



Statistics Research Seminar
Portland State University
April 30, 2009

**Locating CpG Islands with
Statistical Significance**

A. Dittmore, Y. Goda, A. Laughton, J. Minnier
Lewis & Clark College

file prepared by Yung-Pin Chen on
April 29, 2009

Outline

Motivation: questions . . .

What are CpG islands?

Methods: How can . . .

Results: Do CpG . . .

Summary and . . .

Title Page



Page 1 of 37

Go Back

Full Screen

Close

Quit

1. Overview



Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 2 of 37

Go Back

Full Screen

Close

Quit

1. Overview

- Motivation



Outline

Motivation: questions . . .

What are CpG islands?

Methods: How can . . .

Results: Do CpG . . .

Summary and . . .

Title Page



Page 2 of 37

Go Back

Full Screen

Close

Quit

1. Overview

- Motivation
- What are CpG islands?



Outline

Motivation: questions . . .

What are CpG islands?

Methods: How can . . .

Results: Do CpG . . .

Summary and . . .

Title Page



Page 2 of 37

Go Back

Full Screen

Close

Quit

1. Overview

- Motivation
- What are CpG islands?
- Methods: How do CpG islands stand out statistically?
 - † Sequence-defined window and shift parameters
 - † Kullback-Leibler divergence
 - † Truncated Pareto distribution



Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 2 of 37

Go Back

Full Screen

Close

Quit

1. Overview

- Motivation
- What are CpG islands?
- Methods: How do CpG islands stand out statistically?
 - † Sequence-defined window and shift parameters
 - † Kullback-Leibler divergence
 - † Truncated Pareto distribution
- Results: Do CpG islands stand out statistically?
 - † Results on sequence L44140
 - † Results on sequence AL049762



Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 2 of 37

Go Back

Full Screen

Close

Quit

1. Overview

- Motivation
- What are CpG islands?
- Methods: How do CpG islands stand out statistically?
 - † Sequence-defined window and shift parameters
 - † Kullback-Leibler divergence
 - † Truncated Pareto distribution
- Results: Do CpG islands stand out statistically?
 - † Results on sequence L44140
 - † Results on sequence AL049762
- Summary and concluding remarks



Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 2 of 37

Go Back

Full Screen

Close

Quit

2. Motivation: questions to ask



Outline

Motivation: questions . . .

What are CpG islands?

Methods: How can . . .

Results: Do CpG . . .

Summary and . . .

Title Page



Page 3 of 37

Go Back

Full Screen

Close

Quit

2. Motivation: questions to ask

† Are the digits in the expansion of π devoid of any patterns in any base representation?



Outline

Motivation: questions . . .

What are CpG islands?

Methods: How can . . .

Results: Do CpG . . .

Summary and . . .

Title Page



Page 3 of 37

Go Back

Full Screen

Close

Quit

2. Motivation: questions to ask

- † Are the digits in the expansion of π devoid of any patterns in any base representation?
- † Does a biological sequence show any statistically recognizable patterns? What are the biological implications?



Outline

Motivation: questions . . .

What are CpG islands?

Methods: How can . . .

Results: Do CpG . . .

Summary and . . .

Title Page



Page 3 of 37

Go Back

Full Screen

Close

Quit

2. Motivation: questions to ask

- † Are the digits in the expansion of π devoid of any patterns in any base representation?
- † Does a biological sequence show any statistically recognizable patterns? What are the biological implications?
- † A sequence is an ordered array of elements chosen from a given set.



Outline

Motivation: questions . . .

What are CpG islands?

Methods: How can . . .

Results: Do CpG . . .

Summary and . . .

Title Page



Page 3 of 37

Go Back

Full Screen

Close

Quit

2. Motivation: questions to ask

- † Are the digits in the expansion of π devoid of any patterns in any base representation?
- † Does a biological sequence show any statistically recognizable patterns? What are the biological implications?
- † A sequence is an ordered array of elements chosen from a given set.
- † For example, for a DNA sequence, the given set can be { purine, pyrimidine } or { A, G, C, T }.



Outline

Motivation: questions . . .

What are CpG islands?

Methods: How can . . .

Results: Do CpG . . .

Summary and . . .

Title Page



Page 3 of 37

Go Back

Full Screen

Close

Quit

2. Motivation: questions to ask

- † Are the digits in the expansion of π devoid of any patterns in any base representation?
- † Does a biological sequence show any statistically recognizable patterns? What are the biological implications?
- † A sequence is an ordered array of elements chosen from a given set.
- † For example, for a DNA sequence, the given set can be { purine, pyrimidine } or { A, G, C, T }.

For the binary expansion of π (represented in base 2), the given set is { 0, 1 }.



Outline

Motivation: questions . . .

What are CpG islands?

Methods: How can . . .

Results: Do CpG . . .

Summary and . . .

Title Page



Page 3 of 37

Go Back

Full Screen

Close

Quit

The first 3300 digits of binary expansion of π :

```
Out[27]: /$useForm
11. 00100100001111110101010001000100001011010001100001000110100110001001100011
001100010100010111000000001011100000111001101000100101000000100100111000
001000100010100110011110011001100000000100000010111011110101001100011100
0110001001110010110010001001010001010010100000100001111001100011000110100
0000010011011011101111001010100011001101100111001101001110010000011000
11011001100000010101100001010011010111100101011111000010000010111010011
1111100001010101011011010101101010100011100001001000101111001001000001011
0101010111100110001000101111001111100011000110111010001010010000010110
100110100110001011111010101010100001011111101011010101101110111011010
000000101011011011101011011011000110001100111111010101101010100010010
0111111010010101011011011100100100000100010111100010010100011111100
110010010100101000011001100100001110110011001000101101100111011000010
000000000111110011100010100001010000101111100000101100110001101101010
01000001010000011000101010110011100110011010011010001010001111101010
0011111010010010011001110101111000001010001010101110100001110110001
01000110101101001100001100001100011100101010101100100000100001010101
001010111011001110101010100100001011100001001011010101000111011010
10110011100001100001101010100111001010101110010111001000000010011000101
11010001011000001000100100000110000010000101111000010010101000001011
11001001100010100110011001100110011011110000100010111000010010100001010
0001100000001110100001100000011001100100100110000001101001011011000000
00111101000100001111010101000010101101100000001011110000001011100010100
101100100110111000101000101000010101010101010101000010101010101000001
001000100000010000001000000000000000000000000000000000000000000000000000
0010101000010000100101010101000000000000000000000000000000000000000000
0001000010000000000000000000000000000000000000000000000000000000000000
10010110010010010011001110011110011001100110011001100110011001100110011
0110101110100001111100001011010011110000111100100110000101001100101001
11000000100010001001010100001100110110011011100110011001100110010000101
1010110100110110110011010111100010010111110100000010101100110000101
00001000010010011011000010110000000100110010011101101101000010101001
101000010010000000000000000000000000000000000000000000000000000000000000
0111100001000101110101010111010100001010110101101100011011100001001
1000100110000000101101010110001000111001100000000000000000000000000000
0000001110100110010010101010001000010000000000000000000000000000000000
0011110100010000100011001001100000101100001011000010100001010100010000010
0001000000000000000000000000000000000000000000000000000000000000000000
0111100100000011000100110000100001011101101010101101101001010100100100
1100000110010110001100001010101101011000100000000000000000000000000000
1010000010100101101100001010000010111011010001001011000000000000000000
0100010101010101000000010011100001011011011011011000010110101100000100
110111101000111100 ,
```



- Outline
- Motivation: questions ...
- What are CpG islands?
- Methods: How can ...
- Results: Do CpG ...
- Summary and ...

Title Page

◀ ▶

◀ ▶

Page 4 of 37

Go Back

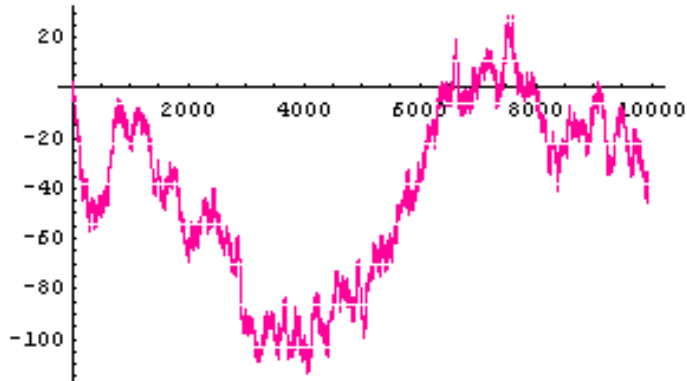
Full Screen

Close

Quit

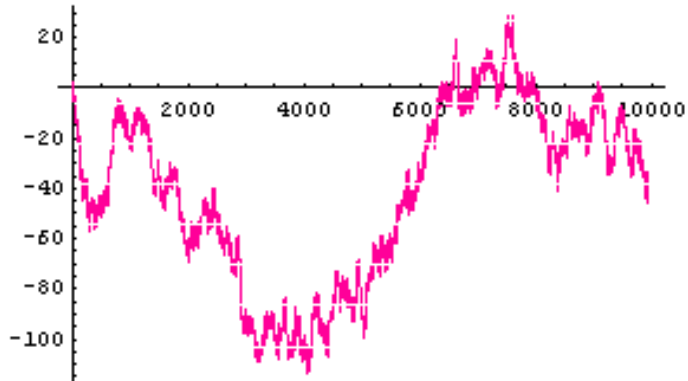
2.1. One-dimensional random walk representation:

A 1-dimensional random walk plot of the first 9966 binary digits of $\pi - 3$:



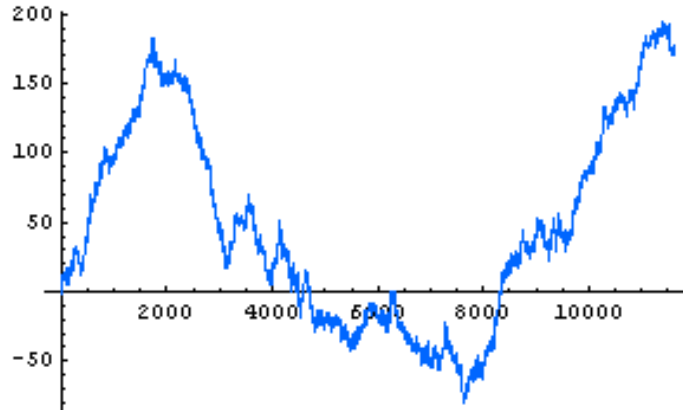
2.1. One-dimensional random walk representation:

A 1-dimensional random walk plot of the first 9966 binary digits of $\pi - 3$:

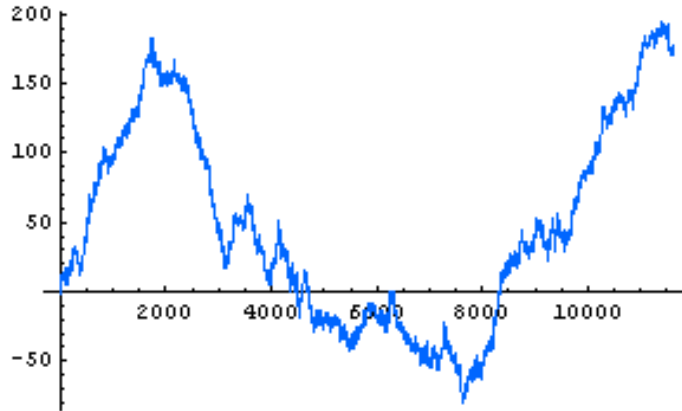


The path randomly walks up one unit if the n th binary place of $\pi - 3$ is 1; otherwise it walks down one unit.

A 1-dimensional random walk plot of the 11624 bases of bacteriophage P4 DNA sequence:



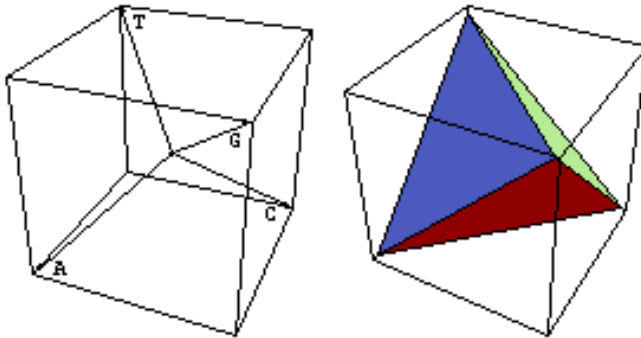
A 1-dimensional random walk plot of the 11624 bases of bacteriophage P4 DNA sequence:



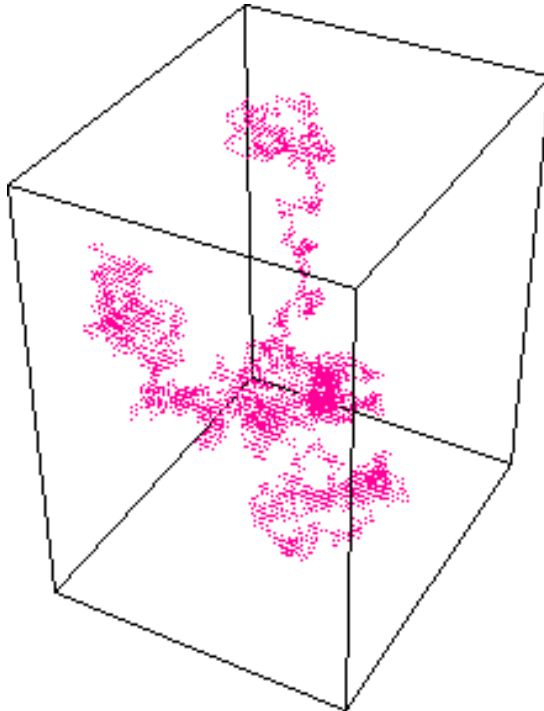
The walk is up 1 unit for a purine base and down 1 unit for a pyrimidine base. Out of the 11624 bases, there are 5901 purines (3008 As and 2893 Gs) and 5723 pyrimidines (2864 Cs and 2859 Ts).

2.2. Three-dimensional random walk representation:

Tetrahedra and the vertex representation of four nucleotides



3-dimensional random walk plot of $\pi - 3$ (in base 4):



Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 9 of 37

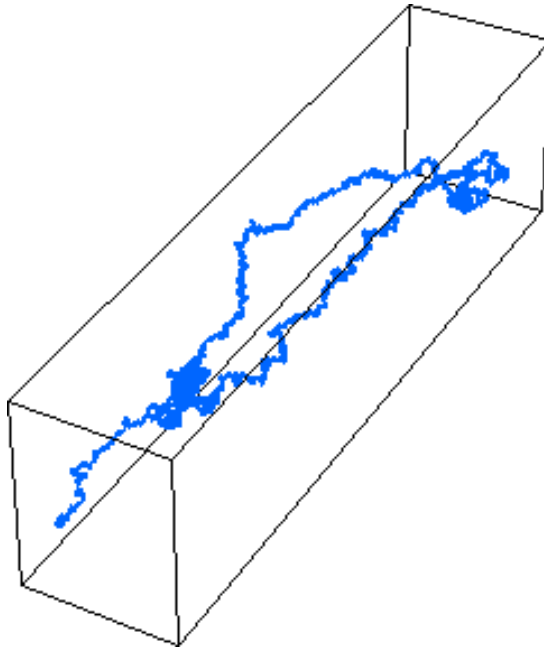
Go Back

Full Screen

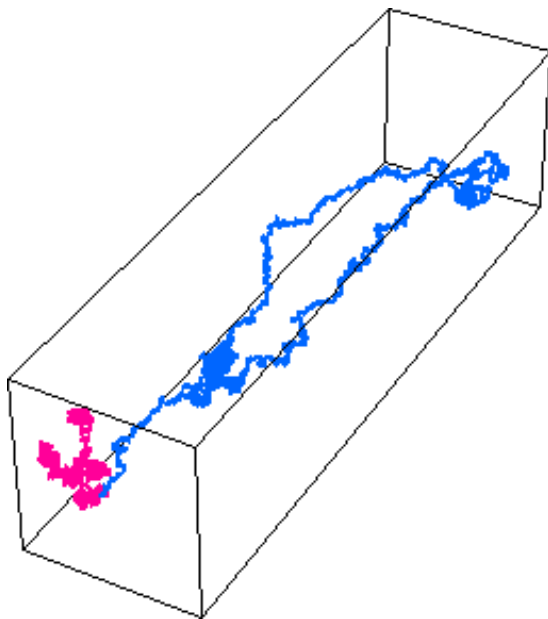
Close

Quit

3-dimensional random walk plot of the 11624 bases of bacteriophage P4 DNA sequence:



Placing the two plots on the same graph:



3. What are CpG islands?



Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 12 of 37

Go Back

Full Screen

Close

Quit

3. What are CpG islands?

† A DNA sequence is a double-stranded helix.

Each strand is linearly composed of four nucleotide bases: A (adenine), T (thymine), G (guanine), and C (cytosine).



courtesy: <http://www.counterbalance.net/cqmedia/dblhlx-body.html>

† Each base is roughly equally used.



Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 13 of 37

Go Back

Full Screen

Close

Quit

- † Each base is roughly equally used.
- † The occurrence percentage of the two adjacent bases C and G is typically low in eukaryotic DNA.



Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 13 of 37

Go Back

Full Screen

Close

Quit

- † Each base is roughly equally used.
- † The occurrence percentage of the two adjacent bases C and G is typically low in eukaryotic DNA.
- † Example: The first sixty bps of the L44140 sequence in Homo sapiens chromosome X region:

aggttcattc g₁₁ctggcagtg tcg₂₃ggagtgc
 cccagagtgg gaagtccg₄₈ag gaattgctgg

Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 13 of 37

Go Back

Full Screen

Close

Quit

- † Each base is roughly equally used.
- † The occurrence percentage of the two adjacent bases C and G is typically low in eukaryotic DNA.
- † Example: The first sixty bps of the L44140 sequence in Homo sapiens chromosome X region:

aggttcattc g₁₁ctggcagtg tcg₂₃ggagtgc
 cccagagtgg gaagtccg₄₈ag gaattgctgg

- † We call this dinucleotide CpG, where the lower case letter “p” indicates that bases C and G are connected by a phosphodiester bond.

Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 13 of 37

Go Back

Full Screen

Close

Quit

† Some statistics of six DNA sequences

There are totally 16 ($= 2^4$) different ordered dinucleotides, and each of them is expected to occur with probability $1/16 = 6.25\%$.



Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 14 of 37

Go Back

Full Screen

Close

Quit

† Some statistics of six DNA sequences

There are totally 16 ($= 2^4$) different ordered dinucleotides, and each of them is expected to occur with probability $1/16 = 6.25\%$.

accession	organism	base pairs	% CpG
AC139751	Mus musculus	183,224	1.362
M63419	Mus musculus	8,735	2.164
AL049762	Homo sapiens	100,575	0.996
AL031723	Homo sapiens	41,255	3.100
AL022327	Homo sapiens	101,270	2.958
L44140	Homo sapiens	219,447	3.290



Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 14 of 37

Go Back

Full Screen

Close

Quit

3.1. What are CpG islands?

† A CpG island is a short DNA fragment that is rich in the CpG dinucleotides.



Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 15 of 37

Go Back

Full Screen

Close

Quit

3.1. What are CpG islands?

- † A CpG island is a short DNA fragment that is rich in the CpG dinucleotides.
- † Gardiner-Garden and Frommer's rule (1987):
 - DNA stretch of length more than 200 bps
 - the ratio of the observed number of CpG dinucleotides to the expected number of CpG dinucleotides exceeds 0.6
 - the G+C content is at least 50%.



Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 15 of 37

Go Back

Full Screen

Close

Quit

3.2. Why do we study CpG islands?

† In the genomes of many higher plants and animals, there is a gene called DNA *methyltransferase* (*Dnmt1*).



Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 16 of 37

Go Back

Full Screen

Close

Quit

3.2. Why do we study CpG islands?

- † In the genomes of many higher plants and animals, there is a gene called DNA *methyltransferase* (*Dnmt1*).
- † *Dnmt1* is an enzyme that can attach a methyl (CH_3) group onto the 5 carbon of cytosine.



Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 16 of 37

Go Back

Full Screen

Close

Quit

3.2. Why do we study CpG islands?

- † In the genomes of many higher plants and animals, there is a gene called DNA *methyltransferase* (*Dnmt1*).
- † *Dnmt1* is an enzyme that can attach a methyl (CH_3) group onto the 5 carbon of cytosine.
- † In vertebrate DNA, this methylation process targets at the cytosines jointed next to guanines by a phosphodiester bond on the same strand.



Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 16 of 37

Go Back

Full Screen

Close

Quit

3.2. Why do we study CpG islands?

- † In the genomes of many higher plants and animals, there is a gene called DNA *methyltransferase* (*Dnmt1*).
- † *Dnmt1* is an enzyme that can attach a methyl (CH_3) group onto the 5 carbon of cytosine.
- † In vertebrate DNA, this methylation process targets at the cytosines jointed next to guanines by a phosphodiester bond on the same strand.
- † The locations of methylated cytosines do not seem to be random.



Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 16 of 37

Go Back

Full Screen

Close

Quit

† A high proportion of methylated cytosines are found in inactive genes.



Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 17 of 37

Go Back

Full Screen

Close

Quit

† A high proportion of methylated cytosines are found in inactive genes.

The non-methylated CpG dinucleotides are often located around the promoters of housekeeping genes or some tissue-specific genes.



Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 17 of 37

Go Back

Full Screen

Close

Quit

† A high proportion of methylated cytosines are found in inactive genes.

The non-methylated CpG dinucleotides are often located around the promoters of housekeeping genes or some tissue-specific genes.

† CpG islands are gene associated and can be used as markers to identify genes

Outline

Motivation: questions . . .

What are CpG islands?

Methods: How can . . .

Results: Do CpG . . .

Summary and . . .

Title Page



Page 17 of 37

Go Back

Full Screen

Close

Quit

† A high proportion of methylated cytosines are found in inactive genes.

The non-methylated CpG dinucleotides are often located around the promoters of housekeeping genes or some tissue-specific genes.

† CpG islands are gene associated and can be used as markers to identify genes

Antequera and Bird used CpG islands to estimate the number of genes in humans and mice.

Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 17 of 37

Go Back

Full Screen

Close

Quit

4. Methods: How can CpG island stand out statistically?

Basic statistical reasoning:

Asking whether a DNA stretch stands out significantly enough to be distinguished from its background as a CpG island.



Outline

Motivation: questions...

What are CpG islands?

Methods: How can ...

Results: Do CpG...

Summary and...

Title Page



Page 18 of 37

Go Back

Full Screen

Close

Quit

4. Methods: How can CpG island stand out statistically?

Basic statistical reasoning:

Asking whether a DNA stretch stands out significantly enough to be distinguished from its background as a CpG island.

Two issues:



Outline

Motivation: questions...

What are CpG islands?

Methods: How can ...

Results: Do CpG...

Summary and...

Title Page



Page 18 of 37

Go Back

Full Screen

Close

Quit

4. Methods: How can CpG island stand out statistically?

Basic statistical reasoning:

Asking whether a DNA stretch stands out significantly enough to be distinguished from its background as a CpG island.

Two issues:

- (1) Window size: determining a proper length of a stretch for screening



Outline

Motivation: questions...

What are CpG islands?

Methods: How can ...

Results: Do CpG...

Summary and...

Title Page



Page 18 of 37

Go Back

Full Screen

Close

Quit

4. Methods: How can CpG island stand out statistically?

Basic statistical reasoning:

Asking whether a DNA stretch stands out significantly enough to be distinguished from its background as a CpG island.

Two issues:

- (1) Window size: determining a proper length of a stretch for screening
- (2) Shift size: determining a proper step size that a stretch is shifted along the sequence while the screening is taken.



Outline

Motivation: questions...

What are CpG islands?

Methods: How can ...

Results: Do CpG...

Summary and...

Title Page



Page 18 of 37

Go Back

Full Screen

Close

Quit

† Sequence-defined window and shift parameters

Example: The first sixty bps of the L44140 sequence in Homo sapiens chromosome X region:

```
aggttcattc   g11ctggcagtg   tcg23ggagtgc  
cccagagtgg   gaagtccg48ag   gaattgctgg
```

1st CpG interarrival: 11

2nd CpG interarrival: $23 - 11 = 12$

3rd CpG interarrival: $48 - 23 = 25$



Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 19 of 37

Go Back

Full Screen

Close

Quit

Some statistics on the CpG interarrivals:

accession	bps	# CpG	max	mean	SD
AC139751	183,224	2,496	984	73.37	99.01
M63419	8,735	189	578	46.20	70.04
AL049762	100,575	1,002	1,309	100.36	140.88
AL031723	41,255	1,279	368	32.22	39.69
AL022327	101,270	2,996	465	33.79	44.51
L44140	219,447	7,220	497	30.38	43.18

Some statistics on the CpG interarrivals:

accession	bps	# CpG	max	mean	SD
AC139751	183,224	2,496	984	73.37	99.01
M63419	8,735	189	578	46.20	70.04
AL049762	100,575	1,002	1,309	100.36	140.88
AL031723	41,255	1,279	368	32.22	39.69
AL022327	101,270	2,996	465	33.79	44.51
L44140	219,447	7,220	497	30.38	43.18

Let I_1, I_2, \dots , and I_v be the lengths of the CpG interarrivals of the sequence. We set

$$w = \text{window size} = \max(I_1, I_2, \dots, I_v)$$

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

† Kullback-Leibler divergence



Outline

Motivation: questions . . .

What are CpG islands?

Methods: How can . . .

Results: Do CpG . . .

Summary and . . .

Title Page



Page 21 of 37

Go Back

Full Screen

Close

Quit

† Kullback-Leibler divergence

Let f_0 and f_1 be two probability mass functions.



- Outline
- Motivation: questions . . .
- What are CpG islands?
- Methods: How can . . .**
- Results: Do CpG . . .
- Summary and . . .

Title Page

◀▶

◀▶

Page 21 of 37

Go Back

Full Screen

Close

Quit

† Kullback-Leibler divergence

Let f_0 and f_1 be two probability mass functions.

Kullback interpreted the logarithmic ratio $\log \frac{f_1(x)}{f_0(x)}$ as the information in the observation x for discriminating $f_1(x)$ against $f_0(x)$.



Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 21 of 37

Go Back

Full Screen

Close

Quit

† Kullback-Leibler divergence

Let f_0 and f_1 be two probability mass functions.

Kullback interpreted the logarithmic ratio $\log \frac{f_1(x)}{f_0(x)}$ as the information in the observation x for discriminating $f_1(x)$ against $f_0(x)$.

The Kullback-Leibler divergence of $f_1(x)$ against $f_0(x)$ is defined to be

$$\text{div}_{\text{K-L}}(f_1, f_0) = \sum_x f_1(x) \log \frac{f_1(x)}{f_0(x)}.$$



Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 21 of 37

Go Back

Full Screen

Close

Quit

Example:

Let $f_0(\text{head}) = 0.4$ and $f_0(\text{tail}) = 0.6$.

Let $f_1(\text{head}) = p$ and $f_1(\text{tail}) = 1 - p$.



Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 22 of 37

Go Back

Full Screen

Close

Quit

Example:

Let $f_0(\text{head}) = 0.4$ and $f_0(\text{tail}) = 0.6$.

Let $f_1(\text{head}) = p$ and $f_1(\text{tail}) = 1 - p$.

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



Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 22 of 37

Go Back

Full Screen

Close

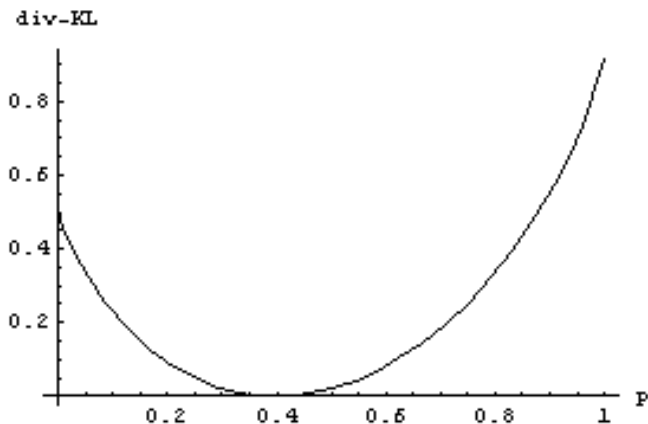
Quit

Example:

Let $f_0(\text{head}) = 0.4$ and $f_0(\text{tail}) = 0.6$.

Let $f_1(\text{head}) = p$ and $f_1(\text{tail}) = 1 - p$.

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



Let $f_0(\text{CpG})$ be the CpG proportion of the bulk DNA.



Outline

Motivation: questions . . .

What are CpG islands?

Methods: How can . . .

Results: Do CpG . . .

Summary and . . .

Title Page



Page 23 of 37

Go Back

Full Screen

Close

Quit

Let $f_0(\text{CpG})$ be the CpG proportion of the bulk DNA.

Let $f_{1,W}(\text{CpG})$ be the observed CpG proportion over window W .



Outline

Motivation: questions . . .

What are CpG islands?

Methods: How can . . .

Results: Do CpG . . .

Summary and . . .

Title Page



Page 23 of 37

Go Back

Full Screen

Close

Quit

Let $f_0(\text{CpG})$ be the CpG proportion of the bulk DNA.

Let $f_{1,W}(\text{CpG})$ be the observed CpG proportion over window W .

Then the Kullback-Leibler divergence of this particular window W is



Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 23 of 37

Go Back

Full Screen

Close

Quit

Let $f_0(\text{CpG})$ be the CpG proportion of the bulk DNA.

Let $f_{1,W}(\text{CpG})$ be the observed CpG proportion over window W .

Then the Kullback-Leibler divergence of this particular window W is

$$\text{div}(f_{1,W}, f_0) = f_{1,W}(\text{CpG}) \log \frac{f_{1,W}(\text{CpG})}{f_0(\text{CpG})} + (1 - f_{1,W}(\text{CpG})) \log \frac{1 - f_{1,W}(\text{CpG})}{1 - f_0(\text{CpG})}.$$



Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 23 of 37

Go Back

Full Screen

Close

Quit

Each window has a divergence value. A high divergence window indicates the proximity of a CpG island.



- Outline
- Motivation: questions...
- What are CpG islands?
- Methods: How can...**
- Results: Do CpG...
- Summary and...

Title Page

◀ ▶

◀ ▶

Page 24 of 37

Go Back

Full Screen

Close

Quit

Each window has a divergence value. A high divergence window indicates the proximity of a CpG island.

The divergence values along the window indices for the L44140 sequence in Homo sapiens chromosome X region:



Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 24 of 37

Go Back

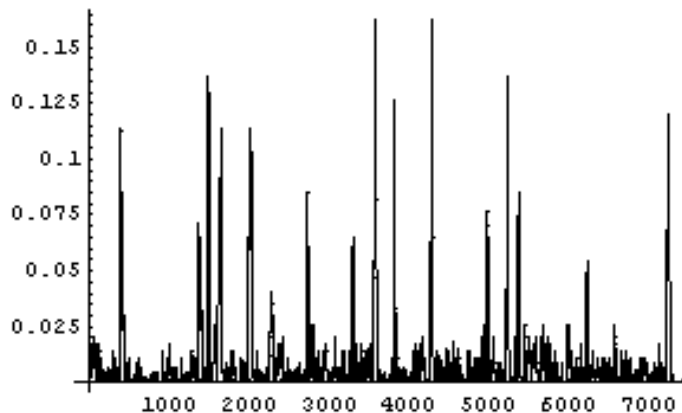
Full Screen

Close

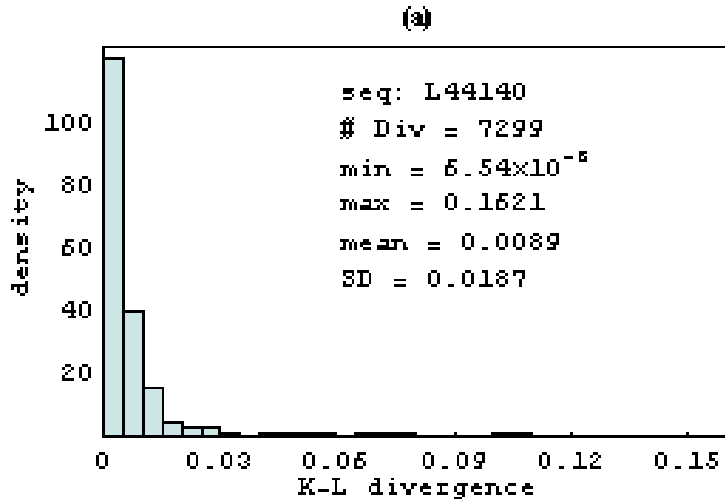
Quit

Each window has a divergence value. A high divergence window indicates the proximity of a CpG island.

The divergence values along the window indices for the L44140 sequence in Homo sapiens chromosome X region:

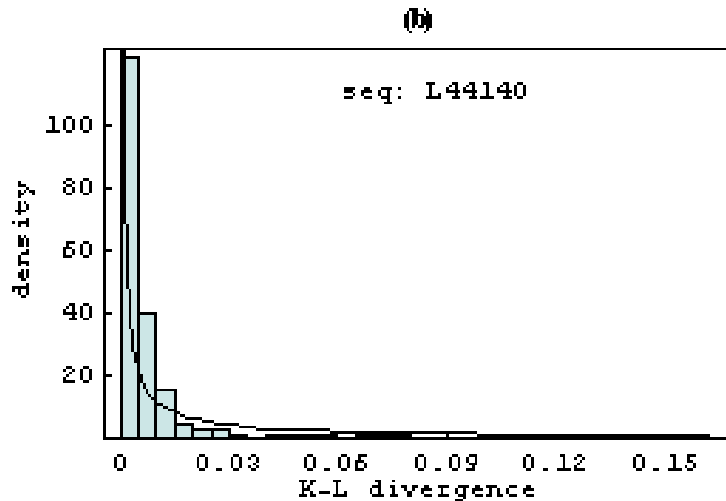


The histogram and some statistics of the divergence values of the L44140 sequence in Homo sapiens chromosome X region:



† Assessing statistical significance:

The Pareto fit of the divergence values of the L44140 sequence in Homo sapiens chromosome X region:



A truncated Pareto distribution has the density

$$f(x) = \frac{r + 1}{\beta^{r+1} - \alpha^{r+1}} x^r, \quad 0 < \alpha < x < \beta,$$

where α is the lower bound parameter of the data range, β is the upper bound parameter, and $r \neq -1$ is the power parameter.

Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 27 of 37

Go Back

Full Screen

Close

Quit

A truncated Pareto distribution has the density

$$f(x) = \frac{r + 1}{\beta^{r+1} - \alpha^{r+1}} x^r, \quad 0 < \alpha < x < \beta,$$

where α is the lower bound parameter of the data range, β is the upper bound parameter, and $r \neq -1$ is the power parameter.

Let x_p be the cutoff of the top $(100 \times p)\%$ of the divergences, i.e.,

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

then

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

Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 27 of 37

Go Back

Full Screen

Close

Quit

† Maximum likelihood estimates (MLEs) of the truncated Pareto distribution



Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 28 of 37

Go Back

Full Screen

Close

Quit

† Maximum likelihood estimates (MLEs) of the truncated Pareto distribution

If X_1, X_2, \dots, X_n are a random sample from the truncated Pareto distribution, then the maximum likelihood estimators (MLEs) $\hat{\alpha}$, $\hat{\beta}$, and \hat{r} of the parameters α , β , and r are:

$$\hat{\alpha} = \min(X_1, X_2, \dots, X_n),$$

$$\hat{\beta} = \max(X_1, X_2, \dots, X_n),$$

and \hat{r} is the unique solution to the equation

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

Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 28 of 37

Go Back

Full Screen

Close

Quit

For the L44140 sequence in Homo sapiens chromosome X region, the MLEs are

$$\hat{\alpha}_{L44140} = 6.54 \times 10^{-6}, \hat{\beta}_{L44140} = 0.1621, \hat{r}_{L44140} = -0.9144.$$



Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 29 of 37

Go Back

Full Screen

Close

Quit

For the L44140 sequence in Homo sapiens chromosome X region, the MLEs are

$$\hat{\alpha}_{L44140} = 6.54 \times 10^{-6}, \hat{\beta}_{L44140} = 0.1621, \hat{r}_{L44140} = -0.9144.$$

The estimated top fifth percentile is

$$\hat{x}_{0.05, L44140} = 0.11498.$$



Outline

Motivation: questions . . .

What are CpG islands?

Methods: How can . . .

Results: Do CpG . . .

Summary and . . .

Title Page



Page 29 of 37

Go Back

Full Screen

Close

Quit

For the L44140 sequence in Homo sapiens chromosome X region, the MLEs are

$$\hat{\alpha}_{L44140} = 6.54 \times 10^{-6}, \quad \hat{\beta}_{L44140} = 0.1621, \quad \hat{r}_{L44140} = -0.9144.$$

The estimated top fifth percentile is

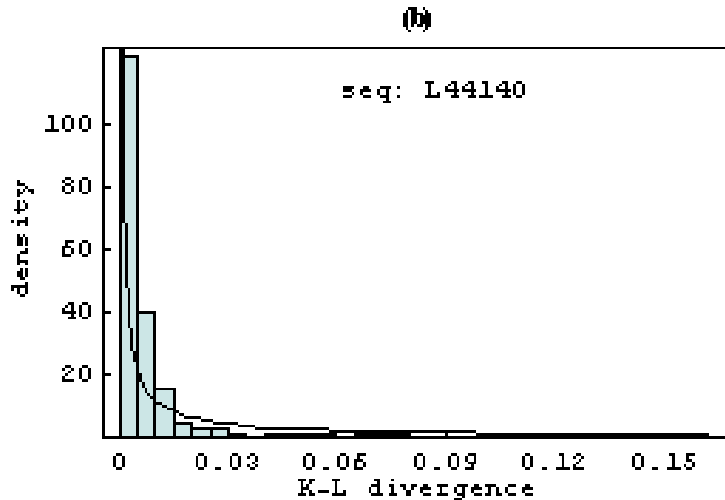
$$\hat{x}_{0.05, L44140} = 0.11498.$$

The fitted truncated Pareto density is

$$\hat{f}_{L44140}(x) = 0.1726x^{-0.9144}, \quad 6.54 \times 10^{-6} \leq x \leq 0.1621.$$

[Outline](#)[Motivation: questions...](#)[What are CpG islands?](#)[Methods: How can...](#)[Results: Do CpG...](#)[Summary and...](#)[Title Page](#)[Page 29 of 37](#)[Go Back](#)[Full Screen](#)[Close](#)[Quit](#)

Again, here is the Pareto fit to the histogram of the divergences of the L44140 sequence in Homo sapiens chromosome X region.



5. Results: Do CpG island stand out statistically?

Idea:

We use the estimated top $(100 \times p)$ th percentile \hat{x}_p to decide high divergence regions for locating CpG islands.



Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 31 of 37

Go Back

Full Screen

Close

Quit

5. Results: Do CpG island stand out statistically?

Idea:

We use the estimated top $(100 \times p)$ th percentile \hat{x}_p to decide high divergence regions for locating CpG islands.

This chosen top $(100 \times p)$ th percentile \hat{x}_p quantifies the statistical significance in locating CpG islands.



Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 31 of 37

Go Back

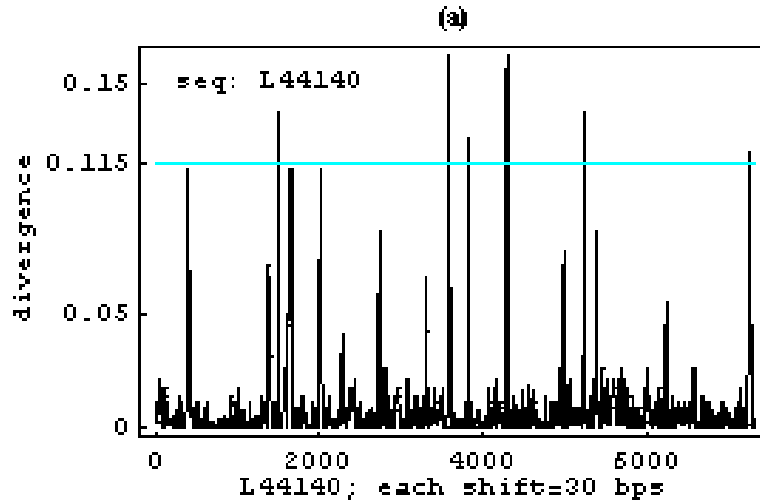
Full Screen

Close

Quit

5.1. Results on sequence L44140

If we use the estimated top fifth percentile $\hat{x}_{0.05, L44140} = 0.11498$ for the L44140 sequence in Homo sapiens chromosome X, we can find six high divergence regions.



We tabulate those six regions with their window indices and base positions below.



region num.	window index	base positions
(i)	1493–1497	44760–45377
(ii)	3565–3578	106920–107807
(iii)	3813–3816	114360–114947
(iv)	4273–4284	128160–128987
(v)	5224–5229	156690–157337
(vi)	7223	216660–217157

Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 33 of 37

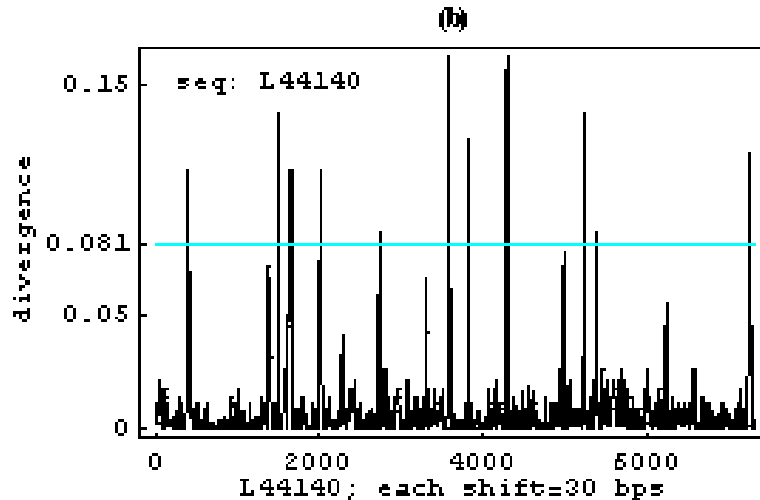
Go Back

Full Screen

Close

Quit

If we use the estimated top tenth percentile $\hat{x}_{0.10, L44140} = 0.08071$ for the L44140 sequence, we can find eleven high divergence regions.



Note: *GenBank* reports 17 CpG islands.



Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 34 of 37

Go Back

Full Screen

Close

Quit

5.2. Results on sequence AL049762

GenBank: It is the human DNA sequence from clone RP1-81F6 on chromosome 1q24.1-25.2. It is the complete sequence containing the 5' end of the gene for HBxAg transactivated protein 2 (XTP2).



Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 35 of 37

Go Back

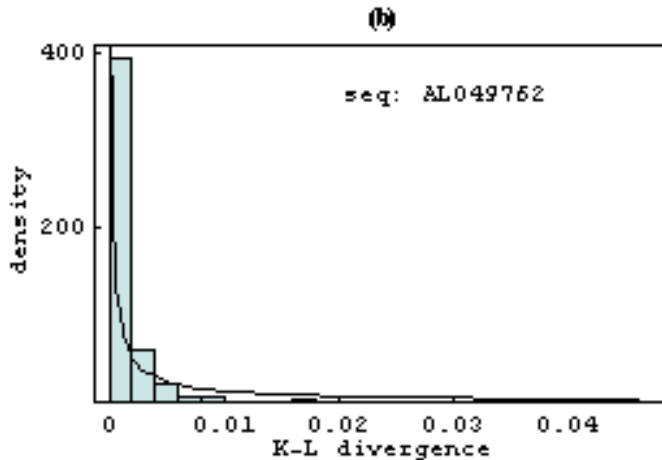
Full Screen

Close

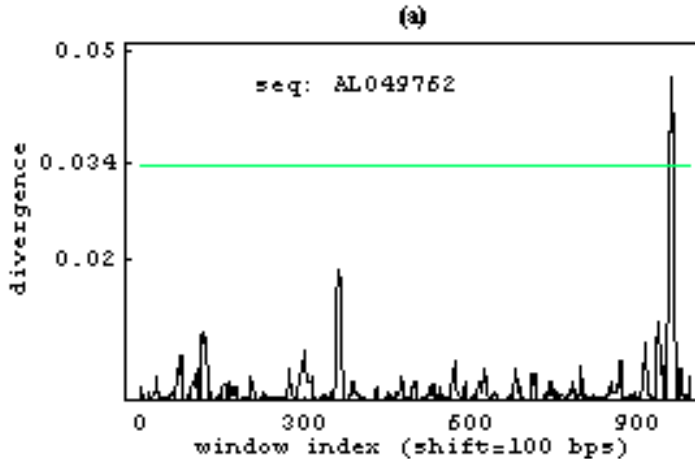
Quit

5.2. Results on sequence AL049762

GenBank: It is the human DNA sequence from clone RP1-81F6 on chromosome 1q24.1-25.2. It is the complete sequence containing the 5' end of the gene for HBxAg transactivated protein 2 (XTP2). Here is the Pareto fit for the divergences.



If we choose the estimated top fifth percentile $\hat{x}_{0.05, AL049762} = 0.03356$, we can find one high divergence region with base positions from 95500 to 97509.



GenBank reports a putative CpG island that has base positions from 96101 to 96822.

Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 36 of 37

Go Back

Full Screen

Close

Quit

6. Summary and concluding remarks



Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 37 of 37

Go Back

Full Screen

Close

Quit

6. Summary and concluding remarks

- CpG islands can be used as markers to identify genes.



Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 37 of 37

Go Back

Full Screen

Close

Quit

6. Summary and concluding remarks

- CpG islands can be used as markers to identify genes.
- We use the window and shift parameters defined by a sequence itself.



Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 37 of 37

Go Back

Full Screen

Close

Quit

6. Summary and concluding remarks

- CpG islands can be used as markers to identify genes.
- We use the window and shift parameters defined by a sequence itself.
- We employ the Kullback-Leibler divergence for locating CpG islands.



Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 37 of 37

Go Back

Full Screen

Close

Quit

6. Summary and concluding remarks

- CpG islands can be used as markers to identify genes.
- We use the window and shift parameters defined by a sequence itself.
- We employ the Kullback-Leibler divergence for locating CpG islands.
- We use truncated Pareto distributions to fit the divergence values and quantify the statistical significance.



Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 37 of 37

Go Back

Full Screen

Close

Quit

6. Summary and concluding remarks

- CpG islands can be used as markers to identify genes.
- We use the window and shift parameters defined by a sequence itself.
- We employ the Kullback-Leibler divergence for locating CpG islands.
- We use truncated Pareto distributions to fit the divergence values and quantify the statistical significance.
- The Kullback-Leibler divergence can be used to measure *linkage disequilibrium*.



Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 37 of 37

Go Back

Full Screen

Close

Quit

6. Summary and concluding remarks

- CpG islands can be used as markers to identify genes.
- We use the window and shift parameters defined by a sequence itself.
- We employ the Kullback-Leibler divergence for locating CpG islands.
- We use truncated Pareto distributions to fit the divergence values and quantify the statistical significance.
- The Kullback-Leibler divergence can be used to measure *linkage disequilibrium*.
- When applying statistics, understanding the context is very important.



Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 37 of 37

Go Back

Full Screen

Close

Quit

6. Summary and concluding remarks

- CpG islands can be used as markers to identify genes.
- We use the window and shift parameters defined by a sequence itself.
- We employ the Kullback-Leibler divergence for locating CpG islands.
- We use truncated Pareto distributions to fit the divergence values and quantify the statistical significance.
- The Kullback-Leibler divergence can be used to measure *linkage disequilibrium*.
- When applying statistics, understanding the context is very important.

THE END



Outline

Motivation: questions...

What are CpG islands?

Methods: How can...

Results: Do CpG...

Summary and...

Title Page



Page 37 of 37

Go Back

Full Screen

Close

Quit