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:

Introduction

Modelling population growth in bacteria is an important problem, as it allows for prediction of bacterial numbers in food vectors etc. This problem has been very well covered in the literature, and many data-sets of bacterial growth are available. As well as this, 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 different 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 they go 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, 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 environment, competition, and also 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. These are of the form:

$$N_t = B_2 t^2 + B_1 t + B_0$$

where t refers to the time. 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 logisitic equation, which takes the form

$$N_t = \frac{N_0 K e^{rt}}{K + N_0(e^{rt} - 1)}$$

where N_0 is initial population size, K is carrying capacity and r is the maximum rate of population growth.

Methods

The population growth data was obtained from 10 different papers [Bae et al., 2014, Bernhardt et al., 2018, Galarz et al., 2016, Gill and DeLacy, 1991, Phillips and Griffiths, 1987, Roth and Wheaton, 1962, da Silva et al., 2018, Sivonen, 1990, Stannard et al., 1985, Zwietering et al., 1994]. 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.

Table 1: 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.

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, it was noticed that there were a number of negative values for population size in the dataset. The IDs which had the 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. It was also noticed that there were a number of negative values for time. This is more plausibly explainable, as they may have started the clock at, for example, -24 hours upon inoculation. In order to deal with this, 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.

The models were fit using the standard linear and non-linear model fits in R. Start values were calculated for the logistic fit as follows:

Model	Formula	k
Quadratic	$N_t = B_2 t^2 + B_1 t + B_0$	3
Cubic	$N_t = B_3 t^3 + B_2 t^2 + B_1 t + B_0$	4
Logistic	$N_t = \frac{N_0 K e^{rt}}{K + N_0 (e^{rt} - 1)}$	3
Gompertz	$\ln(N_t) = N_0 + (K - N_0) e^{-e^{re \frac{t_{lag} - t}{(K - N_0)\ln(10)} + 1}}$	4

Software and Packages

Software Versions

- Ubuntu - 20.04.3 LTS
- Python - 3.8.10
- R - 4.1.1

Packages and Dependencies

- Python
  - pandas
  - scipy
  - matplotlib
  - seaborn
- R
  - ggplot2
  - minpack.lm

Results

Four models were fit to the data: quadratic and cubic linear fits, and two non-linear models, logistic and Gompertz. These models were then compared against each other for each ID, using AIC scores. A table was generated with the tallies of how many IDs each model was most successful for.

A number of possible metrics to compare the models were considered. Akaike Information Criteria (AIC) is widely used to compare models, and gives an estimation of the prediction error. 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. 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. However, the problem with AICc in this sample is that there are cases where $k + 1 \geq n$ (where k is the number of parameters and n is the sample size), meaning that the denominator could be 0 or a negative number. This means that the penalisation term either is undefined or becomes negative, whereas with an extremely low n like this it should add a large penalty.

Another possible method is Bayesian Information Criteria (BIC). BIC is closely related to AIC, but uses a different penalty for the number of parameters, which is larger in most cases [Stoica and Selen, 2004]. Given that sample size is very small in many of the population subsets (117 cases where $n < 10$), and the outlined problems with AICc in this dataset, BIC was chosen as the most suitable estimation of model prediction error.

We can derive the situations where the penalty for additional parameters will be greater in AIC than BIC. This occurs when

$$2 > \ln n \tag{1}$$

$$n < e^2 \tag{2}$$

Since $n < e^2$ in 75 cases,

AIC	$2k - 2\ln(\hat{L})$
BIC	$k\ln(n) - 2\ln(\hat{L})$
AICc	$AIC + \frac{2k^2+2k}{n-k-1}$

Table 2: Tallies of numbers of models for which that model type had the lowest AIC

Model	Tallies
Quadratic	22
Cubic	79
Logistic	2
Gompertz	178

As we can see, the quadratic model was the best fit for the highest number of IDs, followed by the Cubic and Gompertz models. For the cubic model in particular, care was taken to discard instances where the model had a ”perfect fit” due to the number of data points being equal to the number of variables. For the Gompertz model, a fit was found for only 96 of the IDs, whereas cubic and quadratic had 281 fits, and logistic only one less at 280.

Discussion

References

Young-Min Bae, Ling Zheng, Jeong-Eun Hyun, Kyu-Seok Jung, Sunggi Heu, and Sun-Young Lee. Growth Characteristics and Biofilm Formation of Various Spoilage Bacteria Isolated from Fresh Produce. *Journal of Food Science*, 79(10):M2072–M2080, 2014. ISSN 1750-3841. doi: 10.1111/1750-3841.12644.

Joey R. Bernhardt, Jennifer M. Sunday, and Mary I. O’Connor. Metabolic Theory and the Temperature-Size Rule Explain the Temperature Dependence of Population Carrying Capacity. *The American Naturalist*, 192(6):687–697, December 2018. ISSN 0003-0147. doi: 10.1086/700114.

Ana Paula Rosa da Silva, Daniel Angelo Longhi, Francieli Dalcanton, and Gláucia Maria Falcão de Aragão. Modelling the growth of lactic acid bacteria at different temperatures. *Brazilian Archives of Biology and Technology*, 61, October 2018. ISSN 1516-8913, 1678-4324. doi: 10.1590/1678-4324-2018160159.

Liane Aldrighi Galarz, Gustavo Graciano Fonseca, and Carlos Prentice. Predicting bacterial growth in raw, salted, and cooked chicken breast fillets during storage. *Food Science and Technology International*, 22(6):461–474, September 2016. ISSN 1082-0132. doi: 10.1177/1082013215618519.

C. O. Gill and K. M. DeLacy. Growth of Escherichia coli and Salmonella typhimurium on high-pH beef packed under vacuum or carbon dioxide. *International Journal of Food Microbiology*, 13(1):21–30, May 1991. ISSN 0168-1605. doi: 10.1016/0168-1605(91)90132-9.

J. D. Phillips and M. W. Griffiths. The relation between temperature and growth of bacteria in dairy products. *Food Microbiology*, 4(2):173–185, April 1987. ISSN 0740-0020. doi: 10.1016/0740-0020(87)90033-5.

Norman G. Roth and Robert B. Wheaton. Continuity of psychrophilic and mesophilic growth characteristics in the genus arthrobacter. *Journal of Bacteriology*, 83(3):551–555, March 1962. doi: 10.1128/jb.83.3.551-555.1962.

K Sivonen. Effects of light, temperature, nitrate, orthophosphate, and bacteria on growth of and hepatotoxin production by Oscillatoria agardhii strains. *Applied and Environmental Microbiology*, 56(9):2658–2666, September 1990. doi: 10.1128/aem.56.9.2658-2666.1990.

C. J. Stannard, A. P. Williams, and P. A. Gibbs. Temperature/growth relationships for psychrotrophic food-spoilage bacteria. *Food Microbiology*, 2(2):115–122, April 1985. ISSN 0740-0020. doi: 10.1016/S0740-0020(85)80004-6.

P. Stoica and Y. Selen. Model-order selection: A review of information criterion rules. *IEEE Signal Processing Magazine*, 21(4):36–47, July 2004. ISSN 1558-0792. doi: 10.1109/MSP.2004.1311138.

M. H. Zwietering, J. C. de Wit, H. G. A. M. Cuppers, and K. van 't Riet. Modeling of Bacterial Growth with Shifts in Temperature. *Applied and Environmental Microbiology*, 60(1):204–213, January 1994. doi: 10.1128/aem.60.1.204-213.1994.