Research Journal

Kirolos A. Shahat
kshahat@vols.utk.edu

Beginning 19 May 2017

Contents

Friday, May 19 2017

1 Goals for today

- Get LaTeX up and running to begin taking notes
- Study 2007 article and use it to begin learning terminology and understand key concepts
- Study Jeremy Rogers' and Alan Dixon's previous notes and see how they began their research

2 Current Progress and Notes

- Beginning to understand layout and format of LaTeX files.
- Terminology:
  1. Codon - Sequence of three nucleotides that together form a unit of genetic code in a DNA or RNA molecule.
  2. Codon Usage Bias (CUB) - Nonuniform usage of particular synonymous codons within a genetic sequence.
  3. Genome - The genetic material of an organism.
  4. Stochastic Model - A model that allows for random variation of one or more inputs over time.
  5. Stochastic Evolutionary Model of a Protein's Production Rate (SEMPPR) - A way to link CUB and the average protein production rate mechanistically. Essentially the model makes inferences about the production rate of a gene based on its elevation on the fitness landscape of protein production costs.
  6. Polypeptide - Linear amino-acid chain which forms most, or all, of a protein.
  7. Genetic Fitness (n) - The reproductive success of a genotype.
  8. Gene - Represented as a vector of Codons.
- Concepts:
  1. A major cost of a nonsense error is the amount of energy invested into assembling the incomplete polypeptide.

2. Selection on codon usage against nonsense errors should increase with codon position along a sequence because the cost is related to the length. This leads to the prediction of increasing codon bias with codon position.
3. Adaptation of a codon sequence, within SEMPPR, refers to the state of its expected cost of producing a protein relative to the minimal possible cost.
4. The resulting output from SEMPPR is a posterior probability distribution for the protein production rate of a gene based on its observed codon sequence.
5. Incomplete proteins are the result of nonsense errors. The cost of these nonsense errors is a function of their expected number and the length of the incomplete proteins.

- Notes for next time
  1. Figure out math and vector notation for LaTeX.
  2. Continue studying 2007 article
  3. Attempt to get to Jeremy and Alan's labbooks

Monday, May 22 2017

1 Goals for today

- Short review of what I learned on May 19th 2017
- Figure out how to get mathematical notation in LaTeX documents
- After speaking with Hollis, he suggested that I look at the existing FONSE code and cross reference that code with given articles and refer to the layout of the ROC model for implementation so that is where I'll focus my time now and address questions when needed in order of relative importance.

2 Current Progress and Notes

- Terminology:
  1. Elongation - The stepwise addition of amino acids to the growing protein chain.
- Questions:
  1. Not sure what the variables bias_csp or mutation_prior_sd are in the FONSEParameter. mutation_prior_sd is the likelihood that there is a mutation where it is defaulted to 0.35. I believe that the bias_csp is the codon bias based on current position.
  2. Unsure how those values are getting updated or where they show up in the $\eta(\vec{c})$ equations.
- Current Notes:
  1. Running runFONSEmodel.R and saving the output into a file to understand what is going on
  2. Beginning scan of FONSEParameter.h and FONSEParameter.cpp and trying to document where I can from current understanding
  3. Currently looking through the constructors in FONSEParameter.cpp and following the methods that are being called. Currently in initFromRestartFile which leads me to
  4. rep(x, y) returns a vector of size y with all elements as value x in R

5. sd = standard deviation, csp = codon specific parameter.

- Notes for next time:
  1. Ask Dr. Gilchrist what the variables in the FONSE equations represent so that I can understand how to get them/decipher them in code.
  2. Continue trying to decipher the FONSEParameter code and relating them to articles/equations. Mostly get help with initvalues methods and from there it shouldn't be too difficult to follow.

Wednesday, May 24 2017

1 Goals for today

- Continue running runFONSEmodel.R and saving the output to file
- Continue scanning through FONSEParameter.h and FONSEParameter.cpp and documenting where I can from current understanding
- If time allows, get the meanings of the variables from Dr. Gilchrist

2 Current Progress and Notes

- Ran through about 930 iterations of runFONSEmodel.R and stopped it. I have the outputs from that program to follow.
- Currently tracing through the FONSEParameter.cpp and it's leading me to initBaseValuesFromFile method so I am checking it out and documenting where I can.
- Spoke with Dr. Gilchrist and he confirmed that I should be going through the C++ code and trying to understand it. He also gave me an understanding of what the variables represent so that I can understand some of the math in the code when I come across it.
- Going through the init methods in FONSEParameter.cpp and I found that sequenceSummary is implemented using maps to emulate enumerators. I haven't seen what all it is being used for but I think enumerators are constant time when maps are log(n) time to find an element. I think it be a nice speedup if it was changed over. Just a thought for later, potentially.
- I'm adding comments and the code is pretty cluttered, my impulse is to start cleaning it and making it more legible but I'm refraining, for now...