# Unipro UGENE

# **Unipro UGENE Manual**

Version 1.30

May 28, 2018



# **Unipro UGENE Online User Manual**

- About Unipro
  - About UGENE
    - Key Features
    - User Interface
    - High Performance Computing
- Cooperation Download and Installation
  - System Requirements
  - UGENE Packages
  - Installation on Windows
  - Installation on Mac OS X
  - Installation on Linux
    - Native Installation on Ubuntu
      - Native Installation on Fedora
- Basic Functions
- UGENE Terminology
  - UGENE Window Components
    - Welcome Page
      - Project View
      - Task View
      - Log View
      - Notifications
  - Main Menu Overview
  - **Creating New Project**
  - Creating Document
  - **Renaming Object**
  - **Opening Document** 
    - Opening for the First Time
      - Advanced Dialog Options
    - Opening Document Present in Project
    - **Opening Several Documents**
    - Opening Unloaded Documents
    - Opening from Clipboard
  - Opening Containing Folder

  - Exporting Documents Locked Documents

  - Using Objects and Object Views
  - Exporting Objects
    - Exporting Sequences to Sequence Format
    - Exporting Sequences as Alignment •
    - Exporting Alignment to Sequence Format
    - Exporting Nucleic Alignment to Amino Translation
    - Export Sequences Associated with Annotation •
  - Using Bookmarks
  - **Exporting Project**
  - Search in Project
  - **Options Panel**
  - Adding and Removing Plugins
  - Searching NCBI Genbank
  - Fetching Data from Remote Database
  - **UGENE** Application Settings
    - General
    - Resources
    - •
    - Network •
    - File Format
    - Directories
    - Logging •
    - Alignment Color Scheme ٠
    - **External Tools Settings** •
    - Workflow Designer Settings OpenCL
- Sequence View
  - Sequence View Components
  - Global Actions
  - Sequence Toolbars
  - Sequence Overview
  - Zoom View
  - **Details View**
  - Information about Sequence •
    - Manipulating Sequence
      - Show/Hide Sequence View Components
      - Showing Sequence in Multiple Lines
      - Translating Nucleotide Sequence
      - ٠ Zooming Sequence
      - Creating New Ruler
      - Going To Position

- Selecting Sequence Region
- Copying and Pasting Sequence
- Editing Sequence
  - Searching in Sequence
    - Load Patterns from File
    - Search Algorithm
    - Search in Other Settings
    - Annotations Settings
  - Exporting Selected Sequence Region
- Exporting Sequence of Selected Annotations
- Exporting Sequence Image
- Locking and Synchronize Ranges of Several Sequences
- Multiple Sequence Opening
- Annotations Editor
  - "db\_xref" Qualifier
  - Automatic Annotations Highlighting
  - The "comment" Annotation
- Manipulating Annotations
  - Creating Annotation
  - Selecting Annotations
  - **Editing Annotation**
  - Highlighting Annotations
    - Annotations Color
      - Annotations Visability
      - Show on Translation
    - Captions on Annotations
  - Creating and Editing Qualifier
    - Adding Column for Qualifier
  - Copying Qualifier Text Finding Qualifier

  - **Deleting Annotations and Qualifiers**
  - Importing Annotations from CSV
  - Exporting Annotations
- Sequence View Extensions
- Circular Viewer
  - · Circular View Settings
  - 3D Structure Viewer
    - Opening 3D Structure Viewer
    - Changing 3D Structure Appearance

      - Selecting Render Style
         Selecting Coloring Scheme
      - Calculating Molecular Surface
      - Selecting Background Color
        Selecting Detail Level
        Enabling Anaglyph View
    - Moving, Zooming and Spinning 3D Structure
    - Highlight Region on 3D Structure
    - Selecting Models to Display
    - Structural Alignment
    - Exporting 3D Structure Image
    - Working with Several 3D Structures Views
  - Chromatogram Viewer
    - Exporting Chromatogram Data
    - Viewing Two Chromatograms Simultaneously
  - Graphs Package
    - Description of Graphs
    - Graph Settings •
    - Saving Graph Cuttoffs as Annotations
  - Dotplot
    - **Creating Dotplot**
    - Navigating in Dotplot
    - Zooming to Selected Region
    - Selecting Repeat
    - Interpreting Dotplot: Identifying Matches, Mutations, Invertions, etc.
    - Editing Parameters
    - Filtering Results
    - Saving Dotplot as Image

    - Saving and Loading Dotplot Building Dotplot for Currently Opened Sequence
    - ٠ **Comparing Several Dotplots**
- Alignment Editor
  - Overview
    - Alignment Editor Features
    - Alignment Editor Components ٠
    - Navigation
    - Coloring Schemes
      - Creating Custom Color Scheme
    - Highlighting Alignment
    - Zooming and Fonts

- Searching for Pattern
- Consensus
  - Export Consensus
- Alignment Overview
- Working with Alignment
  - Selecting Alignment Region Moving Subalignment
  - Copying and Pasting Subalignment
  - Editing Alignment
- Removing Selection
  Filling Selection with Gaps
  - **Replacing Selected Character**
  - Replacing with Reverse-Complement

  - Replacing with Reverse
    Replacing with Complement
  - Removing Columns of Gaps Removing Sequence

  - Removing All Gaps
  - Undo and Redo Framework
  - Saving Alignment
  - Aligning Sequences
  - Aligning Sequence to this Alignment
  - Pairwise Alignment
    - Working with Sequences List
      - Adding New Sequences
        - Renaming Sequences
        - Sorting Sequences Shifting Sequences •

        - Collapsing Rows
      - Copying Sequences
  - Exporting in Alignment
    - Extracting Selected as MSA
    - Exporting Sequence from Alignment
    - Exporting Alignment as Image
  - Importing APR and ACE Files
- Statistics
  - Distance Matrix
  - Grid Profile
- Advanced Functions
  - Building HMM Profile
- Building Phylogenetic Tree
  - PHYLIP Neighbor-Joining
    - MrBayes
    - ٠ PhyML Maximum Likelihood
- Sanger Reads Editor
  - Sanger Reads Editor Overview
    - Sanger Reads Editor Features
    - Sanger Reads Editor Components
  - Working with Chromatogram
    - Mapping Reads to Reference
      - Alignment Statistics
      - Alignment Appearance
      - Overview and Show and Hide Chromatogram
      - Sanger Reads Consensus
      - Export Chromatogram ConsensusNavigation in Sanger Reads Alignment
        - **Editing Sanger Reads** 
          - Inserting Character

            - Replacing Character and Gap Removing Character and Gap
          - Inserting Gap
          - Removing Gap at the Left
          - Removing All Columns of Gaps
          - Trimming Left End
          - Trimming Right End
          - Renaming Read
          - Removing Read
          - Undo and Redo
      - Exporting Alignment without Chromatograms
- Assembly Browser
  - Import BAM and SAM Files
  - Browsing and Zooming Assembly
    - Opening Assembler Browser Window
    - Assembly Browser Window
    - Assembly Browser Window Components
    - Reads Area Description
    - Assembly Overview Description
    - Ruler and Coverage Graph Description
    - Go to Position in Assembly
    - Using Bookmarks for Navigation in Assembly Data

- Getting Information About Read
- Short Reads Vizualization
  - Reads Highlighting
  - Reads Shadowing
- Associating Reference Sequence
- Associating Variations
- Consensus Sequence
- Exporting
  - Exporting Reads
  - Exporting Visible Reads Exporting Coverage • ٠
  - •
  - **Exporting Consensus**
  - Exporting Consensus Variations
    Exporting Assembly as Image
    Exporting Assembly Region
- Options Panel in Assembly Browser
  - Navigation in Assembly Browser
    - Assembly Statistics
    - Assembly Browser Settings
- Assembly Browser Hotkeys
  - Assembly Overview Hotkeys
    - Reads Area Hotkeys
- Phylogenetic Tree Viewer
  - Tree Settings
    - Selecting Tree Layout and View
    - Modifying Labels Appearance
      - Showing and Hiding Labels
      - Aligning Labels
      - Changing Labels Formatting
    - Adjusting Branch Settings
    - Zooming Tree
    - Working with Clade
      - Selecting Clade
      - Collapsing and Expanding Branches
      - Swapping Siblings
      - Zooming Clade
      - Adjusting Clade Settings
      - Changing Root
    - Exporting Tree Image
    - Printing Tree
- Extensions
  - Workflow Designer
  - DNA Annotator
  - **DNA Flexibility** 
    - Configuring Dialog Settings
    - Result Annotations
  - DNA Statistics
  - DNA Generator
  - **ORF** Marker
  - Remote BLAST
    - Exporting BLAST Results to Alignment
    - Fetching Sequences from Remote Database
  - BLAST and BLAST+
    - Creating Database
      - Making Request to Database
    - Fetching Sequences from Local BLAST Database
  - Repeat Finder
    - Repeats Finding
      - ٠ Tandem Repeats Finding
        - Tandem Repeats Search Result
  - Restriction Analysis
    - Selecting Restriction Enzymes
    - Using Custom File with Enzymes
    - Filtering by Number of Hits
    - Excluding Region
    - Circular Molecule
    - Results
  - Molecular Cloning in silico
    - Digesting into Fragments
       Creating Fragment

      - **Constructing Molecule** 
        - Available Fragments
        - Fragments of the New Molecule
        - Changing Fragments Order in the New Molecule
        - Removing Fragment from the New Molecule •
        - Editing Fragment Overhangs
        - Reverse Complement a Fragment
        - Other Constuction Options
        - Output
      - Creating PCR Product

- In Silico PCR
  - Primers Details
  - Primer Library
- Secondary Structure Prediction
- SITECON
  - SITECON Searching Transcription Factors Binding Sites •
    - Types of SITECON Models
      - Eukaryotic
      - Prokaryotic
  - Building SITECON Model
- Smith-Waterman Search • HMM2
  - **Building HMM2 Model**
  - Calibrating HMM2 Model
  - Searching Sequence Using HMM2 Profile
- HMM3
  - Building HMM Model •
  - Searching Sequence Using HMM Profile
  - Searching Sequence Against Sequence Database
- uMUSCLE
  - MUSCLE Aligning
    - Aligning Profile to Profile with MUSCLE
    - Aligning Sequences to Profile with MUSCLE
- ClustalW
- ٠ MAFFT
- ٠ **T-Coffee**
- ٠ Bowtie
  - **Bowtie Aligning Short Reads** ٠
- Building Index for Bowtie
- Bowtie 2
  - · Bowtie 2 Aligning Short Reads
  - Building Index for Bowtie 2
- BWA
  - Aligning Short Reads with BWA
  - Building Index for BWA
- BWA-SW
  - Aligning Short Reads with BWA-SW
  - Building Index for BWA-SW
- BWA-MEM
  - Aligning Short Reads with BWA-MEM
  - Building Index for BWA-MEM
- UGENE Genome Aligner
  - Aligning Short Reads with UGENE Genome Aligner
  - Building Index for UGENE Genome Aligner
  - Converting UGENE Assembly Database to SAM Format
- CAP3
- SPAdes
- Weight Matrix
  - Searching JASPAR Database
- Building New Matrix
- Primer3
  - RTPCR Primer Design
- Spliced Alignment mRNA and cDNA
- External Tools .
  - Configuring External Tool
- Query Designer
- Plasmid Auto Annotation
- ClustalO
- Kalign Aligning
- Shared Database
  - Configuring Database
  - Connecting to a Shared Database
  - Adding Data to the Database
  - . Database in the Project
  - **Deleting Data**
  - Drag'n'drop in the Database
  - Exporting Objects from the Database
- UGENE Public Storage
- UGENE Command Line Interface
  - CLI Options
  - CLI Predefined Tasks
    - Format Converting Sequences

    - Converting MSA
      Extracting Sequence
    - Finding ORFs ٠
    - Finding Repeats
    - Finding Pattern Using Smith-Waterman Algorithm
    - Adding Phred Quality Scores to Sequence
    - Local BLAST Search
    - Local BLAST+ Search

- Remote NCBI BLAST and CDD Requests
- Annotating Sequence with UQL Schema •
- •
- Building Profile HMM Using HMMER2 Searching HMM Signals Using HMMER2 •
- Aligning with MUSCLE
- Aligning with ClustalW Aligning with ClustalO Aligning with Kalign

- Aligning with MAFFT
- •
- Aligning with MAPP 1 Aligning with T-Coffee Building PFM Searching for TFBS with PFM Building PVM •
- •
- Searching for TFBS with Weight Matrices
  Building Statistical Profile for SITECON
- ٠ Searching for TFBS with SITECON
- Fetching Sequence from Remote Database Gene-by-Gene Report •
- •
- Reverse-Complement Converting Sequences ٠
- Variants Calling
- Generating DNA Sequence
- Creating Custom CLI Tasks
- APPENDIXES
  - Appendix A. Supported File Formats
    - Specific File Formats
      - UGENE Native File Formats • Other File Formats
- Tutorials
- Using BioMart with UGENE
  - Environment requirements
  - Installing UGENE extension on Mozilla Firefox
     Opening data found using BioMart in UGENE

  - Opening BioMart data in UGENE by ID
  - Opening selected data in UGENE

# **About Unipro**

Established in 1992 Unipro company has its headquarters located in Novosibirsk Akademgorodok (the home of Siberian Branch of Russian Academy of Sciences). The company's primary activity is IT outsourcing solutions. To learn more about the company, please, visit the company website.

# **About UGENE**

Unipro UGENE is a free cross-platform genome analysis suite. It is distributed under the terms of the GNU General Public License.

To learn more about UGENE visit UGENE website.

It works on Windows, Mac OS X or Linux and requires only a few clicks to install.

- Key Features
- User Interface
- High Performance Computing
- Cooperation

# **Key Features**

- Creating, editing and annotating nucleic acid and protein sequences
- Search through online databases: NCBI, ENSEMBL, PDB, SWISS-PROT, UniProtKB/Swiss-Prot, UniProtKB/TrEMBL, UniProt(D AS), Ensembl Human Genes (DAS)
- Multiple sequence alignment: ClustalW, ClustalO, MUSCLE, Kalign, MAFFT, T-Coffee
- Online and local BLAST and BLAST+ search
- Restriction analysis with integrated REBASE restriction enzyme database
- Integrated Primer3 package for PCR primers design
- · Search for direct, inverted and tandem repeats in DNA sequences
- Constructing dotplots for nucleic acid sequences
- Search for transcription factor binding sites (TFBS) with weight matrix and SITECON algorithms
- Aligning short reads with Bowtie, Bowtie 2, BWA, BWA-SW and UGENE Genome Aligner
- Contig assembly with CAP3
- Search for ORFs
- Cloning in silico
- 3D structure viewer for files in PDB and MMDB formats, anaglyph view support
- Protein secondary structure prediction with GOR IV and PSIPRED algorithms
- HMMER2 and HMMER3 packages integration
- Building (using integrated PHYLIP and MrBayes packages) and viewing phylogenetic trees
- Local sequence alignment with optimized Smith-Waterman algorithm
- · Combining various algorithms into custom workflows with UGENE Workflow Designer
- · Search for a pattern of various algorithms' results in a nucleic acid sequence with UGENE Query Designer
- Visualization of next generation sequencing data (BAM files) using UGENE Assembly Browser
- PCR in silico
- Spade de novo assembler

# **User Interface**

- · Visual and interactive genome browsing including circular plasmid view
- Multiple alignment editor
- Chromatograms visualization
- 3D viewer for files in PDB and MMDB formats with anaglyph stereo mode support
- Phylogenetic tree viewer
- Easy to use Workflow Designer for custom computational workflows
- · Easy to use Query Designer for analyze a nucleotide sequence using different algorithms at the same time
- Assembly Browser for visualize and efficiently browsing large next generation sequence assemblies

# **High Performance Computing**

- Complete support of modern multicore processors and SSE instructions
- Out of the box support of modern GPUs using NVIDIA CUDA and ATI Stream
- Integrated solutions for Cell Broadband Engine

# Cooperation

- · Can be used for education purposes in schools and universities
- · Features to be included into the next release are initiated by users
- UGENE team is ready for collaboration in related projects, both free and commercial

# **Download and Installation**

UGENE is compatible with the three most common operating systems: Windows, Mac OS X, and Linux. It has some minimum system requirements. If your system fits these requirements, you're welcome to download UGENE from http://ugene.unipro.ru/download. The program can be used and distributed under the terms of GPLv2.

Follow these recommendation to choose which UGENE package to download.

Below you can also find links to the guides on UGENE installation on different operating systems.

- System Requirements
- UGENE Packages
- Installation on Windows
- Installation on Mac OS X
- Installation on Linux
  - Native Installation on UbuntuNative Installation on Fedora

# **System Requirements**

The system requirements for UGENE are these:

- Operating system (32 or 64 bit):
  - Windows XP, Windows Vista, Windows 7, Windows 8 Using a zip package it is possible to use UGENE without administrative rights on Windows
  - Mac OS X 10.5 or later For older Mac OS X versions (PowerPC, 10.4) UGENE version 1.10.3 is available.
  - Linux
    - Ubuntu 12.04 or later
    - Fedora 19 or later
    - If you have another Linux system, you may use a universal binary package
- RAM:
  - 512 Mb RAM is required.
  - At least 2 Gb RAM is recommended.
- Disk space:
  - The minimum required disk space depends on the UGENE package:
    - Standard package: 200-300 Mb
    - Full package: 500-900 Mb
    - NGS package: 21-24 Gb
- Display:
  - It is recommended to set the screen resolution to a value greater than 1280x720.
- Internet:
  - Internet connection is required for some tasks like loading data from online databases.

UGENE takes care to use capabilities of your system: the more RAM and cores you have, the more quickly you'll get results of your calculations.

Also, if you have an OpenCL-capable video card, you can use GPU-optimized versions of the following tools:

- Smith-Waterman Search
- UGENE Genome Aligner

# **UGENE** Packages

Besides selecting an appropriate package for your operating system (Windows, Mac OS X, or Linux; 32 or 64 bit), you should take into account the following considerations.

### Should I download standard, full, or NGS package?

In most cases the full package is the best choice. Exceptions are:

• Use the standard package, if:

- You're going to use only basic UGENE features and don't want to waste Internet traffic
- You have limited disk space
- Use the NGS package, if:
  - You're going to analyze ChIP-Seq data with the Cistrome pipeline

Explanation of the tip above: Some tools are embedded into UGENE as external. To be launched from the UGENE graphical interface, an external tool needs a corresponding executable file. The list of the external tools can be found on this page.

The standard package does not include the tools, whereas the full package include all the required tools.

The NGS package, besides containing the external tools, contains sample data for the Cistrome pipeline (hg19 genome, reference genes, etc.), so you can run it out of the box.

In 2013 we worked on extending of the UGENE NGS framework with three popular pipelines for analyzing NGS data:

- · Variant calling with SAMtools
- RNA-Seq data analysis with Tuxedo
- ChIP-Seq data analysis with Cistrome.

The NGS package was added as the result of this work. We decided to add it as we want our users to be able to use <u>all</u> UGENE features out of the box. However, it appears that the first two pipelines are also available out of the box in the full UGENE package.

The work was supported by grant RUB1-31097-NO-12 from NIAID.

### I have Windows. Should I download installer package or portable zip bundle?

If you have administrative rights on Windows, use the installer package. It will make integration with your Windows system more tight. For example, it will add associations for bioinformatics formats supported by UGENE, so that corresponding files are opened in it by default.

### I have Linux. Which package should I use?

If your Linux is not Ubuntu or Fedora, you can use the installer package too. Otherwise, for more tight integration with the systems, you can install UGENE from corresponding repositories, following these guides:

- Native installation on Ubuntu
- Native installation on Fedora

## Installation on Windows

To install UGENE on Windows:

• Download UGENE Windows installation package:



Packages for Windows Vista, Windows 7 and higher Windows versions

Installers:

- Download <u>Universal</u> installer package
- Download 32-bit <u>Standard</u> or <u>Full</u> installer package
- Download 64-bit <u>Standard</u> or <u>Full</u> installer package

Zip bundles:

- Download 32-bit portable <u>Standard</u> or <u>Full</u> zip bundle
- Download 64-bit portable <u>Standard</u> or <u>Full</u> zip bundle
- Download 64-bit <u>NGS</u> portable zip bundle

· Launch the downloaded \*.exe le and follow the Unipro Setup wizard:

🔂 Unipro Setup		
	Welcome to the UGENE Setup Setup will guide you through the installation of UGENE. It is recommended that you close all other applications before starting Setup. This will make it possible to update relevant system files without having to reboot your computer. Click Next to continue.	
Next > Cancel		

A Be sure that you launch the installer with an administrative Windows account. If you have a problem with installation, try to do the following: right-click on the installer '.exe' le and select Run as administrator item.

Alternatively, to use UGENE without installing:

• Download UGENE zip package:



Packages for Windows Vista, Windows 7 and higher Windows versions

Installers:

- Download <u>Universal</u> installer package
- Download 32-bit <u>Standard</u> or <u>Full</u> installer package
- Download 64-bit <u>Standard</u> or <u>Full</u> installer package

### Zip bundles:

- Download 32-bit portable <u>Standard</u> or <u>Full</u> zip bundle
- Download 64-bit portable Standard or Full zip bundle
- Download 64-bit <u>NGS</u> portable zip bundle
- Unpack it.
- Launch the ugeneui.exe le.

# Installation on Mac OS X

To install UGENE on Mac OS you can use installation package or Disk image file:

### 🗳 Mac OS X

Packages for Mac OS X 10.7 and higher:

- Download 64-bit <u>Universal</u> installer package
  Download 64-bit <u>Standard</u> or <u>Full</u> package
  Download 64-bit <u>NGS</u> package

Note that as the NGS package size is about 4 GB, it is highly recommended to download it using some download management software, for example Folx. Otherwise, if you download it via your web-browser, a short disconnection may cancel the process.

Also, find below UGENE packages for old Mac OS X versions. Note, that they updated from time to time (not in each release).

- PowerPC: the latest available version is 1.10.3 (download).
  Mac OS X 10.4 Tiger (Intel): the latest available version is 1.10.3 (download).
- To use this version one should download and install Qt first.
- Mac OS X 10.5: the latest available version is 1.18.0 (download <u>Standard</u> or <u>Full</u> 32-bit package).
   Mac OS X 10.6: the latest available version is 1.18.0 (download <u>Standard</u> or <u>Full</u> 64-bit package).

Please contact us if you're interested in obtaining the latest version of one of these packages.

- Launch the universal installer package and follow the Unipro Setup wizard.
- Also you can launch \*.dmg le and accept the GNU license agreement. The following window will appear:

00	Unipro UGENE	
X	1 of 2 selected, Zero KB available	-0
-		
- Ar		
Applications	ugeneu	

• To start UGENE click on the ugeneui icon. You can also copy UGENE to the Applications folder by dragging it.

### Installation on Linux

To install UGENE on Linux:

Download UGENE Linux installation package:



Installers:

- Download 32-bit Universal installer
- Download 64-bit Universal installer

Universal binary packages:

- Download 32-bit <u>Standard</u> or <u>Full</u> package
- Download 64-bit Standard or Full package
- Download 64-bit <u>NGS</u> package

• Set the file as executable:

chmod +x [name of the downloaded file]

• Launch the file and follow the Unipro Setup wizard.

Unipro UGENE Installer Setup	8
Setup - Unipro UGENE	
Welcome to the Unipro UGENE Setup Wiza	rd.
Settings	Next > Quit

Alternatively, to use UGENE without installing:



Installers:

- Download 32-bit <u>Universal</u> installer
- Download 64-bit <u>Universal</u> installer

Universal binary packages:

- Download 32-bit <u>Standard</u> or <u>Full</u> package
- Download 64-bit <u>Standard</u> or <u>Full</u> package
- Download 64-bit <u>NGS</u> package

• Unpack the archive. You can use this command:

tar -xf [name of the downloaded \*.tar.gz file]

• Change the working directory to the unpacked UGENE directory:

cd [name of the unpacked directory]

• Launch the UGENE GUI version using the command:

./ugene -ui

or the command line version using the command:

./ugene

Several native packages for specic Linux distributions are also available. UGENE is a part of Ubuntu and Fedora Linux distributions. See the next chapter.

- Native Installation on Ubuntu
- Native Installation on Fedora

### Native Installation on Ubuntu

Ugene packages for different Ubuntu versions are available on the Personal Package Archives (PPA). To start installing and using software from the UGENE PPA do the following steps:

• Open a terminal and enter:



sudo apt-get update

• Now you're ready to start installing UGENE:

sudo apt-get install ugene
To install the non-free UGENE plugins do the following:
sudo apt-get install ugene-non-free
To install the UGENE data do the following:

sudo apt-get install ugene-data

UGENE will appear in the applications list.

8 🖨		
<b>O</b> •	⊗ UGENE	
	Applications	
	U	
	Unipro UGENE	
	📄 Files & Folders	See 128 more results ►
	ugene-1.13.0	ugene
1		

### **Native Installation on Fedora**

Ugene packages for different Fedora versions are available on the Fedora. To start installing and using software do the following:

• Open a terminal and enter:

sudo yum install ugene

• Now the latest available UGENE appears in the applications list.

# **Basic Functions**

- UGENE Terminology •
  - **UGENE** Window Components
    - Welcome Page
    - Project View
    - Task View
    - Log View ٠
    - Notifications
- Main Menu Overview
- **Creating New Project** Creating Document
- Renaming Object
  - **Opening Document** 
    - Opening for the First Time
    - Advanced Dialog Options Opening Document Present in Project
      - **Opening Several Documents**
    - Opening Unloaded Documents •
    - Opening from Clipboard
- Opening Containing Folder
- **Exporting Documents** .
- Locked Documents
- Using Objects and Object Views ٠
- ٠ Exporting Objects
  - Exporting Sequences to Sequence Format
  - Exporting Sequences as Alignment
  - Exporting Alignment to Sequence Format
  - Exporting Nucleic Alignment to Amino Translation
  - Export Sequences Associated with Annotation
- Using Bookmarks
- **Exporting Project**
- Search in Project
- **Options Panel**
- Adding and Removing Plugins Searching NCBI Genbank
- Fetching Data from Remote Database
- **UGENE** Application Settings

  - General Resources
  - ٠ Network
  - File Format
  - Directories

  - Logging
  - Alignment Color Scheme •
  - **External Tools Settings** • Workflow Designer Settings
  - OpenCL

# **UGENE** Terminology

### Project

Storage for a set of data files and visualization options.

### Document

A single file (can be stored on a local hard drive or be a remote web page). Each document contains a set of objects.

### Object

A minimal and complete model of biological data. For example: a single sequence, a set of annotations, a multiple sequence alignment.

### Task

A process, usually asynchronous, that works in background. For example: some computations, loading and writing files.

### Plugin

A dynamically loaded module that adds new functionality to UGENE.

### **Object View**

A graphical view for a single or a set of objects.

### **Project View**

A visual component used to manage active project.

### **Task View**

A visual component used to manage active tasks.

### Log View

A visual component used to show logs.

### Notifications

A visual component used to show notifications. Generally it is used to open tasks reports.

### **Plugin Viewer**

A visual component used to manage plugins.

### **Sequence View**

An Object View aimed to visualize DNA, RNA or protein sequences along with their properties like annotations, chromatograms, 3D models, statistical data, etc.

### Annotation

Additional information about a sequence, identified by its name and the sequence region.

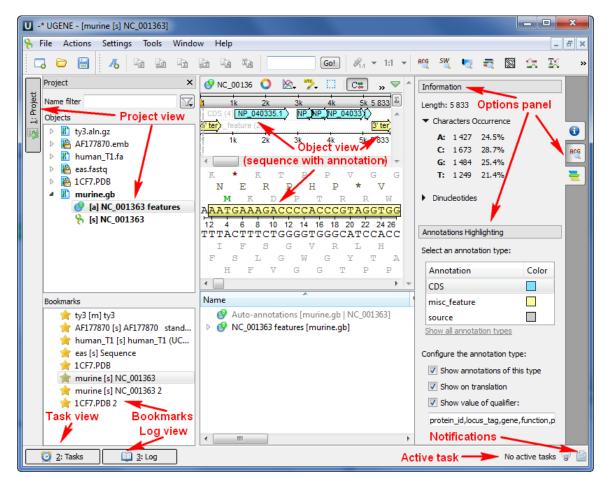
### **Alignment Editor**

An Object View used to visualize and edit DNA, RNA or protein multiple sequence alignments.

### **Options Panel**

An Options Panel it is the panel with different information tabs and tabs with settings for Sequence View and Assembly Browser.

In the image below you can see a typical UGENE window with a Project View and a single Object View window opened:



## **UGENE Window Components**

This chapter describes UGENE main window components Project View, Task View, Log View and the Notifications popup window.

- Welcome Page
- Project View
- Task View
- Log View
   Netifications
- Notifications

### Welcome Page

The Welcome Page is the first page that will appear when UGENE has been launched.

From the Welcome Page you can open files, create sequence, create workflow, open the Quick Start Guide and open recent files directly.

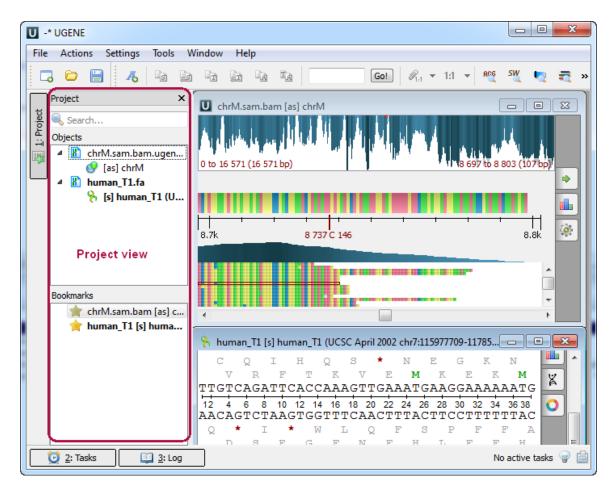
U File Actions Settings Tools Window Help	_ <i>B</i> ×
Welcome to UGENE	
Open File(s)     Create Sequence       Create Workflow     Quick Start Guide	Recent files - 538117317_misc3_copy1.aln - 538117317_misc3.nwk - 538117317_misc3.aln - P027682.nwk - P027682.aln - P19121.gb - P49822.gb Recent projects No opened projects yet
Cite UGENE: "Unipro UGENE: a unified bioinformatics toolkit" Okonechnikov; Golosova; Fursov; the UGENE team Bioinformatics 2012 28: 1166-1167	Follow UGENE:

To return to the Welcome Page go to the Window->Start Page main menu item.

## **Project View**

The *Project View* shows documents and bookmarks of the current *project*. The documents are files added to the project. And the bookmarks are visual view states of the documents. Read *Using Bookmarks* to learn more about bookmarks.

To show/hide the Project View, click the Project button in the main UGENE window:



You can also use the Alt+1 hotkey to show/hide the Project View.

To create a new project, refer to *Creating New Project*. Note that if you have no project created when opening file with a sequence, an alignment or any other biological data, a new anonymous project is created automatically.

### **Task View**

The Task View shows active tasks, for example, algorithms computations.

To show/hide the Task View, click the Tasks button in the main UGENE window:

	Task name	Task state description	Task progress
	Icoading documents	Running	84%
	Øpening view for document: Mycobacterium.sorted	Running	84%
	a 🚡 Adding document to project: Mycobacterium.so	Running	84%
Task View	a 🚡 BAM/SAM file import: Mycobacterium.sorte	Running	84%
	🖌 📔 🚽 🖉 Load BAM info	Finished	100%
	Prepare assembly file to import	Finished	100%
	🚡 Convert BAM to UGENE database (Myco	Importing reads	68%
	•		
0 2: Tasks	3: Log Running task: Loading documents	84%	Tasks: 1 🍚 🧯

The hotkey for showing/hiding the Task View is Alt+2.

The Task name column of the Task View shows the tasks names. Task state description shows the status of the active tasks: Started, Running, Finished and so on. The Task progress column shows the percentage of the tasks progress. If you want to cancel a task, click the red cross button in the Actions column for the task.

### Log View

The Log View shows the program log information.

To show/hide the Log View click the Log button in the main UGENE window:

		Log	×
		[INFO][07:06] Converting assembly from Mycobacterium.sorted.bam to Mycobacterium.sorted.bam.ugenedb started [INFO][07:07] Converting assembly from Mycobacterium.sorted.bam to Mycobacterium.sorted.bam.ugenedb succesfully finished: imported 967272 reads, total time 50 s, pack time 0 s [INFO][07:07] BAMImporter task total time is 53 sec	*
Log View [INFO][07:10] Converting assembly from Mycobacterium.sorted.bam to Mycobacterium.sorted.bam [INFO][07:11] Converting assembly from Mycobacterium.sorted.bam to Mycobacterium.sorted.bam imported 967272 reads, total time 49 s. pack time 0 s		[INFO][07:10] Converting assembly from Mycobacterium.sorted.bam to Mycobacterium.sorted.bam.ugenedb started [INFO][07:11] Converting assembly from Mycobacterium.sorted.bam to Mycobacterium.sorted.bam.ugenedb succesfully finished: imported 967272 reads, total time 49 s, pack time 0 s [INFO][07:11] BAMImporter task total time is 52 sec]	II
			-
Í	💆 <u>2</u> : Tasks 🛛 🛄	3: Log No active tasks 💡	) 歯

The hotkey for this action is Alt+3.

It is possible to configure the *Log View* settings: the level of the log to show (ERROR, INFO, DETAILS, TRACE), the category (Algrorithms, Tasks, etc.), and the format of the log messages (format of the dates, etc.). This settings can be configured in the UGENE *Application Settings*.

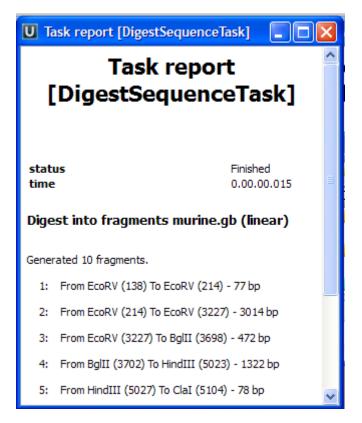
### **Notifications**

The Notifications component shows notifications for tasks reports.

Bookmarks		Notifications	+×
	[INFO][16:54] Converting assembly from Klebsislla.sort.bam to Klei [DETAILS][16:54] Importing assembly 'pkF70' (1 of 3) [DETAILS][16:55] Succesfully imported 136066 reads for assembly [DETAILS][16:55] Importing assembly 'pkf140' (2 of 3) [DETAILS][16:57] Succesfully imported 416287 reads for assembly	pkF70 🚺 [16:27:31] Report for task: '	×
	[DETAILS][16:57] Importing assembly 'pKF94' (3 of 3) [INFO][17:01] Canceling task: Convert BAM to UGENE database (K [INFO][17:01] Canceling task: BAM/SAM file import: Klebsislia.sort.] [INFO][17:01] Canceling task: Convert BAM to UGENE database (K	lebsisli 🐼 [16:26:51] Report for task: '	×
🖸 2: Tasks	LTNEOI [17:01] Canceling task: Adding dogument to project: Klebsis	Notifications	

If a task has finished without errors, the notification is blue. If an error has occured during the task execution, the notification is red. If a warning has occured during the task execution, the notification is yellow.

To open a task report, click on the corresponding notification. See an example of a task report below:



To remove a notification from the Notifications popup window, click the notification cross button.

Note that you can click on the clip button of the Notifications popup window to show the window always on top.

# Main Menu Overview

Menu	Description
File	A set of project level operations. List of operations: new project, new document from text, new workflow, open, open as, open from clipboard, access remote database, search NCBI genbank, connect to UGENE shared database, save all, save project as, export project, close project, recent files, recent projects, exit.
Actions	Various actions associated with the active window. List of operations: go to position, add, copy/paste, analyze, align, cloning, export, remove, edit (for the <i>Sequence View</i> ); go to position, add, copy/paste, colors, highlighting, edit, align, tree, stati stics, view, export, advanced, consensus mode, close active window (for the <i>Alignment Editor</i> ).
Settings	Preferences and plugin settings.
Tools	Various tools. This menu is extended by different plugins. List of operations: sanger data analysis, NGS data analysis, BLA ST, multiple sequence alignment, cloning, primer, search for TFBS, HMMER tools, build dotplot, random sequence generator, show counters, expert discovery, query designer, workflow designer.
Window	A list of active windows and basic manipulations with the windows. List of operations: window layout (multiple or tubbed), close active view, close all windows, tile windows, cascade windows, next window, previous window.
Help	Application help and check for updates. List of operations: open UGENE user manual, open workflow designer manual, open query designer manual, view UGENE documentation online, visit UGENE website, check for updates, op en start page, about.
Unipro UGENE (Mac OS only)	List of operations: about Unipro UGENE, preferences, services, hide Unipro UGENE, hide others, show all, quit Unipro UGENE.

The menus can be dynamically populated with new actions added by plugins. Check the *Plugins* documentation to learn how each plugin affects global and context menus.

# **Creating New Project**

A project store links to the data files, cross-file data associations, and visualization settings.

Below is the description of how to create a new project manually. Note that if you have no project created when opening a file with a sequence, an alignment or any other biological data, a new anonymous project is created automatically.

To create a new project select the File New project menu or click the New project button on the main toolbar.

The dialog will appear:

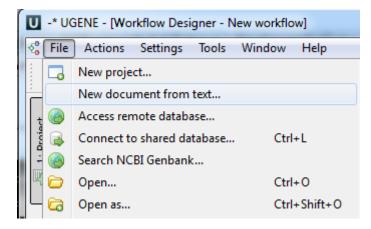
U Create New	/ Project	
Project name	New Project	
Project folder		
Project file	project	
	Create Cancel Help	

Here you need to specify the visual name for the project and the directory and file to store it.

After you click the Create button the Project View window is opened.

### **Creating Document**

To create a new sequence file from a text, select the File New document from text main menu item:



The Create Document dialog appears:

1	U Create Documen	t		?	x
	Paste data here				
l					
	Custom setting:	s			
	Alphabet:		Standard DNA		-
	Skip unknown s	symbols			
	Replace unkno	wn symbols w	iith		
	Document location:				
	Document format:	FASTA		•	•
	Sequence name:	Sequence			
	Save file immediately				
		Create	Cancel	He	lp

You can input the created sequence to the Paste data here field. You can type or paste sequences in FASTA or text format.

The following Custom settings are available:

Alphabet — here you can select the alphabet:

Custom settings		
Alphabet:		Standard DNA 👻
<ul> <li>Skip unknown symbols</li> <li>Replace unknown symbols w</li> </ul>		Standard DNA Standard RNA Extended DNA Extended RNA
Document location:		Standard amino Extended amino All symbols
Document format: FASTA		•

The following alphabets are available: Standard DNA, Standard RNA, Extended DNA, Extended RNA, Standard amino, Extended amino.

Skip unknown symbols / Replace unknown symbols with — you can select either to skip unknown input symbols or to replace them with the specified symbol.

Document location - a location of the created document.

Document format — a format of the created document. Currently available formats are FASTA and Genbank.

Sequence name — a name of the sequence in the created document.

Save file immediately - check this option if you want to save the document immediately after the Create button is pressed.

The created document will be added to the current project and opened in the Sequence View.

# **Renaming Object**

To rename an object in the project tree view use the F2 shortcut or *Rename* context menu item:

Objects		<b>4</b>	500	
4 🖹 murine.gb				
🧏 [s] NC_001 🔗 [a] NC_001		Open view Add to view	) 	(2)
	8	Export/Import Remove selected items	► Del	
		Rename	F2	
		Mark as circular Open containing folder	*	ĸ

# **Opening Document**

UGENE stores information about *documents* you are working with in a *project*. Once a *document* has been opened, the information about it is saved in the current *project*.

- Opening for the First Time
  - Advanced Dialog Options
- Opening Document Present in Project
- Opening Several Documents
- Opening Unloaded Documents
- Opening from Clipboard

## **Opening for the First Time**

To open a *document* that is not yet presented in the current *project* use either an advanced *Open* dialog, a simple open file dialog or just drag the document to the UGENE window.

UGENE automatically detects the format of the document, but if you use the advanced dialog you can choose the format manually.

To open the advanced dialog select one of the following:

- Add Existing document item in the Project View context menu
- File Open As item in the main menu

To simply open the document select one of the following:

- Open item in the main toolbar
- File Open item in the main menu

or drag the file to the UGENE window. Also it is able to drag and drop documents (not objects) between opened UGENEs.

Documents created not by UGENE are *locked*. To be able to edit the document you should save a copy of the document and continue working with the copy.

• Advanced Dialog Options

### **Advanced Dialog Options**

Open the Select Correct Document Format dialog by Add Existing document item in the Project View context menu or by File Open As item in the main menu. The foolowing dialog will appear:

U Select Correct Document Format	? X
Options for human_T1.fa	
FASTA format. Score: 13 (Perfect match)	more
Plain text format. Score: 2 (Low similarity)	more
BED format. Score: 2 (Low similarity)	more
Raw sequence format. Score: 1 (Very low similarity)	more
MSF format. Score: 1 (Very low similarity)	more
Choose format manually Genbank	•
File preview	
>human_T1 (UCSC April 2002 chr7:115977709-117855134) TTGTCAGATTCACCAAAGTTGAAATGAAGGAAAAAATGCTAAGGGCAGCC AGAGAGAGGTCAGGTTACCCACAAAGGGAAGCCCATCAGACTAACAGCGG ATCTCTCGGCAGAAACCCTACAGGCCAGAAGAGAGTGGGGGGCCAATATTC CATATTCTTAAAGAAAAGA	•
OK Cancel	Help

Here you can choose how to interpret the data stored in the file. The format is detected automatically, but you can select it manually.

### **Opening Document Present in Project**

To open a *document* that is already present in the current *project* select it in the *Project View* and click Enter, double-click on it or drag it to an empty space of the UGENE window.

### **Opening Several Documents**

To open several documents that are not yet presented in the current *project* use the *File Open* item in the main menu. The *Select files open* dialog will appear. Select the documents with a help of the *Ctrl* button and click on the *Open* button. The following dialog will appear:

U Sequence Reading Options	? X
The document selected contains multiple sequence instances.	
Please select the way UGENE will read these sequences:	
As separate sequences in sequence viewer	
Merge sequences into a single sequence to show in sequence viewer	
Number of 'unknown' symbols <i>('N 'for nucleic or 'X' for amino)</i> between parts: 10 ba	ses 🔺
O Join sequences into alignment and open in multiple alignment viewer	
Align reads to reference sequence	
File preview:	
>d GACTAGC >a GACTAGC >c GACTAGC >b	* III +
OK Cancel	Help

Select the reading options and click on the OK button.

### **Opening Unloaded Documents**

The [unloaded] prefix indicates that document is unloaded:



To open unloaded document use the *Load selected document* context menu, *Enter* key or double click on the document. To unload the document use the corresponding context menu item.

### **Opening from Clipboard**

It is possible in UGENE opening text data and files by pasting them into the Project View from the clipboard.

Select text data in a browser or in a text editor or file(s) from your system and press Ctrl+C (Cmd+C for Mac OS X).

Switch to UGENE with an opened project, click the project area and press Ctrl+V (Cmd+V).

The copied data or file(s) will be opened in UGENE.

Another way for file(s) is drag'n'drop. Using your mouse, drag and drop the files from the file system to UGENE, and they will be opened by the program.

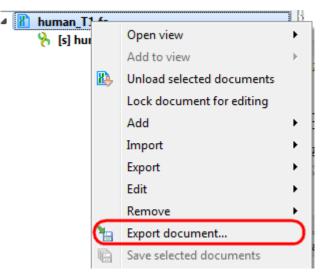
# **Opening Containing Folder**

To open a containing folder of the *document* that is already present in the current *project* select it in the *Project View* and click on the *Open containing folder* context menu item.

# **Exporting Documents**

If a document has a format that supports writing in UGENE (see the Supported File Formats chapter), you can export the document to a new document in a required format.

To do it use the *Export document* item in the context menu:



The following dialog appears:

Export Docume	ent			? <mark>x</mark>
Save to file				
File format FAST	4			•
Compress file				
Add to project				
		Export	Cancel	Help

Here you may select the name of the output file in the *Save to file* field and, optionally, choose the format of the output file in the *File format* fi eld. Use the *Compress file* checkbox to compress the file. The *Add to project* checkbox, checked by default, adds the output file to the current project. After choosing all parameters click the *Export* button.

# **Locked Documents**

The lock icon in the document element indicates that the document can't be modified:



UGENE does not allow modification of some formats that were created not by UGENE.

If UGENE is able only to read a document (see the Supported File Formats chapter), you can export the document objects to a file. To do it use the built-in export utilities.

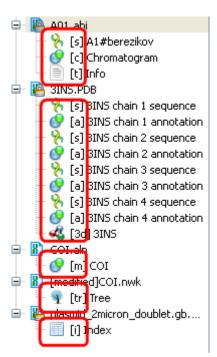
Also, you can export the document objects of unlocked documents.

# **Using Objects and Object Views**

The document always contains one or more objects. An object is a structured biological data that can be visualized by different Object Views.

A single *Object View* can visualize one or several objects of different types. For example a single view can show a sequence, annotations for the sequence, 3D model for the part of the sequence or its chromatogram simultaneously.

The type of an object is indicated by the symbol in the square brackets and the icon near the object:



Below is the list of object types supported by the current version of UGENE.

### **Object types:**

Symbol	Icon	Description
[3d]		A 3D model.
[a]		Annotations for DNA sequence regions.
[as]		An assembly.

[c]	Chromatogram data.
[1]	A file with index information for a set of other, usually large files.
[m]	A multiple sequence alignment.
[s]	A nucleic, protein or raw sequence.
[t]	A plain text.
[tr]	A phylogenetic tree.

You can edit names of particular objects, such as sequence objects, by selecting them in the *Project View* and then pressing F2. To be able to do so, the document containing the target object must be unlocked.

To see the list of all available views for a given object select the object and activate the context menu inside the *Project View* window and select the *Open view* submenu:

Project		×	1 🕑 🛚	VC_	0013	363 s	eque	nce						
Name filter		7			500		1	k		1.5k	040	2k	-	2
Objects			5 tern			2 (2)				NP	040	335	. 1	
🖻 🎦 mur		Open view		۲				ew vie	ew: S	Seque	nce vi	ew		2
		Add to view		Þ								_	<b>`</b>	
	8	Unload selected docur	ments		*	к		г	Р	Р	1	7	G	
		Add		Þ	E		R	Ρ	F	I	P	*	V	
		Import		۲	L	К	D	I	-	Т	R	R		W
		Export		Þ		AA.	GA	ccc	CCA	-	CGI		-	_
		Remove		۲	CT CT	TT TT	10 CT	12 GG0	14 3G1		18 20 GC <i>I</i>			-
	1	Save a copy			F		s	G	١	7	R	L	Н	
	P.	Save selected docume	ents		S	L F	v	G	พ ≆	G G	т Т	ζ Ρ	Т	P
			<											
														_
			Name		NC_	0013	63 fe	ature	s (m	urine.	gb]			

The picture above illustrates an option to visualize the selected DNA sequence object using the Sequence View — a complex and extensible Object View that focuses on visualization of sequence objects in combination with different kinds of related data: sequence annotations, graphs, chromatograms, sequence analysis algorithms. Note, that the Sequence View is described in more details in the separate document ation section.

# **Exporting Objects**

The document objects can be exported into a new document. For more details see the following chapters:

- Exporting Sequences to Sequence Format
- Exporting Sequences as Alignment
- Exporting Alignment to Sequence Format
- Exporting Nucleic Alignment to Amino Translation
- Export Sequences Associated with Annotation

### **Exporting Sequences to Sequence Format**

Select a single or several sequence objects in the Project View window and click the Export Export sequences context menu item:

0	* UGENE - [fasta_example_fasta]	
8	File Actions Settings Tools Window	/ Help _ B ×
	3 🗁 🔡 🔥 🗛 🗛 🗛	Gol Ø <sub>1:1</sub> ▼ 1:1 ▼ <sup>806</sup> S₩ ♥ ₹ »
	Project	🗙 🔗 Phaneropt 🔘 🖄 🧚 🏷 🗌 🖓 🖀 🖓
1: Project	Name filter	So 100 150 200 250 300 350 400 450 500 550 604     So     So
Pre-	Objects	empty
1	▲ K fasta_example_fasta	1 50 100 150 200 250 300 350 400 450 500 550 804
- <b>1</b> 2		
	🗞 [s] Isophya_altaica_EF540820	
	[s] Bicolorana_bicolor_EF540830	30 KTSNSSRIRST
	8 [s] Roeseliana_roeseli	Open view
	S International S Internationa	Add to view
	% [s] Metrioptera_japonica_El % [s] Gampsocleis sedakovii	14 16 18 20 22 24 26 28 30 33
	[5] Oumpsociels_sedukovii_ [5] Deracantha_deracantoic	Unload selected objects AGCTCGGCTTAATCCAGTTG Lock document for editing R A S N P * G
	8 [s] Zychia_baranovi	FT. PT. DV
	😣 [s] Tettigonia_viridissima	SGF TL
	🗞 [s] Conocephalus_discolor	Import •
	🗞 [s] Conocephalus_sp.	Export   Export sequences
	Bookmarks	Edit   Export sequences as alignment
	☆ fasta_example_fasta	Remove
		Save selected documents 250 300 350 400 450 500 550 604 🔻
		A
		· · · · · · · · · · · · · · · · · · ·
-		
	🙋 <u>2</u> : Tasks 🛛 🛄 <u>3</u> : Log	No active tasks 🖙 🧾

The Export Selected Sequences dialog will appear:

U Export Selected Sequences		? <mark>x</mark>
Export to file	C:/work/ugene/data/samples/FASTA/human_T1_region_new.fa	
File format to use	FASTA	•
Export with annotations		
Add document to the project		
Use custom sequence name	human_T1_region	
Convertion options		
Save direct strand	Save complement strand	
Translate to amino alphabet	✓ Save all amino frames	
Use custom translation table	1. The Standard Genetic Code	-
Merge options		
<ul> <li>Save as separate sequences</li> </ul>		
		0
Add gap symbols between seque		UV
	Export Cancel	Help

Here you can select the location of the result file and a sequence file format. You can choose to add newly created document to the current project and use custom sequence name. To do it check the corresponding checkboxes.

Use the Conversion options to choose a strand for saving sequence(s). Also you can translate sequence(s) to amino alphabet.

Also it is possible to specify whether to merge the exported sequences into a single sequence or store them as separate sequences. If you merge the sequences, you're allowed to select the gap symbols between sequences. This is the length of the insertion region between sequences that contain N symbols for nucleic or X for protein sequences.

### Export sequence with annotations

To export sequence with annotations choose Genbank or GFF format. The *Export with annotations* checkbox will be available. Check the checkbox and sequence will be exported with annotations.

# **Exporting Sequences as Alignment**

Suppose, we want to interpret FASTA file as multiple alignment. To do this, select a single or several sequence objects in the *Project View* window, click right mouse button to open the context menu and select the *Export Export sequences as alignment* item:

U -* UGENE - [fasta_example]	
🗞 File Actions Settings Tools Wir	ndow Help _ & X
4	i in ta ta Goi 🖉,₁ ▾ 1:1 ▾ 🕰 💘 💘 »
Project	🗙 🕜 Phaneropt 🔘 🖄 🏷 🗔 🖓 🛣
이 Name filter 	
Cobjects	empty
▲ 🖁 fasta_example	1 50 100 150 200 250 300 350 400 450 500 604 🕕
🖌 👌 [s] Phaneroptera_falcata	
S] Isophya_altaica_EF540	820 <b>*</b> D F <b>*</b> F E P N <b>*</b>
% [s] Bicolorana_bicolor_EF	540830 KTSNSSRIR 🔁
% [s] Roeseliana_roeseli % [s] Montana_monta	
8 [s] Morrana_morra	Add to view
S [s] Gampsocleis_sed	Unload selected objects AGATTAAGCTCGGCTTAATC
😽 [s] Deracantha_dera	
👌 [s] Zychia_baranovi	Lock document for editing E L E L R I L
🗞 [s] Tettigonia_viridi	Add • * N S G F *
% [s] Conocephalus_di	Import •
% [s] Conocephalus_sp % [s] Conocephalus_pq	Export   Export sequences
S [s] Mecopoda elono	Edit Export sequences as alignment
Bookmarks	Remove
🚖 fasta_example 🖷	Save selected documents
	Auto-annotations [fasta_example   Bicolorana
🔯 <u>2</u> : Tasks 🛄 <u>3</u> : Log	No active tasks 💡 🗎

The *Export Sequences as Alignment* dialog will appear where you can point the result alignment file location, to select a multiple alignment file format, to use Genbank "SOURCE" tags as a name of sequences for Genbank sequences and optionally add the created document to the current project:

U Export Sequen	ces as Alignment	X
Export to file		
File format to use	CLUSTALW	-
Add document	to the project	
🔲 Use Genbank 🕷	SOURCE" tags as a name of sequences (for Genbank sequences only	)
	Export Cancel H	lelp

## **Exporting Alignment to Sequence Format**

Select a single object with a sequence alignment in the *Project View* window and click the *Export Export alignment to sequence format* conte xt menu item:

U	-* UGENE - [aln_e	xamp	le [m] aln_	example	e]														-		X	3
	File Actions	Settin	gs Tools	Wind	wob	Help	)														- 8	×
	u 🗁 📄	€,	۹ 🔃		T,	ŵ		4	2	g	è	•	\$7		4							
1 . Droiact			×	Cons	ensi	us		T	C	A (	Ga	n t	Та	t	TA		T	тс	G	A G		
R			.aln					1	2		4	6	8		10	12		14	16	1		
	Copen vie Add to vi Inload se Lock doc Add Import			iew elected	• • •	T T T		A	_	т	T T T A T A	т	ТА	A	Т		G	A	18 18 18	*		
		$\square$	Export				•		E	kpor	t ali	gnn	nent t	o sec	quen	ce f	orm	at				
	Bookmarks		Edit Remove				•		Đ	xpor	t nu	iclei	c alig	nme	nt to	ami	ino	trans	latio	on		
		ŧ.	Save sele	cted do	cume	ents	_	ł												Þ		-
				Find:	<b>(</b>						] •		Ln 1	/3	Co	ol 1 /	604		Pos	1/6	01	6
	🔯 <u>2</u> : Tasks		📜 <u>3</u> : Log														I	No ad	tive	tasks	s 🍚	ė

The Convert Alignment to Separate Sequences dialog will appear:

U Convert Alignm	? <mark>X</mark>	
Export to file		
File format to use	FASTA	•
Add document	to the project	
Gap characters (	-)	
Keep	Trim	
	Export Cancel	Help

Here it is possible to specify the result file location, to select a sequence file format, to define whether to keep or remove gaps ('--' chars) in the aligned sequences and optionally add the created document to the current project.

# **Exporting Nucleic Alignment to Amino Translation**

Select a single object with a nucleic sequence alignment in the *Project View* window and click the *Export Export nucleic alignment to amino translation* context menu item:

U	-* UGEI	NE - (†	test_3_seq	[m] t	est_3_s	eq]												-						X	_
	File /	Action	ns Settin	ngs	Tools	Wind	low	Help	p															- 8	×
	. 6			۹			T,	Ý		9	g	2	•	\$7	₽][⊲	, Tr.									
1: Project		filter ts	Open vi	×	Cons	sensu	s		<b>T</b> t	4		6	8	10	)	12	14	1(	6	18	2(	)	2		
		8	Add to y Unload a Lock do Add Import	select			1	1	TT TC TT	A G	T I	тт	Α		Α	а т	т	c	A	G	с т	G	Α,	23	*
	Export Exp					Export alignment to sequence format Export nucleic alignment to amino translation										4									
	<u>2</u> :	Tasks		<u>3</u>	: Log	F	ind:	<b>(</b>							2	Ln 1	/3	Co	ol 1 /				1/5 tasks	~	<b>6</b>

The Export Nucleic Alignment to Amino Translation dialog will appear:

U Export Nucleic	Alignment to Amino Translation
Export to file	C:/work/ugene/data/samples/CLUSTALW/COI_transl.aln
File format to use	CLUSTALW
Amino translation	1. The Standard Genetic Code
Add document	to the project
Export range	
Whole alignment	nent
<ul> <li>Selected row</li> </ul>	'S
	Export Cancel Help

Here it is possible to specify the result file location, to select a file format and an amino translation, to export whole alignment or selected rows and optionally add the created document to the current project.

### **Export Sequences Associated with Annotation**

In UGENE you can export a sequence associated with an annotation. To do it select the annotation in the *Project View* window and click the *Export/Import Export corresponding sequence* context menu item:

🔺 🚹 human_T1.fa		terminal rep	eat	· · · · · · · · · · · · · · · · · · ·		•••••	•••••		
<ul> <li>% [s] human_T1 (l</li> <li>M murine.gb</li> <li>M human_T1 (l</li> </ul>	JCSC April 2002 1	200 400	600	800	1k 1.2	k 1.4k	1.6k 1	.8k 2k	( 2.2
8 [s] NC_001363									
(P [a] NC_001	Open view		•						
	Add to view		•	т р	_	V	G G		L
	Edit		• D	P	H P T	× R R	-	A S Q	; * A
	Export/Import		•	Ехро	rt annota	tions			CTA
R.	Remove selected ite	ms De	el 📃	Ехро	rt corresp	onding	sequenc	:e	32 GAT
	Open containing fol	der		Impo	ort annota	ations fr	om CSV	file	L
		r s	-	G W	G	Y	T A	L	*
		H F	' V	G	G	т р	P	L	S

The Export Selected Sequences dialog will appear:

U Export Selected Sequences	adde fast farms	-		? x
Export to file	1			
File format to use	FASTA			•
Export with annotations				
Add document to the project				
Use custom sequence name	murine			
Convertion options				
Save direct strand	Save complement strand	Save	both strands	
Save both strands	✓ Save all amino frames			
Use custom translation table	1. The Standard Genetic Code	2		-
Merge options				
Save as separate sequences	Merge sequences			
Add gap symbols between seque	nces			0
		Export	Cancel	Help

Here you can select the location of the result file and a sequence file format. You can choose to add newly created document to the current project and use custom sequence name. To do it check the corresponding checkboxes.

Use the Conversion options to choose a strand for saving sequence(s). Also you can translate sequence(s) to amino alphabet.

Also it is possible to specify whether to merge the exported sequences into a single sequence or store them as separate sequences. If you merge the sequences, you're allowed to select the gap symbols between sequences. This is the length of the insertion region between sequences that contain **N** symbols for nucleic or **X** for protein sequences.

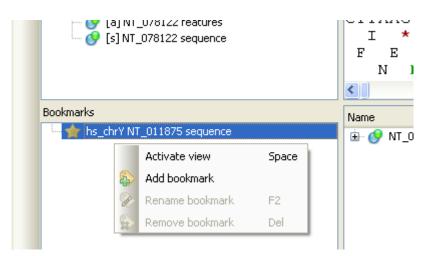
### Export sequence with annotations

To export sequence with annotations choose Genbank or GFF format. The *Export with annotations* checkbox will be available. Check the checkbox and sequence will be exported with annotations.

### **Using Bookmarks**

One of the most important features supported by most *Object Views* is an ability to save and restore visual view state. Saving and restoring visual state of an *Object View* enables rapid switching between different data regions and is similar to bookmarks used in Web browsers.

Initially an Object View is created as transient. It means that its state is not saved. To save current state of a view select an item with the view name in the Bookmarks part of the Project View windows and select the Add bookmark item in the context menu:



For every persistent view UGENE automatically saves the state of the view in the Auto saved bookmark when the view is closed.

Now, by activating bookmarks you can restore the original view state. For example for the Sequence View bookmarks you can store a visual position and zoom scale for the sequence region.

4[2] ♥ 1[6] ♥ 4[2] ♥ 1[6] ♥	NT_025441 features NT_025441 sequence NT_025975 features IT_025975 sequence NT_078122 features IT_078122 sequence		1 CT F	raă c E N	-		TTA	TCT L F S	20 TAA I Q N		AGC D R
- 🖳 Auto	NT_011875 sequence o saved poolsmark		Name 🕀 - (		[_0118]	75 feat	ures (h	is_chr¥	.gbk.gz	]	4
\$ \$ \$	Activate view Add bookmark Rename bookmark Remove bookmark	Space F2 Del									

Use the F2 keyboard shortcut to rename a bookmark. To remove a bookmark press the Delete key.

UGENE has limited set of built-in *Object Views*. Extensions modules or plugins can be used to adjust the existing views or to add new views to the tool.

# **Exporting Project**

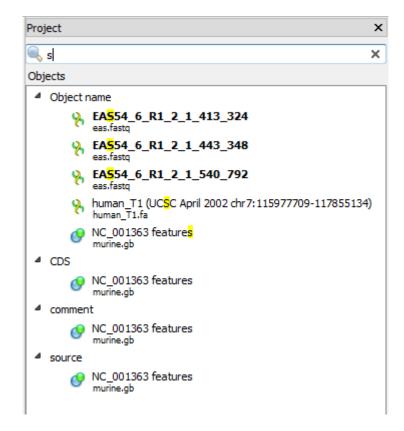
All the opened documents and bookmarks (along with the corresponding views states) can be saved within a project file. To do so, select *File Export Project*. It will invoke the *Export project* dialog, where you can select the destination folder and the project file name.

U	-* UGENE - [NC_014267.1 [s] NC_0142	67]
8	File Actions Settings Tools W	/indow Help
	. 🗁 🗎 🗛 🐚 🙀	19 19 19 19 19 19 19 19 19 19 19 19 19 1
E I: Project	Project × Name filter Objects ▲	VP_003734501.1       rpl34       rpl34       Restriction Sites Map         VP_003734501.1       rpl34       rpl34       rpl34       BamH1: 0 sites         VP_003734501.1       rpl34       rtrl(uag)       rtrl(uag)       rtrl(uag)       rtrl(uag)         vP_003734501.1       sold       sold       rtrl(uag)       rtrl(
	Bookmarks ☆ human_T1 [s] human_T1 ☆ NC_014267.1 [s] NC_014	YP_003734571.1 orf93 YP_003734570.1 orf92 rp0C1 RNA RNA
		Image: Source     Image: Source       Image: Source
	0 2: Tasks 1 3: Log	No active tasks 💡 🕍

To load a saved project later, select File Open and specify the path to the project file.

### **Search in Project**

Use the search field in the project view to search in the whole project:



# **Options Panel**

The Options Panel is available in the Sequence View and in the Assembly Browser. By default, it is closed. To open a tab of the Options Panel click on the corresponding icon at the right side of a Sequence View or Assembly Browser window. To close the tab click again on the

tab icon.

More detailed information about different Options Panel tabs can be found in the following chapters:

- Options Panel in Sequence View
  - Information about Sequence
  - Search in Sequence
  - Highlighting Annotations
- Options Panel in Assembly Browser
  - Navigation in Assembly Browser
  - Assembly Browser Settings
  - Assembly Statistic

# **Adding and Removing Plugins**

A plugin is a dynamically loaded module that adds a new functionality to UGENE.

To manage plugins select the Settings Plugins main menu item. The Plugin Viewer window will appear:

	e <mark>w Project* UGENE - [Pl</mark> e <u>A</u> ctions <u>S</u> ettings <u>T</u> ool	<u> </u>		
: 🗔				
ы	DNA Export	Name 🔶	State	Description
1: Project	Routines to export or align multiple DNA sequences Description of the selected plugin	<ul> <li>BALL</li> <li>BioStruct3D Viewer Plugin</li> <li>BioStruct3D Viewer Plugin</li> <li>Bowtie</li> <li>CUDA Support</li> <li>ChromaView</li> <li>CircularView</li> <li>DNA Annotator</li> <li>DNA Export</li> <li>DNA GraphPack</li> <li>DNA Statistics</li> <li>Dotplot</li> <li>External tool support</li> <li>GORIV</li> <li>HMM2</li> <li>HMM3</li> <li>Kalign</li> <li>MUSCLE</li> <li>ORF Marker</li> <li>Optimized Smith-Waterm</li> <li>Phylip plugin</li> </ul>	On On On On On	A port of BALL framework for m Biological 3D Structure Viewer D An ultrafast memory-efficient sh Utility plugin for CUDA-enabled Chromatograms visualization Enables drawing of DNA sequen This plugin contains routines to <b>Routines to export or align multi</b> This plugin contains a set of gra Provides statistical reports for s Build dotplot for sequences Runs other external tools GORIV protein secondary struct Based on HMMER 2.3.2 package Biological sequence analysis usin HMM profile tools. Plugin is base A port of Kalign package for mul Finds open reading frames (ORF Various implementations of Smit PHYLIP(the PHYLogeny Inferenc
		🥵 Primer 3 🥵 PsiPred	On On	Integrated tool for PCR primers PsiPred protein secondary struc
	2: Tasks 🛛 🛄 3: I	og		»

The window shows the list of available plugins.

To add or remove plugins use the Add plugin and the Remove plugin items available in the Plugin Viewer context menu:

📕 🤹 DNA Anr	notator	On
🕷 DNA Exc	iort	On
🥵 DNA GI	Add plugin	On
🛛 🚭 DNA St	Remove plugin	On
🛛 🍕 Direct So	ocket Transport	On

When you select the *Remove plugin* item for a plugin, the plugin's status is changed to the *to remove after restart* value. The *Remove plugin* is no more available in the context menu of the plugin. Instead the *Enable plugin* item appears in the context menu:

Oirect Se	ocket Transport	_	On
🧶 Dr tpl	A dal aluaia		to remove after restart
🧠 Exter	Add plugin		On
🥞 GORI	Enable plugin		On
🥞 нммz			On

If you select this item the plugin will be enabled again, i.e. it will not be removed after restart. Otherwise, the plugin will not be available after UGENE restart.

# Searching NCBI Genbank

UGENE allows searching data in NCBI GenBank remote database. To do this open the following dialog by *File->Search NCBI Genbank* main menu:

NCBI Sequence Search	8
Term: All fields 😂 🖸 Database: nucleotide 😂	+
Search query:	
Results:	
ID ^ Desc	Size
Result limit: 20	Search
Неір	Close 🖉 Download

To search data in the nucleotide or protein databases enter a general text query to the search field, select the database and click on the *Sear ch* button. You can use a protein name, gene name, or gene symbol directly. Searching for a submitter or author name in the following format will produce the best results.

Use the boolean operator AND to find records that contain every one of your search terms, the intersection of search results.

Use the boolean operator OR to find records that include one of several search terms, the union of search results.

Use the boolean operator NOT to exclude records matching a search term.

To limit results use the Result limit field.

After you click the *Search* button, UGENE searches the biological objects and shows it in the *Results* field. You can download the object(s). Select one or several objects (for selecting several objects use the *Ctrl* button) and click the *Download* button. The dialog will appear:

n Remote Database	? ×
KJ690776	
NCBI GenBank (DNA sequence)	•
[	
gb	•
OK Cancel	Help
	KJ690776 NCBI GenBank (DNA sequence)

After you click the OK button, UGENE downloads the biological objects and adds it to the current project.

# Fetching Data from Remote Database

UGENE allows fetching data from remote biological databases such as NCBI GenBank, NCBI protein sequence database, and some others.

To fetch data select the File Access remote database ... item in the main menu.

The dialog will appear:

• Fetch Data fro	om Remote Database 😣
Resource ID:	
Database:	NCBI GenBank (DNA sequence)
Save to directory:	/home//.UGENE_downloaded
✓ Add to project	
✓ Force download	the appropriate sequence
	k DNA accession number. For example: <u>NC_001363</u> or <u>D11266</u> I multiple items by separating IDs with space or semicolon.
Help	Cancel OK

Here you need to enter the unique id of the biological object and choose a database. The following databases are available: NCBI Genbank (DNA sequence), NCBI protein sequence database, ENSEMBL, PDB, SWISS-PROT, UniProtKB/Swiss-Prot, UniProtKB/TrEMBL. Unique identifiers are different for various databases. For example, for NCBI GenBank such unique id could be Accession Number or NCBI GI number. Optionally, you can add the document to a project and browse for a directory to save the fetched file to.

After you click the OK button, UGENE downloads the biological object (DNA sequence, protein sequence, 3d model, etc.) and adds it to the current project.

If something goes wrong check the Log View, it will help you to diagnose the problem.

## **UGENE** Application Settings

To open UGENE Application Settings dialog choose the Settings Preferences item in the main menu.

To open UGENE Application Settings dialog in Mac OS use the Unipro UGENE->Preferences menu item.

The following settings are available:

- General
- Resources
- Network
- File Format
- Directories
- Logging
- Alignment Color Scheme External Tools Settings
- Workflow Designer Settings

OpenCL

#### General

Application Settings	8
General       General         Resources       Language of User Interface (applied after restart)         Directories       Logging         Alignment Color Scheme       Autodetection         OpenCL       External Tools         Workflow Designer       Fusion         Project       Image: Open last project at startup         Always save new project on exit Image: Open last project at startup       Statistical reports         Enable statistical reports collecting       Updates         Image: Open last project starts       Default settings         Reset settings to default on the next run       Reset settings to default on the next run	
Help	Cancel OK

The following settings are available on the tab:

Language of User Interface (applied after restart) — here you can select UGENE localization. Currently, available localizations are EN and RU. The default value (*Autodetection*) specifies that UGENE should use the operating system regional options to select the localization. This setting is applied only after UGENE is reopened.

Appearance — defines the appearance of the application.

Open last project at startup — if the option is checked, the last project is opened when UGENE is started. Also, you can choose default settings for a saving project.

Enable statistical reports collecting — collects information about UGENE usage and sends it to the UGENE team to help improve the application.

The collected information includes:

- 1. System info: UGENE version, OS name, Qt version, etc.
- 2. Counters info: number of launches of certain tasks (e.g. HMM search, MUSCLE align).

The collected information DOESN'T include any personal data.

Check for updates when UGENE starts - checks for updates when UGENE starts.

Default settings - this option resets the default settings on the next run.

### Resources

<u>/</u>}

Application Settings				? <mark>x</mark>
General Resources	Resources			
Resources Network File Format Directories Logging Alignment Color Scheme External Tools	Application resources Optimize for CPU count Tasks memory limit Threads limit		8 1536Mb 1000	
Genome Aligner Workflow Designer OpenCL				
		ОК	Cancel	Help

On the Resources tab you can set resources that can be used by the application: Optimize for CPU count, Tasks memory limit and Threads limit.

## Network

Resources Network File Format Directories Logging Alignment Color Sch External Tools Genome Aligner Workflow Designer	Preferred Web browser         Image: System default browser         Image: Custom browser         Image: Custom browser         Remote request settings         Remote request timeout         60 sec	
-		
OpenCL	Proxy   Type:   Server:   HTTP   Use authentication with HTTP proxy:   Login   Password   Do not use proxy on following addresses (see   SSL settings   Secure Socket protocol:   SslV3	Port: 0 eparate line for each):

On the Network settings tab of the dialog you can specify Proxy server parameters, select SSL settings and configure the Remote request timeout.

Preferred Web browser — you can use either System default browser or specify some other browser.

## **File Format**

Application Settings		? ×
General Resources Network File Format Logging Alignment Color Scheme External Tools Genome Aligner OpenCL Workflow Designer	File Format Sequence annotations Create annotations for case switchings:	Don't use case annotations <ul> <li></li></ul>
		OK Cancel Help

The Sequence Annotations settings allows to use upper/lower case annotations during the file reading process.

Format options:

- 1. Don't use case annotations (default mode) usual sequence reading and writing.
- 2. Use lower case annotation sequences are read and annotations with names lower\_case are added. When these sequences are written to file then the case becomes like original the file case (the case is saved).
- 3. Use upper case annotation there is a similar behavior but with "upper\_case" annotations.

### **Directories**

Application Settings		8
<ul> <li>Application Settings</li> <li>General Resources Network File Format</li> <li>Directories</li> <li>Logging Alignment Color Scheme OpenCL Workflow Designer External Tools</li> </ul>	Directories  Downloaded data  /home//.UGENE_downloaded  Documents  /home//Documents/UGENE_Data  Temporary files  /tmp  File storage  /home//.UGENE_files  Cleanup storage  Build indexes  /tmp/ugene_tmp/p17364/aligner	
Неір	Cancel	

The following settings are available on the tab:

Downloaded data - specifies the path where files downloaded from the remote databases will be stored.

Documents - the path where will be stored UGENE data.

Temporary files — the path where will be stored temporary files.

File storage — the path where will be stored UGENE files.

Build indexes - the path where will be stored UGENE Genome Aligner indexes.

# Logging

neral sources	Logging				
etwork	Category	TRACE	DETAILS	INFO	ERROR
le Format	< <al>&gt;</al>	Sample text	Sample text	🔽 Sample text	Sample text
ogging lignment Color Scheme	Algorithms				
ternal Tools	Console				
enome Aligner penCL	Core Services				
/orkflow Designer	Input/Output				V
	Performance				<b>V</b>
	Remote Service				
	Scripts				
	Tasks				
	Teamcity Integration				
	Teamcity Log				
	User Actions			V	<b>V</b>
	User Interface				<b>V</b>
	Log format				
	Show date	Date format:	hh:mm		
	Show log level	I 🔲 Show log categ	ory 📝 Enable colo	red log output	
	Save output t	o file:			

On the Logging tab you can select type of log information (ERROR, INFO, DETAILS, TRACE) for each Category that will be output to the Lo g View.

You can select format for each log message by checking the Show date, Show log level and Show log category options.

Log	×
[INFO][09:59] Starting {Open new 'Sequence view'} task [INFO][09:59] Task {Open new 'Sequence view'} finished [INFO][09:59] Starting {Open new 'Sequence view'} task	<b>A</b>
[INFO][09:59] Task {Open new 'Sequence view'} finished [INFO][09:59] Starting {Open new 'Sequence view'} task [INFO][09:59] Task {Open new 'Sequence view'} finished	E

## **Alignment Color Scheme**

U Application Settings	ି ଅନ୍ୟୁ ଅନୁସାର କରି କରି କରି କରି କରି କରି କରି କରି କରି କର
General Resources Network File Format Logging Alignment Color Scheme External Tools Genome Aligner OpenCL Workflow Designer	Alignment Color Scheme Create and modify custom color schemes Directory to save color scheme: Change color scheme Create color scheme Delete
	OK Cancel Help

On the Alignment Color Scheme tab you can create, change and delete custom color schemes.

## **External Tools Settings**

Here you can set the paths to the external tools executable files.

	plication Settings			? ×
	neral Ex	cternal Tools		
Net	work Format	You can select path for external to	pols package	Browse
	gging gnment Color Scheme	Supported tools		
	ernal Tools	Name	Path	*
Ор	nome Aligner enCL rkflow Designer	<ul> <li>Bowtie 2</li> <li>Bowtie 2 index i</li> <li>Bowtie 2 build in</li> <li>Bowtie 2 aligner</li> <li>Bowtie</li> <li>Bowtie build ind</li> <li>Bowtie aligner</li> <li>BLAST+</li> <li>BlastDBCmd</li> <li>RPSBlast</li> <li>TBlastX</li> <li>TBlastX</li> <li>BlastX</li> </ul>	C:\work\ext_tools_win_64-bit\blast+-2.2.28\bin\blastdbcmd.exe C:\work\ext_tools_win_64-bit\blast+-2.2.28\bin\rpsblast.exe C:\work\ext_tools_win_64-bit\blast+-2.2.28\bin\tblastx.exe C:\work\ext_tools_win_64-bit\blast+-2.2.28\bin\tblastx.exe C:\work\ext_tools_win_64-bit\blast+-2.2.28\bin\blastx.exe	
			OK Cancel	Help

# **Workflow Designer Settings**

Use this tab to configure the Workflow Designer settings:

General       Workflow Designer         Resources       Scene appearance         Network       Scene appearance         File Format       Scene appearance         Logging       Alignment Color Scheme         External Tools       Snap to grid         Genome Aligner       Snap to grid         OpenCL       Element style         Workflow Designer       Element font         Morkflow Designer       Mis Shell Dlg 2         Workflow Designer       Track running progress         Image: Track running progress       Track running progress         Image: Directories       Use directory for output files         Image: Directory for custom elements with scripts       Image: Directory for custom elements with command line tools         Image: Directory for included schema elements       Image: Directory for included schema elements	Application Settings	
<ul> <li>Track running progress</li> <li>Enable debugger</li> <li>Directories</li> <li>Use directory for output files</li> <li>Image: Directory for custom elements with scripts</li> <li>Directory for custom elements with command line tools</li> <li>Directory for included schema elements</li> </ul>	Resources Network File Format Logging Alignment Color Scheme External Tools Genome Aligner OpenCL	Scene appearance  Show grid  Snap to grid Element style Element font MS Shell Dig 2 Element background color
Directory for custom elements with command line tools		<ul> <li>✓ Track running progress</li> <li>✓ Enable debugger</li> <li>Directories</li> <li>Use directory for output files</li> </ul>
		Directory for custom elements with command line tools

# OpenCL

If you have a video card that supports OpenCL you can use it to speed up some calculations in UGENE.

To do it install the latest video card driver and check the corresponding check box:

Application Settings		? <mark>x</mark>
General Resources Network File Format Logging Alignment Color Scheme External Tools Genome Aligner OpenCL Workflow Designer	OpenCL The following OpenCL-enabled GPUs are detected. Check the GPUs to use for accelerating algorithms computations. ☑ Intel(R) Corporation Intel(R) HD Graphics 4600 1297 Mb ☑ NVIDIA Corporation GeForce GT 740M 2048 Mb	
		OK Cancel Help

Now you can, for example, use OpenCL optimization for the Smith-Waterman algorithm.

# **Sequence View**

- Sequence View Components
- Global Actions
- Sequence Toolbars
- Sequence Overview
- Zoom View
- Details View
- Information about Sequence
   Manipulating Sequence
  - Show/Hide Sequence View Components
  - Showing Sequence in Multiple Lines
    - Translating Nucleotide Sequence
    - Zooming Sequence
    - Creating New Ruler
    - Going To Position
    - Selecting Sequence Region
    - Copying and Pasting Sequence
    - Editing Sequence
    - Searching in Sequence
      - Load Patterns from File
      - Search Algorithm
      - Search in
      - Other Settings
      - Annotations Settings
    - Exporting Selected Sequence Region
    - Exporting Sequence of Selected Annotations
  - Exporting Sequence Image
  - Locking and Synchronize Ranges of Several Sequences
  - Multiple Sequence Opening
- Annotations Editor
  - "db\_xref" Qualifier
  - Automatic Annotations Highlighting
  - The "comment" Annotation
- Manipulating Annotations
  - Creating Annotation
  - Selecting Annotations
  - Editing Annotation
  - Highlighting Annotations
    - Annotations Color
      - Annotations Visability
      - Show on Translation
      - Captions on Annotations
  - Creating and Editing Qualifier
  - Adding Column for Qualifier
  - Copying Qualifier Text
  - Finding Qualifier
  - Deleting Annotations and Qualifiers
  - Importing Annotations and Qualifiers
     Importing Annotations from CSV
  - Exporting Annotations

# **Sequence View Components**

The Sequence View is one of the major Object Views in UGENE aimed to visualize and edit DNA, RNA or protein sequences along with their properties like annotations, chromatograms, 3D models, statistical data, etc.

For each file UGENE analyzes the file content and automatically opens the most appropriate view.

To activate the *Sequence View* open any file with at least one sequence. For example you can use the \$UGENE/data/samples/EMBL/AF177 870.emb file provided with UGENE.

After opening the file in UGENE the Sequence View window appears:

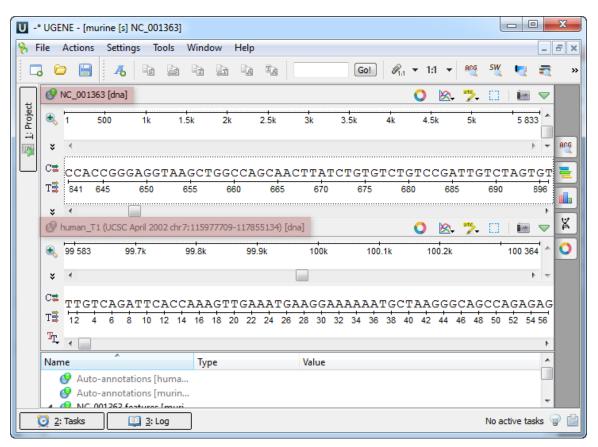
U -	-* UGENE - [murine [s] NC_001363]	
8	File Actions Settings Tools Window Help	×
	🔁 🗁 🔚 📕 🖓 🕼 🎭 🏤 🎰 🖓 🦾 🌚 😡	»
<b></b>	🔮 NC_001363 [dna] 🛛 🔘 🖄 🖓 📋 🗐 🤝	
<u>1</u> : Project	Σ 1 500 1k 1.5k 2k 2.5k 3k 3.5k 4k 4.5k 5k 🔺 5.833	
	CDS (4) NP 040335.1 NP_5977 NP_5977 NP 040336.1 NP	
<b>W</b>	🖕 1 500 1k 1.5k 2k 2.5k 3k 3.5k 4k 4.5k 5k 5833	4
	5245 <b>★</b> [589 bp] →	
	CzGINERPHP*VAS*LK*RHFA	
	GMKDPTRRWQASLSNATLQ	5
	T <sub>L</sub> GGAATGAAAGACCCCCACCCGTAGGTGGCAAGCTAGCTTAAGTAACGCCACTTTGCA	╡
	5 243 5250 5255 5260 5265 5270 5275 5280 5285 5290 5 298	2
	CCTTACTTTCTGGGGTGGGCATCCACCGTTCGATCGAATTCATTGCGGTGAAACGT	
	PIFSGVRLHCALKLLAVKC	
	F S L G W G Y T A L * S L Y R W K A	
	Name Type Value ^	
	▶ misc_feature Misc. Feature 2590	
	▷ ■ misc_feature         Misc. Feature         52455833           ▷ Ø source (0, 1)         ■	
	· · · · · · · · · · · · · · · · · · ·	
	2: Tasks     Image: Second secon	ä

After the view is opened you can see a set of new buttons in the toolbar area. The actions provided by these buttons are available for all sequences opened in the view. In the picture below these buttons are pointed by the "Global actions" arrow.

Below the toolbar there is an area for a single or several sequences. For each sequence a smaller toolbar with actions for the sequence and the following areas are available:

<b>U</b> -	* UGENE - [mi	urine [s] N	IC_001363	]	G	Blobal acti	ons		-			x
8	File Actions	Setting	s Tools	Window	Help						_	Ξ×
	. 🗁 🗎	4				A .	Go	I 🖉 1:1	<ul> <li>■ 161 ×</li> </ul>	acg SW	5	»
	O NC_0013	53 [dna]		— Seque	nce nam	e			0 🖄	***		
1: Project	Σ 1	5004	1k	1.5k	2k 2.5	5k 3k	3.5k	4k	4.5k	5k	5 833	
	Sacil (	1)								T	•	
	5' t	ermina	l repe	at								RCG
		200	400	600	800	1k	1.2k	1.4k	1.6k	1.8k	2 048	2
			Zoom Vi	i_l ew		C	urrent s	oquopo			+ +	
		P D	IJ		тт	GR	* A			ьт	C	
		G !	гт	D P	P P	GG		LA	S N		3 7	X
	TE A Q		ails Vie			R E	V S	W P	A	т т	<b>L</b> C	0
	- <del> </del>			+ +		GGGAGG		TGGCC		CTTAT	CTGTG	
	622 62 GGGT			35 64 CTCCCCT		5 650 CCCTCC	655 ATTCC	060 ACCCC	665 TCCTT	670 מידי בי בי בי בי	676	
	W		30100 G G	V W	W R		L	Q G	A V		R H	
	G	P V	V	S G	G G	P P	L S	A	L L	K D	т	
	•										+	
	Name	~		Туре		Value					*	
	🔺 🧭 n	nisc_featu	ıre (0, 2)									
		l misc_fea		Misc. F	eature	2590					_	
		l misc_fea ource (0,		Misc. F	eature	5245.	5833				-	
	🔯 <u>2</u> : Tasks		<u>3</u> : Log							No ac	tive tasks 🤇	2 👜

An example of the Sequence View with several sequences:



You can change the focus by clicking on the corresponding sequence area. All sequences that are not in focus have the sequence name and icon disabled.

The bottom area of the Sequence View is the Annotations Editor. It contains a tree-like structure of all annotations available for all sequences shown in the Sequence View and can be used to perform various actions on annotations: create a new annotation, modify the existing one, group, sort, etc.

# **Global Actions**



The global action toolbar provides possibility to go to the specified position (in all sequences at the same time).

Also it allows to lock or adjust ranges of sequences in the same Sequence View. See this paragraph for details.

## **Sequence Toolbars**

A brief description of the sequence toolbars buttons is shown on the picture below:

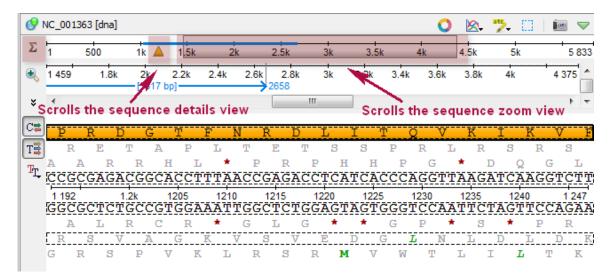
🔮 huma	n_T14	JCSC A	pril 2002 e ann	2 chr7: otati	11597 on de	7709-1 ensit	178551 y graj	34) [dna ph	3]	2.			0	×,	E	Σ		ACG		×
Σ	_	20k		40k		<u> 60k</u>		80k	/	100	k	120	k	140k		160k		180k	20	196
🔍 734		20k Auto	30k matic	40k ann	50k otati	60k onsh	70k nighlig	80k ghting	90k 9	00k Cap	110k oture	120k scre	130k en		150k phs		170k oggle	180k view	200 0	33
	Wra	ap⊴se	utton: qüen	cē	A	A	equer	nce re	gion	н	Sho S 1	w cir	cutar C P	view P	s	Re P	mov P	e sequ P	uence I	) T
			mple mino i CTC1	trans	latio	hs '	* E AGTI	P L CCT	CTC	T Q ACA	U GTC	g FCAF		P 1 P CCC	F F FTC1	S CCT	s ccc	P N CCAA	л П П П П П П П П П	C
	785 54 AAG		54799 GAG2	-	54.8k CGT(	548 CGA		54810 AGGA		815 TGT	5482 CAG2	0 5 AGTI	4825 T G G	5483 GGG		4835 AGGA	548 .GGG	40 GGTI	54 84	
E	K An	ਤ ਜੈno t	E Fansf	* ation	C C	S A 1	T L E	G E E	R 🔸	⊽ 	T I R	E E L	r G	F R G	R E	R G	G G	W G I	D : V	G
R	R	V	R	L	L	*	Ν	R	Ε	С	D	*	V	G 1	K E	EE	G	L	*	

See also:

- Toggling Views
- Capturing Screenshot
- Zooming Sequence
- Showing and Hiding Translations
- Selecting Sequence Region
- Showing Sequence in Multiple Lines

# **Sequence Overview**

The Sequence overview is an area of the Sequence View below the sequence toolbar. It shows the sequence in whole and provides handy navigation in the Sequence zoom view and the Sequence details view.



When the sigma button (in the left part of the *Sequence overview*) is pressed, density of annotations in the sequence is shown. For example in the picture below there are annotations in the parts of the sequence that are marked with dark grey color:



See also:

- Sequence Zoom View
- Sequence Details View

# **Zoom View**

The Sequence zoom view is designed to provide flexible tools for navigation in large annotated sequence regions.

The most Sequence zoom view space is used to visualize annotations for the sequence. The annotations are organized in rows by their names. If two annotations with the same name overlap, an extra row is created. For every row the name and the total number of annotations in the row are shown with a light grey text at the left part of the area.

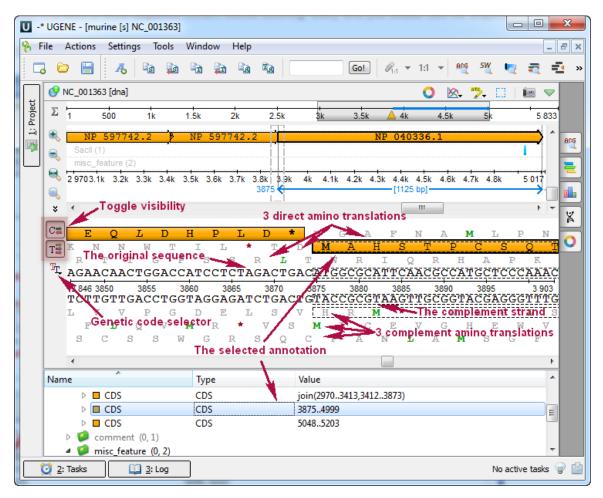
0	NC_014	4267 [dna]										0	0 🖄	***		• -
Σ	1	10k	20k	30k	40k	50k	60k	70k	80k	( 🔺 )	90k	100k	110k	120k	130k	140 426
æ,		LS	C; lar		gle co	yqq							all Bir	ngle (	сору	······
	ncR	NA (1) 💻	🗲 N	ame an	d numl	ber of a	annotat	ions in	the r	ow						
e	rRN	A (6) eat_regi <b>Z</b> (	oming	A	nnotat	ions 🛫		rRNA	rRNA regio	) m						
	tRN/ tmRl	A (30) NA (1)	Rule													
	56 16	7 60	k 62k	64k	66k	68k 7	0k 72k	74k	76k	78k	80k	82k	84k	86k	88k 91	274
									_				ii			
	4															P

Below the annotation rows there is a ruler to show coordinates in the sequence.

# **Details View**

The Sequence details view is a supplementary component of the Sequence overview. It is used to show sequence content without zooming. Every time you double click the sequence in the Sequence overview area or select an annotation, the corresponding sequence position is made visible in the Sequence details view.

For a DNA sequence the Sequence details view automatically shows complement DNA strand and 6 amino translation frames.



See also:

- Navigating the Details View using the Sequence Overview
- Translating Nucleotide Sequence

### Information about Sequence

Context information about a sequence can be found on the *Statistics* tab in the *Options Panel*. All information is contextual, i.e. it shows statistics about the currently selected region (on the selected sequence). The tab includes information about:

- Common statistics
  - Length number of bases in the analyzed sequence
  - GC content the molar percentage of guanine and cytosine bases in an oligonucleotide sequence
  - Molar weight is the sum of the atomic masses of the constituent atoms for 1 mole of oligonucleotide
  - Molar ext. coefficient the molar extinction coefficient is a physical constant that is unique for each sequence and describes the amount of absorbance at 260nm (A<sub>260</sub>) of 1 mole/L DNA solution measured in 1 cm path-length cuvette
  - Melting TM melting temperature is the temperature at which an oligonucleotide duplex is 50% in single-stranded form and 50% in double-stranded form
  - nmole/OD<sub>260</sub> the amount of oligonucleotide in nanomoles that, when dissolved in 1 mL volume, results in 1 unit of absorbance at 260 nm with a standard 1 cm path-length cuvette
  - g/OD<sub>260</sub> the amount of oligonucleotide in micrograms that, when dissolved in 1 mL volume, results in 1 unit of absorbance at 260 nm with a standard 1 cm path-length cuvette
- Characters occurrence
- Dinucleotides occurrence (for sequences with the standard DNA and RNA alphabets)

1       500       tk       1.5k       2k       ak       3.5k       4k       4.5k       5k       5k33       Common Statistics         1       500       tk       1.5k       2k       25k       ak       3.5k       4k       4.5k       5k33       Common Statistics         1       500       tk       1.5k       2k       25k       ak       4.5k       5k33       Common Statistics         1       500       tk       1.5k       2k       25k       ak       4.5k       5k33       Common Statistics         1       500       tk       1.5k       2k       2k       5k33       Ak       4k4       5k5       5833       Ak       F       F       5833       F       F       F       5833       F       F       F       5833       F	🔮 NC_001363 [dna]	Ti T <sub>r</sub> 🖮 🔍 🔍	🌒 🔍 🔻	Statistics	
1       500       1k       1.5k       2k       2.5k       3k       3.5k       4k       4.5k       5k       5.833       G       G       G       Content:       52.12%       Molar Weight:       181765.49 Da         4       A       A       A       V       A       A       C       R       M       Q       A       S       P       M       M       S       H       Molar Weight:       181765.49 Da         V       A       A       V       L       T       K       D       A       C       K       I       T       Molar Weight:       181765.49 Da         V       A       A       V       L       T       K       D       A       C       K       I       T       Molar Veight:       181765.49 Da         V       A       A       V       L       K       D       A       C       K       I       T       Molar Veight:       181765.49 Da       Molar Veight:       181765.49 Da         V       A       A       A       A       A       N       A       I       Image: C       Da       Image: C       Da       Molar Veight:       181765.0000       A			5 833 Σ	<ul> <li>Common Statistics</li> </ul>	
Image: Second status       Image: Second status <t< td=""><td>1 500 1k 1.5k 2k 2.5k 3k 3.5k</td><td>4k 4.5k 5k</td><td>5 833 🔺</td><td>Length: 589</td><td></td></t<>	1 500 1k 1.5k 2k 2.5k 3k 3.5k	4k 4.5k 5k	5 833 🔺	Length: 589	
<ul> <li></li></ul>	€[589 bp]→590				12%
3       3				-	
V       A       A       I       A       V       L       T       M       G       Q       P         TAGCAGCCCATTGCCGTACTGACAAAGGATGCAGGCAAGGCTAACCAAGGCCAC       T       M       G       Q       P       Immole/ODsec       0.15       19/0Dsec       28.61       W         Zey74       2860       2895       2990       2995       3k       3005       3010       3015       3020       3025       3020         A       A       M       A       T       S       V       F       S       A       P       L       S       V       M       P       C       G       S         L       W       Q       R       V       S       L       P       H       L       C       A       M       A       T       S       V       P       A       V       P       A       V       D					
2974       2980       2985       2990       2995       3k       3005       3015       3020       3025       3030         ATCGTCGGTAACGGCATGACTGTTTCCTACGTCGTTCGATTGGTACCCTGTCGGTG       A       M       A       T       S       V       P       C       S       V       P       C       S       V       P       C       S       V       P       C       S       V       P       C       C       D       Dinudeotides       A       C       C       C       C       C       A       L		10		-	<b>I</b>
2 974       2880       2985       2995       3k       3010       3011       3020       3025       3030         ATCGTCGGTAACGGCATGACTGTTTCCTACGTCCGATTGGATGGTACCTGTCGGTG       A       M       A       T       S       V       F       S       V       M       P       C       G       S       M       P       C       G       S       V       M       P       C       G       S       M       P       A       V       A       A       M       A       T       S       V       M       P       C       G       S       U       V       A       V       V       A       V       V       A       V       V       V       V       V       V       V       V       A       V       V       A       V       V       A       V       V       A       V       V       A       V       V       A       V       V       A       V       A       S       A       S       A       S       A       S       A       S       A       S       A       S       A       S       A       A       S       A       S       S       S       A	TAGCAGCCATTGCCGTACTGACAAAGGATGCAGGC	AAGCTAACCATGGO	GACAGCCAC	μg/OD <sub>260</sub> : 28.	61 X
A       M       A       T       S       V       F       S       V       M       P       C       G       S       C       G       Name       C       A       L       W       P       V       A       V       A       S       O       A       S       O       A       S       O       A       S       O       A       S       S       A       S       S       A       S       S       A       A       S       O       A       A       S       O       A       A       S       A       A       S       O       A       A       S       O       A       A       S       O       A       A       S       O       A       A       S       O       A       A       S       O       A       A       S       O       A       A       S       O       A       C       S       A       A       S       O       A       A       S       O       A       A       S       O       A       A       S       O       A       C       S       S       A       S       O       C       S       A       S <td></td> <td></td> <td></td> <td></td> <td></td>					
L L W Q R V S L P H L C A L W P V A V Y C G N G Y Q C L I C A L G H S L W AA: 50 AC: 28 AC: 28 AC: 44 AT: 29 CA: 41 CC: 54 CG: 27 CA: 41 CC: 54 CD: 38 CA: 41 CC: 54 CD: 38 CD: 3					•
I C G N G I Q C L I C A I C C S I C C C C C C C C C C C C C C C	L L W Q R V S L P H L C	A L W P			
Name       Type       Value       AG: 44         Name       Type       Value       AT: 29         Orgen       Orgen       Case 41         Orgen       Case 41       Case 41         Orgen </td <td>Y C G N G Y Q C L I C A</td> <td>L * G H S</td> <td>3 <b>L</b> W</td> <td></td> <td></td>	Y C G N G Y Q C L I C A	L * G H S	3 <b>L</b> W		
Name       Type       Value       AT: 29			4		
▲ ● NC_001363 features [murine.gb]       CC: 54         ▲ ● CDS (0, 4)       CC: 54         ▶ ■ CDS       CDS         ■ CDS       CDS         □ CDS       CDS         ○ comment (0, 1)       TA: 22         ○ misc_feature (0, 2)       Misc. Feature 2.590         ○ misc_feature       Misc. Feature 52455833         ○ Source (0, 1)       T: 29	Name	Туре	Value		
▲ CDS (0, 4)       CG: 27         ▶ CDS       CDS       10422658       CT: 41         ▶ CDS       CDS       join(29703413,       GA: 37         ▶ CDS       CDS       CDS       38754999       GC: 38         ▶ CDS       CDS       CDS       50485203       GG: 37         ▶ CDS       CDS       CDS       50485203       GG: 37         ▶ CDS       CDS       CDS       GG: 32         ■ misc_feature (0, 2)       TA: 22       TA: 22         ▶ misc_feature (0, 2)       Misc. Feature       2.590       TC: 43         ▶ misc_feature       Misc. Feature       52455833       TG: 36         ▶ source (0, 1)       TH: 29       THEP				<b>CA:</b> 41	
▷ ■ CDS       CDS       10422658       CT: 41         ▷ ■ CDS       CDS       join(29703413, GA: 37)         ▷ ■ CDS       CDS       CDS       38754999         ▷ ■ CDS       CDS       CDS       50485203       GG: 37         ▷ ■ CDS       CDS       CDS       50485203       GG: 37         ▷ ■ cost_feature (0, 2)       TA: 22       TA: 22         ▷ ■ misc_feature       Misc. Feature       2590       TC: 43         ▷ ■ misc_feature       Misc. Feature       52455833       TG: 36         ▷ ◎ source (0, 1)       TT: 29       THEIP       Help				CC: 54	
▷ ■ CDS       CDS       join(2970.3413, GA: 37       GA: 37         ▷ ■ CDS       CDS       3875.4999       GC: 38         ▷ ■ CDS       CDS       5048.5203       GG: 37         ▷ ■ cDS       CDS       5048.5203       GG: 37         ▷ ■ misc_feature (0, 2)       TA: 22       TA: 22         ▷ ■ misc_feature       Misc. Feature       2590       TC: 43         ▷ ■ misc_feature       Misc. Feature       5245.5833       TG: 36         ▷ ◎ source (0, 1)       TT: 29       Help		CDC	1042 2659		
▷ ■ CDS       CDS       3875.4999       GC: 38         ▷ ■ CDS       CDS       5048.5203       GG: 37         ▷ ☑ comment (0, 1)       GT: 32       TA: 22         ▷ ■ misc_feature (0, 2)       TA: 22         ▷ ■ misc_feature       Misc. Feature       2590         ▷ ■ misc_feature       Misc. Feature       52455833         ▷ ■ source (0, 1)       TT: 29					
▷ ■ CDS       CDS       50485203       GC. 30         ▷ ☑ comment (0, 1)       GT. 32       GT. 32         □ misc_feature (0, 2)       TA: 22         □ misc_feature       Misc. Feature       2590         □ misc_feature       Misc. Feature       52455833         □ misc_feature       Misc. Feature       52455833         □ misc_feature       TT: 29					
▷ ② comment (0, 1)       GT: 32         ■ misc_feature (0, 2)       TA: 22         □ misc_feature       Misc. Feature       2590         □ misc_feature       Misc. Feature       52455833         □ misc_feature       Misc. Feature       52455833         □ misc_feature       TT: 29					
<ul> <li>▷ □ misc_feature</li> <li>▷ □ misc_feature</li> <li>▷ □ misc_feature</li> <li>▷ ☑ source (0, 1)</li> </ul>	▷ 🧐 comment (0, 1)				
▷ ■ misc_feature     Misc. Feature     52455833     TG: 36       ▷ ♥ ♥ source (0, 1)     TT: 29				<b>TA:</b> 22	
b Source (0, 1) Help	-			TC: 43	
Help		Misc. Feature	52455833		
	v 🕊 source (0, 1)			TT: 29	
					Help
	<		+		

To copy the statistical information about a sequence select it on the *Options Panel* and choose the copy item in the context menu, or use the Ctrl+C shortcut.

# **Manipulating Sequence**

- Show/Hide Sequence View Components
- Showing Sequence in Multiple Lines
- Translating Nucleotide Sequence
- Zooming Sequence
- Creating New Ruler
- Going To Position
- Selecting Sequence Region
- Copying and Pasting Sequence
- Editing Sequence
- Searching in Sequence
  - Load Patterns from File
  - Search Algorithm
  - Search in
  - Other Settings
- Annotations Settings
- Exporting Selected Sequence Region
- Exporting Sequence of Selected Annotations
- Exporting Sequence Image
- Locking and Synchronize Ranges of Several Sequences

#### • Multiple Sequence Opening

### **Show/Hide Sequence View Components**

To show or hide a Sequence View component (Sequence Overview, Zoom View, or Details View) use the following buttons on a sequence toolbar. The fourth button allows one to show/hide all the components at once.



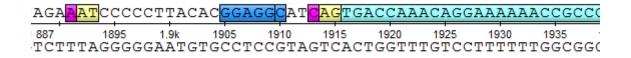
×

If several sequences are opened in the Sequence View, click the cross button

on a sequence toolbar to hide the sequence.

#### Showing Sequence in Multiple Lines

By default, a sequence in the Details View is shown in one line:



To show the sequence in multiple lines, click the Wrap sequence button on the left toolbar of the Details View:

GT	CTC	TGGI	cccc	GCCG	CAT	CCAT	ACCCC	CAGT	GTTT	ACCCI	CACAA	CGTT	CCAGT	AACC	GGGC	ATG
Cat		1760 ACCA	1765		, 70 (3 TT A	1775 GGT A	1780 PGGCG	1785 GTCA			95 1.8 AGTGTT			810 നന്ദര		1820 TAC
T						00111		01011				001111		1100	0000	
T	mme	7	mea	~	ccc	CIIIAI	COUCA	CCAR	remem	amaan	ттсат	CCC T		macc	CCCA	mca
0	-	1825	1830		335	1840	1845	1850						B75		1 885
											AAGTA				GGGT	ACT
	ACA	GA <mark>A</mark> A	TCCC	ссст	TAC	AC <mark>GG</mark>	AGGCA	T <mark>C</mark> AG	GACC	AAACA	AGGAAA	AAAC	ceccc	TTAA	CATG	GCC
		1890	1895		.9k	1905	1910	1915 A C III C I	1920				35 19 3CGGG	940	1945	1 950
	TGT	CTTI	AGG	JGGA.	ATG	TGCC.	recer	AGTCA	ICTGG	TTTGT	COTT	TTTG	30866	AATT	GTAC	666

### **Translating Nucleotide Sequence**

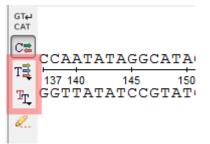
When a nucleotide sequence is opened in the Sequence View, the sequence and its complementary sequence are shown by default in the D etails View.

#### Showing/hiding the amino acid sequences

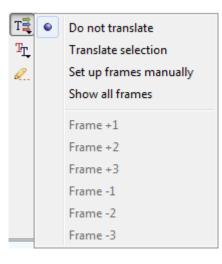
It is possible to translate them and show also the corresponding amino acid sequences:

Α D  $\mathbf{L}$ Q R Τ ь V Ν Y V s W C G Q v т A G Ν М s L С G :TGCAACGGGCAATATGTCTCTGTGTGGA 3 18 20 22 24 26 28 30 32 34 36 38 40 42 44 **ACGTTGCCCGTTATACAGAGACACACCT** V Ρ  $\mathbf{L}$ Ι D R Η Α P т Ρ С Υ Е T H Τ Q Ь C R Α I H R Q т S

The translation settings are available on the left toolbar of the Details View.

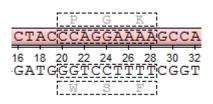


The Show/hide amino acid translations menu allows one to set up the mode of the amino acid sequences visualization:



The following options are available:

- Do not translate hide the amino acid sequences.
- Translate selection translate only a selected region of the sequence and the complementary sequence.



• Set up frames manually — select the reading frames to show. There are three frames for the sequence ("+1", "+2", "+3") and three frames for the complementary sequence ("-1", "-2", "-3"). Note that the complementary frame items are hidden in the menu, if the complementary sequence is hidden.

T		Do not translate	1		
T <sub>T</sub>		Translate selection			
0	۲	Set up frames manually			
		Show all frames			
		Frame +1	A	т	G
		Frame +2	rgca		
	$\checkmark$	Frame +3	18 20 ACGT		
		Frame -1	2 <b>L</b>	P	-
	$\checkmark$	Frame -2	С	R	A
	$\checkmark$	Frame -3			
	_				

• Show all frames — show all amino acid sequences.

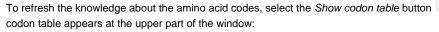
#### Setting the genetic code

The default value for the genetic code for a nucleotide sequence translation is read by UGENE from the sequence file when it is available. One can also set up the genetic code for the sequence using the Select genetic code menu:

T <sub>L</sub>		1. The Standard Genetic Code
0		2. The Vertebrate Mitochondrial Code
		3. The Yeast Mitochondrial Code
		4. The Mold, Protozoan, and Coelenterate Mitochondria and the Mycoplasma Code
		5. The Invertebrate Mitochondrial Code
		6. The Ciliate, Dasycladacean and Hexamita Nuclear Code
		9. The Echinoderm and Flatworm Mitochondrial Code
		10. The Euplotid Nuclear Code
	۲	11. The Bacterial and Plant Plastid Code
		12. The Alternative Yeast Nuclear Code
		13. The Ascidian Mitochondrial Code
		14. The Alternative Flatworm Mitochondrial Code
		15. Blepharisma Nuclear Code
		16. Chlorophycean Mitochondrial Code
		21. Trematode Mitochondrial Code
		22. Scenedesmus obliquus Mitochondrial Code
		23. Thraustochytrium Mitochondrial Code

All analysis routines (like HMMER, ORF finding, etc.) will use this code by default. (

#### Codon table





on the Sequence View global toolbar. The

1st base				2nd ba	ise				3rd base
TSC Dase		U		C		Α		STU Dase	
	UUU	Phenylalanine (Phe, F)	UCU		UAU	Tyrosine (Tyr, Y)	UGU	Cysteine (Cys, C)	U
U	UUC		UCC	Serine (Ser, S)	UAC	<u></u>	UGC	<u></u>	С
U U	UUA		UCA	<u>oenne (oer, oj</u>	UAA	Stop codon (*)	UGA	Stop codon (*)	A
	UUG		UCG		UAG		UGG	<u>Tryptophan (Trp, W)</u>	G
	CUU	Leucine (Leu, L)	CCU		CAU	Histidine (His, H)	CGU		U
с	CUC	<u>Leacine (Lea, L)</u>	CCC	Proline (Pro, P)	CAC	modulie (mo, m)	CGC	Arginine (Arg, R)	С
C	CUA		CCA	<u>Fronne (Fro, F)</u>	CAA	Glutamine (Gln, Q)	CGA		Α
	CUG		CCG		CAG	<u>Giutannine (Gin, Q)</u>	CGG		G
	AUU		ACU		AAU	Asparagine (Asn, N)	AGU	Serine (Ser, S)	U
	AUC	<u>Isoleucine (Ile, I)</u>	ACC	Threonine (Thr, T)	AAC	Asparagine (Ash, N)	AGC	<u>serine (ser, s)</u>	С
A	AUA		ACA	<u>inneonnie (inn, i)</u>	AAA	Lysine (Lys, K)	AGA	Arginine (Arg, R)	Α
	AUG	Methionine (Met, M)	ACG		AAG	Lysine (Lys, K)	AGG	Arginine (Arg, K)	G
	GUU		GCU		GAU	Aspartic acid (Asp, D)	GGU		U
	GUC	Valine (Val. V)	GCC	Alanine (Ala, A)	GAC	Aspartic acid (Asp, D)	GGC	Chusing (Chu, C)	С
G	GUA	<u>Valine (Val, V)</u>	GCA	Aldnine (Ald, A)	GAA	Clutamic acid (Clu. E)	GGA	<u>Glycine (Gly, G)</u>	Α
	GUG		GCG		GAG	Giutamic acid (Glu, E)			G

# **Zooming Sequence**

To zoom a sequence in the Sequence zoom view you can use one of the zoom button on the sequence toolbar:

€,	
Θ,	
€	
0;	

There are standard *Zoom In* and *Zoom Out* buttons. Additionally you can zoom to a selected region using the *Zoom to Selection* button. To restore the default view of the *Sequence zoom view* (when the sequence is not zoomed) use the *Zoom to Whole Sequence* button.

### **Creating New Ruler**

You can create any number of additional rulers by clicking the Ruler Create new ruler context menu item:

R G G C	Go to position Select sequence region New annotation Rename item	Ctrl+G Ctrl+A Ctrl+N	12 12	F E T N I F L K P T F * N Q R TTTTGAAACCAACG 45 1250 1255 AAAACTTTGGTTGC!
s	Copy/Paste Select	•	K	K F G V 1 K S V <b>L</b> S K Q F W R
	Add Analyze Align	, , ,		
	Cloning Export	+ +	F	
	Edit Remove	+		
	Rulers	•		Create new ruler
1	Statistics		<ul> <li>✓</li> </ul>	Show Main Ruler Show Custom Rulers

The following dialog will appear:

Create New Ruler										
Ruler name	New ruler									
Ruler start		99976 韋								
Ruler color	Sample Text									
Create	Cancel	Help								

The new ruler will be shown right above the default one:

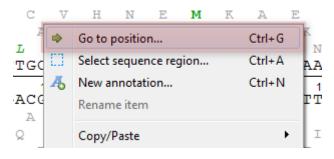
mat_peptide (7	·	mat pep	tide										
matnat_peptid	>							ma	t pept	ide			
Repeat ruler	1) -2.918	, -2000	, -1500	, -1000	, 500	. 0	, 500	, 1k	, 1.5	+ k _ 2k	, 2.5k	, 3k	, 3 71
500	1k	1.5k		2.5k	3k	3.5k	4k	4.5k	5k	5.5k		6.5k	7k 743
I L Y *	G F V F	Y I 7 T	Y P	New ri	iler wit	h a cus	tom offs	et S R	I S S	C L V	R S D J	V L F	L * S
I R	F GTTTT	L P	T ACCC	Q E AGGAA	K AAAGO	P I CAAC	. 14	L I CTCGA	) L ATCTO	L *	GATCI	C S IGTTO	

## **Going To Position**

To go to a position, use the global actions toolbar:



Or use the Go to position context menu or the Actions main menu item.



Also you can use the shortcut Ctrl-G.

### **Selecting Sequence Region**

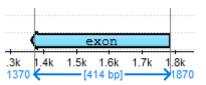
#### Selection in the Sequence View components

Selection of a sequence region is synchronized between different Sequence View components.

In the Sequence Overview the selection is displayed as a blue line:



In the *Zoom View* a blue line with coordinates and the length of the selected region is shown:



In the Details View a rectangle with a dashed border is shown around the selected region:

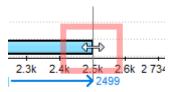
	GCAGA		GG
2505	2510	2515	25
TCGTT	CGTCT	TCGGG	CC

Finally, when one selects an annotation in the Annotations Editor, the corresponding region is also selected in the other components and visa versa.

#### Selecting a region using the mouse

It is possible to select a region of a sequence using the mouse in each component of the Sequence View.

In the Zoom View and the Details View the selection can be adjusted by dragging its left or right border:



#### Selecting a region using the dialog

If information about exact coordinates of a region is available, click the Select sequence region button on the toolbar or Select->Sequence region in the context menu of the Sequence View. Input the coordinates in the Region Selection dialog:

😣 Region Sele	ection		? ×
Single Range	ge Selection		
Region:	Min 1	- 199950	Max
🔘 Multiple Ra	nge Selection		
Multi Region	1199950		
	Go	Cancel	Help

It is possible to input:

- Single Range Selection a common region with a single range. Use the *Min* and *Max* buttons to automatically specify the beginning and the end of the sequence in the corresponding fields.
- *Multiple Range Selection* a region that consists of several join elements, for example, exons of a gene. See "The DDBJ/ENA/GenBank Feature Table Definition" for details.

#### Selecting a region around or between annotations

To select a region between two annotations, for example, an intron region between two exons:

• Click on the annotations, holding the Ctrl/Cmd key:

Ç				ex							$\subseteq$						xon						
1350 1370 <b>&lt;</b>	1.4k	1450	1.5k	1550	1.6k 4 bp]	1650	1.7k	1750	→ 1.8k	: 1850 3 1870	1.9k	1950	280	2050	2.1k	2150	2.2k 30 bp]—	2250	2.3k	2350	2.4k	2450	2.5k

• Click Select->Sequence between selected annotations in the context menu of the Sequence View.

To select a region that contains the annotations in a single range, choose the Sequence around selected annotations item in the menu.

#### **Copying and Pasting Sequence**

The selected sequence region, an annotation sequence or their amino translations can be copied to clipboard:

• By pressing the corresponding buttons in the global toolbar.



- Using the following shortcuts:
  - Ctrl-C copies direct sequence strand
  - Ctrl-T copies direct amino translation
  - Ctrl-Shift-C copies reverse-complement sequence
  - Ctrl-Shift-T copies reverse-complement amino translation
- Using the Copy submenu of the context menu:

W	 Q	A S L S N	A T	÷	Q G M E K Y I	I T E N I		
TGG	•	Go to position	Ctrl+G	TT	GCAAGGCATGGAAAAATACA	1 1		
25 ACC		Select sequence region	Ctrl+A	AA	55 60 65 70 CGTTCCGTACCTTTTTATGT2	75 80 E		
Н	4	New annotation	Ctrl+N	ĸ	C P M S F Y M	V S F L		
A P		Rename item			A L C P F I C L A H F F V Y	L Q S Y Y S L I ;		
	· 📃	Copy/Paste	•		Copy sequence	Ctrl+C		
		Select	•		Copy reverse-complement sequence	Ctrl+Shift+C		
		Add	+		Copy translation	Ctrl+T		
		Analyze	•	Þ	Copy reverse-complement translation	Ctrl+Shift+T		
		Align	•	₿ <u>a</u>	Copy annotation sequence			
		Cloning	•	Ŧ <u>A</u>	Copy annotation sequence translation			
		Export	+	È	Paste sequence	Ctrl+V		
		Edit	+					
		Remove	+	L				
		Rulers	•					

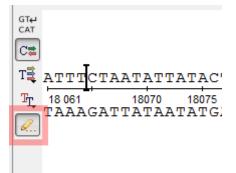
To paste sequence use the corresponding context menu item or Ctrl+V shortcut.

### **Editing Sequence**

If the corresponding document is not locked, it is possible to edit a sequence.

#### Editing mode and annotations settings

To switch on the editing mode for the sequence, select the Edit sequence button on the left toolbar of the Details View:



A special cursor appears in this case, and the sequence can be edited as in a text editor:

- To insert characters to the sequence, move the cursor to the required location and type the characters or paste them using Ctrl + V/ Cmd + V keyboard shortcut.
- To replace a sequence region, select it and insert the new characters.
- To remove a character from the sequence, move the cursor to the required location and press *Backspace* or *Delete*. To remove a region, select it and press one of these shortcuts.

One can also configure the way how annotations located in an edited region should be modified. To open a dialog with the settings select *Edi t*->*Annotations settings on sequence editing* in the *Actions* main menu or in the context menu of the *Sequence View*.

8 Annotation Settings on Sequence Editing	x
<ul> <li>Expand or crop affected annotation</li> </ul>	
Remove affected annotation	
Split (join annotation parts)	
Split (separate annotations parts)	
Recalculate values of qualifiers	
OK Cancel Hel	p

Select one of the following:

- Expand or crop affected annotation an annotation located in an edited region is expanded in case of characters insertion or cropped in case of characters deletion.
- Remove affected annotation all annotations in an edited region are removed.
- Split (join annotation parts) an annotation is split into two join elements (see "The DDBJ/ENA/GenBank Feature Table Definition" for details).
- Split (separate annotation parts) an annotation is split into two annotations.

It is also possible to check the *Recalculate values of qualifiers* option in the dialog. If it is selected, qualifiers of all annotations are parsed on the sequence editing. Qualifiers values that specify coordinates (like "100..200") are re-calculated accordingly to the sequence modification. For example, the value might become "104..204", if four characters have been inserted before the corresponding annotation.

#### Getting reverse-complement, reverse, or complement sequence

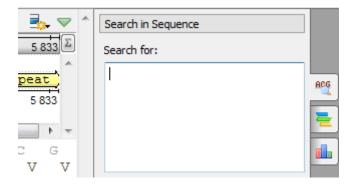
To replace a sequence with its reverse-complement sequence, select Edit->*Reverse-complement sequence* in the *Actions* main menu or in the context menu of the Sequence View. The keyboard shortcut for this action is Ctrl + Shift + R / Cmd + Shift + R.

Use the *Reverse sequence* or *Complement sequence* items in the same menus to replace the sequence with its reverse or complement sequence.

### Searching in Sequence

To search for a pattern(s) in a sequence go to the Search in Sequence tab of the the Options Panel in the Sequence View.

Input the value you want to search in the text field and click the Search button. To search multiple patterns input the patterns separated by a new line in the pattern text field. To add a new line symbol *Ctrl+Enter* may be used. You can input the value as sequence or name of the sequence in the FASTA format and sequence after that.



By default, **misc\_feature** annotations are created for regions that exactly match the pattern. Find below the description of the available settings.

- Load Patterns from File
- Search Algorithm
- Search in
- Other SettingsAnnotations Settings

Load Patterns from File

Load patterns from file								
Path:								

Use this checkbox to load patterns from file. When this option is active the *Search for* field is disabled. **Search Algorithm** 

<ul> <li>Search algorithm</li> </ul>				
Algorithm	Exact 🔹			

This group specifies the algorithm that should be used to search for a pattern. The algorithm can be one of the following:

- InsDel there could be insertions and/or deletions, i.e. a pattern and the searched region can vary in their length. You can specify
  the percentage of the pattern and a searched region match in the field nearby. Note that this value also depends on the pattern
  length and is disabled when the pattern hasn't been specified.
- Substitute a pattern may contain characters different from the characters in the searched region. When this algorithm has been selected you can also specify the match percentage and additionally it is possible to take into account ambiguous bases.
- Regular expression a regular expression may be specified instead of a pattern. For example character '.' matches any character, '.' matches zero or more of any characters. There is also the *Limit result length* option that specifies the maximum length of a result.
- *Exact* find a place where one or several patterns are found within a larger pattern.

Search in

<ul> <li>Search in</li> </ul>	
Strand	Both 🔻
Search in	Sequence 💌
Region	Whole sequence 🔻

In this group you can specify where to search for a pattern: in what region and in which strand (for nucleotide sequences). Also for nucleotide sequences it is possible to search for a pattern on the sequence translations.

Strand - for nucleotide sequences only. Specifies on which strand to search for a pattern: Direct, Reverse-complementary or Both strands.

Search in — for nucleotide sequences you can select the *Translation* value for this option. In this case the input pattern will be searched in the amino acid translations.

Region — specifies the sequence range where to search for a pattern. You can search in the whole sequence, specify a custom region or search in the selected region.

#### **Other Settings**

▼ Other settings	
Remove overlapped results	
Limit results number to:	
100000	×

This group contains additional common settings:

Remove overlapped results - annotates only one of the overlapped results.

*Limit results number to* — limits number of the searched results to the specified value. **Annotations Settings** 

<ul> <li>Save annotation(s) to</li> </ul>
Existing table:
🕑 NC_014267 features [ 👻 🌇
Create new table:
<ul> <li>Annotation parameters</li> </ul>
Group name:
<auto></auto>
Annotation type:
Misc. Feature 🔻
Annotation name:
by type
Description:
Use pattern name

In the Save annotation(s) to group you can set up a file to store annotations. It could be either an existing annotation table object or a new annotation table.

In the Annotation parameters group you can specify the name of the group and the name of the annotation. If the group name is set to <auto> UGENE will use the group name as the name for the group. You can use the '/' characters in this field as a group name separator to create subgroups. If the annotation name is set to by type UGENE will use the annotation type from the Annotation type: table as the name for the annotation. Also you can add a description in the corresponding text field. To use a pattern name for the annotations check the corresponding checkbox.

After that click the *Create annotations* button. The annotations will be created. Also you can see the result statistic and navigation under the *Search for:* field:

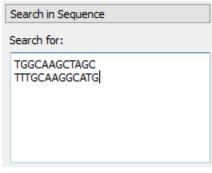
Results: 1/1			
Previous	Next		

#### Searching for one or several patterns and names of the result annotations

If you search for one pattern only, than input the required name into the Annotation name field and leave the Use pattern name check box unchecked.

You can also search for several patterns at a time by:

• Inputting several patterns into the search field (click <Ctrl> + <Enter> keys to insert to a new line):



• Inputting several patterns into the search filed in FASTA format:

Search in S	Search in Sequence			
Search for	:			
>pattern: TGGCAAG				
>pattern2 TTTGCAA0				

Loading patterns from a FASTA file

4

Even when you search for several patterns, names of the found annotations will be identical by default (the name is specified in the Annotatio n name field).

If you want to assign different names to annotations found for different patterns, than you should:

- Input the patterns in FASTA format (the latter two cases above)
- Check the Use pattern name checkbox in the Annotation parameters group

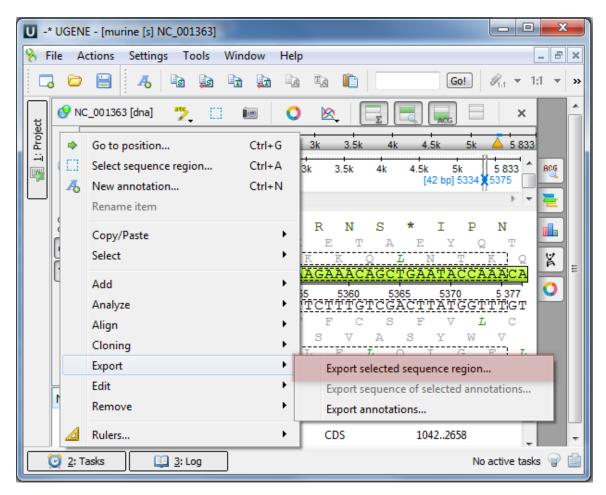
Here is an example of the found annotations in the Annotations Editor:

đ	Annotations [MyDocument_1.gb] *				
	4 🧭	misc_feature (0, 4)			
	$\triangleright$	pattern1	2435		
	$\triangleright$	pattern1	52675278		
	$\triangleright$	pattern2	5062		
	$\triangleright$	pattern2	52935305		
	-				

### **Exporting Selected Sequence Region**

Open a sequence object in the Sequence View and select a region by pressing and moving the left mouse button over the sequence.

Use the Export Export selected sequence region context menu item to save selection into a file of a sequence format.



The Export Selected Sequence Region dialog will appear which is similar to the Export Selected Sequences dialog described here.

## **Exporting Sequence of Selected Annotations**

Open the Sequence View with document that contains annotations. A good candidate here could be any file in Genbank format with both sequence and annotations. Select a single or several annotations or annotation groups in the Annotation editor, click the right mouse button to open the context menu and select the Export Export sequence of selected annotations item:

_							
U -* UG	ENE - [n		Find qualifier				
🗞 File	Action		Invert annotation	selection			_ <i>B</i> ×
	> 🗎		Rename item		F2		Go! 🖉 - 1:1 - 🕫 SW 🐚 🟹 »
		Ē.	Paste annotations		Ctrl+Shift+V	$\downarrow$	
	NC_001				~	- 2	
1: Project	4	*	Go to position		Ctrl+G		4k 4.5k 5k 5.833
			Select sequence re	gion	Ctrl+A		4K 4.0K 0K 0000
- Ref	, 1	<b>4</b>	New annotation		Ctrl+N		4k 4.5k 5k 5 833
¥ L	•		Copy/Paste			•	F 🔫 🜉
GT+			Select			۰ <u>-</u>	
CAT	<b>*</b>		Add				
C	D						D С Ү Н S L К 🗶
T	GAC		Analyze			AG	ACTGTTACCACTCCCTTAAG
Tr	_ HHH		Align			105	
<u>т</u>	CTG		Cloning				TGACĂĂTGGTGAGGGĂATTC
	s		-			S	Q * W E R L
	S			rom remote database		·	
	v		Export			•	Export selected sequence region
	•		Edit			•	Export sequence of selected annotations
Nar	me		Remove			•	Export annotations
	4 🧭		Rulers			• [	
			os	CDS		104226	558
	Þ		DS	CDS		join(297	703413,34123873)
	⊳		DS .	CDS		387549	999
	$\triangleright$		DS	CDS		504852	203
	Þ 🧭	com	ment (0, 1)				
	4 🧰	misc	feature (0.2)				•
2	<u>2</u> : Tasks		🛄 <u>3</u> : Log				No active tasks 💡 🛗

The Export Sequence of Selected Annotations dialog will appear which is similar to the Export Selected Sequences dialog described here.

## **Exporting Sequence Image**

Use a sequence toolbar Export image button to save a screenshot of the sequence:



The Export Image dialog will appear where you should set name, location, export settings and format of the picture:

Export Image	8
Export settings Area Currently viewed Zoomed annotations Sequence details	egion Selected region 🗘 31 - 75
Export to file	
File name	
Format png 🗘 Quality	y 80 🗘
Help	Cancel Export

UGENE supports export to the BMP, JPEG, JPG, PNG, PPM, TIF, TIFF, XBM, XPM and SVG image formats. You can export currently viewed, zoomed annotations or sequence detailes areas. Also you can export whole sequence or custom region. Use the *Region* settings to

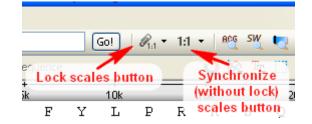
do it.

## Locking and Synchronize Ranges of Several Sequences

An important feature of the Sequence zoom view is the ability to synchronize and lock visual ranges of different sequences shown in the Seq uence View.

This feature is available when there are two or more sequences opened in the same Sequence View.

If we click the *Lock scales* button the second sequence scale will be adjusted to be the same as the focused sequence scale and is locked. Now if we move a scrollbar or use zoom buttons for any of the sequence, visual ranges for the rest sequences will also be adjusted.



To unlock the scales click the same button again.

You may use the Adjust scales button to synchronize scales without locking them.

Note, that if you have a selected sequence region or a selected annotation the scales will be synchronized by the start position of the region or the annotation. If there are no active selection the regions are synchronized by the first visible sequence position on the screen.

### **Multiple Sequence Opening**

To open several sequences use the *File->Open* menu item or *Open* toolbar button and using *Ctrl* select the several sequences and click the *Open* button. The following dialog will appear:

U	Multiple Sequence Reading Mode	9	×
	Separate sequence mode		
	Merge sequence mode		
	Number of unknown symbols ('N 'for nucleic acid or 'X' for amino acid) between sequences $\fbox{10}$		* *
	O Join sequences into alignment		
	New document name		
	✓ Save document		
	1. NC_014267.1.gb		
	2. PBR322.gb		
	3. sars.gb		$\mathbf{V}$
	OK Cancel	He	elp

The following parameters are available:

Separate sequence mode - opens the sequences as separate sequences.

Merge sequence mode - merges sequences into one sequence with selected number of unknown symbols between sequences.

Join sequences into alignment - joins sequences into alignment.

Save document - save document to the selected document.

Also you can change the order of the sequences by up and down arrows.

Choose the parameters and click the Open button. Sequences will be opened in the selected mode.

In the Separate sequence mode sequences will be opened as separate sequences in selected order. You can change the sequences order

by drag and drop in the sequence view.

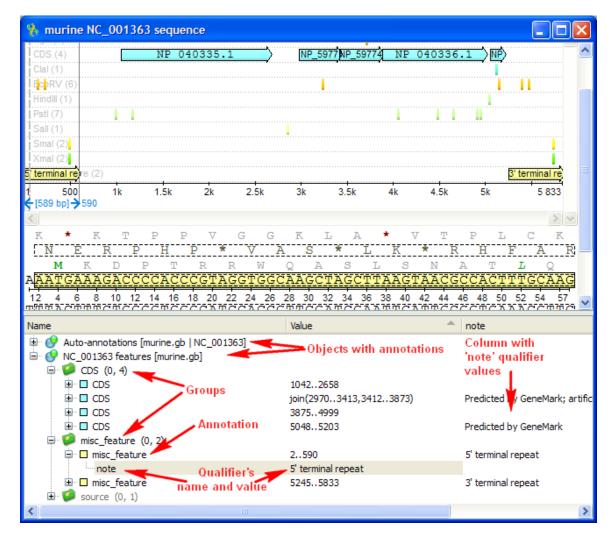
# **Annotations Editor**

The Annotations editor contains tools to manipulate annotations for a sequence. It provides a convenient way to organize, view and modify a single annotation as well as annotation groups.

An annotation for a sequence consists of:

- Name (or key) indicates the biological nature of the annotated feature.
- Location coordinates in the sequence.
- The list of qualifiers qualifiers are the general mechanism for supplying information about annotation. Qualifiers are stored as pairs of (name, value) strings.

Below is the default layout of the Annotations editor with an extra column for the "note" qualifier added:



There are usually several objects with annotations in the Annotations editor. A special Auto-annotations object is always presented for each sequence opened. It contains annotations automatically calculated for the sequence (see *below* for details).

An object contains groups of annotations used by UGENE for logical organization of the annotations. An annotation must always belongs to some group.

For documents created not by UGENE annotations are grouped by their names. For annotations created in UGENE it is possible to use arbitrary group names.

Groups can contain both annotations and other groups. The numbers in the brackets after a group name in the Annotations editor are the count of subgroups and annotations in the current group.

A single annotation is allowed to be presented in several groups simultaneously. An annotation is physically removed from the document when it does not belong to any group.

"db\_xref" Qualifier

- Automatic Annotations Highlighting
- The "comment" Annotation

### "db\_xref" Qualifier

۵

Some files in Genbank format contain the *db\_xref* qualifier. A value of this qualifier is a reference to a database.

	04718 features [sars.gb] 'UTR (0, 1)			
Þ 🧭 5	'UTR (0, 1)			
4 🧭 C	DS (0, 14)			
4	CDS	CDS	join(26513398,13	39821485)
	codon_start		1	
	db_xref		GI:30124074	
	db_xref		GeneID:1489680	
	gene		orflab	
	locus_tag		sars1	

When you click on the value a web page is opened or a file is loaded specified in the reference. The loaded file is added to the current project

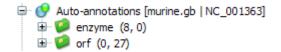
## **Automatic Annotations Highlighting**

Enabling the automatic annotations highlighting allows you to automatically calculate and highlight annotations on each nucleotide sequence opened.

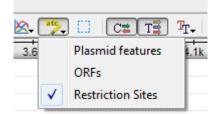
Currently, the following annotations types support the automatic highlighting:

- Open reading frames
- Restriction sites
- Plasmid features

The corresponding groups of annotations found are stored in the Auto-annotations object in the Annotations editor, for example:



To disable/enable the automatic annotations calculations use the Automatic Annotations Highlighting menu button on the Sequence View too lbar:



To create a permanent annotation click on the *Make auto-annotations persistent* context menu item and choose the annotation parameters in the *Create Permanent Annotation* dialog.

### The "comment" Annotation

General information about a file in GenBank or Vector NTI Sequence format, stored in the COMMENT sections of the file, is shown in UGENE in a special *comment* annotation in the *Annotations Editor*.

The information, for example, may include the name of the file author, creation date and last modification date for the file, and so on:



This file is created by Vector NTI http://www.invitrogen.com/ Demo User Thu Dec 23 13:29:41 2010 Thu Dec 23 13:29:41 2010 2xCyPET-3xFLAG-6xHIS/pDONRP2R-P3 (p3')

# **Manipulating Annotations**

- Creating AnnotationSelecting Annotations
- ٠ Editing Annotation ٠
  - Highlighting Annotations
    - Annotations Color
      - Annotations Visability
      - Show on Translation
      - Captions on Annotations
- Creating and Editing Qualifier
- ٠ Adding Column for Qualifier .
- Copying Qualifier Text •
- Finding Qualifier
- Deleting Annotations and Qualifiers ٠ Importing Annotations from CSV
- Exporting Annotations

## **Creating Annotation**

To create a new annotation for the active sequence press the Ctrl-N key sequence, select the New annotation toolbar button or use the Add New annotation or New annotation context menu item:

U -	U -* UGENE - [sars [s] NC_004718]						
8	😽 File Actions Settings Tools Window Help 🛛 🗕 🗗 🗙						
	3 🗁 🖹 🥵 🖪	ľ	10 la la ll	Go! 07:1	T:1      ■      R     R     S		
ين ا	OC_004718 [dna]		2 🖸 🔟	0 🖄			
1: Project	Σ 1 2k 4k 6k	8	k 10k 12k 📥 14k 16k	18k 20k	22k 24k 26k 29 751		
1	🔍 1 2k 4k 6k	•	Go to position	Ctrl+G	22k 24k 26k 29 751 A RC		
	* -	С <b>Д</b>	Select sequence region New annotation	Ctrl+A Ctrl+N			
	CAT G M W K G		Rename item		PLMQSA		
			Copy/Paste	•	N P * C S L T L D A V C ¥		
	TE GGAATGTGGAAAGG		Select	•	ACCCTTGATGCAGTCTGC		
	T <sub>13 315 13320 13325</sub> CCTTACACCTTTCC		Add	+	A New annotation Ctrl+N		
	SHPFT		Analyze	•	Objects with annotations		
	F T S L		Align	•	Qualifier Ins		
			Cloning	•			
	Name		Fetch sequences from remote database	*	A		
	🚱 Auto-annotations [sars.gb		Export Edit		Ξ		
	<ul> <li>MC_004718 features [sars.g</li> <li>\$ 2'UTR (0, 1)</li> </ul>		Remove	•			
	5'UTR (0, 1) CDS (0, 14)		Rulers	۲			
	CDS		Disable 'CDS' highlighting		.485)		
	codon_start 1 db_xref GP-30124074						
	2: Tasks <u>1 3: Log</u>	]			No active tasks 💡 🖆		

This will activate a dialog where to set up annotation parameters:

% Create Annotation				? ×
Annotation type:		Group name	source	*
Glycosylation Site Homeodomain iDNA Insertion Intron J-Region J-Segment Leucine Zipper Domain Loci LTR Mature Peptide Misc. Binding Site Misc. Difference Misc. Feature	Ĩ	Annotation name Description Location      Simple form      Complement      GenBank/E	-	
<ul> <li>Existing table</li> <li>Create new table</li> </ul>	NC_001363 features	s [murine.gb]		• 💽
<ul> <li>Use auto-annotations table</li> </ul>		(	Create Cancel	Help

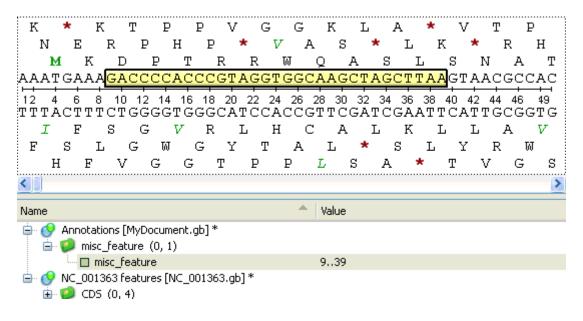
The dialog asks where to save the annotation. It could be either an existing annotation table object, a new annotation table or auto-annotations table (if it is available).

You can also specify the name of the group and the name of the annotation. If the group name is set to <auto> UGENE will use the group name as the name for the group. You can use the '/' characters in this field as a group name separator to create subgroups. If the annotation name is set to *by type* UGENE will use the annotation type from the *Annotation type:* table as the name for the annotation. Also you can add a description in the corresponding text field.

The *Location* field contains annotation coordinates. The coordinates must be provided in the Genbank or EMBL file formats. If you want to annotate complement strand sequence check the corresponding checkbox for the simple format or surround the coordinates with the "complement()" word or press the last button in the corresponding row to do it automatically.

Note, that by default the Location field contains the coordinates of the selected sequence region.

Once the *Create* button is pressed the annotation is created and highlighted both in the *Sequence overview* and the *Sequence details view* a reas:



### **Selecting Annotations**

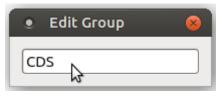
To select one annotation click on it. To select several annotations hold *Ctrl* key while clicking on the annotations. To invert the selection use the *Invert annotation selection* item in the *Annotations editor* context menu.

## **Editing Annotation**

If the document is not locked, it is possible to edit an annotation or an annotation group using the *Rename item* context menu from the *Annot ation Editor* or from the *Sequence View* or with a help F2 key in the *Annotation Editor*. The result of pressing for an annotation:

• Edit Annotation	8
Annotation type:	Annotation name CDS
-10 Signal -35 Signal 3' Clip	E Description
3' UTR 5' Clip	Location
5' UTR Allele	Simple format
Assembly gap Attenuator	1042 - 2658
k bHLH Domain C-Region	Complement
CAAT Signal CDS Cellular	O GenBank/EMBL format
Centromere Conflict	10422658
Help	Cancel Edit

The result of pressing for an annotation group:



## **Highlighting Annotations**

To configure settings of annotation names go to the Annotation Highlighting tab in the Options Panel.

By default the tab shows annotations names of the opened Sequence View.

🔮 NC_00136 🜔 🖄 🥍 🎵 🖸 🕼 🕄 🔀 😵	Annale Kana Litela Kan
	Annotations Highlighting
1 500 1k 1.5k 2k 2.5k 3k 3.5k 🚣 4k 4.5k 5k 5833 🗵	Select an annotation name:
5' termin ature (2)	Annotation Color
1 500 1k 1.5k 2k 2.5k 3k 3.5k 4k 4.5k 5k 5833 3875 (1125 bp) -> 4999	BadI
<	BaeGI
DHPLD*HGAFNAM	
	BaeI 💼
G P S S R L T W R I Q R H	CDS III
GACCATCCTCTAGACTGACATGGCGCATTCAACGCCATG	SacII 📕
3 856 3860 3865 3870 3875 3880 3885 3890 3 894	comment 📃 💽
CTGGTAGGAGATCTGACTGTACCGCGTAAGTTGCCGGTAC	misc_feature
G D E L S V H R M * R W A	source
VMR * VSMACEVGH	Show all annotation names
SWGRSQCPAN <b>L</b> AM	show an annotation maries
	Configure the annotations:
Name Type ^	Show annotations
D CDS CDS	
CDS CDS	Show on translation
CDS CDS	Show value of qualifier:
DE CDS CDS	label,note
V V comment (0, 1)	
4 🤪 misc_feature (0, 2)	
Imisc_feature Misc	
▶	
> Source (0, 1)	Help

If you want to see all annotation names, click the Show all annotation names link. The Previous annotation and Next annotation buttons seek to the previous or to the next annotation of the view correspondingly.

Find below information about annotations names' properties that you can configure.

- Annotations Color
- Annotations Visability
- Show on Translation
- Captions on Annotations

#### **Annotations Color**

To change a color of all annotations of a certain type click on the corresponding color box in the annotations types table and select the required color in the appeared *Select Color* dialog.

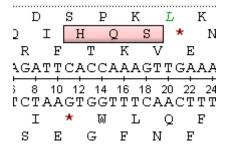
#### **Annotations Visability**

To show/hide annotations with a certain name, select this name in the annotations names table and check/uncheck the *Show annotations* ch eck box below. Another way to show/hide the annotations is to select the *Enable/Disable highlighting* item in the context menu of an annotation.

<b>U</b> -	* UGENE - [sars	[s] N	C_004718]				
8 F	File Actions	Setti	ngs Tools Window Help				_ & ×
	. 🗢 🗎	<b>"</b>	Find qualifier		Go!	Ø <sub>1:1</sub> ▼ 1:1 ▼ 800 S	🦉 💐 🛃 »
	🔮 NC_004718	Ť	Invert annotation selection		>>	Annotations Highlighting	
oject	Σ		Rename item	F2		Select an annotation name:	
🛃 <u>1</u> : Project		Ē.	Paste annotations	Ctrl+Shift+V	29 751	Annotation Colo	
	▲	⇒	Go to position	Ctrl+G	751	3'UTR	ACG
	* •		Select sequence region	Ctrl+A	• •	5'UTR	
		А	New annotation	Ctrl+N	v	CDS 📃	
		Ŭ	C	•	C	comment	
	G C		Copy/Paste Select		A	gene	X
	T≣ GATGO T <sub>T</sub> 13 378		Select	,	12 417	mat_peptide	0
	CTAC		Add	•	ACG	misc_feature	
	IC		Analyze	•	H	Show all annotation names	
	S A H		Align	•	A	Snow all annotation names	
			Cloning	•	-	Configure the annotations:	
	Name		Fetch sequences from remote database	•	-	Show annotations	
	Auto-a		Export Edit		=	Show on translation Show value of qualifier:	
	4 🚱 NC_00		Remove			label	
	> 🧭 3'U ⊿ 🧐 5'U	4					
			Rulers	•		<b>4</b>	<b></b>
	4 🧭 CD		Disable 'CDS' highlighting				
	4 🔳 🤅		CDS	join(265.	.13398		
	•	codo	on_start 	1	*		Help
	过 <u>2</u> : Tasks		🛄 <u>3</u> : Log				No active tasks 💡 📋

#### **Show on Translation**

This option is available for nucleotide sequences only. It specifies to show the annotation on the corresponding amino sequence instead of the original nucleotide sequence in the Sequence Detailed View, for example:



You can enable/disable this option by checking/unchecking the Show on translation checkbox.

#### **Captions on Annotations**

It is possible to show a value of a qualifier of an annotation instead of the annotation type name in the Sequence Zoom View. To enable this option for an annotation type check the Show value of qualifier check box and input the values of the required qualifiers in the text field nearby this check box. See the image below.

🕑 N	C_001	363 [d	0	🖄 🏷	. []]	C	T	$T_{T_{\bullet}}$		æ,	۹	<b>.</b>	»	Annot	ations Highlighting		
1 :	500	1k	1.5k	2k	2.5k	3k	3.5k	4		1.5k	5k	<u>/</u> 5 833	Σ	Select	an annotation name	e:	
1 0.15	5.01	5050	5 41	Show							5750			Ann	otation	Color	ACG
5 245	5.3K	5350	5.4k	5450	5.5k	5550 589 bp]·	5.6k	565	50 :	5.7k	5750	5 833		Bad	I		<u> </u>
•			/	L								•	-	Bae	GI		- =
A	R		D	I	Ċ	GI	x Q	F		L	P	R L		Bael	I		
G	Р	N	R I	t S	V	V	S	S	S	С	Р	G S		CDS			<u> </u>
G	Q	Т	G	Y I	W	*	A	V	Р	A	P	A	R	SacI	т		X
<u>GGC</u> 5 445	CAA	ACA	GGA.	LATCI	GTG	GTA	IGCA	GTT	CC.	rgc	100	GGCTC	9		ment		
			5455 CCT2	5460		465 CATT	5470 CGT		5475 \GGI		480 GGG	5 48 CCGAG		_	_feature		0
Р	W	V	P 3	Z R	Н	Y	A	т	G	A	G	A R	5	sou	-		
G		L	I	D	т	T I	i L			2	_	P E			/ all annotation name		
A	L	С	S	IQ	9 P	L	С	N	R	G	R	S	Р	SHOW	all annotation name	es	
•													•	Config	ure the annotations	s: /	
Nam	e				Туре		Val	ue							Show annotations		
4 🤇	NC	001363	featur	es [m													
¢	· 🧭	source	(0, 1)								1		Ξ		Show on translation		
4	• 🧭	misc_fe	eature	(0, 2)										V 5	Show value of qualifi	ier:	
	4	🗖 miso	_featu	re	Misc.	Feature	524	5583	3					not	e		
		not	e				Sho	ow val	lue of	qualif	fier						
	-	misc	-		Misc.	Feature	25	90							1		
1	2	comm		1)										-	J	-	
4	1	CDS (0	), 4)														
	$\triangleright$	CDS			CDS			8520								Help	
	⊿	CDS			CDS		387	5499	19				Ŧ				

If you input several qualifiers names (separated by comma), then the first found qualifier is taken into account and shown on the annotation.

## **Creating and Editing Qualifier**

To add a qualifier to an annotation select it in one of the Sequence View subviews and press the Insert key, or use the Add Qualifier context menu or the Actions main menu item.

4 🧭 5'UTR	Select	+			
▲ □ 5'UT evic	Add	•	А	New annotation	Ctrl+N
⊿ 🧭 CDS (0	Analyze	•		Objects with annotations	
CDS	Align	•		Qualifier	Ins
cod	Cloning	- • T			
db_:	Export	•		24074	
db_:	Edit	•		):1489680	
gen			1ab		
Ιοςι	Remove	•	rs1		
	Rulers	•	-	assumed that the SARS orf1ab polyprotein (pp1ab)	polyprotein
protein_i	d	N	P_828	3849.2	

The dialog will appear:

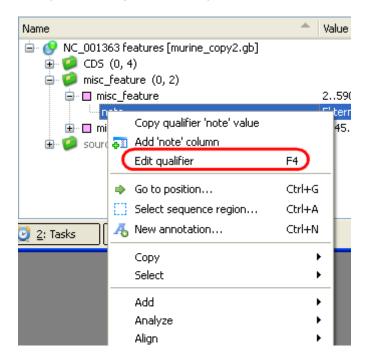
🗞 Add	d New Qualifier	
Name Value	new_qualifier	
	OK Cancel Help	

Here you can specify the name and the value of the qualifier.

You can use the F2 key to rename a qualifier:

9	👌 Rename Qualifier	8	×
	new_qualifier		

To edit a qualifier, select the qualifier and press the F4 key or use the Edit qualifier context menu item:



## Adding Column for Qualifier

It is possible to add a column with the qualifier values to the Annotations editor. To add the column, select the Add '[the qualifier name]' column qualifier context menu item.

### **Copying Qualifier Text**

Use the Copy qualifier '[the qualifier name]' text qualifier context menu item to copy the qualifier value.

#### **Finding Qualifier**

To find a qualifier select annotation(s) or group(s) of annotations and use the Find qualifier context menu.

Name	Value
<ul> <li>Auto-annotations [CVU55762.gb   CVU55762]</li> <li>CVU55762 features [CVU55762.gb]</li> </ul>	
<ul> <li>CDS (0, 2)</li> <li>CDS codon_start db_xref</li> </ul>	<ul> <li>Find qualifier</li> <li>Invert annotation selection</li> <li>Rename item</li> </ul>
gene product protein_id transl_table translation	Image: Go to position     Ctrl+G       Image: Go to position     Ctrl+A       Image: Go to position     Ctrl+A       Image: Go to position     Ctrl+N
▷ □ CDS	Сору

The dialog will appear:

8 Find Qualifier	? ×
Search in: CDS	
Qualifier	
Name:	
Value:	
Match Parameter	
Exact match	
<ul> <li>Contains substring</li> </ul>	
Select all Next Cancel	Help

Here you can specify the name and the value of the qualifier and select the searching parameter: Exact match or Contains substring.

## **Deleting Annotations and Qualifiers**

Selected annotations, groups and qualifiers can be deleted using the Delete key.

To remove an annotation object from the active view, select the object in the *Annotations editor* and press the Shift-Delete. Note that the object will not be removed from the project, but just from the active *Sequence View*. To add object again just drag and drop it to the *Sequence View*.

## Importing Annotations from CSV

It is possible to import annotations for a sequence from an annotations table stored in the CSV format.

To import annotations from a CSV file, right-click on a *Project View* and select *Import Import annotations from CSV*. The following dialog box will appear:

File to read	
Results	
Result file	
File format Genbank	•
☑ Add result file to project	
Column separator value: [,], hex: [2c], length: 1	
File parsing	
Olumn separator	, Guess
◎ Script	Edit
First lines to skip	Do not skip 💻
Skip all lines starts with the text	#
Interpret multiple separators like a single separator (try when see	eparator is a whitespace character)
Remove quotes	
Default annotation name	misc_feature
Results preview	Preview
	Preview
	Freview
	Freview
	Freview
	Preview
	Preview
Raw file preview:	

Basically you need to specify the file to read annotations table from (required):

File to read	D: /projects/dev/ugene/trunk/test/	common data/scenarios/annotations im	port/apps1.csv	ſ
riie to reau	Dr/projecta/dev/dgene/d dnk/teau	_common_data/scenanos/annotations_im	por c/arms1.csv	

And the format of and the path to the file to write the annotations table into (required):

Results		
Result file	D:/projects/dev/ugene/trunk/test/_common_data/scenarios/annotations_import/result.gb	J
File format	Genbank 🗸	

Check Add result file to project to link the annotations to the currently opened sequence.

To use a separator to split the table, check the *Column separator* item and specify the separator symbols. Also you can press *Guess* to try to detect the separator from the input file.

Hie parsing		
<ul> <li>Column separator</li> </ul>	, <b>(</b>	Guess

Alternatively, you can press *Edit* and edit the script which will specify the separator for each parsed line. It is possible to use line number in the script.

Script	U Script Editor	x	Edit
First lines to	Used script		Do not skip 🚖
Skip all lines :	Script text		#
Interpret	//The script parses input line		
Remove	// and returns an array of parsed elements as the result var line; //input line		
Default annota	var lineNum; //parsed line number		misc_feature
Results preview			Preview
	<pre>var firstColumn = [lineNum]; var otherColumns = line.split(" "); result =firstColumn.concat(otherColumns);</pre>		
	Check syntax Save Line: 1		
Daw file provide	Clear Save as		
Raw file previe		lelp	

Using the arrows, you exclude the necessary number of lines at the beginning of the document from parsing. You can also skip all lines that start with the specified text.

First lines to skip	1 line(s) 🤹

By pressing *Preview* one can bring up the view of the current annotations table (which is produced from the input file with the specified parameters values). The input file contents will also be shown at the bottom part of the dialog.

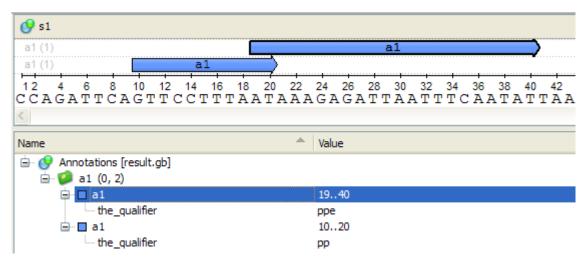
Resul	ts preview			
	[ignored]	[ignored]	[ignored]	[ignored]
a1		10	20	pp
a1		19	40	ppe
		1		
Raw f	file preview:			
	e,start,end,qua	əl 1		
	.0,20,pp .9,40,ppe			

The preview table headline indicates the types of the information contained in the corresponding columns. By default the values are *[ignored]*. To specify a column role, click on the corresponding headline element:

U Select the Role	of the Column	n 8 2	٢
Column role			
Annotation s	tart position	Add offset 0 bp 🛓	]
Annotation e	end position	✓ Indusive	
Annotation I	ength		
Complement	strand mark	Mark value	]
Annotation r	name		-
Annotation g	jroup		
Qualifier			]
Ignore this c	olumn		
	(	OK Cancel Help	

The annotation start and end positions must be specified. It is possible to add an offset to every read start position by checking the Add offset checkbox, and to shorten annotations by one from the end by uncheking the *Inclusive* checkbox.

When all the roles are specified, press *Run*. With the *Add to project* checkbox specified and a *Sequence View* opened, on success you will see the *Sequence View* with annotations linked:



### **Exporting Annotations**

To export annotation(s) open the Sequence View with a document that contains annotations. Select a single or several annotations or annotation groups in the Annotation editor, select the Export Export annotations context menu item.

Also, you can export annotation(s) from the Project View by the Export Export annotations context menu item.

The Export Annotations dialog will appear:

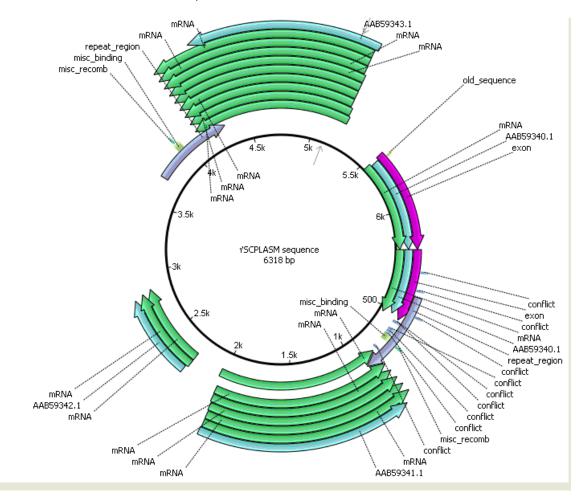
Export Annotations			
Export to file ata/samples/Genbank/murine_annotations.gb			
File format	at GenBank 👻		
✓ Add to project			
Save sequences under annotations			
Save sequence names			
	Help Cancel OK		

Here you can add the document to project, set the path to the file, choose the file format and optionally for CSV format you can save the sequence along with annotations and save sequence names.

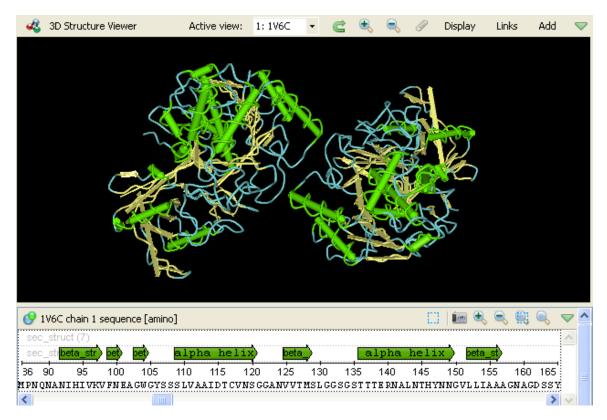
# **Sequence View Extensions**

The functionality of the Sequence View can be significantly increased with Sequence View Extensions. Below is the demonstration its functionality.

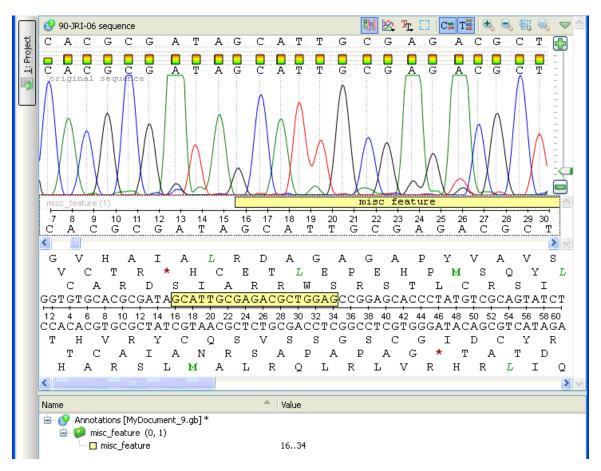
The Circular Viewer shows the circular view of a sequence:



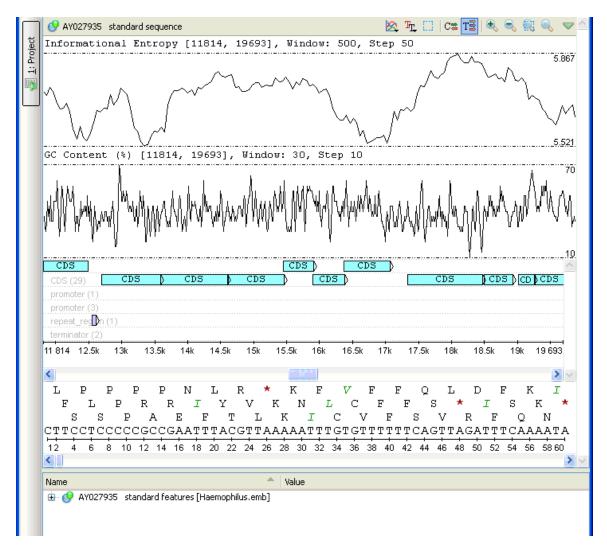
The 3D Structure Viewer adds 3D visualization for PDB and MMDB files:



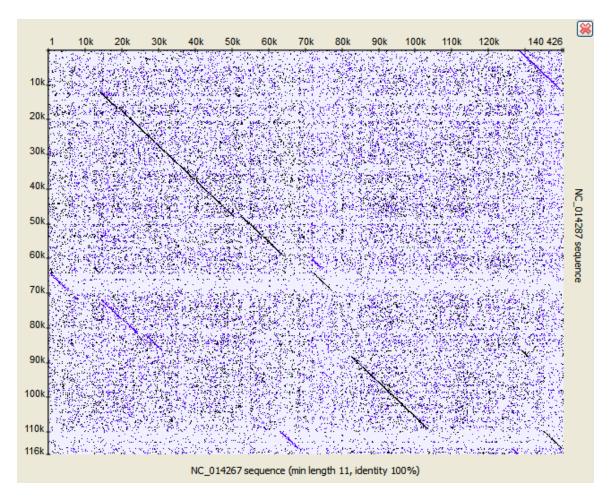
The Chromatogram Viewer adds support for chromatograms visualization and editing:



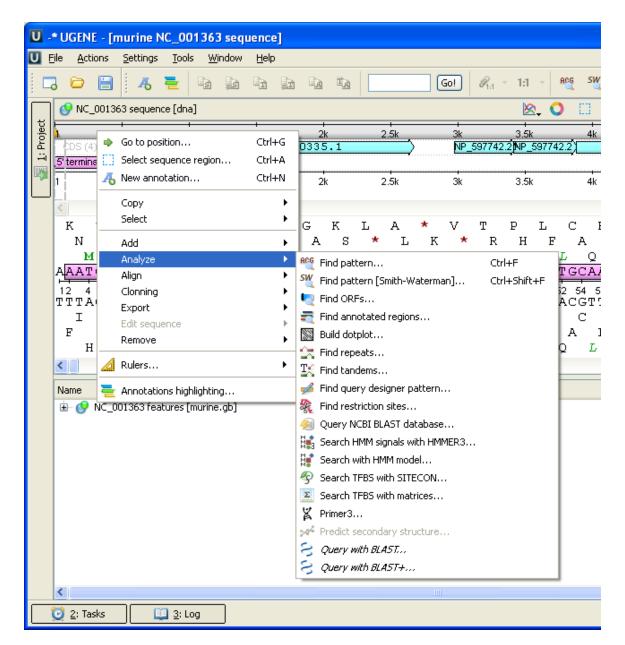
The DNA Graphs Package shows various graphs for sequences:



The Dotplot provides a tool to build dotplots for DNA or RNA sequences.



A number of other instruments add the graphical interface for popular sequence analysis methods:



For details see the next sections of the documentation:

- Circular Viewer
- Circular View Settings
- **3D Structure Viewer** 
  - Opening 3D Structure Viewer
  - Changing 3D Structure Appearance
    - Selecting Render Style
    - Selecting Coloring Scheme •
    - Calculating Molecular Surface
    - Selecting Background Color
    - Selecting Detail Level
    - ٠ Enabling Anaglyph View
    - Moving, Zooming and Spinning 3D Structure
  - Highlight Region on 3D Structure
  - Selecting Models to Display ٠
  - Structural Alignment
  - Exporting 3D Structure Image
  - Working with Several 3D Structures Views
- Chromatogram Viewer
  - Exporting Chromatogram Data
  - Viewing Two Chromatograms Simultaneously
- Graphs Package
  - Description of Graphs •
    - Graph Settings
  - ٠ Saving Graph Cuttoffs as Annotations
- Dotplot
  - Creating Dotplot
  - Navigating in Dotplot

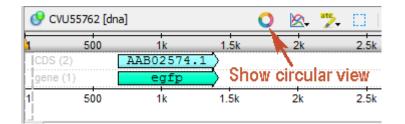
- Zooming to Selected Region
- Selecting Repeat
- Interpreting Dotplot: Identifying Matches, Mutations, Invertions, etc.
- Editing Parameters
- Filtering Results
- Saving Dotplot as Image
  Saving and Loading Dotpl
- Saving and Loading Dotplot
   Building Dotplot for Currently
- Building Dotplot for Currently Opened Sequence
- Comparing Several Dotplots

## **Circular Viewer**

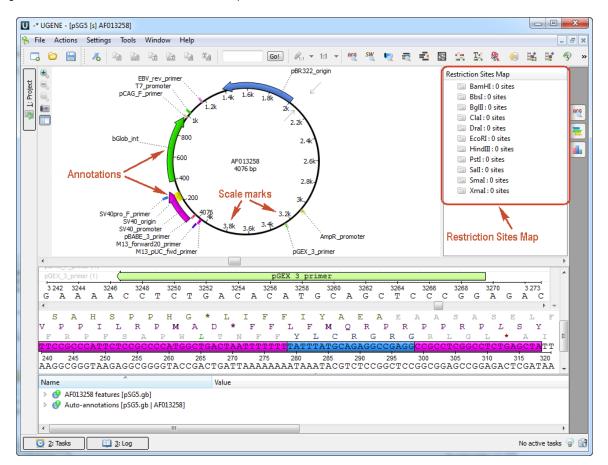
The Circular Viewer plugin provides capability to show the circular view of a nucleotide sequence.

#### Usage example:

Open a nucleotide sequence object in the Sequence View. The Show circular view button is available on the sequence toolbar:



Pressing the button will show the circular view of the sequence:



If you work with file with many sequences the button closes circular views if some circular views are opened and if all circullar views are closed, it opens all of them.

Also you can mark sequences as circular in UGENE by the *Mark as circular* sequence context menu item. When the sequences are marked as *Circular*, the Circular View is automatically opened for them in all opened Sequence View windows.

The *Restriction Sites Map* will appear automatically. To show restriction sites the *Show Restriction Sites* menu should be checked. To hide the map click on the following button:

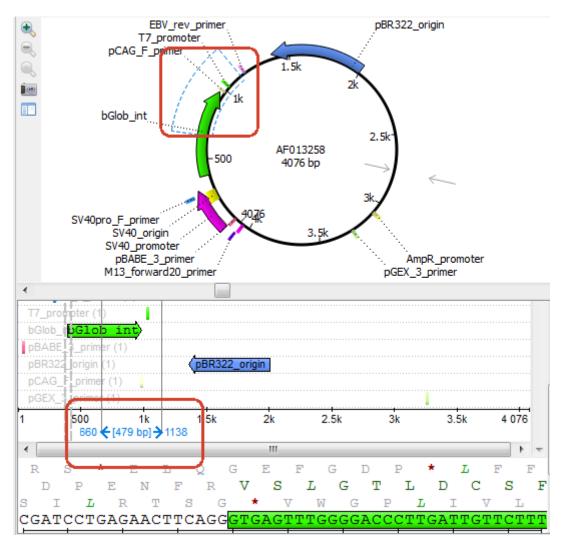


A The Circular Viewer is opened automatically when the Sequence View is opened for a plasmid.

The inner circle represents the sequence clockwise and the scale marks show the corresponding sequence positions. The sequence annotations are represented as curved colored regions at the outer side of the circle.

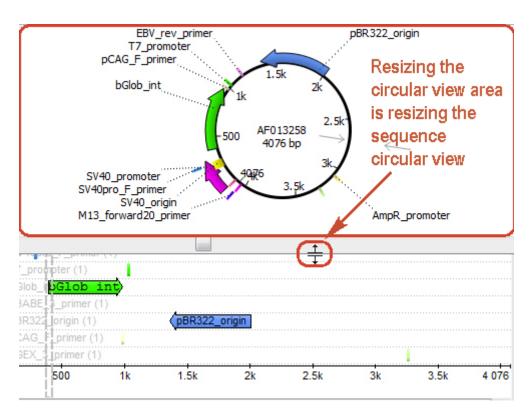
The Circular Viewer helps to navigate within the sequence. You can select an annotation on the circular view and the annotation will also be focused and highlighted in all Sequence View areas: Sequence overview, Sequence zoom view, Sequence details view and Annotations editor.

You can also select a sequence region:



This will also affect the Sequence View. You can select a sequence region with Ctrl and the selection will be inverted.

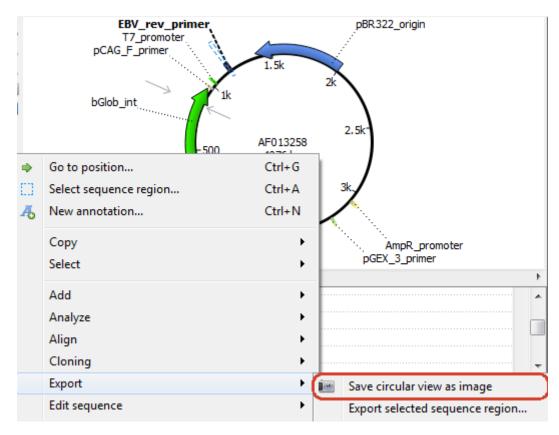
Note that the circular view is zoomed automatically when the Circular Viewer area is resized:



So you can adjust it to an appropriate size.

It is possible to rotate the circular view using the mouse wheel. Also it is possible to shift the start point of a circular molecule by *Edit* sequence -> Set new sequence origin context menu item.

Use the Export Save circular view as image context menu or the Actions main menu item to save the image of the circular view.

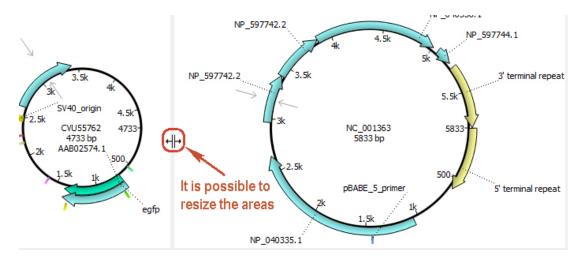


The Export Image dialog will appear:

• Export Image 🛛 😵
Export settings
Include position marker
☑ Include selection
Export to file
File name
Format png 🗘 Quality 80 🗘
Width 1073px 🗘 Height 262px 🗘
Help Cancel Export

Here you can browse for the file name, select the width, height and resolution of the image as well as its format: svg, ps, pdf, bmp, jpeg, jpg, png, ppm, tif or tiff. Also you can include position and selection markers to the image by the corresponding checkboxes.

Note, that if a sequence file contains several sequences it is possible to view the circular views of the sequences in the same *Circular Viewer* area.



You can work with these circular views at the same time.

### **Circular View Settings**

To configure circular view settings go to the Circular View Settings tab in the Options Panel.

Activate the circular view for a sequence and the following settings will appear:

Circular View Se	ettings	
▼ Title Show		RCG
	length	2
	MS Shell Dlg 2	6
Attribute	B	0 ¥
Show	ruler line coordinates	
<ul> <li>Label font si</li> <li>Annotations</li> <li>Label position</li> </ul>	s	
Font size	7	

In the title section you can show or hide title and length, change font, size and attribute.

In the ruler section you can show or hide ruler line and coordinates and change the label font size.

In the annotation section you can select the label position and change the label size.

The following label positions are available:

- · inside all labels are inside of the annotations
- outside all labels are outside of the annotations
- inside/outside if the label can fit the annotation and it is not auto-annotation, it's located inside. Otherwise outside.
- none no labels at all

## **3D Structure Viewer**

The 3D Structure Viewer is intended for visualization of 3D structures of biological molecules.

Using the 3D Structure Viewer you can work with data from the Protein Data Bank (PDB) - a repository for the 3D structural data of large biological molecules, such as proteins and nucleic acids, maintained by the Worlwide Protein Data Bank (wwPDB).

You can work as well with data from the NCBI Molecular Modeling DataBase (MMDB), also known as "Entrez Structure", a database of experimentally determined structures obtained from the RCSB Protein Data Bank.

Find the description of the 3D Structure Viewer' features below.

- Opening 3D Structure Viewer
  - Changing 3D Structure Appearance
    - Selecting Render Style
    - Selecting Coloring Scheme
    - Calculating Molecular Surface
    - Selecting Background Color
    - Selecting Detail Level

- Enabling Anaglyph View
- Moving, Zooming and Spinning 3D Structure
- Highlight Region on 3D Structure
- Selecting Models to Display
   Structural Alignment
- Structural Alignment

•

Exporting 3D Structure ImageWorking with Several 3D Structures Views

## **Opening 3D Structure Viewer**

The 3D Structure Viewer is opened automatically when you open a PDB or MMDB file.

For example, open \$UGENE/data/samples/PDB/1CF7.PDB. The 3D Structure Viewer adds a view to the upper part of the Sequence View:

🤹 3D Structure Viewer	Active view: 1: 1CF7 🔻 Display	Links 🤝
🚱 1CF7 chain 1 sequence [amino]	🖄 🥍 🔂	
	alpha_helix         alpha helix         beta_stra         beta_stra	
* <		F

Notice the *Links* button on the toolbar. When you click the button the menu appears with quick links to online resources with detailed information about the molecule opened:

- PDB Wiki
- RSCB PDB
- PDBsum
- NCBI MMDB

Note that if you're online, you can access the Protein Data Bank directly from UGENE and load a required file by its PDB ID (see *Fetching Data from Remote Database* for details). Hint

Don't forget to select the correct database (PDB) while fetching.

### **Changing 3D Structure Appearance**

This chapter describes how you can change a 3D stucture appearance.

- Selecting Render Style
- Selecting Coloring Scheme
- Calculating Molecular Surface
- Selecting Background Color
- Selecting Detail Level
- Enabling Anaglyph View

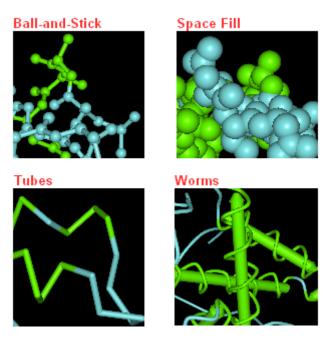
#### **Selecting Render Style**

The following render styles are available:

- Ball-and-Stick
- Space Fill

- Tubes
- Worms

To change the render style select an appropriate item in the *Render Style* menu (it can be found either in the 3D Structure Viewer context menu or in the the *Display* menu on the toolbar).

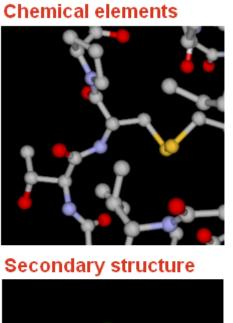


### **Selecting Coloring Scheme**

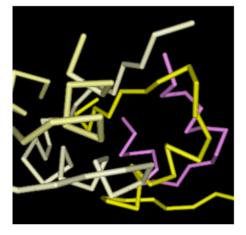
You can select one of the following coloring schemes:

- Chemical Elements
- Molecular Chains
- Secondary Structure
- Simple colors

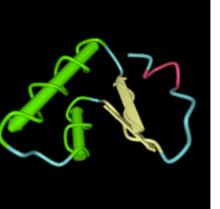
To change the coloring scheme open the Coloring Scheme menu (available in the context menu and in the Display menu on the toolbar).



## Molecular chains



Simple colors





#### **Calculating Molecular Surface**

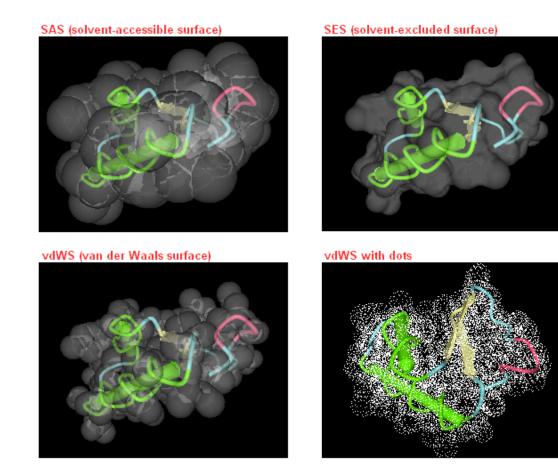
To calculate the molecular surface of a molecule select the *Molecular Surface* item in the 3D Structure Viewer context menu or in the *Display* menu on the toolbar and check one of the following items:

- SAS (solvent-accessible surface)
- SES (solvent-excluded surface)
- vdWS (van der Waals surface)

To remove the molecular surface that has already been calculated select the Off item.

You can also select the Molecular Surface Render Style to modify the calculated molecular surface appearance:

- Convex Map
- Dots



#### **Selecting Background Color**

To change the background color open the Settings dialog (choose the Settings item in the 3D Structure Viewer context menu or in the Displa y menu on the toolbar), press the Set background color button and select a color in the dialog appeared. Selecting Detail Level

To select the detail level of a 3D Structure representation open the *Settings* dialog of the 3D Structure Viewer and drag the *Detail level* slider. **Enabling Anaglyph View** 

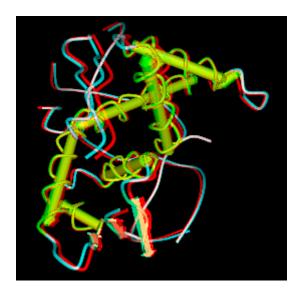
UGENE allows you to view a molecule in the anaglyph mode. To enable the anaglyph view open the *Settings* dialog of the 3D Structure Viewer and check the *Anaglyph view* check box.

You can modify the color settings: select one of the available Glasses colors or set custom colors, swap the colors.

The offset of the color layers can be adjusted by dragging the Eyes shift slider.

🖌 🗹 Anag	yph view	
Eyes shif	ft	
Glasses o	colors	Cyan - Red 🛛 👻
Left	Right	
		Swap colors

See the result the anaglyph view is applied to a molecule below:



## Moving, Zooming and Spinning 3D Structure

A 3D structure can be easily spinned, moved and resized:

- To spin the 3D structure drag the mouse on the 3D structure while holding the left mouse button.
- To move the 3D structure hold the Ctrl keyboard button and drag the mouse with the left button pressed.
- To resize the 3D structure either use the mouse wheel or Zoom In and Zoom Out buttons on the toolbar.

At any time you can restore the default view by pressing the Restore Default View button on the toolbar.



You can also overview the whole structure by spinning it automatically. Select the *Spin* item either in the 3D Structure Viewer context menu or in the *Display* menu on the toolbar to do it.

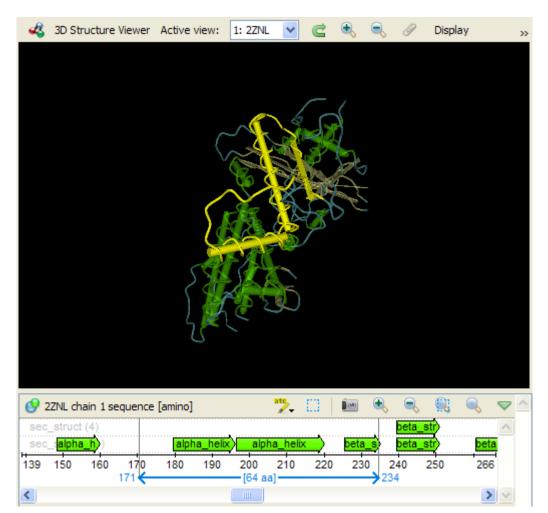
To stop the spinning uncheck the Spin item.

## **Highlight Region on 3D Structure**

When you are selecting a region of a sequence e.g in the Sequence zoom view the corresponding region on the 3D structure is being highlighted while the rest regions of the 3D structure are being shaded.

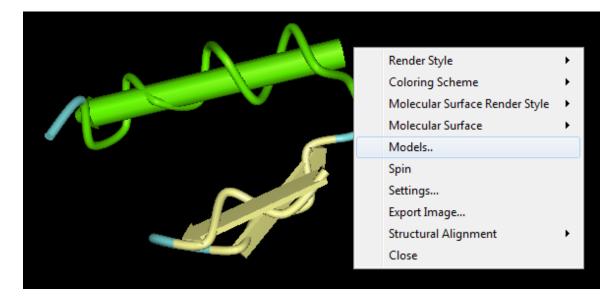
To configure the color of a region selected open the Settings dialog (press the Settings item in the 3D Structure Viewer context menu or in the Display menu on the toolbar to do it), press the Set selection color button and select a color in the dialog appeared.

To adjust the shading drag the Unselected regions shading slider in the Settings dialog.



## **Selecting Models to Display**

When a molecular structure contains multiple models (e.g. NMR ensembles of models), the *Models* item appears in the 3D Structure Viewer context menu and in the *Display* menu on the toolbar. For example, for protein with PDB ID 1ZNF the menu looks as follows:



The dialog will appear:

🚜 Select Models	?	×
		_
		<u>^</u>
4		
▼ 5		
6		
7		
8		
9		
10		
		=
12		=
13		
14		
☑ 15		
<b>1</b> 6		
17		
18		
19		
20		
23		
24		
25		
27		
28		-
All Invert OK Cancel	He	lp

To show all the models check the *All* item. To show only one model check the item and click the *OK* button. To show several models select it and click *OK* button. To show the inverted selection click the *Invert* button and click *OK* button.

## **Structural Alignment**

To use the structural alignment call the Structural alignment->Align with context menu item. The following dialog will appear:

U Structural Alignment		
Algorithm PTools	▼	
Reference	Mobile	
Structure 1CF7  Chain All chains  Region Model 1	Structure 1CF7  Chain All chains Region Model 1	
OK Cancel Help		

Here you can change reference and mobile settings. After that click on the OK button. To reset structural alignment call the *Structural alignment->Reset* context menu item.

## **Exporting 3D Structure Image**

To export a 3D structure image select the *Export Image* item in the 3D Structure Viewer context menu or in the *Display* menu on the toolbar. The *Export Image* dialog will appear:

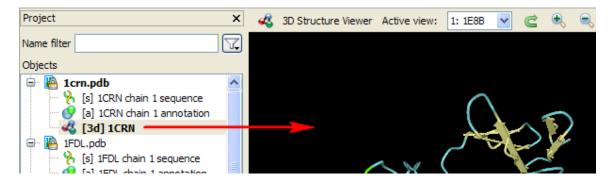
Export Image	8
Export to file	
File name	
Format png 🗘 Quality	
Width 1369px 🗘	Height 275px 🗘
Help	Cancel Export

Here you can browse for the file name, select the width and height of the image as well as its format: svg, png, ps, jpg, jpeg, tiff, tif, pdf, bmp or ppm. For jpg, jpeg formats the quality score parameter is available.

### **Working with Several 3D Structures Views**

To add another view to the 3D Structure Viewer you can:

• Drag a required [3d] object from the *Project View* to the 3D Structure Viewer.

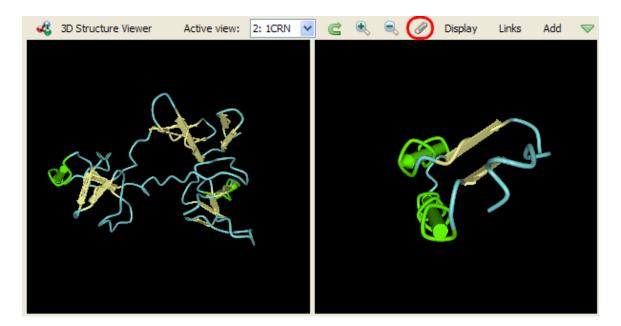


 Press the Add button on the toolbar. The Select Item dialog will appear. Select [3d] objects to add. Hint

Use the Ctrl keyboard button to select several objects.

8	Select Item
	<ul> <li>ICF7.PDB</li> <li>[3d] 1CF7</li> <li>ICRN.PDB</li> <li>[3d] 1CRN</li> </ul>
	OK Cancel

Below you can see the 3D Structure Viewer with two views:

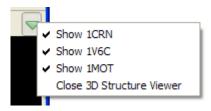


To select an active view click on the view area or select an appropriate value in the Active view combo box on the toolbar.

To synchronize the views press the Synchronize 3D Structure Views sticky button on the toolbar (see the image above). When the button has been pressed the 3D structures are *moved*, *zoomed and spinned* synchronously. Press the button again to stop the views synchronization.

The views that are no more required can be closed by selecting the Close button in the 3D Structure Viewer context menu.

Also you can hide/show views for a while. Use the menu of the green arrow button on the toolbar to do it:



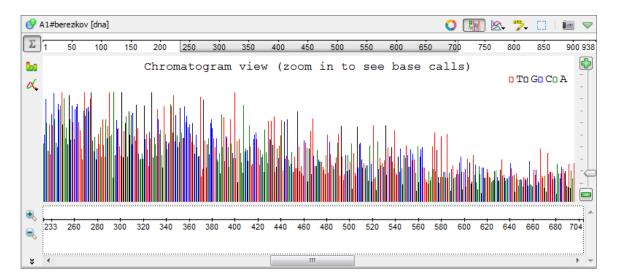
Notice that the 3D Structure Viewer can be closed from this menu.

# **Chromatogram Viewer**

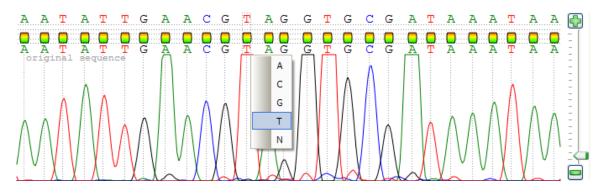
The Chromatogram Viewer plugin brings DNA chromatogram data viewing and editing capabilities into UGENE.

Currently supported chromatogram file formats are ABIF and SCF.

To view a chromatogram, just open an interesting file in UGENE by standard means (e.g. drag&drop the file or press the Ctrl-O shortcut). The *Chromatogram Viewer* is automatically embedded into the generic *Sequence View* if chromatogram data are found, as on the screenshot below:



After zooming in, more chromatogram details are available:



To edit a sequence data, right-click on the chromatogram view and select the *Edit new sequence* item in the appeared context menu. The following dialog will appear:

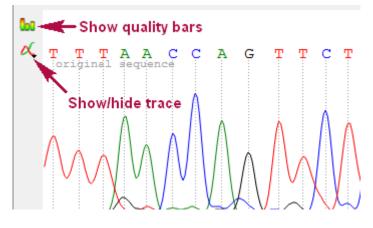
Add New Document	
Document format FASTA	
Document location	
Compress file	
Create Cancel Help	

Select new document format and location and click on the Create button.

The original DNA sequence is not allowed to be changed; however you can add and modify a new sequence stored in a separate file.

The sequence being edited is displayed right above the original one. Symbols can be changed by clicking on interesting value, modifications are shown in bold.

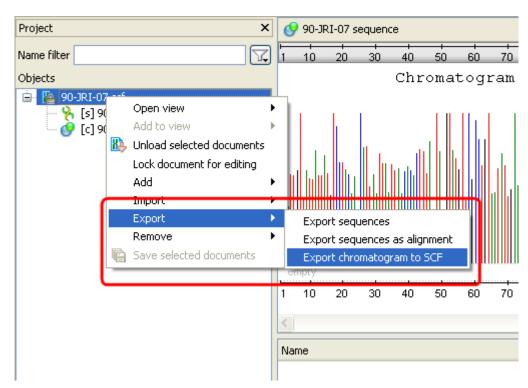
Also you can show/hide different signals of chromatogram and quality bars with a help of the *Show/hide trace* and *Show quality bars* toolbar buttons correspondingly:



- Exporting Chromatogram Data
- Viewing Two Chromatograms Simultaneously

### **Exporting Chromatogram Data**

Open, for example, the \$UGENE/data/samples/SCF/90-JRI-07.srf file. In the *Project View* context menu there is *Export chromatogram to SCF* item:



After clicking on the item, the Export chromatogram file dialog will appear:

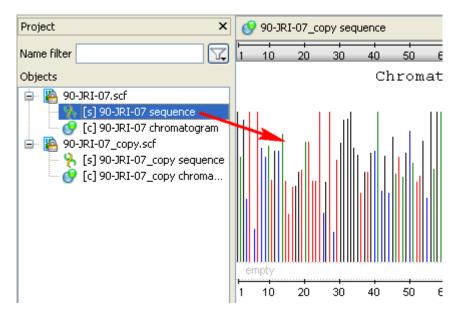
Export Chromatogram File						
Export to file	xport to file C:/work/ugene/data/samples/SCF/90-JRI-07_copy.scf					
File format to use SCF						
Reversed	Complemented					
Add document to the project						
	Export Cancel Help					

Check the *Reversed* and *Complemented* options if you want to create a reverse and complement chromatogram. Press the *Export* button.

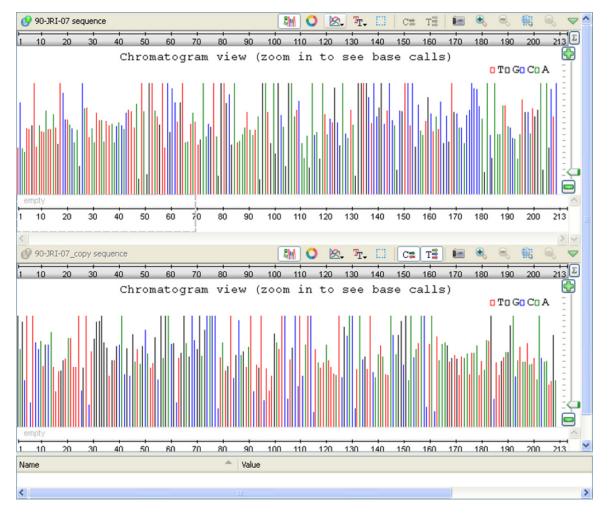
The exported file will be opened in the Sequence View.

## Viewing Two Chromatograms Simultaneously

To add another sequence to the Sequence View, drag the required sequence object from the Project View and drop it in the Sequence View area. (Note that the dragged object is the sequence object, not the chromatogram object.)

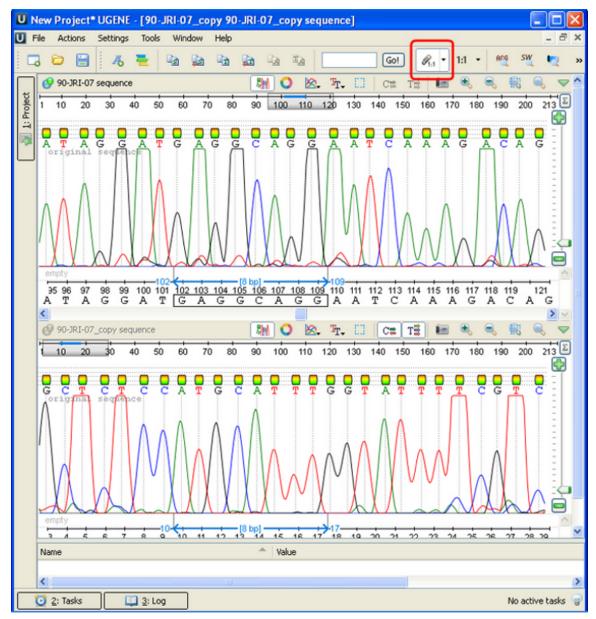


The result will look like this:



You can also use the Lock scales and Adjust scales global actions for the chromatograms.

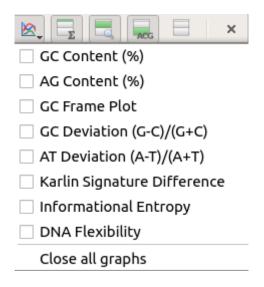
For example if you lock the scales you are able to scroll the sequences simultaneously. Also when you select a sequence region in one sequence, the same region is selected in the second sequence.



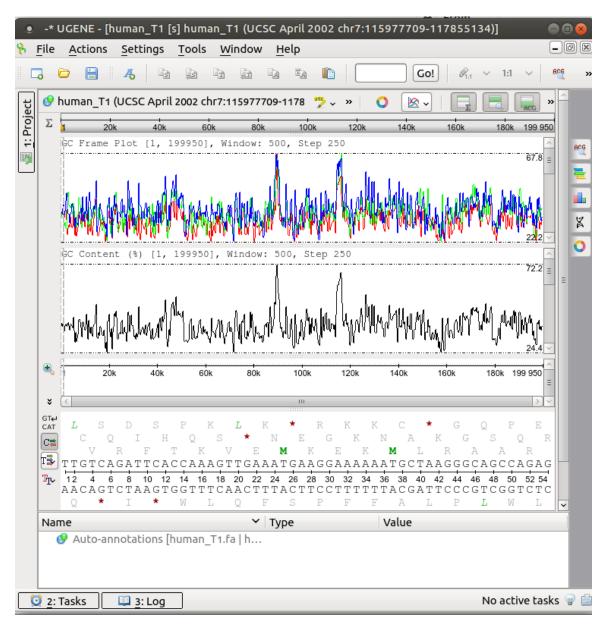
# **Graphs Package**

The DNA/RNA Graphs Package draws contextual graphs for sequences. The DNA/RNA Graphs Package is available for the Standard DNA and Standard RNA alphabets.

Open a sequence in the Sequence View and click the Graphs icon on the toolbar. The popup menu appears:

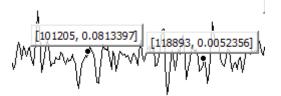


To see a graph select the corresponding graph item in the popup menu. A new area with the graph appears right above the Sequence zoom view:



Each point on a graph is calculated for a window of a specified size. The window is moved along the sequence by a step. See *Graph Settings* for instructions on how to modify these parameters.

It is possible to get information about each point of a graph. When a mouse is moved in the *Graphs* area, a small circle shows on the graph. A coordinates hint shows above it. When you hold Shift and click on a graph, the circle and the hint locks:



To remove it click on the hint. Also, you can delete all labels by *Graph->Delete all labels* context menu. To select all extremum points use the *Graph->Select all extremum points* context menu item.

All graphs are always aligned with the range shown in the Sequence zoom view. It means that if you change the visible range in the overview (either by zooming or scrolling) the graph will also be updated. The minimum and maximum values of the visible range are shown at the right lower and upper corners of the graph.

To close a graph, uncheck its item in the popup menu.

- Description of Graphs
- Graph Settings
- Saving Graph Cuttoffs as Annotations

### **Description of Graphs**

Find below the detailed description of each graph. Note that characters A, C, G and T in the formulas denote the number of corresponding nucleotide in a window.

 DNA Flexibility — searches for regions of high DNA helix flexibility in a DNA sequence. The average Threshold in a window is calculated by the following formula:

```
(sum of flexibility angles in the window) / (the window size - 1)
```

For more detailed information see DNA Flexibility paragraph.

 GC Content (%) — shows the percentage of nitrogenous bases (either guanine or cytosine) on a DNA molecule. It is calculated by the following formula:

(G+C)/(A+G+C+T)\*100

• AG Content (%) — shows the percentage of nitrogenous bases (either adenine or guanine) on a DNA molecule. It is calculated by the following formula:

(A+G)/(A+G+C+T)\*100

- GC Frame Plot this graph is similar to the GC content graph but shows the GC content of the first, second and third position independently. It is most effective in organisms with GC rich genomic sequence but it also works on all microbial sequences.
- GC Deviation (G-C)/(G+C) shows the difference between the "G" content of the forward strand and the reverse strand. GC Deviation is calculated by the following formula:

(G-C)/(G+C)

• AT Deviation (A-T)/(A+T) — shows the difference between the "A" content of the forward strand and the reverse strand. AT Deviation is calculated by the following formula:

(A-T)/(A+T)

 Karlin Signature Difference — dinucleotide absolute relative abundance difference between the whole sequence and a sliding window. Let:

```
f(XY) = frequency of the dinucleotide XY
f(X) = frequency of the nucleotide X
p(XY) = f(XY) / f(X) * f(Y)
p_seq(XY) = p(XY) for the whole sequence
p_win(XY) = p(XY) for a window
```

The Karlin Signature Difference for a window is calculated by the following formula:

```
sum(p_seq(XY) - p_win(XY)) / 16
```

 Informational Entropy — is calculated from a table of overlapping DNA triplet frequencies. The use of overlapping triplets smooths the frame effect. Informational Entropy is calculated by the following formula:

```
-(triplet frequency)*log10(triplet frequency)/log10(2)
```

## **Graph Settings**

To change settings of a graph, select the Graph->Graph settings item in the graph context menu. The Graph Settings dialog appears:

🖄 Graph Settings	2 ×					
Window:	100 🚔					
Steps per window:	100 🚖					
Default color:						
Cutoff for minimum and maximum values						
Minimum:	0.00					
Maximum:	0.00					
ОК	Cancel Help					

The following parameters are available:

Window — the number of bases in a window.

Steps per window - the number of steps in window. The Step is calculated as Window / Steps per window.

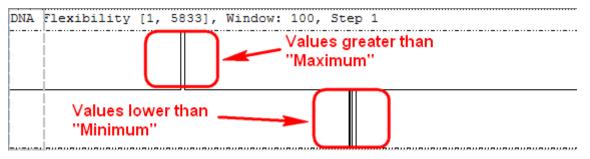
Default color — the default color of line of graph (or lines of graphs for GC Frame Plot).

Checking of the Cutoff for minimum and maximum values checkbox enables the following settings:

Minimum — the minimum value for cutoff.

Maximum --- the maximum value for cutoff.

Select an appropriate minimum and maximum value and click the *OK* button to show the graph of cutoffs. The graph is divided into 2 parts. The upper part shows values greater than the specified *Maximum* value. The lower part of the graph shows values lower than the specified *M inimum* value. For example:



#### Saving Graph Cuttoffs as Annotations

To save graph cuttoffs as annotations select the *Graph->Save cuttoffs as annotations* item in the graph contex menu. The following dialog will appear:

😽 Save Graph Cutof	fs as Annotations	? ×						
Maximum cutoff	52.25 🜩							
Minimum cutoff		16.25 🜩						
Area to annotate	Area to annotate							
Around cutoff v	Around cutoff values							
Between cutoff values								
<ul> <li>Save annotation(s</li> </ul>	▼ Save annotation(s) to							
<ul> <li>Existing table</li> </ul>	• NC_014267 features [NC_014267.1.gb]	▼ 🌇						
Create new ta	ble							
🔘 Use auto-anno	O Use auto-annotations table							
<ul> <li>Annotation parame</li> </ul>	<ul> <li>Annotation parameters</li> </ul>							
Group name	<auto></auto>	*						
Annotation type	Misc. Feature	•						
Annotation name	graph_cutoffs							
Description								
	Save Cancel	Help						

The following parameters are available:

Maximum cutoff - maximum cutoff value.

Minimum cutoff - minimum cutoff value.

Around cutoff values - saves the values around cutoffs values.

Between cutoff values - saves the values between cutoffs values.

In the Save annotation(s) to group you can set up a file to store annotations. It could be either an existing annotation table object, a new annotation table or auto-annotations table (if it is available).

In the Annotation parameters group you can specify the name of the group and the name of the annotation. If the group name is set to <auto> UGENE will use the group name as the name for the group. You can use the '/' characters in this field as a group name separator to create subgroups. If the annotation name is set to by type UGENE will use the annotation type from the Annotation type: t able as the name for the annotation. Also you can add a description in the corresponding text field.

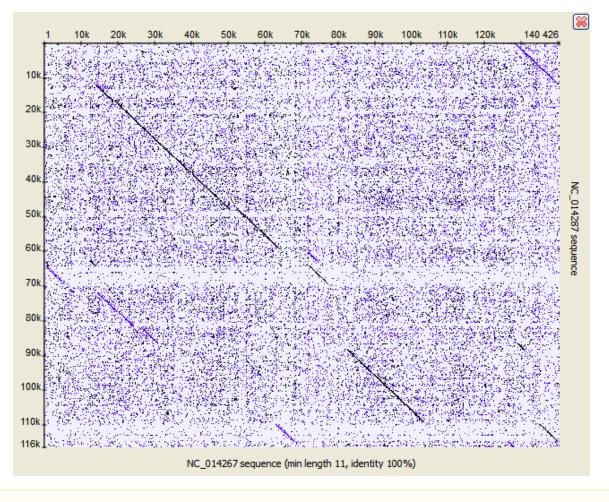
Select the parameters and click on the Save button. The corresponding annotations will be saved.

## Dotplot

The *Dotplot* plugin provides a tool to build dotplots for DNA or RNA sequences. This allows comparing these sequences graphically. Using a dotplot, you can easily identify such differences between sequences as mutations, inversions, insertions, deletions and low-complexity regions.

Also the plugin provides advanced features: comparing multiple dotplots, navigation in a dotplot, dotplots synchronization, saving and loading a dotplot, etc.

An example of a dotplot view:



A The Dotplot plugin uses the Repeat Finder plugin to build a dotplot, make sure you have the Repeat Finder plugin installed.

The Dotplot features are described in more details below.

- Creating Dotplot
- Navigating in Dotplot
- Zooming to Selected Region
- Selecting Repeat
- Interpreting Dotplot: Identifying Matches, Mutations, Invertions, etc.
- Editing Parameters
- Filtering Results
- Saving Dotplot as ImageSaving and Loading Dotplot
- Building Dotplot for Currently Opened Sequence
- Comparing Several Dotplots

### **Creating Dotplot**

To create a dotplot select the Tools Build dotplot main menu item. The Build dotplot from sequences dialog will appear:

U Build Dotplot from Sequences	? ×
File with first sequence	
Join all sequences found in the file	Gap size: 0 🌲
Compare sequence against itself	
File with second sequence	
Join all sequences found in the file	Gap size: 0
	Next Cancel Help

Here you should specify the File with first sequence. Also you should either check the Compare sequence against itself option or select the File with second sequence.

Optionally you can select to *Join all sequences found in the file* (for the first and/or for the second file). If you select to join the sequences you can also select the *Gap size*. The gap of the specified size will be inserted between the joined sequences.

After you press the Next button, the dialog to configure the dotplot parameters will appear:

DotPlot Dotplot parameters	? ×
X axis sequence	human_T1 (UCSC April 2002 chr7:115977709-117855134) 🔻
Y axis sequence	human_T1 (UCSC April 2002 chr7:115977709-117855134) 🔻
	Load Sequence
Search for direct repeats	default
Search for inverted repeats	default
Custom algorithm	Auto 🔻
Minimum repeat length	100bp 🌩 🚺
Repeats identity	100% 🚔 100
	OK Cancel Help

The following parameters are available:

X axis sequence — the sequence for the X dotplot axis.

Yaxis sequence — the sequence for the Y dotplot axis.

If there are several sequences in the specified (the first or the second) file and you haven't selected to join the sequences in the previous dialog, then you can select a sequence in these fields.

If you have selected to Join all sequences found in the file, then you can't select a separate sequence from the file, the joined Sequence can be selected instead.

Search direct repeats — check this option to search for direct repeats in the specified sequences. You can also select the color with which the repeats will be displayed in the picture. The *default* button sets the default color.

Search inverted repeats — check this option to search for inverted repeats in the specified sequences. You can also select the color with which the repeats will be displayed in the picture. The *default* button sets the default color.

Custom algorithm — optionally you can select an algorithm to calculate the repeats:

- Auto
- Suffix index
- Diagonals

A The specified algorithm is provided to the Repeat Finder plugin as an input parameter. In most cases the Auto value is appropriate.

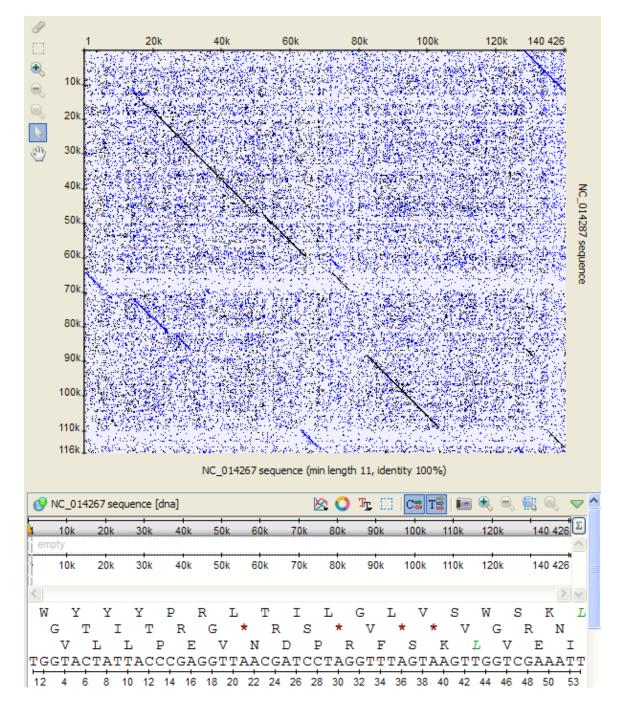
*Minimum repeat length* — allows to draw only such matches between the sequences that are continuous and long enough. For example if it equals to *3bp*, then only repeats will be found that contain 3 and more base symbols.

Press the 1k button to automatically adjust the *Minimum repeat length* value. Such value will be set, that there will be about 1000 repeats found.

Repeats identity - specifies the percents of the repeats identity.

Press the 100 button to set the 100% identity.

After the parameters are set, press the OK button. The dotplot will appear in the Sequence View:



It is a two-dimentional plot consisted of dots.

Each dot on the plot corresponds to a matched base symbol at the "x" position of the horizontal sequence and the "y" position of the vertical sequence.

Visible diagonal lines indicate matches between sequences in the given particular region.

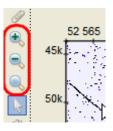
See also:

- Interpreting Dotplot: Identifying Matches, Mutations, Invertions, etc.
- Building Dotplot for Currently Opened Sequence

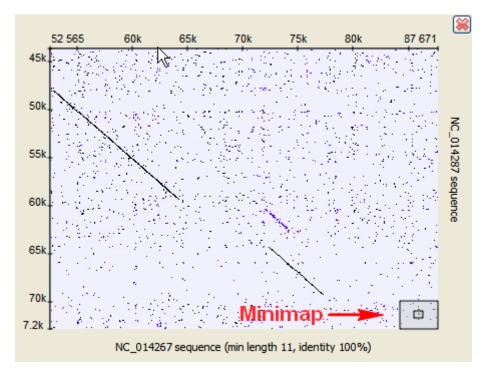
### **Navigating in Dotplot**

To zoom in / zoom out a dotplot you can:

- Rotate the mouse wheel.
- · Press corresponding zoom buttons located on the left:



To move the zoomed region you can:



- Hold the middle mouse button and move the mouse cursor over the zoomed region of the doplot.
- Click on the desired region of the *minimap* in the right bottom corner.
- Activate the Scroll tool, hold the left mouse button and move the mouse cursor over the zoomed region:

# **Zooming to Selected Region**

To select a dotplot region activate the Select tool:

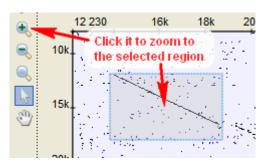
hold down the left mouse button and drag the mouse cursor over the dotplot.

When you select a region on a dotplot the corresponding region is also selected in other Sequence View areas (Sequence details view, Sequ

ence zoom view, etc.).

The opposite is true as well: if you select a region in a Sequence View area, the corresponding region is also selected in the dotplot view.

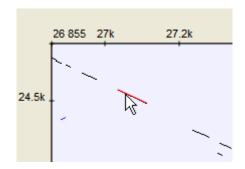
To zoom to the region selected click the Zoom in on the left.



# **Selecting Repeat**

To select a repeat activate the Select tool:

and click on the repeat:



To deselect the repeat either click on other repeat or hold Ctrl and click somewhere on the dotplot.

### Interpreting Dotplot: Identifying Matches, Mutations, Invertions, etc.

Using a dotplot graphic, you can identify such the following differences between the sequences:

#### 1. Matches

A match between sequences looks like a diagonal line on the dotplot graphic, representing the continuous match (or repeat).

#### 2. Frame shifts

#### a. Mutations

*Mutations* are distinctions between sequences. On the graphic they are represented by gaps in diagonal lines. They interrupt matches.

#### **b.** Insertions

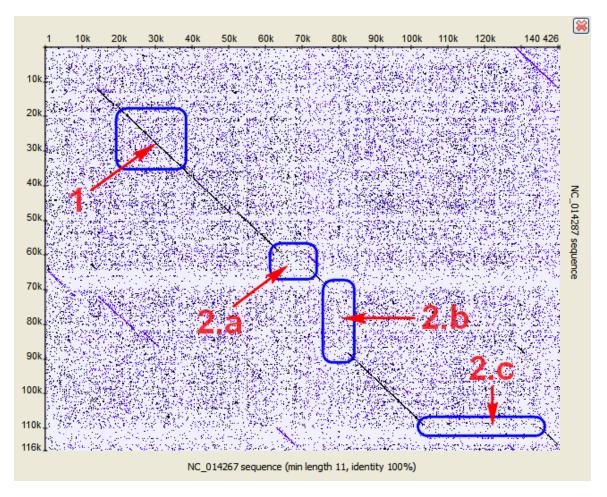
Insertions are parts of one sequence that are missed in the another, while the surrounding parts match. In other words, an insertion is a subsequence that was inserted into a sequence.

Graphically, insertions are represented by gaps which lie only on one axis. A little shift towards the other axis indicates a mutation involved.

#### c. Deletions

A deletion is a subsequence that was deleted from a sequence.

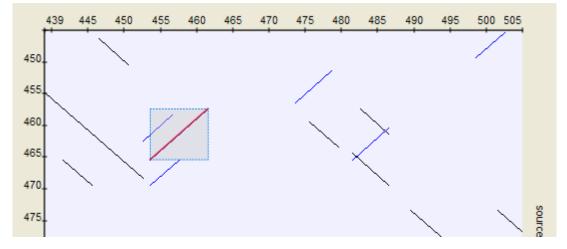
A deletion from sequence A found in sequence B can be considered as an insertion into sequence B and contained in sequence A.



#### 3. Inverted repeats

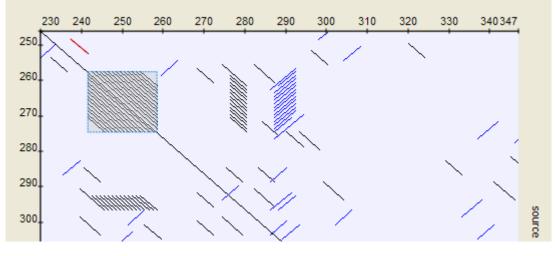
The Dotplot plugin allows to search for inverted repeats as well. Inverted repeats are shown contrary to the direct repeats.

Use the Search direct repeats and Search inverted repeats options of the Dotplot parameters dialog to select which repeats to draw (the dialog is described here).



#### 4. Low-complexity regions

A low-complexity region is a region produced by redundancy in a particular part of the sequence. It is represented on a plot as a rectangular area filled with the matches.



#### Hint

Compare sequence with itself to easily find low-complexity regions in it.

### **Editing Parameters**

It is possible to edit parameters of a built dotplot. Right-click on the dotplot and select the Dotplot Parameters context menu item:

			ھے۔ اندر ان کر اندوں		
	Dotplot	•		Parameters	l
۵	Go to position	Ctrl+G		Save/Load 🕨	
	Select sequence region	Ctrl+A		Remove	
	New annotation	Ctrl+N	$\overline{\mathbf{x}}$	· · · · ·	
	Сору	•	ŀ.		· · · · ·

The parameters dialog will be re-opened. See description of the available parameters here.

### **Filtering Results**

It is possible to find features intersections and filter dotplot results. Right-click on the dotplot and select the *Dotplot Filter results* context menu item. The following dialog will appear:

U DotPlot	? ×
Dotplot parameters           No Filtration           Features Intersection           Intersection Parameters	
Features Selection	
Feature Name	Select All
Sequence X: CVU55762     gene	Different Only
source	Invert Selection
CDS CDS Sequence Y: NC_001363	Clear Selection
<ul> <li>misc_feature</li> <li>source</li> <li>CDS</li> </ul>	
OK	Cancel Help

Select features and click OK button. The filtered dotplot will appear.

# Saving Dotplot as Image

To save a dotplot as image right-click on the dotplot and select the Dotplot Save/Load Save as image context menu item:

-		and a state of the		
Ì	<b>1</b>	Dotplot	•	Parameters
	⇔	Go to position	Ctrl+G	Save/Load  Save as image
		Select sequence region	Ctrl+A	Remove Save :
ì	Æ	New annotation	Ctrl+N	Load
4		Сору	•	
`.		Select		

The following dialog will appear:

Export Image		8
Export settings		
Include area selection		
Include repeat selection		
Export to file		
File name		
Format png 🗘 Quality		80 🗘
Width 1372px -   Height 376px -		DPI 96 🗘
Help	Cancel	Export

Available formats are \*.png, \*.jpg, \*.bmp, \*.jpeg, \*.ppm, \*.tif, \*.tiff, \*.xbm and \*.xpm.

### Saving and Loading Dotplot

To save a dotplot in a native format, right-click on the dotplot and select the Dotplot Save/Load Save context menu item:

-	<u>-</u> -	and a state of the		<u>`</u> .		2		·
Ì	<b>\$</b>	Dotplot	•		Parameters			· `.
	⇔	Go to position	Ctrl+G		Save/Load 🕨		Save as image	
		Select sequence region	Ctrl+A		Remove		Save	
`	<b>A</b>	New annotation	Ctrl+N	_	s - 1 - 5		Load	
4		Сору	•					<u>-</u>
· .		Solart		. ·			· · · · · ·	

The Save Dotplot dialog will appear. A dotplot is saved in a file with the \*.dpt extension.

Later the dotplot can be loaded using the Dotplot Save/Load Load context menu item.

### **Building Dotplot for Currently Opened Sequence**

To build a dotplot for currently opened sequences, create a multiple view containing these sequences. It can be arranged by dragging the corresponding sequence objects (the items strated with the "[s]") into the same Sequence View.

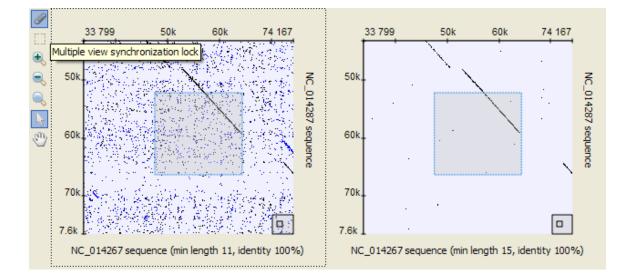
Then right-click on the created view and select the *Analyze Build dotplot* item in the context menu. Every sequence from the current multiple sequence view can be used to build a dotplot.

ho If you need to compare a sequence with itself, you can activate the menu from a single Sequence View.

### **Comparing Several Dotplots**

Dotplots created for the same view are shown at the same view.

If the horizontal and vertical sequences of several dotplots are the same correspondingly, it is possible to lock all zooming and navigating operations for these dotplots. Press the *Multiple view synchronization lock* button on the left.



# **Alignment Editor**

- Overview
  - Alignment Editor Features
  - Alignment Editor Components
  - Navigation
  - Coloring Schemes
    - Creating Custom Color Scheme
  - Highlighting Alignment
  - Zooming and Fonts
  - Searching for Pattern ٠ ٠
  - Consensus
    - Export Consensus
  - Alignment Överview
- Working with Alignment
  - Selecting Alignment Region
  - Moving Subalignment
  - Copying and Pasting Subalignment Editing Alignment
    - Removing Selection
    - Filling Selection with Gaps
    - **Replacing Selected Character**
    - Replacing with Reverse-Complement •
    - Replacing with Reverse •
    - Replacing with Complement
    - Removing Columns of Gaps
    - Removing Sequence
    - **Removing All Gaps**
    - ٠ Undo and Redo Framework
  - Saving Alignment
  - Aligning Sequences
  - Aligning Sequence to this Alignment
  - Pairwise Alignment
    - Working with Sequences List
      - Adding New Sequences
      - Renaming Sequences
      - Sorting Sequences
      - ٠ Shifting Sequences
      - Collapsing Rows
      - Copying Sequences
  - Exporting in Alignment
    - Extracting Selected as MSA
    - Exporting Sequence from Alignment
    - Exporting Alignment as Image
  - Importing APR and ACE Files
- Statistics
  - Distance Matrix
  - Grid Profile
- Advanced Functions
  - Building HMM Profile
- Building Phylogenetic Tree
  - PHYLIP Neighbor-Joining
    - MrBayes
    - PhyML Maximum Likelihood

# Overview

This chapter gives an overview of the Alignment Editor components and explains basic concepts of browsing an alignment.

- **Alignment Editor Features**
- Alignment Editor Components ٠
- Navigation
- **Coloring Schemes** 
  - Creating Custom Color Scheme
- Highlighting Alignment
- . Zooming and Fonts
- Searching for Pattern
- Consensus
- Export Consensus
- Alignment Overview

# **Alignment Editor Features**

The Alignment Editor is a powerful tool for visualization and editing DNA, RNA or protein multiple sequence alignments. The editor supports different multiple sequence alignment (MSA) formats, such as ClustalW, MSF, and Stockholm. The full list of file formats supported in

#### UGENE is here.

The editor provides the interactive visual representation which includes:

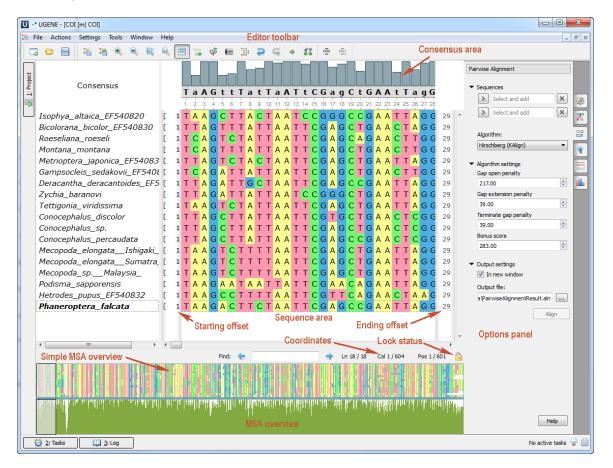
- Navigation through an alignment;
- Optional coloring schemes (for example Clustal, Jalview like, etc.);
- Flexible zooming for large alignments;
- Export publication-ready images for alignment;
- Several consensus calculation algorithms.

Using the Alignment Editor you can:

- Perform multiple sequence alignment using integrated MUSCLE and KAlign algorithms;
- Edit an alignment: delete/copy/paste symbols, sequences, and subalignments;
- Build phylogenetic trees;
- Generate grid profiles;
- Build Hidden Markov Model profiles to use with HMM2/HMM3 tools.

### Alignment Editor Components

Here is the default layout of the editor:



The Alignment Editor components:

For example, let's assume that the coordinate of the first visible base of the row is *N*, but the row contains *K* gaps before the position *N*. The starting offset value will be *N*-*K*. The same rule is true for the ending offset.

You can turn off the Sequence offsets by unchecking the Actions View Show offsets main menu item or View Show offsets context menu item.

### Navigation

The Sequence area provides several flexible ways to navigate through an alignment. The simplest way is to use the mouse and the scrollbars.

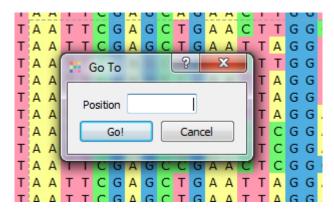
Alternatively, you can use arrow keys on the keyboard to navigate.

The list of hotkeys for quick navigation:

- PageUp to move one screen left.
- PageDown to move one screen right.
- Home to center the starting columns of the alignment.
- End to move to the trailing columns of the alignment
- Hint

if you use Shift key with the hotkeys above you will navigate through the rows. For example, Shift-PageDown will move one screen down.

Finally, you can use the Go to position dialog from the Actions menu, the context menu or the editor toolbar.



Enter the column number (base coordinate) and the view will be centered to the corresponding base.

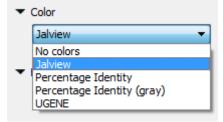
### **Coloring Schemes**

There are various coloring schemes for DNA and amino alphabets available.

To change the scheme, activate the Colors context menu:

		Ctrl+G		Ĭ	ŭ	'	č	<u>^</u>		1
C 🖻	Go to position			G	G	С	С	А		1
C	Add	•		G	G	т	С	А		
	Сору	•	14		<u> </u>	Ľ.	Ŭ	· `		
C	Colors	•		No co	olors					I,
C	Highlighting	•	$\checkmark$	Jalvie	w					
	Edit	•		Perce	ntage	Ider	tity			ľ
Ci	Align	•		Perce	ntage	Iden	tity (	gray)		:
С	Tree	+		UGEN	IE					ŀ
	Statistics	•		Custo	om sc	hem	es		×	Ľ
C	View		<b>-</b>	• •	0	-	C	~	1	-
С	Export	•	0	G	G	А	С	А		1

or use the Highlighting tab of the Options Panel:



• Creating Custom Color Scheme

#### **Creating Custom Color Scheme**

To create custom color scheme use the Colors->Custom schemes->Create new color scheme context menu item. The Application Settings di alog will appear. Click on the Create color scheme button:

General	Alignment Color Scheme
Resources Network File Format Logging Alignment Color Scheme External Tools Genome Aligner OpenCL Workflow Designer	Create and modify custom color schemes Directory to save color scheme: Change color scheme Create color scheme Delete Delete
	OK Cancel Help

The following dilaog will appear:

U Create Alignme	U Create Alignment Color Scheme							
New scheme name	Custom color scheme							
Alphabet	Alphabet Nucleotide							
Use extended m	node							
	Create Cancel Help							

Select the new scheme name, alphabet and click on the Create button. The next dialog will appear for nucleotide extended mode:

U	Color Sch	eme					? <mark>x</mark>	
	-	А	В	С	D	G	Clear Restore last state	
	Η	Κ	Μ	Ν	R	S		
	Т	U	V	W	Х	Y		
OK Cancel Help								

Here you can select a color for each element. Click on the element for it. The new scheme will be created after clicking the *OK* button. The new custom scheme will be available in the *Colors->Custom schemes* context menu.

# **Highlighting Alignment**

TCGAGCCGAATTAGG AΑ CA G ΤА т А Α Т C С GΑ ΤА ΑΑ ССТ GGT G т С ΤА Ctrl+G Go to position ... G С т т А Α CA GG А С С ΤА Add CAGG Т A G А Α С Т ΤА Т C Сору т ΑA CCAGG А G С ТΑ C Т Colors т т А Α Δ т Highlighting ۲ No highlighting т Α Α Δ т Edit ۶ Agreements т А С ΤA Align Þ Disagreements т A С ΤА Tree ۲ Т Α Gaps G ΤА Statistics Α ۲ G Т **d**h ΤA √ Conservation level Т Α G ΤA View ۲ Transitions т Α G ΤА Export Þ Transversions ΤА Т ΤА Advanced • Use dots Т А т ΤА Set this sequence as reference т А GACAACCT т ΤA G G CAACCAGGA G Α ТС Т Т Consensus mode ... CAACCCGGCTG ΤА G А

To apply an alignment highlighting mode, select it in the Highlighting context menu:

or on the Highlighting tab of the Options Panel:

Highlighting	
▼ Reference sequence	
Select and add	٢
▼ Color	K
UGENE	
▼ Highlighting	9
Conservation level 🔹	-
Threshold: 50%	800 H
· · · · · · · · · · · · · · · · · · ·	•
Highlight characters with conservation level:	
⊘ ≤ threshold	
● ≥ threshold	
Use dots Export	

The following modes are available:

- Agreements highlights symbols that coincide with the reference sequence.
- Disagreements highlights nucleotides that differ from the reference sequence.
- Gaps highlights gaps.
- Conservation level highlights conservation level of symbols in a multiple alignment >= or <= treshhold. To select the conservation

parameters use the Highlighting Options Panel tab.

- Transitions highlights transitions.
- Transversions highlights transversions.

To use dots instead of symbols which are not highlighted check the *Use dots* checkbox in the *Options Panel* or use the *Highlighting->Use dots* context menu item.

To select a reference sequence use the Set this sequence as reference context menu or Reference sequence field in the Highlighting tab of the Options Panel.

Also you can export highlighting with a help of the *Export* button in the *Options Panel* or by the *Export->Export highlighted* context menu item. The following dialog will appear:

Export Highlighted to Fil	e	? X
Export to file		
Exported area: from	1 🚔 to	604 🚔
Indexing		
I-based	O-based	
Keep gaps		
📝 Dots instead not highlight	ed	
Export	Cancel	Help

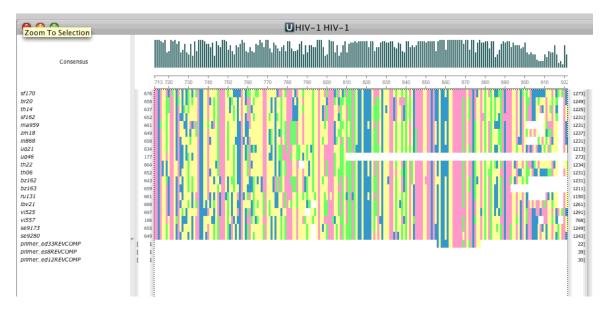
Select file to export, exported area and click on the Export button. The task report will appear in the Notifications.

# **Zooming and Fonts**

To perform zoom operations use the corresponding buttons on the editor toolbar.

								0	GLINE		
	R	0;		T	Ŵ	10	⇔	\$\$	9	Ģ	
ct		0	00								UHIV-1
		Ţ									

By default, the base characters are visible when zooming. But for rather long sequences there is another zoom mode available. In this mode the bases are not shown. This allows viewing very large sequence regions (up to 500 bp).



You can zoom to the selected region by clicking the Zoom to selection button. It is very convenient operation, when the alignment size is

rather large. For example, you can zoom out to some percentage, select an interesting region and then zoom to the selection.

You can change font by clicking the Change font button.

To reset zoom and font click the Reset zoom button.

# **Searching for Pattern**

You can search for a pattern inside an alignment.

Enter a query string in the edit box under the Sequence area.

1	Ν	L	Y	Q	F	G	Е	М	Ι	S	K	K	Т	G	Т	-	F	G	L	F
1	S	L	L	Е	F	G	М	М	I	L	G	Е	т	G	Κ	-	Ν	Ρ	L	Т
1	Н	L	L	Q	F	R	Κ	М	Ι	Κ	Κ	М	т	G	К	-	Е	Ρ	V	Ι
1	н	L	L	Q	F	R	К	М	Ι	К	К	М	т	G	К	-	Е	Ρ	Ι	V
1	È		-	-	3															
		Fir	nd:		4												1		÷	

Press the right arrow to search in the direction "From left to right, from top to bottom". Press the left arrow to search in the direction "From right to left, from bottom to top". If the pattern is found, the result will be focused and highlighted in the *Sequence area*. You can continue the search in any direction from this position.

### Consensus

Each base of a consensus sequence is calculated as a function of the corresponding column bases. There are different methods to calculate the consensus. Each method reveals unique biological properties of the aligned sequences. The *Alignment Editor* allows switching between different consensus modes. To switch the consensus mode go to the *General tab* of the *Options Panel* or activate the context menu (using the right mouse button) or the *Actions* menu and select the *Consensus mode* item and *General tab* will be opened automatically:

General	
▼ Reference sequence	
Select and add	c7.0
- Alizzanak iz 6-	٢
<ul> <li>Alignment info</li> </ul>	8
Alphabet: Standard DNA	•13
Length: 604	
Sequences: 18	-
	¥
▼ Consensus mode	<b>800</b> H
Consensus type:	
Default 🛛 🗘	
Threshold:	
····· 100% _	
✓ Copy to clipboard	
Format:	
CLUSTALW 😂	
Сору	

There are several consensus modes:

- JalView (Default) the mode is based on the JalView algorithm. A character in the consensus is calculated based on the characters in the corresponding column and the value of the threshold:
  - If the percentage value of a character is greater than the threshold, the character is shown in upper case in the consensus.
  - If there are two characters with high frequency in a column, the '+" sign is shown.
  - If the percentage value of a character is high, but lower than the specified threshold, the character is shown in lower case in the consensus.
- ClustalW emulates the ClustalW program behavior.
- Levitsky proposed by Victor Levitsky, this mode calculates the consensus of a DNA alignment, taking into account frequencies of characters in the whole alignment, i.e. not only one column. The algorithm includes these steps:
  - Collect global alignment frequencies for every character using the extended DNA alphabet.
  - For every column select the rarest character in the whole alignment with the percentage in the column greater or equal to the threshold value.
- Simple extended a character of the consensus is calculated based on the characters in the corresponding column. The extended DNA alphabet is used for the result consensus, the threshold value is also taken into account.
- Strict if a character percentage in a column is higher than the specified threshold, the character is shown in the consensus at this position. Otherwise, a gap character ('--') is shown.

Also the *General tab* shows the general information about an alignment and allows one to set up a reference sequence. The following chapter describes how to export a consensus sequence:

Export Consensus

#### **Export Consensus**

To export a consensus sequence use the Export consensus tab of the Options Panel:

Export Consensus	
Export to file	r2+
File format	**
Plain text 🔹	• Q
Info: alphabet of the consensus is	
undefined, the sequence can only be saved into a plain text document.	¥
uocument.	808 II
Export	

The following parameters are available:

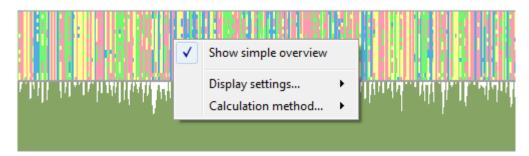
Export to file - the output file path.

*File format* - format of the output file. The possible values depend on the mode, selected for the consensus calculation. For *Default* and *ClustalW* modes the *Plain text* value is only available. For *Levitsky*, *Simple extended* and *Strict* consensus modes sequence formats FASTA and GenBank are also available.

Click on the Export button to export the consensus sequence to the specified output file.

### **Alignment Overview**

The alignment overview is shown automatically in the Alignment Editor. To close the overview click on the *Overview* toolbar button. To show the simple alignment overview use the *Show simple overview* context menu item of the overview.



The following settings of the alignment overview are available:

*Export as image* - you can export multiple alignment overview and simple alignment overview as image. Use this context menu item to do it. In the following dialog select the required parameters and click on the *Export* button:

U Export Image	? ×
Alignment over	
Export to file	
Format	Export Cancel Help

Dispaly settings:

Graph type - sets the graph type: histogram, line graph or area graph.

Orientation - sets the orientation: top to buttom or buttom to top.

Set color - sets the gpaph color.

Calculation method - sets the calculation method: strict, gaps, clustal or highlighting.

To use these settings go to the corresponding context menu items of the alignment overview.

# Working with Alignment

This chapter explains how to work efficiently with the *Alignment Editor*. You will learn how to modify an alignment, remove gaps, align sequences, copy and paste regions, add new sequences and extract subalignments as new alignments.

- Selecting Alignment Region
- Moving Subalignment
- Copying and Pasting Subalignment
- Editing Alignment
  - Removing Selection
  - Filling Selection with Gaps
  - Replacing Selected Character
  - Replacing with Reverse-Complement
  - Replacing with Reverse
  - Replacing with Complement
  - Removing Columns of Gaps
  - Removing Sequence
  - Removing All Gaps
  - Undo and Redo Framework
- Saving Alignment
- Aligning Sequences
- Aligning Sequence to this Alignment
- Pairwise Alignment
  - Working with Sequences List
    - Adding New Sequences
    - Renaming Sequences
    - Sorting Sequences
    - Shifting Sequences
    - Collapsing Rows
    - Copying Sequences
- Exporting in Alignment
  - Extracting Selected as MSA
  - Exporting Sequence from Alignment
  - Exporting Alignment as Image
- Importing APR and ACE Files

# **Selecting Alignment Region**

#### Selection in the Alignment Editor components

In the Name List and the Sequence Area selection is visualized as a rectangle with a dashed border. In the Consensus Area the consensus sequence character of the corresponding columns are shown with light gray background:

Consensus:	lkgtrfiletdasdyaglp														<u> </u>				
	-	39	-	400		402		404		406		408		410		412	-	414	-
LacBicTy3-1	273-	-	-	-	-	A	т	L		т	D	-	s		F	A	-	-	-
AspOryTy3-1	279 -	-	-	-	-	I	к	L	I	т	D	А	s	G	F	А	-	-	-
HisCapTy3-1	279 -	-	-	-	-	s	Q	С	Е	т	D	А	s	G	N	G	-	-	-
BotCinTy3-2	279 -	-	-	-	-	I	м	v	Е	т	D	s	s	v	F	А	-	-	-
HisCapTy3-4	279 -	-	-	-	-	I	Q	I	Е	Ρ	D	А	s	т	F	А	-	-	-
HisCapTy3-5	279 -	-	-	-	-	I	Q	I	Е	s	D	А	F	т	F	т	-	-	-
UncReeTy3-5	277 -	-	-	-	-	I	к	L	к	L	Ν	Ι	s	т	н	т	-	-	-
HisCapTy3-6	281 -	-	-	-	-	L	R	Ι	Е	т	D	G	s	G	v	А	-	-	-
HisCapTy3-2	277 -	-	-	-	-	т	Q	v	Е	т	D	А	s	к	D	G	-	-	-
Pc_Metavir2	274 -	-	-	-	-	F	R	L	Е	С	D	s	s	D	F	А	-	-	-
Pc_Metavir3	274 -	-	-	-	-	F	R	L	Е	С	D	s	s	D	F	А	-	-	-

Information about selection is shown in the statistic bar under alignment.

	Find: 🔙	🚽 L	_n 4 / 26	Col 1 / 708	Pos 1 / 601	Sel 708x10	
--	---------	-----	-----------	-------------	-------------	------------	--

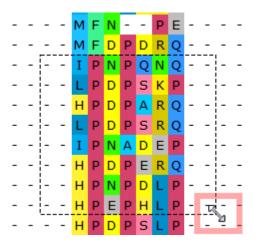
#### Selecting a region using the mouse and keyboard shortcuts

Using the mouse one can select:

- Several consequent sequences in the Name List.
- Several consequent columns in the Consensus Area.
- An alignment region in the Sequence Area.

By holding the *Shift* key it is possible to select a region between two coordinates of a mouse click. For example, if there is a big alignment, one can click on a character (the left upper coordinate of the region), press *Shift*, scroll to the required location of the alignment, and click on another character (the right bottom coordinate of the region).

It is possible to modify a selection in the Sequence Area by dragging the selection rectangle border:



Press the Esc key to clear the selection.

### Moving Subalignment

To move subalignment there are different ways:

- 1. Select a subalignment and drag and drop it. The subalignment will be moved.
- 2. With a help of the *Space*, the subalignment will be moved to the right by the size of the selection. With a help of the *Backspace*, the subalignment will be returned to the first state.
- 3. With a help of the *Ctrl+Space*, the subalignment will be moved to the right by one column. With a help of the *Ctrl + Backspace*, the subalignment will be returned to the first state.

4. With a help of the *Ctrl*, you can join the parts of alignment. When there is a fragment of characters, surrounded by gaps in a row, sel ect continuous block without gaps on *Ctrl* pressed and move the selection to the right/left.

# **Copying and Pasting Subalignment**

To copy current selection click the Copy/Paste Copy selection item in the Actions main menu or the context menu. The hotkey for this action is Ctrl-C.

Also you can copy selection in different formats, for example FASTA, CLUSTALW, RICH TEXT(HTML), NEXUS, MSF.

To copy subalignment in selected format do the following:

- Select the subalignment;
- Select the Copy/Paste->Copy formatted context menu item in the context menu or use hotkey Ctrl+Shift+C.
- The global Options Panel tab will open:

General	
<ul> <li>Reference sequence</li> </ul>	
> Select and add	×
<ul> <li>Alignment info</li> </ul>	
Length:	615
Sequence number:	46
<ul> <li>Consensus mode</li> </ul>	
Consensus type:	
Default	•
Threshold:	2
	100% 🚖
·	100%
Tony to dishaard	
<ul> <li>Copy to clipboard</li> </ul>	
Format:	
Rich text (HTML)	
Сору	

• Select format and press the Copy button.

To paste any of copied selection use the Copy/Paste->Paste context menu item or Ctrl+V shortcut.

# **Editing Alignment**

Select the Edit submenu in the Alignment Editor context menu:

T			ĄΤ			_							GG					-		C	Т	
•	Go to position	Ctrl+G			A G															C	т	
	Add	•			A G															Т		
	Copy/Paste			TG																		
	Colors	•			T G A G																	
	Highlighting	•			TG				_												т Т	
	Edit																	-			Т	
		,		Re	mov	e sel	ecti	on													t,	ł
	Fill selection with gaps																					
	Tree	•		Replace selected character Shift+F										ft+R		t,						
	Statistics	+		Replace selected rows with reverse-complement											E							
	View	+		Replace selected rows with reverse												Ļ						
	Export	+			' eplace									ent							ł	
	Advanced	+	4	Remove columns of gaps Shift+								ft+D	el	t								
	Remove requests																					
	Set this sequence as reference		#=				·														r	
	Consensus mode	\$\$	Re	mov	e all	gap	s	_	_	_	_	_	_	_	_	_	_			J		

The actions available from this menu are described below.

- Removing Selection
- Filling Selection with Gaps
- Replacing Selected Character
- Replacing with Reverse-Complement
- Replacing with Reverse
- Replacing with Complement
- Removing Columns of Gaps
- Removing Sequence
- Removing All Gaps
- Undo and Redo Framework

### **Removing Selection**

To remove a subalignment select it and choose the *Edit Remove selection* item in the context menu or press the Delete key. For Mac OS use the Fn+Delete key instead of the Delete key.

### **Filling Selection with Gaps**

Select a region in the alignment and choose the *Edit Fill selection with gaps* item in the context menu or press the Spacebar. The region is filled with gaps shifting the subalignment from the region to the right.

#### **Replacing Selected Character**

To replace a symbol in the alignment select the nucleotide and press Shift+R for switching the edit mode on.

While the selection is blinking, press a keyboard character that you want to replace the nucleotide with.

After that the edit mode is switched off. For replacing another nucleotide press Shift+R again.

You always can cancel the changes pressing Ctrl+Z.

### **Replacing with Reverse-Complement**

To replace sequence(s) in the alignment with reverse-complement select it and use the *Edit->Replace with reverse-complement* item in the context menu.

#### **Replacing with Reverse**

To replace sequence(s) in the alignment with reverse select it and use the *Edit->Replace with reverse* item in the context menu. **Replacing with Complement** 

To replace sequence(s) in the alignment with complement select it and use the *Edit->Replace with complement* item in the context menu. **Removing Columns of Gaps** 

To remove columns containing certain number of gaps select the Edit Remove columns of gaps item in the context menu. The dialog appears:

Remove Columns of Gaps	? ×
Remove columns with number of gaps	1
Remove columns with percentage of gaps	10%
Remove all gap-only columns	
Remove	cel Help

There are the following options:

Remove columns with number of gaps - removes columns with number of gaps greater than or equal to the specified value.

Remove columns with percentage of gaps - removes columns with percentage of gaps greater than or equal to the specified value.

Remove all columns of gaps — this option is selected by default. It specifies to remove columns from the alignment if they entirely consist of gaps.

Select the option required and press the *Remove* button. **Removing Sequence** 

Select a sequence and call the *Edit Remove sequence(s)* item in the *Actions* main menu or in the context menu to remove sequence from the alignment.

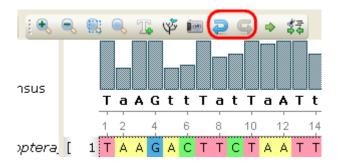
**Removing All Gaps** 

Use the *Edit Remove all gaps* item in the *Actions* main menu or in the context menu to remove all gaps from the alignment. **Undo and Redo Framework** 

The editor tracks all modifications of the aligned sequences.

When a modification happens the current state of the multiple sequence alignments object is being recorded.

You can apply any previous state and redo the modifications using the corresponding buttons on the toolbar:



# **Saving Alignment**

To save current alignment click the Save alignment button, to the the alignment into another file click the Save alignment as button.

U	U -* UGENE - [BL060C3 [m] Contig1]												
33	File	Actio	ons	Settings	Тоо	ls V	Vindo	w H	elp				
		6		-	: <u>e</u>	۹,		9;			Ţ		

# **Aligning Sequences**

The *Alignment Editor* integrates several popular multiple sequence alignment algorithms. Below is the list of available algorithms and links to the documentation:

- Port of the popular *MUSCLE3* algorithm.
- KAlign plugin: effective work with huge alignments.
- ClustalW and MAFFT: these algorithms appeared in the version 1.7.2 of UGENE with the External Tools plugin.

- T-Coffee: this alignment algorithm is available since version 1.8.1 of UGENE with the External Tools plugin.
- ClustalO: Clustal Omega is the latest addition to the Clustal family. It offers a significant increase in scalability over previous versions, allowing hundreds of thousands of sequences to be aligned in only a few hours.

To align sequences choose a preferred alignment method in the Actions main menu, in the context menu or by Align main toolbar button .

Also you may find useful the following video tutorials devoted to the multiple sequence alignment:

- Making a multiple sequence alignment from FASTA file
- Working with large alignments in UGENE
- · Performing profile-to-profile and profile-to-sequence MUSCLE alignments
- Running remote MUSCLE task

### Aligning Sequence to this Alignment

To align a sequence to an opened alignment, click the Align sequence to this alignment toolbar button:



Choose a file with the sequence from the files system, and click Open.

Also, you can add an already opened sequence or sequences to the alignment. To do it, select the sequence object(s) in the Project View an d click the *Align sequence to this alignment* toolbar button. The sequence(s) will be aligned to the alignment automatically.

### Pairwise Alignment

To align two sequences go to the Pairwise Alignment tab of the Options Panel:

Pairwise Alignment	
▼ Sequences	
Select and add	0
> Select and add	1
Algorithm:	
Hirschberg (KAlign) 👻	9
<ul> <li>Algorithm settings</li> <li>Gap open penalty</li> </ul>	800
217.00	
Gap extension penalty	
39.00	
Terminate gap penalty	
39.00	
Bonus score	
283.00	
▼ Output settings	
In new window	
Output file:	
a \PairwiseAlignmentResult.aln	
Align	

Select two sequence from the original alignment, select the parameters and click on the Align button. The following parameters are available:

Algorithm - algorithm of the pairwise alignment. There are two algorithms:

Hirschberg (KAlign) - algorithm has the following parameters:

Gap open penalty - indicates the penalty applied for opening a gap. The penalty must be negative.

Gap extension penalty - indicates the penalty applied for extending a gap.

Terminate gap penalty - the penalty to extend gaps from the N/C terminal of protein or 5'/3' terminal of nucleotide sequences.

Bonus score - a bonus score that is added to each pair of aligned residues.

Smith-Waterman - the following parameters are available:

*Algorithm version* - version of the algorithm implementation. Non-classic versions produce the same results as classic but much faster. To use these optimizations our system must support these capabilities: OPENCL, SSE2 or SW\_classic.

Scoring matrix - scoring matrix.

Gap open penalty - penalty for opening a gap.

Gap extension penalty - penalty for extending a gap.

Output settings - settings of the otput file.

### **Working with Sequences List**

- Adding New Sequences
- Renaming Sequences
- Sorting Sequences
- Shifting Sequences
- Collapsing Rows
- Copying Sequences

### **Adding New Sequences**

You can add new sequences to an alignment using the Add submenu in the Actions main menu or the context menu.

There are two ways to add a new sequence to the current alignment:

- From a file in the compatible format (FASTA, GenBank etc.). The list of the supported data formats can be found here.
- From the current project.

If you activate this item, the following dialog will appear:

Select Item
<ul> <li>ICF7.PDB</li> <li>[s] 1CF7 chain 3 sequence</li> <li>[s] 1CF7 chain 4 sequence</li> <li>[s] 1CF7 chain 4 sequence</li> </ul>
OK Cancel

You will see the *Project View* tree filtered to show only appropriate sequences. Select the items to add and press the *Ok* button. **Renaming Sequences** 

To rename a sequence double click on the name of this sequence and enter a new sequence name in the dialog. **Sorting Sequences** 

To sort sequences by name in the alphabetical order choose the *View Sort sequences by name* item from the *Actions* main menu or the context menu.

### **Shifting Sequences**

To change an order of sequences in a multiple sequence alignment do the following:

- select sequence or sequences in the sequences names list by click or by click and drag correspondingly.
- click and drag on selected region to shift it.

### **Collapsing Rows**

It is able to coolaps the sequential rows. To collapse rows click on the Switch on/off collapsing main toolbar button:



The triangle will appear near collapsed sequences. Click on the triangle to show the whole tree of the collapsed rows.

Consensus	TaAGttTatTaATtCGagC
	1 2 4 6 8 10 12 14 16 18 2
Bicolorana_bicolor_EF540830	TTAGTTTATTAATTCGAGC
Conocephalus_discolor	T T A <mark>G C T T A T T A A T T C G T G C 1</mark>
Conocephalus_percaudata	ΤΤΑ <mark>GCTT</mark> ΑΤΤΑΑΤΤ <mark>CG</mark> ΑGC(
Conocephalus_sp.	T T A <mark>G C </mark> T T A T T A A T T <mark>C G</mark> A <mark>G C</mark> 1
Deracantha_deracantoides_EF540	T T A <mark>G</mark> A T T <mark>G</mark> C T A A T T C G A G C (
Gampsocleis_sedakovii_EF540828	T <mark>C A G A T</mark> T A T T A A T T <mark>C G A G C</mark> 1
Hetrodes_pupus_EF540832	TAAGCCTTTTAATTCGTTC
Isophya_altaica_EF540820	TAAGCTTACTAATCCGGGC
Mecopoda_elongataIshigakiJ	TAAGTCTTTTTAATTCGAGC
Mecopoda elongata Sumatra	TAAGTCTTTTTAATTCGAGC
Mecopoda_spMalaysia_	TAAGTCTTTTAATTCGAGC
Metrioptera_japonica_EF540831	T T A G T C T A C T A A T T C G A G C I
Montana_montana	T C A G T T T A T T A A T T C G A G C T
Phaneroptera_falcata	TAAGACTTCTAATTCGAGCC
Podisma_sapporensis	TAAGAATAATTATTCGAAC
Roeseliana_roeseli	T C A G T C T A T T A A T T C G A G C /
Tettigonia_viridissima	T A A <mark>G T C T A T T A A T T C G A G C</mark> 1
Zychia_baranovi	T T A <mark>G</mark> A T T A T T A A T C C G G G C 1

To update the collapsed groups click on the corresponding main toolbar button . **Copying Sequences** 

To copy one or several sequences do the following:

- Select the sequences in the Sequence list area;
- Select the Copy/Paste Copy selection context menu item in the Sequence area or use hot key combination. Note, that if you activate context menu in the Sequence list area you will lose your current selection.

A C A C
AC
C
C
C
C
C
C
A C
A C
-

To copy consensus sequence use the Copy/Paste Copy consensus item. To copy consensus with gaps use the corresponding menu item.

# **Exporting in Alignment**

- Extracting Selected as MSA
- Exporting Sequence from Alignment
- Exporting Alignment as Image

#### **Extracting Selected as MSA**

It is possible to extract a subalignment and save it as new multiple sequence alignment (MSA).

Select a subalignment and choose the *Export Save subalignment* item in the *Actions* main menu or in the context menu. The following dialog appears:

Extract Selected as MSA
From 3 🔷 to 31 🐳
Selected sequences
<ul> <li>Phaneroptera_falcata</li> <li>Isophya_altaica_EF540820</li> <li>Bicolorana_bicolor_EF540830</li> <li>Roeseliana_roeseli</li> <li>Montana_montana</li> <li>Metrioptera_japonica_EF540831</li> <li>Gampsocleis_sedakovii_EF540828</li> <li>Deracantha_deracantoides_EF540</li> <li>Zychia_baranovi</li> <li>Tettigonia_viridissima</li> <li>Conocephalus_discolor</li> <li>Conocephalus_sp.</li> <li>Conocephalus_percaudata</li> </ul>
Invert selection Select all Clear selection
File name ata/samples/CLUSTALW/COI_subalign1.aln
File format to use CLUSTALW
Add to project Extract Cancel Help

Specify the name and format of the new MSA file in the *File name* and *File format to use* fields. The currently selected region is extracted by default when you press the *Extract* button.

You can change the columns to be extracted using the From and to fields. And change the rows to be extracted by checking/unchecking

required sequences in the Selected sequences list.

Use buttons:

- Invert selection to invert the selection of the sequences.
- Select all to select all sequences.
- *Clear selection* to clear the selection of all sequences.

The Add to project check box specifies to add the MSA file created from the subalignment to the active project.

#### **Exporting Sequence from Alignment**

To export one sequence from an alignment select the sequence in the sequence list or in the sequence area and use the *Export->Save* sequence context menu item. The following dialog will appear:

U Export Selected	d Sequence from Alignment	? <mark>×</mark>
Export to file		
File format to use	FASTA	•
Add document		
Gap characters (	(∼) © Trim	
Neep		
	Export Cancel	Help

Here it is possible to specify the result file location, to select a sequence file format, to define whether to keep or remove gaps ('--' chars) in the sequence and optionally add the created document to the current project.

### **Exporting Alignment as Image**

To export an alignment as image click the *Export as image* button on the editor toolbar or call the *Export->Export as image* context menu item.



The Export Image dialog will appear where you should set name, location, export settings and format of the picture:

Export Image	8
Export settings	
Region Whole alignment	≎ Select
✓ Include sequences' names	
✓ Include consensus	
✓ Include ruler	
Export to file	
File name	
Format png 🗘 Quality	···· 80 🗘
Help	Cancel Export

UGENE supports export to the BMP, CUR, DDS, ICNS, JP2, JPEG, JPG, PBM, PGM, PNG, PPM, TIF, TIFF, WBMP, WEBP, XBM, XPM and SVG image formats. You can export whole alignment or custom region. To select the custom region click on the *Select* button.

# Importing APR and ACE Files

To start working with APR file you can open it in the Alignment Editor in read-only mode or convert it to another format.

To do this, *open* the \*.apr file. The following dialog will appear:

Select Docur	ment Format 🛛 😣
Options for gyrA	A.apr
Open in rea	ad-only mode
○ Convert to	another format:
Save to fiie	ugene/data/samples/APR/gyrA
File format	PHYLIP Sequential
Help	<mark>⊗ ⊆</mark> ancel <u>√</u> <u>O</u> κ

If you choose the first option the file will be opened in the Alignment Editor as multiple sequence alignment in read-only mode. If you choose the second option the file will be converted to a corresponding format and open in UGENE.

The following formats are available: CLUSTALW, FASTA, Mega, MSF, NEXUS, Phylip Interleaved, Phylip Sequential, Stockholm.

To start working with ACE file you can open it in the Alignment Editor or import it to the UGENE database file.

To do this, open the \*.ace file. The following dialog will appear:

• Select Docur	nent Format 🛛 😵
Options for K26.	ace
<ul> <li>Multiple se</li> </ul>	quence alignment in the Alignment Editor
○ Short reads	assembly in the Assembly Browser
Save to file	lata/samples/ACE/K26.ace.ugenedb
Help	<mark>⊗ c</mark> ancel <u>√</u> <u>o</u> ĸ

If you choose the first option the file will be opened in the Alignment Editor as multiple sequence alignment. If you choose the second option the file will be opened in Assembly Browser.

# **Statistics**

To show statistics use the Statistic tab of the Options Panel:

Statistics	
<ul> <li>Reference sequence</li> <li>Select and add</li> </ul>	<u>.</u>
▼ Distances column	1
Show distances column	
Distance algorithm:	
Hamming dissimilarity 🔹	1
Profile mode	800 II
Counts	
Percents	
Exclude gaps	
Automatic updating	
Press button to update	
Hint: select a reference above	

Here you need to select a reference sequence.

You can show/hide distance column by Show distances column checkbox.

Also you can change the distance algorithm, select the profile mode and exclude gaps.

To update the column use Automatic update checkbox or press corresponding button to update.

To generate distance matrix and grid profile see the documentation below:

- Distance Matrix
- Grid Profile

### **Distance Matrix**

Using the Alignment Editor you can also create a distance matrix of a multiple sequence alignment.

To create a distance matrix, use the Statistics Generate distance matrix item in the Actions main menu or in the context menu.

The dialog will appear:

Generate Distance Matrix
Distance algorithm Hamming dissimilarity
Profile mode
Ount
○ Percentage
✓ Exclude gaps
Show group statistics of multiple alignment
Save profile to file
File ENE_Data/COI_distance_matrix_1.html
Hypertext (.HTML)
O Comma separated ( .CSV)
Help Cancel Generate

The following parameters are available:

Distance algorithm - there are two distance algorithms: "Hamming distance" for dissimilarity and "Simple similarity" for similarity.

Profile mode: Counts/Percents - select the Percents to have scores shown as percents in the report. Also, you can Exclude gaps.

Show group statistics of multiple alignment - shows group statistics when the collapsing is switched on.

Save profile to file — allows saving profile to a file in the HTML or CSV format. The CSV format is convenient for further processing in worksheets editors like Excel.

The resulting profile in the HTML mode:

		Isophya_altaica_EF540820		
Phaneroptera_falcata	0	106	118	115
Isophya_altaica_EF540820	106	0	115	119
Bicolorana_bicolor_EF540830	118	115	0	54
Roeseliana_roeseli	115	119	54	0
Montana_montana	116	118	85	75
Metrioptera_japonica_EF540831	113	115	84	72
3ampsocleis_sedakovii_EF540828	128	125	101	97
Deracantha_deracantoides_EF540	110	109	91	92
Zychia_baranovi	100	114	109	112
Tettigonia_viridissima	114	110	104	99
Conocephalus_discolor	123	115	110	116
Conocephalus_sp.	122	114	110	114
Conocephalus_percaudata	130	121	123	120
Mecopoda_elongataIshigakiJ	103	100	107	100
Mecopoda_elongataSumatra_	103	100	107	100
Mecopoda_spMalaysia_	102	101	102	98
Podisma_sapporensis	116	128	120	116
Hetrodes pupus EF540832	152	162	154	146

### **Grid Profile**

Using the Alignment Editor you can create a statistic profile of a multiple sequence alignment.

The alignment grid profile shows positional amino acid or nucleotide counts highlighted according to the frequency of symbols in a row.

To create a grid profile, use the Statistics Generate grid profile item in the Actions main menu or in the context menu.

To learn more about this feature, refer to the DNA Statistics plugin documentation.

# **Advanced Functions**

This chapter is devoted to the advanced functions of the *Alignment Editor*. You will learn how to build a grid profile, export a picture of an alignment and build HMM profiles.

Building HMM Profile

# **Building HMM Profile**

The editor has capabilities to build a Hidden Markov Model profile based on the multiple sequence alignment.

This functionality is based on the Sean Eddy's HMMER package.

To build a HMM profile select the Advanced Build HMMER2 profile or the Advanced Build HMMER3 profile item in the Actions main menu or in the context menu.

Learn more about the HMM tool in the documentation pages of the HMM2 and the HMM3 plugins.

# **Building Phylogenetic Tree**

To build a tree from an alignment either press the *Build Tree* button on the toolbar, select the *Tree Build Tree* item in the alignment context menu or the *Actions Tree Build Tree* item in the main menu.

Ý

Also you can use Tree Settings tab of the Options Panel:



Three methods for building phylogenetic trees are supported:

- 1. The PHYLIP Neighbour-Joining method. The PHYLIP package implementation of the method is used under the hood.
- 2. The MrBayes external tool. Check MrBayes Web Site for more details.
- 3. PhyML Maximum Likelihood method. Check PhyML Maximum Likelihood Web Site for more details.
- PHYLIP Neighbor-Joining
- MrBayes
- PhyML Maximum Likelihood

# **PHYLIP Neighbor-Joining**

The Building Phylogenetic Tree dialog for the PHYLIP Neighbour-Joining method has the following view:

Build Phylogenetic Tree	8 ×			
Tree building method	PHYLIP Neighbor Joining			
Distance Matrix Bootstrapping and Consensus	Tree Display Options			
Distance matrix model	F84			
Gamma distributed rates across sites				
Coefficient of variation of substitution rate among s	ites 0.50			
Transition/transversion ratio	2.00			
Save tree to				
Remember Settings Restore Default				
	Build Cancel Help			

#### The following parameters are available:

Distance matrix model — model to compute a distance matrix. The following values are available for a nucleotide multiple sequence alignment:

- F84
- Kimura
- Jukes-Cantor
- LogDet

The following models are available for a protein alignment:

- Jones-Taylor-Thornton
- Henikoff/Tillier PMB
- Dayhoff PAM
- Kimura

Gamma distributed rates across sites — specifies to take into account unequal rates of change at different sites. It is assumed that the distribution of the rates follows the Gamma distribution.

Coefficient of variation of substitution rate among sites — becomes available if the Gamma distributed rates across sites parameter is checked. Specifies the coefficient of the distribution of the rates.

Transition/transversion ratio - expected ratio of transitions to transversions.

To enable bootstrapping check the Bootstrapping and Consensus Trees group check box. The following parameters are available:

Number of replicates - number of replicate date sets.

Seed — random number seed. By default, it is generated automatically. You can manually change this value in order to make results of different runs (of a tree building) reproducible. The should must be an integer greater than zero and less than 32767 and which is of the form 4n+1, that is, it leaves a remainder of 1 when divided by 4. Any odd number can also be used, but may result in a random number sequence that repeats itself after less than the full one billion numbers. Usually this is not a problem.

Consensus type — specifies the method to build the consensus tree. Select one of the following:

- Strict specifies that a set of species must appear in all input trees to be included in the strict consensus tree.
- *Majority Rule (extended)* specifies that any set of species that appears in more than 50% of the trees is included. The program then considers the other sets of species in order of the frequency with which they have appeared, adding to the consensus tree any which are compatible with it until the tree is fully resolved. This is the default setting.
- M1 includes in the consensus tree any sets of species that occur among the input trees more than a specified fraction of the time (see the *Fraction* parameter below). The *Strict* consensus and the *Majority Rule* consensus are extreme cases of the MI consensus, being for fractions of 1 and 0.5 respectively.
- *Majority Rule* specifies that a set of species is included in the consensus tree if it is present in more than half of the input trees.

Fraction — becomes available when the Consensus type parameter is set to M1. Specifies the fraction.

Display tree in new window - displays tree in new window.

Display tree with alignment editor - displays tree with alignment editor.

Synchronize alignment with tree - synchronize alignment and tree.

Save tree to - file to save the tree built.

Press the *Build* button to build a tree with the parameters selected.

#### **MrBayes**

The Building Phylogenetic Tree dialog for the MrBayes method has the following view:

Build Phylogenetic Tre	e		? ×
Tree building method		MrBayes	•
Model MCMC Di	splay Options		
Chain length	10000		<b>.</b>
Subsampling frequence	1000		* *
Burn-in length	10		*
Heated chains	4		*
Heated chain temp	0.40		*
Random seed	8789		*
Save tree to			
	tore Default	Build Cancel	Help

There are two steps to a phylogenetic analysis using MrBayes:

- 1. Set the evolutionary model.
- 2. Run the Markov chain Monte Carlo (MCMC) analisys.

The evolutionary model is defined by the following parameters:

Substitution model — specifies the general structure of a DNA substitution model. This parameter is available for the nucleotide sequences. It corresponds to the Nst setting of MrBayes. You may select one of the following:

- JC69 (Nst=1)
- HKY85 (Nst=2)
- GTR (Nst=6)

*Rate matrix (fixed)* — specifies the fixed-rate amino-acid model. This parameter is available for amino-acid sequences. The following models are available:

- poisson
- jones
- dayhoff
- mtrev
- mtmam
- wag
- rtrev
- cprev
- vt
- blosum
- equaline

The following parameters are common for nucleotide and amino-acid sequences:

Rate — sets the model for among-site rate variation. Select one of the following:

- equal no rate variation across sites.
- gamma gamma-distributed rates across sites. The rate at a site is drawn from a gamma distribution. The gamma distribution has a single parameter that describes how much rates vary.
- propinv a proportion of the sites are invariable.

• invgamma — a proportion of the sites are invariable while the rate for the remaining sites are drawn from a gamma distribution.

Gamma - sets the number of rate categories for the gamma distribution.

You can select the following parameters for the MCMC analisys:

*Chain length* — sets the number of cycles for the MCMC algorithm. This should be a big number as you want the chain to first reach stationarity, and then remain there for enough time to take lots of samples.

Subsampling frequency — specifies how often the Markov chain is sampled. You can sample the chain every cycle, but this results in very large output files.

Burn-in length — determines the number of samples that will be discarded when convergence diagnostics are calculated.

Heated chains — number of chains will be used in Metropolis coupling. Set 1 to use usual MCMC analysis.

Heated chain temp — the temperature parameter for heating the chains. The higher the temperature, the more likely the heated chains are to move between isolated peaks in the posterior distribution.

Random seed — a seed for the random number generator.

Display tree in new window - displays tree in new window.

Display tree with alignment editor - displays tree with alignment editor.

Synchronize alignment with tree - synchronize alignment and tree.

Save tree to - file to save the built tree.

Press the Build button to run the analysis with the parameters selected and build a consensus tree.

### PhyML Maximum Likelihood

The Building Phylogenetic Tree dialog for the PhyML Maximum Likelihood method has the following view:

Build Phylogenetic Tree						
Tree building method		PhyML Maxir	num Likelihood	•		
Substitution Model	Branch Support	Tree Searching	Display Options			
Substitution model		HKY85	^	•		
Equilibrium frequenci	ies		optimized 🔾 🖲 e	npirical		
Transition / transvers	ion ratio	0	fixed $\bigcirc$ $\odot$ es	stimated		
Proportion of invariable sites		0.00 🌲	fixed $\bigcirc$ $\odot$ es	stimated		
Number of substitution rate categories 4						
Gamma shape parame	eter	0	fixed $\bigcirc$ $\odot$ es	stimated		
Save tree to /home//ugene/data/samples/CLUSTALW/COI.nwk						
Save Settings Restor	e Default					
Help			Cancel	Build		

#### The following parameters are available:

Substitution model parameters - a selection of the Markov model of substitution:

Substitution model - model of substitution.

Equilibrium frequencies - equilibrium frequencies.

Transition/transversion ratio - fix or estimate the transition/transversion ratio in the maximum likelihood framework.

Proportion of invariable sites - the proportion of invariable sites, i.e., the expected frequency of sites that do not evolve, can be

#### **Unipro UGENE Manual, Version 1.30**

fixed or estimated.

Number of substitution rate categories - the number of substitution rate categories.

Gamma shape parameter - the shape of the gamma distribution determines the range of rate variation across sites.

Branch support parameters - the selection of the method that is used to measure branch support:

Use fast likelihood method - use fast likelihood method.

*Perform bootstrap* - the support of the data for each internal branch of the phylogeny can be estimated using non-parametric bootstrap.

Tree searching parameters - the selection of the tree topology searching algorithm:

Make initial tree automatically - initial tree automatically.

*Type of tree improvement* - type of tree improvement.

Set number of random starting tree - the number of random starting tree.

Optimize tree topology - the tree topology is optimised in order to maximise the likelihood.

Optimize branch lengths - optimize branch lengths.

Optimise substitution rate - substitution rate parameters are optimized.

Display tree in new window - displays tree in the new window.

Display tree with alignment editor - displays tree with alignment editor.

Synchronize alignment with tree - synchronize alignment and tree.

Save tree to - file to save the built tree.

Press the Build button to run the analysis with the parameters selected and build a consensus tree.

# **Sanger Reads Editor**

- Sanger Reads Editor Overview
  - Sanger Reads Editor Features
  - Sanger Reads Editor Components
- Working with Chromatogram
  - Mapping Reads to Reference
  - Alignment Statistics Alignment Appearance
  - Overview and Show and Hide Chromatogram
  - Sanger Reads Consensus
  - Export Chromatogram Consensus
  - Navigation in Sanger Reads Alignment
  - Editing Sanger Reads
    - Inserting Character
    - Replacing Character and Gap
      Removing Character and Gap

    - Inserting Gap
    - Removing Gap at the Left
    - Removing All Columns of Gaps
    - Trimming Left End
    - Trimming Right End
    - Renaming Read
    - Removing Read
    - Undo and Redo
  - Exporting Alignment without Chromatograms

### Sanger Reads Editor Overview

This chapter gives an overview of the Sanger Reads Editor components and explains basic concepts of browsing a chromatogram.

- Sanger Reads Editor Features
- Sanger Reads Editor Components

### Sanger Reads Editor Features

Sanger Reads Editor is dedicated to DNA sequence analysis and manipulation. You can trim, map to reference and view Sanger sequencing trace files and create consensus sequences.

The editor provides interactive visual representation which includes:

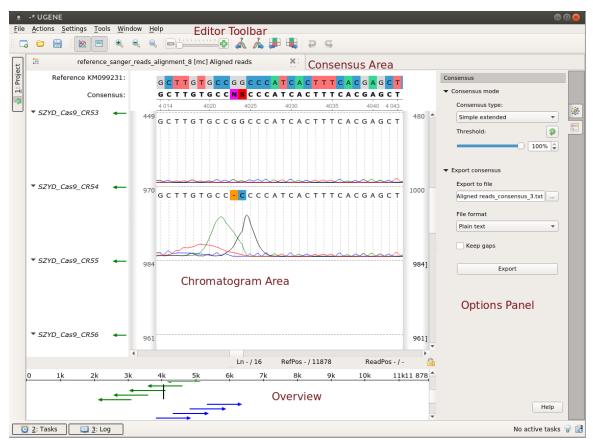
- Navigation through an chromatogram;
- · Flexible zooming;
- · Several consensus calculation algorithms.

Using the Sanger Reads Editor you can:

- · Export alignment without chromatogram;
- Edit chromatogram: remove/replace/trim.

### Sanger Reads Editor Components

Here is the default layout of the editor:



# Working with Chromatogram

This chapter explains how to work efficiently with the Sanger Reads Editor.

- Mapping Reads to Reference
- Alignment Statistics

•

- Alignment Appearance
- Overview and Show and Hide Chromatogram
- Sanger Reads Consensus
  - Export Chromatogram Consensus
  - Navigation in Sanger Reads Alignment
  - Editing Sanger Reads
    - Inserting Character
      - Replacing Character and Gap
      - Removing Character and Gap
      - Inserting Gap
      - Removing Gap at the Left
      - Removing All Columns of Gaps
      - Trimming Left End
      - Trimming Right End
      - Renaming Read
      - Removing Read
      - Undo and Redo
- Exporting Alignment without Chromatograms

### **Mapping Reads to Reference**

To map Sanger sequencing reads to a reference use the *Tools-> Sanger data analysis-> Map reads to reference* main menu item. The following dialog appears:

Map Sanger Reads to Reference	:	8
n <b>put data</b> Reference		
Input a file with a reference seque	nce	
Reads		
		Add Remove
ettings		
Trimming quality threshold	30	<u>^</u>
Mapping min identity	80%	<u>^</u>
esult alignment		
///./UGENE_Data/sanger_reads	_alignment.ug	enedb
✓ Add to project		
Help		🔇 Cancel 🖌 Align

It is required to input:

*Reference sequence* — a file with a single DNA sequence of any supported file format (e.g. FASTA or GenBank). This parameter is mandatory.

*Reads* — a set of files in \*.ab1 or \*.scf format with the Sanger sequencing data to be mapped to the reference sequence. It is required to add at least one read. A read orientation (forward or reverse) will be automatically detected during the mapping.

During the mapping task execution these data are processed as follows. First of all, to enhance the further mapping sensitivity, the low-qualit y ends of the reads are trimmed off. Than the reads are mapped using a UGENE original method that works in two steps: rough mapping of a read using BLAST+ and enhancement of the read alignment with the Smith-Waterman algorithm. After that the percentage similarity of the two sequences is calculated, that is the number of all edit operations (i.e., insertions, deletions and substitutions) required to transform a read sequence into the corresponding region of the reference sequence. Reads that have low similarity with the reference sequence are filtered out.

One can configure the following parameters of the mapping task

*Trimming quality threshold* — all bases at the ends of the reads with quality lower than the specified value are trimmed. To skip the trimming, set this value to zero.

Mapping min similarity — all reads mapped to the reference with lower percentage similarity than the specified value are filtered out.

Read name in result alignment — reads in the result alignment can be named either by names of the sequences in the input files or by the input files names. Set this value to "File name", for example, if the sequences in the input \*.ab1 files have the same name, this will help in distinguishing of the reads in the result alignment.

The result alignment is stored in a native UGENEDB format. One can set up the file location and name in the *Result alignment* field. Note that thereafter it is also possible to export data to standard alignment formats without chromatograms such as FASTA, ClustalW, etc.

To initiate the mapping task execution click the *Map* button in the dialog. Note that when the task is finished, the task statistics can be found in a report, available on clicking the corresponding notification.

### **Alignment Statistics**

To show statistics use the General tab of the Options Panel:

General		
✓ Alignment info		- 88
Length: Sequence number:	11943 13	<u></u>
		806

### **Alignment Appearance**

To perform zoom operations use the corresponding buttons on the editor toolbar.

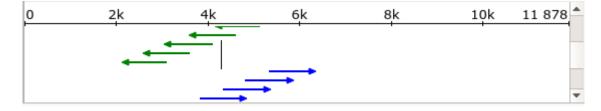


You can zoom in and zoom out by the corresponding toolbar buttons. To reset zoom use the *Reset zoom* button. To increase/decrease peaks height use the corresponding buttons.

You can change font by Appearance->Change characters font context menu item or by the main Actions->Appearance menu item.

### **Overview and Show and Hide Chromatogram**

The chromatogram overview is shown automatically in the Sanger Reads Editor. To close the overview click on the Show overview toolbar button. Drag and drop vertical cursor to change the current position.



To show/hide chromatogram click on the corresponding toolbar button:

### Sanger Reads Consensus

Each base of a consensus sequence is calculated as a function of the corresponding column bases. The Sanger Reads Editor allows switching between different consensus modes: Simple extended and Strict.

The Simple extended algorithm selects the best character from the extended DNA alphabet. Only bases with frequences which are greater than a threshold value are taken into account.

The Strict algorithm returns gap character ('-') if symbol frequency in a column is lower than threshold specified.

To switch the consensus mode go to the Consensus tab of the Options Panel:

Consensus	
Consensus type:	120
Strict	252
Threshold:	8.0
100% 🜲	
▼ Export consensus	
Export to file	
Aligned reads_consensus_3.txt	
File format	
Plain text 👻	
Keep gaps	
Export	

• Export Chromatogram Consensus

Export Chromatogram Consensus

To export consensus sequence use the Consensus tab of the Options Panel:

Consensus	
▼ Consensus mode	
Consensus type:	
Strict 💌	•
Threshold:	
<ul> <li>Export consensus</li> </ul>	
Export to file	
Aligned reads_consensus_3.txt	
File format	
Plain text 💌	
Keep gaps	
Export	

The following parameters are available:

Export to file - here you need to select path for the output file.

File format - format for the output file.

Keep gaps - set this parameter if the result consensus must keep gaps.

When you click on the Export button the consensus sequence will be exported into selected output file.

### **Navigation in Sanger Reads Alignment**

You may jump to previous/next ambiguous character in the reads alignment and edit the characters.

After you finish the editing, use the navigation by previous/next variations to find differences between the reference sequence and the consensus.

Use the main Actions-> Navigation menu, context menu or the following toolbar buttons:



### **Editing Sanger Reads**

Select the Edit submenu in the Sanger Reads Editor context menu.

The actions available from this menu are described below:

- Inserting Character
- Replacing Character and Gap
- Removing Character and Gap
- Inserting Gap
- Removing Gap at the Left
- Removing All Columns of Gaps •
- Trimming Left End
- Trimming Right End
- ٠ **Renaming Read**
- Removing Read ٠
- Undo and Redo

### **Unipro UGENE Manual, Version 1.30**

#### **Inserting Character**

To insert a nucleotide into a read in the Sanger Reads Editor, select a nucleotide or a gap in the corresponding location in the Chromatogram Area and switch to the editing mode by pressing Shift+I, selecting Edit->Insert character/gap in the context menu, or selecting Actions->Edit->Insert character/gap in the main menu. While the cursor is blinking, press a character on the keyboard that you want to insert. After that the editing mode is switched off.

Note that if you press the '-' character on the keyboard while being in the editing mode, a gap is inserted. The other way to insert a gap is by pressing *Space*, see the Inserting Gap chapter.

You can always cancel the changes by pressing *Ctrl+Z*. **Replacing Character and Gap** 

To replace read symbol or gap in the Sanger Reads Editor select the nucleotide and press Shift+R or Edit->Replace character/gap context menu for switching the edit mode on.

While the selection is blinking, press a keyboard character that you want to replace.

After that the edit mode is switched off. For replace another symbol press Shift+R again.

You always can cancel the changes pressing *Ctrl+Z*. **Removing Character and Gap** 

To delete read symbol or gap in the Sanger Reads Editor select the nucleotide and press Del or Edit->Remove character/gap context menu for switching the edit mode on.

While the selection is blinking, press a delete button.

After that the edit mode is switched off. For remove another symbol press Del again.

You always can cancel the changes pressing Ctrl+Z.

**Inserting Gap** 

To insert a gap into a read in the Sanger Reads Editor select a character or a gap in the corresponding location in the Chromatogram Area a nd press Space, select Edit->Insert gap in the context item, or Actions->Edit->Insert gap in the main menu.

Thus, by holding the *Space* key, one can shift a Sanger read to the right. To shift the read to the left hold the *Backspace* key, see the Removi ng Gap at the Left chapter.

You can always cancel the changes by pressing *Ctrl+Z*. **Removing Gap at the Left** 

To remove a gap in the Sanger Reads Editor select the next character or gap after the gap you want to remove and press Backspace. You can also select Edit->Remove gap at the left in the context menu or Actions->Edit->Remove gap at the left.

If there is a region of gaps or it is the beginning of a read, by holding the *Backspace* key, one can shift a Sanger read to the left. To shift the read to the right hold the *Space* key, see the Inserting Gap chapter.

You can always cancel the changes by pressing Ctrl+Z.

#### **Removing All Columns of Gaps**

Use the *Shift+Del* or *Edit Remove all columns of gaps* item in the *Actions* main menu or in the context menu to remove all columns of gaps from the alignment. This can be useful, for example, when an insertion in a read turned out not to be real, it was replaced by a gap, and a column of gaps was left in the alignment.

#### **Trimming Left End**

To trim left end in the Sanger Reads Editor select the next symbol after trimming part and use Ctrl+Shift+Backspace or Edit->Trim left end context menu for switching the edit mode on.

After that the edit mode is switched off.

You always can cancel the changes pressing Ctrl+Z.

**Trimming Right End** 

To trim right end in the Sanger Reads Editor select the previous symbol before trimming part and use Ctrl+Shift+Del or Edit->Trim right end c ontext menu for switching the edit mode on.

After that the edit mode is switched off.

You always can cancel the changes pressing *Ctrl+Z*. **Renaming Read** 

To rename read use F2 or Rename read context menu.

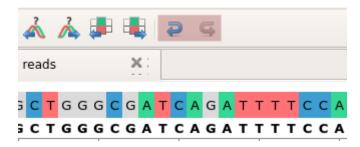
#### **Removing Read**

To remove read use *Delete* keyboard button or *Remove read* context menu. **Undo and Redo** 

The editor tracks all modifications.

When a modification happens the current state is being recorded.

You can apply any previous state and redo the modifications using the corresponding buttons on the toolbar:



#### **Exporting Alignment without Chromatograms**

To export alignment without chromatogram use the Actions->Alignment->Export alignment without chromatogram main menu or the same context menu item. The following dialog will appear:

Export Alignment without Chromatograms	s 😣
Export to file ///UGENE_Data/reference_sanger_r	eads_alignment_10.aln
File format CLUSTALW	•
<ul> <li>Add document to project</li> </ul>	
✓ Include reference sequence	
	Export Cancel Help

Here it is possible to specify the result file location, to select a sequence file format, optionally include consensus and add the created document to the current project.

# **Assembly Browser**

The UGENE Assembly Browser project started in 2010 was inspired by Illumina iDEA Challenge 2011 and multiple requests from UGENE users. The main goal of the Assembly Browser is to let a user visualize and efficiently browse large next generation sequence assemblies.

Currently supported formats are SAM (Sequence Alignment/Map) and BAM, which is a binary version of the SAM format. Both formats are produced by SAMtools and described in the following specification: SAMtools. Support of other formats is also planned, so please send us a request if you're interested in a certain format.

To browse an assembly data in UGENE, a BAM or SAM file should be imported to a UGENE database file. After that you can convert the UGENE database file into a SAM file. The import to a UGENE database file has both advantages and disadvantages. The disadvantages are that the import may take time for a large file and there should be enough disk space to store the database file.

On the other hand, this allows one to overview the whole assembly and navigate in it rather rapidly. In addition, during the import you can select contigs to be imported from the BAM/SAM file. So, there is no need to import the whole file if you're going to work only with some contigs. Note that in the future there are plans to support the other approach as well, namely, when a BAM/SAM file is opened directly.

The Assembly Browser has been tested on different BAM/SAM files from the 1000 Genomes Project and other sources.

Read the documentation below to learn more about the Assembly Browser features.

- Import BAM and SAM Files
  - Browsing and Zooming Assembly
    - Opening Assembler Browser Window
    - Assembly Browser Window
    - Assembly Browser Window Components
    - Reads Area Description
    - Assembly Overview Description
    - Ruler and Coverage Graph Description
    - Go to Position in Assembly
    - Using Bookmarks for Navigation in Assembly Data
- Getting Information About Read
  - Short Reads Vizualization
    - Reads HighlightingReads Shadowing
  - Reads Shadowing
- Associating Reference Sequence
- Associating Variations
- Consensus Sequence
- Exporting
  - Exporting Reads
  - Exporting Visible Reads
    Exporting Coverage
  - Exporting Coverage
  - Exporting Consensus
  - Exporting Consensus Variations
  - Exporting Assembly as Image
  - Exporting Assembly Region
- Options Panel in Assembly Browser
  - Navigation in Assembly Browser
    - Assembly Statistics
    - Assembly Browser Settings
- Assembly Browser Hotkeys
  - Assembly Overview Hotkeys
    - Reads Area Hotkeys

## Import BAM and SAM Files

To start working with an assembly import it to the UGENE database file. To do this, open the assembly file.

For assembly file without header you need to choose a referenece sequence:

U Import SA	M File
Source URL:	C:/work/ugene/data/samples/Assembly/chrM.sam Info
Reference:	
_	file does not contain the header. Please, choose the reference sequence.
Destination U	RL: C:/work/ugene/data/samples/Assembly/chrM.sam.ugenedb
	Import Cancel Help

Select the referense sequence and click Import button.

For other assembly files the following dialog appears:

Import BAM File		2 ×
Source URL: C:/work	/ugene/data/samples/Ass	embly/chrM.sorted.bam Info
Assembly name	Length	URI
1 🔽 chrM	16 571	
Select All Deselect		
Select All Deselect		

The Source URL field in the dialog specifies the file to import. The Info button nearby can be used to obtain additional information about the file.

There is a list of contigs below the Source URL. Check the contigs that you want to import to the database. You can use the Select All, Desel ect All and Invert Selection buttons to manage the selection.

The Destination URL field specifies the output database file.

If you check the *Import unmapped reads*, then all unmapped reads in the assembly (i.e. read with the unmapped flag or without CIGAR) are imported. Note, however, that they are not vizualized in the current UGENE version.

To start the import, click the *Import* button in the dialog. You can see the progress of the import in the *Task View*. To export a UGENE database file into the SAM format, select the *Actions Export assembly to SAM format* item in the main menu.

### Browsing and Zooming Assembly

- Opening Assembler Browser Window
- Assembly Browser Window
- Assembly Browser Window Components
- Reads Area Description

- Assembly Overview Description
- Ruler and Coverage Graph Description
- Go to Position in Assembly
- Using Bookmarks for Navigation in Assembly Data

#### **Opening Assembler Browser Window**

An imported assembly added to the project is shown in the Project View as follows:

Project	×
Name filter	T,
Objects	
Klebsislla.sort.bam.ugenedb	
🚱 [as] pkF70	
🔗 [as] pkf140	

Each [as] object corresponds to an imported contig. When you double-click on an [as] object a new Assembly Browser window with the assembly data is opened. A window for the first assembly object in the list is opened automatically after the import.

#### **Assembly Browser Window**

The opened window contains the list of well-covered regions of the assembly:

Δų.	pitt.	a, ley	ر السلم								4,		At Her	ny.	l, i l.	1.01	1	
0 to 16	571 (16	571 bp	)					1				1.1	1.1	1 to 1	6 571	(16 571	bp)	
																		Ь
· ·	- 1			1	1	1	1	1	1			1	1	-			7	٠
									1	0k				1	15 385	C 239		200
																	_	
			Z	loom in	to see	the reads	s or cho	ose one	of the	well-cove	ered re	gions:						
							Posit	tion Co	verad	<b>_</b>								
						1	6 950		330	-								
						2	7 950		321									
						3	16 25	-	320									
						4	7 850		316									
						5	16 35	-	312									
						6	13 45		308									
						7	12 95		307									
						8	11 75		301									
						9	12 05		300									
						10	13 25		296									
								_										
				_														
			TIP:	Page	up/Pa	ige dow	n — Mo	ve one	page up	)/down ir	the Re	eads Are	ea					
																	Ŧ	
																	÷.	

Note that for large assemblies it may take some time to calculate the overview and the well-covered regions.

To see the reads, either select a region from the list or zoom in, for example, by clicking the link above the well-covered regions or by rotating the mouse wheel.

You can also use the hotkeys. Tips about hotkeys are shown under the list of well-covered regions. To learn about available hotkeys refer to *Assembly Browser Hotkeys*.

### **Assembly Browser Window Components**

An Assembly Browser window consists of:

#### **Assembly Overview**

By default, shows the whole assembly overview. Can be resized to provide an overview of an assembly part.

#### **Reference Area**

Shows the reference sequence.

#### **Consensus Area**

Shows the consensus sequence.

#### Ruler

Shows the coordinates in the Reads Area.

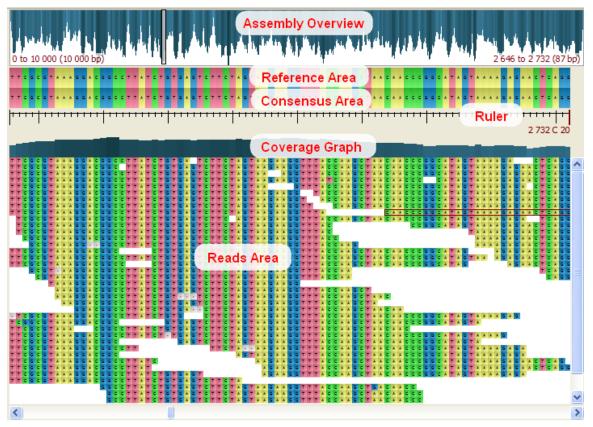
#### **Reads Area**

Displays the reads.

#### **Coverage Graph**

Shows the coverage of the Reads Area.

See the example below:



### **Reads Area Description**

The Reads Area provides a visualization of reads of an assembly part. To zoom in or zoom out, rotate the mouse wheel.

To perform zooming you can also use the Zoom In and Zoom Out buttons on the toolbar or the Actions Zoom In and Actions Zoom Out item s in the main menu.

Also, when you double-click on a read it is zoomed in and moved to the center of the window.

By dragging the mouse while holding the left mouse button you can navigate in the Reads Area.

To navigate long distances in the Reads Area use the Assembly Overview described below.

Other ways to navigate in the assembly are:

- Use the horizontal and vertical scroll bars of the Reads Area
- · Go to a specified position in an assembly

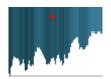
To learn about available hotkeys refer to Assembly Browser Hotkeys.

By default, assembly rendering is optimized while scrolling. While you are moving across an assembly, it shows the assembly in gray color, but when you stop it shows the assembly in different colors. To disable this option uncheck the *Optimize the rendering while scrolling* item in the context menu of the *Reads Area* or *Optimize scrolling* item on the *Assembly Browser Settings* tab of the *Options Panel*.

#### Assembly Overview Description

The Assembly Overview shows a coverage overview of the assembly. The longer the depth of a line in the overview and the deeper the color, the more reads are located in this region.

To open a region of the assembly in the *Reads Area* click on it in the Assembly Overview. On the overview, the selected region is displayed either as a gray rectangle, a red cross or a red rectangle. For example:

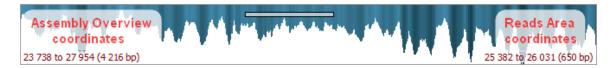


If you hold Shift and select a region on the overview, the overview is zoomed to the selection.

Note that when the Assembly Overview is in focus and you use either the zoom buttons on the toolbar, the zoom items in the *Actions* main menu, or a mouse wheel, the Reads Area is resized appropriately.

The Assembly Overview can also be resized. To zoom in the overview, select either the Zoom in or the Zoom in 100x item in the Assembly Overview context menu. You can scroll the resized overview by dragging the mouse while pressing down the mouse wheel. To zoom out the overview, select the Zoom out item in the context menu. The Restore global overview item in the context menu restores the default overview size when the whole contig overview is shown.

Notice that the Assembly Overview shows the coordinates of the assembly areas visible in the Reads Area and in the Assembly Overview:



To scroll the resized overview, drag the mouse while pressing down the mouse wheel.

To learn about available hotkeys refer to Assembly Browser Hotkeys.

#### **Ruler and Coverage Graph Description**

The *Ruler* shows the coordinates in the *Reads Area*. When you move the mouse cursor in the Reads Area the coordinate of the selected location with the coverage of reads is shown on the ruler in dark red. The Coverage Graph shows the exact coverage of the sequence at each position. For example on the image below the coordinate is 9168 and the coverage of reads is 251.

			· · · · <mark>· · · · · · · · · · · · · · · </mark>		
8.9k	9k	9.1k	9 168 C251	9.3k	

To show/hide the coordinates on the ruler you can click the following button on the toolbar:

111 5k

To show/hide the coverage on the ruler you can click the following button on the toolbar:



Alternatively, you can use the Show coordinates and Show coverage under cursor check boxes located on the Assembly Browser Settings ta b of the Options Panel.

### Go to Position in Assembly

To go to the required position in an assembly use the following field located on the Assembly Browser toolbar.



Input the location and click the Go! button. A similar Go! field is also available on the Navigation tab of the Options Panel.

### Using Bookmarks for Navigation in Assembly Data

Use bookmarks to save and restore visual state of an assembly, for example, position in the assembly, zoom scale, etc.

## **Getting Information About Read**

A read displayed in the *Reads Area* consists of the bases (A, C, G, T). It may also contain the N character that stays for an ambigous base. Depending on the value of the *Cigar* parameter, the read can be shown partially or gaps can be inserted inside the read (see below).

By default when a read is hovered over in the Reads Area a hint appears:

C	С	A	G	G	Т	G	Т	A	С	С	С	Т	G	Т	А	Т	G	C	С	G	G	Т	G	G	Т	G	A	Т	G	С	С	С			С	т	G	С	А	G
	С	A	G	G	т	G	Т	A	С	С	С	т	G	т	А	Т	G	С	С	G	G	т	G	G	Т	G	A	Т	G	С	С	С	G	С	С			С	A	G
C	С	A	G	G	т	G	T	۵	c	c	C	Т	G	Т	Δ	Т	G	C.	C.	G	G	Т	G	G	Т	G	Δ	Т	G	c	c	C	G	C.	ç	т	G	С	A	G
С	С	A	G	G	т	G	т	13	60		2	FC	30	)H	46	A	AX	X:1	8:4	11:	46	58:	19	94	1		A								¢	т	G	С	A	G
С	С	А	G	G	т	G						928															A								¢	т			А	G
С	С	А	G	G	т	G	т					35			A												A								¢	т	G		А	G
C	С	А	G	G	т	G	т					M			А												A								¢	т	G	С		
С	С	А	G	G	т	G	т					cor		em	ê	Ŧ											A								¢	т	G	С	А	G
С	с	A	G	G	т	G											éc	fre	TA	60	4	TO	TA	TO	ē	-C	4	cīc	fe	NT.	6		-C	ćc	4	т	G	С	A	
С	с	A	G	G	т	G	т	N	2	ų:	9E	qu	5	-00	7	-	90	100	1	i Ci	ż	10	14	10	-u	.9	2	90	2	2	G	20		ý.	4	т	G	С	A	G
С	С	A	G	G	т	G	Т	A	С	С	С	т	G	т	A	Т	G	С	С	G	G	Т	G	G	т	G	A	Т	G	с	С	С	G	С	С	т	G	С	A	G
		А	G	G	т	G	т	A	С	С	С	т	G	т	А	т	G	С	С	G	G	т	G	G	т	G	А	т	G	с	с	С	G	с	с	т				

To disable this behaviour click the following button on the toolbar:

Or uncheck the Show pop-up hint check box on the Assembly Browser Settings tab of the Options Panel.

The hint shows the following information about the read:

- Read name
- Location
- Length
- Cigar
- Strand
- Read sequence

The operations in the Cigar parameter are described as follows:

- **M** Alignment match (can be a sequence match or mismatch).
- I Insertion to the reference. Skipped when the read is aligned to the reference, i.e. it is not shown in the Reads Area, but is present in the read sequence.
- D Deletion from the reference. Gaps are inserted to the read when the read is aligned to the reference. For example:

	C /	N A	1 6	a (a	6	A	A	6	А	6	А	C,	C.	6	A	1	А	А	А	1	1	А	C.																1	. 6	A		μ	A	6	А		1	1		1
ſ	C /	A	A G	G	G	A	A	G	А	G	A	С	С	G	Α	Т	А	-	-	-	т	A	С	С	С	G	C	G	С	т	G	Т	С	G	A	А	С	Т	G (	: G	i A	T	A	A	G	i A	т	Т	т	Т	Т
1	C /	A	A G	G G	G	A	A	G	A	F	٨	_	_	_	_	_	٨	۸	٨	т	-	۸	c	c	c	G	c	G	c	т	G	т	c	G	٨	۸	c	т	<u> </u>	- 6	· ^	1	- 0	A	G	i A	Т	Т	Т	T	т
I	C /	A	A G	i G	G	A	A	G	А	9	23	Æ	xa	m	ple	2 5	e	que	en	CE	e F	١.	5	2(:	19	,2	2);	S	eo	on	da	iry	5	eq	ue	n	e	58	69								_				
	C /	A	A G	a G	G	A	A			F	ro	m	58	35	to	<b>5</b>	88	61	Ro	w	: 1	3.		C		G	S	G		т		т	C,											A	G	A	т	т	т	Т	Т
	C /	A	A G	i G	G	A	A	G	А	L	en	gt	h:	52	Д	Т		-	_	-	_	-	-	-			D	e	et	10	n	IS	S	10	W	1	as	ga	ap	<b>S</b> G				A	G	A	т	т	т	т	Т
	C /	A	A G	i G	G	A	A	G	А	C	iq	ar	q	вМ	13	D	31	M																													_				
	С			G	G	A	A	G	А	s	tra	ân	d:	dir	éc	f																												A	G	A	т	т	т	т	
	C /	A	A G	i G	G	A	A	G	А	R	èz	d	SE	- n	iei	nc	e:	C.		GG	GA	AA	GA	G	AC	CG	AT	AT	AC	CC	G	CG	СТ	GT	CG	A	۱C	TG	G/	ATA	AC	A	ш	т							
	C /	A	A G	G	G	А	A	G	А	5	À	ĉ	e	9	A	-	A	A	A	Ŧ	-	A	ć	Ċ	C	G	ć	9	C	-		-	Ċ,	G	À	A	ĉ	T	-		-		-	-	G	A					
										-		-	-	-		-				-	-		-	-	-	-	-	-	-					_								_			_						

- N Skipped region from the reference. Behaves as D, but has a different biological meaning: for mRNA-to-genome alignment it represents an intron.
- **S** Soft clipping (clipped sequences are present in the read sequence, i.e. behaves as **I**).
- H Hard clipping (clipped sequences are not present in the read sequence).
- P Padding (silent deletion from padded reference).
- = Exact match to the reference.
- **x** Reference sequence mismatch.

To copy the information about the read to the clipboard, select the *Copy read information to clipboard* item in the Reads Area context menu. Now you can paste it in any text editor.

To copy the current position of the read select the Copy current position to clipboard item in the Reads Area context menu.

## **Short Reads Vizualization**

There are various modes of reads highlighting and shadowing.

- Reads Highlighting
- Reads Shadowing

### **Reads Highlighting**

To apply a reads highlighting mode, select it in the *Reads highlighting* menu of the *Reads Area* context menu or on the *Assembly Browser Settings* tab of the *Options Panel*. The following modes are available:

- *Nucleotide* shows all nucleotides in different colors. It is used by default.
- Difference highlights gaps and nucleotides that differ from the reference sequence. You should add a reference first for correct displaying of this highlighting.

A	c	c	G	т	т	c	c	G	т	G	G	c	A	A	A	G	c	A	A	A	A	G	т	т	c	A	A	A	A	т	c	A	c	c	A	A	c	т	G	G	c	c	c
+	+	†	+	+	+	+	+	9	0	19		1	40	7	+	+	+	+	+	+	+	1	+	+	+	+	+	+	+	+	+	†	+	+	+	+	+	+	+	+	+	†	t
A	C	C	G	Т	Т	C	C	G	Т	G	G	C	A	A	A	G	C	A	A	A	A	G	Т	Т	C	A	A	A	A	т	A	A	C	C	A	A	C	Т	G	G	C	C	c
A	C	C	G	T	Т	C	C	G	Т	G	G	C	A	A	A	G	C	A	A	A	A	G	Т	T	C	A	A	A	A	T	C	A	C	C	A	A	C	т	G	G	C	C	c
A	C	C	G	т	Т	C	C	G	Т	G	G	C	A	A	A	G	С	A	A	A	A	G	Т	T	C	A	A	A	A	т	С	A	С	С	A	A	C	G	G	G	C	C	c
A	C	С	G	Т	т	C	C	G	Т	G	G	C	A	A	A	G	т	A	A	A	A	G	Т	Т	C	A	A	A	A	т	С	A	C	С	A	A	С	т	G	G	C	C	C
A	C	С	A	т	т	C	C	G	Т	G	G	C	A	A	A	G	С	A	A	A	A	G	Т	Т	C	A	G	A	A	т	С	A	С	С	A	A	С	т	G	G	C	C	C
A	C	С	A	т	т	C	С	G	т	G	G	C	A	A	A	G	С	A	A	A	A	G	Т	Т	C	A	A	A	A	т	A	A	C.	C.	A	A	C	T	G	G	C.	C	C
A	С	С	A	т	Т	С	C	G	Т	G	G	C	A	A	A	G	С	A	A	A	A	G	Т	Т	С	A	G	A	A	т	C	A	С	С	A	A	С	т	G	G	C.	C	C
A	C	С	G	Т	т	C	С	G	Т	G	G	C	A	A	A	G	С	A	A	A	A	G	Т	Т	C	A	A	A	A	т	A	A	C	С	A	A	С	Т	G	G	C.	C	C
A	C	С	G	Т	т	C	С	G	Т	G	G	C	A	A	A	G	С	A	A	A	A	G	Т	Т	C	A	A	A	A	т	C	A	С	С	A	A	С	Т	G	G	C.	C	C
A	C	C	G	Т	T	C	C	G	T	G	G	C	A	A	A	G	C	A	A	A	A	G	Т	T	C	A	A	A	A	т	A	A	С	С	A	A	C	T	G	G	С	C	C
A	C	C	G	Т	T	С	C	G	T	G	G	C	A	A	A	G	C	A	A	A	A	G	Т	T	С	A	G	A	A	т	С	A	С	С	A	A	C	T	G	G	т	C	C
A	C	C	G	Т	T	С	C	G	T	G	G	C	A	A	A	G	C	A	A	A	A	G	Т	T	С	A	G	A	A	т	С	A	С	С	A	A	C	T	G	G	т	C	C
A	C	C	G	Т	T	C	C	G	T	G	G	C	A	A	A	G	C	A	A	A	A	G	Т	T	C	A	G	A	A	т	С	A	C	C	A	A	C	T	G	G	C	C	C
A	C	C	G	С	T	C	C	G	T	G	G	C	A	A	A	G	C	A	A	A	A	G	T	T	C	A	A	A	A	T	C	A	C.	C	A	A	C	T	G	G	C	C	C

• Strand direction — highlights reads located on the direct strand in blue and reads on the complement strand in green.

A	G	A	A	т	G	С	т	G	A	т	т	т	A	С	A	С	т	т	т	т	G	A	G	G	A	т	A	A	т	С	С	G	G	A	т	A	A	G	т	A	A	С	G	С	т	G	С	A
A	G	A	A	т	G	С	т	G	A	т	т	т	A	С	A	С	т	т	т	т	G	A	G	G	A	т	A	A	т	С	С	G	G	A	т			G	т	A	A	с	G	с	т	G	с	A
A	G	A	A	т	G	с	т	G	A	т	т	т	A	с	A	С	т	т	т	т	G	A	G	т	A	т	A	A	т	с	с	G	G	A	т	А		G	т	A	A	A	G	с	т	G	с	A
A	G	A	A	т	G	С	т	G	A	т	т	т	A	С	A	С	т	т	т	т	G	A	G	G	A	т	A	A	т	С	С	G	G	A	т	A	A		т	A	А	с	G	с	т	G	с	A
A	G	A	A	т	G	с	т	G	A	т	т	т	A	С	A	С	т	т	т	т	G	A	G	G	А	т	A	А	т	с	с	G	G	A	т	A	A	G	т	A	A	с	G	С	т	G	с	A
A	G	A	A	т	G	с	т	G	А	т	т	т	А	с	A	С	т		т	т	G	A	G	G	A	т	A	А	т	С	с	G	G	A	т	А	A	G	т	A	А	с	G	с	т	G	с	A
A	G	A	A	т	G	С	т	G	А	т	т	т	А	с	A	С	т	т	т	т	G	A	G	G	A	т	A	А	т	С	С	G	G	A	т	A	A	G	т	A	А	С	G	с	т	G	с	A
A	G	A	A	т	G	с	т	G	А	т	т	т	А	с	A	С	т	т	т	т	G	А	G	G	А	т	A	А	т	с	с	G	G	A	т	А	A	G	т	A	А	с	G	с	т	G	с	A
A	G	A	A	т	G	С	т	G	А	т	т	т	A	С	A	С	т	т	т	т	G		G	G	A	т	A	A	т	С	с	G	G	A	т	A	A	G	т	A	A	с	G	с	т	G	с	A
A	G	A	A	т	G	С	т	G	A	т	т	т	A	С	A	С	т	т	т	т	G	A			A	т	A	A	т	С	С	G	G	A	т	A	A	G	т	A	А	с	G	с	т	G	с	A
A	G	A	A	т	G	С	т	G	A	т	т	т	A	С	A	С	т	т	т	т	G	A	G	G	А	т	A	А	т	с	с	G	G	A	т	A	A	G	т	A	А	с	G	с	т	G	с	A
A	G	A	A	т	G	с	т	G	A	т	т	т	A	С	A	С	т	т	т	т	G	A	G	G	A	т	A	A	т	С	т	G	G	A	т	A	A	G	т	A	A	с	G	С	т	G	с	A
A	G	A	A	т	G	С	т	G	A	т	т	т	A	С	A	С	т	т	т	т	G	A	G	G	A	т	A	A	т	С	С	G	G	A	т	A	A	G	т	A	A	с	G	С	т	G	с	A
		A	A	т	G	с	т	G	A	т	т	т	A	с	A	C	т	т	т	т	G	A	G	G	A	т	A	A	т	с	С	G	G	A	т	A			т	A	A	с	G	с	т	G	с	A

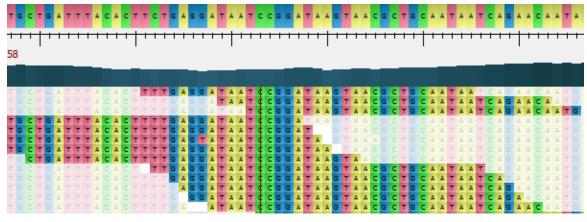
· Paired reads — highlights all paired reads in green. Note that the information about the pair is shown in the hint.

								100	100							100	10					100	100				100	1000	100	100	1000				100				100											
- C	T 6	•	6	т	τ.	c (	•	6	٨	Τ.	٩,	٩.	۰.	e (	1	•	9	•	e (	τ 6		19	•	٩.	٩.	τ.	6 1	-		6	6	Τ.	6 6	-	٩.	τ.	<u>د ۱</u>	r 6	т.	C	G 7	6	т.	т	c (	÷ •	Τŝ	2 G	T C	4
C 1	T G	i A	G	т	т	C (	: A	G	A	т	с,	с,	A	G   1	r (	A.	C.	A	G	тс	Τ.	С	A	C.	c.	т	G 1	- C	Т	G	G	т	6 6	6	с.	т	C 1	r e	т	C	C 1	r 6	π.	G	C (	- C	τ¢	6 G	T C	l
C 1	T G	i A	G	G	Т	A C	: A	G	A	т	C.	C.	A	G 1	r (	A	C	A	G	τc	T	C	A	C.	C.	т	G 1	- C	T	G	G	T	GC	6 G	C.	т	C 1	r 6	т	C (	с т	r c	т	т	C (	÷ 6	ΤĢ	5 G	T C	f,
C 1	T G	i A	G	т	т	C (	: A	G	A	т	C.	C.	A	G _ 1	r . 7	4	C.	4	6	τ. (		10	4	c.	c.	τ.	6	1.0	. 7	6	6	7	<u>a a</u>	1.0	0	τ.	c -	- 6	T	C	c 7	- 6	T	T	6 6	; C	TR	5 6	TC	1
C	T G	A	G	T	T	<b>c</b> (	A	G	A	T	C.	с.	A	s (	50	R	<b>n</b> 1	41	87	0	11	no	2	58	24																					C.	TR	5 G	TC	1
S 1	r c	- A	G	τ.	τ.	c (	: A	G	A	τ.	ς.	۰.	A	c 17		_	100		10.00			1.7	-				6 7		T	G			6 6													C	TC	a a	TC	1
C	t la		G	-	-	c (		G		-	C.	C.		a I	n	on	1.1	15	23	82	86	5	to	) ]	15	23	82	90	)0	R	٥V	V:	84	6												0	TO	a a	TO	ł.
100			6	-	-			a	-	-	-	-		a 1	-	ng	+1		20	-																										1	-		7.0	f,
100			-	-	-			0	-	-	-	2	-																																		-			l
			2	-	-			12					-	10	. ic	ja	r:	36	М																											-				l
-				4	2	2.	10		0	4	2	2	2			ra																																		1
-			-	4	2		1	1	0	4	2	5	2		_																															1			100	1
E .	T G	•		т	т	6 6		-	^	т	۰.	٩,	۰.	۰.	Re	ar	t,	ie	aı	ie	nc	e:	: T	A	CA.	G	AT	CC	A.	GT	C	AC	AC	T	СТ	C	٩C	CT	GI	IC.	ΤG	G	ΓĠ	GC	C.	-	TS		TC	l
- C	7 6	•	6	Ŧ	Ŧ	c (		6	A	т	۰.	٩,	Α.	91		-	15	1				-	6		-		G. 1				-				-	-	-		-	-				_	1	6	TG	2 6	TC	ł
C	T G	A	G	Т	Т	<b>c</b> (		G	A	т	ς.	6	A	G																																6	TG	5 6	TC	ł.
C 1	T G	i A	G	т	т	c (	- A	G	A	т	c,	c,	A	٩ I	Da	ire	ed.	in the	éa	d	Т																								C (	С.	τ¢	5 G	T C	l
C 1	T G	i A	G	т	т	c (	: A	A	A	т	c,	c,	A	6								1	Α.	6	ς.																				C (	C	TG	5 G	TC	ł
C 1	T G	i A	G	т	т	C (	i A	G	A	т	C.	c.	A	6 B	ы	R	01	4	62	υ.	10	99	8	58	54																				C (	C	TR	5 G	TO	l
C 1	T G	i A	G	т	т	c (	: A	G	A	т	C.	с.	A	6 J	-	on	1	15	23	83	04	18	te	-1	15	23	83	ns	27	P	ON N	÷.	20	4												C	TR	5 6	TC	l
C	T G	i C	G	т	т	<b>c</b> (	A	G	A	т	C	C.	A	C							•	2	1	6	5		22	~	-	1			-	G												C	TC	5 G	TC	l
C I	T Q	A	G	T	T	c 1	A	G	A	T	ς.	с.	A	c I	e	ng	t	1:	35	1																										C .	TO	5 G	TC	1
6	T 6		G	-	-	c (	. A	G	A	т	ς.	۰.	A	s (	-	ja	-	19	2.2	5	M.																									C .	7.0	5 G	тс	l
101			a		÷	c (		G				0												-																						10		a a	TIC	l
				-	-			6		-	2	1	4		Бt	ra	nc	: (	CO	mp	ble	me	en	t																						1	-		-	l
1				-	4	2.	1		2	4	2	2													TC	C	-	-	0	AT	NC	-	TC	CT	C	NO	Ar	-A	00	C	NC	A		CI	AG		-			l
				1	2				2		2					d	100	SC	чu	IC	IIC	e.		11	0	C	1		5	41	HC		10	CI.	91	40	MI	H			40	HI	30	4	10	-				ſ
				1	1				^	2		×.	•				-	~				-		-	-			-		100		100							100			10.05	100							ł

## **Reads Shadowing**

Various modes of column highlighting are available from the Reads shadowing item in the context menu of the Reads Area:

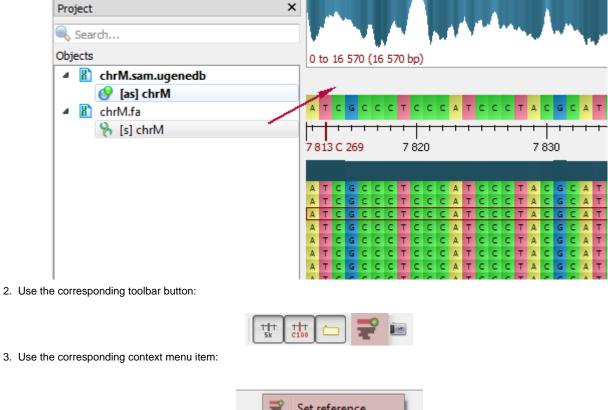
- Disabled highlights all columns of nucleotides.
- *Free* highlights all reads that intersect a given column. In this mode you can lock a position. Click the *Lock here* item in the context menu to do it. To return to a locked position, select the *Jump to locked base* item in the context menu.
- Centered highlights all reads that intersect the column in the center of the screen.



## **Associating Reference Sequence**

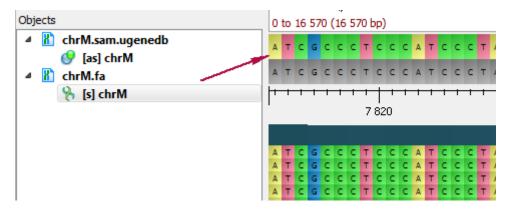
To associate a reference sequence with the assembly use one of the following ways:

1. Open the sequence (the sequence must be loaded) and drag it to the Assembly Reference Area:



₽.	Set	t ref	fere	nce			
	Un	ass	ocia	ate			
-	1	+	+	-	-	-	+
	<b>*</b>		-		Set reference Unassociate		

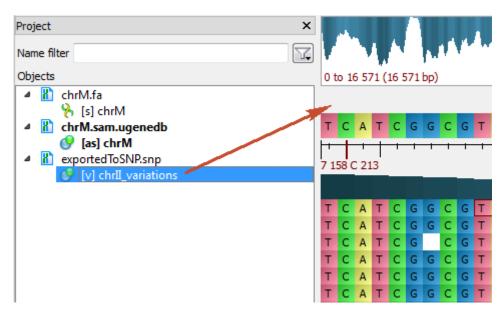
The reference sequence will appear in the Reference Area:



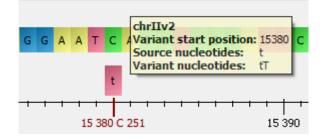
To remove the association, select the Unassociate item in the Reference Area context menu.

### **Associating Variations**

To associate variations with the assembly, open the sequence (the sequence must be loaded) and drag it to the Assembly Reference Area:



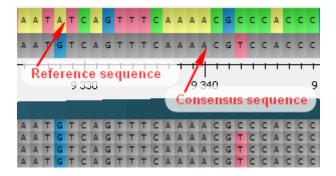
The variations will appear under the Consensus Sequence:



To remove the association, select the Remove track from the view item in the Variations Area context menu.

### **Consensus Sequence**

A consensus sequence can be found in the Consensus Area under a reference sequence. It refers to the most common nucleotide at a particular position.



To choose a consensus algorithm select the Consensus algorithm item either in the context menu of the Consensus Area, in the context menu of the Reads Area or on the Assembly Browser Settings tab of the Options Panel.

The following algorithms are currently available:

- Default shows the most common nucleotide at each position. When there is equal numbers of different nucleotides in a position, the consensus sequence resulting nucleotide is selected randomly from these nucleotides.
- SAMtools uses an algorithm from the SAMtools Text Alignment Viewer to build the consensus sequence. The algorithm takes into account quality values of reads and nucleotides and works with the extended nucleotide alphabet.

To leave only differences between the reference and the consensus sequences highlighted on the consensus sequence, select the Show difference from reference item in the context menu of the Consensus Area or the Difference from reference item on the Assembly Browser Settingstab of the Options Panel:

A	т	с	т	G	G	т	A	G	A	A	A	с	A	с	т	G	G	т	A	с	A	с	G	т	т	т	G
A	т	с	т	G	G	т	A	G	A	A	6	т	G	9	т	G	G	т	A	с	A	с	G	т	т	т	G
(	C	Dif	ffe	er	e	nç	e	s		+	-	1	+	+	+	1	 91	+ .0	+	+	+	+	+	+	+	+	+
A	Т	С	т	G	G	Т	А	G	А	А	A	т	G	С	т	G	G	Т	А	С	А	С	G	Т	т	т	G
А	т	с	т	G	G	т	А	G	A	A	A	т	G	с	т	G	G	т	А	С	А	С	G	т	т	т	G
A	т	с	т	G	G	т	А	G	A	A	A	т	G	с	т	G	G	т	А	С	А	С	G	т	т	т	G
A	т	с	т	G	G	т	A	G	A	A	A	т	G	С	т	G	G	т	A	С	A	С	G	т	т	т	G
A	т	с	т	G	G	т	A	G	A	A	A	т	G	С	т	G	G	т	A	С	A	С	G	т	т	т	G
Δ.	-	-	-	0	0	-	•	C.		٨		T	G	0	T	G	C	T	٨	0	٨	0	C	T	T	T	G

To export a *Consensus Sequence*, right-click on it in the *Consensus Area* and select the *Export Export consensus* item in the context menu. For more information about consensus exporting see *Exporting Consensus*.

### Exporting

- Exporting Reads
- Exporting Visible Reads
- Exporting Coverage
- Exporting Consensus
- Exporting Consensus Variations
- Exporting Assembly as Image
- Exporting Assembly Region

### **Exporting Reads**

To export a read, right-click on it in the Reads Area and select the Export Current read item in the context menu.

The *Export Reads* dialog appears:

U Export Rea	ds ? X	
Export to file		
File format	fasta 🔹	
Add to pro	ject	
	Export Cancel Help	

Select a file to export the read to and the file format. The read can be exported either to a FASTA or FASTQ file.

When the parameters are set click the *Export* button.

The read is exported to the file and if the Add to project check box has been checked it is added to the current project from where you can op en it.

### **Exporting Visible Reads**

To export all reads visible in the Reads Area select the Export Visible reads item in the Reads Area context menu.

The Export Reads dialog appears. The dialog is described in the Exporting Read section.

#### Exporting Coverage

To export a coverage of the assembly, select either the Export coverage item in the Consensus Area context menu.

The Export Coverage dialog appears:

Export the Assembly Cove	rage ? X
Export to:   Format: Per base	Compress the file
Threshold: 1 Additional options	<b>.</b>
Export coverage value	
Export bases quantity	
	Export Cancel Help

Select a file, threshold and format: *Histogram, Per base* or *Bedgraph.* Where *threshold* is the minimum coverage value to export. For *Per base* format the additional options are available: *Export coverage value* or *Export bases quantity* or both of them.

When all the parameters are set click the Export button.

### **Exporting Consensus**

To export a consensus sequence of the assembly, select either the *Export consensus* item in the *Consensus Area* context menu or the *Export Consensus* item in the *Reads Area* context menu.

The Export Consensus dialog appears:

U Export Consensus	;		8 ×
Export to file			
File format	FASTA		•
Sequence name	chrM_consensus		
Consensus algorithm	Default		•
Region Whole seque	nce 🔻	1 -	16571
V Keep gaps			
Add to project			
		Export Car	Help

Select a file and the file format. The consensus can be exported to a FASTA, FASTQ, GFF or GenBank file.

Modify, if required, the exported sequence name and choose the consensus algorithm.

The consensus is exported with gaps if the Keep gaps check box has been checked.

Also you can select the exporting region. It can be either a Whole sequence, a Visible region, or a Custom region.

When all the parameters are set click the Export button.

The consensus sequence is exported to the file and if the Add to project check box has been checked it is added to the current project and opened.

### **Exporting Consensus Variations**

To export a consensus sequence variations of the assembly, select the *Export consensus variations* item in the *Consensus Area* context menu.

The following dialog will appear:

Export Consensus	Variations		? ×
Export to file			
Mode	Variations		•
File format	SimpleSNP		•
Consensus algorithm	Default		•
Region Whole seque	nce 🔻	1 -	16571
Keep gaps			
Add to project			
		Export Cancel	Help

Select a file, mode and the file format. The following modes are available: Variations, Similar and All. Variations can be exported as to a SimpleSNP or VCFv4 file.

Modify, if required, the consensus algorithm.

The consensus is exported with gaps if the Keep gaps check box has been checked.

Also you can select the exporting region. It can be either a Whole sequence, a Visible region, or a Custom region.

When all the parameters are set click the *Export* button.

The consensus sequence is exported to the file and if the Add to project check box has been checked it is added to the current project and opened.

The Export consensus variations feature is available when the reference sequence is associated with assembly.

### Exporting Assembly as Image

To export the visible part of the assembly as an image, select either the Actions Export as image item in the main menu or the following button on the toolbar:

۲

The Export Image dialog appears:

∕∿

Export Image	8
Export to file	
File name	
Format png 🗘 Quality	
Help	Cancel Export

In the dialog you can select the image file name and its format (bmp, jpeg, png, etc.). For some file formats the Quality parameter also becomes available.

When the parameters are set click the OK button.

### **Exporting Assembly Region**

To export an assembly region use one of the following ways:

1. Use the Actions->Export assembly region main menu item:

U <u>F</u> ile	Actions	Settings	Tools	<u>W</u> indow	<u>H</u> elp	
	🔍 Zoom	in in				
	🔍 Zoom	out				L
ਦ <u>-</u>	🖮 Expo	rt as image				
oje	<sup>зам</sup> Ехро	rt assembly	to SAM	format		
d to	📳 Expo	rt assembly	region			L
сс	🚏 Set re	eference				l c
Ссс	Close	active view	,		Ctrl+W	
<u> </u>	+ + + +	+ + + + +	1 1 1			+

2. Use the corresponding toolbar button:



3. Use the corresponding context menu item:

Т	<u>C C T C C T T A C C A T C A A T C A A T T G G C C A C C A A T </u>	G G
Т	Copy read information to clipboard	S G
Т	· · · · · · · · · · · · · · · · · · ·	G G
Т	Copy current position to clipboard AATTGGCCACCAA	5 G
T	Export > Coverage	G
T	Reads highlighting > Consensus	G
T	Reads shadowing > Current read	G
	-	G
Т	Consensus algorithm > Visible reads as sequences	G
Т	✓ Optimize rendering when scrolling Assembly region	G
Т		G
Т	C C C T C C C T T A C C A T C A A A T C A A T T G G C C A C C A A T	G G
Т	C C C T C C C T T A C C A T C A A T C A A T T G G C C A C C A A T	G G

The following dialog will appear:

Export Assembly Region
Export to file
File format UGENE Database
Region Visible 🗘 7825 - 7925
Add to project
Cancel OK

Select a file, region and the file format.

### **Options Panel in Assembly Browser**

- Navigation in Assembly Browser
- Assembly Statistics
- Assembly Browser Settings

### **Navigation in Assembly Browser**

The Navigation tab of the Options Panel in the Assembly Browser includes the list of well-covered regions of the assembly and the field for searching required position.



To learn more about well-covered regions refer to the Assembly Browser Window chapter.

To learn more about searching required position refer to the Go to Position in Assembly chapter.

#### **Assembly Statistics**

The Assembly Statistics tab includes the following Assembly Information:

- Name the name of the opened assembly.
- Length the length of the assembly.

• Reads — the number of reads in the assembly.

Also the tab can include the Reference Information if it is available in the assembly file. For example:

- MD5
- Species
- URI



#### **Assembly Browser Settings**

The Assembly Browser Settings tab includes Reads Area, Consensus Area and Ruler settings.

U -* UGENE - [chrM.sam.bam.sa	am [as] chrM]			
U File Actions Settings T	ools Window Help			_ 8 >
🗔 🗁 📑 🔍 🔍				
U         0 to 16 571 (16 571 bp)           U         0 to 16 571 (16 571 bp)	8.1k 8110	8 133 C 163 8 14	(65 bp)	ssembly Browser Settings Reads Area Reads highlighting: Difference Vou should add reference first for correct displaying of this highlighting Scrolling can be optimized by drawing only reads positions without content while scrolling:
				Vitine Scioling.
ETERIC C C C TT C E C TT C E C C TT C E C C C C				Help No active tasks 💡 🗐

To learn more about Reads Area settings refer to the Reads Area Settings chapter.

To learn more about Consensus see the Consensus Sequence chapter.

To learn more about Ruler see the Browsing and Zooming Assembly chapter.

## **Assembly Browser Hotkeys**

- Assembly Overview Hotkeys
- Reads Area Hotkeys

### **Assembly Overview Hotkeys**

The following hotkeys are available for the Assembly Overview:

Hotkey	Action
Shift + move mouse	Zoom the Assembly Overview to selection
Ctrl + wheel	Zoom the Assembly Overview
Alt + click	Zoom the Assembly Overview in 100x
wheel + move mouse	Move the Assembly Overview

### **Reads Area Hotkeys**

The following hotkeys are available for the Reads Area:

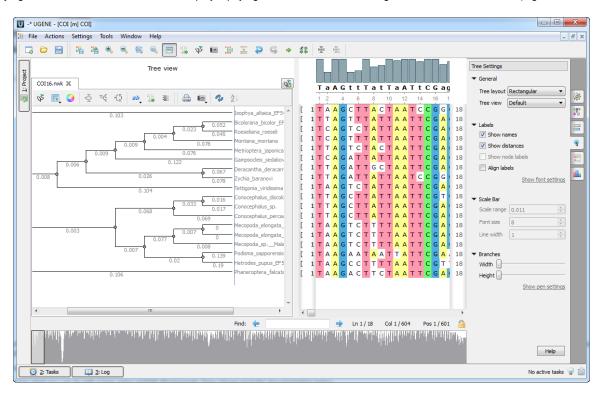
Hotkey	Action
wheel	Zoom the Reads Area
double-click	Zoom in the Reads Area
+/-	Zoom in / zoom out the Reads Area
click + move mouse	Move the Reads Area
arrow	Move one base in the corresponding direction in the Reads Area

# Unipro UGENE Manual, Version 1.30

Ctrl + arrow	Move one page in the corresponding direction in the Reads Area
Page Up / Page Down	Move one page up / down in the Reads Area
Home / End	Move to the beginning / end of the assembly in the Reads Area
Ctrl+G	Focus to the Go to position field on the toolbar

# **Phylogenetic Tree Viewer**

The Phylogenetic Tree Viewer is intended to display a phylogenetic tree built from an alignment or loaded from a file (e.g. a Newick file).



To load a tree from a file follow the instruction described in the Opening Document paragraph or use the Tree settings tab of the Options Panel. For example, you may open the \$UGENE\data\samples\Newick\COI.nwk sample file provided within UGENE package.

To build a tree from a multiple sequence alignment see the Building Phylogenetic Tree paragraph.

To learn what you can do with a tree using UGENE Phylogenetic Tree Viewer read the documentation below.

- Tree Settings
  - Selecting Tree Layout and View
  - Modifying Labels Appearance
    - Showing and Hiding Labels •
    - Aligning Labels
    - Changing Labels Formatting
  - Adjusting Branch Settings
- Zooming Tree

•

- Working with Clade
  - Selecting Clade
    - Collapsing and Expanding Branches
  - Swapping Siblings
  - Zooming Clade
  - Adjusting Clade Settings
  - Changing Root •
- Exporting Tree Image
- **Printing Tree**

## **Tree Settings**

To adjust a tree settings select either the Tree Settings toolbar button or the Tree settings tab of the Options Panel. The Tree settings tab:

	_
Tree Settings	
▼ General	
Tree layout Rectangular	- C.A.
Tree view Default	<u></u>
	15
▼ Labels	
Show names	
Show distances	¥
Show node labels	800
Align labels	
Hide font settings	
Color	
Size 11	
Attributes B / U	
▼ Scale Bar	
Scale range 0.011	
Font size 11	
Line width 1	
Branches     Width Height Hide pen settings	
Color	
Line width 1	

Detailed information about tree setting see below:

- Selecting Tree Layout and View
  Modifying Labels Appearance
  Showing and Hiding Labels
  - - Aligning Labels
      Changing Labels Formatting
- Adjusting Branch Settings

### **Selecting Tree Layout and View**

You can select one of the following tree layouts:

- Rectangular
- Circular
- Unrooted

To do it press the Layout toolbar button and check the required item in the appeared menu or select it in the Tree settings Options Panel tab:



Also you can select one of the following tree view:

- Default
- Phylogram
- Cladogram

#### Modifying Labels Appearance

From this paragraph you can learn how to show/hide taxon and distance labels, align them and change their formatting (font, color, etc.).

- Showing and Hiding Labels
- Aligning Labels
- Changing Labels Formatting

#### **Showing and Hiding Labels**

When you open a tree all available labels are shown by default.

To hide the taxon (sequence name) labels select the Show labels toolbar button or in the Tree settings Options Panel tab uncheck the Show names item.

To hide the distance labels uncheck the Show distances item.

To show/hide node labels (if it is available) use the Show node labels item.

To show the labels again check an appropriate item.

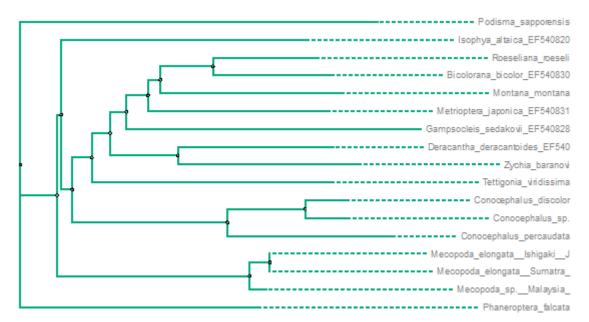
Labels settings in the Options Panel:

•	Labels		
	Show names		
	Show distances		
	Show node labels		
	Align labels		

### **Aligning Labels**

To align a tree labels press the Align labels toolbar button or in the Tree settings Options Panel tab check the Align label item.

See the example of aligning labels below:



#### **Changing Labels Formatting**

To change formatting of a tree labels select the Labels Formatting toolbar button or the Tree settings Options Panel tab:

Color		
Font	Helvetica	-
Size	12	<u>▲</u> ▼
Attributes		B <i>I</i> <u>U</u>

Here you can select color, font, size and attributes (bold, italic, etc.) of the labels.

Note that when a clade has been selected the labels formatting settings are applied to the clade only.

### **Adjusting Branch Settings**

To adjust branch settings select the Branch Settings toolbar button, the Branch Settings context menu item or the Tree settings Options Panel tab.

The following settings are available:

▼ Branches Width Height		Hide pen settings
		<u>Inde per betangs</u>
Color		
Line weight	1	 ▼

Here you can select the color and the line width of the tree branches.

Note that when a clade has been selected the branch settings are applied to the clade only.

## **Zooming Tree**

To change the size of a tree use the Zoom In and Zoom Out toolbar button. You can use the Restore Zooming toolbar button to set the default size.

Or use the corresponding items in the Actions main menu.

See also: Zooming Clade.

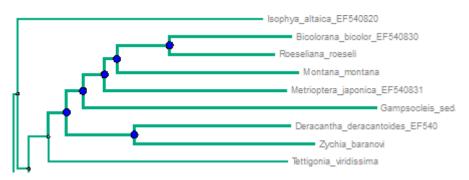
# Working with Clade

This paragraph describes how to select a clade and modify it's appearance.

- Selecting Clade
- Collapsing and Expanding Branches
- Swapping Siblings
- Zooming Clade
- Adjusting Clade Settings
- Changing Root

### **Selecting Clade**

To select a clade click on it's root node:

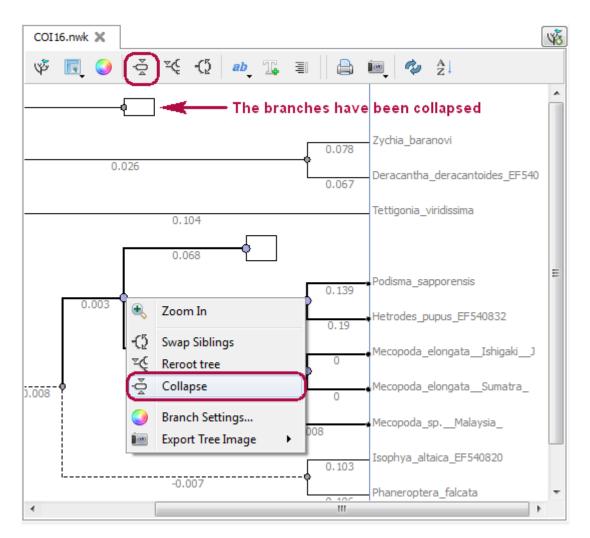


You can see that the corresponding branches are highlighted.

To select several clades at the same time hold the Shift key and click on the root nodes of the clades.

### **Collapsing and Expanding Branches**

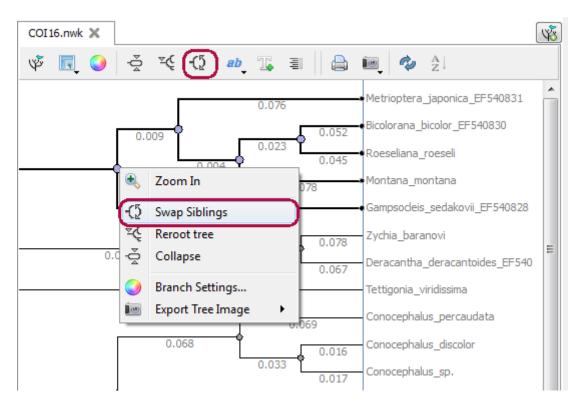
You can hide branches of a clade by selecting the *Collapse* item in the context menu of the clade's root node or use the *Collapse* button on the tree toolbar:



To show the collapsed clade select the Expand item in the node's context menu.

### Swapping Siblings

To rearrange two branches of an internal node, select the *Swap Siblings* item in the node context menu or click the *Swap Siblings* button on the tree toolbar, while the node is selected:



### **Zooming Clade**

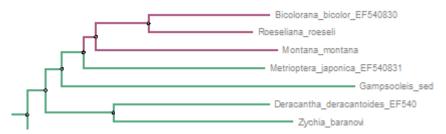
Additionally to other zooming options you can use the Zoom In item in the context menu of the root node of a clade.

### **Adjusting Clade Settings**

When a clade is selected the branch and the labels formatting settigns are applied to the clade only.

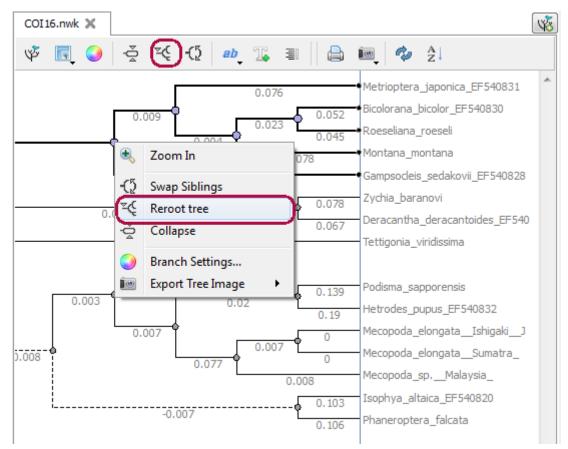
Note that the settings are not applied to the *collapsed* branches (if any).

See an example of changing branch settings for a clade:



### **Changing Root**

To change root of a tree select the root and call the Reroot tree context menu item or use the Reroot tree button on the tree toolbar:



# **Exporting Tree Image**

A tree image can be exported to a raster format (.png, .jpg, .bmp, etc.) or to a vector format (.svg).

Select either the Export Tree Image toolbar button or the Actions Export Tree Image item in the main menu.

In the submenu appeared select the Screen Capture item to save the tree image to a raster format. The standard Save As dialog will appear where you can select the file name and format.

To export a tree image to a vector format select the As SVG item in the Export Tree Image submenu.

## **Printing Tree**

To print a tree select either the Print Tree toolbar button or the Actions Print Tree item in the main menu.

The standard print dialog will appear where you can select a printer to use and specify other settings.

# Extensions

- · Workflow Designer
- DNA Annotator
- **DNA Flexibility** ٠
  - Configuring Dialog Settings Result Annotations
  - **DNA Statistics**
- ٠ **DNA** Generator
- **ORF** Marker

٠

- ٠ Remote BLAST
  - Exporting BLAST Results to Alignment
  - Fetching Sequences from Remote Database
- BLAST and BLAST+
  - Creating Database
  - Making Request to Database
  - Fetching Sequences from Local BLAST Database
- Repeat Finder
  - Repeats Finding
  - Tandem Repeats Finding
    - Tandem Repeats Search Result
- Restriction Analysis
  - Selecting Restriction Enzymes
    - Using Custom File with Enzymes
    - Filtering by Number of Hits
    - Excluding Region
    - Circular Molecule
  - Results
- Molecular Cloning in silico
  - Digesting into Fragments
  - Creating Fragment
  - Constructing Molecule
    - Available Fragments
    - Fragments of the New Molecule
    - Changing Fragments Order in the New Molecule
    - Removing Fragment from the New Molecule
    - Editing Fragment Overhangs •
    - Reverse Complement a Fragment
    - Other Constuction Options ٠
  - Output
  - Creating PCR Product
- In Silico PCR
  - Primers Details
- Primer Library
  Secondary Structure Prediction
- SITECON .
  - SITECON Searching Transcription Factors Binding Sites
     Types of SITECON Models

    - Eukaryotic
    - Prokaryotic
  - Building SITECON Model
- Smith-Waterman Search
- HMM2
  - Building HMM2 Model
    - Calibrating HMM2 Model •
    - ٠ Searching Sequence Using HMM2 Profile
- HMM3
  - Building HMM Model
  - Searching Sequence Using HMM Profile
  - Searching Sequence Against Sequence Database
- uMUSCLE
  - MUSCLE Aligning
  - Aligning Profile to Profile with MUSCLE
  - Aligning Sequences to Profile with MUSCLE
- ClustalW
- . MAFFT
- **T-Coffee** ٠
- ٠ Bowtie
  - Bowtie Aligning Short Reads
  - Building Index for Bowtie
- Bowtie 2
  - Bowtie 2 Aligning Short Reads Building Index for Bowtie 2
- BWA
  - Aligning Short Reads with BWA ٠
  - Building Index for BWA
- BWA-SW
  - Aligning Short Reads with BWA-SW

- Building Index for BWA-SW
- BWA-MEM
  - Aligning Short Reads with BWA-MEM
  - Building Index for BWA-MEM
- UGENE Genome Aligner
  - Aligning Short Reads with UGENE Genome Aligner
  - Building Index for UGENE Genome Aligner
  - Converting UGENE Assembly Database to SAM Format
- CAP3SPAde

٠

٠

- SPAdes
  Weight Mat
  - Weight Matrix
    - Searching JASPAR Database
      Building New Matrix
  - Primer3
  - RTPCR Primer Design
  - Spliced Alignment mRNA and cDNA
- External Tools
  - Configuring External Tool
- Query Designer
- Plasmid Auto Annotation
- ClustalO
- Kalign Aligning

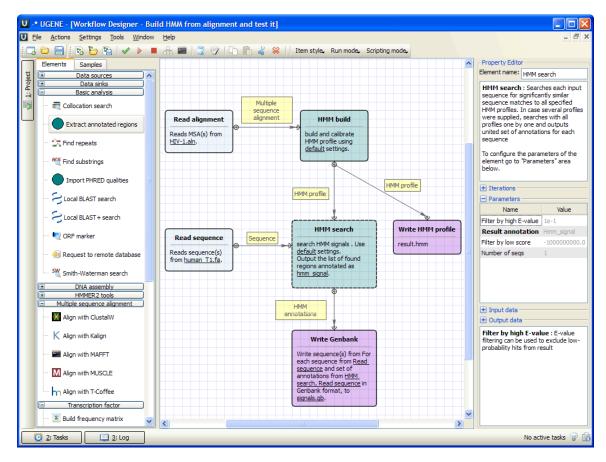
### **Workflow Designer**

The *Workflow Designer* allows a molecular biologist to create and run complex computational workflow schemas even if he or she is not familiar with any programming language.

The workflow schemas comprise reproducible, reusable and self-documented research routines, with a simple and unambiguous visual representation suitable for publications.

The workflow schemas can be run both locally and remotely, either using graphical interface or launched from the command line.

The elements that a schema consists of corresponds to the bulk of algorithms integrated into UGENE. Additionally you can create custom workflow elements.



To learn more about the *Workflow Designer* read the Workflow Designer Manual (follow the link on the UGENE documentation page).

### **DNA** Annotator

The DNA Annotator plugin provides an algorithm to search for sequence regions that contain a predefined set of annotations.

#### Usage example:

Open the Sequence View for a sequence that has annotations. A good candidate here could be any file in Genbank format with a rich set of annotations.

Select the Analyze Find annotated regions item in the context menu. The dialog will appear:

U Find Groups of Annotated Regions	२ ×
CDS - misc_feature - < <click '+'="" add="" annotation="" button="" new="" to="">&gt; + Annotations to search</click>	[444, 1443] [4505, 5504] <b>Results</b>
Region size: 1000 🐳	Save regions as annotations) Clear results
<ul> <li>Direct</li> <li>Complement</li> <li>Both</li> <li>Annotation must fit into region</li> </ul>	
Found 2 regions	Search Cancel Help

Using this dialog you can search for DNA sequence regions that contain every annotation from the list on the left side. The found regions are displayed on the right side of the dialog.

Use the Save regions as annotations button to store the regions as new annotations to the sequence.

## **DNA Flexibility**

To search for regions of high DNA helix flexibility in a DNA sequence, open the sequence in the Sequence View and select the Analyze Find high DNA flexibility regions item in the context menu. Note that only standard DNA alphabet is supported, i.e. the sequence should consist of characters A, C, G, T and N.

The following dialog appears:

U DNA Flexibility		? ×
Search Settings High DNA Flexibilit	Output y Regions Settings	
Window size Window step Threshold	100 1 13.7	Remember Settings Restore Defaults
		 Search Cancel Help

The calculation is made for overlapping windows along a given sequence. If there are two or more consecutive windows with an average flexibility threshold (in each window) greater than the specified *Threshold* parameter, such area is marked by an *annotation*.

The average threshold in a window is calculated by the following formula:

```
(average window threshold) = (sum of flexibility angles in the window) / (the
window size - 1)
```

The following flexibility angles are used during the calculation:

Dinucleotide	Angle	Dinucleotide	Angle
AA	7.6	CA	14.6
AC	10.9	сс	7.2
AG	8.8	CG	11.1
AT	12.5	СТ	8.8
GA	8.2	ТА	25
GC	8.9	тс	8.2
GG	7.2	TG	14.6
GT	10.9	π	7.6

A minimum value is used when N characters is present in a dinucleotide:

- CN, NC, GN, NG, NN: 7.2
- AN, NA, TN, NT : 7.6
- Configuring Dialog Settings
- Result Annotations

### **Configuring Dialog Settings**

In the dialog you can setup the corresponding parameters:

Window size — the number of bases in a window. The window size should be greater than 2. The default value is 100 bp.

Window step — the number of bases used to shift a window. The Window step should be a positive integer. The default value is 1 bp.

Threshold — the threshold value of the twist angle (see above). The default value is 13.7.

You can remember the input values or restore the default values using the Remember Setting and the Restore Defaults buttons.

The annotations names and other parameters can be changed on the Output tab of the dialog:

U DNA Flexibility			? <mark>X</mark>
Search Settings Ou	tput		
- Course and the first (a)			
<ul> <li>Save annotation(s)</li> </ul>	) to		
Existing table		ONC_001363 features [murine.gb] ▼	
Create new ta	ble		
<ul> <li>Use auto-anno</li> </ul>	otations table		
<ul> <li>Annotation parame</li> </ul>	eters		
Group name	<auto></auto>		*
Annotation name	dna_flex		
Description			
		Search Cancel	Help

Once the Search button has been pressed, the annotations for the regions of the high DNA flexibility are created.

#### **Result Annotations**

Each annotation has the following qualifiers:

- area\_average\_threshold average window threshold in the area (i.e. total\_threshold / windows\_number)
- total\_threshold sum of all window thresholds in the area
- *windows\_number* number of windows in the area

🖨 🗖 dna_flex	144156
area_average_threshold	14.672
···· total_threshold	58.689
windows_number	4

M Using the DNA Graphs Package you can see the flexibility graph of a DNA sequence.

## **DNA Statistics**

The DNA Statistics plugin provides exportable statistic reports.

In the current UGENE version, the DNA Statistics plugin provides only Alignment Grid Profile report. The Alignment Grid Profile shows positional amino acid or nucleotide counts highlighted according to the frequency of symbols in a row.

The original idea of the MSA Grid Profile is described in the following paper:

"Alberto Roca, Albert Almada and Aaron C Abajian: ProfileGrids as a new visual representation of large multiple sequence alignments: a case study of the RecA protein family, BMC Bioinformatics 2008, 9:554"

#### Usage example:

13 S S S D G н E Ctrl+G Go to position.. ⇔ S Т Е 13 К γ G Сору Т N. D 13 R γ D Edit R S SS D 14 S C 9 Align -٠ G D Т Т 14 1 Statistics ۲ Generate grid profile 13 Т Т S \_  $\mathbf{V}$ View ۶ 13 Т S G Т E D Δ \_ 0 Y Advanced S G Е 13 A N D S G M Т **T** 14 11 0

Open a sequence alignment in the Alignment Editor and use the Statistics Generate grid profile context menu item.

The dialog will appear:

Generate Alignment Prof	ile 😣
Profile mode	
Ount	
O Percentage	
Custom options	
Show scores for gaps	
Show scores for symbols r	not used in alignment
Skip gaps in consensus po	sition increments
Save profile to file	
File	JUGENE_Data/COI_grid_profile_4.html
Hypertext (.HTML)	
O Comma separated ( .CSV)	
Help	Cancel OK

Here is a brief description of the options that can be set in the dialog:

Profile mode: Counts/Percents - select the Percents to have scores shown as percents in the report.

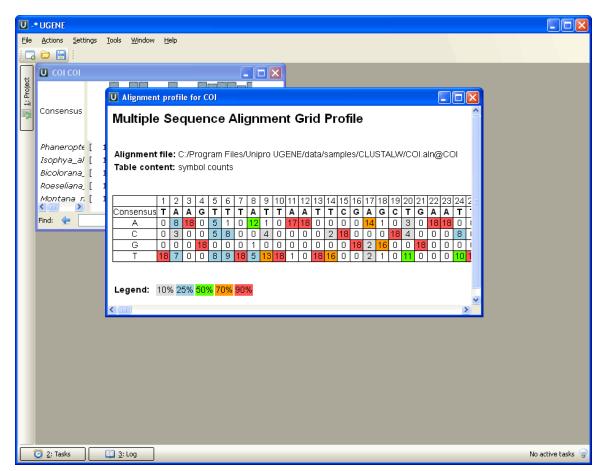
Show scores for gaps --- check this item if you want gap characters ('---') statistics to be shown in the report.

Show scores for symbols not used in alignment — if a symbol is not used in the alignment at all it won't be shown in the report. Check this item to make all symbols of alignment alphabet reported.

Skip gaps in consensus position increments — consensus ruler configuration. If checked the gaps in consensus will not lead to ruler increments.

Save profile to file — allows saving profile to a file in the HTML or CSV format. The CSV format is convenient for further processing in worksheets editors like Excel.

The resulting profile in the HTML mode:



# **DNA Generator**

DNA sequence generator is a tool that generates a random DNA sequence with specified nucleotide content. To generate a random DNA sequence select the *Tools->Generate* sequence item in the main menu. The dialog will appear:

• Generate Seque	ence		8
Parameters			
Length		1000 bp	\$
Window size		1000 bp	•
Number of sequen	ces	1	•
Initialize randon	n generator man	ually 0	
Content			
Reference			
O Base content (%	b)		
A 25 C	25 🗘 G	25 🗘 T 25	
○ GC Skew		0.00	
Output			
Output File			
Format	FASTA		•
✓ Add to project			
Help		Cancel Gener	ate

The following parameters are available:

Length - length of the resulted sequence(s) (using '1000' bp by default).

Window size - the size of a window where set content (using '1000' by default).

Number of sequences - the number of sequences to generate (using '1' by default).

Initialize random generator manually - value to initialize the random generator.

Reference - path to the reference file (could be a sequence or an alignment).

Base content - set the base content percents manually.

GC Skew - set the GC skew of the resulted sequence.

Output file - output file.

Format - output file format (using 'fasta' by default).

Add to project - adds the generated sequence(s) to project.

Once the Search button has been pressed, the sequence(s) are created.

# **ORF Marker**

From this chapter you can learn how to search for Open Reading Frames (ORF) in a DNA sequence. The ORFs found are stored as automatic annotations. This means that if the automatic annotations highlighting has been enabled then ORFs are searched and highlighted for each sequence opened. Refer *Automatic Annotations Highlighting* to learn more.

To open the ORF Marker dialog, select the Analyze Find ORFs item in the context menu.

ORF Marker	8
Settings Output	
Strand	Search Settings Preview
	☑ Min length, bp: 100 🗘 Clear results
<ul><li>Both</li></ul>	Must terminate within region
	Must start with init codon
O Direct	Allow overlaps
	Allow alternative init codons
O Complement	Include stop codon
	☑ Max result 200000
1. The Standard Genetic	: Code
Start codons Alternative start codon Stop codons	ATG s TTG CTG TAA TAG TGA
Region Whole sequence	e 🗘 🚺 - 🚺 199950
Region ^ S	trand Length
0 results found. Help	Cancel OK

The following search settings are available:

*Min length* — ORFs with length lower than *Min length* value will not be found.

Must terminate within region — this option ignores boundary ORFs located beyond the search region.

*Must start with init codon* — item switches the ORF Marker algorithm to the mode when any non-stop amino acid code is interpreted as region start position.

Allow overlaps — alternative (downstream) initiators, when another start codon is located within a longer ORF, i.e. all possible ORFs will be found, not only the longest ones.

Allow alternative init codon — option includes ORFs starting with alternative initiation codons, accordingly to the current translation table.

Include stop codon - includes stop codons into resulting annotations.

The other available parameters are:

DNA-to-Amino translation table defines the way start, alternative start and stop codons are encoded.

Strand — where to search the ORFs: in the direct strand, in the complement strand or in both strands.

Preview — allow to preview the regions, strands and lengths of the found ORFs.

Clear results - becomes available when some results have been found, clears these results.

To set the saving parameters go to the Output tab of the dialog.

Here you can modify the annotations saving parameters (Group name, Description and a file to save the annotation to).

Results:

When the search parameters has been selected and the OK button has been pressed in the dialog, the *auto-annotating* becomes enabled. In the *Annotations editor* the ORFs annotations can be found in the Auto-annotations\orf group.

After the search has been finished you can browse the results, sort them by length, strand or start position and save as annotations to the original sequence in the Genbank format.

For more information about codons use the codon table. To show or hide the table use *Ctrl+T* shortcut or click the *Show codon table* toolbar button menu:



The codon table will appear:

۲	• -* UGENE - [human_T1 [s] human_T1 (UCSC April 2002 chr7:115977709-117855134)]							
8	<u>F</u> ile <u>A</u> ct	ions <u>S</u>	ettings <u>T</u> ools <u>W</u> indow <u>H</u> e	lp				
	□ 🗁 🔚 🔏 🕼 🕼 🕼 🕼 🕼 🛍 🛍 🛍 🚺 🔽 Go! 🖉 : → 1:1 → 🛛 🟘 5½ 🔍 🗨 🗮 🔯							
L	1st base				2nd	base		
1: Project	ISC Dase		U		С		A	
D.		UUU	Rhopydalapipe (Rho, C)	UCU		UAU	Turacina (Tur V)	
<u>н</u>	U	UUC	Phenylalanine (Phe, F)	UCC	Serine (Ser, S)	UAC	Tyrosine (Tyr, Y)	
w.		UUA		UCA	Serine (Ser, S)	UAA	Stop codon (*)	
		UUG		UCG		UAG	<u>3000 00001(-)</u>	
	с	CUU	<u>Leucine (Leu, L)</u>	CCU	<u>Proline (Pro, P)</u>	CAU	Histidine (His, H)	
		CUC		CCC		CAC		
		CUA		CCA		CAA	Glutamine (Gln, Q)	
		CUG		CCG		CAG		
		AUU		ACU		AAU	Asparagine (Asn, N)	
	A	AUC	Isoleucine (Ile, I)	ACC	Threonine (Thr, T)	AAC	<u></u>	
		AUA		ACA	<u></u>	AAA	Lysine (Lys, K)	
		AUG	Methionine (Met, M)	ACG		AAG	<u></u>	
		GUU		GCU	Alanine (Ala, A)	GAU	Aspartic acid (Asp, D)	
	G	GUC	Valine (Val, V)	GCC		GAC	<u></u>	
		GUA	<u>((00, 1)</u>	GCA	<u></u>	GAA	Glutamic acid (Glu, E)	
		GUG		GCG		GAG		

Clicking on a codon name redirects you to Wikipedia to give you a brief description of the corresponding amino acid. Cells of the table are colored according to classes of amino acids.

# **Remote BLAST**

The Remote BLAST plugin provides a capability to annotate sequences with information stored in the NCBI BLAST remote database.

To perform a remote database search open a *Sequence View*, select a sequence region to analyze and click the *Analyze Query NCBI* BLAST database context menu item. If a region is not selected the whole sequence will be analyzed.

	PT 1 1/2		ACG	Find wettern	CHL F
-	Find qualifier			Find pattern	Ctrl+F
	Invert annotation selection		sw	Find pattern [Smith-Waterman]	Ctrl+Shift+F
	Rename item	F2	•	Find ORFs	
	Go to position	Ctrl+G	₹.	Find annotated regions	
			2	Annotate plasmid	
	Select sequence region	Ctrl+A		Build dotplot	
4	New annotation	Ctrl+N	<u>2</u>	Find repeats	
	Copy/Paste	+	T	Find tandems	
	Select	•	1	Analyze with query schema	
	Add	•	R	Find restriction sites	
	Analyze	•	<b>%</b>	Query NCBI BLAST database	
-	Align	•	H M	Find HMM signals with HMMER2	
	Cloning	+	M	Find HMM signals with HMMER3	

The following dialog will appear where you can choose the search options:

Search Through a Remote Database
General options Advanced options
Select the search type: 🔹 blastn 💌 🗷 Search for short, nearly exact matches
Expectation value: 10.000000 📩 🗆 Megablast
Results limit: 20 📩
The database:
Nucleotide collection - nr
The database description:
biological sequence information, such as the amino-acid sequences of different proteins or the nucleotides of DNA sequences. A BLAST search enables a researcher to compare a query sequence with a library or database of sequences, and identify library sequences that resemble the query sequence above a certain threshold.
O Existing document
• New document
✓ Annotation parameters
Group name <auto></auto>
Description
Search timeout 20 min 🛨
Help Cancel Search

General options are:

Select the search type — in the remote databases the *blastn* search is used for nucleotide sequences, *blastp* and *cdd* searches are used for amino sequences.

UGENE also provides a way to use *blastp* and *cdd* searches for nucleotide sequences. This is achieved by translating the nucleotide sequence into the amino sequences.

When a sequence is translated the translation table from the active Sequence View is used. Finally, all 6 translations are used to query the remote database with the selected *blastp* or *cdd* search.

*Expectation value* — this option specifies the statistical significance threshold for reporting matches against database sequences. Lower expect thresholds are more stringent, leading to fewer chance matches being reported.

*Max hits* — the maximum number of hits that will be shown (not equal to number of annotations). The maximum available number is 5000.

Database — the target database.

Search for short, nearly exact matches — automatically adjusts the word size and other parameters to improve results for short queries.

Megablast — select this option to compare query with closely related sequences. It works best if the target percent identity is 95% or more, but it is very fast.

You can see the description of the annotation saving parameters here.

Search timeout - the remote task terminated if the timeout is reached.

There is a little difference in default values of parameters between NCBI Nucleotide BLAST web interface and UGENE:

The web interface uses the *megablast* option by default: the search is fast, but only highly similar sequences are found.
UGENE ignores the option by default: the search may take more time, but all somewhat similar sequences are found.

Check the Megablast option, if you want exactly the same results to be found in UGENE as you had in the NCBI web interface.

Also there is Advanced options tab:

<ul> <li>Search Through a Remote Database</li> </ul>		8
General options Advanced options		
Word size 11 💌	Gap costs	52 -
Match scores 2 -3 💌		
Entrez query		
Filters	Masks	
✓ Low complexity filter	Mask for lookup table only	
Human repeats filter	□ Mask lower case letters	
- Filter results		
Filter by	Select result by	
accession	• EValue	
$\square$ def (filter by definition of annotations)	○ Score	
🗆 id		
L		
Help	Cancel	Search

The view of the Advanced options tab depends on the selected search. For the blastn search it looks like on the picture above.

Word size - the size of the subsequence parameter for the initiated search.

Gap costs — costs to create and extend a gap in an alignment. Increasing the Gap costs will result in alignments which decrease the number of Gaps introduced.

Match scores - reward and penalty for matching and mismatching bases.

*Entrez query* — a BLAST search can be limited to the result of an *Entrez query* against the database chosen. This restricts the search to a subset of entries from that database fitting the requirement of the *Entrez query*. Examples are given below:

protease NOT hiv1[organism] — this will limit a BLAST search to all proteases, except those in HIV 1.

1000:2000[slen] — this limits the search to entries with lengths between 1000 to 2000 bases for nucleotide entries, or 1000 to 2000 residues for protein entries.

*Mus musculus[organism] AND biomol\_mrna[properties]* — this limits the search to mouse mRNA entries in the database. For common organisms, one can also select from the pulldown menu.

10000:100000[mlwt] — this is yet another example usage, which limits the search to protein sequences with calculated molecular weight between 10 kD to 100 kD.

src specimen voucher[properties] — this limits the search to entries that are annotated with a /specimen\_voucher qualifier on the source feature.

all[filter] NOT environmental sample[filter] NOT metagenomes[orgn] — this excludes sequences from metagenome studies and uncultured sequences from anonymous environmental sample studies.

For help in constructing Entrez queries see the Entrez Help document.

Filters — filters for regions of low compositional complexity and repeat elements of the human's genome.

*Masks for lookup table only* — this option masks only for purposes of constructing the lookup table used by BLAST so that no hits are found based upon low-complexity sequence or repeats (if repeat filter is checked).

Mask lower case letters — with this option selected you can cut and paste a FASTA sequence in upper case characters and denote areas you would like filtered with lower case.

Filter by - filters results by accession, by definition of annotations or by id.

Select result by - selects results by EValue or by score.

When the *blastp* search is selected in the general options, the view of the *Advanced options* tab is the following:

<ul> <li>Search Through a Remote Database</li> </ul>		8
General options Advanced options		
Word size 6	Gap costs	11 1 💌
Entrez query		
Matrix BLOSUM62 💌	Service	plain 💌
Filters	Masks	
Low complexity filter	Mask for lookup table only	
🗆 Human repeats filter	Mask lower case letters	
Filter results		
Filter by	Select result by	
accession	<ul> <li>EValue</li> </ul>	
$\Box$ def (filter by definition of annotations)	○ Score	
🗆 id		
Help	Cancel	Search

As you can see there is no Match scores option, but there are Matrix and Service options.

Matrix - key element in evaluating the quality of a pair-wise sequence alignment is the "substitution matrix", which assigns a score for aligning any possible pair of residues.

Service — blastp service which needs to be performed: plain, psi or phi.

The Advanced options tab is not available when the cdd search is selected.

- Exporting BLAST Results to AlignmentFetching Sequences from Remote Database

#### **Exporting BLAST Results to Alignment**

To export BLASt results as alignment select the results in the Annotations Editor and call the Export->Export BLAST result to alignment cont ext menu item. The following dialog will appear:

Export BLAST Result to Multiple Alignment			
Export to file			
File format to use	CLUSTALW	•	
Qualifier to use as name	accession	•	
Add reference to alignment			
Add document to the project			
Help		Cancel Export	

The following parameters are available:

Export to file - name of the new file.

*File format to use -* format of the new file. The following formats are available: CLUSTALW, FASTA, MSF, MEGA, NEXUS, PHYLIP Interleaved, PHYLIP Sequential, Stockholm.

Qualifier to use as name - name of the qualifier. The following qualifiers are available: accession, def, id.

Add reference to alignment - adds a reference to alignment.

Add document to the project - adds the new document to the project.

Select the options and click on the Export button.

### **Fetching Sequences from Remote Database**

Each result annotation found with the *remote BLAST* in UGENE has "accession" and "id" qualifiers that can be used to fetch the corresponding sequences from the NCBI. The prompt way to fetch the sequences of several annotations is the following:

- Select the annotations in the Annotations Editor.
- Open the context menu.
- Choose the Fetch sequences from remote database->Fetch sequences by 'id' from 'blast result' item or Fetch sequences from remote database->Fetch sequences by 'accession' from 'blast result' item.

The following dialog will appear:

Get Sequences by ID		8
The sequences from selected BLA NCBI Genbank by their GI identifie		m
Save to directory:		
Add to project		
	Help <u>C</u> ancel <u>O</u> K	

Select an output path in the dialog and click the OK button.

## **BLAST and BLAST+**

The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

BLAST+ is a new version of the BLAST package from the NCBI.

From UGENE you can use the following tools of the old BLAST package:

- blastall the old program developed and distributed by the NCBI for running BLAST searches.
- formatdb formats protein or nucleotide source databases before these databases can be searched by blastall.

And the following tools of the new BLAST+ package:

- blastn searches a nucleotide database using a nucleotide query.
- **blastp** searches a protein database using a protein query.
- **blastx** searches a protein database using a translated nucleotide query.
- tblastn compares a protein query against a translated nucleotide database (the all six reading frames).
- tblastx translates the query nucleotide sequence in all six possible frames and compares it against the six-frame translations of a nucleotide sequence database.
- makeblastdb — formats protein or nucleotide source databases before these databases can be searched by other BLAST+ tools

BLAST home page: http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE\_TYPE=BlastHome

To make BLAST (or BLAST+) tools available from UGENE:

- 1. Install the required verion of *BLAST* (or *BLAST*+) on your system.
- 2. Set the paths to the executables, you are going to use, on the External tools tab of UGENE Application Settings dialog.

After you've finished this configuration you can access the tools from the Tools BLAST submenu of the main menu.

- Creating Database
- Making Request to Database
- Fetching Sequences from Local BLAST Database

#### **Creating Database**

To format a BLAST database do the following:

- If you're using BLAST open Tools BLAST FormatDB.
- If you're using BLAST+ open Tools BLAST BLAST+ make DB.

The Format database dialog appears:

Format Database		
Input data		
Select input file(s) for formatting database:		
<ul> <li>Or select directory with input files:</li> </ul>		
-File filter		
Include files filter *.fa;*.fasta		
O Exclude files filter *.pal		
Type of file(s)		
• protein O nucleotide		
Output settings		
Select the path to save the database into:		
Base name for BLAST files:		
Title for database file:		
Help Cancel Format		

Here you must select the input files. If all the files you want to use are located in one directory, you can simply select the directory with the files. By default only the files are taken into account with \*.fa and \*.fasta extensions. You can change this by specifying either *Include files filter* or *Exclude files filter*.

You can choose either protein or nucleotide type of the files.

Then you must select the path to save the database file and specify a Base name for BLAST files and a Title for database file.

### **Making Request to Database**

To make a request to a local BLAST database do the following:

- If you're using BLAST open Tools BLAST BLAST Search.
- If you're using BLAST+ open Open Tools BLAST BLAST+ Search.

If there is a sequence opened you can also initiate the request to a local BLAST database from the Sequence View.

- If you're using BLAST select the Analyze Query with BLAST item in the context menu or in the Actions main menu.
- If you're using BLAST+ select the Analyze Query with BLAST+ item in the context menu or in the Actions main menu.

The Request to local BLAST database dialog will appear:

Request to Local BLAST Database			
General options Advanced options Extension options			
Select search: blastn 💌 🗆 Search for short, nearly exact matches			
Expectation value 10.000000 🗄 🗆 Megablast			
Best hits limit: 100 🗄 $\odot$ Both strands $\bigcirc$ Direct $\bigcirc$ Complement			
Database path Database file			
Base name for BLAST DB files			
▼ Save annotation(s) to			
<ul> <li>Existing document</li> <li>MyDocument.gb [Annotations]</li> </ul>			
New document     er/Documents/work/UGENE_data/MyDocument.gb			
<ul> <li>Annotation parameters</li> </ul>			
Group name <auto></auto>			
Description			
Number of CPUs being used 8			
Help   Cancel   Search   Restore to default			

The following general options are available:

Select search - here you should select the tool you would like to use. If the query sequence is a nucleotide sequence then blastn, blast x and tblastx items are available. For a protein sequence the items are blastp and tblastn.

*Expectation value* - this option specifies the statistical significance threshold for reporting matches against database sequences. Lower expect thresholds are more stringent, leading to fewer chance matches being reported.

*Culling limit* - the maximum number of hits that will be shown (not equal to number of annotations). The maximum available number is 5000.

Search for short, nearly exact matches - automatically adjusts the word size and other parameters to improve results for short queries.

Megablast - select this option to compare query with closely related sequences. It works best if the target percent identity is 95% or more, but it is very fast.

Database path - path to the database files.

Base name for BLAST DB files - base name for the BLAST database files.

You can see the description of the annotation saving parameters here.

The following advanced parameters are available:

Request to Local BLAST Database	8
General options Advanced options	Extension options
Word size 11 🗧 Gap costs	2 2
Match scores	1-3
Filters Low complexity filter Human repeats filter	Masks Mask for lookup table only Mask lower case letters
Help	Cancel Search Restore to default

Word size - the size of the subsequence parameter for the initiated search.

Gap costs - costs to create and extend a gap in an alignment. Increasing the Gap costs will result in alignments which decrease the number of Gaps introduced.

Match scores - reward and penalty for matching and mismatching bases.

Filters - filters for regions of low compositional complexity and repeat elements of the human's genome.

*Masks for lookup table only* — this option masks only for purposes of constructing the lookup table used by BLAST so that no hits are found based upon low-complexity sequence or repeats (if repeat filter is checked).

Mask lower case letters — with this option selected you can cut and paste a FASTA sequence in upper case characters and denote areas you would like filtered with lower case.

The view of the Advanced options tab depends on the selected search. For the *blastn* search it looks like on the picture above. When the *bla* stx search is selected in the general options, the view of the Advanced options tab is the following:

Request to Local BLAST Database	8
General options Advanced options	Extension options
Word size 3 Gap co	sts 11 1
Threshold 12.00	
Matrix BLOSUM62  Service	e plain 💌
Filters	Masks
Low complexity filter	Mask for lookup table only
Human repeats filter	Mask lower case letters
Help	Cancel Search Restore to default

As you can see there is no Match scores option, but there are Threshold, Matrix, Composition-based statistics and Service options.

Threshold - threshold for extending hits.

*Matrix* — key element in evaluating the quality of a pair-wise sequence alignment is the "substitution matrix", which assigns a score for aligning any possible pair of residues.

Service — blastp service which needs to be performed: plain, psi or phi.

Composition-based statistics - composition-based statistics.

When the tblastx search is selected in the general options, the view of the Advanced options tab is the following:

Request to Local BLAST Database	8
General options Advanced options	Extension options
	e plain 💌
Filters  Low complexity filter  Human repeats filter	Masks Mask for lookup table only Mask lower case letters
Нер	Cancel Search Restore to default

The following extension options are available:

Request to Local BLAST Database	8
General options Advanced options	Extension options
X dropoff value (in bits)	
For gapped alignment	0
For ungapped extensions	7.00
For final gapped alignment	0
Multiple Hits Window Size	40
Perform gapped alignment	
Help	Cancel Search Restore to default

For gapped alignment - X dropoff value (in bits) for gapped alignment.

For ungapped alignment - X dropoff value (in bits) for ungapped alignment.

For final gapped alignment - X dropoff value (in bits) for final gapped alignment.

Multiple hits window size - multiple hits window size.

Perform gapped alignment - performs gapped alignment.

#### Fetching Sequences from Local BLAST Database

To fetch sequences from local BLAST database use the Fetch sequences from local BLAST database->Fetch sequences by 'id' from 'blast result' context menu item of the blast result. The following dialog will appear:

Fetch Sequences from BLAST Database	8
Query ID gi 538117317 gb CY148232.1	
Select database Database path Base name for BLAST DB files	Database file
Output path UGENE_Data/gi_538117317_gb_CY148232_1fa	
Add to project	
Help	Cancel Fetch

Here you need select a query ID, database, type of file(s) and output path. After that click on the *Fetch* button. To fetch sequences for several annotations at the same time select the blast results with *Ctrl* key and call the *Fetch sequences from local BLAST database->Fetch sequences by 'id' from 'blast result'* context menu item.

# **Repeat Finder**

The *Repeat Finder* plugin provides a tool to search for direct and invert repeats in a DNA sequence. Also it allows to search for tandem repeats.

- Repeats Finding
  - Tandem Repeats Finding
    - Tandem Repeats Search Result

## **Repeats Finding**

#### Usage example:

Open a DNA sequence in the Sequence View and select the Analyze Find repeats... context menu item:

_		- Innaire		
	Analyze •	ACG	Find pattern	Ctrl+F
	Align	SW	Find pattern [Smith-Waterman]	Ctrl+Shift+F
	Cloning •	•	Find ORFs	
	Export •	=	Find annotated regions	
1	Edit sequence	2	Annotate plasmid and custom features	
:	Remove •		Build dotplot	
	Rulers	23	Find repeats	
		T.	Find tandems	J
	Statistics	1	Analyze with query schema	

The dialog will appear that allows specifying repeat parameters and the annotations table document to save the results into:

• Find Repeats		8
Base Advanced		
Repeat finder parameters		
Window size		100bp 🗘 1k
Minimum indentity per window		100% 💭 100
Minimum distance between re	peats	Obp 🌲
☑ Maximum distance between r	epeats	5000bp 🗘
Region to process		
Region Whole sequence	1 -	199950
O Use auto-annotations table	/home/yalgaer/MyDocument.gb	
▼ Annotation parameters		
Group name <auto></auto>		*
Annotation name [repeat_unit		
Description		
Estimated repeats count: 0		
Help		Cancel Start

The dialogues status line displays approximate repeats number that will be found with the current settings.

The Advanced tab provides additional repeats finding options:

• Find Repeats	8
Base Advanced	
Advanced parameters	
Custom algorithm	Auto
Search only for repeats that lie inside of an annotated region	
Search only for repeats that have an annotated region inside	
Filter repeats that have an annotated region inside	
Nested repeats filter algorithm	Disjoint repeats 😂
Search for inverted repeats	
Exclude tandems areas	
Estimated repeats count: 0	
Help	Cancel Start
( ) ( )	Concer

The found repeats are saved and displayed as annotations to the DNA sequence:

🚱 human_T1 (UCSC /	April 2002 chr7:115	977709-1 (	🖄 🏷 🖸	Ca Ta	T <sub>T+</sub> 🔟	چ 🔍 »	^
1 2 <mark>0</mark> k	40k 60k	80k	100k 120k	140k	160k	180k 199 950 🗵	j
repeat_unit (4)		repeat uni	.t	r	epeat uni	t ^	
19 852 19880	19.9k 19 19912	920 19940 [45 bp]	19960	19980 20 19990 - 20	0k 20020 [45 bp]	20 047	
•						F =	
S L S []		S L F		L F L	S L	SL	
LSP	S L S	L C	F S L	SF	S L S		
F S L P	L S I	S V	F L S	L S	L S	LSL	
			10000 40005	1011101	40045 400		
19 903 19910 AGAGAGAGG	19915 19920 AGAGAGAGA		19930 19935 AAAGAGAG	19940 AGAAAGA	19945 199 GAGAGAG		
i	ERE	RQK	E R	E K E	R E	RE	
ERER	ERE	RN	KER	KR	ER	ERE	
ERG	RER	ЕТ	K R E	R E	REF	ER	
•						•	Ŧ
Name	^		Value				
4 🚱 Annotations	[MyDocument_1.	qb] *					
a 🧭 repeat_ur	•						-
repeat	_unit		join(1991219	956,19990200	34)		-
I repeat	_unit		join(6372763	804,65754658	31)		
Image: Provide the second s	_unit		join(1854991	85538,189058	189097)		
I repeat	_unit	_	join(1905331	90577,190640	190684)		Ŧ
III							

# **Tandem Repeats Finding**

To find tandem repeats, select the Analyze Find tandems... context menu item in the Sequence View window.

In the opened dialog you can specify the tandem search parameters, the region to search in and the result parameters:

Tandem finder parar	meters					
Tandem preset					All	•
Min period					1n	▲ ▼
Max period					1000000 n	
Region to process						
Region Whole sequ	uence 🔻		1	-		140426
<ul> <li>Save annotation(s)</li> <li>Existing table</li> <li>Create new tage</li> </ul>		<b>O</b> NC_014267 feat	ures [NC_014267	1.gb]		•
Existing table	ble	<b>O</b> NC_014267 feat	ures [NC_014267	.1.gb]		
<ul> <li>Existing table</li> <li>Create new tag</li> </ul>	<b>ble</b> otations table	<b>O</b> NC_014267 feat	ures [NC_014267	.1.gb]		
<ul> <li>Existing table</li> <li>Create new ta</li> <li>Use auto-anno</li> </ul>	<b>ble</b> otations table	<b>ONC_014267</b> feat	ures [NC_014267	.1.gb]		
<ul> <li>Existing table</li> <li>Create new ta</li> <li>Use auto-anno</li> <li>Annotation parameter</li> </ul>	ble otations table eters <auto></auto>	<b>O</b> NC_014267 feat	ures [NC_014267	.1.gb]		
<ul> <li>Existing table</li> <li>Create new ta</li> <li>Use auto-anno</li> <li>Annotation parame</li> <li>Group name</li> </ul>	ble otations table eters <auto></auto>	<b>O</b> NC_014267 feat	ures [NC_014267	.1.gb]		
<ul> <li>Existing table</li> <li>Create new ta</li> <li>Use auto-anno</li> <li>Annotation parame</li> <li>Group name</li> <li>Annotation name</li> </ul>	ble otations table eters <auto></auto>	<b>W</b> NC_014267 feat	ures [NC_014267	.1.gb]		

The dialog parameters:

Tandem preset — specify the tandem repeats parameters with predefined values by selecting the available preset:

U Find Tandems		? ×
Base Advanced Tandem finder parameters		
Tandem preset		Custom
Min period		All Micro-satellites
Max period		Mini-satellites Big-period tandems Custom
Region to process		K. Market Market State
Region Whole sequence 🔻	1 -	199950

Min period, Max period — the minimum and maximum acceptable repeat length measured in base symbols.

Region to process — specify the region to search in the whole sequence, a custom region or the region of the current selection (if any).

In the Save annotation(s) to group you can set up a file to store annotations. It could be either an existing annotation table object, a new annotation table or auto-annotations table (if it is available).

In the Annotation parameters group you can specify the name of the group and the name of the annotation. If the group name is set to <auto> UGENE will use the group name as the name for the group. You can use the '/' characters in this field as a group name

separator to create subgroups. If the annotation name is set to by type UGENE will use the annotation type from the Annotation type: t able as the name for the annotation. Also you can add a description in the corresponding text field.

Base Advanced	
Advanced parameters	
Algorithm	Suffix array (optimized) 🔻
Minimum tandem size	9
Minimum repeat count	x3
Show overlapped tandems	

Additional search options can be found in the Advanced tab:

Algorithm — the algorithm parameter allows to select the search algorithm. The default and a fast one is optimized suffix array algorithm.

*Minimum tandem size* — the minimum tandem size sets the limit on minimum acceptable length of the tandem, i.e. the minimum total repeats length of the searched tandem.

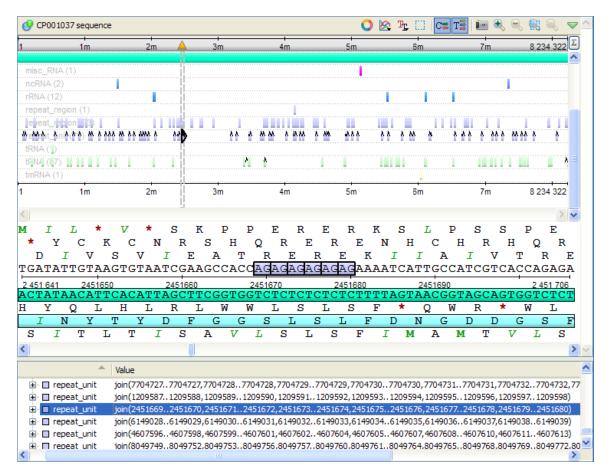
Minimum repeat count — the minimum number of repeats of a searched tandem.

Show overlapped tandems - check if the plugin should search for the overlapped tandems, otherwise keep unchecked.

• Tandem Repeats Search Result

#### **Tandem Repeats Search Result**

An example of the search results for the micro-satellite preset:



The tandem repeats annotations are located side by side.

## **Restriction Analysis**

From this chapter you can learn how to search for restriction sites on a DNA sequence.

The restriction sites found are stored as automatic annotations. This means that if the automatic annotations highlighting is enabled then the restriction sites are searched and highlighted for each nucleotide sequence opened. Refer Automatic Annotations Highlighting to learn more.

Open a DNA sequence in and click the following button on the Sequence View toolbar:

Alternatively, select either the Actions Analyze Find restriction sites item in the main menu or the Analyze Find restriction sites item in the context menu.

The Find restriction sites dialog appears:

Find Restriction Sites	8
Filter by name:	Open enzymes
Name Y Accession Type Sequent	
<ul> <li>▷ A (0, 264)</li> <li>▷ B (2, 917)</li> </ul>	Aaal Axyl Bacl BvuBl Select All
<ul> <li>▷ C (1, 193)</li> <li>▷ D (1, 31)</li> </ul>	Cacl Cvnl Daql Dsp1l Select None
<ul> <li>▷ E (2, 325)</li> <li>▷ F (0, 63)</li> </ul>	Eacl EspHK30I F-CphI F-TevIV Select by length
<ul> <li>▷ G (0, 23)</li> <li>▷ H (1, 312)</li> </ul>	Gall Gsul Invert selection
▶ 1 (0, 61)	I-AchMI I-Vdi1411
Selected enzymes: BamHI,BglII,ClaI,DraI,EcoRI,EcoRV,HindIII,P	Save selection
	REBASE Info
□ Filter by number of results:	
Minimum hits:	1 $\bigcirc$ Maximum hits: 2 $\bigcirc$
Search in:	
Region Whole sequence	\$ 199950
Exc	clude 1- 199950
Total number of enzymes: 4862, selected 11	
Help	Cancel OK

You can see the list of restriction enzymes that can be used to search for restriction sites. Also you can set a region to search for. The information about enzymes was obtained from the REBASE database. For each enzyme in the list a brief description is available (the accession ID in the database, the recognition sequence, etc.). If you're online you can get more detailed information about an enzyme selected by clicking the *REBASE Info* button.

- Selecting Restriction Enzymes
- Using Custom File with Enzymes
- Filtering by Number of Hits
- Excluding Region
- Circular Molecule
- Results

### **Selecting Restriction Enzymes**

To select an enzyme check it in the list. Notice that the enzyme appears in the Selected enzymes area of the dialog.

You can also use the Select All button to select all the enzymes available, the Select None button to deselect all the enzymes.

To select all enzymes with recognition sequence length shorter than the specified value click the Select by length button and input the minimum length in the dialog appeared.

To invert selection click the Invert selection button.

As soon as enzymes are selected you can click the OK button to search for corresponding restriction sites in the sequence.

#### Using Custom File with Enzymes

To load a custom file with enzymes click the Enzymes file button and browse for the file. The file must be of the Bairoch format.

For details about the format refer http://rebase.neb.com/rebase/rebase.f19.html.

To export enzymes use the *Export enzymes* button. You can also save the currently selected enzymes to a file and load saved selection. Click the *Save selection* and *Load selection* buttons correspondingly to do it.

### Filtering by Number of Hits

To filter the results by the number of restriction sites found for an enzyme check the *Filter by number of results* check box and input the minimum value and the maximum value of hits.

## **Excluding Region**

To exclude a sequence region from the search check the *Exclude region* check box and input the start and the end positions of the region. If a subsequence has been selected before opening the dialog you can click the *Selected* button to automatically fill the values with the selected subsequence's start and end positions.

#### **Circular Molecule**

To consider the sequence as circular and be able to search for restriction sites between the end and the beginning of the sequence check the *Circular molecule* option.

Example: Let's consider:

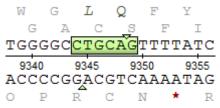
- The sequence is "CTGC ... CAC".
- Aarl restriction enzyme (with recognition sequence "CACCTGC") has been checked.

In this case if the *Circular molecule* option has been checked, the restriction site will be found. If it hasn't been checked, the restriction site won't be found (in this position).

#### Results

When at least one enzyme has been selected and the OK button has been pressed in the dialog, the *auto-annotating* becomes enabled. In the *Annotations editor* the Restriction Sites annotations can be found in the Auto-annotations\enzyme group.

The direct and complement cut site positions are visualized as triangles on an annotation in the Sequence details view.



# Molecular Cloning in silico

This chapter describes a set of tools in UGENE to perform molecular cloning experiments in silico.

This allows you to digest a molecule into fragments, create a fragment from a sequence region and ligate fragments into a new molecule.

- Digesting into Fragments
- Creating Fragment
- Constructing Molecule
  - Available Fragments
  - Fragments of the New Molecule
  - Changing Fragments Order in the New Molecule
  - Removing Fragment from the New Molecule
  - Editing Fragment Overhangs
  - Reverse Complement a Fragment
  - Other Constuction Options
  - Output
- Creating PCR Product

### **Digesting into Fragments**

Open a DNA molecule you want to cut into fragments.

Digestion into fragments is performed using restriction enzymes. So before continuing make sure that the restriction analysis has been performed. Refer chapter *Restriction Analysis* for details.

Select either the Tools Cloning Digest into Fragments item or the Actions Cloning Digest into Fragments item in the main menu or the Clon ing Digest into Fragments item in the context menu.

The Digest Sequence into Fragments dialog appears:

Restriction Sites	Conserved Annotat	ions Outpu	t			
Target Sequence: human_T1 (UCSC April 2002 chr7:115977709-117855134)						
Available enzymes:			Selected enzymes:			
		Add> Add All> Remove Clear Selection				
Circular molecule						
Hint: there are no ava	ailable enzymes. Use "		Restrictions Sites" feat	ure to find them. Help		

On the *Restriction Sites* tab of the dialog you can see the name of the molecule, the list of restriction enzymes found during the restriction analysis that can cut the molecule and the list of enzymes selected to perform the digestion.

To digest the sequence into fragments you should select at least one enzyme.

To move an enzyme to the Selected enzymes list click on it in the Available enzymes list and press the Add button. Note that you can select several items in a list by holding the Ctrl key while clicking on the items.

To select all available enzymes press the Add All button.

To remove enzymes from the Selected enzymes list select them in the list and press the Remove button.

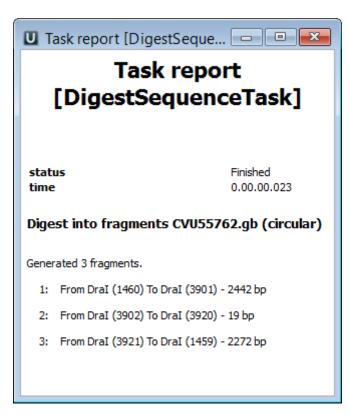
To remove all items from the Selected enzymes list press the Clear Selection button.

On the Conserved Annotations tab of the dialog you can select the annotations that must not be disrupted during cloning.

On the Output tab of the dialog you can select the file to save the new molecule to.

As soon as the required parameters are selected press the OK button. The fragments will be saved as annotations.

Also all the generated fragments are available in the task report:



Refer to Notifications to learn more about task reports.

#### **Creating Fragment**

To create a DNA fragment from a sequence region activate the Sequence View window and select either the Actions Cloning Create Fragment item in the main menu or the Cloning Create Fragment item in the context menu.

The Create DNA Fragment dialog appears:

Create DNA Fragment Fragment Options Output	2 <b>—</b>
Region Whole sequence  Include Left Overhang	1 - 199950
Direct     Reverse-complement	Direct     Reverse-complement
	OK Cancel Help

If a region has been selected you can choose to create the fragment from this region. Otherwise you can either choose to create the fragment from the whole sequence or choose the *Custom* item and input the custom region.

To add a 5' overhang to the direct strand check the *Include Left Overhang* check box and input the required nucleotides. To add a 5' overhang to the reverse strand in addition to the described steps select the *Reverse-complement* item in the same group box.

Similarly, to add a 3' overhang check the *Include Right Overhang* check box, input the required overhang and select either the direct or the reverse-complement strand.

On the Output tab of the dialog you can optionally modify the annotations output settings.

Finally, press the OK button to create the fragment. The fragment will be saved as an annotation.

# **Constructing Molecule**

To construct a new molecule from fragments select the Tools Cloning Construct Molecule item in the main menu.

If a Sequence View window is active you can also select either the Actions Cloning Construct Molecule item in the main menu or the Clonin g Construct Molecule item in the context menu.

The Construct Molecule dialog appears:

	t Molecule		
Construct	tion Output		
Available f	fragments:		
			Add Add All From Project
New molect	cule contents: Fragment	3'	Inverted
•		III	
•	Up Down Rem	III Iove Edit Clear All	
	Up Down Rem		
✓ Annota			
Annota	ate fragments in new molecule		

- Available Fragments
- Fragments of the New Molecule
- Changing Fragments Order in the New Molecule
- Removing Fragment from the New Molecule
- Editing Fragment Overhangs
- Reverse Complement a Fragment
- Other Constuction Options
- Output

**Available Fragments** 

All the fragments available in the current project are shown in the Available fragments list.

You can automatically create a fragment from a DNA molecule from the current UGENE *project*. Click the *From Project* button to do so. The *Select Item* dialog appears with the sequence objects available. Select a sequence and press the *OK* button. After that create a fragment in the appeared *Create DNA Fragment* dialog as described in the *Creating Fragment* paragraph. The fragment created from the sequence appears in the list of available fragments.

**Fragments of the New Molecule** 

The next step is to add required fragments to the new molecule contents.

To add fragments select them in the list of available fragments and click the Add button or by double-click on a fragment.

To add all the fragments click the Add All button.

**Changing Fragments Order in the New Molecule** 

To change the order of fragments in the new molecule select a fragment in the new molecule contents list and click either the *Up* or the *Down* button to move the fragment in the corresponding direction.

**Removing Fragment from the New Molecule** 

To remove a fragment from the new molecule select it in the new molecule contents list and click the Remove button.

To remove all the fragments click the *Clear All* button. **Editing Fragment Overhangs** 

To edit a fragment's overhangs select the fragment in the new molecule contents list and click the *Edit* button.

The Edit Molecule Fragment dialog appears:

Type <ul> <li>Overhang</li> <li>Blunt</li> <li>Custom overhang</li> </ul>
Blunt
Custom overhang
0 3'-5'
Reset
3' AATCGG ITAGCC
5'

Here you can select the type of each DNA end and even input a custom overhang.

The changes you've made are shown in the Preview area of the dialog.

To confirm the changes and close the dialog click the *OK* button. **Reverse Complement a Fragment** 

To reverse complement a fragment check the *Inverted* check box for the fragment in the new molecule contents list. **Other Constuction Options** 

To save the fragments of the new molecule as annotations check the Annotate fragments in new molecule check box.

To make all DNA ends blunt check the Force "blunt" and omit all overhangs check box. All overhangs would be cut in this case.

Check the *Make circular* check box to make the new molecule circular. **Output** 

On the Output tab of the dialog you can select the file to save the new molecule to.

The molecule is opened by default as soon as it is created. To modify this behavior uncheck the Open view for new molecule check box on the same tab.

To save the molecule file to the hard disk immediately after it is created check the *Save immediately* check box. Otherwise it would be stored in memory until you save or remove it.

## **Creating PCR Product**

To create a PCR product from a primer use the Cloning->Create PCR product context menu of primer annotation.

The Create PCR Product dialog appears:

Create PCR product Fragment Options Output	2 <b>-</b>
Region Custom region 🔻	185965 - 186141
Include Left Overhang	Include Right Overhang
Direct     Reverse-complement	Direct     Reverse-complement
	OK Cancel Help

If a primer has been selected you can choose to create the PCR product from this primer. Otherwise you can either choose to create the PCR from the whole sequence or choose the *Custom* item and input the custom region.

To add a 5' overhang to the direct strand check the *Include Left Overhang* check box and input the required nucleotides. To add a 5' overhang to the reverse strand in addition to the described steps select the *Reverse-complement* item in the same group box.

Similarly, to add a 3' overhang check the *Include Right Overhang* check box, input the required overhang and select either the direct or the reverse-complement strand.

On the Output tab of the dialog you can optionally modify the annotations output settings.

Finally, press the OK button to create the PCR product. The PCR product will be saved as an annotation.

# In Silico PCR

#### In Silico PCR Overview

In silico PCR is used to calculate theoretical polymerase chain reaction (PCR) results using a given set of primers (probes) to amplify DNA se quences.

UGENE provides the In silico PCR feature only for nucleic sequences with Standard DNA and Extended DNA alphabets. To use it in UGENE open a DNA sequence and go to the *In silico PCR* tab of the Options Panel:

In Silico PCR					
<ul> <li>Forward prim</li> </ul>	er				
ACGTTACGT	ACGTAC	TACGTAC	STGC		
Tm = 59.54°	C, 26-m	er	C		ACG
Mismatches	0 bp			*	2
<ul> <li>Reverse prim</li> </ul>	er				6
AAAAAACG	TACGTO	GT			X
Tm = 38.25°	°C, 16-m	er	C		0
Mismatches	0 bp			*	
▼ Settings					
3' perfect ma	atch	15 bp		-	
Maximum pro	oduct	5000 bp		* *	
Extract anno	tations	Inner		-	
Show primers de Warning: Self-dimer car Delta G: -15.4 Base Pairs: 11 Find	n be for kcal/m		,		

There are the following parameters:

Forward primer - forward primer.

Reverse primer - on the opposite strand from the forward primer.

Mismatches - mismatches limit.

3' perfect match - specify the number of nucleotides at the 3' end that must not have mismatches.

Maximum product - maximum size of the amplified sequence.

Extract annotations - specify the type of extracted annotations: Inner, All intersected or None.

- Value *Inner* is selected by default. When this value is selected, the extracted PCR product contains annotations from the original sequence, located within the extracted region.
- Value All intersected specifies that all annotations of the original sequence that intersect the extracted region must be extracted as well.
- Value None specifies that annotations from the original sequence must not be extracted.

#### **Choosing primers**

Type two primers for running In Silico PCR. If the primers pair is invalid for running the PCR process then the warning is shown. Also, primer s for the running In silico PCR can be chosen from a primer library. Click the following button to choose a primer from the primers library:

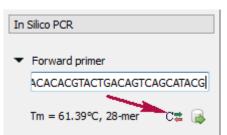
•	Forward primer				
	ACACACGTACTGACAGTCAGCATACG				
	Tm = 61.39%	C, 28-mer	C# 🔒		
	Mismatches	7 bp			

The following dialog will appear:

Name	GC-content (%)	Tm (°C)	Length (bp)	Sequence
Primer	22.22	22	9	AAAAACGT

The table consists of the following columns: name, GC-content (%), Tm, Length (bp) and sequence. Select primer in the table and click the C hoose button.

Click the Reverse-complement button for making a primer sequence reverse-complement:



Click Show primers details for seeing statistic details about primers.

When you run the process, the predicted PCR products appear in the products table.

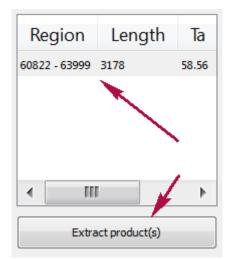
### **Products table**

There are three columns in the table:

- region of product in the sequence
- product length
- preferred annealing temperature

Click the product for navigating to its region in the sequence.

Click the *Extract product*(s) button for exporting a product(s) in a file or use double click for that.



- Primers Details Primer Library

## **Primers Details**

Click Show primers details for seeing statistic details about primers.

Mismatches	10 bp		٢
<ul> <li>Settings</li> </ul>			
Maximum pro	duct	5000 bp	
Show primers de	tails		

The following dialog will appear:

Primers Details			? ×			
Criteria	Valid Values	Forward	Reverse			
% GC	50-60	53.12	41.94			
Tm (°C)	55-80	65.67	60.4			
GC Clamp	>=1 G or C at 3' end	2	4			
Runs	<=4 base runs	3	6			
Self-dimers:						
Delta G: -9.3 kcal/mole Base Pairs: 7						
AGGTCAGGTTACCCAC            A ACTACCCGAAGGGAA						
Delta G: -6.2 kcal/mole Base Pairs: 3						
AAATGAAGGAAA	LAATGCTAAGGGC    A GACCG					
Hetero-dimers:						
Delta G: -16 kcal/	mole Base Pairs: 13					
1 11 1 1	CACAAAGGGAAGCCCATCA	АА				
		01				
		ОК	Help			

This is a dialog with statistic details about primers: melting temperature, GC content, dimers, self-dimers, etc. If a value is not correct for its criteria then it is colored in red.

## **Primer Library**

The primer library is a storage for keeping user primers. The added primers are stored between UGENE sessions.

Go to the Tools->Primer->Primer library context menu to configure the primer library. The following window will appear:

U	U -* UGENE - [Primer Library]					
U	File Actions Se	ttings Tools Win	idow Help			_ & ×
	3 🗁 🗎					
tr tr	Name	GC-content (%)	Tm (°C)	Length (bp)	Sequence	New primer
1: Project	Primer	22.22	22	9	ААААААСӨТ	Edit primer
	Primer 1	52.63	51.09	19	GTCCCACTGTACGTTTACG	Remove primer(s)
	Test Primer	37.5	59.26	32	ACGTTTACGTACGATTCGATACACATAGACAT	Import primer(s)
						Export primer(s)
						Close
						Help
_		63.54				
	🔯 <u>2</u> : Tasks	🛄 <u>3</u> : Log				No active tasks 💡 🚮

Click the New primer button to add a new primer. The following dialog will appear:

U Creat	te New Primer	8	x
þ			3'
Name	New Primer		
	OK Cancel	Help	>

Input the primer sequence and primer name and click on the OK button.

Select the primer and click the Edit primer button to edit primer.

Select the primer in the table (you can use Ctrl and Shift) and click Remove primer(s) button to remove primer(s).

To export primer(s) select it and click the *Export primer(s)* button. The following dialog will appear:

U Export Primers	<u>२</u> ×
Export to: Local file	•
Format: genbank	•
File path:	
	OK Cancel Help

Select file and file format and click on the OK button.

To import primer(s) click the *Import primer(s)* button. The following dialog will appear:

U Import Primers		? <mark>x</mark>
Import from: Local file(s)		
Files:	Ad	dd file(s)
	F	lemove
	Import Cancel	Help

Add one or several files with primer sequences. Note that all sequence formats, supported by UGENE, can be imported, for example, FASTA, GenBank, etc. But the sequences must consist of ACGT characters only.

Click the Import button to import the added files into the primers library.

## **Secondary Structure Prediction**

The Secondary Structure Prediction plugin provides a set of algorithms for the protein secondary structure (alpha-helix, beta-sheet) prediction from a raw sequence.

Currently, available algorithms are:

• GORIV Jean Garnier, Jean-Francois Gibrat, and Barry Robson,"GOR Method for Predicting Protein Secondary Structure from Amino Acid Sequence", in Methods in Enzymology, vol.266, pp. 540 - 553, (1996).

The improved version of the GOR method in J. Garnier, D. Osguthorpe, and B. Robson, J. Mol. Biol., vol. 120, p. 97 (1978).

 PsiPred Bryson K, McGuffin LJ, Marsden RL, Ward JJ, Sodhi JS. & Jones DT. (2005) Protein structure prediction servers at University College London. Nucl. Acids Res. 33(Web Server issue): W36-38.

Jones DT. (1999) Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292: 195-202.

You can access these analysis capabilities for a protein sequence using the *Analyze Predict secondary structure...* context menu item. The dialog will appear:

Secondary Structure	Prediction	8
Algorithm Range Start: 1 📩	GORIV Range End: 82 -	•
Results:		
Region	Structure Type	
Total predicted: 0		
Help	Save Cancel Pr	edict

It supports the following options:

Algorithm - you can choose the preferred algorithm. Currently, "GORIV" and "PsiPred" algorithms are available.

Range start / Range end — select the sequence range for prediction.

*Results* — the visual representation of the prediction results, for example:

Secondary Structure Prediction					
Algorithm Range Start: 1 Results:	• -	GORIV Range End: 82			
Region		Structure Type			
1 [57]	beta_strand				
2 [915]	beta_strand	beta_strand			
3 [2432]	alpha_helix	alpha_helix			
4 [3975]	alpha_helix	alpha_helix			
5 [7581]	beta_strand				
Total predicted:	Total predicted: 5				
Help	Sa	ave Cancel Predict			

Save as annotation — select this button to save the results as annotations of the current protein sequence.

## SITECON

*SITECON* — is a program package for recognition of potential transcription factor binding sites basing on the data about conservative conformational and physicochemical properties revealed on the basis of the binding sites sets analysis.

To cite SITECON use the following article:

"Oshchepkov D.Y., Vityaev E.E., Grigorovich D.A., Ignatieva E.V., Khlebodarova T.M.SITECON: a tool for detecting conservative conformational and physicochemical properties in transcription factor binding site alignments and for siterecognition. //Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W208-12."

UGENE version of SITECON provides a tool for recognition of potential binding sites for over *90 types* of transcription factors. Also UGENE version of SITECON provides a tool for recognition of potential binding sites basing site alignment proposed by user. For the detailed method description see the original SITECON site.

Data about used context-dependent conformational and physicochemical properties are available in the PROPERTY Database.

- SITECON Searching Transcription Factors Binding Sites
  - Types of SITECON Models
    - Eukaryotic
    - Prokaryotic
- Building SITECON Model

## **SITECON Searching Transcription Factors Binding Sites**

To search transcription factor binding sites in a DNA sequence select the Analyze Search TFBS with SITECON... context menu item.

In the appeared search dialog you must select a file with TFBS profile. The profiles supplied with UGENE are placed in the \$UGENE/data/sit econ\_models folder.

After the profile is loaded the threshold-filter is populated with values read from profile. You can use the filter to remove low-scoring regions from the result.

8 SITECON Search				5 ×
File with model				
Threshold				•
				*
				-
Strands			Region	
<ul> <li>Both strands</li> </ul>			Whole sequence	•
O Direct strand			1	199950
Complement st	trand			
Range	Strand	PSUM	First type error	Second type error
Clear results Sa	ve as annotations			
0 results found				
			Search	Close Help

The regions found by SITECON algorithm can be saved as annotations to the DNA sequence in the Genbank format.

Every SITECON profile supplied with UGENE contains complete information about calibration settings provided to UGENE team by the author of SITECON.

The original TFBS alignments used to calculate profiles can be requested directly from the author of SITECON.

## **Types of SITECON Models**

- Eukaryotic
- Prokaryotic

Eukaryotic

Name	Description
CEBP_a	CCAAT-enhancer-binding protein_alpha
CEBP_all	CCAAT-enhancer-binding proteins
CLOCK	Circadian Locomotor Output Cycles Kaput
cMyc_can	Myc (c-Myc) is a regulator gene that codes for a transcription factor. A mutated version of Myc is found in many cancers.
CRE	Cyclic AMP response element
E2F1	Transcription factor E2F1 is a protein that in humans is encoded by the E2F1 gene.
E2F1/DP1sel1	E2F factors bind to DNA as homodimers or heterodimers in association with dimerization partner DP1.

EGR1	Early growth response protein 1
EKLf	Erythroid Kruppel-like Factor
ER2	Estrogen receptor beta
GATA_all	GATA transcription factors are a family of transcription factors characterized by their ability to bind to the DNA sequence "GATA"
GATA-1	GATA-binding factor 1
GATA-2	GATA-binding protein 2
GATA-3	
HMG-1	Trans-acting T-cell-specific transcription factor GATA-3 High-mobility group protein 1
HNF-1	
HNF-3	Hepatocyte nuclear factor 1
	Hepatocyte nuclear factor 3
HNF-4	Hepatocyte nuclear factor 4
IRF	Interferon regulatory factors
isre	Interferon stimulation response element
МуоD	MyoD belongs to a family of proteins known as myogenic regulatory factors (MRFs)
MyOGsel3	Myogenin
NF-1	Neurofibromin 1
NF-E2	Transcription factor NF-E2 45 kDa subunit is a protein that in humans is encoded by the NFE2 gene.
ΝΕΑΤρ	Pre-existing component of the NFAT(Nuclear factor of activated T-cells) transcription complex.
NFkB_all	Nuclear factor kappa-light-chain-enhancer of activated B cells
NFkB_hetero	The p50 (NFKB1)/p65 (RELA) heterodimer is the most abundant form of NF-kB
NFkB_ homo	The c-Rel protein is a member of the NF-kB family of transcription factors and contains a Rel homology domain
Nfy	Nuclear transcription factor Y
Nrf2	Nuclear factor (erythroid-derived 2)-like 2
Oct-1	Octamer transcription factor 1
Oct_all	Octamer transcription factors
p53	Protein 53
PPRF	Paramedian pontine reticular formation
Pu1	Is a protein that in humans is encoded by the SPI1 gene
setCREB	cAMP response element-binding
setCREBzag	cAMP response element-binding
SRE_san	Serum response element
SRF	Serum response factor
STAT1	Signal Transducer and Activator of Transcription 1
STAT	Signal Transducer and Activator of Transcription

TTF1	Thyroid transcription factor 1
USF	Upstream stimulatory factors
уу1	Is a protein that in humans is encoded by the YY1 gene

## Prokaryotic

aC A R	N-acetylgalactosamine repressor, AgaR, negatively controls the expression of the aga gene cluster AgaC is the Enzyme IIC domain of a predicted N-acetylgalactosamine-transporting PEP-dependent phosphotransferase system ArcA transcriptional dual regulator ArgR complexed with L-arginine represses the transcription of several genes involved in biosynthesis and transport of arginine, transport of histidine, and its own synthesis and activates genes for arginine catabolism.
A R	N-acetylgalactosamine-transporting PEP-dependent phosphotransferase system ArcA transcriptional dual regulator ArgR complexed with L-arginine represses the transcription of several genes involved in biosynthesis and transport of arginine, transport of histidine, and its own synthesis and activates genes for
R	ArgR complexed with L-arginine represses the transcription of several genes involved in biosynthesis and transport of arginine, transport of histidine, and its own synthesis and activates genes for
	several genes involved in biosynthesis and transport of arginine, transport of histidine, and its own synthesis and activates genes for
	DNA-binding response regulator in two-component regulatory system with CpxA
	cAMP receptor protein
B	Cysteine B
R	Cytidine Regulator
DR	Deoxyribose Regulator
	DnaA is the linchpin element in the initiation of DNA replication in E. coli.
IR	Fatty acid degradation Regulon
	Factor for inversion stimulation
DC	Operon that encodes two transcriptional regulators
	FNR is the primary transcriptional regulator that mediates the transition from aerobic to anaerobic growth through the regulation of hundreds of genes.
r	Fructose repressor
R	Ferric Uptake Regulation
LR	Galactose repressor
LS	Galactose isorepressor
PR	sn-Glycerol-3-phosphate repressor
TP	Is a member of the GntP family transporters
s	Histone-like nucleoid structuring protein
R	Isocitrate lyase Regulator
	Integration host factor
R1	Iron-sulfur cluster Regulator 1
R3	Iron-sulfur cluster Regulator 3
	LexA represses the transcription of several genes involved in the cellular response to DNA damage or inhibition of DNA replication

Lrp	Leucine-responsive regulatory protein
MALT	Maltose regulator
MARA	Multiple antibiotic resistance
MELR	Melibiose regulator
MEtJ	MetJ represses the expression of genes involved in biosynthesis and transport of methionine
MetR1	MetR participates in controlling several genes involved in methionine biosynthesis [Weissbach91] and a gene involved in protection against nitric oxide
MLC	DgsA, better known as Mlc, "makes large colonies," is a transcriptional dual regulator that controls the expression of a number of genes encoding enzymes of the Escherichia coli phosphotransferase (PTS) and phosphoenolpyruvate (PEP) systems
MODE	Molybdate-responsive transcription factor
NAC	Nitrogen assimilation control
NAGC_new2	N-acetylglucosamine
NANR	N-acetyl-neuraminic acid regulator
NARL2	Nitrate/nitrite response regulator NarL
NARL	Nitrate/nitrite response regulator NarL
NARP	Nitrate/nitrite response regulator NarP
NIRC	NirC is a nitrite transporter which is a member of the FNT family of formate and nitrite transporters
OmpC	OmpC is a member of the GMP family
OxyR	Oxidative stress regulator
РНОВ	PhoB is a dual transcription regulator that activates expression of the Pho regulon in response to environmental Pi
РНОР	Member of the two-component regulatory system phoQ/phoP involved in adaptation to low Mg2+ environments and the control of acid resistance genes
PurR	PurR dimer controls several genes involved in purine nucleotide biosynthesis and its own synthesis
RcsB_1	Regulator capsule synthesis B
RcsB_2	Regulator capsule synthesis B
Rob2	Right origin-binding protein
ROB	Right origin-binding protein
soxS	SoxS is a dual transcriptional activator and participates in the removal of superoxide and nitric oxide and protection from organic solvents and antibiotics
TORR	TorR response regulator
TRPR	Tryptophan (trp) transcriptional repressor
TyrR	Tyrosine repressor

# **Building SITECON Model**

To build a new SITECON model call the *Tools->SITECON->Build new SITECON model from alignment* main menu item. The following dialog will appear:

U SITECON Build	? ×
Input alignment (nucleic) Output model	
Options	
Window size	40bp 👤
Calibration random seed	0
Calibration sequence length	100 Kb 🔻
Weight algorithm	Algorithm N2 💌
Default status	
	Build Cancel Help

Here you need to select a nucleotide alignment and an output model. Optionally, you can change other parameters. After that click on the *Bu ild* button.

## **Smith-Waterman Search**

The Smith-Waterman Search plugin adds a complete implementation of the Smith-Waterman algorithm to UGENE.

To use the plugin open a nucleotide or protein sequence in the Sequence View and select the Analyze Find pattern [Smith-Waterman] item in the context menu. The Smith-Waterman Search dialog appears:

• Smith-Waterman Sea	arch		(
Smith-Waterman parame	eters Input and output	ut	
Enter pattern here			
Search in	Strand		Region
Sequence	О Во	th	Whole sequence 😂
O Translation	Oir	ect	Whole sequence
	⊖ <b>Co</b>	mplement	1 - 199950
Smith-Waterman algor	ithm parameters		
Algorithm version	Scoring matrix	Gap scores	Results filtering strategy
Classic 2 🗘	dna 🗘	Gap open	-10 🗘 Report results filter-int 🗘
Advanced.	View	Gap extension	n -1 ☆ Minimal score 90% ☆
Help			Cancel Search

First of all you need to specify the pattern to search for. The rest parameters are optional:

Search in — select either to search in the sequence or in its translation.

Strand — select the strand to search in: direct, complementary or both strands.

Region - specifies the region of the sequence that will be used to search for the pattern. By default, if a subsequence has been

selected when the dialog has been opened, then the selected subsequence is searched for the pattern. Otherwise, the whole sequence is used. You can also input a custom range.

Algorithm version — version of the algorithm implementation. Non-classic versions produce the same results as classic but much faster. To use these optimizations our system must support these capabilities.

- Classic 2
- SSE2
- CUDA
- OPENCL

Scoring matrix - can be chosen from a bunch of matrices supplied with UGENE. To view a matrix selected click the View button.

Gap open - penalty for opening a gap.

Gap extension — penalty for extending a gap

*Report results* — simple heuristic which allows to filter intersected hits. If it is set to *none*, the algorithm may report large set of almost identical results in the same region.

*Minimal score* — another simple heuristic which measures sequences similarity. It is more convenient than using some abstract scores. If set to 100%, the algorithm will search for exact substring match.

The results of the search are saved as annotations or as multiple alignment. To set the saving parameters go to the *Input and output* tab of the dialog.

If you want to save the results as annotations input *the annotations saving* parameters (*Annotation name*, *Group name*, *Annotation type*, *Description* and a file to save the annotation to). Also you can add qualifier with corresponding pattern subsequences to result annotations. Check the corresponding checkbox for it.

If you want to save the results as multiple alignment select the following parameters:

Smith-Waterman	Search		8
Smith-Waterman pa	rameters Inpu	ut and output	
Save results as	Annotations		•
▼ Save annotation(	s) to		
O Existing table			\$ <b>E</b> }
Oreate new tag	able	/home/yalgaer/MyDocument.gb	
O Use auto-anno	otations table	Image: Solution of the second sec	
<ul> <li>Annotation parar</li> </ul>	neters		
Group name	<auto></auto>		<b>*</b>
Annotation type	Misc. Feature	2	
Annotation name	by type		
Description			
🗆 Add qualifier wil	h correspondir	ng pattern subsequences to result annotations	
Help			Cancel Search

Here you can select a file to save the alignment to (Alignment files directory path parameter).

Using the Set advanced options checkbox you can select the saving options.

You can set the different templates for files names: create your own or create by using the following: [E] — adds a subsequence end position, [hms] — adds a time, [MDY] — adds a date, [S] — adds a subsequence start position, [L] — adds a subsequence length, [SN] — adds a reference sequence name prefix, [PN] — adds a pattern sequence name prefix, [C] — adds a counter.

You can create templates for alignment files names, reference subsequence names, pattern subsequence names and for pattern sequence name:

	File Actio	ns Settir	ngs To	ols Wi	ndow	Help										-	5
	3 🗁 🛛			<b>()</b> ; ()	Ţ	Ý			9	9	٠	\$2	▶][⊲				
1: Project	Project Name filter Objects	uman_T1.fa	3	Refer	> T ence				ensu e	IS			A C	<b>G T A</b> 3 4 5	<b>C G -</b> 6 7 8		_
	P:	[s] huma L_human_1 [m] P1_h	[1_1.aln		2002	Tiu	man _1_7	T1 .	5727	4 57	281	[1 [1	A C	G T A G T A	CGT CG-		8] 7]
			-			Pat	ttern	sub	seq	uen	ce	/	•	-		Þ	
						Find	: 🗲		Alig	nme	nt		<b>-</b>	Ln 1/2	Col 1/8	Pos 1/	/8 (

## HMM2

The HMM2 plugin is a toolkit based on the Sean Eddy's HMMER2 package.

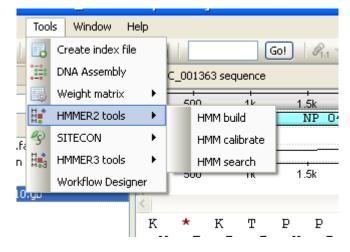
While working on this plugin we were guided by the following principles:

- Make the HMMER2 tools accessible to a wider user audience by providing graphical interface for all supported utilities for most of the platforms.
- Be compatible with the original HMMER2 package.
- Create the high-performance solution utilizing modern multi-core processors and SIMD instructions.

The current version of UGENE provides user interface for three HMM2 tools: HMM build, HMM calibrate and HMM search.

In the original program the corresponding commands are: "hmmbuild", "hmmcalibrate" and "hmmsearch".

To access these tools select the Tools HMMER2 tools submenu of the program main menu:



We highly recommend reading the original HMMER2 documentation to learn how to use utilities provided by the plugin.

SSE2 algorithm is implemented by Leonid Konyaev, Novosibirsk State University. Use of the SSE2 optimized version of the HMM search algorithm with quad-core CPU gives >30x performance boost when compared with the original single-threaded algorithm (single sequence mode).

- Building HMM2 Model
- Calibrating HMM2 Model
- Searching Sequence Using HMM2 Profile

## **Building HMM2 Model**

HMM build tool is used to build a new HMM profile from a multiple alignment.

You can use any alignment file formats supported by UGENE. The output HMM profile format is compatible with the HMMER2 package.

Multiple alignment file:	
File to save HMM profile:	
Expert options	
Name this HMM:	
Default (hmmls) behaviour:	0
Emulate hmmfs behaviour:	0
Emulate hmms behaviour:	0
Emulate hmmsw behaviour:	0
Build Close	Help

The HMM build tool does not automatically calibrate a profile. Use the HMM calibrate tool to calibrate the profile.

#### Calibrating HMM2 Model

The *HMM calibrate* tool reads a HMM profile file, scores a large number of synthesized random sequences with it, fits an extreme value distribution (EVD) to the histogram of those scores, and re-saves the hmm file including the EVD parameters.

To avoid modification of the original HMM file you can select a new location for the calibrated profile.

U HMM Calibrate	- ? <b>-</b> -	x
HMM file:		)
Expert options		
Fix the length of the random sequences to:	0	
Mean length of the synthetic sequences:	325 🌲	
Number of synthetic sequences:	5000 🚖	
Standard deviation:	200.00	
Random seed:	0	
Save calibrated profile to file:		
	Calibrate Close Help	

### Searching Sequence Using HMM2 Profile

The HMM search tool reads a HMM profile from a file and searches the sequence for significantly similar sequence matches.

The sequence must be selected in the Project View or there must be an active Sequence View window opened.

If the selected sequence is nucleic and the HMM profile is built for amino alignment, the sequence is automatically translated and all 6 translations are used to search in.

If a HMM profile is built for nucleic alignment, the search is performed for both strands (direct and complement).

8 HMM Search			? ×
File with HMM profile:			
▼ Save annotation(s	) to		
Existing table		🚱 NC_001363 features [murine.gb]	▼ 🌇
Create new ta	ble		
🔘 Use auto-anno	otations table		
<ul> <li>Annotation parame</li> </ul>	eters		
Group name	<auto></auto>		*
Annotation type	Misc. Signal		•
Annotation name	hmm_signal		
Description			
Expert options			
Filter results with E-	-value greater ti	ien:	1e-1 🌲
Filter results with So	ore lower than:		0.0 🌲
Number of sequence	es in dababase:		1 🔹
Algorithm		SSE of	ptimized 🔻
		Search Close	Help

The search results are stored as sequence annotations in the Genbank file format.

600096 NC_00096	64 sequence								🖄 1	t, C	]   C	T	6	₽, 🔍	, 🔃	۹, ۱
exon (4)																
gene (465)			I													
					po	nA										
hmm_signal (	2)			hmm	sigr	ıal										
intron (2)			i	I												
misc_RNA (2	)		1													
misc_feature	(1)		İ.	i i												
rRNA (30)																
scRNA(1)			1													
tRNA (86)	•••••		I													
2 342 436	2342.6k	2342.7k	234	2.8k	2342	.9k	2343	lk	2343	.1k	2343	.2k	2343.	3k	2 34	13 464
<																>
RGN	I S	* т	V	W	F	K	R	E	I	I	۲ ۲	. ]	R	2	5	I
ΕE	ΤА	K I	i S	G	L	N	V	' ł	ζ	Y	D	Κ	D	N	Q	S
RK	QΙ	N	С	L	V	*	т	*	Ν	т	I	K	т	I	N	$\mathbf{L}$
GAGGAAA	ACAGCT	AAACI	GTC	TGG	TTT.	AAA	CGT	GAI	AT.	ACC	GATA	AA	GAC	AT	CAA	TCT
2 342 750	234276	0	234	2770		2	34278	0		23	42790			2342.8	3k 23	342 80
<								)								
Name						Valu	Je									
	ations [MyD		o] *													
	nm_signal (															
	hmm_signa	l .					2756	23429	992							
							e-007									
	- HMM-mod	el				fn3										
	Score					22.0	_	2257								
	hmm_signa 00964 featu		661			225	7666	22575	944							
⊞ (¥ NC_00	Jugor redu	res (bsub.g	UK]													

All HMM2 UGENE tools work only with files that contain a single HMM model.

The HMM3 plugin is a toolkit based on the Sean Eddy's HMMER3 package.

While working on this plugin we were guided by the following principles:

- Make the HMMER3 tools accessible to a wider user audience by providing graphical interface for all supported utilities for most of the platforms.
- Be compatible with the original HMMER3 package.
- · Create the high-performance solution utilizing modern multi-core processors.

The current version of UGENE provides user interface for three HMM3 tools: HMM3 build, HMM3 search and Phmmer search.

In the original program the corresponding commands are: "hmmbuild", "hmmsearch" and "phmmer".

To access these tools select the Tools HMMER3 tools submenu of the program main menu:

	Tools	Window	, Help					
;	🐻 Cri	eate inde	× file	5	n Ba	ĒA		
-	📑 DN	IA Assemi	bly	F				
P	🐻 We	eight matr	rix 🕨 🕨	2	855134)		0	2
	HM	IMER2 to	ols 🕨	Ţ	80k		100k	
	🍫 SI	FECON	•	L				
	🕍 нм	IMER3 to	ols 🔹 🕨		Build pr	ofile		1
	We	orkflow De	esigner		Search	HMM	signals	Ŀ.
ï	P	ĸ	LF	ï.	Phmme	r sear	ch	n.
						~	17	R.T.

We highly recommend reading the original HMMER3 documentation to learn how to use utilities provided by the plugin.

- Building HMM Model
- Searching Sequence Using HMM Profile
- Searching Sequence Against Sequence Database

## **Building HMM Model**

The *HMM3 build* tool is used to build a new HMM profile from a multiple alignment. You can use any alignment file formats supported by UGENE.

The output HMM profile format is compatible with the HMMER3 package, but it is not compatible with the HMMER2.

The HMM3 build automatically calibrates the target model.

U	Hmm3 Build					? X
	Input and output	Construction strategies	Relative weighting	Effective weighting	E-value calibration	Other
	Input alignment file					
	Build to profile					
				Bu	ild Cancel	Help

The HMM3 configuration dialog provides an easy way to set appropriate search parameters.

Here you can see effective weighting strategies options:

U	Hmm3 Build				8	X
	Input and output	Construction strategies	Relative weighting	Effective weighting	E-value calibration Other	
	Adjust effective	e sequence number to achiev	e relative entropy targe	t		
	Minimum relativ	e entropy/position			0.00	*
	Sigma paramete	er			45.00	*
	O Use number of	single linkage clusters as effe	ctive		Fractional identity cutoff 0.62	×
	O Use number of	sequences as effective				
	Effective seque	ence number for all models to			0.01	A V
				Ви	ild Cancel Help	>

## **Searching Sequence Using HMM Profile**

The HMM3 search tool reads a HMM profile from a file and searches a sequence for significantly similar sequence matches.

The sequence must be selected in the Project View or there must be an active Sequence View window opened.

If the selected sequence is nucleic and profile HMM is built from amino alignment, the sequence will be automatically translated and searched in all possible frames (6 totally).

If a profile HMM is built for nucleic alignment, the search is performed for both strands (direct and complement).

The *HMM3 search* accepts the HMMER2 HMM profiles (amino only) as a backward compatibility feature. An interesting post about using the HMMER2 models with the HMMER3 is available on the Sean Eddy's blog.

HMM3 Search		₹ ?
Input and output	Reporting thresholds Acceleration heuristics Other	
Query profile HMM f	le:	
<ul> <li>Save annotation</li> </ul>	ís) to	
Existing table	• NC_014267 features [NC_014267.1.gb]	-
Create new	able	
🔘 Use auto-an	notations table	
<ul> <li>Annotation para</li> </ul>	neters	
Group name	<auto></auto>	*
Annotation type	Misc. Signal	•
Annotation name	hmm_signal	
Description		
	Run Cancel	Help

For example, reporting thresholds options can be configured using the dialog:

U HMM3 Search	? ×
Input and output Reporting thresholds Acceleration heuristics Other	
Report domains with E-value less than	1E+1 束
Report domains with score greater than	
Score threshold:	0.01
Use profile's GA gathering cutoffs	
Use profile's NC noise cutoffs	
<ul> <li>Use profile's TC trusted cutoffs</li> </ul>	
Number of significant sequences for domain E-value calculation	0.00
Run	Cancel Help

The search results are stored as sequence annotations in the Genbank file format.

	200k	400k	600k	800k	1m	1.2	m 1.4m	1.6m	1.8m	2m	2.2m	2.4m	2.6m	2.8m		
I	2006	4000	OUOK	UUUK		1.2		1.00	1.011	2111	2.2111	2.40	2.011	2.01		
TΥ	PDP	PENV	KWRDR	FANSI	FLTWI	D P P I	KNDGGSR	IKGYI	VERC	PRGS	DKWV	ACGEE	VAETI	KMEV		
088		7095	7.1k	7105	7110		15 7120	7125	7130	) 7	135	7140	7145	7150		
ame							Value									
🔇	Anno	tations [I	MyDocument	:_3.gb]*												
Ė	3 🧭 H	nmm_sign	al (0, 24024	ŧ)												
	÷	🛛 hmm_si	gnal				65946679									
		🛾 hmm_si	-			66956781										
		hmm_si	-				67966882									
		] hmm_si	-				69927076									
	Ē	hmm_si	-				70927177									
			acy per resid	lue			9.76351e-01									
		Bias					3.53754e-02									
			ional e-valu	-			5.96204e-17									
		Enveic	pe of doma	in location			70917177 fn3 Accession				00041					
		HMM r					187	number in i	-ram uaca	Dase; Pr	00041					
		-	egion endent e-va	lua			107 1.89874e-17									
		Score					49.864132									
		hmm_si	anal				72887372									
		hmm si	-				73877473									

() The HMM3 search works only with files that contain a single HMM model.

## Searching Sequence Against Sequence Database

The Phmmer search tool searches for query sequence matches in sequence database, much as BLASTP or FASTA would do.

The Phmmer search works essentially like the HMM3 search does, except you provide a query sequence instead of a query profile HMM.

The database sequence must be selected in the Project View or there must be an active Sequence View window opened.

Select the query sequence in the *Phmmer search* dialog:

Phmmer Search					? X
Input and output	Reporting tresholds	Scoring system	Acceleration	E-value calibration	Other
Query sequence file:					
<ul> <li>Save annotation(s</li> </ul>	s) to				
Existing table					-
Oreate new tag	able C:/Users/yalgaer	/MyDocument.gb			
<ul> <li>Annotation param</li> </ul>	eters				
Group name	<auto></auto>				*
Annotation name	signal				atc
			Sear	rch Cancel	Help

You can set options of the *Phmmer search* by choosing the needed dialog tab. Here you can see the e-value calibration options:

U Phmmer Search					2	x	
Input and output	Reporting tresholds	Scoring system	Acceleration	E-value calibration	Other		
Length of sequence	Length of sequences for MSV Gumbel mu fit						
Number of sequence	es for MSV Gumbel mu fit				200	-	
Length of sequence	s for Viterbi Gumbel mu fi	t			200	-	
Number of sequence	es for Viterbi Gumbel mu	fit			200	-	
Length of sequence	s for Forward exp tail mu	fit			100	-	
Number of sequence	es for Forward exp tail m	u fit			200	-	
Tail mass for Forwar	d exponential tail mu fit				0.0	04 🚔	
			Sear	ch Cancel	He	elp	

The results are stored as sequence annotations in the Genbank file format.

<b>NILLIN</b>										
20	Ok 400k	600k	800k	im	1.2m	1.4m	1.6m	1.8m	2m	2.2m
HEGM	IEYTFRV	SAENK	Y <mark>gvg</mark>	EGLK	SEPIV.	ARHPF	DVPD/	ABBBB	IIVDV	RHDS
40 569	40575 40	580 40	585	40590	40595	40.6k	40605	40610	40615	406
lame					🛆 🛛 Valu	ie				
	- signal - signal - signal - signal - signal - signal	46) y per residi	19		104 136 367 373 405	19856 7110549 6013717 0736782 9737475 8640643 2474e-01				
	Bias Conditio Envelop HMM re: HMM re: Indeper Query s Score	nal e-value e of domair gion ndent e-valu	n location		1.38 1.34 405 804 1.34 fibro -3.5	3025e-02 4634e-01 7940645 858 4634e-01 pnectin_1.2 i13604				
	- 🗖 signal - Accurac - Bias	y per residi	le		7.36	3363708 5235e-01 7314e-03				

The Phmmer search works only with single-sequence databases.

## uMUSCLE

UGENE contains graphical ports of the Robert C. Edgar's MUSCLE tool for multiple alignment.

```
MUSCLE4 is not supported since UGENE version 1.7.2.
∕∿
```

The package is integrated completely, so there is no need in extra files for using it. It is possible to run several multiple alignment tasks in parallel, check the progress and cancel the running tasks safely.

The k-mer clustering part of the MUSCLE algorithm was optimized for multicore systems by Timur Tleukenov, Novosibirsk State ∕∿ Technical University.

- MUSCLE Aligning
- Aligning Profile to Profile with MUSCLE
  Aligning Sequences to Profile with MUSCLE

## **MUSCLE** Aligning

To run the classic MUSCLE use the Align Align with MUSCLE context menu item in the Alignment Editor.

G		Edit 🕨	ΓA	CCACTAATATTAGG	A
G			1-1	<u> </u>	A
G		Align 🕨 🕨	м	Align with MUSCLE	ç
G	ь	Statistics •	М	Align sequences to profile with MUSCLE	Ģ
G		View 🕨	М	Align profile to profile with MUSCLE	e
G		Advanced •	Κ	Align with Kalign	Ģ

The dialog contains the list of MUSCLE modes: MUSCLE default, Large alignment, Refine only.

Align with MUSCLE	? ×
Mode MUSCLE default	•
Mode details:	
The default settings are designed to give the best accuracy	
Command line: muscle	
Advanced options	
☑ Do not re-arrange sequences (-stable)	
Max iterations	2 🚊
Max time (minutes)	5 🚔
Translate to amino when aligning	
Translation table: 1. The Standard Genetic Code	
Region to align	
Whole alignment	
○ Column range 1 ▲ -	604 🚖
Align Cancel	Help

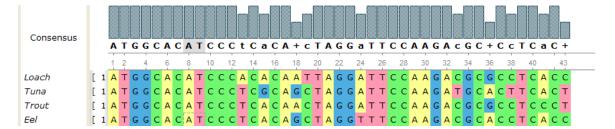
By default UGENE does not rearrange sequence order in an alignment, but the original MUSCLE package does. To enable sequence rearrangement uncheck the Do not re-arrange sequences (-stable) option in the dialog.

One of the improvements to the original MUSCLE package is the ability to align only a part of the model. When the *Column range* item is selected the region of the specified columns is only passed to the MUSCLE alignment engine. The resulted alignment is inserted into the original one with gaps added or removed on the region boundaries.

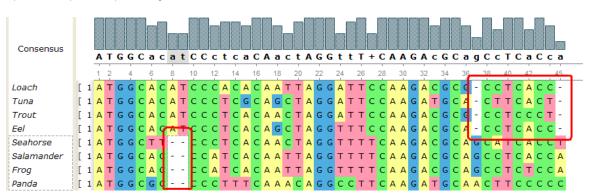
To visually select the column range to align, make a selection in the alignment editor first. Then invoke the MUSCLE plugin. Its column range boundary values will automatically match the given selection.

## Aligning Profile to Profile with MUSCLE

The Align Align profile to profile with MUSCLE context menu item allows to align an existing profile to an active alignment. During this process the MUSCLE does not realign the profiles, but inserts columns with gaps characters only ('—' characters). For example, the alignment in the picture below could be used as a profile:



The same profile after profile-to-profile alignment:



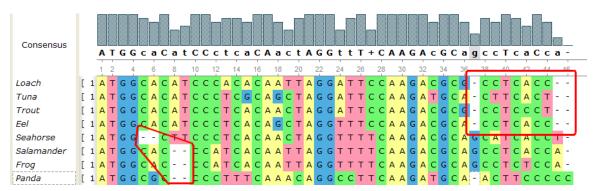
There are two gap columns inserted into the source profile, and two gap columns inserted into the added one. Therefore the profiles columns kept intact and the alignments haven't been changed.

Aligning a profile to the active alignment you will modify the original alignment file, since it will contain 2 profiles after the operation is completed.

## **Aligning Sequences to Profile with MUSCLE**

Another feature provided by the plugin is aligning a set of unaligned sequences to an existing profile. To use this feature select the *Align Align sequences to profile with MUSCLE* context menu item.

This option is not available in the original MUSCLE package (v3.7) and is a new functionality for original MUSCLE users. In this mode each sequence from the input file is aligned to the active profile separately and is merged to the result alignment only after all sequences are processed. For example, the alignment in the picture above can be used as a profile again. And the added profile can be used as a set of sequences. The result of such sequences-to-profile alignment is presented on the picture below:



The original alignment is not modified, only columns with gap ('--') character can be inserted.

The second profile was considered as a set of sequences and therefore is modified.

Note that if a file with another alignment is used as a source of unaligned sequences, the gap characters are removed and each input sequence is processed independently.

This method is quite fast, for example an alignment of 3000 sequences (1000 bases each) to the existing profile takes about 5 minutes on the usual Core2Duo computer.

## ClustalW

Clustal is a widely used multiple sequence alignment program. It is used for both nucleotide and protein sequences. *ClustalW* is a command-line version of the program.

#### Clustal home page: http://www.clustal.org

If you are using Windows OS, there are no additional configuration steps required, as *ClustalW* executable file is included to the UGENE distribution package. Otherwise:

- Install the *Clustal* program on your system.
- Set the path to the ClustalW executable on the External tools tab of UGENE Application Settings dialog.

Now you are able to use *Clustal* from UGENE.

Open a multiple sequence alignment file and select the *Align with ClustalW* item in the context menu or in the *Actions* main menu. The *Align with ClustalW* dialog appears (see below), where you can adjust the following parameters:

Gap opening penalty - cost of opening up a new gap in the alignment. Increasing this value will make gaps less frequent.

Gap extension penalty - cost of every item in a gap. Increasing this value will make gaps shorter.

Weight matrix — specifies a single weight matrix for nucleotide sequences or series of matrices for protein sequences.

For nucleotide sequences the weight matrix selected defines the scores assigned to matches and mismatches (including IUB ambiguity codes), it can take values:

- IUB default scoring matrix used by BESTFIT for the comparison of nucleic acid sequences. X's and N's are treated as matches to any IUB ambiguity symbol. All matches score 1.9; all mismatches for IUB symbols score 0.
- CLUSTALW previous system used by ClustalW, in which matches score 1.0 and mismatches score 0. All matches for IUB symbols also score 0.

For protein sequences it describes the similarity of each amino acid to each other. The following values are available:

- BLOSUM BLOcks of Amino Acid SUbstitution Matrices first introduced in a paper by Henikoff and Henikoff. These matrices appear to be the best available for carrying out data base similarity (homology searches).
- PAM Point Accepted Mutation matrices introduced by Margaret Dayhoff. These have been extremely widely used since the late '70s.
- GONNET these matrices were derived using almost the same procedure as the Dayhoff one (above) but are much more up to date and are based on a far larger data set. They appear to be more sensitive than the Dayhoff series.
- ID identity matrix which gives a score of 1.0 to two identical amino acids and a score of zero otherwise.

Iteration type — specifies the iteration type to use. During the iteration step each sequence is removed in turn and realigned. It is kept if the resulting alignment is better than the one has been made before. This process is repeated until the score converges or until the maximum number of iterations is reached. Available values are:

- NONE specifies not to use iterations.
- TREE specifies to iterate at each step of the progressive alignment.
- ALIGNMENT specifies to iterate on the final alignment.

Max iterations - maximum number of iterations.

Align with ClustalW	8
Advanced options	
Gap opening penalty	15.00
Gap extension penalty	6.66
Weight matrix	IUB 👻
Iteration type	NONE
Max iterations	3
Out sequences order	Input
Protein gap parameters	
Gap separation distance	4
Hydrophilic gaps off	
□ No end gap separation penalty	
Residue-specific gaps off	
Help	Cancel Align

The following parameters are only available for protein sequences:

Gap separation distance — tries to decrease the chances of gaps being too close to each other. Gaps that are less than this distance apart are penalized more than other gaps. This does not prevent close gaps; it makes them less frequent, promoting a block-like appearance of the alignment.

Hydrophilic gaps off — increases the chances of a gap within a run of hydrophilic amino acids.

No end gap separation penalty - treats end gaps just like internal gaps to avoid gaps that are too close.

Residue-specific gaps off — amino acid specific gap penalties that reduce or increase the gap opening penalties at each position in the alignment or sequence. For example, positions that are rich in glycine are more likely to have an adjacent gap than positions that are rich in valine.

## MAFFT

Originally, MAFFT is a multiple sequence alignment program for unix-like operating systems. However, currently it is available for Mac OS X, Linux and Windows. It is used for both nucleotide and protein sequences.

#### MAFFT home page: http://mafft.cbrc.jp/alignment/software

To make MAFFT available from UGENE:

- Install the MAFFT program on your system.
- Set the path to the MAFFT executable on the External tools tab of UGENE Application Settings dialog.

For example, on Windows you need to specify the path to the mafft.bat file.

To use *MAFFT* open a multiple sequence alignment file and select the *Align with MAFFT* item in the context menu or in the *Actions* main menu. The following dialog appears:

Align with MAFFT	8
Advanced options	
Gap opening penalty	1.53
Offset (works like gap extension penalty)	0.00
Maximum number of iterative refinement	0
Help Cancel	Align

The following parameters are available:

Gap opening penalty - Gap opening penalty at group-to-group alignment.

Offset (works like gap extension penalty) — offset value, which works like gap extension penalty, for group-to-group alignment.

Maximum number of iterative refine — specifies the number of cycles of iterative refinement to perform.

## **T-Coffee**

T-Coffee is a multiple sequence alignment package.

#### T-Coffee home page: T-Coffee

To make T-Coffee available from UGENE see the External Tools.

To use *T*-Coffee open a multiple sequence alignment file and select the Align with *T*-Coffee item in the context menu or in the Actions main menu. The following dialog appears:

Align with T-Coffee	8
Advanced options	
Gap opening penalty	-50
Gap extension penalty	0
Number of iterations	0
Help	Cancel Align

The following parameters are available:

Gap opening penalty — indicates the penalty applied for opening a gap. The penalty must be negative.

Gap extension penalty — indicates the penalty applied for extending a gap.

Number of iterations - specifies the number of iterations.

## Bowtie

Bowtie is a popular short read aligner. Click this link to open Bowtie homepage. Bowtie is embedded as an external tool into UGENE.

Open Tools DNA Assembly submenu of the main menu.

Too	ols Window Help				
1	Sanger data analysis		1		
1	NGS data analysis			Reads quality control	ł
S	BLAST		i:	Genome de novo assembly	l
<b>*</b>	Multiple sequence alignment	1.0	] ]	Map reads to reference Build index for reads mapping	ļ
	Primer •	1			

Select the *Align short reads* item to align short reads to a DNA sequence using *Bowtie*. Or select the *Build index* item to build an index for a DNA sequence which can be used to optimize aligning of the short reads to the sequence.

- Bowtie Aligning Short Reads
- Building Index for Bowtie

## **Bowtie Aligning Short Reads**

When you select the Tools DNA Assembly Align short reads item in the main menu, the Align Short Reads dialog appears. Set value of the Align short reads method parameter to Bowtie. The dialog looks as follows:

U Align Sequencing Reads					2 ×
Alignment method		Bo	owtie		•
Reference sequence Result file name Library Single-end  P Short reads	rebuilt index				··· ✓ SAM output
Path				Туре	Order
Parameters			Flags	Add	d Remove
Mode: -n mo	de 🔻 2	* *	Colorspace		
Maq error (maqerr)	70	* *		unding (nomaqrou	
Seed length (seedlen)	28	* *		l orientation (nof -complement orien	
Maximum of backtracks (	maxbts) 800	*		l (tryhard)	(
Descriptors memory usag	e (chunkmbs) 64	* *		ents (best)	
Seed (seed)	0	A V	All alignmer	nts (all)	
			Sta	art Cano	el Help

There are the following parameters:

Reference sequence - DNA sequence to align short reads to. This parameter is required.

Result file name — file in SAM format to write the result of the alignment into. This parameter is required.

Library - single-end or paired-end reads.

Prebuilt index — check this box to use an index file instead of a source reference sequence. The index is a set of 6 files with suffixes .1.ebwt, .2.ebwt, .3.ebwt, .4.ebwt, .rev.1.ebwt, and .rev.2.ebwt. The index is created during the alignment. Also you can *build it manually*.

SAM output — always save the output file in the SAM format (the option is disabled for Bowtie).

Short reads — each added short read is a small DNA sequence file. At least one read should be added.

Short reads length for Bowtie can't be more than 1024.

You can also configure other parameters. They are the same as in the original *Bowtie* (you can read detailed description of the parameters on the Bowtie manual page).

Select one of the following alignment modes:

The -n alignment mode:

When the *-n mode* is selected, *Bowtie* determines which alignments are valid according to the following policy. Alignments may have no more than N mismatches (where N is a number 0-3) in the first L bases (where L is a number 5 or greater, set with *Seed length*) on the high-quality (left) end of the read. The sum of the Phred quality values at all mismatched positions (not just in the seed) may not exceed E (set with *Maq error*). Where qualities are unavailable (e.g. if the reads are from a FASTA file), the Phred quality defaults to 40.

The -v alignment mode:

In *-v mode*, alignments may have no more than V mismatches, where V may be a number from 0 through 3. Quality values are ignored. The *-v mode* is mutually exclusive with the *-n mode*.

The following parameters are available:

Maq error (-maqerr) — maximum permitted total of quality values at all mismatched read positions throughout the entire alignment, not just in the "seed". The default is 70. By default, *Bowtie* rounds quality values to the nearest 10 and saturates at 30. Note that the rounding can be disabled with *No Maq rounding*.

Seed Length (-seedlen) — the number of bases on the high-quality end of the read to which the -n applies. The lowest permitted setting is 5 and the default is 28.

*Maximum of backtracks (-maxbts)* — the maximum number of backtracks (default: 125 without *Best*, 800 with *Best*). A "backtrack" is the introduction of a speculative substitution into the alignment.

Descriptors memory usage (-chunkmbs) — the number of megabytes of memory a given thread is given to store path descriptors in the Best flag. Default: 64. This parameter is available if the Best flag is checked.

Seed (-seed) - pseudo-random number generator.

The following flags are available:

Colorspace (-color) — the input is read in colorspace, colors are encoded as characters A/C/G/T (A=blue, C=green, G=orange, T=red).

No Maq rounding (-nomaqround) — Maq (Mapping and Assembly with Quality) accepts quality values in the Phred quality scale, but internally rounds values to the nearest 10, with a maximum of 30. By default, *Bowtie* also rounds this way. No Maq rounding prevents this rounding in *Bowtie*.

No forward orientation (-nofw) - do not attempt to align against the forward reference strand.

No reverse-complement orientation (-norc) — do not attempt to align against the reverse-complement reference strand.

Try as hard (-tryhard) - try as hard as possible to find valid alignments when they exist, including paired-end alignments.

Best alignments (-best) — make Bowtie guarantee that reported singleton alignments are "best" in terms of stratum (i.e. number of mismatches, or mismatches in the seed for the case of -n mode) and in terms of the quality values at the mismatched position(s).

All alignments (-all) — report all valid alignments per read or pair. Validity of alignments is determined by the alignment policy (combined effects of -n mode, -v mode, Seed length, and Maq error).

Select the required parameters and press the Start button.

#### **Building Index for Bowtie**

To build Bowtie index select the Tools NGS data analysis Build index for reads mapping item in the main menu. The Build Index dialog appears. Set the Align short reads method parameter to Bowtie.

The dialog looks as follows:

U Build Index			? <mark>x</mark>
Align short reads meth	od Bowt	ie	•
Reference sequence Index file name			
	Start	Cancel	Help

There are the following parameters:

Reference sequence — DNA sequence to which short reads would be aligned to. This parameter is required.

Index file name - a file to save the created index to. This parameter is required.

Colorspace (-color) — the input is read in colorspace, colors are encoded as characters A/C/G/T (A=blue, C=green, G=orange, T=red).

## Bowtie 2

Bowtie 2 is a popular ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences. Click this link to open Bowtie 2 homepage. Bowtie 2 is embedded as an external tool into UGENE.

Open Tools Align to reference submenu of the main menu.

Tools Window Help							
	Sanger data analysis	×					
	NGS data analysis	۲		Reads quality control			
-	BLAST			Genome de novo assembly			
<b>1</b>	Multiple sequence alignment Cloning Primer			Map reads to reference Build index for reads mapping			

Select the Align short reads item to align short reads to a DNA sequence.

Or select the Build index item to build an index for a DNA sequence which can be used to optimize aligning of the short reads to the sequence:

- Bowtie 2 Aligning Short ReadsBuilding Index for Bowtie 2

## **Bowtie 2 Aligning Short Reads**

When you select the Tools Align to reference Align short reads item in the main menu, the Align Sequencing Reads dialog appears. Set value of the Align short reads method parameter to Bowtie 2. The dialog looks as follows:

U Align Sequencing Reads				? ×
Alignment method		Bowtie2		•
Reference sequence				
Result file name				
Library Single-end 💌 🔲 Prebuilt index	¢			SAM output
Short reads				
Path			Туре	Order
Parameters Mode	end-to-end ▼	Flags	ignments (no-m	
Number of mismatches	0	No discordant	alignments (no-	discordant)
Seed length (L)	20	No forward or	entation (nofw)	)
Add columns to allow gaps (dpad)	15 *	No reverse-co	mplement orienta	tion (norc)
Disallow gaps (gbar)	4	No overlapping	g mates (no-ove	erlap)
Seed (seed)	0	No mates cont	aining one anothe	er (no-contain)
		St	art Can	Help

There are the following parameters:

Reference sequence - DNA sequence to align short reads to. This parameter is required.

Result file name — file in SAM format to write the result of the alignment into. This parameter is required.

Library - single-end or paired-end reads.

Prebuilt index — check this box to use an index file instead of a source reference sequence. The index is a set of 6 files with suffixes .1.ebwt, .2.ebwt, .3.ebwt, .4.ebwt, .rev.1.ebwt, and .rev.2.ebwt. The index is created during the alignment. Also you can *build it manually*.

SAM output - always save the output file in the SAM format (the option is disabled for Bowtie).

Short reads — each added short read is a small DNA sequence file. At least one read should be added.

You can also configure other parameters. They are the same as in the original *Bowtie 2* (you can read detailed description of the parameters on the Bowtie 2 manual page).

Select one of the following alignment modes:

The --end-to-end alignment mode:

By default, Bowtie 2 performs end-to-end read alignment. That is, it searches for alignments involving all of the read characters. This is also called an "untrimmed" or "unclipped" alignment.

When the --local option is specified, Bowtie 2 performs local read alignment. In this mode, Bowtie 2

# might "trim" or "clip" some read characters from one or both ends of the alignment if doing so maximizes the alignment score.

#### The following parameters are available:

*Number of mismatches (--N)* — sets the number of mismatches to allowed in a seed alignment during multiseed alignment. Can be set to 0 or 1. Setting this higher makes alignment slower (often much slower) but increases sensitivity.

Seed length (-L) — Sets the length of the seed substrings to align during multiseed alignment. Smaller values make alignment slower but more senstive.

Add columns to allow gaps (--dpad) — "Pads" dynamic programming problems by <int> columns on either side to allow gaps.

*Disallow gaps (-gbar)* — disallow gaps within <int> positions of the beginning or end of the read.

Seed (-seed) — use <int> as the seed for pseudo-random number generator.

#### The following flags are available:

*No unpaired alignments (-no-mixed)* — by default, when bowtie2 cannot find a concordant or discordant alignment for a pair, it then tries to find alignments for the individual mates. This option disables that behavior.

*No discordant alignments (-no-discordant)* — by default, bowtie2 looks for discordant alignments if it cannot find any concordant alignments. A discordant alignment is an alignment where both mates align uniquely, but that does not satisfy the paired-end constraints. This option disables that behavior.

No forward orientation (-nofw) — if --nofw is specified, bowtie2 will not attempt to align unpaired reads to the forward (Watson) reference strand.

No reverse-complement orientation (-norc) — if --norc is specified, bowtie2 will not attempt to align unpaired reads against the reverse-complement (Crick) reference strand.

No overlapping mates (-no-overlap) - if one mate alignment overlaps the other at all, consider that to be non-concordant.

No mates containing one another (-no-contain) - if one mate alignment contains the other, consider that to be non-concordant.

Select the required parameters and press the Start button.

#### **Building Index for Bowtie 2**

To build *Bowtie 2* index select the *Tools NGS data analysis Build index for reads mapping* item in the main menu. The *Build Index* dialog appears. Set the *Align short reads method* parameter to *Bowtie 2*.

The dialog looks as follows:

U Build Index			? ×
Align short reads meth	od (	Bowtie2	•
Reference sequence Index file name			
	Sta	rt Cancel	Help

There are the following parameters:

Reference sequence — DNA sequence to which short reads would be aligned to. This parameter is required.

Index file name - a file to save the created index to. This parameter is required.

## **BWA**

*BWA* is a fast light-weighted tool that aligns relatively short reads to a reference sequence. Click this link to open *BWA* homepage. *BWA* is embedded as an *external tool* into UGENE.

Open Tools DNA assembly submenu of the main menu.

Тос	ols Window Help		
1	Sanger data analysis		
	NGS data analysis		Reads quality control
S	BLAST		Genome de novo assembly
	Multiple sequence alignment	-	 Map reads to reference
0	Cloning •	-	
	Primer •		Build index for reads mapping

Select the *Align short reads* item to align short reads to a DNA sequence using *BWA*. Or select the *Build index* item to build an index for a DNA sequence which can be used to optimize aligning of short reads.

- Aligning Short Reads with BWA
- Building Index for BWA

## Aligning Short Reads with BWA

When you select the Tools DNA Assembly Align short reads item in the main menu, the Align Short Reads dialog appears. Set value of the Align short reads method parameter to BWA. The dialog looks as follows:

Align Sequencing Read	ds			8
Alignment method		BWA		0
Reference sequence				
Result file name				
Library Single-end 😂				☑ SAM output
Short reads				
Path		T	уре	Order
Base Options Advanced			Add	Remove
○ Max #diff (-n)	0	Max gap opens (-o)	1	
Missing prob (-n)	0.04	Index algorithm (-a)	autodet	
Seed length (-l)	32	Best hits (-R)	30	<b>^</b>
Non-iterative mode	nalty for long deletions ( · (-N)	-L)		
Help			Cancel	Start

There are the following parameters:

Reference sequence — DNA sequence to align short reads to. This parameter is required.

Result file name — file in SAM format to write the result of the alignment into. This parameter is required.

Library - single-end or paired-end reads.

Prebuilt index — check this box to use an index file instead of a source reference sequence. Also you can build it manually.

SAM output - always save the output file in the SAM format (the option is disabled for BWA).

Short reads - each added short read is a small DNA sequence file. At least one read should be added.

You can also configure other parameters. They are the same as in the original *BWA* (you can read detailed description of the parameters on the BWA manual page). Select one of the following parameters, that correspond to the *-n* option in the original *BWA*.

Max #diff (-n) — maximum edit distance. An integer value should be input.

Missing prob (-n) — the fraction of missing alignments given 2% uniform base error rate. A float value is used.

Seed length (-I) — take the subsequence of the specified length as seed. If the specified length is larger than the query sequence, seeding will be disabled. For long reads, this option is typically ranged from 25 to 35.

*Max gap opens (-o)* — maximum number of gap opens.

Index algorithm (-a) — algorithm for constructing BWA index.

It implements three different algorithms:

- is designed for short reads up to ~200bp with low error rate (<3%). It does gapped global alignment w.r.t. reads, supports
  paired-end reads, and is one of the fastest short read alignment algorithms to date while also visiting suboptimal hits.</li>
- bwtsw is designed for long reads with more errors. It performs heuristic Smith-Waterman-like alignment to find high-scoring local hits. Algorithm implemented in BWA-SW. On low-error short queries, BWA-SW. is slower and less accurate than the *is* algorithm, but on long reads, it is better.
- div does not work for long genomes.

Best hits (-R) — proceed with suboptimal alignments if there are no more than specified number of equally best hits. This option only affects paired-end mapping. Increasing this threshold helps to improve the pairing accuracy at the cost of speed, especially for short reads (~32bp).

Long-scaled gap penalty for long deletion (-L) — long-scaled gap penalty for long deletion.

Non-iterative mode (-N) — disable iterative search. All hits with no more than Max #diff differences will be found. This mode is much slower than the default.

You can also configure the following advanced parameters:

Enable long gaps — checking this box allows one to set the Max gap extentions parameter.

Max gap extensions (-e) — maximum number of gap extensions.

Indel offset (-i) — disallow insertions and deletions within the specified number of base pairs towards the ends.

Max long deletion extensions (-d) — disallow a long deletions within the specified number of base pairs towards the 3'-end.

Max queue entries (-m) — maximum queue entries.

Barcode length (-B) — length of barcode starting from the 5`-end. When the specified length is positive, the barcode of each read will be trimmed before mapping and will be written at the BC SAM tag. For paired-end reads, the barcode from both ends are concatenated.

Threads (-t) — number of threads.

Max seed differences (-k) — maximum edit distance in the seed.

Mismatch penalty (-M) — BWA will not search for suboptimal hits with a score lower than the specified value.

Gap open penalty (-O) — gap open penalty.

Gap extension penalty (-E) — gap extension penalty.

Quality threshold (-q) — parameter for read trimming.

Select the required parameters and press the Start button.

## **Building Index for BWA**

To build *BWA* index select the *Tools NGS* data analysis Build index for reads mapping item in the main menu. The Build Index dialog appears. Set the *Align short reads method* parameter to *BWA*.

The dialog looks as follows:

U Build Index	8 ×
Align short reads method	BWA 👻
Reference sequence	
Index file name	
Index algorithm (-a)	is 🔹
Colorspace (-c)	
	Start Cancel Help

There are the following parameters:

Reference sequence — DNA sequence to which short reads would be aligned to. This parameter is required.

Index file name - file to save index to. This parameter is required.

Index algorithm (-a) — Algorithm for constructing BWA index. Available options are:

It implements three different algorithms

- is designed for short reads up to ~200bp with low error rate (<3%). It does gapped global alignment w.r.t. reads, supports
  paired-end reads, and is one of the fastest short read alignment algorithms to date while also visiting suboptimal hits.</li>
- bwtsw is designed for long reads with more errors. It performs heuristic Smith-Waterman-like alignment to find high-scoring local hits. Algorithm implemented in BWA-SW. On low-error short queries, BWA-SW. is slower and less accurate than the *is* algorithm, but on long reads, it is better.
- *div* does not work for long genomes.

Colorspace (-color) — the input is read in colorspace, colors are encoded as characters A/C/G/T (A=blue, C=green, G=orange, T=red).

## **BWA-SW**

*BWA* is a fast light-weighted tool that aligns relatively short reads to a reference sequence. Click this link to open *BWA* homepage. BWA-SW share similar features such as long-read support and split alignment. *BWA-SW* is embedded as an *external tool* into UGENE.

Open Tools Align to reference submenu of the main menu.

Тоо	Tools Window Help								
	Sanger data analysis								
1	NGS data analysis		Reads quality control						
S	BLAST •		Genome de novo assembly						
	Multiple sequence alignment		Map reads to reference						
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	Cloning	1	Build index for reads mapping						
	Primer •		build index for reads mapping						

Select the *Align short reads* item to align short reads to a DNA sequence using *BWA-SW*. Or select the *Build index* item to build an index for a DNA sequence which can be used to optimize aligning of short reads.

- Aligning Short Reads with BWA-SW
- Building Index for BWA-SW

## Aligning Short Reads with BWA-SW

When you select the Tools Align to reference Align short reads item in the main menu, the Align Sequencing Reads dialog appears. Set value of the Align short reads method parameter to BWA-SW. The dialog looks as follows:

Align Sequencing Reads			<u> </u>
Alignment method		BWA-SW	•
Reference sequence			
Result file name			
ibrary Single-end 🔻			✓ SAM output
Short reads			
Path		Туре	Order
Base Options		Add	Remove
Index algorithm (-a)	bwtsw 🔻	Number of threads (-t)	8
Score for a match (-a)	1	Size of chunk of reads (-s)	1000000 ≑
Mismatch penalty (-b)	3	Score threshold (divided by match score) (-T)	30 🚖
Gap open penalty (-q)	5 🗘	Z-best (-z)	1
Gap extention penalty (-r)	2	Number of seeds to start rev alginment (-N)	5
Band width (-w)	50 🌲	Mask level (-c)	0.50
		Prefer hard clipping in SAM output (-H)	
NOTE: bwa-sw performs ali in FASTA or FASTQ format.	ignment of long Reads should b	sequencing reads (Sanger or 454). It acce be compiled into single file.	pts reads only
		Start Cancel	Help

There are the following parameters:

Reference sequence - DNA sequence to align short reads to. This parameter is required.

Result file name - file in SAM format to write the result of the alignment into. This parameter is required.

SAM output — always save the output file in the SAM format (the option is disabled for BWA).

Short reads - each added short read is a small DNA sequence file. At least one read should be added.

You can also configure other parameters.

Index algorithm (-a) — algorithm for constructing BWA-SW index.

## **Unipro UGENE Manual, Version 1.30**

It implements three different algorithms:

- is designed for short reads up to ~200bp with low error rate (<3%). It does gapped global alignment w.r.t. reads, supports
  paired-end reads, and is one of the fastest short read alignment algorithms to date while also visiting suboptimal hits.</li>
- *bwtsw* is designed for long reads with more errors. It performs heuristic Smith-Waterman-like alignment to find high-scoring local hits. Algorithm implemented in BWA-SW. On low-error short queries, *BWA-SW*. is slower and less accurate than the *is* algorithm, but on long reads, it is better.
- div does not work for long genomes.

Score for a match (-a) - score of a match.

Mismatch penalty (-b) — mismatch penalty.

Gap open penalty (-q) — gap open penalty.

Gap extention penalty (-r) — Gap extension penalty. The penalty for a contiguous gap of size k is q+k\*r.

Band width (-w) - Band width in the banded alignment.

Number of threads (-t) - Number of threads in the multi-threading mode.

Size of chunk of reads (-s) - Maximum SA interval size for initiating a seed. Higher -s increases accuracy at the cost of speed.

Score threshold (divided by much score) (-T) - minimum score threshold.

Z-best (-z) - Z-best heuristics. Higher -z increases accuracy at the cost of speed.

Number of seeds to start rev alignment (-N) - Minimum number of seeds supporting the resultant alignment to skip reverse alignment.

*Mask level (-c)* - Coefficient for threshold adjustment according to query length. Given an I-long query, the threshold for a hit to be retained is  $a^*max\{T,c^*log(I)\}$ .

Prefer hard clipping in SAM output (-H) - use hard clipping in the SAM output. This option may dramatically reduce the redundancy of output when mapping long contig or BAC sequences.

Select the required parameters and press the Start button.

#### **Building Index for BWA-SW**

To build BWA-SW index select the Tools NGS data analysis Build index for reads mapping item in the main menu. The Build Index dialog will appears. Set the Align short reads method parameter to BWA-SW.

The dialog looks as follows:

U Build Index			? ×
Align short reads met	hod	BWA-SW	•
Reference sequence			
Index file name			
Index algorithm (-a)		is	•
Colorspace (-c)			
	5	Start Cancel	Help

There are the following parameters:

Reference sequence — DNA sequence to which short reads would be aligned to. This parameter is required.

Index file name — file to save index to. This parameter is required.

Index algorithm (-a) — Algorithm for constructing BWA index. Available options are:

It implements three different algorithms

- is designed for short reads up to ~200bp with low error rate (<3%). It does gapped global alignment w.r.t. reads, supports paired-end reads, and is one of the fastest short read alignment algorithms to date while also visiting suboptimal hits.</li>
- bwtsw is designed for long reads with more errors. It performs heuristic Smith-Waterman-like alignment to find high-scoring local hits. Algorithm implemented in BWA-SW. On low-error short queries, BWA-SW. is slower and less accurate than the *is* algorithm, but on long reads, it is better.
- *div* does not work for long genomes.

Colorspace (-color) — the input is read in colorspace, colors are encoded as characters A/C/G/T (A=blue, C=green, G=orange, T=red).

## **BWA-MEM**

*BWA* is a fast light-weighted tool that aligns relatively short reads to a reference sequence. Click this link to open *BWA* homepage. BWA-ME M is generally recommended for high-quality queries as it is faster and more accurate. BWA-MEM also has better performance than BWA-backtrack for 70-100bp Illumina reads.

Open Tools Align to reference submenu of the main menu.

Тоо	Tools Window Help							
	Sanger data analysis	•	1					
1	NGS data analysis	•		Reads quality control				
S	BLAST			Genome de novo assembly				
	Multiple sequence alignment							
-	Cloning			Map reads to reference				
	Primer	-		Build index for reads mapping				

Select the *Align short reads* item to align short reads to a DNA sequence using *BWA-MEM*. Or select the *Build index* item to build an index for a DNA sequence which can be used to optimize aligning of short reads.

- Aligning Short Reads with BWA-MEM
- Building Index for BWA-MEM

#### Aligning Short Reads with BWA-MEM

When you select the Tools Align to reference Align short reads item in the main menu, the Align Sequencing Reads dialog appears. Set value of the Align short reads method parameter to BWA-MEM. The dialog looks as follows:

Align Sequencing Reads				8
Alignment method		BWA-MEM		•
Reference sequence				
Result file name				
Library Single-end 💌				🗹 SAM output
Short reads				
Path			Туре	Order
			Add	Remove
Algorithm Scoring Options	Other			1
Min seed length (-k)	19 ÷	Band width	(-w)	100 🗄
Dropoff (-d)	100 ÷	Internall see	eds length (-r)	
Skip seeds threshold (-c)	10000 ÷	Drop chain t	hreshold (-D)	0.50 ÷
Rounds of mate rescues (-m)	100 ÷	Number of t	hreads (-t)	8
🗆 Skip pairing (-P)				
NOTE: bwa mem accepts read	ls only in FAST	A or FASTO for	mat. Reads sl	hould be
compiled into a single file for			inde needs si	
Help			Cance	el Start

There are the following parameters:

Reference sequence — DNA sequence to align short reads to. This parameter is required.

Result file name — file in SAM format to write the result of the alignment into. This parameter is required.

Prebuilt index — check this box to use an index file instead of a source reference sequence. Also you can build it manually.

SAM output — always save the output file in the SAM format (the option is disabled for BWA).

Short reads - each added short read is a small DNA sequence file. At least one read should be added.

You can also configure other parameters.

Index algorithm (-a) — algorithm for constructing BWA index.

It implements three different algorithms:

- *is* designed for short reads up to ~200bp with low error rate (<3%). It does gapped global alignment w.r.t. reads, supports paired-end reads, and is one of the fastest short read alignment algorithms to date while also visiting suboptimal hits.
- *bwtsw* is designed for long reads with more errors. It performs heuristic Smith-Waterman-like alignment to find high-scoring local hits. Algorithm implemented in BWA-SW. On low-error short queries, *BWA-SW*. is slower and less accurate than the *is* algorithm, but on long reads, it is better.
- *div* does not work for long genomes.

*Number of threads (-t)* — number of threads.

*Min seed length* (-*k*) — minimum seed length. Matches shorter than *INT* will be missed. The alignment speed is usually insensitive to this value unless it significantly deviates 20.

Band width (-w) — band width. Essentially, gaps longer than *INT* will not be found. Note that the maximum gap length is also affected by the scoring matrix and the hit length, not solely determined by this option.

Dropoff (-d) — off-diagonal X-dropoff (Z-dropoff). Stop extension when the difference between the best and the current extension score is above  $|i-j|^*A+INT$ , where *i* and *j* are the current positions of the query and reference, respectively, and *A* is the matching score. Z-dropoff is similar to BLAST's X-dropoff except that it doesn't penalize gaps in one of the sequences in the alignment. Z-dropoff not only avoids unnecessary extension, but also reduces poor alignments inside a long good alignment.

Internall seeds length (-r) - trigger re-seeding for a MEM longer than minSeedLen\*FLOAT. This is a key heuristic parameter for tuning the performance. Larger value yields fewer seeds, which leads to faster alignment speed but lower accuracy.

Skip seeds threshold (-c) - discard a MEM if it has more than INT occurence in the genome. This is an insensitive parameter.

Drop chain threshold (-D) - drop chains shorter than FLOAT fraction of the longest overlapping chain.

Rounds of mate rescues (-m) - perform at most INT rounds of mate rescues for each read.

Skip mate rescue (-S) - skip mate rescue.

Skip pairing (-P) - in the paired-end mode, perform SW to rescue missing hits only but do not try to find hits that fit a proper pair.

Score for a match (-A) - matching score.

Mismatch penalty (-B) - mismatch penalty. The sequence error rate is approximately: {.75 \* exp[-log(4) \* B/A]}.

Gap open penalty (-O) - gap open penalty.

Gap extention penalty (-E) - gap extension penalty. A gap of length k costs O + k\*E (i.e. Gap open penalty is for opening a zero-length gap).

Penalty for clipping (-L) - clipping penalty. When performing SW extension, BWA-MEM keeps track of the best score reaching the end of query. If this score is larger than the best SW score minus the clipping penalty, clipping will not be applied. Note that in this case, the SAM AS tag reports the best SW score; clipping penalty is not deducted.

Penalty unpaired (-U) - penalty for an unpaired read pair. BWA-MEM scores an unpaired read pair as scoreRead1+scoreRead2-INT a nd scores a paired as scoreRead1+scoreRead2-insertPenalty. It compares these two scores to determine whether we should force pairing.

Score threshold (-T) - don't output alignment with score lower than score threshold. This option only affects output.

Select the required parameters and press the Start button.

#### Building Index for BWA-MEM

To build BWA-SW index select the Tools NGS data analysis Build index for reads mapping item in the main menu. The Build Index dialog will appears. Set the Align short reads method parameter to BWA-MEM.

The dialog looks as follows:

U Build Index				
Align short reads method		BWA-MEM		
Reference sequence				
Index file name				
Index algorithm (-a)		is		•
Colorspace (-c)				
	s	tart	Cancel	Help

There are the following parameters:

Reference sequence — DNA sequence to which short reads would be aligned to. This parameter is required.

Index file name — file to save index to. This parameter is required.

Index algorithm (-a) — Algorithm for constructing BWA index. Available options are:

It implements three different algorithms

- *is* designed for short reads up to ~200bp with low error rate (<3%). It does gapped global alignment w.r.t. reads, supports paired-end reads, and is one of the fastest short read alignment algorithms to date while also visiting suboptimal hits.
- *bwtsw* is designed for long reads with more errors. It performs heuristic Smith-Waterman-like alignment to find high-scoring local hits. Algorithm implemented in BWA-SW. On low-error short queries, *BWA-SW*. is slower and less accurate than the *is* algorithm, but on long reads, it is better.
- *div* does not work for long genomes.

Colorspace (-c) — the input is read in colorspace, colors are encoded as characters A/C/G/T (A=blue, C=green, G=orange, T=red).

### **UGENE Genome Aligner**

The UGENE Genome Aligner is a fast short read aligner. It aligns DNA sequences of various lengths to the reference genome with configurable mismatch rate.

It is available from the Tools DNA assembly submenu of the main menu.

Тоо	Tools Window Help					
1	Sanger data analysis	١.	1			
1	NGS data analysis	•		Reads quality control		
S	5EAST	•	1	Genome de novo assembly		
	Multiple sequence alignment Cloning		1	Map reads to reference		
	Primer	•		Build index for reads mapping		

Select the *Align short reads* item to align short reads to a DNA sequence or *Build index* item to build an index for a DNA sequence which can be used to optimize aligning short reads to the sequence.

- Aligning Short Reads with UGENE Genome Aligner
- Building Index for UGENE Genome Aligner
- Converting UGENE Assembly Database to SAM Format

## Aligning Short Reads with UGENE Genome Aligner

When you select the Tools DNA Assembly Align short reads item in the main menu, the Align Short Reads dialog appears. Set the Align short reads method parameter to UGENE Genome Aligner. The dialog looks as follows:

Align Sequencing Reads		? ×
Alignment method	UGENE Genome Aligner	•
Reference sequence		
Result file name		
Library Single-end  Prebuilt index Short reads		SAM output
Path	Туре	Order
	.76-	
		Add Remove
Common parameters Advanced parameters Index p	arameters	
Mismatches allowed		
Mismatches number		1 *
Percentage of mismatches		1 *
Align options		
Use GPU-optimization		
Align reverse complement reads		
Use "best"-mode during the alignment		
Omit reads with qualities lower than		20 🚖
	Start	Cancel Help

The following parameters are available:

Reference sequence - DNA sequence to align short reads to. This parameter is required.

Result file name — file in UGENE database format or SAM format (if the box SAM output check), to write the result of the alignment into. This parameter is required.

Prebuilt index — check this box to use an index file instead of a reference sequence. Also you can build it manually.

SAM output — checking this box allows one to save output files in the SAM format. The default format of output files is the UGENE database format (ugenedb).

Short reads — each added short read is a small DNA sequence file. At least one read should be added.

The Aligning Short Reads with UGENE Genome Aligner has no limitation on short reads length.

Common parameters:

Mismatches allowed — check this box to allow mismatches between the reference sequence and a short read. Select one of the following:

## **Unipro UGENE Manual, Version 1.30**

- Mismatches number to set the number of mismatched nucleotides allowed. This parameter can take values: 1, 2 and 3.
- Percentage of mismatches to set the number of mismatches in percents. Note, that in this case the absolute number of
  mismatches can vary for different reads. This parameter can take values: 1 10 %.

### Align options:

- Use GPU-optimization use an openCL-enabled GPU during the alignment (the corresponding hardware should be available on your computer).
- Align reverse complement reads use both: a read and its reverse complement during the alignment.
- Use "best"-mode during the alignment report only about best alignments (in terms of mismatches).
- Omit reads with qualities lower than omit all reads with qualities lower than the specified value. Reads that have no qualities are not omited.

### Advanced parameters:

Maximum memory for short reads — maximum memory usage for short reads. This parameter allows one to decrease the load on the computer on one side and to increase the computer speed of the task on the other side.

- Total memory usage shows the total memory usage.
- System memory size shows the total system memory size.

### Index parameters:

*Reference fragmentation* — this parameter influences the number of parts the reference will be divided. It is better to make it bigger, but it influences the amount of memory used during the alignment.

- Index memory usage size shows the index memory usage.
- Directory for index files temporary directory for saving index files.

You can choose a temporary directory for saving index files for the reference that will be built during the alignment. If you need to run this algorithm one more time with the same reference and with the same reference fragmentation parameter, you can use this prebuilt index that will be located in the temporary directory.

## **Building Index for UGENE Genome Aligner**

You can build an index to optimize short reads alignment using UGENE Genome Aligner. To open the Build Index dialog, select the Tools NGS data analysis Build index for reads mapping item in the main menu. Set value of the Align short reads method parameter to UGENE Genome Aligner.

U Build Index		? <mark>- x</mark>
Align short reads method	UGENE Genome Aligner	•
Reference sequence		
Reference fragmentation		0 Mb
Total memory usage: System memory size:		0 Mb 1536 Mb
	Start Cancel	Help

The dialog looks as follows:

The parameters are the following:

Reference sequence — DNA sequence to which short reads would be aligned to. This parameter is required.

Index file name - file to save index to. This parameter is required.

Reference fragmentation — this parameter influences the amount of parts the reference will be devided. It is better to make it bigger,

but it influences the amount of memory used during the alignment.

Total memory usage - shows the total memory usage.

System memory size — shows the total system memory size.

## **Converting UGENE Assembly Database to SAM Format**

To convert UGENE data base to SAM format click on the *Tools->DNA Assembly->Convert UGENE assembly database to SAM format* conte xt main menu item. The following dialog will appear:

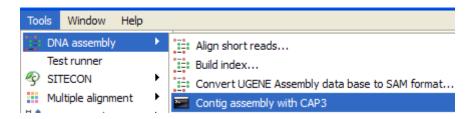
U Convert UGENE Assembly Database to SAM Format						
Assembly database:						
Result SAM file:						
	Convert Cancel Help					

Select assembly and result files and click on the Convert button.

# CAP3

CAP3 (CONTIG ASSEMBLY PROGRAM Version 3) is a sequence assembly program for small-scale assembly with or without quality values. Click this link to open CAP3 homepage. CAP3 is embedded as an *external tool* into UGENE.

Open Tools DNA assembly submenu of the main menu.



Select the Contig assembly with CAP3 item to use the CAP3.

The Contig Assembly With CAP3 dialog appears.

U Contig Assembly with CAP3	? ×
Base Advanced	
Input files (long DNA reads to assembly)	
	Add Remove Remove All
Result contig	
Run Cancel	Help

You can add or remove input files using *Add* and *Remove* buttons. To remove all files click the *Remove all* button. *Input files* are files with a long DNA reads in FASTA, FASTQ, SCF or ABI formats. At least one input file should be added. Input a *Result contig* name and press the *R un* button. *CAP3* produces assembly results in the ACE file format (".ace"). The file contains one or several contigs assembled from the input reads.

The quality scores for FASTA sequences can be provided in an additional file. The file must be located in the same folder as the original sequences and have the same name as FASTA file, but another extension: *.qual*.

Also you can change the following advanced parameters:

Base Advanced			
Clipping for poor regions	Length and percent identity of an over	lap	
Base quality cutoff for clipping (-c)	Overlap length cutoff (-o)	40	
Clipping range (-y) 100	Overlap percent identity cutoff (-p)	90	<b>▲</b>
Quality difference score of an overlap	Other parameters		
	Max number of word matches (-t)	300	. <u></u> T
Base quality cutoff for differences (-b) 20	Band expansion size (-a)	20	* *
Max qscore sum at differences (-d) 200	Max gap length in any overlap (-f)	20	* *
	Assembly reverse reads (-r)		
Similarity score of an overlap			
Match score factor (-m) 2			
Mismatch score factor (-n) -5			
Gap penalty factor (-g) 6			
Overlap similarity score cutoff (-s) 900			

Clipping for poor regions parameters:

Clipping of a poor end region of a read is controlled by parameters *Base quality cutoff for clipping (-c)* (the specified value should be more than 5), and *Clipping range (-y)* (the specified value should be more than 5).

Quality difference score of an overlap parameters:

Base quality cutoff for differences (-b) — if an overlap contains a difference at bases of quality values q1 and q2, then the score at the difference is max(0, min(q1, q2) - b), where b is the specified value. The specified value should be more than 15. The difference score of an overlap is the sum of scores at each difference.

Max qscore sum at differences (-d) — remove an overlap if its difference score is greater than the specified value. The specified value should be more than 20.

Similarity score of an overlap parameters:

The following parameters are used to calculate the similarity score of an overlapping alignment:

*Match score factor* (-m) — a match at bases of quality values q1 and q2 is given a score of m \* min(q1, q2), where m is the specified value. The specified value should be more than 0.

*Mismatch score factor* (-n) — a mismatch at bases of quality values q1 and q2 is given a score of n \* min(q1, q2), where n is the specified value. The specified value should be less than 0.

*Gap penalty factor* (*-g*) — a base of quality value q1 in a gap is given a score -g \* min(q1, q2), where g is the specified value; q2 is the quality value of the base in the other sequence right before the gap. The specified value should be more than 0.

The similarity score is callulated as the sum of scores of each match, each mismatch and each gap. Based on this value and the following value some overlaps are removed:

Overlap similarity score cutoff (-s) — remove overlaps with similarity scores less than the specified value. The specified value should be more than 250.

Length and percent identity of an overlap parameters:

Overlap length cutoff (-o) — minimum length of an overlap (in base pairs). The specified value should be more than 15 base pairs.

Overlap percent identity cutoff (-p) — minimum percent identity of an overlap. The specified value should be more than 65%.

Other parameters:

*Maximum number of word matches (-t)* — an upper limit of word matches between a read and other reads. Increasing the value would result in more accuracy, however this could slow down the program. The specified value should be more than 0.

Band expansion size (-a) — a number of bases to expand a band of diagonals for an overlapping alignment between two sequence reads. The specified value should be more than 10.

*Max gap length in any overlap (-f)* — reject overlaps with a gap longer than the specified value. A small value may cause the program to remove true overlaps and to produce incorrect results. This option may be used by the user to split reads from alternative splicing forms into separate contigs. The specified value should be more than 1.

Assembly reverse reads (-r) — consider reads in reverse orientation for assembly. The default value is "checked".

## SPAdes

SPAdes – St. Petersburg genome assembler. Click this link to open SPAdes homepage. SPAdes is embedded as an *external tool* into UGENE.

Open Tools NGS data analysis.

	Тоо	s Window Help		
		Sanger data analysis	⊁	Gol 🖉 - 1:1 👻 👫 SW 🍡 👼 🛃
1	1	NGS data analysis	•	Reads quality control
1	S	BLAST	•	 Genome de novo assembly
ł		Multiple sequence alignment	•	
	ø	Cloning	►	 Map reads to reference
		Primer	•	 Build index for reads mapping

Select the Genome de novo assembly item to use the SPAdes.

#### The Assemble Genomes dialog will appear.

U Assemble Genom	es			?[	x
Assembly method			SPAdes		•
Output directory					)
Library Single-end	•				
Properties		Left reads		Right reads	
# Type	Orientation	Path		Path	
		Add	move	Add Remove	
Base Options					
Dataset type Running mode	Multi Cell Error Correction and Assembly	ý		Number of threads (-t) 8	
k-mer sizes (-k)	auto			Memory limit GB (-m) 250 🚔	
				Start Cancel Help	

#### The following parameters are available:

Output directory - SPAdes stores all output files in output directory, which is set by the user.

Library - to run SPAdes choose one of the following libraries:

- Single-end
- Paired-end
- Paired-end (Interplaced)
- Paired-end (Unpaired files)
- Sanger
- PacBio

Left reads - file(s) with left reads.

Right reads - file(s) with right reads.

For each dataset in the paired-end libraries you can change type and orientation.

Datasest type - dataset type.

Running mode - running mode.

k-mer sizes (-k) - k-mer sizes.

Number of threads (-t) - number of threads.

Memory limit GB (-m) - memory limit.

## **Weight Matrix**

The Weight Matrix plugin is a tool for solving the problem of a sequence annotating. As well as for the SITECON, the main use case of the plugin is recognition of potential transcription factor binding sites on basis of the data about conservative conformational and physicochemical properties revealed with the binding sites sets analysis.

The Weight Matrix contains a lot of position frequency matrices (PFM 's) and position weight matrices (PWM 's, also known as position specific score matrices — PSSM 's). The matrices came from two wide-known open archives: JASPAR, which contains frequency matrices, and UniPROBE containing weight matrices.

Also the *Weight Matrix* plugin provides a tool for creating specific position frequency and weight matrices from an existing alignment or from a file with several sequences. The created matrix can be used as a profile for the search as well as the JASPAR and UNIPROBE ones.

To search for transcription factor binding sites in a DNA sequence select the *Analyze Search TFBS with matrices* context menu item. The *W* eight matrix search dialog will appear:

Weight Matrix Se	arch			? <mark>x</mark>
Matrix:			0	
Score:	1 I I I I			85%
Search JASPAR data	abase Build new ma	atrix		View matrix
Weight algorithm Be	erg and von Hippel			
Strands		Region		
Both strands		Whole sequen	ce	•
Direct strand			1 -	199950
Reverse completion	ement strand			
Matrix	Minimal score	Algorithm		Load list
				Save list
				Load folder
				Clear list
Range	Matrix	Strand	Score	
Clear results Sav	/e as annotations			
Results found: 0.				
	Add to queue	Search	Close	Help

In the search dialog you must specify a file with PWM or PFM. You can do so by pressing the browse button and selecting the file.

Also you can use the special interface to choose a JASPAR matrix by pressing the Search JASPAR database button.

Alternative way to specify the position weight/frequency matrix is to create a specific one from an alignment or a file with several sequences with the *build a new matrix* tool.

After the profile (the matrix) is loaded, you can adjust the threshold value. The threshold sets the minimal identity score for a result to pass. The more the result score is, the more it is homologically related to the aligned region. By changing the threshold you can filter low- scoring results.

If the loaded matrix is a position frequency matrix, you must also specify the algorithm to build the corresponding position weight matrix which will represent the transcription factor. There are four algorithms available.

Weight algorithm	Berg and von Hippel	•
-Strands	Berg and von Hippel	
Suanus	Log-odds	
O Both strand	Match	
O bourbuland	NLG	
O Direct shares		

Also you can add a selected matrix with the specified *Minimal score* and the *Algorithm* to the matrices list. To do it, select the matrix and other options and press the *Add to queue* button. The plugin will search with all matrices specified in the list.

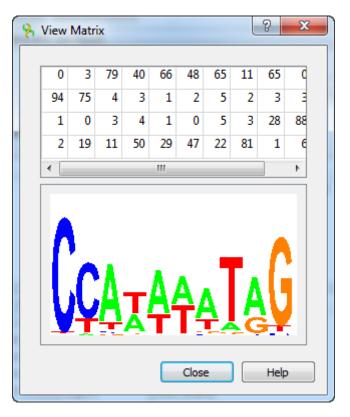
You can use the Save list button to export the list of matrices to a \*.csv file. Later the list can be loaded from the file using the Load list button

The rest options are standard sequence search options: the strand and the sequence region where to search for matches.

After specifying the necessary options press the Search button. The found results will appear in the dialog table. The corresponding results identity scores are in the Score column.

Range 💎	Matrix	Strand	Score	
199944199949	MA0271.1.pfm	Direct strand		31.26%
199943199948	MA0271.1.pfm	Direct strand		62.39%
199942199947	MA0271.1.pfm	Direct strand		53.92%
199941199946	MA0271.1.pfm	Direct strand		26.86%
199940199945	MA0271.1.pfm	Direct strand		26.86%
199939199944	MA0271.1.pfm	Direct strand		14.07%
199938199943	MA0271.1.pfm	Direct strand		57.69%

Also you can see the matrix by using the View matrix button:



The regions found by the weight matrix algorithm can be saved as annotations to the DNA sequence in the Genbank format by pressing the *Save as annotations* button.

After saving, the file with resulting annotations will be automatically added to the current project, and the annotations will be added to the original sequence.

Note that in case of selecting JASPAR or UNIPROBE matrix, the resulting annotations will contain the given matrix properties.

Weight Matrix S Matrix: C:/work/ug	earch jene/data/position_weight_matrix/JASPAR/plants/MA0044.1.pfm	···
Score:	% Create Annotation	View matrix
Weight algorithm Strands	<ul> <li>✓ Save annotation(s) to</li> <li>○ Existing table</li> <li>✓ Image: Create new table</li> </ul>	
<ul> <li>Direct stranc</li> <li>Reverse com</li> </ul>	✓ Annotation parameters Group name <auto></auto>	199950
Matrix MA0044.1.pfm	Annotation name misc_feature	Load list Save list Load folder
		Clear list
199397199405 198975198983	Matrix     Strand     Score       MA0044.1.pfm     Direct strand       MA0044.1.pfm     Direct strand       MA0044.1.pfm     Direct strand       We as annotations     Ve as annotations	85.35% 85.75% 85.29%
Results found: 1379	Add to queue Search Close	Help

See also:

- Searching JASPAR DatabaseBuilding New Matrix

## Searching JASPAR Database

Press the Search JASPAR database button in the Weight matrix search dialog. The following dialog will appear:

Search JASPAR	Database			? ×
Name	ID	Class	Family	*
<ul> <li>vertebrates</li> <li>urochordate</li> <li>plants</li> </ul>	s			н
squamo	sa MA0082.1	Other Alpha-He	MADS	
myb.Ph		Helix-Turn-Helix	Myb	
id1	MA0120.1	Zinc-coordinati	BetaBetaAlpha-zin	c finger
bZIP911	MA0097.1	Zipper-Type	Leucine Zipper	
bZIP910	MA0096.1	Zipper-Type	Leucine Zipper	
abi4	MA0123.1	Beta-Hairpin-Ri	AP2 MBD-like	
TGA1A	MA0129.1	Zipper-Type	Leucine Zipper	
PEND PBF	MA0127.1 MA0064.1	Zipper-Type Zinc-coordinati	Leucine Zipper Dof	
MNR1A	MA0004.1 MA0053.1	Zinc-coordinati	Dof	-
acc		AAC18941		A III
class		Zinc-coordinating	9	
collection		CORE		
comment		-		-
		Select	Cancel	Help

Here the matrices are divided into categories and you can read detailed information of a matrix which is represented by its properties. It could help you to choose the matrix properly.

The matrices provided with UGENE are located in the \$UGENE/data/position\_weight\_matrix folder.

## **Building New Matrix**

To create a position weight or frequency matrix from an alignment or a file with several sequences, press the *Build new matrix* button in the *Weight matrix search* dialog, or select the *Tools Search for TFBS Build weight matrix* main menu item:

Тоо	ls Window Help		-
1	Sanger data analysis	•	Go! 𝒫1:1 ▾ 1:1 ▾ 🗝
1	NGS data analysis	•	
S	BLAST	•	
	Multiple sequence alignment	+	1k 1.5k 2k
- 🔊	Cloning	+	
£	Primer	•	
	Search for TFBS	•	Build SITECON model
IM.	HMMER tools	•	Build weight matrix
	Build dotplot		1k 1.5k 2k
\$	Random sequence generator		

The Build weight or frequency matrix dialog will appear:

U Build Weight or	Frequency Matrix	? <mark>x</mark>
Input file		
Output file		
Statistic options		
Statistic type:	Mononucleic O Dinucleic	
Matrix options		
Matrix type:	<ul> <li>Frequency matrix</li> <li>Weight matrix</li> </ul>	
Weight algorithm	Berg and von Hippel	-
	Start Cancel	Help

The following parameters are available:

Input file — an alignment or a file with several sequences to build the matrix from. The parameter is mandatory.

Output file - the resulting matrix will be saved in this file. The parameter is mandatory.

Statistic type — defines the way in which the statistics will be collected. The *Mononucleic* option is basically good for small alignments, and the *Dinucleic* option must give more appropriate results for big alignments.

Matrix type --- defines the type of the resulting matrix.

If the Frequency matrix option is selected then the frequency matrix will be created and saved into the resulting file.

If the Weight matrix option is selected then the intermediate frequency matrix will be created and then transformed into a weight matrix on basis of the selected Weight algorithm. Then the weight matrix will be saved into the resulting file.

For some input files the colored "Alignment Logo" appears at the bottom of the dialog. It gives the representation of the selected alignment.

U Build Weight or Frequency Matrix
Input file
Output file
Statistic options
Statistic type:      O Mononucleic     O Dinucleic
Matrix options
Matrix type:      O Frequency matrix      Weight matrix
Weight algorithm Berg and von Hippel 💌
Start Cancel Help

1 The "Alignment logo" appears when:

- The input file format is \*.pfm, \*.aln or it is a file with several sequences;
- The size of the input file is small enough.

To start the operation, press the *Start* button. The matrix will be created and saved. If the *Build weight or frequency matrix* dialog was invoked from the *Weight matrix search* dialog, then the matrix also will be chosen as the current profile.

## Primer3

The Primer3 plugin is a port of the Primer3 tool. It is intended to pick primers from a DNA sequence.

To use the Primer3, open a DNA sequence and select the Analyze Primer3 context menu item. The dialog will appear:

Primer Designer						<u></u> ନ୍
Main General Setting	s Internal Oligo	Penalty Weights RT-PCF	Sequence Qual	lity Result Set	tings	
Targets						
Product size ranges 150	-250 100-300 301-400	0 401-500 501-600 601-700 70	1-850 851-1000		Mispriming/Repeat library	NONE
Number to return	5	Max 3' stability		9.00		
Max repeat mispriming	12.00	Pair max repea	t mispriming	24.00	×	
Max template mispriming	12.00	Pair max templ	ate mispriming	24.00	•	
Start codon position						
Pick left primer		Pick hybridization	n probe (internal oligi	o)	Pick right primer	
or use left primer below		or use oligo below			or use right primer below (5	to 3' on opposite strand)
	_					
gion Whole sequence	•	1 -	199950			

All available parameters are the same as in the original Primer3.

However there is one additional feature available which is not originally a part of Primer3 tool. It allows user design primers for RT-PCR experiments by choosing which exons/introns to span with the primer product. This feature is described in detailed below. When you select the parameters you can save and load settings with a help of the corresponding buttons in the right corner of the dialog.

RTPCR Primer Design

## **RTPCR Primer Design**

This feature allows to search for primer pairs that span introns on the genomic sequence or exon junctions on the mRNA sequence.

Note that RT-PCR design is only available for mRNA/cDNA sequences with annotated exons. There are several ways to obtain the cDNA for a corresponding DNA sequence.

- From NCBI or ENSEMBL database.
   For example, one can download the *TMPRSS2 transcript variant 1* from NCBI Genbank using identifier NM\_001135099.1.
   This can be also done from UGENE using option Access remote database or Search NCBI Genbank.
- Align the genomic and cDNA sequences using spliced aligner.
   For this option one must have both genomic and cDNA sequences.
   In UGENE the spliced alignment can be performed using the Spidey tool.
   To run the alignment open the genomic sequence and select action *Align Align to mRNA sequence*.
   The generated exon annotations can be then exported using action *Export Export sequence of selected annotations*

To design primers for your mRNA sequence and go to the RT-PCR tab of the Primer Designer dilaog:

Primer Designer	nii i						8 X
Main General Settings	Internal Oligo	Penalty Weights	RT-PCR	Sequence Quality	Result Settings		
Design primers for RT-P	CR analysis						
mRNA sequence							
Exon annotation name:	exon						
Minimum exon junctio	on overlap size						
At 5' side (bp): 5 🚔							
At 3'side (bp) 5 🌩							
Exon range:							
Primer product must sp	oan at least one intr	on on the correspond	ling genomic	DNA			
Max number of pairs to qu	iery: 1000 ≑						
Max number of pairs to qu	ery: 1000 🖨						
Max number of pairs to qu	ery: 1000 束						
Max number of pairs to qu	ery: 1000 👻						
Max number of pairs to qu	ery: 1000 👻						
Max number of pairs to qu	ery: 1000 🛓						
Max number of pairs to qu	ery: 1000 🛓						
Max number of pairs to qu	ery: 1000 🚖						
Max number of pairs to qu	1000 x						
Max number of pairs to qu	ery: 1000 +						
Max number of pairs to qu	ery: 1000 *						
Max number of pairs to qu Max number of pairs to qu Whole sequence	ery: 1000 *	1 -		199950			

The following parameters are available:

#### Exon annotation name

To detect exon boundaries UGENE searches for exonic annotations. This option allows to set custom name for annotations denoting exons. Default value is "exon"

Minimum exon junction overlap size

If checked, then only the pairs with at least one of the primers overlapping exon junction in the mRNA sequence will be selected.

At 5' side (bp)

Minimum overlap size on the 5' side of the exon junction. Default is 5 bp.

At 3' side (bp)

Minimum overlap size on the 3' side of the exon junction. Default is 5 bp.

#### Exon range

This option allows to limit the sequence region, where the primers are searched for. For example, setting value "3-5" will limit the search to a sequence region consisting of exons 3,4,5 of the transcript, as defined by the order in the sequence. Default value is an empty string, which means that there are no limitations.

#### Span at least one intron

This option makes sure that primer product should span an intron on the genomic sequence i.e the forward and reverse primers must be located in different exons. The option is enabled by default.

#### Max numbers of pairs to query

The algorithm applied in RT-PCR primer design first searches for all available primers in a given sequence. Then it filters the detected pairs to make sure that they satisfy the selected configuration. This option allows to set the maximum number of pairs for the initial search query. Larger number will result in increased sensitivity, but also in a longer running time. Default value is 1000.

**Important**: using the **RT-PCR** primer design tab will reset the values set in the *Exlcuded regions* and *Targets* of the **Main** configuration tab. Additionally if the *Exon range* option is set, the defined sequence region will be ignored.

# Spliced Alignment mRNA and cDNA

UGENE allows to align spliced mRNA/cDNA sequence to genomic sequences.

The default underlying algorithm which is used for the alignment is an external tool called Spidey.

Before running the alignment make sure that Spidey is available and validated in the list of External Tools.

To perform the alignment of a mRNA sequence to a genomic sequence open the the genomic sequence in the Sequence View. Next activate context menu item Align -> Align to sequence to mRNA.

F	Т	S N L K Q	T P	TH I	W C G	I A	s i	* D	L 1 S	L R * 0	R 5 E	s v	L C
V	۵	Go to position	Ctrl+G	м	V	н	L	т	Р	Е	Е	K S	B A
ттc		Select sequence region	Ctrl+A	AT(	GGTG	CAI	ст	GAC	TCCI	GAG	GAGA	AGT	CTGC
5 AAC	<b>4</b>	New annotation	Ctrl+N	TA	55 CCAC	GTA	60 AGA(		65 AGGI	70 ACTCO	75 CTCT	TCAG	80 GACC
E		Сору	+	1	P A	. I	) ;	S	Eζ	Q P	S	т	Q
N 🔹		Select	+	H M	H T	<b>м</b> С	Q R	s V	R G	L S S	L B F	L F D	R G A
		Add	+										
		Analyze	•										
		Align	•		Align	seque	nce to	o mRN	JA				
		Cloning	•		Align	selecte	ed seq	quenc	e regior	1S			
		Export	•		Align	selecte	ed seq	quenc	e regior	ns (amin	o acids)		
		Edit sequence	•		Align	selecte	ed anı	notati	ons				
		Remove	Þ		Align	selecte	ed anı	notati	ons (am	nino acid	ls)		
		Rulers	+	Г									_
		Statistics											

In the list of sequences select the corresponding mRNA sequence and click OK.

The following dialog will appear:

U Save result	to annotation	? ×
Group name	exon	*
Description		
Existing tag	ble	
<b>ORTIGO 13</b>	63 features [murine.gb]	- 🏠
O Create ne	w table	
O Use auto-	annotations table	
	Create Cancel	Help

Here you can set up a file to store annotations. It could be either an existing annotation table object or a new annotation table or auto-annotations table (if it is possible). Also you can modify the group name parameter and add a description.

The resulting alignment will be saved as an annotation with the corresponding name:

-* UGEN	NE - [c Action			; T	Tools	W	/ind	ow	Hel	p																					X 7 X
•			<b>7</b> 6	ß		1	Ì		Ę	<u>a</u>	ĨA					Go!		P <sub>1:1</sub>	Ŧ	1:1	•	ACG	s	<u>v</u>	•2	F					>
1 1 1 1 1 1 1 1 1 1 1 1 1 1	C FGTT 5 13 ACAZ N T	200 200 142 H S FCAT 20 AGTZ M	273 € 273 € 1329 1329 TAC	P E CT( 5 GA(	40 [223] S L ] <del>CTT</del>	00 bp]- Y I FA1	I S FCS AGZ R	00 495 F F 1335	GGA R	P 9 134 G G	700 700 P F F F F G G	T ( H ACZ C 7	800 800 III 2 S 3 3 5 1345		900 900 P L S CT 135 GA R	G W CC P P	G A GC	Q N 355 TI L	T R	1 7 7 136	CT 0	1. 1 G W GG	V S TC 365		4k [2 C <b>V</b> 137 AC H	1.5 263 b ∇ C TG	A CT GA	608 608 608 608 608 707 707 707 707 707 707 707 707 707 7	A A	1	Acg
Nam	ne			-	~							Valu	e																,	-	
	4 🧭	notatio exon exo exo to-ann	(0, 1) n	-								join	[114	12,27	7349	5,13	461	1608	)												
•				1		_																								F	

# **External Tools**

The External Tools plugin allows one to launch an external tool from UGENE.

To use an external tool from UGENE, the tool needs to be installed on the system and the path to it should be properly configured. However, there is no need in the additional configuration, if you've installed the UGENE Full Package, as it already contains all the tools by default.

Otherwise, if you've installed the UGENE Standard Package, you would need to configure an external tool in order to use it. Note that in this case you can download the package with all the external tools from this page.

To learn how to configure an external tool, read below.

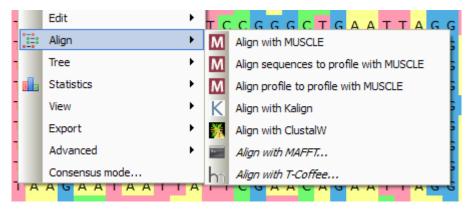
Configuring External Tool

## **Configuring External Tool**

To configure an external tool:

- 1. Make sure the tool is installed on your system.
- 2. Set a path to the tool executable file in UGENE. It can be set on the External Tools tab of the Application Settings dialog.

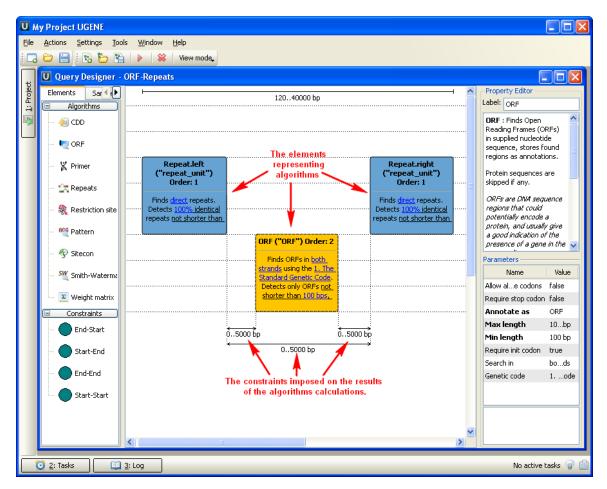
If the path hasn't been set for a tool, UGENE menu items that launch the tool are displayed in italic. For example, on the image below a path for the ClustalW external tool has been set, and paths for MAFFT and T-Coffee has not:



## **Query Designer**

The *Query Designer* allows a molecular biologist to analyze a nucleotide sequence using different algorithms (Repeats finder, ORF finder, Weight matrix matching, etc.) at the same time imposing constraints on the positional relationship of the results obtained from the algorithms.

A user-friendly interface is used to create a schema of the algorithms and constraints.



Alternatively, you can create / edit a schema using a text editor.

When the schema has been created and all its parameters have been set you can run it for a nucleotide sequence. The results are saved as a set of annotations to the specified file in the Genbank format. Also when you have query designer scheme you can analyze a nucleotide sequence from the sequence view with a help of this schema. Call the *Analyze->Analyze with query schema* context menu item for this.

To learn more about the Query Designer read the Query Designer Manual.

# **Plasmid Auto Annotation**

Plasmid Auto Annotation feature allows to automatically annotate possible functional elements of the given sequence such as promoters, terminators, origin of replication, known genes, common primers and other features. Conceptually this functionality is similar to the one offered by PlasMapper software. The database for plasmid auto-annotation is based on the following resource.

To activate Plasmid Auto Annotation upon your sequence use the menu item Analyze Annotate plasmid and custom features. In the appeare

d dialog one can selected the features to search in sequence.

8 Plasmid Auto Annotations	8 ×
Selected features	
V Promoter	V Origin
Terminator	V Primer
Gene	✓ Other features
Regulatory	
Annotate	Cancel Help
[L	

The detected plasmid features are stored as automatic annotations and can be controlled through corresponding menu. Refer *Automatic Annotations Highlighting* to learn more.

The database containing features and their sequences is located in a subfolder of UGENE data folder: data/custom\_annotations/

### plasmid\_features.txt.

## **ClustalO**

Clustal is a widely used multiple sequence alignment program. It is used for both nucleotide and protein sequences. Clustal Omega is the latest addition to the Clustal family. It offers a significant increase in scalability over previous versions, allowing hundreds of thousands of sequences to be aligned in only a few hours. It will also make use of multiple processors, where present.

### Clustal home page: http://www.clustal.org

If you are using Windows OS, there are no additional configuration steps required, as *ClustalO* executable file is included to the UGENE distribution package. Otherwise:

- Install the *Clustal* program on your system.
- Set the path to the ClustalW executable on the External tools tab of UGENE Application Settings dialog.

Now you are able to use ClustaOl from UGENE.

Open a multiple sequence alignment file and select the *Align with ClustalO* item in the context menu or in the *Actions* main menu. The *Align with ClustalO* dialog will appear (see below), where you can adjust the following parameters:

Number of iterations - number of (combined guide tree/HMM) iterations.

Max number guidetree iterations - maximum guide tree iterations.

Max number of HMM iterations ---maximum number of HMM iterations.

Number of CPUs being used - number of processors to use.

Set options automatically - set options automatically (might overwrite some of your options).

U Align with Clustal Omega	? <mark>x</mark>
Iteration	
Number of iterations	1
Max number guidetree iterations	0
Max number of HMM iterations	0
Miscellaneous	
Number of CPUs being used	8 🌻
Set options automatically	
Align Cancel	Help

# **Kalign Aligning**

Kalign is a fast and accurate multiple sequence package designed to align large numbers of protein sequences.

### Kalign home page: KAlign

To use *Kalign* open a multiple sequence alignment file and select the *Align with Kalign* item in the context menu or in the *Actions* main menu. The following dialog appears:

📰 Align with Kalign	? ×
Advanced options	
Gap open penalty	217.00
Gap extension penalty	39.40 🔅
Terminal gap penalty	292.60 🚖
Bonus score	28.30 🗘
Translate to amino when aligning	
Output file 1. The Standard Genetic Code	<b>~</b>
	Align Cancel Help

#### The following parameters are available:

Gap opening penalty - indicates the penalty applied for opening a gap. The penalty must be negative.

Gap extension penalty — indicates the penalty applied for extending a gap.

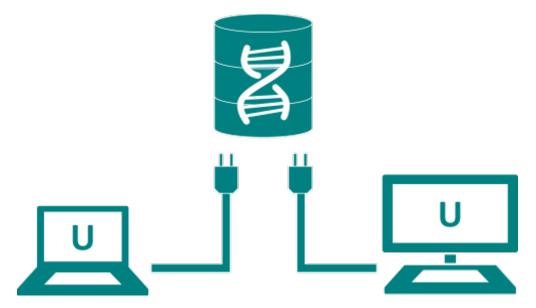
Terminal gap penalty — the penalty to extend gaps from the N/C terminal of protein or 5'/3' terminal of nucleotide sequences.

Bonus score - a bonus score that is added to each pair of aligned residues.

Translate to amino when aligning - translates an alignment to amino when aligning.

# **Shared Database**

The rational storage of biological data is an ever-present issue. It is not only about large data sizes, but also about the requirement of simultaneous access to them by several scientists. For instance, a few researchers from a lab may need to work on the same data, like a set of primers or data produced by sequencing. That information has to be updated and synchronized between different users and kept in a common storage. That is what UGENE Shared Database is intended for.



To start sharing data via UGENE you need to deploy a public database server. MySQL servers are currently supported. See this paragraph f or details about the required server configuration.

After that any UGENE user (who knows the correct login/password, however) can connect to the database. The connected database is shown in the Project View as a document exactly the same way as if the data were located on the local computer.

As described in this paragraph the users can have a read-only access to the database or be able to modify its content. A user with a read-only access can:

- Browse the data in the database
- Open the data in the UGENE views
- Export the data to the local computer

Users with write access, in addition, can:

- Add new objects to the database
- Create new folders to order the data in the database
- Modify the folders hierarchy inside the database (using drag'n'drop)
- Rename objects and folders
- Delete existed objects
- Delete folders

All UGENE instances connected to a database constantly monitors the state of the database and shows changes, made by other users.

UGENE accesses large remote data, such as NGS assemblies, so that only a viewed part of them is loaded to a client computer. So, if you store the assembly data on a server, the data can be browsed in the UGENE Assembly Browser on a local computer almost instantly, without the need to copy the data on the computer or use the hard disk space.

For details see the documentation below:

- Configuring Database
- Connecting to a Shared Database
- Adding Data to the Database
- Database in the Project
- Deleting Data
- Drag'n'drop in the Database
- Exporting Objects from the Database

# **Configuring Database**

To make use of a shared database follow the steps below:

1. Deploy a MySQL database server

We recommend you to download MySQL binaries from the official site. Note that UGENE supports MySQL versions 5.5 and higher. Here you can also find instructions on how to install and launch a MySQL server instance for each platform.

2. Create an empty database

Log in to the MySQL server as a user with administrative privileges (you must be able to create databases and users, and to grant privileges to the created users). In the MySQL console or in your favorite SQL browser execute the following command: > CREATE DATABASE `your\_database\_name`;

3. Create database users

You may probably want to limit possible influence on the shared database by the UGENE users who will use it. In this case create a distinct MySQL user for each UGENE user (or a group of users). In order to do this, execute the following commands: > CREATE USER `user\_nickname` IDENTIFIED BY 'user\_password';

Decide whether the created user is allowed to modify the database content or only to view it. In the first case execute the command below:

> GRANT ALL PRIVILEGES ON your\_database\_name.\* TO `user\_nickname`@'%' IDENTIFIED BY 'user\_password'; and in the second case execute:

> GRANT SELECT ON your\_database\_name.\* TO `user\_nickname`@'%' IDENTIFIED BY 'user\_password';

4. Use the database from a UGENE instance

The database with "your\_database\_name" is now available from a UGENE instance (version 1.14 or higher). It's time to try it out and fill it with some initial data. To do it, open UGENE and connect to the database. As we need to add the data to the database, use "user\_nickname" and "user\_password" of a user with privileges to modify the database. As soon as connection is established, a dd the required data to the database.

From now on the data will be available for all users from this and other UGENE instances who connected to the same database.

## **Connecting to a Shared Database**

To start using the shared database you need to have a running public MySQL database server. Usually the system administrator of your department does it. You should ask him or her to give you the access to a MySQL database. Particularly, you need a few parameters to connect to the database: the IP-address of the server (the computer where a MySQL server is running), a user name for the MySQL database and a password. You can also install a MySQL server by yourself on any public computer you have an access to (even on your workstation), following the steps described in the Configuring Database section.

To connect to the database use the File->Connect to shared database main menu item. The following dialog appears:

U Shared Databases Connections	? ×
UGENE public database	Add Delete Edit
Connect Disconnect	Close Help

To add new connection click on the Add button. The following dialog appears:

U Connectio	n Settings		? ×
Connection n	ame:		
-Database lo	ocation	-Authentication	data
Host:		Login:	
Port:	3306	Password:	
Database:		Remember me	
OK Cancel Help			

Here you need to specify *Host* (IP-address of the server), *Port* (number of the port used by the MySQL server) and *Database* (name of the database). You may also fill *Login* and *Password* fields. Otherwise, you are asked to input them every time you are establishing this connection until you check the *Remember me* box. Click on the *OK* button, then the connection is created and the appropriate item appears in the previous *Shared Database Connections* dialog.

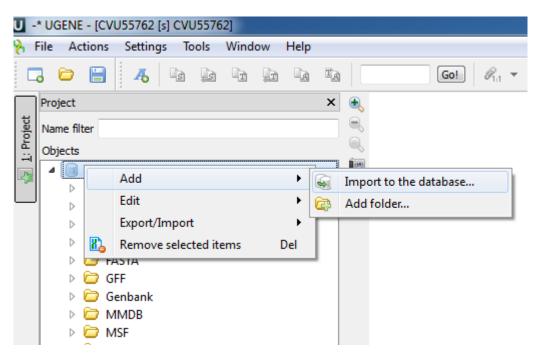
If you want to use already existing connection choose the appropriate item in the *Shared Database Connections* dialog and press the *Conne ct* button. This can also be done by double clicking the item. If the specified database is empty, UGENE has to initialize it. This routine is done only once. In this case you get an appropriate message box, asking whether to initialize the database or not. If you choose Yes the database is populated with UGENE data structures, if *No* it remains empty and UGENE does not connect to it.

If you want to delete some connection select it in the Shared Database Connections dialog and click on the Delete button. You may also edit connection parameters using the Edit button.

An established connection can be terminated by pressing the *Delete* button. The same effect is produced by removing the database document item from Project View.

# Adding Data to the Database

To add data to the database use the Add->Import to the database context menu item of the database in the project tree view. Also you can drag'n'drop it to a shared database folder.



The following dialog will appear:

U Import to the Database		? ×
Item to import	Import to	Add files Add folder Add objects
		General options Remove

Here you can add to the database files, folders or other objects from the current Project View. To do this use corresponding buttons. After specifying your data click on the *Import* button. The data will be imported and appear in the database data tree. Also you can change import settings. To do this click on the *General options* button. The following dialog will appear:

U Database Import Default Options	? ×
Destination folder: /	
Files and folders options	
Process directories recursively	
Keep folders structure	
Create a subfolder for the top level folder	
Create a subfolder for each file	
Import unrecognized files	
Multi-sequence files import policy	
Import as separate sequences	
Merge into a single sequence	
Number of 'unknown' symbols ('N 'for nucleic or 'X' for amino) between parts: 10 bases	A.V.
O Join into alignment	
Documents and objects options	
Create a subfolder for each document	
OK Cancel	Help

Available parameters are described below:

*Process directories recursively* - if this option is checked, the import procedure recreates the hierarchy of the imported directories and all their sub-directories in the database. Otherwise, only the content of the directories, specified for import, is uploaded to the *Destinati on folder* without taking into account any sub-directories.

*Create a subfolder for each file* - if this option is checked, for each file uploaded to the database a new folder is created, having the same name as the file, and the file content is placed in the folder. Otherwise, the file data are imported into the *Destination folder*.

*Import as separate sequences* - if this option is selected and an uploaded file contains several sequences, they are represented by distinct sequence objects in the database after the import is done.

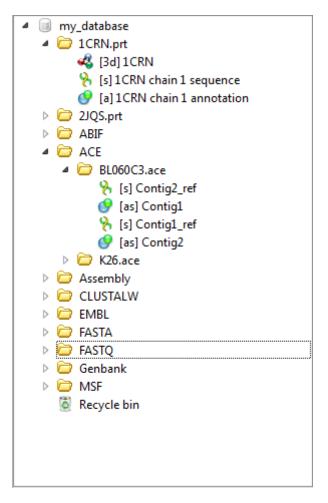
*Merge into a single sequence* - if this option is selected and an uploaded file contains several sequences, they are merged into a single sequence object in the database after the import is done.

Join into alignment - if this option is selected and an uploaded file contains several sequences, they are joined into a multiple sequence alignment object in the database after the import is done.

*Create a subfolder for each document* - if this option is checked, for each document or object uploaded to the database a new folder is created, having the same name as the file, and the data are placed in the folder. Otherwise, the data are imported into the *Destination folder*.

# **Database in the Project**

The database in UGENE Project View looks like as a tree with folders and objects:



You can add a new folder to the database tree. To do that use the *Add->Add folder* database context menu item. To add a subfolder to some existing folder use the *Add->Add folder* folder context menu item. To delete an object or a folder press the *Delete* button or drag'n'drop it to the *Recycle bin*.

In this version of UGENE objects in the database are read-only. Nevertheless, there is a workaround to edit them. First, you need export the objects to files on your computer using the *Export/Import* object context menu. Then you can change that files locally, upload them to database and, finally, delete the originals.

If new data are added to the database by another user or removed from it, UGENE detects this and shows updates automatically in Project View.

# **Deleting Data**

To remove an object or a folder select it and press the Delete button or drag it to the Recycle bin folder.

All removed items are located in the Recycle bin folder.

To delete all files from Recycle bin click on the Empty recycle bin context menu item of the Recycle bin.

To restore objects from the Recycle bin select them and call the Restore selected items context menu item.

When the database is updated outside, UGENE shows these changes on your computer automatically.

(1) You cannot delete any object from *Recycle bin* if it is opened on the other computer. This situation can appear if the object was being viewed by another user when you moved it to *Recycle Bin*.

# Drag'n'drop in the Database

In the database tree you can drag'n'drop objects between folders, folders between folders. Also you can drag'n'drop other objects and documents from project to the database.

# **Exporting Objects from the Database**

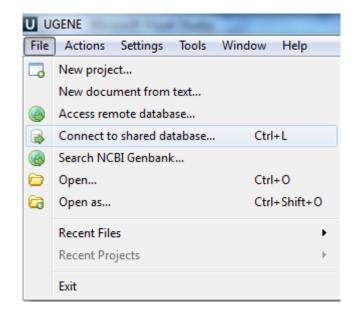
The objects in the database can not be altered though they can be deleted. To edit the objects you need to export them to the project, then make you modifications locally and replace existing originals. More detailed information about exporting you can find here.

# **UGENE Public Storage**

UGENE provides the free-to-use public bioinformatics data storage. This storage keeps DNA sequences of several popular genomes such as human, mouse, drosophila melanogaster, etc. and hundreds of plasmid sequences.

Follow the instructions for accessing the storage:

1. Use the menu File -> Connect to shared database (or press the Ctrl+L shortcut).



2. Choose the predefined "UGENE public database" item and click the Connect button.

U Shared Databases Connections	? ×
UGENE public database	Add
	Delete
	1 Edit
Connect Disconnect	Close Help

3. Browse the storage content.

Project ×
Name filter Read-only access
Objects
<ul> <li>GENE public database</li> <li>Recycle bin</li> </ul>
<ul> <li>genomes</li> </ul>
Arabidopsis thaliana (TAIR 10)
C. elegans (ce6)
🖻 🗁 Drosophila melanogaster (dm3)
🔺 🗁 Escherichia coli str. K-12 substr. MG16
📄 [t] INFO 🔫 🥌
(a) NC_000913 features
🖻 🗁 Human (hg19)
Human Immunodeficiency Virus (HIV
Mouse (mm9)
Mycobacterium tuberculosis (NC_000
Salmonella Enterica(NC_016856.1)
Vibrio cholerae (NC_002505)
Veast (Saccharomyces cerevisiae) (sac
Zebrafish (Danio rerio) (danRer7)
Plasmids

- The storage document is marked with the lock icon 🖾. It means that the storage provides the read-only access. Any data modifications are unavailable for such type of access (importing, removing or replacing of data).
- Each genome folder contains the INFO text object. It is the information about the genome or its source.
- You can export the data to your computer for working with the data locally.
- There are hundreds of plasmids in the storage. Use the name filter for fast navigating and searching an interesting plasmid:

Project	×
Name filter pfr-luc	
[s] AF058756_Cloning vector pFR-Luc,	•
[a] AF058756_Cloning vector pFR-Luc	, compl

The list of available genomes:

- Human (hg19)
- Mouse (mm9)
- Arabidopsis thaliana (TAIR 10)
- C. elegans (ce6)
- Drosophila melanogaster (dm3)
- Escherichia coli str. K-12 substr. MG1655 K12(NC\_000913.3)
- Human Immunodeficiency Virus (HIV-2)
- Mycobacterium tuberculosis (NC\_000962.3)
- Salmonella Enterica(NC\_016856.1)
- Vibrio cholerae (NC\_002505)
- Yeast (Saccharomyces cerevisiae)(sacCer3)
- Zebrafish (Danio rerio) (danRer7)

# **UGENE** Command Line Interface

UGENE command line interface (CLI) was developed keeping in mind the following principles:

- · To make it as easy as popular shell commands.
- To include all significant UGENE features.
- To allow users to add their own commands.

To use UGENE CLI make sure to add the path to the UGENE executable to your %PATH% environment variable.

The general syntax is the following:

```
ugene [[--task=]task_name] [--task_parameter=value ...] [-task_parameter value ...]
[--option[=value]] [-option[ value]]
```

Here:

```
task_name — task to execute, it can be one of the predefined tasks or a task you have created.
```

task\_parameter — parameter of the specified task. Some parameters of a task are required, like in and out parameters of some tasks.

```
option — one of the CLI options.
```

```
See the example below:
```

ugene align --in=COI.aln -out result.aln -log-level-details

- CLI Options
- **CLI Predefined Tasks** 
  - Format Converting Sequences
  - Converting MSA
    - Extracting Sequence
    - Finding ORFs
    - Finding Repeats
    - Finding Pattern Using Smith-Waterman Algorithm
    - Adding Phred Quality Scores to Sequence
    - ٠ Local BLAST Search
    - Local BLAST+ Search
    - Remote NCBI BLAST and CDD Requests .
    - Annotating Sequence with UQL Schema
    - Building Profile HMM Using HMMER2
    - Searching HMM Signals Using HMMER2
    - Aligning with MUSCLE
    - Aligning with ClustalW
    - Aligning with ClustalO Aligning with Kalign

    - Aligning with MAFFT
    - Aligning with T-Coffee
    - **Building PFM**
    - Searching for TFBS with PFM
    - Building **PWM**
    - Searching for TFBS with Weight Matrices
    - Building Statistical Profile for SITECON ٠
    - Searching for TFBS with SITECON
    - Fetching Sequence from Remote Database
    - Gene-by-Gene Report
    - **Reverse-Complement Converting Sequences**

    - Variants Calling Generating DNA Sequence
- Creating Custom CLI Tasks

# CLI Options

--help | -h [<option\_name> | <task\_name>]

Shows help information. For example:

```
ugene --help ## Shows general UGENE CLI help.
ugene -h
ugene --help=<option_name> ## Shows help for the <option_name> option.
ugene --help=<task_name> ## Shows help for the <task_name> task.
ugene -h <task_name>
```

--task=<task\_name> [<task\_parameter>=value ...]

Specifies the task to run. A user-defined UGENE workflow schema can be used as a task name. For example:

```
ugene --task=align --in=COI.aln -out result.aln
ugene --task=C:\myschema.uwl --in=COI.aln --out=res.aln
```

#### --log-no-task-progress

A task progress is shown by default when a task is running. This option specifies not to show the progress.

### --log-level="[<category1>=]<level1> [, ...]"

Sets the log level per category. If a category is not specified, the log level is applied to all categories.

The following categories are available:

- "Algorithms"
- "Console"
- "Core Services"
- "Input/Output"
- "Performance"
- "Remote Service"
- "Scripts"
- "Tasks".

The following log levels are available: TRACE, DETAILS, INFO, ERROR or NONE.

By default, loglevel=ERROR.

For example:

```
ugene --log-level=NONE
ugene --log-level="Tasks=DETAILS, Console=DETAILS"
```

#### --log-format="<format\_string>"

Specifies the format of a log line.

Use the following notations: L - level, C - category, YYYY or YY - year, MM - month, dd - day, hh - hour, mm - minutes, ss - seconds, zzz - milliseconds.

By default, logformat="[L][hh:mm]".

--license

Shows license information.

--lang=language\_code

Specifies the language to use (e.g. for the log output). The following values are available:

CS (Czech)

- EN (English)
- RU (Russian)

### --log-color-output

If log output is enabled, this option make it colored: *ERROR* messages are displayed in red, *DETAILS* messages are displayed in green, *TRACE* messages are displayed in blue.

#### --session-db

Session database is stored in the temporary file that is created for every UGENE run. But it can supplied with the command line argument. If the supplied file does not exest it will be created. The session database file is removed after closing of UGENE.

### For example:

ugene --session-db=D:/session.ugenedb

--version

Shows version information.

--tmp-dir=<path\_to\_file>

Path to teporary folder.

--ini-file=<path\_to\_file>

Loads configuration from the specified .ini file. By default the UGENE.ini file is used.

--genome-aligner

UGENE Genome Aligner is an efficient and fast tool for short read alignment. It has 2 work modes: build index and align short reads (default mode).

If there is no index available for reference sequence it will be built on the fly.

Usage: ugene --genome-aligner { --option[=argument] }

The following options are available:

--build-index Use this flag to only build index for reference sequence.

--reference Path to reference genome sequence

--short-reads Path to short-reads data in FASTA or FASTQ format

--*index* Path to prebuilt index (base file name or with .idx extension). If not set, index is searched in system temporary directory. If --build-index option is applied, index will be saved to specified path.

--result Path to output alignment in UGENEDB or SAM format (see --sam)

--memsize Memory size (in Mbs) reserved for short-reads. The bigger value the faster algorithm works. Default value depends on available system memory.

--ref-size Index fragmentation size (in Mbs). Small fragments better fit into RAM, allowing to load more short reads. Default value is 10.

--n-mis Absolute amount of allowed mismatches per every short-read (mutually exclusive with --pt-mis). Default value is 0.

--pt-mis Percentage amount of allowed mismatches per every short-read (mutually exclusive with --n-mis). Default value is 0.

--rev-comp Use both the read and its reverse complement during the aligning.

--best Report only about best alignments (in terms of mismatches).

--omit-size Omit reads with qualities lower than the specified value. Reads which have no qualities are not omitted. Default value is 0.

--sam Output aligned reads in SAM format. Default value is false.

For example:

```
Build index for reference sequence:
ugene --genome-aligner --build-index --reference=/path/to/ref
Align short reads using existing index:
ugene --genome-aligner --reference=/path/to/ref --short-reads=/path/to/reads
--result=/path/to/result
```

# CLI Predefined Tasks

Using current version of UGENE you can perform the following tasks by running a simple command:

- Format Converting Sequences
- Converting MSA
- Extracting Sequence
- ٠ Finding ORFs
- Finding Repeats
- Finding Pattern Using Smith-Waterman Algorithm
- Adding Phred Quality Scores to Sequence ٠
- Local BLAST Search
- Local BLAST+ Search
- ٠ Remote NCBI BLAST and CDD Requests
- Annotating Sequence with UQL Schema
- Building Profile HMM Using HMMER2
- Searching HMM Signals Using HMMER2
- Aligning with MUSCLE
- Aligning with ClustalW Aligning with ClustalO
- Aligning with Kalign
- Aligning with MAFFT
- Aligning with T-Coffee
- **Building PFM**
- Searching for TFBS with PFM
- Building **PWM**
- Searching for TFBS with Weight Matrices •
- Building Statistical Profile for SITECON ٠
- Searching for TFBS with SITECON
- Fetching Sequence from Remote Database
- Gene-by-Gene Report
- **Reverse-Complement Converting Sequences**
- Variants Calling
- Generating DNA Sequence

## **Format Converting Sequences**

#### Task Name: convert-seq

Converts a sequence from one format to another.

#### Parameters:

- in input sequence file. [String, Required]
- out name of the output file. [String, Required]
- format format of the output file. [String, Optional]

The following values are available:

- fasta
- fastq
- genbank
- gff
- raw

### Example:

ugene convert-seq --in=human\_T1.fa --out=human\_T1.gbk --format=genbank

## **Converting MSA**

Task Name: convert-msa

Converts a multiple sequence alignment file from one format to another.

#### Parameters:

in — input multiple sequence alignment file. [String, Required]

out - name of the output file. [String, Required]

format - format of the output file. [String, Optional]

The following values are available:

- clustal (default)
- fasta
- mega
- msf
- nexus
- phylip-interleaved
- phylip-sequential
- stockholm

#### Example:

ugene convert-msa --in=CBS.sto --out=CBS --format=msf

## **Extracting Sequence**

### Task Name: extract-sequence

Extracts annotated regions from an input sequence.

### Parameters:

in - semicolon-separated list of input files. [String, Required]

out --- output file. [String, Required]

annotation-names - list of annotations names which will be accepted or filtered. [String, Optional]

annotation-names-file - file with annotation names, separated with whitespaces which will be accepted or filtered. [String, Optional]

accumulate - accumulate all incoming data in one file or create separate files for each input. In the latter case, an incremental numerical suffix is added to the file name (using 'True' by default). [Boolean]

*accept-or-filter* — if set to *true*, accepts only the specified annotations, if set to *false*, accepts all annotations except the specified ones. [Boolean, Optional]

*complement* — complements the annotated regions if the corresponding annotation is located on the complement strand. [Boolean, Optional]

extend-left — extends the resulting regions to the left for the specified number of base symbols. [Number, Optional]

extend-right --- extends the resulting regions to the right for the specified number of base symbols. [Number, Optional]

gap-length — inserts a gap of the specified length between the merged annotations.

transl - translates the annotated regions. [Boolean, Optional]

#### Example:

ugene extract-sequence --in=sars.gb --out=res.fa --annotation-names=gene

## **Finding ORFs**

#### Task Name: find-orfs

Searches for Open Reading Frames (ORFs) in nucleotide sequences and saves the regions found as annotations.

#### Parameters:

in - semicolon-separated list of input files. [String, Required]

out - output file with the annotations. [String, Required]

name - name of the annotated regions. [String, Optional, Default: "ORF"]

min-length ---- ignores ORFs shorter than the specified length. [String, Optional, Default: 100]

*require-stop-codon* — ignores boundary ORFs that last beyond the search region (i.e. have no stop codon within the range). [Boolean, Optional, Default: false]

require-init-codon — allows ORFs starting with any codon other than terminator. [Boolean, Optional, Default: true]

*allow-alternative-codons* — allows ORFs starting with alternative initiation codons, accordingly to the current translation table. [Boolean, Optional, Default: false]

#### Example:

ugene find-orfs --in=human\_T1.fa --out=result.gb --require-init-codon=false

### Finding Repeats

#### Task Name: find-repeats

Searches for repeats in sequences and saves the regions found as annotations.

### Parameters:

in - semicolon-separated list of input files. [String, Required]

out -- output file with the annotations. [String, Required]

name - name of the annotated regions. [String, Optional, Default: "repeat\_unit"]

min-length — minimum length of the repeats. [Number, Optional, Default: 5]

identity - percent identity between repeats. [Number, Optional, Default: 100]

min-distance — minimum distance between the repeats. [Number, Optional, Default: 0]

max-distance — maximum distance between the repeats. [Number, Optional, Default: 5000]

inverted — if true, searches for the inverted repeats. [Boolean, Optional, Default: false]

### Example:

ugene find-repeats --in=murine.gb --out=murine\_repeats.gb --identity=99

## Finding Pattern Using Smith-Waterman Algorithm

### Task Name: find-sw

Searches for a pattern in a nucleotide or protein sequence using the Smith-Waterman algorithm and saves the regions found as annotations.

#### Parameters:

- in input sequence file. [String, Required]
- out output file with the annotations. [String, Required]
- name name of the annotated regions. [String, Optional, Default: "misc\_feature"]
- ptrn --- subsequence pattern to search for (e.g. AGGCCT). [String, Required]
- score percent identity between the pattern and a subsequence. [Number, Optional, Default: 90]
- matrix scoring matrix. [String, Optional, Default: "Auto"]

Among others the following values are available:

- blosum62
- dna
- rna
- dayhoff
- gonnet
- pam250
- etc.

The matrices available are stored in the \$UGENE\data\weight\_matrix directory.

filter --- results filtering strategy. [String, Optional, Default: "filter-intersections"]

The following values are available:

- filter-intersections
- none

### Example:

ugene find-sw --in=human\_T1.fa --out=sw.gb --ptrn=TGCT --filter=none

### Adding Phred Quality Scores to Sequence

### Task Name: join-quality

Adds Phread quality scores to a sequence and saves the result to the output FASTQ file.

#### Parameters:

in - input sequence file. [String, Required]

quality - input Phred quality scores file. [String, Required]

out --- output FASTQ file. [String, Required]

#### Example:

ugene join-quality --in=e\_coli.fa --quality=e\_coli.qual --out=res.fastq

## Local BLAST Search

Task Name: local-blast

Performs a search on a local BLAST database using old version of the NCBI BLAST.

BLAST is used as an external tool and must be installed on your system.

#### Parameters:

toolpath — path to the blastall executable. By default, the path specified in the Application Settings is applied. [String, Optional, Default: "default"]

*tmpdir* — directory for temporary files. By default, the path specified in the *Application Settings* is applied. [String, Optional, Default: "default"]

in - semicolon-separated list of input sequence files. [String, Required]

dbpath --- path to the BLAST database files. [String, Required]

dbname - base name of the BLAST database files. [String, Required]

out - output Genbank file, the results of the search are stored as annotations. [String, Required]

name - name of the annotations. [String, Optional, Default: "blast result"]

p — type of the BLAST search. [String, Optional, Default: "blastn"]

The following values are available:

- blastn
- blastp
- blastx
- tblastn
- tblastx

e — expectation value threshold. [Number, Optional, Default: 10]

### Example:

ugene local-blast --in=input.fa --dbpath=. --dbname=mydb --out=output.gb

## Local BLAST+ Search

Task Name: local-blast+

Performs a search on a local BLAST database using BLAST+.

BLAST+ is used as an external tool and must be installed on your system.

#### Parameters:

toolpath — path to an appropriate BLAST executable (e.g. blastn, blastp, etc.). By default, the path specified in the Application Settings is applied. [String, Optional, Default: "default"]

*tmpdir* — directory for temporary files. By default, the path specified in the *Application Settings* is applied. [String, Optional, Default: "default"]

in - semicolon-separated list of input sequence files. [String, Required]

dbpath --- path to the BLAST database files. [String, Required]

dbname — base name of the BLAST+ database files. [String, Required]

out --- output Genbank file, the results of the search are stored as annotations. [String, Required]

name - name of the annotations. [String, Optional, Default: "blast result"]

p-type of the BLAST search. [String, Optional, Default: "blastn"]

The following values are available:

- blastn
- blastp
- blastx
- tblastn
- tblastx

e - expectation value threshold. [Number, Optional, Default: 10]

### Example:

ugene local-blast+ --in=input.fa --dbpath=. --dbname=mydb --out=output.gb

## **Remote NCBI BLAST and CDD Requests**

#### Task Name: remote-request

Performs remote requests to the NCBI. Saves the results as annotations.

#### Parameters:

in - semicolon-separated list of input files. A file can be of any format containing sequences or alignments. [String, Required]

db — database to search in. [String, Optional, Default: "ncbi-blastn"]

The following databases are available:

- "ncbi-blastn" for nucleotide sequences
- "ncbi-cdd" for amino acid sequences
- · "ncbi-blastp" for amino acid sequences

out - output Genbank file. [String, Required]

*eval* — specifies the statistical significance threshold for reporting matches against database sequences. [Number, Optional, Default: 10]

hits - maximum number of hits, that will be shown. [Number, Optional, Default: 10]

name — name of the result annotations. If not set, name will be specified with the "cdd" result or the "blast" result. [String, Optional, Default: "cdd" or "blast"]

short - optimizes search for short sequences. [Boolean, Optional, Default: false]

*blast-output* — path to the file with the NCBI-BLAST output (only for the "ncbi-blastp" and "ncbi-blastn" databases). [Boolean, Optional, Default: the file is not saved]

#### Example:

ugene remote-request --in=seq.fa --db=ncbi-blastp --out=res.gb

## Annotating Sequence with UQL Schema

#### Task Name: query

Annotates a sequence in compliance with a UGENE Query Language (UQL) schema. This allows to analyze a sequence using different algorithms at the same time imposing constraints on the positional relationship of the results.

To learn more about the UQL schemas read the Query Designer Manual.

#### Parameters:

in - semicolon-separated list of input sequence files. [String, Required]

out - output Genbank file with the annotations. [String, Required]

schema — UQL schema. [String, Required]

merge — if true, merges regions of each result into a single annotation. [Boolean, Optional, Default: false]

offset — if merge is set to true, specified left and right offsets for merged annotations. [Number, Optional, Default: 0]

#### Example:

ugene query --in=input.fa --out=result.gb --schema=RepeatsWithORF.uql

#### Task Name: hmm2-build

Builds a profile HMM using the HMMER2 tools.

#### Parameters:

in - semicolon-separated list of input multiple sequence alignment files. [String, Required]

out --- output HMM file. [String, Required]

name — name of the profile HMM. [String, Optional, Default: "hmm\_profile"]

calibrate — enables/disables calibration. [Boolean, Optional, Default: true]

seed - random seed, a non-negative integer. [Number, Optional, Default: 0]

#### Example:

ugene hmm2-build --in=CBS.sto --out=CBS.hmm

## Searching HMM Signals Using HMMER2

#### Task Name: hmm2-search

Searches each input sequence for the significantly similar sequence that matches to all specified profile HMM using the HMMER2 tool.

#### Parameters:

seq - semicolon-separated list of the input sequence files. [String, Required]

hmm - semicolon-separated list of the input HMM files. [String, Required]

out - output file with annotations. [String, Required]

name - name of the result annotations. [String, Optional, Default: "hmm\_signal"]

e-val — e-value that can be used to exclude low-probability hits from the result. [Number, Optional, Default: 1e-1]

score — score based filtering which is an alternative to e-value filtering to exclude low-probability hits from the result. [Number, Optional, Default: -1000000000]

#### Example:

ugene hmm2-search --seq=CBS\_seq.fa --hmm=CBS.hmm --out=CBS\_hmm.gb

## Aligning with MUSCLE

#### Task Name: align

Performs multiple sequence alignment with MUSCLE algorithm and saves the resulting alignment to file. Source data can be of any format containing sequences or alignments.

#### Parameters:

in Input alignment [Url datasets] '2' number max-iterations Maximum of iterations (using by default) [Number] mode - Selector of preset configurations, that give you the choice of optimizing accuracy, speed, or some compromise between the favors 'MUSCLE two. The default accuracy (using default' by default) [Number] range - Whole alignment or column range e.g. 1..100 (using 'Whole alignment' by default) [String] stable Do not rearrange aligned sequences (using 'True' by default) [Boolean] format format of Document output alignment (using 'clustal' by default) [String] out - Output alignment [String]

#### Example:

ugene align --in=test.aln --out=test\_out.aln --format=clustal

## Aligning with ClustalW

#### Task Name: align-clustalw

Multiple sequence alignment with ClustalW.

ClustalW is used as an external tool and must be installed on your system.

#### Parameters:

toolpath — path to the ClustalW executable. By default, the path specified in the Application Settings is applied. [String, Optional, Default: "default"]

tmpdir - directory for temporary files. [String, Optional]

in - semicolon-separated list of input files. [String, Required]

out --- output file. [String, Required]

format - format of the output file. [String, Optional]

#### Example:

ugene align-clustalw --in=COI.aln --out=COI.sto --format=stockholm

## **Aligning with ClustalO**

Task Name: align-clustalo

Create alignment with ClustalO. ClustalO is a general purpose multiple sequence alignment program for proteins.

(1) ClustalO is used as an *external tool* and must be installed on your system.

#### Parameters:

in - Input alignment [Url datasets]

format - Document format of output alignment (using 'clustal' by default) [String]

out - Output alignment [String]

max-guidetree-iterations - Maximum number guidetree iterations (using '0' by default) [Number]

max-hmm-iterations - Maximum number of HMM iterations (using '0' by default) [Number]

iter - Number of (combined guide-tree/HMM) iterations (using '1' by default) [Number]

toolpath - ClustalO location (using the path specified in UGENE by default) [String]

auto - Set options automatically (might overwrite some of your options) (using 'False' by default) [Boolean]

tmpdir - Directory to store temporary files (using UGENE temporary directory by default) [String]

#### Example:

ugene align-clustalw --in=test.aln --out=test\_out.aln --format=clustal

## Aligning with Kalign

Task Name: align-kalign

Multiple sequence alignment with Kalign.

#### Parameters:

in - semicolon-separated list of input files. [String, Required]

out --- output file in the ClustalW format. [String, Required]

#### Example:

ugene align-kalign --in=COI.aln --out=COI\_aligned.aln

### Aligning with MAFFT

#### Task Name: align-mafft

Multiple sequence alignment with MAFFT.

MAFFT is used as an external tool and must be installed on your system.

#### Parameters:

toolpath — path to the MAFFT executable. By default, the path specified in the *Application Settings* is applied. [String, Optional, Default: "default"]

tmpdir - directory for temporary files. [String, Optional]

in - semicolon-separated list of input files. [String, Required]

out --- output file. [String, Required]

format - format of the output file. [String, Required]

op — penalty for opening a gap. [Number, Optional]

ep - penalty for extending a gap. [Number, Optional]

maxiterate --- maximum number of cycles of iterative refinement. [Number, Optional]

#### Example:

ugene align-mafft --in=COI.aln --out=COI\_aligned.aln

## Aligning with T-Coffee

#### Task Name: align-tcoffee

Create alignment with T-Coffee. T-Coffee is a collection of tools for computing, evaluating and manipulating multiple alignments of DNA, RNA, Protein Sequences.

T-Coffee is used as an external tool and must be installed on your system.

#### Parameters:

gap-ext-penalty - Gap Extension Penalty. Positive values give rewards to gaps and prevent the alignment of unrelated segments (using '0' by default) [Number]

gap-open-penalty - Gap Open Penalty. Must be negative, best matches get a score of 1000 (using '-50' by default) [Number]

iter-max - Number of iteration on the progressive alignment: 0 - no iteration (default), -1 - Nseq iterations (using '0' by default) [Number]

toolpath - T-Coffee location (using the path specified in UGENE by default) [String]

tmpdir - Directory to store temporary files (using UGENE temporary directory by default) [String]

in - Input alignment [Url datasets]

format - Document format of output alignment (using 'clustal' by default) [String]

out - Output alignment [String]

Example:

ugene align-tcoffee --in=test.aln --out=test\_out.aln --format=clustal

## **Building PFM**

Task Name: pfm-build

Builds a position frequency matrix from a multiple sequence alignment file.

#### Parameters:

in - semicolon-separated list of input MSA files. [String, Required]

out - output file. [String, Required]

type — type of the matrix. [Boolean, Optional, Default: false]

The following values are available:

- true (dinucleic type)
- false (mononucleic type)

Dinucleic matrices are more detailed, while mononucleic ones are more useful for small input data sets.

#### Example:

ugene pfm-build --in=COI.aln --out=result.pfm

## Searching for TFBS with PFM

#### Task Name: pfm-search

Searches for transcription factor binding sites (TFBS) with position weight matrices (PWM) converted from input position frequency matrices (PFM) and saves the regions found as annotations.

#### Parameters:

seq — semicolon-separated list of input sequence files to search TFBS in. [String, Required]

matrix - semicolon-separated list of the input PFM. [String, Required]

out - output Genbank file.

name - name of the annotated regions. [String, Optional, Default: "misc\_feature"]

type — type of the matrix. [Boolean, Optional, Default: false]

The following values are available:

- true (dinucleic type)
- false (mononucleic type)

Dinucleic matrices are more detailed, while mononucleic ones are more useful for small input data sets.

algo --- algorithm used to convert a PFM to a PWM. [String, Optional, Default: "Berg and von Hippel"]

The following values are available:

- · Berg and von Hippel
- Log-odds
- Match
- NLG

score -- minimum percentage score to detect TFBS. [Number, Optional, Default: 85]

strand - strands to search in. [Number, Optional, Default: 0]

The following values are available:

- 0 (both strands)
- 1 (direct strand)
- -1 (complement strand)

#### Example:

ugene pfm-search --seq=in.fa --matrix=MA0265.1.pfm;MA0266.1.pfm --out=res.gb

## **Building PWM**

#### Task Name: pwm-build

Builds a position weight matrix from a multiple sequence alignment file.

#### Parameters:

in - semicolon-separated list of input MSA files. [String, Required]

out --- output file. [String, Required]

type — type of the matrix. [Boolean, Optional, Default: false]

The following values are available:

- true (dinucleic type)
- false (mononucleic type)

Dinucleic matrices are more detailed, while mononucleic ones are more useful for small input data sets.

algo - algorithm used to build the matrix. [String, Optional, Default: "Berg and von Hippel"]

The following values are available:

- Berg and von Hippel
- Log-odds
- Match
- NLG

#### Example:

ugene pwm-build --in=COI.aln --out=result.pwm

## Searching for TFBS with Weight Matrices

Task Name: pwm-search

Searches for transcription factor binding sites (TFBS) with position weight matrices (PWM) and saves the regions found as annotations.

#### Parameters:

seq - semicolon-separated list of input sequence files to search TFBS in. [String, Required]

matrix - semicolon-separated list of the input PWM. [String, Required]

out --- output Genbank file.

name - name of the annotated regions. [String, Optional, Default: "misc\_feature"]

min-score - minimum percentage score to detect TFBS. [Number, Optional, Default: 85]

strand - strands to search in. [Number, Optional, Default: 0]

The following values are available:

- 0 (both strands)
- 1 (direct strand)
- -1 (complement strand)

#### Example:

ugene pwm-search --seq=input.fa --matrix=Aro80.pwm;Aft1.pwm --out=res.gb

## **Building Statistical Profile for SITECON**

#### Task Name: sitecon-build

Builds a statistical profile for SITECON. It can be later used to search for TFBS.

#### Parameters:

in — semicolon-separated list of input DNA multiple sequence alignment files. An input file must not contain gaps. [String, Required]

out — output file. If several input files have been supplied, then a sitecon profile is built for each input file, i.e. several output files (with different indexes) are generated. [String, Required]

*wsize* — window size. The window is a region of the alignment used to build the profile. It is picked up from the center of the alignment and occupies the specified length. The edges of the alignment beyond the window are not taken into account. The recommended length is a bit less than the alignment length, but not more than 50 bp. [Number, Optional, Default: 40]

clength — length of a random synthetic sequence used to calibrate the profile. [Number, Optional, Default: 1000000]

*rseed* — random seed used to calibrate the profile, e.g. to generate the random synthetic sequence. Use the same value to get the same calibration results twice on the same data. By default, new random seed is generated each time a calibration occurs. [Number, Optional, Default: 0]

walg — specifies to use the Algorithm 2 weight algorithm. In most cases it is not required, but in some cases it can increase the recognition quality. [Boolean, Optional, Default: false]

#### Example:

ugene sitecon-build --in=COI.aln --out=result.sitecon

## Searching for TFBS with SITECON

#### Task Name: sitecon-search

Searches for transcription factor binding sites (TFBS) with SITECON and saves the regions found as annotations.

#### Parameters:

in - semicolon-separated list of input sequence files to search TFBS in. [String, Required]

*inmodel* — input SITECON profile(s). If several profiles have been supplied, searches with all profiles one by one and outputs merged set of annotations for each input sequence. [String, Required]

out - output Genbank file. [String, Required]

annotation-name - name of the annotated regions. [String, Optional, Default: "misc\_feature"]

*min-score* — recognition quality threshold. The value must be between 60 and 100. Choosing too low threshold will lead to recognition of too many TFBS recognised with too low trustworthiness. Choosing too high threshold may result in no TFBS recognised. [Number, Optional, Default: 85]

min-err1 — setting for filtering results, minimal value of Error type I. [Number, Optional, Default: 0]

max-err2 — setting for filtering results, maximum value of Error type II. [Number, Optional, Default: 0.001]

strand - strands to search in. [Number, Optional, Default: 0]

The following values are available:

- 0 (both strands)
- 1 (direct strand)
- -1 (complement strand)

#### Example:

ugene sitecon-search --in=input.fa --inmodel=profile.sitecon --out=res.gb

## Fetching Sequence from Remote Database

#### Task Name: fetch-sequence

Fetches a sequence from a remote database. The supported databases are accesed via alias.

Database	Alias
NCBI Genbank (DNA)	genbank
NCBI Genbank (protein)	genbank-protein
Protein Data Bank	pdb
SwissProt	swissprot
Uniprot	uniprot

#### Parameters:

db — database alias to read from. [String, Required]

id - semicolon-separated list of resource IDs in the database. [String, Required]

save-dir --- directory to store sequence files loaded from the database. [String, Optional]

#### Example:

ugene fetch-sequence --db=PDB --id=3INS;1CRN

## **Gene-by-Gene Report**

#### Task Name: gene-by-gene

Suppose you have genomes and you want to characterize them. One of the ways to do that is to build a table of what genes are in each genome and what are not there.

- 1. Create a local BLAST db of your genome sequence/contigs. One db per one genome.
- 2. Create a file with sequences of genes you what to explore. This file will be the input file for the scheme
- 3. Setup location and name of BLAST db you created for the first genome.
- 4. Setup output files: report location and output file with annotated (with BLAST) sequence. You might want to delete the "Write Sequence" element if you do not need output sequences.
- 5. Run the scheme
- 5\*. Run the scheme on the same input and output files changing BLAST db for each genome that you have.

As the result you will get the report file. With "Yes" and "No" field. "Yes" answer means that the gene is in the genome. "No" answer MIGHT mean that there is no gene in the genome. It is a good idea to analyze al

I the "No" sequences using annotated files. Just open a file and find a sequence with a name of a gene that has "No" result.

#### Parameters:

in - Input sequence file [Url datasets]

final-name - Annotation name used to compare genes and reference genomes (using 'blast\_result' by dafault) [String]

exist-file - If a target report already exists you should specify how to handle that. Merge two table in one. Overwrite or Rename existing file (using 'Merge' by default) [String]

*ident* - Identity between gene sequence length and annotation length in per cent. BLAST identity (if specified) is checked after (using '90.0' percents by default) [Number]

out - Output report file [String]
blast-out - Location of BLAST output file [String]
search-type - Type of BLAST searches (using 'blastn' by default) [String]
db-name - Name of BLAST DB [String]
blast-path - Path to BLAST DB [String]
expected-value - This setting specifies the statistical significance threshold for reporting matches against database sequences (using '10.0' by default) [Number]
gapped-aln - Perform gapped alignment (using 'use' by default) [Boolean]
blast-name - Name for annotations (using 'blast\_result' by default) [String]
tmpdir - Directory for temporary files (using UGENE temporary directory by default) [String]
toolpath - External tool path (using the path specified in UGENE by default) [String]
out-type - Type of BLAST output file (using 'XML (-m 7)' by default) [String]

#### Example:

ugene gene-by-gene --in=human\_T1.fa --out=human\_T1\_report

## **Reverse-Complement Converting Sequences**

#### Task Name: revcompl

Convert input sequence into its reverse, complement or reverse-complement counterpart and write result sequence to file

#### Parameters:

type - Type of operation. Available are 'Reverse Complement', 'Complement' and 'Reverse' (using 'Reverse Complement' by default) [String]

in - Input file [Url datasets]

accumulate - Accumulate all incoming data in one file or create separate files for each input. In the latter case, an incremental numerical suffix is added to the file name (using 'True' by default) [Boolean]

format - Output file format (using 'fasta' by default) [String]

split - Split each incoming sequence on several parts (using '1' by default) [Number]

out - Output file [String]

#### Example:

```
ugene revcompl --in=human_T1.fa --out=human_T1_result.fa --format=fasta --type=reverse
```

### Variants Calling

#### Task Name: snp

Call variants for an input assembly and a reference sequence using SAMtools mpileup and bcftool

#### Parameters:

bam - Input sorted BAM file(s) [Url datasets]

```
ref - Input reference sequence [Url datasets]
```

wout - Out file with variations [String]

bN - A/C/G/T only [Boolean]

- bl List of sites [String]
- ml BED or position list file [String]
- bg Per-sample genotypes [Boolean]
- mC Mapping quality downgrading coefficient [Number]
- bT Pair/trio calling [String]
- mB Disable BAQ computation [Boolean]
- me Gap extension error [Number]
- mE Extended BAQ computation [Boolean]
- bF Indicate PL [Boolean]
- vw Gap size [Number]
- m6 Illumina-1.3+ encoding [Boolean]
- bi INDEL-to-SNP Ratio [Number]
- bA Retain all possible alternate [Boolean]
- vD Max number of reads per input BAM [Number]
- md Max number of reads per input BAM [Number]
- mL Max INDEL depth [Number]
- va Alternate bases [Number]
- v2 BaseQ bias [String]
- vd Minimum read depth [Number]
- v4 End distance bias [Number]
- v3 MapQ bias [Number]
- Q Minimum RMS quality [Number]
- v1 Strand bias [Number]
- *mQ* Minimum base quality [Number]
- mq Minimum mapping quality [Number]
- bd Min samples fraction [Number]
- b1 N group-1 samples [Number]
- bU N permutations [Number]
- bG No genotype information [Boolean]
- ml No INDELs [Boolean]
- mo Gap open error [Number]
- mP List of platforms for indels [String]
- vp Log filtered [Boolean]
- bP Prior allele frequency spectrum. [String]
- bQ QCALL likelihood [Boolean]
- mr Pileup region [String]
- bs List of samples [String]
- mh Homopolymer errors coefficient [Number]
- bt Mutation rate [Number]

mA - Count anomalous read pairs [Boolean]

vW - A/C/G/T only [Number]

#### Example:

```
ugene snp --bam=test.bam --ref=test_ref.fa --wout=test_out.vcf
```

## **Generating DNA Sequence**

Task Name: generate-dna

Generates a random DNA sequence with specified nucleotide content

#### Parameters:

algo - Algorithm for generating (using 'GC Content' by default) [String]

content - Specifies if the nucleotide content of generated sequence(s) will be taken from reference or specified manually (A, G, C, T parameters) (using 'manual' by default) [String]

count - Number of sequences to generate (using '1' by default) [Number]

length - Length of the resulted sequence(s) (using '1000' bp by default) [Number]

a - Adenine content (using '25' percents by default) [Number]

c - Cytosine content (using '25' percents by default) [Number]

g - Guanine content (using '25' percents by default) [Number]

t - Thymine content (using '25' percents by default) [Number]

ref - Path to the reference file (could be a sequence or an alignment) [String]

seed - Value to initialize the random generator. By default (seed = -1) the generator is initialized with the system time (using '-1' by default) [Number]

wnd-size - Size of window where set content (using '1000' by default) [Number]

accumulate - Accumulate all incoming data in one file or create separate files for each input. In the latter case, an incremental numerical suffix is added to the file name (using 'True' by default) [Boolean]

format - Output file format (using 'fasta' by default) [String]

split - Split each incoming sequence on several parts (using '1' by default) [Number]

out - Output file [String]

#### Example:

ugene generate-dna --length=2000 --a=45 --out=test.fa

## **Creating Custom CLI Tasks**

The predefined tasks are actually the Workflow Designer schemas stored in the \$UGENE/data/cmdline directory.

Follow the instructions in the Workflow Designer Manual on how to create a schema and to run it from the command line.

You may also find useful the following video tutorial devoted to the creating of a custom console command:

• Creating custom console command (MUSCLE alignment with various output format)

# **APPENDIXES**

- Appendix A. Supported File Formats
  - Specific File Formats
  - UGENE Native File Formats
  - Other File Formats

## **Appendix A. Supported File Formats**

UGENE is able to read and write files compressed with Unix/Linux gzip utility. You don't have to unpack the files. ⚠

- Specific File FormatsUGENE Native File Formats
- Other File Formats

## **Specific File Formats**

File format	File extension	Read	Write	Comment
ABIF	*.ab1, *.abi, *.abif	+	-	A chromatogram file format. See also: Chromatogra m Viewer
ACE	*.ace	+	-	A file format for storing data about genomic contigs. See also: Alignment Editor
APR	*.apr	+	-	Vector NTI / AlignX Alignment format. See also: Alignment Editor
Bairoch	*.bairoch	+	+	A file format to store enzymes. See also: Restriction Analysis
BAM	*.bam	+	-	Binary compressed SAM format. See also: Assembly Browser
ClustalW	*.aln	+	+	A multiple sequence alignments (MSA) file format. See also: Alignment Editor
EBWT	*.ebwt	+	+	A Bowtie prebuilt index file. See also: <i>Bowtie</i>
EMBL	*.em, *.emb, *.embl	+	-	A rich format for storing sequences and their annotations. See also: Sequence View

FASTA	*.fa, *.mpfa, *.fna, *.fsa, *.fas, *.fasta, *.sef *.seqs	+	+	One of the oldest and simplest sequence file format.
				<b>See also:</b> Sequence View
FASTQ	*.fastq	+	+	A file format used to store a sequence and its corresponding quality scores. It was originally developed at the "Wellcome Trust Sanger Institute". See also: Sequence
				View
Genbank	*.gb, *.gbk, *.gen, *.genbank	+	+	A rich format for storing sequences and associated annotations. See also: Sequence View
GFF	*.gff	+	+	The Gene Finding Format (GFF) format is used to store features and annotations. See also: Sequence View
нмм	*.hmm	+	+	A file format to store HMM profiles. See also: HMM2, HMM 3
MMDB	*.prt	+	-	ASN.1 format used by the Molecular Modeling Database (MMDB). See also: 3D Structure Viewer
MSF	*.msf	+	+	A multiple sequence alignments file format. See also: Alignment Editor
Mega	*.meg, *.meg.gz	+		A multiple sequence alignments file format. See also: Alignment Editor
Newick	*.nwk, *.newick	+	+	A tree file format. See also: Building Phylogenetic Tree, Phyl ogenetic Tree Viewer

Nexus	*.nex *.nxs	+	+	A multiple alignment and phylogenetic trees file format.
				See also: Alignment Editor, Building Phylogenetic Tree, Phyl ogenetic Tree Viewer
PDB	*.pdb	+	-	The Protein Data Bank (PDB) format allows to view the 3D structure of the sequence. See also: 3D Structure Viewer
pDRAW32	*.pdw	+	-	A sequence file format used by pDRAW32 software. See also: Sequence
PFM	*.pfm	+	+	View A file format for a
				position frequency matrix.
				See also: Weight Matrix
Phylip	*.phy	+	+	A multiple alignment file format.
				See also: Alignment Editor
PWM	*.pwm	+	+	A file format for a position weight matrix.
				See also: Weight Matrix
Raw	*.seq	+	+	A raw sequence format.
				See also: Sequence View
SAM	*.sam	+	÷	The Sequence Alignment/Map (SAM) format is a generic alignment format for storing read alignments against reference sequences.
				See also: Assembly Browser, Bowtie, UGEN E Genome Aligner
SCF	*.scf	+	-	It is a Standard Chromatogram Format.
				See also: Chromatogra m Viewer
SITECON	*.sitecon	+	-	A file format to store TFBS profile.
				See also: SITECON

Stockholm	*.sto	+	+	A multiple sequence alignments file format. See also: <i>Alignment</i> <i>Editor</i>
Swiss-Prot	*.txt *.sw	+	-	An annotated protein sequence in format of t h e UniProtKB/Swiss-Prot database. See also: Sequence View
Vector NTI Sequence	*.gb *.gp	+	+	A rich format for storing sequences and associated annotations, produced by Vector NTI software. See also: Sequence View
VCF	*.vcf	+	+	The VCF specifies the format of a text file used for storing gene sequence variations. See also: Assembly Browser

## **UGENE Native File Formats**

File format	File extension	Read	Write	Comment
Dotplot	*.dpt	+	+	Stores a dotplot of a sequence. See also: Dotplot
UGENE database file	*.ugenedb	+	+	UGENE database files stores information for imported BAM or SAM files and can be used for converting this information into a SAM file. See also: Import BAM/SAM File
Short Reads FASTA	*.srfa, *.srfasta	+	+	A multiple sequence alignments file format. See also: Alignment Editor
UGENE Workflow Language	*.uwl	+	+	Human-readable format to store workflows, created in UGENE <i>Work</i> <i>flow Designer</i> . See also: <i>Workflow</i> <i>Designer</i>

UGENE Query Language	*.uql	+	+	Human-readable format to store schemas, created in UGENE <i>Quer</i> <i>y Designer</i> . See also: <i>Query</i> <i>Designer</i>
Workflow element for command line tool	*.etc	+	+	Format for storing workflow elements that can launch an external command line tool. See also: Workflow Designer

## **Other File Formats**

File format / extension	Comment
*.csv	<b>Example of usage:</b> annotations can be exported to this format; the <i>Weight Matrix</i> matrices list can also be saved to this format.
*.html	For example it is used to store reports.
image formats: *.bmp, *.jpg, *.png, *.tiff, *.svg, etc.; *.pdf	These formats are used throughout the program to save screenshots, etc.
*.txt	It is possible to view and modify plain text files in UGENE.

# Tutorials

- Using BioMart with UGENE
  - Environment requirements
  - Installing UGENE extension on Mozilla Firefox
  - Opening data found using BioMart in UGENE
  - Opening BioMart data in UGENE by ID
  - Opening selected data in UGENE

## **Using BioMart with UGENE**

The BioMart system enables scientists to perform advanced querying of a wide range of biological data sources through a single web interface, regardless of the data sources geographical locations.

This tutorial describes how data found through the *BioMart* web interface can be easily opened for further analysis in UGENE by a couple of mouse-clicks.

- Environment requirements
- Installing UGENE extension on Mozilla Firefox
- Opening data found using BioMart in UGENE
- Opening BioMart data in UGENE by ID
- Opening selected data in UGENE

## **Environment requirements**

Currently UGENE extension is available for Mozilla Firefox web browser only. Please make sure to launch UGENE before using the extension!

Follow the instructions below to install the extension.

## Installing UGENE extension on Mozilla Firefox

To install UGENE extension on Mozilla Firefox open Add-ons Manager and select Install Add-on From File item in the settings menu:

Firefox <b>*</b>	
add-ons Manager	+

In the browse dialog select *ugene.xpi* file that you can find in the *Firefox* directory of the UGENE Web Browsers Extensions Package that there is on the Download page.

## Opening data found using BioMart in UGENE

For now there are two options to open data found using BioMart in UGENE:

- Open data by ID, for example, by an Ensembl ID.
- Open selected data.

## **Opening BioMart data in UGENE by ID**

Let's open web site:



Click, for example, on the Proceed to Bio Portal link. The following page will appear:

till BioMart × till BioMart Portal ×	
← → C [] central.biomart.org	
BioMart Central Portal	
IDENTIFIER SEARCH	BIOMART CENTRAL PORTAL Databases: 41
Tools       Gene retrieval     Variant retrieval     Sequence retrieval     ID converter       Cancer genes     Ensembl     Ensembl     Ensembl       Ensembl Bacteria     Ensembl Pungi     Ensembl Metazoa       Ensembl Plants     Ensembl Protists     Mouse Genome Informatics       VEGA     VEGA	CANADA UNITED SINCEOM 2 SFAN 1 CHIA 1 CHIA 1 CHIA 1 CHIA 1 CHIA 1 CHIA 1 CHIA 1 CHIA 1 CHIA 1 CHIA 1 CHIA 1 CHIA 1 CHIA 1 CHIA 1 CHIA 1 CHIA 1 CHIA
DATABASE SEARCH	Click on the map to view the list of databases
Search by type Search by organism Search by database name (A-Z)	
► Genome	
Gene annotation	
Protein sequence and structure	
Interaction and pathways	
Gene expression	

Notice that an example Ensembl ID below the search bar is highlighted (it has a light blue background).

Current version of the UGENE extension allows detecting the following types of identification numbers:

- 1. Ensemble Gene ID
- 2. Ensembl Protein ID
- 3. PDB ID

Right-click on the ID and select Open in UGENE item in the context menu:

# **BioMart Central Portal**

Home

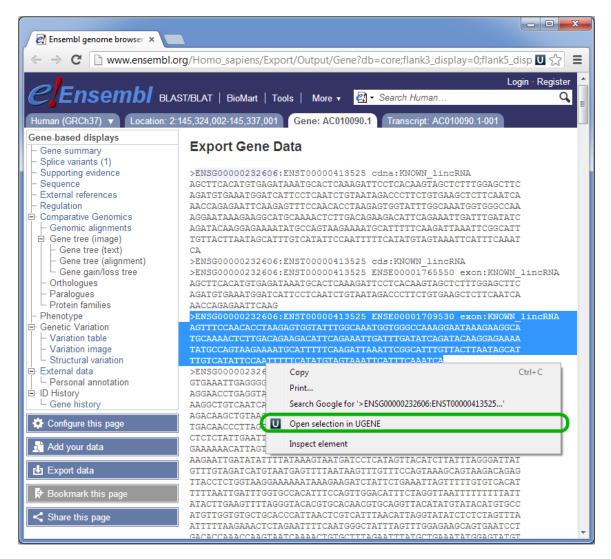


The sequence with the selected ID will be opened in UGENE.

## **Opening selected data in UGENE**

Imagine that you have browsed for required data (e.g. a sequence with annotations) and opened, for example, an html view for the data in a web browser. Now you would like to open the data in UGENE to analyze them in more detail. Or, alternatively, maybe you would like to analyze a certain sequence part.

In this case you select the required data in the web browser window. the Open selected in UGENE item should now appear in the context menu:



The selected data will be opened in UGENE.