Inferring cellular heterogeneity with mixture models for DNA methylation rates

Hugo Barbot*[1], David Causeur[1], Magali Richard[2], and Yuna Blum[3]

[1]Institut de Recherche Mathématique de Rennes – Université de Rennes, Institut National des Sciences Appliquées - Rennes, École normale supérieure - Rennes, Université de Rennes 2, Centre National de la Recherche Scientifique, Institut Agro Rennes ANgers – France

[2]Laboratoire d'Informatique de Grenoble – Institut National de Recherche en Informatique et en Automatique, Centre National de la Recherche Scientifique, Université Grenoble Alpes, Institut Polytechnique de Grenoble - Grenoble Institute of Technology – France

[3]Institut de Génétique et Développement de Rennes – Université de Rennes, Centre National de la Recherche Scientifique, Structure Fédérative de Recherche en Biologie et Santé de Rennes, Structure Fédérative de Recherche en Biologie et Santé de Rennes – France

Abstract

Cellular heterogeneity in biological tissues reflects progression of disease state and is therefore valuable for improving diagnostic and prognosis. Cellular composition of tissues is however difficult to assess from bulk molecular profiles. Specifically, the recorded signals represent aggregated contributions from all constituent cell types. Cell deconvolution is a common approach to unravel the heterogeneous molecular profiles observed in bulk tissues, by inferring the underlying relative abundance of constituent cell populations. Existing approaches assume that bulk omic profiles result from weighted sums of so-called signature cell-specific omic profiles, where weights represent the unknown proportions of those cell types. Consistently, most statistical methods used for cellular deconvolution are based on extensions of the Ordinary Least Squares (OLS) optimization algorithm, under non-negativity and sum-to-one constraints on the estimated mixing coefficients. Using OLS implicitly assumes independence, homoscedasticity and normality of the residual errors, conditions under which OLS optimization guarantees optimal estimation. In cellular deconvolution applied to bulk molecular profile, all three assumptions are highly questionable. Indeed, strong violations of those assumptions may be due to the intrinsic nature of omics data: genomic features are generaly overdispersed and dependencies among them arise from underlying gene regulatory networks.

The goal of this work is to provide a well-defined statistical framework for deconvolution that respects the inherent characteristics of biological data. Among available omics data types, RNA-seq gene expression and DNA methylation are most frequently used for deconvolution. We focus here on DNA methylation data, which provide complementary information to gene expression and are particularly useful when RNA quality is limited or degraded. Whole-genome cell-type specific distributions of DNA methylation rates actually show a latent group structure, that can explain poor estimation accuracy when fitting deconvolution models on the whole genome. Therefore, we propose a mixture of non-negative

*Speaker

beta regression models, estimated via an EM (Expectation-Maximization) algorithm, explicitly accounts for this latent structure. In this framework, selecting the optimal mixture component corresponds to an implicit gene selection step. Therefore, identifying the best latent component is decisive. We propose a component selection criterion that balances two objectives: (i) within-component improvement of the fit of the cell deconvolution model, and (ii) estimation stability evaluated by the condition number of the asymptotic variance-covariance matrix of the cell type proportions estimator. We evaluate the proposed approach through an extensive comparative study on several benchmark datasets: an *in vitro* mixtures of isolated pancreatic cancer cell populations, an *in vitro* mixtures of isolated immune blood cell populations, and *in vivo* blood cell data. Our results demonstrate the strong sensitivity of deconvolution performance to the latent component choice and highlight the significant accuracy gains achieved when using genes from the best-performing component compared to whole-genome deconvolution. Results confirm both the marked sensitivity of cell deconvolution performance to the latent component choice, and highlight the significant accuracy gains using the geneset within the best latent component compared to whole-genome deconvolution. These findings establish mixture modeling of whole-genome methylation data as a promising methodological perspective that enhances both the accuracy and interpretability of cellular deconvolution.

Keywords: Cell deconvolution, Mixture models, Beta regression