BioImg.org: A Catalog of Virtual Machine Images for the Life Sciences

Martin Dahlö[1–3], Frédéric Haziza[1], Aleksi Kallio[4], Eija Korpelainen[4], Erik Bongcam-Rudloff[5] and Ola Spjuth[1–3]

[1]SNIC-UPPMAX, Department of Information Technology, Uppsala University, Uppsala, Sweden. [2]Science for Life Laboratory, Uppsala University, Uppsala, Sweden. [3]Department of Pharmaceutical Biosciences, Uppsala University, Uppsala, Sweden. [4]CSC – IT Center for Science Ltd., Espoo, Finland. [5]SLU-Global Bioinformatics Centre, Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden.

ABSTRACT: Virtualization is becoming increasingly important in bioscience, enabling assembly and provisioning of complete computer setups, including operating system, data, software, and services packaged as virtual machine images (VMIs). We present an open catalog of VMIs for the life sciences, where scientists can share information about images and optionally upload them to a server equipped with a large file system and fast Internet connection. Other scientists can then search for and download images that can be run on the local computer or in a cloud computing environment, providing easy access to bioinformatics environments. We also describe applications where VMIs aid life science research, including distributing tools and data, supporting reproducible analysis, and facilitating education. BioImg.org is freely available at: https://bioimg.org.

KEYWORDS: catalogue, virtual machine image, virtual appliance, container, software repository, cloud computing

CITATION: Dahlö et al. BioImg.org: A Catalog of Virtual Machine Images for the Life Sciences. *Bioinformatics and Biology Insights* 2015:9 125–128 doi: 10.4137/BBI.S28636.

TYPE: Technical Advance

RECEIVED: April 28, 2015. **RESUBMITTED:** June 29, 2015. **ACCEPTED FOR PUBLICATION:** July 05, 2015.

ACADEMIC EDITOR: Thomas Dandekar, Editor in Chief

PEER REVIEW: Seven peer reviewers contributed to the peer review report. Reviewers' reports totaled 1,723 words, excluding any confidential comments to the academic editor.

FUNDING: This work was supported by the Swedish strategic research program eSSENCE; EU COST Action BM1006 – Next Generation Sequencing Data Analysis Network (SeqAhead), EU FP7 AllBio [KBBE.2011.3.6-02], the SNIC-Cloud project, and ELIXIR Finland node at CSC – IT Center for Science for virtualization support.
The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: ola.spjuth@farmbio.uu.se

Introduction

The recent increase in data amounts in the life sciences, driven by technological advances in, eg, massively parallel sequencing[1] and high-throughput proteomics,[2] has made bioinformatics data analysis a bottleneck in many projects.[3] The interdisciplinary nature of projects in biomedicine requires multiple people involved in the analysis, and international consortia have been formed to produce and deposit reference data sets in large public repositories.[4] The field of bioinformatics is rapidly developing and characterized by a large variety of tools and a wide range of publicly available data,[5] and it is common in biological analyses to rely on a number of different applications and data subsets to answer scientific questions. An important task in bioinformatics is the provisioning of data and tools in a simple manner for users to locate and use it. Examples of big data and service providers include EMBL-EBI[6] (http://www.ebi.ac.uk) and NCBI[7] (http://www.ncbi.nlm.nih.gov), supporting the biological community with online access to well-maintained databases and tools. Another important task in bioinformatics is to download and set up a working environment with the necessary tools and data to be able to carry out efficient analysis, but this can be challenging as it can take time and require substantial IT expertise.

With the large number of online resources in bioinformatics, comes the necessity to organize and present them in a way that makes it easy to access them. Examples of catalog services include BioCatalogue for web services[8] (https://www.biocatalogue.org), the myExperiment repository of workflows[9] (http://www.myexperiment.org), and the MetaBase wiki-database of biological databases[10] (http://metadatabase.org).

Virtual machine images. Providing a web service means that you offer, normally for free, computing power and storage on a server connected to Internet, and it can be an undertaking to update and maintain such services over time. Further, the increasing data volumes in molecular biology, produced by, eg, massively parallel sequencing, makes it infeasible to offer such data and compute-intensive services online. The availability of applications as open source when the developer makes the source code of their programs freely available has made it popular to download and install data and tools on local computers and clusters,[11] but with complex dependencies and setups, this can be a daunting task.

A virtual machine image (VMI), also known as *cloud image*, *virtual appliance*, or *simply image*, encapsulates a complete software environment, including operating system, tools, data, and configurations. This means that they can be started

on any operating system as long as there is a compatible virtualization software installed regardless of which operating system is installed in the VMI. Scientists are able to download and run the VMI on a local computer, in a cloud environment such as Amazon EC2[12] (http://aws.amazon.com/ec2) or on a private cloud. Running a VMI instead of using a web service means that users are responsible for the computational and storage resources, but in return get full control of the system. There are not many steps to get started using VMIs. First, you need to install a suitable virtualization software like Virtual Box.[13] The next step is to download an image that is compatible with your virtualization software. The last step is to import the image in the virtualization software, and then start the virtual machine.

VMIs are becoming popular in bioinformatics and their potential for, eg, data analysis is considered to be high.[14] Examples of general VMIs in bioinformatics include Bio-linux[11] (http://environmentalomics.org/bio-linux/) and Cloud-BioLinux[15] (http://cloudbiolinux.org/) that extend a Linux distribution with a large variety of bioinformatics tools. Other examples include the CloVR virtual machine for sequence analysis[16] (http://clovr.org/), and the myChEMBL virtual machine of open data and cheminformatics tools[17] (https://github.com/chembl/mychembl).

With the increasing number of VMIs being made available by the bioinformatics community, comes the obstacle of locating VMIs for specific purposes. There are listings of VMIs for Amazon, eg, AWS Marketplace[18] (https://aws.amazon.com/marketplace), but they do not allow for other formats than Amazon Machine Images and are not targeting the life sciences.

Results

We have developed BioImg.org, an open catalog of VMIs where scientists can publish and share information about VMIs, and other scientists can search for images annotated specifically for bioinformatics. Information about VMIs in BioImg.org is structured as follows: "flavors" are the brand of the VMI, eg, BioLinux, and a flavor can have multiple "versions," which, for example, could be an updated version of the VMI, or different occasions in the case of an educational course image (see Fig. 1). A version is further divided into "groups," where the group names are decided by the uploader, eg, which kind of virtualization platform the image is made for or any other arbitrary category the uploader think fits the image best.

BioImg.org allows for uploading any VMI or container type, and each version can have multiple files attached to it, providing a flexible way to continuously update the catalog when a new version of the VMI is available. Since images usually are several gigabytes in size, uploading them through a web browser is not feasible because of problems with interrupted transfers. When uploading an image, a web-accessible URL needs to be entered so that the image can be retrieved by BioImg.org servers. A custom script for downloading the images and making them available on the site is running on a separate server where all the files are stored. VMIs can have substantial size on disk, and BioImg.org is served by a large file system (100+ terabyte) and a fast 10 Gbit/s network connection to the Swedish University backbone. The VMI upload functionality can resume or restart terminated data transfers, and image providers are encouraged to specify an MD5 or SHA1 checksum to verify the uploaded files. The URL the file was downloaded from is saved and made visible together with other information about the file. Hosting of files on BioImg servers is optional; if an image requires registration at the image providers' homepage before being available for download, there is always the option of adding the flavor or version to BioImg.org with a link to its web page.

BioImg.org, like most repositories, relies on crowd sourcing for reporting problems with the cataloged resources. When the number of images start growing, auditing all new images and keeping track of discovered exploits in the old ones will be too big a task if it is only the maintainers that take care of it. As always when using software or VMIs prepared by others, users should take proper care and test/validate the functionality before relying on them in actual scientific projects.

The site itself uses the web framework Django[19] (https://www.djangoproject.com/) to serve the pages with information about the images.

There are other sites that offer similar services as BioImg.org, eg, The CloudMarket[20] (http://thecloudmarket.com/). The main difference from BioImg.org is that the images hosted at The CloudMarket only run on Amazon EC2, ie, it is not possible to download the image and run it on your own hardware. Another difference is that there are a lot of general VMIs and no specific category for bioinformatics. Other sites like VirtualBoxImages[21] (https://virtualboximages.com) do allow download of the images, but are still too general to make it easy to find images tailored for bioinformatics.

Newer types of virtualization techniques called *containers* exist that are more lightweight than using complete virtual

Figure 1. VMIs are structured in BioImg.org as follows: (1) Flavor: the brand of the VMI. (2) Version: when a flavor is updated, a new version of the flavor is released. (3) Group: the grouping of the files within a version is free for the uploader to decide, for example, virtualization platform or another grouping that makes more sense for the specific version. An example is: *Flavor*: Chipster, *Version*: 2.12.1, *Group*: VirtualBox.

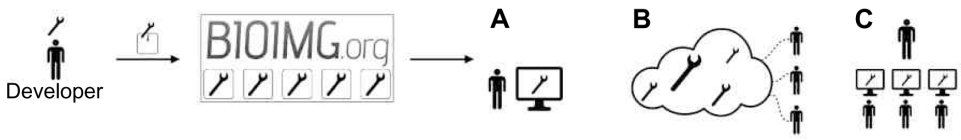

Figure 2. All the three use cases start the same way, with a developer of a tool that packages it as a VMI and uploads it to BioImg.org. (**A**) Describes a researcher who downloads the VMI and runs it on his/her local computer. (**B**) A cloud provider could make the VMI available for its users, and the tool can run in as many instances as the users require. (**C**) A teacher can make sure each student's computer environment is identical to what the laboratory instructions were created after.

images, such as Docker[22] (https://www.docker.com/) and LXD[23] (http://www.ubuntu.com/cloud/tools/lxd). The main difference is that they do not contain a complete operating system, but instead reuse many of the software components already running on the host computer, such as the Linux kernel. This means that there is almost no boot up time to start a container, and their footprint is much smaller since they make use of processes and services that are already running on the host computer instead of starting their own. The whole point of these containers is to be as portable as possible, and so it is not a problem to package them as files. These files can then be uploaded to BioImg.org in the same way as VMIs.

Applications

Distributing bioinformatics tools. Integrated bioinformatics workbenches such as Chipster[24] (http://chipster.csc.fi/) need hundreds of tools and databases to support their functionality. Virtual machines are a good, and often the only, solution for distributing these tool collections outside the original server environment. In Chipster, the bioinformatics tool collection was originally distributed as a platform specific bundle of binaries, but as virtualization technology became available, the bundle was converted into a platform-independent VMI, which resulted in a widespread adoption of the system. Currently, Chipster bundles 200 GB worth of tools and databases into a ready-to-run VMI that is integrated and tested before publishing. The tools can also be used at the command line besides the Chipster GUI.

In biomedical groups, a lot of manual effort goes into building bioinformatics environments that support all the aspects of analysis work of their respective domains. This burden could be eased by creating high-quality tools, packaging them as VMIs, and sharing them within the community. The source code of these tools might be easy to share using existing tools like GitHub[25] (https://github.com/) or Synapse[26] (https://www.synapse.org/), but getting the code to run usually involves compiling and installing dependencies, which often create problems for novice users.

Distributing bioinformatics data. The increasing size of data sets in bioinformatics in many cases necessitates downloading and carrying out analysis on local computing resources. Even though web APIs exist, the latency for millions of web transactions make it infeasible to use public resources directly. Data resources can be downloaded in their native form, such as a database dump, but in this form, there is usually a lot of work for local administrators and bioinformaticians to be able to set up and query the data. VMIs offer a solution where data and associated database software together with associated middlewares or wrappers can be packaged and delivered to users, greatly simplifying the establishment of a local mirror of a data source. An example of this is the myChEMBL virtual machine of open data and cheminformatics tools.[17] As the data sets grow larger, there is a point to not distribute them inside the VMI and instead deliver them separately. If there is an update to the operating system in the VMI, it would be better to just update that part instead of having to download the large data set once again. This is, eg, the way Chipster has divided their files.

Supporting reproducible experiments. VMIs can allow for reproducible analyses where all data, tools, and scripts can be packaged in a VMI by taking a snapshot of the system where the study was performed. This allows for easy sharing of complete experimental workflows and for other scientists to reproduce, compare, and extend analyses.[27] The ENCODE project[28] (http://www.genome.gov/encode/) is a good example of such resource sharing. Currently, the requirements for providing reproducible experiments in scientific journals are low,[29] but here VMIs can be used to facilitate the peer-review process and ultimately the published experiment.

It can be challenging to download and reproduce other scientists' analysis workflows because of, eg, incompatible computer environments, such as a local shared high-performance computing (HPC) center where users often are restricted in the way they are allowed to run programs. Owing to security concerns, HPC systems usually are equipped with firewalls, preventing the users from running any kind of web service. A cloud-based system solves this by isolating the VMs from each other, so that if one VM is compromised it does not affect the whole system. This gives the user more freedom to download VMIs and try them out.

Facilitating bioinformatics education. Anyone who has given a course in any subject that involves computers and software is aware that changes in version numbers can result in a crashing program. If a student tries to run a command from the instructions and it results in an error message, it can be hard for a beginner to figure out what went wrong. The problem in bioinformatics is that many of the new users are not familiar with command-line tools or Linux, so every problem

they encounter with the operating system steals focus from the subject being taught. One way to make sure the environment is identical to what the instructions are made for is to let every student start a VM containing everything needed for the exercise. It reduces the startup time of the exercise and makes sure everything runs smoothly, since it is only necessary to get the virtualization program up and running.

Another point that is improved by using VMs is portability. As long as the host machine has a functioning virtualization program, you can run almost any operating system on it and install any programs needed for the exercise. This greatly decreases the preparation time needed by the teacher and the risk of any last-minute surprises.

Conclusions

Many researchers in the life sciences lack in-house informatics expertise to be able to install all components necessary to run a complex workflow analysis, and VMIs containing a complete analysis system can be of great assistance. Another scenario is when scientists would like to run different types of analysis, eg, first a proteomics, and then a RNASeq analysis, they can download and switch between images without having to reconfigure their computers. The authors also have good experiences from using VMIs in bioinformatics teaching.

Virtual machines are getting increasingly popular in bioinformatics, and we envision BioImg.org to become a valuable resource for image providers and for scientists to locate images. BioImg.org is available free of charge.

Author Contributions

Planning, design and testing: MD, OS. Back-end implementation: MD. Front-end implementation and design: FH. Case studies and testing: EB, AK, EK. All authors contributed to the writing of the manuscript. All authors reviewed and approved of the final manuscript.

REFERENCES

1. Metzker ML. Sequencing technologies – the next generation. *Nat Rev Genet*. 2010;11:31–46.
2. Nilsson T, Mann M, Aebersold R, Yates JR III, Bairoch A, Bergeron JJ. Mass spectrometry in high-throughput proteomics: ready for the big time. *Nat Methods*. 2010;7:681–5.
3. Mardis ER. The $1,000 genome, the $100,000 analysis? *Genome Med*. 2010;2:84.
4. Baxevanis AD. The importance of biological databases in biological discovery. *Curr Protoc Bioinformatics*. 2011;Chapter 1:Unit1.1.
5. Zou D, Ma L, Yu J, Zhang Z. Biological databases for human research. *Genomics Proteomics Bioinformatics*. 2015;13:55–63.
6. McWilliam H, Li W, Uludag M, et al. Analysis tool web services from the EMBL-EBI. *Nucleic Acids Res*. 2013;41:W597–600.
7. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. 2014;42:D7–17.
8. Bhagat J, Tanoh F, Nzuobontane E, et al. BioCatalogue: a universal catalogue of web services for the life sciences. *Nucleic Acids Res*. 2010;38:W689–94.
9. Goble CA, Bhagat J, Aleksejevs S, et al. myExperiment: a repository and social network for the sharing of bioinformatics workflows. *Nucleic Acids Res*. 2010;38:W677–82.
10. Bolser DM, Chibon PY, Palopoli N, et al. MetaBase – the wiki-database of biological databases. *Nucleic Acids Res*. 2012;40:D1250–4.
11. Field D, Tiwari B, Booth T, et al. Open software for biologists: from famine to feast. *Nat Biotechnol*. 2006;24:801–3.
12. *Amazon EC2 – Elastic Compute Cloud*. 2015. Available at: https://aws.amazon.com/ec2/. Accessed April 09, 2015.
13. *VirtualBox*. Available at: https://www.virtualbox.org. Accessed June 26, 2015.
14. Nocq J, Celton M, Gendron P, Lemieux S, Wilhelm BT. Harnessing virtual machines to simplify next-generation DNA sequencing analysis. *Bioinformatics*. 2013;29:2075–83.
15. Krampis K, Booth T, Chapman B, et al. Cloud BioLinux: preconfigured and on-demand bioinformatics computing for the genomics community. *BMC Bioinformatics*. 2012;13:42.
16. Angiuoli SV, Matalka M, Gussman A, et al. CloVR: a virtual machine for automated and portable sequence analysis from the desktop using cloud computing. *BMC Bioinformatics*. 2011;12:356.
17. Ochoa R, Davies M, Papadatos G, Atkinson F, Overington JP. myChEMBL: a virtual machine implementation of open data and cheminformatics tools. *Bioinformatics*. 2014;30:298–300.
18. *AWS Marketplace*. Available at: https://aws.amazon.com/marketplace. Accessed June 25, 2015.
19. *Django – The Web Framework for Perfectionists with Deadlines*. Available at: https://www.djangoproject.com/. Accessed June 24, 2015.
20. *The Cloud Market*. Available at: http://thecloudmarket.com. Accessed April 09, 2015.
21. *VirtualBoxImages.com – VirtualBox Virtual Appliances*. Available at: https://virtualboximages.com/. Accessed June 26, 2015.
22. *Docker – Build, Ship and Run Any App, Anywhere*. Available at: https://www.docker.com/. Accessed April 09, 2015.
23. *LinuxContainers.org – Infrastructure for Container Projects*. Available at: https://linuxcontainers.org/. Accessed April 09, 2015.
24. Kallio MA, Tuimala JT, Hupponen T, et al. Chipster: user-friendly analysis software for microarray and other high-throughput data. *BMC Genomics*. 2011;12:507.
25. *GitHub*. Available at: https://github.com. Accessed June 26, 2015.
26. *Synapse – Contribute to the Cure*. Available at: https://www.synapse.org/. Accessed June 26, 2015.
27. Howe B. Virtual appliances, cloud computing, and reproducible research. *Comput Sci Eng*. 2012;14:36–41.
28. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489:57–74.
29. Rebooting review. *Nat Biotechnol*. 2015;33:319.