Predicting competitions and estimating selection coefficients from growth curves

Yoav Ram, Eynat Deluss-Gur, Uri Obolski, Maayan Bibi, Judith Berman & Lilach Hadany

# Introduction

Many experimental investigators in microbiology, genetics and evolutionary biology use growth curves to estimate fitness. They measure the Optical Density (OD) of one or more populations of cells over several hours or even days to acquire the growth curves (Fig. 1). The simplest way to estimate fitness from these curves is to infer the growth rate: taking the log of the curves during the exponential growth phase, using linear regression to fit a linear line to the data, and taking the slope of the line as a measure of the growth rate (Hall et al. 2014). Growth rates can indeed be proxies of the *selection coefficient s* (Chevin 2011), which is the standard way of measuring relative fitness in population genetics (Crow and Kimura 1970). But in many cases growth curves include additional phases besides the exponential growth phase: a lag phase, a deceleration phase, and a stationary phase. Making sense of combined effects of the different phases is more complicated.

A more robust method to infer fitness is to do competition assays between a strain of interest and a reference strain (for example, a mutant strain and a wildtype strain). From the change in frequency over the duration of the competition, investigators can estimate the *selection coefficient s* of the strain of interest (Wiser and Lenski 2015). Theoretically, this is a much better method to infer fitness, as it directly estimate relative fitness rather than indirectly estimating it from proxy measures such as growth rate. However, competition assays are laborious and expansive, requiring the construction and assaying of genetic or phenotypic markers (Lenski et al. 1991; Hegreness et al. 2006; Levy et al. 2015).

Because competition assays require so much work and/or incur high costs, many investigators do without them and use simpler methods which produce proxies of fitness such as growth rates. However, these proxies of fitness suffer from several disadvantages: many times they fail to capture the full scope of effects contributing to differences in fitness; they are hard to compare between different studies and organisms; and they are hard to use as inputs for population genetics models that could be used to predict dynamics and test hypotheses.

Here we propose a new method to estimate fitness. Our method uses fits growth models to growth curves data and uses the fitted growth models to predict the results of competitions assays. The predicted competitions are then used instead of empirical ones to infer selection coefficients. In addition, we present the results of an experimental test of our method which indicate that it is accurate and reproducible.

We implemented our method using an open source Python package that can be freely used and extended. We also provide a friendly user interface to our method to allow other investigators easy access to analysis of their growth curves.

# Methods

Here we describe the three parts of our computational method – fitting growth models to the growth curves data, using the fitted models to predict the results of competition assays, and inferring selection from the predicted competition results. We also describe the our experimental design which includes the strains and media, the experimental growth curves, and the flow cytometry measurement and analysis.

## Computational framework

### Growth models

Because we are interested in different phases of the growth – the lag phase, the exponential growth phase, the deceleration phase and the stationary phase (Fig. 1) – we use an extension of the classical logistic model called the *Baranyi-Roberts model* (Baranyi and Roberts 1994; Baranyi 1997).

The Baranyi-Growth model is defined by the following one-species ordinary differential equation (see eqs. 1c, 3a, and 5a in (Baranyi and Roberts 1994)):

(1a)

(1b)

, (1c)

where is the population density, is the per capita growth rate, is time, is the adjustment function (see below), is the maximum density, and is a deceleration parameter.

The term is used to decelerate the growth of the population as it nears the maximum density. When the deceleration parameter , the deceleration is the same as in the classical logistic model and the density at the time of the maximum growth rate (the curve's inflection point) is half the maximum density . When or this density is higher or lower (to the right or the left of the inflection point in the logistic model).

The adjustment function is used to describe the adjustment of the population to a new environment. Typically, organisms are grown in overnight culture and are then diluted into fresh media for the growth curve experiment. Therefore, the organisms who were previously adjusted to stationary phase must now adjust to growth, and this requires some time – this is usually called the *lag phase*. The specific adjustment function we use here was suggested by Baranyi and Roberts due to being both computationally convenient and having biological meaning: is the initial amount of some molecule (nutrient, enzyme, etc.) that is required for growth; is the rate in which this molecule is accumulated.

The Baranyi-Roberts differential equation has a closed form solution:

(2a)

. (2b)

We use four versions of the Baranyi-Roberts model. The full model, BR6, is described by eq. 2 and has six parameters. BR5 is the model in which the deceleration parameter is set to one , as in the classical logistic model. This model is useful because the reduced growth during the lag phase might sometimes be inferred as . BR4 is the mode without a lag phase, . This is also known as the Richards model or the generalized logistic model. This model is useful in cases where there is no observed lag phase – either because the organism in question adjusts very rapidly or because the population has been adjusted prior to the growth experiment by priming it in fresh media. The last model is BR3, in which and . This is simply the classic logistic model, . Fig. 2 presents the hierarchical relationship between the models.

### Model fitting and selection

We fit all four models to the mean growth curve of each strain using a least-squares procedure (LMFIT CITATION). The standard deviation at each time point is used as weights for the least-squares procedure so that time points with lower variance are more heavily weighted and therefore are fitted better.

We then calculate the Bayesian Information Criteria (BIC) of each model fit:

,

where is the number of parameters of the model, is the number of time points , is the average growth at time point , and is the expected growth at time point according to the model. We select the model with the lowest BIC.

As a sanity check, we also fit the data using simple linear regression () and check that the BIC of our selected model is larger than the BIC of the linear regression by at least 6 (Kass and Raftery 1995).

We repeat the model fitting procedure for the growth curves data of each strain to produce estimates for all six parameters as well as confidence intervals on these parameters for each of the strains.

### Competition prediction

We introduce the two-species Baranyi-Reoberts model, which, to the best of our knowledge, has not been used before:

(3a)

(3b)

, (3c)

, (3d)

where is parameter of strain – in which we plug in the estimated values from the model fitting procedure. This equation system is then solved by numerical integration, resulting in a prediction for the competition dynamics.

This two-species competition model explicitly assumes that all interactions between the two species (or strains) can be attributed to *resource competition* and therefore captured by the deceleration of growth rate of each species in response to growth of the other species. We do not however assume the same limiting resource or resource efficiency for both species as we use different maximum density for each strain.

### Selection coefficient inference

There are different ways to infer selection coefficients (or relative fitness) from competition dynamics. For example, Chevin (2011) suggests that the selection coefficient of the strain of interest is

where is the frequency of the strain of interest and is the frequency of the strain of interest.

Lenski and colleagues use a different measure:

. (4)

Crow and Kimura (1970) suggest that the frequency of the strain of interest follows a logistic equation:

(5)

Following this suggestion, we fit eq. 5 to the competition prediction after setting to find an estimate for the selection coefficient *s*. Of course, the competition prediction can be readily used to infer fitness in other ways if required.

## Experimental design

### Strains, plasmids and media

We use two strains of *Escherichia coli* –DH5α and TG1.

We insert one of two plasmids to the bacteria. Both plasmids code for resistance to Kanamycin (Kan) and Chloramphenicol (Cap), and they differ by the constitutive flouresence marker they contain – GFP or RFP. The plasmids were inserted into electrocompetent bacteria, which were than grown for one hour in rich media and then overnight. Bacteria cultures were than frozen in -80° Celsius in glycerol.

All growth experiments, as well as overnight culturing, was performed in LB rich medium with Kan and Cap, 30° Celsius, and shaking.

### Growth curve experiment

Bacteria from -80° Celsius glycerol stock were inoculated in 20ml flasks and grown overnight. The next day the overnight cultures were inoculated to a 96-well microplate. Columns 1-4 were inoculated with bacteria with a GFP marker; columns 9-12 with bacteria with a RFP marker; columns 5-8 with both strains of bacteria. Wells G12 an H12 were not inoculated to serve as a blank measurement. Each well contained 100ml of fresh filtered media and 3xx µl of bacteria; columns 5-8 had half the amount of GFP marked bacteria and half the amount of RFP marked bacteria.

The microplate was then placed in an automatic plate reader (Tecan Infinity F200) for 8 hours. Every 10 minutes the plate reader took an OD measurement of the entire plate. Every roughly 60 minutes we stopped the reader and took 1µl sample from row number X into a different row in a separate microplate containing 100 µl of PBS buffer in each well. See supplementary material for the plate reader configuration file.

### Flow cytometry measurements

The second microplate, containing samples of all 3 cultures (GFP, GFP+RFG, RFP) from 8 different time points, was inserted into a Fluorescence Automatic Cell Sorter (FACS) machne (MacsQuant XXX). The FACS machine then quantified the fluorescence of 100,000 cells from each well, roughly 1,000 cells per second, using the X and X filters. See supplementary material for the analysis and measurement definition files.

### Flow cytometry analysis

The flow cytometry data (*fcs* files) were analyzed using the [FlowCytometryTools](http://gorelab.bitbucket.org/flowcytometrytools/) Python package to calculate the frequency of red versus green fluorescent cells at each well at each time point. The frequencies were then averaged over the four wells from columns 5-9. See supplementary material for an IPython notebook of this analysis.

# Discussion

Because the Baranyi-Roberts has both a differential equation form and a closed form solution it is highly advantageous for our use: the closed form is used to fit to growth curve data; the differential equation is used to predict the competitions.