Proposal: Materials and Workshops for Cyberinfrastructure Education in Biology

Proposal for supplemental funding; C. Titus Brown, representing BEACON NSF STC. July 2012

Introduction

As science becomes increasingly computational, the education gap between the skills that most biologists *have* and the skills that they *need* continues to widen. Unlike other natural sciences like physics and chemistry, in biology there is little undergraduate or graduate training in mathematics or computation, and there is a general lack of computational culture in most of the biological sciences. Moreover, this is a chicken-and-egg kind of problem: there is also a general lack of *trainers* and *material* in computational biology.

At a recent NSF-sponsored BIO meeting on Cyberinfrastructure (CI) attended by a number of BIO centers, the need for improved educational materials and practices was highlighted as one of the central challenges in making use of advanced cyberinfrastructure within the biological sciences. Three basic educational topic areas emerged: first, basic computational practice, ranging from effective computer use to advanced topics such as reusing code and highly concurrent programming; second, data use, archiving and reuse, together with effective metadata generation, storage, and analysis; and third, strategies for understanding how biologists can best learn and be taught computation, including development, delivery, and assessment of training materials. (These areas arose from the aggregated list of CI challenges presented by all of the centers; see Appendix 2 for a full list.) Several centers indicated that this kind of education was their *second* biggest challenge, after their primary mission.

In response to this need, several centers, including BEACON, NESCent, and iPlant, have embarked on training programs. BEACON provides integrated interdisciplinary training as part of its first-year graduate courses, and also offers an advanced summer course on next-gen sequence analysis; NESCent regularly runs "hackathons", training workshops, and meetings to help move computational biology forward; and iPlant offers 40-50 workshops a year to help teach members of the plant community about data and metadata as well as effective use of iPlant CI. However, these courses and offerings are tremendously oversubscribed: for example, both the BEACON and NESCent next-gen sequence analysis courses received nearly 250 applicants for their 2012 courses, despite having fewer than 50 total student slots available. The most telling comment came from iPlant, who pointed out that their workshops were very frequent yet still insufficient and oversubscribed: their attendees wanted broader and deeper content, as well as asynchronous options for learning, e.g. online materials and lectures.

We therefore propose a collaborative CI project to (1) extend existing online computational science training material to facilitate self-learning by biologists across a

wide range of expertise; (2) run a number of focused workshops to teach the materials and train others in delivery; (3) develop reusable assessment strategies to study the effect of these materials on learning and help identify unmet learning needs; and (4) host several meetings across a number of centers to develop a list of shared educational needs.

1. Extend existing training material

There are several existing sets of online training material:

First, the Software Carpentry group (http://software-carpentry.org/) has been focused on training scientists in existing computational techniques for well over a decade, and provides many beginning- and mid-level static and video tutorials suitable for scientists who want to become more proficient in scripting, command-line, and other basic skills. Much of the material was developed by Dr. Greg Wilson, who has many years of industrial and programming experience; all of the material is under a Creative Commons license and is hence reusable. The material underpins the twenty one workshops that Dr. Wilson has run over the last year, with seed funding from the Sloan Foundation. The material is not biology focused and has been used to teach scientists across the physical, biological, and social sciences.

Second, the ANGUS site (http://ged.msu.edu/angus/) was developed for Dr. C. Titus Brown's summer workshop on Analyzing Next-Generation Sequencing Data, now in its third year (and NIH funded in 2012 and 2013). The tutorials are designed for biologists that are highly motivated but have little or no background in anything computational, and are primarily focused on supporting presentations during the workshop. Regardless, the site receives about 60,000 visits a year already. The ANGUS materials are also available under a Creative Commons license.

iPlant and NESCent collectively run dozens of training sessions a year, with all of their materials available online. iPlant has served over 500 scientists in their training sessions so far, and integrates materials from all available online resources, including Software Carpentry.

The current materials lack in several areas: first, there is no gentle introduction to common computational approaches such as scripting, for people with no prior programming background; and second, there is almost no material in key advanced areas including testing and the use of version control for collaboration. The Software Carpentry materials are designed for slightly more advanced users than basic bench biologists, while the ANGUS materials are focused almost entirely on cloud computing. In addition, neither set of materials offers in-depth practical tutorials on more advanced subjects

We propose to remedy these key deficiencies by:

a. Updating the existing Software Carpentry material to include both more basic and more advanced material.

b. Expanding this material to cover modern distributed version control systems such as Git and Mercurial, which are more appropriate for most scientific computing projects than Subversion (which is currently taught).
c. Use the IPython Notebook, an "executable notebook" package from Berkeley, to integrate tutorials, demonstrations, and exploration.
d. Create an "instructor's guide" that explains the pedagogical principles and techniques used in the course; this has been requested by instructors, who recognize that they have little or no formal training in teaching.
e. Include teaching guides and "dependency diagrams" that lay out which modules and topics are prerequisite for others.
f. Develop and deliver weekly online tutorials on relevant topics using a combination of Web video and real-time text collaboration tools (such as Etherpad) to deepen and broaden learner's understanding of key topics, strengthen community ties, and develop the next generation of instructors.

A key theme of this proposal is that we have yet to discover the optimal strategy for delivering educational materials: it is clear that such a strategy must include asynchronous materials for self-education, as well as supporting both informal and formal group learning. We will develop and use assessment materials to guide our primary material development.

2. Run a number of workshops.

A number of BIO centers have expressed interest in hosting workshops. During and after the materials development we will fund travel of trainers and TAs to interested centers, help the centers recruit students, and otherwise organize the workshops. In addition to "beta testing" the material live and performing assessments before and after the workshops, this will also help produce a network of faculty, postdocs, and students who are interested and trained in delivering the material, as well as fostering cross-center interactions. Each workshop's content will be specialized for the kinds of research in which the hosting center is engaged.

We include letters of support and collaboration from iPlant and NESCent, and have informal commitments from iDigBio and C-MORE. We anticipate no problems arranging 6-10 workshops: our experience with Software Carpentry suggests that most people are happy to host workshops if trainer travel is provided.

3. Develop assessment materials

Assessment of the materials' effectiveness for both in-person and online training is an essential component of developing high quality and properly targeted material. The Software Carpentry group has recently finished developing and implementing a set of instruments for evaluating computational training for scientists, and applying these instruments across the centers should yield substantial improvements the gains from computational training for scientists. For example, this type of evaluation has helped discover mismatches between expected and actual needs. The most recent round of

Software Carpentry tutorials focused on introducing scientists to SQL for data management, but in practice few participants found SQL to be the right match for their data management needs after the workshop; instead, spreadsheet-like operations were considered more useful. The summary results of assessment for a Software Carpentry workshop held at MSU in May are attached in Appendix 1.
The assessment was performed by StemEd LLC, who has been working with BEACON and Dr. Brown for two years.

There is also substantial opportunity in the area of assessment of online materials. While there is an increasing emphasis on building online materials for asynchronous training, there is also increasing evidence suggesting that we do not yet know how to design or develop effective online courses. By building online materials with assessment built in we can better plan future material development.

Finally, it is unclear how people already expert in one field (biology) go from novice to journeyman to expert in another field (computation). Studying this process with appropriate assessment tools may provide long-term insight into how to retool undergraduate and graduate education in biology to include more computation.

4. Run cross-center workshops to identify shared educational needs and solutions.

The NSF BIO Centers CI meeting resulted in a clear acknowledgement that the centers shared many common educational needs, with some centers better positioned than others to provide solutions. For example, iPlant, NCEAS, and NEON are working on data and metadata standards for long-term archival of many different types of data. BEACON and iPlant are both interested in training scientists to expertly use existing cyberinfrastructure, although iPlant is a provider while BEACON is primarily a user. Many scientists at NESCent and BEACON are involved in open source software development, and NESCent has an already-developed "hackathon" model for hosting workshops on developing software. BEACON is developing algorithms for next-gen sequence analysis data that C-MORE and NEON are interested in using. In many of these cases, several other centers are interested in aspects of educational materials and approaches that are already being developed; in other cases, such as metadata and data archiving standards, more discussion between the centers could result in synergy between efforts.

We will therefore fund two 2-day "working group" meetings to help centers identify shared educational needs, share solutions, and develop plans to address common gaps. These working groups will be held at SESYNC, and will include members from other BIO Centers. One output of these working groups will be reports or white papers that can be used for future planning purposes and that should apply broadly to BIO.

Additional considerations

All materials will be made available under a Creative Commons-Attribution license, together with guides on how to remix and reuse them for center-specific educational efforts.

Appendix 1 – Results of StemEd LLC evaluation of Software Carpentry workshop at MSU

BEACON engaged StemEd LLC to evaluate a Software Carpentry workshop hosted at MSU, May 7-May 9. The summary was very positive and indicated that Software Carpentry improved computational understanding and skills within the target population.

1. Scores on the Perception of Computational Ability scale were calculated for the pre-workshop survey. Overall scale scores were calculated as averages across all six items. *Prior to instruction, participants expressed either no or low ability in computational ability*, with a scale average of 1.73 ± 0.49. We note that this scale was not given post-workshop; we encourage its use as a pre-post measure of change in the future.
2. Scores on the Computational Understanding scale were calculated for both the pre- and post-workshop surveys. Results of the t-test indicate that pre- and post-workshop results are statistically different, $t(36) = 10.2$, $p < .001$, with post-workshop results higher than pre-workshop. *This indicates that participants perceived greater understanding after engagement in the workshop.*
3. Scores on the Python Coding Ability scale were calculated for both the pre- and post-workshop surveys. Results of the t-test indicate that pre- and post-workshop results are statistically different, $t(36) = 8.93$, $p < .001$, with post-workshop results higher than pre-workshop. *This indicates that participants perceived greater coding ability after engagement in the workshop.*
4. *Participants were generally very satisfied with the workshop (Table 6). On average, participants rated the workshop components as Good-Very Good.*
5. *Participants generally felt the workshop met their needs and would overwhelmingly recommend it to others.*
6. Participants were generally positive about the workshop in their open-ended comments. Suggestions for improvement include: more introductions and basic tutorials, PDF of commands diagrams, a reading list for Python and SQL, and videos of specific codes and related processes.

Sixteen participants made a wide range of suggestions for the workshop. In general, comments were positive (e.g., “Great work! You made it easy to follow a complex subject and improve my code!”; “Wonderful workshop. Learned a lot. Instructors explain things very well! Thank you.”). Several participants commented on the fact that installation problems inhibited their ability to engage in the material on the first day. Two participants felt the workshop was too general for those who know Python and SQL, although the majority of other comments related to the complexity of the workshop suggest the material was too advanced for some students. Participants were interested in receiving more introductions and basic tutorials. Suggestions for additional supporting materials included PDF of commands diagrams to clarify structure of git/version control, a reading list for Python and SQL provided before workshop, and videos of specific codes and related processes. One participant specifically indicated that the material was too advanced for novices, and would have appreciated instruction that spent "more time demonstrating how common problems that many of us face can be aided through use of

the software".

We propose to address many of these concerns within the framework of this proposal.

Appendix 2 – CI education needs

Education and development needs identified at the CI meeting include:

- development and adaptation of tools to archive data and metadata from diverse sources to enable data mining
- integration of structured and unstructured data from heterogeneous data sources
- discussion of standard methods for assessing data quality based on standard protocols
- approaches and tools for sharing data within and beyond centers prior to publication
- investigation and understanding of private and public cloud-based approaches
- tools and software for better distributed project participation and management
- visualization approaches
- online/asynchronous training materials
- modeling approaches and practice
- training in open-source-style software development practices
- development of metadata and semantic standards
- information and training on software & data reusability best practices, and minimum information standards
- assessment technologies for how biologists learn computational approaches
- experience in using, adapting, and creating workflows
- development, linkage, and assessment of cross-scale models
- distance teaching approaches and assessment
- approaches to streaming data analysis