Carolina Center for Computational Toxicology

Funded by U.S. EPA Cooperative Agreement STAR RD 833825

Progress Report for Project Period 04/01/09 - 03/31/10 (Year 2)

Report Date: 04/21/10

Ivan Rusyn, Principal Investigator

Project 1:
Predictive Modeling of Chemical-Perturbed Regulatory Networks in Systems Toxicology

Preliminary Data and Work Progress

We have focused on the analysis of ToxCast data and the development of the central part of our mechanistic model for metabolic function. With regard to ToxCast data analysis, we have developed a methodology that allows for the discovery of specific modules (sets of genes, proteins) that show specific relevance to specific subsets of chemicals. This approach is different than typical clustering approaches and is based on frequent itemset methodologies. In addition, it provides the ability to identify subsets based on conditionality constraints. This capability allows us to select modules based on specific endpoints – e.g., Liver Carcinogenicity. The outputs from our work have been passed on to Project 3 for pathway-specific QSAR analysis. We have also completed development of an NMR spectra alignment method that allows the "digitization" of instrument outputs and enables more sophisticated downstream statistical analyses that are of particular relevance to toxicity relevant studies. This approach is currently being made into a web application available to outside researchers. We have also developed a novel approach for studying perturbations to molecular networks. This approach is based on the diffusion of information between network nodes along connecting edges, with edges having associated measures of the relevance of the two connected nodes (e.g., probabilities of functional relevance). We are using this approach to help identify high-relevance genes from eQTL data (with Project 2) as well as in the analysis of ToxCast data. Finally, our mechanistic model of metabolism has been under significant development and we have used it to identify key regulators of the response time in switching to glycogen production in both fed and fasted states. We are currently investigating the incorporation of specific toxicity pathways into this model.

Results to Date

Our work was focused on establishing a biological context for ToxCast data and the preliminary analysis of the relative strengths/weaknesses of the individual assays with regard to functional information. A data mining methodology based on frequent itemset approaches has been developed and applied to the analysis of ToxCast data. Capable of leveraging conditional constraints during the search process, we are able to find sets of assays having the greatest functional relevance to one or more particular endpoints. We note that this approach is different

than traditional approaches such as hierarchical and related clustering approaches in that the focus is on finding specific correlated subsets of data under one or more user-defined constraints (e.g. relevance to a particular disease endpoint). A manuscript describing this work is in preparation. Furthermore, we are investigating the addition of “fuzzy” frequent itemset conditions to this approach.

The outputs of these efforts are tightly integrated with those of Project 3. For instance, this analysis has produced additional pathway-based descriptors which are being incorporated into Project 3 QSAR models. As a result, rather than treating ToxCAST assays as individual variables within these models, with Project 3 we are investigating the utility of grouping sets of assays into groups that together provide greater information content that when used individually. This process will help to establish new predictor variables for use in these and other modeling efforts. A manuscript describing the results of this effort is in preparation with members of Project 3.

Building on our previous work in network inference and modeling, we have developed a gene prioritization method related to recent developments in image segmentation methodologies. This approach (currently called segNET) has many of the features of previously developed methods such as molecular triangularization and eQED, while having greater flexibility and range of application. In the context of image analysis, the segmentation approach we use propagates information between neighboring pixels in a manner analogous to diffusion. Transmission of this information to other pixels in the image is based on the use of Laplace’s equation, which describes a variety of physical flow phenomena including energy, gravity, and electromagnetism. There also exists a discretized version of Laplace’s equation, which may be applied to graph structures including the molecular networks typical of biological data. Using this discrete form of the Laplace equation, information transmission becomes a steady state diffusion problem within a graph. In the biological context of ToxCast and related toxicity data sets, information is transmitted between nodes (genes/proteins/metabolites) within the network as a function of the strength of association between them.

With Project 2, we have begun to apply the segNET method to the analysis of eQTL data available from the experimental studies *in vivo* and *in vitro*. In addition, this approach can be generalized to the identification of cliques of genes or proteins that are involved in the response to perturbation, for example, due to chemical exposure. We are currently applying this approach to ToxCast data.

We have published a novel method in BMC Bioinformatics for comparing phylogenetic trees that allows us to infer protein interactions and aid in the inference of protein interaction networks. This approach can work across phyla and thus has particular relevance for cross-species predictions. In addition to tree comparison, this method can be used to compare network-like structures, so as to determine regions of similarity and/or dissimilarity. An improvement of this method has also been recently developed. This approach improves on the ability to compare trees and network structures. In fact, this approach allows the capability to predict interaction specificity between various protein families. The manuscript describing this advance is in preparation. We have similarly pursued structure-based approaches for constructing protein interaction networks that are of relevance to mapping relationships between species (e.g. between mouse and human).

We have continued our development of a mechanistic model of specific aspects of metabolism. Consisting of liver, adipose, muscle and blood components, this model allows us to examine and predict the results of perturbations to network behavior. For example, using this model we have been able to identify components of the glycogen circuit that appear to be responsible for a dramatic decrease in the time it takes to convert glucose to glycogen when in the fasted state. We are further linking the dynamics of this system to gene expression, as well as genetic variation, through the use of data derived from murine strains including the Collaborative Cross. A manuscript describing this model is in progress.

We have also developed a novel method for the alignment of NMR spectra that are typical outputs of systems-level metabolomic analyses. Published in BMC Bioinformatics, this approach (termed PCANS – Progressive Consensus Alignment of NMR Spectra) has potential use in the improved analysis of metabolomic data sets that are relevant in the analysis of toxicity-relevant data sets. We are currently using it in the analysis of both chemical and pharmacological perturbations and their downstream effects on metabolism in both mice and humans. Furthermore, this methodology allows for a full STOCSY (Statistical Total Correlation Spectroscopy) analysis to be undertaken; an analysis approach whose adoption has been limited due to challenges in computation and visualization. Use of the aligned spectral features output from our approach allows such STOCSY analyses to now be readily undertaken. A short paper describing this advance is in preparation. A web application that allows public users to perform the PCANS spectral alignment has also been developed and is currently being deployed onto a publicly accessible server. The inclusion of this STOCSY analysis will be added after appropriate testing. An application note based on this work is also planned for submission.

Activities for Subsequent Reporting Period

Further effort will be placed on extending our mechanistic model of metabolism so as to incorporate more toxicity-relevant features. In particular, we are investigating the possibility of linking in models of nuclear receptor and/or glutathione function into this model. Either in addition or alternatively, we are also considering incorporating particular pathways or components that have been identified as being relevant as a result of ToxCAST data analysis. Of primary consideration here is the availability of sufficient appropriate data for the development and validation of models. To further enhance the relevance of these modeling efforts with EPA interests, we have begun a collaboration with Dr. Imran Shah's group at NCCT. A longer-term goal of these efforts is to help provide models and/or modeling components of relevance to initiatives such as the Virtual Liver Project.

We will continue to develop and apply our network perturbation methodologies in the study of both genetic data (such as the eQTL and related studies of Project 2), as well as in the study of the effect of chemical perturbations on molecular network function. We have initiated an effort with Project 2 that is focused on the use of our segNET methodology in the analysis of eQTL data collected as part of that project. The primary goal in this project will be to determine if our approach can identify a subset of high-priority genes out of the much larger number of genes typically found in these studies. We are further pursuing its use in the analysis of changes in molecular networks in response to chemical perturbation and will be investigating the use of alternative modeling towards this end.

With Project 2, we are also working on establishing metrics of significance for both the segNET methodology as well as our frequent itemset approach for analyzing perturbed networks. These efforts have the potential to establish statistically meaningful metrics that can be used in future decision-making processes.

We will further enhance the efforts of Project 3 with regard to the creation of informative and biologically relevant descriptors for use in QSAR and similar models. We will do this by continuing to develop our data integration methodologies, currently based on Bayesian approaches to combining disparate data types. Such approaches will generally provide a network-based representation of the (potential) relationships between genes, proteins, and/or chemicals. Given such representations, substantial effort will be placed on the use of such network information to find "hidden" components of network structure that can be inferred from measurements of network properties under different perturbation conditions. Our current focus is on finding the appropriate context in which these methods will have the greatest impact on improving our analysis of chemical perturbation data such as that provided in ToxCast.

The PCANS tool is the first web-accessible tool we will deploy (it is already distributed as a standalone software package) and we plan to continue to migrate associated tools and methodologies to the web for broader utility to the research community. The tree/graph comparison algorithms have currently been made available as standalone software.

Publications Arising From this Project

Papers

1. Choi, K., and Gomez, S.M. (2009) Comparison of phylogenetic trees through alignment of embedded evolutionary distances. BMC Bioinformatics 10:423 (*from submitted to published*).
2. Weinreb, G.E., Kapustina, M.T., Jacobson, K., and Elston, T.C. (2009) In silico generation of alternative hypotheses using causal mapping (CMAP). PLoS One, 4:e5378.
3. Tsygankov D, Liu Y, Sanoff HK, Sharpless NE and Elston TC. (2009) A quantitative model for age-dependent expression of the p16INK4a tumor suppressor. PNAS, 106:16562-16567.
4. Strychalski, W., Adalsteinsson, D., and Elston, T.C. (2010) A cut-cell method form modeling spatial sytems of biochemical reactions in arbitrary cell geometries. Comm Appl Math Comp Sci 5:31-53.
5. Staab, J., O'Connell, T.M. and Gomez, S.M. (2010) Enhancing metabolomic data analysis with Progressive Consensus Alignment of NMR Spectra (PCANS). BMC Bioinformatics 2010, 11**:**123 (*from submitted to published)*.
6. Doolittle, J.M. and Gomez, S.M. (2010) Structural similarity-based predictions of protein interactions between HIV-1 and Homo sapiens. Virology J (*in press*).
7. Strychalski, W., Adalsteinsson, D., and Elston, T.C. (2010) Simulating biochemical networks in complex moving geometries. SIAM J Sci Comput (*in press*).

Posters/Abstracts

1. O'Connell, T.M., Staab, J. and Gomez, S.M. Enhancing metabolomic data analysis with Progressive Consensus Alignment of NMR Spectra (PCANS). 5th International Conference of the Metabolomics Society, Aug 30th – Sep 2nd, 2009, Edmonton, Canada.