Pairwise gene-gene interactions from RNAi perturbation screens: scalability and accuracy of recent machine learning tools

Kieran Elmes[1], Fabian Schmich[2], Ewa Szczurek[3], Niko Beerenwinkel[4,5], and Alex Gavryushkin[1]

[1] Biological Data Science Lab, Department of Computer Science, University of Otago, Dunedin, New Zealand, [2] Roche, Munich, Bavaria, Germany, [3] Institute of Informatics, University of Warsaw, Warsaw, Poland, [4] Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland, [5] SIB Swiss Institute of Bioinformatics, Basel, Switzerland

Introduction

Inference of genetic interactions is challenging. While it is feasible to experimentally perform most pairwise knockouts in simple organisms [1], doing so with the approximately 20,000 genes present in humans would require almost 200 million experiments. Leveraging the combinatorial nature of siRNA knockdowns, we are able to infer pairwise interactions on a large scale using existing statistical tools. We evaluated the performance of two recent tools for interaction detection, `xyz` [3] and GLINTERNET [2], on simulated siRNA screens of 100 genes. Scalability was also tested on simulated sets of up to 4000 genes.

Materials and Methods

We simulate an siRNA–gene perturbation matrix X, choose main effects and interactions, and sample a fitness vector Y. Noise is added to both X and Y to match specific signal-to-noise ratios.

$$n \text{ siRNAs} \overbrace{\begin{bmatrix} 1 & 0 & \cdots & 1 & 1 \\ 0 & & & & \\ \vdots & & \ddots & & \\ 0 & & & & 0 \end{bmatrix}}^{p \text{ genes}} \quad \underbrace{\begin{bmatrix} 2.3 \\ -3.3 \\ \vdots \\ 0.3 \end{bmatrix}}_{Y} \; n \qquad y_k \approx \beta_0 + \sum_i x_{ki}\beta_i + \sum_{i<j} x_{ki}x_{kj}\beta_{i,j}$$

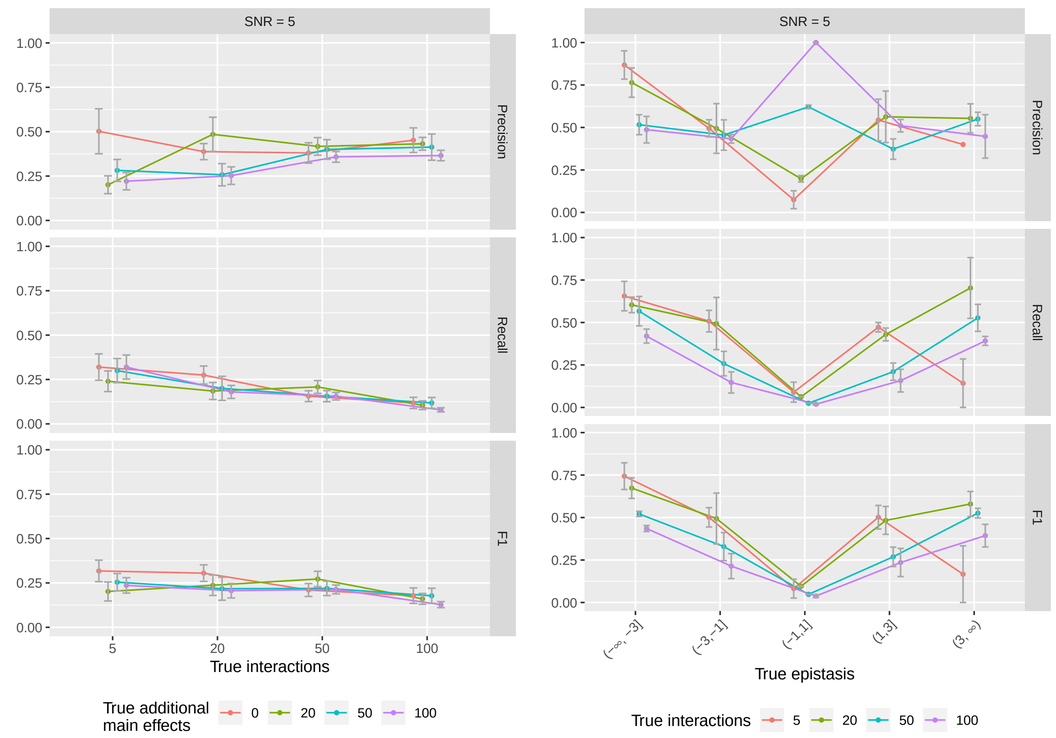
X

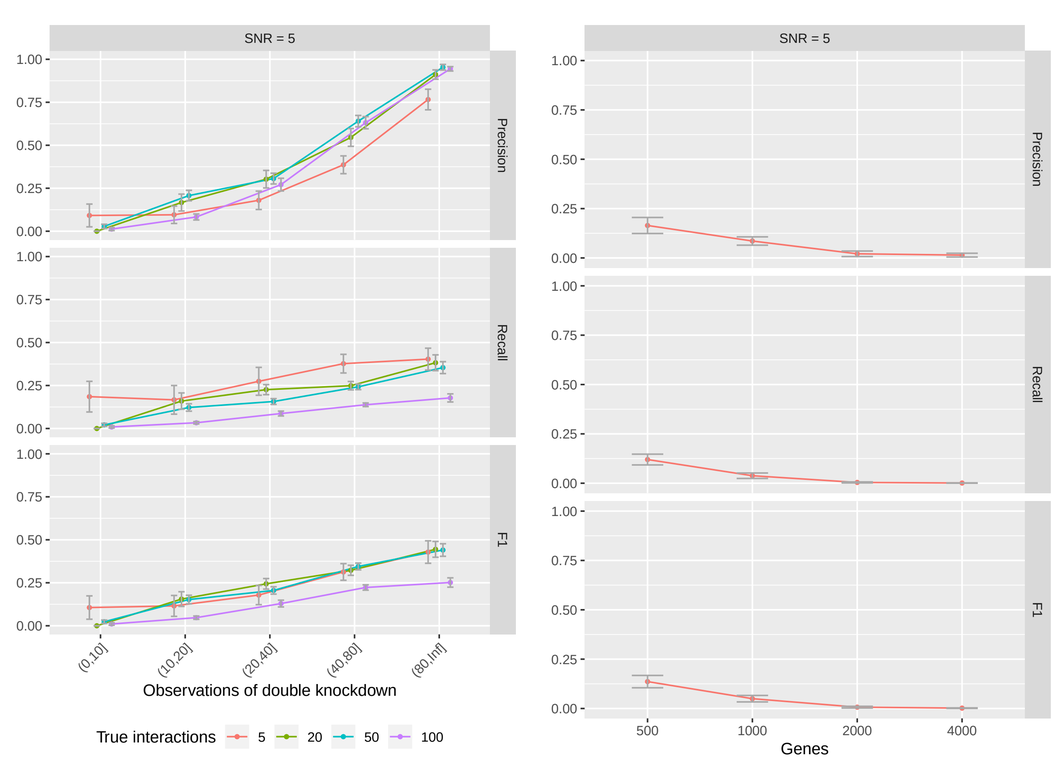
We run both `xyz` and GLINTERNET on the simulated data sets to find interaction coefficients $\beta_{i,j}$. Results are filtered in all cases with the chi-squared test, and we reject values that are not significant at the level of $\alpha = 0.05$.

Results

`xyz`

Given a small data set, `xyz` is able to identify ≈ 10-25% of the interactions, while returning 50-75% false positives. Strong interactions, and those that occur frequently, are significantly more likely to be correctly identified.

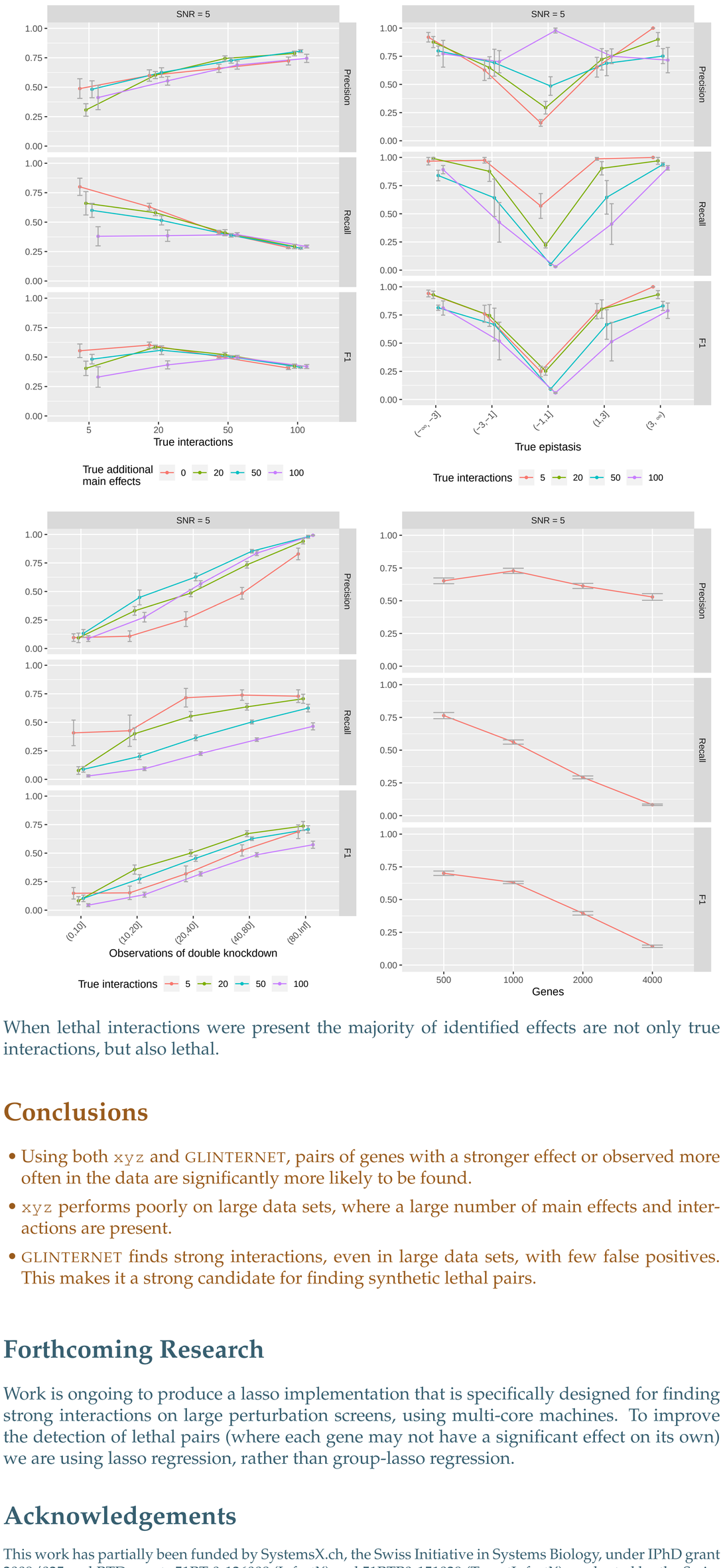




As the size of the data sets and number of effects increase, the vast majority of results become false positives. Strong negative effects are no longer found.

GLINTERNET

On the same data sets, GLINTERNET significantly outperforms `xyz`. 50-75% of the results are correctly identified interactions. Again, both strong interactions and those that occur frequently are significantly more likely to be correctly identified.



When lethal interactions were present the majority of identified effects are not only true interactions, but also lethal.

Conclusions

- Using both `xyz` and GLINTERNET, pairs of genes with a stronger effect or observed more often in the data are significantly more likely to be found.
- `xyz` performs poorly on large data sets, where a large number of main effects and interactions are present.
- GLINTERNET finds strong interactions, even in large data sets, with few false positives. This makes it a strong candidate for finding synthetic lethal pairs.

Forthcoming Research

Work is ongoing to produce a lasso implementation that is specifically designed for finding strong interactions on large perturbation screens, using multi-core machines. To improve the detection of lethal pairs (where each gene may not have a significant effect on its own) we are using lasso regression, rather than group-lasso regression.

Acknowledgements

This work has partially been funded by SystemsX.ch, the Swiss Initiative in Systems Biology, under IPhD grant 2009/025 and RTD grants 51RT-0_126008 (InfectX) and 51RTP0_151029 (TargetInfectX), evaluated by the Swiss National Science Foundation. We acknowledge support from the Royal Society of New Zealand through the Rutherford Discovery Fellowship awarded to AG.

References

[1] Michael Costanzo et al. “The genetic landscape of a cell.” In: *Science* (2010).

[2] Michael Lim and Trevor Hastie. “Learning interactions via hierarchical group-lasso regularization”. en. In: *J. Comput. Graph. Stat.* 24.3 (2015), pp. 627–654.

[3] Gian-Andrea Thanei, Nicolai Meinshausen, and Rajen D Shah. “The xyz algorithm for fast interaction search in high-dimensional data”. In: (Oct. 2016). arXiv: `1610.05108 [stat.ML]`.