioned in sections 2.3.1 and 2.3.2 there is a great need for consistency in the description and definition of genes and proteins because of the variations in the available and used *terminologies*.

Gene ontology (GO; www.geneontology.org) is an effort toward the developing of a ***controlled vocabulary*** applicable to all organisms. The project began as a collaboration between three model organism databases: ***FlyBase*** (Drosophila), the ***Saccharomyces Genome Database*** (SGD) and the ***Mouse Genome Database*** (MGD) in 1998. Since then many databases has been included. ***GO terms*** are organized in structures called *directed acyclic graphs* (DAGs), which differ from hierarchies in that a 'child' (i.e., more *specialized term*) can have many 'parents' (i.e., less specialized terms). Not only a structured vocabulary for genome annotation is build and provided but services (yet preliminary) in an attempt to make *mapping* and *translation tables* between catalogs and ***GO***, although these mapping are only used as a guide.

In ***MineBioText***, we also use gene/protein terms from ***PubMed*** and from ***Ensembl*** (http://www.ensembl.org/index.html) genomics resource. *Ensembl* human gene/protein ***identifiers*** are used as our ***primary*** reference identifier for genes and proteins. All other identifiers for different species are also provided and utilized but will be searched for in the given texts. Substantially gene terms are stored in the structure described below.

3.1.3 Storing Terms: An Intelligent Repository for Gene/Protein References

The informatics community faces several numbers of problems in the structure and organization of data in the biomedical domain. A considerable problem of *ambiguity* arises because of *inexact mapping* of knowledge and linking of variant forms between external representations, machine executable formats and biomedical databases. A further corruption occurs when users try to "fill the gaps" with their own interpretations and terminology. A common problem met is the appearance of more that one terms for a single object, which brings a great need for recognition and linking between the different available biomedical nomenclatures (Hirschman *et al.,* 2002; Ananiadou *et al.,* 2005).

More specifically and for text mining in the biomedical domain, the main problem raised relates to the huge text collections to be manipulated. During pre-processing and parsing of the input abstracts the whole set of gene/protein, terms should be retrieved, for each step of the parsing. A minimal set of abstracts may be estimated to contain about 5×10^5 abstracts with about 200 words each. If we utilize a database for storing terms then, at every step of the search algorithm we would need about 10×10^7 queries to it!

3.1.3.1 An Efficient Data-Structure to Parse-for and Store terms

A primary research task for this thesis is the organization, 'amalgamation' and utilization of different gene/protein terminologies. Towards this direction, and in order to cope with the intrinsic to this task high computational cots we rely and employ the ***Trie***[10] – a special data structure for the storage and retrieval of gene/protein terms.

Tries were introduced in the 1960's by E.Fredkin. As it is stated: "*Trie memory is a way of storing and retrieving information that consists of item-term pairs - information conventionally stored in unordered lists, ordered lists. The main advantages of a TRIE memory over the other memory plans are shorter access time, greater case of addition and updating, greater convenience in handling arguments of diverse lengths, and the ability to take advantage of redundancies in the information stored. The main disadvantage is relative inefficiency in using storage space, but this inefficiency is not great when store is large."* (Fredkin, 1960)

The utilised ***Trie*** data structure was appropriately tuned and customised to meet the ***MineBioText*** needs in terms of space and time complexity requirements. In this context, we rely on recent results reported in (Bodon and Ronyai, 2003l Bodon, 2003; 2004) about the *Trie* data-structure, and especially the finding that *Trie*-like structures are efficient for the construction of frequent *item-sets*, i.e, frequent attribute or, feature or, term (for text mining) combinations. The original problem was stated in (Bodon and Ronyai, 2003) in the context of *association rules mining*, where the main

[10] The term Trie comes from the word "reTRIEval".

step (and the most time- and space-consuming) task is to find frequent occurring item-sets.

A *Trie* is a tree structure in which each transition corresponds to a character of the *keys* in the presented key set *K*. A path from the root state to a leaf state corresponds to a key. A special *endmarker* # is added at the end of each key to distinguish keys such as 'the' and 'then'. Henceforth the state number of the *Trie* is represented as a positive number; the root state is represented as 1. The goto function *g* is introduced to represent a transition labeled character *a* from state *r* to state *t*, and is represented as *g(r, a) = t*. The absence of a transition indicates failure (fail). Figure 4, below, gives an example of a traditional *Trie* for the set *K1* = {*illusion, in, inspiration, installation, instrument*} with *#*. Since retrieval on the *Trie* advances character by character, the worst-case time complexity of retrieval is in proportion to the length of key strings, so the *Trie* is a fast retrieval technique. For example, the retrieval of the key '*in#*' in figure 4 is performed by the confirmations of transitions *g(1, 'i') = 2, g(2, 'n') = 11, g(11, '#') = 12*, in sequence. In the *Trie*, common prefixes of keys can be shared, but after making a branch path, common suffixes of keys cannot be shared. As the *Trie* has many states for a large set of keys, it is important to make the *Trie* more compact (Morimoto *et al.,* 1994).

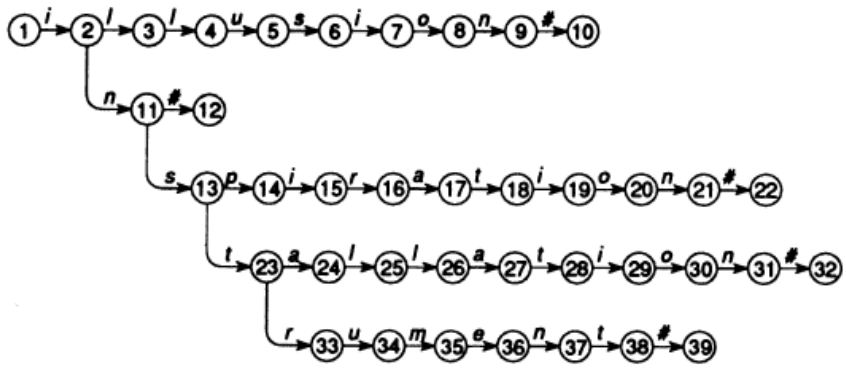


Fig. 4: *The traditional Trie structure*

It was found that the *Trie* structure outperforms approaches based on hash-tree representation approaches. Although hash and B-tree strategies are based on comparisons between keys, a *Trie* structure can make use of their representation as a sequence of digits or alphabetic characters. A *Trie* can search all keys made up from prefixes in an input string without the need to scan the structure more than once. This is so since the Trie advances retrieval character by character, which make up keys. This characteristic makes the *Trie* structure frequently applied to various fields. Examples include the building of dictionaries for natural language processing, searching of reserved words for compilers, the data structure of dynamic hashing tables for database systems and inverted files for text retrieval (Fredkin, 1959; Aho *et al.*, 1983).

However, *Tries* have the disadvantage of having many states for a large set of keys. A DAWG (directed acyclic word graph) is a well-known method of reducing the size of tries (Aho *et al.,* 1983). A DAWG can merge transitions associated with suffixes of the traditional trie structures to reduce the total number of states, but it cannot determine the record for a key correctly. Thus, the applications have been restricted to areas

where record information is not needed. Other schemes for reducing the size of tries were proposed in (Maly, 1976; Ai-Suwaiyel, 1984; Dundas, 1991), but their areas of application are restricted to static keys. Although the approach described in Dundas, 1991) is suited to dynamic key sets, it decreases the size of the resulting *Trie* slightly as one of the side-effects.

- **Towards efficiency: The Revised Trie Structure.** We have employed and implemented a double-chained *Trie* where, the edges of a node are stored in a double connected list. An example of the way this structure is utilised for terms' storage is illustrated in figure 5.

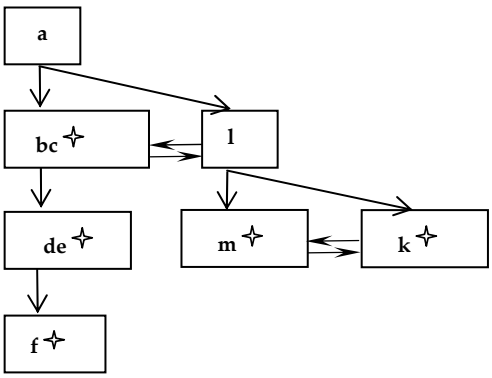


Fig. 5: An example of how terms are stored. *The tree contains terms abc, abcde, abcdef, alm, and alk. Each node contains a unique symbol or group. Each leaf of the tree which contains '✧', is a complete term. A term is composed of all the above ancestors till the root of the tree. All the other nodes simply represent a common symbol of other terms.*

We utilise this structure in an algorithmic process. The time complexity of the algorithm can be approached assuming a word of *n* letters. The search process will seek for the first letter in all the nodes of the built tree in order to figure out its ancestor. The time complexity of this action depends only on the amount of letters contained in the alphabet, suppose *c*. Similarly, all the letters of a term, that are going to be inserted in the tree, will need *c* steps. The first letter of the term takes *c* steps, as well as the second, the third till the n^{th}; Therefore each term of n letters will take *c*n* steps, resulting into a time complexity of O(c*n)[11]. This algorithmic process is followed in the course of parsing the input texts and the identification of gene/protein terms in them.

3.1.4 Text Parsing and Gene/Protein Identification: An Informal Presentation

3.1.4.1 Removal of Common Words.

The first problem in parsing free text references of biomedical content is the removal of *string patterns* that contain *common words* (i.e., words with no semantically relation with the biomedical domain). An efficient way to cope with this problem is to eliminate pre-specified patterns by using list of *common words*, and employing a look-up approach. A dictionary of English common used words is utilized for this purpose

[11] It can be proven that in an implementation without a double connected list this time can be reduce to n*logC.

(http://wordlist.sourceforge.net/) - a collection of twelve dictionaries of English word lists. The comparison of the words contained in the dictionary of gene terms will reveal the terms to expel. Finally parsing of the genes terms is necessary in order to increase sensitivity and reduce parsing time. Note that gene/protein names and symbols are converted into lowercase; with punctuation marks and others symbols removed. As long as the parsing process searches for *single-terms*, a stemming operation is needless, i.e., potential common words within the text will never be reached following the *Trie*-like algorithmic process presented above. All the stored gene/protein terms are stored with this structure, and common used words will never be reached (by the aforementioned algorithmic parsing process) as they are removed. Note that with the MineBioText system the user may customize (add/remove words and/or phrases) the exclusion common-words dictionary to meet her/his needs.

3.1.4.2 Gene/Protein Localization, Recognition, Registration & Representation

The utilized standard gene/protein terminologies. The process of the localization and recognition of terms utilizes various sources on genes/proteins terminologies and nomenclatures. Primarily the inquiry is based on the combination of the 'Ensemble Gene ID' – which consists the primary gene/protein reference key, as well as other identifiers utilized as standard gene/protein ***synonyms***. These gene/protein synonyms are provided from various related gene/protein nomenclature resources: 'GO Id' and '*GO description*' (GO, 2006), 'HUGO id' (HUGO, 2006), 'OMIM' (OMIM, 2006), 'Uniprot id' (Uniprot, 2006), 'UNIPROT/SWISSPROT' (SwissProt, 2006), and EMBL (EMBL, 2006) – all of these are appropriately incorporated and utilized by the MineBioText system.

Adaptation to specific search needs is enabled by the formation of an appropriately formatted input *domain-file*, where the different terminologies, i.e., gene/protein names and symbols and their synonyms (from the utilized nomenclatures) are specified. Respective operations and services are offered by the MineBioText system. An example entry in the domain-file is shown below.

Ensembl Gene ID	UniProt ID	SwiisProt ID	GO description
ENSG00000006831	*ADR2_HUMAN*	*ADR2_HUMAN*	*fatty acid oxidation*

- '*ENSG00000006831*' is the primary gene/protein key-reference.
- '*ADR_HUMAN*' (or '*ADR2*') are utilized as synonyms for the primary gene/protein reference.
- '*fatty*', '*acid*', '*oxidation*' as respective extra synonyms.

Gene/Protein Identification and GO-descriptions. Initialization of the search and term identification process is done by consulting the input domain-file. The specified terminological references are searched in all the texts of the input collections (also specified in the domain-file). This operation is performed on the basic of the algorithmic Trie-based parsing process presented previously, and it is based on the 'Ensemble Gene ID' as the primary key-reference ID, and the consultation of all other gene/protein references referred in the input domain-file.

- A basic contribution of the work reported in this thesis is the use of the respective GO-***descriptions***, as a means to identify genes/proteins not only by their standard terminological encodings and references (e.g., 'ADR2' in the above example which referrers to the standard UniProt naming and encoding of genes/proteins) but, with reference to their functional category as reported in the respective GO-description. In other words we need find and register respective ***gene/protein lexicographic-***

identifiers – '***gli***', which are *descriptive of* (*relates to)* specific gene/protein terms, (e.g., the word '`fatty`' in the above example). It is a process that in a way extends the genes/proteins terms with extra synonyms. Here we are faced with the problem of words (or, roots of words) contained in many gene/protein descriptions. To cope with this problem we follow an intelligent parsing operation of the GO-description (small) texts in order to assess and measure the ***degree*** of ***gli relevance*** – ***gli_r*** of the description-words with respect to the corresponding genes/proteins. We cope with two cases:

- If the *gli* is found in just one single description then its *gli_r* is set to '***1***';
- If the *gli* is located in more than one description, its *gli_r* is computed by the sum of all the previous calculated weight values for it ($SUM_{other\text{-}gli\_r}$), plus 1 divided by the total number of descriptions where the *gli* is found ($Description_{gli}$). Note that we may result into lgi_r values that are greater than *1*. In a more formal setting:

$$gli\_r = SUM_{other\_gli\_r} + (1 \,/\, Description_{gli}).$$

The parsing process and the above formula present a form of *term normalization* (a common approach in information retrieval endeavors). Moreover, during parsing GO-description texts we have to remove the most common *morphological* and *inflexional endings* from the respective words. This requires a stemming operation on particular words (e.g., stop-list), that also checks for uppercase words and covert them to lowercase, an acronyms' recognizer etc. In this context we rely and utilize the *Porter-parser* - appropriately customized and implemented within the MineBioText system (Porter, 1980; 1997).

Binary Vector-based representation of texts. Initially we employ a (binary; i.e., '1' or, '0') *vector-based* approach to register the occurrences of every term in the input texts (as described in section 2.2.1). Whenever a term is met in the text as an 'Ensemble ID' or, other terminological reference or 'id' (from the utilized input terminologies) the respective position in the binary vector is set '*1*'. This value represents the ***significance*** of the term - its ***descriptive power*** for the respective text it is identified. The value is set to '*1*' because of the *uniqueness* of the utilized gene/protein identifiers (HUGO, OMIM, UniProt, Swiss-Prot ID etc) - there is a univocally 1-1 correspondence to the primary Ensembl ID for each gene/protein.

- ***Weighted Vector-based representation***. The most interesting case, also a main contribution of the work reported here, is the assignment of *weights* to the words found in genes/proteins GO-descriptions (identified by the operation described above). In this case, we deviate from the binary vector-based representation and move towards a more 'vague' assessment and registration of located genes/proteins. During parsing (of a given biomedical text-reference), the located words should be tested for their *relevance* with respective genes/proteins. For this purpose, a special process is devised and implemented. It copes with two cases:
  - *The located word matches a lexicographic-identifier (gli)*: its weigh is set equal to the respective (computed and recorded) *gli_r* (described previously); note that in this case, *gli_r* values may be greater that 1.
  - The located word matches a gene/protein term: its weight value is assigned to the largest weight value from all other located words in the text (also taking into account the previous case).

3.1.5 Formal Definitions

The input to MineBioText consist of several parts – all specified in the aforementioned input domain-file: The abstracts (text-references) set that comprise the main input; the gene/protein identifiers from the utilized gene/protein terminological resources; the GO-descriptions of these genes/proteins; a set of English common used words; and a stop-, upper-to-lower-case conversions, acronyms-to-remove etc (for the Porter parsing operation). We introduce different annotations for each set:

- **Definition 1:** *Abstract*. We define an abstract the set of all a_i that belong to A and a_i a subset of Λ where,

$$\Lambda \text{ is a potential set of words,}$$
$$\forall a_i \in A \text{ where } a_i \subset \Lambda,$$
$$a_i = \{\lambda_{i1}, \ldots, \lambda_{iki}\} \text{ and } k_i = |a_i| \text{ the size of } a_i.$$

- **Definition 2:** *Set-of-Abstracts*. Assume $A = \{a_i \ldots a_n\}$ as the finite *set of the abstracts*. Each a_i denotes an abstract from the initial set. The total number of abstracts is denoted as $|A|$.
- **Definition 3:** *Set-of-Terminology-Terms*. We denote the set of *all terms from the utilized gene/protein terminologies/nomenclatures* as T_{nom}; with different instantiations for the each of the corresponding different gene/protein ontologies, e.g., T_{HUGO} for HUGO, $T_{UniProt}$ for UniProt, $T_{SwissProt}$ for SwissProt etc. A single gene/protein terms is denoted with t_x.
- **Definition 4:** *Set-of-All-Terms*. We define the set of all terms - except for ensemble identifiers- as T_X, we can conclude that $T_X = T_{HUGO} \cup T_{EMBL}$.
- **Definition 5:** *Set-of-Ensembl-Terms*. We denote the set of the Ensemble gene/protein identifiers as $S = \{s_1 \ldots s_m\}$, with a single gene/protein identifier denoted as s_i; the size of S is denoted as $|S|$.
- **Definition 6:** *Description (lgi)*. A *description* t_{D_i} is a set of words, and is defined as:

$$\forall\, t_D \in T_D,\ t_D \subset \Lambda$$
$$t_D = \{\lambda_1 \ldots \lambda_N\},\ N = |t_d|$$

- **Definition 7:** Set-of-*Descriptions (lgi)*. A set of descriptions T_D is the set of all t_{D_i} defined as: $t_D \in T_D$ *where* t_D *is a set of words* Λ^K
- **Definition 8:** Set-of-*Common-Exclusion-Words – the Words List*. Is denoted with L, L = *{the set of all English words in the input common-words file}*.

Initially all the gene/protein terms T_D and S are stored. Assume that during parsing, and for each gene/protein contained and located in the set T_X, as well as the terms contained in the descriptions T_D, the Ensembl Gene Identifier is located, selected and characterized as *significant*. We have to define formally these concepts.

- **Definition 9:** *From T to S*. The significance of a gene/protein term is defined as a function: $\forall t_x \in T_x, \exists s_{t_x} \in S$ *such as* $\exists T \rightarrow S$.
- **Definition 10:** Correspondence between S_{T_D} and S. For each GO-description there is a set of significant identifiers (i.e., gli) $S_{t\,D}$ that belongs to S.

$$\forall t_D \in T_D \text{ corresponds an } S_{T_D} \in S$$

3.1.6 Parsing and Trie-structure Utilization

Each S_{t_x} represents a unique identifier and reference to a gene/protein as specified by the utilized standard Ensembl (as the basic gene/protein key-reference) as well as from the other standard gene/protein terminological references. So, definitions 8 and 9 above, establishes a correspondence between a unique Ensembl identifier and identifiers from the other standard gene/protein terminological references (i.e., HUGO, UniProt, Swissprot, EMBL etc). This is necessary because of the need for a 'primary id' which we can use every time we want to refer to a specific gene/protein. Every time we detect a gene/protein term in the abstracts its location in the *Trie* tree data-structure should be found (see section 4.1.3.1). Moreover, each gene/protein standard term identifier has one single and unique position in the vector used to represent a biomedical text-reference but the s_{tx} identifier and their synonyms t_x, possess different positions in the *Trie*-tree. So, we are faced with the problems of storing sets S_x, T_x, and T_D how to link the positions of of s_x , t_x and t_D nodes to the primary reference node.

For the following you may refer to illustrative examples shown in figures 6 and 7. Consider that the input files that contain the initial names of the gene/protein terms are stored in pairs (s_i, t_x). The first name -s_i- is the Ensemble identifier. The second one, t_x, is any other, corresponding to term, synonym-reference (coming from the other gene/protein terminologies) and appears in the Trie-tree. For each new coming gene/protein terms pair, in order to check whether it is already present in the *Trie*-tree or not, we examine if s_i is already stored: if yes, we localize and store –temporarily- the position of the term; if no, we first create a new entry, only for the s_i.

The next step is to insert the second name, t_x, of the specific gene/protein term and create a *link* to the node that contains its corresponding *significant* Ensemble identifier. In this ways, we secure that there is going to be always a connection between each gene/protein term and its corresponding *significant* identifiers.

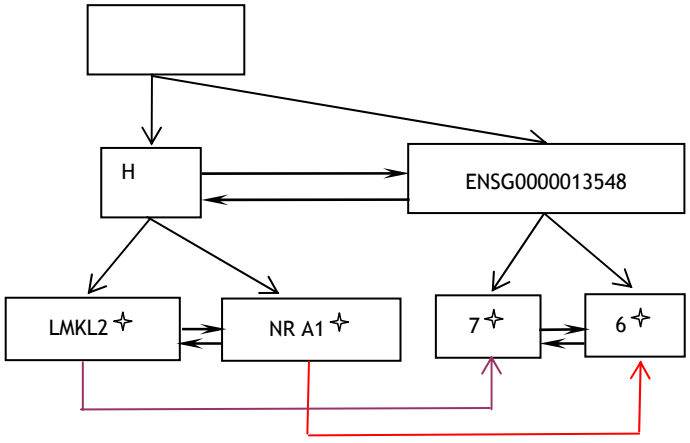


Fig. 6: How a new term is inserted in the TRIE: *Assume that the trie holds ENSG00000135487 and HLMKL2; the ENSG00000135486 and HNRPA1 are to be inserted. The figure shows the state after the insertion where the red line indicates the link that connects the new entered gene terms.*

The second tree holds information about T_D. Each node contains a word that belongs to a description, t_d, of a particular gene/protein term. For each node there is link to a double connected list which holds the Ensembl identifier of the gene/protein to which the word's description belongs. Thereby we can ensure that when a word is met in an

abstract, there is a flexible way –through the connected list- to find t_D that corresponds to the description where the word belongs. Later, in order to assign the weight values to each gene/protein term that indicates its descriptive-power to an abstract, we must be aware whereas if it was met in an abstract or a description. The computed weight values also depend on the number of gene/protein terms that their descriptions contain the specific term.

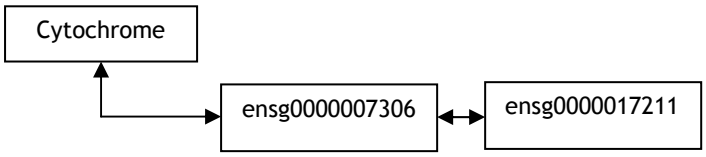


Fig. 7: How identifiers of a common gene/protein term are sited in the TRIE: *Assume gene/protein terms ensg00000073067 and ensg00000172115 and the corresponding descriptions Cytochrome P450 2W1, Cytochrome c. Cytochrome, the common word of the descriptions of the two gene terms is stored as shown.*

The third instance of the *Trie*-structure holds a stop-word list (http://wordlist.sourceforge.net/) used for stemming, containing words in singular and plural form, in all form that they take in different tenses. Before we check whether a term is contained in the abstracts, we examine if it is a stop word. If it is contained in the dictionary, we assume that it is not a gene/protein term name, the term is ignored and the search is blocked.

3.1.7 Computing Gene/protein Weight Values

The formal presentation of the algorithm for the computation and assignment of weight values to gene/protein terms is presented in figure 8, below. An illustrative example of a weight assignment is deployed in Fig.9.

Assume word $k \in a_i$ *where* $a_i \in A$ *(A: set of abstracts)*

If $(k \notin L)$ *and*

If $\exists\, t_x \in T_x$ *such as* $t_x = k$ *Assuming Def.(10)* $\Rightarrow Vt_x = 1$

else if $\exists_1\, t_D \in T_D$

where $T_D = \{ t_D \in T_D:\ \lambda \in t_D \}$, *such that* $\lambda = k$ $\Rightarrow Vt_D = 1$

Def.(11) $\Rightarrow$ *for each* $t_D \in T_D$ *corresponds* $S_{T_D} \in S$

Else if $\exists\, t_{D1}, t_{D2}, \ldots, t_{DN}$ *where* $t_{Di} \in T_D$ *(N: number of descriptions)* $\Rightarrow$

for the significant identifiers that correspond to $t_{D1},\ldots, t_{DN}$: $St_{D1},\ldots St_{DN}$ $\Rightarrow$ $VSt_D = \min(1, VSt_D + 1/n)$

Fig. 8: Assigning Term Weights. *Assume a word a_i that belongs to an abstract. If the word belongs to the No-Word-List (L), the search should terminate. Else if it belongs to the set of terms from HUGO/EMBL references (T_x), we assign the value $V_{tX} = 1$ to the corresponding significant term of the term. If the word is located in a description, we should first figure out whether it belongs in a single description or not; if a_i belongs to **n** descriptions, for each one we first locate the corresponding **significant identifiers**; for each one we add to the previous assigned weight value V_{TX} the 1/n or 1 if the sum is greater than 1.*

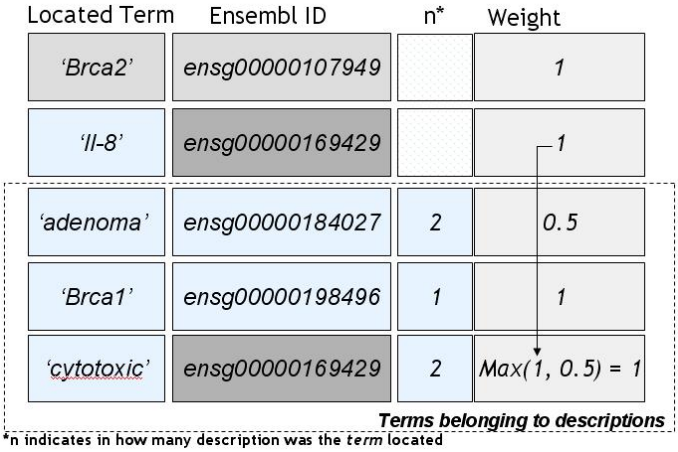


Fig. 9: Assigning Term Weights: *'Brca2'and 'Íl-8' were found as gene terms; 'ádenoma', 'cytotoxic' were detected as part of the 'eng00000184027' and 'ensg00000169429' descriptions' respectively. The assigned values are estimated according to the algorithm described above. Note that for the 'ensg00000169429' the weight is not 1.5.*

3.1.8 Gene/Protein Associations: The MIM measure

In order to estimate the ***strength of*** the ***associations*** between gene/protein terms a well established scoring scheme is used in order to measure how *informative* the associations are. Assigning associations between objects by just locating co-occurrences in literature has been widely used in biology and medicine. However, associations detected lack of specialization and bring out false positives. The significance of reappearance of terms is evaluated using MIM, which originally is based on Shannon's Entropy theory (Shannon, 1948).

Mutual Information Measure – **MIM** has been used to quantify dependencies between variables, including co-occurring terms in text (Dunning, 1993; Conrad and Utt, 1994; Stapley and Bennoit, 2003). Previous work has shown that it is possible to identify implicit relationships by ranking inferred relationships and preferentially examining those at the top list (Wren *et al.*, 2004). In a recent work, an extended version of MIM is introduced and applied on biomedical literature mining (Wren, 2004). Although the co-occurrence of terms in abstracts marks a valid relationship, it is considerable that many co-occurrences of terms within literature do not always mean a biological association. This emerges the need for locating *more informative interconnections between gene/protein terms*.

$$\mathrm{MIM}(i, j) = \sum_{k=0,1} P_{k_i} \times P_{k_j} \times \log \frac{P_{k_i k_j}}{P_k[i] \times P_k[j]}$$

MIM(i,j) between terms *i* and *j*, is computes given: a list of terms and a list of abstracts, and takes in consideration all the terms' occurrence possibilities:

$$MIM\ [i,j] = p_{0_i 0_j} \times log\ \frac{p_{0_i 0_j}}{p_0[i] \times p_0[j]} + p_{0_i 1_j} \times log\ \frac{p_{0_i 1_j}}{p_0[i] \times p_1[j]} + p_{1_i 0_j} \times log\ \frac{p_{1_i 0_j}}{p_1[i] \times p_0[j]} + p_{1_i 1_j} \times log\ \frac{p_{1_i 1_j}}{p_1[i] \times p_1[j]} +$$

Fig. 10: Mutual Information Measure (MIM) for terms i and j. *Where* $p_{0_i}p_{0_j}$ *is the number of the texts that don't contain the term j neither i;* $p_{0_i}p_{1_j}$ *is the number of the texts that don't contain the term i and contain the term j;* $p_{1_i}p_{0_j}$ *is the number of the texts that contain the term i and don't contain the term j;* $p_{1_i}p_{1_j}$ *is the number of the texts that contain the term i and j;* $p_0[i]$ *is the percentage of non occurrence of the term i in the text references;* $p_1[i]$ *is the percentage of occurrence of the term i in the text references;* $p_0[j]$ *is the percentage of non occurrence of the term j in the text references and* $p_1[j]$ *is the percentage of occurrence of the term j in the text references.*

The formula presented in figure 10, below, estimates the co-occurrences between (gene/protein terms) with reference to a given collection of abstracts[12]. The computed MIMs are stored in a file to be used for the construction of gene/protein correlation (or, association) network (next sub-section).

3.1.9 Construction of Gene/Protein Correlations/Associations Network

In this subsection, we present the construction of a *genes/proteins Correlation Network* – ***gpCN***. Here we have to note that literature based inference of a potential 'gene/protein-to-gene/protein' lacks a clear-cut semantic meaning, at least with respect to a potential 'biological interaction' or, 'biological regulation' between them. So, we prefer the use of the word '*correlation*' to refer to potential gene/protein relations because expresses a '*less-causal*' concept, as contrast to 'association' which underlies a potential 'causal' relation. This uncovers the real utility of biomedical literature mining: *evidence-based support to biomedical scientists based on 'interesting hints' to focus and target their research.*

The gpCN construction within ***MineBioText*** is based on the utilization of the computed MIM values between gene/protein terms. The whole process follows three steps:

i. The following input information is provided: the list of terms; the list of abstracts; and a user specified *percentage MIM threshold for the gene/protein terms with top ranked MIM values*. The last input specification is provided in order to *filter-out the gene/protein MIM values that are below the specified MIM threshold*. This is done in order to keep the *most-informative* gene/protein correlations – the user may decide to keep all gene/protein correlations with a threshold value of '0'.

[12] The stored input MIM file is linked with a specific list of Abstracts → *MIM file(Terms, Abstracts)*

ii. After filtering-out, the remaining gene/protein correlations are also examined for their ***strength***. This operation is performed with a careful ***discretization*** of the corresponding MIM values into three correlation strength levels, with the following 'natural' interpretation: ***strong***, ***medium*** and ***weak***. The discretzation of MIM values is based on a method reported in (Lopez *et al.*, 2000), and utilised also in (Potamias *et al.*, 2004). The descritization process has as follows.

- Assume (in a general setting) that a MIM value may be assigned to an (ordered) set of nominal values; assume *n* such values. In the case of *n=3*, value '*1*' is interpreted as of '*weak*', value '*2*' as '*medium*', and value '*3*' as '*strong*'. Define,

$$w_i = \frac{max(MIM) - min(MIM)}{n}$$

where, *min(MIM)*, and *max(MIM)* are the minimum and maximum MIM values, respectively. A MIM value, MIM_v, is descritised to a nominal value, MIM_{nom}, using the following formula:

$$MIM_{nom} = \begin{cases} n & if\ MIM_v = max(MIM) \\ \left\lfloor \frac{MIM_v - min(MIM)}{w_i} \right\rfloor + 1 & else \end{cases}$$

where, $\lfloor fraction \rfloor$ is the integer part of the fraction.

3.1.10 Abstracts/Texts Categorization & Classification

We introduce a novel approach for *text categorization* and text *class/category-prediction* based on term frequency and *supervised learning* techniques.

Class Prediction is achieved by a ***similarity matching formula*** that compares the rankings of weight values – calculated according to the values from algorithm of the assignment of the Term Weights (Fig. 8) attached to an unclassified/test set of abstracts with respect to the rankings of the classified/train set of abstracts. The approach is tested on the *clinico-genomic* based classes, by measuring and ranking the aggregate presence of terms in abstracts containing this word, and classifying the corresponding documents.

Training-phase. Assume a *two-class* (categories) problem, i.e., the task is to classify a set of documents (biomedical abstracts) into two (pre-specified) categories. Let us refer to these classes as '*POS*' and '*NEG*' (for positive and negative, respectively). The process may be generalized to cover *multi-class* cases. Therefore, two sets of abstracts are available, and the documents in each of them are pre-assigned to one of the two classes; this is the *training set*. Training is performed on each of the class-specific set of abstracts from the training set, i.e., MineBioText is called to run two times, and follows two steps.

i. The corresponding abstracts are parsed and for each abstract, the corresponding vector-based representation of it is formed (as presented in section 4.1.4 and 4.1.6). Then MIM computation is performed and the strength of each gene/protein correlation is computed, with all corresponding values being discretized (as described in the previous section). All this information is stored in a file. Note: This file can be called at any time and filtered in case the user wants to specify a different threshold, i.e., keep more- or, less-strong correlations.

ii. The strength values for the significant identifiers is the sum of the weight values (as estimated from the algorithm of Fig. 8 in section 4.1.7), of all the terms identified – computed by the formula, below.

$$s_{A_{train}} = \sum_{i=1\ldots L} Vt_{A_{train_i}}$$

Fig. 11: Calculation of Strength Values. *Assume the set of train abstracts as A_{train}, the number of gene terms located in A_{train} as L, the set of gene terms located in A_{train} as $t_{A_{TRAIN_1}} \ldots t_{A_{TRAIN_L}}$,and their corresponding Ensembl IDs as $S_{A_{TRAIN_1}} \ldots S_{A_{TRAIN_L}}$. If $V_{t_{A_{TRAIN_1}}} \ldots V_{t_{A_{TRAIN_L}}}$ are the assigned values of the terms that were located (from algorithm of Fig.8) then for the i^{th} significant term, $S_{A_{TRAIN_i}}$ located in A_{train} the corresponding strength value of $s_{A_{train}}$ is given above.*

The calculated strength values given by the formula of Fig.11 for all (training) abstracts are sorted and stored. For the estimation of the ***similarity formula*** we will also need to keep the maximum value of the corresponding terms weights (of Algorithm of Fig. 8); these actually consists the *train-results-files*.

Testing-phase. As for the training case, we assume the availability of two class-specific sets of abstracts. MineBioText is called to run on each of these sets separately. For each set, the identified gene/protein terms and their weights are recorded and compose the *train-results-files,* as described above.

For each term identified in the set of test abstracts, we check its occurrence in the saved *train-results-files* as well as its corresponding class-specific ***rank***, i.e., its position in the ordered (by their training strengths) lists of the corresponding file. An illustrative example is shown in figure 12, below. In this example, note that term 't_x' is identified in the test abstracts but <u>not</u> in the training abstracts. In this case, its rank is set equal to *0*. So, we have different ranks for the 'POS', $rank_{POS}(t)$, and for the 'NEG', $rank_{NEG}(t)$ classes, respectively. The formula below computes the *strength,* $strength_{TEST}(t)$ *of a term t,* identified in a test-abstract, with reference to its weight, $weight(t)$, and its corresponding and class-specific strengths, $strength_{POS}(t)$ and $strength_{NEG}(t)$.

$$strength_{TEST}(t) = \frac{rank_{TRAIN\_POS}(t)}{count_{TRAIN\_POS}} - \frac{rank_{TRAIN\_NEG}(t)}{count_{TRAIN\_NEG}} \times weight_{TRAIN} \times \left| \frac{strength_{POS}(t)}{max(strength_{POS})} - \frac{strength_{NEG}(t)}{max(strength_{NEG})} \right|$$

Equation 1. The similarity formula used for classification*. The formula is been applied to each gene term that is found in an unclassified/test abstract. Count is the total number of all the gene terms.*

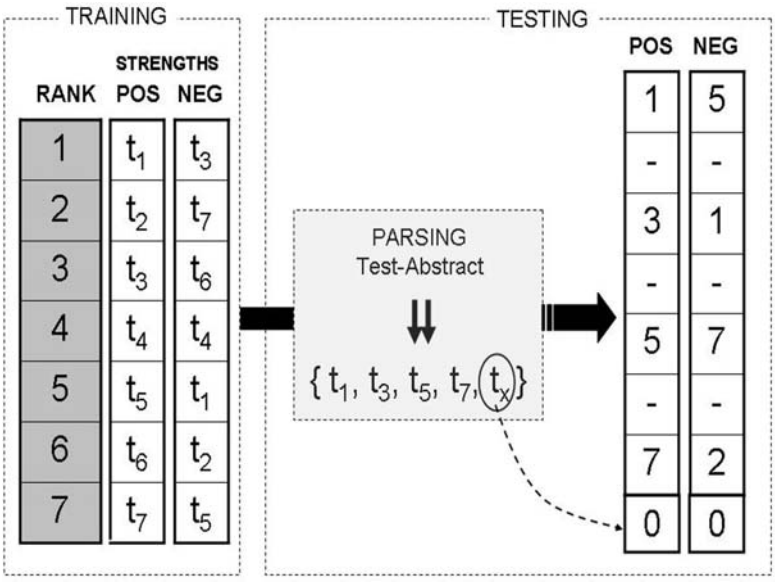


Fig. 12: *Ranking terms identified in the test-abstracts in order to calculate their test-abstracts' strengths.*

For each abstract from the test file, a sum of values estimated by the *similarity formula* of all the located gene terms is compared according to the comparison with zero it is assigned to a class; if the assigned value is greater than 0 it is assigned to ***class A*** else to ***class B***.

If the value of the similarity function was below 0.5 the entry was assigned as unclassified abstract else:

- *If is predicted as A class and actually belongs to class A, we assign it as a true positive.*
- *If is predicted as A class and actually belongs to class B, we assign it as a false positive.*
- *If it is predicted as B class and actually belongs to class A, we assign it as a false negative.*
- *If it is predicted as B class and actually belongs to class B, we assign it as a true negative.*

The accuracy for the prediction of the class A are the corrected predicted as class A divided by the total predictions for class A and class B:

$$\frac{True\ Positives}{False\ Positives + TruePositives} \times 100$$

$$\frac{True\ Negatives}{False\ Negatives + TrueNegatives} \times 100$$

Equation 2, 3*. The prediction accuracy formulas*

The total accuracy of the correct predictions is the correct predictions divided by all prediction made[13]:

$$\frac{True\ Positives + True\ Negatives}{True\ Positives + False\ Positives + False\ Negatives + True\ Negatives} \times 100$$

$$\frac{False\ Positives + False\ Negatives}{True\ Positives + False\ Positives + False\ Negatives + True\ Negatives} \times 100$$

Equation 4, 5*. The total accuracy for the correct and false predictions.*

[13] It has to be noted that all the threshold values mentioned above have been derived empirically to optimize the prediction accuracy rather than reflecting some theoretical model or consideration.

4. MineBioText in Action

In general, the tasks tackled by the ***MineBioText*** system are listed in the following shaded box.

MineBioText Architecture:
Tasks/Sub-Tasks tackled

- **TASK 1** Collect literature and Gene Terminology from PubMed and Ensembl respectively.
- **TASK 2** Post Processing of Data includes:
  - **TASK 2.1** stemming and removal of common used words and patterns
  - **TASK 2.2** locate of gene terms in the abstracts set
  - **TASK 2.3** assignments of weight values to the extracted terms according to the appearance frequency
- **TASK 3** Gene Associations Network Extraction:
  - **TASK 3.1** We estimate term-hoods including occurrences as well as co-occurrences of the Gene Terms in the corpus through the Mutual Information Measure formula based on the weight values estimated in task 2.3.
  - **TASK 3.2** Discretization of the extracted MIM values gives three levels of strength for the identified associations between gene terms. (Strong, Medium and Weak)
  - **TASK 3.3** The gene association's network is extracted through the estimated strength level of the terms, and visualized by TULIP graph tool.
- **TASK 4** Class Prediction through statistical based learning:
  - **TASK 4.1** Using weight values concerning the Term occurrence, calculated in task 2.3, we extract strength values for the gene terms located in the training set of literature.
  - **TASK 4.2** The estimated weigh and strength values are used by a similarity scoring scheme in order to classify the documents.

4.1 *MineBioText* General Architecture

Initially a corpus of data including literature and gene terminology is collected from MEDLINE and PubMed respectively. The post processing of data includes the parsing of the corpus of abstracts and the set of Gene Terms. Unsupervised learning includes the extraction of the occurrences of terms from the abstracts and the estimation of the strength of the associations between them; Mutual Information Measure scoring scheme is used in order to measure how informative associations are by locating co-occurrences of the terms. The next steps include the graphical visualization of the interrelations between the terms, the discretization and revision of the calculated MIM values according to an input threshold. In order to predict the classes of an unclassified set of documents, we propose supervised learning. In supervised learning after the post processing of the input terms and texts, two sets of terms are extracted in order to be used in the prediction of an the unclassified set of abstracts. (The architecture of the system is shown in Fig. 13 and 14).

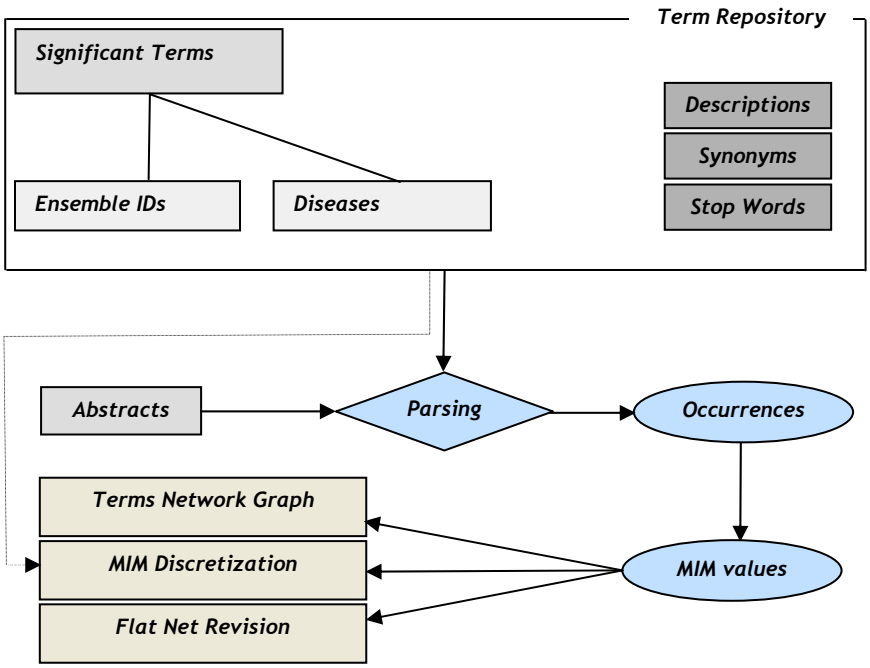


Fig. 13: System Architecture 1, Unsupervised learning. *Input Terms and Abstracts are passed through the Post Processing phase (Parsing). The output graph visualization includes the gene terms or the diseases, according to the significant Term Identifier.*

4.2 Building Gene (association) networks with *MineBioText*

After the post processing of the input is accomplished, the next step is the creation of knowledge (association) network. The intention is to identify the terms sharing implicit associations. The relationships may include binding interactions between the connected objects or biological influence or activations that may an object cause to the other. Each object present to the network will contain link to other information that can provide the system for the specific object. Links are directed and labeled; thus, a network is a directed graph.

The approaches for computing term associations are divided in two categories; those estimating term relationships directly from the co-occurrence frequency and those inferring to term associations from the relevance information through feedback. In the first approach, the semantic relationships are computed from the frequency of the co-occurrences of terms in different documents. The methods are based on the hypothesis that if two or more terms are met in many documents, they are possibly semantically related. (Spark Jones, 1971; van Rijsbergen, 1979; Salton 1989).

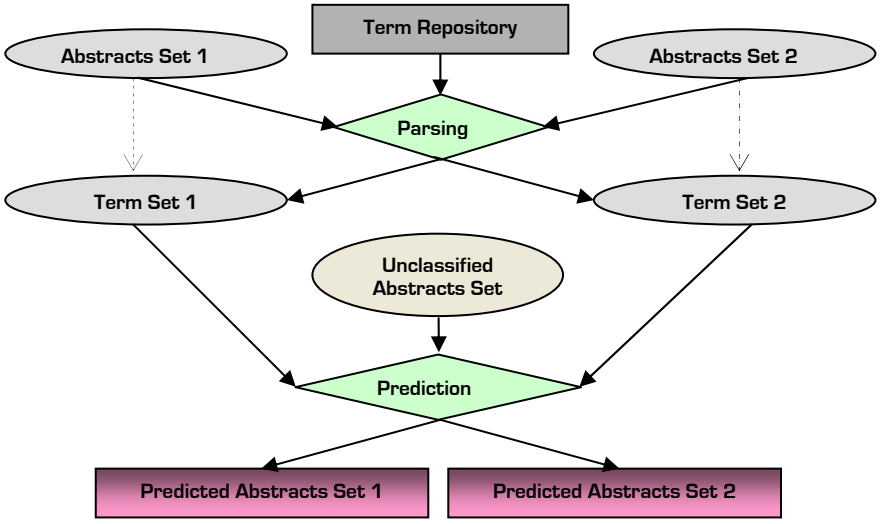


Fig. 14: System Architecture 2, Supervised learning. *Input Terms and Abstracts are passed through the Post Processing phase (Parsing).The output except from the predicted classes of the abstracts, includes the accuracy of the prediction.*

4.2.1 MIM Computation

As described in chapter 3.3 in order to estimate the strength of the associations between the gene terms the formula of MIM has been used. The output of the process is a MIM **file** with the following format:

[i]	level[1]	[j]	level[j]	MIM[i][j]	null
..	....	....	....	...	...
#Terms with upper10% MIM					
#Terms with upper15% MIM					
#Terms with upper20% MIM					
....					
#Terms with upper95% MIM					
#All Terms (100%)					

Fig. 15: The first output file from the computation of MIM. *The first column indicates the first term; the third indicates the second term; and fifth column shows the calculated MIM for the pair of terms. In this phase of MIM, calculation level is -1 and the last column that indicates strengthens of the MIM value, is null. The specific file is sorted according to the MIM values. Terms with upper x% MIM are the x% of the top ranked pairs of terms of the sorted list.*

4.2.2 Construction of Flat Gene Terms Network

After the computation of the MIM values (chapter 3.3) a network is constructed (Fig. 25) indicating the correlations between the gene terms and the potential disease terms located in the collection of input abstracts. This new network actually a filtered MIM file- is generated and stored. Graph visualization software uses specific filtered output in order to make the network (Fig. 25).

[i]	level[1]	[j]	level[j]	MIM[i][j]	S/M/W
..	....	....	....	...	...
#Terms with 3 as discretised value of MIM with 'S'trong interconnection					
#Terms with 2 as discretised value of MIM with 'M'edium interconnection					
#Terms with 1 as discretised value of MIM with 'W'eak interconnection					
#All Terms (in this file)					

Fig. 16: The output MIM file format from the revision phase. *The first and the third columns indicate the pair of terms. The fifth column shows the revised MIM value for the pair of terms. Still all level[i] are '-1'. The format is the same as in fig.12, except that now instead of 'null' one of 'S', 'M', 'W' indicator is put at the last column.*

4.2.3 Revision of Flat Gene Terms Network

After the initial computation the MIM values, given the list of gene terms, the set of input Abstracts, and a user specified network the revision of the network is accomplished.

[i]	level[i]	[j]	level[j]	-1	S/M/W/	null
..	....	....	....	...		

Fig. 17: MIM file format. *Still all level[i] are '-1'. MIM values are now hidden from the user and indicated as '-1' (fifth column). There are 'S", "M', 'W' or, 'null' indications at the last column. The power of the terms' interconnection is also hidden here.*

The abstracts are parsed and the MIM values of all given terms are computed. Based on the computed MIM values, the input network is revised. MIM values are assigned to the original '-1' values and discretized (3.5.2) so that levels *Strong*, *Medium* and *Weak* interconnections are assessed. The extracted graph from the new revised network is constructed (Fig. 25).

	Computed (Replace '-1')			New column			
[i]	level[i]	[j]	level[j]	MIM	S/M/W/	null	S/M/W
..	....		....	....	...		S

Fig. 18: *MIM file format*

4.3 Working with MineBioText: The Graphical User interface

The initial graphical user interface (Fig.19) is divided in two major areas; the input and the processes that are available through it. The input files[14] include the options file, the file that contains the set of abstracts, and the MIM file. The options file gathers all the potential filenames and parameters that will be necessary for the execution of the system; the set of abstracts file contains all the abstracts that will be parsed later and used as train set by the supervised learning algorithm.

The available tasks that the user can introduce, includes the pre-processing phase of the gene terms and the abstracts set; the computation of the Mutual Information Measure for the exported pairs of gene terms that were located in the abstracts through the Abstracts Parsing process; the discretization of the values calculated in the previous step, according to the MIM values file exported and a user specified percentage; the construction of a gene associations network according to the estimated values of MIM given by an input file; the graphic visualization of the associations network; the revision of an existing genes association network according to a given number of levels; and finally the class prediction of a set of unclassified abstracts.

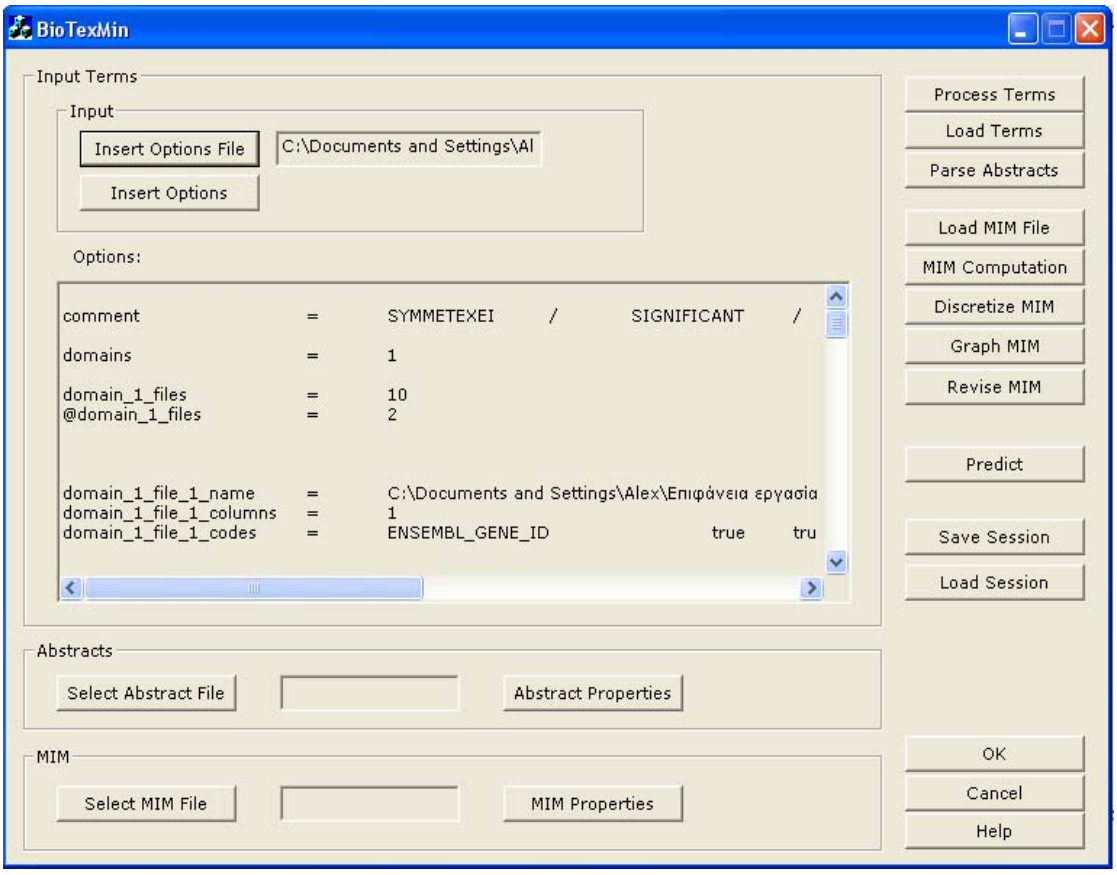


Fig. 19: *The entry MineBioText GUI*

[14] The input is resolved in detail below.

4.3.1 The tasks through the GUI

- ***Determining the options through a file***. Primarily, we should mention that the options can be given either through an options file either by the user through a guided process. In the first case, the file given contains all the essential parameters and filenames (Fig.20). The options file contains the number of domains that can be inserted, the number of files that will be inserted for each domain, the path name of files containing domain gene terms, a determination whether each gene term participates in the processing function, whether is significant or comprises a description. The files containing the terms must include in the first line a label for each column[15]. In addition, the options file contains the pathname of the abstracts' and the stop list's file.

```
Comment =          PARTICIPATES / SIGNIFICANT / DESCRIPTION
Domains = 1
domain_1_files = 2

domain_1_file_1_name    = C:\Data\Gene-Terms\domain1\EnsGeneID-EMBLID-UniProtSwissProtAC-
UniProtAC.tsv
domain_1_file_1_columns = 4
domain_1_file_1_codes   =
                          Ensembl Gene ID    true    true    false
                          EMBL ID            true    false   false
                          UniProt/Swiss-Prot AC      true    false   false
                          UniProt AC                 true    false   false

domain_1_file_2_name    = C:\Data\Gene-Terms\domain1\EnsGeneID-OMIMID-DISEASEDISCR.tsv
domain_1_file_2_columns = 3
domain_1_file_2_codes   =
                          Ensembl Gene ID    true    true    false
                          Disease OMIM ID            true    false   false
                          Disease description        true    false   true

Abstracts               = C:\Data\Abstracts\85-05-Humans-Abstracts-PubMed.txt
Wordlist                = C:\Data\Words\12dicts\2of12inf.txt
```

Fig. 20: *A sample options file*

[15] The labels may be the names from Ensembl. The application includes parsing for them.

- **A guided options selection.** At first place the number of domains must be inserted (Fig. 21). For each domain (browsed by the *next* button), the files containing the gene terms will be specified. Likewise the Stop List File, which contains the list of words that were excluded from the search process in the abstracts, will be selected in this dialog box. This option is defined in the second line of the options file.

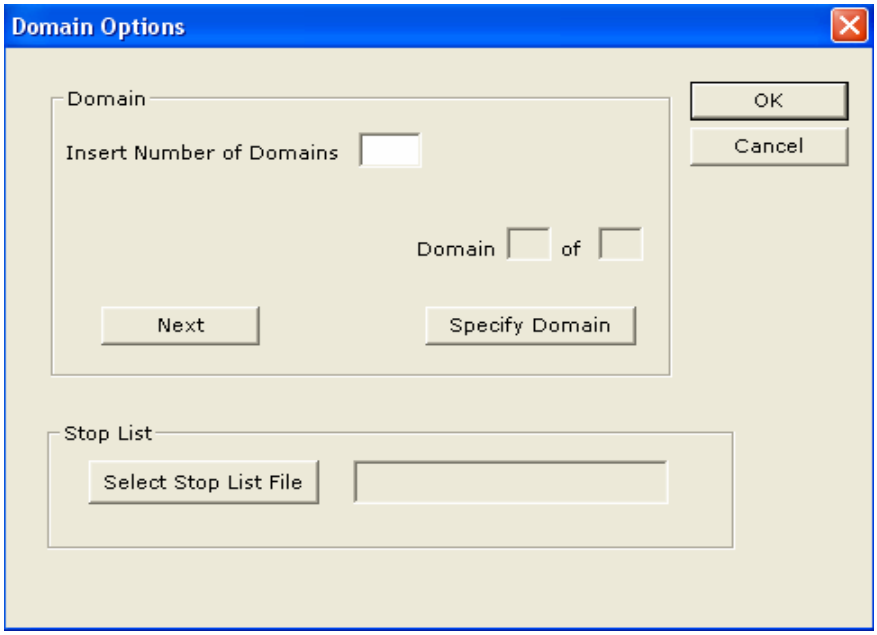


Fig. 21: *Determining the domain options*

The next step of the process includes the selection of the files containing the gene terms. Initially the number of domain files should be specified, in the text box shown in Fig.22. For each file inserted in the second text box, the user should determine the respective properties through the new pop-up dialog window. Buttons "*Next*" and "*Previous*" will gradually advance for all of the given files.

Each file selected in the previous step, contains tab-delimited data. For each column the user must specify the predefined category it belongs to. The user can browse through all the gene terms that were contained. For each gene term contained in the columns, the user must select the "*Column Participation*" if the specific term will actually take part in term extraction process from the abstracts. If the user does not select the "*Column Participation*" the term will not be loaded as a term in the application, and used in the parsing of the abstracts.

The second box determines whether the specific column contains a term that was located in a gene's description or not. Finally the last check box should be checked only for the Ensemble Gene Identifiers.

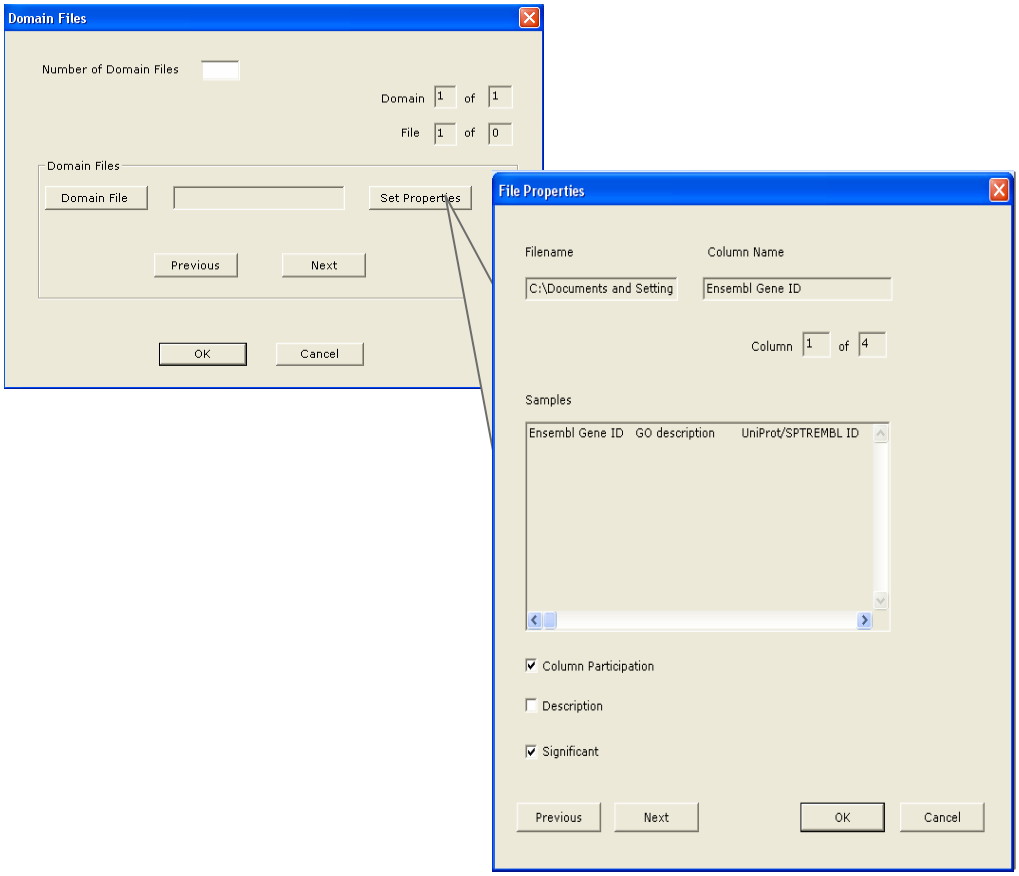


Fig. 22: *The Dialog Windows for the insertion of the domain files and the attributes specification of the contained gene terms*

4.4 Implementation Issues

In this chapter, we present a complete system analysis of our implementation for all methods presented. The test was performed on a PC with 1.8GHz Pentium processor with 512 MB of RAM. In tables 1,2 and 3 we have record the time duration of all methods.

As we have already proven the time complexity of the addition operation in a TRIE data structure is O(c*n). Where n is the length of the word added and c is the number of the letters of the alphabet that the word is written. The efficiency of the data structure can be shown at the initial phase of the implementation where a file that contains all common English words is loaded. This file contains 81.520 words and is loaded in approximately 1 sec. Subsequently, we added all significant words, loaded from Ensembl. The addition of 22.289 significant terms costs in time less than 1 sec.

During synonym term loading, the addition algorithm becomes more complex. For each term added, we look for its significant in the structure. This adds the burrier of the find operation in the addition operation. We also parse large text files making lot I/O operations. These results in a significant increasing of the time needed to store synonym terms. Generally, in order to store 190.376 synonyms we spent approximately eight minutes. Storing words from *GO-descriptions* took approximately 35 seconds. The identified set of *non-common English* words, participating in the description of one or more genes were 22.899 words. Here, in addition with special word identification, we had to find and mark all genes having this special word increasing the relevant time. The parse phase was tested under two different conditions. The first was by having all genes significant identifiers, synonyms plus description words and in the second phase we removed the words from descriptions.

In the first phase, in order to parse the 13.218 abstracts of the leukemia domain we spend 9 minutes. The MIM computations phase took 20 minutes and this was the longest procedure during the whole testing phase. MIM discretization and graph creation took negligibly low amount of time. In contrary, in the second phase where no words from description were in present the time amounts where surprisingly smaller. Instead of 20 minutes, the parsing procedure lasted for 37 seconds and the rest procedures had negligible minor times. This great difference between these two phases comes from the extreme increasing of gene-gene relations when description words are in present. These words seem to be met with high frequency in biomedical abstracts resulting in a huge amount of relation in the order of 10^6. These relations had to be stored bringing the resources of our system in the edge of exhaustion. This procedure justifies the screening phase via MIM discretization that we have introduced. In future work we plan to add methods that are more sophisticated in order to limit the relations identified. Even so, our implementation spends a reasonable time to store and manage these relations.

During classification procedure, our implementation took approximately half a minute to complete the whole task. The classification procedure was performed with all genes significant, synonyms and description words loaded in the data structure. Since our approach calculates only the feature vector of each abstract and not the MIM value the classification performance was completed in almost negligible time.

As a conclusion, we may state that our methods contain sophisticated approaches in term management that consume reasonable computer resources. The MIM computation with description words may sometimes spend extreme amounts of memory. Although

this was expected for extreme inputs like these presented we plan to improve our filtering approached during MIM computation.

Table 1. System Analysis for Gene Associations Network Extraction

Locating Gene Terms & Extracting Gene Associations Network		
Action	***Amount***	***Time***
"Loading Common English words"	81.520 Words	~1 sec
"Loading Significant Terms"	22.221 Terms	~1 sec
"Loading Synonym Terms"	190.376 Terms	8 min 14 sec
"Loading Description"	22.899 Words	~35 sec
"Parsing Leukemia abstracts"	13.218 Abstracts	9 min 19 sec
"MIM computation"	6094129 pairs of terms	20 min
"MIM Discretization"		~3 min 2 sec
"Graph MIM values"		~3 min 2 sec

Table 2. System Analysis for Gene Associations Network Extraction

The procedures without free text descriptions		
"Parsing *Leukemia* abstracts"	13.218 Abstracts	~37 sec
"MIM computation"	7888 pairs of terms	~1 sec
"MIM Discretization"		~2 sec
"Graph MIM values"		~2 sec

Table 3. System Analysis for Classification

Classification	
'Leukemia Cancer'	13.218 Abstracts
'Colon cancer'	4.594 Abstracts
Action	***Time***
"Split abstracts files into test & train"	~30 sec
"Training phase"	~3 min
"Testing phase"	~4 min

5. Validation and Evaluation of MineBioText

PubMed the service of the National Library of Medicine includes over 15 million citations for biomedical articles back to the 1950's (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi).These citations are from MEDLINE and additional life science journals. PubMed includes links to many sites providing full text articles and other related resources. However, the availability of the full text of the document is dependent on the policy of the publisher. For several documents, the provided text includes only the abstract. PubMed provides the user the correlated list of documents given a word based query through an integrated text-search based search and retrieval system called Entrez (Entrez PubMed: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi). The queries that are provided through the web interface can be based either on characteristics of the publication of the documents, either on the contents of the documents. Specifically the search can be based on the author name, on the journal title, on the year of publication or gene/protein names, on diseases, or any keyword that can determine a specific domain, consequently the retrieval of literature from the NCBI db is domain specific.

Ensembl is a joint project between EMBL - EBI and the Sanger Institute to develop a software system that produces and maintains automatic annotation on selected eukaryotic genomes (Ensembl Genome Browser: http://www.ensembl.org/index.html). The site provides free access to all data from the Ensembl project through a variety of available software. Ensembl uses MySQL relational databases to store its information. A comprehensive set of Application Programme Interfaces (APIs) serve as a middle-layer between underlying database schemes and more specific application programs. The APIs aim to encapsulate the database layout by providing efficient high-level access to data tables and isolate applications from data layout changes. Ensembl provides a Perl API and a Java API (Ensj) although the Java one is slightly less complete. An available Ensembl generic data management system we chose to use is BioMart (BioMart Project: http://www.biomart.org/); a data mining tool that can be used with any type of data and provides a build-in support for query optimization. It provides the user a set of filters in order to exclude or include characteristics of the Gene Set. The first step includes the database and dataset selection that actually determines category of the set of genes such as Homo sapiens genes/ Drosophila melanogaster genes etc. In the second step more specific characteristics about the dataset must be specified; the region of the dataset; specifications about the genes such as disease genes, common genes or having specific ids given by the user. The last step of BioMart includes the selection of the Chromosome and Ensembl attributes that will be included in the exported data, as well as some external references such as Protein ID/ GO ID/ HUGO ID/EMBL ID/.

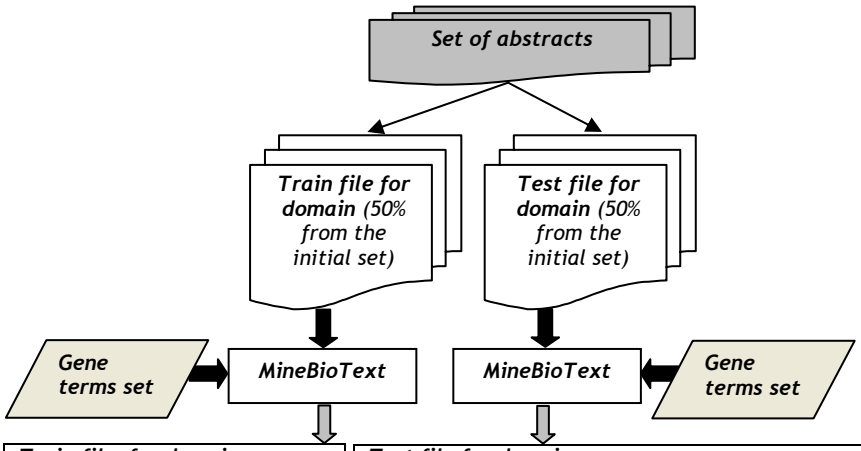
In this thesis, we decided to deal with abstracts because of the conciseness of the information gathered in the specific part of the publication. The specific gene terms located in the abstract of a document usually concerns the research hypothesis or the conclusions of the work. The set of genes that can be referred in a whole biomedical research document has a significant diversity. Therefore, the correlation indicated for the gene terms mentioned in the abstract is more significant for this thesis, than two potential terms located in the full text.

Although the search of citations in can be based on specific gene terms, we chose more generic keywords describing a domain such as *colon* or *breast cancer*. The primary reason was the fact that the set of abstracts should be covering the whole domain that is described by the keyword. The search cannot be based on specific gene terms

because of the need to have an integrated set of documents (full covering of the domain) which through the classification will reveal the more descriptive gene terms for the specific domain. ***MineBioText*** starts from a set of gene terms that were retrieved by Ensembl through a similar domain specific filtering. A primary goal of this thesis is to determine the minimum set of terms that are correlated best with the retrieved abstracts, given a domain generic set of abstracts and gene terms. The user inserts a specific set of gene terms and a domain specific set of abstracts, and queries the most significant correlated documents. A mediate output generated from ***MineBioText*** includes a vector of weighted set of terms. Each weight expresses the frequency of occurrence of the term in the text and the value (as described in chapter 3.4.1) depends on whether the Gene Terms was located in the abstract's set or in the set of descriptions of Genes. The desirable output of our system could be described as the quality of a function that corresponds the set of gene terms to the set of abstracts, in a specified domain.

5.1 A Validation Scenario

In order to evaluate the classification method we focused on two specific diseases, which we have strong belief that are caused by genetic factors. *The first phase* (Fig. 23) of the procedure includes the retrieval of the input set of abstracts that satisfied the query and the respective set of gene terms. The set of abstracts was divided in two parts; one for training set and one for testing. The two groups of retrieved references should not have common parts. In order to avoid two appearances of an abstract, we compared the two sets and removed the common abstracts from one of the sets.



Train file for domain

Ensembl ID	*Strength*
ensg00000169429	15.239394
ensg0000015734	67.928571

Test file for domain

PMID	*Term*	*Ensembl ID*	*Weight value*
15604281'	*'il8'*	*ensg00000169429*	1
15606925	*'mrna'*	*ensg00000157344*	0.071429

Fig. 23: Input and Output of the first phase of Classification in MineBioText. *Initially the set of abstract divided in two parts, the train file and the test file. The Train file is parsed by MineBioText and the desired set of gene terms that are located in, are exported, as well as the sum of the weight values. This is the train file for the specific domain. The remaining set of abstracts is the test file, that as well as train file, is parsed by MineBioText application in order just to locate the terms contained. The exported file (test file for the domain) contains the Primary Identifier of the abstracts taken from Medline (**PMID**), the gene terms as they were located in the abstract, their corresponding Ensembl ID, and their calculated weight values (estimated according to algorithm of Fig.8).*

First, we divided[16] the groups of documents; for example 50% of the abstracts for the training and 50% for the test. The independent groups of references will be input to *MineBioText* separately; for the two *'runs'* -for each domain separately- the Gene Terms (Sx, $T_{x,}$), the '*stop-word*' will be necessary as well, in order to extract the weight values for each Gene Term located in the abstract/train files. (Fig. 23)

The next phase (Fig. 24) applies the *similarity matching metric* (Eq.1) to the *test/unclassified* set abstracts; those have not been classified and comprised the test sets. The next step is to predict the class of each of the documents contained in the test set.

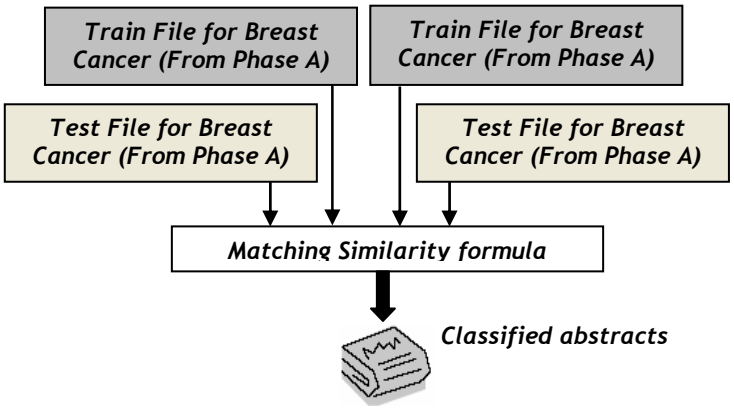


Fig. 24: *The Input for the similarity formula (second phase) of Classification in MineBioText*

5.2 Validation of MineBioText

In order to evaluate the reliability of the approach, we focused on six domain sets including retrieved sets of Abstracts and Gene Terms from PubMed and Ensembl concerning '*Colon*', '*Breast*', '*Leukaemia*', '*Ovarian*' and '*Prostate' Cancer*. The output Gene Associations Networks generated for each domain set as well as the Classification results are mentioned in the sections below.

Domain Set 1. The first set of abstracts include 4.594 for '*Colon*'cancer, 9.278 for '*Breast*'Cancer and 13.218 for '*Leukaemia*', from three keyword based queries: "(colon AND gene)", "(breast AND gene)", "(leukaemia AND gene)". The sets of abstracts were compared in order to exclude the common parts. The removed set from '*Colon*' and '*Leukaemia*' contained 141 abstracts; 252 abstracts from ''*Breast*' and '*Leukaemia*' and 499 from '*Colon*' and '*Breast*' sets. We also retrieved 168.019 Gene Terms from Ensembl (from '*Homo Sapiens*' dataset) including the features of '*Ensembl Gene ID*', '*description*' (from '*Ensembl Attributes*'), '*EMBL ID*', *Hugo* indicated as '*GO ID*' and '*Protein ID*' (from the '*External References*').

[16] The percentage is a user input.

Table 4. The Retrieved Set of Abstracts

The input set of Abstracts	
'Colon' Cancer	4.594
'Breast' cancer	9.278
'Leukemia' cancer	13.218

Table 5. The common set of abstract between the domains that were excluded.

The common set of Abstracts	
'Colon' - *'Breast'* cancer	499
'Colon' - *'Leukemia'* cancer	141
'Breast' - *'Leukemia'* cancer	252

After domain specific literature has been processed and associations were derived through *Mutual Information Measure*, the next step was the creation of knowledge (association) network. The extracted relationships include binding interactions between the connected objects or biological influence or activations that may an object cause to the other. A potential network can reveal relation between gene to gene and gene to disease. Each object present to the network will contain link to other information that can provide the system for the specific object.

Given the list of terms, the set of abstracts (tables 5,6), a user specified network of terms and a user specified percentage (X%) of the gene terms with top ranked MIM values. The MIM file computed in the previous step, including the gene terms and abstracts, is filtered with remaining rows just the ones where the MIM values are over the one that corresponds to the upper X% percentage. For all these remaining rows, the MIM values are discretised to 3 levels; Strong, Medium and Weak.

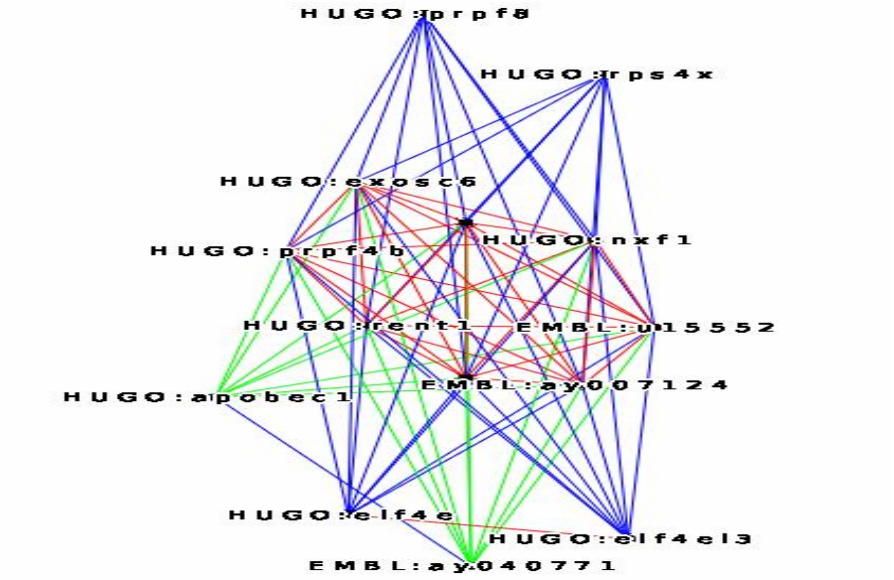


Fig. 25: *A visualized Genes Association network*

In order to visualize the network tulip graph software[17] was used. Tulip is a software system dedicated to the visualization of huge graphs that manages graphs with up to 500,000 elements (node and edges).

The visualized network contain all the information extracted from the search of the set of Gene Terms in the given set of abstracts, including explicitly mentioned relations and facts as well as novel findings extracted from the network. Each node of the networks, shown in the figures above is a Gene Identifier or a Disease. The attached labels for the Gene Terms comprise the corresponding *HUGO identifiers*[18]. Each Gene Term can be connected to another Gene Term as well as to a Disease name[19], according to the co-occurrences indicated by the revised MIM values (presented in chapter 3.5.3).

As it will be shown for the domain sets, 5 & 6 the findings that can be extracted from the Gene Networks include potential relations between Genes and Genes with Diseases. The distilled knowledge can be retrieved from either explicitly mentioned relations and facts or identification of implicit novel patterns in the given set of abstracts, based on the co-occurrences indicated by the revised network of chapter 3.5.3. In the next paragraphs, we mention examples of possible relations and facts between the nodes of a network.

Domain Set 2. The second domain set includes Breast Cancer and Leukaemia. We retrieved 162 abstracts from PubMed concerning the diseases mentioned. The input Gene Terms set and the Abstracts were input to *MineBioText* and the calculated MIM values were revised in order to produce the Visualised Network seen in Fig. 26.

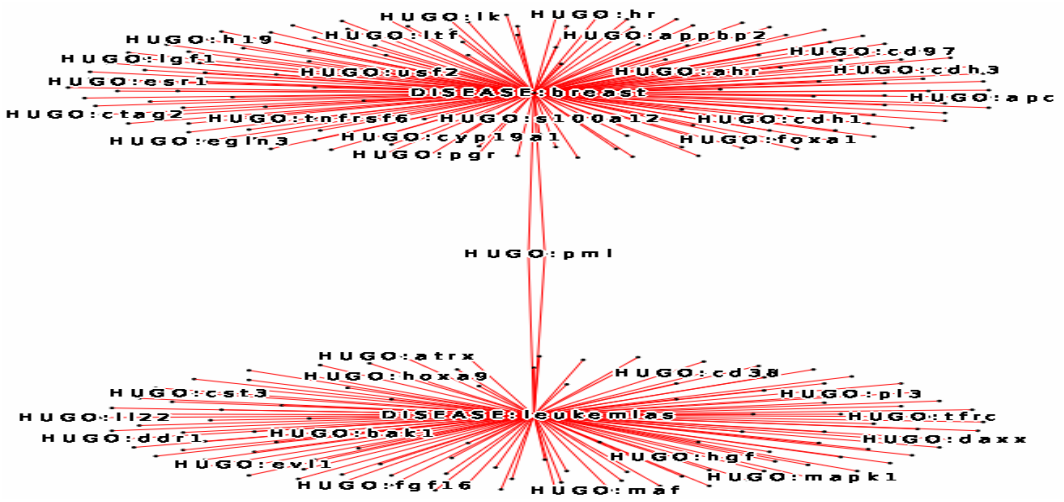


Fig. 26: *The visualized genes association network between Breast Cancer & Leukemia.*

[17] http://www.tulip-software.org/
[18] We noticed that HUGO Identifiers were the most common mentioned in the set of abstracts.
[19] The corresponding significant identifiers for the disease names have been also inserted to the Trie: ENSG10000000001 for leukaemia/leukaemia/leukaemias/leukemias, ENSG10000000002 for colon cancer and ENSG10000000003 for breast cancer.

Domain Set 3. The third domain set concerns '*Breast*' and '*Ovarian*' Cancer including 129 abstracts. The output network is seen in Fig. 27.

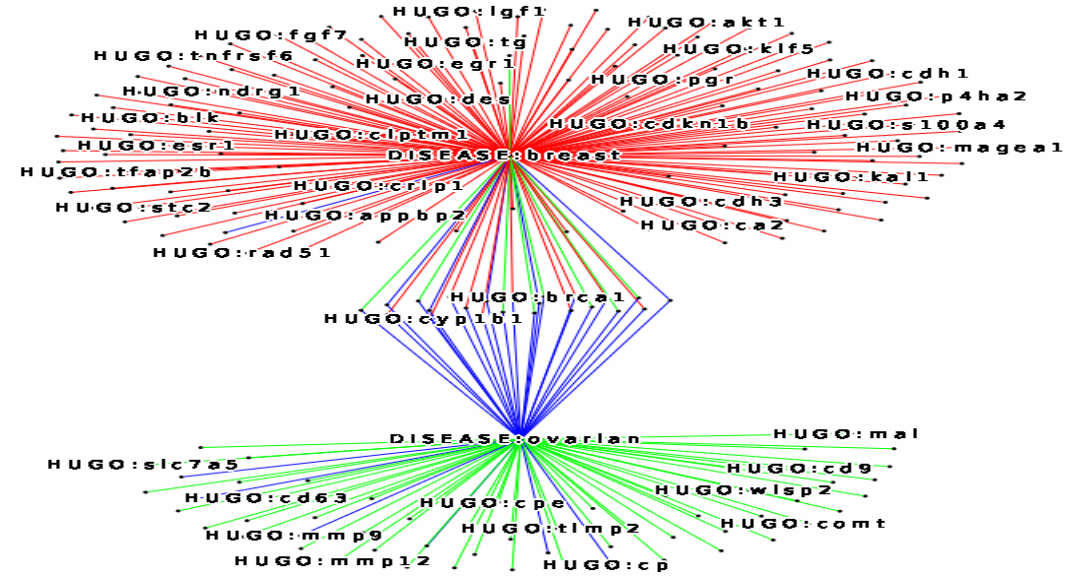


Fig. 27: *The visualized genes association network between Breast & Ovarian Cancer.*

Domain Set 4. The next experiment was made in datasets from '*Breast*' and '*Prostate*' Cancer with about 142 abstracts retrieved. The Gene Associations network from the revised MIM values is shown in Fig.28.

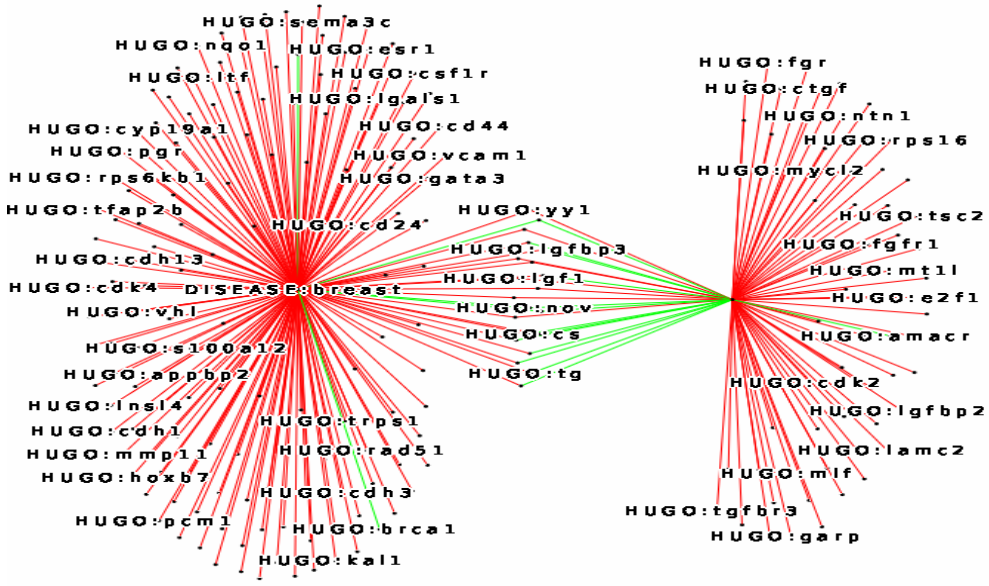


Fig. 28: *The visualized genes association network between Breast & Prostate Cancer.*

Domain Set 5. The domain set for '*Prostate'* and '*Ovarian'* Cancer was made on a set of 86 Abstracts. The output visualized network is shown in Fig. 29.

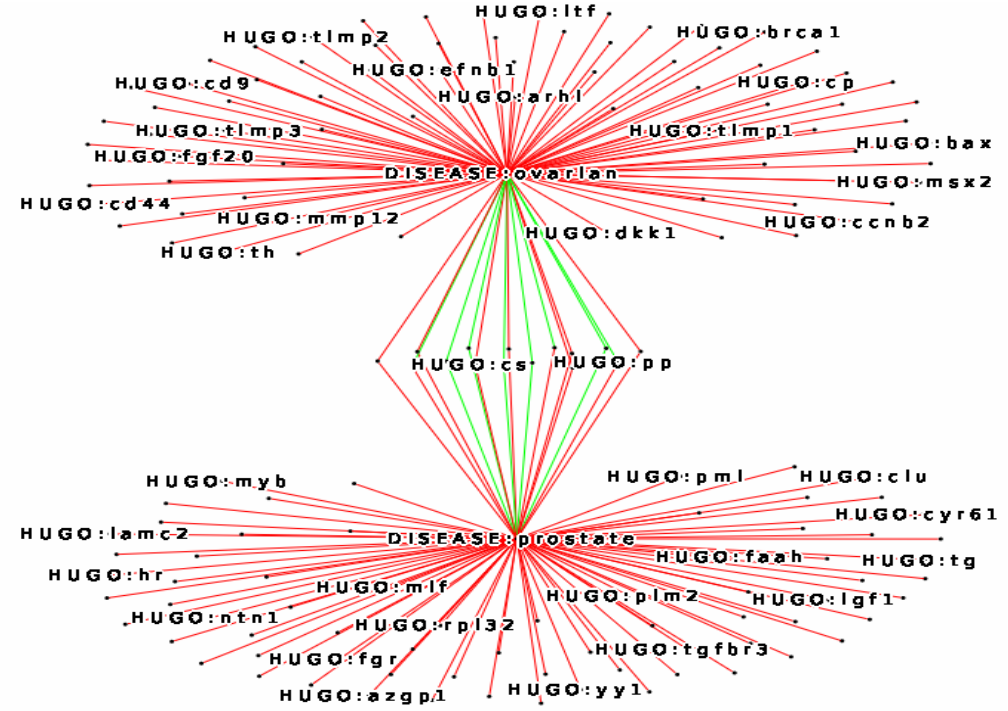


Fig. 29*: The visualized genes association network between Prostate and Ovarian Cancer*

- In Fig. 29, the network indicates the connection between '***Prostate Cancer***' with '***Cxcr4***' and '***Ovarian Cancer'*** with '***Cxcr4'***. Since there are not common publications in the set of abstracts we have retrieved a novel connection revealed**:**
  - '***Cxcr4'***- '***Prostate***'**.** In abstract "*Expression signature of the mouse prostate*", PMID: 16055444, Cxcr4 is explicitly mentioned.
  - ***'Cxcr4'***- '***Ovarian***'**.** In abstract "*Role of immunoreactions and mast cells in pathogenesis of human endometriosis--morphologic study and gene expression analysis.*" *PMID*: 15005245, '*Cxcr4*' is explicitly mentioned.
- Another example of extracted information is that of Fig. 29 where the network reveals the association between '***Prostate Cancer***' and '***Ovarian Cancer***' with Gene Term ***'IL-8'***.
  - ***'IL-8'***- ***'Prostate'***. In abstract "*Identification of genes involved in estrogenic action in the human prostate using microarray analysis.*" PMID: 14667807, ***'IL-8'***is explicitly mentioned.
  - ***'IL-8'***- ***'Ovarian'***. In abstract "*Role of immunoreactions and mast cells in pathogenesis of human endometriosis--morphologic study and gene expression analysis.*" *PMID*: 15005245, '*IL-8'* is explicitly mentioned.

Domain Set 6. The last set includes abstracts about '*Breast*', '*Ovarian*' and '*Prostate*' *Cancer*.

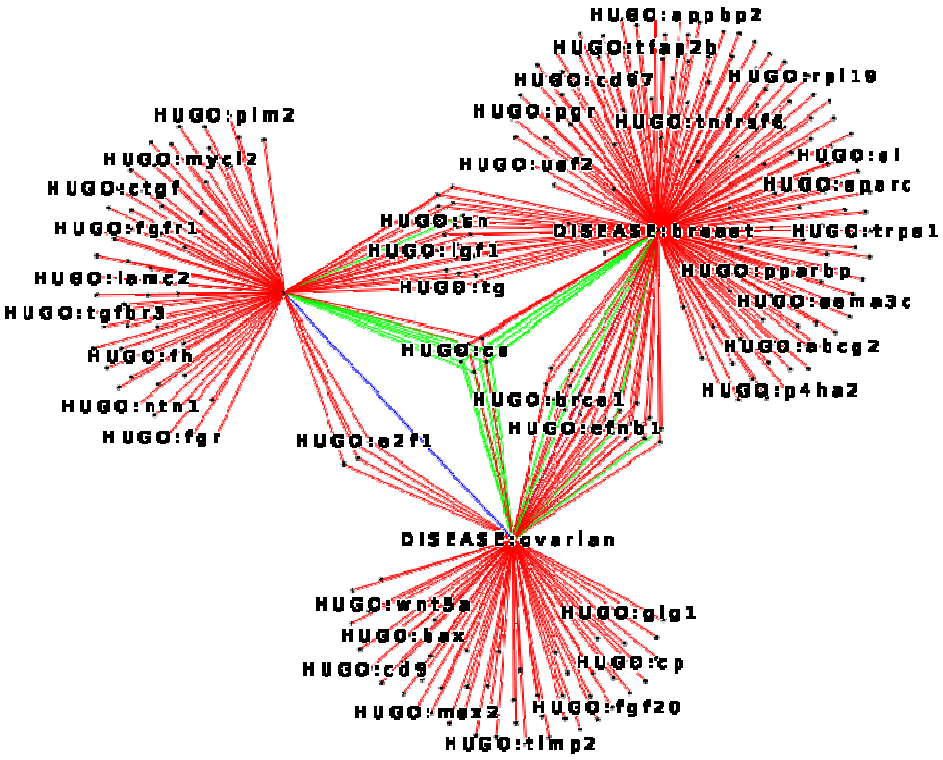


Fig. 30: *The visualized genes association network between Breast- Ovarian and Prostate Cancer.*

- The visualized network of Fig. 30 (Domain 5 including '*Breast*'/'*Ovarian*'/'*Prostate*' Cancer) revealed the connection between Gene Term '***BRCA1***' with '***Breast***' and '***Ovarian***' cancer.
  - '***BRCA1***' - '***Ovarian***'. As seen in abstract "*Gene expression profiles of BRCA1-linked, BRCA2-linked, and sporadic ovarian cancers*", PMID: 12096084 there is no-single reference to "breast cancer".

5.2.1 Towards a Qualified Validation of MineBioText Findings

The results inferred by *MineBioText* may be also validated by reference to other works related to biomedical literature mining findings, and to related ***biology-related experimental findings***. This is crucial for the reliability of the *MineBioText* findings (i.e., correlations between genes/proteins, and between genes/proteins and diseases).

With respect to that we summarise the above *MineBioText* findings, i.e., correlations, into the following figure (figure 31a), and the visualised findings from STRING to figure 31b).

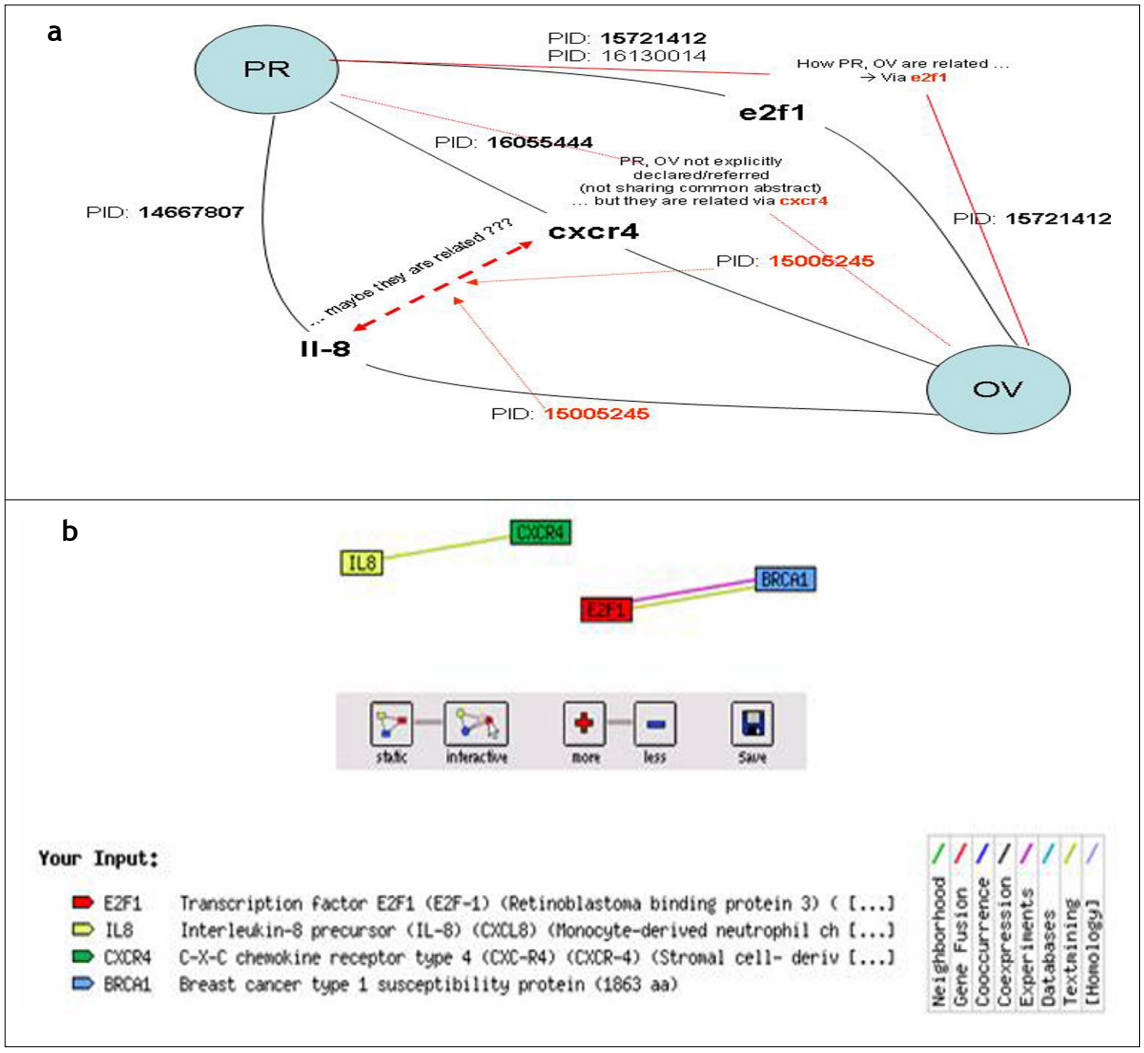


Fig. 31: *The visualized genes association network between Breast- Ovarian and Prostate Cancer.*

We used the STRING Web-based system (http://string.embl.de/) – as systems that registers and visualises gene/protein associations, based on both literature mining and experimental evidence. It can be easily verified the literature-mining as well as the experimental-based validation of the correlations between: genes/proteins 'IL-8', 'cxcr4', 'erf1' and 'brca1' - 'brca1' is also correlated with the 'ovarian-cancer' disease.

The above presentation validates *MineBioText* findings in a ***qualified*** way.

5.3 Biomedical Texts Classification: Evaluation of *MineBioText*

As shown in table 7 for the set of 9278 abstracts retrieved for the domain of '*Breast*' cancer the 4264 where classified. About the '*Colon*' cancer domain from the 4594 set of abstracts retrieved from PubMed, the 2036 where classified and for the '*Leukaemia*' domain the initial set of 13218 gave 6358 classified abstracts.

Classification results between the *'Leukaemia'* and *'Colon'* sets of abstracts, with division 50% of for both of the sets, achieved Total Accuracy = 97.5% and the AUC of the ROC curve was 0.99. The respective results, with the same split of 50% and AUC results between *'Colon'* and *'Breast'* sets achieved 93% and the sets of *'Breast'* and *'Leukaemia'* 90 %.

***Table 6**. The Classified set of Abstracts*

Domain	Set of Abstracts	Classified
'*Breast*' cancer	9278	4264
'*Colon*' cancer	4594	2036
'*Leukemia*' cancer	13218	6358

***Table 7**. The Classification Results for Domain 1*

Dataset	Percentage (train-test)	Total Accuracy	AUC
'*Breast*' - '*Colon*' cancer	50%, 50%	93%	0.9932
'*Colon*' -'*Leukemia*' cancer		97.5%	0.9965
'*Breast*' - '*Leukemia*' cancer		90 %	0.9663

***Table 8**. The Classified set of Abstracts for the Domain Sets 2-6*

Domain	Set of Abstracts	Classified
'*Breast*' -'*leukaemia*' Cancer	162	162
'*Breast*'-'*Ovarian*' Cancer	129	129
'*Breast*'- '*Prostate*' Cancer	142	142
'*Ovarian*'-'*Prostate*' cancer	86	86

As seen in *chapter 4.1.10* according to the estimated value of the similarity formula described in *Eq. 1*, a prediction is indicated as true positive (A/A); false positive (A/B); false negative (B/A) and true negative(B/B). The corresponding results for each domain set are shown in 'Class Prediction' column of table 9.

***Table 9**. The Classification Results for Domain set 2-6.*

Dataset	Percentage (train-test)	Total Accuracy	Class Prediction	AUC
'*Breast*' – '*Leukaemia*' Cancer	50%, 50%	98.148%	A/A: 92 A/B: 0 B/A: 3 B/B: 67	0.999
'*Breast*'-'*Ovarian*' Cancer		98.449%	A/A: 91 A/B: 0 B/A: 2 B/B: 36	0.998
'*Breast*' - '*Prostate*' Cancer		98.591%	A/A: 89 A/B: 2 B/A: 0 B/B: 51	1.000
'*Ovarian*' - '*Prostate*' Cancer		97.674%	A/A: 32 A/B: 2 B/A: 0 B/B: 52	1.000

5.3.1 Evaluation of MineBioText Classification on a TREC-Genomics Task

In order to evaluate the classification process we used an evaluation scheme provided by TREC 2004 Genomics Track. The Text REtrieval Conference (http://trec.nist.gov/) co-sponsored by the National Institute of Standards and Technology (http://www.nist.gov/[20]) and US department of Defense. It was started in 1992 as part of the TIPSTER Text program. Its purpose was to support research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies.

The TREC workshop series has the following goals:

- To encourage research in information retrieval based on large test collections;
- To increase communication among industry, academia, and government by creating an open forum for the exchange of research ideas;

[20] NIST is a federal technology agency that works with industry to develop and apply technology, measurements, and standards.

- To speed the transfer of technology from research labs into commercial products by demonstrating substantial improvements in retrieval methodologies on real-world problems; and
- To increase the availability of appropriate evaluation techniques for use by industry and academia, including development of new evaluation techniques more applicable to current systems.

The first year (2003) of genomics track featured two tasks including ad-hoc retrieval and information extraction; both centered on the Gene reference to function (GeneRIF[21]) and attracted 29 groups who participated. Trec 2004 focused on the standard ad-hoc retrieval task of topics retrieved from biomedical research scientists Medline bibliographic database as well as the categorization of full text documents simulating the task of curators of the Mouse Genome Informatics[22]. TREC Genomics track 2004 differed from the other tracks in that it focused on the biomedical domain; the main goal was to create a large collection of test data used for the evaluation of retrieval systems; focused on biomedical scientists, curators and annotators, providing scientific literature. The track was supplied by resources from National Science Foundation (NSF) Information Technology Research (ITR)[23] and was overseen by a committee of individual with background in IR and genomics.

A total of 33 groups participated in the 2004 Genomics Track, making it the track with the most participants in all of TREC 2004. A total of 145 runs were submitted. For the *ad hoc task*, there were 47 runs from 27 groups, while for the categorization task; there were 98 runs from 20 groups. The runs of the categorization task were distributed across the subtasks as follows: 59 for the triage subtask, 36 for the annotation hierarchy subtask, and three for the annotation hierarchy plus evidence code subtask.

The goal of the ad-hoc retrieval task was to mimic conventional searching. The use case was a scientist with a specific information need searching MEDLINE for relevant articles. The documents were a 10-year subset of MEDLINE full texts (1994-2003). This provided a total of 4,591,008 records, which is about one third of the full MEDLINE database. The ad-hoc retrieval task consisted of 50 topics derived from interviews eliciting information needs of real biologists.

The results for the ad-hoc retrieval system were measured with the classical *recall (Eq.6)* and *precision (Eq.7)* measurements, using the preferred TREC statistic of *mean average precision (Eq.8)* (average precision at each point a relevant document is retrieved, also called MAP). This was done using the standard TREC approach of participants submitting their results in the format for input to Chris Buckley's *trec_eval* program[24].

The second task was divided in three subtasks; the first one focused on the triage of articles with potential experimental evidence warranting the assignments of GO terms; the other two focused on the assignment of the three *GO* categories indicating the assignment of a term within them. Systems were required to classify full-text documents from a two-year span (2002-2003) of three journals. The first year's (2002)

[21] http://www.ncbi.nlm.nih.gov/projects/GeneRIF/GeneRIFhelp.html
[22] Mouse Genome Informatics (MGI) provides integrated access to data on the genetics, genomics, and biology of the laboratory mouse., http://www.informatics.jax.org/
[23] www.itr.nsf.gov/
[24] ftp://ftp.cs.cornell.edu/pub/smart/

documents comprised the training data, while the second year's (2003) documents made up the test data.
In the triage task positive examples were papers designated for *GO* annotation by *MGI*. Negative examples were all papers not designated for *GO* annotation in the operational *MGI* system. For the training data (2002), there were 375 positive examples, meaning that there were 5837-375 = 5462 negative examples. For the test data (2003), there were 420 positive examples, meaning that there were 6043-420 = 5623 negative examples.

5.3.1.1 TREC-Genomics Task Classification Topics

The topics for the ad hoc retrieval task were developed from the information needs of real biologists and modified as little as possible to create needs statements with a reasonable estimated amount of relevant articles (i.e., more than zero but less than one thousand).

The information needs capture began with interviews by 12 volunteers who sought biologists in their local environments. A total of 43 interviews yielded 74 information needs. Some of these volunteers, as well as an additional four individuals created topics in the proposed format from the original interview data.

The aim was to have each information need reviewed more than once but it was only able to do this with some, ending up with 91 draft topics. The same individuals then were assigned different draft topics for searching on *PubMed* so they could be modified to generate final topics with a reasonable number of relevant articles. The track chair made one last pass to make the formatting consistent and extract the 50 that seemed most suitable as topics for the track.

The topics where formatted in XML and had the following fields:

- **ID**: 1 to 50
- **Title**: Abbreviation statement of information need
- **Information Need**: Full statement information need.
- **Context**: Background information to place information need in context

From the 50 total topics, 26 had general genomic interest without any specific gene term included in the topic description. For example:

---

*<**TOPIC**>*
*<**ID**>2</**ID**>*
*<**TITLE**>Generating transgenic mice</**TITLE**>*
*<**NEED**>Find protocols for generating transgenic mice.</**NEED**>*
*<**CONTEXT**>Determine protocols to generate transgenic mice having a single copy of the gene of interest at a specific location.</**CONTEXT**>*
*</**TOPIC**>*

The rest 24 topics was an information need for the functionality of one or more specific genes or proteins. For example:

<*TOPIC*>
<*ID*>1</*ID*>
<*TITLE*>Ferroportin-1 in humans</*TITLE*>
<*NEED*>Find articles about Ferroportin-1, an iron transporter, in humans.</*NEED*>
<*CONTEXT*>Ferroportin1 (also known as SLC40A1; Ferroportin 1; FPN1; HFE4; IREG1; Iron regulated gene 1; Iron-regulated transporter 1; MTP1; SLC11A3; and Solute carrier family 11 (proton-coupled divalent metal ion transporters), member 3) may play a role in iron transport.</*CONTEXT*>
</*TOPIC*>

Although the track specification does not make any distinguishing between these two categories, our implementation is a term oriented classification mechanism and we expect to be more accurate in the **second category**. For each topic a vast collection of documents (abstracts of scientific papers) were created. This collection is named ***'pool (of documents)'***. A *pool* is composed by documents that are characterized as "*Definitely Relevant*", "*Possibly Relevant*" or "*Not Relevant*" by a set of Biology experts. The characterization procedure is quit complex in order to assure a consensus between different types of documents. The detailed procedure can be found in *TREC 2004 genomics track overview*, [Hersh,W. R., and Bhupatiraju, 2004]. The average *pool* size (average size of documents judged by topic) was 976, with a range of 476-1450.

Evaluation Measures. In order to evaluate the characterization ability of a method, TREC proposes the use of the *mean average precision* (MAP) score [Kazuaki K., 2005]. For each topic, we rank our documents in descending order from "most relevant" to "less relevant" according to our scoring schema. For each relevant document retrieved we measure the *Precision* value and the *Average Precision* (AP) shown in Equations 6,7.

$$Precision = \frac{\#\,Relevant\ Documents\ Retrieved}{\#\,Documents\ Retrieved}$$

$$AP = \frac{\sum_{i=1}^{\#\,Relevant\ Documents\ Retrieved} Precision_i}{\#\,Relevant\ Documents}$$

Equation 6,7. *Precision and Average Precision Formulas*

Finally the *Mean Average Precision* (MAP) is the mean value of all *Average Precisions* in all topics.

$$MAP = \frac{\sum_{i=1}^{\#\,Topics} AP_i}{\#\,Topics}$$

Equation 8. *Mean Average Precision.*

The procedure of evaluating the MAP score can be seen in the example of figure 32.

Relevant Documents:

Topic#	Document_ID
1	doc_a
1	doc_b
1	doc_c
1	doc_d
2	doc_k
2	doc_l
2	doc_m

Algorithm Retrieval Results:

#Topic	Document_ID	Score	Rank	Is Relevant?	Precision	Average Precision (AP)	Mean Average Precision (MAP)
1	doc_b	12	1	Yes	1/1 = 1	(1+0.67)/4=0.42	(0.42+0.39)/2=0.41
1	doc_z	8	2	No			
1	doc_a	4	3	Yes	2/3 = 0.67		
2	doc_y	15	1	No		(0.5+0.67)/3=0.39	
2	doc_l	7	2	Yes	1/2 = 0.5		
2	doc_m	6	3	Yes	2/3 = 0.67		

Fig. 32: *The procedure of evaluating MAP score*

5.3.2 MineBioText Classification Results on the TREC-Genomics Task

In order to evaluate our method we had to make some refinements in the classification procedure. First, in this domain we do not have two train/test sets. We only have a ***topic specification*** and a ***pool of documents***. The *topic specification* that forms the research query acts as a unique *train set* and the *pool of documents* as a unique *test set*.

Furthermore words in the query that are not common English words are regarded as ***gene terms*** as well. We slightly changed the set of ***stop-word list*** to exclude words that were contained in some queries but descriptive for the query. These words usually included words of *organs* (i.e., *kidney, heart*) and *diseases* (i.e., *stroke, thrombosis, cancer, tumor*).

As a future work, we plan to avoid this heuristic action by providing a general biomedical dictionary as the Medical Subject Headings (http://www.nlm.nih.gov/mesh/). Moreover whenever a query contained a Gene Term provided by Ensembl, an abstract should obligatory had this term in order to be characterized as "***relevant***". We added this heuristic because especially when a query contained a Gene Term (such as "BRCA1") and a prior common English word (such as "*breast*") the majority of the documents had the common English word but did not had the Ensembl Gene Term producing a lot Retrieved-Irrelevant documents.

Finally, the matching formula used in *MineBioText* (Eq.1), has changed to the formula given in equation 9.

$$strength_{docs}(t) = \frac{rank_{topic}(t)}{count_{topic}} \times weight_{topic} \times \left| \frac{strength(t)}{max(strength)} \right|$$

Equation 9*. The matching similarity formula used where the* ***strength*** *referring to the abstract-test file, in the original formula (Eq.1), is now referring to the* ***pool of document****. In addition, the corresponding* ***train set*** *in this equation refers to the* ***topic specification*** *formed by the research query.*

The sign of the prior formula was the criterion whether an abstract belong to "***class 1***" or to "***class 2***". The new formula does not produce negative values but it estimates "*how strong*" an abstract has common features with the query. Hence, if an abstract yields ***Similarity value*** of zero is considered as '*irrelevant'*, otherwise it is '*relevant'*. These values also used for ranking of the documents in order to estimate the MAP value as we can see in figure 33.

Topic	Gene Related	Pool	Definitely Relevant	Possibly Relevant	Not Relevant	D & P Relevant	MAP Average	**TermGene** AP	Difference
1	YES	879	38	41	800	79	0.31	0.47	0.16
2	NO	1264	40	61	1163	101	0.06	0.11	0.05
3	NO	1189	149	32	1008	181	0.10	0.26	0.16
4	NO	1170	12	18	1140	30	0.03	0.07	0.04
5	NO	1171	5	19	1147	24	0.06	0.11	0.05
6	YES	787	41	53	693	94	0.40	0.36	-0.04
7	NO	730	56	59	615	115	0.20	0.18	-0.02
8	NO	938	76	85	777	161	0.10	0.19	0.09
9	YES	593	103	12	478	115	0.61	0.65	0.03
10	YES	1126	3	1	1122	4	0.58	0.21	-0.37
11	NO	742	87	24	631	111	0.33	0.35	0.02
12	YES	810	166	90	554	256	0.42	0.42	-0.01
13	YES	1118	5	19	1094	24	0.03	0.19	0.16
14	YES	948	13	8	927	21	0.05	0.23	0.18
15	NO	1111	50	40	1021	90	0.14	0.24	0.10
16	NO	1078	94	53	931	147	0.19	0.48	0.29
17	YES	1150	2	1	1147	3	0.09	0.17	0.08
18	YES	1392	0	1	1391	1	0.63	0.85	0.22
19	YES	1135	0	1	1134	1	0.16	0.85	0.69
20	NO	814	55	61	698	116	0.15	0.20	0.05
21	YES	676	26	54	596	80	0.27	0.53	0.26
22	YES	1085	125	85	875	210	0.14	0.35	0.21
23	NO	915	137	21	757	158	0.18	0.52	0.34
24	NO	952	7	19	926	26	0.60	0.82	0.23
25	NO	1142	6	26	1110	32	0.03	0.16	0.13
26	YES	792	35	12	745	47	0.44	0.42	-0.02
27	NO	755	19	10	726	29	0.26	0.40	0.14
28	NO	836	6	7	823	13	0.20	0.22	0.02
29	YES	756	33	10	713	43	0.14	0.31	0.18
30	YES	1082	101	64	917	165	0.21	0.64	0.43
31	YES	877	0	138	739	138	0.10	0.30	0.20
32	NO	1107	441	55	611	496	0.18	0.63	0.45
33	NO	812	30	34	748	64	0.14	0.20	0.06
34	NO	778	1	30	747	31	0.06	0.07	0.00
35	YES	717	253	18	446	271	0.35	0.58	0.23
36	YES	676	164	90	422	254	0.49	0.79	0.30
37	YES	476	138	11	327	149	0.53	0.65	0.12
38	NO	1165	334	89	742	423	0.14	0.31	0.17
39	NO	1350	146	171	1033	317	0.10	0.27	0.17
40	NO	1168	134	143	891	277	0.11	0.34	0.23
41	NO	880	333	249	298	582	0.34	0.61	0.28
42	NO	1005	191	506	308	697	0.16	0.60	0.44
43	NO	739	25	170	544	195	0.12	0.30	0.18
44	NO	1224	485	164	575	649	0.13	0.63	0.50
45	NO	1139	108	48	983	156	0.03	0.13	0.10
46	YES	742	111	86	545	197	0.26	0.48	0.22
47	YES	1450	81	284	1085	365	0.07	0.37	0.31
48	YES	1121	53	102	966	155	0.17	0.35	0.18
49	YES	1100	32	41	1027	73	0.23	0.45	0.22
50	YES	1091	79	223	789	302	0.07	0.26	0.18
Mean		975.06	92.58	72.78	809.70	165.36	0.22	0.39	0.17

Fig.33: Analytical result of methods submitted in 2004 Genomic Track and the relative results of our approach. *The first column shows the ID of each topic. The second shows if a topic contained a query for one or more gene terms. Our approach is a matching term oriented technique and we expect to have better results in gene term related topics. The third column shows the total number of abstract that each topic contains. The forth and fifth column contains the number of abstracts that were characterized as "definitely relevant" and "possibly relevant" respectively by the expert group. During 2004 Genomic Track both "definitely relevant" and "possibly relevant" were managed as relevant documents. The sixth column is the final number of irrelevant documents and the seventh column is the number of relevant documents of each topic. The eighth column is the MAP value that succeeded the 47 submitted runs and the ninth column is the Average Precision succeeded by MineBioText. In the final tenth column is the difference of MineBioText and the MAP of the rest submitted methods.*

At 2004 genomics, track 47 methods (or else "runs") were submitted. In the following table, we provide information about the characteristics of each topic, the average results of the 47 methods submitted, and the results of our method.

As we can see the MAP, value for our experiments is 0.39. If we limit the topics to these that are Gene Related then the MAP value is 0.45, but if we limit the topics to these that are not Gene Related then the MAP value is 0.33. The following table shows the results of the 47 "runs" that submitted in 2004 Genomic Track. In Figure 34 we inserted our results. As we can see we have placed in the second rank.

Rank	Run	MAP			
-	***MineBioText*(GR)***	0.451	23	akoyama	0.216
1	Pllsgen4a2	0.408	24	PDTNsmp4	0.207
-	***MineBioText***	0.390	25	PD50501	0.206
2	uwmtDg04tn	0.387	26	RMITb	0.206
3	Pllsgen4a1	0.369	27	UBgtNormJM1	0.204
4	THUIRgen01	0.344	28	ConversAuto	0.201
5	THUIRgen02	0.343	29	york04g2	0.201
-	***MineBioText*(NGR)***	0.332	30	tgnNecaux	0.195
6	Utaauto	0.332	31	lga1	0.183
7	uwmtDg04n	0.332	32	york04g1	0.179
8	PSE	0.331	33	lga2	0.175
9	tnog3	0.325	34	rutgersGAH1	0.170
10	tnog2	0.320	35	wdvqlxa1	0.158
11	utamanu	0.313	36	wdvqlx1	0.157
12	aliasiBase	0.309	37	DCUmatn1	0.139
13	ConversManu	0.293	38	BioTextAdHoc	0.138
14	RMITa	0.280	39	shefauto2	0.130
15	aliasiTerms	0.266	40	rutgersGAH2	0.130
16	akoike	0.243	41	shefauto1	0.129
17	OHSUNeeds	0.234	42	run1	0.118
18	tgnSplit	0.232	43	MeijiHilG	0.092
19	UIowaGN1	0.232	44	DCUma	0.090
20	tq0	0.228	45	csusm	0.012
21	OHSUAll	0.227	46	edinauto2	0.002
22	LHCUMDSE	0.219	47	edinauto5	0.001

Fig. 34: Results of all 47 "runs" submitted in 2004 Genomics Track plus MineBioText ranked in descending order of MAPs. *The first column is the succeeded rank during Track. The second column is the method's name. More details for each submitted method can be found at "TREC 2004 genomics track overview", [Hersh,W. R., and Bhupatiraju, 2004].* MineBioText *(GR) or* MineBioText *(Gene Related) is our method limited in only gene related topics where* MineBioText *(NGR) or MineBioText (Not Gene Related) is our method limited in not gene related topics.*

Interpretation of Results. As we can see in Fig. 34, our system succeeded the second best result over 47 submitted runs, with an insignificantly small difference from the best run. The main result that has to be discussed is that our system succeeded far better performance than the others in gene related topics. That is because the other methods include *NLP techniques* that are very effective in common *Information Retrieval* domains, but they do not include *Term Matching* techniques as we do. It seems that inserting information from gene annotations is more significant than implementing sophisticated *NLP techniques*. Specifically in the ***Genomics*** domain, we have an increasing amount of sources of *gene terms*, *synonyms*, *descriptions*, *annotations* and *ontologies*. A system that tackles these sources can handle not only classification tasks but can satisfy certain information needs as well. Nevertheless, NLP still plays a central role in text categorization and we plan to include related techniques as a future step.

6. Conclusions and Future Work

Automatic extraction of information from biomedical texts appears as a necessity considering the growing of the massive amounts of scientific literature. According to NLM and the web database system there is an amount of metadata for more than 11 million articles. Despite the great amount, the gathering among specific organisms, the availability through the biomedical literature seems to lack. The common denominator where the main problems are gathered is terminology and lexical coverage. The problem arises from the fact that there is not a standard adopted vocabulary.

According to ***Ensemble***, several naming conventions are entry points into the *Ensemble* database. There seem to be a great need of organizing and centralization of the terminology in the biological domain that calls experts from different but eventually assembled sections of science. Additionally the use of pronouns and definite articles, long, complex or negative sentences or those in which information is implicit can be also inconvenient for a searching algorithm. Term ambiguity can arise from the identification with common English words or bad encoding of human genes. There are a number of efficient, publicly available tools for data processing, storing retrieving information, and analyzing results in the context of existing knowledge, involving techniques from Natural Language Processing (NLP) and data mining. The great amount of literature referring to gene and protein related biological functions raised a great interest for automating the techniques of identification, extraction, management integration and exploitation of knowledge. Despite the great need that emerges from the large amount of bibliography concerning biomedicine, an additional challenge was to deal with the problems arisen in biomedical literature.

Our approach used statistical techniques in order to address the recognition of general terms, rank the discovered Gene Terms and estimate term-hoods. More specifically ***Mutual Information Measure*** was used to estimate the strength of associations between the terms. We proposed *unsupervised learning* in order to achieve extraction of ***genes association network***. The implementation provides ***discretization*** of the MIM values according to an input threshold either the revision of MIM. The output is actually an ***Association's network*** of the gene terms and a graphical visualization of the interrelations between them is provided. After the post processing of the input terms and texts, two sets of terms are extracted in order to be used in the prediction of the unclassified set of abstracts. Finally, class prediction of documents was accomplished through supervised learning.

We presented a general schema for storing and managing biomedical term information. From the associations derived we can conclude that term information contained in the abstracts is essential and valuable source for knowledge extraction. The novel extracted information comprises a primal step for the biomedical research towards the discovery of gene to gene and gene to disease relations and transforms the application to an assistant for a biomedical researcher. The extracted networks contain trivial interrelations between genes as well as new discovered with equivalent weight that comprise a challenge for further biomedical research. The classification method presented can be applied not only in the biomedical literature but can equally be extended to a general text classification domain.

Although the complexity of terminology and nomenclature in the biomedical domain, we proposed an effective approach for the distillation of valuable information and a considerable high classification rate. Intensive experimental to other 'disease' dataset validation is remaining as a future study. Another task we consider as future work is the

revision of the Gene Nomenclature; present and extend the string matching of the Terms in the Set of Input Texts to a fuzzy string matching. Referring to the Terminology used in the experiments, we suggest a more expansive search of the Terms in MESH in order to extend the extraction of the potential relations between in the input abstracts to more object such as Anatomy Terms; Organisms; Diseases; Chemicals and Drugs; Analytical, Diagnostic and Therapeutic Techniques and Equipment; Psychiatry and Psychology Terms; Biological Sciences such as Genetic Processes/Phenomena/ Structures as well as Chemical and Pharmacologic Phenomena; Physical Sciences; Anthropology, Education, Sociology and Social Phenomena; Technology, Food and Beverages; Information Science such as Computing Methodologies, Information Services e.t.c; Persons; Health Care or Geographic Locations.

Towards efficiency some implementations issues should also be consulted; a potential implementation of a double Trie should be introduced in order to reduce the number of states created in the traditional implementation; a more extended search of the recognized words as Gene Terms from the sets of the Gene descriptions with the introduction of a more efficient approach of stemming. The application implemented could also be adjusted and settled in order to accept input from to a more generic suite of mining numerical and biomedical data such as gene expressions, sequences; or protein folding research tools. Towards an effective update of the input Gene Terms, Gene Synonyms, Descriptions, Ontologies and Publications, *MineBioText* should be adjusted as a web interface.

We considered the **efficiency** of a system in which the *Biomedical Annotation* was retrieved and recognized in the set of abstracts, instead of using Natural *Language Processing* techniques to extract the terms. The difficulty is that current technology is not at the level where it can correctly ***identify the relationships*** from a sentence and accurately link them to the genes or other biomedical objects with an ***acceptable accuracy and recall*** across all domains.

In contrast with NLP-techniques, the ***DB-population*** by sets of identified *Gene Terms* cannot be achieved, since biomedical terms are not discovered in text but retrieved from ***Ensembl***. We accept the value of ***term extraction*** so we consider the potential insertion of NLP-techniques as a future work.

References

[1] Aho, A.V., Hopcroft, J.E., and Ullman, J.D. (1983). *Data Structures and Algorithms*, Addison-Wesley, Reading, Mass., pp. 163-169.

[2] Ai-Suwaiyel, M. and Horowitz, E. (1984). Algorithms for trie compaction. *ACM Trans. Database Syst.* 9(2), pp. 243-263.

[3] Allen, J. (1995). *Natural Language Understanding*, Benjamin Cummings.

[4] Ananiadou S., Bodenreider, O., Spasic, I., and Zweigenbaum, P. (2005). Terminologies and Ontologies in Biomedicine: Can Text mining help? *MIE 2005 Workshop 401*.

[5] Ananiadou, S., Bodenreider, O., Spasic, I., and Zweigenbaum, P. (2005). Terminologies and Ontologies in Biomedicine: Can Text mining help? *Medical Informatics in Europe* (MIE 2005) Workshop 401. [http://www.mie2005.net/workshops/ws401.PDF; accessed January 2006].

[6] Andrade, M.A., and Valencia, A. (1997). Automatic annotation for biological sequences by extraction of keywords from MEDLINE abstracts. Development of a prototype system. In *Proc. AAAI Conf. on Intelligent Systems for Molecular Biology (ISMB)*.

[7] Apica, G., Ignjatovicb, T., Boyerb, S. and Russell, R. (2005). Illuminating drug discovery with biological pathways. *FEBS Letters* 579, pp. 1872–1877.

[8] Beitzel, S., Jensen, E., Cathey, R., Ma, L., Grossman, D., and Frieder, O. (2003). Task Classification & Document Structure for Known-Item Search" *IIT at TREC* (TREC 2003), p. 311-320.

[9] Bodenreider O., Rindflesch, T.C., and A. Burgun, A. (2002). Unsupervised, Corpus-Based Method for Extending a Biomedical Terminology. In *Proceedings of the ACL'2002 Workshop "Natural Language Processing in the Biomedical Domain"*, pp. 53-60.

[10] Bodon F. and Ronyai, L. (2003). Trie: an alternative data structure for data mining algorithms. *Computers and Mathematics with Applications*, 38(2003),pp.739—751.

[11] Bodon, F. (2003). A fast APRIORI implementation. *Workshop on Frequent Itemset Mining Implementations' (FIMI'03)*, Melbourne, Florida, USA, 2003.

[12] Bodon, F. (2004). Surprising results of trie-based FIM algorithms. *Workshop on Frequent Itemset Mining Implementations (FIMI'04)*. In Bart Goethals and Mohammed J. Zaki and Roberto Bayardo (Eds), *CEUR Workshop Proceedings*, v. 90, Brighton, UK.

[13] Bonetta, L. (2004). Bioinformatics - from genes to pathways, *Nature Methods* 1(2), p. 169.

[14] Brill, E. (1992). A simple rule-based part-of-speech tagger. In *Proceedings of ANLP-92*, Trento, IT, pp. 152-155.

[15] Bunescu R., Ge, R., Mooney, R.J. (2002). Extracting Gene and Protein Names from Biomedical Abstracts. *Unpublished Technical Note*, March 2002,Available from http://www.cs.utexas.edu/users/ml/publication/ie.html, 2002

[16] Bunescu R., Ge, R., Kate, R.J., Marcotte, E.M., Mooney, R.J., Ramani, A.K., Wong, Y.W. (2005). Comparative Experiments on Learning Information Extractors for Proteins and their Interactions. *Artificial Intelligence in Medicine* 33(2), pp. 139-155.

[17] Bunescu R., Ge, R., Kate, R.J., Mooney, R.J., Wong, Y.M. (2004). Learning to Extract Proteins and their Interactions from MEDLINE Abstracts. In *Proceedings of the 2004 ACM symposium on Applied computing*, pp. 121-127.

[18] Cardie, C. (1997). Empirical methods in information extraction. *AI Magazine* 18(4), 65–80.

[19] Charniak, E. (1993). *Statistical Language Learning*, MIT Press, New Haven, CT.

[20] Ciravegna, F. (2001). Challenges in Information Extraction from Text for Knowledge Management. In *IEEE Intelligent Systems and Their Applications*, (Trend and Controversies). November 2001.

[21] Cohen, W.W., and Singer, Y. (1999). Context-sensitive learning methods for text categorization. *ACM Transaction on Information Systems* 17(2), pp. 141–173.

[22] Collier, N., Nobata, C., and Tsujii, J. (2000). Extracting the Names of Genes and Gene Products with a Hidden Markov Model. In *Proceedings of COLING 2000*, Saarbruecken, pp. 201-207.

[23] Conrad, J.G., and Utt ,M.H. (1994). A System for Discovering Relationships by Feature Extraction from Text Databases. *SIGIR* 1994, pp. 260-270.

[24] Cowie, J., and Lehnert, W. (1996). Information extraction. *Communications of the ACM* 39(1), pp. 80–91.

[25] Crasto C.J., Marenco L.N., Migliore M. et al.: Text mining neuroscience journal articles to populate neuroscience databases. Neuroinformatics 1(3). 215-237 (2003).

[26] Craven, M. and Kumlien. J. (1999). Constructing biological knowledge bases by extracting information from text sources. In T. Lengauer, R. Schneider, P. Bork, D. Brutlag, J. Glasgow, H-W. Mewes, and R. Zimmer, (Eds), *Proceedings of the ISMB99, Seventh International Conference on Intelligent Systems for Molecular Biology*, pp. 77-86. AAAI Press, August 1999.

[27] Daille, B., *et al.* (1994). Towards automatic extraction of monolingual and bilingual terminology. In *Proc. Int. Conf. on Computational Linguistics (COLING-94)*, pp. 515–521.

[28] David, D., Harte, R., Lu, Y., and Chin, D. (2003). Using Natural Language Processing and the Gene Ontology to Populate a Structured Pathway Database. In *Proceedings of the IEEE Computer Society Conference on Bioinformatics (2003)*, pp. 646-647.

[29] Deerwester, S., *et al.* (1990). Indexing by latent semantic analysis. *J. Soc. Inf. Sci.* 41(6), pp. 391-407.

[30] Diaz, C. (2005). INFOBIOMED: A Joint European Effort to Support the Establishment of Biomedical Informatics. *ERCIM News* 60, pp. 16-17.

[31] Dickerson, J.A., Berleant, D., Cox, Z., Qi, W., Ashlock, D., Wurtele, E., and Fulmer, A.W. (2003). Creating Metabolic Network Models using Text Mining and Expert Knowledge. *Computational Biology and Genome Informatics*, (2003), pp. 207-238.

[32] Dingare S., Finkel, J., Nissim, M., Manning, C., and Grover, C. (2004). A System For Identifying Named Entities in Biomedical Text: How Results from Two Evaluations Reflect on Both the System and the Evaluations. *Comparative and Functional Genomics* 6(1-2), pp. 77-85.

[33] Dingare, S., Finkel, J., Manning, C., Nissim, M., Alex, B. (2005). Exploring the Boundaries: Gene and Protein Identification in Biomedical Text. BMC *Bioinformatics* 6, pp.S5

[34] Dumais, S.T. (1990). Enhancing performance in latent semantic (LSI) indexing. *Behavior Research Methods, Instruments and Computers* 23(2), pp. 229–236.

[35] Dumais, S.T., *et al.* (1988). Using latent semantic analysis to improve access to textual information. In *Proc. Conf. Human Factors in Computing (CHI88).*

[36] Dumais, S.T., *et al.* (1998). Inductive learning algorithms and representations for text categorization. In *Proc. 7th Int. Conf. on Information and Knowledge Management (CIKM-98)*, pp. 148–155.

[37] Dumais, S.T., *et al.* (1998). Inductive learning algorithms and representations for text categorization. In *Proc. 7th Int. Conf.on Information and Knowledge Management (CIKM-98)*, pp. 148–155.

[38] Dundas, J.A. (1991). Implementing dynamic minimal-prefix tries. *Software-Practice and Experience* 21, pp. 1027-1040.

[39] Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics* 19, pp. 61-74.

[40] Eisenhaber B., Bork P., Eisenhaber F., Prediction of potential GPI-modification sites in proprotein sequences, JMB (1999), 292 (3), 741-758

[41] EMBL. (2006). EMBL Nucleotide Sequence Database. http://www.ebi.ac.uk/embl/index.html [accessed January 2006].

[42] Ensembl Genome Browser. (2005). http://www.ensembl.org/index.html [accessed January 2005].

[43] Eskin, E., and Agichtein, E. (2004). Combining Text Mining and Sequence Analysis to Discover Protein Functional Regions. *Proceedings of the 9th Pacific Symposium on Biocomputing*, pp. 288-299. 2004.

[44] Evans, W.E., and Relling, M.V. (2004). Moving towards individualized medicine with pharmacogenomics. *Nature* 429, pp. 464-468.

[45] Feldman, R. (1999). Mining Unstructured Data. *KDD Tutorial Notes* 1999, pp.182-236

[46] Finkel, J., Dingare, S., Nguyen, H., and Nissim, M., Manning, C. and Sinclair, G. (2004). Exploiting Context for Biomedical Entity Recognition: From Syntax to the Web. In *Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications* at Coling 2004, pp. 88-91.

[47] Frantzi K., Ananiadou, S., and Mima, H. (2000). Automatic Recognition of Multi-Word Terms: The Cvalue/NC-value method. *International Journal on Digital Libraries* 3(2), pp. 115-130.

[48] Frantzi, T. (1997). Incorporating context information for the extraction of terms. In *Proc. ACL-EACL-97*.

[49] Frawley, W.J., *et al.* (1991). *Knowledge discovery in databases: An overview*, In Piatetsky-Shapiro, G., and Frawley, W.J., (Eds), *Knowledge Discovery in Databases*, pp. 1-27, MIT Press.

[50] Fredkin, E. (1959). *Trie memory*. Informal Memorandum, Bolt Beranek and Newman Inc. Cambridge, Mass., 23 January 1959.

[51] Fredkin, E. (1960). Trie Memory. *CACM* 3(9), pp. 490-499.

[52] Friedman, C., *et al.* (2001). Genies: A natural-language processing system for the extraction of molecular pathways from journal articles. In *Proc. Int. Conf. on Intelligent Systems for Molecular Biology (ISMB)*, S74–S82.

[53] Fukuda, K., *et al.* (1998). Toward information extraction: Identifying protein names from biological papers. In *Proc. Pacific Symposium on Biocomputing (PSB)*, pp. 705–716.

[54] Furnas, G.W., *et al.* (1988). Information retrieval using a singular value decomposition model of latent semantic structure. In *Proc. Int. ACM Conf. on Research and Development in Information Retrieval (SIGIR-88)*.

[55] GO. (2006). Gene Ontology. http://www.geneontology.org [accessed January 2006].

[56] Gondy, L., Chen, H., Martinez, J.D., Eggers, S., Falsey, R.R., Kislin, K.L., Huang, Z., Li, J., Xu, J., McDonald, D.M., and Ng, G. (2003). Genescene: Biomedical Text And Data Mining. *In Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries: Information & retrieval Data mining*, pp. 116-118.

[57] Hachey, B., Nguyen, H., Nissim, M., Alex, B., and Grover, C. (2004). Grounding Gene Mentions with Respect to Gene Database Identifiers. In Blaschke, C. (ed.), *Proceedings of the BioCreative Workshop*, Granada, Spain, 2004.

[58] Hanisch, D., *et al.* (2003). Playing biology's name game: Identifying protein names in scientific text. In *Proc. Pacific Symposium on Biocomputing (PSB)*, pp. 403–411.

[59] Harte R., Lu, Y., Asborn, S., Dehoney, D., and Chin, D. (2003). Refining the Extraction of Relevant Documents from Biomedical Literature to Create a Corpus for Pathway Text Mining. *IEEE Computer Society Bioinformatics Conference* 2003, pp. 644-645.

[60] Hayes, P. (1992). Intelligent high-volume processing using shallow, domain-specific techniques. In *Text-Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval*, pp. 227–242. Lawrence Erlbaum Assoc., Hillsdale, NJ.

[61] Hayes, P., and Weinstein, S. (1990). CONSTRUE: A system for content-based indexing of a database of news stories. In *Proc. 2nd Annual Conf. on Innovative Applications of Artificial Intelligence*.

[62] Hearst, M.A. (1999). Untangling text data mining. In *Proc. 37th Annual Meeting of the Association for Computational Linguistics*, pp. 3–10.

[63] Hersh,W. R., and Bhupatiraju, R. T. TREC 2004 genomics track overview. In Proceedings of TREC.(in press) [http://ir.ohsu.edu/genomics/trec-04-genomics.pdf]

[64] Hirschman, L., Park, J.C., Tsujii, J., Wong, L., and Wu, C.H. (2002). Accomplishments and challenges in literature data mining for biology *Bioinformatics* 18(12), pp. 1553-1561.

[65] Hofmann, T. (1999). Probabilistic latent semantic indexing. In Proc. 22nd ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR-99).

[66] HUGO. (2006). The HUGO Gene Nomenclature. http://www.gene.ucl.ac.uk/nomenclature/ [accessed January 2006].

[67] Hull, R.D., Waldman, L.F. (2003). Recognizing Gene and Protein Function in MEDLINE Abstracts. *12th Text Retrieval Conference* (TREC 2003), pp. 93-97.

[68] Ibushi, K., and Collier, N., and Tsujii, J. (1999). Classification of MEDLINE Abstracts. *Genome Informatics* 10, pp. 290-291.

[69] Iliopoulos, I., Enright, A., and Ouzounis, C. (2001). Textquest: Document clustering of MEDLINE abstracts for concept discovery in molecular biology. *Pac. Symp. Biocomput*. 2001, pp. 384-395.

[70] Jenssen T.K., Laegreid A., Komorowski J., Hovig E.: A literature network of human genes for high-throughput analysis of gene expression. Nat. Genet. 28(1), 21-28 (2001).

[71] Jenssen, T.-K., *et al.* (2001). A literature network of human genes for high-throughput analysis of gene expression. *Nature Genet*. 28, pp. 21–28.

[72] Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *Proc. European Conf. on Machine Learning (ECML-98)*.

[73] Jones, K.S., and P. Willett, P. (Eds). (1997). *Readings in Information Retrieval*. Morgan Kaufmann Publishers, 1997.

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

[75] Larkey, L.S., and Croft, W.B. (1996). Combining classifiers in text categorization. In *Proc. 19th ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR-96)*, pp. 289–297.

[76] Leek, T.R. (1997). Information extraction using hidden Markov models. Master's thesis, Department of Computer Science, University of California, San Diego.

[77] Lewis, D.D. (1995). Evaluating and optimizing autonomous text classification systems. In *Proc. 18th ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR-95)*, pp. 246–254.

[78] Lewis, D.D., and Hayes, P.J. (1994). Guest editorial for the special issue on text categorization. *ACM Transactions on Information Systems*, 12(3).

[79] Lewis, D.D., and Ringuette, M. (1994). A comparison of two learning algorithms for text categorization. In *Proc. 3rd Annual Symposium on Document Analysis and Information Retrieval (SDAIR-94)*, pp. 81–93.

[80] Lewis, D.D., *et al.* (1996). Training algorithms for linear text classifiers. In *Proc. 19th ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR-96)*, pp. 298–306.

[81] Lopez, L.M., I.F. Ruiz, R.M. Bueno and G.T. Ruiz. (2000). Dynamic Discretisation of Continuous Values from Time Series. In R.L. Mantaras and E. Plaza (Eds) *Proc. 11th European Conference on Machine Learning* (ECML 2000), *LNAI* 1810, pp. 290-291.

[82] Maly, K. (1976). Compressed tries. *Commun. ACM* 19(7), pp. 409-415.

[83] Martin-Sanchez, F., Iakovidis, I., *et al.* (2004). Synergy between medical informatics and bioinformatics: facilitating genomic medicine for future health care. J Biomed Inform. 37(1), pp. 30-42.

[84] MEDLINE. (2005). http://www.nlm.nih.gov/databases/databases_medline.html [accessed January 2005].

[85] Mitcell, T.M. (1997). *Machine Learning*. McGraw-Hill companies Inc. ISBN 0-07-115467-1.

[86] Mladenic, D. (1998). PhD thesis. http://www-ai.ijs.si/DunjaMladenic/PhD.html.

[87] Morgan, A., Yeh, A., Hirschman, L., and Colosimo, M. (2003).Gene Name Extraction Using FlyBase resources. In *Proceedings of NLP in Biomedicine, ACL 2003*, Sapporo, Japan, pp. 1-8.

[88] Morimoto, K., Iriguchi, H., and Aoe, J-I. (1994). A Method of Compressing Trie Structures. Software-Practice and Experience 24(3), pp. 265-288.

[89] Mullen, T., Mizuta, Y., and Collier, N. (2005). A baseline feature set for learning rhetorical zones using full articles in the biomedical domain. SIGKDD Explorations 7(1), pp. 52 - 58.

[90] Nervins J.R., Huang E.S., Dressman H., Pittman J., Huang A.T., West M.: Towards integrated clinico-genomic models for personalised medicine: combining gene expression signatures and clinical factors in breast cancer outcomes prediction. Hum. Mol. Genet. 12 (spec. No. 2) R153-R157 (2003).

[91] Ng, S.K., and Wong, M. (1999). Toward Routine Automatic Pathway Discovery from On-line Scientific Text Abstracts. *Genome Informatics Workshop* v.10, pp. 104-112.

[92] Nikolsky, Y., Nikolskaya, N. and Bugrim, A. (2005). Biological networks and analysis of experimental data in drug discovery. *Drug Discovery Today* 10(9).

[93] OMIM. (2005). Online Mendelian Inheritance in man. http://www.ncbi.nlm.nih.gov/Omim/ [accessed January 2006].

[94] Persidis A., Deftereos S., Persidis A., Systems literature Analysis, Pharmacogenomics 5(7), p. 943-947, 2004.

[95] Pittman J., Huang E. Dressman H. et al.: Integrated modeling of clinical and gene expression infromation for personalised prediction of disease otucomes. Proc. Natl. Acad. Sci. USA 101(22) 8431-8436 (2004).

[96] Ponte, J.M., and Croft, W.B. (1998). A language modeling approach to information retrieval. In *Proc. 21st ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR-98)*.

[97] Porter, M.F. (1980). An algorithm for suffix stripping. *Program* 14(3), pp. 130-137. - The Martin porter's algorithm home-page: http://www.tartarus.org/~martin/PorterStemmer/ [accessed january 2006].

[98] Porter, M.F. (1997). An algorithm for suffix stripping (reprint), In K.S. Jones and P. Willet (Eds) *Readings in Information Retrieval*, Morgan Kaufmann Publishers.

[99] Potamias, G. (2001). Adaptive Classification of Web Documents to Users Interests. *LECT NOTES COMPUT SC – LNCS* 2563, pp. 147-158.

[100] Potamias, G., Koumakis, L, and Moustakis, V. (2004). Gene Selection via Discretized Gene-Expression Profiles and Greedy Feature-Elimination. *LNAI* 3025, pp. 256-266.

[101] Potamias, G., Raxenidis, V., and Papadakis, A. (2001). Personalized Classification of Web Documents. In *Procs 8th Panhellenic Conference in Informatics*, Nicosia, Cyprus, vol. 2, pp. 213-222.

[102] Proux D., Rechenmann F., Julliard L., Pillet V. V. Jacq B. Detecting Gene Symbols and Names in Biological Texts : A First Step toward Pertinent Information Extraction. Genome Inform Ser Workshop Genome Inform, 1998;9:72-80

[103] PubGene Gene Database and Tools. http://www.pubgene.org/

[104] Rabiner, L. R. and Juang, B. H., An introduction to hidden Markov models, IEEE ASSP Magazine, (January 1986), pp. 4-15

[105] Rindflesch, T.C., *et al.* (2000). Edgar: Extraction of drugs, genes and relations from the biomedical literature. *Proc. Pacific Symposium on Biocomputing (PSB)*, pp. 514–525.

[106] Russell, S.J., and Norvig, P. (1995). *Artificial Intelligence, A Modern Approach*, chap. 22–23, Prentice Hall, Englewood Cliffs, NJ.

[107] Sahami, M., *et al.* (1996). Applying the multiple cause mixture model to text categorization. *Proc. 13th Int. Conf. on Machine Learning.*

[108] Salton, G. (1989). *Automatic Text Processing*, Addison-Wesley, Reading, MA.

[109] Sathiya, K.S. *et al.* (2002). A Machine Learning Approach for the Curation of Biomedical Literature (for *KDD Cup 2002; Task 1*). *ACM SIGKDD Explorations Newsletter* 4(2), pp. 93-94.

[110] Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys* 34(1), pp. 1–47.

[111] Set, J.A., Haas, J., and Overton, C. (1993). Knowledge Discovery in GenBank. In *Proceedings of the 2nd International Conference on Information and Knowledge Management*, pp. 554-564.

[112] Shannon, C. (1948). A Mathematical Theory of Communication. Reprinted with corrections from *The Bell System Technical Journal* 27, pp. 379–423, 623–656, July, October, 1948 [http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html; accessed January 2006]

[113] Shatkay, H., and Feldman, R. (2003). Mining the Biomedical Literature in the Genomic Era: An Overview. *Journal of Computational Biology* 10(6), pp. 821-855.

[114] Shatkay, H., *et al.* (2002). Information retrieval meets gene analysis. *IEEE Intelligent Systems, Special Issue on Intelligent Systems in Biology* 17(2), pp. 45–53.

[115] Shen, D., Zhang, J., Zhou, G., Su, J., and Tan, C. (2003). Effective Adaptation of Hidden Markov Model based Named Entity Recognizer for Biomedical Domain. In *Proceedings of NLP in Biomedicine, ACL 2003*, Sapporo, Japan, pp. 49-56.

[116] Song, Y.-I., Han, K.-S., Seo, H.-C., Kim, S.-B., Rim, H.-C. (2003). Biomedical Text Retrieval System at Korea University. *12th Text Retrieval Conference* (TREC 2003) p. 368.

[117] Stapley B., Kelley L., and M.Sternberg, 2002. Prediction in the Sub-Cellular Location of Proteins from Text Using Support Vector Machines., Proceeding of the Pacific Symposium on Bio-Computing, PSB(2002)., pp. 374-385.

[118] Stapley, B.J., and Benoit, G. (2000). Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in MEDLINE abstracts. *PSB 2000*, pp. 529-540.

[119] SwissProt (2006). The UniProtKB/Swiss-Prot Protein Knowledgebase (an annotated protein sequence database). http://www.ebi.ac.uk/swissprot/ [accessed January 2006].

[120] Tanabe, L., and Wilbur, W.J. (2002). Tagging gene and protein names in biomedical text. *Bioinformatics* 18(8), pp. 1124-1132.

[121] Tasnim, G.R. (2003). REGEN: Retrieval and Extraction of Genomics Data. *TREC 2003*, pp. 107-116.

[122] UniProt. (2006). The Universal protein resource. http://www.ebi.ac.uk/uniprot/ [accessed January 2006].

[123] Uramoto, N. , Matsuzawa, H., Nagano, T., Murakami, A., Takeuchi, H., and Takeda, K. (2004). A text-mining system for knowledge discovery from biomedical documents. *IBM Systems Journal* 43(3), pp. 516-533

[124] van Rijsbergen, C.J. (1979). *Information Retrieval*, Butterworth, London.

[125] Weed, L.L. (1991) Knowledge Coupling: New premises and new tools for medical care and education. Springer-Verlag, ISBN: 0387975373.

[126] Witten, I.H., *et al*. (1999). *Managing Gigabytes, Compressing and Indexing Documents and Images* (2nd ed.), Morgan-Kaufmann, San Diego, CA.

[127] Wren, J.D, Bekeredjian, R., Stewart, J.A., Shohet, R.V., Garner H.R. (2004). Knowledge discovery by automated identification and ranking of implicit relationships. *Bioinformatics* 20, pp. 389-398.

[128] Wren, J.D. (2004). Extending the mutual information measure to rank inferred literature relationships. *BMC Bioinformatics* 5:145, doi:10.1186/1471-2105-5-145 [http://www.biomedcentral.com/1471-2105/5/145, accessed January 2006].

[129] Yandell, M.D., and Majoros, W.H. (2002). Genomics and natural language processing. *Nature Reviews* 3, pp. 601–610.

[130] Yang, Y., and Chute, C.G. (1994). An example-based mapping method for text categorization and retrieval. *ACM Trans. Inf. Systems* 12(3), pp. 252–277.

[131] Yang, Y., and Liu, X. (1999). A re-examination of text categorization methods. In *Proc. 22nd ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR-99)*, pp. 42–49.

[132] Yildiz, Y.M., Pratt, W. (2005). The Effect of Feature Representation on MEDLINE Document Classification. *American Medical Informatics Association Fall Symposium (AMIA05)*.

[133] Young-In, S., Han, K-S., Seo, H-C., Kim, S-B., and Rim, H-C. (2003). Biomedical Text Retrieval System at Korea University", *12th Text Retrieval Conference* (TREC 2003) p. 368.

[134] Zhang Z., Ng S.K., Interweaver: interaction reports for discovering potential protein interaction partners with online evidence, NucleicAcids Res. 1, 32 (2004).

[135] Zhou G, Zhang J, Su J, Shen D, Tan CL: Recognizing names in biomedical texts: a machine learning approach. Bioinformatics 2004, 20(7):1178-1190.

[136] Zhou, G.D. Recognizing names in Biomedical Texts Using Hidden Markov Model and SVM Plus Sigmod. Proceedings of 2004 Joint Workshop on Natural Language Processing in Biomedicine and its Applications (2004), pp.1-7

[137] Zhou, G.D., Shen, D., Zhang, J., Su, J., Tan, S.H. (2005). Recognition of protein/gene names from text using an ensemble of classifiers, *BMC Bioinformatics* 6(Suppl 1):S7.