The Bioinformatics Open Source Conference (BOSC) 2013

Nomi L. Harris*[1†], Peter J.A. Cock[2†], Brad A. Chapman[3], Jeremy Goecks[4], Hans-Rudolf Hotz[5] and Hilmar Lapp[6†]

[1]Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[2]The James Hutton Institute, Dundee, UK
[3]Bioinformatics Core, Harvard School of Public Health, Boston, MA, USA
[4]Emory University, Atlanta, GA, USA
[5]Friedrich Miescher Institute, Basel, Switzerland
[6]National Evolutionary Synthesis Center (NESCent), Durham, NC, USA
[†]Open Bioinformatics Foundation (OBF) Board of Directors

Abstract

The 14th annual Bioinformatics Open Source Conference (BOSC) was held in Berlin in July 2013, bringing together over 100 bioinformatics researchers, developers and users of open source software. Since its inception in 2000, BOSC has been organised as a Special Interest Group (SIG) satellite meeting preceding the large International Conference on Intelligent Systems for Molecular Biology (ISMB), which is the annual meeting of the International Society for Computational Biology (ISCB). BOSC provides bioinformatics developers with a forum for communicating the results of their latest efforts to the wider research community, and a focused environment for developers and users to interact and share ideas about standards, software development practices, and practical techniques for solving bioinformatics problems. As in previous years, BOSC 2013 was preceded by a Codefest, a two day hackathon that brings together bioinformatics open source project developers and members of the community and allows them to work collaboratively and achieve greater interoperability between tools developed by different groups.

The session topics at BOSC 2013 included several that have been popular in previous years, including Cloud and Parallel Computing, Visualization, Software Interoperability, Genome-scale Data Management, and a session for updates on ongoing open source projects, as well as two new sessions: Translational Bioinformatics, recognizing the growing use of computational biology in medical applications, and Open Science and Reproducible Research. Open Science, a movement dedicated to making all aspects of scientific knowledge production freely available for reuse and extension, not only validates published results by allowing others to reproduce them, but also accelerates the pace of scientific discovery by enabling researchers to more efficiently build on previous work, rather than having to reinvent tools and reassemble data sets.

BOSC typically features two keynote talks by researchers who are influential in some aspect of open source bioinformatics. Our first keynote talk this year was by Cameron Neylon, the Advocacy Director for the Public Library of Science (PLOS), who is a prominent advocate for open science. He discussed the cultural issues that are hindering open science, and how openness in scientific collaborations can generate impact. Our second keynote speaker, Sean Eddy, who is perhaps best known as the author of the HMMER software suite, began his keynote talk with an inspiring history of how he got involved in bioinformatics and proceeded to argue that dedicating effort to thorough engineering in tool development, which is often shunned as incremental, can become the key to creating a lasting impact.

With the increasing reliance of more and more fields of biology on computational tools to manage and analyze their data, BOSC is well positioned to stay relevant to life science, and thus life scientists, for many years to come.

[*] to whom correspondence should be addressed nlharris@lbl.gov

PeerJ PrePrints | https://peerj.com/preprints/83v1/ | v1 received: 20 Oct 2013, published: 20 Oct 2013, doi: 10.7287/peerj.preprints.83v1

1 Introduction

In July 2013, over 100 bioinformatics researchers, developers and users of open source software gathered in Berlin, Germany, to attend the 14th annual Bioinformatics Open Source Conference (BOSC, http://www.open-bio.org/wiki/BOSC_2013). Since its inception in 2000, BOSC has provided bioinformatics developers with a forum for communicating the results of their latest efforts to the wider research community, and a focused environment for developers and users to interact and share ideas about standards, software development practices, and practical techniques for solving bioinformatics problems.

BOSC began as a relatively informal meeting in 2000 at the University of California, San Diego, held in conjunction with the International Conference on Intelligent Systems for Molecular Biology (ISMB) meeting. At the time, the concept of open source itself was far from mainstream in scientific software development. BOSC was a response to the need for a regular face-to-face gathering and knowledge exchange opportunity for the nascent bioinformatics developer communities around the rapidly growing BioPerl (Stajich et al., 2002), Biopython (Cock et al., 2009) and BioJava (Holland et al., 2008) projects. In part to establish a financial sponsor of BOSC, leaders of those communities one year later incorporated the Open Bioinformatics Foundation (OBF, http://www.open-bio.org/) as an umbrella organization supporting infrastructure needs of its member projects. Since then, BOSC has been held every year as an official Special Interest Group (SIG) meeting preceding the annual ISMB conference.

Although in the time since BOSC's early years software sharing as open source code has gained wide acceptance, the size of the audience attracted by the event has remained strong. The bioinformatics projects and communities that present and meet at BOSC have expanded well beyond the founding OBF member projects, and its thematic areas of emphasis can change from year to year to respond to emerging trends and new bioinformatics challenges. Nonetheless, the overall subject of the conference, and its core open-source commitment, have remained consistent since its founding years.

2 BOSC 2013 Topics

The session topics this year included several that have been popular in previous years, including Cloud and Parallel Computing, Visualization, Software Interoperability, Genome-scale Data Management, and a session for updates on ongoing open source projects. BOSC also featured two new sessions this year: Open Science and Reproducible Research (discussed below), and Translational Bioinformatics, recognizing the growing use of computational biology in medical applications.

The dedicated Open Source Bioinformatics Project Update session provides an avenue for ongoing projects to present highlights of their latest capabilities and other community news. By necessity, many of these are bound to be incremental in nature, and providing a forum where these are nonetheless welcome has been a unique role of BOSC from the beginning. As discussed in the Panel Session, this service provides value to the community, because conference presentations are a metric for the recognition of software development efforts. This year two OBF-affiliated Bio* projects gave updates. The BioRuby (Goto et al., 2010) talk focused on the increased community interest in their new BioGem modularization system (Bonnal et al., 2012), while the Biopython talk focused on Python 2 to Python 3 transition plans.

Slides from all of the presentations are available from the BOSC website, along with some of the posters. This year, for the first time, video recordings are also available for selected talks, including the keynotes and panel discussion.

3 Keynotes

Each day of BOSC traditionally starts off with a keynote talk by a person of influence in open source bioinformatics. Our first keynote this year was by Cameron Neylon, the Advocacy Director for the Public Library of Science, who is a prominent advocate for open science. Neylon discussed the cultural issues that are hindering open science, and spoke about the potential of openness in scientific collaborations for generating impact, which he summarized in what has since been dubbed the "Neylon Equation". In this equation, the probability of work having an impact rises proportional to both the interest it generates and the number of people reached, but is inversely proportional to the "friction", or obstacles to using the work. Our second keynote speaker, Sean Eddy, a group leader at the Howard Hughes Medical Institute's Janelia Farm, who is perhaps best known as the author of the HMMER software suite, began his keynote talk with an inspiring history of how he got involved in bioinformatics. He proceeded to argue from his own experience and practices how dedicating effort to thorough engineering in tool development, which is often shunned as incremental, can become the key to creating a lasting impact.

4 Panel Discussion

To stimulate discussion on important controversial or multifaceted topics, BOSC has for several years included a panel, in which panelists representing a range of viewpoints answer questions from the audience. This year's panel was on Strategies for Funding and Maintaining Open Source Software. It was led by moderator Brad Chapman, with panelists Peter Cock (The James Hutton Institute), Sean Eddy (Janelia Farm), Carole Goble (University of Manchester), Scott Markel (Accelyrs), and Jean Peccoud (Virginia Bioinformatics Institute) together spanning a range of academic and commercial perspectives.

To secure continued funding for a software project, researchers must be able to demonstrate its impact. The panelists agreed that traditional publications, and tracking their citations, still play an important role in publicizing and demonstrating the use of one's software, but they are not the only metric for impact. Panelist Sean Eddy described the "Deletion Phenotype" to evaluate research impact: were a researcher's work to be deleted from the scientific record, would there be an observable "phenotype", i.e., impact on the field? The panel explored ways to quantify usage of one's software as a measure of impact. Social coding sites such as GitHub and BitBucket leverage distributed version control systems to record a network of contributors, a direct proxy for uptake. Similarly, mailing list or web-forum activity can reflect the vitality of a community. Metrics such as download counts or user tracking are still useful, but more critical is showing the specific impact of the work on research. Aside from tracking citation, this is primarily achieved by successful use cases, often appended to grant applications as letters of support. Panelist Carole Goble suggested that to open future funding opportunities, lobbying of government agencies is important, including advocating for the categorization of scientific software as infrastructure and, in consequence, the costs of software development as a capital expenditure.

5 Codefest

The BOSC Codefest brings together open source project developers and the community, providing the chance to work collaboratively for two days prior to the conference. Over 30 developers participated in this year's Codefest, hosted by Humboldt-Universitt zu Berlin (Möller et al., 2013). Meeting in person is a valuable complement to traditional online distributed teamwork, allowing more intensive discussions and social bonding that continues into the BOSC meeting. Projects accomplished by attendees included the extension of existing open-source tools, development of standards for provenance tracking, and integration of infrastructure management, visualization and parallelization frameworks (Chapman, 2013). A key outcome was increased interoperability between tools, an essential component of doing large scale science in rapidly evolving research areas.

6 From Open Source to Open Science

In contrast to common practice in bioinformatics during the early years of BOSC a decade ago, it has almost become a norm for cutting edge bioinformatics tools to be released under an open source licence. Some journals in the field even make it a prerequisite for publication that authors make their source code available (Prlić & Lapp, 2012). With software being increasingly important in all areas of science, the principles of open source software have served as a fundamental building block for Open Science. Open Science is a movement dedicated to making all aspects of scientific knowledge production freely available for reuse and extension, including scientific data, methods, and analyses.

In response to the increasing traction that this movement has gained, including in the life sciences, BOSC has, for the first time, included this year a session devoted explicitly to Open Science. One of the objectives of Open Science is the wider issue of making published research reproducible. Aside from openness in software licensing, this also includes openness of data, and unhindered access to scientific papers themselves (Open Access). When researchers can freely access publications and the source code and data that support them, it becomes possible for them to recreate the steps that the authors went through to reach their conclusions, and to then go beyond them. In this way, Open Science not only stands to provide the value of validating published results by recreating them, but also to accelerate the pace of scientific discovery, by enabling researchers to more effectively build on the results of previous work, rather than having to reinvent tools and reassemble data sets.

7 The Future of BOSC

With the increasing reliance of more and more fields of biology on computational tools to manage and analyze their data, BOSC seems assured to stay relevant to life science, and thus life scientists. In fact, conceptually similar events aimed at fostering community and knowledge exchange around open software and standards have sprung up in other fields of biology, such as the meetings of the Genomics Standards Consortium (GSC) (`http://gensc.org`) for the metagenomics community, and the Informatics for Evolution, Phylogenetics, and Biodiversity (iEvoBio) Conference for evolution, ecology, and biodiversity science (`http://ievobio.org`). The arrangement of holding BOSC as a SIG of the much larger ISMB meeting has had considerable benefits, both for attendees, many of whom also attend ISMB or other SIGs, as well as for the organizers of BOSC, who do not need to worry about on-the-ground and registration logistics. There may be opportunities in future years to facilitate cross-pollination with other biological informatics communities also faced with Big Data challenges by occasionally holding BOSC jointly with the meetings of those communities typically not represented among the ISMB audience.

Acknowledgements

We thank all the current and past members of the BOSC organizing and abstract review committees, the Codefest organizers and hosts, and our panelists, speakers and delegates for making every BOSC a stimulating and informative meeting. We are grateful to Eagle Genomics for repeatedly sponsoring student travel awards, and to ISCB and the ISMB meeting organizers, especially Steven Leard, for their part in making BOSC happen.

References

Bonnal, Raoul J.P., Jan Aerts, George Githinji, Naohisa Goto, Dan MacLean, Chase A. Miller, Hiroyuki Mishima, Massimiliano Pagani, Ricardo Ramirez-Gonzalez, Geert Smant, Francesco Strozzi, Rob Syme, Rutger Vos, Trevor J. Wennblom, Ben J. Woodcroft, Toshiaki Katayama & Pjotr Prins. 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.

Chapman, Brad A. 2013. Summary from Bioinformatics Open Science Codefest 2013: Tools, infrastructure, standards and visualization. `http://bcbio.wordpress.com/2013/07/18/summary-from-bioinformatics-open-science-codefest-2013-tools-infrastructure-standards-and-visualization/` (blog post).

Cock, Peter J A, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski & Michiel J. L. de Hoon. 2009. Biopython:

freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25(11). 1422–1423. doi:10.1093/bioinformatics/btp163.

Goto, Naohisa, Pjotr Prins, Mitsuteru Nakao, Raoul Bonnal, Jan Aerts & Toshiaki Katayama. 2010. BioRuby: bioinformatics software for the ruby programming language. *Bioinformatics* 26(20). 2617–2619. doi:10.1093/bioinformatics/btq475.

Holland, Richard C. G., T. A. Down, M. Pocock, A. Prlić, D. Huen, K. James, S. Foisy, A. Dräger, A. Yates, M. Heuer & M. J. Schreiber. 2008. BioJava: an open-source framework for bioinformatics. *Bioinformatics* 24(18). 2096–2097. doi:10.1093/bioinformatics/btn397.

Möller, Steffen, Enis Afgan, Michael Banck, Peter Cock, Matus Kalas, Laszlo Kajan, Pjotr Prins, Jacqueline Quinn, Olivier Sallou, Francesco Strozzi, Torsten Seemann, Andreas Tille, Roman Valls Guimera, Toshiaki Katayama & Brad Chapman. 2013. Sprints, Hackathons and Codefests as community gluons in computational biology. *EMBnet.journal* 19(B). `http://journal.embnet.org/index.php/embnetjournal/article/view/726`.

Prlić, A. & H. Lapp. 2012. The PLOS Computational Biology software section. *PLoS Computational Biology* 8(11). e1002799. doi:10.1371/journal.pcbi.1002799.

Stajich, Jason E, David Block, Kris Boulez, Steven E Brenner, Stephen A Chervitz, Chris Dagdigian, Georg Fuellen, James G R Gilbert, Ian Korf, Hilmar Lapp, Heikki Lehväslaiho, Chad Matsalla, Chris J Mungall, Brian I Osborne, Matthew R Pocock, Peter Schattner, Martin Senger, Lincoln D Stein, Elia Stupka, Mark D Wilkinson & Ewan Birney. 2002. The Bioperl toolkit: Perl modules for the life sciences. *Genome Research* 12(10). 1611–1618. doi:10.1101/gr.361602.

PeerJ PrePrints | https://peerj.com/preprints/83v1/ | v1 received: 20 Oct 2013, published: 20 Oct 2013, doi: 10.7287/peerj.preprints.83v1