Chapter 4
Data Warehousing of Life Science Data

Benjamin Kormeier and Klaus Hippe

Abstract Increasingly, scientist have begun to collect biological data in different information systems and database systems that are accessible via the internet, which offer a wide range of molecular and medical information. Regarding the human genome data, one important application of information systems is the reconstruction of molecular knowledge for life science data. In this review paper, we will discuss major problems in database integration and present an overview of important information systems. Furthermore, we will discuss the information reconstruction and visualization process based on that integrated life science data. These database integration tools will allow the prediction for instance of protein–protein networks and complex metabolic networks.

Keywords Data warehouse · Life science · Database integration

4.1 Introduction

The diverse research areas of molecular biology generate a variety of publicly available data stored in molecular biology databases. These databases are global via information systems and are mostly publicly available. In recent years, the number of molecular biology databases has increased exponentially. There are currently 1641 databases providing information from different categories (Rigden and Fernández 2020).

The importance of data integration has been known in bioinformatics for several years. Therefore, it is essential for scientists to analyze and process information from different and distributed systems. The molecular biological data has a high degree of semantic heterogeneity because the data comes from a series of experiments. In molecular biology, complex problems are tackled that rely on an immense and

B. Kormeier (✉) · K. Hippe
FH Bielefeld, University of Applied Sciences, Interaktion 1, Bielefeld, Germany
e-mail: bkormeie@techfak.uni-bielefeld.de

M. Chen, R. Hofestädt (eds.), *Integrative Bioinformatics*,
https://doi.org/10.1007/978-981-16-6795-4_4

diverse amount of data. The number of databases and the data they contain are increasing steadily, which means that data distribution and high redundancy cannot be excluded. For these reasons, it is important to develop data warehouse systems for keeping consistent and non-redundant data.

4.1.1 Aims and Scope

The integration of life science and biological data from heterogeneous, autonomous, and distributed data sources is an important task in bioinformatics. The challenge is to integrate huge data sets regarding the large heterogeneity of the databases on the semantic and technical level (Kormeier 2010). Therefore, relevant integration approaches in the field of data warehouses as well as modeling and simulation software approaches will be introduced. We will focus on several widely used data warehouse approaches. Furthermore, some selected tools for modeling and visualizing of biochemical pathways will be presented in this review paper.

4.2 Molecular Database Integration

The integration of data sets from data sources with different heterogeneities is a challenge not only in the economy but also in research and science. Especially, in the life sciences, numerous biological datasets are experimentally generated, which have significant heterogeneities in various domains. The storage, deployment, and administration of these data are usually done by molecular biology databases. Usually, these databases are freely available, distributed worldwide, and linked together by explicit cross-references. There are also significant differences in the structuring of data, accessibility, and copyright. One aspect of bioinformatics is the implementation of applications with which help an effective data integration of molecular biology databases are made possible. The goal of the data integration is to realize a database that has a uniform data structure and provides all the necessary data from the data sources. The data sources usually have different schemas, which is why schema transformation and schema integration are necessary. After that, the actual integration of the data stocks from the respective data sources takes place. The data is analyzed and validated so that inconsistencies and duplicates are identified and eliminated. During this data cleanup, merging and completion of incomplete data sets can also be done. As a result of this data fusion, a complete data set is realized to provide more information than the original data records from the data sources. The resulting consistent and structured database enables an efficient and global view of all data sources from the data sources. However, the merging of databases from different data sources is linked to three basic problems that will be described in the following sections.

4.2.1 Distribution, Autonomy, and Heterogeneity

With the help of specific software solutions, the integration of data from different data sources is realized. Such systems usually have different integration architectures, which successfully overcome the three basic problems of data integration. The distribution, autonomy and heterogeneity of a data source represent these basic problems and are also described as an orthogonal dimension of data integration (Leser and Naumann 2007).

One problem that needs to be addressed in data integration is the global distribution of data sources. Usually, the databases are provided by different systems and are geographically distributed. Because of this different localization of the data, a distinction is made between the physical and the logical distribution. With the help of a materialized integration architecture, the problem of physical distribution can be overcome. The provision of metadata and data cleansing methods by the integration system enables the removal of the logical distribution.

The autonomy of a data source is usually unavoidable, because the responsible organization of a data source usually uses its own development strategies and technologies. The term autonomy in connection with the data integration means that the data source can autonomously decide on the provision, the access possibilities, and the copyright of the data. In addition, autonomy is responsible for different problems of heterogeneity. In Conrad (1997), the different types of autonomy are discussed in detail.

The main problem that needs to be addressed in data integration is heterogeneity. If two information systems do not provide identical methods, models, and structures for accessing the database, these are called heterogeneous. Different kinds of heterogeneity are defined according to (Leser and Naumann 2007) as follows: technical heterogeneity, syntactic heterogeneity, data model heterogeneity, structural heterogeneity, schematic heterogeneity, semantic heterogeneity (Kormeier 2010). Autonomy is primarily responsible for heterogeneity, but distribution can also create heterogeneity. It is possible to force specific properties to be homogenous by restricting autonomy of a data source. This can be achieved by standards in exchange formats, interfaces, and protocols.

4.2.2 Approaches of Database Integration

The development of an integrated database system is a complex task, particularly, when a large number of heterogeneous databases have to be integrated. Hence, an elaborate blueprint of the architecture of the system is essential. However, another non-trivial problem is the availability of databases that should be integrated. Generally, there exists two architectures for integration. They are divided into materialized integration and virtual integration (Kormeier 2010). The main difference between the two integration architectures is the location of the relevant databases during integration. A materialized integration architecture is a central and persistent database and copies all the necessary data from the data sources into the database. In contrast, a virtual integration architecture does not have such a database and therefore does not copy any data. Therefore, the integrated and homogenous data set of a virtual integration architecture only exists virtually and has to be realized again for all requests. However, there are also hybrid architectures that have materialized and virtual data sets.

Different approaches of database integration have been frequently discussed and reviewed since the beginning of the millennium. The most important are the following three approaches besides data warehouses:

- Hypertext navigation systems. HTML frontends linked to molecular biological databases.
- Federated database systems and mediator-based systems. Virtual integration does not store any data in a global schema. Federated systems integrate multiple autonomous database systems into a virtual single federated database. Usually, each database is interconnected via a computer network. The databases may be geographically decentralized. In comparison to federated database systems, multi-database systems do not have a global schema, rather these systems interactively generate queries for several databases at the same time (Kormeier 2010).

4.2.3 Data Warehouses (DWH)

In this section, we want to have a closer look at data warehouses. Data warehouses are one of the widely used architectures of materialized integration. Usually, data warehouses are used in the field of information management. In particular data analysis, data mining and long-term storage of business intelligence in companies are the major advantages of data warehouse systems. In bioinformatics data, warehouses are usually used for data integration (Kormeier 2010). There is no consistent definition of the DWH term. While different consortia such as the OLAP Council are trying to standardize the DWH term, the first definition was given by Inmon (1996):

> *A data warehouse is a subject oriented, integrated, non-volatile, and time variant collection of data support of management's decision.*

Only through the data warehouse process can a data warehouse system accomplish the various issues. This dynamic process is responsible for the acquisition, storage and analysis of the data. The data warehouse process can be divided into the following four steps:

1. In the first step, the component extraction, transformation and loading are used, which are summarized under the term ETL components (Günzel and Bauer 2009). This step is called the ETL process and is responsible for extracting the data sets from the data sources and transforming them. Furthermore, the ETL process is responsible for loading the structured data into the DWH.
2. The persistent storage of data in DWH will be realized in the next step. However, some analyzes or specialized applications do not need all the data, so that can be realized in the so-called data marts.
3. These data marts represent a specific view of the DWH and are created in the third step.
4. The analysis and evaluation of the databases take place in the last step. The results are then provided to the different applications.

The key benefits of the materialized integration architecture are efficient data cleansing, unrestricted query capabilities, and good query performance. The disadvantage of this integration architecture may under certain circumstances be the timeliness of the database. However, this aspect always has to be considered in the context of the respective analysis or question, because not every topic needs up-to-date data. The relevance of the data is particularly important for complex analyzes of the financial markets. In the context of molecular biology research, updating the database every quarter is sufficient. The data sources are usually molecular biology databases and their updating is usually done every quarter.

4.3 Related Data Integration Approaches

In this chapter, relevant integration approaches in the field of data warehouses will be introduced. Furthermore, related visualization approaches for molecular networks and life science data will be discussed.

4.3.1 Data Integration Approaches

In the literature, data integration approaches in bioinformatics are divided into the following classes (Leser and Naumann 2007):

- Indexing systems: SRS (Sequence Retrieval System) (Etzold et al. 1996), Entrez (Kaps et al. 2006), and BioRS (Wheeler et al. 2004; Maglott et al. 2007).
- Multi-databases: OPM (Object Protocol Model) (Chen and Markowitz 1995), DiscoveryLink (Haas et al. 2001), and BioKleisli (Davidson et al. 1997).

- Ontology-based integration: TAMBIS (Transparent Access to Multiple Bioinformatics Information Sources) (Stevens et al. 2000), ONDEX (Köhler et al. 2006), and CoryneRegNet (Pauling et al. 2012).
- Data warehouse: Atlas (Shah et al. 2005), BioWarehouse (Lee et al. 2006), Columba (Trissl et al. 2005), Biozon (Birkland and Yona 2006), Booly (Do et al. 2010), JBioWH (Vera et al. 2013), Unison (Hart and Mukhyala 2009), and SYSTOMONAS (Choi et al. 2007). However, under certain aspects, CoryneRegNet, and ONDEX can also be assigned to this category.

Due to the already mentioned advantages (e.g., performance, availability of data, and simple conception), the data warehouse technology has become established in bioinformatics. Most of the applications were developed for specific molecular-biological questions, which means that they could not be used in other projects and their questions, or only through extensive extensions of the respective software solution. Atlas, BioWarehouse, Columba, ONDEX, and CoryneRegNet use the data warehouse technique for data integration, whereas CoryneRegNet and ONDEX provide a web service. Atlas, BioWarehouse, and ONDEX provide a software infrastructure for data integration, rather Columba, CoryneRegNet. They provide a web interface and therefore they are directly useable (Kormeier 2010). In addition, the database of many systems is out of date or no longer available, so important information is not available to the user. In particular, the complexity and flexibility of the respective software as well as the attitude of the project financing are responsible for it. In recent years, a plenty of systems have been implemented and made available to the user.

4.4 Data Warehouse for Life Science Data Warehouse

In the previous sections of this chapter, several principles and approaches for database integration and network visualization were introduced. A couple of the principles of the introduced integration systems are well suited to be used within the database integration system that will be presented. Particularly the functions of the software toolkit BioDWH (Töpel et al. 2008) will be illustrated. Furthermore, DAWIS-M.D. 2.0 (Hippe et al. 2011), a web-based data warehouse approaches based on the BioDWH integration toolkit, will be described.

4.4.1 BioDWH: A General Data Warehouse Infrastructure for Life Science Data Integration

BioDWH is an open source software toolkit, which can be used as a general infrastructure for integrative bioinformatics research and development. The advantages of the approach are realized by using a Java-based system architecture and

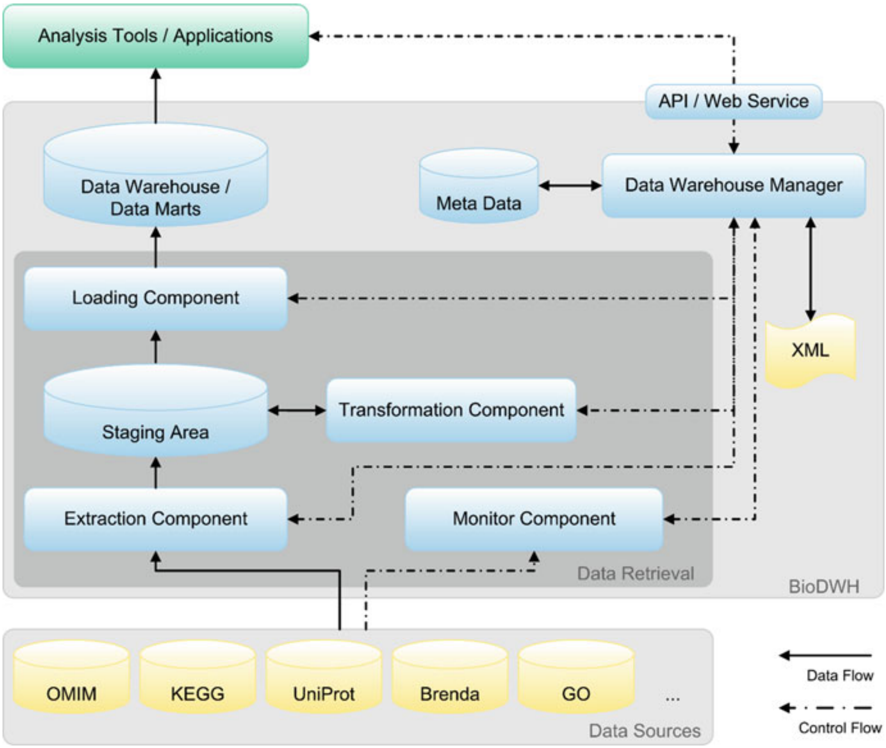

Fig. 4.1 Schematic illustration of the BioDWH system architecture following the general data warehouse design (Hippe 2014)

object-relational mapping (ORM) technology (Kormeier 2010). Figure 4.1 shows the reference architecture of a data warehouse system. This architecture is the foundation of the system architecture of BioDWH. Basically, the system consists of the Data Retrieval module, the Data Warehouse Manager, and a Graphical User Interface (GUI). The user is able to control the infrastructure via GUI and XML configuration files. Core of the system is the Data Retrieval component that is composed of Loading, Transformation, Extraction Component, i.e., the Parser (ETL), Monitor component, and the Staging Area. The parser library provides a large number of ready-to-use-parsers for biological and life science databases which are available, such as UniProt, KEGG, OMIM, GO, Enzyme, BRENDA, OMIM, Reactome, iProClass, and more. Using the BioDWHParser interface, it is easy to create own tailored parser. To achieve independence from the RDBMS, a persistence layer is necessary. Therefore, a well-engineered object-relational mapping framework called Hibernate was used as a persistence layer. Hibernate performs well and is independent from manufacturers like MySQL, PostgreSQL, or Oracle. Thus, the Hibernate framework fits perfectly into the infrastructure of the BioDWH application.

The system is realized as a Java-based open source application that is supported on different platforms with an installed Java Runtime Environment (JRE). Nowadays, Java is very popular and usually installed on most of the computers. Additionally, Java is available on most platforms such as Windows, Linux, and MacOS. Thus, Java applications have a high degree of platform independence. Moreover, Java applications over flexible software solutions that can be provided to a large audience. In this way the software solutions can become widely used (Kormeier 2010).

Another feature of BioWH is an implemented easy-to-use Project Wizard that supports the user or administrator to configure a DWH integration process in four steps. No additional knowledge in database systems or computer science is necessary. The whole configuration starting from database connection settings, via parser configuration to monitor configuration, is supported by the graphical user interface. In background the BioDWH infrastructure is running with multiple threads which means it is possible to run several download processes, uncompress processes, or integration processes in parallel (Kormeier 2010). Finally, a logging mechanism watches the integration process and starts a simple recovery process to guarantee a consistent state of the data warehouse.

4.4.2 DAWIS-M.D. 2.0: A Data Warehouse System for Metabolic Data

One of the major challenges in bioinfomatics is the integration and management of data from different sources and their presentation in a user-friendly format. DAWIS-M.D. 2.0 is a platform-independent data warehouse information system for metabolic data. The information system integrates data from 13 widely used life science databases (KEGG, EMBL-Bank, Transfac, Transpath, SCOP, JASPAR, EPD, UniProt, HPRD, GO, BRENDA, ENZYME, and OMIM). The information of integrated databases is divided into 13 various biological domains (Compound, Disease, Drug, Enzyme, Gene, Gene Ontology, Genome, Glycan, Pathway, Protein, Reaction, Reactant Pair, and Transcription Factor), which are available via the graphical user interface of the web application. The data warehouse architecture (Fig. 4.2) provides a platform-independent web interface that can be used with any common web browser. The system enables intuitive search of integrated life science data, simple navigation to related information as well as visualization of biological domains and their relationships. To ensure maximum up-to-dateness of the integrated data the BioDWH data warehouse infrastructure including a monitor component is used. The persistence layer of DAWIS-M.D. 2.0 uses the ORM technique, whereby the application layer is independent from database layer. Thereby, it is possible to support different database management systems. The DAWIS-M.D. 2.0 data warehouse incorporates the advantages of a navigation and informational system and builds a bridge to the network editor approach VANESA

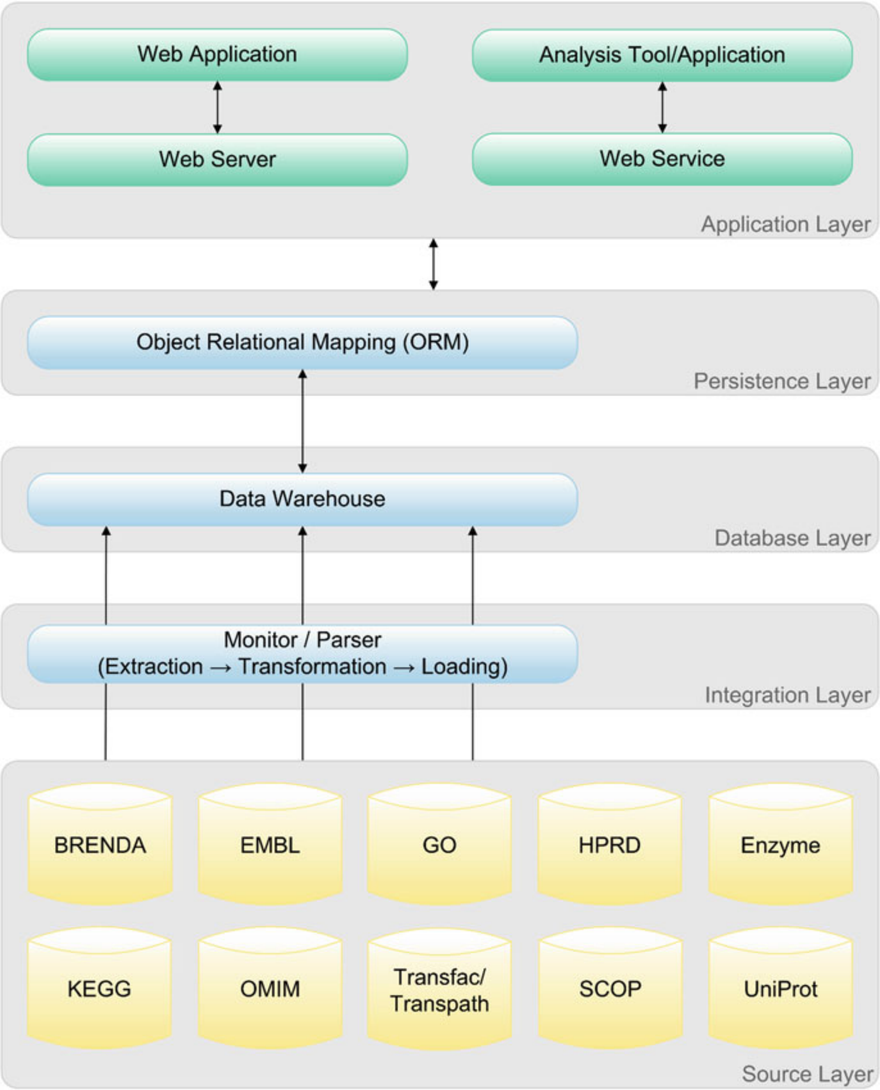

Fig. 4.2 Schematic representation of the DAWIS-M.D. n-layer system architecture from the original heterogeneous data sources to the web application layer (Hippe 2014)

(Brinkrolf et al. 2014). Hence, it is possible to browse through the integrated life science data and bring the information into a modeling and visualization environment. Therefore, it is easy for the scientists to search information of interest, find relationships and interactions between different biomedical domains and bring them for editing, manipulation, and analyzing directly into the VANESA network editor. Finally, the scientists gain a better understanding of complex biological problems and are able to develop new theoretical models for further experiments. DAWIS-M.D. 2.0 is available at https://agbi.techfak.uni-bielefeld.de/DAWISMD/ (Hippe et al. 2011).

4.5 Summary

Different research domains of life sciences by different experimental methods generate an immense and diverse amount of data. Usually, such data is stored in database systems. For a comprehensive and efficient usage of the data, it is necessary to integrate the distributed and heterogeneous data and provide them for further analysis to the researcher. Moreover, the user needs to be supported by applicable tools for navigation within the integrated data sets that support an efficient and precise processing of the data. The number of molecular databases has been continuously increasing in the last decade (Töpel et al. 2008). Today, approximately 1641 publicly available databases and information systems for life science data are listed in the NAR catalogue (Rigden and Fernández 2020). This is mainly due to technological progress and computer-aided laboratory automation.

Transparency, integrity, semantic correctness, and non-redundancy are classical requirements of integration and therefore very important. However, other requirements gain importance in life science data integration, such as an efficient access to the increasing amount of data which should be, but is not always up-to-date. Furthermore, solutions for complex and changing schemata in life science data are required. Hence, the challenge was to combine diverse and multiple data and to bring them into a homogenous, consistent state. The new system should be flexible and also applicable in general for any other project. For that purpose, BioDWH data warehouse software kit is developed as a Java-based open source application for building life science data warehouses using common relational database management systems. By using the object-relational mapping (ORM) technology, it is no longer necessary to select the local database management system based on the restrictions of the integration software. BioDWH provides a number of ready-to-use parsers to extract data from public life science data sources and to store the information in a data warehouse. The integration process is supported by an easy-to-use graphical user interface that makes it possible to integrate any supported database in a few steps into a local database (Töpel et al. 2008).

DAWIS-M.D. 2.0 is a publicly available web-based system that integrates data from 13 different biomedical databases and divided the integrated data from the different data sources into 13 biomedical domains (Hippe et al. 2011). This data warehouse information system provides an integrated and consistent view of large-scale biomedical data. Additionally, relationships and interactions between multiple data sets and biomedical domains are identified and displayed (Janowski 2013). The advantages of the DAWIS-M.D. 2.0 application are the usability, performance, high level of platform independence, and wide range of life sciences information and biological knowledge (Hippe et al. 2011). Furthermore, the system is connected by a "remote-control" to the VANESA network editor to easily visualize and analyze biological networks from data of interest.

Software solutions that provide visualization, analysis services, and an information management framework are in high demand among scientist as already discussed. It is not surprising that many groups over the world have contributed to the task of developing such software frameworks. Therefore, a DWH system to search integrated life science data and simple navigation called DAWIS-M.D. 2.0 as a base for a modeling and visualization system called VANESA were implemented (Hippe et al. 2011).

In conclusion, in this chapter, we presents a powerful and flexible data warehouse infrastructure BioDWH that can be used for building project-specific information systems, such as DAWIS-M.D. 2.0. Finally, the system was the basis for network modeling and pathway reconstruction in different scientific projects. The presented applications are in use since more than one decade within several projects as well as in ongoing in-house projects.

Acknowledgments This work could not have been carried out without the support of many people at Bioinformatics/Medical Informatics Department of Bielefeld University.

References

Birkland A, Yona G (2006) BIOZON: a system for unification, management and analysis of heterogeneous biological data. BMC Bioinform 7(1):70

Brinkrolf C, Janowski SJ, Kormeier B, Lewinski M, Hippe K, Borck D, Hofestädt R (2014) VANESA—a software application for the visualization and analysis of networks in system biology applications. J Integr Bioinform 11(2):239

Chen IMA, Markowitz VM (1995) An overview of the object protocol model (opm) and the opm data management tools. Inf Syst 20(5):393418

Choi C, Munch R, Leupold S, Klein J, Siegel I, Thielen B, Benkert B, Kucklick M, Schobert M, Barthelmes J, Ebeling C, Haddad I, Scheer M, Grote A, Hiller K, Bunk B, Schreiber K, Retter I, Schomburg D, Jahn D (2007) SYSTOMONASan integrated database for systems biology analysis of pseudomonas. Nucleic Acids Res 35:D533537

Conrad S (1997) Föderierte Datenbanksysteme—Konzepte der Datenintegration. Springer, Berlin

Davidson SB, Overton GC, Tannen V, Wong L (1997) BioKleisli: a digital library for biomedical researchers. Int J Digit Libr 1(1):36–53

Do L, Esteves F, Karten H, Bier E (2010) Booly: a new data integration platform. BMC Bioinform 11(1):513

Etzold T, Ulyanov A, Argos P (1996) SRS: information retrieval system for molecular biology data banks. Methods Enzymol 266:114128

Günzel H, Bauer A (2009) Data-Warehouse-Systeme. dpunkt.verlag, Heidelberg

Haas LM, Schwarz PM, Kodali P, Kotlar E, Rice JE, Swope WC (2001) Discoverylink: a system for integrated access to life sciences data sources. IBM Syst J 40(2):489511

Hart RK, Mukhyala K (2009) Unison: an integrated platform for computational biology discovery. In: Pacific symposium on biocomputing, pp 403–414

Hippe K (2014) Identifikation von potenziellen Transkriptionsfaktorbindestellen in Nukleotidsequenzen basierend auf einem Data-Warehouse-System. Bielefeld University, Bielefeld

Hippe K, Kormeier B, Janowski SJ, Töpel T, Hofestädt R, DAWIS-M.D. (2011) 2.0—a data warehouse information system for metabolic data. In: Proceedings of the 7th International Symposium on Integrative Bioinformatics

Inmon WH (1996) Building the data warehouse. Wiley, Indianapolis

Janowski SJ (2013) VANESA—a bioinformatics software application for the modeling, visualization, analysis, and simulation of biological networks in systems biology applications. Bielefeld University, Bielefeld

Kaps A, Dyshlevoi K, Heumann K, Jost R, Kontodinas I, Wolff M, Hani J (2006) The BioRS(TM) integration and retrieval system: an open system for distributed data integration. J Integr Bioinform 3(2)

Köhler J, Baumbach J, Taubert J, Specht M, Skusa A, Rüegg A, Rawlings C, Verrier P, Philippi S (2006) Graph-based analysis and visualization of experimental results with ONDEX. Bioinformatics 22:13831390

Kormeier B (2010) Semi-automated reconstruction of biological networks based on a life science data warehouse. Bielefeld University, Bielefeld

Lee TJ, Pouliot Y, Wagner V, Gupta P, Stringer-Calvert DW, Tenenbaum JD, Karp PD (2006) BioWarehouse: a bioinformatics database warehouse toolkit. BMC Bioinform 7:170

Leser U, Naumann F (2007) Informationsintegration. dpunkt Verlag, Heidelberg

Maglott D, Ostell J, Pruitt KD, Tatusova T (2007) Entrez gene: gene-centered information at NCBI. Nucleic Acids Res 35(suppl 1):D26–D31

Pauling J, Röttger R, Tauch A, Azevedo V, Baumbach J (2012) Coryneregnet 6.0—updated database content, new analysis methods and novel features focusing on community demands. Nucleic Acids Res 40:D610–D614

Rigden DJ, Fernández XM (2020) The 2021 Nucleic Acids Research database issue and the online molecular biology database collection. Nucleic Acids Res 49(D1):D1–D9. https://doi.org/10.1093/nar/gkaa1216

Shah SP, Huang Y, Xu T, Yuen MM, Ling J, Ouellette BF (2005) Atlas—a data warehouse for integrative bioinformatics. BMC Bioinform 6:34

Stevens R, Baker P, Bechhofer S, Ng G, Jacoby A, Paton NW, Goble CA, Brass A (2000) TAMBIS: transparent access to multiple bioinformatics information sources. Bioinformatics 16:184185

Töpel T, Kormeier B, Klassen A, Hofestädt R (2008) BioDWH: a data warehouse kit for life science data integration. J Integr Bioinform 5(2):93

Trissl S, Rother K, Müller H, Steinke T, Koch I, Preissner R, Frömmel C, Leser U (2005) Columba: an integrated database of proteins, structures, and annotations. BMC Bioinform 6:81

Vera R, Perez-Riverol Y, Perez S, Ligeti B, Kertész-Farkas A, Pongor S (2013) JBioWH: an open-source Java framework for bioinformatics data integration. Database 2013:bat051

Wheeler DL, Church DM, Edgar R, Federhen S, Helmberg W, Madden TL, Pontius JU, Schuler GD et al (2004) Database resources of the National Center for biotechnology information: update. Nucleic Acids Res 32(suppl 1):D35–D40