Bioinformatics, 33(9), 2017, 1370–1378
doi: 10.1093/bioinformatics/btw673
Advance Access Publication Date: 18 January 2017
Original Paper

Data and text mining

PhenoCurve: capturing dynamic phenotype-environment relationships using phenomics data

Yifan Yang,[1,†] Lei Xu,[2,3,†] Zheyun Feng,[2] Jeffrey A. Cruz,[3] Linda J. Savage,[3] David M. Kramer[3,4,*] and Jin Chen[5,*]

[1]Department of Epidemiology and Biostatistics, [2]Department of Computer Science and Engineering, [3]Department of Energy Plant Research Laboratory and [4]Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI 48824, USA and [5]Institute of Biomedical Informatics, University of Kentucky, Lexington, KY 40536, USA

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Jonathan Wren

Received on April 1, 2016; revised on September 12, 2016; editorial decision on September 27, 2016; accepted on October 27, 2016

Abstract

Motivation: Phenomics is essential for understanding the mechanisms that regulate or influence growth, fitness, and development. Techniques have been developed to conduct high-throughput large-scale phenotyping on animals, plants and humans, aiming to bridge the gap between genomics, gene functions and traits. Although new developments in phenotyping techniques are exciting, we are limited by the tools to analyze fully the massive phenotype data, especially the dynamic relationships between phenotypes and environments.
Results: We present a new algorithm called PhenoCurve, a knowledge-based curve fitting algorithm, aiming to identify the complex relationships between phenotypes and environments, thus studying both values and trends of phenomics data. The results on both real and simulated data showed that PhenoCurve has the best performance among all the six tested methods. Its application to photosynthesis hysteresis pattern identification reveals new functions of core genes that control photosynthetic efficiency in response to varying environmental conditions, which are critical for understanding plant energy storage and improving crop productivity.
Availability and Implementation: Software is available at phenomics.uky.edu/PhenoCurve
Contact: chen.jin@uky.edu or kramerd8@cns.msu.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Plants have a truly remarkable ability to harness energy from the sun to produce food. That is called photosynthesis. Understanding how plants optimize or regulate photosynthesis in response to a continuously changing environment is essential towards developing strategies to improve crop yields (Kramer and Evans, 2011). Recent development of high-throughput photosynthetic phenotyping platforms allows continuous measurements of various photosynthetic parameters over developmental time scales (days to weeks) and under dynamically changing conditions (e.g. light intensity and temperature), providing rich data sets for phenotype characterization (see details in Fig. 1) (Baker *et al.*, 2007; Cruz *et al.*, 2016; Houle *et al.*, 2010; Rascher *et al.*, 2011; Subramanian *et al.*, 2013). From the large volume of phenotype data biologists expect to identify genes ancillary to optimizing photosynthetic performance,

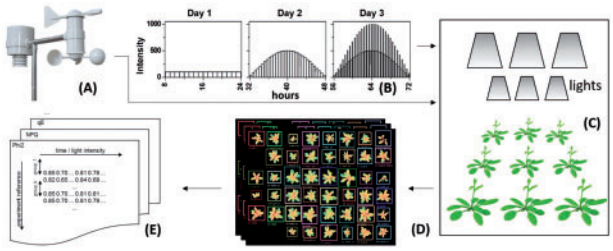

Fig. 1. Workflow of high-throughput plant photosynthesis phenotyping. Environmental factors measured in the real world **(A)** or synthesized **(B)** are played back in specially-designed chambers **(C)** outfitted with arrays of sophisticated lighting and imaging sensors, to produce false-color fluorescence images of plants **(D)**, which are then converted into matrices of photosynthesis parameters **(E)**. In a matrix, each row *i* is a plant, each column *j* is a time point that is associated with a specific environmental condition, and each value is a photosynthesis phenotype of plant *i* at time point *j*

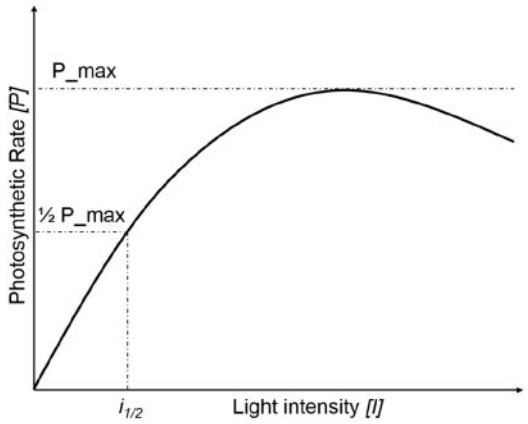

Fig. 2. Hyperbolic PI curve

growth and yield under dynamic growth conditions, especially those related to abiotic or biotic stress, and to design more focused experiments to fully define their functions.

Machine learning and computer vision algorithms have been recently developed for phenotype information retrieval, data quality control, and knowledge discovery (Gao *et al.*, 2016; Green *et al.*, 2012; Tessmer *et al.*, 2013; Xu *et al.*, 2015; Yin *et al.*, 2014a,b). For example, computational tools are available to identify temporal patterns from phenomics data (Yang and Leskovec, 2011), to group traits by phenotypes existing (Gao *et al.*, 2016), and to predict unknown gene functions (Vlasblom *et al.*, 2015). Ideally, these tools would be useful for predicting which genotypes will perform the best in specific *real world* growth environments.

Nevertheless, new tools are required to precisely model the influence of environment (e.g. abrupt changes in light intensity or temperature) on phenotype (e.g. excessive losses in photosynthetic efficiency and/or increases in photodamage). As has been emphasized in the literature, it is crucial to *simultaneously* measure and correlate both phenotypes and environmental factors to arrive at a holistic characterization of plant performance (Großkinsky *et al.*, 2015; Kutsukake *et al.*, 2012; Walter *et al.*, 2015; Wong *et al.*, 2012). In plants, photosynthesis must respond to changing environment to provide the optimal amount of energy to meet the needs of the organism, in the correct forms, without producing toxic byproducts. In this context, photosynthesis can be viewed as a set of integrated modules that form a self-regulating network that is sensitive to changes in both environmental parameters (e.g. light intensity) and metabolic or physiological factors (Kramer and Evans, 2011).

Studying the complex relationships between phenotypes and environments that define the interacting photosynthetic modules poses several computational challenges. First, phenotype–environment relationships are usually learned using data-driven approaches such as linear regression or curve fitting. However, (i) it is difficult to choose the best fitting function (Serôdio and Lavaud, 2011); (ii) it is difficult to incorporate biological knowledge needed to adequately describe the complex responses involved; (iii) purely data-driven approaches may be significantly affected by bias and noise in phenomics data (Osborne and Overbay, 2004).

Second, if a phenotype–environment relationship has already been well studied, researchers tend to apply the known biological model *directly* (Eilers and Peeters, 1988). However, biological models are usually simple and static, while the phenotype–environment relationships in the real world are often complex and dynamic. For example, when light intensity is low, photosynthesis activity is positively correlated with light, but such relationships can be reversed if the light intensity is so large that it causes photoinhibition (see Fig. 2). It is thus inappropriate to directly apply a theoretical model to the real data that vary constantly over time and condition (Xu *et al.*, 2015).

Third, the photosynthesis phenotypes are often measured under dynamic environmental conditions, over a relatively long period, and on many plants with vastly different genetic backgrounds. This broad range of data variation adds another level of complexity to the problem. In summary, novel algorithms are required to explore complex phenotype–environment relationships that enable researchers to model phenotypes, environments and genetic diversity simultaneously.

In this article, we present *PhenoCurve* to explore dynamic phenotype–environment relationships with three major advantages over existing approaches:

i. Although phenotype and environment are measured separately with different techniques, they are biologically correlated. Studying the complex relationships between them may reveal patterns that cannot be discovered by only using phenotype data.
ii. PhenoCurve divides the whole dataset into reliable and unreliable parts, and then optimizes the phenotype–environment relationships on the unreliable part.
iii. In contrast to purely data-driven approaches, PhenoCurve can effectively incorporate biological knowledge thus significantly improving its performance.

In the following content, we demonstrate the effectiveness of PhenoCurve by identifying the dynamic relationships between a key photosynthesis phenotype Φ_{II} (steady state quantum yield of photosystem II) and light intensity. PhenoCurve can be easily extended for other phenotype data with a simple modification.

2 Related work

In this section we introduce the biological background for modeling the relationships between light and photosynthesis using the photosynthesis-irradiance (PI) curve, as well as the existing computational approaches on curve fitting and regression.

2.1 PI curve

The PI curve is a graphical representation of the empirical relationship between light and photosynthesis (MacIntyre *et al.*, 2002). As a derivation of the Michaelis-Menten kinetics, one of the best-known models of enzyme kinetics, PI is modeled as a hyperbolic curve as shown in Figure 2 (Chou and TaLaLay, 1981; Dowd and Riggs,

1965; Menten and Michaelis, 1913), indicating that there is a generally positive correlation between light intensity and photosynthetic rate. The PI curve has been applied successfully to photosynthesis phenotype data under limited sets of conditions, e.g. to explain ocean-dwelling phytoplankton photosynthetic response to changes in light intensity (Jassby and Platt, 1976).

Let P be the photosynthetic rate at a given light intensity $[I]$ (or denoted as i), the PI function is given by Equation (1):

$$P = \frac{[I] \times P_{\max}}{i_{1/2} + [I]} \tag{1}$$

where $P_{\max}$ is the maximum potential photosynthetic rate per individual, and $i_{1/2}$ is half-saturation parameter representing the amount of light to produce half of maximum photosynthesis. Note that Equation (1) only models the positive correlation part of the PI curve.

By describing the photosynthetic rate P using linear electron flow, i.e. $P = \Phi_{II} \times i$, we can rewrite Φ_{II} with respect to time t:

$$\Phi_{II}(t, i_{1/2}) = \frac{\max(\Phi_{II})}{1 + \frac{i(t)}{i_{1/2}}} \tag{2}$$

where t is a time point, $\Phi_{II}(t, i_{1/2})$ and $i(t)$ represent the steady state quantum yield of photosystems II and light intensity at t, respectively, and $\max(\Phi_{II})$ is the maximal Φ_{II} within the whole day. In general, Φ_{II} decreases as i increases, reflecting a combination metabolic congestion as the electron transfer reactions exceeds the capacity of downstream biochemistry, and feedback down regulation. See details in Xu *et al.* (2015).

Over a brief time periods, $i_{1/2}$ remains constant. However, it may change gradually and smoothly over a long time period with the gradual changes of multiple environmental factors such as light intensity, temperature, concentrations of CO_2 and O_2 or acclimation process such as gene expression and developmental factors, aging or the accumulation of damage (Cruz *et al.*, 2016). Such phenotypic plasticity indicates an organism can express a range of phenotypes when exposed to different environments (Nicotra *et al.*, 2010; Price *et al.*, 2003).

2.2 Sliding window-based curve fitting methods

Under gradually varying environments, $i_{1/2}$ is assumed to change continuously. A sliding window approach may be appropriate to apply. Using a sliding window approach, we can divide the whole phenotype data $(t, i(t), \Phi_{II}(t))$ into overlapped temporal windows along t, and then employ a curve fitting or regression method to compute a local $i_{1/2}$ value for each window. $i_{1/2}$ at time t is determined by the data in the window centered around t. Finally, all the local $i_{1/2}$ values are merged to capture the global phenotype–environment relationship. Note that there is no explicit boundary between curve fitting and regression; while the former pertains to the fitting optimization itself, the latter focuses more on statistical inference (Motulsky and Christopoulos, 2004).

Curve fitting is a commonly used method to model the relationships among two or more variables (Motulsky and Christopoulos, 2004). Mathematically, it is a process to tune the parameters of a known mathematical function f to achieve the best fit to a series of data points, where in our case f is a function to describe the underlying biological relationship between phenotype and environment. The Levenberg-Marquardt algorithm (LMA), aka the damped least-squares method, has been widely used for nonlinear least squares calculations for solving generic curve-fitting problems (Levenberg, 1944). LMA interpolates between the Gauss-Newton algorithm and the method of gradient descent, aiming to find a local minimum (Bates and Watts, 1988; Holland and Welsch, 1977). If the fitting function f is unknown, some nonparametric smoothing techniques such as robust regression like robust locally weighted regression (LOWESS) (Cleveland, 1979) and kernel smoother models like locally linear regression (LLR) (Gupta *et al.*, 2008), are often used for estimating a smooth curve from observations. In this way, nonlinear relationships between phenotypes and environmental factors can be learned purely from data.

2.3 Bayesian linear model with normal inverse gamma prior

Although sliding window-based curve fitting methods optimize local fitting, they simply ignore the global continuity of $i_{1/2}$. Thus, they may be sensitive to noise in raw phenotype data in local windows, resulting in inaccurate phenotype–environment relationships. Performance improvement may be achieved by using the Bayesian linear model with normal inverse gamma (NIG) prior (Gelman, 2006).

Given phenomics data in window $W_j = \{(t_k, \Phi_{II,k}, i_k) : \forall k \in [j, j+n]\}$, where $n+1$ is the window width, we first transfer Equation (2) to its linear form for easier representation:

$$\frac{\max(\Phi_{II})}{\Phi_{II,k}} - 1 = \frac{1}{i_{1/2}} i_k + \varepsilon_k \tag{3}$$

where $\Phi_{II,k}$, i_k are Φ_{II} and light intensity at t_k, and ϵ_k is the error term associated with t_k distributed as normal distribution $N(0, \sigma^2)$.

Second, similar to Section 2.2, we adopt a sliding window approach to estimate $i_{1/2}$ and σ^2 for each temporal window using linear regression methods (Freedman, 2009). Given a threshold of the linear regression reliability R^2 (Holland and Welsch, 1977), we can classify all windows $D = \{W_j : 1 \le j \le N\}$ into two groups, i.e. reliable data $D^{(r)}$ and unreliable data $D^{(u)}$, where $^{(r)}$ and $^{(u)}$ stand for 'reliable' and 'unreliable'; N is the number of windows; $D^{(r)} \cup D^{(u)} = D$; and $D^{(r)} \cap D^{(u)} = \varnothing$.

Third, in order to derive $\hat{i}_{1/2}$ in each unreliable window, we assume that the prior $(i_{1/2}, \sigma^2)^\top$ follows the NIG distribution, i.e. $i_{1/2}, \sigma^2 \sim \text{NIG}(\mu, V, a, b)$, where μ, V, a, b is the set of hyper parameters of NIG. Subsequently $i_{1/2}, \sigma^2 | D^{(u)}$ is also distributed as NIG yet with different parameters (μ^*, V^*, a^*, b^*). Assuming that the priors of parameters for both $D^{(r)}$ and $D^{(u)}$ follow the same single mode NIG distribution with (μ, V, a, b), we estimate the hyper parameters $(\hat{\mu}, \hat{V}, \hat{a}, \hat{b})$ using $D^{(r)}$ and apply them on $D^{(u)}$. Finally, $i_{1/2}$ of each temporal unreliable window can be obtained by using the following steps, 1) obtaining $(\hat{\mu}^*, \hat{V}^*, \hat{a}^*, \hat{b}^*)$ by both the shared hyper parameters $(\hat{\mu}, \hat{V}, \hat{a}, \hat{b})$ and local data $D^{(u)}$ in each window, 2) obtaining the marginal posterior distribution $\Pr(i_{1/2} | D^{(u)})$ by integrating out σ^2, 3) maximizing the marginal posterior distribution to obtain the estimates $\hat{i}_{1/2}$. See more details at Gelman (2006) and Bolstad (2013).

In this model, the global continuity of $i_{1/2}$ is guaranteed due to the single mode of the NIG distribution (Bolstad, 2013). However, assumptions about constant hyperparameters could be too rigid, because the parameters for $D^{(r)}$ and $D^{(u)}$ may be under different prior distributions due to different environmental stresses.

3 Materials and Methods

To explore the dynamic phenotype–environment relationships without the technical limitations in curve fitting and Bayesian NIG

methods, we present *PhenoCurve* based on regularized polynomial regression.

PhenoCurve has four components, as showed in Figure 3. First, it splits the raw phenotype and environment data into highly overlapped temporal windows using a sliding window approach, obtaining D and allowing for modeling the gradual change of $i_{1/2}$. Second, it employs a non-linear curve fitting method to compute $i_{1/2}$ for each temporal window, and classifies D into two groups by the results, i.e. reliable one $D^{(r)}$ and unreliable one $D^{(u)}$, based on each R^2. Third, using $D^{(r)}$, it estimates $i_{1/2}$ for each unreliable window with regularized polynomial regression. Finally, it optimizes the $i_{1/2}$ values for all unreliable windows using local data in that window, resulting in increased reliability in curve fitting.

We introduce all the four steps in the following content. In addition, an illustrative example in Figure 4 is used to demonstrate the key steps in PhenoCurve. Some of the sampled Φ_{II} are far from the ground truth because of unavoidable noise during measurement (see the corresponding light intensities in Supplementary Fig. S1). Using this simple case, we demonstrate how PhenoCurve optimizes both the local fitting of Φ_{II} and the global trend of $i_{1/2}$ thus recovering the real Φ_{II} values under dynamic environmental conditions. Note the ground truth is hidden from the program and is only used for performance evaluation.

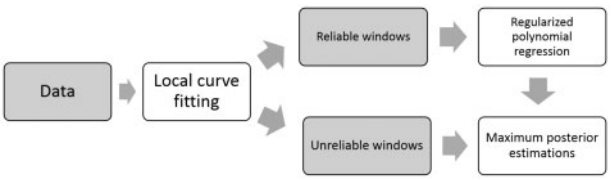

Fig. 3. The workflow of PhenoCurve. The gray boxes are data types, and the white boxes are processes

3.1 Data separation with sliding window

Due to biological constraints, the sampling rate of the phenotype is usually much lower than that of the environmental factors (Cruz *et al.*, 2016). Subsequently, we split the raw data into highly overlapped temporal windows solely based on the phenotype data, resulting in window set $D = \{W_j : 1 \leq j \leq N\}$, where each temporal window $W_j = \{(t_k, \Phi_{II,k}, i_k) : \forall k \in [j, j+n]\}$ has n pairs of phenotype and environment values, and the values in each pair are measured at exactly the same time (or are close enough to each other). Each temporal window shares $n - 1$ values with the previous and the next window. The window width satisfies two conditions: (i) $i_{1/2}$ remains relatively constant within each temporal window, and (ii) there are enough data in each temporal window for inferring the value of $i_{1/2}$.

3.2 Local curve fitting

The next step is to infer the half light parameter $i_{1/2,j}$ in each window W_j. The value of $i_{1/2,j}$ can be estimated using the least square curve fitting (Freedman, 2009). The general idea is to identify parameters in a fitting function to minimize the sum of all the

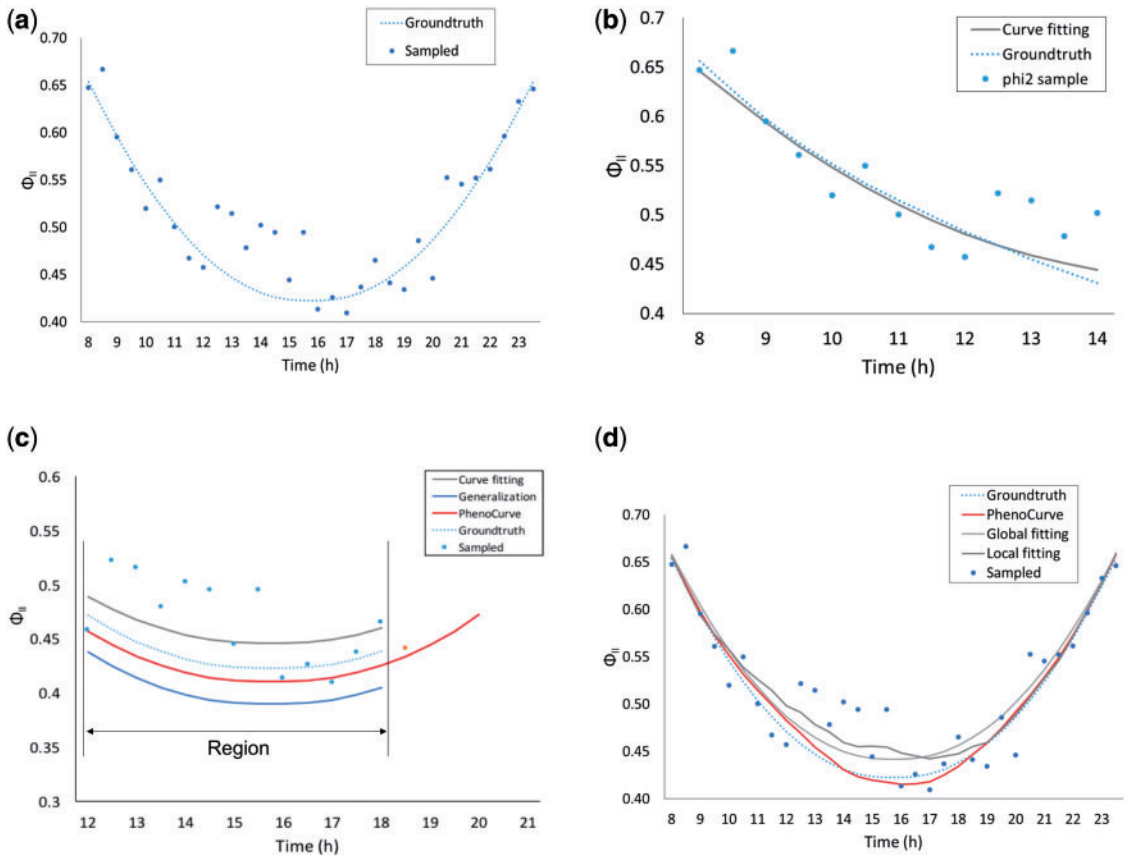

Fig. 4. An illustrative example of PhenoCurve. It demonstrates that PhenoCurve optimizes both the local fitting and the global trend of $i_{1/2}$ thus recovering the real Φ_{II} values under dynamic environmental conditions

square of errors between the predicted and the observed values. Given all the n phenotype–environment pairs in W_j, we estimate $i_{1/2,j}$ by:

$$\hat{i}_{1/2,j} = \underset{i_{1/2}\in \mathrm{R}}{\operatorname{argmin}} \sum_{k=j}^{j+n} \left(\Phi_{II,k} - \frac{\max(\Phi_{II})}{1+\frac{i_k}{i_{1/2}}} \right)^2 \tag{4}$$

The fitting procedure also yields a R^2 score, indicating the level of reliability of the fitting. Based on R^2, we can divide all windows $D = \{W_j : 1 \le j \le N\}$ into two groups, i.e. the reliable windows $D^{(r)} = \{W_j^{(r)} : 1 \le j \le N^{(r)}\}$ and the unreliable windows $D^{(u)} = \{W_j^{(u)} : 1 \le j \le N^{(u)}\}$, where $D^{(r)} \cup D^{(u)} = D$, $D^{(r)} \cap D^{(u)} = \varnothing$, and $N^{(r)} + N^{(u)} = N$.

Figure 4B shows the curve fitting results on one of the reliable temporal windows in our running example, where the fitted curve (grey solid line) matches well with the ground truth (dotted line). However, the fitted curve (black solid line) on one of the unreliable temporal windows, as shown in Figure 4C, is far from the ground truth (dotted line).

3.3 Regularized polynomial linear regression model

Using the half light parameter in the reliable temporal windows, denoted as $\hat{i}_{1/2,j}^{(r)}$, $1 \le j \le N^{(r)}$, we build a regularized polynomial linear regression model, aiming to generate a l-degree polynomial smooth curve of $i_{1/2}$, where l is the given degree of polynomial (Thrampoulidis *et al.*, 2015). In a regularized polynomial linear regression model, the degree of one represents a linear model, the order of two represents a quadratic form, and the order of three and above works for arbitrary shapes. Note using higher-degree polynomial would risk in over-fitting (Bishop, 2006).

Given all reliable windows $W_j^{(r)} = \{(t_k^{(r)}, i_k^{(r)}, \Phi_{II,k}^{(r)}) : \forall k \in [j, j+n]\}, 1 \le j \le N^{(r)}$, we adopt the l-degree polynomial linear regression model to generate the smooth curve of $i_{1/2}$. The procedure is described as follows.

Let $X_j^{(r)} = (1, t_j^{(r)}, i_j^{(r)}, \Phi_{II,j}^{(r)}, t_j^{(r)2}, i_j^{(r)2}, \Phi_{II,j}^{(r)2}, \ldots, t_k^{(r)l}, i_j^{(r)l}, \Phi_{II,j}^{(r)l})$ denote the j-th row of data matrix $\mathbf{X}^{(r)} = (X_1^{(r)}, \ldots, X_{N^{(r)}}^{(r)})^\top$, and let $\mathbf{y}^{(r)}$ denote the vector of all $\hat{i}_{1/2,j}^{(r)}$ in the reliable windows, $1 \le j \le N^{(r)}$, the regression model is given by:

$$\mathbf{y}^{(r)} = \mathbf{X}^{(r)}\beta + \eta \tag{5}$$

where η is distributed as a standard multivariate normal distribution.

To solve Equation 5 while avoiding overfitting, a sparsity model is required (Tibshirani, 1996). To this end, Lasso or its generalization form called elastic net is often considered (Zou and Hastie, 2005). Since elastic net is more flexible than Lasso, we consider elastic net regularization for the above polynomial linear regression model as following:

$$\hat{\beta} = \underset{\beta\in \mathrm{R}^{3l+1}}{\operatorname{argmin}} ||\mathbf{y}^{(r)} - \mathbf{X}^{(r)}\beta||^2 + \lambda_1||\beta||_1 + \lambda_2||\beta||^2 \tag{6}$$

where both λ_1 and λ_2 are tuning parameter, $||.||_1$ and $||.||$ are L_1 and L_2 norms of vector, respectively. By Lagrangian duality, the model with elastic net penalty is suggested to be given by:

$$\hat{\beta} = \underset{\beta\in \mathrm{R}^{3l+1}}{\operatorname{argmin}} ||\mathbf{y}^{(r)} - \mathbf{X}^{(r)}\beta||^2 \quad \text{subject to} \quad J(\beta) \le s \tag{7}$$

where s is a user-specified parameter, $J(\beta) = (1-\alpha)||\beta||_1 + \alpha||\beta||^2$ and weight $\alpha = \frac{\lambda_2}{\lambda_1+\lambda_2}$.

Usually, tuning parameters $\alpha \in [0,1]$ and $s \ge 0$ can be chosen via well-established methods like cross-validation (Hastie *et al.*, 2005). When $\alpha = 0$, the elastic net penalty becomes lasso penalty (Tibshirani, 1996), which selects at most $\min\{\dim(\beta), N^{(r)}\}$ variables. If l increases and $N^{(r)}$ decreases, the number of selected variable is bounded by $N^{(r)}$. When $\alpha \in (0,1)$, it is the elastic net penalty. Although it can select variables without limitation on the lower bound and encourages group effects, the solution is yielded via a stage-wise LARS-EN algorithm (Zou and Hastie, 2005) rather than in a closed form formula. When $\alpha = 1$, the elastic net penalty becomes bridge regression model (Fu, 1998), which fits any variable selection situation because of general penalty form (Park and Yoon, 2011). On the other hand, bridge regression model yields the closed form solution $\hat{\beta} = (\mathbf{X}^{(r)\top}\mathbf{X}^{(r)} + \lambda_2\mathbf{I})^{-1}\mathbf{X}^{(r)\top}\mathbf{y}^{(r)}$ where $\mathbf{I}$ is the unit matrix of size $3l+1$.

Applying $\hat{\mathbf{i}}_{1/2}^{(r)} = \mathbf{X}^{(r)}\hat{\beta}$ to data in unreliable windows, we obtain a predicted half-saturation parameter for both reliable and unreliable windows, denoted as $\hat{\mathbf{i}}_{1/2} = (\hat{\mathbf{i}}_{1/2}^{(r)\top}, \hat{\mathbf{i}}_{1/2}^{(u)\top})^\top = (\hat{i}_{1/2,1}^{(r)}, \ldots, \hat{i}_{1/2,N^{(r)}}^{(r)}, \hat{i}_{1/2,1}^{(u)}, \ldots, \hat{i}_{1/2,N^{(u)}}^{(u)})^\top$. $\hat{\mathbf{i}}_{1/2}$ ensures smooth hidden state variables under dynamic environments.

The lower solid (blue) line in Figure 4C shows the results on one of the unreliable temporal windows. In this window, several sampled Φ_{II} values are far from the ground truth (blue dots), which results in biased local curve fitting results (gray line) with $R^2 = 0.70$. The polynomial regularization (blue solid line), on the contrary, is independent of the local data, which suggests the actual curve should be lower than the locally fitted curve.

3.4 Maximum posterior estimations

Given the phenotype data in the jth unreliable temporal window, we re-estimate its local half light parameter $\hat{i}_{1/2,j}^{(u)*}$. Our aim is to identify the half light parameter that fit best with both the local data in the jth unreliable temporal window and the estimated half light value $\hat{i}_{1/2,j}^{(u)}$ learned from the regularized polynomial linear model in the previous step.

Mathematically, for the fixed jth unreliable window, we rewrite Equation (2) with error terms as

$$\Phi_{II,k}^{(u)} = \frac{\max(\Phi_{II})}{1 + i_k^{(u)}/i_{1/2,j}^{(u)*}} + \varepsilon_k^{(u)} \tag{8}$$

for $j \le k \le n+j$, where $\varepsilon_k^{(u)}$ follows $N(0, \sigma_1^2)$. It yields

$$p(\Phi_{II,k}^{(u)} | i_k^{(u)}, i_{1/2,j}^{(u)*}) \propto \exp\left[-\frac{1}{2\sigma_1^2} \left(\Phi_{II,k}^{(u)} - \frac{\max(\Phi_{II})}{1 + i_k^{(u)}/i_{1/2,j}^{(u)*}} \right)^2 \right] \tag{9}$$

Assume $i_{1/2,j}^{(u)*}$ have a normal prior with hypoparameters $\hat{i}_{1/2,j}^{(u)}$ and σ_2^2, that is,

$$p(i_{1/2,j}^{(u)*}) \propto \exp\left[-\frac{1}{2\sigma_2^2} \left(i_{1/2,j}^{(u)*} - \hat{i}_{1/2,j}^{(u)} \right)^2 \right] \tag{10}$$

The posterior of $i_{1/2,j}^{(u)*}$ is given by

$$p(i_{1/2,j}^{(u)*} | W_j^{(u)}) = \frac{p(i_{1/2,j}^{(u)*}, W_j^{(u)})}{p(W_j^{(u)})} \propto p(i_{1/2,j}^{(u)*}, W_j^{(u)})$$

$$= p(i_{1/2,j}^{(u)*}) p(W_j^{(u)} | i_{1/2,j}^{(u)*}) = p(i_{1/2,j}^{(u)*}) \prod_{k=j}^{n+j} p(\Phi_{II,k}^{(u)} | i_k^{(u)}, i_{1/2,j}^{(u)*}) \tag{11}$$

where $p(W_j^{(u)}|i_{1/2,j}^{(u)*})$ is the joint likelihood function. Thus,

$$\log p(i_{1/2,j}^{(u)*}|W_j^{(u)}) \propto \log p(i_{1/2,j}^{(u)*}) + \log \prod_{k=j}^{n+j} p(\Phi_{II,k}^{(u)}|i_k^{(u)}, i_{1/2,j}^{(u)*})$$

$$= -\frac{\left(i_{1/2,j}^{(u)*} - \hat{i}_{1/2,j}^{(u)}\right)^2}{2\sigma_2^2} - \sum_{k=j}^{n+j} \frac{\left(\Phi_{II,k}^{(u)} - \frac{\max(\Phi_{II})}{1+i_k^{(u)}/i_{1/2,j}^{(u)*}}\right)^2}{2\sigma_1^2} \quad (12)$$

The final equation that can be solved by maximizing the posterior estimation is given by:

$$\hat{i}_{1/2,j}^{(u)*} = \underset{i_{1/2}\in \mathrm{R}}{\operatorname{argmin}} \left(\frac{(i_{1/2} - \hat{i}_{1/2,j}^{(u)})^2}{2\sigma_2^2} + \sum_{k=j}^{j+n} \frac{\left[\Phi_{II,k}^{(u)} - \frac{\max(\Phi_{II})}{1+i_k^{(u)}/i_{1/2}}\right]^2}{2\sigma_1^2} \right) \quad (13)$$

One of the drawbacks of the sliding window approach is the fixed window width. If the majority data in a window happen to be noisy, the fitted curve will not be reliable no matter what model is used. Hence, we develop a simple procedure to expand each unreliable window by adding the closest phenotype–environment pair located in a reliable window, thus increasing the reliability of curve fitting.

In the example in Figure 4C, the middle solid (red) curve shows the results of maximizing the posteriori estimation on one of the unreliable temporal windows. By taking into consideration both the local Φ_{II} data and the global trend of $i_{1/2}$, the optimized curve is the closest to the ground truth. Besides, phenoCurve significantly increases the curve reliability R^2 by extending the scope of window to include the closest Φ_{II} value (orange dot) that is located in a reliable window. Figure 4D shows the results of PhenoCurve on the complete dataset. For comparison, the results of global and local fitting are also included. Clearly PhenoCurve (red line) is the closest to the ground truth while the global/local fitting (two gray lines) are significantly affected by the noisy inputs. See details in Supplementary Table S1.

4 Experimental results

We evaluate the performance of PhenoCurve on both the real and simulated phenotype data regarding fitting performance and fitting reliability. We also compare PhenoCurve with six existing methods, i.e. (i) the direct computation with PI function, (ii) the one-window curve fitting (ONE), (iii) the sliding-window based curve fitting (WIN), (iv) the kernel smoothing method using LLR, (v) the Bayesian linear model with NIG prior and (vi) robust LOWESS, all introduced in the Related Work Section.

4.1 Experimental data

We first test the performance of PhenoCurve using the real plant photosynthetic phenotype data consisting of 331 Arabidopsis thaliana plants (330 confirmed T-DNA insertion mutants and wild-type (Col-0) used as a reference, each with at least four biological replicates) (Ajjawi *et al.*, 2010; Alonso *et al.*, 2003). In the biological experiment, all the plants are evenly sampled at 32 time-points during 16 h. All the environmental factors except light are constant. Following a sinusoidal curve, light intensity changes gradually from 39 to 500 μmol m^{-2} s^{-1} then goes back to 39 μmol m^{-2} s^{-1} (details in Supplementary Fig. S1). The photosynthetic phenotype values vary dramatically across plants, reflecting potential differences in development, stress responses or regulation of processes such as stomatal conductance, photodamage, and storage of photosynthate (Cruz *et al.*, 2016). In PhenoCurve, we set three window sizes to include 6, 12 or 18 measurements, respectively. A window with 18 measurements is viewed approximately as half day, a window with 12 measurements is viewed approximately as a quarter of a day, and a window with 6 measurements is viewed approximately as a period of 3 h.

Second, we test the performance of PhenoCurve using synthetic data, which are generated in three steps. First, we randomly generate a vector of $i_{1/2}$ that changes gradually over time. Second, we reconstruct the Φ_{II} phenotype data using the vector of $i_{1/2}$, the same vector of light as the real data, and the PI function (MacIntyre *et al.*, 2002). Third, we randomly add noise (levels vary from 5 to 15%) to the phenotype data. This process has been repeated 2000 times to generate the synthetic data (see Supplementary Table S2).

Threshold $R^2 = 0.9$ is used on both the real and the synthetic datasets to define the reliable and unreliable windows. The highest order of polynomial term is set to two by using cross-validation.

4.2 Evaluation criteria

We define four criteria for systematic performance evaluation. We apply all of them to data in unreliable windows, data in reliable windows, and data in the whole windows.

First, coefficient of determination, denoted as R^2, is often used as the main criteria for measuring whether a curve fitting is adequate (Cameron and Windmeijer, 1997; Holland and Welsch, 1977). In our case, for window W_j, R^2 is defined as:

$$R^2 = 1 - \frac{\sum_{k=j}^{j+n} (\Phi_{II,k} - \hat{\Phi}_{II,k})^2}{\sum_{k=j}^{j+n} (\Phi_{II,k} - \bar{\Phi}_{II,j})^2} \quad (14)$$

where $\hat{\Phi}_{II,k}$ is the fitted $\Phi_{II,k}$ at time t_k, and $\bar{\Phi}_{II,j}$ is the averaged $\hat{\Phi}_{II,k}$ in temporal window W_j. In Equation (14), R^2 measures the fraction of the total variation in the phenotype data that can be explained by the curve. Higher values indicate that the curve fits the data better. If $R^2 = 1.0$, all points lie exactly on the curve with no scatter.

Second, we compute the smoothness of $i_{1/2}$ of each window. For a continuous curve, smoothness can be measured using high order of derivatives. For discrete values (which is our case), we measure all the angles formed by adjacent temporal windows:

$$\text{smoothness}(i_{1/2}) = \frac{1}{N} \sum_{j=2}^{N-1} [\alpha_j \le T_\alpha] \quad (15)$$

where $\alpha_j = |\arctan(\hat{i}_{1/2,j+1}^* - \hat{i}_{1/2,j}^*) - \arctan(\hat{i}_{1/2,j}^* - \hat{i}_{1/2,j-1}^*)|$ represents the angle difference centered around all windows, T_α is a user given angle threshold, and $[X] = 1$ if the condition X is satisfied, otherwise $[X] = 0$. In our experiment, $T_\alpha = 30^\circ$. A higher value indicates that the curve is smoother.

Finally, for synthetic data, we compute both $\Delta\Phi_{II}$ and $\Delta i_{1/2}$ and use both of them as evaluation criteria. $\Delta\Phi_{II}$ is the sum of all the absolute differences between every ground truth phenotype value and its corresponding value on the fitted curve. $\Delta i_{1/2}$ is defined as the sum of all the absolute differences between every ground truth $i_{1/2}$ value and its corresponding parameter of the fitted PI curve. Both of the criteria are the smaller, the better. Note that the last two criteria are only applicable to the synthetic data for which the ground truth is known. Similarly, not all the evaluation criteria are applicable for all the methods to compare. Please refer to details in Supplementary Section 1.

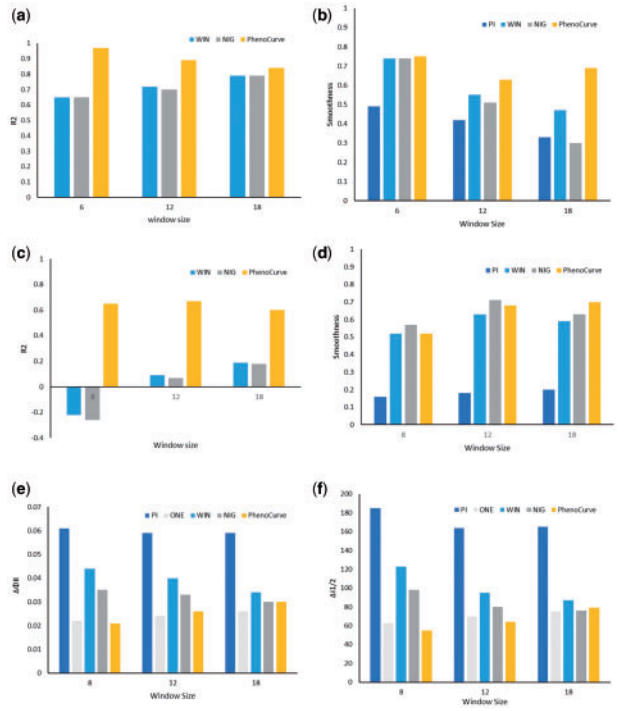

Fig. 5. Coefficient of determination R^2 and smoothness of $i_{1/2}$ on all the unreliable windows of the real data **(a,b)** and the synthetic data **(c,d)**. $\Delta\Phi_{II}$ and $\Delta i_{1/2}$ on the synthetic data **(e,f)** (Color version of this figure is available at *Bioinformatics* online.)

Table 1. Coefficient of determination R^2 and $\Delta i_{1/2}$ on the unreliable windows of synthetic phenotype data

Window Size		8	12	18
$\Delta i_{1/2}$	PhenoCurve	0.02	0.03	0.03
	LOWESS	0.05	0.05	0.05
R^2	PhenoCurve	0.65	0.67	0.60
	LOWESS	0.75	0.67	0.69

$\Delta i_{1/2}$ is the smaller the better.

4.3 Experimental results on real data

We run PhenoCurve on the real photosynthesis phenotype data using three different window sizes: 6, 12 and 18. The highest degree of polynomial term was set to two via using cross-validation. We compare the performance of PhenoCurve with the existing methods using the same parameters.

The experimental results indicate that PhenoCurve has more reliable results than the existing methods. Figure 5a shows that for the unreliable windows, the averaged coefficient of determination (R^2) increases 32% from 0.65 to 0.97, compared with WIN and NIG. With the increase of window size, the R^2 values of all the existing methods are steadily increased from 0.65 to 0.79 but are still lower than the R^2 of PhenoCurve (0.84).

With the increase of window size from 6 to 18, the curve smoothness of all the existing methods decreases significantly from 0.74 to 0.47, while the smoothness of PhenoCurve (varying between 0.63 and 0.75) is insensitive to window size (see Fig. 5b).

On the complete dataset (see Supplementary Fig. S3a and b), which includes both data in the reliable and the unreliable windows, the same trend remains, except that the results of WIN and NIG are smoother than PhenoCurve when the window size is small. Directly using the PI results in the least smooth curve, because PI is sensitive to noise.

Table 2. The robustness test of PhenoCurve with multiple noise and bias rates

Noise rates		5%	10%	15%
Unreliable windows	R^2	67.07%	86.57%	118.37%
	smoothness	7.14%	7.35%	3.28%
	$\Delta\Phi_{II}$	−83.33%	−53.85%	−45.95%
	$\Delta i_{1/2}$	−35.29%	−50.00%	−41.67%
All the windows	R^2	24.42%	36.49%	45.76%
	smoothness	20.00%	4.88%	5.13%
	$\Delta\Phi_{II}$	−33.33%	−31.82%	−27.27%
	$\Delta i_{1/2}$	−15.07%	−25.37%	−21.74%

The values are the performance improvements by comparing PhenoCurve with the local curve fitting method (WIN).

4.4 Experimental results on synthetic data

We compare the performance of PhenoCurve and the five existing methods on the synthetic phenotype data using three different window sizes. The synthetic data is introduced in Section 4.1. The results of PhenoCurve are in Supplementary Table S3.

The experimental results indicate that PhenoCurve is more reliable than the compared methods. Figure 5c shows that on the unreliable windows, the R^2 values of WIN and NIG are negative because of the noisy synthetic phenotype data, and PhenoCurve significantly increases the R^2 value from −0.26 to 0.65 (91% improvement). On the whole synthetic data (see Supplementary Fig. S4a), PhenoCurve also increases the R^2 value from 0.32 to 0.74 (42% improvement). Regarding the smoothness test, all the tested algorithms (except PI) have similar performance (see Fig. 5d and Supplementary Fig. S4b).

Figure 5c and Supplementary Figure S5a and b show that in all the tests PhenoCurve has overall smaller residues than PI, WIN and NIG, indicating the Φ_{II} curve generated by PhenoCurve is the closest to the ground truth, as demonstrated in Supplementary Figure S4D. The results show that PhenoCurve is overall the best except that when the window size is large (18) all the methods (except PI) have the similar performance.

We also compared PhenoCurve with the nonparametric smoothing technique LOWESS on all the unreliable windows of the synthetic data. The results in Table 1 show that although LOWESS has a higher R^2 than that of PhenoCurve, it always yields higher errors of $i_{1/2}$ than PhenoCurve does due to the lack of the ability to adopt a knowledge model to reduce the number of parameters.

4.5 Robustness test

In order to test the robustness of PhenoCurve, we run PhenoCurve on a synthetic dataset with three different levels of noise (5, 10 and 15%). Table 2 includes the performance improvement at each noise rate by comparing PhenoCurve and the local curve fitting method (WIN), the most commonly used model. It shows that PhenoCurve can constantly improve performance, especially the curve reliability (R^2). It also indicates that PhenoCurve is more robust than the local curve fitting method with the increase of the noise rate. Since both $\Delta\Phi_{II}$ and $\Delta i_{1/2}$ are the smaller, the better, the negative ratios in Table 2 indicate the results of PhenoCurve is closer to the ground truth than WIN at all the settings.

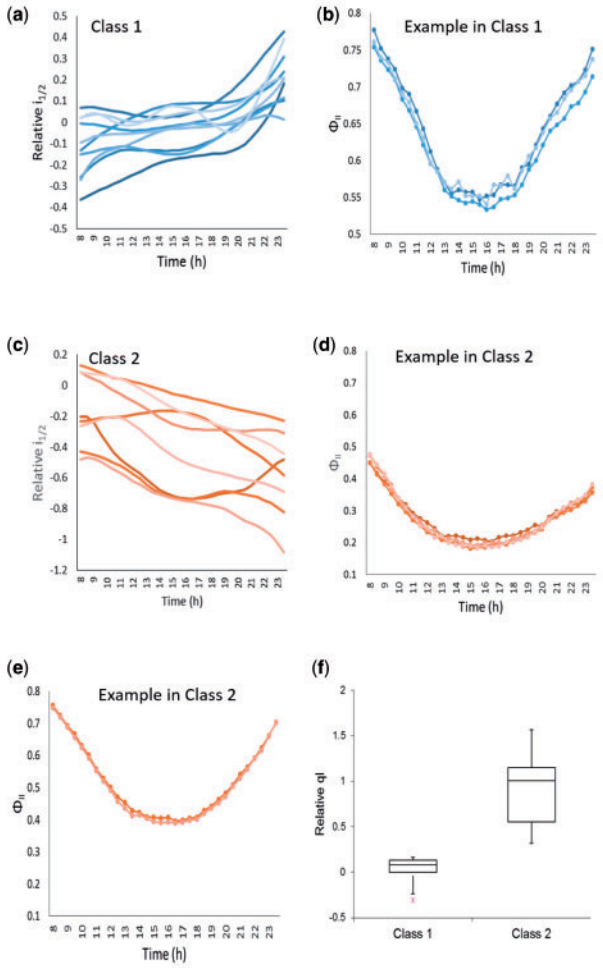

Fig. 6. Trend analysis of $i_{1/2}$ reveals two types of hysteresis patterns in Φ_{II}, which have a strong correlation with photoinhibition *qI*

4.6 Biological application

Diurnal hysteresis in photosynthesis is a phenomenon where the photosynthetic performance of an organism is asymmetric (e.g. lower in the afternoon than in the morning or vice versa) even at symmetric environmental conditions. This phenomenon has been widely observed in phytoplankton, macroalgae, and higher plants (Levy *et al.*, 2004). Mutant lines in which the photosynthetic hysteresis changes may indicate that a loss of function impacts photosynthesis by influencing the rate of photodamage, rate of repair, sink capacity, and so on. However, photosynthetic hysteresis is difficult to quantify especially under dynamic conditions.

By identifying the key parameter $i_{1/2}$ governing the photosynthesis phenotype–environment relationship, we test whether employing PhenoCurve can solve the problem. The rationale is that $i_{1/2}$ is an indicator of how sensitively photosynthetic rate responds to light intensity (Equation 1)—a smaller $i_{1/2}$ means that photosynthesis saturates at lower light intensity. This level of control is exerted primarily by modulation of Φ_{II} (Equation 2). Decreases in $i_{1/2}$ yields a relative decreases in Φ_{II}, assuming max Φ_{II} is relatively constant.

Given all the $i_{1/2}$ values of all the 330 Arabidopsis mutant lines obtained in Section 4.3, we compute their relative $i_{1/2}$ values by comparing each $i_{1/2}$ value with the corresponding value of wild-type plants (Col-0) in the same flat (a flat is a set of plants phenotyped together). And then we compute the trend of the relative $i_{1/2}$ for each mutant using linear regression.

From the data, we identify two types of hysteresis patterns, i.e. class 1 and class 2 shown in Figure 6a and c, respectively. In both figures, each line represents a serial of relative $i_{1/2}$ values of a mutant line. Clearly, classes 1 and 2 have the opposite trend. One of the mutant lines in class 1 is shown in Figure 6b, where each line represents the Φ_{II} of a biological replicate. It shows that the phenotype measures are highly reproducible, and the photosynthetic efficiency recovers in the afternoon, indicating the mutant line can efficiently repair any photodamage caused by high light during mid-day, or in the previous day. In class 2, a case study (Fig. 6d) shows low Φ_{II} in the afternoon, suggesting photoinhibition caused by strong light at noon. See Supplementary Figure S2 for the hysteresis patterns.

In order to verify the discovery, we conduct a biological experiment to measure *qI* (photoinhibition) of the same mutants under the same conditions. Figure 6f shows the logged fold change of *qI*, and it indicates that the plants in class 2 have significantly higher photoinhibition than those in class 1. For example, the logged fold change of *qI* of the mutants in Figure 6B and E are 0.12 and 1.57, respectively. Another case study in class 2 (Fig. 6e), on the contrary, has insignificant *qI* (0.3) but is hysteresis, suggesting either congestion of linear electron flow or PSI damage.

In summary, learning the phenotype–environment relationships with PhenoCurve simplifies hysteresis pattern detection, thus enabling biologists to discover the mechanisms that regulate responses to dynamic environments. Although the experiment is conducted using smooth and symmetric light conditions for easy visualization and validation, the model can be easily extended for experiments with rapid environmental perturbations.

5 Conclusion

In plants, photosynthesis is the primary energy source for metabolism and growth. With the large volume of photosynthesis phenotype data that has been collected, normalized and cleaned, biologists expect to immediately identify mutant lines with efficient photosynthetic machinery, and quickly generate and test biological hypotheses that may lead to a new breakthrough in bio-energy research. To meet the growing needs, we develop a new tool called PhenoCurve to study the dynamic relationships between phenotypes and environments using biological knowledge. PhenoCurve splits the whole phenotype and environment data into highly overlapped temporal windows, employs the PI curve and non-linear curve fitting methods to calculate $i_{1/2}$ for the reliable windows, and then optimizes the $i_{1/2}$ for the unreliable windows using regularized polynomial regression and maximum posterior estimations. The results on both the real and the synthetic data show that PhenoCurve is significantly better than the existing methods. We also demonstrate that PhenoCurve can be directly used for hysteresis pattern detection. We will extend PhenoCurve to model multiple phenotypes. A technical challenge is that with more data types, the number of parameters rapidly increases. We will also generalize PhenoCurve for broader applications including growth prediction and early disease detection without relying on the PI function.

Funding

This research was supported by the National Science Foundation (1458556), the US Department of Energy (DE-FG02-91ER20021) and MSU Center for Advanced Algal and Plant Phenotyping.

Conflict of Interest: none declared.

References

Ajjawi,I. *et al*. (2010) Large-scale reverse genetics in arabidopsis: case studies from the chloroplast 2010 project. *Plant Physiol*., **152**, 529–540.

Alonso,J. *et al*. (2003) Genome-wide insertional mutagenesis of arabidopsis thaliana. *Science*, **301**, 653–657.

Baker,N. *et al*. (2007) Determining the limitations and regulation of photosynthetic energy transduction in leaves. *Plant Cell Env*., **30**, 1107–1125.

Bates,D., and Watts,D. (1988) *Nonlinear Regression: Iterative Estimation and Linear Approximations*. Wiley Online Library, New York.

Bishop,C. (2006) *Pattern Recognition and Machine Learning*. Springer, New York.

Bolstad,W. (2013) *Introduction to Bayesian Statistics*. John Wiley & Sons, Hoboken, New Jersey.

Cameron,C., and Windmeijer,F. (1997) An r-squared measure of goodness of fit for some common nonlinear regression models. *J. Econometrics*, **77**, 329–342.

Chou,T., and TaLaLay,P. (1981) Generalized equations for the analysis of inhibitions of michaelis-menten and higher-order kinetic systems with two or more mutually exclusive and nonexclusive inhibitors. *Eur. J. Biochem*., **115**, 207–216.

Cleveland,W. (1979) Robust locally weighted regression and smoothing scatterplots. *j Am Stat Assoc*, **74**, 829–836.

Cruz,J. *et al*. (2016) Dynamic environmental photosynthetic imaging (depi) reveals emergent phenotypes related to the environmental responses of photosynthesis. *Cell Syst*., **2**, 365–377.

Dowd,J., and Riggs,D. (1965) A comparison of estimates of michaelis-menten kinetic constants from various linear transformations. *J. Biol. Chem*., **240**, 863–869.

Eilers,P., and Peeters,J. (1988) A model for the relationship between light intensity and the rate of photosynthesis in phytoplankton. *Ecol. Model*, **42**, 199–215.

Freedman,D. (2009) *Statistical Models: Theory and Practice*. Cambridge University Press, Cambridge.

Fu,W. (1998) Penalized regressions: the bridge versus the lasso. *J. Comput. Graph. Stat*., **7**, 397–416.

Gao,Q. *et al*. (2016) Inter-functional analysis of high-throughput phenotype data by nonparametric clustering and its application in photosynthesis. *Bioinformatics*, **32**, 67–76.

Gelman,A. (2006) Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian Anal*., **1**, 515–534.

Green,J. *et al*. (2012) Phenophyte: a flexible affordable method to quantify 2d phenotypes from imagery. *Plant Methods*, **8**, 1–12.

Großkinsky,D. *et al*. (2015) Plant phenomics and the need for physiological phenotyping across scales to narrow the genotype-to-phenotype knowledge gap. *J. Exp. Bot*., **66**, 5429–5440.

Gupta,M. *et al*. (2008) Adaptive local linear regression with application to printer color management. *Image Proces., IEEE Trans*, **17**, 936–945.

Hastie,T. *et al*. (2005) The elements of statistical learning: data mining, inference and prediction. *Math. Intell*., **27**, 83–85.

Holland,P., and Welsch,R. (1977) Robust regression using iteratively reweighted least-squares. *commun Stat-Theor M*, **6**, 813–827.

Houle,D. *et al*. (2010) Phenomics: the next challenge. *Nat. Rev. Genet*., **11**, 855–866.

Jassby,A., and Platt,T. (1976) Mathematical formulation of the relationship between photosynthesis and light for phytoplankton. *limnol Oceanogr*, 540–547.

Kramer,D., and Evans,J. (2011) The importance of energy balance in improving photosynthetic productivity. *Plant Physiol*., **155**, 70–78.

Kutsukake,M. *et al*. (2012) An insect-induced novel plant phenotype for sustaining social life in a closed system. *Nat. Comm*., **3**, 1187.

Levenberg,K. (1944) A method for the solution of certain non–linear problems in least squares. *Q. J. Appl. Math*., **2**, 164–168.

Levy,O. *et al*. (2004) Diurnal hysteresis in coral photosynthesis. *Mar. Ecol-Prog. Ser*., **268**, 105–117.

MacIntyre,H. *et al*. (2002) Photoacclimation of photosynthesis irradiance response curves and photosynthetic pigments in microalgae and cyanobacteria. *J. Phycol*., **38**, 17–38.

Menten,L., and Michaelis,M. (1913) Die kinetik der invertinwirkung. *Biochem. Z*, **49**, 333–369.

Motulsky,H., and Christopoulos,A. (2004) *Fitting Models to Biological Data Using Linear and Nonlinear Regression: A Practical Guide to Curve Fitting*. Oxford University Press, Oxford.

Nicotra,A. *et al*. (2010) Plant phenotypic plasticity in a changing climate. *Trends Plant Sci*., **15**, 684–692.

Osborne,J., and Overbay,A. (2004) The power of outliers (and why researchers should always check for them). *Pract. Assess. Res. Eval*., **9**, 1–12.

Park,C., and Yoon,Y. (2011) Bridge regression: adaptivity and group selection. *J. Stat. Plan Infer*., **141**, 3506–3519.

Price,T. *et al*. (2003) The role of phenotypic plasticity in driving genetic evolution. *R. Soc. Lond. B*, **270**, 1433–1440.

Rascher,U. *et al*. (2011) Non-invasive approaches for phenotyping of enhanced performance traits in bean. *Funct. Plant Biol*., **38**, 968–983.

Serôdio,J., and Lavaud,J. (2011) A model for describing the light response of the nonphotochemical quenching of chlorophyll fluorescence. *photosynth Res*, **108**, 61–76.

Subramanian,R. *et al*. (2013) A high throughput robot system for machine vision based plant phenotype studies. *Mach. Vision Appl*., **24**, 619–636.

Tessmer,O. *et al*. (2013) Functional approach to high-throughput plant growth analysis. *BMC Syst. Biol*., **7(Suppl 6)**, S17.

Thrampoulidis,C.S. *et al*. (2015) Regularized linear regression: A precise analysis of the estimation error. In *Learning Theory Conf*. Paris, France, pp. 1683–1709.

Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. Roy. Stat. Soc. B. Met*., **58**, 267–288.

Vlasblom,J., *et al*. (2015) Novel function discovery with genemania: a new integrated resource for gene function prediction in escherichia coli. *Bioinformatics*, **31**, 306–310.

Walter,A. *et al*. (2015) Plant phenotyping: from bean weighing to image analysis. *Plant Methods*, **11**, 14.

Wong,M. *et al*. (2012) Prediction of susceptibility to major depression by a model of interactions of multiple functional genetic variants and environmental factors. *mol Psychiatr*, **17**, 624–633.

Xu,L. *et al*. (2015) Plant photosynthesis phenomics data quality control. *Bioinformatics*, **31**, 1796–1804.

Yang,J., and Leskovec,J. (2011) Patterns of temporal variation in online media. In *Web Search and Data Mining, ACM Intl Conf on*, pp. 177–186.

Yin,X., *et al*. (2014a) Multi-leaf alignment from fluorescence plant images. In *Appl of Comp Vis, IEEE Winter Conf. on*, pp. 437–444. Steamboat Springs CO.

Yin,X., *et al*. (2014b) Multi-leaf tracking from fluorescence plant videos. In *Img Proc, IEEE Intl. Conf. on*, pp. 408–412. Paris, France.

Zou,H., and Hastie,T. (2005) Regularization and variable selection via the elastic net. *J. Roy. Stat. Soc. B*., **67**, 301–320.