Making Sense of Genome-scale Experimental Data

A Statement for the 2012 NSF Discovery Informatics Workshop

Lawrence Hunter, University of Colorado School of Medicine

Post-genomic molecular biology is faced with a critical challenge in *making sense of the results of genome-scale experiments* that implicate dozens or hundreds of genes and gene products in the system under study. While an experimenter is likely to be intimately familiar with many of the previously known or hypothesized genes* relevant to the studied phenomena, the point of doing genome-scale science is to identify genes that were *not* previously believed to play a role. Frequently, many of those implicated genes have been previously studied, but in areas of biology remote from the context of the motivating experiment. Unfortunately, *it remains difficult to use existing resources to build an accurate, detailed and comprehensive story of the roles, interactions and activities of the large groups of genes* that result from genome-scale studies (Mirel 2009). This is not for lack of trying. Various resources have been created to support such sense-making activities–for example, summary databases such as Entrez Gene (Maglott, Ostell et al. 2011) provide useful information about each gene; enrichment analysis tools such as GOStat (Beissbarth and Speed 2004), DAVID (Huang da, Sherman et al. 2009) or Martini (Soldatos, O'Donoghue et al. 2010) attempt to characterize the functional classes that are overrepresented in a large group of genes; and visualization systems map gene lists to various kinds of networks (e.g. Cytoscape (Shannon, Markiel et al. 2003) or the popular commercial system, Ingenuity Pathway Assistant).

Molecular biology research faces a profound challenge. Experimental methodologies taking advantage of dramatic reductions in the cost of nucleic-acid sequencing, high-density microarrays, multiplex mass spectrometry, high-throughput protein-protein interaction screens, and related technologies have unleashed a flood of valuable data sets at genomic scale. The relevant scientific literature is also expanding rapidly, and there are many millions of peer-reviewed publications potentially important in the interpretation of these high-throughput results. Furthermore, curated databases containing information relevant to analyses such as genes and gene products, genomic locations, sequences and their variants, biological processes, phenotypes and diseases, intermolecular interactions, and pathways have grown dramatically. *All of these data and knowledge must be semantically integrated and methods to effectively reason over it developed if researchers are to have any hope of making sense of this ever-increasing flood of information.*

Valuable as they are, manually maintained databases are often incomplete, and likely to always lag information reported in the literature. Existing natural language processing methods can identify information about genes and some other phenomena directly from the literature. However, *existing approaches to information retrieval and information extraction do not yet meet the needs of researchers trying to make sense of genomic data.*

Interpreting a genome-scale data set is an extended, multifaceted cognitive process, which ultimately involves the creation of a coherent, useful, and well-supported set of claims regarding why a particular set of experimental results arose. Although the interpretation of each particular genome-scale dataset is unique, there are many tasks and operations that appear repeatedly, for example, grouping genes or gene products (say, by functional class or involvement in a physiologic process), identifying connections

* Broadly construed to include genes, gene products, genetic polymorphisms, etc.

among different groups (such as signaling or regulatory interactions), and succinctly summarizing hypothesized mechanisms in figures or diagrams. These recurring tasks and relationships among them can be mapped to more abstract and generic descriptions from the vocabulary of computer science, which in turn facilitates the design of a visual encoding and set of interactions that provide meaningful support for this interpretive work. As both the datasets and the relevant background knowledge are too large to be comprehended as a whole, a central activity is to filter and highlight, tasks for which existing and future knowledge base, reasoning and NLP techniques could be particularly useful.

There is a significant national need to create innovative methods and systems that will dramatically improve the ability of a broad array of researchers to interpret the results of genome-scale experiments. From a computational perspective, their sense-making task requires research in—and integration of—three previously disparate sorts of computational methods: (1) knowledge representation and reasoning, (2) natural language processing, and (3) information visualization. New research into computational reasoning is required to create formal methods that can assemble many potentially conflicting sources of knowledge into a unified view, assess which aspects of this knowledge are most likely to be of interest to a researcher analyzing a dataset, and link those aspects together into potential explanations of the phenomena observed in the experiment. The goal of research into computational reading in this context should be to create gene-centric representations of the current state of gene-centric biological knowledge from the scientific literature, captured in a formally well-defined knowledge-base. The goal of research into information visualization in this context should be to create methods that support interactive, knowledge-driven analysis of large datasets with novel visualizations and innovative approaches to navigating the scientific literature, respecting the constantly changing nature of scientific understanding. The research in these individual areas needs to be linked in order to provide an over-arching framework and relevant analyses of scientific users that can drive the design of software that effectively supports the cognitive activities of researchers interpreting genome-scale experimental results.