33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Identification of CpG islands in DNA Sequences Using Matched Filters

Rajasekhar Kakumani[1], M. Omair Ahmad[1], *Fellow*, *IEEE* and Vijay Devabhaktuni[2], *Senior Member, IEEE*
[1]Department of Electrical and Computer Engineering, Concordia University, 1455 de Maisonneuve Blvd. West Montreal, Quebec, H3G1M8, Canada
[2]Department of Electrical Engineering and Computer Science, University of Toledo, MS 308, 2801 W. Bancroft St., Toledo, OH 43606.
E-mail: omair@ece.concordia.ca

***Abstract*—CpG islands (CGIs), rich in CG dinucleotides, are usually located in the promoter regions of genes in DNA sequences and are used as gene markers. Identification of CGIs plays an important role in the analysis of DNA sequences. In this paper, we propose a new digital signal processing (DSP) based approach using matched filters for the identification of CGIs. We also formulate a new/reliable CGI identification characteristic replacing the several existing probability transition tables for CGIs and non-CGIs. The peaks in matched filter output, obtained by correlating the CGI characteristic with the DNA sequence to be analyzed, accurately and reliably identify CGIs. This approach is tested on a number of DNA sequences and is proved to be capable of identifying CpG islands efficiently and reliably.**

I. Introduction

CpG islands (CGIs) in DNA sequences have considerably high frequency of CG dinucleoetides as compared to non-CGIs. The CGIs are associated with promoter regions of most genes and known to influence gene expression hence playing an important role in the identification of promoters and genes in DNA sequences [1][2]. It is known that, CGIs occur in and around the promoter regions of 50% to 60% of human genes, including most housekeeping genes (the genes which are essential for general cell functions) [3]. CGIs have also contributed significantly to our understanding of the epigenetic causes of cancer. In cancer cells, CGIs are found to undergo a dense hypermethylation leading to gene silencing. Hence, the DNA methylation profiles in CGIs can be used for early detection of cancer [4]. These are some of the reasons which make identification of CGIs indispensable for genome analysis and annotation.

The initial CGI identification methods [5][6] relied on the following three characteristics of a CGI: (i) length of a CGI is at least 200 bp, (ii) G and C nucleotide content in a CGI is $\geq$ 50%, and (iii) observed CpG to expected CpG ratio (o/e) in a CGI is ≥ 0.6. Typically, the length of a CGI varies from a few hundred to a few thousand base pairs (bp), but rarely exceeds 5000 bp. Later on, sophisticated methods [7] utilizing two Markov chain models, one for CGIs and the other for non-CGIs, were proposed. These two models differ in their respective model parameters which characterize the difference in transition probabilities between successive nucleotides in CGIs and non-CGIs respectively. In these methods, a DNA segment is defined as a CGI, if the log-score computed using the Markov model for a CGI is greater than that computed using the Markov model for a non-CGI. The parameters used for modeling CGIs and non-CGIs play a crucial role in identifying CGIs. However, CGI identification methods using different Markov model parameters sometimes produce inconsistent results. Digital filters for identification of CGIs have also been proposed with considerable success. These methods are similar to Markov chain methods but use digital filters to compute weighted log-score to identify CGIs. The method proposed in [8] uses a bank of IIR low-pass filters to identify CGIs by looking at the weighted log-scores of all filters together. This method is computationally demanding as it employs large number of filters in the bank. From the above discussion, it is evident that the CGI identification methods and the criteria used therein play an important role in the identification of CGIs. Therefore, there is a need for developing fast and efficient computational methods using more reliable CGI identification characteristic.

In this paper, we propose a new DSP based approach using matched filters for identifying CGIs in DNA sequences. A new CGI identification characteristic is formulated which removes the ambiguity associated with the choice of the transition probability tables employed in some of the methods. Matched filters are then used to identify CGIs by detecting the corresponding locations of the CGI identification characteristic present in DNA sequences. The

approach is tested on several CGIs belonging to already annotated DNA sequences obtained from [10]. It is shown that the algorithm is simple to implement and yet able to identify CGIs reliably and efficiently as compared with other existing DSP based CGI identification methods.

II. Identification of CpG Islands

In this section a brief review of some of the existing DSP based CGI identification methods is presented.

A. Markov Chain Approach

In this method, a DNA sequence X of length N, represented as $X = \{x(n), x(n+1), \ldots, x(n+N-1)\}$ where each symbol $x(n) \in \{A, C, T, G\}$, is considered as a first-order Markov chain [7]. This is due to the fact that the probability of a particular nucleotide occurring at $x(n+1)$ depends only on the nucleotide having occurred at $x(n)$. The transition probabilities for the CGI and non-CGI Markov models are given in Table I and Table II respectively [7]. $p^{+}_{\beta\gamma}$ and $p^{-}_{\beta\gamma}$ are the probabilities of transition from a nucleotide β to a nucleotide γ in a CGI and a non-CGI respectively. These transition probabilities $p^{\pm}_{\beta\gamma}$ are calculated using

$$p^{\pm}_{\beta\gamma} = \frac{n^{\pm}_{\beta\gamma}}{\sum_{k\in\{A,T,G,C\}} n^{\pm}_{\beta k}} \tag{1}$$

where, $n_{\beta\gamma}$ is the number of $\beta\gamma$ dinucleoetides in the DNA sequence. In a CGI, the probability of transition from the nucleotide base C to the base G is higher in comparison with that in a non-CGI.

The probability of observing a windowed sequence $X_n = \{x(n), x(n+1), \ldots, x(n+L-1)\}$ of length L, assuming that it belongs to a CGI is given by

$$\begin{aligned} P(X_n \mid \text{CGI}) &= P(\{x(n),\ldots,x(n+L-1)\} \mid \text{CGI}) \\ &= \prod_{i=0}^{L-1} p^{+}_{x(n-1+i)x(n+i)} \end{aligned} \tag{2}$$

Similarly, the probability of observing X_n assuming it belongs to a non-CGI is

$$\begin{aligned} P(X_n \mid \text{non-CpG}) &= P(\{x(n),\ldots,x(n+L-1)\} \mid \text{non-CGI}) \\ &= \prod_{i=0}^{L-1} p^{-}_{x(n-1+i)x(n+i)} \end{aligned} \tag{3}$$

If the value of $P(X_n \mid \text{CpG}) > P(X_n \mid \text{non-CpG})$, then it is concluded that the sequence X_n belongs to a CGI. Otherwise, it is more likely to be a non-CGI. Alternatively, by formulating a log-likelihood ratio, given by

TABLE I
Transition Probabilities Inside the CGI Region

$p^{+}_{\beta\gamma}$	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

TABLE II
Transition Probabilities Inside the Non-CGI Region

$p^{-}_{\beta\gamma}$	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

$$S(n) = \frac{1}{L}\log\frac{P(X_n \mid \text{CpG model})}{P(X_n \mid \text{non-CpG model})}, \tag{4}$$

it can be concluded that the sequence X_n belongs to a CGI or otherwise depending on whether $S(n)$ is greater than or less than zero.

B. IIR Low-pass Filter Approach

Byung-Jun Yoon *et al.* [8] have noted that the log-likelihood ratio given in (4) can be expressed as:

$$\begin{aligned} S(n) &= \frac{1}{L}\log\prod_{n=0}^{L-1}\frac{p^{+}_{x(n-1)x(n)}}{p^{-}_{x(n-1)x(n)}} \\ &= \frac{1}{L}\sum_{i=0}^{L-1} y(n+i) \\ &= \lambda(n) * h_{avg}(n) \end{aligned} \tag{5}$$

where $\lambda(n)$ is a sequence representing the log-likelihood ratio of a single transition given by

$$\lambda(n) = \log\left(\frac{p^{+}_{x(n-1)x(n)}}{p^{-}_{x(n-1)x(n)}}\right) \tag{6}$$

and, $h_{avg}(n)$ is a simple averaging filter given by

$$h_{avg}(n) = \begin{cases} 1/L & \text{if } -L+1 \le n \le 0 \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

Then, they proposed using a bank of M filters such that the transfer function in the k^{th} (k = 0, …, M-1) channel is given by

$$H_k(z) = \frac{1-\alpha_k}{1-\alpha_k z^{-1}}, \tag{8}$$

where, $0 < \alpha_0 < \alpha_1 < ... < \alpha_{M-1} < 1$. The log-likelihood ratio obtained from the output of the k^{th} channel is given by

$$S_k(n) = \lambda(n) * h_k(n). \tag{9}$$

The values of $S_k(n)$ obtained for all k and n are then used to obtain a two-level contour plot. The bands corresponding to $S_k(n) > 0$ determine the locations of CGIs.

III. Proposed Approach

Matched filters are used for detection of signals of known shape but unknown gain by maximizing the signal to noise ratio. Let's consider that a DNA sequence X, is mapped to a corresponding binary indicator sequence X_{CG} [9], and the CGI identification characteristic be formulated using a binary sequence $\Phi = \{\phi(n)\}$. Now, in the proposed approach matched filters are employed to determine the degree to which X_{CG} resembles Φ. By employing an appropriate threshold on the matched filter output, we can arrive at a decision whether X_{CG} is a CGI or not.

The following subsections explain in detail the steps involved in identification of CGIs in a DNA sequence using matched filters.

A. Numerical Mapping of DNA Sequences

Identification of CGIs involves determination of G and C content in a DNA sequence. Hence, we define a new binary indicator sequence $X_{CG} = \{x_{CG}(n)\}$, where $1 \le n \le N$, such that $x_{CG}(n) = 1$ indicates the presence of the nucleotides C or G, and $x_{CG}(n) = 0$ indicates their absence at the location n in a DNA sequence of length N. For example, the DNA sequence X = {ATCCGAAGTATAACGAA} maps to the binary indicator sequence X_{CG} = {0011100100000100}.

B. CGI Identification Characteristic

A CGI contains frequent occurrence of CG dinucleotide and at least 50% of its nucleotide content is due to C and G. These two properties of CGIs can be combined to formulate a CGI identification characteristic in the form of a binary sequence Φ = {1100110011…001100}. The length of Φ is chosen to be equal to the length of the sliding window, L, used. Obviously, as CG dinucleotides occur more frequently in a CGI, the 1's in Φ appear as sets of two and the number of 1's constitute at least 50% of the elements in Φ.

C. Matched Filtering

The input DNA sequence X, of length N, is first mapped to an appropriate binary numerical sequence X_{CG}. A sliding window of length L is used to evaluate if each of the windowed sequences $X_n = \{x_{CG}(m)\}$ where n = 1, 2,…, N-L+1 and $m = n, n+1, \ldots, n+L-1$, belong to a CGI or a non-CGI. This is accomplished by correlating the CGI characteristic, Φ, with the windowed sequence, X_n. The matched filter output of the n^{th} window is given by

$$y(n) = \sum_{m=n}^{n+L-1} \phi(m-n+1) x_{CG}(m). \tag{10}$$

If the value of $y(n)$ is greater that an appropriate threshold η, then the windowed sequence X_n is considered to belong to a CGI. Consequently, all the peaks which are above the threshold η in the plot of the smoothened and normalized matched filter output $y(n)$ versus the base location index n, are the locations of CGIs in the DNA sequence X.

IV. Results

The proposed CGI identification approach based on matched filters is tested on the DNA sequence L44140 (GenBank accession number) from the human chromosome X. The sequence is of length 219447 bp and has 17 CGIs whose exact locations can be obtained from the NCBI website [10]. Performance of the proposed approach is then compared with the existing DSP based CGI identification approaches such as Markov chain method [7] and IIR low-pass filters method [8]. Fig. 1 shows the comparative performance of prediction of one of the CGIs in L44140, located between the nucleotides 3095 and 3426, by the three approaches mentioned above. A sliding window of length L = 100 is considered for all the methods.

Fig. 1(a) shows the performance of Markov chain approach, where log-likelihood ratio $S(n)$ is plotted against base index of the sequence, n. The transition probability parameters given in Table I and Table II are used to calculate $S(n)$. All the adjacent base locations, n, with $S(n) > 0$ are considered as a CGI. The exact location of the CGI (3095 to 3426) is shown by the horizontal line at the threshold $S(n)$ = 0. One of the major drawbacks of this approach is the presence of a lot of false positives that falsely categorize non-CGIs into CGIs.

The contour plot in Fig. 4(b) shows the performance of IIR low-pass filter method [8] where the filter coefficient α is plotted against base index, n, of the sequence. The transition probability parameters given in [8] are used to calculate $S(n)$. The orange/red regions located between base pairs 3000 and 3250 in contour plot denote the locations of CGIs as they correspond to the regions with $S(n) > 0$. It can be seen that the exact locations of the CGIs (boundaries of the orange/red regions) are difficult to obtain. Moreover, the method is computationally expensive as it involves plotting of contour plots for values of α varying from 0.95 to 0.99 with a step size of 0.001.

Fig. 1(c) shows the performance of the proposed matched filter based scheme in predicting the CGIs. Unlike the above mentioned methods, the proposed approach utilizes a unique

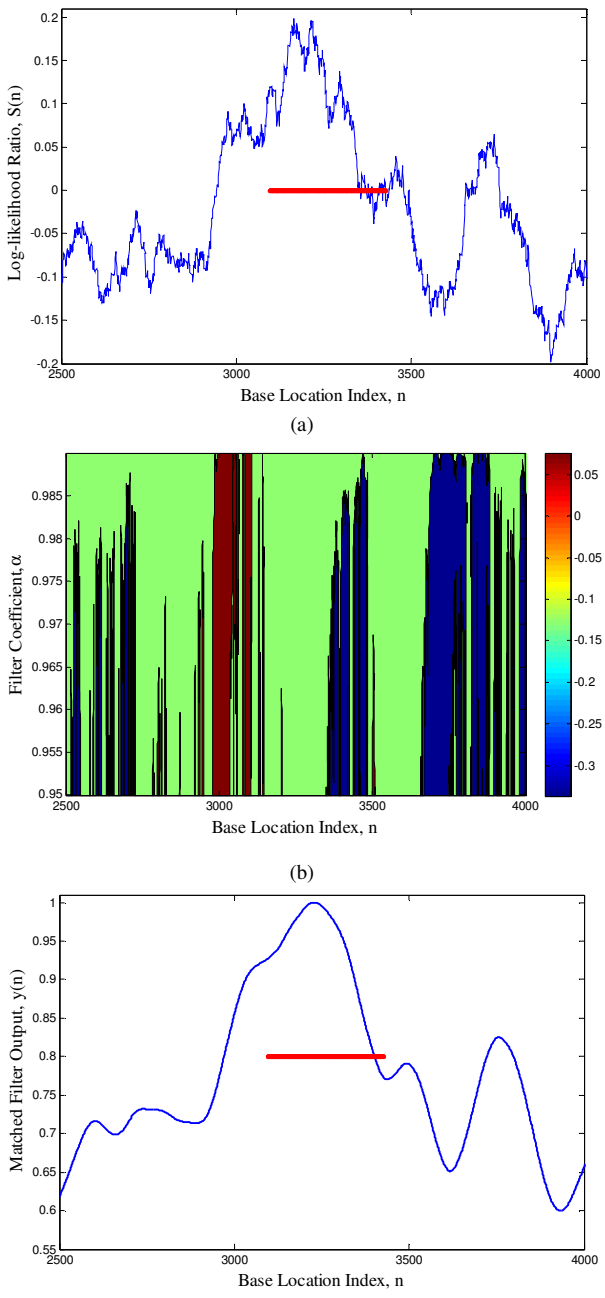

Fig. 1. Comparison of CGI identification using (a) Markov model method, (b) IIR low-pass filter method and (c) the Proposed matched filter approach.

CGI identification characteristic, $\Phi = \{\phi(n)\}$, instead of the probability transition tables. Effectiveness of our scheme is readily seen in Fig. 1(c), which depict more contrasting peak with less number of false positives as compared to the other two methods. A threshold of $\eta = 0.80$ is used to identify CGIs in this method. The proposed approach eliminates the ambiguity associated with the choice of probability transition tables by using a simple yet reliable CGI identification characteristic $\Phi = \{\phi(n)\}$. Moreover, the proposed approach is computationally more efficient as compared to the other two methods as it involves calculating simple correlation and the lookup tables in the form of transition probabilities are eliminated. Finally, the {sensitivity (*Sn*), and specificity (*Sp*)} of the Markov model method, IIR low-pass filter method and the proposed matched filter approach for the test sequences considered are {0.55, 0.60}, {0.58, 0.61} and {0.73 and 0.76} respectively.

IV. Conclusion

In this paper, an efficient and reliable DSP based approach for identification of CGIs is presented. It has been shown that the CGI prediction accuracy of the existing methods is greatly affected by the choice of the transition probability tables for CGIs/non-CGIs. The unique CGI identification characteristic proposed in our approach removes the ambiguity in choosing the appropriate probability transition tables. A matched filter has been employed to identify the locations of CGIs based on the peaks obtained by correlating the proposed CGI characteristic with the DNA sequence. Simulation results of the proposed approach have shown superior prediction accuracy in terms of sensitivity and specificity as compared with the other two DSP based CGI prediction methods.

Acknowledgement

This work was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada and in part by the Regroupement Stratgique en Microlectronique du Qubec (ReSMiQ).

References

[1] F. Larsen, G. Gundersen, R. Lopez, and H. Prydz, "CpG islands as gene markers in the human genome," *Genomics*, vol. 13, no. 4, pp. 1095–1107, 1992.

[2] F. Antequera and A. Bird, "CpG islands as genomic footprints of promoters that are associated with replication origins," *Current Biology*, vol. 9, pp. 661–667, 1999.

[3] F. Antequera and A. Bird, "Number of CpG islands and genes in human and mouse," *Proceedings of the National Academy of Sciences of USA*, vol. 90, no. 24, pp. 11995–11999, 1993.

[4] J. P. Issa, "CpG island methylator phenotype in cancer," *Nature*, vol. 4, no. 12, pp. 988-993, 2004.

[5] M. Gardiner-Garden and M. Frommer, "CpG islands in vertebrate genomes." *Journal of molecular biology*, vol. 196, no. 2, p. 261, 1987.

[6] D. Takai and P. Jones, "Comprehensive analysis of CpG islands in human chromosomes 21 and 22," *Proceedings of the national academy of sciences*, vol.99, no.6, p.3740-3745, 2002.

[7] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological sequence analysis*. Cambridge Univ. Press Cambridge, 1998.

[8] B. Yoon and P. Vaidyanathan, "Identification of CpG islands using a bank of IIR low-pass filters," *Proceedings of 11th Digital Signal Processing Workshop*.

[9] R. F. Voss, "Evolution of Long-range Fractal Correlations and 1/f noise in DNA base sequences," *Physical Review Letters*, vol. 68, pp. 3805-3808, June 1992.

[10] National Center for Biotechnology Information (NCBI) database. Available: http://www.ncbi.nlm.nih.gov/.