MEETING ABSTRACTS

UT-ORNL-KBRIN Bioinformatics Summit 2011

Memphis, TN, USA. 1-3 April 2011

Edited by Eric C Rouchka, Robert M Flight and Ramin Homayouni

Published: 5 August 2011

These abstracts are available online at http://www.biomedcentral.com/1471-2105/12?issue=S7

INTRODUCTION

A1
Proceedings of the Tenth Annual UT-ORNL-KBRIN Bioinformatics Summit 2011
Eric C Rouchka1, Robert M Flight2, Ramin Homayouni3
1Department of Computer Engineering and Computer Science, University of Louisville, Duthie Center for Engineering, Louisville, KY 40292, USA;
2Department of Anatomical Sciences and Neurobiology, University of Louisville, Louisville, KY 40292, USA;
3Department of Biological Sciences, University of Memphis, Memphis, TN 38152, USA
E-mail: eric.rouchka@louisville.edu

BMC Bioinformatics 2011, 12(Suppl 7):A1

The University of Tennessee (UT), the Oak Ridge National Laboratory (ORNL), and the Kentucky Biomedical Research Infrastructure Network (KBRIN), have collaborated over the past decade to share research and educational expertise in bioinformatics. One result of this collaboration is the joint sponsorship of an annual regional summit to bring together researchers, educators and students who are interested in bioinformatics from a variety of research and educational institutions. This summit provides unique opportunities for collaboration and forging links between members of the various institutions. This year, the Tenth Annual UT-ORNL-KBRIN Bioinformatics Summit was held at the University of Memphis, Tennessee from April 1-3, 2011. A total of 225 participants pre-registered for the summit, with 146 from various Tennessee institutions and 60 from various Kentucky institutions. A number of additional participants came from universities and research institutions from other states and countries, e.g. University of Arkansas Medical Sciences, University of British Columbia, University of Cincinnati, Iowa State University, etc. Seventy-four registrants were faculty, with an additional 73 students, 50 staff, and 28 postdoctoral participants.

The conference program consisted of three days of presentations. The first day included a pre-summit of talks by researchers supported by the Kentucky Biomedical Research Infrastructure Network (KBRIN), and three workshops covering the topics Next-Generation Sequencing, the workflow platform Galaxy, and GeneMANIA. The next two days were dedicated to scientific presentations divided into three plenary sessions on Pediatric Genetics, Pharmacogenomics and Systems Biology. Each session also included short talks selected from the submitted poster abstracts.

Pre-summit KBRIN session: Dr. Eric Rouchka started the pre-summit KBRIN session with an update on the supplement to the primary KBRIN grant awarded in 2009. Thanks to the funding provided by the supplement, five postdoctoral research associates have been hired, along with two masters level bioinformatics staff positions. In addition, a weekly seminar series has been implemented that includes externally invited speakers with strong research programs in bioinformatics.

Following was a series of short talks from various University of Louisville researchers: "Developing an Analysis Pipeline for FT-ICR-MS Isotopologue Data from Stable Isotope Resolved Metabolomics (SIRM) Experiments" (Hunter Moseley), "Time Series Classifier Model for miR-mRNA Relationships" (Jovan D. Rebollo-Mendez), "Assessing Variations in NGS Data" (Alex Kemper), "A Systems Based Approach to Find Protein Interactions Across Tissues" (Fahim Mohammadi), "categoryCompare: High-Throughput Data Meta-Analysis Using Gene Annotations" (Robert M. Flight), "Collaborations between Biology/Bioinformatics" (Benjamin J. Harrison).

Friday workshops: Jon Armstrong of Cofactor Genomics (St. Louis, MO; http://www.cofactorgenomics.com) started the official summit program with an excellent overview of the past, present and future of DNA sequencing, beginning with the Maxam-Gilbert [1] and Sanger [2] methods and transitioning to the approaches used by next-generation sequencers. Among the technologies covered were pyrosequencing [3], reversible dye terminator technology [4], sequencing by ligation [5], and single molecule real time sequencing [6]. As Jon walked through each of the technologies, he also sought to explain what types of experiments are best suited for each NGS platform, based on the strengths and weaknesses of each technology. Given the rapidly growing number of NGS machines available, the information was extremely useful for both those planning NGS experiments and those analyzing the resulting data. Following up, Dr. James Taylor from Emory University gave a workshop on the workflow system GALAXY [7,8]. Dr. Taylor started with a comprehensive explanation of the many research issues motivating the development of GALAXY as an easily modifiable, reproducible workflow system for both biologists and bioinformaticians working with high-throughput data. He then proceeded to explain and demonstrate many of GALAXY’s features by performing real world analyses using data from the UCSC genome browser [9]. Notable features include the ability to share workflows with others, the ability to generate results pages (similar to full blown publications) with workflows embedded in the document and accessible to anyone viewing the page, and the recent ability for NGS core lab implementations to couple sample requests to workflows so that data processing can occur as soon as the data becomes available. As a final note, Dr. Taylor also demonstrated how easy it is to set up GALAXY instances in Amazons EC2 cloud computing system, enabling one to take advantage of cloud computing systems [10].

The final workshop was given by Dr. Quaid Morris from the University of Toronto on the use of GeneMANIA [11] for pathway and network analysis. GeneMANIA uses gene association networks to assign probable functions of genes based on guilt by association: those genes that share connections (annotations or interactions) probably share other attributes as well. GeneMANIA is available either as a Cytoscape plugin [12,13] or on the web. Session I: pediatric genetics: The first session began with opening remarks by Dr. Ramin Homayouni from the University of Memphis. Dr. Homayouni provided a 10 year retrospective of the summit, its beginnings and growth, and the many collaborations and new bioinformatic tools that have resulted from those attending the conference. The Pediatric Genetics session was truly underway with Dr. Hakon Hakonarson of The Children’s Hospital of Philadelphia discussing his research on determining the genetic underpinnings of complex pediatric disorders. As a way to try and cope with the large number of rare disorders, as a way to try and cope with the large number of rare
3Prop extends its previous work of predicting gene function given limited annotation information [11,45]. This approach has been applied to identify the gender of users of social media sites by positively weighting their genders and negatively weighting their friends’ genders.

The second talk of the session was the final plenary speaker, David Galas, from the Institute for Systems Biology who presented “Genetics in the Age of Sequencing: Converging on Complexity.” In this talk, Dr. Galas discussed the transformation of human genetics as a result of whole genome sequencing and associated computational approaches. A discussion ensued about the advances that genome wide association studies (GWAS) have led into the understanding of complex diseases [46], including 1212 published genome-wide associations for 210 traits as of 12/2010 (http://www.genome.gov/gwastudies). However, the point was made that familial-based sequencing can be a powerful technology for complex disease association, as demonstrated by a family of four sequencing projects in which two siblings and their parents were sequenced [47]. Dr. Galas discussed the results of this project which was based on two recessive Medallian disorders displayed in the siblings: Miller syndrome and primary ciliary dyskinesia. By looking at familial sequencing, the ability to detect recombination events with high precision is possible, which can lead to more directed analysis of Parabellum at a gene level. Dr. Galas expanded upon additional current research at the Institute for Systems Biology, including a multigenerational sequencing project and a project to sequence 600 individuals (all in families) focusing on diseases such as Huntington’s and congenital heart defects.

Posters and short talks: The poster session was held on day two before the main banquet. Forty-eight posters were on display, all from a variety of different research areas. A number of posters were also selected for short talks in each session. The talks and the presenting authors are listed below for each session. For full author lists and abstracts see the rest of the supplement.


Future plans: The 2012 Bioinformatics summit will return to the state of Kentucky in the spring of 2012. Potential focus areas include current technological trends in molecular biology, applications of next-generation sequencing, and systems biology.

Acknowledgements: We would like to thank the additional Conference Program Committee members Nigel Cooper (University of Louisville), Dan Goldowitz (University of British Columbia), Julia Krushkal (University of Tennessee-Memphis), Mike Langston (University of Tennessee-Knoxville), Terry Mark-Major (University of Tennessee-Memphis), Cynthia Peterson (University of Tennessee-Knoxville), Claire Rinehart (Western Kentucky University) Arnold Stromberg (University of Kentucky), Rob Williams
(University of Tennessee-Memphis) and Zhongming Zhao (Vanderbilt University) for organizing an outstanding scientific program. In addition, I want to thank the conference organization details. Funding for the UT-ORNL-KBRIN CloudMan: delivering cloud compute clusters. BMC Bioinformatics 2010, 11(Suppl 12):S4.

References


Genome-wide analyses of exonic copy number variants in
Science
89
From disease association to risk assessment: an optimistic
BMC Med Genet
7
469
PLoS Genet
19
Thyroid
3
J Cardiovasc Pharmacol
459
2009,
5
2009,
Copy number variation at 1q21.1 associated with
32. Gai X, Xie HM, Perin JC, Takahashi N, Murphy K, Wenocur AS, D
31. Bucan M, Abrahams BS, Wang K, Glessner JT, Herman EI, Sonnenblick LI,
Glessner JT, Chiavacci R, Stanley C, Monos D, Grant SF, Polychronakos C,
Hakonarson H. Comparative genetic analysis of inflammatory bowel disease and type 1 diabetes implicates multiple loci

31. Bucan M, Abrahams BS, Wang K, Glessner JT, Herman EI, Sonnenblick LI,
30. Magalhaes TR, Correia C, Abrahams BS, Almeida J, Bacchelli E, Bader GD,
23. Bucan M, Abrahams BS, Wang K, Glessner JT, Herman EI, Sonnenblick LI,
22. Magalhaes TR, Correia C, Abrahams BS, Almeida J, Bacchelli E, Bader GD,
20. Magalhaes TR, Correia C, Abrahams BS, Almeida J, Bacchelli E, Bader GD,
17. Wang LS, Hranilovic D, Wang K, Lindquist E, Yucabia L, Petkovic ZB,
15. Wang LS, Hranilovic D, Wang K, Lindquist E, Yucabia L, Petkovic ZB,
2. Wang LS, Hranilovic D, Wang K, Lindquist E, Yucabia L, Petkovic ZB,

15. Wang LS, Hranilovic D, Wang K, Lindquist E, Yucabia L, Petkovic ZB,
2. Wang LS, Hranilovic D, Wang K, Lindquist E, Yucabia L, Petkovic ZB,
Previously, a quantitative trait loci (QTL) for alcohol preference on chromosome 2 in a C57BL/6J-BG (B6) background has been identified. The overlap of two of interval specific congenic recombinant strains (ISCRS) strains reduced the QTL interval into a 3.4 mbp region.

Results: By using the keyword alcohol, we identified a total of 39 genetic elements in the region between markers D2Mit56 and D2Mit10. Among these genetic elements, we found seven with potential function in alcohol preference (Table 1). We then examined the SNPs, insertions and deletions, and gene expression levels of those seven genes.

Conclusions: Our current data suggest that the Atf2 and Titin genes are potentially the most alcohol relevant genes. However, further experiments and examination are still needed to confirm their candidacy. Several other candidate genes are also in the process of being identified.

Acknowledgments: Support for this research is partially from the NIAAA (1R01 AA016342). NIH, the Veterans Administration Medical Center, and DNA Discovery Core, University of Tennessee, Memphis, TN.

Table 1 Candidate genes for alcohol preference on Chr 2

<table>
<thead>
<tr>
<th>ENSEMBL ACCESSION</th>
<th>SYMBOL</th>
<th>FULL NAME</th>
<th>SNPS</th>
<th>INSERTIONS</th>
<th>DELETIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSMUSG00000027104</td>
<td>AT2</td>
<td>ACTIVATING TRANSCRIPTION FACTOR2</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENSMUSG00000027109</td>
<td>SP3</td>
<td>TRANS-ACTING TRANSCRIPTION FACTOR3</td>
<td>1</td>
<td>G/A</td>
<td>AT(72784944)</td>
</tr>
<tr>
<td>ENSMUSG00000006494</td>
<td>PDK1</td>
<td>PYRUVATE DEHYDROGENASE KINASE, ISOENZYME 1</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENSMUSG00000009207</td>
<td>LNP</td>
<td>LIMB AND NEURAL PATTERNS</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENSMUSG00000027107</td>
<td>CHRNA1</td>
<td>CHOLINERGIC RECEPTOR, NICOTINIC, ALPHAPOLYPEPTIDE1</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENSMUSG00000018770</td>
<td>ATP5G3</td>
<td>ATP SYNTHASE, H+ TRANSPORTING, MITOCHONDRIAL FO COMPLEX, SUBUNIT C, (SUBUNIT 9), ISOFORM 3</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A novel noise handling method to improve clustering of gene expression patterns

Anindya Bhattacharya1*, Rajat K De2

1Center for Integrative and Translational Genomics, University of Tennessee Health Science Center, Memphis, TN, 38163, USA; 2Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India

E-mail: anindyamail@rediffmail.com

BMC Bioinformatics 2011, 12(Suppl 7):A3

Background: Cluster analysis of gene expression data is a useful tool for identifying biologically relevant groups of genes that show similar expression patterns under multiple experimental conditions. Performance of clustering algorithms is largely dependent on selected similarity measure. Efficiency in handling outliers is a major contributor to the success of a similarity measure. In gene expression data, there may be pairs of genes that have completely different expression values over a few samples under certain experimental condition(s), although they exhibit similar behavior over the other samples. Depending on the algorithms, these outliers are either placed in single element clusters (hierarchical clustering), are allowed to be in a cluster that is more similar compared to others (partitioning clustering) or they may be completely discarded from grouping (density-based, grid-based and graph-based clustering). In all these cases outliers affect the outcome of a clustering result. Measurement errors or conditional changes during microarray experiments may cause a single sample, if not more, differing in expression level to a great extent compared to the other samples. Expression value of the single or a very few outlier samples may cause a gene to be an outlier. We formulate a new weighted function based method to reduce the effect of outliers on similarity measures. The better the similarity measure is in measuring similarity between genes in the presence of outliers, the better the performance of the clustering algorithms will be in forming biologically relevant groups of genes.

Results: The effectiveness of the weighted function based method has been demonstrated with the clustering algorithms, viz., K-means [1], Minimization of Disagreement (MIND) [2], Divisive Correlation Clustering Algorithm (DCCA) [3], Average Correlation Clustering Algorithm (ACCA) [4] and Bi-Correlation Clustering Algorithm (BCCA) [5] on a yeast gene expression dataset (Yeast Cheng and Church dataset from Yeast Functional Genomics Database [http://yfgdb.princeton.edu/]). Assessment of the results has been done by using P-values on functional annotations. P-values less than 5.0 × 10-7 are reported as enriched functional...
categories. Figure 1 shows the number of functionally enriched attributes in the most enriched clusters obtained by each of the clustering and biclustering algorithms on the yeast gene expression dataset. The results suggest that the new weighted function based method significantly improves performance of all the cases, in terms of finding biologically relevant groups of genes.

References

A4 Identifying the key genes and pathways in the progression of hepatitis C virus induced hepatocellular carcinoma using a systems biology approach

Siyan Zheng1, Zhongming Zhao1,2
department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, 37232, USA; 1department of Cancer Biology, Vanderbilt University Medical Center, Nashville, TN, 37232, USA.
E-mail: zhongming.zhao@vanderbilt.edu
BMC Bioinformatics 2011, 12(Suppl 7):A4

Background: Incidence of hepatitis C virus (HCV) induced hepatocellular carcinoma (HCC) has been increasing in many developed countries including the United States and Europe during the recent years. Although many efforts have been made to understand the pathogenesis, the picture of its progression still remains elusive.

Materials and methods: We developed a systematic approach to identify deregulated biological networks in HCC by integrating gene expression profiles [1] with high-throughput protein-protein interaction data [2]. Samples were grouped into five disease stages including normal, cirrhotic, dysplastic, early and advanced HCC. For each pair of consecutive stages, we compared gene expressions and then mapped these measures to the protein interaction network. Responsive subnetworks were then identified from these node weighted networks. The searching algorithm is adapted from a previous study [3], which expands the seed graphs under constrains of several parameters.

Results: Four networks were identified including precancerous networks (normal-cirrhosis and cirrhosis-dysplasia) and cancerous networks (dysplasia-early HCC, early-advanced HCC). A summary of these networks is shown in Table 1. An independent dataset was used for network validation. Statistical significance of these networks was assessed within three hypotheses. Little overlap was observed between precancerous and cancerous networks, in contrast to a substantial overlap within precancerous or cancerous networks. Network functions were annotated with Gene Ontology biological process using hypergeometric distribution based enrichment analysis. Significant functions were then assembled into a module map in temporal order. The apoptosis gene ZBTB16 was highlighted by examining the module map, which shows a negative expression pattern with c-myc. Network analysis led to the identifications of key genes and pathways by developmental stage, such as LCK signaling pathways in cirrhosis, MMP genes and TIMP genes in dysplastic liver, and CD2-C-mediated cell cycle signaling in early and advanced HCC.

Conclusions: Our study uncovers a temporal spectrum of functional deregulation and prioritizes key genes and pathways in the progression of HCV induced HCC. Despite the confirmation of much knowledge in the pathogenesis of this disease, these findings also provide additional insights for further investigations.

Acknowledgements: We thank Drs. Scott Hiebert, William Tansey, Jingchun Sun and Peillin Jia and Mr. Jeffery Ewers for helpful discussions.

References
4. Jordan M Utley, Elizabeth E Howell, Jeroome Baudry, Robert J Hinde: Graduate School of Genome Science and Technology, University of Tennessee, Knoxville, TN, 37996, USA; 1department of Biochemistry, Cellular, and Molecular Biology, University of Tennessee, Knoxville, TN, 37996, USA; 2department of Bioinformatics, Cellular, and Molecular Biology, University of Tennessee, Knoxville, TN, 37996, USA; 3Center for Molecular Biophysics, Oak Ridge National Laboratory, TN, 37831, USA; 4department of Chemistry, University of Tennessee, Knoxville, TN, 37996, USA. E-mail: jharri43@utk.edu
BMC Bioinformatics 2011, 12(Suppl 7):A5

A5 Determining anion-quadrupole interactions among protein, DNA, and ligand molecules

Jason B Harris1,2, David D Jenkins1, Jonathan Rayles1, Stephanie Rickett2, Jordan M Utley1, Elizabeth E Howell1,1, Jeroome Baudry1,1, Robert J Hinde1,1
1Graduate School of Genome Science and Technology, University of Tennessee, Knoxville, TN, 37996, USA; 2Graduate School of Genome Science and Technology, University of Tennessee, Knoxville, TN, 37996, USA; 3Department of Biochemistry, Cellular, and Molecular Biology, University of Tennessee, Knoxville, TN, 37996, USA.
BMC Bioinformatics 2011, 12(Suppl 7):A5

Background: An extensive search through the Protein Databank (about 4500 nonredundant structures) was previously completed within our lab to analyze the energetic and geometric characteristics of an understudied molecular interaction known as an anion-quadrupole (AQ) interaction. Such an interaction occurs when the positively charged edge of an aromatic ring, resulting from a quadruple moment (i.e., a dual dipole moment), renders the aromatic molecule noncovalently bound to a nearby anionic molecule. The study considered a very limited scenario of molecules that can participate in AQ interactions, consisting of the phenyl group of a phenylalanine (phe) amino acid as the aromatic participant and the carboxylate group of an aspartate (asp) or glutamate (glu) amino acid as the anionic participant. The results revealed anion-quadrupole pairs to be prevalent within most of the protein structures. It was also observed that the interaction energy for AQ pairs was heavily dependent on the angle between the anion and plane of the aromatic ring, favoring a more planar interaction.

In light of these critical observations being made from such a limited scenario, only phe-glu and phe-asp pairs and in a reduced sample set of the PDB, we are now continuing this work of identifying AQ interactions using a greatly expanded strategy. We are following these four aims: 1. Optimizing the AQ-search program to run in a semi-parallel fashion and on a large cluster of processors in order to handle larger analyses, 2. Adding to our search additional anionic participants which will include

Table 1(abstract A4) Overview of the responsive networks

<table>
<thead>
<tr>
<th>Network</th>
<th>#Genes</th>
<th>#Interactions</th>
<th>#DEGs</th>
<th>#Hub interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal- cirrhosis</td>
<td>55</td>
<td>67</td>
<td>53</td>
<td>42 (62.7%)</td>
</tr>
<tr>
<td>Cirrhosis-dysplasia</td>
<td>38</td>
<td>50</td>
<td>37</td>
<td>35 (70.0%)</td>
</tr>
<tr>
<td>Dysplasia –early HCC</td>
<td>60</td>
<td>65</td>
<td>53</td>
<td>37 (56.9%)</td>
</tr>
<tr>
<td>Early-advanced HCC</td>
<td>68</td>
<td>98</td>
<td>59</td>
<td>79 (80.6%)</td>
</tr>
</tbody>
</table>

1Differentially expressed genes (DEGs) were identified as genes with up or down regulation fold change ≥ 2 and student t test P value ≤ 0.01.
2Hub interaction number refers to the total number of interactions involving hub genes.
3Hub genes were defined to have at least 5 interactions in each network.
non-protein structures such as DNA and small ligands. 3. Studying a subset of the AQ pairs with molecular dynamics simulations in buried and solvent exposed environments to observe non-static behavioral traits as well as the reproducibility of AQ interactions by force field parameters.

4. Building an online database for public access to our data and search program.

Acknowledgments: We would like to acknowledge the NSF-IGERT traineeship, Scalable Computing and Leading Edge Innovative Technologies (SCALE-IT), for providing the resources for this project.

A6
Integrative biclustering of heterogeneous datasets using a Bayesian nonparametric model with application to chemogenomics
Dazhuo Li, Eric C Rouchka
Department of Computer Engineering and Computer Science, University of Louisville, Louisville, KY, 40292, USA
E-mail: eric.rouchka@louisville.edu
BMC Bioinformatics 2011, 12(Suppl 7):A6

Motivation: The identification of protein function and the prediction of ligand-target interaction is an active research field that is facilitated by means of categorizing ligands and proteins into biologically sensible groups. Because of the pharmacological fact that related drugs can bind to receptors without obvious sequence or structural similarity, it is appropriate to categorize proteins based not only on their sequence or structures but also on the chemical structure and the phenotypic side-effect of their ligands. In chemogenomic studies where the complete set of ligands for a protein is not known a priori, integrating the de novo detection of interacting ligand and protein groups into the categorization process can guide the process towards more biologically sensible solutions.

Results: We present the Weighted Infinite Relational Model (WIRM) that jointly detects biologically sensible ligand groups and protein groups by integrating the clustering of various data types including chemical compound descriptors, protein sequences, ligand-target bindings and pharmaceutical effects. WIRM takes advantage of the Bayesian nonparametric paradigm for integrating multiple data types, for allowing for missing values (e.g. unknown ligand-target interaction) in the data, for automatically inferring the number of clusters without explicit model comparison, and for predicting the ligand-target interactions. Because some of these data types, to varying degrees, may suggest relationships having no implication for ligand-target interactions or for biological sensible ligand and protein groups, WIRM allows different types of data to have different weights based on prior knowledge of their quality or relevance.

Conclusion: We apply WIRM to the ion channel proteins and G-protein-coupled receptors. We validate its performance using functional annotation and ligand-target interaction. We also test the relationship among multiple data types by varying the weights which indicate the impact of each data type on the model. The categories and interactions inferred by WIRM both confirm known biology and suggest novel predictions.

Acknowledgements: This work was supported in part by the National Institutes of Health (P20RR16481, P20RR16481S1, P30ES014443) and the Department of Energy (DE-EM0000197). Its contents are solely the responsibility of the authors and do not represent the official views of NCRR, NIEHS, NIH, or DOE.

A7
mDAG: a web-based tool for analyzing microarray data with multiple treatments
Vinhthuy T Phan1*, Nam S Vo2, Thomas R Sutter3
1Department of Computer Science, University of Memphis, Memphis, TN, 38152, USA; 2Bioinformatics Program, University of Memphis, Memphis, TN, 38152, USA; 3Department of Biological Sciences, University of Memphis, Memphis, TN, 38152, USA
E-mail: vphan@memphis.edu
BMC Bioinformatics 2011, 12(Suppl 7):A7

Background: In microarray experiments involving multiple treatments, pairwise comparisons between all pairs of treatments are desirable but expensive. To cope with this, we previously introduced a method that performed all pairwise comparisons in a post hoc manner. This method employs directed graphs to represent gene response to pairs of treatments. It has been applied and found useful in identifying and differentiating genes sharing similar functional pathways [1,2].

Results: mDAG is a web-based software based on this method. mDAG allows users to upload microarray data in GCT format through a web interface. From this data, the application performs calculations to assess graphical patterns to genes and outputs images and textual data for further analyses. These graphical patterns carry specific meanings in terms of how genes respond to pairs of treatments. The software can be used online or off-line.

Acknowledgements: This work was supported by the Center for Alternatives to Animal Testing at the John Hopkins school of Public Health, Project CAAT-2011-18.

References

A8
Protected Research Information Management Environment (PRIME) provides a secure open source data management option for clinical and scientific research
Teeradache Vangteeravat1, Venkateswarra Ra Nagisetty1, Matthew N Anyanwu2, Emin Kuscu3, Ian M Brooks3, Chanchai S McDonald4
Biomedical Information Sciences Unit, Clinical & Translational Science Institute, University of Tennessee Health Science Center, Memphis, TN 38163
E-mail: tvangte@uthsc.edu
BMC Bioinformatics 2011, 12(Suppl 7):A8

Meeting abstract: The Biomedical Information Sciences Unit (BISU) of the UT-CTSI developed a secure, centralized information management system that is accessed through web applications. This system, called Slim-Prim [1,2], is mounted on an Oracle server housed within the UT-ITS and currently hosts three multi-site clinical trials and several institutional trials. In addition, Slim-Prim houses databases of clinical and laboratory information and provides a centralized location for several physician-generated surveys. In the majority of these cases users have approached the BISU and asked for assistance with their data management needs. Slim-Prim [1,2] and the BISU provide a cost effective mechanism for funded groups to handle their data management needs. However, we are also aware that many smaller groups are in desperate need of secure yet flexible database solutions; thus there is the need for a portable, secure and open source instance of Slim-Prim. To this end we have developed and released an open source instance of the Slim-Prim system under the Affero GPL v3 (http://www.gnu.org/) called "Protected Research Information Management Environment (PRIME)." The PRIME instance is written in PHP, based on an object-oriented programming (OOP) architecture, and JavaScript with inverted file indexing search algorithms to render the graphical user interface (GUI) (http://www.php.net/). PRIME has a user-friendly interface that is menu driven with help features that assist the user in navigating through the application. PRIME exploits the MySQL database system and allows powerful SQL queries across tasks through using fully normalized table structures (http://www.mysql.com/). It is a web-enabled application that runs on most popular browsers. It is provided to the clinical and translational science community as a free and easily customized database solution.

Acknowledgement: This work was supported by funds from the University of Tennessee Clinical Translational Science Institute (CTSI).

References
Phage Eco-Locator: a web tool for visualization and analysis of phage genes in metagenomic data sets

Ramy K Aziz,1,2*, Bhakti Dwivedi,3 Mya Breitbar,4 Robert A Edwards1,4
1Department of Computer Science, San Diego State University, San Diego, CA, 92182, USA; 2Department of Microbiology and Immunology, Faculty of Pharmacy, Cairo University, Cairo, Egypt; 3College of Marine Science, University of South Florida, St. Petersburg, FL, USA; 4Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, USA
E-mail: ramy.aaziz@salmonella.org

BMC Bioinformatics 2011, 12(Suppl 7):A9

Background: Bacteriophages, viruses that infect bacteria, are the most abundant biological entities on our planet, and their nucleic acids constitute a substantial proportion of total DNA in Earth's ecosystems [1,2]. While the advent of metagenomic methods has allowed the rapid and efficient investigation of microbial and viral communities [3-5], there has not been a comprehensive comparative analysis of phage genes and genomes present in all sequenced ecosystems [6,7]. To examine the abundance and distribution of phage genes in environmental metagenomic sequences, we developed a web-based tool, Phage Eco-Locator (http://www.phantome.org/eco-locator) that screens all publicly available sequenced metagenomes for a user-defined phage genome, or all phage genomes within a user-selected metagenomic sample.

Materials and methods: The tool relies on pre-calculated tBLASTX searches in which metagenomic sequence reads are the input query and all phage genomes are the BLAST database [8]. For optimization, several BLAST parameters have been tested, and the best results are obtained when all tBLASTX matches above a threshold E-value of 10^-5 are included as positive hits. Positive hits are then mapped to phage genome scaffolds and visualized in two different types of plots: one representing sequence hits at different similarity scores (Fig. 1; upper panel) and another representing the coverage density over phage nucleotides (Fig. 1; lower panel).

Results: All 588 phage genomes available in the PhAnToMe database (http://www.phantome.org) (as of January 1, 2011) were screened against 296 de-replicated metagenomic libraries. The graphical output was translated into metrics representing phage abundance, extent and breadth of distribution, and coverage density and evenness. Applying these metrics to all samples demonstrated a pervasive, yet uneven, distribution of phage genes in metagenomic libraries and allowed the separation of phage genomes into distinct groups. The analyses also showed a tendency for phage genomes to prevail in environments similar to their original isolation source, where their bacterial hosts are expected to thrive (e.g., cyanophages in aquatic samples and halophages in hypersaline environments).

Conclusion: Phage Eco-Locator effectively allows the global analysis of all phage sequences in metagenomes while also permitting gene-level analysis of individual phage genomes. In the future, application of this tool to sequences from a wide range of ecosystems will enhance our understanding of the factors controlling phage biogeography and environmental selection.

Acknowledgments: This work was supported by the PhAnToMe grant from the NSF Division of Biological Infrastructure to RAEE (NSF DBI-0850356) and MB (NSF DBI-0850206).

References

Pharmacomicrobiomics or how bugs modulate drugs: an educational initiative to explore the effects of human microbiome on drugs

Ramy K Aziz1,2*, Rama Saad3, Mariam R Rizkallah4
1Department of Microbiology and Immunology, Faculty of Pharmacy, Cairo University, 11562 Cairo, Egypt; 2Department of Computer Science, San Diego State University, San Diego, CA, 92182, USA; 3Department of Biology, School of Sciences and Engineering, The American University in Cairo, 11835 Cairo, Egypt; 4Open Source Technologies Department, Information Technology Institute, 12577 Giza, Egypt
E-mail: ramy.aaziz@egybio.net

BMC Bioinformatics 2011, 12(Suppl 7):A10

Background: Pharmacogenomics investigates how variations within the human genome affect the action and disposition of drugs as well as drug tolerance [1]. Yet, variations within the human genome do not fully account for the tremendous phenotypic variations observed between individuals. Human-associated microbes, which exceed the human cells in number, significantly contribute to the effective human gene pool, and their combined genomes (known as the human microbiome) have not gained attention until recently. The Human Microbiome Project was launched in 2007 to catalogue the tremendous diversity of cultured and uncultured human-associated microbial communities residing in different human tissues, and to study the effect of microbial genes and genomes on human health and disease [2,3]. However, the effect of these microbes on drugs remains largely unexplored. Since microbes have complex metabolism, including an extraordinary ability to metabolize xenobiotics [4-6], they are expected to play a pivotal role in modulating the action, disposition, and toxicity of drugs with which they interact in different sub-ecosystems within the human body [7].

Materials and methods: The Pharmacomicrobiomics initiative (http://pharmacomicrobiomics.org) is a research-based educational web platform that aims at exploring how microbes modulate drugs. The project was launched as an educational platform to introduce bioinformatics and microbial genomics to pharmacy students while
benefiting the research community. The first step of this project was mining existing literature and extracting known microbe-drug interactions using a combination of keywords in an iterative process. The second step was the manual curation of the extracted literature data and their classification by drug classes, microbial families, and body systems (e.g., Table 1). The third step is the creation of a relational database that includes the microbes at different body sites and their effects on drugs’ pharmacokinetic and pharmacodynamic properties. Finally, participating students screen and attempt to isolate fecal microbes that alter a specific drug, and each student selects a drug class and a microbial species within a body site to examine their complex interaction in vitro.

Conclusion: The literature-mining steps of the pharmacomicrobiomics project have resulted in the initiation of a continuously updated web portal maintained by students (http://pharmacomicrobiomics.org/papers and http://pharmacomicrobiomics.com/examples.html). The project is expected to build a knowledge base that allows interested students and scholars, in the future, to predict the behavior of untested members of drug classes or unstudied microbial species, and to design laboratory experiments for testing these predictions.

References

A11
ColonyTrak: a web tool and database system for managing experimental animal models
Laasya Vadlamudi1, Lynn A Jones2, Ramin Homayouni1,2*
1Bioinformatics Program, University of Memphis, Memphis, TN, 38152, USA; 2Department of Biology, University of Memphis, Memphis, TN, 38152, USA
E-mail: rhomayon@memphis.edu
BMC Bioinformatics 2011, 12(Suppl 7):A11

Background: Genetically modified animal models are useful for understanding mechanisms of human disease and for development of

Table 1 (abstract A10) Effects of gut microbes on drugs

<table>
<thead>
<tr>
<th>CID</th>
<th>Drug</th>
<th>Body Site</th>
<th>Microbial effects</th>
<th>NCBI PMID</th>
</tr>
</thead>
<tbody>
<tr>
<td>64982</td>
<td>Baicalin, (Potential antioxidant, anti-inflammatory and liver tonic)</td>
<td>Gut</td>
<td>Gut microbes hydrolyze baicalin and enhance its absorption. Absence of gut microbiota resulted in lower levels of baicalin in plasma following oral administration [8].</td>
<td>11197087</td>
</tr>
<tr>
<td>2724385</td>
<td>Digeoxin, (Cardiac glycoside)</td>
<td>Gut</td>
<td>Eubacterium lentum is responsible for the difference in metabolite concentration of digoxin between North Americans and Southern Indians [9].</td>
<td>2759492</td>
</tr>
<tr>
<td>1794427</td>
<td>Chlorogenic acid, (Antioxidant)</td>
<td>Gut</td>
<td>Variation in gut microbiome alters chlorogenic acid metabolism [10].</td>
<td>12771329</td>
</tr>
<tr>
<td>1983</td>
<td>Asetaminophen, (Analgesic and antipyretic)</td>
<td>Gut</td>
<td>Acetaminophen toxicity is associated with elevated levels of p-cresol produced by some bacterial communities [4].</td>
<td>19667173</td>
</tr>
<tr>
<td>9064</td>
<td>(+)-catechin and (-)-epicatechins, (Anti-oxidants)</td>
<td>Gut</td>
<td>In germ-free rats, (+)-catechins and (-)-epicatechins resulted in increase in the levels of liver CYP450 2C11, and (+) catechins caused elevation in the specific activity of liver UGT-Chloramphenicol [11].</td>
<td>12659723</td>
</tr>
<tr>
<td>5734</td>
<td>Zonisamide, (Anticonvulsant)</td>
<td>Gut</td>
<td>Gut microbiota reduce zonisamide into 2-sulfomoyacetylphenol, Levels of 2-sulfomoyacetylphenol reportedly increased upon re-inoculation of germ-free rats with gut microbiota [12].</td>
<td>9231340</td>
</tr>
</tbody>
</table>
new therapeutics. Maintaining animal colonies and performing genetic experiments requires careful record keeping. Many commercial or publicly available data management tools exist for animal recordkeeping, however they are rather complicated and may require technical expertise to install and maintain [1,2]. Consequently, biologists often use spreadsheets or database tools that are tied to a single computer, which limits accessibility by multiple users at different locations. Here, we have developed a user-friendly tool called ColonyTrak, designed for the biologist, to enable data management from any remote computer using a web browser.

Materials and methods: ColonyTrak was developed using a MySQL database, HTML (for GUI), CSS (for style sheets), PHP (server side scripting language), AJAX (for autofilling and auto increment features), and Javascript (for displaying alert messages or data validation) in a manner that is transparent to the end user. Registered users have administrator rights and, through a simple interface, are able to manage projects, users and read/write privileges (Figure 1). Users are able to enter a variety of animal data such as date of birth, genotype, lineage information, mating schemes from any location using a web browser. Several features such as auto-fill and auto-increment are implemented to make data entry easier.

Results and conclusions: ColonyTrak provides a secure multi-level user environment to enter, store and retrieve animal data. Importantly, the tool allows multiple users to simultaneously query the database with simple or complex queries and export the results in tabular format. Queries can be formatted by using a combination of pull-down menus to select the colony label, project label and/or gender, followed by specification of the dates corresponding to date of birth, date of death, mating date, or date of weaning (Figure 2). In addition, the user can define specific output variables. The query results are displayed in tabular html format or can be downloaded in an Excel format (Figure 3). Animals which are in mating cages are displayed in the first table (tan color) followed by animals in holding cages (blue color). The values in each html column can be sorted by clicking on the column headings.

We posit that ColonyTrak will enhance research productivity and the efficiency with which biologist manage experimental animal models. ColonyTrak is available for free at http://binf1.memphis.edu/ColonyTrak/.

Acknowledgments: This work was supported by the University of Memphis Bioinformatics Program and the Assisi Foundation of Memphis.

References

A12 Comparative studies of high-throughput biological graphs
Jonathan Reyes1,*, Charles Phillips2
1Graduate School of Genome Science & Technology, UT-ORNL, Oak Ridge, TN, 37830, USA; 2Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, Knoxville, TN, 37996, USA
E-mail: jreyes@utk.edu
BMC Bioinformatics 2011, 12(Suppl 7):A12

Background: The exponential growth of biological data has given rise to new and difficult challenges. Because large data is often dealt with, it is inefficient to infer from each individual characteristics of a given dataset. Bioinformaticists are developing quantitative techniques to analyze and interpret key data properties. Graph algorithms can provide powerful and intuitive insight on such properties [1]. Using this approach, we collect
biological data from transcriptomic and protein-protein interaction (PPI) sources. These data can be represented as a correlation matrix, where the rows are the vertices and the columns are the edges. We will analyze these graphs, and describe their differing structural characteristics.

Materials and methods: We are using a high throughput method for graphical exploration of genomic and proteomic data. Experimental datasets are extracted from the public databases Biomart and Gene Expression Omnibus (GEO) [2,3]. R [4] and MATLAB are used to develop algorithms that compute and compare various structural characteristics. We specifically developed an in-house script used to output essential histograms and unweighted/weighted edges. We are currently developing protocols to analyze the comparison of transcriptomes and PPI sources.

Acknowledgements: We express gratitude towards Jay Snoddy and Michael Langston for the ideas that led us to pursue this bioinformatics investigation.

References

A13
MSCTrees: a mean-shift based toolkit for cluster analysis of phylogenetic trees
Wei Li 1, Jerzy W Jaromczyk 1
1Department of Computer Science, University of Kentucky, Lexington, KY, 40506, USA; 2Department of Biology, University of Kentucky, Lexington, KY, 40506, USA

Background: Mean shift, an iterative technique for identifying the local maxima of a probability density function, has been successfully used as a clustering method in computer vision and image processing. We apply the mean shift technique to the high dimensional space of phylogeny trees. The basic idea behind this technique is to, given a set of sample points, shift each point in the direction of the gradient of the underlying density function in an iterative manner until the points converge to the natural maxima of the density function and form natural clusters [1]. We have developed software named MSCTrees based on a variant of the mean shift method, called the adaptive mean shift [2], to perform cluster analysis on a set of multidimensional data points corresponding to phylogenetic trees.

Methods: MSCTrees has two components: a C program called ms_cluster which implements a clustering algorithm based on the adaptive mean shift method, and a Perl script called cluster_trees.pl, which converts phylogenetic trees to multidimensional data points and calls ms_cluster to perform cluster analysis on the resulting points. The ms_cluster program, developed in C for optimized performance, takes a set of multidimensional data points as input, and outputs the clusters of the input points together with the cluster centers. The ms_cluster performs the following steps: 1) calculate the adaptive bandwidth for each data point using the k-Nearest Neighbor (ANN) method; 2) initialize a set of points using the values of the original data points; 3) shift the set of initialized points to new locations based on the mean shift vectors computed at each point; 4) repeat step (3) until all points have converged; 5) merge points that have converged to the same locations into clusters. Four auto-optimized (and user-definable) parameters have been implemented to control the mean shift clustering process.

The cluster_trees.pl script uses the BioPerl modules to parse a set of phylogenetic trees as the input. It maps a phylogenetic tree to a multidimensional data point by calculating the pair-wise distances between the leaves of the tree as the dimensional values of the resulting point. The script produces as output clusters of phylogenetic trees resulting from the clustering of their corresponding data points.

Results and conclusion: We tested MSCTrees with a well-known gopher-louse data set, which contains two sets of phylogenetic trees (101 trees each) for 15 species of gophers and 15 species of lice, respectively. Separate cluster analyses were performed on the two sets of trees, followed by a cluster analysis on the combined tree sets. A significantly reduced number of clusters was obtained from the combined data, which suggests similarity between the two tree distributions and is consistent with the known co-evolutionary relationship between gophers and lice. The pilot results demonstrate that the MSCTrees tool has strong potential for effective high dimensional cluster analysis of phylogenetic trees. We are also investigating other phylogenetic applications, such as identifying gene transfers via outlier detection.

Acknowledgements: This work was supported by NIH Research Project Grant Program (R01) from the Joint DMS/BIO/NIGMS Math/Bio Program under Grant No. 1R01GM086888-01 and the National Science Foundation under Grant No. 0814194.

References

A14
Next-gen sequencing of multi-drug resistant Acinetobacter baumannii at Nashville General Hospital at Meharry
Leon Dent 1, Nahed Ismail 2, Steven Robinson 1, Gary Rogers 3, Siddharth Piapat 4, Dana Marshall 5
1Surgery, Meharry Medical College, Nashville, TN, 37208, USA; 2Pathology, Meharry Medical College, Nashville, TN, 37208, USA; 3Student, Hampton University, Hampton, VA, 23668, USA; 4Electrical Engineering and Computer Science, University of Tennessee, Knoxville, TN, 37996, USA; 5Microarray & Bioinformatics Core, Microbiology & Immunology, Meharry Medical College, Nashville, TN, 37208, USA
E-mail: dmarshall@mmc.edu

BMC Bioinformatics 2011, 12(Suppl 7):A14

Background: Acinetobacter baumannii is a nonfermentative Gram-negative bacillus, which easily acquires antibiotic resistance determinants and causes life-threatening nosocomial infections [1]. Multi-drug resistant (MDR) strains are common therefore, empirical treatment choices are limited. More knowledge is needed regarding genetic diversity patterns and resistance phenotypes in a given clinical setting. Our goal is to identify the resistance genotypes of A. baumannii at Nashville General Hospital and correlate them with MDR phenotypes [1].

Materials and methods: A. baumannii isolate MMC#4 is sensitive to tobramycin with a possible extended-spectrum beta-lactamase phenotype. It was compared to baumannii reference strains using a next-gen sequencing methodology. Single-end sequencing was conducted on an Illumina Genome Analyzer II system at the Vanderbilt University Genome Technology Core (https://gtc.vanderbilt.edu/gtc/tech). Assembly was conducted at the Meharry Microarray and Bioinformatics Core using BowTie Aligner Software. Gene level annotation was conducted using Cufflinks software at the University Of Tennessee at Knoxville.

Results: Initial sequencing yielded 5,250,420 single end reads at 43bp each, totaling 225.76 Mb (Mega bases). The reads were aligned to six MDR baumannii reference strains and a fully drug susceptible strain (SDF). Of the 5.2 million total reads, 4.4 million (~85%) aligned to MDR baumannii strain ACICU with an average coverage depth of 43.96x fold. Gene level annotation using A. baumannii MDR strain AB0057 as a genome reference revealed sequence reads mapping to 3,209 genes or hypothetical ORFs of the ~3,800 total genes/ORFs in baumannii strain AB0057.

Conclusions: Strand-to-reference next-gen DNA sequencing of an MDR bacillus isolate showed roughly 58% coverage of the NCIBC genome by at least one sequence read and a depth of ~44X. Given that the genome size of A. baumannii ranges from 3.2Mb in strain STY (sensitive) to 3.9Mb in the MDR AYE strain, we are confident in the proper assembly of a significant portion of the genome. There are six complete assemblies of A. baumannii in the NCBI Genome Project data base, as well as ten “in progress”, allowing a true strain-to-reference approach utilizing the already assembled genomes as a scaffold for newly acquired sequences.
Although 100% assembly is not likely given the limitations of the short-read sequencing methodology, we would expect to have the majority of the isolate genome unambiguously mapped to a reference strain or assembled into contigs large enough to contribute to the genome databases. The information gained using this technology will lead to rapid and better diagnostics, guide empiric treatment and help people infected with this emerging pathogen.

Acknowledgements: This research was supported by The Meharry Translational Research Center (MeTRC) Grant Number U54RR026140-01 and The Research Centers in Minority Institutes (RCMI) Grant Number 5G12MD007584-15 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH). The research was also supported in part by Vanderbilt CTSA grant 1 UL1 RR024975 from the National Center for Research Resources, National Institutes of Health. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of NCRR or NIH.

References:

A15

Gene expression based prototype for automatic tumor prediction
Atiq Islam1, Khan M Iftekharuddin*1, Sloufsm E George3
1Ebay Applied Research, Ebay Inc., San Jose, CA 95125, USA; 2Department of Electrical and Computer Engineering, University of Memphis, Memphis, TN 38152, USA; 3Department of Mathematical Sciences, University of Memphis, Memphis, TN 38152 USA. E-mail: iftekhar@memphis.edu
BMC Bioinformatics 2011, 12(Suppl 7):A15

Background: Automatic detection of tumors is a challenging task due to the heterogeneous phenotypic and genotypic behaviors of cells within tumor types [1-3]. In recent years, a number of research endeavors have been reported in literatures that exploit microarray gene expression data to predict tissue/tumor types with high confidence [3-14]. However, in predicting tissue types, the above mentioned works neither explicitly consider correlation among the genes nor the probable subgroups within the known groups. In this work, our primary objective is to develop an automated prediction scheme for tumors based on DNA microarray gene expressions of tissue samples.

Material and methods: The workflow to build the tumor prototypes is shown in Fig. 1. Considering various sources of variation in array measures, we estimate tumor-specific gene expression measures using a two-way ANOVA model. Then, marker genes are identified using Wilcoxon [15] and Kruskal-Wallis [16] test. We then group the highly correlated marker genes together. Then, we obtain eigen-gene expression measures [10] from each individual gene. At the end of this step, we replace the gene expression measurements with eigen-gene values that conserve correlations among the strongly correlated genes. We then divide the tissue samples of known tumor types into subgroups. The CS measure [17] is exploited to obtain the optimal number of gene groups and tissue subgroups within each tissue type. The centroids of these subgroups of tissue samples represent the prototype of the corresponding tumor type. Finally, any new tissue sample is predicted as the tumor type of the closest centroid.

Results: To evaluate the proposed tumor prediction scheme, five different gene microarray datasets [3,5,7-9] are used, all of which were obtained using Affymetrix technology. We use leave-one-out cross validation method. Table 1 shows a summary of our experimental results for all the datasets. We provide relevant intermediate results along with the final classification accuracy. Finally, Table 2 shows the performance comparison between our proposed prediction scheme and the methods discussed in original works [3,5,7-9] wherein the corresponding datasets are published. We also compare our classification accuracies with those of a Supervised Clustering method [4] for completeness.

Conclusions: In this work, we propose a novel, seamless, and integrated technique of automatic tumor detection using Affymetrix microarray gene expression data. We appropriately normalize the data by estimating tumor-specific gene expression measures using an ANOVA model. Furthermore, our novel tumor prediction scheme explores molecular information such as probable correlations among genes and probable unknown subgroups within known tumor types. We demonstrate the efficacy of our proposed scheme using five different Affymetrix gene expression datasets.

Acknowledgements: The research in this paper is supported in part through research grants (RG-01-0125, TG-04-0026) provided by the Whitaker Foundation with Khan M. Iftekharuddin as the principal investigator.

References:
Table 1 (abstract A15) Experimental results with different dataset

<table>
<thead>
<tr>
<th>Dataset</th>
<th>No. of Samples</th>
<th>No. of Gene in each chip</th>
<th>No. of Marker genes with q-value &lt; 0.05</th>
<th>No. of eigen-gene expression</th>
<th>No. of tissue subgroups</th>
<th>Classification Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brain Tumor: A [3]</td>
<td>42</td>
<td>6,817</td>
<td>1179</td>
<td>150</td>
<td>Medullo: 5</td>
<td>92%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Gioma: 5</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>AT/RTs: 5</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Normal: 2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>PNET: 3</td>
<td></td>
</tr>
<tr>
<td>Brain Tumor: B [3]</td>
<td>34</td>
<td>6,817</td>
<td>29</td>
<td>11</td>
<td>Classic: 5</td>
<td>97%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Desmoplastic: 3</td>
<td></td>
</tr>
<tr>
<td>Brain Tumor: C [3]</td>
<td>60</td>
<td>6,817</td>
<td>550</td>
<td>88</td>
<td>Survivor: 5</td>
<td>98%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Deceased: 4</td>
<td></td>
</tr>
<tr>
<td>Colon Cancer [5]</td>
<td>62</td>
<td>6,500</td>
<td>104</td>
<td>37</td>
<td>Normal: 7</td>
<td>97%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Tumor: 9</td>
<td></td>
</tr>
<tr>
<td>Prostate Cancer [9]</td>
<td>102</td>
<td>12,600</td>
<td>410</td>
<td>76</td>
<td>Normal: 5</td>
<td>99%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Tumor: 9</td>
<td></td>
</tr>
<tr>
<td>Leukemia [7]</td>
<td>72</td>
<td>7,129</td>
<td>60</td>
<td>20</td>
<td>All: 7</td>
<td>99%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>AML: 5</td>
<td></td>
</tr>
<tr>
<td>Breast Cancer [8]</td>
<td>38</td>
<td>7,129</td>
<td>109</td>
<td>38</td>
<td>ER +: 9</td>
<td>97%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ER -: 20</td>
<td></td>
</tr>
</tbody>
</table>

Table 2 (abstract A15) Comparison of methods

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Original works</td>
<td>83%</td>
<td>97%</td>
<td>78%</td>
<td>90%</td>
<td>90%</td>
<td>N/A</td>
<td>95%</td>
</tr>
<tr>
<td>Supervised Clustering [4]</td>
<td>88%</td>
<td>N/A</td>
<td>N/A</td>
<td>84%</td>
<td>95%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Our Method</td>
<td>92%</td>
<td>97%</td>
<td>98%</td>
<td>97%</td>
<td>99%</td>
<td>99%</td>
<td>97%</td>
</tr>
</tbody>
</table>


A16
categoryCompare: high-throughput data meta-analysis using gene annotations
Robert M Flight1, Jeffrey C Petruska1, Benjamin J Harrison1, Eric C Rouchka1*
1Department of Anatomical Sciences and Neurobiology, University of Louisville, Louisville, KY, 40292, USA; 2Department of Computer Engineering and Computer Science, University of Louisville, Louisville, KY, 40292, USA
E-mail: ericrouckha@louisville.edu
BMC Bioinformatics 2011, 12(Suppl 7):A16

Background: Many current DNA microarray and other high-throughput data meta-analysis studies concentrate on deriving a concordant list of genes across many experiments to discover the true genes responsible for a particular disease process or biological pathway or cellular response (Figure 1A). However, by concentrating on the genes in common, similarities or differences that exist at a pathway or process level may be missed. Results: We describe a meta-analysis approach that allows comparison and contrast of gene lists at the level of categorical annotation (pathway or Gene Ontology annotations). This categorical evaluation compares enriched annotations between gene lists (Figure 1B), and displays the results graphically to allow intuitive visualization and exploration of the similarities and differences. False discovery correction via simulation is implemented to control for the effect of different sized gene lists as inputs.

Conclusions: The approach was tested using two gene lists, genes involved in the response to denervation in muscle (a literature compendium), and in skin (experimentally determined). Using the categorical comparison highlights known biological processes that are
common in the two cases, while also allowing one to easily see areas of difference that are not apparent from examining the gene lists alone.

**Availability:** categoryCompare is available as a Bioconductor package, and a web interface (using RApache) has also been developed to facilitate use in the wider research community.

**Acknowledgements:** The authors gratefully acknowledge the Christopher and Dana Reeve Foundation, KY Spinal Cord and Head Injury Research Trust, the Paralyzed Veterans of America, and NIH Grants: P20RR016481-10, P20RR016481-09S1, P30ES014443-04 for funding.

---

**A17**

**Bioinformatic challenges for proteomic biomarkers of cancer**

David L Tabb1,*, Daniel C Liebler2

1Biomedical Informatics, Vanderbilt University, Nashville, TN, 37232-8575, USA; 2Biochemistry, Vanderbilt University, Nashville, TN, 37232-6350, USA

E-mail: david.l.tabb@vanderbilt.edu

*BMC Bioinformatics 2011, 12(Suppl 7):A17

**Background:** Proteomic biomarkers are sets of proteins that may be detected in biofluids for disease detection, prognosis, and treatment selection. Vanderbilt University is taking part in two National Cancer Institute initiatives to establish a basis for clinical biomarkers. The first, Clinical Proteomic Technology Assessment for Cancer (CPTAC), characterizes proteomic technologies for discovery and validation of biomarkers. The second, the Early Detection Research Network (EDRN), attempts to develop new sets of biomarkers and evaluate their effectiveness in clinically relevant samples. The quest for biomarkers is fraught with challenges for both “bench” experiments and bioinformatics. Here, we examine sources of discrepancy at several levels. The search engines that identify peptides from LC-MS/MS experiments differ significantly in the set of spectra that they identify from experiments. Parsimonious protein assembly may prune out proteins in one experiment and retain them in another. Protein differentiation tools may yield divergent results, even when starting from the same data sets. Verifying biomarkers through targeted proteomics has only recently been supported by tools that can work across instruments from different vendors. Translating cancer biology knowledge into clinically relevant tests for this disease will require care at all these levels.

**Figure 1(abstract A16)**

A - Usual method of high-throughput experiment meta-analysis comparing gene lists (L1 and L2) directly. B - categoryCompare compares the gene lists on the basis of enriched annotations.

---

**A18**

**Integration of bioinformatics tools in candidate gene prioritization of co-regulated gene sets in Saccharomyces cerevisiae**

Vida Abedi1,*, Mohammed Yeasin1,2, Thomas R Sutter1,3

1Department of Electrical and Computer Engineering, University of Memphis, Memphis, TN, 38152, USA; 2Bioinformatics Program, College of Arts and Science; University of Memphis, Memphis, TN, 38152, USA; 3Department of Biological Science, University of Memphis, Memphis, TN, 38152, USA

E-mail: vidaabedi@hotmail.com

*BMC Bioinformatics 2011, 12(Suppl 7):A18

The availability of massive amounts of heterogeneous and distributed biological data has prompted the development of a wide range of data analysis and data mining tools in the area of bioinformatics. However, due to the nature of the biological data, performing a specific analysis by combining such tools can be complicated and cumbersome. Yet, integration of number of tools can provide complementary information.

**Figure 1(abstract A18)**

Experimental design of the candidate gene prioritization process. Information filtering is organized in three levels. At level 1, all the 142 genes are considered in the analysis. At the end of level 1, three sets of genes are obtained: 1) genes that are part of relevant GO categories, 2) genes for which there is significant amount of literature, and 3) genes that are part of enriched GO categories. All uncharacterized genes from the three lists are extracted and passed to the second level of prioritization. In addition, genes with at least two supporting evidences will also be forwarded to the second level of exploration. The filtered gene set from level 2 is used as input in level 3, where physical and genetic interaction among these genes are further explored. The resulting sets of genes will be the uncharacterized genes and the genes with at least two supporting evidences, which are then prioritized further if they are interrelated with at least one physical or genetic interaction.
and improve the efficiency of the data analysis to further our understanding and knowledge discovery. The development of an integrated software platform can considerably enhance the usability of such tools and benefits the research communities at large. Towards that goal, this study focuses on systematically integrating a number of tools for analyzing Saccharomyces cerevisiae data in order to improve candidate gene prioritization from microarray data using evidences from complementary sources.

Microarray data from a recent study by Ouyang et al.[1] was used to evaluate the proposed framework. An array of free and open source bioinformatics tools were used to develop the Saccharomyces Integrated Software Platform (SISP). In particular, sources of information used in this analysis include literature data, Gene Ontology, physical and genetic interaction data as well as pathway information. SISP has the strength of combining prior knowledge with user-defined weighting of different sources of evidence. Access to the integrated tool will be facilitated by a user-friendly web interface with options including data query, import, export, analysis and visualization.

The set of 142 genes from the microarray experiment was systematically reduced to sixteen genes (Figure 1); four out of the sixteen genes were highly ranked based on various sources of information. The sixteen genes were part of thirteen inter-related pathways, with eight genes playing major roles in those pathways. This integrated analysis exploited key features of essential information and the identification of key inter-related pathways and genes. Integration of bioinformatics tools allows merging complementary sources of information which are critical to the identification of candidate genes for further experimental validation.

Reference

A19
Statistical analysis of microarray gene expression data from a mouse model of toxoplasmosis
Shrikant Pawar1,2, Cheryl D Davis1,2, Claire A Rinehart1,2
1Department of Biology, Western Kentucky University, Bowling Green, KY 42101, USA; 2Bioinformatics and Information Science Center, Western Kentucky University, Bowling Green, KY 42101, USA. E-mail: pawar1550@gmail.com
BMC Bioinformatics 2011, 12(Suppl 7):A19

Background: Toxoplasmosis, caused by the protozoan parasite Toxoplasma gondii is a major cause of morbidity and mortality in patients with AIDS and an important cause of miscarriage, stillbirth and congenital disease in newborns. Previous studies have provided evidence that dietary supplementation with vitamin E and selenium is harmful during experimental toxoplasmosis in mice, whereas a diet deficient in vitamin E and selenium results in decreased numbers of tissue cysts in the brain and dramatically reduced brain pathology. The overall goal of the present study was to determine the impact of dietary supplementation with antioxidants on gene expression in the brains of non-infected mice and in mice infected with T. gondii using microarray analysis. RNA was isolated from the brains of C57BL/6 mice, and an Agilent Oligo Whole Mouse Genome Microarray (Agilent Technologies, Inc.) was performed. A total of 48 chips were normalized by Z ratios and the Data Driven Harr Fisch Normalization methods. Differentially expressed genes were identified by applying thresholds to identify significant values and the results were compared between the normalization methods. These differentially expressed genes and their respective fold change ratios were used in Ingenuity Pathway Analysis (IPA) software to analyze the pathways involved with these genes.

Acknowledgements: Support from the National Center for Research Resources NIH Grant Number 2 P20 RR-16481 and from the WKU Bioinformatics and Information Science Center is gratefully acknowledged.

A20
Meta-analysis of gene expression changes in response to radiation exposure
1Department of Computer Engineering and Computer Science, University of Louisville, Louisville, KY, 40292, USA; 2Department of Anatomical Sciences and Neurobiology, University of Louisville, Louisville, KY, 40292, USA; 3Department of Bioengineering, University of Louisville, Louisville, KY, 40292, USA; 4Department of Medicine, University of Louisville, Louisville, KY, 40292, USA. E-mail: eric.rouchka@louisville.edu
BMC Bioinformatics 2011, 12(Suppl 7):A20

Background: Given NASA's recent focus on long-duration space travel, potential adverse effects on astronauts of ionizing radiation need to be minimized. Levels of exposure for astronauts on such flights can be high enough to cause damage to DNA, possibly causing mutation and cancer. By analyzing the mRNA expression levels of genes exposed to different doses of gamma radiation in laboratory experiments, it can be determined which genes act as biomarkers of dosage-specific radiation exposure which can then be used on lab-on-a-chip (LOC) diagnostic platforms for early detection of radiation exposure.

Methods: Currently, our group is incorporating three genes known to be involved in double-stranded DNA damage and repair, including: p21, p53, and H2AX. While these genes should show up in response to ionizing radiation exposure, they are not ideal biomarkers due to their lack of specificity. Therefore, by using meta-analysis, our goal is to gain further insight into additional biomarkers for radiation exposure. More than 40 publicly-available data sets appropriate to our study were obtained from the online repository, Gene Expression Omnibus (GEO) [1]. These gene expression experiments were subsequently categorized by radiation type and dose. Meta-analysis was performed using a combination of statistical tests using the Differential Expression via Distance Synthesis (DEDS) package [2].

Results: Preliminary meta-analysis of these publicly available datasets yields potential biomarkers from the P53 signaling pathway (CDKN1A, GADD45A, MDM2, PMAIP1), stress response transcription factors (ATF3, JUN, JUNB, JUND), and cell surface receptors (CD69, CD70, CD83). Additional microarray experiments involving irradiated blood samples are underway. The analysis of both publicly available data and our own datasets will yield a broader picture of genes most sensitive to exposure of ionizing radiation for use as biomarkers on LOC diagnostic platforms for early detection of radiation exposure, leading to subsequent treatment.

Acknowledgements: Members of the University of Louisville Research Group for Diagnosing and Mitigating Human Exposure to Radiation Using Micro-Nanotechnology include Robert W. Cohn, John Eaton, William D. Ehringer, Andre M. Gobin, Andrea S. Gobin, Balaji Panchapakesan, Eric C. Rouchka, Palaniappan Sethu, and Robert S. Keayton. This work was supported by the National Aeronautics and Space Administration (NNX10A36G) and the National Institutes of Health (NIH) (P20RR016481, P30ES01443). Its content is solely the responsibility of the authors and does not necessarily represent the official views of NASA, NCRR, NIAMS, or NIH.

References