Nigam Shah, MBBS, Ph.D.
Assistant Professor of Medicine
Stanford University

Contributing Statement for Discovery Informatics Workshop

February 2-3, 2011

With the advent of high-throughput technologies, there is an abundance of individual data types such as gene and protein sequences, gene expression data, protein structures, protein interactions and annotations. At the same time, there is a shortage of tools and methods that can handle this information and allow a scientist to draw meaningful inferences. A significant amount of time and energy is spent in merely locating and retrieving information, rather than thinking about what that information *means*. On the clinical side, the emphasis on comparative-effective research has renewed interest in capturing patient data in machine-interpretable form. As research gets more data intensive, the need for tools for thought gets more acute in molecular biology, in medicine and in the translational road connecting the two.

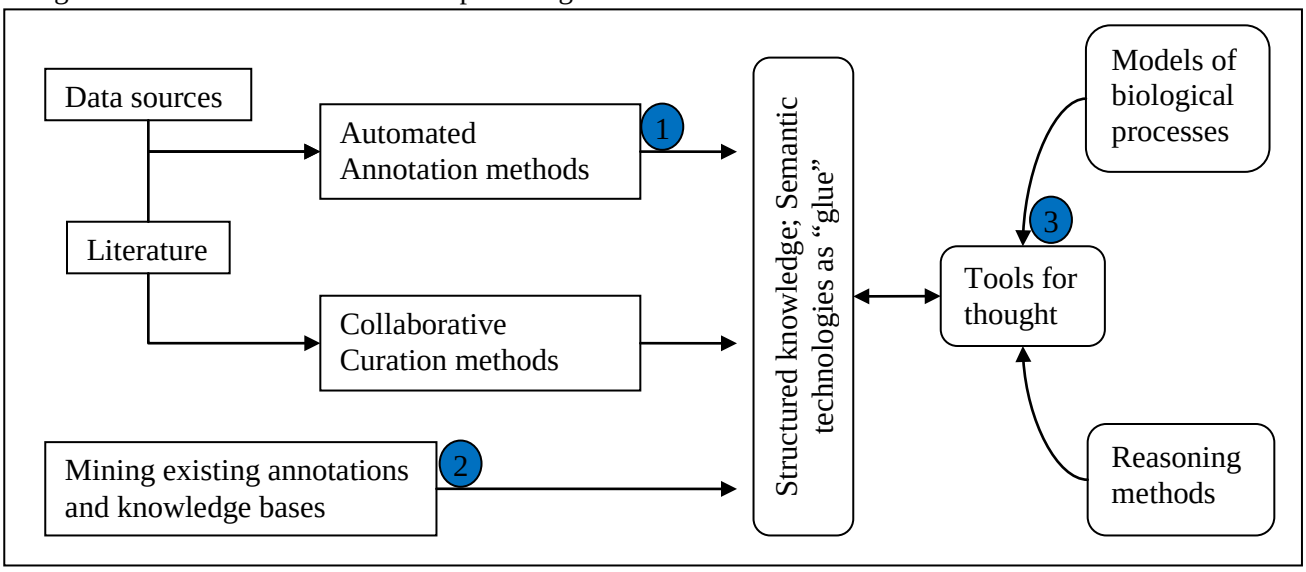
To advance discovery, it is essential to create *tools for thought* in biomedicine, which enable scientists to ask "what if" questions about a system, form explanations, and make and evaluate predictions. It is clear that biomedical computing will need to evolve to address the growing disparity between the massive production of data and the small amounts of knowledge being extracted from these data. The main limiting factor, however, is the formal representation of biological systems and we are quite far from having a consistent representation for biomedical knowledge. Semantic technologies, which standardize both data identification and the knowledge representation, offer a means to enable tools for thought.

The current methods in biomedical informatics that attempt to address this lack of tools for thought can be grouped into two coarse categories: 1) data-centric methods, where quantitative methods are used to spot trends and patterns in large datasets; 2) knowledge-centric methods, where formal knowledge management methods are used to reason about a biological system to guide further exploration. Semantic Web technologies are emerging as the key enabling technology to bridge these two categories and begin to address the data-knowledge gap.

In order to accomplish the goal of creating tools for thought that accelerate discovery, **it is essential to use bio-ontologies as well as semantic web technologies for organizing biomedical information to make it actionable**; for which we need:

1. To develop methods and tools for automating the creation of structured knowledge
2. To develop methods for mining annotations of existing biomedical datasets
3. To develop methods for integrating prior knowledge with current datasets for hypothesis based query and reasoning

The figure below outlines the relationships among these needs.



Medical School Office Building, Room X-229; 1265 Welch Road, Stanford, CA 94305
Phone: 650-725-6236 · Fax: 650-725-7944

Nigam Shah, MBBS, Ph.D.
Assistant Professor of Medicine
Stanford University

1) Methods and tools for automating the creation of structured knowledge

Currently the creation of structured content is primarily via the process of manual curation, which is unlikely to scale. One of the simplest curation tasks is that of annotating entities of interest, such as proteins and tissue-samples, with an ontology term that denotes the disease condition that the entity relates to. Such simple annotation of gene products with Gene Ontology (GO) terms produces the widely known and used Gene Ontology Association files. There is no logical reason that such annotation should be limited to just GO terms. It is possible to create such annotations using disease terms; however doing so manually would cost too much [1]. It is essential to develop automated methods that use existing concept recognizers, natural language processing tools and existing bio-ontologies to allow researchers to create ontology based annotations for a wide variety of datasets and use cases.

2) Methods for mining annotations of existing biomedical datasets

Currently, GO annotations form the cornerstone of high-throughput data analysis, especially for analyzing gene expression data for over-representation (or enrichment) of specific GO-terms in a set of differentially expressed genes. There is no reason that such analyses be restricted to the Gene Ontology. Similar to asking the question: *which biological process is over-represented in my set of interesting genes*; it is possible, using automatically generated disease annotations, to ask the question: *which disease (or class of disease) is over-represented in my set of interesting genes or proteins*. Researchers have shown that by analyzing protein annotations it is possibly to identify mutation types that are over- or under-represented in specific disease [2, 3]. Such analyses can allow identification of general classes of drugs, disease and test results that are enriched in a certain cohort of patients—such as enrichment of heart attack in patients with rheumatoid arthritis, who took Vioxx before 2005 [4].

3) Methods for integrating prior knowledge with new datasets for hypothesis based query and reasoning

Creation and analysis of GO-style annotations is invaluable for summarizing results of large scale experiments. However, as research gets more data intensive it is increasingly difficult to integrate current knowledge about the relationships within biological systems to formulate and evaluate hypotheses about a large number of molecular entities. Annotation analysis alone does not enable a researcher to determine whether such hypotheses are consistent internally or with data, to refine inconsistent hypotheses or to understand the implications of complicated hypotheses—i.e. it does not serve as a tool for thought.

Among the researchers attending the workshop, there have been several efforts to address this bottleneck [5, 6]. In previous work, I developed a system for evaluating alternative hypotheses about biological process models by presenting to the user an explicit listing of what assumptions and relationships must hold in order for their model of a biological process to be true [7]. However, scaling such efforts and putting all the existing data about a particular biological system into a framework that allows a scientist to understand, manipulate and evaluate the relationships among the data remains a major challenge that needs to be addressed [8].

1. Baumgartner, W.A., Jr., et al., *Manual curation is not sufficient for annotation of genomic databases.* Bioinformatics, 2007. **23**(13): p. i41-8.
2. Mort, M., et al., *In silico functional profiling of human disease-associated and polymorphic amino acid substitutions.* Human mutation, 2010. **31**(3): p. 335-46.
3. Lependu, P., M.A. Musen, and N.H. Shah, *Enabling enrichment analysis with the Human Disease Ontology*. Journal of biomedical informatics, 2011.
4. LePendu, P., et al. *Annotation Analysis for Testing Drug Safety Signals*. in *The 14th Bio-Ontologies SIG meeting at ISMB 2011*. 2011. Viena, Austria.
5. Tipney, H.J., et al., *Leveraging existing biological knowledge in the identification of candidate genes for facial dysmorphology*. BMC Bioinformatics, 2009. **10 Suppl 2**: p. S12.
6. Ciccarese, P., et al., *An open annotation ontology for science on web 3.0.* Journal of Biomedical Semantics, 2011. **2 Suppl 2**: p. S4.
7. Racunas, S.A., et al., *HyBrow: a prototype system for computer-aided hypothesis evaluation.* Bioinformatics, 2004. **20**(suppl_1): p. i257-264.
8. Callahan, A., M. Dumontier, and N.H. Shah, *HyQue: evaluating hypotheses using Semantic Web technologies.* Journal of Biomedical Semantics, 2011. **2 Suppl 2**: p. S3.

Medical School Office Building, Room X-229; 1265 Welch Road, Stanford, CA 94305
Phone: 650-725-6236 · Fax: 650-725-7944