ISSN: 2455 – 4200
International Journal of Engineering Research and Modern Education
Impact Factor 6.525, Special Issue, April - 2017
6th National Conference on Innovative Practices in Construction and Waste Management
On 25th April 2017 Organized By
Department of Civil Engineering, Sri Ramakrishna Institute of Technology, Coimbatore, Tamilnadu

DETECTION OF CANCER BIOMARKER

M. Bagyarani*, A. Kunthavai, V. Vidhyashree*, A. Sumathi* & C. Kalaivani***
* Lecturer/CE, CIT Sandwich Polytechnic, Coimbatore, Tamilnadu
** Assistant Professor, Coimbatore Institute of Technology, Coimbatore, Tamilnadu

Cite This Article: M. Bagyarani, A. Kunthavai, V. Vidhyashree, A. Sumathi & C. Kalaivani, "Detection of Cancer Biomarker", International Journal of Engineering Research and Modern Education, Special Issue, April, Page Number 194-200, 2017.

Abstract:
The Advancement of Information technology has grown faster and faster. The data acquired and stored as a digitized form exceeds a lot to a peak apart from the imagination. It challenges the research analysts for knowledge discovery. New methods or techniques are necessary to extract useful information from the raw data collected in repositories. Bioinformatics is the area where we analyze the biological information about the living organisms which helps in major areas like disease prediction. Many people are affected by a crucial disease i.e., cancer. Biomarker the indicator of disease plays a vital role in diagnosing of the presence of disease. The presence of cancer disease can be detected using the measurable indicator called cancer biomarker. The recent research focus on analyzing the gene data and the high dimensional data stored as text in repositories. The paper concentrates on the second domain to retrieve useful information for the detection of presence of disease. The useful data from the repositories are mined using powerful tool is developed. The potential biomarker can be extracted for cancer detection. The datasets are taken from the PubMed and the tool extracts useful information to detect the presence of cancer disease. Now-a-days the number of patients affected by the disease like cancer increased. The death rate increased particularly for the patients affected by prostrate and breast cancer. The paper takes the input of prostate and breast cancer data from the repository. The proposed system gives the promising and effective result.

Key Words: Cancer Biomarker, Text Analysis, Text Mining, Information Extraction & Knowledge Discovery

1. Introduction:

General: Networks do exist in most of the domains as it is very difficult to find a domain or an area where the entities interact with other. Better abstractions of the entities covered by the domain and connecting the relationship among entities are reflected by networks. The network association is growing rapidly and helps the researcher to study and compares the association and connectivity in a fair manner. Genes, micro RNA, proteins like biological information or context are constructed using association networks between the above data. Relationship between molecules or drugs and a molecule are built using association networks. High quality molecular association network (absorbed information) makes the researchers to compare and address the molecular therapeutic/toxicological profiles of various drugs when the molecular drug related information available for the processing of data to extract useful information for disease prediction. Analysing and Extracting of regulatory relationship between genes makes way to detect the presence of disease and understand its impact. The technology advancement and the experimental growth were done to determine interactions and associations between molecules. The experiment enriched the scientific literature with more articles to tackle series health related problems and reports interesting results. Development of automated techniques is required to connect the mined information from the literature to the specific disease data. The manual information mining from the repository is impossible as the data grows larger and larger to save time and money. Various text mining algorithms were developed for the above said difficulties. Text-mining algorithms for relations extraction can be classified into simple statistical co-occurrence, pattern matching [17], and full-sentence parsing. When automatic extraction of biomedical information is done, redundancy of the relations was extracted. The accuracy of the analysis and the result goes low. To overcome the problem, automatically extracted relations were aggregated and proposed. The abstracts of the scientific publications contain the efficient source of scientific biomedical information as they are widely available and accessible. The precise and the description of the information in the actual article are present in the abstract. It is reliable when the text- mining is applied to the abstract as it contains direct description. Apart from the simple statistical method, several approaches with advanced and complicated natural language processing methods have been proposed. A tool named Bio Inf Builder (BIB) is developed to build association networks based on templates and rules describing building links. More restricted parameters are only allowed in the tool for providing further precision. The weighted links are predefined on the level of each keyword is also supported by this tool BIB. The input for the BIB is the raw text documents in various formats, in addition with the nodes for building and binding the links. BIB extracts information from the abstracts, articles and documents with the help of predefined words related to the disease. The input is taken from the National centre for Biotechnology Information (NCBI). The BIB is tested with association network for the prostate and breast cancer. Topological and biological analysis is also conducted. Building the association network by extracting abstracts is described in the following various stages are described in the following sections. Result analysis of the networks are conducted, presented and discussed.

2. Related Work:

Data analysis requires new sophisticated techniques in the rapid advancement in the technology. Data analysis highly influenced disease biomarker discovery. The data which are extracted from the scientific data warehouse and the microarray data play as a valuable source. A robust outcome can be produced by combining these two data. The paper concentrates on the extraction of useful information of detection of disease from the scientific warehouse by implementing network construction and analysis. This section reviews about the survey related to the microarray data as well as the scientific bioinformatics data analysis.

Gene Expression Data Analysis: Discovering of novel biomarker for cancer is experimented with thousands of proteins and

genes in a single experiment in simultaneous examination in this paper [13]. The paper predicts the general behaviour of the tumor using robust technology. DNA microarray technology gives details about the different types of cancer disease. High dimensional data of biological information can be handled using proteomic and genomic technologies. The serial analysis of DNA and mRNA is enhanced with the parallel analysis [19]. This paper classifies the tumor based on the behaviour of the gene [19], [25]. This paper analyzes Proteomic and spectrometry is considered as the acquired information from mRNA is insufficient to get the details or functionality of proteins. Unique proteins reveal the subtype of the cancer disease. The unique behaviour of the protein can be discovered by spectrometry and also the tumor subtype of cancer disease. A robust technology used to measure the proteins in terms of quantitative and qualitative [3]. This paper explains about the statistical and machine learning methods to detect disease biomarker. In this paper the author explained about the difficulties faced when the full gene sequence of the human is taken for consideration. The noise or the irrelevant data should be avoided to get the useful information regarding the disease. In the high dimensional microarray data of gene selection a proper method is followed to mine the disease related data [14]. Entropy based method used in this paper for gene selection [6]. Correlation based method used in this paper for the gene selection [10]. Unsupervised feature selection method used to select the gene in this paper [24]. The entropy based method filters out the features and minimizes the heuristics. Highest discrimination value is used in the correlation approach. Finally, we need to mention the commonly used SVD method which finds the singular values for the purpose of feature selection. Different gene selection algorithm was proposed to get the disease detection biomarker and also to classify it in subtypes. The tumor can be cancerous or non-cancerous, so detecting the cancerous tumor by the proper gene selection using artificial intelligence matters. Artificial Intelligent based approaches are used as classification algorithms. Naive Bayes algorithm is a popular algorithm for classification. It uses probabilistic induction for assigning class labels for tuple testing. The next popular classification algorithm is the decision tree algorithm. To avoid over-fitting of the above algorithms by pruning, heuristics is used to avoid errors. The correlation based approach is used to avoid noise and bias with two methods nearest neighbour and clustering method. Defining the distance measure between the known trained vectors and testing vector is the nearest neighbour algorithm. The training sample cancer profile is taken for testing in this project to find the cancer disease prediction. The clustering based approach based on training samples into different groups and the samples with value different from the threshold are removed. The grouping of genes which shares specific functions of tumor class can be viewed using a clustering algorithm two dimensional hierarchical clustering. One of the recent methods to infer knowledge from microarray data is gene network. Apriori known gene selection is included for driving the modules of gene expression. Discrete formalism is also used by some papers to observe gene expression. Logical discrete formalism was also used for gene selection may be explained by perturbation of the network. The paper analyzes different steady state of the biological reaction networks as positive and negative influences and named the instance as Metareg formalism used to predict different patterns of gene expression from the network structure [7]. A method with Different types of quantitative data correlates with the interaction graphs has been developed in this paper. The distance between the pairs of genes is considered to identify the activated path [23].

Biomedical Text Analysis: Extracting useful health information from warehouse underlies numerous advances in the biomedical research called biomedical text mining. Dynamic knowledge base is essential for the publishing work on biomedical literature databases and MEDLINE. In the fine level as well in the coarse level, large scale literature, text mining bridges highly specialized biomedical research subject and interactive effort in understanding complex biological problems [26]. Biomedical data analysis using text mining is reviewed. The fundamental problem in text mining is Name entity recognition and the simple task is to find the biological entities like genes, proteins etc [11]. Lexicon is a method used to find out the biological entities in this paper [12]. In this paper [22] rule based mining is used and the statistically-based approached used in this paper [32]. Independent assumption between instances is the problem in Name entity recognition system and is measured using F-score and F-measure [15]. The classification mainly classifies the accurate and meaningful database annotations and interested documents [5]. In the paper [4], Probabilistic laten categorizer, one of the statistical data approaches is used. The rate of biomedical literature increases a key task of extracting synonym and abbreviation. The name synonym lists from automated means [30], other specified forms [15], [28] are used as the approach for text mining here the biological information extraction. The protein and gene names are used mostly in a standard format, so it will be easy to retrieve whereas some non standard names are also widely used in the literature gives difficulty for interactive biomedical research. Manually generated templates are used in this paper [31] as the relationship extraction plays a vital area after synonym extraction. The relationship between the gene and protein has to be detected or extracted. The relationship and the integration between the gene and the protein is found using automatic template in this paper [2]. In the paper [8] the author used clustering technique to mine textual information of genes. The author in the paper [20] discussed the topological information, data mining technique used to extract keywords between entities. Biological supporting texts are necessary to extract the relationship and the type and it is a challenging task [27]. Integrative framework was used for text mining in the papers [16] and [9].

3. Pubmed Abstracts Extraction:

BIB is designed in such a way that it extracts any number of abstracts form the pubmed using keywords. It does the searching process automatically by scanning the database in the pubmed using the given keyword. The results are saved physically in the hard drive for further network building process. The actual result results in row text format. The abstract extraction using set of keywords has to be provided by the user only. Even the pubmed search engine is taken for the study; a small portion of the database is build as association network. BIB provides flexible local storage and to build a small portion of the database. The local storage feature improves the processing speed to build the network. The tool converts the pubmed abstract database to XML files links to the pubmed website location. BIB parses the XML file and results are retrieved locally

ISSN: 2455 – 4200
International Journal of Engineering Research and Modern Education
Impact Factor 6.525, Special Issue, April - 2017
6th National Conference on Innovative Practices in Construction and Waste Management
On 25th April 2017 Organized By
Department of Civil Engineering, Sri Ramakrishna Institute of Technology, Coimbatore, Tamilnadu

4. Association Network Building:

Few parameters have to be set before association network as the abstracts are retrieved and ready for mining. The nodes for which the network has to be build should be indicated by the user. The key nodes, keywords, rules, dates, weights and the rules should also be indicated. Graphical User Interface (GUI) is used by the BIB for the nodes. The nodes here considered as the genes and may be the CSV files. The list of genes and its association network is explained in the following section. The rules of the BIB based on the binding links between genes using keywords. BIB is given with the keywords depending on the objective of the network builder and the users need. If the gene regulatory gene is represented using association network, the keyword given is 'regulate' or 'interact'. The rules or the keywords will indicate which binding link should be associated with genes or nodes. The result produced by the early stage of the BIB used to retrieve relation between keywords and nodes using full sentence parsing algorithm. BIB allows the individual user to add weight for each keyword in the same document and also in the same paragraph which reflects the thickness of the binding link. The manual assignment of rules or keywords can be done using BIB and can also be imported from the text file. To create a specific format for the representation of keywords along with their weights proves the ability.

<Keyword1>[Document:<DocumentWeight>,Paragraph:<ParagraphWeight>]
||/&&
<Keyword2>[Document:<DocumentWeight>,Paragraph:<ParagraphWeight>]
... ...
<Keyword(n)>[Document:<DocumentWeight>,Paragraph:<ParagraphWeight>]

The additional feature included and supported by the BIB is setting key nodes. The user can assign different weight values for binding the links for one or more genes or nodes within the network. The previous knowledge about the specific gene should be considered to build the gene regulatory network, to find some other gene has more impact on building the interactive association network. When the feature has no meaning, can be neglected and assigning key node is not mandatory. BIB allows the user to set specific publication dates looking for. Different association networks in different time of publicationscan be building with the ability of considering dates during building network. This will lead to compare different published information. The thickness of the binding links between nodes is indicated by the weights. Nodes, keywords, key nodes and dates are assigned with the weights. The flexibility of the network with different entities increases the network association. The resulting network can be adjusted by the user depending on the network building objective. The frequency of the occurrence of keyword increases the thickness of the binding links in addition to the thickness by its weight.

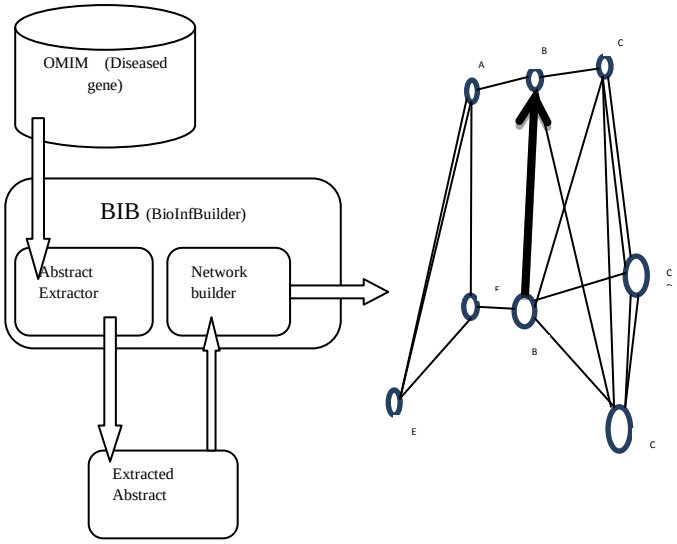

Figure 1: Bio Inf Builder Working Flow diagram

5. Prostate Cancer Network:

The prostate cancer abstract from the pubmed website is extracted to build the gene network by the BIB. The input from the OMIM database from the pubmed is taken as it contains wide range of diseased genes. Our project has chosen 8 genes related to the prostate cancer. 1,623 abstracts were extracted from the pubmed. The tool picks up the keyword to develop the frequency count of the words extracted from the abstract. The weights associated to the genes and keywords are equivalent. The keywords are picked and the frequency count is developed for the abstract extracted. All genes and keywords have equivalent weights associated with them. The network builder is run using the input abstract from the pubmed by specifying parameters as shown in Table I. The network contained 8 nodes and 15 edges; the time taken for building the network is 7 minutes. The output is generated with corresponding weight.

6. Breast Cancer Network:

The process described in the section V is done as same for the breast cancer regulatory network. Here also 8 genes related to breast cancer is chosen for extraction. 15,021 abstracts related to the breast cancer is extracted. Here also all the genes and the keywords have equivalent weights associated with them. The network is build using the input abstract from the pubmed

ISSN: 2455 – 4200
International Journal of Engineering Research and Modern Education
Impact Factor 6.525, Special Issue, April - 2017
6th National Conference on Innovative Practices in Construction and Waste Management
On 25th April 2017 Organized By
Department of Civil Engineering, Sri Ramakrishna Institute of Technology, Coimbatore, Tamilnadu

by specifying the parameters given in the Table II. The resulting node has 8 nodes and 17 edges. The processing time for building the association network for the breast cancer is 8 minutes. The output is generated and displayed with the corresponding weights.

Figure 2: Abstract Extraction in BIB

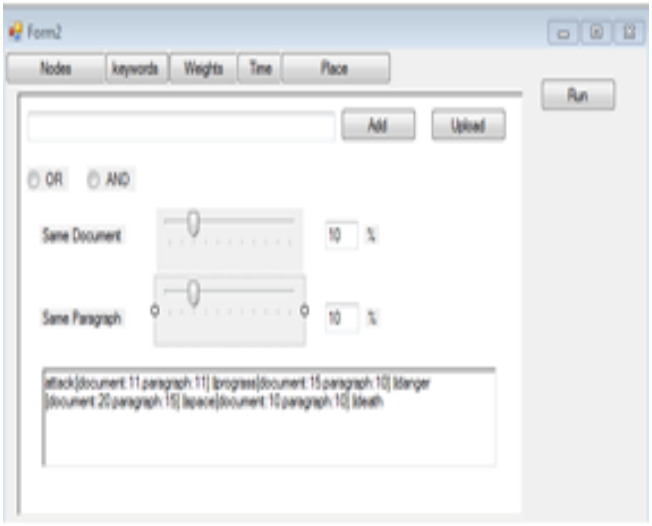

Figure 3: BIB Network builde

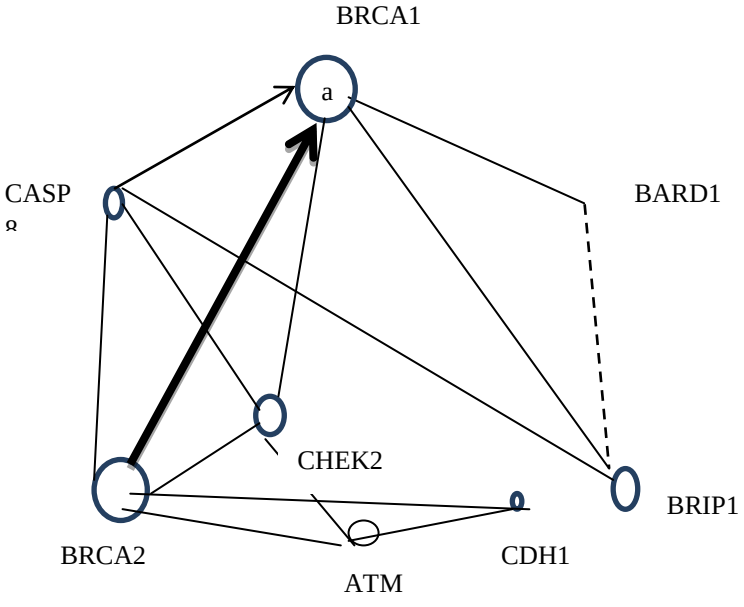

Figure 4: Prognosis prostate cancer network

ISSN: 2455 – 4200
International Journal of Engineering Research and Modern Education
Impact Factor 6.525, Special Issue, April - 2017
6th National Conference on Innovative Practices in Construction and Waste Management
On 25th April 2017 Organized By
Department of Civil Engineering, Sri Ramakrishna Institute of Technology, Coimbatore, Tamilnadu

Table 1: Genes and the Keywords Taken for Building Prostate Cancer Association Network

Genes	Keywords
ATM	Regulate
BARD1	Interact
BRCA1	Activate
BRCA2	Prognostic
BRIP1	Biomarker
CASP8	
CDH1	
CHEK2	

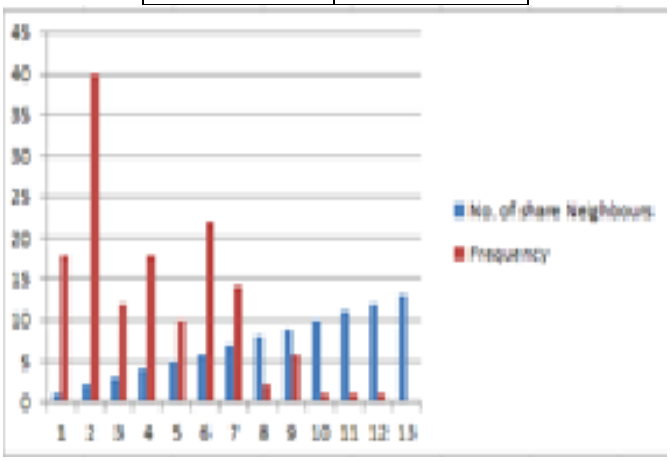

Figure 5: Prognosis prostate cancer network

Table 2: Genes and the Keywords Taken for Building Breast Cancer Association Network

Genes	Keywords
AR	Regulate
BRCA1	Interact
BRCA2	Activate
CD82	Prognostic
CDH1	Biomarker
CHEK2	
EHBP1	
ELAC2	

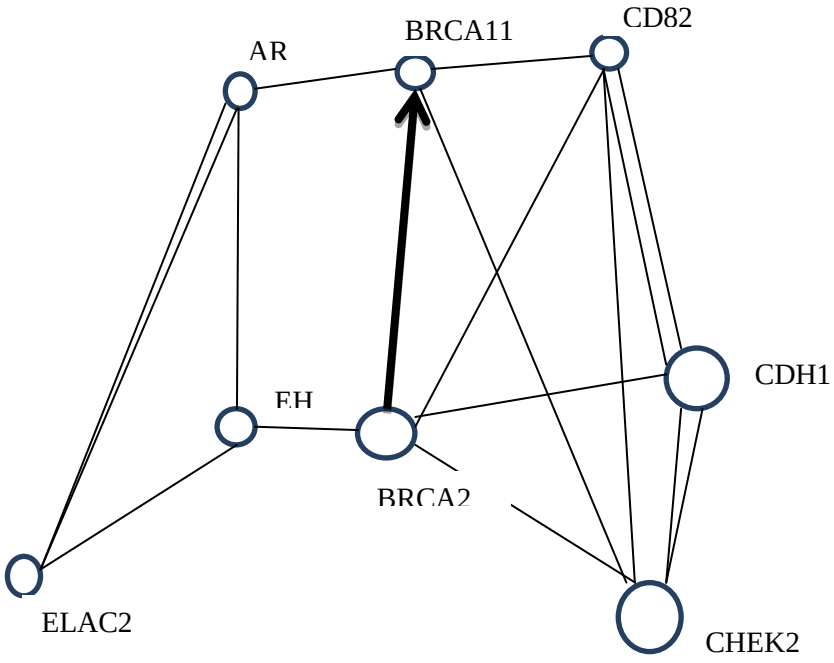

Figure 6: Prognosis breast cancer network

ISSN: 2455 – 4200
International Journal of Engineering Research and Modern Education
Impact Factor 6.525, Special Issue, April - 2017
6th National Conference on Innovative Practices in Construction and Waste Management
On 25th April 2017 Organized By
Department of Civil Engineering, Sri Ramakrishna Institute of Technology, Coimbatore, Tamilnadu

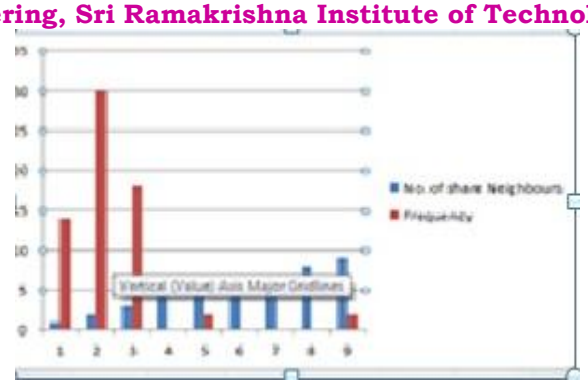

Figure 7: Number of shared neighbors in Breast cancer gene network.

7. Analysis:

The clustering coefficient for the prostate cancer gene network, the clustering coefficient is 0.733 as shown in figure 4. The figure 5 indicates that the genes are highly inter connected. The clustering coefficient of the breast cancer network is 0.721 is slightly lower than the prostate cancer gene network as shown in figure 6. The network is densely connected and the distribution of the number of shared neighbours is high as shown in figure 7In prostate and breast cancer, mutation of BRCA1 and BRCA2 genes are highly regulating and prognosis for the disease.

8. Summary, Conclusions and Future Work:

Advancement in the technological growth and the number of experiments to find out the associations between bio-molecular entities in the automated way has been developing to extract in the literature. To meet the essential and current issue to build the association networks based on rules and templates by assigning weights, BIB is developed. The precision is increased in the tool by allowing only the restricted parameters. BIB also supports predefined weighted links on each level. BIB takes raw text document as input in various formats with the nodes to build links between them. BIB is able to extract any number of abstracts in the pubmed search engine. A set of keywords are given to the tool and it scans the database automatically. Two regulatory networks for the breast cancer and the prostate cancer are constructed for testing and validate BIB. The result with good performance is produced by the tool BIB. Topological and biological analysis was conducted on both the network. Analysis of the prognosis cancer network showed that the mutation of the BRCA1 and BRCA2 plays avital role in the prostate and the breast cancer. The future work analyzes the other types of diseases

9. References:

1. D. Applet, et al. SRI International FASTUS system: Muc-6 test results and analysis. Proc. of the Message Understanding Conference, pp.237-248, 199.
2. A. Cohen, et al. Using co-occurrence network structure to extract synonymous gene and protein names from medline abstracts. BMC Bioinformatics, 6(1):103, 2005.
3. A. Domon, R. Aebersold. Mass spectrometry and protein analysis. Science; 312(5771):212-7. 2006.
4. P.B. Dobrokhotov, et al. Combining NLP and probabilistic categorisation for document and term selection for Swiss-Prot medical annotation. Bioinformatics, 19 Suppl 1, 2003.
5. I. Donaldson, et al. Prebind and textomy - mining the biomedical literature for protein-protein interactions using a support vector machine. BMC Bioinformatics, 4(1):11, 2003.
6. U. Fayyad and K. Irani. Multi-interval discretization of continuous-valued attributes for classification learning, Proc. of IJCAI, pp.1022-1029, 1993.
7. Gat-Viks, A. Tanay and R. Shamir. Modeling and analysis of het-erogeneous regulation in biological networks. Journal of Computational Biology, 11(6):1034-49, 2004.
8. P. Glenisson, et al. Evaluation of the vector space representation in text-based gene clustering. In Proc of PSB, pp.391–402, 2003.
9. P. Glenisson, et al. TXTGate: profiling gene groups with text-based information. Genome Biology, 5:R43+, 2004.
10. L. Huiqing, L. Jinyan and W. Limsoon. A Comparative Study on Feature Selection and Classification Methods Using Gene Expression Profiles and Proteomic Patterns. Genome Informatics 13:51-60, 2002.
11. D. Hanisch, et al. Playing biology's name game: identifying protein names in scientific text. Proc.of PSB, pp.403–414, Lihue, Hawaii, 2003.
12. L. Hirschman, A.A. Morgan and A.S. Yeh. Rutabaga by any other name: extracting biological names. Journal of Biomedica Informatics, 35(4):247–259, Aug. 2002.
13. V. Kulasingam and E.P. Diamandis. Strategies for discovering novel cancer biomarkers through utilization of emerging technologies. Nature Clinical Practice Oncology. 2008. (10):588-99.
14. Y. Lu and J. Han. Cancer classification using gene expression data. Information Systems; 28(4):243-68, 2003.
15. H. Liu and C. Friedman. Mining terminological knowledge in large biomedical corpora. Proc.of PSB, pp.415-426, 2003.
16. S. Novichkova, S. Egorov and N. Daraselia. Med Scan, a natural language processing engine for MEDLINE abstracts. Bioinformatics, 19(13):1699– 1706, Sept. 2003.
17. T. Ono, et al. Automated extraction of information on protein - protein interactions from the biological literature.

18. J. H. Park, et al. Protein Expr. Purif. 22, 60-6, 2001.
19. C. M. Perou, et al. Molecular portraits of human breast tumours. Nature. 2000; 406(6797):747-52.
20. S. Raychaudhuri, H. Schutze, and R. B. Altman. Using text analy-sis to identify functionally coherent gene groups. Genome Research, 12(10):1582–1590, 2002.
21. T. Sekimizu, H. S. Park, T. Jun'ichi. Identifying the interaction between genes and gene products based on frequently seen verbs in Medline abstracts. Genome Inform Ser Workshop, 9:62-71, 1998.
22. L. Tanabe and W.J. Wilbur. Tagging gene and protein names in biomed-ical text. Bioinformatics, 18(8):1124–1132, Aug. 2002.
23. J. P. Vert and M. Kanehisa. Extracting active pathways from gene expression data. Bioinformatics 2003, 19(Suppl 2):II238-II244.
24. R. Varshavsky, et al. Novel unsupervised feature filtering of biological data. Bioinformatics, 22, e507-e513, 2006.
25. B. Weigelt, et al. Molecular portraits and 70-gene prognosis signature are preserved throughout the metastatic process of breast cancer. Cancer Research. 2005; 65(20):9155-8.
26. M. Weeber, et al. Text-based discovery in biomedicine: the architecture of the DAD-system. Proceedings / AMIA ... Annual Symposium. AMIA Symposium, pages 903–907, 2000.
27. H. Xu, et al.Facilitating cancer research using natural language process-ing of pathology reports. Studies in health technology and informatics, 107(Pt 1):565–572, 2004.
28. Y. Xu, Z. Wang, Y. Lei, Y. Zhao, and Y. Xue. Mba: a literature mining system for extracting biomedical abbreviations. BMC Bioinformatics, 10(1):14, 2009.
29. A. Yakushiji, et al. Event extraction from biomedical papers using a full parser. Proc. of PSB. 6, 408-419 2001.
30. H. Yu and E. Agichtein. Extracting synonymous gene and protein terms from biological literature. Bioinformatics, 19 Suppl 1(suppl 1):i340–i349, July 2003.
31. H. Yu, et al. Automatic extraction of gene and protein synonyms from MEDLINE and journal articles. Proc AMIA Symp, pages 919–923, 2002.
32. G. Zhou, et al. Recognizing names in biomedical texts: a machine learning approach. Bioinformatics, 20(7):1178–1190, 2004.