BMI 210 Project Milestone

Julia Daniel, Eric Cramer, Gaby Steiner, Akhila Moturu

Abstract

Our project aims to build a resource to guide biomedical or biological science researchers in devising statistically appropriate and well designed studies. To make this tool, we will create an ontology of study designs, technologies, statistical methods, and how they should be used in the context of a specific goal and data set. We will use user input to identify correct study types, pertinent wet lab technologies, and appropriate statistical methods given the data generated from the study. The goal of this tool will be to guide new biomedical researchers to design good studies.

We are creating an ontology of study designs, associated wet lab technologies, and implied statistical methods (for example, where each of these are superclasses with many subclasses associated with each other - a study design has_some technology which_uses some statistical analysis). A front end tool will accept user input to determine which study type, biomedical technology, and statistical analysis combination would be optimal for producing a study with sound conclusions.

Background and Motivation

One of the most serious issues currently facing the scientific community is faulty experimental design. As noted in an Atlantic article published in 2010, “much of what medical researchers conclude in their studies is misleading, exaggerated, or flat-out wrong” (Freedman, 2010). As was concluded by doctor and researcher John Ioannidis, this issue is due in large part to poor biomedical study design, incorrect implementation of study types, or misinterpretation of statistical measures used to characterize results (Ioannidis et al. 2005). At any stage of the experimental design process, a researcher must make numerous nuanced decisions in order to

generate valid and useful data. There are many different types of biomedical wet-lab and clinical experimental methods, each of which yields different data with a specific scope of soundly derivable conclusions. The constraints and strengths of a given experimental process strongly impacts the reliability of a study's outcome. However, optimally designing an experimental process can be extremely challenging, in part due to the many new biomedical technologies created in the last ten years, the misunderstanding of what conclusions can be drawn from certain experimental methods, and a general lack of understanding surrounding statistical methods.

Currently, there have been a few efforts to standardize protocols and build ontologies that can generate workflows for various experimental needs. Dr. Yolanda Gil at the USC Information Sciences Institute, for example, has led the development of the WINGS workflow system, a toolkit that allows users to generate custom workflows and share workflows publicly with other scientists in order to maintain consistent procedures. The same lab has also worked on the DISK ontology, a knowledge system that automates hypothesis testing by applying relevant data analytics procedures to existing datasets. In addition, a variety of guidelines exist for researchers to follow in order to generate data that is reliable and replicable. In particular, the FAIR principles outline key characteristics that data should have in order to contribute to modern scientific sharing and understanding. While these principles do not outline specific experimental workflows, they provide a standardized goal for data generated during such experiments to meet in order to be useful to - and accessible by - the broader scientific community.

Methods

Ontology

The domain of our ontology consists of wetlab techniques and tools that are most relevant to research in molecular/cell biology. The ontology should capture knowledge of experimental design that can be utilized in a typical biological lab, such as that found in a university setting; standard techniques will be captured, but not necessarily new state-of-the-art techniques or

extremely costly techniques that do no often appear in the literature. It should also be noted that the ontology is meant as a tool to researchers with a well-formed hypothesis, which assumes that the researcher has already designed the most preliminary aspects of their experiment, such as what cell lines or animal model they intend to use. Thus, domain knowledge does not include categorical knowledge of things like animal models, cell lines, fluorescent probes, etc. Rather, the scope of the ontology encompasses knowledge about what can be done to pre-determined experimental species to test a hypothesis. More specifically, domain knowledge includes the names of standard experimental methods, what type of species is required for each method (protein, DNA, RNA, etc.), what types of data the methods produce, and what types of data can be used to support specific conclusions.

The ontology itself will be designed top-to-bottom. Top Concepts include general concepts that describe what species a researcher intends to utilize, and Bottom Concepts describe specific wet lab methods. Refer to Table 1 for details.

Description of Concepts:	Examples of Concepts:
➢ Thing ➢ Biology Experiment ➢ What you're working with ➢ More specifically… ➢ What needs to be done ➢ More specifically… ➢ Specific Methods	➢ Thing ➢ Biology Experiment ➢ Nucleic Acid ➢ DNA ➢ Purification ➢ Specific Purification ➢ SCODA DNA purification

Table 1: Example organization within ontology

It should be noted that classes will not be disjoint, as many specific methods can be utilized for a variety of purposes (purification and detection, for example).

Relationships will consist of necessary links that tie these concepts together such that they form an experimental process. As we have defined it, an experimental process involves a series of tests (experimental methods) to be performed on an experimental species to accomplish a particular goal. The relationships should capture the links of this definition. For example, the

ontology will have the following relationships: [Method 1, i.e. Bradford Assay] *testsFor* [Goal, i.e. Nonspecific Detection]; [Method 2, i.e. PCR] *usedOn* [Species, i.e DNA].
Axioms will describe information about concepts that is necessary to form logical entailments about what types of conclusions can be inferred from what type of experiments. For example, Specific Methods will contain 2 boolean attributes, “showsNecessity” and “showsSpecificity”, that indicate whether or not a given method can be used to show biological necessity and/or specificity. Python rules will use this information in order to accomplish our ultimate goal of inferring what specific experimental methods should be used to test a hypothesis in question.

We performed a literature review to create competency questions, which mainly consist of hypotheses tested in a range of molecular/cell biology papers that we built up into more general questions and assertions. We also referred to Wikepedia’s catalogue of Biology Techniques and Tools to help us define the scope of our ontology, and specify what concepts our ontology should include.

Problem-Solving Methods

As stated above, we plan to build the ontology in Protege and thoroughly model the relationships and attributes. From this OWL file, we have the ability to export it as an XML file or a CSV file. We then plan on building a Python program that interacts and interprets this file to read the class hierarchy and properties and parses this into a data structure. The program will run an algorithm in order to reason over the ontology. The user will be guided through a series of questions about the requirements of the experiments and results.

Example script (not an actual output of program):

```
Program: Will you be working with proteins? [Y/N]
User: N
Program: In what time frame do you want the results? [x days/x months/x years]
User: 1 day
…
Program: You should do: immunofluorescence assay
```

Based on the inputted experimental and data-collection requirements, the algorithm we write will make decisions about what type of experiment fits best with the inputted information by the user. We have two options for building this algorithm. One option is to use forward chaining in order to reason over the requirements (Russell & Norvig 2009). The program will continue to ask questions until it has narrowed down the experiment type to a single guess, or returns unknown if none of the requirements are able to be met by the ontology. The other option is to use backward chaining; starting from the researcher's hypothesis, we work backwards to define the data needed to support each step in the scientific reasoning that could lead to the hypothesis being supported or refuted; this type of PSM is used in automated theorem-provers (Russell & Norvig 2009). We will use the problem-solving method that is the most optimal for our project, depending on the underlying structure of the ontology and ease of implementation, components of which remain to be determined. Alternatively we may combine both methods and use opportunistic reasoning (Simina et al. 1995) if it offers either a computational or implementation advantage. (We would appreciate feedback from the TAs on this point.)

If time permits, we will build a GUI overlay that will facilitate interaction with and visualization of our ontology. We will use Flask or Tkinter to build a web application that will add a graphical interactivity to the program. A user could visualize the entire ontology, or could input particular terms and characteristics to see only the concepts and relationships they are interested in. Finally, they could click to examine any particular node in the ontology for more information.

Evaluation

We plan to use two different approaches to measure the performance of our project, and combine the results with different weighting for each.

Our first and more heavily weighted method of evaluation will be presenting our system to experts (researchers, research assistants, doctoral fellows, etc.). They will use our tool to evaluate

hypotheses and experiments which they have previously completed in the lab, and were successful in obtaining valid results. We will then measure success (i.e. a "good" result) if the output of the program matches the type of experiment and structure which the testing researcher had used. We define a "bad" result as when the output of the program does not match the experimental structure and design choices which the user needs. This method of evaluation will be more heavily weighted because this method most closely mimics what our project is intended to accomplish.

Our second and more weakly weighted method of evaluation will be using published research studies with validated and invalidated findings to test our project. We will collect a variety of such papers from various different domains within the scope of our ontology, identify the hypothesis being tested, and use that hypothesis with our tool to develop a study design or structure. We will then compare those results to the actual methods used in the paper to determine if our project produces either identical, antithetical, or analogous results in terms of producing a valid or invalid study.

The advantage of this second method is we may test our tool similarly to a classification algorithm (machine learning) - we can use the number of correctly matched or "classified" studies to build a confusion matrix. From the confusion matrix, we can calculate our project's specificity and sensitivity, or provide analytics such as an ROC or precision-recall curve to evaluate our project quantitatively. Since this method does not closely mirror the project's intended use, we will weight it less in the final evaluation. However, the project team would like to posit that this could be a valid use for our proposed system, and that such a system could be used to re-evaluate previously published studies should it be deemed a reliable tool.

Results

Results to be reported in final write-up.

Discussion and Future work

Discussion to be included in final write-up.

References

- PLOS One article about why published research findings are false by John Ioannidis. http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124
- Atlantic article publicizing John Ioannidis's research in PLOS One. https://www.theatlantic.com/magazine/archive/2010/11/lies-damned-lies-and-medical-science/308269/
- Semantic Workflows: https://www.isi.edu/~gil/research/workflows.html
- DISK hypotheses ontology http://disk-project.org/
- Princeton Review MCAT Subject Review
- FAIR data principles: https://www.force11.org/group/fairgroup/fairprinciples
- Hayes-Roth, Frederick; Donald Waterman; Douglas Lenat (1983). Building Expert Systems. Addison-Wesley. ISBN 0-201-10686-8
- Marin D. Simina et al. "Opportunistic Reasoning: A Design Perspective" in Proceedings of the Seventeenth Annual Conference of Cognitive Science edited by Johanna D. Moore, 1995 ISBN 0-8058-2159-7, page 78
- Russell & Norvig 2009, p. 337, https://www.ics.uci.edu/~rickl/courses/cs-171/aima-resources/Artificial%20Intelligence%20A%20Modern%20Approach%20(3rd%20Edition).pdf

Division of Labor

Gaby Steiner is managing the biomedical research component, contributing to ontology design and will also be a poster contributor. Eric Cramer will be the Data Wrangler, contributing to the ontology design and management and will also be a poster contributor. Julia Daniel will

contribute to the computer science portion: ontology management and front-end tool design, and will also be a poster contributor. Akhila Moturu will contribute to the computer science portion: ontology management and front-end tool design, and will also be a poster contributor and presenter.