Shared bioinformatics database within Unipro UGENE

Ivan Protsyuk[1], **Mikhail Fursov**[2]
iprotsyuk@unipro.ru, mfursov@unipro.ru
[1]Novosibirsk State University, Novosibirsk, Russia
[2]Center of Information Technologies "UniPro", Novosibirsk, Russia

Unipro UGENE [Okonechnikov et al. 2012] is an open-source bioinformatics toolkit that integrates popular tools along with original instruments for molecular biologists within a unified user interface. It has grown to a platform that can cope with a large variety of tasks including multiple alignment, phylogeny, functional annotation, *in silico* cloning, protein structure analysis and TFBSs recognition. To perform various types of analysis UGENE gathers computational algorithms, visualization capabilities and advanced workflows.

Nowadays most bioinformatics desktop applications, including UGENE, make use of local data models when processing different types of data. Namely, a user can work with files stored on his/her local machine only. Such an approach may cause inconvenience to scientists working cooperatively and relying on the same data. The most obvious issue arises from the need to make multiple copies of certain resources for every workplace. After that the problem of synchronization comes into play. There are tools that provide the needed capabilities, for instance, CLC Bioinformatics Database or Geneious but they are proprietary and quite expensive.

Recently we focused on delivering collaborative work into the UGENE user experience. Currently several UGENE installations can be connected to a designated shared database and use it simultaneously as if it was an ordinary file possibly containing a huge amount of data. Objects of each data type supported by UGENE such as sequences, annotations and multiple alignments can now be easily imported from or exported to remote storage. The storage itself is a regular database server (so far only MySQL is supported) available for external connections so even an inexpert user can deploy it. All its data are stored using a fully relational approach within a self-developed database schema. Thus, UGENE is able to provide access to shared data for a few users located, for example, in the same lab or institution.

Furthermore, UGENE maintains integrity with its visualization capabilities as well. Not only can data be transferred via the database between computers, but it also may be displayed and modified permanently just as if it were located in a normal file. Thereby smooth transition is achieved for the end-user between his local storage and the shared one.

This work itself presents a basis for further development in various directions. First, UGENE Workflow Designer coupled with a shared database may operate remotely and do jobs for clients aiming to process the shared data. Therefore, this opens an opportunity for using easy-to-install computational clusters requiring the UGENE suite only. Another feature is the saving of full user contexts between successive UGENE launches to an embedded local database. That implies tracking and permanent storing of states for all the open files, views and databases involved in a user session. In this way, UGENE may improve the convenience and productivity of a biologist's workplace even more.

The first version of the UGENE tool, supporting shared databases, is going to be released by the end of June 2014.

Project Web Site: http://ugene.unipro.ru/
Software and source code: http://ugene.unipro.ru/download.html
License: GNU General Public License v.2