Bioinformatics, 31(11), 2015, 1796–1804
doi: 10.1093/bioinformatics/btu854
Advance Access Publication Date: 22 January 2015
Original Paper

OXFORD

Data and text mining

Plant photosynthesis phenomics data quality control

Lei Xu[1], Jeffrey A. Cruz[2], Linda J. Savage[2], David M. Kramer[2,3,*] and Jin Chen[1,2,*]

[1]Department of Computer Science and Engineering, [2]Department of Energy Plant Research Laboratory and [3]Department of Biochemistry and Molecular Biology, Michigan State University, MI, East Lansing 48824, USA

*To whom correspondence should be addressed

Associate Editor: Robert F. Murphy

Received on August 21, 2014; revised on December 1, 2014; accepted on December 23, 2014

Abstract

Motivation: Plant phenomics, the collection of large-scale plant phenotype data, is growing exponentially. The resources have become essential component of modern plant science. Such complex datasets are critical for understanding the mechanisms governing energy intake and storage in plants, and this is essential for improving crop productivity. However, a major issue facing these efforts is the determination of the quality of phenotypic data. Automated methods are needed to identify and characterize alterations caused by system errors, all of which are difficult to remove in the data collection step and distinguish them from more interesting cases of altered biological responses.

Results: As a step towards solving this problem, we have developed a coarse-to-refined model called dynamic filter to identify abnormalities in plant photosynthesis phenotype data by comparing light responses of photosynthesis using a simplified kinetic model of photosynthesis. Dynamic filter employs an expectation-maximization process to adjust the kinetic model in coarse and refined regions to identify both abnormalities and biological outliers. The experimental results show that our algorithm can effectively identify most of the abnormalities in both real and synthetic datasets.

Availability and implementation: Software available at www.msu.edu/%7Ejinchen/DynamicFilter

Contact: jinchen@msu.edu or kramerd8@cns.msu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Plants capture sunlight to fix CO_2 into energy rich molecules, thus supplying our ecosystem with O_2 and essentially all of its biological energy, including 100% of our food. Recent work has focused on improving the efficiency of photosynthesis to meet our growing needs for food and fuel (Bonner, 1962; Kramer and Evans, 2011; Von Caemmerer and Farquhar, 1981). To develop efficiency-boosting mechanisms that reduce energy losses or enhance CO_2 delivery to cells during photosynthesis, advanced technologies in high-throughput plant photosynthetic phenotyping and pheno-informatics have been developed (Cruz *et al.*, 2014; Houle *et al.*, 2010; Tessmer *et al.*, 2013; Zhu *et al.*, 2010). These technologies have allowed plant photosynthesis phenotypic variability to be characterized and to be related to putative biological functions, leading to a better understanding of the underlying mechanisms that control photosynthetic properties under various environmental conditions. Plant phenomics is a first-class asset for understanding the mechanisms regulating energy intake in plants (Fiorani and Schurr, 2013; Rascher *et al.*, 2011).

Plant phenotyping systems monitor photosynthetic performance for many plants both continuously and simultaneously. Phenomics datasets are large and continue to grow as we increase duration of sampling and resolution. Yet despite the size and richness of the data, small clusters of erroneous values, which give the appearance

of real differences in biological responses, can skew the analysis towards an invalid interpretation (Herbert *et al.*, 2004). There are several ways in which a measurement can be in error: errors originating from instrumentation malfunctions, biased values from miscalibrated sensors and inevitable errors of precision. All these issues compromise the downstream data analysis tasks. Given the value of clean data for any operation, the ability to improve data quality is a key requirement for effective knowledge mining from large-scale phenotype data.

In this article, we focus on data abnormalities detection, which is a type of measurement errorto demonstrate how clean phenotype data can be obtained. Similar to sensor data, abnormalities in plant phenotype data deviate significantly from expected patterns and are visible outliers in the whole dataset (Shanahan, 2005; Subramaniam *et al.*, 2006). The majority of abnormalities in plant phenotyping originate from instrumentation malfunctions (e.g. loss of sensor synchronization during measurement) or non-biological statistical outliers caused by data collection limitations (e.g. deterioration of signal-to-noise ratio for a sample as it progresses through the experiment).

Data abnormalities are often viewed as outliers in the whole dataset. Recent work has shown the effectiveness of applying data mining techniques, especially outlier detection, for the purpose of data cleaning (Maletic and Marcus, 2000), making it possible to automate the cleansing process for a variety of domains (Chu *et al.*, 2005; Ebaid *et al.*, 2013; Mayfield *et al.*, 2010; Pearson, 2002). In these methods, by detecting the minorities of values that do not conform to the general characteristics of a given data collection, outliers are identified and are considered violations of association rules or other patterns in the data. However, the existing models are not suitable for phenotype data cleaning. These methods, while applied to phenotype data, may remove outliers including both measurement errors and true biological discoveries, since true biological discoveries, to some extent, are outliers as well. Furthermore, detecting abnormalities from long time-series phenotype data requires handling a high temporal dimension, which increases the model complexity.

To identify and remove abnormalities in phenotype data and to minimize the deletion of biological discoveries, we have developed a coarse-to-refined residual analysis algorithm, called *dynamic filter*. Dynamic filter has three key steps: (i) identify abnormal candidates at the coarse level, (ii) refine abnormality identification in a projected feature space and (iii) iteratively identify abnormalities at the refined level. Dynamic filter can speed up the data preparation process and make it more effective. Such improvements will minimize time-consuming and labour-intensive data preparation and increase the significance and confidence in biological discoveries. In summary, our model has the following advantages:

- To our knowledge, dynamic filter is the first work to integrate biological constraints with time-series phenotype data for data cleaning.
- Our model can identify both abnormalities and biological discoveries.
- Dynamic filter outperforms the existing solutions by optimizing the fitness between phenotype data and biological constraints.

2 Background

Data cleaning is the process of identifying incorrect or corrupted records in a dataset. The goal of data cleaning is to ensure an accurate representation of the real-world constructs to which the data refer. Removing impurities from data is traditionally an engineering problem, where *ad hoc* tools made up of low-level rules (such as detecting syntax errors) and manually tuned algorithms are designed for specific tasks (such as the elimination of integrity constraints violations) (Muller and Freytag, 2005). Detection and elimination of complex errors representing invalid values, however, go beyond the checking and enforcement of integrity constraints. They often involve relationships between two or more attributes that are very difficult to uncover and describe by integrity constraints. Recent work has shown the effectiveness of applying techniques from statistical learning for the purpose of data cleaning. In particular, outlier detection methods have made it possible to automate the cleansing process for a variety of domains (Chu *et al.*, 2013; Ebaid *et al.*, 2013; Koh *et al.*, 2007; Maletic and Marcus, 2000; Mayfield *et al.*, 2010; Pearson, 2002).

However, none of the existing outlier-detection based methods are suitable for phenotype data cleaning. First, both biological discoveries and errors of detection are difficult to separate from distribution. Second, the cohesiveness rule used in temporal data cleaning is not applicable for the phenotype data, because (i) a non-cohesive time-serial could represent an interesting phenotype pattern rather than an error; (ii) all the observations at the same time point may be similarly affected by a systematic abnormal event (Muller and Freytag, 2005).

Alternatively, rather than checking the raw values, residue analysis can be employed to model the differences between the real values and the theoretical curve, which is usually derived from biological constraints such as generalized light reactions (Jassby and Platt, 1976; MacIntyre *et al.*, 2002). This is often called the *goodness-of-fit* model. The goodness-of-fit based data cleaning models can be classified into two categories. First, statistical distribution characters such as mean, standard deviation, confidence interval or range have been used to find unexpected values indicating possible invalid values (Maletic and Marcus, 2000). Such simple methods can be efficiently applied to big data. However, these parameters (such as mean) are inclined to be biased by abnormalities with large deviations. Since it does not take into account local characteristics of data, there is a risk of mislabelling a range of normal data as abnormalities and vice versa. Second, combined data-mining techniques are used to identify patterns that apply to most residual records. A pattern is defined by a group of residuals that have similar characteristics (behaviour for certain percentage of the fields in the dataset). Outliers are then identified as values that do not conform to the patterns in the data. Among them, the Hampel filter uses the median of neighbouring observations as a reference value and looks for local outliers in a streaming data sequence (Pearson, 2002, 2005). While the Hampel filter is suitable for temporal data cleaning, it assumes that the data are independent and identically distributed, which is not valid under dynamic environmental conditions.

It should be noted that while the goodness-of-fit based data cleaning models focus on the modelling of deviation, they are not aware that the theoretical curve, which is used as the reference, may not always be precise. Typically, theoretical curves derived from biological knowledge are simple compared with the real-world situation. It is therefore inappropriate to directly use the imperfect theoretical curve to infer abnormalities.

In this article, we develop a coarse-to-refined residual analysis model called dynamic filter to effectively identify abnormalities in plant photosynthesis phenotype data. Our model derives a theoretical curve from the photosynthetic biological constraints; adjusts the theoretical curve to fit the phenotype data via optimization and studies the deviations of individual phenotype values from

theoretical curve. The resulting patterns in residuals indicate abnormalities, which are types of errors of detection, and the optimized theoretical curves reveal true biological outliers.

3 Methods

In this section, we first introduce the theoretical curve of time-series steady-state quantum yield data and then introduce a framework for abnormality detection.

In this article, the time-series steady-state quantum yield of photosystems II (denoted as Φ_{II}) is chosen for abnormality detection for three reasons. First, Φ_{II} can be readily measured using fluorescence video imaging making it useful for high-throughput phenotyping. Second, because it reflects light-driven electron transfer, it can be used as an indicator of photosynthetic rates and efficiency, albeit with the caveat that it reflects the sum of CO_2 fixation, photorespiration and other processes (Ögren and Evans, 1993). Finally, Φ_{II} is a good demonstration of the approach because it tends to follow, to a reasonable degree, relatively simple saturation behaviours. Given an adequate model, the cleaning procedure described in the manuscript may also be applied to other photosynthetic parameters like non-photochemical quenching (NPQ), which can display complex behaviours.

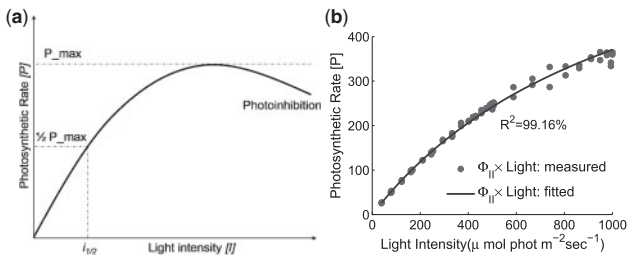
3.1 Theoretical photosynthetic curve

An abnormality in residual analysis is an observation exhibiting a large difference between the theoretical value and the observed value and may indicate a data entry error from the phenotyping sensors. To derive the theoretical curve, we model Φ_{II} with the photosynthesis-irradiance (PI) curve (see Fig. 1a) (Jassby and Platt, 1976; MacIntyre *et al.*, 2002).

As a derivation of Michaelis–Menten kinetics, one of the best-known models of enzyme kinetics in biochemistry (Menten and Michaelis, 1913), PI is modelled as a hyperbolic curve (see Fig. 1a) in Equation (1), revealing the empirical relationship between solar irradiance and photosynthesis (MacIntyre *et al.*, 2002).

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

where P is photosynthetic rate at a given light intensity, $P_{\max}$ is the maximum potential photosynthetic rate per individual, $[I]$ is a given light intensity and $i_{1/2}$ is half-saturation constant. Figure 1a shows the generally positive correlation between light intensity and photosynthetic rate. The PI curve has already been applied successfully to explain ocean-dwelling phytoplankton photosynthetic response to changes in light intensity (Jassby and Platt, 1976) as well as terrestrial and marine reactions.



Hyperbolic PI curve. Measured photosynthetic rates of Col-0 and the fitted PI curve.

Fig. 1. PI curve. (**a**) Hyperbolic PI curve. (**b**) Measured photosynthetic rates of Col-0 and the fitted PI curve. R^2 is computed based on the unexplained variance

We describe the photosynthetic rate P in terms of linear electron flow (Kramer and Evans, 2011) and associate both temporal steady-state quantum yield of photosystems II Φ_{II} and temporal light intensity i with time t, as shown in Equation (2):

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

where t is a time point in a user-defined temporal region T ($t \in T$); $\Phi_{II}(t)$ and $i(t)$ represent the steady-state quantum yield of photosystems II and light intensity at t; $\max(\Phi_{II})$ is the maximal Φ_{II} in T; and the half-saturation constant $i_{1/2}$ is the light intensity at which the photosynthetic rate proceeds at half $P_{\max}$. See proof in Supplementary Section S1.

One may reasonably ask if the NPQ or photoinhibition would affect the theoretical model for light saturation. In fact, NPQ has (surprisingly) little effect on the relationship between Φ_{II} and light intensity, as can be readily seen in the fact that the Φ_{II} light saturation curves for wild type and the *npq4* mutant of Arabidopsis are essentially identical despite large differences in qE (i.e. rapidly reversible photoprotection of NPQ) (Li *et al.*, 2000). The reason for this apparent disconnect is that, at high light, the slowest step in the light reactions of photosynthesis occurs subsequent to light absorption at the cytochrome *b6f* complex and is finely regulated by the pH of the lumen (Takizawa *et al.*, 2007). Light absorption become rate limiting only at NPQ levels much higher than those observed here. The biological role of NPQ under most conditions appears to be in regulating electron transfer but in preventing the build up of reactive intermediates within the photosystem II reaction centre (Muller *et al.*, 2001). Thus, the effects of moderate levels of NPQ and photoinhibition should have little effect on the behaviour of the wild-type system. However, under extreme conditions of in mutant lines with altered behaviour producing high levels of NPQ or photoinhibition, we expect to see behaviour that deviates from that produced by the model. These instances will be detected as outliers and flagged for further investigation of possible biological discoveries.

Consequently, the half-saturation constant $i_{1/2}$ can be learned using all $\Phi_{II}(t)$ and $i(t)$ in T with a non-linear regression method (Seber and Wild, 2003). Note that the half-saturation constant can be dramatically different between plants and between leaves in plants. Thus, the general shape of the curve is typically maintained but not its maximal or half-saturation light intensity.

Finally, given $i_{1/2}$, the residual value at each time point t is defined as

$$rsd(t) = \Phi_{II}(t) - \Phi'_{II}(t, i_{1/2}) \tag{3}$$

where $rsd(t)$ is the residual value at time t; and $\Phi_{II}(t)$ is the observed value and $\Phi'_{II}(t)$ is the theoretical value of steady-state quantum yield at t calculated using Equation (2).

We note that there are multiple models for PI curves, which give similar responses to light (de Lobo *et al.*, 2013; Govindjee *et al.*, 2005; Lambers *et al.*, 2008; Long and Hällgren, 1993; Zeinalov). In this article, we chose the Michaelis–Menten kinetics model because it is convenient to use and fits plant photosynthesis rate data well (see Fig. 1b). It should be noted that an important feature of our approach is that these alternative models can be easily added or substituted for comparison.

3.2 Framework of dynamic filter

Dynamic filter is a coarse-to-refined residual analysis approach, which has three major steps as shown in Figure 2. We define

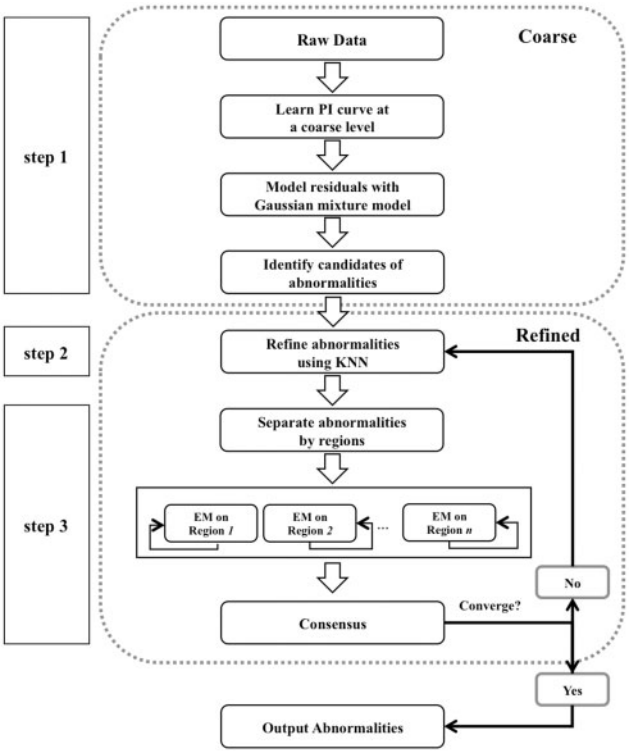


Fig. 2. The framework of dynamic filter (DF)

abnormalities using a definition to that proposed in Mayfield *et al.* (2010):

DEFINITION 1: Abnormality. *Let* $\{\Phi_{nor}\}$ *be a set of normal phenotype data, and* $\{rsd_{nor}\}$ *be the corresponding residual set. An abnormality* Φ_{abn} *is a phenotype value whose residual falling off the* α *confidence interval of the major normal distribution of* $\{rsd_{nor}\}$.

Note that confidence interval $\alpha = 99\%$ is commonly used in literature (Mayfield *et al.*, 2010; Sohn *et al.*, 2005), but is adjustable by users. In this article, by adopting the concept of confidence interval, we assume that (i) the majority of the phenotype values are correct, and (ii) they form the major distribution in the residual data, which is also distinctly different from the distribution(s) of the residual data of the abnormalities.

Step 1. Coarse process to identify abnormal candidates

Given a set of phenotype data Φ_{II}, we adopt Equation (2) to generate the theoretical values of steady-state quantum yields for each plant, denoted as $\{\Phi'_{II}\}$, by using the whole time-serial as temporal region T, aka the coarse level. For the dataset used in Section 4, the smallest value of time interval is 10 min, and the scale of T in the whole dataset is 3 days. Consequently, we generate the residual data of all plants $\{rsd\}$ using Equation (3), and model them using a Gaussian mixture model (GMM) (see details in Section 3.3 and example in Fig. 3a). Finally, we generate the abnormality candidate set $\{\Phi_{abn}\}$ with Definition 1. In Figure 3b, solid points are abnormality candidates in the coarse process. Clearly, because of the simplified PI curve model, not all the abnormality candidates are correctly identified.

Step 2. K-Nearest Neighbors (KNN) process to refine abnormality identification

Abnormality candidates may have certain intrinsic patterns of distribution highly related to certain ranges of feature space. For example,

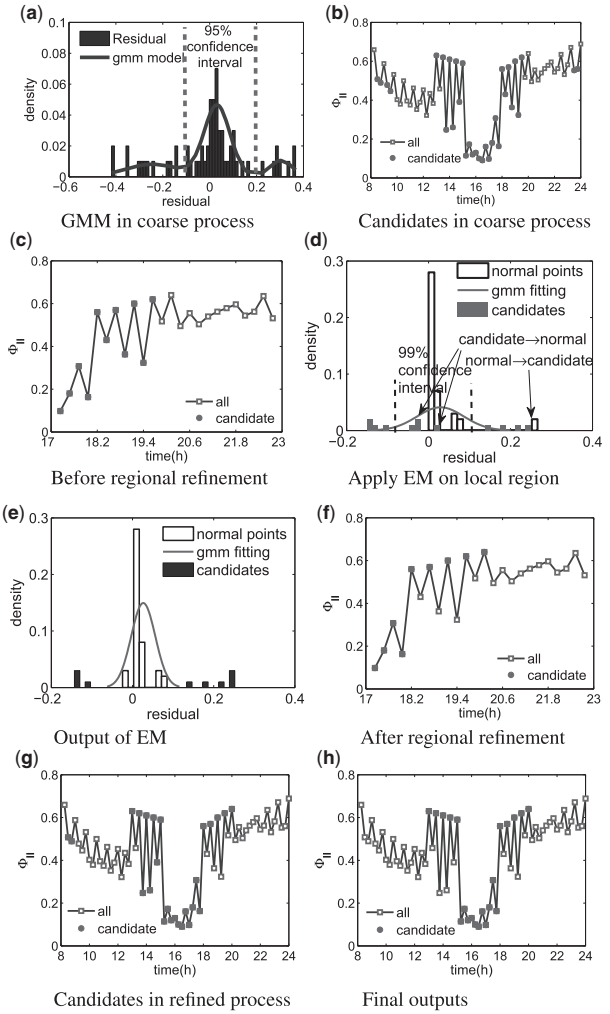


Fig. 3. An example of DF. The solid points are abnormalities and the hollow points are normal values. (**a**) GMM in coarse process. (**b**) Candidates in coarse process. (**c**) Before regional refinement. (**d**) Apply EM on local region. (**e**) Output of EM. (**f**) After regional refinement. (**g**) Candidates in refined process. (**h**) Final outputs

accidental dysfunction of data-capturing devices may cause abnormalities concentrated around some regions, which form statistical patterns on the distribution plot of the feature space. From a statistical viewpoint, abnormalities should be away from normal values in the feature space, and values with similar features tend to have the same labels. This leads to a refinement process to exploit the patterns of abnormalities candidates on selected feature space, and to make use of these patterns to refine abnormality identification, as described in Algorithm 1.

Specifically, we first select the optimal features from Φ_{II}, *rsd*, *i*, *t*, etc., in which abnormalities and normal values are maximally separated. To solve this feature reduction problem, linear discriminant analysis (LDA) is adopted to get the principal components of the optimal feature space (Algorithm 1 line 9–20, see details in Section 3.3). Second, we apply *K*-nearest-neighbour approach on the selected feature space, such that each abnormality candidate will be relabelled as its majority label of *k*-nearest-neighbours (Algorithm 1 line 4–7) (Altman, 1992).

Step 3. Refined process to identify abnormalities in local regions

Because the theoretical values $\{\Phi'_{II}\}$ are learned with the simplified PI curve model at the coarse level, not all the assignments of the

Algorithm 1 KNN process to refine results

1: **procedure** Refine(Ψ, C, k)
2: ▷ Ψ is original feature space, C is the set of labels (abnormality or normal)
3: $\Psi_{proj} \leftarrow$ FeatureSelection (Ψ, C)
4: **for** ψ_i in Ψ_{proj} **do**
5: $C_i \leftarrow$ majority label of k-nearest-neighbours
6: **end for**
7: **return** C
8: **end procedure**
9: **procedure** FeatureSelection(Ψ, C)
10: $\mu \leftarrow \frac{1}{|C|}\sum \psi$
11: **for** i from 1 to 2 **do** ▷ process both kinds of labels in C
12: $\Psi_i \leftarrow$ Features of i_{th} label
13: $n_i \leftarrow |\Psi_i|;\ \mu_i \leftarrow \frac{1}{n_i}\sum \psi_i$
14: $SW_i \leftarrow \frac{1}{n_i}\sum(\psi_i - \mu_i)(\psi_i - \mu_i)^T$
15: **end for**
16: $SW \leftarrow \sum_{i=1}^{2} SW_i$
17: $SB \leftarrow \sum_{i=1}^{2} \frac{n_i}{|C|}(\mu_i - \mu)(\mu_i - \mu)^T$
18: $\Psi_{proj} \leftarrow eig(SB/SW) \cdot \Psi$
19: **return** Ψ_{proj} ▷ Ψ_{proj} is projected space
20: **end procedure**

Algorithm 2 EM optimization on each local region r

procedure EM_Optimization(Φ, i, α)
2: ▷ Φ is phenotype values in a local region, i is light, α is confidence interval
Let Φ_{nor} and Φ_{abn} be normal values and abnormalities in Φ
4: **repeat**
E-step:
6: $[rsd_{nor}, i_{1/2_{nor}}] \leftarrow$ PI_CurveFitting (Φ_{nor}, i) ▷ Eq. 2
$[\mu_{nor}, \sigma_{nor}] \leftarrow$ GMM(rsd_{nor})
8: $[rsd_{min}, rsd_{max}] \leftarrow$ getConfidenceInterval $(\mu_{nor}, \sigma_{nor}, \alpha)$
M-step:
10: $rsd_{abn} \leftarrow$ getResidual($\Phi_{abn}, i_{1/2_{nor}}$) ▷ Eq. 3
$[\Phi_{nor}, \Phi_{abn}] \leftarrow$ UpdateCandidate($rsd_{nor}, rsd_{abn}, rsd_{min}, rsd_{max}$)
12: **until** Φ_{nor} and Φ_{abn} are stable
end procedure

abnormal candidates are correct. Consequently, we separate the abnormal candidates $\{\Phi_{abn}\}$ to temporal checking regions (see Definition 2) and refine abnormality identification in each region.

DEFINITION 2: Temporal Checking Region. *A checking region r consists of at most m normal values flanking the selected abnormal candidates, depending on data availability, denoted as* $\{\Phi_{nor}\}$, *and at most n abnormal candidates such that the last abnormal candidate is constrained to be at most l-timepoints away from the first one, denoted as* $\{\Phi_{abn}\}$.

In Definition 2, m, n and l are user-defined parameters that determine the size of a temporal checking region. A check region has at most $m + l - n$ normal values and at most n abnormities. Note that abnormal candidates can be continuous or discontinuous, and two checking regions may share common normal values.

In the refined process, an expectation-maximization (EM) process is employed to repeatedly optimize the results in each temporal region r. Pseudo-code of the EM process is shown in Algorithm 2. In the E step, using the local normal values $\{\Phi_{nor}\}$ in checking region r as inputs, we regenerate the theoretical values $\{\Phi'\}$ with Equation (2). Then the residuals $\{rsd\}$ for both the abnormal candidates $\{\Phi_{abn}\}$ and the normal values $\{\Phi_{nor}\}$ are regenerated using Equation (3) (Algorithm 2 line 6–8). In the M step, we redefine the abnormal candidate set $\{\Phi_{abn}\}$ with the statistical distribution of the new residual data $\{rsd\}$ according to Definition 1. Specifically, a value falls off the confidence interval threshold of the major distribution of the normal residual values will be moved to $\{\Phi_{abn}\}$; and if an abnormal candidate is within the confidence interval threshold of the major distribution of the normal residual values, it will be labelled as normal and be moved to $\{\Phi_{nor}\}$ (Algorithm 2 line 10 and 11).

The EM process will stop when the label assignment is stable. Figure 3(c–f) shows the iterative process in a checking region. Since checking regions may share common values, the results from different regions may be conflicted. For example, a phenotype value is identified as an abnormality in one region but is considered a normal value in another region. To solve conflicts and consequently improve performance, we employ an information sharing process in the end of the EM process to broadcast all the local results to all the checking regions. If conflict exists, voting results will be used to redefine abnormal candidates in the selected feature space (Step 2), and the EM process will rerun on the new checking regions. The process will repeat till the results converge. Figure 3g and h demonstrate that all the abnormalities are identified.

3.3 Related Works

We introduce the GMM and the LDA used in Section 3.2 as follows.

3.3.1 Gaussian mixture model

A GMM is a parametric probability density function represented as a weighted sum of Gaussian component densities (Reynolds, 2009). GMMs are commonly used as a parametric model of the probability distribution of continuous features (Reynolds, 2009). The probability density function is given by the equation:

$$p(x|\lambda) = \sum_{i=1}^{M} \omega_i g(x|\mu_i, \Sigma_i) \quad (4)$$

where x is a D-dimensional continuous-valued vector, ω_i, $i = 1, \ldots, M$, are the mixture weights, and $g(x|\mu_i, \Sigma_i), i = 1, \ldots, M$ are the component Gaussian densities. Each component density is a D-variate Gaussian function of the form:

$$g(x|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1}(x - \mu_i)\right\} \quad (5)$$

with μ_i be the mean vector and Σ_i be the covariance matrix $(i = 1, \ldots, M)$. The mixture weights satisfy the constraint that $\sum_{i=1}^{M} \omega_i = 1$. GMM parameters are estimated from training data using the maximum likelihood parameter estimation or maximum a posteriori estimation (Reynolds, 2009). In this article, residuals are 1D scalar data, we use μ_i and σ_i to represent the mean and variance of residuals.

3.3.2 LDA for feature selection

LDA is a method used in statistics, pattern recognition and machine learning to find a linear combination of features, which

characterizes or separates two or more classes of objects or events, such that the inter-class variance is maximized and the intra-class variance is minimized (Webb, 2002). The resulting combination may be used as a linear classifier, or more commonly, for dimensionality reduction before later classification. In this article, we seek combination of features, with which normal values (one class) are centred around one area, while abnormalities (another class) are centred around a distinctively separated area.

Suppose there are C classes, and each class has n_i points, mean μ_i and intra-class variance Σ_i. Then the inter-class variance may be defined by the sample covariance of the class means:

$$\mathrm{SB} = \sum_{i=1}^{C} \frac{n_i}{|C|}(\mu_i - \mu)(\mu_i - \mu)^T \tag{6}$$

and the intra-class variance of whole dataset is $\mathrm{SW} = \sum_{i=1}^{C} \mathrm{SW}_i$ (McLachlan, 2004). The class separation in a direction $\vec{\omega}$ in this case will be given by:

$$S = \frac{\vec{\omega}^T \mathrm{SB} \vec{\omega}^T}{\vec{\omega}^T \mathrm{SW} \vec{\omega}^T} \tag{7}$$

The objective function is to maximize S and it can be shown that when $\vec{\omega}$ is the eigenvector of $\mathrm{SW}^{-1}\mathrm{SB}$, S will have maximized value corresponding to eigenvalue (Rao, 1948).

4 Experiment

We compared dynamic filter on both real and synthetic datasets with two widely used data cleaning algorithms: (i) a statistical approach that classifies abnormalities based on standard variance (Maletic and Marcus, 2000) and (ii) Hampel filter that identifies abnormalities based on digress from median of trends (Pearson, 2002, 2005). Note that all the three methods were applied on the same phenotype residual data for a fair comparison.

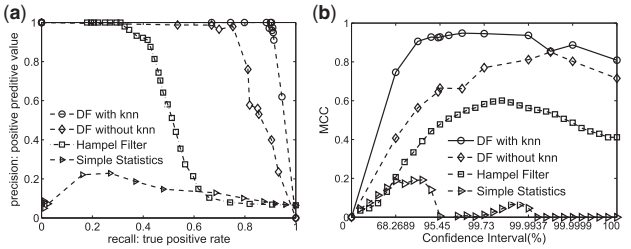
For performance evaluation, we used both the precision-recall curve and the Matthews correlation coefficient (MCC) (Baldi *et al.*, 2000). The MCC that can appropriately represent a confusion matrix is computed with:

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \tag{8}$$

4.1 Real phenotype dataset

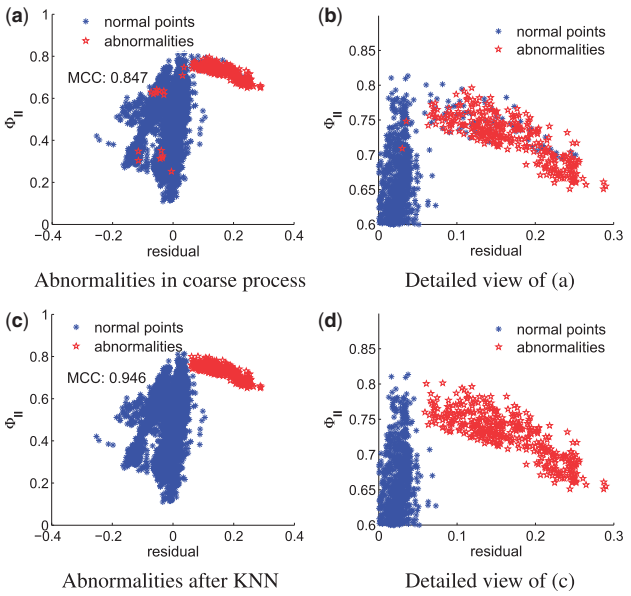
We first tested the performance of dynamic filter using the plant photosynthetic phenotype data consisting of 106 *Arabidopsis thaliana* plants (confirmed T-DNA insertion mutants and wild types) sampled at 64 time points under dynamic light conditions (Ajjawi *et al.*, 2010; Alonso *et al.*, 2003). The 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 (Kramer and Evans, 2011). Experts went through the data and manually marked the ground truth of abnormalities, and found the error rate is 6.5%.

The experimental results shown in Figure 4a indicated that dynamic filter is significantly better than the other two approaches in the precision-recall curve. Specifically, dynamic filter yields Area Under Curve (AUC) as high as 0.964, higher than the AUC of simple statistics and Hampel filter (0.147 and 0.543, respectively). Figure 4b shows our model is also significantly better according to MCC. Furthermore, it shows that dynamic filter is insensitive to the



Precision-Recall curves MCC w.r.t. confidence interval threshold

Fig. 4. Performance evaluation of precision-recall and MCC on real dataset. DF represents dynamic filter. (**a**) Precision-recall curves. (**b**) MCC w.r.t. confidence interval threshold



Abnormalities in coarse process Detailed view of (a)

Abnormalities after KNN Detailed view of (c)

Fig. 5. Performance improvement by applying the KNN refinement process. (**a**) Abnormalities in coarse process. (**b**) Detailed view of (a). (**c**) Abnormalities after KNN. (**d**) Detailed view of (c)

selection of the confidence interval threshold, which is distinctly different from the other algorithms that rely on well-picked parameters.

Note that the AUC of dynamic filter without KNN is 0.862 (Fig. 4a), implying that KNN refinement (Step 2) is a key component of dynamic filter. Specifically, Figure 5 shows how KNN refinement improved the performance of data cleansing. On the Φ_{II} versus residual plot shown in Figure 5a (detailed visualization on Figure 5b), some isolated normal values are misclassified as abnormalities, and certain abnormalities misclassified as normal values. Clearly, these values do not conform with the most nearby values. By applying KNN refinement, this misclassification is effectively corrected (Fig. 5c and d).

We systematically tested the performance of the different components of dynamic filter. Figure 6 shows the performance improvement by comparing dynamic filter with a model without KNN refinement (v5), iteration of EM (v4), consensus on all regions (v3), reassignment of normal values and abnormalities in EM (v2) or even without the whole refined process (v1). It implies that the refined process, especially the KNN and EM refinement, is the key of performance improvement.

Figures 7 and 8 show case studies on the real data. In Figure 7, the experiment was run on a wild-type reference plant, Arabidopsis Col-0.

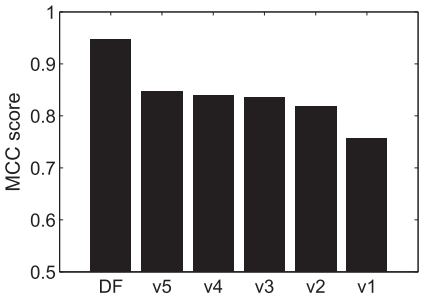


Fig. 6. Performance comparison. Each version corresponds to a different version of DF without: KNN refinement (v5), iteration of EM (v4), consensus on regions (v3), reassignment of normal/abnormal labels in EM (v2), or the whole refined process (v1)

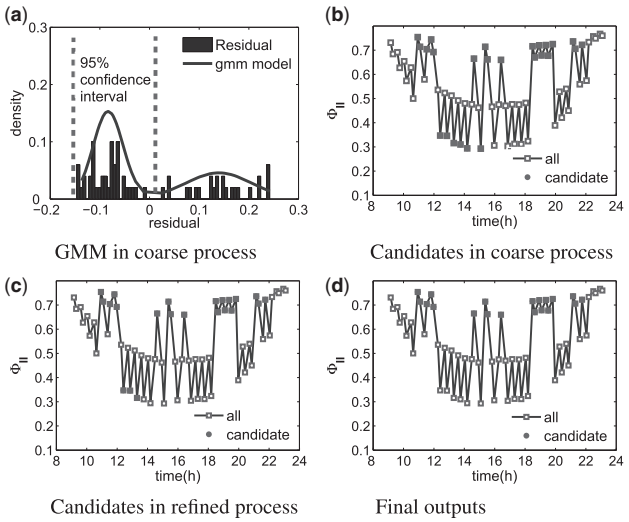


Fig. 7. A case study on the real data shows that DF correctly identifies all the abnormalities. (**a**) GMM in coarse process. (**b**) Candidates in coarse process. (**c**) Candidates in refined process. (**d**) Final outputs

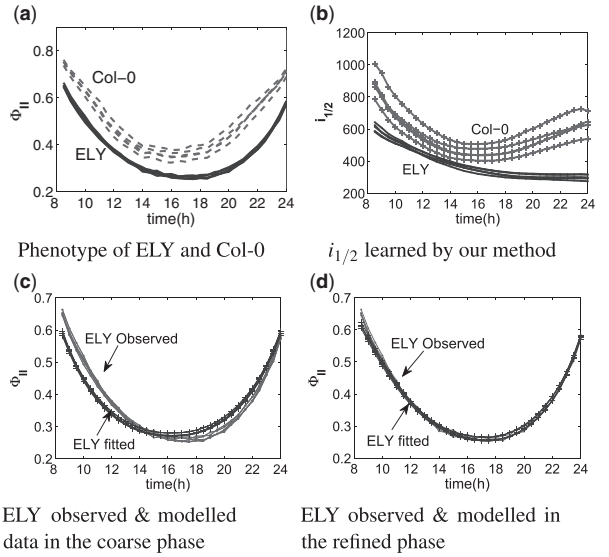


Fig. 8. A case study on the real data shows that DF identifies true biological discoveries under the diurnal light condition. Lines with the same marker represent biological replicates. (**a**) Phenotype of ELY and Col-0. (**b**) $i_{1/2}$ learned by our method. (**c**) ELY observed & modelled data in the coarse phase. (**d**) ELY observed & modelled in the refined phase

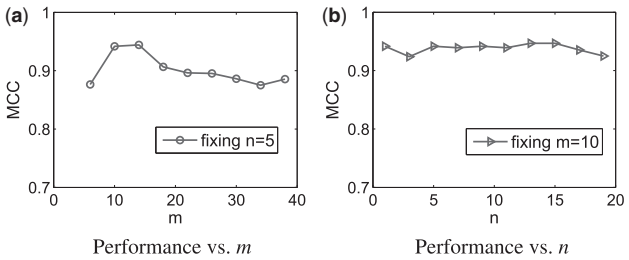


Fig. 9. Performance test on temporal checking region size. (**a**) Fixing max number of abnormalities and varying max number of normal values; (**b**) Fixing max number of normal values and varying max number of abnormalities. (a) Performance versus m. (b) Performance versus n

In the coarse process, the residual analysis was applied to identify the abnormal candidates (Fig. 7a and solid points in Fig. 7b). Clearly, six solid points on the bottom were incorrectly labelled as abnormalities, which were gradually corrected in the refined process (Fig. 7c and d). Figure 8a shows a true biological discovery on the real data. Our screen revealed accession ELY exhibiting photosynthetic characteristics markedly different from the reference (Col-0). It would however be labelled as abnormal and subsequently deleted by the existing outlier-detection based data cleaning methods, resulting in over-clean problem. Dynamic filter identifies ELY correctly and suggests that the differences in its quantum yield are caused by the monotone decrease of $i_{1/2}$ regardless the change of sunlight (see Fig. 8b). The non-negligible deviation between the observed values and the theoretical curve learned from the coarse phase of dynamic filter (see Fig. 8c) implies the theoretical model is simple compared with the real-world situation. Instead of directly use the PI curve to infer abnormalities, we optimize the fitting results in the refined phase of dynamic filter, resulting in almost perfect match between the observed values and the theoretical curve (see Fig. 8d).

Furthermore, we varied the size of the temporal checking region and compared the performance in Figure 9. The results in Figure 9a reveal that dynamic filter achieves the best performance when m is between 10 and 15. This number allows enough training data for the refinement process, meanwhile avoiding NPQ variation over long time interval. Figure 9b shows that performance of dynamic filter is relatively stable against max number of abnormalities n, implying that robustness of dynamic filter is high.

4.2 Synthetic dataset

Since the true biological discoveries in the real data are unknown, we further tested dynamic filter on serials of synthetic datasets. The synthetic datasets were generated by varying four parameters systematically: lights and $i_{1/2}$ being smoothly or abruptly changed, abnormalities being continuously or discontinuously distributed, and error ratio being low or high. Furthermore, we added variations representing abnormalities and biological discoveries (different $i_{1/2}$ values) in the synthetic datasets. In total, 63 kinds of synthetic datasets in nine groups were generated, and for each kind of synthetic data, we repeatedly generated 100 datasets.

Figure 10 shows the robustness of dynamic filter on different synthetic datasets generated under nine different settings. The performance is evaluated using MCC on both abnormalities and on biological outliers. Each figure represents synthetic data generated under different settings (see details in supplementary section S2). Each point in Figure 10 represents a MCC score of biological

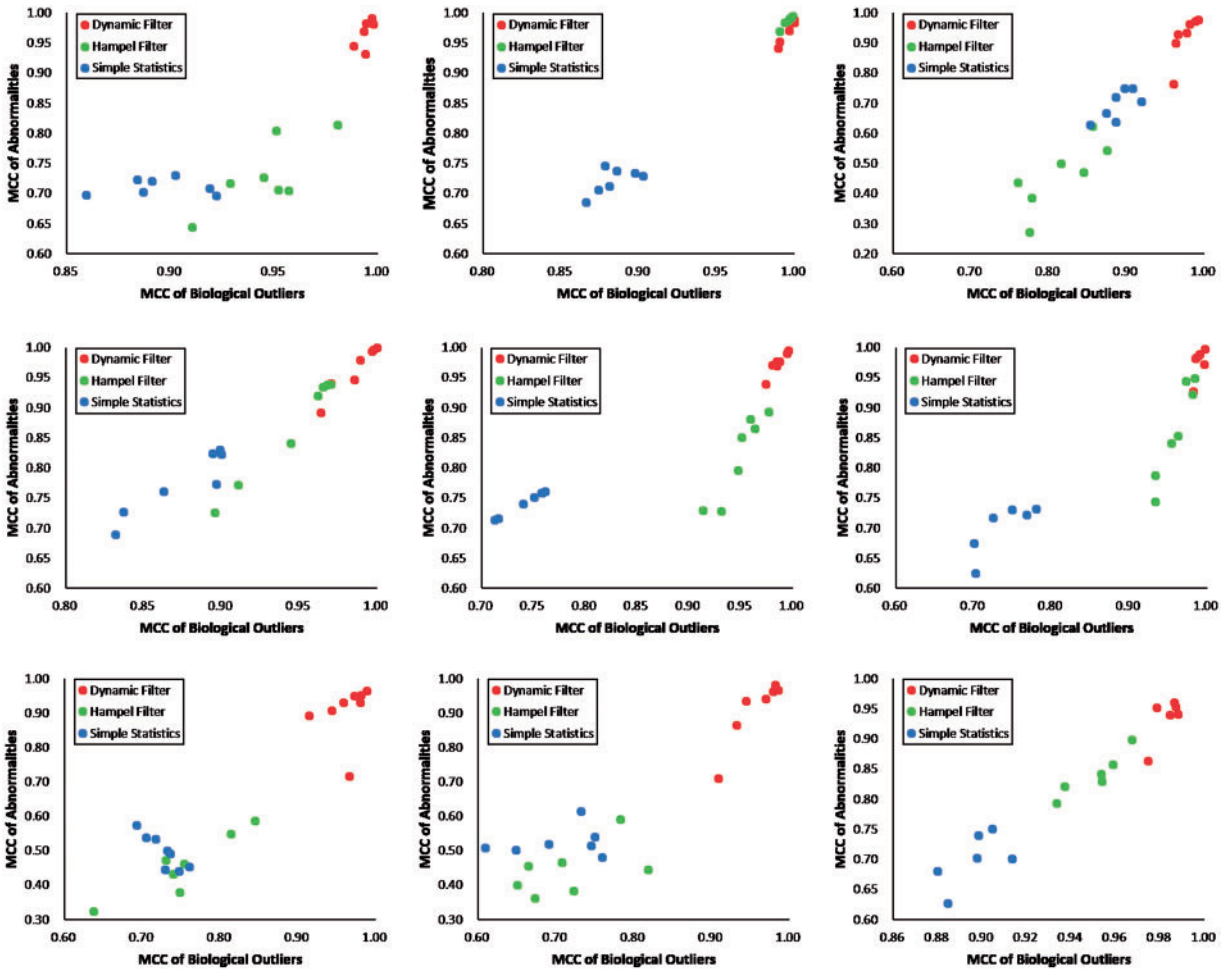


Fig. 10. The MCC of biological discoveries and abnormalities on synthetic data

discovery identification at x-axis and a MCC score of abnormality identification at y-axis. The highest possible value is (1.0,1.0). The experimental results show that dynamic filter (red circle) is better than the other two methods in almost all the synthetic datasets. This is because dynamic filter can identify and remove abnormalities while reserving biological discoveries (see supplementary Tables S1 and S2 for performance comparison on MCC and true positive rate, respectively).

5 Conclusion

With an aim towards identifying targets for improving energy yield, advanced technologies in high-throughput plant photosynthetic phenotyping have been developed (Cruz *et al.*, 2014; Houle *et al.*, 2010). These systems can be used to quantify photosynthetic behaviour in genetically diverse populations and to draw relationships among genotype, phenotype and biological function, leading to a better understanding of the underlying mechanisms that control the photosynthetic properties under various environmental conditions (Fiorani and Schurr, 2013; Rascher *et al.*, 2011). As a consequence of the long-time high-throughput plant phenotyping, the scale of plant phenomics data grows exponentially. However, the quality of phenotype data may be skewed by sources of noise that are difficult to remove in the data collection step.

The purpose of plant phenotyping is to discover phenotype values that are significantly different from a reference. But phenotype values leading to biological discoveries may be obscured by abnormal values caused by errors during detection. To ensure high data quality, effective data cleaning should be considered a primary task. However, since advanced data cleaning algorithms are primarily based on indiscriminate outlier detection, they may remove both abnormalities and biological discoveries not separable in the data distribution.

We have developed a new coarse-to-refined model called dynamic filter to effectively identify both abnormalities and biological discoveries by adopting a widely used photosynthetic model. Specifically, dynamic filter is a residual analysis approach by dynamically tracing statistical distributions of all samples rather than individuals, and incorporating EM for performance optimization in refined checking regions.

We note that certain events, such as transient changes in growth environment, could introduce signals similar to growth lighting malfunction, which could be wrongly labelled as abnormalities by dynamic filter. Therefore, instead of automatically deleting all the predicted abnormalities, we send all of them to domain experts for confirmation. Meanwhile, all raw data are kept for any rollback operation.

Experimental results show that our model is significantly better than the existing data cleaning tools on both real-phenomics data and synthetic data. Dynamic filter may have a wide impact because of the rapid increase of large-scale phenotyping technologies. It should be noted that although we used a photosynthesis-specific

curve, the model itself is independent of actual biological constraints. In principle, our approach can be used to clean data for any number of phenotypes as long as suitable theoretical curves can be derived for their behaviour. Implementation for new use cases would involve substituting the appropriate theoretical curve into the program, calculating the residuals of fits to the datasets and optimizing the fitting procedure as described in Figure 2.

Funding

This research was supported by Chemical Sciences, Geosciences and Biosciences Division, Office of Basic Energy Sciences, Office of Science, U.S. Department of Energy [award number DE-FG02-91ER20021], and by Center For Advanced Camelina Oils, Advanced Research Projects Agency - Energy, U.S. Department of Energy [award number DE-AR0000202, sub 21018-MI].

Conflict of Interest: none declared.

References

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

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

Altman,N.S. (1992) An introduction to kernel and nearest-neighbor nonparametric regression. *Am. Stat.*, **46**, 175–185.

Baldi,P. *et al.* (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**, 412–424.

Bonner,J. (1962) The upper limit of crop yield this classical problem may be analyzed as one of the photosynthetic efficiency of plants in arrays. *Science*, **137**, 11–15.

Chu,F. *et al.* (2005) Data cleaning using belief propagation. In: *Proceedings of the 2Nd International Workshop on Information Quality in Information Systems*, ACM, New York, NY, USA, pp. 99–104.

Chu,X. *et al.* (2013) Holistic data cleaning: putting violations into context. *ICDE*, 458–469.

Cruz,J.A. *et al.* (2014) Dynamic environmental photosynthetic imaging (depi) reveals emergent phenotypes related to the environmental responses of photosynthesis. *Nat. Biotech.*, in revision.

Ebaid,A. *et al.* (2013) Nadeef: a generalized data cleaning system. *VLDB Endowment*, **6**, 1218–1221.

Fiorani,F. and Schurr,U. (2013) Future scenarios for plant phenotyping. *Annu. Rev. Plant Biol.*, **64**, 267–291.

Govindjee *et al.* (2005) *Discoveries in Photosynthesis*. Springer.

Herbert,K.G. *et al.* (2004) Bio-ajax: an extensible framework for biological data cleaning. *ACM SIGMOD Record*, **33**, 51–57.

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

Jassby,A.D. and Platt,T. (1976) Mathematical formulation of the relationship between photosynthesis and light for phytoplankton. *Am. Soc. Limnol. Oceanogr.*, **21**, 540–547.

Koh,J.L.Y. *et al.* (2007) Correlation-based detection of attribute outliers. In: Kotagiri,R. Radha Krishna,P. *et al.* (eds), *Advances in Databases: Concepts, Systems and Applications*, Springer Berlin Heidelberg, pp. 164–175.

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

Lambers,H. *et al.* (2008) Response of Photosynthesis to Light. In: *Plant Physiological Ecology*. Springer, New York, pp. 26–47.

Li,X.P. *et al.* (2000) A pigment-binding protein essential for regulation of photosynthetic light harvesting. *Nature*, **403**, 391–395.

Lobo,F.de.A. *et al.* (2013) Fitting net photosynthetic light-response curves with microsoft excel a critical look at the models. *Photosynthetica*, **51**, 445–456.

Long,S.P. and Hällgren,J.E. (1993) Measurement of CO_2 assimilation by plants in the field and the laboratory. In: D.O. Hall *et al.* (ed.) *Photosynthesis and Production in a Changing Environment*. Chapman and Hall, pp. 129–167.

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

Maletic,J.I. and Marcus,A. (2000) Data cleansing: Beyond integrity analysis. In: *IQ*, pp. 200–209.

Mayfield,C. *et al.* (2010) ERACER: a database approach for statistical inference and data cleaning. In: *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, ACM, New York, NY, USA, pp. 75–86.

McLachlan,G. (2004) *Discriminant analysis and statistical pattern recognition*. **Vol. 544**. John Wiley & Sons, Hoboken, New Jersey.

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

Muller,H. and Freytag,J.C. (2005) *Problems, methods, and challenges in comprehensive data cleansing*. Professoren des Inst. Fur Informatik, Humboldt University Berlin.

Muller,P. *et al.* (2001) Non-photochemical quenching. A response to excess light energy. *Plant Phisiol.*, **125**, 1558–1566.

Ögren,E. and Evans J.R. (1993) Photosynthetic light-response curves. *Planta*, **189**, 182–190.

Pearson,R.K. (2002) Outliers in process modeling and identification. *IEEE T. Contr. Syst T.*, **10**, 55–63.

Pearson,R.K. (2005) *Mining imperfect data: Dealing with contamination and incomplete records*. Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania.

Rao,C.R. (1948) The utilization of multiple measurements in problems of biological classification. *J. R. Stat. Soc.*, **10**, 159–203.

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

Reynolds,D. (2009) Gaussian mixture models. In: Li, S. Z. Jain, A. *Encyclopedia of Biometrics*, Springer US, pp. 659–663.

Seber,G. and Wild,C. (2003) *Nonlinear Regression*. Wiley-Interscience, Hoboken, New Jersey.

Shanahan,M. (2005) Perception as abduction: turning sensor data into meaningful representation. *Cognitive Sci.*, **29**, 103–134.

Sohn,H. *et al.* (2005) Structural damage classification using extreme value statistics. *J. Dyn. Syst-T Asme.*, **127**, 125–132.

Subramaniam,S. *et al.* (2006) Online outlier detection in sensor data using non-parametric models. In: *Proceedings of the 32Nd International Conference on Very Large Data Bases (VLDB)*, VLDB Endowmen, Seoul, Korea, pp. 187–198.

Takizawa,K. *et al.* (2007) The thylakoid proton motive force in vivo. quantitative, non-invasive probes, energetics, and regulatory consequences of light-induced pmf. *Biochim. Biophys.*, **1767**, 1233–1244.

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

Von Caemmerer,S. and Farquhar,G.D. (1981) Some relationships between the biochemistry of photosynthesis and the gas exchange of leaves. *Planta*, **153**, 376–387.

Webb,A.R. and Copsey,K.D. (2011) Fishers Criterion – Linear Discriminant Analysis. In: Webb,A.R. *et al.* (eds), *Statistical Pattern Recognition*. Wiley, Hoboken, New Jersey.

Zeinalov,Y. (2005) Mechanisms of photosynthetic oxygen evolution and fundamental hypotheses of photosynthesis. In: Pessarakli,M. (ed.) *Handbook of Photosynthesis*, CRC Press, Boca Raton, Florida.

Zhu,X. *et al.* (2010) Improving photosynthetic efficiency for greater yield. *Annu. Rev. Plant Biol.*, **61**, 235–261.