DOI 10.18699/PlantGen2019-168

Unified automated information system for the formation of highly structured plant genomes and proteomes

Shlikht A.*, Kramorenko N.
Far Eastern Federal University, Vladivostok, Russia
** e-mail: schliht@mail.ru*

Formation of new genomes requires creation of effective ways of extraction, storage, analysis and interpretation of omics data. Currently existing file systems with scripting languages require a fairly deep knowledge of bioinformatics, which makes it difficult to use such systems are not bioinformatics. The report considers the unified automated information system for the formation of highly structured representations of plant genomes and proteomes based on the technology of databases and knowledge bases. The developed system is equipped with an intelligent interface to work with omics data on the basis of semantically understandable terms (DNA, genes, transcripts, exons, introns, mRNA, proteins, metabolites, reactions, metabolic pathways, etc.), forming ontological knowledge of the subject area, and does not require programming knowledge. The system is based on the restructuring of the primary omics data of the world portals (NCBI, EnsemblPlants, etc.), represented by text formats (FASTA, TXT, XML), indexed highly structured database format with storage on a local server or personal computer. The system can also work with experimental data obtained during genome sequencing and/or proteome mass spectrometry. The restructuring and local storage of data, in turn, ensure the system's autonomous operation and high performance. Periodically, the data is updated in an automated mode with connection to the world portal. The system has extensive functionality of working with omics data: access DNA, genes, transcripts, proteins up to nucleotide or amino acids with their coordinates; search and transformation omics data; modeling of signaling and metabolic pathways; the search for motifs of transcription factors; modeling of mutations at the level of genes, transcripts, proteins. Setting up the system for a new genome is carried out in an automated mode and includes the following stages: loading file data from ftp-portals; restructuring and indexing the loaded omics data format of databases; coding data; automatic translation of genes into proteins. Minimum memory requirements: RAM from 4 GB, disk memory of 500 GB. The system can be used for both research and education.