Integrated Bioinformatics

Exam I

Skyler Alexander Kuhn

BNFO 601: Exam 1

Virginia Commonwealth University

Dr. Paul Fawcett and Dr. Jeff Elhai

October 1, 2016

On my honor I have neither given nor received help on this assignment in accordance with the VCU Honor Policy, and I have read and reviewed the Honor Policy prior to attempting submitting this report.

2b) Are there particular codons for which there were large or perhaps unexpected differences between the frequencies from simulated and actual PCC7120 DNA? If so, describe for which codons you see major discrepancies (consider more than just start and stop codons here folks!). In particular, which amino acids did these encode? What are your thoughts about what might underlie the discrepancy?

Yes, there were several large differences between of the codon frequencies of the simulated and actual PCC1720 DNA. Based off of the algorithm that I employed and the given GC content of 42%, we would expect to produce a sequence with the following relative nucleotide fractions: G = 0.21, C = 0.21, A = 0.29, and T = 0.29. That being said, if our sequence is 1000 nucleotides long (we did not do this), it will contain 160 more A and T nucleotides than G and C nucleotides. Again, this is assuming that the pseudorandom generator that is imported from the random module is functional. But we have a pretty large sample size so due to the central limit theorem we can assume normal variance, right? Anyways, let's forget about Statistics for a moment. The largest difference can be seen in codons that encode for the stop codons. That is to be expected seeing as how I did not program my bias_analyzer to inhibit that codon's generation within the sequence (you said it was okay if we do this). The next highest difference was observed in the codon GAA (E, Diff: -28.3), which was significantly underrepresented in our simulated DNA. Glutamic acid is an amino acid which has charged side-chains. This amino acid is also very charged and hydrophilic. It can typically be seen facing the outside in an aqueous environment. That be being said, having the correct ratio of this specific codon could play a crucial role in the development of this protein's overall structure. The next highest difference was observed in the codon CAA (Q, -22.1). Like Glutamic acid, Glutamine is polar and also hydrophobic. But unlike Glutamic acid, Glutamine has polar side chains. Again, this codon is unrepresented; in the actual DNA sequence having the correct ratio of this specific codon could play a crucial role in the development of this protein's overall structure and in the development of transmembrane proteins (just a thought). As far as why is this codon more heavily expressed than say "CAG" (the other synonymous codon), it may have to do with differential expression of the tRNA that actually pairs to codon or CAA may infer increased translational efficiency. After generations of codon optimization this codon may increase the organism's fitness.

2c) Do you see a relationship in the real world data between those amino acids that have the most synonymous codons (six each in the case of arginine, serine, and leucine) and how often those amino acids appear in actual proteins? How does this compare to what you observe for your simulated proteins?

Yes, there does seem to be some relationship between increasing the number of synonymous codons and how often those amino acids appear in the actual proteins. As the number of synonymous codons increases the observed difference seems to decrease. In what we observed in our simulated DNA, I believe this is due more to permutations and combinatorics more than anything else. As the number of synonymous codons increases, the overall *weight* of each synonymous decreases (weight being a metric of its cumulative summation).

2d) Are there any commonalities or emerging themes evident among particular over or under represented codons that you observe? If so speculate as to cause.

Please see question "2b" for my thoughts on commonalities among underrepresented codons. As far as overrepresented codons, both of the codons that encode for the protein "C" or Cysteine occur at higher than actual levels. Paired cysteines allow for the formation of disulfide bridges. That being said, the relative abundance of this specific protein may be heavily influenced optimization related to protein folding. Having too many Cysteines could be deleterious to the process protein folding and the resulting protein's function. For the most part, there tends to be more/less overrepresented codons that have side chains that are either polar/nonpolar.

2e) Does (and should) bulk GG% percentage alone explain all of the codon bias you observe in your simulated sequence (you can arrive at this result logically)? How about in the case of the actual Anabaena protein data? Explain your reasoning for this answer.

It should be noted that the codon frequencies that were generated would-- more than likely-- be ever so slightly closer to the actual values if we did not allow multiple stop codons (per sequence).

Yes, the GC% that we defined does explain the codon bias that was observed in our simulated sequences (that and technically python's random number generation). That is, again, what was creating our sequences after all. But it should also be noted that there are no outside forces such as natural selection for translational optimization, thermodynamics, differential tRNA expression, or quantum mechanics that could have influence that generation of this sequence. All of points mentioned above could technically influence the observed codon frequencies. Using GC content to generate accurate coding regions is not the best approach. If we were generating random intergenic regions within a eukaryotic genome that is one thing, but using this approach to generate sequences within prokaryotic coding regions is poor at best. That being said, the selective pressures associated with ensuring coding regions encode for a functional product are too high to just ignore (or not include in our simulation).

2f) In actual prokaryotic DNA, is GC% uniform at each codon position, or does it vary? How about in your simulated DNA? If there is a difference, which of the codon positions would expect would most closely follow the overall GC% and why? What feature of the genetic code might help explain this observation? How might position specificity of GC% have affected your simulation outcomes? As always in biology, it's useful to couch your answers in terms of selection.

Our simulation seems to refutes the notion that in actual prokaryotic DNA the GC content is **not** uniform at each codon position. It varies. The resulting differences that were observed in the simulated DNA versus the actual DNA further support this. Codon composition and the order of the nucleotide triplet making up the codon supersedes the overall generalization brought about by our rudimentary efforts to simulate DNA through GC content alone. And to say the least, it is not a sufficient method to accurately replicate coding regions of DNA. In the simulated DNA, GC content would be preserved at each codon position. That is just due to the algorithm that was employed to generate the random sequence (as each nucleotide was added to the growing string, it had a equiprobable chance of being added at each codon position).

It would be expected that the that the first two codons would be the most conserved due to thermodynamic stability when bonding to a tRNA (with the the third being most volatile). Francis Crick's Wobble Hypothesis touches on this idea. He stated that the first two bases in the codon create the coding specificity. And as he stated, this is because they typically form Watson-Crick pairs that bond strongly to the anticodon of the tRNA. Several studies have shown the importance of Wobble base pairing in the formation of functional proteins. Because the first two nucleotides in a codon hold more specificity, it would have been better if the GC content at those two positions were weighted more favorably than the last position. This is something that we did not employ in our random sequence generator.

Again, the algorithm that we employed to generate a random sequence of a coding region was naive at best. Using GC content alone to replicated coding regions in this organism produced observed differences that would not be acceptable for further testing. As previously discussed, GC content is not equiprobably distributed at each codon position. This makes sense for several reasons, and it is partly due to the fact that the first two codons need to bond strongly to confer the proper stability when bonding to its tRNA. That is not say GC content does not have its place. It is useful in metagenomic studies to determine what is in a sample, but for generating random coding regions, it fails to provide the specificity that is necessary to do so. Again, there are generations upon generation upon generations of selective pressures that have resulted in the actual sequence in the organism, and the GC content at each codon position reflects this. I would actually be pretty impressed if our attempts resulted in a functional protein.

Also a lot of bacteria-- especially bacteria that have short generation times (30 mins)-- have streamlined and optimized genomes. It is energetically costly to maintain anything that is unnecessary. Certain codons may be favored over others due to translational optimization that has, again, occurred over millions of generations. Translational optimization leads to faster translation rate and increased accuracy, which again leads to increased organismal fitness. There are no selective pressures in our simulation... ***And with that, if we are to create a truly accurate simulation, we must first create the universe!***