# Cleaning, Integrating, and Warehousing Genomic Data from Biomedical Resources

**Fouzia Moussouni**

**Université de Rennes 1 France**

**Laure Berti-Équille**

**Institut de Recherche pour le Développement, France**

## 1. Introduction

Four biotechnological advances have been accomplished of the last decade: *i)* sequencing of whole genomes giving chance to the discovery of thousands of genes, *ii)* functional genomics using high-throughput DNA microarrays to measure expression of each of these genes in multiple physiological and environmental conditions, *iii)* scaling of proteins using Proteome to map all the proteins produced by a genome, and *iv)* the dynamics of these genes and proteins in a network of interactions that gives life to any biological activity and phenotype. These major breakthroughs resulted in massive collection of data in the field of Life Sciences. Considerable efforts have been made to sort out, curate and integrate every relevant piece of information from multiple information sources in order to understand complex biological phenomena.

Biomedical researchers spend a phenomenal time to search data across heterogeneous and distributed resources. Biomedical data are indeed available in several public databanks: banks for genomic data (DNA, RNA) like Ensembl, banks for proteins (polypeptides and structures) such as *SWISS-PROT*[1], generalist databanks such as *GenBank*[2], *EMBL*[3] (*European Molecular Biology Laboratory*) and *DDBJ*[4] (*DNA DataBank of Japan*). Other specialized databases exist today to describe specific aspects of a biological entity, including structural data of proteins (*PDB*[5]), phenotype data (*OMIM*[6]), gene interactions (*KEGG*[7]) and gene expression data (*ArrayExpress*[8]). Advances in communication technologies enabled these databases to be worldwide accessible by scientists via the Web. This has promoted the desire to share and integrate the data they contain, for connecting each biological aspect to another, e.g., gene sequence to biological functions, gene to partners, gene to cell, tissue and body locations, signal transductions to phenotypes and diseases, etc. However, semantic heterogeneity has been a major obstacle to the interoperability of these databases, moving to semantic scale the structuring efforts of biomedical information. Since then, interoperability, *i.e.*, the linking of distributed and heterogeneous information items, has become a major problem in bioinformatics. Besides, biological data integration is still error-prone and difficult to achieve without human intervention.

In spite of these barriers, we have assisted in the last decade to an explosion of data integration approaches and solutions to help Life Sciences researchers to interpret their results, test and generate new hypothesis. In high throughput bio-technologies like DNA-chips, data warehouse solutions encountered a great success, because of constant needs to store locally the delivered gene expression data, confront and enrich them with data extracted from other sources, for multiple possibilities of novel analyses.

The Life Sciences data sources are supplied by researchers. They are also accessed by researchers to interpret their results and generate new hypotheses.

However, in case of insufficient mechanisms for characterizing the quality of the data they contain, such as: truthfulness, accuracy, redundancy, inconsistency, completeness, and freshness, data is considered by scientists as a "representation" of reality. Many imperfections in the data are not detected or corrected before integration and analysis. In this context, tremendous amount of data warehouse projects integrate data from various heterogeneous sources, having different degrees of quality and trust. Most of the time, the data are neither rigorously chosen nor carefully controlled for data quality. Data preparation and data quality metadata are recommended but still insufficiently exploited for ensuring quality and validating the results of information retrieval or data mining techniques (Berti-Équille and Moussouni, 2005).
Most-used on-line databanks for Life Sciences are riddled with errors and lots of factors will cause them. The three major sources of data quality problems are the following:

- ***Heterogeneity of data sources:*** Public molecular databases (GenBank, Swiss-Prot, DDBJ, EMBL, PIR, among others) are large and complex artifacts. They already integrate data from multiple sources, and transform it using various programs, scripts and manual annotation procedures that are neither traced, nor documented and reproducible, and that change over time. Extensive duplication, repeated submissions of the sequences to the same or different databases and cross-updating of databases accelerate the propagation of errors within and across the main on-line databanks.

- ***Free-ruled data annotation***: Biological data come from journal literature and direct author submissions for otherwise unpublished sources. There are usually no content restrictions for the submitters or collaborators to present their data to the databanks, even allow them claim patents, copyrights, or other intellectual property rights in all or a portion of the data with very few checking or assessment of the information content validity. Data entry errors can be easily introduced due to the lack of standardized nomenclature, variations in naming conventions (synonyms, homonyms, and abbreviations). In addition, information content may have different interpretations.

- ***Instrumentation/Experimental errors.*** The tools driving the current automated, high-throughput sequencing systems are not infallible. Even a 1% error rate will produce

10 mistakes in every 1000 bases generated by the machine. Due to the unboundary information feature of coding and origin region in genomic sequence data, the researchers of molecular biology have to extract the relevant data from them when performing analysis and addressing specific research. Any data problem or error in the symbol sequences and repetitions may cause misleading and wrong data analysis results or misinterpretations.

- ***Inadequacy of data quality control mechanisms and scalability issues***. Since the data sizes of major public databanks have been increasing exponentially, (e.g., GenBank contains approximately 126,551,501,141 bases in 135,440,924 sequence records in the traditional GenBank divisions and 191,401,393,188 bases in 62,715,288 sequence records in the WGS division as of April 2011), manual data curation still predominates, despite its high cost and obvious problems of scalability (Baumgartner et al., 2007). Systematic approaches to data checking and cleaning are lacking (Buneman et al., 2008).

A wide range of data quality problems may emerge at any time during data life cycle (*i.e.*, data acquisition, assembly, transformation, extraction, integration, storage, internal manipulation, etc.) from primary raw experiment databases to large public databanks and specialized laboratory information management systems (LIMS).
Careful data cleaning and data preparation are very necessary prerequisites to any process of knowledge discovery from integrated biological data.

In this chapter, we review the literature on data integration in the Life Sciences with a particular focus on the approaches that have been proposed to handle biological data quality problems (Section 2). We propose a classification of data quality problems in biomedical resources and we present some of preprocessing solutions that can be practically implemented before any task of data mining (Section 3). Based on our previous work on data cleaning, integration and warehousing of biomedical data, we present the lessons we've learnt and the approach we've implemented in practice (Section 4). Finally, we'll conclude this chapter with some challenging research directions for biomedical data preprocessing and integration (Section 5).

# 2. Related work

The first generations of data integration systems for the Life Sciences were based on flat file indexing (e.g., SRS[9], DBGet[10], Entrez[11], Atlas[12]), multi-database query languages (Kleisli, OPM, P/FDM), and federated databases (DiscoveryLink, BioMediator, caGRID). Recent systems are now mediation systems (or mediators) that consist in connecting fully autonomous distributed heterogeneous data sources. Mediators do not assume that integrated sources will all be relational databases. Instead, integrated resources can be various database systems (relation, object-relational, object, XML, etc.), flat files, etc. The integration component of mediation is in charge of (1) providing a global view of integrated resources to the user, (2) proving the user with a query language to query integrated resources, (3) executing the query by collecting needed data from each integrated resource, and (4) returning the result to the user. For the user, the system provides a single view of the integrated data as it was a single database. Several mediation systems have been designed for domain specific integration of biomolecular data, providing non-materialized views of biological data sources. They include:

- BioKleisli (Davidson et al., 1997, Buneman at al., 1998) and its extensions K2 (Davidson et al., 2001) and Pizzkell/Kleisli (also known as Discovery Hub, Wong, 2000)
- the multi-database system based on the Object Protocol Model (OPM) (Chen & Markovitz, 1995) to design object views (Chen et al., 1997) and its Object-Web Wrapper (Lacroix 2002),
- the DiscoveryLink (Haas et al., 2001),
- P/FDM (Kemp et al., 1999, 2000) and
- TAMBIS (Baker et al., 1998).

Indeed mediation systems often offer an internal query language that allows the integration of (new) resources (data and tools) in addition to a user's query language that is used by biologists to access, analyze, and visualize the data. Existing mediation approaches rely on traditional database query languages (e.g., SQL, OQL). As a particular example of ontology-based integration, TAMBIS (Baker et al., 1998) is primarily concerned with overcoming semantic heterogeneity through the use of ontologies. It provides users an ontology-driven browsing interface. Thus it restricts the extent to which sources can be exploited for scientific discovery.

To summarize, these systems have made many inroads into the task of data integration from diverse biological data sources. They all rely on significant programming resources to adjust to specific scientific tasks. They are also difficult to maintain and provide user's query language that requires programming ability (such as SQL, OQL, Daplex, etc.) and significantly limit the query capabilities.

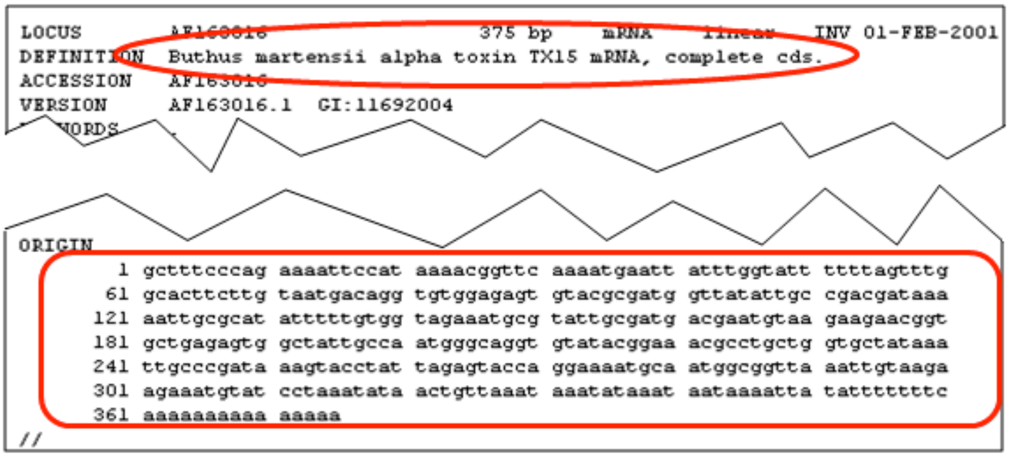
However, none of the existing systems allows the management of data quality metadata and none of them offers the flexibility of customization for ETL (Extract-Transform-Load) or data preprocessing tasks. These functionalities may be partially covered by emerging scientific workflow management systems (Cohen-Boulakia & Leser, 2011; Ives, 2009) emphasizing data provenance as a critical dimension of biological knowledge discovery (Cohen-Boulakia & Tan, 20009).

# 3. Typology of data quality problems in biomedical resources

We can classify data quality problems occurring in biomedical resources into the following categories illustrated with relevant examples:

- ***Redundancy***: Redundant or duplicated data are mainly caused by over-submission. This category is due to overlapping annotations and replication of identical sequence information, e.g., the same sequence can be submitted to different databases or submitted several times to the same database by different groups, and/or the protein sequence may be translated from the duplicate nucleotide sequence and several records may contain fragmented or overlapping sequences with more or less complete sequences. The redundancy problem often comes along with partial incompleteness of records and more generally it is caused by the evolving nature of knowledge. Extensive redundancy is caused by records containing fragmented or overlapping sequences with more complete sequences in other records (see Example 1 for illustration).

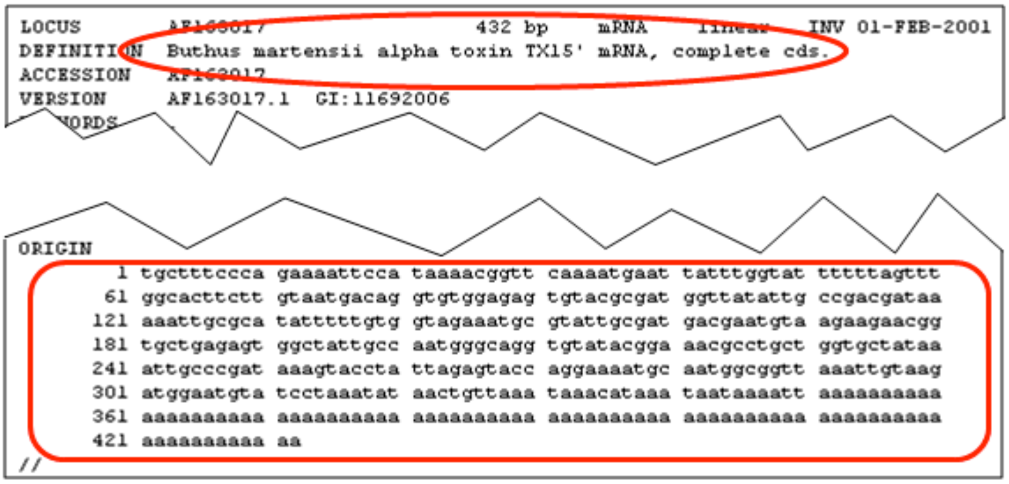
  **Example 1. *Redundancy.*** Consider two records describing the same biological entity, GI:11692004 and GI:11692006 respectively from NCBI nucleotide databank presented in Figure 1. The only difference between the two records relies on the sequence length. The record GI:11692006 provides additional irrelevant bases “a”.



```
LOCUS       AF163016                 375 bp    mRNA    linear   INV 01-FEB-2001
DEFINITION  Buthus martensii alpha toxin TX15 mRNA, complete cds.
ACCESSION   AF163016
VERSION     AF163016.1  GI:11692004
KEYWORDS    .

ORIGIN
        1 gctttcccag aaaattccat aaaacggttc aaaatgaatt atttggtatt ttttagtttg
       61 gcacttcttg taatgacagg tgtggagagt gtacgcgatg gttatattgc cgacgataaa
      121 aattgcgcat atttttgtgg tagaaatgcg tattgcgatg acgaatgtaa gaagaacggt
      181 gctgagagtg gctattgcca atgggcaggt gtatacggaa acgcctgctg gtgctataaa
      241 ttgcccgata aagtacctat tagagtacca ggaaaatgca atggcggtta aattgtaaga
      301 agaaatgtat cctaaatata actgttaaat aaatataaat aataaaatta tatttttttc
      361 aaaaaaaaaa aaaaa
//
```

http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&val=11692004



```
LOCUS       AF163017                 432 bp    mRNA    linear   INV 01-FEB-2001
DEFINITION  Buthus martensii alpha toxin TX15' mRNA, complete cds.
ACCESSION   AF163017
VERSION     AF163017.1  GI:11692006
KEYWORDS    .

ORIGIN
        1 tgctttccca gaaaattcca taaaacggtt caaaatgaat tatttggtat tttttagttt
       61 ggcacttctt gtaatgacag gtgtggagag tgtacgcgat ggttatattg ccgacgataa
      121 aaattgcgca tatttttgtg gtagaaatgc gtattgcgat gacgaatgta agaagaacgg
      181 tgctgagagt ggctattgcc aatgggcagg tgtatacgga aacgcctgct ggtgctataa
      241 attgcccgat aaagtaccta ttagagtacc aggaaaatgc aatggcggtt aaattgtaag
      301 atggaatgta tcctaaatat aactgttaaa taaacataaa taataaaatt aaaaaaaaaa
      361 aaaaaaaaaa aaaaaaaaaa aaaaaaaaaa aaaaaaaaaa aaaaaaaaaa aaaaaaaaaa
      421 aaaaaaaaaa aa
//
```

http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&val=11692006

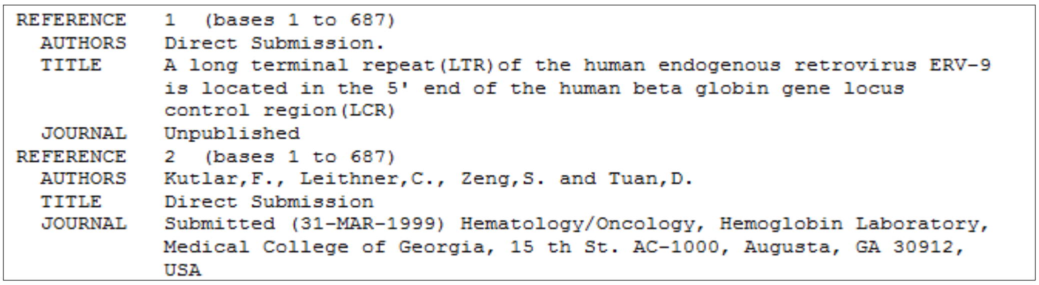
**Figure 1.** Example of two redundant records with uninformative sequence portions

- ***Incompleteness***: Paradoxically, over-submission does not prevent from submission of incomplete records and fragmented information from one record to another with potentially overlapping or conflicting data.
- ***Inconsistency***: Multiple database records of the same nucleotide or protein sequences contain inconsistent or conflicting feature annotations. This category includes data entry errors, misspelling errors, mis-annotations of sequence functions, different expert interpretations, and inference of features or annotation transfer based on best matches of low sequence similarity. Problematic data that lack of domain consistency, such as contaminated data existing in coding region due to unsure

reasons, outdated, missing and discrepant annotations comparing with other databanks. Various kinds of inconsistency may occur:

- ***Syntax errors:*** The syntax errors are violations of syntactic constraints on particular format/fields of the databank record.
- ***Semantics errors:*** Semantics errors contain data field discrepancy, invalid data content identified either by the databank flatfile format or other NCBI specifications. For examples, invalid MedLine or PubMed numbers, invalid reference number, etc. Another type of error is the mis-use of fields when data content does not correspond to the field usage (see Example 2).
- ***Naming ambiguities:*** The manifestation of synonyms, homonyms and abbreviations results in information ambiguities which cause problems in biological entity identification and keyword searching. For example, BMK stands for "Big Map Kinase", "B-cell/myeloid kinase", "bovine midkine", as well as for "Bradykinin-potentiating peptide". The scorpion neurotoxin BmK-X precursor has a permutation of synonyms. It is also known as "BmKX", "BmK10", "BmK-M10", "Bmk M10", "Neurotoxin M10", "Alpha-Neurotoxin TX9", and "BmKalphaTx9".
- ***Undersized/oversized fields***: Sequences with meaningless content can be found in protein records queried using Entrez to the major protein or translated nucleotide databases: these are protein sequences shorter than four residues and sequences shorter than six bases. The undersized fields may alter the entity identification: e.g., "M" is the synonym of the protein "ACTM_HELTB" (record GI:1703137) but "M" also corresponds to 1,389,441 records on NCBI protein database.
- ***Cross-annotations with conflicting values***: Multiple database records of the same nucleotide or protein sequences may contain conflicting feature annotations, data entry errors, mis-annotation of sequence functions, different expert interpretations, and inference of features or annotation transfer based on best matches of low sequence similarity (see Example 2).
- ***Putative information***: Functional annotation sometimes involves searching for the highest matching annotated sequence in the database. Features are then extrapolated from the most similar known searched sequences. In some cases, even the highest matching sequence from database search may have weak sequence similarities and therefore does not share similar functions as the query sequence. "Blind" inference can cause erroneous functional assignment.

**Example 2. *Inconsistency***. Consider the bibliographic reference provided in the record GenBank: AF139840.1 presented in Figure 2.



```
REFERENCE   1  (bases 1 to 687)
  AUTHORS   Direct Submission.
  TITLE     A long terminal repeat(LTR)of the human endogenous retrovirus ERV-9
            is located in the 5' end of the human beta globin gene locus
            control region(LCR)
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 687)
  AUTHORS   Kutlar,F., Leithner,C., Zeng,S. and Tuan,D.
  TITLE     Direct Submission
  JOURNAL   Submitted (31-MAR-1999) Hematology/Oncology, Hemoglobin Laboratory,
            Medical College of Georgia, 15 th St. AC-1000, Augusta, GA 30912,
            USA
```

**Figure 2.** Example of mis-use of the bibliographic references field

This record and sequence information has been directly submitted to GenBank and they don't correspond to a peer-reviewed publication *stricto sensu*.

- ***Irrelevancy***: Less meaningful, nonsense or irrelevant data existing in free-text field of annotation or description, e.g., coding region, which intervene with the target analysis. Some values of finer granularity may be concatenated and automatically imported into a data field of coarser granularity. These values are so-called misfielded (see Example 3).
    - ***Uninformative features or data***: A profuse percentage of the unknown residues (“X”) or unknown bases (“N”) can reduce the complexity of the sequence and thus, the information content of the sequence.
    - ***Contaminated data***: Introns and exons must be non-overlapping except in cases of alternative splicing. But in some erroneous records, nucleotide sequences have overlapping intron/exon region and some sequences can possibly be contaminated with vectors commonly used for the cloning.

**Example 3. *Irrelevancy.*** Consider the following `DEFINITION` field of the protein record AAB25735.1 (http://www.ncbi.nlm.nih.gov/protein/AAB25735.1): it includes the species, the sequence length, etc. These additional information items are irrelevant and mis-fielded.

```
DEFINITION neurotoxin, NTX [Naja naja=Formosan cobra, ssp. atra, venom, Peptide, 62 aa]
```

- ***Obsolescence***: Instead of checking existing records related to the biological entity of interest and updating one of them, users may prefer to submit a new record. This may increase not only the inter-record redundancy and overlaps in the databank but it also has two consequences, first on increasing the difficulty to achieve entity resolution and correctly group together the records that may be truly related to the same biological entity, and second on keeping out-of-date records with misleading or no longer valid knowledge elements.

Table 1 summarizes a categorization of potential intra-record data quality problems into categories and the fields they can affect in a traditional record content.

| Categories | Data quality problems | Record Fields | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Global Identifier | Definition | Taxonomy | References | Cross-Links | Feature annotations | Raw data |
| Inconsistency | Typo/ Mis-spelling | | X | | X | | | X |
| | Format violation | | X | | X | | X | X |
| | Ambiguous naming (homonyms, synonyms, abbreviations) | | X | X | X | | X | X |
| | Mis-fielded values | | X | | X | | X | X |
| | Undersized/ over-sized field | | X | | | | X | X |
| | Measurement error , Contaminated data | | | | | X | X | X |
| | syntax errors and format violations | | X | X | X | X | X | X |
| Irrelevancy | Putative information | | X | | | | X | X |
| | Uninformative data | | X | | X | | X | X |
| Incompleteness | Incomplete data / default values | | X | X | X | X | X | X |
| Obsolescence | Out-of-date data | X | X | | X | X | X | X |

***Table 1**. Categorization of potential intra-record data quality problems.*

Since redundancy can be observed from a group of records, it can be classified as inter-record data quality problem. In the next table, we present the existing solutions for consolidating data both at the intra- and inter-record levels. These solutions are based on integrity, format and constraint checking, comparative analysis and duplicate detection depending on the type of data quality problem.

| Categories | Data quality problems | Attribute-based solutions | Intra-record solutions | Inter-record solutions |
|---|---|---|---|---|
| Inconsistency | Typo/ Mis-spelling | Dictionary look-up | Constraint checking | Duplicate detection |
| | Ambiguous naming (homonyms, synonyms, abbreviations) | | Entity resolution | |
| | Mis-fielded values | Integrity constraints | Constraint checking | |
| | Format violation | | Formatting ETL | Schema remapping |
| | Undersized/ over-sized field | | Size checking | Comparative analysis |
| | Measurement error , Contaminated data | Vector screening, sequence structure parser | Constraint checking | Comparative analysis |
| | syntax errors and format violations | Format checking | | |
| Irrelevancy | Putative information | Keywords search | | Comparative analysis |
| | Uninformative data | Constraint checking | Constraint checking | |
| Incompleteness | Incomplete data / default values | | | |
| Obsolescence | Out-of-date data | | | |

***Table 2**. Practical solutions to biological data quality problems.*

# 4. Cleaning, integrating and warehousing biomedical data

Within this specific context, the aim of this section is to report on our experience during the design of GEDAW, the Gene Expression Data Warehouse (Guérin et al., 2005) and the implementation of the biomedical data integration process in the presence of syntactic and semantic conflicts. We will precisely point out on the lessons learned from data pre-processing and propose the different but complementary solutions we have adopted for quality aware data integration.

## *4.1. Lessons learned from integrating and warehousing biomedical data on liver genes and diseases*

Liver diseases, including those from infectious, alcoholic, metabolic, toxic and vascular etiology, are a major public health problem. They are frequently complicated by the occurrence of acute liver failure or the development of cirrhosis and liver cancer which

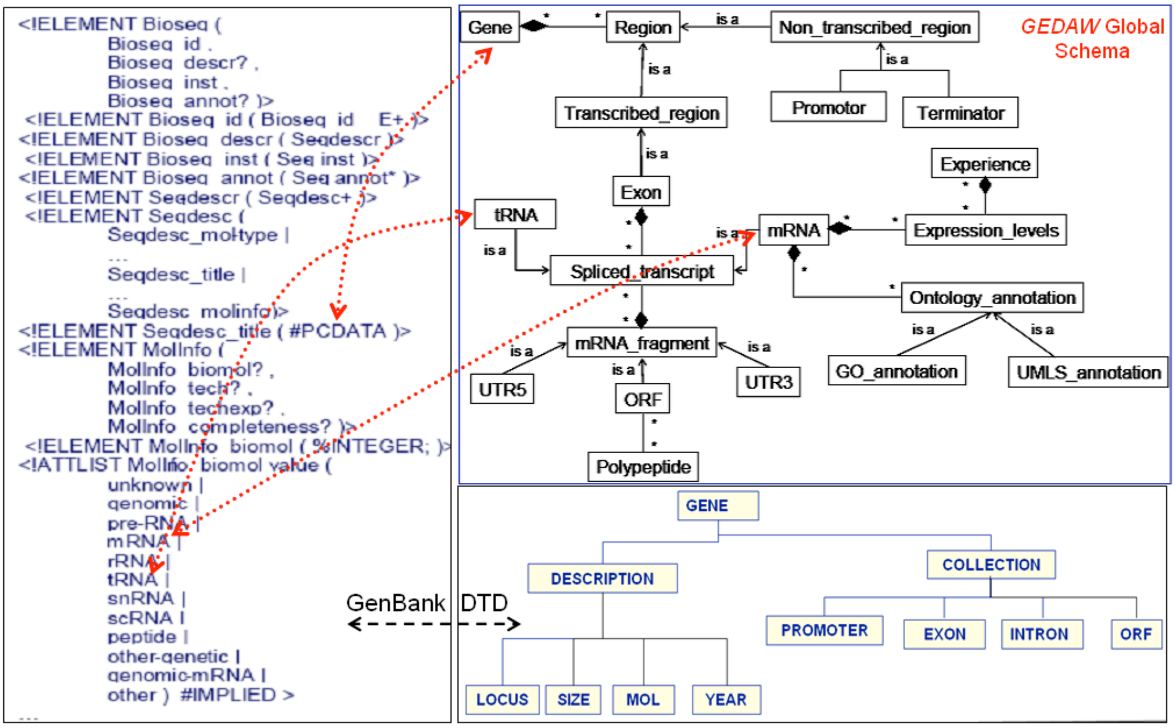
shorten life expectancy. Molecular mechanisms involved in the occurrence of these diseases and of their complications are still not well known. Ongoing researches focus on identifying new relative molecular mechanisms leading to new diagnostic and therapeutic tools.
One way to study liver diseases and correlated complications is the use of DNA-Chips technologies for high-throughputs transcriptome study. Using this technology, thousands of genes can be studied simultaneously, in order to find out the subset of genes that are abnormally expressed in injured tissues, and that gives an attractive big turn in delivering new knowledge on gene networks and regulation mechanisms.

However, the data generated on gene expression are massive and involve difficulties in their management and analysis. Furthermore, for the interpretation of a single gene expression measurement, the biologist has to consider the available knowledge about this gene on different databanks, including its chromosomal location, relative sequences with promoters, molecular function and classification, biological processes, gene interactions, expressions in other physio-pathological situations, clinical follow-ups and an increasingly important bibliography.

The Gene Expression DAta Warehouse *GEDAW*, we have developed at the *National Medical Research Institute* (INSERM), stores data on genes expressed in the liver during iron overload and liver pathologies. Relevant information from public databanks, DNA chips home experiments and medical records have been integrated, stored and managed in GEDAW for globally analyzing the delivered gene expression measurements.

*GEDAW* aimed at *in silico* studying liver pathologies by using expression levels of genes in different physiological situations, enriched with annotations extracted from the variety of the scientific data sources, ontologies and standards in Life Sciences and medicine. For the case of GenBank, each record, usually associated to a gene, describes the genomic sequence with several annotations and is identified by a unique accession number. It may also be retrieved by keywords (cf. Figure 5. GenBank screen shots for HFE Gene). Annotations may include the description of the genomic sequence: function, size, species for which it has been determined, related scientific publications and the description of the regions constituting the sequence (codon start, codon stop, introns, exons, ORF, etc.).

However, designing a single global data warehouse schema that integrates syntactically and semantically many heterogeneous Life Sciences data sources is a challenging task. Only structured and semi-structured data sources were used to integrate *GEDAW*, using a *Global As View* (GAV) schema mapping approach and a rule-based transformation process from a given source schema to the global schema of the data warehouse (cf. Figure 3). As an almost hands-off integration method, this technique was quite advanced at this time, comparing to previous developed warehouses like (Paton et al., 2000) for which yeast data were completely flat.



```
<!ELEMENT Bioseq (
            Bioseq id ,
            Bioseq descr? ,
            Bioseq inst ,
            Bioseq annot? )>
 <!ELEMENT Bioseq id ( Bioseq id    E+ )>
<!ELEMENT Bioseq  descr ( Seqdescr )>
 <!ELEMENT Bioseq  inst ( Seq inst )>
<!ELEMENT Bioseq  annot ( Seq annot* )>
 <!ELEMENT Seqdescr ( Seqdesc+ )>
 <!ELEMENT Seqdesc (
            Seqdesc_moltype |
            ...
            Seqdesc_title |
            ...
            Seqdesc  molinfo)>
<!ELEMENT Seqdesc_title ( #PCDATA )>
<!ELEMENT MolInfo (
            MolInfo  biomol? ,
            MolInfo  tech? ,
            MolInfo  techexp? ,
            MolInfo  completeness? )>
 <!ELEMENT MolInfo  biomol ( %INTEGER; )>
<!ATTLIST MolInfo  biomol value (
            unknown |
            genomic |
            pre-RNA |
            mRNA |
            rRNA |
            tRNA |
            snRNA |
            scRNA |
            peptide |
            other-genetic |
            genomic-mRNA |
            other )  #IMPLIED >
...
```

**Figure 3.** Mapping GenBank DTD to *GEDAW*

Figure 3 gives a synthesized Class diagram of *GEDAW* and some correspondences with the *GenBank* DTD (e.g., *Seqdes_title* and *Molinfo* values were extracted, transformed and migrated to other description attributes of the class *Gene* in the *GEDAW* global schema). The GEDAW system presented in (Guérin et al., 2005) allows massive import of biological and medical data into an object-oriented data warehouse that supports transcriptome analyses specific to the human liver. It focused on the relevant genomic, biological and medical resources that have been used to build GEDAW. The integration process of the full sequence annotations of the genes expressed was performed by parsing and cleaning the corresponding XML description in GenBank, transforming the recorded genomic items to persistent objects and storing them in the warehouse. This process is almost systematic because another aspect related to the conciliation of duplicate records has been added. Elements of formalization of expertise rules for mapping such data were given. This ongoing work is still a difficult problem in information integration in Life Sciences and has not yet satisfied answers by classical solutions proposed in existing mediation systems. In order to lead strong analysis on expressed genes and correlate expression profiles to liver biology and pathological phenotype, a second way of annotation has been added to the integration process.

## *4.2. Data quality-aware solutions*

Different input data sources have been considered during the built of GEDAW: *i) GenBank* for the genomic features of the genes, *ii)* annotations of genes in biomedical ontologies and terminologies (such as *UMLS*[13], *MeSH*[14] and *GO*[15]), and *iii)* gene expression measurements generated in different physiological conditions.
Because gene expression data is massive (more than two thousands measures per experiment and a hundred of experiments per gene), the use of schema integration in our case – *i.e.*, the replication of the source schema in the warehouse - would highly burden the data warehouse.
By using a Global as View (*GAV)* mapping approach for integrating one data source at a time (cf. Figure 3 for *GenBank*), we have minimized as much as possible the problem of identification of equivalent attributes. The problem of equivalent instance identification is still complex to address. This is due to general redundancy in the occurrence of a biological entity even within one data source. As we pointed out in Section 3, biological databanks may have inconsistent values of equivalent attributes referring to the same real-world object. For example, in GenBank, there are more than 10 data forms associated to the same human HFE gene, a central gene associated to iron uptake! Obviously the same segment could be a clone, a marker or a genomic sequence.

This is mainly due to the fact that Life Sciences researchers can submit any biological information to public databanks with more or less formalized submission protocols that usually do not include names standardization or data quality controls. Erroneous data may be easily entered and cross-referenced. Even if some tools propose clusters of records (like *LocusLink*[16] for *GenBank*, more recently called *EntryGene*) to identify a same biological concept across different biological databanks for being semantically related, biologists still must validate the correctness of these clusters and resolve interpretation of differences between records.

Entity resolution and record linkage is required in this situation. It is even augmented and made more complex due to the high-level of expertise and knowledge it requires *(i.e.*, difficult to formalize because related to many different sub-disciplines of biology, chemistry, pharmacology, and medical sciences). After the step of biological entity resolution, data are scrubbed and transformed to fit the global data warehouse schema with the appropriate standardized format for values, so that the data meets all the validation rules that have been decided upon by the warehouse designer. Problems that can arise during this step include null or missing data; violations of data type, non-uniform value formats, and invalid data.

### 4.2.1. Biological entity resolution and record linkage

As the first preprocessing step for data integration, the process of entity identification, resolution and record linkage has to be performed using a sequence of increasingly sophisticated linkage techniques, described in the following, and also additional knowledge bases, ontologies and thesaurus (such as *UMLS Metathesaurus and MeSH-SR* vocabulary*),* each operating on the set of records that were left unlinked in the previous phase:

1. Linkage based on exact key matching: *i.e.*, based on gene names and cross-referenced accession numbers (for instance between a gene from *HGNC*[17] and a protein in *SWISS-PROT*),

2. Linkage based on nearly exact key matching (*i.e.*, based on all the synonyms of a term and all the identifiers of a gene or gene product in *HGNC*, the *UMLS Metathesaurus* and *MeSH-SR* and in the cluster of records proposed by *EntryGene*),

3. Probabilistic linkage based on the full set of comparable attributes (*i.e.*, based on the search for information about a gene or a gene product: the set of concepts related to this gene in the *Gene Ontology* (*Molecular Function* (F), *Biological Process* (P) and *Cellular Component* (C)) and the set of concepts related to the gene in *UMLS* and *MedLine*[18] abstracts (including chemicals & drugs, anatomy, and disorders),

4. Search for erroneous links (false positives),

5. Analysis of residual data and final results for biological entity resolution.

As an example, consider data related to *Ceruloplasmin*, a gene expressed mainly in the liver and involved in iron metabolism through its ferroxidase activity, which is dependent of the copper charge of the protein. Relative disease, called *Aceruloplasminemia*, is a genetic disease responsible of iron overload (Loreal et al., 2002). The level of plasmatic ceruloplasmin is modulated during various chronic liver diseases (Laine et al., 2002).

As shown in Figure 4, a first phase of linkage based on a search of *Ceruloplasmin* in *GOA*[19] database and *HGNC* provides related terms and returns the corresponding accession numbers in *GeneEntry* (1356) or SWISS-*PROT*, approved gene name (*Ceruloplasmin ferroxidase*), and gene symbol (*CP*). The accession number can then be used to find information in external sources.

Another search of the term on *Gene Ontology* returns the set of concepts of each of the categories F, P and C. From the *UMLS* context, terms associated to *Ceruloplasmin* in the *Metathesaurus* and terms that co-occur with *Ceruloplasmin* in *MedLine* are extracted and *MedLine* abstracts are made accessible.

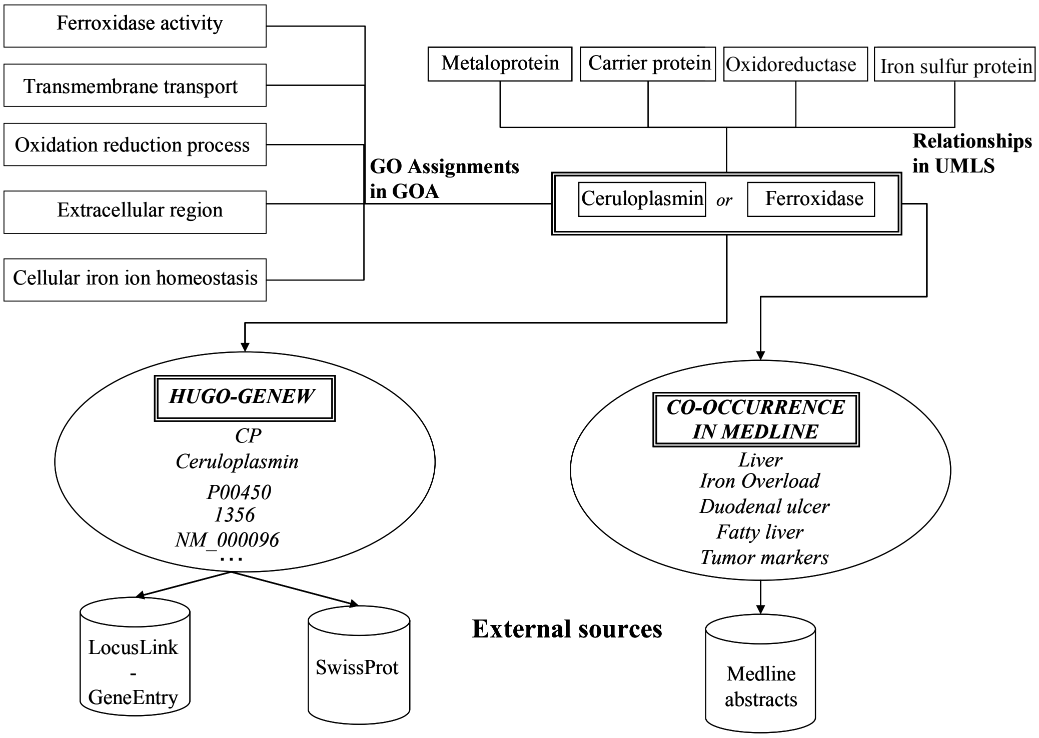


**Figure 4**. Entity resolution and record linkage of *Ceruloplasmin* gene

Indeed, in our experience, combining medical and molecular biology knowledge provides valuable information about genes, e.g., *Ceruloplasmin* is involved in molecular functions such as iron transport mediation, and has relationships to diseases like, *Iron overload* and *Duodenal ulcer*. It can be used to support various tasks to cluster genes according to their properties. Moreover, integration is required for better understanding of disease-molecular data relationships. All these functionalities are presented with more details in (Guérin et al., 2006).

## 4.2.2. Biomedical data scrubbing and conflict resolution

In order to define an appropriate data aggregation of all the available information items resulting from the previous step of biological entity resolution, data conflicts have to be resolved using rules for mapping the source records and conciliating different values recorded for a same concept.

Mapping rules have been defined to allow data exchange from public databanks to *GEDAW*. Apart from experimental data, public information items are automatically extracted by scripts using the *DTD* (*Document Type Definition*) of the data source translated into the *GEDAW* conceptual data model.

Three categories of mapping rules were proposed for *GEDAW*: 1) structural mapping rules, 2) semantic mapping rules and 3) cognitive mapping rules according to the different knowledge levels involved in the biological interpretation of data.

***Structural mapping rules*** are defined at the schema level according to the *GEDAW* model by identifying the existing correspondences with relevant DTD elements; e.g., in Figure 3, the *Seqdesc_title* element in *GenBank* DTD is used to extract the attribute *Name* of the gene and the *MolInfo_biomol* value to determine the type of molecule.

***Semantic and cognitive mapping rules*** are used for data unification at the instance level: several rules may use available tools for determining analogies between homologous data (such as sequence alignment). The result of the *BLAST* algorithm (*Basic Local Alignment Search Tool)* implemented as a set of similarity search programs allows considering that two genomic sequences match.

The nomenclature provided by the entity resolution and record linkage phase, described in the previous section is also considerably used to conciliate duplicate records, based on several ontologies, like *UMLS* that covers the whole biomedical domain and *Gene Ontology*™ *(GO)* that focuses on genomics, as well as additional terminologies, as that provided by the *HUman Genome Organisation (HUGO) Gene Nomenclature Committee (HGNC)* to resolve synonymy conflicts.

More semantic mapping rules are built using this information during the integration process. For example, the *Gene-ID* is used to cluster submitted sequences (DNA, mRNA and Proteins) associated to a same gene with cross-referenced records in *GeneEntry databank* and the official gene name along with its aliases to relate different gene name appearances in literature. These aliases are also stored in the data warehouse and used to tackle the mixed or split citation problems similar to those studied by (Lee et al., 2005) in Digital Libraries.

**Example 4.** Three distinct records are obtained from *GenBank* Nucleotide databank by querying the DNA sequence for the human gene *HFE*, as partially presented in Figures 5, 6 and 7 respectively.

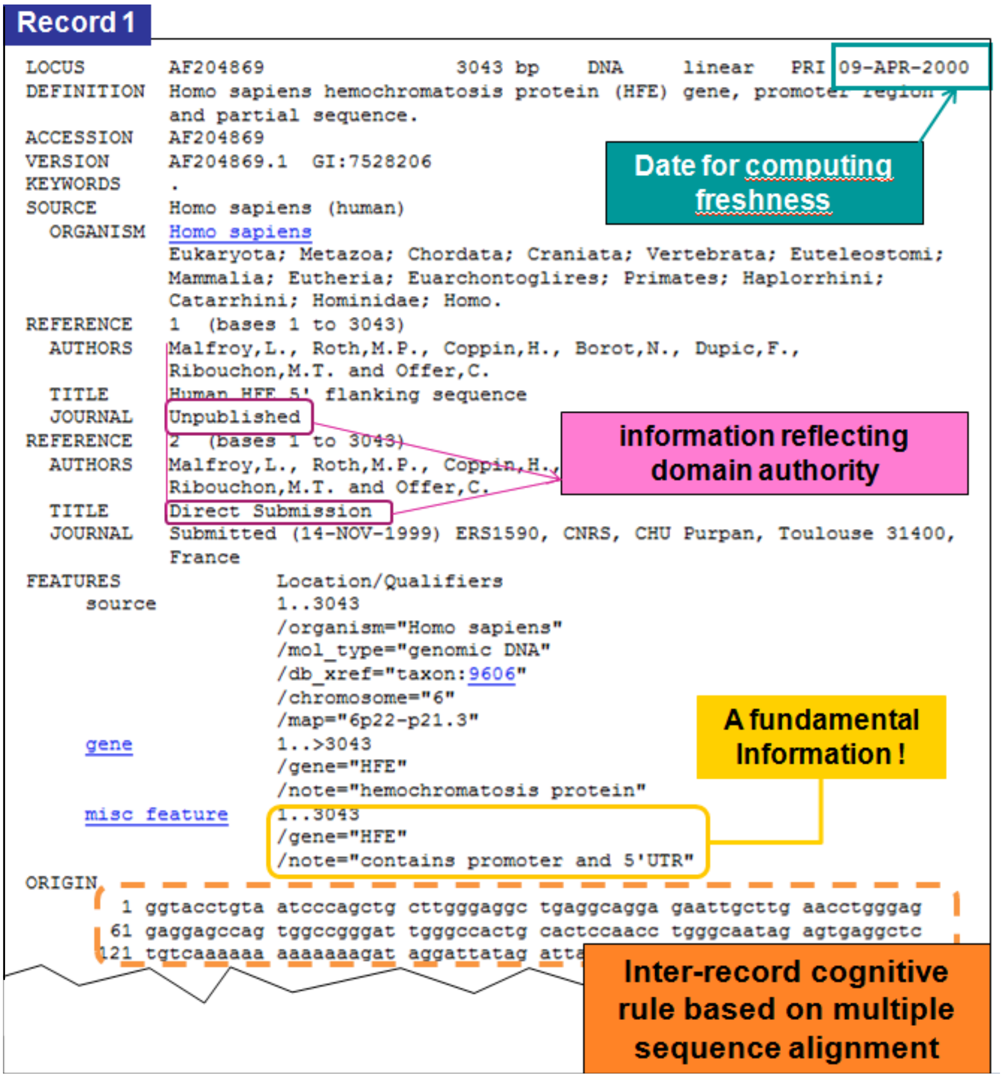
- A first record 1 identified by the accession number *AF204869* describes a partial gene sequence (size = 3043) of the *HFE* gene[20] with no annotation but one relevant and fundamental information item about the position of the promoter region at [*1..3043]* in the "*misc_feature"* field which cannot be found in the other records.
- A second record 2 identified by the accession number *AF184234* describes a partial sequence (size = 772) of the protein precursor of *HFE* gene[21] with a detailed but incomplete annotation.
- The third record 3 identified by the accession number *Z92910* describes the complete gene sequence (size = 12146) of the *HFE* gene[22] with a complete annotation.

We need to integrate this information and to evaluate the quality of these three records because they are complementary regarding to the biological topic of interest (*i.e.*, *HFE* human gene). The first record has a relevant data item that the other records do not have, the second record overlaps the third one regarding the gene sequence but provide more detailed annotations and the third record is complete regarding the gene sequence. This example shows the main quality criteria we use: *i.e.* completeness, relevancy and detail level of annotation.

In this example, using the *BLAST* algorithm for determining the sequence alignment between the two sequences of the records 2 and 3 shows 100% of alignment. This indicates that the sequence in both records 2 and 3 are perfectly identical and can be merged. The detailed annotation of record 2 can be concatenated with the more complete annotation of record 3 in the data warehouse.

Several cognitive mapping rules may be used in this example for conciliating data such as the position offset: in the record 3 the fourth exon is located at position 6494 and in the record 2 this same exon is located at the relative position 130, thus using overlapping information that identifies the same entities, we can deduce the position offset and use the following cognitive rule such as:

*record(AF18423)/exon[number>=4]/position = record(Z92910)/exon[number >=4]/position – 6364*



Record 1

```
LOCUS       AF204869                3043 bp    DNA     linear   PRI 09-APR-2000
DEFINITION  Homo sapiens hemochromatosis protein (HFE) gene, promoter region
            and partial sequence.
ACCESSION   AF204869
VERSION     AF204869.1  GI:7528206
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 3043)
  AUTHORS   Malfroy,L., Roth,M.P., Coppin,H., Borot,N., Dupic,F.,
            Ribouchon,M.T. and Offer,C.
  TITLE     Human HFE 5' flanking sequence
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 3043)
  AUTHORS   Malfroy,L., Roth,M.P., Coppin,H.,
            Ribouchon,M.T. and Offer,C.
  TITLE     Direct Submission
  JOURNAL   Submitted (14-NOV-1999) ERS1590, CNRS, CHU Purpan, Toulouse 31400,
            France
FEATURES             Location/Qualifiers
     source          1..3043
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:9606"
                     /chromosome="6"
                     /map="6p22-p21.3"
     gene            1..>3043
                     /gene="HFE"
                     /note="hemochromatosis protein"
     misc_feature    1..3043
                     /gene="HFE"
                     /note="contains promoter and 5'UTR"
ORIGIN
        1 ggtacctgta atcccagctg cttgggaggc tgaggcagga gaattgcttg aacctgggag
       61 gaggagccag tggccgggat tgggccactg cactccaacc tgggcaatag agtgaggctc
      121 tgtcaaaaaa aaaaaaagat aggattatag atta
```

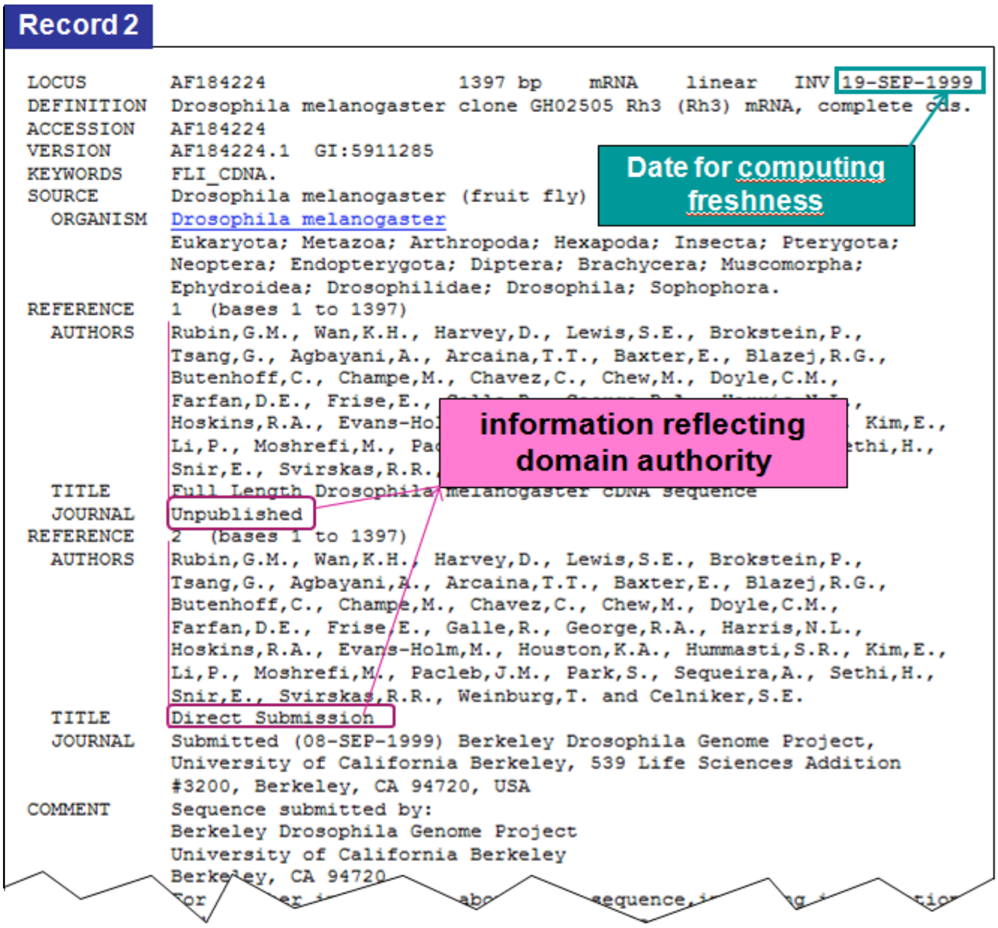
Date for computing freshness

information reflecting domain authority

A fundamental Information !

Inter-record cognitive rule based on multiple sequence alignment

**Figure 5.** *GenBank* Screen Shot for HFE Gene: Record AF204869

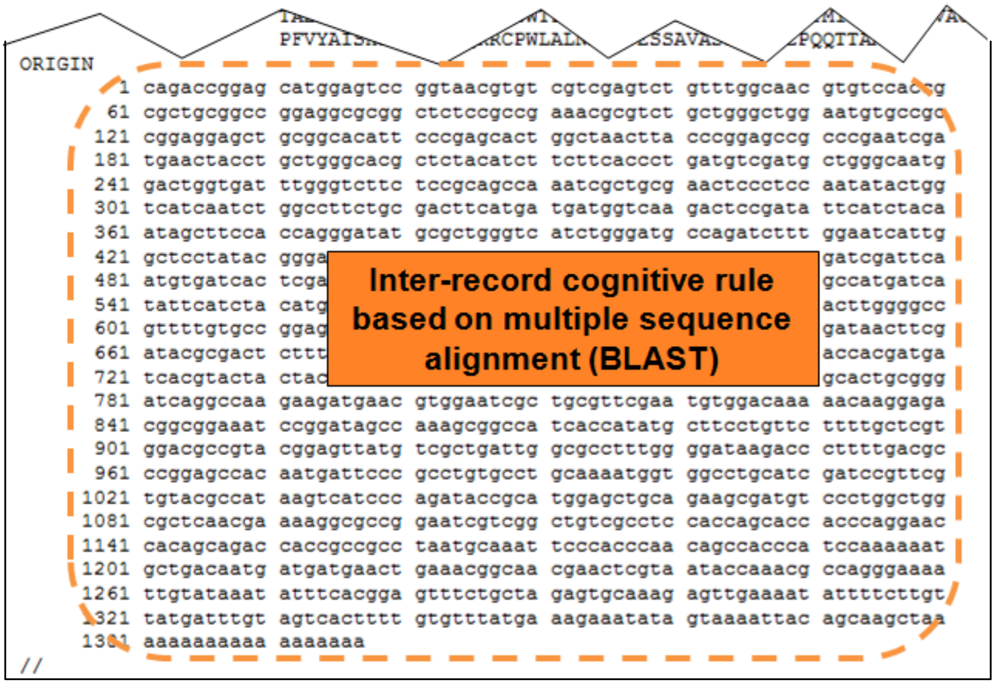


Record 2

```
LOCUS       AF184224                1397 bp    mRNA    linear   INV 19-SEP-1999
DEFINITION  Drosophila melanogaster clone GH02505 Rh3 (Rh3) mRNA, complete cds.
ACCESSION   AF184224
VERSION     AF184224.1  GI:5911285
KEYWORDS    FLI_CDNA.
SOURCE      Drosophila melanogaster (fruit fly)
  ORGANISM  Drosophila melanogaster
            Eukaryota; Metazoa; Arthropoda; Hexapoda; Insecta; Pterygota;
            Neoptera; Endopterygota; Diptera; Brachycera; Muscomorpha;
            Ephydroidea; Drosophilidae; Drosophila; Sophophora.
REFERENCE   1  (bases 1 to 1397)
  AUTHORS   Rubin,G.M., Wan,K.H., Harvey,D., Lewis,S.E., Brokstein,P.,
            Tsang,G., Agbayani,A., Arcaina,T.T., Baxter,E., Blazej,R.G.,
            Butenhoff,C., Champe,M., Chavez,C., Chew,M., Doyle,C.M.,
            Farfan,D.E., Frise,E.,                                   ,
            Hoskins,R.A., Evans-Ho                                  Kim,E.,
            Li,P., Moshrefi,M., Pa                                  ethi,H.,
            Snir,E., Svirskas,R.R.
  TITLE     Full Length Drosophila melanogaster cDNA sequence
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 1397)
  AUTHORS   Rubin,G.M., Wan,K.H., Harvey,D., Lewis,S.E., Brokstein,P.,
            Tsang,G., Agbayani,A., Arcaina,T.T., Baxter,E., Blazej,R.G.,
            Butenhoff,C., Champe,M., Chavez,C., Chew,M., Doyle,C.M.,
            Farfan,D.E., Frise,E., Galle,R., George,R.A., Harris,N.L.,
            Hoskins,R.A., Evans-Holm,M., Houston,K.A., Hummasti,S.R., Kim,E.,
            Li,P., Moshrefi,M., Pacleb,J.M., Park,S., Sequeira,A., Sethi,H.,
            Snir,E., Svirskas,R.R., Weinburg,T. and Celniker,S.E.
  TITLE     Direct Submission
  JOURNAL   Submitted (08-SEP-1999) Berkeley Drosophila Genome Project,
            University of California Berkeley, 539 Life Sciences Addition
            #3200, Berkeley, CA 94720, USA
COMMENT     Sequence submitted by:
            Berkeley Drosophila Genome Project
            University of California Berkeley
            Berkeley, CA 94720
            For
```

Date for computing freshness

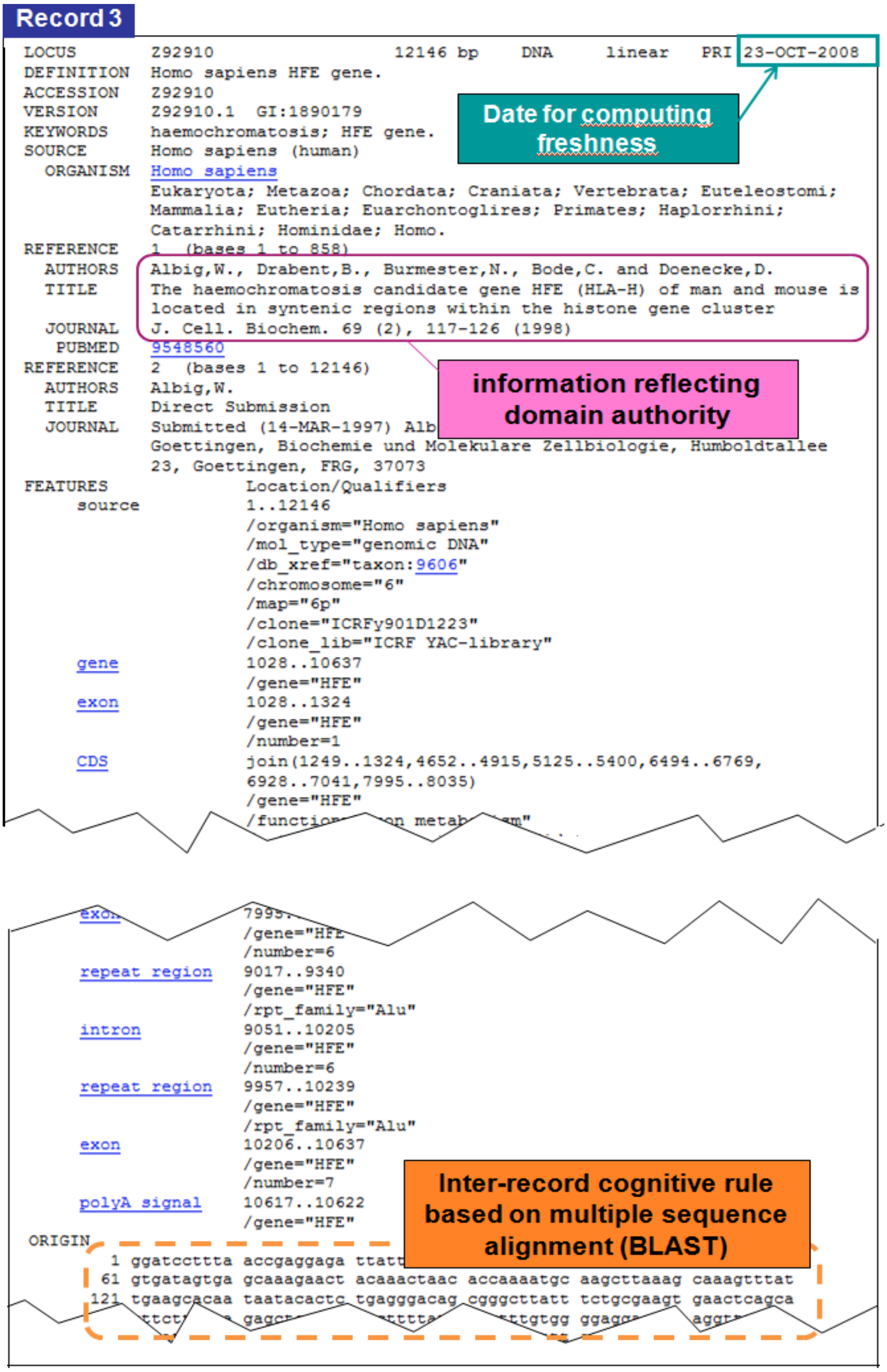
information reflecting domain authority



```
                                    TA                      WT                               MI          VA
                                    PFVYAIS                RCPWLALN          SSAVAS         PQQTTA
ORIGIN
        1 cagaccggag catggagtcc ggtaacgtgt cgtcgagtct gtttggcaac gtgtccaccg
       61 cgctgcggcc ggaggcgcgg ctctccgccg aaacgcgtct gctgggctgg aatgtgccgc
      121 cggaggagct gcggcacatt cccgagcact ggctaactta cccggagccg cccgaatcga
      181 tgaactacct gctgggcacg ctctacatct tcttcaccct gatgtcgatg ctgggcaatg
      241 gactggtgat ttgggtcttc tccgcagcca aatcgctgcg aactccctcc aatatactgg
      301 tcatcaatct ggccttctgc gacttcatga tgatggtcaa gactccgata ttcatctaca
      361 atagcttcca ccagggatat gcgctgggtc atctgggatg ccagatcttt ggaatcattg
      421 gctcctatac ggga                                         gatcgattca
      481 atgtgatcac tcga                                         gccatgatca
      541 tattcatcta catg                                         acttggggcc
      601 gttttgtgcc ggag                                         gataacttcg
      661 atacgcgact cttt                                         accacgatga
      721 tcacgtacta ctac                                         gcactgcggg
      781 atcaggccaa gaagatgaac gtggaatcgc tgcgttcgaa tgtggacaaa aacaaggaga
      841 cggcggaaat ccggatagcc aaagcggcca tcaccatatg cttcctgttc ttttgctcgt
      901 ggacgccgta cggagttatg tcgctgattg gcgcctttgg ggataagacc cttttgacgc
      961 ccggagccac aatgattccc gcctgtgcct gcaaaatggt ggcctgcatc gatccgttcg
     1021 tgtacgccat aagtcatccc agataccgca tggagctgca gaagcgatgt ccctggctgg
     1081 cgctcaacga aaaggcgccg gaatcgtcgg ctgtcgcctc caccagcacc acccaggaac
     1141 cacagcagac caccgccgcc taatgcaaat tcccacccaa cagccaccca tccaaaaaat
     1201 gctgacaatg atgatgaact gaaacggcaa cgaactcgta ataccaaacg ccagggaaaa
     1261 ttgtataaat atttcacgga gtttctgcta gagtgcaaag agttgaaaat attttcttgt
     1321 tatgatttgt agtcactttt gtgtttatga aagaaatata gtaaaattac agcaagctaa
     1381 aaaaaaaaaa aaaaaaa
//
```

Inter-record cognitive rule based on multiple sequence alignment (BLAST)

**Figure 6.** *GenBank* Screen Shos for HFE Gene: Record AF184224



Record 3

```
LOCUS       Z92910                  12146 bp    DNA     linear   PRI 23-OCT-2008
DEFINITION  Homo sapiens HFE gene.
ACCESSION   Z92910
VERSION     Z92910.1  GI:1890179
KEYWORDS    haemochromatosis; HFE gene.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 858)
  AUTHORS   Albig,W., Drabent,B., Burmester,N., Bode,C. and Doenecke,D.
  TITLE     The haemochromatosis candidate gene HFE (HLA-H) of man and mouse is
            located in syntenic regions within the histone gene cluster
  JOURNAL   J. Cell. Biochem. 69 (2), 117-126 (1998)
   PUBMED   9548560
REFERENCE   2  (bases 1 to 12146)
  AUTHORS   Albig,W.
  TITLE     Direct Submission
  JOURNAL   Submitted (14-MAR-1997) Alb
            Goettingen, Biochemie und Molekulare Zellbiologie, Humboldtallee
            23, Goettingen, FRG, 37073
FEATURES             Location/Qualifiers
     source          1..12146
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:9606"
                     /chromosome="6"
                     /map="6p"
                     /clone="ICRFy901D1223"
                     /clone_lib="ICRF YAC-library"
     gene            1028..10637
                     /gene="HFE"
     exon            1028..1324
                     /gene="HFE"
                     /number=1
     CDS             join(1249..1324,4652..4915,5125..5400,6494..6769,
                     6928..7041,7995..8035)
                     /gene="HFE"
                     /functio   on metab   sm"

     exon            7995.
                     /gene="HFE
                     /number=6
     repeat_region   9017..9340
                     /gene="HFE"
                     /rpt_family="Alu"
     intron          9051..10205
                     /gene="HFE"
                     /number=6
     repeat_region   9957..10239
                     /gene="HFE"
                     /rpt_family="Alu"
     exon            10206..10637
                     /gene="HFE"
                     /number=7
     polyA_signal    10617..10622
                     /gene="HFE"
ORIGIN
        1 ggatccttta accgaggaga ttatt
       61 gtgatagtga gcaaagaact acaaactaac accaaaatgc aagcttaaag caaagtttat
      121 tgaagcacaa taatacactc tgagggacag cgggcttatt tctgcgaagt gaactcagca
```

Date for computing freshness

information reflecting domain authority

Inter-record cognitive rule based on multiple sequence alignment (BLAST)

**Figure 7.** *GenBank* Screen Shot for HFE Gene: Record Z92910

### 4.2.3. Database profiling and data quality metrics

Several information quality dimensions with their related metrics can be then defined, computed, and associated as metadata to the data extracted from biological databanks. These metadata can be very useful for data integration, knowledge pre- and post-filtering. We have categorized them into three categories (cf. Table 3):

- Bio-knowledge-based quality metadata such as originality, domain authority of the authors who submitted the sequence,
- Schema-based quality metadata such as local and global completeness, level of details, intra- and inter-record redundancy,
- Contextual quality metadata such as freshness, and consolidation degree.

| ***Category*** | ***Quality Criterion*** | ***Target*** | ***Definition*** |
|---|---|---|---|
| Bio-Knowledge-based Quality Criteria | Originality | Data items and sub-items per record | Considering a set of records related to the same bio-entity (i.e., entity identification resolved), the originality of a data (sub-) item in a record set is defined by its occurrence frequency and its variability based on the normalized standard deviation of the edit distance between the considered strings. |
| | Domain Authority | Record | Domain authority is a grade in [0,1] that is computed depending on the status of the reference (*Published*, *Submitted, Unpublished*), the number of referenced submissions of the authors in the record and of the user-grade defined on the journal and authors reputations of the most recent reference of these authors. |
| Schema-based Quality Criteria | Local Completeness | Record | Local completeness is defined by the fraction of the number of items and sub-items with non null values on the total number of items and sub-items in the local data source schema (DTD). |
| | Global Completeness | Record | Global completeness is defined by the fraction of the number of items and sub-items with non null values provided by a source on the total number of items and sub-items in the global schema of the data warehouse. |
| | Level of Detail | Data items and sub-items per record | Level of detail is the number of sub-items per item described with non null values by a local source normalized by the total of possible sub-items in the data source schema. |
| | Intra-Record Redundancy | Record | Intra-record redundancy is defined by the fraction of items and sub-items in the record that are approximately the same based on the edit or q-grams distance functions or other semantic and cognitive rules |
| | Inter-Record Redundancy | Record Set of the same bio-entity | Inter-record redundancy is defined by the fraction of items and sub-items in the record set that are approximately the same based on edit or q-grams distance functions, BLAST or other sequence alignment techniques or other cognitive rules. |
| Contextual Quality Criteria | Freshness | Record | Freshness is defined by the difference between the current date and the publication date of the record |
| | Consolidation Degree | Data items and sub-items per record | Consolidation degree is defined by the number of inter-record redundancies and overlaps. |

***Table 3.*** *Computing Data Quality Metadata for Documenting Biomedical Sources Before Integration*

## *4.3. Ontology-based Approaches*

Semantic Web anticipates the use of ontologies to facilitate data sharing over the web, and ontologies are proposed as a solution to conciliate and attain as much as possible heterogeneity between data sources. As a result, the use of ontologies for semantic driven data integration to build multiple data warehouses, that combine and analyse different sorts of data was promising.

Two major events have urged the development of ontologies in Life Sciences: i/ a strong emergence of large volume of data represented heterogeneously in multiple data sources and ii/ increasing motivation to world-wide share these data on the web.

Following the publication of the genome sequences and their various annotations, the use of bio-ontologies became essential to deal with the heterogeneity of data and sources. Bio-ontologies helped to unify different definitions, improve data quality and promote data sharing and exchange.

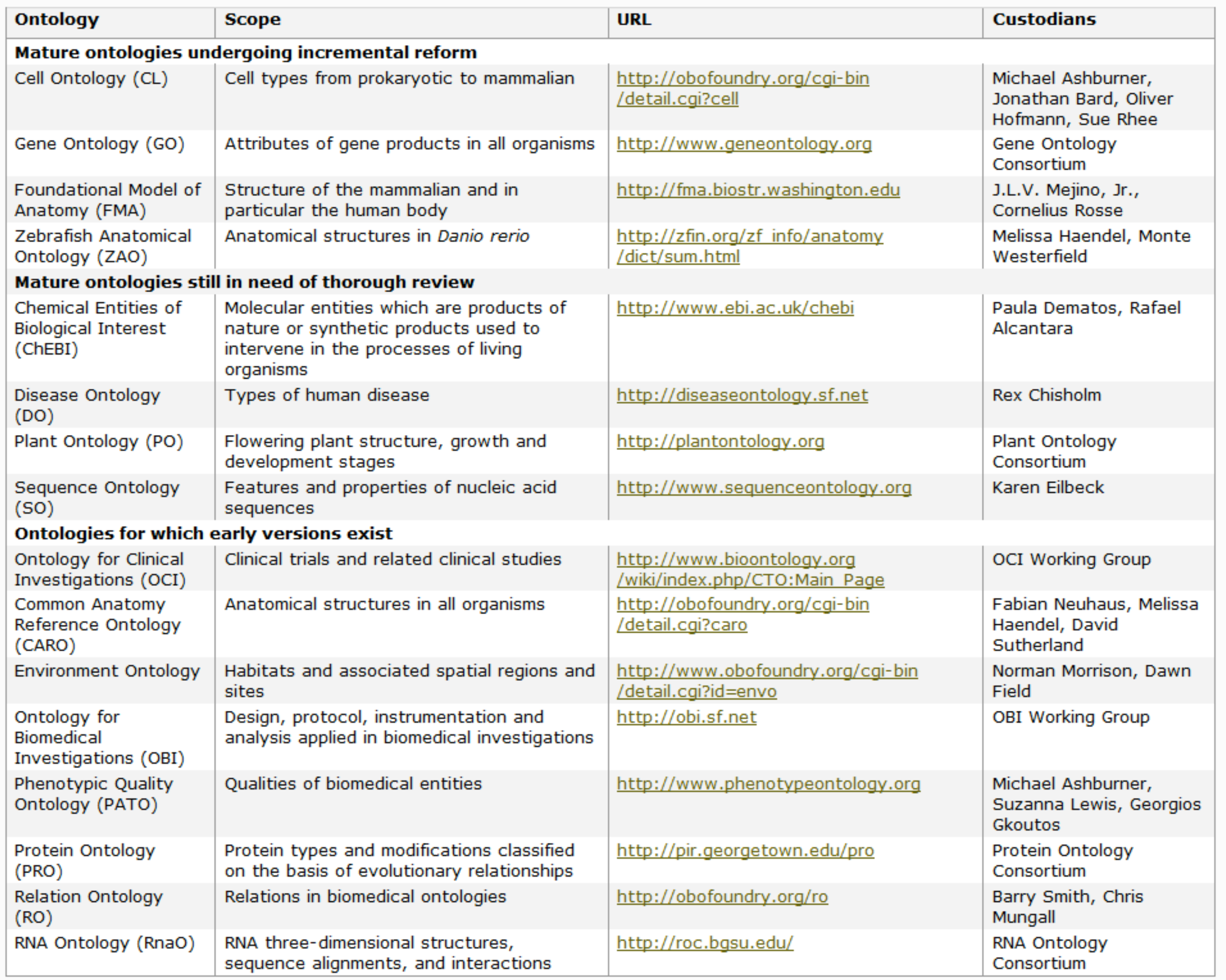
Paradoxically, it is the medical informatics community that has first developed strategies to facilitate and improve access to biomedical knowledge using ontologies. Thus, the NLM (National Library of Medicine) has developed the Unified Medical Language System (UMLS), a rich knowledge base qualified as a medical ontology of more than one million of concepts and developed by the unification of 60 biomedical terminologies (Bodenreider 2004).

Thus, previous achievements on ontologies in the medical domain had a direct impact in the bioinformatics community. The understanding of functional genomic data being also one of the challenges of modern medicine, the two communities have joint their efforts in the development of bio-ontologies.

While Gene Ontology has rapidly turned-out to be the leading Ontology in functional genomics, other ontologies have emerged as a response to a constant need to formalize the various fields of Life and Health Sciences. Consequently, the Open Biological and biomedical *Ontologies foundry*[23] (OBO) archives a collection of bio-ontologies in a standard format. A strong community involvement was crucial to avoid as much as possible redundancy and ensure that only single ontologies for each area are placed in the public domain.

As shown in Table 4, the OBO Foundry supports various domain knowledge of Life and Health Sciences, and includes ontologies like: Gene Ontology, Pathway Ontology, Disease Ontology, Systems Biology Ontology, and Chemical Entities of Biological Interest (CHEBI) Ontology (Smith et. al., 2007).

Shared ontologies are used to conciliate and to attain as much as possible data conflicts. Various standards in Life Sciences have been developed to provide domain knowledge to be used for semantically driven integration of information from different sources.



| Ontology | Scope | URL | Custodians |
|---|---|---|---|
| **Mature ontologies undergoing incremental reform** | | | |
| Cell Ontology (CL) | Cell types from prokaryotic to mammalian | http://obofoundry.org/cgi-bin/detail.cgi?cell | Michael Ashburner, Jonathan Bard, Oliver Hofmann, Sue Rhee |
| Gene Ontology (GO) | Attributes of gene products in all organisms | http://www.geneontology.org | Gene Ontology Consortium |
| Foundational Model of Anatomy (FMA) | Structure of the mammalian and in particular the human body | http://fma.biostr.washington.edu | J.L.V. Mejino, Jr., Cornelius Rosse |
| Zebrafish Anatomical Ontology (ZAO) | Anatomical structures in *Danio rerio* | http://zfin.org/zf_info/anatomy/dict/sum.html | Melissa Haendel, Monte Westerfield |
| **Mature ontologies still in need of thorough review** | | | |
| Chemical Entities of Biological Interest (ChEBI) | Molecular entities which are products of nature or synthetic products used to intervene in the processes of living organisms | http://www.ebi.ac.uk/chebi | Paula Dematos, Rafael Alcantara |
| Disease Ontology (DO) | Types of human disease | http://diseaseontology.sf.net | Rex Chisholm |
| Plant Ontology (PO) | Flowering plant structure, growth and development stages | http://plantontology.org | Plant Ontology Consortium |
| Sequence Ontology (SO) | Features and properties of nucleic acid sequences | http://www.sequenceontology.org | Karen Eilbeck |
| **Ontologies for which early versions exist** | | | |
| Ontology for Clinical Investigations (OCI) | Clinical trials and related clinical studies | http://www.bioontology.org/wiki/index.php/CTO:Main_Page | OCI Working Group |
| Common Anatomy Reference Ontology (CARO) | Anatomical structures in all organisms | http://obofoundry.org/cgi-bin/detail.cgi?caro | Fabian Neuhaus, Melissa Haendel, David Sutherland |
| Environment Ontology | Habitats and associated spatial regions and sites | http://www.obofoundry.org/cgi-bin/detail.cgi?id=envo | Norman Morrison, Dawn Field |
| Ontology for Biomedical Investigations (OBI) | Design, protocol, instrumentation and analysis applied in biomedical investigations | http://obi.sf.net | OBI Working Group |
| Phenotypic Quality Ontology (PATO) | Qualities of biomedical entities | http://www.phenotypeontology.org | Michael Ashburner, Suzanna Lewis, Georgios Gkoutos |
| Protein Ontology (PRO) | Protein types and modifications classified on the basis of evolutionary relationships | http://pir.georgetown.edu/pro | Protein Ontology Consortium |
| Relation Ontology (RO) | Relations in biomedical ontologies | http://obofoundry.org/ro | Barry Smith, Chris Mungall |
| RNA Ontology (RnaO) | RNA three-dimensional structures, sequence alignments, and interactions | http://roc.bgsu.edu/ | RNA Ontology Consortium |

***Table 4.*** *The OBO Foundry Ontologies (April 2007)*

Unfortunately, the one way that was massively used to integrate life science data using ontologies, is through the annotation of the multiple sorts of data in genomics (gene sequences and proteins) using the common vocabulary carried by these ontologies. But the great success of this approach has led to proliferation of bio-ontologies that again has created obstacles to data integration. In some sorts, the OBO foundry consortium has emerged to overcome this problem (Smith et. al., 2007).

More ideally, the aim of such ontologies in the context of data integration would be of granting a model of biological concepts that can be used to form a semantic framework for querying the heterogeneous Life Sciences sources or for systematizing annotation of experimental results. As an experience, the TaO ontology (TAMBIS ontology), that describes a wide range of Life Sciences concepts and their relationships, provided such framework. Rather than materializing bio-data in integrated data warehouses, the TAMBIS project aimed to providing a single and transparent access point for Life Sciences information through the use of a mediating ontology (Baker et al., 1998). Queries are written in terms of TaO ontology concepts and converted to queries to appropriate sources.

More recently, there exists an extraordinary number of bioinformatics applications (Erson et al., 2010) that are based on ontology as a background domain knowledge and a unified model against Life Sciences resources to remediate data annotation, data integration, and data heterogeneity. However, ontology development and maintenance is time-consuming and requires constant investment from expert curators. Open collaborative platforms enable the wider scientific community to become involved in developing and maintaining them, but raises concerns regarding the quality and correctness of the information added (Hoehndorf et al., 2009).

# 5. Conclusions and perspectives

Many data sources in the biomedical domain are renowned for containing data of sometimes poor quality. This is due to the experimental nature of the field, the quickly changing knowledge landscape, the high redundancies in experiments performed often leading to contradicting results, and the difficulties in properly describing the results of an experiment in a domain as complex as molecular biology. Furthermore, it was often observed that data quality problems multiply when data of low quality are integrated and re-used for annotation.

Based on our past experience of building the biomedical data warehouse *GEDAW* (*Gene Expression Data Warehouse*) that stores all the relevant information on genes expressed in the liver during iron overload and liver pathologies (*i.e.*, records extracted from public databanks, data generated from DNA chips home experiments, data collected in hospitals and clinical institutions as medical records), we presented some lessons learned, data quality issues in this context and current solutions we proposed for quality-aware integrating and warehousing our biomedical data. In this chapter, we gave an overview of data quality problems and solutions relevant to any preprocessing approach and also elements for data quality-awareness for the complex processes of integrating and warehousing biomedical data.

With regards to the limits of any data warehousing approach, it is relevant to generate quality metadata at the preprocessing and pre-integration stage, as long as the whole data integration process (from the original data sources into the destination data warehousing system) stays feasible automatically and with a reasonable performance. The final data filtering task has generally to be performed by the expert on the delivered annotations or data analysis before their storage in the warehouse by using multiple data quality criteria, like the authoritativeness of the information source or the credibility of the authors of the submitted record, for instance.

Quality in the results of data mining and knowledge discovery from biomedical resources critically depends on the preparation and on the quality of analyzed datasets. Indeed biomedical data mining processes and applications require various forms of data preparation, correction and consolidation combining complex data transformation operations and cleaning techniques, because the data input to the mining algorithms is assumed to conform to “nice” data distributions, containing no missing, inconsistent or incorrect values. This leaves a large gap between the available “dirty” data and the available machinery to process and analyze the data for discovering added-value knowledge and decision making in Life Sciences.

The aspects of measuring data quality and detecting hot-spots of poor quality constitute very challenging research directions for the Bioinformatics community. These include analyzing contradicting values in the case of duplicate entries and detecting hard-to-catch errors. Such an erroneous data is one whose value looks perfectly legitimate. Yet, if we examine this value in conjunction with other attribute values, the value appears questionable. Detecting such dubious values is a major problem in data cleaning but it becomes much harder in complex domains such as Life Sciences.

# 6. References

Ailamaki, A., Kantere, V. and Dash, D. Managing Scientific Data. Communications of the ACM, **53**(6):68-78, 2010.

Baker, P., Brass, A., Bechhofer, S., Goble, C., Paton, N. & Stevens, R. (1998). TAMBIS: Transparent Access to Multiple Bioinformatics Information Sources. An Overview. Proc. of the Sixth International Conference on Intelligent Systems for Molecular Biology.

Baumgartner Jr, W.A., Cohen, K.B., Fox, L.M., Acquaah-Mensah, G., and Hunter, L., Manual curation is not sufficient for annotation of genomic databases. Bioinformatics, 2007. **23**(13): p. i41.

Berti-Équille, L., Moussouni, F. (2005). Quality Aware Integration and warehousing of Genomic Data. Proc. of the International Conference on Information Quality, Boston, US.

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 32 Database issue:D267-70.

Brochhausen, M., Spear, A.D., Cocos, C., Weiler, G., Martín, L., Anguita, A., Stenzhorn, H., Daskalaki, E., Schera, F., Schwarz, U., Sfakianakis, S., Kiefer, S., Dörr, M., Graf, N., Tsiknakis, M. (2010). The ACGT Master Ontology and its Applications -Towards an Ontology-Driven Cancer Research and Management System. J Biomed Inform. 2011 Feb., **44**(1):8-25. Epub 2010 May 11.

Buneman, P., Crabtree, J., Davidson, S., Tannen, V. & Wong, L. (1998). BioKleisli. BioInformatics.

Buneman, P., J. Cheney, W.-C. Tan, and S. Vansummeren. Curated Databases. Proc. of the International Conference PODS 2008. Vancouver, Canada.

Chen, I.A., Kosky, A.S., Markowitz, V.M. & Szeto, E. (1997). Constructing and Maintaining Scientific Database Views. Proc. of the 9th Conference on Scientific and Statistical Database Management.

Chen, I.A. & Markowitz, V.M. (1995). An Overview of the Object-Protocol Model (OPM) and OPM Data Management Tools. Information Systems, 20(5), pp 393—418.

Cohen-Boulakia, S., Tan, W.C., Provenance in Scientific Databases, In Encyclopedia of Database Systems, 2009.

Cohen-Boulakia, S., Leser, U., Next Generation Data Integration for the Life Sciences. Tutorial presented at the International Conference of Data Engineering, ICDE 2011.

Davidson, S., Crabtree, J., Brunk, B., Schug, J., Tannen, V., Overton, C. & Stoeckert, C. (2001). K2/Kleisli and GUS: Experiments in integrated access to genomic data sources. IBM Systems Journal, 40(2), 512-531.

Erson, E., Cavusoglu, M. (2010). Design of a framework for modeling, integration and simulation of physiological models. Proc. of IEEE Eng. Med. Biol. Soc., pp. 1485-9.

Gertz M., Managing Data Quality and Integrity in Federated Databases, Proc. of the IFIP, Second Working Conference on Integrity and Internal Control in Information Systems, pages 211–230, Kluwer, B.V., 1998.

Guérin, E., Marquet, G., Burgun, A., Loréal, O., Berti-Équille, L., Leser, U., Moussouni, F. (2005). Integrating and Warehousing Liver Gene Expression Data and Related Biomedical Resources in GEDAW. Proc. of the International Workshop on Data Integration in Life Sciences, Lecture Notes in Bioinformatics, 2005, 3615, pp. 158-174.

Guérin, E., Marquet, G., Chabalier, J., Troadec, M.B., Guguen-Guillouzo, C., Loréal, O., Burgun, A., Moussouni, F. (2006) Combining biomedical knowledge and transcriptomic data to extract new knowledge on genes, online Journal of Integrative Bioinformatics, 3(2).

Haas, L., Kodali, P., Rice, J., Schwarz, P. & Swope, W. (2000) Integrating Life Sciences Data - With a Little Garlic. In IEEE International Symposium on Bio-Informatics and Biomedical Engineering (BIBE), Published by IEEE Press, Washington, DC.

Hoehndorf, R., Bacher, J., Backhaus, M., Gregorio, SE Jr, Loebe, F., Prüfer, K., Uciteli, A., Visagie, J., Herre, H., Kelso, J., (2009). BOWiki: an ontology-based wiki for annotation of data and integration of knowledge in biology. BMC Bioinformatics. 2009 May 6,10 Suppl 5:S5.

Ives, Z.G. (2009). Data Integration and Exchange for Scientific Collaboration. Proc. of the International Workshop on Data Integration in Life Sciences, DILS 2009. Manchester, UK.

Jenkinson, A.M., et al. (2008). Integrating Biological Data – The Distributed Annotation System. Proc. of the International Workshop on Data Integration in Life Sciences, DILS 2008. Evry, France.

Kemp, G., Robertson, C. & Gray, P. (1999). Efficient access to biological databases using CORBA. CCP11 Newsletter, 3.1(7).

Kemp, G., Angelopoulos, N. & Gray, P. (2000) A Schema-based Approach to Building a Bioinformatics Database Federation, IEEE International Symposium on Bio-Informatics and Biomedical Engineering (BIBE), Washington, DC.

Lacroix, Z. (2002) Biological data integration: wrapping data and tools. to be published in IEEE Transactions on Information Technology in Biomedecine.

Lacroix Z., Cartik R., Mork P., Rifaieh R., Wilkinson M., Freire J., Cohen-Boulakia S., (2009). Biological Resource Discovery, Encyclopedia of Database Systems Springer US (Ed.) 220-223.

Laine, F., Ropert, M., Lan, C.L., Loreal, O., Bellissant, E., Jard, C., Pouchard, M., Le Treut, A. and Brissot, P. (2002) Serum ceruloplasmin and ferroxidase activity are decreased in HFE C282Y homozygote male iron-overloaded patients. J. Hepatol, 36(1), 60-65.

Lee, D., Von, B.-W., Kang, J., Park, S., (2005). Effective and Scalable Solutions for Mixed and Split Citation Problems in Digital Libraries, Proc. of 2nd Intl. ACM Workshop on Information Quality in Information Systems (IQIS'05), Baltimore, USA.

Loréal O, Turlin B, Pigeon C, Moisan A, Ropert M, Morice P, Gandon Y, Jouanolle AM, Vérin M, Hider RC, Yoshida K, Brissot P., (2002) Aceruloplasminemia: new clinical, pathophysiological and therapeutic insights. J. Hepatol. 2002 Jun, 36(6):851-6.

Marquet, G., Burgun, A., Moussouni, F., Guérin, E., Le Duff, F., Loréal, O., BioMeKe: An Ontology-Based Biomedical Knowledge Extraction System Devoted to Transcriptome Analysis. Journal of Studies in Health Technology and Informatics (2003), 95:80-5.

Mercadé, J., Espinosa, A., Adsuara, J.E., Adrados, R., Segura, J., Maes, T., (2009). Orymold: ontology based gene expression data integration and analysis tool applied to rice. BMC Bioinformatics. 2009 May 23, 10:158.

Missier P., Paton N. and Li P., Workflows for Information Integration in the Life Sciences. Lecture Notes in Computer Science, Volume 6585/2011, 215-225, 2011.

Müller, H., Naumann, F. and Freytag, J.-C. (2003). Data Quality in Genome Databases. Proc. of the International Conference on Information Quality, Boston, US.

Paton N.W., Khan, S., A. Hayes, Moussouni, F., Brass, A., Eilbeck, K., Goble, C.A., Hubbard, S., Oliver, S.G., (2000). Conceptual Modelling on Genomic Information. Bioinformatics Journal, Vol 16, No 6, 548-558.

Smith B., et al., (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology 25, pp. 1251-1255.

Wong, L. (2000). Kleisli, its Exchange Format, Supporting Tools, and an Application Protein Interaction Extraction. Proc. of the IEEE International Symposium on Bio-Informatics and Biomedical Engineering (BIBE), Published by IEEE Press, Washington D.C.

# 7. Netography

1. SWISS-PROT: http://www.expasy.org/sprot

2. GenBank: http://www.ncbi.nlm.nih.gov/genbank/

3. EMBL: http://www.ebi.ac.uk/embl/

4. DDBJ: http://www.ddbj.nig.ac.jp/

5. PDB: http://www.pdb.org/pdb/home/home.do

6. OMIM : http://www.ncbi.nlm.nih.gov/omim

7. KEGG: http://www.genome.jp/kegg/pathway.html

8. ArrayExpress: http://www.ebi.ac.uk/arrayexpress/

9. SRS: http://srs.ebi.ac.uk/

10. DBGet: http://www.genome.jp/dbget/

11. Entrez: http://www.ncbi.nlm.nih.gov/Entrez/

12. Atlas: http://www.genatlas.org/

13. Unified Medical Language System® (UMLS): http://www.nlm.nih.gov/research/umls/

14. MeSH: http://www.nlm.nih.gov/research/mesh

15. Gene Ontology™ (GO): http://www.ontologos.org/IFF/Ontologies/Gene.html

16. LocusLink: http://www.ncbi.nlm.nih.gov/LocusLink

17. HGNC (Human Gene Nomenclature Database): http://www.genenames.org/

18. MedLine, PubMed: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed

19. GOA (Gene Ontology Database): http://www.geneontology.org/GO.database.shtml

20. NCBI Record AF204869: http://www.ncbi.nlm.nih.gov/nuccore/af204869

21. NCBI Record AF184224: http://www.ncbi.nlm.nih.gov/nuccore/af184224

22. NCBI Record Z95910: http://www.ncbi.nlm.nih.gov/nuccore/z92910

23. OBO Foundry Paper: http://www.obofoundry.org/