

Learning Genetic Pathways Using Bayesian Networks and Qualitative Probabilistic Networks**Huang, Zan¹, Chen, Hsinchun¹, Su, Hua¹, Marshall, Byron B.¹, Smith, Benjamin. L.¹, Watts, George W.², Martinez, Jesse D.²****¹Department of Management Information Systems, University of Arizona, Tucson, AZ, USA; ²Arizona Cancer Center, University of Arizona, Tucson, AZ, USA**

Recent advances in microarray technologies have made possible large-scale gene expression analyses based on simultaneous measurements of thousands of genes. A reverse engineering approach has been used to extract gene regulatory networks in order to reveal the structure of transcriptional gene regulation processes. Although the resulting networks have the potential to help researchers propose and evaluate new hypotheses, their usefulness is limited by its low statistical power due to the inherent dimensionality problem (insufficient samples) in microarray data. Many researchers have proposed the incorporation of existing biological knowledge and other types of data to improve the statistical power of regulatory network learning. In this study, we are particularly interested in developing a principled and scalable framework for incorporating existing biological knowledge into regulatory network learning processes.

Most previous studies that incorporated biological knowledge to enhance learning from gene expression data used small sets of known gene regulatory relations. There are two reasons for the limited scale of these studies: the limited human-encoded knowledge base of gene regulatory relations and the limited overlap between genes showing significant variation in experimental data and genes that are involved in recorded relations. We propose an integrated learning framework that uses an extensive regulatory relation network (knowledge network) built from various types of knowledge sources. These knowledge sources include: human-encoded databases, ontologies, and regulatory relations automatically extracted from literature. We employ a well-established formalism, *qualitative probabilistic networks*, to support reasoning based on this network to provide relevant information for regulatory network learning from gene expression data. Two specific joint learning approaches under this general framework are proposed. (1) A *transitive effect* approach: incorporate transitive regulatory relations derived from the knowledge network into the Bayesian networks learning process. (2) A *synthesized expression* approach: use simulated “expression” data of relevant biological entities derived from the knowledge network to enhance the Bayesian networks learning process. Using the simulated expression data, we expand the regulatory network analysis to also include unobserved biological entities. With this expanded set of “genes” (observed genes and biological entities with simulated data), additional known relations can be incorporated into the Bayesian networks learning process. Both approaches have the potential to incorporate large amounts of existing biological knowledge into the learning process to derive more accurate regulatory networks from the gene expression data. We have experimented with a human gene expression dataset and relevant knowledge sources. The expert evaluation indicated that both approaches resulted in more interesting regulatory relations that may suggest good hypotheses for further experiments.

This research is supported by grant NIH/NLM, 1 R33 LM07299-01, 2002-2005, “Genescene: A Toolkit for Gene Pathway Analysis.”