Chapter XVI
Verification of Uncurated Protein Annotations

Francisco M. Couto
Universidade de Lisboa, Portugal

Mário J. Silva
Universidade de Lisboa, Portugal

Vivian Lee
European Bioinformatics Institute, UK

Emily Dimmer
European Bioinformatics Institute, UK

Evelyn Camon
European Bioinformatics Institute, UK

Rolf Apweiler
European Bioinformatics Institute, UK

Harald Kirsch
European Bioinformatics Institute, UK

Dietrich Rebholz-Schuhmann
European Bioinformatics Institute, UK

ABSTRACT

Molecular Biology research projects produced vast amounts of data, part of which has been preserved in a variety of public databases. However, a large portion of the data contains a significant number of

errors and therefore requires careful verification by curators, a painful and costly task, before being reliable enough to derive valid conclusions from it. On the other hand, research in biomedical information retrieval and information extraction are nowadays delivering Text Mining solutions that can support curators to improve the efficiency of their work to deliver better data resources. Over the past decades, automatic text processing systems have successfully exploited biomedical scientific literature to reduce the researchers' efforts to keep up to date, but many of these systems still rely on domain knowledge that is integrated manually leading to unnecessary overheads and restrictions in its use. A more efficient approach would acquire the domain knowledge automatically from publicly available biological sources, such as BioOntologies, rather than using manually inserted domain knowledge. An example of this approach is GOAnnotator, a tool that assists the verification of uncurated protein annotations. It provided correct evidence text at 93% precision to the curators and thus achieved promising results. GOAnnotator was implemented as a web tool that is freely available at http://xldb.di.fc.ul.pt/rebil/tools/goa/.

INTRODUCTION

A large portion of publicly available data provided in biomedical databases is still incomplete and incoherent (Devos and Valencia, 2001). This means that most of the data has to be handled with care and further validated by curators before we can use it to automatically draw valid conclusions from it. However, biomedical curators are overwhelmed by the amount of information that is published every day and are unable to verify all the data available. As a consequence, curators have verified only a small fraction of the available data. Moreover, this fraction tends to be even smaller given that the rate of data being produced is higher than the rate of data that curators are able to verify.

In this scenario, tools that could make the curators' task more efficient are much required. Biomedical information retrieval and extraction solutions are well established to provide support to curators by reducing the amount of information they have to seek manually. Such tools automatically identify evidence from the text that substantiates the data that curators need to verify. The evidence can, for example, be pieces of text published in BioLiterature (a shorter designation for the biological and biomedical scientific literature) describing experimental results supporting the data. As part of this process, it is not mandatory that the tools deliver high accuracy to be effective, since it is the task of the curators to verify the evidence given by the tool to ensure data quality. The main advantage of integrated text mining solutions lies in the fact that curators save time by filtering the retrieved evidence texts in comparison to scanning the full amount of available information. If the IT solution in addition provides the data in conjunction with the evidence supporting the data and if the solutions enable the curators to decide on their relevance and accuracy, it would surely make the task of curators more effective and efficient.

A real working scenario is given in the GOA (GO Annotation) project. The main objective of GOA is to provide high-quality GO (Gene Ontology) annotations to proteins that are kept in the UniProt Knowledgebase (Apweiler et al., 2004; Camon et al., 2004; GO-Consortium, 2004). Manual GO annotation produces high-quality and detailed GO term assignments (i.e. high granularity), but tends to be slow. As a result, currently less than 3% of UniProtKb has been confirmed by manual curation. For better coverage, the GOA team integrates uncurated GO annotations deduced from automatic mappings between UniProtKb and other manually curated databases (e.g. Enzyme Commission numbers or InterPro domains). Although these assignments