Patterns

CellPress
OPEN ACCESS

Preview

Modeling in systems biology: Causal understanding before prediction?

Szilvia Barsi[1] and Bence Szalai[1,*]
[1]Department of Physiology, Faculty of Medicine, Semmelweis University, Budapest, Hungary
*Correspondence: szalai.bence@med.semmelweis-univ.hu
https://doi.org/10.1016/j.patter.2021.100280

Babur et al. (2021) developed the *CausalPath* tool to infer causal signaling interactions in high-throughput proteomics data that may foster mechanical understanding from large-scale biological datasets.

Recent advancements of high-throughput technologies allow acquisition of large-scale biological datasets of different modalities, like transcriptomics, (phospho) proteomics, or metabolomics (generally "omics" data), even on the level of single cells. While these datasets promise unique opportunities to understand molecular mechanisms behind biological phenotypes in health and disease, their correct interpretation is complicated by several factors. At first, standard analysis methods in most cases return only lengthy lists of differentially expressed or phenotype-correlated genes or proteins, which hamper the effort to gain mechanistic insight about the observed phenotype. Also, the high dimensionality of experimental data (e.g., ~20,000 in the case of transcriptomics) makes it complicated to distinguish between simple correlations and causal associations—for understanding and therapeutic interventions, the latter is essential.

The main aim of different systems biology modeling and analysis techniques is to overcome these limitations. Generally, these approaches can be classified as knowledge- or data-driven ones (Figure 1). Knowledge-driven methods use in most cases extensive, curated lists of gene sets form connected biological processes or pathways and use statistical methods (with or without explicit pathway information) to find overrepresentation/enrichment of these gene sets in biological datasets.[1] These methods tend to give more biological insights than simple lists of differentially expressed genes, thus they are more appropriate for hypothesis generation. However, in most cases the used gene sets are too general to identify real causal information from data. On the other side, data-driven methodologies, including machine learning models, focus on predictive performance. Predictive performance of systems biology models are important from several points. At first, predictive models can be important in different fields of biology from drug discovery to patient stratification. Also, one can argue, if some biological phenotype is predictable from omics data, that means that the prediction model identifies the underlying biological mechanisms. However, these later claims are unfortunately overrated: machine learning models can learn some technical biases and confounding factors of the analyzed datasets, which foster prediction performance but hamper biological understanding and generalization.[2] Also, several of the best performing machine learning models are "black-box" models, meaning it is complicated to derive the exact prediction mechanisms from them, which also prevents biological interpretation.

Recently, several new methods were developed to bridge these differences between knowledge- and data-driven methodologies.[3–5] These "causal reasoning tools" connect prior-knowledge networks (like signaling pathways or gene regulatory networks) with genome scale gene expression or proteomics measurements and use statistical tools to identify contextualized, sample-specific signaling network alterations and thus causal effects explaining the observed data. These methods have been shown to better estimate pathway activity changes than classical knowledge-driven methodologies in different benchmarks.

Babur et al. (2021)[6] added a new, interesting methodology to this later toolset. *CausalPath* uses kinase/phosphatase—substrate and transcription factor—regulated gene relationships from the Pathway Commons database to create graphical patterns. These graphical patterns are causal associations like "$Kinase_A$ is active when phosphorylated on site P_1. Active $Kinase_A$ phosphorylates $Protein_B$ on site P_2." These kinds of graphical patterns are matched with measurements like "$Kinase_A$ is phosphorylated on site P_1, and $Protein_B$ is phosphorylated on site P_2", leading to causal conjectures like "$Kinase_A$ phosphorylates $Protein_B$ in the given dataset", identifying the potential causal way of signaling. *CausalPath* also tests the statistical significance of the derived results using a data label permutation-based approach. In their paper, the authors test their methodology in different cancer related datasets, and they successfully identify mechanisms of action of different ligands and drugs from proteomics data.

The results of Babur et al. (2021)[6] also highlight the importance of using the correct type of prior knowledge with the corresponding omics modality. When they used gene regulatory networks with proteomics data, the inferred causal networks were not statistically significant, while using the same prior-knowledge network with gene-expression data resulted in significant causal associations. These results also highlight a general problem of systems biology modeling: given the higher abundance of transcriptomics datasets (compared to phosphoproteomics, for example), gene expression data are more frequently used in modeling studies. However, the used prior-knowledge networks are defined on the level of protein activities (pathways) in most cases. As the

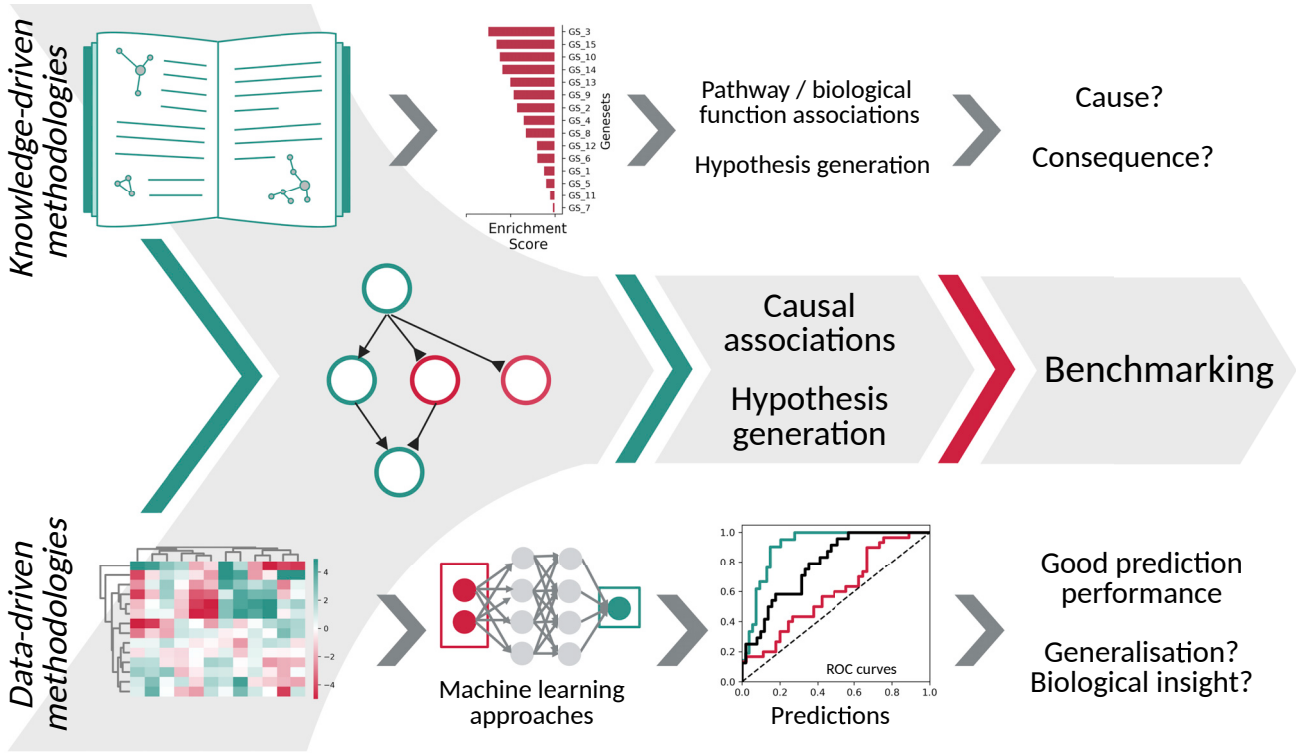

Figure 1. Schematic representation of systems biology modeling directions

Knowledge-driven methods (top) use literature-curated gene sets of functionally related genes and perform some kind of overrepresentation/enrichment analysis using them. The enriched gene sets can help to interpret associations with different biological mechanisms; however, causal interactions are hard to be identified. Data-driven methods (bottom) use statistical/machine-learning methods to predict biological phenotypes. While these methods reach good predictive performance, their generalization and ability to gain mechanistic insight is limited in several cases. Causal reasoning methods (middle) use prior-knowledge network information together with data to identify contextualized causal signaling networks. The identified causal interactions can be used for hypothesis generation; however, future benchmarking of these methods is needed. Figure was created with BioRender.com.

association between gene expression and protein abundance/activity can be modest, using gene-expression data with pathway networks can lead to incorrect interpretation of the results.[7] These considerations, and also the results of Babur et al. (2021),[6] suggest the crucial importance of using matching prior-knowledge networks and data, like gene regulatory networks with transcriptomics and signaling networks with (phospho)proteomics. Correct integration of different types of prior-knowledge networks and data types also promises to identify causal associations in multi-omics datasets.[8]

While currently the most important aspect of causal reasoning tools is biological hypothesis generation, assessing the predictive performance or causal reasoning tools is also crucial for benchmarking the different methods to select the best-performing ones. In their paper, Babur et al. (2021)[6] compared their method to several existing ones, which is a good first step toward this direction. However, as more and more related tools are developed, it is crucial to perform unbiased, independent benchmarking. A bottleneck for this benchmarking is high-quality data where causal associations are already known. For this purpose, perturbation data (where the general cause of changes is given by the used perturbation, i.e., drug, genetic manipulation etc.) looks most suitable,[9] but off-target effects of perturbations (drugs, small interfering RNA [siRNA]) can complicate method evaluation. Nevertheless, large-scale benchmarking projects, like Dialogue on Reverse Engineering Assessment and Methods (DREAM) Challenges[10] can foster the development and assessment of causal reasoning systems biology tools in the future.

ACKNOWLEDGMENTS

B.S. was supported by the Premium Postdoctoral Fellowship Program of the Hungarian Academy of Sciences (460044).

DECLARATION OF INTERESTS

B.S. receives consultant fees from Turbine Ltd.

REFERENCES

1. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550.
2. Eid, F.-E., Elmarakeby, H.A., Chan, Y.A., Fornelos, N., ElHefnawi, M., Van Allen, E.M., Heath, L.S., and Lage, K. (2021). Systematic auditing is essential to debiasing machine learning in biology. Commun Biol *4*, 183.
3. Bradley, G., and Barrett, S.J. (2017). CausalR: extracting mechanistic sense from genome scale data. Bioinformatics *33*, 3670–3672.
4. Liu, A., Trairatphisan, P., Gjerga, E., Didangelos, A., Barratt, J., and Saez-Rodriguez, J. (2019). From expression footprints to causal pathways: contextualizing large signaling networks with CARNIVAL. NPJ Syst. Biol. Appl. *5*, 40.
5. Paull, E.O., Carlin, D.E., Niepel, M., Sorger, P.K., Haussler, D., and Stuart, J.M. (2013). Discovering causal pathways linking genomic events to transcriptional states using Tied

Diffusion Through Interacting Events (TieDIE). Bioinformatics *29*, 2757–2764.

6. Babur, Ö., Luna, A., Korkut, A., Durupinar, F., Siper, M.C., Dogrusoz, U., Vaca Jacome, A.S., Peckner, R., Christianson, K.E., Jaffe, et al. (2021). Causal interactions from proteomic profiles: Molecular data meet pathway knowledge. Patterns *2*. https://doi.org/10.1016/j.patter.2021.100257.
7. Szalai, B., and Saez-Rodriguez, J. (2020). Why do pathway methods work better than they should? FEBS Lett. *594*, 4189–4200.
8. Dugourd, A., Kuppe, C., Sciacovelli, M., Gjerga, E., Gabor, A., Emdal, K.B., Vieira, V., Bekker-Jensen, D.B., Kranz, J., Bindels, E.M.J., et al. (2021). Causal integration of multi-omics data with prior knowledge to generate mechanistic hypotheses. Mol. Syst. Biol. *17*, e9730.
9. Keenan, A.B., Jenkins, S.L., Jagodnik, K.M., Koplev, S., He, E., Torre, D., Wang, Z., Dohlman, A.B., Silverstein, M.C., Lachmann, A., et al. (2018). The Library of Integrated Network-Based Cellular Signatures NIH Program: System-Level Cataloging of Human Cells Response to Perturbations. Cell Syst. *6*, 13–24.
10. Gabor, A., Tognetti, M., Driessen, A., Tanevski, J., Guo, B., Cao, W., Shen, H., Yu, T., Chung, V., Signaling, S.C., et al. (2021). Cell-to-cell and type-to-type heterogeneity of signaling networks: Insights from the crowd. bioRxiv. https://doi.org/10.1101/2021.03.23.436603.