Academy of Scientific Research and Technology
27th National Radio Science Conference
Faculty of Electronic Engineering, Menoufia Univ., Menouf, Egypt
16-18 March 2010

Identifying Candidate Informative Genes for Biomarker Prediction of Liver Cancer

Nagwan M. Abdel Samee[1], Nahed H. Solouma [2], Mahmoud Elhefnawy[3], Abdalla S. Ahmed[4], Yasser M. Kadah [5]

[1]*Computer Engineering Department, Misr University for Science and Technology, Egypt*
[2]*Engineering Applications Department, NILES, Cairo University, Giza, Egypt*
[3]*Informatics and System Department, NRC, Giza, Egypt*
[4,5]*Systems and Biomedical Engineering Department, Cairo University, Giza, Egypt*

Abstract

Biomarker determination is a very important issue in medical research. Knowing the biomarker, better prognosis and diagnosis could be reached. Experimental prediction of biomarkers is lengthy and very costly. Recently, many studies have been conducted in the area of bioinformatics for In-silico prediction of biomarkers. This is assumed to save a lot of money while enabling us to explore many genetic issues. In this paper we provide a method to extract the most likely genes that could be considered as informative genes in microarrays of Hepatocellular Carcinoma. The used dataset is composed of fifty samples (19 normal and 31 abnormal samples). We extract the informative genes as those having the highest entropy. Those genes are considered as candidate biomarkers. The genes that have the highest mutual information with the well-known tumor biomarker Alpha fetoprotein gene are considered as candidate biomarkers. The results are validated by examining the presence of the experimentally known biomarkers among the extracted ones to get a confidence level.

1. Introduction

DNA—in the form of genes—carries the information necessary for making proteins in all known life. Which proteins are created, and in what quantity and at what time, determines the structure, function, and behavior of all cells. Genes are expressed by being transcribed into RNA, and this transcript may then be translated into protein. That protein has a role in the cell function. So, if the supposed expression of a specific gene changes, this will lead to a defect in cell function that depend on it and this imbalance could lead to the emergence of a particular disease. It follows that genes that play a role in the occurrence of a specific disease would have a variation in their expression from normal to sick samples.

Biomarkers are biological molecules produced in our bodies that are indicative of the presence of a specific disease. They are essential for diagnosis, prognosis and detection of diseases. Prediction of new biomarkers is done in biological lab. But we need more time and money to do an experiment to discover a new biomarker. So, the Insilco prediction of biomarkers can help in discovering several biomarkers with minimum effort and cost. This is performed here by making analysis of gene expression values measured by microarrays. The DNA microarray is a recent technology that can be used to measure expression levels of thousands of genes simultaneously. It takes advantage of hybridization properties of nucleic acid and uses complementary molecules attached to a solid surface, referred to as *probes*, to measure the quantity of specific nucleic acid transcripts of interest that are present in a sample, referred to as the *target*. The molecules in the target are labeled, and a specialized scanner is used to measure the amount of hybridized target at each probe, which is reported as intensity. Various manufacturers provide a large assortment of different platforms. The different platforms can be divided into two classes: *high-density oligonucleotide array* (Affymetrix) and *two-color spotted* platforms.

Statistical methods have a great role in analyzing different values of gene expression values measured from several samples under different conditions. Estimating variations in gene expression values was performed by the help of entropy estimation [1]. Moreover, the relationships between genes according to their expression values were inferred by evaluating their mutual information [2]. So, in this study we are using these techniques to predict new biomarkers of liver cancer. This is performed through the following three steps. Firstly the

variations in gene expression values are estimated by evaluating its entropy. Secondly, Genes with high entropy are considered as informative genes. Finally, the mutual information between a well known biomarker of liver cancer, Alpha fetoprotein (Afp), and the informative genes are evaluated. Genes with high mutual information with Afp are considered novel biomarkers of liver cancer.

2. Methodology

If the expression values of a gene are constant in normal and tumor samples then this gene does not play a role in the appearance of that tumor. We can call that gene non informative gene. Informative genes are genes that have variations in their expression from normal to tumor sample. Variations in measurement can be measured by the help of entropy estimation. The entropy of a discrete random variable X with possible values $\{x_1, x_2, x_3, \ldots . x_n\}$ can be calculated by equation 1.

$$H(X) = -\sum_{k=1}^{n} p(x_k) \log\, p(x_k) \qquad (1)$$

In equation 1, p is the probability mass function of X. But the probability mass function for gene expression values is not known, so an entropy estimator, the empirical entropy [3], is used. Moreover, the gene expression values need to be discretized into a number of intervals, bins, before evaluating its entropy. The empirical estimator is also called Maximum likelihood defined by equation 2.

$$H(X) = -\sum_{k=1}^{n} \frac{nb(x_k)}{m} \log \frac{nb(x_k)}{m} \qquad (2)$$

In equation 2, $nb(x_k)$ is the number of data points in bin k, m is number of observations (the number of gene expression values) and n is the number of bins. The number of bins that is selected in this study equals two. The reason of selecting that number is that the informative genes should have similar expression values in normal samples than those in tumor samples.

The relationships between Afp gene and the informative genes are evaluated using Mutual information. The Mutual information measures the amount of information that can be obtained about one random variable by observing another [4]. Formally, the mutual information of two discrete random variables X and Y can be defined as in equation 3.

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \frac{p(x,y)}{p_1(x)\, p_2(y)} \qquad (3)$$

In equation 3, p(x,y) is the joint probability distribution function of X and Y, and p1(x)and p2(y)are the marginal probability distribution functions of X and Y respectively. Intuitively, mutual information measures the information that X and Y share: it measures how much knowing one of these variables reduces our uncertainty about the other. For example, if X and Y are independent, then knowing X does not give any information about Y and vice versa, so their mutual information is zero. At the other extreme, if X and Y are identical then all information conveyed by X is shared with Y: knowing X determines the value of Y and vice versa. As a result, in the case of identity the mutual information is the same as the uncertainty contained in Y (or X) alone, namely the entropy of Y (or X: clearly if X and Y are identical they have equal entropy). So, Mutual information can be equivalently expressed as in equation 4.

$$\left.\begin{aligned} I(X;Y) &= H(X) - H(X \mid Y) \\ &= H(Y) - H(Y \mid X) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned}\right\} \quad (4)$$

In equation 4, H(X) and H(Y) are the marginal entropies, H(X|Y) and H(Y|X) are the conditional entropies, and H(X,Y) is the joint entropy of X and Y. Intuitively, if entropy H(X) is regarded as a measure of uncertainty about a random variable, then H(X|Y) is a measure of what Y does not say about X. This is "the amount of uncertainty remaining about X after Y is known", and thus the right side of the first of these equalities can be read as "the amount of uncertainty in X, minus the amount of uncertainty in X which remains after Y is known", which is equivalent to "the amount of uncertainty in X which is removed by knowing Y". This corroborates the intuitive meaning of mutual information as the amount of information (that is, reduction in uncertainty) that knowing either variable provides about the other.

3. Microarrays dataset

Hepatocellular Carcinoma (HCC) is a major type of liver cancer. It rises as a consequence of underlying liver diseases such as viral hepatitis and liver cirrhosis. Hepatitis B virus (HBV), hepatitis C virus (HCV) and intakes of alcohol are widely recognized as the three major etiological factors of HCC. But, HCV is a predominant cause of HCC. RNA expression data for liver samples from subjects with HCC as a complication of HCV cirrhosis are used in this study. Fifty microarray samples are downloaded from Gene Expression Omnibus (GEO) [5]. Thirty one of theses samples are for subjects with HCV cirrhosis and HCC. The remaining nineteen samples are for normal subjects. These data are collected on the Affymetrix HG-U133A 2.0 platform. The raw data in ".CEL" format are collected from GEO and an up-to-date probe set definition (.CDF file) based on Entrez Gene sequence is used in place of the Affymetrix original probe set definition provided by Bioconductor [6].

4. Gene expression data processing and analysis

In this work, the prediction of new biomarkers from Affymetrix microarrays is done through four modules. These modules are illustrated in Figure 1.

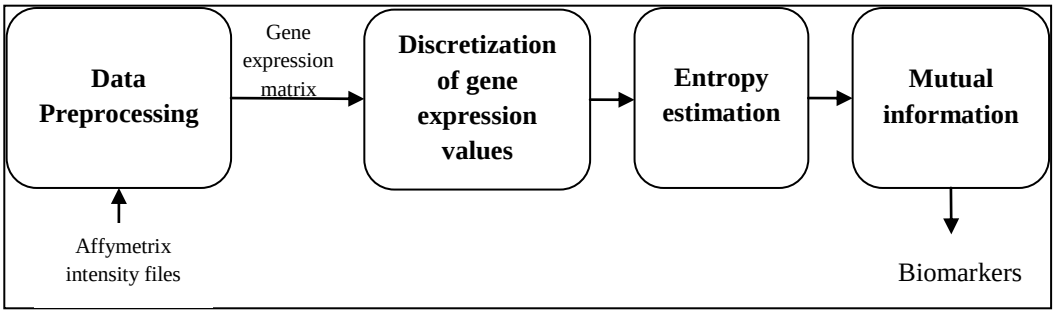

Figure 1. The basic modules for prediction of biomarkers using evaluation of entropy and mutual information

Firstly, the data is preprocessed to generate a gene expression matrix. Rows of that matrix are genes, columns are microarrays samples and entries of the matrix are gene expression values. Here, the Affymetrix gene chips are preprocessed using RMA method [7] implemented by Affy package [8]. Secondly, the gene expression values are discretized using the equal width discretization method [9]. Thirdly, the empirical estimator is calculated to discover the informative genes and finally the mutual information with Afp is evaluated.

5. Results and discussion

The entropy of a number of 22,277 genes was calculated. Genes with highest entropy values were selected using the histogram of the entropy which is illustrated in figure 2.

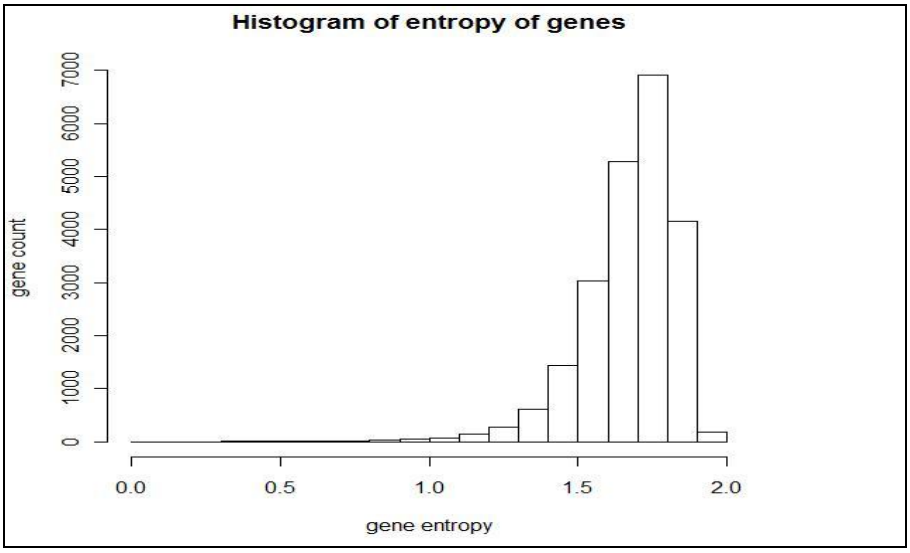

Figure 2. Histogram of genes entropy values

As noticed from figure 2, the number of genes that have the highest entropy values is 7000 genes. Those genes are considered as informative genes that could be used in predicting novel biomarkers of liver cancer. Investigating the experimentally-detected biomarkers, we found that these biomarkers (genes) are of higher rank in the ranked mutual information arrays. So, the mutual information between Afp and those genes are evaluated. Genes that have high mutual information with Afp could be considered as new biomarkers.

6. Conclusion

In conclusion, we can extract new tumor biomarkers from microarray datasets by extracting the informative genes. Informative genes are those having the highest entropy. These genes are biomarker candidates. Based on the relation between an informative gene and a well-known tumor marker (Alpha fetoprotein gene in case of HepatoCelluar Carcinoma), we can predict some new biomarkers as those having the highest mutual information with the known biomarker.

References

[1] Paninski, L.," Estimation of entropy and mutual information", *Neural Computation*, 15(6):1191-1253, 2003.
[2] Butte, A.J., Kohane, I.S.," Mutual Information Relevance Networks: Functional Genomic Clustering Using Pairwise Entropy Measurments.", Pac Symp Biocomput,418-429, 2000.
[3] Olsen, C., Meyer, P.E., Bontempi, G.,"On the Impact of Entropy Estimator in Transcriptional Regulatory Network Inference", In 5th International Workshop on Computational Systems Biology (WSCB 08) , 2008.
[4] Meyer, P.E., Lafitte, F. and Bontempi G., "minet: A R/Bioconductor Package for Inferring Large Transcriptional Networks Using Mutual Information", BMC Bioinformatics, 9:461, 2008.
[5]GEO: http://www.ncbi.nlm.nih.gov/geo.
[6] The BioConductor Project [http://bioconductor.org/].

[7] Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U. and Speed, T.P., "Exploration, normalization, and summaries of high density oligonucleotide array probe level data", Biostatistics, 4:249-64, 2003.
[8] Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., et al., " Bioconductor: open software development for computational biology and bioinformatics", Genome Biol, 5:R80, 2004.
[9] Liu, H., Hussain, F., Tan, CL., and Dash, M., " Discretization: An Enabling Technique", Data Mining and Knowledge Discovery, 2002.