Phenomenological Models Outperform Mechanistic Models for Estimation of Messy Biological Data

Eamonn Murphy

Department of Life Sciences
Silwood Park
Imperial College London

Friday 3rd December 2021

Word Count: 2222

Abstract

Modelling bacterial population growth is an important problem, and a variety of potential models have been proposed to explain this. Many of the mechanistic models proposed include common assumptions around the important parameters, such as initial population size and carrying capacity. In this project, I fit 4 different models, quadratic and cubic polynomials, Logistic and Gompertz, to a diverse dataset of bacterial populations. Model fitting was carried out successfully, and Gompertz was the best fitting model, when assessed using AICc (Akaike Information Criterion with correction). However, the linear phenomenological models fit better than the mechanistic models overall on the dataset, indicating that phenomenological models outperform mechanistic models for fitting to messy biological data.

Introduction

Modelling population growth in bacteria is an important problem, as it allows for many applications, including prediction of bacterial numbers in food, which is important for spoilage (?). As well as prediction, population growth models can help to improve our mechanistic understanding of bacterial growth. This problem has been very well covered in the literature, and many data-sets of bacterial population growth are available. A large number of models have been proposed for bacterial growth, including a number of linear and non-linear models. In this mini-project, I have attempted to fit a number of phenomenological and mechanistic models to a provided data-set of bacterial growth, taken from 10 different published papers. These data include a variety of species, growth mediums and temperatures.

There are a number of important proposed parameters which help to explain the growth of a bacterial population. The first is that, upon introduction to a new environment, population size tends to follow a sigmoidal (or S-shaped) curve (?). This happens as bacteria must first adapt to their new environment before beginning to grow and expand. These adaptations can include up-regulating new sets of genes, such as the lac operon, when bacteria are transferred from a glucose to lactose environment (?). The parameters that explain the shape of this curve include the initial population size ($N0$), the length of the lag phase (modelled by t_{lag}), the highest rate of population growth reached (r_{max}) and the carrying capacity of the environment(K). The carrying capacity is the maximum possible population that can be sustained in that environment, and explains underlying variables such as the nutrient density and type, the size of the envi-

Table 1: Equations for the Bacterial Growth Models

Model	Formula	k (number of parameters)
Quadratic	$\ln(N_t) = B_2t^2 + B_1t + B_0$	3
Cubic	$\ln(N_t) = B_3t^3 + B_2t^2 + B_1t + B_0$	4
Logistic	$N_t = \frac{N_0Ke^{rt}}{K+N_0(e^{rt}-1)}$	3
Gompertz	$\ln(N_t) = N_0 + (K - N_0)e^{-e^{re\frac{t_{lag}-t}{(K-N_0)\ln(10)}+1}}$	4

t - time; N_t - population size at time t; K - carrying capacity; r - highest rate of population growth reached; t_{lag} - delay until population growth, or x-axis intercept of the tangent r

ronment, competition, and other underlying characteristics of the bacterial species that you are studying (**?**).

The simplest set of models which can explain bacterial growth are linear quadratic and cubic models (formulae in Table **??**). These are what would normally be referred to as phenomenological models; they do not attempt to apply any of our knowledge about how bacterial populations grow and what determines it, and simply try to find a best fit to the existing data. However, there is some form of mechanistic explanation to any reasonably fitting model; for example, the intercept B_0 will likely be an approximation of the starting population size, while the slope B_1 will be a rough approximation of the rate of growth during the exponential phase.

There are also quite a large number of non-linear mechanistic models which have been developed to try to explain bacterial growth. One of the simplest is the Logistic equation outlined in Table **??** (**?**). This equation will fit a sigmoidal curve to the data, using the parameters N_0, K and r_{max}. However, the Logistic equation does not include any parameter to model the lag phase, and as such it will be inaccurate for populations which have been recently inoculated, particularly those which have been transferred between media with significantly different environments. Another sigmoidal model which does include a parameter for the lag phase, t_{lag}, is the Gompertz equation (see Table **??**).

The purpose of this study is to explore the differences between phenomenological and mechanistic models. Logic would assume that the mechanistic models would fit better, as they are built specifically to explain bacterial growth, and contain parameters that are

Table 2: Format of Columns in Dataset

PopBio	Population or biomass measurement.
Temp	Temperature at which the microbe was grown (degrees Celsius).
Time units	Units time is measured in.
PopBio units	Units population or biomass are measured in.
Species	Species or strain used.
Medium	Medium the microbe was grown in.
Rep	Replicate within the experiment.
Citation	Citation for the paper in which the study was recorded.

designed to explain the underlying biological reasons for the shape of the growth curve. However, perhaps the freedom of the parameters in the phenomenological models to fit the specific perturbations of the dataset may allow them to get a better fit. Alongside this, we may have made incorrect assumptions about the factors determining bacterial growth, meaning that our mechanistic models are of the wrong shape or are insufficient to explain a real biological dataset. By fitting the outlined models to the growth curves in this dataset, I will attempt to explore these questions.

Methods

The population growth data was obtained from 10 different papers (**??????????**). The analysis on the resulting dataset was carried out using 3 scripting languages, Python, R and Bash. Initial data visualisation and wrangling was carried out using Python. Following this, further data wrangling was carried out in R, and model fitting for the different proposed models was carried out.

Data Cleaning and Preparation

A number of important choices were made when cleaning the data. First, a unique ID was assigned to each combination of citation, temperature and species, in order to identify separately each population / experiment. Following this, the subsets containing negative population values were removed from consideration, as I assumed these must have been measurement errors (since they are biologically impossible), and thus all of the population size measurements in that dataset could be suspect. I further cleaned the data by removing subsets with ≤ 6 data points. Given that the models have up to

4 parameters, models fit on data with only an equal or slightly greater number of data points will be overfit to random fluctuations in the data. Many of the other subsets will also encounter this problem given the sample sizes, but given that 41% of the populations have less than 10 samples, I decided to draw the cutoff at 6. I standardised the time in each subset so that the smallest time value was always equal to zero and other times were proportional to that. Negative values for time are plausible, as they may have started the clock at, for example, -24 hours upon inoculation.

One of the papers (?) included experimental replicates. I first tested that the replicates were not different from each other. In order to do this, I fit a simple quadratic regression model on time with replicate as an extra term. No significant differences were found (range of p $0.49 - 0.94$). Therefore, I decided to include the data as single subsets, discarding the replicate information from further modelling. Many of the models should be fit on the natural log of the population size. However, there were some population size values of 0 in the dataset. To deal with this, a small positive constant was added to all population values in the affected subsets, which is a common method as outlined in ?. After all data cleaning was completed, a total of 240 population subsets were included.

Model Fitting

Quadratic and cubic models were fit using standard linear least squares fitting in R (i.e. minimising the sum of the squares). I fit the Logistic and Gompertz models using non-linear least squares, and start values were needed for the free parameters. These were estimated as follows for the Logistic model:

N_0	Minimum population size in subset
K	2× maximum population size in subset
r_{max}	Iteratively determined preset value

Sampling was attempted in order to avoid model fitting to local minima, but did not improve accuracy, aso was discarded in order to improve computational efficiency.

For the Gompertz model, sampling was needed for the start values in order to find good model fits. This was performed as outlined below:

N_0	Normal distribution around minimum population size
K	Normal distribution around 2× maximum population size
r_{max}	Uniform distribution between iteratively determined presets
t_{lag}	Timepoint of maximum value of second order derivative

Table 3: Package versions and their usage

	Version	Usage
Python	3.8.10	
pandas	0.25.3	To inspect and clean data using an intuitive and clean method
R	4.1.1	
ggplot2	3.3.5	To plot data and make appealing figures
minpack.lm	1.2.1	To use Levenberg-Marquardt algorithm for fitting of non-linear models
AICcmodavg	2.3.1	To calculate AICc scores for models

The models generated by each sampling run were then compared using AICc, and the best model was kept for each subset. Model fits were checked manually by plotting. Following this, a number of tables and figures were generated for the final comparison.

Computing Tools

A combination of Bash, Python and R scripting was used for this project. Scripting was carried out on Ubuntu version 20.04.3 LTS. Bash was used for the final script to run the whole project, including LaTeX compiling. Python was used for some initial data cleaning and preparation. After this, R was used for further data cleaning, model fitting and figure generation. The reason for this is that R has well implemented functions for these use cases, as well as having the adaptable plotting package ggplot which allows for easily controllable and aesthetically pleasing figure generation. Packages used and their reasoning is outlined in Table **??**.

Results

A number of possible metrics to compare the models were considered. Akaike Information Criteria (AIC) and Bayesian Information Criterion (BIC) are two information theory methods which use the estimated log-likelihood of a model for comparison, giving an estimate of the prediction error. Lower values mean that the model is a better fit for the data. A modified version of AIC can be used when sample size is small, named AICc (AIC with correction for small sample sizes). This method was developed as there is a substantial probability of overfitting using standard AIC with small sample sizes

Table 4: Formulae for AIC, BIC and AICc

AIC	$2k - 2\ln(\hat{L})$
BIC	$k\ln(n) - 2\ln(\hat{L})$
AICc	$AIC + \frac{2k^2+2k}{n-k-1}$
$\ln[\hat{L}(\theta\|x)]$	$-\frac{n}{2}\ln(\frac{RSS}{n})$

k - number of parameters; n - sample size; $\hat{L}$ - Estimated likelihood; RSS - Residual Sum of Squares

(**?**). Overfitting refers to the situation where a model is overly sensitive to the random fluctuations of that specific dataset, and thus has little predictive or diagnostic power. Given the prevalence of small sample sizes in this dataset (only two subsets with > 100 data points), AICc seemed the most appropriate measure of model performance.

In order to compare model performance across the whole dataset, a simple tally was calculated, which counted the number of subsets for which each model was the best performing by AICc (lowest score). All of the scores had a difference of > 2, meaning they are significantly different by the general rule of thumb for AIC scores. The most successful model was Gompertz, which was the best fit for 39.6% of the subsets (see Figure **??**). However, the linear, phenomenological models outperformed the mechanistic models in general; 59.6% of the models were best fit by a quadratic or cubic model, against 40.4% by Gompertz or Logistic.

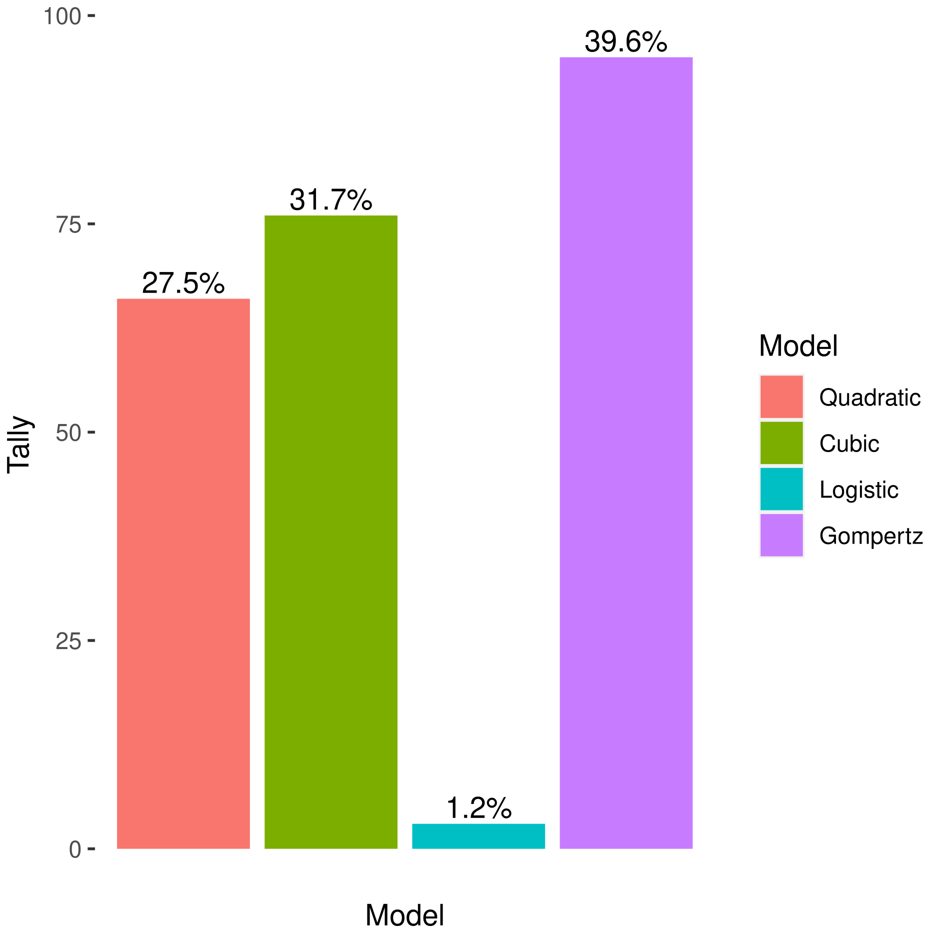

Figure 1: Number of subsets for which each model was the best fit by AICc. Each model is labelled with the percentage of the dataset for which it is the best fit.

Discussion

The above results display some interesting features and distinctions of phenomenological and mechanistic models. This is a typical messy biological dataset, including clearly implausible values, small sample sizes, and population subsets which do not meet any of our expectations for how a bacterial population will grow (see Figure ?? for an example). In this case, it makes sense that the linear, phenomenological models will perform better. Since these models do not make any assumptions about the shape of the data, they can fit better to whatever random fluctuations have determined the recorded measurements. These do not just include real fluctuations in the population size, but also measurement,

../results/scatterplots/246.png

Figure 2: Example of typical messy population data from subset ID 246

recording and transcribing errors.

On the other hand, the mechanistic models have a certain predetermined shape because of the parameters of the model, based on assumptions about how bacterial populations grow and develop. As outlined in ?, we attempt to find sufficient parameters to reduce the endless numbers of possible biological variables to a number of higher-level entities, which each group together many of these possible variables into a single, more simple metric. However, if we have made incorrect assumptions, we may have insufficient parameters which exclude many important variables, meaning that the mechanistic model is less adaptable to data of unexpected shape. For example, many growth curves will display a death phase after exhaustion of available nutrients, whereby the population will begin to decline (?).

Another advantage of mechanstic models is that they allow for more interpretation and understanding. We may be able to look at the parameters for two different datasets in a mechanistic model, and therefore make some deductions about the differences between the datasets. For example, we may be able to state that one species tends to have a longer lag phase than another, or that the carrying capacity of one medium is less than anothers. These kinds of deductions are not possible with phenomenological models, as it is unclear what the underlying meaning of the parameters is in a biological sense.

Regarding the predictive power of the models, it is difficult to tell which will be more powerful. In one sense, it is possible that the linear models are more susceptible to overfitting, since the model has more freedom to vary its shape to match the data. However, since these models do perform better by the AICc metric, it is also possible that the non-specified parameters have more freedom to reflect underlying structures in the data which are not captured by the specified metrics. In any case, it is not possible to assess predictive validity using this type of study design. To assess predictive power, you must withhold a portion of the dataset to test your pre-fit model upon.

By fitting the 4 models above to a real biological dataset, filled with messy data, errors and biologically implausible values, I was able to show the relative strengths of mechanistic and phenomenological models. Although the mechanistic models, Gompertz in particular, performed reasonably well, they were unable to explain the unexpected growth curves which could be found in this dataset. On the other hand, the phenomenological models, with more freedom to match their curves to the fluctuations of the dataset, performed better. This indicates that phenomenological models are better than mechanistic when used to estimate messy biological data. Besides being less able to explain experimental and measurement errors, it is likely that the mechanistic models above, Logistic and Gompertz, do not meet the basic criteria for sufficiency of parameters outlined in **?** seminal paper.