BAYESIAN DATA INTEGRATION: A FUNCTIONAL PERSPECTIVE

Curtis Huttenhower and Olga G. Troyanskaya
Department of Computer Science, Lewis-Sigler Institute for Integrative Genomics,
Princeton University Princeton, NJ 08544, USA *

Accurate prediction of protein function and interactions from diverse genomic data is a key problem in systems biology. Heterogeneous data integration remains a challenge, particularly due to noisy data sources, diversity of coverage, and functional biases. It is thus important to understand the behavior and robustness of data integration methods in the context of various biological functions. We focus on the ability of Bayesian networks to predict functional relationships between proteins under a variety of conditions. This study considers the effect of network structure and compares expert estimated conditional probabilities with those learned using a generative method (expectation maximization) and a discriminative method (extended logistic regression). We consider the contributions of individual data sources and interpret these results both globally and in the context of specific biological processes. We find that it is critical to consider variation across biological functions; even when global performance is strong, some categories are consistently predicted well, and others are difficult to analyze. All learned models outperform the equivalent expert estimated models, although this effect diminishes as the amount of available data decreases. These learning techniques are not specific to Bayesian networks, and thus our conclusions should generalize to other methods for data integration. Overall, Bayesian learning provides a consistent benefit in data integration, but its performance and the impact of heterogeneous data sources must be interpreted from the perspective of individual functional categories.