International Journal of Advanced Scientific Innovation
ISSN: 2582-8436 Volume 05 Issue 06, June 2023

Harmonizing Multi-Omics for Enhanced Machine Learning

Praveen Kumar N K, Nayan Murthy

Jain University, Bengalore, Karnataka, India

[1]p.kumarnk@gmail.com

Abstract - The proliferation of high-throughput technologies has yielded an abundance of omics data, spanning diverse biological layers such as genomics, epigenomics, transcriptomics, proteomics, and metabolomics. Machine learning algorithms have harnessed this data deluge, yielding diagnostic and classification biomarkers. However, prevailing biomarkers predominantly rely on single omic measurements, overlooking the potential insights from multi-omics experiments that encapsulate the entirety of biological complexity. To fully exploit the wealth of information embedded in different omics layers, effective multi-omics data integration strategies become imperative. This minireview categorizes recent integration methods/frameworks into five strategies: early, mixed, intermediate, late, and hierarchical. Our focus is on delineating challenges and exploring existing multi-omics integration strategies, with a keen emphasis on their application in machine learning.

Keywords - ***High-throughput technologies, Omics data, Machine learning algorithms, Multi-omics experiments, Data integration strategies***

I. INTRODUCTION (SIZE 10 &BOLD)

The emergence of cost-effective and potent screening technologies [1] has ushered in a new era of extensive biological data, paving the way for advancements in therapeutics and personalized medicine [2]. Variances in treatment effectiveness and adverse effects among individuals, attributable to factors like age, sex, genetics, and environmental influences (e.g., anthropometric and metabolic status, dietary habits, lifestyle [3,4]), underscore the importance of precision medicine. The objective is to tailor interventions based on individual biological information [5]

Clinical and omics data can be sourced directly from databases or gathered through screening technologies for applications such as disease analysis [6], class prediction [7], biomarker discovery [8], disease subtyping, enhanced system biology understanding [9], and drug repurposing. Each omics data type represents a distinct "layer" of biological information, such as genomics, epigenomics, transcriptomics, proteomics, and metabolomics, offering complementary perspectives on biological systems or individuals. Historical single-omics studies aimed to uncover the causes of pathologies and guide appropriate treatments, but current understanding acknowledges the complexity of diseases involving intricate molecular pathways with interactions across different biological layers.

To navigate existing approaches, a classification system is essential for selecting suitable methods and identifying best practices. Zitnik et al. (2019) [10] categorized integration into horizontal and vertical types. This mini-review focuses on vertical integration, where each omics dataset shares the same rows (samples) but different variables (omics features). We assume that the datasets are already processed, normalized, or scaled based on their omics type. Existing general reviews on vertical integration [11] often categorize methods by mathematical aspects, such as Bayesian, network-based, deep learning-based, kernel-based, or matrix factorization-based methods.

II. Contributions

Multiple goals, including sample classification, disease subtyping, and biomarker discovery, can be achieved with multiple omics datasets. However, integrating these datasets, each with the same rows (representing samples) and different columns (representing biological variables), poses challenges. Machine learning (ML) models are commonly employed, but integrating multiple noisy and high-dimensional datasets requires careful consideration. Various integration strategies have been developed, each with its pros and cons. Assuming proper pre-processing of each dataset, a simple approach involves assembling datasets through sample-wise concatenation, creating a matrix used as input for ML models

III. Related Work

In multi-omics analysis, dimensionality reduction becomes crucial to decrease noise and simplify datasets. This optional step can be applied regardless of the chosen integration strategy, but some strategies (like early and intermediate integration) often benefit from prior dimensionality reduction Two approaches exist: feature selection, which removes noisy and redundant variables, and feature extraction, combining original variables into new and more meaningful ones [12]. In early integration, dimensionality reduction should consider the concatenated matrix to incorporate all omics. If performed separately on each dataset, there's a risk of information loss, placing it under another integration strategy. The following sections outline commonly used methods in both approaches, with specific reviews available for further details

Most omics datasets have high dimensionality, especially challenging in multiomics studies due to the number of datasets[13]. Feature selection addresses this by identifying a smaller set of features that retains relevant information while reducing dimensionality. This not only enhances computing efficiency but also improves model performance, interpretability, and mitigates the risk of overfitting. Feature selection can also address the block scaling problem by balancing the number of features in each omics block when many variables are removed.

Feature extraction (FE) methods strive to transform input features into a fresh set of variables, encompassing linear or non-linear combinations of the original features. The primary aim is to extract features in a way that preserves pertinent information while minimizing noise and redundancy [14]. While beneficial for exploratory data visualization and unveiling crucial features, FE methods introduce a trade-off by potentially compromising the interpretability of a model, given that the extracted features no longer directly represent biological measurements

These FE methods can be independently applied to each omics dataset, facilitating integration and block scaling in a mixed integration approach, or they can be implemented on concatenated multi-omics datasets in the context of early integration [15]. The resultant extracted features can serve as inputs for machine learning (ML) models or clustering. However, these strategies may inadvertently lead to redundancy and suboptimal results. Intermediate methods aim to overcome these challenges by concurrently analyzing datasets, yielding FE methods capable of considering all variables simultaneously

III. Methodology

Early integration involves consolidating all datasets into a unified matrix, amplifying the number of variables while maintaining the same number of observations. Challenges arise due to the resulting intricate, noisy, and high-dimensional matrix, making the learning process challenging. Imbalances in size between omics datasets may introduce learning biases, and early integration might overlook the distinctive data distribution of each omics layer, potentially guiding ML models toward irrelevant patterns. Despite these drawbacks, early integration remains popular due to its simplicity, ease of implementation, and the capacity to directly reveal interactions between different layers
Hierarchical integration incorporates regulatory relationships among diverse omics layers, mirroring the modular organization at the molecular level. This strategic approach leverages prior knowledge from interaction databases and scientific literature to enhance integration. Challenges in multi-omics integration are systematically addressed for each dataset, leveraging the organized nature of omics to mitigate integration complexities

Recent integration methodologies often involve modifying each dataset independently before integration. While informative, this approach may result in information loss and render models susceptible to noise. Early and intermediate integration strategies mitigate these concerns by simultaneously considering all datasets, yet challenges persist in effectively utilizing the resulting large matrix. Hierarchical integration, tailored to specific omics types, exhibits limitations in generalizability. As multi-omics research gains prominence, the identification of optimal practices and strategies becomes imperative. Benchmark studies, particularly those encompassing diverse ML models, are essential to steer future research and applications in multi-omics integration

In this mini-review, we presented the different strategies available to handle multi-omics datasets integration. Most integration approaches developed in recent years tend to first modify and transform each dataset using different machine learning models known as Mixed integration, in order to reduce their complexities and heterogeneities and facilitate their subsequent integration and analysis. While it can give informative results, each dataset is transformed

independently, potentially resulting in a loss of information and a final model that can still suffer from noise or redundant information. Ideally, at any point of the learning process, each omics dataset should be assessed while considering the other datasets, so that the complementary information could be best exploited. The early and intermediate integration strategies do solve this problem by integrating all datasets beforehand, but the large matrix resulting from an early integration is difficult to exploit by most ML models and intermediate integration often relies on unsupervised matrix factorization, which has difficulty incorporating the considerable amount of pre-existing biological knowledge. Another methodology, hierarchical integration, is explicitly designed with the prior understanding of how the different omics layers interact with each other. However, only few such methods have been developed and are often tailored for specific omics types, which makes them less generalizable than other approaches. Additionally, they are dependent on prior data, which prevents them from exploring and discovering new biological mechanisms and pathways.

With the ever-growing access to biological data, multi-omics research will be performed more and more often, and it is urgent that we identify the best practices, tools and strategies for their integration. In that aspect, benchmark studies are also particularly useful and should be done more frequently. With the notable exception of Herrmann et al. (2020) [168] which focused on survival prediction methods for multi-omics data, most benchmarks focus on clustering and dimensionality reduction methods [14]. Thorough comparisons of other ML models have not been made for multi-omics datasets, and we have yet to know if the deep learning prowess made in other fields of pattern recognition can be reproduced in bioinformatics.

IV. CONCLUSIONS

In conclusion, the field of mult i-omics integration is at the forefront of advancing our understanding of complex biological systems and holds immense potential for applications in therapeutics and personalized medicine. Feature extraction (FE) methods play a crucial role in transforming input features, enabling the creation of new variables that capture relevant information while mitigating noise and redundancy. However, it's important to recognize that the interpretability of models may be compromised as the extracted features no longer directly represent biological measurements.

Various integration strategies, such as early and intermediate integration, address the challenges of combining diverse omics datasets. Early integration, despite its drawbacks of increased complexity and potential bias, remains widely utilized for its simplicity and ability to directly uncover interactions between different layers. Intermediate integration strategies, on the other hand, offer a more nuanced approach by jointly analyzing datasets, providing a means to consider all variables simultaneously

The hierarchical integration strategy, incorporating regulatory relationships, reflects the modular organization at the molecular level. While this approach enhances integration by leveraging prior knowledge, it is currently limited in generalizability and application to specific omics types.

As the volume of biological data continues to grow, the importance of identifying best practices, tools, and strategies for multi-omics integration becomes increasingly urgent. Benchmark studies, particularly those encompassing diverse machine learning models, are essential to guide future research in this evolving field. While advancements have been made, further exploration and validation of integration methods are required to ensure robust and reliable outcomes in multi-omics research. Overall, the ongoing progress in multi-omics integration holds promise for unraveling intricate biological mechanisms and paving the way for more personalized and effective medical interventions

REFERENCES

[1] Labory J et al. Multi-omics approaches to improve mitochondrial disease diagnosis: challenges, advances, and perspectives. Front Mo l Biosci 2020;7:590842.

[2] Leon-Mimila P, Wang J, Huertas-Vazquez A. Relevance of multi-omics studies in cardiovascular diseases. Front Cardiovasc Med 2019;6:91.

[3] Jamil IN et al. Systematic multi-omics integration (MOI) approach in plant systems biology. Front Plant Sci 2020;11:944.

[4] Zitnik M et al. Machine learning for integrating data in biology and medicine: principles, practice, and opportunities. Inf. Fusion 2019;50:71–91.

[5] Higdon R et al. The promise of multi-omics and clinical data integration to identify and target personalized healthcare approaches in autism spectrum disorders. OMICS 2015;19:197–208.

[6] Sharifi-Noghabi H, Zolotareva O, Collins CC, Ester M. MOLI: multi-omics late integration with deep neural networks for drug response prediction. Bioinformatics 2019;35:i501–9.

[7] Tini G, Marchetti L, Priami C, Scott-Boyer M-P. Multi-omics integration—a comparison of unsupervised clustering methodologies. Briefings Bioinf. 2019;20:1269–79.

[8] Mantini G, Pham TV, Piersma SR, Jimenez CR. Computational analysis of phosphoproteomics data in multi-omics cancer studies. Proteomics 2021;21: e1900312.

[9] Dahal S, Yurkovich JT, Xu H, Palsson BO, Yang L. Synthesizing systems biology knowledge from omics using genome-scale models. Proteomics 2020;20: e1900282.

[10] Ahmed Z. Practicing precision medicine with intelligently integrative clinical and multi-omics data analysis. Hum. Genomics 2020;14.

[11] Chauvel C, Novoloaca A, Veyre P, Reynier F, Becker J. Evaluation of integrative clustering methods for the analysis of multi-omics data. Brief. Bioinform. 2020;21:541–52.

[12] Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucl Acids Res 2018;46:10546–62.

[13] Franco EF et al. Performance comparison of deep learning autoencoders for cancer subtype detection using multi-omics data. Cancers 2021;13.

[14] Menyhárt O, Gy}orffy B. Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis. Comput Struct Biotechnol J 2021;19:949–60.

[15] Nicora G, Vitali F, Dagliati A, Geifman N, Bellazzi R. Integrated multi-omics analyses in oncology: a review of machine learning methods and tools. Front Oncol 2020;10:1030.