Title and author(s). Genomic Similarities between Species: CSE 140 HW 9

Garrett Quesnell and Lauren Glass

Summary of research questions.

1. How similar are human mitochondrial DNA sequences to other species' mitochondrial DNA sequences?
    a. We are trying to determine if humans have common mitochondrial DNA with other species. This will indicate if humans had a close common ancestor with that organism.
2. How similar are the proteins expressed in human mitochondrial DNA with other organisms?
    a. The proteins in mitochondrial DNA that are expressed show the behavior of the cell from the given genetic code. We want to examine how the behavior of cells is similar across species.
3. What is GC content of each species mitochondrial DNA?
    a. In biological theory, GC content indicates DNA with a high GC content is more stable than DNA with a low GC content. High GC content DNA is used more often in biological processes than low GC content DNA.
4. What is the GC content of each human chromosome?
    a. The GC content of human chromosomes might indicate how often they are translated in biological processes.

Motivation and background.

Genome analysis matters because it shows us how similar the genetic makeup of one organism is to another. Mitochondrial DNA comes from the section of a cell called the mitochondria that is responsible for converting energy into nutrition for a cell. In humans, the mitochondrial DNA is the smallest chromosome with only 37 genes and approximately 16,600 base pairs. It is the first part of the human genome to exist. The study of mitochondrial DNA has applications for medical purposes, evolutionary theory, anthropological value, and general scientific purposes.

The motivation behind this project was to work with a smaller version of an organism's genome to explore the basic types of genomic analysis being done in laboratories. We were inspired by a recent scientific accomplishment featured in the news. Geneticists were able to trace back a deceased South Carolinan man's Y chromosome to an ancestor approximately 338,000 years old. This new date of the origin of modern humans is thousands of years older than the previous

estimates. We wanted to investigate the human genetic evolutionary track ourselves. We decided to compare organisms mitochondrial DNA because it gives unique insight into the evolutionary process and is reasonably sized data for our project to utilize.

Another motivation is to make large pieces of genomes easy to process. The program will be generic enough to process DNA anywhere from 10 bases all the way to a billion bases dependent on memory and processing time. It can be used in more complicated research projects having to do with genome sequencing. It could also be used for processing small pieces of DNA for gene splicing.

Dataset.

The dataset we will use is from 'http://genome.ucsc.edu/' in the form of text documents with letters representing each base of genomic DNA. Most of our analysis will use mitochondrial DNA, which exists for the following organisms: human, chimpanzee, orangutan, dog, horse, yeast, chicken, cow, mouse, rat, zebrafish, finch, pufferfish, pig, medaka fish, D. simulans (a type of fly), C. elegans (a nematode worm), and opossum. Part of our program downloads this data from an FTP server and converts it into a usable representation depending on what organism the user wants to analyze.

Methodology (algorithm or analysis).

1st research question: How similar are human mitochondrial DNA sequences to other species' mitochondrial DNA sequences?

1. Read mitochondrial DNA data for two organisms.
2. Take DNA sequence and compare it base for base against the other DNA sequence.
3. If the DNA is not aligned, check to see if there were mutations by shifting the DNA forward or backward up to two bases away.
4. Count how many bases are similar.
5. Divide total count by the length of the DNA sequence to get a percentage.

2nd research question: How similar are the proteins expressed in human mitochondrial DNA with other organisms?

1. Convert DNA to mRNA by finding the reverse transcript of the DNA, exchanging T's for U's.
2. Convert mRNA to proteins by looking for start codon "AUG" and translating until a stop codon "UAG", "UGA", or "UAA".

3. Add these amino acid sequences to a set.
4. Find percentage similarity between species.

3rd research question: What is GC content of each species mitochondrial DNA?

1. Read mitochondrial DNA data.
2. Count all G's and C's in genome.
3. Divide this count by the total number of G's, C's, A's, and T's in the sequence.

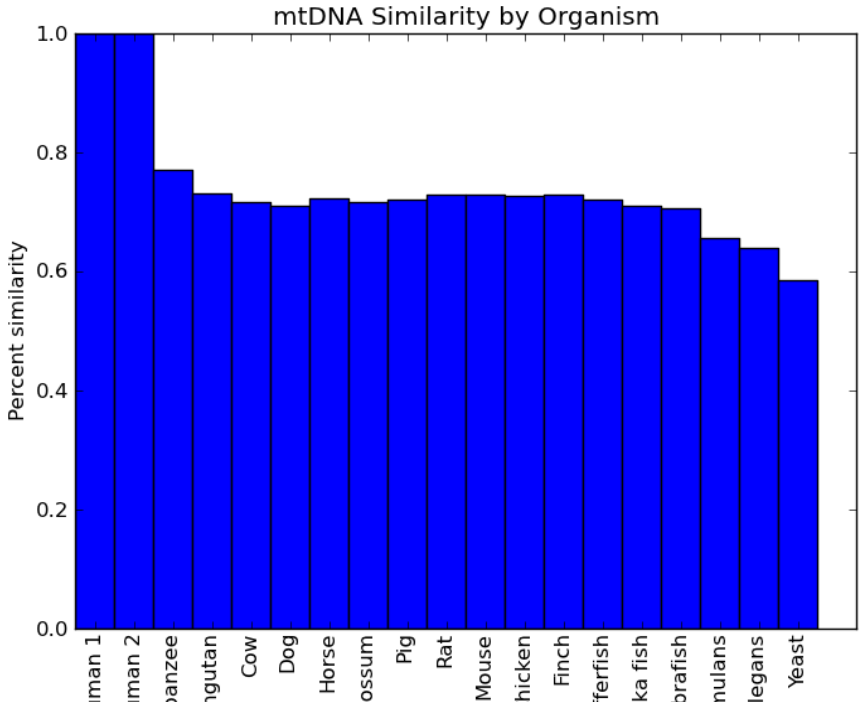
4th research question: What is the GC content of each human chromosome?

1. Read chromosome DNA data.
2. Count all G's and C's in genome.
3. Divide this count by the total number of G's, C's, A's, and T's in the sequence.

Results.

1. How similar are human mitochondrial DNA sequences to other species' mitochondrial DNA sequences?

   Similarity ranking by mtDNA sequence: ['Human 2', 'Chimpanzee', 'Orangutan', 'Mouse', 'Finch', 'Rat', 'Chicken', 'Horse', 'Pufferfish', 'Pig', 'Opossum', 'Cow', 'Dog', 'Medaka fish', 'Zebrafish', 'D. simulans', 'C. elegans', 'Yeast']

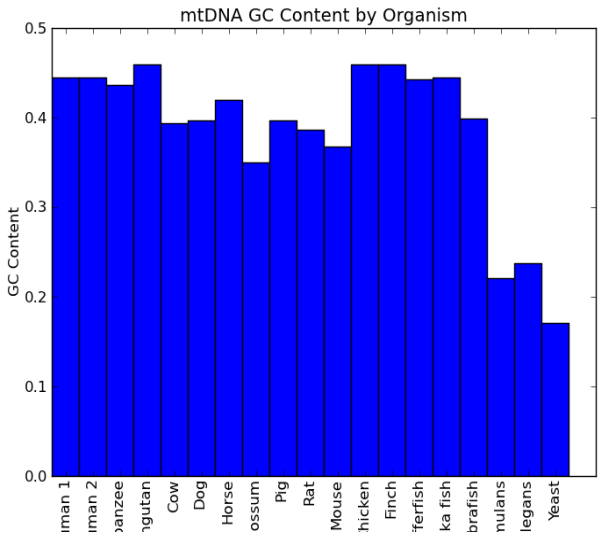


2. How similar are the proteins expressed in human mitochondrial DNA with other organisms?

   The mitochondria from all species uses exactly the same proteins as the human 1 mitochondria.

3. What is GC content of each species mitochondrial DNA?

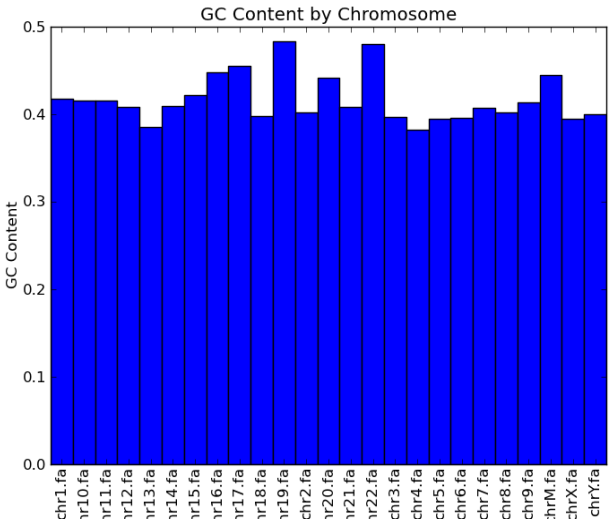
   Human 2's GC content: 0.44490707217, Chimpanzee's GC content: 0.436907278768, Orangutan's GC content: 0.459393939394, Cow's GC content: 0.394210171981, Dog's GC content: 0.396580583453, Horse's GC content: 0.419602664906, Opossum's GC content: 0.35, Pig's GC content: 0.396696678791, Rat's GC content: 0.386970124532, Mouse's GC content: 0.367484662577, Chicken's GC content: 0.459644730567, Finch's GC content: 0.459416162335, Pufferfish's GC content: 0.442424610895, Medaka fish's GC content: 0.444690397846, Zebrafish's GC content: 0.39934927999, D. simulans' GC content: 0.220797435384, C. elengans' GC content: 0.237839797028, Yeast's GC content: 0.171100489625



4. What is the GC content of each human chromosome?

   chr1.fa GC content: 0.41743925494
   chr2.fa GC content: 0.402437822769
   chr3.fa GC content: 0.396942735339
   chr4.fa GC content: 0.382478885127
   chr5.fa GC content: 0.395162873815
   chr6.fa GC content: 0.396109109954
   chr7.fa GC content: 0.407513079318
   chr8.fa GC content: 0.401756859767
   chr9.fa GC content: 0.413168420226
   chr10.fa GC content: 0.4158487647

chr11.fa GC content: 0.415656506994
chr12.fa GC content: 0.408119842742
chr13.fa GC content: 0.385265421248
chr14.fa GC content: 0.408871533266
chr15.fa GC content: 0.422009527244
chr16.fa GC content: 0.447894266109
chr17.fa GC content: 0.455404600677
chr18.fa GC content: 0.397849705916
chr19.fa GC content: 0.483603159663
chr20.fa GC content: 0.441257223847
chr21.fa GC content: 0.4083253987
chr22.fa GC content: 0.479880724054
chrX.fa GC content: 0.39496335821
chrY.fa GC content: 0.399650465762
chrM.fa GC content: 0.44490707217



Based on our program, we found that human mitochondrial DNA differs substantially from the other organisms. However, we found that the proteins that make up the mitochondrial DNA are the same. This shows that although the species has significant differences, the behavior of their cells is essentially the same (based on mitochondrial DNA).

The GC content of more evolved species like humans is higher than the GC content of less evolved species like yeast. Biologically this may indicate that the DNA of the more evolved species with the higher GC content is more stable. This would make sense because the species with higher GC content reproduce less quickly and have longer lifespans.

The important conclusion to draw from our project is the similarities in mitochondrial DNA between species. We can begin to understand how the human evolutionary tree developed based on the percentage similarities in DNA sequences. For instance, it makes sense that the species with the highest sequence similarities to humans are primates and other vertebrates. The results give an ordered list of most closely related species to least closely related species to humans. One discrepancy is that the Finch and Chicken similarity rankings seem to be high based on phylogenetic trees. We would expect the Finch and Chicken to be less related than all the mammals. This inconsistency could be attributed to a poor similarity algorithm or common divergent evolution. Overall, our program indicates that somewhere along the evolutionary track, all life is related.

Reproducing your results.

In order to obtain the data and run the analysis, one must only run the python program named “main.py” through the python interpreter. The prompt will ask you for a valid email in order to access the online database. The program is designed to download available mitochondrial DNA for certain organisms we picked to compare.

The program will then print the results for each of the following species: human 2, chimpanzee, orangutan, cow, dog, horse, opossum, pig, rat, mouse, chicken, finch, pufferfish, medaka fish,

zebrafish, d. simulans, c. elegans, and yeast. The results will include the expressed protein similarity, the mitochondrial DNA sequence similarity, and the GC content.
Next the program will download the human 1 chromosome files to calculate their GC content. This analysis takes up a decent amount of memory. If the computer running the analysis has the available memory, the program will then print the GC content of the whole human chromosome. The program will also display all results graphically.

Collaboration: We did not receive help from outside of our group.

Reflection: We both enjoyed being able to work on a project that we devised on our own. It made us use all of the topics from class and apply them in a way that made sense to us.

Sources:
http://en.wikipedia.org/wiki/Mitochondrial_DNA
http://www.cell.com/AJHG/abstract/S0002-9297(13)00073-6