Bridging the gap between AI and the life sciences: towards a standardized multi-omics data type

Laurin Herbsthofer, Monika Oberhuber, Barbara Prietl, Pablo López García[*]
CBmed - Center for Biomarker Research in Medicine
Stiftingtalstraße 5
8010 Graz (Austria)

Abstract

Omics data are key for the understanding of life and for improving human health but the contributions of AI in the field of multi-omics analysis are scarce when compared to single omics or medical imaging. We believe the major reason behind this fact is the lack of a standardized multi-omics data type. In this position paper, we introduce this problem, discuss some controversial aspects, and sketch a possible solution, as biomedical researchers clearly realized that there can be no real precision medicine without truly integrated multi-omics analysis and are desperately calling for collaboration. Our proposed multi-omics data type would provide a standardized way of storing raw and preprocessed multi-omics data together with preprocessing methods, therefore greatly simplifying data analysis and facilitating the participation of AI practitioners.

1 The need for a multi-omics data type in AI

Omics is a generic term that encompasses a number of disciplines in biology including genomics, transcriptomics, proteomics, metabolomics, and others, all of which are key for the study of life. These disciplines use high-end technology to obtain measurements that often result in massive amounts of data stored in a plethora of raw file formats that require heavy pre-processing [1]. The vast majority of analyses performed on these datasets has been at individual omics level, (e.g., searching for associations between genomic mutations and diseases) but recent studies have revealed that inner relationships between omics and outcome are much more complex than anticipated and simplistic models work worse than expected in practice [2].

Part of the reason for this complexity is the multi-layered and inter-dependent nature of omics. Some links between omics are well understood (e.g., DNA to mRNA transcription, or mRNA to protein translation) but others are not. However, all omics disciplines are important to get a full understanding of biological processes. Therefore, and given the advances and affordability of technology, the idea of collecting and simultaneously analyzing multiple omics data from individuals has gained a lot of traction. The advantages of multiple measurements from a single biological entity are many and drive single-cell multi-omics analysis: confounding factors are eliminated and unambiguous inference (such as genotype-phenotype) is possible [3].

A common approach of doing multi-omics analysis following fusion methods [4] and AI is to (1) merge all omics datasets, ideally having matched subjects for all of them, (2) extract relevant features, and (3) train a model for predicting a specific outcome in a supervised manner while trying to minimize overfit, given the classic p»n problem found in these datasets [5]. This was the approach we followed in a study to predict colon cancer stage II recurrence, based matched multi-omics (genomics, proteomics, metabolomics) and immunological data from 73 patients. As a result, we were able to

[*] Corresponding author: `pablo.lopez-garcia@cbmed.at`

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

reduce the initial 17,000 features to 52 relevant ones, and obtain excellent predictive performance when compared to using individual omics data only.

However, in this and other projects with academia and industry we found a recurrent issue: the urgent need for a standardized multi-omics data type that reflects the underlying multi-layered reality of biology and allows for easy data sharing, consumption, and analysis using AI. We believe that the lack of such data type is the main reason behind the worrying figure of 0.58% of multi-omics publications linking to reproducibility platforms [6] and the virtual non-existence of such datasets in platforms like Kaggle [7]. To the best of our knowledge, a standardized multi-omics data type does not currently exist, and how to make multi-omics data accessible not only to AI practitioners but to researchers in general is still a matter of debate [8].

2 Criticism

2.1 Multi-omics datasets are expensive to obtain and process

With historical figures of years of research and millions spent for sequencing the human genome, costly liquid chromatography-mass spectrometry metabolomics readings, and specialized laboratories for omics available in research centers only, it is questionable that multi-omics analysis is viable from an economic perspective.

Indeed, some omics technologies still remain complex and are available only in research facilities, but others have become widely available, affordable, and accessible. Take, for example, genomics, where commercial companies are offering 30x whole genome sequencing and access to raw file download for less than $300 directly to end consumers [9], or cost and time reductions in metabolomics which might make them feasible for use in clinical routine [10]. Given the current pace of technology, it would not be surprising that research centers or companies start offering affordable multi-omics services in the same way they already offer whole genome sequencing to end consumers. This is a scenario familiar to AI practitioners in terms of computation, with healthy competition between cloud service providers, offering extremely powerful, scalable, and affordable services for data storage and analysis, to the point where in many cases it is difficult to justify having on-premise expensive servers or clusters. There is no reason to believe it will not happen differently with omics laboratories and technology.

2.2 Single omics data types are sufficient in practical applications

For over a decade, the fact that targeted therapies based on genetic analysis, mutation-disease associations, and similar successful use cases with single omics, one would ask why the need to introduce extra layers of omics.

As mentioned in the introduction, all omics disciplines are strongly interconnected and form a series of layers that are key for the study of life. That is why many important clinical questions (e.g. colon cancer relapse risk or refractory carcinoma therapies) have not been answered yet by considering only the genetic layer but require additional biological layers to capture the biological processes are at play. The fact that the majority of studies and therapies are based on single omics corresponds mostly to technological and economical limitations of the time when they were started. Once multi-omics studies started to gain popularity and datasets released [8], researchers clearly realized that there can be no real precision medicine without truly integrated multi-omics analysis and are desperately calling for collaboration [11].

2.3 A multi-omics data type might introduce extra complexity for AI development

AI projects where omics data processing is required often require a multidisciplinary team that includes not only AI experts but often bioinformaticians and DevOps. The introduction of a new data type is risky because it would add an extra layer of complexity to a scenario which is already very complex.

Such criticism sounds convincing, but experience and literature [11] shows otherwise. When trying to apply AI to omics datasets, there is a first step to transform raw biological data into some kind of tabular format accessible to researchers that needs to be again transformed into a machine-readable AI-ready dataset (generally a tidy dataset [12]). After that, multi-omics datasets need to be analyzed together in a way that is still open to debate. In our opinion, it precisely is the lack of standards that explains the worrying figures mentioned before (0.58% of multi-omics publications linking to reproducibility platforms and the virtual non-existence of such datasets in platforms like Kaggle). On the one hand, people generating multi omics datasets are not necessarily computer scientists, software developers, or AI experts and are uncomfortable releasing datasets that are difficult to consume; and vice versa: it could take weeks to months before AI practitioners are comfortable understanding omics raw files or unprocessed datasets, discouraging them from collaborating or being interested in solving such problems. Compare this with the huge contributions of AI to the field of medical imaging, the main reason being that the data type is understandable and easy to consume.

3 Proposal

It is well known that every multi-omics dataset requires different pre-processing steps. These can be a daunting task, considering that bioinformaticians typically specialize in only one or two omics techniques. We therefore propose a standardized multi-omics dataset that would provide pre-processed data and/or associated methods to turn raw data into usable features in a single software package. Aside from facilitating collaboration, this would encourage the sharing of datasets collected at different institutions, make them more comparable, and increase the potential training data for AI applications.

A draft of how it would look like in pseudo-code is given below.

```
import multiOmicsFormat as mof

Data = mof.downloadPatients([001:009])
NGS = Data.ngs.get_features()
Metabolomics = Data.metabolomics.get_features()
Target = Data.clinical.get_features('disease')

Model = randomForest.fit([NGS, Metabolomics], Target)
Features = Model.relevantFeatures()
```

4 Conclusion

Omics data are key for the understanding of life and for improving human health but the plethora of raw formats and specialized bioinformatics knowledge required to bring them to a state where AI algorithms can be applied limit their use. Moreover, not only one but multiple layers of omics data are often needed to provide the answers the life science community is looking for, further complicating the issue. We believe that the lack of a standardized data type has severely limited the contribution of the AI community in the multi-omics area, which pales in comparison to contributions in others like single omics or medical imaging. Current trends in technology and biomedical research indicate that the number of multi-omics datasets will increase and they will become essential to develop precision medicine. So will the cost of opportunity for the AI community to contribute if no action is taken.

In this paper, we introduced this problem, discussed some controversial aspects, and proposed the creation of a standardized multi-omics data type to be consumed by AI algorithms. Our proposed multi-omics data type would provide a standardized way of storing raw and preprocessed multi-omics data together with preprocessing methods, therefore greatly simplifying data analysis and making it available to AI practitioners. We hope this paper shows the desperate need for a multi-omics data type that bridges the gap between omics and AI and accelerates research and development of precision medicine.

References

[1] Griffin PC, Khadake J, LeMay KS, Lewis SE, Orchard S, Pask A, Pope B, Roessner U, Russell K, Seemann T, Treloar A. Best practice data life cycle approaches for the life sciences. F1000Research. 2017;6.

[2] Singh RS, Gupta BP. Genes and genomes and unnecessary complexity in precision medicine. NPJ genomic medicine. 2020 May 4;5(1):1-9.

[3] Macaulay IC, Ponting CP, Voet T. Single-cell multiomics: multiple measurements from single cells. Trends in Genetics. 2017 Feb 1;33(2):155-68.

[4] Baldwin E, Han J, Luo W, Zhou J, An L, Liu J, Zhang HH, Li H. On fusion methods for knowledge discovery from multi-omics datasets. Computational and structural biotechnology journal. 2020 Jan 1;18:509-17.

[5] Biswas N, Chakrabarti S. Artificial intelligence (AI)-based systems biology approaches in multi-omics data analysis of cancer. Frontiers in Oncology. 2020;10.

[6] Krassowski M, Das V, Sahu SK, Misra BB. State of the field in multi-omics research: From computational needs to data mining and sharing. Frontiers in Genetics. 2020;11.

[7] Kaggle. https://www.kaggle.com/

[8] Conesa A, Beck S. Making multi-omics data accessible to researchers. Scientific data. 2019 Oct 31;6(1):1-4.

[9] Nebula Genomics. https://nebula.org/whole-genome-sequencing-dna-test/

[10] Bordag N, Zügner E, López-García P, Kofler S, Tomberger M, Al-Baghdadi A, Schweiger J, Erdem Y, Magnes C, Hidekazu S, Wadsak W. Towards fast, routine blood sample quality evaluation by Probe Electrospray Ionization (PESI) metabolomics. medRxiv. 2021 Jan 1.

[11] Olivier M, Asmis R, Hawkins GA, Howard TD, Cox LA. The need for multi-omics biomarker signatures in precision medicine. International journal of molecular sciences. 2019 Jan;20(19):4781.

[12] Wickham H. Tidy data. Journal of statistical software. 2014 Sep 12;59(1):1-23.