<| Main Contents (Index.ipynb) |>

Appendix: The computing Miniproject

Contents

We have talked a lot about workflows and confronting models with data. It's time to do something concrete with all the techniques you have been learning.

The CMEE Miniproject gives you an opportunity to try the "whole nine yards" of developing and implementing a workflow and delivering a "finished product" — where you ask and answer a scientific question in biology (potentially involving multiple sub-questions/hypotheses). It will give you an opportunity to perform a "dry run" of executing your actual dissertation project, and you may use it to trial some of the techniques and/or explore some of the data/theory you might use in your Dissertation project.

Objectives

The general question you will address is: *What mathematical model best fits an empirical dataset?*

You may think of this as testing a set of alternative hypotheses — every alternative hypothesis is nothing but an alternative model to describe an observed phenomenon, as you will have learned in the lectures on model fitting.

The Project

You will choose a dataset and set of alternative models from the options provided to you.

Please read the papers in the Readings section — these will help orient you in the right direction for tackling your miniproject.

The Miniproject must satisfy the following criteria (and follow the accompanying guidelines):

1. It should employ all the biological computing tools you have learned so far: shell (bash) scripting, git, LaTeX, R, and Python. Using these tools, you will build a workflow that starts with the data and ends with a written report (in LaTeX). How you choose the different tools (e.g., Python vs R) is your choice. This is part of the assessment.
2. *At least* two different models (hypotheses) must be fitted to the data. The models should be fitted and selected using an appropriate method. Specifically, irrespective of the problem/dataset you choose (see below), use Nonlinear Least Squares (NLLS) to fit ≥ 2 alternative models to data, followed by model selection using AIC and BIC (read the Johnson and Omland 2005 paper). You may choose additional means for model comparison/selection beyond these.*
3. The project should be fully reproducible. Write a script that "glues" the workflow together and runs it, from data processing to model fitting to plotting (e.g., in R) to compilation of the LaTeX written report (*More detailed instructions on report below*). Look back at the TheMulQuaBio to see how you would run the different components. For example, we have covered how to run R and compile $LaTeX$ using the `subprocess` module in Python. The assessor should be able to run just this script to get everything to work without errors. Use Python or to write this main script. If using bash, call it `run_MiniProject.sh` and if using Python, called it `run_MiniProject.py` .

You will be given lectures and practicals on model fitting before you start on your Miniproject.

The Report

The report should,

- be written in LaTeX using the article document class, in 11pt (any font will do, within reason!).
- be double-spaced, with *continuous* line numbers.
- have a title, author name with affiliation and wordcount (next point) on a separate title page.
- have an introduction with objectives of the study, and appropriate additional sections such as methods, data, results, discussion, etc.
- should contain in the *Methods* a sub-section called "Computing tools" which states briefly how each of the three scripting language (bash, R, Python) and what packages within them were used and a justification of why.
- must contain $\leq$3500 words *excluding the contents of the title page, references, and Figure or Table captions+legends*; there should be a word count at the beginning of the document (typically using the `texcount` package).
- have references properly cited in text and formatted in a list using bibtex.

For the writeup, you probably should read the *general* (*not* word count, formatting etc.) dissertation writing guidelines given in the Silwood Masters Student Guidebook.

Submission

Add, commit and push all your work to your bitbucket repository using a directory called `MiniProject` at the same level as the Week1, Week2 etc. directories, by the Miniproject deadline given in your course guidebook.

At this stage you are not going to be told you how to organize your project — that's part of the marking criteria (see next section).

Marking criteria

Equal weightage will be given to the code+workflow and writeup components — each component will be marked to a max of 100 pts and then rescaled to a single mark / 100 using equal weightage

The assessor will be looking for the following while assessing your submission:

- A well-organized project where code, results, data, etc., are easy to locate, inspect, and use. In the project's README also include:
  - Any dependencies or special packages the user/marker should be aware of
  - What each package you used is for
  - Version of each language used
- A project that runs smoothly, without any errors once the appropriate script (i.e., `run_MiniProject.py` or `run_MiniProject.sh`) is called.
- A report that contains all the components indicated above in "The Report" subsection, with some original thought and synthesis in the **Introduction** and **Discussion** sections.
- Quality of the presentation of the graphics and tables in your report, as well as any plots showing model fits to the data.
- The marking criteria you may refer to is the summative marking criteria (./MARKING_CRITERIA.pdf).

The Model Fitting Problems

You can pick from one of the following three options.

Thermal Performance Curves

The question

How well do different mathematical models, e.g., based upon biochemical (mechanistic) principles vs. phenomenological ones, fit to the thermal responses of metabolic traits?

This is currently a "hot" (no pun intended!) topic in biology. On the *ecological side*, because the temperature-dependence of metabolic rate sets the rate of intrinsic $r_{\max}$ (papers by Savage et al., Brown et al.) as well as interactions between species, it has a strong effect on population dynamics. In this context, note that 99.9% of life on earth is ectothermic! On the *evolutionary side*, the temperature-dependence of fitness and species interactions also means that warmer environments may have stronger rates of evolution. This may be compounded by the fact that mutation rates may also increase with temperature (papers by Gillooly et al.).

The Data

The dataset is called `ThermRespData.csv` . It contains a subset of the full "BioTraits" database. This subset contains hundreds of "thermal responses" for growth, respiration and photosynthesis rates in plants and bacteria (both aquatic and terrestrial). These data were collected through lab experiments across the world, and compiled by various people over the years. The field names are defined in a file called `BiotraitsTemplateDescription.pdf` , also in the `data` directory. The two main fields of interest are `OriginalTraitValue` (the trait values responding to temperature), and `ConTemp` (the temperature). Individual thermal response curves can be identified by `FinalID` values --- each `FinalID` corresponds to one thermal performance curve.

The Models

All the following parameters and variables are in SI units.

There are multiple models that might best describe these data. The simplest are the general quadratic and cubic polynomial models:

$$B = B_0 + B_1 x + B_2 x^2 \tag{1}$$

$$B = B_0 + B_1 x + B_2 x^2 + B_3 x^3 \tag{2}$$

These are phenomenological models, with the parameters B_0, B_1, B_2 and B_3 lacking any mechanistic interpretation. x is the independent variable (in this case Temperature, T)

Another phenomenological model option is the Briere model (Appendix-ModelFitting.ipynb#The-TPC-models):

$$B = B_0 T (T - T_0) \sqrt{T_m - T} \tag{3}$$

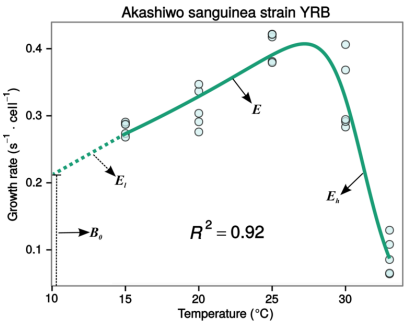
Where T is temperature, T_0 and T_m are the minimum and maximum feasible temperatures for the trait (below or above which the traits goes to zero), and B_0 is a normalization constant.

In contrast, the Schoolfield model (paper is in `readings` directory of TheMulQuaBio repo) is a mechanistic option that is based upon thermodynamics and enzyme kinetics:

$$B = \frac{B_0 e^{\frac{-E}{k}\left(\frac{1}{T} - \frac{1}{283.15}\right)}}{1 + e^{\frac{E_l}{k}\left(\frac{1}{T_l} - \frac{1}{T}\right)} + e^{\frac{E_h}{k}\left(\frac{1}{T_h} - \frac{1}{T}\right)}} \tag{4}$$

Please also have a look at the Delong et al 2017 paper, which lists this and other mechanistic TPC models (see Readings). You may choose additional models listed in that paper for comparison, if you want.

Here, k is the Boltzmann constant (8.617×10^{-5} eV · K^{-1}), B the value of the trait at a given temperature T (K) (K = °C + 273.15), while B_0 is the trait value at 283.15 K (10°C) which stands for the value of the growth rate at low temperature and controls the vertical offset of the curve. E_l is the enzyme's low-temperature de-activation energy (eV) which controls the behavior of the enzyme (and the curve) at very low temperatures, and T_l is the at which the enzyme is 50% low-temperature deactivated. E_h is the enzyme's high-temperature de-activation energy (eV) which controls the behavior of the enzyme (and the curve) at very high temperatures, and T_h is the at which the enzyme is 50% high-temperature deactivated. E is the activation energy (eV) which controls the rise of the curve up to the peak in the "normal operating range" for the enzyme (below the peak of the curve and above T_h).



Example of the Sharpe-Schoolfield eqn, that is, 1 and 2, . 4) fitted to the thermal response curve of a (replaceiological trait. < /figcaptio x with resource abundance.>

In many cases, a simplified Schoolfield model would be more appropriate for thermal response data, because low temperature inactivation is weak, or is undetectable in the data because low-temperature measurements were not made.

$$B = \frac{B_0 e^{\frac{-E}{k}(\frac{1}{T} - \frac{1}{283.15})}}{1 + e^{\frac{E_h}{k}(\frac{1}{T_h} - \frac{1}{T})}} \tag{5}$$

In other cases, a different simplified Schoolfield model would be more appropriate, because high temperature inactivation was not detectable in the data because measurements were not made at sufficiently high temperatures:

$$B = \frac{B_0 e^{\frac{-E}{k}(\frac{1}{T} - \frac{1}{283.15})}}{1 + e^{\frac{E_l}{k}(\frac{1}{T_l} - \frac{1}{T})}} \tag{6}$$

Note that the cubic model (Equation 2) has the same number of parameters as the the reduced Schoolfield models (eq. 5 & 6). Also, the temperature parameter (T) of the cubic model (Equation 2) is in °C, whereas the Temperature parameter in the Schoolfield model is in K.

Functional Responses

The Question

How well do different mathematical models, e.g., based upon foraging theory (mechanistic) principles vs. phenomenological ones, fit to functional responses data across species?

In ecological parlance, a functional response is the relationship between a consumer's (e.g., predator) biomass consumption rate and abundance of the target resource (e.g., prey). Functional responses arise from fundamental biological and physical constraints on consumer-resource interactions (e.g., Holling 1959, Pawar et al, 2012), and determine the rate of biomass flow between species in ecosystems across the full scale of sizes, from the smallest (e.g., microbes) to the largest (e.g., blue whales). Functional responses also play a key sole in determining the stability (responses to perturbations) of the food webs that underpin ecosystems.

The Data

The dataset is called `CRat.csv`. It contains measurements of rates of consumption of a single resource (e.g., prey, plants) species' by a consumer species (e.g., predators, grazers). These data were collected through lab and field experiments across the world. The field names are defined in a file called `BiotraitsTemplateDescription.pdf`, also in the `data` directory. The two main fields of interest are `OriginalTraitValue` (co trait values responding to), and `...` (the resource abundance). Individual functional response curves can be identified by `ID` values --- each `ID` corresponds to one curve.

The Models

All the following parameters and variables are in SI units.

The fundamental measure of interest (the response variable) is consumption rate (c). This is expressed in terms of biomass quantity or number of individuals of resource consumed *per unit time per unit consumer* (so units of Mass (or Individuals) / Time).

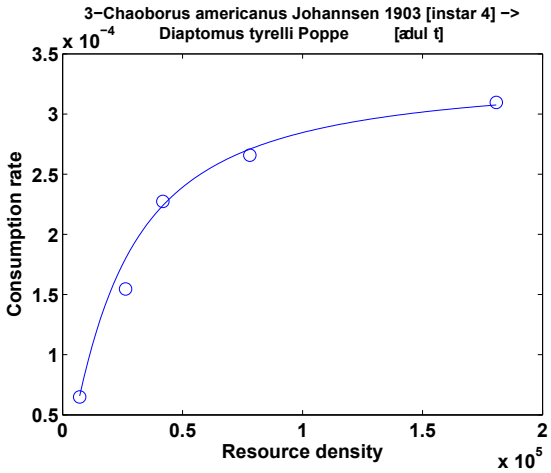
Again, the simplest mathematical models you can use are the phenomenological quadratic and cubic polynomial models, that is eqns. 1 and 2 (replace x with resource abundance).

Then, there is the more mechanistic Holling Type II model (Holling, 1959):

$$c = \frac{ax_R}{1 + hax_R} \tag{7}$$

Here, x_R is resource density (Mass / Area or Volume), a is consumer's search rate (Area or Volume / Time), and h is handling time of the consumer for that resource (time taken to overpower and ingest it).

Below is an example FR curve from the dataset you have been given with the Type II model fitted to it.

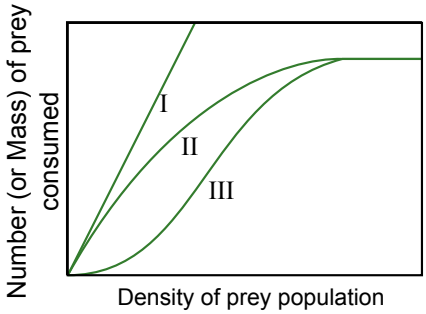


Example of the a Type II model (eqn. 7) fitted to a functional response of a consumer on a resource.

There is also the less-mechanistic "generalized" functional response model:

$$c = \frac{ax_R^{q+1}}{1 + hax_R^{q+1}} \tag{8}$$

Where everything is same as 7, but the additional parameter q (dimensionless) is a shape parameter that allows the shape of the response to be more flexible/variable, from "Type I" to "Type III". This model is less mechanistic because it includes a phenomenological parameter q which does not have a formal biological meaning. Note that if $q = 0$, eqn (8 becomes same as the Type II model (eqn. 7)).



The range of functional responses captured by the generalized functional response model (eqn. 8).

There are other models for functional responses as well (some more mechanistic), that define parameters of the functional response in terms of body size of predator and prey (Pawar et al 2012).

Population Growth

The Question

How well do different mathematical models, e.g., based upon population growth (mechanistic) theory vs. phenomenological ones, fit to functional responses data across species?

Fluctuations in the abundance (density) of single populations may play a crucial role in ecosystem dynamics and emergent functional characteristics, such as rates of carbon fixation or disease transmission. A population grows exponentially while its abundance is low and resources are not limiting (the Malthusian principle). This growth then slows and eventually stops as resources become limiting. There may also be a time lag before the population growth really takes off at the start. We will focus on microbial (specifically, bacterial) growth rates. Bacterial growth in batch culture follows a distinct set of phases; lag phase, exponential phase and stationary phase. During the lag phase a suite of transcriptional machinery is activated, including genes involved in nutrient uptake and metabolic changes, as bacteria prepare for growth. During the exponential growth phase, bacteria divide at a constant rate, the population doubling with each generation. When the carrying capacity of the media is reached, growth slows and the number of cells in the culture stabilises, beginning the stationary phase. Traditionally, microbial growth rates were measured by plotting cell numbers or culture density against time on a semi-log graph and fitting a straight line through the exponential growth phase – the slope of the line gives the maximum growth rate (r_{max}). Models have since been developed which we can use to describe the whole sigmoidal bacterial growth curve.

The Data

The dataset is called `LogisticGrowthData.csv` . It contains measurements of change in biomass or number of cells of microbes over time. These data were collected through lab experiments across the world. The field names are defined in a file called `LogisticGrowthMetaData.csv` , also in the `data` directory. The two main fields of interest are `PopBio` (abundance), and `Time` . Single population growth rate curves can be identified by as unique temperature-species-medium-citation-replicate combinations (concatenate them to get a new string variable that identifies unique growth curves).

The Models

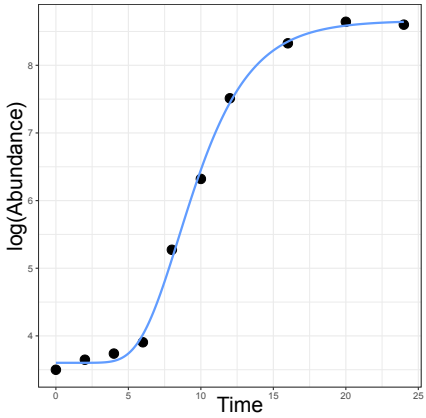
All the following parameters and variables are in SI units.

Yet again, the simplest mathematical models you can use are the phenomenological quadratic and cubic polynomial models, that is eqns. 1 and 2 (replace x with Time).

A classical model mechanistic to a degree (Recall the Modelling Lecture) is the logistic equation:

$$N_t = \frac{N_0 K e^{rt}}{K + N_0(e^{rt} - 1)} \tag{9}$$

Here N_t is population size at time t, N_0 is initial population size, r is maximum growth rate (AKA $r_{\max}$), K is carrying capacity.



An example population growth curve dataset to which the modified Gompertz model (Zwietering et. al., 1990) has been fitted.

Other popular models are the modified Gompertz model (Zwietering et. al., 1990), the Baranyi model (Baranyi, 1993), and the Buchanan model (or three-phase logistic model; Buchanan, 1997). The Buchanan model in particular is capable of capturing the lag phase before the population starts growing exponentially, often seen in microbial population growth. Examples of fitting these models are available in the model fitting appendix (./Appendix-ModelFitting.ipynb#Population-growth-rate-example).

The modified Gompertz model (eq. 10) has been used consistently through the literature as a good description of bacterial growth. Here maximum growth rate (r_{max}) is the tangent to the inflection point, λ is the x-axis intercept to this tangent and A is the asymptote ($A = \ln\left(\frac{N_{max}}{N_{min}}\right)$):

$$N_t = A \cdot \exp\left\{-\exp\left[\frac{r_{max}e}{A}(\lambda - t) + 1\right]\right\} \tag{10}$$

The Baranyi model (eq. 11) introduces a new dimensionless parameter h_0 which represents the initial physiological state of the cells. The length of the lag phase is determined by the value of h_0 at inoculation and the post-inoculation environment. Thus the definition of lag is independent from the shape of the growth curve, and the effect of the previous environment is separated from the effects of the present environment. This allows modelling growth without a lag period following inoculation from media favourable to growth to new media also favourable to growth. One formulation given is:

$$N_t = N_{min} + r_{max}A_t - \ln\left(1 + \frac{e^{r_{max}A_t} - 1}{e^{N_{max} - N_{min}}}\right), \tag{11}$$

where:

$$A_t = t + \frac{1}{r_{max}} \cdot \ln\left(\frac{e^{-r_{max}t} + h_0}{1 + h_0}\right). \tag{12}$$

In this model, r_{max} and h_0 can be related to obtain the lag time, λ:

$$\lambda = \frac{\ln\left(1 + \frac{1}{h_0}\right)}{r_{max}}, \tag{13}$$

taking us back to the original same four parameters as Gompertz.

The Buchanan model, or "three-phase logistic model"' (eq. 14) is very simple and makes three assumptions; 1. growth rate during lag phase is zero, 2. growth rate during exponential phase is constant, 3. growth rate during stationary phase is zero. This is not a good description of the shape of the curve (no curvature), but the argument is that it captures the growth parameters well without the need for a more complicated model.

$$N(t) = \begin{cases} N_{min} & \text{if } t \leq t_{lag} \\ N_{max} + r_{max} \cdot (t - t_{lag}) & \text{if } t_{lag} < t < t_{max} \\ N_{max} & \text{if } t \geq t_{max} \end{cases} \quad (14)$$

Here, t_{max} is the time at which N_{max} is reached.

Additional models and questions you can tackle

In all three options above, you may try to tackle fitting to additional models you find in the literature. Some readings have been provided for each of the three data types. In addtion, you may wish to tackle some other hypotheses or explore patterns by considering additional covariates. For example,

Do different models fit different types of thermal performance curves (e.g., Photosynthesis vs Respiration)?

Do different taxa show different functional responses?

Does temperature or taxon identity affect which population growth rate model fits best?

You may also want to revisit the results of another paper that has done comparisons of the models you have chosen with your new dataset.

Suggested Workflow

You will build a workflow that starts with the data and ends with a report written in LaTeX. I suggest the following components and sequence in your workflow (you can choose to do it differently!):

- A Python or R script that imports the data and prepares it for NLLS fitting, with the following features:
  - It should create unique ids so that you can identify unique datasets (e.g., single thermal responses)
  - It may filter out datasets with less than x data points (where x is the minimum number of data points needed to fit the models).
  - It should deal with missing, and other problematic data values
- The script should add columns containing starting values of the model parameters for the NLLS fitting (how will you get these?)
- Save the modified data to a new csv file.

A Python (or R) script that opens the new modified dataset (from step 1) and does the NLLS fitting. For example, if you choose Python for this, it might have the following features:

- Uses lmfit — look up submodules `minimize`, `Parameters`, `Parameter`, and `report_fit`. *Have a look through* http://lmfit.github.io/lmfit-py (http://lmfit.github.io/lmfit-py), especially http://lmfit.github.io/lmfit-py/fitting.html#minimize (http://lmfit.github.io/lmfit-py/fitting.html#minimize)\ You will have to install lmfit using pip or `easy_install` - use sudo mode. In addition to the lmfit example in class, you may want to look for others online.
- Will use the try construct because not all runs will converge. Recall the `try` example from R.
- The more data curves you are able to fit, the better — that is part of the challenge!
- Will calculate AIC, BIC, R^2, and other statistical measures of fit (you decide what you want to include)
- Will export the results to a csv that the plotting R script (next item) can read.
- A R script that imports the results from the previous step and plots every curve with the two (or more) models (or none, if nothing converges) overlaid — all plots should be saved in a single separate sub-directory. *Use ggplot for pretty results!*
- LaTeX source code that generates your report.
- A Python script (saved in Code) called run_MiniProject.py that runs the whole project, right down to compilation of the LaTeX document.

Doing all this may seem a bit scary at the start. However, you need to approach the problem systematically and methodically, and you will be OK. here are some suggested first steps to get you started:

- Explore the data in R and get a preliminary version of the plotting script without the fitted models overlaid worked out. That will also give you a feel for the data.
- Explore the two models – be able to plot them. Write them as functions in your python script, because that's where you will use them (step 2 above) (you can use matplotlib for quick and dirty plotting and then suppress those code lines later).
- Figure out, using a minimal example (say, with one, "nice-looking" thermal performance, functional response, or population growth curve/dataset) to see how the python `lmfit` module works. We can help you work out th minimal example, including the usage of try to catch errors in case the fitting doesn't converge.

One thing to note is that you may need to do the NLLS fitting on the logarithm of the function (and therefore, the data) to facilitate convergence.

Readings

All these papers are in pdf format in the Readings directory on TheMulQuaBio repository.

General

- Levins, R. (1966) The strategy of model building in population biology. Am. Sci. 54, 421–431.
- Johnson, J. B. & Omland, K. S. (2004) Model selection in ecology and evolution. Trends Ecol. Evol. 19, 101–108.
- Motulsky, H. & Christopoulos A. (2004) Fitting models to biological data using linear and nonlinear regression: a practical guide to curve fitting. Oxford University Press, USA.
- Bolker, B. M. et al. (2013) Strategies for fitting nonlinear ecological models in R, AD Model Builder, and BUGS. Methods Ecol. Evol. 4, 501–512.

Thermal Performance Curves

- Schoolfield, R. M., P. J H Sharpe, and C. E. Magnuson. 1981. “Non-Linear Regression of Biological Temperature-Dependent Rate Models Based on Absolute Reaction-Rate Theory.” Journal of Theoretical Biology 88 (4): 719–31. https://doi.org/10.1016/0022-5193(81)90246-0 (https://doi.org/10.1016/0022-5193(81)90246-0).
- Dell, Anthony I, Samraat Pawar, and Van M Savage. 2011. “Systematic Variation in the Temperature Dependence of Physiological and Ecological Traits.” Proceedings of the National Academy of Sciences of the United States of America 108 (26): 10591–10596. https://doi.org/doi (https://doi.org/doi): 10.1073/pnas.1015178108.
- DeLong, J. P., J. P. Gibert, T. M. Luhring, G. Bachman, B. Reed, A. Neyer, and K. L. Montooth. 2017. “The Combined Effects of Reactant Kinetics and Enzyme Stability Explain the Temperature Dependence of Metabolic Rates.” Ecology and Evolution 7 (11): 3940–50. https://doi.org/10.1002/ece3.2955 (https://doi.org/10.1002/ece3.2955).

Functional responses

- Holling, C. S. 1959. “Some Characteristics of Simple Types of Predation and Parasitism.” The Canadian Entomologist 91 (7): 385–98. https://doi.org/10.4039/Ent91385-7 (https://doi.org/10.4039/Ent91385-7).
- Holling, C S. 1966. “The Functional Response of Invertebrate Predators to Prey Density.” Mem. Entomol. Soc. Canada 48 (48): 1–86.

- Pawar, Samraat, Anthony I. Dell, and Van M. Savage. 2012. “Dimensionality of Consumer Search Space Drives Trophic Interaction Strengths.” Nature 486 (7404): 485–89. https://doi.org/10.1038/nature11131 (https://doi.org/10.1038/nature11131).

Population Growth

- Zwietering, M. H., I. Jongenburger, F. M. Rombouts, and K. Van’t Riet. 1990. “Modeling of the Bacterial Growth Curve.” Applied and Environmental Microbiology 56 (6): 1875–81.
- Buchanan, R. L., R. C. Whiting, and W. C. Damert. 1997. “When Is Simple Good Enough: A Comparison of the Gompertz, Baranyi, and Three-Phase Linear Models for Fitting Bacterial Growth Curves.” Food Microbiology 14 (4): 313–26. https://doi.org/10.1006/fmic.1997.0125 (https://doi.org/10.1006/fmic.1997.0125).
- Micha, Peleg, and Maria G. Corradini. 2011. “Microbial Growth Curves: What the Models Tell Us and What They Cannot.” Critical Reviews in Food Science and Nutrition. https://doi.org/10.1080/10408398.2011.570463 (https://doi.org/10.1080/10408398.2011.570463).