Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Plant Biology

More effort — more results: recent advances in integrative 'omics' data analysis

Dhivyaa Rajasundaram[1,2,3] and Joachim Selbig[1,2]

The development of 'omics' technologies has progressed to address complex biological questions that underlie various plant functions thereby producing copious amounts of data. The need to assimilate large amounts of data into biologically meaningful interpretations has necessitated the development of statistical methods to integrate multidimensional information. Throughout this review, we provide examples of recent outcomes of 'omics' data integration together with an overview of available statistical methods and tools.

Addresses
[1] Institute of Biochemistry and Biology, University of Potsdam, Karl-Liebknecht-Str. 24-25, D-14476 Potsdam-Golm, Germany
[2] Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, D-14476 Potsdam-Golm, Germany

Corresponding author: Selbig, Joachim (jselbig@uni-potsdam.de)
[3] Current address: Laboratory of Bioinformatics and Genomics, Department of Crop and Environmental Sciences, Virginia Tech, 24061 Blacksburg, VA, USA.

Current Opinion in Plant Biology 2016, **30**:57–61

This review comes from a themed issue on **Genome studies and molecular genetics**

Edited by **Yves Van de Peer** and **J Chris Pires**

For a complete overview see the Issue and the Editorial

Available online 15th February 2016

http://dx.doi.org/10.1016/j.pbi.2015.12.010

Omics integration going forward

Plants as living organisms hold an exceptional place as dynamic components of our world that shape and are, in turn shaped by the environment. Complex functions of plant systems such as growth, cell-division, systemic response to perturbations, and emerging phenotypes arise from the intrinsic properties, hierarchical organization, and interaction of thousands of system components. With recent advances in high-throughput 'omics' techniques, it is now possible to generate system-level measurements for virtually all types of cellular components to query unchartered frontiers in the knowledge of plant growth, metabolism and response to environmental cues. As a result, the growing repertoire of 'omics' experiments is providing researchers with a wide spectrum of data. However, the increasing amount of data renders any successive data analysis a difficult task. In addition, understanding the interplay between the components of a biological system posits the demand for integration of data arising from different system-levels. Despite the challenges associated, 'omics' integration studies designed to characterize *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, and human systems have pervaded literature in recent years. The integrative systems approach has been gaining large attention of plant biologists in the last few years, concomitant with the increase in large amounts of molecular data. Here, we review some of the emerging aspects in integrative plant systems research together with their technological advances. We will also highlight a number of recent studies that successfully integrate 'omics' data and finally, the article will conclude with a discussion of the challenges that face the field as well as future directions.

Overview of different approaches

Clearly, the multi-dimensional data generated from high-throughput techniques need systematic approaches that inherently require integration of heterogeneous information (Figure 1). With the availability of data from several 'omics' experiments, often integrative analysis is exercised for two purposes: first, a descriptive analysis to find underlying relationship between the data sets and second, to predict a certain response using one or more explanatory data sets. Indeed, the past few years have seen a multitude of methods for integrative analysis of two data sets and some of the excellent reviews cover topics such as integrated network analysis in plants [1,2], co-expression tools for plant biology [3], biochemical pathway or ontology based integration [4], and machine learning for big data analytics in plants [5]. In addition, there are numerous methods that have been developed to integrate more than two data sets at a time and implemented in **R** for ecological data analysis, food quality, assessment and behavioral research.

Of late, some of these methods are used to corroborate and harness the complexity of 'omics' data sets. Table 1 represents the most commonly used statistical methods with available **R** packages.

Specific examples of integrated analysis categories

In this section, we provide details and specific examples for the different categories of integrative analysis of 'omics' data sets.

Integrative inference of 'omics' data generated from different cellular levels

Availability of 'omics' experiments studying different cellular levels yield messenger ribonucleic acid (mRNA),

Glossary

Canonical correlation analysis: CCA is a multivariate statistical method that seeks to quantify the strength of the relationship by maximizing the correlation between the two sets of variables.

Expression Conservation score (EC): EC score is derived by calculating the PCC between the adjacent edge weights, i.e., the co-expression relationship of two networks thus capturing the similarity of gene neighborhoods.

Gene set enrichment analysis: A computational method that determines whether an a priori defined set of genes shows statistically significant concordant difference between two biological states.

MapMan: A user-driven tool that displays a large data set onto diagrams of metabolic pathways or other processes.

Multiple co-inertia analysis: MCIA is used to describe several data sets observed on the same set of observations by recovering the maximum total variance from each data set.

Multiblock continuum redundancy: An extension of multiple redundancy analysis and used in the prediction of several explanatory data sets and a response data set. It is mostly used when the level of multi-collinearity is high within the explanatory data sets.

Multiblock partial least squares regression: This method is an extension of standard PLS and used in cases where there are several explanatory data sets and one response data set.

Multiblock redundancy analysis: mbRA is useful in the prediction of 'k' exploratory data sets and a response data set. Within the 'k' data sets, important variables which have an impact on the response dataset is also provided.

Multiple factor analysis: MFA is an extension of principal component analysis tailored to handle multiple data sets that measure sets of variables collected on the same observations.

O2PLS: This method enables integrative analysis of data sets by separating joint information across multiple analytical platforms from systemic variation that is unique to each platform.

Paintomics: Serves as a web tool for the integration and visualization of transcriptomics and metabolomics data.

Partial least squares: PLS is used in cases of ill conditioned linear regression models. The method aims to find a linear relationship between two sets of variables and the prediction factor is achieved by extracting a set of orthogonal factors called latent variables.

Partial triadic analysis: PTA is seen as the simplest of the STATIS family and is the PCA of a series of PCA's. The aim is to analyze a series of 'k' tables having the same observations and same variables.

Regularized generalized canonical correlation analysis: RGCCA is used for the analysis of relationships between 'k' data sets and is an extension of CCA. RGCCA also identifies subsets of variables of each data set which are active in their relationships with the other data sets.

Sparse partial least squares: sPLS method imposes sparsity within the context of PLS and thereby carry out dimension reduction and variable selection simultaneously. Sparsity is imposed within the context of PLS and thereby carries out dimension reduction using PLS and variable selection using LASSO penalization.

Sparse generalized canonical correlation analysis: SGCCA is an extension of RGCCA with L1 penalty to account for variable selection.

STATIS: STATIS is an extension of PCA tailored to handle multiple data sets that measure sets of variables collected on the same observations. **dual-STATIS** is a variant of STATIS where the same variables are measured on different sets of observations. **DISTATIS** handles multiple data sets collected on the same observations and generalizes metric multidimensional scaling to three-way distance matrices. **ANISOSTATIS** extends STATIS to give specific weights to each variable rather than to the whole data set. **K + 1 STATIS** predicts the relationship between 'k' data sets and one external data set. As an extension of STATIS, this method highlights the relationship between the 'k' data sets as well as determine the data set which is best related to the external dataset. Details of other variants of STATIS with their suitable applications is provided in an excellent review by Abdi *et al.* [24••].

Tukey Honest Significance Difference: A post hoc test performed after the analysis of variance (ANOVA). Tukey HSD can clarify to the researcher which groups among the sample have significant differences.

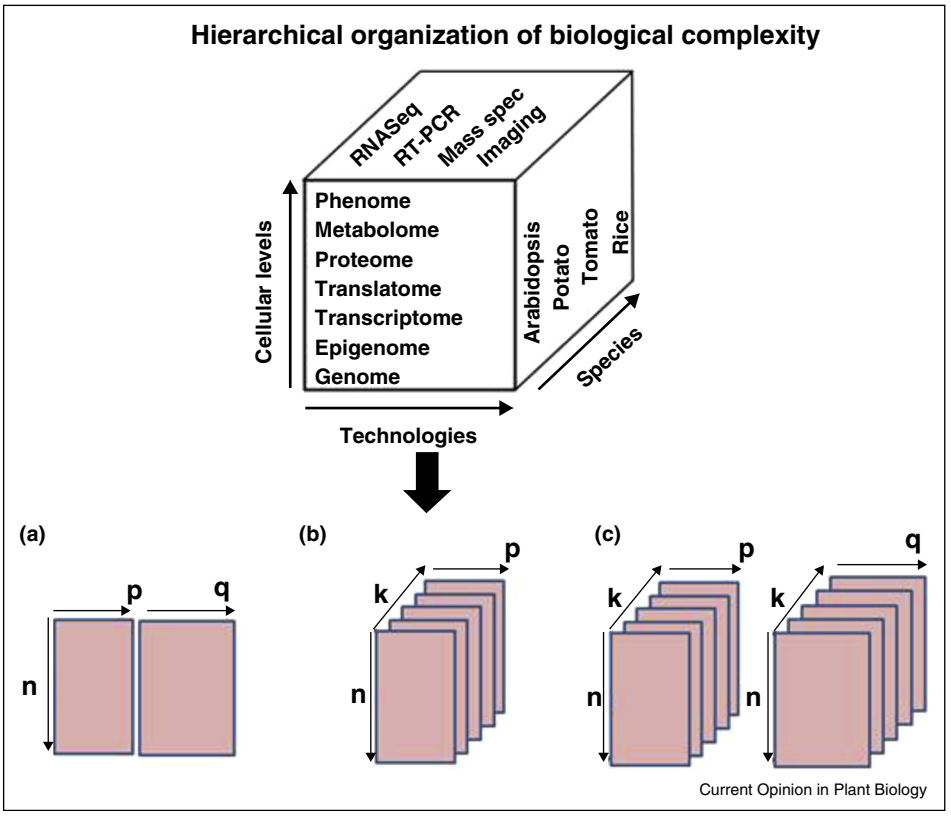
micro RNA (miRNA), proteins, and metabolite profiles that result in multiple levels of quantitative information. A typical integrative analysis scenario includes data from two system-levels, say transcriptome and proteome, or transcriptome and metabolome. There is a growing body of literature describing simple correlation methods such as Pearson or Spearman correlation coefficient for integrating and comparing data from two different system-levels [15]. However, a recent study by Rajasundaram *et al.* [16•] employed cell-type-specific data sets of the Arabidopsis root transcriptome and translatome for a systematic assessment of the degree of co-ordination and divergence between the two levels of cellular organization. The computational analysis considered correlation and variation of expression at the global and single cell level. Moreover, the authors provide insight into the degree of co-regulatory relationships that are preserved across the different system-levels using expression conservation scores. Through a series of Tukey Honest Significant Difference tests (Tukey HSD), the cell-type centric analysis elucidates whether transcriptional and translational patterns are conserved across multiple cell-types and are then displayed as network motifs. In addition, characterization of the biological processes of the genes identified in each step of the analysis was done by gene set enrichment analysis (GSEA). This example of integrative 'omics' analysis exemplifies a novel descriptive analysis pipeline implementing several statistical approaches.

In yet another study by Rajasundaram and colleagues [17], relationships between the polysaccharide (glycan) rich cell walls of cotton fibers and their phenotypic characteristics were established using data from the glycome and phenome level, respectively. Here, the authors employed canonical correlation analysis (CCA) to obtain a global view of association between the system-levels. Additionally, sparse partial least squares regression (sPLS) was used to be able to predict cell wall polysaccharides linked with fiber characteristics. With the use of predictive statistical approaches to integrate different 'omics' data sets, this analysis thus discovered correlations that are in line with already known biological functions and others for which the biological relevance is still to be tested. Such kinds of analysis in commercially important cotton lines help to provide insights into the developmental polysaccharides that are essential to obtain high quality fibers.

Integrative analysis of multiple 'omics' data sets

Majority of recent plant science studies investigate multiple 'omics' level in parallel and hence require methods to facilitate integration of multi-omics data sets. One of the applications of integrative 'omics' analysis using orthogonal partial least squares (O2PLS) multivariate regression method investigated different light condition

Figure 1



The complexity of biological data is multi-dimensional owing to advances in high-throughput technologies. Heterogeneity of the generated data is attributed to the measurement of different cellular levels using a wide range of techniques across various plant species. A schematic representation of possible types of data analysis problems is depicted. **(a)** Depicts the most common form of integrative analysis problems wherein two data sets with 'n' observations and variables 'p' and 'q' are analyzed. **(b)** Represents the case where there are 'k' data sets with the same 'n' observations and the same number of variables 'p'. This is an example of a typical descriptive type of integrative analysis. **(c)** Illustrates the situation where there are two sets of 'k' data sets where 'k' can pertain to time points or experimental conditions on 'n' observations with variables 'p' or 'q'.

induced effects on wildtype hybrid aspen measured in parallel for metabolite and transcript abundances [18•]. The authors identified transcripts and metabolites that exhibit strong multivariate correlation patterns, and the O2PLS method has the distinctive capability to identify unique and common variations in and between data sets. This was then further extended to handle multiple data sets and a study by Srivastava *et al.* [19••] proposed to use

Table 1

Some of the most commonly used descriptive and predictive integrative analysis methods with their corresponding R packages are listed here

Methods	R packages
Multiple canonical correspondence analysis	Vegan [6]
Canonical correlation analysis (CCA), regularized generalized canonical correlation analysis (RGCCA)	mixOmics [7], RGCCA [8]
Multiple co-inertia analysis (MCIA)	ADE4 [9], omicade4 [10]
Multiple factor analysis (MFA)	FactoMineR [11]
Principle component analysis extensions (STATIS, dual-STATIS, DISTATIS, ANISOSTATIS etc.)	ADE4 [9], MExPosition [12]
Partial triadic analysis (PTA)	ADE4 [9]
Partial least squares regression (PLSR), sparse partial least squares regression (sPLS), multiblock partial least squares regression (MBPLSR)	mixOmics [7], pls [13], ADE4 [9], PLS-2.1.0 [13]
Sparse regularized generalized canonical correlation analysis (SGCCA)	SGCCA [14], mixOmics [7]
K + 1 STATIS	ADE4 [9], MExPosition [12]
Multiblock redundancy analysis (mbRA), multiblock continuum redundancy (MCR)	ADE4 [9]

O2PLS for integration of transcriptomic, proteomic, and metabolomic data. Here, an understanding of system-level responses to oxidative stress response in plants was investigated. Furthermore, the identified genes, proteins, and metabolites and their associated pathways were visualized using Paintomics and MapMan [20,21]. These free softwares serve to map and visualize the genes, proteins, and metabolite measurements in Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (using Paintomics) and key affected processes not described by KEGG was illustrated using MapMan.

Integrative analysis of time resolved experiments to identify tightly regulated functions

In time-series analysis, estimation of the transition point in several stages is established to detect time lagged relationships in biological systems. Several methods such as correlation networks, clustering, and classification techniques have been proposed and most widely used to study time lagged profiles. However, there is an increasing need for application of methods which specify the extent to which each dimension of the data set reflects the perturbation. To this end, descriptive methods such as STATIS and variants of STATIS could serve as powerful tools depending on the number of data sets under study and the sampled time points (Figure 1). Recently, one such application of STATIS was demonstrated by Klie *et al.* [22] wherein they analyzed several transcriptome profiles from Arabidopsis over the same set of genes under varying experimental conditions sampled at several time points. The authors illustrated the time-resolved response of Arabidopsis to changing temperature conditions and identified components and pathways which could be under tight control in plant systems. Furthermore, methods such as multiblock partial least squares (mbPLS), and multiblock continuum redundancy (mbCR) has the potential to deal with regression problems involving data sets profiled across different time points, and assess the key drivers at the variable level.

Multi-faceted approach to understand functional variability in natural populations

One of the ambitious objectives of natural variation studies is to understand the genetic bases of complex traits with adaptive implications and how plants sustain the formation of ecotypes through ecological evolution. Dell'Acqua *et al.* [23] used a multi-faceted approach to exploit population genetics, landscape genomics, and genome wide association studies in *Brachypodium distachyon*. The benefits of this study are 3-fold: Firstly, 82 Brachypodium individuals sampled from nine different ecological locations have high intra-population homozygosity and a high-level of inter population genetic diversity; secondly, sampling locations were monitored using geographical information systems to obtain climatic data for each individual together with spatial distribution of genetic diversity; and finally, genotyping by sequencing approach provided a genome wide representation of molecular diversity in the collected individuals. The multi-faceted approach to investigate the structuration and diversity of the Brachypodium collection includes: Firstly, spatial pattern of genetic diversity was assessed using spatial principal component analysis (PCA) which enables the differentiation of global and local spatial structures; secondly, outlier detection was used to identify the loci with clear adaptive significance to climate; and finally, CCA was used to evaluate the relationship between climate gradients and molecular data. The joint analysis led to the discovery of 15 genes involved in *B. distachyon* adaptation. In addition, the authors also concluded that transposable elements were differentially distributed across the genomes of local groups and some with a pattern matching the climatic diversity of the sampling.

Conclusion and future directions

The systematic integrative analysis of heterogeneous data envisages the relationship between and within different biological layers for extensive knowledge discovery. Most of the plant related mechanisms and functions are very complex and vary among the same plant species, different tissues or even the same tissue at different developmental stages. Regardless of the biological question under analysis, some of the most commonly used data integration methods in plant biology are mainly designed to analyze two data sets at a time. However, the data explosion from large biological systems requires us to adopt effective integrative statistical approaches that can be extended to integrate and visualize several data sets at a time. Hence, it is essential to adopt a multi-disciplinary approach to yield unprecedented views on different aspect of plant systems.

Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7 2007-2013) under Grant Agreement number 263916. This paper reflects the author's views only. The European Community is not liable for any use that may be made of the information contained herein.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Fukushima A, Kanaya S, Nishida K: **Integrated network analysis and effective tools in plant systems biology**. *Plant Syst Synth Biol* 2014, **5**:598.

2. Fukushima A, Kusano M, Redestig H, Arita M, Saito K: **Integrated 'omics' approaches in plant systems biology**. *Curr Opin Chem Biol* 2009, **13**:532-538.

3. Usadel B, Obayashi T, Mutwil M, Giorgi FM, Bassel GW, Tanimoto M, Chow A, Steinhauser D, Persson S, Provart NJ: **Co-expression tools for plant biology: opportunities for hypothesis generation and caveats**. *Plant Cell Environ* 2009, **32**:1633-1651.

4. Wanichthanarak K, Fahrmann JF, Grapov D: **Genomic, proteomic, and metabolomic data integration strategies**. *Biomark Insights* 2015, **10**:1-6.

5. Ma C, Zhang HH, Wang X: **Machine learning for Big Data analytics in plants**. *Trends Plant Sci* 2014, **19**:798-808.

6. Dixon P: **VEGAN, a package of R functions for community ecology**. *J Veg Sci* 2009, **14**:927-930.

7. Gonzalez I, Le Cao KA, Dejean S: *mixOmics: 'omics' Data Integration Project*. 2011:. http://www.mixomics.org.

8. Tenenhaus A, Tenehaus M: **Regularized generalized canonical correlation analysis**. *Psychometrika* 2011, **2**:257-284.

9. Dray S, Dufour AB: **The ade4 package: implementing the duality diagram for ecologists**. *J Stat Softw* 2007, **22**:1-20.

10. Meng C, Culhane A, Gholami AM: **A multivariate approach to the integration of multi-omics data sets**. *BMC Bioinformatics* 2013, **15**:162.

11. Sebastian L, Josse J, Husson F: **FactoMineR: an R package for multivariate analysis**. *J Stat Softw* 2008, **25**:1-18.

12. Beaton D, Fatt CRC, Abdi H: **An ExPosition of multivariate analysis with the singular value decomposition in R**. *Comput Stat Data Anal* 2014, **72**:176-189.

13. Mevik BH, Wehrens R: **The PLS package: principal component and partial least squares regression in R**. *J Stat Softw* 2007, **18**:1-24.

14. Tenenhaus A, Philippe C, Guillemot V, Le Cao KA, Grill J, Frouin V: **Variable selection for generalized canonical correlation analysis**. *Biostatistics* 2014, **15**:569-583.

15. Vanholme R, Storme V, Vanholme B, Sundin L, Christensen JH, Goeminne G, Halpin C, Rohde A, Moreel K, Boerjan W: **A systems biology view of responses to lignin biosynthesis perturbations in *Arabidopsis***. *Plant Cell* 2012:24.

16. • Rajasundaram D, Selbig J, Persson S, Klie S: **Co-ordination and divergence of cell-specific transcription and translation of genes in arabidopsis root cells**. *Ann Bot* 2014, **114**:1109-1123.

Studying the root cells of *Arabidopsis thaliana* allowed positing a novel pipeline to systematically investigate and integrate the different levels of information available at the global and single-cell level. The conducted analysis also confirms that previously identified key transcriptional activators of secondary cell wall development display highly conserved patterns of transcription and translation across the investigated cell-types.

17. Rajasundaram D, Runavot J-L, Guo X, Willats WGT, Meulewaeter F, Selbig J: **Understanding the relationship between cotton fiber properties and non-cellulosic cell wall polysaccharides**. *PLOS ONE* 2014, **9**:e112168.

18. • Bylesjö M, Eriksson D, Kusano M, Moritz T, Trygg J: **Data integration in plant biology: the O2PLS method for combined modeling of transcript and metabolite data**. *Plant J Cell Mol Biol* 2007, **52**:1181-1191.

The O2PLS multivariate regression method can be used for combining 'omics' types of data. With this methodology, systematic variation that overlaps across analytical platforms can be separated from platform-specific systematic variation.

19. •• Srivastava V, Obudulu O, Bygdell J, Löfstedt T, Rydén P, Nilsson R, Ahnlund M, Johansson A, Jonsson P, Freyhult E *et al.*: **OnPLS integration of transcriptomic, proteomic and metabolomic data shows multi-level oxidative stress responses in the cambium of transgenic hipl-superoxide dismutase Populus plants**. *BMC Genomics* 2013, **14**:893.

The proposed data evaluation strategy shows an efficient way of compiling more than two data sets arising from multiple 'omics' platforms. OnPLS method proves to be efficient in handling complex multi-omics data sets and will serve to be a base study for integration of 3 data sets at a time.

20. Garcia-Alcalde F, Garcia-Lopez F, Dopazo J, Conesa A: **Paintomics: a web based tool for the joint visualization of transcriptomics and metabolomics data**. *Bioinformatics* 2011, **27**:137-139.

21. Thimm O, Blasing O, Gibon Y, Nagel A, Meyer S, Kruger P, Selbig J, Muller LA, Rhee ASY, Stitt M: **MapMan: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes**. *Plant J* 2004, **37**:914-939.

22. Klie S, Caldana C, Nikoloski Z: **Compromise of multiple time-resolved transcriptomics experiments identifies tightly regulated functions**. *Front Plant Sci* 2012:3.

23. Dell'Acqua M, Zuccolo A, Tuna M, Gianfranceschi L, Pè ME: **Targeting environmental adaptation in the monocot model *Brachypodium distachyon*: a multi-faceted approach**. *BMC Genomics* 2014, **15**.

24. •• Abdi H, Williams LJ, Valentin D, Bennani-Dosse M: **STATIS and DISTATIS: optimum multitable principal component analysis and three way metric multidimensional scaling**. *Wiley Interdisc Rev Comput Stat* 2012, **4**:124-167.

This excellent review provides a detailed account of the mathematical aspects of the available multiblock methods. Many different multi-block methods exist in literature across multiple disciplines, but locating these and particularly to match the most appropriate method to answer a research question present challenges. A comprehensive manual of different methods is provided, outlining a description of the methods, preprocessing strategies used, as well as robustness and stability issues.