Automated workplace of bioinformatics

Shlikht, A.G., Kramorenko N.V.

Far Eastern Federal University, 690950, Vladivostok, Suhanova St., 8, schliht@mail.ru

Presents Automated biomedical information system functioning as the Automated workplace (AWP) of bioinformatics, allows you to automatically simulate and analyze various situations related to the human genome.

Traditional interactive representation of the genome in a text format (FASTA, FASTQ) is not always convenient for the analysis of molecular omics data. In particular, information on a reference genome is available via the Internet poorly structured and do not provide a comprehensive automatic analysis of a set of genes, transcripts, exons, introns and thus excluded aggregations of data across the genome.

The AWP information about the genome is converted into a highly structured relational data format that allows for the automatic mode of operation with the data. Primary data from portals NCBI [1], KEGG [2] (dbSNP, CliVar, OMIM, etc), presented in the form of unstructured text files, human chromosomes, FTP data file have been converted to indexed relational structured presentation with the numbering of each nucleotide in the chromosomes. Moreover in the database (DB) were entered the coordinates and parameters of genes, exons, introns on the basis of data of the reference genome NCBI. Note that it always remains in the DB hyperlink on the primary interactive data portals. The obtained representation of the primary data allowed us to use the power of database technology and knowledge to transformed raw data of the human genome. The representation in the database format made it easy to simulate various structural changes in genes, transcripts, introns, exons, automatically getting various modifications and isoforms of proteins, their charge, mass, primary and secondary structure. Based on this technology automatically solve problems related to the analysis of DNA, RNA, genes, exons, introns, mutations, diseases, construction of genetic networks, protein-protein interaction. The transition from text representation to a relational index performance demanded a significant increase in memory, but gave a big gain in time and allowed to pass to the automatic mode of operation [3] and to solve problems that cannot be solved interactively. Automatically linkages and relationships between genes, proteins, their corresponding enzymes and reactions and metabolites that are supported by these enzymes. All the enzymes, reactions, metabolic pathways and metabolites are also represented in the relational data format.

A relational presentation also allows you to automatically produce comprehensive statistical analysis of non-coding parts of the genome, in particular the analysis of promoter parts for all genes simultaneously, motifs in introns, search numerous repetitions in non-coding parts of the genome. Automatically search patterns across the genome and the issuance of a research protocol. Such tasks cannot be solved with interactive access to world portals. Of course, the primary data holders are also able to perform similar studies as part of their data structures, but, unfortunately, no access to structured information and primary data, which excludes the possibility of conducting a comprehensive research.

The data model developed by the DB contains the following main objects: the version of the genome, DNA, chromosome, gene, transcript, exon, intron, a nucleotide, polymorphism, messenger RNA, amino acid, protein, direct and inverse spiral, biomarker, primer, promoter, motif, enzyme, reaction, metabolite, metabolic path, a peptide, disease, etc. Associative relationships between objects: genome - chromosome, chromosome - gene, gene - transcript, transcript - exon/intron, gene - promoter, DNA - motif, gene - gene, gene - protein, protein - protein, gene - enzyme, enzyme - reaction, enzyme - coenzyme, reaction - metabolic pathway, reaction - metabolite, gene - polymorphism, disease - mutation, marker - primer, marker - disease, etc.

The nucleotide composition of each chromosome are presented in separate tables with the records of nucleotides, with key – number of the nucleotide in a chromosome. The formation is invariant to a specific chromosome (table) the query is generated via an SQL-query that provides a unified analysis of any chromosome. Dictionary of genes created throughout the genome with reference to the corresponding chromosome and contains the coordinates of the gene. Through a system of views and stored procedures are being implemented numerous analytical tasks in genome: formation of proteins on the basis of the chain (chromosome – gene – transcript – exons – splice – messenger RNA – protein); automatic modeling of various mutant situations in the genome, transcriptome, exomics, proteome, followed by statistical analysis of intermediate and final results; analysis of unencrypted parts of genes, introns; finding and determination of the coordinates of motives; the definition of the primary and secondary structure of proteins, charge, mass and other properties; the formation of the stoichiometric matrix chemical reactions of metabolic processes, etc.

Our approach is an alternative to the portals on the basis of large data centers to a wider audience. Our system is designed for a small audience and is implemented on less powerful computers. Moreover, our approach allows to restructure the primary data and further analysis of them by their own algorithms without recourse to the portals. The important point is updating the data as the primary data portals are constantly updated with varying frequency. To update the data developed special programs running in the background automatic mode. These programs allow us to make fundamental changes on primary data.

The AWP bioinformatics is implemented on client-server technology with many visual forms, reflecting the results of research. The whole AWP with the database can reside on a personal computer, allowing you to use it as a portable tool, bioinformatics, physician, genetics research, and practical training purposes.

1. Website NCBI. URL: http://www.ncbi.nlm.nih.gov/

2. Website KEGG. URL: http://www.genome.jp/kegg/

3. А.Г.Шлихт, Н.В.Краморенко (2012) Интегрированная биоинформационная система поддержки постгеномных исследований, мониторинга среды обитания и здоровья человека, *Биоинформатика и молекулярное моделирование, сборник трудов I международной Интернет-конференции, Казань*, 69–73.