Community-driven computational biology with Debian and Taverna

S. Möller[1-3,*], H. Krabbenhöft[2], A. Tille[3], D. Paleino[3,4], A. Williams[5], K. Wolstencroft[5], C. Goble[5], C. Plessy[3]

[1]University Clinics of Schleswig-Holstein, Department of Dermatology, [2]University of Lübeck, Institute for Neuro- and Bioinformatics, Ratzeburger Allee 160, 23530 Lübeck. Germany; [3]Debian Linux Society, [4]Università degli Studi di Palermo, Dipartimento di Scienze Stomatologiche, Via del Vespro 129, 90127 Palermo, Italy, [5]University of Manchester, Oxford Road, Manchester, M13 9PL, UK

Availability: http://taverna.nordugrid.org and http://www.taverna.org.uk under the terms of the GPL, http://debian-med.alioth.debian.org

Computational biology manifests itself in many flavours. It comprises the data analysis and -management of sequences, structures, the observed and synthetical variants of the prior, static or dynamic interactions, and serves the modelling of biological processes in physiological and pathophysiological conditions. The field gained an enormous momentum over the past two decades. The information gathered today covers biological properties of many organisms and serves as a reference and general source for derived work also for neighbouring disciplines. Biologist, physicians and chemists all started using bioinformatics tools, data and models in their routine. The latest trend is to integrate the thinking of engineers and physicis, who construct compounds in silico to later prove the predicted function in the lab. The approach became known as synthetic biology and is perceived by many to allow a fluent transition towards nano-technologies. With research questions becoming increasingly complex, they demand the interaction of highly specialised disciplines. This leads to a steady increase in the number of non-redundant tools and databases that researchers need to interact with - both the computational developer and the biological users.

The dependency of the biological research community on such services will increase over the upcoming years. The strong computational demands of the services, and the sheer complexity of the research fosters the collaboration of individuals from many sites, computationally in form of grid and cloud computing, but also between computationally and biologically primed groups. To maintain the software installation consistently is barely achievable for dedicated individuals; the sharing of such across various platforms and institutional boundaries is the driving force behind the here presented work of the Debian Linux community.

Debian is an open society of enthusiasts around the globe who collaborate on packaging free software for the Linux and FreeBSD kernels. Packages are prepared by individuals and uploaded to the distribution's main servers for auto-building on today's most prominent platforms, thus rendering them available from mobiles to supercomputers and for all common processors. For complex suites or as a principle, packagers have an option to share their effort as part of a community. This process is aided by portalsauto-prepared by the infrastructure of the Debian blends. Packages invite feedback from users with the Bug Tracking System. Around 80000 users have allowed the counting of their applications via Debian's Popularity-Contest initiative. Separately counted are installations of packages that are forwarded to derived distributions. The most prominent of these is Ubuntu, for which more than 1.3 million users are reporting. Packages are described verbosely and are translated to many languages. More formally they may be selected by manual assignment of terms from a controlled vocabulary.

Technical constraints for the packaging are laid out in the Debian Policy document. Changes to it are discussed on the project's mailing lists and may be subject to voting by contributors to the distribution. The Ubuntu Linux distribution adopts the Debian packages for their own software "universe" and as such considerably contributes to the dissemination of the efforts. The computing world experiences continuous transitions, e.g. these days from 32 to 64 bit. Upcoming is an increased acceptance for energy-saving ARM- and MIPS-based operating systems of the mobile world and some special highly parallel systems. With Debian's packages being auto-built on all these different hardware platforms, one can expect continuity during such transitions, and similarly find consistent setups in the typical heterogeneous research infrastructures. This is of particu- lar benefit for distributed computations and contributes to the strong adoption of Debian and Ubuntu for cloud computing.

Packaging is most successful, i.e. up-to-date and tested, when it is derived from the packager's daily routine. For computational biology, the community now faces the challenge to scale with the steady increase in complexity: the number of contributors to the packaging needs to match the number of programs that users expect to be available. The group maintenance of applications is one such approach that seeks to lower the entry hurdle for the packaging by mutual training and the distribution of work according to expertise and interests. It also helps the integration of the software developers themselves with the community, e.g. for AutoDock and BALLView: the software developers follow the distribution's bug reports directly, and may contribute a description of their package or were invited to upload their own packages directly to the distribution's servers rather than offering them on their respective home page.

With an increasing number of packages available, the interaction between those tools becomes more and more of concern. This addresses the establishment of workflows comprising tools from many packages, but from the distribution's perspective it is also the challenge to work on the exact same version of public databases. The sharing of input between multiple applications is an ongoing work, for which many bioinformatics groups around the globe have provided solutions independently. To tap into that wealth of experiences and use it to share the effort to maintain the infrastructure is our impetus.

The distribution's software packages allows the tools included in those packages to be referenced and shared. The UseCase plugin developed as part of the EU KnowARC project extends the Taverna Workflow Workbench to take the description of such tools and include invocations of them within a Taverna workflow. The tools can be configured to run locally, or on a remote machine accessed via secure credentials such as ssh or grid certificates. Multiple invocations of a service can be achieved by the calling of the corresponding tool on a number of nodes at the same time, thus allowing faster running of the workflow over a distributed network of machines. So a workflow developer can write and test a workflow on small amounts of data locally and then by a simple change of configuration, run the workflow on a grid or cloud on much larger data sets. Workflows can include, not only tools within a packaged distribution, but also calls to other services such as WSDL operations, queries of a BioMart database or invocations of R scripts. The workflows can be uploaded to the myExperiment website and shared either publicly or with specific groups of people. The workflows can be downloaded and run, edited or included as part of a wider overall workflow. The development of workflows and the sharing of expertise via the myExperiment website based upon the creation of packaged distributions of tools, allows the collaboration of the Linux and Bioinformatics communities with great future potential.

With Taverna as a workflow engine and as a data transporter, to work locally in a most efficient manner, one also needs to have the data locally accessible - with the right indices and APIs and (especially) in the right version. For clouds, locally may now mean remote to the user's location, and it allows for the sharing of the data. The Debian community has prepared a small utility, getData, that knows how to download the most recent versions of a series of common databases, checks for the availability of a series of bioinformatics tools, and performs the respective indexing. When collaborating in clouds, the users can also ensure that any manual updates of databases are performed only once for the instant direct benefit of all other users.

To conclude, the dynamics of all three contributors, i.e. Linux distribution, Cloud infrastructure and workflow suite, are forming a symbiosis towards a readily usable infrastructure for performing and sharing biologically inspired research. The clouds bring considerable relief to smaller research groups, allowing them to think large, with the (optional) gained confidence through immediately available expert collaborators.

*Corresponding author: moeller@debian.org