Patterns

CellPress
OPEN ACCESS

Preview

An information theoretic approach for the inference of Boolean networks and functions from data: BoCSE

David Murrugarra[1,*] and Alan Veliz-Cuba[2]
[1]University of Kentucky, Lexington, KY, USA
[2]University of Dayton, Dayton, OH, USA
*Correspondence: murrugarra@uky.edu

https://doi.org/10.1016/j.patter.2022.100617

Building predictive models from data is an important and challenging task in many fields including biology, medicine, engineering, and economy. In this issue, Sun et al.[1] present a method for the inference of Boolean networks along with practical applications.

Boolean networks have been used successfully in modeling several dynamical processes including the bistable behavior of the *lac Operon*,[2] the interaction of pancreatic cancer cells with their microenvironment,[3] and the yeast-to-hyphal transition of the yeast *Candida albicans*.[4] Moreover, there is a growing set of Boolean network models in the Cell Collective database covering processes such as signaling, regulation, and cancer.[5] A Boolean model f with n variables is typically characterized by two objects: a directed graph or wiring diagram with n vertices that describes how variables affect each other and n Boolean functions that indicate how variables depend on each other. In contrast to other quantitative modeling approaches such as models based on ordinary differential equations, Boolean networks can be seen as qualitative models that focus on the mechanisms underlying the interactions and the nonlinear features of biological systems. The dynamics of a Boolean network is given by a graph with 2^n vertices (all binary strings with n entries) and directed edges from x to y if the Boolean network can transition from x to y (in the synchronous case this means $f(x) = y$). Two problems of interest in Boolean modeling are (1) the forward problem of dynamics prediction, that is, being able to efficiently predict the dynamics of a Boolean network from its wiring diagram or from the Boolean functions without the need of exhaustive simulation, and (2) the inverse problem of learning or reverse-engineering a Boolean network from partial information. These two problems remain largely unsolved but have been studied with widely different approaches using tools from algebraic geometry, computational algebra, information theory, etc.[1,2,6–8]

A primary challenge in building predictive models from temporal data or input-output data is selecting the appropriate network and the regulatory functions that fit the data. With the increase of available data in repositories and databases, several data-driven approaches for equation learning (EQ) have been developed. Specially, for models based on ordinary differential equations, software and mathematical theory are available and are currently being used for several applications.[9] However, the existing methods for continuous models usually require large amounts of data to provide an accurate model, which limits their applicability on datasets relevant to biological and biomedical applications where relatively few time-point measurements are available. To take full advantage of the potential of mathematical models, new data-driven approaches and software that will work well even in the case of limited data needs to be developed. For time courses with few time points, EQ for Boolean models such as the one introduced in this issue by Sun et al.[1] provides an attractive alternative.

Reverse engineering approaches can be broadly classified into two groups depending on what they find: algorithms that learn the wiring diagram[2] and algorithms that learn the Boolean rules.[6,8] In Veliz-Cuba,[2] the authors used algebraic geometry to study the inverse problem when data are already discrete and noise is negligible. They used an algebraic object (ideal of polynomials) to encode all functions that fit the data without listing the functions. This encoding at the wiring diagram level combined with a kind of factorization (primary decomposition) allowed the authors to find all wiring diagrams for which there are Boolean networks that fit the data. In Liang et al.,[7] the authors introduced REVEAL (reverse engineering algorithm for inference of genetic network architectures) that has been used for benchmarking purposes with other newer methods. In this issue, Sun et al.[1] present a method called Boolean optimal causation entropy (BoCSE). The BoCSE method is based on an optimization procedure of the mutual information between possible input and output vectors which they can solve efficiently. They have assessed the predictive ability of their method using random networks. Additionally, they applied their method to construct a binary classifier for an automated diagnosis of urinary disease using clinical data. Likewise, they applied the Boolean function inference to obtain an automated cardiac SPECT diagnosis using patient data. They also used their method to rank the slots in the game tic-tac-toe. Finally, they applied their approach to devise a classifier of risk causality of loans that result in default using publicly available data.

The limitations of the existing network inference methods are those related to each component in the inference process. For instance, most of these methods will take as inputs discrete or discretized data. Therefore, the discretization can create issues due to noise in the measurements and can lead to overfitting. Another limitation is that it is not known how much data is needed to guarantee that the predicted model is “close” to the true network unless the network is known *a priori*. Finally, we highlight the need of a user-friendly software that can take real data and output a model that can used to make predictions. Although some tools exist that address the inference process,[8,1,7] the code is either unavailable, not editable, or not in a ready-to-use format.

ACKNOWLEDGMENTS

D.M. was partially supported by a Collaboration grant (850896) from the Simons Foundation. A.V.-C. was partially supported by the Simons Foundation grant 516088.

DECLARATION OF INTERESTS

The authors have declared that no competing interests exist.

REFERENCES

1. Sun, J., AlMomani, A.A., and Bolt, E. (2022). Data-driven learning of Boolean networks and functions by optimal causation entropy Principle (BoCSE). Patterns *3*, 100631.
2. Veliz-Cuba, A. (2012). An algebraic approach to reverse engineering finite dynamical systems arising from biology. SIAM J. Appl. Dyn. Syst. *11*, 31–48. https://doi.org/10.1137/110828794.
3. Plaugher, D., Aguilar, B., and Murrugarra, D. (2022). Uncovering potential interventions for pancreatic cancer patients via mathematical modeling. J. Theor. Biol. *548*, 111197. https://doi.org/10.1016/j.jtbi.2022.111197.
4. Wooten, D.J., Zanudo, J.G.T., Murrugarra, D., Perry, A.M., Dongari-Bagtzoglou, A., Laubenbacher, R., Nobile, C.J., and Albert, R. (2021). Mathematical modeling of the Candida albicans yeast to hyphal transition reveals novel control strategies. PLoS Comput. Biol. *17*, e1008690. https://doi.org/10.1371/journal.pcbi.1008690.
5. Helikar, T., Kowal, B., McClenathan, S., Bruckner, M., Rowley, T., Madrahimov, A., Wicks, B., Shrestha, M., Limbu, K., and Rogers, J.A.; The Cell Collective: Toward an open and collaborative approach to systems biology (2012). BMC Syst. Biol. *6*, 1–14.
6. Laubenbacher, R., and Stigler, B. (2004). A computational algebra approach to the reverse engineering of gene regulatory networks. J. Theor. Biol. *229*, 523–537. https://doi.org/10.1016/j.jtbi.2004.04.037.
7. Liang, S., Fuhrman, S., and Somogyi, R. (1998). Reveal, a general reverse engineering algorithm for inference of genetic network architectures. Biocomputing *3*.
8. Dimitrova, E., Garcia-Puente, L.D., Hinkelmann, F., Jarrah, A.S., Laubenbacher, R., Stigler, B., Stillman, M., and Vera-Licona, P. (2011). Parameter estimation for Boolean models of biological networks. Theor. Comput. Sci. *412*, 2816–2826. https://doi.org/10.1016/j.tcs.2010.04.034.
9. Brunton, S.L., Proctor, J.L., and Kutz, J.N. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. Proc. Natl. Acad. Sci. USA *113*, 3932–3937. https://doi.org/10.1073/pnas.1517384113.