(I deliver this content through the course website then do a short introduction in lecture to orient the students.)

Bioinformatics: CpG Islands

Objectives:

- Students will determine the frequency and location of CpG islands in Eukaryotic promoters.
- Students will download and manipulate large DNA sequence files from an on-line genome resource and analyze the sequences using on-line programs.
- Students will invent a way to display the combined data from multiple promoter sequences

Modern biologists need to be be able to extract useful information from large, complex datasets. This is especially true in the area of genetics where fast, cheap DNA sequencing has become a central tool. In this assignment you will be challenged to solve problems related to the acquisition, analysis, and display of a large complex dataset. (The directions are intentionally left somewhat vague to encourage you to innovate!)

Background: CpG Islands

"CpG" (also called a "CG dinucelotide") is a spot in a strand of DNA where a C nucleotide is followed by a G nucleotide. The "p" in CpG refers to the phosphodiester bond between the nucleotides. Do not confuse CpG with a CG base pair (see figure below).

CpGs are important because they are targeted by enzymes called DNA methyltransferases. These enzymes attach a methyl group ($-CH_3$) to the C in the CpG dinucleotide. As we will discuss in class CpG methylation is important because it can change the shape of the DNA to influence gene expression (Scitable Article on DNA Methylation).

CpG dinucleotides occur less frequently in vertebrate genomes that would be predicted based on the %C and %G in these genomes (You can easily pick out CpG in the figure below showing the frequency of all 16 dinucleotides).

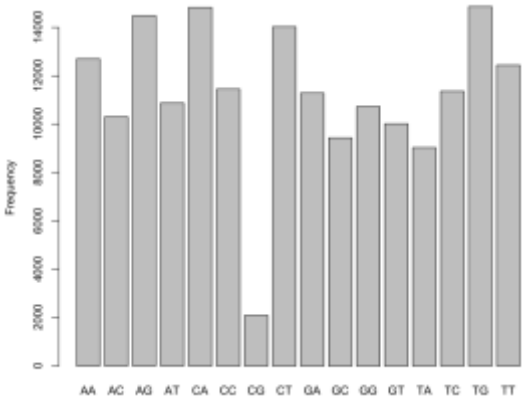
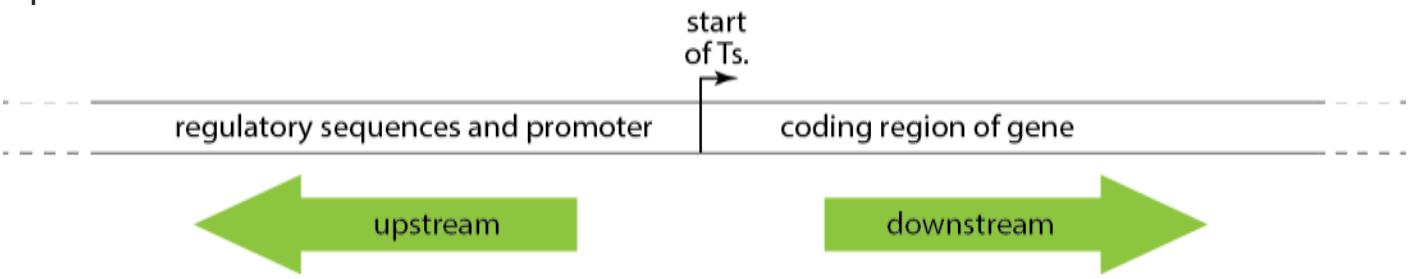


image grom mintgene.

However, regions of DNA can be found in which CpG is very abundant. These clusters are known as **CpG islands**. These islands can become heavily methylated (as the methytransferase enzymes add -CH_3 to the CpG sequences). When methylated, CpG islands become tightly coiled limiting protein access to the DNA.

In this activity you will search for CpG islands in the DNA sequences 5000 base pairs (bp) upstream of the genes of a vertebrate organism of your choice. The image below shows a simple map of a gene. The bent arrow in the middle represents the place where transcription starts. The coding sequence (the part of the gene that actually encodes a protein) is found downstream of the start of transcription. The promoter and other regulator sequences that control the expression of the gene are found upstream of the start of transcription. It is the upstream region that you will search for CpG islands.



Step 1: Downloading Sequence Files

Since promoter regions are of general interest to molecular biologists most genome projects create supplemental files that contain only the sequence upstream of each gene in the genome.

1 Go to the UCSC Genome Download Site
(http://hgdownload.cse.ucsc.edu/downloads.html)

2 Choose a vertebrate genome from the list and click on the name

3 Click on the link to the most recent Full Data Set for your organism.

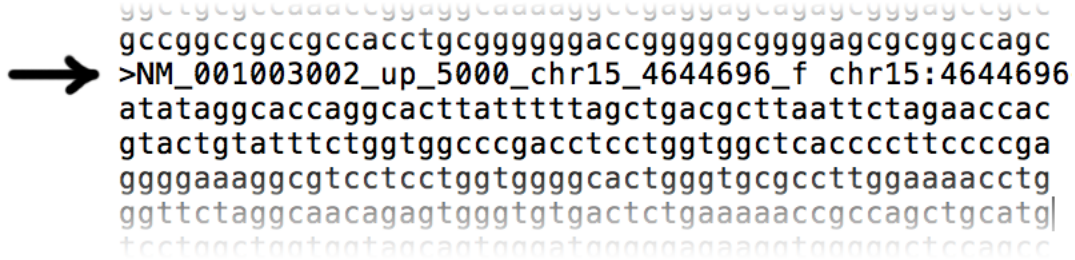
4 Scroll down to the list of available files and look for one called

upstream5000.fa.gz. (If you don't see a file with this name, go back to step 2 and choose a different organism.)

5 Click on the upstream5000 file to begin downloading this file to your computer. This is a large text file so don't be surprised if it takes a few minutes to download.

6 Open the file using a text editor (TextEdit on Mac, NotePad on most PCs). Enjoy scrolling through millions of nucleotides!

The file you opened is in fasta format...this is a common format for DNA sequence files. As you scroll through your file you will see lines of 50 nucleotides occasionally broken up by a line beginning with ">"



This line marks the start of a new sequence. The lines starting with ">" provide information about the sequence that follows. In the case of the sequences you will be analyzing the first nucleotides (atataggc... in the case of the figure shown above) are the nucleotides farthest from the start of transcription...these nucleotides are 5000 bp upstream of the start of transcription (position 1 in your file corresponds to 5000 bp upstream, position 2 corresponds to 4999 bp upstream, etc.). Note: You may see sequence files that contain a large number of n (lines of "nnnnnnnnnnnnnnnnnnnnnnnnn..."). DO NOT use sequences with n's in the subsequent steps..."n" represents an unknown nucleotide, usually the result of incomplete or inaccurate sequencing. Use only sequences consisting entirely of a, g, c and t.

Step 2: Searching for CpG Islands

There are a number of software packages that can search for CpG islands. For this activity I've chosen a tool developed at the University of Taiwan because it seemed easy to use and I liked the output. In general CpG searching involves the following steps (NOTE: you don't need to do the calculations...the site does it for you. I just want you to know what the program is doing.)

- Count all the nucleotides in a 200 bp "window" starting with nucleotides 1-200.
- Count all C and G nucleotides in the window.
- Determine the frequency of CpG dinucleotides in the window.

- If the %CG is greater than 50% and if the number of observed CpG dinucleotides is greater than the number of expected CpGs then the sequence is identified as a CpG island.
- The window moves 1 nucleotide down the sequence and the calculations are repeated. This continues until the whole sequence has been analyzed in this way and all CpG islands are identified.

•

To analyze a sequence (you should analyze at least 20 sequences for this assignment!)

1 From your fasta file, copy the entire 5000 bp sequence for one gene (**DO NOT** include the information line beginning with "<".)

2 Go to the Database of CpG Islands (http://dbcat.cgm.ntu.edu.tw)*

3 Paste the sequence into the small window on the left sidebar.

4 Click "Go!"...in a few seconds your results should be displayed.

Step 3: Display Your Data

In the genomic era it can be a challenge to display the large amounts of data to which researchers have access. Your goal here is to create a single, clear image that summarizes the results of analyzing multiple (minimum of 20) upstream sequences. These are the criteria:

- A single image that summarizes the location and frequency of CpG islands in the 5000 bp upstream of the start of transcription in your organism of choice. This can be generated with a computer or by hand.
- A figure legend with this title (in bold) "**CpG Islands in the Promoter Regions of** Genus species, **common name** (customize for your organism)". The title should be followed by a few sentences describing how to navigate the image, how the sequence data was obtained, the number of sequences analyzed to create the image, the percent of sequences analyzed that contained CpG islands, and any other information that would help a geneticist understand your data.
- The image and figure legend must fit on a single page...save this as a PDF and upload to PolyLearn it using the link below.

There are 8 points possible for this assignment. It is expected to take you 1-3 hours...before you start copying and pasting sequences, give some thought to a good system for recording your observations that will make it easier to display visually in your figure. Grading is based on the following criteria:

4 points...Downloading and accessing sequence files and obtaining CpG

data
2 points...Using your ingenuity to create a clear, information-rich image.
2 points...Figure legend (guidelines above) and correct interpretation of data.

Some samples of student work are on the following pages:

Makenna Clark
5/31/17
BIO 351

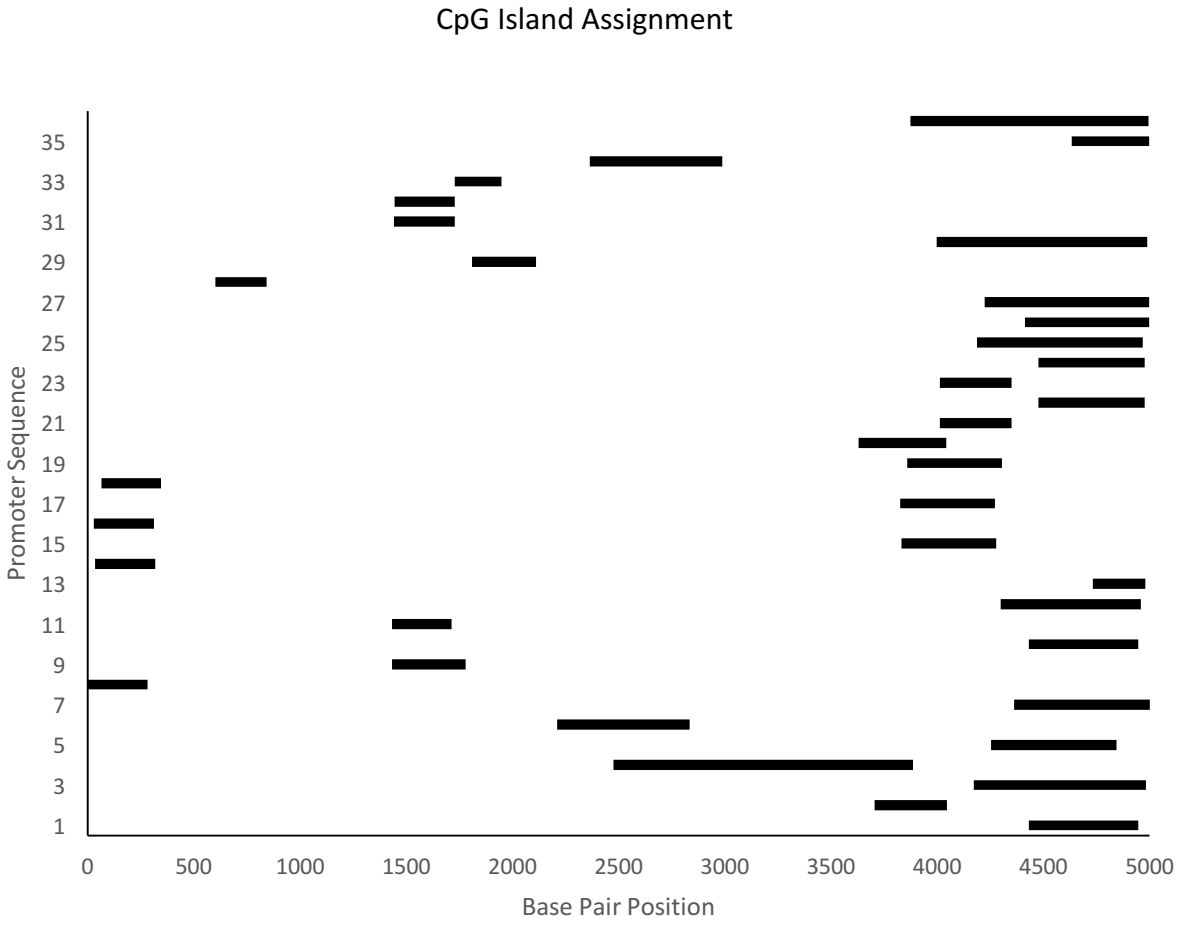
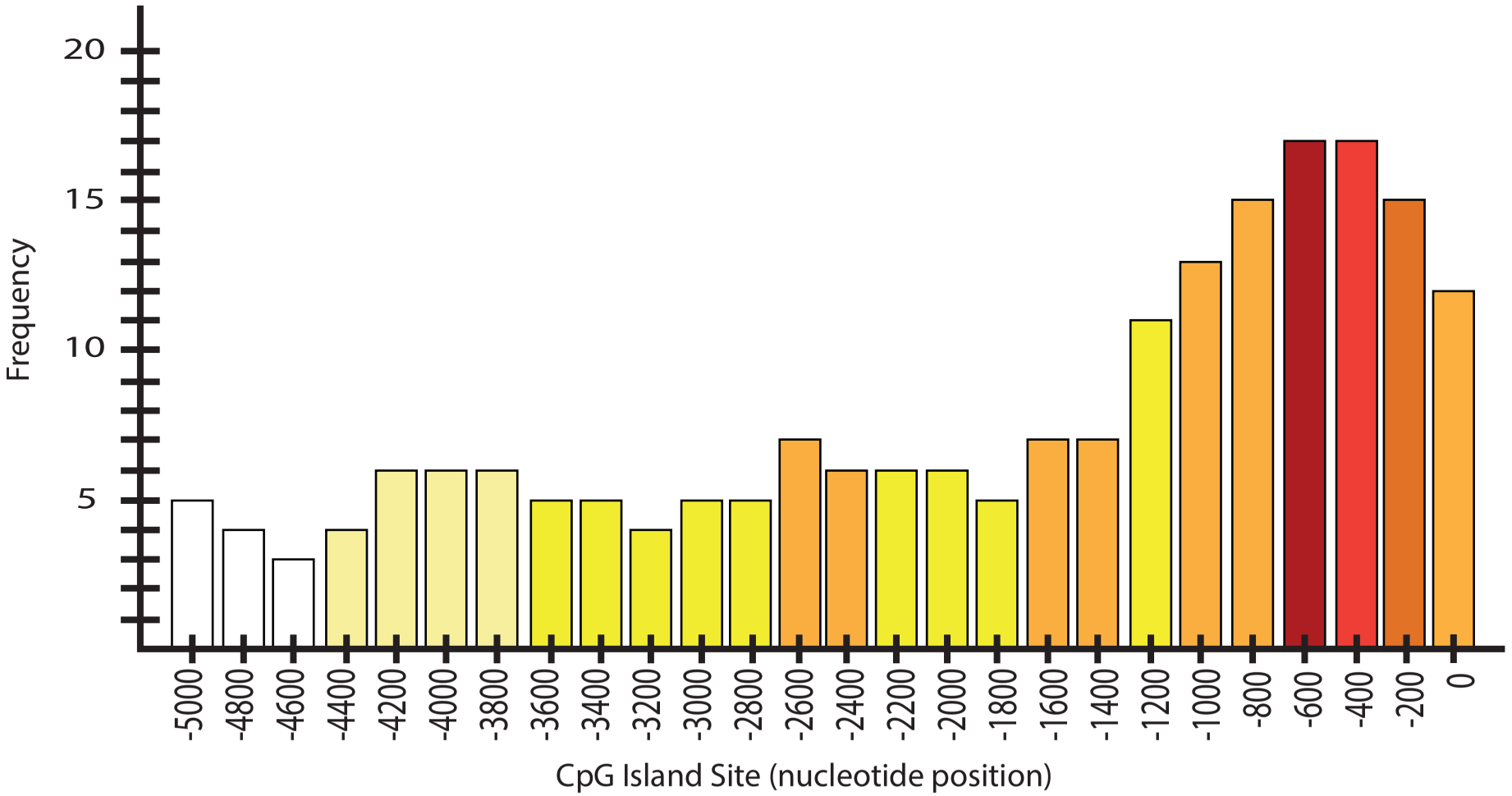


Fig 1: CpG Islands in the Promoter Regions of *Callithrix* (common name, Marmoset). The x-axis shows the base pair position of the upstream promoter sequence and the y-axis shows the number of CpG islands found from analyzing promoter sequences. The bars on the graph show the length of the CpG island and their relative location. The sequence data was provided by UCSC Genome Bioinformatics from March of 2009. I analyzed a total of 30 sequences and roughly 76% of those contained a CpG island that I could use to create this image. By lining up the CpG islands you can see the frequency of the location of the CpG island in many upstream promoter sequences. This data suggests that most CpG islands are found in the -4500 to -5000 base region, this correlates to -1 to -500 in regards to the start of transcription.



CpG Islands in the Promoter Regions of *Felis catus*, Domestic Cat

In order to create the figure shown above, the genome of the domestic cat (provided by the University of California, Santa Cruz Genome Browser) was sequenced in segments of 5000 nucleotides using software developed by the University of Taiwan. Thirty sequences were completed and the site of the CpG island was recorded to create a chart displaying the frequency of CpG islands in certain areas of the promoter regions of the domestic cat genome. Of the thirty sequences, twenty five of them contained island (83.3%). The color of each of the bars in the chart is also significant, the more red the color implies that there was a higher frequency of "dense" CpG islands in these regions. As can be seen from the image, position -600 had the highest frequency of "dense" CpG islands and position -4800 had no "dense" CpG islands.

CpG Islands in the Promotor Regions of Bos taurus, common name cow

The 5,000 base pair sequences shown are the promoter region in different genes, where position -1 represents the base pair leading up to the start of transcription. The sequences were obtained from University of California, Santa Cruz's Genome Bioinformatics website. The sequences were then input in to a CpG island finder site that identified the locations of CpG islands. The highlighted regions in the figure below represent the locations of CpG islands from 23 upstream sequences. The darker green regions represent the denser island between two island in the same sequence. Of the 23 sequences looked at, 61 percent contained CpG islands.

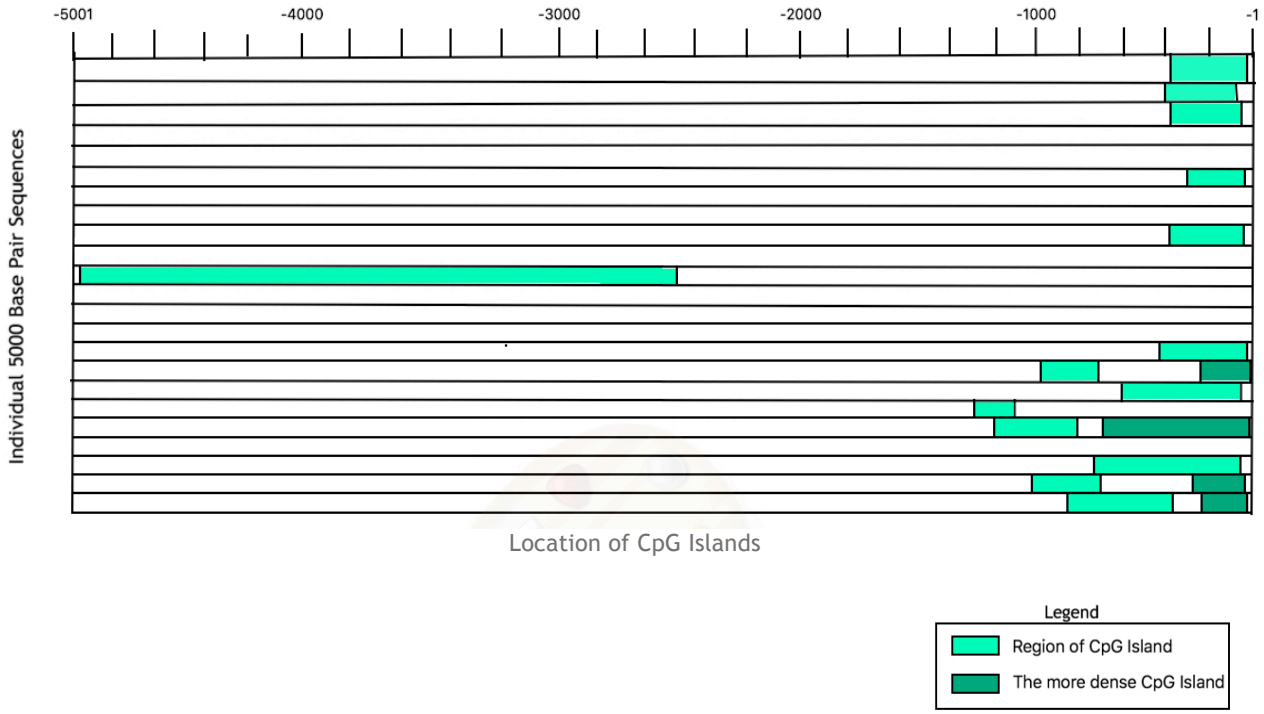


Figure 1. Location of CpG islands in 23 5,000 base pairs sequences upstream from the start of transcription in Bos taurus, or commonly cow.

HAROLD the HEDGEHOG

CpG Islands in the Promotor Regions of Erinaceidae Erinaceinae, Hedgehog

This image was created using data from UCSC Genome Bioinformatics analyzing 20, 5000 bp regions on CHR1 upstream from the start of transcription. Each of the Hedgehogs large spikes represents one of these sequences. The smaller spikes protruding from the others represent the CpG islands found in the sequences. The relative width, height, and position on the spike represent the length, GC content, and proximity to the start of transcription, respectively. The closer to the point the closer to the start it is. Smaller spikes between large ones represent large jumps (>5 million bp) on the chr.