Introducing a Novel Simulation Tool for Interconnected Differential Expression Signatures and Its Application to Benchmarking

Catalina GONZALEZ GOMEZ[1,2], Manuel ROSA-CALATRAVA[2], Julien FOURET[1]
[1]Signia Therapeutics, 60 Avenue Rockefeller, 69008, Lyon, France.
[2]Laboratoire de Virologie et Pathologies humaines, 7-11 rue Guillaume Paradin, 69008, Lyon, France.
Corresponding authors: catalina.gonzalez-gomez@univ-lyon1.fr julien.fouret@signiatherapeutics.com

ABSTRACT

Pharmaceutical research has long used **differential gene expression signatures** to study external stimuli like pathogenic determinants or small molecule treatments. These signatures measure expression values for multiple tags and are often compared using the **concept of connectivity** [1, 2]. Despite the scientific community's efforts [3, 4 ,5, 6, 7] to produce unbiased datasets for evaluating connectivity-based methods for drug identification and repurposing, the **limits of benchmarking data** hinder their effectiveness.

To address this, we developed **a simulation method to generate pairs of connected differential expression signatures**, that is based on a **three layers decomposition** and relies on a **statistical framework with different levels of parametrization**. We benchmarked **seven connectivity scores methods from the literature** [8] using our simulated signatures. We then evaluated the capacity of each method to retrieve the most connected signatures for a specific query, using the **area under the precision-recall curves (AUPRC)** [9, 10]. Moreover, we introduced a novel application perspective by training a Siamese Neural Network with our simulated data to predict the connectivity score.

Overall, our method is a significant advance in pharmaceutical research, **providing a reliable way to simulate connected differential expression signatures**. It will help develop and **evaluate algorithms for comparing signatures** to find the most connected or reversed, leading to more effective drug repurposing. An open-source version of the package will be released at the end of 2023.

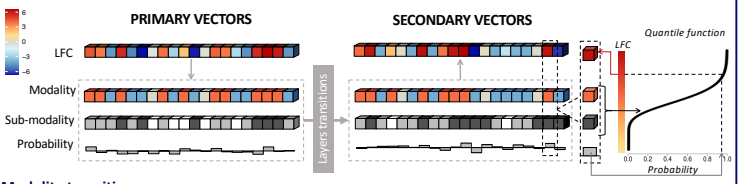
SIMULATION ALGORITHM AND DATA PROPERTIES

Motivation : limits of benchmarking data (labeled pairs based on therapeutic class, efficacity on a specific pathology, …) for evaluating connectivity-based methods for drug identification and repurposing.

Goal : generate differential expression signatures that are associated with a predefined connectivity score.

Differential expression signatures are represented by **log2 fold-change vectors** (LFC). In order to model the LFC distribution while linking the vectors with a known connectivity score (CS), we have implemented a three layers decomposition based on the following features :

1. **Modality** $\in \{UP, NR^*, DOWN\}$**,** represents the deregulation status of a gene
2. **Sub-modality** $\in [1; N]$**,** represents a rank that maps to a quantile function that models the LFC values for genes within that sub-modality. This mapping is specific to each modality.
3. **Probability** $\in [0; 1]$, is the probability that the distribution function will be equal to the final LFC value.



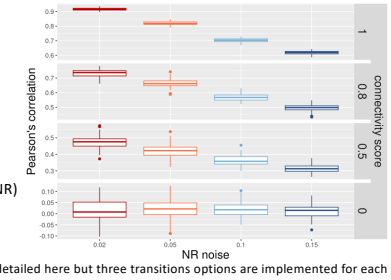
Modality transition

Transition matrix

	UP	NR	DOWN
UP	$\frac{1}{2}(1+c)\left(1-\frac{\gamma}{2}\right)$	$\frac{\gamma}{2}$	$\frac{1}{2}(1-c)\left(1-\frac{\gamma}{2}\right)$
NR	$\frac{\gamma}{2}$	$1-\gamma$	$\frac{\gamma}{2}$
DOWN	$\frac{1}{2}(1-c)\left(1-\frac{\gamma}{2}\right)$	$\frac{\gamma}{2}$	$\frac{1}{2}(1+c)\left(1-\frac{\gamma}{2}\right)$

Notations:
- **c: connectivity score**
- **γ: noise** introduced by the non-deregulated* (NR) genes

Properties



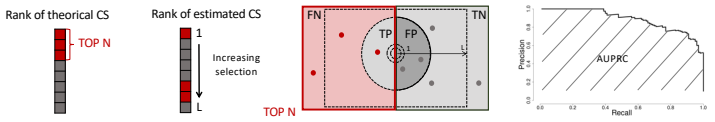
Note: sub-modality and probability transitions are not detailed here but three transitions options are implemented for each layer.

APPLICATION IN BENCHMARKING

Datasets

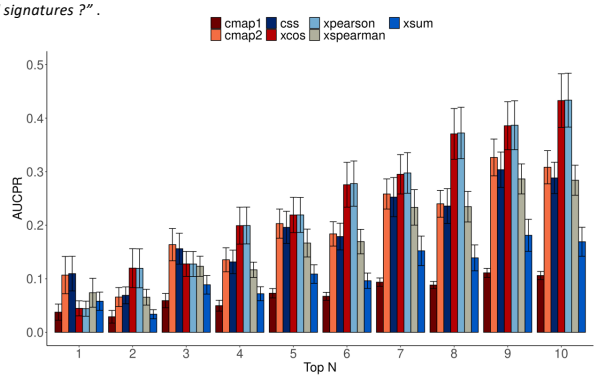
Replicated sets of **1000 secondary signatures (profiles DB)** were generated from a **simulated primary signature (query),** with each **query-profile pair (L=1000)** having a unique CS. **To account for inference bias**, we simulated experimental replicates (n=3) per signature and subsequently a read-counts matrix and employed **DESeq2 [11] to re-estimate the final differential expression signature**.

These **re-inferred signatures** are then used to estimate the CS using seven different methods, those estimations are then compared to the theorical CS used during the simulation stage.



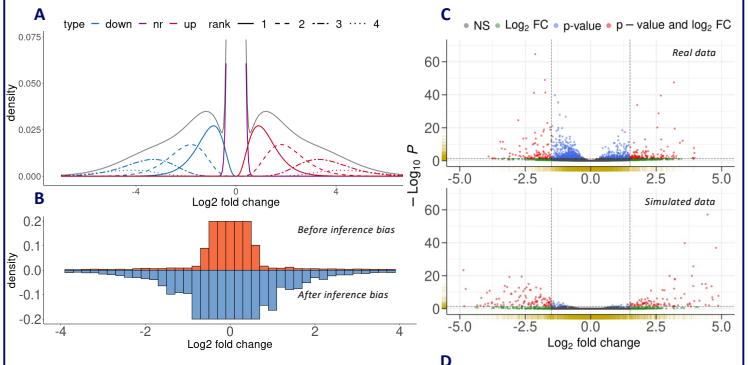
Evaluation metric

The evaluation metric choose for this benchmarking was the AUPRC. The question we wanted to answer in this specific benchmarking was *"How well does the standard CS estimators could select the top N of most connected signatures ?"* .



Bars represent the mean value over the 20 replicates and error bars represent the confidence interval computed at 95% level.

SIMULATED VS REAL DATA



A. Multimodal distribution used to model the primary LFC values.

B. Impact of Inference Bias on Primary LFC Distribution

C. Comparison of volcano plots: real data VS simulated primary distribution (post-inference bias).

D. Dependency of estimated Pearson's correlation on the percentage of DEG, for simulated and real data with a known connectivity of 1 (experimental replicates).

PERSPECTIVES

Use of Siamese Neural Networks

The availability of training data for neural networks in the prediction of a phenotypic effect (such as therapeutic effect) is limited and biased depending on the chosen estimation method. Simulated data provides a promising solution for training deep neural networks before using real data for fine-tuning

Integration of systems biology : co-expression networks / functional networks

The integration of systems biology information during the simulation process as well as in the development of novel methods estimating the CS holds immense potential for advancing drug repurposing research (introduce different connectivity scores by functional categories; convolutional graphs; …).

CONCLUSION

- Our innovative simulation tool enables the exploration of diverse patterns, generating datasets that encompass a **wide range of properties and traits**. These datasets serve **as robust challenges for methods estimating connectivity scores**.
- **Extreme Pearson's correlation and Cosine similarity**-based estimators, showcase notable performance in our benchmarking. Nevertheless, these results are influenced by the properties of the simulated data.
- Current results highlight the **potential of simulated data** on drug repurposing by overcoming limitations of real-world data. However, it is crucial to integrate biological functions during the simulation process to ensure the generation of more realistic and **biologically relevant datasets**.

ACKNOWLEDGMENTS

The authors gratefully acknowledge all the members of the Virpath and Signia Therapeutics who have made valuable contributions to the development of this project. Special thanks to the Bioinformatics team and their intern for their dedicated efforts and insightful contributions. Authors also acknowledge the calculus infrastructure and support offered by the PSNM (ENS).

Financial support for this project was provided by the French National Research and Technology Agency (ANRT) through the Cifre program.

1. Lamb, J. et al. (2006), Science, 313
2. Subramanian, A. et al. (2017) Cell, 171(6)
3. Cheng, J. et al. (2014) Genome Medicine, 6(12)
4. Yang, C. et al. (2022) eLife, 11,
5. Lin, K. et al. (2020 Briefings in Bioinformatics, 21(6)
6. Cheng, J. et al. (2012) Biocomputing 2013
7. Cheng, J. and Yang, L. (2013) 2013 IEEE
8. Samart, K. et al. (2021) Briefings in Bioinformatics, 22(6)
9. Davis, J. and Goadrich, M. (2006) International conference on Machine learning
10. Wouters, et al. (2020) JAMA, 323(9)
11. Love, et al. (2014) Genome Biol 15