2004 IEEE 11th Digital Signal Processing Workshop
& IEEE Signal Processing Education Workshop

IDENTIFICATION OF CPG ISLANDS USING A BANK OF IIR LOWPASS FILTERS

Byung-Jun Yoon and P. P. Vaidyanathan

Dept. of Electrical Engineering
California Institute of Technology, Pasadena, CA 91125, USA
E-mail: bjyoon@caltech.edu, ppvnath@systems.caltech.edu

ABSTRACT

It has been known that biological sequences such as the DNA sequence display different kinds of patterns depending on their biological functions. This statistical difference can be exploited for identifying the region of interest, such as the protein coding regions or CpG islands, in a new biological sequence that has not been annotated yet. A region of particular interest is the CpG island, which is a region in a DNA sequence that is rich in the dinucleotide CpG, since it is known that they can be used as gene markers. There have been several computational methods for identifying CpG islands, each with its own strength and weakness. In this paper, we propose a novel scheme for detecting CpG islands in a genomic sequence, which is based on a bank of IIR lowpass filters. The proposed method is capable of identifying CpG islands efficiently at a low computational expense. Simulation results are included where appropriate to demonstrate the idea.

1. INTRODUCTION

It has been known that biological sequences such as the DNA sequence display different kinds of patterns depending on their biological functions. For example, it has been observed that the protein-coding regions in a genomic sequence show a period-3 behavior [1]. On the other hand, non-coding regions display long-range correlations instead of short-term periodicities [2]. This statistical difference can be exploited for identifying the region of interest (e.g. protein-coding regions, CpG islands, etc.) in a new biological sequence that has not been annotated yet. In fact, given the huge amount of genomic data that is available in the public domain these days, computational methods for analyzing biological sequences have become increasingly popular. Some of the interesting applications of computational methods in biological sequence analysis can be found in [3], [4] and a tutorial overview can be found in [5].

A region that is of particular interest is the CpG island, which is a region in a DNA sequence that is abundant in the dinucleotide CpG. This dinucleotide is usually denoted as CpG in order to distinguish it from the C-G base pair across the two strands in a DNA double helix. The CpG islands are of our interest for many biological reasons. For example, it has been shown that CpG islands can be used as gene markers, since they are located upstream of the transcription start regions of many genes [6]. Analysis of human genome shows that all housekeeping genes (which are genes that are expressed in all cells throughout the body, and produce proteins that are necessary for basic maintenance and cellular functions) and 40% of the tissue-specific genes are associated with CpG islands [6]. This makes them useful landmarks for identifying protein-coding regions in the human genome. Moreover, experiments have shown that methylation of CpG islands plays an important role in gene silencing [7], genomic imprinting [8], carcinogenesis [9], etc.

Due to the frequent occurrence of CpG dinucleotides as well as the high G+C content in such regions, several techniques have been proposed for identifying CpG islands in a genomic sequence each with its own strength and weakness [3], [10], [11], [12]. In this paper, we consider the Markov chain model elaborated in [3]. Based on this model, we propose a novel scheme for detecting CpG islands. The proposed method is based on a digital signal processing technique that uses a bank of IIR lowpass filters. Despite the simplicity of the proposed method, it is capable of identifying CpG islands efficiently at a very low computational cost.

2. IDENTIFICATION OF CPG ISLANDS

The first large-scale computational analysis of CpG islands traces back to the work of Gardiner-Garden and Frommer in 1987 [10]. They defined CpG islands as regions of at least 200 bp length, with a G+C content higher than 50% and the observed CpG to expected CpG ratio equal to or above 0.6. The exact definition of CpG islands is somewhat arbitrary, since the choice of the cut-off parameters can have a critical impact on which regions are included in the definition of the CpG islands. For example Takai and Jones redefined the CpG island as a region of DNA whose length is at least 500bp with a G+C content equal to or above 55% and observed CpG to expected CpG ratio above 0.65 [11]. Using this definition, they could find regions that are more likely to be associated with the 5' regions of genes while excluding most of the so-called *Alu*-repeats [11].

In addition to these simple schemes, there are other interesting approaches that make use of more sophiscated - hence, more powerful - techniques [3], [12], [13]. For example, the DNA sequence can be modeled using Markov chains. In [3], the CpG islands and the remainder of the genomic sequence are modeled separately using two Markov chains with different statistics, i.e. different transition probabilities. Given a short DNA sequence, we compute the log-score of the sequence for each model and compare the scores to choose the more likely one. This allows us to decide whether the region belongs to a CpG island or not.

Let us consider a sequence of nucleotides $x(n) \in \{A, C, T, G\}$. We assume that $x(n)$ forms a 1st-order Markov chain where the probability of each symbol $x(n+1)$ depends only on the current symbol $x(n)$. Now, let us denote the transition probability from a base β to a base γ in a CpG island and in a non-CpG island region as $p^+_{\beta\gamma}$ and $p^-_{\beta\gamma}$ respectively. For example, p^+_{AC} is the probability

Work supported in part by the ONR grant N00014-99-1-1002, USA.

that the next symbol will be a C given that the current symbol is an A inside a CpG island. Using these notations, the probability of observing the sequence $x(n)x(n+1)\cdots x(n+L-1)$, assuming that it belongs to a CpG island and that the previous symbol was $x(n-1)$ can be written as

$$\begin{aligned} & P(n|\text{CpG}) \\ &= P(x(n)\cdots x(n+L-1)|x(n-1), \text{CpG model}) \\ &= \prod_{i=0}^{L-1} p^{+}_{x(n-1+i)x(n+i)} \end{aligned} \tag{1}$$

Similarly, the probability of observing this sequence assuming it belongs to a non-CpG island region is

$$\begin{aligned} & P(n|\text{non-CpG}) \\ &= P(x(n)\cdots x(n+L-1)|x(n-1), \text{non-CpG model}) \\ &= \prod_{i=0}^{L-1} p^{-}_{x(n-1+i)x(n+i)} \end{aligned} \tag{2}$$

If $P(n|\text{CpG})$ is greater than $P(n|\text{non-CpG})$, we can conclude that the DNA sequence $x(n)x(n+1)\cdots x(n+L-1)$ belongs to a CpG island. Otherwise, it is more likely that the sequence does not belong to a CpG island. Therefore, if we define

$$S(n) = \frac{1}{L}\log\frac{P(n|\text{CpG})}{P(n|\text{non-CpG})}. \tag{3}$$

which is the log-likelihood ratio normalized by the length of the sequence, $S(n) > 0$ implies that the given DNA sequence is more likely to belong to a CpG-island, whereas $S(n) < 0$ implies that the sequence probably belongs to a non-CpG island region.

Despite the simplicity of this idea, it has been shown that this method works considerably well [3]. To see this, consider the following example. First, we took a DNA sequence of length 219447 from the human chromosome X (GenBank accession number L44140) that has been already annotated, and computed the transition probabilities for the two regions. These are shown in Table 1 and Table 2. Each row in the table contains the transition probabilities from a specific base to each of the four bases. For example, the first row of Table 1 contains the probabilities that each of the four bases will follow the base A inside CpG islands. Therefore, every row in the tables adds up to unity. By comparing Table 1 and Table 2, we can find one interesting fact about the transition probabilities. In Table 1, we can see that the probability that a C will be followed by a G is very high inside the CpG islands, resulting in many CpG dinucleotides. However, this is not the case outside the CpG islands. Table 2 shows that it is rather unlikely that a C will be followed by a G. This is a result of the methylation process which mutates a C into a T with a relatively high chance whenever it finds a CpG dinucleotide [14]. As a consequence, CpG dinucleotides appear much less frequently than they are expected. However, this methylation process is suppressed in the CpG islands, hence we can find more CpG dinucleotides than usual [14].

$p^{+}_{\beta\gamma}$	A	C	G	T
A	0.1598	0.2914	0.4247	0.1241
C	0.1299	0.3862	0.3093	0.1746
G	0.1425	0.3675	0.3675	0.1225
T	0.0758	0.3742	0.3687	0.1813

Table 1. Transition probabilities inside the CpG island region.

$p^{-}_{\beta\gamma}$	A	C	G	T
A	0.2499	0.2209	0.3526	0.1766
C	0.2810	0.3352	0.0941	0.2897
G	0.2159	0.2586	0.3397	0.1858
T	0.1283	0.2624	0.3594	0.2499

Table 2. Transition probabilities in the non-CpG island region.

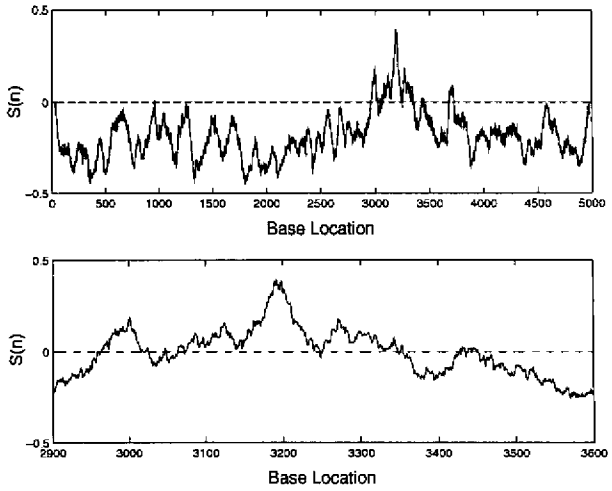

Fig. 1. (Top) CpG island prediction result using the Markov chain method. (Bottom) Magnified plot.

Figure 1 shows the prediction result of CpG islands based on this approach. Between the base locations 1 and 5000, there is only one CpG island of length 332 between 3095 and 3426. At this location, $S(n) > 0$ which implies that it is very likely that this region overlaps with a CpG island. Outside this region, $S(n)$ is mostly negative although there are some fluctuations. This plot shows that the CpG/non-CpG regions can be reasonably discriminated by looking at the sign of $S(n)$. However, if we look into the plots more closely, we can find that there are a lot of fluctuations around zero, resulting in many unwanted zero-crossings. The bottom plot in Figure 1 shows the magnified plot around the CpG island. We can see that the region around base location 3000 has positive values although it doesn't belong to a CpG island. Moreover, there are several zero-crossings inside the CpG island. Apparently, this is not what we expect. In the next section, we propose a new method that can relieve these problems and improve the prediction results.

3. IDENTIFYING CPG ISLANDS USING A BANK OF IIR LOWPASS FILTERS

When using the method elaborated in the previous section, it is not very obvious how to choose the window size L for computing $S(n)$. This is an important issue, since the choice of the window size can have a significant effect on the detection results. Larger windows usually enhance the reliability of the result but degrade the resolution of the output. On the contrary, smaller windows are

able to catch up with the changes of the statistical properties very quickly, but $S(n)$ may fluctuate around zero more often, thereby making the identification results less reliable.

Let us consider again the log-likelihood ratio $S(n)$ in (3). If we define $y(n)$ as the log-likelihood ratio of a single transition, that is,

$$y(n) = \log\left(\frac{p^{+}_{x(n-1)x(n)}}{p^{-}_{x(n-1)x(n)}}\right), \tag{4}$$

then S(n) can be rewritten as

$$\begin{aligned} S(n) &= \frac{1}{L}\sum_{i=0}^{L-1} y(n+i) \\ &= y(n) * h_{ave}(n). \end{aligned} \tag{5}$$

Here, $h_{ave}(n)$ is a simple averaging filter that is defined as

$$h_{ave}(n) = \begin{cases} \frac{1}{L} & -L+1 \le n \le 0 \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

Note that $h_{ave}(n)$ can be viewed as a simple lowpass filter. Instead of using a single filter that is rectangular in shape, we may use a bank of M filters, where each filter is a lowpass filter with a different bandwidth. By looking at the outputs altogether, we may be able to make a better decision on the locations of the CpG islands. This idea is shown in Fig. 2.

One way to construct such a filter bank is to use a one-pole filter

$$H_k(z) = \frac{1-\alpha_k}{1-\alpha_k z^{-1}} \tag{7}$$

in the k-th channel, such that $0 < \alpha_0 < \alpha_1 < \cdots < \alpha_{M-1} < 1$. This corresponds to $h_k(n) = (1-\alpha_k)\alpha_k^n u(n)$ in the time-domain. Choosing the filter $h_k(n)$ in this way results in weighted averaging of $y(n)$, where the more recent inputs are given larger weights than the past ones. Also note that $\sum_n h_k(n) = 1$, serving as a proper weighting function.

In order to demonstrate the idea, consider the following simulation. We chose α_k $(k = 0, 1, \ldots, 40)$ from 0.95 to 0.99 by increasing its value by 0.001, and $H_k(z)$ was chosen as (7). We computed $y(n)$ defined as (4) from the same input sequence (the human chromosome X) that we used for the averaging method in section 2. Then we filtered $y(n)$ using the filters $h_k(n)$ to obtain

$$S_k(n) = y(n) * h_k(n) \tag{8}$$

for all k. We combined these outputs altogether to obtain the contour plot shown in Fig. 3. The contour plot in Fig. 3 clearly shows the band which corresponds to the CpG island located between

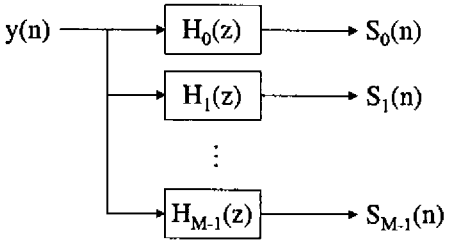

Fig. 2. A bank of M lowpass filters with different bandwidths.

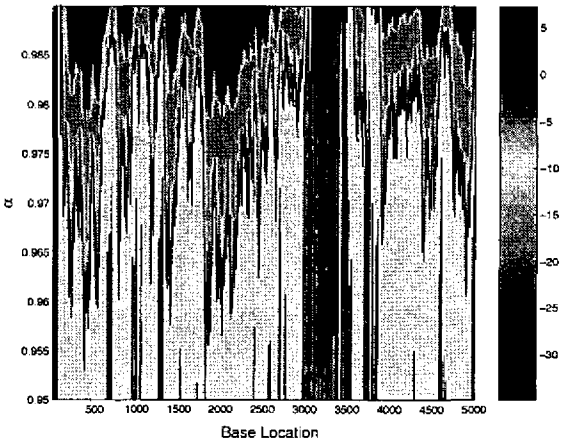

Fig. 3. Contour plot of the outputs $S_k(n)$. The red band in the middle clearly indicates the existence of a CpG island.

3095 and 3426 (colored in orange and red). This is more prominent in Fig. 4, which is a two-level contour plot of the same output $S_k(n)$. The level curves are located at $S_k(n) = 0$, and the shaded area indicates where $S_k(n)$ is positive. From Fig. 4, we can see that as α_k increases, there are less fluctuations around zero. For example, for $\alpha_k = 0.99$ there is only one small region inside the shaded band where $S_k(n)$ goes below zero. On the contrary, for $\alpha_k = 0.95$ there are more than 10 regions inside the band where $S_k(n)$ is negative. One more interesting thing to note in Fig. 4 is the fact that the shaded region slightly bends to the right as α_k increases. This shows that the response time of the filter $h_k(n)$ is longer for larger α_k. If α_k is large, the past samples are given more weights than when we use a smaller α_k. Effectively, this means that more samples of $y(n)$ are taken into account in computing $S_k(n)$. This allows us to obtain a smoother output with less fluctuation, but at the same time, the filter is slower in catching up with the changes in the input statistics. So, there is a trade-off between the responsiveness of the filter and the stability of the out-

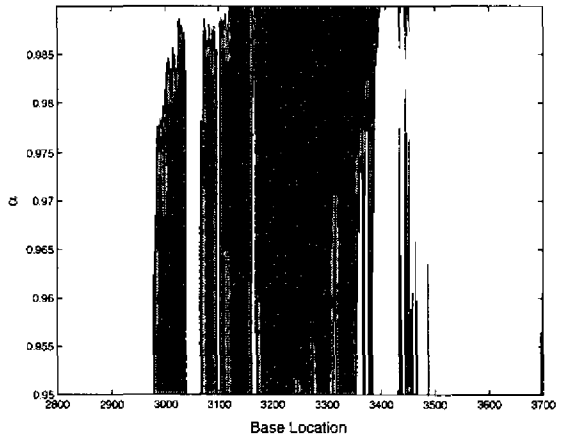

Fig. 4. Two level contour plot of the outputs $S_k(n)$. The level curve is located at zero, separating the positive and the negative regions.

put. This is indeed a very similar problem to that of choosing the window size L when using the averaging method in section 2, and this is the reason why we have to look at the outputs altogether instead of depending on a single output.

4. PREDICTION OF THE CHANGE POINTS

Now that we are given a number of outputs corresponding to different α_k, how can we predict the start and end points of the CpG islands more accurately? In order to answer this question, let us first consider the following. Assume that we have two distinct regions - a CpG island and a non-CpG island region - each of which can be modeled using a first-order Markov chain with different statistics. Now, using a rectangular window of length L, let us compute the weighted sum of the log-likelihood ratio $y(n)$ inside the window. Since we are using a rectangular window in this case, all $y(n)$ are weighted equally as in (5). Initially, let us assume that this window is inside the first region and doesn't overlap with the second region at all. Then the expectation of $S(n)$ can be simply written as

$$\begin{aligned} E[S(n)] &= \frac{1}{L}\sum_{i=0}^{L-1} E[y(n+i)] \\ &= \frac{1}{L}(L \cdot E[y(n)]) \\ &= E[y(n)]. \end{aligned} \tag{9}$$

Therefore, the expectation of $S(n)$ inside the CpG-islands will be $E[S(n)|\text{CpG}] = E[y(n)|\text{CpG}] > 0$, and $E[S(n)|\text{non-CpG}] = E[y(n)|\text{non-CpG}] < 0$ outside the CpG-islands. Now, consider gradually shifting the window to the right one-by-one. At some point, it will cross the change point between those two regions, and the window will have an overlap with both regions as shown in Fig. 5. If we let k be the length of the part of the window that overlaps with the second region, the expectation of $S(n)$ can be written as

$$\begin{aligned} E[S(n)] &= \frac{1}{L}\sum_{i=0}^{L-1} E[y(n+i)] \\ &= \frac{1}{L}\Big((L-k)\cdot E[y(n)|\text{region 1}] \\ &\qquad +k \cdot E[y(n)|\text{region 2}]\Big) \\ &= \frac{L-k}{L}E[y(n)|\text{region 1}] + \frac{k}{L}E[y(n)|\text{region 2}] \end{aligned} \tag{10}$$

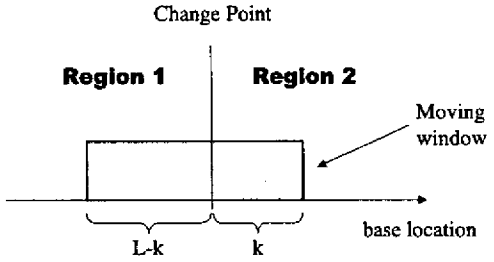

Fig. 5. A rectangular window of length L located around the change point.

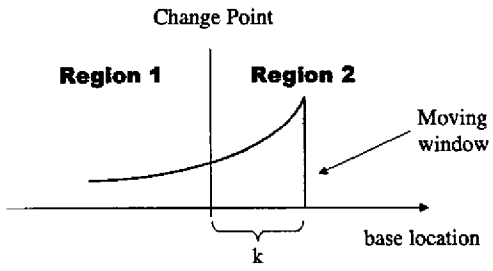

Fig. 6. An exponentially decaying window located around the change point.

where the sign of $E[S(n)]$ depends on L, k, $E[y(n)|\text{region 1}]$ and $E[y(n)|\text{region 2}]$. As we shift the window further, the overlap with the first region will decrease, and finally the whole window will be located inside the second region. Since the sign of $E[S(n)]$ in each region is different, at some point we can observe the change of sign of $E[S(n)]$. Let us denote the k $(0 < k < L)$ that satisfies $E[S(n)] = 0$ as k^*. If we solve for k^*, we get

$$k^* = \frac{E_2 L}{E_2 - E_1} \tag{11}$$

where $E_i = E[y(n)|\text{region } i]$. In fact, this is the point where we can first recognize the change between regions in the given sequence. Therefore, k^* can be viewed as the delay of the detection algorithm, and it can be computed once we know E_1 and E_2. Since we know the statistics of the respective Markov chains used for modeling the CpG/non-CpG island region, we can compute $E^+ = E[S(n)|\text{CpG}]$ and $E^- = E[S(n)|\text{non-CpG}]$ that are needed for computing the delay. We have

$$E^+ = \sum_{\beta,\gamma\in\{A,C,G,T\}} p^+_{\beta\gamma} \log\left(\frac{p^+_{\beta\gamma}}{p^-_{\beta\gamma}}\right) \tag{12}$$

and

$$E^- = \sum_{\beta,\gamma\in\{A,C,G,T\}} p^-_{\beta\gamma} \log\left(\frac{p^+_{\beta\gamma}}{p^-_{\beta\gamma}}\right). \tag{13}$$

Using the transition probabilities in Table 1 and Table 2, we get $E^+ = 0.0427$ and $E^- = -0.0412$. Therefore, when entering a CpG island from a non-CpG island region, we expect a delay of

$$k^* = \frac{E^+ L}{E^+ - E^-} = 25.04 \tag{14}$$

where $L = 51$. Similarly, when we leave a CpG island and enter a non-CpG island region, the expected delay is

$$k^* = \frac{E^- L}{E^- - E^+} = 25.96. \tag{15}$$

As we are using a rectangular window, we expect the delay to be around $L/2$, which is indeed the case.

Now let us consider using an exponential window defined as (7). This window is shown in Fig. 6. Again, we want to find the k that satisfies $E[S(n)] = 0$. In this case, $E[S(n)]$ can be written as

$$\begin{aligned} E[S(n)] &= \sum_{i=-\infty}^{L-1} (1-\alpha)\alpha^{L-1-i} E[y(n+i)] \\ &= \alpha^k E_1 + (1-\alpha^k)E_2. \end{aligned} \tag{16}$$

If we solve for the k^* that makes $E[S(n)] = 0$, we get

$$k^* = \log_\alpha \left(\frac{E_2}{E_2 - E_1} \right). \quad (17)$$

Based on the transition probabilities in Table 1 and Table 2, we get the plot of the delay k^* as a function of α as shown in Fig. 7. Now that we have computed the expected delays corresponding to different values of α, let us compare these with the actual zero-crossing points. We generated a random sequence of A, C, G and T based on the transition probabilities in Table 1 and Table 2. We computed $S_k(n)$ for $0.95 < \alpha_k < 0.99$ and computed the zero-crossing points. Figure 8 shows the level curves for $S(n) = 0$ with the theoretical curve obtained from (17). It can be seen that the theoretical curves are very close to the actual curves. Therefore, in order to predict the changing point of the two regions more accurately, we may first find the level curves for $S(n) = 0$ and find the theoretical curve of the zero-crossing points that matches the actual curve best. From this, we can make up for the delay and predict the actual change point more precisely.

5. CONCLUDING REMARKS

In this paper, we proposed a novel scheme for predicting CpG islands, which is based on a bank of IIR filters. It was shown that this method can locate the CpG islands efficiently at a low computational cost. The purpose of the paper was in proposing an idea that may improve the prediction results and therefore we haven't compared the performance with the existing systems. In order to compare this method with other existing algorithms, we may have to fine-tune the idea by incorporating the knowledge about typical lengths of CpG islands, etc. We may also use lowpass filters with better passband/stopband details to improve the prediction results. These are possible directions for future research.

6. REFERENCES

[1] E. N. Trifonov and J. L. Sussman, "The pitch of chromatin DNA is reflected in its nucleotide sequence", Proc. Nat. Acad. Sci. USA 77 (1980) 3816-3820, .

[2] C. K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, H. E. Stanley, "Long-range correlations in nucleotide sequences", Nature 356 (1992) 168-170.

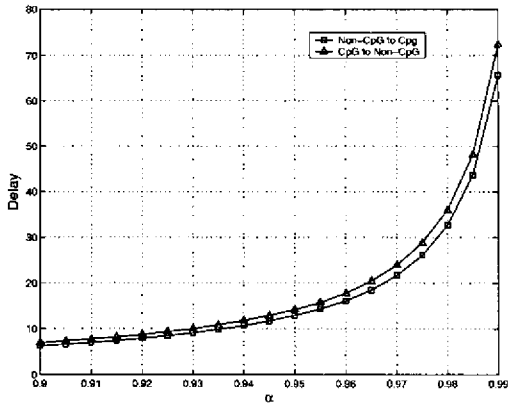

Fig. 7. The delay k^* corresponding to the value α_k.

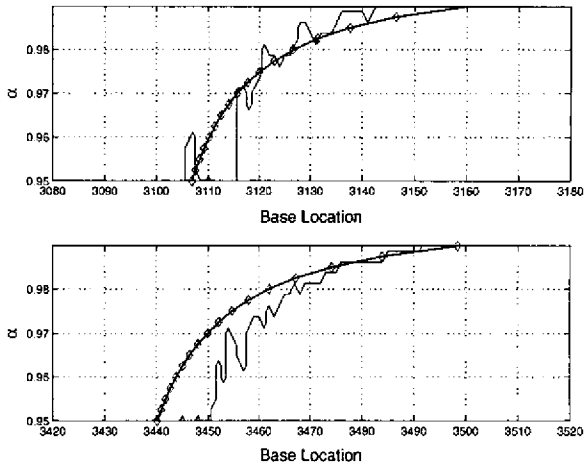

Fig. 8. Actual level curves for $S(n) = 0$ against the theoretical curve (thick line with diamonds) computed from the transition probabilities. (Top) Region changes from a non-CpG island region to a CpG island. (Bottom) Region changes from a CpG-island to a non-CpG island region.

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

[4] D. Anastassiou, "Genomic signal processing", IEEE Signal Processing Magazine, July 2001, pp. 8-20.

[5] P. P. Vaidyanathan and Byung-Jun Yoon, "The role of signal-processing concepts in genomics and proteomics", Journal of the Franklin Institute 341(2003) 111-135.

[6] F. Larsen, G. Gundersen, R. Lopez, H. Prydz, "CpG islands as gene markers in the human genome.", Genomics 13(4) (1992) 1095-1107.

[7] A. Bird, "DNA methylation patterns and epigenetic memory", Genes & Development 16 (1999) 6-21.

[8] R. Feil and S. Khosla, "Genomic imprinting in mammals: an interplay between chromatin and DNA methylation?", Trends in Genetics 15(11) (1999) 429-474.

[9] P. A. Jones and P. W. Laird, "Cancer-epigenetics comes of age", Nature Genetics 21 (1999) 163-167.

[10] M. Gardiner-Garden and M. Frommer, "CpG islands in vertebrate genomes", Journal of Molecular Biology 196 (1987) 261-282.

[11] D. Takai and P. A. Jones, "Comprehensive analysis of CpG islands in human chromosomes 21 and 22", PNAS 99(6) (2002) 3740-3745.

[12] E. C. Rouchka, R. Mazzarella, and D. J. States, "Computational Detection of CpG Islands in DNA", Technical Report, Washington University, Department of Computer Science, WUCS-97-39, 1997.

[13] N. Dasgupta, S. Lin and L. Carin, "Sequential modeling for identifying CpG island locations in human genome", IEEE Signal Processing Letters 9(12) (2002) 407-409.

[14] A. Bird, "CpG islands as gene markers in the vertebrate nucleus", Trends in Genetics 3 (1987) 342-347.