Data integration or die: The importance of biologist input in efficiently sharing data

October 6 2015

Researchers reviewed the importance of biologist input in efficiently sharing data. Credit: TGAC

Vicky Schneider, 361° Division at The Genome Analysis Centre, along with UK and European partners, has reviewed key aspects of standards

and formats of biological data to highlight the importance of data integration and management tools for biologists.

Data format structural standards are critical to the intrinsic value of analyses, with regard to retrieval, sharing, validation, reproducibility, and particularly, integration and interpretation.

Integrating data is imperative for the advancement of research; blending results of diverse disciplines is often an essential step in answering meaningful biological questions. To achieve this, standards should be implemented at the source of the data for the sake of efficiency, particularly since the datasets are constantly increasing in size, and it may be almost impossible to achieve unification further downstream.

In order to engage the biologist community, the aim of the scientific paper is to familiarise experimental biologists with definitions and terms used by computational biologists, to foster cooperation towards cohesive data flow pipelines. Four main classes of data format are identified, (tables, FASTA, Genbank and tag-structured), a major step in defining how the multitude might be curated.

Data integration in biological research is centred on standards adoption promising easier conversion between data/file formats. The scale and infrastructure of a given database determine whether it should be stored in a centralised or distributed manner, with a trade-off against the difficulty of updating or querying, respectively. Either way, when the data needs to be (further) integrated (with other data), the computational burden of unifying formats should be eased wherever possible.

Ideally biologists should work with bioinformaticians and computer scientists to get more involved with standardising their data structures, reducing the ongoing issue of database management and programming tools to parse data. This will boost biological research, gaining a more

robust structure for data analysis.

Senior Author, Dr Vicky Schneider, Head of the 361° Division at TGAC, said: "Data integration should not just rely on software engineers and computational scientists, but needs to be driven by the actual users whose communities need to define, adopt and use standards, ontologies and annotation best practice. Therefore, it is particularly important for the biological research community to get acquainted with the conceptual basis of data integration, its limitations, challenges and terminology."

Senior Author, Dr Allegra Via, Assistant Professor in the Biocomputing Group of Sapienza, University of Rome, added: "The importance of biologists in data integration is huge. They are those who produce and analyse data, which need to be shared for a better science. There cannot be data sharing without good practice in data integration."

The paper, titled: "Data Integration in Biological Research: An overview" is published in PubMed. The publication is a collaborative effort between TGAC, Department of Informatics at Ionian University, the ELIXIR Hub and Biocomputing Group, Sapienza University.

Provided by The Genome Analysis Centre