Biotechvana

SOFTWARE ARTICLE. Biotechvana Bioinformatics: 2011-SOFT3. ISSN 1988-7957

GPRO: the professional tool for management, functional analysis and annotation of omic sequences and databases

Futami, R.[1], Muñoz-Pomer, A.[1,2], Viu, J.M.[1], Domínguez-Escribà, L.[1], Covelli, L.[1], Bernet, G.P.[1], Sempere, J.M.[2], Moya, A.[3,4], Llorens, C.[1]

1- Biotechvana, Parc Cientific de la Universitat de València
2- Departamento de Sistemas Informáticos y Computación (DSIC), Universitat Politècnica de València
3- Unidad Mixta de Investigación en Genómica y Salud del Centro Superior de Investigación en Salud Pública (CSISP)-Universitat de València (Instituto Cavanilles de Biodiversidad y Biología Evolutiva)
4- CIBER en Epidemiología y Salud Pública (CIBEResp)

Corresponding author: carlos.llorens@biotechvana.com

Availability: Available online February 28th, 2011 at http://www.biotechvana.com/software/gpro

Summary: In this article we present the first version of Gypsy Database PROfessional (GPRO 1.0) a software for annotation, data processing, management and analysis of DNA/RNA and protein databases (including host genes, repeats and mobile genetic elements).

Remarks: GPRO ia a standalone, installable multifunction software coupled online with a pipeline hosted at the Gypsy Database (GyDB) of Mobile Genetic Elements. The "software-pipeline" combination implements a worksheet-based annotation management system that lets users map and annotate multiple distinct sequences, simultaneously. Resulting annotations can be based on different standards such as RefSeq databases and commonly accepted ontology vocabularies (gene ontology, clusters of ortholog groups). The tool also implements a suite of software utilities to provide simplicity in diverse production tasks such as the creation, edition, analysis and management of sequences of up to two gigabases, annotation projects, and sequence databases.

Availability: GPRO is distributed by Biotech Vana S.L. at: [URL 1]. A 30 days free trial version is available.

Keywords: Bioinformatics | Computational biology

INTRODUCTION

Nowadays, the analysis of omic data (genomics, proteomics, metabolomics, etc.) is closely connected to the availability of comparative information stored in different online initiatives [1-7]. A variety of software applications and web servers (see for instance [8-12]) are available for massive analysis and annotation of biological information derived from omic projects by assigning function, taxonomy and biological categories to the newly-characterized sequences. Indeed, in the contemporary post-genomic biomedical era, advances in next generation sequencing technologies (NGS) have allowed projects to be undertaken in which it is possible to obtain and analyze thousands and millions of sequencing reads simultaneously. However, the management and editing of the multiple distinct files generated from these projects is still a daunting task. Thus, it is necessary to implement automatic pipelines and tool suites to generate protocols capable of massive analysis not only of genes/proteins but also of repeat variations (see [13-18]), mobile genetic elements (MGEs) [19-23], domains/modules [24-26], exon-intron segmentation [27], ontology [28-30] and complexity [31,32]. On the other hand, bioinformaticians usually spend a lot of needless time in concatenating distinct scripts usually designed *ad hoc* to automate the management and labeling of data files and sequence information contained therein. Taking these aspects of omic research into primary consideration, we have developed Gypsy Database PROfessional (GPRO) a software project for the annotation, data processing, management and functional analysis of DNA/RNA and protein databases. GPRO consists of installable multifunction software coupled online with an omic pipeline installed on a high-end computing server hosted at the Gypsy Database (GyDB) of Mobile Genetic Elements [1], enabling users to run intensive computation jobs in remote private sessions. The combination "software-pipeline" implements a powerful annotation management system that lets the users map, annotate and analyze multiple distinct sequences simultaneously using the most common tools ([33] and [URL 2]), vocabularies of gene ontology (GO) and ortholog (COG/KOG) classification [28,34,35], and mobile genetic element (MGE) databases [36,37]. In addition, GPRO implements a suite of tools providing simplicity and versatility to the management of files and folders. On the other hand, the tool also deals with the molecular analysis of DNA/RNA and protein sequences allowing sequences of up to two gigabases to be edited, translated and analyzed as well as retrieving ORFs and sequence motifs from edited sequences.

OVERVIEW

The GPRO is an hybrid between academic initiatives such as BLAST2GO [11] and commercial tools such as Geneious [URL 3]. It consists of installable multifunction software coupled online with an omic pipeline installed on a high-end computing server hosted at GyDB [1] to enable users to run intensive computing jobs in remote private sessions. There follows a quick overview of the distinct functions of the software (for more extensive and specific details, see the user guide accompanying the tool). Basically, GPRO is divided into four major components: PIPELINE, MENU, WORKSHEET and LAYOUT.

The PIPELINE (Figure 1) includes the following services or utilities:

1. 50GB hard disk space, which will increase periodically to guarantee enough computational space to fit the requirements of the most demanding projects.

2. A guaranteed quality of service and distributed CPU bandwidth for high-throughput computing analyses, which provides the user with the maximum available processing capacity on the cluster.

3. A user account in the remote computing cluster for running intensive computing analyses in private sessions.

4. A SSH client for logging into a user's private account on the remote computing cluster and sending commands for launching automated analysis tools.

5. An FTP client system organized as a remote file-tree manager for transferring analysis files between the client computer and the remote cluster user's account. Users can upload sequence files to be processed on the remote cluster and download the result files generated to their local computer.

6. A database compiler tool for BLAST [33] and HMMER [URL 2] servers.

7. A graphical front-end for launching BLAST and HMMER automated batch analyses using precompiled or user-generated databases. These tools can be launched in unattended mode, notifying the user by email when the job is finished.

8. A script for processing BLAST and HMMER result files in XML format for automated generation of a CSV (comma separated values) format report and its associated Fasta sequences according to an E (Expected) value cutoff defined by the user.

9. A script for retrieving ontology and taxonomical databases from the most common public servers to append annotation and functional analysis to the mapped sequences of your database project.

The MENU (Figure 2) is located at the top of the software interface and integrates the following commands:

1. DATABASES: this entry allows users to choose a specific custom-directory folder, or to open sequence files and databases (only in FASTA format), GenBank accessions, and worksheets for managing databases.

2. DIRECTORY: The software uses a special root folder called directory to manage the distinct folders and files of an annotation project (see "Layout" Section). Using this tab, users can show or hide the directory at the left of the GPRO's window.

3. EDITOR: this command launches two editing programs. One is a "Database (text) editor" associated with distinct utilities of the editing, mining and management of sequence database files. The other is an implementation of TIME [38], a sequence editor for displaying, analyzing and editing protein and nucleotide sequences of up to 2×10^9 bases (two gigabases or amino acids).

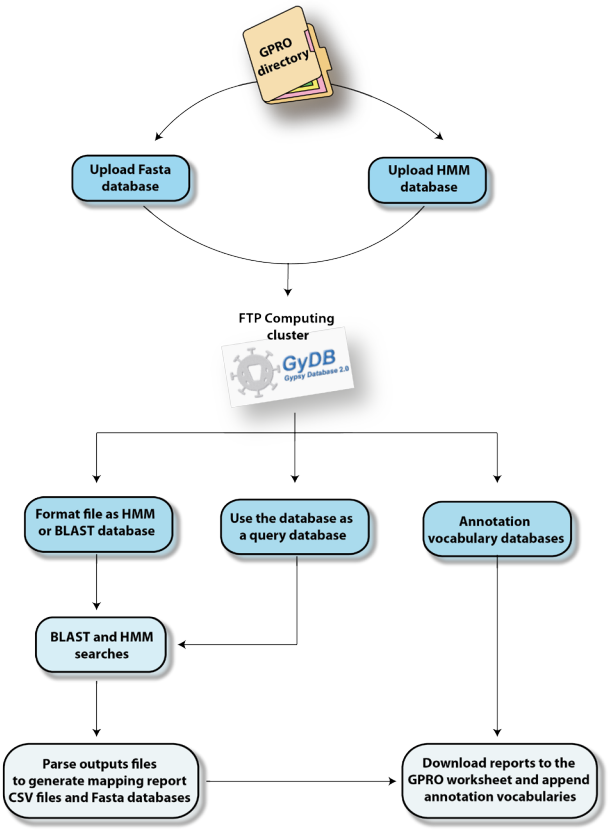

Figure 1. Pipeline flowchart

4. OMIC ANALYSES: this entry lets users exploit the pipeline for transferring sequence and HMM (Hidden Markov Model) database files from their computers to their accounts within the remote computing clustering in order to: format BLAST databases; perform BLAST/HMM searches against these or other databases; process the outputs into FASTA and annotation files.

5. ALIGNMENT ANALYSIS: Utilities for creating HMM profiles [39], Majority Rule Consensus (MRC) sequences and sequence logos [40] from the input of multiple sequence alignments.

6. MANAGEMENT: this option launches a suite of scripts to manage files and folders in different ways. For instance, users can join, split, and rearrange files, folders and their contents. Users can also execute specific data mining searches in these files and folders and then export the results to new files and folders.

7. PREFERENCES: for selecting the user's preferences with regard to diverse issues (FTP and pipeline connection, etc.).

8. HELP: menu entry for accessing the technical support service, the user guide, and the corporative information of GPRO.

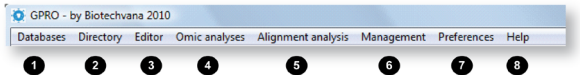

Figure 2. Menu tool bar and functions. Numbers indicate the eight tabs associated with each function as described in the text.

The WORKSHEET (Figure 3) is a grid of cells arranged in numbered rows (one for each sequence) and columns to provide information regarding mapping and/or functional analysis annotation (name, accession, function, size, species, E values, annotation vocabulary, etc.). Columns and rows are editable and can be rearranged by right-clicking over the columns. The worksheet implements a horizontal menu with diverse commands to run and handle annotation. There follows a description of each utility:

1. FILE: a drop-down menu allowing users to create and remove rows and files, and export databases from an annotation project on the basis of a selection of rows and/or columns.

2. SEARCH AND REPLACE: for searching and replacing terms in the worksheet.

3. SORTING/FILTERING: command to sort by ascending or descending order in a column, depending on statistical values and terms.

4. IMPORT: to join two or more annotation projects. For instance, users can import sequence clusters associated to a particular sequence (i.e. paralogs, orthologs, clades, etc.) or can also combine two worksheets to create a new database with common features (for instance, lineages of a particular virus).

5. EXPORT OPTIONS: to select which columns must be shown or hidden in the worksheet (the columns can be selected using the mouse).

6. ANNOTATION: to invoke the pipeline for appending functional and ontology terms to previously mapped sequences using annotation systems such as Gene Ontology (GO) [28] and Clusters of Ortholog Groups (COG and KOG) [34,35]. This tab also allows the user to switch between accessions and identifiers (IDs) provided by distinct institutions and classificatory initiatives such as the European Molecular Biology Laboratory (EMBL) [URL 4], GenBank [5], DNA Databank of Japan (DDBJ) [URL 5], Universal Protein Resource (UniProt) [6], InterPro [7], etc.

7. SELECT: to select and/or remove specific rows in a worksheet by using different selection criteria such as key terms, expected values and statistics, and grid color. Cells can be colored and arranged on the basis of distinct criteria (mapping, annotation, function, statistics, etc.).

8. ASSOCIATE DATABASE: users can also associate a sequence Fasta database with a related worksheet of reference to make common changes in both the worksheet and the database simultaneously.

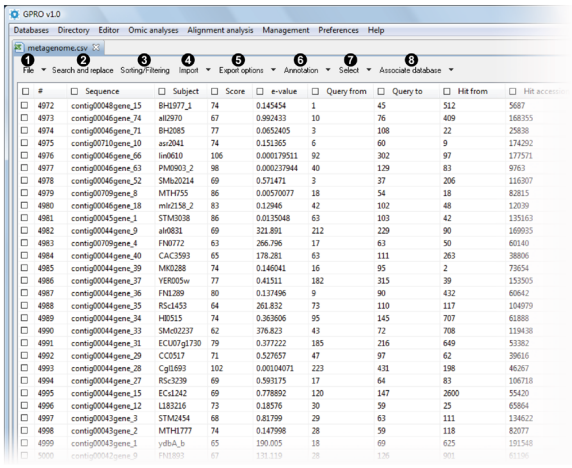

Figure 3. Worksheet screenshot. Numbers on the worksheet horizontal menu bar indicate the utilities that GPRO allows.

The LAYOUT (Figure 4) is organized in four intuitive window-based sections:

1. MAIN DESKTOP: central working space for editing files and databases, managing annotation projects and analyzing sequences.

2. DIRECTORY: users can define a major directory for storing databases and annotation projects. This directory can be shown (and hidden) to the left of the main desktop and organizes files and folders as a hierarchical file-tree that lets users visualize, select and drag any item from the Directory to other sections of the tool by simply using the mouse.

3. FTP: File Transfer Protocol (FTP) allows users to upload and download files and folders from the Directory to the remote user account for running the GPRO pipeline.

4. FASTA EXPLORER: this is a window-based utility coupled with the "Database Editor" that lets users have visual control and manage the names of the sequences in a text-edited database.

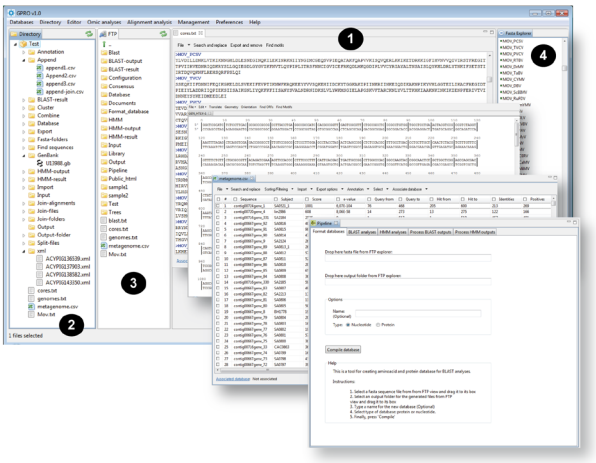

Figure 4. GPRO layout organization and interface implementation. Numbers indicate the four window-based sections as described in the text and displayed in the figure as follows: (1) Main desktop; (2) Directory; (3) FTP; (4) Fasta Explorer.

INSTALLATION

GPRO runs on personal computers and workstations as a standalone program. This tool is distributed as an installer for Windows XP/Vista/7 (32 bit and 64 bit), a self-extracting disk image for Mac OS X 10.5 or later (64 bit), and a compressed tarball archive for Linux 2.6 kernel series or later (32 bit and 64 bit).

REQUIREMENTS

GPRO requires Java 6 or later [URL 6]. The minimum system requirements for GPRO are a computer with a Pentium 4 1.5 GHz or AMD Athlon XP 1500+ processor or higher with at least 1 GB of RAM, although 4 GB are recommended. GPRO is coupled with a remote computational cluster for running omic analyses or for assigning functional annotation categories. To perform these tasks the tool requires an internet connection.

CONCLUDING REMARKS

In summary, GPRO makes data management easier and more productive. The current GPRO version offers a wide array of analytical tools for annotation manipulation, data mining and sequence tagging. However, GPRO is a tool in continuous progress. We are preparing a new release (version 1.1) in which we will address the implementation of new commands oriented to evaluate reads' quality files provided by the distinct available technologies (Illumina, Roche 454, Helicos, Sanger, Solid) and other functionalities such as: (a) statistics and graphic representations; (b) a genome browser for superimposing predictive models over sequences; (c) learning algorithms to concatenate Open Reading Frame (ORF) repeats, and MGE and virus features (cassettes and domains) to automate the annotation of intron-exon organized genes, multidomain proteins, full-length MGEs and endogenous viruses. Upon this new implementation, we are quite receptive of users' feedback. So if you have any suggestion about new commands and utilities you think might be of general interest when working with omic data, please do not hesitate let us to know. You can make the comments using the "Suggestions and bugs" form available in the "Help" tab of the main GPRO menu and send them to us.

ACKNOWLEDGMENTS

GPRO 1.0 has been partly supported by Grants IDI-20100007 from CDTI (Centro de Desarrollo Tecnológico Industrial) and PTQ-09-01-00020 and PTQ-09-01-00670 from MICINN (Ministerio de Ciencia e Innovación) in Spain.

Funding to pay the Open Access publication charges for this article was provided by the University of Valencia

LICENSE AND DISTRIBUTION

LITERATURE

1. Llorens C, Futami R, Bezemer D, Moya A: The Gypsy Database (GyDB) of Mobile Genetic Elements. *Nucleic Acids Research (NAR)* 2008, **36**: 38-46.

2. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J: Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* 2005, **110**: 462-467.

3. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV et al.: The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 2003, **4**: 41.

4. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM et al.: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000, **25**: 25-29.

5. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: GenBank. *Nucleic Acids Res* 2010, **38**: D46-D51.

6. Uniprot Consortium: The universal protein resource (UniProt). *Nucleic Acids Res* 2008, **36**: D190-D195.

7. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D et al.: InterPro: the integrative protein signature database. *Nucleic Acids Res* 2009, **37**: D211-D215.

8. Al-Shahrour F, Diaz-Uriarte R, Dopazo J: FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* 2004, **20**: 578-580.

9. Kuzniar A, Lin K, He Y, Nijveen H, Pongor S, Leunissen JA: ProGMap: an integrated annotation resource for protein orthology. *Nucleic Acids Res* 2009, **37**: W428-W434.

10. Milone DH, Stegmayer GS, Kamenetzky L, Lopez M, Lee JM, Giovannoni JJ et al.: *omeSOM: a software for clustering and visualization of transcriptional and metabolite data mined from interspecific crosses of crop plants. *BMC Bioinformatics* 2010, **11**: 438.

11. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M: Blast-2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 2005, **21**: 3674-3676.

12. Lewis SE, Searle SMJ, Harris N, Gibson M, Iyer V, Richter J et al.: Apollo: a sequence annotation editor. *Genome Biology* 2002, **3**: research0082.

13. Ambrosini A, Paul S, Hu S, Riethman H: Human subtelomeric duplicon structure and organization. *Genome Biol* 2007, **8**: R151.

14. Tan JC, Tan A, Checkley L, Honsa CM, Ferdig MT: Variable Numbers of Tandem Repeats in Plasmodium falciparum Genes. *J Mol Evol* 2010, **71**: 268-278.

15. Gemayel R, Vinces MD, Legendre M, Verstrepen KJ: Variable Tandem Repeats Accelerate Evolution of Coding and Regulatory Sequences. *Annu Rev Genet* 2010.

16. Davison J, Tyagi A, Comai L: Large-scale polymorphism of heterochromatic repeats in the DNA of Arabidopsis thaliana. *BMC Plant Biol* 2007, **7**: 44.

17. Okada K, Yamazaki Y, Yokobori S, Wada H: Repetitive sequences in the lamprey mitochondrial DNA control region and speciation of Lethenteron. *Gene* 2010, **465**: 45-52.

18. Ball EV, Stenson PD, Abeysinghe SS, Krawczak M, Cooper DN, Chuzhanova NA: Microdeletions and microinsertions causing human genetic disease: common mechanisms of mutagenesis and the role of local DNA sequence complexity. *Hum Mutat* 2005, **26**: 205-213.

19. Kapitonov VV, Jurka J: A universal classification of eukaryotic transposable elements implemented in Repbase. *Nat Rev Genet* 2008, **9**: 411-412.

20. Eickbush TH, Jamburuthugoda VK: The diversity of retrotransposons and the properties of their reverse transcriptases. *Virus Res* 2008, **134**: 221-234.

21. Glockner G, Szafranski K, Winckler T, Dingermann T, Quail MA, Cox E et al.: The complex repeats of Dictyostelium discoideum. *Genome Res* 2001, **11**: 585-594.

22. Gonzalez J, Petrov D: Genetics. MITEs--the ultimate parasites. *Science* 2009, **325**: 1352-1353.

23. Craig NL, Craigie R, Gellert M, Lambowitz AM: Mobile DNA II. Washington, DC.: ASM Press; 2002.

24. Apic G, Russell RB: Domain recombination: a workhorse for evolutionary innovation. *Sci Signal* 2010, **3**: e30.

25. Malik HS, Eickbush TH: Modular evolution of the integrase domain in the Ty3/Gypsy class of LTR retrotransposons. *J Virol* 1999, **73**: 5186-5190.

26. Lang A, Szilagyi K, Major B, Gal P, Zavodszky P, Perczel A: Intermodule cooperativity in the structure and dynamics of consecutive complement control modules in human C1r: structural biology. *FEBS J* 2010, **277**: 3986-3998.

27. Lynch M: The origins of eukaryotic gene structure. *Mol Biol Evol* 2006, **23**: 450-468.

28. The Gene Ontology Consortium: Gene ontology: tool for the unification of biology. *Nat Genet* 2000, **25**: 25-29.

29. Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V et al.: Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2009, **37**: D5-15.

30. Chen F, Mackey AJ, Stoeckert CJ, Jr., Roos DS: OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res* 2006, **34**: D363-D368.

31. Capy P: Evolutionary biology. A plastic genome. *Nature* 1998, **396**: 522-523.

32. Lynch M, Conery JS: The origins of genome complexity. *Science* 2003, **302**: 1401-1404.

33. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W et al.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997, **25**: 3389-3402.

34. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS et al.: The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res* 2001, **29**: 22-28.

35. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV et al.: The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 2003, **4**: 41.

36. Llorens C, Muñoz-Pomer A, Futami R, Moya A: The GyDB Collection of Viral and Mobile Genetic Element Models. In Biotechvana Bioinformatics. Biotechvana,Valencia; 2009:CR: GyDB Collection.

37. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J: Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* 2005, **110**: 462-467.

38. Muñoz-Pomer A, Futami R, Covelli L, Dominguez-Escriba L, Bernet GP, Sempere JM et al.: TIME a sequence editor for the molecular analysis of DNA and protein sequence samples . *Biotechvana Bioinformatics*: 2011-SOFT2 2011.

39. Eddy SR: Profile hidden Markov models. *Bioinformatics* 1998, **14**: 755-763.

40. Schneider TD, Stephens RM: Sequence Logos - A New Way to Display Consensus Sequences. *Nucleic Acids Research* 1990, **18**: 6097-6100.

URLS

1. GPRO Web Site: http://www.biotechvana.com/software/gpro
2. HMMER3: http://hmmer.janelia.org
3. GENEIOUS: http://www.geneious.com
4. EMBL: http://www.ebi.ac.uk/embl
5. DDBJ: http://www.ddbj.nig.ac.jp
6. JAVA language: http://www.java.com