Visualization of experimental design & workflows in biological experiments

Eamonn Maguire  Philippe Rocca-Serra  Min Chen  Susanna-Assunta Sansone

Oxford e-Research Center, University of Oxford, UK

ABSTRACT

Evidence based science mandates full disclosure of scientific data and ancillary experimental metadata payload. Only this data provenance can warrant the possibility of reviewing findings and claims. Reports should therefore contain information about biological materials, their treatments and the molecular dimensions being surveyed and the methods to perform those measurements.

The ISA-Tab format [1] provides a generic grammar to organize and structure experimental information and linking it to matrices of results. While spreadsheet based representation is popular as a natural interface, devising new means to graph experiments and sample processing workflows can allow for more immediate understanding of overall experiment design. In particular by mimicking the practice of drawing experimental groups and applying the technique of dimensionality reduction, we are working at devising more engaging, compact and informative renderings than techniques currently available.

We posit that the same techniques could be used to plan experiments and drive data acquisition (therefore providing game changing means to data management).

KEYWORDS: Bioinformatics Visualization, Time Series Data, Tabular Data, Illustrative Visualization.

1 EXPERIMENTAL DESIGN OVERVIEWS: A PICTURE OR 1000 WORDS?

Hypothesis evaluation and testing are at the core of scientific practice. In this world, the notion of design of experiment is key. Declaring variables, sample sizes and procedure should be the norm. Yet, extracting this information from public resources is seldom easy, let alone the task of presenting it in a meaningful way.

While making a judgment call over experimental design choices should remain the realm of scientific debate, providing the means to present essential information about experimental design attributes such as nature of replication, variables and their level is a reasonable endeavor.

We have developed a simple view (see figure 1) to show distribution of biological replicates across treatment conditions in our ISAcreator tool [1]. Limitations of this current method are that there is no timeline support or concept of treatment intensity. Aware of such limitations, we will expand on this initial work to create a reader-driven, narrative visualization [2] capable of telling the underlying story behind an experiment in as little or as much detail as the reader desires.

The preliminary idea for such a visualization is shown in figure

2 EXPERIMENTAL WORKFLOWS

Workflows are made up of protocol application events occurring over the course of the experiment and enacted on physical objects (biological specimen, populations) or digital object (data files, images). Recording them properly vouches for reproducibility of experiments and helps perform analysis downstream. Some effort has gone into visualizing experimental

---

2. This illustration is describing visually what is represented in text in [3,4] and shows just how much easier it would be for a user/reviewer/curator to extract the experimental design when depicted as an image than through plain text. However, this initial design needs to be made general so as to represent all types of experiments.

---

2. Display of the experimental design for an example toxicogenomics investigation. Treatment of rats with 10 compounds at low, medium and high doses with 3 sacrifice time points, taking blood, urine, liver and kidney samples.
workflows in the past with the general mechanism for doing so being the graph layout.

Most of these visualizations do not scale well (see figure 3a) and even simple workflows are difficult to represent using this approach [6, 7].

One of the better methods we have seen for this is in the SIDR data repository study report [5] like that shown in figure 3b. At present, their experimental workflow graph is drawn manually; doing so allows filtering of unnecessary information in advance by trained curators.

Elaborating on this approach, we wish to provide an automated mechanism to visualize the experimental workflow, but in a way differing from techniques tried thus far. Since this visualization is largely time oriented, lessons can be learned from efforts in visualizing genealogical [9] data for example since many of the operations we wish to represent in this visualization are similar. For instance, splitting of protocols and merging of samples onto one DNA microarray can be translated to what is seen in genealogical trees with marriage and divorce.

3 WORK PLAN

In the coming months, we will investigate ways to make our proposals scale and devise a prototype implementation to generate interactive visualizations of the experimental design and workflows and link these visualizations together.

We envisage a few challenges during design and implementation phases, namely:

1. Process content for visualization: through use of the ISA-Tab [1] format as the sole entry format to generate these visualizations, extraction of key information for the visualization is relatively straightforward. However, being able to distinguish between treatment group intensities and timeline data (for example) for any given ISA-Tab submission is challenging owing to the inconsistent semantics encompassing much experimental metadata. We envisage use of semantic tagging and ontological frameworks to aid content processing;

2. Devising layout algorithms: Layout algorithms are a non-trivial problem and coming up with algorithms to create these visualizations will require better investigation or current techniques or generation of algorithms to do this task. Moreover, we need to take into consider current practices by biologists when they draw their experimental designs so as to increase uptake of a visualization which can do it for them;

3. Resolving scaling issues: Overcrowding resulting from increased dimensionality is a problem very prevalent in biological visualizations. In general, it can be addressed through identification of patterns and presenting abstracted views to the user but knowing what and how to abstract is a problem which will need to be solved; and

4. Iconographic ambiguity: can be addressed through definition of conventions to enable creation of generic icons and investigating how the biologists represent concepts when drawing their experimental designs will be an important task.

ACKNOWLEDGMENTS

Funding: BBSRC, NERC (BB/I000917/1, BB/I025840/1).

REFERENCES