Trends in Cell Biology

CellPress OPEN ACCESS

Spotlight

Data-driven bioinformatics to disentangle cells within a tissue microenvironment

Jesper N. Tegner [1,2,3,4,*] and David Gomez-Cabrero [1,5,6]

Molecular profiling of clinical tissue samples is at the core of precision medicine. Yet, to elucidate the contribution of mixed cell types and detect changes in cell populations in response to infections or drugs is challenging. Recent advances using machine learning promise to learn explanatory models directly from data.

Since the human genome project, we have increasingly powerful tools for deep molecular characterization of genes, proteins, and metabolites. Contrasting tissue samples from patients and controls enables the discovery of biomarkers and insights into the effects of therapies. Yet, tissue microenvironments are a complex mixture of cell types. A palette of intratumoral immune cells can predict tumor progression and recurrence [1]. Decomposing the sample signal from a microenvironment has attracted tremendous interest in the clinical genomics community [2]. It is referred to as cellular deconvolution in bioinformatics [3].

We can view the cellular deconvolution problem as a molecular version of the 'cocktail party effect'. Imagine when standing in a crowd, you can focus your attention on the speech of one particular person within the crowd. This is commonly referred to as the cocktail party effect and is an example of the blind source separation problem [4]. To decompose the tissue signals into its sources amounts to not only finding (or focusing the attention to) 'one singular voice' (cell type) out of a crowd, but rather finding all different voices (cell types and their dialects) contributing to the sound (the tissue signal) you are hearing (profiling). To illustrate the problem in its simplest form, let us imagine obtaining a composite signal such as a number 42. The take-home message is that this problem cannot be solved uniquely, even with extreme (unrealistic) assumptions. Let us assume we know that there are precisely two sources and we have an accurate description of the sources (their signature). Yet, with sources number 2 and number 3, there are still seven distinct linear solutions yielding 42, assuming that there is no complicated interaction between the sources, just simple linear addition.

A large body of computational research has been devoted to deconvolute tissue profiles into cellular sources [3,5]. The leading idea of signature-based methods such as CIBERSORT or xCELL is that each cell type has a specific molecular expression profile (Figure 1). This 'makes' the problem tractable, similar to the '42 problem'. However, since the profiles are subject to biological variability and technical noise, standard regression techniques are used to 'fit the data' to the tissue profile (i.e., forcing a unique solution). The first limitation is that the cellular signature of a given cell type, say a T cell, may change between a healthy and a disease state, thus altering the proportions of cells. In contrast to static molecular signatures, molecular profiles are a continuum depending on the cellular state, biological condition, and cell type rather than a set of discretely clean profiles [6,7]. Secondly, the assumption of the different sources' effectively independent (e.g., linear) contribution is unrealistic. It is difficult for this class of methods relying on regressing over predefined molecular profiles to be robust over a broad range of clinical conditions, sequence platforms, batch effects, experimental conditions, and sample types [3]. Thus, the assumptions required for 'solving' the 42 problem are unrealistic from a biological standpoint.

Instead, we would like to learn unbiasedly about molecular profiles and their nonlinear combinations from data. This is where we would expect machine learning to be helpful. In short, to select 'relevant' features (genes) for regression while also creating new features that are optimal for the regression task. Recently, a deep neural network (DNN) method (Scaden) used bulk RNA-seq mixtures simulated together with single-cell data [8] (Figure 1). The DNN was trained on *in silico* bulk RNA-seq data derived from single-cell profiles to predict the cell type proportions from a bulk expression of cellular mixtures. Yet, DNN methods are extremely data-hungry. Furthermore, a specific limitation is that such training also assumes that the testing and training conditions must match across different conditions.

Thus, we seek a data-driven method without having to assume predefined molecular features, which at the same time can be trained such that individual training and testing conditions are respected. This is the contribution of the data augmentation method using an *in silico mixing (*DAISM)-DNN method [9]. Their strategy to overcome data limitations is to train the system on data derived from a DAISM strategy (Figure 1). The method generates additional training samples by computing weighted combinations of: (i) 'calibrated data' (known cell type composition); and (ii) a single sample derived from 'simulated pseudobulk derived from single cell or purified bulk-RNA samples'. When pseudobulk single cell-derived data is considered, the pseudo profile is derived from 500 cells. Cells are randomly selected following the precomputed proportion. The final pseudobulk is calculated from the aggregation for all cells and a transcripts per million normalization. When pseudobulk bulk derived is considered, a weighted

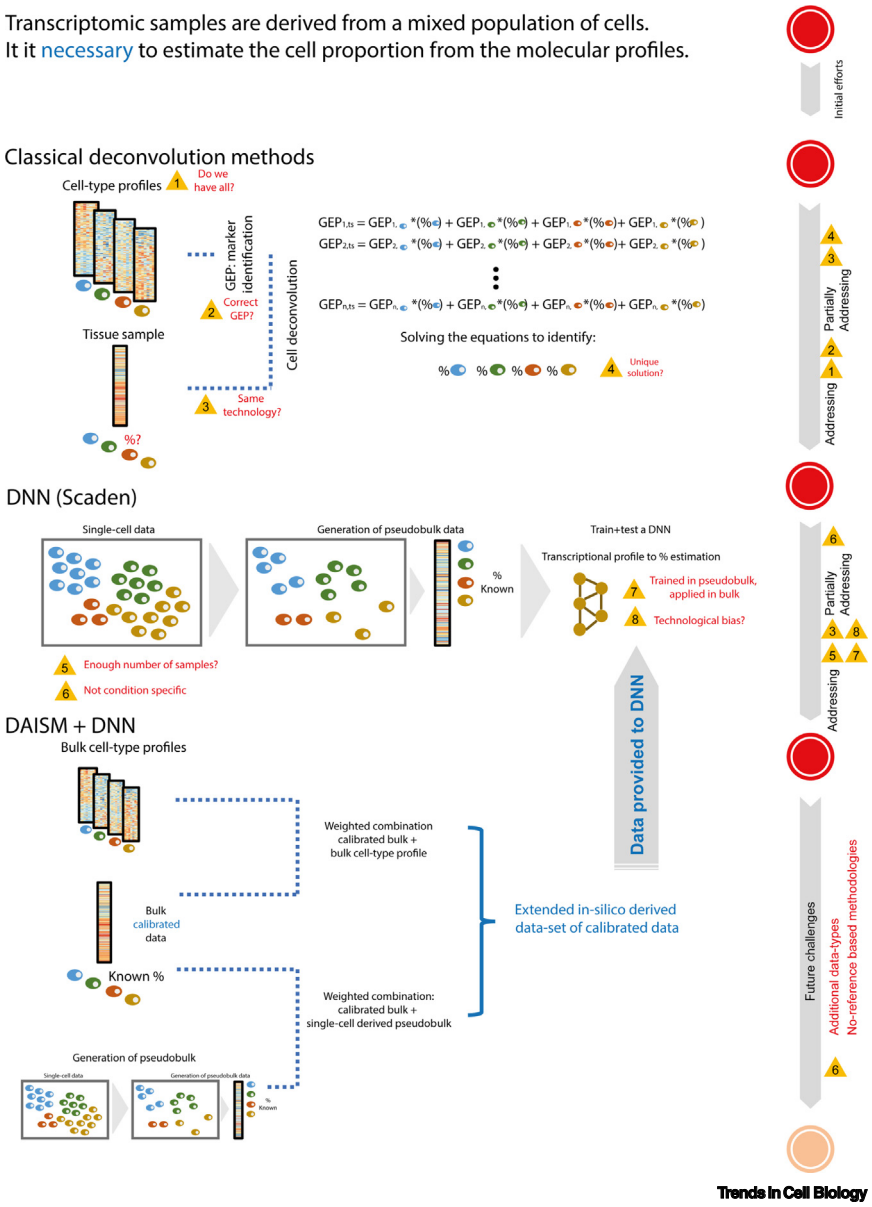

Figure 1. Cell proportion estimation algorithms. Top: flow of method development starting with 'marker-based methodologies'. Middle: Scaden, a deep neural network (DNN) formulation. Bottom: data augmentation strategies as implemented in data augmentation method using an *in silico mixing* (DAISM)-DNN. A numbered yellow triangle identifies the limitations of each method. At the right side of the figure, the improvements provided by the method directly below are identified. Addressed (partially addressed) denotes limitations overcomed (partially overcomed) by the method below. The total set of markers for all cell types is indicated by n. Abbreviations: GEP, gene expression profiles.

combination of the cells types is computed using the precomputed proportions.

The cell type composition and the associated transcriptomic profiles are known in both cases. Therefore, the new sample will be a weighted combination of them and, as a result, a new data point can also be considered a calibrated data point. Metrics such as Pearson correlation between predicted and fluorescence activated cell sorting-derived cell type proportions, Lin's concordance correlation coefficient, and root mean square error were used to quantify the performance of DAISM-DNN. They show that DAISM-DNN robustly outperforms the signature regression-based methods and Scaden. DAISM-DNN also generates reliable models using only a limited number of calibrated samples, making it into a 'train once, reuse many times' system, which is 'platform agnostic'.

The DAISM-DNN demonstrates the power of machine learning for deconvoluting mixed tissue profiles while respecting that different experimental conditions and disease states alter the molecular profiles of cells. Next, to extend their framework (https://github.com/xmuyulab/DAISM-XMBD) to other data modalities, such as DNA methylation and assay for transposable-accessible chromatin using sequencing data, appears natural. Yet, the current generation of machine learning techniques used for cellular deconvolution requires a supervised learning setup in that it requires annotated training data. One exciting avenue for future work is to develop unsupervised machine learning techniques in combination with ideas from signal source separation to supplement the analysis when limited by the availability of condition-specific annotated molecular profiles. This may open the door to disease- and condition-specific mathematical modeling [10].

Acknowledgments

J.N.T and D.G-C acknowledge support from King Abdullah University of Science and Technology.

Declaration of interests

The authors declare no competing interests.

[1]Bioscience Program, Bioengineering Program, Biological and Environmental Science and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia
[2]Computer Science Program, Statistics Program, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia

[3]Unit of Computational Medicine, Department of Medicine, Center for Molecular Medicine, Karolinska Institutet, Karolinska University Hospital, L8:05, SE-171 76, Stockholm, Sweden
[4]Science for Life Laboratory, Tomtebodavägen 23A, SE-17165, Solna, Sweden
[5]Mucosal and Salivary Biology Division, King's College London Dental Institute, London, UK
[6]Translational Bioinformatics Unit, Navarrabiomed, Complejo Hospitalario de Navarra (CHN), Universidad Pública de Navarra (UPNA), IdiSNA, Pamplona, Spain

*Correspondence:
jesper.tegner@kaust.edu.sa (J.N. Tegner).

https://doi.org/10.1016/j.tcb.2022.03.009

References

1. Avila Cobos, F. *et al.* (2020) Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nat. Commun.* 11, 5650
2. Bindea, G. *et al.* (2013) Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. *Immunity* 39, 782–795
3. Gomez-Cabrero, D. *et al.* (2016) High-specificity bioinformatics framework for epigenomic profiling of discordant twins reveals specific and shared markers for ACPA and ACPA-positive rheumatoid arthritis. *Genome Med.* 8, 124
4. Kiselev, V.Y. *et al.* (2019) Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat. Rev. Genet.* 20, 273–282
5. Lin, Y. *et al.* (2022) DAISM-DNNXMBD: highly accurate cell type proportion estimation with in silico data augmentation and deep neural networks. *Patterns* 3, 100440
6. Menden, K. *et al.* (2020) Deep learning-based cell composition analysis from tissue expression profiles. *Sci. Adv.* 6, 1–12
7. Mohammadi, S. *et al.* (2017) A critical survey of deconvolution methods for separating cell types in complex tissues. *Proc. IEEE* 105, 340–366
8. Pal, M. *et al.* (2013) Blind source separation: a review and analysis. In *International Conference Oriental COCOSDA*, pp. 1–5
9. Stein-O'Brien, G.L. *et al.* (2021) Forecasting cellular states: from descriptive to predictive biology via single-cell multiomics. *Curr. Opin. Syst. Biol.* 26, 24–32
10. Tegnér, J.N. *et al.* (2009) Computational disease modeling - fact or fiction? *BMC Syst. Biol.* 3, 56