Annex I: Data Integration

Description

Data integration is one of the main challenges in bioinformatics. Although the term “data integration” often appears in research, especially in life science, a consolidated definition is still missing. We report here several definitions of “data integration” which will help in understanding the central concept and how it evolved.

In the beginning, the “data integration” problem was restricted to database research. Ziegler et al. describe the integration problem as the aim at “combining selected systems so that they form a unified new whole and give users the illusion of interacting with one single information system” (Ziegler & Dittrich 2004). A higher level definition comes from Leser and Naumann which defines “data integration” as “a redundancy-free representation of information from a collection of data sources with overlapping contents” (Leser & Naumann 2007). More recently, Gomez-Cabrero et al. describe “data integration as the use of multiple sources of information (or data) to provide a better understanding of a system/situation/association/etc” (Gomez-Cabrero et al. 2014).

Out of these three options, the last definition is the most general and complete. The authors emphasized the use of several sources to discover new insights that are hidden when looking at a single source. We propose to adopt this definition of “data integration” and consider the two main challenges associated with it: 1) finding relevant data sources (data discovery); 2) using collected data to produce new findings (data exploitation) (Weidman and Arrison 2010). Additionally, we tackle the problem of data provenance, which arises once data sources are integrated and allow to trace the origin of each piece of information.

Data discovery

The data discovery challenge entails the search of relevant data sources for describing a system. Nowadays, finding data sources for a specific problem is easy. However, finding a relevant source for a system of interest is becoming more and more challenging. The ease of publishing data through the Web has contributed to an explosion of online resources and publicly accessible datasets. Furthermore, the picture becomes more and more fragmented as each sub-discipline provides its data representation. In biology, the problem is not only arising from aiming to combine different omics but also to reconcile information even within the same field. For example in glycomics, several different formats are available to describe glycan structures.
We list here the main three:

- IUPAC format which is regulated by the International Union of Pure and Applied Chemistry
- GlycoCT format which has been developed under the EUROCarbDB project
- WURSC format which is developed in Japan and used for the structure repository GlyTouCan.

The creation of more and more heterogeneous data sources leads to what has been called “A loose federation of bio-nation” (Goble and Stevens 2008). In this context, the use of standards and community guidelines, which are going to be tackled in “Data integration in bioinformatics” paragraph, is the only way to stop the data fragmentation smoothing the way towards an efficient data discovery.

Data exploitation

Data exploitation is the process of exploring a set of data sources to provide new insights. Before starting the data exploitation, scientists must have a complete overview of each dataset including the units of measure and more subtle aspects such as environmental conditions, equipment calibrations, preprocessing algorithms, etc (Weidman and Arrison 2010). Having an in-depth knowledge of the data is the only way to avoid the phenomenon of “garbage-in-garbage-out” that could affect the data exploitation outcome.

Once each data set has been fully clarified, researchers can focus on developing methodologies to analyse the data. Methodologies can vary according to the different data types. In this regards, we distinguish between “similar” and “heterogeneous” types. According to Hamid et al. we consider “similar type” when they are produced by the same underlying source, for example, they are all gene expression data sets. If multiple data sources, like gene expression and protein quantification data sets, are taken into account, we refer to data as “heterogeneous type” (Hamid et al. 2009). Although researchers are developing more and more hybrid methodologies for data integration, a set of general techniques will be presented in the next paragraph.

The creation of data explorative tools is the last but not the least step of the data exploitation process. The development of user-friendly interfaces to navigate the outcome of data exploitation can decree the success of the whole process. Although scientist are usually putting most of their attention on methodologies and algorithms, the design of custom visualisation can be time-consuming especially due to data heterogeneity and data volume.

Data provenance

With the integration of more and more data sources, understanding the derivation of each piece of information becomes an issue. This is referred to as the data provenance issue in the literature. As in the case of data integration, data provenance has a different definition according to the field. In the context of database systems, Buneman et al. describe it as “the description of the origins of a piece of data and the process by which it arrived in a Database” (Buneman, Khanna, & Wang-Chiew 2001). However, data provenance is not only related to the data but also to the processes that led to the creation of the data. For this reason, Greenwood et al. (Greenwood et al. 2003) define data provenance as part of metadata which records “the process of biological experiments for e-Science, the purpose and results of experiments as well as annotations and notes about experiments by scientists”.

We characterise data provenance as the record of the origin and all transformations that every dataset has undergone. It is quite a central question since the corresponding solutions enable users to trace back to the original data. Additionally, this concern for tracing information back provides the means to understand the different workflows that have been applied to integrate each dataset. For example, if a user is interested in removing all datasets that are produced from mouse, using the data provenance metadata, he/she can exclude these datasets and use the rest of information to perform new analyses (Masseroli et al. 2014). Data provenance is also essential in the assertion of the quality of a dataset, especially in fields where information is treated manually and the quality can vary according to the curator.

Category

1. News