PI(s) Grant: DARPA ASDF

Institution(s): Harvard Medical School

Grant Number: ARO contract W911NF018-1-0124

Title of Grant: An Automated Scientific Discovery Framework

Abstract Authors:

John A. Bachman, Benjamin M. Gyori, Klas Karis, and Peter K. Sorger

Abstract Text:

Biology is currently grappling with the challenge of integrating large datasets with a body of scientific knowledge that has grown too large and complex for any single scientist to read or understand. The size and scope of large-scale data compendia ("big data") has created analytical challenges that require new solutions, ideally ones that make use of aggregated scientific knowledge.

To address this problem we have developed the Integrated Network and Dynamical Reasoning Assembler (INDRA), a system that automatically assembles mechanistic models from pathway databases, literature, and expert knowledge expressed in natural language. INDRA draws on three existing natural language processing systems and uses a modular architecture to build different types of models from a variety of sources. Mechanisms are extracted from each source format and converted into *Statements,* a normalized representation of biological mechanisms. To identify redundancies and overlaps, Statements are sorted into a hierarchy graph that identifies which are generic (e.g., "MEK phosphorylates ERK") and which are more specific (e.g., "MEK1 phosphorylated at serine 218 and S222 phosphorylates ERK1 at threonine 185"). The reliability of each mechanism is scored probabilistically based on the sources and frequency of extractions. Manual evaluation of the system on a corpus of papers indicated that this assembly process can reliably extract previously uncurated protein-protein interactions and post-translational modifications, and eliminates many of the errors that arise in automated model construction. A key feature of this approach is that the assembled models are not only broad in scope but also mechanistic, capturing information about sites of post-translational modification and necessary molecular context.

To evaluate the ability of INDRA to systematically generate explanations of high-throughput data, we assembled a rule-based executable model to explain a previously published dataset of the phospho-proteomic response of a melanoma cell line to 12 different drugs. Static analysis of the model allowed us to identify possible mechanistic paths linking drug targets to experimentally observed effects on phospho-protein abundances. The model generated biochemically plausible explanations for 20 of the 22 largest effects in the data (91%). In a second study, we are evaluating the effectiveness of our text-mined network of ~5 million unique interactions to identify mechanisms mediating correlations between gene knockouts in the Broad Cancer Dependency Map dataset. Starting with 48,000 correlations involving 340 cancer-related genes (Pearson correlation > 0.3), we found that 75% of the correlations can be explained with a single intermediate linking gene, either as a common downstream target or part of a pathway. Taken together, this work shows the potential of automatically assembled models to systematically explain high-throughput data, generating mechanistic hypotheses and identifying unexplained phenomena.