Assembly and annotation pipeline for microorganism genomes from sequence to gene networks

Demenkov P.S.[1,2]*, Lashin S.A.[1,2], Ivanisenko V.A.[1,2]
[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
[2] Kurchatov Genomic Center of the Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
* demps@bionet.nsc.ru

Acknowledgments

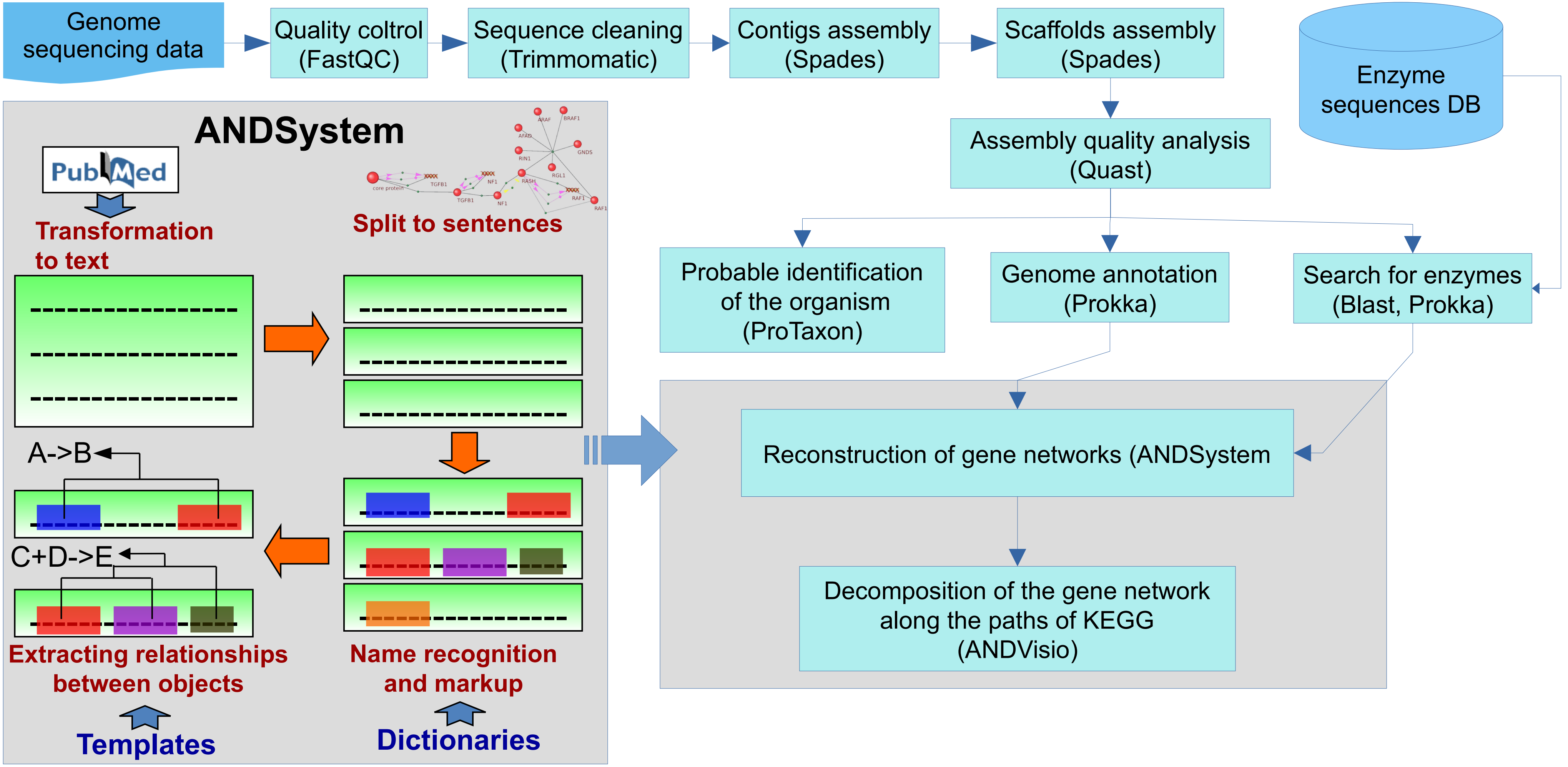
Work was funded by the Ministry of Science and Higher Education of the Russian Federation project “Kurchatov Center for World-Class Genomic Research” number 075-15-2019-1662 from 2019-10-31.

Introduction

Microorganisms of natural and technogenic ecosystems represent an inexhaustible pool of metabolic pathways for the utilization of some compounds and the biosynthesis of others. Modern experimental technologies of genetics and microbiology provide an effective search in microbiological collections for new strains of microorganisms promising for biotechnological applications, as well as full sequencing of their genomes. As a result of experimental studies, large amounts of genetic data are generated. Their analysis requires the involvement of modern information and computer technologies, which make it possible, on the basis of genomic information, to reconstruct the gene networks that control the production of target substances.

This poster proposes a software pipeline for assembly, annotation of sequenced genomes of microorganisms and reconstruction of their gene networks.

Software pipeline for assembly, annotation of sequenced genomes of microorganisms and reconstruction of their gene networks.



Genome sequencing data

Quality coltrol (FastQC)

Sequence cleaning (Trimmomatic)

Contigs assembly (Spades)

Scaffolds assembly (Spades)

Enzyme sequences DB

Assembly quality analysis (Quast)

Probable identification of the organism (ProTaxon)

Genome annotation (Prokka)

Search for enzymes (Blast, Prokka)

ANDSystem

PubMed

Transformation to text

Split to sentences

Name recognition and markup

Extracting relationships between objects

A->B

C+D->E

Templates

Dictionaries

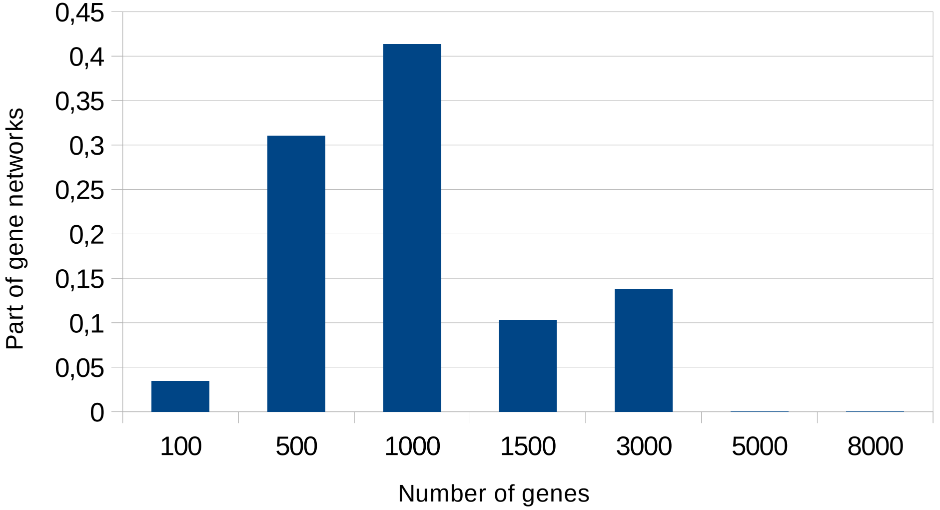
Reconstruction of gene networks (ANDSystem

Decomposition of the gene network along the paths of KEGG (ANDVisio)

Results

The developed pipeline was used on the data of the collection of microorganisms of the Kurchatov Genomic Center. To date, 1005 genomes have been processed. The average genome length was 6,045,404 bp. The average number of genes in the genome is 5545. KEGG gene networks and metabolic pathways were reconstructed for all analyzed genomes.

Distribution of gene networks by the number of genes Involved in them for sequenced microorganisms



Distribution of the number of KEGG pathways according to the proportion of enzymes found in the B-2403 genome among pathway participants