BIOINFORMATICS APPLICATIONS NOTE
Vol. 19 no. 1 2003
Pages 153–154

An integrated genetic data environment (GDE)-based LINUX interface for analysis of HIV-1 and other microbial sequences

T. De Oliveira[1,*], R. Miller[2], M. Tarin[1] and S. Cassol[1]

[1]Molecular Virology and Bioinformatics, Africa Centre/University of Natal, Durban, South Africa and [2]Inpharmatica, Ltd., London, UK

Received on February 26, 2002; revised on June 5, 2002; July 9, 2002; accepted on July 24, 2002

ABSTRACT

Motivation: Sequence databases encode a wealth of information needed to develop improved vaccination and treatment strategies for the control of HIV and other important pathogens. To facilitate effective utilization of these datasets, we developed a user-friendly GDE-based LINUX interface that reduces input/output file formatting.

Design and Results: GDE was adapted to the Linux operating system, bioinformatics tools were integrated with microbe-specific databases, and up-to-date GDE menus were developed for several clinically important viral, bacterial and parasitic genomes. Each microbial interface was designed for local access and contains Genbank, BLAST-formatted and phylogenetic databases.

Availability: GDE-Linux is available for research purposes by direct application to the corresponding author. Application-specific menus and support files can be downloaded from (http://www.bioafrica.net).

Contact: toliveira@mrc.ac.za

INTRODUCTION

Recent advances have lead to an unprecedented increase in HIV-1 and other microbial sequence data. Most of the newly acquired information has been deposited in public repositories such as the Los Alamos HIV Sequence Database (http://hiv-web.lanl.gov), the Influenza Sequence Database (Macken *et al.*, 2001) and the Sanger (parasite-specific) Database (http://www.sanger.ac.uk/Projects/L_major/). These datasets constitute a rich source of sequence and epitope data for use in evolutionary and vaccine development studies.

New bioinformatics tools are needed to effectively manage, analyse and interpret these rapidly emerging datasets. Stand-alone software programs edit, align and perform phylogenetic analyses, but are often difficult to locate and use. In addition, installation of new programs is time consuming and may require extensive expertise to manipulate the computer's operating system, select the correct bioinformatics tools and construct an appropriate dataset.

To facilitate data mining, we developed a bioinformatics 'workbench' that combines the flexibility of the Genetic Data Environment (GDE) with the power of LINUX. GDE is a 'front-end' sequence analysis program originally developed for Sun UNIX^{TM} systems with an OpenView X Window manager (Smith *et al.*, 1994). GDE was subsequently adapted to Seqlab, a graphical user interface (GUI) that incorporates most of the software distributed in the GCG toolset (Womble, 2000). The main advantage of this software is its ability to 'wrap' around a wide range of programs and display the program output. LINUX, a re-implementation of UNIX^{TM}, is one of the most frequently-used operating systems in bioinformatics. It is portable (suitable for use with laptops and Intel 386-based systems) and supports clustering.

The combined GDE-LINUX reduces the complexity and repetitive nature of input/output formatting and facilitates development of user-defined local databases that are population- and/or disease-specific. Although easily adapted to any small-sized (10–20 Kb) sequence, GDE-LINUX is particularly useful for the diagnosis and evolutionary analysis of viral (HIV, HBV, HCV, HHV-8), bacterial (*M. tuberculosis*) and parasitic (*Schistosoma*, *Leishmania*) pathogens. Sequence databases and menu files are available at (http://www.bioafrica.net/GDElinux/GDEmicrobial.html).

SYSTEM AND METHODS

New software is integrated into GDE-L by editing a unique menu file that controls the menu appearance (.GDEmenus). This same menu file is responsible for sending commands and input files to applications such as BLAST, CLUSTAL W, PAML and readseq. Since the code for menu editing is similar to the code used to run the software command line, the incorporation of new functions is simple and requires knowledge of shell programming (cat, sed and variables). All formats

[*]To whom correspondence should be addressed.

supported by readseq sequence conversion software are readily accepted in GDE-L. If a software input format is supported, and the command line options and output are known, integration is easy.

The newly-created .GDEmenus files are then copied to a home directory. Accessory PERL scripts permit the integration of new sequence and BLAST-formatted datasets without editing the menu control file. The copying of pathogen- or disease-specific menu files allows each user to create his/her own custom-designed bioinformatics interface, update and change the interface and adapt it to his/her own research needs. Each new database gives rapid, local access to an entire set of Genbank sequences, BLAST-formatted and phylogenetic databases.

On-line tools have also been integrated into GDE-L. By using a web browser interface, investigators can access on-line software, send sequences to remote sites and search on-line databases (Genbank, PubMed). Researchers simply copy the sequence from GDE and paste it into the web browser. PERL scripts allow integration of novel web resources.

HIV-1 AS A MODEL INTERFACE

As with all studies of genetic diversity, sequence analysis of HIV-1 involves the use of multiple bioinformatics tools. Each new sequence must be edited, aligned and screened against previously amplified sequences to rule out cross-contamination between samples. The order of these steps is highly variable and dependent on the hypothesis to be tested. Studies may involve screening the *pol* gene for drug resistance, identifying biologically important variants circulating in a given population, searching for recombinants, monitoring the changing dynamics of subtype distribution, or identifying subtype-specific transmission patterns.

All of the sequence-specific databases, phylogenetic datasets and programs needed to study the diversity and molecular phylogeny of HIV-1 have been integrated into a single GDE-Linux interface. Local GenBank databases are accessed with XYLEM (Fristensky, 1993) and searched using either a FASTA or BLAST program. The BLAST dataset contains information on sequence location, contamination detection, and genetic subtype. Additional datasets, such as information on African HIV-1 strains, drug resistance and subgenomic regions are easily imported into the same interface.

Other bioinformatics programs installed in the HIV-1 interface include: sequence clustering (Phrap) and Stack Pack (Miller *et al.*, 1999); phylogenetic programs, including Phylip (Felsenstein, 1989); PAML (Yang, 1997); Bootscanning (Salminen *et al.*, 1995) for detecting recombination and TipDate (Rambaut, 2000) for measuring rates of evolution.

DISCUSSION

GDE-LINUX is an efficient, user-friendly interface that facilitates access to a broad spectrum of bioinformatics tools without the complexities of input/output file formatting between modules. The system is easily adapted to the needs of individual researchers, providing them with the expertise they need to analyse and annotate their own sequences. The novel use of scripts makes it easier to add, maintain and update these sequence-specific databases and to access on-line tools. The open source nature for distributing interface menus allows the scientific community to add new menus and applications as required. The ability to create highly specialized, population-specific databases, that can be easily analysed at the local level, eliminates the need for continual access to large international databanks. This is important in resource-constrained settings where bandwidth and access to high-speed internet are problematic. Our current research is focused on using GDE-L to better understand the epidemiological behaviour of HIV-1 in southern Africa, identify CTL escape mutants and develop novel vaccine and therapeutic strategies for the prevention and treatment of HIV/AIDS in developing countries.

REFERENCES

Felsenstein,J. (1989) PHYLIP — Phylogeny inference package (Version 3.2). *Cladistics*, **5**, 164–166.

Fristensky,B. (1993) Feature expressions: creating and manipulating sequence datasets. *Nucleic Acids Res.*, **21**, 5997–6003.

Macken,C., Lu,H., Goodman,J. and Boykin,L. (2001) The value of a database in surveillance and vaccine selection. In Cox,O.N. and Hampson,W.A. (eds), *Options for the Control of Influenza IV*. Elsevier Science, A.D.M.E Amsterdam, pp. 103–106.

Miller,R.T., Christoffels,A.G., Gopalakrishnan,C., Burke,J., Ptitsyn,A.A., Broveak,T.R. and Wide,W.A. (1999) A comprehensive approach to clustering of expressed human gene sequences: the sequence tag alignment and consensus knowledge base. *Genome Res.*, **9**, 1143–1155.

Rambaut,A. (2000) Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics*, **16**, 395–399.

Salminen,M., Carr,J., Burke,D. and McCutchan,F. (1995) Identification of breakpoints in intergenotypic recombinants of HIV-1 by Bootscanning. *ARHR*, **11**, 1423–1425.

Smith,S.W., Overbeek,R., Woese,C.R., Gilbert,W. and Gillevet,P. (1994) The genetic data environment an expandable GUI for multiple sequence analysis. *Comput. Appl. Biosci.*, **10**, 671–675.

Womble,D.D. (2000) GCG: The Wisconsin Package of sequence analysis programs. *Methods Mol. Biol.*, **132**, 3–22.

Yang,Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.*, **13**, 555–556.