Improving Deconvolution Methods in Biology through Open Innovation Competitions: an Application to the Connectivity Map

Author 1 Author 2 ...

Last updated: May 14, 2019

Abstract

Report results fo open innovation competition aimed at solving a gene-related deconvolution problem.

Keywords: biology; open innoation competitions; crowdsourcing; deconvolution; gene expressions; cell lines.

Contents

1 Introduction

Deconvolution problems arise in a variety of settings, such as regression, natural language, image, and signal processing, and methods developed to solve these problems have applications in fields as diverse as genetics, microscopy, geology, astronomy, engineering, among others.

Many studies have shown significant benefits for drug discovery from the systematic analysis of large data repositories of gene-expression profiles [Refs]. Traditional gene-expression profiling multianalyte methods, such as Luminex profiling technology, however, are limited by the type and number of available analytes [Refs]. Therefore, the cost of big data generation in biology remains prohibitive.

- Compound profiling is essential step in drug discovery
- biomedical big data generatioon is costly.
- CMap addressed problem by using the same Luminex beads by two types of mRNA probes requires a reliable deconvolution approach. Furthermore, biases introduced by batch effects need subtle normalization and quality control methods.

In biomedical research, our focus here, deconvolution problems are common in multianalyte assay methods. These methods are widely used to do X, Y and Z. In general terms, multianalyte assay methods are based on microspheres with different fluorescence decay times. This feature can be used to do X, Y, and Z. [EXPLAIN BRIEFLY CMap PROBLEM]. One problem with existing approaches is that they [....] In mathematical terms, the problem is to solve the equation of the form: y = x * z + e where y is the recorded signal, x is the signal we want to obtain, z is a confounding signal, and e is noise. Solving this equation is difficult because we do not observe e and z. [ADAPT THIS GENERAL TO CMAP] Many solutions have been developed: Parametric (maximum likelihood) Non-Parametric (k-means), blind deconvolution For CMAP: Current solution based on k-means achieves good performance. However, potential limitations are: computational time is high (e.g., ~20 minutes) documented examples of detection mistakes Theoretically it is biased when signals overlap Trade-off between accuracy and computation time unknown and perhaps suboptimal To identify accurate methods we launched an open challenge that allowed a rapid exploration of different approaches. Key ingredients of there challenges are: training and testing dataset benchmark solution to improve

1.1 Statistical deconvolution of gene-specific expression profiles.

Assume fluorescent-intensity values X_{ij} for beads $i = 1, 2, \ldots, n$ and analytes $j = 1, 2, \ldots, J$, and gene-specific proportions w_{ik} for beads $i = 1, 2, \ldots, n$ and genes $k = 1, 2, \ldots, K$. Our model of analyte fluourescent intensity is:

$$X_{ij} = \sum_{k=1}^{K} w_{ik} h_{kj} + e_{ij}.$$

where h_{ik} is the gene-expression value for genes $k = 1, 2, \ldots, K$ and analytes $j = 1, 2, \ldots, J$.

For the UNI detection method, the gene-specific proportions are such that each analyte has only one gene. Hence, $w_{ik}^{\text{uni}} = 1$ when $j = k$, and it is zero otherwise. This implies that each sample can detect at most J different genes under the UNI method.

For the DUO detection method, the gene-specific proportions are such that each analyte is paired with two genes in 1:2 ratio. Hence, pick an element $g \in G^2$ from the set G^2 of all non-overlapping subsets of size two of the gene set G. For each pair of genes in g associated with an analyte j, we have: $w_{i1}^{\text{duo}} = 2/3$, $w_{i2} = 1/3$ and is zero otherwise.

1.2 To read

- Compound signature detection on LINCS L1000 big data used a fuzzy c-means Gaussian Mixture Model (GMM) to process raw L1000 data, showing better performance compared to KNN. This method is described below:

  To deconvolute such overlapped peaks, we assumed that the fluorophore intensities of each analyte type (corresponding to a specific mRNA type) had a Gaussian distribution. The distribution of the mixture of analytes GeneH(i) and GeneL(i) corresponding to the expression levels of GeneH and GeneL, respectively, should be subject to a bimodal Gaussian distribution, with the proportion of 1.25 to 0.75. We initialized the estimations of the two Gaussian distributions using buzzy c-means clustering [11] and estimated the GMM parameters using the Nelder-Mead method [12]. Thus, the overlapped peaks were deconvoluted as the two estimated Gaussian peaks and the expression levels of the two genes sharing the same analyte were extracted. Mathematical details are included in the Supplementary Methods (the GMM model).

- Deconvolution of linear systems by constrained regression and its relationship to the Wiener theory
- Efficient Bayesian-based multiview deconvolution
- A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis
- Gene expression deconvolution in linear space
- Cell type–specific gene expression differences in complex tissues

2 References