Gene-level Inference of Regulatory effects As Factorizations of Functions of Expressions (GIRAFFE)

Soel Micheletti[1, 2], Alexander Marx[2], Julia Vogt[2], John Quackenbush[1,3, 4], Jonas Fischer[1], Panagiotis Mandros[1]

[1] Harvard T.H. Chan School of Public Health (Boston, MA), [2] ETH Zurich (Zurich, Switzerland), [3] Dana-Farber Cancer Institute (Boston, MA), [4] Channing Division of Network Medicine, Brigham and Women's Hospital (Boston, MA).

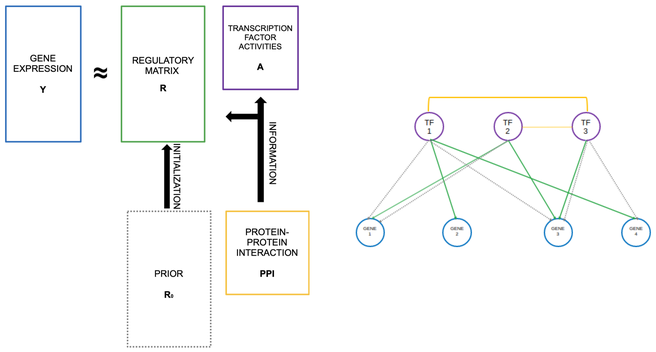
Objectives

➜ Inference of Gene regulatory Networks (GRN) is essential to:
- understand important cellular processes,
- provide insights into development and progression of disease,
- design appropriate interventions.

➜ There is a wide scope for improvements in GRN inference:
1. **Accuracy**: GRN inference on large and noisy human data is complex, and scores around the random baseline are often reported.
2. **Interpretability**: distinguishing activating from inhibiting regulation opens avenue for discovery in case-control studies.
3. **Scalability**: scaling up beyond a few hundreds genes enables proper validation on human data.

✯ We solve these by employing a **scalable matrix factorization** and **interpretable linear approach** that integrates **biological domain knowledge for better accuracy**.

Model development



➜ In high-dimensional datasets, GRN inference as matrix factorization is an **underdetermined problem**.

✯ **GIRAFFE** = **G**ene-level **I**nference of **R**egulatory effects **A**s **F**actorization of **F**unctions of **E**xpressions enforces **alignment with prior biological knowledge to overcome this issue**, and it effectively finds a biologically meaningful solution by minimizing the following objective:

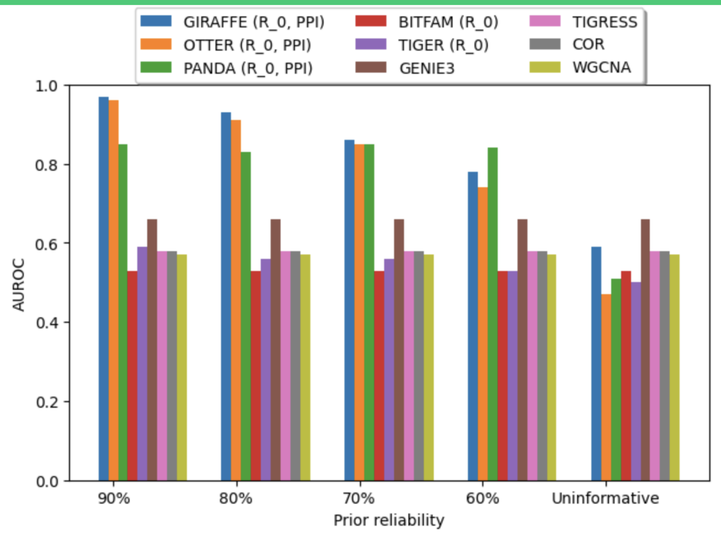
$$\begin{aligned}\underset{R,A\geq 0}{\arg\min}\quad & \alpha||Y - R\cdot A||_F^2 && \text{(Reconstruction error)}\\ & +\beta||R^T\cdot R - P||_F^2 && \text{(PPI projection)}\\ & +\gamma||R\cdot R^T - C(Y)||_F^2 && \text{(Co-expression projection)}\\ & +\delta||A\cdot A^T - P||_F^2 && \text{(TFs cooperation term)}\\ & +\lambda||R||_F^2, && \text{(Regularization)}\end{aligned}$$

where $\{\alpha, \beta, \gamma, \delta\}$ are tuned with a loss rebalancing approach that ensures that all objectives are satisfied.

Supplemental information

← **Thesis** (May, 2023)
← References & acknowledgements
Contact information →
GitHub repo →

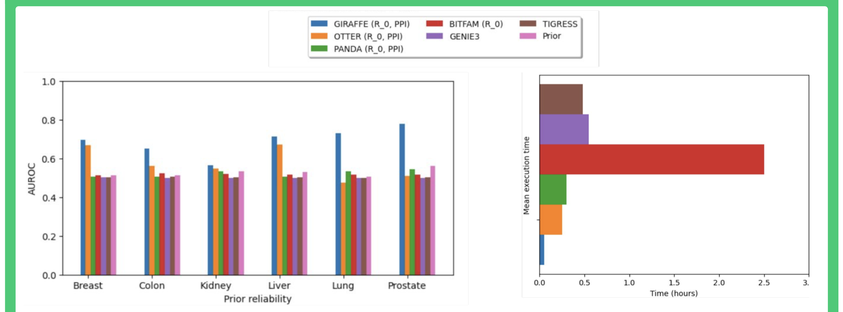
Improved regulation estimates *in-silico*



✯ For prior reliability $\geq 60\%$, methods incorporating **PPI** (GIRAFFE, PANDA, OTTER) perform significantly better.

✯ GIRAFFE **significantly outperforms** currently used methods (e.g. GENIE3, WGCNA, etc), and it is the most **robust** prior-based method.

Improved GRN inference in multiple cell lines



➜ Tested GRN inference methods using ChIP-seq data as ground-truth.

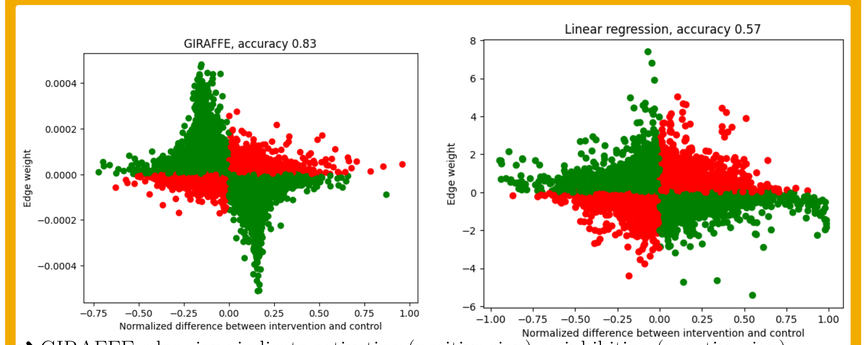
✯ GIRAFFE achieves the highest scores across all tissues, showing that it is able to better **capture the heterogeneity of cellular processes in different tissues**.

✯ GIRAFFE is at least five times faster, potential to **scale up** with the next generation of genome-wide data and providing significant advantage in GRN inference.

References

1. Chen et al.(2022). Joint inference of transcription factor activity and context-specific regulatory networks. *bioRxiv*, pages 2022–12.
2. Gao et al. (2021). A bayesian inference transcrip- tion factor activity model for the analysis of single-cell transcriptomes. *Genome Research*, 31(7):1296–1311.
3. Glass et al.(2013). Passing messages between biological networks to refine predicted interactions. *PloS one*, 8(5):e64832.
4. Weighill et al. (2021). Gene regulatory network inference as relaxed graph matching. *In Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10263–10272.
5. Haury et al. (2012). Tigress: trustful inference of gene regulation using stability selection. *BMC systems biology*, 6(1):1–17.
6. Huynh-Thu et al.(2010). Inferring regulatory networks from expression data using tree-based methods. *PloS one*, 5(9):e12776
7. Langfelder et al. (2008). Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):1–13.
8. Mandros et al. (2024). node2vec2rank: Large Scale and Stable Graph Differential Analysis via Node Embeddings and Ranking. `https://github.com/pmandros/node2vec2rank`.

Interpretable: regulation sign on yeast



➜ GIRAFFE edge signs indicate activating (positive sign) or inhibiting (negative sign) regulation. We used interventional TFKO data to validate GIRAFFE interpretation.

✯ GIRAFFE **effectively discriminates activating from inhibiting regulation**.

✯ **New opportunities** to better understand biological phenomena in case-control studies with GRNs, as distinguishing activating from inhibiting regulation yields larger differences for sign changing TF-gene interactions.

Analysis of liver hepatocellular carcinoma (LIHC)

✯ GIRAFFE scales to 25000+ genes, while remaining **predictive** and **interpretable**.

✯ GIRAFFE validates known drivers of LIHC and **facilitates** the **discovery** of biologically meaningful pathways.