Name: Jon Abrahamson
ID: 107084898

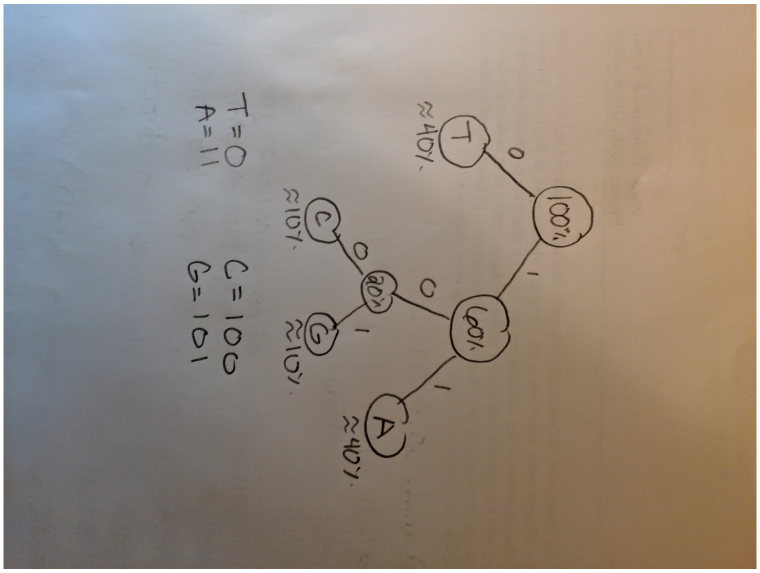
CSCI 3104, Algorithms **Profs. Grochow & Layer**
Explain-It-Back 5 **Spring 2019, CU-Boulder**

At a recent infections disease seminar, you hear about how the speaker is sequencing millions of *Plasmodium falciparum* genomes (the human malaria parasite) in order to better characterize how patients respond to treatment. In the presentation, the speaker complained to the audience that although they are making the data sets a small as possible by only using 2 bits to encode the 4 nucleotides of the genomes their IT department is still struggling to store all the data. Later in the presentation, you learn that the Plasmodium falciparum genome is "AT-rich." That is, over 80% of the nucleotides in the genome are either A or T. Please help this team understand how they can leverage Plasmodium falciparums AT-richness to help their IT department deal with the influx of data.

Name: Jon Abrahamson
ID: 107084898

CSCI 3104, Algorithms **Profs. Grochow & Layer**
Explain-It-Back 5 **Spring 2019, CU-Boulder**

By creating an optimal Huffman code and the corresponding encoding tree, we can reassign the coded bits to represent each of the nucleotides. We want the least frequent nucleotides to have the the most bits to encode while at the same time having the more frequently occurring nucleotides have the minimum amount of bits to encode. We are told that the genome is AT rich, meaning that 80 percent of the nucleotides in the genome are either an A or a T. With this knowledge we can try to re encode the 4 nucleotides so that we can have the minimum amount of bits to encode the A or T nucleotide. Although the scientists have already been able to maintain a small data set encoding the A and T with 2 bits each, we can make it so one of these two nucleotides can be encoded by a single bit and the other remain coded by 2 bits. Attached below is the corresponding Huffman tree and each genomes corresponding encoding. You can see that we have decreased one of the "rich" nucleotides (A or T) to one bit and kept the other at two. Additionally, we have increased the number of bits of the other 2 nucleotides (C and G) from two bits to three bits. While this makes encoding the C and G nucleotides cost more data storage, due to the the fact that the A and T nucleotides occur far more frequently, the algorithm will actually save data. For example if we have genome ATCTTATTAG (80 percent of the genome contains A or T), then under the new system there would be a total of 17 bits (T=1 bit, A=2 bits, C=3 bits, and G=3 bits) and under the old system there would be a total of 20 bits (2 bits each times 10 nucleotides). As the genomes get larger, the more data that is saved. So, when these scientists are sequencing millions of very large genomes, this small change to the encoding of the nucleotides results in significant saving of data storage.



100%
0
T
≈40%
1
60%
0
20%
0
C
≈10%
1
G
≈10%
1
A
≈40%
T=0
A=11
C=100
G=101