COMPARATIVE ANALYSIS OF SUPERVISED INTEGRATIVE METHODS FOR MULTI-OMICS DATA

Alexei Novoloaca,[1,*] Camilo Broc,[1] Laurent Beloeil,[1] Wen-Han Yu[2] and Jérémie Becker[1]

[1], BIOASTER Research Institute, 40 avenue Tony Garnier, 69007, Lyon, France

[2], Bill & Melinda Gates Medical Research Institute, Cambridge, Massachusetts, USA

***Corresponding author. alexei.novoloaca@bioaster.org**

Camilo Broc: Camilo.Broc@bioaster.org

Laurent Beloeil: Laurent.Beloeil@bioaster.org

Wen-Han Yu: wenhan.yu.dr@gmail.com

Jérémie Becker: Jeremie.Becker@bioaster.org

Résumé

Les récents progrès dans les technologies de séquençage, de spectrométrie de masse et de cytométrie ont permis de collecter plusieurs types de données omiques à partir d'un seul échantillon. Ces vastes ensembles de données ont conduit à un consensus croissant selon lequel une approche holistique est nécessaire pour identifier de nouveaux biomarqueurs candidats et dévoiler les mécanismes sous-jacents à l'étiologie des maladies, clé de la médecine de précision. Bien que de nombreuses revues et évaluations aient été menées sur les approches non supervisées (Bersanelli et al., 2016), leurs homologues supervisés ont reçu moins d'attention dans la littérature et aucun standard de référence n'a encore émergé (Krassowski et al., 2020).

Dans ce travail, nous présentons une comparaison approfondie d'une sélection de cinq méthodes, représentatives des principales familles d'approches intégratives (factorisation de matrice, méthodes à noyaux multiples, apprentissage ensembliste et méthodes basées sur les graphes). Comme contrôle non-intégratif, une forêt d'arbre de décision a été exécutée sur des ensembles de données concaténés et séparés. Les méthodes ont été évaluées à la fois sur des données simulées et réelles, ces dernieres étant soigneusement sélectionnées pour couvrir différentes applications médicales (maladies infectieuses, oncologie et vaccins) et différentes types d'omiques. Un ensemble de quinze scénarios de simulation a été conçu à partir des données réelles pour explorer un espace de paramétres vaste et réaliste (par exemple, taille de l'échantillon, dimensionnalité, déséquilibre de classes, taille de l'effet).

La comparaison entre les approches intégratives et non intégratives a montré que bien que des performances comparables aient été trouvées sur les simulations, les méthodes basées sur des variables latentes surpassaient généralement les méthodes non-intégratives sur les données réelles. Les forces et les limites de ces méthodes seront discutées en détail ainsi que des lignes directrices pour les futures applications.

Mots-clés : analyse comparative ; intégration de données ; données multi-omiques ; modèles de prédiction ; analyse supervisée

Abstract

Recent advances in sequencing, mass spectrometry and cytometry technologies have enabled researchers to collect multiple 'omics data types from a single sample. These large datasets have led to a growing consensus that a holistic approach is needed to identify new candidate biomarkers and unveil mechanisms underlying disease aetiology, a key to precision medicine. While many reviews and benchmarks have been conducted on unsupervised approaches (Bersanelli et al. 2016 [1]), their supervised counterparts have received less attention in the literature and no gold standard has emerged yet (Krassowski et al. 2020 [2]).

In this work, we present a thorough comparison of a selection of five methods, representative of the main families of integrative approaches (matrix factorization, multiple kernel methods, ensemble learning and graph-based methods). As non-integrative control, random forest was performed on concatenated and separated data types. Methods were evaluated both on simulated and real-world datasets, the latter being carefully selected to cover different medical applications (infectious diseases, oncology and vaccine) and data modalities. A set of fifteen simulation scenarios were designed from the real-world datasets to explore a large and realistic parameter space (*e.g.* sample size, dimensionality, class imbalance, effect size).

The comparison between integrative and non-integrative approaches showed that although comparable performances were found on simulations, latent variable models generally outperformed non-integrative methods on experimental data. The strengths and limitations of these methods will be discussed in detail as well as guidelines for future applications.

Key words: benchmark; data integration; multi-omics data; prediction models; supervised analysis

Benchmark workflow

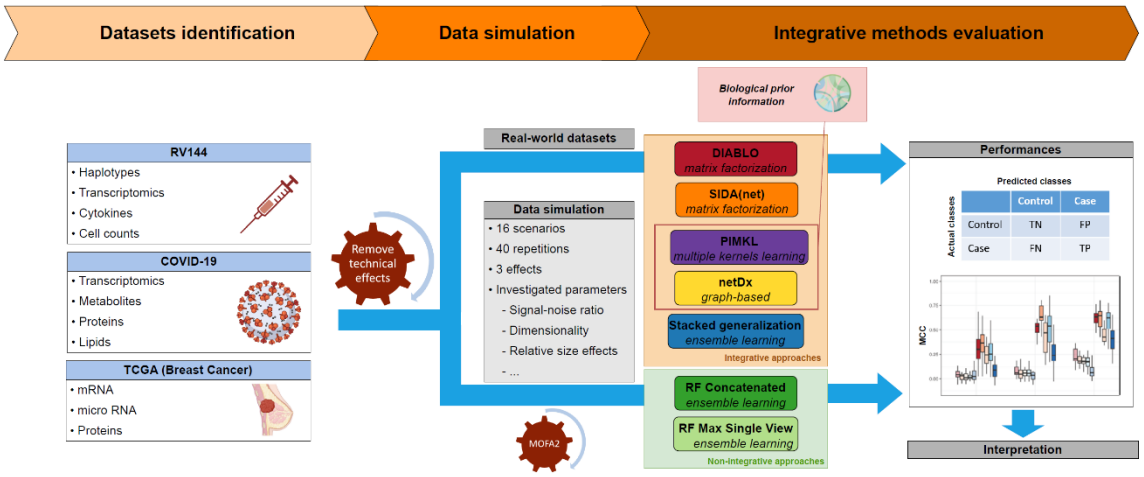

Figure 1 Three multi-omics datasets, covering distinct medical applications, were selected. A reference simulation scenario was designed using signal-to-noise ratio (SNR) and sparsity levels estimated from real-world datasets. 14 alternative scenarios were also generated by modifying class imbalance, SNR, dimensionality, relative importance of effects, etc. A selection of five integrative approaches, representative of existing methods, were evaluated on both real-world and simulated data based using MCC.

Context

The continuous progress made in omics technologies have reshaped our understanding of human biology. While single omics analyses have produced valuable insights, most common human diseases associated with high mortality (e.g. type 2 diabetes, cardiovascular disease) still lack effective therapeutic strategies. Moreover, diverse large-scale cancer projects, such as TCGA [3], ICGC [4], COSMIC [5], consistently demonstrated the power of data integration in patient stratification.

With a growing interest in extracting multi-omics features associated with health-related outcomes, an in-depth understanding of current supervised integrative approaches is much needed. In this context, a wide variety of integrative approaches have been introduced to address one or some of the following goals: (i) patient stratification, (ii) prediction of clinical outcome and (iii) identification of molecular mechanisms acting across molecular layers.

Methods

The current literature commonly distinguishes six main families of methods:

- matrix factorization
- multiple kernel learning
- network-based methods
- bayesian
- ensemble learning
- deep learning

We selected five methods (see Table 1) representative of the main families of integrative approaches (matrix factorization, multiple kernel methods, ensemble learning and graph-based methods) that we evaluated on a set of 15 simulations (see Table 2) exploring a large and realistic parameter space (e.g. sample size, dimensionality, confounding effects, effect size) and on real-world datasets. The datasets choosen were carefully selected to cover different medical applications (infectious diseases, oncology and vaccine).

As non-integrative control, two alternatives based on Random Forest (RF) were also included in this benchmark to evaluate the added value of data integration: RF_Concat concatenates omics layers sample-wise and evaluate the overall performance, while RF_Max_Single_View consists in evaluating RF on each modality and keeping only the highest classification performance.

	Name	Underlying approach	Prior information	Implementation
Integrative methods	DIABLO	Sparse generalized CCA	No	R package *mixOmics*
	SIDANet	Combination of LDA and CCA	Yes	R package *SIDA*
	PIMKL	Multiple kernel learning	Yes	Python script
	netDx	Integrated patient Similarity network	Yes	R package *netDx*
	Stacked generalization	Ensemble of weak learners	No	R package *SuperLearner*
Non-integrative methods	RF_Concat RF_Max_Single_View	Random Forest on concatenated data separated data	No	R package *randomForest*

Table 1 Summary of the methods selected in the benchmark.

Classification performance criterion

The Matthews Correlation Coefficient (MCC) is a popular metric used to evaluate the performance of binary classifiers. In a recent study, Chicco *et al.* [6] advocated for its use over accuracy and F1-score due to its robustness in imbalanced setting and invariance for class swapping. The MCC is defined as:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}$$

where TP and TN are the number of true positives and true negatives; FP and FN the number of false positives and false negatives. Like Pearson's correlation coefficient, MCC ranges between -1 and 1. ±1 reflects perfect (mis)classification, while 0 indicates random classification. To ensure an unbiased comparison, all methods were evaluated in 5-fold cross-validation.

Simulation scenarios

Scenario	Number of samples (cases, controls)	Number of features per omic	Main factor(s)	Fraction of signal features per omic	Overlap across factors
Reference	**80 (40, 40)**	**1000, 240, 60**	**All equal**	**0.1, 0.1, 0.1**	**0.5**
n/5	16 (8, 8)				
px5		5000, 1200, 300			
CaseControl_1:7	80 (10, 70)				
High_Main_MO			High main MO		
High_Conf_MO_Overlap			High confounding MO		0.95
Main_MO_2Smallest_Omics			Main MO kept in the		
Main_MO_1Largest_Omic			smallest/largest omic(s)		
High_Fraction_Signal_Feat				0.3, 0.3, 0.3	
nx5	400 (200, 200)				
p/5		200, 48, 12			
High_Conf_SO_Overlap			High confounding SO		0.95
Main_MO_1Smallest_Omic			Main MO kept in the		
Main_MO_2Largest_Omics			smallest/largest omic(s)		
Noise				0, 0, 0	

Table 2 15 scenarios were generated from real-world datasets. The reference scenario is defined by 2 classes of 40 samples each, 3 omics with p = (1000, 240, 60) variables and 3 factors (a main multi-omics, Main MO and two confounding factors acting at the single- and multi-omics levels, Conf SO, Conf MO). For the other scenarios, only the deviations from the reference are indicated.

Results on simulated data

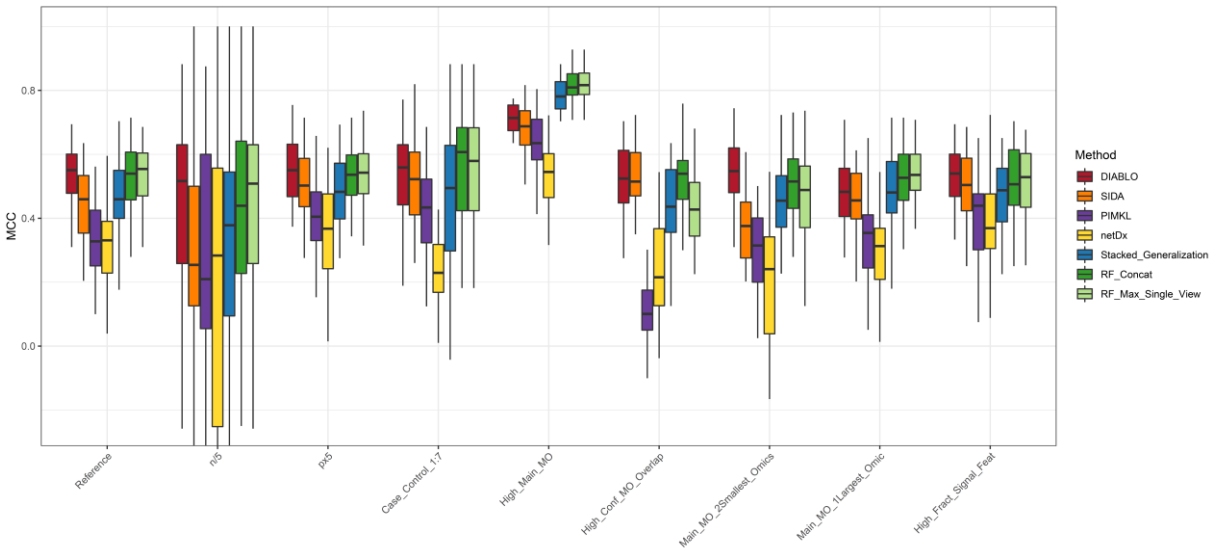

Figure 2 Method comparison on simulated data. Integrative approaches were evaluated on main 9 simulation scenarios. Two non-integrative methods (RF_Concat, RF_Max_Single_View) were also included to quantify the added-value of data integration. For each scenario, 40 repetitions were generated, on which, MCC was computed in 5-fold cross-validation.

Results on real-world data

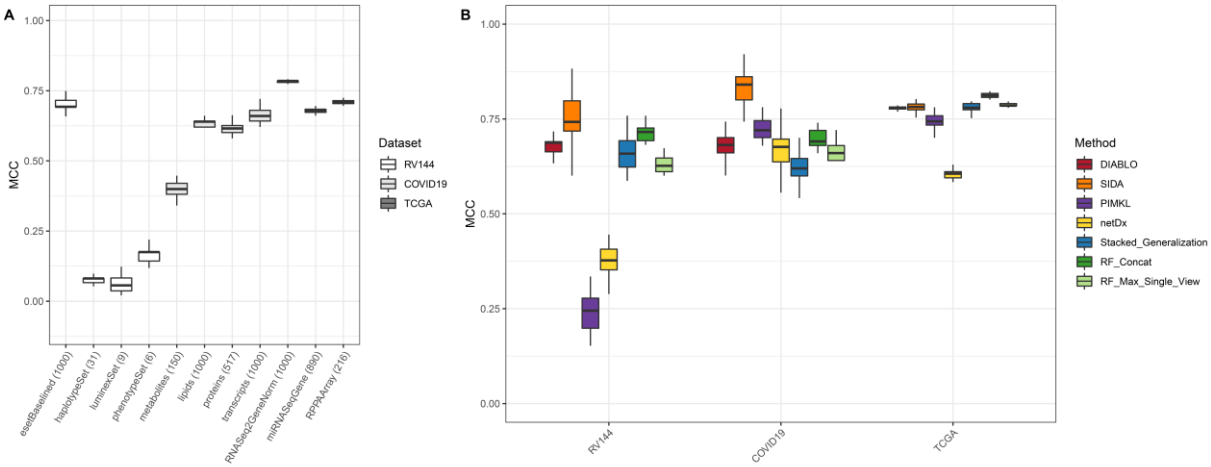

Figure 3 Method comparison on 3 real-world datasets. Prediction performance (A) on individual omic using Random Forest or (B) integrative methods. MCC was computed on 40 repetitions of 5-fold cross-validation.

Key points

- Supervised integrative methods have received little attention in the literature so far. In this work, five supervised methods spanning major families of integrative approaches are thoroughly evaluated.
- Non-integrative approaches (Random Forest based) were further included to elucidate the conditions in which data integration provides a clear advantage.
- In many simulation scenarios, Random Forest and latent variable models lead to comparable performances.
- When the main multi-omics effect is present in a subset of views, integrative approaches demonstrate their superiority. Conversely, when the multi-omics effect is strong, Random Forest outperforms its integrative counterparts.
- Among integrative approaches, latent variable models lead to the best performance on simulated (DIABLO) and experimental (SIDA) data.

References

1. Matteo Bersanelli et al. Methods for the integration of multi-omics data: mathematical aspects. BMC Bioinformatics, 17(S2):S15, December 2016.
2. Michal Krassowski et al. State of the Field in Multi-Omics Research: From Computational Needs to Data Mining and Sharing. Frontiers in Genetics, 11:610798, December 2020.
3. Katarzyna Tomczak, Patrycja Czerwi´nska, and Maciej Wiznerowicz. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemporary Oncology/Wsp´olczesna Onkologia, 2015(1):68–77, 2015. Publisher: Termedia.
4. Thomas J. Hudson (Chairperson), Warwick Anderson, Axel Aretz, et al. International network of cancer genome projects. Nature, 464(7291):993–998, April 2010. Number: 7291 Publisher: Nature Publishing Group.
5. John G. Tate, Sally Bamford, Harry C. Jubb, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Research, 47(D1):D941–D947, January 2019.
6. Davide Chicco and Giuseppe Jurman. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics, 21(1):6, January 2020.