Annex III: Data Integration in Bioinformatics

Description

All data integration techniques presented in the previous paragraphs need touch points to be implemented. In bioinformatics, a diversity of efforts have been carried out to provide standards and, as a matter of fact, touch points across data sources. In this paragraph, we identify some areas of study which are crucial to enforce standardisation and encourage data integration.

Standards

In life science, where data can be represented in many ways, widely adopted standards provide the only ground for data exchange and data integration. To show why standards are so relevant, we take the example of the amino acid naming convention. The 20 amino acids have a standard name that is recognised worldwide. Additionally, each amino acid has a one and three letter codes that are used by biologists around the globe. If a European biologist talks at a conference in Asia about the S amino acid, everyone in the audience understands it is about Serine.

Nowadays, lots of initiatives for developing standards are arising. We took a list of the most famous from a paper published by Lapatas et al. which is available in Table I (Lapatas et al. 2015). We will not explore each of these initiatives, but we provide URLs for more information.

To conclude, we want to strive the importance of standards for data sharing. Standards facilitate data re-use, limiting the work needed to integrate different data sources and the waste of potential datasets.

List of data standard initiatives. Courtesy of “Data integration in biological research : an overview” (Lapatas et al. 2015).

- **OBO** The Open Biological and Biomedical Ontologies www.obofoundry.orgEstablish a set of principles for ontology development to create a suite of orthogonal interoperable reference ontologies in the biomedical domain PMID=17989687
- **CDISC** Clinical data interchange standards consortium www.cdisc.orgEstablish standards to support the acquisition, exchange, submission and archive of clinical research data and metadata PMID=23833735
- **HUPO- PSI** Human Proteome Organisation- Proteomics Standards Initiative www.psidev.info Defines community standards for data representation in proteomics to facilitate data comparison, exchange and verification PMID=16901219
- *Alliance* Global Alliance for Genomics and Health genomicsandhealth.org Create interoperable approaches to catalyze projects that will help unlock the great potential of genomic data PMID=24896853
- *COMBINE* Computational Modeling in Biology co.mbine.orgCoordinate the development of the various community standards and formats for computational models PMID=25759811

- *MSI* Metabolomics Standards Initiative msi-workgroups.sourceforge.net

- Define community-agreed reporting standards, which provided a clear description of the biological system studied and all components of metabolomics studies PMID=17687353
- **RDA** Research Data Alliance rd-alliance.org Builds the social and technical bridges that enable open sharing of data across multiple scientific disciplines

Ontologies

In the last twenty years, several ontologies have been created in the biological and biomedical fields (Bard & Rhee 2004 ; Hoehndorf, Schofield, & Gkoutos 2015 ; Kelso, Hoehndorf & Prüfer 2010). In philosophy, an ontology describes “what exists”, whereas, in life science, it represents what exists in a specific context, for example, diseases (Turk 2006). **An ontology is defined as a collection of concepts and relationships used to characterise an area of concern.**

To consolidate and coordinate the rapid spread of ontologies, in 2001, Ashburner and Lewis established The Open Biomedical Ontology (OBO) consortium. The OBO ontologies form the basis of OBO Foundry, *a collaborative experiment based on the voluntary acceptance by its participants of an evolving set of principles that extend those of the original OBO* (Smith et al. 2007). As stated in its website, the OBO Foundry “is a collective of ontology developers that are committed to collaboration and adherence to shared principles”, and its mission “is to develop a family of interoperable ontologies that are both logically well-formed and scientifically accurate”.
OBO contains ten ontologies (June 2018) which are member ontologies like Gene Ontology (GO) and more than a hundred candidate ontologies. To become member, a candidate ontology has to be developed using OBO’s shared principles and validated by OBO members (Quesada-Martínez et al. 2017).
The growing use of OBO ontologies allows connecting more and more datasets using techniques like the semantic web. This enhances data integration and fosters the creation of the web of data.

Formats and reporting guidelines

As data is increasingly generated by high throughput techniques, computers, equipped with appropriate software, are required to store and analyse the information produced. In this scenario, data formats play a critical role as they provide instructions to store data in a file. However, the scarcity of well-designed data formats gives birth to many different standards that hamper data exchange and data integration. Therefore, bioinformaticians are forced to build converters, spending more time to clean data than to analyse them. Currently, to store Next Generation Sequence (NGS) data, there are six common file formats : FASTQ, FASTA, SAM/BAM, GFF/GTF, BED, and VCF (Zhang 2016).
To ease and foster data sharing, converters between different data format have been developed in genomics, proteomics, etc. In this context, a good example is provided by the PRoteomics IDEntifications (PRIDE) project (Vizcaíno et al. 2016). PRIDE is a centralised repository for proteomics data which includes protein and peptide identification as well as post-translational modifications. Since mass spectrometry data can be encoded with several formats, the PRIDE development team have developed PRIDE Converter. This tool converts a large amount of mass spectrometry encoding formats into a valid PRIDE XML ready for submission to the PRIDE repository (Barsnes et al. 2009).
With the gradual maturation of “omics”, a huge step forward has been made by adopting clear guidelines to describe and depositing datasets. In this context, the first example of a concrete guideline is represented by The Minimum Information About a Microarray Experiment (MIAME) (Brazma et al. 2001) . MIAME defines “the minimum information required to ensure that microarray data can be easily

interpreted and that results derived from its analysis can be independently verified". The use of MIAME facilitates the creation of public repositories as well as the development of data analysis tools.
Following the example of MIAME, in 2007, Taylor et al. published The minimum information about a proteomics experiment (MIAPE) (Taylor et al. 2007). Nowadays, proteomics guidelines are defined by HUPO Proteomics Standards Initiative (Hermjakob 2006) which additionally proposes data formats and controlled vocabularies (http://www.psidev.info).
In general, these guidelines focus on defining the content and the structure of necessary information to describe a specific experiment. Although they do not provide any technical solution for storing data, some of them suggest standard file formats.
To conclude, the use of minimum information guidelines together with suggested data formats enhance the data integration progress and the reusability of datasets.

Identifiers

An identifier is a short list of characters which identifies a data entry. For example, UniProt (Bateman et al. 2017) is using accession numbers, i.e. stable identifiers, to identify entries. When two or more entries are merged, all the accession numbers are kept. In this case, one is the "Primary (citable) accession number" whereas the others become "Secondary accession numbers". To avoid any source of uncertainty, it is not possible that one accession number refers to multiple proteins.
In life science, the information is spread across multiple databases, and each of them has developed its identifiers. This leads to a multitude of identifiers to describe the same biological concept. However, to facilitate data integration, databases have cross-referenced their entries with external resources (See chapter 1 – Link integration). In UniProt, for example, each protein has a cross-reference section which contains all external identifiers related to the same protein (3D structures, protein family, genome annotation, etc) (Figure 4).
If all life science databases used the same identifier to characterise a biological concept, data integration would be facilitated. However, the use of identifiers from established databases, like UniProt or GenBank, in research papers is already a sign of progress.
In genomics, for example, most journals oblige researchers to deposit newly obtained DNA and amino acid sequences to a public sequence repository (DDBJ/ENA/Genbank – INSDC) as part of the publication process (Benson et al. 2018). Although this is happening in advanced fields like genomics and proteomics, disciplines like glycomics are lagging behind.

Visualisation

In biology, data visualisation is an essential part of the research process. Scientists have always relied on different visualisation means to communicate experimental results. Some domains of biology like phylogeny (Allende, Sohn, & Little 2015) and pathway analysis (Tanabe & Kanehisa 2012) have created specific visualisations that, nowadays, are considered a standard (i.e. phylogenic trees). The spreading of high throughput technologies has complicated the panorama. The increasing quantity of data and the integration of heterogeneous information have created new challenges for visualisation experts.
For example, the rise of the next generation sequencing and the resulting availability of genome data has prompted the need for new custom visualisations to show sequence alignments, expression patterns or entire genomes (Gaitatzes et al. 2018).
Data visualisation is also becoming a crucial resource in the integration of multiple resources. General purpose tools are available to overlay data from different data sources. An example is Cytoscape

(Shannon et al. 2003), an open source software platform for visualising interaction networks and biological pathways. Cytoscape gives the possibility to overlay networks with gene expression profiles, annotation and other quantitative and qualitative data.

Category

1. News