Can Big Data Analytics Recapitulate Biology? A Survey of Multi-omics Data Integration Approaches.

Amrit Singh[1,2]*, Casey P. Shannon*[2]*, Kim-Anh Lê Cao*[3]*, and Scott J. Tebbutt*[2,4]

ORCID iDs: 0000-0002-7475-1646, 0000-0002-5687-3156, 0000-0003-3923-1116, 0000-0002-7908-1581

[1]Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, British Columbia, Canada.
[2]Prevention of Organ Failure (PROOF) Centre of Excellence, Vancouver, BC, Canada.
[3]Melbourne Integrative Genomics, School of Mathematics and Statistics, The University of Melbourne, Melbourne, Australia.
[4]Department of Medicine (Respiratory Division), University of British Columbia, Vancouver, BC, Canada.

Abstract

Advances in high throughput technologies, and associated reduction in costs, have enabled simultaneous profiling of many biological compartments, and the collection of many data types from biological specimens. These high dimensional datasets, in turn, have necessitated the development of novel integrative methods to consider the data in a holistic manner, a paradigm sometimes termed “systems biology”. In this book chapter, we survey current approaches for the integration of multiple omics high-dimensional datasets obtained on the same set of individuals. These approaches employ a variety of methods such as factorization, message passing, multi-block methods, generalized canonical correlation analysis, classification and regression algorithms, network-constrained and Bayesian methods. These approaches can be categorized into data-driven approaches which are based only on empirical data, or knowledge-based approaches which also incorporate known biological knowledge in order to improve biological interpretability. A distinction is made between unsupervised approaches that seek common relationships between datasets, and supervised approaches that use labeled data in order to identify discriminatory patterns (*e.g.* between phenotypic groups). We describe the objectives of each approach and give examples of their application to multiple omics data, along with important limitations.

1 Introduction

High-throughput molecular and cellular analytical platforms are inundating researchers with high-dimensional multi-omics data. A major challenge of modern biological research is to integrate within and across these very complex datasets to derive biological insight. Integrating across multiple data sources can improve our understanding of a complex system[1] and may lead to more reliable hypothesis-generation. For example, in biology, a given molecular pathway may entail the structural and/or functional interplay between various molecules from messenger RNA transcripts to proteins and metabolites. Biological functions may be modified by genetic (*e.g.* mutations and polymorphisms), epigenetic (*e.g.* methylation and microRNA), and environmental regulatory factors. Simultaneous molecular profiling of tens to millions of molecules (later referred to as 'variables' or 'features') from different biological compartments is necessary to capture this complexity.

Data integration of high-dimensional omics (*e.g.* transcriptomics, proteomics, metabolomics) datasets may be carried out using *unsupervised analyses* which disregard sample label information (such as disease status). On the other hand, *supervised analyses* can be used to identify patterns, trends, and associations that discriminate between qualitatively distinct groups of specimens. Integrative analyses may also incorporate additional data from curated databases such as protein-protein interactions, canonical pathways, transcription factor binding data, *etc*. We refer to approaches that leverage such data as *knowledge-based*, whereas approaches that solely rely on empirical data are referred to as *data-driven*. The biological question, sample size, and types of data available, should inform the type of integrative analysis applied, though many exploratory avenues may be open to the investigators.

We define two broad categories of integrative studies: 1) P-integration and 2) N-integration[2,3]. P-integration integrates multiple datasets that were generated for the same set of P variables. For example, multiple laboratories interested in a specific research area may perform whole genome profiling of different cohorts of individuals. The integrative analysis would constitute a combined analysis of the different datasets for that given omic source. While this is an important application, we will not discuss P-integration in this chapter. Rather, we will focus specifically on N-integration approaches that are being used to integrate different datasets from multiple omics sources, obtained from the same set of individuals. The need for such methods is timely as multiple omics datasets are becoming increasingly common (*e.g.* The Cancer Genome Atlas (TCGA); 7 data-types including SNP (Single Nucleotide Polymorphisms), CNV (Copy Number Variants), DNA methylation, gene expression microRNA expression for over 30 cancer types[4]).

We present several approaches for analyzing multi-omics data, including factorization methods, message passing algorithms, methods for multi-block data analysis and generalized canonical correlation analysis (CCA), network-based methods, Bayesian methods and classification and regression algorithms (Figure 1). Many of the techniques covered in this chapter can be considered as belonging to the field of machine learning[5,6]. These integrative techniques are used to "learn" holistic patterns from multi-omics data. We categorize the methods for multi-omics data integration into *unsupervised* and *supervised* analyses, as well as *data-driven* and *knowledge-based* (Figure 1). Our survey is not meant to provide a comprehensive review of the literature on multi-omics data integration, but rather to introduce the reader to the different analytical approaches. Table 1 provides a brief summary including the data-types, number of samples and variables for each method discussed in this chapter.

The chapter is organized as follows: Section 2 covers methods for unsupervised analyses, and is further divided into data-driven methods (2.1) and knowledge-based methods (2.2). Section 3 discusses methods for supervised analyses, including data-driven (3.1) and knowledge-based (3.2) methods. We conclude in Section 4 and address remaining knowledge gaps, limitations and areas for further development.

[INSERT FIGURE 1]

2 Methods for unsupervised multi-omics data integration

In this section, we describe methods that integrate multiple high-dimensional omics datasets without using additional phenotypic information on the biological samples.

2.1 Data-driven methods for unsupervised multi-omics data integration

Joint Non-negative Matrix Factorization: Joint Non-negative Matrix Factorization (joint NMF) was proposed to identify subsets of correlated variables across different biological layers within all or a subset of samples[7]. Joint NMF extends NMF by projecting multiple datasets onto the same coordinate axes. Zhang *et al.*[7] applied Joint NMF to mRNA (GE), miRNA (ME) and DNA methylation data (DM) from 385 ovarian cancer samples, in order to identify correlated multi-dimensional (md-) modules. On average, each md-module consisted of 239.6 mRNAs, 13.8 miRNAs and 162.3 methylation markers. Any given md-module was defined as functionally homogeneous if it was enriched in at least one GO biological process category, with a q-value < 0.05. Among the 200 md-modules, 80 GE, 12.5 ME and 62.7% DM were identified as functionally homogeneous. Combining the GE, ME and DM dimensions resulted in 93% of the md-modules being functionally homogenous. Further, a significant proportion of genes were adjacent to DNA methylation markers in the same md-modules as well as genes targeted by miRNAs. Stratifying patients with/without strong association with particular md-modules indicated significant differences with respect to median survival times.

Joint NMF is a useful method to uncover coherent features across multi-omics data while reducing the complexity of the data, as it models the global structure of the different data-types. Md-modules consist of regulatory relationships that span different biological layers but an arbitrary threshold is required to determine/define md-module membership. To avoid such *ad-hoc* thresholds, would require the incorporation of sparsity constraints that result in variable selection.

Multiple Co-Inertia Analysis: Similar to joint NMF, Multiple Co-Inertia Analysis (MCIA) is a dimension reduction technique that also projects several datasets (also referred to as blocks) onto the same dimensional space in order to improve biological insight. Meng *et al.*[8] applied MCIA to transcriptomics (from multiple platforms) and proteomics data from 59 cell lines from nine difference tissues as part of the NCI-60 cancer cell line project. Cancer lines from the same tissue of origin clustered closely across all datasets with a stronger similarity between the transcriptomic datasets than between the transcriptomic and proteomic datasets. The first principal component (dimension) separated cell lines based on epithelial and mesenchymal characteristics and the associated loadings (variable importance scores) consisted of larger weights (in absolute value) for epithelial and mesenchymal genes. Integrative analysis of the transcriptomics and proteomics data indicated greater statistical significance for the top identified pathways (e.g. *leukemia extravasation signaling pathway*) compared to the analysis of the transcriptomics data alone.

Since each dimension is associated with a loading vector where the weights for all variables are non-zero, variable selection using penalization constraints can improve biological interpretability. Overall MCIA is a useful method to unravel common trends across multi-omics datasets and may improve biological insights and help generate novel hypotheses.

Spare Multi-Block Partial Squares (sMB-PLS): Transcriptional output (gene expression) is under strict regulatory control through many factors such as the genome (*e.g.* copy number variants), and epigenome (miRNA and DNA methylation). MB-PLS maximises the covariance between multiple datasets and a response matrix of interest. Sparse MB-PLS (sMB-PLS) extends MB-PLS[9] by allowing for variable selection such that a smaller set of variables from each omic dataset can be identified. sMB-PLS was used to identify sets of regulatory factors that affect the expression of a specific collection of mRNA transcripts[10]. Li *et al.*[10]*,* used sMB-PLS to identify gene expression modules using copy number variants (CNV), DNA methylation (DM), microRNA expression (ME) from 230 ovarian TCGA samples. Md-modules on average consisted of 30 samples, 45 CNVs, 42 DM marks, 5 miRNAs, and 44 genes. Modules obtained using conventional methods such as biclustering algorithms missed variables from one (59% of modules) or two (22% of modules) datasets. Concatenation of the datasets and then applying sPLS resulted in modules that missed at least one dataset (47% of modules), where 17% of modules only consisted of variables from one dataset. The md-modules consisted of variables with significantly greater functional homogeneity compared to md-modules identified using randomized data. sMB-PLS is a flexible method for the integration of multi-omics datasets as well as performing variable selection from each dataset.

Similarity Network Fusion: Patient phenotyping is useful in identifying subgroups of individuals with similar characteristics in heterogeneous diseases such as cancer and other complex diseases. Similarity Network Fusion (SNF) was developed to integrate similarity matrices (also called networks) into a “fused” similarity matrix (based on message passing theory) which can then be used to perform sample clustering[11]. Wang *et al.*[11], applied SNF to mRNA, miRNA and DNA methylation data from five cancer-types from TCGA. Spectral clustering was applied to the patient similarity matrix of each data-type independently and to the patient-similarity network produced by SNF. Patient subgroups identified using SNF were more tightly correlated compared to clusters identified on each data-type separately. The use of local affinity may have led to a reduction in noise (correlations between patient clusters) in SNF compared to spectral clustering applied to each data-type separately. Half of the patient similarities were supported by two data-types, 17.2% by three data-types and one-third by only one data-type. The subtypes identified from the SNF analysis resulted in significantly different survival profiles as compared to the single dataset cluster analyses. SNF also outperformed iCluster (a joint latent variable model for integrative clustering) on survival analysis and risk of death prediction using the METABRIC[12] dataset.

SNF requires several hyperparameters (similarity index, scaling factors) that must be specified by the user, or tuned (which can be burdensome) and have an impact on the resulting clusters. A limitation of SNF is that the method uses all variables in each data-type, therefore, coherent patterns that exist between subsets of variables from each data-type are missed. SNF would benefit from variable selection and supervised analysis extensions.

Bayesian Consensus Clustering (BCC): Bayesian Consensus Clustering[13] (BCC) provides a statistical framework for integrative clustering of samples simultaneously using all omics datasets. BCC uses a mixture model to identify a data-specific clustering that is conditional on an overall clustering. Simulation studies were used to demonstrate the reduced error of BCC in identifying correct cluster membership, compared to separate and joint (concatenation of datasets) cluster analysis using a Dirichlet mixture model. The joint method performed well when there was a perfect agreement between datasets, whereas the separate method performed well when there was no agreement between datasets. BCC was also applied to breast cancer data and the identified clusters were different but not independent from those identified using the PAM50 classification (Her2, Basal, LumA and LumB). Most subjects in BCC cluster 1 corresponded to the Basal subtype, whereas BCC cluster 2 consisted of a unique set of LumA samples which had lower copy number variants (as measured by the fraction of genome altered), and better survival times. This type of LumA cluster has been previously identified by other studies, on independent datasets[12,14].

BCC is a useful method in identifying a consensus clustering of samples that is represented across the different omic datasets. BCC is able to handle model uncertainty and borrows information across multiple omic datasets. However, since BCC makes prior distributional assumptions, strong deviations from these distributions may result in a consensus clustering with little agreement between omic datasets. Nevertheless, to infer whether or not BCC could be applicable to a given set of multi-omics data, the adherence parameter can be used to determine the level of coherency between each dataset-specific clustering and the overall consensus clustering.

2.2 Knowledge-based methods for unsupervised multi-omics data integration

The previous section focused primarily on data-driven methods that look for coherent patterns between multi-omics data. This section introduces unsupervised knowledge-based methods.

Sparse Network-regularized Multiple Non-negative Matrix Factorization (SNMNMF):

MicroRNAs play an important regulatory role in mRNA translation by binding to the 3' untranslated regions of their targets. The functional roles of miRNAs may be better understood by incorporating curated interactions with empirical data from expression studies. SNMNMF is a computational framework for identifying co-modules (sets of miRNAs and mRNAs) by incorporating miRNA and mRNA expression data, as well as PPI and DNA-protein interaction data[15]. Zhang *et al.*[15], applied SNMNMF to mRNA and miRNA expression form 385 ovarian samples from TCGA. 49 miRNA-gene co-modules were identified, each consisting on average of 3.8 miRNAs and 78 mRNAs per module. The anti-correlation between miRNAs and mRNAs were statistically significant in 69.4% of the modules as compared to randomized miRNA-gene co-modules. 11 of the modules were significantly enriched with at least one miRNA cluster (miRNAs located within 50kb in the genome). Significantly higher numbers of enriched GO biological processes were identified using the 49 co-modules compared to random modules. Stratifying patients based on co-module activity resulted in three groups that often differed with respect to their survival characteristics.

SNMNMF can functionally annotate groups of miRNA:mRNA pairs, and is a useful hypothesis-generating tool for further mechanistic work using *in vitro* or *in vivo* studies. The method may be extended to other types of data such as drug interaction data, copy number variants *etc*. However, the combination of additional sources of data will require the use of additional penalties that may increase the computational cost of tuning additional hyperparameters.

__Passing Attributes between Networks for Data Assimilation (PANDA):__ PANDA is a message-passing algorithm that incorporates a cooperativity network (e.g. protein-protein interaction data), coregulatory network (e.g. co-expression network using gene expression) and a regulatory network (e.g. motif data comprising of transcription factor (TF) and gene interactions) in order to determine a consensus regulatory network. Glass *et al.*[16], applied PANDA to gene expression data in yeast, TF motif (sequence) data and protein-protein interaction data from BIOGRID. The jackknife procedure was used to remove 10% of edges in the cooperativity, regulatory and coregulatory networks in the initial and final networks identified using PANDA, and compared to gold-standard networks in order to compute the performance based on the area under the receiver operating characteristic curve (AUROC). The gold standard for these networks was based on interaction experiments to generate cooperativity (*e.g*. co-fractionation, co-localization) networks, ChIP-chip data for regulatory networks, and co-targeted in Chip-chip (if both gene pairs had a binding site associated with a particular transcription factor) for coregulatory networks. The resulting PANDA networks were significantly more predictive of all network types as determined using experimental data as compared to the initial networks. Furthermore, PANDA was competitive with other network reconstruction approaches in predicting regulatory networks using regulator knock-out, cell-cycle and stress-response datasets.

Significant improvements have recently been made to PANDA for computational efficiency[17]. However, further exploration is needed in order to determine the stability of these networks. Although PANDA was more predictive in estimating networks compared to motif data alone, increases in the AUROCs were modest (between 3-8%). Further extensions of PANDA for the identification of differential regulatory networks may provide additional utility in studies with multiple phenotypic groups.

Reconstructing Integrative Molecular Bayesian Networks (RIMBANET): Bayesian networks are directed graphs, where the nodes represent random variables (*e.g.* omic features) and edges represent conditional probabilities between nodes and their parents (also called probabilistic causal networks). Thousands of networks are generated using Monte Carlo simulations (optimized using the Bayesian Information Criterion) and combined into a consensus network (RIMBANET). Zhu, Sova and Xu *et al.*[18]*,* used six data-types to construct probabilistic causal networks. 16 of 59 quantified metabolites had significant log odds (FDR <0.05) and 12 of 16 metQTLs overlapped with four previously identified eQTL hot spot regions[19]. Gene transcripts and metabolites linked to the first two eQTL hot spots recapitulated known biological processes (leucine and pyrimidine biosynthesis pathways). This Bayesian network approach was useful in attributing functional roles to an uncharacterised eQTL hot spot 3, by linking the metabolites isoleucine, threonine and valine as well as the gene *CHA1* which encodes a serine/threonine deaminase. However, no causal regulator was identified for this hot spot which may suggest that the genetic variation affected protein levels directly instead of changes in the mRNA levels. Protein-coding variants in genes in this hot spot were identified and tested experimentally by knocking out each candidate gene and comparing its metabolite levels with those in wild-type yeast strains. Knockout of the vacuolar transport regulatory gene *VPS9* resulted in changes to threonine, isoleucine, valine and serine concentrations, suggesting its role as a potential causal regulator of the eQTL hot spot 3.

Bayesian networks have proven useful in accurately predicting complex cell regulatory processes which have been validated experimentally. The use of both high throughput data as well as known interactions from biological databases has proven useful in recapitulating complex cell behaviour. However, as these processes are most likely disease-specific, such methods should be applied independently to data from each disease condition.

3 Methods for supervised multi-omics data integration

In the previous section, the focus was primarily on maximizing the coherence across multi-omics data, regardless of sample groupings. The methods described in this section incorporate phenotypic information in order to identify discriminatory patterns between sample groupings. For example, biological networks across multiple biological compartments that exist in healthy subjects may become dysregulated in cancer subjects[20]. These patterns may result in the identification of biomarker signatures that are predictive of phenotypic traits, either in continuous (*e.g.* age) or categorical forms (*e.g.* cancer vs. controls).

3.1 Data-driven methods for supervised multi-omics data integration

Integrative Bayesian Analysis of Genomics (iBAG): This method uses a two-component hierarchical model to first determine the effects of methylation and other components (*e.g*. copy number variants) on gene expression and then jointly uses both components to predict a continuous outcome such as survival time. Wang *et al.*[21], used iBAG to integrate gene and methylation data of glioblastoma multiforme (GBM) cancer samples. iBAG was compared to non-integrative methods (non-INT) where a model was constructed using only the gene expression data, and an additive model (ADD), where both gene and methylation data were combined into one joint explanatory matrix in order to predict patient survival times. The C-index (generalization of the AUROC) of the iBAG model was higher than both the non-INT and ADD models, for both training and test datasets, albeit with overlapping 95% confidence intervals. iBAG identified 22 novel genes associated with survival, not previously associated with GBM. The iBAG framework offers the flexibility of incorporating additional sources of information such as pathway information, and additional data-types.

Classification and regression algorithms

Concatenation-based classifiers: A simple approach for multi-omics classification panel is to combine data matrices of different omics into one joint matrix. Classification algorithms such as penalized regression, support vector machines (SVM), and random forest can then be applied to identify a set of molecular features that best predict a given outcome of interest. Although useful, combining datasets of different data-types poses a challenge. For example, different scales of various data-types lead to differences in the relative effect sizes, such that some data-types may over power the signal from other data-types[22]. A recent study using multi-omics datasets (copy number variation, gene expression, and proteomics) from breast cancer tumours demonstrated that a classifier developed using only the proteomics dataset outperformed classifiers developed using other omics datasets as well as multi-omics (data-fusion) methods such as concatenation, multiple kernel learning methods, and random forest[23].

Ensemble classifiers: A predictive model (classifier) is constructed independently for each omic dataset and is combined (ensembled) using various classification rules, such as average or majority vote schemes[6]. Ensemble classifiers often have lower error rates compared to single base classifiers. This is possible if the error rate of the single classifiers is less than 50% and if the single classifiers are independent with uncorrelated errors[24]. Gunther *et al.*[25], developed proteogenomic classifiers that could discriminate acute rejection from non-rejection samples in a kidney transplant study. The ensemble classifiers out-performed the individual classifiers based on the area under the receiver operating characteristic curve using both the average and majority vote aggregation methods. Model stacking is another method to combine model predictions of multiple classifiers using additional classifiers resulting in meta-classifiers[26].

Multi-block methods and extensions to generalized canonical correlation analysis: The presence of multiple datasets observed on the same set of individuals leads to the natural use of methods for the analysis of multi-block data[27–31]. Sparse generalized canonical correlation analysis (sGCCA) combines the power of multi-block data analysis with well-defined criteria to optimize and the flexibility of Projection to Latent Structures, which incorporates *a priori* relationships between datasets[30]. sGCCA is a dimension reduction approach which maximizes the sum of the pairwise covariances of connected (specified by the user) datasets. Tenenhaus *et al.*[32], applied sGCCA to genomic data (gene expression, copy number variants and a qualitative variable encoding the location of the tumours; central nuclei or brain stem) from 53 tumour samples of children with pediatric high-grade gliomas. The components generated by sGCCA were used to build a predictive model using Bayesian discriminant analysis. The performance of this approach was comparable to existing approaches such as penalized LDA (linear discriminant analysis)[33], supervised CCA[34] and regularized GCCA[30].

Supervised Multi-block Sparse Matrix Analysis (SMSMA) was recently proposed[35], and can be viewed as a supervised extension of sMB-PLS in section 2.1. SMSMA was applied to SNP data, imaging data (magnetic resonance imaging, and positron emission tomography) and a continuous outcome variable measuring dementia in patients with Alzheimer's Disease (AD). The components of the SNP and imaging data were used in a multiple logistic regression model to predict AD and healthy controls. Using a 10-fold cross-validation, the AUROC was 0.762 and 0.952 for the SNP and imaging data respectively. Many SNPs with the largest contributions to the principal components were found near previously associated AD risk genes such as *BIN1*, *APOE*, *CCR2*, *LOC651924*, and *IL13*. Therefore, SMSMA is a useful method to obtain a subset of multi-omic features that can discriminate between phenotypic groups. Methods for multi-block data analysis and extensions to generalized CCA offer an extremely flexible methodology for different study designs and are becoming commonly used for the analysis of multi-omics data.

3.2 Knowledge-based methods for supervised multi-omics data integration

We describe methods that incorporate experimental data (*e.g.* gene expression) and data from curated databases such as gene-gene, or miRNA-gene interactions. They may also be augmented using concatenation or ensemble based approaches for the integration of multiple omic datasets.

Network Smoothed T-Statistic Support Vector Machines (stSVM): Biomarker panels (combination of single omic features) can be developed using the top-ranked features based on some univariate ranking procedure such as a *t*-test, or ANOVA. Adjusting this ranking based on biological connectivity between variables may capture the complexity of biological networks that exist in heterogeneous systems. The stSVM method combines univariate ranking based test statistics with a network graph and selects the top variables from the resulting modified ranking to develop a SVM classifier. stSVM[36] was applied to gene expression data for several cancer types (breast, ovarian, prostate and prostate) in order to predict survival times (dichotomized into 2 classes) and on average outperformed sgSVM (SVM trained using significant genes, false discovery rate (FDR) <5%), aepSVM (average gene expression of KEGG pathways), PAC (pathway activity classification), RRFE (reweighted recursive feature elimination) and netRank (modification of Google's PageRank to rank genes based on differential expression and network connectivity). The features selected using stSVM were consistently selected in the cross-validation folds due to their connectivity in the protein-protein network. Enrichment analysis of the selected features corroborated existing cancer related pathways *(e.g. Pathways in cancer*, *Prostate Cancer*, *ERBB signaling*).

Although stSVM had the top consensus ranking across the four analyzed datasets, its cross-validated performance was well in the range of the other methods. Further, feature stability could be due to the strong influence of the networks used which may overpower the univariate ranking such that the same features were repeatedly selected. However, stSVM provides a biomarker panel which improved interpretability in terms of biological functionality as well as strong classification performance.

Generalized Elastic Net: Instead of univariate ranking of features prior to classifier development, variable selection can be induced through regularization of the loss function in linear models. Generalized Elastic Net (GELnet) extends the elastic net penalty[37] to regularize generalized linear models, through the use of penalties for individual features and pairs of features based on domain knowledge. Sokolov *et al.*[38], demonstrated that GELnet outperformed Elastic Net[37] with respect to the reconstruction error rate, when the same graphical Gaussian model was used to simulate both the gene expression and signaling network. GELnet was applied to mRNA expression data from 54 cancer cell lines as well as their sensitivity profiles to 74 compounds. GELnet achieved a lower root mean square error for 22 of the 74 drugs as compared to Elastic Net which may suggest the incomplete knowledge of the underlying biological mechanisms that is captured in curated genetic pathway databases. Therefore, GELnet is limited by incomplete pathway information and may not always out-perform Elastic Net if the network information is not associated with the expression data.

Adaptive group-regularized ridge regression (GRridge): Variable selection can significantly improve the predictive performance and interpretability of classification algorithms. This can be further improved by incorporating additional information across groups of omic variables which together have zero or non-zero coefficients. GRridge uses an empirical Bayes method to estimate group-specific penalties based on grouping data (*e.g.* annotations). This procedure leads to a larger difference between small and large regression coefficients, which may be used to perform variable selection. Furthermore, only one global penalty is needed to be tuned using cross-validation. Van de Wiel *et al.*[39], applied GRridge to methylation data from 20 normal cervical tissue and 17 CIN3 (cervical intraepithelial neoplasia high-grade precursor lesions) tissue biopsies. Methylation probes were grouped based on their distance from CpG islands: CpG island (CpG), North Shore (NSe), South Shore (SSe), North Shelf (NSf) and South Shelf (SSf) and Distant (D). The estimates of group-specific penalties were large for all groups except for the CpG and SSe group, resulting in a GRridge model containing only CpG and SSe probes. GRridge (AUROC=0.92) outperformed ridge (AUROC=0.86), adaptive ridge (AUROC=0.84) and the group-lasso (AUROC=0.79) in discriminating CIN3 from normal samples. The advantage of tuning a lower number of hyperparameters offers GRridge computational advantages over the group lasso, as well as reduced bias since all variables are kept non-zero.

5 Summary and concluding remarks

In this book chapter, we described current approaches for the integration of multi-omics data, focusing on N-integration. We categorized these methods into *unsupervised* and *supervised* analyses, as well as methods that use experimental data only, and those that also incorporate curated data of known molecular relationships. Each of these methods are unique, answering different questions from the data and should be used when appropriate, and in conjunction with one another. Therefore, application of each method will often result in different results and conclusions.

While the methods surveyed in this chapter are impressive in their ability to model the highly complex regulatory interplay of biological systems, some aspects of this complexity are still underserved. Integration of microbiome[40], metagenome[41], environmental exposures[42] *etc.*, are of particular interest. In addition, methods that can accommodate more complex study designs, such as repeated measures designs for longitudinal studies[43], and cross-over repeated measure studies[44] would be highly desirable. Finally, since the end user of these algorithms are most likely biomedical researchers, visual aids for the complex outputs of these algorithms would be useful to enable model interpretation and assessment[45]. Lastly, these algorithms must be user-friendly and be made freely available through open source software.

Each method and general approach has limitations, both at the mathematical and computational level, and all rely on underlying assumptions as to the quality, accuracy and precision of the original datasets, as well as additional issues that plague the phenotyping of individuals in natural populations. Even so, in combination with thoughtful study design, well controlled standard operating protocols for both sample acquisition and processing, and an ability to work in the exploratory space offered by numerous and complementary computational approaches, the era of multi-omics data integration is becoming more established and has tremendous opportunity for increasing biological insights, hypothesis-generation, and knowledge. The hypotheses generated through multi-omics data integration must be followed up with experimental studies in order to isolate true biological relationships from purely spurious results.

[INSERT TABLE 1]

Acknowledgements

We thank the biosignatures development team at the PROOF Centre of Excellence for continuous discussions and feedback on the contents of this book chapter.

Figure captions

Figure 1. Methods for multi-omics data integration. The integrative methods discussed in this book chapter are based on four major categories, supervised *versus* unsupervised analyses, and data-driven *versus* knowledge-based methods. All methods are further grouped into different method-types. For each method-type, specific examples are listed along with the package in the R statistical computing language in the form; Name of method: `R-package` Methods with a superscript "a" have been implemented in Matlab and their source code is linked in their respective publications. iBAG[b] is implemented in R and its associated R-package and R-based Shiny web application can be found on the authors webpage[46]. The software used to construct Bayesian networks (superscript "c") is called Reconstructing Integrative Molecular Bayesian Networks (RIMBANET) implemented in Perl[47]. All other methods have been implemented in R and can be obtained *via* the comprehensive archive network (CRAN, https://cran.r-project.org/), Bioconductor (https://www.bioconductor.org/), or github (https://github.com/).

Table 1. Summary of studies describing methods for multi-omics data integration covered in this chapter

Method	Disease	Datasets	# of variables	# of samples	Reference
Data-driven unsupervised analyses					
Joint Non-negative Matrix Factorization (NMF)	Ovarian cancer	DNA methylation	2,008 probes	385	7
		gene expression	2,985 genes		
		miRNA expression	270 miRNAs		
Multiple Co-Inertia Analysis (MCIA)	NCI-60: 59 cancer cell lines	gene expression:			8
		Agilent	11,051 genes	-	
		HGU95	8,803 genes		
		Hgu133	9,044 genes		
		Hgu133 plus 2.0	10,382 genes		
		protein expression	7,150 proteins	-	
Sparse Multi-Block Partial Least Squares (sMB-PLS)	Ovarian cancer	copy number variants	31,324 variants	230	10
		DNA methylation	14,735 probes		
		gene expression	15,846 genes		
		miRNA expression	799 miRNAs		

Similarity Network Fusion (SNF)	Glioblastoma multiforme (GBM)		DNA methylation	1,491 probes	215	11
			gene expression	12,042 genes		
			miRNA expression	534 miRNAs		
	Other cancers: breast, kidney, lung, colon		DNA methylation, gene and miRNA expression	534 miRNAs in GBM to 21,578 methylated genes in lung and colon cancer	92-215	
Bayesian Consensus Clustering (BCC)	Breast cancer		DNA methylation	574 probes	348	13
			gene expression	645 genes		
			miRNA expression	423 miRNAs		
			protein expression	171 proteins		
Knowledge-based unsupervised analyses						
Sparse Network-Regularized joint Non-negative Factorization (SNMNMF) method	Ovarian cancer	Empirical data	gene expression	12,456 genes	385	15
			miRNA expression	559 miRNAs		
		Curated data	protein-protein and DNA-protein interaction data	31,949 gene-gene interactions	-	
			predicted miRNA-gene interactions	243,331 miRNA-gene interactions	-	

Passing Attributes between Networks for Data Assimilation (PANDA)	*Saccharomyces cerevisiae* (yeast)	Empirical data	gene expression	2555 genes	-	16
		Curated data	transcription factor (TF) motifs in sequence data	53 TFs	-	
			protein-protein interactions	135,415 TF-gene interactions	-	
Reconstructing Integrative Molecular Bayesian Networks (RIMBANET)	*Saccharomyces cerevisiae* (yeast)	Empirical data	DNA variation	-	-	18
			gene expression	3,662 genes	-	
			metabolite expression	56 metabolites	120 yeast segregants	
		Curated data	protein-DNA binding	119 TFs	-	
			protein-protein interaction	-	-	
			metabolite-protein interaction	2,252 metabolite-protein interactions	-	
Data-driven supervised analyses						
	Glioblastoma multiforme	DNA methylation		6,890 probes		

<table>
<tr><td><b>Integrative Bayesian Analysis of Genomics (iBAG)</b></td><td></td><td colspan="2">gene expression</td><td>7,785 genes</td><td>201</td><td>21</td></tr>
<tr><td rowspan="4"><b>Classification and regression algorithms: Concatenation-based classifiers</b></td><td rowspan="4">Breast cancer</td><td colspan="2">copy number variation</td><td>24,174 variants</td><td rowspan="4">77</td><td rowspan="4">23</td></tr>
<tr><td colspan="2">gene expression</td><td>16,525 genes</td></tr>
<tr><td colspan="2">protein expression</td><td>12,553 proteins</td></tr>
<tr><td colspan="2">phosphoprotein expression</td><td>32,939 phosphoproteins</td></tr>
<tr><td rowspan="2"><b>Classification and regression algorithms: Ensemble-based classifiers</b></td><td rowspan="2">Acute kidney rejection</td><td colspan="2">gene expression</td><td>27,306 genes</td><td rowspan="2">32</td><td rowspan="2">25</td></tr>
<tr><td colspan="2">proteomics</td><td>147 protein groups</td></tr>
<tr><td rowspan="2"><b>Sparse Generalized Canonical Correlation Analysis (SGCCA)</b></td><td rowspan="2">Pediatric high-grade gliomas</td><td colspan="2">copy number variants</td><td>1,229 variants</td><td rowspan="2">53</td><td rowspan="2">32</td></tr>
<tr><td colspan="2">gene expression</td><td>15,702 genes</td></tr>
<tr><td rowspan="2"><b>Supervised Multi-block Sparse Matrix Analysis (SMSMA)</b></td><td rowspan="2">Alzheimer’s disease</td><td colspan="2">genotypes</td><td>549,709 SNPs</td><td rowspan="2">100</td><td rowspan="2">35</td></tr>
<tr><td colspan="2">imaging data</td><td>2,122,945 features</td></tr>
<tr><td colspan="7"><b>Knowledge-based supervised analyses</b></td></tr>
<tr><td rowspan="2"><b>Network Smoothed T-Statistic Support</b></td><td rowspan="2">Breast, ovarian prostate</td><td rowspan="2">Empirical data</td><td>gene expression</td><td rowspan="2">-</td><td rowspan="2">79-228</td><td rowspan="2">36</td></tr>
<tr><td>miRNA expression</td></tr>
</table>

Vector Machine (stSVM)		Curated data	protein-protein interaction	610,185 gene-gene interactions		
			pathway data	17,518 gene-gene interactions		
			miRNA-gene interactions	-		
Generalized Elastic Net (GELnet)	Breast cancer	Empirical data	gene expression	9,984 genes	54	[38]
		Curated data	pathway data	-		
Adaptive group-regularized ridge regression (GRridge)	Cervical cancer	Empirical data	DNA methylation data	40,000 probes	37	[39]
		Curated data	CpG genomic location	6 probe locations		

References

1. Gomez-Cabrero, D. *et al.* Data integration in the era of omics: current and future challenges. *BMC Syst. Biol.* **8,** I1 (2014).

2. Tseng, G., Ghosh, D. & Zhou, X. J. *Integrating Omics Data*. (Cambridge University Press, 2015).

3. Rohart, F., Gautier, B., Singh, A. & Le Cao, K.-A. mixOmics: an R package for omics feature selection and multiple data integration. *bioRxiv* 108597 (2017).

4. The TCGA Research Network. The Cancer Genome Atlas.

5. Bishop, C. M. *Pattern recognition and machine learning*. (Springer, 2006).

6. Lin, E. & Lane, H.-Y. Machine learning and systems genomics approaches for multi-omics data. *Biomark. Res.* **5,** (2017).

7. Zhang, S. *et al.* Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Res.* **40,** 9379–9391 (2012).

8. Meng, C., Kuster, B., Culhane, A. C. & Gholami, A. M. A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics* **15,** 162 (2014).

9. Kowalski, B. R. & Wangen, L. E. A multiblock partial least squares algorithm for investigating complex chemical systems. *J. Chemom.* **3,** 3–20 (1989).

10. Li, W., Zhang, S., Liu, C.-C. & Zhou, X. J. Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. *Bioinformatics* **28,** 2458–2466 (2012).

11. Wang, B. *et al.* Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* **11,** 333–337 (2014).

12. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* (2012). doi:10.1038/nature10983

13. Lock, E. F. & Dunson, D. B. Bayesian consensus clustering. *Bioinformatics* **29,** 2610–2616 (2013).

14. Jonsson, G. *et al.* Genomic subtypes of breast cancer identified by array-comparative genomic hybridization display distinct molecular and clinical characteristics. *Breast Cancer Res.* **12,** R42 (2010).

15. Zhang, S., Li, Q., Liu, J. & Zhou, X. J. A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules. *Bioinformatics* **27,** i401–i409 (2011).

16. Glass, K., Huttenhower, C., Quackenbush, J. & Yuan, G.-C. Passing messages between biological networks to refine predicted interactions. *PLoS ONE* **8,** e64832 (2013).

17. van IJzendoorn, D. G. P., Glass, K., Quackenbush, J. & Kuijjer, M. L. PyPanda: a Python package for gene regulatory network reconstruction. *Bioinformatics* **32,** 3363–3365 (2016).

18. Zhu, J. *et al.* Stitching together multiple data dimensions reveals interacting metabolomic and transcriptomic networks that modulate cell regulation. *PLoS Biol.* **10,** e1001301 (2012).

19. Zhu, J. *et al.* Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nat. Genet.* **40,** 854–861 (2008).

20. Ha, M. J., Baladandayuthapani, V. & Do, K.-A. DINGO: differential network analysis in genomics. *Bioinformatics* **31,** 3413–3420 (2015).

21. Wang, W. *et al.* iBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics* **29,** 149–159 (2013).

22. Aben, N., Vis, D. J., Michaut, M. & Wessels, L. F. A. TANDEM: a two-stage approach to maximize interpretability of drug response models based on multiple molecular data types. *Bioinformatics* **32,** i413–i420 (2016).

23. Ma, S., Ren, J. & Fenyö, D. Breast cancer prognostics using multi-omics data. *AMIA Summits Transl. Sci. Proc.* **2016,** 52 (2016).

24. Dietterich, T. G. Ensemble methods in machine learning. in *Multiple Classifier Systems* 1–15 (Springer Verlag, 2000).

25. Günther, O. *et al.* A computational pipeline for the development of multi-marker bio-signature panels and ensemble classifiers. **13,** 326 (2012).

26. Whalen, S. & Pandey, G. A comparative analysis of ensemble classifiers: case studies in genomics. in 807–816 (IEEE, 2013). doi:10.1109/ICDM.2013.21

27. Lohmöller, J.-B. *Latent variables path modeling with partial least squares*. (Heidelberg: Physica-Verlag, 1989).

28. Wold, H. Partial least squares. in *Encyclopedia of statistical sciences* **6,** 581–591 (New York: Wiley, 1985).

29. Tenenhaus, M., Vinzi, V. E., Chatelin, Y.-M. & Lauro, C. PLS path modeling. *Comput. Stat. Data Anal.* **48,** 159–205 (2005).

30. Tenenhaus, A. & Tenenhaus, M. Regularized generalized canonical correlation analysis. *Psychometrika* **76,** 257–284 (2011).

31. Rohart, F., Gautier, B., Singh, A. & Cao, K.-A. L. mixOmics: An R package for ‘omics feature selection and multiple data integration. *PLOS Comput. Biol.* **13,** e1005752 (2017).

32. Tenenhaus, A. *et al.* Variable selection for generalized canonical correlation analysis. *Biostatistics* **15,** 569–583 (2014).

33. Witten, D. M. & Tibshirani, R. Penalized classification using Fisher’s linear discriminant. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **73,** 753–772 (2011).

34. Witten, D. M., Tibshirani, R. & Hastie, T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10,** 515–534 (2009).

35. Kawaguchi, A. & Yamashita, F. Supervised multiblock sparse multivariable analysis with application to multimodal brain imaging genetics. *Biostatistics* **00,** 1–15 (2017).

36. Cun, Y. & Fröhlich, H. Network and data integration for biomarker signature discovery via network smoothed t-statistics. *PLoS ONE* **8,** e73074 (2013).

37. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67,** 301–320 (2005).

38. Sokolov, A., Carlin, D. E., Paull, E. O., Baertsch, R. & Stuart, J. M. Pathway-based genomics prediction using generalized elastic net. *PLoS Comput Biol* **12,** e1004790 (2016).

39. van de Wiel, M. A., Lien, T. G., Verlaat, W., van Wieringen, W. N. & Wilting, S. M. Better prediction by use of co-data: adaptive group-regularized ridge regression. *Stat. Med.* **35,** 368–381 (2016).

40. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102,** 15545–15550 (2005).

41. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44,** W90–W97 (2016).

42. Avey, S. *et al.* Multiple network-constrained regressions expand insights into influenza vaccination responses. *Bioinformatics* **33,** i208–i216 (2017).

43. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning*. (Springer, 2001).

44. Ambroise, C. & McLachlan, G. J. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl. Acad. Sci.* **99,** 6562–6566 (2002).

45. IOM, (Institute of Medicine). *Evolution of translational omics: lessons learned and the path forward*. (National Academy Press, 2012).

46. Bourgon, R., Gentleman, R. & Huber, W. Independent filtering increases detection power for high-throughput experiments. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 9546–9551 (2010).

47. Cawley, G. C. & Talbot, N. L. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.* **11,** 2079–2107 (2010).

48. Lê Cao, K.-A. *et al.* MixMC: A multivariate statistical framework to gain insight into microbial communities. *PLOS ONE* **11,** e0160169 (2016).

49. Heintz-Buschart, A. *et al.* Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nat. Microbiol.* **2,** 16180 (2016).

50. Robinson, O. *et al.* The pregnancy exposome: multiple environmental exposures in the INMA-Sabadell birth cohort. *Environ. Sci. Technol.* **49,** 10632–10641 (2015).

51. Straube, J., Gorse, A.-D., PROOF Centre of Excellence Team, Huang, B. E. & Lê Cao, K.-A. A linear mixed model spline framework for analysing time course 'omics' data. *PLOS ONE* **10,** e0134540 (2015).

52. Liquet, B., Lê Cao, K.-A., Hocini, H. & Thiébaut, R. A novel approach for biomarker selection and the integration of repeated measures experiments from two assays. *BMC Bioinformatics* **13,** 325 (2012).

53. González, I., Lê Cao, K.-A., Davis, M. J. & Déjean, S. Visualising associations between paired 'omics' data sets. *BioData Min.* **5,** 1–23 (2012).

54. Baladandayuthapani, V. iBAG page. Available at: http://odin.mdacc.tmc.edu/~vbaladan/Veera_Home_Page/iBAG_page.html. (Accessed: 14th August 2017)

55. Zhu, J. RIMBANET for Bayesian network reconstruction. Available at: http://research.mssm.edu/integrative-network-biology/RIMBANET/RIMBANET_overview.html. (Accessed: 14th August 2017)