Augmenting Metadata with Models of Experimental Methods: Filling in the Gaps

Scott Colby and Mark Musen
Stanford University, Stanford, CA, USA

Introduction

Two of the fundamental requirements of scientific research are that findings be **repeatable** by particular group or investigator and that they are also **reproducible** by independent investigators following the same procedures.

In a *Nature* survey, 90% of respondents agree that there is a significant or slight "crisis" of reproducibility and more than 60% reported **failing to reproduce** at least one experiment.

Methods are stored and communicated in a variety of forms, each with its own level of granularity. We have identified **three distinct levels of abstraction** of these representations.

Future Directions

Continued work in the **identification of important motifs** and steps in protocols as they are written at all three levels is necessary. The examination of a large number of laboratory notebooks, journal article methods sections, and abstracts will facilitate this process.

Our data model will be **integrated** with existing metadata repositories, including CEDAR, to enhance the **searchability and comparability** of data sets online. Methods for the automatic performance of experiments will also be investigated.

Results

Enzyme-catalyzed expressed protein ligation, methods excerpt:

The Abstract

"Expressed protein ligation is a valuable method for protein semisynthesis that involves the reaction of recombinant protein C-terminal thioesters with N-terminal cysteine (N-Cys)-containing peptides, but the requirement of a Cys residue at the ligation junction can limit the utility of this method. Here we employ subtiligase variants to efficiently ligate Cys-free peptides to protein thioesters. Using this method, we have more accurately determined the effect of C-terminal phosphorylation on the tumor suppressor protein PTEN."

The Methods Section

"4. Immobilize the fusion protein fusion on chitin resin and wash to remove impurities."

The Lab Notebook

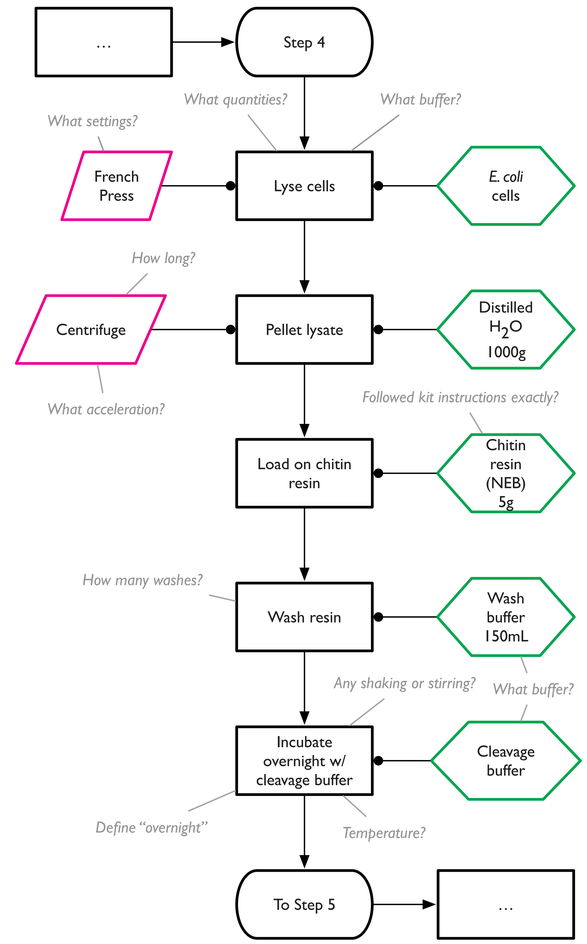
"*E. coli* cells were lysed by French press, the lysate was pelleted (17,500 × g, 15 min, 4 °C), and the supernatant was loaded onto 5 ml of chitin resin (NEB). Resin was washed with 150 ml wash buffer (250 mM NaCl, 25 mM HEPES, 0.1% Triton X-100, pH 7.5) then incubated overnight in cleavage buffer (250 mM NaCl, 50 mM HEPES, 300 mM MESNA, pH 7.5)."

What's missing at each level?

In the **abstract,** only a very basic description of the method is provided in one sentence. Even a domain expert would not be able to reconstruct the experiment from this information alone.

In the **methods section,** the experiment is enumerated as a list of steps. Take step 4: investigators familiar with chitin resin affinity purification could probably carry out this portion of the experiment from just this view of the method, but their selection of buffer, centrifugation acceleration, or other details would likely differ in some respects from the experiment that the publication describes.

In the **supplementary information/lab notebook section,** the most detailed description of the experiment is provided. Even here, however, some domain knowledge is assumed. Cell lysation by French press is a multi-step technique that has been collapsed into a single clause. One must also know that "NEB" refers to a supplier of chitin resin affinity purification kits and must find the kit's specified protocol to learn what exactly took place. The buffer preparations are also only described in a shorthand fashion.



Methods section view flowchart for step 4

References

Baker, M. *Nature.* 2016. **533,** 452–454. https://doi.org/10.1038/533452a

Chalk, S. *J Chem Inform.* 2016. **8.** https://doi.org/10.1186/s13321-016-0168-9

King, R. D.; Whelan, K. E.; Jones, F. M.; *et al. Nature.* 2003. **427,** 247–252. https://doi.org/10.1038/nature02236

Musen, M. A.; Bean, C. A.; Cheung, K.; *et al. J Am Med Inform Assoc.* 2015. **22,** 1148-52. https://doi.org/10.1093/jamia/ocv048

Henager, S. H.; Chu, N.; Chen, Z.; *et al. Nat Meth.* 2016. **13,** 925-927. https://doi.org/doi:10.1038/nmeth.4004

Acknowledgments

We would like to thank Prof. Vijay Pande for invaluable advice on the directions to take and all of the Musen group. Further thanks goes to all the Stanford students who provided the primary data for this research through access to their laboratory notebooks. CEDAR is supported by grant U54 AI117925 awarded by the National Institute of Allergy and Infectious Diseases through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov).