Data Mining Technique Applied To DNA Sequencing

R. Jamuna

Abstract

CpG islands (CGIs) play a vital role in genome analysis as genomic markers. Identification of the CpG pair has contributed not only to the prediction of promoters but also to the understanding of the epigenetic causes of cancer. In the human genome [1] wherever the dinucleotides CG occurs the C nucleotide (cytosine) undergoes chemical modifications. There is a relatively high probability of this modification that mutates C into a T. For biologically important reasons the mutation modification process is suppressed in short stretches of the genome, such as ‘start’ regions. In these regions [2] predominant CpG dinucleotides are found than elsewhere. Such regions are called CpG islands. DNA methylation is an effective means by which gene expression is silenced. In normal cells, DNA methylation functions to prevent the expression of imprinted and inactive X chromosome genes. In cancerous cells, DNA methylation inactivates tumor-suppressor genes, as well as DNA repair genes, can disrupt cell-cycle regulation. The most current methods for identifying CGIs suffered from various limitations and involved a lot of human interventions. This paper gives an easy searching technique with data mining of Markov Chain in genes. Markov chain model has been applied to study the probability of occurrence of C-G pair in the given gene sequence. Maximum Likelihood estimators for the transition probabilities for each model and analogously for the model has been developed and log odds ratio that is calculated estimates the presence or absence of CpG islands in the given gene which brings in many facts for the cancer detection in human genome.

Keywords: technique, DNA, sequencing, genome, human genome
I. Introduction

CpG islands (CGIs) play a vital role in genome analysis as genomic markers. Identification of the CpG pair has contributed not only to the prediction of promoters but also to the understanding of the epigenetic causes of cancer. In the human genome [1] wherever the dinucleotides C-G occurs the C nucleotide (cytosine) undergoes chemical modifications. There is a relatively high probability of this modification that mutates C into a T. For biologically important reasons the mutation modification process is suppressed in short stretches of the genome, such as ‘start’ regions. [2] In these regions predominant CpG dinucleotides are found than elsewhere. Such regions are called CpG islands. The most current methods for identifying CG islands suffered from various limitations and involved a lot of human intervention which can be easily searched with Markov Chain based Java coding developed in this paper. Java program has been created to identify the presence or absence of CpG Islands in the given Genome sequence. Therefore, given an annotated training data set, our coding has the capability to find other specific nucleotides sequences in DNA. DNA methylation is an effective means by which gene expression is silenced. In normal cells, DNA methylation functions to prevent the expression of imprinted and inactive X chromosome genes. In cancerous cells, DNA methylation inactivates tumor-suppressor genes, as well as DNA repair genes, can disrupt cell-cycle regulation. DNA methylation in instances such as these has enormous impact in the prevention and treatment of human cancers.

II. Data mining with markov chains model

A Markov chain is a model that generates sequences in which the probability of a symbol depends only on the previous symbol. A Markov chain model [3] is defined by (a) a set of states, Q, which emit symbols and (b) a set of transitions between states. States are represented by circles and transitions are represented by arrows. Each transition has an associated transition probability, 

\[ a_{ij}, \]

which represents the conditional probability of going to state j in the next step, given that the current state is i. The sum of all transition probabilities from a given state must equal 1. The row values are summed up to unity in Table[1.1].
III. Finite Markov Chain

An integer time stochastic process, consisting of a domain $D$ of $m > 1$ states \{s$_1$... s$_m$\} and

1. Consider a $m$ dimensional initial distribution vector
   \[(p(s_1)... p(s_m))\].

2. It has $n \times n$ transition probabilities matrix $M = (a_{ijkl})$

For example, in a domain $D$ can be the letters \{A, C, T, G\}, $p(A)$ the probability of A to be the 1st letter in a sequence, and $a_{AG}$ the probability that G follows A in a sequence. For each integer $n$, a Markov Chain [4] assigns probability to sequences $(x_1...x_n)$ over $D$ (i.e, $x_i \in D$) as follows:

\[p((x_1, x_2,...,x_n)) = p(X_1 = x_1)\prod_{i=2}^{n} p(X_i = x_i \mid X_{i-1} = x_{i-1})\]
\[= p(x_1)\prod_{i=2}^{n} a_{x_{i-1},x_i}\]

Similarly, $(X_1,..., X_i ...)$ is a sequence of probability distributions over $D$.

There is a rich theory which studies the properties of such “Markov sequences” $(X_1... X_i ...)$.

| Table[1.1] |
|---|---|---|---|
| A | G | C | T |
| A | 0.95 | 0 | 0.05 | 0 |
| G | 0.2 | 0.5 | 0 | 0.3 |
| C | 0 | 0.2 | 0 | 0.8 |
| T | 0 | 0 | 1 | 0 |

Each directed edge say $A \rightarrow G$ is associated with the positive transition probability from $A$ to $G$. 
IV. Modeling CpG islands in Genome

In human genomes the pair CG often transforms to (methyl-C) G which often transforms to TG.

Hence the pair CG appears less than expected from what is expected from the independent frequencies of C and G alone. [5] Due to biological reasons, this process is sometimes suppressed in short stretches of genomes such as in the start regions of many genes. These areas are called CpG islands (p denotes “pair”).

The “-” model: Use transition matrix $A^- = (a^-_{st})$, Where:

$$a^-_{st} = \text{(the probability that } t \text{ follows } s \text{ in a non CpG island)}$$

V. Maximum Likelihood estimators for the transition probabilities

Maximum Likelihood estimators for the transition probabilities for each model is calculated with sample CpG – islands of a human DNA one estimates the following transition probabilities[6].

The following table shows:

a. Two Markov chain models: CpG islands (the ‘+’ model)

b. Remainder of the sequence (the ‘-’ model). Each row sums to 1. Tables are asymmetric.

Table of frequencies: Table [1.2]

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0.180</td>
<td>0.274</td>
<td>0.426</td>
<td>0.120</td>
</tr>
<tr>
<td>C</td>
<td>0.171</td>
<td>0.368</td>
<td>0.274</td>
<td>0.188</td>
</tr>
<tr>
<td>G</td>
<td>0.161</td>
<td>0.339</td>
<td>0.375</td>
<td>0.125</td>
</tr>
<tr>
<td>T</td>
<td>0.079</td>
<td>0.355</td>
<td>0.384</td>
<td>0.182</td>
</tr>
<tr>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0.300</td>
<td>0.205</td>
<td>0.285</td>
<td>0.210</td>
</tr>
<tr>
<td>C</td>
<td>0.322</td>
<td>0.298</td>
<td>0.078</td>
<td>0.302</td>
</tr>
<tr>
<td>G</td>
<td>0.248</td>
<td>0.246</td>
<td>0.298</td>
<td>0.208</td>
</tr>
<tr>
<td>T</td>
<td>0.177</td>
<td>0.239</td>
<td>0.292</td>
<td>0.292</td>
</tr>
</tbody>
</table>
To use the model for discrimination one calculates the log-odds ratio

\[
S(x) = \log\left( \frac{P(x | \text{mod} +)}{P(x | \text{mod} -)} \right) = \sum_{j=1}^{J} \log \frac{a_j^{x+h_j}}{a_{J+1}^{x+h_{J+1}}}
= \sum_{j=1}^{J} \beta_j^{x+h_j}
\]

References


Birney, E. "Hidden Markov models in biological sequence analysis". July 2001:

Boufounos, Petros, Sameh El-Difrawy, Dan Ehrlich. "Hidden Markov Models for DNA Sequencing".

