Implementation of Data Mining in Bioinformatics for Extracting Enzymes Names from Literature

Dr. V. S. Gulhane [1], Prof. L. K. Gautam [2], V. G. Sahu[3]

[1,2,3]SGBAU, Amravati

Abstract

This paper highlights some of the basic concepts of bioinformatics and data mining. The major research areas of bioinformatics are highlighted. The application of data mining in the domain of bioinformatics is explained. It also highlights some of the current challenges and opportunities of data mining in bioinformatics.

Keywords: Datamining, Bioinnformatics, ProteinSequencesAnalysis, BioinformaticsTools.

1. INTRODUCTION

In recent years, rapid developments in genomics and proteomics have generated a large amount of biological data. Drawing conclusions from these data requires sophisticated computational analyses. Bioinformatics, or computational biology, is the interdisciplinary science of interpreting biological data using information technology and computer science. The importance of this new field of inquiry will grow as we continue to generate and integrate large quantities of genomic, proteomic, and other data. A particular active area of research in bioinformatics is the application and development of data mining techniques to solve biological problems. Analyzing large biological data sets requires making sense of the data by inferring structure or generalizations from the data. Examples of this type of analysis include protein structure prediction, gene classification, cancer classification based on microarray data, clustering of gene expression data, statistical modeling of protein-protein interaction, etc. Therefore, we see a great potential to increase the interaction between data mining and bioinformatics.

2. BACKGROUND THEORY

a. CLUSTERING

Cluster analysis or clustering is the task of grouping a set of objects in such a direction that objects in the same group are called a cluster. It is a primary task of explanatory data mining a common technique for statistical data analysis used in various fields including machine learning, pattern, picture analysis, data retrieval & Bioinformatics. In clustering method, targets of the dataset are grouped into clusters, in such a way that groups are almost different from each other and the objects in the same group or cluster are very alike to each other. Unlike Classification, in which previously defined set of categories are faced, but in Clustering there are no predefined set of classes which means that resulting clusters are not recognized before the implementation of clustering algorithm. These clusters are extracted from the dataset by grouping the objects in it.

b. TEXT MINING

Text mining refers to the process of deriving high quality of information from the text documents.It is a challenging task to help the users in finding what the user's actually want from the number of text documents. It is quite tough to deal with the text which is in unstructured form. The main purpose of the text mining is to finding the interesting information from the natural language text. To answer the complicated questions and to do the web searches with intelligence is the main aim of the text mining tools. Text mining uses the automation methods for achieving the common knowledge which is available in text documents. Text mining process is as shown in following fig.1

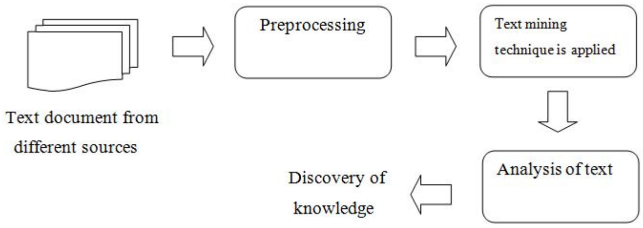

Text document from different sources
Preprocessing
Text mining technique is applied
Analysis of text
Discovery of knowledge

Figure 1 Text mining process.

c. DATA MINING IN BIOINFORMATICS

The two "high-level" primary goals of data mining, in practice, are prediction and description. The main tasks well suited for data mining, all of which involves mining meaningful new patterns from the data, are:

- *Classification:* Classification is learning a function that maps (classifies) a data item into one of several predefined classes. The Classifier Algorithm are: Support Vector Machine (SVM) is one of the machine learning techniques for Text Categorization, Naive Bayes Classifier, K-Nearest Neighbor.
- *Estimation:* Given some input data, coming up with a value for some unknown continuous variable.
- *Prediction:* Same as classification & estimation except that the records are classified according to some future behavior or estimated future value).
- *Association rules*: Determining which things go together, also called dependency modeling.
- *Clustering:* Segmenting a population into a number of subgroups or clusters.
- *Description & visualization:* Representing the data using visualization techniques.
- Learning from data falls into two categories: directed (“supervised”) and undirected (“unsupervised”) learning.

The first three tasks – classification, estimation and prediction – are examples of supervised learning. The next three tasks – association rules, clustering and description & visualization – are examples of unsupervised learning. In unsupervised learning, no variable is singled out as the target; the goal is to establish some relationship among all the variables. Unsupervised learning attempts to find patterns without the use of a particular target field. The development of new data mining and knowledge discovery tools is a subject of active research. One motivation behind the development of these tools is their potential application in modern biology.

Applications of data mining to bioinformatics include gene finding, protein function domain detection, function motif detection, protein function inference, disease diagnosis, disease prognosis, disease treatment optimization, protein and gene interaction network reconstruction, data cleansing, and protein sub-cellular location prediction. For example, microarray technologies are used to predict a patient’s outcome. On the basis of patients’ genotypic microarray data, their survival time and risk of tumor metastasis or recurrence can be estimated. Machine learning can be used for peptide identification through mass spectroscopy. Correlation among fragment ions in a tandem mass spectrum is crucial in reducing stochastic mismatches for peptide identification by database searching. An efficient scoring algorithm that considers the correlative information in a tunable and comprehensive manner is highly desirable.

d. WHAT ARE ENZYMES?

Enzymes are Protein Catalyst that increase the velocity of chemical reaction and are not consumed during the reaction they catalyze. Enzymes do not cause reaction to take place, but they greatly enhance the rate of reaction that would proceed much slower in their absence. Without enzymes, reaction would be to slow too maintain life.

The International Union of Bio-Chemistry (IUB) classified enzymes into six main types

- Oxidoreductases
- Transferases
- Hydrolases
- Lyases
- Isomerases
- Ligases

3. LITERATURE REVIEW

The term bioinformatics was coined by Paulien Hogewegn in 1979 for the study of informatics processes in biotic systems. It was primary used since late 1980s has been in genomics and genetics, particularly in those areas of genomics involving large-scale DNA sequencing. Bioinformatics can be defined as the application of computer technology to the management of biological information. Bioinformatics is the science of storing, extracting, organizing, analyzing, interpreting and utilizing information from biological sequences and molecules. It has been mainly fuelled by advances in DNA sequencing and mapping techniques. Over the past few decades rapid developments in genomic and other molecular research Technologies and developments in information technologies have combined to produce a tremendous amount of information related to molecular biology. The primary goal of bioinformatics is to increase the understanding of biological processes. Some of the grand area of research in bioinformatics includes:

- *Sequence analysis*

Sequence analysis is the most primitive operation in computational biology. This operation consists of finding which part of the biological sequences are alike and which part differs during medical analysis and genome mapping processes. The sequence analysis implies subjecting a DNA or peptide sequence to sequence alignment, sequence databases, repeated sequence searches, or other bioinformatics methods on a computer.

- *Genome Annotation*

In the context of genomics, annotation is the process of marking the genes and other biological features in a DNA sequence. The first genome annotation software system was designed in 1995 by Dr. Owen White.

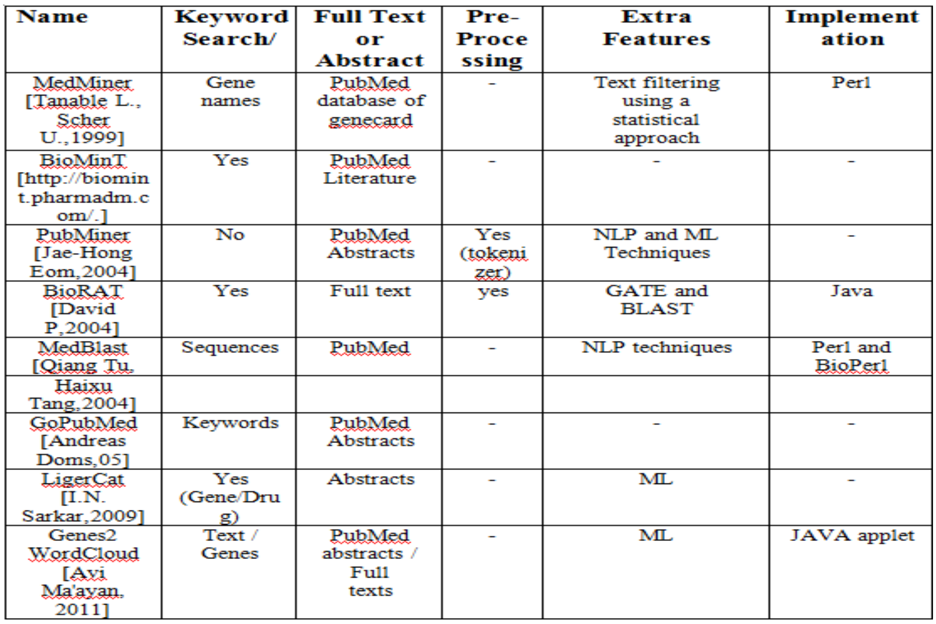
- *Analysis of gene expression*

The expression of many genes can be determined by measuring mRNA levels with various techniques such as microarrays, expressed cDNA sequence tag (EST) sequencing, serial analysis of gene expression (SAGE) tag sequencing, massively parallel signature sequencing (MPSS), or various applications of multiplexed in-situ hybridization etc. All of these techniques are extremely noise-prone and subject to bias in the biological measurement. Here the major research area involves developing statistical tools to separate signal from noise in high-throughput gene expression studies.

- *Protein structure prediction*

The amino acid sequence of a protein (so-called, primary structure) can be easily determined from the sequence on the gene that codes for it. In most of the cases, this primary structure uniquely determines a structure in its native environment. Knowledge of this structure is vital in understanding the function of the protein. For lack of better terms, structural information is usually classified as secondary, tertiary and quaternary structure. Protein structure prediction is one of the most important for drug design and the design of novel enzymes. A general solution to such predictions remains an open problem for the researchers.

There are number of tools designed for Text Mining applied to bioinformatics problems. A summary of some of the presented tools, adapted from [29], is presented in figure 2



Name	Keyword Search/	Full Text or Abstract	Pre-Processing	Extra Features	Implementation
MedMiner [Tanable L., Scher U.,1999]	Gene names	PubMed database of genecard	-	Text filtering using a statistical approach	Perl
BioMinT [http://biomint.pharmadm.com/.]	Yes	PubMed Literature	-	-	-
PubMiner [Jae-Hong Eom,2004]	No	PubMed Abstracts	Yes (tokenizer)	NLP and ML Techniques	-
BioRAT [David P,2004]	Yes	Full text	yes	GATE and BLAST	Java
MedBlast [Qiang Tu,	Sequences	PubMed	-	NLP techniques	Perl and BioPerl
Haixu Tang,2004]					
GoPubMed [Andreas Doms,05]	Keywords	PubMed Abstracts	-	-	-
LigerCat [I.N. Sarkar,2009]	Yes (Gene/Drug)	Abstracts	-	ML	-
Genes2 WordCloud [Avi Ma'ayan, 2011]	Text / Genes	PubMed abstracts / Full texts	-	ML	JAVA applet

Figure 2 Summary of Tools

4. Analysis of Problem

It is of capital importance for every researcher to be aware of the work that has been done in his research area. With the advent of the Internet, the amount of information available to everyone is usually overwhelming. One might think that researchers could now get easy access to all related work. There is, however, a very difficult problem that needs to be solved before we reach that situation. Accessing the right information amidst the overwhelming amount of documents available in the Internet is quite difficult, in most cases. The volume of research publications, in almost all areas of knowledge, has been growing at a phenomenal rate. Nowadays, most publications are available on the Internet, some completely and some partially. The overload of research publications can hinder researchers due to the time spent looking for the real interesting or relevant" publications. Usually, to find these publications a researcher uses the traditional keyword-based search engines and, as a result, faces a huge list of publications to read with a large number of irrelevant publications. To tackle this problem, research in Text Mining and Knowledge Extraction has been applied to literature mining to help researchers to identify the most relevant publications out of a great amount resulting from simple search strategies. Although the traditional search engines are quite useful for specific queries, when applied to more complex searches they have strong limitations. For this reason, scientists have focused their attention on Text Mining techniques (information retrieval, information extraction and data mining) as a way to solve this problem. Text Mining helps gathering, maintaining, interpreting and discovering knowledge needed for research in a efficient way. By adding meaning to text, these techniques produce a more structured analysis of textual knowledge than simple word searches.

In the figure 2 for e.g 1) PubMiner, This system allows the visualization of the results in a graph, where the nodes represent the names of genes and of proteins and the arcs represent the possible interactions . 2) MedBlast uses BLAST to find abstracts linked to homologous sequences but can also find abstracts with the gene and organism name derived from annotated protein entries.

5. Proposed Work

Our proposed system uses K-Mean algorithm to extract the Enzymes names from the document and their properties from the literature and store them in the database for analysis.

Our work is motivated by the approach:

- Collect paper from the PubMed Database.
- Investigate & Apply Different retrieval techniques that will be used for the retrieval of relevant documents from large repositories.
- Construct K-Mean Classifier
- Use Classifier to get relevant Enzyme form File Database
- Display Count From relevant file Database.

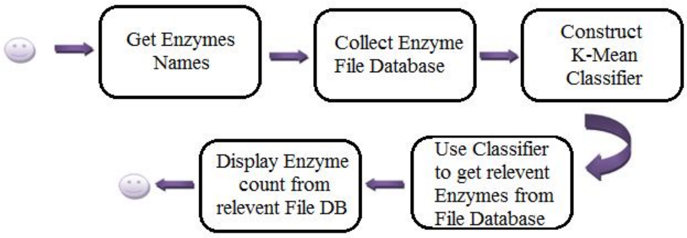

Get Enzymes Names
Collect Enzyme File Database
Construct K-Mean Classifier
Use Classifier to get relevent Enzymes from File Database
Display Enzyme count from relevent File DB

Figure 3 Proposed Model for Extracting Enzyme name from Literature

a. Tf-idf Weighting

Before being able to run k-means on a set of text documents, the documents have to be represented as mutually comparable vectors. To achieve this task, the documents can be represented using the tf-idf score. The tf-idf, or term frequency-inverse document frequency, is a weight that ranks the importance of a term in its contextual document corpus. Term frequency is calculated as normalized frequency, a ratio of the number of occurrences of a word in its document to the total number of words in its document. It's exactly what it sounds like, and conceptually simple, and can be thought of somewhat like a fraction of the document that is a particular term. The division by the document length prevents a bias toward longer documents by "normalizing" the raw frequency into a comparable scale. The inverse document frequency is the log (no matter the base, because it scales the function by a constant factor, leaving comparisons unaffected) of the ratio of the number of documents in the corpus to the number of documents containing the given term. Inverting the document frequency by taking the logarithm assigns a higher weight to rarer terms.

Multiplying together these two metrics gives the tf-idf, placing importance on terms frequent in the document and rare in the corpus.

6. Result

We have used data sets, composed by more than 90 relevant papers related to enzymes name and uses k-mean algorithm to extract enzyme name and found that k-mean algorithm take less time on larger dataset.

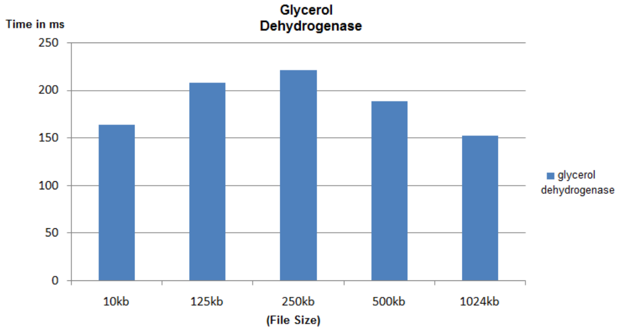


Figure 4 Time Complexity for Searching Enzyme using K-mean

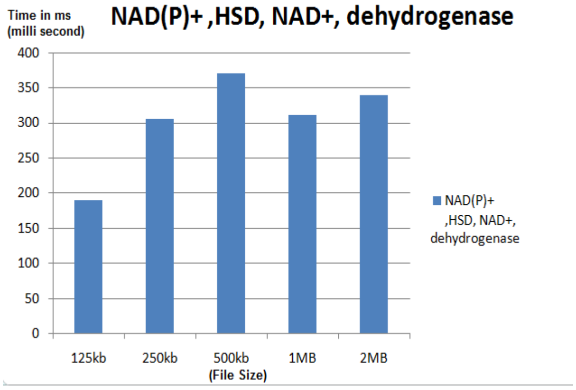


Figure 5 Time Complexity for Searching no. of Enzymes using K-mean

In the above Figure 2 shows time complexity for searching enzyme, as the size of document increases the k-mean takes less time for searching in the relevant document. In Figure 3 shows time complexity for searching no. of enzyme, as the document size increases it takes more time for searching.

7. Conclusion

K means algorithm gave good results for large data-sets. In our case we have worked on k means clustering of database. The future work may benefit with less computational time as compared to previous work as database with records are increasing day by day and there is a need of data clustering on large databases.

References

[1] Aluru, S., ed. (2006), “Handbook of Computational Molecular Biology. Chapman & Hall/Crc”.

[2] Baxevanis, A.D.; Petsko, G.A.; Stein, L.D. and Stormo, G.D., eds. (2007), “Current Protocols in Bioinformatics”. Wiley.M. Clerc, “The Swarm and the Queen: Towards a Deterministic and Adaptive Particle Swarm Optimization,” In Proceedings of the IEEE Congress on Evolutionary Computation (CEC), pp. 1951-1957, 1999. (conference style)

[3] Berson, Alex, Smith, Stephen and Threaling, Kurt, “Building Data Mining Application for CRM”, Tata McGraw Hill.

[4] Gilbert, D. (2004), “Bioinformatics software resources. Briefings in Bioinformatics, Briefings in Bioinformatics”.

[5] Han and Kamber (2006). “Data Mining concepts and techniques, Morgan Kaufmann Publishers”.

[6] Hirschman, Lynette; C. Park, Jong; T., Junichi, Wong, L. and H. Wu., Cathy (2002), “Accomplishments and challenges in literature data mining for biology”, BIOINFORMATICS REVIEW, Vol. 18 no. 12, 1553–1561

[7] Hand, D. J.; Mannila, H. and Smyth, P. “Principles of Data Mining, MIT Press”.

[8] Jiong, Lei Liu; Yang, A. and Tung, K. H (2005). "Data Mining Techniques for Microarray Datasets", Proceedings of the 21st International Conference on Data Engineering (ICDE 2005).

[9] Lee, Kyoungrim. (2008), "Computational Study for Protein-Protein Docking Using Global Optimization and Empirical Potentials", Int. J. Mol. Sci. 9, 65-77.

[10] Luis, T.; Chitta; B. and Kim, S. (2008). "Fuzzy c-means clustering with prior biological knowledge, Journal of Biomedical Informatics".

[11] Liu, H.; Li, J. and Wong, L. (2005), "Use of Extreme Patient Samples for Outcome Prediction from Gene Expression Data", Bioinformatics, vol. 21, no. 16, pp. 3377–3384

[12] Mewes, H.W.; Frishman, D.; X.Mayer, K. F.; Munsterkotter, M., Noubibou , O.; Pagel, P. and Rattei, T. (2006)]. Nucleic Acids Research, 34, D169.

[13] Mount, D. W. (2002), "Bioinformatics: Sequence and Genome Analysis Spring Harbor Press".

[14] Nayeem, Akbar; Sitkoff, Doree, and Krystek, Jr., Stanley. (2006)**,** "A comparative study of available software for highaccuracy homology modeling: From sequence alignments to structural models", Protein Sci. April; 15(4): 808–824

[15] N., Cristianini and M., Hahn. (2006), "Introduction to Computational Genomics, Cambridge University Press". ISBN 0-5216- 7191-4.

[16] Pevzner, P. A. (2000). "Computational Molecular Biology: An Algorithmic Approach The MIT Press".

[17] Richard, R.J. A. and Sriraam, N. (2005), "A Feasibility Study of Challenges and Opportunities in Computational Biology: A Malaysian Perspective", American Journal of Applied Sciences 2 (9): 1296-1300.

[18] Soinov, L. (2006). "Bioinformatics and Pattern Recognition Come Together. Journal of Pattern Recognition Research", (JPRR), Vol 1 (1) p.37-41

[19] SJ, Wodak and Janin, J. (1978), "Computer Analysis of Protein-Protein Interactions". Journal of Molecular Biology 124 (2): 323–42.

[20] Tang, Haixu and Kim, Sun, "Bioinformatics: mining the massive data from high throughput genomics experiments, analysis of biological data": a soft computing approach, edited by Sanghamitra Bandyopadhyay, Indian Statistical Institute, India.

[21] Yang, Qiang, "Data Mining and Bioinformatics": Some Challenges, http://www.cse.ust.hk/~qyang

[22] Zaki , J.; Wang , T.L. and Toivonen, T.T. (2001), "BIOKDD01: Workshop on Data Mining in Bioinformatics".

[23] Zhang, Yanqing; C., Jagath, Rajapakse, "Machine Learning in Bioinformatics", Wiley, ISBN: 978-0-470-11662-3

[24] Sophia Ananiadou, Douglas B B. Kell, and Jun-Ichi I. Tsujii, "Text mining and its potential applications in systems biology", Trends Biotechnol, 2006.

[25] (http://en.wikipedia.org/wiki/Vector_space_model)

[26] (http://en.wikipedia.org/wiki/Tfidf)

[27] (http://en.wikipedia.org/wiki/Text_corpus)

[28] http://jonathanzong.com/blog/2013/02/02/kmeansclusteringwithtfidfweights.

[29] Dieng-Kuntz R., Khelif K., and Barbry P. Mining biomedical texts to generate semantic annotations, 2007.