[Briefings in Bioinformatics, software review submission, Sept. 2002, d. gilbert]

Pise, software for building bioinformatics webs

Abstract

Pise is interface construction software for bioinformatics applications that run by command-line operations. It creates common, easy to use interfaces to these for the Web, or other uses. It is adaptable to new bioinformatics tools, and offers program chaining, Unix system batch and other controls, making it an attractive method for building and using your own bioinformatics web services.

Keywords

Bioinformatics web, Perl, sequence analysis, interface builder

Introduction

Bioscientists use Internet web services, the “fourth” operating system, for many if not most data analyses today. This is because of their accessibility, common simple interface of documents and usage forms, and often free services providing many up-to-date databases and analysis tools. While web interfaces to molecular biology analyses are not always the best choice, when they can handle the job, they may be preferred to a program running on your Macintosh, Windows or Unix system.

Many bioinformatics analysis programs have been developed with a command-line method for use. This both avoids the high cost to developers of designing a usable human interface, and allows flexibility for running on many computer systems in ways that can be integrated with other computer programs. This approach is flexible, but leaves to others the task of making such programs easily usable for a bioscientist.

Over the past several years, a number of projects have addressed this problem of human interfaces for command-line tools. The Pasteur Institute Software Environment named ***Pise***[1] is a very robust program for adapting molecular biology software to web and other human-use interfaces. Designed and written by Catherine Letondal, it runs on Unix computers with a standard web server and Perl interpreter. It is freely available from the Pasteur Institute under a GNU license for open-source software.

If you pick any dozen bioinformatics tools, you will often find a dozen different and difficult to learn command operations. Pise addresses this by providing common web form structure and syntax for use of all the programs. One common limitation of web interfaces is an inability to integrate many tasks where you wish to pass your data and intermediate results from one to the next task, such as from multiple sequence alignment to phylogenetic analysis. Pise addresses this need. The Pise software objectives are aimed at producing usable human and computer interfaces, including ability to chain or ‘pipe’ together several analysis programs in logical progressions of analyses. The outcome provides homogeneous and easy to use web interface to biology tools.

There is no need to rewrite the original software; Pise is flexible in its options for command, input and output to and from these programs. These can include small scripts (in Perl or other languages) to modify as needed parameters to make them suitable for the target program. Pise is well designed for adding new biology programs, including steps and examples on how you can add your favorite application. For many bioinformatics centers and groups, you will find the selection of already-supported applications is comprehensive. It can be useful for biologists who want a simple web interface to analyses on their own Unix workstation (including MacOSX and Linux). Along with bio-software interfacing, Pise includes methods for handling bio-databases as well as inter-conversions among data formats, a common step for analyses.

Details of operation

The general structure of Pise interface builder is to take program description documents (in XML), parse these into Perl modules with common functions for handling Web to program calls. These Perl modules are then used to make specific web documents and CGI scripts for handling each program, as diagrammed in Figure 1. This architecture compartmentalizes tasks to allow a software manager to add new bio-applications simply by defining new XML program descriptions, or a developer can add new interface generators (such as the new interface for the SeWeR[3] JavaScript package), without recourse to changing other parts of Pise.

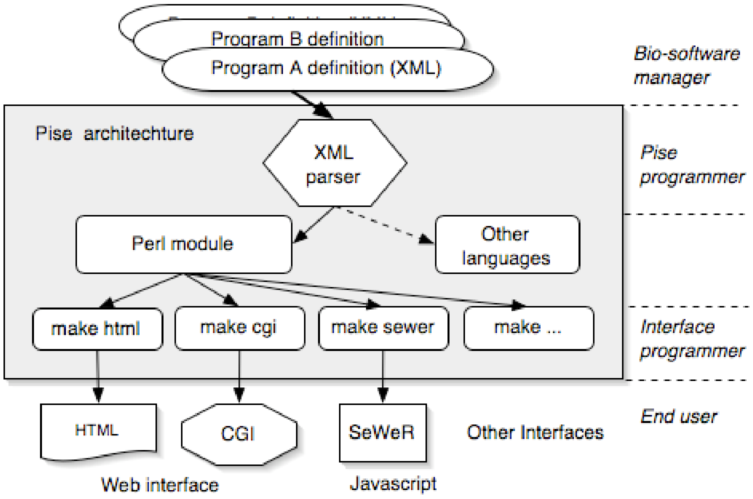

Figure 1 Diagram of the Pise interface generator architecture, adapted fromreference [1]

XML documents are used to describe the interfaces to molecular biology software. These include descriptions of the software and its options, the option or parameter syntax and content (e.g. sequence input data, integer or string values, or a choice list of terms), and program interface conversions needed for using options (such as a line of Perl code to convert value from a web form to the option value needed for the program). The XML documents are read by Pise to generate a combination of HTML web pages, Perl library and web CGI interface software. There is documentation on how to construct new program descriptions, and using one of the many included in this package gives a software manager the needed starting point.

There are over 350 program interfaces included with Pise, covering most public bioinformatics packages and tools, including the EMBOSS and PHYLIP packages. Others of note are ClustalW, HMMER, NCBI and Washington U. BLAST versions, Glimmer, Genscan, MFOLD, to name a few. Table 1 summarized the categories of programs available with a Pise web interface.

Table 1. Categories and counts of Pise interfaces (adapted from http://bioweb.pasteur.fr)

Program category	***No.***
Sequences Alignments and Comparisons	43
(BLAST, Pairwise, Multiple, Structural alignment, HMMs)	
Databases searches	2
DNA sequence analysis	48
(Restriction enzymes, transcription factors, repeats, codon usage, primers, and more)	
Search Genes and Coding Regions	18
Motifs and Pattern Search	51
Phylogeny	34
Protein Sequence Analysis	30
(Pattern searching, Protein features and enzyme kinetics)	
RNA analysis	12
Structure analysis	31
Sequence format conversions	12
Sequence tools	16

Piping, or transferring the results of one program to be input for analysis in another program, possibly with a chain of several analyses, is a growing need for bioinformatics automated analyses. This function is handled well by Pise, by marking data types that are transferable from a program. As well, Pise allows one to save several steps of analyses as a macro script that one can repeat as needed. As an example, one common bioinformatics analysis is to align multiple sequences, find the phylogenetic distance and best trees of the alignment, then draw the trees. This often is done with three or more programs. With a Pise web, one can pipe the results of *ClustalW* multiple alignment into PHYLIP phylogenetic analyses *dnadist* and *fitch*, then draw these with *drawtree*.

The web interface for Pise is designed to handle common needs for a service where the analyses may take hours or days to run. Long term access to results is provided to end users, by storing data and result files for several days, in a configurable way, and providing the user with hyperlinks to these results. Redundant submissions by an end user are detected and avoided intelligently. A recent addition, the ability to use the Portable Batch System (PBS) for managing concurrent, web-submitted analysis processes is a very good feature for setting up molecular biology applications on any shared (public or intranet) server. This allows the administrator to control the process uses of computer resources, and to schedule the time or relative amount of computing resources allocated to a program.

A basic level of Unix and Perl familiarity is required of the computer administrator who will install and manage Pise. As a prerequisite, install those molecular biology tools that you want to use with it. The Pise system leaves this to you, but provides links for finding these, including the Pasteur Institute's and others molecular biology software collections. Good sources for finding public bioinformatics programs include the BioNetBook[4], and the GenomeWeb[5]. Installation requires that you edit a few configuration files to set directory paths to the biology software and web server on your computer.

Installation steps are documented, starting with required and optional prerequisite software, which include free Perl libraries, an XML parser, and standard bioinformatics programs for data and document conversions. During a test, this reviewer ran into a few hang-ups that can occur when installing Unix programs: configuration settings needed adjustment; the Make program did not create certain directories, and he forgot to follow instructions fully. After making corrections, and installing programs to test, running bio-programs via workstation web server was fully successful.

Beyond the web, Pise provides other useful interfaces. *Pise-Bioperl* integration allows one to use the many programs known to Pise with BioPerl[6] project tools, a very handy combination for the beginning and advanced bioinformatician. *Pise-python* provides similar integration with BioPython and the Python programming environment. *Pise-SeWeR*: SeWeR[3] is a JavaScript-based web interface to sequence analysis tools which has attractive features that improve on a web HMTL interfaces. Pise recently added an interface builder for SeWeR that allows one to add new bio-applications to it.

Discussion

Pise provides an important middle layer to integrate many diverse bioinformatics tools through web and other interfaces. Among bioinformatics web services available for public use, well-integrated examples include the Australian ANGIS service[7], and its commercial offshoot, BioNavigator[8]. The Canadian Bioinformatics Resource[9] has developed a well-designed and integrated service, along with other national and regional bioinformatics centers in the European Union, including HGMP-RC and other EMBnet[10] centers, those in the Asian-Pacific region and elsewhere. Pise does not provide a complete framework to match these, but it handles much of the core construction of such web interfaces. It also offers working groups, laboratories and individuals a route to building one's own custom bioinformatics web.

The XML-based program descriptions in Pise provide a good starting point for common descriptions of computer interfaces to many standard bioinformatics tools. Other bioinformatics developers can use these same XML documents for interface development, as this reviewer has. They stand with or ahead of EMBOSS's ACD, W2H[2] and GCG package configuration files for computable program interface descriptions.

Areas where additions to Pise software could make it more useful for a bioinformatics web include: providing web page structure for organizing the bioinformatics programs, enhancing installation and update methods, and providing for user-account access. The first task can be done 'by-hand' by the person installing Pise, but can be tedious unless one has a starting point of documents for organizing and categorizing these bio-web tools. The Pasteur Institute provides one such organization in its bioweb service[12]. With many programs written in Perl and other languages, installation often requires additional packages. Aspects of installation could be streamlined, and methods to check and update Pise interfaces when new versions of EMBOSS or other molecular biology packages are installed would make useful additions.

Pise avoids the need for each user to log on to an account, in favor of a simpler 'click-and-run' approach. This makes it more accessible to many scientists. However there are instances where a user-account login procedure, such as used with W2H[2], has advantages in allowing one to maintain and manage data and result files. With Pise, you necessarily manage these files on your workstation, uploading and downloading as needed through the web interface.

Overall, if you want to build, or have someone build for you, a web interface to biology applications, look at Pise. This reviewer installed Pise on Solaris server computers and a MacOS X workstation with minimal problems, and recommends it as an alternative to tedious command-line typing, used either as a web service or via Pise-BioPerl programs. If you like what you see, there is a mailing list[1], and as with other open-source bioinformatics software, you can get involved to help extend and enhance it. Thanks to José Valverde for comments on this review.

Don Gilbert
Biology Department, and
Center for Genomics and Bioinformatics
Indiana University,
Bloomington, Indiana 47405 USA
Tel: +1 812 855 0587
E-mail: gilbertd@bio.indiana.edu

References

1. Letondal, C. (2001). 'A Web interface generator for molecular biology programs in Unix', *Bioinformatics*, Vol. 17(1): pp. 73-82. URL: http://www.pasteur.fr/~letondal/Pise/
2. Senger, M., Flores, T., Glatting, K.-H., Hotz-Wagenblatt,A. and Suha, H. (1998). 'W2H: WWW interface to the GCG sequence analysis package'. *Bioinformatics*, Vol. 14: pp. 452-457. URL: http://www.w2h.dkfz-heidelberg.de/
3. Basu M K (2001) 'SeWeR: a customizable and integrated dynamic HTML interface to bioinformatics services.' *Bioinformatics*. 17(6): 577-578. URL: http://iubio.bio.indiana.edu/webapps/SeWeR/

4. BioNetBook, URL: http://www.pasteur.fr/recherche/BNB/bnb-en.html
5. GenomeWeb, URL: http://www.hgmp.mrc.ac.uk/GenomeWeb/
6. Mangalam, H. (2002) 'The Bio* toolkits – a brief review.' Briefings in Bioinformatics 3(3): 296-302.
7. Australian National Genomic Information Service (ANGIS), URL: http://www.angis.org.au/new/index.html
8. Bionavigator from Entigen, URL: http://www.bionavigator.com/
9. Canadian Bioinformatics Resource, URL: http://www.cbr.nrc.ca/
10. European Molecular Biology network, URL: http://www.embnet.org/
11. European Molecular Biology Open Software Suite (EMBOSS), URL: http://www.emboss.org/
12. Pasteur Institute Bioweb, URL: http://bioweb.pasteur.fr/