MTH 499 CSUMS Final Report: Comparison of Two DNA Sequences

Laura Jean Champagne
Undergraduate Studet
Dept. of Mathematics
UMass Dartmouth
Dartmouth MA 02747
Email: Laura.Champagne@gmail.com

May 10, 2011

Abstract

This work compares the statistical differences of two genes in the human genome, Amelogenin, X isoform (AMELX) and Amelogenin, Y isoform (AMELX). Primarily using MATLAB, the focus was on learning the software's capabilities and using its features to compare AMELX and AMELY, possibly identifying any unique patterns to either, or both, genes. Basic statistical analysis of the data sets showed slight differences for the base pair sequences and the amino acid sequences of the two genes. A weak test for randomness was also applied to each set of data, results showing weakly non-random for both sets of data.

1 Original Abstract, February 17, 2011

This project examines the randomization of DNA sequences based off of current genome sequencing projects. Using various computational programs such as R, MATLAB, Mathematica and/or Python, different mathematical tests for randomness will be applied to DNA sequences in an attempt to find patterns or other unique properties. DNA sequences used are obtained from the Virtual Library of Genetics.

2 Background Information

This work compares two genes, AMELY and AMELX. This gene codes for tooth enamel and is gender linked. In human females, only the AMELX (Amelogenin, X chromosome) is present and active. Males, although they have both a Y and an X chromosome only display AMELY. Because of this gene's unique

gender-dependence trait, I was interested in examining the statistical differences between the two genes. I was originally looking to find any sort of pattern to dispel randomness in the gene's sequence but first needed to understand the basic mechanisms of MATLAB in order to use it's more complicated and advanced features.

2.1 Base Pairs and Amino Acids

The base pair sequence for each gene was taken from GenBank. AMELX is approximately 7900 base pairs and AMELY just over 8100. The sequence is composed of long seemingly random orders of adenine, A; cytosine, C; guanine, G; and thymine, T, which were converted into 1, 2, 3 and 4, respectively, before any mathematical analysis was started.

When the gene is transcribed into a protein, every third set of base pairs is transcribed into an amino acid. These sets of three base pairs, or codons, are more significant as the corresponding amino acid. There are 4^3, or 64 possible combinations of codons, yet there are only 20 amino acids. Some amino acids, such as Leucine, are coded by as many as 6 codons, others, such as Methionine, are only coded for by one unique codon. This initially presented a challenge in transcribing the base sequence into the more significant sequence of amino acids.

3 Methodology

3.1 Original Goals

At first, I wanted to mimic research done by Gary Davis, PhD on Microplasma genetalium, a bacterium with one of the smallest known genomes. In his research, Gary Davis discovered a relationship between every 3rd and every 4th amino acid. My first thought was to write a similar program to see if there was a similar relationship in my data. However, although I started out with great aspirations, I quickly realized my limitations. This course is the first course I've ever had with MATLAB, or any type of programming for that matter. I was lucky enough to get my hands on a MATLAB Primer, a step-by-step guide to the program. Once I reconciled with myself that I needed to take baby steps, I worked through the book's first four chapters, on basic functions MATLAB can do. Once I was more confident with MATLAB's syntax and basic commands, I skipped ahead to the chapters on two dimensional plots. By this point, I could see some of the logic in the syntax. It took me over a month to get comfortable with basic MATLAB, but I'm really glad I did this. Until this point in my college career, I had viciously avoided any type of programming and computer-based mathematics. This was a missing link in my education. I have a strong background in Biology, Chemistry and Mathematics, yet I didn't have the link to use the Mathematical software to analyze scientific work. Looking back, all those times I used Excel for lab reports I could have easily used MATLAB, or

something similar, to explain and graphically show my results. A command of MATLAB would have also helped me to deepen my explanation of my results, as I have the Math background to explain and understand many more analytical processes than I used.

3.2 Formatting

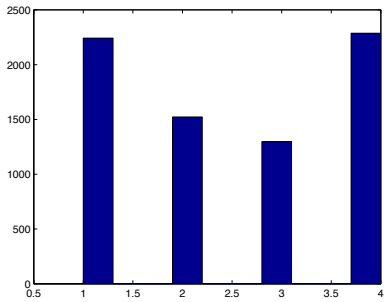
The first step I took was to alter the sequences from the letters A, C, G, and T to the numbers 1, 2, 3 and 4, respectively. This, I did in Word by using the command Find and Replace. I then used Mathematica's partition command to partition the sequence into sets of three, not overlapping numbers to represent the codons. After unsuccessfully trying to write a program to convert the codons into the appropriate amino acids, I discovered that MATLAB has a significant amount of preinstalled functions for biological studies. One of those functions, nt2aa, transcribes a sequence of nucleotides to amino acids with ease, even recognizing the stop codon and mutations within the code that make for erroneous results.

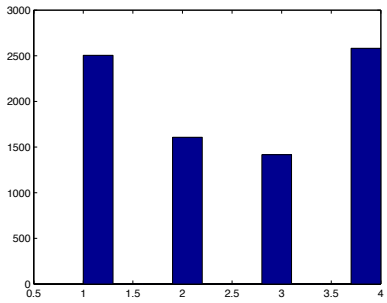
For some unexplainable reason, I didn't check to see if MATLAB has a corresponding feature to translate the amino acids into integers before then returning into Word to do Find and Replace. I painstakingly switched all 20 amino acid letters into the numbers 1-20 for both AMELX and AMELY. Once I had moved on and played with my data a bit, I realized I had edited the same set of data, AMELX, twice, and saved it as both AMELX and AMELY. Thankfully, once I noticed my mistake I also did a bit more research and found aa2int, which converts the nucleotide sequence into numbers. I stepped back at this point and retraced a lot of my steps, being much more careful to organize and properly label my data. Although frustrating, it was so much easier the second time around, as I had then spent the time with the MATLAB Primer and understood the syntax of MATLAB much better.

4 Numerical Results

4.1 Histogram of A, C, G, T

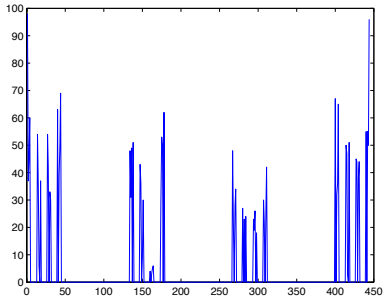
These first two histograms are a simple comparison between the total base pairs of AMELX (fig. 1) and AMELY (fig. 2). In this graph, the x-axis represents the bases adenine, cytosine, guanine and thymine as 1, 2, 3 and 4, respectively. The AMELY gene has an extra 1,000 base pairs, which can be noticed by looking closely at the variation in the y-axis. Otherwise, the graphs are very similar. There is a significant discrepancy in the quantity of each base pair for both genes. It is expected that a fully random sample would have a near equal amount of each base in each gene.

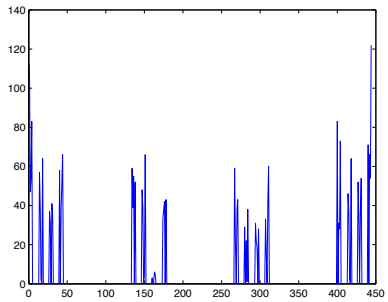




4.2 Histogram of codons

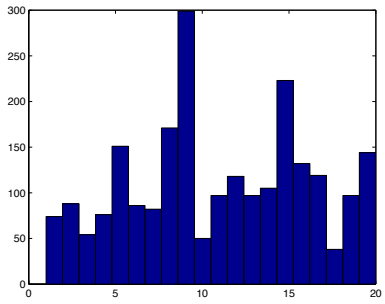
The second comparison I did was to use Mathematica to partition the data into non-overlapping sets of three. This was the first step in working on the amino acids. As the x-axis shows, the data sets range in triplets from 111 to 444, only using the numerals 1, 2, 3 and 4. These graphs are essentially useless because they don't truly represent the amino acid building blocks of the resulting proteins. Certain sets of codons represent the same amino acid.

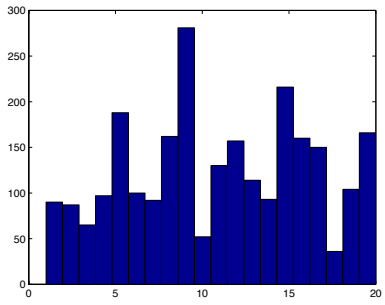




4.3 Histogram of Amino Acids

After using MATLAB to transcribe the nucleotide base pairs into the correct amino acid sequence, I was able to edit the data and produce these histograms. I had a bit of trouble editing the data correctly; my first two graphs of the amino acids were identical because I had mixed up my data and used the AMELY data twice. Although MATLAB has a function to change amino acid sequences into numerals, I still had to further edit that data (syntax) to use it in the histogram function. Both genes have a high quanitity of leucine (#9). Although they have similar shapes, there is a bit of difference in the relative quantities of asparagine (#11) and glutamine (#13).

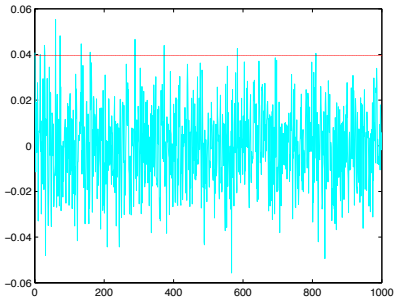


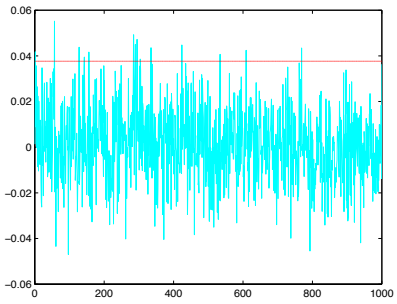


4.4 Correlagram

X-corr.eps Y-corr.eps This program was adapted from an original program written by Gary Davis as a weak test for randomness. This program systematically goes through the long data set and compares a segment of the sequence to the rest of the sequence, sliding over base by base. It produces a plot of the difference in the data relevant to the variance of the data. A straight line on the graph signifies the standard error of the data.

Both the AMELX and AMELY correlagram graphs show slight non-randomness. Had the majority of the blue line been above the standard error line (red), it would have indicated a significantly non-random data set. It is important to remember that the correlagram program was run individually for the AMELY and AMELX data sets; they were not compared to each other with this program.





5 The Experience

CSUMS was a great class to take. Although my work was not extremely significant mathematically, it was an incredible learning experience for me. I have avoided both programming and independent research for my entire college career. As I wrote under my Original Goals, I now have a bit of confidence in using tools such as MATLAB and can see many ways in which to employ it. I'm entering a PhD program in the fall and am certain I would use MATLAB over Excel to analyze research results. MATLAB still has a bit of a learning

curve for me, but I can see how the benefits of working with it far outweigh the challenges. I am at a point with my research that if I were able to continue with it, I feel I would make significant progress. I spent the first month of the semester finding a project. Once I settled on comparing the two genes in the eyes of randomness, I had a wealth of great ideas. It took me a bit of frustration with MATLAB before I realized I needed to step back and first learn the basics. After I took the time to learn the fundamentals of MATLAB, I was able to manipulate and work with my data with ease.

CSUMS provided me with my first opportunity to attend and participate in a research conference. Although my individual experience at UMass Amherst for the Statewide Undergraduate Research Conference could have been better, the overall opportunity and experience as valuable. I was also very grateful for the various opportunities throughout the semester such as speakers and local conferences. This semester, CSUMS and Statistics (which was taught by my advisor in a similar way), made me realize the potential in our Mathematics department. In my struggles with MATLAB I reached out to a few students for tutoring and I was amazed by the support I received. Although I felt my work was elementary to some of the other projects in the class, people never made me feel insignificant or stupid. I appreciate the camaraderie in the class, between students and faculty.

6 Future Work

I plan to add my work to the blog I created for statistics, laurajeanstats.wordpress.com. I graduate and will not be able to enroll in CSUMS to continue my work, but perhaps another student would want to. As I learned very quickly, there's no need to reinvent the wheel. What I have done this semester may be of interest and help to someone else

7 Appendix

Include code for correlagram?

8 References

Davis, G. (2/2007) Ultra Weak Deterministic structure in the chromosome of Mycoplasma genetalium [PowerPoint Slides] retrieved from author via email

The GenBank NCBI Reference Sequence of the AMELY gene I reference for this paper is: $NC_000024.9$

The GenBank NCBI Reference Sequence of the AMELX gene I reference for this paper is: $NC_000023.10$

Sigmon, K. & Davis, T., MATLAB Primer, 7th ed. 2004. CRC Pres.

9 Acknowledgements

I thank Gary Davis, my advisor, for his sense of humor and his patience. While I was busy feeling as though I was floundering, he was supporting me and turning me to the right people for tutoring and support. I also thank Sigal Gottleib for her open mind and supportive attitude. I also thank NSF for the funding which defrayed the lost work hours for the UMass Amherst conference and the hours I left work to come to tutoring. This was an incredible opportunity, and I'm very appreciative to everyone who works hard to make it a reality for us.