

Unipro



UGENE

Решение практических задач с помощью UGENE

**Пособие для школы-семинара молодых ученых
«Вычислительные задачи молекулярной биологии и платформа UGENE»**

30 мая – 2 июня 2011



Содержание

Содержание	2
Введение	3
Часть I. Знакомство с UGENE	4
1. Общие сведения о UGENE	4
2. Практическая задача: Исследование неизвестного вируса	5
3. Практическая задача: Работа с данными секвенирования	11
4. Практическая задача: Поиск гена в последовательности	15
5. Практическая задача: Построение вычислительных схем	18
Часть II. Запуск задач на кластере HГУ	26
1. Где в UGENE прописать адрес кластера	26
2. Как запустить схему на кластере	28
3. Примеры схем	28
Часть III. Работа с модулем Expert Discovery в UGENE	31
1. Общие сведения о модуле Expert Discovery	31
2. Практическая задача: Поиск комплексных сигналов на выровненной выборке	32
Заключение	41

Введение

Данное пособие содержит вспомогательные материалы для решения практических задач, представленных на школе-семинаре «Вычислительные задачи молекулярной биологии и платформа UGENE».

Для запуска задач используется текущая сборка UGENE (r.643).

В первую часть пособия включено решение следующих практических задач:

1. Исследование неизвестного вируса:

В данной задаче исследуется последовательность неизвестного вируса. Рассматривается поиск гомологов с помощью удаленного запроса BLAST, загрузка последовательностей с NCBI, множественное выравнивание последовательностей, построение филогенетических деревьев.

2. Работа с данными секвенирования:

Делается краткий обзор работы с данными секвенирования (в формате BAM) с помощью UGENE Assembly Browser: просмотр данных, экспорт ридов в FASTA файл.

3. Поиск гена в последовательности:

Приводится пример поиска составного сигнала с помощью схемы UGENE Query Designer.

4. Построение вычислительных схем:

Рассматривается 2 примера построения вычислительных схем. Также рассматривается запуск схемы из командной строки и использование скриптов для задания значения параметра.

Во второй части пособия описывается необходимая информация о запуске схем на кластере НГУ. Приводятся примеры схем.

Третья часть пособия содержит описание и пример использования системы Expert Discovery, встроенной в UGENE, позволяющей размечать протяженные районы генов, отвечающие за регуляцию транскрипции.

Часть I. Знакомство с UGENE

1. Общие сведения о UGENE

Что такое UGENE

UGENE – свободное кроссплатформенное бионформационное программное обеспечение.

В UGENE интегрированы десятки известных биоинформационных инструментов и алгоритмов, доступных как через графический интерфейс, так и через командную строку.

Используя встроенный дизайнер вычислительных схем, различные инструменты и алгоритмы могут быть скомпонованы в вычислительную схему.

Чтобы узнать больше:

- <http://ru.wikipedia.org/wiki/UGENE>
- <http://ugene.unipro.ru/>

Где можно взять UGENE

Последнюю версию UGENE всегда можно свободно скачать со следующей страницы:

- <http://ugene.unipro.ru/rus/download.html>

На данной странице можно скачать пакеты для операционных систем Windows, Linux, Mac OS X, и др. Также доступен исходный код продукта (распространяется на условиях [GPLv2](#)).

Можно также скачать одну из последних “предрелизных” сборок UGENE:

- <http://ugene.unipro.ru/rus/snapshot.html>

Документация (на английском языке) доступна на следующей странице:

- <http://ugene.unipro.ru/documentation.html>

2. Практическая задача: Исследование неизвестного вируса

Что есть

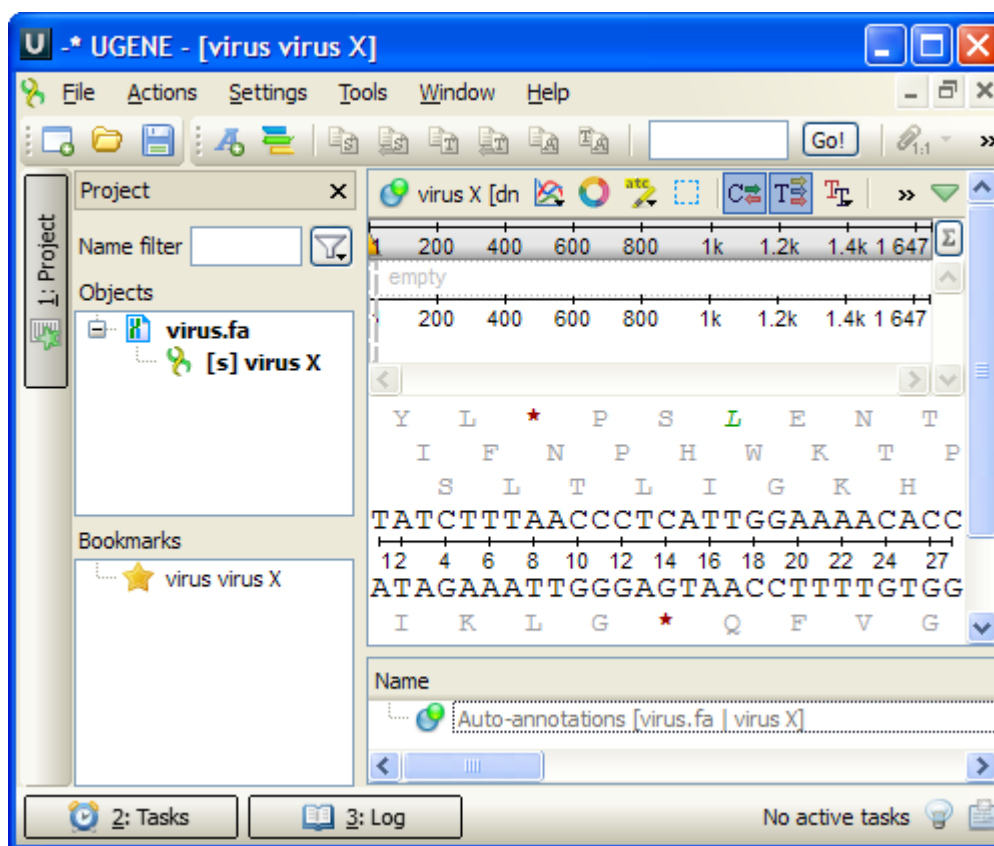
Имеется ДНК последовательность некоторого неизвестного вируса в формате FASTA: "virus.fa".

Что требуется

Найти гомологи для данной последовательности, выровнять полученные последовательности и построить филогенетическое дерево.

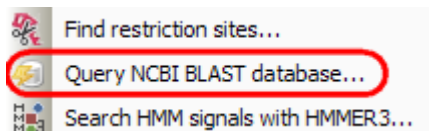
Как это сделать

1. Открыть "virus.fa" в UGENE:

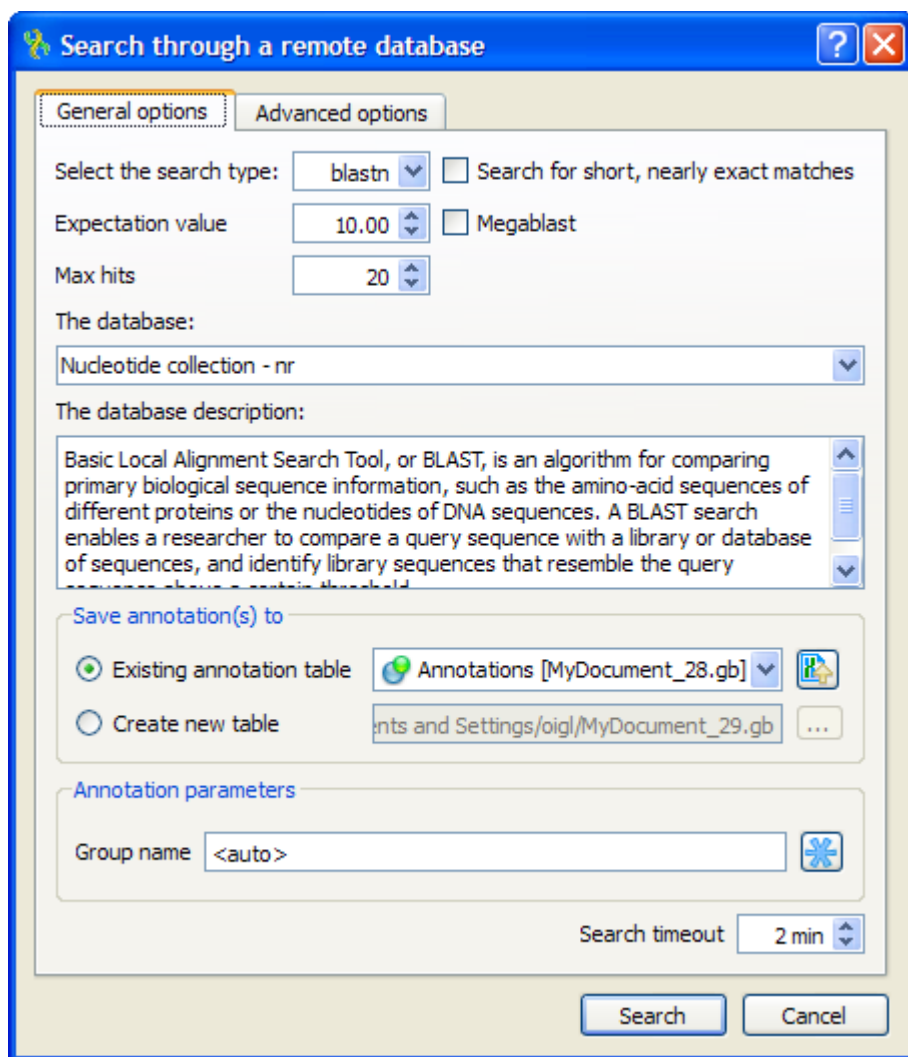


2. Найти гомологи с помощью удаленного запроса [BLAST](#):

- Нажмите правую кнопку мыши и выберите "Analyze > Query NCBI BLAST database" в появившемся контекстном меню.

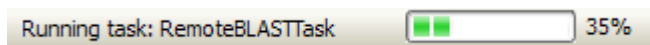


- Для того, чтобы начать поиск достаточно нажать “Search” в появившемся диалоге. При необходимости можно также задать параметры поиска, отличные от значений по умолчанию.

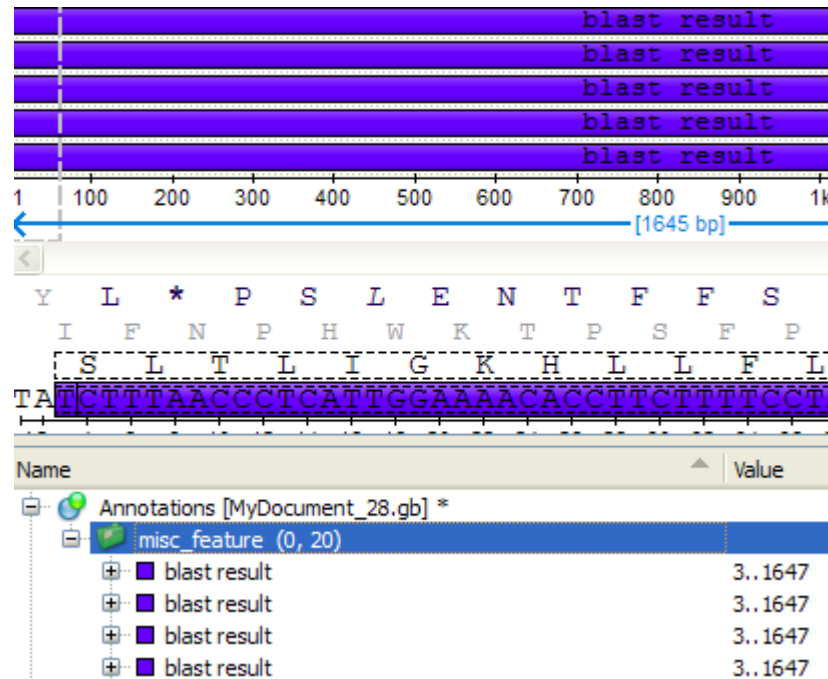


Внимание: Поиск производится в удаленной базе данных, поэтому для успешного выполнения данного пункта требуется доступность интернета. В случае отсутствия интернета, можете воспользоваться локальным поиском BLAST, описание которого не входит в данное пособие.

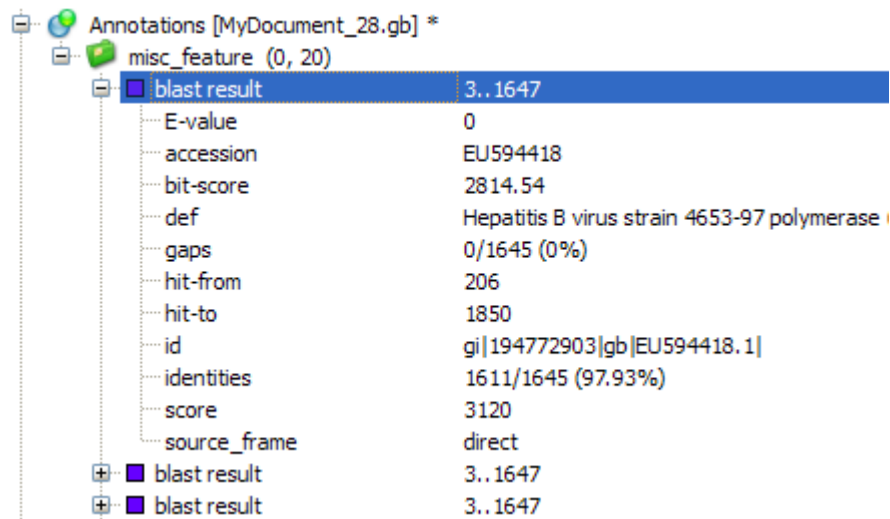
- После того, как поиск был начат, за ходом его выполнения можно следить, например, в нижней части окна UGENE:



- По завершении поиска последовательность вируса будет проаннотирована:

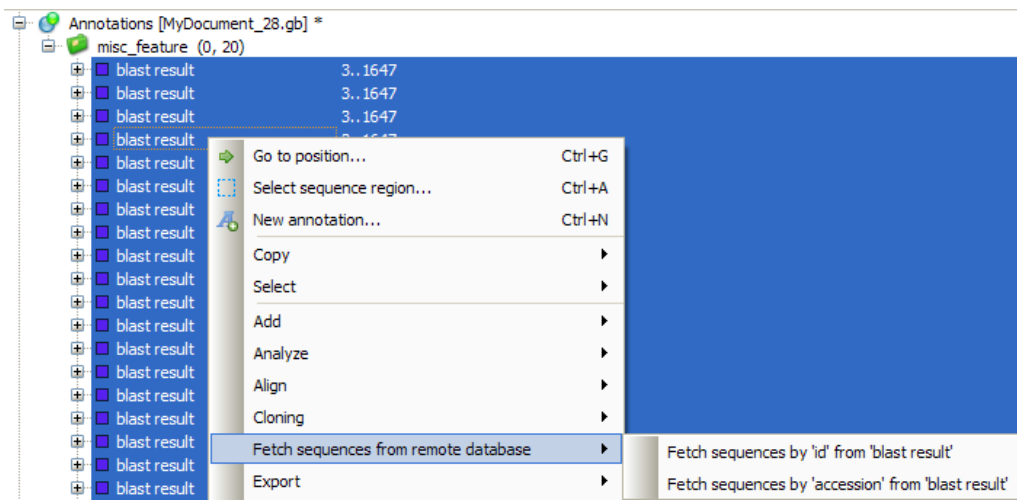


- Чтобы посмотреть подробную информацию о каком-нибудь результате поиска, раскройте соответствующий узел дерева аннотаций:

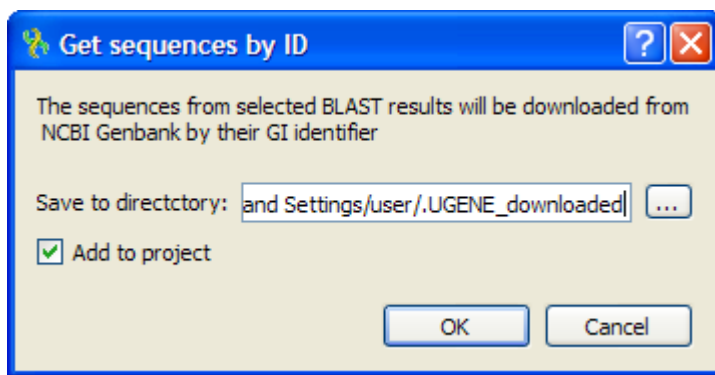


3. Загрузить последовательности гомологов из [NCBI GenBank](https://www.ncbi.nlm.nih.gov/genbank/):

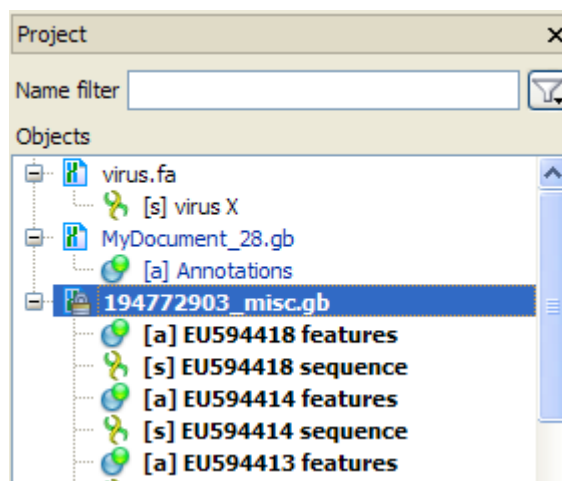
- Для этого выделите аннотации в нижней части окна (называемой "Annotations Editor") как показано на рисунке ниже и выберите "Fetch sequences from remote database > Fetch sequences by 'id' from 'blast result'" в контекстном меню:



- В появившемся диалоге можно выбрать папку, куда загрузить файлы. Оставьте включенной опцию “Add to project” и нажмите “OK”:



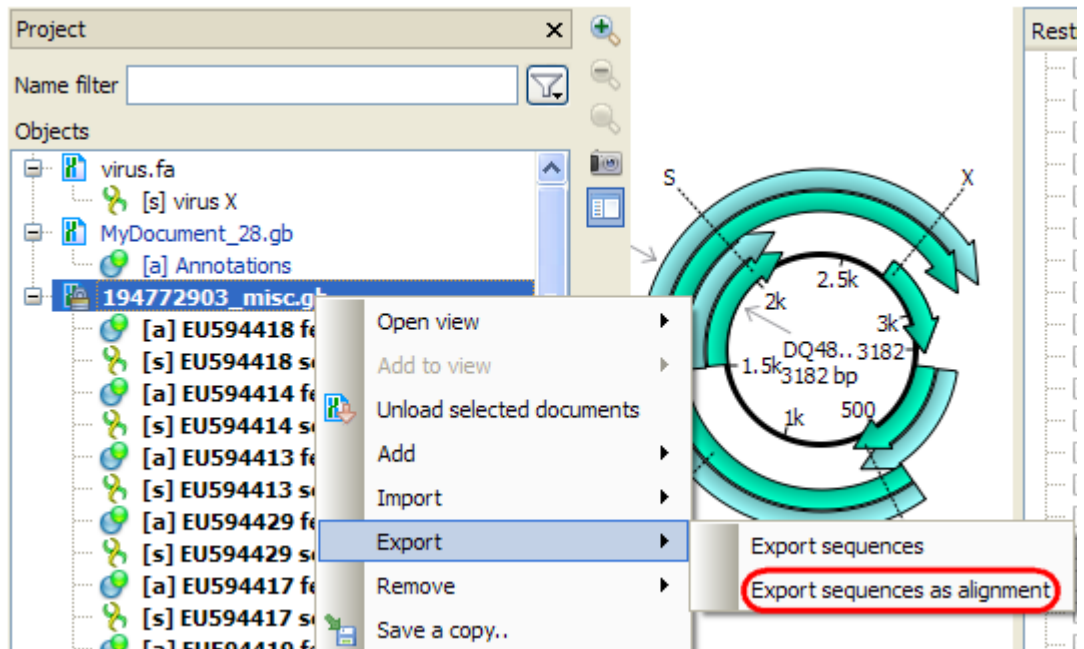
- После того, как последовательности загрузятся, в проекте появится новый GenBank файл с этими последовательностями:



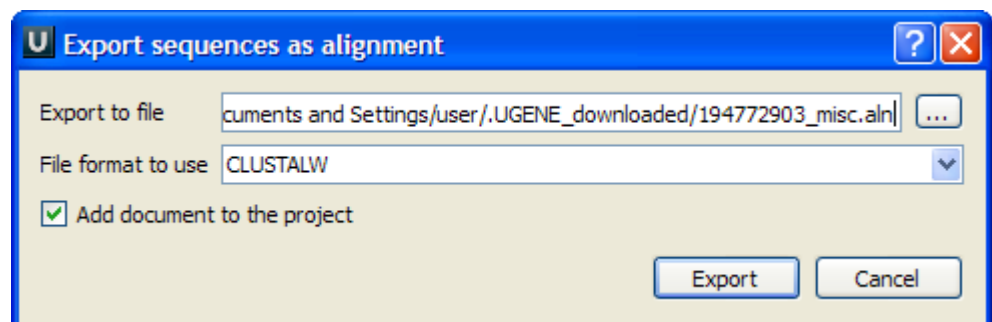
Внимание: Для успешного выполнения данного пункта требуется доступность интернета.

4. Экспортировать последовательности в формат множественного выравнивания:

- Нажмите правой кнопкой мыши на файл с последовательностями и выберите “Export > Export sequences as alignment” в контекстном меню:



- В появившемся диалоге можно оставить значения параметров (имя файла множественного выравнивания и его формат) по умолчанию и нажать “Export”:



5. Выровнять последовательности:

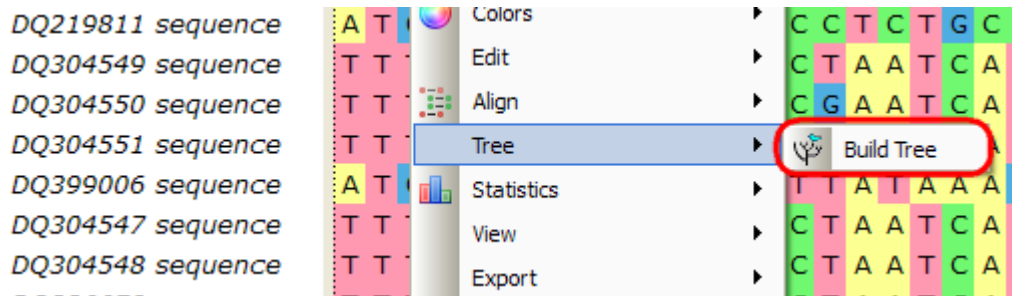
- В контекстном меню множественного выравнивания выберите “Align > Align with MUSCLE”:



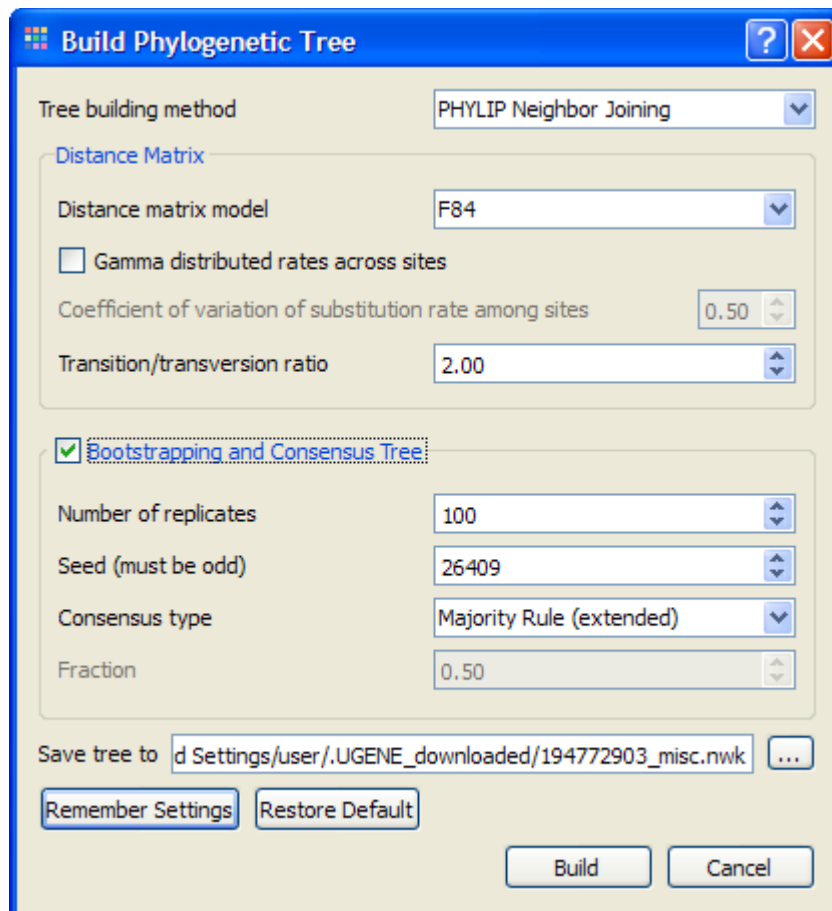
- В появившемся “Align with MUSCLE” диалоге нажмите на кнопку “Align”.

6. Построить филогенетическое дерево:

- Выберите “Tree > Build Tree” в контекстном меню множественного выравнивания:



- В диалоге “Build Phylogenetic Tree” нажмите “Build”:



Для построения дерева используется метод “Neighbor Joining” с различными моделями для подсчета матрицы расстояний, реализованный в пакете [PHYLP](#). При построении дерева может быть применен бутстреп-анализ.

3. Практическая задача: Работа с данными секвенирования

Что есть

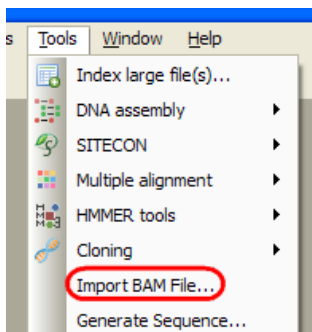
Имеются данные секвенирования в формате BAM (“example-alignment.sorted.bam”), файл индекса (“example-alignment.sorted.bam.bai”) и референтная последовательность (“example-sequence.fasta”).

Что требуется

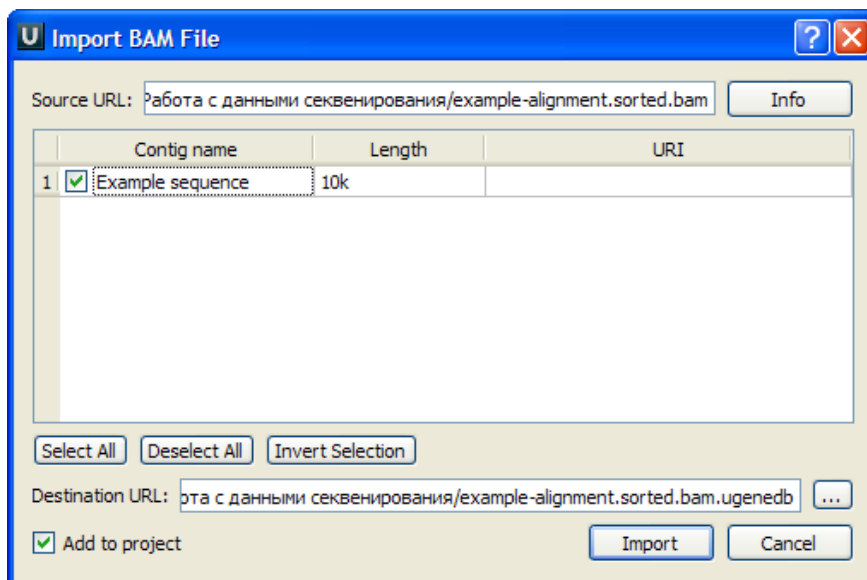
Отобразить имеющиеся данные, экспортировать часть данных в формат FASTA.

Как это сделать

1. Открыть UGENE.
2. Импортировать BAM файл:
 - В главном меню выберите “Tools > Import BAM File”:



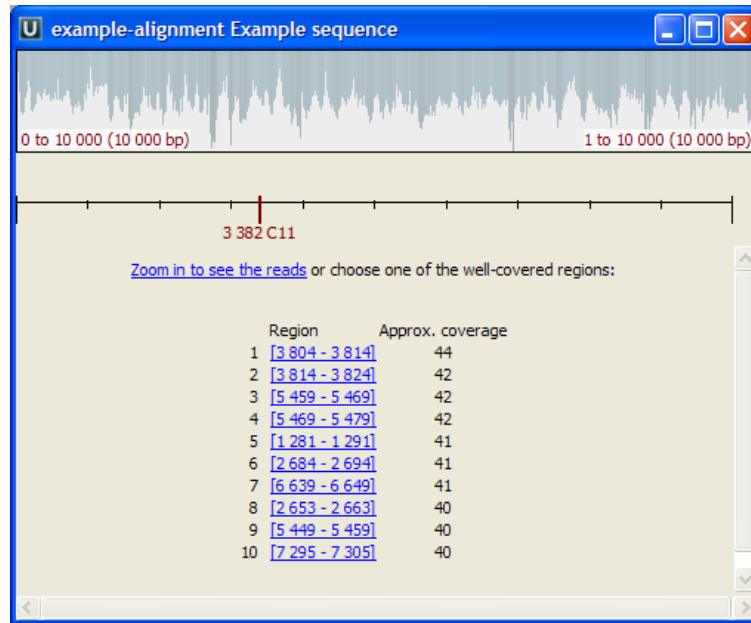
- В появившемся диалоге выберите файл “example-alignment.sorted.bam” и нажмите “Open”.
- В диалоге “Import BAM File” нажмите “Import”:



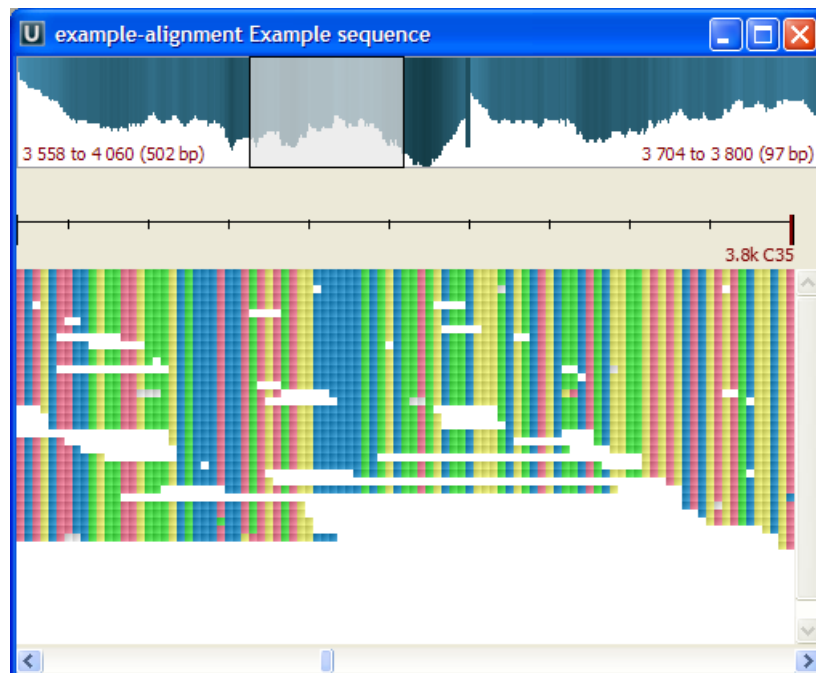
Замечание: В данном примере импорт занимает мало времени. В реальной ситуации может потребоваться некоторое время для импорта данных.

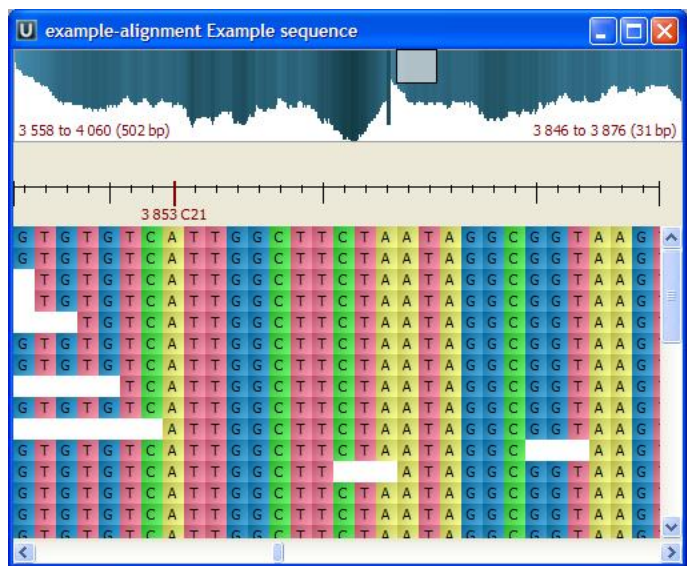
3. Отобразить данные:

- В следующем окне нажмите на первом регионе – этот регион имеет максимальное покрытие короткими последовательностями (“ридами”) в открытом контиге BAM файла.

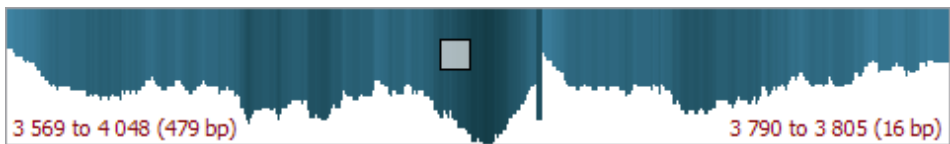


- В открывшемся окне покрутите колесо мыши, чтобы увеличить данные:

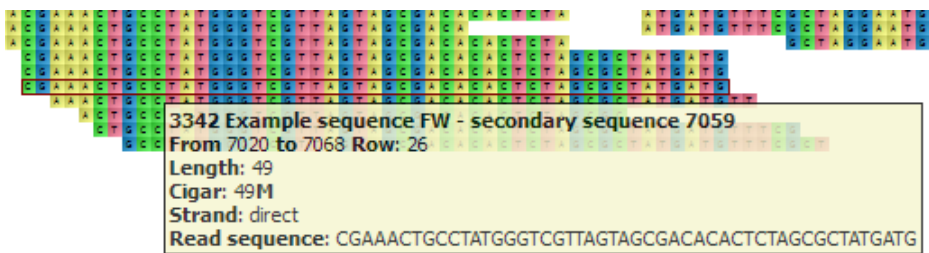




Замечание: Сверху в окне отображается “Assembly Overview”, оно показывает покрытие рядами всего контига или его части (то есть, “Assembly Overview” также можно приближать/удалять). Более подробно см. документацию по Assembly Browser.

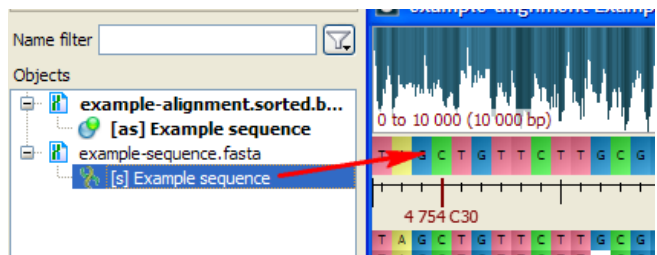


- Чтобы узнать информацию о каком-нибудь ряде подведите курсор мыши к нему:



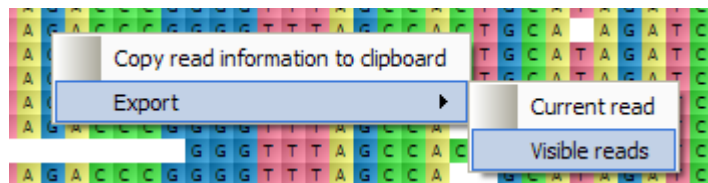
4. Добавить референтную последовательность:

- Откройте “ example-sequence.fasta” в UGENE.
- Перетащите последовательность в “Reference Area”:



5. Экспортировать данные в FASTA:

- Выберите “Export > Visible reads” в контекстном меню:



- В появившемся “Export Reads” диалоге введите имя файла в поле “Export to file” и нажмите кнопку “Export”. Файл, содержащий видимые риды, добавится к проекту.

4. Практическая задача: Поиск гена в последовательности

Что есть

ДНК последовательность Escherichia_coli (“NC_000913.gb”).

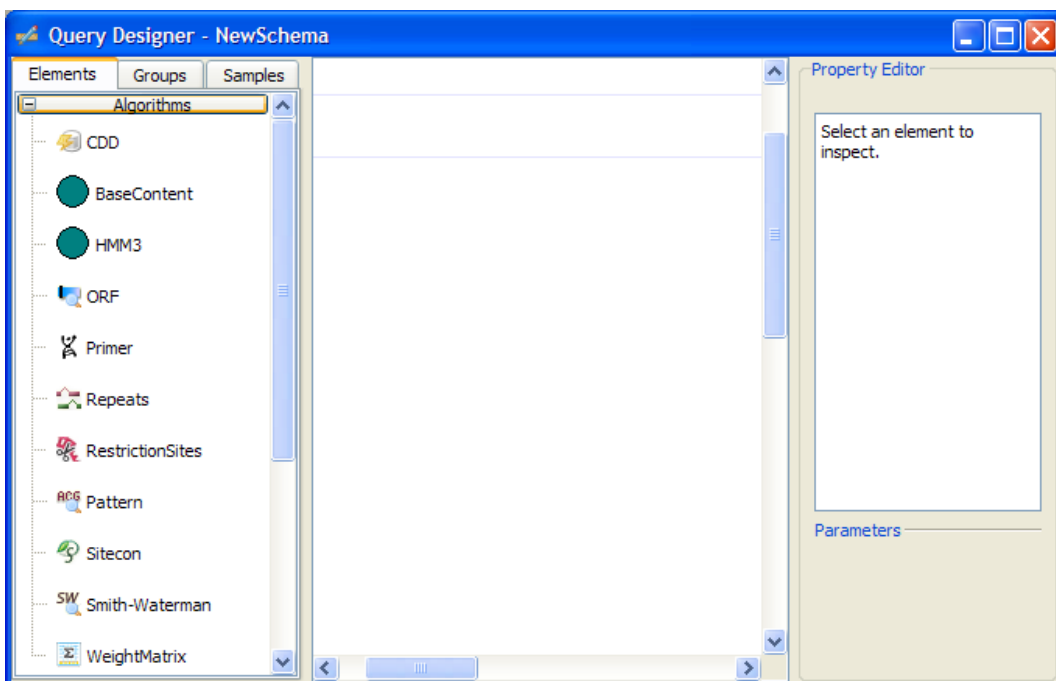
Что требуется

Найти в последовательности и аннотировать места, где потенциально может находиться ген.

Как это сделать

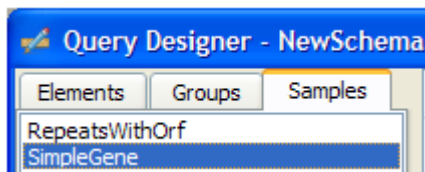
1. Запустить Query Designer в UGENE:

Чтобы открыть окно Query Designer выберите “Tools > Query Designer” в главном окне UGENE.



2. Открыть схему поиска гена:

- Выберите вкладку “Samples”:

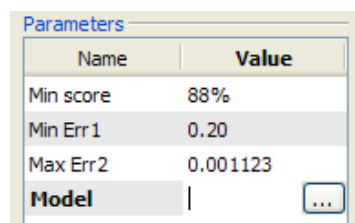


- Дважды щелкните мышью на “SimpleGene”, чтобы открыть схему.

3. Задать параметры схемы:

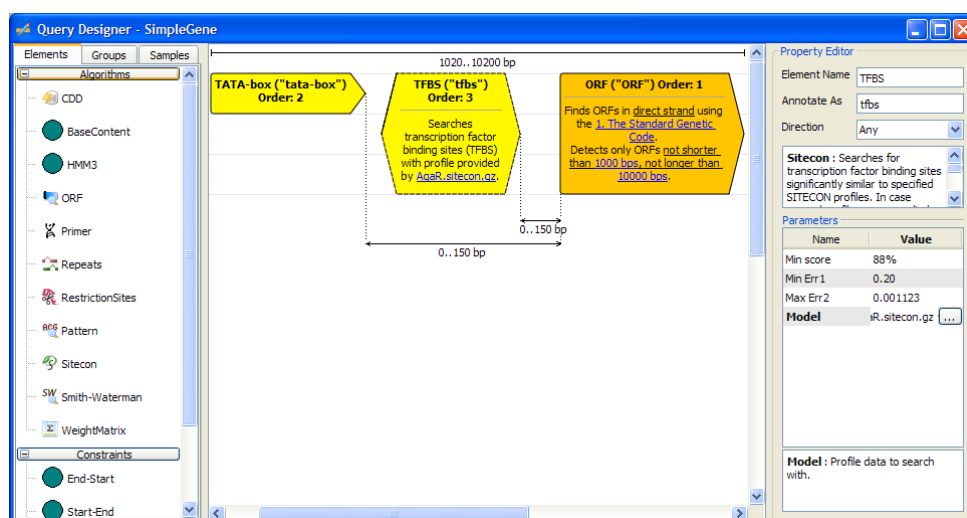
В данной схеме достаточно задать модель для поиска сайтов связывания транскрипционных факторов.

- Выберите элемент схемы “TFBS”.
- Справа в параметрах нажмите на значение параметра “Model”:



Name	Value
Min score	88%
Min Err 1	0.20
Max Err 2	0.001123
Model	...

- Нажмите на кнопку “...” , выберите модель “AgaR.sitecon.gz” в открывшемся диалоге и нажмите “Open”.



Query Designer - SimpleGene

Elements Groups Samples

Algorithms

- CDD
- BaseContent
- HMM3
- ORF
- Primer
- Repeats
- RestrictionSites
- Pattern
- Sitecon
- Smith-Waterman
- WeightMatrix

Constraints

- End-Start
- Start-End

1020..10200 bp

TATA-box ("tata-box") Order: 2

TFBS ("tfbs") Order: 3

ORF ("ORF") Order: 1

Searches transcription factor binding sites (TFBS) with profile provided by [AgaR.sitecon.gz](#).

Finds ORFs in direct strand using the [1. The Standard Genetic Code](#). Detects only ORFs not shorter than 1000 bps, not longer than 10000 bps.

0..150 bp

0..150 bp

Property Editor

Element Name: TFBS

Annotate As: tfbs

Direction: Any


Sitecon: Searches for transcription factor binding sites significantly similar to specified SITECON profiles. In case...

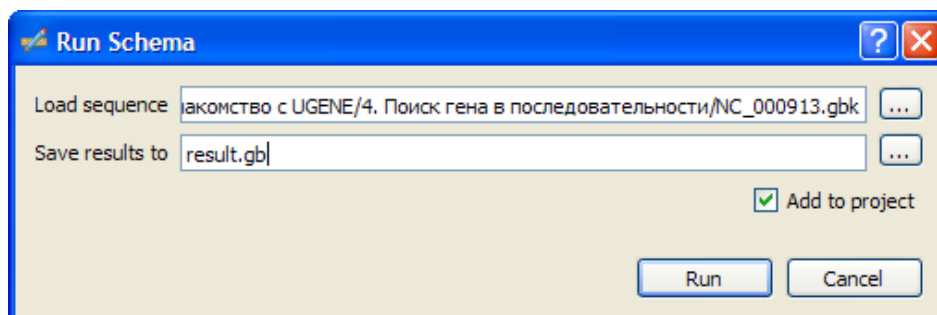
Parameters

Name	Value
Min score	88%
Min Err 1	0.20
Max Err 2	0.001123
Model	r.sitecon.gz ...

Model: Profile data to search with.

4. Запустить схему на выполнение:

- Нажмите кнопку  на панели задач.
- В “Run Schema” диалоге загрузите последовательность “NC_000913.gbк”, укажите имя для файла с результатами:



Run Schema

Load sequence: ...

Save results to: ...

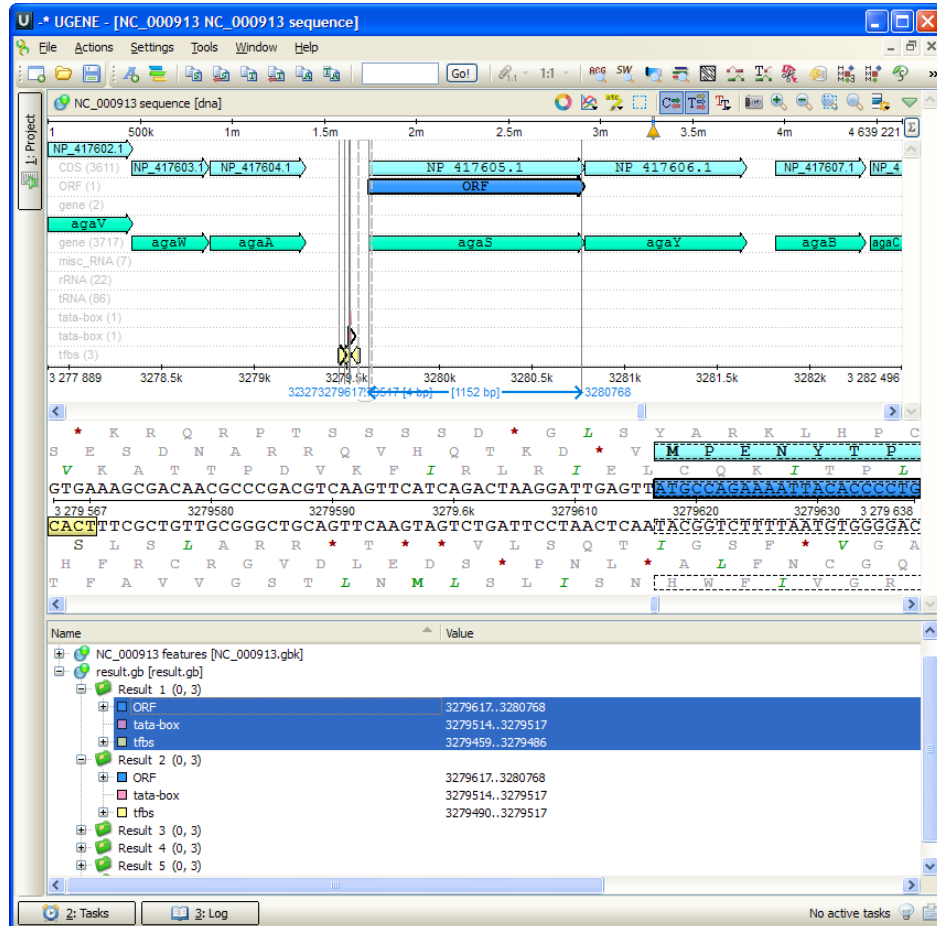
Add to project

Run Cancel

- Нажмите “Run”.

5. Посмотреть результаты:

После того, как схема выполняется откроется новое окно Sequence View, содержащее данную последовательность:



Найденные результаты сохранялись как аннотации в файле result.gb:



5. Практическая задача: Построение вычислительных схем

Что есть

1. Приходится часто вручную выполнять множественное выравнивание набора последовательностей.
2. Стоит задача разделить multi-FASTA файл со множеством последовательностей на отдельные FASTA файлы.

Что требуется

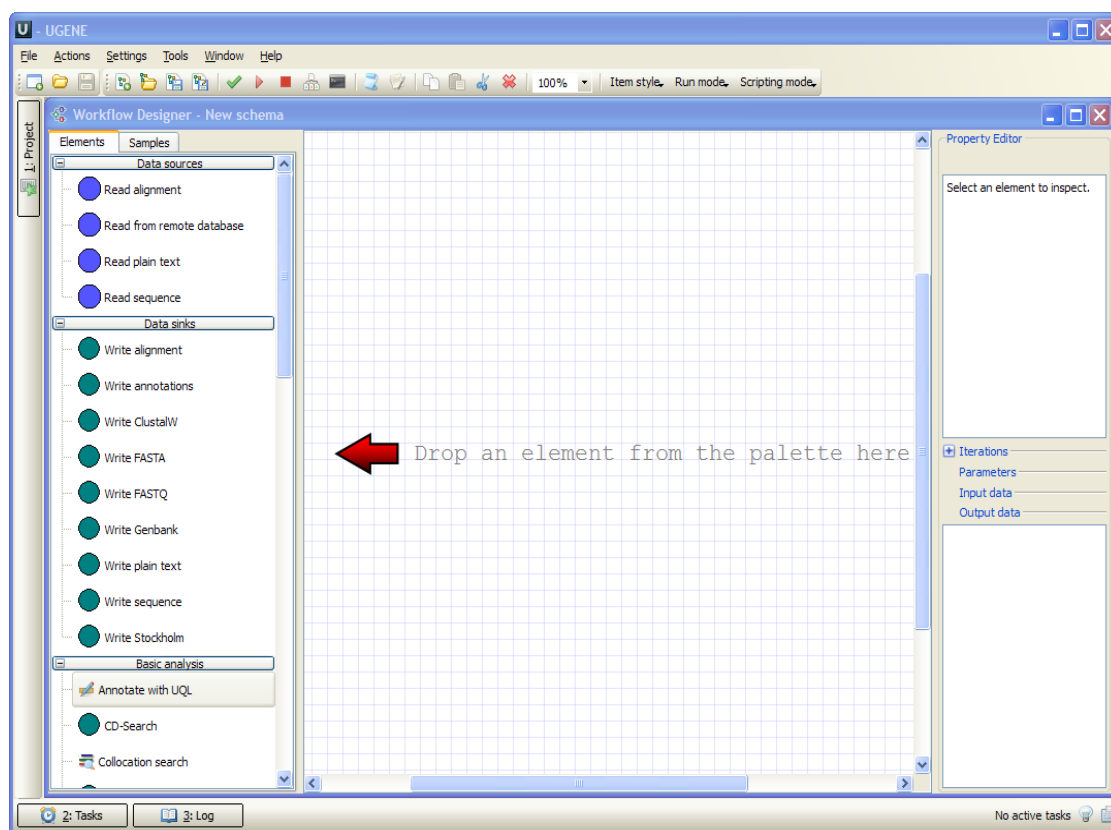
1. Автоматизировать первую задачу – запускать ее из командной строки, задавая имена файлов как параметры.
2. Решить вторую задачу. Обеспечить возможность ее автоматизации.

Как сделать 5.1

Чтобы автоматизировать множественное выравнивание последовательностей надо:

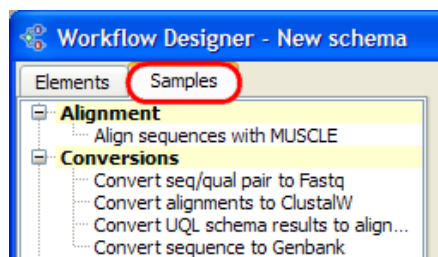
1. **Запустить Workflow Designer в UGENE:**

Чтобы открыть окно Workflow Designer выберите “Tools > Workflow Designer” в главном окне UGENE. Откроется следующее окно:

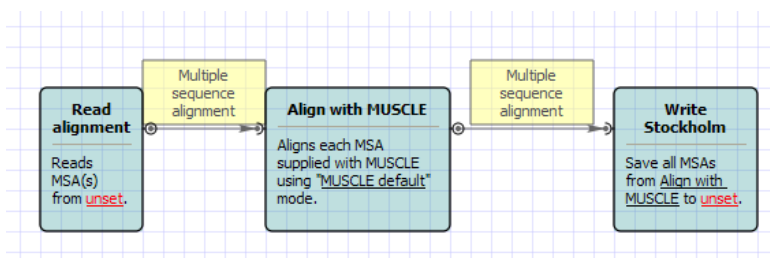


2. Открыть схему-пример “Align sequences with MUSCLE” :

- Выберите вкладку “Samples”:



- Дважды щелкните мышью по “Align sequences with MUSCLE” (см. выше). Откроется следующая схема:

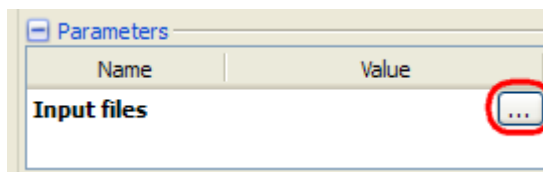


Замечание: В данном примере рассматривается выравнивание с помощью MUSCLE, однако доступны и другие элементы схемы для множественного выравнивания последовательностей – выравнивание с помощью ClustalW, Kalign и др.


3. Запустить схему из графического интерфейса: (этот пункт можно пропустить)

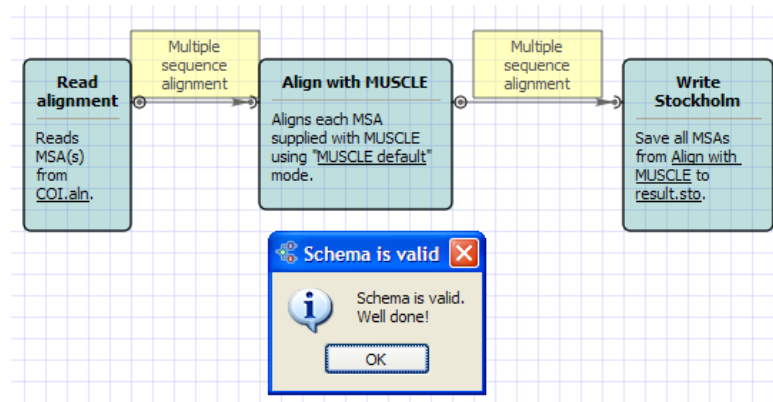
Для проверки, что схема работает попробуем запустить ее из графического интерфейса.


- Выберите элемент “Read alignment” и в параметрах (справа) нажмите на поле значения параметра “Input files”, нажмите на появившуюся кнопку “...” :

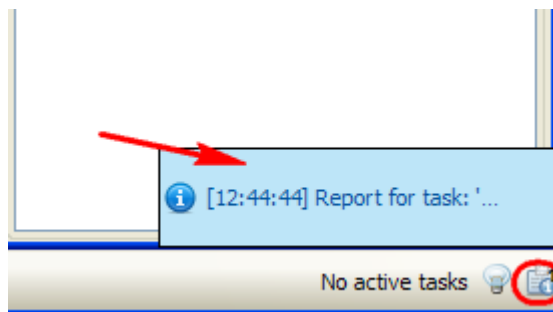


- В появившемся диалоге выберите один или несколько файлов в формате множественного выравнивания, например можно использовать файл “\$UGENE\data\samples\CLUSTALW\COI.aln”. Здесь “\$UGENE” – директория куда был установлен UGENE, например на Windows это соответствует “C:\Program Files\Unipro UGENE\data\samples\CLUSTALW\COI.aln”.

- Точно так же выберите элемент “Write Stockholm” и укажите какое-нибудь имя файла результата (параметр “Output file”).
- Нажмите кнопку  на панели задач – схема корректна, ошибок при ее валидации не возникло:



- Нажмите “OK” в “Schema is valid” диалоге
- Нажмите кнопку  на панели задач чтобы запустить схему.
- Когда схема выполнится нажмите на отчет в правом нижнем углу окна:



- Открывшийся отчет содержит ссылку на файл с результатом:

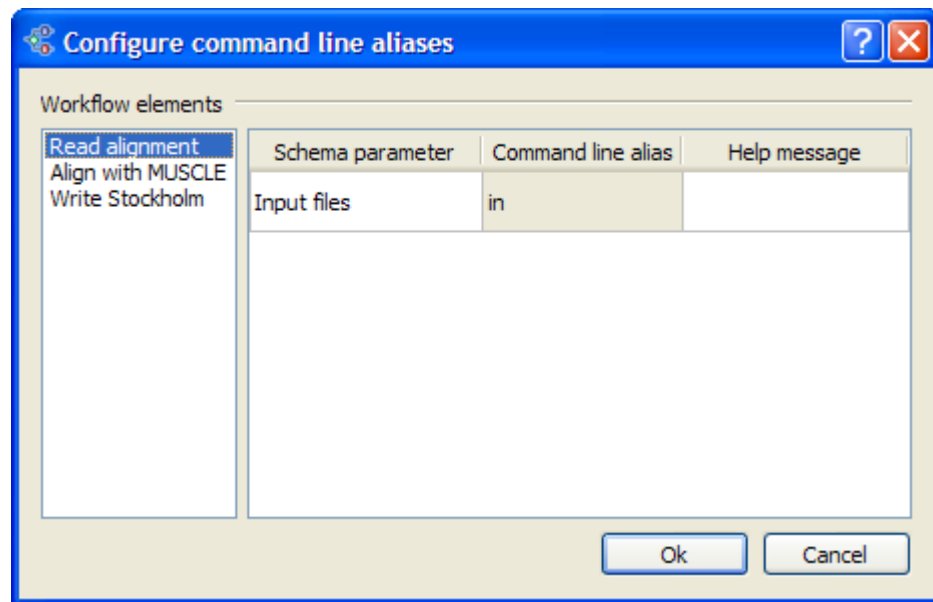
Output files:
<C:/Program Files/Unipro/UGENE/data/samples/CLUSTALW/result.sto>

4. Отредактировать параметры запуска схемы из командной строки:

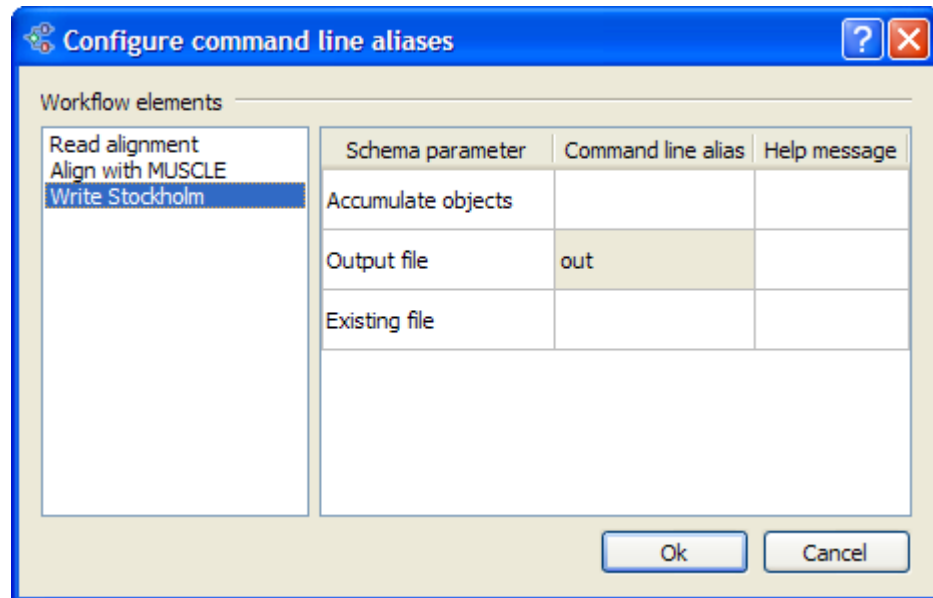
Итак, схема запускается из графического интерфейса, теперь зададим “алиасы” для параметров схемы, то есть названия параметров, которые будут использоваться при запуске схемы из командной строки.


- Нажмите на кнопку  на панели задач.

- В появившемся диалоге “Configure command line aliases” для параметра “Input files” элемента “Read alignment” задайте алиас “in”:

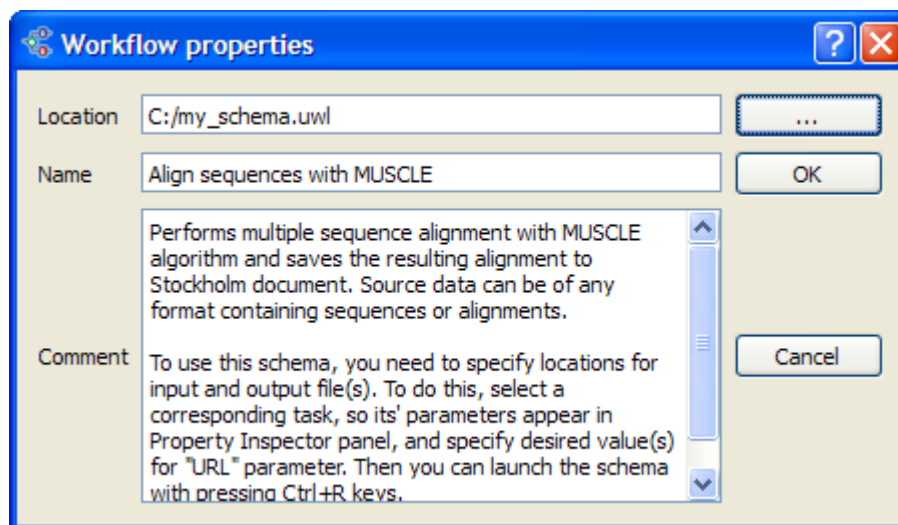


- Выберите элемент “Write Stockholm” и задайте алиас “out” для параметра “Output file”:



- Нажмите “Ok”.
- Нажмите на кнопку  на панели задач чтобы сохранить схему.

- В диалоге “Workflow properties” задайте “Location” (где будет сохранена схема):



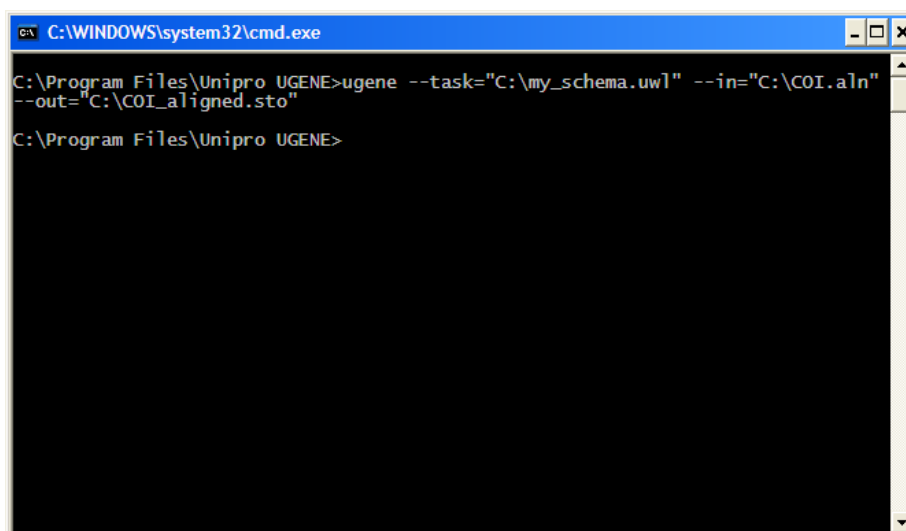
- Нажмите “OK”.

5. Запустить схему из командной строки:

- Откройте командную строку (например, в Windows можно запустить “cmd”).
- Для простоты примера положим “COI.aln” на диск “C:”.
- Запустите следующую команду:

ugene --task="путь к схеме" --in="входной файл" --out="выходной файл", то есть:

ugene --task="C:\my_schema.uwl" --in="C:\COI.aln"--out="C:\COI_aligned.sto"

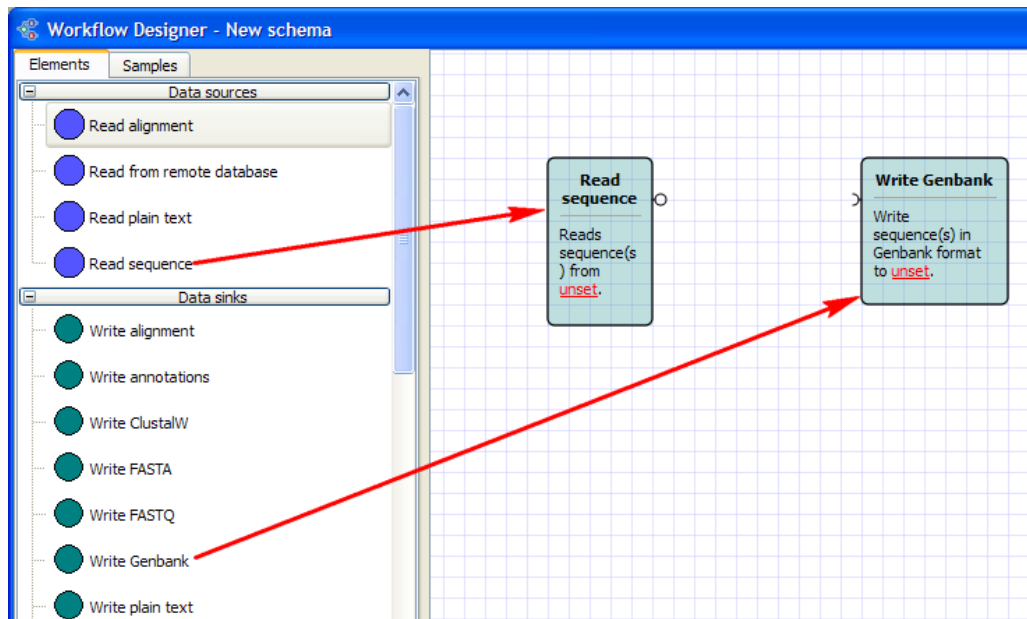


- Теперь можно открыть выходной файл “COI_aligned.sto” в UGENE.

Как сделать 5.2

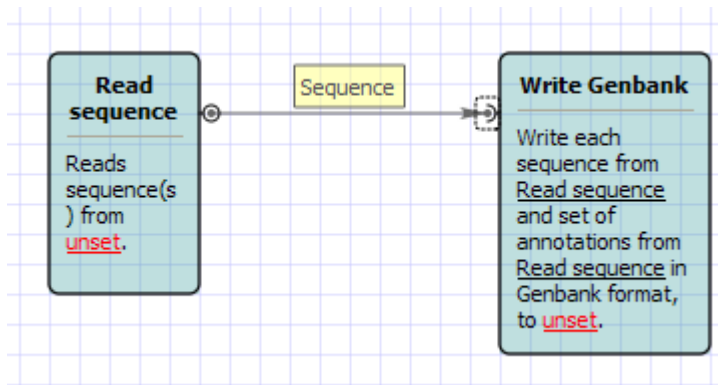
1. Запустить Workflow Designer в UGENE.
2. Перетащить необходимые элементы на сцену:

В данной схеме нам потребуется считывать последовательности и записывать их (будем записывать их в формате GenBank), поэтому перетащите элементы “Read sequence” и “Write Genbank”:



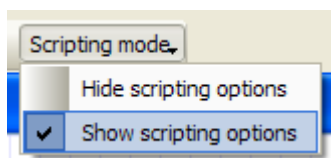
3. Соединить элементы:

Считываемые последовательности необходимо перенаправить на запись. Для этого соедините выходной порт элемента “Read sequence” с входным портом элемента “Write Genbank”:



4. Задать с помощью скрипта имена выходных файлов:

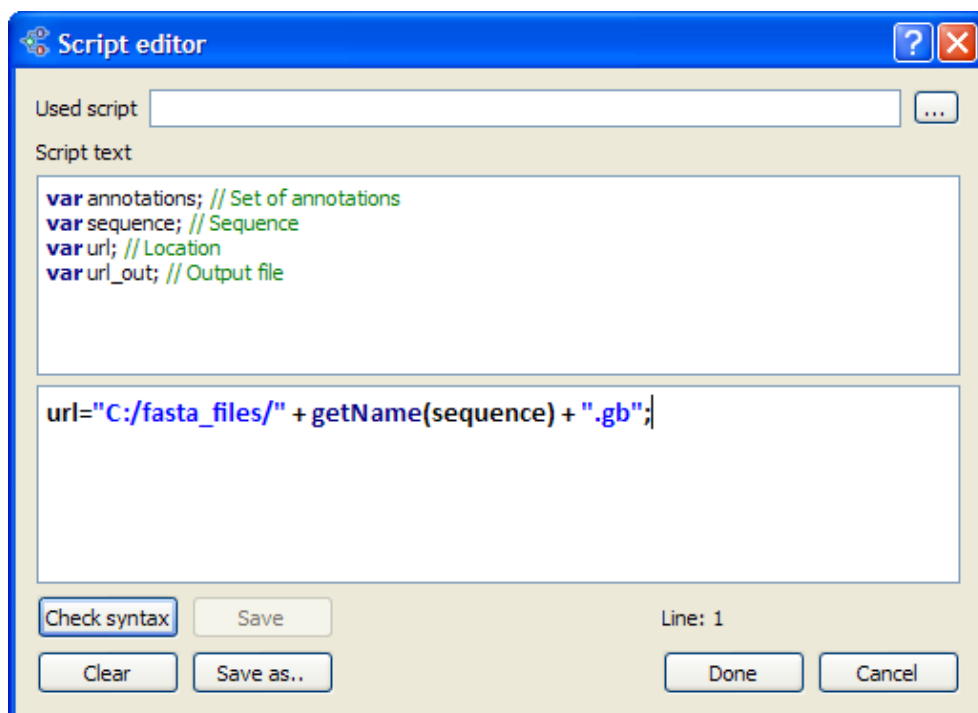
- На панели задач выберите значение “Show scripting options” для “Scripting mode”:



- Выберите элемент “Write Genbank” на схеме и в параметрах в колонке “Script” выберите значение “user script” для параметра “Output file”:

Name	Value	Script
Accumulate objects	True	N/A
Output file		no script
Existing file	Rename	no script
		user script

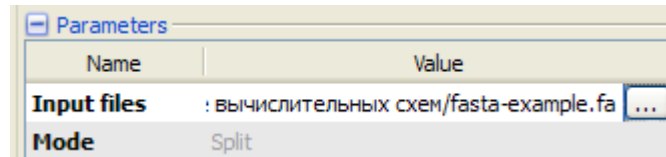
- В появившемся диалоге “Script editor” введите:
“url="C:/fasta_files/" + getName(sequence) + ".gb";”




- Нажмите “Done” чтобы сохранить настройки и закрыть диалог.

5. Задать входные данные:

Выберите элемент “Read sequence” и укажите входной multi-FASTA файл (или несколько файлов) в параметре “Input files”.



6. Выполнить схему:

- Нажмите кнопку  на панели задач чтобы запустить схему.
- После выполнения схемы файлы последовательностей будут находиться в указанной в скрипте папке: “C:\fasta_files”.

Часть II. Запуск задач на кластере HГУ

В этот разделе будут приведены краткие сведения о запуске задач на кластере HГУ.

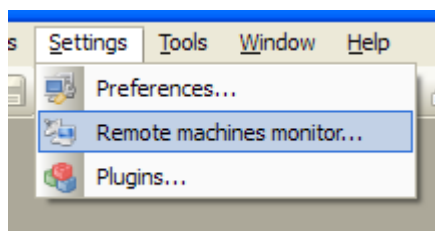
Чтобы запустить задачу на кластере надо:

1. Создать соответствующую вычислительную схему с помощью Workflow Designer.
2. Прописать адрес кластера в UGENE.
3. Запустить схему из Workflow Designer на удаленное выполнение.

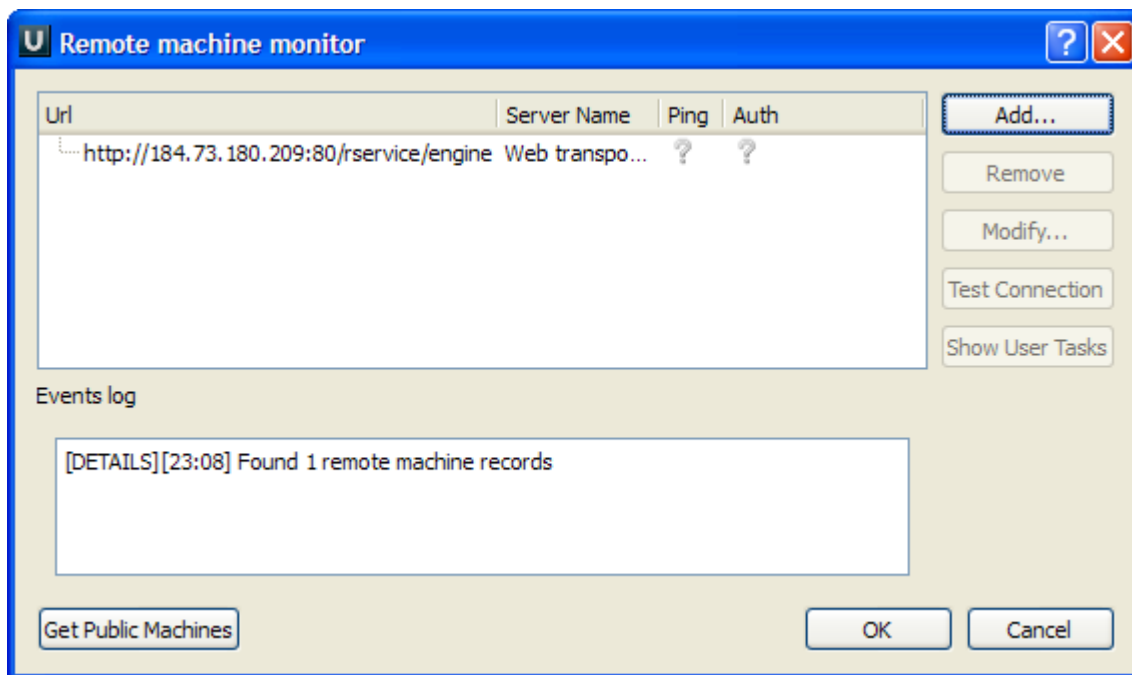
Описание пункта 1 не входит в данный раздел, однако ниже будут приведены несколько задач с примерами схем. Остальные два пункта описаны ниже.

1. Где в UGENE прописать адрес кластера

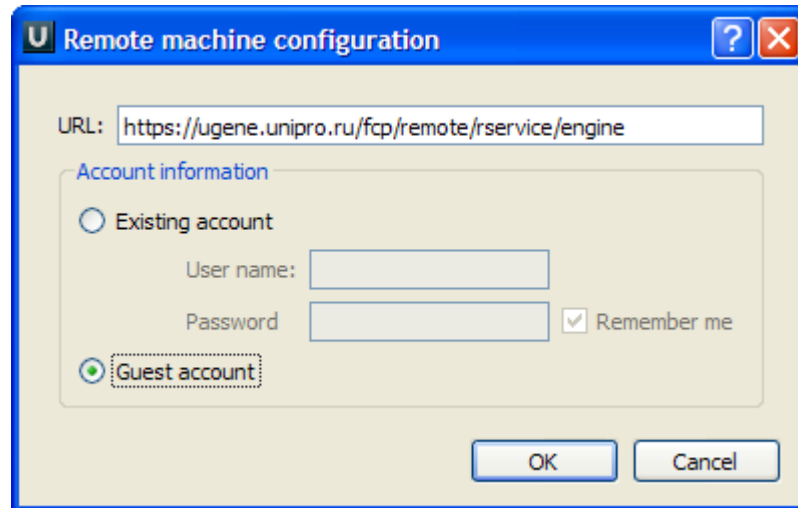
В главном окне UGENE выберите “Settings > Remote machines monitor”:



Появится “Remote machine monitor” диалог:

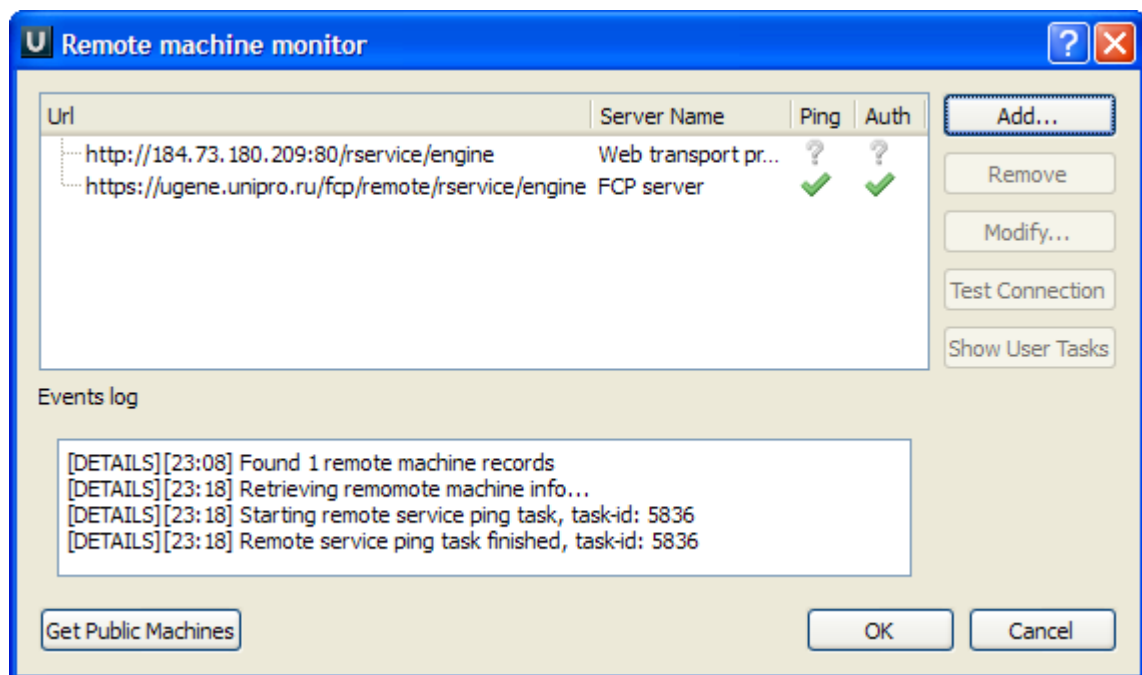


Нажмите кнопку “Add” и введите следующее значение в появившийся диалог:
“https://ugene.unipro.ru/fcp/remote/rservice/engine”



Вы можете зарегистрироваться на сервисе “ <https://ugene.unipro.ru/fcp/remote/user>” и указать в “Remote machine configuration” диалоге данные своего аккаунта (см. “Existing account”). Тогда запускаемые задачи можно будет отслеживать с помощью этого сервиса.

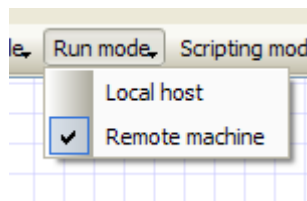
После нажатия кнопки “OK” в диалоге “Remote machine monitor” появится новая запись:




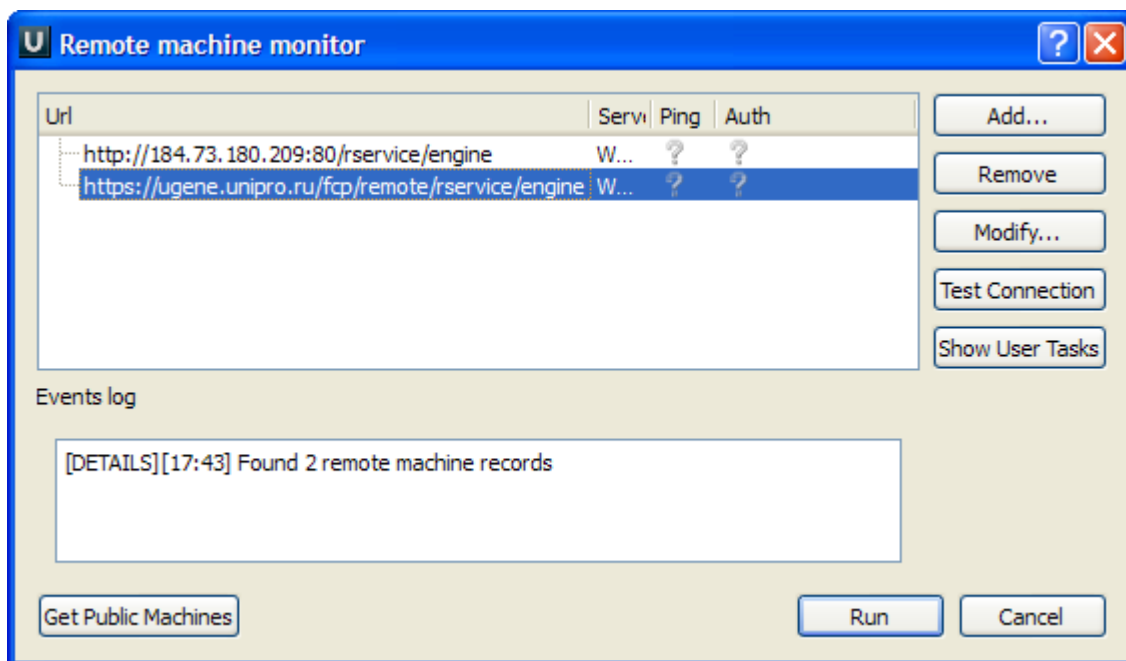
Нажмите “OK” чтобы закрыть диалог. Указанные настройки будут сохранены даже после перезапуска UGENE.

2. Как запустить схему на кластере

Откройте схему в Workflow Designer. На панели задач задайте значение “Run mode” равным “Remote machine”.



Нажмите на кнопку , также расположенную на панели задач, чтобы запустить схему. Появится “Remote machine monitor” диалог. Выберите в нем введенный ранее адрес кластера и нажмите “Run”:



3. Примеры схем

Пример 1. Поиск гена

Данные:

annotate_with_uql.uwl	Схема для запуска.
hs_ref_chr19_region_7mb.fa.gz	Входная последовательность, часть человеческой хромосомы 12.
simple_gene.uql	SimpleGene схема из примеров Query Designer с выбранной моделью Eklf.

Алгоритм запуска:

- Откройте схему “**annotate_with_uql.uwl**”.
- Укажите входной файл “**hs_ref_chr19_region_7mb.fa.gz**” (параметр “Input files” элемента “Read sequence”).
- Укажите UQL схему “**simple_gene.uql**” (параметр “Schema” элемента “Annotate with UQL”).
- Укажите выходной файл, например “**result.gb**” (параметр “Output file” элемента “Write Genbank”).
- Запустите схему удаленно (см. описание выше).

Замечание: В Query Designer схеме путь к SITECON модели (например Eklf) должен быть прописан как путь на кластере.

Пример 2. Выравнивание с помощью MUSCLE

Данные:

muscle.uwl	Схема для запуска.
Fungi.aln	Множественное выравнивание.

Алгоритм запуска:

- Откройте схему “**muscle.uwl**”.
- Укажите входной файл “**Fungi.aln**” (параметр “Input files” элемента “Read alignment”).
- Укажите выходной файл, например “**result.aln**” (параметр “Output file” элемента “Write alignment”).
- Запустите схему удаленно (см. описание выше).

Пример 3. Поиск паттерна

Данные:

sw_search.uwl	Схема для запуска.
hs_ref_chr19_region_7mb.fa.gz	ДНК последовательность в формате FASTA.

Алгоритм запуска:

- Откройте схему “**sw_search.uwl**”.
- Укажите входной файл “**hs_ref_chr19_region_7mb.fa.gz**” (параметр “Input files” элемента “Read sequence”).

- Укажите выходной файл, например **“result.gb”** (параметр **“Output file”** элемента **“Write Genbank”**).
- Запустите схему удаленно (см. описание выше).

Пример 4. Выравнивание ридов

Данные:

genome_aligner.uwl	Схема для запуска.
NC_008253.fna	Референтная последовательность (расположена на кластере).
e_coli_10000snp.fa	Короткие последовательности или риды.

Алгоритм запуска:

- Откройте схему **“genome_aligner.uwl”**.
- Укажите входной файл с ридами **“e_coli_10000snp.fa”** (параметр **“Input files”** элемента **“Read sequence”**).
- Укажите выходной файл, например **“result.sam”** (параметр **“Output file”** элемента **“Write alignment”**).
- Запустите схему удаленно (см. описание выше).

Замечание: Для того чтобы минимизировать объем передаваемых данных при запуске схемы, референтная последовательность была заранее загружена на кластер. В дальнейшем планируется расширять библиотеку доступных референтных геномов и предоставить пользователю возможность загружать их вручную.

Пример 5. Поиск BLAST

Данные:

blast_nr.uwl	Схема для запуска.
1CF7_region.fa	Аминокислотная последовательность в формате FASTA.

Алгоритм запуска:

- Откройте схему **“blast_nr.uwl”**.
- Укажите входную последовательность **“1CF7_region.fa”** (параметр **“Input files”** элемента **“Read sequence”**).
- Укажите выходную последовательность, например **“result.gb”**.
- Запустите схему удаленно (см. описание выше).

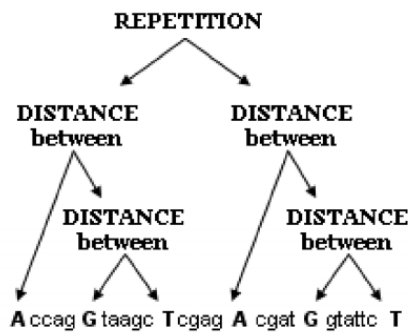
Часть III. Работа с модулем Expert Discovery в UGENE

1. Общие сведения о модуле Expert Discovery

Что такое Expert Discovery

Система Expert Discovery позволяет исследовать и аннотировать протяженные районы генов, отвечающие за регуляцию транскрипции, в частности, находить сайты связывания различных транскрипционных факторов.

Для исследования регуляторных областей применяются так называемые “комплексные сигналы”. Комплексный сигнал может быть представлен в виде дерева, состоящего из “элементарных сигналов” и условий, например:



Элементарным сигналом может являться некоторая короткая последовательность или в частном случае буква, как например показано на рисунке ('A', 'G', 'T', 'A', 'G', 'T').

Условие, накладываемое на сигналы может быть одним из следующих:

- **Distance:** Задано *min* и *max* расстояние между сигналами (элементарными или комплексными).
- **Repetition:** Сигнал должен повторяться от N_{min} до N_{max} раз. Задано также *min* и *max* расстояние между соседними повторами.
- **Interval:** Сигнал должен находиться в интервале от *min* до *max*.

Также при работе программы задается три набора выборок последовательностей:

- **Позитивные:** в данных последовательностях комплексный сигнал должен присутствовать.
- **Негативные:** в данных последовательностях комплексный сигнал отсутствует (с определенной долей вероятности).
- **Контрольные:** последовательности, проверяемые на наличие комплексного сигнала.

Таким образом, обнаруживаются комплексные сигналы, отличающие позитивную выборку от негативной. Частота встречаемости каждого сигнала в позитивной выборке значимо

отличается от таковой в негативной выборке. Качество полученных сигналов проверяется на контрольной выборке.

Где можно взять Expert Discovery

Большая часть функциональности оригинальной программы “ExpertDiscovery” встроена в UGENE в качестве модуля “Expert Discovery” (альфа версия).

Оригинальная версия программы “ExpertDiscovery”, а также статьи и документация по ней доступны по следующей ссылке: <http://www.math.nsc.ru/AP/ScientificDiscovery/index.html>

2. Практическая задача: Поиск комплексных сигналов на выровненной выборке

Что есть

Позитивные (“positive_learning.fa”), негативные (“negative_learning.fa”) и контрольные (“control.fa”) последовательности в формате FASTA.

Что требуется


Обучить программу отличать объекты позитивной выборки от объектов негативной выборки. Комплексные сигналы необходимо автоматически сгенерировать из букв, и применить их совокупность к контрольным последовательностям (процедура распознавания).

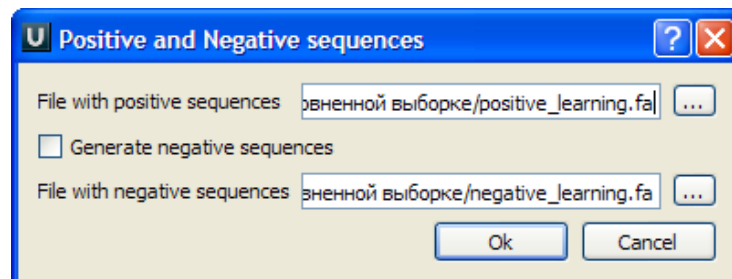
Как это сделать

1. Запустить Expert Discovery в UGENE:

Чтобы открыть окно Expert Discovery выберите “Tools > Expert Discovery (alpha)” в главном окне UGENE.

2. Загрузить выборки:

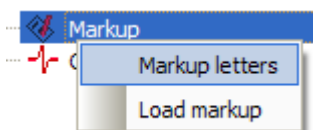
- Выберите кнопку  на панели инструментов.
- В появившемся диалоге “Positive and Negative sequences” выберите файлы с позитивной и негативной выборками:



- В следующем диалоге "Positive and Negative sequences markup" нажмите "Cancel", так как в данном примере комплексные сигналы будут сгенерированы автоматически.

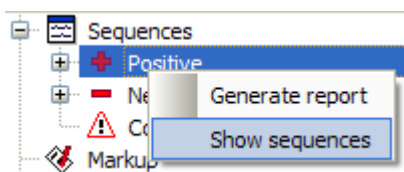
3. Выбрать буквы в качестве элементарных сигналов:

Для этого нажмите правой кнопкой мыши на пункте "Markup" в окне Expert Discovery и выберите "Markup letters" в появившемся контекстном меню:

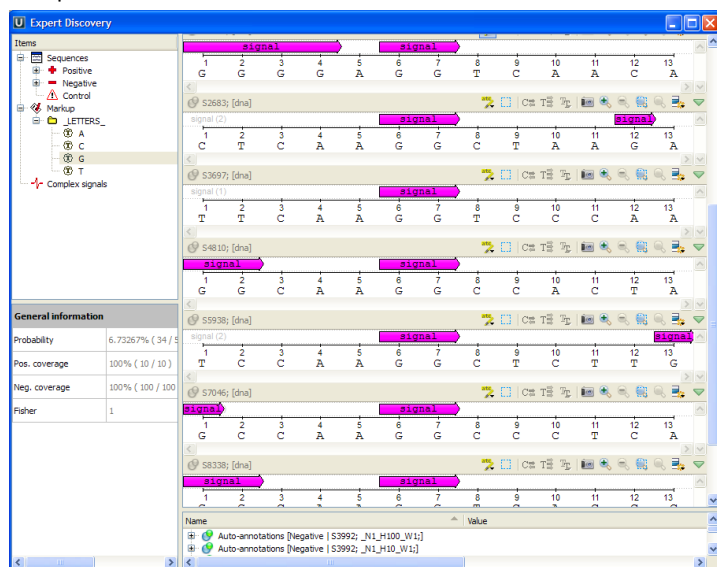


4. Показать последовательности: (этот пункт можно пропустить)

- Чтобы показать первые последовательности в позитивной выборке, выберите "Show sequences" в контекстном меню "Positive":



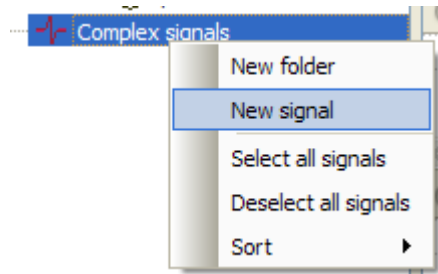
- Чтобы добавить еще "незагруженную" последовательность дважды щелкните мышью на ней.
- Выберите, например, букву (т.е. в данном случае элементарный сигнал) 'G' в "Markup > _LETTERS_". Все буквы 'G' отображаются на последовательностях в виде аннотаций:



Замечание: Обратите внимание на то, что выборка выровнена.

5. Создать сигнал вручную: (этот пункт можно пропустить)

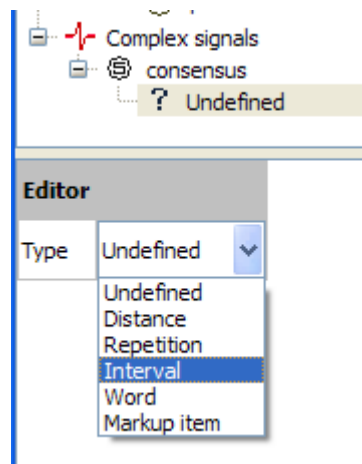
- Выберите “New signal” в контекстном меню “Complex signals”:



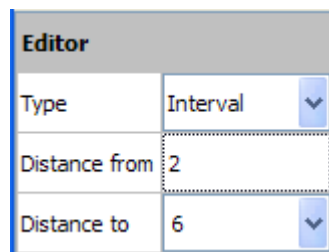
- Щелкните дважды мышью на имени созданного комплексного сигнала и переименуйте его в “consensus”:



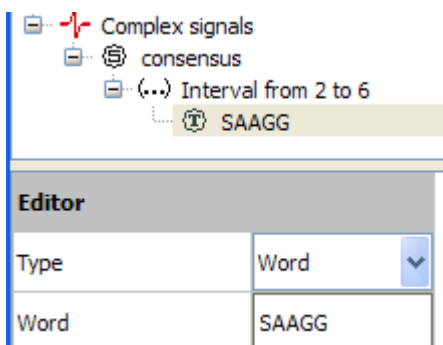
- Выберите подпункт “Undefined” для созданного сигнала:



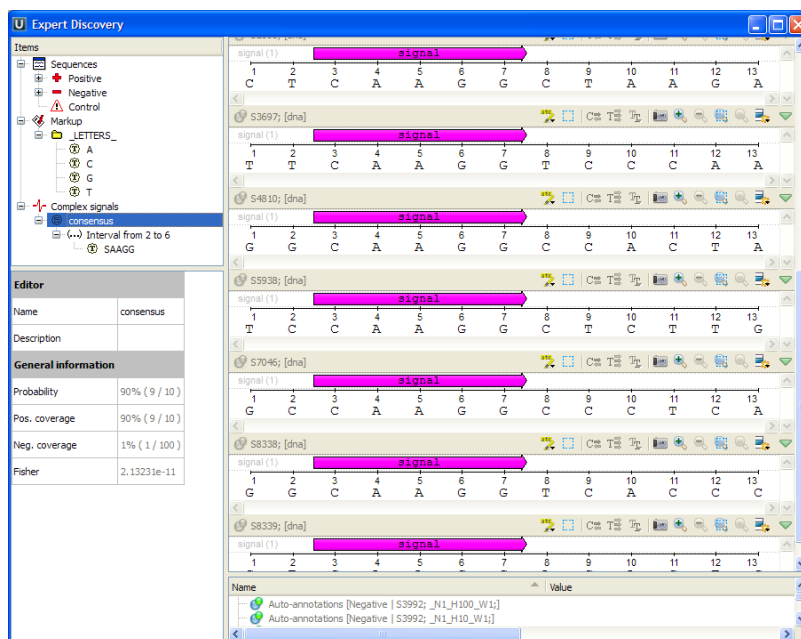
- Ниже задайте тип “Interval”, а также значения “2” и “6”:



- На втором уровне вложенности задайте слово “SAAGG” (в 15-символьном коде):

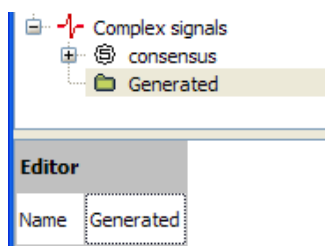



- Выберите созданный комплексный сигнал чтобы отобразить заданное слово в указанном интервале:

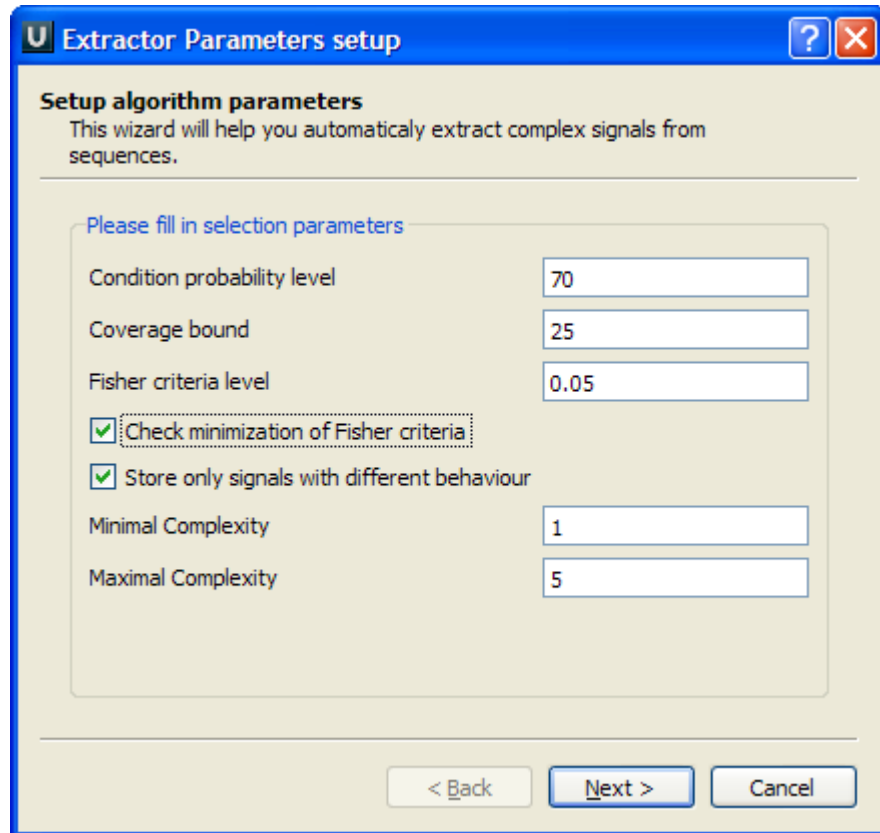


6. Автоматически сгенерировать сигналы:

- Создайте новую папку для комплексных сигналов (выберите “New folder” в контекстном меню “Complex signals” чтобы сделать это) и переименуйте ее в “Generated”:



- Нажмите кнопку  на панели задач.
- В появившемся диалоге “Extractor Parameters setup” измените значение параметра “Condition probability level” на “70”, отметьте галочку “Check minimization of Fisher criteria”:



Параметры в этом диалоге имеют следующее значение:

Condition probability level – порог условной вероятности, $P = a_{11} / (a_{10} + a_{11})$, где a_{11} - общее количество реализаций сигнала на позитивной выборке, a_{10} - общее количество реализаций сигнала на негативной выборке.

Coverage bound – уровень покрытия. Задаёт ограничение на количество реализаций сигнала на позитивной выборке. Например, в данном случае (см. выше) комплексный сигнал будет учитываться в случае, если он встретится более чем на 25 процентах позитивных выборок.

Fisher criteria level – порог уровня статистической значимости по точному критерию Фишера. Показывает на сколько статистически значим сигнал, то есть на сколько велика вероятность случайного возникновения данного сигнала.

Check minimization of Fisher criteria – минимизировать уровень статистической значимости сигнала.

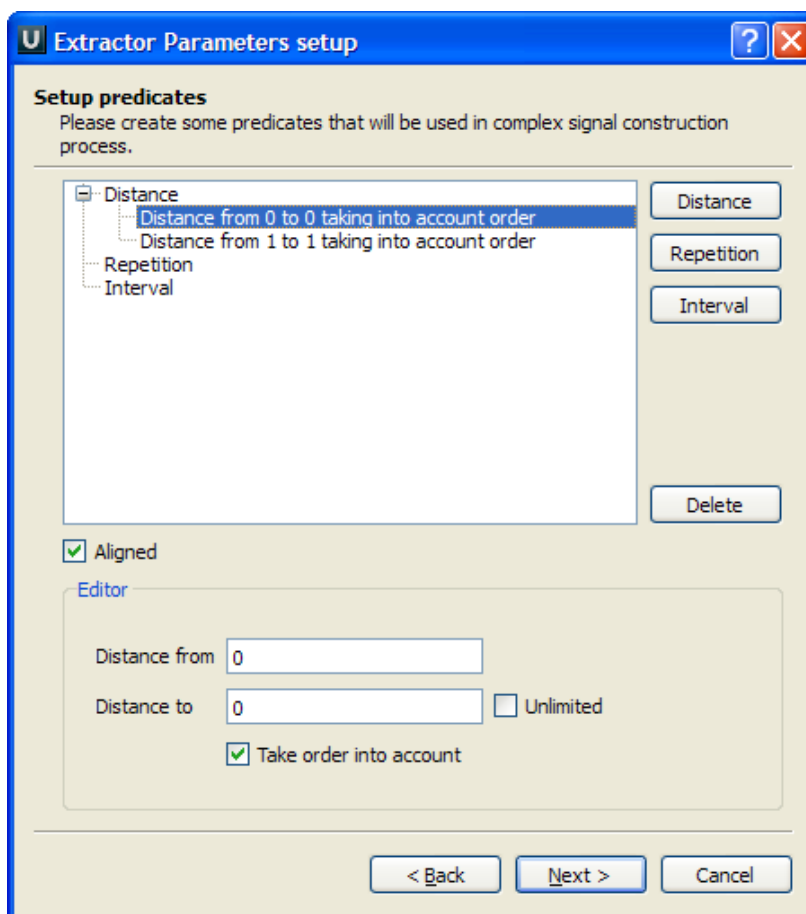
Store only signals with different behavior – отсеиваются сигналы с одинаковым поведением на выборках, то есть если 2 комплексных сигнала встречаются одинаковое количество раз на позитивных и негативных выборках, то второй сигнал не учитывается. Опция применяется чтобы избежать дублирования результатов когда один и тот же комплексный сигнал может быть представлен с помощью разных деревьев.

Minimal complexity – минимальная сложность комплексного сигнала.

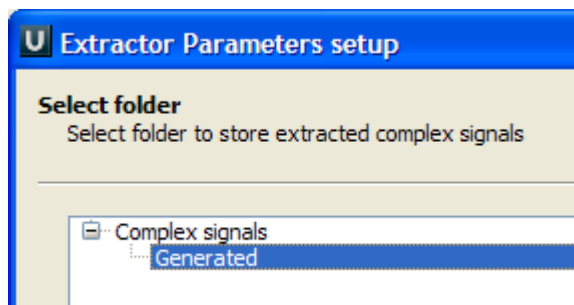
Соответствует минимальному количеству элементарных сигналов в дереве комплексного сигнала.

Maximum complexity – соответственно, максимальная сложность комплексного сигнала.

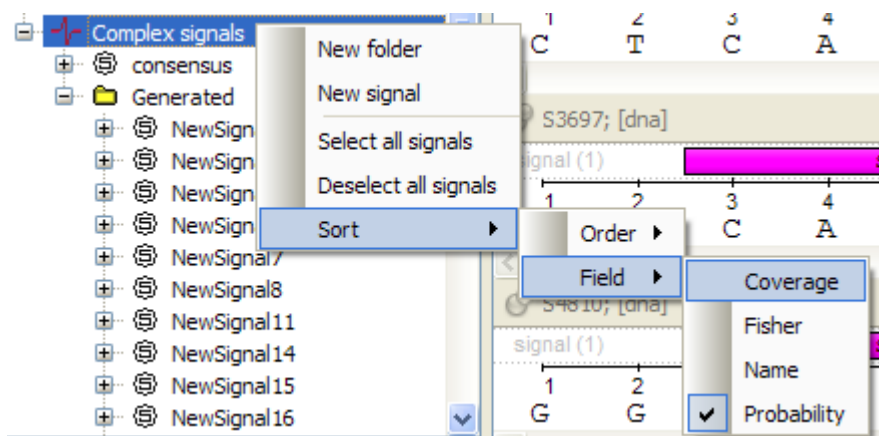
- Нажмите “Next”.
- В появившемся окне дважды нажмите кнопку “Distance” чтобы добавить 2 предиката расстояния.
- Выберите первый созданный предикат и задайте значение “0” для параметров “Distance from” и “Distance to” (предварительно убрав галочку “Unlimited”).
- Для второго предиката задайте значение “1” для обоих свойств:



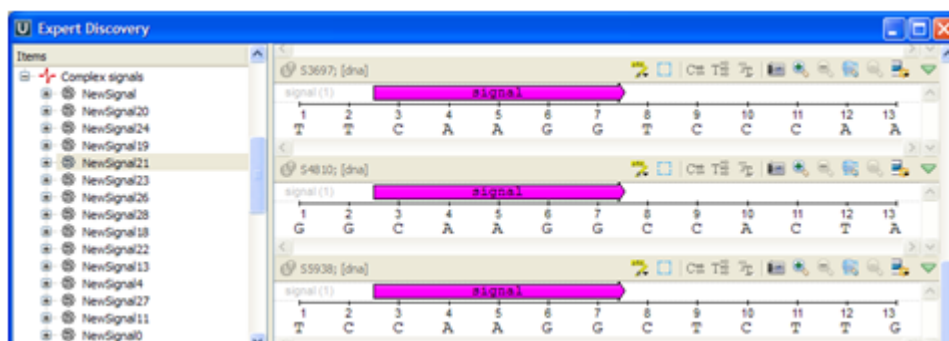
- Оставьте отмеченной галочку “Aligned”, чтобы при анализе учитывать, что выборки выровнены (см. также пункт 4 данной практической задачи).
- Нажмите “Next”.
- В следующем окне выберите ранее созданную папку “Generated”:




- Нажмите “Finish”.
- Отсортируйте комплексные сигналы по степени покрытия (выберите “Sort > Field > Coverage” в контекстном меню “Complex signals”):

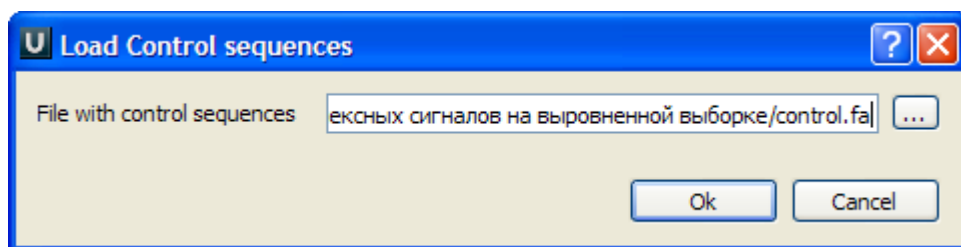


Замечание: Обратите внимание, что среди сигналов с наибольшим покрытием есть автоматически выделенный сигнал, который размечает ту же область, что и тот, который был создан вручную.




7. Загрузить контрольную выборку:

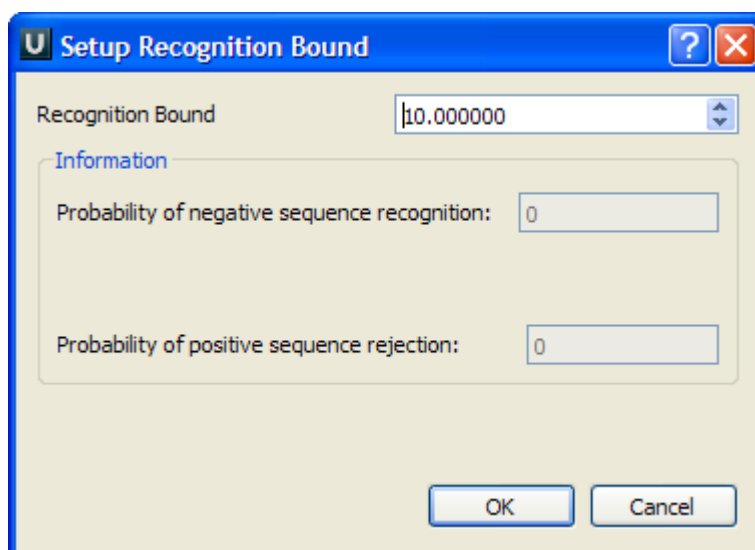
- Выберите кнопку  на панели задач.
- В появившемся диалоге выберите файл с контрольными выборками:



- Нажмите "OK".

8. Установить порог распознавания:

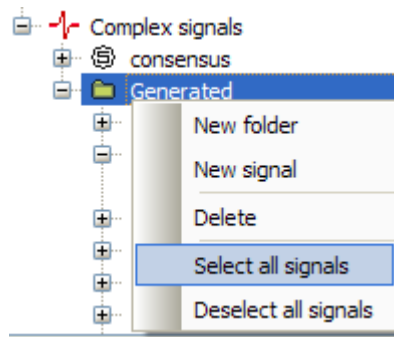
- Выберите кнопку  на панели задач.
- Установите значение "10" для "Recognition Bound". В этом окне можно следить за величиной ошибок первого и второго рода для текущего порога.




- Нажмите "OK".

9. Сгенерировать отчет:

- Выделите все сгенерированные сигналы, выбрав пункт "Select all signals" в контекстном меню папки "Generated":



- Выберите кнопку  на панели задач.
- Укажите имя файла отчета в появившемся диалоге. Отчет будет сохранен в формате HTML.
- Откройте отчет с помощью браузера (например Internet Explorer):

Control base

Total sequences: 9

Recognized sequences: 4

Sequences with zero score: 0

Details:

Sequence No	Sequence Name	Score	Result
1	S4929;	3.98898	Not recognized
2	S3292;	19.4139	Recognized
3	S3986;	39.3105	Recognized
4	S4803;	4.80811	Not recognized
5	S2638;	68.7333	Recognized
6	S2639;	5.59842	Not recognized
7	S5938;	23.0181	Recognized
8	S6368;	2.19722	Not recognized
9	S8076;	5.59842	Not recognized

Заключение

В данном пособии представлены лишь некоторые из возможностей UGENE, из числа тех что показались нам актуальными для представления на школе-семинаре. Многие темы упомянуты вскользь, другие вовсе не затронуты (например, клонирование или работа с хроматограммами).

Мы постоянно стремимся сделать наш продукт более доступным для пользователей и предлагаем воспользоваться нашими ресурсами для более полного знакомства с UGENE:

- Документация UGENE (<http://ugene.unipro.ru/documentation.html>)
- Подкаст UGENE (<http://ugene.unipro.ru/rus/podcast.html>)
- Форум (<http://ugene.unipro.ru/forum/>), в том числе на русском языке (<http://ugene.unipro.ru/forum/YaBB.pl?board=russian>)
- Система контроля задач UGENE (<https://ugene.unipro.ru/tracker>)

Вопросам и предложениям всегда рады по адресу ugene@unipro.ru.