Linux distributions for bioinformatics:
an update


Antonia Rana[1] and Fabrizio Foscarini

Joint Research Centre, European Commission

Introduction

The article provides an updated view on the world of Linux distributions tailored for bioinformatics analysis. The main driver for producing these distributions is to provide an easy-to-use, user friendly environment for non IT specialised users without strong requirements on the knowledge of the technology. Most commonly, intended users of these distributions are students of bioinformatics-related courses. Around 2007, quite a number of Linux distributions, which wrapped almost all available open source tools for bioinformatic analysis, appeared on the Internet. Most of them were assembled by universities and their main purpose was to use them as a tool for teaching and learning. Live CDs which did not require installation were particularly useful for this purpose. The Linux distribution around which almost all of them were built was Knoppix [3], a Linux flavour whose main characteristic was, in fact, that it was an easy customizable live distribution. In the latest two years, new technology trends have emerged in the world of Linux distributions addressed at novice users: the ability to boot from a USB flash drive (practically replacing the CD/ROM, not requiring a CD drive, providing data persistency and reusable) and the availability of distributions as Linux environments to be run as a virtual machine, in parallel with the host operating system, a feature which has the advantage of giving occasional users or students the possibility to use their usual environments while becoming familiar with a new operating system. This is reflected in the Linux distributions for bioinformatics that we have reviewed in this article. A trend that has been noticed in respect with the review we made in 2007 is the tendency to use Ubuntu as base distribution which is in fact replacing Knoppix and to provide the bioinformatics bench environment also as a virtual machine which can be run inside the popular VMWare environment in parallel with the host operating system. While reviewing the distributions in this article we have paid particular attention to their user friendliness and ease of use.

Bio-Linux

Bio-Linux [4], developed and distributed by the NERC Environmental Bioinformatics Centre, has evolved since our review in 2007, its home page has also changed. Its developers describe it as "*…a fully featured, powerful, configurable and easy to maintain bioinformatics workstation*" and in fact it is rich with applications and documentation. In its current versions, 5.0, the most notable new features are the possibility to boot it from a USB stick, as well as a LiveDVD and to install it on the hard disk. The Linux distribution on which it is based has also changed from Debian to Ubuntu. This change benefits from all the features and advantages of Ubuntu over Debian, without losing the Debian characteristics since Ubuntu is also based on Debian.

Bio-Linux provides about 500 bioinformatics programs. The complete list is available on its website. The structure and organisation of the bioinformatics programs has not changed since our last review: the bioinformatics applications are accessible via a submenu (**Bioinformatics**) of the Applications menu. Bioinformatics software is installed under `/usr/local/bioinf`. The general layout includes a directory with the base name of the package, under which a directory for each update of the software is installed. This makes it easy for users to locate packages which must be run using the command line.

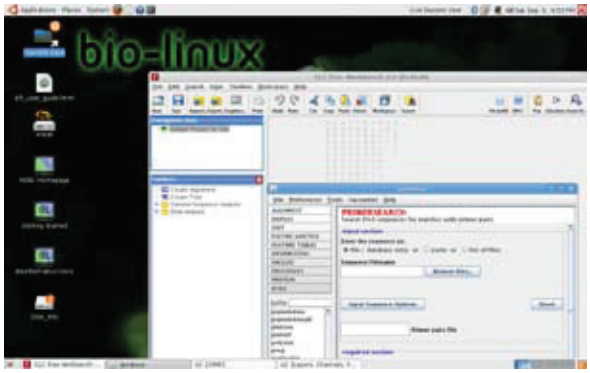
We have tested the LiveDVD version. Hardware recognition went fast without any problem or configuration required from the user. However, during the whole start-up process we did not see

1 The views expressed are purely those of the writer and may not in any circumstances be regarded as stating an official position of the European Commission.

any information about what the system is doing, we are only shown a progress bar. This can make you feel somehow uncomfortable until the whole process ends: if you use it on a PC where you have data, you want to see what is happening, make sure that the start-up process is progressing, etc.

Once started the system is fast, compatible with the need to access the DVD media when running a new program. Support for the most popular LAN, wireless and bluetooth drivers are provided. The start screen displays the icons: **Getting started** which illustrates the system, **Bioinformatics docs** for easy access to the Bio-Linux bioinformatics documentation system, **NEBC Homepage**, **Install** and **Sample data**. They are very useful to become familiar with the system and how to use it. The Install facility is also conveniently located on the desktop for easy installation on the hard drive. This is a characteristics shared also by all the other distributions which are based on Ubuntu. The security of the system is guaranteed by the installation of a personal firewall (iptables) and ssh for secure remote login.

Bio-Linux has a very rich suite of bioinformatics programs, but what is also very important is that it provides very extensive documentation on the bioinformatics software as well as on the system itself, how to install new packages, how to update it, how to install a bootable USB stick, etc. Its website is also very informative with a lot of useful information.



Download: http://nebc.nox.ac.uk/tools/bio-linux/bl_download

BioBrew and NPACI Rocks with BioRoll

The BioBrew (http://biobrew.bioinformatics.org/) distribution has not been upgraded since our last review (version 4.1.3). In fact, already in 2007 the NPACI Rocks cluster distribution with its optional BioRoll package which contains bioinformatics software looked like a candidate to its replacement. NPACI Rocks is a Linux distribution tailored for clusters and was the operating system underlying BioBrew. The main feature that distinguished BioBrew from all the other distributions was that it provided "off-the-shelf" cluster functionality. Currently, this capability can be implemented using NPACI Rocks (current version 5.2) and installing on top of it, its optional package called BioRoll which comprises a large set of bioinformatics tools.
Download: www.rocksclusters.org

DNALinux

DNALinux [5] is the distribution among those reviewed in 2007 that has, more than the others, radically changed. The most notable is that it is no longer distributed as a LiveCD/DVD but only as a virtual machine that can be run inside the VMware player on a Windows OS. The virtual machine bundles together the operating system and the bioinformatics application. Similarly to a live distribution, a virtual machine does not require installation on a dedicated computer and in addition, it can be run in parallel with the host operating system, so you can continue using your PC while running your bioinformatics application in DNALinux.

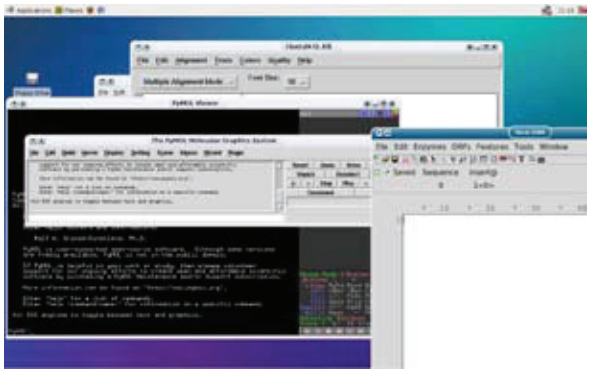
This approach shares with the live CD approach the advantage that users don't need to modify existing installations, it also shares its main disadvantage: the relative low speed of loading applications, in addition running a virtual machine implies higher memory usage.

Another difference from the DNALinux version reviewed in 2007, which was a live CD, is the Linux distribution it is based on. Slax has been replaced with Xubuntu, the light version of the popular Ubuntu linux distribution. Its authors motivate the choice of Xubuntu over Ubuntu as the first is faster thanks to the lighter desktop environment it uses.

The latest version of DNALinux is also included in the book Python for Bioinformatics[2], for this reason it is also called DNALinux Virtual Desktop Py4Bio.

DNALinux provides a large number of preinstalled bioinformatics software, the complete list is available at http://www.dnalinux.com/installed-software.html. However we have not found a menu or easy access indications for them. Some of the packages are located in the home directory of the user that is logged into the virtual machine. Some of the tools with a graphical user interface are available under the **Science** or **Education** menus.

DNALinux can be downloaded only using the bit torrent protocol. This can result in longer download times and may be not an optimal solutions for environments in which bit torrent traffic is blocked.



Download: http://www.dnalinux.com/

Open Discovery

OpenDiscovery is a new Linux distributions for bioinformatics which was not available at the time of our first review. According to its authors [6], besides providing the usual bioinformatics software (e.g. sequence analysis), OpenDiscovery has been developed with the capability to perform complex tasks like molecular modelling, docking and molecular dynamics. Like Bio-Linux, OpenDiscovery is capable of booting from USB flash drives, live DVD and can be installed on hard disk. Unlike most of the "updated" bioinformatics distributions which seem to prefer Ubuntu over knoppix, the choice which was popular in 2007, OpenDiscovery has chosen Fedora as its base distribution. Open Discovery integrates a comprehensive range of bioinformatics software. The complete list is available on its homepage.

We tested the live version of OpenDiscovery and, again with a look at user friendliness, we have noted that the startup process is very straightforward although, as for Bio-Linux, seeing only a progress bar instead of the usual list of Linux start-up operations makes us a little nervous when running a live CD. The startup is quick, though and at the end we have a Linux desktop which, without any problem, has started up our wireless network interface and plugged into the network.

The desktop is not rich: you will see the home icon, the classical **Computer** icon and the option to install the system on the hard drive. There is no *Getting started* information or shortcut to the bioinformatics application which would make it easier for novice users to become familiar with the environment. The security of the system is increased with the presence of common security tools such as: a personal firewall (iptables), and secure remote login (ssh). The firewall is configured to exclude any incoming connection.

Download: http://opendiscovery.org.in/

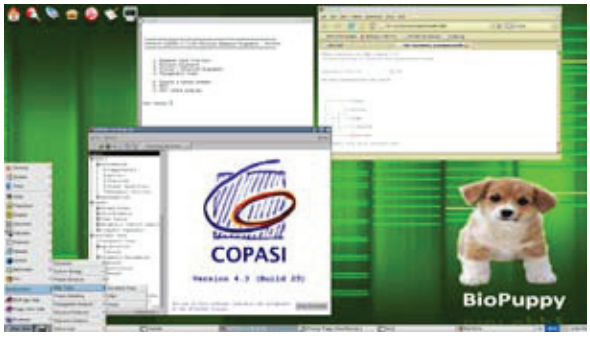
BioPuppy

BioPuppy [7] is also a newly found Linux distribution for bioinformatics which is released still in beta version. It is based on a Linux distribution known as LinuxPuppy the main advantage of which is its compactness. It contains all the tools available in the basic LinuxPuppy plus bioinformatics tools.

As for the other distributions which have been updated or new in this review, according to [7] it can be run as a live CD, from a USB stick or installed on the hard disk. We downloaded the live CD version, however testing it was not possible: the start-up process is not as straightforward

2 Sebastian Bassi, Python for Bioinformatics, Chapman & Hall

as for the other distributions, hardware recognition requires the intervention of the user who is asked to provide information about the monitor for instance. The overall process was not successful. This test was done on a notebook, as different hardware might behave differently we also tested it on a netbook (Acer AspireOne with an USB CD drive) and on a desktop PC. The netbook gave the same result, while BioPuppy started on the desktop PC. The problem seems to be related to the fact that being a very small distribution the choice of drivers available is quite limited. However, BioPuppy comes with a personal firewall installed and has its own package manager to add/delete software (.pet packages). Although it is a very compact distribution, the most popular bioinformatics tools are included (e.g. EMBOSS, HMMER, Clustal-W, Clustal-X, blast, Garlix, Phylip).



Download: http://biopuppy.org

BioSLAX

BioSLAX was just emerging when we did our first review. It is now available as a Live CD/Live DVD and bootable from a USB flash drive as well as, of course, for installation on a hard disk. It is released by the National University of Singapore and it is, in fact, an evolution of the APBioKnoppix and APBioKnoppix2 that we reviewed in 2007. BioSlax is rather different from the other two, though, being based on the SLAX linux distribution (a compressed Slackware flavour of the Linux Operating System).

According to its authors [8], SLAX was chosen over knoppix because knoppix was found not very easily expandable: in order to update an application or add a new one, a new remastering of the distribution was required. This made the distribution highly inflexible. On the other hand, SLAX works by overlaying "application modules" on top of the base Linux OS, thus making the dedicated bioinformatics distribution built on top of it, modular. Applications can be made into modules which can be inserted either dynamically or via a special folder in the BioSLAX USB/DVD distribution.

The current version of BioSLAX is 7.5 and it is based on SLAX 6. It is available for download in four formats:

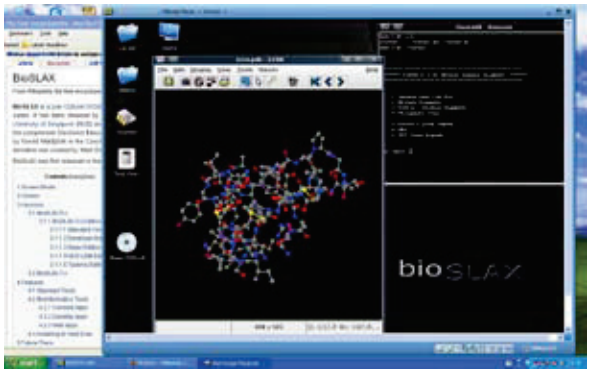
- **Power Developer DVD**: this is the full version with the complete suite of bioinformatics applications which includes also development tools (compilers and Linux kernel headers required for compilation of new applications). This comes as a Live DVD but it can be installed using the BioSLAX installer provided;
- **BioSLAX for NUS LSM courses**: This version is the full power developer version customized for students and teachers at the National University of Singapore;
- **BioSLAX for VMWare (LSM version)**: Again the same 7.5 version but created as a virtual machine for use within the virtualisation software VMWare;
- **BioSLAX with Taverna**: In this case the standard 7.5 distribution includes Taverna for workflow management.

We had some problems in testing the LiveDVD version (the over 800MB ISO file is an ISO CD format and not a DVD and could not be burned). We were interested, on the other hand, to test the virtual machine version.

The startup of the virtual machine is quite fast and it is very handy if one has to use the tools sporadically to have your stardard PC environment underlying the virtual machine environment although one needs to get used to the keyboard keys combination to get in and out of the virtual machine.

The desktop presents three icons with links to **cgi-bin**, **htdocs**, **home** and **system**. Bioinformatics tools are easily found in the dedicated **BioSLAX** menu which is conveniently further organised into five submenus: **Documentation**, **Console Apps**, **Desktop Apps**, **WebApps** and **BioSLAX Installer**. **Console Apps** provides access to all the tools that are run on the command line (e.g. blast, clustalW, EMBOSS, phylip, primer3, etc), by clicking on one of the menu items in this section, a console window is opened with the PWD set to the directory of the launched application where the executable is located. **Desktop Apps** comprises all the applications which have

a graphical user interface (e.g. ClustalX, jEMBOSS, NJPlot, Pymol, etc.) and finally **WebApps** starts web based applications (e.g. wEMBOSS). In the latter case, Firefox is started opening the launched application. It is interesting to notice that Firefox is equipped with a bookmark toolbar which provides easy and convenient access to bioinformatics-related websites such as Entrez, Bioinformatics.org, NCBI, etc.



Download: http://www.bioslax.com

BioconductorBuntu

BioconductorBuntu [9], and the distribution in the next section, are somehow different from the others we have described so far in that they have been tailored to a specific type of bioinformatics analysis: DNA microarray analysis using web-based tools. BioconductorBuntu is also a custom distribution of Ubuntu Linux. It has been created to simplify the process of setting up a microarray processing environment collecting together all the necessary analysis tools for this task in an easily installable and distributable format. The distribution is available as a live CD but, as for the other distribution based on Ubuntu, it is easily installable on the hard disk.

BioconductorBuntu provides a user friendly web-based graphical user interface to many of the tools developed by the Bioconductor Project (hence the name of the distribution). Because many of the tools it provides are accessible via a web interface, the best use of this distribution is to install it on a server and allow network access for microarray analysis. The python scripting environment underlying the Bioconductor modules facilitates the server side integration of additional modules.

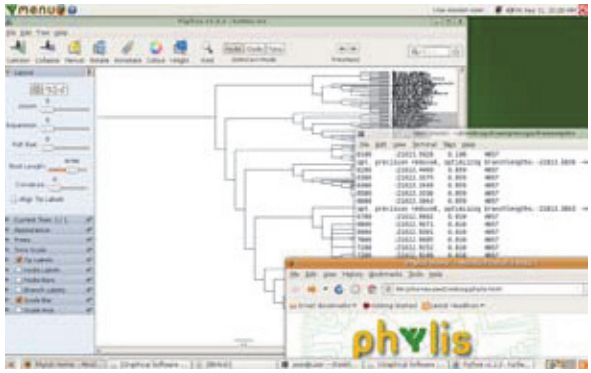
Download: http://www3.it.nuigalway.ie/agolden/bioconductor/version1/biocBuntu.iso

phyLIs

Similarly to the previous one, phyLIS (Phylogenetic Linux for Informatics and Systematics) [10] is also dedicated to a particular type of bioinformatics analysis, namely phylogenetics and phyloinformatics and contains the open source tools that are useful for this type of analysis. The distribution, first released in 2008 and based on Ubuntu, has been configured to include most commonly used phylogenetic software, it is a light distribution streamlined to focus on phyloinformatic research so that computational power is used most effectively for this purpose. Several CPU intensive programs are also available in their parallel version (MPI), so that with the proper hardware processing speed can be improved. Although it does not contain all the range of bioinformatics tools, PhyLIS contains popular scripting languages including Perl (with BioPerl), Python (with BioPython), and R.

PhyLIS is distributed as a live CD and, like all the other distributions reviewed in this article, it can be also installed on the hard disk. It is based on the Ubuntu Linux distribution.

The system is well documented both on the website and on the desktop where a folder called **Examples** contains sample files for the tools installed and the documentation file phyLIS.html, is available which provides an introduction and the list of the software installed also with the indication of its location and command name you would use to start the tool from the command line. A folder called **Graphical software** also located on the desktop, provides easy access to tools with a graphical user interface.



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

Package repositories

We did not find new package repositories, on the contrary some that were available in 2007 are no longer available on the internet. Those which are still available were updated, others have restricted their access to registered users. The status of the tools version is collectively shown for all distributions in table 1.

Debian Med

Debian Med has been updated to be included in the latest release of the Debian operating system (Lenny). Although this project is mainly dedicated to medical informatics and medical imaging, the set of bioinformatics tools included in the set of packages is increasing. In the last version, EMBOSS has been included together with one of its web interfaces, EMBOSS Explorer, which allows you to use EMBOSS either locally or on the network. All the major free programs for multiple sequence alignment and structural bioinformatics have also been included. All the programs for sequence analysis and bioinformatics are collected in the med-bio package.

Download: http://www.debian.org/devel/debian-med/index.html

Distributions that were not updated

Of the distributions we reviewed in 2007, some were not updated but are still available in the same version on their websites (this is the case for BioKnoppix, BioBrew, Vlinux, VigyaanCD, Quantian, and Goebix), others are not available any longer and in some cases the original websites are also not available (this is the case for AR.EMBNET, DebianBioinformatics, BioLand, APBioKnoppix2 and BioLinux-BR).

Conclusions

After two years we have had again a close look at what the open source world makes available to scientists and students who need a bioinformatics workbench for their analysis. We have noted that, like we anticipated two years ago, updating the base distribution and the bioinformatics tools can be an issue. An aspect which is important if selecting a distribution of choice is the documentation and the availability of a "getting started" introduction. Also important is a

Table 1.

	Blast	Bioperl	ClustalX/ CLustalW	EMBOSS	Glimmer	HMMER	Phylip	Primer3	T-Coffee	Gromacs
Bio-Linux	2.2.19-1	1.4	1.83-3	6.0.1-6	2.13-4	2.3.2-5	3.68-3	1.1.4-0	6.30-1	
DNAlinux	2.2.20	1.5.2	1.83	5.0.0	2.13	2.3.2-3	3.67-1	1.1.1-1	5.31-1	3.3.3-2
Vlinux			1.83	2.9.0	2.0	2.1.1	3.6b	0.9	1.37	3.2.1
Bioknoppix		1.2.1	1.82	2.8.0			3.5/3c			
APBioknoppix2		1.4	1.83	3.0.0		2.3.2			3.27	
Vigyaan		1.4	1.83	2.10.0	2.13					3.2.1
Quantian	2.2.12	1.4	1.83 (+ClustalW-MPI)			2.1.4	3.61		2.50	3.3-2
GöBIX			1.83	4.0.0					4.9.3	
Biorpms	(ncbi-6.1)	1.4	1.83	3.0.0		2.3.2	3.6a3-5	1.0.0	2.03	
Biolinux[3]	2.2.8	1.4	1.83	2.9.0			3.61	0.9		
Rocks + Bio roll	(ncbi 6.1.4)	1.5.1	2.0.11	6.0.1	3.02	2.3.2	3.66	1.0.0	7-81	4.0.4
BioSLAX	2.2.17		1.83	3.0.0		2.3.2	N/A	N/A	3.9.3	
AR.EMBNET	2.2.9		1.83	2.10.0						
BioPuppy		1.4	1.83-1		2.13-1	2.3.2-5	3.67	1.0.0	2.0.3	
PhyLIS	2.2.17	1.5.2	1.83	5.0.0-2	2.13-1	2.3.2-3	3.67	1.1.1	5.31-1	
DebianMed	2.2.21		2.0.10	6.1.0	3.0.2	2.3.2	3.68	1.1.4	5.7.2	4.0.5
Package Current Version*	2.2.21 July 2009	1.6.0 Jan 2009	2.0.11 Apr 2009	6.1.0 July 2009	3.0.2 May 2006	2.3.2	3.68 Aug 2008	1.1.4 Apr 2008	8.06 July 2009	4.0.5 May 2009

Black: current and updated
Blue: not updated
Red: no longer available for download
[3] http://www.biolinux.org/wiki/index.php/Main_Page available only to registered users

clear indication of where the tools are located and how to launch them. Having a dedicated menu certainly helps the novice users who are likely to be the target of live or USB based distributions.

From a technological point of view, we have noted a trend towards using Ubuntu or its variations as the base distributions with a few exceptions (Fedora and SLAX) and to provide the distribution as a virtual machine as the only choice or in addition to the historical live CD. In Table 1 we have summarised the current versions for the distributions we have discussed here keeping also those which we discussed in 2007. While some tools tend to be quite static, others do evolve. If you are looking for a distribution that you will install in your laboratory, choosing one which can be easily and frequently updated can make a difference.

References

[1] Rana, A., 2007, Linux for bioinformatics: dedicated distributions for processing of biological data – Part 1: Live distributions, EMBnet.News Vol.13 No.2

[2] Rana, A., Foscarini, F., 2007, Linux for bioinformatics: dedicated distributions for processing of biological data – Part 2: Repositories and Complete Systems, EMBnet.News Vol.13 No.3

[3] knoppix, http://www.knoppix.net

[4] Bio-Linux 5.0, http://nebc.nox.ac.uk/tools/bio-linux/bio-linux-5.0

[5] Bassi, S. and Gonzalez, V.. 2007, DNALinux Virtual Desktop Edition. Available from Nature Precedings <http://dx.doi.org/10.1038/npre.2007.670.1> (2007)

[6] Vetrivel, Umashankar; Pilla, Kalabharath, 2008, Open discovery: an integrated live Linux platform of Bioinformatics tools.(Software), Bioinformation, January 1, 2008

[7] BioPuppy Linux, http://biopuppy.org/

[8] BioSLAX, http://bioslax.com

[9] Geeleher, P., Morris, D., Hinde, J.P., and Golden A., BioconductorBuntu: a Linux distribution that implements a web-based DNA microarray analysis server, Bioinformatics 2009 25(11):1438-1439

[10] Thomson, R. C., phyLIs: A simple GnU/Linux Distribution for phylogenetics and phyloinformatics, Evolutionary Bioinformatics 2009:5 91–95