Machine learning for Big Data analytics in plants

Chuang Ma[1*], Hao Helen Zhang[2], and Xiangfeng Wang[1,3]

[1] School of Plant Sciences, University of Arizona, 1140 E. South Campus Drive, Tucson, AZ 85721, USA
[2] Department of Mathematics, University of Arizona, 617 North Santa Rita Ave, Tucson, AZ 85721, USA
[3] Department of Plant Genetics and Breeding, College of Agronomy and Biotechnology, China Agricultural University, Beijing 100193, China

Rapid advances in high-throughput genomic technology have enabled biology to enter the era of 'Big Data' (large datasets). The plant science community not only needs to build its own Big-Data-compatible parallel computing and data management infrastructures, but also to seek novel analytical paradigms to extract information from the overwhelming amounts of data. Machine learning offers promising computational and analytical solutions for the integrative analysis of large, heterogeneous and unstructured datasets on the Big-Data scale, and is gradually gaining popularity in biology. This review introduces the basic concepts and procedures of machine-learning applications and envisages how machine learning could interface with Big Data technology to facilitate basic research and biotechnology in the plant sciences.

Big Data technology and machine learning

'Big Data' (large datasets) are frequently encountered in the modern biological sciences (see Glossary). The three 'V' features of Big Data – velocity, volume, and variety – have catalyzed the development of innovative technical and analytical strategies to cope with the data [1]. As a result of rapid advances in high-throughput data generation technologies, biologists have entered the Big Data era [2]. Although the cost of data generation is no longer a major concern for genome-wide research, the computational efficiency of analyzing terabytes or even petabytes of data has become a bottleneck [3,4]. The plant science community is seeking novel solutions to the three grand challenges of Big Data: scalable infrastructure for parallel computation; management schemes for large-scale datasets; and intelligent data-mining analytics, which are similar to the challenges faced by any research field producing Big Data (Box 1). The Apache Hadoop ecosystem, which offers a suite of libraries and tools for data storage, access and automated parallel processing, is considered a promising platform that solves the first two infrastructural problems of Big Data [5–9]. 'Machine learning', an emerging multidisciplinary field of computer science, statistics, artificial intelligence, and information theory, is particularly favored by data scientists for exploiting information hidden in Big Data [10]. The unique features of Big Data – massive, high dimensional, heterogeneous, complex or unstructured, incomplete, noisy, and erroneous – have seriously challenged traditional statistical approaches, which are mainly designed for analyzing relatively smaller samples [1]. Furthermore, large biological systems can be so complex that they cannot be adequately described by traditional statistical methods (e.g., classical linear regression analysis, and correlation-based statistical analysis) developed based on hypothesized or prespecified distribution of data, whereas many modern learning techniques are data-driven and able to provide more feasible solutions. For example, popular machine learning approaches such as support vector machines and classification trees do not presume the distribution for data [11,12].

In biology, most of the methods used in genome-wide research are based on statistical testing and designed for analyzing a single experimental dataset. The data explosion introduced by modern genomics technologies requires biologists to rethink data analysis strategies and to create powerful new tools to analyze the data. In recent decades, machine learning has been envisaged by life scientists as a high-performance scalable learning system for data-driven discovery. The effective performance of machine learning has been demonstrated by the Big-Data-scale exploration of an aggregation of various data sources from the encyclopedia of DNA elements (ENCODE) and model organism encyclopedia of DNA elements (modENCODE) projects in animals [13–15]. However, machine learning has not been widely used for analyzing large datasets in plants [12,16,17]. With the success of the iPlant Collaborative (http://www.iplantcollaborative.org) in building a central supercomputing infrastructure and a data consortium for the plant science community [18], this is now an opportune time for plant scientists to take advantage of Big Data technology to address plant-specific problems in their basic research.

Despite the promising potential of machine learning, it is often misunderstood or misused by biologists, mainly owing to their insufficient knowledge of machine learning and the complexity of the biological systems under study. Therefore, the primary goals of this review are to introduce the basic concepts and procedures of machine learning in biology and to envisage how machine learning could

Corresponding author: Wang, X. (xwang1@cals.arizona.edu).

Keywords: Big Data; machine learning; large-scale datasets; plants.

**Current address:* College of Life Sciences, Northwest Agriculture and Forest University, Yangling, Shaanxi 712100, China.

1360-1385/

http://dx.doi.org/10.1016/j.tplants.2014.08.004

Glossary

Active learning: a machine-learning approach that iteratively update training dataset by strategically selecting informative data for obtaining a classifier with high prediction performance.

Adaptive boosting (AdaBoost): a machine-learning approach that iteratively increases the weight of misclassified samples for boosting weak classifiers to be a stronger classifier.

Apache Hadoop: a framework that allows the automated parallel storing and processing of data on a large cluster of computing nodes.

Apache Mahout: a project that aims to build a scalable machine-learning library running on Hadoop for Big Data analysis.

Attributes (or features, inputs, independent variables, predictors): a set of numerical or categorical quantities used to describe an example.

Big Data: a popular term describing large datasets with the features of high velocity, volume, and variety, which are difficult to process using traditional database management and analytical methods.

Big Data scale: 1 Exabyte (EB) = 1000 Petabytes (PB) = 1 000 000 Terabytes (TB) = 1 000 000 000 Gigabytes (GB).

Cloud service: a new type of use-on-demand data computing and storage paradigm that enables users to build time-consuming applications and manage large datasets on many commodity computing nodes.

Evaluation metric: a criterion used to measure the performance of a learned model.

Examples (or instances, samples): the objects from which a model is learned or on which a model will be applied for prediction.

F-score (or F-measure): a measure that can be used to find the optimized threshold of machine-learning models with both high precision and recall.

Hadoop ecosystem: a group of Hadoop-related Big Data storage, access, processing and analysis utilities, including HBase, Spark, Hive, Pig, Sqoop, and Mahout.

Hadoop Distributed File System (HDFS): a distributed file system developed for accessing and processing the data stored with Hadoop in a parallel manner.

Hot deck and cold deck imputation: two techniques for handling missing data. Hot deck imputation replaces missing data with substituted values randomly selected from similar samples in the same dataset. By contrast, cold deck imputation selects values from other datasets.

Kernel methods: machine-learning algorithms that transform the features of samples into a higher-dimensional space using kernel functions, such as polynomial function and radial basis function.

MapReduce: a programming model that enables users to easily write programs supporting automated parallel processing distributed on multiple computing nodes.

Matthews correlation coefficient (MCC): a measure used in machine learning to evaluate the prediction performance of two-class classifiers by taking into account true positives, true negatives, false positives and false negatives.

Model (or learner, learning model): a machine-learning algorithm that assigns an output to an example described with a set of attributes.

Next-generation sequencing technology: a technology that produces DNA or RNA fragments with the capacity of high throughput, scalability, speed, and resolution.

Not Only Structured Query Language (NoSQL): a new data management system that enables the storage and manipulation of data through the construction of highly reliable, scalable and distributed databases.

Output (or response, outcome, dependent variable): the outcome of a learning problem. The output can be a categorical label (qualitative) or a continuous value (quantitative).

Principle component analysis (PCA): a statistical technique that eliminates redundancy by converting data into a set of linearly uncorrected variables (i.e., principle components).

Random forest (RF): a modern machine learning algorithm that constructs with an ensemble of decision trees for classification and regression problems.

Rhadoop: an R package that provides an application programming interface (API) for running R scripts with Hadoop.

Support vector machine (SVM) classifiers: models that built with support vector machine algorithm to perform the classification of positive and negative samples in a high-dimensional space.

Training dataset: a set of examples used to learn the model.

Tuning dataset: a set of examples used to select and validate the model.

Testing dataset: a set of examples used to assess the generalization performance of a learned model.

interface with Big Data technology to facilitate basic research and applied biotechnology in plants.

Basics for building a machine-learning system

Machine learning refers to the process of teaching computers the ability to automatically extract important information from examples to achieve improved prediction or search capabilities for associations and/or patterns in data [19]. Machine learning is a multidisciplinary field incorporating computer science, statistics, artificial intelligence, and information theory. The basic definitions and concepts of machine learning (attributes, evaluation metric, examples, model, output, training dataset, testing dataset, and tuning dataset) are defined in the Glossary.

Based on the goal of learning tasks, machine-learning algorithms are organized taxonomically. Two major algorithms are 'supervised learning' and 'unsupervised learning' [20–22]. Supervised learning takes place when the training examples are labeled with their known outputs. For the example of identifying salt-responsive genes in *Arabidopsis thaliana* shown in Box 2, the label of each example (i.e., gene) in the training dataset is +1 or –1, which is known to the learning model for indicating stress-responsive or non-stress-responsive genes, respectively. Unsupervised learning is used when we observe only attributes for the training examples but not their outputs (i.e., examples are unlabeled, or are labeled but unknown to the learning model). There are other types of machine-learning algorithms that are more complex or a hybrid of different algorithms. For example, semisupervised learning handles both labeled and unlabeled data (i.e., only partial examples in training dataset are labeled) [22], with online learning the model learns sequentially from infinite data streams [23], and active learning is designed to strategically select the most representative examples to be manually labeled [23].

Supervised learning can be further divided into classification and regression based on whether the output is qualitative (categorical) or quantitative (continuous) [21]. The goal of classification is to predict labels of examples, whereas regression involves the estimation of a trend and the prediction of real-valued outputs. In binary classification problems, we typically use the labels +1 or –1 to denote the membership of an example: examples from the two classes are referred to as positive samples or negative samples, respectively.

Commonly encountered machine-learning problems in the real world include:

- Classification (also known as pattern recognition): the problem of learning a classifier that assigns labels (or membership) to new unlabeled examples.
- Regression: the problem of estimating the relation between real-valued outputs and attributes to make predictions.
- Clustering: the task of grouping data such that examples in the same group (called a cluster) are more similar to each other than to those of other groups.
- Recommendation: the task of prioritizing examples based on the attributes of interest.
- Dimensionality reduction: the problem of transforming attributes in a high-dimensional space to a space of fewer dimensions.
- Network analysis: the study of exploring associations between systems components for understanding the biological function of individual components and elucidating the behaviors of biological systems.
- Density estimation: the problem of estimating the probability density function for a population based on the observed data.

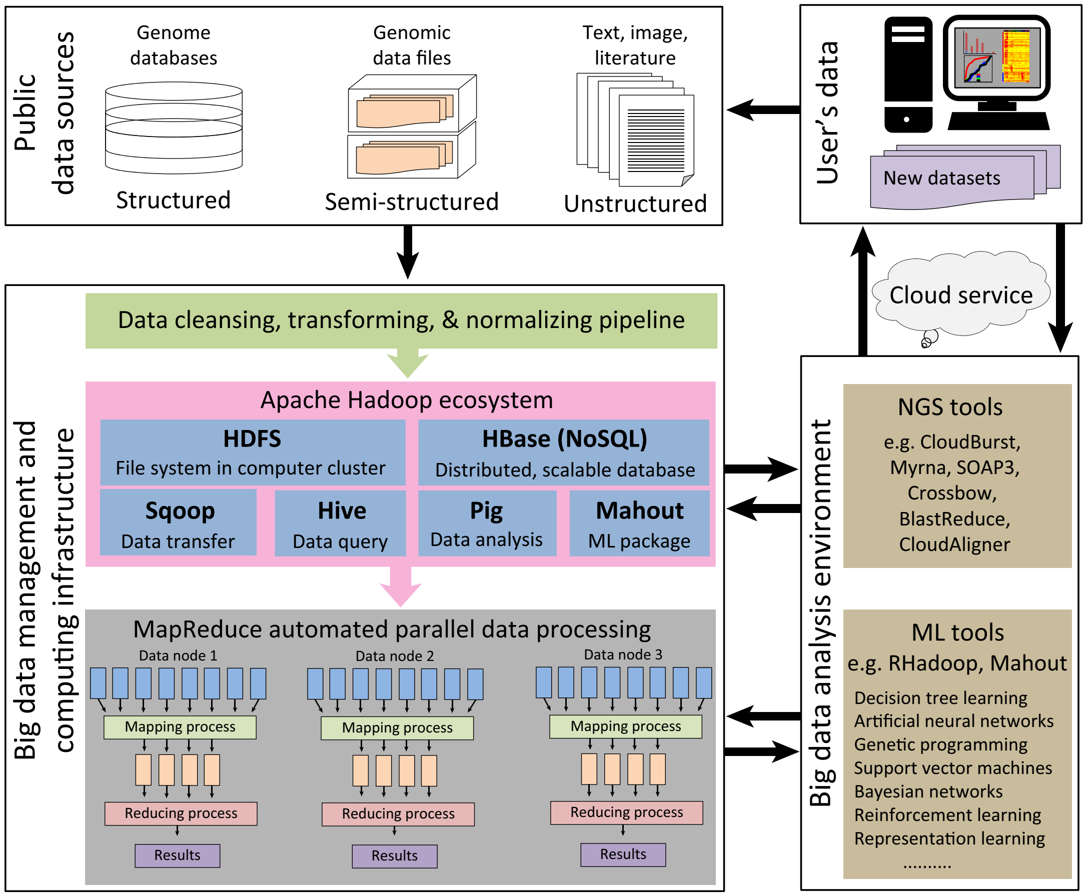
Building a machine-learning system is a complex, multistep, and sometimes recursive process consisting of three basic steps. First, the raw data are preprocessed to generate a dataset with better representation and quality (data preprocessing). Data preprocessing includes cleansing, normalization, transformation, handling of missing data, and feature extraction and selection. Data processing is an important step in machine learning that removes irrelevant and redundant information or noisy and unreliable data. Second, a predictive or descriptive model is learned or trained by using an appropriate machine-learning algorithm that is selected based on the data and learning task (model training). Model training typically involves estimating the parameters and determining the model structure from the data using numerical algorithms (model optimization). Third, the generalization performance of a learned model is evaluated with evaluation metrics (model evaluation). A popular evaluation technique is cross-validation, which is useful to avoid overfitting [24]. Various metrics are used for measuring the overall quality of a machine-learning system in terms of, for example, its prediction accuracy, reliability, and computing time. For example, receiver operating characteristic (ROC) analysis is commonly used to evaluate the discriminative power of the classification model at different thresholds, and a grand score – area under the curve (AUC) of the ROC plot – gives a summary of the overall performance.

For many complex biological systems, various types of information and prior knowledge are available and provide valuable extra information from different perspectives.

Box 1. Big Data infrastructure for the plant science community

Big Data technology generally refers to three aspects of technical innovation that cope with super-large datasets: namely, automated parallel computation, data management schemes, and data mining analytics (Figure I). To set up Big Data infrastructure for the plant science community will require the construction of:

- A centralized Big Data computing infrastructure on top of high-performance computer clusters. The open-source Apache Hadoop ecosystem is such an integrative platform and is composed of Hadoop operation commands, the MapReduce programming model, the Hadoop distributed file system (HDFS), and a variety of utilities for warehousing disparate forms of structured, semi-structured, and unstructured datasets.
- A Big Data storage domain to integrate plant genome databases and public datasets, as well as automated pipelines to preprocess, transform and query data in super-large-volume datasets.
- An analysis environment provided as a use-on-demand cloud service by integrating popular bioinformatics software and machine-learning-based applications that support the Hadoop computing infrastructure (Table I).



TRENDS in Plant Science

Figure I. The general components of Big Data technology.

Table I. Representative machine-learning algorithms and R packages

Learning tasks	Algorithms and methods	Software packages in R
Classification	Nearest neighbor methods, linear discriminant analysis, quadratic discriminant analysis, logistic regression, naive Bayes classifier, support vector machine (SVM), classification trees, neural nets	knn, knn1, glm2, kernsvm, svmpath, CART, e1071, nnet, gcdnet, tree, randomForest, sda, rda, penalizedLDA
Regression	Least squares, linear models, ridge regression, additive models, generalized additive model, nearest neighbor methods, regression trees, project pursuit regression kernel methods, local regression, splines, wavelet smoothing, Bayesian models	lm, gam, knn, splines, locfit, mgcv, polyspline, earth, cosso
Clustering	K-means clustering, spectral clustering, hierarchical clustering, self-organizing maps, association rules, multidimensional scaling, independent component analysis, local multidimensional scaling	Kclust, cluster, fastcluster, sparseBC, sparcl, pvclust
Feature selection	Best subset selection, forward selection, least angle regression, shrinkage methods: lasso, elastic net, group lasso, fused lasso, sure independence screening	Regsubsets, LAR, glmnet, elasticnet, glmpath, gglasso, Sparsenet, penalizedLDA
Dimensionality reduction	Principal component analysis (PCA), factor analysis, kernel PCA, partial least squares	pca, pls, mda, elasticnet, lpc
Ensemble learning	Boosting methods, bagging, random forest, additive regression	Adaboost, randomForest, ada, adabag, erboost, mart
Network analysis	Gaussian graphical models, Bayes networks	bnlearn, JGL, GGMselect, lvgm, gRain, gRim, gRbase
Density estimation	Kernel density estimation	KernSmooth

Determining how to integrate the information into the model learning is an important component of machine learning (knowledge integration). In Box 2, we use a machine-learning method to discover salt-responsive genes in *Arabidopsis* as an example to explain the basic machine-learning terms and to showcase how to build a machine-learning system to identify the genes expressed in response to salt stress by learning the expression patterns of a set of known salt-responsive genes collated from the literature.

Pitfalls (and remedies) in machine learning for biology

High-dimension, low-sample size (HDLSS) data

HDLSS data are common in plant and other science studies [25]. Such data contain a large number of attributes and a relatively smaller number of training examples, which tends to cause overfitting: the model fits the training data too well but has poor performance on testing data [26]. In addition, many of the collected attributes are irrelevant (weakly correlated with the output) or redundant (highly correlated with each other). Their inclusion may make the learning process unstable and yield a model with large variance and poor discriminative power. Therefore, dimension reduction is necessary when using HDLSS data.

Dimension reduction includes feature selection (or attribute selection) and feature extraction. Feature selection is the process of selecting the subset of relevant or important features. Feature selection may take place at the data preprocessing or model learning step. When the number of features is too high, correlation analysis is often used to preselect or to screen features prior to model building [27]. Feature extraction is used to create new features by the transformation or function of raw features. One popular feature extraction procedure is principal component analysis (PCA), which extracts a small set of directions (called leading principal components; PCs) to represent the data and achieves great dimension reduction. One drawback of PCA is being difficult to interpret because each PC is a linear combination of all raw features. Recently, the sparse PCA analysis [28] was proposed to facilitate interpretation by producing PCs with sparse loadings, which involve a smaller number of raw features than traditional PCA methods.

Missing or incomplete data

Missing or incomplete data are commonly encountered in machine learning. For example, a subset of genes on microarrays may miss expression values for certain technical or biological reasons. A naïve method to address missing data is to delete examples with incomplete values; however, this is not considered appropriate because 'missing' may itself be important information [29]. In practice, there are different types of mechanisms for missing data, including missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR) [30]. If the data are MCAR, then the analysis based on removing missing data would increase estimation variance but may still be unbiased, since the remaining data are still representative of the population. However, if the data are NMAR or the 'missingness' occurs systematically, removing missing values will lead to samples which do not represent the population, and then the analysis based on the remaining data would produce bias.

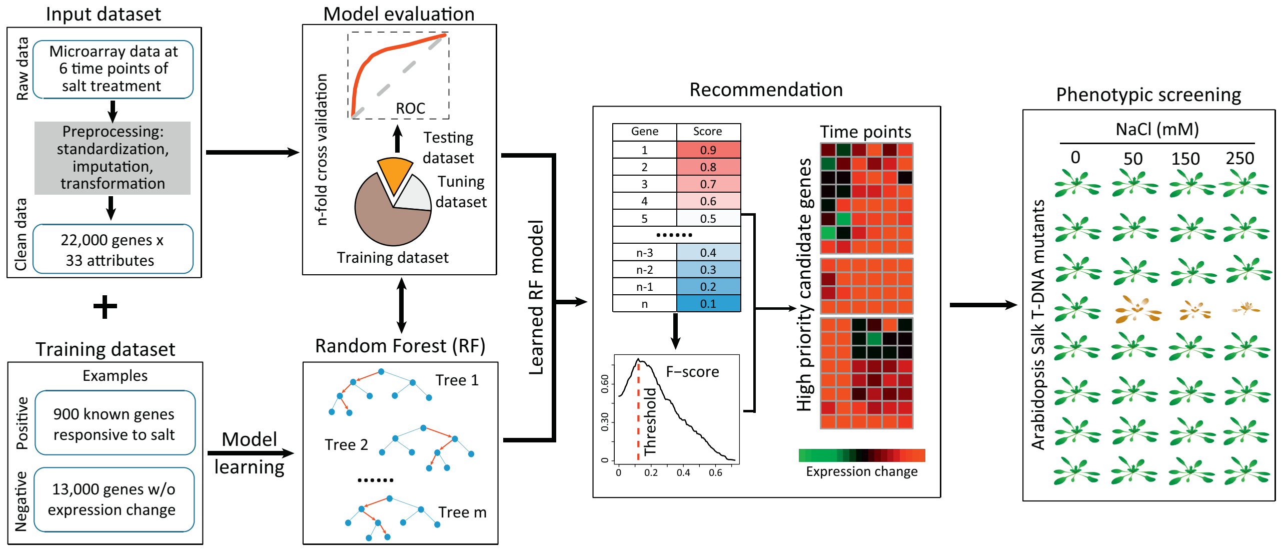
A common way to handle incomplete data is imputation, a process of replacing missing data with substitute values [30]. For example, a predicted value is assigned to the gene with missing data by considering local or global correlations among genes in the data [31,32]. Various imputation techniques have been developed, such as hot deck and cold deck imputation, mean imputation, regression imputation, last observation carried forward imputation, and stochastic imputation [33]. Multiple imputation is a commonly used method in machine learning. Rather than replacing each missing value with one randomly imputed value, multiple imputation replaces each value with several imputed values that reflect our uncertainty about the imputation model [30].

Missing benchmark tools

Benchmarking is a way to evaluate and improve machine-learning algorithms. To make the evaluation and

Box 2. Building a machine-learning system for gene discovery

To explain the basic machine-learning terms and procedures for building a supervised machine-learning system, we demonstrate the use of a supervised machine-learning method to discover genes responsive to salt stress and recommend high-priority candidate genes for phenotypic screening experiments (Figure I). The raw data are composed of 22 000 genes (examples) probed with an Affymetrix microarray whose expression values were profiled at six time points over a 24-h period, while root tissue was subjected to salt treatment. The raw data were first preprocessed by normalization to remove technical variations, imputation to replace missing values, and the transformation of six time-point values to 33 numeric statistics (attributes) to characterize the degree of expression change. Then, the clean dataset was converted to an input matrix of 22 000 genes in rows and 33 attributes in columns. The training set includes 900 known salt-related genes (labeled examples) as positive samples and 13 000 genes without expression change as negative samples. We treat this analysis as a classification problem, and use the random forest (RF) classifier (machine-learning model) to determine whether a gene was related to salt-induced response. The RF classifier builds a series of decision trees by resampling positive and negative samples, which are used to train and validate the final model. The *n*-fold cross-validation (evaluation metric) is used to assess the generation performance of the RF model (model validation and evaluation). The entire gene matrix is randomly split into *n* subsets; each of which keeps a roughly equal number of positive and negative samples. For each cross-validation, *n*–1 subsets are firstly merged into one dataset and then divided into two parts: the training dataset and the tuning dataset. Optimized parameters, at which the RF classifier has high precision and recall, are obtained by training on the training dataset and testing on the tuning dataset. The remaining subset is used as the testing dataset to evaluate the trained RF classifier's prediction power by the ROC curve that plots the variation of TPR (y axis: true positive rate) versus FPR (x axis: false positive rate) at all possible prediction scores. After *n* rounds of cross-validation, an average AUC is calculated to represent the predictive power of the trained RF model. All the genes are ordered based on the priority scores assigned by the trained RF model. The genes with higher priority scores than a threshold are recommended as stress-related candidate genes, and the optimal threshold is determined when the maximal F-score is achieved. The recommended high-priority candidate genes with available Salk transfer DNA (T-DNA) mutation lines are functionally tested by phenotypic screening experiments.



TRENDS in Plant Science

Figure I. Using machine learning to discover salt stress-related genes in *Arabidopsis*.

comparison fair, effective, and efficient, benchmarking analysis should use a set of reliable and unbiased data, a system of quantitative and qualitative evaluation metrics, and a comprehensive comparison scheme [34]. Benchmarking tools with a list of comprehensive evaluation functionalities are particularly important to understand the advantages and limitations of each algorithm. For example, benchmarking analysis has been used for machine-learning-based prediction of promoter regions [35], protein–protein interactions [36], and protein docking sites [37]. A well-designed benchmarking process helps to: select the optimal machine-learning algorithm for the biological problem under study, validate the learned model, and improve its performance. However, most existing benchmark analyses only focus on particular performance aspects, such as the computing speed or prediction accuracy, while other important algorithmic features, such as stability, reliability, and accessibility, are usually neglected [34].

Integrating data of multiple types and sources

In biology, there are typically multiple types and sources of data to describe the various aspects of the biological system under study. For example, a plant system can be characterized by its DNA variation, gene expression, protein levels and sites, protein–protein interactions, and disease-associated traits. Determining how to combine various data sources to build an integrated model is an important yet challenging problem. Owing to the different nature and platforms of these large-scale data, new sophisticated techniques are needed to synthesize information and to integrate knowledge [38].

Imbalanced classification problems

In biology, one prevalent problem in classification analysis is the imbalance of subclasses in the data, namely, when certain classes have substantially more examples than do other classes. This problem is particularly prevalent in high-dimensional datasets [39]. An extreme class imbalance can mislead the classifier by tending to overlearn the majority class and to perform poorly in the prediction of the minority class [40–42].

There are several solutions to the imbalanced classification problem at both the data and algorithm levels [40]. At the data level, resampling techniques can be used to create a balanced dataset through either oversampling the minority class or undersampling the majority class [43,44]. Alternatively, the issue can be solved algorithmically by modifying the classification model. For example, the cost-sensitive learning algorithm can be used to construct weighted support vector machine (SVM) classifiers. The binary cost-sensitive SVM classifier minimizes the risk associated with unequal misclassification errors: it first estimates weighting factors for two classes based on their sample proportions in the training dataset, and then the classifier is trained by imposing a larger penalty onto one type of misclassification error. Other methods such as Adaptive Boosting (AdaBoost), kernel methods, active learning, and ROC-select have been shown to address this issue effectively [41,45,46]. However, if the two classes are sufficiently imbalanced, then the use of recognition-based machine learning instead of two-class classification methods is recommended [47].

Choice of evaluation metrics

The selection of appropriate evaluation methods is extremely important in machine learning to obtain a fair and comprehensive performance assessment for a machine-learning system. Effective evaluation should be able to assess the model prediction accuracy, determine important factors for success, suggest improvement methods, and identify new applications.

Commonly used evaluation metrics in machine learning include threshold-based metrics [e.g., accuracy, sensitivity, specificity, precision, recall, F-score, and Matthews correlation coefficient (MCC)], ordering-based metrics (e.g., AUC of the ROC plot), and probability-based metrics (e.g., the root-mean-square error) [48]. Threshold-based metrics are generally used for discrete classification, such as a decision tree model in which only label information is assigned to each feature. However, these metrics are not effective for scoring classifiers because imbalanced classes may yield a misleading result [39,40,49]. Ordering-based metrics can assess model performance at all possible thresholds and, thus, are not influenced by the imbalanced class problem because the proportions of positive and negative classes are not considered [50]. Ordering-based metrics are usually combined with *n*-fold cross-validation that randomly divides the dataset into training and testing samples for *n* rounds of evaluation to avoid the overtraining of the evaluated classifiers [16,51]. The inappropriate use of cross-validation methods can overestimate the trained machine-learning models, yielding unusually high AUC values that are even greater than 0.99 [52]. A proper way to optimize parameters is a so-called ‘double-cross-validation-loop’ method that further splits the training set for an inner *n*-fold cross-validation [52,53].

Potential machine-learning applications for studies in plants

Machine-learning methodologies have been used in many areas of large-scale data analysis for a broad variety of utilities in genomics, transcriptomics, proteomics, and systems biology [20,21,52]. However, to date, animal studies have been the focus of most machine-learning-based applications, and only a limited number of machine-learning-based applications have been used in plant science studies. In addition, most plant machine-learning models have been trained based on the genomic attributes and known patterns of gene activity in *Arabidopsis*. To ensure performance, species-specific models trained in different plant genomes are required. Moreover, the accumulation of genomic resources, large-scale experimental data and knowledge collated from the literature suggest that there is an urgent demand for systematic applications of machine-learning to solve plant-specific problems in an integrative fashion. Based on previous experiences of using machine learning in both plants and animals, machine learning has promising applications in the following three representative areas of plant genome research: genome assembly and genome annotation, integrative inference of the gene regulatory network, and integrative prediction of gene function.

Genome assembly and genome annotation

Machine learning has been widely used in many areas of genomic analysis, such as genome assembly, genomic variation detection, genome-wide association studies, and the *in silico* annotation of genomic loci that encode, for example, protein-coding genes, transposable elements, noncoding RNAs, miRNAs and targets, transcription factor binding sites, *cis*-regulatory elements, enhancer and silencer elements, and mRNA alternative splicing sites [54–64]. Although most machine-learning applications were developed for animals, many of them are readily applicable to plants. For example, to improve the assembly quality of the *Drosophila mojavensis* and *Escherichia coli* genomes based on shotgun sequencing reads, machine learning was used to detect assembly errors caused by repetitive DNA sequences [62,63]. This technique would be highly useful in the assembly of large, repeat-rich crop genomes with Hadoop-based parallel computing, such as the ~5 Gb barley (*Hordeum vulgare*) genome and ~17 Gb wheat (*Triticum aestivum*) genome, with which the traditional *de novo* assemblers can barely cope. Machine-learning methods have also been used in parsing polyploid plant genomes. In wheat, machine learning has been used to discriminate the highly similar gene homologs encoded in the three subgenomes [65]. Considering that polyploidy is widespread in the plant kingdom, this type of application could potentially be used to determine the expression levels and genotypes of gene homologs in polyploid plants. Another potential use of machine learning in plants is to identify alternatively spliced mRNA isoforms assisted by RNA-Seq data, although alternative splicing is not as prevalent as it is in animals. In the honey bee (*Apis mellifera*), machine-learning-based logistic regression has been used to identify 16 023 alternative splicing events in 5644 genes based on attributes extracted from DNA motifs around splice sites and RNA-Seq read alignments [64]. Considering that more and more RNA-Seq data and genome sequences are available in plants, this application could potentially be used to estimate the frequency and function of alternative splicing events in the plant kingdom.

Integrative inference of the gene regulatory network

Biological functions are fulfilled through the complex molecular interactions of genes in a cell [66]. The network representation of these interactions, such as the regulatory relations between transcription factors and target genes, can greatly aid biologists in forming hypotheses and identifying candidate genes for experimental testing. A regulatory network can be reconstructed *in silico* based on the coexpression patterns of genes. However, a connection of two genes in such a network is only a hint of the potential association of their functions and does not necessarily reflect direct regulation [16]. The machine-learning-based incorporation of multiple types of regulatory evidence from various data sources into expression-based networks has become a trend to increase the fidelity of inferred regulatory interactions. For instance, both supervised (e.g., SVM) and unsupervised (e.g., Sum Rule) machine-learning methods have been used in reconstructing the regulatory network in *Drosophila melanogaster* [15]. Multiple types of data were integrated to infer the edges that potentially represent regulatory interactions, including data from conserved *cis*-regulatory motifs, chromatin immunoprecipitation coupled with next-generation sequencing (ChIP-Seq) analysis of epigenetic modification and transcription factor binding sites, and gene coexpression patterns [15]. In model plants, such as *Arabidopsis*, rice (*Oryza sativa*) and maize (*Zea mays*), numerous similar datasets are available, making it feasible to apply machine-learning-based integrative inference to a regulatory network. Another example of a machine-learning application in a network is the prediction of cell-type-specific transcription factor binding sites in human cell lines, which used an SVM classifier that synthetically considered the gene expression change, DNA sequence preferences, and associated chromatin contexts [67]. This work established a model that combined the static sequence context with dynamic expression activity to reconstruct cell-type-specific regulatory networks. Another machine-learning algorithm that

Table 1. List of large machine-learning software packages

	Weka[a]	scikit-learn[b]	SHOGUN[c]	mlPy[d]	Mlpack[e]	Apache Mahout[f]	ml-hadoop[g]	MLlib[h]	Oryx[i]
Graphical user interface	Yes	No	No	No	No	No	No	No	No
Main language	C++	Python	C++	Python	C++	Java	Python; Java	Python; Java; Scala	Java
Hadoop-based	No	No	No	No	No	Yes	Yes	Not necessary	Yes
Parallelization	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes
Preprocessing	Yes	Yes	Yes	Yes	No	No	No	No	No
Feature selection	Yes	Yes	No	Yes	No	No	No	No	No
Handling missing values	Yes	Yes	No	No	No	No	No	No	No
Performance evaluations	Yes	Yes	Yes	Yes	No	No	No	No	No
Visualization	Yes	Yes	No	Yes	No	No	No	No	No
Classification[j]	NB;NN;RBFN; DT;RF; SVM	NB; NN; DT;RF; SVM; LDA	NB; DT; SVM; LDA	SVM; LDA	NB;	LR; NB; RF	NB	NB; LR; SVM	RF
Regression[k]	SVR; RVM; GP	SVR;KRR; GP	SVR; KRR;GP	SVR; KRR			Multiple linear regression	Linear regression	RF regression
Clustering[l]	HC; K-mean	HC; K-mean	HC; K-mean	HC; K-mean	K-mean	HC; K-means;	K-mean	K-means	K-means
Plants (P), animals (A) or microorganisms (M)	P; A; M	P; A; M	P; A; M						

[a] Weka 3: data mining software in Java (http://www.cs.waikato.ac.nz/ml/weka).

[b] scikit-learn: machine-learning in Python (http://scikit-learn.org/).

[c] SHOGUN 3.2.0 (http://www.shogun-toolbox.org/).

[d] mlpy: machine-learning Python (http://mlpy.sourceforge.net/).

[e] mlpack: a scalable C++ machine-learning library (http://www.mlpack.org).

[f] Apache Mahout™ (http://mahout.apache.org).

[g] ml-hadoop: Hadoop implementation of machine-learning algorithms (https://code.google.com/p/ml-hadoop).

[h] Spark MLlib (http://spark.apache.org/mllib).

[i] Oryx (https://github.com/cloudera/oryx).

[j] Representative classification-related machine-learning algorithms: DT, decision tree; LDA, linear discriminant analysis; RL, logistic regression; NB, naive Bayes; NN, neural network; RBFN, RBF network; RF, random forest; SVM, support vector machine.

[k] Representative regression-related machine-learning algorithms: GP, Gaussian processes; KRR, kernel ridge regression; RVM, relevant vector machine; SVR, support vector regression.

[l] Representative clustering-related machine-learning algorithms: HC, hierarchical clustering; K-mean, k-mean clustering.

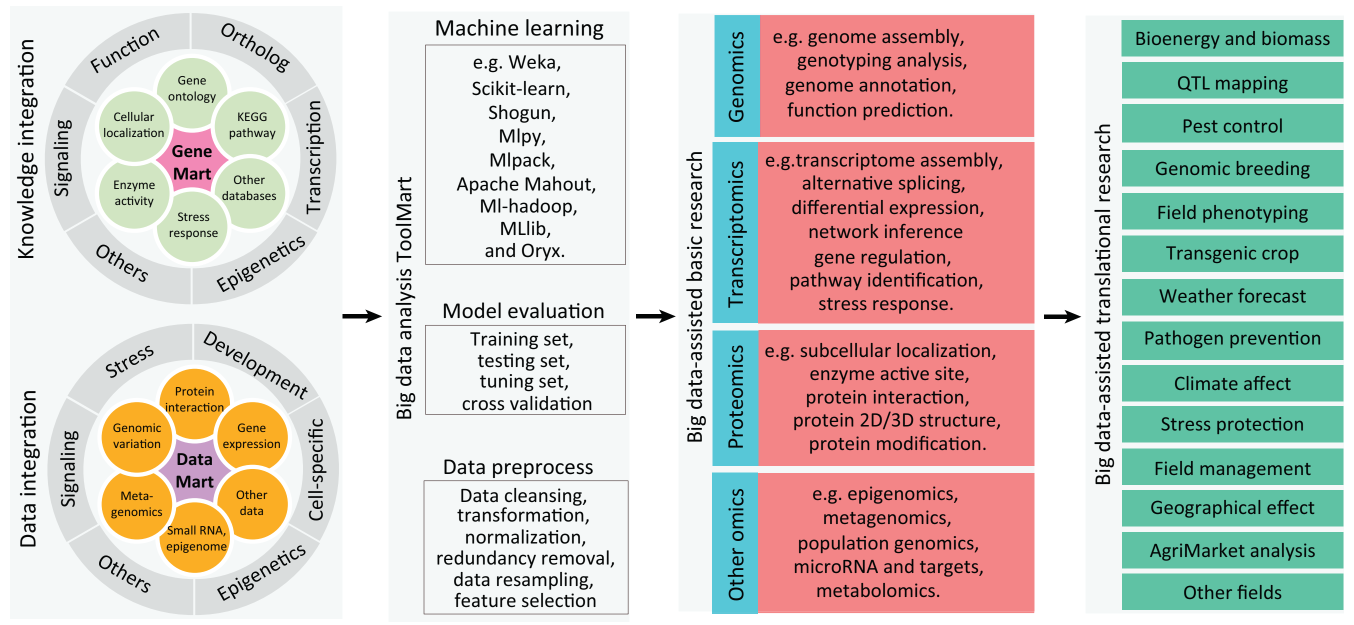
has been applied in budding yeast for studying stress response [68], named the input–output Hidden Markov Model (HMM) – a variant of the traditional HMM that calculates the output and transition probabilities conditioned on the input variables – could be used to infer dynamic stress-responsive networks in *Arabidopsis*, providing a novel way to elucidate molecular interactions between plants and environmental stress. In machine learning, network analysis consists of a family of methods used to elucidate the association of objects, which have been used to infer molecular interactions in biological networks [69–72]. For example, neural networks and Bayesian networks are useful to explore and identify structure and dynamics of gene networks from large-scale genomic data [72].

Integrative prediction of gene function

In *Arabidopsis*, only 16% of the 28 775 genes have been functionally studied, and 46% were annotated with putative functions deduced from homology-based prediction [73]. In crops and other non-model plants, a large proportion of genes still lack functional annotations, even for those plants with available genome sequences [74]. The prediction of gene function through the machine-learning-based integration of multiple data sources is a novel strategy in animals [75,76]. Ensemble learning, which combines a suite of supervised machine-learning methods, including SVM, Random Forest, artificial neural network, and Bayesian Markov random field, has been used to predict gene functions from numerous types of evidence in addition to homology-based searches, such as the results from text

Box 3. Machine learning as a service for Big-Data-assisted research

Hadoop-based machine-learning tools are integrated and interfaced with Big Data management and computing infrastructure, providing services to the plant science community in terms of both Big-Data-assisted basic research and translational research. This platform is composed of GeneMart, DataMart, and ToolMart (Figure I). GeneMart is a knowledge base of plant genes that ingests, distills, integrates and categorizes various sources of information from public genome databases. GeneMart uses a set of rules to assign labels to genes based on their functional attributes. Genes and attribute labels are mapped as keys and values, respectively; easily adaptable to the MapReduce framework. DataMart stores public omics data, categorized based on experimental design and data type. Datasets in the DataMart are labeled with the same rules for the purpose of compiling training, tuning and validating datasets for machine-learning model training, optimization and evaluation, respectively. ToolMart supplies software applications for data preprocessing, model validation, and a suite of machine-learning packages. Based on the data types and learning problems, the appropriate machine-learning models are recommended. Users can accomplish model training, model optimization, and model evaluation through a visualized interface. The outcomes of the machine-learning analysis of existing data are categorized by fields of research (e.g., genomics, transcriptomics, and proteomics) and classes of biological questions. Plant scientists can access these outcomes from a visual web interface to assist their basic research. The trained machine-learning models are supplied to users to apply to new omics datasets. In addition, this platform provides a pool of machine-learning resources to translate findings from basic research to applied biotechnology and to any field of agricultural research assisted by Big Data technology. Machine learning can be used to build models for a broad variety of utilities, including the estimation of the genomic breeding values of candidate lines incorporated with genome-wide association study data [85]; crop production management incorporated with weather and climate data [86]; pathogen prevention incorporated with metagenomic data [87]; biotic and abiotic stress protection incorporated with environmental data; gene-trait association prediction incorporated with quantitative trait locus (QTL) mapping data [88]; pattern recognition in high-throughput field phenotyping [89]; and agricultural supply and demand forecasting incorporated with market data.


Knowledge integration
Function
Ortholog
Signaling
Transcription
Others
Epigenetics
Gene ontology
Cellular localization
KEGG pathway
Gene Mart
Enzyme activity
Other databases
Stress response
Data integration
Stress
Development
Signaling
Cell-specific
Others
Epigenetics
Protein interaction
Genomic variation
Gene expression
Data Mart
Meta-genomics
Other data
Small RNA, epigenome
Big data analysis ToolMart
Machine learning
e.g. Weka, Scikit-learn, Shogun, Mlpy, Mlpack, Apache Mahout, Ml-hadoop, MLlib, and Oryx.
Model evaluation
Training set, testing set, tuning set, cross validation
Data preprocess
Data cleansing, transformation, normalization, redundancy removal, data resampling, feature selection
Big data-assisted basic research
Genomics
e.g. genome assembly, genotyping analysis, genome annotation, function prediction.
Transcriptomics
e.g.transcriptome assembly, alternative splicing, differential expression, network inference gene regulation, pathway identification, stress response.
Proteomics
e.g. subcellular localization, enzyme active site, protein interaction, protein 2D/3D structure, protein modification.
Other omics
e.g. epigenomics, metagenomics, population genomics, microRNA and targets, metabolomics.
Big data-assisted translational research
Bioenergy and biomass
QTL mapping
Pest control
Genomic breeding
Field phenotyping
Transgenic crop
Weather forecast
Pathogen prevention
Climate affect
Stress protection
Field management
Geographical effect
AgriMarket analysis
Other fields

TRENDS in Plant Science

Figure I. A platform for machine learning-based Big Data analytics in plants.

mining data in the published literature, attributes extracted from sequence context, patterns of gene expression, and interactions with functionally known proteins [75]. In addition, an array of information regarding the product of a gene – the protein – may also help to deduce its function. In proteome annotation, the machine-learning-based prediction of 2D and 3D structure, subcellular localization, protein family delineation, protein–protein interaction, and the functional effect of protein mutation has been applied in animals and/or plants [76–84]. Most crop plants are monocots and lack a reverse genetic system such as the tDNA mutational lines in *Arabidopsis*, therefore, the machine-learning-based *in silico* integrative prediction of gene function and protein property may greatly expedite the discovery of agriculturally important genes in the post-genomics era.

Perspective: interfacing machine learning with Big Data

Owing to the needs of Big-Data-scale analysis, many open-source machine-learning software packages have been reengineered with the MapReduce function (Table 1). These tools, together with public genome databases and high-throughput data in plants, will be interfaced with the Apache Hadoop ecosystem as a centralized Big Data analytical platform to serve the plant science community (Box 3). This development may not be that far off given that hardware fundamentals for such a platform are already available in the iPlant Collaborative. Hadoop can be built on top of existing high-performance computing clusters in iPlant, such as myHadoop, which allows Hadoop jobs to be run on high-performance computing clusters (https://github.com/glennklockwood/myhadoop). In the long run, the practical steps that Big Data bioinformaticians need to take are: first, upgrade the algorithms of existing software using the MapReduce programming framework, and then develop novel metadata schemes to manage the annotational and experimental information associated with a plant gene. The experimental information in particular, which includes knowledge integration to build a 'GeneMart' and data integration to build a 'DataMart', may require a hybrid of computer and human curation efforts using a system of unified labeling rules to categorize genes and datasets based on functional descriptions and experimental designs, respectively. This method is practical for the purpose of compiling training example sets in a machine-learning analysis, in which genes and data attributes can be correspondingly retrieved from the GeneMart and DataMart based on the same category labels. In addition, a data preprocessing pipeline with unified algorithms and rules to cleanse, transform, and normalize data should be developed. The heterogeneous nature of biological data should be well considered when building the pipeline because, for example, sequence polymorphic data, gene expression data, proteomic data, epigenomic data, small RNA data, and transcription factor binding data are present in different forms.

The central component of the Big Data analytical platform is a 'ToolMart' that supplies a variety of machine-learning models that are categorized based on the types of learning problems, expected outcomes, and forms of analyzed data in plants. These models are pretrained with the examples in the GeneMart and DataMart, and are provisioned to users to use in multiple aspects of machine-learning analyses with different types of high-throughput data such as that listed in Box 3. Machine learning comprises a complex family of methods, and it is powerful only if a learning problem is appropriately defined and the correct model is selected. However, the dearth of essential machine-learning knowledge and the difficulty in translating certain biological problems into machine learning problems have impeded its use by biologists. Furthermore, the above-discussed pitfalls also obstruct an objective evaluation of the performance of the model. Under these circumstances, a compromise is the ensemble learning offered by this integrated machine-learning environment. Ensemble learning uses multiple approximate learning models to achieve enhanced performance without precisely understanding the internal structure of a machine-learning black box. Assisted by Big Data technology, this platform may not only facilitate basic plant biology research but also provide a computational framework with a pool of machine-learning tools for various aspects of translational research that make knowledge from basic science applicable in a timely manner to agricultural biotechnology and field practices (Box 3).

Acknowledgments

This work was supported by U.S. National Science Foundation grants DBI-1261830 (X.W.), NSF DMS-1309507 (H.Z.) and DMS-1418172 (H.Z.).

Disclaimer statement

The authors declare that they have no conflicts of interest relevant to this paper.

References

1 Berman, J.J. (ed.) (2013) *Principles of Big Data: Preparing, Sharing, and Analyzing Complex Information*, Elsevier, (Morgan Kaufmann

2 Marx, V. (2013) Biology: the big challenges of big data. *Nature* 498, 255–260

3 Brauer, E.K. *et al.* (2014) Next-generation plant science: putting big data to work. *Genome Bio.* 15, 301

4 Schatz, M.C. (2012) Computational thinking in the era of big data biology. *Genome Bio.* 13, 177

5 Schumacher, A. *et al.* (2014) SeqPig: simple and scalable scripting for large sequencing data sets in Hadoop. *Bioinformatics* 30, 119–120

6 Nordberg, H. *et al.* (2013) BioPig: a Hadoop-based analytic toolkit for large-scale sequence data. *Bioinformatics* 29, 3014–3019

7 Langmead, B. *et al.* (2009) Searching for SNPs with cloud computing. *Genome Bio.* 10, R134

8 Niemenmaa, M. *et al.* (2012) Hadoop-BAM: directly manipulating next generation sequencing data in the cloud. *Bioinformatics* 28, 876–877

9 Zou, Q. *et al.* (2013) Survey of MapReduce frame operation in bioinformatics. *Brief. Bioinform.* http://dx.doi.org/10.1093/bib/bbs088

10 Ratner, B. (ed.) (2011) *Statistical and Machine-learning Data Mining: Techniques for Better Predictive Modeling and Analysis of Big Data* (2nd edn), CRC Press, Taylor & Francis Group

11 Bassel, G.W. *et al.* (2012) Systems analysis of plant functional, transcriptional, physical interaction, and metabolic networks. *Plant Cell* 24, 3859–3875

12 Bassel, G.W. *et al.* (2011) Functional network construction in *Arabidopsis* using rule-based machine learning on large-scale data sets. *Plant Cell* 23, 3101–3116

13 Roy, S. *et al.* (2010) Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* 330, 1787–1797

14 Bernstein, B.E. *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74

15 Marbach, D. *et al.* (2012) Predictive regulatory models in *Drosophila melanogaster* by integrative inference of transcriptional networks. *Genome Res.* 22, 1334–1349

16 Ma, C. *et al.* (2014) Machine learning-based differential network analysis: a study of stress-responsive transcriptiomes in *Arabidopsis*. *Plant Cell* 26, 520–537

17 Van Landeghem, S. *et al.* (2013) The potential of text mining in data integration and network biology for plant research: a case study on *Arabidopsis*. *Plant Cell* 25, 794–807

18 Goff, S.A. *et al.* (2011) The iPlant collaborative: cyber infrastructure for plant biology. *Front. Plant Sci.* 2, 34

19 Mjolsness, E. and DeCoste, D. (2001) Machine learning for science: state of the art and future prospects. *Science* 293, 2051–2055

20 Larranaga, P. *et al.* (2006) Machine learning in bioinformatics. *Brief. Bioinform.* 7, 86–112

21 Tarca, A.L. *et al.* (2007) Machine learning and its applications to biology. *PLoS Comput. Bio.* 3, e116

22 Zhao, N. *et al.* (2014) Determining effects of non-synonymous SNPs on protein–protein interactions using supervised and semi-supervised learning. *PLoS Comput. Bio.* 10, e1003592

23 Bordes, A. *et al.* (2005) Fast kernel classifiers with online and active learning. *J. Mach. Learn. Res.* 6, 1579–1619

24 Japkowicz, N. and Shah, M., eds (2011) *Evaluating Learning Algorithms: A Classification Perspective*, Cambridge University Press

25 Devijver, P.A. and Kittler, J., eds (1982) *Pattern Recognition: A Statistical Approach*, Prentice Hall

26 Hall, P. *et al.* (2005) Geometric representation of high dimension, low sample size data. *J. R. Stat. Soc. B* 67, 427–444

27 Saeys, Y. *et al.* (2007) A review of feature selection techniques in bioinformatics. *Bioinformatics* 23, 2507–2517

28 Taguchi, Y.H. and Murakami, Y. (2013) Principal component analysis based feature extraction approach to identify circulating microRNA biomarkers. *PloS ONE* 8, e66714

29 Pearson, K. (1901) On lines and planes of closest fit to systems of points in space. *Philos. Mag.* 2, 559–572

30 Little, R.A. and Rubin, D.B., eds (2002) *Statistical Analysis with Missing Data* (2nd edn), John Wiley and Sons

31 Liew, A.W. *et al.* (2011) Missing value imputation for gene expression data: computational techniques to recover missing data from available information. *Brief. Bioinform.* 12, 498–513

32 Aittokallio, T. (2010) Dealing with missing values in large-scale studies: microarray data imputation and beyond. *Brief. Bioinform.* 11, 253–264

33 Haukoos, J.S. and Newgard, C.D. (2007) Advanced statistics: missing data in clinical research – part 1: an introduction and conceptual framework. *Acad. Emerg. Med.* 14, 662–668

34 Aniba, M.R. *et al.* (2010) Issues in bioinformatics benchmarking: the case study of multiple sequence alignment. *Nucleic Acids Res.* 38, 7353–7363

35 Abeel, T. *et al.* (2009) Toward a gold standard for promoter prediction evaluation. *Bioinformatics* 25, i313–i320

36 Martin, J. (2014) Benchmarking protein–protein interface predictions: why you should care about protein size. *Proteins* 82, 1444–1452

37 Hwang, H. *et al.* (2010) Protein–protein docking benchmark version 4.0. *Proteins* 78, 3111–3114

38 Linn, M.C. (2006) The knowledge integration perspective on learning and instruction. In *The Cambridge Handbook of the Learning Sciences* (Sawyer, R.K., ed.), pp. 243–264, Cambridge University Press

39 Blagus, R. and Lusa, L. (2010) Class prediction for high-dimensional class-imbalanced data. *BMC Bioinformatics* 11, 523

40 Zhao, X.M. *et al.* (2008) Protein classification with imbalanced data. *Proteins* 70, 1125–1132

41 Gudys, A. *et al.* (2013) HuntMi: an efficient and taxon-specific approach in pre-miRNA identification. *BMC Bioinformatics* 14, 83

42 Chawla, N.V. *et al.* (2004) Editorial: special issue on learning from imbalanced data sets. *SIGKDD Explorations* 1–6

43 Zang, Q. *et al.* (2013) Binary classification of a large collection of environmental chemicals from estrogen receptor assays by quantitative structure–activity relationship and machine learning methods. *J. Chem. Inf. Model.* 53, 3244–3261

44 Furey, T.S. *et al.* (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* 16, 906–914

45 Zheng, W. *et al.* (2014) An ensemble method for prediction of conformational B-cell epitopes from antigen sequences. *Comput. Biol. Chem.* 49, 51–58

46 He, H. and Garcia, E.A. (2009) Learning from imbalanced data. *IEEE Trans. Knowledge Data Eng.* 1263–1284

47 Yousef, M. *et al.* (2008) Learning from positive examples when the negative class is undetermined – microRNA gene identification. *Algorithms Mol. Bio.* 3, 2

48 Japkowicz, N. and Shah, M., eds (2011) *Evaluation Learning Algorithm: A Classification Perspective*, Cambridge University Press

49 Lertampaiporn, S. *et al.* (2013) Heterogeneous ensemble approach with discriminative features and modified-SMOTEbagging for pre-miRNA classification. *Nucleic Acids Res.* 41, e21

50 Fawcett, T. (2006) An introduction to ROC analysis. *Pattern Recogn. Lett.* 27, 861–874

51 Zou, C. *et al.* (2011) Cis-regulatory code of stress-responsive transcription in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. U.S.A.* 108, 14992–14997

52 Kelchtermans, P. *et al.* (2014) Machine learning applications in proteomics research: how the past can boost the future. *Proteomics* 14, 353–366

53 Wessels, L.F. *et al.* (2005) A protocol for building and evaluating predictors of disease state based on microarray data. *Bioinformatics* 21, 3755–3762

54 Ruffalo, M. *et al.* (2012) Accurate estimation of short read mapping quality for next-generation genome sequencing. *Bioinformatics* 28, i349–i355

55 Yip, K.Y. *et al.* (2013) Machine learning and genome annotation: a match meant to be? *Genome Bio.* 14, 205

56 DePristo, M.A. *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43, 491–498

57 Yandell, M. and Ence, D. (2012) A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.* 13, 329–342

58 Leclercq, M. *et al.* (2013) Computational prediction of the localization of microRNAs within their pre-miRNA. *Nucleic Acids Res.* 41, 7200–7211

59 Sherwood, R.I. *et al.* (2014) Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nat. Biotechnol.* 32, 171–178

60 St Laurent, G. *et al.* (2013) Genome-wide analysis of A-to-I RNA editing by single-molecule sequencing in Drosophila. *Nat. Struct. Mol. Biol.* 20, 1333–1339

61 Shlyueva, D. *et al.* (2014) Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* 15, 272–286

62 Choi, J.H. *et al.* (2008) A machine-learning approach to combined evidence validation of genome assemblies. *Bioinformatics* 24, 744–750

63 Palmer, L.E. *et al.* (2010) Improving de novo sequence assembly using machine learning and comparative genomics for overlap correction. *BMC Bioinformatics* 11, 33

64 Li, Y. *et al.* (2013) TrueSight: a new algorithm for splice junction detection using RNA-seq. *Nucleic Acids Res.* 41, e51

65 Brenchley, R. *et al.* (2012) Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature* 491, 705–710

66 Middleton, A.M. *et al.* (2012) Modeling regulatory networks to understand plant development: small is beautiful. *Plant Cell* 24, 3876–3891

67 Arvey, A. *et al.* (2012) Sequence and chromatin determinants of cell-type-specific transcription factor binding. *Genome Res.* 22, 1723–1734

68 Gitter, A. *et al.* (2013) Linking the signaling cascades and dynamic regulatory networks controlling stress responses. *Genome Res.* 23, 365–376

69 Petrey, D. and Honig, B. (2014) Structural bioinformatics of the interactome. *Annu. Rev. Biophys.* 43, 193–210

70 Marbach, D. *et al.* (2012) Wisdom of crowds for robust gene network inference. *Nat. Methods* 9, 796–804

71 Dehmer, M. and Basak, S.C., eds (2012) *Statistical and Machine Learning Approaches for Network Analysis*, John Wiley and Sons

72 Lee, W.P. and Tzou, W.S. (2009) Computational methods for discovering gene networks from gene data. *Brief. Bioinform.* 10, 408–423

73 Lamesch, P. *et al.* (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* 40, D1202–D1210

74 Rhee, S.Y. and Mutwil, M. (2014) Towards revealing the functions of all genes in plants. *Trends Plant Sci.* 19, 212–221

75 Radivojac, P. *et al.* (2013) A large-scale evaluation of computational protein function prediction. *Nat. Methods* 10, 221–227

76 Yachdav, G. *et al.* (2014) PredictProtein – an open resource for online prediction of protein structural and functional features. *Nucleic Acids Res.* 42, W337–W343

77 Wang, Z. and Xu, J. (2013) Predicting protein contact map using evolutionary and physical constraints by integer programming. *Bioinformatics* 29, i266–i273

78 Faraggi, E. and Kloczkowski, A. (2014) A global machine learning based scoring function for protein structure prediction. *Proteins* 82, 752–759

79 Bradford, J.R. *et al.* (2010) GO-At: in silico prediction of gene function in *Arabidopsis thaliana* by combining heterogeneous data. *Plant J.* 61, 713–721

80 Kaundal, R. *et al.* (2010) Combining machine learning and homology-based approaches to accurately predict subcellular localization in *Arabidopsis*. *Plant Physiol.* 154, 36–54

81 Reumann, S. *et al.* (2012) PredPlantPTS1: a web server for the prediction of plant peroxisomal proteins. *Front. Plant Sci.* 3, 194

82 Lingner, T. *et al.* (2011) Identification of novel plant peroxisomal targeting signals by a combination of machine learning methods and in vivo subcellular targeting analyses. *Plant Cell* 23, 1556–1572

83 Agrawal, G.K. *et al.* (2010) Plant secretome: unlocking secrets of the secreted proteins. *Proteomics* 10, 799–827

84 Park, Y. and Marcotte, E.M. (2011) Revisiting the negative example sampling problem for predicting protein–protein interactions. *Bioinformatics* 27, 3024–3028

85 Ornella, L. *et al.* (2014) Genomic-enable prediction with classification algorithm. *Heredity* 112, 616–626

86 Ehret, D.L. *et al.* (2011) Neural network modeling of greenhouse tomato yield, growth and water use from automated crop monitoring data. *Comput. Electron. Agric.* 79, 82–89

87 Verma, R. and Melcher, U. (2012) A support vector machine based method to distinguish proteobacterial proteins from eukaryotic plant proteins. *BMC Bioinformatics* 13, S9

88 Moore, J.H. *et al.* (2010) Bioinformatics challenges for genome-wide association studies. *Bioinformatics* 26, 445–455

89 Eliceiri, K.W. *et al.* (2012) Biological imaging software tools. *Nat. Methods* 9, 697–710