Facing the Challenges of Data Integration in Biosciences

Aiguo Li[1]

Abstract— **Data integration in molecular biology and clinical science has become imperative for providing the comprehensive information extraction in systems biology. In this review we evaluate the evolution and characteristics of biological databases and examine existing approaches to data integration in bioscience. Strengths and weaknesses of these approaches are identified by surveying several successful examples in biological data integration. We point out the challenges faced and possible solutions in biological data integration on various levels while contrasting the efforts of data integration in biosciences with those in industry.**

Index Terms **- data integration, federation, warehouse, and bioscience.**

INTRODUCTION

Science has historically been divided into disciplines for the convenience of learning and of managing the complexity. As a result, the scientific data have been traditionally collected and managed by categories. With the advent of high throughput technologies in biology, the data accumulation has been growing exponentially; with the current state of web technologies and data management technologies in computer science, it becomes possible to understand organisms systematically by comprehensive data analysis and information extraction [1] [2]. The requirements for comprehensive information extraction and data analysis across diverse disciplines are becoming emergent for the need of our understanding in system biology [3] [4]. The current data marts or databases, most of which are originally designed for the purpose of data storage or repository, become incompetent in answering comprehensive questions related to discovery or decision-making in large scale [4] [5]. The efforts of data integration through data warehouse or data federation are needed to answer the complex analytical queries in the bioscience [6]. Although well structured solutions for data integration in industry already exist, the technologies for biological data integration are still in its infant stage and special care needs to be taken with respect to the unique properties of biological data [7] [8].

Manuscript received May 1, 2006.
A. G. Li is with the Neuro-oncology Branch, National Cancer Institute, National Institute of Health, Bethesda, MD 20892 USA (phone:301-435-1454; fax:301-480-4743; e-mail: liai@mail.nih.gov)[1]

CHARACTERISTICS OF BIOLOGICAL DATABASES

The characteristics of biological databases are derived from their unique origin and history. Originally, most biological databases were devised by a group of scientists who have limited database background. The major purpose of these databases was data storage rather than information extraction [8]. Furthermore, biological data is hierarchical by nature and its data types are tightly correlated with the specific technologies of data acquisition [6]. As a result, the databases are sporadic, data types are heterogeneous and span diverse domains [4] [6]. For example, biological data for human species traverse multiple levels, such as organism, organ, tissue, cell, organelle and pathways or networks, and span diverse domains, such as genomics, transcriptomics, proteomics, pheneomics, localizeomics, ORFeomics, pharmacogenomics, pharmacogenetics clinical trials, etc [2]. The nature of biological data, plus its unique evolution over history leads to some special features of molecular databases, which is summarized in detail below.

1. Molecular data are highly heterogeneous due to the inherent complexity in biological system and a wide array of technologies used to study them. The new terminology "omics" is a real-life reflection of this reality [2]. The classification of databases based on the contents of data can easily be divided into the following main categories: genome database, gene databases, gene expression databases, protein databases, protein-protein interaction databases, pathway databases, etc. Each of these database types themselves can be easily divided into several subtypes of databases. For example, the protein database includes protein sequence databases, protein structure databases, protein signature, domains, profile databases. Table 1 listed some examples of the major molecular databases categorized as gene and transcripts, gene expression, genomes, protein sequences, protein domains and motifs, ptotein-protein interactions, protein structures and pathways.
2. The data volume is large with unique data types, and data accumulation is on-going and far from complete. For instance, the estimated human gene number in total is 20,000 to 25,000 [9]. Without considering individual differences or ethnic differences, theoretically, a completed gene expression profiling database should contain expression profiles of all these gene in all human organs/tissues, cell types in particular cases, covering various development stages or time lines without considering any stress effects. Viewed in the context of other molecular types such as DNA and proteins and the

various types of technologies used to study them, the volume of biological data becomes extremely high. The data types are dictated by the biotechnologies used in experiments [6]. Data types in bioscience include the common data types in industry, such as integer, real number and other character string, also it has some unique data types, such as DNA sequences, graphics, images of 2-D gels and of immunohistology, with which special care needs to be taken to store and to analyze them [6] [8] [10].

3. Data sources in bioscience are highly dynamic. The data dimensions in biosciences are expanding rapidly as a result of the development and the innovation of new technologies. To pace with these changes, new data types or databases are emerging all the time and the existing databases continuously restructure their formats to incorporate the new data, which leads to the multiple generations/releases of legacy databases yearly. For example, GenBank, one of the major genome, gene and protein data repository system, make their release bimonthly [1].
4. Biological data structure is highly hierarchical by nature. For example, a gene is a fragment of DNA in a chromosome and a chromosome is located in the nucleus of a cell, this gene will encode one or more proteins through one or more mRNAs. These proteins will likely function in one or many pathways in various tissues. This type of deep hierarchical structure is very common in biology and it could be difficult to model and inefficient to query using traditional relational models [5] [10].
5. Lack of standardization in data formats and in controlled vocabularies in scientific domains. Molecular databases are highly heterogeneous due to their original formation and history. As a result, the database schema of the similar biological data types by different databases will be quite different due to the technologies used. For example, data formats and data types for gene expression profiling from affymetrix oligonucleotide arrays can be quite different from those of cDNA arrays. Much more differences will be found in the gene expression profiling from serial analysis of gene expression (SAGE) technologies. Although all these three technologies are used in detecting gene expression levels in tissues or cells, one will find that the across-platform comparison is almost impossible at the data-analysis level. Additionally, data formats vary over different domains and over different projects. The vocabularies in describing biological objects are ambiguous due to the fact of widely used synonyms and homonyms. We end up with a vast mosaic of databases in one biological domain with different formats typically using non-standard query software specific for that particular database [4]. These databases and systems often do not have an explicit database schema, which is conventionally considered as a formalized catalogue of all interrelated tables in a database with well-defined attributes and well-structured indices of these tables, which is prevalent in industry databases [5] [11].
6. The database management applications and data-access tools for biological databases are at their infant stages [8] [10]. Lack of standardization in data formats and the dynamics in data types hamper the development of application tools in biological database management system [6]. Hence the retrieval efficiency is low and complicated, and heterogeneous applications need to be developed to handle the information extraction and analysis [11].
7. Data annotation using external sources through hypertext has been a successful solution for integrating relevant external sources from diverse domains with the advance in web technology [12]. Hence, the hypertext constitutes a part of the database contents and provides added meaning to biological entities. Additionally, this type of point and click interfaces attract biologists simply because it is easy to learn [8]. Nevertheless, one of the disadvantages in hypertext is its vulnerability to the ambiguity in identifiers or terminology system [4] and the reliability on the internet accessibility.

REQUIREMENTS FOR BIOLOGICAL DATABASES AND APPLICATIONS

The unique features of data or databases in biosciences hold some interesting requirements from biologists to the databases in biosciences. For example, to answer comprehensive biological queries one often needs to traverse a wide range of object domains from many heterogeneous databases and a user must click through many interfaces and must make efforts to manage intermediate results [6] [13]. This situation constitutes a challenge to biologists and bioinformaticians. Customized applications on top of traditional database management system that interfaces with a large number of databases are needed to achieve the satisfactory query results [4] [6] [8]. Furthermore, a data integration technology that recognizes which parts of two data sources have the same meanings or overlapped domains is desirable. The detailed requirements of databases in biosciences can be specified below. 1) The heterogeneous features of biological databases require that data models and database management systems in biosciences are capable of handling data types and are flexible in dealing with data types [8]. Otherwise the possible constraints of data types and values placed on databases and database management systems could result in the exclusion of unexpected types and values in biosciences [14] and further diminish the reliability of query results or data analysis results. 2) The highly dynamic feature of biological databases challenges the database and application development community in biology to support database schema evolution and data object migration for improving information flow between generations/releases of databases [6] [11]. Currently, the ability to extend the database schema to meet the requirements of frequent changes in the biological setting is unsupported in most relational and object database management systems [10]. However, this sort of tracking in history is important for biological researchers to be able to access and verify previous results. Therefore, mechanisms for aligning different biological databases with similar contents or different versions of formats should be supported. Data alignment tools, and data integration tools based on the various biological workflow for legacy databases should be available

[6] [8]. 3) The deep hierarchical nature of biological data causes some concerns whether relational schemas will meet the challenge for efficient data representation and retrieval in highly integrated data warehouse [10]. Object-oriented relational database schemas or object database schemas considering their large potential to model hierarchical structures probably meet these needs better. However, the progress in developments of object database management systems is slower than our expectation and its system query tools and optimizers are still in its early development phase [14] [15]. 4) Most of the molecular databases are in the category of deep, primary databases with point solution (Table 2), which basically means that databases are highly diversified and data types are heterogeneous and the system application tools on the top of the databases are simple. Isolated data are of less useful in biological systems and the meaningful information extraction needs highly integrated databases with complex contents and intensive data analysis [4]. On the other hand, data quality is the underlying foundation for reliable query results. It is an essential requirement for biologists to have curated knowledge bases derived from reliable systems. It has been proven that data from single high throughput method is more error prone than those obtained by integrating data from multiple approaches [2]. Currently the federated data is poorly transformed which leads to erroneous results. Overall, the needs of biologists for high quality data to answer complex queries are not met yet [4] [11].

COMPARISON BETWEEN DATABASES IN INDUSTRY AND BIOSCIENCES

The major differences between industrial and biological databases and their management systems can be summarized as below. First, biological databases are still in their infant stage and data itself are still growing exponentially both in quantity and new data types are emerging all the times [1] [2]. Industry data is increasing in quantity, but the data types are relatively stable [10]. Secondly, an industry database can be conventionally considered as a set of formalized interrelated tables in a database with well defined attributes for each table and well structured index on attributes [10]. On the top a commercial database management system sits and efficient data retrieval can be achieved with minor efforts in customization [5] [10]. On the other hand, the majority of the biological databases exist without conforming to data formats and contents within or across knowledge domains [8]. The function of the database management system needs to be substantially extended to meet the need of the requirements of information extraction and database management. Thirdly, databases in bioscience are dispersed geologically over various biological domains, although initial efforts of data integration have already started and the promising results have been seen in spotted areas [6] [16]. However, there are no system solutions or approaches to solve the problem yet. On the other hand, in industry a well structured solution has matured for a data warehouse construction using various sets of tools or systems for data extraction from data warehouse, data transformation, data loading and data analyzing [5]. In biological fields the development of data management application tools are still in their infant phases.

CLASSIFICATIONS OF MOLECULAR DATABASES

Databases in biosciences have been classified based on initial motivation and goals with which they were designed and built [11], and the contents of databases [14]. Although there are overlaps, grey areas and confusion occasionally, the existing classification does help us, both bioinformaticians and biologists, to figure out at what stage or level a database is and how sophisticated the database management system or application tool are on the top of the database. The classification provides us with some quick guidelines to find out which categories and which database we should use to fulfill our tasks in daily work. The major classifications of molecular databases and their applications are described in detail below.

1. Primary versus secondary databases

A group of scientists from 3rd Millennium Inc [11] separate biological databases into primary and secondary databases in terms of the original goals of designers to build them [11]. Primary databases are mainly used for data repository and archival although almost all of them have some simple data retrieval functions with various complexities. The primary database contains experimental results with certain degrees of integration presented as annotation of primary data. The typical example of a primary database is GenBank and its main function is for nucleotide sequence data repository from experimental labs or projects, although the internal interpretation of the data and standardization of the format defined by GenBank is enforced during the submission [1]. On the other hand, secondary databases are designed to provide a curate review of experimental data. It often contains data from several databases sources or publicly available sources, such as publications. The Pfam database is a typical secondary database that contains protein sequence structures (i.e. protein signature, domain) extracted automatically or manually from primary protein databases [17].

2. Deep versus broad databases

Based on the nature of database contents Cornell et al. divide molecular databases into deep or broad databases [14]. A broad database stores a single kind of data, but collects such data from many organisms. The example of a typical broad database is: SwissProt, storing protein from all organisms [18]. As for broad databases, the functional focus is mainly on browsing and visualizing rather than querying and analyzing. On the other hand, a deep database focuses on one or a small number of species, but stores many different kinds of data generally including both sequences and functional data. The typical examples of deep databases are SGD [19], YPD [20] and GIMS [14]. The original intention of deep databases is to answer complex queries through certain degrees of data integration.

3. Point solution versus general solution databases

According to the original design purposes, Wong et al. separate biological databases and their systems into point solution and general solution databases [8]. The system design goals for a point solution database are to address predefined specific

problems or questions and often data sources are small with limited scalability. In contrast, for a general solution database there are neither predefined data sources nor questions to ask during the system design phase. Hence, the extensibility in incorporating additional data source and flexibility in answering general queries are considered during the design phases [8].

Based on the above classification, the major molecular databases and their applications can be categorized as shown in Table 2.

SUCCESSFUL EXAMPLES OF DATA INTEGRATION EFFORTS IN BIOSCIENCES

Different from industry, in which the conventional solutions and technologies for data integration are mature and stable for the purpose of decision making, in bioscience arena the data types and data volumes are changing frequently, which presents a serious effort on the localized database maintenance and data model updating in data warehouse [2] [3]. Probably for that reason, the solution of data integration in molecular biology at the early stage was primarily focused on the data federation approach. In this approach, resource databases in the system remain dispersed or autonomous and the federation system facilitates accessibility to a wide range of databases using an application wrapper which often has a common data model and the source database is mapped to the shared data schema on the fly through the internet [7]. The characteristics of molecular databases make the data federation a crucial solution for biological data integration. SRS [21], DiscoveryLink [22], K2/Kleisli [6], and OPM [23] are the examples of data federations. The strength of data federation lies in the relief of the burden of keeping up with the maintenance and updating of the source databases, but the weakness is the difficulty in data cleansing and the limitation of the network performance. Additionally, due to the fact of heavy hypertext contents in biological database, the link-driven federation has been an important component in data federation [8] [24]. The successful example of this special type of federation is SRS [21]. In the data warehouse approach all data sources from diverse databases are pooled into one physical database with a common data model. The father of the data warehouse, Inmon, defines a data warehouse as "a collection of integrated subject-oriented databases designed to support the decision support system function, where each unit of data is relevant to some moment in time" [7]. The typical example of data warehouses in biosciences are the UCSC genome browser [25], GUS [6] and REMBRANDT[26].

A brief survey and examination of several data integration examples are detailed in this section.

1. Successful examples of data federation

SRS (Sequence Retrieval System) marketed by Lion Bioscience of Heidelberg, Germany, is one of the earliest data federation systems with wrappers for over about 400 source databases including biological sequences, metabolic pathways and literature abstracts [21]. SRS is built on the indexing technology of flat files using Icarus as a parsing language. A front end interface simplifies the query formulation and the result view from source databases. Additionally, on top of SRS, Lion Bioscience provides the SCOUT suite of applications that integrates with SRS with each SCOUT application designed for an analysis of a given biological domain [21]. However, SRS itself is more a navigational tool with limited data joining and data restructuring capabilities. It does not offer additional applications in organizing or transforming the retrieved results in a way that might be needed for setting up an analytical pipeline. For this reason, SRS is considered as a link-driven data retrieval system instead of data integration tool [8] [11]. Different from SRS, DiscoveryLink, an IBM system, possesses an explicit relational data model, which facilitates the relational data representation and query formulation with standard SQL [22]. The wrappers in DiscoveryLink are written in C++, which falls into the industry standard also. DiscoveryLink's strength lies in its use of SQL as its query language and its sophisticated query optimization technology [8]. Different from SRS, DiscoveryLink is designed to query relational databases rather than to access flat files. Kleisli, developed by the University of Pennsylvania, provides a view integration environment through data federation approach [8]. A global nested relational data model allows the relationships in source data to be mapped [8]. The wrapper for data source is written in Collection Programming Language and the query language is an extension of SQL called sSQL. Java and Perl application programming interfaces are available for writing applications that interact with the middleware. Strengths of K1 are that it does not require the data schema to be available and has a nested relational data model and a data exchange format that external databases and software systems can easily translate into. K2, descendant of Kleisli, inherits the nested data model, but use SQL as the query language for fast and completed data access of SQL query formulation instead of sSQL [8].

2. The successful examples of data warehouse

University of California Santa Cruz (UCSC) Genome Browser database is a data warehouse originally developed to support the human genome project and is now one of the main genomic browsers including sequences from human and mouse currently across a couple of releases [25]. Besides genomic sequences, the EST and mRNA sequences are also stored in the database and extensive annotations from sequence alignments as well as from external resources can be easily retrieved from this system [25] [27]. The Integr8 project, a jointly efforts among several European bioinromatics groups, is aimed at creating a framework that integrates genomics and proteomics data into one database and provides an entity-centric view of integrated data [13]. This project utilized advanced application tools in bioinformatics fields and brought together approximately 25 legacy databases in a diverse biological domains [16] [27]. The heterogeneous data is modeled using a unified data model in universal modeling language and a n-tier architecture is implemented on top of the underlying relational database [27]. This is a broad database with current contents of data from over 240 different species [13]. Genomics Unified Schema (GUS) is a data warehouse containing DNA, RNA and protein sequences and annotation information from major legacy databases [6]. GUS transforms the sequence-centric entities in external source database into gene-centric entities

and the identification of erroneous annotation was done during the data transformation [6]. Repository of molecular brain neoplasia data (REMBRANDT, http://rembrandt-db.nci.nih.gov), a recently jointed effort between the Neuro-oncology branch and the National Cancer Institute Center for Bioinformatics of NIH, is targeted to develop a national molecular, genetics and clinical database containing primary brain tumor data from large number of patients. Furthermore, molecular classification schema for primary brain tumor will be established by intensive data analysis of these data. Currently, the data warehouse contains gene expression profiling data from cDNA microarray and affymetrix oligoneucleiotide microarrays as well as genetic abnormality profiling data using Affymetric SNP arrays as well as clinical data of about 200 patients. On the backend REMBRANDT is a n-tier architecture, the conceptual model is designed using universal modeling language and implemented using J2EE. The unique feature of REMBRANDT lies in its emphasis on addressing the complex queries through computationally intensive data analysis in the context of primary data integration. Additionally, the extension with caBIO and caCORE will greatly improve the annotation capabilities of the biological entities in the data warehouse [28]. An recent business call between IBM and Lion with intention to integrate SRS to DiscoveryLink can potentially bring the data integration to another level both in data contents and in performances because of the complementary features of these two systems.

XML

With the tremendous diversity of data elements and frequent changes in data types in the biological domain, an extensive and flexible data model for data storage, representation and data exchanges has to be used. XML, a meta language supporting the specification of other languages [29], is becoming increasingly important for meeting this challenge [30]. XML is able to represent the hierarchical structure in biological data, and to uniformly represent ontological data, presenting additional strength to meet with challenges in biological data integration [31]. Native XML database systems such as PharmGKB [32] have been developed to manage the biological data, which indicates the feasibility using XML to handle the complex relationships [32] [33]. Alternatively, XML and relational databases are integrated into each other, seemingly which is more attractive and popular in the biological data integration [34]. A collection of extended markup languages specifically representing a given type of biological data have been developed. The systems biology markup language (SBML) is a standard exchange format for computational models of biochemical networks [31]. The microarray gene expression markup language (MAGE-ML) is designed to describe microarray designs, microarray experiment designs, gene expression data and data analysis results and is a widely used XML standard for describing and exchanging information among microarray community [35]. Annotated gel markup language (AGML) and human proteome markup language (HUP-ML) are created to describe proteomic analysis data from 2-D gel experiment and mass spectrometry data [5] [36]. The Omic Space Markup Language (OSML) for representing a variety of omic knowledge is used in data integration of several omic domains [37]. However, the difficulty with XML is to accommodate and to represent complex relationships. Furthermore, XML formats, like flat file formats, can be large, and complex, making data access inefficient and impractical in large scale data integration [4] [23].

CHALLENGES OF DATA INTEGRATION IN BIOSCIENCE

The problems of modeling, storing and querying data in bioscience is not solved satisfactorily yet [3]. One of the challenges we face is to represent the relationships in bioscience in a precise and unambiguous manner. Obviously, developing a single global data schema for data integration seems impossible and difficult [4]. One proposal for solving this problem is to develop mediated schemas which focus and represent one domain of knowledge each and to further integrate into a mediated schema to represent the global and complex knowledge domains [4]. This approach is called peer data management system [4]. Although no successful examples have been reported yet it presents a hopeful approach for solving the problem [3]. It is unlikely that one satisfactory solution will solve all the problems in biological data integration. Alternatively, the current data integration efforts should probably focus on identifying the common domains with emphasis on either well studied organisms, such as yeast, or mature knowledge domains, such as genomics related to disease. The successful example of this kind is PharmGBK [33]. Aligning the efforts of data integration will avoid the difficulties in designing global complex data schema and nurture the maturity in challenging issues such as controlled vocabulary, and data representation, as well as providing localized solutions to the demanding needs in biosciences. Secondly, lack of standardizations and centralized authorities for naming systems represent another challenge of data integration in biosciences. Relational data models or object relational data models will ensure the industry standard for fast retrieval and comprehensive access. However, precise, unambiguous relationships must be enforced for relational model and data must be complete and structured. Unfortunately, our understanding of relationships and our knowledge in biosciences are rarely precise and complete. There are sporadic efforts for nomenclature standardizations in molecular biology field. For example, the HUGO gene nomenclature committee focuses on standardizing human gene symbols [38] [39] and the RefSeq database has its own systematic identifiers for their curated non-redundant sequences [40]. The reality is that we are still lacking a centralized authority or conformation for the entire field of biosciences and the semantic inconsistency is bottlenecking our data modeling and representation. The efficiency and accuracy of our data analysis and query formulations are hampered by this reality in existing systems [24]. Thirdly, the visualization of query results in an intuitive way and in a style that appeals to biologists represents an additional challenge considering the complexity of data elements and the diversities of knowledge domains in search results. For example, to visualize the sequence of a particular gene that encodes several isoforms and its protein sequence domains or motifs as well as structures

often requires a user to navigate through many user interfaces. Simplifying this type of multiple-clicks navigations not only represents a challenge to UI designers but also to programmers. My personal experience with query result visualizations comes from my involvement in the development of a commercial product, the Vector PathBlazer™ system in Invitrogen Inc [41], which was an early product tempted to integrate metabolic, signal transduction and regulatory pathways from several legacy sources. One of the major challenges we faced was to present the complicated protein-protein interaction network in an intuitive way and dynamically present the biological pathway in a style that biologists are familiar with without sacrificing the information contents. Vector PathBlazer™ implemented many presentation styles in graphic theories to visualize the network or pathways, but it never reached the point to satisfy the biologist's requirements for many reasons. The complexity of a backend data model could also limit the visualization if the degree of details is not modeled properly.

In conclusion, our understanding of biological systems and relationships among enormous data elements and knowledge domains will be a long time effort. There is no single system that will provide all the solutions in bioinformatics arena now or in the future. Biological data integration efforts in the near future should focus on areas with demanding needs or with relatively matured data accumulations such as model organisms. One of the hot spot has been and will continue to be the integration of clinical data with biological data for the purpose of data-driven biomedical diagnostics, therapy and drug discovery and understanding the molecular mechanisms of the disease. The data-driven genome medicine will be the underlying foundation for individualized health care and surely will improve the efficiency and accuracy for the new drug discovery in the future [42] [43].

Kimball et al. predict that future data warehouses in industry will consist of dozens or hundreds of separates machines with widely different operation systems and database systems [5]. He holds the opinion that these machines can share a uniform architecture of conformed dimensions, which will allow them to be fused into a coherent whole if designed properly [5]. This prophecy probably reflects the future of the biological databases also. However, the particular characteristics of biological databases and user requirements will always have some influence on database models, database management systems and application tools in the future.

ACKNOWLEDGMENT

The author wishes to thank Dr. Howard A. Fine for his support and for his valuable comments. Special thank also goes to Dr. Maarten Leerkes for his reading this article and for his valuable comments.

REFERENCES

[1] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler, "GenBank," *Nucleic Acids Res*, vol. 34, pp. D16-20, 2006.

[2] H. Ge, A. J. Walhout, and M. Vidal, "Integrating 'omic' information: a bridge between genomics and systems biology," *Trends Genet*, vol. 19, pp. 551-60, 2003.

[3] B. Louie, P. Mork, F. Martin-Sanchez, A. Halevy, and P. Tarczy-Hornoch, "Data integration and genomic medicine," *J Biomed Inform*, 2006.

[4] L. D. Stein, "Integrating biological databases," *Nat Rev Genet*, vol. 4, pp. 337-45, 2003.

[5] R. Kimball, L. Veeves, M. Ross, and W. Thornthwaite, *The data warehouse lifecycle toolkit*: John Wiley & Sons, Inc, 1998.

[6] S. B. Davidson, J. Crabtree, B. Brunk, J. Schug, V. Tannen, C. Overton, and C. Stoeckert, "K2/Kleisli and GUS: experiments in integrated access to genomic data sources," *IBM Systems Journal*, vol. 40, pp. 1-23, 2001.

[7] W. H. Inmon, C. Imhoff, and R. Sousa, *Corporate information factory*,, 2nd edition ed: Wiley, 2001.

[8] L. Wong, "Technologies for integrating biological data," *Brief Bioinform*, vol. 3, pp. 389-404, 2002.

[9] International Human Genome Sequencing Consortium, "Finishing the euchromatic sequence of the human genome," *Nature*, vol. 431, pp. 931-45, 2004.

[10] R. Elmasri and B. S. Navathe, *Fundamentals of database systems*: Addison-Wesley, 2000.

[11] 3rd Millennium Inc, "Practical data integration in pharmaceutical R/D: strategies and technologies," in *White paper*, May, 2002.

[12] G. A. Thorisson, A. V. Smith, L. Krishnan, and L. D. Stein, "The International HapMap Project Web site," *Genome Res*, vol. 15, pp. 1592-3, 2005.

[13] M. Pruess, P. Kersey, and R. Apweiler, "The Integr8 project--a resource for genomic and proteomic data," *In Silico Biol*, vol. 5, pp. 179-85, 2005.

[14] M. Cornell, N. W. Paton, C. Hedeler, P. Kirby, D. Delneri, A. Hayes, and S. G. Oliver, "GIMS: an integrated data storage and analysis environment for genomic and functional data," *Yeast*, vol. 20, pp. 1291-306, 2003.

[15] M. Stonebraker, P. Brown, and D. Moore, *The next great wave in DBMS technology. In Object-relational DBMSs: Tracking the next great wave*: Morgan Kaufmann Publisher Inc, 1999.

[16] Nature technology feature, "Bioinformatics: Bring it all together," *Nature*, vol. 419, pp. 751-757, 2002.

[17] A. Bateman, E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S. R. Eddy, S. Griffiths-Jones, K. L. Howe, M. Marshall, and E. L. Sonnhammer, "The Pfam protein families database," *Nucleic Acids Res*, vol. 30, pp. 276-80, 2002.

[18] B. Boeckmann, A. Bairoch, R. Apweiler, M. C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider, "The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003," *Nucleic Acids Res*, vol. 31, pp. 365-70, 2003.

[19] K. R. Christie, S. Weng, R. Balakrishnan, M. C. Costanzo, K. Dolinski, S. S. Dwight, S. R. Engel, B. Feierbach, D. G. Fisk, J. E. Hirschman, E. L. Hong, L. Issel-Tarver, R. Nash, A. Sethuraman, B. Starr, C. L. Theesfeld, R. Andrada, G. Binkley, Q. Dong, C. Lane, M. Schroeder, D. Botstein, and J. M. Cherry, "Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related sequences from other organisms," *Nucleic Acids Res*, vol. 32, pp. D311-4, 2004.

[20] M. C. Costanzo, M. E. Crawford, J. E. Hirschman, J. E. Kranz, P. Olsen, L. S. Robertson, M. S. Skrzypek, B. R. Braun, K. L. Hopkins, P. Kondu, C. Lengieza, J. E. Lew-Smith, M. Tillberg, and J. I. Garrels, "YPD, PombePD and WormPD: model organism volumes of the BioKnowledge library, an integrated resource for protein information," *Nucleic Acids Res*, vol. 29, pp. 75-9, 2001.

[21] E. M. Zdobnov, R. Lopez, R. Apweiler, and T. Etzold, "The EBI SRS server--new features," *Bioinformatics*, vol. 18, pp. 1149-1150, 2002.

[22] L. M. Haas, J. E. Rice, P. M. Schwarz, W. C. Swope, P. Kodali, and E. Kotlar, "DiscoveryLink: A system for integrated access to life sciences data sources," *IBM Systems Journal*, vol. 40, pp. 489-511, 2001.

[23] I. M. Chen and V. M. Markowitz, "An overview of the object-protocol model (OPM) and OPM data management tools," *Informatics System*, vol. 20, pp. 395-418, 1995.

[24] R. Nagarajan, M. Ahmed, and A. Phatak, "Database challenges in the integration of biomedical data sets," *Proceedings of the 30th VLDB conference, Toronto, Canada, 2004*, pp. 1202-1213, 2004.

[25] D. Karolchik, R. Baertsch, M. Diekhans, T. S. Furey, A. Hinrichs, Y. T. Lu, K. M. Roskin, M. Schwartz, C. W. Sugnet, D. J. Thomas,

R. J. Weber, D. Haussler, and W. J. Kent, "The UCSC Genome Browser Database," *Nucleic Acids Res*, vol. 31, pp. 51-4, 2003.
[26] REMBRANDT, "REMBRANDT: enpowering the translational research for brain tumor studies."
[27] P. J. Kersey, L. Morris, H. Hermjakob, and R. Apweiler, "Integr8: enhanced inter-operability of European molecular biology databases," *Methods Inf Med*, vol. 42, pp. 154-60, 2003.
[28] P. A. Covitz, F. Hartel, C. Schaefer, S. De Coronado, G. Fragoso, H. Sahni, S. Gustafson, and K. H. Buetow, "caCORE: a common infrastructure for cancer informatics," *Bioinformatics*, vol. 19, pp. 2404-12, 2003.
[29] M. V. Mannino, *Database design, application development and administration*: McGraw-Hill Irwin.
[30] X. Wang, R. Gorlitsky, and J. S. Almeida, "From XML to RDF: how semantic web technologies will change the design of 'omic' standards," *Nat Biotechnol*, vol. 23, pp. 1099-103, 2005.
[31] A. Finney and M. Hucka, "Systems biology markup language: Level 2 and beyond," *Biochem Soc Trans*, vol. 31, pp. 1472-3, 2003.
[32] M. Hewett, D. E. Oliver, D. L. Rubin, K. L. Easton, J. M. Stuart, R. B. Altman, and T. E. Klein, "PharmGKB: the Pharmacogenetics Knowledge Base," *Nucleic Acids Res*, vol. 30, pp. 163-5, 2002.
[33] T. E. Klein and R. B. Altman, "PharmGKB: the pharmacogenetics and pharmacogenomics knowledge base," *Pharmacogenomics J*, vol. 4, pp. 1, 2004.
[34] F. Achard, G. Vaysseix, and E. Barillot, "XML, bioinformatics and data integration," *Bioinformatics*, vol. 17, pp. 115-25, 2001.
[35] P. T. Spellman, M. Miller, J. Stewart, C. Troup, U. Sarkans, S. Chervitz, D. Bernhart, G. Sherlock, C. Ball, M. Lepage, M. Swiatek, W. L. Marks, J. Goncalves, S. Markel, D. Iordan, M. Shojatalab, A. Pizarro, J. White, R. Hubley, E. Deutsch, M. Senger, B. J. Aronow, A. Robinson, D. Bassett, C. J. Stoeckert, Jr., and A. Brazma, "Design and implementation of microarray gene expression markup language (MAGE-ML)," *Genome Biol*, vol. 3, pp. RESEARCH0046, 2002.
[36] R. Stanislaus, C. Chen, J. Franklin, J. Arthur, and J. S. Almeida, "AGML Central: web based gel proteomic infrastructure," *Bioinformatics*, vol. 21, pp. 1754-7, 2005.
[37] Y. Hasegawa, M. Seki, Y. Mochizuki, N. Heida, K. Hirosawa, N. Okamoto, T. Sakurai, M. Satou, K. Akiyama, K. Iida, K. Lee, S. Kanaya, T. Demura, K. Shinozaki, A. Konagaya, and T. Toyoda, "A flexible representation of omic knowledge for thorough analysis of microarray data," *Plant Methods*, vol. 2, pp. 5, 2006.
[38] H. M. Wain, M. J. Lush, F. Ducluzeau, V. K. Khodiyar, and S. Povey, "Genew: the Human Gene Nomenclature Database, 2004 updates," *Nucleic Acids Res*, vol. 32, pp. D255-7, 2004.
[39] R. B. MacIntosh, "Hugo Obwegeser: forty years later," *N Y State Dent J*, vol. 71, pp. 42-4, 2005.
[40] K. D. Pruitt, T. Tatusova, and D. R. Maglott, "NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins," *Nucleic Acids Res*, vol. 33, pp. D501-4, 2005.
[41] Informax, "For biological pathways analysis Vector PathBlazer," 2004.
[42] C. Sander, "Genomic medicine and the future of health care," *Science*, vol. 287, pp. 1977-8, 2000.
[43] H. F. Willard, M. Angrist, and G. S. Ginsburg, "Genomic medicine: genetic variation and its impact on the future of health care," *Philos Trans R Soc Lond B Biol Sci*, vol. 360, pp. 1543-50, 2005.

Table 1. Examples of molecular databases by categories in biosciences

Category	Names	Databases Contents	Types	URL
Genes & transcript	EMBL	DNA sequences and derived protein sequences hosted by EBI	Flat file	http://www.ebi.ac.uk/embl/
	UniGene	Clustering of human, mouse, and rat DNA and EST sequences into gene-oriented, non-redundant clusters hosted by NCBI	ASN.1	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene
Gene expr.	GEO	A gene expression and hybridization array repository system hosted by NCBI	R[1]	http://www.ncbi.nlm.nih.gov/geo/
	ArrayExpress	A public microarray data repository system hosted by EBI	R[1]	http://www.ncbi.nlm.nih.gov/geo/
Genome	MGD	Mouse genome data hosted by Jackson Laboratory	Flat file	http://www.informatics.jax.org/
	UCSC	Human, mouse, rat genomic sequences hosted by UCSC	R[1]	http://genome.ucsc.edu/
Protein	SWISS-PROT	Protein sequences and annotation hosted by EBI	Flat file	http://us.expasy.org/
Protein domains & motifs	InterPro	Protein families, domains and functional sites in which identifiable features found in known proteins hosted by EBI	R[1]	http://www.ebi.ac.uk/interpro/
	Pfam	Protein sequence alignments and domain profiles hosted by Sanger Institute	R[1]	http://pfam.wustl.edu/
PPI	BIND	Biomolecular interaction database hosted by Blueprint Institute of Mount Sanai Hospical, Canada	Object	http://www.bind.ca/
	DIP	Protein-protein interaction database hosted by Harward Hughes Medical Institute	R[1]	http://dip.doe-mbi.ucla.edu/
Protein structures	MMDB	Curated protein structures, related sequences and literatures hosted by NCBI	Flat file	http://www.ncbi.nlm.nih.gov/Structure/
	PDB	Experimentally determined 3-D structures of biological macromolecules hosted by RCSB	R[1]	http://www.rcsb.org/pdb/
Pathways	TransPath	Signal transduction pathways and reactions hosted by BioBase	Object	http://www.biobase.de/pages/products/transpath.html

[1]: relational database

Table 2. Classification of molecular databases

	Name	Pri	Sec	D	B	PS	GS	Rep	Bro	Vis	Query	Ana
Genes & transc	GenBank	✓			✓	✓		✓	✓	✓	✓	
	EMBL	✓			✓	✓		✓	✓	✓	✓	
	DDBJ	✓			✓	✓		✓	✓	✓	✓	
	RefSeq		✓		✓	✓			✓	✓	✓	
	UniGene		✓		✓	✓			✓	✓	✓	
	dbEST	✓			✓	✓		✓	✓	✓	✓	
Genomes	Flybase	✓		✓		✓		✓	✓	✓	✓	
	MGD	✓		✓			✓	✓	✓	✓	✓	
	SGD	✓		✓			✓	✓	✓	✓	✓	
	UCSC		✓				✓		✓	✓	✓	
Expr	GEO	✓			✓		✓	✓	✓	✓	✓	
	ArrayExpress	✓			✓		✓	✓	✓	✓	✓	✓
	REMBRANDT	✓		✓		✓		✓	✓	✓	✓	✓
	GXD	✓		✓		✓		✓	✓	✓	✓	
Proteins	Swissprot	✓			✓	✓		✓	✓	✓	✓	
	Trembl	✓			✓	✓		✓	✓	✓	✓	
	Enzyme	✓			✓	✓		✓	✓	✓	✓	
	UniProt	✓			✓	✓		✓	✓	✓	✓	
	MIPS	✓			✓	✓		✓	✓	✓	✓	
	BRENDA	✓			✓	✓			✓	✓	✓	✓
Nome	HUGO	✓		✓		✓			✓	✓	✓	
	Enzyme nome.	✓			✓	✓		✓	✓	✓	✓	
P seq. & motifs	InterPro		✓		✓	✓			✓	✓	✓	✓
	Prosite	✓			✓	✓			✓	✓	✓	
	ProDom	✓			✓	✓			✓	✓	✓	
	SMARTS	✓			✓	✓			✓	✓	✓	
	PRINTS	✓			✓	✓			✓	✓	✓	
	BLOCKS	✓			✓	✓			✓	✓	✓	
	PFAM								✓	✓	✓	
Protein structure	PDB	✓		✓		✓		✓	✓	✓	✓	
	MMDB		✓		✓	✓		✓	✓	✓	✓	
	FSSP/Dali		✓		✓	✓			✓	✓	✓	
	SCOP		✓		✓	✓			✓	✓	✓	
	CATH		✓		✓	✓			✓	✓	✓	
	HSSP								✓	✓	✓	
PPI	YPD		✓	✓		✓		✓	✓	✓	✓	
	BIND	✓			✓	✓		✓	✓	✓	✓	
	DIP	✓			✓	✓		✓	✓	✓	✓	
	MINT	✓		✓		✓		✓	✓	✓	✓	
Pathways	PathDB	✓			✓	✓		✓	✓	✓	✓	✓
	TransPath	✓			✓	✓			✓	✓	✓	✓
	Kegg		✓		✓	✓		✓	✓	✓	✓	
	MPW	✓			✓	✓		✓	✓	✓	✓	
	UM-BBD	✓			✓	✓		✓	✓	✓	✓	
Annot	Gene Ontology	✓			✓		✓	✓	✓	✓	✓	
	Taxonomy		✓		✓	✓		✓	✓	✓	✓	
	OMIM	✓			✓	✓		✓	✓	✓	✓	

Note: the classifications are based on design goals and the contents of the databases as well as the applications on the databases. Abbreviations in table: pri-primary database, sec-secondary database, D-deep database, B-broad database, PS- point solution, GS-general solution, rep-repository, bro-browser, vis-visualizing, ana-analysis, transc-transcripts, expr-expression, nome-nomenclature, ppi-protein-protein interaction, annot-annotation.