Sprints, Hackathons and Codefests as community gluons in computational biology

Steffen Möller✉, Enis Afgan, Michael Banck, Peter J. A. Cock, Matus Kalas, Laszlo Kajan, Pjotr Prins, Jacqueline Quinn, Olivier Sallou, Francesco Strozzi, Torsten Seemann, Andreas Tille, Roman Valls Guimera, Toshiaki Katayama, Brad Chapman

Universtiy of Lübeck Department of Dermatology, Lübeck, Germany

Received 31 July 2013; **Accepted** 10 September 2013; **Published** 14 October 2013

Competing interests: the authors have declared that no competing interests exist.

Abstract

Sprints, Hackathons and Codefests are all names used for informal software developer meetings, especially popular in open source communities. These meetings, along side more traditional conferences, are a vital part of the international network of interactions between software developers working in bioinformatics and computational biology, and complement purely online interactions such as project mailing lists, online chat, web forums and more recently voice and video calls. This paper lays out how the events are organised and presents an overview on their achievements.

Motivation and Objectives

The challenge for everyone is to be aware of existing implementations of a particular desired functionality and the compatibility with the local infrastructure. Strategically, it is beneficial to know other contributors to the externally maintained library, and to ensure that contributions are integrated with the remaining code in the best future-compatible way and with the least possible redundancies.

To help achieve these goals, the Bioinformatics Open Source Conference (BOSC) was established in 2000 by the Open Bioinformatics Foundation Bio* project members as an international venue for showcasing new projects and progress, and for developers world-wide to meet in person. To support team building and help communication, BOSC adopted Birds-of-a-Feather (BoF) sessions, i.e. group meetings of one-two hours.

Methods

A series of longer BioHackathons have been held since 2002 (Stajich *et al.*, 2002). This is short for “biologically motivated code hacking marathons”. This initiative evolved into the annual BioHackathons in Japan, organised every year since 2008 with Japanese and key foreign Open Source developers attending (Katayama *et al.*, 2010; Katayama *et al.*, 2011; Katayama *et al.*, 2013).

BOSC’s Codefests run as a precursor to an international conference, i.e. BOSC and ISMB, and so is more international. The Sprints take particular effort to invite bioinformaticians local to the event. The BioHackathons have been organised as an invitational event with the loose intention of encouraging the participants to collaborate towards a given theme.

The Codefest is about new developments, but also about helping legacy code to remain compatible with new file formats and/or libraries. The role of Debian Med and Bio-Linux is largely that of an observer, re-distributor, extra pairs of eyeballs during packaging (involving the recompilation) and that of a bridge while answering or forwarding reports by users and/or downstream developers. The same individuals that package for the distribution may also contribute to the packaged project itself.

A main driving force for bringing all those biological tools to a Linux distribution is to save resources by avoiding the compilation, know the versions installed to be tested with the right set of dependencies, and thus allow for more complex combinations of those tools – to create and refine biological workflows. The binaries can already be integrated (Krabbenhöft *et al.*, 2008) with Taverna (Wolstencroft *et al.*, 2013) and remote resources be added (Möller *et al.*, 2010). Bio-Linux ships with Taverna as it comes from the developers’ website. And it offers Galaxy (Goecks *et al.*, 2010), the latter packaged so nicely that it should also ship with Debian.

The original BioHackathons in 2002 and 2003 were mainly dedicated to interoperability in handling sequence data amongst the Bio* projects. BioPerl (Stajich *et al.*, 2002), BioJava (Prlić *et al.*, 2012), Biopython (Cock *et al.*, 2009), and BioRuby and BioGems (Goto *et al.*, 2010; Bonnal *et al.*, 2012) groups worked together to develop common sequence object models, APIs for the BioSQL database and Web services. This ensured fundamental bioinformatic functionality would be compatible among those four programming toolkits.

The first years of the BioHackathon meetings in Japan focused on Web services and interoperability (Katayama *et al.*, 2010; Katayama *et al.* 2011) and later moved to improving life science data integration with Semantic Web technologies (Katayama *et al.*, 2013), reflecting the perceived needs of the biomedical community to move from work flows towards integration of data resources, ontology, semantics and reasoning.

Debian Med (Möller *et al.*, 2010) and Bio-Linux (Field *et al.*, 2006) provide the necessary glue for distribution of individual tool updates back to the wider community. This is achieved by packaging and distributing the tools in the context of these larger tool repositories. Debcamp is an unconference, a place where people meet and work on specific topics, either alone or in teams. The Debcamp concept was later generalised to the Debian Sprints, weekend gatherings of a small number of individuals to address a specific technical challenge. The Debian Med Sprints, because of the heterogeneity of applications while working with many similar types of data, take the idea further. Every winter, general invites are sent to the mailing list to convene at a European coastal town in a family-run hotel, historically resulting in 20 to 25 attendees with expertise across many scientific fields. The first Sprint in 2011 achieved the admirable goal of synchronising Bio-Linux with Debian Med.

Results and Discussion

Every event held keeps a description of its progress and achievements on a dedicated web page: Codefest[1], Debian Med[2], BioHackathon[3]. Having 20+ talented and motivated individuals with shared interests together for two or more days is always special. Such events can be organised at any level with a large enough user base, like in universities, and in all regions of the world. They combine individualised training, social networking, technical contributions and help prepare scientific discoveries.

The two day Codefests and Sprints are often too short to allow every issue to be resolved or to complete forming a consensus. One commonly observes subgroups to dive deeply at one particular topic and stick to it throughout the event. This is excellent for the participants, but difficult for other contributors to synchronise and approach with their concerns. At one week long, as their name suggests, the BioHackathon events in Japan are more of a marathon than a sprint, and allow more interaction between groups - but are more expensive to organise.

Since Open Source software developers spread across the globe already collaborate by communicating online via distributed source-code repositories, mailing lists, chat and other means, the time and expense of travelling to meet up in person may seem like a waste - even case of the BOSC Codefest there is no additional travel for those already attending the main conference. However, physical meetings bring an edge to productivity, including temporarily avoiding day to day workplace duties, and the opportunity to see software and infrastructure problems from outside your local needs. Also, meeting in person temporarily solves the problems of cross time zone collaborations. This is particularly acute for contributors in Australasia communicating with Europeans or Americans, where live interactions like conference calls must be often scheduled outside normal office hours, and any conversation by E-mail can takes days. This is often inefficient, and can be a barrier for promoting international collaboration on Open Source projects when the development speed matters and intensive communication is needed in early brain storming.

Meeting physically also helps build inter-personal relationships and can motivate attendees to follow-up on issues they might not tackle otherwise. One feels a joint strength and confirmation. However, there is also a joint network of remote experts, also outside pure Bioinformatics, that one can rely on.

We feel that to further increase acceptance of the Open Source infrastructures, even though these are already accepted as a commodity, we need to find ways to further ease an adoption of the technology and pave the way for user contributions. The Debian Med Sprints have tutorials and general overviews on software. Future Codefests will likely consider including these too,

[1] http://www.open-bio.org/wiki/Codefest
[2] http://wiki.debian.org/DebianMed/Meeting
[3] http://www.biohackathon.org/

and possibly borrow an idea from the Google Code-in: small well-described yet unresolved tasks tackling real-life problems for the participants to complete within a few hours.

We can point to specific examples of software developments and bug fixes made during the developer meetings described, and in some cases meeting report publications. However, the true worth is more intangible in the form of the community itself, new and strengthened collaborations, and the spread of ideas and best practice - both scientific and for software development.

Acknowledgements

All scientific groups are thanked that support contribution to Open Source software. The events' hosting institutions and funding agencies are thanked for providing facilities and support. O'Reilly, Electric Genetics, Apple Asia, the Debian project and the employing institutions are thanked for contributing to the travel costs of the events. The authors are listed alphabetically by surname, with the exception of the first and last authors who initiated the meetings.

References

Bonnal RJ, Aerts J, *et al.* (2012) Biogem: an effective tool-based approach for scaling up open source software development in bioinformatics. *Bioinformatics* **28** (7):1035-1037. doi: 10.1093/bioinformatics/bts080

Cock PJ, Antao T, *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**(11):1422-1423. doi: 10.1093/bioinformatics/btp163

Field D, Tiwari B, *et al.* (2006) Milo: Open software for biologists: from famine to feast, *Nat Biotechnol.* **24**:801-803. doi:10.1038/nbt0706-801

Goecks J, Nekrutenko A, Taylor J and The Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* **11**(8):R86. doi:10.1186/gb-2010-11-8-r86

Goto N, Prins P, *et al.* (2010) BioRuby: bioinformatics software for the Ruby programming language. *Bioinformatics* **26** (20):2617-2619. doi: 10.1093/bioinformatics/btq475

Katayama T, Arakawa K, *et al.* (2010) The DBCLS BioHackathon: standardization and interoperability for bioinformatics web services and workflows. The DBCLS BioHackathon Consortium*. *J Biomed Semantics* **1**(1):8. doi: 10.1186/2041-1480-1-8.

Katayama T, Wilkinson MD, *et al.* (2011) The 2nd DBCLS BioHackathon: interoperable bioinformatics Web services for integrated applications. *J Biomed Semantics* **2**(2):4. doi: 10.1186/2041-1480-2-4.

Katayama T, Wilkinson MD, *et al.*, (2013) The 3rd DBCLS BioHackathon: improving life science data integration with Semantic Web technologies. *J Biomed Semantics* **4**(1):6. doi: 10.1186/2041-1480-4-6.

Krabbenhöft HN, Möller S, Bayer D. (2008) Integrating ARC grid middleware with Taverna workflows. *Bioinformatics* **24**(9): 1221-1222. doi:10.1093/bioinformatics/btn095

Möller S, Krabbenhöft HN, *et al.* (2010) Community-driven computational biology with Debian Linux. *BMC Bioinformatics* **11**(S-12): S5. doi:10.1186/1471-2105-11-S12-S5

Prlić A, Yates A, *et al.* (2012) BioJava: an open-source framework for bioinformatics in 2012. *Bioinformatics* **28**(20):2693-2695. doi: 10.1093/bioinformatics/bts494

Stajich JE, Block D, *et al.* (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.* **12**(10):1611-1618. doi: 10.1101/gr.361602

Wolstencroft K, Haines R, *et al.* (2013) The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. *Nucleic Acids Res* **41** (W1): W557-W561. doi:10.1093/nar/gkt328