BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 8 2012, pages 1166–1167
doi:10.1093/bioinformatics/bts091

Genome analysis Advance Access publication February 24, 2012

Unipro UGENE: a unified bioinformatics toolkit

Konstantin Okonechnikov*, Olga Golosova, Mikhail Fursov and the UGENE team
Unipro Center for Information Technologies, 6/1 Lavrentyev avenue, Novosibirsk, Russia, 630090

Associate Editor: John Quackenbush

ABSTRACT

Summary: Unipro UGENE is a multiplatform open-source software with the main goal of assisting molecular biologists without much expertise in bioinformatics to manage, analyze and visualize their data. UGENE integrates widely used bioinformatics tools within a common user interface. The toolkit supports multiple biological data formats and allows the retrieval of data from remote data sources. It provides visualization modules for biological objects such as annotated genome sequences, Next Generation Sequencing (NGS) assembly data, multiple sequence alignments, phylogenetic trees and 3D structures. Most of the integrated algorithms are tuned for maximum performance by the usage of multithreading and special processor instructions. UGENE includes a visual environment for creating reusable workflows that can be launched on local resources or in a High Performance Computing (HPC) environment. UGENE is written in C++ using the Qt framework. The built-in plugin system and structured UGENE API make it possible to extend the toolkit with new functionality.

Availability and implementation: UGENE binaries are freely available for MS Windows, Linux and Mac OS X at http://ugene.unipro.ru/download.html. UGENE code is licensed under the GPLv2; the information about the code licensing and copyright of integrated tools can be found in the LICENSE.3rd_party file provided with the source bundle.

Contact: ugene@unipro.ru

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on October 16, 2011; revised on January 28, 2012; accepted on February 19, 2012

1 INTRODUCTION

The number of available bioinformatics tools has grown dramatically over the past two decades. However, as a result of increasing complexity of the tools and a lack of rigorous software engineering practices, using, installing and configuring a bioinformatics application, along with ensuring that it is communicating well with other programs, can be a challenging task. Moreover, a large number of computational methods and tools are available only as source-code packages, command-line utilities which can run only in a Linux/Unix environment, web-services (Stockinger *et al*., 2008), or programming libraries (Mangalam, 2002), and cannot be easily accessed by biologists who do not have the necessary skills. A number of existing tools also lack user documentation and have limited support.

*To whom correspondence should be addressed.

Some solutions have been proposed to improve the situation. There are bioinformatics tools which are multiplatform, easy to install and hide the underlying complexity behind an easy-to-use GUI interface. Usually, these tools are focused on a particular niche. Good examples of such tools are the Artemis genome browser (Rutherford *et al.*, 2000), Jalview multiple alignment editor (Waterhouse *et al.,* 2009), Tablet (Milne *et al.*, 2010) and IGV (Robinson *et al.*, 2011) next-generation sequencing data visualization packages. Applications such as Galaxy (Goecks *et al.*, 2010) or Taverna (Hull *et al.*, 2006) solve communication issues between tools by providing easily accessible capabilities to manage bioinformatics workflows, but still require significant effort to set up on a local server. There are also closed-source commercially available programs, such as VectorNTI (www.invitrogen.com), Geneious (www.geneious.com) or CLCBio (www.clcbio.com) that could act as a bioinformatics 'Swiss army knife'.

In this article, we introduce Unipro UGENE, a multiplatform open-source application, which integrates popular bioinformatics algorithms and tools, providing both graphical and command-line interfaces. It has an active user community and support from a team of professional software developers. To perform complex computational experiments, UGENE allows the construction of reusable workflow diagrams.

2 IMPLEMENTATION

UGENE currently supports reading and writing in >20 popular biological data formats. Among them are FASTA, GenBank, Clustal, GFF and SAM/BAM. It is capable of working with local files and accessing data from key biological online databases including NCBI GenBank, UniProt and PDB. To simplify management of biological datasets, UGENE uses the concept of a project, an abstract structure that contains information about files and visualization settings. A project can be saved or exported with all data to a specified location and restored at any time. By using the point-and-click visual interface, one can manipulate groups of items in the project.

The toolkit allows the visualization of biological objects such as annotated DNA/RNA and protein sequences, multiple sequence alignments, phylogenetic trees, chromatograms, macromolecular 3D structures and NGS assemblies. All visualization components use common principles in providing functionality for users, and allow export of publication-quality images. Most of the visualization components allow editing of biological objects. For example, the sequence viewer allows visualizing and editing of genomes in linear and circular modes, and includes an interactive annotations editor (Fig. 1).

UGENE incorporates a large library of computational methods. The library currently consists of >30 methods and is constantly growing. The integrated tools and algorithms solve a variety of

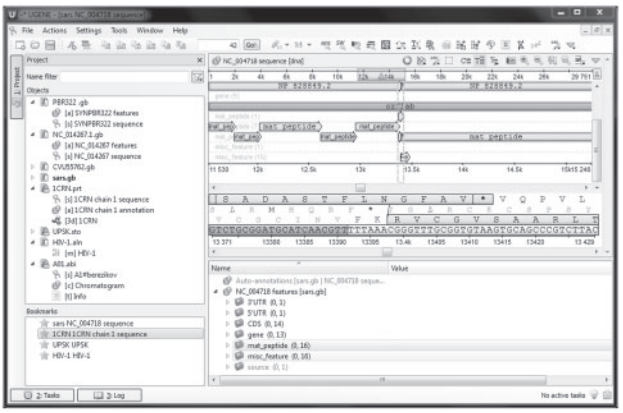

Fig. 1. The project view and sequence viewer, showing GenBank sequence NC_004718 (available from the samples provided with UGENE).

bioinformatics tasks that include a pattern search, local sequence alignment, search for repeats, multiple sequence alignment, HMM profile tools, restriction sites analysis, primer design, short read alignment and many others.

There are two kinds of UGENE tools: those directly integrated into the UGENE application and external tools. To use an external tool, it must be available as a binary executable. All needed binaries can be downloaded from the UGENE website as a separate package or as a part of the UGENE Full Package.

A key advantage of UGENE is that all of the included algorithms are adapted to use the internal UGENE data model. This allows one to avoid manual data conversion between the tools' input and output. The algorithm library consists not only of popular tools and methods such as MUSCLE (Edgar, 2004) or Bowtie (Langmead *et al.*, 2009), but also includes unique computational methods developed by the UGENE team or contributed by other researchers. New versions of third-party tools are tracked by UGENE developers and, typically, updates are applied in every major release. A test base containing >5000 automated tests helps to avoid potential regressions in the integrated code.

Some of the integrated algorithms are adapted to utilize multicore systems, special processor instructions and GPUs. For example, the integrated Smith–Waterman algorithm, developed by the UGENE team, has the multi-threaded and vectorized version for multicore processors and the GPU-optimized version. All optimized algorithms are included in the standard installation package and accessible for users 'out-of-the-box'.

One of the key components of UGENE is the Workflow Designer, a visual tool for building complex analysis pipelines. As UGENE is a standalone application, one does not need to install any additional components or upload and download any data to use the Workflow Designer. This factor, along with the intuitive and user-friendly interface, reduces the entry threshold for new users. For advanced users, the Workflow Designer provides capabilities to create custom workflow elements in C++, QtScript programming languages or from any external application by customizing the input and output with the step-by-step wizard. Every user-designed workflow can be run from the command line. This feature makes it easy to run workflows in an HPC environment, and/or incorporate them into users' scripts. A more detailed overview of the Workflow Designer's features and its comparison to Taverna and Galaxy can be found in Section S1 in the Supplementary Materials. The Workflow Designer file format is described in Section S2 in the Supplementary Materials.

Workflows constructed using the Workflow Designer can be offloaded to a remote computational resource running UGENE Remote Service—a custom HPC facility. There is a publicly available beta of the service utilizing the Amazon EC2 cloud. More details are available in Section S3 in the Supplementary Materials.

UGENE is written in C++ using the Qt4 library. The toolkit has a highly modular structure and a built-in plugin system. Therefore, its components can be reused in other applications, and UGENE can be extended with new functionality. The Unipro UGENE project was started in 2008, and has already gathered a strong community of users. According to our statistics, the UGENE binary packages for different operating systems are downloaded ∼1000 times per month. Owing to the efforts of the community, the toolkit has been integrated into major Linux distributions: Ubuntu, Fedora, Debian (Möller *et al.*, 2010) and others. New features suggested by users have the highest priority in the project road map. For better interaction with the community, there is an open bug-tracking system (https://ugene.unipro.ru/tracker) and a forum.

ACKNOWLEDGEMENTS

Unipro UGENE is developed by the UGENE team: Alexey Varlamov, Yuri Vaskin, Ivan Efremov, German Grehov, O.G., Denis Kandrov, Kirill Rasputin, Maxim Syabro, Timur Tleukenov and M.F. We are grateful to all UGENE users for their suggestions and feedback. Special thanks to Agustín Ure for his active participation in the project development.

Conflict of Interest: none declared.

REFERENCES

Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

Goecks,J. *et al.* (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible and transparent research in the life sciences. *Genome Biol.*, **11,** R86.

Hull,D. *et al.* (2006) Taverna: a tool for building and running workflows of services. *Nucleic Acids Res.*, **34**, W729–W732.

Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

Mangalam,H.J. (2002) The Bio* Toolkits—a brief overview. *Brief. Bioinform.*, **3**, 296–302.

Milne,I. *et al.* (2010) Tablet—next generation sequence assembly visualization. *Bioinformatics,* **26**, 401–402.

Möller,S. *et al.* (2010) Community-driven computational biology with Debian Linux. *BMC Bioinformatics*, **11** (Suppl. 12), S5.

Robinson,J.T. *et al.* (2011) Integrative genomics viewer. *Nat. Biotechnol.,* **29**, 24–26.

Rutherford,K *et al.* (2000) Artemis: sequence visualization and annotation. *Bioinformatics,* **16**, 944–945.

Stockinger,H. *et al*. (2008) Experience using web services for biological sequence analysis. *Brief. Bioinform.*, **9**, 493–505.

Waterhouse,A.M. *et al.* (2009) Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics,* **25**, 1189–1191.