18 August 18th

18.1 Goals for the day

Goals from Last Time

1. *Analyze my.cubappr.r*
   (a) *Add in print statements, then rerun the example*
   (b) *my.drawBConditionalAll.??????*
2. *(Optional) Study R user manual more?*

Additional Goals

3. **Sign up and connect GitHub** (TOP PRIORITY)
4. LaTeX changes
5. **Read Gilchrist 2014 paper** (HIGH PRIORITY)
6. Look at REU Results (a1 is questionable?)
7. Workflow Tracker (like doxygen) for R?

18.2 Progress/Notes

18.2.1 Analyze my.cubappr.r

18.2.2 Study R user manual

18.2.3 Connect GitHub

Connected. Username ozway. Forked cubfits. Added "worklog" directory with August.

1. cd /worklog
2. git commit -a
3. git push

18.2.4 LaTeX changes

Added augustweek.tex, a .tex layout for analyzing the week's work. Not sure how to number it, for now I'll do august18th-22nd.tex, section 1. For September, I may replace the monthly summary august.tex with all of the weekly summaries. That will be more concise. I can't see any reason to do a big summary of the whole month, it's likely better to separate out the weeks.

18.2.5 Read Gilchrist 2014

Strangeish. "Researchers strongly believe that genomic sequences encode a trove of biologically important information." I thought it was a given. Is there a faction of biologists who don't believe in DNA?

A question strikes me: What do we already know about CUB, and what are we going to learn about CUB?

"...highly expressed genes should show the strongest bias for codons with shorter pausing times and error rates [Ikemura, 1981b, 1985, Sharp and Li, 1986, 1987a]. As a result, the patterns of CUB observed within a genome should contain a significant amount of information about a gene's expression level, specifically the average rate at which proteins are synthesized from the ORF. Further, because low expression genes are under very weak selection to reduce η, their patterns of CUB should provide information on the mutational biases experienced within a genome."

~~As I understand it, we have one direction of data? We have data about gene expression levels, and want to produce information about mutation bias and pausing times.~~ We have BOTH directions of information.

"Using the Saccharomyces cerevisiae S288c genome as an example, we demonstrate that our model can be used to accurately estimate differences in codon specific mutation biases and contributions to η without the need for gene expression data."

But do we also have information about pausing times, and we want to find out their expression levels? What would be the model where we have no information about gene expression levels? Is it just trying to work backwards from the results of the first model?

Answer: We have both, and can use the program to solve for either of them. With or without $\vec{X}$ (the ϕ values), there are (apparently) reliable results for mutation bias and ROC.

Questions:

- What is $\vec{X}$? Based on the reading, $\vec{X}$ is a set of estimates for ϕ, but there is a sizable section about the estimation of ϕ given $\vec{X}$. Why isn't that estimation perfect? what am I missing here?

- What is the input to the code when given no $\vec{X}$ estimates? Does the code first approximate the protein synthesis rate (ala Figure 3(a))? Does it use information about the ROC or NSE?

- ϕ is more closely correlated with and without $vecX$ for higher values of ϕ. That seems promising? When actually applied to the genome, (say S.cerevisiae), the model will mainly be dealing with genes with higher expression levels (it's just math). Is this valid? Also, why does $\delta\eta$ seem to be LESS accurate at higher expression levels? It's not unsurprising, but it is a contrast.

- Does any of the documentation have things about the code structure? Dr. Gilchrist said it would be in the supplemental materials, but I still have no idea what B is, or drawBConditional

18.2.6 REU Results

18.2.7 Workflow for R

18.3 Future Goals

1. Continue analyzing my.cubappr.r
2. Look for Wei-Chen code on the NSE model
3. Look at REU results (esp. a1 discussion?)
4. Look into workflow programs for R
5. Read about Data Structures in R user manual