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

Author 1 Author 2 ...

Last updated: Aug 22, 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

TODOS:

- Intro AB
- Methdos TN
- Figure Scatter runtime vs accuracy for top10 TN
- Add data

Here's a figure 1.

1 Introduction

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

Using an assay called L1000, The Connectivity Map (CMap) group at the Broad Institute has developed a novel approach that matches pairs of genes to the same Luminex beads to double the count of profiled genes per bead, thus lowering costs [Subramanian 2017]. A central component of this approach is to quantify gene-type frequency of beads, and then statistically deconvolve and compare gene type-specific average expression profiles for pairs of mixed gene samples.

This type of deconvolution problems are ubiquitous and have a long history in biology [refs]. For example, deconvolution problems arise when trying to identify cell type–specific gene expression differences in complex tissues [Shen-orr et al., 2010]; or in the discovery of target proteins of small molecules [Jung, 2015].

Common approaches are parametric (mixture models) and non-parametric. CMap's current solution to this problem is an algorithm, called "D-Peak," based on a K-means clustering [Refs]. This method automatically partitions a set of gene-expression measurements into k clusters, taking the mean of the two largest clusters, assigning the mean value of the largest cluster to the gene in higher proportion and the mean value of the second largest cluster to the gene in . This works well but has several problems as well. [List problems.] Including time. The trade-off between accuracy and computation time is currently unknown.

Alternative methods are well known, such as Gaussian etc. But it would have required substantial resources to experiment with these alternative approaches (more than what already done) and to adapt new to our data. Moreover, impossible an exhaustive search for all available approaches to try; and the combination of these different approaches.

Instead, we used an open innovation competition as a research tool to engage a variety of computer scientist, software developers and bioinformatics in the problem. This approach allows simultaneous exploration of competing approaches tailored to our problem, at no cost.

2 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 [. . . .]

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

2.1 L1000 Experimental Scheme

The L1000 assay uses Luminex bead-based fluorescent scanners to detect gene expression changes resulting from in vitro perturbation experiments [Subramanian 2017]. In a single experiment, CMap makes 488 measurements, each made by a different colored bead. Each measurement produces an intensity histogram (a list of integers), which characterizes the expression of two distinct genes in the given sample (for a total of 488 x 2 = 976 genes). The genes are mixed in 2:1 ratio. Thus, the areas under the peaks have 2:1 ratio, which enables association of each peak with a specific gene. In the ideal case, each histogram consists of two peaks (see Figure 1), each corresponding to a single gene.

After data collection, the expression values of the two genes are computationally extracted in a process called ‘peak deconvolution,’ described further in the next section.

2.2 Statistical deconvolution of gene-specific expression profiles.

In each sample, 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.

2.2.1 Benchmark k-means solution

CMap's current solution to this problem is based on a K-means clustering algorithm called dpeak that works as follows:

For each measurement, the dpeak partitions the list of realizations into K>=2 distinct clusters and identifies two of the clusters whose ratio of membership is as close as possible to 2:1. The algorithm then takes the median intensity of each of the two clusters, assigning these values to the appropriate gene (i.e., matching clusters with more observations to the gene mixed in higher proportion).

After deconvoluting each sample on a plate, dpeak then uses the plate-wide distributions to perform adjustments on a per-well basis, correcting peaks that may have been misassigned (see Appendix).

Known problems with the current approach are that the K-means algorithm is generally a biased and inconsistent estimator of the peaks of a bimodal distribution and it fails sometimes to detect peaks with few observations or it incorrectly identifies these peaks as 3rd peaks and disregarding them. Another limitation is that it is computationally costly and, hence, expensive for data generation at large scale (current Matlab implementation takes about 30 minutes on a 12-core server to process one set of 384 experiments).

2.3 Data generation

To generate data for this contest, we profiled six 384-well perturbagen plates, each containing mutually exclusive sets of compound and shRNA treatments. Multiple treatment types were used to ensure avoid potentially over-fitting to any one. The compound and shRNA perturbagen plates were arbitrarily grouped into pairs, and to avoid any potential 'information leakage' each pair was profiled in a different cell line. The resulting lysates were amplified by ligation mediated amplification (LMA, Subramanian 2017). The amplicon was then split and detected in both one-gene-per-bead (UNI) and two-genes-per-bead (DUO) detection modes. The three pairs of data were arbitrarily assigned to training, testing, and holdout categories, where in each case the UNI data served as the ground truth.

The data so generated were then split into three subsets called: training, provisional-testing, and system-testing. Training and provisional-testing data were made available for all the contestants to develop and validate their solutions, while system-testing data were secured to evaluate competitors' last submissions, which was used to award the prizes, and to avoid overfitting as well.

In UNI detection mode, each of the 978 landmark genes is measured individually on one of the 500 Luminex bead colors. In order to capture all 978 genes, two detection plates are used, each measuring 489 landmarks. The two detection plates' worth of data are then computationally combined to reconstruct the full 978-gene expression profile for each sample, but deconvolution is not required.

By contrast, in the DUO detection scheme two genes are measured using the same bead color. Each bead color produces an intensity histogram which characterizes the expression of the two distinct genes. In the ideal case, each histogram consists of two peaks each corresponding to a single gene. The genes are mixed in 2:1 ratio, thus the areas under the peaks have 2:1 ratio, which enables the association of each peak with the specific gene.

Table 1: Data generated

Category	Type
training	Compounds
provisional testing	Compounds
system testing	Compounds
training	shRNA
provisional testing	shRNA
system testing	shRNA

2.4 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

3 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 [. . . .]

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

3.1 L1000 Experimental Scheme

The L1000 assay uses Luminex bead-based fluorescent scanners to detect gene expression changes resulting from treating cultured human cells with chemical or genetic perturbations [Subramanian 2017]. In a single experiment, CMap makes 488 measurements, each made by a different colored bead. Each measurement produces an intensity histogram (a list of integers), which characterizes the expression of two distinct genes in the given sample (for a total of 488 x 2 = 976 genes). The genes are mixed in 2:1 ratio. Thus, the areas under the peaks have 2:1 ratio, which enables association of each peak with a specific gene. In the ideal case, each histogram consists of two peaks (see Figure 1), each corresponding to a single gene.

After data collection, the expression values of the two genes are computationally extracted in a process called 'peak deconvolution,' described in the next section.

In UNI detection mode, each of the 978 landmark genes is measured individually on one of the 500 Luminex bead colors. In order to capture all 978 genes, two detection plates are used, each measuring 489 landmarks. The two detection plates' worth of data are then computationally combined to reconstruct the full 978-gene expression profile for each sample, but deconvolution is not required.

By contrast, in the DUO detection scheme two genes are measured using the same bead color. Each bead color produces an intensity histogram which characterizes the expression of the two distinct genes. In the ideal case, each histogram consists of two peaks each corresponding to a single gene. The genes are mixed in 2:1 ratio, thus the areas under the peaks have 2:1 ratio, which enables the association of each peak with the specific gene.

3.2 Statistical deconvolution of gene-specific expression profiles.

In each sample, 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.

3.2.1 Benchmark k-means solution

CMap's current solution to this problem is based on a k-means clustering algorithm called dpeak that works as follows:

For each measurement, the dpeak partitions the list of realizations into $k >= 2$ distinct clusters and identifies two of the clusters whose ratio of membership is as close as possible to 2:1. The algorithm then takes the median intensity of each of the two clusters, assigning these values to the appropriate gene (i.e., matching clusters with more observations to the gene mixed in higher proportion).

After deconvoluting each sample on a plate, dpeak then uses the plate-wide distributions to perform adjustments on a per-well basis, correcting peaks that may have been misassigned (see Appendix).

Known problems with the current approach are that k-means is generally a biased and inconsistent estimator of the peaks of a bimodal distribution [ref]. It also sometimes fails to detect peaks with few observations or it incorrectly identifies these peaks as extraneous and disregards them. Another limitation is that it is computationally expensive (current Matlab implementation takes about 30 minutes on a 12-core server to process one set of 384 experiments).

3.3 Data generation for contest

To generate data for this contest, we profiled six 384-well perturbagen plates, each containing mutually exclusive sets of compound and shRNA treatments. Multiple treatment types were used to ensure avoid potentially over-fitting to any one. The compound and shRNA perturbagen plates were arbitrarily grouped into pairs, and to avoid any potential 'information leakage' each pair was profiled in a different cell line. The resulting lysates were amplified by ligation mediated amplification (LMA, Subramanian 2017). The amplicon was then split and detected in both one-gene-per-bead (UNI) and two-genes-per-bead (DUO) detection modes. The three pairs of data were arbitrarily assigned to training, testing, and holdout categories, where in each case the UNI data served as the ground truth.

The data so generated were then split into three subsets called: training, provisional-testing, and system-testing. Training and provisional-testing data were made available for all the contestants to develop and validate their solutions, while system-testing data were secured to evaluate competitors' last submissions, which was used to award the prizes, and to avoid overfitting as well.

Table 2: Data generated

Category	Type
training	Compounds
provisional testing	Compounds
system testing	Compounds
training	shRNA
provisional testing	shRNA
system testing	shRNA

3.4 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

4 Results

4.0.0.1 Participation

The contest attracted xxx participants, who made xxx code submissions (a median of xxx per person).

Fig. 1. Participation stats (Submission counts)

4.0.0.2 Overall accuracy and speed.

Fig. 2. (A) Leaderboard (all scores) (B) Disaggregated scores for top 10 (barplot for mean of the 2 plates by submission and for each metric) (C) Scatter plot runtime vs accuracy (mean of AUC and Correlation)

Table 1. Summary contestant solutions (top 5 methods)

Explain accuracy as measured in the coontest (slide p. 124). And then explain, KD additional test of accuracy (slides p. 128). Results are good on both.

(How far froom the max achievable improvement in accuracy (down-sampling uni)?)

Discrepancy between genes with high/low bead counts.

4.0.0.2.1 Clustering Submissions.

Do methods overlap? Not at a level that we care about.

Figure 3. (A) Clustering by genes (high ovverlap); (B) TS1-2 Seem to be clustering by method (C) Differences mitigated after standard normalization procedure

4.0.0.3 Ensambles.

Figure 3. (A) Scatterplot runtime vs accuracy for ensamble (slides p. 163)

Speed vs accuarcy trade-off. Integration one or multiple methods?

4.0.0.4 Minors:

- signs of ovverfitting (compare traing vs testing)

5 Discussion

Summary of the results presented in the methods section.

Discussion generality of the solutions

- Novel? Have any of these solutions previously been applied to deconvolution problems?
- Specific to this problem or general to others?

Discuss implications of these methods for CMap production

- Preliminary results on past data conversion
- Directions for pipeline integration and generation of future data
- Cost savings
- Implementation strategy and outcomes
- Increase in data processing throughput

6 Display items

7 References

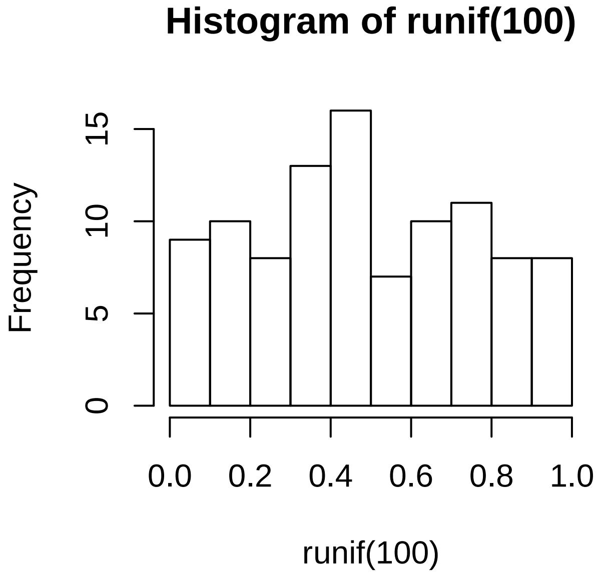

Figure 1: Random picture