Mining Scientific Data from Pub-Med Database

G .Charles Babu
Professor, Dept. of CSE
HITS, Bogaram

Dr. A.GOVARDHAN
Professor and Director of Evaluation
JNTU Hyderabad

***Abstract*- The continuous, rapidly growing volume of scientific literature and increasing diversification of inter-disciplinary fields of science and their answers to unsolved problems in medical and allied fields of science present a major problem to scientists and librarians. It should be recalled in this aspect that today as many as 4800 scientific journals exist in the internet of which some are online only. The list of journals located in subject citation indexes in Thomson Reuters can be obtained from the website. From researchers' point of view, the problem is amplified when we consider today's competition where we may not be able to spend time on experimental work merely because of already published information. Therefore, considering these facts partly and the volume of serials on the other, a study has been initiated in evaluating the scientific literature published in various journal sources. The scope of the study does not permit inclusion of all periodicals in the extensive fields of biology and hence a text mining routine was employed to extract data based on keywords such as bioinformatics, algorithms, genomics and proteomics. The wide availability of genome sequence data has created abundant opportunities, most notably in the realm of functional genomics and proteomics. This quiet revolution in biological sciences has been enabled by our ability to collect, manage, analyze, and integrate large quantities of data.**

I. INTRODUCTION

Scientific discovery in genomics and related biomedical disciplines increased the amount of data and information [3] whereas text mining provide useful tools to assist in the curation process [4] in extracting relevant information using automatic techniques, text-mining and information-extraction approaches [5]. Text literature is playing an increasingly important role in biomedical discovery.

Most text mining applications require the ability to identify and classify words, or multi-word terms, that authors use in an article. Several strategies have been tried to recognize biological entity names in articles.

Some methods rely on protein and gene databases to assemble dictionaries of protein names. Most of these methods were developed for abstracts, because abstracts are readily available for millions of articles (e.g., PubMed)[6]. To support data interpretation, bioinformatics tools were utilized to identify relevant information from literature databases. On the other hand, success has been achieved in developing biomedical literature mining software using semantic analyses to automatically extract information [7]. This method uses a pattern discovery algorithm to identify relevant keywords in abstracts.

In this paper, we present segregated information of journals that contain or publish data on bioinformatics, proteomics and genomics. Keyword searches in PubMed database with a list of countries and their involvement in research publications have also been presented. Most of the articles in bioinformatics journals are often technology centred, focusing on rapidly evolved techniques for analysis of sequences, structures and phylogenies [8]. Some articles emphasized on data integration and analysis with data-driven data management for integrative bioinformatics systems [9]

For the purposes of investigation, the evaluation was confined to the scientific journals hosted in PubMed only [1]. It is obvious that in compiling the information on the volume of data published in journals and that even the most careful check could not exclude the possibility of errors; however it is understood that the influence of such errors is minimal considering the huge volume of information in PubMed database.

II. MATERIALS AND METHODS

NCBI PubMed literature database was selected for the study. Initially a generalized search without any limits was employed to retrieve articles related to *bioinformatics* and *computational biology*. As search results indicated the presence of keyword anywhere in the article (title, abstract, address, keywords and text), a more stringent search criterion was employed to identify the number of articles appeared when a search performed either by individual or in combinations of keywords by limiting the search within Title and Abstract.

Title and abstract only search were considered in this study because the Title field in some articles refers to the most important keywords relative to the subject. Therefore, a validated disparity in information retrieved through text mining limited to Titles and Abstract terms only.

Articles belonging to bioinformatics, computational biology are explicitly reported in journals, some may have the term in Title/Abstract while some are representative of the field without keywords. Therefore, though a myriad of pertinent articles are located; preference is given to the two search techniques: Title and Abstract.

Title/Abstract is selected as limit to search the database in order to overcome false hits and to identify true positives. Therefore, an article is considered true positive only if the keyword is explicitly identified in Title/Abstract.

Records without abstracts are counted as true positives only if title contained the keywords [10]. Finally, year wise growth in number of articles in each field was carried to find out the enormous amount of data deposited in PubMed.

III. Results and Discussion

A generalized search in NCBI PubMed literature database, on March 28th 2012, using *bioinformatics* as keyword resulted in 97618 articles, of which 42.9 % are free full text and 15.9 % constitute review only articles. On the other hand, a search for *computational biology* articles in PubMed resulted in 79965 articles, of which 39.7 % and 17.9 % constitute free full text and review articles (see Table I).

TABLE I: DISTRIBUTION OF MAXIMUM NUMBER OF ARTICLES IN PUBMED DATABASE

Keyword	Total no. of articles	Free full text	Review
Bioinformatics	97618	41966	15551
Computational biology	79965	31820	14370

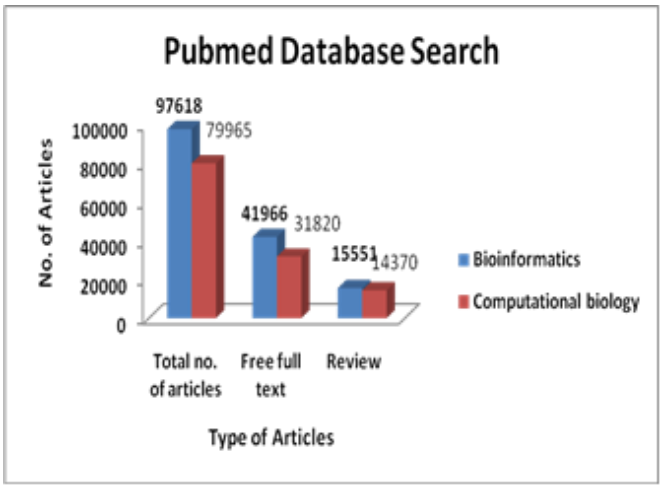

Fig. I PubMed database search with *bioinformatics* and *computational biology* as keywords.

Fig 1. illustrates the experimental results when the PubMed database was searched with *bioinformatics* and *computational* biology as keywords. The numbers over each bar represent the total number of articles from each field.

However, a more stringent search with Title/Abstract as key words revealed 11728 *bioinformatics* articles (11.9% of wild search as given in Table-1) and *computational biology* 2608 articles (3.6%) (See Table II).

TABLE II: DISTRIBUTION OF MAXIMUM NUMBER OF ARTICLES IN PUBMED WITH TITLE/ABSTRACT AS LIMIT

Keyword	Total no. of articles	Free full text	Review
Bioinformatics	11728	5664	1726
Computational biology	2934	1386	716

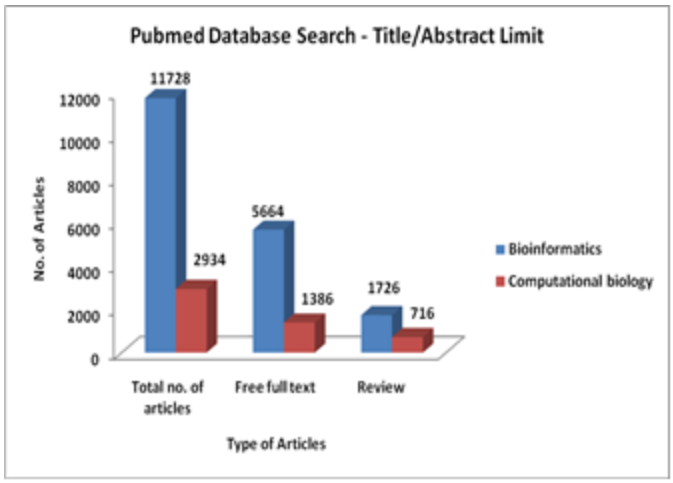

Fig. 2 PubMed database search with Title/Abstract limit for the two fields.

Fig 2. illustrates the experimental results when the PubMed database search with Title/Abstract limit for the two fields. The numbers over each bar represent the total number of articles from each field.

Boolean operator search enabled in PubMed database was used to extract combined keywords (See Table III). This shows the impact of these two ever-growing areas in sharing information and influencing the research publications.

TABLE III: NUMBER OF ARTICLES RETRIEVED IN A BOOLEAN SEARCH FROM PUBMED DATABASE

Boolean Operator	Total no. of articles	Free full text	Review
AND	80679	31873	14381
OR	97736	42018	15562

TABLE IV, Fig. 3 illustrate the annual data of the articles published on Bio-Informatics in the PubMed database.

TABLE IV : THE ANNUAL DATA OF THE ARTICLES PUBLISHED ON BIO-INFORMATICS IN THE PUBMED DATABASE.

Year	Total no. of articles	Free full text	Review
2000	1506	512	278
2001	2245	716	498
2002	3295	1209	826
2003	4674	1840	953
2004	6745	2989	1248
2005	8568	3802	1438
2006	9526	4106	1634
2007	10582	4753	1716
2008	11035	5371	1757
2009	13229	6782	1833
2010	14687	7202	2059
2011	14151	4908	1707

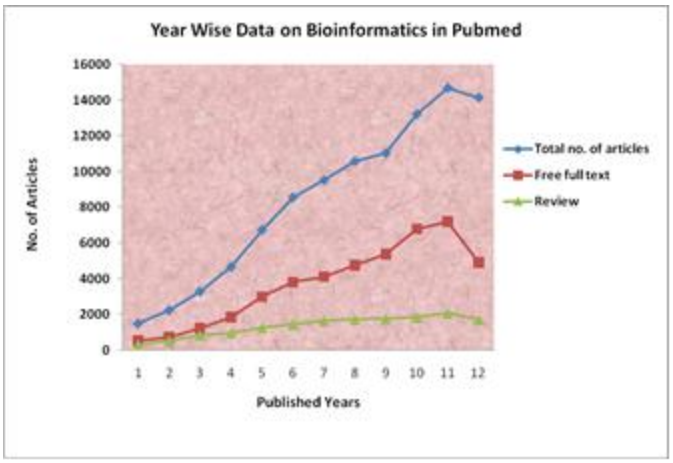

Fig. 3 Annual data on bioinformatics articles published in PubMed

IV. CONCLUSION

From the report, it can be emphasized that text mining is a useful alternative considering the enormous amount of data present in literature database such as PubMed. From an informatics perspective, integrated literature database like PubMed provides new insights for research in areas such as bioinformatics and computational biology. Though many research and review papers aimed at these two fields and as keywords are limited to Title/Abstract only, data suggests the phenomenal rise in number of papers in their respective fields. Therefore, from the work reported here, it can be suggested that scientific literature and approaches towards text mining have greater impact on data integration that support research for potential gains in life sciences and enable to understand the literature database applications.

REFERENCES

[1] http://science.thomsonreuters.com/mjl/

[2] JohnQuackenbush *"Open-Source Software Accelerates Bioinformatics."* Genome Biology 4: p 336, 2003.

[3] Hersh W, Bhupatiraju RT, Corley S *"Enhancing Access To The Bibliome: The TREC Genomics Track."* Medinfo p 773–777, 2004.

[4] Yeh AS, Hirschman L, Morgan AA *"Evaluation Of Text Data Mining For Database Curation"* Lessons Learned From The KDD Challenge Cup. Bioinformatics 19: Suppl 1:i331–339, 2003.

[5] Krallinger M, Erhardt RA, Valencia A *"Text-Mining Approaches In Molecular Biology And Biomedicine"* Drug Discovery Today 10 p 439-445, 2005 .

[6] Shi L, Campagne F. *"Building A Protein Name Dictionary From Full Text: A Machine Learning Term Extraction Approach."* BMC Bioinformatics 6 p 88, 2005.

[7] Chaussabel D.*"Biomedical Literature Mining: Challenges And Solutions In The 'Omics' Era"*. American Journal of Pharmacogenomics 4 p383-393, 2004.

[8] David B. Searls *"Using Bioinformatics In Gene And Drug Discovery".* Drug Discovery Today 5 p135-143, 2000.

[9] Jain, E. and Jain, K. *"Integrated Bioinformatics – High Throughput Interpretation Of Pathways And Biology"*. Trends Biotechnology 19,p 157–158, 2001.

[10] Kaveh G. Shojania, Lisa A. Bero,*"Taking Advantage Of The Explosion Of Systematics Reviews: An Efficient MEDLINE Search Strategy"*. Effective Clinical Practice,July/August 2001.

AUTHORS PROFILE

He received his M.Tech from JNTU in 1997 and B.Tech from KLCE in 1997.He is working as a Professor & Head in Dept. of CSE , Holy Mary Institutte of Technology & Science, Bogaram, Hyedarabad, India.

Received Ph.D in Computer Science & Engg from JNTU in 2003. M.Tech from JNTU in 1994.B.E from Osmaina University in 1992. He is working as a Director of Evaluation in JNTU Hyderabad.

He has published around 120 papers in various national and international Journals/conferences. His research of interest includes Data Mining, Information Retrieval & Search Engines.