BioFOSS: a survey of Free/Open Source Software in Bioinformatics

Kirby Shabaga
University of Victoria
kshabaga@uvic.ca

Daniel M. German
Department of Computer Science
University of Victoria
dmg@uvic.ca

Abstract

This paper discusses the current state of Free/Open Source Software (F/OSS) projects in the field of academic bioinformatics. The paper reports on a survey of the Bioinformatics journal that enumerates the number of Application Notes published between volumes 2004-20-17 and 2005-21-7. The purpose of this survey is to determine what percentage of bioinformatics applications are made available under open source licenses. Bioinformatics includes tools, databases, and organizations to support them. An overview is given for the EMBOSS project, the Open Bioinformatics Foundation, and GenBank. In addition, a short discussion of Linux distributions tailored to the needs of bioinformaticians is provided.

1. Introduction

This paper presents a survey of the major Free/Open Source Software projects available in the field of academic bioinformatics. We start by describing how F/OSS has benefited two important bioinformatics projects: the Human Genome Project, and the Sequencing of the SARS virus.

The journal Bioinformatics, published by the Oxford University Press, is required reading for those in the field of bioinformatics. We present a survey of Application Notes published during a one-year interval that reveals the number of applications available, and analyses their licenses, identifying those that are F/OSS.

An overview and lines of code (LOC) analysis of EMBOSS, a major F/OSS project, is detailed in section 4. As with all F/OSS projects, a community has formed around the needs of bioinformaticians. Section five introduces the reader to the Open Bioinformatics Foundation (O|B|F) and provides a LOC analysis of four of the projects that it sponsors. As with most scientific computing fields, access to current data in a timely and efficient manner is essential to allow researchers to ask and answer questions. GenBank is one of three major data repositories for genetic information; section six describes this repository. Finally, section seven provides motivation for, and enumerates a number of turnkey Linux distributions that are tailored to the requirements of bioinformaticians.

2. Open source in bioinformatics

The field of bioinformatics is relatively young and has emerged in the last two decades as an important component of the biotechnology and pharmaceutical industries. Various definitions exist for the word bioinformatics and are sometimes confused with or used in addition with the term computational biology (which deals only with algorithms). The Oxford English Dictionary gives us this definition:
"The science of information and information flow in biological systems, esp. of the use of computational methods in genetics and genomics."

Ouzounis and Valencia have written a review of the history of bioinformatics [1]. In a nutshell, bioinformatics includes applications and their related algorithms created for the purpose of analyzing biological data. Examples of common applications include those for searching very large databases, comparing (computing similarity of) multiple nucleotide (DNA) and protein (amino acid) sequences, and for predicting the location of genes within a genome. In addition there are hundreds of applications available for performing predictions of biologically significant regions within a genome (e.g. promoter regions). Other applications provide visualizations of biological data and assist researchers in determining protein function and protein-protein interaction within an organism.

F/OSS has been beneficial to bioinformatics. In 1986 The Human Genome Project (HGP) was initiated with the ambitious task of sequencing the human

Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)
0-7695-2517-1/06 $20.00 © 2006 IEEE

genome. This project was described at the time as a "big science" project along the lines of the historically significant Apollo missions; it was a "moon shot". The purpose of the project was to list the complete sequence of DNA that describes man; this consists of some 3.2 billion nucleotides (fundamental building blocks of DNA). The HGP is a global project involving multiple sequencing centres and countless numbers of technicians, researchers, support staff, and computer scientists. An early and important contribution to the HGP involved tools written in the open source software language Perl. During 1996, in a Perl Journal article, Lincoln Stein, a researcher at Cold Spring Harbor Laboratory, explains how Perl saved the Human Genome Project [2]. The problem was that the vast amount of raw sequencing data generated by different centres did not follow a common record format. The solution, as simple as it may sound, was to generate a set of Perl scripts that would translate proprietary formats to a common format known as CAF. Although this task might have been accomplished with any number of computer languages, it is significant that Perl was chosen. In the early years of bioinformatics many Perl scripts were written and shared by researchers. Because of its string manipulation abilities and relative simplicity, Perl was the language of choice among computer scientists and non-computer scientists alike.

A second example of BioFOSS is found in the article "Sequencing the SARS Virus" by Martin Krzwinski [3]. This article describes how the Genome Sciences Centre released the complete sequence of the SARS Virus. The centre accomplished this task in only five days using a combination of commodity hardware and F/OSS software. Software included RedHat Linux, MySQL relational database management system, Apache web server, and well-known BioFOSS applications BLAST, Phred, Phrap, and Consed. In addition, custom Perl scripts and Java based applications were designed, written, and deployed at the centre. A custom interface to the hardware sequencers was also written showing that the entire workflow from sequencing to assembly to web publishing was performed by F/OSS applications.

3. Bioinformatics' application notes

The journal Bioinformatics began publication in April 1985. As the field has developed, changes were made to the journal and it now currently includes three main sections: Editorial, Original Papers, and Application Notes. The Application Notes section is dedicated to announcements of software for bioinformatics. Recently, the Application Notes section has been further organized into application categories such as genome analysis, phylogenetics, and sequence analysis. This evolution within the journal structure is an indication of how quickly this field is developing and how applications are moving from generic to specific in nature.

The journal includes a "notes to authors" section both within the hardcopy journal and on the journal web site. In March 2005 the instructions for section "Applications Notes" included this sentence: "The software or data should be freely available". We were interested to know how many of these applications were truly open source, and of those, what was the license they were using. We analyzed the application notes from Volume 20 issue 7 (July 2004) to volume 21 issue 7 (July 2005).

During this period 113 application notes were published. Twenty-nine of these notes were for web-based applications and were ignored by this study. Nine of the remaining 84 application notes had broken URL's and could not be evaluated. 75 applications remained.

Unfortunately the Journal does not require authors to indicate the license under which the software is available. Of these 75 applications, 5 did not make their source code available. We were able to download source code for 36 applications. The remaining 34 applications did not make their source available, but did not indicate if it was not available either (the authors were not contacted to clarify this). This breakdown is listed in Table 1. The results show that less than half of the application notes provide easy access to their source code.

Table 1. Source code availability

Count	Source code availability
29	Web based
9	URL provided in the Journal was broken
5	Source code explicitly not available
36	Source code available for download
34	Unknown

Unfortunately the application note does not usually describe the license under which the application is available. On 6 April 2005 one of the authors contacted the editorial office to clarify what was meant by "freely available" and if it was possible to add an additional field to journal papers stating what license the software was released under.

The executive editor, Alex Bateman responded by email on 7 April 2005. His response was:

"Thank you for your mail. You are right that the current wording freely available is not very specific. This means that the software does not cost anything, and in general we only require it si [sic] free for academics. We are in the process of rewriting our scope and it is likley [sic] this will contain wording to encourage source code to be available. We will continue to monitor and

update our instructions in this area and thank you for your comments!"

In July 2005, the website containing instructions to authors was rewritten to include:

"Software or data must be freely available to non-commercial users. Availability must be clearly stated in the article. Authors must also ensure that the software is available for a full TWO YEARS following publication. Web services should not require mandatory registration by the user. Additional Supplementary data can be published online-only by the journal, or may be provided on the author's web site.

If describing software, the software should run under nearly all conditions on a wide range of machines."

The modifications made to the instructions are welcome; however, it would be beneficial if licensing terms were also required prior to publication.

Of the 36 application notes that provided easy access to source code, a variety of licenses were found (Table 2). The majority of the proprietary licenses were from academic institutions. Surprisingly, four of the applications do not state any licensing terms at all.

Table 2. License types

Count	License Type
14	GPL
9	Proprietary
4	LGPL
4	Unknown; not available
1	BSD
1	Common Public License
1	DARPA BioCOMP Open Source License
1	MIT
1	Public Domain Notice

Licensing information for application notes is important for both philosophical reasons and for practical reasons. Some researchers choose only to use F/OSS tools; in many cases this may be due to limited funds available by the researcher. F/OSS applications "level the playing field" for researchers from lesser economic regions of the world. In addition it allows a way for those interested (and capable) to contribute improvements back to the community. After all, there is no way to predict where the next "big discovery" may come from.

An additional practical reason for licensing is discussed in section seven of this paper. There are a growing number of turnkey Linux distributions made available to novices and researchers who lack dedicated support staff. Those distributions must include F/OSS applications in order to gain maximum exposure.

4. EMBOSS: a major project

In March 1996 the EMBOSS (European Molecular Biology Open Software Suite) project was undertaken by the Human Genome Mapping Project/Rosalind Franklin Centre for Genomics Research (HGMP/RFCGR) in the United Kingdom. Of interest, one of the defining requirements was the need to release a software suite as open source to the community. This may have been partly due to the fact the Human Genome Project would be releasing their data publicly and therefore it made sense to provide tools to analyze this data freely as well. EMBOSS is briefly described in a 2000 article in Trends in Genetics by Rice et al [4].

The number of target users defined for the EMBOSS project was in the tens of thousands, including ten thousand registered academic researchers from HGMP/RFCGR and some thirty thousand users across thirty countries from the EMBnet service. In addition to academic users, the target users included pharmaceutical and biotechnology companies.

Due to the large scale of this application suite and significant user base, EMBOSS was a well-funded project and staffed by senior applications designers. The project was officially accepted in August 1997 and work began in earnest in November 1997. EMBOSS is a collection of software applications built upon a well-defined software framework and published APIs (AJAX and NUCLEUS).

EMBOSS comprises some 300 applications and supports multiple Graphical User Interfaces (GUIs) and Application Programming Interfaces (APIs). It is one of the mainstays of bioinformaticians and is included in virtually all of the bioinformatics Linux distributions as mentioned in section seven of this paper. In addition, it is included as part of combined hardware/software turnkey systems such as the Apple Workgroup Cluster for Bioinformatics by Apple Computer, Inc. It estimated that at least 10,000 sites have installed this software suite.

As with any successful large-scale software application, a number of standards were defined at project startup time. These included: GPL/LGPL licensing, ANSI C standard code (and now includes C++ and Java), support for all common Unix platforms, extensibility, user interface guidelines, code documentation, and testing. The EMBOSS project demands nightly builds of the entire applications suite and includes two thousand plus test cases. In addition, software documentation and code documentation are validated and indexed.

The major releases of EMBOSS are: 15 July 2000 – EMBOSS 1.0.0, 15 July 2002 – EMBOSS 2.0.0, and 15 July 2005 – EMBOSS 3.0.0.

In July 2005 funding for the EMBOSS project was cut-off. In anticipation of this event, the project was moved to SourceForge. On 23 July 2005 a snapshot of the application suite was analyzed using GNU/Linux utilities and the StatCVS tool. Table 3 shows the

number of source code files (by programming language) and basic lines of code statistics. Files were filtered by programming language (based on file extension) using variations of the command line:

```
find . -name "*.h" | xargs wc | sort -n -k 1,2.
```

Table 3. EMBOSS source code analysis

		Lines of Code (LOC)				
File	Count	Total	Min	Max	Avg	Median
.h	94	18067	15	1554	111	192
.c	309	324312	26	23054	359	1050
.pl	35	31173	17	1351	108	274
.java	138	46120	19	1610	192	334

Since 10 February 2002 access to the daily snapshot of EMBOSS has been made available publicly via CVS. A complete snapshot taken on 23 July, 2005 was processed by StatCVS version 0.2.2. The project is significant in size (approaching 3 million lines of code), however, it must be noted that the snapshot is not limited to just source code. It also includes test data files, documentation and images. All files were included in the StatCVS analysis as they are relevant to and part of the EMBOSS project. The EMBOSS project is approaching nine thousand files with an average file size of approximately 300 lines.

StatCVS results suggest that from October 2000 to July 2004 the project was increasing at a steady rate. As the number of files increased so do the lines of code, as one would expect. A sharp rise at July 2004 occurs when the CVS repository was moved to SourceForge and additional non-source code files (E.g. test cases, test data, documentation) were added to the project.

5. Open Bioinformatics Foundation (O|B|F)

The Open Bioinformatics Foundation (OBF) is an example of a second major project, however, it exists to provide administrative support to existing projects rather than to develop new applications. Unfortunately, complete historical background information is not available from the organization website. An email was sent on 8 March 2005 to the board asking for more detailed information on the activities and support provided by the organization; no reply has been received. All information in this section is derived from the organization website [5] and from the projects that it lists as supporting.

It is not stated when the organization was founded, however, a newsletter is available for 2001 and 2002; it appears that the organization has existed since the winter of 2001. No newsletter after 2002 is available.

From the contents of the Winter 2001 newsletter it appears that an informal “umbrella” group of volunteers developed from the BioPerl project into the OBF. At that time, funds of US$7000 were administered by a small business account owned by the BioPerl project. It is not clear where startup funding came from; there is, however, mention of some financial support coming from Sun Microsystems, Inc. for an upcoming conference along with a breakdown of income and expenses for the previous years conference.

It is assumed that the majority of funds for the organization are received from registration to BOSC (Bioinformatics Open Source Conference). The first BOSC conference was in 1999. Additional support, presumably hardware and/or software tools, comes from the following supporters: WYETH, Sun Microsystems, Inc., Apple Computer, Inc., O’Reilly & Associates, and Electric Genetics. In addition to these listed supporters, individual projects may receive support from other sources. For example, the BioJava website thanks Genetics Institute, Inc. and the Compaq Bioinformatics Solutions Centre for bandwidth and hardware.

The mission statement for the OBF is:

“The Open Bioinformatics Foundation is a non profit, volunteer run organization focused on supporting open source programming in bioinformatics.”

The organization was incorporated in the state of Delaware, USA as a not-for-profit company and has a pending application with the US IRS for tax-exempt status as a 501(c)(3) non-profit foundation.

A very brief explanation and history from the organization website reads:

“The foundation grew out of the volunteer projects Bioperl, BioJava and Biopython and was formally incorporated in order to handle our modest requirements of hardware ownership, domain name management and funding for conferences and workshops.”

The organization lists three main activities:

1. Underwriting and supporting the BOSC conferences
2. Organizing and supporting developer-centric "hackathon" events
3. Managing our servers, colocation facilities, bank account & other assets

As of August 2005 this organization lists ten sponsored projects. Web hosting and source code repository services are provided for the EMBOSS and BioPathways projects. The remaining eight projects are: BioPerl, BioJava, BioPython, BioRuby, BioPipe, BioSQL / OBDA, MOBY, and DAS. These projects provide a library of functions to aid in the development of bioinformatics applications, data repositories and nascent workflow applications.

The board of directors numbers six and includes members from sponsored projects (BioPerl, BioPython) along with academic (UC Berkeley) and

commercial (BioTeam, Dalke Scientific) representation. Four of the members have ties to the BioPerl project. Based on the organization news site, BioPerl appears to either be the most active project, or at least is the most represented one. Table 4 provides statistics for the four sponsored projects relating to specific programming languages.

Table 4. OBF projects: source code analysis

Project	Files	Total	min	max	avg	mdn
		Lines of Code (LOC)				
BioJava *.java	2178	323363	4	2815	148	90
BioPerl *.pl	101	19836	14	4245	196	85
BioPython *.py	599	131606	1	14499	220	108
BioRuby *.rb	127	37506	2	1559	295	197

From Table 4 we can see that the median lines of code for all four projects are close to 100. The notable exception is the BioRuby project that is closer to 200 lines of code per file and may be an indication of the immaturity of the project and maybe less modular, or an indication of an inherent difference in the Ruby language itself.

From June 23-24 2005 the sixth annual BOSC (Bioinformatics Open Source Conference) took place in Detroit, Michigan as one of the Special Interest Groups (SIG) of the 13th International Conference on Intelligent Systems for Molecular Biology (ISMB). A detailed program was made available prior to the conference and lists sixteen abstracts on topics to be presented. In addition to these, about twenty "lighting talks" and software demonstrations were scheduled.

6. GenBank: a major public database

In addition to applications, the field of bioinformatics relies on public, free access to information. This information includes the vast amount of raw nucleotide sequences generated by sequencing centres and research labs around the world.

GenBank [6] is one of three public databases that provide simple and free access to anyone interested. The other two related databases are EMBL Data Library in the UK and the DNA Data Bank of Japan (DDBJ). Collectively, these three databases comprise the International Nucleotide Sequence Database Collaboration.

GenBank is maintained by the National Center for Biotechnology Information (NCBI). NCBI is a division of the National Library of Medicine (NLM) located on the campus of the National Institutes of Health (NIH) in Bethesda, MD, USA. NCBI places no restrictions on the use or distribution of the GenBank data.

As sequencer technology has improved and the scale of projects has increased, notably the Human Genome Project, the amount of raw data available to the public is immense. By February 2004, GenBank included more than forty million sequences totaling forty-five billion base pairs.

Updates and additions to GenBank occur daily and a downloadable version of the database is generated every two months. The size of the database is measured in billions of base pairs. It has experienced dramatic growth since the announcement of the first complete chromosome being sequenced by the Human Genome Project in 1999.

At the time of research, version 148 of the flat file database required approximately 172 GB (sequence files only) or 189 GB (including the 'short directory', 'index' and related text files). Access to the database does not require registration and does not require any monetary transaction.

GenBank may be accessed by a number of methods; through a web-based interactive search tool (Entrez), programmatically with an application programming interface, and by direct ftp download of the flat file database.

7. Linux distributions

There exist a number of live and full Linux distributions tailored to the requirements of bioinformaticians. The Canada Genome Bioinformatics newsletter in March 2005 describes three full distributions and three live distributions [7]. A quick search reveals that at least two more live distributions are also available (Table 5). It is likely that this table is not complete, and that other distributions exist.

These Linux distributions allow researchers who are not tech savvy, or do not have dedicated support staff in their labs, to experience the best of what BioFOSS has to offer. This is especially true in the case of live distributions and allows a researcher who does not have access to a Unix ™ machine to experiment with bioinformatics; many software applications do not run on Microsoft Windows XP.

Table 5. Bioinformatics linux distributions

Distribution	Full	Live	Based on
BioBrew	✓		NPACI Rocks
Bio-Linux	✓		Debian GNU/Linux
BioLand	✓		Fedora Core 2
Vlinux		✓	Knoppix 3.3
Vigyaan		✓	Knoppix v3.7
Bioknoppix		✓	Knoppix
Dnalinux		✓	Slax 4.1.4
Quantian		✓	Knoppix 3.6

The number of available distributions tailored to the bioinformatician suggests that there is a need to supply turnkey systems for researchers. This is not surprising given that many academic research labs may not have the funding available to hire full-time (or even part-time) systems administrators and/or software developers. While many researchers may have the technical knowledge to install and maintain a system it is not their primary task and takes valuable time away from their research. In addition, it is interesting to note how many of these distributions are live. These are directly or indirectly targeted towards non-computer savvy users. As mentioned earlier in this paper, many of the applications available for bioinformatics research were first written (or are only available) on Unix ™ systems. Providing live distributions to the bioinformatics community allows the large base of Microsoft Windows equipped labs access to these applications.

A problem found in live distributions is how researchers can archive their data from session to session. There may be an opportunity for providing such a service to the community. An interesting idea may be to integrate the gmailfs (Google's Gmail filesystem) into one of these distributions and therefore allow researchers to have network storage available to them from any network computer they are working on. Of course, space is always an issue when dealing with very large sequences.

A final observation is that none of these distributions originate from North America. This may suggest that either F/OSS is more prevalent outside of North America or that North American researchers have access to more funding and therefore do not need these distributions.

8. Conclusions

Given the culture that most academic research occurs under, particularly when funded through limited public grants, F/OSS and bioinformatics is a natural match. In addition, the field of bioinformatics and the emergence of relatively inexpensive workstation computers arrived almost simultaneously. Combine this with the requirements for string manipulation, database, etc. and quite early a collection of ad hoc tools were being crafted in laboratories around the world. As researchers shared these tools and improved upon them collaboration became an obvious benefit in order to reduce the workload.

A terrific example of what may be achieved within the bioinformatics F/OSS community is the EMBOSS suite of tools that is described in section four.

As the field of bioinformatics continues to mature so too does the sophistication of BioFOSS projects. In particular, it is becoming clear to many that it is possible to combine many of these projects into a very powerful and useful workbench. An example of this is the Taverna project [8].

BioFOSS is quickly maturing as is evidenced by the formation of the Open Bioinformatics Foundation and the Bioinformatics Open Source Conference.

9. References

[1] Ouzounis, C. A., Valencia A., Early Bioinformatics: the birth of a discipline – a personal view. *Bioinformatics*, 19, 7 (Mar. 2003), 2176-2190.

[2] Stein, L., How Perl Saved the Human Genome Project. *The Perl Journal*, 1, 2 (Feb. 1996).

[3] Krzywinski, M., Sequencing the SARS Virus. *LINUX Journal*, 115 (Nov. 2003).

[4] Rice P., Longden, I., Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, 16, 61 (June 2000), 276-277.

[5] Open Bioinformatics Foundation. http://www.open-bio.org/, June 2005, Accessed July 2005.

[6] Benson, D. A., Karsch-Mizrachi I., Lipman, D. J., Ostell J., Wheeler, D. L., GenBank: update. *Nucleic Acids Research*, 32, Database issue, (Jan. 2004), D23-D26.

[7] Tiwari, B., Gearing up for Bioinformatics. Canadian Bioinformatics Help Desk Newsletter, 34, (Mar. 2005).

[8] Oinn T., Addis M., Ferris J., Marvin D., Senger M., Greenwood M., Carver T., Glover K., Pocock M. R., Wipat A., Li P., Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, 20, 17 (June 2004), 3045-3054.