Review of statistical learning methods in integrated Omics studies
(An integrated information science)

Irene Suilan, Zeng and Thomas, Lumley
Department of Statistics,
Faculty of Science,
University of Auckland
New Zealand

Key words: statistical learnings, integrated Omics, exploratory learning, regression, network learning
Abstract

Integrated Omics is becoming a new channel for investigating the complex molecular system in modern biological science and sets a foundation for systematic learning for precision medicine. The statistical/machine learning methods that have emerged in the last decade for Integrated Omics are not only innovative but also multidisciplinary with integrated knowledge in biology, medicine, statistics, machine learning and artificial intelligence. Here we review the non-trivial classes of learning methods from the statistical aspects and streamline these learning methods within the statistical learning framework. The intriguing findings from the review are that the methods used are generalizable to other disciplines with complex systematic structure, and the integrated Omics is part of an integrated information science which has collated and integrated different types of information for inferences and decision makings.

We review the statistical learning methods of exploratory and supervised learning from 39 publications. We also discuss the computational method including a comparison between Bayesian and non-Bayesian approach, the sparsity issues and regularization approach to dealing with sparsity and singularity when there are fewer observations than the number of features (variables). Lastly, we provide insights in the meta-analysis method used in the integrated omics science. For the completeness of the review, a table of currently available software and packages from 23 publications for Omics are summarized in the appendix.

Exploratory methods

Dimension reduction

Dimension reduction is an important multivariate statistical technique; it is used to identify latent structure which is not observable but presented in the observations that are results of these structures. The number of dimensions or factors of the latent structure needs to be less than the number of variables, and the groupings of variables or weighted combinations of all variables are the statistical representations defined for the latent structure.

Principal component and factor analyses are two elementary statistical techniques for dimension reduction. In the literature for integrated omics, dimension reduction methods have presented several variations from principal components analysis and factor analysis. These variations include multiple factor analysis (MFA) (Sanchez, et al. 2012), consensus PCA (CPCA) and multiple block principal component analysis (MBPCA) (Hassani, et al. 2013). It also presents an extension of factor analysis Non-negative matrix factorization (Yang and Michailidis 2015).

Principal components analysis and its variations
Hassani, et al. 2010 firstly introduced a consensus principal components analysis method (CPCA) for multiple Omic datasets referred as blocks and three validation tools in 2010. The block represents one type of omics measurement, and multiple blocks are collected from same biological samples. Authors use the genetic fingerprinting data and metabolite fingerprinting Fourier Transform Infrared (FTIR) spectra as an example which subdivides spectra into blocks of polysaccharide region and fingerprint region, protein region and the fatty acid regions.

CPAC utilizes the algorithm NIPALS (Miyashita, et al. 1990, Wold 1966) to identify the latent structure from the combined measurement data. NIPALS is presented to identify latent structure parameters including the block and global loading scores, block scores and global scores iteratively (Figure 1.). To reduce the subjectivity of exploratory analysis, authors present three validation methods, Root Mean Square Error (RMSE), uncertainty $t$-test and stability plot. RMSE is used to identify the correct number of PCA through the RSME plot; uncertainty $t$-test which uses a $t$-statistics estimated from loading coefficients assesses if the measurement significantly contributes to the CPCA; stability plot, which assesses any outlying observations, are proposed at the block level and the global level.

In 2013, the same group of authors (Hassani, et al. 2013) introduced the concepts of multi-block principal component analysis and compared three different deflations strategies (an approach to derive the structure of data represented by a vector such as the Eigenvectors (Mavroeidis and Marchiori 2011) for multi-block principal component analysis). The multi-block principal component analysis is an iterative algorithm using multiple PCAs to find the optimal latent components. Deflation occurs by subtracting the data values from the predicted values and the resultant residual values are used for the next iteration to derive new principal components. Conesa, et al. 2010 propose a multiway approach to identify the underlying components that interconnect with different omics variables. They use a dimension reducing technique TUCKERS method for intra-omics analysis and the N-partial least squares (N-PLS) for inter-omics analysis. The different omics datasets comprise functional genomics measurements of transcriptomics, metabolomics and physiological datasets. TUCKER3 method is suggested to be an appealing data integration strategy because it can accommodate the structure of the data from a multifactorial design experiment (i.e. time x treatment x protein expression), and N-PLS can infer the relationships between biomolecular measurements in multidimensional space.

Factor analysis and its variations

In contrast to principal component analysis which projects the observations into the new latent structuralized space, factor analysis identifies the latent structures that can be used to form (or explain) the observed data. Sanchez, et al. 2012 introduce Multiple Factor Analysis (MFA) that can be used to reduce dimension and integrate supplementary information with the original omics datasets in a common space. MFA starts from a PCA on each block (type) of data and followed
by jointly analyzing the singular-value normalized data using the global PCA. The normalized singular value represents the square root of the first Eigenvalue. Sanchez, et al. 2012 suggest that using MFA is expected to avoid the $n<<p$ problem and is suitable for different types of omics datasets.

**Non-negative matrix factorization and others**

As a dimension reduction technique, conventional NMF method decomposes the data matrix by using a latent factor Matrix $W$ and a basic component data matrix. NMF is similar to PCA, but using non-negative constraints instead of orthogonality constraints. Its solution is less uniquely defined but more interpretable for the non-negative omics measurements. Yang and Michailidis 2015 introduced an integrated non-negative matrix factorization method (iNMF) to handle the heterogeneous multiple omics datasets and reduce the overall dimensions. A joint conventional NMF decomposes $m$ multiple non-negative data matrices by using the non-negative common latent factor matrix $W$ and $m$ basic non-negative component data matrix $H$, assuming the $m$ datasets have common latent structures. Examples of the non-negative data matrix can be miRNA and gene expression. "iNMF" adds the variation in the latent factor matrix $W$, and uses a penalty term to control the variation for latent factor matrix across different data matrices. In contrast to the "orthogonality" constraints approach used in PCA, partial least squares and canonical correlations analysis which maintains the center of mass, the "iNMF" uses constraint over "non-negativity" for a better interpretation. Both approaches are to identify the best approximations for the original datasets. iNMF has also been extended to cope with sparsity using a sparsity parameter in the penalty term. These methods were proposed for expression datasets with continuous measurements.

The other streamlines in dimension reduction include Serra, et al. 2015 who combine dimension reduction and cluster analysis to multiple genomic datasets. The algorithm involves prototype extraction and ranking which aims to reduce dimension by filtering variables using variance, and rank the prototype based on their abilities to separate classes. As an application of using different dimension reduction techniques, Su, et al. 2015 presents a method to integrate image-omics and functional omics data for the classification of breast cancer staging. They demonstrate an improvement of 3% in classifications using the integrated data compared to using the image-omic data only.

**Clustering methods**

In Integrated Omics, clustering methods appear to be the commonly used approach for subjects or features partitioning. It is a useful tool to provide exploratory view of the underlying clusters pattern. The dataset from multi-Omics can have a complex data topology; new strategies are required to identify the
partitioning structure of the integrated information. Apart from the conventional clustering approach utilizing different distance measures, newly proposed methods use maximum likelihood method and some include penalized terms to control for complexity in feature selections. Among these studies, Newman and Cooper 2010, Aibar, et al. 2015 introduce and modify the stochastic clustering method: self-Organizing map (SOM) that has been used in Cartography in geography. Shen, et al. 2013 and Kim 2015 used the latent variable approach with penalty terms to optimize the likelihood for cluster memberships. Sharma, et al. 2016 use iterative maximized likelihood method to cluster both categorical and continuous variables.

**Iterative maximum likelihood based approaches**

Newman and Cooper 2010 presented an unsupervised clustering technique which bases on the Self-Organizing Map (SOM) (Figure 2), a stochastic clustering method to reduce the number of dimensions and preserve the local topology of gene expressions. Initial SOM measures the similarity of adjacent nodes, and derive the dissimilarity surface (error matrix). The error matrix is used to identify borders of clusters and group similar data points and separate dissimilar data points iteratively. The AutoSome method uses density equalization, which is a technique of cartography, to ensemble these graphical features output from SOM, and to rescale the SOM output lattices. The density equalization treats nodes of high errors with high density and forces these nodes separating from each other; conversely, it treats nodes of low errors with low density and aggregates them. A minimal spanning tree method is then built from the rescaled nodes to identify the final clusters solution. Using the similar approach, Aibar, et al. 2015 applied the SOM in transcriptomics samples from three real datasets: Myelodysplastic Syndrome (MDS), Alzheimer’s disease and colorectal cancer, to classify patients from different disease stages.

Sharma, et al. 2016 propose a maximum likelihood based clustering approach that can be applied to both categorical and continuous data. In system biology, this method can be applied to microarray expression and SNP data. It uses stepwise iterative method to identify the optimal solution that maximizes the likelihood for the n class clusters following the data topology. The iterative algorithm includes: Initialize the cluster members, shift one sample from one cluster to another and recalculate the total likelihood of m clusters based on the new mean and covariance matrices of each cluster. The proposed likelihood method utilizes both the distance measures and variance components in the samples.

**Regularization based methods to control for complexity in feature selections**

Regularization or Penalty constraints are one common approach in statistical modeling for controlling complexity and achieve precision when the number of observations is far beyond the number of features, or when the real associations between molecular features are known to be much smaller than all the possible associations.
Shen, et al. 2013 propose a penalty based clustering method (iCluster) to identify the number of clusters and membership of clusters for the Integrated genetic and genomic features (copy number variation, DNA methylation, SNP). The main idea is to treat the latent variables of clusters as missing information and use expectation and maximization algorithm to estimate parameters of the penalized complete data likelihood. The penalty term induces sparsity in the weighting matrix for the latent variables and achieves simplicity of the clusters. The paper introduces three types of penalty functions, namely, lasso, elastic net and fused lasso to control the number of clusters.

Kim 2015 proposes group penalty method for group structured and tight integrative clustering in which group lasso is presented as an updated version of icluster (Shen, et al. 2013). Under the penalized regression framework, the joint penalty complete log-likelihood was extended by adding a group lasso penalty term. Because it is possible that multiple feature modules share the same feature, for example, two miRNA regulate the same gene. The group lasso regularization, which is based on multiple feature modules, contains overlapped features (i.e. mRNA, CNV, two methylations) and maintains the biological information in the model building.

Chi, et al. 2017 create a convex biclustering method to partition samples and features under a regulation penalty path. They utilize the distance-based measurements for clusters and iteratively shrink both the column (features) and samples (rows) simultaneously. The biclustering method is motivated by solving problems in the high-dimensional genome data and can be extended to use in the omics study for 2-dimension partition problems.

**Network learning methods**

Networks composing of nodes and arcs provide an advanced tool to demonstrate the interactions between large numbers of variables (molecular features) in Integrated Omics. In network learning theories, variables are presented as nodes, causal relation or associations are presented as the arcs or edges between nodes. The graphical model and Bayesian network provides probabilistic estimates between nodes in these networks (Alpaydin 2010). The learning methods for causal and conditional dependent networks can be used to investigate the multilayer associations and causal relations between Omics features in integrated Omics studies. When the causal relations are not the focus, matrix based statistics are used to measure the associations between the linked datasets. The existing method for omics datasets includes canonical correlation and RV. Developments in matrix statistics for integrated omics blossomed in the last decade include the maximal first order partial correlation coefficients (MF-PCcor) and adjusted RV.

*Estimating associations between Omics datasets*
Kayano, et al. 2013 introduce ranking-based maximal first-order partial-correlation coefficients (MF-PCcor) to estimate the associations within the metabolite network and cope with outlying samples. The partial correlation coefficient bases on the normalized rank of the expression data and the maximal first-order partial correlation estimates the edges between metabolites.

Mayer CD and colleagues present an unbiased estimate of matrix correlation-adjusted RV coefficient (Mayer, et al. 2011). RV was originally used as a similarity index for two matrices, it is a generalization of the correlation from two variables to two datasets. It is the ratio between trace of cross-correlation matrices’ product and trace of squared correlation matrices’ product:

\[ RV(X, Y) = \frac{\text{tr}(R_{XX}R_{XY}^T)}{\text{tr}(R_{XX}^2R_{YY}^2)} \]

The adjusted RV, an unbiased estimate of the matrix correlation, replacing the squared correlation with the conventional adjusted R squares in linear regression. The adjusted RV is applied to multiple system biology datasets for the identification of biologically meaningful subgroups and can be used as the input for clustering and multi-scaling analysis.

**Estimating structure of multilayer networks formed by integrated Omics datasets**

Angione, et al. 2016 introduce the multi-layer network (Multiplex) method for the integrated omics data. It is known that iCluster and similarity network fusion (SNF) method are not designed for the analysis of cross-omics data (Angione, et al. 2016). iCLUSTER does not scale all measurements and needs pre-selection of genes. SNF only creating aggregated layers from genes. Angione proposes a method to model the linkage between Genotype and Phenotypes. It constitutes multiplex networks of transcriptomics and fluxomics (a duplex) and fuses the two networks into one using a weighting network fusion approach. The proposed method uses a linear program to map the gene expression onto the metabolite model. Network with two layers is constructed with nodes representing environmental condition and edges representing similarity between nodes regarding genes or metabolites expression. The final derived single network is used to identify clusters of conditions with similarities. The weighted fusion approach of multiplex networks uses the weight to reflect the importance of gene or metabolite to the nodes (environmental conditions). Figure 3 provides the visualization map of the multiplex fusion algorithm.

Mosca and Milanesi 2013 present a network analysis method similar to Angione, et al. 2016, to integrate biological components and their interactions from multiple omics datasets. They propose to use molecular interactions and multiple objectives (MO) for the simultaneous optimization, based on statistical criteria at the network level and component level. Different statistical criteria are set for different objective functions in the multiple objective optimizations. Of these
criteria, hypervolume indicator, which presents the volume of the dominated portion (suboptimal points) of the objective space, is used as the quality measure of MO optimization process. The introduced algorithm integrates a weighted network from multiple omics datasets and optimizes the weighted networks. Cun and Frohlich 2014 presents netClass method of joining networks using smoothing approach. It uses smoothing method (kernel based smoothing network diffusion) on the feature-wise marginal statistics over the structure of a joint protein-protein and miRNA-target gene interaction graph. Random walk kernel is used for smoothing and a permutation test is used to select features of each dataset. The package provides an analytical tool to integrate miRNA and mRNA expression data, with protein-protein interactions and miRNA-target gene information.

Apart from developing new learning methods, some studies applied the existing methods into integrated omics. One typical study of these applications is Peñagaricano, et al. 2015 who applied Bayesian network (R package bnlearn) to explore the causal networks underlying fat deposition and muscularity in pigs, using genotype, transcriptomic and phenotype datasets. The study group introduces an integrated analysis using marginal associations between genotypic and phenotypic traits (genotype and phenotype data) via pQTL, marginal associations between genotypic and expression traits (genotype and transcript expression mRNA data) via eQTL, and identify the co-localized joint significant eQTL and pQTL from the mapping analysis. They provide a summary of several methods to infer the causal genotype-phenotype network. One of the causal structure learning techniques is the inductive causation algorithm (IC) and its extended version Incremental Association Markov Blanket (IAMB). IC starts with determining a pair of variable (A and B)'s conditional associations given all other variables, by searching any possible subsets of other variables as the dependency set; it follows by the second set of conditional independent tests including the adjacent variable C of A and B. The resultant partially directed graph are then filled with undirected edges as many as possible so long as that there is no new V structure and new directed edge formed. The Extended version of the IC- IAMB algorithm includes a screening process to identify the Markov blanket of every variable X. IAMB involves a set of conditional independent test for a pair of variable X and Y given subset W; it reduces the computation complexity without compromising accuracy.

There are study groups only providing tools for building a network and visualizing these networks. Appendix one includes a summary of these tools. One example is the BisoGenet (Martin, et al. 2015), which is a network building tool assigning biological functional relations of protein and protein, protein and genes, based on a local in-house database “SysBiomics”. This server provides network building and visualization functions, given input entities nodes and edges.

**Regression-based methods**
In the integrated omics literature, the regression equations are set for explaining inter or intra-system relation and interactions. The strategies of parallel or sequential regressions are sometimes used with constraints. Parallel regressions is chosen to model causal relations between multiple molecular responses (i.e. metabolites and genes) on continuous or categorical scale and their interacting effects as well as factors of interests, i.e. pathway membership. Multivariate responses technique is not suitable due to the necessity of including inter-response relations in the explanatory factors of these models. One example of these inter-response relation is, an active pathway membership of gene affects metabolites involved in the same pathway.

**Parallel regressions**

The parallel regressions are used in different Omics responses to explain intersystem responses simultaneously. One example is the model proposed by Jauhiainen A. et al to integrate transcriptomic and metabolomics data to make an informed pathway-level decision(Jauhiainen, et al. 2012). Authors propose two linear models to describe responses of the gene and metabolite expression on pathway memberships. The fixed and random effect metabolite linear models include the pathway membership of gene presented by the regression coefficients from its parallel linear model; the mixed model includes random effects on the metabolite level. The random term allows the effects from unselected genes in the pathway being measured as these genes could post effects on the metabolite even if they are not selected at the gene level. The model selection occurs at two levels, firstly to select differentially expressed genes and subsequently which genes are allowed to influence the metabolites expression. Secondly, on the global pathway level to pick out the active pathways.

Poisson, et al. 2011 introduced two joint tests for gene expression and metabolite information using two parallel logistics regressions. The gene expression and metabolite information are fitted in separate logistic regression, both of which predict the probability being in the interested gene or metabolite set S. The first test involves a 2-degree of freedom Wald test on the resultant regression coefficients. The second test is an enrichment test statistics using the sum of square statistics for gene and metabolite which are constructed as a 2-dimensional vector$(W^G_S, W^M_S)$, by permutation. A similar enrichment strategy was given by Pey J et al (Pey, et al. 2013) who used an optimized pathway analysis model enriched by the classification based on up-or-down regulated gene/protein expression. The optimization is divided into three stages to minimize the associations between flux and reactions in the classes. Results of gene expression measured by transcriptomics and protein data measured by proteomics are used to infer the forming of pathways.

**Sequential regressions**

Acharjee, et al. 2014 present a sequential analytical approach starting from using the random forest to screen variables from individual “omic” dataset, followed by
further selection of the redundant variables via eQTL (quantity trait linkage). One advantage of the study is using the well-known regulatory genetic and metabolic pathway to validate the method. The method in the study is applied to transcriptomic (mRNA), proteomic (2d GEL) and metabolomics (LC - and GC- MS) data. Sequentially, the model starting with a random forest method implemented by R package randomForest, authors propose a permutation test to determine the metabolite/protein/RNA significance for predicting the trait. Secondly, the integrated linkage map is used and implemented via R package metanetwork. Lastly, the final selected gene, protein and metabolites including the trait are used to construct the network. The network’s nodes are formed by the aforementioned molecules and traits, and the edge are representing the strength of the interactions measured by regularized partial correlations.

**Partial least squares (PLS)**
Partial least squares is a multivariate technique used to identify latent structures of both predictors and responses by maximizing the covariance between them. It is widely used in the integrated omics study. Since S. Wold introduced the NIPALS algorithm for PCA and Partial least squares (PLS) in chemometrics in the 80’S, NIPALS became the popular computer algorithm for PLS. Lê Cao, et al. 2008 proposed the sparse PLS by utilizing lasso penalization in 2008 for Integrated Omics. Sparse PLS (SPLS) optimizes the square error terms with a penalty term of loading vectors of response matrix Y and predictor matrix X.

Fonville, et al. 2010 introduced the extended orthogonal signal correction (OSC) PLS (O-PLS) 2010. It weights the predictor variables by using the orthogonal components in the covariance matrix between the response (Y) and the predictor variables (X). O-PLS filters out the “structure noise” bases on the covariance matrix for Y and X. It becomes one of the popular approaches in metabolomics due to its easy interpretation. A similar idea named N-PLS was also given by Conesa, et al. 2010, N-partial least squares (N-PLS) construct data array responses (Y) from multiple Omics platform and the predictor data blocks (X) curated from another type of Omics platform in the multifactorial (N) spaces. It finds latent spaces that can maximize the covariance between X and Y and decomposing X from the improved version. Authors proposed a gene selection procedure using a gene associated parameter p that reflects the contribution of each gene (Figure 4).

Chen and Li 2016 present three stochastic discrete dynamic equations to describe the relations among gene, proteins, miRNAs and DNA methylation. These stochastic dynamic equations provide quantitative predictions of measurements of message RNA, micro RNA, and protein expression at a specific time point. The quantitative measurements involve their expression level at time t, interactions respectively for miRNA-mRNA, protein-protein, and the degradation of mRNA, as well as rate of miRNA- mRNA coupling. These stochastic equations describe the inter-molecular relations included in protein-protein interaction, miRNA and gene
regulatory network, and the measurement errors. In addition to the three stochastic equations, an extra equation for path gene protein is added to construct the integrated genetic and epigenetic cellular network (IGECN). The regulatory and interactive parameters included in these four dynamic equations are evaluated by using temporal data and solved by the constrained least square parameter estimation problem.

Another example is given by Pavel, et al. 2016 who integrate three types of molecular data: mutation, copy number variation and gene expression via a fuzzy system score for each gene and sample. Biological rules are created based on the defined categories of these three molecular data sets. A fuzzy logic modelling is used to cluster and subtype discovery, and to recover many known suppressor genes and oncogenes and subtypes in colorectal cancer cells.

**Biological knowledge Enrichment learning**

As defined by machine learning literature, supervised learning method uses response variable and training data or prior knowledge, to provide a prediction for response variable. In statistical learning, the prior knowledge can either be used to set prior in the Bayesian model or inform the model selection. Bayesian modelling provides the essential framework to incorporate know information in analysis. It is called supervised learning in the context of prediction, because the “true” value \( Y \) is part of the training data. In the context of estimating causal relationships between omics variables, however, the value of \( Y \) is not the goal of the analysis and these do not involve knowing the true causal relations.

Pavel, et al. 2016 give examples of using the biological knowledge for forming biological rules in the cluster. Poisson, et al. 2011 introduced enrichment tests learned by biological knowledge to jointly evaluate gene expression and metabolite abundance. Nguyena and Hob 2012 proposed a semi-supervised machine learning method to identify disease-related genes via the publically available database. The method starts with identified disease-related proteins from the public databases which provide known biological information for the proteins to be analyzed. The included database are UniProt, Gene Ontology, Pfam, Interdom, Reactome, and gene expression. The proteins of interests are divided into two groups according to the integrated information from these databases: disease related or not related; After the division, data are extracted and pre-processed according to the feature functions, namely, the protein sequence length, keywords appeared in the database related to each protein, enzyme function, protein interaction with disease, protein pathways involvements, protein domain involvements and DDI (interDom) involvements. The final procedure uses the Gaussian random field and harmonic functions to learn a new set of the disease gene. Gomez-Cabrero, et al. 2015 uses rank statistics to identify the most correlated gene markers of the comorbidities of COPD patients via public data and then introduced the Relative risk (RR) and correlations for binary variables to cluster the disease. Kamburov, et al. 2011 present a web-based tool IMPaLA for
joint pathway analysis of transcriptomic or proteomic and metabolomics data from multiple datasets. Joint P-value is given for multiple datasets comprising metabolites and genes/proteins in the learning.

**Discussion**

Computation (using Bayesian versus not using Bayesian)

In machine learning literature, supervised learning includes utilizing Bayesian statistical approaches to integrating prior knowledge in the current observations. Bayesian network is another way of using prior information. Sharma, et al. 2016 uses the prior probability of cluster belongings in their iterative maximal likelihood algorithm for estimating the posterior probability of clusters membership. Although their computation method does not use the classical posterior samplings (i.e. MCMC approach), they have employed the Newton-Rapson gradient ascending method to find the optimal estimates which have integrated the priors information. iCluster is another example of using Bayesian approach for identifying cluster membership but using EM algorithm in the computation, icluster (Shen, et al. 2013) requires prior knowledge of the number of clusters. MDI (Kirk, et al. 2012), uses the multinomial mixture model which requires prior knowledge of mixture probability, and utilize the Gibb samplings.

Bayesian network combines network theories and probability models. The required priors are normally the prior distributions of gene activities. Paradim (Vaske, et al. 2010) and Conexic (Akavia, et al. 2010) are two algorithms that utilize Bayesian network and Paradim uses EM algorithm in the computation of the unknown factor graph parameters.

Bayesian method is essential as a knowledge driven approach when there are known prior knowledge (i.e. pathway or network structure) (Ahmad and Fröhlich 2016), compared to the data-driven method.

Sparsity and regularization

Sparsity commonly exists in the models of integrated omics datasets. The sparsity is one natural feature in biological pathway or function network which are caused by only a few real existing relations between different molecules. On the other hand, the sparsity of the associations makes the modeling possible when the number of the subjects is smaller than the number of features. When there is no prior knowledge available about the associations in the dataset, regularization is a mathematical method to obtain a balanced solution to overfitting with a favorable computing speed. It uses penalized approach, i.e. regression with additional penalty term such as the absolute value ($l_1$) or the combination of $l_1$ and sum of squares. Pineda, et al. 2015 applied the penalized regression in combined omics datasets that has genetic variants, DNA methylation, and gene expressions. The penalized regressions were useful to investigate if the changes
in gene expression were modified and or confounded by genetic variants or DND methylations given the known facts that most of the associations are null or not relevant. There are several other integrated omics studies that have applied the penalized regressions (Ahmad and Fröhlich 2016). Using penalized regression yields a higher precision. It provides approximations to the regression coefficients of the predictors.

Meta-analysis

Meta-analysis is used as an approach for data integration on the summarized statistics level, it is also used as a tool to combine with analysis integrated from different studies on the level of individual observation which is called the mixed approach (Sanchez, et al. 2012). In the latter application, meta-analysis is used to provide information to refine the analysis of a new study (Kannan, et al. 2016). The semi-supervised method proposed by Nguyena T and Ho T-B summarized in section 2.3 is one example of the mixed approach which uses meta-analysis to provide prior biological information from multiple publically available databases to update results in the later analysis. These meta-analyses utilize both statistical and biological inputs in the integration. Kim 2015 extends a meta-analytical framework for Principal Component Analysis (PCA). The aim of using the meta-analytical framework of PCA is to utilize multiple datasets to provide the common PCs. Two methods are presented to summarize the common PCs: 1) decomposition of sum-of-variance-decomposition and 2) minimization of sum-of-squared cosine-(SSC)-maximization. Sum-of-variance-decomposition uses the weighted sum of covariance matrices from \( m \) datasets to find the common Eigenvector matrix. Minimization of sum-of-squared cosine-(SSC)-maximization uses \( m \) Eigenvectors derived from the multiple datasets (studies) to form an Eigenvector matrix.

The publically available software packages for meta-analysis includes CNAmet (Louhimo and Hautaniemi 2011), Rtopper (Tyekucheva, et al. 2011), iClusterPlus (Mo. Q and Shen 2016) and the STATegra (Consortia 2017) Bioconductor package.

Closing remarks

The presented methods for integrated Omics are not only innovative but also diverse. The selections of analytical techniques are primarily determined by the research questions sought to answer. New methods are created for providing better strategies to integrate different omics measurements from different technical platforms that have both inter and intra platforms variabilities. Streamline of these methods give us a clear vision of how the statistical framework has been built and extended to cohesive with other sciences in analytical applications in Omics studies. Future research requires more uniformed structure and methods in networks estimation and prediction for mixed types of measurements, and more applications in precision medicines.
Acknowledgement
The authors would like to express their gratitude to Ms Vivian Ward who helped us to visualize the analytical processes for the review.

References

Acharjee A, Bjorn Kloosterman B, Visser R and Maliepaard C (2014) Integration of multi-omics data for prediction of phenotypic traits using random forest. Statistical Methods for Omics Data Integration and Analysis, Heraklion, Crete, Greece
Hassani S, Martens H, Qannari E, Hanafi M, Borge G and Kohler A (2010) Analysis of -omics data: Graphical interpretation- and validation tools in multi-


Multi-block principal component analysis

(a) Metabolite
Protein expression
Gene expression

(b) Metabolite
Protein expression
Gene expression

X

\( X \tilde{P}_i^b \)

(new block score matrix T)

(c) \( \left( \begin{array}{c} T_1^T \\ \vdots \\ T_n^T \end{array} \right) \)

(d) \( T \times \tilde{W} \rightarrow \tilde{t} \) New global score for the next iteration

regress

Normalised to length 1

block loadings

block W (normalised to 1)
Self Organised Map (SOM)

(a) Dense network

(b) U-Matrix

Nodes

I  K

\[ \begin{array}{ccc}
     I & U_{1,2} & U_{1,3} & \ldots & U_{1,K} \\
     U_{2,1} & U_{2,2} & \ldots & U_{2,K} \\
     U_{3,1} & U_{3,2} & \ldots & \ldots \\
    \vdots & \vdots & \ddots & \vdots \\
     U_{K,1} & \ldots & \ldots & \ldots \\
\end{array} \]

Dissimilarity between adjacent nodes

Density-equalising cartograms

(c) Low error

Low error

High error

Low

High

(d) Minimised spanning tree to reduce complexity

Simplified nodes of tree
Multiplex fusion (Similarity network fusion)
N-Partial Least Squares (N-PLS)

Gene expression

Metabolite expression

Physiology expression

NPLS1

NPLS2

NPLS3
<table>
<thead>
<tr>
<th>Names</th>
<th>Language used</th>
<th>Analytical functions</th>
<th>Include Visualization</th>
<th>Provide Public Databases</th>
<th>Omics techniques*</th>
<th>Designed for integrative omics analysis ††</th>
<th>Involved statistical models</th>
<th>Involved for human study</th>
<th>Design ed for human study</th>
<th>Web-base open source</th>
</tr>
</thead>
<tbody>
<tr>
<td>OmicKriging** (Wheeler, et al. 2013)</td>
<td>R</td>
<td>It is designed for predicting complex traits (quantitative and qualitative) by leveraging and integrating similarity in genetic and large scale omics.</td>
<td>No</td>
<td>NA</td>
<td>miRNA, mRNA, T, SNP and other large-scale omics</td>
<td>Yes (subject level)</td>
<td>Yes. It uses an algorithm to optimize the composited similarity matrix which integrates different omics correlation matrices.</td>
<td>Yes</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>TranSMART+Gala xy+MINERVA: a combination pipeline** (Satagopam, et al. 2016)</td>
<td>NA</td>
<td>TranSMART repository provides integration of low dimensional clinical data and high dimensional molecular datasets, with built-in data mining and analysis applications. Galaxy provides analytical pipeline; and MINERVAR provides conceptualized and visualization for molecular interaction networks</td>
<td>Yes</td>
<td>eTRIKS, <a href="http://www.etriks.o">www.etriks.o</a> rg</td>
<td>GE,T,P,M (not specified)</td>
<td>Yes (subject level)</td>
<td>Yes. Galaxy uses the R Bioconductor packages limma.</td>
<td>Yes</td>
<td>Yes. Galaxy is a web server and cloud bench</td>
<td></td>
</tr>
<tr>
<td>OmicsAnalyzer (Stoltmann, et al. 2013)</td>
<td>JAVA</td>
<td>As a plug in for cystoscope, it has the functions of mapping different datasets, estimating associations and visualizations.</td>
<td>Yes</td>
<td>No</td>
<td>NA</td>
<td>Yes (molecular ID level)</td>
<td>Yes</td>
<td>Not specified</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>VANTED (Rohn, et al. 2012)</td>
<td>JAVA</td>
<td>VANTED is a framework providing essential functions for system biology. It has 7 tasks including data integration, visualization and data analysis for correlation, clustering, differential analysis and enrichment analysis. It also computes some topological features for the network.</td>
<td>Yes</td>
<td>Connect to network database: MetaCrop, KEGG, RIMAS</td>
<td>Not specified</td>
<td>Yes (molecular ID level)</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>The DNA Microarray Interomics Analysis Platform (Waller, et al. 2015)</td>
<td>R</td>
<td>It provides data process function and focuses on the integration of these two types: 1) lipidomics and transcriptomics; 2) mi-RNA and mRNA; The integration analysis includes mi-RNA and mRNA interaction</td>
<td>Yes</td>
<td></td>
<td>murine nutrigenomics dataset; Normal Human Dermal Fibroblasts (NHDF)</td>
<td>GE, miRNA, T,L, miRNA-mRNA interaction</td>
<td>Yes (subject level)</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td><strong>Lemon-Tree</strong>  (Bonnet, et al. 2015)</td>
<td><strong>JAVA</strong></td>
<td>It is a modular network software. It provides a function (ganesh) for model-base Gibb sampler to infer co-expression modules and condition clusters within each modular. It also provides regulator program that forms decision trees with nodes of the regulator at the expression level. Function tight-cluster will build clusters formed by consensus modules of genes.</td>
<td>Yes</td>
<td>TCGA glioblastoma expression and copy-number data</td>
<td>Yes (subject level)</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
<td></td>
</tr>
<tr>
<td>**integreOomics **  (Lê Cao, et al. 2009)</td>
<td><strong>R</strong></td>
<td>It provides regularized canonical correlation analysis, Sparse Partial Least squares regression</td>
<td>Yes</td>
<td>No</td>
<td>M,L,C</td>
<td>Yes (subject level)</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td><strong>Mayday SeaSight</strong>  (Battke and Nieselt 2011)</td>
<td><strong>JAVA with an built-in R terminal</strong></td>
<td>Mayday has the daily used methods for array analysis. It includes cluster, differentiation analysis, and machine learning methods. It also has a terminal connection with R which facilitates usage of R functions.</td>
<td>Yes</td>
<td>KEGG, MetaCyc</td>
<td>GE,T</td>
<td>SeaSight provides the integrative function for GE and next generation sequence data (at the experiment level)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td><strong>DASS-GUI</strong>  (Hollunder, et al. 2010)</td>
<td><strong>C++</strong></td>
<td>It provides two modes: 1) Calculation mode: DASS.cs which identifies the closed set, and DASS.pv which calculates the statistical significance of the derived closed sets. 2) Analytical mode: Pattern</td>
<td>No</td>
<td>No</td>
<td>NA</td>
<td>No</td>
<td>Yes. It uses biclustering and other data mining method.</td>
<td>Yes</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Analysis Tool</td>
<td>Language(s), Platforms</td>
<td>Specific Features</td>
<td>Data Types</td>
<td>Source Availability</td>
<td>Relevance</td>
<td>Analysis Scope</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------</td>
<td>------------------------</td>
<td>-------------------</td>
<td>------------</td>
<td>----------------------</td>
<td>-----------</td>
<td>----------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GeneTrail2</td>
<td>Optimized C++ library based on Boost, Eigen 3 and GMP</td>
<td>It provides differential expression tests at the identifier level and set level; It also provides multiple tests corrections. Its gene set and phenotype strategies use an optimal permutation method to reduce computing time.</td>
<td>No, T, M, P, GE, miRNA</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OmixAnalyzer</td>
<td>Java, R, Perl</td>
<td>It includes differential analysis (t-test and ANOVA), cluster and functional enrichment, provides figures and reports. It targets mid-size systems biology project.</td>
<td>Yes, No</td>
<td>GE, EX, P (on its way)</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Specmine</td>
<td>R</td>
<td>Identification of metabolites, univariate (corr, regression, ANOVA), multi-variate (robust PCA, cluster), machine learning and feature selection (classification and regression, validation)</td>
<td>Yes, No</td>
<td>M, S</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ImDEV</td>
<td>R and Visual Basic</td>
<td>It provides functions to execute multivariate R functions from Excel. It includes MDS methods (Cluster, PCA, PLS) and 2/3 Dimensional visualizations.</td>
<td>Yes, No</td>
<td>M, C</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XMRF</td>
<td>R</td>
<td>Fitting Markov Networks to a wide range of high-throughput genomics data</td>
<td>Yes, No</td>
<td>GE</td>
<td>Yes</td>
<td>No</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PathVisioRPC</td>
<td>Allowed access from R, Perl, Python, Java, C, C++, PHP</td>
<td>A Remote Procedure Call for PathVisio, provides a link/Communicating between the interface (PathViso) and the statistical analytical tools (scripts). PathVisioRPC wraps pathVisio functionality into XMLRPC functions which can be implemented in many languages for execution. The R package of PathVisio is RPathVisio</td>
<td>Yes, it is provided from PathVisio</td>
<td>GE, M, T</td>
<td>No</td>
<td>No</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COBRApy</td>
<td>Python,</td>
<td>It uses constrained modeling to</td>
<td>Yes, No</td>
<td>GE, M</td>
<td>No</td>
<td>No</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Notes:**
- **No** indicates the feature is not available.
- **Yes** indicates the feature is available.
- **Not specified** indicates the feature is not specified in the provided information.
- **Not applicable** indicates the feature is not applicable in the context provided.
<table>
<thead>
<tr>
<th>Tool</th>
<th>Language</th>
<th>Description</th>
<th>Gene Ontology</th>
<th>Database</th>
<th>Use Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matlab (Ebrahim, et al. 2013)</td>
<td>Matlab</td>
<td>MatLab represent the complex biological process of metabolism and gene expression in a pathway. Constrained based modeling includes a biological system constraint which is defined by the objective function, and usually, linear programming is used as the analytical method.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3Omics (Kuo, et al. 2013)</td>
<td>Perl, PHP</td>
<td>Correlation networking, Co-expression, phenotyping, pathway enrichment, and GO (Gene Ontology) enrichment.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paintomics (García-Alcalde, et al. 2011)</td>
<td>Perl, Python</td>
<td>A joint visualization tool for Transcriptomic and Metabolomics Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COEUS (Lopes and Luís Oliveira 2012)</td>
<td>Jena, Java</td>
<td>It is a data integration software, a new semantic web framework No</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cytoscape (Smoot, et al. 2011)</td>
<td>JAVA</td>
<td>A popular tool for biological network visualization and data integration Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Software Package</td>
<td>Language</td>
<td>Description</td>
<td>Mainly for visualization</td>
<td>Supported Datasets</td>
<td>Additional Information</td>
</tr>
<tr>
<td>-----------------------------------</td>
<td>----------</td>
<td>------------------------------------------------------------------------------</td>
<td>---------------------------</td>
<td>--------------------</td>
<td>----------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>MGV (May day Graph Viewer)</td>
<td>JAVA</td>
<td>JAVA is an extension of the platform mayday</td>
<td>No</td>
<td>T,M,P,GE</td>
<td>No</td>
</tr>
<tr>
<td>Symons and Nieselt 2011</td>
<td></td>
<td>It provides visualizations for cluster comparison between studies, cross datasets biological pathway, gene models, and probe centric view.</td>
<td>Yes, mainly for visualization</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Omix (Droste, et al. 2011)</td>
<td>OVL script</td>
<td>A customized visualization tool for metabolic network</td>
<td>Yes</td>
<td>KEGG</td>
<td>No</td>
</tr>
<tr>
<td>MVBioDataSim (Fratello, et al. 2015)</td>
<td>R</td>
<td>It is a multi-view genomic data simulator</td>
<td>No</td>
<td>No</td>
<td>GE</td>
</tr>
<tr>
<td>ATHENA** (Kim, et al. 2013)</td>
<td>implemented in C++ and uses the libGE(version 0.206) and GAlib (version 2.4.7) genetic algorithm library</td>
<td>Grammatical evaluation neural network is used to analyze associations between single, multiple level genetic interactions and clinical outcomes. It includes: a) variable/feature selection; 2) model main and interactions effects predicting clinical outcomes; 3) interpretation prepared for further bioinformatics</td>
<td>No</td>
<td>(TCGA) data portal ovarian cancer</td>
<td>CNA,GM, miRNA ,GE, C Machine learning method: Extension of Artificial Neural Network</td>
</tr>
</tbody>
</table>


** Software package that has functions to integrate clinical data and omics data, and provides advanced statistical techniques for integrated data analysis.

†† Integration occurs at the molecular level: the input data are IDs of gene, protein, and metabolite and merged by these ID; results are derived by using public databases (i.e. pathway enrichment analysis via information of KEGG). Integration occurs at the subject level: the input data are an original expression or sequence variables from the same subject, data are merged by subject ID.
References:


