Inferring Perturbation Profiles of Cancer Samples

Authors of the Paper:
Martin Pirkl and Niko Beerenwinkel
Department of Biosystems Science and Engineering, ETH Zurich, Basel 4058, Switzerland
Swiss Institute of Bioinformatics, Basel 4058, Switzerland

Poster Made by:
ZIQIziqiZ

10.18129/B9.bioc.nempi

Abstract

- **Goal:** Infer perturbation profiles of cancer driver genes.
- **Model:** NEMπ - a nested effects model extension, which utilizes the causal network between genetic aberrations and gene expression data.
- **Results:** The proposed model performed much better on a simulated dataset and a CRISPR scRNA-seq dataset compared to other classification models.

Introduction

Task Introduction:

- Replication project of the paper - *Inferring perturbation profiles of cancer samples.*
- Given gene expression data, samples are labeled by their perturbations.
- Infer perturbation profiles - which gene is perturbed in which sample.

Importance of the Problem:

- Overcome existing challenges: difficulty in observing gene perturbations directly.
- Overcome existing challenges: difficulty in measuring molecular profiles (indirect indication of perturbation).
- Identify cancer driver genes.
- Distinguish different types of cancer.
- Determine treatments accordingly.

Summary of the Project:

- Replicated validating NEMπ on a simulated dataset - recovered unobserved perturbation profiles.
- Replicated validating NEMπ on a CRISPR scRNA-seq dataset.
- Replicated running a support vector machine model, a neural net model, and a random forest model on both datasets for comparison.
- Extension: explored adding a bootstrap function and the reasons for the variability of NEMπ's results.

Datasets

- Simulated Dataset

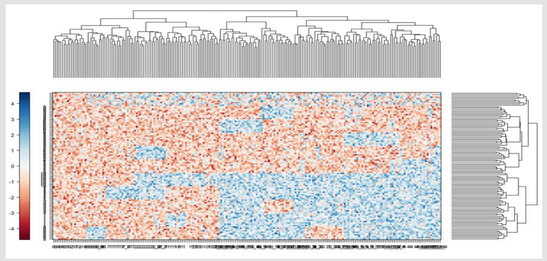


Figure 1: Heatmap of the data expression matrix. Blue corresponds to E-genes, and red indicates the genes with no effect. The shades of colors suggest the values of log odds.

- CRISPR scRNA-seq Dataset

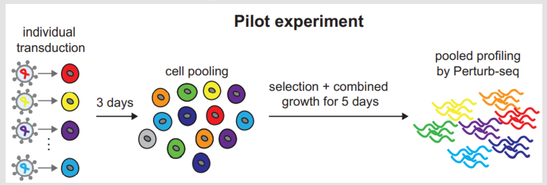
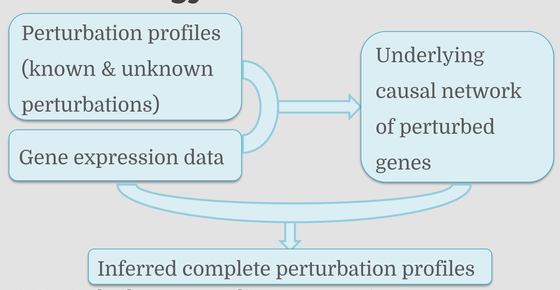


Figure 2: Pilot experiment schematic. (Source: Adamson, B. et al. (2016) A multiplexed single-cell crispr screening platform enables systematic dissection of the unfolded protein response. Cell, 167, 1867–1882.e21.)

Methodology



- Perturbed genes -> adjacency matrix
- Effect genes -> binary attachment matrix
- Adjacency & attachment -> gene expression profiles
- Compare gene expression profiles to real data
- Expectation maximization algorithm -> iteration

Results

Simulation Study - Infer Unobserved Perturbation Profiles:

- Simulate single-cell data from a random mixture of networks.
- Parameters: P-genes, E-genes, the number of samples, uninformative samples, multiple perturbations.
- Standard deviation of Gaussian noise (1 or 3)
- Percentage of removed perturbation profiles (10% or 50%)

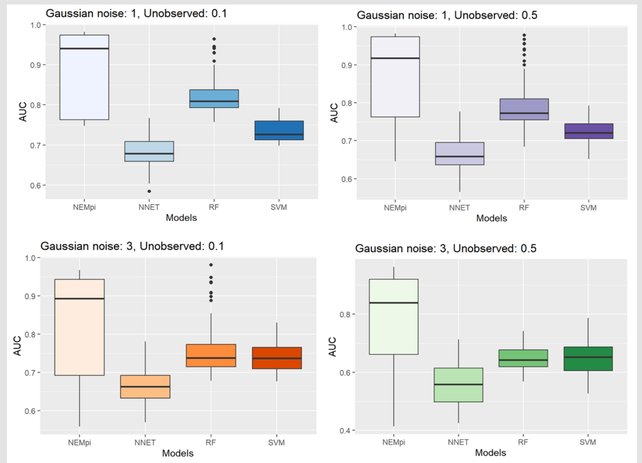


Figure 3: AUC of NEMπ, support vector machine, neural network, and random forest when inferring unobserved perturbation profiles with different Gaussian noises and different percentages of unobserved perturbation profiles.

CRISPR scRNA-seq Data Study:

- Matrix with genes as the names of rows and cell barcodes as the names of columns.
- Preprocess the sparse matrix, compute the log odds of gene expression data.
- Compare the actual perturbation profiles with the derived perturbation matrix.

Table 1: Classification performance of NEMπ, support vector machine, neural network, and random forest on CRISPR scRNA-seq data

NEMπ	SVM	NNET	RF
0.87	0.71	0.77	0.78

Extension

NEMπ + Bootstrap Function:

- Resample data from original datasets.
- Grid search: choose the optimal number of bootstraps.
- Generate more stable results.

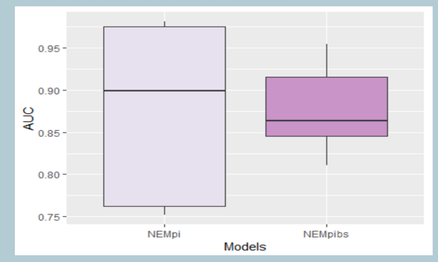


Figure 4: AUC of NEMπ and NEMπ bootstrap function with 8 bootstraps, Gaussian noise of 1, and 0.1 unobserved perturbation profiles.

Discussion

- NEMπ performs better than other supervised classification models, such as the neural network, the random forest, and the support vector machine.
- NEMπ is able to infer perturbation profiles accurately even if the dataset has high noise and a high percentage of unobserved perturbations.
- NEMπ performs well on simulated datasets and real-life CRISPR scRNA-seq datasets.
- Applying a bootstrap function to NEMπ helps generate more stable and accurate results.

Summary

- NEMπ infers perturbation profiles accurately.
- Importance of causal network of perturbed genes.
- Limitation: interpretation of the causal network.