Multifunctionality drives gene characterization: A re-evaluation of hubs and promiscuity in gene function prediction

Jesse Gillis and Paul Pavlidis

Centre for High-Throughput Biology and Department of Psychiatry, University of British Columbia, Vancouver BC, 177 Michael Smith Laboratories 2185 East Mall, University of British Columbia, Vancouver, BC, V6T1Z4.

e-mail: jesse.gillis@gmail.com

Many previous studies have shown that by using variants of “guilt-by-association”, gene function predictions can be made with very high statistical confidence. In these studies, it is assumed that the “associations” in the data (e.g., protein interaction partners) of a gene are necessary in establishing “guilt”. In this work, we show that multifunctionality, rather than association, is a primary driver of gene function prediction. We first show that knowledge of the degree of multifunctionality alone can produce astonishingly strong performance when used as a predictor of gene function. We then demonstrate how multifunctionality is encoded in gene interaction data (such as protein interactions and coexpression networks) and how this can feed forward into gene function prediction algorithms. We find that high-quality gene function predictions can be made using data that possesses no information on which gene interacts with which. By examining a wide range of networks from mouse, human and yeast, as well as multiple prediction methods and evaluation metrics, we provide evidence that this problem is pervasive and does not reflect the failings of any particular algorithm or data type. We propose computational controls that can be used to provide more meaningful control when estimating gene function prediction performance. We suggest that this source of bias due to multifunctionality is important to control for, with widespread implications for the interpretation of genomics studies.