Multivariate Statistical Methods for High-Dimensional Multiset Omics Data Analysis

Attila Csala • Aeilko H. Zwinderman

Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, Amsterdam, The Netherlands

Author for correspondence: Attila Csala, Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, Amsterdam 1105 AZ, The Netherlands. Email: a@csala.me

Doi: http://dx.doi.org/10.15586/computationalbiology.2019.ch5

Abstract: This chapter covers the state-of-the-art multivariate statistical methods designed for high-dimensional multiset omics data analysis. Recent biotechnological developments have enabled large-scale measurement of various biomolecular data, such as genotypic and phenotypic data, dispersed over various omics domains. An emergent research direction is to analyze these data sources using an integrated approach to better model and understand the underlying biology of complex disease conditions. However, comprehensive analysis techniques that can handle both the size and complexity, and at the same time can account for the hierarchical structure of such data, are lacking. An overview of some of the developments in multivariate techniques for high-dimensional omics data analysis, highlighting two well-known multivariate methods, canonical correlation analysis (CCA) and redundancy analysis (RDA), is provided in this chapter. Penalized versions of CCA are widespread in the omics data analysis field, and there is recent work on multiset penalized RDA that is applicable to multiset omics data. How these methods meet the statistical challenges that come with high-dimensional multiset omics data analysis and help to further our understanding of the human condition in terms of health and disease are presented. Additionally, the current challenges to be resolved in the field of omics data analysis are discussed.

In: *Computational Biology.* Holger Husi (Editor), Codon Publications, Brisbane, Australia. ISBN: 978-0-9944381-9-5; Doi: http://dx.doi.org/10.15586/computationalbiology.2019

Keywords: canonical correlation analysis; high-dimensional data analysis; integrative omics data; multivariate statistics; redundancy analysis

INTRODUCTION

High-throughput sequencing methods such as the Affymetrix GeneChip 1994, Illumina SNP genotyping 2001 and Illumina BeadChip 2005 have provided the possibility of collecting millions of molecular variables (i.e., biomolecular data) from biological samples (1). Simultaneously, developments in knowledge databases including the Kyoto Encyclopedia of Genes and Genomes 1995, Human Genome Project 2003 and 1000 Genomes Project 2015, along with the formation of large biobanks such as the Estonian Genome Project 2000 and the UK Biobank 2006, have provided new means to store and manage biomolecular data. National computing services and leading data science companies have established large-scale computer facilities (e.g., Globus Genomics 2013, Helix Nebula 2013 and European Open Science Could 2019) to enable routine access and analysis of extremely large databases (2, 3). Many biomedical research institutions have established biobanks to store and manage both organic tissue and in silico data of patients on genetic and genomic variations, epigenetic measurements, and gene- and protein-expressions in various tissues, along with disease phenotypes and treatment response (1, 3).

These technological developments in the biomedical field, sometimes collectively referred to as the biotechnological revolution, have created new opportunities to better understand the human condition in terms of health and disease. The development and application of statistical methods that aim to analyze and understand large-scale biomolecular data is referred to as the field of biomolecular big data analysis. The topic of this chapter is omics data analysis, which is a subfield of biomolecular big data analysis. Omics data analysis aims to analyze and understand large-scale biomolecular data from more than one omics data source, where omics is shorthand for a range of -omics domains such as genomics, epigenomics, transcriptomics, proteomics, lipidomics, metabolomics and microbiomics. The field of omics data analysis has two main objectives (4–6):

(i) To understand the underlying biology of disease conditions with emphasis on mechanisms and etiology
(ii) To improve our ability to predict, prevent and treat disease conditions (i.e., translational medicine).

While there has been considerable progress on these objectives for simple monogenic disease conditions (7), such progress has been slow for complex poly- and omnigenic disease conditions (5, 8, 9). The main reason for the relatively low progress in complex conditions is often attributed to the lag between the technologies to collect such vast amounts of biomolecular data and the techniques to analyze and understand such data (10). Current technologies can measure vast amounts of data on simple as well as on complex disease conditions. However, complex conditions presumably have multifaceted underlying biological pathways that the current techniques are unable to model from the available large-scale data sources (9, 11, 12).

Advancements in biotechnology offer the possibility to routinely collect, store and analyze high-dimensional omics data. The high-dimensionality of such biomolecular data refers to the routine practice of collecting biomarkers and disease phenotypes (i.e., biomolecular variables) on a large-scale, often measured in the thousands to millions, while the number of available samples (i.e., patients) is usually measured mostly in the hundreds (i.e., variables >> samples). The collection and analysis of vast numbers of biomolecular variables is hoped to help biomedical scientists to better understand the human condition in terms of health and disease. The main goal of omics data analysis is to model biological pathways in biomolecular data sources in such a way that the biological pathways best model the genetic architecture and the overall underlying biology of disease conditions (8). The resulting biological pathway models then can be used to understand the mechanisms and etiology of disease conditions and ultimately be used to improve our ability to treat such conditions. In light of these possibilities, many scientists believe that personalized medicine at an extremely detailed molecular scale will be possible in the near future (13, 14).

This chapter provides an overview of the development of techniques that are aimed at analyzing and understanding large-scale biomolecular data, with emphasis on multivariate techniques for omics data analysis. Multivariate techniques can: (i) handle the simultaneous analysis of multiple high-dimensional omics data sources, (ii) provide biologically interpretable results, (iii) have well-defined objective functions (no-black box methods) and (iv) preferably have open source software implementations. A perspective on the gap between the technologies that collect, store and manage large-scale biomolecular data and the techniques that analyze and understand such data (i.e., the technology-technique gap) is provided. The four periods in the history of omics data analysis (Table 1) that are well distinguishable in terms of paradigm shifts and the way the biomedical scientific community approaches large-scale biomolecular data are described. Although there are various statistical methods available to analyze omics data, many of them do not meet certain requirements. Thus, the so-called supervised machine learning techniques, which require labeled data for classification (15, 16), are excluded. An excellent review that describes supervised and unsupervised techniques can

TABLE 1 The four periods of development of multivariate techniques and the associated paradigm shifts

Period	Time	Technique	Paradigm Shift
1	Early 2000s	Univariate approach	Associating one or a subset of biomarkers with a single-disease phenotype
2	Late 2000s	Multivariate approach	Associating subset of biomarkers and disease-phenotypes with each other
3	Early 2010s	Multiset multivariate approach	Associating subsets of biomarkers and disease-phenotypes with each other from various data sources
4	Late 2010s	Hierarchical multiset multivariate approach	Associating one or a subset of dependent disease-phenotypes with subsets of independent biomarkers from various data sources

be found in Ref. (17). Also, methods that can be considered multivariate techniques but do not have well-defined objective functions are excluded (12, 18–20). Overall reviews on multivariate techniques for omics data analysis can be found in Refs. (14, 21–25).

EARLY 2000s: THE UNIVARIATE APPROACH

Historically, most techniques focus on analyzing the association between a single disease phenotype and one, or a subset of, biomarker(s) from a particular omics data source. This approach has been widespread since the early 2000s in genome-wide association studies (GWASs) (7). The study published in 2002 by Ozaki et al. on myocardial infarction is widely regarded as the first successful GWAS study (26). Generally, a GWAS aims to analyze the association between a single disease phenotype and one or a subset of biomarkers, which translates to a monothematic model (1). This is often referred to as the univariate approach, since there is only a single dependent variable, namely a disease phenotype, that is associated with one or a subset of independent variables, namely the biomarker(s). Biological pathways modeled by the univariate model are then composed by a single disease phenotype and one or a subset of biomarker(s). This univariate approach, especially in the GWAS framework, has made considerable contributions to biomarker discovery for monogenic and genetically complex conditions (8, 27). However, many biomedical scientists argue that univariate approaches are suboptimal for the pursuit of objectives (i) and (ii) mentioned above, especially when applied to data collected on patients with complex poly- or omnigenic conditions (1, 8, 9, 11).

LATE 2000s: MULTIVARIATE APPROACHES

Complex poly- or omnigenic conditions have complex biological pathways, composed of multiple biomarkers that can be associated with more than one disease phenotype. That is, biological pathways of complex conditions can be best modeled in omics data by associating multiple biomarkers with multiple disease phenotypes. The emergence of this hypothesis resulted in the development of multivariate techniques for omics data analysis, since some multivariate techniques are able to associate multiple disease phenotypes with multiple biological markers.

Penalized canonical correlation analysis

Among the first multivariate statistical methods that were developed for omics data analysis are the modified versions of canonical correlation analysis (CCA). CCA is a well-known multivariate technique that aims to subtract linear combinations of variables (i.e., canonical variates) from two data sources, in a way that the canonical variates maximally correlate with each other (28). The objective function of CCA is:

$$arg \max_{\boldsymbol{a},\boldsymbol{b}} cor(\boldsymbol{Xa},\boldsymbol{Yb}), \tag{1}$$

where ***X*** denotes the first data source and ***Xa*** denotes a linear combination of the variables from ***X***, and ***Y*** denotes the second data source and ***Yb*** denotes a linear combination of the variables from ***Y***. ***Xa*** and ***Yb*** are the canonical variates, and the correlation between the canonical variates is called the canonical correlation. Thus, the objective function of CCA is to maximize the canonical correlation.

CCA applied to omics data results in a set of biomarkers from one omics data source that maximally correlates with a set of biomarkers or disease phenotypes from a second data source. Note that CCA does not distinguish between dependent and independent variables. Also, CCA, in its organic form, is not applicable to omics data, since the high-dimensional nature of omics data (i.e., variables >> samples) causes CCA to fail to subtract canonical variates from the data sources. Modified versions of CCA that solve this issue have started to appear from the late 2000s, among them are penalized canonical correlation analysis (penalizedCCA) (29), regularized canonical correlation analysis (rCCA) (30), sparse canonical correlation analysis (sCCA) (31) and penalized canonical correlation analysis (pCCA) (32). These studies applied a form of penalization to the organic CCA framework, which makes penalized forms of CCA applicable to high-dimensional data and, in most cases, results in a model that includes only a subset of the original variables from the data sources (i.e., variable selection) (33). Variable selection is a desirable property when the original variables are too numerous to be interpretable in the results of the analysis, which is exactly the case with omics data. The exact properties of variable selection depend on the type of penalization applied to CCA, and an overview on penalization methods can be found in Ref. (34). In general, penalized forms of CCA have the same objective function as the generic CCA, that is, it aims to maximize the correlation between linear combinations of two (sub)sets of variables. Applying penalized forms of CCA to omics data results in a model with a (sub)set of biomarkers that maximally correlate with a (sub)set of disease phenotypes or biomarkers penalizedCCA, sCCA and pCCA facilitate variable selection, while sCCA uses a penalization form that makes it applicable to high-dimensional data but does not facilitate variable selection.

Penalized partial least squares regression

Other multivariate statistical methods that were developed in the late 2000s for omics data analysis are modified versions of partial least squares regression (PLS). PLS is a set of general least squares regression techniques applied in an iterative algorithmic framework, and, in fact, CCA is a special case of PLS (35). In general, PLS techniques aim to subtract two sets of linear combinations of variables (i.e., latent variables) from two data sources in a way that the covariance between the latent variables is maximized (36). The objective function of PLS is:

$$arg \max_{\boldsymbol{a},\boldsymbol{b}} cov(\boldsymbol{Xa},\boldsymbol{Yb}), \qquad (2)$$

where ***X*** denotes the first data source and ***Xa*** denotes a linear combination of the variables from ***X***, and ***Y*** denotes the second data source and ***Yb*** denotes a linear combination of the variables from ***Y***. ***Xa*** and ***Yb*** are the latent variables. The objective function of the generic PLS is to maximize the covariance between the latent variables. While this objective function can be modified based on the regression techniques used in the iterative framework (35), the early applications of PLS to omics data aimed to maximize the covariance between the latent variables.

PLS applied to omics data results in a linear combination of biomarkers between two data sources that have maximum covariance with each other. Similar to CCA, PLS in its organic form is not applicable to omics data, since high-dimensional data (i.e., variables >> samples) cause the general least squares regression techniques in PLS to fail to subtract linear combinations from the data sources. Lê Cao et al. introduced a penalized version of PLS, called the sparse PLS (sPLS), to solve this issue (37). Other PLS-based methods are sparse partial least squares regression (sPLSR) (38), sparse PLS-discriminant analysis (sPLS-DA) (39) and two-way orthogonal PLS (O2PLS) (40). sPLS, sPLSR, sPLS-DA and O2PLS facilitate variable selection, which is a desirable property, as discussed above in the case of penalized CCA.

EARLY 2010S: MULTISET MULTIVARIATE APPROACHES

From the mid-2010s, the need has become apparent for multiset techniques that are able to analyze multiple sets of omics data sources simultaneously (i.e., integrated or multiset techniques). The developments of such methods were motivated by the hypothesis that biological pathways are composed of a collection of biomarkers and disease phenotypes that are not constrained to one or two biological domains. This hypothesis was probably influenced by the relatively new field of systems biology.

Systems biology advocates that properties of biological organisms can be best modeled by assessing its multiple components and the interactions of its various biological domains simultaneously (41). Thus, system biology claims that system properties, such as the function and mechanism of complex conditions, can be better assessed through a system-wide approach (i.e., integrating and analyzing different parts of an organism simultaneously) in contrast to the so-called reductionist approach (i.e., analyzing different parts of an organism separately). Translating this to omics data analysis, one may hypothesize that techniques constrained to one or two omics domains result in a monothematic type of knowledge and possibly miss modeling system-wide properties of complex conditions. In fact, omics domains are not discrete and separable biological entities, as the reductionist approach advocates, but they can rather be better conceptualized as different biomolecular data sources measuring the manifestation of particular biological pathways across different biological sections in the organism. In other words, various omics data sources can be seen as measurements of biomarkers and disease phenotypes of particular conditions present in the patient, dispersed over various biomolecular sections. Therefore, for complex poly- and omnigenic conditions, integrated analysis of multiple omics data sources should be favored (1).

Generalized penalized canonical correlation analysis

The simultaneous analysis of multiple omics domains created the anticipation that multiset techniques will enable better biological pathway models through the discoveries of biomarkers and disease phenotypes that are dispersed over multiple biomolecular domains (42). One group of such multiset techniques is based on generalized penalized CCA (43), which is the generalization of penalized CCA to multiple data sources. The objective function of generalized penalized CCA is similar to that of CCA in Equation 1, but instead of maximizing the canonical correlation of two canonical variates, it maximizes the canonical correlation of multiple canonical variates

$$arg \max_{\boldsymbol{a}_1,\ldots,\boldsymbol{a}_j} \sum\nolimits_{j,k=1,\, j\neq k}^{J} c_{ij}\; cor\left(\boldsymbol{X}_j \boldsymbol{a}_j, \boldsymbol{X}_k \boldsymbol{a}_k\right), \tag{3}$$

where $\boldsymbol{X_j}$ denotes the *j*th data source and $\boldsymbol{X_j a_j}$ denotes a linear combination of the variables from $\boldsymbol{X}_{j.}$ $\boldsymbol{X_j a_j}$ is the *j*th canonical variate and c_{jk} indicates whether two data sources are connected; $c_{jk} = 1$ if $\boldsymbol{X_j}$ and $\boldsymbol{X_k}$ are connected and 0 otherwise (43).

Generalized penalized CCA applied to omics data results in multiple sets of biological variables that maximally correlate with each other, thereby enabling the simultaneous analysis of multiple biomarkers and disease phenotypes that are dispersed over multiple omics domains. Variations of generalized penalized CCA for omics data analysis started to appear in the mid-2010s, among them are generalized CCA (gCCA) (44), sparse generalized canonical correlation analysis (sGCCA) (45) and data integration analysis for biomarker discovery using latent components (DIABLO) (46). sGCCA and DIABLO facilitate variable selection, while gCCA does not.

Penalized multi-block partial least squares regression

Another group of multiset techniques belong to the extended versions of penalized PLS. These techniques, called multi-block penalized PLS, have a similar objective function to that of penalized PLS in Equation 2 (as generalized penalized CCA relates to penalized CCA). We omit the equation, as it is almost identical to Equation 3, but instead of the correlation, the covariances between the multiple latent variables are maximized. Multi-block penalized PLS applied to omics data results in multiple sets of biomarkers or disease phenotypes that have maximum covariance with each other. Some of the early applications of multi-block penalized PLS to omics data analysis are sparse Multi-Block PLS (sMBPLS) regression (47) and Sparse multi-block PLSR (Sparse MBPLSR) (48). Both sMBPLS and Sparse MBPLSR facilitate variable selection.

A summary of multivariate methods for one-, two-, and multiset omics data analysis can be found in (23). These multiset methods, based on CCA and PLS, are able to detect multiple highly associated biomarkers and disease phenotypes dispersed over multiple biological domains. Note that all the multivariate techniques described so far are aiming to maximize either the correlation or covariance between linear combinations of (sub)sets of biomarkers and disease phenotypes. Therefore, they can at best be used to pursue our understanding of the mechanisms of complex disease. However, in order to understand disease etiology, analyzing the correlation and covariance between linear combinations of subsets of variables is not sufficient (4, 5, 11).

LATE 2010s: HIERARCHICAL MULTISET MULTIVARIATE APPROACHES

Since the mid-2010s, the need for techniques that are not only able to help detect correlated biomarkers and biological pathways of disease phenotypes, but also could aid in detection of causal relationships and understanding disease etiology, has become more apparent (4, 5, 11). This need was motivated by the hypothesis that omics domains have an inherent hierarchical relationship in terms of possible interactions. One of the earliest hypotheses for such a hierarchical relationship model for biomolecular domains, called the Central dogma of molecular biology, was published in the 1970s, sketching plausible interactions between what we call today genomics, transcriptomics and proteomics (49). The Central dogma postulates that genetic information is transferred from genomics to proteomics through transcriptomics. As of today, there are multiple hypotheses on the possible hierarchical structure between the various omics domains, with most implying a genetic information flow from the genome to the phenome (11). In other words, there is a hierarchical structure between genome and phenome in terms of the phenome being dependent on the genome. Thus, in order to better understand disease etiology for complex conditions, multiset multivariate techniques that are able to account for a hierarchical structure between omics domains in terms of dependent and independent data sources should be favored. Redundancy analysis (RDA), the multivariate equivalent of regression analysis, accounts for the genetic information flow in omics domains by distinguishing between dependent and independent omics data sources.

Penalized multi-block redundancy analysis

RDA can be seen as the multivariate extension of univariate regression analysis. RDA aims to subtract linear combinations of independent variables (i.e., latent variables) from a data source in a way that the latent variables explain the most variance in a second dependent data source (50). The objective function of RDA is:

$$\mathit{arg}\,\max_{\boldsymbol{a}} \sum\nolimits_{q=1}^{Q} \mathit{cor}\left(\boldsymbol{Xa}, \boldsymbol{y}_{\mathbf{q}}\right)^{2}, \tag{4}$$

where $\boldsymbol{X}$ denotes the independent data source, $\boldsymbol{Xa}$ denotes a linear combination of the variables from $\boldsymbol{X}$ and $\boldsymbol{y}_{\mathbf{q}}$ denotes the qth variable from the dependent data source (with a total of Q variables). $\boldsymbol{Xa}$ is a latent variable, and the sum of the squared correlations between the latent variable and all the variables of $\boldsymbol{Y}$ is called the redundancy index. Thus, the objective function of RDA is to maximize the redundancy index. Note that RDA maximizes the sum of squared pairwise correlations between a linear combination of variables from an independent data source and between variables of a dependent data source. The aim of RDA is then to find a linear combination of the independent variables that explains the most variance in all the dependent variables. Similarly, we could describe the CCA (or PLS) techniques we presented earlier as techniques aiming to explain maximum variance in their canonical variate (or latent variable) pairs. But the CCA and PLS techniques do not distinguish between dependent and independent data sources, since in Equation 1, and in Equation 2, the objective function is maximized with respect to the canonical variates, and latent variables, from both data sources, and thus, the variables in both data sources are regarded as independent variables. In Equation 4, the objective function of RDA is maximized with respect to the latent variable of $\boldsymbol{X}$, and the variables from $\boldsymbol{Y}$ are not transformed and are regarded as the dependent variables.

RDA applied to omics data results in a set of independent biomarkers from one data source that explains the most variance in the dependent disease phenotypes from a second data source. RDA accounts for the hierarchical structure between data sources in terms of dependent and independent variables. RDA, in its organic form, is not applicable to omics data, since high-dimensional data cause RDA to fail to subtract latent variables from the independent data source. Similarly, as with CCA and PLS, this can be solved by introducing penalization to RDA. The first penalized RDA, called regularized linear redundancy analysis (regRDA), appeared in the late 2000s (51), and its first application to omics data analysis, called sparse redundancy analysis (sRDA), was in the late 2010s (52). sRDA facilitates variable selection and regRDA does not.

Penalized RDA is able to account for the hierarchical structure between two data sources, and its multiset extension is able to account for the hierarchical structure between multiple data sources. The objective function of multiset penalized RDA is similar to that of RDA in Equation 4, but instead of maximizing the redundancy index between the independent latent variable and all the dependent variables, it maximizes the sum of redundancy indices of multiple latent variable with all the dependent variables (53):

$$arg \max_{\boldsymbol{a}_1,\dots,\boldsymbol{a}_j} \Sigma_j^J \Sigma_q^Q cor\left(\boldsymbol{X}_j \boldsymbol{a}_j, \boldsymbol{y}_q\right)^2, \tag{5}$$

where $\boldsymbol{X}_j$ denotes the jth independent data source and $\boldsymbol{y}_q$ denotes the qth variable from the dependent data source (with a total of Q variables). $\boldsymbol{X}_j \boldsymbol{a}_j$ denotes the jth linear combination of the variables from $\boldsymbol{X}_j$.

Multiset penalized RDA applied to omics data results in multiple sets linear combinations of independent biomarkers that explain the most variance in the dependent disease phenotypes. Therefore, multiset penalized RDA enables the simultaneous analysis of multiple biomolecular variables that are dispersed over multiple omics domains, while it accounts for the hierarchical structure between the data sources. One application of multiset penalized RDA is multiset sparse redundancy analysis (multi-sRDA) (53), which facilitates variable selection. A summary of the multivariate methods reviewed in this text can be found in Table 2.

CONCLUSION

We examined the state-of-the-art techniques aimed to analyze and understand large-scale biomolecular data. As also reported by others, we likewise identified a technology–technique gap, namely the gap between technologies to collect, store and manage large-scale biomolecular data and the techniques to analyze

TABLE 2 Multivariate statistical methods for high-dimensional omics data analysis, a chronological overview

Name	Multiset	Variable selection	Hierarchical	Year	Reference
Penalized CCA (pCCA)	no	yes	no	2007	(28)
Regularized CCA (rCCA)	no	no	no	2008	(29)
Sparse PLS (sPLS)	no	yes	no	2008	(36)
Sparse CCA (sCCA)	no	yes	no	2009	(30)
Penalized CCA (pCCA)	no	yes	no	2009	(31)
Sparse partial least squares regression (sPLSR)	no	yes	no	2009	(37)
Sparse PLS-discriminant analysis (sPLS-DA)	no	yes	no	2011	(38)
Regularized generalized CCA (rGCCA)	yes	no	no	2011	(42)
sparse Multi-Block PLS (sMBPLS) regression	yes	yes	no	2012	(46)
Generalized CCA (gCCA)	yes	no	no	2014	(43)
Sparse generalized canonical correlation analysis (sGCCA)	yes	yes	no	2014	(44)
Sparse multi-block PLSR (Sparse MBPLSR)	yes	yes	no	2015	(47)
Two-Way Orthogonal PLS (O2PLS)	no	yes	no	2016	(39)
Sparse RDA (sRDA)	no	yes	yes	2017	(51)
Multiset sRDA	yes	yes	yes	2018	(52)
Data Integration Analysis for Biomarker discovery using Latent cOmponents (DIABLO)	yes	yes	no	2019	(45)

The first column contains the names, column *Multiset* indicates whether the method is applicable for multiple omics sets, column *Variable selection* indicates whether the method facilitates variable selection and column *Hierarchical* indicates whether the method is able to account for the hierarchical structure between omics data sources. This table is complementary to and based on the tables that can be found in (23).

and understand such data. We described four periods in the history of omics data analysis that are well distinguishable in terms of paradigm shifts in the way the biomedical scientific community approaches large-scale biomolecular data. We highlighted some of the main effects of these major paradigm shifts on the advancement of the omics data analysis field. The main motivation to switch from univariate to multiset multivariate techniques is that analytical techniques constrained to one or two omics domains result in a monothematic type of knowledge and likely miss modeling system-wide properties of complex conditions. Omics domains are not discrete and separable biological entities as reductionist-type approaches. They should be conceptualized as various biomolecular data sources measuring the manifestations of biological pathways across various biological sections in an organism. Therefore, various omics domains can be seen as sources for biomarkers and disease phenotypes of particular conditions present in patients, dispersed over various biomolecular sections. We described multiset multivariate methods that aim to identify associated biomarkers and disease phenotypes dispersed over various biomolecular sections and therefore provide optimized biological pathway models of complex conditions. Therefore, to pursue objectives (i) and (ii) mentioned in the introduction section for complex poly- and omnigenic conditions, multiset multivariate techniques should be favored over univariate ones. To pursue objective (ii), techniques that aim to identify causal associations should be favored. We describe techniques that aim to identify causal relationships by modeling the hierarchical structure between omics domains in terms of interactions between biomarkers and disease phenotypes from various omics domains. As of today, there are multiple hypotheses on the possible hierarchical structure between the various omics domains, and most of these hierarchical structures aim to model the genetic information flow from the genome to the phenome. We conclude that, in order to pursue objectives (i) and (ii) for complex conditions, a prominent research direction for the omics data analysis field is the development and application of hierarchical multiset multivariate approaches.

Conflict of interest: The authors declare no potential conflict of interest with respect to research, authorship and/or publication of this chapter.

Copyright and permission statement: To the best of our knowledge, the materials included in this chapter do not violate copyright laws. All original sources have been appropriately acknowledged and/or referenced. Where relevant, appropriate permissions have been obtained from the original copyright holder(s).

REFERENCES

1 Manzoni C, Kia DA, Vandrovcova J, Hardy J, Wood NW, Lewis PA, et al. Genome, transcriptome and proteome: The rise of omics data and their integration in biomedical sciences. Brief Bioinform. 2018 Mar 1;19(2):286–302. http://dx.doi.org/10.1093/bib/bbw114

2. Berger B, Peng J, Singh M. Computational solutions for omics data. Nat Rev Genet. 2013 Apr 18;14(5):333–46. http://dx.doi.org/10.1038/nrg3433

3. Langmead B, Nellore A. Cloud computing for genomic data analysis and collaboration. Nat Rev Genet. 2018 Jan 30;19(4):208–19. http://dx.doi.org/10.1038/nrg.2017.113

4. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol. 2017 Dec 5;18(1):83. http://dx.doi.org/10.1186/s13059-017-1215-1

5. Gallagher MD, Chen-Plotkin AS. The post-GWAS era: From association to function. Am J Hum Genet. 2018 May;102(5):717–30. http://dx.doi.org/10.1016/j.ajhg.2018.04.002

6. Pingault JB, O'Reilly PF, Schoeler T, Ploubidis GB, Rijsdijk F, Dudbridge F. Using genetic data to strengthen causal inference in observational research. Nat Rev Genet. 2018;19(9):566–80. http://dx.doi.org/10.1038/s41576-018-0020-3

7. Visscher PM, Goddard ME, Derks EM, Wray NR. Evidence-based psychiatric genetics, AKA the false dichotomy between common and rare variant hypotheses. Mol Psychiatry. 2012;17(5):474–85. http://dx.doi.org/10.1038/mp.2011.65

8. Timpson NJ, Greenwood CMT, Soranzo N, Lawson DJ, Richards JB. Genetic architecture: The shape of the genetic contribution to human traits and disease. Nat Rev Genet. 2017 Dec 11;19(2):110–24. http://dx.doi.org/10.1038/nrg.2017.101
9. Wray NR, Wijmenga C, Sullivan PF, Yang J, Visscher PM. Common disease is more complex than implied by the Core Gene Omnigenic Model. Cell. 2018 Jun;173(7):1573–80. http://dx.doi.org/10.1016/j.cell.2018.05.051
10. Alyass A, Turcotte M, Meyre D. From big data analysis to personalized medicine for all: Challenges and opportunities. BMC Med Genomics. 2015;1–12. http://dx.doi.org/10.1186/s12920-015-0108-y
11. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype–phenotype interactions. Nat Rev Genet. 2015;16(2):85–97. http://dx.doi.org/10.1038/nrg3868
12. Yan J, Risacher SL, Shen L, Saykin AJ. Network approaches to systems biology analysis of complex disease: Integrative methods for multi-omics data. Brief Bioinform. 2017 Jun 30;19(June 2017): 1370–81. http://dx.doi.org/10.1093/bib/bbx066
13. Karczewski KJ, Snyder MP. Integrative omics for health and disease. Nat Rev Genet. 2018 Feb 26;19(5):299–310. http://dx.doi.org/10.1038/nrg.2018.4
14. Kim M, Tagkopoulos I. Data integration and predictive modeling methods for multi-omics datasets. Mol Omics. 2018 Feb 12;14(1):8–25. http://dx.doi.org/10.1039/C7MO00051K
15. Li Y, Wu FX, Ngom A. A review on machine learning principles for multi-view biological data integration. Brief Bioinform. 2018;19(2):325–40.
16. Camacho DM, Collins KM, Powers RK, Costello JC, Collins JJ. Next-generation machine learning for biological networks. Cell. 2018;173(7):1581–92. http://dx.doi.org/10.1016/j.cell.2018.05.015
17. Huang S, Chaudhary K, Garmire LX. More is better: Recent progress in multi-omics data integration methods. Front Genet. 2017 Jun 16;8(JUN):1–12. http://dx.doi.org/10.3389/fgene.2017.00084
18. Min S, Lee B, Yoon S. Deep learning in bioinformatics. Brief Bioinform. 2016 Jul 29;18(5):bbw068. http://dx.doi.org/10.1093/bib/bbw068
19. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, et al. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface. 2018 Apr;15(141):142760. http://dx.doi.org/10.1098/rsif.2017.0387
20. Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. A primer on deep learning in genomics. Nat Genet. 2019 Jan 26;51(1):12–8. http://dx.doi.org/10.1038/s41588-018-0295-5
21. Zierer J, Menni C, Kastenmüller G, Spector TD. Integration of "omics" data in aging research: From biomarkers to systems biology. Aging Cell. 2015 Dec;14(6):933–44. http://dx.doi.org/10.1111/acel.12386
22. Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, Castellani G, et al. Methods for the integration of multi-omics data: Mathematical aspects. BMC Bioinformatics. 2016 Dec 20;17(S2):S15. http://dx.doi.org/10.1186/s12859-015-0857-9
23. Meng C, Zeleznik OA, Thallinger GG, Kuster B, Gholami AM, Culhane AC. Dimension reduction techniques for the integrative analysis of multi-omics data. Brief Bioinform. 2016;17(October 2015): 628–641, bbv108. http://dx.doi.org/10.1093/bib/bbv108
24. Dihazi H, Asif AR, Beißbarth T, Bohrer R, Feussner K, Feussner I, et al. Integrative omics—From data to biology. Expert Rev Proteomics. 2018 Jun 3;15(6):463–6. http://dx.doi.org/10.1080/14789450.2018.1476143
25. Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: Review and cancer benchmark. Nucleic Acids Res. 2019 Jan 25;47(2):1044. http://dx.doi.org/10.1093/nar/gky1226
26. Ozaki K, Yozo O, Aritoshi I, Akihiko S, Ryo Y, Tatsuhiko T, et al. Functional SNPs in the Lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. Nat Genet. 2002;32(4):650–4. http://dx.doi.org/10.1038/ng1047
27. Mills MC, Rahal C. A scientometric review of genome-wide association studies. Commun Biol. 2019 Dec 7;2(1):9. http://dx.doi.org/10.1038/s42003-018-0261-x
28. Hotelling H. Relations between two sets of variates. Biometrika. 1936 Dec 1;28(3/4):321. http://dx.doi.org/10.2307/2333955
29. Waaijenborg S, Zwinderman AH. Penalized canonical correlation analysis to quantify the association between gene expression and DNA markers. BMC Proc. 2007;1 Suppl 1:S122. http://dx.doi.org/10.1186/1753-6561-1-S1-S122
30. Gonzalez I, Déjean S, Martin P, Baccini A. CCA : An R Package to extend canonical correlation analysis. J Stat Softw. 2008;23(12):1–14. http://dx.doi.org/10.18637/jss.v023.i12

31. Parkhomenko E, Tritchler D, Beyene J. Sparse canonical correlation analysis with application to genomic data integration. Stat Appl Genet Mol Biol. 2009 Jan 6;8(1):1–34. http://dx.doi.org/10.2202/1544-6115.1406
32. Witten DM, Tibshirani RJ. Extensions of sparse canonical correlation analysis with applications to genomic data. Stat Appl Genet Mol Biol. 2009 Jan 9;8(1):1–27. http://dx.doi.org/10.2202/1544-6115.1470
33. Tibshirani R. Regression selection and shrinkage via the Lasso. J Roy Stat Soc Ser B. 1996;58:267–88. http://dx.doi.org/10.1111/j.2517-6161.1996.tb02080.x
34. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B (Statistical Methodol.). 2005 Apr;67(2):301–20. http://dx.doi.org/10.1111/j.1467-9868.2005.00503.x
35. Esposito Vinzi V, Russolillo G. Partial least squares algorithms and methods. Wiley Interdiscip Rev Comput Stat. 2013 Jan;5(1):1–19. http://dx.doi.org/10.1002/wics.1239
36. Esposito Vinzi V, Chin WW, Henseler J, Wang H, editors. Handbook of partial least squares. Berlin, Heidelberg: Springer Berlin Heidelberg; 2010.
37. Lê Cao K-A, Rossouw D, Robert-Granié C, Besse P. A sparse PLS for variable selection when integrating omics data. Stat Appl Genet Mol Biol. 2008 Jan 18;7(1):35. http://dx.doi.org/10.2202/1544-6115.1390
38. Chun H, Keleş S. Expression quantitative trait loci mapping with multivariate sparse partial least squares regression. Genetics. 2009 May;182(1):79–90. http://dx.doi.org/10.1534/genetics.109.100362
39. Lê Cao K-A, Boitard S, Besse P. Sparse PLS discriminant analysis: Biologically relevant feature selection and graphical displays for multiclass problems. BMC Bioinformatics. 2011 Dec 22;12(1):253. http://dx.doi.org/10.1186/1471-2105-12-253
40. Bouhaddani S El, Houwing-Duistermaat J, Salo P, Perola M, Jongbloed G, Uh H-W. Evaluation of O2PLS in Omics data integration. BMC Bioinformatics. 2016;17 Suppl 2(2):11. http://dx.doi.org/10.1186/s12859-015-0854-z
41. Tavassoly I, Goldfarb J, Iyengar R. Systems biology primer: The basic methods and approaches. Essays Biochem. 2018 Oct 26;62(4):487–500. http://dx.doi.org/10.1042/EBC20180003
42. Haas R, Zelezniak A, Iacovacci J, Kamrad S, Townsend S, Ralser M. Designing and interpreting "multi-omic" experiments that may change our understanding of biology. Curr Opin Syst Biol. 2017 Dec;6(September):37–45. http://dx.doi.org/10.1016/j.coisb.2017.08.009
43. Tenenhaus A, Tenenhaus M. Regularized generalized canonical correlation analysis. Psychometrika. 2011 Apr 17;76(2):257–84. http://dx.doi.org/10.1007/s11336-011-9206-8
44. Shen C, Sun M, Tang M, Priebe CE. Generalized canonical correlation analysis for classification. J Multivar Anal. 2014 Sep;130:310–22. http://dx.doi.org/10.1016/j.jmva.2014.05.011
45. Tenenhaus A, Philippe C, Guillemot V, Le Cao K-A, Grill J, Frouin V. Variable selection for generalized canonical correlation analysis. Biostatistics. 2014 Jul 1;15(3):569–83. http://dx.doi.org/10.1093/biostatistics/kxu001
46. Singh A, Shannon CP, Gautier B, Rohart F, Vacher M, Tebbutt SJ, et al. DIABLO: An integrative approach for identifying key molecular drivers from multi-omics assays. Birol I, editor. Bioinformatics. 2019 Jan 18;35(January):1–8, 3055–3062.
47. Li W, Zhang S, Liu C-C, Zhou XJ. Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. Bioinformatics. 2012 Oct 1;28(19):2458–66. http://dx.doi.org/10.1093/bioinformatics/bts476
48. Karaman İ, Nørskov NP, Yde CC, Hedemann MS, Bach Knudsen KE, Kohler A. Sparse multi-block PLSR for biomarker discovery when integrating data from LC–MS and NMR metabolomics. Metabolomics. 2015 Apr 14;11(2):367–79. http://dx.doi.org/10.1007/s11306-014-0698-y
49. Crick F. Central dogma of molecular biology. Nature. 1970 Aug 8;227(5258):561–3. http://dx.doi.org/10.1038/227561a0
50. van den Wollenberg AL. Redundancy analysis an alternative for canonical correlation analysis. Psychometrika. 1977 Jun;42(2):207–19. http://dx.doi.org/10.1007/BF02294050
51. Takane Y, Hwang H. Regularized linear and kernel redundancy analysis. Comput Stat Data Anal. 2007 Sep;52(1):394–405. http://dx.doi.org/10.1016/j.csda.2007.02.014
52. Csala A, Voorbraak FPJM, Zwinderman AH, Hof MH. Sparse redundancy analysis of high-dimensional genetic and genomic data. Bioinformatics. 2017 Oct 15;33(20):3228–34. http://dx.doi.org/10.1093/bioinformatics/btx374
53. Csala A, Hof MH, Zwinderman AH. Multiset sparse redundancy analysis for high-dimensional omics data. Biom J. 2018 Nov;61:1–18, 406–423. http://dx.doi.org/10.1002/bimj.201700248