GeneLink: Supplemental Information and Analysis

Draft

Greg O'Brien[1], Robert Shields[2], Natalie Maricic[2], Robert A. Burne[2],
Stephen J. Hagen[1]

February 28, 2018

Abstract

GeneLink, a program for robust fitting and large-scale comparison of growth curves, attempts to address the issues surrounding cell growth variability and multivariate comparisons of big data sets. We created a novel way for clustering bioinformatics data by combining multiple cluster results, offering improved flexibility and performance over existing methods. Testing several data sets, we obtain results better than current clustering methods. For the purpose of this paper, we will be using optical density at 600nm as an indicator of the total viable cell density.

1 Clustering

In simple terms, clustering is the process of grouping similar objects together. Many applications of clustering exist to help organize data into meaningful relationships. Internet search engines, genetics, artificial intelligence, image-object recognition, investment banking and fraud detection are the most wildly uses of clustering. Often the type of clustering is strongly dependent on the application and the data, requiring trial and error.

GeneLink, a software tool for analyzing growth curves and clustering similar experiments together. Made available to the end-user, a large assortment of clustering algorithms are

automatically combine into a robust, agglomerative, hierarchal result. This approach, not used in any other tools, removes the guess-work out of clustering.

Internet search engines use clustering to group similar URLs. In genetics, clustering is used for finding families of genes. AI programs also use clustering for processing raw data, such as image object recognition. Clustering has also proved useful in fraud detection and investment banking.

2 Growth Dynamics

2.0.1 Lag Phase is the first growth stage encountered by colonizing bacteria. Cells exposed to new environments may require additional inter-cellular machinery, such as enzymes, for growth causing a delay. Similarly, previously stressed cells may also need additional time to exit Survival Phase. A systematic trade-off exist between cells growing early vs cells investing additional resources to produce better machinery. The result is a stochastic population modeled by exponential growth, at early times. Stochastic populations offer a survival strategy that takes advantage of quickly changing environments while maximizing growth under favorable conditions. Lag Time is thought to be independent of cell concentration [**?**]. With low cell concentrations, exponential growth only results in a few additional cells, where it may be difficult to accurately determine Lag Time.

2.0.2 Growth Phase occurs as cells begin growing, exiting Lag Phase. Cells divides at a rate limited by the slowest enzymatic step in the cell metabolism [1]. With bottlenecks in metabolism, energy sources often influence the maximum growth rate achieved. Other factors affecting cellular growth rates include ambient temperature, PH, and nutrient concentrations [1]. Competing cells may release toxins, signaling molecules or waste products that may also change cellular activity. [say something about stochasticity?]

2.0.3 Stationary Phase exists as an equilibrium based on nutrient and cell concentrations. The number of cells remains constant, with the number of dying cells equaling the number of new cells produce by division. The cell concentration at which Stationary Phase

occurs is considered the carrying capacity of the environment. Cells can change their surroundings to increase the of time spent in Stationary Phase, sometimes resulting in biofilms. In biofilms, a complex extracellular matrix protects cells form external stressors. Biofilms serve as physical barriers trapping nutrients and blocking competing populations. Additionally, in biofilms, ph, nutrient and oxygen levels can vary widely over short distances [2]. When environments do not have a steady supply of incoming nutrients decreasing nutrients and accumulating waste products often results in a shorter Stationary Phase, with Death Phase quickly following after carrying capacity is reached.

2.0.4 Death Phase occurs when insufficient nutrients are unable to support further growth for the majority of cells. A decreasing total cellular population results, although some cells continue to grow. Additionally if cells have reached a critical size, they may still divide, even if they are no longer growing. [3] Nutrients and growth rates are often non homogeneous. The later, allowing some cells, with slower growth rates to still grow and divide in Death Phase. By-products from degrading cells may either serve as further nutrients for surviving cells or contain toxins and signal molecules, suppressing further growth. During Death Phase, a more nefarious survival strategy may occur, cannibalism. Cells will engage in same-cell warfare to acquire nutrients, beneficial DNA, and decreased competition overall.

2.0.5 Survival Phase takes place once an equilibrium is reached for remaining cells and nutrients. Surviving cells enter a period of decreased metabolic activity as a trade-off for prolonged survival. Cells will lie dormant, some fortifying their cellular membrane, waiting for favorable growth conditions to return. There is an inherent delay for cells to wake up from Survival Phase, termed the Lag Time, as previously discussed.

2.0.6 Environmental factors often determine the shape of growth curves. Cell growth is limited by the slowest metabolic step. [1] As consequence, the media in which cells grow affects the growth rate. The cell environment is also usually changing and cells may alter their metabolic activity once a certain threshold is reached. [] Additionally, hysteresis, the

previous state of the cells can also impact future activity. Cells may have a decreased Lag Time if they were previously grown in the same media verses cells who have never been exposed to a type media. []

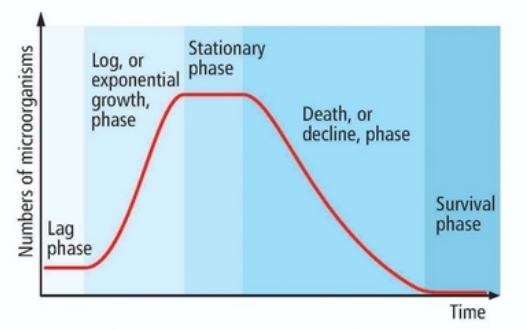

Figure 1: Graph showing different phases of a growth curve. It can be often difficult to decide where one phase finishes and the next one begins. Additionally phases can overlap if only some of the population changes their growth activity. The steepest slop of Growth Phase is often related to the rate limiting step of cellular metabolism [1]

3 Curve fitting techniques

3.0.1 Least squares fitting, by definition, is an optimization problem. While several strategies can be used to find minima, most fitting routines rely on minimizing differences between a model and data, known as residuals. Similar to the model, residuals form a plane in multidimensional space. The goal is to find the minimum of the plane. For simple functions this is easy where an exact, explicit solution exist, such as drawing a line across 2 points. However, if there are 2 or more points than free fit parameters, the solution is often non-exact (due to Cramer's rule). The problem becomes more difficult if saddle points and local minimas exist (as false global minimums). Additionally, with bounded searches, boundaries can be mistaken for global solutions. Bounded parameters also are usually not indicative of physical models. To address local minimas, most fit procedures start with an initial guess that is near the true solution.

3.0.2 Linear fit algorithms (not to be confused with linear regression) use the first derivative of individual single variables in the residual plane to locate minimums. Levenberg-

Marquardt uses the Jacobian to determine the path of steepest decent towards the minimum. Full Newton methods uses higher ordered derivatives to more accurately determine minimums. In most uses, these methods continue fitting until a criterion is meet. The criterion can be a given size of residuals, changes in parameters below a value, the number of iterations or CPU time.

3.0.3 Models determine the quality of fit, therefore, it is important to pick an accurate model. If the model doesn't vary over the domain of the data, it may be difficult to minimize. The number of free fit parameters is also important. Too many free parameters will over fit the data or lead to degeneracy in fit parameters. An example of over fitting are high order polynomials where additional parameters have no direct existential meaning.

3.0.4 Bootstrapping is a commonly used method for assessing fit quality. Data points are selected at random and used in the fit procedure. Over multiple trials, the fit will vary slightly because the data changes. The variance of the fit parameters indicates how robust the variables are. More robust parameters indicate more precise fits. The distribution of bootstrap parameters can indicate issues too. A good fit should have a narrowly centered maximum of bootstrap trials. Highly skewed distributions are indicative of outliers or systemic fighting errors. A bimodal distribution may indicate too many fit parameters or too few data points.

3.0.5 Additional measures of goodness: The F-Test is a statistical evaluation to determine if a model has too many free parameters used. [4]. If there are $\boldsymbol{n}$ data points for two models, with $\boldsymbol{p_1}$ and $\boldsymbol{p_2}$ free parameters, the F-statistic is given by,

$$\boldsymbol{F_{12}} = \frac{(\frac{\boldsymbol{rss_1 - rss_2}}{\boldsymbol{p_2 - p_1}})}{(\frac{\boldsymbol{rss_2}}{\boldsymbol{n - p_2}})} \tag{1}$$

where, $\boldsymbol{rss}$ is the residual sum square of each fit model. If the F-statistic is greater than the F-distribution of $\boldsymbol{p_2 - p_1, n - p_2}$ degrees of freedom at a desired significance level ($\boldsymbol{\alpha}$), we accept the new model as a more significantly accurate model.

χ^2 statistics can be additionally used to access robustness. Similar to the T-test, the χ^2 statistic is compared against a χ^2 distribution for a given number of replicates, r, to determine a P-value.

Coefficients of variation $c.o.v. = \sigma/\mu$) are also useful for describing robustness, where σ and μ are the standard deviation and mean for the data set. $c.o.v.$ of bootstrap and experimental replicates should be both small and different experiments should be large, indicating robustness and distinctness respectively. [previous sentence should be reworded?]

3.1 Clustering evaluation

3.1.1 Clustering types my be divided up in two distinct categories, Hierarchical (HC) and Non-Hierarchical clustering (NHC). HC methods examine differences in variables. Variables are treated as dimensions in a multivariate space with a metric applied, determining distances in the space. Groups are determined by the linkage method of distances, such as nearest neighbors. Groups are either combined in to larger groups (bottom up clattering) or divided into smaller sub-groups (top down clustering). Finally a tree is constructed showing the linkage of groups. One output of most clustering algorithms is a cluster vector, where the index corresponds to the original object and values indicate the associated groups.

There are several advantages to both types of clustering. HC can be quick and produce mappable trees for simple models, known as dendrograms (the Latin translation being, tree diagrams). Depending on the data, sometimes HC has trouble with more complicated relationships, such as negative correlation and inhibition. NHC may also be better for subspace clusters that span several dimensions, but do not cover the same vector space in every dimension.

The most common HC method uses a euclidean metric and forms groups based off the nearest neighbor. The most common NHC is K-means. Data is also treated as a multidimensional variable space, where centroids for groups are determined. Centroids can also be determined with different metrics.

There are several ways to evaluate clustering results and determine their cogency.

3.1.2 Errors may occur in 2 possible ways with clustering. False-negatives (FN) occur when the null hypothesis is rejected, but is true. In clustering, this would be removing an object from the correct group and placing it in another group, such as having only 3 aces, instead of 4. False-positives (FP) occur when the null hypothesis is false, but is calculated to be true. With clustering, FPs occur when you have placed a foreign object in the wrong group, such as 4 aces and a "2" (playing card). FNs will increase with forming more numerous, smaller clusters and FPs will increase with fewer, larger clusters.

3.1.3 F-scores, a non biased measure of validity, does not favor fewer or more clusters, while not technically needing to know the exact number of clusters either. F-scores compare both the rate of FNs, FPs and true positives (TP).

$$\boldsymbol{F_\beta = \frac{(1+\beta^2)\cdot \text{true positive}}{(1+\beta^2)\cdot \text{true positive} + \beta^2 \cdot \text{false negative} + \text{false positive}}} \tag{2}$$

With $\boldsymbol{\beta = 1}$, the F-score is maximized when the number of clusters is equal to the number of distinct elements. $\boldsymbol{\beta}$, the ratio between FNs and FPs may also be adjusted based on the application.

4 Object Filtering

5 Test Data

The first, Data set I, consisted of transposon insertion mutants [**?**]. Individual genes were removed from the bacterial chromosome, by inserting antibiotic resistance genes. Strains were selected via a P-Gal blue white assay. OD_{600} were recorded in the presence of spectinomycin in biological replicates of 3 for up to 24 hours, with and without CSP (an extra cellular signaling molecule [6]). Data set I performed the most poorly in terms of clustering ($\boldsymbol{\bar{f}_1 = .52}$). The transposon polarity was found to have an impact on observed phenotypes, for some genetic regions. The second, Data set II contained TNseq mutants, where a transposon is randomly inserted into chromosomes [7]. Mutated cells where grown with and

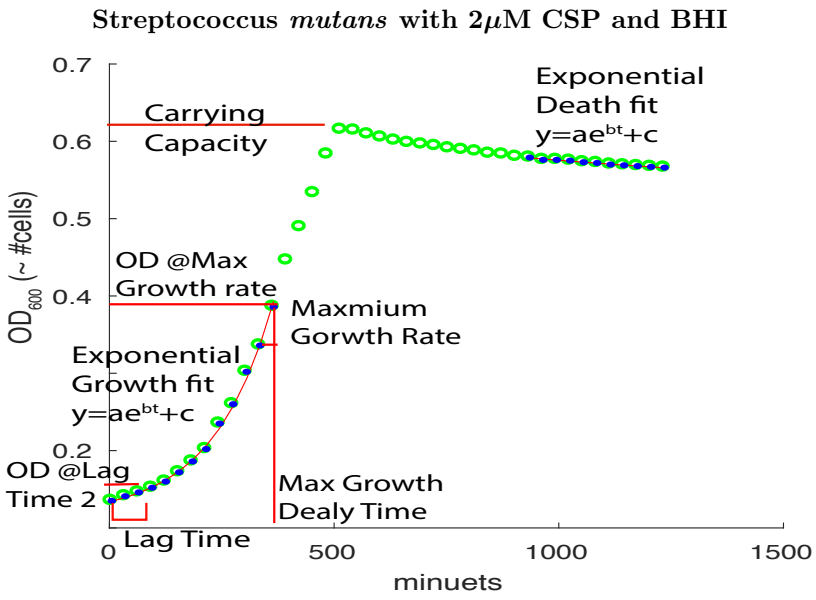

Figure 2: A growth curve used as a control during a representative fit procedure. The measured points are green circles. Exponential fits are determined by using points with blue centers. The Growth Phases are considered as the region from the minimum, to the point of inflection ($\boldsymbol{d^2y/dt^2 = 0}$). The carrying capacity is treated as the maximum. Points near the carrying capacity are not used. The points used for the Death Phase fit start at the mid-time point between the carrying capacity and the last observation time to the last time point. Here, a majority of the cells are assumed to be in the Death Phase. The Lag Time may be determined by using 3 different methods. Method 1 is a direct calculation of Lag Time, $\boldsymbol{-ln(a)/b}$. Method 2 uses the intersection of the initial starting OD and the steepest slop (maximum growth rate). Method 3 only considers the time required for the first OD change of 0.1. For better time resolution, Method 3 incorporates a spline fit, instead of using only the measured OD. The curvature (not shown) is calculated by the method outlined in [5]

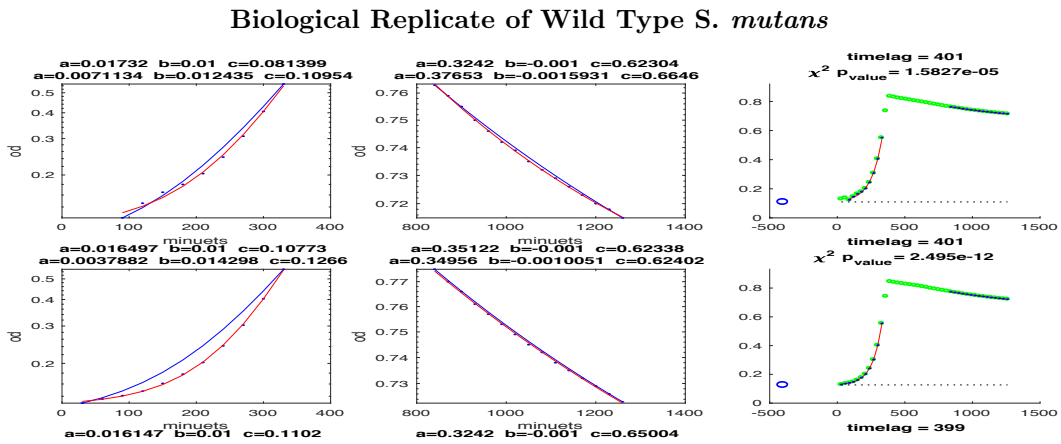

Figure 3: Fits of 2 biological replicates. An example output of GeneLink. Experimental variability causes no two estimates to be the same, demonstrating the need of a robust analysis method for biological replicates.

without the molecule CSP and then ran though PCR. Mutants with increased transposon expression under CSP were selected for. OD_{600} was measured up to 24 hours, with and without CSP. Data set II contained the least noise and performed well in terms of clustering accuracy. Data set III consisted of strains with selected growth defects grown under various conditions; BHI, 25mM paraquat, .003% H2O2, ph=5.5, 42c, CSP 2uM, and without an oil overlay. Data set III contained high noise levels, along with several cultures not growing or reaching carrying capacity in 24 hours. [Probably best if removed for the purposes of a supplemental paper?]

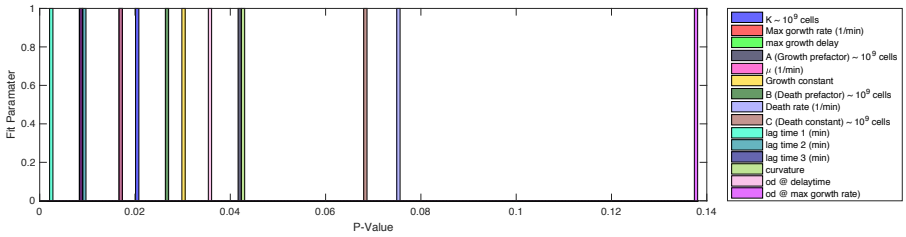

Figure 4: Average of the p-values determined via the $\boldsymbol{\chi^2}$ distribution, identifying accurate parameters. Low p-values suggest we can be confident ($\mathbf{1 - P_{value}}$ confidence) that if we replicated an experiment, we would obtain the same values.

5.1 Results

6 Discussion

Some clustering algorithms may perform well over data, however, the best results are usually obtained by combining results. The explanation is only certain relationships are found with certain cluster routines. This is significant because it demonstrates a flexible program outperforming well established individual methods. Furthermore, the combined approach is applicable not only to growth data, but is suitable for clustering a wide range data types. Most clustering methods specificity requires a fundamental understanding data and clustering methods to maximize clustering results. Operation of the program does not require detailed knowledge of clustering, allowing it to be used by a wider verity of people.

6.1 F-scores

$$\boldsymbol{F_\beta score} = \frac{\boldsymbol{(1+\beta^2)\cdot true\ positives}}{\boldsymbol{(1+\beta^2)\cdot true\ positives + \beta^2 \cdot false\ negatives + false\ positives}} \tag{3}$$

F-scores were determined to be the most reliable indicators of clustering quality when compared to other metrics. Algorithms usually only clustered a few genes. Combining cluster results increased resolution of gene relationships. New information was contained using additional clustering methods, even with lower F-scores. Cluster vectors are weighted by their F-score(s), giving priority to better quality results. Relationships identified using several cluster algorithm are weighted higher by adding their respective F-score(s) to the corresponding similarity matrix element. The F-score's weighting factor ($\boldsymbol{w1 or w2}$) along with optional cutoffs may be adjusted by users, for more or less weighting.

$$\bar{\boldsymbol{S}}(i,ii) = \frac{1}{\boldsymbol{max(S)}} \sum_{\boldsymbol{j}}^{\boldsymbol{algo}} \sum_{\boldsymbol{i}}^{\boldsymbol{genes}} \sum_{\boldsymbol{ii}}^{\boldsymbol{genes}} \boldsymbol{F_1^{w1}(i)} \times \boldsymbol{F_{similar}^{w2}(ii)} \times \boldsymbol{\delta_{v_i(i),v_i(ii)}} \tag{4}$$

Table 1: F-scores (averaged vs combined)

	F_1	F_{sim}
Average	.59	.55
Combined	.67	.63

6.2 Fitting

Continuous function models for fitting were additionally investigated, but these performed significantly worse using the F-Test. 5 models for fitting were examined that invoked 3 different fitting algorithms. Logistic, Gompertz, modified Gompertz, Richards, and piecewise exponentials were tested using linear fitting, Levenberg Marquardt and full Newton algorithms[8]. The modified Gompertz algorithm performed well with Levenberg Marquardt; however, this method struggled to accurately determine growth rates with early growth phase noise. Piecewise exponential fits performed best with full Newton fitting methods, however these were more CPU intensive, than the linear fitting procedure. Linear fitting was found to be relatively robust and significantly quicker and thus was chosen for GeneLink.

6.3 Performance

GeneLink is compiled in C++ via Matlab® Compiler 6.4 and is fully parallelizable. The strengths of the program are robust fit parameters, such as maximum growth rate, average growth rate, and the prefactor of the Death Phase. The program additionally is able to place data from separate growth conditions in separate clusters (figures not shown), demonstrating the ability to handle several data types together.

The limitations GeneLink included optical aberrations negatively impacting accuracy. Low frequency (¿10 mins.) of OD measurements also hindered accuracy. Systematic errors were encountered with a BioscreenC instrument. Early data points were relatively noisy. To address noise at early times, the minimum is used as the starting point for exponential fits. The Lag Time was difficult to determine due it occurring at early time points, at low optical densities and high relativley higher noise. Lag Time inconsistency was addressed

by applying three separate methods with the best determined by p-values of replicates. GeneLink may additionally benefit in speed from openGL and Cuda® support in future releases.

7 Acknowledgments

7.1 Funding

Funding for this project was provided by NIH grant #######

7.2 Collaborators

7.3 Conflicting Interest

This program was developed for analysis of data in Shields, 2017[**?**] and Marcic, 2017[**?**]

References

[1] J Monod. The growth of bacterial cultures. *Annual Review of Microbiology*, 3(1):371–394, 1949.

[2] Hans-Curt Flemming, Jost Wingender, Ulrich Szewzyk, Peter Steinberg, Scott A. Rice, and Staffan Kjelleberg. Biofilms: an emergent form of bacterial life. *Nat Rev Micro*, 14(9):563–575, 09 2016.

[3] Zhilin Qu, James N. Weiss, and W. Robb MacLellan. Coordination of cell growth and cell division: a mathematical modeling study. *Journal of Cell Science*, 117(18):4199–4207, 2004.

[4] MH Zwietering, Il Jongenburger, FM Rombouts, and K Van't Riet. Modeling of the bacterial growth curve. *Applied and environmental microbiology*, 56(6):1875–1881, 1990.

[5] József Baranyi and Terry A. Roberts. A dynamic approach to predicting bacterial growth in food. *International Journal of Food Microbiology*, 23(3):277 – 294, 1994. Special Issue Predictive Modelling.

[6] Lauren Mashburn-Warren, Donald A Morrison, and Michael J Federle. A novel double-tryptophan peptide pheromone controls competence in streptococcus spp. via an rgg regulator. *Molecular microbiology*, 78(3):589–606, 2010.

[7] Tim van Opijnen, Kip L Bodi, and Andrew Camilli. Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nat Meth*, 6(10):767–772, 10 2009.

[8] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University Press, New York, NY, USA, 3 edition, 2007.