ISSN : 0974 - 7435 Volume 10 Issue 9

2014

BioTechnology

An Indian Journal

FULL PAPER

BTAIJ, 10(9), 2014 [3772-3779]

Applications domain driven data mining methodology in bioinformatics

Yadan Li[1], Qinghua Bai[1], Zhicheng Chen[2]

[1]Tongji University, Siping Road 1239, Shanghai, (CHINA)

[2]Shanghai Finance University, Shanghai, (CHINA)

E-mail : liyadan2@gmail.com; tjqbai@vip.sina.com; woods@126.com

ABSTRACT

Bioinformatics is the science of managing, mining, and interpreting information from biological sequences and structures.Although data mining in the domain of bioinformatics is popular, the two areas have largely been developing separately. Thereis a strong and challenging need to mine for more informative and actionable knowledge in bioinformatics. To respond thisrequirements, this study tries to probe domain driven data mining methodology in the field of bioinformatics, and tries toprovide a new thought for bioinformatics in future research.

KEYWORDS

Bioinformatics; Data mining; Domain driven data mining methodology; Actionable knowledge.

INTRODUCTION

In recent years, rapid developments in genomics and proteomics have generated a large amount of biologicaldata. Drawing meaningful results from these data requires sophisticated computational analyses.Bioinformatics, or computational biology, is the interdisciplinary science of interpreting biological datausing information technology and computer science,and etc. 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[1].

A particular active area of research in bioinformatics is the application and development of data mining techniques to solve biological problems by now. 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. However, extant Data mining is presumed as an automated process that produces automatic algorithms andtools without human involvement and the capability to adopt to external environment constraints[2]. We know how to classify biological sequences (SVM, Neural Nets, Decision Trees, Rules), know howto cluster biological entities (Bi-clustering, K-means, hierarchical), know how to select features (PCA,LDA, SVM-RFE), as a result, although many patterns are mined from the data set through these technologies, few are satisfied the real needs and applications of bioinformatics. For instance, many knowledge discovered by data mining technology from one set could not help another sample.

Bioinformatics is the science of managing, mining, and interpreting information from biological sequences and structures. Advances such as genome-sequencing initiatives, micro arrays, proteomics, and functional and structural genomics have pushed the frontiers of human knowledge. In a sense, human knowledge cannot involved in the process of data mining process usually.

In addition, data mining has been advancing in strides in recent years, with high-impact applications from marketing to science. Although researchers have spent much effort on data mining for bioinformatics, the two areas have largely been developing separately[3].

Accordingly, in order to generate actionable knowledge satisfied genuine needs of bioinformatics completely, a great potential to increase the interaction between data mining and bioinformatics, this study tries to probe domain driven data mining methodology in the field of bioinformatics, and tries to provide a new thought for bioinformatics in future research.

RELATED WORK

Bioinformatics

The term bioinformatics was coined by Paulien Hogeweg in 1979 for the study of informatic processesin biotic systems. It was primary used since late 1980s has been in genomics and genetics, particularlyin those areas of genomics involving large-scale DNA sequencing[1,4].

Bioinformatics can be defined as the application of computer technology to the management of biologicalinformation. Bioinformatics is the study of applying computational methods to large amount ofbiological information in order to facilitate in biology and medicine. It has been mainly fueled by advancesin DNA sequencing and mapping techniques. Over the past few decades rapid developments ingenomic and other molecular research technologies and developments in information technologies havecombined to produce a tremendous amount of information related to molecular biology. The primarygoal of bioinformatics is to increase the understanding of biological processes. Some of the grand areaof 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 partdiffers during medical analysis and genome mapping processes.

Genome annotation

In the context of genomics, annotation is the process of marking the genesand other biological features in a DNA sequence.

Analysis of gene expression

The expression of many genes can be determined by measuringmRNA levels with various techniques such as microarrays, expressed cDNA sequence tag (EST)sequencing, serial analysis of gene expression (SAGE) tag sequencing, massively parallel signaturesequencing (MPSS), or various applications of multiplexed in-situ hybridization etc.

Analysis of protein expression

Gene expression is measured in many ways including mRNA andprotein expression, however protein expression is one of the best clues of actual gene activity sinceproteins are usually final catalysts of cell activity. Protein microarrays and high throughput (HT)mass spectrometry (MS) can provide a snapshot of the proteins present in a biological sample.

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. Knowledge of this structureis vital in understanding the function of the protein. For lack of better terms, structural informationis usually classified as secondary, tertiary and quaternary structure. Protein structure prediction isone of the most important for drug design and the design of novel enzymes.

Comparative genomics

Comparative genomics is the study of the relationship of genome structureand function across different biological species. Gene finding is an important application ofcomparative genomics, as is discovery of new, non-coding functional elements of the genome.

Modeling biological systems

Modeling biological systems is a significant task of systems biologyand mathematical biology. Computational systems biology aims to develop and use efficient algorithms,data structures, visualization and communication tools for the integration of large quantitiesof biological data with the goal of computer modeling.

High-throughput image analysis

Computational technologies are used to accelerate or fully automatethe processing, quantification and analysis of large amounts of high-information-contentbiomedical images. Modern image analysis systems augment an observer's ability to make measurementsfrom a large or complex set of images. Biomedical imaging is becoming more importantfor both diagnostics and research.

Data mining

Data mining named knowledge discovery, as well as its synonyms knowledge discovery, is frequentlyreferred to the literature as the process of extraction interesting information or patterns from data[5]. ata mining is not specific to any industry. It requires intelligent technologies and the willingness toexplore the possibility of hidden knowledge that resides in the data.

Data Mining approaches seem ideally suited for Bioinformatics(The process of knowledge discoverycan be seen in Figure 1), since it is data-rich, but lacks a comprehensive theory of lifea¸r´s organization at themolecular level. There are many methods of data mining shown as follows:

Classification

Classification is learning a function that maps a data item into one of several predefinedclasses.

Estimation

Given some input data, coming up with a value for some unknown continuous variable.

Prediction

Same as classification except that the records are classified according to some futurebehavior.

Association rules

Discover the high frequency pattern and discover which things appear frequentlyand simultaneously.

Clustering

Segmenting a population into a number of subgroups or clusters.

Visualization

Representing the data using visualization techniques.

While there are countless researchers, especially recent researchers, working on designing efficientdata mining technique and algorithms. Data mining is a data driven trial-and-error process[6,], aidsto extract patterns in data without human involvement. Knowledge discovery overemphasized by innovativealgorithm-driven research can not meet the needs of real world[8]. For the domain of Bioinformatics,human knowledge involved in the process of mining is very important. For instance, learn humanunderstandable rules that can define the epigenetic process in cancer and embryonic stem cell[9].

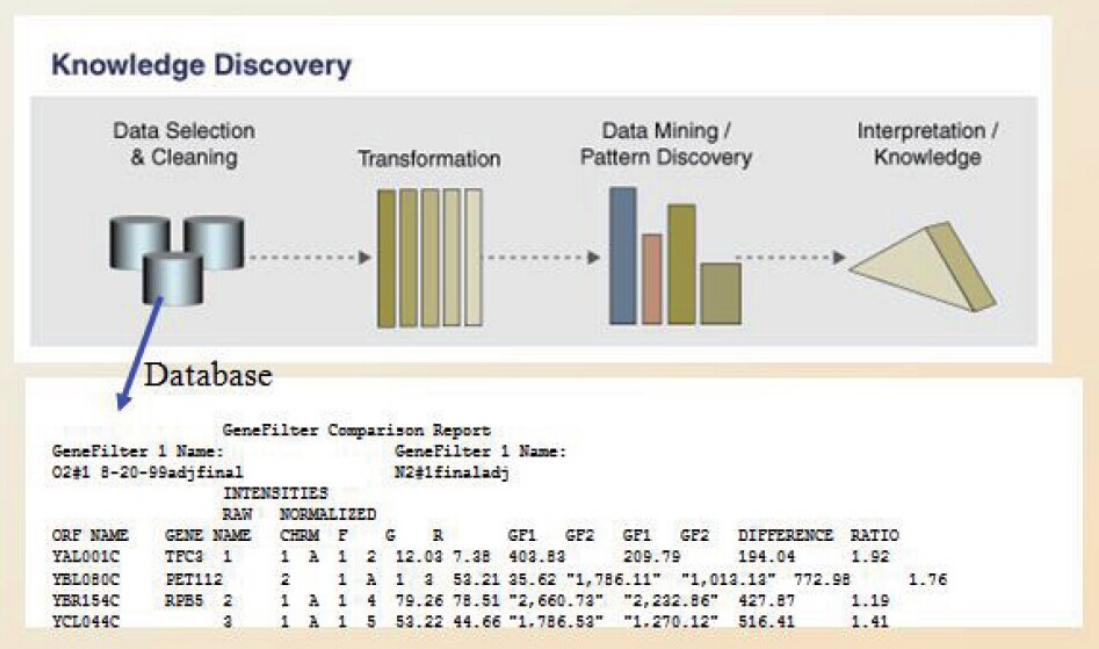

Figure 1 : The process of knowledge discovery for bioinformatics

Nowadays, the primary challenges are moving data-driven to domain-driven and focus on discoveringinteresting and actionable knowledge. A new methodology on top of the traditional data-centered patternmining framework, is called Domain Driven Data Mining Methodology.

DOMAIN DRIVEN DATA MINING METHODOLOGY

Domain Driven Data Mining Methodology targets to overcome three types of contradiction existing in traditional data mining. The first one is the rule vs. interestingness, that is, discovered rules are not useful or interesting to user, as well as the rules are not desired; the second is the rule vs. actionability, that is, there is a gap between rule and real world applications; the last is the rule vs. data, it can be discussed from the following two aspects, one is that data don't contain the information that user required usually, the other is many useless or bad data, which cover the values of meaningful data. To deal with the contradiction mentioned, Domain Driven Data Mining Methodology caters for the effective involvement of intelligence, such as domain knowledge and expert experience, surrounding actionable knowledge discovery in meeting real world needs.

Attributes of domain driven data mining methodology

Generally, data-driven data mining system can generate a glut of knowledge, most of which are of no interest to the experts. Domain Driven Data Mining Methodology is moving data-driven to domain driven and focuses on discovering interesting and actionable knowledge. Thus, some attributes, such as interestingness, actionability and domain knowledge [7,10], need to be involved during the process of knowledge discovery.

Attribute 1 Interestingness

The knowledge discovered is unexpected or desired to the decision makers. It is strongly dependent on the application domain, expert knowledge, as well as experience.

Attribute 2 Actionability

It refers to the knowledge mined can suggest concrete and profitable action to the decision makers. It is actually relied on the application domain.

Attribute 3 Domain knowledge

Domain knowledge, the knowledge which is valid and used directly for a pre-selected domain of human behavior and experience or an autonomous computer activity. It is dependent on domain expert, user, environment, context, etc.

Combined rule mining

Combined rule mining is employed to find more actionable knowledge usually. A combined rule is composed of multiple heterogeneous itemsets from different datasets. Combined patterns take the forms of combined association rules, combined rule pairs and combined rule clusters[5], which are defined as follows.

Definition 1 Combined association rule

Assume that there are *m* database *Di* (i=1,⋯,m). Assume*Ii* to be the set of all items in database *Di* and $\forall\ i \neq j$, $Ii \cap Ij = \phi$. A combined association rule *R* is inthe form of

$$A1 \wedge A2 \wedge \cdots \wedge AK \rightarrow T \quad (1)$$

Where $Ai \subseteq Ii$ (i=1,…,m) is an itemset in dataset *Di*, $T \neq \phi$ is a target item or class and $\exists i \neq j$, $Ai \neq \phi$, $Aj \neq \phi$.

Definition 2 Combined Rule Pair

Assume that *R*1 and *R*2 are two combined rules and that theirleft sides can be spilt into two parts, *U* and *V* , where *U* and *V* are respectively itemsets from *IU* and*IV* (*I*={I_i}, $IU \subset I$, $IV \subset I$, $IU \neq \phi$, $IV \neq \phi$ and $IU \bigcap IV \neq \phi$. If *R*1 and *R*2 share a same *U* but havedifferent *V* and different right sides, then they build a combined rule pair *P* as

$$\mathbf{P} := \begin{cases} \mathbf{R_1 : U \wedge V_1 \rightarrow T_1} \\ \mathbf{R_2 : U \wedge V_2 \rightarrow T_2} \end{cases} \quad \textbf{(2)}$$

Where $U \neq \phi$, $V1 \neq \phi$, $V2 \neq \phi$, $T1 \neq \phi$, $T \neq \phi$, $U \bigcap V1 \neq \phi$, $U \bigcap V2 \neq \phi$, $V1 \bigcap V2 \neq \phi$ and $T1 \bigcap T2 \neq \phi$.

Definition 3 Combined Rule Cluster

A combined rule cluster *C* is a set of combined associationrule based on a combined rule pair *P*, where the rules in *C* share a same *U* but have different *V* in theleft side.

$$\mathbf{C} := \begin{cases} \mathbf{U \wedge V_1 \rightarrow T_1} \\ \mathbf{U \wedge V_2 \rightarrow T_2} \\ \mathbf{...} \\ \mathbf{U \wedge V_n \rightarrow T_n} \end{cases} \quad \textbf{(3)}$$

Where $U \neq \phi$; $\forall i$, $Vi \neq \phi$, $U \bigcap Vi \neq \phi$; and $\forall i \neq j$, $Vi \bigcap Vj = \phi$.

Based on traditional *Support*, *Confidence* and *Lift*, two new lifts are designed as follows for measuringthe interestingness of combined association rules[5].

$$\mathbf{Lift_U(U \cap V \rightarrow T) = \frac{Conf(U \cap V \rightarrow T)}{Conf(V \rightarrow T)} = \frac{Lift(U \cap V \rightarrow T)}{Lift(V \rightarrow T)}} \quad \textbf{(4)}$$

TABLE 1 : Dataset

Patient ID	stem cells	Disease	Patient ID	stem cells	Disease
1	P_3, P_4	Y	1	P_4	Y
2	P_3, P_5	N	2	P_4, P_5	Y
2	P_2, P_4, P_5	N	3	P_2, P_3, P_5	Y
3	P_3, P_4, P_7	Y	4	P_3, P_4	N
4	P_5	N	4	P_2, P_4	N

TABLE 2 : Patient demographic data

Patient ID	Gender	Patient ID	Gender
1	F	2	F
3	M	4	M

$$\mathbf{Lift_V(U \cap V \rightarrow T) = \frac{Conf(U \cap V \rightarrow T)}{Conf(U \rightarrow T)} = \frac{Lift(U \cap V \rightarrow T)}{Lift(U \rightarrow T)}} \quad \textbf{(5)}$$

Based on the above two new lifts, the interestingness of combined association rules is defined as

$$\mathbf{I_{rule}(U \cap V \rightarrow T) = \frac{Lift_U(U \cap V \rightarrow T)}{Lift(U \rightarrow T)}} \quad \mathbf{(6)}$$

Irule indicates whether the contribution of *U* (or *V*) to the occurrence of *T* increases with *V* (or *U*) asprecondition. The value of *Irule* falls between 0 and infinity. When *Irule*>1, the higher *Irule* is, the moreinteresting the rule is.

For a rule cluster *C* composed of *n* combined association rules *R*1, *R*2,… ,*Rn*, itsinterestingness isdefined as

$$\mathbf{I_{cluster}(C) = \max_{i \neq j, R_i, R_j \in C, T_i \neq T_j} I_{pair}(R_i, R_j)} \quad \mathbf{(7)}$$

A NUMERICAL EXAMPLE

A simplified numerical example is used to demonstrate the implementation procedure of domain drivendata mining. Take identification of disease between parts and gender as an example. There are eight partsof stem cells(*P*1,*P*2, *P*3, *P*4, *P*5, *P*6, *P*7, *P*8). Suppose human understandable pattern is a key factor in the identification of patient. This example only considers patient demographic dataset and the numberattributes of dataset (See TABLES 1 and 2). The concrete procedure can be summarized as follows:

– Assume*Minsupport*1=0.2 and*Minconfidence*1=0.4. Thus, single association rules can be shownas follows: $F \rightarrow Y$, $F \rightarrow N$, $M \rightarrow Y$, $M \rightarrow N$, $P3 \rightarrow Y$, $P3 \rightarrow N$, $P4 \rightarrow Y$, $P4 \rightarrow N$, $P5 \rightarrow Y$ and $P5 \rightarrow N$.

TABLE 3 : Combined association rules

Rules	*Support*	*Confidence*	*Lift*	$Lift_1$	$Lift_2$	I_{rule}
$F \wedge P_4 \rightarrow Y$	0.3	0.75	1.5	1.3	1.25	1.0
$F \wedge P_5 \rightarrow N$	0.2	0.67	1.3	1.1	1.7	1.4
$M \wedge P_3 \rightarrow Y$	0.2	0.67	1.3	1	1.7	1.3
$M \wedge P_4 \rightarrow N$	0.2	0.67	1.3	1.3	1.1	1.1

TABLE 4 : Combined rule pairs

Pairs	Combined rules	I_{pair}
$Pair_1$	$F \wedge P_3 \rightarrow Y$	1.3
	$F \wedge P_4 \rightarrow N$	1.3
$Pair_2$	$M \wedge P_4 \rightarrow Y$	1.1
	$M \wedge P_4 \rightarrow N$	1.1
$Pair_3$	$F \wedge P_4 \rightarrow Y$	1.4
	$F \wedge P_5 \rightarrow N$	1.4

– Assume $Minsupport_2$=0.2, $Minconfidence_2$=0.6. According to equations (1, 4, 5, 6), combinedassociation rules can be shown in TABLE 3.

– Based on equations (2, 3, 7), combined rule pairs can be shown in TABLE 4.

– Finally, The actionable patterns indicate that *P*3 is the key part leading to the disease of male patient.Meanwhile, *P*4 is the key part leading to the disease of female patient.

CONCLUSION

Bioinformatics and data mining are developing as interdisciplinary science. Data mining approachesseem ideally suited for bioinformatics, since bioinformatics is data-rich but lacks a comprehensive theoryof life's organization at the molecular level.

However, although many patterns are mined from the data set through data mining technologies, feware satisfied the real needs and applications of bioinformatics. For instance, the existing problem is therange of levels the domains of expertise present amongst potential users, so it can be difficult for thedatabase curators to provide access mechanism appropriate to all. Since data miningapplications haveurgent requirements for discovering actionable knowledge to decision makers in bioinformatics, datacenteredtraditional data mining cannot satisfy the needs fully. Thus, DDDM is developed to provide asystematic overview in acquiring actionable knowledge through human mining system involved ubiquitousintelligence such as expert intelligence and human intelligence.

To conclude, this study proposesDDDMtowards the fields of bioinformatics for actionable knowledge.In particular, a numerical example is employed to verify its effectiveness. Furthermore, more experimentswill be confirmed in future research.

REFERENCES

[1] K.Raza; Application of Data Mining in Bioinformatics, Indian Journal of Computer Science and Engineering, **1**, 114–118 **(2010)**.

[2] L.B.Cao, C.Q.Zhang; Domain-driven, actionable knowledge discovery, IEEE Intelligent Systems, **22**, 78–88 **(2007)**.

[3] J.Y.Li, L.S.Wong, Q.Yang; Data Mining in Bioinformatics, IEEE Intellignet Systems, **20**, 16–18 **(2005)**.

[4] R.J.A.Richard, N.Sriraam; A Feasibility Study of Challenges and Opportunities in Computational Biology: AMalaysian Perspective, American Journal of Applied Sciences, **2**, 1296–1300 **(2005)**.

[5] Y.C.Zhao, H.F.Zhang, L.B.Cao, C.Q.Zhang, H.Bohlscheid; Combined Pattern Mining: From Learned Rules toActionable Knowledge,Lecutre Notes in Computer Science, **5360**, 393–403 **(2008)**.

[6] L.B.Cao, C.Q.Zhang; Domain-driven Data Mining: A Practical Methodology, International Journal of Data Warehousingand Mining, **2**, 49–65 **(2006)**.

[7] L.B.Cao, P.S.Yu, C.Q.Zhang, Y.C.Zhao; Domain Driven Data Mining, Springer-Verlag, **(2010)**.

[8] Q.Yang, J.Yin, C.X.Ling, T.L.Chen; Extracting actionable knowledge from decision trees. IEEE Transactions onKnowledge and Data Engineering, **19**, 43–56 **(2007)**.

[9] H.Liu, J.Li, L.Wong; Use of Extreme Patient Samples for Outcome Prediction from Gene Expression Data, Bioinformatics, **21**, 3377–3384 **(2005)**.

[10] L.B.Cao, D.Luo, C.Zhang; Knowledge Actionability: Satisfying Technical and Business Interestingness, International Journal of Business Intelligence and Data Mining, **2**, 496–514 **(2007)**.