Phenomenological models fit logistic growth data better than mechanistic models

Computational Methods in Ecology and Evolution MRes

Miniproject

Yewshen Lim

Department of life Sceinces

Imperial College London

y.lim20@imperial.ac.uk

Wordcount: 2109

Abstract

Under optimal conditions, the growth of bacterial populations exhibit a consistent pattern. Models have been developed to describe the growth curves. There are two broad categories of models, phenomenological and mechanistic. Here we apply three phenomenological and one mechanistic models on the empirical bacterial growth data to compare fits between the two groups and within the mechanistic group. All models were able to fit onto more than 85% of the data subsets. The modified Gompertz model was the best fit 45% of the data subsets. The cubic and logistic models were best fit for 22% and 23% of the data subsets respectively. The Baranyi model best fit only about 10% of the data subsets. Goodness-of-fit tests show that the phenomenological models fit much better than the mechanistic model, up to 90% of the data subsets. None of the models differ in performance on data at different temperatures and growth mediums. Even though they differ in performance, phenomenological and mechanistic models should still be used in conjunction, rather than against each other when trying to understand complex systems.

1 Introduction

Under optimal conditions, the growth of bacterial populations exhibit a consistent pattern. They can usually be described in three parts, a lag phase, an exponential phase then a stationary phase (figure 1) (McKellar and Lu 2004), though there is a final part, the mortality phase which is often left out because it is of little concern. In the lag phase, the abundance of resources trigger the activation of transcription in the individual cells which prepares the cells for growth. These transcriptional mechanisms increase nutrient uptake and adjust metabolic activity. When the cells are ready, the curve proceeds to the exponential phase, which corresponds to the constant rate of division of the cells, and doubles the population with every subsequent generation. Finally, the curve proceeds into the stationary phase when the population reaches the carrying capacity of the resources, and the growth rate slows to a halt. Bacterial growth is heavily investigated in food sciences, for its role in food spoilage, where food poisoning a major risk (Peleg and Corradini 2011). Models have been developed to describe the growth curves. These models for bacterial growth have evolved from plotting straight lines along the exponential phase to being able to describe the entire curve. However, no single model is able to best fit every dataset. There exists inherent variation from the different factors involved, many of which might still not be understood.

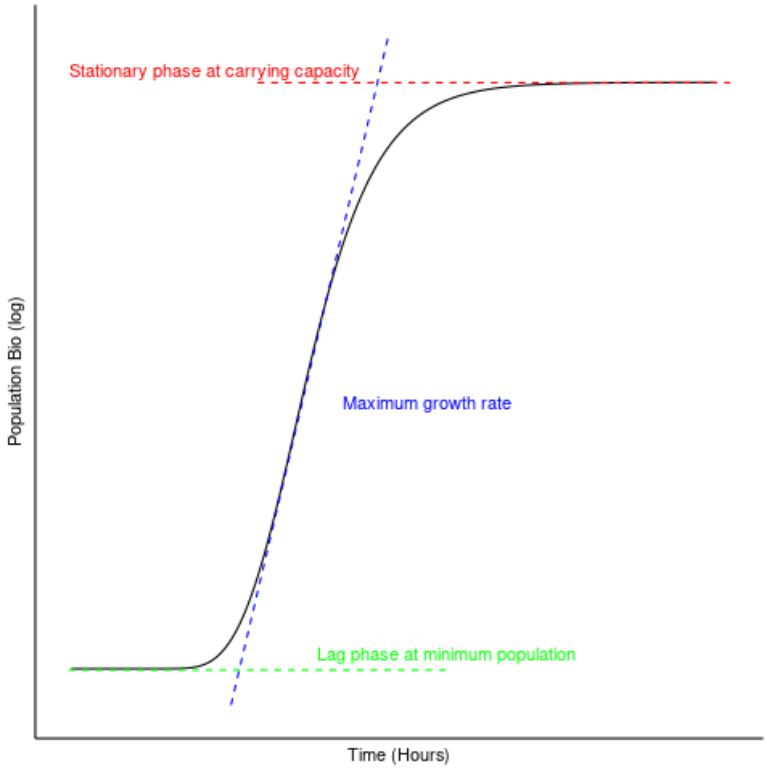

Figure 1: A growth curve of population against time

There are two broad categories of models, phenomenological and mechanistic. Phenomenological models are purely mathematical, the parameters have no biological basis and they describe the statis-

tically significant and non-random patterns or phenomena in empirical data. Mechanistic models have biological meaning, where the parameters are obtained from the data itself and relate to the processes involved.

Here we apply three phenomenological and one mechanistic models on the empirical bacterial growth data to compare fits between the two groups and within the mechanistic group. The models are a cubic polynomial, the logistic (Verhulst) model (McKendrick and Pai 1912), the modified Gompertz model (Zwietering et al. 1990) and Baranyi model (Baranyi and Roberts 1994).

Cubic polynomial:

$$N = ax^3 + bx^2 + cx + d \tag{1}$$

A phenomenological model, the cubic polynomial equation with parameters *a, b, c* and *d*. These parameters have no biological meaning and are simple mathematically coherent.

Logistic (Verhulst) model:

$$N_t = \frac{N_0 * N_{max} * e^{r_{max} * t}}{N_{max} + N_0 * (e^{r_{max} * t} - 1)} \tag{2}$$

A phenomenological model, the logistic (Verhulst) model involves the following parameters: minimum population N_0, maximum population N_{max} and maximum growth rate r_{max}. Even though these parameters sound like they do have biological meaning, they do not.

Modified Gompertz model:

$$N_t = N_0 * (\frac{N_{max}}{N_0})^{e^{-e^{\frac{e^1 * r_{max} * (t_{lag} - t)}{log \frac{N_{max}}{N_0}} + 1}}} \tag{3}$$

A phenomenological model, the modified Gompertz model involves an additional parameter, the duration of lag phase t_{lag}.

Baranyi Model:

$$h = r_{max} * t_{lag} \tag{4}$$

$$A = t + \frac{1}{r_{max}} * log(e^{-r_{max} * t_{lag}} + e^{-h} - e^{-r_{max} * t - h}) \tag{5}$$

$$N_t = N_0 + r_{max} * A - log(1 + \frac{e^{r_{max} * A} - 1}{e^{N_{max} - N_0}}) \tag{6}$$

A mechanistic model, the Baranyi model shares the same parameters as the modified Gompertz model. However, h is a parameter specifying the initial physiological state of the organism which influences the duration of lag phase.

2 Methods

2.1 Data

Empirical datasets from ten different publications were analysed. The dataset (LogisticGrowthData.csv) and its metadata (LogisticGrowthMetaData.csv) are available here. This collection of experiments contain bacterial growth data from 45 different species (including subspecies and variants), grown at 17 different temperatures on 18 different mediums.

2.2 Computing Tools

2.2.1 Data exploration and preparation

Python v3.8.6 was used to perform data exploration and preparation. The package *pandas v.1.1.4* allowed for easy manipulation of the dataset in a dataframe format. The dataset was organised into different IDs using the combination of species name, temperature (in degrees celcius), medium they were grown in, replicates (within each publication) and their sources. This produced a total of 299 different unique IDs. A minimum of five datapoints are required to fit the models, the unique ID data subsets were screened and subsets with less than five datapoints were removed.

2.2.2 Model Fitting and Analysis

R v.4.0.2 was used to perform the model fitting and analysis. The package *minpack.lm v.1.2* provided a simple and straightforward way to fit the models. A loop was used to perform the following steps on every unique ID. First, the population numbers were $\log_{10}$ transformed. The initial starting values, N_0, N_{max}, r_{max} and t_{lag} were then extracted. A function was written to generate a sample of 100 normally distributed values with a standard deviation of one around each of the initial starting values. These sampled values were then passed to each model as the starting parameters for optimisation. To prevent the models from arriving at mathematically suitable but incorrect values, the parameters were also bounded, within the ranges as shown below:

Minimum population: $(N0 - |2 * N0|) \cup (N0 + |2 * N0|)$
Maximum population: $(Nmax - |2 * Nmax|) \cup (Nmax + |2 * Nmax|)$
Maximum growth rate: $(rmax - |1.5 * rmax|) \cup (rmax + |1.5 * rmax|)$
Lag phase duration: $(tlag - |2 * tlag|) \cup (tlag + |2 * tlag|)$

The non-linear least square (NLLS) method, *nls.lm* from *minpack.lm*, was used to fit these four models.

From the 100 starting values, each model is fitted up to 400 times and is optimised by minimising the residual sum of squares (RSS). All the successful fits were recorded to a separate dataframe for each model. The parameters generated by individual best fits were also recorded for plotting the fitted curves. The outputs from each model were used to calculate the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and RSS. The individually best fitting model's starting values, AIC, BIC and RSS were selected and recorded for comparison between models. A cubic polynomial fit was also performed using the linear model (lm) method from the same package. The lowest AIC was used to select the best fitting model (Johnson and Omland 2004) for each unique ID. The best fitting curve for each unique ID were then plotted over the data points for visualisation.

2.3 Visualisation and report writing

Visualisation of plots were all performed using the package *ggplot2* in *R* so as to directly use the results from the previous analysis script. The report was written in LaTeX as it allows fine control of typesetting.

2.3.1 Project execution script

Bash was used to write and execute the above scripts into a clear and reproducible workflow. Bash was chosen due to the ease of executing the *R*, *Python* and compiling *LaTeX* scripts altogether.

3 Results

3.1 Model fits

Figure 2 shows that all models were able to fit onto more than 85% of the data subsets. The phenomenological cubic model fit the most number of data subsets at 98.9%, while the other two phenomenological logistic and modified Gompertz, and mechanistic Baranyi model fit to a similar number of data subsets, at 90.9%, 90.9% and 87.6% respectively.

Figure 3 shows the percentage of best fits for each model on the data subsets. The modified Gompertz model was the best fit for close to half of the data subsets at 45%, much higher than the other three models. The cubic and logistic models were the best fit for a similar number of data subsets at 22% and 23% respectively. The Baranyi model performed the worst, best fitting only about 10% of the data subsets. Both G-test of goodness-of-fit and chi-square goodness-of-fit tests show that in this

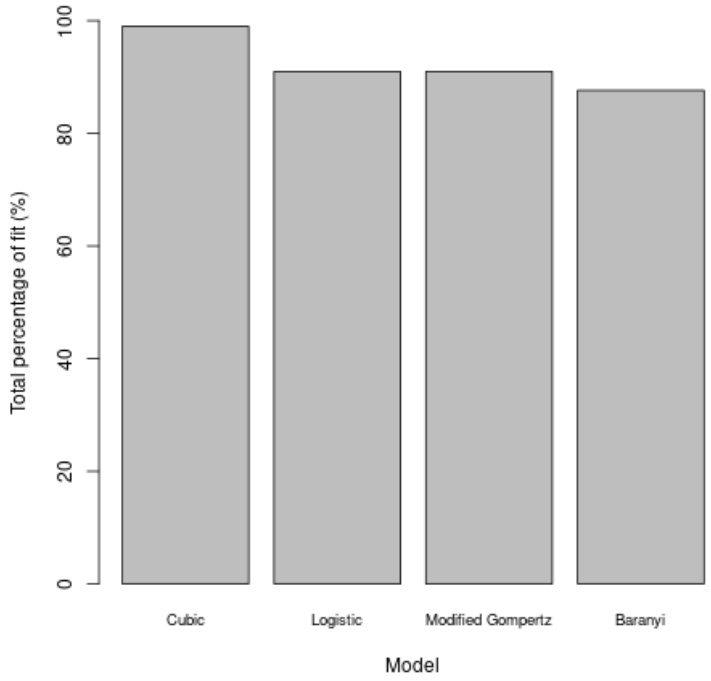

Figure 2: Percentage of fitting success of each model onto the data subsets

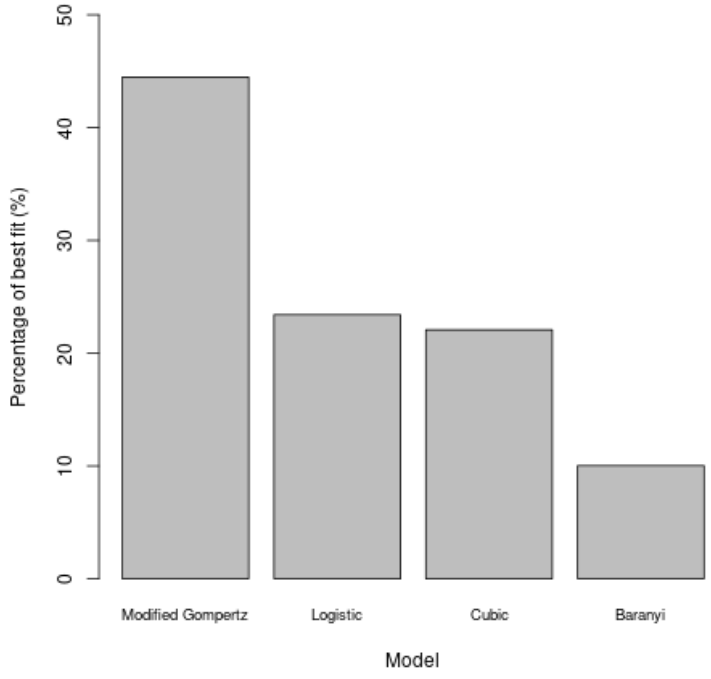

Figure 3: Percentage of best fitting model on the data subsets

study, the phenomenological models fit much better than the mechanistic model ($p < 2.2 \times 10^{-16}$ for both), where they best fit 90% of the data subsets.

3.2 Data subsets

The individual IDs amount to 299 different subsets. Table one shows the proportion of fits by the different temperatures. Temperatures ranged from 0 to 30 degrees celcius and were binned into three different levels, from 0 to 10, 11 to 20 and more than 21 degrees celcius. G-test of goodness-of-fit shows that none of the models differ in performance on data at different temperatures ($p > 0.05$). From the six plots in figure four, we can see that when the temperature increases from 2°C to 20°C, the time taken to reach maximum population shortens from more than 500 hours to more than 50 hours. Visually, the plots also perform similarly across the six different temperatures.

	Temperature Bins (degress celcius)	Baranyi	Cubic	Logistic	Modified Gompertz
1	1-10	10.1% (12)	24.4% (29)	21.8% (26)	43.7% (52)
2	11-20	10.2% (10)	21.4% (21)	22.4% (22)	45.9% (45)
3	21-30	9.8% (8)	19.5% (16)	26.8% (22)	43.9% (36)
4	Total	10.0% (30)	22.1% (66)	23.4% (70)	44.5% (133)

Table 1: Fit proportions of models on the three different temperature bins.

Table two shows the proportions of fits by the different mediums. Mediums were binned into solid and liquids. Exact test of goodness-of-fit shows that none of the models differ in performance on either solid or liquid mediums ($p > 0.05$).

	Medium State	Baranyi	Cubic	Logistic	Modified Gompertz
1	Liquid	12.2% (19)	20.5% (32)	29.5% (46)	37.8% (59)
2	Solid	7.7% (11)	23.8% (34)	16.8% (24)	51.7% (74)
3	Total	10.0% (30)	22.1% (66)	23.4% (70)	44.5% (133)

Table 2: Fit proportions of models on the two different states of mediums.

4 Discussion

4.1 Mechanistic vs Phenomenological models

From our results, we can conclude that overall, phenomenological models perform better than the mechanistic model. This was expected because phenomenological models provide the fundementals, and identify patterns which mechanistic models are then built upon. The biologically meaningful parameters are also likely the reason behind mechanistic fitting worse than phenomenological ones, where the non biological parameters are less restricted by assumptions and undertainty (Bokulich 2011; Chowell et al. 2016; Eskola and Parvinen 2007) and possibly due to insufficient understanding

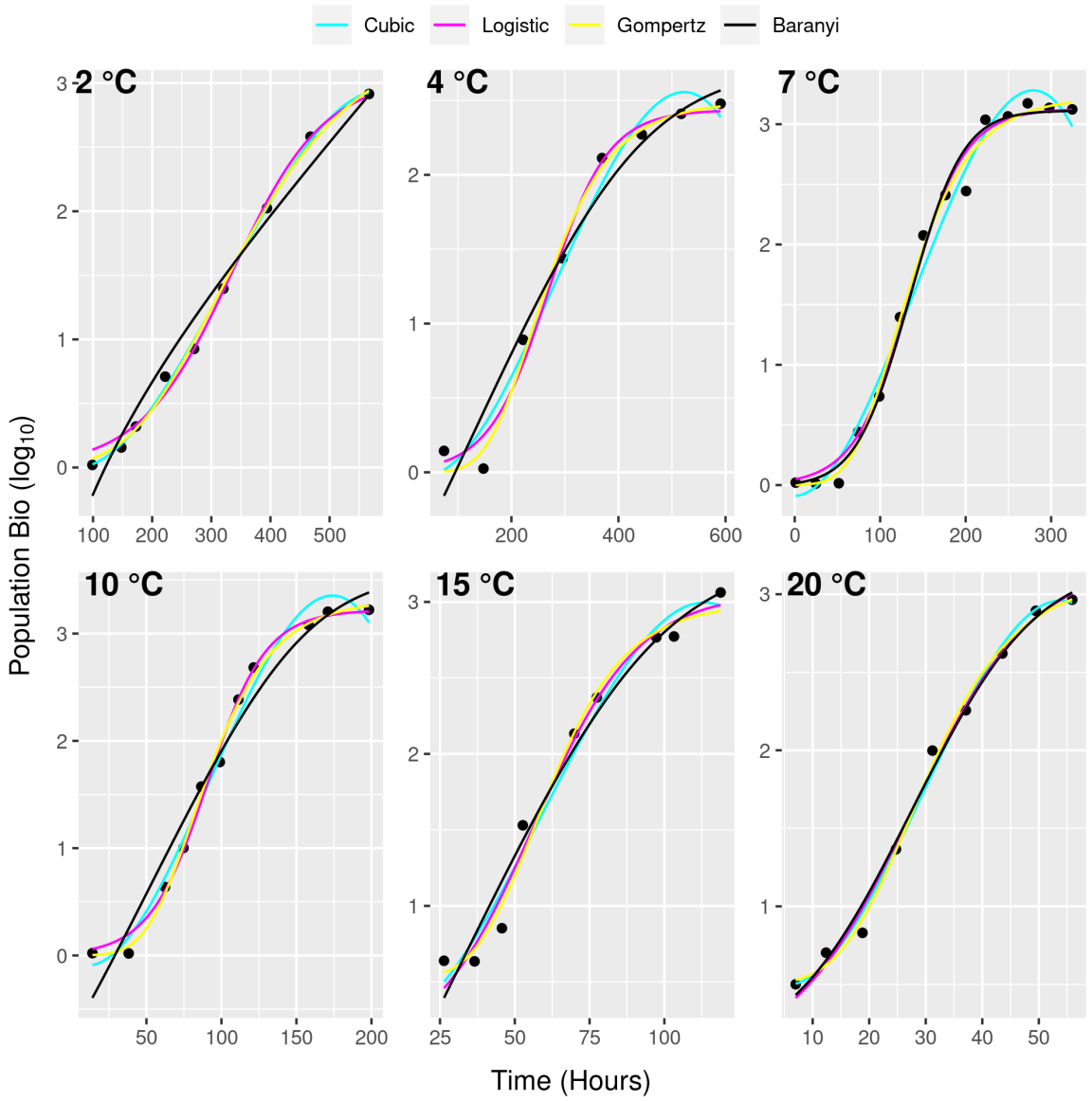

Figure 4: Model fits on *Staphylococcus spp.* at six different temperatures

of bacterial growth mechanisms biologically. Phenomenological models are used to generate hypotheses, which then mechanistic models can be developed to validate. Mechanistic models perform better because they seek to be mathematically relevant, thus they do not identify how the underlying mechanisms work. The mechanistic Baranyi model performed the worst out of all the models. This could mean that the Baranyi model requires much more improvement, as it failed to perform on these empirical datasets. Another consideration is that the Baranyi model might be highly specific, which means it would not fit best on most of the data subsets, but this also points to the Baranyi model not being a suitable general model. These results also show that if we were to only compare between the modified Gompertz and cubic model, these two models perform similarly. However, if any of the phenomenological models were compared with the Baranyi model, then there would be a significant difference in performance. This puts emphasis on using and comparing more than one model, and both types of models for analysis, so as to not be restricted by the performance of any single model.

4.2 Temperature and growth medium

In our analysis, temperature and medium does not affect the performance of the models on the dataset. This could be attributed to the two different factors not affecting the general shape of the bacterial growth curve. Temperature (within survival ranges) could affect the growth rate, but a lower or higer growth rate only changes the gradient of the growth phase, and does not affect the lag or stationary phase, which can be seen from figure four, where the increase in temperatures shorten the time required for the population to reach the maximum.

5 Conclusion

In conclusion, phenomenological models perform better than mechanistic models. However, this should not discount the use of mechanistic models, as they are the ones which validate the hypotheses generated by the phenomenological models. Phenomenological are also able to fit better due to the possibility of taking on additional parameters (since they are arbitrary), resulting in them being able to fit better than mechanistic models. For example, a polynomial model with up to x^n will be able to fit a broad range of data mathematically. The lack of biological meaning stops additional parameters from being developed as they do not end up providing any insight. Phenomenological and mechanistic models should be used in conjunction, rather than against each other when systems get increasingly complicated (McMeekin et al. 2013). Although this analysis showed that temperature and growth medium does not affect performance of fits, that is not a conclusion, and it is recommended to always consider multiple models across temperatures and growth medium. Additionally, although model comparison often use AIC and BIC as a criteria, they are not without limitations. It is also recommended to explore alternatives to supplement AIC and BIC for model comparison.

References

Baranyi, József and Terry A Roberts (1994). "A dynamic approach to predicting bacterial growth in food". In: *International journal of food microbiology* 23.3-4, pp. 277–294.

Bokulich, Alisa (2011). "How scientific models can explain". In: *Synthese* 180.1, pp. 33–45.

Chowell, Gerardo et al. (2016). "Using phenomenological models to characterize transmissibility and forecast patterns and final burden of Zika epidemics". In: *PLoS currents* 8.

Eskola, Hanna TM and Kalle Parvinen (2007). "On the mechanistic underpinning of discrete-time population models with Allee effect". In: *Theoretical Population Biology* 72.1, pp. 41–51.

Johnson, Jerald B and Kristian S Omland (2004). "Model selection in ecology and evolution". In: *Trends in ecology & evolution* 19.2, pp. 101–108.

McKellar, Robin C and Xuewen Lu (2004). *Primary models*. Boca Raton, Florida: CRC Press.

McKendrick, A. G. and M. Kesava Pai (1912). "XLV.—The Rate of Multiplication of Micro-organisms: A Mathematical Study". In: *Proceedings of the Royal Society of Edinburgh* 31, 649–655.

McMeekin, Tom et al. (2013). "Predictive microbiology theory and application: Is it all about rates?" In: *Food Control* 29.2, pp. 290–299.

Peleg, Micha and Maria G Corradini (2011). "Microbial growth curves: what the models tell us and what they cannot". In: *Critical reviews in food science and nutrition* 51.10, pp. 917–945.

Zwietering, MH et al. (1990). "Modeling of the bacterial growth curve". In: *Applied and environmental microbiology* 56.6, pp. 1875–1881.