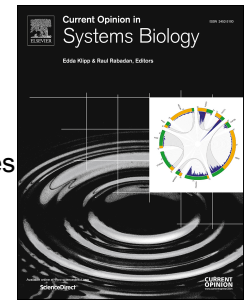
Accepted Manuscript



Learning with unknowns: analyzing biological data in the presence of hidden variables

Claudia Battistin, Benjamin Dunn, Yasser Roudi

PII: S2452-3100(16)30030-0

DOI: 10.1016/j.coisb.2016.12.010

Reference: COISB 10

To appear in: *Current Opinion in Systems Biology*

Please cite this article as: Battistin C, Dunn B, Roudi Y, Learning with unknowns: analyzing biological data in the presence of hidden variables, *Current Opinion in Systems Biology* (2017), doi: 10.1016/j.coisb.2016.12.010.

Learning with unknowns: analyzing biological data in the presence of hidden variables

Claudia Battistin

Kavli Institute for Systems Neuroscience and Centre for Neural Computation, NTNU, Trondheim

Benjamin Dunn

Kavli Institute for Systems Neuroscience and Centre for Neural Computation, NTNU, Trondheim

Yasser Roudi

Kavli Institute for Systems Neuroscience and Centre for Neural Computation, NTNU, Trondheim

Institute for Advanced Studies, Princeton

Abstract

Despite our improved ability to probe biological systems at a higher spatiotemporal resolution, the high dimensionality of the biological systems often prevents sufficient sampling of the state space. Even with large scale datasets, such as gene microarrays or multi-neuronal recording techniques, the variables we are recording from are typically only a small subset, if wisely chosen, representing the most relevant degrees of freedom. The remaining variables, or the so called *hidden variables*, are most likely coupled to the observed ones, and affect their statistics and consequently our inference about the function of the system and the way it performs this function. Two important questions then arise in this context: *which variables should we choose to observe and collect data from? and how much can we learn from data in the presence of hidden variables?* In this paper we suggest that recent algorithmic developments rooting in the statistical physics of complex systems constitute a promising set of tools to extract relevant features from high-throughput data and a fruitful avenue of research for coming years.