Research Journal

Undergraduate Research Assistant Notes

Walker Bussey-Spencer
wbusseys@vols.utk.edu

Beginning 22 July 2016

Contents

22 July 2016

1 C. Elegans Life Stages

Walker Bussey-Spencer

- Looking at data with original model accounting for noise and include an intercept
- Created 2 non-overlapping sets out of the life stages
  - First set includes more data points and is called lower hierarchy
  - Second set includes fewer, more encompassing data points and is called higher hierarchy
- Running the MCMC with each of these different selected data sets.
- Meeting with Dr. Gilchrist:
- On the far right of Figure 1 we derived the weighting coefficient/expression coefficient omega as a function of time and ATP cost importance.
- Then, to handle the fact that we do not necessarily know the length of time stages or the definite occurrence of each time stage, we defined omega in terms of the probability of observing a given time stage for a certain amount of time divided by the total proabability.
- There was also a factor of the cost importance in the particular stage compared to probably the overall cost importance.
- We established that the probability of observing a certain life stage was = 1 for the life stages which are inevitable for the C. Elegans, specifically: the embryonic stages, L1, L4, and Adult stage.
- The probability of observing dauer, l2 dauer, and post dauer were all equal, and also were equal to 1 - probability of entering L2/L3.
- The last formula we arrived at was that the probability of a life stage is equal to the the product of the omega, the total time, and the average ATP cost divided by the product of the expected value of the time in the life stage and the life stage specific q value.

- Since we know the expected time value and the omega value, we can obtain ratio values for the different q values, as the total time and average q can cancel out.
- We also can look at the assumption that the embryonic stages all have the same relative q values since there is a limited amount of energy in the egg that does not change in a ratio between life stages.
- The first step now is to write two codes that include all the dauer stages and differ only in the accounting for the two aforementioned non-overlapping sets.
- To Do:
  - Address how the measurements in a stage that consists of substages is generated.
  - Find out where Cedric got the data.
  - Look at how the mass of the worm changes with each life stage to gain insight into the q values.
  - Make sure sum of weighting constraints = 1.
  - Write the code for each set.

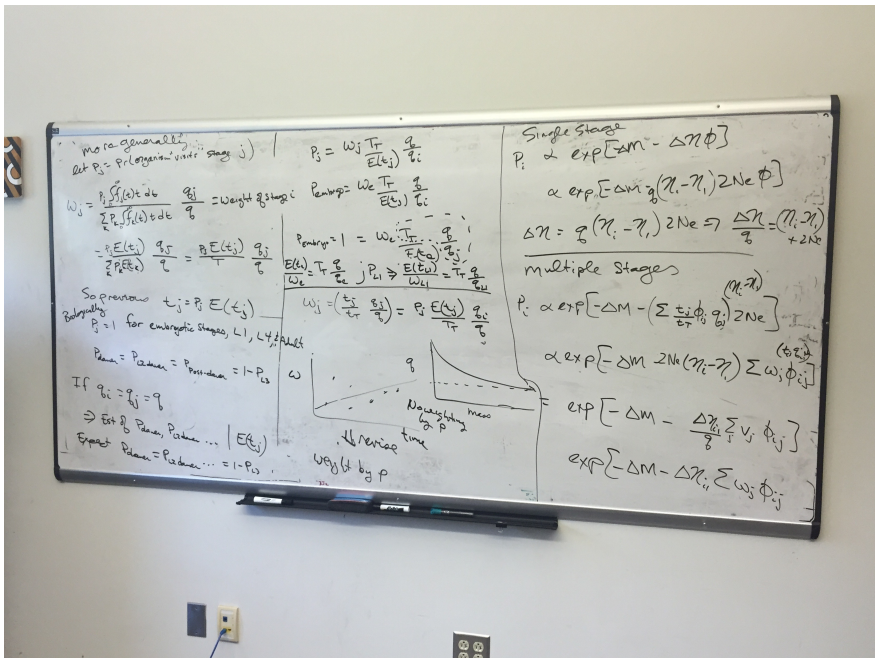

Figure 1: July 22nd Derivation of Expression Coefficient as a Multivariable Function

25 July 2016

1 C. Elegans Life Stages

- Morning Meeting:
  - Look up what unit testing is.
  - Look up mortality Leslie matrix.
  - Look up Hawk and Dove game.
- Trying to gain comfort and ability with this latex program and to become more adept at making notes.
- Asked Cedric where he got the life stages data and he gave me the link to the website:
- https://www.ebi.ac.uk/gxa/experiments/E-MTAB-2812
- Will try to type up the mathematics tomorrow.

26 July 2016

1 C. Elegans Life Stages

- Finished making code for both the higher nonoverlapping set with Dauer and the lower nonoverlapping set with Dauer both allow for an intercept α term.
  - Lower includes: 4-cell, gastrulating, enclosing, 3-fold, fully-elongated, L1, L2, L3, L4, adult, L2D, Dauer, Post-Dauer.
  - Higher includes: proliferating, elongating, fully-elongated, L1, L2, L3, L4, adult, L2D, Dauer, Post-Dauer.
- Running the Model currently.
- Looking into how the Empirical Life Stage Data was obtained.
- Name of the Experiment we got the data from: E-MTAB-2812 - Deep sequencing of the Caenorhabditis elegans transcriptome using RNA isolated from various developmental stages under various experimental conditions RW0001 - uninfected worms.
- Does it matter that the data comes from various strains?
- There is a slight variation in the different type of cells that each run looked at.
- Although most give the cell type as organism, there is also some data taken from specifically somatic cells and neuronal motor cells.
  - These different cell types have very different
  - The study also includes varying ages for some of their life stages.
    - 3-fold Embryo has 12 runs that have ages ranging from 500-710 minutes old, but each run consists of a 150 minute span.
    - Elongating Embryo has 15 runs that have ages ranging from 350-620 minutes old, but each run consists of a 150 minute span.
    - Enclosing Embryo has 5 runs that have ages ranging from 170-350 minutes old with each run consisting of a 150 minute span.
    - Fully Elongated Embryo has 17 runs, 9 of which do not have ages provided, with the other 8 ranging from 590-830 minutes and each spanning 150 minutes.

    * Gastrulating Embryo has 6 runs, 1 of which does not have an age provided, with the other 5 ranging from 80-300 minutes old, and all but one spanning 190 minutes each but the exception spans 150 minutes.
    * Late Cleavage has 7 runs that range from 230-470 minutes old, each spanning 150 minutes.
    * Proliferating Embryo has 28 runs, but only 14 runs have ages provided. These ages range from 0 - 200 minutes with each spanning 150 minutes.
    * There is no age data provided for the runs of 4-Cell Embryo, Adult, Dauer, Embryo, L1, L2, L3, L4, L2D-dauer, Post-Dauer.
    * Total of 201 runs analyzed.
- Analysis Methods: (Directly Quoted)
- Pipeline Version: iRAP 0.6.1p9
- Analyzed Libraries: Single-end only
- Filtering Steps:
    * Step 1- Discard reads below minimum quality threshold.
    * Step 2- Check of bacterial contamination; discard offending reads.
    * Step 3- Discard reads with common uncalled characters (e.g. N)
    * Step 4- Remove reads from pair-end libraries that were orphaned by filtering steps 1-3.
- Read Mapping: Against genome reference (Ensembl Metazoa release: 26) tophat2 version: 2.0.12
- Quantification: htseq2 version: 0.6.1p1
- Normalized Counts Per Gene:
    * (FPKMs) are calculated from the raw counts by iRAP.
    * These are averaged for each set of technical replicates, and then quantile normalized within each set of biological replicates using limma.
    * Finally, they are averaged for all biological replicates (if any)
- Model finished running and I have the results located in files that I will push to github.
- I will run the other model tonight as it would be interrupted by my shutting my computer when I leave later.
- TO DO:
- Run Model with other set of data.
- Continue to research about the empirical data.
- Start writing code to compare the results.

27 July 2016

1 C. Elegans

- I ran the model for the higher hierarchy with dauer included last night and recorded the results in the file "Model_Results_Higher_With_Dauer".

- The results from yesterday's run of the lower hierarchy with dauer are in the file "Model_Results_Lowest_With_Dauer".

- The mean log likelihood for the higher hierarchy with dauer was -57252.98 while the mean log likelihood for the lower hierarchy with dauer was -49394.46.

- In these two files I have the mean value for each weighting coefficients–called beta in the file–and the values of the 2.5% and 97.5% quantiles.

- *** Would overlapping gene use in different life stages affect the weighting and artificially increase or decrease the values in life stages that show similar transcription activity?... e.g. proliferating..

- I would assume that all the cell activity is similar throughout the replication processes in each of the proliferating stages.

- Getting an expected time equal to the full time of proliferation for the 4-cell stage is a hypothesis in the lower hierarchy.

- I further catalogued all of the ages given by the experiment into the same excel file that has the original life stage and order numbers.

- That file name is "Data/Ordered.Time.Spent.Lifestages.Celeg_1.csv".

- Below I will attach two PDFs that have the graphic results from each trial run.

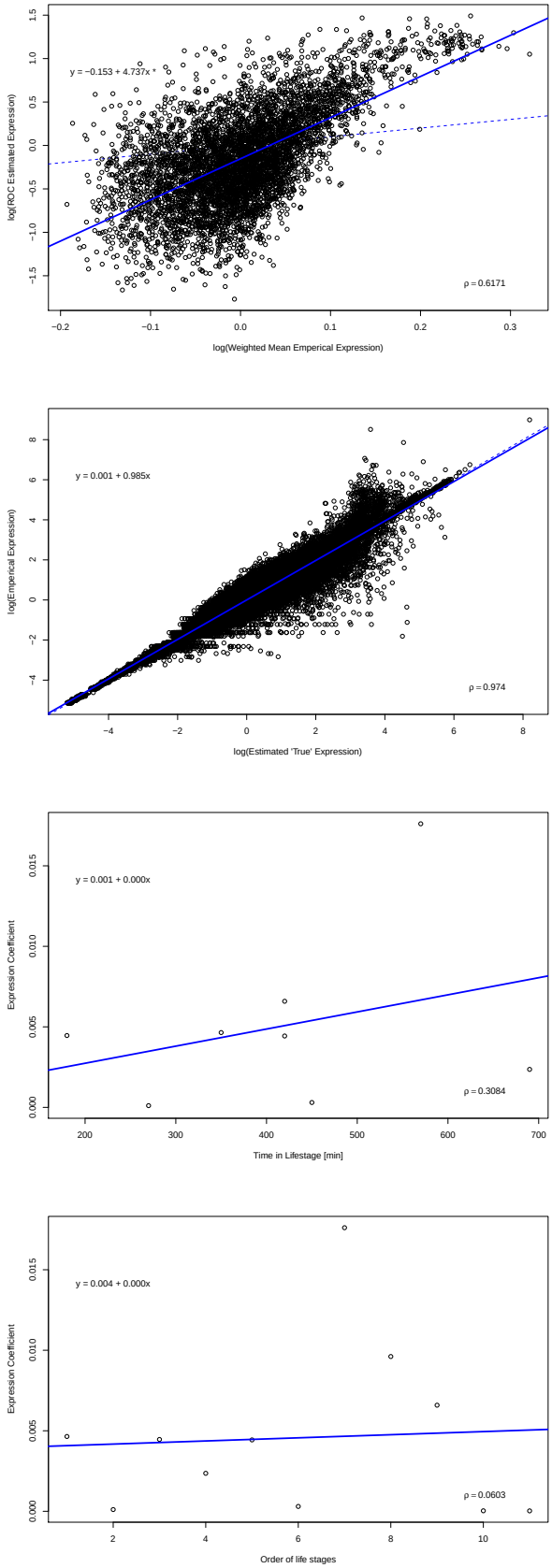

Figure 1: July 27th Graphs From Higher Hierarchy With Dauer Model Run

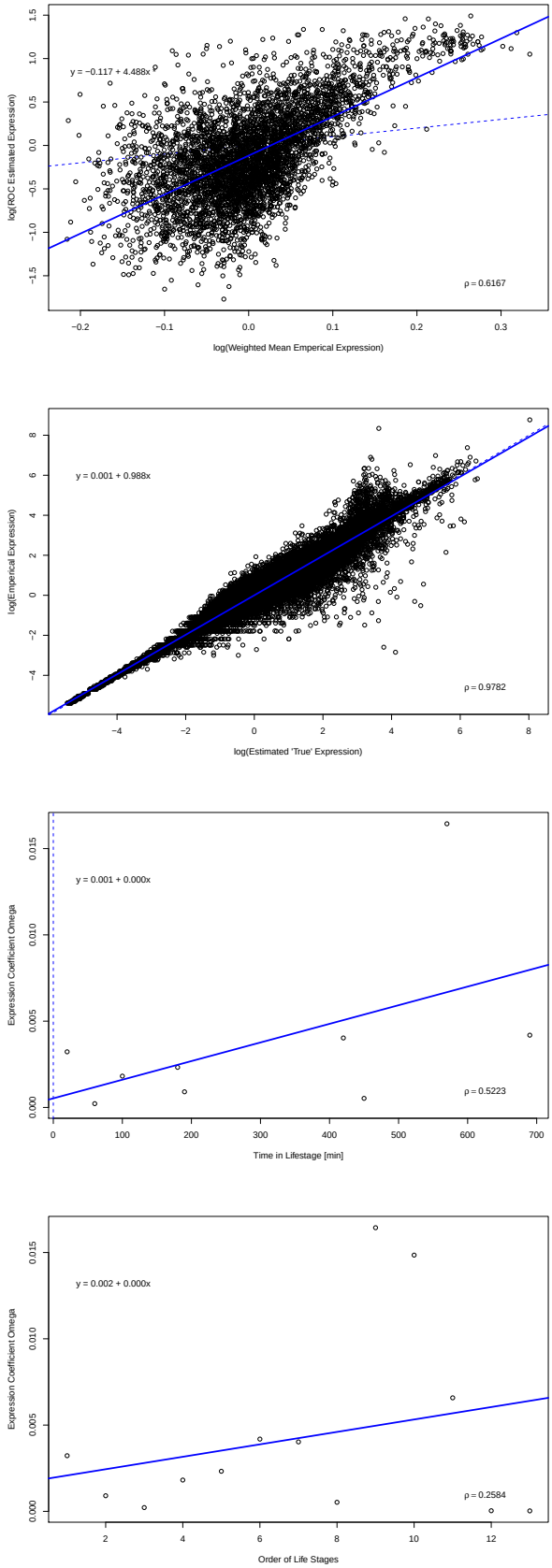

Figure 2: July 27th Graphs From Lower Hierarchy With Dauer Model Run

- **** Need to figure out how to display all 5 of the graphs that are on each pdf by splitting them onto multiple pages so they are not too small to read.
- The first graph for each set of Figures charts the log of the Roc Estimated Expression on the y-axis and the log of the Weighted Mean of the Observed data. *** Need to look at this
- The second graph displays the log Empirical Expression–the observed data– versus the log Estimated True Expression.
- The third graph displays the Time in Lifestage plotted with the Expression Coefficients
- The fourth graph displays the ordering of the life stages.
- I put these expected life stage durations and orderings in the file
- "Data/Ordered.Time.Spent.Lifestages.Celeg_1.csv".
- The fifth graph displays the original log Roc expression data plotted against the log Mean Empirical Expression.