CIDMS: Cardiac Integrated Database Management System for Cardiac Systems Biology

Pradeep K. Sreenivasaiah, Bharathi Mohanraju, Do Han Kim*
Department of Life Science, Gwangju Institute of Science and Technology (GIST) Gwangju, Korea
*email: dhkim@gist.ac.kr

Introduction

A major goal of systems biology is to produce comprehensive models that can predict cellular functions from numerous experimental data such as gene expression profiles, protein-protein interactions (PPI) and profiling of metabolites. It is estimated approximately 20,000 to 27,000 genes are expressed in cardiovascular system[1]. The number will be much higher for proteome, as they are subjected to alternative mRNA splicing, RNA editing and post translational modifications[2]. Abundant amount of data describing function, expression, regulation, interactions for many of the known genes/proteins in heart under physiological and pathophysiological conditions already exist and more will be accumulated[3]. Although excessive data nourishes systems biological approaches, it becomes harder to use as the amount grows. Problems have arisen because data are not easily accessible; they are distributed across heterogeneous resources, fragmented, redundant, inconsistent and incomprehensive, thus not suitable for producing quality models.

Although a number of large public databases are being operated to facilitate management of the rapidly expanding data, most of them are specialized for catering one type of data[5]. The presently available databases provide very minimal information about heart relevant genes and proteins and, none of the databases provides an integrated view of qualitative and quantitative (i.e. kinetic data, models) data sets relevant to heart and thus they are of limited use for cardiac systems biology research[3, 6].

Results

To alleviate the above problems and to provide bioinformatical resource that is helpful for research in cardiac systems biology, we have developed Cardiac Integrated Database Management System (CIDMS). CIDMS is a web based system that facilitates easy storage, retrieval and exploration of heart relevant information. Currently CIDMS holds mammalian (man, mouse, rat) information for >8000 heart relevant and >70 heart specific genes/proteins, >120 disease associations, >60 pathways and several mathematical models. All these information has been curated from publications, mined from public databases, obtained by contributions from the community and collaborators, and also from our own high throughput experiments. We have enlisted axioms that are used as guide lines to acquire, annotate and categorize information hosted on CIDMS. See Figure 1 for details of CIDMS architecture.

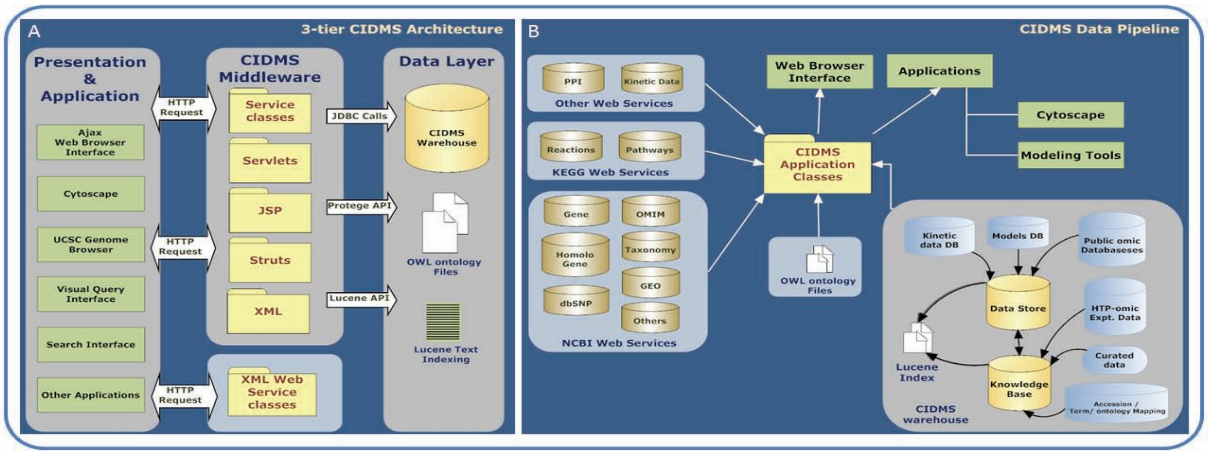

Figure 1: (**A**) 3-tier architecture of CIDMS based on J2EE technology, designed for leverage of efficient access to distributed information. Used open source packages include; Sun Java Enterprise Server, Struts library, Ajax, Java 1.5 jdk, dhtmlx, Lucene, XML, Apache axis, MySql etc. (**B**) CIDMS dataflow. By combining the data warehouse concept and asynchronous Ajax web service calls, up-to-the-minute information is made available. Using accessions, ontology (GO, Disease, Pathway) and custom curated keyword information from >20 databases are integrated and presented in well organized and concise formats. A variety of display options makes the database easily accessible via a web interface.

Features

- Integration of heterogeneous data.
- Access to up-to-the-minute information as a result of combining the data warehouse concept and web service calls using Ajax (Figure 1).
- Multiple data integration approaches, which add flexibility allowing incorporating of more number of datasets (Figure 2).
- Value added domain specific annotations from the source datasets.
- Powerful text searches including Boolean operators using Lucene library[7]
- Modular architecture that can be scaled to incorporate additional datasets, web service accesses, integration methods, external application integration etc.
- GO browser interface implemented as dynamic tree provides easy navigation to the terms and easy access to heart relevant proteins that are annotated to each GO terms. Further use of protégé API in its implementation, gives opportunity to employ Description Logic Reasoners (inference based search engine). (Future release).

Significance

- CIDMS serves as a foundation for hypothesis driven systems biological research and for uncovering previously unsuspected relationships among cardiac genes, their involvement in different mechanisms of physiological and pathophysiological processes. We are committed to make CIDMS richer in terms of content and features.

Availability

CIDMS can be accessed freely at http://203.237.50.48/CIDMS.

References

1. Dempsey, A. A., V. J. Dzau, et al. (2001). "Cardiovascular genomics: estimating the total number of genes expressed in the human cardiovascular system." J Mol Cell Cardiol 33(10): 1879-86
2. McGregor, E. and M. J. Dunn (2006). "Proteomics of the heart: unraveling disease." Circ Res 98(3): 309-21.
3. Winslow, R. L. and M. S. Boguski (2003). "Genome informatics: current status and future prospects." Circ Res 92(9): 953-61.
4. Safran, M., I. Solomon, et al. (2002). "GeneCards 2002: towards a complete, object-oriented, human gene compendium." Bioinformatics 18(11): 1542-3.
5. The Molecular Biology Database Collection: 2006 update
6. Le Novere, N. (2006). "Model Storage, exchange and integration." BMC Neurosci 7 Suppl 1: S11.
7. http://lucene.apache.org/java/docs/

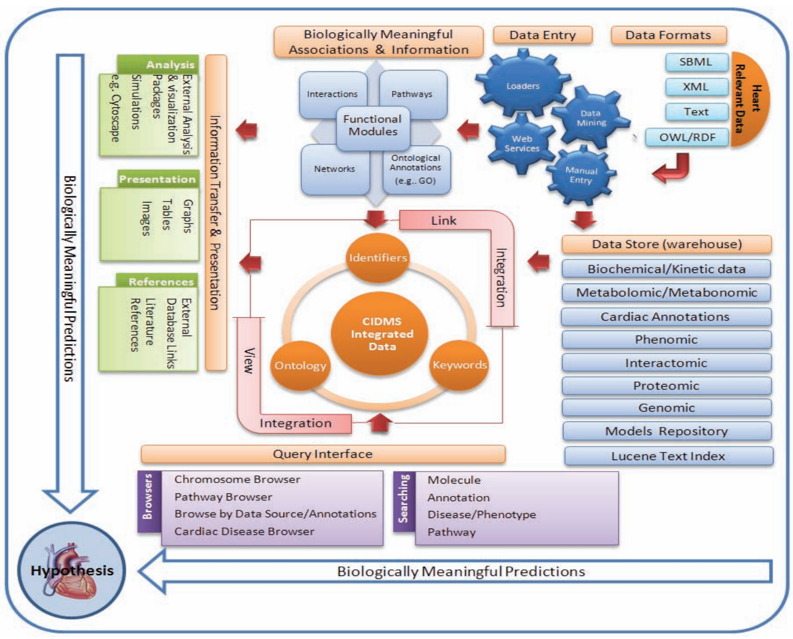

Figure 2: The modules of CIDMS that are relevant to the user (system biologist) perspective. Each module has well defined functionality and the arrows (red) represent the flow of data and information. CIDMS employs three integration strategies. Link integration is the first integration strategy by providing extensive cross-references to external databases. Second strategy for integration is data warehousing. Data from several public sources are loaded into a unified data model. View integration is the final strategy. At this level queries are formulated to fetch information from the warehouse as well as by several external databases by web services calls. Data is fetched, integrated and returned to user in concise textual and graphical formats. Also this information can be transferred to applications such as cytoscape for advanced analysis and visualization. CIDMS currently supports browsing of genes on a particular chromosome, phenotypes, pathways, models and by source annotations. Some of the browser modules have graphical output interface. For examples Chromosome Location browser displays each chromosome along with locations of individual feature in a genome. This map can be interactively explored to see gene details on a particular location. CIDMS also supports searching by accession numbers and text with Boolean operators. Searches can be made for a molecule, pathway, disease, categories, ontology, source annotations, models etc. This integrated availability of information will help in predictions and ultimately framing biologically meaningful hypothesis.