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.

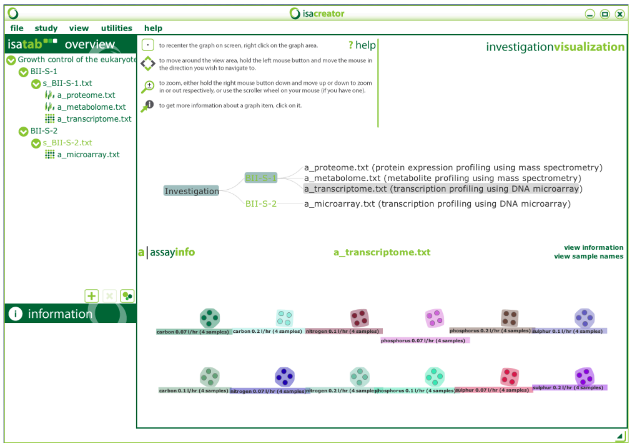
1 Experimental Design Overviews: A Picture or 1000 Words?

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*. 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


isacreator
file
study
view
utilities
help
isatab overview
Growth control of the eukaryote
BII-S-1
s_BII-S-1.txt
a_proteome.txt
a_metabolome.txt
a_transcriptome.txt
BII-S-2
s_BII-S-2.txt
a_microarray.txt
information
? help
investigationvisualization
Investigation
BII-S-1
BII-S-2
a|assayinfo
a_transcriptome.txt
view information
view sample names
carbon 0.07 l/hr (4 samples)
carbon 0.2 l/hr (4 samples)
nitrogen 0.1 l/hr (4 samples)
phosphorus 0.07 l/hr (4 samples)
phosphorus 0.2 l/hr (4 samples)
sulphur 0.1 l/hr (4 samples)
carbon 0.1 l/hr (4 samples)
nitrogen 0.07 l/hr (4 samples)
nitrogen 0.2 l/hr (4 samples)
phosphorus 0.1 l/hr (4 samples)
sulphur 0.07 l/hr (4 samples)
sulphur 0.2 l/hr (4 samples)

to recenter the graph on screen, right click on the graph area.

to move around the view area, hold the left mouse button and move the mouse in the direction you wish to navigate to.

to zoom, either hold the right mouse button down and move up or down to zoom in or out respectively, or use the scroller wheel on your mouse (if you have one).

to get more information about a graph item, click on it.

a_proteome.txt (protein expression profiling using mass spectrometry)
a_metabolome.txt (metabolite profiling using mass spectrometry)
a_transcriptome.txt (transcription profiling using DNA microarray)
a_microarray.txt (transcription profiling using DNA microarray)

Figure 1. Graphical view of treatment groups shown as circular blobs whose size depends on the number of replicates in such groups.

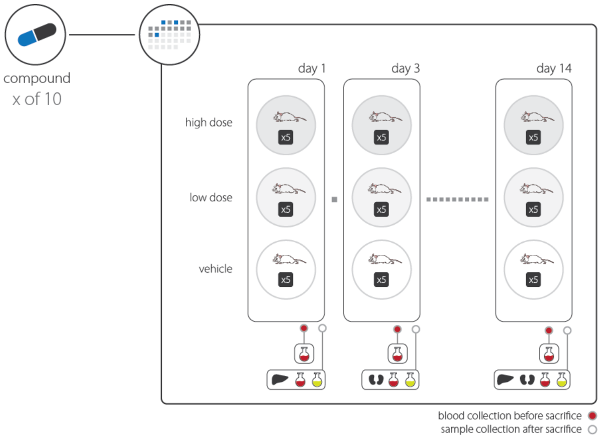

compound
x of 10
day 1
day 3
day 14
high dose
low dose
vehicle
x5
blood collection before sacrifice
sample collection after sacrifice

Figure 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.

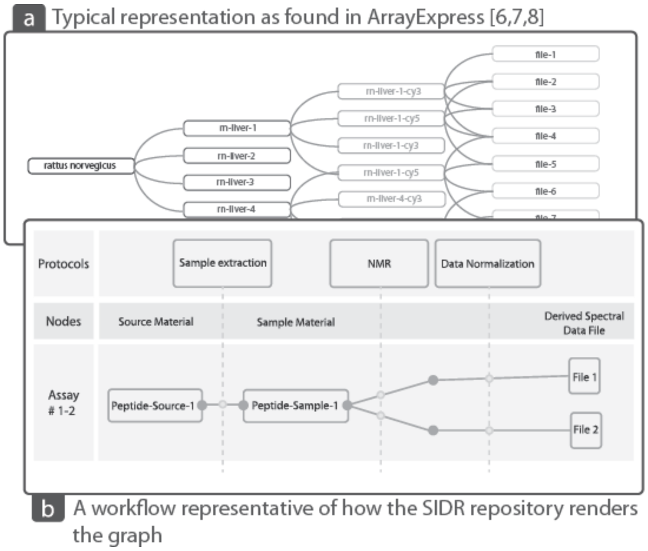
2 Experimental Workflows

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

- e-mail: {eamonn.maguire, philippe.rocca-serra, min.chen, susanna-assunta.sansone}@oerc.ox.ac.uk 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.


a Typical representation as found in ArrayExpress [6,7,8]

b A workflow representative of how the SIDR repository renders the graph

Figure 3. **a)** Typical display of the experimental workflow for an experiment as rendered in ArrayExpress [8] for [7].This approach breaks quickly when there are lots of samples and protocols. **b)** SIDR has manually created an image like this showing the experimental workflow [5]. From our investigation, we have found that this, in our opinion is the best current representation of an experimental workflow since it is compact and concise.

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

[1] Rocca-Serra et al, ISA software suite, Bioinformatics 2010 15:26.
[2] Segal, E. and Herr, J. Narrative Visualization: Telling Stories with Data. IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis), 2010.
[3] Study FP001RO: Evaluation of the Acute Toxicity, Gene Expression, Protein Expression, Metabolite Production, Clinical Chemistry and Pathology Profile Following an Oral Administration of Compound R2717 to Rats, http://www.ebi.ac.uk/bioinvindex/study.seam?studyId=BII-S-8.
[4] Collins B.C. *et al*, Use of SELDI MS to discover and identify potential biomarkers of toxicity in InnoMed PredTox: a multi-site, multi-compound study, Proteomics 2010 10:8
[5] SIDR, Experiment report document http://sidr-dr.inist.fr/RepoSidr/sabnp_curmi-1/reports/report_s_I19L-2.pdf
[6] E-BUGS-65 Experimental workflow, http://www.ebi.ac.uk/arrayexpress/files/E-BUGS-65/E-BUGS-65.biosamples.png
[7] E-MTAB-346 Experimental workflow, http://www.ebi.ac.uk/arrayexpress/files/E-MTAB-346/E-MTAB-346.biosamples.png
[8] ArrayExpress repository, http://www.ebi.ac.uk/arrayexpress/
[9] Nam Wook, K. *et al*. Tracing Genealogical Data with TimeNets. Advanced Visual Interfaces, 241–248, 2010