Подсистема «Геномика»: исследование полноразмерных секвенированных геномов организмов.

Структура документа (оглавление).

1. Цель и задачи подсистемы «Геномика».

Геномика – недавно возникшее комплексное направление науки, объектом изучения которой являются геномы всех организмов.

Геномику можно разделить на 3 основных раздела:

- структурная геномика – содержание и организация геномной информации;
- функциональная геномика – реализация информации, записанной в геноме, от гена – к признаку;
- сравнительная геномика – сравнительные исследования содержания и организации геномов разных организмов;

Все эти разделы геномики вносят вклад в фундаментальную биологию (индивидуальное развитие, эволюция), здравоохранение, сельское хозяйство и биотехнологию.

Итог структурной геномики – получение последовательности нуклеотидов (сиквенс от англ. sequence), которая представляла бы полностью каждую из хромосом с первого нуклеотида до последнего (секвенированием в русскоязычной литературе называют процесс определения последовательности нуклеотидов, этот термин также является калькой с английского названия).

Одно из направлений геномики - воссоздание суммарной карты метаболических путей живого, состоящей из частных метаболических карт, характерных для каждого организма.

Выявление в разных геномах определенных наборов генов метаболических функций позволяет предположить, функциональную связь генов этого набора в едином участке метаболической цепи.

Критерий функциональной связи между генами, особо хорошо работающий на бактериях, основан на сохранении соседства одних и тех же (по сиквенсу) генов у разных видов бактерий. У бактерий нередко бывает, что группа генов, расположенных вместе, отвечает за группу последовательных этапов метаболизма. Такая группа генов регулируется на уровне транскрипции единым образом и называется оперон (единица операции - транскрипции). Часто последовательность расположения генов в опероне совпадает с последовательностью метаболических этапов. Для эукариот соседнее расположение функционально связанных генов не типично, но, хоть такие гены и разбросаны у них по геному, скоординированная регуляция транскрипции есть и эукариот.

На данный момент просеквенировано несколько сотен геномов бактерий и геномы нескольких эукариот. Теперь мы знаем, что у бактерий размеры генома не бывают меньше 0,5 миллионов пар нуклеотидов, а максимальный размер генома около 10 миллионов п.н., у дрожжей (эукариотический организм)– порядка 12 миллионов, у червя нематоды – 97 млн., а у человека – 3 миллиарда пар нуклеотидов. А число генов у про- и эукариот различается уже в меньшее число раз. Минимальное количество генов у бактерии микоплазмы – 470 штук, у дрожжей – 6000, у нематоды – 19000, а у человека около 20000, то есть от нематоды и мухи по количеству генов мы не сильно отличаемся. Количество хромосомной ДНК, приходящейся на один ген у бактерий -1000 п.н. то есть гены упакованы очень плотно; у дрожжей – 2000 п.н., и кое-где гены разделены некоторым пространством; у нематоды – 5000 п.н. на ген и появляются пространства внутри генов – интроны; у человека – 30000 п.н. У нас в геноме большие межгенные пространства и большие пространства внутри генов, которые не переходят в зрелую РНК.

Заметим, все эти организмы по размерам зрелых транскриптов не сильно отличаются. В зрелой РНК белок-кодирующий участок занимает обычно основную часть последовательности. Часть генов кодируют РНК, с которой белок вообще не синтезируется. Перед белок-кодирующей последовательностью в зрелой мРНК расположены участки регуляции трансляции, а после белок кодирующей последовательности – участки, определяющие стабильность (время жизни РНК). У прокариот последовательности перед и после белок-кодирующей части гораздо короче, чем у эукариот. Так что по размерам РНК все организмы ближе, чем по размерам генов, а по размерам белков – еще ближе.

2. Использование методов и подходов биоинформатики в геномике организмов: структура подсистемы «Геномика» и детальное руководство по ее применению.

2.1. Описание GenomeBrowser.

GenomeBrowser – программный комплекс, позволяющий во вновь полученной последовательности (сиквенсе) генома:

- найти открытые рамки считывания (ОРС), возможно кодирующие белки,
- выявить опероны, состоящие из ОРС,
- найти промоторы,
- сайты связывания транскрипционных факторов (ССТФ),
- характерные олигонуклеотиды – кандидаты на функциональные сайты,
- при наличии информации о высокоэкспрессирующихся мРНК в данном организме можно оценить интенсивность экспрессии всех секвенированных ОРС,
- построить профиль плотности повторов и увидеть их местоположение.

Описание работы с GenomeBrowser.[1]

2.1.1. Начало работы

Для того чтобы начать работу, нужно выбрать файл, содержащий необходимую последовательность в нуклеотидном представлении. Это можно сделать при помощи меню "File|Open".

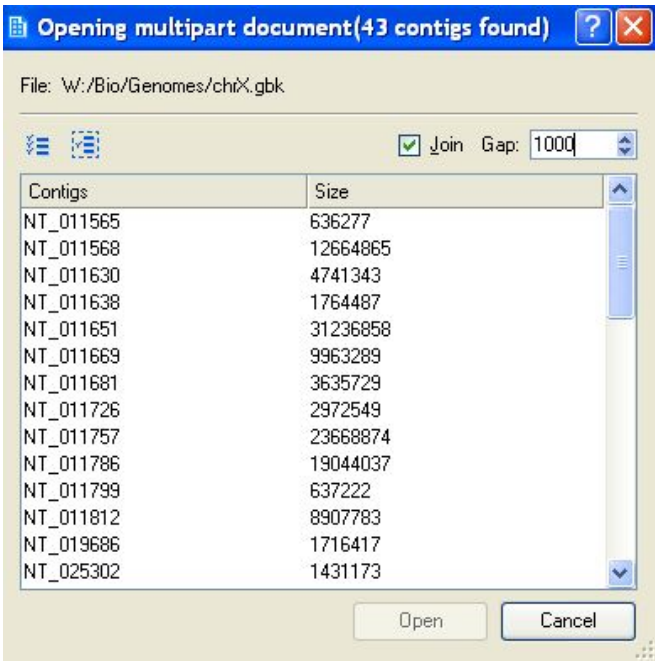


Рисунок 1.

Для файлов, содержащих несколько документов, появится диалог с уточнением, какой из документов необходимо открыть. В случае если эти документы являются частями контигами (contigs) одной последовательности и имеет смысл открыть их одновременно, то необходимо выбрать опцию "Join contings", задать размер пропуска между контигами "Gap size" (пропуск будет заполнен символами 'N') и выделить нужные контиги для открытия. GenomeBrowser может открывать только файлы, содержащие нуклеотидное представление последовательности. При открытии документа будут автоматически загружены и отображены все находящиеся в нем аннотации. GenomeBrowser может работать с большинством основных форматов представления последовательностей и аннотаций:

- ABI
- EMBL
- Fasta

[1] Подробнее описание работы с Genome Browser смотри в Руководстве пользователя АСНИ-01, раздел 4.2.2.1.

- Genbank
- CHR
- DMX

2.1.2. Основные компоненты GenomeBrowser.

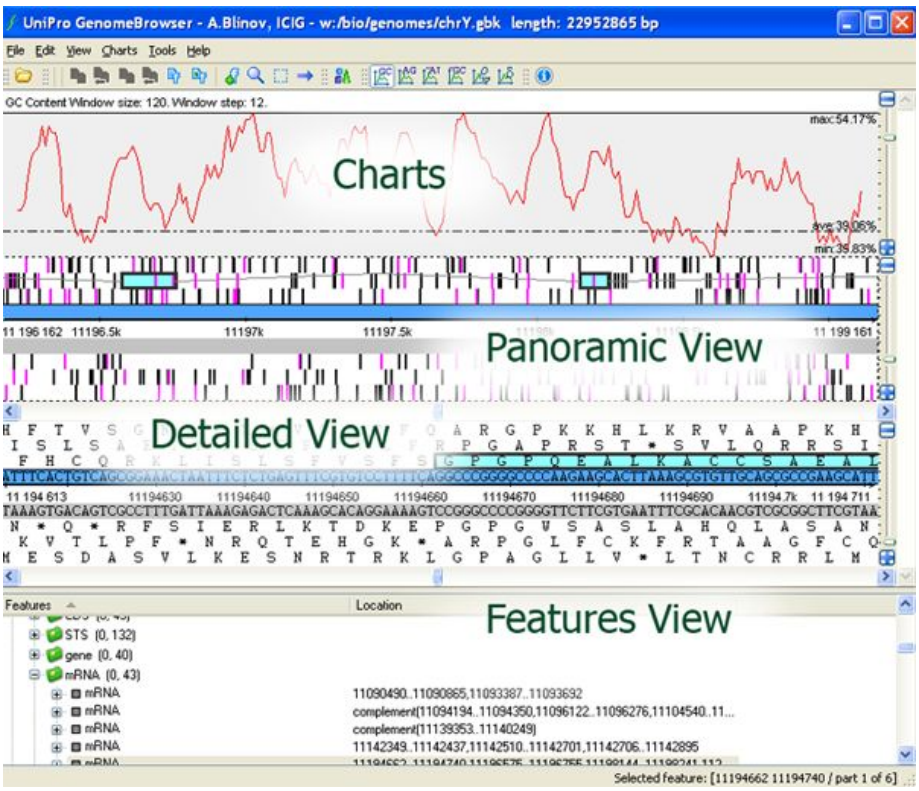


Рисунок 2.

Отображение последовательности

Отображение последовательности осуществляется при помощи двух взаимосвязанных компонент: Panoramic View и Detailed View. Каждый из них предоставляет возможности навигации, масштабирования, выделения и копирования заданного участка последовательности. Panoramic и Detailed View логически дополняют друг друга: первый всегда содержит область, отображаемую во втором, и помогает легче установить положение участка, отображаемого в Detailed View, относительно всей последовательности.

Feature Table View

Feature Table View предоставляет возможности визуализации и управления загруженными на данный момент аннотациями. Аннотация (feature) является описанием части последовательности и состоит из имени, координат и набора свойств. В соответствии с терминологией, принятой в Genbank, имя аннотации называется key, а набор свойств представлен парами "имя-значение" и именуется qualifiers. Аннотации могут быть как загружены, так и выгружены в процессе работы, одновременно с документом, которому они принадлежат.

Feature Table View предоставляет различные способы группировки аннотаций и выделение отдельных значений qualifier в дополнительные столбцы с возможностью

сортировки. Синхронизация текущего выделения Feature Table View с Detailed View позволяет одновременно просматривать как свойства, так и сам аннотированный участок в контексте последовательности и других аннотаций.

Графики контекстных статистик

В GenomeBrowser реализована возможность построения графиков наиболее часто используемых контекстных характеристик. Графики включаются при нажатии соответствующей кнопки панели инструментов, либо при выборе одного из пунктов меню Graph. В каждый момент времени на графике отображается контекстная статистика об участке последовательности, видимом на Panoramic View.

Пользуясь контекстным меню, можно сохранять выбранный график в виде изображения в формате PNG, либо в виде набора значений для последующего использования в любом редакторе электронных таблиц. С GenomeBrowser поставляются только самые базовые типы графиков. Их число может быть существенно увеличено при помощи динамически подключаемых модулей, доступных на сайте GenomeBrowser.

2.1.3. Работа с последовательностями.

Отображение последовательности осуществляется при помощи двух взаимосвязанных компонент: Panoramic View и Detailed View. Каждый из них предоставляет возможности навигации, масштабирования, выделения и копирования заданного участка последовательности. Panoramic и Detailed View логически дополняют друг друга: первый всегда содержит область, отображаемую во втором, и помогает легче установить положение участка последовательности, отображаемой в Detailed View, относительно всего документа.

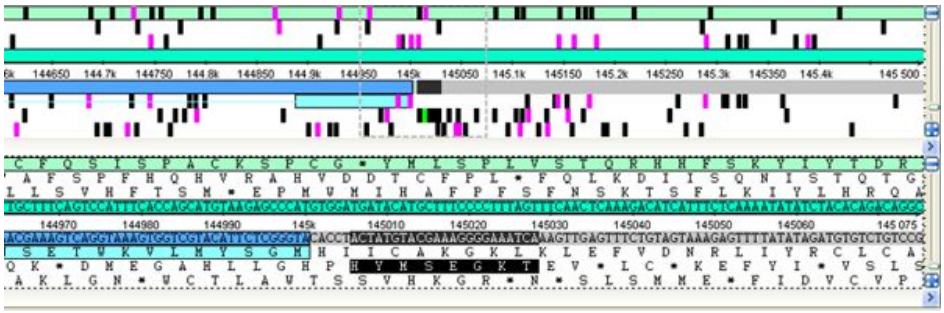


Рисунок 3.

2.1.4. Изменение масштаба и навигация

В GenomeBrowser реализован механизм масштабирования и навигации, делающий процесс навигации и анализа по большим последовательностям простым и понятным. После загрузки последовательности Detailed View установлен в режим максимального увеличения, в котором виден текст последовательности: прямой и комплементарный стренд и 6 возможных трансляций. Горизонтальный scrollbar, расположенный снизу Detailed View, позволяет перемещаться по последовательности, сохраняя текущий уровень увеличения. Для того, чтобы увеличить видимую область последовательности, необходимо использовать slider на правой стороне Detailed View, либо кнопку "-" сверху от slider. Уменьшая масштаб отображаемого участка, мы можем увеличивать отображаемую область последовательности до тех пор, пока не будет отображена вся последовательность. Текст последовательности

доступен только на максимальном уровне увеличения. Подсветка аннотаций будет происходить в любом из режимов увеличения.

Дополнительные возможности, предоставляемые Panoramic и Detailed View:
Переход на заданную координату в последовательности

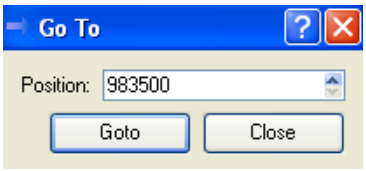


Рисунок 4.

Для того, чтобы позиционировать просмотр последовательности на заданную координату, можно воспользоваться диалогом перехода, который доступен через кнопку "GoTo" в панели инструментов, через пункт "GoTo" меню "Edit" или используя комбинацию горячих клавиш Ctrl+G.

Выделение и копирование участка последовательности

Для копирования нужного участка последовательности, прежде всего, необходимо сделать этот участок выделенным. Последовательность может быть выделена путем перемещения указателя мыши над нужным участком с нажатой левой кнопкой, либо при помощи кнопки "Select Region". Одновременно с выделением последовательности будет происходить и выделение трансляции, относящейся к выделенному участку. Помимо выделения последовательности можно выделить видимую аннотацию, для этого достаточно просто нажать на аннотированный участок последовательности. В этом случае под выделением будут пониматься все участки последовательности, составляющие данную аннотацию.

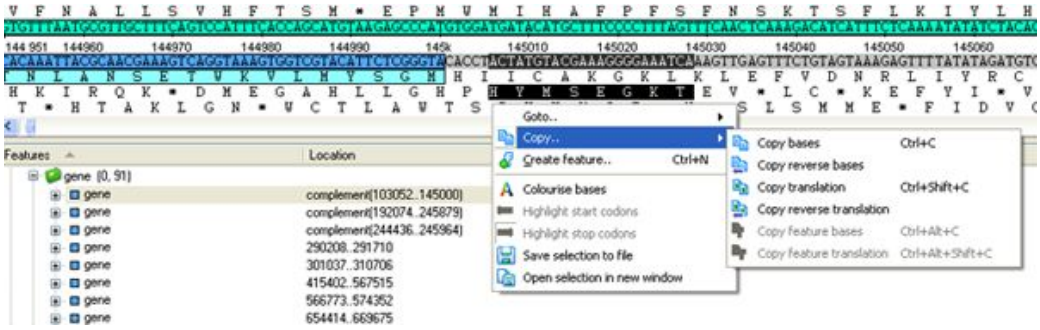


Рисунок 5.

Существует несколько способов копирования выделенного участка последовательности: копирование последовательности, копирование трансляции и копирование обратной последовательности и трансляции (reverse). Под обратным копированием подразумевается направление по отношению к визуальному представлению. Так, чтобы корректно скопировать участок из комплементарного стренда последовательности, необходимо выделить участок в комплементарной области и выбрать пункт меню "Copy reverse sequence". Полученный в результате reverse-complement участок будет являться прямым по отношению к биологическому прочтению последовательности и обратным по отношению к визуальному представлению. Копирование выделенного участка может проводиться как при помощи пунктов меню "Edit", так и при помощи кнопок на toolbar или "горячих клавиш".

Copy Bases - копировать отмеченную часть последовательности (так же доступно по Ctrl+C)
Copy Translation - копировать отмеченную трансляцию (так же доступно по Ctrl+Shift+C)
Copy Reverse Sequence - копировать отмеченную часть последовательности задом наперед
Copy Reverse Translation - копировать отмеченную трансляцию задом наперед

Точный и неточный поиск образца в последовательности

GenomeBrowser позволяет искать как точные, так и неточные вхождения образца в последовательность. При неточном поиске допускается несовпадение ряда символов в искомом и найденном участках. Для активации диалога поиска необходимо нажать кнопку "Find..." на панели инструментов, либо использовать комбинацию горячих клавиш "Ctrl-F".

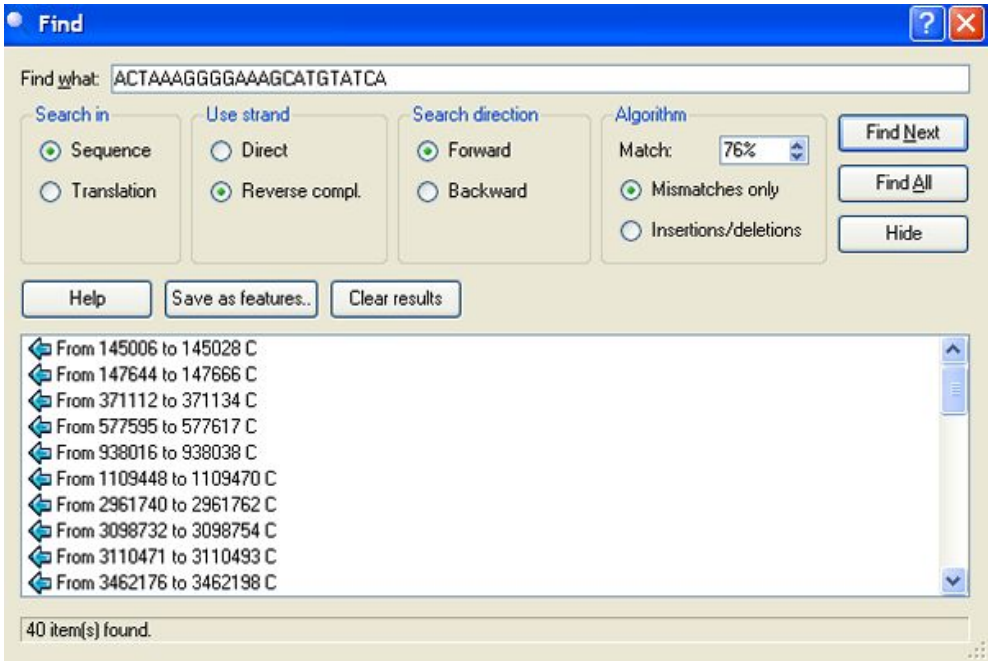


Рисунок 6.

Возможно, выполнять поиск как по двум из нуклеотидных представлениям последовательности, так и по любой из 6 трансляций. Кэширование результатов поиска позволяет легко возвращаться к найденным участкам, либо сохранять их в качестве аннотаций к последовательности для последующего использования. Алгоритм поиска оптимизирован для многопроцессорных систем и может работать в фоновом режиме.

2.1.5. Работа с аннотациями

Аннотацией выделенного участка последовательности является набор характеристик, описывающих свойства данного участка. Аннотации принято давать имя, определяющим собой тип описываемого элемента и набор свойств вида "имя-значение". В качестве аннотируемого участка, возможно, рассматривать как отдельную подстроку последовательности, так и непересекающийся набор подстрок.

GenomeBrowser является удобным инструментом для работы с аннотациями и предоставляет возможности редактирования, сохранения, сортировки и анализа данных, представленных в аннотациях.

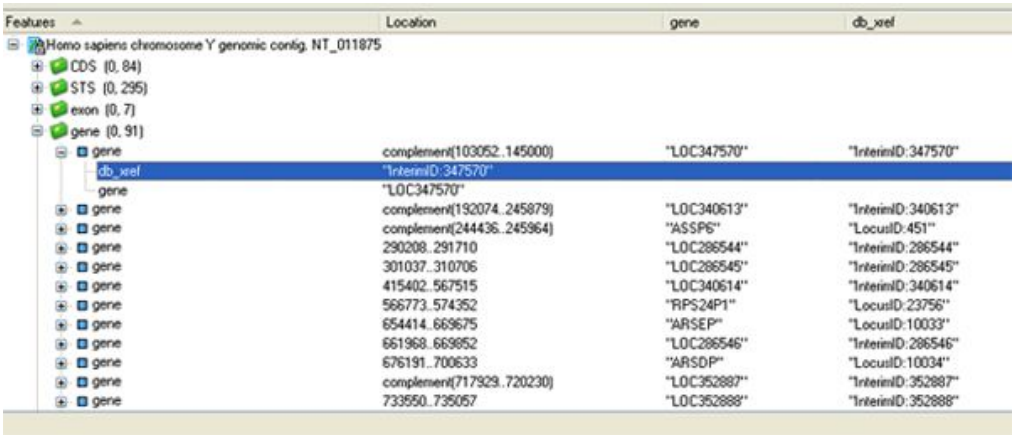


Рисунок 7.

Загрузка и сохранение аннотации

Аннотации к последовательности могут храниться во множестве файлов, разделенные удобным для пользователя способом. В каждый момент времени для открытого документа, возможно, подгрузить новые аннотации, либо выгрузить загруженные. При открытии документа с последовательностью, содержащего аннотации, все найденные в нем аннотации загружаются автоматически. Для того, чтобы в процессе работы загрузить дополнительные аннотации, нужно воспользоваться меню File->Load Features , либо контекстным меню Feature Table View. Опция "Unload Features" становится доступной при выделенном документе с аннотациями в Feature Table View.

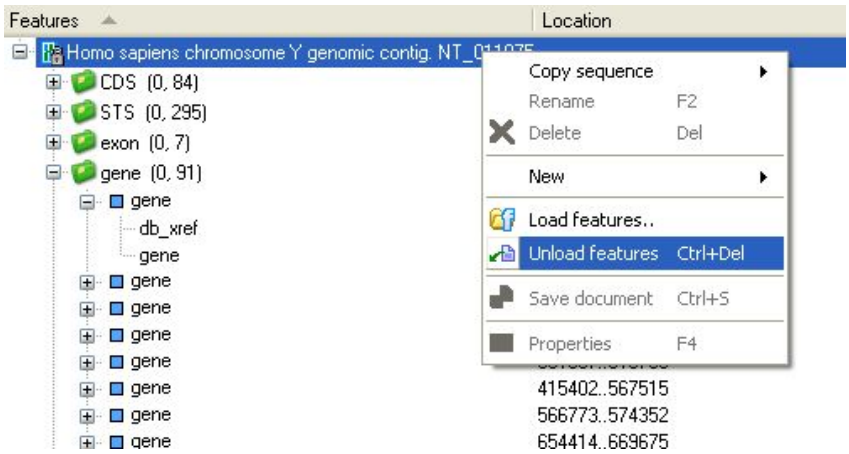


Рисунок 8.

Документы с аннотациями, созданные вне GenomeBrowser, не могут быть модифицированны. Основанием для этого ограничения является то, что при получении обновленного документа с аннотациями от 3-rd party пользователь не потеряет собственные наработки, сохраненные в отдельной файле. Документы, содержащие аннотации, созданные пользователем, могут быть изменены, сохранены, выгружены и загружены вновь без каких-либо ограничений.

Создание собственных аннотаций

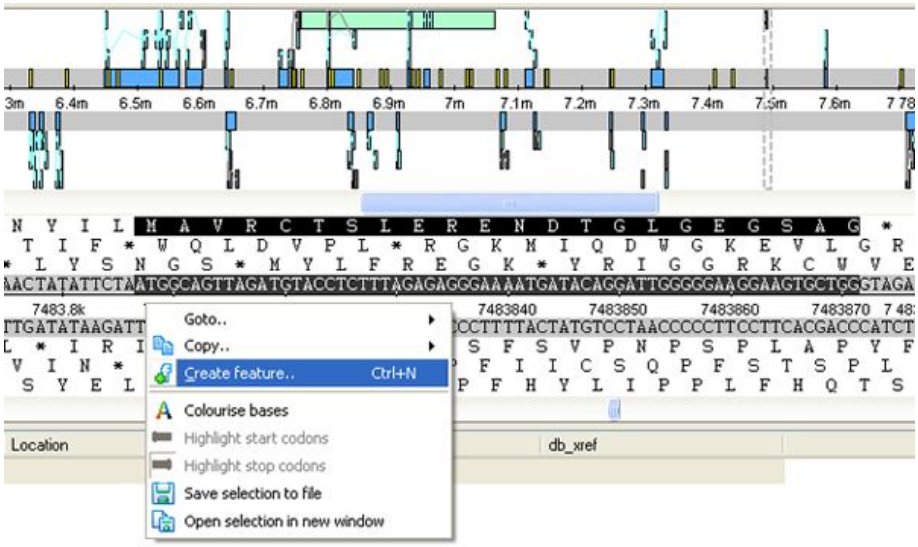


Рисунок 9.

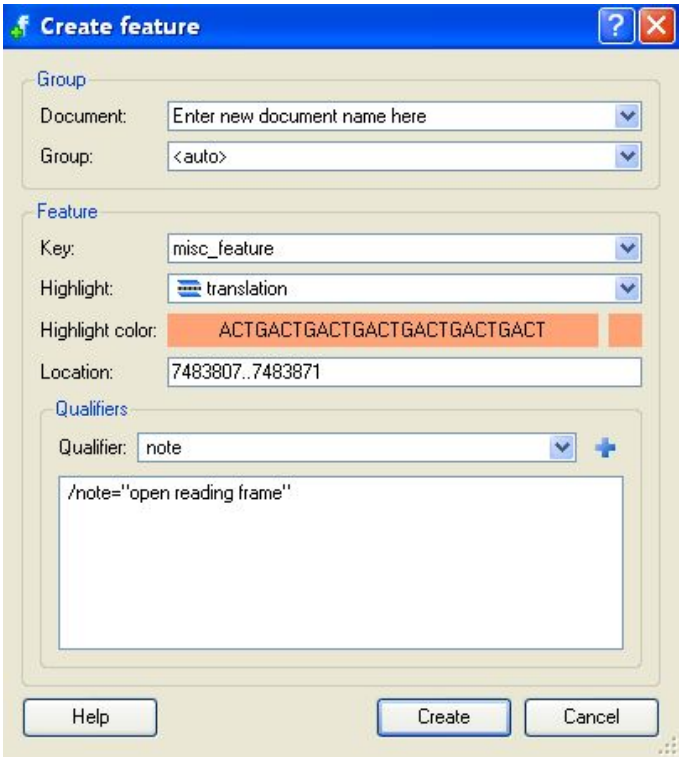


Рисунок 10.

Создание собственных аннотаций, возможность быстро возвращаться к интересующим участкам и вовлечение аннотаций в процесс автоматического анализа помогает значительно упростить и ускорить работу исследователя. Для создания аннотации в GenomeBrowser достаточно выделить участок последовательности и выбрать пункт "Create New Feature" из контекстного меню.

При этом вызывается диалог создания аннотации, позволяющий создать аннотацию с интересующими характеристиками.

Группа

Group - название группы, которой будет пренадлежать новая аннотация. Если группа с заданным именем не существует, то она будет создана.

Highlighting - тип подсветки аннотаций, принадлежащих группе (sequence - подсветка располагается на последовательности, translation - подсветка располагается на соответствующей трансляции).

Highlight color - цвет подсветки аннотаций, принадлежащих группе. Для задания цвета нажмите кнопку справа.

Аннотация

Key - название аннотации. Можно выбрать название аннотации из существующих, либо задать самому.

Location - позиция аннотации.

Add qualifier - список квалификаторов аннотации, которые можно добавить. Для того чтобы добавить выбранный квалификатор, нажмите кнопку ✚. Если квалификатор отсутствует в списке, то его можно ввести вручную.

После нажатия на кнопку Create будут созданы:
документ с заданным именем, если он не существовал;
группа с заданным именем, если она не существовала;
аннотация с заданным именем и параметрами.
Аннотация будет выделена.

Сортировка и группировка аннотаций

GenomeBrowser предлагает несколько режимов для визуального представления аннотаций и различные варианты сортировок. Переключение режимов представления аннотаций может быть осуществлено при помощи пунктов контекстного меню Feature Table View Group by Document, Arrange/Arrange by Group и Arrange/Feature меню Arrange. Кроме этого, аннотации могут быть отсортированы по любому из их свойств. Для этого следует воспользоваться контекстным меню для конкретного свойства аннотации и выбрать пункт Add Column '...'. Эта функция позволяет создать новый столбец, отображающий значения свойств всех загруженных аннотаций с тем же именем, что и у выбранного свойства. Для созданного столбца станут доступны опции сортировки по значению, дающие дополнительные возможности группировке.

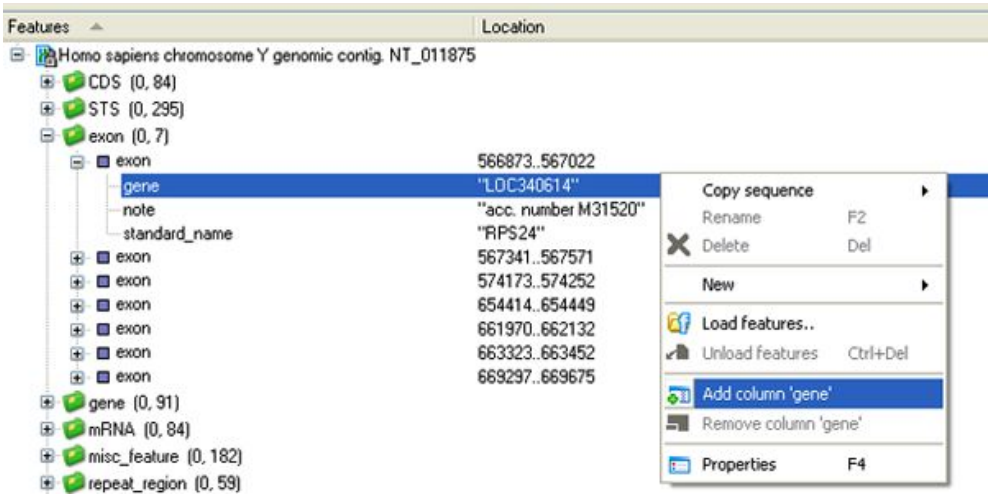


Рисунок 11.

2.1.6. Контекстные статистики

В GenomeBrowser реализована возможность построения графиков для наиболее часто используемых характеристик. Графики включаются при нажатии кнопки панели инструментов , либо при выборе одного из пунктов меню "Graph".

Все графики привязаны к области последовательности, отображаемой на "Panoramic View", и будут пересчитываться каждый раз, когда на "Panoramic View" изменяется отображаемый участок.

Типы графиков

GenomeBrowser может отображать 5 различных контекстных статистик для участка последовательности в "Panoramic View":

- GC Content (%) - Процентное содержание GC.
- AG Content (%) - Процентное содержание AT.

- AT Deviation (A-T)/(A+T) - Разность содержания A на прямой и комплементарной цепях ДНК.
- GC Deviation (G-C)/(G+C) - Разность содержания G на прямой и комплементарной цепях ДНК.
- Features Coverage. - Плотность покрытия последовательности аннотациями.

Число статистик может быть существенно увеличено при помощи подключаемых модулей, описанных в разделе Работа с подключаемыми модулями.

Основные параметры графиков

GenomeBrowser использует два типа алгоритмов для построения графиков - оконные (window) и обычные(standard). Стандартные алгоритмы отличаются от оконных методом подсчета значений. В простом алгоритме величины подсчитываются непосредственно в точках. При использовании оконного алгоритма общий вид графика задается тремя характеристиками - масштабом, шагом и размером окна. Масштаб графика определяется значением, заданным во фрейме масштабированного просмотра (вертикальная полоса прокрутки). Размер окна (регулируется вертикальной полосой прокрутки на фрейме графика) - это длина отрезка, который берется для подсчета значения точки графика. Соответственно, чем больше размер окна, тем выше точность. Шагом окна называют расстояние, на которое сдвигается окно для подсчета очередной точки. Эта величина выставляется автоматически в зависимости от размера окна и не регулируется пользователем.

Экспорт статистических данных

Пользуясь контекстным меню можно сохранять построенные графики в виде изображения (Export As Picture...), либо в виде набора значений для каждой из точек графика (Export As CSV...).

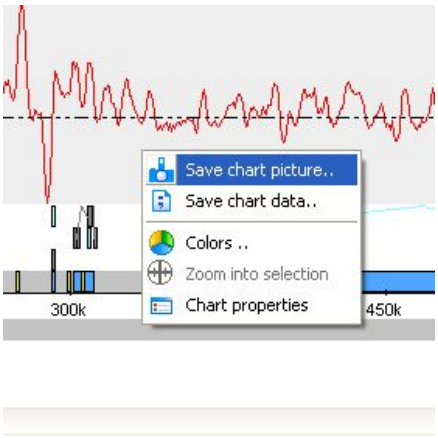


Рисунок 12.

В последнем случае возможен импорт данных в редактор таблиц для их последующей обработки.

2.2. Описание плагинов GenomeBrowser

Плагины GenomeBrowser (подключаемые модули) являются программами, встраивающимися в интерфейс GenomeBrowser и существенно расширяющими его возможности.

Описание работы с программными компонентами[2]

2.2.1. Подключение новых плагинов

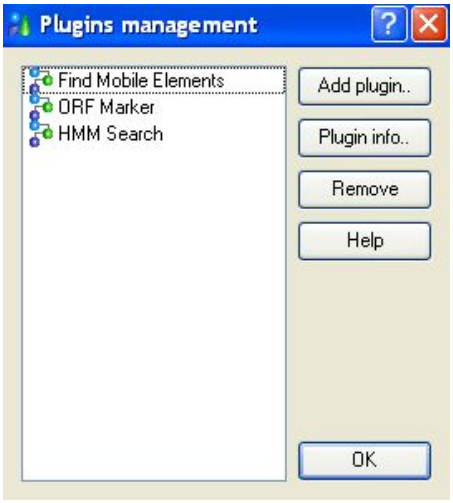


Рисунок 13.

2.2.2. ORFMarker – плагин для выявления открытых рамок считывания (ОРС) в геномных последовательностях.

Открытая рамка считывания начинается с AUG кодона – старт-кодона трансляции и заканчивается одним из 3-х стоп-кодонов. Открытая рамка считывания может кодировать какой-нибудь белок, хотя и не обязательно. Для найденных открытых рамок считывания впоследствии можно искать гомологи в хорошо изученных организмах и делать выводы о функции кодируемого белка.

Для решения задачи выявления открытых рамок считывания система GenomeBrowser использует плагин **ORFMarker**, который выполняет поиск открытых рамок считывания в прямой (5'->3') и комплементарной (3'->5') цепях с заданной пользователем минимальной длиной. Пользователь также выбирает используемую таблицу генетического кода и один или несколько инициирующих кодонов (ATG, GTG, TTG).

Выявление открытых рамок считывания может проводиться в последовательности ДНК длиной не менее минимальной длины (задаваемый параметр) и не более 10 000 000 нуклеотидов.

Анализируемая последовательность должна быть считана в систему Genome Browser согласно пункту 4.2.2.1 руководства пользователя. После подключения плагина **ORFMarker**, в *панели команд* и *меню инструментов* главного окна программы появится дополнительная команда запуска плагина (Рис. 14, 15):

[2] Подробнее описание работы с плагинами Genome Browser смотри в Руководстве пользователя АСНИ-01, раздел 4.2.2.1.

Рисунок 14.

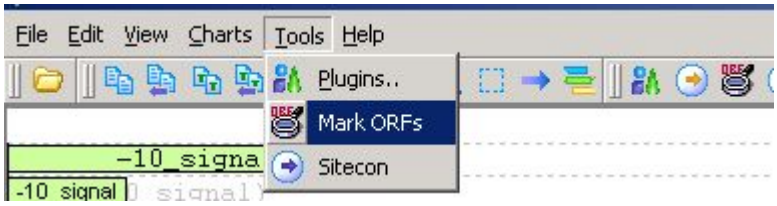


Рисунок 15.

После запуска **ORFMarker** в GenomeBrowser появится диалоговое окно, в котором пользователь может указать необходимые для поиска ОРС параметры (Рис. 16.):

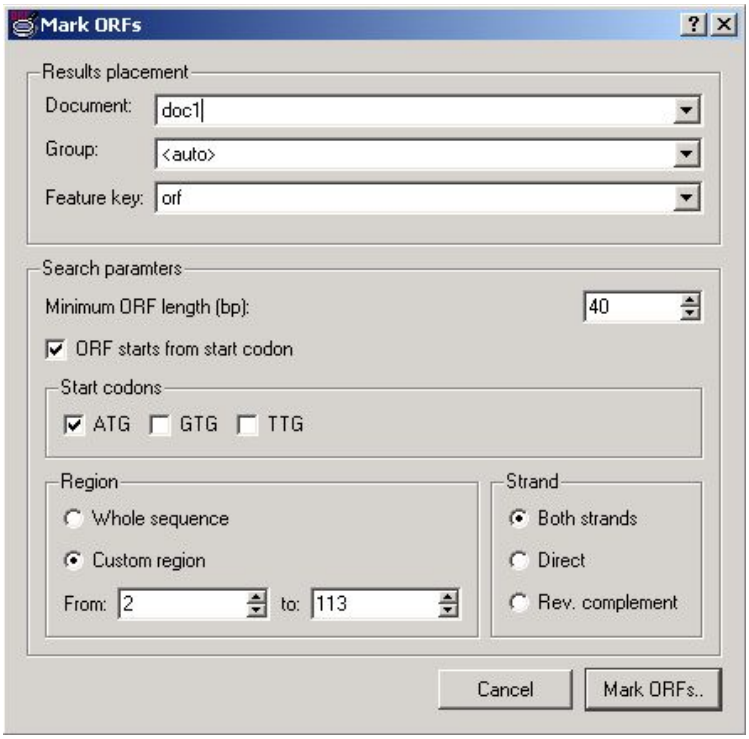


Рисунок 16. Диалоговое окно ORF Marker

В разделе *Results placement* задаются параметры размещения найденных открытых рамок считывания: В поле *Document* пользователь может указать имя документа (новое или существующее), в которое будет помещена группа найденных ОРС. Имя группы задается в поле *Group*. Имя, которое будет присвоено найденным ОРС, можно выбрать из выпадающего меню *Feature Key*, или ввести явно.

В разделе *Search Parameters* задаются нужные пользователю параметры поиска ОРС. В текстовом поле *Minimum ORF Length (bp)* указывается минимальная необходимая длина найденной ОРС (100 по умолчанию). Если пометка *ORF starts from start codon* отмечена, то поиск ОРС будет осуществляться только с начальных кодонов.

В подразделе *Start codons* можно определить последовательности нуклеотидов, которые будут приниматься за стартовые кодоны, отметив нужные пометки из возможных *ATG*, *GTG*, *TTG*.

В подразделе *Region* задаются границы поиска ОРС. Если выбрана опция *Whole sequence*, то поиск будет осуществляться по всей последовательности нуклеотидов. Если выбрана опция *Custom Region*, поиск будет осуществлен только по выбранной пользователем части последовательности, верхняя и нижняя границы которой задаются в полях *From* и *To* соответственно.

В подразделе *Strand* задается направление поиска по последовательности. Если выбрана опция *Direct*, поиск будет проводиться по прямой последовательности, если *Rev. complement* – то по комплементарной, и если *Both strands* – то в обоих направлениях.

По нажатию кнопки *Mark ORFs..* начнется поиск ОРС, диалоговое окно закроется, и результаты поиска будут отображены в *панели аннотаций*. По нажатию кнопки *Cancel* диалоговое окно закроется, и поиск осуществлен не будет.

Результатом работы утилиты является список найденных ОРС в *панели аннотаций* в окне GenomeBrowser (Рис. 17).

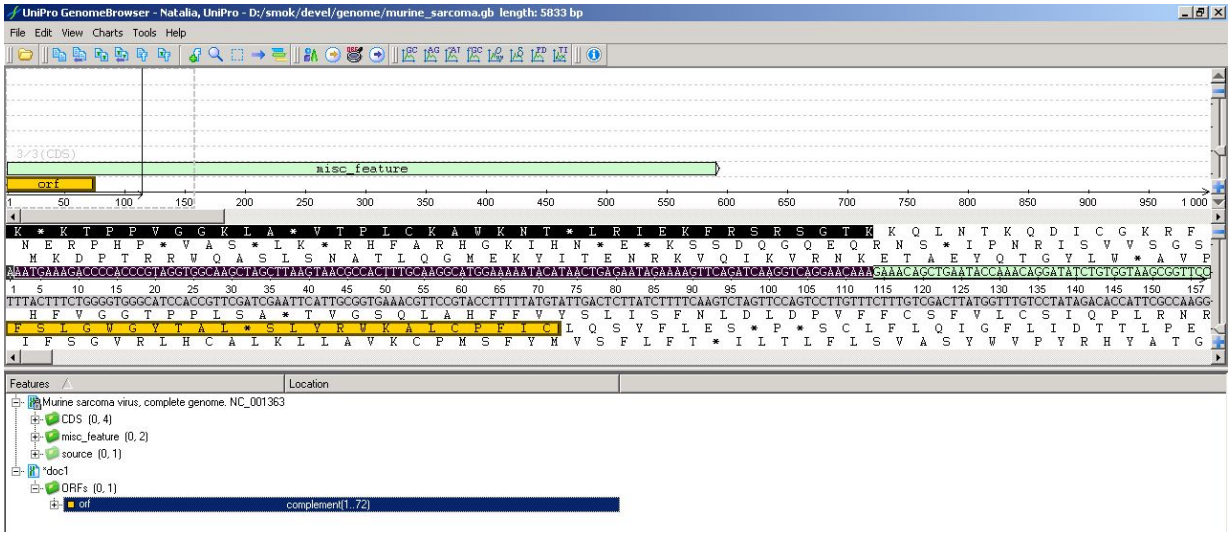


Рисунок 17. Результат работы утилиты ORF Marker.

2.2.3. OperonSet - плагин для реконструкции оперонной структуры геномов прокариотических организмов.

Оперон – это совокупность нескольких близко расположенных друг к другу открытых рамок считывания, с которых считывается единая матричная РНК с одного и того же (обычно) промотора. Алгоритм работы программы основывается на анализе расстояний между генами (кодирующими частями, открытыми рамками считывания), их ориентации, а также информации о том, в каком виде ДНК присутствует в данном организме (линейная либо кольцевая). В ходе работы программы гены (кодирующие части, открытые рамки считывания), расстояние между которыми меньше либо равно заданного значения (в п.н.) вносятся в одну оперонную структуру. Таким образом, результатом работы программы является определение оперонной структуры геномов прокариотических организмов.

Для работы программы **OperonSet** необходим файл, содержащий нуклеотидную последовательность и ее первичную аннотацию в формате GenBank (gbk формат).

Программа исполняется в средах операционных систем Linux и Windows.

После запуска **OperonSet** в **GenomeBrowser** появится диалоговое окно, в котором пользователь может указать: расстояний между генами (кодирующими частями, открытыми рамками считывания). В ходе работы программы гены (кодирующие части, открытые рамки считывания), расстояние между которыми меньше либо равно заданного значения (в п.н.) вносятся в одну оперонную структуру. Результатом работы программы является список найденных оперонов в *панели аннотаций* и картина их локализации в окне GenomeBrowser.

2.2.4. ArgoViewer - плагин для поиска промоторов в бактериальных генах.

Промоторные районы генов являются одним из наиболее важных регуляторных элементов, определяющих уровень экспрессии генов. Сборка базального транскрипционного комплекса и особенности транскрипции прокариотических генов зависят от контекстной и структурной организации корового промотора и присутствия в 5-регуляторном районе гена сайтов связывания транскрипционных факторов (ССТФ).

Большинство имеющихся подходов к распознаванию промоторов основано на выявлении потенциальных ССТФ с помощью весовых матриц, консенсусов или других методов.

Однако, несмотря на разнообразие предложенных подходов, распознавание промоторов на основе анализа ССТФ не обеспечивает необходимой точности поиска этих регуляторных районов. По-видимому, это объясняется тем, что экспериментально доступная информация о сайтах связывания транскрипционных факторов не описывает всего разнообразия ССТФ, функционирующих в составе промоторов.

Следует указать на существенную причину, затрудняющую создание точных методов распознавания промоторов - огромное разнообразие вариантов их структурно-функциональной организации. Это обстоятельство затрудняет поиск общих контекстных закономерностей, на основе которых может осуществляться распознавание промоторов.

Для решения задачи распознавания промоторов в геномах бактерий используется плагин **ArgoViewer**, позволяющий проводить распознавание регуляторных районов генов на основе олигонуклеотидных мотивов. Метод основан на сравнении представленности и характера распределения мотивов в рассматриваемой последовательности и последовательностях регуляторных районов генов. Анализируемая последовательность должна быть считана в систему Genome Browser согласно пункту 4.2.2.1. руководства. Распознавание промоторов может проводиться в последовательности ДНК длиной не менее 100 и не более 10 000 000 нуклеотидов.

После запуска **ArgoViewer** в **GenomeBrowser** появится диалоговое окно (Рис. 18), в котором пользователь может указать: необходимый для распознавания тип промоторов (выпадающий список *Site*), путь к файлу с профилем (поле *Profile* и кнопка *Browse*), объединять ли смежные точки профиля, попавшие в срез (пометка *Join adjacent points*), порог ошибки (таблица *Error level*), а также документ и группу, в которые будут сохраняться аннотации (поля *Document* и *Group* соответственно).

После нажатия кнопки *Create* из файла, указанного в поле *Profile*, будет загружен профиль (если файла не существует, то профиль будет построен и сохранен в этот файл). Тот же файл с профилем может быть использован позже для разметки этой последовательности сайтами этого типа с другим порогом. Далее программа выберет те точки профиля, значение которых больше либо равно указанному пороговому значению (столбец *Threshold* в таблице *Error level*). Чем выше порог, тем выше достоверность найденных промоторов, тем меньше потенциальных сайтов будет обнаружено. Соответсвующие цифры в строках соответствуют стандартным характеристикам качества распознавания - уровням недопредсказания (столбец I type) и перепредсказания (столбец II type).

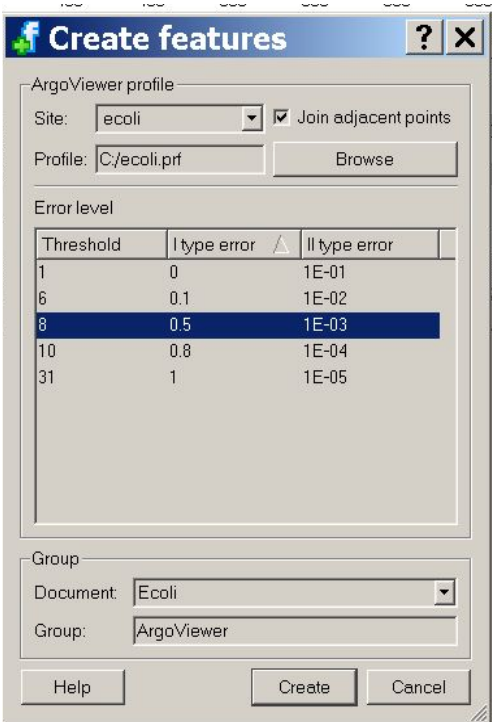


Рисунок 18. Диалоговое окно параметров ArgoViewer.

Если пометка *Join adjacent points* не была отмечена, каждая точка профиля, попавшая в срез, будет представлена как аннотация со свойствами *score* (значение профиля в этой точке) и *transcription* (нуклеотид соответствующий этой точке). Если же отметить пункт *Join adjacent points*, то программа будет объединять соседние точки среза и сохранять их как аннотации со свойствами: *max_score* (максимальное значение профиля на объединенном участке среза), *max_score_position* (позиция точки с максимальным значением) и *transcription* (последовательность нуклеотидов, попавших в данный участок среза).
Результатом работы программы является картина локализации предсказанных промоторов в окне GenomeBrowser (Рис. 19).

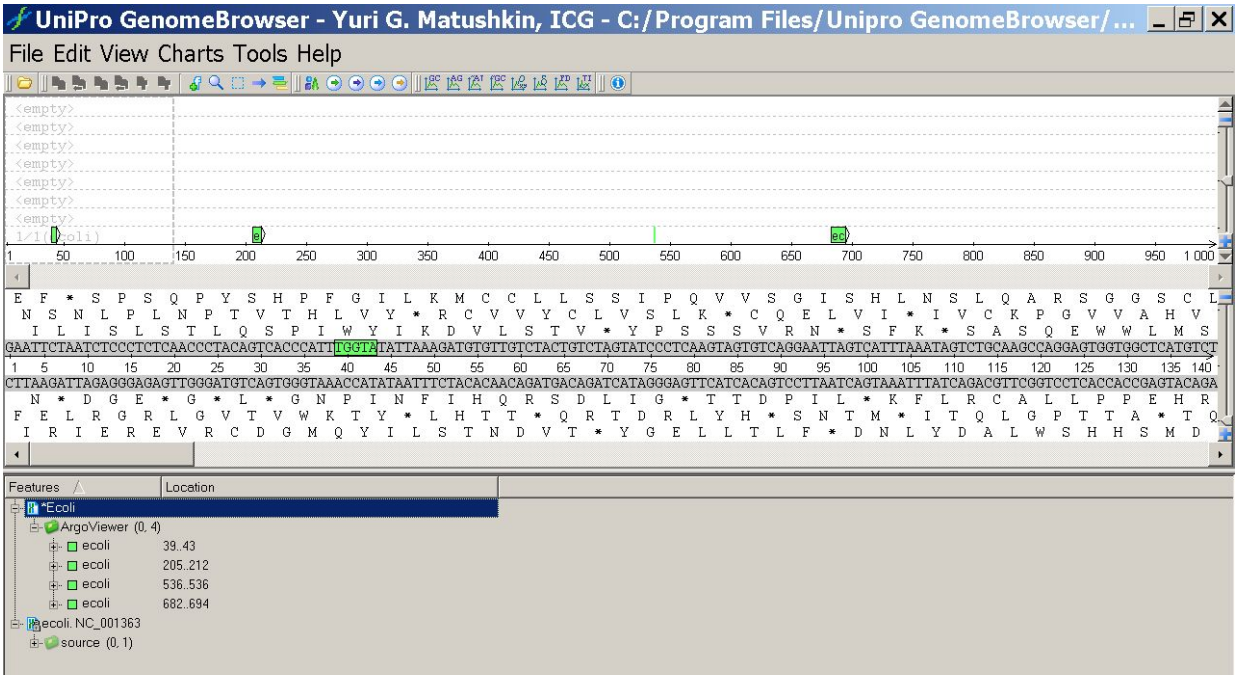


Рисунок 19. Локализация предсказанных промоторов в окне GenomeBrowser

2.2.5. SITECON - плагин для распознавания потенциальных сайтов связывания транскрипционных факторов.

Как правило, статистический анализ выборок сайтов связывания транскрипционных факторов позволяет выявлять общие контекстные характеристики, используемые для распознавания потенциальных сайтов. Однако данные о контекстно–зависимых конформационных и физико–химических свойствах может также эффективно быть использованы как для анализа, так и для распознавания сайтов связывания транскрипционных факторов, поскольку локальная конформация сайтов играет роль в узнавании транскрипционным фактором сайта связывания (Starr et al., 1995). Впервые зависимость конформации ДНК от контекста была обнаружена Dickerson and Drew (1981), путем рентгеноструктурного анализа додекамеров ДНК. Растущее число данных структурного анализа показало неоднородность конформационных и физико–химических свойств и их зависимость от нуклеотидной последовательности (Frank et al., 1997, Suzuki et al., 1997).

Для ряда позиций в выравнивании выборки сайтов связывания транскрипционного фактора можно выделить набор контекстно - зависимых конформационных и физико–химических свойств, которые остаются константными для всех вариантов сайтов выборки, несмотря на различие контекста. Этот факт, прежде всего, является следствием того, что определяемая контекстом локальная конформация молекул ДНК является одним из факторов специфичности ДНК – белкового узнавания (Meierhans et al., 1997). Иными словами, для успешного связывания DNA сайта с конкретным белковым фактором ряд свойств в конкретных участках сайта должны иметь определенные фиксированные значения. Определяется это, прежде всего, специфичностью механизмов ДНК–белковых взаимодействий (Рис. 20) для конкретного комплекса ДНК- белок (Oshchepkov et al., 2004). Именно эти свойства будут иметь близкие значения во всех вариантах различных по контексту сайтов выборки, и тем самым, анализ, направленный на поиск вариаций свойств в сайтах выявит их малую дисперсию в конкретных позициях выборки. Полный набор данных

о консервативных конформационных и физико-химических свойствах сайтов тем самым отражает специфичность взаимодействия ДНК последовательности с конкретным белком, и может быть эффективно использован для распознавания потенциальных сайтов связывания.

Программа SITECON использует подход для распознавания потенциальных сайтов связывания транскрипционных факторов, основанный на выявлении набора консервативных контекстно–зависимых конформационных и физико–химических свойств, определенных для коротких участков выравнивания функциональных последовательностей ДНК и дальнейшего сравнения выявленных консервативных свойств со свойствами анализируемой последовательности.

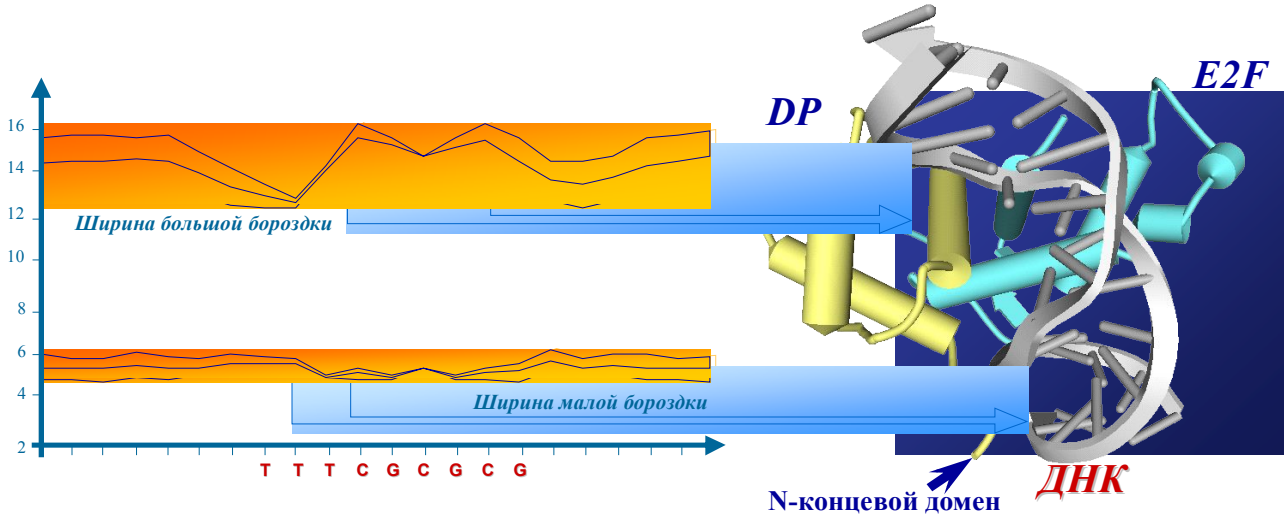


Рис 20. Выявленные SITECON значимо консервативные значения размеров бороздок в комплексе гетеродимера E2F/DP с ДНК согласуются с данными о механизме связывания, полученными с помощью рентгеноструктурного анализа ДНК-белкового комплекса. Ширины большой и малой бороздок увеличены в местах контактов с распознающими альфа-спиралями и N-концевым доменом соответственно.

Интерфейсы всех разработанных методов распознавания являются универсальными и однотипными. Рассмотрим решение задачи распознавания сайта связывания Frur в протяженной геномной последовательности на примере программы SITECON.

После запуска плагина **SITECON** в **GenomeBrowser** появится диалоговое окно (Рис. 17), в котором пользователь может указать: необходимый для распознавания тип сайтов связывания транскрипционных факторов (выпадающий список *Site*), путь к файлу с профилем (поле *Profile* и кнопка *Browse*), обединять ли смежные точки профиля, попавшие в срез (пометка *Join adjacent points*), порог ошибки (таблица *Error level*), а также документ и группу, в которые будут сохраняться аннотации (поля *Document* и *Group* соответственно).

После нажатия кнопки *Create* из файла, указанного в поле *Profile*, будет загружен профиль (если файла не существует, то профиль будет построен и сохранен в этот файл). Тот же файл с профилем может быть использован позже для разметки этой последовательности сайтами этого типа с другим порогом. Далее программа выберет те точки профиля, значение которых больше либо равно указанному пороговому значению (столбец *Threshold* в таблице *Error level*). Чем выше порог, тем выше достоверность найденных сайтов, тем меньше потенциальных сайтов будет обнаружено. Соответствующие цифры в строках соответствуют стандартным характеристикам качества распознавания - уровням недопредсказания (столбец I type) и перепредсказания (столбец II type).

Если пометка *Join adjacent points* не была отмечена, каждая точка профиля, попавшая в срез, будет представлена как аннотация со свойствами *score* (значение профиля в этой точке) и *transcription* (нуклеотид соответствующий этой точке). Если же отметить пункт *Join*

adjacent points, то программа будет объединять соседние точки среза и сохранять их как аннотации со свойствами: *max_score* (максимальное значение профиля на объединенном участке среза), *max_score_position* (позиция точки с максимальным значением) и *transcription* (последовательность нуклеотидов, попавших в данный участок среза).

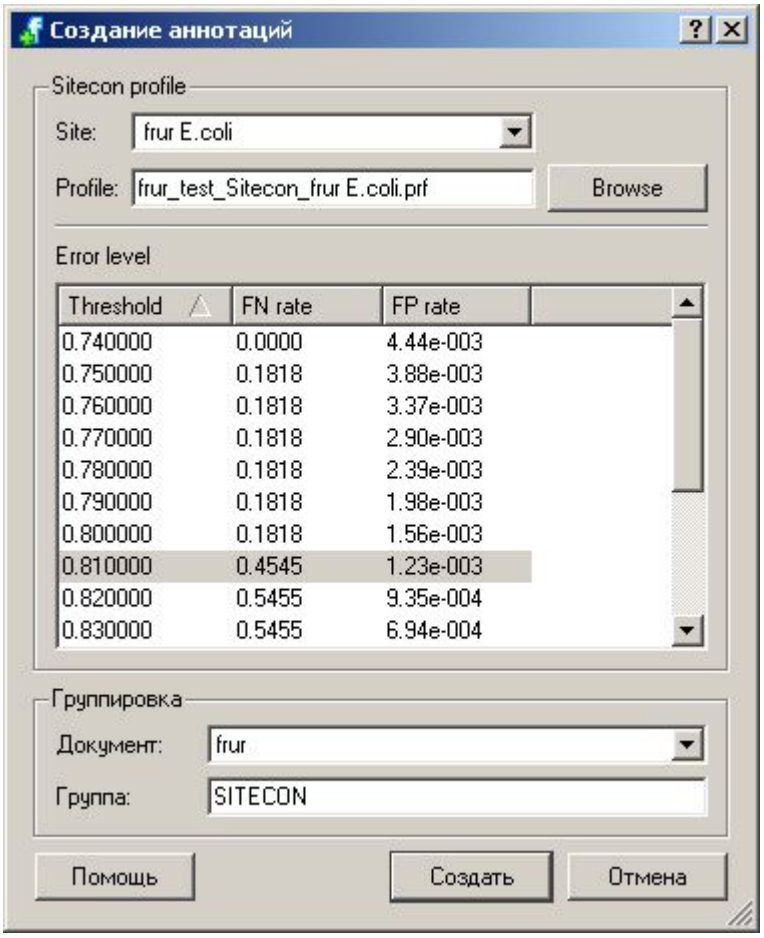


Рисунок 21. Диалоговое окно параметров Sitecon

Результатом работы программы является картина локализации предсказанных сайтов связывания транскрипционных факторов в окне GenomeBrowser (Рис. 22).

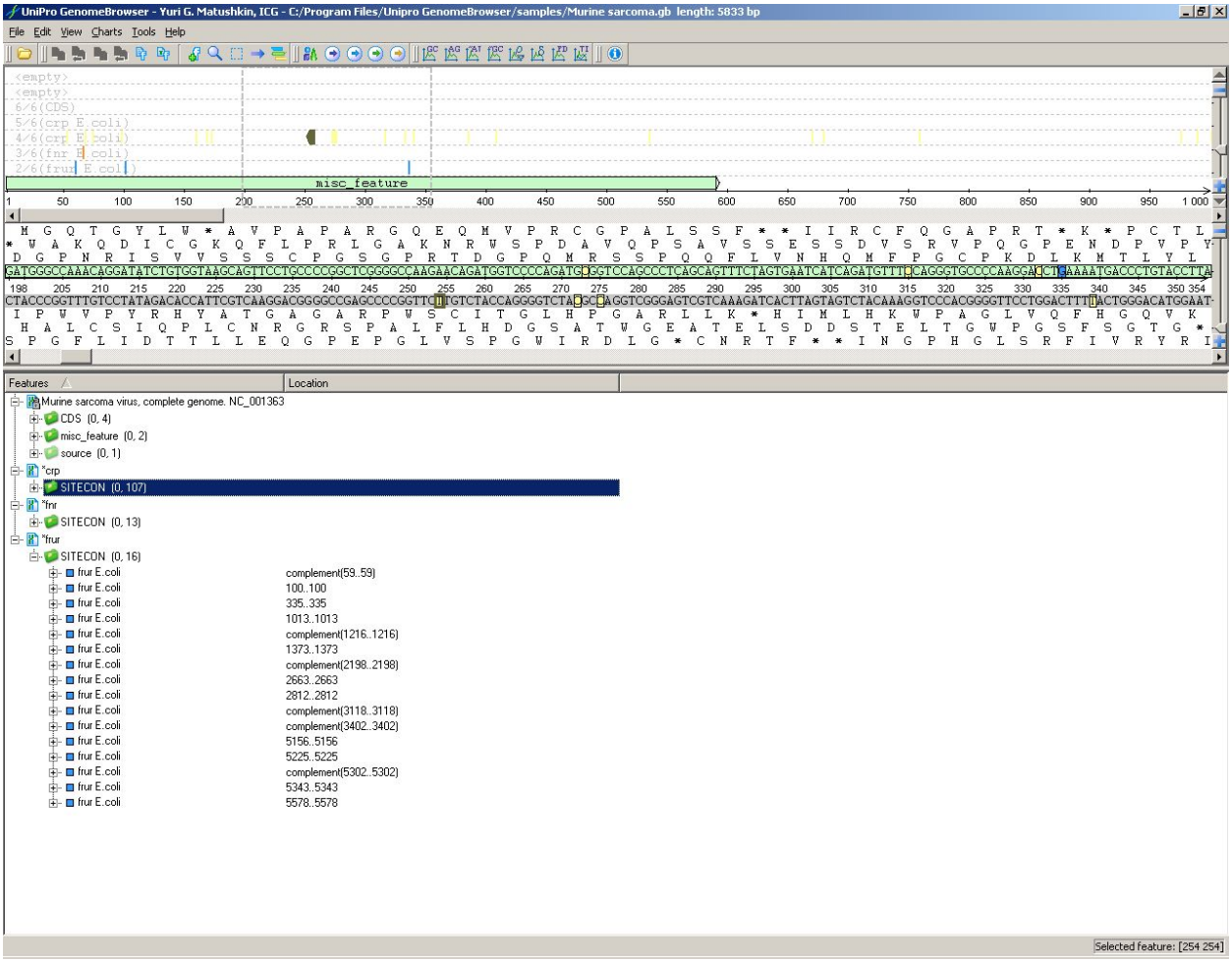
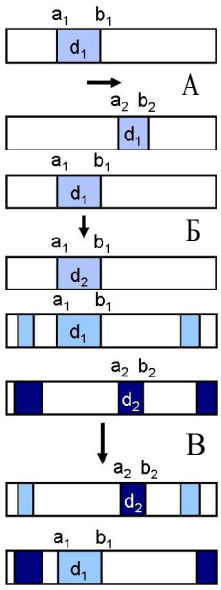


Рисунок 22. Результат работы программы распознавания сайтов связывания транскрипционных факторов Sitecon.

2.2.6. SiteGA плагин для распознавания сайтов связывания транскрипционных факторов.

Транскрипция генов эукариот является важнейшим этапом регуляции их экспрессии. Важную роль в контроле транскрипции эукариот играют 5'-регуляторные районы генов. Эти районы имеют блочно-иерархическую организацию, при этом в качестве отдельных структурных элементов выступают сайты связывания транскрипционных факторов (ССТФ). Обычно каждый ССТФ представляется небольшим (~10-15 п.о.) участком ДНК, необходимым для ее взаимодействия с ТФ. Разработка методов распознавания функциональных ССТФ очень важна для аннотации генов эукариот.

Метод распознавания ССТФ SiteGA представляет собой поиск набора локально-позиционированных динуклеотидов (ЛПД). Задачу поиска набора ЛПД решает генетический алгоритм (ГА), использующий популяцию особей, представляющих собой наборы ЛПД. Каждый ЛПД особи характеризуется положением в пределах рассматриваемого района ДНК, а также типом динуклеотида (АА, ТТ, и т.д.). Работа ГА начинается с того, что для каждой особи популяции случайным образом задаются типы и положения всех ЛПД. Затем ГА итеративно производит циклы мутаций и рекомбинаций. Мутация меняет положение (рис. 23А) или тип (рис. 23Б) одного ЛПД особи. Рекомбинация осуществляет обмен ЛПД между двумя разными особями (рис. 23В).



А) мутация «изменение положения ЛПД» (ЛПД выделен цветом). $[a_1, b_1]$ – начальное положение ЛПД, $[a_2, b_2]$ – конечное положение ЛПД, d_1 –тип динуклеотида, который не изменяется;

Б) мутация «изменение типа ЛПД» (ЛПД выделен цветом). d_1 – начальный тип ЛПД, d_2 – конечный тип ЛПД, $[a_1, b_1]$ – положение ЛПД, которое не изменяется;

В) рекомбинация между двумя особями «обмен отдельными ЛПД». 1 и 2 – родительские особи (ЛПД отмечены разным цветом), 1' и 2' – дочерние особи. Для участвующих в обмене ЛПД отмечены положения $[a_1, b_1]$ и $[a_2, b_2]$, а также типы d_1 и d_2.

Рисунок 23. Примеры элементарных операций генетического алгоритма

В ГА максимизируемым параметром приспособленности особи является расстояние R^2 Махаланобиса, рассчитываемое по частотам ЛПД. По полученному набору ЛПД с помощью методики дискриминантного анализа строится функция распознавания ССТФ.

Плагин SiteGA имеет такой же интерфейс, как и плагин SITECON.

2.2.7. CONSENSUS – Плагин для поиска сайтов связывания транскрипционных факторов (TF-сайтов) в бактериальных генах.

Распознавание функциональных сайтов является ключевым эпизодом аннотации геномных ДНК (Fickett, 1996; Hausler, 1998). МакКлур (Hawley and McClure, 1983) впервые ввел, на примере "-35" и "-10" боксов промоторов E.coli, математическую модель для прогноза качественного влияния замен нуклеотидов на биологические свойства функциональных сайтов S длины L, консенсус:

$$\text{Score}(S=\{s_1,s_2,\ldots,s_j,\ldots s_{L-1},s_L\}) = \Sigma_{1\le i\le K}\,\Sigma_{1\le j\le L}\, w_{i,j}\times\delta(s_j=e_i), \quad (1)$$

$$w_{i,j}= \begin{cases} 1, \text{ если } N_{i,j} \ge N_0 + \mu\times\sigma(N_{i,j}); \\ 0, \text{ если иначе;} \end{cases} \quad (2)$$

$$[\ \text{Score}(S^{\#}) \ge \Omega\ ] \Rightarrow \{\ S^{\#} - \text{результат предсказания "потенциальный сайт"}\}, \quad (3)$$

здесь: δ(истина)=1, δ(ложь)=0; $w_{i,j}$– так называемые «позиционный вес» i-ого мономера e_i алфавита $\{e_1, e_2, \ldots, e_i, \ldots, e_K\}$ объема K в j-ой позиции сайта; N – количество известных вариантов сайта; $N_{i,j}$ – число вариантов сайта, где i-ый мономер e_i алфавита занимает j-ую позицию ($s_j=e_i$); $N_0=N/K$ – оценка априорно ожидаемого значения $N_{i,j}$ при случайной, равномерной и независимой встречаемости любого мономера в любой позиции сайта; $\sigma(N_{i,j})$ – стандартное отклонение $N_{i,j}$ для N известных вариантов сайта; μ - эмпирический множитель, чьи значения были установлены с помощью распределения Пуассона равными 3 и 6 соответственно для умеренно- и сильно-консервативных мономеров в позициях сайтов; Ω - так называемое "пороговое значение предсказывающего правила (формула 3)".

В тоже время является хорошо известным, что многие функциональные сайты ДНК и РНК характеризуются специфическим олигонуклеотидным составом. В частности, акцепторные сайты сплайсинга имеют так называемый «полипиримидиновый тракт» (Kel et al., 1994), в состав сайта 3'-концевого разрезания-полиаденилирования мРНК входит так называемый «G/U-богатый элемент» (McDevitt et al., 1986), эукариотический TATA-бокс часто погружен в "CpG-остров" (Bucher, 1990). Однако, до сих пор отсутствуют компьютерные методы выявления и анализа сайт-специфического олигонуклеотидного состава для последующего применения этих закономерностей к распознаванию сайтов.

Поэтому мы обобщили формулу (1) на случай сайт-специфического олигонуклеотидного состава ДНК. Идея нашего обобщения состоит в применении указанных формул не только к алфавиту канонических нуклеотидов $E_1^0=\{e_1^0=A, e_2^0=T, e_3^0=G, e_4^0=C\}$ (IUPAC-IUB, 1971), но и, по аналогии, к "алфавитам олигонуклеотидов".

В Табл.1 определены 26 олигонуклеотидных алфавитов, которые исследованы ниже. Например, алфавит динуклеотидов $E_2^{W,S}=\{e_1^{W,S}=WW, e^{W,S}{}_2=WS, e^{W,S}{}_3=SW, e^{W,S}{}_4=SS$; где W="A или T", S="G или C"} мог бы, например, акцентировать внимание на учете термодинамических свойств ДНК-дуплекса, специфических для исследуемого сайта.

В свою очередь, другой алфавит динуклеотидов $E_2^{R,Y}=\{e^{R,Y}{}_1=RR, e^{R,Y}{}_2=RY, e^{R,Y}{}_3=YR, e^{R,Y}{}_4=YY$; где R="A или G", Y="T или C"} позволил бы, в соответствие с правилом Колладина-Диккерсона (Calladine, 1982), сфокусировать анализ сайта на свойствах конформации его B-спирали ДНК.

При анализе каждый сайт был представлен выборкой $\{S_1...S_n...S_N\}$ из N нуклеотидных последовательностей $S_n=s_{1,n}...s_{i,n}...s_{L,n}$ длины L, которые были известны из экспериментов (здесь, $s\in\{A, T, G, C\}$). При этом для всех последовательностей ДНК каждого сайта было осуществлено множественное выравнивание с помощью стандартного метода потенциала Гиббса (Lawrence, 1994). Затем рассматривался алфавит $E_m^{\#}=\{Z_1, ..., Z_j, ..., Z_K\}$из K олигонуклеотидов $Z_j=\{z_{1,j}z_{2,j}...z_{m,j}\}$ длины m (здесь, $z\in\{A, T, G, C, W=(A, T), S=(G, C), R=(A, G), Y=(T, C), M=(A, C), K=(T, G), N=(A, T, G, C)\}$ (IUPAC-IUB, 1971)).

	Код	Алфавит $E_m^{\#}=\{Z_1, ..., Z_j, ..., Z_K\}$ из K олигонуклеотидов типа # длины m	K
1	E_1^0	A, T, G, C (IUPAC-IUB, 1971)	4
Обобщение алфавита E_1^0 на специфичность ближайших соседей (Kondrakhin et al., 1994)			
2	E_2^0	AA, AT, AG, AC, TA, TT, TG, TC, GA, GT, GG., GC, CA, CT, CG, CC	16
3	E_3^0	AAA, AAT, AAG, AAC, ATA, ..., CGG, CGC, CCA, CCT, CCG, CCC	64
Обобщение алфавитов E_2^0 и E_3^0 на чередование нуклеотидов (настоящая работа)			
4	$E_3^{A,T,G,C,N}$	ANA, ANT, ANG, ANC, TNA, TNT, ..., GNC, CNA, CNT, CNG, CNC	16
5	$E_5^{A,T,G,C,N}$	ANANA, ANANT, ANANG, ..., CNCNA, CNCNT, CNCNG, CNCNC	64
Обобщение алфавитов 1-5 на термодинамические свойства ДНК (настоящая работа)			
6	$E_1^{W,S}$	W, S	2
7	$E_2^{W,S}$	WW, WS, SW, SS	4
8	$E_2^{W,S}$	WWW, WWS, WSW, WSS, SWW, SWS, SSW, SSS	8
9	$E_4^{W,S}$	WWWW, WWWS, WWSW, WWSS, ..., SSWW, SSWS, SSSW, SSSS	16
10	$E_3^{W,S,N}$	WNW, WNS, WNS, SNS	4
11	$E_5^{W,S,N}$	WNWNW, WNWNS, WNSNW, ..., SNWNS, SNSNW, SNSNS	8
12	$E_7^{W,S,N}$	WNWNWNW, WNWNWNS, WNWNSNW, ..., SNSNSNW, SNSNSNS	16
Обобщение алфавитов 1-5 на конформационные свойства ДНК (настоящая работа)			
13	$E_1^{R,Y}$	R, Y	2
14	$E_2^{R,Y}$	RR, RY, YR, YY	4
15	$E_3^{R,Y}$	RRR, RRY, RYR, RYY, YRR, YRY, YYR, YYY	8
16	$E_4^{R,Y}$	RRRR, RRRY, RRYR, RRYY, ..., YYRR, YYRY, YYYR, YYYY	16
17	$E_3^{R,Y,N}$	RNR, RNY, YNR, YNY	4
18	$E_5^{R,Y,N}$	RNRNR, RNRNY, RNYNR, ..., YNRNY, YNYNR, YNYNY	8
19	$E_7^{R,Y,N}$	RNRNRNR, RNRNRNY, RNRNYNR, ..., YNYNYNR, YNYNYNY	16
Обобщение алфавитов 1-5 на электростатические свойства ДНК (настоящая работа)			
20	$E_1^{M,K}$	M, K	2
21	$E_2^{M,K}$	MM, MK, KM, KK	4
22	$E_3^{M,K}$	MMM, MMK, MKM, MKK, KMM, KMK, KKM, KKK	8
23	$E_4^{M,K}$	MMMM, MMMK, MMKM, MMKK,..., KKMM, KKMK, KKKM, KKKK	16
24	$E_3^{M,K,N}$	MNM, MNK, KNM, KNK	4
25	$E_5^{M,K,N}$	MNMNM, MNMNK, MNKNM, ..., KNMNK, KNKNM, KNKNK	8
26	$E_7^{M,K,N}$	MNMNMNM, MNMNMNK, ..., KNKNKNM, KNKNKNK	16

Таблица 1. Олигонуклеотидные алфавиты, исследованные в настоящей работе.

Обозначения: M=”A или C”, K=”G или T”, R=”A или G”, Y=”T или C”, W=”A или T”, S=”G или C”, N=”A или T или G или C” (IUPAC-IUB, 1971).

В этих обозначениях, число вариантов сайта, где i-ый олигонуклеотид Z_i длины m из алфавита $E_m^{\#}=\{Z_1, ..., Z_j, ..., Z_K\}$ объема K занимает j-ую позицию (2) обобщалось:

$$N_{i,j} = \Sigma_{1\le n\le N} \Pi_{1\le k\le m} \delta(s_{j+k-1,n} \in z_{k,j}). \quad (4)$$

Поскольку новые алфавиты из Табл.1 содержат разное количество олигонуклеотидов, то формула (4) применялась к алфавитам №№ 1, 6, 7, 10, 13, 14, 17, 20, 21 и 24 при N≥8; к алфавитам №№ 8, 11, 15, 18, 22 и 25 при N≥25; к алфавитам №№ 2, 4, 9, 12, 16, 19, 23 и 26 при N≥65; и, наконец, к алфавитам №№ 3 и 5 при N≥200. Аналогично мы обобщили формулу (1), которая в этих новых обозначениях имеет следующий вид:

$$\text{Score}\ (S=\{s_1,s_2,\ldots,s_j,\ldots\ s_{L-1},s_L\}) = \Sigma_{1\le i\le K}\ \Sigma_{1\le j\le L-b+1}\ w_{i,j}\times\Pi_{1\le k\le m}\ \delta(s_{j+k-1,n}\in z_{k,j}), \quad (5)$$

здесь: $w_{i,j}$ - позиционные веса олигонуклеотидов (формула 2).

Для преодоления несоизмеримости мы сделали дополнительно еще один необходимый шаг обобщения консенсуса от случая традиционного алфавита канонических нуклеотидов E_1^0={A, T, G, C} (IUPAC-IUB, 1971) к исследуемому здесь случаю алфавита $E_m^{\#}=\{Z_1, ..., Z_j, ..., Z_K\}$ олигонуклеотидов определенного типа "#" и длины m. С этой целью мы определили единую шкалу измерения всех величин, вычисляемых по формуле (5) для исследемого сайта при любых алфавитах:

$$\left\{\left\{\left\{\begin{array}{l}\text{Score}^0\ (S) = (\text{Score}\ (S) - \Delta) / \gamma, \quad (6)\\ \Delta = (M^0_{S\in SITE}[\text{Score}(S)] + M^0_{S\in RAND}[\text{Score}(S)])/2;\\ \gamma = (M^0_{S\in SITE}[\text{Score}(S)] - M^0_{S\in RAND}[\text{Score}(S)])/2;\end{array}\right.\right.\right.$$

$$\text{Score}_\Sigma(S) = [\Sigma_{1\le h\le H}\ \text{Score}_h^0\ (S)\ ] / H; \quad (7)$$

$$[\ \text{Score}_\Sigma\ (S^{\#}) \ge 0\ ] \Rightarrow \{\ S^{\#} - \text{результат предсказания "потенциальный сайт"}\}, \quad (8)$$

здесь: Δ, γ - нормировочные коэффициенты, с помощью при которых средние значения сходства для всех известных сайтов, $M^0_{S\in SITE}[\text{Score}^0(S)]$, и для 10^3 случайных ДНК, $M^0_{S\in RAND}[\text{Score}^0(S)]$, устанавливаются равными соответственно "+1" и "-1"; H – количество используемых олигонуклеотидных алфавитов (Табл. 1); Ω=0 - порог "предсказывающего правила (формула 3)" в случае этой новой нормировки.

Формулы (6-8) основаны на использовании Математической теории принятия аддитивных решений (Fishburn, 1970). С помощью этих новых алфавитов для сайтов связывания ДНК с 8 белковыми факторами транскрипции мы построили консенсусы (формула 1-8), которые были интегрированы в программу CONSENSUS.

Установлено, что при распознавании сайтов, ошибка II рода консенсуса канонических нуклеотидов E_1^0={A, T, G, C}, уменьшалась в соответствие с Центральной предельной теоремой благодаря учету дополнительных олигонуклеотидных алфавитов.

Плагин **CONSENSUS** имеет такой же интерфейс, как и плагин **SITECON**.

2.2.8. MATRIX – плагин для поиска сайтов связывания транскрипционных факторов (TF-сайтов) в бактериальных генах.

Распознавание функциональных сайтов является ключевым эпизодом аннотации геномных ДНК (Fickett, 1996; Hausler, 1998). МакКлур (Hawley and McClure, 1983) впервые ввел, на примере "-35" и "-10" боксов промоторов E.coli, математическую модель для прогноза качественного влияния замен нуклеотидов на биологические свойства функциональных сайтов S длины L, консенсус:

$$\left\{\begin{array}{l}\text{Score}\ (S=\{s_1,s_2,\ldots,s_j,\ldots\ s_{L-1},s_L\}) = \Sigma_{1\le i\le K}\ \Sigma_{1\le j\le L}\ w_{i,j}\times\delta(s_j=e_i), \quad (1)\\ w_{i,j}= \left\{\begin{array}{l}1, \text{если } N_{i,j} \ge N_0 + \mu\times\sigma(N_{i,j}); \quad (2)\\ 0, \text{если иначе;}\end{array}\right.\\ [\ \text{Score}\ (S^{\#}) \ge \Omega\ ] \Rightarrow \{\ S^{\#} - \text{результат предсказания "потенциальный сайт"}\}, \quad (3)\end{array}\right.$$

здесь: δ(истина)=1, δ(ложь)=0; $w_{i,j}$– так называемые «позиционный вес» i-ого мономера e_i алфавита $\{e_1, e_2, \ldots, e_i, \ldots, e_K\}$ объема K в j-ой позиции сайта; N – количество известных вариантов сайта; $N_{i,j}$ – число вариантов сайта, где i-ый мономер e_i алфавита занимает j-ую позицию ($s_j=e_i$); $N_0=N/K$ – оценка априорно ожидаемого значения $N_{i,j}$ при случайной, равномерной и независимой встречаемости любого мономера в любой позиции сайта; $\sigma(N_{i,j})$ – стандартное отклонение $N_{i,j}$ для N известных вариантов сайта; μ - эмпирический множитель, чьи значения были установлены с помощью распределения Пуассона равными 3 и 6 соответственно для умеренно- и сильно-консервативных мономеров в позициях сайтов; Ω - так называемое "пороговое значение предсказывающего правила (формула 3)".

Затем, для более детального описания функциональных сайтов, Стаден (Staden, 1984) ввел матрицы $\{f_{i,j}\}_{1\le i\le K, 1\le j\le L}$ размером K*L оценок частот встречаемости $f_{i,j}$ каждого i-ого мономера e_i из алфавита $\{e_1, e_2, \ldots, e_i, \ldots, e_K\}$ в каждой j-ой позиции исследуемого сайта:

$$f_{i,j} = (N_{i,j} + 1) / (N + K), \qquad (4)$$

здесь, в качестве примера, приведена так называемая «байесовская» оценка таких частот.

Благодаря этой идее МакКлур и соавт. (Mulligan et al., 1984) усовершенствовали свою модель (формула 2) для прогноза уже не только качественных, но и количественных эффектов мутаций с помощью линейно-аддитивной аппроксимации "Homology Score":

$$w_{i,j}=f_{i,j}. \qquad (5)$$

Оказалось, что вычисленные с помощью этой самой простой формулы (5) величины Score(S) сайта S связывания промоторов E.coli с σ-субъединицей РНК-полимеразы ("-35" и "-10"-боксы), имеют статистически достоверную линейную корреляцию с величинами "селективности" РНК-полимеразы, измеренными в эксперименте (Mulligan et al., 1984).

В тоже время является хорошо известным, что многие функциональные сайты ДНК и РНК характеризуются специфическим олигонуклеотидным составом. В частности, акцепторные сайты сплайсинга имеют так называемый «полипиримидиновый тракт» (Kel

et al., 1994), в состав сайта 3'-концевого разрезания-полиаденилирования мРНК входит так называемый «G/U-богатый элемент» (McDevitt et al., 1986), эукариотический TATA-бокс часто погружен в "CpG-остров" (Bucher, 1990). Однако, до сих пор отсутствуют компьютерные методы выявления и анализа сайт-специфического олигонуклеотидного состава для последующего применения этих закономерностей к распознаванию сайтов.

Поэтому мы обобщили формулы (1-5) на случай сайт-специфического олигонуклеотидного состава ДНК. Идея нашего обобщения состоит в применении указанных формул не только к алфавиту канонических нуклеотидов $E_1{}^0=\{e_1{}^0=A, e_2{}^0=T, e_3{}^0=G, e_4{}^0=C\}$ (IUPAC-IUB, 1971), но и, по аналогии, к "алфавитам олигонуклеотидов".

В Табл.1 определены 26 олигонуклеотидных алфавитов. Например, алфавит динуклеотидов $E_2{}^{W,S}=\{e_1{}^{W,S}=WW, e^{W,S}{}_2=WS, e^{W,S}{}_3=SW, e^{W,S}{}_4=SS$; где W="A или T", S="G или C"$\}$ мог бы акцентировать внимание на учете термодинамических свойств ДНК-дуплекса, специфических для исследуемого сайта. В свою очередь, другой алфавит динуклеотидов $E_2{}^{R,Y}=\{e^{R,Y}{}_1=RR, e^{R,Y}{}_2=RY, e^{R,Y}{}_3=YR, e^{R,Y}{}_4=YY$; где R="A или G", Y="T или C"$\}$ позволил бы, в соответствие с правилом Колладина-Диккерсона (Calladine, 1982), сфокусировать анализ сайта на свойствах конформации его B-спирали ДНК.

При анализе каждый сайт был представлен выборкой $\{S_1 \ldots S_n \ldots S_N\}$ из N нуклеотидных последовательностей $S_n=s_{1,n}\ldots s_{i,n}\ldots s_{L,n}$ длины L, которые были известны из экспериментов (здесь, $s\in\{A, T, G, C\}$). При этом для всех последовательностей ДНК каждого сайта было осуществлено множественное выравнивание с помощью стандартного метода потенциала

Гиббса (Lawrence, 1994). Затем рассматривался алфавит $E_m^{\#}=\{Z_1, ..., Z_j, ..., Z_K\}$из K олигонуклеотидов $Z_j=\{z_{1,j}z_{2,j}...z_{m,j}\}$ длины m (здесь, $z\in\{$A, T, G, C, W=(A, T), S=(G, C), R=(A, G), Y=(T, C), M=(A, C), K=(T, G), N=(A, T, G, C)} (IUPAC-IUB, 1971)). В этих обозначениях, «байесовскую» оценку частот встречаемости (формулы 4 и 5) обобщалась:

$$w_{i,j}=f_{i,j}=(1+\Sigma_{1\le n\le N}\,\Pi_{1\le k\le m}\,\delta(s_{j+k-1,n}\in z_{k,j}))\,/\,(N+K). \qquad (6)$$

Поскольку новые алфавиты из Табл.1 содержат разное количество олигонуклеотидов, то формула (4) применялась к алфавитам №№ 1, 6, 7, 10, 13, 14, 17, 20, 21 и 24 при N≥8; к алфавитам №№ 8, 11, 15, 18, 22 и 25 при N≥25; к алфавитам №№ 2, 4, 9, 12, 16, 19, 23 и 26 при N≥65; и, наконец, к алфавитам №№ 3 и 5 при N≥200. Аналогично мы обобщили формулу (1), которая в этих новых обозначениях имеет следующий вид:

$$\text{Score}\,(S=\{s_1,s_2,\ldots,s_j,\ldots\,s_{L-1},s_L\})=\Sigma_{1\le i\le K}\,\Sigma_{1\le j\le L-b+1}\,w_{i,j}\times\Pi_{1\le k\le m}\,\delta(s_{j+k-1,n}\in z_{k,j}). \qquad (7)$$

Для преодоления несоизмеримости мы сделали дополнительно еще один необходимый шаг обобщения весовых матриц от случая традиционного алфавита канонических нуклеотидов $E_1^0=\{$A, T, G, C} (IUPAC-IUB, 1971) к исследуемому здесь случаю алфавита $E_m^{\#}=\{Z_1, ..., Z_j, ..., Z_K\}$ олигонуклеотидов определенного типа "#" и длины m. С этой целью мы определили единую шкалу измерения всех величин, вычисляемых по формуле (5) для исследемого сайта при любых алфавитах:

$$\left\{\begin{array}{l}\left\{\begin{array}{l}\left\{\begin{array}{l}\text{Score}^0(S)=(\text{Score}(S)-\Delta)/\gamma, \qquad (8)\\ \Delta=(M^0_{S\in SITE}[\text{Score}(S)]+M^0_{S\in RAND}[\text{Score}(S)])/2;\\ \gamma=(M^0_{S\in SITE}[\text{Score}(S)]-M^0_{S\in RAND}[\text{Score}(S)])/2;\end{array}\right.\\ \text{Score}_\Sigma(S)=[\Sigma_{1\le h\le H}\,\text{Score}_h^0(S)]\,/\,H; \qquad (9)\end{array}\right.\\ [\,\text{Score}_\Sigma(S^{\#})\ge 0\,]\Rightarrow\{\,S^{\#}-\text{результат предсказания "потенциальный сайт"}\}, \qquad (10)\end{array}\right.$$

здесь: Δ, γ - нормировочные коэффициенты, с помощью при которых средние значения сходства для всех известных сайтов, $M^0_{S\in SITE}[\text{Score}^0(S)]$, и для 10^3 случайных ДНК, $M^0_{S\in RAND}[\text{Score}^0(S)]$, устанавливаются равными соответственно "+1" и "-1"; H – количество используемых олигонуклеотидных алфавитов (Табл. 1); Ω=0 - порог "предсказывающего правила (формула 3)" в случае этой новой нормировки.

Формулы (8-10) основаны на использовании Математической теории принятия аддитивных решений (Fishburn, 1970). С помощью этих новых алфавитов для сайтов связывания ДНК с 8 белковыми факторами транскрипции мы построили позиционно-весовые матрицы (формула 1-10) и интегрировали их в программу MATRIX. Установлено, что при распознавании сайтов, ошибка II рода позиционно-весовых матриц канонических нуклеотидов $E_1^0=\{$A, T, G, C} уменьшалась в соответствие с Центральной предельной теоремой благодаря учету дополнительных олигонуклеотидных алфавитов.

Плагин **MATRIX** имеет такой же интерфейс, как и плагин **SITECON**.

2.2.9. ArgoMotifs – плагин для выявления консервативных мотивов в геномных последовательностях.

Выявление консервативных мотивов может проводиться в наборе из не более чем 500 анализируемых последовательностей ДНК одинаковой длины длиной не более 1000 нуклеотидов.

Для решения задачи выявления наборов вырожденных олигонуклеотидных мотивов, специфичных для регуляторных элементов генов (в том числе в 5' проксимальной области)

прокариот система **GenomeBrowser** использует плагин **ArgoMotifs**. Программа **ArgoMotifs** основана на методе кластеризации схожих совершенных олигонуклеотидов, входящих в состав различных регуляторных последовательностей с итерационным построением для каждого класса олигонуклеотидов результирующего консенсуса.

Анализируемая выборка последовательностей должна быть считана в систему Genome Browser согласно пункту 4.2.2.1 руководства пользователя.

После запуска **ArgoMotifs** в **GenomeBrowser** появится диалоговое окно, в котором пользователь может указать:

- начальная позиция анализируемого окна
- размер анализируемого окна
- длину рассматриваемого мотива;
- число разрешенных несовпадений (расстояние Хэмминга);
- долю присутствия мотива в выборке;
- значимость мотивов;
- лимит присутствия мотива в случайной выборке
- размер алфавита
- путь к файлу с таблицей выявленных мотивов (поле *Table* и кнопка *Browse*),
- а также документ и группу, в которые будут сохраняться аннотации (поля *Document* и *Group* соответственно).

После нажатия кнопки *Create* из файла, указанного в поле *Table*, будет загружена таблица (если файла не существует, то таблица будет построена и сохранена в эти файлы), содержащая список выявленных мотивов и их характеристики.

Результатом работы программы является список найденных мотивов и их характеристики (Рис. 24).

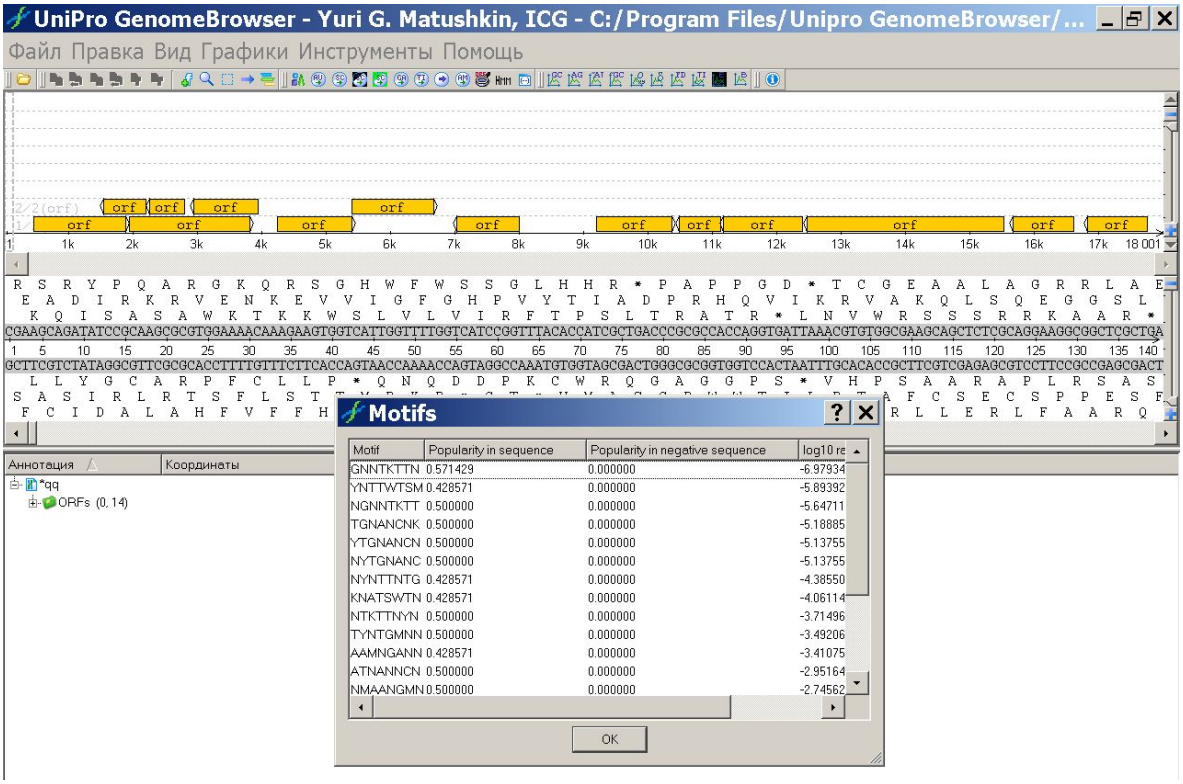


Рисунок 24. Результат работы программы поиска консервативных мотивов ArgoMotifs.

2.2.10. TransIndex – плагин для предсказания эффективности элонгации трансляции мРНК на основе их первичной структуры

Было показано, что индексы эффективности элонгации (EEI) хорошо кореллируют с уровнем экспрессии генов для большинства известных одноклеточных организмов, включая эукариотов, и корреляция имеет очень высокий уровень достоверности (Likhoshvai V.A., Matushkin Yu.G., 2002). Это объясняется тем, что при расчете EEI учитывается возможное образование вторичных структур на мРНК, которые играют важную роль в трансляции.

Для оценки корреляции 5 форм индекса EEI с эффективностью экспрессии гена использован тот факт, что в одноклеточных организмах рибосомные гены входят в список самых высоко экспрессирующихся, и по ним, как по маркерам, можно судить о том, насколько эффективно работает индекс.

В организме выделяется набор генов кодирующих белки, участвующих в сборке рибосом и непосредственно связанных с ними. Это делается с помощью специально созданной программы распознавания и разметки аннотированных рибосомных генов, по комбинациям ключевых слов в аннотации гена.

Для каждого из 5 форм индекса EEI вычисляются среднее и стандартное отклонение позиций рибосомных генов в ряду генов упорядоченных по увеличению эффективности элонгации (Рис. 25).

Пары среднее – стандартное отклонение характеризуют трансляционные свойства организма: (M_1,R_1), (M_2,R_2), (M_3,R_3), (M_4,R_4), (M_5,R_5).

Рисунок 25. Схема расположения рибосомных генов среди остальных генов организма. Гены упорядочены по убыванию индекса EEI(i), i=1,..,5.

Критерием того, что индекс адекватно оценивает эффективность элонгации, служит достоверно высокое значение M_i и/или достоверно низкое значение R_i. В первом случае рибосомные гены в среднем занимают высокое положение в иерархии уровней экспрессии среди других генов, во втором – их уровни экспрессии хорошо скоординированы друг с другом (Рис. 26).

Рисунок 26. Иллюстрация положения среднего и стандартного отклонения для рибосомных генов в случае их распознавания как высокоэкспрессирующихся.
Для каждого гена вычисляется его индекс EEI в 5 формах, и гены упорядочиваются по возрастанию индекса, служащего мерой эффективности трансляции.

Программа **TransIndex** позволяет качественно и количественно описать стратегию оптимизации трансляции, применяемую заданным одноклеточным организмом. Составляет список самых трансляционно-оптимальных генов организма, что позволяет предсказывать эффективность экспрессии генов. Необходимость в создании программы диктуется отсутствием в мире полноценных программ для оценки эффективности экспрессии генов на основе единого теоретически обоснованного подхода. Разработка такого комплекса необходима для теоретических исследований и практических биоинженерных задач.

TransIndex – программа для предсказания эффективности элонгации трансляции мРНК на основе их первичной структуры, классификацию бактериальных организмов по особенностям контекстной и структурной организации мРНК, значимым для эффективности трансляции.

Для предсказания эффективности элонгации трансляции мРНК и классификации бактериальных организмов используется программа **TransIndex**. Программа всем открытым рамкам считывания (ОРС) ставит в соответствие вещественное число от 0 до 10, которое является мерой эффективности трансляции. Кроме того, программа относит исследуемый геном к одной из 5-ти трансляционных групп, что позволяет оценивать роль кодонного состава и вторичных структур в элонгации трансляции.

Анализ может проводится для последовательности ДНК длиной не менее 1000 и не более 10 000 000 нуклеотидов, содержащей не менее 10 и не более 20000 ОРС.

Анализируемая последовательность должна быть считана в систему **Genome Browser** согласно пункту 4.2.2.1 руководства. После подключения плагина **TransIndex**, в *панели команд* появится дополнительная команда запуска плагина **TransIndex** и кнопка графика **TransIndex** (Рис. 27):

Рисунок 27.

После запуска **TransIndex** в GenomeBrowser появится окно создания аннотаций, в котором можно указать необходимые для создания аннотаций параметры (Рис. 28):

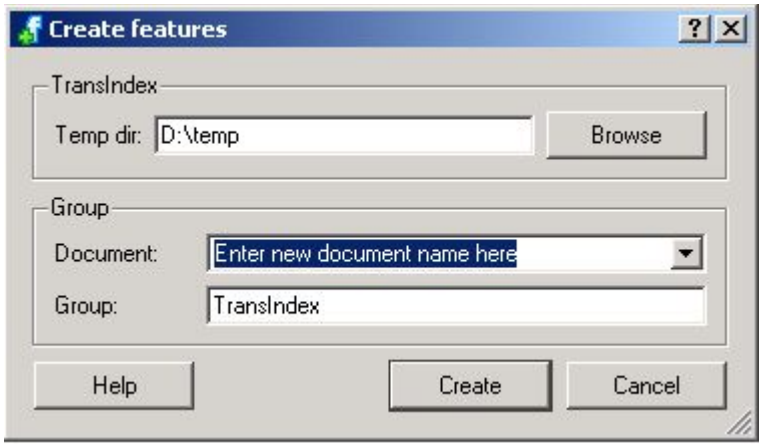


Рисунок 28.

В разделе *TransIndex* указывается временная директория, которая будет использована для создания временных файлов. Нажав на кнопку *Browse*, пользователь может с помощью стандартного диалогового окна выбрать директорию для хранения временных файлов. Полный путь выбранной директории отобразится в текстовом поле *Temp dir*. По умолчанию используется временная директория в домашнем каталоге пользователя.

В разделе *Group* пользователю предоставляется возможность указать имена создаваемого документа и группы аннотаций (текстовые поля *Document* и *Group* соответственно). Если документ с указанным в поле *Document* именем уже существует, указанная группа будет добавлена к существующему документу. Если группа с указанным в поле *Group* именем уже существует, новые аннотации будут добавлены к существующей группе.

По нажатию кнопки *Cancel*, создания аннотаций не происходит, окно создания аннотаций закрывается. По нажатию кнопки *Create* начинается работа **TransIndex**. Если в процессе работы произошла какая-либо ошибка, открывается диалоговое окно с сообщением об ошибке. После успешного завершения работы открывается диалоговое окно с информацией о результатах (Рис. 29):

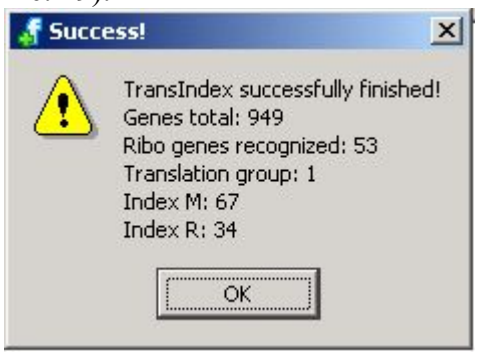


Рисунок 29.

Результатом работы программы является сводная таблица характеристик генома значимым для эффективности трансляции (Рис. 30) (данные сохраняются в аннотации *summary*), а также полный список аннотированных ОРС с соответствующими индексами EEI (аннотации *ORF*).

Рисунок 30. Сводная таблица характеристик генома значимым для эффективности трансляции.

Каждая аннотированная ОРС получает значение индекса EEI (аннотация *eei*, имеющая значение от 0 до 10.0), характеризующие эффективность ее элонгации.

Кнопка построения графика (TI) на панели графиков дает возможность построения ступенчатого графика EEI всех ОРС генома (Рис. 31).

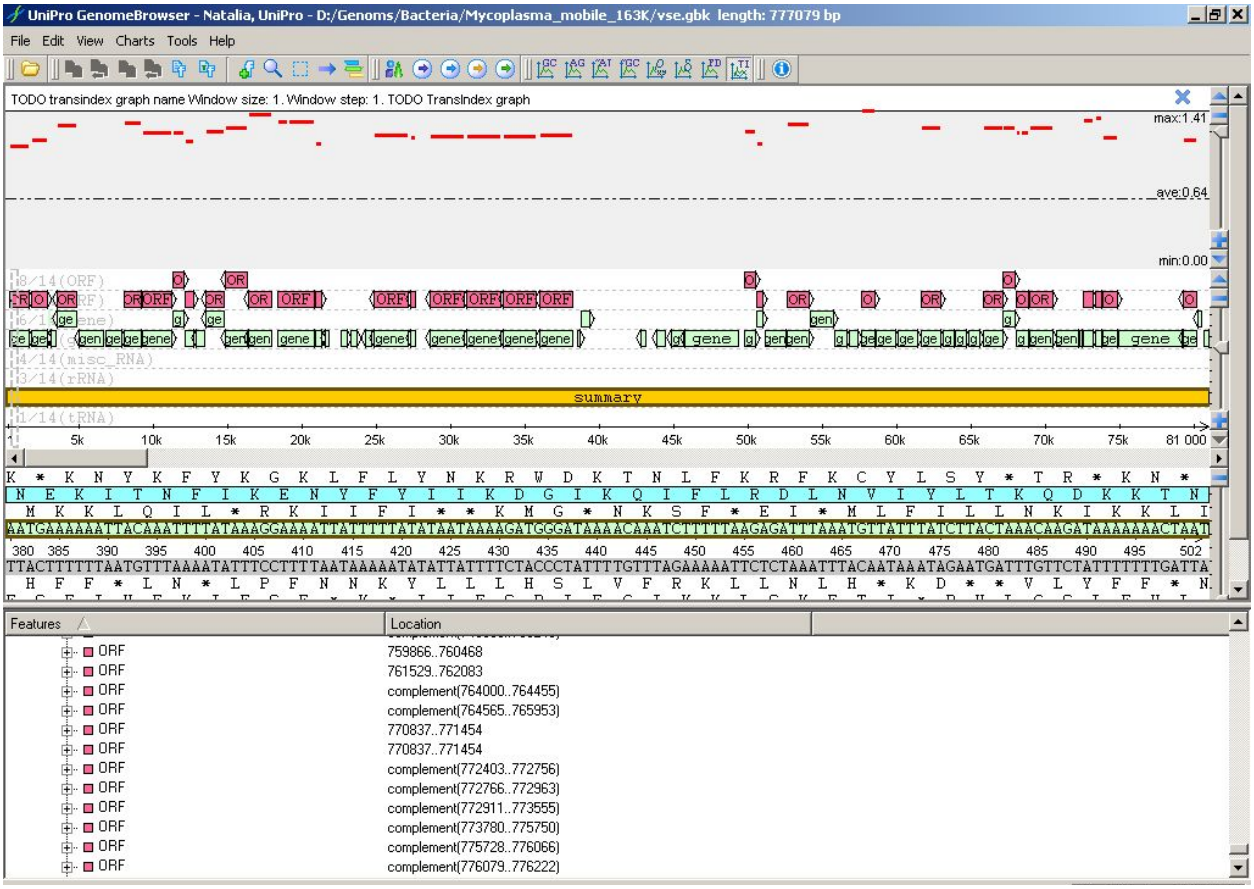


Рисунок 31. Общий вид окна с результатами работы TransIndex: график EEI (вверху), аннотации *ORF* (в центре и внизу).

2.2.11. Eagle – плагин для поиска повторов в заданной последовательности, строит профиль насыщенности повторами.

Программа выполняет поиск повторов между двумя геномными нуклеотидными последовательностями с заданной пользователем степенью вырожденности, оценивает статистические параметры повторов. Последовательность может быть одна, тогда анализируются повторы внутри нее.

Повтор – это полное (совершенный повтор) или частичное (несовершенный повтор) копирование последовательности нуклеотидов. В длинных последовательностях, тем более в целых геномах, повторы встречаются часто. Их изучение позволяет выделять все родственные гены и участки геномов (возможно некодирующие), минимальная длина и сходство которых (процент гомологии) определено пользователем.

Реализован алгоритм быстрого поиска совпадений с учетом *l*-грамм (коротких слов длины *l*). Нуклеотидные последовательности кодируются цифрами в 4-значном коде (по 4 нуклеотидам). По одной из последовательностей строится *l*-граммное разложение - для каждого слова выписываются в отдельную структуру (список) все позиции в которых оно встретилось. Вторая последовательность обрабатывается так - берется нуклеотидное слово, сравниваются все его вхождения в первую последовательность и оценивается длина совпадающего фрагмента, если она велика, повтор записывается в таблицу. За счет использования списка (*l*-граммного разложения) достигается линейная в зависимости от длины генома скорость работы алгоритма.

Затем таблица анализируется, и выбираются только непересекающиеся повторы максимальной длины.

После запуска **Eagle** в **GenomeBrowser** появится диалоговое окно, в котором пользователь может указать:

- необходимые для выявления типы повторов: прямой, обратный(симметричный), прямой комплементарный и обратный комплементарный (инвертированный), обозначенные соответственно D (direct), S (symmetric), C (complementary), I (inverted),
- путь к файлам с профилем и таблицей выявленных повторов (поля *Profile, Table* и кнопки *Browse*),
- минимальную длину выявляемых повторов, значение ограничено снизу размером 10 п.н. и со значением по умолчанию 20 п.н.
- а также документ и группу, в которые будут сохраняться аннотации (поля *Document* и *Group* соответственно).

После нажатия кнопки *Create* из файлов, указанных в полях *Profile и Table*, будут загружены профиль и таблица (если файлов не существует, то профиль и таблица будут построены и сохранены в эти файлы). Далее программа, используя аннотации из списка выявленных повторов и профиль, отмечает выявленные повторы и строит профиль плотности выявленных повторов

Результатом работы программы является картина локализации найденных повторов в окне GenomeBrowser и профиль плотности распределения выявленных повторов.

3. Полезные ссылки:

1. Bucher P. Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. // J. Mol. Biol. 1990. V. 212. P. 563-578.
2. Calladine C.R. Mechanics of sequence-dependent stacking of bases in B-DNA. // J. Mol. Biol. 1982. V.161. 343-352.
3. Dickerson T.D., Drew H.R. Structure of B-DNA dodecamer. II. Influence of base sequence on helix structure. // J Mol Biol. 1981. V. 149. P. 761–86.
4. Fickett J.W. Finding genes by computer: the state of the art. // Trends Genet. 1996. V. 12. P. 316-320.
5. Fishburn P.C. Utility theory for decision making. New York, Jonh Wiley & Sons. 1970.
6. Frank D.E., Saecker R.M., Bond J.P., Capp M.W., Tsodikov O.V., Melcher S.E., Levandoski M.M., Record M.T., Jr. Thermodynamics of the interactions of Lac repressor with variants of the symmetric Lac operator: effects of converting a consensus site to a non-specific site. // J Mol Biol. 1997. V. 267. P. 1186–206.
7. Haussler D. Computational genefinding. // Trends Guide in Bioinformatics, 1998. V.1, P. 12-15.
8. Hawley D.K. and McClure W.R. Compilation and analysis of Escherichia coli promoter DNA sequences. // Nucleic Acids Res. 1983. V.11. N. 8. P. :2237-2255.
9. IUPAC-IUB commission on biochemical nomenclature (CBN). Abbreviations and symbols for nucleic acids, polynucleotides and their constituents. // J. Mol. Biol. 1971. V. 55. N. 3. P. 299-310.
10. Kel A.E., Ponomarenko M.P., Likhachev E.A., Orlov Y.L., Ischenko I.V., Milanesi L., Kolchanov N.A. SITEVIDEO: a computer system for functional site analysis and recognition. Investigation of the human splice sites. // Comput. Appl. Biosci., 1993. V. 9, P. 617-627.

11. Kondrakhin, Y.V., Shamin, V.V., Kolchanov, N.A. Construction of a generalized consensus matriN for recognition of vertebrate pre-mRNA 3'-terminal processing sites. // Comput. Appl. Biosci. 1994. V. 10. P. 597-603.
12. Lawrence C. Toward the unification of sequence and structural data for identification of structural and functional constraints. // Comput. Chem. 1994. V. 18. P. 255-258.
13. Likhoshvai V.A., Matushkin Yu.G. 2002. Differentiation of single-cell organisms according to elongation stages crucial for gene expression efficacy. // *FEBS Lett.* V. **516** P. 87-92.
14. Meierhans D., Sieber M., Allemann R.K. High affinity binding of MEF-2C correlates with DNA bending. // Nucleic Acids Res. 1997. V. 25. P. 4537–44.
15. McDevitt M.A., Hart R.P., Wong W.W., Nevins J.R. Sequences capable of restoring poly(A) site function define two distinct downstream elements. // EMBO J. 1986. V.5. P. 2907-2913.
16. Oshchepkov D.Yu., Turnaev I.I., Pozdnyakov M.A., Milanesi L., Vityaev E.E., Kolchanov N.A. SITECON—A tool for analysis of DNA physicochemical and conformational properties: E2F/DP transcription factor binding site analysis and recognition. In N.Kolchanov and R.Hofestaedt (ed.) // *Bioinformatics of genome regulation and structure.* Kluwer Academic Publishers, Boston/Dordrecht/London. 2004. P. 93-102.
17. Starr D.B., Hoopes B.C., Hawley D.K. DNA bending is an important component of site-specific recognition by the TATA binding protein. // J Mol Biol. 1995. V. 250. P. 434–46.
18. Suzuki M., Amano N., Kakinuma J., Tateno M. Use of 3D structure data for understanding sequence-dependent conformational aspects of DNA. // J Mol Biol. 1997. V. 274. P. 421–35.
19. Mulligan M.E., Hawley D.K., Entriken, R., and McClure W.R. Escherichia coli promoter sequences predict in vitro RNA polymerase selectivity. // Nucleic Acids Res. 1984. V.12. P. 789-800.
20. Staden R. Method for calculating the probabilities of finding patterns in sequences. // Comput. App. Biosci. 1989. V.5. N.2, P. 89-96.