MiMI: Michigan Molecular Interactions

Adriane Chapman, Magesh Jayapandian, Cong Yu, H.V. Jagadish
Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, USA

ABSTRACT

There is a proliferation of data sources in biology. A complete understanding of a biological problem often requires the integration of multiple data sources, each providing insights on certain aspects of the problem. Furthermore, different sources often represent data in different ways, even when they cover the same information. Researchers interested in a particular biological problem are forced to search for and understand multiple, often conflicting sources, and piece the jigsaw puzzle of information together for themselves.

The Michigan Molecular Interactions Database (MiMI) attempts to relieve scientists of this burden (MiMI, 2005). By integrating popular, well-known datasets, MiMI combines all the power of each individual dataset, like BIND (Bader *et al*., 2003), and multiplies their benefits to individual researchers by merging them with other known facts from diverse datasets. By such integration, MiMI shows scientists when facts are corroborated by similar facts from different datasets, and when facts are contradicted among datasets. Moreover, the provenance of each data item has been annotated, allowing scientists to view information from only the sources they trust. The following is a brief description of some of the underlying concepts found in MiMI, as well as a detailed list of datasets used to generate it.

It is imperative for data to be recorded even when the confidence in it is low. As a result, many interaction databases place an interaction in the public domain even if it is supported by only one experiment. However, this forces scientists to search through multiple databases for conflicting or corroborating evidence. For instance, protein interaction information can be generated using yeast-2-hybrid, immunoprecipitation, individual experiments, as well as many others, each having a different degree of accuracy (Bader *et al*., 2002; von-Mering *et al*., 2002). MiMI helps scientists with this evidence-searching task by integrating all information from participating data sources through the process of *deep merging*. As a result, redundant data are removed and related data are combined. Moreover, the provenance of each piece of information is tracked throughout the system, allowing scientists to choose which data to trust (Buneman *et al*., 2002). MiMI currently has 65,153 molecules and 151,021 interactions, and is the result of integrating BIND (Bader *et al*., 2003), DIP (Xenarios *et al*., 2002), (HPRD, 2005), (GRID, 2005), (Pfam, 2005), (InterPro, 2005) and GO (GO Consortium *et al*., 2001).

Through integration, MiMI creates a synergistic effect, and allows users to ask more advanced questions that each of its component databases can not answer independently. Because of the provenance used in MiMI, scientists can easily determine where conflicting information comes from, and whether to trust it or not. We also demonstrate that MiMI can be used to predict possible interactions. For instance, using homology data from Pfam and interaction data from BIND or DIP, it is possible to write queries that extract possible interactions.

MiMI's functionalities would be limited if it required users to understand the schema that the data was stored under. We have developed a user interface which allows the rawest of beginners to ask complex queries. Utilizing a simple point and click method, combined with form boxes, our user interface generates complete query statements for the user. It also allows advanced users to edit these queries and generate their own. Additionally, MiMI output complies with the PSI format, allowing users to take advantage of industry tools for viewing interactions.

We have built MiMI using Timber (Jagadish *et al*., 2002), a native XML database. This gives MiMI the power and security of all the traditional database features such as transactions, indexing and logging, yet allows the MiMI data model to be flexible to change when biological understanding modifies the attributes of an interaction.

Availability: MiMI is currently available at the University of Michigan, http://www.eecs.umich.edu/db/mimi. It is covered under the GNU General Public License (GPL). All data reported by MiMI is in the public domain.

Contact: {apchapma, jmagesh, congy, jag} @umich.edu

REFERENCES

Bader, G., D. Betel, C. Hogue (2003) BIND - The Biomolecular Interaction Network Database, *Nucleic Acids Research*, **31:1**, 248-250.

Bader, G., C. Hogue (2002) Analyzing Yeast Protein-Protein Interaction Data Obtained from Different Sources, *Nature*, **20**, 991-997.

Buneman P., S. Khanna, K. Tajima, W.-C. Tan (2002) Archiving Scientific Data, *ACM SIGMOD*, 1-12.

Gene Ontology Consortium (2001) Creating the gene ontology resource: design and implementation, *Genome Res*, **8**, 1425-1433.

The GRID: The General Repository for Interaction Datasets, Samuel Lunenfeld Research Institute, http://biodata.mshri.on.ca/grid/servlet/Index/.

Human Protein Reference Database, Johns Hopkins University and The Institute of Bioinformatics, http://www.hprd.org/.

InterPro, European Bioinformatics Institute, http://www.ebi.ac.uk/interpro/.

Jagadish, H.V., S. Al-Khalifa, A. Chapman, L. Lakshmanan, A. Nierman, S. Paparizos, J. Patel, D. Srivastava, N. Wiwatwattana, Y. Wu, C. Yu (2002) TIMBER: A Native XML Database, *The VLDB Journal*, **11:4**, 274-291.

MiMI, University of Michigan, http://www.eecs.umich.edu/db/mimi.

Pfam, Sanger Institute, http://www.sanger.ac.uk/Software/Pfam/index.shtml/.

von-Mering, C., R. Krause, B. Snel, M. Cornell, S. Oliver, S. Fields, P. Bork (2002) Comparative assessment of large-scale data sets of protein-protein interactions, *Nature*, **417**, 399-403.

Xenarios, I., Ł. Salwínski, X. J. Duan, P. Higney and S.-M. Kin, D. Eisenberg (2002) DIP, The Database of Interacting Proteins: A Research Tool for Studying Cellular Networks of Protein Interactions, *Nucleic Acid Research*, **30:1**, 303-305.