Analysis of Liver Cancer DNA Sequence Data using Data Mining

N.Senthil Vel Murugan
Department of Mathematics
St.Joseph's College of Engineering, Chennai

V.Vallinayagam, PhD.
Department of Mathematics
St.Joseph's College of Engineering, Chennai

K. Senthamarai Kannan, PhD.
Department of Statistics
Manonmaniam Sundaranar University, Tirunelveli

T. Viveka
Department of Computer Science
Lord Jagannath College of Engg. &
Technology

ABSTRACT

Data mining is flourishing during recent years and it is establishing itself as a major discipline in Computer Science and Statistics with industrial relevance. Data mining is finding interesting structure in databases [2]. DNA is an extraordinary chip data with thousands of attributes which represents the gene expression values [4]. The DNA data set are stored in huge biological databases for several purposes [1].Cancer occurs when lumps of cells usually group together to form tumors. A growing tumor becomes a lump of cancer cells that can destroy the normal cells around the tumor and damage the body's health tissues. In the last two decades the researchers have drawn much attention about liver cancer. Liver cancer is a disease in which malignant cells form in the tissues of the liver. It is relatively rare form of cancer but has a high mortality rate. The aim of this paper is to analyze the liver cancer DNA sequence data using the generalization of Kimura Models and Markov Chain. The reasonable results verify the validity of our method.

Keywords: DNA; Data Mining; Liver Cancer; Markov Chain

1. INTRODUCTION

Data mining is the latest field which Interfaces computer Science and Statistics using advantages in both the fields, like extracting information from large databases. Now data mining serves as a subset of Statistics. It can assist clinical diagnosis to make informed decision and improve health service. Data mining makes use of ideas, tools, and methods from other areas such as database technology.

Cancer is actually a group of many related diseases that all have to do with cells. Cells are the very small units that make up all living things, including the human body. Cancer cells that are not normal grow and spread very fast. Cancer cells usually group or clump together to form tumors. A growing tumor becomes a lump of cancer cells that can destroy the normal cells around the tumor and damage the body's health tissues. The study tries to extract meaningful information large experimental data sets Present evidence, based on systematic studies of data from Gen Bank database.

The largest and perhaps the most resilient of all the organs in the body, the liver is also one of the most mysterious. It is in fact responsible for over 500 functions including regulating sex hormones, controlling cholesterol levels and vitamin and mineral supplies, warding off viruses and disposing of toxic material from the blood. It is also the only organ that has the ability to regenerate. Liver cancer is the third most deadly cancer worldwide. Liver cancer is a disease in which malignant cells form in the tissues of the liver. It is relatively rare form of cancer but has a high mortality rate.

Deoxyribonucleic acid (DNA) micro-arrays present a powerful means of observing thousands of gene terms levels at the same time. DNA is a one-dimensional fragment, made of two paired strands, coiled around each other as a double helix and held together by hydrogen bonds that connect a linear sequence of complementary pairs of bases. There are four types of bases, referred as C, G, A, T; the bond pairs are G-C and A-T. DNA inhabits in the nucleus of cells. A gene is a part of DNA, which include the formula for the chemical composition of one extracting protein. The genome holds the collection of all the genes that code for the entire proteins that an organism wants and produces.

2. DESCRIPTION OF MODEL

As Described by Kimura (1980) [3] in molecular evolution, the nucleotide substitutions in Eukaryotes were best described by Markov chains with continuous time. In these cases, the four DNA bases {A, T, G, C} are generated by a Markov chain with continuous time with transition rates $\{\alpha,\ \beta,\ \gamma,\ \varsigma,\ \in,\ \kappa,\ \lambda,\ \sigma\}$ as described below:

$$\begin{array}{c} \\ A \\ G \\ C \\ T \end{array}\begin{array}{c} \begin{array}{cccc} A & G & C & T \end{array} \\ \begin{pmatrix} 1-\alpha-\gamma-\lambda & \alpha & \gamma & \lambda \\ \in & 1-\in-\gamma-\lambda & \gamma & \lambda \\ \varsigma & \kappa & 1-\beta-\varsigma-\kappa & \beta \\ \varsigma & \kappa & \sigma & 1-\varsigma-\kappa-\sigma \end{pmatrix} \end{array}$$

The matrix P(t) of transition probabilities from states at time 0 to states at time t. We will show the matrix P(t) has four distinct real eigen values $\{v_i,\ i=1,\ 2,\ 3,\ 4\}$, so that B is diagonable. Let M_1 be a 4×4 matrix defined by:

$$M_1 = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \qquad M_1^{-1} = \begin{pmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \text{ and}$$

$$C = M_1^{-1} P(t) M_1$$

$$= \begin{pmatrix} 1-\alpha-\gamma-\lambda-\varsigma & \alpha-\kappa & \gamma-\sigma & 0 \\ \in-\varsigma & 1-\in-\gamma-\lambda-\kappa & \gamma-\sigma & 0 \\ 0 & 0 & 1-\beta-\varsigma-\kappa-\sigma & 0 \\ \varsigma & \kappa & \sigma & 1 \end{pmatrix}$$

Since the characteristic function of P(t) is

$$\phi(x) = \left| M_1^{-1} P(t) M_1 - x I_4 \right| = \left| C - xI_4 \right| \;\ldots\ldots\ldots\ldots (1)$$

$$\phi(x) = (1-x)\,(1-\beta-\varsigma-\kappa-\sigma-x)$$
$$[(1-\alpha-\gamma-\lambda-\varsigma-x)\,(1-\in-\gamma-\lambda-\kappa-x) - (\alpha-\kappa)\,(\in-\varsigma)]$$

$$\ldots\ldots\ldots\ldots (2)$$

A special king of Markov Process is a Markov Chain where the system can occupy a finite or countable infinite number of states e_1, e_2, …..e_j, ……such that the future evolution of the process, once it is in a given state, depends only on the present state and not on how it arrived at that state. The transition probability matrix together with the initial probability distribution completely specifies a Markov Chain. Consider a stochastic process associated with a parameter t. suppose t takes values 1, 2, 3, .and let X_1, X_2… be the corresponding random variables. Suppose further that each of X1, X_2… is discrete random variable, with X_{11}, X_{12},…. as the values of X_{11}, $X_{21,}$ $X_{22,}$…. as the values of X_2 so on. Thus, the set $\{x_{\alpha\beta},\ \alpha = 1,2,3,\ldots\ldots\ \ \beta = 1,2,3,\ldots.\}$ is the state space of the stochastic process being considered. Suppose this set is a finite set, say A = {a_1, a_2, …..a_m}, so that each is equal to some $x_{\alpha\beta}$, is equal to some a_i, in the set A. thus here {1, 2, 3, ….} is the index set {a_1, a_2, …..a_m} are the states and A is the state space for the process being considered. The process is evidently a discrete-state-discrete-parameter process. Now, suppose X_1, X_2, … are such that the conditions as

(i) The values taken by X_2 depend upon the values taken by X_1.

(ii) The values taken by X_3 depend upon the values taken X_2 but not upon the values taken by X_1 and so on.

Suppose that a Markov chain has transition matrix P and that at time t the probability that the process is in state E_j is ϕ_j, j = 1, 2, 3, 4, ……s. This implies that the probability at time t+1 the process is in state j is $\varphi_j = \sum_{k=1}^{s} \phi_k\, p_{kj}$, j = 1, 2, 3, 4, ……s.---------------(3). In this case we say that the probability distribution $(\phi_1, \phi_2, \phi_3, \ldots\ldots \phi_s)$ is stationary; i.e. it has not changed between times t and t+1, and therefore will never change.

3. ANALYSIS

The matrix P(t) of transition probabilities from states at time 0 to states at time t. Thus the Transition probability matrix P(t) is given by

$$P(t) = \begin{array}{c|cccc} & A & G & C & T \\ \hline A & 0.3855 & 0.2517 & 0.2173 & 0.1455 \\ G & 0.2018 & 0.4354 & 0.2173 & 0.1455 \\ C & 0.2240 & 0.2108 & 0.35 & 0.2152 \\ T & 0.2240 & 0.2108 & 0.2372 & 0.328 \end{array} \text{ and}$$

$$C = \begin{array}{c|cccc} & A & G & C & T \\ \hline A & 0.1615 & 0.0409 & -0.02 & 0 \\ G & -0.022 & 0.2246 & -0.02 & 0 \\ C & 0 & 0 & 0.1128 & 0 \\ T & 0.2240 & 0.2108 & 0.2372 & 1 \end{array}$$

The eigen values of the transition matrix P is given in (2) are $\{v_1 = 0.312,\ \ v_2 = 0.1583;\ v_3 = 0.402,\ \ v_4 = 0.4031\}$

Let $\pi = \{\pi_1, \pi_2, \pi_3, \pi_4\}$ be a steady state distribution of the Markov Chain. Then $\pi P = \pi$

$$\pi = (\pi_1, \pi_2, \pi_3, \pi_4) \begin{pmatrix} 0.3855 & 0.2517 & 0.2173 & 0.1455 \\ 0.2018 & 0.4354 & 0.2173 & 0.1455 \\ 0.2240 & 0.2108 & 0.35 & 0.2152 \\ 0.2240 & 0.2108 & 0.2372 & 0.328 \end{pmatrix}$$
$$= (\pi_1, \pi_2, \pi_3, \pi_4)$$

$$-0.6145\pi_1 + 0.2018\pi_2 + 0.2240\pi_3 + 0.2240\pi_4 = 0\ldots..(4)$$
$$0.2517\pi_1 - 0.5646\pi_2 + 0.2108\pi_3 + 0.2108\pi_4 = 0\ldots.(5)$$
$$0.2173\pi_1 + 0.2173\pi_2 - 0.65\pi_3 + 0.2372\pi_4 = 0\ldots\ldots..(6)$$
$$0.1455\pi_1 + 0.1455\pi_2 + 0.2152\pi_3 - 0.672\pi_4 = 0\ldots\ldots(7)$$
$$\pi_1 + \pi_2 + \pi_3 + \pi_4 = 1 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(8)$$

Solve the above equations (4) to (7), we get

$$\pi_1 = 0.9091\,\pi_2 \ \ldots\ldots\ldots\ldots\ldots\ldots\ldots..(9)$$
$$\pi_3 = 1.2148\,\pi_2 \ \ldots\ldots\ldots\ldots\ldots\ldots\ldots(10)$$
$$\pi_4 = 0.3780\,\pi_2 \ \ldots\ldots\ldots\ldots\ldots\ldots\ldots.(11)$$

Substituting these values in (5), we get

$$\pi_1 = 0.2596, \qquad \pi_2 = 0.2856,$$
$$\pi_3 = 0.3469, \qquad \pi_4 = 0.1079$$

i.e. $\lim\limits_{n\to\infty} P^{(n)} = \{0.2596 \quad 0.2856 \quad 0.3469 \quad 0.1079\}$

Thus in the long run, the variable should spent about 25.96% of the time in state A, about 28.56% of the time in state G, about 34.69% of the time in state C and 10.79% of the time in state T.

4. CONCLUSION

The given study focuses at the level of biological modules, rather than individual genes, an approach that produces results that are biologically interpretable and statistically robust. The study thus tries to use biological knowledge in developing analytic techniques. From the point of view of long-term averages, over a long time period the random variable should spent about 25.96% of the time in state A, about 28.56% of the time in state G, about 34.69% of the time in state C and 10.79% of the time in state T. It reveals that the percentage is approximately same for all the states. Hence In future, the following symptoms are observed it may lead to liver cancer.

5. REFERENCES

[1] Alfred Ultsch et. al., 2004 Knowledge Discovery in DNA Microarray data of cancer patients with emergent self-organizing maps, ESANN'2004 proceedings, pg. 501-506

[2] Fayad, U.M et al., 1996 Advances in knowledge discovery and data mining, AAAI/MIT press.

[3] Kimura. M, 1980 A simple method for estimating evolutionary rate in a finite population due to mutational production of neutral and nearly neutral base substitution through comparative studies of nucleotide sequences, J. Molec. Biol. Pg. 111-120.

[4] Langdon et al., 2004, Genetic programming for mining DNA chip data from cancer patients, University College, London

[5] Tan Wai-Yuan,(2002) "Stochastic Models with Applications to Genetics, Cancers, Aids and other Biomedical systems", World Scientific Publishing Co. Ltd.

[6] Warren .J. Ewens et. al (2004), "Statistical methods in Bioinformatics", Springer Publication, New Delhi