UGENE toolkit: Bioinformatics “Swiss army knife”

Yuriy Vaskin

Novosibirsk State University, Novosibirsk, Pirogova St. 2, 630090, vaskin90@gmail.com

M. Fursov

NCIT Unipro, Novosibirsk, Lavrentieva ave. 6/1, 630090, mfursov@unipro.ru

In the age of information technologies it is not surprising that biologists use special computer programs in their day-to-day work. Some of the programs are huge systems like Geneious (www.geneious.com) designed to help in solving a large number of tasks. Some of them are just small scripts like SAMtools (http://samtools.sourceforge.net/) which can only filter files of a particular format. Diversity of biological problems is the reason of such a variety in software. For instance, in the fields of Next Generation Sequencing biologists have a multistep and complex pipeline and gigabytes of data they want to process or even visualize. In that case a computer program, which implements that pipeline, must be aware of such an amount of data and it must take advantage of computational resources available for a biologist. On the other hand, the biologist can go down to nucleotides and construct molecular vectors or analyze specific genomic regions in the context of multiple sequence alignment. These tasks require from software good visualization capabilities and convenient user interface. But all these techniques just show the same biological objects but at different angles. Thus, computer programs should share that idea by providing unified workspaces with tools that are linked together.

UGENE is a multiplatform free open-source toolkit which integrates popular algorithms and tools with graphical and command-line interfaces [1]. All the instruments share similar interfaces, data structures and logic. The unified toolkit covers a wide range of biological tasks: sequence alignment, functional annotation of sequences, phylogeny, NGS data processing, genome assembly, etc. For advanced data analysis UGENE provides Workflow Designer to build computational schemes.

Currently there are a lot of data formats and types which are spread throug various biological databases. UGENE supports reading and writing in more than 30 data formats. Users can open local files or download their data from remote databases like GenBank, PDB, etc. Using the plugin system of web-browsers UGENE can recognize items on BioMart pages then automatically download and open the items.

The toolkit provides different visualization capabilities for different data types with features and algorithms specific for the type. Among these types are DNA/RNA/protein sequences, multiple alignments, 3D protein structures, phylogenetic trees and NGS data. For instance, in a special window, sequences of any size can be viewed, edited and annotated with elements like HMM signals, TFBSs, repeats, restriction sites, etc. While in another window of Assembly Browser biologists can instantly navigate their assembly data with a full coverage graph.

Often biological data analyses involve multistep processes which can be automated. UGENE Workflow Designer implements that idea of computational pipelines. A scheme can be created with a few mouse clicks and each block of the scheme is a computational algorithm that might be an optimized one which takes advantage of user's hardware. Such schemes can be saved in files and shared with other users. Workflow Designer includes various blocks: input/output, algorithmic, data filtering/multiplexing and special types of blocks that can represent an external tool or a user defined script. In the library of sample pipelines there are schemes designed to perform common tasks and which implement well-known NGS data pipelines like Cistrome[2], Tuxedo[3], SAMtools variant calling[4]. There is also an option to stop a computational process and check medium data.

1. K. Okonechnikov, O. Golosova, M. Fursov, the UGENE team (2012) Unipro UGENE: a unified bioinformatics toolkit, *Bioinformatics,* **28**: 1166-1167.

2. T. Liu, J. A. Ortiz1. (2011) Cistrome: an integrative platform for transcriptional regulation studies, *Genome Biology,* **12**: R83.

3. C. Trapnell, A. Roberts (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks, *Nature Protocols,* **7**: 562–578.