International Research Journal of Management, IT & Social Sciences
Available online at https://sloap.org/journals/index.php/irjmis/
Vol. 2 No. 7, July 2015, pages: 15~19
ISSN: 2395-7492
https://sloap.org/journals/index.php/irjmis/article/view/314

Data Mining Technique Applied to DNA Sequencing

R. Jamuna [a]

Article history:

Received: 20 March 2015
Accepted: 30 May 2015
Published: 31 July 2015

Keywords:

DNA;
Genome;
Technique;
Sequencing;
Human genome;

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 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, 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 the human genome.

Author correspondence:
R. Jamuna,
Associate Professor, Department of Computer Science S.R. College, Bharathidasan University, Trichy,
Email address: *rjamuna2002@yahoo.co.in*

1. Introduction

Bridges, S. M., & Vaughn, R. B. (2000, October), 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, Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996), 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

[a] Department of Computer Science S.R. College, Bharathidasan University, Trichy

mutation modification process is suppressed in short stretches of the genome, such as ‘start’ regions. Han, J., Pei, J., & Kamber, M. (2011), 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

Liao, S. H., Chu, P. H., & Hsiao, P. Y. (2012), 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. Ngai, E. W., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011), 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 an enormous impact in the prevention and treatment of human cancers.

2. Research Methods

The present study applied qualitative methods. All data is analyzed descriptively. It is used a paraphrase to explain, elaborate, and explore regarding the phenomenon belonging. The conclusion is the last remarked based on the previous discussion and result.

3. Results and Analysis

3.1 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. Romero, C., & Ventura, S. (2007), Markov chain model 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,

The 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].

3.2 Finite Markov Chain

An integer time stochastic process, consisting of a domain D of m>1 states {s_1… s_m } and

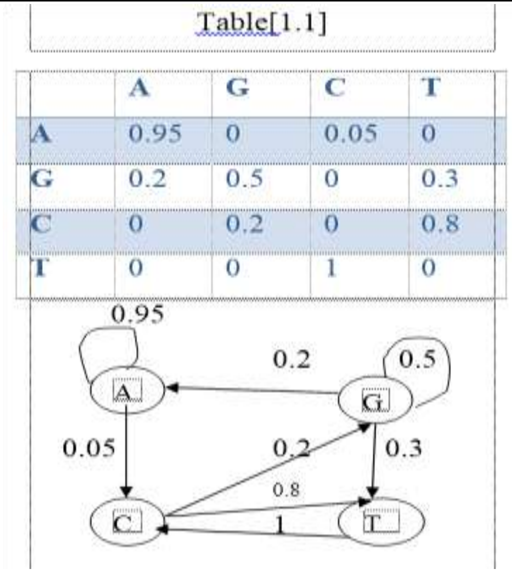
a) Consider a *m* dimensional initial distribution vector (p (s_1)... p(s_m)).
b) It has n×n transition probabilities matrix M= (a_{sisj})

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 a probability to sequences (x_1...x_n) over ***D*** (i.e, x_i ***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 says A→G is associated with the positive transition probability from A to G.

3.3 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. 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} = (the probability that t follows s in a non CpG island)

3.4 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) The remainder of the sequence (the '-' model). Each row sums to 1.Tables are asymmetric Table 2

Table 2
Table of frequencies

+	**A**	**C**	**G**	**T**
A	**0.180**	**0.274**	**0.426**	**0.120**
C	**0.171**	**0.368**	**0.274**	**0.188**
G	**0.161**	**0.339**	**0.375**	**0.125**
T	**0.079**	**0.355**	**0.384**	**0.182**
-	**A**	**C**	**G**	**T**

A	**0.300**	**0.205**	**0.285**	**0.210**
C	**0.322**	**0.298**	**0.078**	**0.302**
G	**0.248**	**0.246**	**0.298**	**0.208**
T	**0.177**	**0.239**	**0.292**	**0.292**

To use the model for discrimination one calculates the log-odds ratio

$$S\left(x\right)=\log\left(\frac{P\left(x\,|\operatorname{mod}el+\right)}{P\left(x\,|\operatorname{mod}el-\right)}\right)=\sum_{i=1}^{L}\log\frac{a^{+}_{x_{i-1}x_i}}{a^{-}_{x_{i-1}x_i}}$$

$$=\sum_{i=1}^{L}\beta_{x_{i-1}x_i}$$

4. Conclusion

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 an enormous impact in the prevention and treatment of human cancers.

Conflict of interest statement and funding sources

The author(s) declared that (s)he/they have no competing interest. The study was financed by personal funding.

Statement of authorship

The author(s) have a responsibility for the conception and design of the study. The author(s) have approved the final article.

Acknowledgments

The author would like to thank the reviewer for their consideration to the further process of the peer review. The author as well as thanks to the editor for their support, valuable time, and advice. Last but not least, the author thanks all researcher for their contribution as the references to the present article.

References

Bridges, S. M., & Vaughn, R. B. (2000, October). Fuzzy data mining and genetic algorithms applied to intrusion detection. In *Proceedings of 12th Annual Canadian Information Technology Security Symposium* (pp. 109-122).

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Liao, S. H., Chu, P. H., & Hsiao, P. Y. (2012). Data mining techniques and applications–A decade review from 2000 to 2011. *Expert systems with applications*, *39*(12), 11303-11311.

Ngai, E. W., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, *50*(3), 559-569.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*, *33*(1), 135-146.