Data Integration in Genomic Medicine: Trends and Applications

Contribution of the IMIA Working Group on Informatics in Genomic Medicine

J. A. Seoane, J. Dorado¹, V. Aguiar-Pulido, A. Pazos

¹ IMIA Informatics in Genomic Medicine Working Group Chair

Information and Communications Technologies Department, School of Computer Science, University of A Coruña A Coruña, Spain

Summary

Objectives: In a near future, each person will incorporate his/her own sequenced genome in his/her electronic health record. In that precise moment, genomic medicine will be fundamental for clinical practice, as an essential key of personalized medicine. All the genomic data, as well as other 'omics' and clinical data necessary for personalized medicine, are stored in several distributed databases. Research and patient care require each time more biomedical data integration of several distributed heterogeneous datasources.

Methods: This work develops a comprehensive review of the most relevant works in biomedical data integration, specifically in genomic medical data, analyzing the evolution of architecture and integration techniques during the last 20 years, and its usage.

Conclusion: Most of these solutions, based on cross-linking, data warehouse or federated approaches, are suitable for specific domains. However, none of the models found in the literature is completely appropriate for a general biomedical data integration problem.

Keywords

Genetic database, integrated advanced information management systems, information storage and retrieval, federated systems, data warehouse

Yearb Med Inform 2012;117-25

Introduction

Exponential accumulation of genomic and biomedical data has the promise to advance biomedical knowledge. This would allow the development of new generations of personalized biomedicine.

Each year, Nucleid Acid Research publishes a Web Server Issue [1] with a list of Bioinformatics Links Directory, with a comprehensive list of web servers and databases that store, visualize, integrate and analyze biomedical data. In the last 2011 number, there are nearly 1900 datasources. Biomedical researchers waste important amounts of time querying several datasources in order to obtain the information they need.

Many researches feel overwhelmed by task that involve searching and integrating data from heterogeneous databases because they must navigate through different databases and select or filter the data they require. The complexity of this task indicates the need of developing biomedical data integration systems which allow simplifying this task.

Integrating data consists essentially in making queries on different data sources, which can be either related databases or semi-structured data sources connected to a network, then join the data and present it in a homogeneous way. There is a lot of literature [2-20] about data integration in the

area of biomedical informatics and informatics. In this review, technologies that can be used to solve data integration problems in genomic medicine are described.

Background

Data integration has been a challenge in computer science during the past 30 years and especially in genomic medicine, the past 20 years. There are several approaches to data integration, but they can be mainly grouped in three categories: 1) The link integration approaches, where there is a web server or centralized software that maps the links between the different databases. 2) Data warehouse approaches, where all the data from different datasources are stored in a central repository under a common schema. 3) Finally, federation approaches where the data remain distributed in the original datasources and the integration is performed developing a common model and mapping the original data to this common model. From the user point of view, the systems can be classified into: 1) web applications without installation and configuration, easy to use but impossible to expand. 2) software toolkits or architectures with important installation and configuration steps, but with expansion capabilities.

During this section, the most significant works applied to genomic medicine will be described chronologically.

One of the most important and famous web tool for biomedical research, which provides access to more than 40 databases, is Entrez [21, 22]. This cross-link integration system was developed by the National Centre of Biotechnology Information, mapping relationships between individual entries in databases like Pubmed and other molecular databases.

In 1994 the integrated genomic database [23] becomes in one of the pioneers in the field of genomic data integration systems. This toolkit uses a local data warehouse to store the data from several databases using a global schema.

The EBI-SRS (Sequence Retrieval System) server [24, 25] has become in one of the most important cross-link integration web applications. This system connects using links of hundreds of databases and allows navigating through them. It was originally aimed at facilitating the access to biological sequences databases, initially stored in text files. Lately, they include an object-oriented design to represent more complicated biological classes. One of the main advantages of this integration system is its speed. The SRS retrieve the data directly from flat files without suffering an integration process of complex relational databases management systems. The problem of this system is that it does not easily support additional links between databases that were not explicitly described previously.

BioKleisli [26, 27] is a framework designed to provide read access to multiple data source in complex structured data. It is based on Kleisi, exporting the data of each datasource in query-time to a text format called CPL and presenting this in a homogeneous way. This is one of the first attempts of a federated data integration approach in biomedicine. A similar solution using Kleisi was developed in [28].

In Imai's work [29], a multi agent system approach has been followed to

implement a federated model in this toolkit. This architecture is composed by one wrapper-agent for each datasource and a user-agent for the result presentation as a mediator. The authors use an object oriented approach for model representation. The authors compare their system with the TSIMMIS system, because it's the first biomedical data integration work that implements the wrapper mediator pattern.

Another important field in data integration is the homogeneous integration of heterogeneous clinical trials. The Nadkarni et al. approach [30] from the Center of Medical Informatics in Yale proposed a client/server architecture to stores this data from clinical trials in entity-attribute-value form. It allows querying the systems for this data in several standard reports or flat text.

TAMBIS(Transparent Access to Multiple Bioinformatics Information Sources) [31, 32] was one of the most famous toolkits designed to integrate biological information services. It uses wrapper mediator architecture. The mediator uses an ontology as conceptual model, first described in GAIL and now in DAML + OIL. The wrappers are implemented over the remote object architecture CORBA. This system was the first one to allow the query formulation over a conceptual model using an ontology, and further translation of this query to the original datasources. This architecture has been used later in other ontology-mediation-based systems.

The Datafoundry [33] approach is a warehouse architecture approach that introduces the term of mediation to create and maintain the warehouse in a hybrid mediated-warehouse architecture. The integration process is developed by linking related interdatabase data, as opposed to integrating them by defining a single correct view. It allows providing consistent views of instance data when appropriate while still maintaining the heterogeneity between the instances. The system works with several databases, including PDB, SwissPROT, SCoP and dbEST.

The Distributed Annotation Server (DAS) [34] is an annotation system based on distributed storage without centralized supervision. That means that each peer annotates and stores a portion of the sequence. An external viewer can search the sequence and retrieve an integrated view of the whole annotated sequence. DAS sometimes is not considered as a biomedical data integration system, because there is a specific protocol and format to deal with the sequence. Actually, DAS can be considered as a protocol of a distributed database system more than a federated system.

The IBM's middleware Discoverylink [35] allows data integration from heterogeneous life-sciences datasources. This system is based on Garlic, a federated database management system prototype and DataJoiner, an IBM federated database management product for relational data sources. The first one uses wrappers that allows modular architecture and facilitates integration of new datasources, and the second has a probed technology for federation relational data sources and query optimization.

The toolkit presented by Mork [36] is based on a mediated schema, represented as a graph, where each node represent an entity, and the edges connecting this nodes represent the relationships between the entities. It's based on six main entities: phenotype, gene, locus, nucleotide, protein and structure. This mediated scheme is very flexible and can be extended both in new entities and new data sources easily. The data model was implemented describing each entity with a DTD and the queries were performed using XML to get the data from the original datasources through the query engine Tuskwila.

The Genome Information Management System (GIMS) [37] integrate genome, transcriptome, mutation, phenotype and protein interactions in a warehouse environment. The system uses an object database to store the functional and sequence data and uses the same object model as the global model. This system is specifically targeted at analysis tasks.

Biomolquest [38] is a web search engine instead an integration system. This search engine relates data from protein databases as PDB, Swiss-PROT, ENZYME or CATH. This approach stores these databases in a central repository and provides cross-reference links to the data of their databases. This approach cannot be considered as a warehouse because there is no new integration model but the cross-reference tables.

The ISYS platform [39] is a toolkit for biomedical data integration based on the mediator approach but using components. Instead of defining a highly-detailed schema, in order to maximize the flexibility of the system; the general model developed is based on Java interfaces to define the essential entities in highly general and abstract terms. Each component can extend and implement these entities.

The model based on mediation proposed by Ludascher [40] is another semantic architecture to solve the data integration problem. In this model, the views are defined and executed at the level of conceptual models instead of in a structured level. This system uses domain maps, semantic nets of concepts and relationships, used to mediate across sources. The data are translated to XML and then integrated into the mediator through a conceptual model. The mapping and the queries are defined in XML too.

The Target Informatics Net [41] is an integration system which integrates heterogeneous data sources like Mouse Genome Database, Gene Expression Database, Gen Bank, SwissProt, Pubmed, GeneCards, the results of software like Blast or Prosites and other local databases. These datasources come from relational database, flat files, web sites or applications. This approach is based in a hybrid approach that combines the federated model with the local copies of the warehouse approach. All the datasources are copied and updated regularly locally. After that, they use a federated approach to access all this heterogeneous data. The commu-

nication between the local databases and the query processor is done using the remote object architecture CORBA. The federated data model is implemented in an object oriented multi-database model OPM that provides tools for mapping and querying.

Starting as a database to store genotype and phenotype data from pharmacogenetic studies, the PharmGKB project [42-44] has been growing in the last 12 years, becoming one of the most important resources of biomedical data. This web portal is currently formed by 5 main entries (Genes, Pathways, Diseases, Variants and Drugs & small molecules). The architecture is based in a warehouse approach and the object model is based in main entities like genes, drugs, sequence, assays, etc. and exposes a presentation using web pages or web services.

In the same way that Imai, Karasavvas et al. [45, 46] propose a Multi-Agent bioinformatics integration system. In contrast to Imai's approach, the MAS Karasavvas model uses the agents not only to implement the wrapper/mediator pattern, but also to implement decision making for constructing and executing the query in the decision agents.

Information Integrator [47] is a commercial software package of IBM. It was an extension of the IBM's relational database DB2, to provide federated access to different datasources, the successor of the IBM DiscoveryLink. The main advantage of Information Integrator is the inclusion of a module called Life Sciences Data Connect, which consist in a series of wrappers to access to biological non-relational data sources. When using relational database management systems in the information integration process, many technical problems arise related with the semantics.

In the SEMEDA (Semantic Meta Database) [48] a 3-tiered architecture was developed. It consists of a relational database to store ontologies, database metadata and semantic database definitions. The biomedical datasources are accessed by means of JDBC if relational

or using BioDataServer otherwise. The system uses JSP to dynamically generate the presentation tier. During the integration process, SEMEDA uses two kind of ontologies; general ontologies like GO to unify entries between databases and a SEMEDA custom ontology which defines databases at the scheme level and acts like the data model.

Diseasecard [49, 50] is a public web portal that provides a single entry point to access medical and genetic information about rare diseases. The system integrates (cross-linking) several databases, relating concepts and constructing "cards" for each disease. Then more information about the content of this card can be obtained through the original data source.

The EnsMart system [51] is based on a warehouse approximation, capturing and transforming the data from primary datasources and storing it in a local database (data mart). This system has two types of front-end to provide interaction: a web-based tool and a standalone tool with an API for interaction with other applications. This system was originally designed to interact with the databases of the EBI's Ensemble (different species genomic annotation, functional annotation and expression), but it can work with external datasources.

The Biomediator system [52] (formerly GeneSeek) is a federated approach where the system defines and traverses a graph where the nodes represent datasource instances and the edges represent instances of the relationships that connect entities in one or more sources in the schema. The system uses an annotated mediated schema to describe the biomedical entities. Each user can create a completely customize mediated schema to describe her/his needs. Users annotate the schema by adding additional information about data sources and their relationships. This system uses protégé to store the catalog of sources and mapping between them. The wrapper datasources in Biomediator are generalized, exposing all the data, and not mapping it onto a particular mediated schema.

The approach followed by LIMBO [53] is based on a data warehouse, where store contents of the heterogeneous databases are stored following a common data structure. The data that fulfill the warehouse can be obtained either from flat files or databases. The most remarkable feature of this system is the use of a generalized database structure, using three main tables, DATA, RELATION and METADATA. The advantages of this structure are that no structure needs to be developed to provide integrate access to complex data and changes in the of data sources structure do not affect the target data structure.

Robinson and Rahayu [54] propose a federated integration system for genome databases based in mediation. In this approach, the Bioinformatic Sequence Markup Language (BSML) is used to interchange data between the layers of the system.

The Query Integrator System [55-57] is a federated approach based in a semantic mediator. The queries are described in the mediator in a common language and translated into the specific datasource syntax. The system is composed of three servers: the Integrator Server (IS), the data source server (DSS) and the ontology server (OS). The DSS capture the schema of the datasource, creating a schema metadata repository that is mapped onto ontologies. The IS creates a specific user-oriented ontology, related with the domain of the problem. This ontology is mapped with all the DSS ontologies that represent the metadata of the datasources. OS maintain an integrated schema and one or more controlled vocabularies that are used within the federation. This system incorporates the positive aspects of TAMBIS and adds new ones, such as the definition of database ontologies.

In 2002 the European Commission funded the project TEMBLOR, coordinated by the EBI, which aims to provide resources for patient, protein-protein interaction, structural and microarray data. This consortium developed an integration layer, the Integr8 [58, 59] (58,

59). This approach uses the warehouse method, building APIs of the different databases and storing the data from many sources in a single site using controlled vocabularies.

Atlas [60] is an extensible data warehouse that establishes common relational data models as integration models. Each of these common data models stores information about the same entity (i.e. molecular interaction, genes, sequences, etc.) that is related with the other entities using the relational model. The authors develop an API to populate the warehouse with data from new databases. The warehouse can be queried using SQL or APIs in different languages like Java, Perl or C++. All data in the system are annotated with the original datasource to avoid information loss during the integration process, causing semantic inconsistencies and replicated data.

The Biowarehouse toolkit [61] is an open source system for developing warehouses using a common database management system (Oracle or MySQL). All the data of public heterogeneous data sources are loaded in the warehouse schema, which was designed studding the schemas of the most common databases to be integrated. Biowarehouse has been widely used in several biomedical research projects, becoming the most recommended warehouse toolkit for biomedical purposes.

Biozon [62] is a web portal based on a data warehouse, devoted primarily to structural biology, but with genomic information too. It's based on relating fundamental biological objects, mapping the source databases following a tightly integrated hybrid hierarchical-graph based schema. The data model of Biozon uses a combination of expressive core graph model supported with a class hierarchy in each graph element. This solution is flexible and can be adjusted to several biological problems. Biozon incorporates analysis tools like Blast, EST analysis, etc.

Apart of the integration task, the ONDEX [63] framework includes sequence analysis, data mining and graph

analysis. The integration model is based on a graph structure which uses ontologies, importing firstly the data from the databases using parses developed ad-hoc and aligning/mapping then the data taking into account semantic information. This mapping is performed between concepts instead of creating new entities.

The Cancer Research Database [64] is based on a hybrid model which combines a warehouse with a federation approach through a mediator design. The mediator is used with larger data that changes often, high availability sources and flexible timing constraints in processing time, while the warehouse is used for smaller data that change periodically, sources that can be unavailable often and with very stringent timing constraints. The schema design is based on two main entities, DockingResults and AssayResults, which are related with Experiments, Mutants and Molecules tables.

Ontofusion [65] is another approach to mediation biomedical datasource integration, using a multi-agent system architecture and ontologies. The architecture of the system is the most popular in information integration with MAS, with an agent acting as mediator, and one wrapper agent for each datasource, using the JADE multi-agent platform. Ontofusion uses a hybrid query translation, where each datasource is represented by an individual schema (virtual schema). These virtual schemas are generated by means of a mapping process, where the user assigns each element from a datasource to concepts in a domain ontology. The system can use several domain ontologies. Then, all the virtual schemas are unified in a global virtual schema automatically.

The Semantic Web-Enabled Data Integration (SWEDI) [66] is a approach of integration which uses ontological models, in particular, OWL, to link small models from datasources to a larger semantic model, using a bottom-up approach. First, one or several knowledge model is created capturing the domain specific concepts and rela-

tionships in OWL. Then the datasource data is transformed to RDF, and this RDF data is linked to knowledge models. After that, a common domain may be selected and finally, the system constructs and run a semantic query. The results are returned to a web interface.

Linkhub [67] is a semantic software architecture designed to integrate data, using a federated hierarchical model of “hub of hubs”. The datasources are connected in small local hubs which cover similar topics, and then these hubs are connected in a bigger hub with a federated architecture. The datasources can be accessed using SQL or RBF. Based on a semantic RDF graph of biomedical identifiers and their relationships, it can build complex cross-database queries and semantic data navigation. It’s been tested using Uniprot and East Structural Genomics Consortium.

BioDWH [68] is another toolkit for biomedical data warehouse development. One important feature of this system is that it maintains the data always updated, monitoring the original datasources, and updating the warehouse when necessary, following some update strategies. The system has some wrappers for the most common datasources, like Uni-prot, KEGG, OMIM, GO, Enzyme, BRENDA, PDB, MINT, SCOP, EMBL-Bank or PubChem. The data model follows an object schema, using the object-relational mapping middleware Hibernate.

DebugIT [69] is a 7th European Framework Project that aims to improve the detection and elimination of bacteria using IT technologies. One of the scopes of the project was to integrate heterogeneous clinical and biological sources using a semantic infrastructure. All the datasources are accessed using SPARQL and the results are presented in RDF. The system uses wrappers that extract the data from local datasources, transform it to the global model and store it locally. The core of the system is a federated database that links all the local stored normalized data and providing access to a simple query point.

The GeNS [70] is a data warehouse system that allows integrating the most common biomedical datasources, like EMBL, UniProt (SwissProt and TrEMBL), ExPASy (PROSITE and ENZIME), NCBI (Entrez, Taxonomy, Pubmed, GenBank and OMIM), Biomart, ArrayExpress, InterPro, GO, KEGG and PharmGKB. The main feature of GeNS is that the schema was built using both generalization and specialization: 1) A generalization schema to access to a large number of datasources in order to include as much different data as possible. 2) A specialization schema to access to a limited number of unaltered datasets in order to address specific issues. GeNS stores all data into the warehouse using a general schema and then uses a top-layer meta-model where all entities are represented for accessing to specific data.

MartView is a web portal which gives full access to many different heterogeneous datasources through a warehouse BioMart [71]. This system implements a service layer (MartService) that provides easy programmatic access to all the resources of Biomart through web services. The flexibility of the architecture developed in Biomart has turned it into one of the most popular solutions. Biomart has been extended for use in several software packages like BioConductor, DAS, Galaxy, Cytoscape or Taverna.

BioDBNet [72] is an online web resource created with the intention of integrating more than 100 different databases cross-referencing identifiers to do the integration task. The authors consider this system as a warehouse, but the absence of a global schema identifies this system more like a cross-reference system. One of the features is the possibility to convert the identifiers of one entity in one database into the identifier of another database.

The architecture of the Prostate Cancer Information System [73] is a federation system of two main databases: the prostate cancer database and a tumor registry. A mapping server is

also included acting as a mediator that stores the mappings between the ontology and the databases, that has been previously describe in a declarative language D2RQ. The last part of the architecture is the data query subsystem, which consists in a SPARWL interface that allows users to submit the queries through a web interface.

The Khaos Ontology-based mediator Framework [74] is a federated integration architecture that uses a Semantic Directory to register and manage ontologies. This Semantic Directory is used to build a common model represented by ontologies.

BioXM [75] uses an object-oriented semantic integration model to build a knowledge management system. This uses semantic networks to link “elements”, “relations”, “annotations” and “context” concepts with the aim of build the model or can use previously created ontologies. These semantic networks of relationships allow detecting connections, extracting patterns and answering complex questions. To optimize performance, data can be managed alternatively within a relational database or integrated from external data sources, combining the advantages of data warehouse and federated approach. The federation of external datasources uses the Biomax software BioRS [76].

Another example of hybrid approaches is the DW4TR [77] (Data Warehouse for translational Research), which combines, with a semantic model, a warehouse for the integration of clinical data, that includes clinical data, medical image, genomic analyses, etc., and a federation for the inclusion of external data.

Discussion

In the previous sections many integrations applications and integration toolkits has been analyzed. Table 1 summarizes these web applications and toolkits, ordered chronologically.

Table 1 List of all data integration approaches analyzed, with the date, architecture, cites in (a) Scopus, (b) Pubmed and (c) Google Scholar, web system or toolkit and URL if available

Name	Year	Architecture	Cites			System	URL
			a	b	c		
NCBI [21]	1990	Cross-link	6	8	29	Web portal	www.ncbi.nlm.nih.gov/Entrez
Integrated Genome Database [23]	1994	Data warehouse	24	4	62	Toolkit	
SRS [24, 25]	1996	Cross-link	277	90	471	Web portal	srs.ebi.ac.uk
BioKleisli [26]	1997	Federated system	68	NO	NO	Toolkit	
Imai et al. [29]	1997	Federated multi-agent	2	NO	4	Toolkit	
Nadkarni et al. [30]	1998	Federated	28	28	78	Toolkit	
TAMBIS [31]	1998	Mediator	41	28	216	Toolkit	Inaccessible
DataFoundry [33]	2000	Warehouse + mediator	19	0	32	Toolkit	
DAS [34]	2001	Distributed protocol	233	145	381	Protocol	
DiscoveryLink [35]	2001	Federated	129	NO	267	Toolkit	
Mork et al. [36]	2001	Mediator	20	9	74	Toolkit	
GIMS [37]	2001	Warehouse	NO	NO	32	Toolkit	
BioMolQuest [38]	2001	Search engine	11	2	19	Web portal	Inaccessible
ISYS [39]	2001	Mediator	35	10	71	Toolkit	
Ludascher et al. [40]	2001	Semantic mediator	15	NO	101	Toolkit	
Target Informatics Net [41]	2001	Federated w/ local copies	24	4	47	Toolkit	
PharmGKB [42-44]	2001	Warehouse	135	39	205	Web portal	www.pharmgkb.org
Karasavvas et al. [45]	2002	Multi agent mediator	NO	NO	6	Toolkit	
Information Integrator [47]	2003	Federated	0	0	0	Toolkit	
SEMEDA [48]	2003	Federated	68	9	99	Web portal	Inaccessible
DiseaseCard [49, 50]	2004	Cross-link	0	0	9	Web portal	www.diseasecard.org
Ensembl [51]	2004	Warehouse	254	160	374	Web portal	www.ensembl.org/Ensembl
BioMediator [52]	2004	Federated	5	4	37	Toolkit	
LIMBO [53]	2004	Warehouse	12	3	30	Toolkit	
Robinson and Rahayu. [54]	2004	Federated	1	NO	1	Toolkit	
QIS [55-57]	2004	Mediator	9	9	14	Toolkit	
Integr8 [58, 59]	2005	Warehouse	96	47	131	Web portal	www.ebi.ac.uk/integr8
Atlas [60]	2005	Warehouse	57	31	116	Toolkit	
BioWarehouse [61]	2006	Warehouse	55	29	102	Toolkit and test web portal	biowarehouse.ai.sri.com/PublicHouseOverview.html
BIOZON [62]	2006	Warehouse	17	10	33	Web portal	www.biozon.org
ONDEX [63]	2006	Warehouse	45	25	92	Toolkit	

Cancer Research Database [64]	2007	Warehouse + mediator	1	0	3	Toolkit	
Ontofusion [65]	2007	Multi-agent federation	25	5	34	Toolkit	
SWEDI [66]	2007	Semantic warehouse	12	9	27	Web portal	Inaccessible
LinkHub [67]	2007	Semantic federation	17	7	37	Web portal	hub.gersteinlab.org
BioDWH [68]	2008	Warehouse	5	0	14	Toolkit	
DebugIT [69]	2008	Semantic federation	15	1	22	Toolkit	
GeNS [70]	2009	Warehouse	0	NO	9	Toolkit	
BioMart [71]	2009	Warehouse	124	102	197	Web portal	www.biomart.org
BioDBnet [72]	2009	Cross-link	4	3	7	Web portal	biodbnet.abcc.ncifcrf.gov
Prostate Cancer Information System [73]	2009	Semantic federation	6	1	10	Prototype	
KA-SB [74]	2009	Semantic federation	2	0	3	Toolkit	
BioXM [75]	2011	Semantic warehouse + federation	1	1	3	Toolkit + Web portal	http://www.biobridge.eu/bio/
DW4TR [77]	2011	Warehouse + federation	0	0	0	Prototype	

In the result tables, we can observe the impact of these applications using several citing measures (Scopus, Pubmed and Google Scholar). We used only the number of cites of the first publication, ignoring further publications or revisions. This restriction makes this measure a guideline of impact, instead of a scientific measure of impact.

Many approaches are close to solving the problem in a specific field, but none of the solutions can be completely generalized to any genomic medicine problem. This is due to several problems of biological and biomedical data. Firstly there is a huge amount of biomedical or biological databases and, among them, many are not correctly annotated. Biomedical data are inherently complex, due to their evolved nature (biological concepts are representations of naturally evolved systems) [78]. This complexity makes difficult to define boundaries of these biomedical concepts, so it's very complex to represent them and compute upon. This complexity does not affect only the concepts, but also the relationship between the concepts. Another reason is the huge variability present in the data. This feature makes difficult to deal with biomedical data with static sys-

tems that cannot evolve with the data and with the schema of the data. Using the classification proposed by Hernandez et al [79], different approaches in data integration will be discussed.

The link integration approach is very used in the biomedical community, but some authors don't consider it like a pure integration architecture [80]. This is because it does not add enough functionality to deal with complex queries and does not behave well with large amounts of data of changing nature. There must be a high cohesion in the different datasources that compound it.

In the case of the warehouse or federated approach, it is possible that any of these pure models could be applied to a complex genomic medicine integration task. Analyzing the approaches in this review, it can be observed that there exists a trend towards changing from centralized warehouse models to federated semantic models, with the improvements in communication, database management systems and in ontology engineering. Another possible evolution of the integrated systems is the mixed model of federated and warehouse, or federated models with local copies of the datasources.

The main advantages of the warehouse model are the centralized nature, allowing more scalability, performance in query execution and more data control. In contrast, federated models are more complex solutions, and present more latency in data retrieval and lack of control. The main advantage of the federated model is that in centralized approaches, the global model has been done ad-hoc, so the federated model is more flexible and can access to the original datasources. This is particularly important in biomedicine, because these complex data are important not only for "what" represent, but also for "how" they are represented. The warehouse approach eliminates this information from the global schema. Data warehouses have a high cost in terms of defining how the data has to be integrated, how the information is stored and how the different data types match with the general model. Arenson [47] points out that using data warehouses will be impossible for molecular biology until the field reaches a point where there is general agreement on what types of information are interesting and how they relate with each other. Finally, another advantage of the federated system is the update of the data. The federated sys-

tems are always updated, but in the data warehouse systems the data must be updated frequently.

The federated systems are a better solution in problems that can change over time, where new entities or features can be added. Warehouse approaches are suitable for monolithic problems or with high performance requirements.

After analyzing the impact of these systems, we found that the most widely used are the integration systems where the user has no responsibilities (web portals instead of toolkits) like NCBI or EBI resources, ENsMart, PharmGKB or Biomart. Tambis and Sameda are the most used federated system while the most used warehouses are Atlas and Biowarehouse.

Conclusion

In this review, several integration works related with biomedical data have been analyzed chronologically, highlighting the main features of each of them. Most of these works are suitable for a particular solution, but not for the general case. When a new biomedical data integration problem must be faced, it is essential to understand the complexity of the biomedical data and to analyze the possible solutions, evaluating each of the features of the integration problem and the datasources that must be integrated. Integration problems cannot be solved without a correct semantic annotation, mapping and querying.

Acknowledgments

This work is supported by the following projects: "Ibero-American Network of the Nano-Bio-Info-Cogno Convergent Technologies", Ibero-NBIC Network (209RT-0366) funded by CYTED (Spain), "Development of new image analysis techniques in 2D Gel for biomedical research" (ref.10SIN105004PR) funded by Xunta de Galicia and RD07/0067/0005, funded by the Carlos III Health Institute.

References

1. Brazas MD, Yim DS, Yamada JT, Ouellette BF. The 2011 Bioinformatics Links Directory update: more resources, tools and databases and features to empower the bioinformatics community. *Nucleic Acids Res* 2011 Jul;39(Web Server issue):W3-7.
2. Williams N. How to get databases talking the same language. *Science* 1997 Jan 17;275(5298):301-2.
3. Sujansky W. Heterogeneous Database Integration in Biomedicine. *J Biomed Inform* 2001;34(4):285-98.
4. Antezana E, Kuiper M, Mironov V. Biological knowledge management: the emerging role of the Semantic Web technologies. *Briefings in Bioinformatics* 2009 July 1, 2009;10(4):392-407.
5. Wong L. Technologies for integrating biological data. *Brief Bioinform* 2002 December 1, 2002;3(4):389-404.
6. Stein LD. Integrating biological databases. *Nat Rev Genet* 2003 May;4(5):337-45.
7. Köhler J. Integration of life science databases. *Drug Discovery Today: BIOSILICO*. [Integration of life science databases]. 2004 march 2004;2(2): 9.
8. Philippi S, Kohler J. Addressing the problems with life-science databases for traditional uses and systems biology. *Nat Rev Genet* 2006 Jun;7(6):482-8.
9. Brazhnik O, Jones JF. Anatomy of data integration. *J Biomed Inform* 2007;40(3):252-69.
10. Louie B, Mork P, Martin-Sanchez F, Halevy A, Tarczy-Hornoch P. Data integration and genomic medicine. *J Biomed Inform* 2007;40(1):5-16.
11. Burgun A, Bodenreider O. Accessing and integrating data and knowledge for biomedical research. *Yearb Med Inform* 2008;91-101.
12. Cheung K-H, Kashyap V, Luciano JS, Chen H, Wang Y, Stephens S. Semantic mashup of biomedical data. *J Biomed Inform* 2008;41(5):683-6.
13. Cheung KH, Yip KY, Townsend JP, Scotch M. HCLS 2.0/3.0: health care and life sciences data mashup using Web 2.0/3.0. *J Biomed Inform* 2008 Oct;41(5):694-705.
14. Goble C, Stevens R. State of the nation in data integration for bioinformatics. *J Biomed Inform* 2008 Oct;41(5):687-93.
15. Philippi S. Data and knowledge integration in the life sciences. *Brief Bioinform* 2008 November 1, 2008;9(6):451.
16. Akula SP, Miriyala RN, Thota H, Rao AA, Gedela S. Techniques for integrating -omics data. *Bioinformatics* 2009;23(6):284-6.
17. Ruttenberg A, Clark T, Bug W, Samwald M, Bodenreider O, Chen H, et al. Advancing translational research with the Semantic Web. *BMC Bioinformatics* 2007;8 Suppl 3:S2.
18. Zhang Z, Cheung KH, Townsend JP. Bringing Web 2.0 to bioinformatics. *Brief Bioinform* 2009 Jan;10(1):1-10.
19. Cheung KH, Frost HR, Marshall MS, Prud'hommeaux E, Samwald M, Zhao J, et al. A journey to Semantic Web query federation in the life sciences. *BMC Bioinformatics* 2009;10 Suppl 10:S10.
20. Lambrix P, Strömbäck L, Tan H. Information Integration in Bioinformatics with Ontologies and Standards. In: Bry F, Maluszynski J, editors. *Semantic Techniques for the Web*. Springer Berlin / Heidelberg; 2009. p. 343-76.
21. Benson D, Boguski M, Lipman D, Ostell J. The National Center for Biotechnology Information. *Genomics* 1990 Feb;6(2):389-91.
22. Wheeler DL, Chappey C, Lash AE, Leipe DD, Madden TL, Schuler GD, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2000;28(1):10-4.
23. Ritter O, Kocab P, Senger M, Wolf D, Suhai S. Prototype implementation of the integrated genomic database. *Comput Biomed Res* 1994 Apr;27(2):97-115.
24. Etzold T, Ulyanov A, Argos P. SRS: information retrieval system for molecular biology data banks. *Methods Enzymol* 1996;266:114-28.
25. Zdobnov EM, Lopez R, Apweiler R, Etzold T. The EBI SRS server—recent developments. *Bioinformatics* 2002 Feb;18(2):368-73.
26. Davidson SB, Overton C, Tannen V, Wong L. BioKleisli: a digital library for biomedical researchers. *International Journal on Digital Libraries* 1997;1(1):36-53.
27. Kolatkar PR, Sakharkar MK, Tse CR, Kiong BK, Wong L, Tan TW, et al. Development of software tools at Bioinformatics Centre (BIC) at the National University of Singapore (NUS). *Pac Symp Biocomput* 1998:735-46.
28. Chung SY, Wong L. Kleisli: a new tool for data integration in biology. *Trends Biotechnol* 1999 Sep;17(9):351-5.
29. Imai T, Matsuda H, Sekihara T, Nakanishi M, Hashimoto A, editors. Implementing an integrated system for heterogeneous molecular biology databases with intelligent agents. *Communications, Computers and Signal Processing, 1997 '10 Years PACRIM 1987-1997 - Networking the Pacific Rim' 1997 IEEE Pacific Rim Conference on*; 1997.
30. Nadkarni PM, Brandt C, Frawley S, Sayward FG, Einbinder R, Zelterman D, et al. Managing attribute—value clinical trials data using the ACT/DB client-server database system. *J Am Med Inform Assoc* 1998 Mar-Apr;5(2):139-51.
31. Baker PG, Brass A, Bechhofer S, Goble C, Paton N, Stevens R. TAMBIS—Transparent Access to Multiple Bioinformatics Information Sources. *Proc Int Conf Intell Syst Mol Biol* 1998;6:25-34.
32. Stevens R, Baker P, Bechhofer S, Ng G, Jacoby A, Paton NW, et al. TAMBIS: transparent access to multiple bioinformatics information sources. *Bioinformatics* 2000 Feb;16(2):184-5.
33. Critchlow T, Fidelis K, Ganesh M, Musick R, Slezak T. DataFoundry: information management for scientific data. *IEEE Trans Inf Technol Biomed* 2000 Mar;4(1):52-7.
34. Dowell RD, Jøkerst RM, Day A, Eddy SR, Stein L. The distributed annotation system. *BMC Bioinformatics* 2001;2:7.
35. Haas LM, Schwarz PM, Kodali P, Kotlar E, Rice JE, Swope WC. DiscoveryLink: a system for integrated access to life sciences data sources. *IBM Syst J* 2001;40(2):489-511.
36. Mork P, Halevy A, Tarczy-Hornoch P. A model for data integration systems of biomedical data applied to online genetic databases. *Proc AMIA Symp* 2001:473-7.

37. Cornell M, Paton NW, Shengli W, Goble CA, Miller CJ, Kirby P, et al., editors. GIMS-a data warehouse for storage and analysis of genome sequence and functional data. Bioinformatics and Bioengineering Conference, 2001 Proceedings of the IEEE 2nd International Symposium on; 2001.
38. Bukhman YV, Skolnick J. BioMolQuest: integrated database-based retrieval of protein structural and functional information. Bioinformatics 2001 May;17(5):468-78.
39. Siepel A, Farmer A, Tolopko A, Zhuang M, Mendes P, Beavis W, et al. ISYS: a decentralized, component-based approach to the integration of heterogeneous bioinformatics resources. Bioinformatics 2001 January 1, 2001;17(1):83-94.
40. Ludascher B, Gupta A, Martone ME, editors. Model-based mediation with domain maps. Data Engineering, 2001 Proceedings 17th International Conference on; 2001 2001.
41. Eckman BA, Kosky AS, Laroco J, Leonardo A. Extending traditional query-based integration approaches for functional characterization of post-genomic data. Bioinformatics 2001 July 1, 2001;17(7):587-601.
42. McDonagh EM, Whirl-Carrillo M, Garten Y, Altman RB, Klein TE. From pharmacogenomic knowledge acquisition to clinical applications: the PharmGKB as a clinical pharmacogenomic biomarker resource. Biomark Med 2011 Dec;5(6):795-806.
43. Klein TE, Chang JT, Cho MK, Easton KL, Ferguson R, Hewett M, et al. Integrating genotype and phenotype information: an overview of the PharmGKB project. Pharmacogenetics Research Network and Knowledge Base. Pharmacogenomics J 2001;1(3):167-70.
44. Gong L, Owen RP, Gor W, Altman RB, Klein TE. PharmGKB: an integrated resource of pharmacogenomic data and knowledge. Curr Protoc Bioinformatics 2008 Sep;Chapter 14:Unit14 7.
45. Karasavvas K, Burger A, Baldock R. A Multi-agent Bioinformatics Integration System with Adjustable Autonomy. In: Ishizuka M, Sattar A, editors. PRICAI 2002: Trends in Artificial Intelligence: Springer Berlin / Heidelberg; 2002. p. 171-87.
46. Karasavvas KA, Baldock R, Burger A. Bioinformatics integration and agent technology. J Biomed Inform 2004;37(3):205-19.
47. Arenson AD. Federating data with Information Integrator. Brief Bioinform 2003 December 1, 2003;4(4):375-81.
48. Köhler J, Philippi S, Lange M. SEMEDA: ontology based semantic integration of biological databases. Bioinformatics 2003 December 12, 2003;19(18):2420-7.
49. Oliveira JL, Dias G, Oliveira I, Rocha P, Hermosilla I, Vicente J, et al. DiseaseCard: A Web-Based Tool for the Collaborative Integration of Genetic and Medical Information. In: Barreiro JM, Martin-Sanchez F, Maojo V, Sanz F, editors. Biological and Medical Data Analysis: Springer Berlin / Heidelberg; 2004. p. 409-17.
50. Dias GS, Oliveira JL, Vicente J, Martin-Sanchez F. Integrating medical and genomic data: a successful example for rare diseases. Stud Health Technol Inform 2006;124:125-30.
51. Kasprzyk A, Keefe D, Smedley D, London D, Spooner W, Melsopp C, et al. EnsMart: a generic system for fast and flexible access to biological data. Genome Res 2004 Jan;14(1):160-9.
52. Donelson L, Tarczy-Hornoch P, Mork P, Dolan C, Mitchell JA, Barrier M, et al. The BioMediator system as a data integration tool to answer diverse biologic queries. Stud Health Technol Inform 2004;107(Pt 2):768-72.
53. Philippi S. Light-weight integration of molecular biological databases. Bioinformatics 2004 January 1, 2004;20(1):51-7.
54. Robinson A, Rahayu W. Genome Database Integration. In: Laganà A, Gavrilova ML, Kumar V, Mun Y, Tan CJK, Gervasi O, editors. Computational Science and Its Applications – ICCSA 2004: Springer Berlin / Heidelberg; 2004. p. 443-53.
55. Marenco L, Wang TY, Shepherd G, Miller PL, Nadkarni P. QIS: A framework for biomedical database federation. J Am Med Inform Assoc 2004 Nov-Dec;11(6):523-34.
56. Lam HY, Marenco L, Clark T, Gao Y, Kinoshita J, Shepherd G, et al. AlzPharm: integration of neurodegeneration data using RDF. BMC Bioinformatics 2007;8 Suppl 3:S4.
57. Marenco L, Wang R, Nadkarni P. Automated Database Mediation Using Ontological Metadata Mappings. J Am Med Inform Assoc 2009 September 1, 2009;16(5):723-37.
58. Kersey PJ, Morris L, Hermjakob H, Apweiler R. Integr8: enhanced inter-operability of European molecular biology databases. Methods Inf Med 2003;42(2):154-60.
59. Kersey P, Bower L, Morris L, Horne A, Petryszak R, Kanz C, et al. Integr8 and Genome Reviews: integrated views of complete genomes and proteomes. Nucleic Acids Res 2005 Jan 1;33(Database issue):D297-302.
60. Shah SP, Huang Y, Xu T, Yuen MM, Ling J, Ouellette BF. Atlas - a data warehouse for integrative bioinformatics. BMC Bioinformatics 2005;6:34.
61. Lee TJ, Pouliot Y, Wagner V, Gupta P, Stringer-Calvert DW, Tenenbaum JD, et al. BioWarehouse: a bioinformatics database warehouse toolkit. BMC Bioinformatics 2006;7:170.
62. Birkland A, Yona G. BIOZON: a hub of heterogeneous biological data. Nucleic Acids Res 2006 Jan 1;34(Database issue):D235-42.
63. Kohler J, Baumbach J, Taubert J, Specht M, Skusa A, Ruegg A, et al. Graph-based analysis and visualization of experimental results with ONDEX. Bioinformatics 2006 Jun 1;22(11):1383-90.
64. Bichutskiy VY, Colman R, Brachmann RK, Lathrop RH. Heterogeneous Biomedical Database Integration Using a Hybrid Strategy: A p53 Cancer Research Database. Cancer Inform 2007;2:277-87.
65. Alonso-Calvo R, Maojo V, Billhardt H, Martin-Sanchez F, Garcia-Remesal M, Pérez-Rey D. An agent- and ontology-based system for integrating public gene, protein, and disease databases. J Biomed Inform 2007;40(1):17-29.
66. Post LJJ, Roos M, Marshall MS, van Driel R, Breit TM. A semantic web approach applied to integrative bioinformatics experimentation: a biological use case with genomics data. Bioinformatics 2007 November 15, 2007;23(22):3080-7.
67. Smith AK, Cheung KH, Yip KY, Schultz M, Gerstein MK. LinkHub: a Semantic Web system that facilitates cross-database queries and information retrieval in proteomics. BMC Bioinformatics 2007;8 Suppl 3:S5.
68. Topel T, Kormeier B, Klassen A, Hofstadtl R. BioDWH: a data warehouse kit for life science data integration. J Integr Bioinform 2008;5(2).
69. Lovis C, Colaert D, Stroetmann VN. DebugIT for patient safety - improving the treatment with antibiotics through multimedia data mining of heterogeneous clinical data. Stud Health Technol Inform 2008;136:641-6.
70. Arrais J, Pereira J, Fernandes J, Oliveira J, editors. GeNS: a biological data integration platform; 2009.
71. Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, et al. BioMart—biological queries made easy. BMC Genomics 2009;10:22.
72. Mudunuri U, Che A, Yi M, Stephens RM. bioDBnet: the biological database network. Bioinformatics 2009 Feb 15;25(4):555-6.
73. Min H, Manion FJ, Goralczyk E, Wong Y-N, Ross E, Beck JR. Integration of prostate cancer clinical data using an ontology. J Biomed Inform 2009;42(6):1035-45.
74. Roldan-Garcia Mdel M, Navas-Delgado I, Kerzazi A, Chniber O, Molina-Castro J, Aldana-Montes JF. KA-SB: from data integration to large scale reasoning. BMC Bioinformatics 2009;10 Suppl 10:S5.
75. Maier D, Kalus W, Wolff M, Kalko SG, Roca J, Marin de Mas I, et al. Knowledge management for systems biology a general and visually driven framework applied to translational medicine. BMC Syst Biol 2011;5:38.
76. Kaps A, Dyshlevoi K, Heumann K, Jost R, Kontodinas I, Wolff M, et al. The BioRS(TM) Integration and Retrieval System: An open system for distributed data integration. J Integr Bioinform 2006;3(2):44.
77. Hu H, Correll M, Kvecher L, Osmond M, Clark J, Bekhash A, et al. DW4TR: A Data Warehouse for Translational Research. J Biomed Inform 2011 Dec;44(6):1004-19.
78. Bernstam EV, Smith JW, Johnson TR. What is biomedical informatics? J Biomed Inform 2010 Feb;43(1):104-10.
79. Hernandez T, Kambhampati S. Integration of biological sources: current systems and challenges ahead. ACM SIGMOD Record 2004;33(3):51-60.
80. Davidson SB, Overton C, Buneman P. Challenges in integrating biological data sources. J Comput Biol 1995;2(4):557-72.

Correspondence to:

Julián Dorado
Information and Communications Technologies Department
School of Computer Science
Campus de Elviña s/n.
15071 A Coruña, Spain
Tel: +34 981167000 ext 1239
Fax: +34 981167160
E-mail: julian@udc.es