Bioinformatics, 32(1), 2016, 67–76
doi: 10.1093/bioinformatics/btv515
Advance Access Publication Date: 3 September 2015
Original Paper

Systems biology

Inter-functional analysis of high-throughput phenotype data by non-parametric clustering and its application to photosynthesis

Qiaozi Gao[1,†], Elisabeth Ostendorf[2,†], Jeffrey A. Cruz [2], Rong Jin[1,*], 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, East Lansing, MI 48824, USA

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Received on May 12, 2015; revised on August 25, 2015; accepted on August 25, 2015

Abstract

Motivation: Phenomics is the study of the properties and behaviors of organisms (i.e. their phenotypes) on a high-throughput scale. New computational tools are needed to analyze complex phenomics data, which consists of multiple traits/behaviors that interact with each other and are dependent on external factors, such as genotype and environmental conditions, in a way that has not been well studied.

Results: We deployed an efficient framework for partitioning complex and high dimensional phenotype data into distinct functional groups. To achieve this, we represented measured phenotype data from each genotype as a cloud-of-points, and developed a novel non-parametric clustering algorithm to cluster all the genotypes. When compared with conventional clustering approaches, the new method is advantageous in that it makes no assumption about the parametric form of the underlying data distribution and is thus particularly suitable for phenotype data analysis. We demonstrated the utility of the new clustering technique by distinguishing novel phenotypic patterns in both synthetic data and a high-throughput plant photosynthetic phenotype dataset. We biologically verified the clustering results using four Arabidopsis chloroplast mutant lines.

Availability and implementation: Software is available at www.msu.edu/~jinchen/NPM.

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

Contact: jinchen@msu.edu, kramerd8@cns.msu.edu or rongjin@cse.msu.edu

1 Introduction

The plummeting cost of genome sequencing has ushered genomics into a new era. However, using genomics data alone is insufficient to answer most major questions in biology, because genomics data cannot explain how genes control complex behaviors (Cobb *et al.*, 2013; Shendure and Ji, 2008; Wagner and Zhang 2013). The next wave in biology will be to phenotype organisms and then integrate genomes with phenomes to identify genes that may govern phenotype and responses to the varying environment (Butte and Kohane, 2006). Especially, progress in phenotyping technologies is required to accelerate genetic mapping and gene discovery in genome-wide association studies for dissecting the genetic architecture of important traits underlying phenotypes (Yang *et al.*, 2014).

It thus is important to develop computational methods to examine potentially interacting phenotypes (i.e. measurable traits/behaviors) of a genetically diverse population for a given species in order to identify meaningful differences in biological function. The database PhenomicDB has been built to collect phenotypes of multiple species in free-text format with the goal to gain insights into the genetic origin of diseases (Groth *et al.*, 2007). Phenoclustering was then

proposed to perform meta-analysis on PhenomicDB using advanced natural language processing methods, resulting in clusters of genes by the similarity of their phenotypes that can be further explored for genotype–phenotype interaction analysis (Groth *et al.*, 2010, 2011). However, since high-throughput phenotyping quantitatively assesses multiple phenotypes of many traits simultaneously, it often generates a large amount of data rather than qualitative descriptions. New approaches to analyze quantitative phenomics data are needed. One application is in bioenergy research, where computational techniques have been developed to detect, collect and study photosynthesis and related traits, such as growth, under non-laboratory conditions and in high throughput (Cruz *et al.*, 2015; Tessmer *et al.*, 2013; Xu *et al.*, 2015).

In general, photosynthesis phenotypes can be altered in two ways, either through the changes of peripheral processes of the photosynthesis regulatory network or the modification of central components of the network (Diner and Rappaport, 2002). Changes in peripheral processes tend to preserve regulatory relationships within the network. In contrast, altering the central components of photosynthesis is likely to perturb the relationships between key regulatory processes, leading to a different correlation among photosynthesis phenotypes. In this study, we focus on data analysis for the second case, to which we refer as *inter-functional* analysis for photosynthesis phenotypes data.

By taking advantage of the systematic knockout of genes encoding chloroplast-targeted proteins in *Arabidopsis thaliana*, a model plant, we have conducted a large-scale phenotype screen on the single knockouts and the references to understand the underlying molecular functions of the knockout genes in plant photosynthesis (Cruz *et al.*, 2015; Lu *et al.*, 2011). In our experiment, plants with single gene knockouts have been profiled by non-invasively measuring three key photosynthetic parameters, i.e. ϕ_{II}, q_E and q_I (that reflect photosynthesis efficiency in photosystem II, active energy dissipation, and photoinhibition, respectively), in response to fluctuating light conditions, resulting in a large volume of plant phenotyping data that allow us to determine the downstream effects of the knockout genes on photosynthesis in dynamic conditions (Kramer and Evans, 2011). More specifically, multiple fluorescence images (Fig. 1a) are taken for every plant as the light intensity is varied over time and the photosynthesis parameters (Fig. 1b) are derived from each image. As a result, each plant is characterized by a sequence of three-dimensional vectors with each dimension corresponding to a different photosynthesis parameter. The objective of this study is to automatically identify gene groups with their photosynthesis phenotypes significantly different from reference based on the collected data.

A natural first step is to examine the extremes, i.e. mutant lines that exhibit significantly different photosynthesis characters from the wild type in all of the three photosynthesis parameters. Although this simple technique has been shown to be effective for identifying potential biomarkers and drug targets, it is not an effective approach for exploiting the full potential of experimental results (Eisen *et al.*, 1998). In this work, we take a holistic approach for inter-functional photosynthesis data analysis. We develop an efficient workflow to effectively identify novel phenotypes based on mutant lines sharing similar measurements of photosynthesis parameters. The identified gene groups and their associated phenotypes are used to derive insight into the functions of and interactions between related genes. Based on existing annotations and literature for the well-studied genes, we can associate biological processes with each phenotype pattern, which in turn can be used to reveal the functions for the poorly characterized genes (Jaqaman *et al.*, 2007).

Although numerous clustering algorithms have been developed, most of them require vector representation of data and thus cannot be applied directly to our problem where each plant is represented by a sequence of vectors (Bar-Joseph, 2004; McNicholas and Murphy, 2010; Nascimento *et al.*, 2012; Qian *et al.*, 2001; Schliep *et al.*, 2003). One simple way to handle the challenge is to run standard clustering algorithms (e.g. *k*-means and hierarchical clustering) against all the vectors measured for different plants, and determine the cluster member for each plant by taking the major vote among the measured vectors. Another approach is to concatenate the sequence of vectors measured for each plant into a single vector, and apply standard clustering algorithms to the derived vectors. The main problem with these two approaches is that they are unable to effectively explore the correlation among different photosynthesis parameters revealed by the set of vectors measured for each plant. This is particularly important for photosynthesis data analysis because it is the correlation between photosynthesis parameters that helps understand the regulatory role of the mutated genes on the photosynthesis processes.

We propose to address the above challenges of phenotype data analysis by developing a novel clustering technique named **Non-parametric modeling**, or **NPM** for short. We introduce a

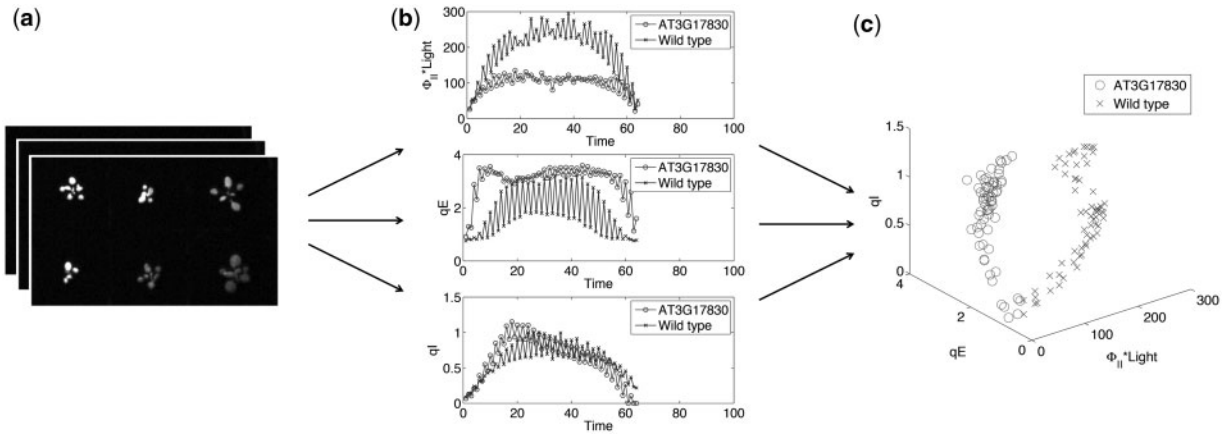

Fig. 1. The process of extracting data from photosynthesis phenotype experiment. **(a)** a series of fluorescence images, **(b)** the derived photosynthesis parameters ($\phi_{II} \times$ *light*, q_E, q_I) for two genotypes, i.e. mutant line AT3G17830 and the wild type (used as reference) under dynamic environmental conditions, **(c)** the photosynthesis parameters of the same plants in 3D space. $\phi_{II} \times$ *light* refers to a more advanced photosynthesis parameter named Linear Electron Flux

cloud-of-points representation for each mutant line, with each point corresponding to a vector in the sequential phenotype measurements taken for the mutant line, where the correlation among photosynthesis parameters is captured by the 'shapes' of clouds. Based on the theory of kernel density estimation (KDE), we then apply a non-parametric clustering technique to group plants into the same cluster if their clouds share similar shapes (Cybakov, 2009).

When compared with the existing clustering algorithms, the key advantage of NPM is that it does not make any assumption about the underlying data distribution and thus is particularly suitable for phenotype data analysis. Our empirical study shows promising results of the proposed clustering technique on both synthetic and real photosynthesis datasets. We have biologically verified the clustering results using four Arabidopsis mutant lines for nuclear encoded chloroplast proteins. The experimental results demonstrate that NPM is able to generate testable hypotheses, which lead to new biological discoveries. We emphasize that although our empirical study is limited to photosynthesis data, the non-parametric clustering algorithm developed in this work is general and can be applied to other domains (e.g. gene expression data analysis).

2 Related work

Clustering algorithms have been widely used in biological data analysis. One well known example is gene expression clustering (Eisen *et al.*, 1998; Herwig, 1999; Tamayo *et al.*, 1999). Numerous clustering algorithms have been designed to exploit temporal correlations and trends in gene expression time series (Bar-Joseph 2004; Bar-Joseph *et al.*, 2003; Costa *et al.*, 2004; Qian *et al.*, 2001; McNicholas and Murphy, 2010; Nascimento *et al.*, 2012; Ramoni *et al.*, 2002; Schliep *et al.*, 2003; Sivriver *et al.*, 2011). Clustering algorithms have also been successfully applied to biological network analysis, which aims to discover and identify gene functional groups (Cingovska *et al.*, 2012; Wang *et al.*, 2013; Wu *et al.*, 2002), and to sequence analysis [e.g. sequence alignment (Corpet, 1988), motif detection (Lones and Tyrrell, 2007) and protein/DNA sequence clustering (Enright and Ouzounis, 2000; Li and Godzik, 2006)]. Most previous studies of clustering algorithms in biological data analysis are based on parametric models (Bakar and Watada, 2008; Ben-Dor *et al.*, 1999). In contrast, the proposed method adopts non-parametric density estimation for data clustering that does not make parametric assumptions about the underlying distribution.

A well-known non-parametric clustering method is mean-shift (Carreira-Perpiñán, 2006; Comaniciu, 2002), which is widely used in computer vision and image processing. Other examples of non-parametric clustering are the scale-space method (Roberts, 1997; Wilson and Spann, 1990) and spectral clustering (Von Luxburg, 2007). The scale-space clustering method was pioneered in Wilson and Spann (1990), and further studied in Roberts (1997) using density estimation. Spectral clustering is mostly based on the theory of manifold learning (Ma and Fu, 2012), and has been widely used for image segmentation. However, since most non-parametric clustering algorithms require vector representation of data, they cannot be applied directly to our problem.

The theoretical foundation of the proposed work is based on KDE (Parzen, 1962; Rosenblatt, 1956). KDE was first introduced in the 1950s (Rosenblatt, 1956). The consistency of KDE was shown initially in Parzen (1962) and its convergence rate for finite samples was later proved (Nadaraya, 1965; Stute, 1982). A more complete literature review for KDE can be found in Cybakov (2009).

3 Non-parametric clustering for phenotypic data analysis

In this work, we develop a novel clustering algorithm NPM for inter-functional phenotype data analysis. Below, we will present phenotype data representation and framework of non-parametric clustering, and then describe an efficient algorithm for solving the related optimization problem.

3.1 Data representation

The center of NPM is a cloud-of-points representation. Let $\mathcal{M}_1, \ldots, \mathcal{M}_m$ be a collection of m genotypes. Let p be the number of photosynthesis parameters. For each genotype $\mathcal{M}_i$, we denote by $\mathbf{x}_1^i, \ldots, \mathbf{x}_{n_i}^i$ the sequence of multi-dimensional vectors measured for $\mathcal{M}_i$, where n_i is the number of valid measurements for $\mathcal{M}_i$ and each $\mathbf{x}_j^i \in \mathbb{R}^p$ includes the measurements of p photosynthesis parameters derived from the jth measurement for $\mathcal{M}_i$.

Because the focus of this study is to examine the dependence among different photosynthesis parameters, we will ignore the sequential order among vectors $\{\mathbf{x}_j^i\}_{j=1}^{n_i}$ and simply characterize the photosynthesis property of genotype $\mathcal{M}_i$ by the set of data points $\mathcal{D}_i = \{\mathbf{x}_1^i, \ldots, \mathbf{x}_{n_i}^i\}$, which we refer to as *cloud-of-points representation*. Figure 1c shows examples of the cloud-of-points representation for the sequence of ϕ_{II}, q_{E} and q_{I} measurements that are given in Figure 1b.

3.2 A framework for non-parametric clustering

Following the standard framework of mixture models (McLachlan and Peel, 2004), we assume that there are K different underlying distributions, where each distribution is introduced to capture a different 'shape' of the cloud-of-points representation, and all the data points observed in the cloud-of-points representation are drawn independently from one of the K distributions.

More specifically, let $f_1(\cdot), \ldots, f_K(\cdot)$ be the density functions for the K underlying distributions, and let $\mathbf{P} = (P_1, \ldots, P_K)$ be the prior probabilities for choosing each distribution. Then, for mutant line $\mathcal{M}_i$, the likelihood of observing the cloud-of-points representation $\mathcal{D}_i = \{\mathbf{x}_j^i\}_{j=1}^{n_i}$ is then given by

$$\Pr(\mathcal{D}_i) = \sum_{j=1}^{K} P_j \Pr(\mathcal{D}_i | f_j) = \sum_{j=1}^{K} P_j \prod_{k=1}^{n_i} f_j(\mathbf{x}_k^i). \tag{1}$$

Following the framework of maximum likelihood estimation (Scholz, 1985), we find the optimal density functions $\{f_j(\cdot)\}_{j=1}^{K}$ by solving the optimization problem

$$\max_{f_1, \ldots, f_K, \mathbf{P}} \sum_{j=1}^{m} \log \Pr(\mathcal{D}_i) \tag{2}$$

where $\Pr(\mathcal{D}_i)$ is given in Equation (1).

The main challenge arises from solving the optimization problem in Equation (2), in which the variables to be optimized are *functions* (or vectors of infinite dimension). This is in contrast to most optimization problems where the variables are of finite dimension (Fletcher, 2013). The type of optimization problem in Equation (2) is often referred to as *variational optimization* in the literature of operations research. One simple approach towards variational optimization is to make specific assumptions about the density functions $\{f_j(\cdot)\}_{j=1}^{K}$, a common approach adopted in the study of mixture models (McLachlan and Peel, 2004). In the simple approach, a parametric family of distributions, e.g. the well-known Gaussian mixture model, is specified assuming that each $f_j(\cdot)$ is a Gaussian

distribution. However, this simple approach could be problematic in our study, since the complicated shaped cloud-of-points representation (see Fig. 4) may not fit well to any member in the specified distribution family. Therefore, we adopt the non-parametric density estimation approach, because the key advantage of non-parametric approach is that it avoids the parametric assumption of the underlying distribution (Cybakov, 2009).

3.3 Efficient computational algorithm

We propose to solve the optimization problem in Equation (2) by exploring the theory of KDE. Our approach is based on the Nadaraya-Watson method for density estimation (Cybakov, 2009). More specifically, let $\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_\ell$ be the ℓ landmark points, where landmark points are randomly sampled from the collection of data points for all mutant lines. Using the landmark points, we approximate each density function f_j as:

$$f_j(\mathbf{x}) = \frac{1}{\ell} \times \sum_{q=1}^{\ell} \alpha_j^q \kappa(\mathbf{x} - \hat{\mathbf{x}}_q) \tag{3}$$

where $\boldsymbol{\alpha}_j = (\alpha_j^1, \ldots, \alpha_j^\ell)$ are parameters to be determined for f_j, $\kappa(\mathbf{x}) = (\pi h^2)^{-p/2} \exp(-|\mathbf{x}|^2/h^2)$ is a multi-variate Gaussian distribution and h is the kernel bandwidth. Since f_j is a density function, we have $\alpha_j^q \geq 0$, $q \in [1, \ell]$ and $\sum_{q=1}^{\ell} \alpha_j^q = \boldsymbol{\alpha}_j^\top \mathbf{1} = 1$, where $\mathbf{1}$ is a vector of all ones. Using the approximation in (3), the likelihood function $\Pr(\mathcal{D}_i)$ is rewritten as:

$$\Pr(\mathcal{D}_i) = \sum_{j=1}^{K} P_j \prod_{k=1}^{n_i} \left(\frac{1}{\ell} \sum_{q=1}^{\ell} \alpha_j^q \kappa(\mathbf{x}_k^i - \hat{\mathbf{x}}_q) \right). \tag{4}$$

We can approximate the variational optimization problem in Equation (2) into the following normal optimization problem:

$$\max_{\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_K, \mathbf{P}} \mathcal{L}(\{\boldsymbol{\alpha}_j\}_{i=1}^K, \mathbf{P}) = \sum_{i=1}^{m} \log \Pr(\mathcal{D}_i) \tag{5}$$

where $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_K \in \{\boldsymbol{\alpha}_j \in \mathbb{R}_+^p : \alpha^\top \mathbf{1} = 1\}$ and $\Pr(\mathcal{D}_i)$ is given in (4).

We developed an EM algorithm to efficiently optimize the objective function in Equation (5). For each mutant line $\mathcal{M}_i$, we introduce a hidden variable $Z_i \in [K]$ to indicate which cluster (or density function) $\mathcal{M}_i$ belongs to. In the E-step, we estimate the posterior probability $\Pr(Z_i = j|\mathcal{D}_i)$ as:

$$\Pr(Z_i = j|\mathcal{D}_i) \propto P_j \prod_{k=1}^{n_i} \left(\sum_{q=1}^{\ell} \alpha_j^q \kappa(\mathbf{x}_k^i - \hat{\mathbf{x}}_q) \right) \tag{6}$$

In the M-step, we update the prior probabilities $\mathbf{P}$ and parameters $\{\boldsymbol{\alpha}_j\}_{j=1}^K$ by:

$$P_j = \frac{1}{n} \sum_{i=1}^{m} \Pr(Z_i = j|\mathcal{D}_i) \tag{7}$$

$$\alpha_j^q \propto \sum_{i=1}^{n} \Pr(Z_i = j|\mathcal{D}_i) \sum_{k=1}^{n_i} \frac{\tilde{\alpha}_j^q \kappa(\mathbf{x}_k^i - \hat{\mathbf{x}}_q)}{\sum_{q'=1}^{\ell} \tilde{\alpha}_j^{q'} \kappa(\mathbf{x}_k^i - \hat{\mathbf{x}}_{q'})} \tag{8}$$

where $\tilde{\boldsymbol{\alpha}}_1, \ldots, \tilde{\boldsymbol{\alpha}}_K$ are parameters obtained in the previous iteration of EM algorithm.

Algorithm 1 highlights the key steps of our EM algorithm. In line 2–5, data are first normalized to prevent large measurement dominating others. In line 7, the landmark points are sampled. Then, the EM iteration part is in line 9–17. E- and M-step are repeated until the log-likelihood converges. The output are K density functions $f_1, \ldots, f_K$. Each mutant line is assigned to one density function that contributes the largest likelihood to its data points.

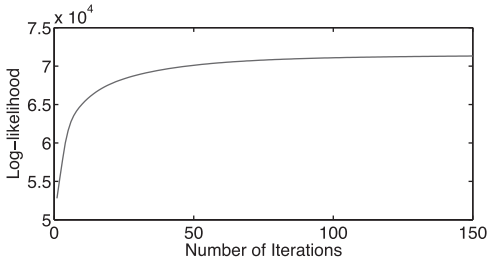

Fig. 2. Increase in log-likelihood against iteration number when running EM algorithm

Figure 2 shows how the log-likelihood function $\mathcal{L}(\{\boldsymbol{\alpha}_i\}_{i=1}^K, \mathbf{P})$ is improved over iterations, when running our EM algorithm on photosynthesis data (see details in Section 4). Because it is well-known that an EM algorithm could be trapped in a bad local optimum, we run Algorithm 1 multiple times and choose the solution with the largest log-likelihood.

Algorithm 1. EM algorithm for non-parametric clustering

Input: Cloud-of-points representation for mutant lines $\mathcal{D}_1, \ldots, \mathcal{D}_m$, the number of photosynthesis parameters p, the kernel bandwidth h, the number of landmark points ℓ, the number of clusters K, and the stop criterion ϵ.

1: *// Normalize data points*
2: **for** $k = 1, \ldots, p$ **do**
3: Compute the minimum and maximum values for the kth photosynthesis parameter
4: Normalize the kth photosynthesis measures to $[0, 1]$ by subtracting them by minimum value and dividing them by the difference between maximum and minimum value
5: **end for**
6: *// Select landmark points*
7: Randomly sample ℓ data points from the collection of data points for all mutant lines, denoted by $\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_\ell$
8: *// EM algorithm*
9: Initialize $\tilde{\boldsymbol{\alpha}}_1, \ldots, \tilde{\boldsymbol{\alpha}}_K$ with numbers randomly chosen from $[0, 1]$
10: Initialize $\tilde{\mathcal{L}} = -\infty$ and $\Delta\mathcal{L} = +\infty$
11: **while** $\Delta\mathcal{L} > \epsilon$ **do**
12: Update posterior probability $\Pr(Z_j = k|\mathcal{D}_j)$ using (6) for every mutant line $\mathcal{M}_j$ and every cluster k
13: Update the coefficients $\alpha_1, \ldots, \alpha_K$ using (8) and $\mathbf{P}$ using (7)
14: Set $\tilde{\boldsymbol{\alpha}}_j = \boldsymbol{\alpha}_j, j = 1, \ldots, K$
15: Compute the log-likelihood, $\mathcal{L} = \sum_{i=1}^{m} \log \Pr(\mathcal{D}_i|F)$
16: Update $\Delta\mathcal{L} = \mathcal{L} - \tilde{\mathcal{L}}$ and $\tilde{\mathcal{L}} = \mathcal{L}$
17: **end while**

Return $f_1, \ldots, f_K$ with $f_j(\mathbf{x}) = \sum_{q=1}^{\ell} \alpha_j^q \kappa(\mathbf{x} - \hat{\mathbf{x}}_q)/\ell$

One fundamental assumption of NPM is that density function $f_j(\cdot)$ can be approximated by Equation (3). Theoretic study of KDE can easily show that any smooth bounded density function can be well approximated by Equation (3) provided that the number of sampled landmark data points ℓ is sufficiently large. Please refer to Theorem S1 in Supplementary Material. We finally note that the density functions for most well-known distributions (e.g. Gaussian

and Laplacian distributions) are smooth and bounded, and thus assuming density functions to be smooth and bounded is a very minor assumption in real-world applications.

4 Experiments

For performance evaluation, we compared NPM with a series of baselines that adopt two different data representation methods other than the cloud-of-point representation used in NPM. The first method is to use concatenated vectors, i.e. concatenating the set of data vectors measured for each genotype into a single vector, and then applying to the derived vectors three standard clustering algorithms: (i) *k*-means (MacQueen, 1967) (Kmeans), (ii) hierarchical clustering (Johnson, 1967) (Hierarchy) and (iii) using principal component analysis (Jolliffe, 1986) to reduce the dimensionality of the concatenated vectors and then applying *k*-means (PCA + Kmeans), where the reduced dimensionality was set to 3 in our experiments. The second data representation method is to simply use the original vectors: clustering all the vectors measured for different genotypes, and then determining the cluster membership for each genotype by taking the majority vote among the vectors measured for that genotype. Two baseline algorithms were developed based on this method: (i) applying *k*-means to all the measured vectors (Kmeans-split) and (ii) applying Gaussian mixture model to all the measured vectors (GMM-split).

Except for the number of clusters K, there are three parameters to be specified in NPM: the stopping criterion ϵ for EM process, the kernel bandwidth h for KDE and the number of landmark points ℓ. Within a cross-validation dataset (see Section 4.2 for details), a grid search was applied to find the best kernel bandwidth h that resulted in the maximal log-likelihood. A similar approach is applied to determine the number of landmark points ℓ. For parameter ϵ, we simply set $\epsilon = 1$ in our experiment.

4.1 Plant photosynthesis phenotyping experiment

Photosynthesis involves the net movement of electrons generated from the oxidation of water through a series of electron carriers to a chemical reductant NADPH (LEF, linear electron flux). These reactions are driven by light energy that is funneled to two chlorophyll containing complexes, Photosystem II (PSII) which extracts electrons from water and Photosystem I (PSI) which donates electrons to the terminal acceptor Ferredoxin and ultimately NADP+to form NADPH. Electron transfer between PSII and PSI is essentially connected by the cytochrome b6f complex (cyt b_6f) and mobile electron carriers, plastoquinone/plastoquinol (PQ/PQH$_2$) and plastocyanin. LEF is tightly coupled to the formation of an electrochemical proton gradient (*pmf*, proton motive force) between the lumen (inner thylakoid compartment) and the stroma. First, the oxidation of water by PSII releases protons into the lumen. Second, during the sequential reduction of PQ to PQH_2 in the stroma-localized quinone binding sites of PSII and b_6f, protons are taken from the stroma and are released into the lumen upon oxidation of PQH_2 to PQ at the lumen-localized quinol binding site on cyt b_6f. *Pmf* has an electric field component ($\Delta\psi$) and a chemical gradient component (ΔpH), which are thermodynamically equivalent but kinetically separable. Its dissipation through the chloroplast ATP synthase is coupled to the phosphorylation of ADP to form ATP. The amplitude of *pmf* depends on the rate of proton accumulation and on the resistance (inverse of the proton conductivity, gH^+) of the chloroplast ATP synthase to proton flux. Both NADPH (reducing power) and ATP (chemical bond energy) are used by downstream metabolism, primarily the fixation of carbon dioxide into sugars.

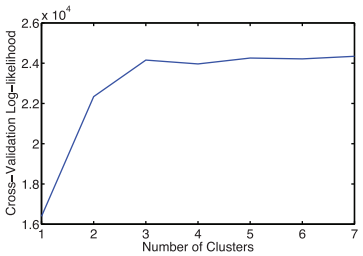

Fig. 3. Cross-validation likelihood of different number of clusters on photosynthesis phenotype data

The formation of high energy states and reactive intermediates during photosynthesis can lead to damage to the photosystems (photoinhibition) and/or formation of reactive oxygen species. Indeed, photoinhibited PSII centers accumulate in a light intensity dependent manner and have been linked to the long-lived component of non-photochemical quenching (q_I). However, LEF is self-regulating; as excitation pressure increases, electron transfer slows (in effect photosynthetic efficiency, ϕ_{II}, decreases) and excess is energy actively dissipated (quenching increases) to minimize the accumulation of reduced intermediates. More specifically, the ΔpH component of *pmf* (acidification of the lumen) slows the oxidation of PQH2 at cyt b_6f and at the same time induces two pH sensitive components of q_E quenching. Under stress, inhibition of LEF may limit the formation of *pmf*. For example, under severe CO_2 limitation, as the acceptor pool (NADP+) diminishes in the absence of carbon fixation, LEF will slow to a rate that cannot support a significant ΔpH. In such cases, a higher ΔpH can be sustained by: (i) alternative pathways that supplement proton uptake such as cyclic electron flux, (ii) increasing the fraction of *pmf* stored as ΔpH, and/or 3) decreasing proton conductivity through the ATP synthase, gH^+ [as reviewed in Baker (2008); Kramer *et al.* (2004)].

In our experiment, we screened a total of 338 chloroplast-targeted single mutant lines in *A. thaliana*. To ensure the reliability of the results, multiple biological replicates (four or more) were examined for each mutant line, so the total number of plants added up to 1, 499, which is also the number of instances used in follow-up analysis. A sample of the phenotype dataset is available at Supplementary Table S1. For each plant, three photosynthesis parameters (ϕ_{II}, q_E and q_I) were measured ($P = 3$ for our algorithm). The photosynthetic parameter, Φ_{II} (quantum yield of photochemistry), when multiplied by the ambient light intensity (*light*) yields a quantity proportional to the LEF rate. The quenching parameters, q_E (rapidly relaxing in the dark, Δph dependent) and q_I (slowing relaxing, related to photoinhibition), are two separate photoprotective mechanisms that dissipate excess light energy (Cruz *et al.*, 2015).

4.2 Clustering results using NPM

Given the photosynthesis phenotype data, we adopted the cross-validation method suggested in Smyth (2000) to decide the number of clusters K. We used 3-fold cross-validation, i.e. at each time, NPM was fit on two thirds of the data (as training set), and the likelihood of the fitted model on the remaining one third data (as test set) was computed. Higher log-likelihood means that the model is better fitted to data. Figure 3 shows that when K is larger or equal to 3, the cross-validation likelihood is near its maximum value. Therefore, we set K to be 3.

Subsequently, three clusters of plants were identified by NPM. For the convenience of visualization, we plotted the clustering results in 2D spaces, with coordinates corresponding to two

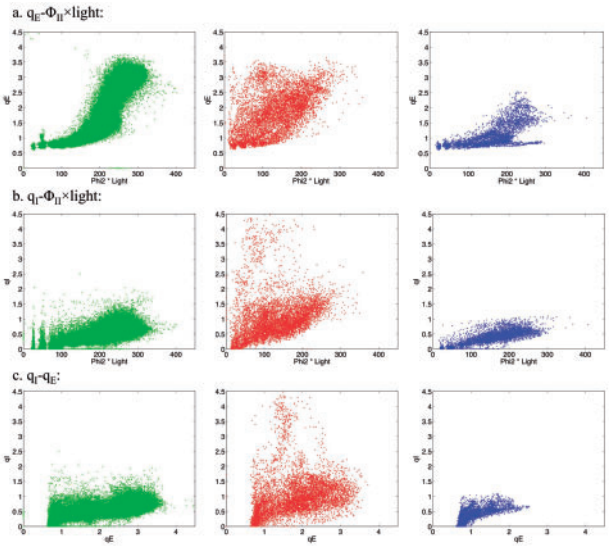

Fig. 4. Three clusters of Arabidopsis chloroplast mutant lines identified by NPM on the plant photosynthesis phenotype data. The data points are plotted in 2D spaces: **(a)** $q_E - (\phi_{II} \times light)$ space, **(b)** $q_I - (\phi_{II} \times light)$ space, **(c)** $q_I - q_E$ space

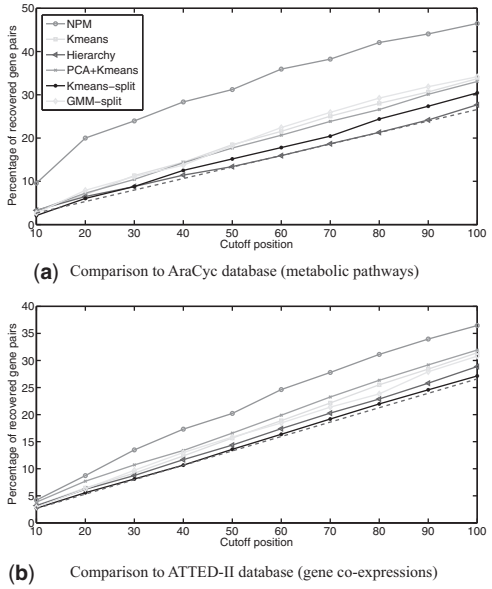

Fig. 5 Evaluation results of different clustering algorithms for the photosynthesis data based on the percentage of recovered similar gene pairs (vertical axis) derived from databases AraCyc **(a)** and ATTED-II **(b)**. The horizontal axis is the cutoff ranking position that is varied from 10 to 100. The dashed line represents the results of random selection

parameters each time (see Fig. 4). To better understand the overall properties of each cluster, we focus on the q_E versus $(\phi_{II} \times light)$ plot in Figure 4a. For all plants, changes in q_E and $(\phi_{II} \times light)$ are positively correlated with the changes in light intensity. The clustering results of the sampled phenotype dataset are available at Supplementary Table S2.

The first cluster (green) in this plot contains 87% of plants, including all the wild type plants. We refer to it as the cluster of *wt-like*. When compared with the other two clusters, the *wt-like* cluster exhibits a stronger correlation between q_E and $(\phi_{II} \times light)$, which reflects the sigmoidal relationship between the ΔpH component of *pmf* generated by photosynthetic electron flux and lumen pH-dependent processes required for q_E quenching, i.e. violaxanthin deepoxidase activation and PsbS protonation (Takizawa *et al.*, 2007).

When compared with the first cluster, the mutant lines in the second cluster (red) tend to have significantly higher q_E sensitivity to electron flux, suggesting that in these mutant lines *pmf* builds to a larger extent due to more restricted proton flux through the ATP synthase lower gH^+) or that the fraction of *pmf* stored as ΔpH is larger. We refer to it as the cluster of *qE-sensitive*. On the other hand, mutant lines in the third cluster (blue) show normal $(\phi_{II} \times light)$ values but significantly lower q_E values than the *wt-like* cluster, indicating that these mutant lines may lack the capacity to accumulate a significant ΔpH or other biochemical components essential for q_E. We refer to it as the cluster of *low-qE*. Similar phenotype partition can be found in the $q_I - (\phi_{II} \times light)$ space (Fig. 4b) and $q_I - q_E$ space (Fig. 4c).

4.3 Performance evaluation

Because a significant portion of the chloroplast-targeted proteins have unknown functions, the optimal gene grouping is unknown, making it difficult to directly measure the clustering performance of NPM (Lamesch *et al.*, 2012). To perform objective evaluation of the clustering results, we computed the gene-to-gene similarities based on the clustering result, denoted as *photosynthesis phenotype-based similarity* (PPS). In parallel, we derived another kind of gene-to-gene similarity score based on the biological knowledge from metabolic pathways or existing gene expression data, denoted as *extra information-based similarity* (EIS). We then measured how well the two sets of similarities are aligned with each other. Our assumption is that a good clustering result of mutant lines should lead to a PPS that well aligns with the EIS. Below we describe the computation of PPS and EIS, respectively.

For each mutant line, there are at least four biological replicates and each replicate may be grouped into a different cluster. By taking the average of the cluster membership vectors for the replicates, we derived a three dimensional vector representation $\mathbf{z}_i$ for each mutant line $\mathcal{M}_i$. The PPS between any two mutant lines $\mathcal{M}_i$ and $\mathcal{M}_j$ was then computed as the similarity between their vector representations:

$$\mathrm{PPS}(\mathcal{M}_i, \mathcal{M}_j) = \mathbf{z}_i^\top \mathbf{z}_j. \tag{9}$$

The biological knowledge base AraCyc (Zhang *et al.*, 2010), (a manually curated database providing the metabolic pathway information for 46 out of the 338 genes in our study) was used as the extra information. More precisely, two genes form a similar gene pair if they are in the same metabolic pathway and we get a total number of 91 similar gene pairs. Therefore two mutant lines have a high EIS value if they form a similar gene pair:

$$\mathrm{EIS} = \begin{cases} 1, & (\mathcal{M}_i, \mathcal{M}_j) \in \text{ similar gene pairs} \\ 0, & \text{otherwise}. \end{cases} \tag{10}$$

Next, we measure how well the PPS is aligned with the EIS derived from AraCyc. For each mutant line $\mathcal{M}_i$ that appears in the *similar gene pairs*, we ranked all the other genes in descending order of their PPS with $\mathcal{M}_i$, and recorded the percentage of the AraCyc-based similar gene pairs that were recovered by the top ranked genes with the cutoff rank varied from 10 to 100. Based on our assumption, the larger the percentage of recovered similar gene pairs, the better the clustering result will be.

Figure 5a summarizes the percentage of similar gene pairs recovered by different clustering algorithms using AraCyc. We also

Table 1. Results of Kolmogorov-Smirnov test on the photosynthesis data using extra information derived from databases AraCyc and ATTED-II

Database	Eval. criterion	Kmeans	Hierarchy	PCA + Kmeans	Kmeans-split	GMM-split	Random selection
AraCyc	Same pathway	9.4×10^{-10}	1.5×10^{-14}	5.7×10^{-9}	1.7×10^{-9}	4.0×10^{-29}	1.9×10^{-50}
ATTED-II	Co-expression	1.3×10^{-5}	1.2×10^{-10}	4.7×10^{-8}	2.0×10^{-21}	7.3×10^{-13}	3.0×10^{-51}

The *P*-values indicate the statistical differences between NPM and other clustering methods

included in Figure 5a the result of random selection for each mutant line as the baseline (highlighted by a dashed line). It shows that the performance of NPM is significantly better than the other clustering methods and the random ranking (Kolmogorov-Smirnov test, see *P*-values in Table 1) (Press *et al.*, 1992).

Because the EIS computed using AraCyc only covers a portion of the genes in our study, we repeated the same evaluation steps with the gene co-expression scores in the ATTED-II database (Obayashi *et al.*, 2011). Although gene co-expression data is not as accurate as manually curated data, it includes most of the genes in our study, and has been well recognized as a reliable source for performance evaluation (Harr and Schlötterer, 2006). We used the mutual rank (MR) scores to compute EIS from ATTED-II. More specifically, we set the threshold of MR to be 20, and gene pairs with MR scores above this threshold are considered to be ATTED-based similar gene pairs. We get a total number of 465 similar gene pairs. The results in Figure 5b indicate that the clustering result generated by NPM has a better correlation with the biological similarities derived from transcriptional co-expressions than the other clustering methods. A Kolmogorov-Smirnov test verifies that NPM is statistically significantly better than all baselines with a significance level of 0.01 on the ATTED-II database (see Table 1).

In summary, the performance evaluations using both the metabolic pathways and the gene co-expression values show that the clustering result generated by NPM better correlates with external biological information than the other clustering methods.

4.4 Biological significance

We examined the biological significance of selected clustering results, as shown in Figure 4. Dozens of interesting mutant lines were discovered from the inter-functional clustering analysis. As a demonstration, we discuss two mutant lines in the following text.

The first one is *cfq* (AT3G24530; coupling factor quick recovery). The *cfq* mutant line harbors a point mutation on the gamma subunit of the chloroplast ATP synthase, which alters its regulation. In our analysis, *cfq* belongs to the *qE-sensitive* cluster (red) which showed lower values of both ϕ_{II} and q_E than wild type. This phenotype is consistent with excess enzymatic activity of the chloroplast ATP synthesis (high gH^+), which minimizes *pmf* formation and subsequently the capacity for photoprotection via the q_E mechanism and the ability to slow electron transfer through the cyt b6f complex. Consequently, Photosystem I in these plants is more susceptible to photodamage particularly under fluctuating light, as confirmed by independent biochemical and spectroscopical methods (Wu *et al.*, 2007).

The second example is the t-DNA insertion line AT1G44575 (npq4) of the pigment-binding protein PsbS. It is associated with photosystem II antenna complexes of higher plants and part of the qE mechanism. In our data, the mutant line had normal ϕ_{II} but very low q_E. Its biased distribution was assigned by NPM to the *low-qE* cluster (blue). This is reasonable because it is an essential component of the qE mechanism, but does not affect the regulation of light-induced electron transfer (Lamesch *et al.*, 2012).

These two examples demonstrate that the clustering approach can distinguish mutant lines with overlapping effects on multiple phenotypes. In this case, both *cfq* and *npq4* showed low q_E but had distinct effects on electron transfer.

We observe in Figure 4a that the mutant lines in the *qE-sensitive* cluster have restricted proton flux through the ATP synthase, resulting in high q_E sensitivity, and the mutant lines in the *low-qE* cluster (blue) have low quenching capacity, causing significantly low q_E but normal ϕ_{II} values. Therefore, we hypothesize that skew from the *wt-like* cluster in the q_E versus ($\phi_{II} \times light$) plot may be used as an indicator for altered activity or regulation of the ATP-synthase.

To verify the hypothesis, we chose four function unknown mutant lines from the cluster results with the most distinct phenotypes to further explore the relationships among multiple photosynthesis parameters. From the *qE-sensitive* cluster, we chose the lines *SALK*_106162 and *SALK*_072581, which are T-DNA insertion mutagen of AT2G29180 and AT4G33520, respectively. AT4G33520 encodes a putative metal-transporting P-type ATPase, and is involved in copper ion transmembrane transport. AT2G29180 is a function unknown chloroplast-targeted protein. From the *low-qE* cluster, we chose *SALK*_114469, a T-DNA insertion mutagen of AT1G79040, which encodes for the PsbR subunit of photosystem II. For comparison, we also chose *SALK*_044616 (AT1G65230) from the *wild-type like* cluster.

We first repeated ϕ_{II} and q_E measurements under several light intensities using an Integrated Diode Emitter Array (IDEA) spectrophotometer, which also allows high-resolution absorbance measurements of other photosynthetic parameters that are not yet possible to perform using high throughput methods (Hall *et al.*, 2013). Figure 6 shows the re-measured plants have the same relationships between q_E and ($\phi_{II} \times light$) as our high-throughput phenotype values, indicating the reproducibility of our data is high.

We also measured proton-flux across the thylakoid membrane, thus the ATP-synthase activity using the same device (Figure 7). As expected, *wt-like* candidate shows no difference in ATP synthase activity compared with wild type (Col-0). Both candidates from the *qE-sensitive* cluster have a decreased proton-flux, vH^+, compared with wild type. The lower proton conductivity, gH^+, compared with wild-type, observed in *SALK*_072581 is consistent with downregulation of ATP synthase activity to limit proton flux and favor *pmf* formation. Contrary to expectations, *SALK*_106162 shows marginally higher gH^+ despite lower vH^+, consistent with lower steady state *pmf* in this mutant compared with wild-type. In this case, the sensitivity of the q_E response may be due either an increase in the fraction of *pmf* stored as ΔpH or increased sensitivity of the quenching response to lumen *pH*. The candidate from the *low-qE* cluster showed similar proton flux rates and conductivities compared with wild type, indicating the loss of q_E may be due to other factors, e.g. a decrease in $\Delta pH/pmf$ ratio or loss of a component that influences quenching efficiency. In summary, the behavior of the distinct

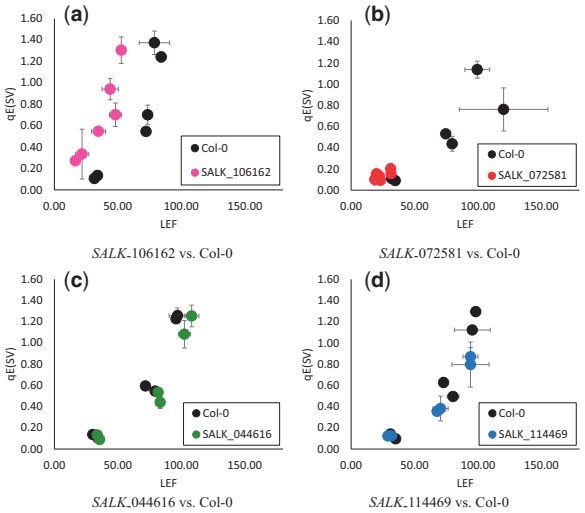

Fig. 6 Verification of the photosynthetic parameters for wild type (Col-0, used as the reference) and four t-DNA insertion mutant lines selected from the NPM clustering results. Measurement of q_E and *LEF* was repeated (proportional to $\Phi_{II} \times light$) using an IDEA spectrophotometer. *SALK*_106162 and *SALK*_072581 are from the *qE-sensitive* cluster. *SALK*_044616 is from the *wild-type like* cluster. *SALK*_114469 is from the *low-qE* cluster

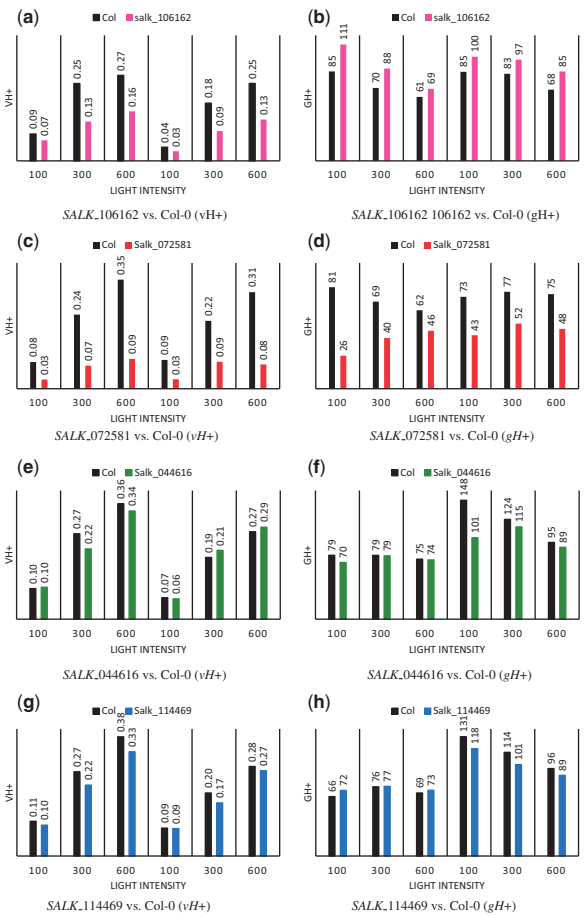

Fig. 7. Experimental verification of the gene groups identified by NPM. The proton flux vH^+ and the voltage-gated proton conductance gH^+ was determined for the four selected t-DNA insertion mutant lines compared with wild type (Col-0). *SALK*_106162 and *SALK*_072581 are from the *qE-sensitive* cluster, *SALK*_044616 is from the *wild-type like* cluster, and *SALK*_114469 is from the *low-qE* cluster

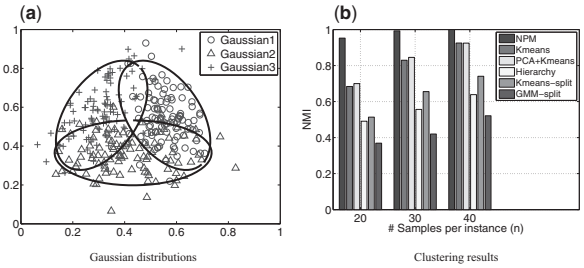

Fig. 8. The three Gaussian distributions used in the simulated experiment and the clustering results for NPM and five baseline clustering methods

clusters relative to the *wt-like* cluster may be indicative of the functions (or loss of function) responsible for the phenotype.

4.5 Experimental results for simulated data

In the second experiment, we created a synthesized dataset with 120 instances. The cloud-of-points representation for each instance $\mathcal{M}_i$ is generated as follows: we randomly chose two out of the three Gaussian distributions (Figure 8a), and sampled n data points from the two Gaussians, where n is varied from 20, 30 to 40. There are totally 3 classes, i.e. Gaussian 1&2, Gaussian 2&3 and Gaussian 1&3. For example, if the two chosen distributions are Gaussian 1 and Gaussian 2, we label this instance $\mathcal{M}_i$ as Gaussian 1&2. The synthetic dataset is available at Supplementary Table S3. We measured the clustering performance by Normalized Mutual Information (Priness *et al.*, 2007) defined as $NMI(\Omega, C) = 2I(\Omega, C)/(H(\Omega) + H(C))$, where Ω is the ground truth class label partition, C is the data partition generated by a clustering algorithm, $I(\Omega, C)$ measures the mutual information between the two partitions, and $H(\Omega)$ and $H(C)$ measure the entropy of the two partitions, respectively.

Figure 8b summarizes the clustering performance for NPM and baseline algorithms. As expected, we observed that the clustering performance of all algorithms improves with increasing number of samples. We also observe that NPM significantly outperforms the baseline algorithms. The poor performance of Kmeans-split and GMM-split suggests that applying clustering algorithms directly to the measured vectors is not appropriate for the cloud-of-points representation. The fact that PCA + Kmeans does not significantly outperform Kmeans suggests that blind dimensionality reduction may not be sufficient for handling high dimensionality in data clustering.

5 Conclusion

Sophisticated phenomics measurements are becoming increasingly important for the discovery of important biological functions and genes (Bilder *et al.*, 2009). Due to the complexity of phenomics datasets, the conventional clustering approaches often fail to identify genes of distinguished functions because they have to make parametric assumptions about the underlying data distribution. There is thus an urgent need for new tools to analyze large and complex phenomics datasets. The NPM method presented here represents an important advance over existing clustering algorithms because it avoids the parametric assumption of data distribution. A demonstration of our method on plant photosynthesis data shows that the proposed technique is effective in capturing mutant lines with similar photosynthesis profiles in comparison to the conventional clustering algorithms such as k-means and hierarchical clustering. In the future, we plan to further improve our analysis by (i) conducting biochemical experiments to validate novel genes with the new phenotypes, and

(ii) developing new clustering algorithms that also take into account the sequential order among photosynthesis measurements.

Funding

This research was supported by the U.S. Department of Energy, Office of Science, Basic Energy Sciences [DE-FG02-91ER20021] for work at the PRL by JC and DMK on data collections and analyses, DOE [DE-AR000202] for experimental work at PRL by EO, the National Science Foundation [award number 1458556] for work by JC on algorithm development, and the MSU Center for Advanced Algal and Plant Phenotyping for development and use of phenotyping tools.

Conflict of Interest: none declared.

References

Bakar,R.B.A. and Watada,J. (2008) Biological clustering method for logistic place decision making. In: Lovrek,I. *et al* (eds). *Knowledge-Based Intelligent Information and Engineering Systems*. Springer, Berlin Heidelberg, pp. 136–143.

Baker,N.R. (2008) Chlorophyll fluorescence: a probe of photosynthesis in vivo. *Annu. Rev. Plant Biol.*, **59**, 89–113.

Bar-Joseph,Z. *et al*. (2003) Continuous representations of time-series gene expression data. *J. Comput. Biol.*, **10**, 341–356.

Bar-Joseph,Z. (2004) Analyzing time series gene expression data. *Bioinformatics*, **20**, 2493–2503.

Ben-Dor,A. *et al*. (1999) Clustering gene expression patterns. *J. Comput. Biol.*, **6**, 281–297.

Bilder,R.M. *et al*. (2009) Cognitive ontologies for neuropsychiatric phenomics research. *Cogn. Neuropsychiatry*, **14**, 419–450.

Butte,A.J. and Kohane,I.S. (2006) Creation and implications of a phenome-genome network. *Nat. Biotech.*, **24**, 55–62.

Carreira-Perpiñán,M.Á. (2006) Fast nonparametric clustering with Gaussian blurring mean-shift. Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh PA, pp. 153–160.

Cobb,J.N. *et al*. (2013) Next-generation phenotyping: requirements and strategies for enhancing our understanding of genotype-phenotype relationships and its relevance to crop improvement. *Theor. Appl. Genet.*, **126**, 867–887.

Cingovska,I. *et al*. (2012) *Protein Function Prediction by Clustering of Protein-Protein Interaction Network. ICT Innovations 2011*, Springer, Berlin Heidelberg, pp. 39–49.

Comaniciu,D. (2002) Image segmentation using clustering with saddle point detection. Proceedings of the IEEE International Conference on Image Processing, **vol. 3**, pp. 297–300, New York.

Corpet,F. (1988) Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res.*, **16**, 10881–10890.

Costa,I.G. (2004) Comparative analysis of clustering methods for gene expression time course data. *Genet. Mol. Biol.*, **27**, 623–631.

Cruz,J.A. *et al*. (2015) Dynamic environmental photosynthetic imaging (DEPI) reveals emergent phenotypes related to the environmental responses of photosynthesis. *Nat. Biotech.*, in press.

Diner,B.A. and Rappaport,F. (2002) Structure, dynamics, and energetics of the primary photochemistry of photosystem II of oxygenic photosynthesis. *Annu. Rev. Plant Biol.*, **53**, 551–580.

Eisen,M.B. *et al*. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl. Acad. Sci.*, **95**, 14863–14868.

Enright,A.J. and Ouzounis,C.A. (2000) GeneRAGE: A robust algorithm for sequence clustering and domain detection. *Bioinformatics*, **16**, 451–457.

Fletcher,R. (2013). *Practical Methods of Optimization*, 2nd ed. John Wiley & Sons, Chichester.

Groth,P. *et al*. (2007) PhenomicDB: a new cross-species genotype/phenotype resource. *Nucleic Acids Res.*, **35**(Suppl. 1), D696–D699.

Groth,P. *et al*. (2010) Phenoclustering: online mining of cross-species phenotypes. *Bioinformatics*, **26**, 1924–1925.

Groth,P. *et al*. (2011) Phenotype mining for functional genomics and gene discovery. *Silico Tools for Gene Discovery*. Springer, New York, pp. 159–173.

Hall,C.C. *et al*. (2013) Photosynthetic measurements with the idea spec: An integrated diode emitter array spectrophotometer/fluorometer. In: *Photosynthesis Research for Food, Fuel and the Future*. Springer, Berlin Heidelberg, pp. 184–188.

Harr,B. and Schlötterer,C. (2006) Comparison of algorithms for the analysis of Affymetrix microarray data as evaluated by co-expression of genes in known operons. *Nucleic Acids Res.*, **34**, e8.

Herwig,R. *et al*. (1999) Large-scale clustering of cDNA-fingerprinting data. *Genome Res.*, **9**, 1093–1105.

Jaqaman,K. *et al*. (2007) Phenotypic clustering of yeast mutants based on kinetochore microtubule dynamics. *Bioinformatics*, **23**, 1666–1673.

Johnson,S.C. (1967) Hierarchical clustering schemes. *Psychometrika*, **32**, 241–254.

Jolliffe,I.T. (1986) *Principal Component Analysis*. Springer-Verlag, New York.

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

Kramer,D.M. *et al*. (2004) Dynamic flexibility in the light reactions of photosynthesis governed by both electron and proton transfer reactions. *Trends Plant Sci.*, **9**, 349–357.

Lamesch,P. *et al*. (2012) The arabidopsis information resource (tair): improved gene annotation and new tools. *Nucleic Acids Res.*, **40**, D1202–D1210.

Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Lones,M.A. and Tyrrell,A.M. (2007) Regulatory motif discovery using a population clustering evolutionary algorithm. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, **4**, 403–414.

Lu,Y. *et al*. (2011) Chloroplast 2010: A database for large-scale phenotypic screening of Arabidopsis mutants. *Plant Physiol.*, **155**, 1589–1600.

Ma,Y. and Fu,Y. (2012) *Manifold Learning Theory and Applications*. CRC Press, London.

MacQueen,J. (1967) Some methods for classification and analysis of multivariate observations. *Proceedings of the 5th Berkeley Symposium on mathematical statistics and probability*, **vol. 1**, pp. 281–297.

McLachlan,G. and Peel,D. (2004) *Finite Mixture Models*. Wiley, New York.

McNicholas,P.D. and Murphy,T.B. (2010) Model-based clustering of microarray expression data via latent Gaussian mixture models. *Bioinformatics*, **26**, 2705–2712.

Nadaraya,E.A. (1965) On non-parametric estimates of density functions and regression curves. *Theor. Probab. Appl.*, **10**, 186–190.

Nascimento,M. *et al*. (2012) Bayesian model-based clustering of temporal gene expression using autoregressive panel data approach. *Bioinformatics*, **28**, 2004–2007.

Obayashi,T. *et al*. (2011) ATTED-II updates: Condition-specific gene coexpression to extend coexpression analyses and applications to a broad range of flowering plants. *Plant Cell. Physiol.*, **52**, 213–219.

Parzen,E. (1962) On estimation of a probability density function and mode. *Ann. Math. Stat.*, **33**, 1065–1076.

Press,W.H. *et al*. (1992) Kolmogorov-Smirnov Test. In: *Numerical Recipes in FORTRAN: The Art of Scientific Computing*, 2nd edn, Cambridge University Press, Cambridge, England, New York, pp. 617–620.

Priness,I. *et al*. (2007) Evaluation of gene-expression clustering via mutual information distance measure. *BMC Bioinformatics*, **8**, 111.

Qian,J. *et al*. (2001) Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions. *J. Mol. Biol.*, **314**, 1053–1066.

Ramoni,M.F. *et al*. (2002) Cluster analysis of gene expression dynamics. *Proc. Natl. Acad. Sci. USA*, **99**, 9121–9126.

Roberts,S.J. (1997) Parametric and non-parametric unsupervised cluster analysis. *Pattern Recogn.*, **30**, 261–272.

Rosenblatt,M. (1956) Remarks on some nonparametric estimates of a density function. *Ann. Math. Stat.*, **27**, 832–837.

Schliep,A. *et al*. (2003) Using hidden Markov models to analyze gene expression time course data. *Bioinformatics*, **19**(Suppl 1), i255–i263.

Scholz,F.W. (1985) Maximum likelihood estimation. In: Kotz,S. *et al*, (eds.) *Encyclopedia of Statistical Sciences*, 2nd edn. Wiley, Hobokenm NJ, pp. 4629–4639.

Shendure,J. and Ji,H. (2008) Next-generation DNA sequencing. *Nat Biotech*., **26**, 1135–1145.

Sivriver,J. *et al*. (2011) An integrative clustering and modeling algorithm for dynamical gene expression data. *Bioinformatics*, **27**, i392–i400.

Smyth,P. (2000) Model selection for probabilistic clustering using cross-validated likelihood. *Stat. Comput.*, **10**, 63–72.

Stute,W. (1982) A law of the logarithm for kernel density estimators. *Ann. Probab*., **10**, 414–422.

Tamayo,P. *et al*. (1999) Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA*, **96**, 2907–2912.

Takizawa,K. *et al*. (2007). The thylakoid proton motive force in vivo. Quantitative, non-invasive probes, energetics, and regulatory consequences of light-induced pmf. *BBA-Bioenergetics*, **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 Luxburg,U. (2007) A tutorial on spectral clustering. *Stat. Comput*., **17**, 395–416.

Wagner,G.P. and Zhang,J. (2013) The pleiotropic structure of the genotype–phenotype map: The evolvability of complex organisms. *Nat. Rev. Gen*., **12**, 204–213.

Wang,H. *et al*. (2013) Function-function correlated multi-label protein function prediction over interaction networks. *J. Comput. Biol*., **20**, 322–343.

Wilson,R. and Spann,M. (1990) A new approach to clustering. *Pattern Recogn*., **23**, 1413–1425.

Wu,L.F. *et al*. (2002) Large-scale prediction of Saccharomyces cerevisiae gene function using overlapping transcriptional clusters. *Nat. Genet*., **31**, 255–265.

Wu,G. *et al*. (2007) A point mutation in atpC1 raises the redox potential of the Arabidopsis chloroplast ATP synthase γ-subunit regulatory disulfide above the range of thioredoxin modulation. *J. Biol. Chem*., **282**, 36782–36789.

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

Yang,W. *et al*. (2014) Combining high-throughput phenotyping and genome-wide association studies to reveal natural genetic variation in rice. *Nat. Commun*., **5**, 5087.

Zhang,P. *et al*. (2010) Creation of a genome-wide metabolic pathway database for Populus trichocarpa using a new approach for reconstruction and curation of metabolic pathways for plants. *Plant Physiol*., **153**, 1479–1491.