A peer-reviewed version of this preprint was published in PeerJ on 19 April 2018.

Growth Score: A single metric to define growth in 96-well phenotype assays

Daniel A Cuevas [Corresp., 1] , **Robert A Edwards** [1]

[1] Computational Science Research Center, San Diego State University, San Diego, CA, USA

Corresponding Author: Daniel A Cuevas
Email address: dcuevas@sdsu.edu

High-throughput phenotype assays are a cornerstone of systems biology as they allow direct measurements of mutations, genes, strains, or even different genera. High-throughput methods also require data analytic methods that reduce complex time-series data to a single numeric evaluation. Here, we present the Growth Score, an improvement on the previous Growth Level formula. There is strong correlation between Growth Score and Growth Level, but the new Growth Score contains only essential growth curve properties while the formula of the previous Growth Level was convoluted and not easily interpretable. Several programs can be used to estimate the parameters required to calculate the Growth Score metric, including our *PMAnalyzer* pipeline.

1 ***Growth Score*: a single metric to define growth in 96-well**

2 **phenotype assays**

3 Daniel A. Cuevas[1], Robert A. Edwards[1]

4 [1] Computational Science Research Center, San Diego State University, San Diego, CA, USA

5

6 Corresponding Author:

7 Daniel Cuevas

8 5500 Campanile Drive, San Diego, CA, 92182, USA

9 Email address: dcuevas08@gmail.com

Abstract

10

11   High-throughput phenotype assays are a cornerstone of systems biology as they allow direct

12   measurements of mutations, genes, strains, or even different genera. High-throughput methods

13   also require data analytic methods that reduce complex time-series data to a single numeric

14   evaluation. Here, we present the Growth Score, an improvement on the previous Growth Level

15   formula. There is strong correlation between Growth Score and Growth Level, but the new

16   Growth Score contains only essential growth curve properties while the formula of the previous

17   Growth Level was convoluted and not easily interpretable. Several programs can be used to

18   estimate the parameters required to calculate the Growth Score metric, including our

19   *PMAnalyzer* pipeline.

Introduction

20

21   Bacterial growth of homogenous cultures is commonly described through the Monod growth

22   phases (Monod, 1949). Specifically, the three major phases of growth are lag phase, where

23   growth rate is zero and bacterial density is constant at the initial measurement; exponential

24   phase, where growth rate is at its maximum value; and stationary phase, where growth rate is

25   zero and bacterial density is constant at its maximum yield. The Growth Level (GL) has been

26   used to quantitatively measure the amount of growth displayed by a bacteria liquid culture

27   (Cuevas & Edwards, 2017). In *PMAnalyzer* it is calculated following the least-squares fitting of

28   the Zwietering logistic model (Zwietering et al., 1990):

$$\hat{y} = y_0 + \frac{A - y_0}{1 + \exp\left[\frac{4\mu}{A}(\lambda - t) + 2\right]}.$$

30   Using the modeled growth curve ($\hat{y}$), the starting absorbance (y_0), and the biomass yield value

31   (A), GL can be calculated as

32
$$GL = \frac{n}{\sum\limits_{i}^{n} \frac{1}{x_i}},$$

33   where $x_i = (\hat{y}_i - y_0) + amplitude = (\hat{y}_i - y_0) + (A - y_0).$

34   *GL* is a variation on the harmonic mean where the logistic growth curve is weighted by the

35   amount of biomass the bacteria culture attained, or the difference in bacterial density, during the

36   course of the experiment (represented as *amplitude* in the *GL* formula). This amplitude-weighted

37   metric performs well for differentiating growing data from growth curves that display no growth.

38   The *GL* provides threshold values that can be used to ascribe qualitative labels or classes to

39   growth, ranging from **no growth** to **very high growth**.

40   There are several mathematical drawbacks of the *GL* formula. The harmonic mean is

41   commonly used to average values of rates; however, absorbance data are not rate measurements.

42   In addition, using the entire growth curve in *GL* is implicitly affected by the lag time of the

43   bacteria that is dependent on numerous biological properties including ageing of cells, activation

44   of enzymes, metabolic adaptation, and other regulatory mechanisms (Monod, 1949; Robinson et

45   al., 1998; Rolfe et al., 2012). The simplicity of the three phase growth curve is lost in the *GL*

46   calculation.

47   Here, we propose a new calculation to simplify the quantitative meaning of the level of

48   growth. The Growth Score, *GS*, is defined by three parameters of the Zwietering (Zwietering et

49   al., 1990) bacterial growth curve

50                              $$GS = (A - y_0) + 0.25\mu.$$

51   *GS* uses the starting absorbance, y_0, biomass yield, A, and the maximum growth rate, μ, to

52   compute a score for a growth curve. Without any explicit dependency on the fitted values or

53   implicit dependencies on lag time, *GS* is clearly understood by its growth parameters. In

54   addition, *GS* performs similarly to *GL*, primarily because of *GL*s strong dependency on yield in

55   defining a quantitative measurement of growth.

56 Simulating growth curves

57   A set of 1,000 growth curves were generated using a total time $t = 50$ hours and the Zwietering

58   logistic model. Uniformly-distributed random values were selected for each growth curve

59   parameter. Distribution ranges for each parameter are: starting absorbance, $y_0 = [0.05, 0.10]$; lag

60   time, $\lambda = [0, 20]$; biomass yield, $A = [0.1, 1.2]$; and maximum growth rate, $\mu = [A/t, 1.1A]$.

61   Growth rate range is limited by the biomass yield to provide a realistic result. The first value

62   represents an organism achieving the biomass yield at the end of the experiment; when *A* is large

63   this simulates very slow growth. The second value represents a rate where yield is obtained

64   within one hour; when *A* is large this simulates very fast growth. For cases when *A* is small,

65   growth rate is negligible, therefore has minimal influence on *GS*.

66 Comparison of Growth Level and Growth Score

67   The distributions of *GL* and *GS* are illustrated in the Fig. 1 histograms. The range for *GL* is much

68   larger than *GS*. Both metrics demonstrate a slight left-skewness but *GL* was slightly higher at

69   0.143 compared to *GS* at 0.078. The growth score should be left skewed because a low

70   asymptote correlates with low growth rates and results in a low growth score, but the reverse is

71   not necessarily true, high growth score values may result from high or low growth rates.

72   Differences in metrics are displayed in Fig. 1 with *GL* shown along the x-axis and the

73   corresponding *GS* value along the y-axis, along with a linear regression line, and growth classes.

74   A strong correlation is shown between *GL* and *GS* with a Pearson correlation coefficient of 0.98

75   (p-value < 0.001) (Fig. 1). The qualitative classes are defined in *PMAnalyzer* as value thresholds.

76   Given these *GL* classes, *GS* class thresholds are also indicated in Fig. 1 as horizontal dotted lines.

77   A total of 191 out of the 1,000 growth curves were classified differently between the two

78   metrics, however, only 20 growth curves had a change from "-" to "+" or vice versa. To further

79   demonstrate the separation of the growth data, average growth curves are plotted according to

80   their classification in Fig. 2. Growth curves were generated using the Python 3.6 programming

81   language (Python Software Foundation. Python Language Reference, version 3.6. Available at

82   https://www.python.org), NumPy version 1.13.1 (http://www.numpy.org/) (Walt, Colbert &

83   Varoquaux, 2011), pandas version 0.20.3 (http://pandas.pydata.org) (McKinney, 2010), and

84   Seaborn version 0.8.1 (http://seaborn.pydata.org/). See Supplementary Material 1 for the Python

85   code that generates the random curves, Supplementary Material 2 for the Jupyter Notebook

86   version, and Supplementary Material 3 for a PDF version of the Jupyter Notebook.

87   Biomass yield has been successful in discriminating growth from no growth using the GL

88   formula. However, that calculation lacked direct incorporation of the maximum growth rate, a

89   useful component in measuring fitness. Intuitively, faster growth rates should indicate a bacteria

90   that is more capable of consuming nutrients and proliferating. Yet, a higher rate of growth does

91   not solely demonstrate fitness, evidenced by the abundance of slow-growing bacteria throughout

92   the environment (e.g., r and K growth strategies (Pianka, 1970)). Nor is it a strong feature to

93   ultimately predict larger bacterial densities. For example, two bacteria with the same phenotype

94   can reach the same density, even if one grows at half the rate of the other. Essentially, the ability

95   to reach that potential at a faster rate indicates some biological advantage but has less

96   significance than yield. In GS maximum growth rate is included at an amount of 25% in order for

97   biomass accumulation to result in the primary component of growth. Faster growth rates can

98   occur within a short time frame, resulting in low yield and high rate and, therefore, causing a

99   disproportionate GS if the magnitude of the growth rate was not reduced.

100   Here, we have introduced the Growth Score, a new parameter for describing bacterial

101   growth in 96-well phenotypic assays. The Growth Score provides three distinct advantages over

102   other metrics used to describe growth: First, it only uses growth curve properties (yield and

103   growth rate) in its calculation, in contrast our previous Growth Level was averaged over time and

104   was thus heavily influenced by the length of the experiment. Second, Growth Score can be used

105   with results from other software that also performs growth curve modeling or parameterization

106   (DuctApe (Galardini et al., 2014), GCAT (Bukhman et al., 2015), grofit (Kahm et al., 2010;

107   Vaas et al., 2012), OmniLog Biolog Phenotype MicroArrays (Borglin et al., 2012; Vaas et al.,

108   2012) without the need for the raw data, whereas growth level would at least need lag time and

109   /or the raw spectrophotometry data. Finally, the time independence of Growth Score is also a

110   benefit over measurements like Area Under the Curve (AUC) employed by some software. AUC

111   is subject to similar biases as growth level—longer experiments directly affect how lag time and

112   growth rate mathematically influence the AUC calculation.

113

Figure Legends

FIGURE 1. Growth Level and Growth Score correlation. The linear relationship results in a Pearson correlation coefficient of 0.98. Distributions are plotted as marginal histograms for each metric. Point colors represent the *GL* classes whereas dotted lines represent the proposed *GS* lower thresholds for each colored class.

FIGURE 2. Growth curves distributions. Average growth curves colored by *GL* and *GS* classes. One hundred cycles of bootstrapping were performed for each class, indicated by the transparent traces surrounding the average line. A) *GL* classes. B) *GS* classes.
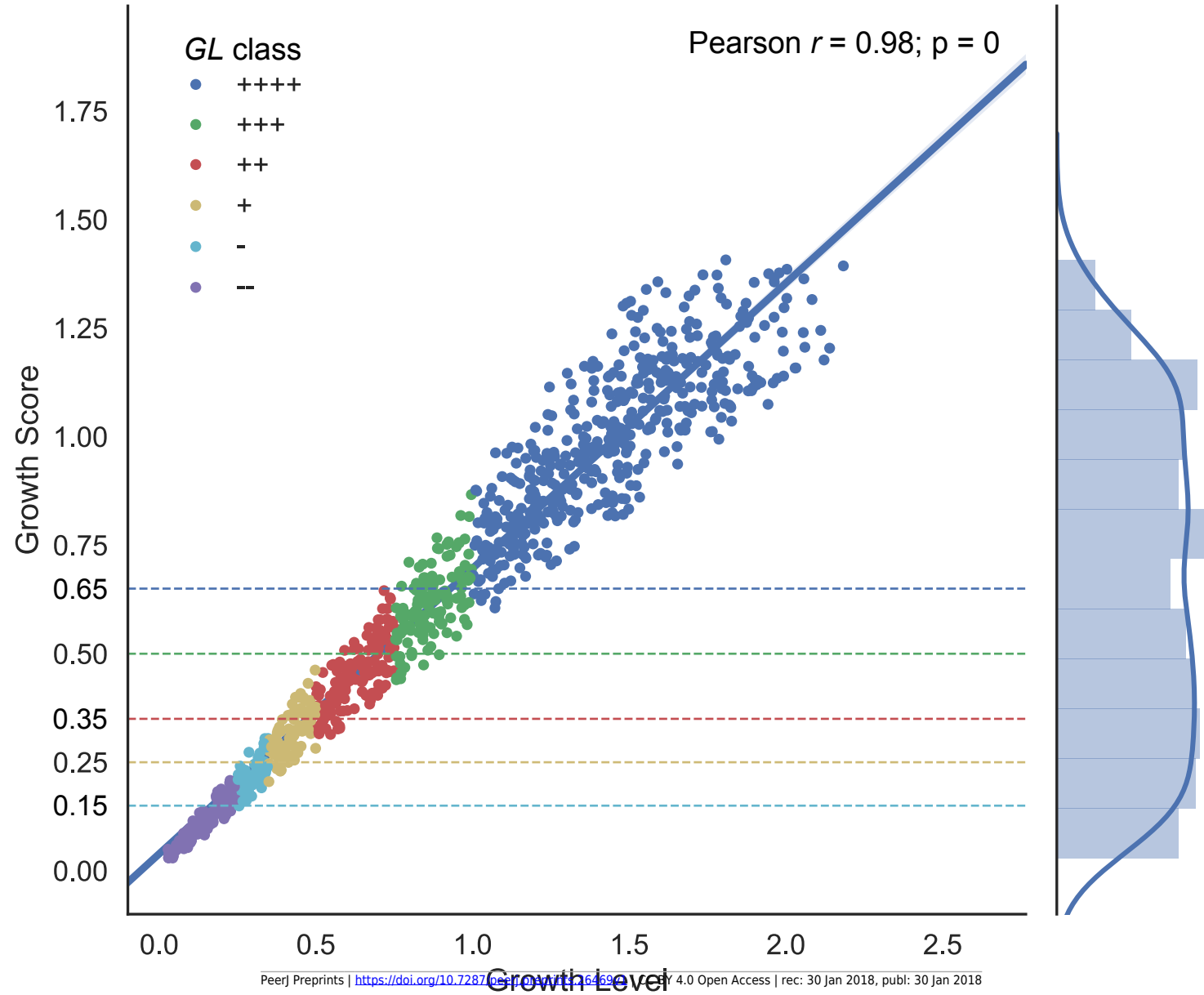
References

Borglin S., Joyner D., DeAngelis KM., Khudyakov J., D'haeseleer P., Joachimiak MP., Hazen T. 2012. Application of phenotypic microarrays to environmental microbiology. *Current opinion in biotechnology* 23:41–48. DOI: 10.1016/j.copbio.2011.12.006.

Bukhman YV., DiPiazza NW., Piotrowski J., Shao J., Halstead AGW., Bui MD., Xie E., Sato TK. 2015. Modeling Microbial Growth Curves with GCAT. *BioEnergy Research* 8:1022–1030. DOI: 10.1007/s12155-015-9584-3.

Cuevas DA., Edwards RA. 2017. PMAnalyzer: a new web interface for bacterial growth curve analysis. *Bioinformatics* 33:1905–1906. DOI: 10.1093/bioinformatics/btx084.

Galardini M., Mengoni A., Biondi EG., Semeraro R., Florio A., Bazzicalupo M., Benedetti A., Mocali S. 2014. DuctApe: A suite for the analysis and correlation of genomic and OmniLog™ Phenotype Microarray data. *Genomics* 103:1–10. DOI: 10.1016/j.ygeno.2013.11.005.

Kahm M., Hasenbrink G., Lichtenberg-Fraté H., Ludwig J., Kschischo M. 2010. grofit: Fitting Biological Growth Curves with R. *Journal of Statistical Software* 33. DOI: 10.18637/jss.v033.i07.

McKinney W. 2010. Data Structures for Statistical Computing in Python. In: *Proceedings of the 9th Python in Science Conference*. 51–56.

Monod J. 1949. The Growth of Bacterial Cultures. *Annual Review of Microbiology* 3:371–394. DOI: 10.1146/annurev.mi.03.100149.002103.

Pianka ER. 1970. On r- and K-Selection. *The American Naturalist* 104:592–597. DOI: 10.1086/282697.

Robinson TP., Ocio MJ., Kaloti A., Mackey BM. 1998. The effect of the growth environment on the lag phase of Listeria monocytogenes. *International Journal of Food Microbiology* 44:83–92. DOI: 10.1016/S0168-1605(98)00120-2.

Rolfe MD., Rice CJ., Lucchini S., Pin C., Thompson A., Cameron ADS., Alston M., Stringer MF., Betts RP., Baranyi J., Peck MW., Hinton JCD. 2012. Lag Phase Is a Distinct Growth Phase That Prepares Bacteria for Exponential Growth and Involves Transient Metal Accumulation. *Journal of Bacteriology* 194:686–701. DOI: 10.1128/JB.06112-11.

Vaas LAI., Sikorski J., Michael V., Göker M., Klenk H-P. 2012. Visualization and Curve-Parameter Estimation Strategies for Efficient Exploration of Phenotype Microarray Kinetics. *PLoS ONE* 7:e34846. DOI: 10.1371/journal.pone.0034846.

Walt S van der., Colbert SC., Varoquaux G. 2011. The NumPy Array: A Structure for Efficient Numerical Computation. *Comput. Sci. Eng.* 13:22–30.

Zwietering MH., Jongenburger I., Rombouts FM., van 't Riet K. 1990. Modeling of the Bacterial Growth Curve. *Applied and Environmental Microbiology* 56:1875–1881.

Figure 1(on next page)

Growth Level and Growth Score correlation.

The linear relationship results in a Pearson correlation coefficient of 0.98. Distributions are plotted as marginal histograms for each metric. Point colors represent the *GL* classes whereas dotted lines represent the proposed *GS* lower thresholds for each colored class.

Figure 2(on next page)

Growth curves distributions.

Average growth curves colored by *GL* and *GS* classes. One hundred cycles of bootstrapping were performed for each class, indicated by the transparent traces surrounding the average line. A) *GL* classes. B) *GS* classes.

A)

B)