Title (80 characters max)

Organizing Scientific Hackathons: Lessons Learned

Short Title (25 characters max)

Hackathon Lessons

Name and contact information for Project Leader, and any Co-Leaders

Arlin Stoltzfus, arlin@umd.edu, leader
Rutger Vos, rutgeraldo@gmail.com, co-leader

Project Summary (250 words max)

In the technology world, a “hackathon” (hacking + marathon) is an intensive multi-day bout of computer programming that stresses spontaneity, practical outputs, and new interactions. NESCent was the lead sponsor for 8 scientific software programming hackathons from 2006 to 2014, each with ~25 participants. The HIP (Hackathons, Interoperability, Phylogenies) working group staged 3 of these hackathons; working group members were on the organizing teams of all 8 hackathons. We propose an additional meeting to review NESCent’s use of hackathons as a form of education, outreach and technology development. We will gather and review information on 8 hackathons and their outcomes, discuss the factors leading to positive and negative outcomes, and generate guidance resources based on the lessons learned. We will disseminate this “how to” knowledge broadly and in various forms (open access publication, screencast, slideshow).

Public Summary (250 words max)

In the technology world, a “hackathon” (hacking + marathon) is an intensive multi-day bout of computer programming that stresses spontaneity, practical outputs, and new interactions. NESCent was the lead sponsor for 8 scientific software programming hackathons from 2006 to 2014. This working group, which includes hackathon organizers and participants, aims to review NESCent’s use of hackathons as a form of education, outreach and technology development. We will gather and review information on 8 hackathons and their outcomes, discuss the factors leading to positive and negative outcomes, and generate guidance resources based on the lessons learned. We will disseminate this “how to” knowledge broadly and in various forms (open access publication, screencast, slideshow).

Introduction and Goals

The aim of this project is to evaluate the outcomes of NESCent’s hackathons, so that we can share the resulting knowledge of best practices with the scientific community. A hackathon is an intensive workshop in which computer programmers who normally work separately meet up to pursue shared goals in software engineering. This idea is particularly promising for scientific programmers, who often work alone and do not have time or funding to team up with others to build shared resources. For sponsors who would like to promote work on a specific theme, a small hackathon entails spending just $30K to purchase 6-person-months of effort from talented and enthusiastic programmers, while raising the profile of the sponsor’s theme in a broad

community.

NESCent has sponsored 8 hackathons from 2006 to the present, most taking place in Durham. Each event was organized by a leadership team that decided on a theme, planned, budgeted, advertised, recruited participants, managed the event, and organized follow-up efforts. At each of these weeklong events, 25 to 30 participants (grads, post-docs, faculty) organized into a handful of teams to work on projects of their own design. Often a project takes a first stab at implementing an idea that members have been thinking about, but which has not yet been tried, e.g., implementing a database of trees as a triple-store, fetching images by species name to augment visualization of a species tree, or annotating a set of trees using a logical schema and a controlled vocabulary. Other projects augment an existing codebase with new features, or focus on interoperability between two resources.

The HIP (Hackathons, Interoperability, Phylogenies) working group has organized 3 hackathons (the third is the OpenTree hackathon taking place in September, 2014). The first two "Phylotastic" hackathons led to a highly accessed publication (Stoltzfus, et al., 2013)[1] and a collaborative proposal to NSF for 3 years of funding to develop an automated delivery system for expert knowledge of species phylogeny (Stoltzfus, Pontelli and O'Meara, pending). Before HIP existed, roughly the same group organized a hackathon focused on achieving interoperability of data resources via web services and common standards. This sustained focus on an interoperability strategy has stimulated the growth of an emerging web of interoperable resources that promises major benefits to the evolution community in the coming years. HIP includes individuals from the leadership teams of each of NESCent's 5 other hackathons, and we also have organized and participated in hackathons sponsored by other organizations. Together, we have the combined experience of co-organizing more than 8 hackathons, monitoring dozens of projects, and shepherding 3 hackathon-based manuscripts to publication.

We propose to make a deliberate effort of evaluating the impact of NESCent's hackathons. Hackathons generate software products, and other tangible outcomes such as annotations, how-to documents, draft standards, screencasts and occasionally, publications. They also have a variety of less tangible impacts, such as exposing participants to best practices in scientific programming (e.g., code versioning, open-source licensing), and building a sense of community. We have records of participants, projects, schedules, and in some cases, detailed meeting reports (e.g., https://www.nescent.org/wg_evoinfo/Hackathon_Report).

Of course, we already have drawn some lessons from our experiences. We know what kind of meeting space and furnishings we need. We know how to recruit a diverse participant pool. We

[1] Stoltzfus, A., H. Lapp, N. Matasci, H. Deus, B. Sidlauskas, C.M. Zmasek, G. Vaidya, E. Pontelli, K. Cranston, R. Vos, C.O. Webb, L.J. Harmon, M. Pirrung, B. O'Meara, M.W. Pennell, S. Mirarab, M.S. Rosenberg, J.P. Balhoff, H.M. Bik, T.A. Heath, P.E. Midford, J.W. Brown, E.J. McTavish, J. Sukumaran, M. Westneat, M.E. Alfaro, A. Steele, and G. Jordan, *Phylotastic! Making tree-of-life knowledge accessible, reusable and convenient.* BMC Bioinformatics, 2013. **14**: p. 158.

know that a team needs at least 3, but no more than 7. We know that the process by which teams self-organize on day 1 can run off the rails if not done right. We know that hackathons promote collaborations and that repeated hackathons can have a long-term impact on building cohesion and awareness, and on promulgating new technology and standards.

However, we feel there is more to learn by taking the time to gather and review information on past hackathons, by discussing best practices with each other, and by seeking feedback from colleagues who have been involved in hackathons but are not part of the phyloinformatics community that has grown up at NESCent. We have a variety of questions about projects, participants, and processes. How many software products are still available, or in use? How many prototypes led to a fuller implementation? How many participants remain active in the community? What was behind the biggest successes and failures? What helps "newbies" succeed? What was the effect of meeting logistics? How do we maximize effectiveness of the team-forming activity on day1? How do participants view the hackathon experience years after the event?

Many of these questions will not lead to clear answers. Our aim is to work through the available information, guided by a list of questions, in search of the lessons that are relatively clear. We will then take this knowledge and express it as guidance for participants, organizers, and sponsors, and disseminate it by various means, including a publication (e.g., PLoS Computational Biology "10 simple rules" series). This guidance will explain the entire process of staging a hackathon— deciding on a theme, planning, soliciting and evaluating applications, managing the event, and organizing follow-up— with a focus on critical points where our experience will be useful to others who may wish to sponsor, organize, or join a hackathon.

Proposed Activities

1. Organization and planning. Upon approval of this proposal, the PIs will recruit 6 more people committed to our goals, mostly from past hackathon organizing teams, but including at least 1 member who is not an organizer, and 1 who has experience with hackathons but has not been part of our community. We will convene this group (via videoconference) and give the newly formed group a chance to re-think the strategy in this proposal (experience indicates that this is necessary for buy-in, and to take advantage of unique perspectives and unanticipated knowledge that the members are bringing).

2. Research. Prior to the face-to-face meeting, the group will convene via videoconference to develop a short list of guiding questions for evaluating the projects, participants and processes. Each member will be assigned one of the 8 NESCent hackathons to study and will produce a tabular report.

3. Sharing of information, discussion and analysis. At the face-to-face meeting, each member will present an analysis of their assigned hackathon. After drawing initial lessons and refining the questions based on the results, members will return to studying their assigned

hackathons for clarification and further evaluation. The group will then reconvene for further discussion, with the aim of assembling a list of lessons learned. Analysis will take up about 30 % of the meeting time.

4. Composition. At the face-to-face meeting, when the group has learned lessons, attention will turn to composing materials for dissemination, including slides, a screencast, and a short manuscript for publication. Most of the meeting will be devoted to this activity. There is also a considerable amount of source data on the hackathons to be organized, much of which will be done by the PIs prior to the meeting.

5. Dissemination. The manuscript providing hackathon guidance will be submitted for publication, probably as a PLoS Computational Biology "10 Simple Rules" paper. The availability of the guidance documents, templates, slides and screencasts will be advertised widely in the phyloinformatics and bioinformatics community via email lists and social media.

Participating Fields and Partial List of Proposed Participants.

Arlin Stoltzfus (molecular evolution; phyloinformatics)
Rutger Vos (phyloinformatics; bioinformatics)
Enrico Pontelli (computer science, phyloinformatics)
To be named, 3 individuals who helped organize past NESCent hackathons
To be named, experienced hacker not part of the NESCent-associated community
To be named, hackathon participant who is not an organizer

Rationale for NESCent support

Although one argument for this project would be to make a record of NESCent's achievements for posterity, a more compelling reason is simply that the group's outputs will further NECent's aim to promote synthesis and synthesis activities in the evolutionary biology community. The group to be assembled for this meeting has a considerable amount of experience organizing scientific hackathons, certainly more than any other group in the area of evolutionary informatics. Hackathons build community and promote collaborativeness, while building software—sometimes exactly the kinds of software (and standards) that require a collaboration to build. By sharing our experience with the rest of the evolution and biodiversity communities, we will enable others to achieve these outcomes more easily.

Collaborations with other NESCent Activities

We will invite the participation of all NESCent staff who have been involved in organizing hackathons, particularly the two who have been on the organizing team for most of NESCent's hackathons.

Anticipated IT Needs

We require only a wireless network with internet connectivity, and a meeting room with a projector. Our products will all be archived and disseminated via github.

Proposed Timetable

The activities of the group will take place mainly during a 4-day meeting at NESCent in January of 2015, but there also will be pre-meeting activities (organization and preparation), and post-meeting follow-up activities. The pre-meeting activities will be conducted by individual participants, and as a group, during 3 scheduled videoconferences in the months prior to the meeting. At the face-to-face meeting, the group will begin with sharing the results of background research, and discussion. After a period spent planning and deciding, the group will begin working on specific outcomes. The resources generated at the meeting will be made available for comment, but will not be in a final, polished state. We aim to solicit feedback and finish up the outputs via individual effort in the 8 weeks following the meeting.

Outcomes

1. A project archive (as a github repository) with all resources generated for the project.

2. Electronic documents encoding
    - a guide to NESCent's hackathons and their products, with commentary and links to further information (i.e., expanding on http://informatics.nescent.org/wiki/Main_Page)
    - succinct guidance for organizers, participants and sponsors, in the form of
        - a "10 simple rules" document
        - a slideshow
        - a screencast
        - additional instructions as needed
    - templates for advertisements, applications, and schedules

3. Dissemination of information via
    - formal publication of a guidance document in a widely read scientific venue
    - broadcasting the availability of resources via web sites, email lists, and social media

Arlin Stoltzfus, Ph.D.

A. Professional preparation

INSTITUTION AND LOCATION	DEGREE	YEAR(s)	FIELD OF STUDY
Grinnell College, Iowa, USA	B.A., *c.laude*	1985	English
University of Iowa, Iowa, USA	Ph.D.	1991	Biology
Dalhousie Univ., Halifax, Canada	Post-Doctoral	1999	Molecular Evolution

B. Appointments

2006-present Associate Professor, Institute for Bioscience and Biotechnology Research (formerly CARB), University of Maryland and the National Institute of Standards and Technology, Rockville, MD

1999-2006 Assistant Professor, CARB, University of Maryland and the National Institute for Standards and Technology, Rockville, MD

C. Selected products

Five publications closely related to the proposed project

1. Gopalan, V., Qiu, W.G., Chen, M.Z., and **Stoltzfus, A**. 2006. Nexplorer: Phylogeny-based exploration of sequence family data. *Bioinformatics, 22:120-121*.
2. T. Hladish, V. Gopalan, C. L. Liang, W. G. Qiu, P. J. Yang, and **A. Stoltzfus**. 2007. Bio::NEXUS: a Perl API for the NEXUS format for comparative biological data. *BMC Bioinformatics* **8**:191.
3. F. Prosdocimi, B. Chisham, E. Pontelli, J. D. Thompson, **A. Stoltzfus**, 2009. Initial Implementation of a Comparative Data Analysis Ontology (CDAO). *Evolutionary Bioinformatics* 5: 47-66.
4. **Stoltzfus A**, O'Meara B, Whitacre J, Mounce R, Gillespie EL and others. 2012. Sharing and Re-use of Phylogenetic Trees (and associated data) to Facilitate Synthesis. *BMC Research Notes* 5: 574.
5. **Stoltzfus A**, Lapp H, Matasci N, Deus H, Sidlauskas B and others. 2013. Phylotastic! Making tree-of-life knowledge accessible, reusable and convenient. *BMC bioinformatics* **14**:158.

Five other publications

1. **Stoltzfus, A**., Spencer, D.F., Zuker, M., Logsdon, J.M., Jr., and Doolittle, W.F. 1994. Testing the exon theory of genes: the evidence from protein structure. *Science* **265:** 202-207.
2. **Stoltzfus, A**. 1999. On the possibility of constructive neutral evolution. *J Mol Evol* **49:** 169-181.
3. Qiu, W.G., Schisler, N., and **Stoltzfus, A**. 2004. The evolutionary gain of spliceosomal introns: sequence and phase preferences. *Mol Biol Evol* **21:** 1252-1263.
4. **Stoltzfus, A** and Yampolsky, Y. 2009. Climbing Mount Probable: Mutation as a cause of non-randomness in evolution. *J. Heredity* **100** (5): 637-47.
5. McCandlish, D. and **Stoltzfus, A**. 2014. Modeling evolution using the probability of fixation: history and implications. *Quart. Rev. Biol.* , in press.

Synergistic Activities

Co-leader of the NESCent HIP (Hackathons, Phylogeny, Informatics) working group, 2011 to present, focusing on the Phylotastic project to deliver tree-of-life knowledge

Co-leader of the NESCent Evolutionary Informatics working group, 2006 to 2009, focusing on improving interoperability through standards and technology (evoinfo.nescent.org).

Co-organizer of 6 NESCent hackathons from 2006 to 2014, empowering early-career

researchers with the skills and connections to improve interoperability.
Developer and project leader, Bio::NEXUS, an open-source Perl API for the NEXUS file format.
Developer and project leader, 2007 to 2012, Comparative Data Analysis Ontology (CDAO, www.evolutionaryontology.org)

D. Collaborations and Other Affiliations

Collaborators: Michael E. Alfaro (UCLA), James P. Balhoff (National Evolutionary Synthesis Center), Holly M. Bik (UC Davis), Brian O'Meara (Department of Ecology & Evolutionary Biology, Joseph W. Brown (Institute for Bioinformatics and Evolutionary Studies (IBEST), University of Idaho), Jason A Caravas (Wayne State University), Brandon Chisham (New Mexico State University), Karen Cranston (National Evolutionary Synthesis Center), Helena Deus (National University of Ireland), Emily L. Gillespie (Marshall University), Luke J. Harmon (University of Idaho), Tracy A. Heath (UC Berkeley), Mark T Holder (University of Kansas), Greg Jordan (Paperpile), Sudhir Kumar (Arizona State University), Hilmar Lapp (National Evolutionary Synthesis Center), Jim Leebens-Mack (University of Georgia), Naim Matasci (University of Arizona), David McCandlish (University of Pennsylvania), Emily Jane McTavish (University of Texas at Austin), Peter E. Midford (National Evolutionary Synthesis Center), Siavash Mirarab (University of Texas at Austin), John Moult (University of Maryland), Ross Mounce (University of Bath, UK), Ryan W. Norris (Lock Haven Univ.), Maryam Panahiazar (University of Georgia), Matthew W. Pennell (University of Idaho), Megan Pirrung (University of Colorado Denver), Enrico Pontelli (New Mexico State University), Anurag Priyam (Indian Institute of Technology Kharagpur, India), Ajith Ranabahu (Wright State University), Dan F. Rosauer (Yale), Michael S. Rosenberg (Arizona State University), Amit P. Sheth (Wright State University), Brian Sidlauskas (Oregon State University), Aaron Steele (U.C. Berkeley), Cory L. Strope (North Carolina State University), Jeet Sukumaran (University of Kansas), Gaurav Vaidya (University of Colorado Boulder), Rutger Vos (NCB Naturalis), Campbell O. Webb (Harvard), Mark Westneat (Field Museum of Natural History), Jamie Whitacre (NMNH), Xuhua Xia (University of Ottawa, Canada), Guoqin Yu (National Cancer Institute), Christian M. Zmasek (Sanford-Burnham Medical Research Institute).

Graduate and Post-Doctoral Advisors: W. Ford Doolittle (Dalhousie University, semi-retired), Roger Milkman (deceased).

Thesis Advisor and Post-graduate-Scholar Sponsor: Weigang Qiu (Assoc Prof, Hunter College, CUNY), Chengzhi Liang (CSHL), Danny DeKee (no scientific affiliation), Vivek Gopalan (Lockheed Martin NIAID Bioinformatics Support), Guoqin Yu (NCI), Ryan Norris (Asst Prof, Lock Haven).