Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Chen et al., J Biomet Biostat 2012, 3:5
DOI: 10.4172/2155-6180.1000148

Research Article Open Access

# Locating CpG Islands with Kullback-Leibler Divergence

**Yung-Pin Chen*, Andrew Dittmore, Yasuhiro Goda, Alicia Laughton and Jessica Minnier**

*Department of Mathematical Sciences, Lewis & Clark College, Portland, USA*

**Abstract**

A CpG island is a short contiguous DNA subsequence that is rich in CG dinucleotides. CpG islands are often located around the promoters of housekeeping genes and have been found associated with certain tissue-specific genes. This observation indicates that they can be used as markers to identify genes. The information about the locations of CpG islands can also help us understand a gene regulation process called methylation. In this report, we propose a statistical method for locating CpG islands. Our method employs the Kullback-Leibler divergence. We use the given DNA sequence to determine a window size and a shift size for computing the divergence values along a DNA segment. A region in the proximity of a CpG island should contain consecutive windows with high divergence values. The distribution of the Kullback-Leibler divergence values can be suitably fitted by a truncated Pareto distribution. We estimate the parameters of the truncated Pareto distribution via the maximum likelihood principle. Then the fitted distribution is applied to locate regions with a divergence value exceeding a threshold level of significance. To assess the accuracy of our method, we compare our results to the putative CpG islands found in four well-studied mouse and human DNA sequences. The comparison suggests our approach consistently yields reliable predictions of CpG island locations.

**Keywords:** CpG island; Methylation; Kullback-Leibler divergence; Truncated Pareto distribution

## Introduction

### What are CpG islands and why do we study them?

A CpG island is a short genomic region that has a higher frequency of CpG dinucleotide than other regions.In vertebrate DNA, CpG dinucleotides are observed to occur much less frequently than expected by chance [1]. Also, typically, CpG dinucleotides are not uniformly distributed in mammalian DNA. For example,McClelland and Ivarie [2] examined 15 mammalian genes and reported that CpG dinucleotides are much richer in the 5'-anking sequences than in the 3'-anking regions. Tomso and Bell [3] demonstrated that CpG dinucleotides are substantially overrepresented at polymorphic sites within the human genome based on a comprehensive computational survey of 1.9 million human single nucleotide polymorphisms. Together, these studies indicate that CpG dinucleotides do not occur in a completely random manner.

### CpG dinucleotides and methylation

In the genomes of many higher plants and animals, there is a gene called DNA methyltransferase (Dnmt1). Dnmt1 is an enzyme which can attach a methyl (CH3) group onto the 5-carbon of cytosine. This methylation process only targets the cytosine joined with guanine by a phosphodiester bond on the same strand. Some people refer to this 5-methyl-C as the fifth base in DNA ([4], p.26). It is reported that approximately 70% of the CpG dinucleotides in eukaryotic chromosomal DNA are methylated ([1]; [5], p.147).

However, the locations of methylated cytosine do not seem to be random. A high proportion of methylated cytosines are found in inactive genes, whereas the CpG dinucleotides located around the promoters of housekeeping genes or around some tissue-specific genes are often not methylated [6,7]. This observation has at least two important implications. First, CpG islands are gene-associated and can be used as markers to identify genes [8-15]. For example, according to the results in [15], about 70% of the identi_ed CpG islands are associated with the human genes. Antequera and Bird [16] used some distinct features of CpG islands to estimate the number of genes in humans and mice by counting the CpG islands. In [17], Davuluri, Grosse and Zhang implemented discriminant functions to predict CpG-related and non-CpG-related first exons in the human genome. The second implication relates to the regulation of methylation (see [1], [18-20]). It has been understood since the 1960s that methylated cytosines are associated with transcriptionally inactive genes ([4], p.26). The effects of methylation include silencing tumor-suppressor genes [21,22], activating growth-stimulating genes [23], and human X-inactivation [24].

### Goals of our work

A commonly accepted rule for deciding CpG islands in vertebrate genomes was proposed in 1987 by Gardiner-Garden and Frommer [6]. They took a DNA stretch of more than 200 basepairs (bps) in length, and checked to see if the ratio of the observed number of CpG dinucleotides to the expected number of CpG dinucleotides exceeded 0.6 and if the G+C residue content was at least 50%. Takai and Jones, in their 2002 paper [25], did a comprehensive analysis of CpG islands in human chromosomes 21 and 22, and they found that a region longer than 500 bps with G+C residue content larger than 55% and the ratio \observed CpG/expected CpG" no less than 0.65 is more likely to be associated with the 5' regions of genes. We need to consider three issues for locating CpG islands.

The first issue is how to determine a proper length of a window for screening and a proper step size at which a window is shifted along the sequence while screening. These two quantities are called, respectively, the window and the shift. As Cuadrado, Sacristan, and Antequera [11] pointed out that CpG islands are quite variable in terms of sequence

***Corresponding author:** Yung-Pin Chen,Department of Mathematical Sciences, Lewis & Clark College, Portland, USA, E-mail: ychen@lclark.edu

**Received** May 24, 2012; **Accepted** June 21, 2012; **Published** June 23, 2012

**Citation:** Chen YP, Dittmore A, Goda Y, Laughton A, Minnier J (2012) Locating CpG Islands with Kullback-Leibler Divergence. J Biomet Biostat 3:148. doi:10.4172/2155-6180.1000148

lengths and positions relative to the transcription initiation sites, it seems unlikely to have a fixed window size or a fixed shift size that will work uniformly well in screening different DNA sequences for locating CpG islands. To tackle this problem, we propose an algorithm that uses a DNA sequence itself to determine both the window size and the shift size.

The second issue is how to measure the signal strength of a CpG island. The proportion of CpG dinucleotides and the G+C residue content are two basic and important measures. In a 1997 online manuscript by Rouchka, Mazzarella, and States [26], a logarithmic probability score is computed for each region, and the breakpoints of CpG islands are located based on the scores. Davuluri, Grosse, and Zhang [17] defined the maximum of the CpG percentages in different sliding windows of 201 bps to be the CpG score. Here, we approach this problem from the perspective of information theory. We treat the CpG proportion of the entire bulk DNA as the background noise distribution and the observed proportion of CpG dinucleotides over a chosen window as the signal distribution. We ask whether or not the DNA composition within a window under screening contains a signal strong enough to stand out from its bulk as a portion of a CpG island. The Kullback-Leibler divergence is employed to measure the strength of the signal distribution contrasting with the noise distribution. A larger divergence value indicates stronger evidence for the presence of a CpG island signal.

The third issue is how to establish a rule that can soundly decide whether or not the CpG island signal within a DNA stretch is statistically significant so that it is unlikely to observe such a signal by chance alone. Currently, both the thresholds for deciding CpG islands based on the ratio “observed CpG/expected CpG" and on the G+C residue content are only a rule of thumb. Many CpG island searchers allow users to specify their thresholds [27]. The quantification of the statistical significance of the observed evidence of a CpG island is lacking. A main goal of our work is to provide a sound statistical procedure to quantify this significance.

We will compare our results to the putative CpG islands found in four well-studied mouse and human DNA sequences. The statistical method we propose here can reliably find high divergent subregions that are probable to be within a CpG island or in the proximity of a CpG island. We want to remark that a statistical method cannot and is not intended to find the exact site of a CpG island with the exact size. The biological interpretation of the presence of a CpG island and how a CpG island is functionally related to other regions should be examined and determined by geneticists.

### Outline

This report is organized as follows. In Section 2, we explain our statistical reasoning and method for locating CpG islands. In Section 3, we choose four well-studied DNA sequences, one from mouse and three from human, to demonstrate how our method works. At the same time, we assess the accuracy of our method by comparing our results to the putative CpG islands reported in those sequences. Some basic compositional statistics of the four sequences are given in Figure 1. All the DNA sequences are achieved from the GenBank website http://www.ncbi.nlm.nih.gov of the National Center for Biotechnology Information (NCBI) dated before January 2006. More detailed information about each sequence can be found at the NCBI website. We will explain and compare our results with the current findings on CpG islands for each of the four sequences. We conclude this report with a brief discussion on a more general problem-sequence segmentation.

## Materials and Methods

### How can CpG islands stand out statistically?

CpG islands are short DNA stretches that are generally rich in CpG dinucleotides, and the G+C content in each stretch is often relatively high. These numerical features provide a basis for quantitative methods of locating CpG islands. Di_erent speci_cations can yield di_erent results in locating CpG islands. A fundamental statistical question is to ask whether a DNA stretch stands out significantly enough to be distinguished from its bulk as a CpG island.

In this section, we present our reasoning and method for locating CpG islands in three parts. First, we explain how we use a DNA sequence itself to determine the window size and the shift size. Second, we introduce the Kullback-Leibler divergence as a statistical measure of the strength that a CpG island signal departs from the background bulk DNA. We build a profile of the Kullback-Leibler divergence values as windows are shifted along the DNA sequence. A region of consecutive windows with high divergence values should be in the proximity of a CpG island. Third, we show that the divergence values can be well-fitted by a truncated Pareto distribution. We estimate the parameters associated with the truncated Pareto density function by the maximum likelihood principle. The fitted truncated Pareto distribution then is applied to locate regions with a divergence value exceeding the 95th percentile.

### Sequence-defined window and shift

We can consider a DNA sequence as a realization of a stochastic renewal process where an occurrence of a CpG dinucleotide is regarded as a renewal. The segment of base pairs between two CpG dinucleotides, excluding the CpG dinucleotide preceding it but including the CpG dinucleotide tailing it, is called a CpG interarrival. Let us take sixty base pairs (from 961 to 1020) of the M63419 sequence, the mouse leukemia

| accession | organism | base pairs | % of A | % of T | % of C | % of G | % of CpG |
|---|---|---|---|---|---|---|---|
| M63419 | Mus musculus | 8,735 | 21.820 | 23.721 | 27.018 | 27.441 | 2.164 |
| AL022327 | Homo sapiens | 101,270 | 23.848 | 20.712 | 27.692 | 27.748 | 2.958 |
| AL031723 | Homo sapiens | 41,255 | 22.717 | 19.433 | 29.509 | 28.341 | 3.100 |
| AL049762 | Homo sapiens | 100,575 | 30.783 | 27.346 | 21.213 | 20.658 | 0.996 |

**Figure 1:** Compositional statistics of four DNA sequences.

inhibitory factor gene, listed below to demonstrate the definition.

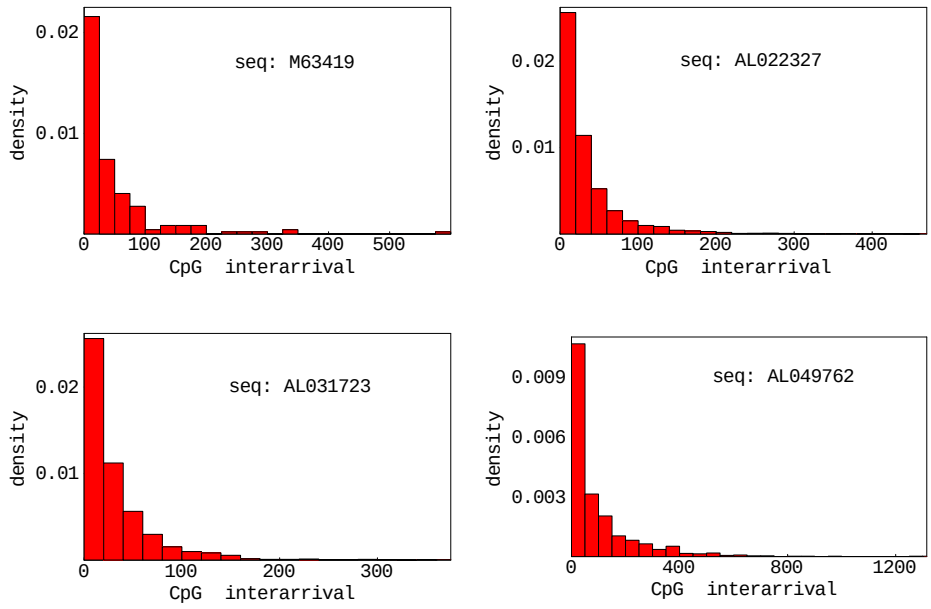
aatgaaggtc ttggcccg$_{977}$cag gtaaatccat gcg$_{993}$ccg$_{996}$ggcc

g$_{1001}$cg$_{1003}$atttaag agtcccg$_{1017}$gct

Suppose we begin with the 961st base pair, which is an adenine. The first CpG dinucleotide ends at the 977th base position, so we say the length of the first CpG interarrival is 977-960 = 17. Now the process starts anew at the 978th base pair, which is a cytosine. The second CpG dinucleotide ends at the 993rd base position, so the length of the second CpG interarrival is 993-977 = 16. Likewise, the third CpG interarrival has length 996-993 = 3, the fourth CpG interarrival has length 1001-996 = 5, the fifth CpG interarrival has length 1003-1001 = 2, and the sixth CpG interarrival has length 1017-1003 = 14. Each CpG interarrival includes the CpG dinucleotide at its end, and all CpG interarrivals are non-overlapping. The M63419 sequence has totally 189 CpG interarrivals, with maximum 578 bps, mean 46:196 bps, and standard deviation 70:043 bps.

Figure 2 displays the histograms and statistics of the CpG interarrivals of the four DNA sequences discussed in this report. It is interesting to see that those histograms indicate distributions with a negative power or an exponential density curve. We expect that a region in the proximity of a CpG island is denser in CpG dinucleotides than other regions, and hence it is more likely to observe smaller CpG interarrivals in such a region. How do we choose a proper window size for examining this? How many base positions should we shift from a window to the next window for screening along the DNA sequence? On one hand, a window should be wide enough to include a signi_ cant portion of a CpG island. On the other hand, a window should not be too wide, so that the CpG island signal within the window can be statistically detected from the bulk DNA. The shift size for sliding a window along the DNA sequence determines the number of times of screening. As each DNA sequence can have CpG islands vary in terms of lengths, it seems reasonable to adopt an algorithm that uses the sequence itself to determine the window size and the shift size.

Consider a DNA sequence consisting of L base pairs. Let I1, I2,.., and Iv be the lengths of the CpG interarrivals of the sequence, where v denotes the number of CpG dinucleotides in the sequence. Note that I1 + I2 + …+ Iv ≤ L. If v = 0, then we would report that no CpG islands are found. If v = 1, we set the window size to be the maximum length of the CpG interarrivals. This ensures that at least one CpG dinucleotide will be observed within each window for all windows preceding the last CpG dinucleotide. It also allows a moderately large window span for including a significant portion of a CpG island, if the window under screening lies in the vicinity of a CpG island. At the same time, as the data suggest, it is not unreasonably large to let the bulk base composition distribution outweigh the CpG island signal inside each window. Another reason for setting the window size to be the maximum CpG interarrival length is due to a technical point for computing the Kullback-Leibler divergence, which will be explained later. For the shift size, we choose a value such that the number of screenings is roughly equal to the number of CpG dinucleotides. So the shift size is taken to be the rounded value of the mean length of the




| accession | base pairs | # CpG | maximum | mean | SD |
|---|---|---|---|---|---|
| M63419 | 8,735 | 189 | 578 | 46.196 | 70.043 |
| AL022327 | 101,270 | 2,996 | 465 | 33.792 | 44.512 |
| AL031723 | 41,255 | 1,279 | 368 | 32.222 | 39.691 |
| AL049762 | 100,575 | 1,002 | 1,309 | 100.363 | 140.883 |

**Figure 2:** Histograms and statistics of the CpG interarrivals of four DNA sequences.

CpG interarrivals. Now let w and h denote the window and shift sizes, respectively. That is, we have set

$$w = \text{window size} = \max(I_1, I_2, ..., I_v) \text{ and} \quad (2.1)$$

$$h = \text{shift size} = \text{round}[\frac{1}{v}(I_1 + I_2 + ... + I_v)]. \quad (2.2)$$

For the M63419 sequences, it has v = 189 CpG dinucleotides out of a total of L = 8735 base pairs. The window size is w = 578 bps, and the shift size is h = 46 bps. For the four DNA sequences listed in Figure 2, the window parameter ranges from 368 to 1309 bps, and the shift parameter runs from 32 to 100 bps. Note that the two quantities defined in (2.1) and (2.2) appear to be positively correlated. Kullback-Leibler divergence. The Kullback-Leibler divergence is used in information theory to measure the statistical distance between two distributions. Let $f_0$ and $f_1$ be two discrete probability mass functions. The quantity $f_0(x)$ is the likelihood of observing a random outcome x from a background noise distribution $f_0$. The quantity $f_1(x)$ is the likelihood of observing the same random outcome x from a signal distribution $f_1$ that needs to be detected from the noise. Kullback [28] interpreted the logarithmic ratio log $[f_1(x)=f_0(x)]$ as the information in the observation x for discriminating $f_1(x)$ against $f_0(x)$. The Kullback-Leibler divergence of the distribution $f_1$ against the distribution $f_0$ is defined to be

$$div_{\text{K-L}}(f_1, f_0) = \sum_x f_1(x) \log \frac{f_1(x)}{f_0(x)}. \quad (2.3)$$

The Kullback-Leibler divergence is nonnegative for any distributions $f_0$ and $f_1$. A larger divergence indicates the two distributions are more distinct from each other. For example, if the noise has the probability masses on dichotomous outcomes \head" and \tail" with $f_0$(head) = 0:4 and $f_0$(tail) = 0:6, then the Kullback-Leibler divergence for discriminating the distribution with $f_1$(head) = p and $f_1$(tail) = 1 - p against $f_0$ is

$$div_{\text{K-L}}(f_1, f_0) = p \log \frac{p}{0.4} + (1-p) \log \frac{1-p}{0.6}.$$

This Kullback-Leibler divergence is a function of p, and it is plotted in Figure 3. Note that the minimum divergence value zero is attained at p = 0:4. We will simply call a value obtained from (2.3) a divergence instead of Kullback-Leibler divergence throughout this report, and symbolize it by div. The main idea of employing the divergence to differentiate a CpG island from its bulk DNA is to use the entire CpG proportion of bulk DNA as the background noise distribution $f_0$ and the observed CpG proportion over a specific window as the signal distribution $f_1$. Let $f_0$(CpG) be the CpG proportion of the bulk DNA, and let $f_1$;W (CpG) be the observed CpG proportion over window W, then the divergence for this particular window W is

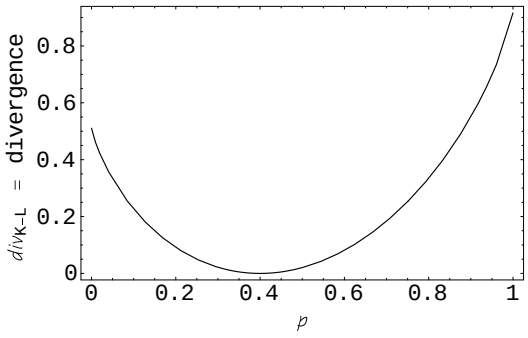



**Figure 3:** K-L divergence function of *p*.

$$div(f_1, \text{w}, f_0) = f_1\text{w(CpG)} \log \frac{f_1\text{w(CpG)}}{f_0\text{(CpG)}} + (1 - f_1, \text{w(CpG)}) \log \frac{1 - f_1\text{w(CpG)}}{1 - f_0\text{(CpG)}}. \quad (2.4)$$

The divergence of a window that is richer in CpG dinucleotides than its bulk DNA tends to be larger, and it indicates that the window may contain, or be contained by, or substantially overlap a CpG island. The window (whose size is chosen to be the maximum CpG interarrival length) is moved along the sequence by a particular shift size (set to be the rounded value of the mean CpG interarrival length) until there is no space to move. The divergence is evaluated each time when the window is shifted. This establishes a list of divergences for the DNA sequence under screening. For example, the M63419 sequence has 189 CpG dinucleotides out of 8735 base pairs, so the CpG proportion of the entire sequence is

$$f_0\text{(CpG)} = \frac{189}{8735 - 1} \approx 0.0216.$$

The reason why we subtract 1 from 8735 is that 8735 bps can form 8734 dinucleotides. The window size we choose for the M63419 sequence is 578 bps. Within the first window of 578 bps, there are 6 CpG dinucleotides, so the observed CpG proportion within the first window is

$$f_1, \text{w(CpG)} = \frac{6}{578 - 1} \approx 0.0104.$$

It follows from (2.4) that the divergence for the first window is

$$0.0104 \times \log(\frac{0.0104}{0.0216}) + (1 - 0.0104) \times \log(\frac{1 - 0.0104}{1 - 0.0216}) \approx 0.0037 \quad (2.5)$$

A technical reason for choosing the maximum CpG interarrival to be the window size is that the observed CpG proportion within each window is assured to be positive so the logarithmic values are well-defined in the divergence formula (2.4). (We can also define $\lim_{x \to 0+} x \times \log x = 0$.) Now the shift size is set to be 46 bps, the rounded mean CpG interarrival length, for the M63419 sequence. We move the first window down 46 bps and compute the divergence of this shifted window using (2.4). We repeat this procedure as we move. How many shifts do we make before we stop? If we let k be the number of shifts of size h bases, then it should satisfy the inequality (w + kh) ≤ L, where L is the length of the DNA sequence. It implies that the number of shifts k is the largest integer not exceeding (L -w)=h. Recall that L = 8735, w = 578, and h = 46 for the M63419 sequence, so the number of shifts is k = 177. Because there is a starting window before shifting, there are totally k + 1 = 177 + 1 = 178 evaluations of divergence, which is close to 189, the number of CpG dinucleotides in the sequence.

## Truncated Pareto distribution

Is the divergence 0:0037 shown in (2.5) significantly large enough to infer that the first window of 578 bps of the M63419 sequence is within or overlapping a CpG island? If we slide the window down for 20 shifts of size 46 bps each, the divergence of this particular window is 0.0732. Is this divergence significantly large? We need to examine the distribution of the divergences to assess the statistical significance. The histogram and basic statistics of the 178 values of divergence for the M63419 sequence are shown in Figure 4(a). The data suggest that the distribution of the divergences can be suitably described by a truncated Pareto distribution with a negative power parameter. A truncated Pareto distribution has the density function

$$f(x)=\begin{cases}\dfrac{r+1}{\beta^{r+1}-\alpha^{r+1}}x^{r}, & 0<\alpha<x<\beta,\ \ if\ r\neq -1;\\ \dfrac{1}{\ln\beta-\ln\alpha}x^{r}, & 0<\alpha<x<\beta,\ \ if\ r=-1.\end{cases} \tag{2.6}$$

where α is a positive parameter for the lower bound of the data range, β is a parameter for the upper bound, and r is a power parameter. (Note that when r < -1, the denominator $\beta^{r+1}$ - $\beta^{r+1}$ becomes negative and the density is still positive.) Let $X_1$, $X_2$,..., $X_n$ be a random sample from the truncated Pareto distribution in (2.6), and let $\alpha, \beta and\ r$ be the maximum likelihood estimators (MLEs) of the parameters $\alpha, \beta and\ r$ respectively. Then

$$\hat{\alpha}=\min(X_1,X_2,...X_n),\ \ \hat{\beta}=\max(X_1,X_2,...X_n), \tag{2.7}$$

and $\hat{r}$ is the unique solution to the equation (with r as the unknown)

$$\frac{1}{r+1}-\frac{\beta^{r+1}\log\hat{\beta}-\hat{\alpha}^{r+1}\log\hat{\alpha}}{\hat{\beta}^{r+1}-\hat{\alpha}^{r+1}}=-\frac{1}{n}\sum_{i=1}^{n}\log X_i. \tag{2.8}$$

A reference on the MLEs related to a truncated Pareto distribution can be found in [29]. We like to remark that an MLE may be biased for the true parameter value. For the M63419 sequence, the values of the MLEs of the three parameters are

$$\hat{\alpha}_{\mathrm{M63419}}=1.70\times10^{-5},\hat{\beta}_{\mathrm{M63419}}=0.0788,\ \text{and}\ \hat{r}_{\mathrm{M63419}}=-0.9553, \tag{2.9}$$

and the fitted truncated Pareto density is

$$\hat{f}_{\mathrm{M63419}}(x)=0.1593x^{-0.9553},1.70\times10^{-5}\le x\le 0.0788. \tag{2.10}$$

We plot the fitted truncated Pareto density curve, together with the histogram of the divergences, in Figure 4(b). Once we obtain a fitted truncated Pareto distribution of the divergences, we can use it to assess the statistical significance of a divergence value because a larger divergence indicates stronger evidence in favor of the presence of a CpG island. That is, we can locate CpG islands by identifying those high divergence regions with a predetermined threshold level of likelihood. Let 0 < p < 1 be a probability. We can ask if a given divergence is in the top (100 × p)% of the data. Now let X be a random variable that has the truncated Pareto distribution with the density function given in (2.6). If xp is the cutoff for the divergences in the top (100 × p)% of the data, then we have

$$\Pr\{X>x_p\}=\int_{x_p}^{\beta}\frac{r+1}{\beta^{r+1}-\alpha^{r+1}}x^{r}dx=p.$$

This equation leads to

$$x_p=[(1-p)\beta^{r+1}+p\alpha^{r+1}]^{1/(r+1)}. \tag{2.11}$$

We can replace the parameters α,β and r in equation (2.11) by the MLEs $\hat{\alpha}$, $\hat{\beta}$ $\hat{r}$ and given in (2.7) and(2.8), respectively, to get an estimated top (100 × p)th percentile $\hat{x}_p$. In our work, we particularly choose the significance level to be p = 0:05. That is, we set the 95th percentile to be the threshold level. So we find those regions with a divergence at least as large as the top fiifth percentile ^x0:05, and those regions are reported to be in the proximity of a CpG island. For the M63419 sequence, substituting the MLEs in (2.9) to (2.11), the resulting estimated top fifth percentile of the divergence distribution is

$$\hat{x}_{0.05,\mathrm{M63419}}=[0.95\times0.0788^{(-0.9553+1)}+0.05\times(1.70\times10^{-5})^{(-0.9553+1)}]^{1/(-0.9553+1)}=0.0553. \tag{2.12}$$

For the first window of 578 bps of the M63419 sequence, the divergence 0.0037 shown in (2.5) is lower than the top fifth percentile given in (2.12), so it is not statistically significant enough to be considered to lie within the proximity of a CpG island. However, after we make 20 shifts of size 46 bps, we will have a divergence value 0.0732, which exceeds the top fifth percentile, and therefore we can consider this particular region of 578 bps to be in the proximity of a CpG island with 5% level of significance.

## Results: Do CpG Islands Stand Out Statistically?

We will present our results on CpG islands for each of the four DNA sequences listed in Figure 1. For each sequence, we start with a graph that shows the divergences against the window index, and we will call it a divergence plot. On each of the divergence plots we draw a horizontal line to indicate the top fifth percentile of the divergences. A divergence above the cutoff line indicates evidence in favor of being in the proximity of a CpG island at 5% level of significance. To demonstrate visually how well the distribution of the divergences can be fitted by a truncated Pareto distribution, we display the histogram of the divergences and the fitted truncated Pareto density curve like the one shown in Figure 4. Within the displaying frame of each histogram, we show the basic statistics that include the sequence accession code, and the number, minimum, maximum, mean, and standard deviation of the divergences. We also present the density function of each fitted truncated Pareto distribution whose parameters are replaced by the maximum likelihood estimates. We report and interpret the high divergence regions we locate, and compare our predictions with the current statistics and results on CpG islands for those DNA sequences.

Before presenting the results on each sequence, we first explain how

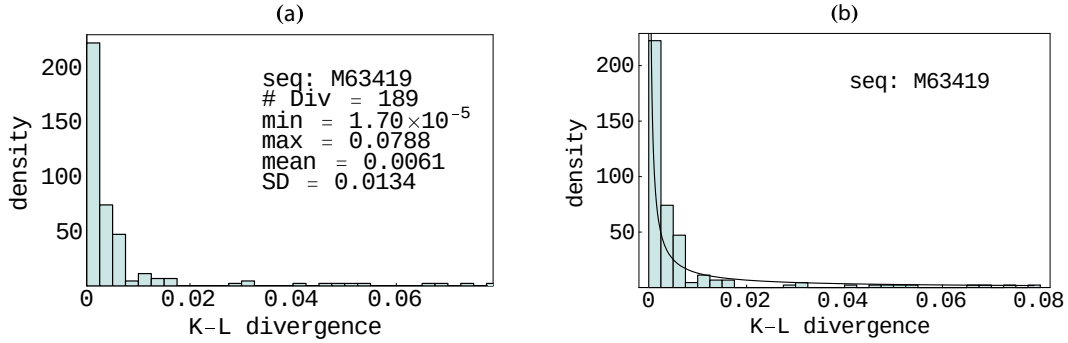



**Figure 4:** (a) Statistics and Histo**gram of the divergences of M63419; (b) Pareto fit of the divergences**

we convert from “window index" to “base position." We have adopted the symbols *w* and *h* for the window size and shift size in (2.1) and (2.2), respectively. Suppose we _nd a region with divergences consistently larger than the threshold determined bythe 5% level of signi_cance, and the region consists of consecutive windows from index i to index j. Then the beginning window i starts at base position (i - 1)×h, and the ending window j stops at base position w + (j - 1) × h. Therefore we have the following conversion formula:

windows of size *w* from *i* to *j* with shift size *h*

$$\Rightarrow \text{base positions from } (i-1)\times h \text{ to } w+(j-1)\times h \tag{3.1}$$

## Results on M63419

We have shown some results for the M63419 sequence, the mouse leukemia inhibitory factor gene, when we explained our method in Section 2. Here we give a recap. The M63419 sequence has 189 CpG dinucleotides out of 8735 bps. The maximum CpG interarrival is 578 bps (so the window size w = 578), and the mean CpG interarrival is 46.20 bps (so the shift size h = 46). There are totally 178 evaluations of divergence. We show the divergence plot in Figure 5(a), and all the divergences are below 0.08 with mean 0.0061 and standard deviation 0.0134.

The MLEs of the parameters of the truncated Pareto distribution, the fitted density function, and the estimated top fifth percentile of the divergence distribution are respectively given in (2.9), (2.10), and (2.12).There is only one region with divergences exceeding the top fifth percentile $\hat{x}_{0.05,\mathrm{AL}0222327} = 0.0853$. This region consists of windows indexed from 20 to 23. According to the conversion formula in (3.1) with window size w = 578 and shift size h = 46, the region is converted into the following base positions.

M63419 high divergence region: base positions from 874 to 1590

This sequence was analyzed in [12], which proposed that the span of nucleotides 1313-1458 is a part of a CpG island. The region we locate is wider, and it has 439 bps more in the upstream and 132 bps more in the downstream.

## Results on AL022327

According to the information provided by GenBank, AL022327 is the human DNA sequence from clone RP3-355C18 on chromosome 22q13.3. The sequence consists of 101270 bps of which there are 2996 CpG dinucleotides. The maximum CpG interarrival is 465 bps and the mean CpG interarrival is 33.79 bps. There are totally 2965 evaluations of divergence, and the divergence plot is given in Figure 6(a).

The maximum divergence does not exceed 0.1286, and the mean and the standard deviation of the divergences are 0.0055 and 0.0125, respectively. The histogram of the divergences shown in Figure 7(a) is highly skewed to the right. The MLEs of the parameters of the truncated Pareto distribution and the estimated top fifth percentile are

$$\hat{\alpha}_{\mathrm{AL}022327} = 5.98\times10^{-56}, \hat{\beta}_{\mathrm{AL}022327} = 0.1286, \hat{r}_{\mathrm{AL}022327} = -0.9578, \text{and } \hat{x}_{0.05,\mathrm{AL}022327} = 0.0853. \tag{3.2}$$

The fitted truncated Pareto density is

$$\hat{f}_{\mathrm{AL}022327}(x) = 0.1340x^{-0.9578}, 5.98\times10^{-6} \le x \le 0.1286,$$

which is displayed with the histogram in Figure 7(b).

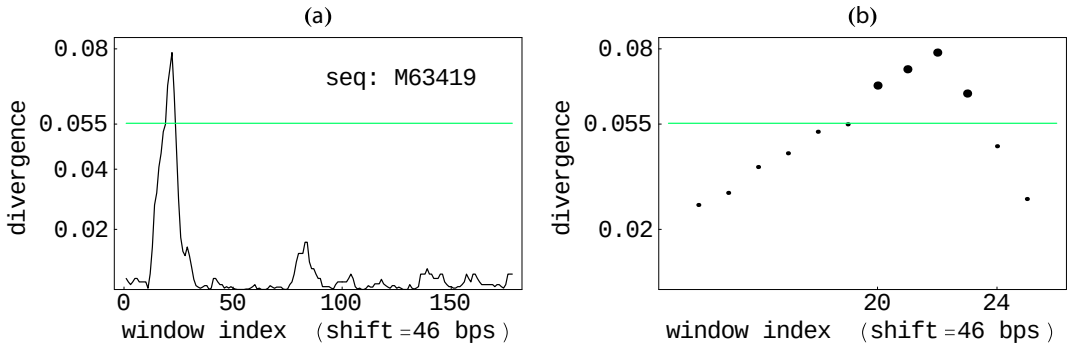



**Figure 5:** Divergence plot of M63419.

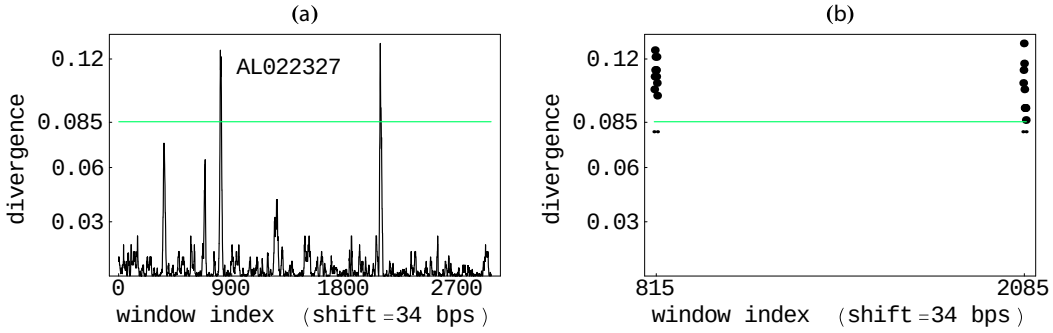



**Figure 6:** Divergence plot of AL022327.

We find two regions with divergences exceeding the top fifth percentile $\hat{x}_{0.05,AL049762} = 0.0895$. The first region contains windows with indices from 808 to 818, which is converted to base positions from 27438 to 28243. The second region has window indices from 2081 to 2090, which is converted to base positions from 70720 to 71491. We summarize the results below.

1st AL022327 high divergence region: base positions from 27438 to 28243

2nd AL022327 high divergence region: base positions from 70720 to 71491

GenBank reports seven putative CpG islands. However, only six are listed with non-experimental evidence, and we tabulate them below.

| CpG island 1 | CpG island 2 | CpG island 3 | CpG island 4 | CpG island 5 | CpG island 6 |
|---|---|---|---|---|---|
| 11973 -13056 | 22848-23801 | 27337-28417 | 42172-43633 | 44147-44700 | 70722-71673 |

Note that our first predicted CpG island is a subregion of the third putative CpG island. They differ by 101 bps in the upstream and 174 bps in the downstream. Our second predicted CpG island overlaps the sixth putative CpG island, with two base pairs difference in the upstream and 182 bps difference in the downstream. Nevertheless, the divergence plot in Figure 6 shows five prominent spikes of unequally high divergences. We have set the significance level at 5%, which may be too stringent to detect all the CpGislands reported in GenBank.

## Results on AL031723

The AL031723 sequence, reported in GenBank, is the human DNA sequence from clone LA16c-439A6 on chromosome 16. It has 41255 bps of which there are 1279 CpG dinucleotides. The maximum CpG interarrival is 368 bps and the mean CpG interarrival is 32.22 bps. There are totally 1278 evaluations of divergence. The divergence plot is given in Figure 8(a). The divergences are no more than 0.149 with mean 0.0064 and standard deviation 0.0171. The histogram of the divergences is shown in Figure 9(a). The MLEs of the parameters of the truncated Pareto distribution and the estimated top fifth percentile are

$$\hat{\alpha}_{AL031723} = 1.79\times10^{-5}\ , \hat{\beta}_{AL031723} = 0.1488,\ \hat{r}_{AL031723} = -1.0274, \text{and}\ \ \hat{x}_{0.05,AL031723} = 0.0895 \quad (3.4)$$

The fitted truncated Pareto density is

$$\hat{f}_{AL031723}(x) = 0.0927x^{-1.0274}, 1.79\times10^{-5} \le x \le 0.1488, \quad (3.5)$$

which is plotted along with the histogram in Figure 9(b).

For the AL031723 sequence, there are two regions found to have divergences exceeding the top fifth percentile $\hat{x}_{0.05,AL031723} = 0.0895$. The first region goes from window 789 to window 806, and the second region covers windows indexed from 1159 to 1161. They are converted to the following based positions.

1st AL031723 high divergence region: base positions from 25216 to 26128

2nd AL031723 high divergence region: base positions from 37056 to 37488

GenBank reports that there are three non-experimental CpG islands. The first has base positions from 18929 to 19548, the second goes from 25202 to 26372, and the third goes from 36891 to 37694.

Although the divergence plot in Figure 8 indicates some evidence

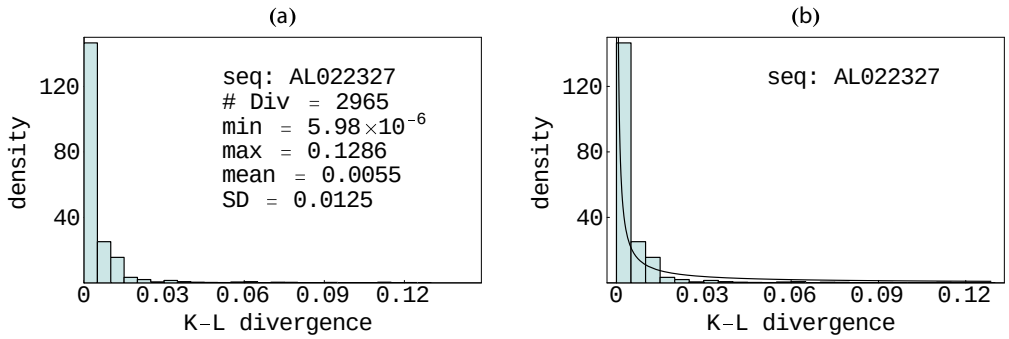



**Figure 7:** (a) Statistics and histogram of the divergences of AL022327; (b) Pareto fit of the divergences

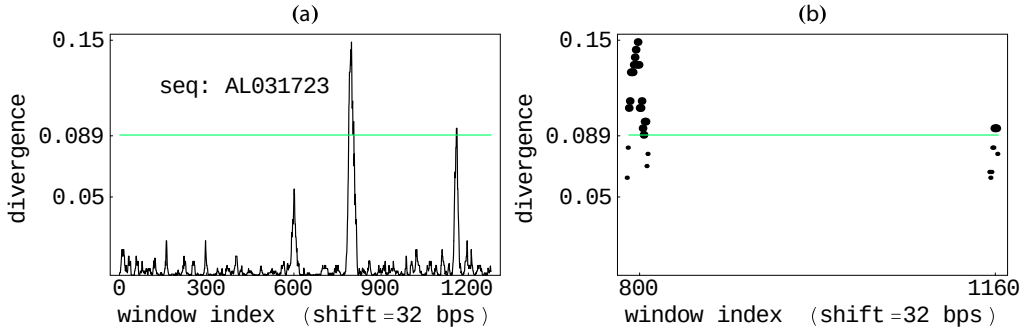



**Figure 8:** Divergence plot of AL031723.

for a CpG island within the proximity of the first non-experimental CpG island reported by GenBank, it is not significant at the level of 5% as we specify. Each of the other two reported putative CpG islands contains the corresponding high divergence region we report as a subset. The second putative CpG island has 14 bps longer in the upstream and 244 bps shorter in the downstream than ours, and the third putative CpG island has 265 bps longer in the upstream and 206bps shorter in the downstream than ours.

### Results on AL049762

According to GenBank, AL049762 is the human DNA sequence from clone RP1- 81F6 on chromosome 1q24.1-25.2. The sequence consists of 100575 bps, and there are 1002 CpG dinucleotides in it. The maximum CpG interarrival is 1309 bps and the mean CpG interarrival is 100.36 bps. There are totally 993 evaluations of divergence, and the divergence plot is given in Figure 10(a). The divergences are in a range of 0.046 with mean 0.0018 and standard deviation 0.0045. The histogram of the divergences is shown Figure 11(a). The MLEs of the parameters of the truncated Pareto distribution and the estimated top fifth percentile are

$$\hat{\alpha}_{\text{AL049762}} = 2.92\times10^{-8}, \hat{\beta}_{\text{AL049762}} = 0.0461, \hat{r}_{\text{AL049762}} = -0.8609, \text{and } \hat{x}_{0.05,\text{AL049762}} = 0.0336 \quad (3.6)$$

The fitted truncated Pareto density is

$$\hat{f}_{\text{AL049762}}(x) = 0.2475x^{-0.8609}, 2.92\times10^{-8} \le x \le 0.0461, \quad (3.7)$$

which is displayed with the histogram in Figure 11(b).

There is only one region with divergences exceeding the top fifth percentile $\hat{x}_{0.05,\text{AL049762}} = 0{:}0336$. This region goes from window 956 to window 963, and it is converted to the base positions as below.

AL049762 high divergence region: base positions from 95500 to 97509

There is only one putative CpG island, base positions from 96101 to 96822, reported in the *GenBank* data for the sequence. It is a subregion of the high divergence region we locate. Our region stretches 601 bps more in the upstream and 687 bps more in the downstream.

## Concluding Remarks

CpG islands can be used as markers to identify genes and help gain information about the methylation process. We employ the Kullback-Leibler divergence as a statistical measure for discriminating CpG islands in a DNA sequence from its bulk. We also develop an algorithm that uses a DNA sequence itself to determine the window size and the shift size for computing the Kullback-Leibler divergence values. In addition, we propose truncated Pareto distributions to quantify how statistically confident we are in locating CpG islands.

The use of truncated Pareto distribution is strongly suggested by the histograms of the Kullback-Leibler divergence values. We have empirically discovered that all histograms appear to be well-fitted by a negative power density curve as long as the window size and the shift size are properly chosen. Our empirical exploration sheds light on fitting the distribution of divergence values. A direction of our future work is to investigate alternative statistical models for describing the distribution of the Kullback-Leibler divergence values.

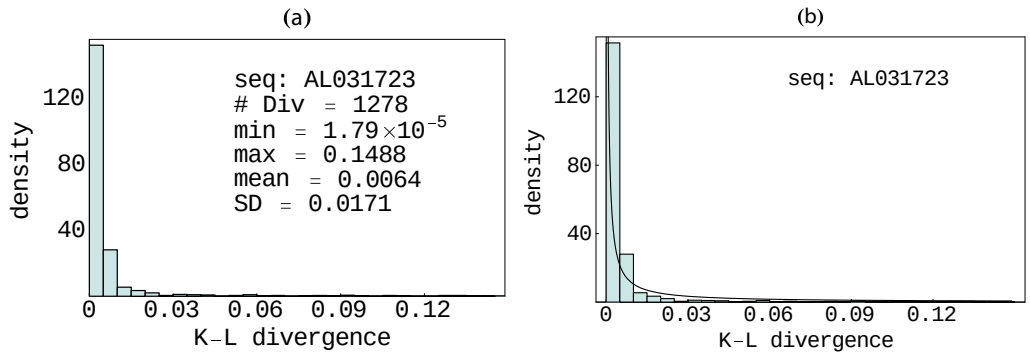



**Figure 9:** (a) Statistics and histogram of the divergences of AL031723; (b) Pareto fit of the divergences.

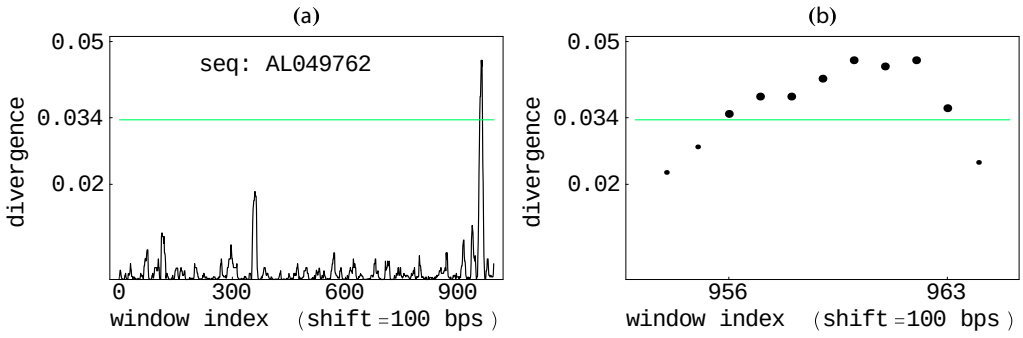



**Figure 10:** Divergence plot of AL049762.

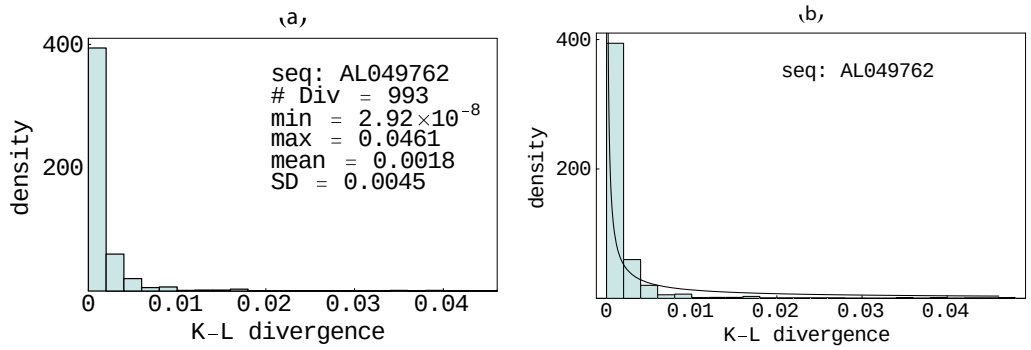



**Figure 11:** (a) Statistics and histogram of the divergences of AL049762; (b) Pareto fit of the divergences.

As we explained in the introduction, the method we propose is not aimed at locating the exact site of a CpG island with the exact size, but instead finding the regions of high divergences with their statistical significances quantified. The regions we locate with the 5% threshold level of significance can overlap, contain, or be contained in those putative CpG islands reported in the *GenBank*. Overall, our results show consistently reliable predictions of the CpG island locations.

Locating CpG islands can be regarded as a special case of a general problem of sequence segmentation that tries to divide a sequence in a meaningful way. For example, many bioinformatics researchers have been working on statistical methods and models for segmenting a DNA sequence into regions of different biological functions. Finding CpG islands seems to be more straightforward than many other sequence segmentation problems because it is simply based on the CpG composition. Many segmentation methods and algorithms are mathematically intriguing and can be computationally expensive. In particular, it is difficult to either quantify the uncertainty involved or give meaningful biological interpretations. In this report, we have manifested an approach for locating CpG islands with statistical significance.

## Acknowledgments

We would like to thank the John Rogers Science Research Program of Lewis & Clark College for funding our research. Specially we want to thank Drs. Deborah Lycan, Greta Binford, and Greg Hermann in the Biology Department for answering many of our questions.

## References

1. Bird A (1986) CpG-rich islands and the function of DNA methylation. Nature 321: 209-213.
2. McClelland M, Ivarie R (1982) Asymmetrical distribution of CpG in an `average' mammalian gene. Nucleic Acids Research 10: 7865-7877.
3. Tomso DJ, Bell DA (2003) Sequence context at human single nucleotide polymorphisms: overrepresentation of CpG dinucleotide at polymorphic sites and suppression of variation in CpG islands. J Mol Biol 327: 303-308.
4. Campbell AM, Heyer L (2007) Discovering Genomics, Proteomics, & Bioinformatics. Benjamin Cummings.
5. Watson JD, Gilman M, Witkowski J, Zoller, M (1992) Recombinant DNA. (2nd edn), W.H. Freeman and Company.
6. Gardiner-Garden M, Frommer M (1987) CpG islands in vertebrate genomes. J Mol Biol 196: 261-282.
7. Yamashita R, Suzuki Y, Takagi, T, Sugano S, Nakai K (2003) Genome wide analysis reveals strong correlation between CpG islands and tissue-specificity. Genome Informatics 14: 404-405.
8. Alcalay M, Toniolo D (1988) CpG islands of the X chromosome are gene associated. Nucleic Acids Res 16: 9527-9543.
9. Ashikawa I (2001) Gene-associated CpG islands in plants as revealed by analyses of genomic sequences. Plant J 26: 617-625.
10. Cross SH, Bird AP (1995) CpG islands and genes. Curr Opin Genet Dev 5: 309-314.
11. Cuadrado M, Sacristán M, Antequera F (2001) Species-specific organization of CpG island promoters at mammalian homologous genes. EMBO Rep 2: 586-592.
12. Kaspar P, Dvorák M, Bartůněk P (1993) Identification of CpG island at the 5' end of murine leukemia inhibitory factor gene. FEBS Lett 319: 159-162.
13. Larson F, Gundersen G, Lopez R, Prydz H (1992) CpG islands as gene markers in the human genome. Genomics 13: 1095-1107.
14. Strichman-Almashanu LZ, Lee RS, Onyango PO, Perlman E, Flam F, et al. (2002) A genome-wide screen for normally methylated human CpG islands that can identify novel imprinted genes. Genome Res 12: 543-554.
15. Wang Y, Leung, FC (2004) An evaluation of new criteria for CpG islands in the human genome as gene markers. Bioinformatics 20: 1170-1177.
16. Antequera F, Bird A (1993) Number of CpG islands and genes in human and mouse. Prod Natl Acad Sci USA 90: 11995-11999.
17. Davuluri RV, Grosse I, Zhang MQ (2001) Computational identification of promoters and first exons in the human genomes. Nature Genetics 29: 412-417.
18. Bird A (1999) DNA methylation de Novo. Science 286: 2287-2288.
19. Bird A (2002) DNA methylation patterns and epigenetic memory. Genes Dev 16: 6-21.
20. Ohlsson R, Kanduri C (2002) New twists on the epigenetics of CpG islands. Genome Res 12: 525-526.
21. Pieper RO, Patel S, Ting SA, Futscher BW, Costello JF (1996) Methylation of CpG island transcription factor binding sites is unnecessary for aberrant silencing of the human MGMT gene. J Biol Chem 271: 13916-13924.
22. Robert MF, Morin S, Beaulieu N, Gauthier F, Chute IC, et al. (2003) DNMT1 is required to maintain CpG methylation and aberrant gene silencing in human cancer cells. Nat Genet 33: 61-65.
23. Yoon BJ, Herman H, Sikora A, Smith LT, Plass C, et al. (2002) Regulation of DNA methylation of Rasgrf1. Nat Genet 30: 92-96.
24. Ke X, Collins A (2003) CpG islands in human X-inactivation. Ann Hum Genet 67: 242-249.
25. Takai D, Jones PA (2002) Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proc Natl Acad Sci USA 99: 3740-3745.
26. Rouchka EC, Mazzarella R, States DJ (1997) Computational detection of CpG islands in DNA.
27. Takai D, Jones PA (2003) The CpG Island Searcher: A new WWW resource. In Silico Biol 3: 235-240.
28. Kullback S (1959) Information Theory and Statistics. John Wiley and Sons Inc, New York.
29. Aban I, Meerschaert MM, Panorska AK (2006) Parameter estimation for the truncated Pareto distribution. J Am Stat Assoc 101: 270-277.