Подсистема «Регуломика»: Распознавание и анализ регуляторных геномных последовательностей эукариот

Структура документа (оглавление)

1. Цель и задачи подсистемы «Регуломика»..... 2
2. Использование методов и подходов биоинформатики в в исследовании регуляторных геномных последовательностей: структура подсистемы «Регуломика» и детальное руководство по ее применению..... 2
2.1. Информационные компоненты подсистемы «Регуломика»..... 3
Структурно-функциональная организация транскрипционных регуляторных районов..... 3
Описание работы с TRRD..... 4
Матрицы сайтов связывания транскрипционных факторов..... 8
Описание работы с БД ARTSITE •..... 10
Структурно-функциональная организация регуляторных последовательностей, контролирующих транскрипцию генов прокариот..... 12
Описание работы с ProkaTEX..... 13
2.2. Программные компоненты подсистемы «Регуломика»..... 20
2.2.1. Программные компоненты, позволяющие выявлять свойства регуляторных геномных последовательностей..... 20
Построение профилей конформационных и физико-химических свойств ДНК..... 20
Описание работы DNAProp..... 21
Выявление повторов в геномных последовательностях..... 24
Описание работы Eagle..... 24
Построение нуклеосомного потенциала последовательности ДНК..... 27
Описание работы Recon..... 28
2.2.3. Программные компоненты для анализа структурно-функциональной организации промоторов коэкспрессирующихся генов..... 30
Обнаружение иерархических комплексных сигналов, построение модели регуляторных районов..... 30
Описание работы ExpertDiscovery..... 32
Выявление консервативных мотивов..... 41
Описание работы ARGO_Motifs..... 41
2.2.4. Программные компоненты для •распознавания регуляторных последовательностей (сайтов связывания транскрипционных факторов и промоторов)..... 45
Распознавание сайтов связывания транскрипционных факторов на основе анализа консервативных конформационных и физико-химических свойств ДНК в выровненных последовательностях обучающей выборки..... 45
Описание работы SITECON..... 47
Распознавания сайтов связывания транскрипционных факторов на основе анализа взаимных зависимостей частот встреч локально-позиционированных динуклеотидов в ССТФ..... 49
Описание работы SiteGA..... 50
Программа CONSENSUS распознавания сайтов связывания транскрипционных факторов на основе поиска в исследуемой последовательности паттернов, сходных с паттернами в обучающих выборках..... 53
Описание работы CONSENSUS..... 56

Программа MATRIX распознавания сайтов связывания транскрипционных факторов (TF-сайтов) на основе весовых матриц..... 59
Описание работы MATRIX..... 62
Оценка количественной величины транскрипционной активности ДНК в логарифмических единицах свободной энергии Гиббса.....65
Описание работы ddG_MATRIX_INR.....68
Распознавания промоторов в геномных последовательностях методом ARGO..... 70
Описание работы ArgoViewer..... 71
3. Полезные ссылки.....75

1. Цель и задачи подсистемы «Регуломика»

Подсистема «Регуломика» предназначена для анализа структурно-функциональной организации протяженных транскрипционных регуляторных районов генов эукариот. Она позволяет: а) накапливать и анализировать экспериментальные данные о структурно-функциональной организации транскрипционных регуляторных районов генов эукариот; б) проводить в регуляторных районах поиск функциональных сигналов (ССТФ, комплексные сигналы и т.д.) и оценивать функционально значимые для регуляции экспрессии свойства (нуклеосомный потенциал, и т.д.); в) исследовать характеристики промоторов и других протяженных регуляторных районов (профили сложности, нуклеосомный потенциал, комплексные сигналы); г) осуществлять распознавание промоторов с использованием выявленных закономерностей организации.

Результаты, полученные с помощью программных компонент подсистемы «Регуломика», будут полезны для решения следующих задач: 1) исследования закономерностей структурно-функциональной организации транскрипционных регуляторных районов генов эукариот; 2) поиска новых генов-мишеней определенных транскрипционных факторов в геномах различных видов; 3) функциональной аннотации генома, включая выявление регуляторных сигналов (ССТФ, комплексных сигналов, построение профиля нуклеосомного потенциала), предсказание промоторов и их функциональная классификация.

2. Использование методов и подходов биоинформатики в в исследовании регуляторных геномных последовательностей: структура подсистемы «Регуломика» и детальное руководство по ее применению

Подсистема «Регуломика» представляет собой информационно-программный комплекс, предназначенный для функциональной аннотации транскрипционных регуляторных районов генов эукариот, а также исследования характеристик промоторов и их распознавания на основе выявленных закономерностей за счет решения следующих задач:

1) Накопление экспериментальных данных о структурно-функциональной организации транскрипционных регуляторных районов генов эукариот и сайтов связывания транскрипционных факторов;
2) Исследование конформационных и физико-химических свойств геномных последовательностей;
3) Выявление повторов и консервативных мотивов в регуляторных геномных последовательностях;
4) Распознавание сайтов связывания транскрипционных факторов в геномных последовательностях;
5) Распознавание промоторов в геномных последовательностях и выявление их принадлежности к определенному функциональному классу;

6) Разметка регуляторных последовательностей функциональными сигналами и функционально значимыми свойствами.

Ниже будут рассмотрены компоненты подсистемы «Регуломика» и приведено детальное (пошаговое) описание использования этих компонентов.

2.1. Информационные компоненты подсистемы «Регуломика».

Структурно-функциональная организация транскрипционных регуляторных районов

Транскрипция является ключевым событием, инициирующим экспрессию генов. В клетках эукариот транскрипция осуществляется тремя типами РНК-полимераз. РНК полимераза I транскрибирует гены 18S, 5.8S и 28S рибосомных РНК (рРНК). РНК полимераза III транскрибирует гены транспортных РНК (тРНК), 5S рРНК, и нескольких малых ядерных РНК (мяРНК). И, наконец, самый многочисленный класс РНК - мРНК, а также малые РНК (кроме U6), транскрибируются РНК полимеразой II. Синтез мРНК эукариот является строго регулируемым процессом. Значительная часть информации, определяющей специфическую программу управления транскрипцией отдельного гена, транскрибируемого РНК полимеразой II, кодируется в его регуляторных районах. Функциональность регуляторных районов определяется наличием в их составе коротких нуклеотидных последовательностей (сайтов), специфично взаимодействующими с регуляторными белками - транскрипционными факторами (ТФ). Уникальный набор сайтов связывания транскрипционных факторов (ССТФ) в регуляторных районах гена представляет собой программу, определяющую возможные паттерны его экспрессии. Реализуется эта программа при участии специфических для каждой клеточной ситуации наборов ТФ, взаимодействующих с ДНК в регуляторных единицах генов. К числу регуляторных единиц генов относятся промоторы, энхансеры, сайленсеры, инсуляторы и локус-контролирующие районы, локализованные в 5’ и 3’-фланкирующих районах генов, интронах, некодирующих экзонах. Комбинаторное взаимодействие между транскрипционными факторами, связывающимися с уникальным набором сайтов в регуляторных районах гена и обеспечивает ген-специфичную активность транскрипции.

Транскрипционные факторы влияют на интенсивность транскрипции, взаимодействуя с белками базального транскрипционного комплекса напрямую или через другие белки (коактиваторы или корепрессоры) (Featherstone, 2002). Кроме того, взаимодействие некоторых транскрипционных факторов с ДНК отражается на степени ацетилирования и метилирования гистоновых белков, плотности нуклеосомной укладки и уровне метилирования ДНК (Kouzarides 2002, McCabe *et al*., 2005, Agalioti, 2000), что также в значительной степени определяет интенсивность транскрипции генов.

Структурно-функциональная организация регуляторных районов генов, транскрибируемых РНК полимеразой II., характеризуется исключительной сложностью. Наличие альтернативных промоторов, удаленных регуляторных районов, содержащих энхансеры и сайленсеры, является обычным для множества исследованных в настоящее время генов. Сайты связывания транскрипционных факторов в пределах регуляторной единицы (промотора, энхансера, сайленсера) могут быть организованы в функциональные модули, определяющие тот или иной паттерн экспрессии гена (Linder, Prough, 1993). Еще одной функционально значимой характеристикой являются множественные старты транскрипции. Эта информация может быть очень важной, поскольку, зачастую отдельные старты транскрипции одного промотора используются для формирования транскриптов в различных тканях или в различных условиях (под действием индукторов, на разных стадиях развития организма и т.д.) (de Mena 1999). В ряде случаев координированная экспрессия нескольких транскрибируемых РНК полимеразой II генов из одного локуса обеспечивается за счет функционирования LCR (locus control regions) (Festenstein and Kioussis, 2000). LCR могут

включать энхансеры, позитивные или негативные регуляторные районы, а также сайты связывания транскрипционных факторов (Подколодная и Степаненко, 1997).

Все это с очевидностью свидетельствует о том, что для описания интегральной системы регуляции транскрипции необходима информация обо всей полноте регуляторных элементов гена. Создание коллекций экспериментально выявленных данных о регуляторных элементах всех уровней абсолютно необходимо как для формирования представлений о природе регуляции транскрипции отдельных генов, так и для создания компьютерных методов предсказания регуляторных элементов, построения генных сетей и функциональной аннотации геномов. Transcription Regulatory Regions Database (TRRD) разработана с целью комплексного описания регуляции транскрипции генов эукариот, транскрибируемых РНК полимеразой II.

В базе данных TRRD (Kolchanov N.A. et al., 2002), представлены структурные характеристики регуляторных районов генов эукариот, а также особенности функционирования регуляторных элементов и специфика экспрессии генов в зависимости от стадии развития организма, стадии клеточного цикла, вида клеток и степени их дифференцировки, внешних воздействий и т.д. TRRD содержит только экспериментально подтвержденные данные, которые внесены в базу экспертами-биологами на основании анализа научной литературы.

Описание работы с TRRD

Поиск информации в TRRD ведется с использованием стандартной поисковой формы системы SRS по конкретным полям базы TRRD, содержащим информацию нужного типа. Рассмотрим два примера поиска: 1) Поиск информации о регуляторных районах в БД TRRD; 2) Поиск информации о сайтах связывания транскрипционных факторов БД TRRD.

Пример 1. Поиск информации о регуляторных районах в БД TRRD

Чтобы осуществить поиск регуляторных районов конкретного гена, необходимо сформулировать запрос по названию гена и реализовать его по полям «GeneName», «GeneName_Brief», «GeneName_Full», по виду – по полю «Species»

Для этого необходимо войти в подсистему «Регуломика» и с помощью мыши кликнуть надпись «TRRDGENES4» в разделе «Формирование выборок регуляторных геномных последовательностей». В результате станет доступен вход в таблицу TRRDGENES4 в системе SRS (Рис. 1). На этой странице необходимо выбрать опцию «Search» и перейти к стандартной поисковой форме по таблице TRRDGENES в системе SRS (Рис. 2).

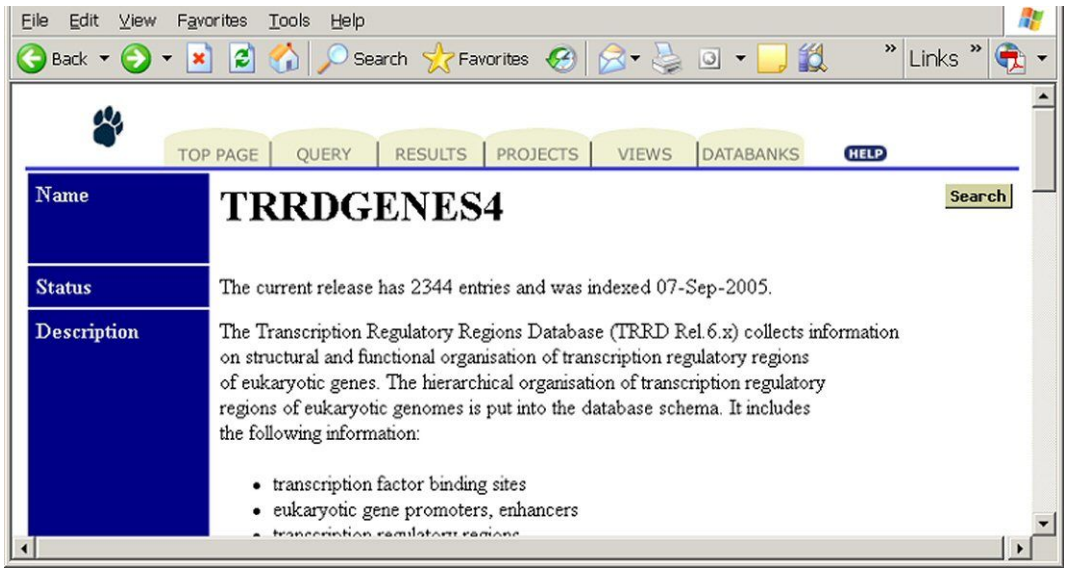


Рисунок 1. Вход в таблицу TRRDGENES в системе SRS

Выполнение операции начинается с ввода названия гена и вида в стандартную поисковую форму в системе SRS (Рис. 2). Для ввода короткого названия гена выбирается тип поля “GeneName_Brief” , а для названия вида выбирается поле “Species“. Опция “Combine search with” устанавливается в положение “AND”. В списке “Select fields to display” помечаются поля, которые будут выдаваться в результате поиска: “Species“, “GeneName_Full”, “RegUnitAC”, “RegUnit”.

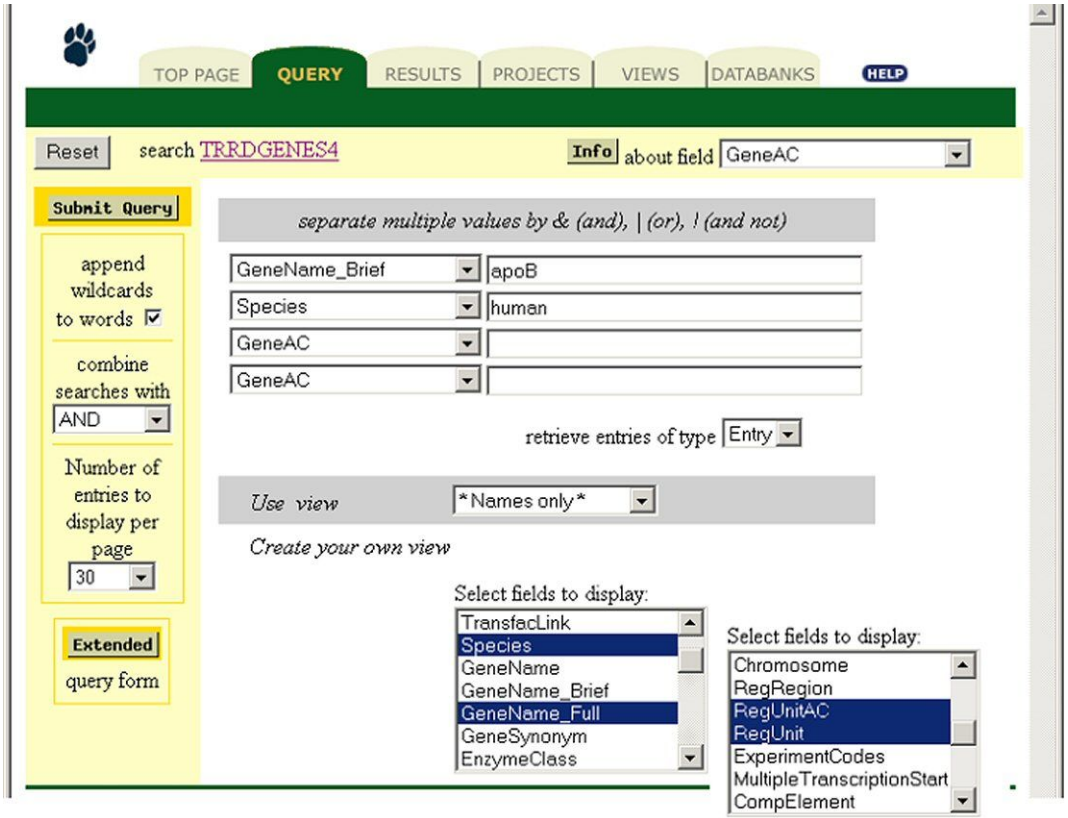


Рисунок 2. Ввод ключевых слов для поиска (название гена и вида) в поисковую форму по таблице TRRDGENES в системе SRS.

После ввода ключевых слов необходимо нажать клавишу “Submit Query”, в результате чего появляется HTML-страница с таблицей, в которой будут перечислены регуляторные районы гена (Рис. 3.) Поскольку данный запрос был описан на примере гена B человека (короткое название «apoB»), для которого по данным литературы изветно 9 регуляторных единиц, в результате запроса будут получены 9 номеров доступа к регуляторным единицам гена (см. колонку RegUnitAc, Рис. 3)

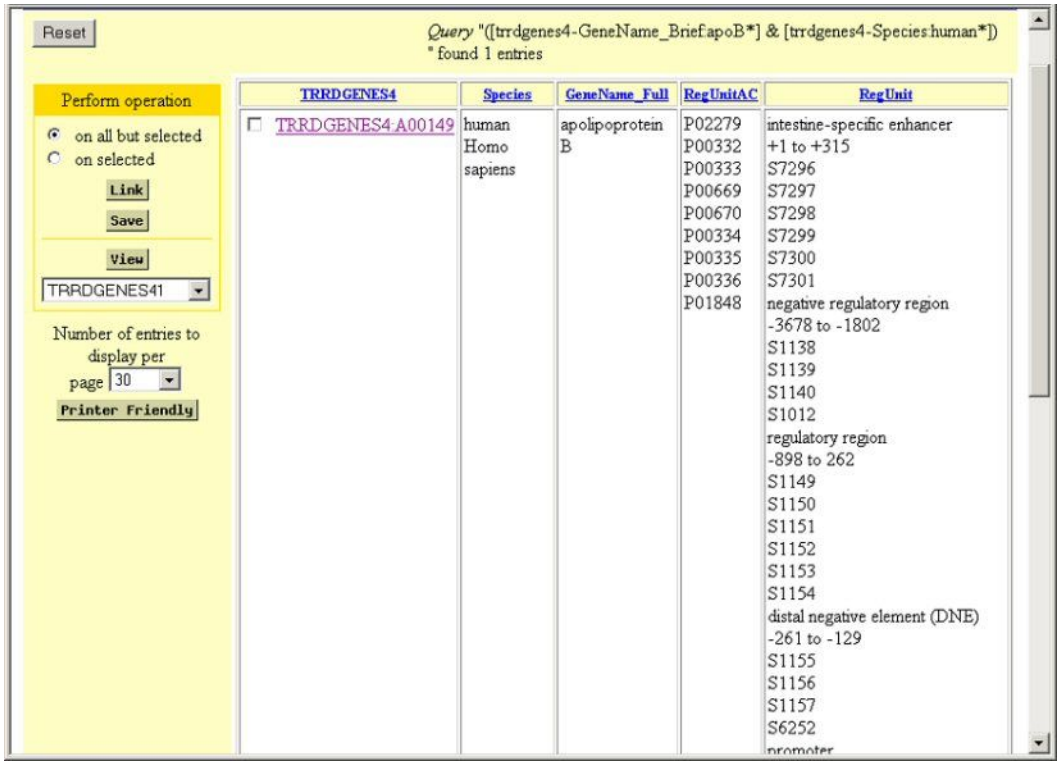


Рисунок 3. Пример HTML-страницы, содержащей результат запроса по таблице TRRDGENES.

Для завершения работы с данным запросом и перехода к новому запросу надо выбрать гипрссылку TRRDGENES4 в названии первой колонки таблицы и вернуться во вход в таблицу TRRDGENES в системе SRS (Рис. 2)

Пример 2. Поиск информации о сайтах связывания транскрипционных факторов БД TRRD.

Чтобы осуществить поиск сайтов связывания транскрипционных факторов необходимо сформулировать запрос по названию транскрипционного фактора и реализовать его по полю « FactorName».

Для этого необходимо войти в подсистему «Регуломика» и с помощью мыши кликнуть надпись «TRRDSITES4» в разделе «Формирование выборок регуляторных геномных последовательностей». В результате станет доступен вход в таблицу TRRDSITES4 в системе SRS (Рис. 4). На этой странице необходимо выбрать опцию «Search» и перейти к стандартной поисковой форме по таблице TRRDSITES в системе SRS (Рис. 5).

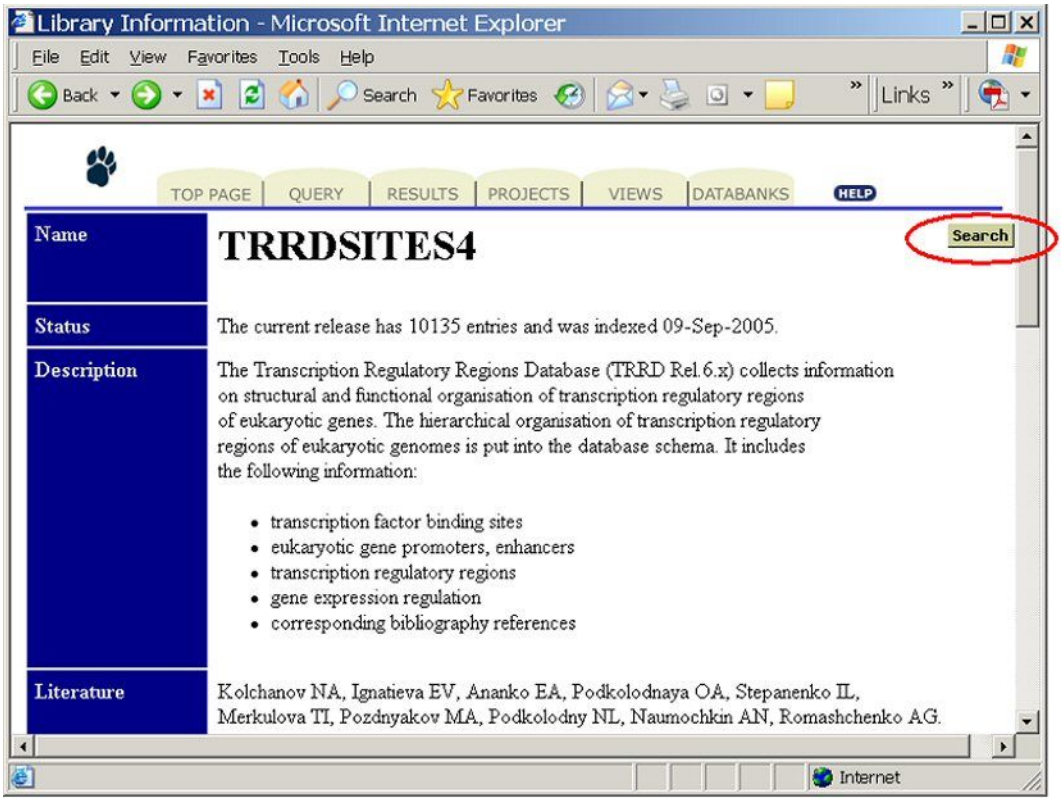


Рисунок 4. Вход в таблицу TRRDSITES в системе SRS

Для ввода названия транскрипционного фактора выбирается тип поля “ FactorName”. В списке “Select fields to display” помечаются поля, которые будут выдаваться в результате поиска: “SiteAC“, “SiteName”, “GeneID”, “RegUnitAC”, “FactorName” (Рис. 5).

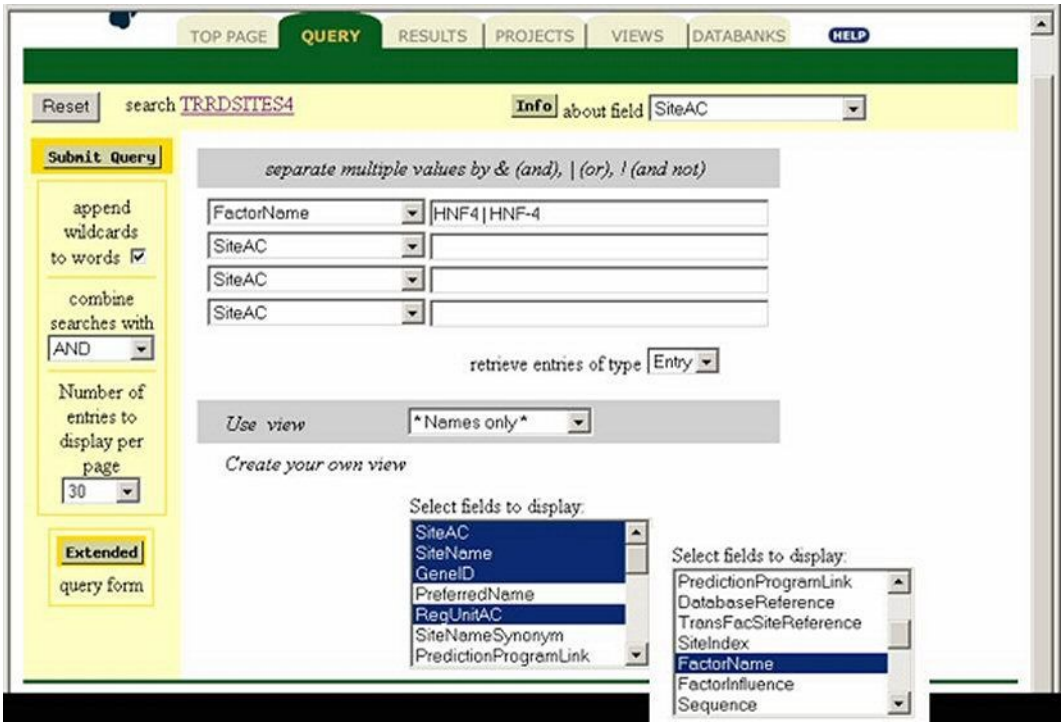
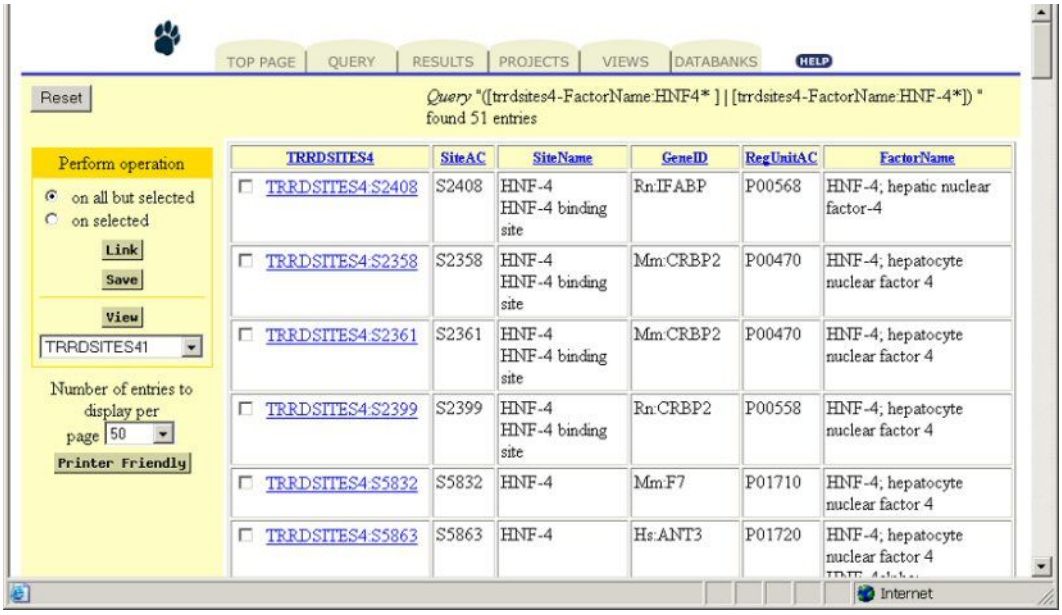


Рисунок 5. Ввод ключевого слова для поиска (название транскрипционного фактора) в поисковую форму по таблице TRRDSITES в системе SRS.

После ввода ключевых слов необходимо нажать клавишу “Submit Query”, в результате чего появляется HTML-страница с таблицей, в которой будут перечислены сайты связывания транскрипционных факторов заданного типа (Рис. 6).

Поскольку процедура поиска сайтов связывания транскрипционных факторов описана на примере сайтов HNF4, запрос реализован по ключевым словам HNF4 и HNF-4, объединенным оператором OR (то есть “HNF4 | HNF-4”). Результатом запроса будет (для релиза TRRD от 09/09/2005) информация о 51 сайтах связывания этого фактора, номера которых представлены в колонке SiteAC (Рис. 6).



TRRDSITES4	SiteAC	SiteName	GeneID	RegUnitAC	FactorName
TRRDSITES4:S2408	S2408	HNF-4 HNF-4 binding site	Rn:IFABP	P00568	HNF-4; hepatic nuclear factor-4
TRRDSITES4:S2358	S2358	HNF-4 HNF-4 binding site	Mm:CRBP2	P00470	HNF-4; hepatocyte nuclear factor 4
TRRDSITES4:S2361	S2361	HNF-4 HNF-4 binding site	Mm:CRBP2	P00470	HNF-4; hepatocyte nuclear factor 4
TRRDSITES4:S2399	S2399	HNF-4 HNF-4 binding site	Rn:CRBP2	P00558	HNF-4; hepatocyte nuclear factor 4
TRRDSITES4:S5832	S5832	HNF-4	Mm:F7	P01710	HNF-4; hepatocyte nuclear factor 4
TRRDSITES4:S5863	S5863	HNF-4	Hs:ANT3	P01720	HNF-4; hepatocyte nuclear factor 4

Рисунок 6. Пример HTML-страницы, содержащей результат запроса по таблице TRRDSITES.

Для завершения работы с данным запросом и перехода к новому запросу надо выбрать гипрссылку TRRDSITES4 в названии первой колонки таблицы и вернуться во вход в таблицу TRRDSITES в системе SRS (Рис. 5)

Матрицы сайтов связывания транскрипционных факторов

Ключевым звеном тонкой регуляции экспрессии генов является структура регуляторных последовательностей промоторов генов, определяющая спектр возможных воздействий со стороны регуляторов транскрипции. Поэтому неудивительно, что поток информации по изучению механизмов регуляции транскрипции и структуры промоторов не ослабевает. Более того, развитие новых технологий создает дополнительные возможности для накопления и анализа информации по структуре регуляторных районов генов. В связи с этим, в последнее десятилетие все большее внимание уделяется созданию баз данных по различным отраслям знаний, в том числе и по регуляции транскрипции, как у эу-, так и у прокариот (Wingender et al., 2001; Kolchanov et al., 2002; Lescot et al., 2002; Praz et al., 2002; Munch et al., 2003). Одной из таких баз является база TRRD – база транскрипционных регуляторных районов генов эукариот. Ее создание началось более десяти лет назад в Институте цитологии и генетики

COPAH и к настоящему времени в ней накоплена значительная информация по структуре промоторов различных генов эу- и прокариот (Kolchanov et al., 2002). Однако, в последнее время, появилось достаточно много информации о структуре сайтов связывания для различных транскрипционных факторов, информация о которых не могла быть занесена в базу TRRD, так как противоречила основным принципам ее создания, но которая позволила бы значительно расширить наши знания о структуре последовательностей, взаимодействующих с транскрипционными факторами. Речь идет о данных по структуре сайтов связывания различных транскрипционных факторов, полученных на основе селекции *in vitro* синтезированных последовательностей ДНК на способность связывания с определенным транскрипционным фактором. О важности накопления этих знаний говорит появление баз данных SELEX_DB и JASPAR, которые пытаются систематизировать эту информацию (Ponomarenko et al., 2000, 2002; Sandelin et al., 2004), однако информационное содержание этих баз к настоящему времени невелико, 116 и 111 входов, соответственно. Более того, объем данных в SELEX_DB с 2000 года не увеличился (Ponomarenko et al., 2000, 2002). Накопление подобной информации идет также и в базе TRANSFAC, одной из наиболее значительных баз по регуляции транскрипции (Matys et al., 2003). К сожалению, большая часть этих данных доступна только в коммерческом релизе базы TRANSFAC. Поэтому не удивительно, что, имея значительный объем информации о структуре природных сайтов в базе TRRD, была поставлена задача объединить данные, полученные с помощью *in vitro* технологий и данные, накопленных к настоящему времени в базе TRRD, и создать базу ArtSite.

ArtSite - это база данных по структуре сайтов связывания транскрипционных факторов (ТФ) у про- и эукариот (Khlebodarova T.M. et al., 2005). Особенности структуры сайтов связывания для различных ТФ описываются в базе ArtSite с помощью матриц, которые построены на основе выравнивания представительных выборок нуклеотидных последовательностей сайтов связывания ТФ. В качестве примера в таблице 1 приводится матрица для природных последовательностей сайтов связывания USF1. Выборки сайтов созданы на основе как природных, так и *in vitro* синтезированных последовательностей, описанных в оригинальной литературе и выявленных с помощью различных методов селекции.

Нуклеотид	Позиция нуклеотида									
	-5	-4	-3	-2	-1	0	+1	+2	+3	+4
A	**22**	6	0	**48**	2	4	2	1	**19**	4
C	10	**23**	**51**	1	**36**	10	1	0	8	**23**
T	2	**17**	0	2	3	0	**48**	4	5	9
G	**18**	9	1	1	11	**37**	1	**47**	**20**	**16**
Консенсус	**R**	**Y**	**C**	**A**	**C**	**G**	**T**	**G**	**R**	**S**

Таблица 1. Матрица сайта связывания транскрипционного фактора USF1 представленная в базе данных ArtSite (номер входа в БД ArtSite - NS00022).

В базе ArtSite присутствует два типа входов для матриц природных сайтов связывания ТФ. Первые получены в результате селекции клонированных геномных фрагментов ДНК, и в этом случае формат описания входа практически не отличается от вышеописанного, за исключением того, что не заполняется поле описания последовательности синтетической матрицы (МА). Вторые получены на основе выравнивания сайтов связывания, экстрагированных из базы TRRD. В этом случае каждый вход получен на основе аннотирования большого числа оригинальных статей, в которых использованы транскрипционные факторы разного происхождения, поэтому соответствующие поля,

касающиеся аннотированной статьи (AU, TI, SO, VL, IS, PG, YR и ML) и вида ТФ (TS), не заполняются. Также не заполняются поля, описывающие селекционный эксперимент (MX, RS), а в поле NM указывается номер сайта в базе TRRD.

Описание работы с БД ARTSITE •

Поиск информации о базе данных ARTSITE ведется с использованием интерфейса ARTSITE по конкретным полям базы ARTSITE, содержащим информацию соответствующего типа. Например, запрос по названию фактора можно реализовать по полям «Transcription factor name & synonyms», по названию домена– по полю «Domain», по авторам публикации по полю «Authors» и т.д. Также можно выбирать поиск по публикации из которой было взято описание данных в карточку ARTSITE по полям «Authors», «Journal», «Pages», «Year»,«Volume»,«PubMed ID», Номеру доступа карточки в базе данных «Accession number», и консенсусу сайта связывания в поле «Consensus». Возможно одновременное задание до трех критериев для поиска.

Для выполнения операции надо войти в подсистему «Регуломика» и с помощью мыши кликнуть надпись «Матрицы сайтов связывания транскрипционных факторов» в списке баз данных подсистемы. В результате появится HTML-страница главной страницы базы данных, которая показана на Рис. 7. На этой странице необходимо выбрать опцию «Search by» в одном из трех впадающих меню выбрав поле по которому необходимо осуществить поиск, после чего ввести одно из ключевых слов для осуществления запроса.

Для поиска по названию транскрипционного фактора нужно выбрать тип поля «Transcription factor name & synonyms» и ввести, например, название транскрипционного фактора YY1.

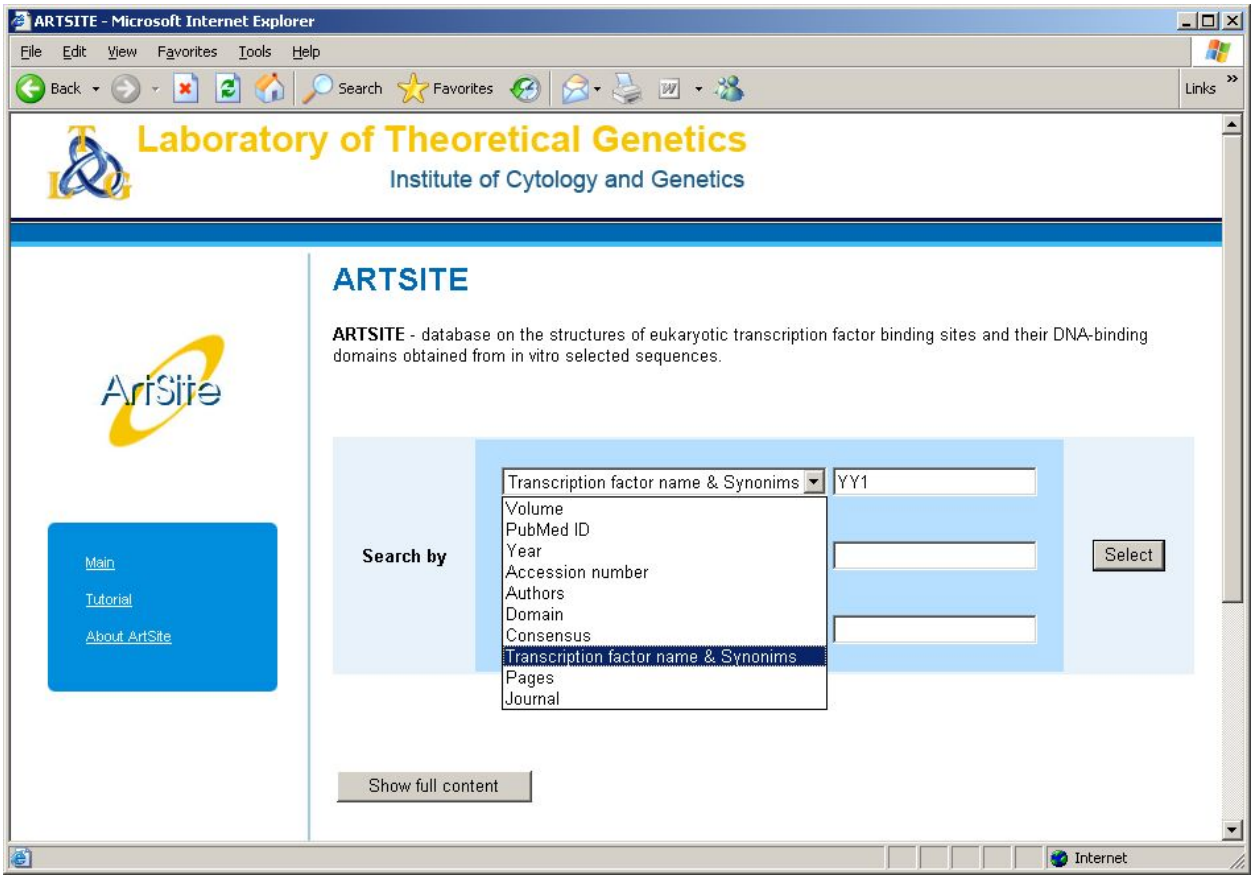
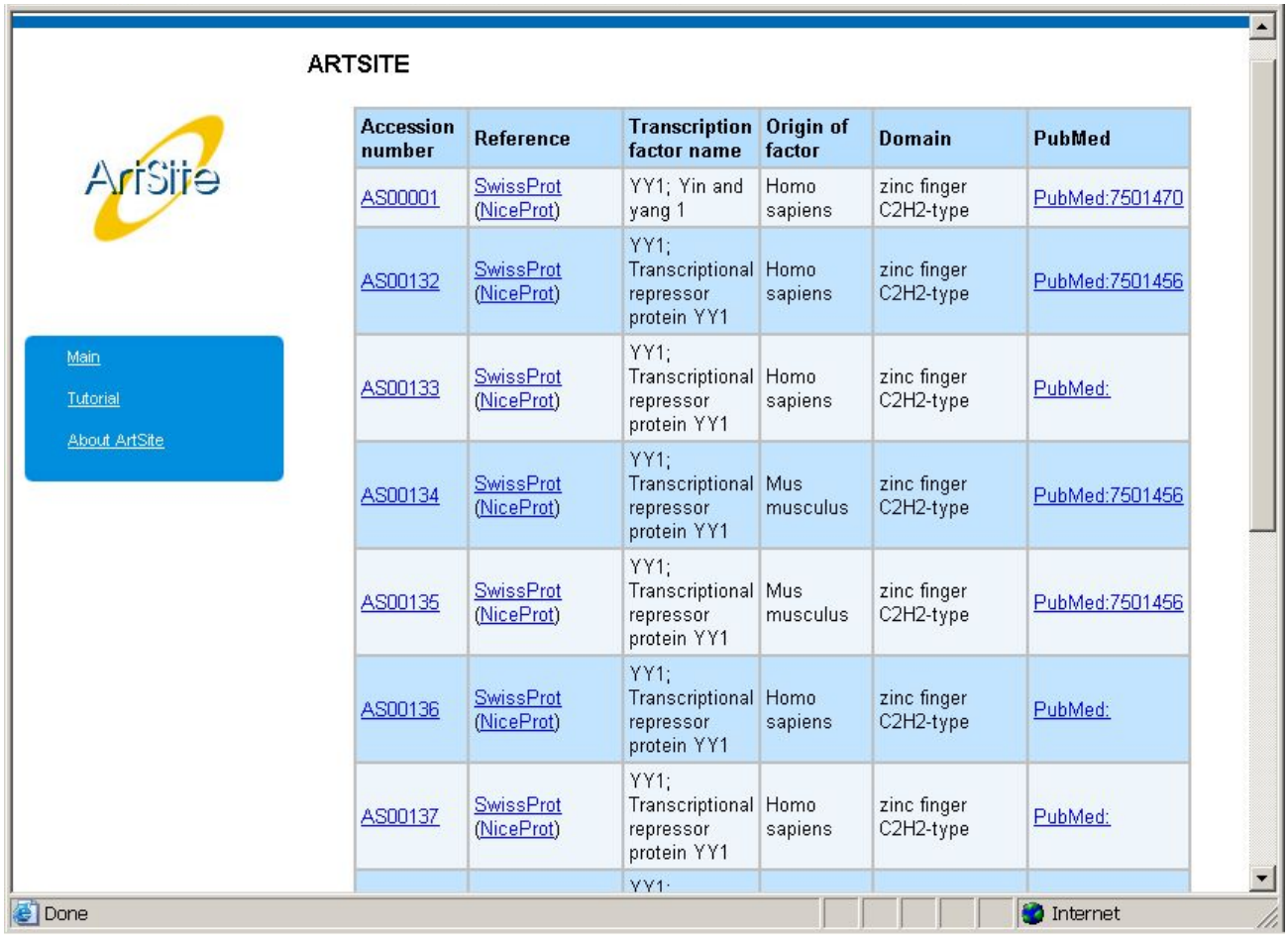


Рисунок 7. Выбор поля для поиска и ввод ключевых слов для поиска (название транскрипционного фактора) в поисковую форму ARTSITE

После ввода ключевых слов необходимо нажать клавишу “Select”, в результате чего появляется HTML-страница с таблицей, в которой будет приведены карточки, включающие информацию о сайтах связывания YY1 (Рис. 8)



Accession number	Reference	Transcription factor name	Origin of factor	Domain	PubMed
AS00001	SwissProt (NiceProt)	YY1; Yin and yang 1	Homo sapiens	zinc finger C2H2-type	PubMed:7501470
AS00132	SwissProt (NiceProt)	YY1; Transcriptional repressor protein YY1	Homo sapiens	zinc finger C2H2-type	PubMed:7501456
AS00133	SwissProt (NiceProt)	YY1; Transcriptional repressor protein YY1	Homo sapiens	zinc finger C2H2-type	PubMed:
AS00134	SwissProt (NiceProt)	YY1; Transcriptional repressor protein YY1	Mus musculus	zinc finger C2H2-type	PubMed:7501456
AS00135	SwissProt (NiceProt)	YY1; Transcriptional repressor protein YY1	Mus musculus	zinc finger C2H2-type	PubMed:7501456
AS00136	SwissProt (NiceProt)	YY1; Transcriptional repressor protein YY1	Homo sapiens	zinc finger C2H2-type	PubMed:
AS00137	SwissProt (NiceProt)	YY1; Transcriptional repressor protein YY1	Homo sapiens	zinc finger C2H2-type	PubMed:

Рисунок 8. Пример HTML-страницы, содержащей результат запроса в БД ARTSITE.

После получения данных в таблице возможно воспользоваться линком в первой колонке (Accession number, например, AS00001) для получения полного набора данных, содержащихся в одной карточке ARTSITE (Рис. 9).

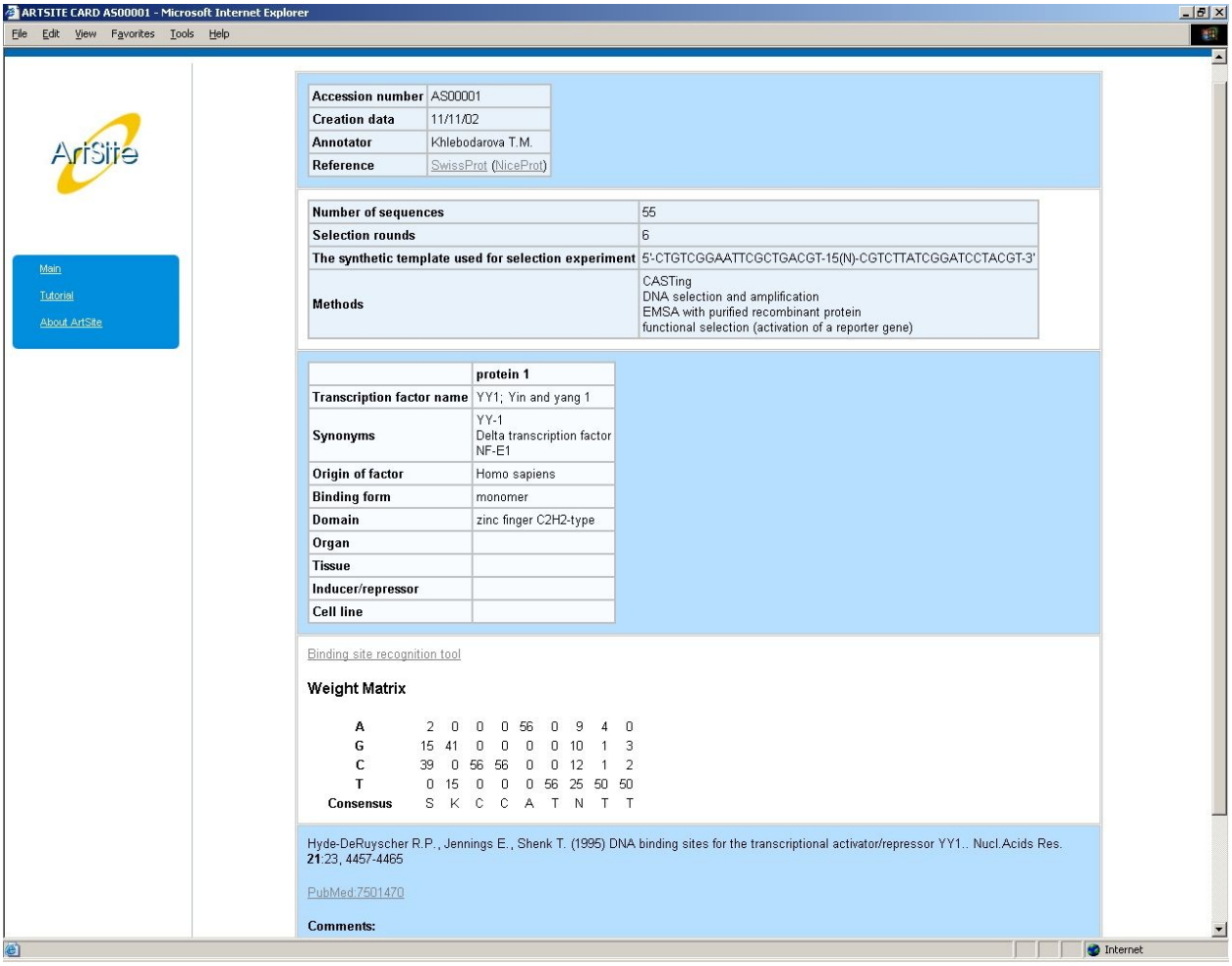


Рисунок 9. Пример HTML-страницы, содержащей полную запись карточки в БД ARTSITE.

Для завершения работы с данным запросом и перехода к новому запросу необходимо вернуться на главную страницу БД, используя закладку «Main» в левой части экрана.

Структурно-функциональная организация регуляторных последовательностей, контролирующих транскрипцию генов прокариот

Особенностью регуляции транскрипции генов прокариот является организация генов в опероны. Оперон – это единая транскрипционная единица, включающая несколько генов, имеющая общий промотор и оператор. Промотор оперона распознается ДНК-полимеразой при инициации транскрипции. С оператором, фрагментом ДНК, связываются репрессоры или активаторы, регулирующие уровень транскрипции мРНК генов, входящих в оперон. Оперон может содержать внутренние промоторы, регулирующие отдельные гены оперона, или состоять из одного гена. Типичный промотор прокариот, как и эукариот состоит из мотивов, позиционированных на определенном расстоянии относительно старта транскрипции. Прокариотический промотор содержит два основных мотива, расположенных на расстоянии –10 и –35 относительно старта транскрипции (Harley C.B. and Reynolds R.P., 1987).

Экспрессия генов прокариот зависит от условий внешней среды, и тип регуляции оперонов может быть негативной или позитивной. Негативная или позитивная регуляция оперона осуществляется при связывании белка репрессора или соответственно активатора другого регуляторного гена, имеющего свой промотор либо общий с регулируемыми генами, с оператором. Другой тип регуляции характерен для индуцибельных оперонов с негативной и позитивной регуляцией. В первом случае регуляторный репрессорный белок обычно связан

со своим сайтом в промоторе и подавляет транскрипцию генов оперона. Молекула индуктор связывается с репрессором, изменяет его конформацию, что делает его неспособным связываться с ДНК, и транскрипция генов включается. Такой тип регуляции характерен для негативных индуцибельных промоторов. Активация индуцибельных промоторов с позитивной регуляцией осуществляется при взаимодействии индуктора с активатором. Белок-активатор не способен связываться с ДНК, но при образовании комплекса с индуктором, он меняет свою конформацию и связывается со своим сайтом на ДНК и включает процесс транскрипции. У позитивного репрессорного типа промоторов подавление транскрипции происходит, если активатор, усиливающий транскрипцию генов оперона, претерпевает конформационные изменения при связывании с индуктором, что приводит к его диссоциации от ДНК.

База данных ProkaTEX предназначена для накопления экспериментальных данных о структуре транскрипционных регуляторных районов генов у различных видов бактерий, а также данных по экспрессии генов в зависимости от стадии роста бактериальной культуры, источников питания и дыхания, различных внешних воздействий и т.д. Вход в базу соответствует одному гену и содержит информацию об оперонах, генах и регуляции транскрипции генов прокариот

Комплексное описание регуляции транскрипции генов прокариот, представленное в базе ProkaTEX, является основой для решения большого числа актуальных задач биоинформатики и представляет интерес для исследователей, занимающихся анализом закономерностей структурно-функциональной организации районов регуляции транскрипции генов прокариот; распознаванием сайтов связывания транскрипционных факторов; изучением механизмов транскрипции; реконструкцией генных сетей; моделированием молекулярно-генетических процессов в бактериальной клетке; интерпретацией данных microarray и т.д.

База данных ProkaTEX содержит информацию

- о гене, включая название гена, синонимы, название оперона, в структуру которого он входит, положение гена на хромосоме, уникальный номер гена в геноме бактерии, ссылки на базы (GenBank/EMBL, SWISS-PROT, BioCyc и др.), позиции сайтов инициации и терминации трансляции
- о структуре регуляторных районов гена (5'-фланкирующие районы), стартах транскрипции (с информацией о множественных стартах транскрипции), последовательностях регуляторных единиц (промоторов), и их взаимном расположении, функции и индуцибельности
- о сайтах, входящие в регуляторные единицы, их последовательностях, позициях сайтов, важных для их функционирования
- о транскрпционных факторах, связывающихся с сайтом, их влиянии на транскрипцию гена
- о паттернах экспрессии гена и регуляторных единицах и сайтах, ответственных за осуществление отдельных паттернов экспрессии гена, индукторах, времени их действия и уровне экспрессии гена.

Описание работы с ProkaTEX

SRS (Sequence Retrieval System) используется как основной инструмент поиска и навигации по базе данных ProkaTEX. Информация распределена по пяти таблицам, которые представлены на рис. 10: PROKATEXGENES (основное описание гена), PROKATEXSITES (сайты связывания транскрипционных факторов), PROKATEXFACTORS (транскрипционные факторы), PROKATEXEXP (описание паттернов экспрессии), PROKATEXBIB (ссылки на

публикации). Использование различных запросов к базе позволяет получить список всех генов данного вида, по ключевому слову, экспрессирующегося на определенной стадии клеточного цикла, фазе роста клеточной культуры, индуцируемым или подавляемым определенным веществом. Получить список сайтов связывания транскрипционного фактора, а также генов, транскрипция которых регулируется данным транскрипционным фактором, что подтверждается определенными экспериментами. Возможны различные комбинации запросов, например, ген, индуцируемым каким-либо веществом и регулируемый определенным транскрипционным фактором.

Databanks Available

11-Aug-2006 22:51

Data Bank	Release	No Entries	Indexing Date	Group	Availability
ProkaTEXGENES		614	11-Aug-2006	Prokariot Database	ok
ProkaTEXBIB		3250	11-Aug-2006	Prokariot Database	ok
ProkaTEXEXP		3198	11-Aug-2006	Prokariot Database	ok
ProkaTEXFACTORS		1982	11-Aug-2006	Prokariot Database	ok
ProkaTEXSITES		2442	11-Aug-2006	Prokariot Database	ok
SENSORGENE		106	06-Jul-2006	Gene Sensor Database	ok
SENSORBIB		186	06-Jul-2006	Gene Sensor Database	ok
SENSORINDUCTOR		223	06-Jul-2006	Gene Sensor Database	ok
SENSORPROMOTER		120	06-Jul-2006	Gene Sensor Database	ok
SENSORFACTOR		43	04-Jul-2006	Gene Sensor Database	ok
SENSORSITE		155	30-Jun-2006	Gene Sensor Database	ok
TRRDGENES4		2344	07-Sep-2005	Transcription Regulatory Regions Database	ok
TRRDSITES4		10135	09-Sep-2005	Transcription Regulatory Regions Database	ok
TRRDUNITS4		3490	07-Sep-2005	Transcription Regulatory Regions Database	ok
TRRDFACTORS4		9087	07-Sep-2005	Transcription Regulatory Regions Database	ok
TRRDEXP4		14407	07-Sep-2005	Transcription Regulatory Regions Database	ok
TRRDBIB4		7609	07-Sep-2005	Transcription Regulatory Regions Database	ok
TRRDLCR		14	21-Mar-2005	Transcription Regulatory Regions Database	ok
TRRDPOT		2374	21-Mar-2005	Transcription Regulatory Regions Database	ok
TRRDPROMOTERS		763	15-Sep-2005	Transcription Regulatory Regions Database	ok
GN_GENE		1006	11-Feb-2004	GeneNet Database	ok

Рис. 10. Вход в таблицы базы ProkaTEX в системе SRS. ProkaTEXGENES – общая информация о гене; ProkaTEXEXP - описание паттернов экспрессии; ProkaTEXSITES - сайты связывания транскрипционных факторов; ProkaTEXFACTORS - транскрипционные факторы; ProkaTEXBIB - ссылки на публикации

Для поиска интересующей информации в базе ProkaTEX необходимо выбрать одну из таблиц. Например, выберем таблицу ProkaTEXEXP, которая содержит информацию об индуцибельности промоторов бактериальных генов.

Поиск по таблице ProkaTEXEXP дает информацию о паттернах экспрессии генов, индуцируемых одним индуктором (RI), с определенным уровнем экспрессии (RL), на определенной стадии развития организма (RD). Для поиска такого рада информации нужно составить соответствующие запросы, используя словари, выбрав соответствующие ссылки. Например, для поиска промоторов генов с высоким уровнем экспрессии генов при низком pH используется последовательность действий, представленных на рис.11. В результате этого поиска будет получен список паттернов экспрессии, соответствующих заданным условиям. В записи идентификатора каждого паттерна из полученного списка содержится ссылка на номер гена, для которого данный паттерн экспрессии описан. Эта ссылка позволяет получить информацию о промоторах генов и их последовательностях.

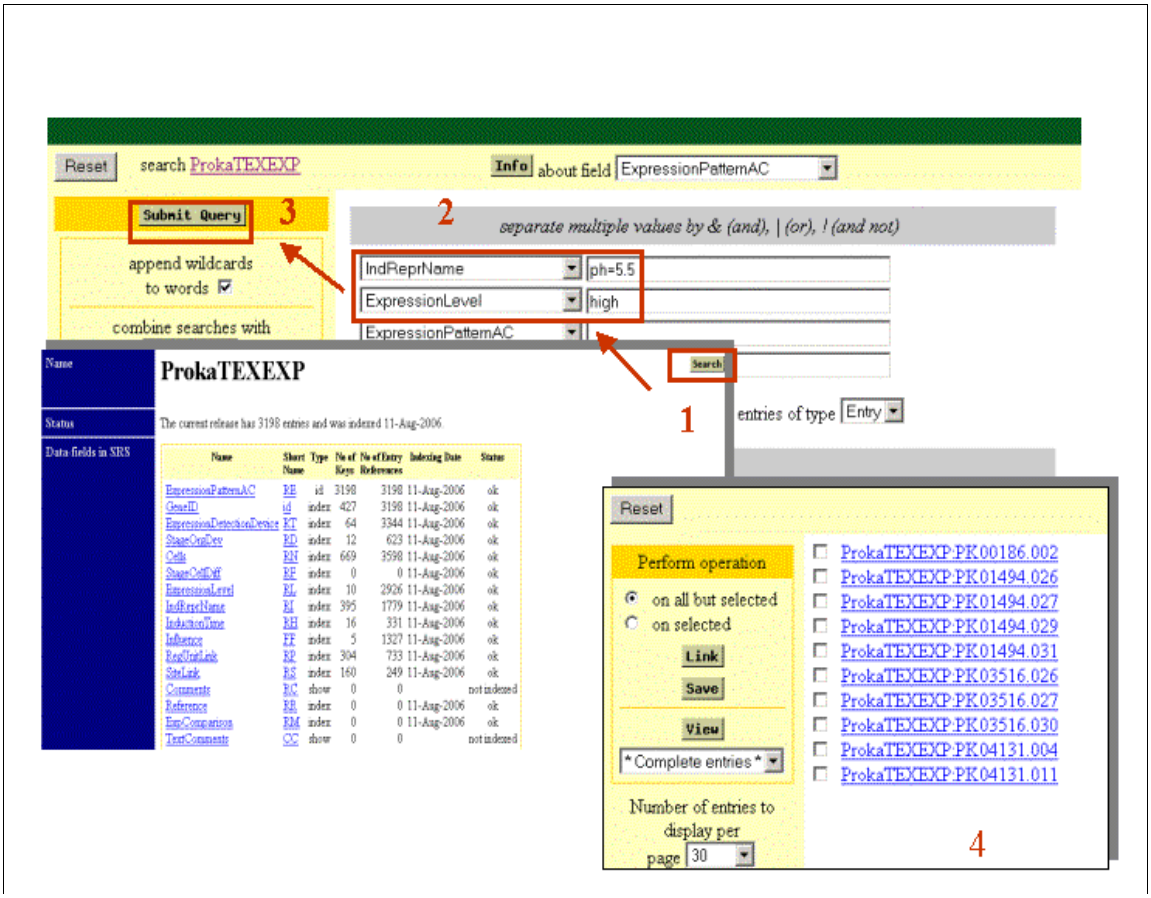


Рис. 11. Вход в базу ProkaTEXEXP и последовательность действий для поиска промоторов генов с высоким уровнем экспрессии генов при низком pH.

Информация о генах содержится в таблице ProkaTEXGENES базы данных ProkaTEX. Таблица ProkaTEXGENES содержит общую информацию о гене, которая позволяет точно его идентифицировать, т.е. название гена, синонимы, название оперона, в структуру которого он входит, положение гена на хромосоме, уникальный номер гена в геноме бактерии ссылки на базы (GenBank/EMBL, SWISS-PROT, BioCyc и др.), позиции сайтов инициации и терминации трансляции а так же информацию о структуре регуляторных районов гена (5'-фланкирующие районы) и стартах транскрипции (с информацией о множественных стартах транскрипции). Благодаря большому количеству индексированных полей в базе данных ProkaTEX, возможно осуществлять различные варианты поиска информации.

Один из примеров поиска в таблице ProkaTEXGENES представлен на рисунке 12. Кликнув мышью на индексированное поле KeyWords, выходим в окно поиска по ключевым словам. Здесь мы можем сделать поиск по любому ключевому слову, или, нажав кнопку List Values, получить список всех ключевых слов и словари клеток, стадий развития, уровней экспрессии и т.д.. Выбрав ключевой слово, например «amino acid biosynthesis and metabolism» получаем список генов, продукты которых участвуют в биосинтезе и метаболизме аминокислот (Рис. 13).

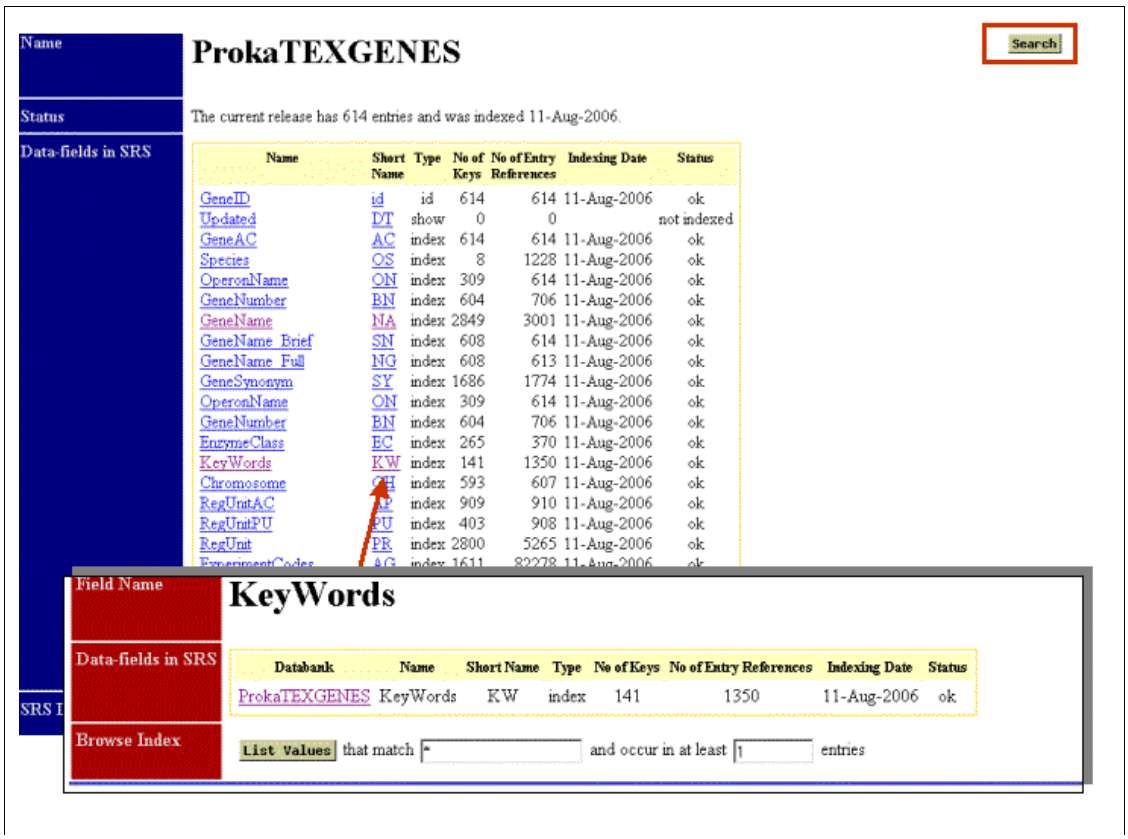


Рис.12. Поиск по базе ProkaTEXGENES по ключевым словам

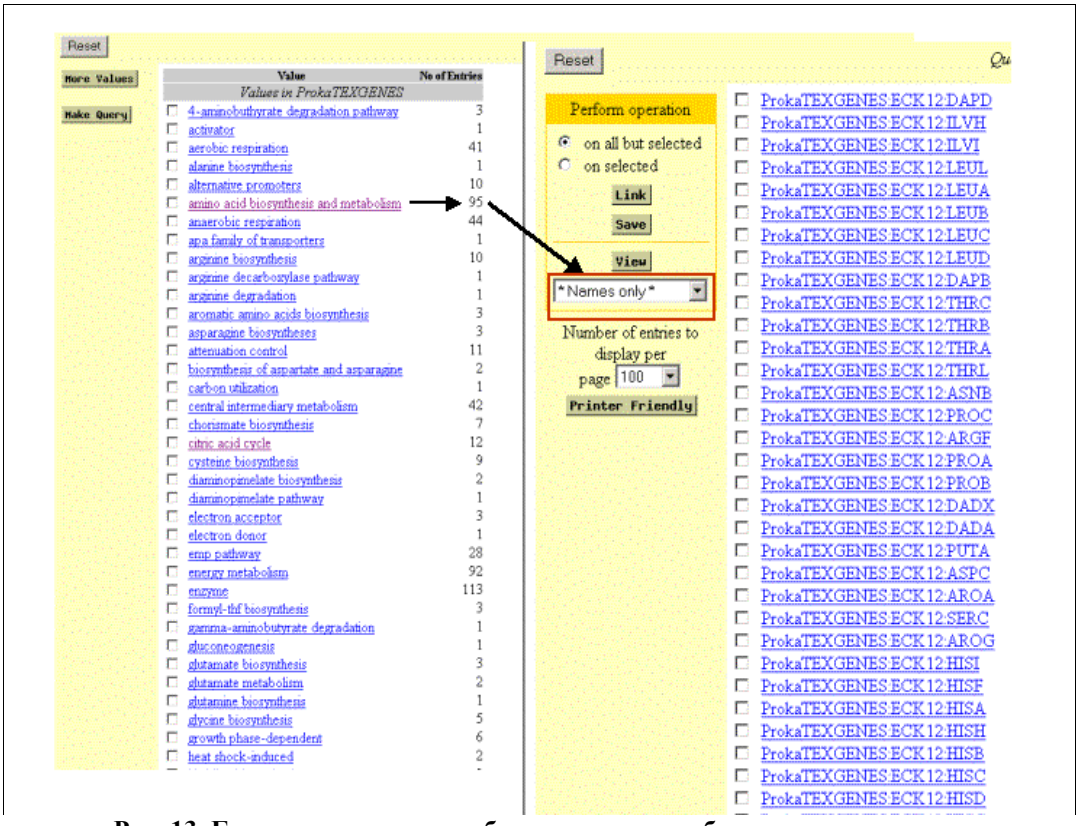


Рис. 13. Гены, участвующие в биосинтезе и метаболизме аминокислот

Система позволяет получить список идентификаторов генов или полные входы, соответствующие описанию отдельных генов из этого списка. Кнопка **«Save»** позволяет сохранить весь список генов в текстовом формате или html. Для этого необходимо выбрать **Complete entries** и нажать кнопку **View** (рис. 14)

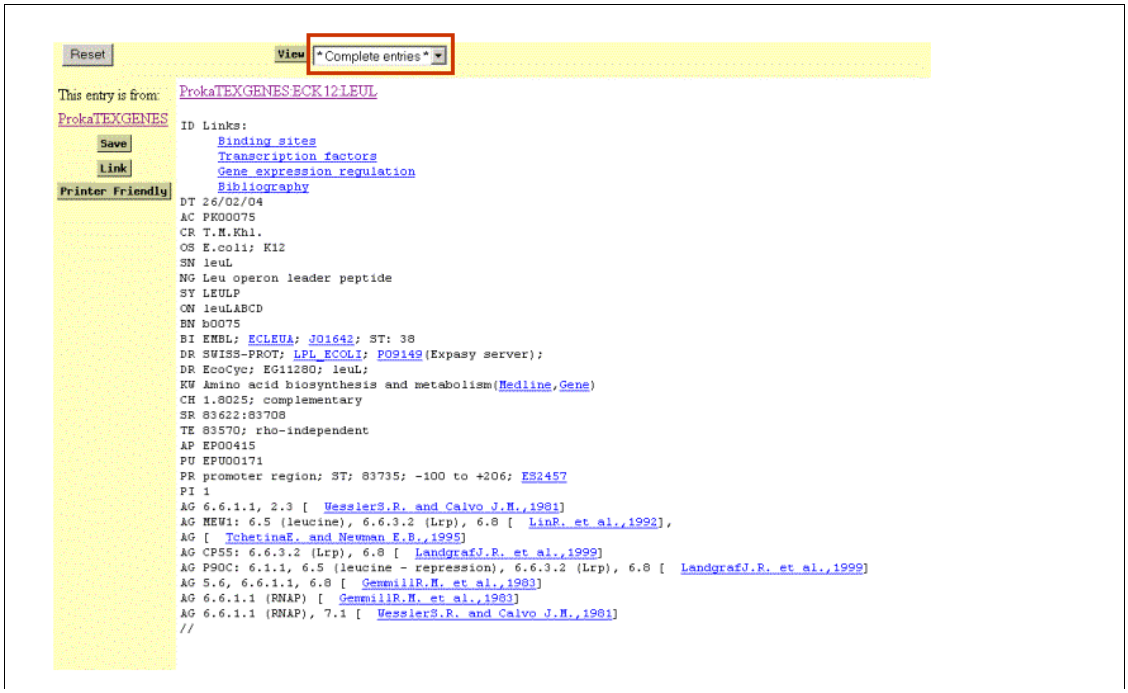


Рисунок 14. Вход в базу ProkaTEXGENES для гена leuLE.coli

Поле AG Таблицы ProkaTEXGENES и ProkaTEXSITES содержит список экспериментов, которые были проведены для идентификации промоторов и характеристики ДНК последовательностей cis-элементов и trans-активирующих белков, регулирующих транскрипцию гена. Поле включает название штамма, коды экспериментов и ссылку на оригинальную публикацию.

Картирование стартов инициации транскрипции	S1 nuclease protection (code 5.6), Primer extension assay (code 5), RNase protection analysis (code 5.5), RACE (rapid amplification cDNA 5'end) (code 5.2)
Функциональный анализ промоторов	Делеционный анализ (code 6.1.1), мутантный анализ (code 6.2) и делеционный с экзонуклеазами: exonuclease III (code 2.1), alpha exonuclease (code 2.2), endonuclease digests, restriction endonuclease protection (code 2.3).
Анализ индуцибельности промотора	Индуктор (репрессор) и анализ мутантов (code 6.5 + 6.1.1 или 6.1.2, 6.2)

На рис. 15 представлен вариант запроса к базе, позволяющий выявить гены с экспериментально подтвержденными стартами транскрипции (эксперимент 5, 5.2, 5.5, 5.6). Если в окне ключевых слов задать интересующее нас слово, например «citric acid cycle», можно получить выборку промоторов генов, продукты которых участвуют в цикле Кребса, задав другие слова можно получить выборки других промоторов, отобранных из базы по ключевым словам.

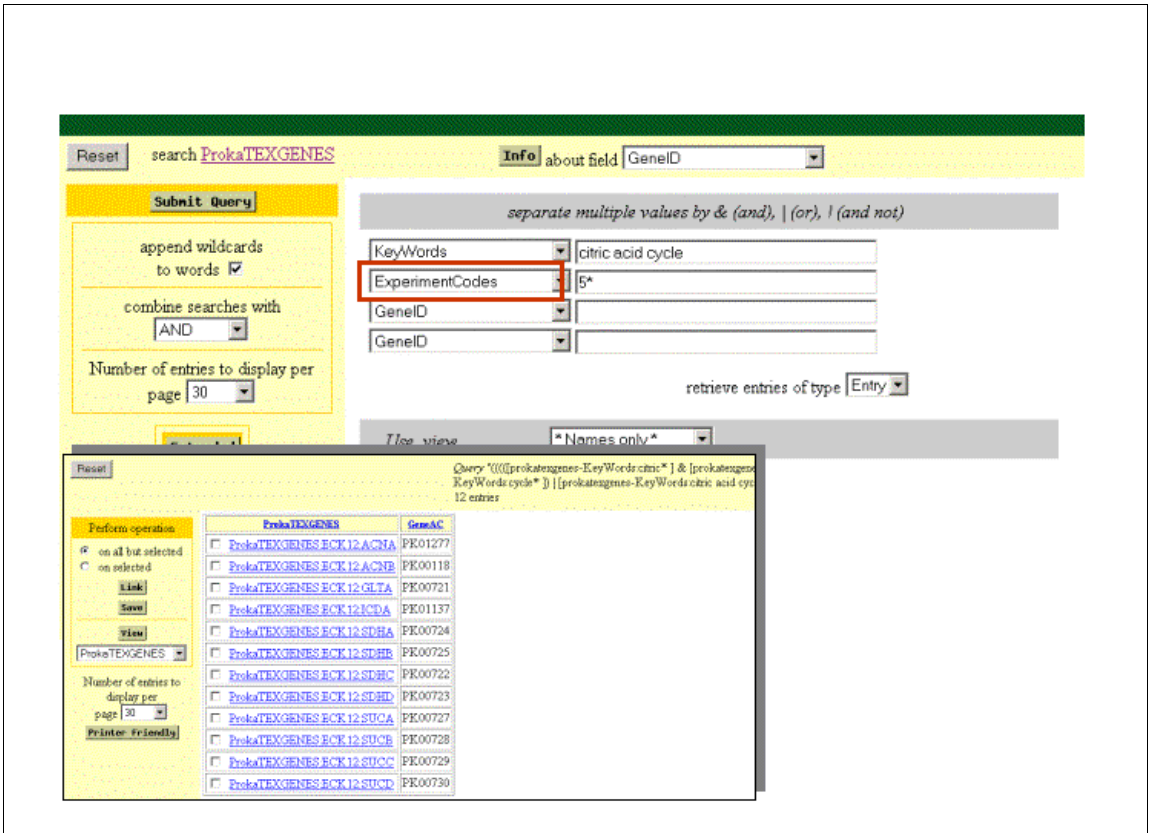


Рис. 15. Поиск промоторов генов белков, участвующих в цикле Кребса, с экспериментально показанными стартами инициации транскрипции

Используя возможности системы SRS, позволяющие устанавливать связи между отдельными таблицами базы данных, можно получить информацию о паттернах экспрессии этих генов или транскрипционных факторах, регулирующих процесс. Для этого нужно нажать кнопку «Link» , отметить соответствующую базу и нажать кнопку «Submit link» (рис.16).

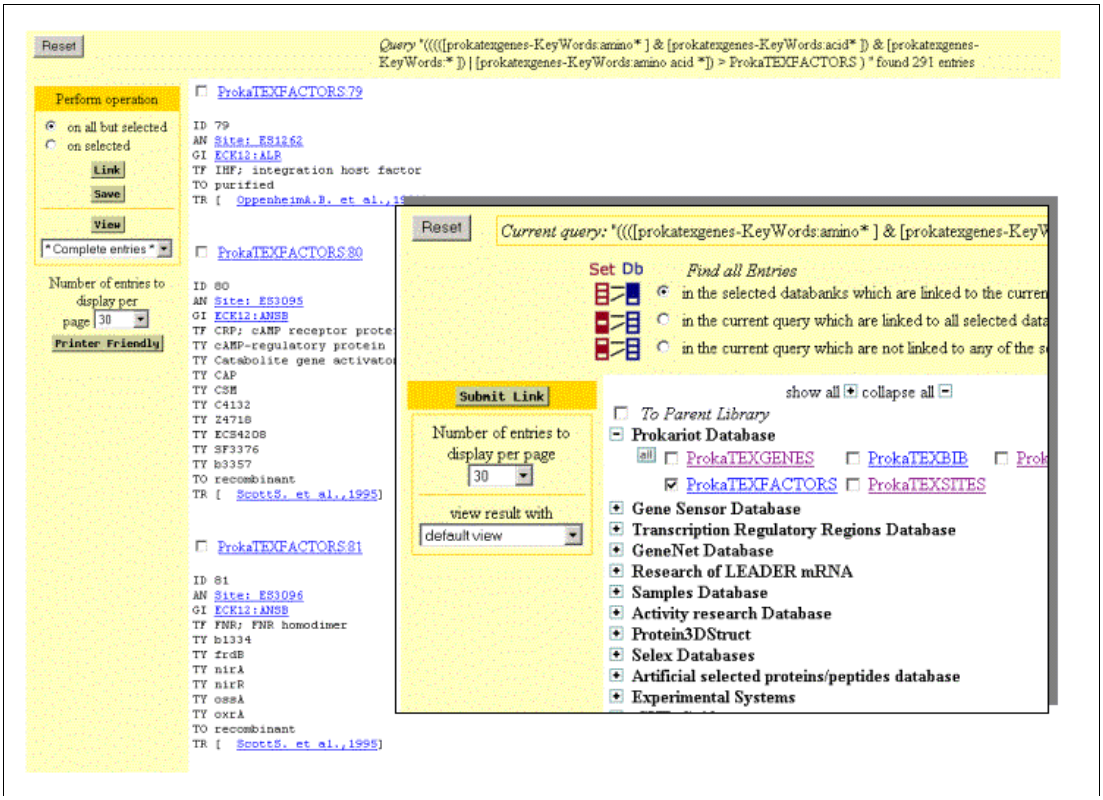


Рис. 16. Информация о транскрипционных факторах, регулирующих гены биосинтеза и метаболизма аминокислот.

Таблица ProkaTEXSITES позволяет создавать выборки сайтов связывания транскрипционных факторов, что является основой для разработки методов распознавания определенных типов сайтов. На рисунке 17 проставлен пример поиска информации в таблице ProkaTEXSITES. Здесь задан поиск сайтов, связывание с которыми транскрипционного фактора Fnr, приводит к увеличению уровня транскрипции гена.

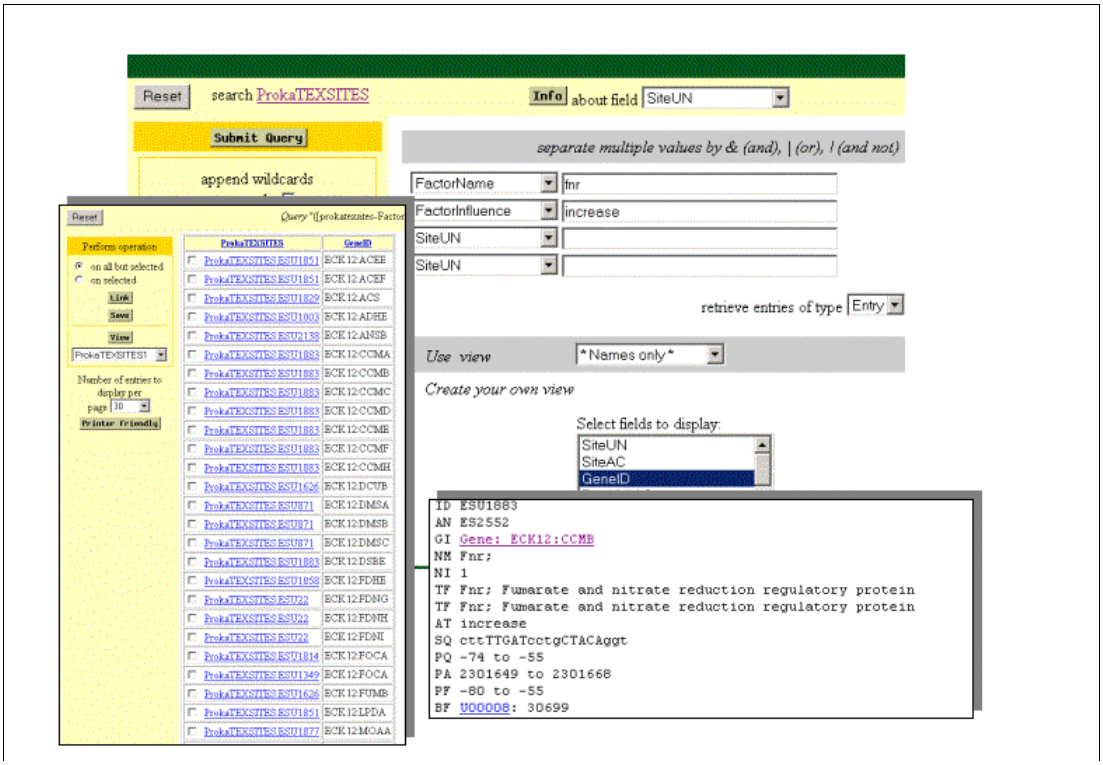


Рис. 17. Выборка сайтов транскрипционного фактора Fnr

Результатом поиска может быть список сайтов с идентификатором гена, в регуляторной области которого он обнаружен, или полное описание этих сайтов.

Пользователь может сохранить выборку сайтов Fnr, выбрав в окошке **Complete entries** и нажав кнопку **Save**, затем еще раз **Save**. Выборка может быть использована для распознавания сайтов связывания фактора Fnr.

2.2. Программные компоненты подсистемы «Регуломика»

2.2.1. Программные компоненты, позволяющие выявлять свойства регуляторных геномных последовательностей.

Построение профилей конформационных и физико-химических свойств ДНК

Программа DNAProp предназначена для построения профилей конформационных и физико-химических свойств ДНК. Конформация ДНК определяется взаимной ориентацией оснований в пределах комплиментарной пары и расположением соседних комплиментарных пар. Показано, что структура двойной спирали ДНК гетерогенна и ее конформация зависит от локального нуклеотидного контекста (Зенгер, 1987). Таким образом, конформационное состояние ДНК является функцией её контекста. Это же относится и к физико-химическим свойствам ДНК, например, к температуре плавления. Например, согласно экспериментальным данным, конформационный параметр, описывающий взаимную ориентацию соседних комплиментарных пар в спирали ДНК (угол спирального вращения твист, рис.18), специфичен для каждого динуклеотида.

Программа DNAProp использует динуклеотидные параметры 38 конформационных и физико-химических свойств ДНК (табл. 2) для вычисления профилей этих свойств для последовательностей ДНК.

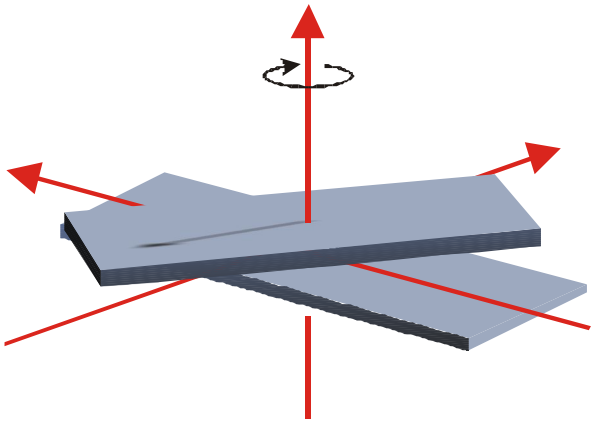


Рисунок 18. Пример конформационного свойства ДНК: «угол твист»– угол спирального вращения ДНК.

Динуклеотид	Угол твист (градусы)
AA	38.9
AT	33.81
AG	32.15
AC	31.12
TA	33.28
TT	38.9
TG	41.41
TC	41.31
GA	41.31
GT	31.12
GG	34.96
GC	38.5
CA	41.41
CT	32.15
CG	32.91
CC	34.96

Таблица 2. Динуклеотидные параметры для конформационного свойства ДНК «угол твист»– угол спирального вращения ДНК.

Описание работы DNAProp

Выявление конформационных и физико-химических свойств с использованием метода DNAProp возможно, если введена последовательность ДНК в алфавите «ATGCatgc» длины L не менее 2 нт, и не более 100.000 нт., при этом длина L также должна превосходить размер окна усреднения W (значение в поле Window size), то есть $2 \leq W \leq L$.

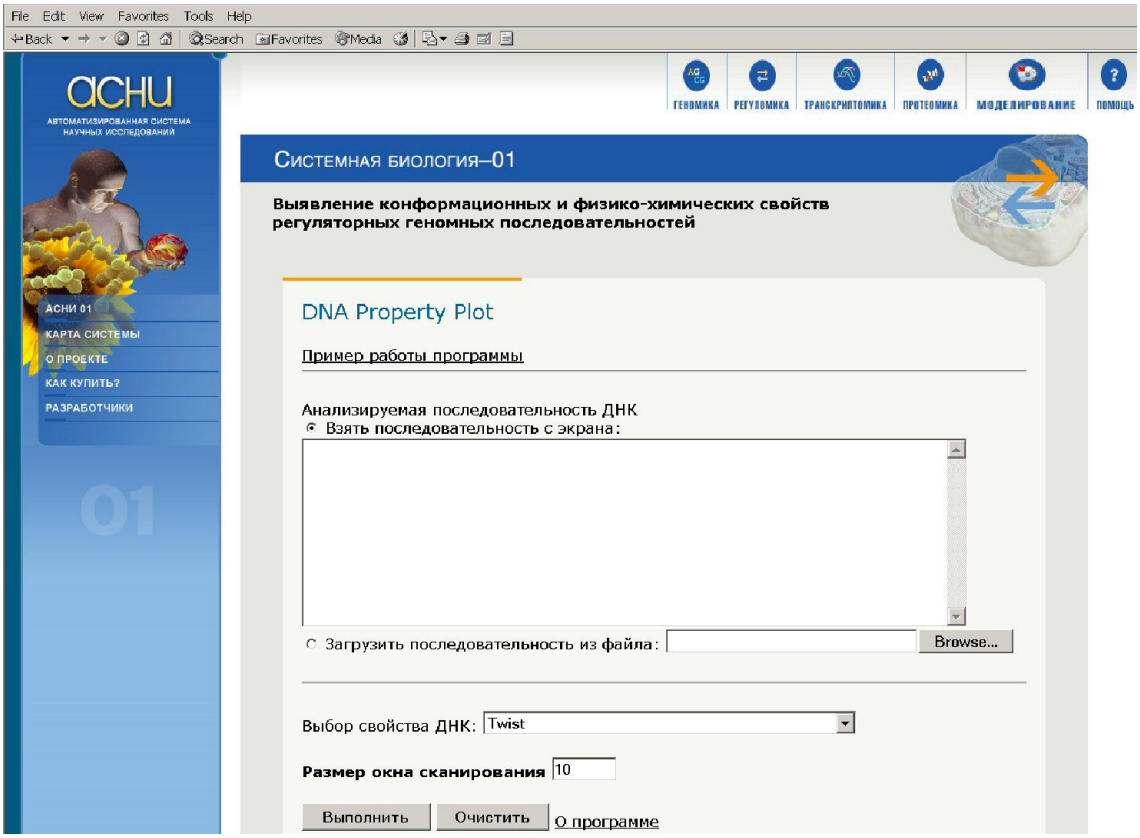


Рисунок 19. Интерфейс программы DNAProp

Необходимо войти в подсистему «Регуломика» и с помощью мыши выбрать соответствующий раздел («Конформационных и физико-химических») в подзаголовке «Выявление свойств регуляторных геномных последовательностей». В результате появится HTML-страница с интерфейсом этой операции.

Далее нужно ввести последовательности ДНК для анализа (в простом формате) в специальное окошко (Рис. 19 «Взять последовательность с экрана») или указать файл с последовательностями ДНК (кнопка ‘Browse’).

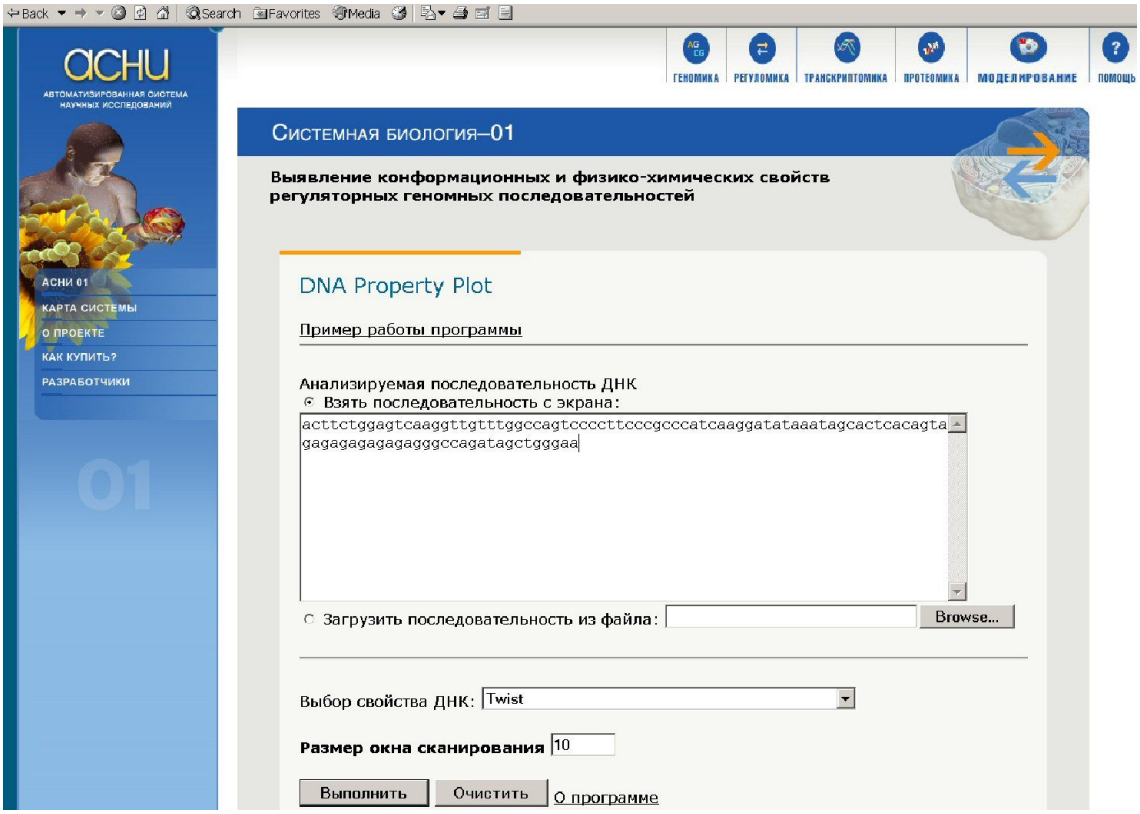


Рисунок 20. Интерфейс программы DNAProp. Заполнено поле входной последовательности.

2. Указать конформационное и физико-химическое свойство ДНК (Рис. 20, опция ‘Выбор свойства’);
3. Указать размер сканирующего окна W (Рис. 20, опция ‘Размер окна сканирования’);
4. Нажать кнопку ‘Выполнить’.

Выходные данные представляют в графическом виде профиль, который для каждой позиции в последовательности отмечает соответствующее значение свойства ДНК (Рис. 21).

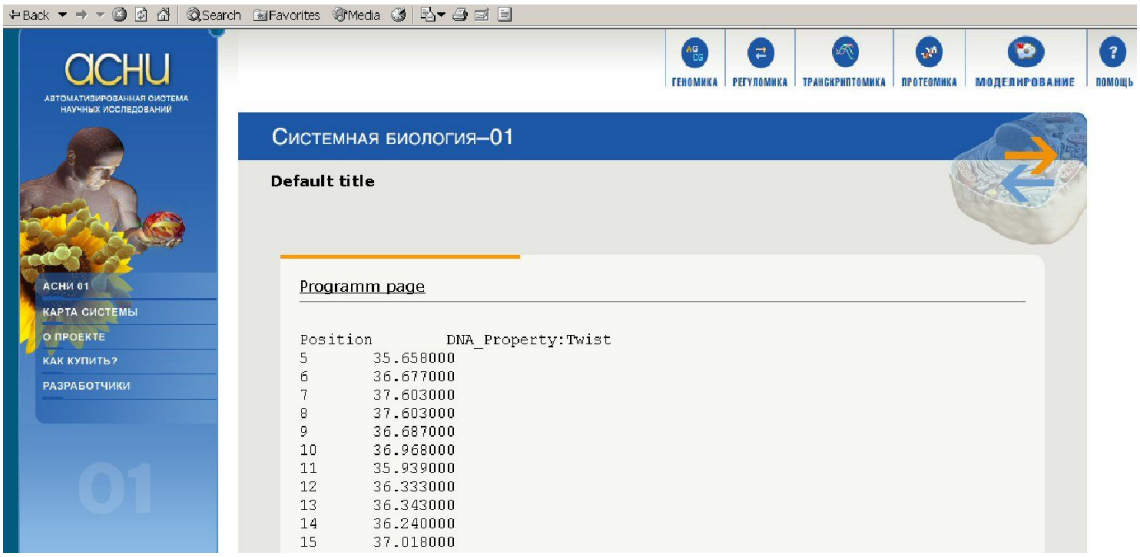


Рисунок 21. Выходные данные программы DNAProp.

Для завершения работы с текущими данными интерфейса надо нажать клавишу ‘Очистить’: в результате чего осуществляется подготовка к работе с другими данными (восстановятся все значения по умолчанию).

Выполнение операции занимает не более 1 минут для последовательности максимальной длины (до 100 000 нуклеотидов)"

Выявление повторов в геномных последовательностях

Большую роль при анализе геномов играет поиск, классификация и в дальнейшем выявление функции повторов. Повторы могут варьировать по степени несовпадений и являться совершенными или несовершенными.

Повторенная ДНК состоит из нуклеотидных последовательностей различной длины и состава, которые встречаются в геноме несколько раз. Размер части генома, занятой повторяющимися последовательностями, широко варьирует между видами. У дрожжей он достигает 20%, у млекопитающих до 60% всей ДНК повторяется. У растений процент повторенных последовательностей может превышать 80%.

По взаимной ориентации в структуре ДНК различаются прямые, инвертированные, симметричные повторы, палиндромы, комплементарные палиндромы и т.д. Также сильно варьирует и длина повторов, степень их повторяемости, и характер распределения в геноме; наконец, расположение повторов ДНК может иметь достаточно сложную структуру, когда короткие повторы находятся внутри в более протяженных или частично пересекаются с ними их и т.п.

Можно классифицировать последовательности ДНК в геноме на несколько классов по степени повторенности – от уникальной до сильно повторенной (Britten, Kohne, 1968). Важную роль в геноме могут играть палиндромные повторы, способные образовывать двуцепочечные шпильки. Наиболее высоко повторенные последовательности, как правило, имеют длину от нескольких нуклеотидов до нескольких сотен п.н., умеренные повторы составляют более длинные последовательности от сотен до тысяч п.н. С биологической точки зрения могут быть интересны также несовершенные повторы.

Известны несколько классов повторов в геномах:

Это long interspersed elements (LINE, длина 6 т.п.о.); short interspersed elements (SINE, 100-400 п.о.); LTR retrotransposons (1.5-11 т.п.о.,); BNA transposons (80-3000 п.о.,), сегментарные дупликации, дупликации генов и генные семейства, множественные копии мобильных элементов, простые повторы (микросателлиты и минисателлиты) Палидромы, внутрибелковые повторы (Heringa, 1998) и др.

Программа EAGLE предназначена для быстрого поиска повторов в геномных последовательностях. В программе реализован алгоритм быстрого поиска совпадений в протяженных нуклеотидных последовательностях на основе использования *l*-грамм (коротких слов длины *l*). Программа позволяет с высокой эффективностью проводит поиск всех четырех типов повторов с заданным количеством несовпадений, предоставляя пользователю исчерпывающие статистические данные о найденных в анализируемой последовательности повторах.

Описание работы Eagle

Выполнение операции является возможным при использовании в качестве входных данных линейной символьной последовательности ДНК, составленной из букв латинского алфавита { A, T, G, C, a, t, g, c } и имеющей длину не менее 100 и не более 10000000 символов.

Подготовка к выполнению операции: надо войти в подсистему «Регуломика» и с помощью мыши кликнуть надпись «Выявление повторов в геномных последовательностях» в списке операций этой подсистемы. В результате появится HTML-страница с интерфейсом этой операции, который показан на Рис. 22.

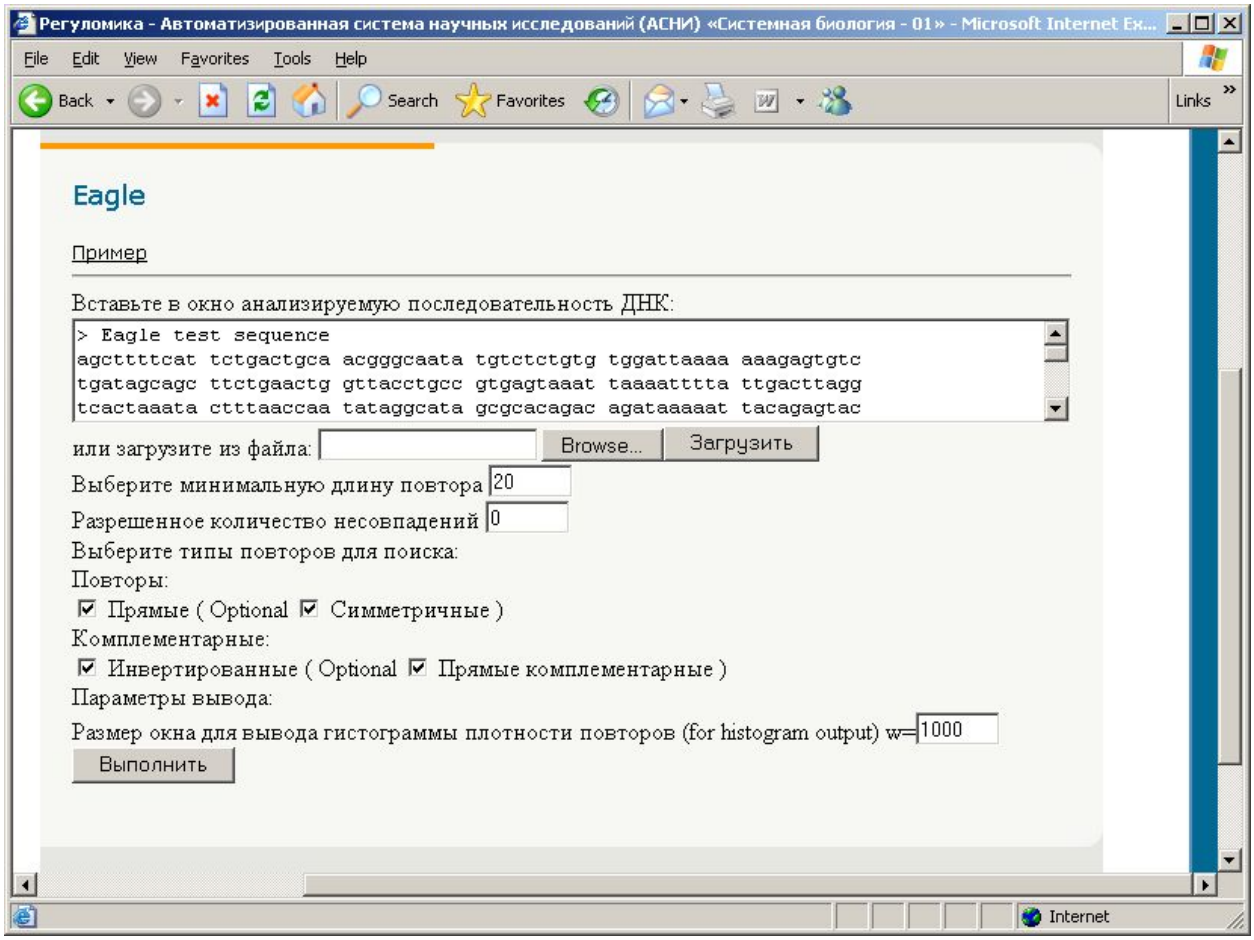


Рисунок 22. HTML-интерфейс операции, подготовленный к началу ее выполнения.

Выполнение операции начинается с ввода нуклеотидной последовательности ДНК. Для этой цели используется раздел “Вставьте в окно анализируемую последовательность ДНК” HTML-интерфейса операции, как это показано на Рис.23 для случая режима ввода данных, который установлен по-умолчанию.

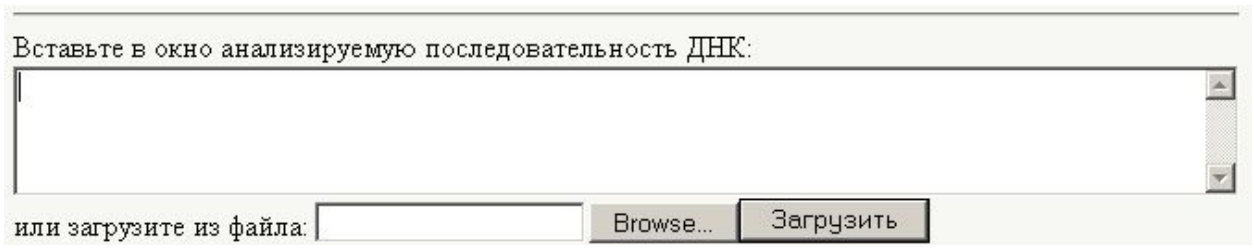


Рисунок 23. Ввод нуклеотидной ДНК в режиме "from Screen (cut & paste)", который установлен по-умолчанию.

В качестве альтернативного способа ввода анализируемой нуклеотидной последовательности из текстового файла используется ИНТЕРНЕТ-браузер путем выбора режима ввода из файла с пользовательского компьютера после нажатия клавиши “Browse”.

Необходимо выбрать минимальную длину повтора (например, 20, рис.24)

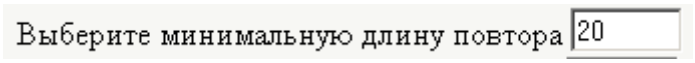


Рисунок 24 Выбор минимальной длины повтора ДНК.

Следующим действием выполнения операции является выбор типов повторов для поиска – Прямые ; Симметричные; Инвертированные и Комплиментарные (Рис. 25)

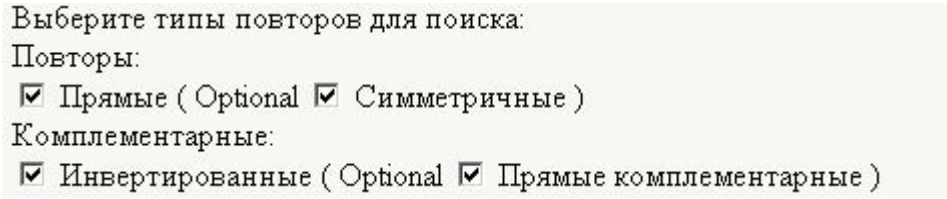


Рисунок 25. Выбор типов повторов для поиска.

После ввода данных, как было описано выше, необходимо нажать клавишу “Выполнить”, в результате чего появляется, как это показано ниже на Рис.26, HTML-страница с таблицей выявленных повторов и их статистическими данными.

Указана общая длина последовательности, параметры поиска,

таблица с количествами каждого из типов повторов и общим их числом,

таблица с максимальными длинами повторов каждого типа,

таблица длин повторов, и другие статистические данные по найденным повторам в последовательности.

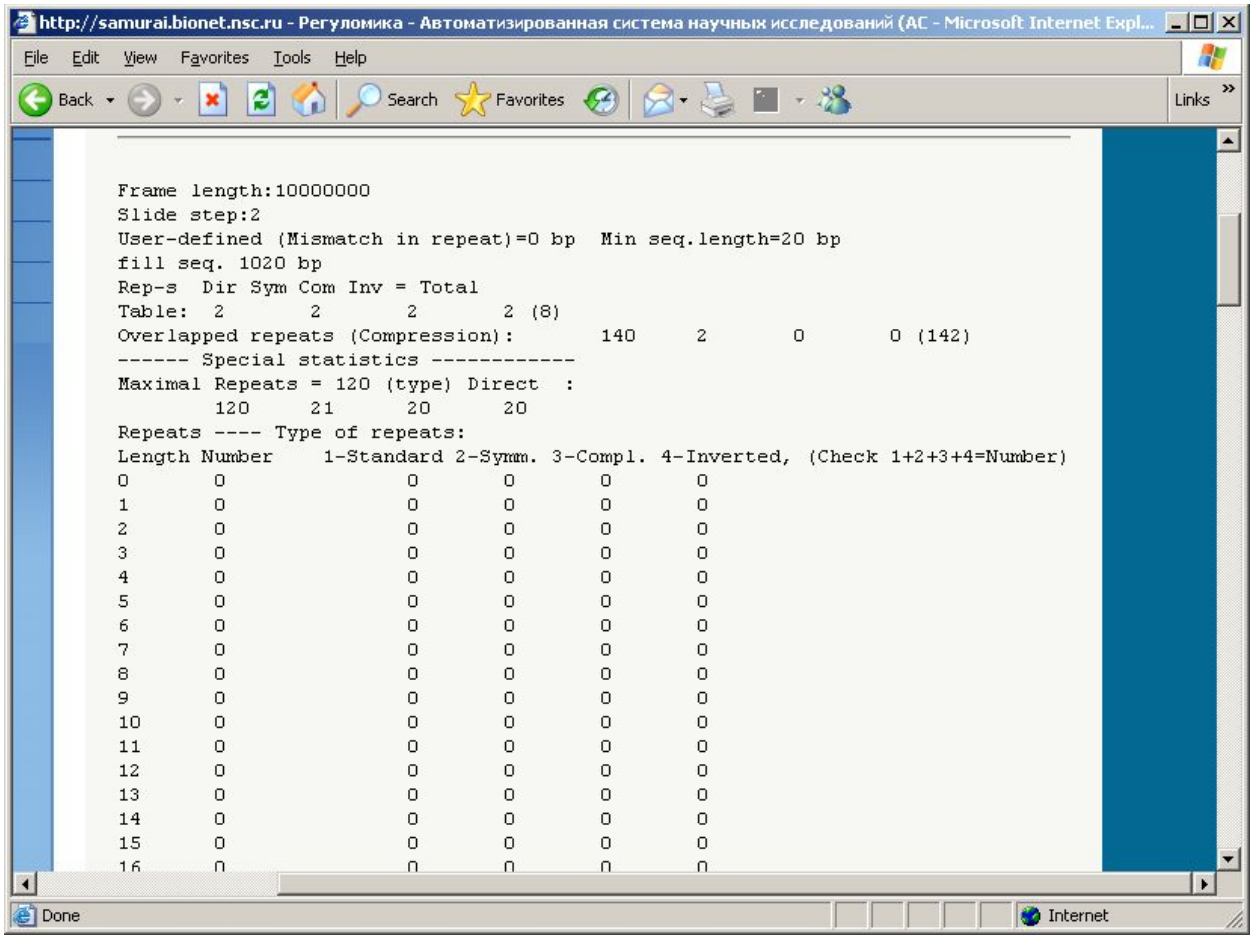


Рисунок 26. Пример HTML-страницы, содержащей таблицу выявленных повторов .

Для завершения работы с текущей (введенной) нуклеотидной последовательностью ДНК надо нажать ссылку “program page” страницы результатов, в результате чего осуществляется подготовка к повторному выполнению этой операции с другой последовательностью ДНК.

Выполнение операции занимает не более 10 минут для последовательности максимальной длины (до 10 000 000 нуклеотидов)"

Построение нуклеосомного потенциала последовательности ДНК

Характерная особенность геномов эукариот – упаковка ДНК в хроматин, базовым уровнем организации которого являются нуклеосомы (Сингер и Берг, 1998). Нуклеосомный уровень компактизации хроматина представляет собой ДНК сравнительно небольшой длины (145-147 п.о.), закрученную вокруг белковой глобулы (октамер или гистоновый кор), состоящую из восьми гистонов (по два белка H2A, H2B, H3, H4). Многочисленные данные, полученные в экспериментальных и компьютерных исследованиях показали, что расположение сайтов формирования нуклеосом (СФН) в геномной ДНК не случайно, а контролируется особым контекстно-конформационным кодом (van Holde, 1989).

Метод распознавания СФН Recon основан на учёте частот динуклеотидов. Рассмотрение динуклеотидов обусловлено тем, что их распределение вдоль нуклеотидной последовательности обеспечивает возникновение локальных конформационных особенностей двойной спирали ДНК, значимых для функционирования геномных

последовательностей. Поэтому предполагается, что рассмотрение динуклеотидов позволит косвенным способом учесть те конформационные особенности ДНК СФН, которые определяют структуру двойной спирали, значимую для ДНК-гистоновых взаимодействий. При построении метода распознавания Recon осуществляется поиск разбиения СФН на отдельные участки, которые характеризуются более гомогенным нуклеотидным контекстом по сравнению с контекстом всего СФН. При поиске оптимального разбиения сначала случайным образом задается произвольное разбиение (Рис. 27А), которое затем подвергается последовательным изменениям (Рис. 27Б-Д).

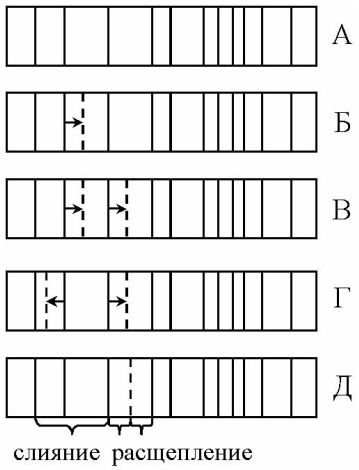


Рисунок 27. Типы элементарных операций, используемых при поиске оптимального разбиения функционального района на участки. Пунктиром отмечены положения изменившихся границ разбиения.
А) начальное разбиение; Б) сдвиг одной границы; В) сдвиг участка; Г) центрально симметричный сдвиг границ участка; Д) расщепление и слияние.

В качестве параметра оценки качества разбиения использовано расстояние R^2 Махаланобиса, рассчитываемое по динуклеотидным частотам для участков разбиения. По полученному разбиению с помощью методики дискриминантного анализа строится функция распознавания СФН.

Описание работы Recon

Анализ и распознавание сайтов формирования нуклеосом с помощью метода Recon возможно, если ведена нуклеотидная последовательность ДНК длины L не менее 160 нт в алфавите «ATGCatgc» и не более 100 000 нт. Допускаются также символы “Nn”, но вблизи их сосредоточения сайты не могут быть предсказаны, то есть данные символы игнорируются программой.

Необходимо войти в подсистему «Регуломика» и выбрать надпись «Анализ и распознавание сайтов формирования нуклеосом» в списке операций этой подсистемы. В результате появится HTML-страница с интерфейсом этой операции, который показан на Рис. 28.

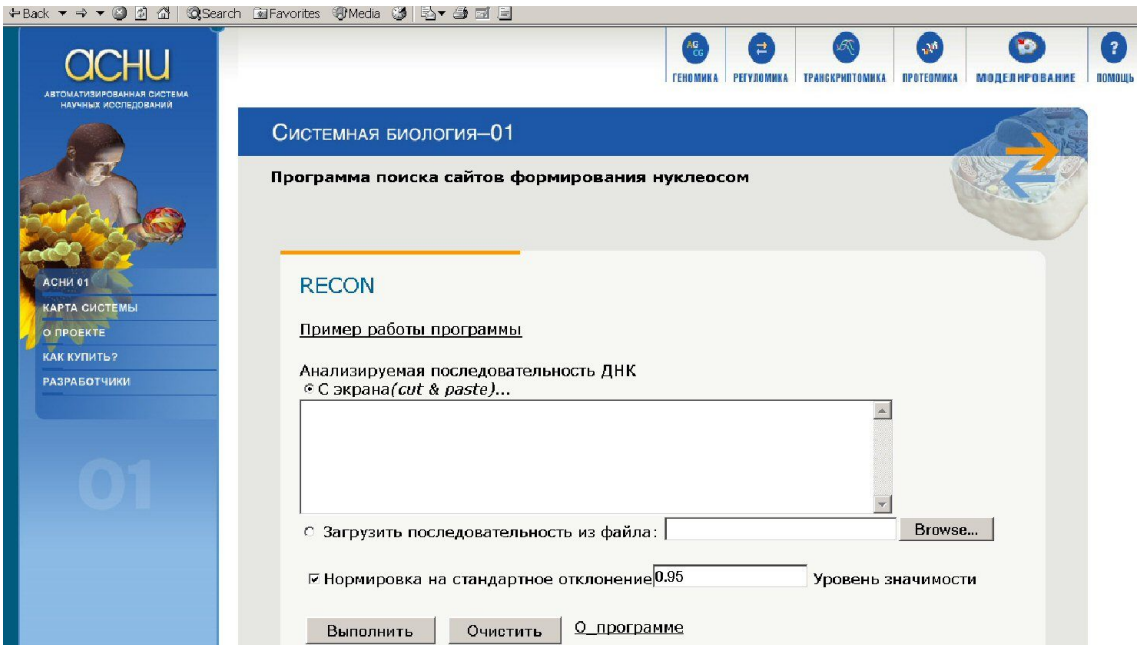


Рисунок 28. Интерфейс программы Recon

Далее нужно ввести последовательности ДНК для анализа (в простом формате) в специальное окошко (Рис. 29, «Анализируемая последовательность ДНК С экрана(cut & paste)...») или указать файл с последовательностями ДНК (кнопка ‘Browse..’);

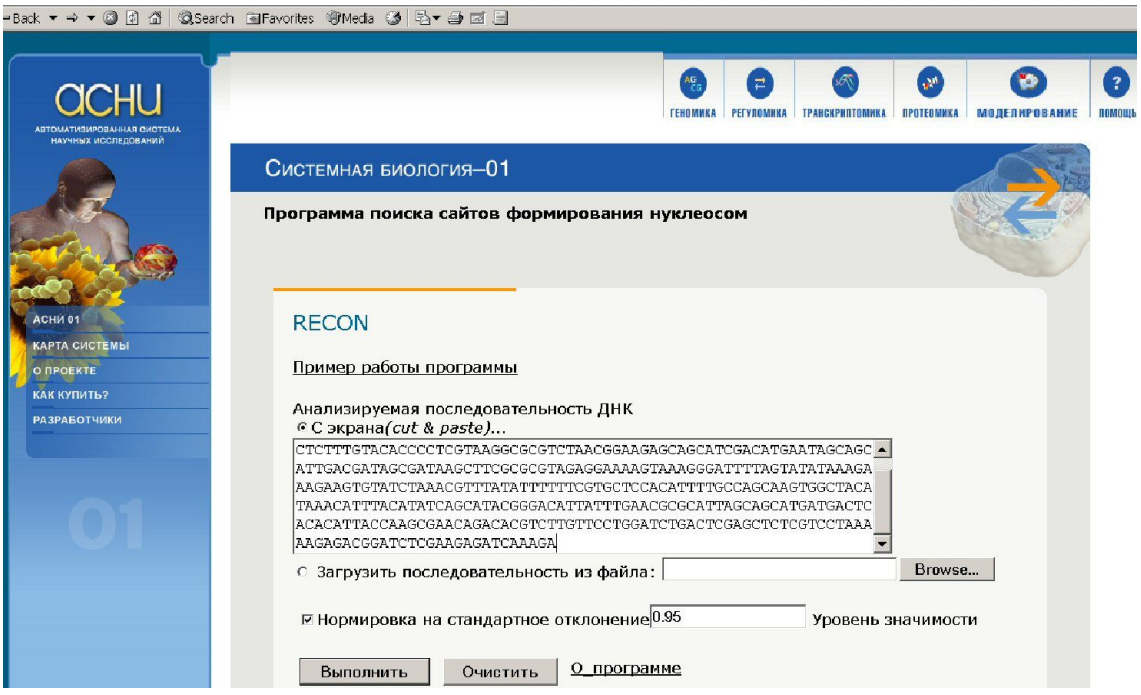


Рисунок 29. Интерфейс программы Recon. Заполнено поле входной последовательности.

Затем нужно указать включение или выключение нормировки выходного профиля по стандартному отклонению (опция ‘Нормировка на стандартное отклонение’), при включении этой опции максимальное значение в выходном профиле +1 соответствует наибольшей вероятности формирования нуклеосомы с центром в заданной позиции, иначе – этому

событию соответствуют более близкие к +1 значения, а значения профиля не ограничены сверху значением +1;

Также нужно указать порог достоверности предсказаний (Уровень значимости, оставить приведенное по умолчанию значение 0.95);

Далее нажать кнопку ‘Выполнить’, получаются выходные данные в следующем виде (Рис. 30).

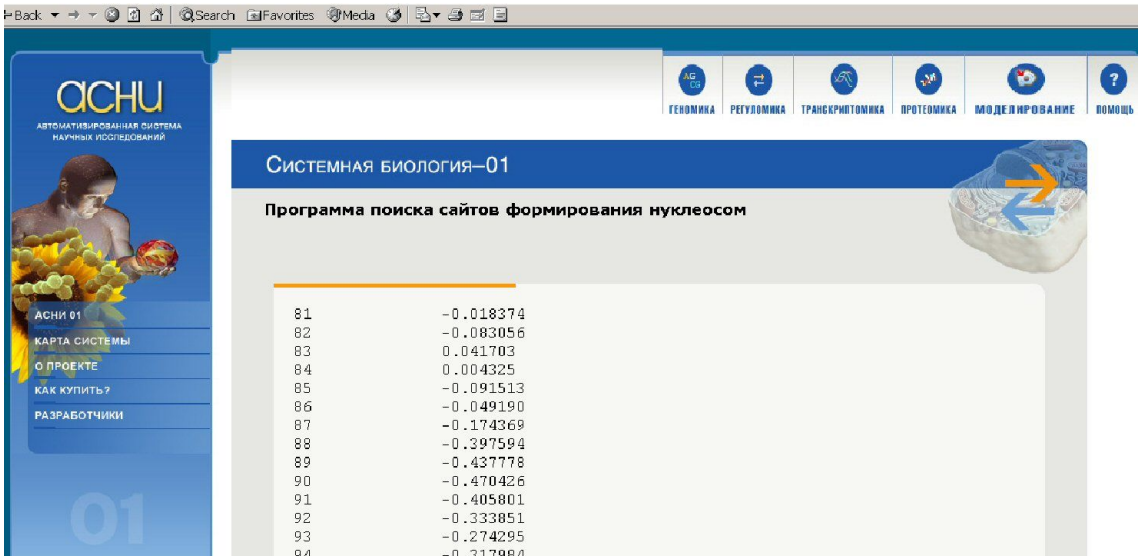


Рисунок 30. Выходные данные программы Recon.

Для завершения работы с текущими данными интерфейса надо нажать клавишу ‘Очистить’: в результате чего осуществляется подготовка к работе с другими данными (восстановятся все значения по умолчанию).

Выполнение операции занимает не более 1 минуты для последовательности максимальной длины (100 000 нуклеотидов)

2.2.3. Программные компоненты для анализа структурно-функциональной организации промоторов коэкспрессирующихся генов

Обнаружение иерархических комплексных сигналов, построение модели регуляторных районов

Промоторы генов можно рассматривать как молекулярные переключатели, которые активируют («включают») либо приостанавливают транскрипцию генов эукариот (Zhang, 1998). Промоторы содержат сайты связывания транскрипционных факторов (ССТФ) - короткие участки ДНК, достаточно консервативные для специфичного распознавания соответствующими белковыми транскрипционными факторами. Присутствие и локализация ССТФ в регуляторных районах генов соответствует ткане- и стадиеспецифичным характеристикам генной экспрессии в организме. Встает задача определения как самих ССТФ в последовательности ДНК, так и предсказания расположения промотора, определяющего регуляцию экспрессии гена по паттерну расположения определенных ССТФ (Qiu, 2003).

В рамках подхода “Relational Data Mining” к методам извлечения знаний (Kovalerchuk and Vityaev, 2000; Kovalerchuk *et al.*, 2001, Vityaev E., Kovalerchuk B. (2004); Vityaev and Kovalerchuk 2005a, 2005b; см. также веб-сайт «Scientific Discovery»

http://www.math.nsc.ru/AP/ScientificDiscovery) разработана интерактивная система ExpertDiscovery, позволяющая пользователю (эксперту-биологу) иерархически конструировать комплексные сигналы, просматривать расположение этих сигналов в последовательностях ДНК и определять их статистические параметры на анализируемой контрастной выборке данных. Комплексные сигналы определяются рекурсивно на основе первичных сигналов.

Первичными сигналами *S* могут быть:

(1) контекстные сигналы - последовательности символов в 15-и буквенном алфавите (олигонуклеотиды) (Vishnevsky and Kolchanov, 2005);

(2) потенциальные функциональные сайты, предсказанные по гомологии (или посредством весовой матрицы) с аннотированными последовательностями из специализированных молекулярно-биологических баз данных (Поздняков и др., 2001);

(3) участки с консервативными для сайтов связывания конформационными или физико-химическими характеристиками (углы двойной спирали между соседними нуклеотидами, температура плавления ДНК) (Oshchepkov *et al*., 2004);

Комплексный сигнал описывается иерархически и по индукции:

- первичный сигнал является комплексным сигналом;

- ориентация комплексного сигнала (прямая, симметричная, инвертированная) является комплексным сигналом;

- повторение комплексного сигнала *N* раз ($2 \le N_{min} \le N \le N_{max}$) является комплексным сигналом. При этом расстояние между соседними копиями сигнала принадлежит заданному пользователем диапазону;

- принадлежность комплексного сигнала некоторому интервалу относительно старта транскрипции (либо начала фазированной последовательности) является комплексным сигналом;

- упорядоченная пара комплексных сигналов S_1 и S_2 с расстоянием между ним, варьирующим в определенном интервале является комплексным сигналом;

Полученные в процессе работы системы комплексные сигналы эффективно применяются для распознавания регуляторных районов генов эукариот.

Пример комплексного сигнала, обнаруженного программой ExpertDiscovery, и его визуализация приведёны на Рис.31. Промоторные последовательности фазированы относительно старта транскрипции, идентификационные номера генов, соответствующие базе данных TRRD (Kolchanov *et al*., 2002), приведены слева.

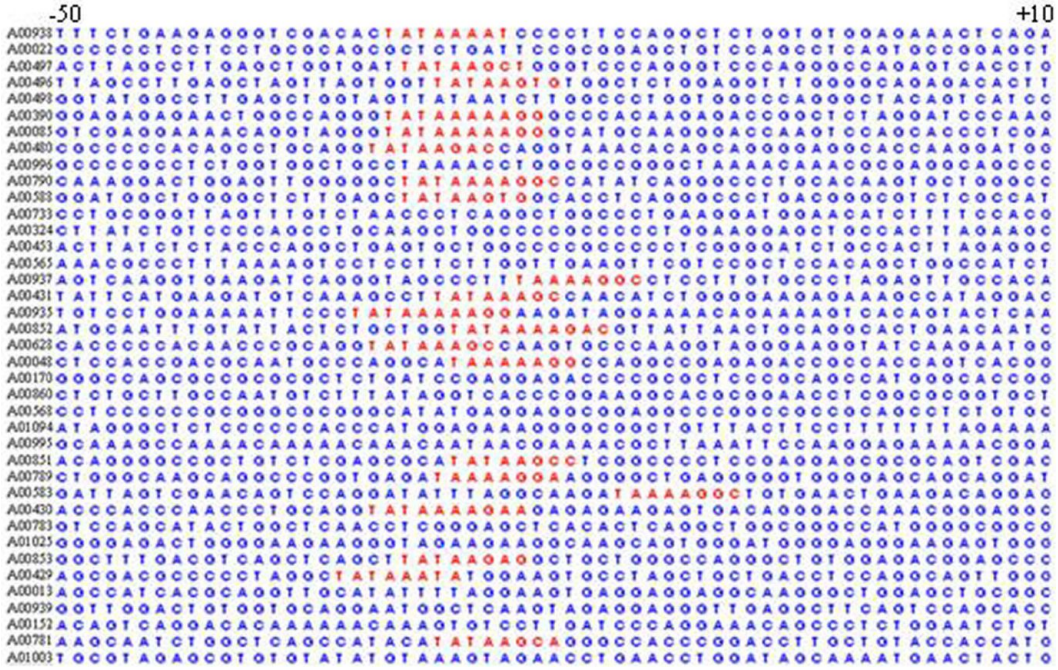


Рис. 31. Комплексный сигнал, соответствующий расположению TATA-бокса в промоторах генов эндокринной системы. Слева приведены идентификационные номера генов, соответствующие базе данных TRRD.

Описание работы ExpertDiscovery

Входными данными для программы являются две выборки последовательностей в формате FASTA: Positive (выборка промоторных районов, Negative (выборка случайных последовательностей, или промоторов других групп), и разметка последовательностей сигналами DBRec или ArgoMotifs.

Для загрузки обучающих данных используется команда меню File->New или кнопка  панели инструментов. Процесс загрузки производится с помощью мастера. На первой странице (step 1) приведено пояснение последующих шагов. На второй странице (step 2) предлагается указать имя файла с позитивной выборкой последовательностей (Рис. 32). На третьей – имя аналогичного файла с негативной выборкой.

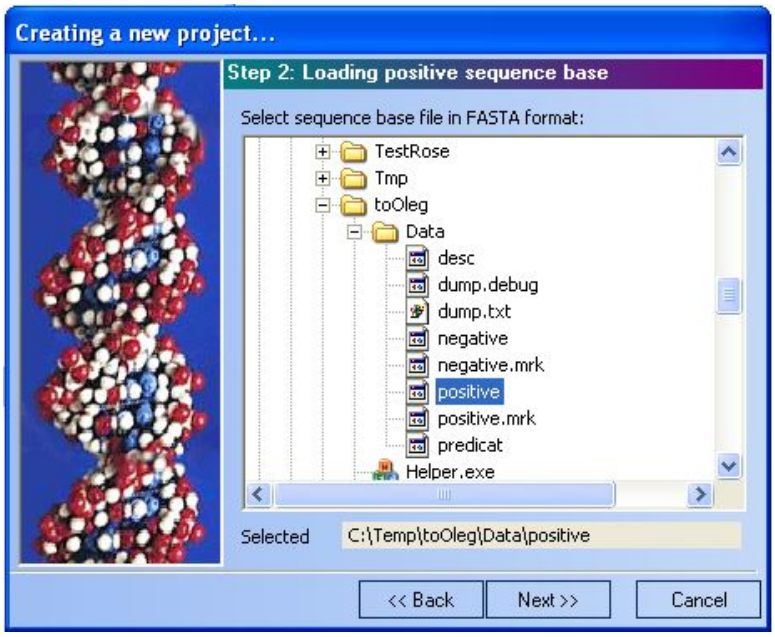


Рис. 32. Загрузка файла с позитивной выборкой последовательностей

Загрузка разметки обучающих последовательностей.

Сразу после загрузки последовательностей на экране появится мастер для загрузки разметки. Можно отказаться от его услуг, нажав кнопку Cancel. Разметка может быть загружена отдельно при помощи команды контекстного меню, всплывающего при щелчке правой кнопкой мыши по группе Markup панели проекта (рис. 33).

Рис. 33. Контекстное меню загрузки разметки.

Мастер загрузки разметки содержит четыре шага. На первом шаге содержится описание последующих шагов. На втором требуется указать файл с разметкой позитивной последовательности. На третьем – негативной последовательности. На последнем шаге требуется указать файл с описанием имён внешних сигналов, либо выбрать опцию для автоматической генерации этой информации по разметке.

В случае разметки DBREC размеченные обучающие последовательности сразу загружаются с помощью *File<Convert DBRec.* (см. контрольный пример)

Создание и редактирование комплексного сигнала.

Для создания комплексного сигнала необходимо воспользоваться контекстным меню, которое появляется при щелчке правой кнопкой мышки на группе «*Complex signals*» панели проекта или одной из папок в этой группе. В меню требуется выбрать пункт «*New signal*» (рис. 34). Появится новый сигнал с именем «*NewSignalXX*», где XX – число необходимое для обеспечения уникальности имени сигнала.

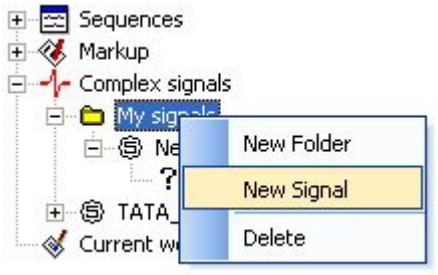


Рис. 34. Контекстное меню для создания нового сигнала

В силу своего определения комплексный сигнал представим в виде дерева, в узлах которого стоят операции, а листья являются терминальными символами. Именно эта форма представления была взята за основу интерфейса. Вначале работы дерево содержит корень с именем сигнала и один неопределённый элемент. После щелчка на неопределённом элементе на панели свойств можно установить тип этого элемента и его параметры (Рис.35).

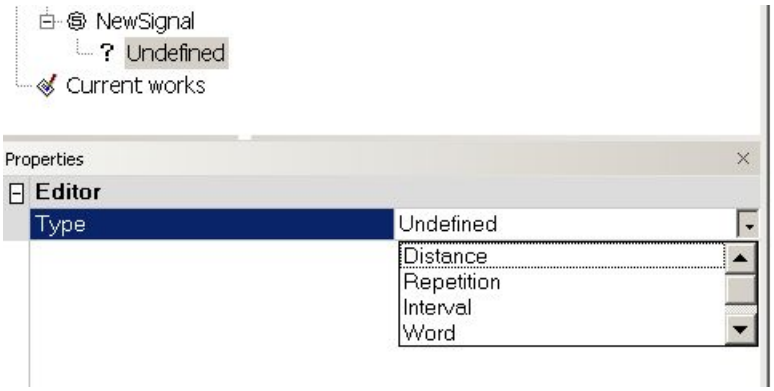


Рис. 35. Установление типа неопределённого элемента.

В программе присутствуют следующие типы узлов комплексного сигнала:

- Операция дистанция (Distance). На панели свойств для этой операции можно установить способ измерения дистанции (Distance type), диапазон расстояний которые могут быть между аргументами операции (Distance from и Distance to) и указать, важен ли порядок аргументов (Order).
- Операция повторения (Repetition). Свойствами этой операции являются искомое количество повторений (от Count from до Count to), дистанция между этими повторениями (от Distance from до Distance to) и способ измерения дистанции (Distance type).

- Операция интервал (Interval). Для этой операции на панели свойств устанавливается диапазон позиций последовательности (от Interval from до Interval to), в которых ищется аргумент операции.
- Слово (Word) в 15-ти буквенном коде.
- Элемент разметки (Markup Item). Соответствует терминальному символу, являющемуся внешним сигналом. Параметрами является семейство (Family) внешнего сигнала и его имя (Name).

Если тип элемента – слово, или элемент разметки, то эта ветвь дерева закончилась. В случае унарной операции текущий узел получает ещё одного неопределённого потомка, в случае бинарной операции – двоих. Так продолжается до тех пор, пока комплексный сигнал полностью не определён, т.е. пока не закончатся неопределённые элементы дерева (Рис.36).

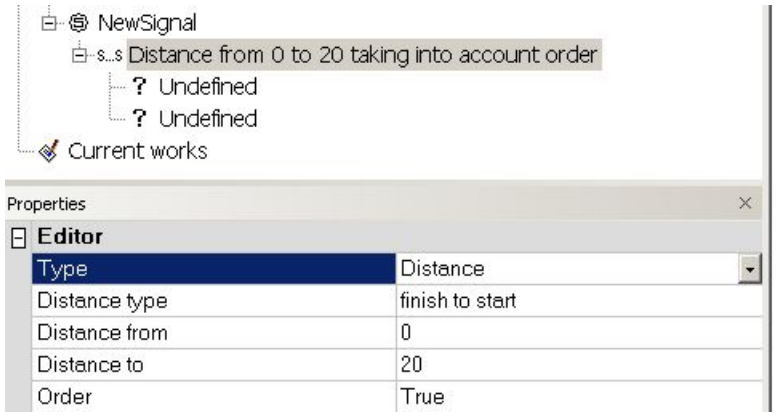


Рис. 36. Задание бинарной операции Distance.

Для автоматического построения комплексных сигналов используется пункт меню *Edit->Extract signals...* или кнопка  панели инструментов. Для запуска процесса появляется мастер, в котором на первом шаге устанавливаются параметры отбора результатов и отсева неперспективных гипотез. Параметры ветвления сигнала вынесены на вкладку Advanced (рис. 37).

При отборе комплексного сигнала оценивается:

- условная вероятность сигнала (Condition probability level)
- покрытие позитивной выборки (Coverage bound)
- значимость сигнала по критерию Фишера (Fisher criteria level)

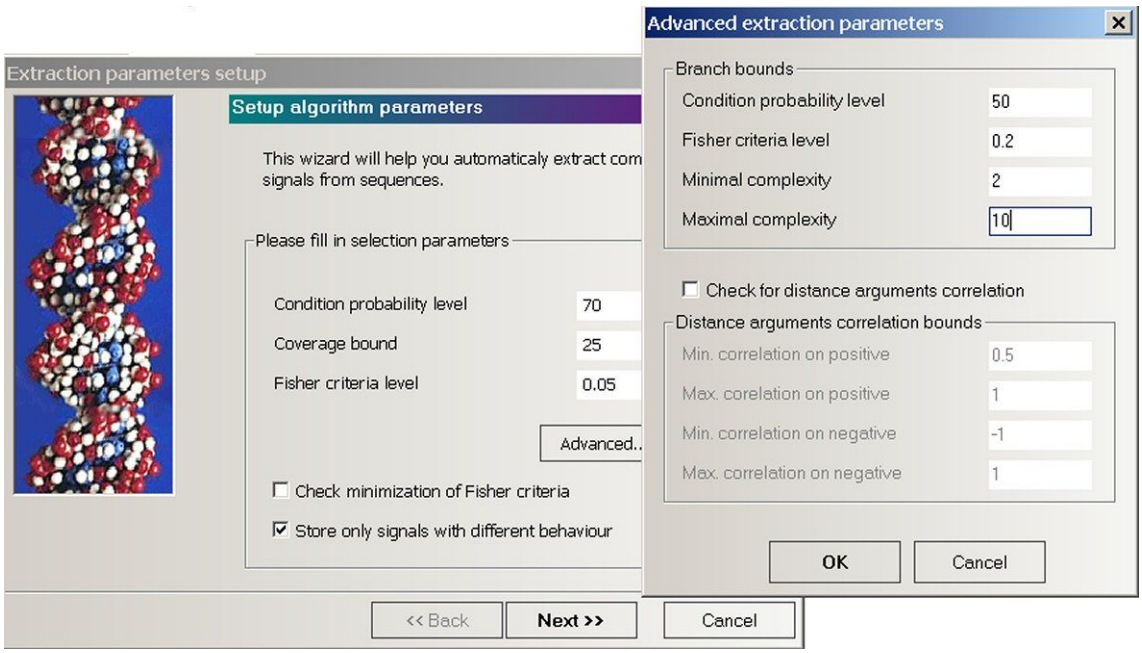


Рис. 37. Установка параметров сигнала и параметров ветвления.

Параметры ветвления:

- \- условная вероятность ветвления (Condition probability level)
- значимость сигнала по критерию Фишера (Fisher criteria level)
- минимальная сложность сигнала (Minimal complexity)
- максимальная сложность сигнала: (Maximal complexity)

На втором шаге требуется задать предикаты, которые будут использоваться при построении комплексных сигналов (рис. 38).

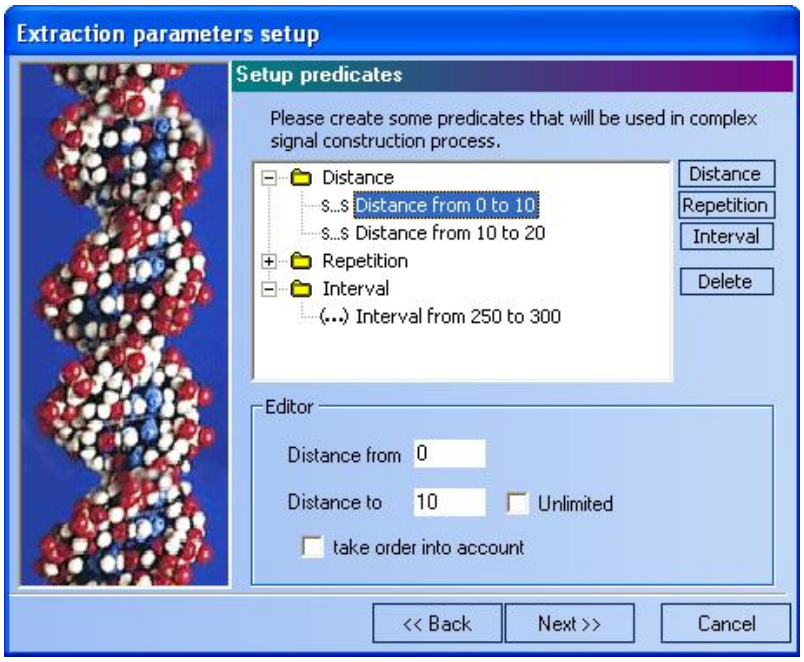


Рис. 38. Окно для спецификации предикатов

Чтобы задать предикат, необходимо создать его, используя одну из кнопок Distance, Repetition или Interval. После чего установить в окне Editor его параметры и перейти к спецификации следующего предиката.

На последней странице мастера требуется указать папку из ветки «*Complex signals*» проекта, в которую будет помещаться результат.

После завершения мастера в списке текущих задач появится новая задача с именем Extracting signals (x%), где x – процент завершения этой задачи (Рис. 39).

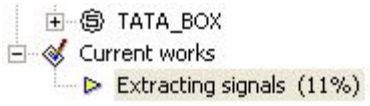


Рис. 39. Отображение созданного процесса в списке текущих работ

Применение совокупности комплексных сигналов, получение отчёта о распознавании.

Для применения полученной совокупности комплексных сигналов к обучающей позитивной и негативной выборке, и контролю, нужно нажать правой кнопкой мыши на папку, в которую происходила генерация сигналов, и выбрать опцию “Select all signals” (Рис. 40)

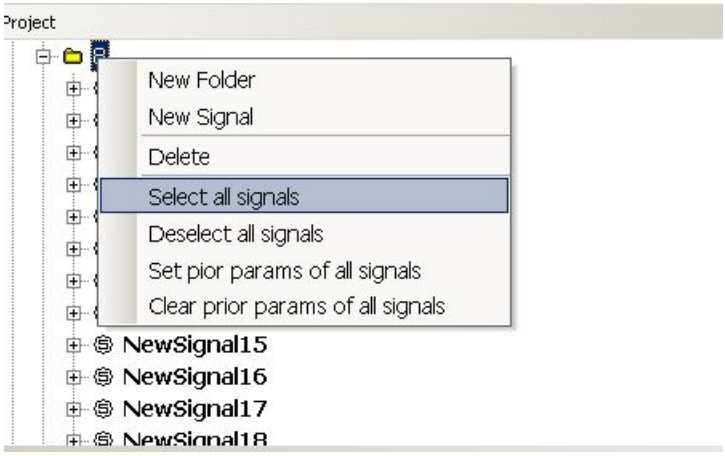


Рис. 40 Выделение комплексных сигналов.

Процедура выбора порога распознавания реализуется с помощью кнопки  панели инструментов. При нажатии этой кнопки появляется мастер, который позволяет менять порог распознавания, указывая на зависимость ошибок первого и второго рода (Рис. 41).

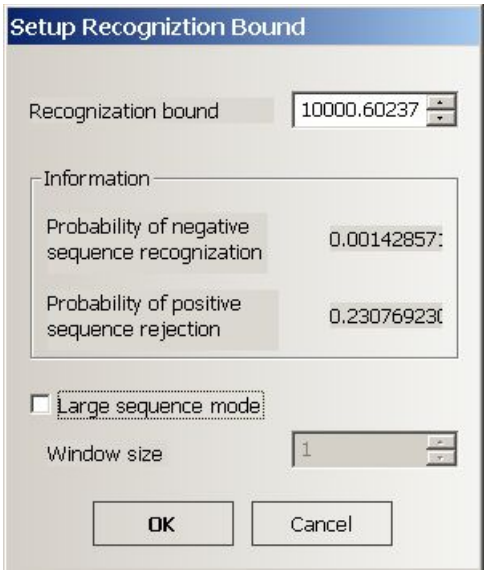


Рис. 41 Выбор порога распознавания.

Генерация отчёта о распознавании осуществляется с помощью кнопки  панели инструментов. Программа сохраняет отчёт в HTML формате (Рис. 42).

Positive base

Total sequences: *39*
Recognized sequences: *36*
Details:

Sequence No	Sequence Name	Score	Result
1	A00938_P01203__M22853__promote	3571.88	Recognized
2	A00022_P00562__X77137__Promote	15896.9	Recognized
3	A00497_P00686__J05245__promote	34771.7	Recognized
4	A00496_P00688__M63125__promote	4222.11	Recognized
5	A00498_P00685__D00161__promote	12609.6	Recognized

Рис. 42 Отчёт о распознавании.

Рассмотрим работу программы на примере выявления комплексных сигналов в группе промоторов интерферон-регулируемых генов

Для этого на вход программе в качестве позитивной выборки подаётся группа интерферон-регулируемых промоторов (promoter inter), в качестве негативной – группы промоторов клеточного цикла (promoter cell cycle) и липидного метаболизма (promoter LM). С помощью кнопки *File<Convert DBRec* программа позволяет выбрать и загрузить размеченные последовательности промоторов (Рис. 43).

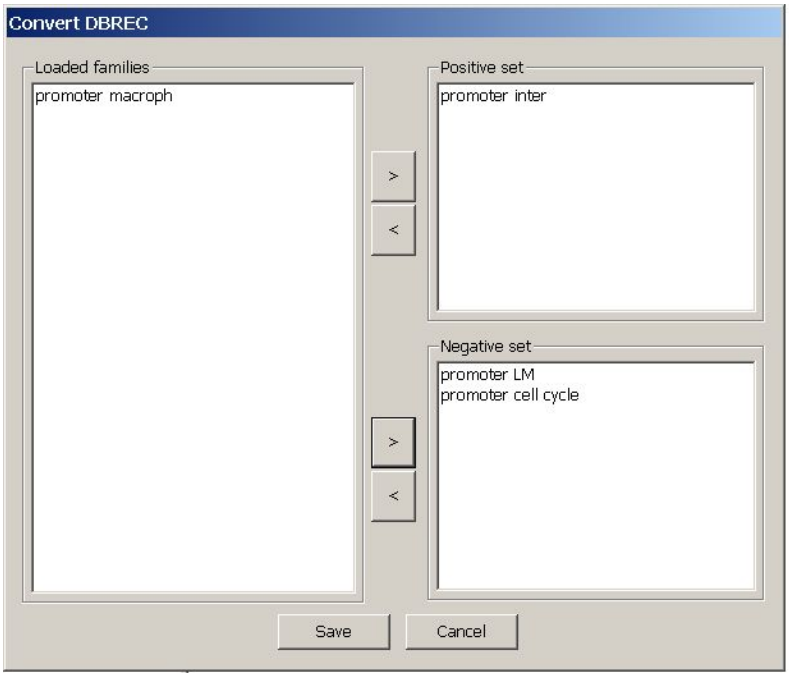


Рис. 43 . Загрузка позитивной и негативной выборки обучающих последовательностей, размеченных DBREC.

Создание и редактирование комплексного сигнала осуществляется в соответствии с описанием в пункте 2 из «Руководства пользователя».

Для инициализации процесса генерации комплексных сигналов требуется создать папку в ветке «*Complex signals*» проекта (см. пункт 2, Руководство пользователя), в которую будет помещаться результат, и задать параметры отбора результата, и параметры ветвления (Рис. 44).

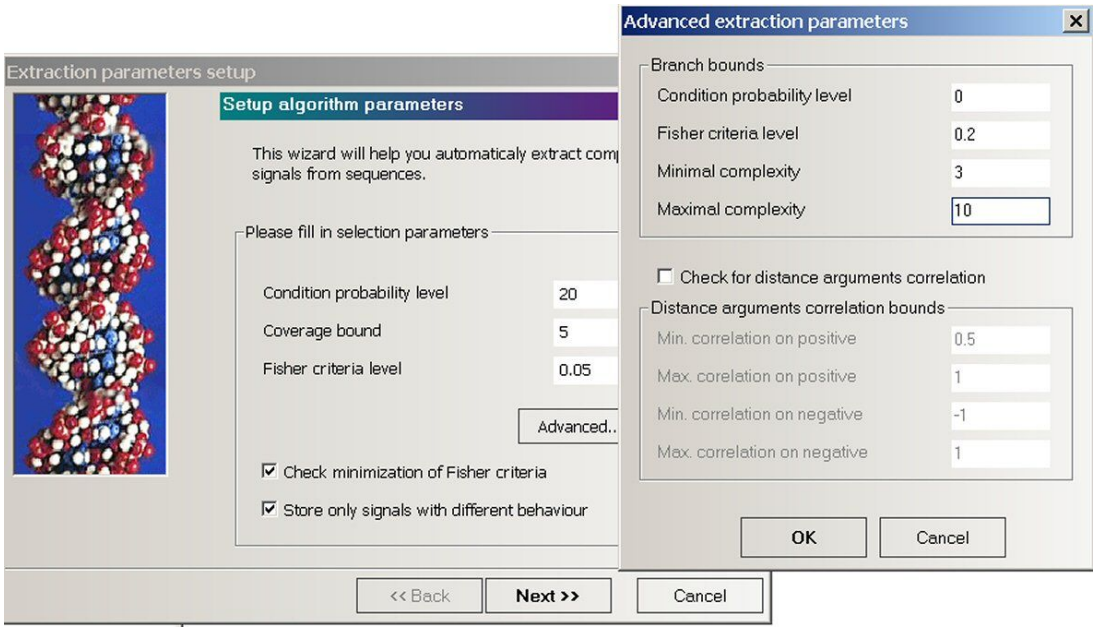


Рис. 44. Установка параметров сигнала и параметров ветвления.

На втором шаге требуется задать предикаты (Смотри пункт 4, Руководства пользователя), которые будут использоваться при построении комплексных сигналов (рис. 45).

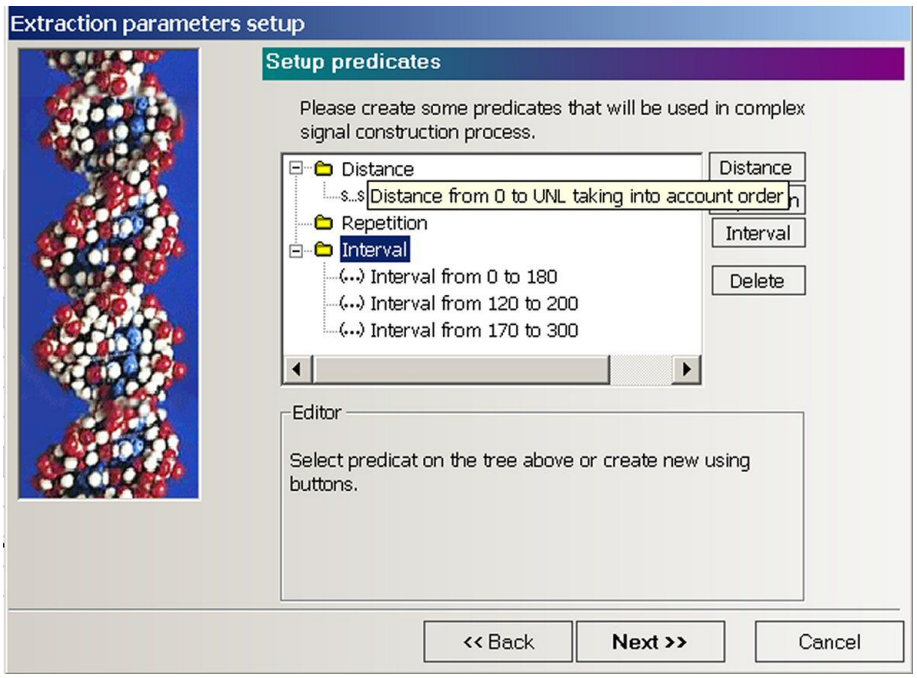


Рис. 45. Спецификация предикатов.

Применение совокупности комплексных сигналов, получение отчёта о распознавании. Пример сигнала из полученной совокупности, и его параметры приведёны на рисунке 46.

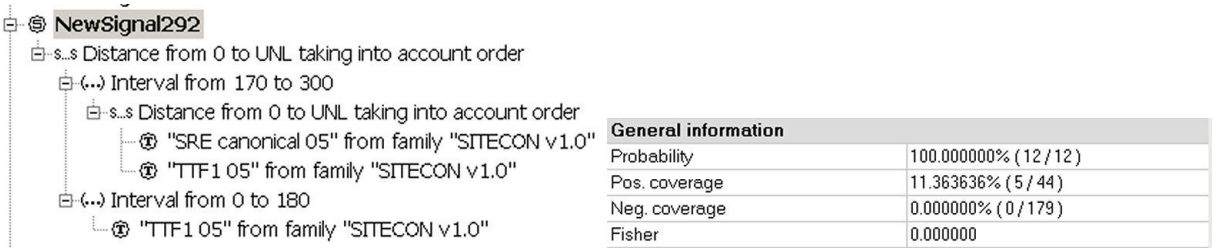


Рис. 46. Иерархическое дерево и основные параметры сигнала.

На рисунке 47 показано, как данный сигнал проецируется на позитивную выборку.

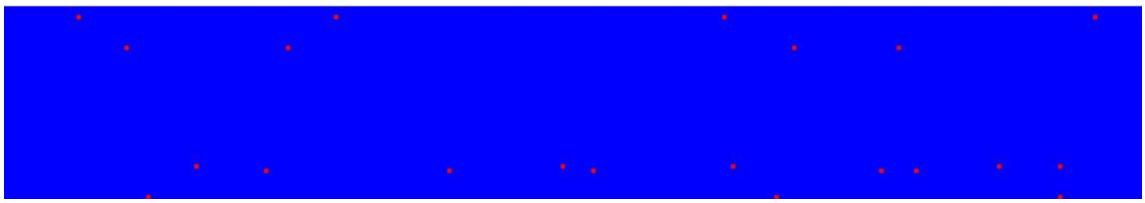


Рис. 47. Отображение сигнала на позитивную выборку.

Выявление консервативных мотивов

Выявление функциональных сигналов необходимо для понимания структурно-функциональной организации регуляторных районов генов эукариот и для разработки методов предсказания структуры гена. Было показано, что учет особенностей распределения олигонуклеотидов вдоль регуляторных районов является эффективным способом повышения точности их распознавания (Fickett, J.W. and Hatzigeorgiou, A.C. 1997). Следует отметить, что выявляемые значимые олигонуклеотиды часто имеют контекст, сходный с известными сайтами связывания белков. Заметим также, что олигонуклеотидный состав, характерный для различных регуляторных районов, может быть обусловлен не только присутствием в этих районах сайтов связывания, но может определяться некоторыми контекстно-зависимыми особенностями локальной конформации ДНК. Итак, олигонуклеотидные мотивы являются достаточно общим типом контекстных сигналов, способных кодировать как сайты связывания, так и конформационные особенности регуляторных районов генов. Нами разработан метод выявления функциональных сигналов, основанный на кластеризации схожих совершенных олигонуклеотидов, входящих в состав различных регуляторных последовательностей с итерационным построением для каждого класса олигонуклеотидов результирующего консенсуса (Vishnevsky O.V., and Kolchanov N.A., 2005).

Описание работы ARGO_Motifs

Выявление мотивов программой ARGO_Motifs может проводиться в наборе из не более чем 100 анализируемых последовательностей ДНК одинаковой длины длиной не более 100 нуклеотидов. Анализируемая выборка последовательностей должна быть в FASTA- формате.

Необходимо войти в подсистему «Регуломика» и выбрать функцию «Выявление консервативных мотивов в выборках последовательностей ДНК» в списке операций этой подсистемы. В результате появится HTML-страница с интерфейсом этой операции, который показан на Рис. 48.

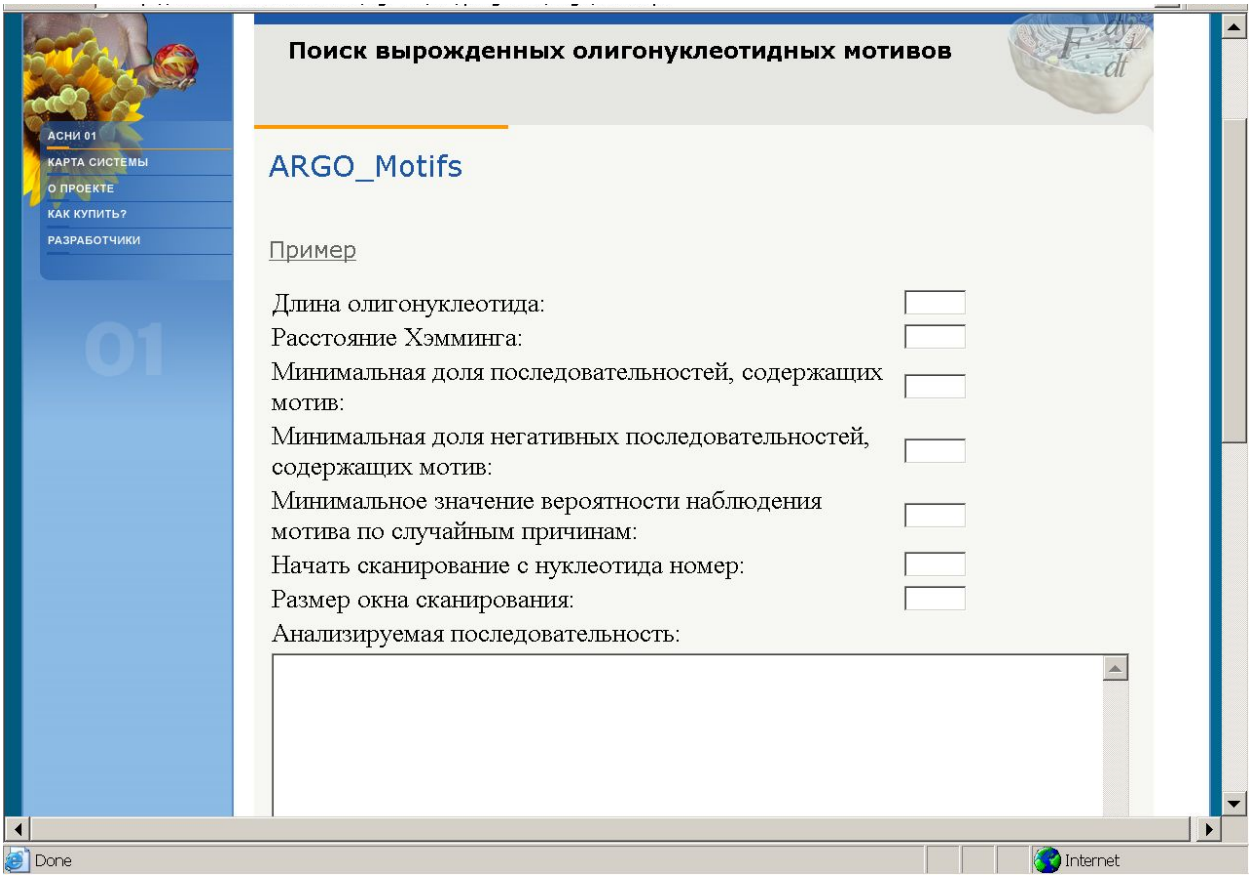


Рисунок 48. Web-интерфейс ввода программы выявления вырожденных олигонуклеотидных мотивов ARGO_Motifs.

Выполнение операции начинается с ввода набора анализируемых последовательностей ДНК. Выборка должна быть записана в FASTA-формате и содержать последовательности одинаковой длины (Рис. 49). Далее необходимо задать следующие параметры:

1. длину олигонуклеотидов, которые будут выявляться;
2. расстояние Хэмминга, которое рассчитывается как количество несовпадающих букв в олигонуклеотидах кластера.
3. минимальную долю анализируемых последовательностей, содержащих найденный олигонуклеотидный мотив.
4. минимальное значение вероятности наблюдения выявленного вырожденного олигонуклеотидного мотива в анализируемой выборке по случайным причинам (граничного значения десятичного логарифма биномиальной вероятности)
5. начало анализируемого окна в выборке.
6. размер окна сканирования
7. код (либо вырожденный 15-ти буквенный IUPAC код, либо совершенный 4-х буквенный).
8. указать, нужно ли проводить поиск мотивов в том числе и в комплементарных цепях последовательностей.

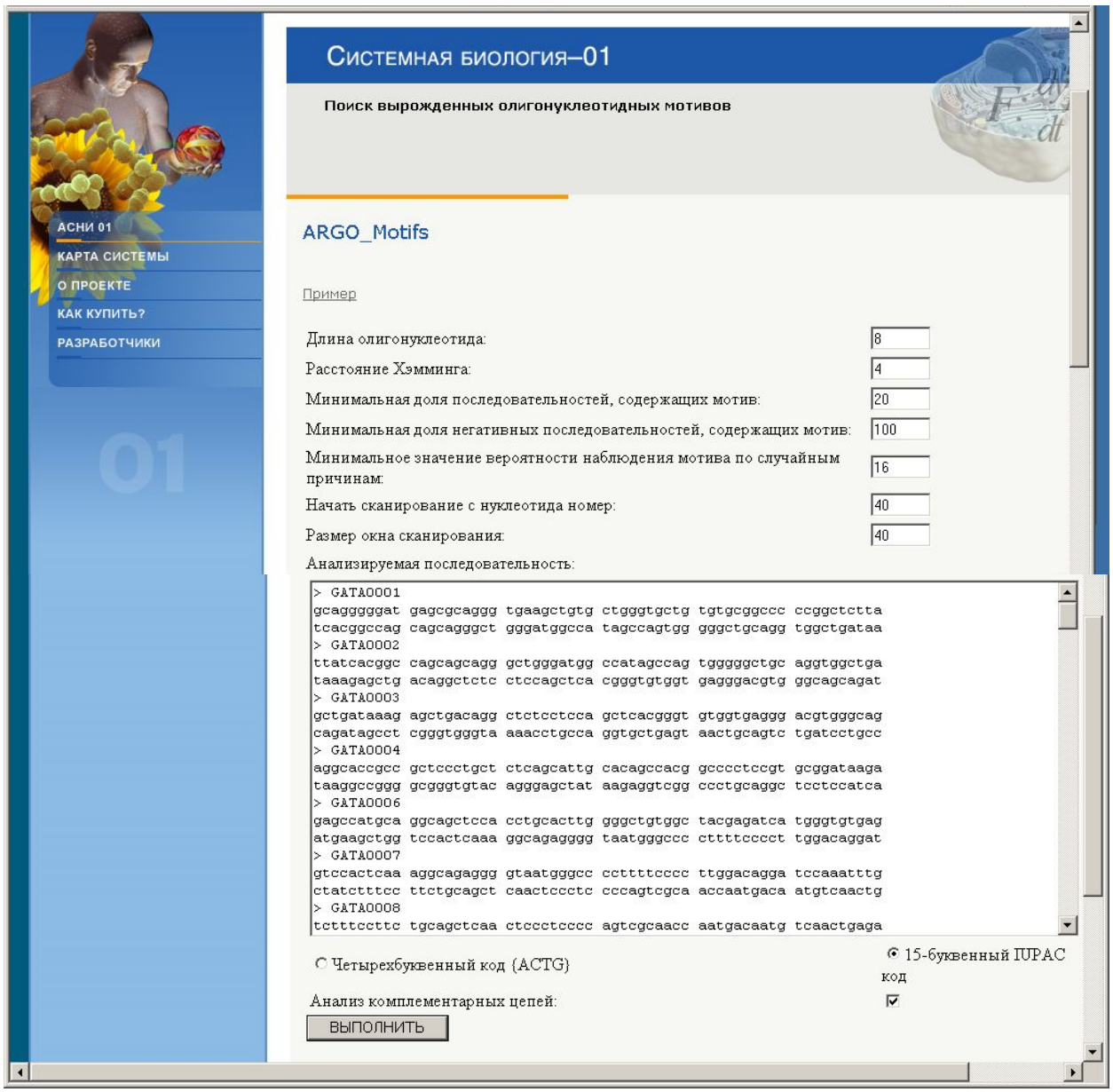


Рисунок 49. Пример установки параметров в Web-интерфейс ввода программы выявления вырожденных олигонуклеотидных мотивов ARGO_Motifs.

После ввода данных, как было описано выше, необходимо запустить программу, используя кнопку запуска (в нижней части экрана) в результате чего появляется, как это показано ниже на Рис. 50, HTML-страница со списком найденных мотивов и их характеристиками, а также картина расположения найденных мотивов вдоль анализируемых последовательностей (Рис.51).

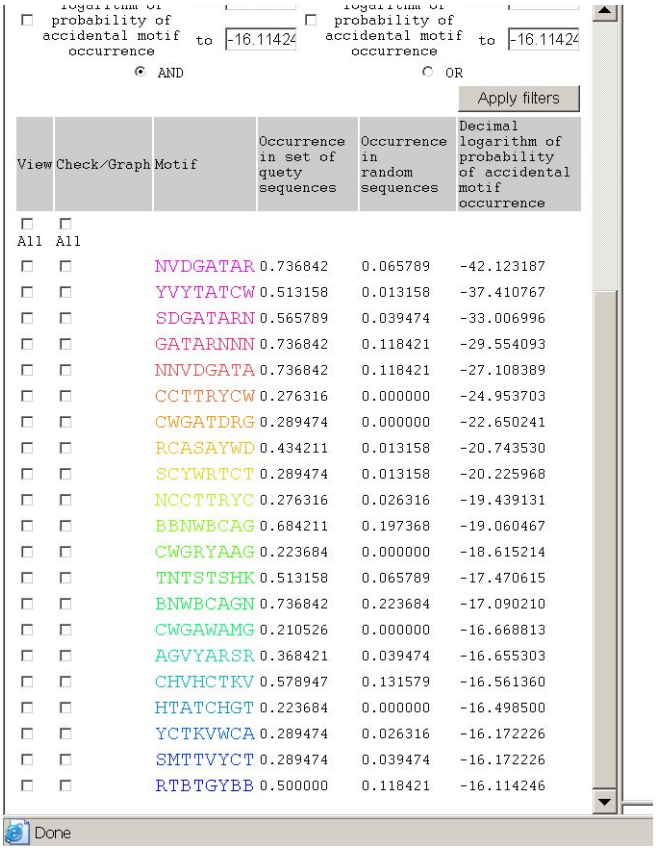
HTML-страница со списком найденных мотивов и их характеристиками (Рис. 50) включает в себя таблицу, состоящую из четырех колонок:

Первая колонка содержит найденные мотивы.

Вторая – частоту присутствия найденных мотивов в анализируемой выборке.

Третья – частоту присутствия найденных мотивов в выборке случайных последовательностей.

Четвертая – десятичный логарифм наблюдения мотивов в анализируемой выборке по случайным причинам.



View	Check/Graph	Motif	Occurrence in set of quety sequences	Occurrence in random sequences	Decimal logarithm of probability of accidental motif occurrence
All	All				
		NVDGATAR	0.736842	0.065789	-42.123187
		YVYTATCW	0.513158	0.013158	-37.410767
		SDGATARN	0.565789	0.039474	-33.006996
		GATARNNN	0.736842	0.118421	-29.554093
		NNVDGATA	0.736842	0.118421	-27.108389
		CCTTRYCW	0.276316	0.000000	-24.953703
		CWGATDRG	0.289474	0.000000	-22.650241
		RCASAYWD	0.434211	0.013158	-20.743530
		SCYWRTCT	0.289474	0.013158	-20.225968
		NCCTTRYC	0.276316	0.026316	-19.439131
		BBNWBCAG	0.684211	0.197368	-19.060467
		CWGRYAAG	0.223684	0.000000	-18.615214
		TNTSTSHK	0.513158	0.065789	-17.470615
		BNWBCAGN	0.736842	0.223684	-17.090210
		CWGAWAMG	0.210526	0.000000	-16.668813
		AGVYARSR	0.368421	0.039474	-16.655303
		CHVHCTKV	0.578947	0.131579	-16.561360
		HTATCHGT	0.223684	0.000000	-16.498500
		YCTKVWCA	0.289474	0.026316	-16.172226
		SMTTVYCT	0.289474	0.039474	-16.172226
		RTBTGYBB	0.500000	0.118421	-16.114246

Рисунок 50. Список мотивов, выявленных программой ARGO_Motifs в тестовом примере.

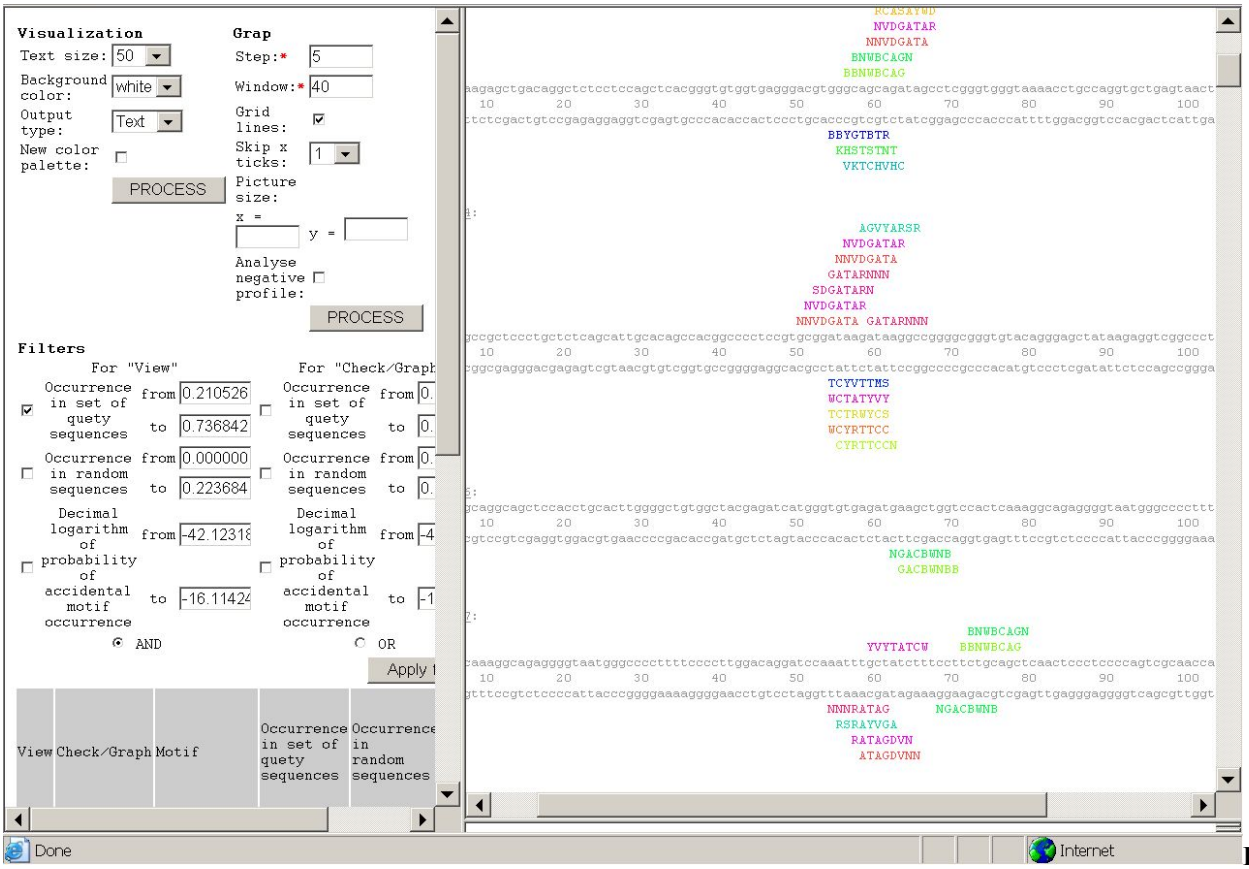


Рисунок 51. Окно локализации найденных олигонуклеотидных мотивов на анализируемых последовательностях.

Для завершения работы с текущей (введенной) выборкой нуклеотидных последовательностей надо вернуться на главную страницу программы, в результате чего осуществляется подготовка к повторному выполнению этой операции с другой выборкой последовательностей ДНК.

2.2.4. Программные компоненты для •распознавания регуляторных последовательностей (сайтов связывания транскрипционных факторов и промоторов)

Распознавание сайтов связывания транскрипционных факторов на основе анализа консервативных конформационных и физико-химических свойств ДНК в выровненных последовательностях обучающей выборки.

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. 52) 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 консевативных 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.

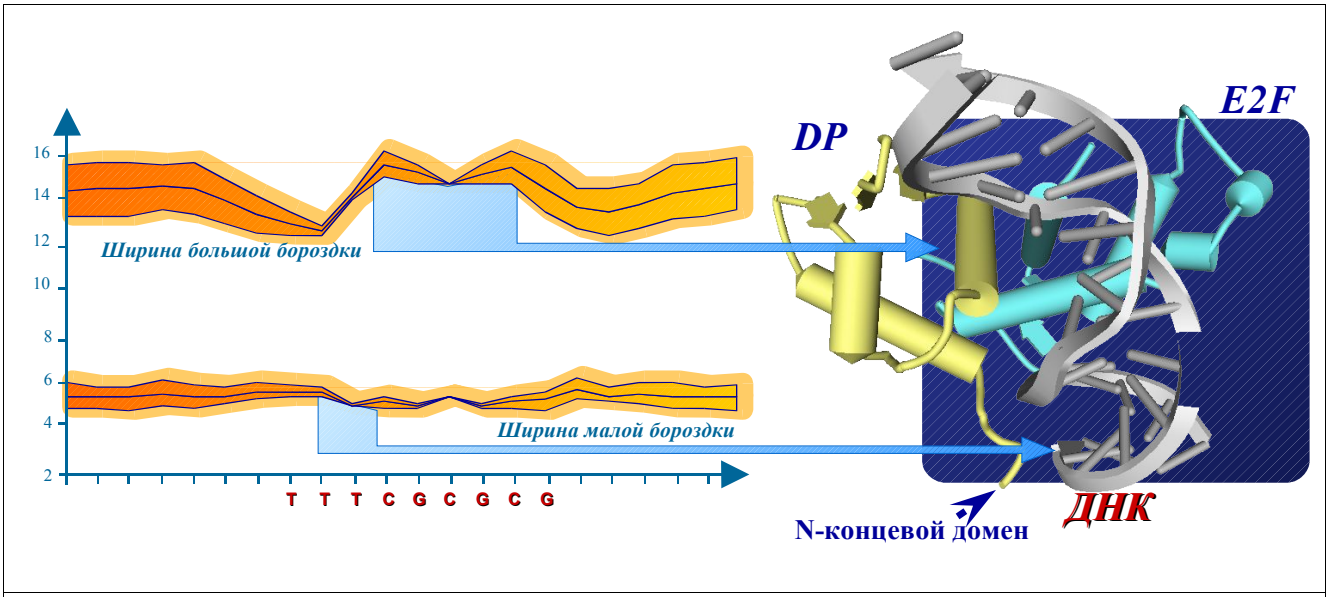


Рисунок 52. Выявленные SITECON значимо консервативные значения размеров бороздок в комплексе гетеродимера E2F/DP с ДНК согласуются с данными о механизме связывания гетеродимера с ДНК, полученными с помощью рентгеноструктурного анализа ДНК-белкового комплекса. Ширины большой и малой бороздок увеличены в местах контактов с распознающими альфа-спиралями и N-концевым доменом соответственно.

Описание работы SITECON

Выполнение операции является возможным при использовании в качестве входных данных линейной символьной последовательности ДНК, составленной из букв латинского алфавита { A, T, G, C, a, t, g, c } и имеющей длину не менее 100 и не более 1000000 символов.

Подготовка к выполнению операции надо войти в подсистему «Регуломика» и выбрать операцию «распознавание сайтов связывания транскрипционных факторов на основе данных о консервативных конформационных и физико-химических характеристиках» в списке операций этой подсистемы. В результате появится HTML-страница с интерфейсом этой операции, который показан на Рис.53.

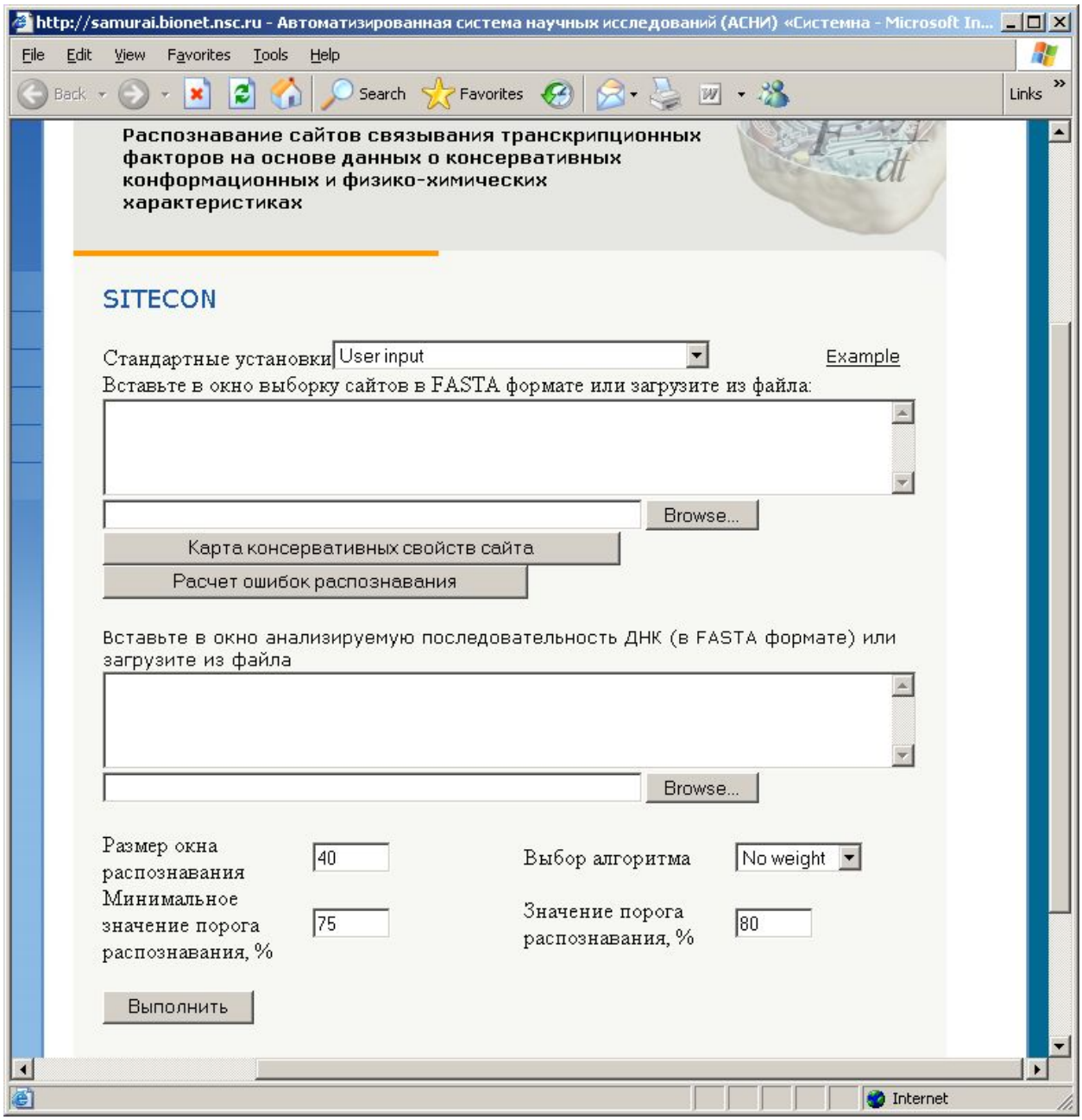


Рисунок 53. HTML-интерфейс операции, подготовленный к началу ее выполнения.

Выполнение операции начинается с ввода нуклеотидной последовательности ДНК. Для этой цели используется раздел “Вставьте в окно выборку сайтов в FASTA формате или загрузите из файла” HTML-интерфейса операции, как это показано на Рис. 54 для случая режима ввода данных, который установлен по-умолчанию.

Рисунок 54 Ввод нуклеотидной ДНК в режиме "from Screen (cut & paste)", который установлен по-умолчанию.

В качестве альтернативного способа ввода анализируемой нуклеотидной последовательности из текстового файла используется ИНТЕРНЕТ-браузер путем выбора режима ввода из файла с пользовательского компьютера после нажатия клавиши “Browse”.

Следующим действием выполнения операции является выбор транскрипционного фактора, чьи сайты связывания необходимо распознать во введенной нуклеотидной последовательности ДНК, как это показано ниже на Рис. 55 с помощью окна “Select transcription factor”.

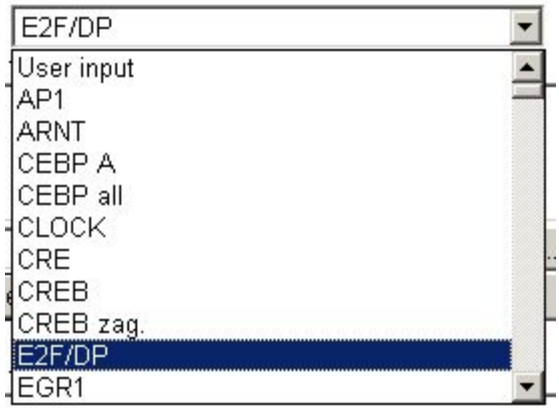


Рисунок 55. Выбор транскрипционного фактора, чьи сайты связывания необходимо распознавать во введенной нуклеотидной последовательности ДНК.

Следующим действием выполнения операции является выбор порога распознавания (уровня конформационного сходства). Чем выше порог распознавания, тем выше сходство распознаваемых сайтов с обучающей выброкой, тем меньше сайтов будет распознано. Порог не должен превышать 100% и не должен быть меньше 50%. (Рис. 55).

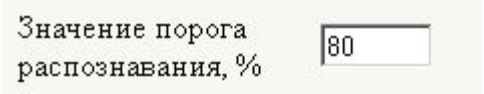


Рисунок 56 Выбор порога конформационного сходства между распознаваемыми сайтами и ДНК-сайтами, составляющими обучающую выборку. Порог не должен превышать 100% и не должен быть меньше 50%

После ввода данных, как было описано выше, необходимо нажать клавишу “Выполнить”, в результате чего появляется, как это показано ниже на Рис. 57, HTML-страница с таблицей распознанных сайтов с указанием позиции в которой они расположены, уровня конформационного сходства с обучающей выборкой, на какой цепи ДНК был обнаружен сайт и его последовательность.

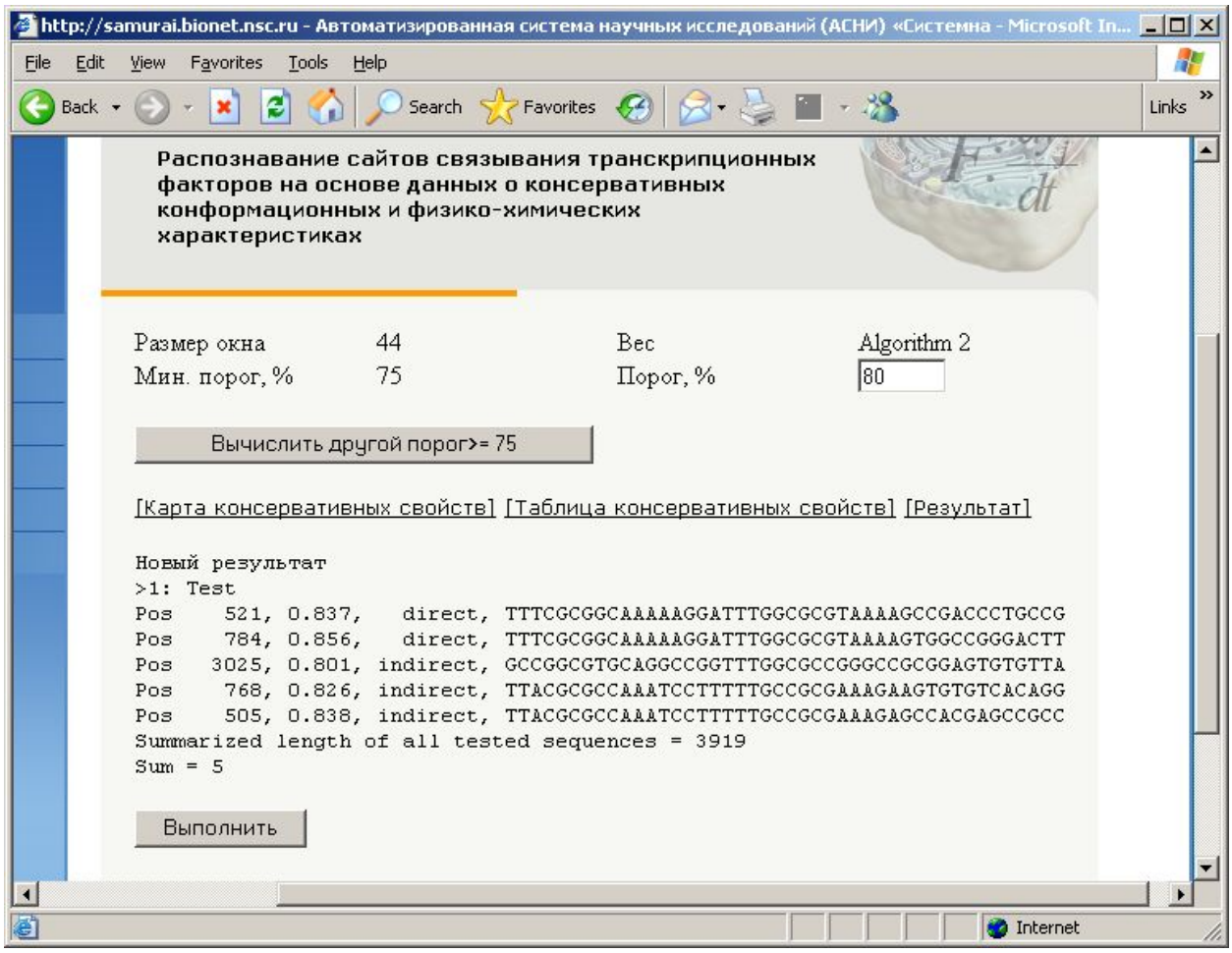


Рисунок 57. Пример HTML-страницы, содержащей предсказание сайтов связывания выбранного транскрипционного фактора во введенной нуклеотидной последовательности ДНК.

Для завершения работы с текущей (введенной) нуклеотидной последовательностью ДНК надо нажать клавишу “ Выполнить ”: в результате чего осуществляется подготовка к повторному выполнению этой операции с другой последовательностью ДНК.

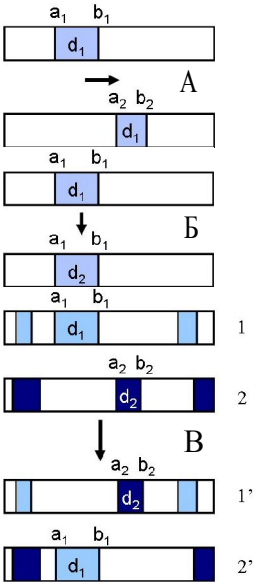
Выполнение операции занимает не более 10 минут для последовательности максимальной длины (до 1 000 000 нуклеотидов)"

Распознавания сайтов связывания транскрипционных факторов на основе анализа взаимных зависимостей частот встреч локально-позиционированных динуклеотидов в ССТФ.

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) (цис-элементы, Сингер и Берг, 1998). 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. 58А) или тип (рис.58Б) одного ЛПД особи. Рекомбинация осуществляет обмен ЛПД между двумя разными особями (рис.58В).



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 58. 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.

Описание работы SiteGA

Для того, чтобы провести распознавание сайтов связывания транскрипционных факторов на основе анализа взаимных зависимостей частот встреч локально-позиционированных динуклеотидов в ССТФ методом SiteGA, необходимо ввести нуклеотидную последовательность ДНК длиной не менее 93 нт, но не более 100 000 нт., или выборку последовательностей в FASTA формате, длина каждой из которых заключена в указанных выше рамках. Последовательности ДНК должна быть введена в алфавите «ATGCatgc», допускаются также символы “Nn”, но вблизи их сосредоточения сайты не могут быть предсказаны, то есть данные символы игнорируются программой.

Для каждого типа сайтов связывания транскрипционных факторов указать пороги.

Доступ к программе SiteGA осуществляется через подсистему «Регуломика», в которой необходимо выбрать надпись «распознавание сайтов связывания транскрипционных факторов на основе анализа взаимных зависимостей частот встречаемости локально-позиционированных динуклеотидов в ССТФ». В результате появится HTML-страница с интерфейсом этой операции, который показан на рисунке ниже (Рис. 59).

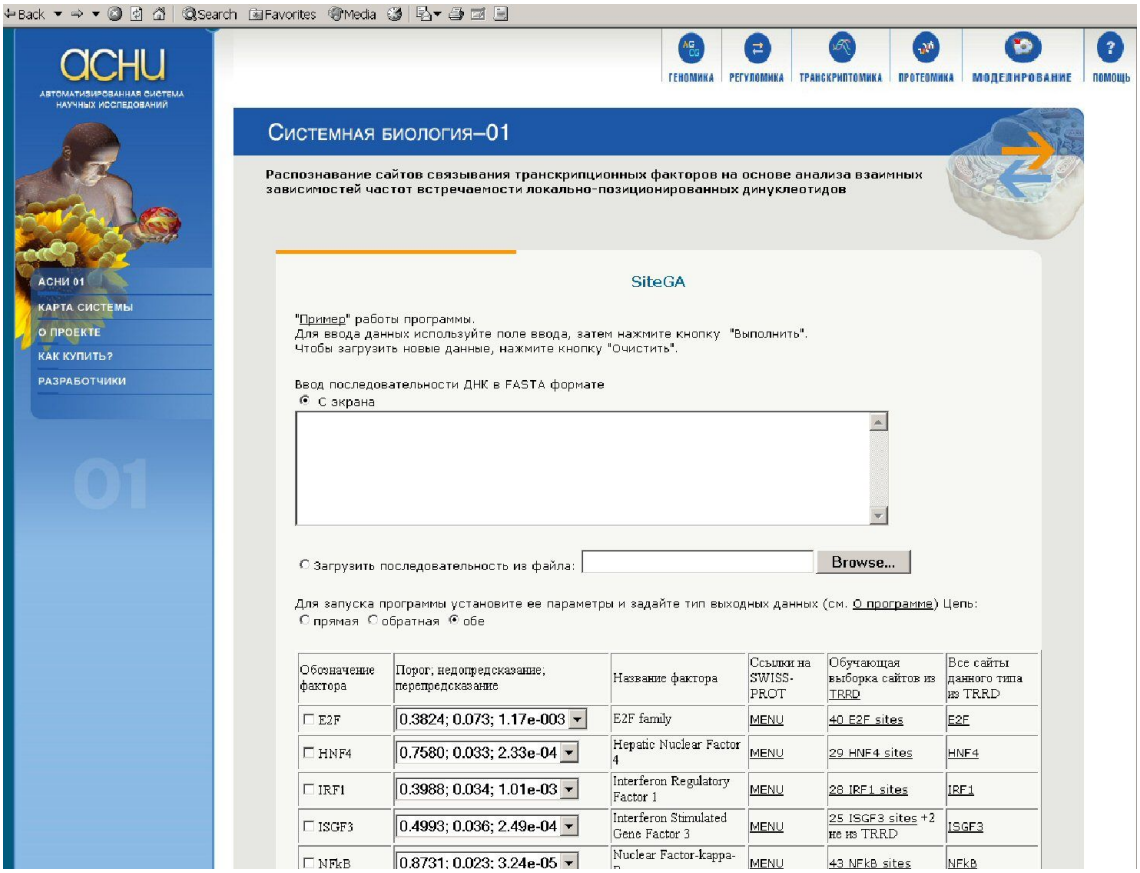


Рисунок 59. Интерфейс программы SiteGA.

Далее нужно

- ввести последовательности ДНК для анализа (одну или несколько в FASTA формате) в специальное окошко (Рис. 60, «Ввод последовательности ДНК в FASTA формате С экрана») или указать файл с последовательностями ДНК (кнопка ‘Browse’);

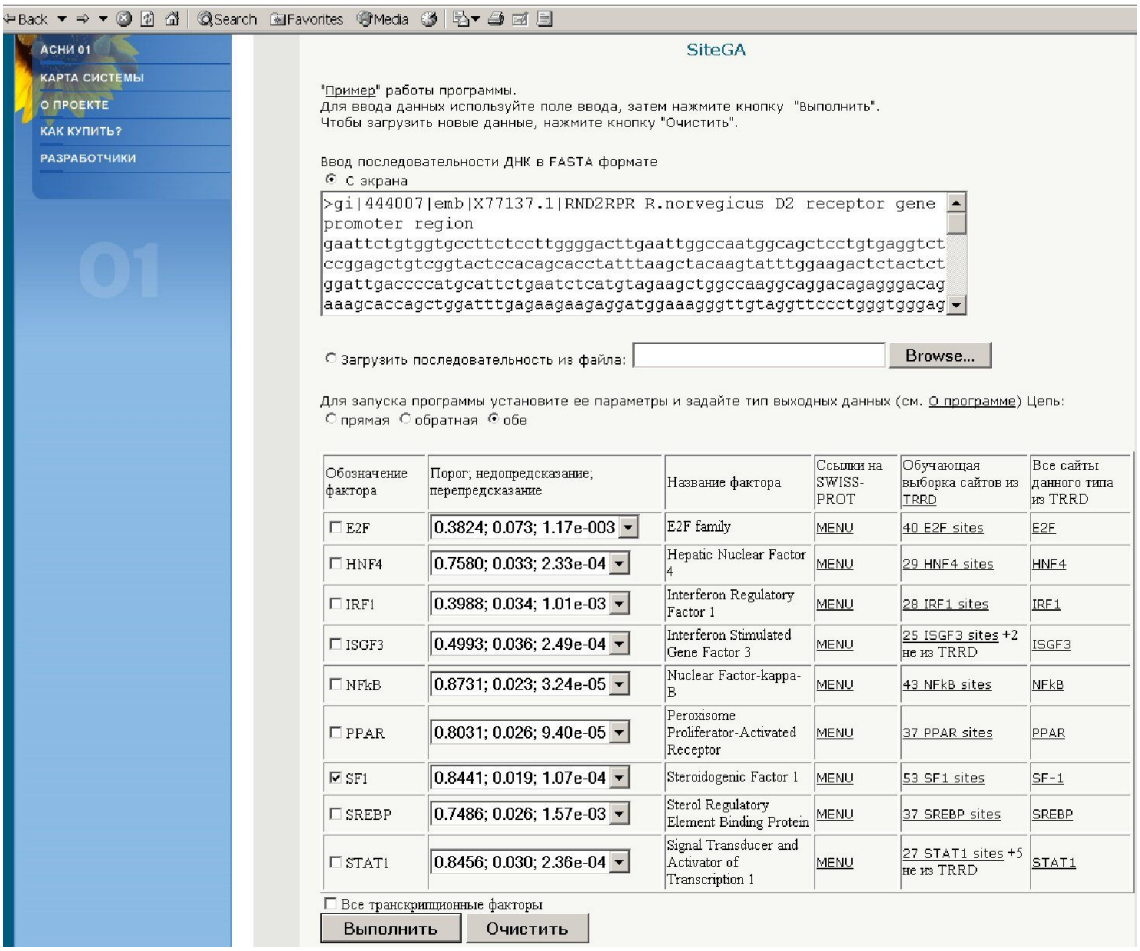


Рисунок 60. Интерфейс программы SiteGA. Заполнено поле входных последовательностей.

- указать цепи ДНК (опция Цепь: прямая, обратная или обе);
- указать типы сайтов (первая колонка в нижней части интерфейса представляет типы предложенных сайтов), или выбрать все типы (метка ‘Все транскрипционные факторы’);
- указать пороги для каждого типа сайтов (можно оставить выбранные по умолчанию значения);
- нажать кнопку ‘Выполнить’.

Выходные данные программы организованы в виде текстового файла (Рис. 61).

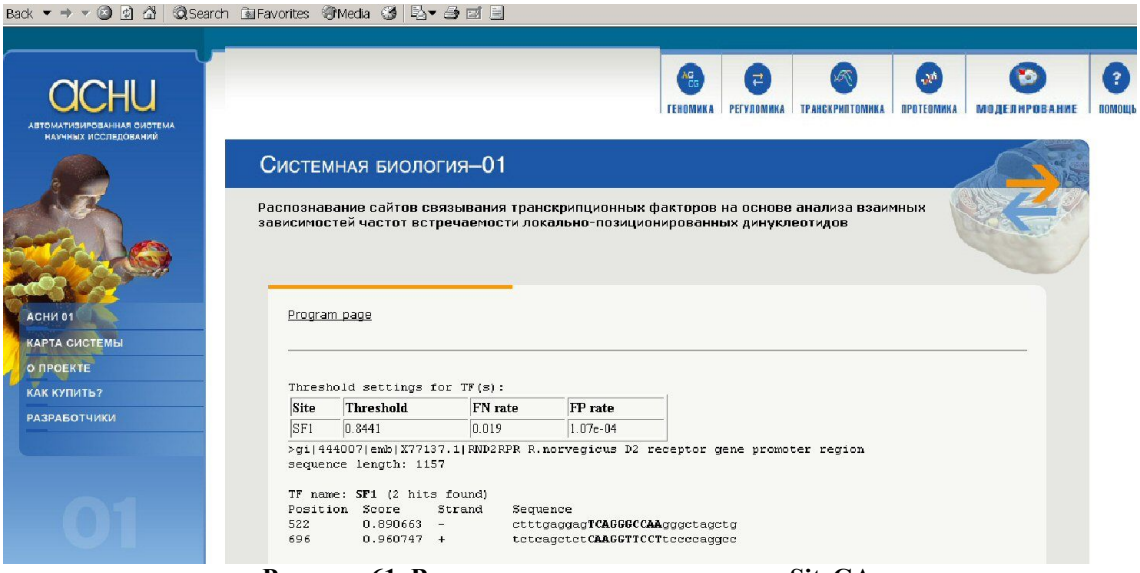


Рисунок 61. Выходные данные программы SiteGA.

Выходные данные содержат:

- Название анализируемой последовательности (>gi|444007|emb|X77137.1|RND2RPR R.norvegicus D2 receptor gene promoter region);
- Длину анализируемой последовательности (1157);
- Далее для каждой проанализированной цепи ДНК приводится таблица (Рис. 61), в которой для каждого предсказанного сайта отмечены:
- Общее число предсказанных сайтов данного типа ТФ;
- Позиция в последовательности, соответствующая центру сайта и значение функции распознавания;

Для завершения работы с текущими данными интерфейса необходимо нажать клавишу “Очистить”, в результате чего осуществляется подготовка к работе с другими данными (восстановятся все значения по умолчанию).

Программа CONSENSUS распознавания сайтов связывания транскрипционных факторов на основе поиска в исследуемой последовательности паттернов, сходных с паттернами в обучающих выборках

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:

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

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)".

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).

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.3 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)).

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:

	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 3. Олигонуклеотидные алфавиты, исследованные в настоящей работе.

<u>Обозначения</u>: 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).

Поскольку новые алфавиты из Табл.3 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:

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

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

Для преодоления несоизмеримости мы сделали дополнительно еще один необходимый шаг обобщения консенсуса и 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:

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

$$Score_\Sigma(S) = [\Sigma_{1\le h\le H}\, Score_h^0(S)\,]/H; \quad (7)$$

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

здесь: Δ, γ - нормировочные коэффициенты, с помощью при которых средние значения сходства для всех известных сайтов, $M^0_{S\in SITE}[Score^0(S)]$, и для 10^3 случайных ДНК, $M^0_{S\in RAND}[Score^0(S)$, 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 21 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 соответствии с Центральной предельной теоремой благодаря учету дополнительных олигонуклеотидных алфавитов.

Описание работы CONSENSUS

Распознавание сайтов связывания транскрипционных факторов на основе поиска паттернов, сходных с паттернами в обучающих выборках является возможным при использовании в алфавита { A, T, G, C, a, t, g, c } и имеющей длину не менее 32 и не более 30000 качестве входных данных линейной символьной последовательности ДНК, составленной из букв латинского символов.

Подготовка к выполнению операции: надо войти в подсистему «Регуломика» и выбрать раздел «распознавание сайтов связывания транскрипционных факторов на основе поиска в исследуемой последовательности паттернов, сходных с паттернами в обучающих выборках» в списке операций этой подсистемы. В результате появится HTML-страница с интерфейсом этой операции, который показан на рисунке ниже (Рис. 62).

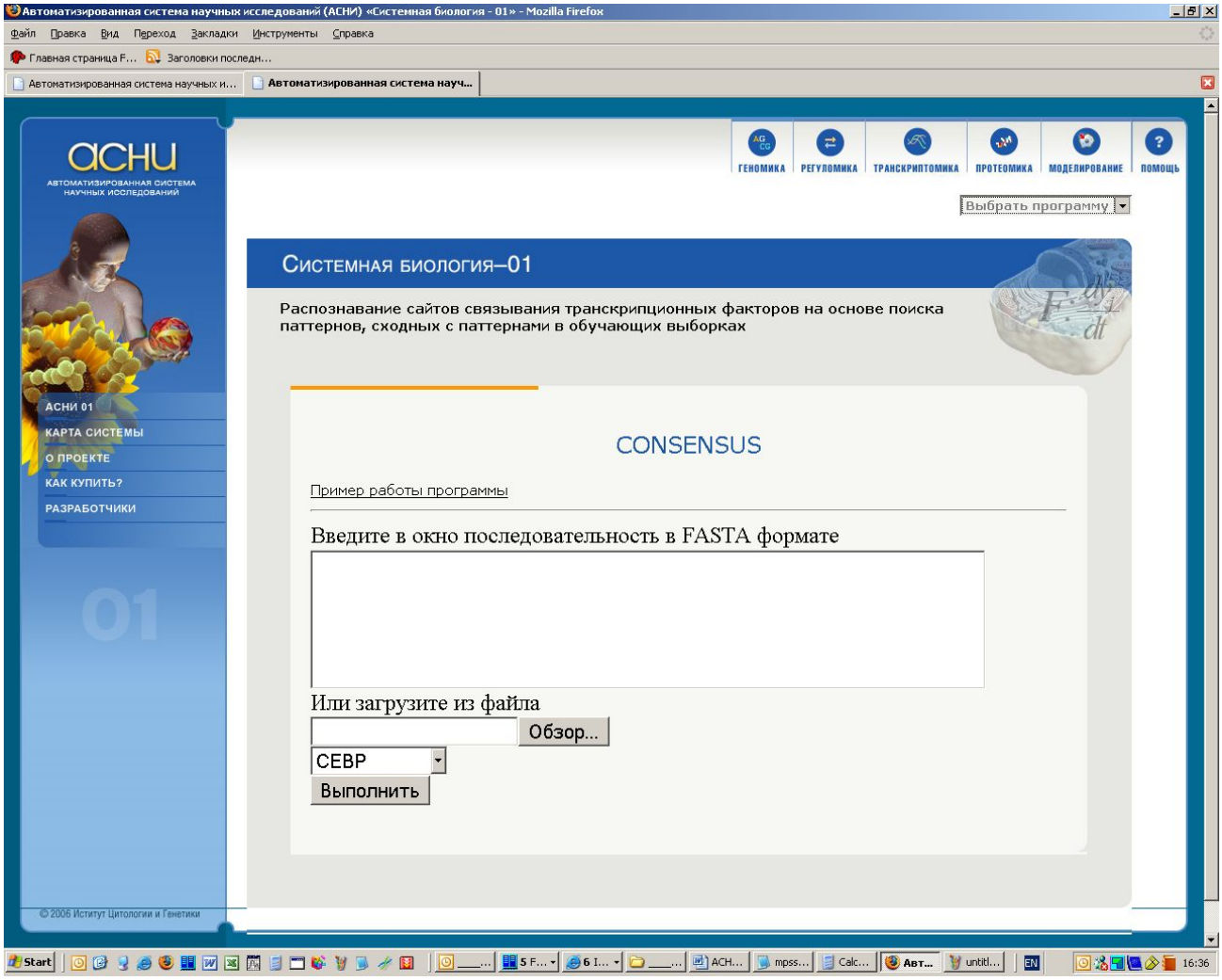


Рисунок 62. HTML-интерфейс операции, подготовленный к началу ее выполнения.

Выполнение операции начинается с ввода нуклеотидной последовательности ДНК. Для этой цели используется раздел “ Введите в окно последовательность в FASTA формате ” HTML-интерфейса операции, как это показано на рисунке ниже (Рис. 63).

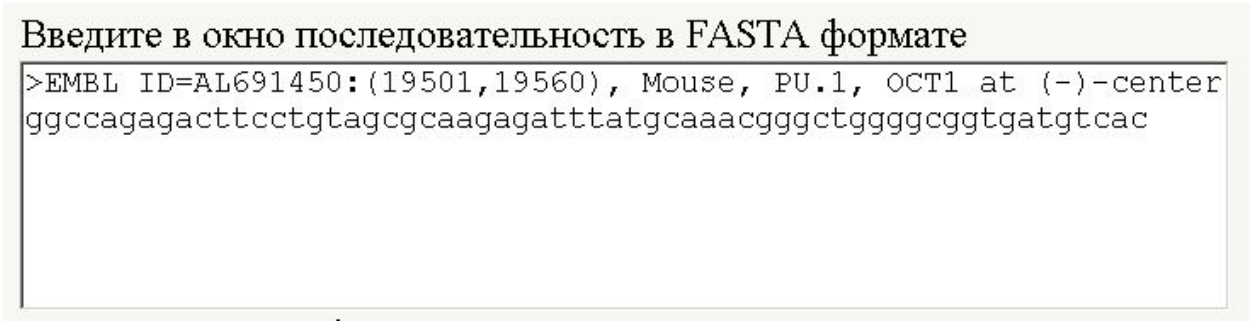


Рисунок 63. Ввод нуклеотидной ДНК в режиме " Введите в окно последовательность в FASTA формате ".

В качестве альтернативного способа ввода нуклеотидной последовательности полиморфного варианта анализируемой мРНК из текстового файла используется ИНТЕРНЕТ-браузер путем выбора режима ввода “Или загрузите из файла” и последующего нажатия клавиши “Обзор”.

Следующим действием выполнения операции является выбор транскрипционного фактора, чьи сайты связывания необходимо распознать во введенной нуклеотидной последовательности ДНК, как это показано ниже на рисунке (Рис. 64):

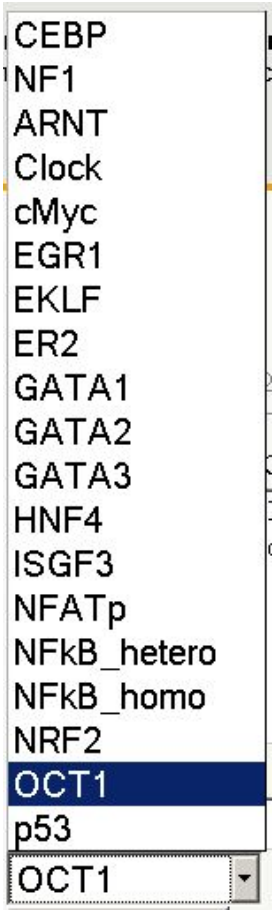


Рисунок 64. Выбор транскрипционного фактора, чьи сайты связывания необходимо распознавать во введенной нуклеотидной последовательности ДНК.

После ввода данных, как было описано выше, необходимо нажать клавишу “Выполнить”, в результате чего появляется, как это показано ниже на рисунке (Рис. 65), HTML-страница с текстовым форматом выдачи результата..

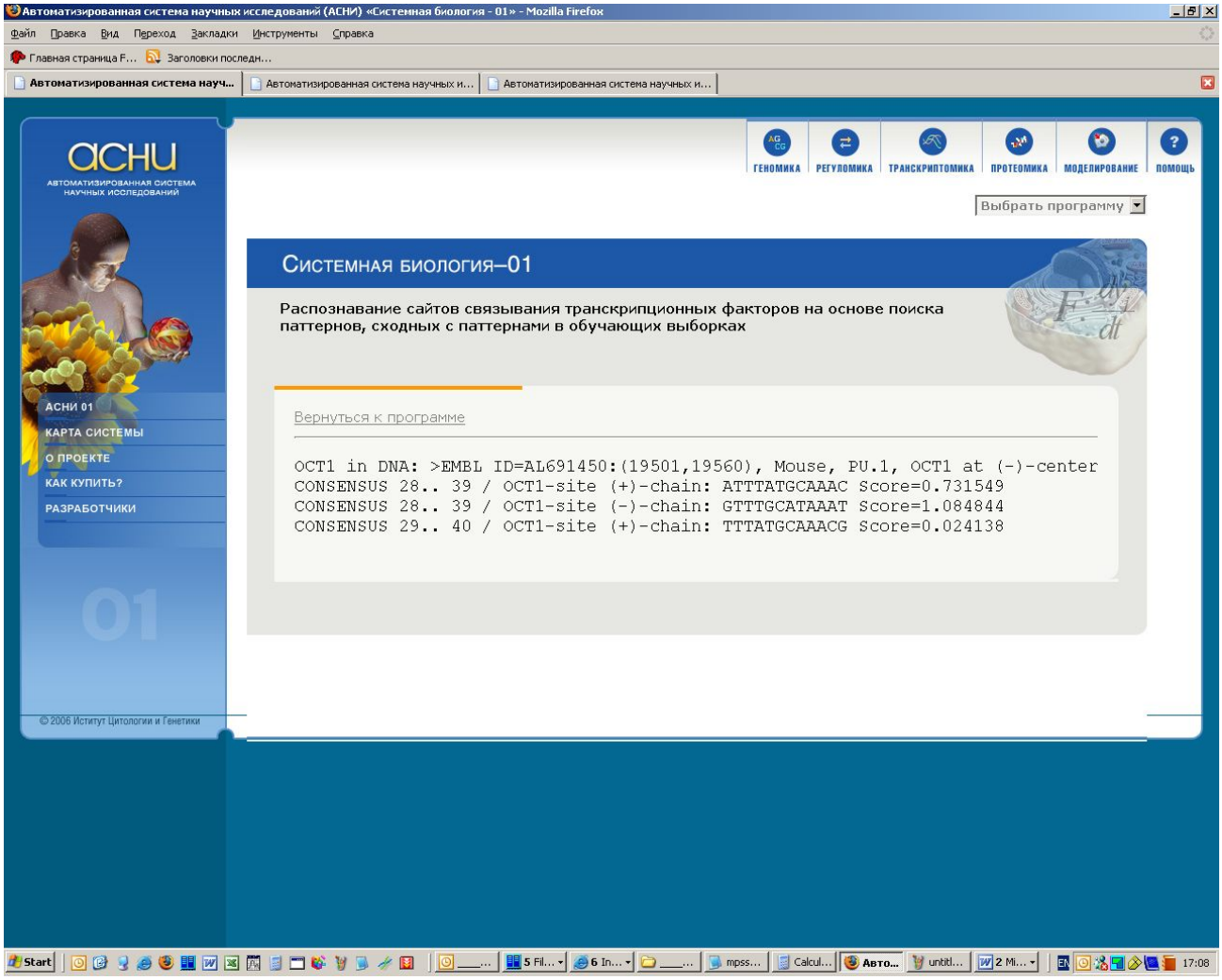


Рисунок 65. Пример HTML-страницы, содержащей предсказание сайтов связывания выбранного транскрипционного фактора во введенной нуклеотидной последовательности ДНК.

Здесь в первой строке приведено названий исследуемого сайта и название анализируемой нуклеотидной последовательности ДНК. В каждой следующей строке описано 1 (один) потенциальный сайт, распознанный программой. Это описание следует так называемому GeneBank-подобному формату

Программа MATRIX распознавания сайтов связывания транскрипционных факторов (TF-сайтов) на основе весовых матриц

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:

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

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:

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

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”:

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

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”.

	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 4. Олигонуклеотидные алфавиты, исследованные в настоящей работе.

<u>Обозначения</u>: 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 Таблице 4 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 $\{S_1...S_n...S_N\}$ 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)). В этих обозначениях, «байесовскую» оценку частот встречаемости (формулы 4 и 5) обобщалась:

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

Поскольку новые алфавиты из Табл.4 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:

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

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:

$$\left\{\left\{\left\{\begin{array}{l} \text{Score}^0(S) = (\text{Score}(S) - \Delta)/\gamma, \\ \Delta = (M^0_{S\in SITE}[\text{Score}(S)] + M^0_{S\in RAND}[\text{Score}(S)])/2; \\ \gamma = (M^0_{S\in SITE}[\text{Score}(S)] - M^0_{S\in RAND}[\text{Score}(S)])/2; \end{array}\right.\right.\right. \quad (8)$$

$$\text{Score}_\Sigma(S) = [\Sigma_{1\le h\le H} \text{Score}_h^0(S)] / H; \quad (9)$$

$$[\ \text{Score}_\Sigma(S^{\#}) \ge 0\ ] \Rightarrow \{\ S^{\#} - \text{результат предсказания "потенциальный сайт"}\}, \quad (10)$$

здесь: Δ, γ - нормировочные коэффициенты, с помощью при которых средние значения сходства для всех известных сайтов, $M^0_{S\in SITE}[Score^0(S)]$, и для 10^3 случайных ДНК, $M^0_{S\in RAND}[Score^0(S)]$, устанавливаются равными соответственно "+1" и "-1"; H – количество используемых олигонуклеотидных алфавитов (Табл. 4); 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 60 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.

Описание работы MATRIX

Распознавание сайтов связывания транскрипционных факторов в эукариотических геномных ДНК с помощью матриц частот олигонуклеотидов является возможным при использовании в качестве входных данных линейной символьной последовательности ДНК, составленной из букв латинского алфавита { A, T, G, C, a, t, g, c } и имеющей длину не менее 32 и не более 30000 символов.

Подготовка к выполнению операции надо войти в подсистему «Регуломика» и с помощью мыши кликнуть надпись «распознавание сайтов связывания транскрипционных факторов на основе весовых матриц» в списке операций этой подсистемы. В результате появится HTML-страница с интерфейсом этой операции, который показан на Рис. 66.

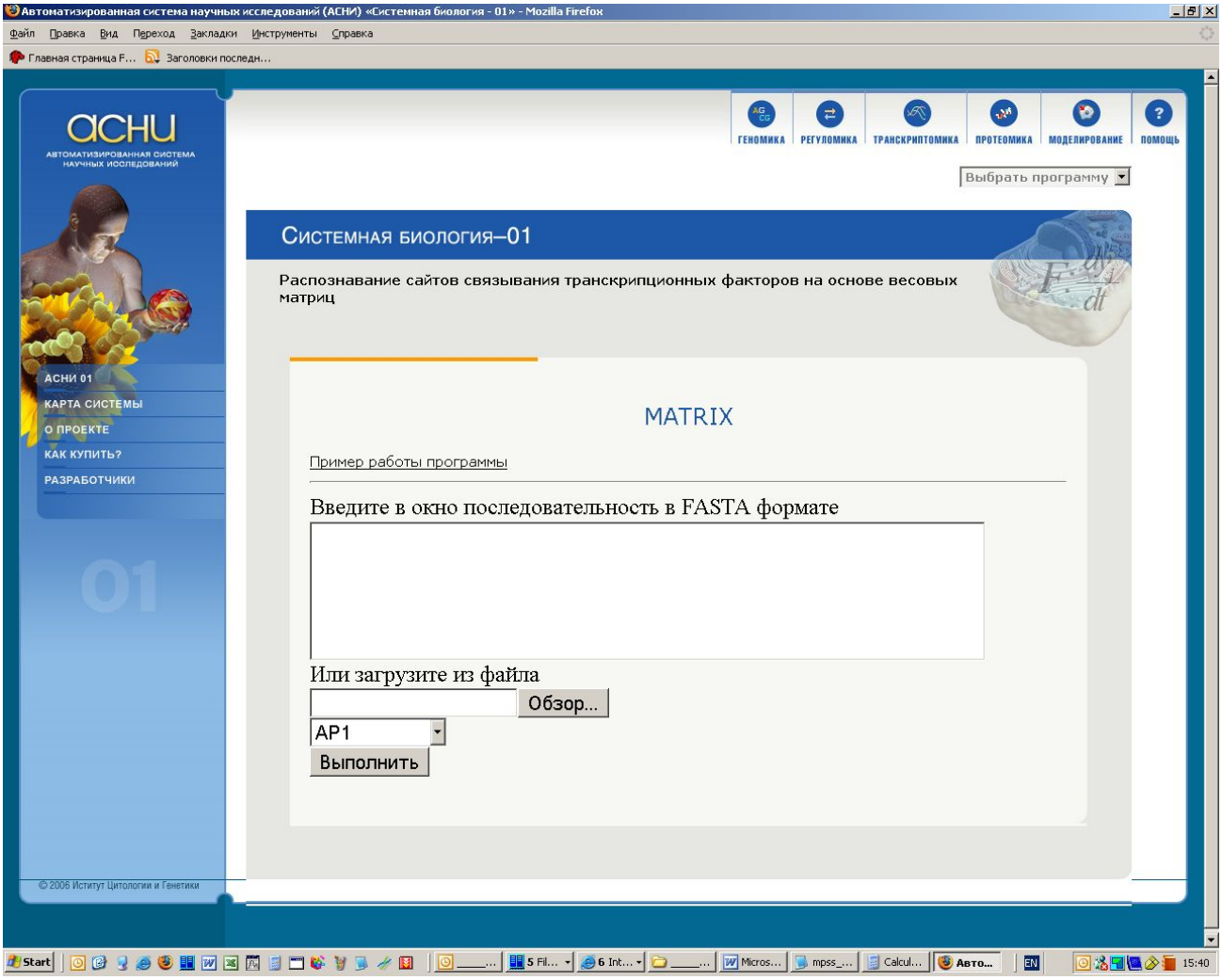


Рисунок 66. HTML-интерфейс операции, подготовленный к началу ее выполнения.

Выполнение операции начинается с ввода нуклеотидной последовательности ДНК. Для этой цели используется раздел “Введите в окно последовательность в FASTA формате” HTML-интерфейса операции, как это показано на Рис. 67.

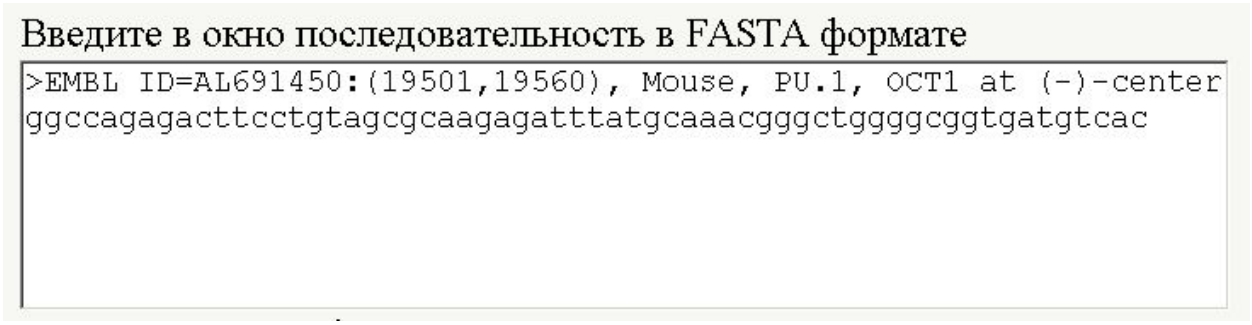


Рисунок 67. Ввод нуклеотидной ДНК в режиме "Введите в окно последовательность в FASTA формате ".

В качестве альтернативного способа ввода нуклеотидной последовательности полиморфного варианта анализируемой мРНК из текстового файла используется

ИНТЕРНЕТ-браузер путем выбора режима ввода “Или загрузите из файла” и последующего нажатия клавиши “Обзор”.

Следующим действием выполнения операции является выбор транскрипционного фактора, чьи сайты связывания необходимо распознать во введенной нуклеотидной последовательности ДНК, как это показано ниже на Рис. 68.

Рисунок 68. Выбор транскрипционного фактора, чьи сайты связывания необходимо распознавать во введенной нуклеотидной последовательности ДНК.

После ввода данных, как было описано выше необходимо, нажать клавишу “Выполнить”, в результате чего появляется, как это показано ниже на Рис. 69, HTML-страница с текстовым форматом выдачи результата.

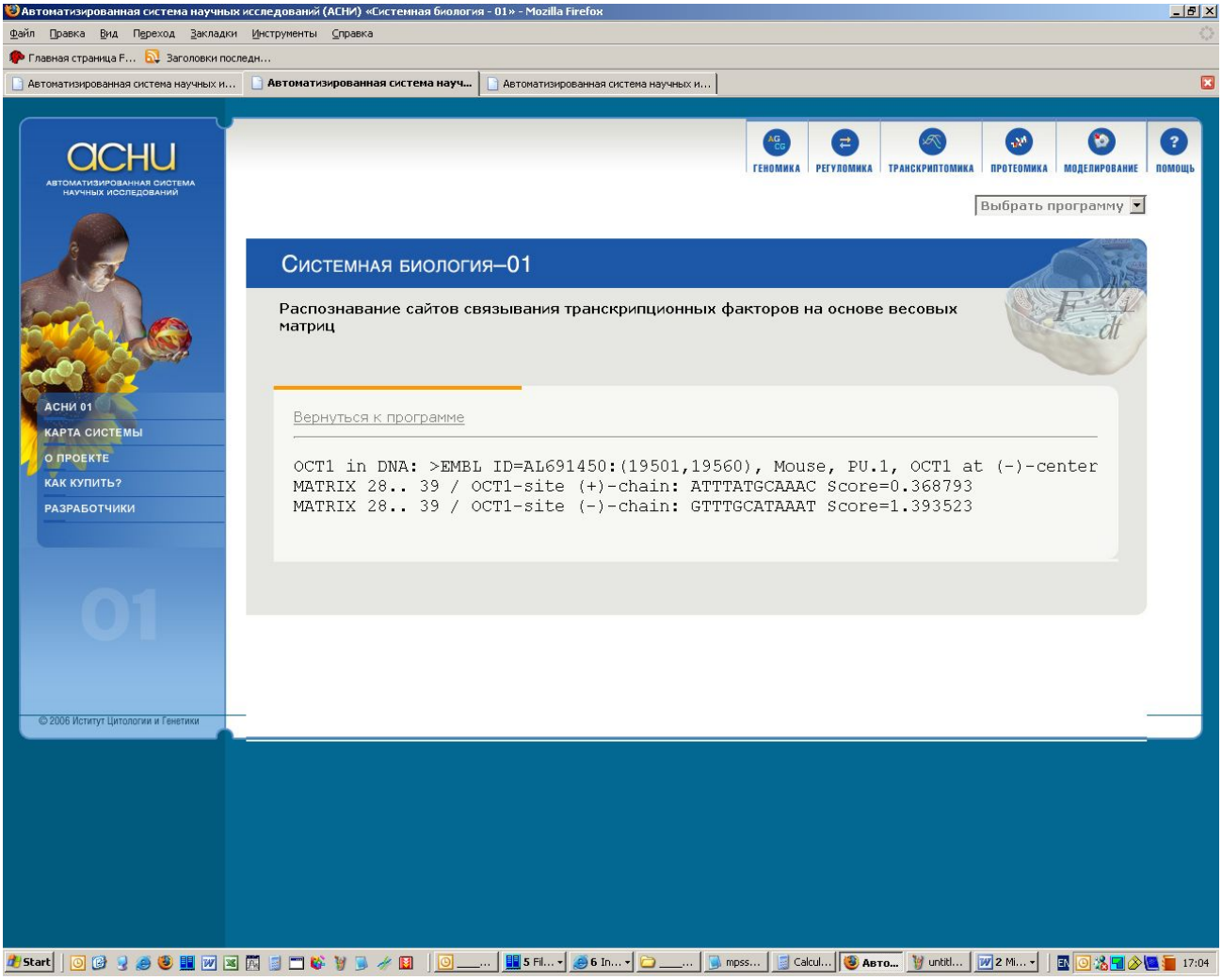


Рисунок 69. Пример HTML-страницы, содержащей предсказание сайтов связывания выбранного транскрипционного фактора во введенной нуклеотидной последовательности ДНК .

Здесь в первой строке приведено названий исследуемого сайта и название анализируемой нуклеотидной последовательности ДНК. В каждой следующей строке описано 1 (один) потенциальный сайт, распознанный программой. Это описание следует так называемому GeneBank-подобному формату.

Оценка количественной величины транскрипционной активности ДНК в логарифмических единицах свободной энергии Гиббса

Проблема предсказания количественных величин Act(S) биологической активности по нуклеотидным последовательностям S ДНК была впервые сформулирована МакКлуром (Mulligan *et al.*, 1984) с целью прогноза “селективности” РНК-полимеразы к промоторам E.coli по степени сходства сайтов связывания ДНК с σ70-субъединицей РНК-полимеразы (“-35” и“-10” боксы) с эволюционно-консервативным паттерном этих сайтов длины L:

$$\begin{cases} \mathrm{Act}(S=\{s_1,\ldots,s_j,\ldots,s_L\}) = a_0 + a_1 \times X(S=\{s_1,\ldots,s_j,\ldots,s_L\}), & (1) \\ X(S=\{s_1,\ldots,s_j,\ldots,s_L\}) = \mathrm{Score}(S)\{s_1,\ldots,s_j,\ldots,s_L\}) = \Sigma_{1\le i\le K}\,\Sigma_{1\le j\le L}\, w_{i,j}\times\delta(s_j=e_i), & (2) \end{cases}$$

$$w_{i,j} = f_{i,j} = (N_{i,j} + 1) / (N + K), \quad (3)$$

здесь: Score(S) – сходство ДНК-сайта с паттерном всех известных таких сайтов (в данном случае, линейно-аддитивная оценка сходства, называнная МакКлуром "Homology Score"); a_0, a_1 – коэффициенты простой регрессии; δ(истина)=1, δ(ложь)=0; $w_{i,j}$ - «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; $f_{i,j}$ – оценки частоты встречаемости каждого i-ого мономера e_i в каждой j-ой позиции сайта (в данном случае приведена "байесовская оценка"); N – число известных вариантов сайта; $N_{i,j}$ – число вариантов сайта, в которых i-ый мономер e_i занимает j-ую позицию ($s_j=e_i$)..

Сараи и соавторы (Tanikawa et al., 1993) заменили приближенное оценивание (формулы 2 и 3) позиционно-весовых вкладов $w_{i,j}$ их измерением в прямом эксперименте:

$$w_{i,j} = \ln(\Phi[\mathrm{Site}_{i,j}^{\mathrm{MUT}}=s^0_1,\ldots,s^0_{j-1},s_j^{\mathrm{MUT}}=e_i,s^0_{j+1},\ldots,s^0_L] / \Phi[\mathrm{Site}^{\mathrm{WT}}=s^0_1,\ldots,s^0_j,\ldots,s^0_L]), \quad (4)$$

здесь: Ф[Site] – измеренное в эксперименте значение биологической активности сайта (например, для сайта связывания ДНК с белковым фактором транскрипции, так называемая «трансляционная активность» определенного промотора, который регулируется этим фактором); $\mathrm{Site}^{\mathrm{WT}}=s^0_1,s^0_2,\ldots,s^0_j,\ldots,s^0_L$ – последовательность ДНК природного варианта исследуемого сайта; $\mathrm{Site}_{i,j}^{\mathrm{MUT}}=s^0_1,s^0_2,\ldots,s^0_{j-1},s_j^{\mathrm{MUT}}=e_i,s^0_{j+1},\ldots,s^0_L$ - последовательность мутантного варианта с точечной заменой ровно одного нуклеотида в j-ой позиции природного сайта, где природный нуклеотид s^0_j заменен заданным каноническим мономером e_i в качестве мутантного нуклеотида s_j^{MUT}.

В тоже время накапливается, однако, все больше экспериментальных данных, что, например, в случае транскрипционной активности, ее величины, вычисляемые путем суммирования линейно-аддитивных вкладов каждого нуклеотида в каждой позиции сайта ДНК, не коррелируют с величинами этой биологической активности, измеренными экспериментально (Roulet et al., 1998; Robison et al., 1998; Shulzaberger and Schneider, 1999). Оказалось (Weston, 1992; Hyde-DeRuyscher et al., 1995; Javahery et al., 1994), что многокомпонентность транскрипционной машины может стать причиной достоверного отсутствия корреляций как между разными биологическими активностями (например, "выход мРНК" и "сродство ДНК-белок"), так и для определенной активности в разных условиях ее измерения (например, в разных тканях). Для эукариот, внешними факторами, способными изменять биологическую активность любого сайта ДНК, могут оказаться также нуклеосомная упаковка ДНК вблизи сайта и взаимодействие белков хроматина с регуляторными белками, связывающими этот сайт (Godde et al., 1995; Edmondson and Roth, 1996). Наконец, в числе факторов, способных влиять на биологическую активность любого сайта ДНК, были обнаружены как синергетические, так и антагонистические взаимодействия различных регуляторных белков, ткане-специфически связывающих ДНК вблизи этого сайта в так называемых "композиционных единицах" (French et al., 1998). Кроме того, существенно, что получаемая с помощью позиционно-весовых матриц $\{w_{i,j}\}$ линейно-аддитивная аппроксимация корреляций "последовательность-активность" не дает их содержательной интерпретации: остается неясным, какие особенности ДНК, по какому молекулярному механизму влияют на биологическую активность сайта (Jonsson et al., 1993). Все эти эксперименты указывают, что линейно-аддитивное приближение (Berg and von Hippel, 1987) корреляций последовательность-активность сайтов ДНК требует своего дальнейшего развития для адекватного понимания многофакторных биологических активностей ДНК.

Программа ddG_MATRIX_INR использует для прогноза экспериментальную матрицу позиционно-аддитивных свободных энергий Гиббса (формула 4) из работы (Kraus et al., 1996), корректируя ее учетом кооперативного взаимодействия соседних нуклеотидов путем

усреднения двух конформационных углов изгиба оси Direction и кручения Twist динуклеотидных шагов B-спирали ДНК (номенклатура (EMBO, 1989)).

$$\begin{cases} Act(S=\{s_{-3},\ldots,s_j,\ldots,s_{+4}\}) = a_0 + a_1\times Act_{ADDITIVE}(S) + a_2\times Act_{B\text{-}HELIX}(S); & (5) \\ Act_{B\text{-}HELIX}(S)= [\Sigma_{-3\le j\le +3} w_{i,j}\times\delta(s_j\neq e^0_j)] \times [b_0 + b_1\times Direction(S) + b_2\times Twist(S)], & (6) \end{cases}$$

$$Direction(S) = [\ \Sigma_{-3\le i\le +2}\ Direction(s_is_{i+1})\ ] / 5, \quad (7)$$

$$Twist(S) = [\ \Sigma_{-2\le i\le +3}\ Twist(s_is_{i+1})\ ] / 5, \quad (8)$$

здесь: (a_0, a_1, a_2) и (b_0, b_1, b_2) – наборы коэффициентов линейной регрессии, вычисляемые стандартным способом (Forster and Ronr, 1979) по экспериментальную матрицу позиционно-аддитивных свободных энергий Гиббса из работы (Kraus et al., 1996); $\{e^0_{-3},\ldots,e^0_j,\ldots,e^0_{+4}\}$ = {GTTATTT} – общепринятый в качестве стандарта природный INR-элемент главного позднего промотора вируса SV40 (Kraus et al., 1996); Direction(s_is_{i+1}) и Twist(s_is_{i+1}) - экспериментальные величины для динуклеотидов s_is_{i+1}, взятые из работ (Shpigelman et al., 1993) и (Suzuki et al., 1996), соответственно.

Конформационные углы изгиба оси Direction и кручения Twist динуклеотидных шагов B-спирали ДНК для формул (6-8) были автоматически выбраны системой ACTIVITY (Ponomarenko et al., 1997) из числа всех 38 известных конформационных и физико-химических параметров B-спирали ДНК (Ponomarenko et al., 1999), усредняемых на всех возможных фрагментах INR-элемента. В процессе этого выбора ACTIVITY для каждого свойства ДНК и для каждого района его усреднения в пределай сайта тестирует 11 статистических критериев, в числе которых ранговые, линейные и знаковые корреляции. При этом с помощью метода bootstrap (Hayes et al., 1989) анализируется семь поднаборов сочетаний «среднее значение свойства ДНК – транскрипционная активность». Каждая такиая **n**-ая проверка ($1\le\mathbf{n}\le77$) заключается в оценке уровня статистической значимости α_n корреляции, которая затем преобразуется, в терминах нечетких логик Задэ (Zadeh, 1965), в оценку обоснованности корреляции «среднее значение свойства ДНК – транскрипционная активность»:

$$u_n = \begin{cases} \mathbf{1}, & \text{если } \alpha_n\le\mathbf{0.01}; \\ \mathbf{1.3 - 28.3}\alpha_n + \mathbf{55.6}\alpha_n^2, & \text{если } \mathbf{0.1}\ge\alpha_n\ge\mathbf{0.01}; \\ \mathbf{-1}, & \text{если } \alpha_n\ge\mathbf{0.1}. \end{cases} \quad (9)$$

Каждому достоверной корреляции ($\alpha_n<0.05$) формула (№№7) дает положительную оценку ее обоснованности u_n от 0 до 1, недостоверной - отрицательную оценку от -1 до 0. Для каждого сочетания «среднее значение свойства ДНК – транскрипционная активность» получается 77 таких частных оценок обоснованности $\{u_n\}_{1\le n\le77}$ наличия какий-либо корреляций между ними, которые, согласно теории принятия решений (Fishburn, 1970), усредняются в интегральную оценку обоснованности:

$$U = [\ \Sigma_{1\le n\le77}\ u_n\ ] / 77. \quad (10)$$

Формула (10) дает каждому сочетанию «среднее значение свойства ДНК – транскрипционная активность» тем большую оценку обоснованности U корреляций между ними, чем больше наблюдается количество таких достоверных корреляций и чем выше уровень статистической их значимости α_n. Поэтому самая высокая положительная оценка обоснованности U указывает именно то свойство B-спирали ДНК и участок его усреднения в пределах INR-элемента, где было самое лучшее согласие средних значений этого свойства с

экспериментальными величинами транскрипционной активности. Угол кручения Twist (формула 8) был выявлен в качестве наиболее скоррелированного с повреждающими заменами нуклеотидов, приводящих к почти полному исчезновению транскрипционной активности. Угол изгиба оси Direction (формула 7) был наиболее скоррелирован с нейтральными заменами нуклеотидов, приводящими к слабой модуляции транскрипционной активности вблизи ее природного уровня.

Описание работы ddG_MATRIX_INR

Оценка количественной величины транскрипционной активности ДНК (в логарифмических единицах свободной энергии Гиббса) с помощью программы ddG_MATRIX_INR является возможным при использовании в качестве входных данных линейной символьной последовательности ДНК, составленной из букв латинского алфавита { A, T, G, C, a, t, g, c } и имеющей длину не менее 7 и не более 30000 символов.

Подготовка к выполнению операции: надо войти в подсистему «Регуломика» и с помощью мыши кликнуть надпись «ddG_MATRIX_INR - оценка количественной величины транскрипционной активности ДНК (в логарифмических единицах свободной энергии Гиббса)» в списке операций этой подсистемы. В результате появится HTML-страница с интерфейсом этой операции, который показан на Рис. 70.

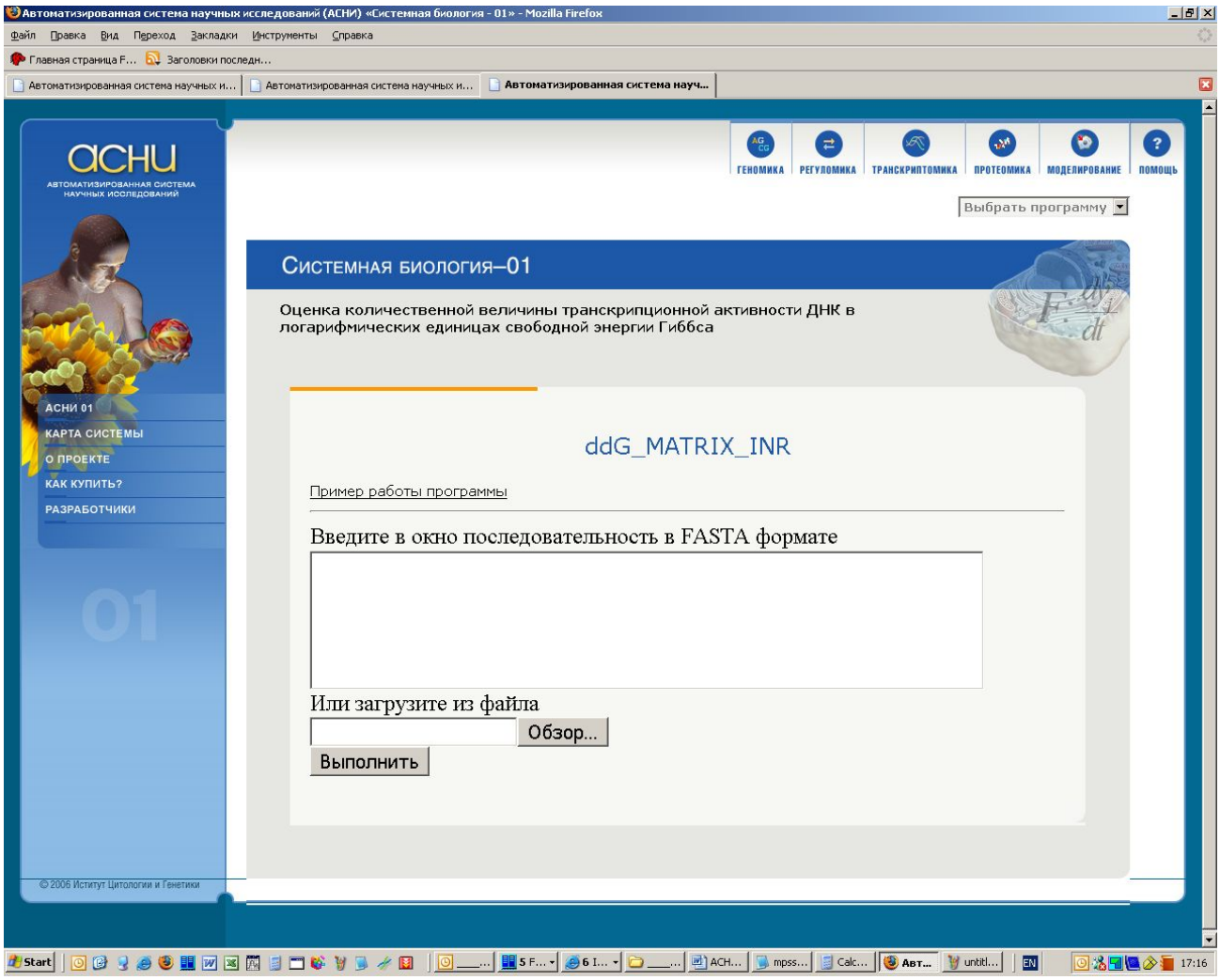


Рисунок 70. HTML-интерфейс операции, подготовленный к началу ее выполнения.

Выполнение операции начинается с ввода нуклеотидной последовательности исследуемой ДНК. Для этой цели используется раздел “Введите в окно последовательность в FASTA формате”, как это показано на Рис. 71.

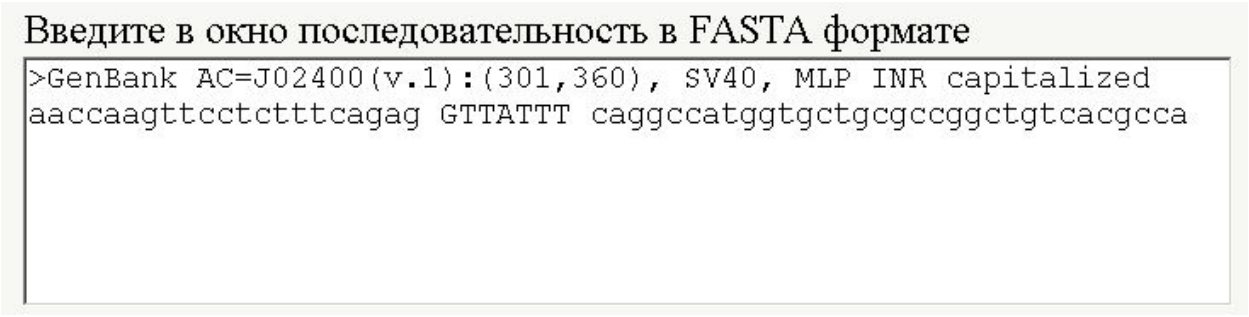


Рисунок 71. Ввод нуклеотидной последовательности ДНК в режиме "Введите в окно последовательность в FASTA формате ".

В качестве альтернативного способа ввода нуклеотидной последовательности ДНК из текстового файла используется ИНТЕРНЕТ-браузер путем выбора режима ввода “ Или загрузите из файла” и последующего нажатия клавиши “Обзор”.

После ввода данных как было описано выше необходимо нажать клавишу “Выполнить”, в результате чего появляется HTML-страница с текстовым форматом выдачи результата, как это показано ниже на Рис. 72.

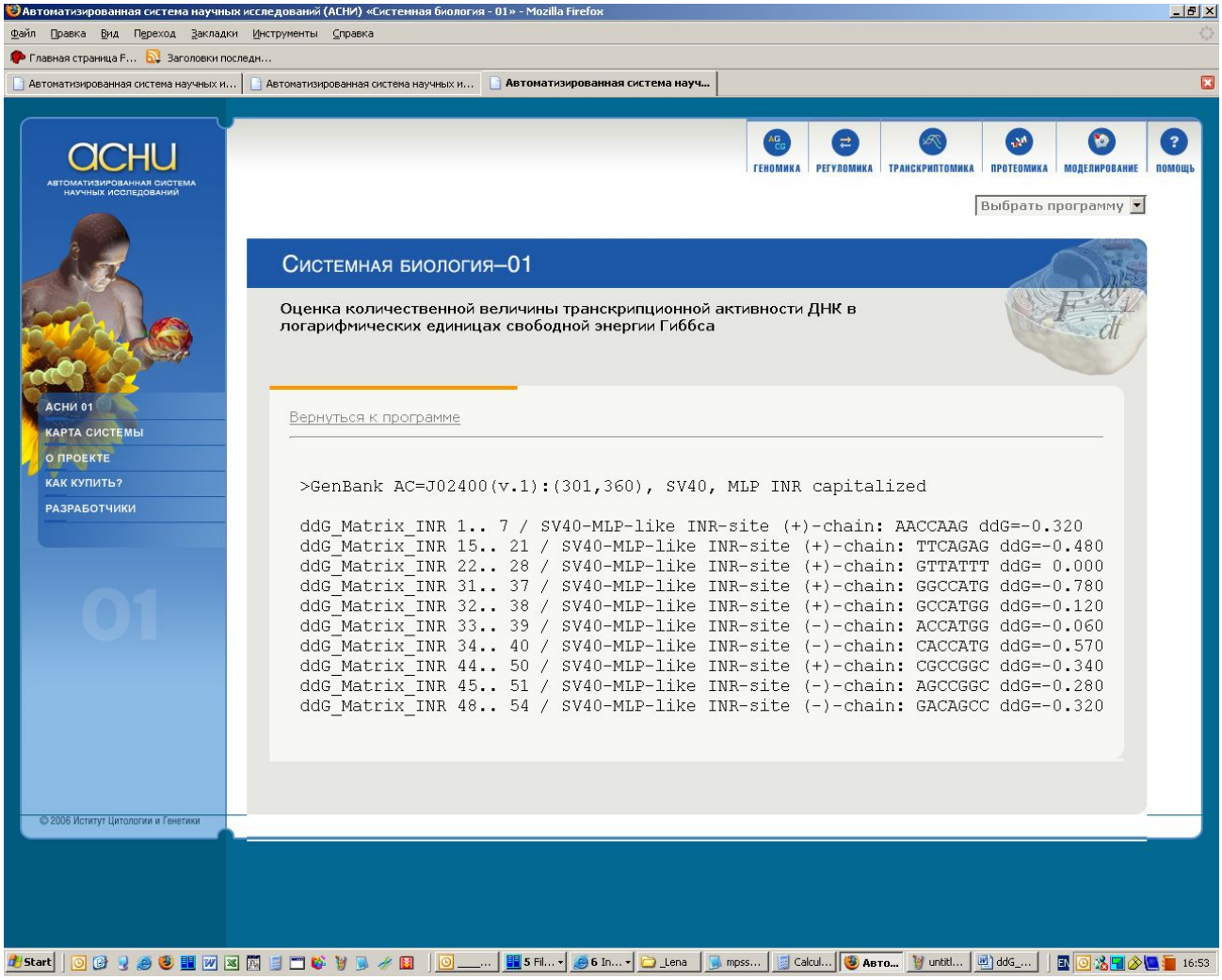


Рисунок 72. Пример HTML-страницы с текстовым форматом выдачи результата, включая оценки количественных величин транскрипционной активности ДНК (в логарифмических единицах свободной энергии Гиббса).

Здесь в первой строке приведено название анализируемой нуклеотидной последовательности ДНК. В каждой следующей строке описано 1 (один) потенциальный INR-сайт, распознанный программой и, в конце строк, оценка количественной величины его транскрипционной активности ДНК (в логарифмических единицах свободной энергии Гиббса). Это описание следует так называемому GeneBank-подобному формату.

Распознавания промоторов в геномных последовательностях методом ARGO

В связи со значительными успехами проектов секвенирования геномов в последнее время особенно большое значение приобретают методы теоретического анализа и предсказания регуляторных особенностей генетических последовательностей. 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) (Krivan,W. and Wasserman,W., 2001).

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 (Kolchanov, N.A., et al., 2002). Это обстоятельство затрудняет поиск общих контекстных закономерностей, на основе которых может осуществляться распознавание промоторов.

Для решения задачи распознавания промоторов в геномах эукариот используется программа ArgoViewer, позволяющая 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.

Описание работы ArgoViewer

Для распознавания промоторов может подаваться ДНК длиной не менее 100 и не более 100000 нуклеотидов.

Анализируемая последовательность должна быть в FASTA-формате.

Необходимо войти в подсистему «Регуломика» (смотри рис. 4.2.4-1) и выбрать раздел «Распознавание промоторов» в списке операций этой подсистемы. В результате появится HTML-страница с интерфейсом этой операции, который показан на Рис. 3.

Рисунок 73. Web-интерфейс ввода программы распознавания промоторов ARGO_Viewer.

Web-интерфейс программы ARGO_Viewer включает в себя интерфейс ввода данных для анализа (Рис. 73), окно состояний задач (Рис. 74), окно вывода результатов (Рис. 75).

Интерфейс ввода данных (Рис.73) содержит:

1. возможность задания имени задачи, которое необходимо для точной ее идентификации среди других задач, одновременно выполняемых на сервере. В том случае, если имя задачи пользователем не задано, ей автоматически присваивается имя “Unnamed №”, где №- автоматически увеличиваемый номер задачи.
2. окно ввода анализируемой последовательности. Данные должны быть записаны в FASTA-формате.
3. Выбор одного из типов распознаваемых промоторов (по тканеспецифичности экспрессии координируемых данными промоторами генов)

Рисунок 74. Окно состояний задач.

Окно состояний задач (Рис. 74) содержит список заголовков считаемых задач и информацию о статусе каждой задачи. В том случае, когда задача еще не досчитана, выдается сообщение “not finished”. Как только анализ задачи заканчивается, ее статус меняется на “finished” и пользователь может ознакомиться с результатами проведенного анализа. Если по каким-то причинам анализ задачи завершился с ошибкой, выдается протокол расчета с типом ошибки.

Кроме того, пользователь имеет возможность удалять из списка не интересующие его задачи, а также многократно просматривать результаты анализа разных задач.

Результатом работы программы является картина локализации предсказанных промоторов в окне (Рис. 75) и таблица расположения потенциальных стартов транскрипции (Рис. 76).

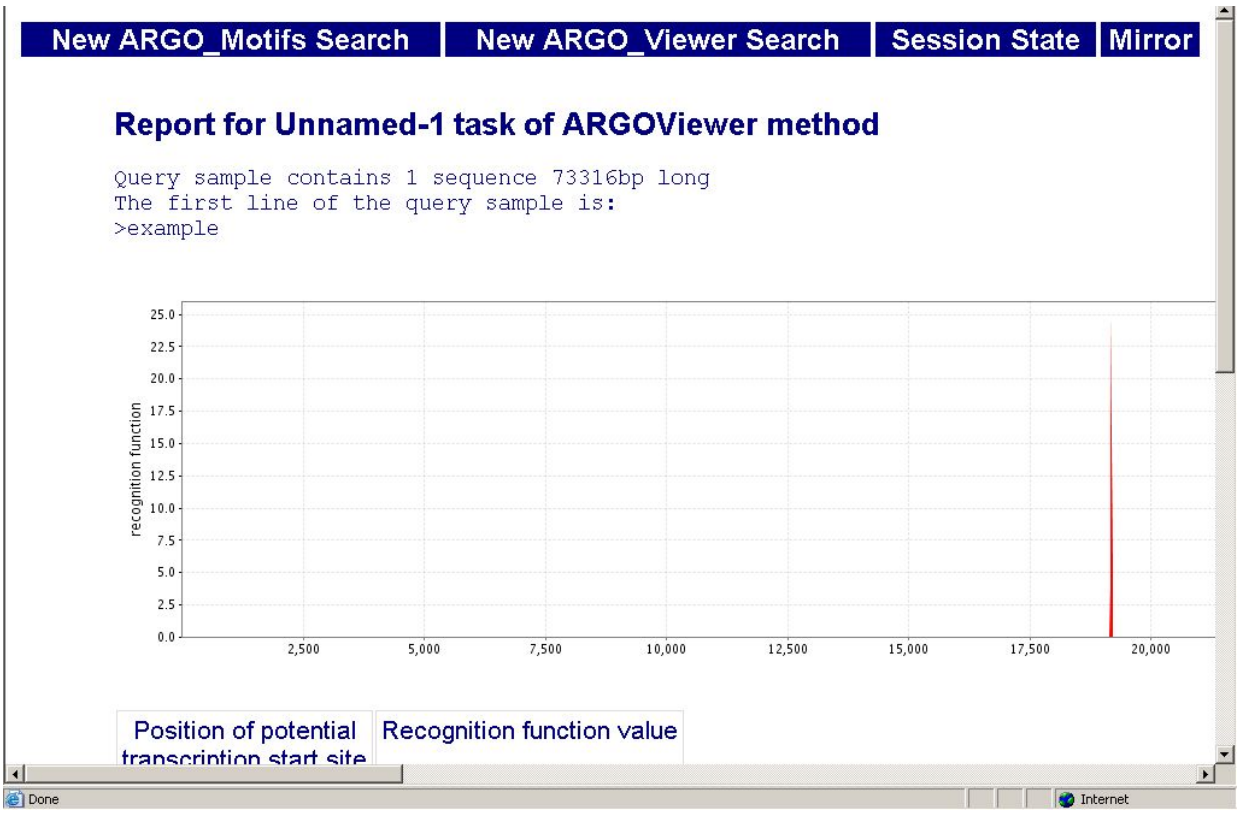
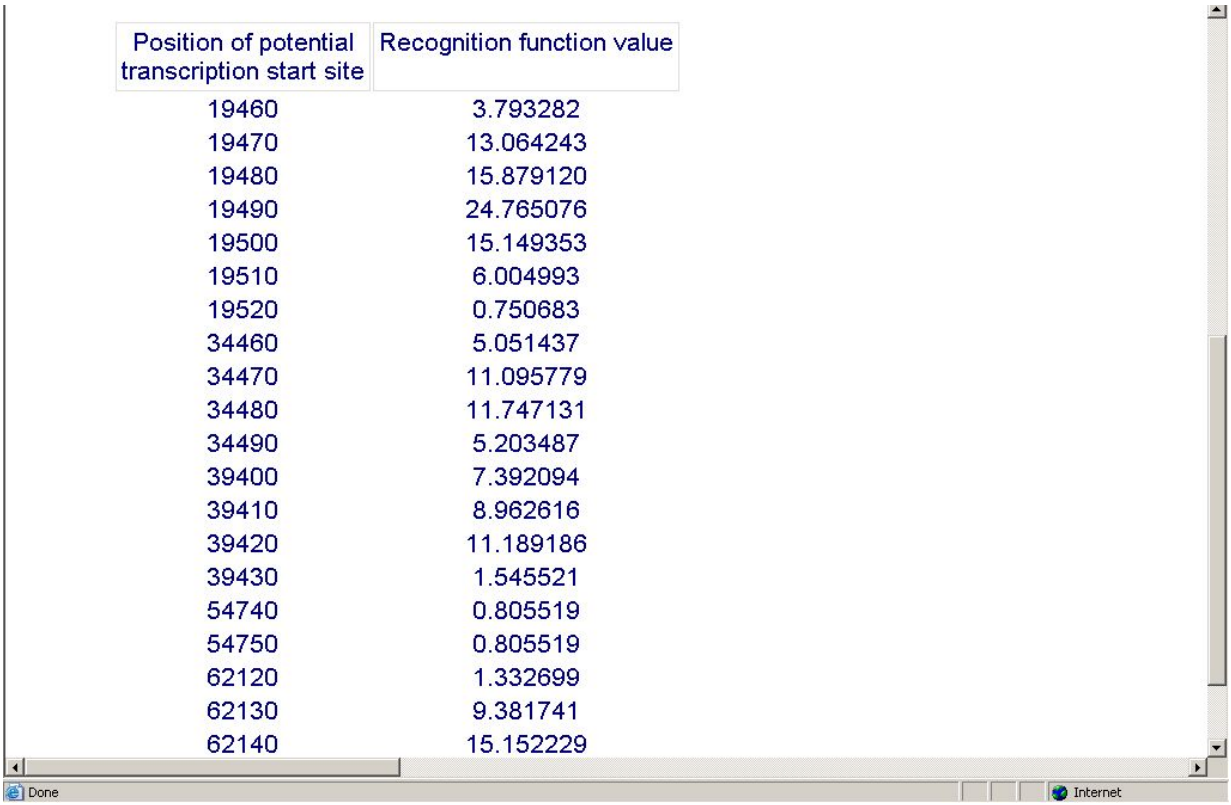


Рисунок 75. Профиль функции распознавания промоторов в протяженной геномной последовательности.



Position of potential transcription start site	Recognition function value
19460	3.793282
19470	13.064243
19480	15.879120
19490	24.765076
19500	15.149353
19510	6.004993
19520	0.750683
34460	5.051437
34470	11.095779
34480	11.747131
34490	5.203487
39400	7.392094
39410	8.962616
39420	11.189186
39430	1.545521
54740	0.805519
54750	0.805519
62120	1.332699
62130	9.381741
62140	15.152229

Рисунок 76. Таблица предполагаемых позиций стартов транскрипции.

Окно вывода результатов (рис 75 и 76) включает в себя:

1. Название задачи (рис. 75)
2. краткое описание параметров анализируемой выборки. (рис. 75)
3. таблицу предполагаемых позиций стартов транскрипции в анализируемой геномной последовательности, состоящую из двух колонок.
   Первая колонка содержит найденные позиции. Вторая – вес найденной позиции, который дает представление о достоверности того, что предполагаемая позиция действительно является стартом транскрипции заданного типа.
4. профиль функции распознавания промоторов заданного типа в анализируемой геномной последовательности (рис. 76).

3. Полезные ссылки.

1) Зенгер В.. Принципы структурной организации нуклеиновых кислот. 1987Москва: “Мир”.
2) Подколодная О.А., Степаненко И.Л. Механизмы транкрипционной регуляции эритроид-специфичных генов. //*Молекулярная* биология. 1997. V. 31. N. 4 P. 671-83.
3) Поздняков М.А., Витяев Е.Е., Ананько Е.А., Игнатьева Е.В., Подколодная О.А., Подколодный Н.Л., Лаврюшев С.В., Колчанов Н.А. Сравнительный анализ методов распознавания потенциальных сайтов связывания транскрипционных факторов. // Молекулярная Биология. 2001. 35 (6), 961-969.
4) Сингер М., Берг П. Гены и геномы. 1998. Москва: “Мир”.
5) Agalioti T., Chen G., Thanos D. Deciphering the transcriptional histone acetylation code for a human gene .*// Cell*. 2002. V. 111. N. 3. P. 381-392.
6) Berg O.G. and von Hippel P.H. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. // J. Mol. Biol., 1987. V. 193, P. 723-750.
7) Britten RJ, Kohne DE. Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. // Science. 1968. V. 161 N. 841 P. 529-40.
8) Heringa J. Detection of internal repeats: how common are they?// Curr. Opin. Struct. Biol. 1998. V. 8. P. 338-345.
9) 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.
10) Calladine,C.R. Mechanics of sequence-dependent stacking of bases in B-DNA. // J. Mol. Biol. 1982. V.161, 343-352.
11) de Mena R.I., Fernandez-Moreno M., Bornstein B., Kaguni L.S., Garesse R. Structure and regulated expression of the delta-Aminolevulinate synthase gene from Drosophila melanogaster.// *J.Biol.Chem.* 1999. V. 274 N. 52 P. 37321-37328.
12) 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.
13) Edmondson D.G. and Roth S.Y. Chromatin and transcription. // FASEB J., 1996. V. 10, P. 1173-1182. EMBO Workshop // EMBO J. 1989. V. 8. P.1-5.
14) Featherstone M. Coactivators in transcription initiation: here are your orders. // Curr Opin Genet Dev. 2002. V. 12. N. 2. P.149-55.
15) Festenstein R., Kioussis D. Locus control regions and epigenetic chromatin modifiers. // *Curr Opin Genet Dev.* 2000. V. 10. N. 2. P. 199-203.

16) Fickett, J.W. Finding genes by computer: the state of the art. // Trends Genet. 1996. V. 12, P. 316-320.
17) Fickett, J.W., Hatzigeorgiou, A.C. Eukaryotic promoter recognition. //Genome Res. 1997, V. 7. P. 861-878.
18) Fishburn, P.C. Utility theory for decision making. New York, Jonh Wiley & Sons. 1970.
19) Forster E. and Ronr B. Methoden der korrelations- und regressions analyse. Verlag Die Wirtschaft, Berlin. 1979.
20) 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.
21) French K., Quandt K. and Werner T. Muscle actin genes: a first step towards computational classification of tissue specific promoters. // In Sylico Biology, 1998. V. 1. N. 1. P. 29-38.
22) Godde J.S., Nakatani,Y. and Wolffe A.P. The amino-terminal tails of the core histones and the translational position of the TATA box determine TBP/TFIIA association with nucleosomal DNA. // Nucleic Acids Res. 1995. V. 23, P. 4557-4564.
23) Haussler, D. Computational genefinding. // Trends Guide in Bioinformatics. 1998. V.1, P. 12-15.
24) 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.
25) Hayes K.G., Perl M.L. and Efron B. Application of the bootstrap statistical method to the tau-decay-mode problem. // Phys. Rev. D. Part. Fields. 1989. V. 39, P. 274-279.
26) Hyde-DeRuyscher R., Jennings E. and Shenk T. DNA binding sites for the transcriptional activator/repressor YY1. // Nucleic Acids Res., 1995. V. 23. P. 4457-4465.
27) 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.
28) Javahery R., Khachi A., Lo K., Zenzie-Gregory B. and Smale S.T. DNA sequence requirements for transcriptional initiator activity in mammalian cells. // Mol. Cell. Biol. 1994. V. 14. N. 1. P. 116-127.
29) Jonsson J., Norberg T., Carlsson L., Gustafsson C. and Wold S. Quantitative sequence-activity models (QSAM) - tools for sequence design. // Nucleic Acids Res. 1993. V. 21, P. 733-739.
30) 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.
31) Khlebodarova T.M., Oshchepkov D.Y., Podkolodnaya O.A., Miginsky D.S., Ananko E.A., Ignatieva E.V. ARTSITE DATABASE: Comparison of *in vitro* selected and natural binding sites of eukaryotic transcription factors. In: Bioinformatics of Genome Regulation and Structure II. Ed. By N. Kolchanov and R. Hofestaedt, 2005. Springer Science+Business Media, Inc.,
32) Kolchanov N.A., Ignatieva E.V., Ananko E.A., Podkolodnaya O.A., Stepanenko I.L., Merkulova T.I., Pozdnyakov M.A., Podkolodny N.L., Naumochkin A.N., Romashchenko A.G. Transcription Regulatory Regions Database (TRRD): its status in 2002. // Nucleic Acids Res. 2002. V. 30. N. 1. P. :312-7.
33) 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.

34) Kouzarides T. Histone methylation in transcriptional control. // *Curr Opin Genet Dev.* 2002. V. 12. N. 2 P. 198-209.
35) Kovalerchuk B., Vityaev E. Data Mining in Finance: Advances in Relational and Hybrid methods, Kluwer Academic Publishers. 2000. P.308.
36) Kovalerchuk B., Vityaev E., Ruiz J.F. Consistent and Complete Data and "Expert" Mining in Medicine, In: Medical Data Mining and Knowledge Discovery (Book chapter), Springer, 2001. P. 238-280.
37) Kraus R.J., Murray E.E., Wiley S.R., Zink N.M., Loritz K., Gelembiuk G.W., Mertz J.E. Perimentally determined weight matriN definitions of the initiator and TBP binding site elements of promoters. // Nucleic Acids Res. 1996. V.24. N. 8. P. 1531-1539.
38) Krivan,W. and Wasserman,W. A predictive model for regulatory sequences directing liver-specific transcription. //*Genome Res.* 2001. V. **11**. P. 1559–1566.
39) 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.
40) Lescot M., Dehais P., Thijs G., Marchal K., Moreau Y., Van de Peer Y., Rouze P., Rombauts S. PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. //Nucleic Acids Res. 2002. V. 30, P. 325-327.
41) Linder M.W, Prough R.A. Developmental aspects of glucocorticoid regulation of polycyclic aromatic hydrocarbon-inducible enzymes in rat liver. // Arch Biochem Biophys. 1993. V. 302. N. 1 P. 92-102.
42) Matys V., Fricke E., Geffers R., Gossling E., Haubrock M., Hehl R., Hornischer K, Karas D., Kel A.E., Kel-Margoulis O.V., Kloos D.U., Land S., Lewicki-Potapov B., Michael H., Munch R., Reuter I., Rotert S., Saxel H., Scheer M., Thiele S., Wingender E. TRANSFAC: transcriptional regulation, from patterns to profiles. //Nucleic Acids Res**.** 2003. V. **31**. P. 374-378.
43) 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.
44) Meierhans D., Sieber M., Allemann R.K. High affinity binding of MEF-2C correlates with DNA bending. //Nucleic Acids Res 1997; 25:4537–44.
45) 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, 789-800.
46) Oshchepkov D.Y., Vityaev E.E., Grigorovich D.A., Ignatieva E.V., Khlebodarova T.M. SITECON: a tool for detecting conservative conformational and physicochemical properties in transcription factor binding site alignments and for site recognition// Nucleic Acids Res. 2003. V. 32 W208-12.
47) Munch R., Hiller K., Barg H., Heldt D., Linz S., Wingender E., Jahn D. PRODORIC: prokaryotic database of gene regulation. // Nucleic Acids Res. 2003. V. 31. P. 266-269.
48) 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.// Bioinformatics of genome regulation and structure. Kluwer Academic Publishers, Boston/Dordrecht/London N.Kolchanov and R.Hofestaedt (ed.) 2003. P. 93-102.
49) Qiu P. Recent advances in computational promoter analysis in understanding the transcriptional regulatory network// Biochem Biophys Res Commun. 2003. V. 309 N. 3. P. 495-501.
50) Staden R. Method for calculating the probabilities of finding patterns in sequences. // Comput. App. Biosci., 1989, V.5, N.2, P. 89-96.

51) Ponomarenko M.P., Kolchanova A.N., Kolchanov N.A. Generating programs for predicting the activity of functional sites. // J. Comput. Biol., 1997. V. 4, P. 83-90.
52) Ponomarenko J.V., Orlova G.V., Ponomarenko M.P., Lavryushev S.V., Frolov A.S., Zybova S.V., Kolchanov N.A. SELEX_DB: an activated database on selected randomized DNA/RNA sequences addressed to genomic sequence annotation. //Nucleic Acids Res. 2000. V. **28. P.** 205-208.
53) Ponomarenko J.V., Ponomarenko M.P., Frolov A.S., Vorobyev D.G., Overton G.C., and Kolchanov N.A. Conformational and physicochemical DNA features specific for transcription factor binding sites. // Bioinformatics. 1999. V. 15, N. 7/8, P. 654-668
54) Praz V., Perier R., Bonnard C., Bucher P. The Eukaryotic Promoter Database, EPD: new entry types and links to gene expression data. //Nucleic Acids Res. 2002. V. 30. P. 322-324.
55) Robison K., McGuire A.M. and Church G.M. A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K-12 genome. // J. Mol. Biol., 1998. V. 284, P. 241-254.
56) Roulet, E., Fisch, I., Bucher, P., and Mermod, N. Evaluation of computer tools for prediction of transcription factor binding sites on genomic DNA. // In Silico Biology, 1998. V.1, P. 21-28.
57) Sandelin A., Alkema W., Engstrom P., Wasserman W.W., Lenhard B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. // Nucleic Acids Res. 2004. V. 32. P. D91-94.
58) Shpigelman E.S., Trifonov E.N. and Bolshoy A. CURVATURE: software for the analysis of curved DNA. // Comput. Appl. Biosci. 1993. V. 9. P. 435-440.
59) Shulzaberger R.K. and Schneider T.D. Using sequence logos and informational analysis of Lrp DNA binding sites to investigate discrepancies between natural selection and SELEX. // Nucleic Acids Res. 1999. V. 27, P. 882-887.
60) 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.
61) 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.
62) Suzuki M., Yagi N. and Finch J.T. Role of base-backbone and base-base interactions in alternating DNA conformations. // FEBS L. 1996. V. 379, P. 148-152.
63) van Holde KE. Chromatin. 1989. New York: Springer-Verlag.
64) Vishnevsky O.V. and Kolchanov N.A. ARGO: a web system for the detection of degenerate motifs and large-scale recognition of eukaryotic promoters. //Nucleic Acids Res. 2005. V. 33. P. 417-422
65) Vityaev E., Kovalerchuk B. Empirical Theories Discovery based on the Measurement Theory. // Mind and Machine. 2004. V. 14. N. 4 P. 551-573.
66) Vityaev E., Kovalerchuk B. Data Mining For Financial Applications.//Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers, In: O. Maimon and L. Rokach (eds.). 2005a. Kluwer Academic Publishers, Boston, forthcoming.
67) Vityaev E., Kovalerchuk B. Relational Methodology for Data Mining and Knowledge Discovery. Sixteenth International Workshop on Database and Expert Systems Applications// 1st International Workshop on Philosophies and Methodologies for Knowledge discovery (22-26 August 2005, Copengagen, Denmark), IEEE Computer Society, 2005b. P. 725-729.
68) Zhang M.Q. Identification of human gene core-promoters in silico. //Genome Res. 1998. V. 8. P. 319-326.
69) Tanikawa J., Yasukawa T., Enari M., Ogata K., Nishimura Y., Ishii S., Sarai A. Recognition of specific DNA sequences by the c-Myb protooncogene product: role of three repeat units in the DNA-binding domain. // Proc. Natl. Acad. Sci. U.S.A. 1993. V.90. N.20. P. 9320-9324.

70) Weston K. Extension of the DNA binding consensus of the chicken c-Myb and v-Myb proteins. // Nucleic Acids Res., 1992. V. 20, P. 3043-3049.
71) Wingender E., Chen X., Fricke E., Geffers R., Hehl R., Liebich I., Krull M., Matys V., Michael H., Ohnhauser R., Pruss M., Schacherer F., Thiele S., Urbach. The TRANSFAC system on gene expression regulation. // Nucleic Acids Res. 2001. V. 29. P. 281-283.
72) Zadeh L.A. Fuzzi sets. // Information and Control. 1965. V. 8. P. 338-353.