Quality checking of data in biomolecular databases

Ghisalberti G[1], Masseroli M[1], Tettamanti L[1]

Motivation

Data quality is a growing important issue in bioinformatics due to the rapidly increasing amount of experimental data and knowledge available in numerous distributed biomolecular databases. They provide extremely valuable information, but only partially curated by experts and mostly computationally derived. Inconsistencies and several kinds of errors no rarely exist in such data. Thus, the effective use of these data to derive new knowledge, or to support the interpretation of experimental results, requires their integration and correction of the errors and inconsistencies that they include. Here we illustrate the data quality techniques that we implemented to test the quality of genomic and proteomic annotation data from numerous biomolecular databases, which we integrated in the data warehouse of our GFINDer system (http://www.bioinformatics.polimi.it/GFINDer/). We focused on the assessment and improvement of the accuracy and consistency (two fundamental data quality dimensions) of the integrated biomolecular annotations. To this aim, we implemented a set of automatic procedures that ensure the best possible quality of the data integrated in the GFINDer data warehouse.

Methods

In order to analyze the quality of data from Entrez Gene, eVOC, GO, GOA, KEGG, Reactome, IPI, UniProt and other several different biological databases integrated in the GFINDer data warehouse, we implemented a set of quality controls that test these data for a variety of different types of errors and inconsistencies. Among others, they check data structure and completeness, ontological data consistency, ID format and evolution, and consistency of data from single and multiple sources. We designed and implemented automatic procedures in Java programming language that verify absence of inappropriate missed data and inconsistent data structures. They also automatically check the correctness of ontological data, i.e. they verify if these data describe a topologically correct ontological graph. Furthermore, our developed procedures automatically identify and syntactically check the numerous different types of IDs present in data from biomolecular databases. To this aim, we adopted a set of regular expressions that describe the correct ID formats. We used them to recognize ID type and provenance, and to control their correct semantic use. By taking advantage of available ID historical

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

data, our automatic procedures also control ID evolution, which often occur in subsequent updates of the biomolecular databases. Furthermore, they check data from both single and multiple databases for duplicates and presence of similar entries. In order to identify redundant or mismatching data, we also implemented automatic cross-controls among data imported from different sources. When multiple independent sources provide overlapping data, we use such overlaps to verify the information they provide and increase its likelihood.

Results

Our implemented data quality automatic procedures identified numerous data errors and inconsistencies in the data provided by several biological databases. The adopted regular expressions identified numerous IDs with wrong format or inconsistent semantic assignment, including several RefSeq IDs provided by the Entrez Gene database. The considered ID history data enabled us to reconcile and make effectively usable many gene and protein annotation data. Cross-comparison of data from different sources, by checking and taking advantage of relationship loops among annotation data, allowed verifying both consistency and completeness of different data sources. 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 tested consistency of GO annotations of proteins and their codifying genes 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 were not comprised in the GO annotations of the protein codifying genes, including also 2,221 (35.02%) protein annotations with evidence stronger than that inferred from electronic annotation (IEA). The implemented data quality procedures demonstrated effective in detecting errors and inconsistencies in the data provided by biomolecular databases and unveiling unexpected information patterns, which might lead to biological discoveries. We reported all identified data errors and inconsistencies to the curators of the original databases from where the data were retrieved. In the majority of cases, the identified issues were corrected in subsequent updating of the original database, demonstrating 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.

Contact e-mail

masseroli@elet.polimi.it