PhyloStars: SMBE Satellite Workshop to Develop a Socially Networked Bioinformatics Community-of-practice

Proposers

- Arlin Stoltzfus (arlin@umd.edu), IBBR, National Institute of Standards and Technology, 9600 Gudelsky Drive, Rockville, MD 20850
- Mark W. Westneat (mwestneat@fieldmuseum.org), Biodiversity Synthesis Center, Field Museum of Natural History, Chicago, IL 60605
- Hilmar Lapp (hlapp@nescent.org), National Evolutionary Synthesis Center (NESCent), Durham, NC 27705

This proposal is submitted on behalf of HIP (Hackathons, Phylogenies, Interoperability), a NESCent working group led by PIs Arlin Stoltzfus, Enrico Pontelli, and Rutger Vos and 7 other team leaders associated with projects such as Encyclopedia of Life (M. Westneat), iPlant (N. Matasci), DateLife.org (B. O'Meara) and Open Tree of Life (K. Cranston). The full roster of team leaders, and links to projects, are on the working group's wiki. The working group is supported by NESCent, iPlant and the Biodiversity Synthesis Center of EoL.

Workshop summary

We propose to engage the SMBE community, particularly early-career researchers, with a participant-driven meeting aimed at (1) developing and disseminating topical bioinformatics solutions that serve recognized needs of researchers in SMBE-relevant fields and (2) developing the capacity to meet this ongoing need in a self-sustaining way, via an online, global, socially networked community-of-practice (hashtag #phylostars).

The explosion in the scale and scope of data affecting SMBE-relevant fields (e.g., molecular evolution, phylogenetics, comparative genomics, molecular population genetics) is accompanied by a parallel explosion in the use of computers to automate aspects of data analysis— including mundane aspects such as format conversions and manipulations of lists, trees, and alignments, as well as specialized aspects such as implementing statistical inference models and managing complex workflows.

The fact that this explosion is recent, ongoing, and dynamic, has distinctive implications for developing a community of practice that effectively addresses daily technical challenges. We all know the problem. A researcher may get stuck, for days or weeks at a time, or make scientific compromises to avoid a technical hurdle. When researchers— often grads and post-docs— get stuck, where do they turn? Most scientific software is poorly documented. Mentors and co-workers frequently do not have the answers. Textbooks are no help. Computer-oriented workshops (Woods Hole, Bodega Bay, etc) are fantastic, but this is not a scalable strategy to educate an entire generation of researchers. Yet scientists all over the world are facing the same challenges— sometimes it is just a matter of connecting with someone else who already solved the problem, or could solve it in a few minutes.

This is a community problem, and an appropriately dynamic community-based solution is emerging already: leveraging online social networking and coding tools to identify, discuss, and solve challenges as they arise. At stackoverflow.com, hardcore programmers post everything from inscrutable bugs to integrative challenges. Others respond, add comments, and vote up valuable answers. A search on “phylogen*” reveals dozens of questions, some with valuable answers, e.g., how to draw 2 trees back-to-back in R to compare branchings, how check if 2 tree topologies are equivalent, how to parse a phyloXML file in RapidXML and C++. While stackoverflow is restricted to coding, BioStars.org (“bioinformatics explained”) welcomes questions about how to find and use data (e.g., "Where can I download a phylogeny of flowering plants?") and how to use tools (e.g., “How does CodeML assign internal branch numbers?”). BioStars has hundreds of questions that match “phylogen*”, many unanswered. The phylogenetics community for the “R” language has a highly active r-sig-phylo mailing list.

To empower this emerging solution, we propose a community-wide process culminating in a week-long meeting that will leverage (1) the energy, knowledge and skill of the SMBE community, particularly early-career researchers (programmers and non-programmers) involved in computer-aided research, (2) web-based social coding sites like BioStars.org that support and incentivize knowledge-sharing communities, and (3) our years of experience staging hackathons (intensive collaborative programming events) and working with scientist-programmers to develop interoperability technology[1].

We will use the registered hashtag #phylostars (a Phyloinformatics Community of Practice) to coordinate information globally across the web. The general scheme is as follows:

- 20 weeks ahead, we solicit sharing of tagged Q & A via BioStars.org
- 16 weeks ahead, we assess community needs evident on BioStars.org
- 12 weeks ahead, we issue a call for participation in the kickstart meeting
- 10 weeks ahead, we rank applications and issue invitations
- we re-issue the community-wide call to share tagged Q & A via BioStars.org
- at the meeting, we encourage teams to pursue one of 3 types of goals
    - develop, document, and disseminate new solutions to existing challenges
    - consolidate existing solutions in a critical topic area (e.g., tree viz)
    - explore strategies for a sustainable SMBE informatics community of practice

We will reach the target community through email lists (e.g., evoldir), social networking (tweet, g+), and (as approved), the SMBE web site. We know how to recruit diverse groups of highly skilled early-career scientists (e.g., group photo from recent hackathon with 40% women & minorities). We will use a time-tested hackathon model for a 5-day meeting with up to 30 participants: on day 1, after brief informational talks and demonstrations, teams self-assemble; for the remaining days, teams work separately, with occasional communal sharing of progress and challenges. Rather than specify teams and goals in advance, we use a bottom-up, OpenSpace-inspired process to facilitate the spontaneous assembly of teams— a process in which ideas are pitched, critiqued, and adjusted iteratively and interactively among participants. This ensures that each team includes the right people to accomplish that team's goals. Teams are coached and monitored to ensure a rigid focus on completing **tangible** outcomes (e.g., working Open Source code, a publicly shared how-to document, a draft manuscript).

The anticipated outcomes of this community-wide process include

- **Topical solutions to dozens (hundreds?) of technical challenges.** Most participants will address actual bioinformatics challenges emanating from the SMBE community, post answers to BioStars.org, and re-share these using social networking tools. As a metric of success, we can count the number of #phylostar-tagged questions answered.
- **Education and awareness**. By becoming directly involved, all participants will become aware of social coding resources and how to use them. They will better understand how to find resources and use them to generate solutions shared within a community.
- **Networking.** All participants will have the chance to meet potential collaborators, make tech-relevant connections online, and join relevant mailing lists and social media groups.
- **A community-wide strategy.** Our planned solicitations involve the whole SMBE community in submitting #phylostar-tagged questions to BioStars and raising awareness of this approach to sharing knowledge. To consolidate our gains, we will publish a short manuscript describing the workshop and the phylostars strategy.

[1] This proposal is submitted on behalf of HIP (see "Proposers"), the NESCent working group whose main hackathon project is Phylotastic, an on-the-fly delivery system for tree-of-life knowledge. HIP is the successor of EvoInfo, the working group that staged 4 hackathons and spawned core interoperability technology including the NeXML format and the Comparative Data Analysis Ontology (CDAO).

Financial summary

We propose a 30 to 35 person meeting in Chicago at the Biodiversity Synthesis Center at the Field Museum (BioSync), with a cost-share budget of $50,000, of which $40,000 is requested from SMBE and $10,000 is committed from BioSync. The $10,000 matching funds are guaranteed from BioSynC through the end of 2013 for this purpose. Average participant costs for a 5 day meeting in Chicago (travel, lodging, food, etc) are $1,500 to $2,000, including $300 to 500 domestic airfare, $750 to $1000 lodging and $250 food and local transport. We anticipate that full support will be provided to the majority of participants. From past experience we expect 5 to 10 individuals will be able to pay some or all of their costs.