Quality controls in integrative approaches to detect errors and inconsistencies in biological databases

Giorgio Ghisalberti[1], Marco Masseroli[1,2], Luca Tettamanti[1]

[1] Dipartimento di Elettronica e Informazione, Politecnico di Milano,
Piazza Leonardo da Vinci 32, 20133 Milano, Italy

Summary

Numerous biomolecular data are available, but they are scattered in many databases and only some of them are curated by experts. Most available data are computationally derived and include errors and inconsistencies. Effective use of available data in order to derive new knowledge hence requires data integration and quality improvement. Many approaches for data integration have been proposed. Data warehousing seams to be the most adequate when comprehensive analysis of integrated data is required. This makes it the most suitable also to implement comprehensive quality controls on integrated data. We previously developed GFINDer (http://www.bioinformatics.polimi.it/GFINDer/), a web system that supports scientists in effectively using available information. It allows comprehensive statistical analysis and mining of functional and phenotypic annotations of gene lists, such as those identified by high-throughput biomolecular experiments. GFINDer backend is composed of a multi-organism genomic and proteomic data warehouse (GPDW). Within the GPDW, several controlled terminologies and ontologies, which describe gene and gene product related biomolecular processes, functions and phenotypes, are imported and integrated, together with their associations with genes and proteins of several organisms. In order to ease maintaining updated the GPDW and to ensure the best possible quality of data integrated in subsequent updating of the data warehouse, we developed several automatic procedures. Within them, we implemented numerous data quality control techniques to test the integrated data for a variety of possible errors and inconsistencies. Among other features, the implemented controls check data structure and completeness, ontological data consistency, ID format and evolution, unexpected data quantification values, and consistency of data from single and multiple sources. We use the implemented controls to analyze the quality of data available from several different biological databases and integrated in the GFINDer data warehouse. By doing so, we identified in these data a variety of different types of errors and inconsistencies; this enables us to ensure good quality of the data in the GFINDer data warehouse. We reported all identified data errors and inconsistencies to the curators of the original databases from where the data were retrieved, who mainly corrected them in subsequent updating of the original database. This contributed to improve the quality of the data available, in the original databases, to the whole scientific community.

1 Introduction

Rapid progression of biomedical knowledge is being fostered by the explosion of available biomolecular data and the development of computational methods to extract reliable information from them. Molecular biology databases are continuously increasing in number (more than 1,150 in January 2009 [1]) and in coverage of the included biomolecular entities (e.g. genomic DNAs, genes, transcripts, proteins), as well as of their described structural and

[2] To whom correspondence should be addressed. E-mail: marco.masseroli@polimi.it

functional biomedical features. They provide extremely valuable information, but scattered across numerous data sources, only partially curated and mostly computationally derived, which is known to include errors and inconsistencies. Effective use of available information to support the interpretation of experimental results and derive new biomedical knowledge hence requires their integration and quality improvement. These, or at least the selection of higher quality data among those available, are particularly required for the numerous computationally derived data.

Several different approaches for data integration have been proposed, including data warehousing, multi databases, federated databases, information linkage and mediator based solutions. They have been implemented in systems such as Biowarehouse [2], TAMBIS [3], DiscoveryLink [4], SRS [5], or Biomediator [6], respectively. Data warehousing seams the most adequate when the data to be integrated are numerous and efficient and comprehensive analysis and mining of the integrated data is required [7]. This approach requires that information from the distributed databases to be integrated are automatically retrieved and processed in order to create and maintain updated an integrated and consistent collection of originally distributed data. This makes data warehousing one of the most suitable methods also to implement comprehensive quality controls on the integrated data. In particular, it makes possible to effectively and efficiently check data errors and inconsistencies, both within a single data source and among multiple data sources integrated in the data warehouse.

2 GFINDer Data Warehouse

To effectively take advantage of the numerous genomic and proteomic information sparsely available in many heterogeneous and distributed biomolecular databases accessible via the Internet, we previously developed the Genome Function INtegrated Discoverer (GFINDer) project (http://www.bioinformatics.polimi.it/GFINDer/) [8], [9]. GFINDer is a publicly available Web system used by several thousands worldwide scientists (we counted about 110,000 accesses from more than 6,000 distinct IP addresses in the last 5 years). GFINDer supports comprehensive statistical enrichment analysis and data mining of functional and phenotypic annotations of large-scale lists of user-classified genes, such as those identified by high-throughput biomolecular experiments. It automatically retrieves annotations of several functional and phenotypic categories from different sources, identifies the categories enriched in each class of a user-classified gene list and calculates statistical significance values for each category. Moreover, GFINDer enables the functional classification of genes according to mined functional categories and the statistical analysis of the classifications obtained, aiding better interpretation of high-throughput experiment results.

As other similar systems such as DAVID (http://david.abcc.ncifcrf.gov/) and FatiGO+ (http://www.fatigo.org/), GFINDer is based on a multi-organism genomic and proteomic data warehouse (GPDW). In the GPDW several controlled terminologies and ontologies, which describe gene and gene product related biomolecular processes, functions and phenotypes, are imported and stored together with their associations (annotations) with genes and proteins of several organisms. In the GPDW all such data from several different databases are integrated by interconnecting the imported annotations to the genes and proteins they refer to by means of their provided IDs and cross-references.

To ease maintaining updated and extending the information integrated in the GFINDer data warehouse, which are retrieved from many databases frequently updated, we designed and implemented several automatic procedures in Java programming language. We also implemented a set of data quality control approaches to check the imported data for a variety

of possible errors and inconsistencies, and ensure the best possible quality of the data integrated in subsequent updating of the data warehouse.

3 Quality Controls of Integrated Data

A range of techniques, such as source trustworthiness, error localization and correction, record linkage, and others, exist in the literature to assess and improve the quality of data [10], [11]. Batini and colleagues [12] provide a neat systematic and comparative description of existing methodologies for different data quality improvement strategies. Following a data-driven strategy, in this work we focus on the assessment and improvement of two fundamental data quality dimensions of integrated biomolecular data, namely accuracy and consistency. Other two important data quality dimensions, i.e. completeness and timeliness, which depend on the selected data sources that are being integrated and on the updating frequency of the integration process, are also considered.

In order to detect the variety of possible errors and inconsistencies that no rarely exist in subsequent updating of data available from public biomolecular databases, we implemented a set of automatic procedures for data quality checking and applied them to improve quality of data integrated in the GFINDer data warehouse. Among other features, these procedures check data structure and completeness, ontological data consistency, ID format and evolution, unexpected data quantification values, and consistency of data from single and multiple sources.

3.1 Data structure and completeness

Most of biological databases provides the whole of their data, or part of them, within text files in different formats, including flat, tabular, XML and RDF formats. This allows reimplementing locally the entire database, or part of it, and integrating its data with those from other databases by automatically parsing the data file contents and importing them in a local data warehouse, such as our GPDW. In subsequent versions of such files, data vary according to the data updated in the database. Also changes in data file structure are not infrequent and can produce erroneous data import and integration. To check consistency of both data and their data file structure, we created automatic procedures in which strict checking of data parsed from source data files is enforced and assured by the created data parsers. This enables us to automatically verify absence of inappropriate missed data and inconsistent data structures, and monitor data structure modifications (e.g. variation of expected tags in XML data files, or data columns in tabular data files) in new versions of source data files.

Furthermore, whenever syntactic or semantic information (e.g. tags in XML file format, or an explanatory header in tabular file format) are available for a source data file, we use them for semantic identification and control of imported data. When data structure modifications or incompatible new data are found, such data are not imported in the local data warehouse and warnings are shown for their supervised management. Adaptability to source data structure modifications requires supervision since not enough semantic information is generally available for such data to enable safe automatic implementations.

3.2 Ontological data consistency

Numerous data available in biological databases are ontological data describing different aspects of current biomolecular knowledge. Being part of an ontology, the relationships

among these data should form a hierarchical tree or a direct acyclic graph. Correct structure of these trees and graphs is paramount to support exact analysis and expansion of gene and protein direct annotations provided by many annotation databases.

In order to check the internal consistency of ontological data, we developed automatic procedures which verify that the ontological data describe a topologically correct graph (i.e. a graph that contains a single root and no cycles, and, for hierarchical trees, in which no multiple parents exist for each graph node). When inconsistencies are detected, they are automatically pointed out to be managed in a supervised way.

Furthermore, when data relationships are declared as symmetrical and contain enough redundancy to allow checking them, our developed procedures automatically check their internal consistency and completeness. When inconsistencies are detected, they are shown and, if the developed automatic procedures are set to do so, the missed relationship data are automatically filled in. For example, the Entrez Gene database [13] provides both *Related pseudogene* and *Related functional gene* relationships, which are defined as symmetrical, between genes. Thus, if the relationship about gene A being a related pseudogene of gene B exists, but the relationship about gene B being a related functional gene of gene A is not available, then the provided data are internally inconsistent or incomplete (either the first relationship is wrong, or the second one is missed). The same logical inspection is used to check for external consistency (i.e. among data provided by different data sources).

3.3 ID format and evolution

Several different types of IDs or accession numbers are used to identify data in biological databases. Usually, each database adopts its own IDs and provides their mapping or association with the IDs of other most recognized databases (e.g. Entrez Gene IDs or Ensembl IDs for genes, and Ensembl IDs, RefSeq IDs or UniProt IDs for proteins) [13], [14], [15], [16]. The different types of IDs have several different formats, but most of them has a well defined numerical or alphanumerical format, with a fixed or variable number of digits or characters. These formats can be described with regular expressions; some databases like RefSeq and UniProt provide regular expressions describing their ID formats (Table 1).

ID source	ID type	Regular expression
RefSeq	DNA sequence ID	AC_[0-9]{6}(\.[0-9]+)?
RefSeq	DNA sequence ID	N[CGSTW]_[0-9]{6,9}(\.[0-9]+)?
RefSeq	DNA sequence ID	NZ_[A-Z]{4}[0-9]{8}(\.[0-9]+)?
RefSeq	Transcript ID	[NX][MR]_[0-9]{6,9}(\.[0-9]+)?
RefSeq	Protein ID	[ANXYZ]P_[0-9]{6,9}(\.[0-9]+)?
UniProt	Protein ID	[A-NR-Z][0-9][A-Z][A-Z0-9][A-Z0-9][0-9] (\.[0-9]+)?
UniProt	Protein ID	[OPQ][0-9][A-Z0-9][A-Z0-9][A-Z0-9][0-9] (\.[0-9]+)?

Table 1: Example of available regular expressions describing some biological ID formats

Since many operations performed on biological database data regard ID matching and linking, correct ID format and exact identification of ID type is paramount. In order to automatically identify, syntactically check and manage the different ID types, we both used the available regular expressions and defined a set of other regular expressions (Table 2), which describe the different ID types and their acceptable formats. Furthermore, in order to discriminate

different ID types with same format, our developed automatic procedures also check the name of the source that provides the IDs whenever it is available associated with the IDs. This allows semantically identifying the different ID types and controlling their correct use.

ID source	ID type	Regular expression
Entrez Gene	Gene ID	[0-9]+
Ensembl	Gene ID	ENS[A-Z]{3}G[0-9]{11}
Ensembl	Gene ID	ENSG[0-9]{11}
IPI	Protein ID	IPI[0-9]{8}
Gene Ontology	Biological process, Molecular function, or Cellular component ID	GO:[0-9]{7}
KEGG	Pathway ID	[0-9]{5}
InterPro	Protein family or domain ID	IPR[0-9]{6}
Expasy	Enzyme ID	EC:[0-9]+(\.[0-9]+|-){0,3}

Table 2: Example of defined regular expressions describing some biological ID formats

Besides checking ID format, whenever possible our developed procedures also control ID evolution. This is an important aspect since biological databases have different updating frequencies, and database IDs (or their associations) can vary among different updating versions. Thus, at a given point of time association data between different databases (e.g. annotation data) can refer to a version of the related database data different from the one directly available from the related database. This might be a significant issue when all these data are integrated together. In order to minimize such an issue, besides carefully selecting association data providers, whenever ID history data are available, we used them to automatically reconcile database IDs from different providers.

3.4 Unexpected data quantification values

Implemented data checking automatic procedures also perform data quantifications that can identify possible data errors, inconsistencies or redundancies. For example, they check for data duplicates within data both from a single database and also from different databases.

Furthermore, implemented checking quantifications can also highlight unexpected information patterns. These not only may unveil inconsistencies among the data provided by different sources, or even the same data source, but might also foster the generation of data driven hypotheses, which might potentially lead to biological discoveries and new biomedical knowledge. For example, in the integrated data from multiple organisms and sources, quantifications of the number of genes that codify a single or more proteins, or the number of different proteins codified by one or more genes, can give insight on gene homolog and alternative splicing phenomena.

3.5 Consistency of data from multiple sources

Besides checking consistency within data imported from a single source, we also implemented automatic cross-controls among data imported from different sources. This enable us to identify redundant or mismatching data and ensure best possible quality of data integrated in the GFINDer data warehouse. When multiple independent sources provide

overlapping data, we use such overlaps to verify the information they provide and increase its likelihood. When inconsistencies are detected, they are automatically pointed out.

Cross-comparison of data from different sources is also performed by checking and taking advantage of relationship loops among imported data. This can help in both verifying consistency and completeness of different data sources, and unveiling unexpected information patterns possibly leading to biological discoveries.

4 Errors and Inconsistencies Detected in Available Data

By using the above described implemented automatic procedure to analyze the quality of data available from several different biological databases integrated in the GFINDer data warehouse we identified a variety of different types of errors and inconsistencies, some of which are following reported.

4.1 Data structure and completeness

By checking completeness and structure of data files provided by several different biological databases, we identified both inappropriate missed data and inconsistent data structures. For example, we check the *gi2ipi.xrefs* file from the International Protein Index (IPI) database [17] (ftp://ftp.ebi.ac.uk/pub/databases/IPI/), which provides mapping between protein IPI IDs and the Entrez Gene ID of their codifying gene. In the version of this file downloaded from IPI on July 30th, 2009, we found 510 IPI IDs (on a total of 768,148, about 0.07%) lacking of the Entrez Gene ID of their codifying gene. This may indicate some erroneously missing data, or may also indicate that the encoding genes (and/or their Entrez Gene ID) of the proteins identified by those IPI IDs are not known yet.

Among other data file formats, our developed automatic quality control procedures can also test conformity of data file structure in Open Biomedical Ontology (OBO) format [18], which is used by several data sources to provide ontology data. In the OBO format, data are structured in rooms: a room is a labelled part of the file in which an object of a particular type is described together with its attributes. The label of each room and of each object attribute in the room must have a defined structure and value, in order to correctly recognize the file content. Yet, sometimes these labels present a wrong unexpected value. For example, in the OBO *pathology.obo* file version 2.9 downloaded on July 30th, 2009 from the eVOC database [19] (http://www.evocontology.org/) we found an attribute label with a value (*exact_synonymexact_synonym*) that is not described in the eVOC ontology documentation (Figure 1).

```
[Term]
id: EV:0400005
name: Down's syndrome
is_a: EV:0400002 ! genetic disorders
exact_synonymexact_synonym: "Down syndrome" []
```

Figure 1: Excerpt of the July 30th, 2009 version of the eVOC *pathology.obo* file including the incorrect attribute label *exact_synonymexact_synonym* shown in the box

Furthermore, our implemented data quality procedures automatically identified data file structure variations in subsequent versions of data files provided by some databases. These included an increasing number of data columns in some tabular data file provided by the

Entrez Gene database, or XML tag modifications in the XML file provided by the UniProt database.

4.2 Ontological data consistency

Our consistency checking of ontological data provided by the Gene Ontology [20], eVOC and KEGG [21] databases generally did not identify ontological inconsistencies. Only in the OBO *anatomicalsystem.obo* file version 2.9 downloaded on July 30th, 2009 from the eVOC database we found that the eVOC anatomical system ontology term ID *EV:0100106* was hierarchically associated with two different parent terms of the ontology, i.e. term ID *EV:0100101* and term ID *EV:0100105* (Figure 2). This is not consistent with the hierarchical tree structure of the eVOC anatomical system ontology, which requires that each ontology term has no more than a single parent.

```
[Term]
id: EV:0100106
cdnalib: SEMVNOT05 11716
cdnalib: SEMVTDA01 11717
xref_analog: SWP:TS-0919
name: seminal vesicle
def: "Either of a pair of pouchlike structures posterinferior to
the urinary bladder of males. They secrete an alkaline, viscous
fluid that constitutes a significant proportion of the fluid
that ultimately becomes semen." [ISBN:0-471-36692-7]
is_a: EV:0100101 ! male reproductive system
is_a: EV:0100105 ! vas deferens
```

Figure 2: Excerpt of the July 30th, 2009 version of the eVOC *anatomicalsystem.obo* file including the ontological term ID *EV:0100105* incorrectly hierarchically associated with two parent terms (shown in the box)

Also our checking of the symmetrical ontological data provided by some databases, such as the *Related pseudogene* and *Related functional gene* relationships in the Entrez Gene *gene_group* file (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/), did not find any internal symmetrical inconsistency or incompleteness.

4.3 ID format and evolution

By using specific regular expressions to automatically check the format of all IDs imported in the GFINDer data warehouse, we can both identify syntactical errors in ID formats and recognize ID type and provenance in order to control their correct semantic use. The performed controls identified several inconsistent ID semantic assignments and ID format errors. For example, in the version of the *gene2accession* tabular file downloaded Entrez Gene on July 30th, 2009 we found a great number of semantic inconsistencies, including:

- A set of 312 RefSeq IDs of genomic nucleotide sequences that were incorrectly inserted in the *RNA_nucleotide_accession* column instead of in the *genomic_nucleotide_accession* column of the file, and hence provided as RefSeq IDs of RNA nucleotide sequences (Figure 3)
- A set of 3 IDs in the *protein_accession* column of the file that were not recognized as protein IDs: one of them (*NC_005847*) is the RefSeq ID of a genomic nucleotide sequence; the other two IDs (*AE009443_1.1* and *AE009443_2.1*) are probably GeneBank genomic nucleotide sequence IDs with wrong format (incorrectly including the additional *_1* or *_2* characters); all three IDs were incorrectly inserted in the

protein_accession column instead of in the *genomic_nucleotide_accession* column of the file.

tax_id	GeneID	status	protein accession.version	protein gi	genomic nucleotide accession.version	genomic nucleotide gi
3055	5715037	MODEL	XP_001689876	159463292	NW_001843471.1	159464175
3055	5715038	-	158283417	DS496108.1	158283400	-
3055	5715038	MODEL	XP_001689429.1	159462398	NW_001843471.1	159464175
3055	5715039	-	158283418	DS496108.1	158283400	-
3055	5715039	MODEL	XP_001689430.1	159462400	NW_001843471.1	159464175
3055	5715040	NA	NW_001843471.1	159464175	223205	224618
3055	5715041	-	158283865	DS496108.1	158283400	-
3055	5715041	MODEL	XP_001689877.1	159463294	NW_001843471.1	159464175
3055	5715042	-	158283866	DS496108.1	158283400	-
3055	5715042	MODEL	XP_001689878.1	159463296	NW_001843471.1	159464175
3055	5715043	-	158283419	DS496108.1	158283400	-
3055	5715043	MODEL	XP_001689431.1	159463402	NW_001843471.1	159464175

Figure 3: Excerpt of the July 30th, 2009 version of the Entrez Gene *gene2accession* file including the IDs with incorrect semantic assignment shown in the boxes

Furthermore, in the version of the *ec2go* tabular file downloaded from GOA [22] on July 30th, 2009 (ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/external2go/) we found an enzyme ID with wrong format (EC:3.6.1.52e); the regular expression we created for the Expasy enzyme IDs (Table 2) recognized the additional character *e* incorrectly included in the ID (Figure 4).

```
EC:3.6.1.43 > GO:dolichyldiphosphatase activity ; GO:0047874
EC:3.6.1.44 > GO:oligosaccharide-diphosphodolichol diphosphatase activity ; GO:0047430
EC:3.6.1.45 > GO:UDP-sugar diphosphatase activity ; GO:0008768
EC:3.6.1.52e > GO:diphosphoinositol-polyphosphate diphosphatase activity ; GO:0008486
EC:3.6.1.6 > GO:nucleoside-diphosphatase activity ; GO:0017110
EC:3.6.1.7 > GO:acylphosphatase activity ; GO:0003998
EC:3.6.1.8 > GO:ATP diphosphatase activity ; GO:0047693
```

Figure 4: Excerpt of the July 30th, 2009 version of the GOA *ec2go* file including the EC ID with wrong format shown in the box

Regarding ID evolution, we considered the ID history data provided by the Entrez Gene and Gene Ontology databases in order to manage their ID evolution and update available annotation data including their obsolete IDs. On July 30th, 2009 existed 796,802 discontinued Entrez Gene IDs (14.99% of all Entrez Gene IDs, with 96,346 (12.09%) of them that had been replaced with a new successor ID), and 1,121 discontinued Gene Ontology IDs (3.83% of all Gene Ontology IDs; all of them had been replaced with a new successor ID).

By checking the inclusion of Entrez Gene and Gene Ontology discontinued IDs in available annotation data and in case updating them to their current successor ID, on July 30th, 2009 we could update and make effectively usable numerous gene and protein annotations, including:

- 472 annotations of 187 Entrez Gene IDs to 104 KEGG biochemical pathways (0.71% of all such annotations provided by KEGG, and 90.78% of all the Entrez Gene IDs and 96.30% of all the KEGG biochemical pathways they included)
- 48,541 annotations of 982 Entrez Gene IDs with 330 eVOC ontology terms (49.71% of all such associations provided by eVOC in the version 2.9 of the annotation data with its four main ontologies that we considered (which describe the gene expression in human anatomical systems, cellular types, developmental stages, and pathologies [23]), and 42.57% of all the Entrez Gene IDs and 96.21% of all the eVOC terms they included)

- 11,380 associations of 4,364 Entrez Gene IDs with 9,428 UniSTS IDs (1.57% of all such associations provided by Entrez Gene, and 30.28% of all the Entrez Gene IDs and 36.17% of all the UniSTS IDs they included)

The multiple cases of discontinued IDs present in annotation data are due to the asynchronous updating of the different databases managing the IDs used in such annotations. Interestingly, discontinued IDs of a given data source may be present also in external annotation data provided by that same data source (such as the discontinued Entrez Gene IDs included in the mapping data between Entrez Gene IDs and UniSTS IDs provided by the Entrez Gene itself).

4.4 Unexpected data quantification values

By checking the presence of redundancies in data from a single database, our implemented automatic procedures found a few redundant data, including the two duplicated entries shown in Figure 5. Much more redundancies were found in data, especially annotation data, from different data sources.

```
05110   Vibrio cholerae infection
05111   Vibrio cholerae pathogenic cycle
05120   Epithelial cell signaling in Helicobacter pylori infection
05130   Pathogenic Escherichia coli infection - EHEC
05131   Pathogenic Escherichia coli infection - EPEC
05130   Pathogenic Escherichia coli infection - EHEC
05131   Pathogenic Escherichia coli infection - EPEC
05200   Pathways in cancer
05210   Colorectal cancer
05211   Renal cell carcinoma
```

Figure 5: Excerpt of the July 30th, 2009 version of the KEGG *map_title.tab* file, which provides pathway IDs and the correspondent pathway name; two redundant entries are shown in the boxes

In looking for redundancies, our implemented automatic procedures also check for the presence of similar entries that differ only for one or more secondary fields, which for example are included only in a data source. In this case, the different fields of the entries are merged together to produce a single non redundant entry.

Gene count (Entrez Gene IDs)	Number of proteins (UniProt IDs)
1	19,728
2	70
3	14
4	3
5	3
9	1
14	1

Table 3: Number of human proteins resulted to be encoded by one or more genes (*Gene count*)

Among the several checking quantifications implemented in our automatic procedures to scan the multi-organism data integrated in the GFINDer data warehouse for unexpected information patterns, we point out those recording the number of proteins encoded by one or

more genes; in Table 3 are shown the results for the human organism. In the obtained results some single protein IDs have been interestingly found associated with IDs of several different genes, thus describing the same protein as codified by all such genes. Thorough investigations of such results can evaluate the membership of these multiple genes encoding the same protein to a single gene family, thus contributing in the classification of homolog genes.

4.5 Consistency of data from multiple sources

Our implemented quality control automatic procedures also checks relationship loops among data and execute cross-checking among data imported from different sources. This enables us to verify data consistency and completeness, and helps unveiling unexpected information patterns possibly leading to biological discoveries. For example, on the assumption that, if a protein is annotated to a Gene Ontology term, the gene that codifies that protein must be annotated to that Gene Ontology term as well, we test consistency of GO annotations of proteins and their codifying genes integrated in the GFINDer data warehouse. By checking cross-references existing between Gene Ontology, UniProt and Entrez Gene databases, we found that 6,342 (3.98%) GO annotations (regarding 2,012 different GO terms) of 1,811 human proteins provided in the *gene_association.goa_human* file (downloaded on July 30th, 2009 from the GOA database) were not comprised in the GO annotations of the protein codifying genes provided in the *gene2go* file (downloaded on July 30th, 2009 from the Entrez Gene database). These protein GO annotations included also 2,221 (35.02%) annotations with evidence stronger than that inferred from electronic annotation (IEA).

5 Conclusions

The data warehousing approach, implemented to construct and maintain updated our GFINDer data warehouse, integrates efficiently large quantities of various biomolecular information sparsely available in numerous databases. It provides support to both increasing coverage of information partially provided by single data sources, and better controlling their correctness and completeness. The implementation on top of such integrative approach of the techniques for data quality assessment and improvement that are available in the literature ensures best possible quality of the data integrated in subsequent updating of the data warehouse. It also allows detecting errors and inconsistencies in the data provided by biological databases that are integrated. Among these techniques, syntactic checking of IDs by using regular expressions can highlight errors both in ID format and in their semantic use. Taking advantage of ID history data allows reconciling a great number of association data that involve different types of IDs. Reported results demonstrate both the relevant number of data from different databases that are not aligned due to the asynchronous updating of such databases, and the importance of using ID history data to reconcile as much as possible these misaligned data. Focused quantifications of integrated data identify redundant data imported from different data sources and also highlight unexpected information; the 14 different human genes that were found encoding the same identical gene products (http://www.uniprot.org/uniprot/P62805) are just an example. Cross-validation of data from multiple sources and analysis of data relationship loops proved effective in assessing completeness and consistency of data from different sources. Among others, they identified the lack of annotations in the Gene Ontology annotations of human genes provided by the Entrez Gene database. The use of such human Gene Ontology annotations without cross-checking their completeness would result in considering much less gene annotations than that known, with the consequent considerably less power of the analysis performed on them (e.g. the enrichment analysis of gene expression experiment results).

We reported all identified data errors and inconsistencies to the curators of the original databases from where the data were retrieved, who showed great collaboration and interest. In the majority of cases, the identified issues were corrected, or are in the process to be corrected, in subsequent updating of the original database. This demonstrates the relevance of our quality control effort in contributing to improve the quality of data available, in the original databases, to the whole scientific community.

The work here presented regards quality assessment and improvement of data available in biological databases, focusing mainly on annotation data. Yet, the data quality techniques here discussed, or similar ones, could be also successfully applied to improve integrity of the numerous experimental data produced by high-throughput molecular biology experiments.

References

[1] M. Y. Galperin and G. R. Cochrane. Nucleic Acids Research Annual Database Issue and the NAR Online Molecular Biology Database Collection in 2009. *Nucleic Acids Res*. 37(Database issue): D1-D4, 2009.

[2] T. J. Lee, Y. Pouliot, V. Wagner, P. Gupta, D. W. Stringer-Calvert, J. D. Tenenbaum and P. D. Karp. BioWarehouse: A Bioinformatics Database Warehouse Toolkit. *BMC Bioinformatics* 7: 170, 1-14, 2006.

[3] R. Stevens, P. Baker, S. Bechhofer, G. Ng, A. Jacoby, N. W. Paton, C. A. Goble and A. Brass. TAMBIS: Transparent Access to Multiple Bioinformatics Information Sources. *Bioinformatics* 16: 184-185, 2000.

[4] L. M. Haas, J. E. Rice, P. M. Schwarz, W. C. Swops, P. Kodali and E. Kotlar. DiscoveryLink: A System for Integrated Access to Life Sciences Data Sources. *IBM Systems Journal* 40: 489-511, 2001.

[5] T. Etzold, A. Ulyanov and P. Argos. SRS: Information Retrieval System for Molecular Biology Data Banks. *Methods Enzymol*. 266: 114-128, 1996.

[6] E. Cadag, B. Louie, P. J. Myler and P. Tarczy-Hornoch. Biomediator Data Integration and Inference for Functional Annotation of Anonymous Sequences. *Pac Symp Biocomput*. 343-354, 2007.

[7] M. Masseroli, S. Ceri and A. Campi. Integration and Mining of Genomic Annotations: Experiences and Perspectives in GFINDer Data Warehousing. In: N. W. Paton, P. Missier, C. Hedeler (eds). *Data Integration in the Life Sciences. 6th International Workshop, DILS 2009, proceedings; 2009 Jul 20-22; Manchester, UK*, pp. 88-95. Berlin, D: Springer-Verlag, 2009. LNCS (Lecture Notes in Bioinformatics; vol 5647).

[8] M. Masseroli, D. Martucci and F. Pinciroli. GFINDer: Genome Function INtegrated Discoverer through Dynamic Annotation, Statistical Analysis, and Mining. *Nucleic Acids Res*. 32: W293-W300, 2004.

[9] M. Masseroli. Management and Analysis of Genomic Functional and Phenotypic Controlled Annotations to Support Biomedical Investigation and Practice. *IEEE Trans. Inf. Technol. Biomed*. 11(4): 376-385, 2007.

[10] C. Batini and M. Scannapieco. *Data Quality: Concepts, Methodologies and Techniques*. Springer, 2006.

[11] S. E. Madnick, R. Y. Wang, Y. W. Lee and H. Zhu. Overview and Framework for Data and Information Quality Research. ACM *J. Data Inform. Quality* 1(1): 2, 1-22, 2009.

[12] C. Batini, C. Cappiello, C. Francalanci and A. Maurino. Methodologies for Data Quality Assessment and Improvement. *ACM Comput. Surv*. 41(3): 16, 1-52, 2009.

[13] D. Maglott, J. Ostell, K. D. Pruitt and T. Tatusova. Entrez Gene: Gene-centered Information at NCBI. *Nucleic Acids Res*. 35(Database issue) D26-D31, 2007.

[14] T. J. Hubbard, B. L. Aken, S. Ayling, B. Ballester, K. Beal, E. Bragin, S. Brent, Y. Chen, P. Clapham, L. Clarke, G. Coates, S. Fairley, S. Fitzgerald, J. Fernandez-Banet, L. Gordon, S. Graf, S. Haider, M. Hammond, R. Holland, K. Howe, A. Jenkinson, N. Johnson, A. Kahari, D. Keefe, S. Keenan, R. Kinsella, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, K. Megy, P. Meidl, B. Overduin, A. Parker, B. Pritchard, D. Rios, M. Schuster, G. Slater, D. Smedley, W. Spooner, G. Spudich, S. Trevanion, A. Vilella, J. Vogel, S. White, S. Wilder, A. Zadissa, E. Birney, F. Cunningham, V. Curwen, R. Durbin, X. M. Fernandez-Suarez, J. Herrero, A. Kasprzyk, G. Proctor, J. Smith, S. Searle and P. Flicek. Ensembl 2009. *Nucleic Acids Res*. 37(Database issue): D690-D697, 2009.

[15] K. D. Pruitt, T. Tatusova and D. R. Maglott. NCBI reference sequences (RefSeq): a Curated Non-Redundant Sequence Database of Genomes, Transcripts and Proteins. *Nucleic Acids Res*. 35(Database issue): D61-D65, 2007.

[16] UniProt Consortium. The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res*. 37(Database issue): D169-D174, 2009.

[17] P. J. Kersey, J. Duarte, A. Williams, Y. Karavidopoulou, E. Birney and R. Apweiler. The International Protein Index: an Integrated Database for Proteomics Experiments. *Proteomics* 4(7): 1985-1988, 2004.

[18] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, OBI Consortium, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S. A. Sansone, R. H. Scheuermann, N. Shah, P. L. Whetzel and S. Lewis. The OBO Foundry: Coordinated Evolution of Ontologies to Support Biomedical Data Integration. *Nat Biotechnol*. 25(11): 1251-1255, 2007.

[19] J. Kelso, J. Visagie, G. Theiler, A. Christoffels, S. Bardien, D. Smedley, D. Otgaar, G. Greyling, C. V. Jongeneel, M. I. McCarthy, T. Hide and W. Hide. eVOC: A Controlled Vocabulary for Gene Expression Data. *Genome Res*. 13(6A): 1222-1230, 2003.

[20] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, G. Sherlock and The Gene Ontology Consortium. Gene Ontology: Tool for the Unification of Biology. *Nat. Genet*. 25(1): 25-29, 2000.

[21] M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu and Y. Yamanishi. KEGG for linking genomes to life and the environment. *Nucleic Acids Res*. 36(Database issue): D480-D484, 2008.

[22] E. Camon, M. Magrane, D. Barrell, V. Lee, E. Dimmer, J. Maslen, D. Binns, N. Harte, R. Lopez and R. Apweiler. The Gene Ontology Annotation (GOA) Database: Sharing

Knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res*. 32(Database issue): D262-D266, 2004.

[23] M. Ceresa, M. Masseroli and A. Campi. A Web-enabled Database of Human Gene Expression Controlled Annotations for Gene List Functional Evaluation. In: A. Dittmar and J. Clark (eds.) *29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS 2007)*, pp. 394-397. Stoughton, WI: The Printing House, 2007.