Available online at www.elixirpublishers.com (Elixir International Journal)

Pharmacy

Elixir Pharmacy 55A (2013) 13235-13238

ISSN: 2229-712X

A comparative study on blood glucose spectra values to predict dominating patterns using bio-data mining

Monisha Swaminathan, S. Monisha, Dhanya Biju and Roshni Pal
Department of Bio-Medical Engineering, Alpha College of Engg, Chennai, T.N, India.

ARTICLE INFO

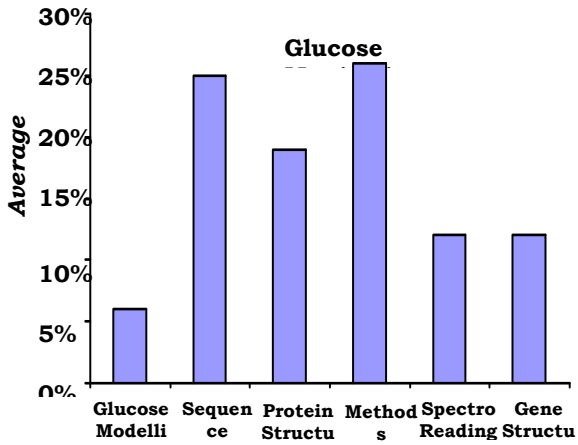
Article history:
Received: 10 September 2012;
Received in revised form:
1 February 2013;
Accepted: 19 February 2013;

Keywords
Blood Glucose Measuring,
Spectra Values,
Pattern Based Classification.

ABSTRACT
Pattern Mining is the process of extracting or mining the patterns from very large amount of biological datasets. Utilization of Data mining algorithms can reveal biological relevant associations between different genes and gene based expression. In Data Mining, several techniques are available for predicting frequent patterns. One among the technique is association rule mining algorithm; which can be applied for solving the crucial problems faced in the field of biological science. From the literature, various algorithms have been employed in generating frequent patterns for distinct application. These algorithms have some limitations in predicting frequent patterns, such as space, time complexity and accuracy. In order to overcome these drawbacks, the study is made on existing algorithms for generating frequent patterns from the biological sequences. The literature survey gives a significant number of methods were generated for predicting associative patterns. The proposed system has to be developed for solving problems in archiving glucose samples of many patients. Biological sequence may be a collection of DNA sequence, Gene expression sequence or Protein sequence for a specific viral disease. Amino acids are the building blocks of proteins. Proteins are organic compounds made up of amino acids arranged in a linear chain and folded into a globular form. And it will also satisfy some factors such as: time complexity, space and predict accurate solution to the required problem. With the help of these three factors into consideration and efficient algorithm can be identified for predicting the dominating amino acids for any kind of specific biological implication.

Introduction

Data mining refers extracting or "mining" knowledge from large amount of data. It is defined as "the process of Discovering meaningful new correlations, patterns and trends by digging into large amounts of data stored in warehouses". Data Mining is called as Knowledge Discovery in Databases (KDD).As data sets have grown in size and complexity, the modern technologies of computers, networks and sensors have made data collection and organization much easier. However, the captured data needs to be converted into information and knowledge to become more useful. Data Mining is the entire process of applying computer-based methodology, including new techniques for knowledge discovery from data. Data Mining approaches seem ideally suited for Biological Data Mining, since it is data-rich, but lacks a comprehensive theory of life's organization at the molecular level. The extensive databases of biological information create both challenges and opportunities for development of novel KDD methods. Mining biological data helps to extract useful knowledge from massive datasets gathered in biology, and in other related life sciences areas such as medicine and neuroscience. A crucial challenge in the future of bioinformatics involves putting that data to work. Now life scientists hope to plan large experiments, collect lots of data, analyse it, compare data between experiments, and eventually combine all of that information to improve basic theories, biotechnology, and medicine. The average trends in bioinformatics research areas are shown in the following Fig.1.



RELATED WORK AND EXISTING MODEL

In the past two decades the challenges faced by the research pupil in the field of biomedical is for an explosive growth of biomedical data (i.e.,ranging from those collected in pharmaceutical studies and diabetes) investigations to those identified from genomics and proteomic research by discovering sequential patterns, gene functions, and protein-protein interactions. The rapid progress of biotechnology and bio data analysis methods has led to the emergence and fast growth of promising new field coined as Bio-Data Mining. Applications

Tele:
E-mail addresses: bmestudents@yahoo.com

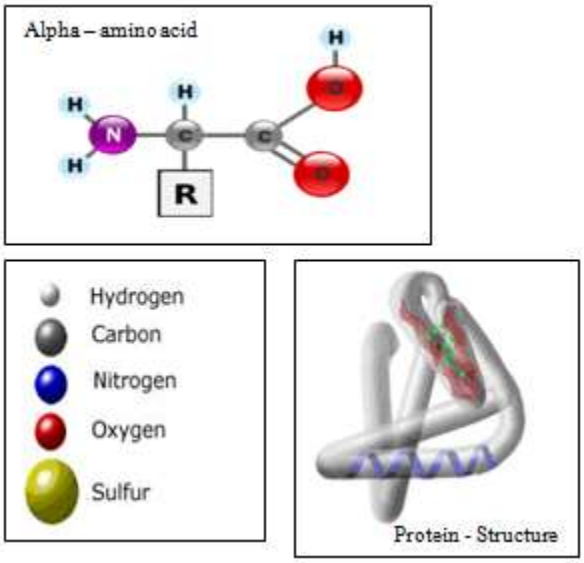
of data mining to Bio-Data Mining includes gene finding, protein function domain detection, function motif detection, protein function inference, disease diagnosis, disease prognosis, disease treatment optimization, protein and gene interaction network reconstruction, data cleansing, and protein sub-cellular location prediction.

Amino acids

Amino acids play central roles both as building blocks of proteins and as intermediates in metabolism. The chemical properties of the amino acids of proteins determine the biological activity of the protein. Proteins not only catalyse all or most of the reactions in living cells, they control virtually all cellular process. The general structure of a α-amino acid is depicted in Fig.2, Which represents the amino group on the left, the carboxyl group on the right and R a side chain to each amino acid.

Protein Structure

Proteins are an important class of biological macromolecules present in all biological organisms, made up of elements such as carbon, hydrogen, nitrogen, phosphorous, oxygen and sulphur. The elements of a protein and the tertiary structure of protein are depicted in Fig. 3. There are four distinct aspects of a protein structure such as Primary structure, Secondary structure, Tertiary structure and Quaternary structure.



Amino acids are the building blocks of proteins. Proteins are large molecules composed of one or more chains of amino acids in a specific order. The order is determined by the base sequence of nucleotides in the gene that codes the protein Amino acids combine in a condensation reaction that releases water and the new amino acid residue that is held together by a peptide bond. Proteins are defined by their unique sequence of amino acid residues; amino acids can be linked in varying sequences to form a vast variety of proteins. Twenty standard amino acids are used by cells in protein biosynthesis, and these are specified by general genetic code. The Table I illustrates the list of essential amino acids and Table II shows the list of non- The one-letter and three-letter codes for amino acids used in the knowledgebase are those adopted by the commission on Biochemical Nomenclature of the IUPAC-IUB.

Bioinformatics Data Sets for Diabetes Patients.

Bioinformatics is the science and technique for organizing and analysing biological data. Bioinformatics is conceptualizing biology in terms of molecules and applying informatics techniques to understand and organize the information associated with these molecules, on a large scale.

Table I. List of essential amino acids

Amino Acid	3-Letter	1-Letter
Arginine	Arg	R
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
[illegible]	[illegible]	[illegible]
Methionine	Met	M
Phenylalanine	Phe	F
Threonine	Thr	T
Tryptophan	Trp	W
Valine	Val	V

TABLE II. LIST OF NON-ESSENTIAL AMINO ACIDS

Amino Acid	*3-Letter*	*1-Letter*
Alanine	Ala	A
Asparagine	Asn	N
Aspartate	Asp	D
Cysteine	Cys	C
Glutamate	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Proline	Pro	P
Serine	Ser	S
Tyrosine	Tyr	Y

The Swiss-Prot group develops, annotates and maintains the UniProtKB/Swiss-Prot protein sequence database, the most widely used protein information resource in the world. The Bioinformatics Database group also develops and maintains other databases including PROSITE, a database of protein families and domains, and ENZYME, a database of enzyme nomenclature. The group also co-heads the development and maintenance of the ExPASy proteomics website. UniProtKB consists of two sections namely,

- UniProtKB/Swiss-Prot. - Protein sequence database is manually annotated and is reviewed
- UniProtKB/TrEMBL. - Protein sequence database is automatically annotated and is not reviewed

Biomedical Process for Glucose Isolation & Mining

Association Rule Mining in Genomics proposed by Anandhavalli [7] has two major goals for analysing massive genomic data for glucose patients: (i) To determine how the expression of any particular gene might affect the expression of other genes, (ii) To determine what genes are expressed as a result of certain cellular conditions using association and clustering concepts. The author selected an efficient algorithm to facilitate these analysis, the number of passes were not a major factor to be considered. Finally the author has concluded that the number of genes in one single transaction was very large. The Apriori Property of Sequence Pattern Mining with Wildcard Gaps proposed by Fan Min [8] has an alternative definition of the number of offset sequences by adding a number of dummy characters at the tail of sequence. Data miners designed pattern growth algorithms to obtain frequent patterns with periodical wildcard gaps, where the pattern frequency was defined as the number of pattern occurrences divided by the number of offset sequences. With the proposed definition, these uninteresting patterns were no longer frequent and the Apriori property holds good, hence Apriori algorithm can mine all frequent patterns with minimal endeavor.

Chien-Hua Wang [9] has proposed Fuzzy Frequent Pattern growth (FFP-growth) to derive from fuzzy association rules. In this approach first, fuzzy partition methods have been applied to decide a membership function of quantitative value for each transaction item and then implement FFP-growth to deal with

the process of data mining. This FFP-growth need not to generate candidate item-sets and improves the efficiency of repetitious database scanning. Compared with other pattern mining algorithms the proposed one has achieved better executive efficiency.

A Vector Operation Based Fast Association Rules Mining Algorithm proposed by Zhi Liu [10] has proposed a vector operation based association rule mining algorithm (V_Apriori algorithm) which solves the problem of multiple scanning of the database. With this algorithm the transaction database need to be scanned only one time to generate the boolean matrix which is stored in bit mode, so that memory space is greatly saved. The frequent itemsets are predicted through the AND operation on the vectors in the matrix, and the number of the candidates itemsets were reduced significantly. Compared with the traditional Apriori algorithm the new V_Apriori algorithm has been improved on time and space factor. The author [11] has proposed a data mining system for the assessment of heart event related risk factors using association analysis based on the apriori algorithm. The events investigated were: myocardial infarction (MI), percutaneous coronary intervention (PCI), and coronary artery bypass graft surgery (CABG).

There are several factors that contribute to the development of a coronary heart event. These risk factors may be classified into two categories: (i) Not-modifiable includes factors that cannot be altered by intervention such as age, gender, patient family history and genetic attributes,(ii) Modifiable currently includes smoking, elevated cholesterol , hypertension, and diabetes. This can be monitored / lowered with the doctor's advice and medications so that the incidence of heart episodes can be lowered. The risk of Type 1&2 (diabetes) of a patient may be reduced through a proper control of these factors by EUROASPIRE I, II, and III surveys. Thus, data mining could help in the identification of high and low risk subgroups of patients, a decisive factor for the selection of therapy, i.e. medical or surgical.

PROPOSED – GLUCOSE SPECTRA SYSTEM

From the existing methodologies it is not easy to solve crucial problems faced in the field of biological science in case of emergencies. In the Bio-medical field huge volumes of data are predominantly increasing time to time. The existing algorithm doesn't satisfy the researchers or find the solution to the real time problem under various situations due to the time complexity, space and accuracy. These problems could be rectified by the proposed system which is illustrated in Fig.4 and Fig.5 in two steps;

- Find an efficient algorithm based on three factors such as time, memory and precision.
- Find the dominating frequent patterns from biological sequences using Association rule mining algorithms.

A significant number of methods have addressed the clustering of protein sequences and most of them can be categorized in three major groups: hierarchical, graph-based and partitioning methods.

Among the various sequence clustering methods, hierarchical and graph-based approaches have been widely used. Although partitioning clustering techniques are extremely used in other fields, few applications have been found in the field of protein sequence clustering. It is not fully demonstrated if partitioning methods can be applied to protein sequence data and if these methods can be efficient compared to the published methods.

Association rule mining algorithms are generally meant for mining frequent item-sets.

The frequent item-sets could be generated in two ways, such as follows;

- Mining frequent patterns using candidate generation;
- Apriori algorithm
- Dynamic Item-set:

a) Counting algorithm

Mining frequent patterns without candidate

- Frequent Pattern-Tree growth algorithm

The proposed model PROCAD in Fig. 1.0 illustrate the comparative status of an efficient algorithm and in turn to discover dominating patterns from known to unknown sequence. Fig. 5 focuses on efficiently applying Association rule mining algorithms over protein sequence datasets, and predicts frequent patterns with candidate generation and without candidate generation from the comparative study which results in identifying an efficient algorithm as illustrated in Fig. 6.

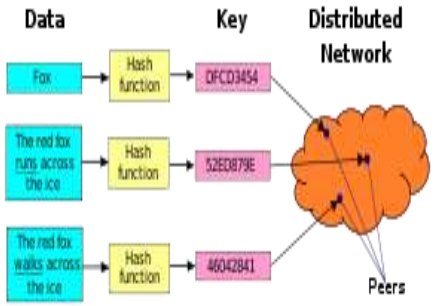


Figure. 1.0 Peer to Peer in Distributed Hospital Network

Distributed hash tables (DHTs) are a class of decentralized distributed systems that provide a lookup service similar to a Responsibility for maintaining the mapping from keys to values is distributed among the nodes, in such a way that a change in the set of participants causes a minimal amount of disruption. This allows DHTs to scale to extremely large numbers of nodes and to handle continual node arrivals, departures, and failures. DHTs form an infrastructure that can be used to build peer-to-peer networks. Notable distributed networks that use DHTs include BitTorrent's distributed tracker, the Kad network, the Storm botnet, YaCy, and the Coral Content Distribution Network.

DHT-based networks have been widely utilized for accomplishing efficient resource discovery [4][5] for grid computing systems, as it aids in resource management and scheduling of applications. Resource discovery activity involves searching for the appropriate resource types that match the user's application requirements. Recent advances in the domain of decentralized resource discovery have been based on extending the existing DHTs with the capability of multi-dimensional data organization and query routing.

Majority of the efforts have looked at embedding spatial database indices such as the Space Filling Curves (SFCs) including the Hilbert curves, Z-curves, k-d tree, MX-CIF Quad tree and R*-tree for managing, routing, and indexing of complex Grid resource query objects over DHT networks. Spatial indices are well suited for handling the complexity of Grid resource queries. Although some spatial indices can have issues as regards to routing load-balance in case of a skewed data set, all the spatial indices are more scalable in terms of the number of hops traversed and messages generated while searching and routing Grid resource queries.

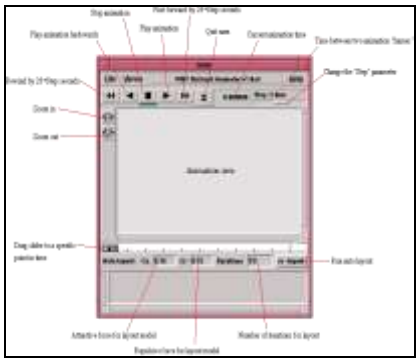


Figure. 2.0 Individual Layered Model

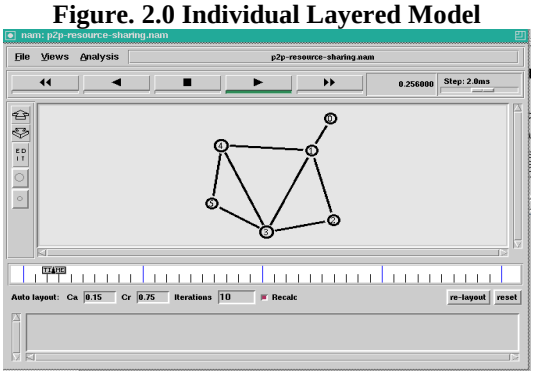


Figure. 3.0 Resource Sharing

Figure. 4.0 Layered model in system (Protein Structure)

From the study of literature, it is known that an efficient algorithm is required to predict frequent pattern. The survey concludes that frequent item-sets could be generated from a clustered protein sequence which causes the viral disease in human. Among the generated frequent item-sets few amino acids may be found to be strongly associated. Using the results retrieved from the clustered protein sequence, a focus has to be given on the most dominating amino acids by forming association rules. Various protein sequences could be applied on the proposed system which is being developed for identifying the dominating amino acids.

Further investigation will be involved by mining frequent item-sets with and without candidate generation and their results will be compared with the predicted results. In future this work could be extended to other protein sequence which causes other viral disease like flue, Dengue Fever, viral fever, swine flu, etc. Finally this work helps and it is more beneficial in preparing medicines to cure the disease caused by these viral infections during the case of emergency.

CONCLUSIONS

In this Project, we presented an analytic framework to evaluate the latencies associated with file replication in P2P systems. The main contribution of the project is a Unstructured p2p to evaluate the file transfer delay at the peers. Our model accounts for the query search times and peer characteristics like the number of simultaneously allowed downloads at a peer, file popularity, number of copies of the file, etc.

REFERENCES

- Bashed Abdullah, SantosHenrique, Validatin R Tam: Interface with CVD and the cardio stheth theory, Global Journal of Bio-science and cardio vascular disease,Vol.10, pp.68–73(2008).
- Claraand Zannad F, Briancon S, Juilliere Y, Mertes PM, Villemot JP: Telephone reminders improve adolescent clinic attendance:a randomized controlled trial Pediatric Child Health care, pp.79–83(2008).
- Christa Amrita Pal, Victor W, Pratim Lina Datta: Telemedicine Diffusion in a developing country for AIDS, IEEE Trans of Information.Tech,Vol.9, pp.59–65(2008).
- Christopher Sangal A.K: Tele communication backbone for telemedicine, pp.61–67(2008).

Ebe Ralph Grove, Department of computer science: Internet based expert system, International Journal on Expert System, Vol.17, pp.129–135(2008).

- Francis Stamper, Todd Maxwell JR: A software engineering approach to the design of a medical expert system, Proceedings of 4th Inter. Conference on S.E,Vol.15, pp.341–348(2008).
- Elisa Solomon Scott D, Dobson Joanna, Pocock Stuart, Skali Hicham, McMurray John J V, Granger Christopher B,Yusuf Salim, Swedberg Karl, Young James B,Michelson Eric L, Pfeffer Marc: Influence of nonfatal hospitalization for Diabetes Mellitus on subsequent mortality in patients with chronic heart failure, pp.174–189(2009).
- Francis Burton Committee to Update the 2001 Guidelines for the Evaluation and Management of Diabetes Mellitus and guidelines for the diagnosis and management of chronic heart failure and adult throat cancer diagnosis, pp.138–157(2009).
- John Teslia, Raffaele Cappelli, Dario Maio: The Christina theory for the Diagnosis and Treatment of Chronic Heart Failure of the ESC, Guidelines for the diagnosis and management of diabetes mellitus.1093/ eurheartj, pp.122–128(2009).
- Joshuva Goldberg Lee R, Piette John D, Walsh Mary Norine, Frank Theodore A, Jaski Brian E, Smith Andrew L, Rodriguez Raymond, Mancini Donna M, Hopton Laurie A, Orav E John, Loh Evan: Randomized trial of a daily electronic home monitoring system in patients with advanced diabetes mellitus, pp.57–63(2009).
- Rose Violet Konstanta D, Herzog R: Continuous Monitoring of Vital Constants and Diseases for Mobile Users - The Mobile Health Approach, IEEE EMBS, pp.82–83(2010).