

Bioinformatics Brew: the cross-platform package manager for open-source bioinformatics tools

Mark Howison
Computing & Information Services
Brown University
Providence, RI 02912 USA
mhowison@brown.edu

Aaron Shen
Center for Computation & Visualization
Brown University
Providence, RI 02912 USA
aaron_shen@brown.edu

ABSTRACT

Research across the life sciences is increasingly reliant on computational analyses that use a growing and diverse ecosystem of open-source bioinformatics packages. In fact, a review in 2012 of short-read aligners included over 70 such tools in that area alone [2], leading one bioinformatician to call for an “embargo on short read alignment tools” [5]. In an effort to help tame this wilderness, we recently developed an open-source package manager for bioinformatics software called Bioinformatics Brew (BiB).

BiB grew out of two related needs. First, it has allowed us to automate the process of rebuilding the entire bioinformatics software environment on our research computing cluster at Brown [4], which consists of 179 packages (270 when including the multiple versions installed for each package). We have also used BiB to deploy bioinformatics images for use in the cloud on Amazon Web Services, and for installing dependencies for the web-based bioinformatics platform Galaxy [3]. Second, BiB makes it easier for us to distribute our own open-source phylogenomics pipeline, Agalma [1]. Agalma calls 15 open-source bioinformatics tools across 52 pipeline stages, and some stages require specific versions of tools. By specifying an exact software environment, we greatly reduce the possibility of users inadvertently missing or using an incorrect version of a dependency.

In addition to meeting these two needs, we designed BiB around the following broader aims:

Accommodate the long tail of available packages. While package management is not a new problem in the history of open-source software deployment, the solutions that exist in modern Linux distributions have an unfortunate bias when it comes to bioinformatics software. They contain well-known packages like BLAST and EMBOSS, but there is a long tail of bioinformatics tools that are not in wide enough use to be curated and included in OS-level package repositories.

Encourage community participation. BiB uses a simple recipe system based on INI files that we hope makes it accessible and will lead to community participation from bioinformatics developers. BiB is hosted at BitBucket (<https://bitbucket.org/mhowison/bib>) and anyone can contribute a new or modified recipe by forking the main git repository and initiating a pull request. This approach is based on the successful HomeBrew (<http://brew.sh>) package manager for Mac OS X. BiB currently includes recipes for 49 packages covering sequence alignment and assembly, phylogenetic analysis, and quality control and storage for next-generation sequencing data.

Free up researchers to focus on more important computational tasks. Researchers already have to learn many computational skills to use open-source bioinformatics tools: navigating the command line, using regular expressions, writing scripts, and more. We should not add to this list the low-level details of how to navigate open-source

build systems like GNU autotools and cmake. The collection of recipes in BiB encapsulates the expertise of system administrators and software developers, so that this expertise does not have to be duplicated by researchers.

Make it easier to reproduce computational analyses. As analyses become more complex, it is increasingly difficult to keep track of the exact dependencies and installation steps, and to communicate them to other researchers who may want to reproduce or extend a published analysis. BiB provides an automated and reproducible way for researchers to rebuild the software environment used in an analysis. For instance, tying the deployment of Agalma to BiB allows us to summarize the installation of all dependencies with a single command.

Record the history of available packages. Because BiB recipes are updated with multiple entries for successive versions of a package, they provide a rich source of *metadata* about the available software in the field at different snapshots in time. Each recipe documents the URLs and checksums for downloadable source code and installers, and the steps needed to compile and install them.

AVAILABILITY

Bioinformatics Brew is available under a Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0) from <https://bib.bitbucket.org>.

Supported operating systems: Ubuntu 13.10, 14.04; CentOS 6.5; Mac OS X 10.9, 10.10.

REFERENCES

- [1] Dunn CW, Howison M, Zapata, F. 2013. Agalma: an automated phylogenomics workflow. *BMC Bioinformatics* 14: 330. doi:10.1186/1471-2105-14-330
- [2] Fonseca NA, Rung J, Brazma A, *et al.* 2012. Tools for mapping high-throughput sequencing data. *Bioinformatics* 28: 3169–3177. doi:10.1093/bioinformatics/bts605
- [3] Goecks J, Nekrutenko A, Taylor J, *et al.* 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology* 11: R86. doi:10.1186/gb-2010-11-8-r86
- [4] Howison M, Shen A, Loomis A. 2013. Building Software Environments for Research Computing Clusters. In *Proceedings of the 27th Large Installation System Administration Conference (LISA 13)*. <https://www.usenix.org/conference/lisa13/technical-sessions/presentation/howison>
- [5] Watson, M. December 28, 2012. An embargo on short read alignment tools. *opionomics*. Accessed on March 3, 2015. <https://biomickwatson.wordpress.com/2012/12/28/an-embargo-on-short-read-alignment-software/>