Statistical Discovery of Transcriptomic Cancer Signatures using Multimodal Local Search

Emile Zakiev[1] Johann Dreo[1,2] Mara Santarelli[1] Benno Schwikowski[1]

[1]Computational Systems Biomedicine Laboratory, Department of Computational Biology,
[2]Bioinformatics and Biostatistics hub, Université Paris Cité,
Institut Pasteur, 25–28 rue du Docteur Roux, 75015, Paris, France,
{ezakiev, johann.dreo, mara.santarelli, benno.schwikowski}@pasteur.fr

Keywords: Transcriptomics, Multimodal Optimization, Local Search, Cancer, scRNA-seq.

Cancer exhibits *intra-tumor heterogeneity* of gene expression that can be profiled using *single-cell RNA sequencing* technology. 'Precision' treatment strategies attempt to exploit this for treatment recommendations tailored to patient-specific heterogeneity patterns in specific sets of genes (*signatures*). Identification of treatment-relevant signatures requires their observation across multiple patients. However, *inter-patient heterogeneity* makes direct integration of gene expression data across patients, and subsequent identification of signatures, unreliable.

For the case of glioblastoma, Neftel *et al.* [1] have circumvented this problem using an approach that first heuristically determines a candidate set of patient-specific signatures, and then identifies similar candidates across patients, to obtain biologically validated results.

We expand and improve upon this approach using: (i) a statistically well-founded approach to score general signatures in individual patients, (ii) robust ranks instead of normalized RNA expression levels, (iii) a straightforward extension of the patient-specific score to a global score across all patients, and (iv) a *gradient* structure of the global score function.

Since this binary partitioning problem is NP-complete, we use a randomized optimization method in a multimodal setup, across 10,000 runs. Our greedy algorithm identifies signatures by starting from a random gene set, then iteratively moves to the best signature in its neighborhood, using an efficient partial evaluation of the objective function, until a local optimum is found.

Out of the seven 50-genes signatures that we found in the *glioblastoma* data set of 7,167 genes and 6,855 cells, five had a high degree of similarity to all eight of the metaprograms from the original study, with some of our signatures including genes from two metaprograms simultaneously. Gene set enrichment analysis of one of the remaining two signatures identified a specific neuronal process that is biologically plausible within the biological context.

Our approach is free from ad-hoc thresholds, simple, transparent, robust, and can yield biologically plausible results. We believe that our approach allows for bypassing the need for a complicated process of generating individual signatures in every sample and their further integration, and hence represents a useful addition to the tool belt of methods tackling the signature search problem.

References

[1] C. Neftel et al. An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma. *Cell*, 178(4):835–849, 2019.