Modelling Competition for Nutrients between Microbial Populations Growing on Solid Agar Surfaces

Author: Daniel Boocock; Supervisor: Dr Conor Lawless

August 16, 2016

1 1 INTRODUCTION

The bacteria *Eschericia Coli* and yeast *Saccharomyces cerevisae* are unicellular organisms studied as a model prokaryote and eukaryote respectively. They grow in colonies, where cells may (be clones originating from a single cell or) belong to different genetic strains originating from different individual cells. In favourable conditions, growth is exponential and this makes growth rate a major component of fitness; faster growing strains quickly come to dominate the population. At a certain point growth becomes limited and a stationary phase is reached. For unicellular organisms, growth rate is equal to cell cycle progression rate and all of the genetic information must be copied before each division. As a result, evolutionary pressure has led to rapidly dividing organisms with compact genomes of essential genes. These genes have been conserved in other species over billions of years of evolution, which is, in part, what makes *E. Coli* and *S. cerevisae* useful as model species. The eukaryote *S. cerevisae*, is particularly useful for the study of other eukaryotes such as humans.

The growth rate of microbial organisms is measurable and is often used to determine fitness. In experiments, cell cultures are commonly grown in two types of medium: on the surface of a nutrient rich solid agar and in a liquid mixture containing nutrients. (REMOVE: In spot tests (phenotypic array), cultures are pinned or inoculated on the surface of a solid agar containing nutrients. In liquid culture assays, cultures are mixed in a liquid medium containing nutrients.) In both cases cultures are incubated and growth is observed. Identical strains can grow differently between the two mediums and disagreement in fitness estimates is currently an issue Baryshnikova *et al.* (2010a) (I couldn't find a paper specifically talking about this issue but they have a correlation plot Fig2a where correlations are worse with a liquid culture study by Jasnos and Korona; in fact the Baryshnikova paper Fig3c seems to say that they had strong correlation in their "high-resolution liquid growth profiling study"). I do not focus on this issue and exclusively study fitness screens using solid agar.

Fitness estimates can be used to infer genetic interaction or drug response and high-throughput methods allow this to be conducted on a genome-wide scale (see e.g. Costanzo *et al.* (2010); Andrew *et al.* (2013)). In a typical genetic interaction screen a strain is made with a mutation in a query gene. Double mutants are created by introducing a second deletion in this strain. By comparing the growth of double mutants with a control containing a neutral deletion, genetic interactions can be inferred. If a strain is fitter than the control then the deletion is said to suppress the defect of the query gene. If a strain is less fit than the control then the deletion is said to enhance the defect of the query gene. Either scenario suggests that the two genes interact and have a related function. Due to redundancy, single deletions are often non-lethal. (Remove: Knock downs and conditional mutations can also be used.) This has allowed Costanzo *et al.* (2010) to explore genetic interactions for $\sim$75% of the *S. cerevisae* genome.

Synthetic Genetic Array (SGA) and Quantitative Fitness Analysis (QFA) are high-throughput methods for obtaining quantitative fitness estimates of microbial cultures grown on solid agar (Baryshnikova *et al.*, 2010b; Banks *et al.*, 2012). Typically one query gene and replicates of several deletions are pinned or inoculated in a rectangular array on a solid agar plate. Many plates with different query genes and deletions are grown in high-throughput to explore whole genomes. I study data from QFA which refers to quantitative estimation of fitness by measurement and fitting of growth curves. In a typical QFA procedure liquid cultures are incolulated onto solid agar (containing nutrients (already mentioned above)) in a 16x24 rectangular array of 384 spots. Inoculum density can be varied to capture more or less of the growth curve and the most dilute cultures are inoculated with $\sim$100 starting cells (Addinall *et al.*, 2011). Plates are grown in incubation and removed to be photographed at timepoints throughout growth. Photographs are of whole plates and growth typically covers several days to capture both the exponential and stationary growth phases. Colonyzer (Lawless *et al.*, 2010) processes optical density measurements in photographs to produce a timecourse of cell density estimates for each culture. In pasts analysis, the logistic growth model was independently fit to the timecourse of each culture and fitness estimates were defined in terms of parameters of this model: the growth constant r and carrying capacity K. In contrast, SGA typically uses a larger array of 1536 pinned cultures and a single endpoint assay of culture area to quantify growth. The the differential form and solution of the logistic model (Verhulst, 1845) (probably don't need this reference) are given in Equations 1, where C represents cell density and C_{t_0} is cell density at time zero.

$$\dot{C} = rC\left(1 - \frac{C}{K}\right) \tag{1a}$$

$$C(t) = \frac{KC_{t_0}e^{rt}}{K + C_{t_0}(e^{rt} - 1)} \tag{1b}$$

The logistic model is a simple mechanistic model describing self-limiting growth and has a sigmoidal solution. Growth begins exponentially with rate rC and curtails as the population size increases and cells begin to compete for space and nutrients (remove: or interact in some other way). Cell density reaches a final carrying capacity K at the stationary

phase. In QFA, nutrients must diffuse through agar to reach cells growing on the surface. It is plausible that the carrying capacity K represents the point at which nutrients either run out or growth becomes limited by the diffusion of nutrients and is approximately stationary. Fitting the logistic model to QFA data requires plate level or culture level parameters for C_{t_0} and culture level parameters for r and K making 769 or 1152 parameters per 384 culture plate.

//Could remove and just discus MDR when I get to the results// The growth constant r could be used as a fitness measure. However, Addinall *et al.* (2011) define a more complicated fitness measure as the product of Maximum Doubling Rate (MDR) and Maximum Doubling Potential (MDP) which they calculate from logistic model parameters. MDR measures the doubling rate at the beginning of the exponential growth phase, when growth is fastest, and MDP is the number of divisions which a culture undergoes from inoculation to the stationary phase.

$$MDR = \frac{r}{log\left(\frac{2(K-C_0)}{K-2C_0}\right)} \tag{2a}$$

$$MDP = \frac{log\left(\frac{K}{C_0}\right)}{log(2)} \tag{2b}$$

To improve the quality of fits, QFA now uses the generalised logistic model which requires an extra shape parameter for each culture. Standard and generalised logistic model r are not equivalent so comparison relies on MDR and MDP as fitness measures. The analysis of QFA data using both models is available through the QFA R package (Lawless *et al.*, 2016). //Could remove and just discus MDR when I get to the results//

//Could remove//Addinall *et al.* (2011) used QFA and *S. cerevisae* to screen for genes involved in telomere stability which is related to ageing and cancer and has implications for human health and disease. Hits from this study have been successfully followed to discover new biology (Holstein *et al.*, 2014). (To be honest I have no idea what they found in that paper. Maybe I should leave this to biologist and out of my report? We had a more general focus. Obviously I will mention the Addinall paper when I describe p15 in the methods.) //Could remove//

Since QFA aims to determine differences in the fitness of microbial strains from measurements of differences in growth, fast and slow growing cultures are often grown side-by-side. Figure 1 shows a section of a QFA plate from a study by Addinall *et al.* where this is the case. (Cultures were inoculated with approximately equal cell density but have grown at different rates to visibly different sizes after $\sim$2.5 days.) It is likely that nutrients diffuse along gradients between fast and slow growing neighbours causing growth to appear faster or slower than if it were independent. Further support for such an effect comes from the experiment shown in Figure 2 where the same cultures are grown in alternate columns on two separate plates but with cultures added or removed from the neighbouring columns in between. Cultures in Figure 2a), where neighbours were removed, grew faster and larger (how much? I can look at the data myself) than the same cultures in Figure 2b), where neighbours were added. This suggests that an interaction between neighbours is present and may be affecting fitness estimates. Current QFA analysis using the logistic model assumes that cultures grow independently and ignores possible competition effects between neighbours. The sigmoidal curve of the logistic model poorly fits QFA data in many cases and this may be due to competition effects. I aim to fit a network model of nutrient dependent growth and diffusion to QFA data to try to correct for competition and increase the accuracy and precision of fitness estimates.

Could explain the difference in dilute and more concentrated cultures. In the image captions or elsewhere?

Could also talk about quorum sensing and amonia when I get to competition.

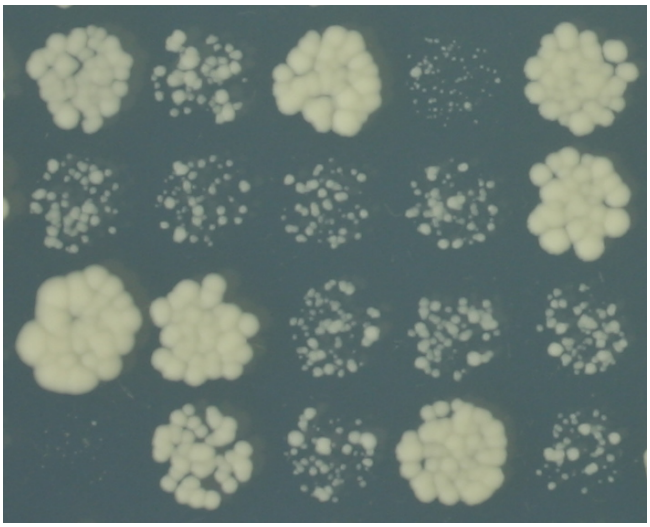

Figure 1: 4x5 section of a QFA plate. *Taken from a 16x24 format solid agar plate inoculated with dilute S. cerevisae cultures. Image captured at $\sim$2.5d after inoculation and incubation at 27°C.*

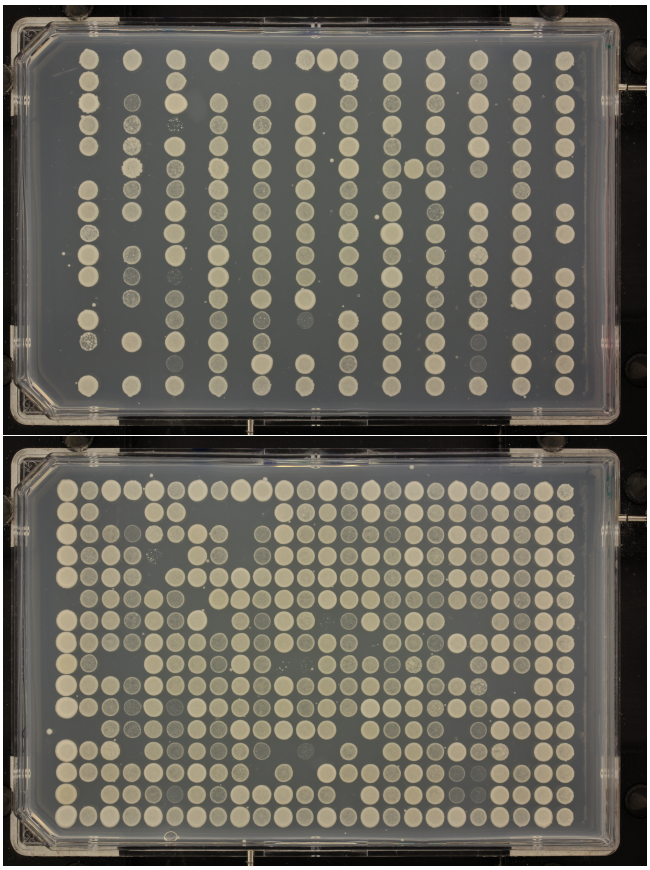

Figure 2: QFA experiment designed to examine competition. *A) QFA plate inocluated with a more concentrated S. Cerevisae inoculum (no cells inocluated on alternate columns). B) Same as in A, but with strains of similar growth rate inoculated in the positions missing in A.*

Competition effects could be dealt with experimentally by randomising the location of cultures on repeated plates. This does not require explicit knowledge or modelling of the source of competition but reduces throughput, so, if possible, a modelling approach is desirable. Poisoning of cultures by a signal molecule such as ethanol, which *S. cerevisae* produces in the metabolism of sugars by fermentation, is another possible source of competition. QFA does not measure nutrients or signal, so if more than one source of competition exists, it becomes very difficult to fit a model and randomisation may be the best approach. QFA data for edge cultures is noisy due to reflections from plate edges. This is only partially corrected for by Colonyzer (Lawless *et al.*, 2010) and as a result data for edge cultures is usually discarded. Addinall *et al.* (2011) grow repeats of a neutral deletion in edge locations, rather than leaving them empty, because of concerns about competition. In an SGA study, Baryshnikova *et al.* (2010a) use statistical techniques to correct for competition between fast and slow growing neighbours in end-point assays of culture area. I hope that modelling competition for nutrients explicitly will better correct for competition using fewer repeats. QFA uses more information than SGA by fitting whole growth curves, rather than a single endpoint assay, so a modelling approach promises to be more powerful. Furthermore, modelling may identify and explain the source of competition. Simulation of an accurate model will allow comparison of experimental designs and exploration of ways to reduce competition effects.

//Diffusion Equation: I am probably going to have to repeat this when I get to the discussion so I could just leave until then.// Reo and Korolev (2014) use a diffusion equation model to simulate nutrient dependent growth of a single bacterial culture on a pertri dish in two-dimensions. It would be too computationally intensive to fit a similar model to a full QFA plate in three-dimensions, especially if the model is to be used to process many plates from high-throughput experiments. Therefore, a simpler model of nutrient diffusion is required. //Diffusion Equation//

Lawless (link blog) proposed a model of nutrient dependent growth and competition (3,4), hereinafter the competition model, using mass action kinetics and network diffusion. A schematic of the model is drawn in Figure 3. He represents the nutrient dependent division of cells with the reaction equation,

$$C + N \xrightarrow{b} 2C, \tag{3}$$

where C is a cell, N is the amount of nutrient required for one cell division, and b is a rate constant for the reaction. (The identity of the limiting nutrient N is unknown but possible candidates are sugar and nitrogen.) He defines separate reactions (3) with growth constant b_i for each culture, indexed i, on a plate and uses mass action kinetics to derive rate equations for the amount of cells and nutrients associated with each culture, C_i and N_i. This gives the rate equation for C_i (4a) and the first term in the rate equation for N_i (4b).

$$\dot{C}_i = b_i N_i C_i, \tag{4a}$$

$$\dot{N}_i = -b_i N_i C_i - k_n \sum_{j \epsilon \delta_i} (N_i - N_j). \tag{4b}$$

To arrive at the full competition model, he models the diffusion of nutrients along gradients between a culture i and its closest neighbours δ_i by the second term in (4b), where k_n is a nutrient diffusion constant. This can also be expressed as a series of reactions of the form

$$N_i \xrightarrow{k_n} N_j \tag{5a}$$

$$N_j \xrightarrow{k_n} N_i, \tag{5b}$$

and modelled with mass action kinetics. Unlike the logistic model (1), the competition model has no analytical solution, and must instead be solved numerically. If k_n is set to zero, the competition model reduces to the mass action equivalent of the logistic model, hereinafter the mass-action logistic model, (and has the same sigmoidal solution). (In this limit, parameters of the competition model can be converted in terms of parameters the logistic model (see methods section)). When the competition model is fit to QFA data, C_i is observed and N_i is hidden. Inoculum density, C_{t_0}, is often below detectable levels. By assuming that incolulum density is the same for all cultures and that nutrients are distributed evenly throughout the agar at time zero, plate level initial values of cells and nutrients, C_{t_0} and N_{t_0}, can be used. k_n is assumed to

be constant across the plate but must be inferred. There is a growth constant, b_i, for each of 384 cultures on a typical QFA plate making 387 parameters in total. The competition model shares more information between cultures and has less than half the number of parameters of either the standard or generalised logistic model (Banks *et al.*, 2012; Lawless *et al.*, 2016). (If I remove the section on MDR and the generalised logistic model above I will need to add a line of explanation here.)

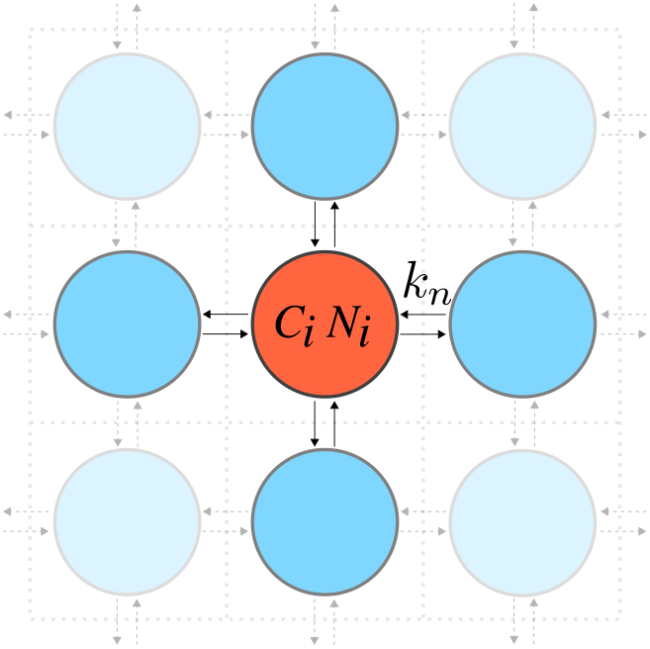

Figure 3: Schematic of the competition model. *Each circle represents a culture, indexed i, growing in a rectangular array on the surface of a nutrient containing solid agar. Arrows represent a network of nutrient diffusion along gradients between cultures. C_i - amount of cells; N_i - amount of nutrients; k_n - plate level nutrient diffusion constant; darker blue circles δ_i - closest neighbours of culture i.*

In QFA, populations begin with ∼100 cells and quickly grow to reach thousands of cells so a deterministic approximation appears valid. Mass action kinetics applies to reactions in a well stirred mixture and is perhaps less valid for a culture growing on solid agar. However, a mass action approximation has been successful in other situations where this assumption is questionable: in the Lotka-Volterra model of predator-prey dynamics (Berryman, 1992) and in signalling and reaction models inside cells (Aldridge *et al.*, 2006; Chen *et al.*, 2010). The order of a reaction also affects the rate equation and the identity and quantity of the nutrient molecule in the (3) is unknown. Reaction (3) also assumes that all nutrients are converted to cells and includes no model of metabolism. I justify the use of the competition model because in the independent limit it has the same solution as the logistic model which has long been used to model microbial growth. Studying the competition model may help us to understand the nature of QFA experiments and, if some assumptions do not hold, it could be used as a first step in developing a more accurate model. Furthermore, collectively fitting the competition model involves a the large number of parameters and data points and will require many simulations to be run. This necessitates the use of an approximate model for computational feasibility. This is especially true if the model is to be used in the analysis of high-throughput data. It is hoped that even an approximate model will be able to measure more reliable growth parameters and better estimate fitness. This will increase the power to infer genetic interaction and drug response which could lead to further discoveries. (For an example of a sucessful QFA study and follow up using the logistic model see Addinall *et al.* (2011) and Holstein *et al.* (2014)).

2 2 METHODS

To analyse QFA data I created a Python package (“CANS”) for model composition, model simulation, parameter inference, and visualisation of results. It accepts cell density timecourses for any size rectangular array. CANS can produce SBML models to document results of parameter inference or for independent validation. It is relatively simple to create and simulate new models, if they involve reactions between species within cultures or between neighbouring cultures, and to fit these provided an initial guess. The CANS package, containing source code for the results in this paper, is available at https://github.com/lwlss/CANS.

2.1 2.1 Tools, solving, and fitting

2.1.1 Solving

I use CANS to parse QFA data after processing with Colonyzer (Lawless *et al.*, 2010). Colonyzer processes series of time-stamped whole-plate images, using integrated optical density measurement as a proxy for cell density, to produce timecourses of cell density for all cultures on a plate. I use these cell density estimates, which have arbitrary units, throughout my analysis. CANS numerically solves models using one of two methods. The first is slower and uses SciPy’s integrate.odeint to solve models written in Python at user supplied timepoints. I vectorised code using NumPy to optimise solving of the competition model by this method. For solving a plate of 384 cultures with cell density observations at 10 unevenly spaced time points, I found an approximately 10 times further increase in speed using the Python bindings for the package libRoadRunner. libRoadRoadrunner’s RoadRunner.simulate requires models to be written in SBML so I wrote code using the libSBML Python API to automatically generate SBML versions of the competition model for any size plate.

Unlike SciPy’s odeint, libRoadRunner only simulates at uniformly spaced timepoints. To fit QFA timecourses, which do not have fixed intervals, requires simulated cell amounts at the observed timepoints. For the analysis in (P15 section), where each timecourse has only 10 timepoints, I simulated using a function call between each pair of adjacent timepoints. This method was slower for the analysis in (Stripes section) where each timecourse had around 50 timepoints. To increase speed I used SciPy’s interpolate.splrep to make a 5th order B-spline of cell density timecourses with smoothing condition $s = 1.0$. I evaluated the spline for cell density using SciPy’s interpolate.splev at 15 evenly spaced intervals from time zero

to the time of the last QFA observation. I then solved these timecourses with one call to RoadRunner.simulate.

2.1.2 Fitting

To fit the competition model to QFA data, I made maximum likelihood estimates of parameters using a normal model of measurement error. This used the L-BFGS-B constrained minimisation algorithm from SciPy's integrate package. I determined stopping criteria so that parameters of full-plate simulated data sets, with a small amount of simulated noise, were recovered with high precision.

To fit each plate I used a range of initial parameter guesses (described in Section (guessing section)).

Table 1: Parameter bounds. *Used for fitting the competition model to P15 and the Stripes and Filled plates. Bounds on N_{t_0} were applied to both $N^I_{t_0}$ and $N^E_{t_0}$ for internal and edge cultures.*

Parameter	Lower	Upper
C_{t_0}	guess x 10^{-3}	guess x 10^3
N_{t_0}	guess / 2	guess x 2
k_n	0.0	10.0
b (all cultures)	0.0	∞

//Delete this note: I was supposed to change the parameter bounds on b after fitting p15 (once we had a rough idea) but forgot. I was using the same script to fit the logistic equivalent model where we do get large b and need heuristic checks. This probably hasn't made any difference though as we don't get stupidly high b values and fits look quite good. Imaginary neighbour guessing of b does use stricter bounds.//

Table 1

I bound all growth constants b to be above zero and parameters to be between zero and above zero and made sure that I always had fits where no parameters hit the bounds.

QFA data for the Stripes plate (sec ref) contained observations for cultures that were known to be empty. When fitting the competition model (4), I set growth constant b to zero for these cultures and removed them from the objective function.

2.2 2.2 Parameter conversion

(Could move to discussion: The identity of the nutrient molecule is unknown and it is not clear whether metabolism of the nutrient molecule will have a significant effect. If necessary a metabolism reaction could also be modelled.)

When k_n is set to zero, the competition model (4) reduces to the mass action logistic model which has the same sigmoidal solution as the standard logistic model. In this limit, it is possible to equate cells of both models and convert parameters using (6) (see Conor's blog for a derivation).

$$r_i = b_i(C_{t_0} + N_{t_0}) \tag{6a}$$

$$K = (C_{t_0} + N_{t_0}) \tag{6b}$$

The reaction equation of the competition model (3) assumes that all nutrients are converted to cells. This implies that all cultures starting with the same amount of nutrients reach the same final amount of cells. Therefore, to fit the mass action logistic model to QFA data, it is necessary to allow N_{t_0} to vary for each culture which is not physical and, in which case, the mass action logistic model has the same number of parameters (769) as the standard logistic model. (Probably repetition: When I fit the competition model I collectively fit the timecourses of all cultures on a plate using a plate level N_{t_0} and 387 parameters.) Figure 4 shows fits of a single culture on a larger 16x24 format plate using both models. This culture grew faster than its neighbours (not shown) and, according to the competition model, competed for more nutrients. Figure 4a shows the mass-action logistic model fit where N_{t_0} is estimated as being approximately equal to the final cell amount, or carrying capacity K. Figure 4b shows the

competition model fit with a plate level N_{t_0} and $kn > 0$. Re-simulating with k_n set to zero gives the dashed mass action logistic model curves which are corrected for competition. We can therefore obtain the corrected logistic model r_i and K_i of these curves by converting from competition model estimates of b_i, C_{t_0}, and N_{t_0}. N.B. b is the same for both the solid and dashed curves in Figure 4b.

Competition model C_{t_0} and N_{t_0} are the same for all cultures on a plate. Therefore, by the conversion equations (6), all cultures on a plate have the same carrying capacity K and all $b_i \propto r_i$ by the same factor. Similarly, MDP is the same for all cultures and all $b_i \propto MDR_i$ by the same factor (see Equation 2). Therefore, b is equivalent to all common QFA fitness measures, r, MDR, and $MDR * MDP$ (see e.g. Addinall *et al.* (2011) and Lawless *et al.* (2016)). This makes b a very convenient fitness measure for the competition model; we need not convert to logistic model parameters to compare the fitness rankings of cultures on the same plate. To compare competition model fitness rankings between different plates we can of course use b. However, this is not equivalent to comparing r or MDR as different plates may have different C_{t_0} and N_{t_0}.

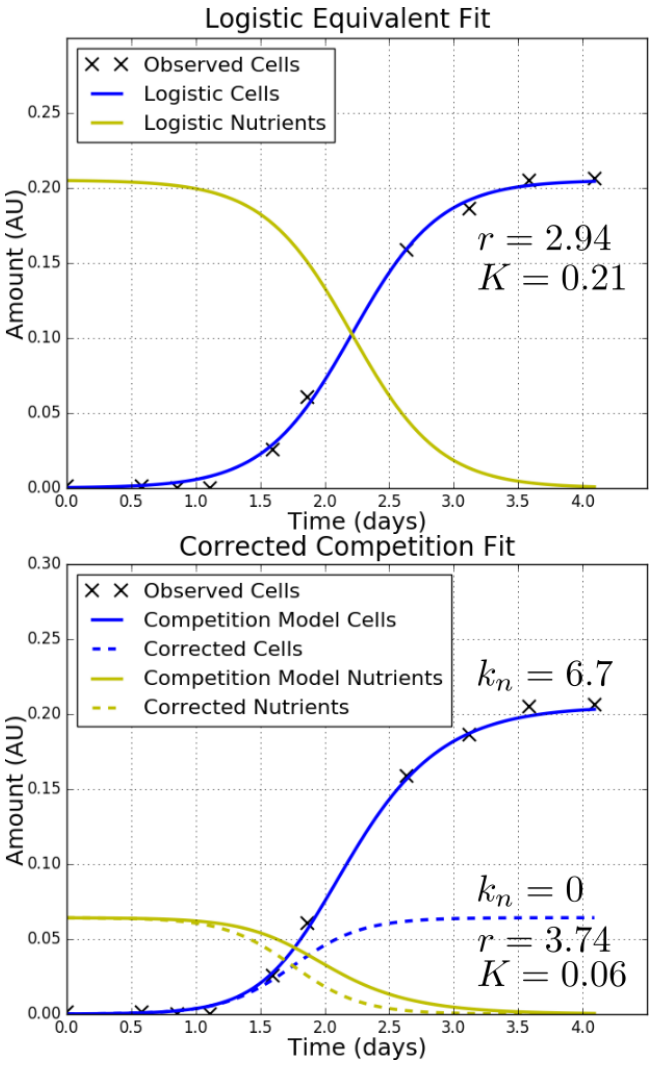

Figure 4: Using the competition model to correct for competition. *Fits are to culture (R10, C3) of P15 which grew faster and reached a higher final cell density than its neighbours (not shown). According to the competition model, this is because this culture competed for more nutrients. To reach the same final cell density, the logistic equivalent model requires a higher amount of starting nutrients for this culture and a different amount for each neighbour. The correction to the competition model simulates how growth would have appeared without competition and allows us to return parameters r and K of the logistic model.*

2.3 2.3 Making an initial guess

2.4 2.4 Development of a genetic algorithm

2.5 2.5 Model comparison using a single QFA plate

2.6 2.6 Cross-plate calibration and validation

REFERENCES

Addinall, S.G. *et al.* (2011) Quantitative fitness analysis shows that nmd proteins and many other protein complexes suppress or enhance distinct telomere cap defects. *PLoS Genet*, **7**, 4, 1–16.

Aldridge, B.B. *et al.* (2006) Physicochemical modelling of

cell signalling pathways. *Nature cell biology*, **8**, 11, 1195–1203.

Andrew, E.J. *et al.* (2013) Pentose phosphate pathway function affects tolerance to the g-quadruplex binder tmpyp4. *PLoS ONE*, **8**, 6, 1–10.

Banks, A. *et al.* (2012) A quantitative fitness analysis workflow. `http://www.jove.com/video/4018/a-quantitative-fitness-analysis-workflow`.

Baryshnikova, A. *et al.* (2010a) Quantitative analysis of fitness and genetic interactions in yeast on a genome scale. *Nature Methods*, **7**, 12, 1017–24. Copyright - Copyright Nature Publishing Group Dec 2010; Last updated - 2014-09-19.

Baryshnikova, A. *et al.* (2010b) Synthetic genetic array (sga) analysis in saccharomyces cerevisiae and schizosaccharomyces pombe. *Methods in enzymology*, **470**, 145–179.

Berryman, A.A. (1992) The orgins and evolution of predator-prey theory. *Ecology*, **73**, 5, 1530–1535.

Chen, W.W. *et al.* (2010) Classic and contemporary approaches to modeling biochemical reactions. *Genes & development*, **24**, 17, 1861–1875.

Costanzo, M. *et al.* (2010) The genetic landscape of a cell. *science*, **327**, 5964, 425–431.

Holstein, E.M. *et al.* (2014) Interplay between nonsense-mediated mrna decay and {DNA} damage response pathways reveals that stn1 and ten1 are the key {CST} telomere-cap components. *Cell Reports*, **7**, 4, 1259 – 1269.

Lawless, C. *et al.* (2010) Colonyzer: automated quantification of micro-organism growth characteristics on solid agar. *BMC Bioinformatics*, **11**, 1, 1–12.

Lawless, C. *et al.* (2016) *qfa: Tools for Quantitative Fitness Analysis (QFA) of Arrayed Microbial Cultures Growing on Solid Agar Surfaces*. R package version 0.0-42/r678.

Reo, Y.J. and Korolev, K. (2014) Modeling of Nutrient Diffusion and Growth Rate in Bacterial Colonies.

Verhulst, P. (1845) Recherches mathematiques sur la loi d'accroissement de la population. *Nouveaux memoires de l'Academie Royale des Sciences et Belles-Lettres de Bruxelles*, **18**, 14–54.