Open Source Bioinformatics Workbench Options for Life Science Researchers

Tarun Kant

Biotechnology Laboratory, FGTB Division
Arid Forest Research Institute, New Pali Road, Jodhpur 342005, India
Email: tarunkant@icfre.org

Abstract: It is unimaginable to think of a life science researcher who does not use a modern day computer as an aid to his research. In fact, modern day research cannot go on without computers playing some role in their research endeavours. The degree of this role may vary from use of embedded microchips in various intelligent equipments in use to complex data analysis. Today's researcher is surely computer savvy. With high throughput cutting-edge technology, a biologist is generating sea of data which needs high speed computing power to process and analyze it and bring out a logical interpretation. Bioinformatics tools are available today for all areas of life sciences. Most of the tools available are open source and freely available. And all these tools are now available under the umbrella of open source Linux OS platform. This review throws light on the available Linux distributions that can be used effectively as bioinformatics workbenches. [New York Science Journal 2010;3(10):82-87]. (ISSN: 1554-0200).

Abbreviations: GNU - GNU's Not Unix; **GPL** - GNU General Public License; **OS** - Operating System; **OSS/FS** - Open Source Software/Free Software; **FOSS** - Free and Open Source Software; **TCO** - Total Cost of Ownership

Keywords: Linux; bioinformatics, FOSS, sequence alignment, phylogeny

1. Introduction

Open Source Software / Free Software (OSS/FS) (also abbreviated as FLOSS or FOSS) has risen to great prominence. Briefly, OSS/FS programs are programs whose licenses give users the freedom to run the program for any purpose, to study and modify the program, and to redistribute copies of either the original or modified program (without having to pay royalties to previous developers). (Wheeler, 2007). The basic idea behind open source is very simple: When programmers can read, redistribute, and modify the source code for a piece of software, the software evolves. People improve it, people adapt it, people fix bugs. And this can happen at a speed that, if one is used to the slow pace of conventional software development, seems astonishing (OSI,2003).

The GNU (recursive acronym for GNU is Not Unix) project was born in January 1984 and over the next decade, it created a variety of critical tools that formed a portion of the operating system. The FSF was created a year later to promote Free Software and the GNU project. However, up until 1991, the GNU project had yet to produce a totally free software system due to a missing critical piece: the kernel. The kernel is the heart of the operating system. In 1991, Linus Torvalds, who at the time was a second year graduate student at the University of Helsinki, wrote and distributed a Unix-like kernel. In the manner of FOSS development, it was distributed widely, improved upon and soon adapted to become the core of the GNU/Linux operating system (Wong and Sayo, 2004).

Linux thus refers to the family of Unix-like computer operating systems using the Linux kernel. Linux has grown into a full-fledged operating system, which is known for its stability, scalability, configurability and most of all reliability for mission-critical jobs. Present day Linux is built and supported by a large international community of developers and users dedicated to Free and Open Source Software. Hundreds of Linux based operating systems are available for free today.

The popular myth surrounding Free/Open Source Software is that it is always "free"— that is, "free of charge." To a certain degree this is true. No

true FOSS application charges a licensing fee for usage. Most FOSS distributions (Red Hat, SuSE, Debian, etc.) can be obtained at no charge off the Internet. On a licensing cost basis, FOSS applications are almost always cheaper than proprietary software However, licensing costs are not the only costs of a software package or infrastructure. It is also necessary to consider personnel costs, hardware requirements, opportunity costs and training costs. Often referred to as TCO, these costs give the clearest picture of the savings from using FOSS (Wong and Sayo, 2004).

2. Linux and Bioinformatics

Most of the bioinformatics tools available are open source and developed by research and programmer community. Since Linux OS is also an outcome of community driven FOSS philosophy, the integration of various bioinformatics software tools with it becomes a natural choice. Unlike proprietary software packages which comes with a limiting license agreement and financial implications, the software (both Linux OS and most of the bioinformatics tools) under the FOSS category are available under a GNU GPL. The GNU General Public License is a free, copyleft (contrary to copyright) license for software and other kinds of works. The licenses for most software and other practical works are designed to take away users' freedom to share and change the works. By contrast, the GNU General Public License is intended to guarantee users' freedom to share and change all versions of a program- to make sure it remains free software for all its users (FSF, 2007). Hence most of the bioinformatics workbenches integrate various freely available bioinformatics tools with various Linux distributions. This proves to be a very effective strategy and very successful as well. The following paragraphs discuss various Linux flavours (better known as Linux distributions) specifically designed and developed with bioinformatics tools to become an effective bioinformatics workbench. These workbenches come as completely loaded operating system with all required software precompiled and ready to use. Most of these distributions are in the form of a Live-media (CD/DVD/Pen Drive) meaning that any compatible computer can be directly booted from these media and a user has a full fledge ready to use working environment before him. There is no need to add any more software. The computer on which the live-media is being run remains untouched and the resident OS is not altered while running the live-media. A user can however also install the complete system onto a PC side-by-side an existing OS or as a dedicated single OS for use as a Bioinformatics workstation.

3. Linux based bioinformatics workbenches

i. Bio-Linux:

In 2002, the UK's Natural Environment Research Council (NERC), in support of its Environmental Genomics Program, established the NERC Environmental Bioinformatics Centre (NEBC) (Field et al 2005). As part of its remit, NEBC developed and deployed NEBC Bio-Linux. Its user community includes bioinformaticians, software developers, system administrators and most importantly, biological researchers, many of whom were new to both Linux and bioinformatics. Bio-Linux is an ideal system for scientists handling and analysing biological data (Field at al, 2006)

Bio-Linux provides both standard and cutting edge bioinformatics tools on a Linux base. It is powerful, configurable and easy to maintain. Bio-Linux has been customised for ease of use and provides an ideal system for scientists handling and analysing biological data. Bio-Linux also comes with many development tools, providing a solid base for bioinformatics software development. Bio-Linux 6.0 is a fully featured, powerful, configurable and easy to maintain bioinformatics workstation. Bio-Linux provides more than 500 bioinformatics programs on an Ubuntu Linux 10.04 base. There is a graphical menu for bioinformatics programs, as well as easy access to the Bio-Linux bioinformatics documentation system and sample data useful for testing programs. Bio-Linux not only has a very rich suite of bioinformatics software packages, but it also provides very exhaustive documentation on the bioinformatics software as well as on the system itself.

ii. BioBrew Linux

BioBrew is a collection of open-source applications for life scientists and an in-house project at Bioinformatics.Org. The BioBrew Roll for Rocks can be used to create Rocks/BioBrew Linux, a distribution customized for both cluster and bioinformatics computing: it automates cluster installation, includes all the HPC software a cluster enthusiast needs, and contains popular bioinformatics

applications. BioBrew Linux is an open source Linux distribution based for bioinformaticists and life scientists. While it looks, feels, and operates like ordinary Red Hat Linux, BioBrew Linux includes popular cluster software e.g. MPICH, LAM-MPI, PVM, Modules, PVFS, Myrinet GM, Sun Grid Engine, gcc, Ganglia, and Globus, and popular bioinformatics software e.g. the NCBI toolkit, BLAST, mpiBLAST, HMMER, ClustalW, GROMACS, PHYLIP, WISE, FASTA, and EMBOSS. It runs on everything from notebook computers to large clusters.

Website: http://bioinformatics.org/biobrew/

iii. DNALinux

DNALinux is a is a SLAX-based Linux distribution with bioinformatics software pre-installed (Bassi and Gonzalez, 2007). On top a usual Linux programs, Bioinformatics software included are: ApE- A Plasmid Editor, Biopython, Blast, Emboss, FinchTV, NCBI Toolkit, Polyxmass, primer3 (and a working web interface, primer3plus), Rasmol, Readseq and many more. DNALinux Virtual Desktop (VD) is also available. DNALinux VD is a preconfigured virtual machine (VM) with applications targeted for bioinformatics (both DNA and protein analysis). This virtual machine runs on top of the free VMWare Player. This player can run in Windows machines since the VM works inside the player, DNALinux VD doesn't disturbe the host computer there are two separated working environment (DNALinux and Windows together on the same computer).

Website: http://www.dnalinux.com/

iv. Bioknoppix

The University of Puerto Rico High Performance Computing facility (HPCf) and the Puerto Rico Biomedical Research Infrastructure Network (BRIN-PR) have release Bioknoppix. Bioknoppix is a live CD linux, based on KNOPPIX, and specialized to include tools for bioinformatics. Bioknoppix does not need to be installed on the computer, making it a perfect tool for workshops and demos. Some of the software included in the 0.3 release are EMBOSS 2.8.0, jemboss, artemis, clustal, Cn3D, ImageJ, BioPython, Rasmol, Bioperl, Bioconductor. Bioknoppix has however been discontinued.

Website: http://bioknoppix.hpcf.upr.edu/

v. Biolinux-BR

BioLinux-BR is a project directed to the scientific community. The intention is to create a Linux distribution for people with little familiarity with the installation of the operational system and mainly for people that do not know how they must proceed to unpack a program, compile and install it correctly. For these reasons, this is a Linux system that aims to be easy to use and still offering packages that will be part of the BioLinux-BR.

Website: http://glu.df.ibilce.unesp.br/

vi. Vlinux

VLinux Bioinformatics workbench is a Linux distribution for Bioinformatics. It is easy to use, no installation required, CD-based distribution based on Knoppix 3.3. It includes a variety of sequence and structure analysis packages.It is an Open source product released under the GNU GPL License.

Website: http://bioinformatics.org/vlinux/index.html

vii. Vigyaan

Vigyaan is an electronic workbench for bioinformatics, computational biology and computational chemistry. It has been designed to meet the needs of both beginners and experts. VigyaanCD is a live Linux CD containing all the required software to boot the computer with ready to use modeling software. VigyaanCD v1.0 is based on KNOPPIX v3.7. At present the following ready to use software comes on VigyaanCD: Arka/GP, Artemis, Bioperl, BLAST (NCBI-tools), ClustalW/ClustalX, Cn3D, EMBOSS tools, Garlic, Glimmer, GROMACS, Ghemical, GNU R, Gnuplot, GIMP, ImageMagick, Jmol, MPQC, MUMer, NJPlot, Open Babel, Octave, PSI3, PyMOL, Ramachandran plot viewer, Rasmol, Raster3D, Seaview, TINKER, XDrawChem, Xmgr and Xfig. GNU C/C++/Fortran compilers and additional Linux tools (such as ps2pdf) are also available.

Website: http://www.vigyaancd.org/

viii. BioPuppy

BioPuppy is an minimal Linux OS (about 250 MB) and electronic workbench for bio-informatics and computational biology. It has been designed to meet the needs of beginners, learners, students, staffs and Research scholars. BioPuppy is available as a live CD cum installation CD [and in USB Pen drive] and containing all the required software to boot the computer with ready to use bio-informatics tools. BioPuppy is based on the Puppy Linux.

BioPuppy contains all the tools available in the basic puppy linux and bioinformatics tools. It provides both standard and cutting edge bioinformatics software tools on a Linux base. It is powerful, configurable and easy to maintain. It has been customized for ease of use and provides an ideal system for scientists handling and analyzing biological data. The Biopuppy is a light weight version of bio OS. BioPuppy included a large number of bioinformatics programs, programming libraries, in addition graphical menus for many of the bioinformatics software. BioPuppy desktop is composed of links and applications for many bioinformatics softwares. The system also includes a comprehensive, categorized and searchable documentation system for bioinformatics software.

Website: http://biopuppy.org/index.html

ix. Discovery2

Open Discovery V.1 came out in 2008 as a live Linux distribution developed with the capability to perform complex tasks like molecular modeling, docking and molecular dynamics in a swift manner. Furthermore, it is also equipped with complete sequence analysis environment and is capable of running windows executable programs in Linux environment. Open Discovery portrays the advanced customizable configuration of fedora (Vetrivel and Pilla, 2008). After the success of Open Discovery V.1 an updated version called Discovery 2 has been launched that enables parallel computing over current generation multi-core processors, bringing cluster computing speeds to the desktop. With MPI compiled GROMACS , Mr Bayes and tools like Avogadro , Pymol uses all the cores of the processor simultaneously .

Website: http://opendiscovery.org.in/

x. BioSLAX

BioSLAX is a live CD/DVD suite of bioinformatics tools that has been released by the resource team of the BioInformatics Center (BIC), National University of Singapore (NUS). Bootable from any PC, this CD/DVD runs the compressed SLACKWARE flavour of the LINUX operating system also known as SLAX.

Website: http://www.bioslax.com/

xi. BioconductorBuntu

BioconductorBuntu is a custom distribution of Ubuntu Linux that automatically installs a server-side microarray processing environment, and provides a user friendly web-based graphical user interface to many of the tools developed by the Bioconductor Project, whether locally or across a network. Installation is a turn-key procedure, simply based on booting off the installation CD or image file. In its current version, several microarrays analysis pipelines are supported including oligonucleotide (e.g. Affymetrix Genechips), dual or single dye (e.g. Exqion miRNA arrays) experiments, with the existing set of preprocessing methods for normalization, background correction and so on are easily expanded. The entire system itself is designed to be extensible, by server side integration of further relevant Bioconductor modules as required, facilitated by its straightforward pipeline construction using the underlying Python scripting environment. This makes BioconductorBuntu particularly flexible as regards the development of user-friendly processing procedures to facilitate the analysis of next-generation sequencing datasets. The system is best installed on a dedicated network server, allowing any number of registered individuals connected to the same LAN to make use of its capabilities (Geeleher et al , 2009).

Website:
http://www3.it.nuigalway.ie/agolden/bioconductor/version1/biocBuntu.iso

xii. phyLIs

PhyLIS is a free GNU/Linux distribution that is designed to provide a simple, standardized platform for phylogenetic and phyloinformatic analysis. The operating system incorporates most commonly used phylogenetic software, which has been pre-compiled and pre-configured, allowing for straightforward application of phylogenetic methods

and development of phyloinformatic pipelines in a stable Linux environment. The software is distributed as a live CD and can be installed directly or run from the CD without making changes to the computer. PhyLIS aims to simplify the process of carrying out complex phylogenetic analyses and has utility both for individual researchers and for teaching environments. The operating system presents a large suite of tools in a stable platform and should be useful for system administrators performing many installations. However, it is also simple enough to use that individual researchers with little previous Linux experience can employ it effectively. PhyLIS is under active development and undergoes periodic updates every six months to incorporate new versions of software and minor bug fixes. (Thomson, 2009).

Website: http://www.eve.ucdavis.edu/rcthomson/phylis/

xiii. Debian Med

The Debian-Med project is a custom Debian Linux distribution created to provide a co-ordinated operating system and collection of available free software packages that are well-suited for the requirements for medical practices and research. Debian-med packages have been included in Ubuntu (and derivative) operating system distribution repositories. The Debian Med project presents packages that are associated with medicine, pre-clinical research, and life science. Its developments are mostly focused on three areas for the moment which are - medical practice, imaging and bioinformatics. Over the previous years, several initiatives have spawned that address the scientific disciplines like chemistry or bioinformatics. Debian Med is not a competition to these efforts but a platform to present the packages to the community as a DebianPureBlends.

xiv. Quantian Scientific Computing Environment

Quantian is a scientific computing environment (Eddelbuettel, 2003). Quantian is a remastering of Knoppix, the self-configuring and directly bootable cdrom/dvd that turns any pc or laptop into a full-featured Linux workstation. Quantian differs from Knoppix by adding a large number of programs of interest to applied or theoretical workers in quantitative or data-driven fields. Bioinformatics tools such all packages from the BioConductor project, as well as bioperl, biopython and applications such as blast2, clustalw, ImageJ, and hmmer are included. But Quantain is not just a bioinformatics workbench. It also has R (including essentially all packages from CRAN), GSL, the GNU Scientific Library (GSL), the Grass geographic information system, TeXmacs for wysiwyg scientific editing, Cernlib, a large number of programs and libraries from the CERN particle physic lab and much more.

Website: http://dirk.eddelbuettel.com/quantian.html

4. Conclusion

Bioinformatics software bundled with GNU/Linux OS offers a better and economical way of setting up a powerful bioinformatics workbench. Linux OS today come with intuitive desktop environments which are easy to use and hardly have any learning curve. And since Linux is virtually resistant to viruses, it is surely a better choice for any life science researcher wanting to carry out bioinformatic research. The choice of these fully loaded operating systems is vast and growing every day. Thus open source bioinformatics workbench is easy to setup and can be quickly become a part of a modern biotechnology laboratory, with very low TCO.

Corresponding Author:
Dr. Tarun Kant
Scientist D & E-Champion
Biotechnology Laboratory, FGTB Division
Arid Forest Research Institute
New Pali Road, Jodhpur 342005 India
Phone: +912912729162 (O)
Email: tarunkant@icfre.org

References

1. Bassi S and Gonzalez V. DNALinux Virtual Desktop Edition. 2007; Available from Nature Precedings <http://dx.doi.org/10.1038/npre.2007.670.1>
2. Eddelbuettel D Quantian: A scientific computing environment. In Proceedings of the 3rd International Workshop on Distributed Statistical Computing DSC 2003; URL http://www.ci.tuwien.ac.at/Conferences/DSC-2003/Proceedings/Eddelbuettel.pdf. ISSN 1609-395X.
3. Field D, Tiwari B, Booth T, Houten S, Swan D, Bertrand N, Thurston M. Open software for

biologists: from famine to feast. 2006; Nature Biotechnology 24, 801 - 803
4. Field D, Tiwari B, Snape J. Bioinformatics and Data Management Support for Environmental Genomics. 2005; PLoS Biol 3(8): e297
5. FSF - "The GNU General Public License Version 3". Free Software Foundation. 2007; Retrieved July 21, 2009.
6. FSF - The Free Software Definition; http://www.fsf.org/philosophy/free-sw.html; Internet; accessed on July 20, 2010
7. Geeleher, P., Morris, D., Hinde, J.P., and Golden A., BioconductorBuntu: a Linux distribution that implements a web-based DNA microarray analysis server. 2009; Bioinformatics 25(11):1438-1439
8. OSI - Open Source Initiative; available from http://www.opensource.org; Internet; cited November 8, 2003
9. Thomson RC. PhyLIS: A Simple GNU/Linux Distribution for Phylogenetics and Phyloinformatics .2009; Evolutionary Bioinformatics 5: 91-95
10. Vetrivel U and Pilla K. Open discovery: an integrated live Linux platform of Bioinformatics tools. 2008; Bioinformation 3(4): 144-146
11. Wheeler, David, “Why OSS/FS? Look at the Numbers!”; available from http://www.dwheeler.com/oss_fs_why.html; Cited ; April 16, 2007.
12. Wong K and Sayo P. Free/Open Source Software – A General Introduction. UNDP-APDIP, Kuala Lampur, Malaysia. 2004; ISBN 983-3094-00-7

11/20/2010