Project Milestone 4 – Technical Brief Draft

To: John P. Watson, President, NovelEnzymes Inc.

From: ENGR 132 Team 30

RE: Data Analysis

Date: 12/1/19

Introduction

NovelEnzymes Inc. has sent a memo that asked us to create algorithms in order to help them determine an initial starting point for enzyme pricing which is done by our function that is supposed to be able to create a model to determine the price of each of the different types of enzymes tested. We can ensure that our algorithm is correct by comparing our v0 values with the PGO-X50 Reference Values given to us. Some constraints to this project are that the executive function must be fully automated, the executive function must have no inputs or outputs, and this technical brief is due by 1:30 pm on 12/2/19.

The algorithm works by taking a tested optimized range of data points of the reacted enzyme (17 unit range for noisy and 11 for clean) and averaging the values within that range in order to produce smooth data to plot on a graph. It then moves onto the next set of data points to smooth until it reaches the end of the enzyme data. For example, the first set of data includes 1st-17th data points(for noisy data), and the second set of data includes 2nd -18th data points.

Our first key decision in improving the algorithm was how to approach noisy data; we decided to find a moving average for all of the data or to group data and then average them. This presented another issue: how large should the data grouping be without losing data while also minimizing the effects of outlying data? This impacted the quality of our solution by also potentially removing any crucial data points that could have potentially been able to create a more accurate graph. This can be seen within M2 Figures 3.2-3.6 where, although there is an accurate graph that gives the general model for the velocity of the equation, it is seen that there is a loss of data as a result of our current method. That brings us to our next critical decision to decrease data loss.

A second critical decision was optimizing the width range that the algorithm averaged either the noisy or clean data. We ran a test to calculate the r^2 value of a model using data grouped by different sizes and we determined that the optimal averaging size would be the one with the corresponding r^2 value closest to one. This width value ended up to be a value of 17 for the noisy data and 12 for the clean data. What we also did to improve the moving average algorithm was to instead make the width algorithm move one index value at a time. To elaborate, our original method moved the width range for the noisy data from 1-17 to 18-34. This created a tremendous loss of data, so to account for this we changed the algorithm to move the width range by one index at a time such that it would move from 1-17 to 2-18 and so on. It can be clearly seen in M4 Table 3 that the algorithm has improved by 6.11% and 64.73% for clean and noisy data respectively.

The final key decision we made to improve our algorithm was in calculating the Km and Vmax using the Hanes-Woolf plot[1]. Using a linearized model and v0 values, the slope of the model is

equivalent to the inverse of Vmax, while the y-intercept is Km/Vmax as seen in Appendix A of this document.

Parameter Identification Procedure

Our algorithm starts off by importing all the test results for the enzymes and, based on the enzyme inputted into the algorithm, selects the columns to use from the data. The algorithm then begins a looping structure for each column of the test results, sending both the original test and duplicate test data into our smoothing function, which returns a smoothed column for the test results and smoothed column for time. The looping structure continues to calculate the v0 values for the test by dividing each value in the smoothed time column by each value in the smoothed test results, this is done with both the original and corresponding duplicate test. These v0 values are then used to calculate Km and Vmax in the executive function. The executive function linearizes the v0 values using the Hannes-Wolf method, then plots them against the Substrate data (Marasović et al., 2017). The Vmax is then calculated by dividing 1 by the slope of the plot, and the Km value is found by dividing the y-intercept of the plot by the slope of the plot. For a visual of the Hannes method, Appendix A illustrates this linearization method.

Results

As you can see from the example execution located in Appendix B, our algorithm returns ten initial reaction velocity values, the maximum reaction velocity, and Michaelis-Menten constant for each of the five enzymes, with the highest maximum reaction velocity being that of enzyme-E at 1.7027, though it is also the second most expensive in suggested price, the most being enzyme-A at $414.94 per pound. We have attached the suggested prices for each enzyme and displayed the equation for the model used to interpret price points in the figure. The goodness of fit statistics for this model has also been included, with the r^2 value being 0.969 and the SSE and SST values being 0.137 and 4.445, respectively. With the calculated statistics, we believe the model should be effective in predicting prices accurately and, since none of the enzymes fell outside the data range, you should find all the suggested prices for the enzymes to be useful.

Interpretation

The error that we can account for in the algorithm is the loss of data with our current smoothing model. Because it is a moving average, there is data loss in the creation of smooth data because it must take the average of the width and combine them into one value. As a result of this, the total data will be decreased by a factor of how many iterations the code runs until it has gone and averaged all the data. In essence, it can be defined as *(smoothed data points) = (number of data points) / iterations*. Evidence of this is seen within our code when it takes the file Data_PGOX50_clean or Data_PGOX50_noisy with an index value of 1222 per test in the file and creates a resulting smoothed dataset that will have an index value of 1222/iterations. The correct optimized width will vary depending on if it is working with either smooth or noisy data.

What NovelEnzymes can honestly say about their products performance is the enzymes function to a specific precision. This goes by saying that their margin of error is low so they have consistent results to show for their customers. Based on the data gathered the noisy data isn't too bad with an r^2

value of 0.969. This being said, they can say that they can guarantee that their product is not going to stray too far from any data collected.

References

Marasović, M., Marasović, T., & Miloš, M. (2017). Robust Nonlinear Regression in Enzyme Kinetic Parameters Estimation. *Hindawi Journal of Chemistry*, *2017*, 1–12. DOI: 10.1155/2017/6560983

APPENDIX A:

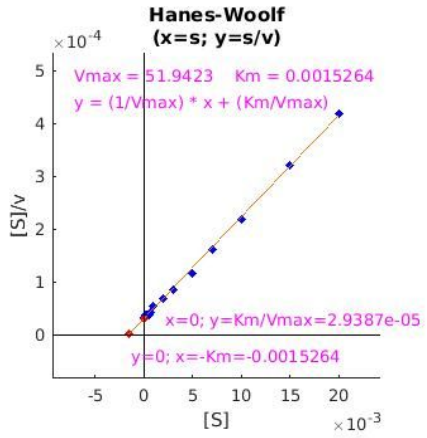

Image Source: https://bit.ly/2quI9YQ

APPENDIX B

```
Command Window
  Pricing Function for Enzymes (Exponential):
  Price(USD($)/lb) = 1950.922 * 10 ^ (-0.004 * Km(uM))

  Goodness of fit for pricing function:
  SSE: 0.137  |  SST: 4.445  |  r^2: 0.969

  ------------------------------------------------
  Parameters for Enzyme NextGen-A:

  v0 values (uM/min):    0.0235    0.0464    0.0890    0.1685    0.2922    0.4440    0.6091    0.7475    0.8364    0.9086

  Vmax (uM/min): 0.9737  |  Km (uM): 150.3015  |  SSE: 0.00016

  Recommended Price: $414.94 per lb
  ------------------------------------------------
  Parameters for Enzyme NextGen-B:

  v0 values (uM/min):    0.0096    0.0191    0.0381    0.0710    0.1383    0.2271    0.3669    0.5192    0.6961    0.7389

  Vmax (uM/min): 0.8763  |  Km (uM): 337.2487  |  SSE: 0.00195

  Recommended Price: $60.51 per lb
  ------------------------------------------------
  Parameters for Enzyme NextGen-C:

  v0 values (uM/min):    0.0245    0.0488    0.0928    0.1752    0.3266    0.5021    0.7191    0.9094    1.0496    1.1423

  Vmax (uM/min): 1.2471  |  Km (uM): 185.0924  |  SSE: 0.00002

  Recommended Price: $289.99 per lb
  ------------------------------------------------
  Parameters for Enzyme NextGen-D:

  v0 values (uM/min):    0.0208    0.0404    0.0789    0.1542    0.2987    0.4894    0.7597    1.0418    1.2733    1.4180

  Vmax (uM/min): 1.6277  |  Km (uM): 288.6323  |  SSE: 0.00025

  Recommended Price: $99.83 per lb
  ------------------------------------------------
  Parameters for Enzyme NextGen-E:

  v0 values (uM/min):    0.0382    0.0741    0.1436    0.2392    0.4844    0.7362    1.0230    1.2794    1.4490    1.5735

  Vmax (uM/min): 1.7027  |  Km (uM): 167.2828  |  SSE: 0.00065

  Recommended Price: $348.36 per lb
  Elapsed time is 2.808763 seconds.
fx >>
```