Application of Variable Order Markov Models to Identifying CpG Islands

Zhenqiu Liu [1] **Dechang Chen** [2] **Jaques Reifman** [3]

Keywords: DNA sequence, CpG island, Markov chain, PST, identification, classification

1 Introduction

Identifying the location and function of human genes in a long sequence of genome is difficult due to lack of sufficient information about genes. Much research has been conducted in identifying CpG islands in DNA sequences using different models. First order Markov model and hidden Markov model (HMM) are among the most popular tools currently used [5]. Because of complexity of the real life sequence, the short memory assumption of the first order Markov chain usually is not satisfied. The HMM model, on the other hand, is more complex and can be slow for complex problems. It has been proved that HMMs can not be trained in polynomial time in the alphabet size. In addition, the algorithm of HMM can only be guaranteed to converge to a local minimum. Here we introduce one alternative model called variable order Markov chain. In the variable order Markov chain, the order of the Markov chain is not fixed [1], [4]. Variable order Markov models can be explained by probability suffix automata [2], [3], [6]. They are more succinct than higher order Markov chains and can be used to overcome the drawback that the size of Markov chain grows exponentially with its order. In addition, variable order Markov models are easy to compute and usually have high identification accuracies.

2 Discrimination with Markov models

A Markov model is fully defined by its states and state transition matrix. A variable order Markov chain is derived from the probability suffix tree. In a probability suffix tree, nodes are defined by the occurrence of symbols in DNA sequence. The nodal relationship is based on the parent being a suffix of its children. Each node also contains the probability of symbols that follow the given symbol in the sequence. An advantage of suffix tree model is that it is parsimonious. The tree order is kept to a minimum through excluding nodes that do not provide more stochastic information. The variable order Markov model is created entirely from the information provided in the probability suffix tree. The first step is to add all of the leaf nodes in the suffix tree to the Markov chain as states. The second step is to create a state that corresponds to the root of the suffix tree, and add any intermediate nodes from the root to leaf as additional new states. The third step is to find the transition probability for the given Markov states.

Markov models can be used to identify the CpG islands for DNA sequences. In order to do so, we need to train two Markov models separately: one for the CpG island, the other for the non-CpG island. For simplicity, denote CpG and non-CpG regions by '+' and '-',

[1]Bioinformatics Cell, TATRC, 110 North Market Street, Frederick, MD 21703, USA. E-mail: liu@bioanalysis.org

[2]Preventive Medicine and Biometrics, Uniformed Services University of the Health Sciences, 4301 Jones Bridge Road, Bethesda, MD 20814, USA. E-mail: dchen@usuhs.mil

[3]Bioinformatics Cell, TATRC, U.S. Army Medical Research and Materiel Command, Ft. Detrick, MD, USA. E-mail: reifman@tatrc.org

respectively. Let a_{ij}^{+} and a_{ij}^{-} represent the transitional probability for the CpG island non-CpG island, respectively. Given a test DNA sequence x of length n, we can discriminate the CpG island from non-CpG island by using the following log-likelihood ratio:

$$R(x) = \log \frac{P(x|model+)}{P(x|model-)} = \sum_{i=1}^{n} \log \frac{a_{ij}^{+}}{a_{ij}^{-}} = \sum_{i=1}^{n} \log a_{ij}^{+} - \sum_{i=1}^{n} \log a_{ij}^{-}.$$

If $R(x) > C$, where C is a predetermined positive constant, the sequence is the CpG island.

3 Computational results

We have identified the CpG islands from hundreds of DNA sequences and found that simple models can usually lead to high prediction accuracies. Here we present an example. One test sequence is a human collagen alpha-1-IV and alpha-2-IV genes, exons 1-3 (HSCOLAA). This sequence has 2184 symbols. We split it into subsequences with 100 symbols. The step for the window to move forward is 1. HMM, the first order Markov chain (MC1), the third order Markov model (MC3), and variable order Markov model (VMC) are all applied to the same sequence. The testing results for the HSCOLAA sequence are given in Table 1, which indicates that the simple variable Markov model with 58 states has the best performance.

Table 1: **Identification with Different Order Markov Chains**

True Islands	MC1	HMM	MC3	VMC
49-877	15-858	1-862	7-833	46-887
953-1538	908-1460	908-1445	916-1444	919-1489
1765-2100	1910-2015	1859-2011	1858-2007	1769-2007

References

[1] Apostolico, A. and Bejerano, G. 2000. Optimal amnesic probabilistic automata or how to learn and classify proteins in linear time and space. *RECOMB 2000.*

[2] Bejerano, G. and Yona, G. 1999. Modeling protein families using probabilistic suffix trees. *RECOMB 99* 15-24.

[3] Kermorvant, C. and Dupont, P. 2002. Improved smoothing for probabilistic suffix trees seen as variable order Markov chains. *ECML'02*, pp. 1-27.

[4] Laird, P. and Saul, R. 1994. Discrete sequence prediction and its applications. *Machine Learning,* 15:43-68.

[5] Lio, P. and Vannucci, M. 2000. Finding Pathogenecity islands and gene transfer events in genome data. *Bioinformatics*, vol. 16, no. 10-2000, pp. 932-940.

[6] Ron, D., Singer, Y. and Tishby, N. 1996. The power of amnesia: learning probabilistic automata with variable memory length. *Machine Learning*, 25, 117-142.