BioManager: the use of a bioinformatics web application as a teaching tool in undergraduate bioinformatics training

Sonia Cattley and Jonathan W. Arthur

Submitted: 6th June 2007; Received (in revised form): 22nd July 2007

Abstract
The completion of the human genome project, and other genome sequencing projects, has spearheaded the emergence of the field of bioinformatics. Using computer programs to analyse DNA and protein information has become an important area of life science research and development. While it is not necessary for most life science researchers to develop specialist bioinformatic skills (including software development), basic skills in the application of common bioinformatics software and the effective interpretation of results are increasingly required by all life science researchers. Training in bioinformatics is increasingly occurring within the university system as part of existing undergraduate science and specialist degrees. One difficulty in bioinformatics education is the sheer number of software programs required in order to provide a thorough grounding in the subject to the student. Teaching requires either a well-maintained internal server with all the required software, properly interfacing with student terminals, and with sufficient capacity to handle multiple simultaneous requests, or it requires the individual installation and maintenance of every piece of software on each computer. In both cases, there are difficult issues regarding site maintenance and accessibility. In this article, we discuss the use of BioManager, a web-based bioinformatics application integrating a variety of common bioinformatics tools, for teaching, including its role as the main bioinformatics training tool in some Australian and international universities. We discuss some of the issues with using a bioinformatics resource primarily created for research in an undergraduate teaching environment.

Keywords: biomanager; undergraduate; bioinformatics; interface; integrated; teaching

INTRODUCTION
Over the last decade, bioinformatics has rapidly advanced in importance as a field of study. A large number of specialist undergraduate and postgraduate degrees have appeared in universities and colleges across the globe demonstrating a perceived need for bioinformatics knowledge and training both now and in the immediate future [1, 2]. These courses fall in one of two streams. There are degrees structured to include intensive training in both biology and computer sciences, in order to produce individuals with a truly cross-disciplinary set of skills who are able to create or modify bioinformatic software applications to address questions in life science research programs. The alternate stream aims to provide life science researchers with an applied knowledge of how to work existing programs, apply them to specific situations and competently interpret the results [3]. These students generally do not have a background in computer science and are usually not familiar with the Unix or Linux operating system under which many bioinformatics
applications are run. Furthermore, many of these students are interested in making effective and appropriate use of bioinformatics applications for furthering their biological research, rather than seeking to understand the information technology and algorithms in, and of, themselves. This latter stream is generally included as part of a biotechnology or general biological science degree. BioManager has been utilized as a training environment within this latter stream in Australia, New Zealand and Malaysia.

Students graduating in molecular biology, veterinary science, agricultural science and medical science, are expected to have a rudimentary knowledge in the application of some basic bioinformatic programs. At the most rudimentary level, this would include familiarity with several common, online data repositories, such as UniProt [4] and GenBank [5], as well as the use of BLAST [6] (Basic Local Alignment Sequence Tool) for identifying sequence similarities in both DNA and protein sequences. More usefully, the researcher should also receive training in multiple sequence alignment, polymerase chain reaction (PCR) primer design, restriction mapping, evolutionary phylogeny, gene detection, microarray analysis, protein structure and function prediction, proteomic protein identification and characterization, motif searching and sequence assembly. Depending on their area of research their may be additional areas, where training is required or more depth is required in certain of the above areas. While the theory behind these techniques can be adequately covered through standard lecture communication, the practical applications require a degree of ‘hands-on’ computer interaction, in order to sufficiently learn the concepts.

There are a number of different ways a course can be structured around these multiple applications. Many, although not all, of the software applications are made available as web applications by the institute or group providing the application. Students can access each application as needed through the corresponding web site. This method has several difficulties. In most instances, each application is hosted on an individual site. In order to cover all the applications, students must visit each different site in turn. Some students find this makes it difficult to maintain concentration, as they are constantly moving between web sites, and many more find it frustrating and inconvenient. It is also difficult for educators to construct meaningful biological examples by combining two or more applications in the solution of a particular problem, as output from one application needs to be cut and pasted to be used as input for another, often with reformatting of the data required in between the cut and the paste. The educator must also constantly revise the training material to reflect changes in the Uniform Resource Locator (URL) of the resource. Finally, many online resources are not hosted on dedicated servers and thus may periodically become unavailable or may not return results in a reasonable space of time due to fluctuating load. In a study of the stability and persistence of URLs published in MEDLINE, Wren [7] found only about 63% of URLs were consistently available with another 19% available intermittently. Sometimes, the conduct of the class itself, where multiple students are submitting similar analyses at roughly the same time, can trigger a collapse in the server or a significant delay in response time.

An alternative solution is to download and install the various bioinformatics applications on a local, central server at the institution conducting the teaching or independently install the software on multiple computers within the class room. Once installed, access to the central server is generally configured to be available only through a series of computers or terminals used by the students and networked with the central server or through a specific portal created for the purpose of offering the training. Thus, student access is limited to class time and this may hamper the ability of some students to complete extra work, such as assignments, using the programs [8]. Individual installations on every machine are also problematic. While technology to create and propagate machine images can ameliorate the need for a repeated, laborious installation on every machine, such technology often still requires input from a specialist in systems administration to manage the process. Furthermore, students still have limited access to the applications for homework and assignments outside of class time. In small departments, or where the bioinformatics component consists of only a minor component of the subject, the time and cost involved in setting up and maintaining such a system is difficult to justify.

Sydney Bioinformatics and one of its predecessors, the Australian Genomic Information Centre (AGIC), through the Australian National Genomic Information Service (ANGIS, www.angis.org.au) has provided a UNIX (1991–present) and web-based (1996–present) bioinformatics interface for academic,
government, non-profit and industry researchers since 1991. Whilst its main purpose is to provide for the needs of Australian researchers, it is available throughout the world and already used in New Zealand and Malaysia. The current web interface, BioManager, was installed in 2001 and has over 2800 active research-based users. This system is a web application and is thus potentially accessible anywhere in the world via the Internet. The application currently runs on a group of servers including a Sun E450 (2 × 440 MHz CPUs, 2 GB RAM), a Sun E3500 (6 × 400 MHz CPUs, 2 GB RAM, 2 TB HD) and six Sun Blade 100s (each 480 MHz CPU, 1 GB RAM) although, at the time of writing a major upgrade of the hardware to two Sun V890 (8 × dual core 1.2 GHz CPUs with 32 GB RAM) backed by a 4 TB high performance storage area network was underway. BioManager integrates over 280 bioinformatic programs and sequence databases (a list can be found at http://www.usyd.edu.au/sydney/bioinformatics/pdfs_docs/bioman_programs_and_dbs.pdf) from a range of different software packages, making it more comprehensive than graphical user interfaces to single packages such as EMBOSS explorer and wEMBOSS, which integrate the applications from the popular EMBOSS package. The installed packages are periodically updated as new versions become available, in order keep the analyses completed through BioManager up to date. Furthermore, the system is built in a modular fashion to allow BioManager staff to continue to expand the list of packages available through BioManager as required. In addition, the provision of a storage facility to contain user sequence and results files allows BioManager to provide a seamless workstation for bioinformatics analyses.

While the main focus of BioManager has been in providing a service for researchers, academics using the system have begun to use the interfaces in their undergraduate classes. This has happened gradually, starting with the UNIX interfaces, and progressing through other interfaces as they have been made available. Using BioManager precludes the need for program installation or maintenance by the educator or the student. The online interface allows the students to access the interface from their home computer, facilitating further study or the completion of assignments outside the class room. The BioManager servers are dedicated to this purpose ensuring the URL of the software is constant and load on the server is properly managed and not subject to fluctuations due to the use of other non-related applications on the same server. Finally, the integrated nature of the BioManager environment provides the student with the same ‘look and feel’ to the different applications, aiding in concentration and reducing frustration and inconvenience, and enables examples involving multiple applications to be constructed by automatically moving output from one application into a second application via the correct data transformations.

UNDERGRADUATE APPLICATIONS

At present there are 28 courses across 17 universities around the world using the BioManager interface as part of their bioinformatics undergraduate training (Table 1).

The BioManager interface allows teachers to create tutorials using a series of standard bioinformatic programs incorporated into a web interface. Thus, no specialist software is required to be installed on the student computers: only a Java-enabled internet browser under a Windows, Apple or Linux/Unix operating system. The current version is known to work on Internet Explorer v6 and v7 as well as Firefox, together comprising about 91% of browser usage (http://www.w3schools.com/browsers/browsers_stats.asp). Furthermore, its wide

<table>
<thead>
<tr>
<th>University</th>
<th>Number of students</th>
</tr>
</thead>
<tbody>
<tr>
<td>Charles Darwin University</td>
<td>12</td>
</tr>
<tr>
<td>Flinders University</td>
<td>104*</td>
</tr>
<tr>
<td>Monash University</td>
<td>414*</td>
</tr>
<tr>
<td>Monash University, Malaysia</td>
<td>83</td>
</tr>
<tr>
<td>Otago University, New Zealand</td>
<td>52*</td>
</tr>
<tr>
<td>Royal Melbourne Institute of Technology</td>
<td>26*</td>
</tr>
<tr>
<td>University of Adelaide</td>
<td>60</td>
</tr>
<tr>
<td>University of Melbourne</td>
<td>78*</td>
</tr>
<tr>
<td>University of New England</td>
<td>26*</td>
</tr>
<tr>
<td>University of South Australia</td>
<td>26*</td>
</tr>
<tr>
<td>University of Sydney</td>
<td>260*</td>
</tr>
<tr>
<td>University of Technology, Sydney</td>
<td>26*</td>
</tr>
<tr>
<td>University of Wollongong</td>
<td>50</td>
</tr>
<tr>
<td>University of Western Sydney</td>
<td>26*</td>
</tr>
<tr>
<td>University of Western Australia</td>
<td>52*</td>
</tr>
<tr>
<td>University of Queensland</td>
<td>20</td>
</tr>
<tr>
<td>Victorian University of Technology</td>
<td>26*</td>
</tr>
</tbody>
</table>

*Estimated on the basis of the number of BioManager logins assigned for the class.
use as a research tool, means it is used in numerous environments, including remote areas of regional Australia with less reliable internet connection and reduced access to bandwidth, and sites outside of Australia. There have been few, if any, reported difficulties arising from the specific environment where BioManager has been used.

The interface, and all generated data, is centrally housed on a server at the University of Sydney. Access is via the web using a login and password assigned by the tutor. Each student is given a unique account and all files created during the class are stored in this account. The account can be accessed at any time and at any location. The tutor can share a series of files with all students and students can share files between each other (to facilitate working on group assignments). Bioinformatics analyses are run on the central server. All analyses are completed via a queuing mechanism, being directed to the most appropriate available processor. Students do not need to wait until the job has finished but can come back to their account at a later time to view the results (Figure 1). As the computationally intensive analyses are run via a queuing mechanism the user interface is relatively 'light' and can be accessed worldwide over a regular internet connection.

One of the main features of BioManager is the automatic formatting of an output file from one program in such a way that it can be used as the input for a related program even where the two programs are not from the same package. For instance, a multiple sequence alignment generated through ClustalW [9] can be used as input to a phylogenetic program from the Phylip [10] package. All formatting requirements between suites of programs are dealt with automatically by the BioManager application. This prevents the need to download and install applications to handle the format changes, where such applications exist, or perform the changes manually, where they do not. It also allows the student to concentrate on the biological analysis without regard to data format transformation.

Figure 1: A screenshot of the BioManager workbench. Selecting the GO option will extract all suitable programs to analyse the selected file.
Students wishing to understand more about the underlying file formats can still access the raw file formats as required.

BioManager also allows both program directed (‘I want to create a phylogenetic tree—what program can I use to do that and which of my data files can be used as input?’) and data directed (‘I have a sequence I want to know more about—what program or analysis can I use on this file?’) approaches using either the Program Index or the Workbench, respectively. This helps guide the student to the required data or alternatively make the student aware of different ways of analysing their data they may not have been aware of previously.

BioManager saves the history of all analyses completed. This history can be displayed by viewing an output and selecting the History view option. In addition to the ancestry of the result (sequences and other output files generated along the path to the final result), the parameters used in the generation of the select result are recorded (Figure 2). This allows students to review their work and more fully understand the process of moving from their original data to the biological conclusions generated by their analysis.

In feedback from the universities concerned, all classes have some formal class time dedicated to the interface, where students submit the same analyses at roughly the same time from multiple computers as they work through set exercises. At other times, students either work on the system in their own time or continue working at pre-determined class times but on their own projects. Here, job submissions are more sporadic and there is not the en-masse submission as seen in the initial class or classes. In more than half of current classes, students are given assignments based on the use of BioManager.

The University of Sydney offers a number of Units of Study that include teaching in the field of bioinformatics. Some of these units make use of the BioManager interface (e.g. Molecular Biotechnology—MOBT3101, Molecular Genetics and Inheritance—BIOL5001, Pharmacokinetics and Pharmacogenetics—PHAR3630, Cell Biology 1B—VETS1018 and Genetics and Biometry—VETS2009) while others do not.
(e.g. Bioinformatics—INIM5006), mainly according to the content of the unit of study. For example, BioManager currently focuses on genomic, comparative genomics and sequence-based bioinformatics. Thus, teaching in other areas of bioinformatics, such as microarray analysis, currently makes use of other software packages. Our experience has shown a number of advantages for both teachers and students in making use of the BioManager environment. These have been outlined earlier. Indeed, in at least one unit of study relying on bioinformatics software outside the BioManager environment, a strong preference is given to the use of online resources. This, at least, reduces the burden of systems administration to install and upgrade the bioinformatics software and databases on all the machines in the class room, which remains substantial even after the use of imaging technology to rapidly transfer an installed environment from one machine to another.

Informal feedback received from academic staff associated with the current 28 courses across 17 universities, and others who have used BioManager for teaching over the last 15 years (see Ai et al. [11] for an example of an approach to teaching including BioManager), has largely echoed our experience of the benefits of using a system like BioManager for teaching while, of course, also providing feedback on the BioManager application to be used in further development of the system. While beyond the scope of this article, a formalized comparison of different modes of teaching delivery would be informative for further planning of teaching strategies.

**USING A RESEARCH TOOL AS A TEACHING AID**

As stated earlier, the BioManager application was designed for researchers and not as a teaching tool. Hence, a number of issues specific to the teaching situation arise when this research tool is used as a teaching aid. From communications with class tutors using the system, and our own experience, many of these issues are relatively easily addressed to improve the learning experience of the students.

**Lengthy analyses**

Some bioinformatics analyses, such as searching for sequence homologues using the local sequence alignment algorithm of BLAST and some phylogenetics analyses, take a reasonably long time to complete, even when utilizing powerful hardware. This is due to the inherent, computationally intense nature of the algorithms themselves. In a research situation, this length of calculation is understood and factored into the individual research plan. In a class situation, even an analysis taking up to 20 min may cause difficulties in regard to teaching. Given the output of these analyses are often needed as input for the next part of the student exercise, slow analyses may limit the amount of material covered in the class and break student concentration on the task.

The creation of a ‘course database’, a subset of sequences in the full GenBank and SWISS-PROT databases, to be used in place of the standard databases in class situations, helps to make teaching of these lengthy analyses more timely and practical. This database will be significantly smaller (by a factor of 100), so BLAST jobs will be completed in a more timely fashion. The sequences to be placed into these databases will be chosen based on sequences used in the current classes. In terms of the statistical significance of resultant hits, the situation is analogous to running BLAST against Swiss-Prot as opposed to the entire UniProt database: the change in database size impacts on the actual numbers but not the principles of how to assess the statistical significance of results. Running BLAST using the Course DNA Database produces a result in <30 s as opposed to 30 min with a large-class, en-masse submission. The result is obviously not an exhaustive search against the available sequence information but is sufficient for a class demonstration and training purposes. In the event there is an assignment requiring a more intensive search, there is still the option of using the full GenBank database. However, as most assignments are completed in the students own time, the longer analysis times are usually not as critical.

**General introductory tutorials**

 Educators are usually more interested in teaching the students how to answer biological questions using the bioinformatics resources available, rather than basic instruction in the BioManager interface itself. However, a lack of knowledge about the BioManager interface and specific tutorials on how to correctly operate individual programs (e.g. phylogeny, protein analysis, etc.) within the interface can impede on the ability of students to grasp the computational biology being demonstrated in the course. While class requirements are different
between universities, the use of the BioManager interface is a common need of all courses. To address this need we have created an introductory tutorial, freely available for download (http://www.usyd.edu.au/sydneybioinformatics/pdfs_docs/bioman_intro.pdf). This tutorial can be used as part of the first formal class session. It is planned for additional tutorials to be added at a later stage.

**Rogue analyses**

Care must be taken to ensure students do not send in untenable jobs that will use up valuable CPU with no chance of completing successfully within the time frame of the class. These rogue analyses occur in a number of ways. In one situation, the student may unknowingly submit inappropriate input data or choose inappropriate search parameters. For example, multiple sequence alignments where the resultant alignment is well over the 10,000 base pair limit set by the program can unproductively use CPU time on the server. In some cases, the interface can check for the validity of input and reject, or at least question, inappropriate jobs. However, it is common for input to be technically valid but computationally untenable. In the example mentioned earlier, the individual sequences may be within the 10,000 base pair limit but the resulting alignment exceeds the limit and stalls the sequence alignment program. Additionally, where many parameters are required for a particular application, the interface is precluded from determining all possible invalid combinations of input parameters by the combinatorics of the situation.

In a second situation, the student may simply not be aware of the approximate time scale of the requested analysis. This usually occurs in the context of assignments or less directed exercises, where students ‘try out’ different analyses. This can lead to jobs still being processed beyond the end of the class time, when they have ceased to be needed but continue to use valuable CPU time.

Education about appropriate limits for various programs can help address this situation in both of the above circumstances. In addition, we are able to monitor CPU usage during class sessions to capture accidental submission of inappropriate analyses.

**Excess parameters**

As a research tool, BioManager provides the full range of options and parameter selections for the various bioinformatics applications. This is essential to ensure researchers can use the applications in the most effective way. However, the range of options can, at times, be confusing to a student who is learning to use the application for the first time.

This can be addressed by preparing the classes in such a way as to concentrate on the main parameters and options. In addition, we are currently exploring the possibility of modifying the parameter selection pages to reduce the number of options available when the tool is accessed by a class account to allow students to focus on the key parameters.

**Communicating with student**

From time to time, issues arise where the operation or maintenance of the BioManager application may impact on student’s completion of class work or assignments. For example, if a student submits an inappropriate analysis that becomes a rogue analysis requiring a system administrator to stop the analysis, it is important the student be informed so they know the analysis was not completed and they do not submit the analysis again until the correct use of the application and limits on the analysis are understood.

To address this issue, students can personalize their accounts by adding their name and email address. This allows Sydney Bioinformatics staff to contact them directly in the event that there is an issue regarding one of their analyses. Educators are encouraged to have students personalize their accounts in this manner. In addition, a Forum has been created for students and researchers alike (http://forums.angis.org.au/) to post questions regarding the use of BioManager and an email helpdesk (help@angis.org.au) also exists. This provides students with additional help in regard to the use and operation of BioManager to augment the guidance provided by their educator.

**Load balancing for research and teaching**

One area of major concern is the potential for large classes to interfere with the use of BioManager by researchers due to the large load on the system caused by such classes. During heavy periods, when there are more than one large class using the system, and when the class is formal (i.e. each student is completing the same analyses at roughly the same time in line with the class exercises), it is possible that researchers may experience a delay while class jobs are being processed.
This can be addressed by limiting the number of analyses students can simultaneously submit. Additional analyses then remain queued until the earlier analyses are complete.

Cost
The operation of BioManager is funded through the support provided by the University of Sydney, occasional research grants and small, annual subscription fees charged to users. The service is not operated for profit.

In order to facilitate the use of BioManager as a teaching tool, training accounts are provided at a significantly reduced cost. A department is entitled to a number of free training accounts equivalent to the number of research accounts it operates. The costs are small in comparison with typical teaching budgets, especially if the true cost of alternatives such as paying computer support staff to install and maintain numerous software applications on multiple machines is taken into account.

Clearly, students trained in bioinformatics through the use of BioManager, will need to purchase a subscription should they wish to continue to use BioManager in their subsequent employment (if the department they are subsequently employed in does not already have a subscription). Again, the cost of subscription is relatively small compared to common research lab budgets, particularly if the subscription is shared with other labs in the same department. In situations where the subscriptions cannot be afforded, students can in many cases fall back on the online versions of many applications. When being trained on BioManager with its common, consistent interface, the student is able to focus on the purpose of each application, the meaning of the input parameters and the correct interpretation of output data, rather than needing to learn the multitude of different user interfaces. Once the application is understood, the student should find it much easier to return to the regular web interface and understand the particular idiosyncrasies of the native interface.

CONCLUSIONS
The BioManager interface has been well received by the students and educators currently using it in 28 courses across 17 universities around the world. In particular, the ability to start an analysis and return to review the results at a later time has been seen as advantageous by time-poor students. While the use of BioManager, created as a research tool, as a teaching aid raises a number of additional issues, various strategies are relatively easily put in place to address these. There have been some marked improvements since the above changes have been implemented. The ability to expose students to a wide range of bioinformatics applications, through the one consistent interface, without either the student or educator having to manage the installation and maintenance of all these programs makes BioManager a convenient resource for undergraduate teaching and other training programs in bioinformatics. Furthermore, the automatic translation of the output from one application into the format for input into another application allows educators to focus students on more complex and meaningful exercises in computational biology requiring the sequential use of several different applications.

Key Points
- The ability to use bioinformatics resources effectively in life science research requires training on a broad range of different software applications.
- Installation and maintenance of the many different software applications is difficult, time-consuming and often expensive. Using online interfaces not supported by dedicated servers, is unreliable and confusing to the student due to the vast range of different user interfaces and access points.
- BioManager provides a consistent user interface to a wide range of bioinformatics applications. Installation and maintenance of software is not required by either the educator or the student and the application is reliable due to the dedicated server required for its use as a research tool. Automated translation of output from one software application into the required format for a second application allows the development of classes focused on the use of bioinformatics to solve real biological problems.

Funding
ANGIS, including BioManager, is subsidized by the University of Sydney and additionally supported by Australian and international researchers and teachers through subscription. The hardware upgrade described is funded by an Australian Research Council LIEF grant (LE0668549).

Acknowledgements
The authors would like to specifically thank Dr Catherine Abbott of Flinders University, Dr Alex Andrianopoulos of the University of Melbourne and Dr Neville Firth of the University of Sydney for providing specific feedback on their experiences of BioManager in a teaching environment. We also acknowledge other BioManager users who have provided feedback on the use of BioManager as both a research and teaching tool.
Information about BioManager access can be found at http://www.usyd.edu.au/sydneybioinformatics/angis/biomanager.shtml. Sydney Bioinformatics is a Centre of the University of Sydney. A free trial period is offered for international (non-Australian) departments. For further information contact Sonia Cattley (scattley@angis.org.au).

References