SysBioCube: A Data Warehouse and Integrative Data Analysis Platform Facilitating Systems Biology Studies of Disorders of Military Relevance

Sudhir Chowbina[1], Rasha Hammamieh[2], Raina Kumar[1], Nabarun Chakraborty[2], Ruoting Yang[1], Uma Mudunuri[1], Marti Jett[2], Joseph M. Palma[3], Robert Stephens[1]
[1]Advanced Biomedical Computing Center, Frederick National Laboratory for Cancer Research/SAIC-Frederick Inc., Frederick, MD
[2]US Army Center for Environmental Health Research (USACEHR), Frederick, MD
[3]USA Medical Research and Materiel Command, Fort Detrick, MD

Abstract

SysBioCube is an integrated data warehouse and analysis platform for experimental data relating to diseases of military relevance developed for the US Army Medical Research and Materiel Command Systems Biology Enterprise (SBE). It brings together, under a single database environment, pathophysio-, psychological, molecular and biochemical data from mouse models of post-traumatic stress disorder and (pre-) clinical data from human PTSD patients.. SysBioCube will organize, centralize and normalize this data and provide an access portal for subsequent analysis to the SBE. It provides new or expanded browsing, querying and visualization to provide better understanding of the systems biology of PTSD, all brought about through the integrated environment. We employ Oracle database technology to store the data using an integrated hierarchical database schema design. The web interface provides researchers with systematic information and option to interrogate the profiles of pan-omics component across different data types, experimental designs and other covariates.

Introduction

Growing amounts of disparate and complex molecular biology and pathophysio- psychological data necessitate the development of a holistic framework to integrate and analyze these data. A systems biology approach is needed to integrate multiple data types, including the clinical and pre-clinical findings, various –omic results, and modeling to identify the relevant biological network of diseases and diagnostic, prognostic and therapeutic markers. For a growing research program generating high throughput molecular profiling and *in silico* data from human & rodent model systems, we have devised a single integrated web platform to facilitate data storage, sharing and analysis.

The diseases of military relevance are the injuries caused by environmental extremes and infectious diseases that commonly occur in combat situation. The long-term repercussions of those impacts are of considerable interest. Such diseases include infections, coagulopathy, heat stroke, traumatic brain injury and, post-traumatic stress disorder (PTSD).

Life threatening trauma that often comes in a repetitive manner complicates the disease management process in military community attributing certain aspects distinct from the civilian community's disease profile. Therefore a meaningful consideration of the military relevant diseases is imperative. A data repository and analysis system that encompasses the majority of the experimental and clinical data of human and rodent models and provides a system-wide view of the data will enhance collaboration and hypothesis generation in this study of disorders of military relevance.

Method

Data integration: Data collected was in different formats and Python parsers were developed to convert them into a tab-delimited text format suitable to be uploaded into an Oracle database. Data compatibility was ensured by using a standard data extraction, transformation, and loading (ETL) process characteristic of data warehousing-based data integration approaches. Staging tables were used to store the initial pre-processed and clean data before the final deployment tables.

Database creation: A database schema was built by establishing a systematic navigation among different datasets. The data in the warehouse can broadly be classified into clinical information, experimental data and general annotation data. Clinical information contains donor demographics and disease evaluation reports. Experimental data includes comprehensive pan-omics outputs, pathophysio- and psychological results and brain imaging data.

The relevant information from experimental data is extracted into appropriate tables meant to hold data types such as EXPERIMENTAL METADATA, NORMALIZED DATA and ANALYZED DATA. Separate tables are used to hold DATA ANNOTATIONS and ANNOTATION METADATA. There are various in-house databases for pathway, gene ontology, protein-protein interactions and other database used to annotate and enrich experiment datasets to glean more associative biological information.

In addition, we assign a combined code for experiments carried out for a specific assay or measurement. To achieve mapping between experimental data and metadata, these codes are then mapped to specific study samples along with the body-part/tissue from which the study material was extracted.

The structure permits integrated queries across the database tables. For example, one can query the microarray and the patient co-variate information to retrieve differentially-expressed genes correlated with a particular phenotype.

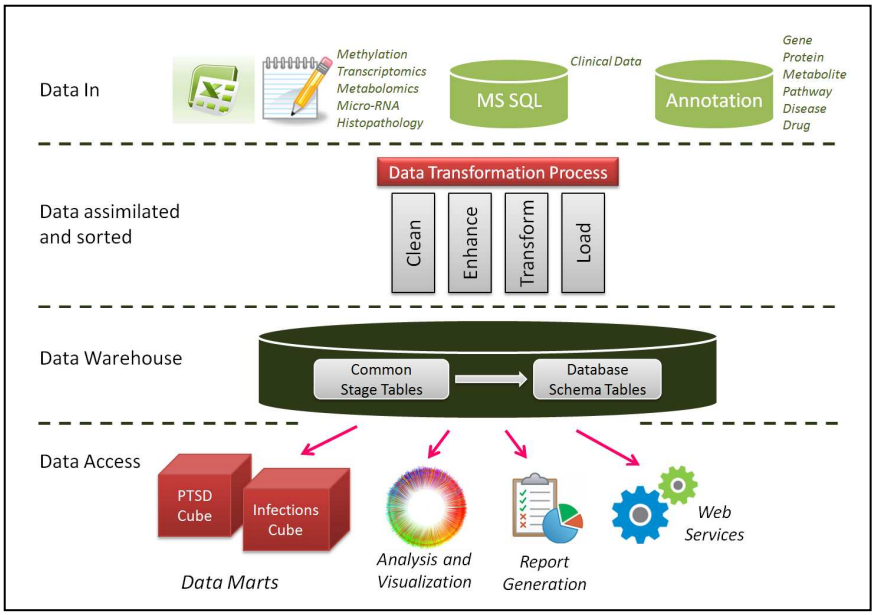

Figure 1. SysBioCube software application architecture

Results and Discussion

The current version of SysBioCube web interface offers three main features for users to (a) browse multiple data types, (b) query/analyze using visual data mining tools, and (c) explore user-defined associations across two different data types.

(a) Browse option

The users can browse multiple data types (Figure 2) such as mouse behavioral studies, neurohistology, histopathology, transcriptomics, epigenomics, metabolomics and physiological information.

Currently, we have gene expression microarray and DNA methylation data in GEO-formatted spreadsheet format with metadata and processed data matrix in two separate sheets in a single excel file. The metadata sheet contains minimum information (such as sample annotation, experiment design, annotation of the arrays and analysis methods) about the experiments.

We understand the need for adopting data standards and ontologies. This is a work in progress and we will move towards adopting common data elements and standard terminologies.

(b) Visual data mining tools

In addition to browsing the above data sets as table views, the users can explore them using visual data mining tools (Figure 3). Tools provided include interactive gene and methylation profiles, interactive heatmaps, cytoscape network views, integrative genomics viewer (IGV), and protein-protein interaction matrices.

The interactive gene expression profile (Figure 3A) was developed using Highcharts javascript library (http://www.highcharts.com/). The profile gives the users an overall view of a single gene expression from different tissues collected across the experiment groups in a single chart. The website also provides an option to include multiple genes.

Integrative Genomics Viewer (IGV)[1], is a high-performance desktop tool for interactive visual exploration of different large-scale genomic and clinical data. It offers dynamic interaction at all scales of genome resolution, from whole genome to base pairs. Visualizing the PTSD human gene expression data along with clinical co-variate annotation allows users to perform real-time sorting and to discover predictive co-variates differentiating PTSD positive and negative patients based on relevant gene expression (Figure 3B).

Cytoscape Web is a library that helps visualize dynamic interactions generated using user defined criteria, as graphs with annotated nodes and edges. The library leverages HTML5 technologies to provide a standards-compliant, consistent cross-browser method of displaying interactive graphs[2]. The protein interaction network (Figure 3C) is constructed using the gene of interest overlaid with proteomics data. The annotation information is obtained from the Biological General Repository for Interaction Datasets (BioGRID) which contains curated physical and genetic interactions from all major model organisms[3].

The protein-protein interaction matrix (Figure 3D) attempts to capture the protein interactions derived from the genes that match the user-defined parameters. The matrix is generated using genes with expression fold-change greater than 1.2 in plasma of the user-specified PTSD mouse group. The matrix is also ordered to show the proteins having the most interacting proteins at the top of the matrix. Thus, the matrix provides co-expressed protein modules in a phenotype.

The interactive heatmap (Figure 3E) was implemented with canvasXpress (http://www.canvasxpress.org). This is a javascript library based on the <canvas> tag implemented in HTML5. The heatmaps allow dynamic zooming, tooltip, and a pop-up table when clicked on a cell in the heatmap. The heatmap allows users to visualize genomics/transcriptomics data from PTSD human datasets alongside with the patient phenotype annotation.

(c) Associations across different data types

Integrated approaches that combine genome, transcriptome, proteome, epigenome and metabolome profiling have become important as they provide better understanding of the biological systems[4-6]. An effective way to interpret complex omics experiments is the combined visualization of the experimental data with existing knowledge. Networks and charts are effective tools to support human reasoning, and when different data-types are presented in the context of the biological pathways much information can be gained about the biological mechanisms.

For instance, we implement an approach to find associations between metabolomics and transcriptomics data (Figure 4) by visualizing the datasets as a network. The user can filter and/or sort the database based on particular experiment group, tissue types and regulation profiles such as fold change and p-value. A network is then derived from the metabolites and genes that match the user input criteria. The edge information of the network (or the link between a metabolite and gene) is derived from the Edinburgh Human Metabolic Network (EHMN) pathways[7].

Future Directions

In future, we plan to: (a) incorporate human and rodent models data related to trauma-related coagulopathy and heat stroke into SysBioCube; (b) develop predictive computational models that offer better diagnostic, therapeutic, preventive strategies, and informed decision making; (c) enhance data visualization techniques to visualize correlation across data-types, multi-species comparative genome data and probe complex biological networks across the species.

Conclusion

We have developed SysBioCube, an integrative knowledge base that will help investigators access, visualize and analyze comprehensive information about the military-relevant diseases.. When fully developed, it will be a one-stop portal for accessing and mining data generated from army research centers, and will help establish links between molecular, biological and clinical data.

Software Availability

The software will not be made available as a whole, partly because we are using open-source tools for analysis and visualization. However, we plan to have a public-accessible version of the website with access to limited PTSD data sets or as made available by investigators.

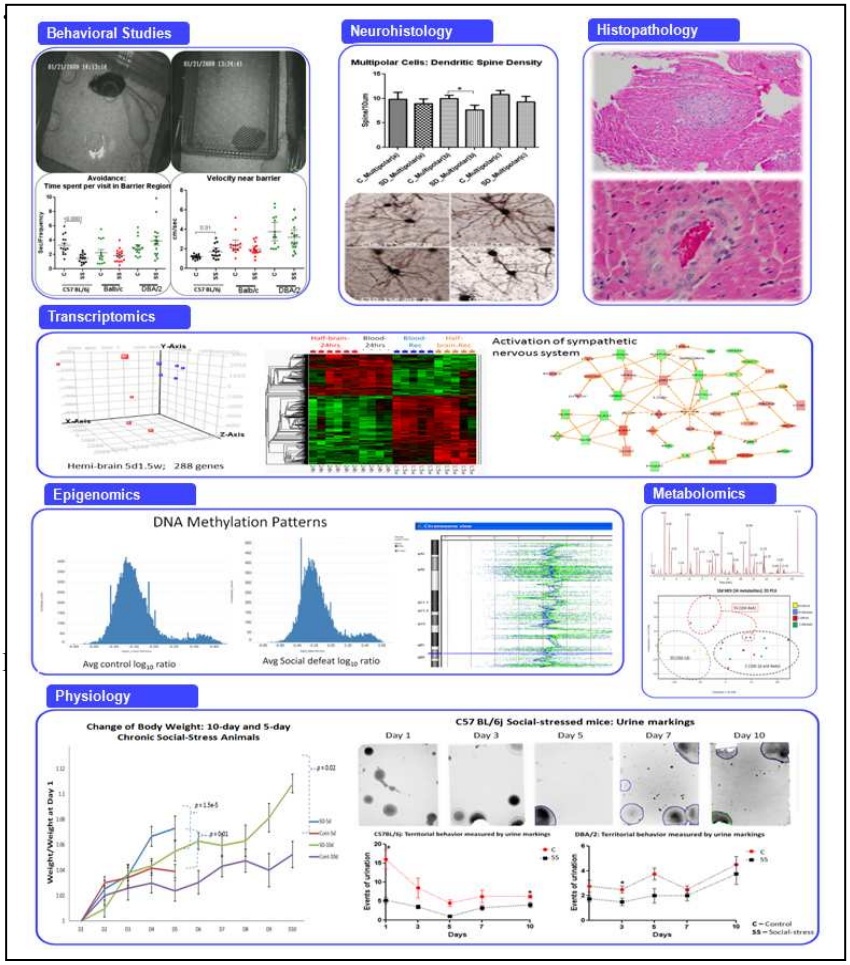

Figure 2. Browse multiple data-types.

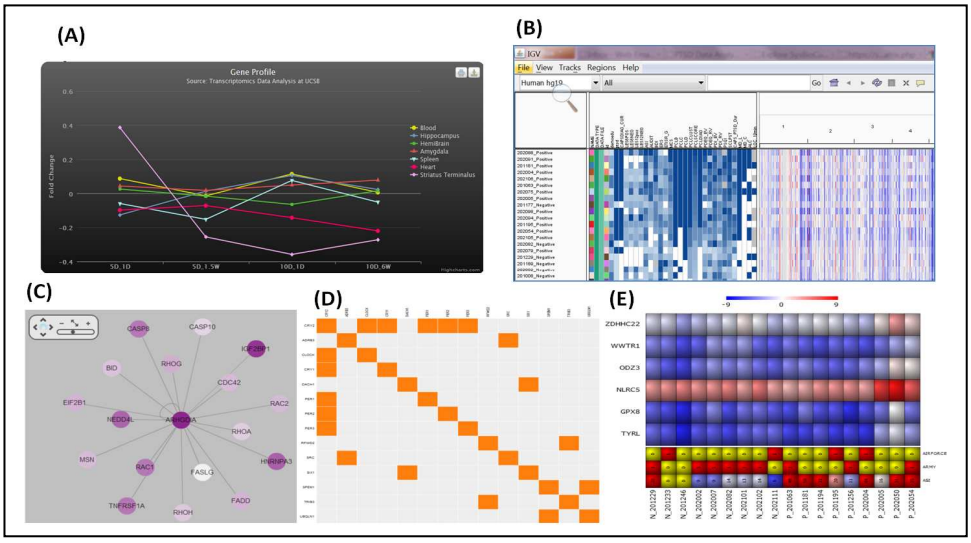

Figure 3. Visual data mining tools. (A) Interactive gene expression profile; (B) Integrative Genomics Viewer; (C) Protein interaction network using Cytoscape Web plugin; (D) Protein-protein interaction matrix; (E) Interactive heatmap to visualize gene expression and clinical annotations.

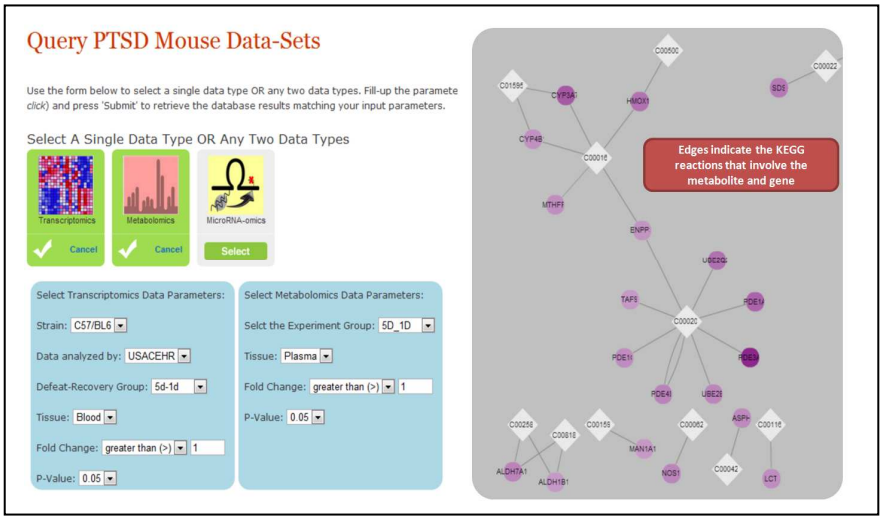

Figure 4. Query of multiple data types; for example, to determine network associations between Transcriptomics and Metabolomics.

References

1. Robinson JT, Thorvaldsdottir H, Winckler W, et al. Integrative genomics viewer. Nat Biotechnol. 2011; 29(1):24-6.
2. Lopes CT, Franz M, Kazi F, et al. Cytoscape Web: an interactive web-based network browser. Bioinformatics. 2010; 26(18):2347-8.
3. Stark C, Breitkreutz BJ, Chatr-Aryamontri A, et al. The BioGRID Interaction Database: 2011 update. Nucleic Acids Res. 2011; 39(Database issue):D698-704.
4. Cho K, Shibato J, Agrawal GK, et al. Integrated transcriptomics, proteomics, and metabolomics analyses to survey ozone responses in the leaves of rice seedling. J Proteome Res. 2008; 7(7):2980-98.
5. Garcia-Alcalde F, Garcia-Lopez F, Dopazo J, et al. Paintomics: a web based tool for the joint visualization of transcriptomics and metabolomics data. Bioinformatics. 2011; 27(1):137-9.
6. Xie L, Weichel B, Ohm JE, et al. An integrative analysis of DNA methylation and RNA-Seq data for human heart, kidney and liver. BMC Syst Biol. 2011; 5 Suppl 3:S4.
7. Hao T, Ma HW, Zhao XM, et al. Compartmentalization of the Edinburgh Human Metabolic Network. BMC Bioinformatics. 2010; 11:393.