We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Chapter 6

Approaches to Access Biological Data Sources

Assia Rharbi, Khadija Amine, Zohra Bakkoury,
Afaf Mikou, Anass Kettani and Abdelkader Betari

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/48740

1. Introduction

In recent years, technological revolutions in genomics and proteomics have revolutionized the work of researchers in molecular biology. Through various techniques of data generation, they have at their hand in the web a very large amount of information contained in public and heterogeneous data sources. Each source has content organized around a particular data type like sequences in Uniprot (for proteins) and Genbank (for gene and mRNA), protein structure in PDB (Protein Data Bank) and publications in biomedical Medline. Their content is heterogeneous in the sense that a similar data can be represented differently in two data sources (eg different names for the same gene). More data sources have a variety in terms of structure, and there are sources of structured data, such as relational databases or semi-structured sources like XML and unstructured sources such as databases composed of flat files. That is to say that a biologist who wishes to obtain information from these sources have to question these one by one, then copy and analyze the data collected, and manage redundancy, complementarities of the information and inconsistencies. Today, one of the greatest challenges of bioinformatics is to enable biologists to effectively access multiple data sources, each with a different pattern. Various approaches have been adopted to unify access to various data sources given a query. Several systems have been produced from data warehouses, a federation of databases or mediators.

In this work, we are interested in mediation systems. Such systems offer to the user a uniform and centralized view of distributed data, this view may also reflect a more abstract, condensed, qualitative data and therefore more meaningful to the user. These mediation systems are also very useful in the presence of heterogeneous data, because they seem to use a homogeneous system.

We aim to assist biologists in their research through the development of a generic tool for the integration of heterogeneous genomic data distributed over the web, and we are placed in a very particular context that is the study of cardiovascular disease and especially familial

hypercholesterolemia. This is a disorder of high LDL ("bad") cholesterol that is passed down through families, which means it is inherited. This disease is caused by a genetic mutation of certain lipoproteins. Indeed, these lipoproteins (called LDL) carry the 2/3 of cholesterol circulating in the blood; they deliver cholesterol to tissues by a system of recognition between Apo lipoprotein Band a receiver: the LDL receptor (lock and key system) that allows the entry of LDL and their cholesterol content in cells. When the LDL receptor (LDL-R) is weak (about one mutation), LDL accumulates in the blood and artery walls causing familial hypercholesterolemia (HF). So knowing these different mutations by biologists, can greatly facilitate the molecular screening of the disease and therefore to find the proper treatment. However, to answer such a query: "What are the mutations that cause familial hypercholesterolemia (HF)?" The biologist hasto make a fastidious search in disparate and heterogeneous databases which requires a considerable investment time.

This chapter is structured as following:

- First we present the background of the project
- Second we focus on the problem of heterogeneity of data sources and biological characteristics of these sources.
- Third we present the state of the art of data integration, problems and constraints of this integration and the various existing approaches to solve this problem.
- And fourth we expose studied scenario, the realization and perspectives.

2. General context

Since the completion of the human genome sequencing in April 2003, we observe the accumulation of an outsize amounts of genomic and proteomic data on the web often syntactically and semantically heterogeneous and difficult to capitalize.

Information about genes provides access to their corresponding proteins. In addition, all diseases are associated with alterations in the structure or function of such proteins. A good knowledge of protein structure provides insight into their function.

Bioinformatics has become an important tool to explore genomic data by relying heavily on computer systems. It suggests methods and software's for biological data storage and processing. Actually, it is acquiring and organizing data, developing software for the analysis, comparison and modeling of these data and analysis results produced by bioinformatics software to infer new biological knowledge, in collaboration with biologists.

This work contributes to facilitate to biologists searching among heterogeneous and distributed data in public and / or private data sources on the web. In particular, it helps them to analyze proteins, by building a platform for integrating biological data. This will provide a tracking system to target special proteins involved in a disease known as familial hypercholesterolemia and thus, to better understand the biological activity of these macromolecules.

Familial hypercholesterolemia disease results from mutations in the LDLR gene. The LDLR gene provides instructions for making a protein called a low-density lipoprotein receptor.

This type of receptor binds to particles called low-density lipoproteins (LDLs), commonly known as bad cholesterol. By removing low-density lipoproteins from the bloodstream, these receptors play a critical role in regulating cholesterol levels. When the LDL receptor (LDL-R) is deficient, LDL accumulated in arteries induces the familial hypercholesterolemia (HF) pathology. So, in biology knowing these different mutations can greatly facilitate the molecular screening of the disease and thus find appropriate treatment.

3. Biological data sources

Number of data sources and tools available to biologists on the web has grown dramatically in recent years. This huge number of available data along with heterogeneous information generated wide variety of access interfaces, and also a profound heterogeneity.

3.1. Genomic databases

There are two types of databanks, those that correspond to a set of heterogeneous data so-called "databases" and those more homogeneous established around a specific theme.

Also, to avoid confusion we will distinguish between semantic databases, general [2] and specialized [3]databases.

For specific requirements related to the activity of a group, or to bibliographic compilations, many specific databases were created in laboratories. In some cases, these databases have been developed continuously; others have not been updated and disappeared as they represented a specific need. Still others are unknown or poorly known and are waiting for further investigation.

All these specialized databases of interest may vary considerably from one base to another according to their size. In most of the case, these bases correspond to a combination compared of generalist databases such as: Swiss-Prot, GenBank. DDBJ (DNA Data Bank of Japan), EMBL (European Molecular Biology Laboratory) which are used very often. It is important to know, that according to the field of activity or the genomics research, the surveyed banks are not necessarily the same. The genomic libraries contain various information that may include:

- Characteristics of proteins or genes such as localization of the gene in the cell: LocusLink5, the 3D structure of protein: Protein Data Bank (GDP) and Molecular Modeling Database (MMDB) or its biological function. More specifically, some databases contain information about a specific family of protein such as “Enzyme8” which include exclusively enzyme type proteins.
- Some phenotypes (specific genes, morphological feature, clinical syndrome ...) or more specifically some genetic diseases: Online Mendelian Inheritance in Man (OMIM);
- Specific species or families of species: FlyBase, Reptilia, Saccharomyces Genome Database (SGD), Mouse Genome Database (MGD);
- The medical literature (banks abstracts): Medline, PubMed.

Table1 and Table 2 give two examples of genomic databases along with protein database.

Designation	Location	Roles	Comments	Web sites and references
Nucleic bases				
EMBL	European Bioinformatics Institute (EBI) Europe	More than 1 million records (January 1998) for more than 15,500 species. The predominant species: Homo sapien,Caenorhabditis elegans, Saccharomyces cerevvisae ...	Information's search tools: SRS, System Retrieval System and via a web interface on EBI, through BLAST et FASTA software	Accessible via the web site: http://www.ebi.ac.uk/ebi_docs/embl_db/ebi/topembl.html
Specific genomic resources				
SGD Saccharomyces Genome Database	Works of Cherry and al, 1998.	Online resources on molecular biology and S.cerevisiae genetic	Numerous research help functions on line	Accessible via the web site: http://genome-www.stanford.edu/Saccharomyces/

Table 1. Genomic databases

Designation	Location	Roles	Comments	Web sites and references
Primary databases				
PIR Protein Information Resource	National Biomedical Research Foundation	Sequences collecting to detect evolutionary relationship between proteins	The current structure includes 4 compartments : PIR1, PIR2, PIR3 and PIR4	Accessible via the web site http://nbrfa.georgetown.edu/pir/
Composite databases				
NRDB Non-redundant Database	National Center Biotechnology Information USA	Composed of GenPept (derived from GenBank), PDB sequences, SWISS-PROT, SPupdate, PIR, and GenPeptupdate (update of GenPept). NRDB is the default database of BLAST and NCBI service	Accessible via the web site: http://www.ncbi.nlm.nih.gov/Web/NRDB/	NRDB Non-redundant Database

Table 2. Protein databases

3.2. Characteristics of biological data sources

The diversity of information distributed sources and their heterogeneity are the one of the main problem that the web users have to face. This heterogeneity may result from the size or structure of the sources (structured sources: relational databases, partially structured sources: XML documents, or unstructured: texts), the access mode and query, or semantic heterogeneity: between concept maps, and implicit or explicit underlying ontology's.

Biological sources have a large heterogeneity at different levels:

- Syntactic: because of the different formats for describing the content sources usually ASN.1 (formal notation for describing data transmitted via exchange protocols), (eg Enter), but also more standard formats such as XML (eg GenBank).
- Semantic which covers several aspects. First, it concerns the focus. However, each base focuses on a type of biological object (eg, the focus of Swiss-Prot is the protein, the focus of GenBank is gene, and the PDB is the 3D protein structure).
- Then, according to the base, the same information is not represented with the same level of detail: some bases are generalists (eg Swiss-Prot in general on proteins) while others are more specialized (eg SGD (Saccharomyces Genome Database) on yeast proteins).
- The final aspect of semantic heterogeneity is related to the diversity of nomenclature modes. Different vocabularies are used to annotate the sequences and the reliance on such annotations is seldom complete. Moreover, within a same database, there are a several names for each single entity (protein, gene). The name of an entity may depend on the disease to which it is linked or to its inventor.
- Source query language: another form of heterogeneity comes from query languages. Languages are often simple forms (combinations of words to search in a text), in the case of portals or simple databases. But one can also find structured languages such as SQL or OQL.
- Protocols for collecting data that are different such as CGI / HTTP or FTP. Access to web sources is limited to the entry forms and their underlying programs
- The tools offered by the Web: there are many tools for text searching and sequence comparison algorithms such as BLAST (Basic Local Alignment Search Tool), FASTA15 or LASSAP16.

4. State of the art of approaches to integration

A data integration system remedies to the problems associated with the expansion of public data sources by giving the possibility to have a unified view of them. Such a system is the interface between user and data sources simplifying requests to perform (a request to query all sources covered by the system). The user is not obliged to know where the data are and how they are structured.

4.1. Current integration approaches

There are two major approaches for integration of information: (1) the data warehouse (DW) or materialized approach and (2) virtual approach (mediator based).In DW approach, huge

amount of historic data is stored in the DW. In the virtual approach, on the other hand, the data is not materialized, but rather is globally manipulated using views. Each of these approaches is suitable in some kinds of applications.

4.1.1. Data warehouse

DW is a powerful tool for decision support and querying the data because it explicitly stores information from heterogeneous sources locally. However, some external data, such as new product announcements from opponents and currency exchange rates, may be needed to support the accuracy of the business decisions. We should not neglect the importance of such data to avoid the problems of incompleteness, inexact, or sometimes wrong results. Warehousing huge and frequently changed information is a big challenge for the following reasons.

Firstly, since the data in the DW is loaded in snapshots and the DW is a huge information repository. Secondly, as the data sources change frequently, the maintenance becomes a complicated and costly issue

Here are two examples of using data warehouses:

- Genomics Unified Schema, GUS [4] is a system for creating a data warehouse focused on molecular biology;
- Gene Expression Data Warehouse, GEDAW [5] is a warehouse dedicated to the analysis of the transcriptome of human liver.

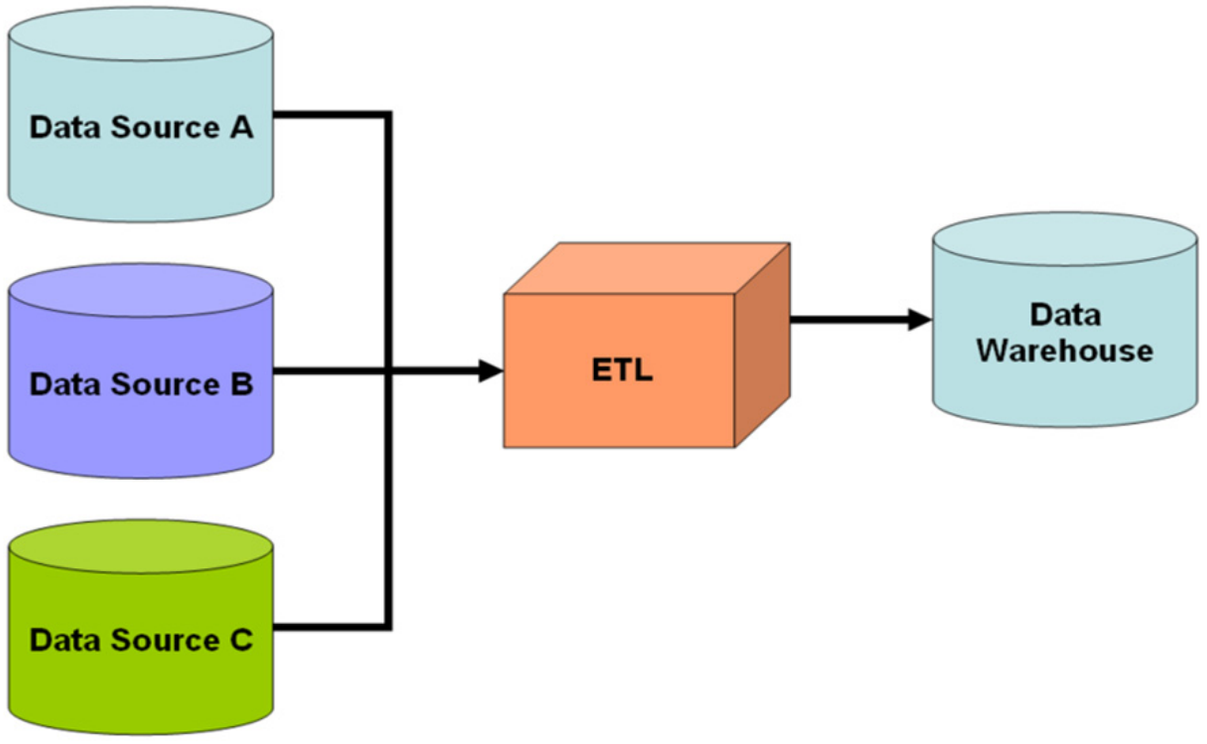

Figure 1. Simple schematic for a data warehouse

4.1.2. The virtual approach (mediator based)

In this approach, the actual data resides in the sources, and queries against the integrated 'virtual' view will be decomposed into sub queries and posed to the sources. This approach is preferred over the materialized approach DW when the information sources change very often. On the other hand, the DW approach may be desired when a quick query answer is required and the information sources change rarely.

The most important step in the construction of a mediator is the creation of the global schema. The mapping consists on the relations between the global schema and local sources. Specification of this mapping, depending on the method, determines the difficulty of query reformulation and the facility of adding or removing sources within the system. Two methods are commonly used to determine the global schema

- GAV (Global As View) approach: In this approach, each concept of the global schema is mapped to a query over data sources. In other words, when the user presents his/her query over the integrated schema, the data corresponds to a concept in the integrated schema, which can actually be answered from the data sources through a specific query. The query processing in GAV is easy, since it just unfold each concept in the integrated schema in the user query with the associated query over the sources, but this approach does not help much when the sources change or grow very often, since these factors affect the mappings and require restricting the integrated schema.
- LAV (Local As View) approach: LAV approach defines the mapping in the other way around; each concept in the data sources is defined in terms of a query over the integrated schema. This makes query processing more difficult, since in this case, the system does not know explicitly how to reformulate the concepts in the integrated view expressed in the user query in terms of the data sources. On the other hand, changes or incremental growth in the sources will not lead to reconstruction of the integrated schema, and need only to modify the mappings

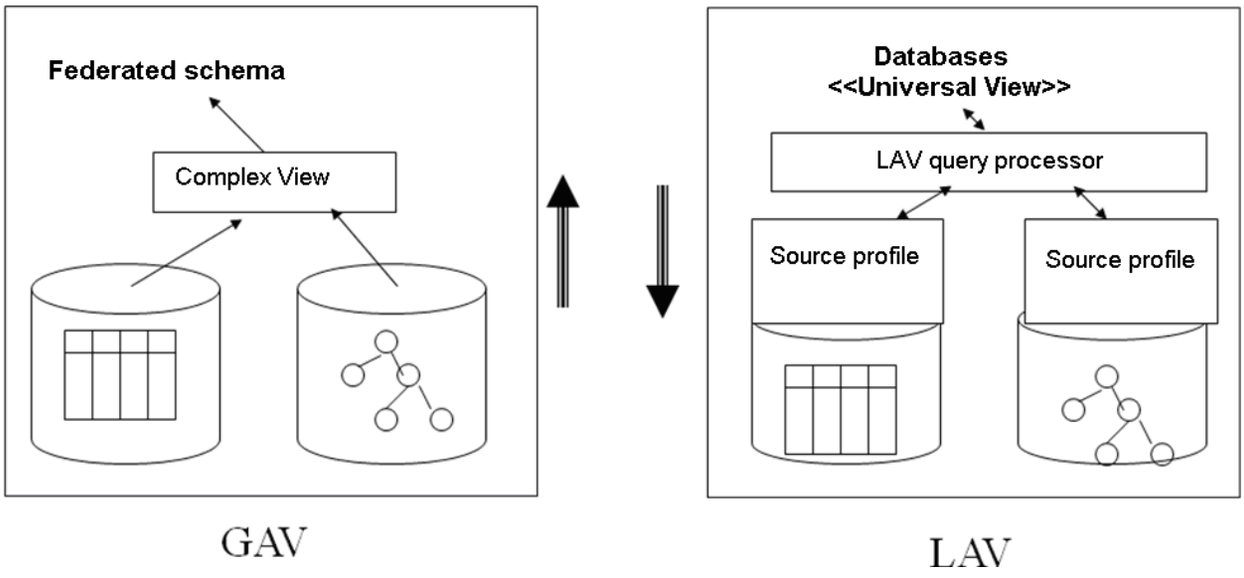

Figure 2. GAV vs. LAV

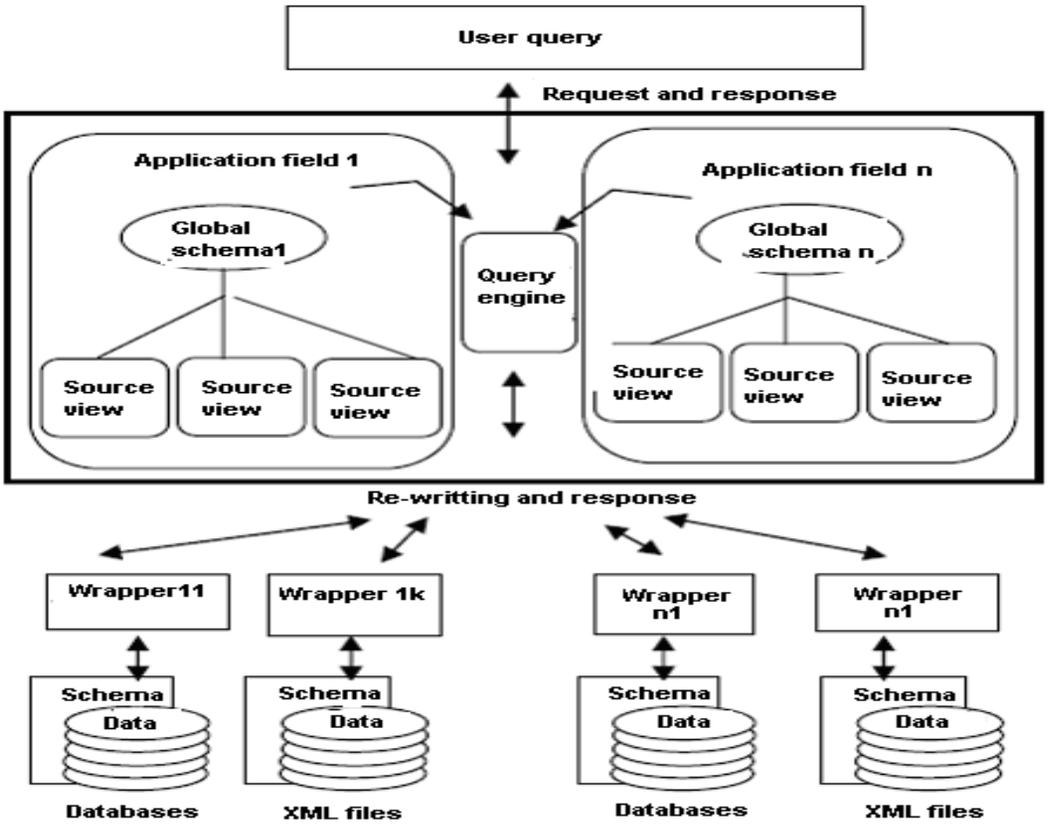

Figure 3. Architecture of a mediator

In fact, these two approaches are not opposite, but complementary; depending on the problem to be solved. To integrate a few sources, most of which are stable, better to use the GAV method. By cons, as part of a large-scale integration, the LAV method is preferable as a material change at a local source with little or no impact on the global schema.

Two examples of systems integration based mediator:

- *Tambis (Transparent Access To Multiple Bioinformatics Information Sources) [6]* is an integration system coupled to an ontology that allows for better interoperability between sources;

K2/BioKleisli [7] is a system based on CPL (Collection Programming Language) is a query language for high-level querying multiple sources.

4.1.3. The multi agents approach

This approach was used in GID-IGC *(Integrated Genomic Database - Genome Information System) project.* The proposed architecture uses a network of agents communicating each with other via CORBA and KQML. All have a specific function, such as *EIA* (External Interface Agent) that manages the user interface, or *SCA* (Dial Selector

Agent) witch decompose the global query into sub-queries for local data sources. This approach is very modular and easily extensible.

4.1.4. Navigating between sources

This approach is based on what users usually do when searching for information on the web, which involves a search page to page by clicking the mouse. In practice, queries generated for this type of tool are converted into path expressions. The data banks are then integrated based on their cross-references. These expressions can answer the query of the user according to different levels of satisfaction.

A reference is a link between two data sources (Figure 4), a bridge between the information relating on the same object or the same concept. It can be done through an identifier of an external source or a URL (Unified Resource Locator). If the link can be browsed in both directions it is a cross-reference ("cross-reference").

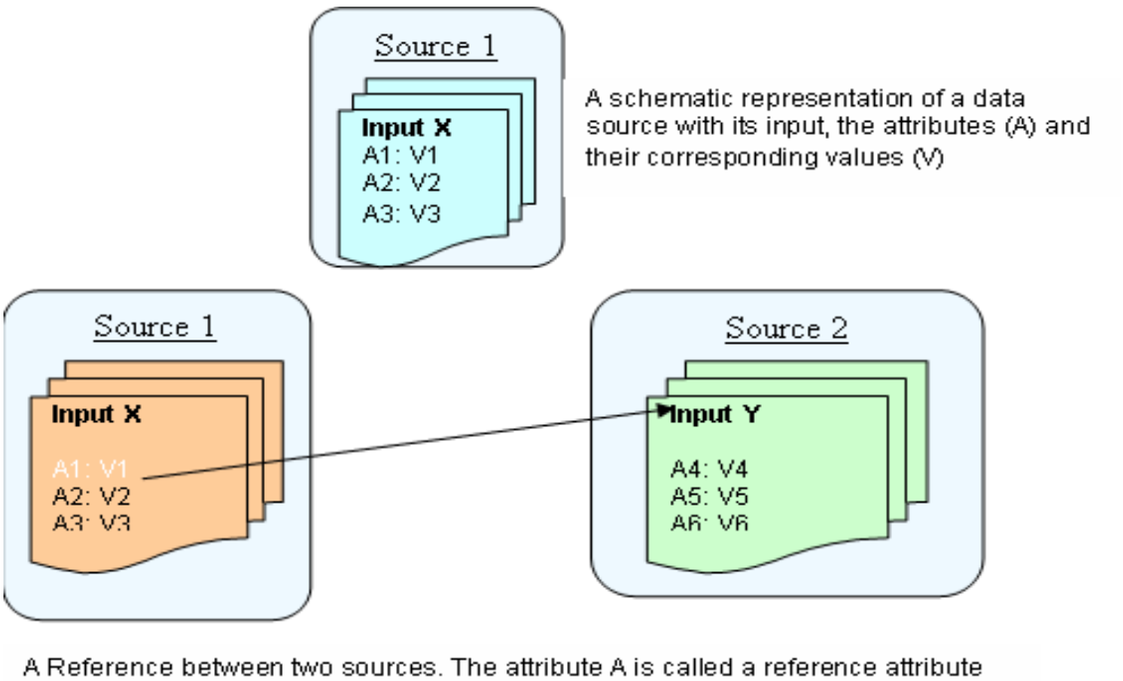

Figure 4. Navigating between sources

4.2. Adopted approach

In this work, we are interested in mediation systems. Such systems offer a uniform and centralized view of distributed data. This view may also reflect a more abstract, condensed, qualitative data and therefore more meaningful to the user. These mediation systems are also very useful in the presence of heterogeneous data, because they seem to use a homogeneous system.

In this architecture, each component provides a set of features, which, together will help to satisfy the user request at the end.

- The mediator is a software module that directly receives the user's request. It has to locate the necessary information to answer the query, resolve schematic and semantic conflicts, query different sources and integrate the partial results in a consistent and coherent

response. This is the most complex component but only one instance of it is necessary (unlike multiple adapters). It provides access to multiple data sources as if it was a single one and offers this consultation through multiple languages and ontologies.

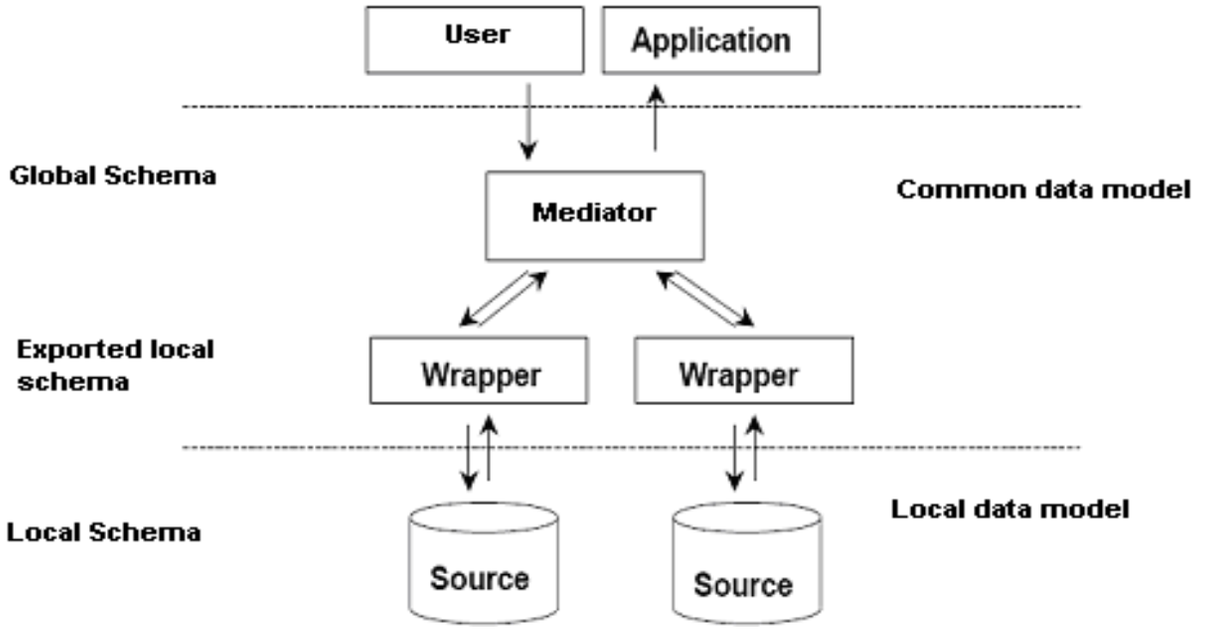

Figure 5. Adopted approach

This is a crucial component that allows a local system to distribute its information to a community of users.

- The adapter allows the presentation of data in the mediation's syntactic format. So it's an interface for querying a database using a standardized language (pivot language).
- Data source: Represent the sources and banks of biological information. A data source can be described by its:
  - Location: Reference, communication protocol, access technique (JDBC, ODBC API), support (DBMS, web pages)
  - Type of data it manages: structured (relational, object), semi-structured (XML, OEM), unstructured (image, multimedia)
  - Ability to query: SQL, OQL, search
  - Results Format: XML, HTML, relationships, texts

5. Studied scenario

After exploring the different integration systems that solve the problem of heterogeneous biological data sources, this section describes the scenario we have chosen to work on.

5.1. Biologist's need

The objective of our work is to develop an integration system for biological data with an application on familial hypercholesterolemia. Such system should facilitate access to

multiple data sources available on the Web, in a transparent and uniform way by giving biologists a single virtual source that summarizes the sites of interest to the application.

In order to satisfy this biologist's need we studied their current way to work. We first focused on existing tools, data sources they use and their functional specifications.

Tools: they use, mainly:

- CHARMM (Chemistry at Harvard Macromolecular Mechanics) [8]: This program offers a wide choice for the production and analysis of molecular simulations. It simulates the standard energy minimization of a given structure and the production of a molecular dynamics trajectory.
- VMD (Visual Molecular Dynamic) [9]: This software is used to visualize the molecules available on the web.

Data sources: The focus was mainly on the following sources:

- PDB (Protein Data Bank) [10]: It's a worldwide collection of data on three-dimensional structure of biological macromolecules: Proteins and nucleic acids. The PDB is the primary source of structural biological data. It allows access to 3D structures of pharmaceutical interest proteins.
- PubMed [11]: it's a free search engine giving access to the MEDLINE bibliographic database, gathering citations and abstracts of biomedical research.

5.2. Adopted scenario

The adopted scenario consists on building a local database "homemade" that would include unorganized data already available in the biologist's laboratory and for our particular case data related to LDL receptor mutations. This goes along with the research we're going to do on the Web using the mediation system:

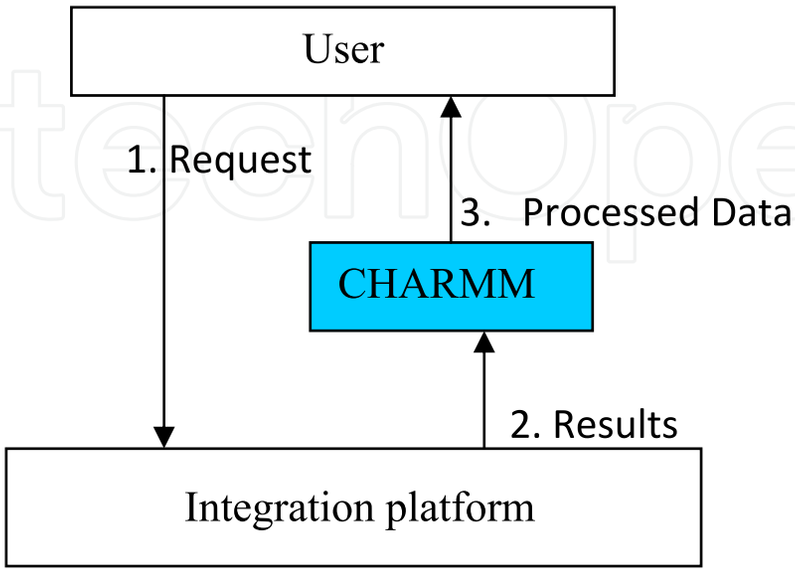

Figure 6. The adopted scenario

The integration platform is the mediation system. It will query data sources related to cardiovascular diseases especially familial hypercholesterolemia namely: PDB and PubMed. The results of the query will be processed by the tool CHARMM before being presented to the user [12, 13].

5.3. Selected data sources

PubMed [8]: Is the leading bibliographic data search engine of all fields of biology and especially medicine. It was developed by the National Center for Biotechnology Information (NCBI), and is hosted by the National Library of Medicine U.S. National Institutes of Health. PubMed is a free search engine giving access to the MEDLINE bibliographic database, gathering citations and abstracts of biomedical research.

The MEDLINE database in April 2007 had more than 15 million citations from 1950 published in 5000 biomedical journals (journals in biology and medicine) distinct. It is the database of reference for biomedical sciences. As with other indexes, including a citation in PubMed has no content. In addition to MEDLINE, PubMed also provides access to:

- OldMedline for articles before 1966
- Citations of all articles, even "irrelevant" (that is to say, covering topics such as plate tectonics or astrophysics) from certain MEDLINE journals, primarily those published in major newspapers of general science or biochemical (such as Science and Nature, for example).
- Citations being listed before indexing in MEDLINE or MeSH, or passage or status "off topic"
- Older citations selected for MEDLINE journal from which they arise (when they are supplied electronically by the publisher)
- Articles submitted to PubMed Central for free

Most citations include a link to the full article when it is available (eg PubMed Central). PubMed is a search engine that allows users to search in the MEDLINE database; this information is also available from private organizations such as Ovid and Silverplatter, among others. PubMed is free since the mid-1990s. For optimal use of PubMed, it is necessary to have an understanding of his core, MEDLINE, and especially the MeSH vocabulary used for indexing articles in MEDLINE.

We can also find in PubMed information about the log, which can search by title, subject, short title, NLM ID, ISO abbreviation, and ISSN (International Standard Serial Number) written and electronic. The database "newspaper" includes all newspapers Enter Base.

The major interest of these bibliographic databases is that:

- Their bodies are used to identify recent publications in scientific journals.
- They help to establish bibliographies (lists of relevant articles) on a subject or author.
- They are portals to access full text documents available on the Internet.
- The bibliographic databases used to find references to documents, select, print or export them to other software. They may also propose to order documents or provide access to full text.

PDB (Protein Data Bank): The databank on proteins of Research Collaborator for Structural Bioinformatics, more commonly known as Protein Data Bank or PDB is a worldwide collection of data on the three-dimensional (or 3D structure) of biological macromolecules : protein essentially, and nucleic.

Founded in 1971 by Brookhaven National Laboratory, the Protein Data Bank was transferred in 1998 to the Research Collaborator for Structural Bioinformatics (RCSB), which consists of Rutgers University, the University of Wisconsin at Madison, National Institute of Standards and Technology (NIST) and the "San Diego Supercomputer Centre." The PDB originally contained (in 1971) 7 structures. The number of structures deposited has grown since the 1980s. Indeed, at that time, the crystallographic techniques have improved, the structures determined by NMR have been added, and the scientific community has changed its view on data sharing.

The PDB containedon 28-04-2008, 50480 structures. The data are from the original pdb format, and in recent years are also mmCif format, specifically developed for structural data from the PDB. From 2000 to 3000 structures are added each year. The bank contains files for each molecular model. These files describe the exact location of each atom of the macromolecule studied, that is to say, the Cartesian coordinates of the atom in a three-dimensional coordinate.

Each model is referenced in the bank by a unique identifier to 4 characters, the first is always a numeric character, the next three being alphanumeric characters. This identifier is called "**pdb code**".

Several formats exist for PDB files:

The PDB format: it is the original format. The guide of this format has been revised several times; the current version is version 2.2[14], which has existed since 1996. Originally pdb format was dictated by the width and the use of punch cards for computers. Consequently, each line contains exactly 80 characters.

Pdb file format is a text file where each column has its meaning: Each parameter is positioned so immutable. Thus, the first 6 columns, that is to say the first 6 characters for a given line, determine the scope of the file. Found for example in the fields " TITLE_ "(That is to say, the title of the macromolecule of interest)," KEYWDS "(The keywords of the entry)," EXPDTA "Which provides information on the experimental method used," SEQRES "(The sequence of the protein under study)," ATOM_ "Or" HETATM "Fields containing all information related to a particular atom.

Pdb format limitations: Format in 80 columns pdb files is relatively restrictive. The maximum number of atoms in a pdb file is 99999, since there are only 5 columns allocated for the numbers of atoms. Similarly the number of residues per chain is at most 9999: There are only 4 columns allowed for this figure. The number of channels is limited to 62: A single column is available, and possible values are one of the 26 letters of the alphabet in upper or lower case, or one of the digits 0 through 9. As this format has been defined, these limitations did not seem restrictive, but they have been taken several times during the deposition of extremely large structures, such as viruses, ribosome, and multienzyme complexes.

MmCIF format:The growing interest in the development of database and electronic publications in the late 1980s has created the need for a more structured, standardized, open-ended and high quality data from the PDB. In 1990, the International Union of Crystallography IUCr extended to macromolecules data representation used to describe crystal structures of molecules of low molecular weight. This representation is called CIF, for Crystallographic Information File. The dictionary mmCIF (macromolecular Crystallographic Information File) published in 1996, was then developed.

In MmCIF format, each field of each section of a pdb file is represented by a description of a characteristic of an object, which includes both the name of the characteristic (eg _struct.entry_id), and the content of the description (pdb code: 1cbn). Which we can call "name-value". It is easy to convert, without loss of information, an mmCIF file format pdb, since all information is directly analyzed. It is not possible, however, to completely automate the conversion of a pdb file format mmCIF, since many mmCIF descriptors are either absent from the PDB file, either in this field " REMARK "Who can not always be analyzed. The contents of fields " REMARK " is indeed separated according to different mmCIF dictionary entries, in order to preserve the completeness of the information contained in such Materials and Methods section (crystal characteristics, refinement method ...) or in the description of the biologically active molecule or other molecules (substrate, inhibitor, ...)

The mmCIF dictionary contains over 1700 entries, which are much safer not all used in a single PDB file. All field names are preceded by the character "underscore"(_), In order to differentiate the values themselves. Each name corresponds to an mmCIF dictionary entry, where the characteristics of the object are exactly defined.

Pdbml format: This format is pdbml adaptation to XML data format bps and contains the entries described in the dictionary "PDB Exchange Dictionary". This dictionary contains the same entries as the mmCIF dictionary, in order to take into account all data managed and distributed by the PDB. This format can store much more information on models than pdb format.

Data retrieval: The files describing molecular models can be downloaded from the website of the PDB and visualized using various software such as Rasmol [15], Jmol [16], chime [17] or an extension VRML [18] (plugin) a browser. The website of the PDB also contains resources for teaching, on structural genomics and other useful software.

5.4. The global schema

By studying and exploring the previous sources and by combining data from genome sources, we have identified all data that define the dictionary related to familial hypercholesterolemia disease (Table 3). From this data dictionary and business rules (as defined and established by experts in the field of biology), we extracted the major biological entities useful for our study. These entities are not independent and form a semantic graph with nodes reflecting relationships between these entities.

Property	Description
code_biblio	Library code
auteur_biblio	Author of publication
date_biblio	Date of publication
volume_biblio	Volume of structure
langue_biblio	Language
contribution_biblio	Contribution
journal_biblio	Newspaper
revue_scint_biblio	Journal
livre_biblio	Book
Cd_proceeding_biblio	CD procedure
nom_recepteur	Name of receiver
nom_mutation	Name of mutation
classe_mutation	Class of mutation
nom_proteine	Name of the protein
longueur_proteine	Length of the protein
type_proteine	Protein type
structure_summary	Summary of the structure
structure_title	Under the structure
nom_molecule	Name of the molecule
author_name	Name of author
date_depot	Date Filed
date_release	Date of publication
derniere_release	Last update
Resolution	Resolution
Compound	Compound
Classification	Classification
molecule_chain_type	Channel Type
experimental_methode	Experimental method of resolution (RX, NMR)

Table 3. Data Dictionaries

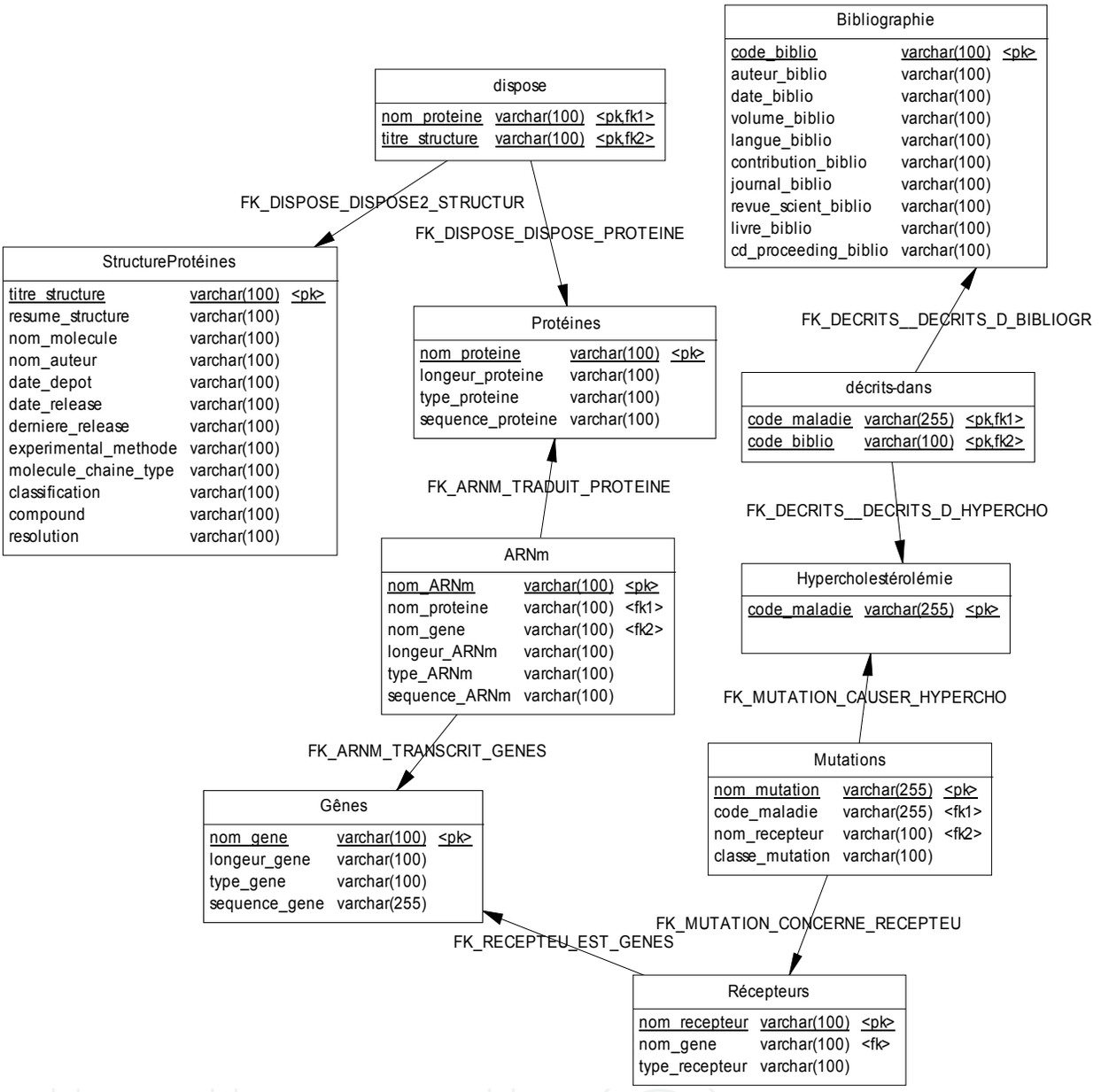

Figure 7. Global Schema

From the global schema, it is possible to make our request and submit it to SQL mediator for treatment. For example, the query that givesthe associated protein mutated gene and publications on familial hypercholesterolemia is expressed as follows in SQL:

```
Select nom_proteine, journal_biblio, auteur_biblio, date_biblio, langue_biblio
From mRNA, gene g, b bibliography
Where a.nom_gene = g.nom_gene
And g.nom_gene in (select nom_gene from recepteurs r, mutation m
                   Where r.nom_recepteur = m.nom_recepteur)
```

For its execution, the query is first submitted to the mediator which is responsible for locating sources and queries them through the wrapper or the associated adapter.It should

be noted that the only access point to our sources for interrogation is a web form that, once processed through a wrapper, gives us the local sources that we describe below.

5.5. Analysis of the query

In the global query, 6 attributes are involved: Nom_proteine, nom_gene, journal_biblio, auteur_biblio, date_biblio, langue_biblio shown in the following table along with the sources (PubMed (S1), and PDB (S2).

Attributes	Lists of sources
journal_biblio, auteur_biblio, date_biblio,langue_biblio	S1
nom_proteine	S2
nom_gene	S2, S1

Table 4. Identification of sources

From these sources, we can extract a local schema, for example:

S1_L (journal_biblio, auteur_biblio, date_biblio, langue_biblio, nom_gene),

S2_L (nom_proteine, nom_gene)

From a programming point of view, S1_L and S2_L represent wrapper sources.

Next section describes the realization in which we develop wrappers, submit queries to the mediator, and combine the final result to be presented to the user.

6. Realization

We define four steps in the realization:

- The development of wrapper
- Definition of "global schema"
- Correspondence between local tables and global schema
- Results analysis

6.1. Step 1: Development of wrappers

A wrapper is a program that envelops the execution of another program in the way that the environment can be more suitable. The mediator requests the various databases via wrappers that will extract information from websites of interest. It is necessary to create a wrapper for each specific database.

The sources that we identified in section 4 will be integrated through wrappers. There are different types of wrappers depending on the type of pages they incorporate. These can be either text files or XML files (Extensible Mark-up Language). It is necessary to know the structure of these files and know where the information is located (after any tag, for example). Developing wrappers is linked to functional specifications of the sources presented earlier.

Wrappers allow therefore the extraction of data to be represented in tables. Indeed, we declare the objects and their attributes for each site based on data provided. From all this information, local schemas (relational) for each of these databases are established.

Various programs were written in java. Even if a wrapper has been created for each database, they all have the same main structure. To fill out the fields of tables, the wrapper accesses the Web site to integrate the page and look for keywords behind which is the value to extract. Wrappers are of two types depending on the format of the sources : Either text wrappers or XML wrappers (Figure 8).

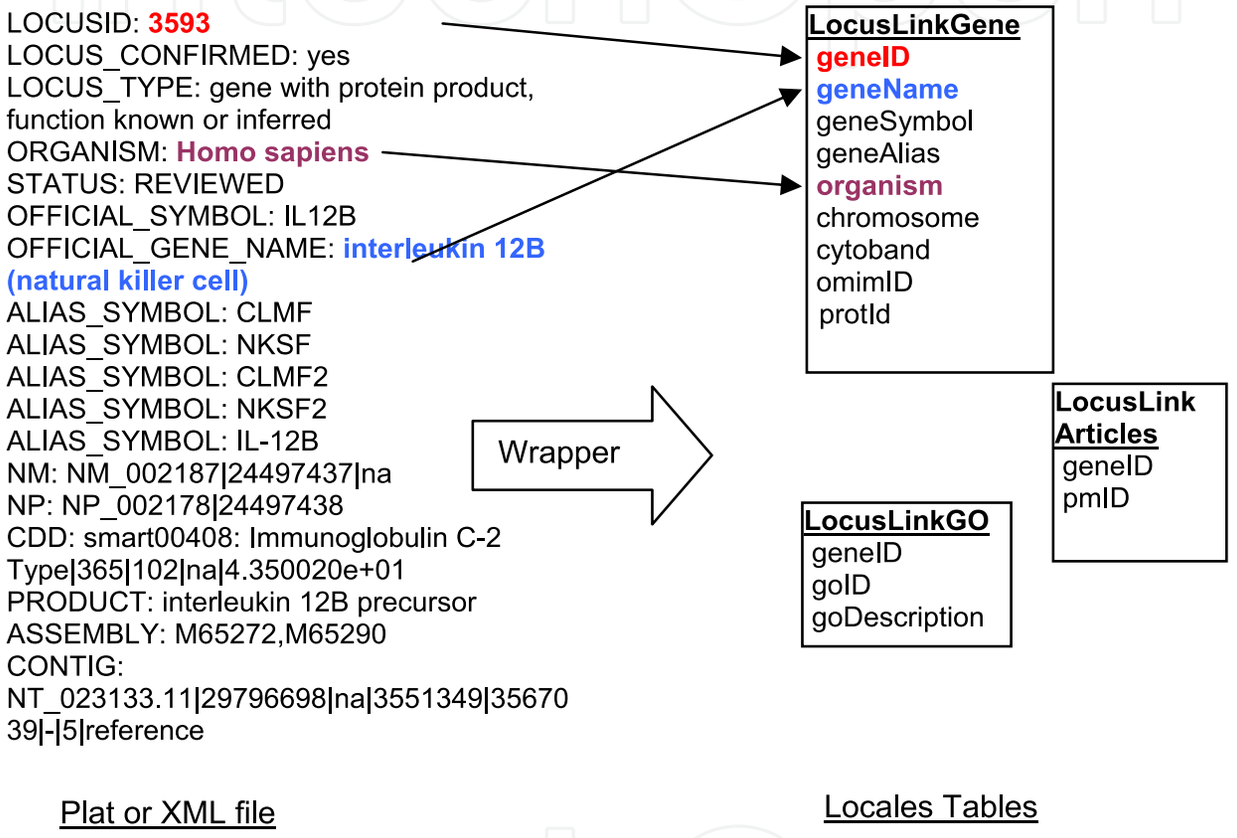

Figure 8. Presentation of a wrapper

Finally, a program that generates and initializes (gives the starting values for all wrappers) is created to coordinate everything. This program is also written in Java and integrates all the wrappers and their relationships. We thus obtain a set of local tables (Provisional) performed by the wrappers.

6.2. Step 2: Definition of "global schema"

It is therefore necessary to build the global schema that will be the only interface for user. Indeed, the user does not know absolutely how the data are integrated. The global schema is a set of relational tables that are defined using local tables (for information). This schema was introduced in the previous section.

6.3. Step 3: Matches between local tables and global schema

As the various local tables have been filled by the wrappers and the global schema has been established, we should now define the correspondence rules between them in order to implement global schema with the extracted information from local schemas. The problem is that several sources may correspond to a business table (we must then join conditions on these tables) or otherwise a source may have several tables trades. For this, we use the Medience server tool.

Medience Server (Figure 5.2) is a complete environment that treats all matching problems (different formats, different representation of business information, dispersal of information described in a single business table). It is a "virtual database", because it does not store information but analyse the user needs. This tool will serve as a mediator that is to say that it will be the unique interface for the user as it will both integrate databases, present data and also offers possibility to loop and see only some information tables of interest to the user. The use of this tool goes in three steps:

1. The first step consists on recording data sources and creating associated source tables using source files provided by wrappers. The global schema is also implemented by local tables. We define the attributes of all tables.
2. In the second part we define the correspondence rules from source tables to the global schema tables. The supply of each table in the global database is defined from the records of source tables. It is thus possible to standardize results coming from various sources by the definition of a standard type.
3. The final step is the verification of all components and installation of all matching rules to make it operational (Figure 9).

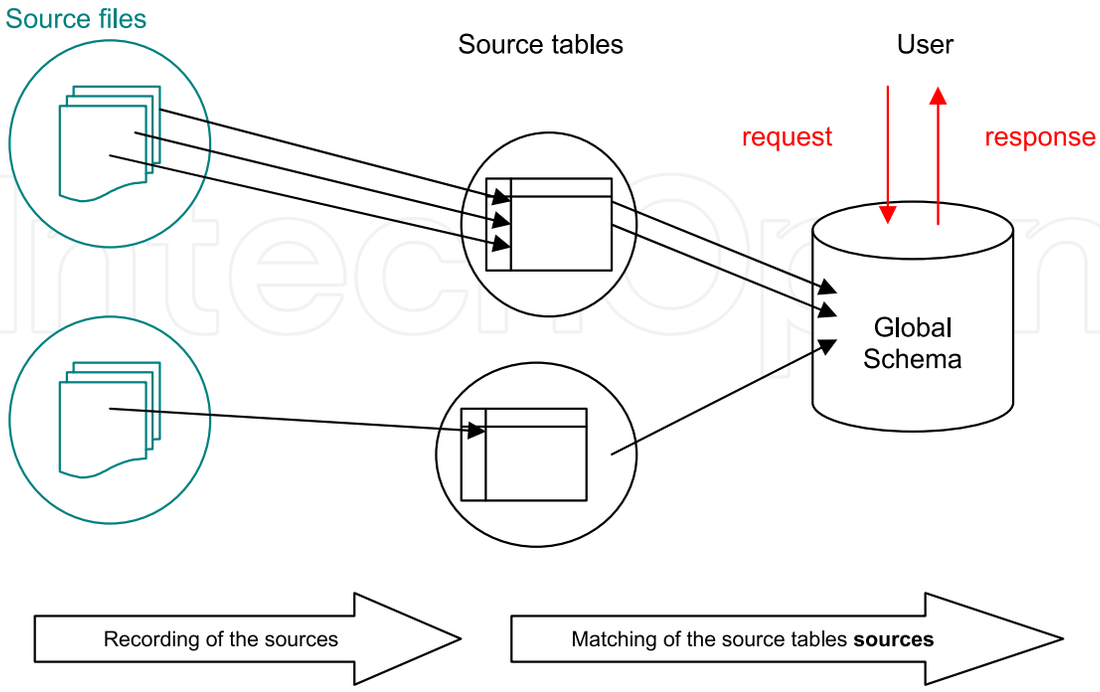

Figure 9. Architecture of Medience

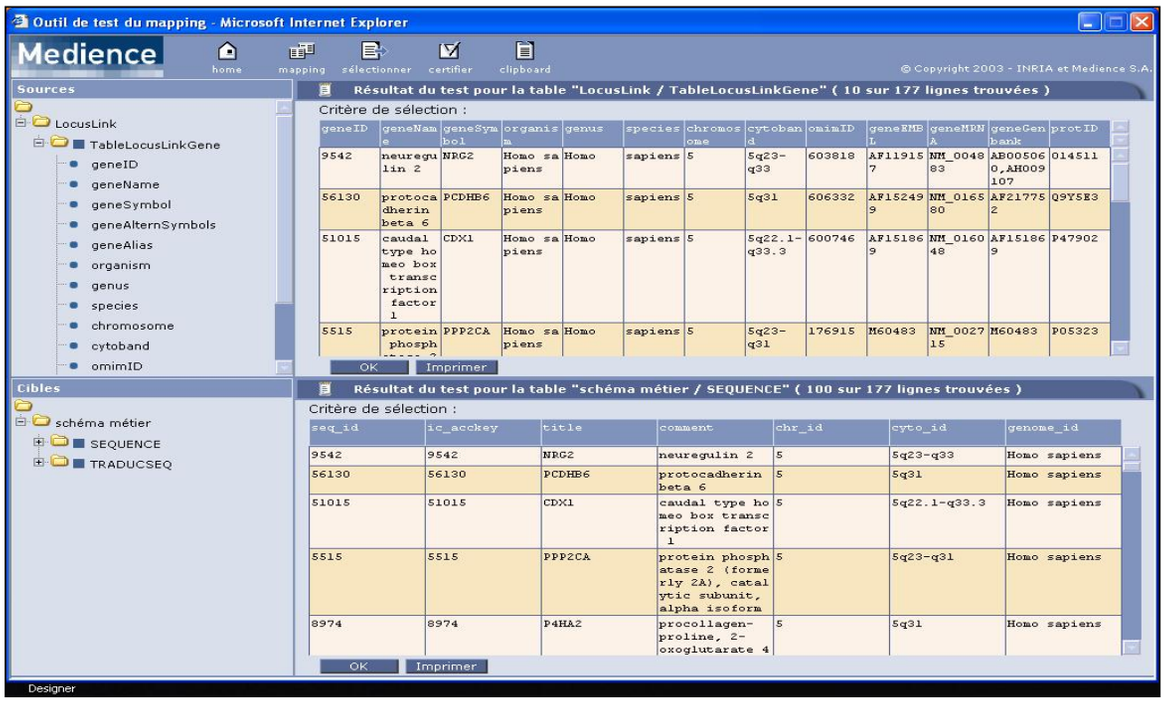

Figure 10. Anexample of Medience interface [19]

6.4. Step 4: Data analysis

It is therefore possible through a platform like Medience to integrate data sources (BD, Excel files, and text files) and view the results in a tabular form. Now, we can process to the analysis of the results. For this, the definition of demand in terms of mining must be decided: How can we use the data provided. Medience offers the possibility to ask tables on the global schema in SQL way. It also offers the ability to define views on these tables and keep a small part that is particularly interesting. It is possible to use our tool to answer the question like what is the protein associated with the mutated gene responsible for familial hypercholesterolemia and related publications.

7. Conclusion

The objective of this work was to develop a system for integrating biological data with an application on familial hypercholesterolemia disease. Such a system should facilitate access to multiple data sources available on the Web, in a transparent and uniform way, giving biologists a single virtual source that summarizes all relevant data sources for the application.

This chapter describes the solution adopted to achieve such a system, where the main elements have been identified, and a computer deployment scenario developed. Among different existing integration approaches, we adopted the mediator approach to integrate data sources. In this approach the most important step is the construction of the global schema as the mediator has to process queries at runtime in order to integrate data sources. We first studied the biologists' needs by exploring different scenarios and we identified with their help various data sources involved.

A study of these sources was necessary in order to build our global schema. From the diagram established, we formulated our SQL query as we built various adapters associated with different sources and at the end we have submitted this request to the mediator for treatment.

As prospects, we have to implement and test this solution and combine the final result of the mediator and that of the tool CHARMM before presenting to the user.

We are currently expanding the platform by integrating other proteins involved in cardiovascular diseases which are the main cause of mortality in the world. In particular, we are investigating a protein called paraoxonase-1 (PON1) which plays an important role in the cardiovascular diseases prevention.

PON1 is an HDL associated enzyme synthesized in the liver and distributed in the blood. It catalyzes the hydrolysis of modified lipids in both HDL (known as good cholesterol) and LDL (known as bad cholesterol) particles and protects them from oxidative modifications, and subsequently reducing the risk of atherosclerosis.

Further bioinformatics analysis including molecular simulations are performed on the PON1 enzyme to better understand the structure activity relationship and also to explore the mutated proteins (genetic polymorphism associated with heart disease) responsible for the weak activity revealed trough the clinical study in both diabetic and coronary patients from Morocco.

Author details

Assia Rharbi and Zohra Bakkoury
Equipe : AMIPS Ecole Mohammadia des Ingénieurs,
Université Mohammed V, Agdal, Rabat – Morocco

Afaf Mikou
Laboratoire GAIA, Spectroscopie Faculté des Sciences Ain Chock –
Université Hassan II, Casablanca – Morocco

Khadija Amine
Laboratoire GAIA, Spectroscopie Faculté des Sciences Ain Chock -
Université Hassan II, Casablanca – Morocco
Laboratoire de Recherche sur les Lipoprotéine et l'Athérosclérose, Unité de Recherche Associée au CNRS-URAC 34-, Faculté des Sciences Ben Msik-Casablanca, Université Hassan II Mohammedia, Morocco

Anass Kettani
Laboratoire de Recherche sur les Lipoprotéine et l'Athérosclérose, Unité de Recherche Associée au CNRS-URAC 34-, Faculté des Sciences Ben Msik-Casablanca, Université Hassan II Mohammedia, Morocco

Abdelkader Betari
ENSA Oujda, Université Mohammed Premier Oujda, Morocco

8. References

[1] M. El Messal, K. Aït Chihab, R. Chater, JC. Vallvé, F. Bennis, A. Hafidi, J. Ribalta, M. Varret, M. Loutfi, JP. Rabès, A. Kettani, C. Boileau, L. Masana, A. Adlouni. Familial Hypercholesterolemia in Morocco: first report of mutations in the LDL receptor gene. J Hum Genet.48 (4):199-203, 2003

[2] http://www.dsi.univ-paris5.fr/bio2/autof2/cha2_1.htm : Bases de données biologiques / Banques généralistes

[3] http://www.dsi.univ-paris5.fr/bio2/autof2/cha2_2.htm : Bases de données biologiques / Banques spécialisées

[4] www.gusdb.org : Genomics Unified Schema (GUS)

[5] Emilie Guérin, Gwenaëlle Marquet, Anita Burgun, Olivier Loréal et Fouzia Moussouni, GEDAW : un environnement intégré pour l'analyse du Transcriptome, JOBIM 2005

[6] http://www.cs.man.ac.uk/~stevensr/tambis/ : TAMBIS

[7] Thomas Hernandez, Subbarao Kambhampati, Integration of Biological Sources: Current Systems and Challenges Ahead, SIGMOD september 2004

[8] http://www.charmm.org/: CHARMM (Chemistry at HARvard Macromolecular Mechanics)

[9] http://www.ks.uiuc.edu/Research/vmd/: VMD (Visual Molecular Dyamic)

[10] www.rcsb.org/pdb/

[11] http://www.ncbi.nlm.nih.gov/sites/entrez?db=PubMed&itool=toolbar

[12] Assia Rharbi, Zohra Bakkoury, Afaf Mikou, Anass Kettani, Abdelkader Betari, and Omar Boucelma, Intégration des données génomiques pour la maladie d'hypercholestérolémie familiale, Journées Scientifiques en Bio-informatique (JSB'2007)

[13] Assia Rharbi, Zohra Bakkoury, Afaf Mikou, Anass Kettani, Abdelkader Betari, and Omar Boucelma, Intégration des données appliquée au domaine biologique, Cinquième Conférence sur les Systèmes Intelligents : Théories et Application (SITA'08)

[14] http://www.rcsb.org/pdb/file_formats/pdb/pdbguide2.2/guide2.2_frame.html

[15] http://rasmol.org/

[16] http://jmol.sourceforge.net/

[17] http://fr.wikipedia.org/w/index.php?title=Chime&action=edit&redlink=1

[18] http://www.w3.org/MarkUp/VRML/

[19] F.-M. Colonna, Thèse : "Intégration de données hétérogènes et distribuées sur le Web et applications à la biologie", Université Paul Cézanne (Aix-Marseille III), Décembre 2008