International Journal of Early Childhood Special Education (INT-JECS)
DOI: 10.48047/INTJECSE/V14I7.212 ISSN: 1308-5581 Vol 14, Issue 07 2022

ENHANCED MODELING OF GENE EXPRESSION DATA USING MICRO ARRAY METHODOLOGY

Dr.T.SHANMUGAVADIVU MCA.,Ph. D.,
Assistant Professor in Computer Science, Arulmigu Palaniandavar Arts College for Women, Palani.

ABSTRACT : The huge amount of genomic data generated not only leads to a demand on the computer science community to help store, organize and index the data, but also leads to a demand for specialized tools to view and analyze the data. The main role of bioinformatics was to create and maintain databases to store biological information, such as nucleotide and amino acid sequences. With more and more data generated, nowadays, the most pressing task of bioinformatics has moved to analyze and interpret various types of data, including nucleotide and amino acid sequences, protein domains, protein structures and so on. To meet the new requirements arising from the new tasks, researchers in the field of bioinformatics are working on the development of new algorithms (mathematical formulas, statistical methods and etc) and software tools which are designed for assessing relationships among large data sets stored.

Keywords : Bagging decision tree, optimal data set, supervised learning

1. INTRODUCTION

In recent years, rapid developments in genomics and proteomics have generated a large amount of biological data. Bioinformatics, or computational biology, is the interdisciplinary science of interpreting biological data using information technology and computer science. The primary goal of bioinformatics is to increase the understanding of biological processes. Analyzing large biological data sets requires making sense of the data by inferring structure or generalizations from the data. The past few decades witness an explosive growth in biological information generated by the scientific community. This is caused by major advances in the field of molecular biology, coupled with advances in genomic technologies. In turn, the huge amount of genomic data generated not only leads to a demand on the computer science community to help store, organize and index the data, but also leads to a demand for specialized tools to view and analyze the data. "Biology in the 21st century is being transformed from a purely lab-based science to an information science as well". As a result of this transformation, a new field of science was born, in which biology, computer science, and information technology merges to form a single discipline. This is bioinformatics. "The ultimate goal of bioinformatics is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned".

One of the basic characteristics of life is its diversity. Everyone can notice this by just observing the great differences among living creatures. Despite this diversity, the molecular details underlying living organisms are almost universal. Every living organism depends on the activities of a complex family of molecules called proteins. Proteins are the main structural and functional units of an organism's cell. Proteins and nucleic acids are both called biological macromolecules, due to their large size compared to other molecules. Important efforts towards understanding life are made by studying the structure and function of biological macromolecules. The branch of biology concerned in this study is called molecular biology. Both proteins and nucleic acids are linear polymers of smaller molecules called monomers. The term sequence is used to refer to the order of monomers that constitute a macromolecule. A sequence can be represented as a string of different symbols, one for each monomer. There are twenty protein monomers called amino acids. There exist two nucleic acids, deoxyribonucleic acid (DNA) and ribonucleic acid (RNA), composed by four different monomers called nucleotides. DNA is the genetic material of almost every living organism. RNA has many functions inside a cell and plays an important role in protein synthesis.

2. PROBLEM DEFINITION

To make use of original biological and clinical data in the data mining process, we follow the regular process flow in data mining but with emphasis on three steps of feature manipulation, viz. feature space generation,

feature selection and feature integration with learning algorithms. These steps are important in dealing with biological and clinical data.

(i) Some biological data, such as DNA sequences, have no explicit features that can be easily used by learning algorithms. Thus, constructing a feature space to describe original data becomes necessary.

(ii) Quite a number of biological and clinical data sets possess many features. Selecting signal features and removing noisy ones will not only largely reduce the processing time and greatly improve the learning performance in the later stage, but also help locate good patterns that are related to the essence of the study. For example, in gene expression data analysis, feature selection methods have been widely used to find genes that are most associated with a disease or a subtype of certain cancer.

(iii) Many issues arising from biological and clinical data, in the final analysis, can be treated as or converted into classification problems and then can be solved by data mining algorithms.

3. MOTIVATION

At the beginning, the main role of bioinformatics was to create and maintain databases to store biological information, such as nucleotide and amino acid sequences. With more and more data generated, nowadays, the most pressing task of bioinformatics has moved to analyze and interpret various types of data, including nucleotide and amino acid sequences, protein domains, protein structures and so on. To meet the new requirements arising from the new tasks, researchers in the field of bioinformatics are working on the development of new algorithms (mathematical formulas, statistical methods and etc) and software tools which are designed for assessing relationships among large data sets stored, such as methods to locate a gene within a sequence, predict protein structure and/or function, understand diseases at gene expression level and etc. Motivated by the fast development of bioinformatics, this thesis is designed to apply data mining technologies to some biological data so that the relevant biological problems can be solved by computer programs. The aim of data mining is to automatically or semi-automatically discover hidden knowledge, unexpected patterns and new rules from data. There are a variety of technologies involved in the process of data mining, such as statistical analysis, modeling techniques and database technology. During the last ten years, data mining is undergoing very fast development both on techniques and applications. It's typical applications include market segmentation, customer profiling, fraud detection, (electricity) loading forecasting, credit risk analysis and so on. In the current post-genome age, understanding floods of data in molecular biology brings great opportunities and big challenges to data mining researchers. Successful stories from this new application will greatly benefit both computer science and biology communities. The researcher would like to call this discovering biological knowledge "in silico" by data mining.

4. BIO INFORMATIC LED OF INFORMATON TECHNOLOGY:

The evolution of bioinformatics led to an interdisciplinary field at the intersection of biology, computer science, and information technology. The organization of data in such a way that allow us to access existing information and to submit new entries as they are produced.

- The development of tools that help in the analysis of data.
- The use of these tools to analyze the individual systems in detail, in order to gain new biological insights.

The field of bioinformatics has many applications in the modern day world, including molecular medicine, industry, agriculture, stock farming, and comparative studies. Bioinformatics in 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.

5. GENE EXPRESSION AND DNA SEQUENCE DATA

Gene expression profiles and DNA sequence data. For gene expression profiles applies the method to solve two kinds of problems: phenotype classification and patient survival prediction. In these two problems, genes serve as features. Since profile data often contains thousands of genes, the researcher put forward a new feature selection method GSA and GFS algorithms to identify genes most related to the problem. GSA and GFS conduct three-phase of gene filtering. First, it selects genes using an entropy-based discretization algorithm, which generally keeps only 10% of discriminating genes. Secondly, these remaining genes are further filtered by Wilcoxon rank sum test, a non-parametric statistic alternative to the t-test. Genes passing this round of filtering are automatically divided into two groups: one group consists of genes that are highly expressed in one type of samples (such as cancer) while

another group consists of genes that are highly expressed in another type of samples (such as non-cancer). In the third phase, correlated genes in each group are determined by Pearson correlation coefficient test and only some representatives of them are kept to form the final set of selected genes. When applying learning algorithms to classify phenotypes, we focus on classifiers built on the idea of an ensemble of decision trees, including the newly published CS4, as well as state-of-the-art Bagging, Boosting and Random forests.

More than one thousand tests are conducted on six published gene expression profiling data sets and one proteomic data set. To compare the performance of these ensembles of decision tree methods with those widely used learning algorithms in gene expression studies, experimental results on support vector machines (SVM) and k-nearest neighbour (k-NN) are also collected. SVM is chosen because it is a representative of kernel function. k-NN is chosen because it is the most typical instance-based classifier. To demonstrate the main advantage of the decision tree methods, the researcher present some of decision trees induced from data sets. These trees are simple, explicit and easy to understand. For each classifier, besides GSA and GFS, the researcher also tries features selected by several other entropy-based filtering methods. Therefore, various comparisons of learning algorithms and feature selection methods can be addressed. In the study of using gene expression profiles to predict patient survival status, we present a new idea of selecting informative training samples by defining “long-term” and “short-term” survivors. After identifying genes associated with survival via GSA and GFS, a scoring model built on SVM is worked out to assign risk score to each patient. Kaplan-Meier plots for different risk groups formed on the risk scores are then drawn to show the effectiveness of the model. Another biological domain to which the proposed 3-step feature manipulation method is applied is the recognition of functional sites in DNA sequences, such as Translation Initiation Sites (TIS) and polyadenylation (poly(A)) signal. In this study, we put our emphasis on feature generation k-gram nucleotide acid or amino acid patterns are used to construct the feature space and the frequency of each pattern appearing in the sequence is used as value. Under the description of the new features, original sequence data are then transformed to frequency vector data to which feature selection and classification can be applied. In TIS recognition, the researcher tests the methods on three independent data sets. Besides the cross validation within each data set, we also conduct the tests across different data sets. In the identification of poly(A) signal, we make use of both public and our own collected data and build different models for DNA and mRNA sequences. In both studies, the researcher achieves comparable or better prediction accuracy than those reported in the literature on the same data sets. In addition, the researcher also verifies some known motifs and finds some new patterns related to the identification of relevant functional sites.

The main contributions of this research are:

(1) Articulating a 3-step feature manipulation method to solve some biological problems;

(2) Putting forward a new feature selection strategy to identify good genes from a large amount of candidates in gene expression data analysis;

(3) Presenting a new method for the study on patient survival prediction, including selecting informative training samples, choosing related genes and building an SVM-based scoring model;

(4) Applying the proposed techniques to published gene expression profiles and proteomic data, and addressing various comparisons on classification and feature selection methods from a large amount of experimental results;

(5) Pointing out significant genes from each analyzed data set, comparing them with literature and relating some of them to the relevant diseases;

(6) Recognizing two types of functional sites in DNA sequence data by using k-gram amino acid or nucleotide acid patterns to construct feature space and validating learning models across different independent data sets.

6. LITERATURE REVIEW:

Literature Review contains a critical analysis and the integration of information from a number of sources, as well as a consideration of any gaps in the literature and possibilities for future research.

(i) CLASSIFICATION BASED LEARNING

Classification learning can deal with more than two class instances. In practice, the learning process of classification is to find models that can separate instances in the different classes using the information provided by training instances. Thus, the models found can be applied to classify a new unknown instance to one of those classes. Putting it more prosaically, given some instances of the positive class and some instances of the negative class, and it can be used as a basis to decide if a new unknown instance is positive or negative . This kind of learning is a process from general to specific and is supervised because the class membership of training instances is clearly

known. In contrast to supervised learning is unsupervised learning, where there are no pre-defined classes for training instances. The main goal of unsupervised learning is to decide which instances should be grouped together, in other words, to form the classes.

(ii) SUPPORT VECTOR MACHINES

Support Vector Machines (SVM) is a kind of a blend of linear modeling and instance-based learning which uses linear models to implement nonlinear class boundaries. It originates from research in statistical learning theory . An SVM selects a small number of critical boundary samples from each class of training data and builds a linear discriminant function (also called maximum margin hyperplane) that separates them as widely as possible. The selected samples that are closest to the maximum margin hyperplane are called support vectors.

7. GENE FILTER (GFA) AND GENE SELECTION (GSA) ALGORITHMS

Gene Filter Algorithm GFA is a new strategy to conduct feature selection, mainly aiming to find significant genes in supervised learning from gene expression data. In this algorithm, the researcher combine the above presented methods of entropy measure and Wilcoxon rank sum test, as well as Pearson correlation coefficient test together to form a three-phase feature selection process.

Gene Filter Algorithm (GFA)

Step-1: k=1

Step-2: Rank all feature in group F on class entropy in an ascending order,

$f_1, f_2, \ldots f_1$.

Step-3: Let $S_k=\{f_1\}$ and remove f_1 from F.

Step-4: For each $f_i (i > 1)$

Calculate Pearson correlation coefficient $r(f_1, f_i)$;

If $r(f_1, f_i) > r_c$

Add f_i into S_k and remove it from F;

Step-5: k=k+1 and goto step 2 until F=∅.

Gene Selection Algorithm (GSA)

Step-1: Select a statistic which will be used to measure differences between classes.

Step-2: Determine the threshold of the statistic according to significant level ∝.

Step-3: Calculate the test statistic for each of total features

Step-4: Get the number of features selected by the threshold record as w.

Step-5: For i^{th} permutation test iteration (i=1,2…….,t):generate a pseudo data set by randomly permuting the class labels of all the samples, calculate the same test statistic for every feature, record how many features are selected by the threshold ,denote it as k_i.

Step-6: Compute the percentage of features selected during the permutation test, $p= \frac{\sum_{i=1}^{t} k_i}{t \times m}$ calculate p×w to be the expected number of false positive

8. PATIENT SURVIVAL PREDICTION ALGORITHM

The researcher carefully form the training set samples by selecting only short-term survivors who died within a short period and long-term survivors who were still alive after a relatively long follow-up time. This idea is motivated by our belief that short-term and long-term survivors are more informative and reliable (than those cases in between) for building and understanding the relationship between genes and patient survival. In the second step, GSA is used to identify genes most associated with survival. In the third step, a linear kernel Support Vector Machine (SVM) is trained on the selected samples and genes to build a scoring model. The model assigns each validation sample a risk score to predict patient survival.

Patient Survival Prediction Algorithm

Step-1 : Read n samples.
Step-2 : Select training samples.
Step-3 : If training samples long-term and short term then
Step-4 : Identify genes
Step-5 : Genes related to survival
Step-6 : Build SVM scoring function and form risk groups
Step-7 : Assign risk score and risk group to each sample
Step- 8 : Draw Kaplan–Meler curves

9. RESULTS AND DISCUSSIONS

All biomedical data contain explicit signals or features as those in the classification problems raised by gene expression profiling. For example, DNA sequences and protein sequences represent the spectrum of biomedical data that possess no explicit features. Generally, a genomic sequence is just a string consisting of the letters “A”, “C”, “G”, and “T” in a “random order”. DNA process can be divided into two stages: transcription and translation.

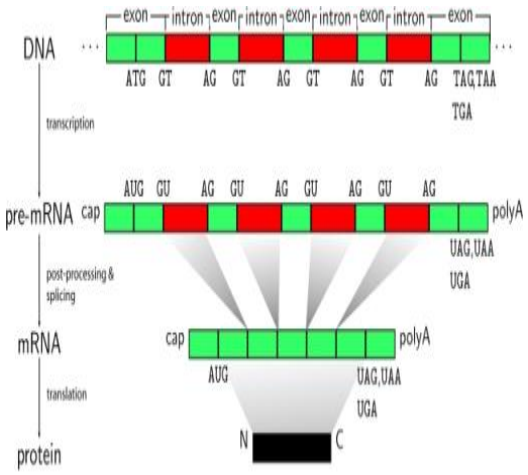

Figure-1: Process of protein synthesis

Transcription: In this stage, the information in DNA is passed on to RNA. This takes place when one strand of the DNA double helix is used as a template by the RNA polymerase to create a messenger RNA (mRNA). Then this mRNA moves from the nucleus to the cytoplasm. In fact, in the cell nucleus, the DNA with all the exons and introns of the gene is first transcribed into a complementary RNA copy named “nuclear RNA” (nRNA). This is indicated as “primary transcription” shown in Figure-1. Secondly, non-coding sequences of base pairs (introns) are eliminated from the coding sequences (exons) by RNA splicing. The resulting mRNA is the edited sequence of nRNA after splicing. The coding mRNA sequence can be described in terms of a unit of three nucleotides called a codon.

Translation: In this stage, the information that has been passed to RNA from DNA is used to make proteins. At the initiation phase of translation, ribosome binds to the mRNA when it reaches an AUG (adenine, uracil, guanine) sequence on the RNA strand in a suitable context. The ribosome is made of protein and ribosomal RNA (rRNA). The start codon AUG is called translation initiation site (TIS) and is only recognized by the initiator tRNA (transfer RNA). After binding to the mRNA, the ribosome proceeds to the elongation phase of protein synthesis by sequentially binding to the appropriate codon in mRNA to form base pairs with the anticodon of another tRNA molecule. Hence, with the ribosome moving from codon to codon along the mRNA, amino acids are added one by one, translated into polypeptide sequences. At the end, the newly formed strand of amino acids (complete polypeptide) is released from the ribosome when a release factor binds to the stop codon. This is the termination phase of translation.

The functional sites in DNA sequences include transcription start site (TSS), translation initiation site (TIS), coding region, splice site, polyadenylation (cleavage) site and so on that are associated with the primary structure of genes. Recognition of these biological functional sites in a genomic sequence is an important bioinformatics application. The researcher proposes a 3-step work flow as follows. In the first step, candidate features are generated using k-gram nucleotide acid or amino acid patterns and then sequence data are transformed with respect to the newly generated feature space. In the second step, a small number of good features are selected by a certain algorithm. In the third step, a classification model is built to recognize the functional site.

The researcher proposed a machine learning methodology to identify functional site in biological sequences. The researcher's method comprises three sequential steps: (1) generating candidate features using k-gram nucleotide acid patterns or amino acid patterns and then transforming original sequences respect to the new generated feature space; (2) selecting relevant features using certain GFA feature selection algorithm; and (3) building classification model to recognize the functional site by applying classification techniques to the selected features. The researcher idea is different from traditional methodologies because it generates new features and also transforms the original nucleotide sequence data to k-gram frequency vectors. The feature selection step does not only greatly shorten the running time of classification program, but also help to obtain explicit important features around the functional site and lead to a more accurate prediction. The researcher applied our idea to predict translation initiation site (TIS) and polyadenylation signal (PAS) in DNA and mRNA sequences. For each application, both public data sets and our own extracted sequences were used to test the effectiveness and robustness of the method. The experimental results achieved are better than those reported previously using the same data sets (if available). The important features captured are highly consistent with those reported in the literature. Most importantly, the researcher not only conducted the cross validation within the individual data sets separately, but also established the validation across the different data sets. The success of such a validation indicates that there are predictable patterns around TIS or PAS.

The researcher designed a series of experiments on the three data sets:

- Conducting computational cross validations in data set I and data set II separately.
- Selecting features and building classification model using data set I. Applying the well trained model to data set II to obtain a blind testing accuracy.
- Incorporating the idea of ribosome scanning into the classification model.
- Applying the model built in experiment-b to genomic sequences.

10. CONCLUSION :

The researcher successfully makes use of data mining technologies to solve some problems arising from biological and clinical data. The researcher have articulated explicitly the 3-step frame work of feature generation, feature selection and feature integration with learning algorithms and demonstrated its effectiveness when dealing with phenotype classification and patient survival prediction from gene expression data, and functional sites recognition in DNA sequences. From large amount of experiments conducted on some high-dimensional gene expression data sets, the researcher clearly observe the improvements on performances of all the classification algorithms under the proposed feature selection scenarios. Among these gene identification methods, the researcher claim GFA algorithm is an effective approach. In the aspect of classification algorithms, no single algorithm is absolutely superior to all others, though SVM achieves fairly good results in most of tests. Compared with SVM,

decision tree methods can provide simple, comprehensive rules and are not very sensitive to feature selections. Among the decision tree methods, the newly implemented CS4 achieves good prediction performance and provides many interesting rules. Feature generation is important for some kinds of biological data. The researcher illustrates this point by properly constructing new feature space for functional sites recognition in DNA sequences. Some of the signal patterns identified from the generated feature space is highly consistent with related literature or biological knowledge. The rest might be useful for biologists to conduct further analysis.

REFERENCES:

[1] P. Agarwal and V. Bafna. The ribosome scanning model for translation initiation: implicationsfor gene prediction and full-length cDNA detection. Proceedings of 6th Internationa lConference on Intelligent Systems for Molecular Biology, pages 2–7, June1998.

[2] Y. Aissouni, C. Perez, B. Calmels, and P.D. Benech. The cleavage/polyadenylation activity triggered by a U-rich motif sequence is differently required depending on the poly(A)site location at either the first or last 3'-terminal exon of the 2'-5' oligo(A) synthetase gene. Journal of Biological Chemistry, 277:35808–35814, 2002.

[3] A.A. Alizadeh, M.B. Eisen, R.E. Davis, C.Ma, I.S. Lossos, A. Rosenwald, J.C. Boldrick,H. Sabet, T. Tran, X. Yu, J.I. Powell, L. Yang, G.E. Marti, T. Moore, J.Jr. Hudson, L. Lu,D.B. Lewis, R. Tibshirani, G. Sherlock, W.C. Chan, T.C. Greiner, D.D. Weisenburger,J.O. Armitage, R. Warnke, and L.M. Staudt. Distinct types of diffuse large B-cell lymphomaidentified by gene expression profiling. Nature, 403:503–511, 2000.

[4] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine.Broad patterns of gene expression revealed by clustering analysis of tumor and normalcolon tissues probed by oligonucleotide arrays. Proceedings of National Academy ofSciences of the United States of American, 96:6745–6750, 1999.

[5] S.A. Armstrong, J.E. Staunton, L.B. Silverman, R. Pieters, M.L. den Boer, M.D.Minden, S.E. Sallan, E.S. Lander, T.R. Golub, and S.J. Korsmeyer. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nature Genetics, 30(1):41–47, 2002.

[6] Ghosh, D. and Chinnaiyan, A.M. (2002). Mixture Modeling of Gene Expression Data from Microarray Experiments. Bioinformatics, 18, 275-286.