Microbial growth and models: how

synchronizable are they?

PokMan HO

- Department of Life Sciences, Faculty of Natural Sciences,
- Imperial College London



Approximate Word Count: 2040

Microbial growth and models: how synchronizable are they?

PokMan HO (CID: 01786076)

Abstract

Phenological models have different features, assumptions and explanatory uses. Hence choosing the correct and fittest model for your experimental data is crucial. This project has used published microbial population sizes data from ten papers to do non-parametric categorical analyses (i.e. Kruskal tests) and parameter evaluations (with PCA). The result shows that Verhulst (classical) and modified Gompertz are general models while Baranyi and Buchanan are specific ones. They have unobservable yet statistical differences in parameter responses. Although phenological models are not significantly better than polynomials on data descriptions, choosing the right model can potentially help discussing the data properties in microbial population of interest.

19 Introduction

28

29

- Phenological models for microbial growth are expected to fit microbial population data. Yet due to different reasons, models developed and published from one sample may not fit the others. These reasons can be data variabilities, confounding factors, inaccurate assumptions or models being too-specific. This data-mining project is aimed at comparing and contrasting published phenological models on microbial population size data, highlighting which is a better model under what conditions. The hypotheses are:
- published phenological models are significantly better fits than polynomials for describing microbial growths;
 - appropriate phenological model(s) is/are identifiable through distinguishable microbial population size parameter values; and

- parameter values of each phenological model from successful fits are clusters with welldefined boundaries between models.
- Also, two biological hypotheses will be put forward:
- Microbial species combine with growth media is specific to certain AIC-optimal models

 (i.e. meeting the "minimal AIC +2" criteria); and
- Microbial species combine with growth media has its own parameter range when modeled.

36 Methods

- Microbial population sizes data were given, sourced from ten different publications (Table 1).
- The collection contained different microbial clades growing at various conditions for varying
- 39 times. Also, these data were recorded in multiple time and population units. Some of the
- 40 population data were direct counts while some were not.
- Experimental microbial population growth data library were divided into individual data
- subsets through six filters ("Temperature (in °C)", "Microbial clade", "growth substrate mate-
- ⁴³ rials", "experimental replicate number", "population data recording unit" and "data source").
- 44 Records with data unit "OD_595" were scaled into optical density percentages (i.e. data*100)
- to facilitate general analyses workflow. Independent (or explanatory) variable was "Time (hr)"
- and dependent (or response) variable was "population size".
- Some raw data were recorded in minutes (instead of hour). This record artifact was not cor-
- 48 rected because of two reasons: 1. shape of curves were the main concern instead of independent
- variable's scale; and 2. the unit was consistent within each data subset.

50 Model assessment

- 51 Six candidate models were assessed, four phenological and two polynomial equations. They
- were "Verhulst (classical)"², "modified Gompertz"³, "Baranyi"⁴, "Buchanan"⁵, "quadratic" and
- 53 "cubic". Non-linear least square (NLLS) method was used only on the four phenological models
- 54 and linear approach was taken for the two polynomials. Starting values selection (phenological
- models only) was described below:
- Initial (N0) and final (K) population sizes were selected to be the minimum and maximum
- 57 values of each data subset respectively. Maximum growth rate (r.max) and relative time lag

58 (t.lag) were obtained through recursive linear modelling with shrinking independent range (from 59 both maximum and minimum). In this project 5% was chosen as the shrinking threshold 60 assuming this resolution was sufficient for initiating NLLS fits. The final slope value (i.e. 61 r.max) would be a positive finite number with the highest R^2 value (i.e. best-fit slope on 62 considered data). The x-intercept for that slope was the t.lag. Time which this r.max linear 63 model intersected with K was regarded as the time achieving carrying capacity (t.K). Population 64 data was then classified into three groups (gx) according to the time: $g1 \le t.lag < g2 < t.K$ 65 $\le g3$. Inputs for phenological models were listed below (popn & time were the dependent and 66 independent variables respectively):

```
Verhulst (classical): popn = f(N0, K, r.max, time)
modified Gompertz: popn = f(N0, K, r.max, time, t.lag)
Baranyi: popn = f(N0, K, r.max, time, t.lag)
Buchanan: popn = f(N0, K, r.max, time, t.lag, gx)
```

All test starting values were than sampled from normal distribution with mean as the estimated value, standard deviation (sd) of 1, absolute the number if it was sampled as a negative.

The sd value was chosen because of different reasons for each parameters. No and K were directly extracted from the raw experimental data. Hence a small sd was precise assuming this extraction was an accurate estimate. r.max was a guesstimated value. So a large sd would be sufficient for getting the "true" value with fair accuracy. 100 trials were done as a optimal value under a trade-off between efficiency and accuracy.

AIC^{6–8} was used to select optimal parameter values within each phenological model and best candidate model(s) for a data subset. All available parameter sets were included in principal component analysis (PCA). AIC tolerance threshold was expanded to min(AIC)+2¹ to incorporate more accepted models for analyses.

79 Statistical analysis

Kruskal test was used for identify the best-fit model among all included model because the count was categorical and not assumed being normally-distributed. Pairwise Nemenyi comparisons would be carried out to identify the best test if p-value of the test was significant.

Parameter weights were assessed across phenological models by PCA R-way analysis method on natural-logged parameter data. Datasets would be expected clustering together if parameter(s) were representing the observed data. Then with Kruskal-Wallis test, each parameter was

- tested for statistical differences across phenological models. Post-hoc Tukey pairwise comparisons would be carried out upon significance.
- Kruskal test was used to determine whether factors "microbial clade" and "growing media type" can be separated. Since it was not (p-val < 0.01), Kruskal test on whether "microbial clade was having specific type of optimal model" was using "optimal model(s)" as response variable and the combined factors of "microbial clade" and "growing media type" was the independent variable. Each parameter (i.e. N0, K, rmax and tlag) was also tested using Kruskal test on the combined independent variable to test significance.

94 Main Assumptions

- there was no negative population changes throughout experiments from source publications. Data not fitted this assumption were set to zeros;
- all parameter estimates converged to global optimal using NLLS method.

98 Computing tools

R (ver 3.6.0)⁹ was used with packages "minpack.lm" ¹⁰ (NLLS), "stats" ⁹ (Kruskal test and PCA) and "PMCMR" ¹¹ (post-hoc Nemenyi pairwise comparisons). Python (ver 3.7.3)¹² was used with package "subprocess" ¹² (streamline project workflow).

102 Results

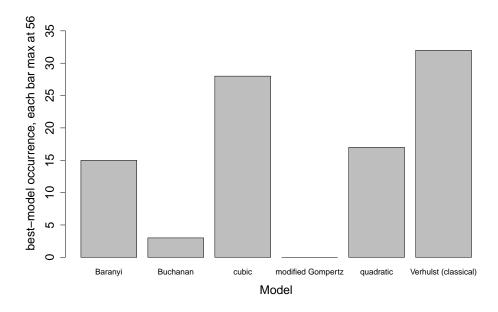


Figure 1: Barplot showing the number of "best model" identification under AIC model-selection methods with "Kruskal-Wallis rank sum test" statistic $X^2=5$, df = 5, p = 0.42

From Fig.1, large fluctuations between each model to be described as "best-fit" were observed.

However the occurrence difference was not statistical significant. Among the counts, there were

39 datasets with more than one "best-fit" models. Verhulst (classical) and cubic were the top

two models selected as "best-fit" for the 56 datasets (32 for Verhulst (classical) and 28 for

cubic). There are 10 datasets calling both "best-fit" at the same trial. Between Baranyi and

quadratic, the counts were 15 and 17 respectively with 6 datasets calling both models "best-fit".

The only outstanding performance was from modified Gompertz, which 0 datasets were called

it as "best-fit".

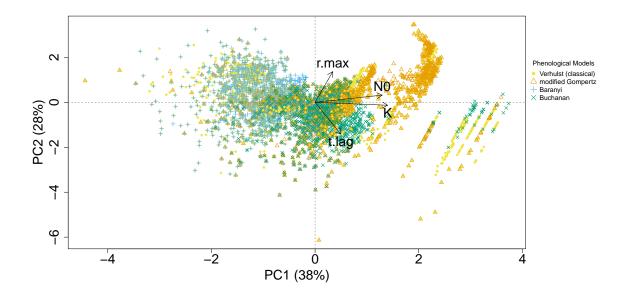


Figure 2: Biplot of Principal Component Analysis (PCA) comparing phenological models using estimated parameter values with "minimal AIC +2" valuations.

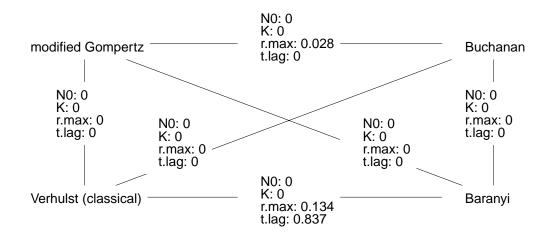


Figure 3: P-value summary between models on the four parameters under post-hoc Tukey-Dist pairwise comparison from Kruskal-Wallis Test. Kruskal tests for all four factors were significant (N0: $X^2=410.33$, df = 3, p-value = 0; K: $X^2=1575.16$, df = 3, p-value = 0; r.max: $X^2=201.75$, df = 3, p-value = 0; t.lag: $X^2=1418.47$, df = 3, p-value = 0).

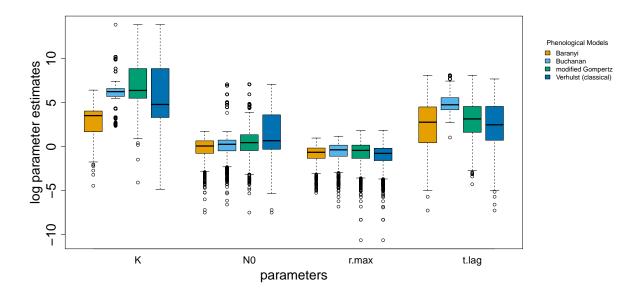


Figure 4: Boxplot of log parameter values grouped by phenological models. Statistical results were summarized in Fig.3

In Fig.2, principal component 1 (PC1) was capturing 38 % variability. It was composed 111 approximately by 0.65 No, 0.7 K, 0.17 r.max and 0.25 t.lag. PC2 was capturing 28 % variability. 112 It was composed approximately by 0.16 No. -0.06 K, 0.69 r.max and -0.71 t.lag. There were 113 51 datasets with phenological models fitting, although they may not be the "best-fit" ones. 114 Datasets 23, 27, 36, 52, 53 were strictly limited to polynomial-fitting (Fig.5). Verhulst (classical) 115 was having the widest neutral coverage across parameter space (Fig.2,4). All other three models 116 (modified Gompertz, Baranyi and Buchanan) were generally modelling within the Verhulst 117 (classical) coverage (Fig.4). modified Gompertz was the second widest coverage model but 118 Verhulst (classical) was evaluated better if both equations fitted the same dataset (Fig.1). 119 More successful trials were towards positive responses for N0, K and r.max. Baranyi was 120 a specific model more specified in describing datasets with negative responses towards most 121 parameter factors (all except r.max). Baranyi had the strictest r.max acceptance for successful 122 NLLS modelling (Fig.4). Buchanan had the narrowest parameter ranges in most parameters 123 (all except r.max, Fig.4). Datasets describable by this model were generally neutral responses 124 towards all four parameters (Fig.2). In the analysis for individual parameters, the parameter 125 value ranges overlapping between phenological models (Fig.4). Hence the differences were not 126 observable although most "differences" were statistically significant (Fig. 3).

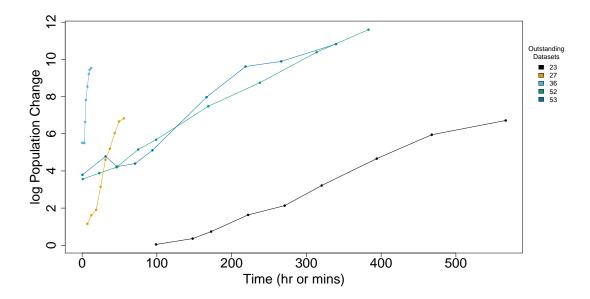


Figure 5: Line plot of datasets restricted to polynomial fits. Dataset details could be found in Table 1

Using Kruskal test, microbial clade was significantly correlated with the experimental growth 128 media ($X^2 = 36.12$, df = 17, p-value = 0). Hence downstream analyses had taken both factors 129 as a combined independent variable. Based on the combined factor, Kruskal test were done using 130 dependent variable of "optimal model type" ($X^2 = 28.74$, df = 38, p-value = 0.86), "N0" (X^2 131 =40.54, df =27, p-value =0.05), "K" ($X^2=46.21$, df =27, p-value =0.01), "r.max" ($X^2=40.54$), "r.max 132 = 29.9, df = 27, p-value = 0.32) and "t.lag" ($X^2 = 45.87$, df = 27, p-value = 0.01). Among 133 the parameters, r.max was the only factor without significance. However posthoc Nemenyi test 134 showed that none of the factor pairs (out of 28 pairs) were statistically significant although the 135 Kruskal test showed otherwise. 136

Discussion

143

AIC is the most suitable approach (compare with BIC and R²) for model evaluation within and between models in this project because AIC 1. is accurate with small sample size^{13,14} and sparse data¹⁴; 2. does not assume a "true model" was under examination^{15–17}; and 3. take number of parameters into evaluation consideration⁶. Also none of the phenological models used are "nested"¹⁸, leaving AIC as the only appropriate model-selection method.

Phenological models were not always better than polynomials. Since experimental data is

highly variable, some may not fit assumptions of these models (Fig. 5). Even among the fitted 144 ones, phenological models are not statistically better than polynomials (Fig.1). Reasons for 145 data unable to be fitted by a ecological model can be due to different reasons within three 146 categories: unfit model, unfit data and assumptions in model not met. The chance for having 147 "unfit model" is small because there are multiple fits in other datasets summarized in Fig.2 148 for each phenological models. Some datasets are clearly "unfit data" (e.g. dataset 36, 52 149 and 53; Fig.5). They share similar properties: few record points, data line-up in a fairly 150 linear manner and not much curvatures are observed. These are not typical logistic growth 151 The reasons generating these data can be due to the insufficient experiment time 152 (hence no carrying capacity plateau has observed), improper culture-handling before the start 153 of experiments (hence no initial population sizes recorded as the microbial clade has already 154 adapted to the experiment environment) and/or coarse data record intervals (hence too many 155 population fluctuation features missed within recording intervals). One or more reasons can 156 potentially related to the data in Fig.5. The other two datasets (i.e. datasets 23 and 27) may 157 fall into the third category. By observations, these two datasets can potentially be described by 158 phenological models. However the reasons for unsuccessful NLLS modelling can be due to the 159 small data sizes. No and K values can be extrapolated but not calculated. There are only one 160 data point on each end of the line indicating a curve direction change. As phenological models 161 need multiple data records to support N0 and K values as parameters, these two datasets can 162 probably only satisfy the r.max requirement. Hence these two datasets are also considered not 163 phenological model-friendly. 164

Between phenological models, some of them may be highly specific (e.g. Baranyi and 165 Buchanan) or general (e.g. Verhulst (classical) and modified Gompertz). Although PCA re-166 sult shows significant separation of model properties (Fig.2,3,4), these differences are not ob-167 servable in the data (Fig.4). The clustering of NLLS parameter estimates is also not having 168 defined boundaries (Fig.2), increasing the difficulty of choosing the most appropriate pheno-169 logical model for a newly-generated microbial population sizes dataset. Due to the different phenological model properties towards their own parameters (Fig.2), the PCA graph can be a 171 reference for determining the best phenological model on those new data in the future. However 172 Fig.2 can only be used literally when the experimental temperature, substrate and population 173 unit of the microbial clades all fall within the categories or ranges listed in Table 1. If not, the 174 same methodology can be used with suitable published data to generate the PCA result for

reference. Generally speaking, Verhulst (classical) will be sufficient to explain most behaviours
of the population. modified Gompertz can be used when authors want to incorporate t.lag into
description, although this factor may be redundant (Fig.1). Using Fig.2 as reference, one can
pinpoint the data response to the parameters and refine their model-of-choice to a more specific
one if necessary.

On biological side, the analyses show that microbial identity and the growth media are statistically significant as expected. Hence they should be analysed as one factor because there is no method to isolate one from the other. Yet this combined factor does not have statistically significant indicative number on their initial, climax population size, their maximum growth rate nor the lag phase duration.

186 Conclusion

Published phenological models were data-specific, which none of them were found significantly performing better than the others. Although most of the parameter values are significantly different between models, their ranges are superimposing with one another. Phenological models correlate with parameters differently, but the correlations are unobservable through plotting a log-linear logistic growth curve. There were assumptions embedded within phenological models which have limited its ability to describe data without a distinct sigmoid shape. Biologically, microbial identity has no indication on how they grow in laboratory conditions. Hence these parameters, if needed for researches, have to be measured case-by-case.

195 Code and Data Availability

All scripts and data used for this report were publicity available at GitHub.

197 References

- 1. Burnham, K. P. & Anderson, D. R. Multimodel inference: understanding AIC and BIC in model selection. Sociological methods & research 33, 261–304 (2004).
- 200 2. McKendrick, A. & Pai, M. K. XLV.—the rate of multiplication of micro-organisms: a

 201 mathematical study. *Proceedings of the Royal Society of Edinburgh* **31**, 649–653 (1912).

- 202 3. Gil, M. M., Brandão, T. R. & Silva, C. L. A modified Gompertz model to predict microbial
- inactivation under time-varying temperature conditions. Journal of Food Engineering 76.
- Bugdeath, 89 -94. ISSN: 0260-8774. http://www.sciencedirect.com/science/article/
- pii/S0260877405003389 (2006).
- ²⁰⁶ 4. Baranyi, J, McClure, P., Sutherland, J. & Roberts, T. Modeling bacterial growth responses.
- Journal of industrial microbiology 12, 190–194 (1993).
- 5. Buchanan, R., Golden, M. & Whiting, R. Differentiation of the effects of pH and lactic or
- acetic acid concentration on the kinetics of Listeria monocytogenes inactivation. Journal
- of Food Protection **56**, 474–478 (1993).
- 6. Johnson, J. B. & Omland, K. S. Model selection in ecology and evolution. *Trends in ecology*
- 212 & evolution 19, 101–108 (2004).
- 7. Akaike, H. in Selected papers of hirotugu akaike 199–213 (Springer, 1998).
- 8. Burnham, K. & Anderson, D. Model selection and multimodel inference: a practical
- information-theoretic approach. Ecological Modelling.
- 9. R Core Team. R: A Language and Environment for Statistical Computing R Foundation
- for Statistical Computing (Vienna, Austria, 2019). https://www.R-project.org/.
- 10. Elzhov, T. V., Mullen, K. M., Spiess, A.-N. & Bolker, B. minpack.lm: R Interface to
- the Levenberg-Marquardt Nonlinear Least-Squares Algorithm Found in MINPACK, Plus
- Support for Bounds R package version 1.2-1 (2016). https://CRAN.R-project.org/
- package=minpack.lm.
- 222 11. Pohlert, T. The Pairwise Multiple Comparison of Mean Ranks Package (PMCMR) R
- package (2014). https://CRAN.R-project.org/package=PMCMR.
- 224 12. Python Core Team. Python: A dynamic, open source programming language Python Soft-
- ware Foundation (Delaware, United States, 2019). https://www.python.org/.
- 226 13. Acquah, H. D.-G. Comparison of Akaike information criterion (AIC) and Bayesian in-
- formation criterion (BIC) in selection of an asymmetric price relationship. Journal of
- Development and Agricultural Economics 2,001-006 (2010).
- 229 14. Kuha, J. AIC and BIC: Comparisons of assumptions and performance. Sociological methods
- 230 & research **33**, 188–229 (2004).

- ²³¹ 15. Aho, K., Derryberry, D. & Peterson, T. Model selection for ecologists: the worldviews of AIC and BIC. *Ecology* **95**, 631–636 (2014).
- ²³³ 16. Vrieze, S. I. Model selection and psychological theory: a discussion of the differences ²³⁴ between the Akaike information criterion (AIC) and the Bayesian information criterion ²³⁵ (BIC). Psychological methods 17, 228 (2012).
- Yang, Y. Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika* **92**, 937–950 (2005).
- Wang, Y. & Liu, Q. Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of stock—recruitment relationships. *Fisheries Research* 77, 220–225 (2006).
- ²⁴¹ 19. Schwarz, G. Estimating the dimension of a model. Ann. Stat. 6, 461–464 (1978).
- ²⁴² 20. Kelley, C. T. Iterative methods for optimization (SIAM, 1999).
- 243 21. Turchin, P. Complex population dynamics: a theoretical/empirical synthesis (Princeton university press, 2003).

Appendix Appendix

Table 1: Table showing dataset id details in this project.

id	$T^{o}C$	clade	substrate	replicate	Source	Pop unit
1	5	Chryseobacterium balustinum	TSB	1	1	OD_595
2	5	Enterobacter sp	TSB	1	1	OD_595
3	5	Pantoea agglomerans	TSB	1	1	OD_595
4	5	Bacillus pumilus	TSB	1	1	OD_595
5	5	Clavibacter michiganensis	TSB	1	1	$OD_{-}595$
6	5	Pseudomonas fluorescens	TSB	1	1	$\mathrm{OD}\-595$
7	5	Acinetobacter clacoaceticus	TSB	1	1	$OD_{-}595$
8	5	Stenotrophomonas maltophilia	TSB	1	1	OD_595
9	5	Klebsiella pneumonia	TSB	1	1	OD_595
10	5	Dickeya zeae	TSB	1	1	OD_595
11	5	Pectobacterium carotovorum	TSB	1	1	$\mathrm{OD}\-595$
12	15	Chryseobacterium balustinum	TSB	1	1	$OD_{-}595$
13	25	Chryseobacterium balustinum	TSB	1	1	OD_595

14	35	Chryseobacterium balustinum	TSB	1	1	$OD_{-}595$
15	5	Tetraselmis tetrahele	ESAW	1	2	N
16	5	Tetraselmis tetrahele	ESAW	2	2	N
17	5	Tetraselmis tetrahele	ESAW	3	2	N
18	5	Tetraselmis tetrahele	ESAW	4	2	N
19	5	Tetraselmis tetrahele	ESAW	5	2	N
20	8	Tetraselmis tetrahele	ESAW	1	2	N
21	16	Tetraselmis tetrahele	ESAW	1	2	N
22	32	Tetraselmis tetrahele	ESAW	1	2	N
23	2	Staphylococcus sp	Raw Chicken Breast	1	3	CFU
24	4	Staphylococcus sp	Raw Chicken Breast	1	3	CFU
25	7	Staphylococcus sp	Raw Chicken Breast	1	3	CFU
26	10	Staphylococcus sp	Raw Chicken Breast	1	3	CFU
27	20	Staphylococcus sp	Raw Chicken Breast	1	3	CFU
28	2	Staphylococcus sp	Salted Chicken Breast	1	3	CFU
29	2	Staphylococcus sp	Cooked Chicken Breast	1	3	CFU
30	2	Pseudomonas sp	Raw Chicken Breast	1	3	CFU
31	2	Aerobic Psychotropic	Raw Chicken Breast	1	3	CFU
32	2	Aerobic Mesophilic	Raw Chicken Breast	1	3	CFU
33	8	Spoilage	Vacuum Beef Striploins	1	4	N
34	8	Escherichia coli	Vacuum Beef Striploins	1	4	N
35	8	Salmonella Typhimurium	Vacuum Beef Striploins	1	4	N
36	10	Spoilage	C02 Beef Striploins	1	4	N
37	12	Spoilage	Vacuum Beef Striploins	1	4	N
38	30	Spoilage	Vacuum Beef Striploins	1	4	N
39	6	Serratia marcescens	Pasteurised Skim Milk	1	5	N
40	6	Serratia marcescens	UHT Skim Milk	1	5	N
41	6	Serratia marcescens	Pasteurised Full-fat Milk	1	5	N
42	6	Serratia marcescens	UHT Full-fat Milk	1	5	N
43	6	Serratia marcescens	Pasteurised Double Cream	1	5	N
44	6	Serratia marcescens	UHT Double Cream	1	5	N
45	0	Arthrobacter sp	TGE agar	1	6	CFU
46	37	Arthrobacter sp	TGE agar	1	6	CFU
47	0	Arthrobacter aurescens	TGE agar	1	6	CFU
48	0	Arthrobacter citreus	TGE agar	1	6	CFU
49	0	Arthrobacter globiformis	TGE agar	1	6	CFU
	i .					

50	0	Arthrobacter simplex	TGE agar	1	6	CFU
51	8	Lactobacillus plantarum	MRS broth	1	7	N
52	4	Weissella viridescens	MRS broth	1	7	N
53	4	Lactobacillus sakei	MRS broth	1	7	N
54	15	Oscillatoria agardhii	Z8	1	8	DryWeight
55	15	Pseudomonas sp	APT Broth	1	9	CFU
56	10	Lactobaciulus plantarum	MRS	1	10	N

²⁴⁶ "Source" column publication key:

247

- Bae, Y.M., Zheng, L., Hyun, J.E., Jung, K.S., Heu, S. and Lee, S.Y., 2014. Growth characteristics and biofilm formation of various spoilage bacteria isolated from fresh produce. Journal of food science, 79(10), pp.M2072-M2080.
- Bernhardt, J.R., Sunday, J.M. and O'Connor, M.I., 2018. Metabolic theory and the temperature-size rule explain the temperature dependence of population carrying capacity. The American naturalist, 192(6), pp.687-697.
- Galarz, L.A., Fonseca, G.G. and Prentice, C., 2016. Predicting bacterial growth in raw, salted, and cooked chicken breast fillets during storage. Food Science and Technology International, 22(6), pp.461-474.
- Gill, C.O. and DeLacy, K.M., 1991. Growth of Escherichia coli and Salmonella typhimurium on high-pH beef packed under vacuum or carbon dioxide. International journal of food microbiology, 13(1), pp.21-30.
- Phillips, J.D. and Griffiths, M.W., 1987. The relation between temperature and growth of bacteria in dairy products. Food Microbiology, 4(2), pp.173-185.
- Roth, N.G. and Wheaton, R.B., 1962. Continuity of psychrophilic and mesophilic growth characteristics in the genus Arthrobacter. Journal of bacteriology, 83(3), pp.551-555.
- Silva, A.P.R.D., Longhi, D.A., Dalcanton, F. and Aragão, G.M.F.D., 2018. Modelling the growth of lactic acid bacteria at different temperatures. Brazilian Archives of Biology and Technology, 61.
- Sivonen, K., 1990. Effects of light, temperature, nitrate, orthophosphate, and bacteria on growth of and hepatotoxin production by Oscillatoria agardhii strains. Appl. Environ. Microbiol., 56(9), pp.2658-2666.

Stannard, C.J., Williams, A.P. and Gibbs, P.A., 1985. Temperature/growth relationships for psychrotrophic food-spoilage bacteria. Food Microbiology, 2(2), pp.115-122.
 Zwietering, M.H., De Wit, J.C., Cuppers, H.G.A.M. and Van't Riet, K., 1994. Modeling of bacterial growth with shifts in temperature. Appl. Environ. Microbiol., 60(1), pp.204-213.