**Introduction[[1]](#footnote-2)**

DNA is double stranded (a double helix) and made up of base PAIRS. Adenine on one strand (represented with an "A") always pairs with thymine (represented with a "T") on the other strand. These are called A~T pairs regardless of which strand has the "A" and which the "T." Similarly, cytosine on one strand (represented with a "C") always pairs with guanine (represented with a "G") on the other strand, creating G~C pairs.

Scientists often represent DNA strands with a string of letters like this:

ATATTTGAAAGCTGTGTCTGTAAACTGATGGCTAACAAAACTAG

This string of letters represents only one strand, or one half of the DNA molecule. There is no need to write down the other strand because as we just described above, a "G" in one strand means there is automatically a "C" in the other strand, just as a "C" in one strand implies that the other contains a "G".

Now think of the human genome and all of the genes in it as a VERY large set of blueprints. Each blueprint is an instruction set for assembling one part or piece of a cell. Almost every cell in your body carries the same set of blueprints -- so what makes a cell in your brain different from a cell in your stomach? A neuron and a stomach lining cell are very different in their morphologies (how they look) and their functions (what job they do). The different shapes and functions are a result of the fact that those two cells use different portions of the complete blueprint set to construct themselves. The neuron uses the blueprints for parts involved in brain signaling, while the stomach cell does not. The stomach cell makes parts for secreting stomach enzymes to help in food digestion, while the brain cell does not. These kinds of blueprints (or genes) are often called "tissue specific" because they are used in some body parts and not in others. However, there are also some blueprints that are used in every cell in the body because the parts they represent are needed in every cell (like pieces used during cell division or making energy). These kinds of blueprints (or genes) are often called "housekeeping" because they represent a basic need of every cell and they "keep up" the basic functions of the cell.

How does the cell know which blueprints to use? Each gene (or blueprint) is has its own control panel that acts as a group of switches affecting when (during an organism's development), where (in the body) and how much a particular blueprint is used. Scientists are still working hard at being able to identify all of the pieces of a gene's control panel. One important part of the control panel that we know a lot about is the "promoter." The number and pattern of As, Ts, Gs and Cs in a promoter is important in determining whether the switch will act like a "housekeeping" switch or a "tissue specific" switch. As of today, scientists are just beginning to understand why this is true.

How to calculate the G~C pair content of a DNA sequence:

1. Count the total number of G's and C's. Try it for this sample:

ATATTTGAAAGCTGTGTCTGTAAACTGATGGCTAACAAAACTAG

You should get a total count of 15 (9 G's and 6 C's).

1. Now count the total number of letters (bases). You should get 44.
2. %G~C pair content = (Count of G's and C's / Total count of all bases)\*100  
     
   So, for our sample, the %G~C pair content = (15/44)\*100 or 34%

**Project Part I**

**%G~C Content Calculator Program**

First, you will develop a Java program that computes the %G~C Content of a DNA sequence.

Your program should:

1. Prompt the user to enter a DNA sequence and read it in as a String.
2. Count the total count of all bases (As, Ts, Gs, and Cs). If the program detects characters other than A, T, G, and C, it will notify you, and will not count those characters in the total.
3. Count the total number of As, Ts, Gs, and Cs.
4. Compute the %G~C content.
5. Display the result for step 2, 3, and 4.

**Project Part II**

**Data Analysis**

In the second part of the project, you will calculate the %G~C content for each sequence in the given text files. Each text file contains a sequence of 100,000. This sequence is broken into sequences of 1000 per line. Keep in mind these are partial DNA sequences for three housekeeping promoters and three tissue specific promoters and actual DNA sequences are 100 orders of magnitude larger. You should record the data in an Excel spreadsheet that has three columns: the type of gene, the sequence name, and the %G~C content. Once you are done with recording the data, create a bar chart of the data.

**Housekeeping Promoters:**

1. Heat Shock Protein 90 (HSP90): When proteins get over-heated, their folding and conformation gets messed up which often affects their function. Heat Shock Proteins repair the unfolded proteins back to their working state.

HSP90.txt

1. Glucose-6-phosphate Dehydrogenase (G6PD): This molecule is a member of a team that helps protect each cell from agents that damage important proteins.

G6PD.txt

1. Beta-actin (ACTB): Actin proteins help the cell make an internal "skeleton" that maintains the cell's proper shape.

ACTB.txt

**Tissue Specific Promoters:**

1. Bone Morphogenetic Protein 5 (BMP5): Bone morphogenetic proteins help induce the growth of new bone.

BMP5.txt

1. Hemoglobin Beta (HBB): Part of hemoglobin which carries iron molecules in blood cells.

HBB.txt

1. GABA Receptor A1 (GABRA1): An important receptor of chemical signals that travel only in the brain.

GABRA1.txt

**Submission**

Your submission for the project should be a zipped folder that includes:

1. The %G~C Content CalculatorJava program (.java file)
2. An Excel spreadsheet that records the data.
3. A word document that
   1. Briefly describes your program and it's purpose
   2. Includes the bar chart from part II: Data Analysis
   3. Answers this question: Does housekeeping promoters have lower, higher, or similar G~C content than tissue specific genes?

1. This project and material are derived from sciencebuddies.org [↑](#footnote-ref-2)