HOW WELL MECHANISTIC MODELS PERFORM IN BACTERIAL POPULATION GROWTH AND HOW TEMPERATURES AFFECT THE MODEL SELECTION

Chalita Chomkatekaew

MSc Computational Methods in Ecology and Evolution

December 2021

Contents

1 Abstract

Microbial organisms are the most abundant organism on the planet. They have numerous implications for many scientific fields, for instance, food safety, biogeochemical conversion, and diseases. However, bacterial growth is yet understood. In the current study, I compared and evaluated the performance of four mathematical models; modified Gompertz, Logistic and two polynomial models to a large dataset of bacteria growth curves. The most frequently model that performed best is the Gompertz model, followed by Logistic, Quadratic and Cubic models, respectively. Different temperatures also result in different model choices where the Gompertz model is preferred in two extreme temperature ranges. Taken together, mechanistic models, particularly the Gompertz model, have the potential to describe and forecast different types of the growth curve, where others perform better in specific growth conditions. It is thus important to note that all models should be used in conjunction with cautions.

Introduction

Prediction of bacterial fitness and dynamics is an imperative tool to understand a role of which microbes in different conditions, ranging from the biogeochemical processes (Schlesinger 1997), food spoilage (Odeyemi et al. 2020), to host-health status and diseases (Feng et al. 2020). In a resource-limited environment, a normal bacteria growth curve is described in 4 stages; "Lag phase", "Exponential phase", "Stationary phase", and "A mortality phase" (Figure 1). The lag phase is where the bacteria prepare the machinery for proliferation, followed by the exponential phase in which bacteria divide at a constant rate. These two initial phases, thus, give a sigmoidal semi-logarithm curve. The bacteria growth then reaches a plateau once the carrying capacity has been reached. Lastly, the death phase is characterised by the exponential decrease in the bacteria cells (McKellar and Lu 2004). However, the death phase is rarely considered in the mathematical models for bacterial growth kinetic, for instance, in the food industry, most food would have been spoiled before the abundance of bacteria reaches the exponential and/or stationary phase. In contrast, bacteria can proliferate in a range of ecological factors. Where temperature is one of the most important factors to control bacteria growth dynamics. Of which the higher temperature could affect the duration of lag phase and growth rate (Bronikowski, Bennett, and Lenski 2001). Understanding how temperature affects bacteria growth and the ability to accurately predict the consequences are vital in bacteria adaptations, particularly in food storage and disease severity.

An example of the idea growth curve
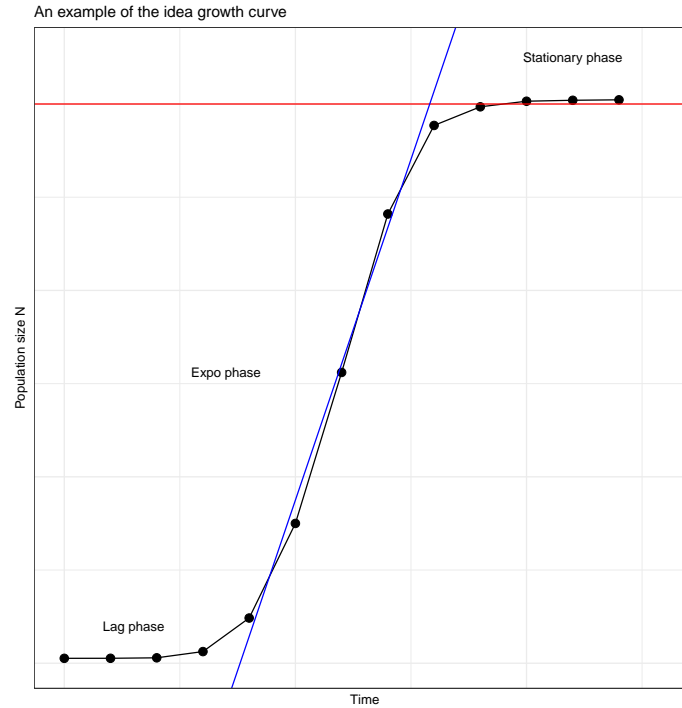
Figure 1: The bacterial population growth in a close-system. Denote the mortality phase was not shown here.

Mathematical modelling has become more widely used to understand large and complex mechanisms in biological systems as well as in bacteria growth dynamics (For review, Jong et al. 2017). There are two types of models; Mechanistic and phenomenological models. Mechanistic models have biological definitions and attempt to elucidate the mechanisms of the empirical data, however, they require some starting values to fit the models. Whereas the latter type is seeking to describe the relationship between the variables instead with minimum parameters used. Both types of models have been proposed to predict the bacterial growth dynamics, however, mechanistic models such as the Gompertz model are preferred over the phenomenological models owing to the direct inference with biological implications the model can provide (Peleg and Corradini 2011).

In this study, I leveraged four existing mathematical models, two mechanistic models; Gompertz (Zwietering et al. 1990)and Logistic models (Peleg 1997), and two phenomeno-logical models (Quadratic and Cubic polynomial models) to assess the model fitness to the published dataset of 285 growth curves as a proof of concept study. Additionally, I aimed to test for correlation between the models and the temperature as a growth condition in an attempt to deduct model preference in each temperature range.

Method and Material

Data

The dataset was collected from 10 previously published papers downloaded from a Github repository (mhasoba 2021). The data includes the growth curves from 45 unique species, and cultured in 17 different temperatures and 18 mediums.

Computing Tools

Data preparation

Data preparation were performed using based *R version 4.1.1* and *Tidyverse* package *version 1.3.1.* The subset data of a single growth curve were characterised using a unique combination of temperature, species, medium, number of replications, and the literature citation. The time points and population abundance that less than zero were removed as they are biologically impossible, followed by the transformation of the population abundance into a log scale.

Mathematical Models

Mathematical models used in this study are defined in R by the following:

Mechanistic Model - Non-linear Least squares

Modified Gompertz model (Zwietering et al. 1990)

$$log(N_t) = N_0 + (K - N_0)e^{-e^{r_{max}exp(1)\frac{t_{lag}-t}{(K-N_0 log(10))}+1}} \tag{1}$$

Logistic (Verhurst) model (Peleg 1997)

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

68 Where N_0 is the initial population abundance. r_{max} is the maximum growth rate and K

69 is the carrying capacity or the maximum abundance the population can have. While N_t

70 is the population size at time t. Additionally, t_{lag} is the time taken before, in this study,

71 bacterial population transition to the exponential phase. Note that only the Gompertz

72 model considers t_{lag} as a parameter.

73 **Phenomenological Model**

74 Quadratic polynomial model

$$y = ax^2 + bx + c \tag{3}$$

75 Cubic polynomial model

$$y = ax^3 + bx^2 + cx + d \tag{4}$$

76 **Model fitting**

77 To fit growth models to the prepared data, I used *minpack.lm* package *version 1.2-1* in

78 R with nlsLM function to fit the Non-linear least squares models; Gompertz and Logistic

79 models. The starting values were estimated and then sampled to fit the model for 100

80 permutations to increase the likelihood for more robust and optimised estimation. I then

81 calculate the Akaike Information Criterion (AIC) of each model permutations to select

82 the optimised parameters with minimum AIC value.

$$AIC = -2log(L(\hat{\theta})) + 2K \tag{5}$$

83 Where $log(L(\hat{\theta}))$ is the maximum log-likelihood of the estimated value and K is the number

84 of parameter in each model.

85 The linear regression models (Quadratic and Cubic polynomial models) were fitted to

86 each subset data using lm function of based R $stats$ package, followed by AIC calculation

87 (Equatuion 5). All the selected parameters and its AIC values were then stored and saved

88 in a file for further analyses. Visualised inspections were also performed to manually check

89 the model performance (Figure S1).

Model selection and analyses

91 The second order derivative of AIC (AICc) were then calculated from AIC obtained

92 previously to correct the bias of the sample size. The model with the most minimum

93 AICc was chosen in each subset data as the best model. I also calculate the Bayesian

94 information criterion or BIC to validate the selected models. To infer the probability

95 of the best model is better than other model tested, Weight Akaike of each model in

96 respective to the best model were calculated. In addition, I then calculated the normalised

97 probability between the best model and second best to find the likelihood of the model

98 chosen (Burnham and Anderson 2002).

$$AICc = AIC + \frac{2K(K+1)}{n-K-1} \tag{6}$$

99 Different between the models

$$\Delta AICc_i = AIC_i - AIC_{min} \qquad (7)$$

100 Weight Akaike

$$W_i(AIC) = \frac{exp(-\frac{1}{2}\Delta AICc_i)}{\sum_{r=1}^{R} exp(-\frac{1}{2}\Delta AICc_R})} \qquad (8)$$

101 Normalised probability

$$Prob = \frac{W_{best}(AIC)}{W_{second}(AIC) + W_{best}(AIC)} \qquad (9)$$

102 **Temperature and model selection analyses**

103 I dissect the effect of temperatures to model selection. The temperatures are grouped

104 into 4 ranges; 0-10 °C, 11-20 °C, 21-30 °C, 31-40 °C. The mean AICc and percentage of

105 each model selected in different temperature bins were calculated and visualised.

106 **Data visualisation and report**

107 Data visualisation was performed in R using ggplot2 package. The report was written in

108 LaTex.

109 Data availability and reproducible

110 The source code and data for this study analyses are available at https://github.com/ChalitaCh/

111 under CMEECourseWork repository. The instruction to run the code for reproducible

112 analyses is also available at the provided repository.

Results

Mechanistic models are preferred

The mathematical models were used to fit the unique 285 growth curve to evaluate the fitness of the models across the dataset. Overall, 90% (n = 256) of the growth curves were able to successfully fit the dataset and used for model selection. Having calculated the AICc score to select the best model, both mechanistic models, Logistic and Gompertz models, generally performed best in comparison with the phenomenological models given the tested dataset (Table 1 and Figure 2A). While the frequency of the best model chosen between the Gompertz and Logistic models are virtually the same, the average AICc score of Gompertz model was the lowest.

Similarly, the average BIC revealed the parallel outcomes by the Gompertz model had the lowest mean BIC score as well as it was selected as the best model in 154 growth curves (57.9%). Even though the Logistic model, which had the second-lowest BIC score, the frequency of the model selected as the best model is almost halved in comparison to the model selection using AICc scores (Table 1 and Figure 2A). While the Cubic polynomial model performed the worst out of the 4 models tested when evaluated with the AICc, the Cubic model satisfied the dataset better than the Quadratic model, thus, ranked as the third-best model with BIC scores.

Models	N.AIC	N.BIC	MeanAICc	MeanBIC	MeanProb	MinProb	MaxProb
Cubic	14	39	-4.282	-11.582	0.753	0.501	1
Gompertz	107	154	-20.608	-27.265	0.899	0.506	1
Logistic	104	57	-16.976	-18.344	0.783	0.505	1
Quadratic	31	16	-1.119	-3.015	0.769	0.502	1

Table 1: The overall selected models with minimum AICc and BIC values, The normalised probability of W(AICc) of each model denoted as MeanProb also illustrated here.

High likelihood of selected model is the best model

The normalised probability using the Akaike Weight was also performed to assess the likelihood of the selected model is the best model. Consistent with previously observed in the AICc value, the Gompertz model has the highest mean likelihood ($P = 0.90$) to fit the empirical subset data. Interestingly, despite having fitted to only 17.2 % of the fitted dataset, the Quadratic and Cubic polynomial models were able to successfully support the data with the mean normalised probability of 0.748 and 0.716, respectively (Figure 3B).
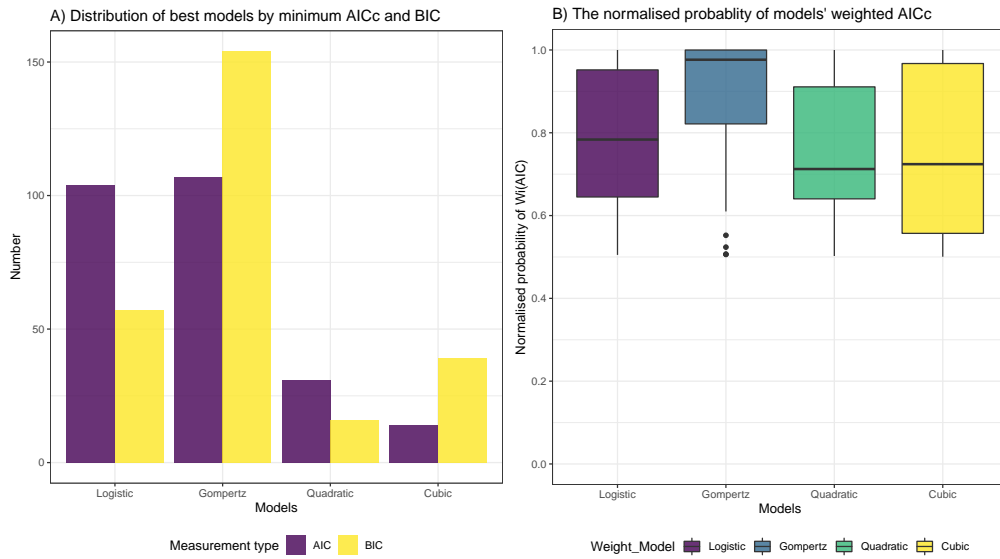
Figure 2: The overall percentage of chosen models and the normalised probability range of each model. The higher the probability is, the likelihood the model is the best model

Lower temperature range affects the model selection

Here, I investigated the effect of temperature on the choice of models based on their goodness of fit. Having grouped the dataset into different temperature bins, the average AICc values of Gompertz model performed best in three temperature ranges. However, the frequency of Gompertz model selected is only in 0-10 °C and 31-40 °C. Whereas the Logistic model fitted the growth curve better in a temperature range of 11 – 20 °C with an average AICc score of -12.28 when compared between the tested models (Table 2). Intriguingly, the Logistic model is the most model chosen in 11-30 °C (n = 67 out of 126). Despite fewer number of fit to the growth curves, both polynomial models appear to perform better in the temperature at 0-10 °C in comparison to other temperature ranges (Figure 3).

Temperature	N	MeanGAICc	MeanLAICc	MeanQAIC	MeanCAIC
0-10	100	-25.661	-17.133	-7.901	-13.695
11-20	85	-10.729	-12.28	3.311	1.7
21-30	41	-17.483	-16.856	3.754	3.757
31-40	30	-36.024	-29.924	2.273	-0.84

Table 2: The average AICc and BIC values of models tested between the temperature bins.
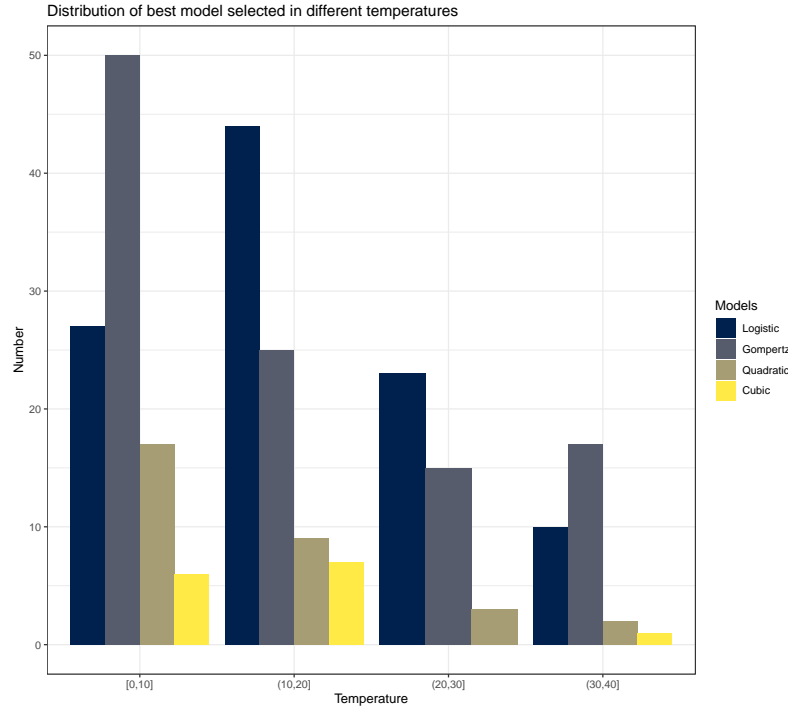
Figure 3: The percentage of model with best performance in each temperature range

Discussion

Bacterial population growth is a vital mechanism to varieties of scientific fields, where understanding the activity and dynamics of the population could provide insights into many biological questions. However, the growth mechanisms in bacteria are highly complex and not fully understood. Computation modelling and analyses have become increasingly desirable approaches to understand this process where the laboratory experiments are la-

borious and difficult to mimic the environment in nature. Here, I dissected four different mathematical models to evaluate the performance and fitness of a large dataset of 285 growth curves. In addition, I attempted to assess the effect of temperature ranges on the model selection.

The models were able to successfully fit to about 90 % of the dataset, where Gompertz model fits more accurately in the majority of the dataset (41-58 %), followed by the Logistic model. This is consistent with the model selection using AICc and BIC values. The normalised probability of the model was also calculated to illustrate the confidence of chosen models. (Figure 2) While both polynomial models were able to fit only 13.4 - 21.1 % of the dataset. The trend of AICc and BIC between the two models are inconsistent. Although both accounting for the goodness of fit, the complexity of the model, as well as the sample size, the development framework of these measurements are different while the BIC penalty term is larger (Wagenmakers and Farrell 2004). Further investigation is needed to understand the discrepancy between them.

In concordance with a previous report (Peleg and Corradini 2011), both Gompertz and Logistic models are preferred than the phenomenological models as 1) its parameters have biological implication rather than seeking to fit the curve to the dataset 2) Both models have an "S-shaped" feature that can accommodate the slow growth of bacteria during the transition between the lag phase and exponential phase. Not only having performed well in the model fitting for bacteria population dynamics, but the Gompertz model is also widely implicated in many biological studies, including the prediction of the tumour cells growth (Benzekry et al. 2014). In contrast, the defined Logistic model (Equation 2) does not account for the long lag time and, thus, may affect the model fitness to growth

13

curve in bacteria with a slow growth rate. Several researchers attempt to develop the modified logistics regression to account for the long delay before the exponential phase (Oscar 2005; Koseki and Nonaka 2012), however, such modification could be costly to the parameters estimated.

Temperature is an important factor to modulate the bacteria growth dynamics, particularly to the bacterial physiological. In this study, I briefly investigated the consequences of temperature on model selection. Given the test dataset, Gompertz model described the data best at the upper and lower temperature ranges. On the other hand, Logistic model is more frequently (58 %, n = 126) selected as the best model in a temperature range of 11-30 °C (Figure 3). At lower temperatures, the enzyme kinetic is slower, resulting in a longer doubling time. The extreme example is from a bacterium isolated from Siberian permafrost which doubled in the abundance after culture for 39 days (Bakermans et al. 2003). With the defined function of logistic model, we can thus infer that it is challenging for the model to describe the data sub-optimal temperature for bacterial growth as above-mentioned. However, further investigation to understand the model choice in other temperature ranges is limited as the majority of data is biased towards the lower temperatures.

Conclusion

In conclusion, the Gompertz model is preferred and has the best goodness of fit to the given bacterial population data. Whereas the logistic and phenomenological models have the potential to perform well in sub-types of the growth curve. Thus, suitable model selection to various growth trajectories in bacteria is crucial. Universal model development for

14

growth curve to account for multiple factors, such as gene expression and host status, is needed to accurately predict the outcome in a 'real world' setting.

References

Bakermans, Corien et al. (2003). "Reproduction and metabolism at - 10°C of bacteria isolated from Siberian permafrost". en. In: *Environmental Microbiology* 5.4. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1046/j.1462-2920.2003.00419.x, pp. 321–326. ISSN: 1462-2920. DOI: `10.1046/j.1462-2920.2003.00419.x`. URL: `https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1462-2920.2003.00419.x` (visited on 12/02/2021).

Benzekry, Sébastien et al. (Aug. 2014). "Classical Mathematical Models for Description and Prediction of Experimental Tumor Growth". en. In: *PLOS Computational Biology* 10.8. Publisher: Public Library of Science, e1003800. ISSN: 1553-7358. DOI: `10.1371/journal.pcbi.1003800`. URL: `https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003800` (visited on 12/02/2021).

Bronikowski, Anne M., Albert F. Bennett, and Richard E. Lenski (2001). "Evolutionary Adaptation to Temperature. Viii. Effects of Temperature on Growth Rate in Natural Isolates of Escherichia Coli and Salmonella Enterica from Different Thermal Environments". en. In: *Evolution* 55.1. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.0014-3820.2001.tb01270.x, pp. 33–40. ISSN: 1558-5646. DOI: `10.1111/j.0014-3820.2001.tb01270.x`. URL: `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.0014-3820.2001.tb01270.x` (visited on 12/03/2021).

Burnham, Kenneth P. and David Raymond Anderson (2002). *Model selection and multi-model inference: a practical information-theoretic approach.* 2nd ed. OCLC: ocm48557578. New York: Springer. ISBN: 978-0-387-95364-9.

Feng, Lihui et al. (Feb. 2020). "Identifying determinants of bacterial fitness in a model of human gut microbial succession". en. In: *Proceedings of the National Academy of Sciences* 117.5. Publisher: National Academy of Sciences Section: Biological Sciences, pp. 2622–2633. ISSN: 0027-8424, 1091-6490. DOI: `10.1073/pnas.1918951117`. URL: `https://www.pnas.org/content/117/5/2622` (visited on 12/03/2021).

Jong, Hidde de et al. (Nov. 2017). "Mathematical modelling of microbes: metabolism, gene expression and growth". In: *Journal of The Royal Society Interface* 14.136. Publisher: Royal Society, p. 20170502. DOI: `10.1098/rsif.2017.0502`. URL: `https://royalsocietypublishing.org/doi/10.1098/rsif.2017.0502` (visited on 12/03/2021).

Koseki, Shige and Junko Nonaka (Sept. 2012). "Alternative Approach To Modeling Bacterial Lag Time, Using Logistic Regression as a Function of Time, Temperature, pH, and Sodium Chloride Concentration". In: *Applied and Environmental Microbiology* 78.17, pp. 6103–6112. ISSN: 0099-2240. DOI: `10.1128/AEM.01245-12`. URL: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3416635/` (visited on 12/02/2021).

McKellar, Robin C. and Xuewen Lu, eds. (2004). *Modeling Microbial Responses in Food.* en. 0th ed. CRC Press. ISBN: 978-0-203-50394-2. DOI: `10.1201/9780203503942`. URL: `https://www.taylorfrancis.com/books/9780203503942` (visited on 12/03/2021).

mhasoba (Nov. 2021). *The Multilingual Quantitative Biologist.* original-date: 2017-10-13T06:15:34Z. URL: `https://github.com/mhasoba/TheMulQuaBio` (visited on 12/03/2021).

Odeyemi, Olumide Adedokun et al. (2020). "Understanding spoilage microbial community and spoilage mechanisms in foods of animal origin". en. In: *Comprehensive Reviews in Food Science and Food Safety* 19.2. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1541-4337.12526, pp. 311–331. ISSN: 1541-4337. DOI: 10.1111/1541-4337.12526. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1541-4337.12526 (visited on 12/03/2021).

Oscar, T. P. (Dec. 2005). "Development and validation of primary, secondary, and tertiary models for growth of Salmonella Typhimurium on sterile chicken". eng. In: *Journal of Food Protection* 68.12, pp. 2606–2613. ISSN: 0362-028X. DOI: 10.4315/0362-028x-68.12.2606.

Peleg, M. (Aug. 1997). "Modeling microbial populations with the original and modified versions of the continuous and discrete logistic equations". eng. In: *Critical Reviews in Food Science and Nutrition* 37.5, pp. 471–490. ISSN: 1040-8398. DOI: 10.1080/10408399709527785.

Peleg, Micha and Maria G. Corradini (Dec. 2011). "Microbial growth curves: what the models tell us and what they cannot". eng. In: *Critical Reviews in Food Science and Nutrition* 51.10, pp. 917–945. ISSN: 1549-7852. DOI: 10.1080/10408398.2011.570463.

Schlesinger, William H. (1997). *Biogeochemistry: an analysis of global change*. 4th ed. SanDiego: Elsevier. ISBN: 978-0-12-814608-8.

Wagenmakers, Eric-Jan and Simon Farrell (Feb. 2004). "AIC model selection using Akaike weights". en. In: *Psychonomic Bulletin & Review* 11.1, pp. 192–196. ISSN: 1531-5320. DOI: 10.3758/BF03206482. URL: https://doi.org/10.3758/BF03206482 (visited on 12/02/2021).

269 Zwietering, M. H. et al. (June 1990). "Modeling of the Bacterial Growth Curve". In:

270 *Applied and Environmental Microbiology* 56.6, pp. 1875–1881. ISSN: 0099-2240. URL:

271 `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC184525/` (visited on 12/03/2021).

Supplementary
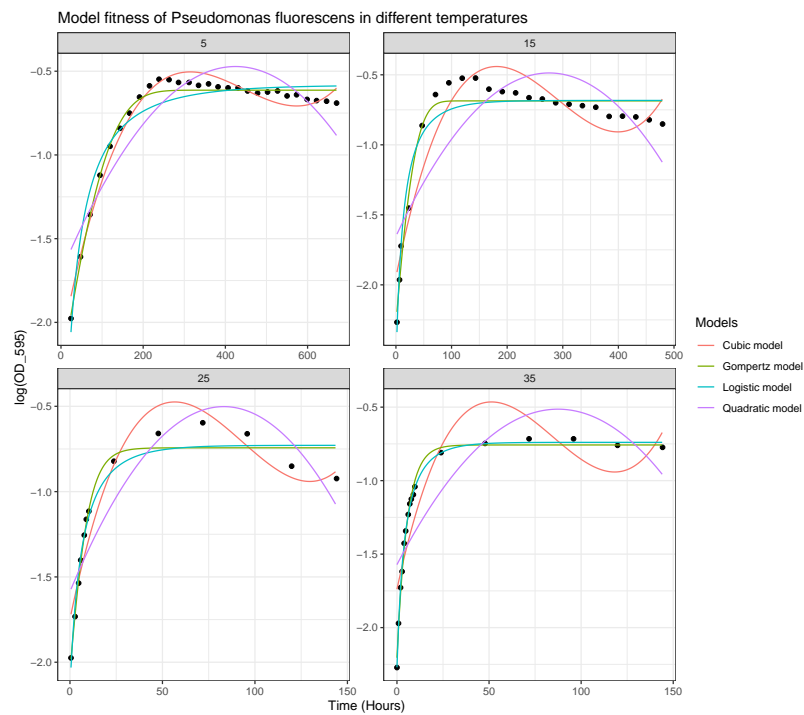
Figure S1: An example of manually inspection of model fitness in the dataset. Here is the *P. fluorescens* growth curves in different temperature range