Indonesian Journal of Electrical Engineering and Computer Science
Vol. 5, No. 2, February 2017, pp. 472 ~ 478
DOI: 10.11591/ijeecs.v5.i2.pp472-478

Classification algorithm for Gene Expression Graph and Manhattan Distance

N. Sevugapandi[1], C. P. Chandran[2]
[1]Research and Development Center, Bharathiar University, Coimbatore, Tamilnadu
[2]Computer Science, Ayya Nadar Janaki Ammal College, Sivakasi, Tamilnadu
Corresponding author, e-mail: snsevugapandi@gmail.com[1], drcpchandran@gmail.com[2]

Abstract

This proposed method focus on these issues by developing a novel classification algorithm by combining Gene Expression Graph (GEG) with Manhattan distance. This method will be used to express the gene expression data. Gene Expression Graph provides the optimal view about the relationship between normal and unhealthy genes. The method of using a graph-based gene expression to express gene information was first offered by the authors in [1] and [2], It will permits to construct a classifier based on an association between graphs represented for well-known classes and graphs represented for samples to evaluate. Additionally Euclidean distance is used to measure the strength of relationship which exists between the genes.

Keywords: *Data Mining, DNA Micro array, Gene Ontology, KEGG pathway*

1. Introduction

1.1. Data Mining

Data mining is defined as the non-trivial process of searching and analyzing data in order to find implicit but potentially useful information. Let D = {d_1... d_n} be the dataset to be analyzed. The data mining process is described as the process of finding a subset D' of D and hypotheses H_U(D',C) about D' that a user U considers useful in an application context C. D' have fewer data elements than D, but it also have a lower dimensionality (m'). In databases the data is partitioned into relations or object classes. D is considered as a union of relations R_1... R_k each has its own dimensionality (m_1... m_k)[1].

1.2. Bioinformatics

Bioinformatics is the Science of integrating, managing, mining and interpreting information from biological datasets at genomic, metabalomic, proteomics, phylogenetic and cellular or whole organism levels.

According to (National Institute of Health) NIH organization, the Bioinformatics and Computational Biology have been defined as “Bioinformatics is research and development or application of computational tools and approaches for expanding the use of biological, medical, health data including those to acquire store, organize, active, analyze or visualize such data” [2].

1.3. Genomics

DNA (Deoxyribonucleic Acid) is a molecule encoding the genetic instructions used in the development and functioning of all known living organisms many viruses. DNA is one of the three major macromolecules that are essential for all known forms of life.

Genetic information is encoded as a sequence of nucleotides (Guanine, Adenine, Thymine, and Cytosine) recorded using the letters G, A, T, and C. Most DNA molecules are double-stranded helices, consisting of two long polymers of simple units called nucleotides with the nucleo bases (G, A, T, C) attached to the sugars. DNA is well-suited for biological information storage, since the DNA backbone is resistant to cleavage and the double-stranded structure provides the molecule with a built-in duplicate of the encoded information [3].

Received September 2, 2016; Revised December 25, 2016; Accepted January 11, 2017

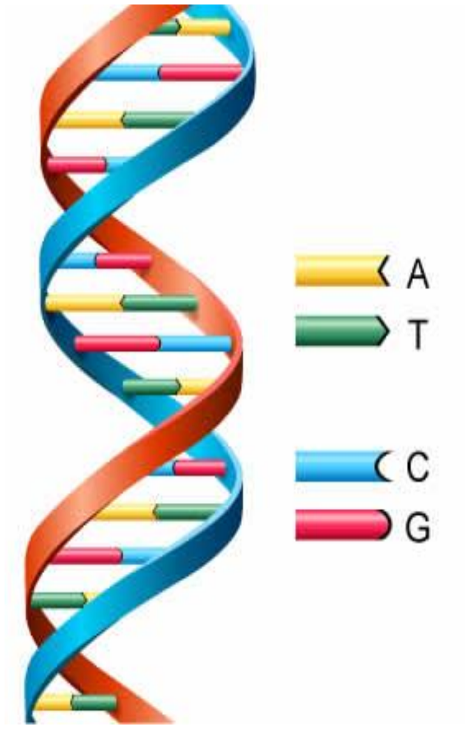

Figure 1. Structure of DNA

1.3.1. Nucleic Acids

Nucleic acids are large biological molecules essential for all known forms of life. They include DNA. Together with proteins, nucleic acids are the most important biological macromolecules; each is found in abundance in all living things, where they function in encoding, transmitting and expressing genetic information.

The nucleic acids Deoxyribonucleic acid DNA are polymers of nucleotides, arranged in a specific sequence. To form macromolecular polymers, nucleotides are joined between the 3' and 5' carbon atoms in their sugar moiety by a phosphodiester bond, giving rise to a nucleic acid with a sugar–phosphate 'backbone' to which is attached a series of bases in a specific order. Hydrogen bonding between pairs of bases can occur, leading to the formation of double-stranded polymers if the sequences are complementary [4].

1.3.2. Base pairs

Basepairs are the building blocks of the DNA double helix, and contribute to the folded structure of both DNA. Dictated by specific hydrogen bonding patterns, Watson-Crick base pairs (Guanine-Cytosine and Adenine-Thymine) allow the DNA helix to maintain a regular helical structure that is independent of its nucleotide sequence. The complementary nature of this based-paired structure provides a backup copy of all genetic information encoded within double-stranded DNA. Figure 2 shows the pairs of ATGC.

The regular structure and data redundancy provided by the DNA helix make DNA an optimal molecule for the storage of genetic information, while base-pairing between DNA and incoming nucleotides provide the mechanism through which DNA polymerase replicates DNA transcribes. Many DNA-binding proteins can recognize specific base pairing patterns that identify particular regulatory regions of genes.

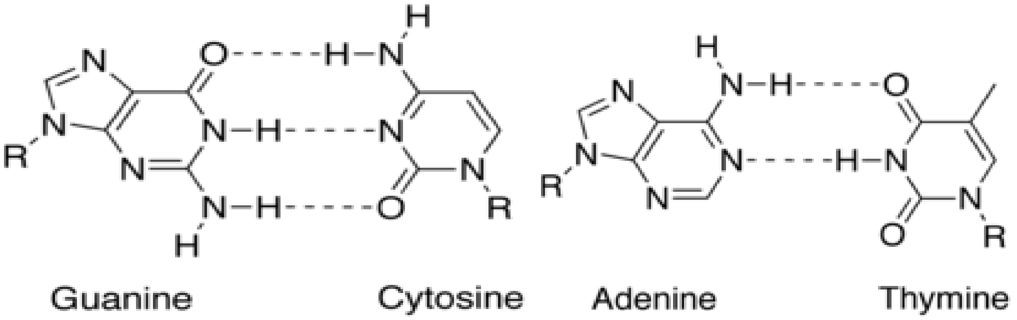

Figure 2. Base pairs of ATGC

1.3.3. Gene Expression Data

Gene expression refers to a complex series of processes in which the information encoded in a gene is used to produce a functional product such as a protein that dictates cell function. It involves several different steps through which DNA is converted to an RNA which in turn is converted into a protein or in some cases RNA, for example, genes encoding the necessary information for transfer RNAs and ribosomal RNAs (tRNAs and rRNAs). The information flow from DNA to RNA to protein can be controlled at several points helping the cell to adjust the quality and quantity of resulting proteins and thus self-regulate its functions. Thus, regulation of gene expression is a critical step in determining what kind of proteins and how much of each protein is expressed in a cell.

1.3.4. DNA Motifs

Complex designs are often created using a relatively small set of common building blocks called motifs. DNA self-assembly can exploit this same design principle to hierarchically create more sophisticated aperiodic structures.

There are many possible DNA motifs and the focus here is on only a few in the context of the target nanostructure. Motifs include junctions that enable three or more double stranded helices of DNA to interact and thus form specific structures (e.g., a triangle, a corner, and so on). Another important motif is a single strand of DNA protruding from a doublestranded helix called a sticky-end.

Two motifs with complementary sequences on their sticky-ends will bind to form a composite motif. Composite motifs may also have embedded sticky-end motifs and thus can also bind with other composite motifs to form another, larger, composite motif. This results in a hierarchical structure for motifs.

1.3.5. Various Gene Expression Techniques

Analytical methods may be used to examine mRNA expression levels or differential mRNA expression. Some examples of these techniques are listed below.

Serial Analysis of Gene Expression (SAGE): SAGE is a technique used to create a library of short sequence tags which can each be used to detect a transcript. The expression level of the transcript can be determined by assessing how many times each tag is detected. This technology enables comprehensive expression analysis across the genome [5].

DNA microarray: Also known of as biochip or DNA chip, a DNA microarray is a solid surface to which a collection of microscopic DNA spots are attached. The microarrays are used to determine expression levels across a large number of genes or to perform genotyping across different regions of a genome [6].

RNA Seq: This refers to methods used to measure the sequence of RNA molecules. Examples include shotgun sequencing of cDNA molecules acquired from RNA through reverse transcription and technologies used to sequences.

RNA molecules from a biological sample so that the primary sequence and abundance of each RNA molecule can be determined.

Tiling arrays: A tiling array is a type of microarray chip, with labelled DNA or RNA targets hybridized to probes attached to a solid surface. However, the probes used differ to those used with traditional microarrays. Rather than known sequences or predicted genes being probed, tiling arrays probe for sequences known to be present in a contigious region.

2. Previous Works

J. Breiman, et al., [3] proposed model was used for formulating the prediction model from the car dataset. This prediction model was made by continuously portioning the data in data space. At the last, the portioned result is shown as a decision tree. Classification tree was based on Decision tree, which was typically intended for dependent variables to acquire a countable number of random values along with possible prediction error. The prediction error was considered in terms of penalty for wrong classification. Regression tree algorithms are likely as classification tree algorithms for dependent variables to acquire continuous or ordered distinct values. In regression tree method, prediction error will be calculated by measuring the squared variation among the actual and predicted values [7].

H. Zhang, et al., [4] proposed a deterministic forest methodology by means of continuous tree partition from microarray gene expression data. The key attribute for this methodology is to get better prediction accuracy rate and also to identify the appropriate genes for selecting tumor cell among datasets. Deterministic classification is more likely as random forest tree algorithm but deterministic provide accurate classification when compared with random forest. In deterministic classification, initial level classification of gene population is carried out by choosing predetermined number for splitting criterion. Then perform scrutinizing operation on repeatedly occurring pairs of genes to get more needed data from dataset. This scrutinizing process should disclose the value of the genes in classifying cancer tissues in the relevant data sets [8].

V. N. Vapnik, et al., [5] proposed a well-known classification algorithm called support vector machine (SVM) algorithm to collect necessary information from DNA microarray data. This algorithm is famous for its accuracy and robust nature. The SVM will provide better performance while using multi category of biological scrutiny in cancer diagnosis from microarray gene expression data. It is capable of analyzing both sample and gene expression data in order to explore the wrongly classified data result. SVM provide significant performance. The main drawback of this method, that SVM failed to detect out of class samples and also restricts to provide inadequate clinical diagnostics application [9].

J. Khan, et al., [6] purpose of this paper is to develop a model for classifying cancers into exact diagnostic groups based on gene expression data by means of artificial neural networks (ANNs). Initially ANNs was trained by using round blue-cell tumors. This cancer diagnostic can fit into four different categories. Based on this training, ANNs appropriately classified the entire sample gene and also identify the most relevant gene for classification. This analysis had shown the possible applications of these ANNs method for cancer diagnosis.

3. Results and Discussion

The proposed weight based gene expression graph performs classification of diseases by means of gene expression data along with carrying weight at the edge. In graph G(V,E), where V represents the gene and the E represents the weighted edge. The weight present in the edge determines the strength of relationship among the gene.

The graph is built from untreated microarray scanned image data to go through the following stages such as preprocessing, data modeling, classification and validation. In microarray technology, gene is identified by types of expression levels. The first level, cy3 represents the healthy condition of gene and the second expression level, cy5 represents the diseased condition of gene. This gene expression is usually attained from the *n* samples are arranged as matrix called gene expression matrix. In preprocessing stage, this is done by means of normalization techniques.

Normalization is particularly useful to train the input values for classification algorithm, this normalization techniques helps to speed up the learning phase in classification algorithm. There are many methods for data normalization. For that, we consider the Standardized normal distribution techniques along with LN ratios of each gene. It can be computed by using equation (1), (2), (3) and (4).

$$\text{LNratio(x)}=\frac{\ln(\text{cy5(x)/cy3(x)})}{\log 5} \tag{1}$$

$$\text{Mg} = \frac{1}{n}(x_1 + x_2 + x_3 + \cdots + x_n) \tag{2}$$

$$\text{Sg} = = \frac{1}{n}(|x_1 - \text{Mg}| + |x_2 - \text{Mg}| + |x_3 - \text{Mg}| + \cdots + |x_n - \text{Mg}|) \tag{3}$$

$$\text{Z(LNratio(x)},\text{M}_\text{g},\text{S}_\text{g})=\frac{\text{LNratio(x)} - \text{M}_\text{g}}{\text{S}_\text{g}} \tag{4}$$

Where, Mg represents the mean, S_g represents the standard deviation and LNratio(x) represents the input value x. For example, Table 1 Consider the expression levels training set to compute standardized normal distribution.

Table 1. Expression levels training set

Gene	a		b		C	
Tissue	Diseased	Healthy	Diseased	Healthy	Diseased	Healthy
Sample 1	30	10000	100	12000	2000	10
Sample 2	50	30000	50	2000	1000	1099
Sample 3	30	5000	13000	80	4000	150
Sample 4	100	6000	80	15000	6000	25

Table 2. LNratio and Normalization of the training data set

Gene	LNratio(x)		Normalization(LNratio(x),M_g,S_g)		Mean	Standard Deviation
	a	B	A	b	M_g	S_g
S1	-8.31	-6.85	-0.86	-0.64	-2.53	6.74
S2	-9.15	-5.28	-1.37	-0.13	-4.85	3.15
S3	-7.32	7.28	-1.5	0.97	1.55	5.92
S4	-5.86	-7.49	-0.62	-0.88	-1.83	6.45

Then relevancy between genes is computed by finding the relevancy of each gene by calculating Relevance Count (RC). Based on TRC value, the gene relevancy is categorized into 3 types. They are more relevant gene (1), quieted gene (-1) and irrelevant gene (0). RC is calculated by using threshold, cy5 and cy3 values. A threshold ε can be specified by using well known methods for example, t test. Total Relevancy Count (TRC) is by doing summation of all RC value in each gene. Table 3 shows the TC value for each gene.

$$\text{RC }(\varepsilon, \text{cy5, cy3}) = \begin{cases} 1 & \text{Z}\big(\text{LNratio(x)}, \text{M}_\text{g}, \text{S}_\text{g}\big) > \varepsilon \\ 0 & -\varepsilon \le \text{Z}\big(\text{LNratio(x)}, \text{M}_\text{g}, \text{S}_\text{g}\big) \le \varepsilon \\ -1 & \text{Z}\big(\text{LNratio(x)}, \text{M}_\text{g}, \text{S}_\text{g}\big) < -\varepsilon \end{cases} \tag{5}$$

Table 3. TRC value for threshold $\varepsilon = 0.5$

Gene	A	b	c
Sample 1	-1	-1	1
Sample 2	-1	0	1
Sample 3	-1	1	1
Sample 4	-1	-1	1
TRC	-4	-1	4

The weight (v_a, v_b) between the genes are calculated by applying Modified Manhattan distance formula. It is defined in equation (6).

$$\text{WG}\ (v_a, v_b) = |v_{a1} * v_{b1}| + |v_{a2} * v_{b2}| + |v_{a3} * v_{b3}| + \ldots + |v_{an} * v_{bn}| \quad (6)$$

Table 4. Represents the weight for gene expression tree as matrix

Gene	A	B	c
A	0	3	4
B	3	0	3
C	4	3	0

5. Conclusion

This paper proposes a Modified Manhattan distance based weighted GEG classifier to classify a gene expression data. The experimental outcome confirms that the efficiency of the proposed method gives better result than that of current methods. The proposed method can also appropriately distinguish out-of-class samples. In addition to that, the proposed method could properly classify samples in the relevant classes. The proposed method is going to improve graph based data structure by adding Euclidean distance to determine the relationships among genes. In this method, relevant genes are extracted by means of weight assigned to each gene, where larger weights point out a healthier relationship between two genes. As a result, the proposed method can decrease the cost acquired by classifying inappropriate genes. One of the major works of the proposed method having ability to appropriately categorizes into the relevant classes and also properly distinguishes out of class samples. In addition, it will significantly reduce computation time taken for classifying the data.

References

[1] A Benso, S DiCarlo, G Politano, L Sterpone. *A Graph Based Representation of Gene Expression Profiles in DNA Microarrays.* Proc. IEEE Symp. Computational Intelligence in Bioinformatics and Computational Biology (CIBCB). 2008: 75-82.

[2] Benso, S Di Carlo, G Politano, L Sterpone. *Differential Gene Expression Graphs: A Data Structure for Classification in DNA Microarrays.* Proc. Eighth IEEE Int'l Conf. BioInformatics and BioEng. (BIBE). 2008.

[3] J Breiman, L ad Friedman, CJ Stone, R Olshen. Classification and Regression Trees. New York, NY, USA: Talyor and Francis. 1984.

[4] H Zhang, CY Yu, B Singer. *Cell and tumor classification using gene expression data: Construction of forests.* Proc. Nat. Acad. Sci. USA. 2003; 100(7): 4168-4172.

[5] VN Vapnik. An Overview of Statistical Learning Theory. *IEEE Trans. Neural Networks*. 1999; 10(5): 988-999.

[6] J Khan, JS Wei, M Ringner, LH Saal, M Ladanyi, F Westermann, F Berthold, M Schwab, CR Antonescu, C Peterson, PS Meltzer. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Med.* 2001; 7(6): 673-679.

[7] Benso, S Di Carlo, G Politano. A cDNA microarray gene expression data classifier for clinical diagnostics based on graph theory. *IEEE/ACM Trans. Comput. Biol. Bioinformat.* 2011; 8(3): 577-591.