IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified

Vol. 6, Issue 3, March 2017

Review on Data Mining Techniques in Bioinformatics for Extracting Enzymes Names from Literature

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

HOD, Sipna College of Engg. and Tech., Amravati, India[1]

Lecturer, Sipna College of Engg and Tech, Amravati, India [2]

Sipna College of Engg and Tech, Amravati, India [3]

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, Bioinformatics, Protein Sequences Analysis, Bioinformatics Tools.

I. 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.

II. LITERATURE SURVEY

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:

A. 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.

B. 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.

C. 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.

D. 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.

III. 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:

A. Classification: Classification is learning a function that maps (classifies) a data item into one of several predefined classes. The Classifier Algorithm is: Support Vector Machine (SVM) is one of the machine learning techniques for Text Categorization, Naive Bayes Classifier, K-Nearest Neighbor.
B. Estimation: Given some input data, coming up with a value for some unknown continuous variable.
C. Prediction: Same as classification & estimation except that the records are classified according to some future behaviour or estimated future value).
D. Association rules: Determining which things go together, also called dependency modelling.
E. Clustering: Segmenting a population into a number of subgroups or clusters.
F. Description & visualization: Representing the data using visualization techniques.
G. 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.

IV. APPLICATION OF DATA MINING

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 tumour 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.

V. CONCLUSION

Bioinformatics and data mining are developing as interdisciplinary science. Data mining approaches seem ideally suited for bioinformatics, since bioinformatics is data-rich but lacks a comprehensive theory of life’s organization at the molecular level. However, data mining in bioinformatics is hampered by many facets of biological databases, including their size, number, diversity and the lack of a standard ontology to aid the querying of them as well as the heterogeneous data of the quality and provenance information they contain. Another problem is the range of levels the domains of expertise present amongst potential users, so it can be difficult for the database curators to provide access mechanism appropriate to all. The integration of biological databases is also a problem. Data mining and bioinformatics are fast growing research area today. It is important to examine what are the important research issues in bioinformatics and develop new data mining methods for scalable and effective analysis.

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

DOI 10.17148/IJARCCE.2017.6367

[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

DOI 10.17148/IJARCCE.2017.6367