**Abstract**

**Introduction**

In the field of Biology, mathematical models can be used to explain or describe natural phenomena and observations. This method of modelling can be applied especially in the field of Ecology and Evolution. Models are a simplistic method to describe the general behaviour of long-term macro-biological events noticed in nature (Johnson and Omland 2004). A field in Ecology is Population Ecology, and simplified models can validate the dynamics of population growth and how different factors and conditions affect the population numbers over time. The models chosen are generalised models describing bacterial growth. Therefore, they disregard the capacity for evolution, classifying the organisms as homogenous (Levins 1996). Finding the best fit will be achieved by fitting suitable models and evaluating the best fit to the data using either maximum likelihood or least squares (Johnson and Omland 2004). Model fitting has become an alternative approach to the traditional null hypothesis tests when carrying out scientific experimentation. If carried out properly, parameters and overall function of the model are used to make biological inferences on the patterns recorded (Johnson and Omland 2004). The challenge, however, is finding the appropriate model and taking into considerations the assumptions and trade-offs of the different models.

Models aim explain the important aspects of the data, thus have many assumptions. When choosing a model, it is important to account for the trade-offs between generality, realism and precision, focusing on model that sustain generality and realism, as they are able to quantitatively explain long-term trends that the data reveals (Levins 1996). In other words, the purpose of the model, and what is being investigated is an important factor. In addition to this, ensuring that the parameters have biological significance rather than just statistical significance (Johnson and Omland 2004). For this project, identifying the correct model is achieved through the process of testing multiple existing mathematical and biology models on the data and finding the model that best fits the data using the analytical method of Maximum Likelihood. The outcome should be plausible models that for most observed data, consistently describes the patterns in these observations mechanistically or empirically.

The focus of this project is to test multiple non-linear regression models on microbial population data. Identifying the best model(s) that best fit and or describe the pattern of the data. Microbes grow through a process known as binary fission, doubling the population continually (Webb 1986). This exponential growth coupled with limiting factors gives microbial population growth a sigmoidal shape which can be divided into four phases (Peleg and Corradini 2011). The trend begins with birth, a starting population (*N0*) and growth where limiting factors is not yet inhibiting, thus having an exponential growth curve, with a growth rate notation of *r* (Zwietering et al. 1990). Limiting factors such as space, nutrients and particularly carrying capacity can inhibit growth, resulting in the rate of growth to decrease and eventually reaches saturation, becoming asymptotic. The asymptotic value being the carrying capacity, with the notation *N\_max* (Zwietering et al. 1990).

Four mechanistic models will be tested on the dataset. These are:

Logistic (formula)  
Baryani (formula)

Gompertz (formula)

Buchanan (formula)

Model fits be will be assessed using the Akaike Information Criterion (AIC), which is a relative comparison where the best fit is the model with the minimum value (Vrieze 2012).

An additional investigation on the correlation between growth rate and temperature is also carried out. Given the range of temperature from 0 degrees Celsius to 37 degrees Celsius, the expectation is to have a positive linear relationship between increasing temperature and growth rate. By using linear regression of temperature and the natural logarithm of the growth rate, for each of the unique species and mediums will give insight if there is a correlation between temperature and growth rate.

**Method**

In this project, published data was provided from the works of multiple experiments on microbial population and growth. Experiments manipulated with different growth conditions including mediums such as Tryptic Soy Broth, Chicken Breast, and Tryptone Glucose Extract Agar, temperature range for the microbes to grow and divide, and recorded through different measurements techniques including *CFU*, *OD\_595* or count (*N*). Population numbers and time were recorded discretely at different time intervals. The data were stored in a .csv file, LogisticGrowthData.csv. All four mechanistic models were fitted on the data.

***Computing Tools***

***Python***

Python software was used to wrangle the data. The data provided was in one large .csv file and needed to be manipulated before being filtered. Combination of functions and commands were used to wrangle the data and remove unwanted data points, as well as converting to log space. Pandas and Numpy were the two packages used for the purpose of data import and carrying out mathematical manipulations of the dataset.

***R***

R software was used to filter the dataset and carry out data analysis. Software package *dplyr* was used to filter the dataset and create unique ID’s, which were saved as *.rda* files. The mechanistic equations to be fitted on the data and functions to find optimal starting value for the parameters were done in R. Using *minpack.lm* package, non-linear regression analysis was carried out using the *nlsLM()* function as well as plotting the different unique analysis.

***Bash***

Bash script was used to run the different python and R scripts, including the latex file, *.*tex, to convert the latex into pdf.

***Data Wrangling and Filtering***

The data was first wrangled before being filtered. All abundance recordings of OD\_595 were first multiplied by 100. Next the data was treated in linear scale, therefore negative time and population values were either removed or shifted. A row containing an outlier value of -668.2839 was first removed from the dataset, before being shifted by the next minimal value of -0.003434656. By shifting the dataset, it would make the negative population values positive. The row containing the second minimal value was also removed as the log10 of the data is undefined. In addition, negative time values were set to zero as it is not realistic to have a negative time recordings. The data was then transformed to a log scale, by taking the common logarithm of the population abundance. This is to fit the criteria of three of the models and make the data less skewed. Although Logistic model is not in log scale, the output value is logged to maintain the linearity with the data and so that the parameters maintain biological meaning.

After wrangling and tidying the data, it was filtered based on the categories: species, medium, temperature, citation and replicate. It is important to note that method of collecting population count was not a factor in the model fitting and assessing of best model(s). The unique datasets were saved in *.rda* files which included the unique ID’s and a data frame of the time intervals, labelled *Time* and the common logarithm of the population count at each discrete time interval, labelled *Abundance*.

The four mechanistic models were then fitted to all the different sub-datasets. Of the four, the Logistic formula is derived from bacterial growth properties, where *r\_max* is the maximum growth rate during the exponential phase and *N\_max* represents the carrying capacity, while the other three models are based on mathematical derivations. However, since all the models utilise the same parameters, their values can be interpreted to have the same biological interpretations (Levins 1996). Using R Programming Language, each of the models will be tested on the data and the Akaike Information Criterion (AIC) will be the statistic to analyse the model of best fit.

***Fitting the models***

The R software will be used to test the models and their fits. Using the package *minpack.lm*, the *nlsLM()* function is used to fit the non-linear mechanistic models. Since it is non-linear regression fitting, parameter starting values must be estimated within range of the global maximum or minimum respectively as unlike linear regression, the function can find an optimal value, but only be a local rather than a global minimum or maximum. To fit the Baranyi, Gompertz and Buchanan models, it requires inputting four starting parameter values (*N\_0, t\_lag, r\_max, N\_max*) for the parameters of the equation, while the Logistic model only requires three of the parameters, not including *t\_lag*.

***Finding Parameters***

*N\_0* is the starting population value at the first-time recording, thus the first data point. This was found using the *min()* function of the abundance column of the dataset. The carrying capacity (*N\_max*) is the largest value found in the dataset, which is the maximum value of the dataset, hence using the *max()* fucntion. The growth rate, r\_max, is the greatest gradient value of the derivative found at the point *N\_max*/2, otherwise known as the inflection point. To find the estimate of this parameter, a function utilising a while loop was used. It describes the rate of change of the population over time. Starting from the first data point and progressing to the final two data points, each iteration calculated the slope between adjacent data points, saving all results in a vector and the greatest value was taken to be the *r\_max*. For example, for the first filtered data set: replicate 1 of Pseudomonas sp. in APT Broth at 12 degrees Celsius, the first two datapoints are 0.1386796 at time 0 and 1.5635348 at time 18.14733 respectively. The slope of would be:   
(Math) = 0.07851596  
All calculated slopes are stored in a vector and the *which()* and *max()* function retrieves the greatest value.

The models include a time lag, which is a delay in time of bacteria growth. The estimation is finding the intersection between the lag phase, which can be represented by the first population value, and exponential growth phase of the curve (Peleg and Corradini 2011). This was calculated by the intersection of y=N\_0 and the line with gradient *r\_max*. Using the y-intercept value from the *r\_max* line that produced the greatest slope, rearranging of the linear equation leads to the formula:

*t\_lag=(N0 – y\_intercept)/r*

Species and medium were the factors accounted for when identifying the correlation between temperature and growth rate. For each of the unique ID’s, the growth rates, temperatures, Species and growth medium were recorded in a large matrix. The matrix was subset to the different unique ID’s and linear regression analysis was done to analyse the correlation.

***Fitting Analysis***

All models were fitted on all of the data and the best fits were analysed using the Akaike Information Criterion (AIC), which measures the amount of lost information when fitting the models (Posada and Buckley 2004). We are using AIC to assess the model fit as it is able to analyse nested and non-nested model fits as well as being less penalising for additional parameters unlike BIC (Posada and Buckley 2004). In this case, the four mechanistic models which all have similar parameters with *t\_lag* being the additional parameter.

***Additional investigation***

Aside from identifying the best fit model, another aspect being investigated is the correlation between growth rate and temperature. The models used despite their derivations contain the same parameters, having a unified theory for the bacterial growth behaviour (Levins 1996). Therefore, biological meanings can be interpreted by the parameters and investigated further. In a separate R script, growth rate values were collected again for each of the unique data sets, as species and medium are factors that can also affect growth rate. For each of the species grown in their respective mediums at a range of temperatures, growth rate was recorded for each of the temperatures and linear regression analysis was used to analyse the correlation between the two variables.

**Results**

Filtering and wrangling the data resulted in 305 different unique sub-datasets being produced, saving the data frames of time and common logarithm of the population, and the respective unique species ID in *.rda* files. The plots to show the trend in population over time has time along the x-axis and *Abundance* along the y-axis. Plots revealed an inconsistent distribution of the dataset, some sigmoidal, log distributed or completely random. All starting parameter values were generated for 298 of the 305 datasets and any negative growth rate slopes (*r\_max*) was set to *NA*.

The models fit approximately 79% of the data. Of the 305 datasets, 241 total model fittings were successfully produced. The 241 fittings were distributed among the four models, with majority of the relative best fits were by the Gompertz model, with 99 of the 241, while the next best fit was the Baranyi model with 65, and Logistic and Buchanan with 50 and 27 respectively. The Buchanan model had 105 fits, Baranyi model had 218 fits, Gompertz model had 211 fits and finally the Logistic model had the most with 241 fits.

In the additional investigation of the correlation between temperature and growth rate, visualisation shows that the growth rate is positively correlated with temperature, with an average value of 0.82. This suggests strong correlation between growth rate and temperature, that as temperature increases, growth rate will increase as well. When plotting the log growth rate against the temperature values, overall the distribution tends to show a positive correlation. Some of the plots have an exponential increasing shape, possibly suggesting some underlying physiological mechanism not explored.

**Discussion**

**Wrangled data and starting parameters**

As seen from the figures later, when visualising the data, the common log of the abundance produced sigmoidal functions as well as log distribution, which we expect. Distribution of the datasets were not consistent as some plots revealed no pattern, having a random distribution of points along the axis, unable to derive starting parameter values and overall fit the models. Furthermore, some of the datasets revealed a decline once the asymptotic value has been reached. This is the death phase which is ignored in the model fitting process (Zwietering et al. 1990).

Getting all required starting parameters is necessary for the model fittings to be produced. Not all the datasets followed a sigmoidal or log distribution, which justifies how some *r\_max* and as a result ­*t\_lag* values were not being produced. Although the Logistic model only requires three of the starting parameters, the growth rate parameter is the determining factor of all the models fitting. Starting population (*N\_0*) and carrying capacity (*N\_max*) are easily derived from the datasets however, growth rate and time lag are mathematical derivations thus have the possibility of having *NA* valuesdepending on the distribution of datapoints. The growth rate coefficient represents the max slope between points, which is the derivative along the inflection point. While the *t\_lag­* requires the y-intercept of *r\_max* slope as well as the actual slope value to derive it.

**Model fitting**

All four models were nested, utilising similar parameters with one parameter *t\_lag* separating the Logistic model from the others. The Logistic model is derived from external limiting factors including resource and density. These external factors play limit the exponential growth of the population, hence producing the sigmoidal shape where it reaches an asymptotic value known as the carrying capacity (Webb 1986). The Baranyi model is a differential model which accounts biological factors of individual microbial growth. It takes into account environmental variation and the physiological limiting factors, specifically the limits in the rate of biochemical reactions within the microbe (Buchanan et al. 1997; Grijspeerdt and Vanrolleghem 1999). The Buchanan model is a simplified version of the Gompertz model and Baranyi model, which accounts for biological variability, considering individual physiological factors and population factors such as adaptation at each of the growth phases to produce a three-phase linear model (Buchanan et al. 1997).

Overall, the results show that the Gompertz model was the most robust model, having the greatest number of goodness of fit. In other words, for the provided dataset, it is the best model to describe the population dynamics of microbial growth over time. This is not surprising as it is a primary level model, which only describes change population over time (Grijspeerdt and Vanrolleghem 1999). The Gompertz model is mathematical derivations to form the sigmoidal relationship (Buchanan et al. 1997). Despite its parameters, it is an empirical model not accounting for any biological process. The robustness of the model is from the parameters. The parameters account for shape or curvature of the fit, and location along the axis, allowing the model to shift the fit along the x-axis or y-axis and maintain its overall shape (Tjørve and Tjørve 2017). Furthermore, similar to the method of deriving the growth rate from the datasets, the growth rate in the Gompertz model is derived along the inflection point.

**AIC**

All of these can be seen through the AIC values produced. Depending on the dataset, the AIC values produced negative or positive ranges of values, and the best model fit was the model that produced the minimum value relative to the other fits. The table shows that the Logistic, Baranyi and Gompertz models were more dynamic in fitting the data compared to the Buchanan model. AIC and BIC cannot explain how “true” a specific model is as it is only comparing relatively between models

**Temperature and Growth Rate Correlation**

The histogram shows that as the correlation between temperature and growth rate is positive. A linear regression was used as temperature is treated to be a continuous variable. In addition to the correlation coefficient, if we use a significance value of 5%, using fisher’s method for independent test statistics, the resulting p-value is much less than the significant level. This means that the slopes are significant, not random and support the positive correlation between temperature and growth rate. It is important to note that different species do have different temperature tolerances, which describes the different distributions of the natural log growth rates.

**Conclusion**

-three parameter model recommended over four (Zwietering et al. 1990)

-data showed a decline (death phase) and models did not account for mortality phase

-although the derivations are empirical, can give biological meaning to the parameters making

them mechanistic