## Author

Before Vadim Bichutskiy approached Dr. Lathrop about becoming his faculty mentor, he did his homework. He read Dr. Lathrop's publications and was familiar with his interdisciplinary approach to research. Vadim suggests that anyone interested in undergraduate research should do the same before they choose a mentor, so they can find a professor whose interests are similar to their own. Now a graduate student at UCI's Donald Bren School of Information and Computer Sciences, Vadim plans to continue doing research to develop new drugs and improve health care. When he's not in the lab, Vadim enjoys swimming, weight lifting, skiing, and playing chess.

## Key Terms

- Data Warehousing
- Database Schema
- Heterogeneous Database Integration
- Mediation
- p53
- Tumor Suppressor

# Heterogeneous Biomedical Database Integration Using a Hybrid Strategy

**Vadim Bichutskiy**
*Information and Computer Science*

## Abstract

p53 is a central tumor suppressor protein that is involved in cell cycle regulation. It is estimated that more than 50% of human cancers have p53 inactivated due to gene mutations. Therefore, the ability to restore function to p53 could have an enormous impact on cancer treatment. This project developed a heterogeneous database to support the search for small molecules that restore function to p53. We used a hybrid strategy that combines the data warehousing and mediation approaches to data integration. The database integrates small molecule data, such as computational docking results, with functional and structural assay results. It unites different research laboratories into a common framework, accessible through the Internet by all involved researchers. It provides for increased productivity and more efficient data sharing. This results in improved chances of finding small molecules that restore function to p53 and can be developed into new anti-cancer drugs.

## Faculty Mentor

This paper is an elegant example of how faculty-mentored undergraduate research can be both an excellent learning experience for the undergraduate and a real contribution to interdisciplinary attacks on important scientific and technological problems of our day. The topic of this paper, heterogeneous databases, is a difficult and important problem in computer science. The application is to p53, which is an important medical problem because it is a central cancer-related protein. Over half of all human cancers have a mutation in the p53 gene, which inactivates a central tumor-suppressor pathway. Thus, the undergraduate research activity behind this paper addressed challenging problems on several fronts, and its successful completion proved to be both interesting and useful.

**Richard Lathrop**
*Donald Bren School of Information and Computer Sciences*

## Introduction

The central tumor suppressor protein p53 provides a potential target for new anti-cancer drugs. Toward this end, different p53 data is produced by four collaborating research laboratories at the University of California, Irvine (UCI): computational docking (in the Donald Bren School of Information and Computer Sciences), small molecule synthesis (Department of Chemistry), functional assays (School of Medicine), and structural assays (School of Biological Sciences).

Our goal is to support their efforts by developing a heterogeneous database that integrates each laboratory's data. Heterogeneous database integration is a challenging topic, which is important in several application domains (Karasavvas et al., 2004), and is one of the most important computer science problems today (Zhou et al., 1995). It is especially difficult in bioinformatics because of the inherent complexity of the domain, where (a) most rules have exceptions; (b) there is rich variety in data, from DNA and protein sequences, to three-dimensional images, to text files; and (c) there are complex relationships between structures (Karasavvas et al., 2004).

We interviewed relevant researchers to find out about their data and designed a database schema that captured it. Then, we integrated the heterogeneous databases. Finally, we wrote queries and stored procedures to retrieve the requested information. The schema is capable of storing all of the data. We are currently integrating data from the four laboratories listed above.

### *p53 Background*

The tumor suppressor protein p53 helps prevent uncontrolled cell growth. It has been identified as the main protein that protects humans from cancer. It performs its function by acting as a transcription factor, a protein needed to initiate transcription, for genes involved in DNA repair, cell cycle arrest, and apoptosis (programmed cell death) (Baroni et al., 2004). Thus, p53 prevents a cell from passing on mutations due to DNA damage. If mutations increase and the cell survives through many divisions—as is more likely if the p53 gene is defective or missing—cancer may result (Campbell and Reece, 2002). In fact, it is estimated that more than 50% of human cancers have p53 inactivated due to gene mutations (Friedler et al., 2002). Furthermore, cancers with inactive, mutant p53 are difficult to treat because they are especially resistant to radiation and chemotherapy (Bullock and Fersht, 2001; Bykov et al., 2002).

The p53 protein is a tetramer consisting of the amino-terminal transactivation, core DNA-binding, carboxy-terminal tetramerization, and regulatory domains. Approximately 95% of cancerous mutations lie in the core DNA-binding domain and prevent p53 from binding to DNA, which it must do to perform its function (Bullock and Fersht, 2001). Further, 75% of p53 gene mutations are single missense mutations, in which a single base substitution in the protein coding region of a gene results in an amino acid substitution in the protein (Bullock and Fersht, 2001). Therefore, mutant p53 is mostly a full-length protein with a single amino acid change in the core DNA-binding domain (Bullock and Fersht, 2001).

Gene mutations in p53 can be divided into several categories. Among them, DNA-contact mutations (R248, R273) result in loss of DNA-binding residues and "structural mutations" result in structural changes to the core DNA-binding domain (R175, G245, R249, R282). The six most frequent cancerous mutations are the "hot-spots" R175H, G245S, R248Q, R249S, R273H, and R282W (Friedler et al., 2002). Figure 1 shows the core DNA-binding domain of p53 protein bound to DNA. The six most frequently mutated amino acids in human cancers are labeled and shown in yellow. All of these residues are important for p53 binding to DNA.

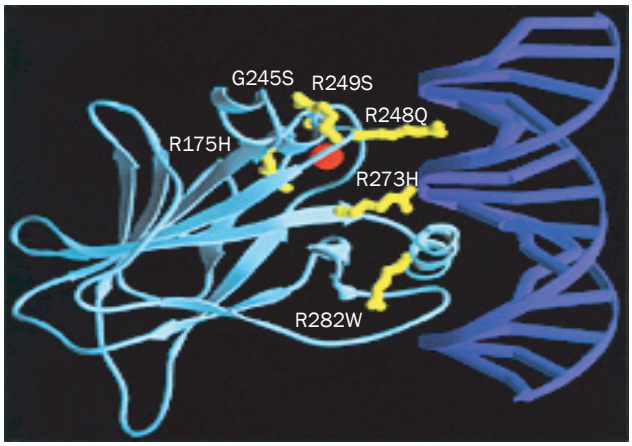


Figure 1
Core DNA-Binding Domain of p53 Protein (Modified from Cho et al., 1994)

### *Database Background*

Biomedical data is often distributed among multiple databases, which frequently have different schemas and are implemented with different technologies (Markowitz and Ritter, 1995). A database schema includes the gross structure and constraints on the database. Database heterogeneities, or differences, can make access to information difficult (Sujansky, 2001). Thus, the need arises for heterogeneous databases. A heterogeneous database unites various

databases, which support different schemas and technologies, by providing a uniform database schema and querying capabilities that integrate distributed data (Sujansky, 2001). The process of integrating data from multiple, heterogeneous sources is called heterogeneous database integration (Sujansky, 2001). In database integration, stored procedures and views are often created to facilitate querying. Stored procedures are operations stored within the database server that are available to clients. Views are "virtual" tables that are not physically stored in the database. Views provide alternate ways of looking at a database to different users. Because of database differences, heterogeneous database integration is a difficult but important problem in biomedicine (Sujansky, 2001).

Heterogeneous database integration is a topic that has been studied in the database research community for many years. However, no preferred solution or consensus of approach currently exists (Widom, 1996). The three most common approaches to heterogeneous database integration are (Garcia-Molina et al., 2000): federated systems (Sheth and Larson, 1990), data warehousing (Widom, 1995), and mediation (Weiderhold, 1992; Domenig and Dittrich, 1999).

Federated systems use decentralized architectures in which one-to-one connections are implemented between all pairs of databases that need to share data. These connections allow a database D1 to query another database, D2, in terms that D2 can understand. Software components are written to translate queries between the databases.

In data warehousing, data from each source is extracted, merged, and stored in a centralized repository (warehouse). The warehouse is a database with a global schema that combines the schemas of the sources. Queries on the system are evaluated at the warehouse without accessing the original sources. Client updates to the warehouse are usually not allowed since they are not reflected in the original sources and would make the warehouse inconsistent with the sources. Instead, the warehouse is updated from the data in the sources. There are multiple policies for updating the warehouse (Garcia-Molina et al., 2000). Implementations of data warehousing approach include Squirrel (Hull and Zhou, 1996) and WHIPS (Hammer et al., 1995).

In mediation, a module called a "mediator" accepts a query from the client, determines the sources needed to answer the query, and decomposes the query into subqueries for each required source. The subqueries are translated to the source-specific query language via modules called "wrappers." The results from the sources are translated back into the common query language by the wrappers. Finally, the mediator obtains results from the wrappers, combines them, and returns the final answer to the client. Mediation can be query-centric or source-centric (Li, 2001). Mediation is one of the most common approaches to data integration in bioinformatics (Karasavvas et al., 2004). Systems that use the mediator approach include TSIMMIS (Chawathe et al., 1994), Information Manifold (Kirk et al., 1995), SIMS (Arens et al., 1996), and Carnot (Singh et al., 1997).

### *Hybrid Strategy*

Ashish (2000) proposed a hybrid strategy to data integration that combines the benefits of the data warehousing and mediation approaches. In the hybrid strategy, part of the data is fetched on demand as in the mediation approach, but other data is collected, stored, and integrated in the warehouse. Many of the complex, large-scale database applications of the future will require both mediation and data warehousing (Widom, 1996). While the best approach varies with the application, it is likely to be a hybrid strategy based on the combination of different approaches (Eckman, 2003). However, the hybrid strategy is less commonly discussed in the literature.

## Materials and Methods

The database was developed using Microsoft SQL Server 2000, a client-server database management system (DBMS) used to create, modify and query a database. We conducted requirements analysis by interviewing p53 researchers. We needed to know what data each laboratory produced, what information they wanted to see from other laboratories, and what queries they wanted to perform on the database. Based on these interviews, we designed a database schema. We integrated computational docking data via a connection to the research laboratory's PostgreSQL database. The connection allowed us to query the laboratory's database. Then we integrated functional assay data by importing it into the project's SQL Server database.

### *Project Description*

Query responses were gathered and categorized from the four research laboratories. Investigators included Dr. Darren Holmes (Department of Chemistry), Dr. Richard Chamberlin (Department of Chemistry), Dr. Felix Grun (School of Biological Sciences), S. Joshua Swamidass (Donald Bren School of Information and Computer Sciences and School of Medicine), and John Coroneus (School of Biological Sciences). We also collaborated with Dr. Rainer Brachmann (School of Medicine), Dr. Pierre Baldi (Donald Bren School of Information and Computer

Sciences and School of Medicine), and Dr. Richard Lathrop (Donald Bren School of Information and Computer Sciences and Department of Biomedical Engineering).

The goal of the project was to develop a heterogeneous database that integrates various data produced by four collaborating research laboratories at UCI. The data relates to experiments performed on small molecules with the goal of finding small molecules that restore function to p53 and can be developed into new anti-cancer drugs.

Computational docking is performed on a library of compounds to identify small molecules that have the potential of restoring function to p53. Each docking experiment is done on a receptor, which is typically a protein. A receptor has binding sites to which small molecules bind. A molecule may have multiple conformations, the three-dimensional shape defined by angles of rotation about the molecule's bonds. The conformation of a molecule is specified in a mol2 file. Docking uses computer algorithms to model how a molecule binds to a binding site of the receptor. The result of a docking experiment is a score that measures the ability of the molecule to bind to the receptor's binding site. Molecules with the best scores are synthesized. Finally, the synthesized molecules are assayed on the p53 mutants to determine what effect they have on the mutants.

The database will allow p53 researchers to share data, provide their results, and facilitate efficient communication across laboratories in a common and convenient framework accessible through the Internet. Thus, the database would greatly improve the chances of discovering new anti-cancer drugs. However, the database should not simply reproduce the sources' data. The database would store only the results of experiments on small molecules that at least one laboratory feels have potential of restoring function to p53 and, thus, would be interesting to other laboratories. In addition, the database should not store all of the data features. It should store only the most important attributes of the data.

## *Database Design*

After the interviews, we developed a database schema that captured the data produced by all of the research laboratories. The database design schema is shown in Figure 2. Each molecule ("Molecules" table) may have more than one conformation ("Conformations" table) and it may come from more than one source ("Sources" table). There are two types of experiments ("Experiments" table) that are done on molecules: docking and assays. We stored the results ("DockingResults" and "AssayResults" tables) of these experiments. Each type of experiment is done on a particular p53 mutant ("Mutants" table) and has a score ("Scores" table) associated with it.

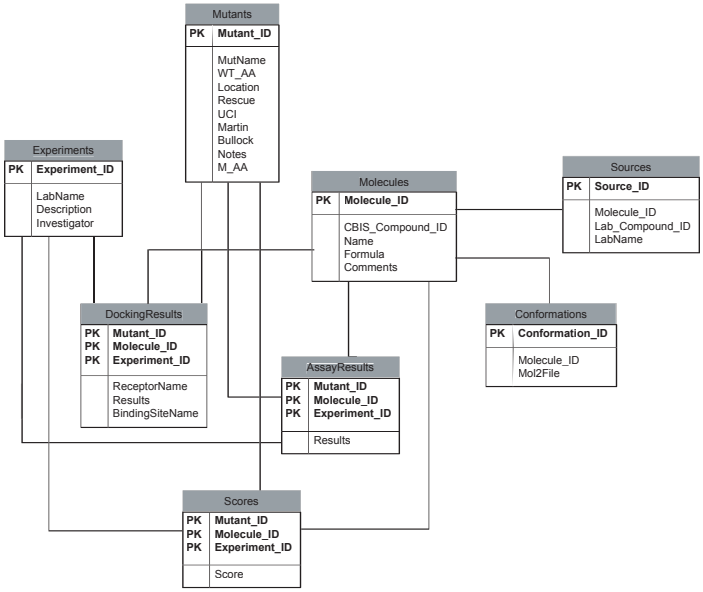


Figure 2
p53 Database Design Schema

The schema design proved to be a difficult task. We had difficulty determining exactly what data was needed and how to store it. As a result, we developed several designs before we found a good solution. During the schema design, we realized that the data consisted of docking and assay results and each result was associated with different conditions: experiments, mutants and molecules. Thus, the schema follows a design pattern of having results tables related to "Experiments," "Mutants" and "Molecules" tables. This makes the schema flexible to changes. If a new laboratory were to become part of the project that produces some data, we would create a new results table with relationships to "Experiments," "Mutants" and "Molecules" tables. Alternatively, if a new condition were created, we would create a new condition table with relationships to all results tables.

## *Data Integration*

Following the development of the database schema, the next step was to input and integrate into the database the data from all of the research laboratories. Figure 3 shows the project's system architecture and the hybrid strategy to data integration.

We integrated computational docking data into the database by connecting the project's Microsoft SQL Server database to the laboratory's PostgreSQL database via Open DataBase Connectivity (ODBC), a standard database

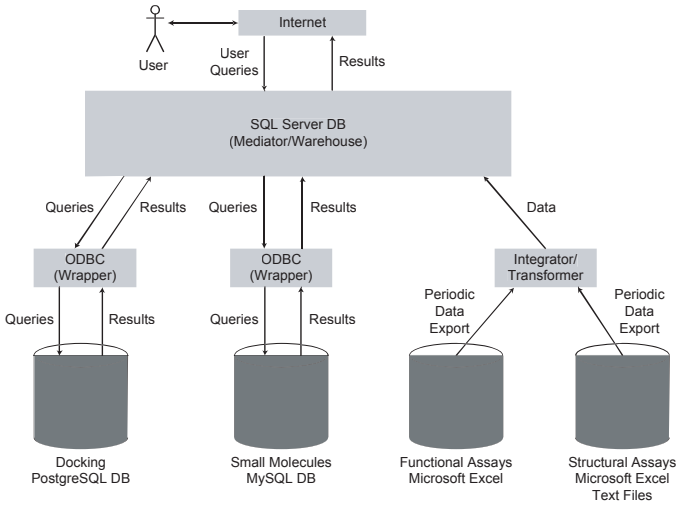


Figure 3
System Architecture and the Hybrid Strategy to Data Integration

access method that makes it possible to access any data from any application. The project database runs on the Windows server while the laboratory database runs on the Linux server. We linked the laboratory's database server with the project's database server via the "Linked Servers" feature of Microsoft SQL Server. Then, the laboratory built a query-based view that contained the docking information that we captured in the project database. Finally, we queried the view to obtain the necessary information. Figure 4 shows the results of a query that retrieved computational docking scores from the laboratory's database. The scores consisted of a molecule ID (mid), receptor ID (rid), score, and score name.

We integrated functional assay data by importing it into the project's database. Because functional assay data was stored

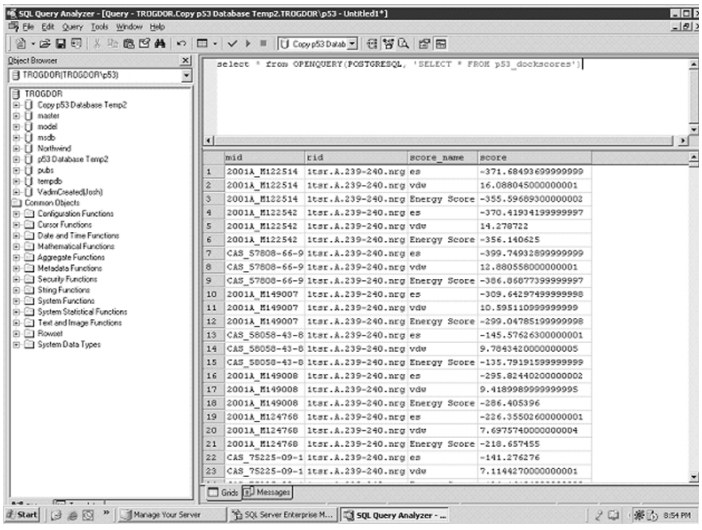

Figure 4
The Results of a Query that Returned p53 Docking Scores Data

in Microsoft Excel spreadsheets and changed periodically, we imported it manually into the project's database via physical file transfer. Data transformation was performed on the functional assay data via complex queries before it was inserted into the database tables.

The project's SQL Server database is both a mediator and a data warehouse. Computational docking data is queried on demand by accessing the laboratory's PostgreSQL database without importing it into the project's database. The ODBC driver for PostgreSQL acts as a wrapper that translates a query written in SQL Server into a query that PostgreSQL understands. Thus, for docking data, the project's database acts as a mediator. On the other hand, functional assay data is imported into the SQL Server database, which acts as a warehouse.

## Hybrid Strategy Guiding Principles

There are many ways to create a hybrid database architecture. One alternative to the project's hybrid strategy is to treat the data warehouse like a data source with a mediator architecture (Voisard and Jurgens, 1998). It is a major design decision to determine what part of the data to import into the warehouse and what part to query on demand. We considered the following principles in the design of a hybrid strategy to data integration:

*Changes to data.* Data that changes often should be queried on demand. The mediator approach was chosen for docking and small molecules because this data changes often.

*Size of the data.* Larger data should be queried on demand. The docking and small molecule data potentially could be extremely large. Thus, we chose the mediator approach for this data.

*Availability of sources.* Data from sources that are always available should not be in the warehouse. The interviews indicated that the laboratories' PostegreSQL and MySQL databases would be reliable most of the time.

*Required query processing time.* The shorter the required response, the greater the need to import data into the warehouse. Although queries on docking and small molecules may take some time to process, we decided to query them on request because of the potentially large size of query results and the frequent changes to the data.

*Predictability of queries.* Predictable queries on data that does not change often should be written in advance with the

results stored in the warehouse. In addition, knowledge of the queries to be performed on the system should be taken into consideration during global schema design. Some of the important queries to be performed on the system were provided by the researchers. We implemented these queries as stored procedures to call from the Internet. In addition, we tried to design the schema in such a way that the joining of tables would be as efficient as possible.

In theory, one would like to use all of the principles in an optimal way. In practice, however, there are trade-offs that need to be made between the principles. For example, we decided to use the mediator approach for docking and small molecule data even though some queries may take some time to process. On the other hand, the size of the data and the frequency of data updates clearly favored this approach.

## Conclusion

The database allowed p53 researchers to share data, provide their results, communicate with each other more efficiently, and improve their chances of finding new anti-cancer drugs. This project demonstrates that a hybrid strategy is a viable approach to heterogeneous database integration in biomedicine. It combines the benefits of the data warehousing and mediation approaches. Mediation should be used for data that changes rapidly such as p53 docking scores, for queries on large amounts of data, and for sources that are reliable. Data warehousing should be used for data that changes rarely such as functional assays, for data with predictable queries, and for queries requiring high performance but not necessarily over the most recent state of the information. Furthermore, this project could lead to further applications of the hybrid strategy to heterogeneous database integration in bioinformatics.

Developing a database that integrates data from different sources is a difficult task; requirements analysis is essential. Multiple design iterations are inevitable. It is important to understand what data needs to be stored, what data is produced by each source, and what queries are to be performed on the database. In heterogeneous database development, schema design is the most critical, challenging and time-consuming phase. The design difficulty is compounded in that some sources store their data in various software systems such as Microsoft Excel and text files. This also makes data integration more difficult because there is no standard method of connecting to the data.

In the future, we hope to integrate the rest of the p53 data, small molecule synthesis, and structural assays into the database. We would like to write stored procedures that allow for automatic data updates of the database. Furthermore, we hope to develop a user-friendly Web front-end that allows p53 researchers to query the database and obtain necessary information.

## Acknowledgments

I would like to express my deepest appreciation to my advisor Professor Richard H. Lathrop without whom this work would have never been possible. Professor Lathrop has had the greatest influence on me as a researcher. His ability to recognize critical issues and focus on the "big picture" are the qualities that I seek. His guidance, encouragement and confidence in me have given me many opportunities to work with the most accomplished researchers in our field.

I am grateful to Richard Colman, with whom I have worked on this project. He gave me important advice and guided me during the course of this project. I would also like to thank other people with whom I have collaborated on this project: Dr. Pierre Baldi, Dr. Rainer Brachmann, Dr. Richard Chamberlin, John Coroneus, Dr. Felix Grun, Dr. Darren Holmes, and S. Joshua Swamidass.

## Works Cited

Arens, Y., C. Hsu, and C. A. Knoblock. "Query Processing in the SIMS Information Mediator." Readings in Agents. Eds. Michael N. Huhns and Munindar P. Singh. San Francisco: Morgan Kaufmann, 1998. 82-90.

Ashish, N. "Optimizing Information Mediators By Selectively Materializing Data." Ph.D. Thesis. Computer Science Department. University of Southern California, March 2000.

Baroni, T. E., T. Wang, H. Qian, L. R. Dearth, L. N. Truong, J. Zeng, A. E. Denes, S. W. Chen, and R. K. Brachmann. "A Global Suppressor Motif for p53 Cancer Mutants." Proceedings of the National Academy of Sciences of the United States of America 101.14 (2004): 4930-4935.

Bullock, A. N. and A. Fersht. "Rescuing the Function of Mutant p53." Nature Reviews 1 (2001): 68-76.

Bykov, V., N. Issaeva, A. Shilov, M. Hultcrantz, E. Pugacheva, P. Chumakov, J. Bergman, K. G. Wiman, and G. Selivanova. "Restoration of the Tumor Suppressor Function to Mutant p53 by a Low-Molecular-Weight Compound." Nature Medicine 8.3 (2002): 282-288.

Campbell, N. A., and J. B. Reece. Biology Ed. Beth Wilbur. 6th ed. San Francisco: Benjamin Cummings, 2002.

Chamberlin, R. Personal Communication. 14 February 2004.

Chawathe, S., H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman, and J. Widow. "The TSIMMIS Project: Integration of Heterogeneous Information Sources." 10th Meeting of the Information Processing Society of Japan (IPSJ) 1994: 7-18.

Cho, Y., S. Gorina, P. D. Jeffrey, and N. P. Pavletich. "Crystal Structure of a p53 Tumor Suppressor-DNA Complex: Understanding Tumorigenic Mutations." Science 265 (1994): 346-355.

Coroneus, J. Personal Communication. 17 February 2004.

Domenig, R. and K. Dittrich. "An Overview and Classification of Mediated Query Systems." ACM SIGMOD Record 28 (1999): 63-72.

Eckman, B. A. "A Practitioner's Guide to Data Management and Data Integration in Bioinformatics." Bioinformatics: Managing Scientific Data. Eds Zoe Lacroix and Terence Critchlow. San Francisco: Morgan Kaufmann, (2003): 35-73.

Friedler, A., L. O. Hansson, D. B. Veprintsev, S. M. V. Freund, T. M. Rippin, P. V. Nikolova, M. R. Proctor, S. Rudiger, and A. R. Fersht. "A Peptide that Binds and Stabilizes p53 Core Domain: Chaperone Strategy for Rescue of Oncogenic Mutants." Proceedings of the National Academy of Sciences of the United States of America 99 (2002): 937-942.

Garcia-Molina, H., J. D. Ullman, and J. Widom. Database System Implementation Upper Saddle River: Prentice Hall, 2000.

Grun, F. Personal Communication. 1 April 2004.

Hammer J., H. Garcia-Molina, J. Widom, W. Labio, and Y. Zhuge. "The Stanford Data Warehousing Project." IEEE Bulletin of the Technical Committee on Data Engineering 18 (1995): 41-48.

Holmes, D. Personal Communication. 20 January 2004.

Hull, R. and G. Zhou. "A Framework for Supporting Data Integration Using the Materialized and Virtual Approaches." Proceedings of the ACM SIGMOD International Conference on Management of Data 1996: 481-492.

Karasavvas, K.A., R. Baldock, and A. Burger. "Bioinformatics Integration and Agent Technology." Journal of Biomedical Informatics 37 (2004): 205-219.

Kirk, T., A. Y. Levy, Y. Sagiv, and D. Srivastava. "The Information Manifold." AAAI Symposium on Information Gathering in Distributed, Heterogeneous Environments 1995.

Li, C. "Query Processing and Optimization in Information-Integration Systems." Ph.D. Thesis. Computer Science Department. Stanford University. August 2001.

Markowitz, V. M. and O. Ritter. "Characterizing Heterogeneous Molecular Biology Database Systems." Journal of Computational Biology 2 (1995): 547-556.

Sheth, A.P. and J. A. Larson. "Federated Databases for Managing Distributed, Heterogeneous, and Autonomous Databases." ACM Computing Surveys 22 (1990): 183-236.

Singh, M. P., P. E. Cannata, M. N. Huhns, N. Jacobs, T. Ksiezyk, K. Ong, A. P. Sheeth, C. Tomlinson, and D. Woelk. "The Carnot Heterogeneous Database Project: Implemented Applications." Distributed and Parallel Databases Journal 5 (1997): 207-225.

Sujansky, W. "Heterogeneous Database Integration in Biomedicine." Journal of Biomedical Informatics 34 (2001): 285-298.

Swamidass, J. S. Personal Communication. 26 February 2004.

Voissard, A. and M. Jurgens. "Geospatial Information Extraction: Querying or Quarrying?" Technical Report. International Computer Science Institute, Berkeley, CA. April 1998.

Weiderhold, G. "Mediators in the Architecture of Future Information Systems." IEEE Computer 25 (1992): 38-49.

Widom, J. "Research Problems in Data Warehousing." Proceedings of the 4th International Conference on Information and Knowledge Management (CIKM) 1995: 25-30.

---. "Integrating Heterogeneous Databases: Lazy or Eager?" ACM Computing Surveys 28 (1996).

Zhou, G., R. Hull, R. King, and J. Franchitti. "Supporting Data Integration and Warehousing Using H2O." IEEE Data Engineering Bulletin, Special Issue on Materialized Views and Data Warehousing 18 (1995): 29-40.