http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 10 (2): 434-439, 2011
ISSN 1812-5638 / DOI: 10.3923/itj.2011.434.439

A Proposed Design for a Suite of Bioinformatics Analysis Tools

[1]S. Zainudin, [1]A.R.M. Hashim, [1]Z. Shukur, [2]L.K. Keong, [2]M.N. Hamid and [2]Z.A.M. Hussein
[1]Faculty of Information Science and Technology,
[2]Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600, Bangi, Selangor, Malaysia

Abstract: Many bioinformatics analysis tools and software has been developed by the bioinformatics research community to support analysis of various types of biological data. Students involved in the bioinformatics program at the tertiary level are usually given various biological data to analyse to familiarise themselves with the bioinformatics techniques during their course of study. This study will discuss a proposed design for a suite of bioinformatics analysis tools to support the analysis and learning at the undergraduate level. The suite will be available as a package of Ubuntu Remix Linux on a netbook computer. The small size of a netbook computer made it a handy and ideal platform for students to utilize the suite of analysis tools.

Key words: Bioinformatics tools suite, computational analysis, Netbook

INTRODUCTION

Bioinformatics research that combined computational analysis of biology and computer science approach has evolved significantly. Currently, many genomic scale projects have taken advantage of the sequence technology evolution. Massive amounts of biological data such as DNA microarray and protein sequences are produced by the sequencing machines. However, the analysis on these biological data still has ample room for improvements as better analysis tools are being developed (Rice *et al*., 2000).

Research students involved in the bioinformatics program are typically provided with biological data to analyze with. Along the way, these students will familiarize themselves with various analysis tools available in bioinformatics. Various bioinformatics analysis tools and software have been developed by the bioinformatics research community. The tools are capable of performing advanced analysis for different facets of bioinformatics analysis. These tools are typically free for public use and are based on the UNIX or Linux operating system.

The study will discuss a proposed design for a bioinformatics Analysis Tools Suite for tertiary level students. The prospective users for this suite are the undergraduate students and research students from the bioinformatics Program at Faculty of Science and Technology at The National University of Malaysia (UKM). The development of this suite is a collaboration between Faculty of Information Science and Technology and Faculty of Science and Technology. The suite is designated as a component of a larger product entitled Intelligent Desktop for College Students (Sufian *et al*., 2010). The tools to be included in the suite will be the tools for basic analysis of biological data such as BLAST, CLUSTALX and others. The proposed suite will be installed on a netbook computer assigned to an undergraduate student at UKM. Netbook computers are increasingly popular due to their reasonable price and compact size that are ideal for mobile computing. Besides the price and size, a netbook contain optimized processing, is handy and portable enough to fulfill the current market demands for mobile computing. As the computing technology trend is now moving towards cloud computing, netbook is projected as a possible client device of the cloud computing system.

Ubuntu Remix is a version of Ubuntu Linux Desktop reengineered for an optimized processing on a netbook. The look, feel and design structure of its graphical user interface are clean and tidy as it is made to satisfy the resolution display constraint as small as 10 inches. The preinstalled applications of Ubuntu Remix are similar with those in the Desktop version. Most of the applications in Linux are based on C or C++ programming language and the same applies to its system kernel.

This study shall explain about the proposed design for a bioinformatics Analysis Tools suite on a Java platform. Specifically it is a Java application suite that will be run on netbook computer with Ubuntu Remix Linux as the operating system platform.

PROBLEM STATEMENT

Undergraduate students in bioinformatics program usually perform basic analysis using publicly available

Corresponding Author: S. Zainudin, Faculty of Information Science and Technology, Selangor, Malaysia

analysis tools for a multitude of biological data. These tools are typically hosted in distributed servers and accessible via the Web. In order to provide these students with convenient access to a uniform set of tools, this study proposed a design for a centralized platform for a suite of bioinformatics analysis tools. This suite will combine a few of the bioinformatics analysis tools most often used by courses under the bioinformatics program. The suite will also benefit any research students who need to perform basic bioinformatics analysis. The proposed suite will be available as a pre-installed package in a Linux system.

The design will ensure that the targeted students will have a standard analysis environment to perform the analysis on their biological data. This suite will help students to analyze their biological data accurately and helps student in getting used to the analysis of bioinformatics. Students can access these tools at any time since it is pre-installed on a netbook.

LITERATURE REVIEW

A similar system compared to the proposed suite is the KDE Bioscience (Lu *et al*., 2006). It is a platform for bioinformatics analysis based on Java. This platform contained analysis tools such as FASTA, CLUSTAL and others and proved the workflow to integrate these different tools together. Since, KDE is based on Java, it has a flexible and extensible architecture that makes it an ideal informatics environment for future bioinformatics or systems biology research (Lu *et al*., 2006).

According to Lu *et al*. (2006), the power of KDE Bioscience comes from the integrated algorithms and data sources. Other novel calculations and simulations can be integrated to augment the current sequence analysis functions with its generic workflow mechanism. Due to this flexible and extensible architecture, KDE Bioscience makes an ideal integrated informatics environment for future bioinformatics or systems biology research

Another similar system is Biobench (Albayraktaroglu *et al*., 2005). Biobench is a bioinformatics suite that combined phylogenetic analysis, multiple sequence alignment, sequence profile searching, genome-level alignment and sequence assembly. Biobench is a benchmark suite that represents a diverse set of bioinformatics applications. The first version of BioBench includes applications from different application domains, with a particular emphasis on mature genomics applications (Albayraktaroglu *et al*., 2005). Analysis tools contained in Biobench are BLAST (Thompson *et al*., 1997), FASTA, PHYLIP (Altschul *et al*., 1997), CLUSTAL W (Felsenstein, 1995), MUMMER and TIGR.

Geneious (http://www.geneious.com/) is an integrated and extendable software platform for the organization and analysis of biological data that runs on all major operating systems. This revolutionary

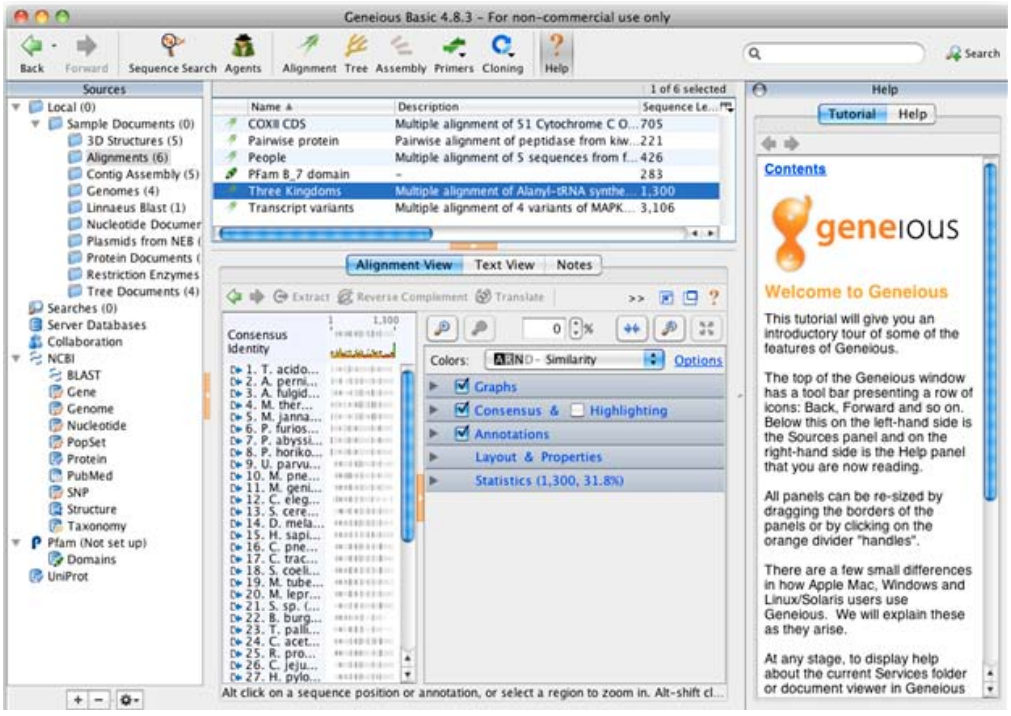

Fig. 1: Geneious software suite screenshots

bioinformatics tool combines industry-leading DNA and protein analysis tools into a single package that is both ultra-powerful and easy to use. Users are able to search, organize, analyze and visualize genomic and protein information of any size via a single powerful desktop program. This affordable product combined major DNA and protein sequence analysis tools into one software solution. Geneious is available for all major software platforms and has user friendly features that enable any biologist to access and utilize bioinformatics tools. A screenshot of Geneious is provided in Fig. 1.

We have analyzed the main features from these three systems and come up with the features for the proposed design for the suite. Further discussion of the features will be in the following section. Most of the available bioinformatics tools are free and many are written in PERL, offer as web-services and open-source. The issue here is learning how to properly install and use the different software. Hence, our proposed solution will be based on the following requirements. The suite will be developed using free Linux platform using a simple one step installation. There will be a standard interface for the different bioinformatics tools with automatic updated for the applications and databases.

RESEARCH METHODOLOGY

Firstly, this section will discuss design requirements for the suite based on the findings from literature review. The bioinformatics analysis tools suite will be based on Java language. To ensure extensibility and availability for multiple platforms, Java would be the language of choice for the development of this application suite. The suite will combine a few of commonly-used bioinformatics analysis tools such as FASTA and BLAST among others considered necessary to satisfy the students' analysis needs.

In certain process of bioinformatics analysis suite such as similarity searching, sequence data from the database is required for processing. The output from processing will be available for further analysis. The proposed suite will query data from a collection of database stored locally in the netbook's storage disk. The database shall be updated regularly from the open source bioinformatics database provider. However, under certain circumstances, users need the latest data in an instant since updating locally installed database is time consuming. A long time is needed because this is because the bioinformatics database is usually very huge in size with multiple databases to be updated. Hence, the suite will provide a choice to access data directly from the database provider. Accessing data stored locally on the storage disk is a must, since analysis might be performed offline.

Since, the application suite contains a collection of bioinformatics analysis tools, the graphical interface is designed to be an Interactive Development Environment (IDE) application. A common IDE would have a set of toolbars, a number of windows with list of process to be conducted and the settings or properties windows to configure a certain process with customized parameters for instance. To minimize implementation time for suite development, the design will be based on Rich Client Platform. Netbeans available at http://www.netbeans.org is our platform of choice to accomplish this. Netbeans modular architecture is manipulated to combine the collection of bioinformatics analysis tools.

Graphical user interface can be programmed easily with the features of What You See Is What You Get that Netbeans offers. This allows customization of JSwing component by dragging and dropping components onto a readymade canvas. In order to enable the graphical interface to request and respond to the analysis tools, a toolkit or library preferably programmed in Java language is needed. BLAST tool officially provides a C/C++ library (Altschul *et al*., 1997). An available open source library written in Java is the BioJava toolkit available from http://en.wikipedia.org/wiki/BioJava. BioJava also support other tools like CLUSTAL (Thompson *et al*., 1997) for the sequence alignment. This made BioJava the ideal choice to be integrated into the proposed design. For now, the focus is to integrate BLAST and CLUSTAL tools from BioJava into a suite prototype.

For the hardware requirements, the suite is designed and developed for the LINUX operating system. Basically the minimal requirement of the system to execute this application suite would be a 512 MB of RAM, 1 GB storage disk size, a netbook version of Linux and web access. A web server will also be provided to handle the web services that connected to the users of the application suite. The web server will update the version of application suite, the bioinformatics tools and databases. For web servers the minimum specification required would be a 2 GB RAM and 500 GB of disk space with an operating system of Linux.

Bioinformatics analysis suite will have a uniform graphical user interface to provide a standard framework to users. The graphical user interface will have user-friendly features to ease users' analysis works. The output from the suite prototype will be presented in an appropriate form and will be saved in a suitable format for future analyses. The suite also supports the feature to

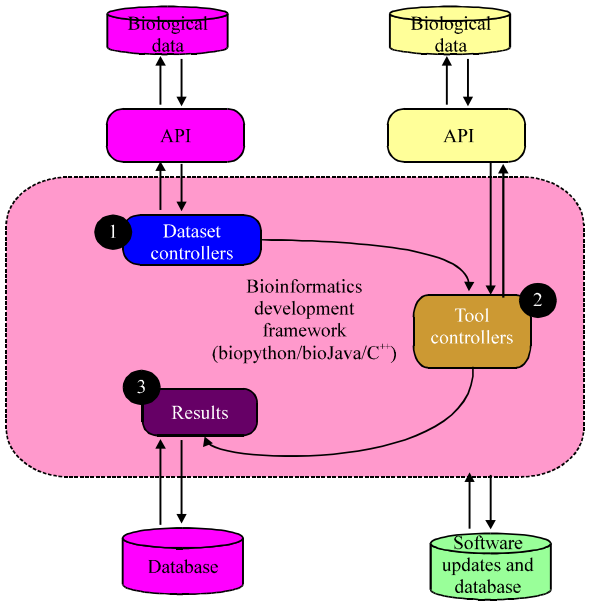

Fig. 2: Bioinformatics analysis tools suite framework design

Table 1: Input and output formats for bioinformatics tools

Bioinformatics tool	File format
Blast	Input > fasta format or text file. Output < fasta format or text file with figure.
ClustalX	Input > fasta format or text file. Output < aln or dnd or phylip or ps or pdf or fasta.
T-Coffee	Input > fasta format or text file. Output < aln or dnd or phylip or ps or pdf or fasta.
Phylip	Input > Inline txt. Output < Outline txt.
Transeq	
Translate	Input > DNA sequence in text form. Output < Protein sequence in text form.
InterProScan	Input > Protein sequence in text form. Output < Figure.
Rasmol	Input > pdb. Output < Figure or py.

save the analysis output into XML file format to ensure that it is readable on other applications that support XML. The graphic user interface was designed using Java Xswing Application Programming Interface (API) that has been integrated in Netbeans IDE Platform libraries. User interface contain five tabs containing basic bioinformatics tools. Those five tabs are Translate (translating nucleotide sequence into protein sequence), Blast (identification of nucleotide and protein sequence via database), Clustal (aligning nucleotide and protein sequences), Motif/Domain (prediction of protein functional) and Visualization (protein 3D structure viewer and analyse). Figure 2 shows the proposed framework design of this applications suite development.

This enable the analysis result to be shared and ready to run on most of bioinformatics applications. The expected input and output to the suite are contained in Table 1.

RESULTS

Nowadays, portable mini-sized computer or also known as Netbook are easily available at an affordable cost. This prompted the development of an integrated bioinformatics suite package specifically using Bio-Java based code on a Netbook platform for the usage of undergraduate students and researchers in UKM. The purpose of this software is to embed basic bioinformatics tools into a package. This paper mainly explains about the Graphic User Interface (GUI) bioinformatics Suite.

Based on the proposed design, a prototype for the suite is being developed. Figure 3 shows a screenshot of the main GUI of the prototype. The prototypes contains panel windows as the space of display. Certain panel such as analysis window panel would require a larger view. We use the hide windows feature commonly used in most of IDE application to hide or show windows. These enable the analysis to use maximum possible display space.

This can be visualized in Fig. 4. The screenshot shows only one panel windows that took up all the display view of the applications while other panel are put in minimized mode with a label icon that are placed at the border side of applications. In other words, those panels are made dockable to out the analysis tab in maximum view. Student can easily each tools being run for analysis by having tabs of panel that are placed to each other. Each tab represents a single analysis tool. In example

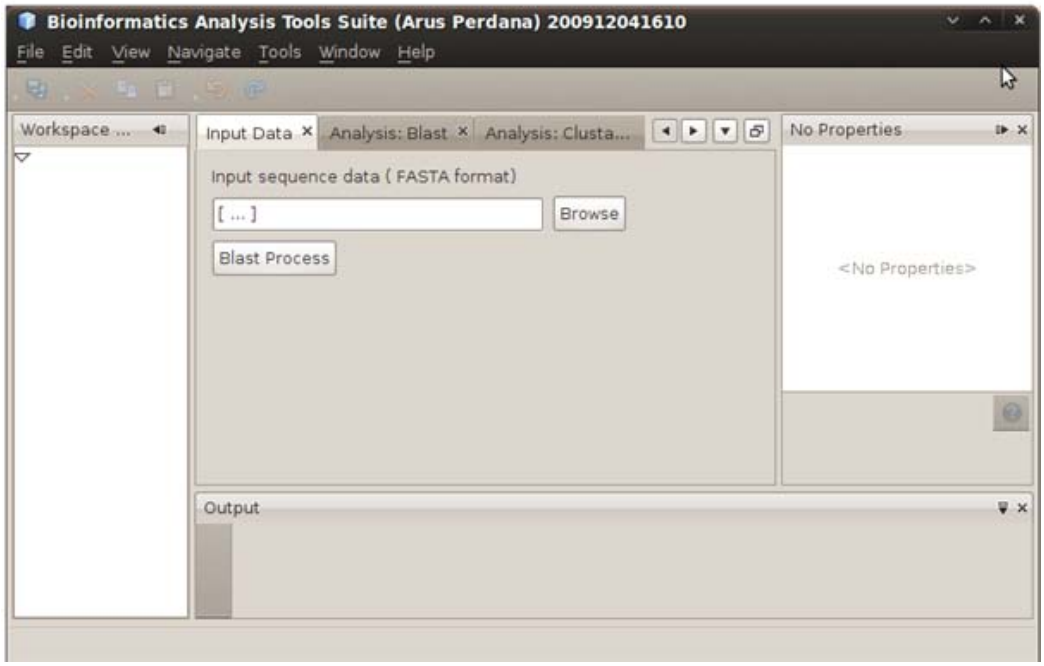

Fig. 3: Graphical user interface windows of the bioinformatics analysis suite

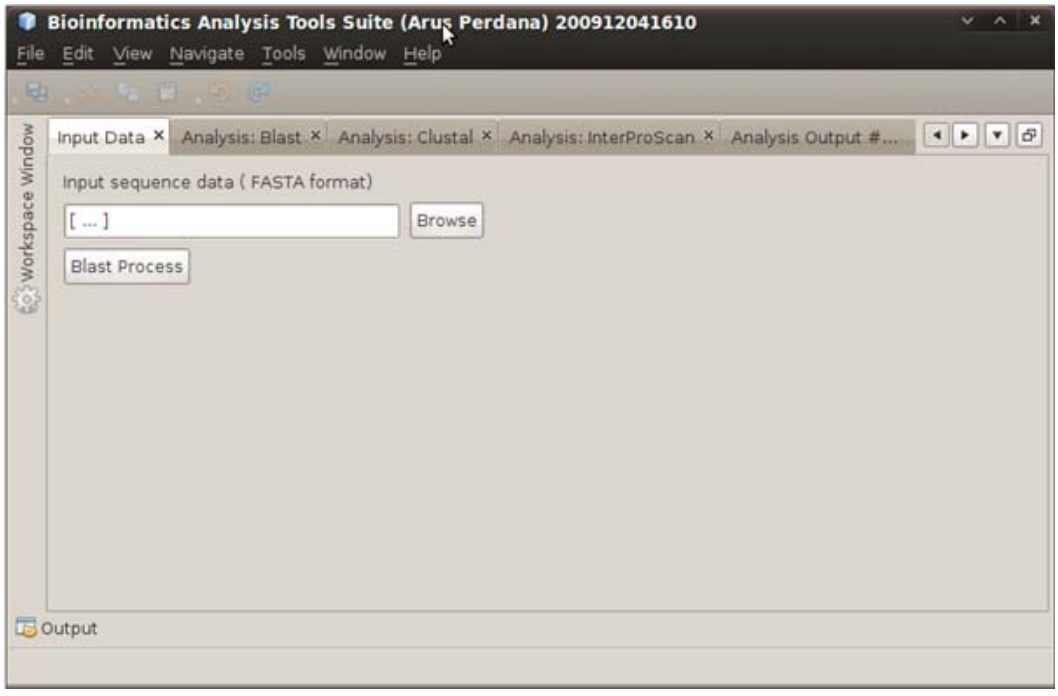

Fig. 4: The application suite with one panel windows viewed in maximized mode while putting other panels in minimized mode

Tab1 would hold the process carried by the BLAST tool, Tab2 by the CLUSTA tool, Tab3 by PHYLIP tool, Tab4 by the InterProScan tool (Zdobnov and Apweiler 2001) and so on.

CONCLUSIONS

This study has presented a proposed design for a suite of bioinformatics Analysis Tools which will be available on a Java platform. The development is still at the early stage where the focus is to complete the first prototype for this proposed system. Completing the prototype is crucial since it involves integrating different components into a common platform. Suite of bioinformatics should be able to assist undergraduate or research students to perform basic bioinformatics analysis based on their designed experiment routine. The suite should provides an insilico workbench for them to perform bioinformatics analysis in a more organized, stable and integrative environment. To conclude, the design presented in this paper is the roadmap which will be followed towards the complete development of the bioinformatics Analysis Tools Suite.

REFERENCES

Albayraktaroglu, K., A. Jaleel, X. Wu, M. Franklin, B. Jacob, C.W. Tseng and D. Yeung, 2005. BioBench: A benchmark suite of bioinformatics applications. Proceedings of the IEEE International Symposium on Performance Analysis of Systems and Software, March 20-22, IEEE Computer Society, pp: 2-9.

Altschul, S.F., F. Stephen, T.L. Maden, A.A. Shaffer and Z. Jinghui *et al*., 1997. Gapped blast and psi-blast a new generation of protein database search programs. Nucleic Acids Res., 25: 3389-3402.

Felsenstein, J., 1995. PHYLIP (Phylogeny inference package), version 3.57c. Department of Genetics, SK-50, University of Washington, Seattle, WA.

Lu, Q., P. Hao, V. Curcin, W. He and Y.Y. Li *et al*., 2006. KDE bioscience: Platform for Bioinformatics analysis workflows. J. Biomed. Informatics Elsevier Sci., 34: 440-450.

Rice, P., I. Longden and A. Bleasby, 2000. Emboss: The european molecular biology open software suite. Trends Genet., 16: 276-277.

Sufian, I., A.B. Marini and S. Zarina, 2010. Architecture of seMeja desktop system. Proceeding of the 4th International Symposium on Information Technology, June 15-17, Kuala Lumpur, Malaysia, pp: 1073-1075.

Thompson, J.D., T.J. Gibson, F. Plewniak, F. Jeanmougin and D.G. Higgins, 1997. The CLUSTAL_X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res., 25: 4876-4882.

Zdobnov, E.M. and R. Apweiler, 2001. Platform for the InterProScan-An integration signature-recognition methods in InterPro. Bioinformatics, 17: 847-848.