

# Подсистема «Компьютерная Транскриптомика»: анализ профилей экспрессии генов.

## Структура документа (оглавление).

1. Цель и задачи подсистемы «Компьютерная Транскриптомика».....	1
Транскриптом – совокупность транскриптов генов.....	1
Задачи транскриптомики. ....	2
2. Структура подсистемы «Компьютерная Транскриптомика» и детальное руководство по ее применению. ....	3
2.1. Информационные компоненты подсистемы «Компьютерная Транскриптомика».....	4
2.1.1. Описание Базы знаний по профилям генной экспрессии (БЗ-ПГЭ).....	4
Описание работы с БЗ-ПГЭ.....	4
2.2. Программные компоненты подсистемы «Компьютерная Транскриптомика».....	9
2.2.1. Система ввода данных. ....	9
2.2.1.1. Формат файлов.....	9
2.2.1.2. Описание работы системы ввода данных .....	11
2.2.2. Программные компоненты для предварительной математической обработки, фильтрации и статистического анализа биочиповых экспрессионных данных .....	14
2.2.2.1. Описание работы сценария «Вычисление статпараметров: среднего, дисперсии, отклонения, вариации».....	15
2.2.3. Выявление дифференциально экспрессирующихся генов .....	18
2.2.3.1. Описание работы сценария «Расчет критерия кратной разницы» .....	19
2.2.3.2. Описание работы сценария «Расчет t-критерия Уэлша».....	22
2.2.3.3. Описание работы сценария «Выявление конститутивных профилей экспрессии генов».....	24
2.2.3.4. Описание работы сценария «Выявление индуцибельных профилей экспрессии генов».....	27
2.2.3.5. Описание работы сценария «Выявление циклических профилей экспрессии генов».....	28
2.2.4. Кластеризация и классификация профилей транскрипционной экспрессии генов. ....	30
2.2.4.1. Описание работы сценария «Иерархическая кластеризация профилей экспрессии генов».....	33
2.2.4.2. Описание работы сценария «Анализ профилей экспрессии генов методом главных компонент (РСА)» .....	35
2.2.4.3. Описание работы сценария «Анализ профилей экспрессии генов с помощью самоорганизующихся карт признаков (SOM)».....	37
2.2.5. Выявление генов с коррелирующей экспрессией .....	40
2.2.5.1. Описание работы сценария «Поиск генов-представителей SOM-кластеров» ..	41
2.2.5.2. Описание работы сценария «Вычисление матрицы межгенных корреляций» ..	43
3. Полезные ссылки. ....	45

## 1. Цель и задачи подсистемы «Компьютерная Транскриптомика».

### Транскриптом – совокупность транскриптов генов.

Совокупность транскриптов всех генов, экспрессирующихся в какой-либо клетке на какой-либо стадии ее развития, называется транскриптом клетки. Соответственно,

существуют также транскриптомы ткани, органа и организма. Поскольку транскрипты являются продуктами транскрипции генов, их совокупность представляет собой первый уровень фенотипа, т.е. первый уровень развертывания и реализации генетической информации, заключенной в геноме (Hack, 2004; Mendes Soares, Valcarcel, 2006).

Структура транскриптома сложно организована и постоянно изменяется, поскольку зависит от стадий клеточного цикла, от типа клеток и тканей, от стадий развития организма, от состояния (норма-болезнь) тканей и органов, от наличия внешних сигналов как для самой транскрипции генов, так и для различных пост-транскрипционных процессов – сплайсинга, редактирования, взаимодействия с микроРНК (miRNA) или короткими интерферирующими РНК (siRNA). Иными словами, транскриптому изначально присуща пространственная дифференциальность и высокая динамичность в распределении транскриптов разных генов и изоформ транскриптов отдельного гена (Hu et al., 2005; Rhodes, Chinnaiyan, 2005).

### **Задачи транскриптомики.**

Исследование структуры и динамики транскриптома, лежащих в основе формирования второго уровня фенотипа – протеома клеток, тканей и т.д., является задачей транскриптомики.

Некоторые задачи транскриптомики в то же время являются также и задачами функциональной геномики - в той мере, в какой информация, необходимая для формирования структуры транскриптома, закодирована в геноме и может быть выявлена при его изучении.

Задачи функциональной геномики, перекрывающиеся с задачами транскриптомики, заключаются в выявлении и исследовании условий для формирования характерной для клеток определенных типов структуры транскриптома, предопределенной генетическими программами развития, т.е. зашифрованными в геноме сигналами. Эти условия реализуются через взаимодействие клеточной биохимической машины с наследуемыми вместе с геномом регуляторными сигналами: (1) транскрипции (промоторы, сайленсеры, энхансеры, граничные элементы/инсуляторы, сигналы завершения синтеза РНК-полимеразой и т.д.), (2) сплайсинга (сайты сплайсинга, регуляторы альтернативного сплайсинга, транс-сплайсинга) и (3) пост-транскрипционных процессов (сайты взаимодействия с miRNA или siRNA, редактирования, цитоплазматической локализации и т.д.).

Собственные задачи транскриптомики состоят в выявлении и изучении не зависящих от геномных сигналов факторов, влияющих на формирование структуры и динамики транскриптома, и, в конечном счете, протеома.

Применение методов биоинформатики на уровне анализа транскриптома, а именно при исследовании структуры транскриптов и их дифференциального временного и пространственного распределения в клетках и организмах, позволяет:

- реконструировать коды, заключенные в геноме (кооперация с геномикой);
- выявлять информацию в виде сигналов и кодов, необходимую для формирования протеома (кооперация с протеомикой).

Существует несколько распространенных методов для широкомасштабного исследования транскриптома - «серийный анализ экспрессии генов» - Serial analysis of gene expression (SAGE) (Velculescu et al., 1995; 2000) , «прочитанные фрагменты экспрессированных последовательностей» - Expressed Sequence Tags (ESTs) (Okubo et al., 1992; Lindlof, 2003), «Массовое одновременное секвенирование характерных фрагментов» Massively Parallel Signature Sequencing (MPSS) (Brenner et al., 2000).

Однако на настоящее время только метод ДНК-биочипов или ДНК микроматриц (DNA microarray, DNA biochip, oligonucleotide microarray, cDNA microarray) является средством

общегеномного и высокопродуктивного исследования транскриптома (Cheung et al., 1999; Duggan et al., 1999; Greenfield, 2000; Deyholos, Galbraith, 2001; Мирзабеков, 2003; Zhu, 2003; ). ДНК-биочипы – это миниатюризированные матрицы или подложки, на которых в определенном порядке распределены фрагменты ДНК, соответствующие отдельным генам или их частям. Такие организованные микроматрицы позволяют проводить эксперименты по одновременному анализу структуры и экспрессии тысяч генов с помощью параллельной гибридизации (Stears et al., 2003). Высокоразвитые методы преобразования результатов этих экспериментов в цифровые данные и методы компьютерной обработки последних обеспечивают возможность анализировать и сопоставлять экспрессию таких массивов генов во множестве экспериментальных условий (Holloway et al., 2002; Moreau et al., 2003; Allison et al., 2006). Таким образом получается статическая информация об экспрессии генов (в какой ткани или типе клеток, на какой стадии, при каком воздействии и т.д.) и динамическая информация об экспрессии генов при сопоставлении данных отдельных экспериментов (Anderle et al., 2003).

## **2. Структура подсистемы «Компьютерная Транскриптомика» и детальное руководство по ее применению.**

Подсистема «Компьютерная транскриптомика» является аналитическим инструментом для исследования качественных и количественных характеристик экспрессии генов и реконструкции ген-генных взаимодействий при формировании биомедицински и биотехнологически значимых фенотипических признаков эукариотических и прокариотических организмов.

Подсистема «Компьютерная Транскриптомика» представляет собой информационно-программный комплекс, предназначенный для проведения анализа профилей экспрессии генов с целью их кластеризации и классификации, выявления групп коэкспрессирующихся генов в результате решения задач:

1. проводить предварительную математическую обработку файлов с данными ДНК-биочип-экспериментов, их фильтрацию и статистический анализ;
2. накапливать и визуализировать профили транскрипционной экспрессии генов;
3. выявлять дифференциально экспрессирующиеся гены,
4. проводить кластеризацию и классификацию профилей экспрессии генов;
5. выявлять гены с коррелирующей экспрессией,
6. сохранять списки выявленных групп генов в базе знаний;
7. предоставлять интерфейс для запрашивания, доставки и просмотра данных по профилям экспрессии генов.

Подсистема рассчитана на пользователей с различной подготовкой в области анализа биочиповых данных. Начинаящие пользователи могут просматривать результаты анализа имеющихся в БЗ-ПГЭ данных по заранее заготовленным сценариям карт, графиков и таблиц. Пользователи с некоторым опытом анализа биочиповых данных могут помещать свои таблицы в хранилище данных, обрабатывать их по имеющимся сценариям, менять некоторые параметры в сценариях и скачивать через Интернет результаты обработки в виде текстовых файлов и рисунков на свой локальный компьютер.

Основными компонентами подсистемы являются данные, сценарии их обработки и средства визуализации получаемых результатов.

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

## **2.1. Информационные компоненты подсистемы «Компьютерная Транскриптомика».**

### **2.1.1. Описание Базы знаний по профилям генной экспрессии (БЗ-ПГЭ).**

База знаний по профилям экспрессии биомедицински- и биотехнологически-значимых генов (БЗ-ПГЭ) предназначена для хранения данных и метаданных об измеренных с помощью экспрессионных биочипов профилях экспрессии биомедицински и биотехнологически значимых генов человека, животных, растений, микроорганизмов.

База знаний ПГЭ содержит:

1) описание биочиповых исследований:

- содержание, целей и задач исследования;
- типа дизайна, плана и условий проведения исследования;
- типа и обозначения биочипов, использованных в аннотируемых исследованиях;
- названия и обозначения образцов;
- структуры и содержания файлов с числовыми данными;

2) описание профилей генной экспрессии, полученных в аннотированных биочиповых исследованиях:

- результатов измерений содержания транскриптов генов в определенном исследовании;
- результатов анализа профилей генов, их кластеризации и классификации.

БЗ-ПГЭ содержит также файлы результатов анализа файлов с числовыми данными и профилей генов в графической форме.

Заполнение БЗ-ПГЭ проводится путем обработки информации из баз экспрессионных биочиповых данных GEO-NCBI (Barrett et al., 2005), ArrayExpress-EBI (Parkinson et al., 2005), SMD (Ball et al., 2005), придерживающихся стандартов описания биочиповых экспериментов (Stoeckert et al., 2002), необходимых для достижения сравнимости и биологической интерпретативности результатов биочиповых экспериментов (Moreau et al., 2003).

*Описание работы с БЗ-ПГЭ.*

### **Поиск документов о биочиповых исследованиях, аннотированных в БЗ-ПГЭ, и о профилях генной экспрессии, установленных в этих исследованиях.**

Получение информации возможно, если в БД содержится такая информация. БЗ-ПГЭ настоящей версии содержит в основном данные по исследованию липидного метаболизма человека и мыши, дифференцировки адипоцитов и миоцитов, а также демонстрационные данные по профилированию экспрессии генов пациентов болезни Хантингтона с помощью двух биочип-платформ.

Если пользователь не имеет своих подготовленных файлов или желает сначала ознакомиться с процедурой анализа данных в подсистеме «Компьютерная Транскриптомика», он может выбрать из списка биочиповых исследований, аннотированных в БЗ-ПГЭ, какое-либо исследование, ознакомится с содержанием описания исследования. Пользователь также может затем загрузить в свой проект файлы из демонстрационного проекта, посвященного этому исследованию, чтобы получить представление о форматах

исходных файлов, об алгоритмах, примененных для анализа данных, и о форматах файлов с результатами анализа. Пользователь также может провести собственный анализ этих данных, меняя предоставленные параметры.

Для входа на страницу поисковой системы БЗ-ПГЭ пользователь сначала должен завести свой профиль на странице входа в подсистему «Компьютерная Транскриптомика» (Рис. 1).

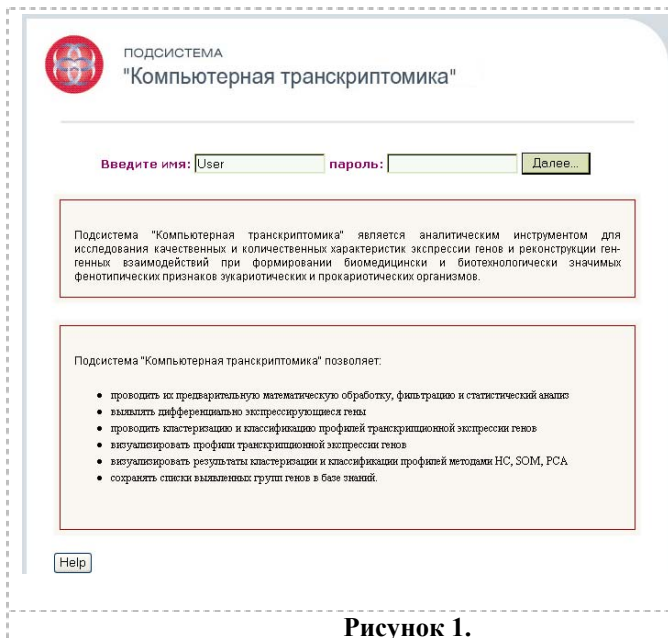


Рисунок 1.

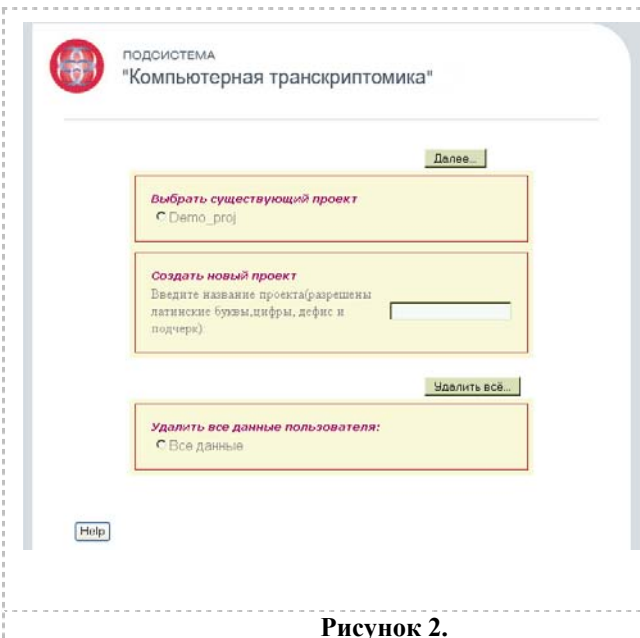


Рисунок 2.

Набрав свой идентификатор и пароль, зарегистрированные для пользователя системы «Системная биология», и ознакомившись с возможностями подсистемы «Компьютерная Транскриптомика», пользователь нажимает на кнопку «Далее» и перемещается на страницу заведения личного проекта (Рис. 2).

Пользователь может выбрать демо-проект, чтобы разобраться с форматами исходных файлов, а также исследовать алгоритмы, примененные для анализа данных, и файлы с результатами анализа.

После введения название своего проекта (в соответствии с указаниями слева) и нажатия на кнопку «Далее» страница обновляется с появлением строки с введенным названием проекта (Рис. 3).

На этой странице пользователь может выбирать между заведенными в его профиле проектами и переходить к анализу данных в этих проектах. Он также может добавлять новые проекты или удалять старые.

Нажатие на кнопку «Далее» перемещает пользователя на страницу выбора задачи, на которой после нажатия на знак «+» в строке с названием «База данных БЗ-ПГЭ (Профили генной экспрессии)» открывается ссылка «Просмотр информации и поиск по условию» (Рис. 4).

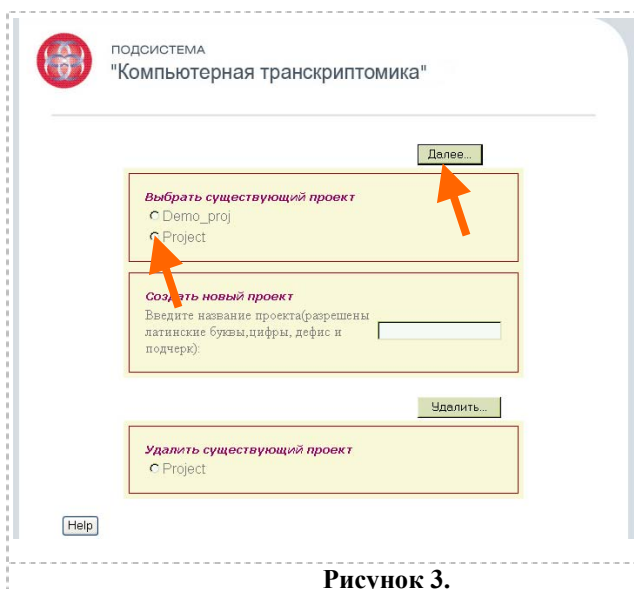


Рисунок 3.

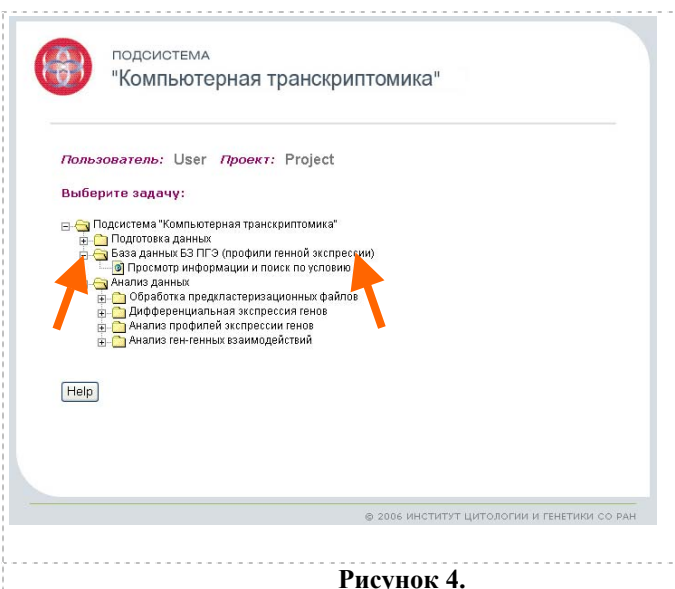


Рисунок 4.

Нажатие на ссылку «Просмотр информации и поиск по условию» переносит пользователя на страницу поисковой системы БЗ-ПГЭ (Рис. 5).

Пользователю предлагается два способа перейти к интересующему его биочип-исследованию:

- Первый способ позволяет перейти к просмотру документов БЗ-ПГЭ, описывающих проаннотированные биочип-исследования, для общего ознакомления с содержанием проаннотированных экспериментов и оформления цели поиска (Рис. 6).
- Второй способ позволяет перейти к просмотру профилей экспрессии какого-либо определенного гена, полученного в проаннотированных биочип-исследованиях с участием проб этого гена (Рис. 11).

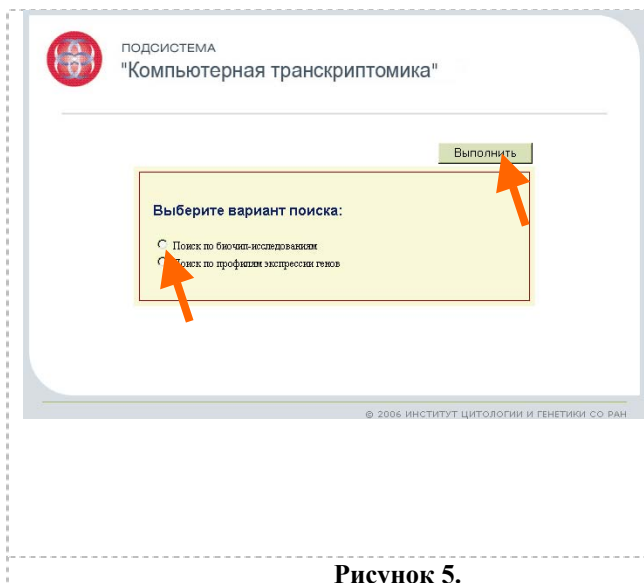


Рисунок 5.

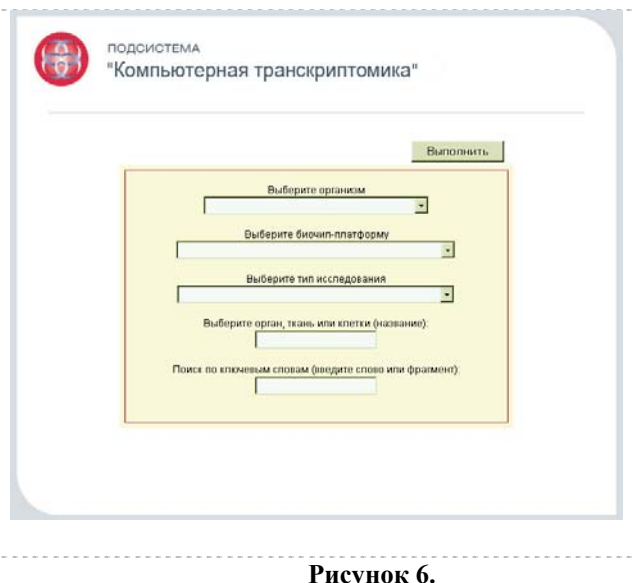


Рисунок 6.

На странице «Поиск по биочип-исследованиям» предлагается отселектировать документы с описанием проаннотированных биочип-исследований для просмотра. Можно провести селекцию: (1) по названию организма, гены которого в виде проб биочипа или образцы были использованы в эксперименте; (2) по названию биочип-платформы, с помощью которой проводили эксперименты; (3) по названию или обозначению органа, ткани или клеток, использованных для приготовления образцов. Можно также провести контекстный поиск по любому слову, например, имени автора эксперимента или названию журнала, где были опубликованы результаты эксперимента, и т.д. (Рис. 6).

Например, пользователя интересуют биочип-исследование, опубликованное автором «Borovecki». Он набирает имя автора в поле «Поиск по ключевым словам (введите слово или фрагмент)». После нажатия на кнопку «Далее» пользователю предлагается документ с описанием соответствующих биочип-исследований (Рис. 7).

подсистема  
"Компьютерная транскриптомика"

Выполнить

Выберите организм

Выберите биочип-платформу

Выберите тип исследования

Выберите орган, ткань или клетки (название):

Поиск по ключевым словам (введите слово или фрагмент):  
Borovecki

Рисунок 7.

подсистема  
"Компьютерная транскриптомика"

Результаты поиска по биочип-исследованиям

Найдено исследований: 2

Условие на поиск:  
Ключевое слово (фрагмент) Borovecki

Наименование исследования:  
Huntington's disease: peripheral blood expression profile (HG-U133A) [Homo sapiens].  
Analysis of blood samples of 5 presymptomatic and 12 symptomatic Huntington's disease (HD) patients. Studies suggest that gene expression may be altered in a variety of tissues in HD, including peripheral blood. Results identify potential markers for HD.

Цель исследования

pub. links  
PubMed ID: 16043692  
Borovecki F, Lovrecic L, Zhou J, Jeong H, Then F, Rosas HD, Hersch SM, Hogarth P.

Рисунок 8.

После просмотра содержания документа с описанием биочип-исследования пользователю предлагается возможность переслать эти данные в свой проект в рамках подсистемы «Компьютерная Транскриптомика» для проведения анализа этих данных и получения представления об алгоритмах и процедурах анализа биочиповых данных (Рис. 9).

GSM30589: N10  
GSM30590: N11  
GSM30591: N12  
GSM30592: N13  
GSM30593: N14

Число образцов: 31

Название групп: 5 presymptomatic and 12 symptomatic Huntington's disease (HD) patients versus 14 healthy controls.

Группа образцов

Переслать данные эксперимента в проект

Наименование исследования:  
Huntington's disease: peripheral blood expression profile (Codetlink Uniset 20K) [Homo sapiens].  
Analysis of blood samples of 5 presymptomatic and 12 symptomatic Huntington's disease (HD) patients. Studies suggest that gene expression may be altered

Цель исследования

Рисунок 9.

подсистема  
"Компьютерная транскриптомика"

Пользователь: User Проект: Project

Выберите задачу:

- Подсистема "Компьютерная транскриптомика"
  - Подготовка данных
  - База данных БЗ ПГЭ (профили генной экспрессии)
  - Анализ данных
    - Обработка предластеризационных файлов
    - Дифференциальная экспрессия генов
    - Анализ профилей экспрессии генов
    - Анализ ген-генных взаимодействий

Help

© 2006. ИНСТИТУТ ЦИТОЛОГИИ И ГЕНЕТИКИ СО РАН

Рисунок 10.

Нажатие на ссылку «Переслать данные эксперимента в проект» перемещает пользователя на главную страницу заведенного пользователем проекта с выбором задач (Рис. 10), предоставляющую переход к программным компонентам подсистемы «Компьютерная Транскриптомика».

Если пользователь решил поискать биочип-исследование по обозначению/названию гена или пробы, представляющей ген на биочипе, то, выбрав пункт «Поиск по профилям экспрессии генов» и нажав на кнопку «Выполнить» (Рис. 11), он перемещается на страницу поиска по этим идентификаторам (Рис. 12).

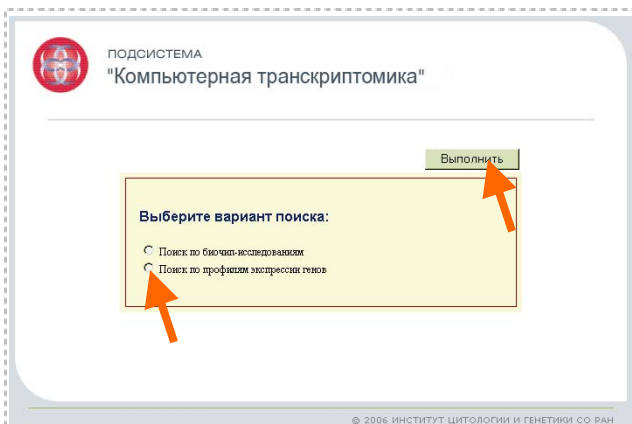


Рисунок 11.

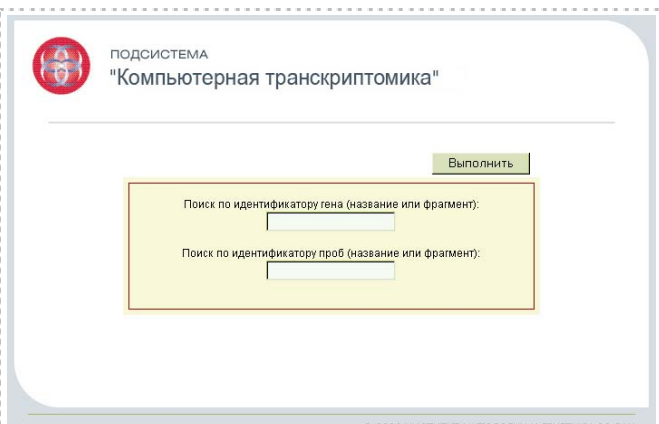


Рисунок 12.

Поиск можно провести, вставив соответствующее ключевое слово в окошко под названием «Поиск по идентификатору гена (название или фрагмент):» или в окошко под названием «Поиск по идентификатору проб (название или фрагмент):». Например, пользователь ввел идентификатор референсной мРНК (NM\_001533) для гена «Homo sapiens heterogeneous nuclear ribonucleoprotein L (HNRPL), transcript variant 1» (Рис. 13). Поисковая система возвращает страницу с результатом поиска (Рис. 14).

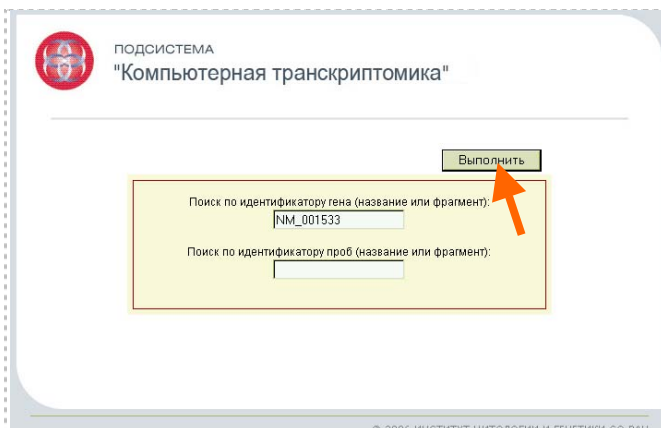


Рисунок 13.

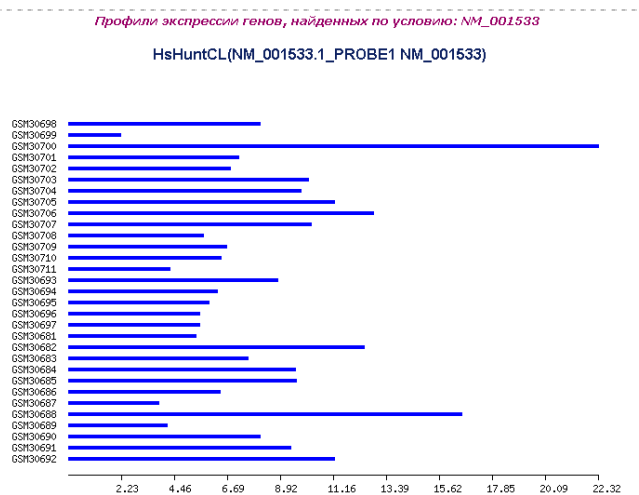


Рисунок 14.

Если профиль гена, полученный в биочип-исследовании, обозначенном над гистограммой, заинтересовал пользователя, он может вернуться к странице с входом в БЗ-



ПГЭ и, используя поиск среди документов описания биочип-исследований, перейти к подробному изучению материалов этого исследования, а затем, загрузив данные этого исследования, - к их анализу с помощью программных компонент подсистемы «Компьютерная Транскриптомика».

После выполнения запланированных процедур пользователь имеет возможность сохранить результаты работы подсистемы с помощью стандартных средств интернет-браузера в виде текстового, графического или html-файла.

## **2.2. Программные компоненты подсистемы «Компьютерная Транскриптомика»**

Программные компоненты подсистемы «Компьютерная Транскриптомика» формируют 11 независимых сценариев в зависимости от задачи. Как только пользователь заводит в своем профиле новый проект по анализу данных биочиповых-исследований и вводит название этого проекта, заводится отдельная папка, в которую копируются все сценарий со ссылками на все необходимые входные, промежуточные и выходные файлы, названия которых содержат название проекта в виде «корня», а специфика файлов задается «суффиксами» и расширениями.

О структуре подсистемы, о назначении программных компонент, формате файлов можно узнать из вспомогательного документа, доступного по ссылке при кнопке «Help» на главной странице подсистемы с выбором задач или по ссылкам, ассоциированным с названием задач в строках со знаками «+» и позволяющим перейти напрямую к разделу вспомогательного документа, относящемуся к соответствующей задаче.

### **2.2.1. Система ввода данных.**

Программные компоненты подсистемы «Компьютерная Транскриптомика» имеют свою специфическую систему ввода данных. Для работы необходимо загрузить основной файл с данными биочип-эксперимента и вспомогательные файлы с информацией об взаимоотношениях между образцами.

Пользователь должен сам подготовить свои файлы в каком-либо текстовом редакторе в соответствии с рекомендациями, изложенными ниже.

#### **2.2.1.1. Формат файлов**

1. Основной файл с данными биочип-эксперимента является текстовым ASCII файлом с табуляторами в качестве разделителей. Основной файл с данными содержит таблицы “профили экспрессии генов (строки) – образцы (колонки)”. Первые несколько столбцов основного файла с данными содержат идентификаторы проб, соответствующих генам, и текстовую вспомогательную информацию о пробах, следующие – числовые данные для образцов. Первая строка содержит заголовки для столбцов – названия типа информации для текстовых столбцов и идентификаторы образцов для числовых столбцов (Рис. 15). Файлу присваивается любое название, поскольку после загрузки его в пользовательский проект он будет переименован по названию проекта.

ID_REF	IDENTIFIER	TYPE	GSM30698	GSM30699
1985786CB1_PROBE1	NM_020357	GCand	1.40618	2.51101
480951.23_PROBE1	NM_016061	GCand	3.46037	6.1926
NM_000700.1_PROBE1	NM_ANXA1	GCand	4.11427	3.7463
NM_001530.1_PROBE1	NM_001530	GCand	1.14972	1.44073
NM_005406.1_PROBE1	NM_005406	GCand	0.561349	0.790164
NM_005642.2_PROBE1	NM_005642	GCand	2.08548	3.22073
NM_005767.1_PROBE1	NM_005767	GCand	0.523997	0.383103
X68560_PROBE1	X68560	GCand	0.540588	1.01563
000106CB1_PROBE1	NM_002489	~	0.768772	1.21582

**Рисунок 15.**

В колонку «TYPE» пользователь может вставить для генов любой код, описывающий биологические характеристики гена, например, что этот ген относится к группе генов «домашнего хозяйства», или этот ген интересен как ген-кандидат, выявленный в ряде экспериментов, и т.д. Если пользователь ввел какие-либо свои коды в эту колонку, ему необходимо провести сортировку всей таблицы по этой колонке, чтобы сгруппировать вместе строки, содержащие одноименные коды. Используя эти коды биологической значимости генов, пользователь может проследивать ассоциацию интересующих его генов с разными группами или кластерами генов, получающимися в результате выполнения алгоритмов подсистемы.

Данные основного файла должны быть не логарифмированными и могут представлять собой как значения сигналов одноцветной гибридизации, т.е. абсолютные значения экспрессии генов, так и значения ОТНОШЕНИЙ сигналов двухцветной гибридизации, т.е. относительные значения экспрессии генов. Если у пользователя есть файл с данными в виде логарифмированных значений, ему необходимо будет прологарифмировать их и в таком виде загрузить в подсистему.

2. Вспомогательный файл «Файл группировки» является текстовым ASCII файлом с табуляторами в качестве разделителей. Первый столбец содержит идентификаторы образцов, второй - текстовую информацию о принадлежности образцов к той или иной экспериментальной группе (код группы). Первая и вторая строка содержат информацию о количестве строк с идентификаторами образцов, третья и четвертая – о количестве столбцов с данными о группировке, т.е. 1. Пятая строка содержит заголовок для столбцов с названиями типа информации (Рис. 16). Файл может иметь любое название, поскольку после загрузки его в пользовательский проект он будет переименован по названию проекта с добавлением суффикса «Symptom».

```

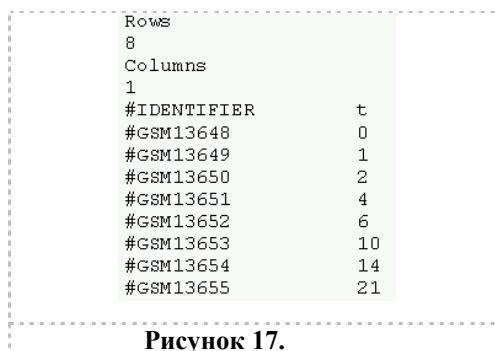
Rows
31
TextCols
1
ID_REF    GROUP
GSM30698 normal
GSM30699 normal
GSM30700 normal
GSM30701 normal

```

**Рисунок 16.**

Образцами могут быть, например, пробы разных тканей или органов, взятые у разных людей, животных, или клеточные культуры, или все эти категории после некоторых экспериментальных воздействий, на разных стадиях развития и т.д.

Если пользователь желает загрузить данные биочип-эксперимента, относящегося к типам «временная серия» (time series или time course) или «дозовый эффект» (dose effect), в которых возможно выявить индуцибельные и циклические гены, он должен подготовить еще один вспомогательный файл – файл «Временной шкалы». Этот файл является текстовым ASCII файлом с табуляторами в качестве разделителей. Первый столбец содержит идентификаторы образцов, второй – числовые данные со значениями точек временного отсчета. Первая и вторая строка содержат информацию о количестве строк с идентификаторами образцов, третья и четвертая – о количестве столбцов со значениями точек временного отсчета, т.е. 1. Пятая строка содержит заголовки для столбцов (Рис. 17). Файл может иметь любое название, поскольку после загрузки его в пользовательский проект он будет переименован по названию проекта с добавлением суффикса «Scale».



```
Rows
8
Columns
1
#IDENTIFIER      t
#GSM13648        0
#GSM13649        1
#GSM13650        2
#GSM13651        4
#GSM13652        6
#GSM13653       10
#GSM13654       14
#GSM13655       21
```

**Рисунок 17.**

#### 2.2.1.2. Описание работы системы ввода данных

Система ввода данных начинается после заведения в профиле пользователя проекта (см. пункт 2.1.).

Набрав свой идентификатор и пароль, зарегистрированные для пользователя системы «Системная биология», и ознакомившись с возможностями подсистемы «Компьютерная Транскриптомика», пользователь нажимает на кнопку «Далее» и перемещается на страницу заведения личного проекта (см. пункт 2.1., Рис. 2).

После введения название своего проекта, следуя указаниям слева, и нажатия на кнопку «Далее» страница обновляется с появлением строки с введенным названием проекта (пункт 2.1., Рис. 3). Нажатие на кнопку «Далее» перемещает пользователя на страницу выбора задачи (Рис. 18), на которой после нажатия на знак «+» в строке с задачей «Подготовка данных» появляется строка с предложением перейти на страницу «Загрузка файлов биочип-исследования на сервер» (Рис. 19).

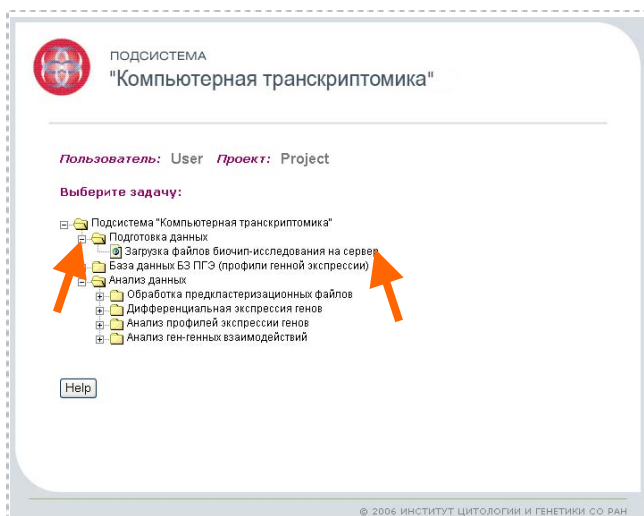


Рисунок 18.

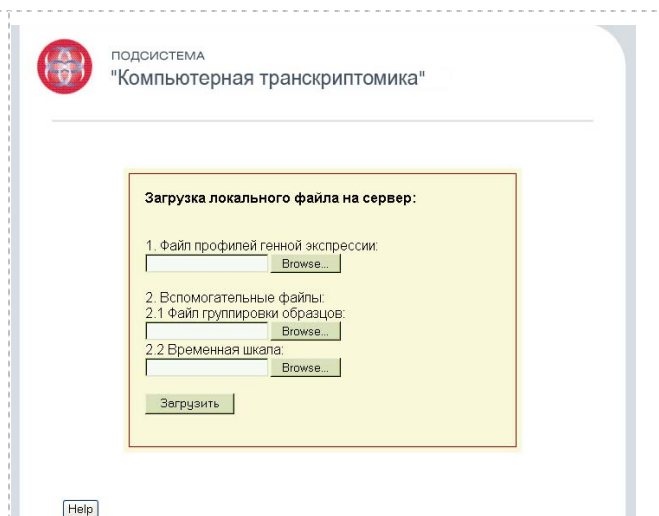


Рисунок 19.

На этой странице предоставляются средства файл-менеджера для загрузки своих файлов с данными и описанием биочип-исследования:

- нажатие на кнопку «Browse» после первого окошка с заголовком «1. Файл профилей генной экспрессии:» позволяет указать путь к основному файлу с данными (см. пункт 2.2.1.1.);
- нажатие на кнопку «Browse» после второго окошка с заголовком «2. Вспомогательные файлы: 2.1 Файл группировки образцов:» позволяет указать путь к вспомогательному файлу «Файл группировки» (см. пункт 2.2.1.1.);
- если пользователь собирается анализировать данные биочип-эксперимента, относящегося к типам «временная серия» (time series или time course) или «дозовый эффект» (dose effect), в которых возможно выявить индуцибельные и циклические гены, он должен нажать на кнопку «Browse» после третьего окошка с заголовком «2.2 Временная шкала:», чтобы указать путь к вспомогательному файлу «Временной шкалы» (см. пункт 2.2.1.1.).

После указания пути к двум обязательным файлам и файлу для анализа временных серий пользователь должен дать команду на их загрузку нажатием на кнопку «Загрузить» (Рис. 20).

После этого пользователь перемещается на страницу извещения, что необходимые файлы успешно загружены, что определяется, в частности, по соответствию названия файлов названию заведенного пользовательского проекта и сохранению размеров заведенных файлов размерам исходных (Рис. 21.).

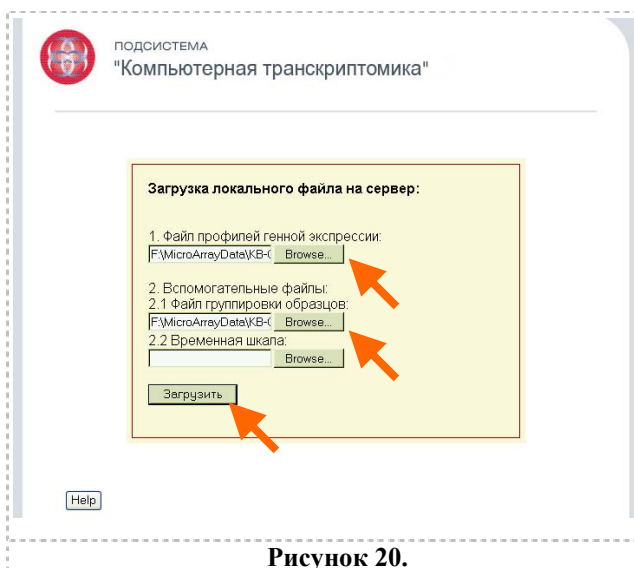


Рисунок 20.

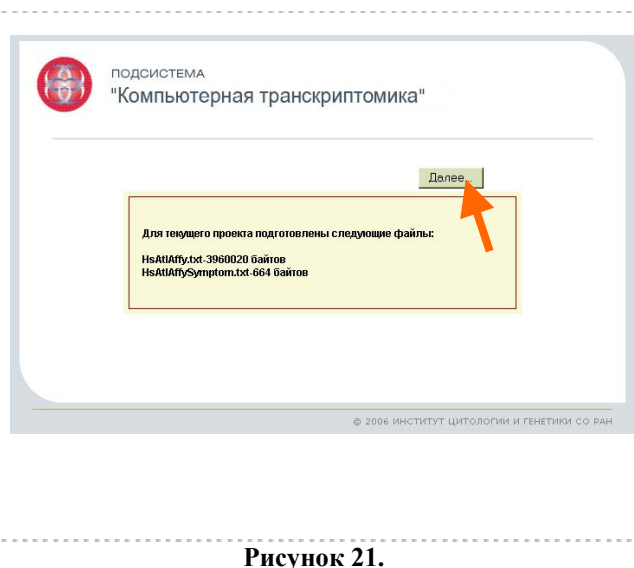


Рисунок 21.

Если пользователь по ошибке повторно произведет загрузку своих файлов, то сервер выдаст сообщение о том, что файлы уже были загружены (Рис. 22). Если пользователь попытается запустить расчетный сценарий, не загрузив файлы с данными, то сервер выдаст сообщение о том, что файлы не были загружены

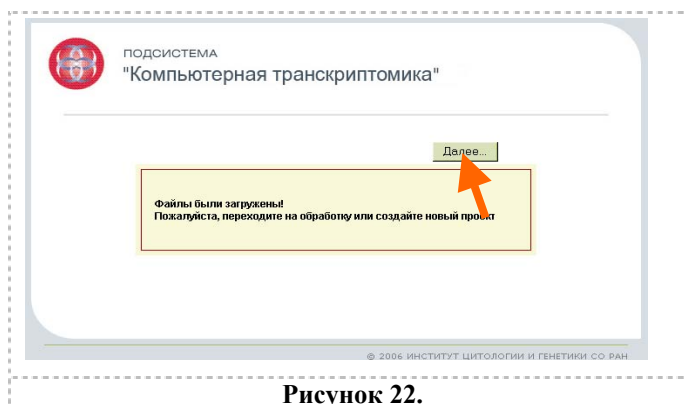


Рисунок 22.

Все наименования файлов внутри проекта генерируются так, чтобы они имели общий «корень», идентичный названию профиля, и различные суффиксы, отражающие сценарии расчетов, в которых эти файлы были сгенерированы. Это защищает файлы от произвольного перезаписывания и помогает пользователю ориентироваться среди множества открытых окон с файлами результатов работы нескольких сценариев.

Нажав на кнопку «Далее» пользователь перемещается на страницу выбора задачи для анализа (Рис. 23).

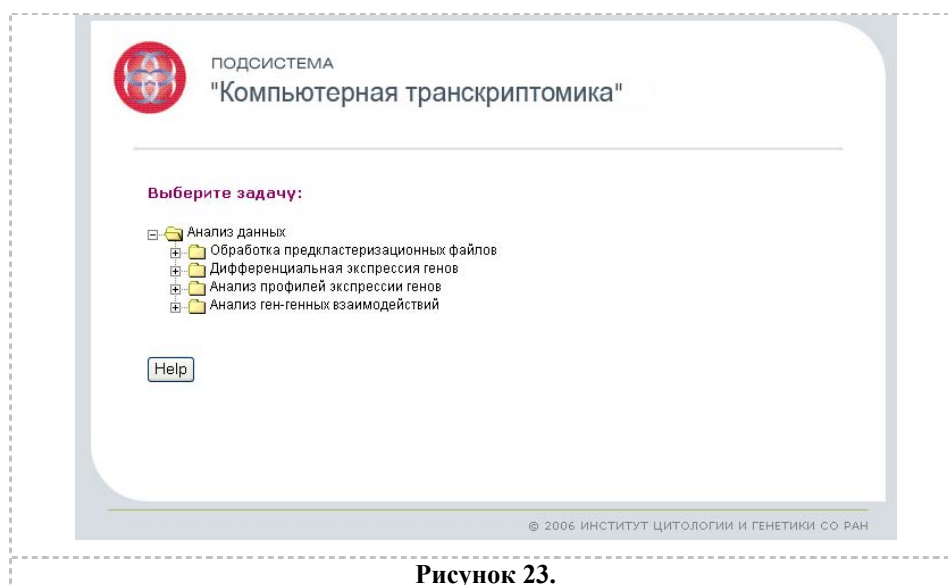


Рисунок 23.

Программные компоненты подсистемы «Компьютерная Транскриптомика» в заведенном пользовательском проекте готовы к выполнению анализа загруженных данных.

В подсистеме «Компьютерная Транскриптомика» пользователю предоставляются возможность:

- провести обработку предкластеризационных файлов с помощью программных компонент для предварительной математической обработки, фильтрации и статистического анализа биочиповых экспрессионных данных;
- провести выявление и статистическую оценку различий в экспрессии генов с помощью нескольких алгоритмов, реализованных программными компонентами для выявления дифференциально экспрессирующихся генов;
- провести кластеризацию или многомерный анализ биочиповых данных с помощью нескольких алгоритмов, реализованных программными компонентами для анализа профилей экспрессии генов;
- провести выявление и статистическую оценку корреляций между экспрессией генов с помощью нескольких алгоритмов, реализованных программными компонентами для анализа ген-генных взаимодействий.

### 2.2.2. Программные компоненты для предварительной математической обработки, фильтрации и статистического анализа биочиповых экспрессионных данных

Необходимой частью всех систем по анализу данных биочиповых экспериментов является предварительная обработка данных квантификации сигналов гибридизации. Предварительная обработка состоит в аналитических или преобразовательных процедурах, необходимых, чтобы данные стали пригодны для детального и продуктивного анализа. Как правило, эти процедуры включают логарифмическое преобразование числовых данных, фильтрацию малоинформативных данных, фильтрацию профилей с пропущенными данными, а также централизацию-стандартизацию-нормализацию таблиц с данными. Последние процедуры, хотя они имеют определенные отличия друг от друга, в целом предназначены для устранения систематических ошибок в данных, что необходимо для того, чтобы можно было корректно сравнивать между собой данные по результатам отдельных гибридизаций. В области анализа биочиповых данных эти преобразования данных принято обозначать одним термином – нормализация (Passanen et al., 2003).

Настоящий подраздел подсистемы представляет собой комплекс программных компонент для математической обработки биочиповых экспрессионных данных, вычисления статистических параметров массива данных и визуального контроля качества массива данных. После процедур логарифмического преобразования числовых данных, фильтрации малоинформативных данных, фильтрации профилей с пропущенными данными, а также централизацию и стандартизации таблиц с данными вычисляются общепринятые статистические характеристики: среднее  $\bar{x} = \sum x_i / N$ , дисперсия  $s^2 = \sum (x_i - \bar{x})^2 / N$ , стандартное отклонение  $s = \sqrt{s^2}$ , коэффициент вариации  $cv = s/\bar{x}$  (Айвазян и др., 1987; Животовский, 1991; Гайдышев, 2001). Вычисления проводятся отдельно по строкам и столбцам.

Результаты вышеназванных преобразований данных можно наглядно представить с помощью графика статистических параметров профилей образцов по генам, так называемого «box plot». На этом графике для результатов гибридизации каждого образца показаны положения для среднего, для одного и двух стандартных отклонений, для максимального и минимального значений.

### 2.2.2.1. Описание работы сценария «Вычисление статпараметров: среднего, дисперсии, отклонения, вариации»

После нажатия на знак «+» в строке с названием «Обработка предкластеризационных файлов» открывается ссылка «Вычисление статпараметров: среднего, дисперсии, отклонения, вариации» (Рис. 24), нажатие на которую приводит к открытию страницы для выполнения сценария с этой задачей (Рис. 25).

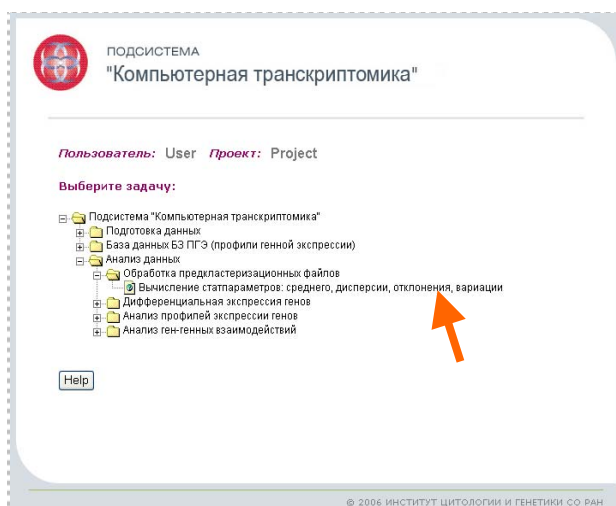


Рисунок 24.

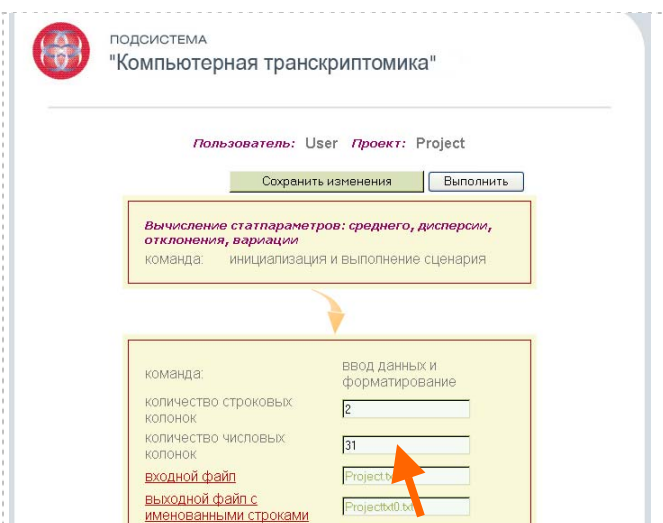


Рисунок 25.

Пользователь должен проверить соответствие значения для числа образцов, присутствующее в окошке «количество числовых колонок», значению, указанному или в вспомогательном файле «Файл группировки» (см. пункт 2.2.1.1.), если файлы были подготовлены и загружены пользователем со своего компьютера, или в документе описания биочип-исследования, если файлы были загружены пользователем из БЗ-ПГЭ. Если числа совпадают, то можно нажать на кнопку «Выполнить» для реализации сценария расчетов (Рис. 26).

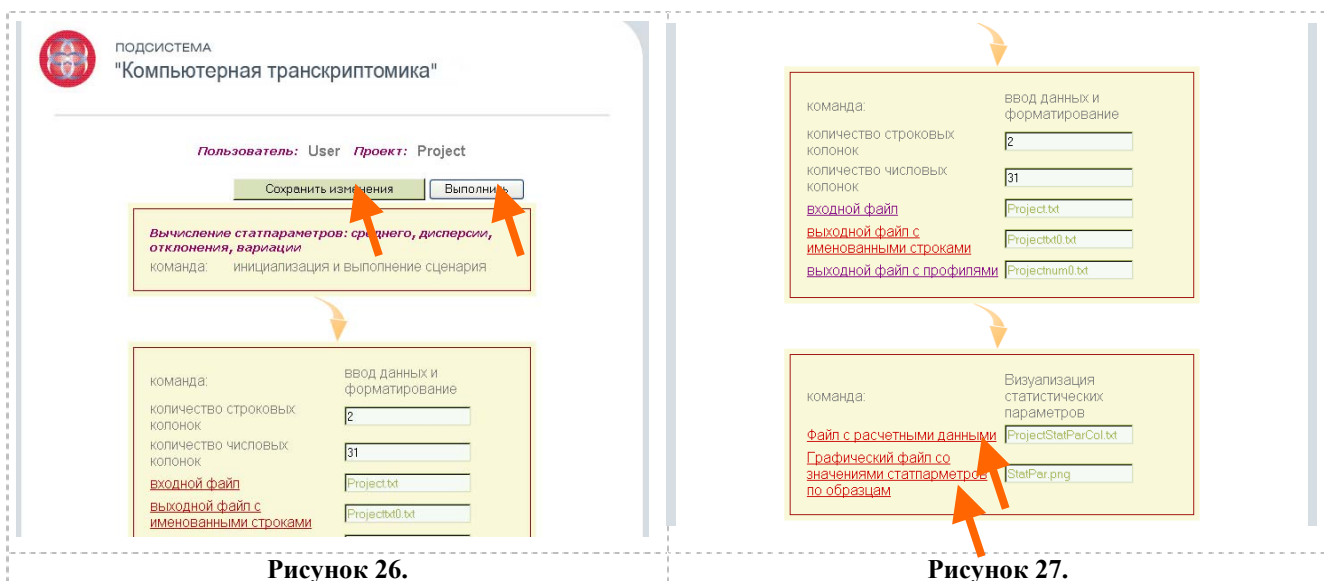


Рисунок 26.

Рисунок 27.

В случае, если количество образцов во входном файле с данными не соответствует параметру «количество числовых колонок» или значение «количество строчковых колонок» не соответствует структуре входного файла с данными, а пользователь не обратил на это внимания, при попытке выполнить сценарий расчетов сервер выдает предупреждение (Рис. 2.2.2.1.5).

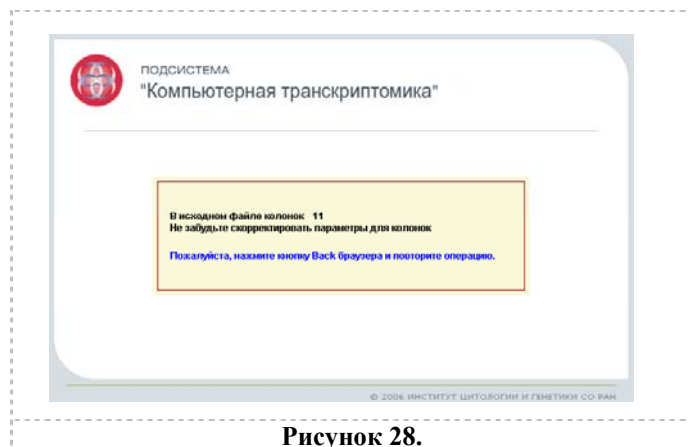


Рисунок 28.

В случае редактирования окошек с параметрами «количество числовых колонок» или «количество строчковых колонок» необходимо затем нажать на кнопку «Сохранить изменения». После того, как сервер выполнит эту команду, о чем можно проследить по окошку состояния и сообщениям в интернет-браузере, можно нажать кнопку «Выполнить» для реализации сценария расчетов (Рис. 26). Страница с сценарием обновляется, и теперь внизу страницы можно посмотреть результаты расчетов, нажав на ссылку «Файл с расчетными данными» (с суффиксом «\*StatParCol.txt») или на ссылку «Графический файл со значениями статпараметров по образцам» («\*StatPar.png») (Рис. 27).

В первом случае будет выведена страница с результатами расчетов в виде табличных числовых данных в окне, позволяющем прокрутку для просмотра (Рис. 29).

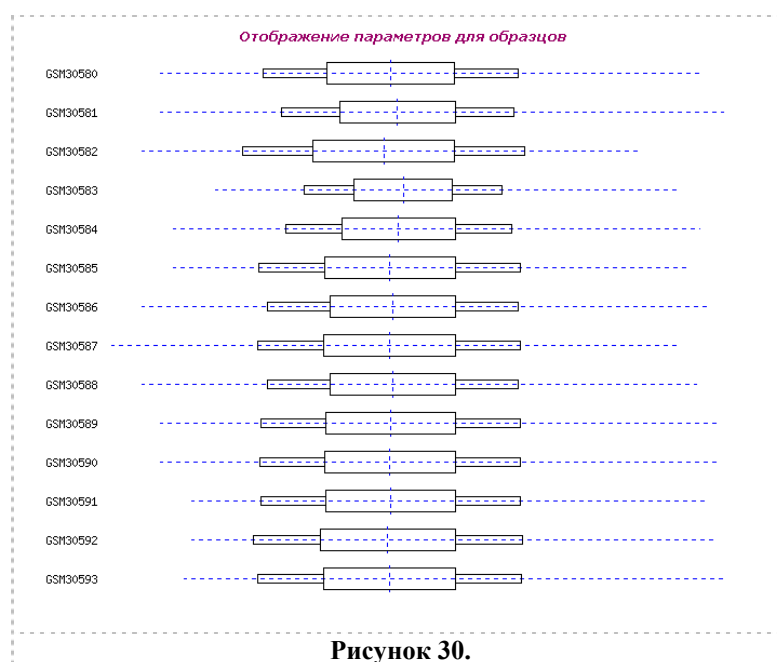


**ProjectStatParCol.txt**

ID_REF	Минимум	Максимум	Среднее	Стандартное отклонение	Дисперсия	Кoeffициент_вариации
GSM30580	-1.73697	15.5863	5.6165	2.04351	4.17594	36.3841
GSM30581	-1.73697	16.2571	5.83294	1.85882	3.45522	31.8677
GSM30582	-2.32193	13.5677	5.39103	2.25249	5.07369	41.7821
GSM30583	0	14.7552	6.01665	1.58479	2.51155	26.34
GSM30584	-1.32193	15.5181	5.87099	1.80821	3.26963	30.7991
GSM30585	-1.32193	15.1551	5.59372	2.09289	4.38019	37.415
GSM30586	-2.32193	15.7823	5.68684	2.00078	4.00312	35.1826
GSM30587	-3.32193	14.8533	5.56695	2.10632	4.43659	37.8362
GSM30588	-2.32193	15.4002	5.67719	2.00923	4.03702	35.3914
GSM30589	-1.73697	16.1103	5.61003	2.07594	4.30953	37.0041
GSM30590	-1.73697	16.0799	5.59641	2.09035	4.36957	37.3516
GSM30591	-0.736966	15.7543	5.61636	2.07768	4.31674	36.9933
GSM30592	-0.736966	15.9555	5.52335	2.15342	4.63724	38.9876

**Рисунок 29.**

Во втором – в виде графика, построенного на основании рассчитанных величин, в форме «box plot» (Рис. 30). На этом графике для результатов гибридизации каждого образца показаны положения для среднего, для одного и двух стандартных отклонений, для максимального и минимального значений.



Если пользователь попытается просмотреть файлы с результатами не запустив перед этим сценарий расчетов, то сервер будет выдавать предупреждение (Рис. 31).

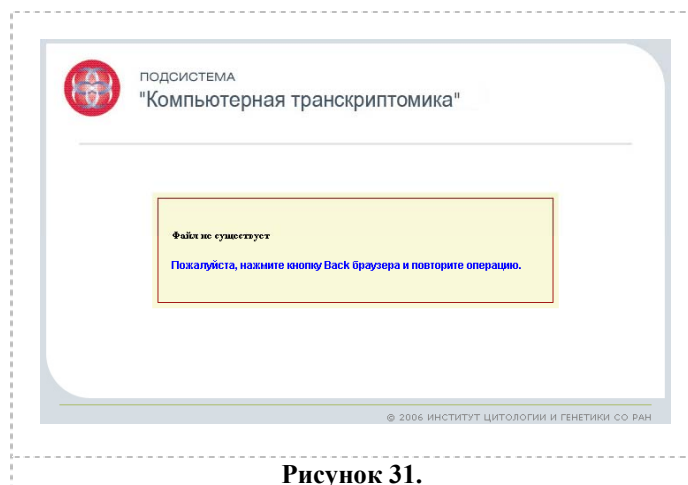


Рисунок 31.

### 2.2.3. Выявление дифференциально экспрессирующихся генов

Этот раздел подсистемы представляет собой комплекс программных компонент сравнительного статистического анализа для выявления и валидации статистически значимых различий в количественных оценках экспрессии генов.

Предложено много методов для выявления дифференциально экспрессирующихся генов (Nadon, Shoemaker, 2002). Один из самых простых является метод отбора по критерию кратной разницы (FC (fold change) cutoff), т.е. отбор таких генных профилей, для которых разница в значениях экспрессии генов между двумя образцами или между значениями сигналов при двухцветной гибридизации для одного образца превышает определенный порог. Обычно используют значения порога от 1.8 до 3.0 (Mutch et al., 2002). Применение критерия двукратной разницы (two-fold criterium), являющегося наиболее часто используемым частным случаем критерия кратной разницы, позволяет предварительно наметить дифференциальную разницу в транскрипционной активности генов даже исходя из данных, полученных в результате одной двухцветной гибридизации, т.е. в случае, если не было использовано ни технических, ни биологических реплик в эксперименте. Применение этого критерия реализовано в сценарии «Расчет критерия кратной разницы».

Однако этот критерий не способен учитывать различия в значениях фона, (неравномерность фона), нестабильность и изменчивость измерений интенсивности сигналов и много других технических факторов, присущих биочиповой технологии. Поэтому часто дает много ложно-положительных результатов (Claverie, 1999; Mutch et al., 2002). Поэтому необходимо использовать статистические методы для выявления дифференциальной разницы в экспрессии генов и оценки значимости выявленных различий (Nadon, Shoemaker, 2002). Например, очень часто используют t-тесты, в частности критерий Уэлша (Крамер, 1975; Quackenbush, 1999; Гайдышев, 2001). В отличие от критерия Стьюдента критерий Уэлша не требует предположения о равенстве неизвестных дисперсий, в нем фактически проверяется гипотеза об отклонении нормально распределенной разницы средних от нуля. Если нужно найти профили, значения которых на этих группах максимально различаются, тогда для каждого профиля вычисляется критерий Уэлша – по формуле:

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

Таблица критических значений для различных уровней значимости одна и та же для обоих критериев (Крамер, 1975; Гайдышев, 2001). Профили сортируются по значению t.

(Необходимо проверять оба конца получаемого списка, так как на одном конце собираются профили, для которых  $\bar{x}_1 > \bar{x}_2$ , на другом – наоборот). Во вспомогательном файле «Файл группировки» для каждого образца задан код группы, в которую он входит. Таким образом, находятся профили, значения которых на этих группах максимально различаются. Это реализовано в сценарии «Расчет t-критерия Уэлша».

Частным случаем выявления дифференциально экспрессирующихся генов является выявление генов, не проявляющих дифференциальную экспрессию, т.е. генов с конститутивной экспрессией. Выявление таких генов позволяет наметить, какие гены могут быть использованы в качестве стандартов или нормализаторов в экспериментах по валидации результатов биочиповых-исследований независимыми методами, например, реальновременной ОТ-ПЦР. Для того, чтобы выявить такие гены, для исходного массива данных (минуя процедуры центрирования и нормирования) вычисляется коэффициент вариации, проводится фильтрация по определенному его значению и отбираются гены с коэффициентом ниже порога (сценарий «Выявление конститутивных профилей экспрессии генов»).

Если проводится анализ данных временных серий из биочиповых экспериментов (дифференцировка клеток, реакция клеток на воздействие) или дозовых эффектов, то можно выявить индуцибельные гены, т.е. экспрессия которых достоверно меняется в зависимости от времени или дозы воздействия. Для того, чтобы выявить такие гены, для каждого профиля вычисляется отношение максимального перепада значений к стандартному отклонению профиля (сценарий «Выявление индуцибельных профилей экспрессии генов»).

При анализе данных временных серий можно также выявить гены, проявляющие циклическую экспрессию (Whitfield et al., 2002; Chen, 2005). Задача состоит в том, чтобы для каждого профиля найти максимально выраженную частоту колебаний и оценить степень ее проявления. Алгоритм изложен в (Wichert et al., 2004). В качестве параметра вводится шаг  $\tau$  – значение дискретизации временной оси. Период  $T$  вида  $\tau \cdot 2^n$  выбирается таким, чтобы не превысить диапазон временных отсчетов  $t_N - t_0$ . Для Фурье-частот с периодами  $T, T/2, T/3 \dots$  вычисляется спектральная плотность. Для каждого профиля в выходной файл выводятся g-статистика – доля спектра на максимально выраженной частоте, период, соответствующий этой частоте, и спектральные плотности всех Фурье-частот. (сценарий «Выявление циклических профилей экспрессии генов»).

### *2.2.3.1. Описание работы сценария «Расчет критерия кратной разницы»*

Сценарий «Расчет критерия кратной разницы» применяется только для данных, представляющих собой отношение значений интенсивности. Обычно такие данные получаются при двухцветной гибридизации. Для того чтобы провести расчеты по сортировке исходной таблицы с данными биочип-эксперимента в соответствии с критерием кратной разницы, пользователь нажимает на знак «+» в строке с названием «Дифференциальная экспрессия генов» (Рис. 32), после чего на появившуюся ссылку «Расчет критерия кратной разницы», в результате чего появляется соответствующая страница (Рис. 33).

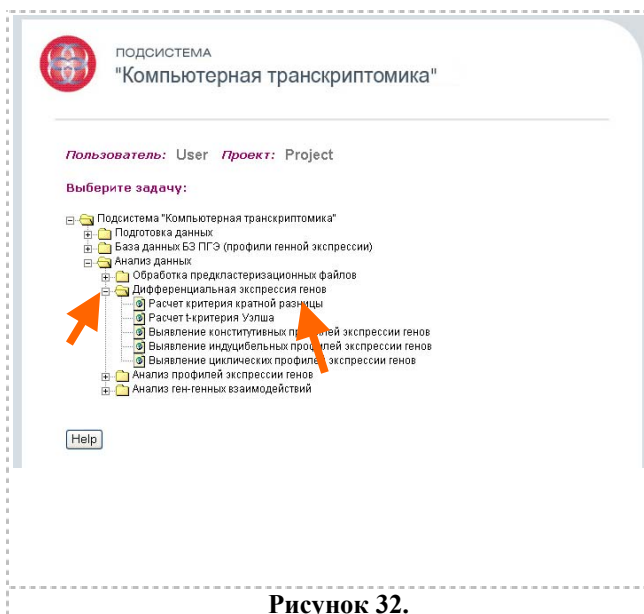


Рисунок 32.

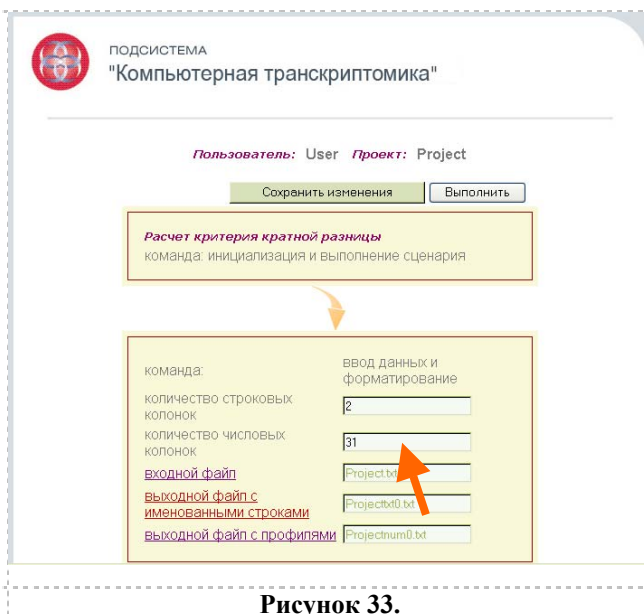


Рисунок 33.

Пользователь должен проверить соответствие значения для числа образцов, присутствующее в окошке «количество числовых колонок», значению, указанному или во вспомогательном файле «Файл группировки» (см. пункт 2.2.1.1.), если файлы были подготовлены и загружены пользователем со своего компьютера, или в документе описания биочип-исследования, если файлы были загружены пользователем из БЗ-ПГЭ.

Если числа совпадают, то можно нажать на кнопку «Выполнить» для реализации сценария расчетов (Рис. 34). В случае редактирования этого окошка необходимо затем нажать на кнопку «Сохранить изменения» (Рис. 34).

После того, как сервер выполнит эту команду, о чем можно проследить по окошку состояния и сообщениям в интернет-браузере, можно нажать кнопку «Выполнить» для реализации сценария расчетов (Рис. 34). Страница с сценарием обновляется, и теперь внизу страницы можно просмотреть результаты расчетов, нажав на ссылки «выходной числовой файл(выше порога)», или «выходной текстовый файл(выше порога)», или «выходной числовой файл(ниже порога)», или «выходной текстовый файл(ниже порога)» (Рис. 35).

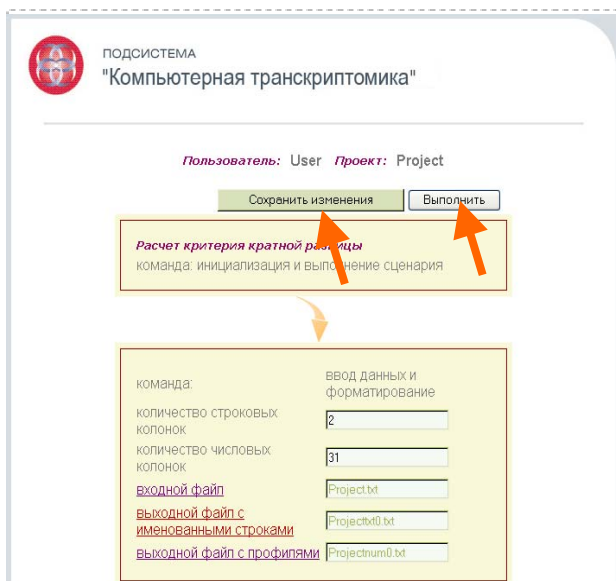


Рисунок 34.

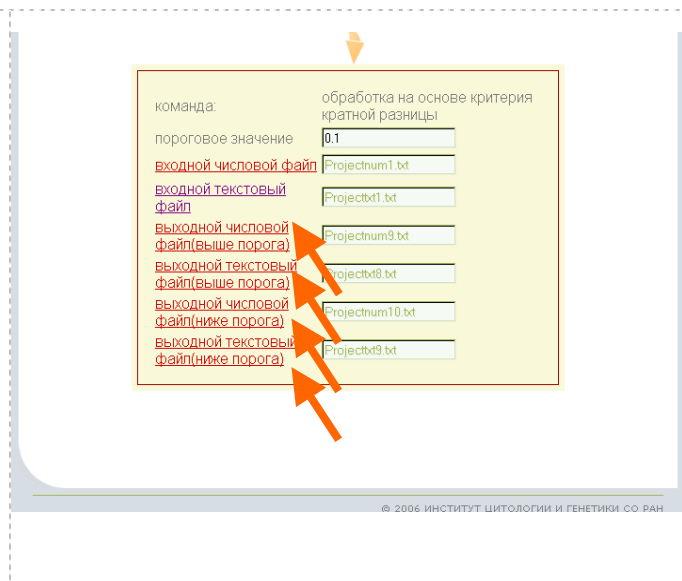


Рисунок 35.

Файлы числовые содержат числовые результаты для пересортированных профилей (Рис. 36), а текстовые – пересортированные колонки с идентификаторами этих профилей (Рис. 37).

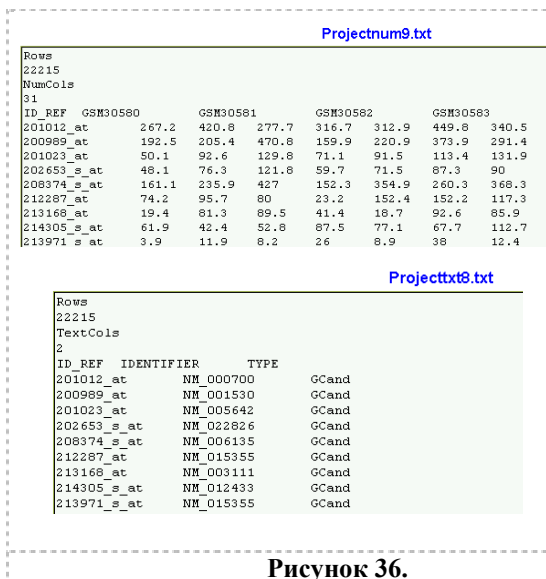


Рисунок 36.

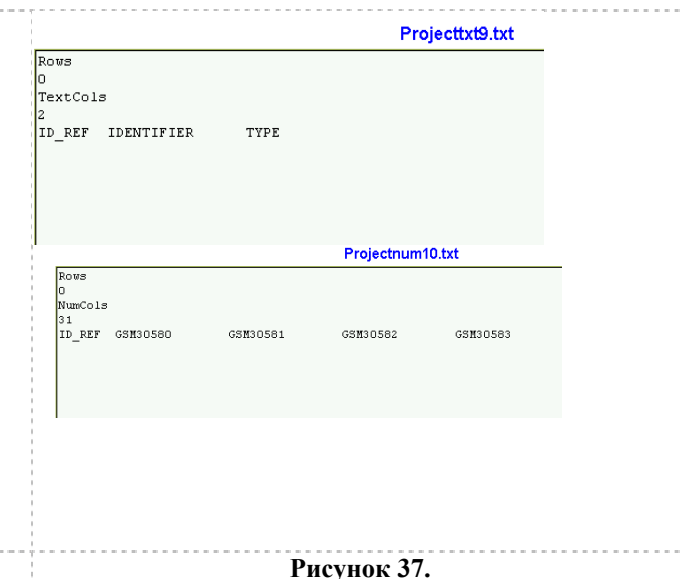


Рисунок 37.

При выполнении расчетов по этому сценарию пользователь может сам установить значение порога для разделения массива профилей (Рис. 38). Параметр «порог» в данном случае является значением отношения интенсивности сигналов (для двухканальной гибридизации), разделяющим массив профилей на две части: выше порога (высокие значения отношения, т.е. гены, проявляющие разницу в экспрессии) и ниже порога (низкие значения отношения, т.е. гены, не проявляющие разницу в экспрессии). По умолчанию порог равен 0.1. После введения нового значения необходимо нажать на кнопку «Сохранить изменения» (Рис. 39).



Рисунок 38.

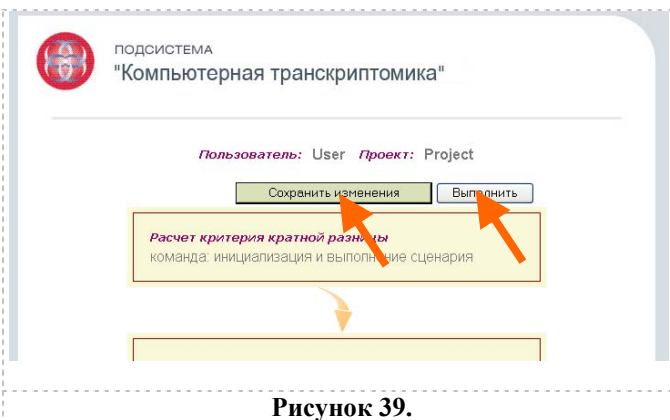


Рисунок 39.

После того, как сервер выполнит эту команду, о чем можно проследить по окошку состояния и сообщениям в интернет-браузере, можно нажать кнопку «Выполнить» для реализации сценария расчетов (Рис. 39). Страница с сценарием обновляется, и теперь внизу страницы можно просмотреть новые результаты расчетов, нажав на ссылки «выходной числовой файл(выше порога)», или «выходной текстовый файл(выше порога)», или «выходной числовой файл(ниже порога)», или «выходной текстовый файл(ниже порога)».

### 2.2.3.2. Описание работы сценария «Расчет t-критерия Уэлша»

Для того чтобы провести расчеты по сортировке исходной таблицы с данными биочип-эксперимента в соответствии с результатами расчета t-критерия Уэлша, пользователь нажимает на знак «+» в строке с названием «Дифференциальная экспрессия генов», после чего на появившуюся ссылку «Расчет t-критерия Уэлша» (Рис. 40), в результате чего появляется соответствующая страница (Рис. 41).

Пользователь должен проверить соответствие значения для числа образцов, присутствующее в окошке «количество числовых колонок», значению, указанному или в вспомогательном файле «Файл группировки» (см. пункт 2.2.1.1.), если файлы были подготовлены и загружены пользователем со своего компьютера, или в документе описания биочип-исследования, если файлы были загружены пользователем из БЗ-ПГЭ.

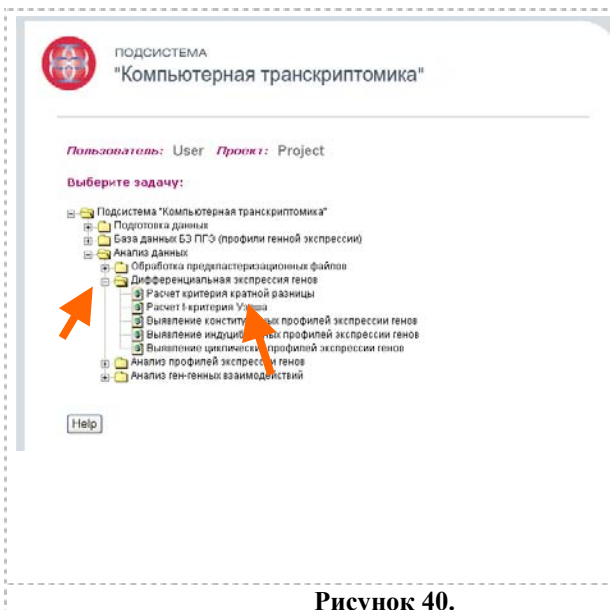


Рисунок 40.

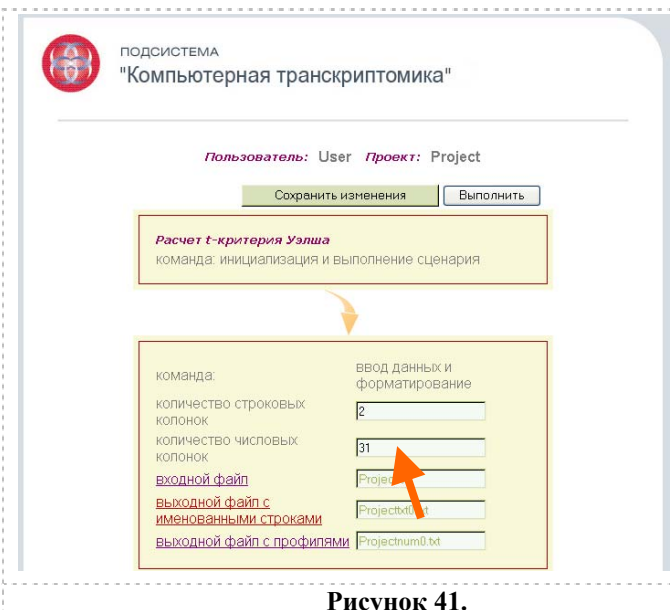
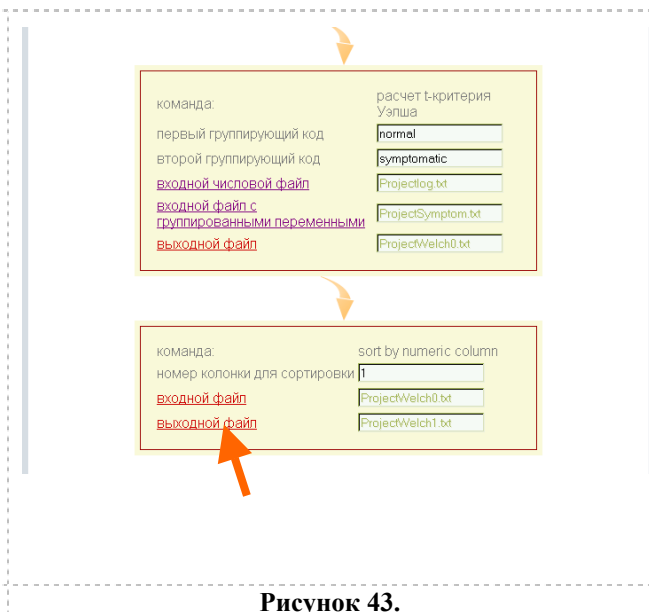
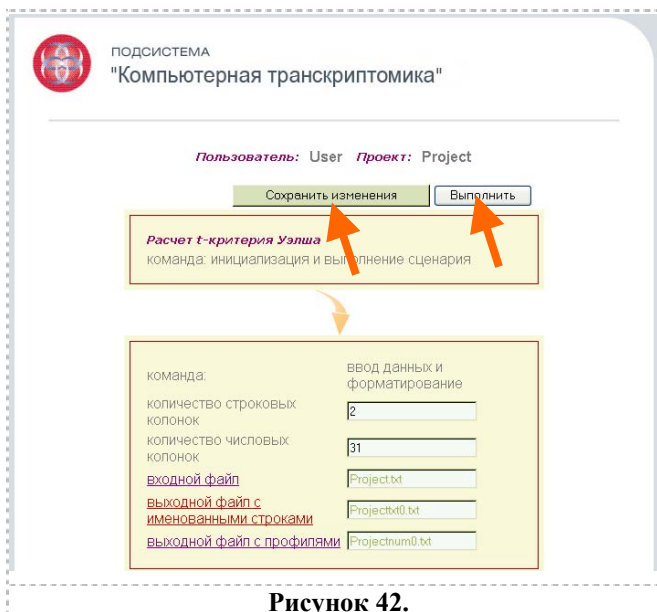


Рисунок 41.

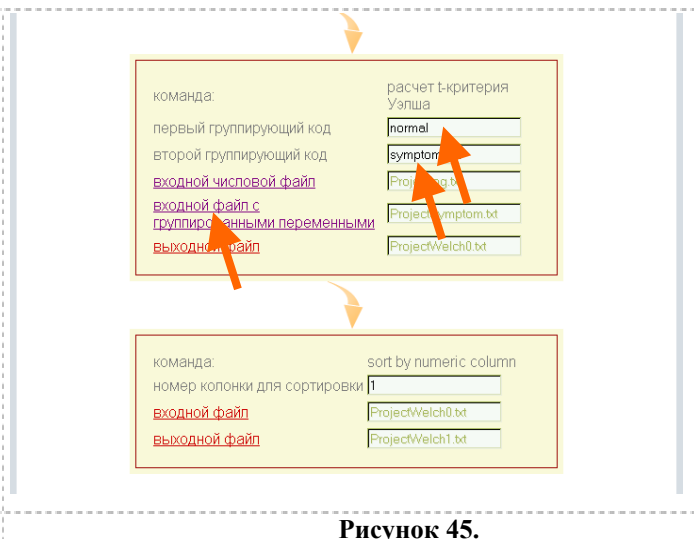
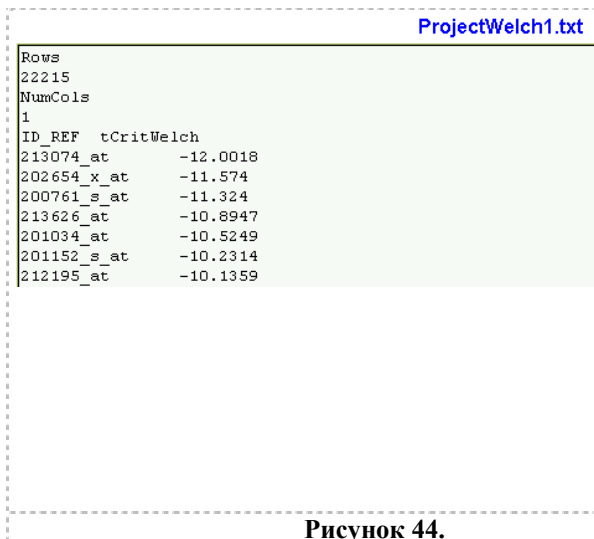
Если числа совпадают, то можно нажать на кнопку «Выполнить» для реализации сценария расчетов (Рис. 42).

В случае редактирования этого окошка необходимо затем нажать на кнопку «Сохранить изменения» (Рис. 42). После того, как сервер выполнит эту команду, о чем можно проследить по окошку состояния и сообщениям в интернет-браузере, можно нажать кнопку «Выполнить» для реализации сценария расчетов (Рис. 42). Страница со сценарием обновляется, и теперь внизу страницы можно просмотреть результаты расчетов, нажав на ссылку «выходной файл» (с суффиксом «\*Welch1.txt») (Рис. 43).



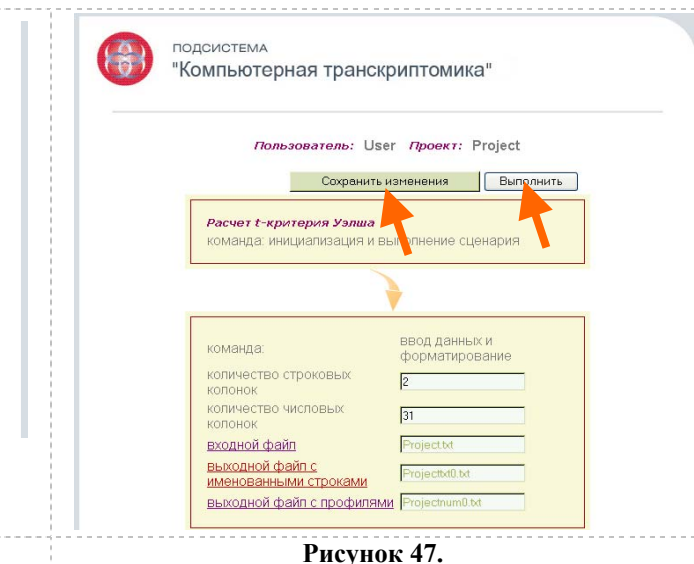
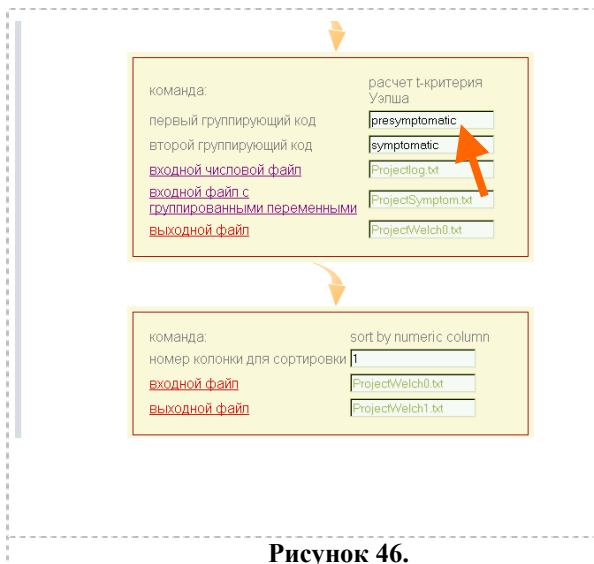
Файл с результатом представляет собой список генов, отсортированных в соответствии с рассчитанными значениями критерия для указанных групп образцов (Рис. 44).

Пользователь может изменить коды групп для выведения файлов с расчетами различий между другими группами образцов. Для этого он редактирует окошки «первый группирующий код» и/или «второй группирующий код» (Рис. 45). Чтобы быть уверенным в точном написании кодов групп, пользователь может в отдельном окне браузера просмотреть входной «Файл группировки» (см. раздел 2.2.1.1.), скопировать оттуда коды и вставить их в соответствующие окошки на странице сценария.



В случае изменения содержания окошек для группирующих кодов (Рис. 46) необходимо затем нажать на кнопку «Сохранить изменения» (Рис. 47).

После того, как сервер выполнит эту команду, о чем можно проследить по окошку состояния и сообщениям в интернет-браузере, можно нажать кнопку «Выполнить» для реализации сценария расчетов (Рис. 47).



Страница со сценарием обновляется, и теперь внизу страницы можно просмотреть новые результаты расчетов, нажав на ссылку «выходной файл» (с суффиксом «\*Welch1.txt»). Файл с новым результатом представляет собой список генов, отсортированных в соответствии с рассчитанными значениями критерия для указанных групп образцов.

### 2.2.3.3. Описание работы сценария «Выявление конститутивных профилей экспрессии генов»

Для того чтобы выявить конститутивные профили после фильтрации исходной таблицы с данными биочип-эксперимента в соответствии с результатами расчета



коэффициента вариации, пользователь нажимает на знак «+» в строке с названием «Дифференциальная экспрессия генов», после чего на появившуюся ссылку «Выявление конститутивных профилей экспрессии генов» (Рис. 48), в результате чего появляется соответствующая страница (Рис. 49).

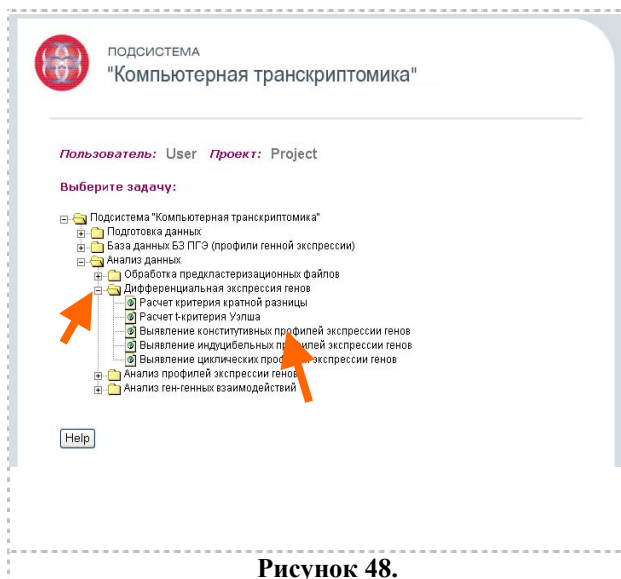


Рисунок 48.

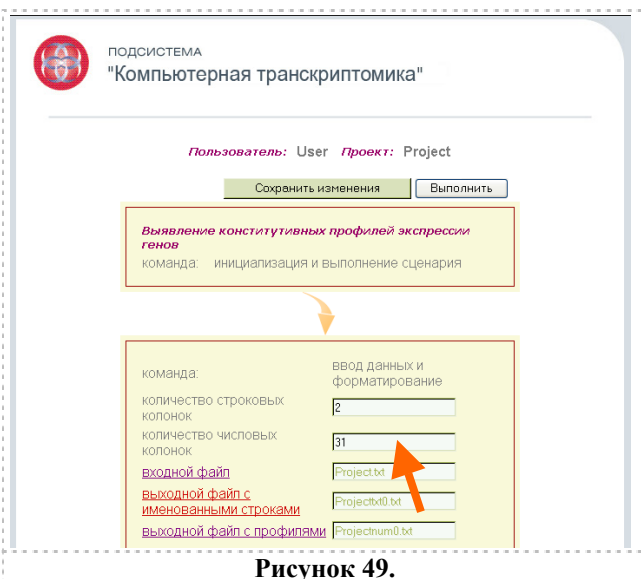


Рисунок 49.

Пользователь должен проверить соответствие значения для числа образцов, присутствующее в окошке «количество числовых колонок», значению, указанному или в вспомогательном файле «Файл группировки» (см. пункт 2.2.1.1.), если файлы были подготовлены и загружены пользователем со своего компьютера, или в документе описания биочип-исследования, если файлы были загружены пользователем из БЗ-ПГЭ.

Если числа совпадают, то можно нажать на кнопку «Выполнить» для реализации сценария расчетов (Рис. 50). В случае редактирования этого окошка необходимо затем нажать на кнопку «Сохранить изменения» (Рис. 50).

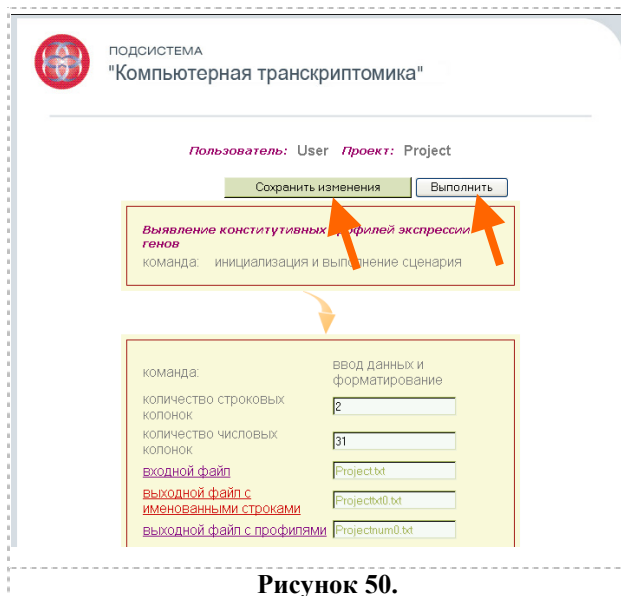


Рисунок 50.

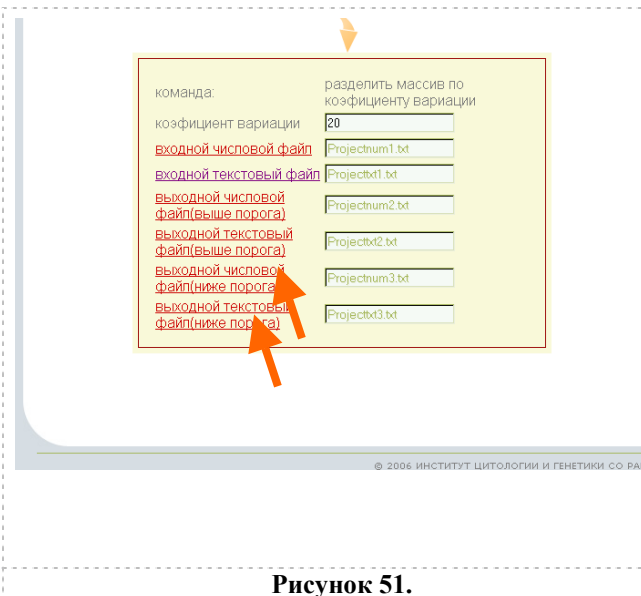


Рисунок 51.

После того, как сервер выполнит эту команду, о чем можно проследить по окошку состояния и сообщениям в интернет-браузере, можно нажать кнопку «Выполнить» для реализации сценария расчетов (Рис. 50). Страница со сценарием обновляется, и теперь внизу страницы можно просмотреть результаты расчетов, нажав на ссылки «выходной числовой файл(ниже порога)» и «выходной текстовый файл(ниже порога)» (Рис. 51).

**Projectnum3.txt**

```

Rows
343
NumCols
31
ID_REF GSM30580      GSM30581      GSM30582      GSM30583
79005_at      181      242      175.9      172.7      252.7      160      175.7
71933_at      575.4      447.9      503.4      393.7      509.9      516      494
65591_at      159.4      158.4      168.9      198      142.6      222.6      214.6
64064_at      918.8      574.8      830.4      448.5      575.6      682.9      658.2
60528_at      339.2      246.2      342.8      367      237.6      407.1      308.1
56197_at      1279.2      562.7      1102.3      722.4      906.8      1378.2      992.8
5214_
5177_
5027_
Projecttxt3.txt
Rows
493:
485:
TextCols
2
ID_REF IDENTIFIER TYPE
79005_at      NM_024881      ~
71933_at      NM_006522      ~
65591_at      NM_020839      ~
64064_at      NM_018384      ~
60528_at      NM_005090      ~
56197_at      NM_152766      ~
52169_at      NM_153335      ~
51774_s at      AW014299      ~

```

Рисунок 52.

**Projectnum3.txt**

команда: разделить массив по коэффициенту вариации

коэффициент вариации

входной числовой файл

входной текстовый файл

выходной числовой файл(выше порога)

выходной текстовый файл(выше порога)

выходной числовой файл(ниже порога)

выходной текстовый файл(ниже порога)

Рисунок 53.

Файлы с результатом выявления конститутивных профилей представляют собой списки числовых значений («числовой») и идентификаторов («текстовый») генов, отсортированных в соответствии с рассчитанными значениями коэффициента вариации и отфильтрованных по пороговому значению (Рис. 52). По умолчанию это значение равно 20.

Параметр, который может изменять пользователь – это коэффициент вариации, т.е. отношение среднего квадратичного отклонения к среднему в процентах. Если пользователь желает смягчить или ужесточить процедуру фильтрации, он может изменить значение коэффициента вариации выше или ниже, соответственно. Для этого он должен ввести новое значение параметра в окошко «коэффициент вариации» (Рис. 53).

подсистема "Компьютерная транскриптомика"

Пользователь: User Проект: Project

Сохранить изменения Выполнить

**Выявление конститутивных профилей экспрессии генов**

команда: инициализация и выполнение сценария

команда: ввод данных и форматирование

количество строковых колонок

количество числовых колонок

входной файл

выходной файл с именованными строками

выходной файл с профилями

Рисунок 54.

**Projectnum3.txt**

```

Rows
6848
NumCols
31
ID_REF GSM30580      GSM30581      GSM30582      GSM30583
91920_at      223.2      132.8      212.7      454.5      202.6      199.3      264.5
91816_f_at      84.3      52.2      64.9      70.9      59      49.8      60.4
91684_g_at      144.9      135.5      87.1      74.4      67.4      88.4      114.4
91682_at      260.7      216.2      155.5      252.5      226      219.8      258.3
90610_at      237      221.9      200.7      89.3      182.6      273.8      303.9
90265_at      520.6      622.9      485.7      383      453.8      455.2      571.7
8181_
7900_
Projecttxt3.txt
Rows
6848
TextCols
2
ID_REF IDENTIFIER TYPE
91920_at      NM_198427      ~
91816_f_at      NM_203304      ~
91684_g_at      NM_019037      ~
91682_at      A1571298      ~
90610_at      NM_002319      ~
90265_at      NM_006869      ~
81811_at      NM_013232      ~
79005_at      NM_024881      ~

```

Рисунок 55.

В случае редактирования этого окошка необходимо затем нажать на кнопку «Сохранить изменения» (Рис. 54). После того, как сервер выполнит эту команду, о чем можно проследить по окошку состояния и сообщениям в интернет-браузере, можно нажать кнопку «Выполнить» для реализации сценария расчетов (Рис. 54). Страница со сценарием обновляется, и теперь внизу страницы можно просмотреть новые результаты расчетов, нажав на ссылки «выходной числовой файл(ниже порога)» и «выходной текстовый файл(ниже порога)» (Рис. 55).

#### 2.2.3.4. Описание работы сценария «Выявление индуцибельных профилей экспрессии генов»

Для того чтобы выявить индуцибельные профили пользователь должен работать с данными биочип-эксперимента, относящегося к типам «временная серия» (time series или time course) или «дозовый эффект» (dose effect). Процедура состоит в фильтрации исходной таблицы с в соответствии с результатами расчета коэффициента вариации между соседними образцами. Чтобы воспользоваться сценарием по выявлению индуцибельных профилей экспрессии генов пользователь нажимает на знак «+» в строке с названием «Дифференциальная экспрессия генов», после чего на появившуюся ссылку «Выявление индуцибельных профилей экспрессии генов» (Рис. 56), в результате чего появляется соответствующая страница (Рис. 57).

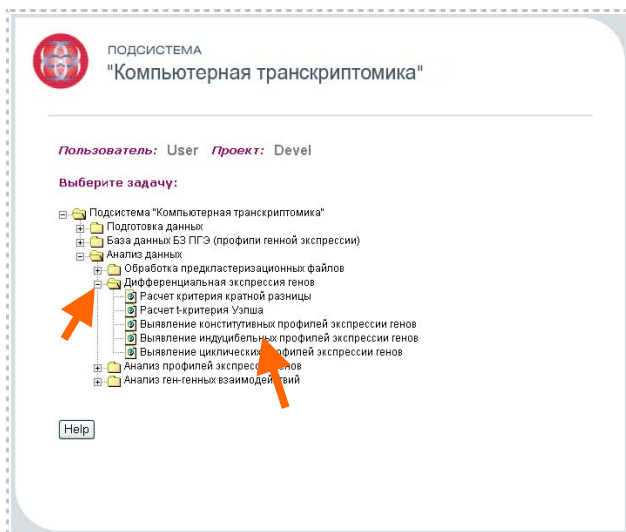


Рисунок 56.

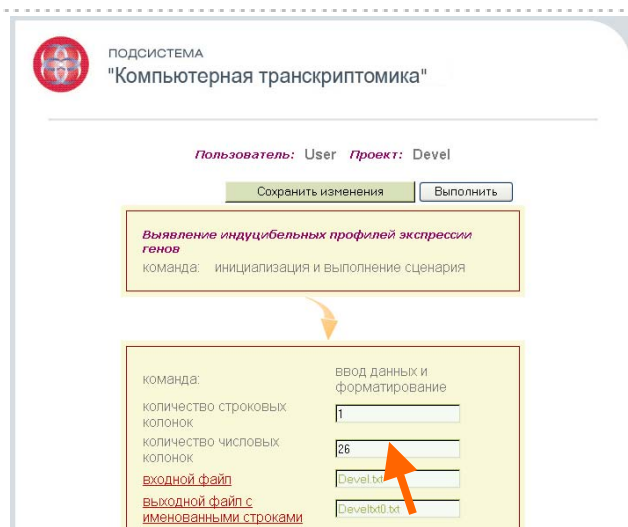


Рисунок 57.

Если числа совпадают, то можно нажать на кнопку «Выполнить» для реализации сценария расчетов (Рис. 58). В случае редактирования этого окошка необходимо затем нажать на кнопку «Сохранить изменения» (Рис. 58).

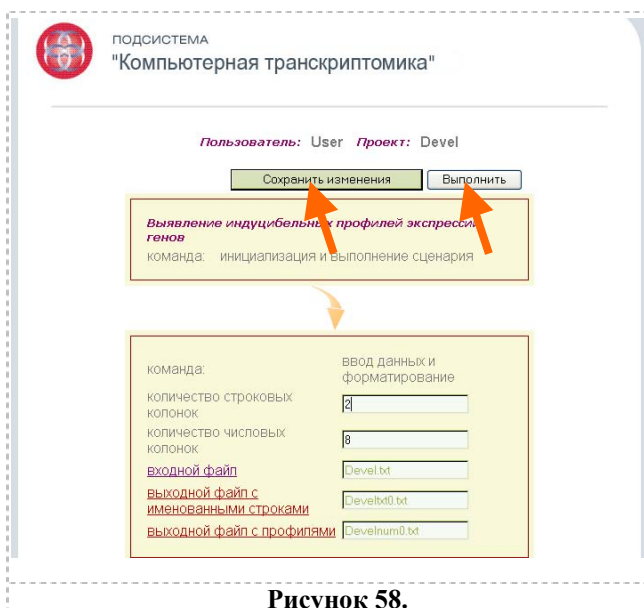


Рисунок 58.

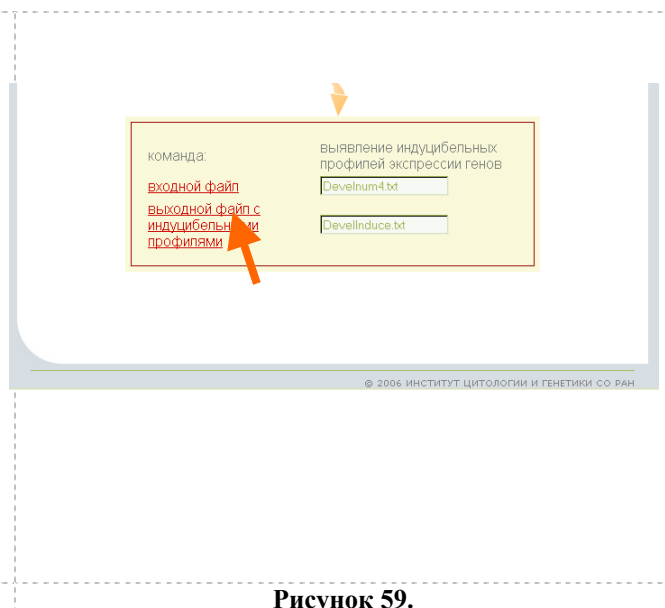


Рисунок 59.

После того, как сервер выполнит эту команду, о чем можно проследить по окошку состояния и сообщениям в интернет-браузере, можно нажать кнопку «Выполнить» для реализации сценария расчетов (Рис. 58). Страница со сценарием обновляется, и теперь внизу страницы можно посмотреть результаты расчетов, нажав на ссылку «выходной файл с индуцибельными профилями» (Рис. 59).

**DevelInduce.txt**

ID	REF	Induce	GSM13648	GSM13649	GSM13650	GSM13651
AAAP	1.58219		-0.715581	-0.156193	0.276791	
PSMA5	1.49568		0.476432	-0.0523695	-0.0150762	
ATP5F1	1.34461		0.603834	0.128444	0.26131	
TOP1	2.97306		0.300659	0.124403	0.159801	
CD34	2.68782		-0.136697	-0.132617	0.562299	
TCF3	1.93938		-0.0249469	-0.284586	-0.195873	
TOP1	2.38499		-0.288541	-0.224858	-0.516133	
HNRPL	2.06252		-0.101413	-0.0691678	0.0635699	
CYC1	1.58843		0.374306	0.634019	0.072423	

Рисунок 60.

Выходной файл (с суффиксом \*.Induce.txt) (Рис. 60) представляет список генов с расчетами отношения максимального перепада значений для этого гена к стандартному отклонению профиля, что интерпретируется как степень индуцибельности профилей.

### 2.2.3.5. Описание работы сценария «Выявление циклических профилей экспрессии генов»

Сценарий «Выявление циклических профилей экспрессии генов» применим только к таким массивам профилей, у которых образцами выступают временные отсчеты  $t_0, \dots, t_N$ . Чтобы воспользоваться сценарием по выявлению циклических профилей экспрессии генов пользователь нажимает на знак «+» в строке с названием «Дифференциальная экспрессия генов», после чего на появившуюся ссылку «Выявление циклических профилей экспрессии генов» (Рис. 61), в результате чего появляется соответствующая страница (Рис. 62).

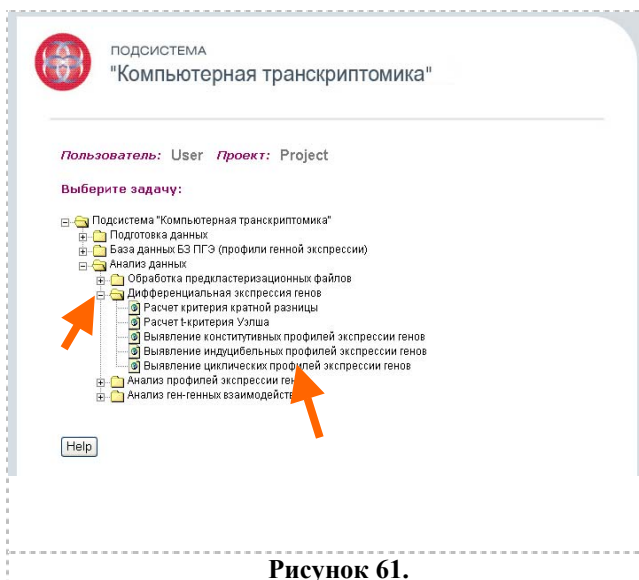


Рисунок 61.

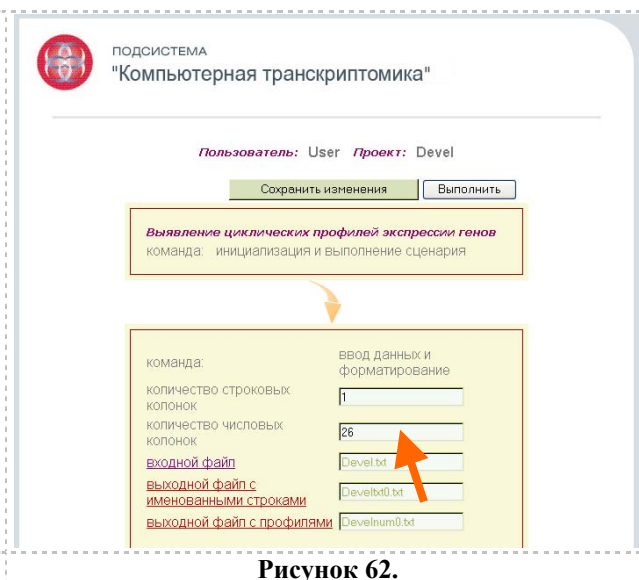


Рисунок 62.

Если числа совпадают, то можно нажать на кнопку «Выполнить» для реализации сценария расчетов (Рис. 63). В случае редактирования этого окошка необходимо затем нажать на кнопку «Сохранить изменения» (Рис. 63).

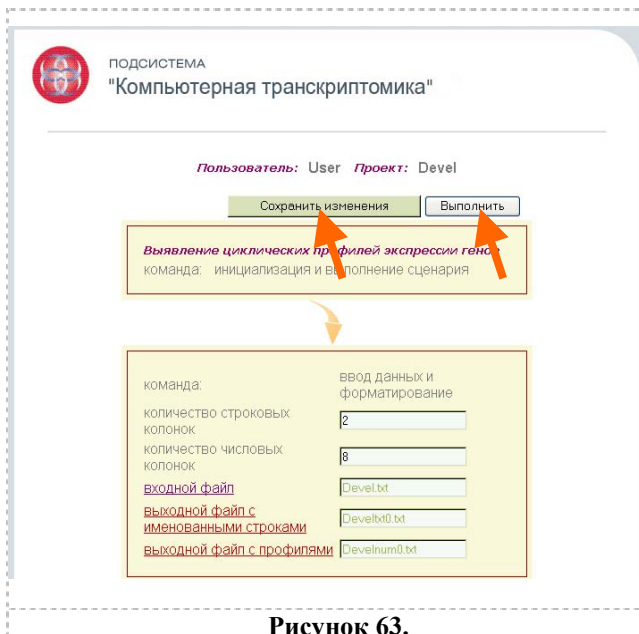


Рисунок 63.

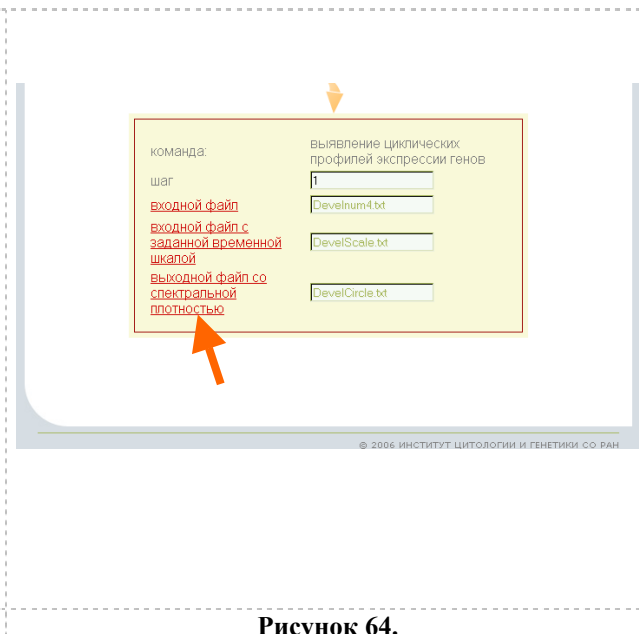
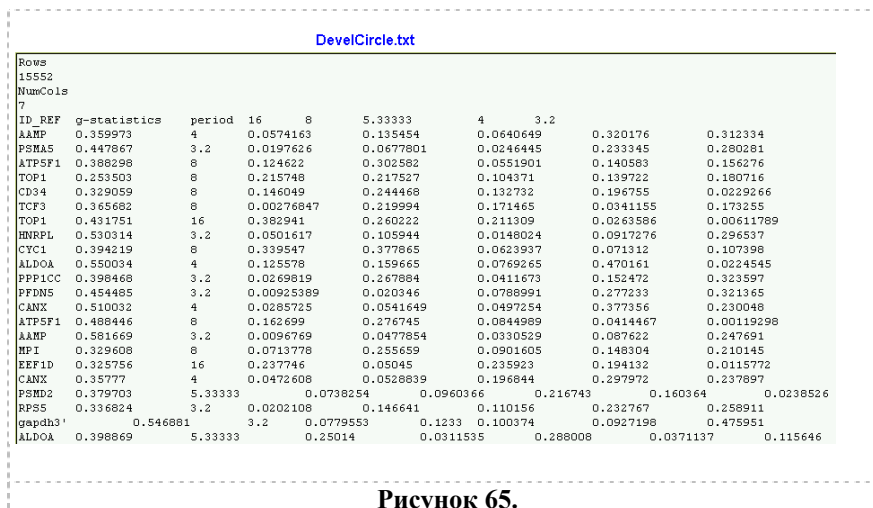


Рисунок 64.

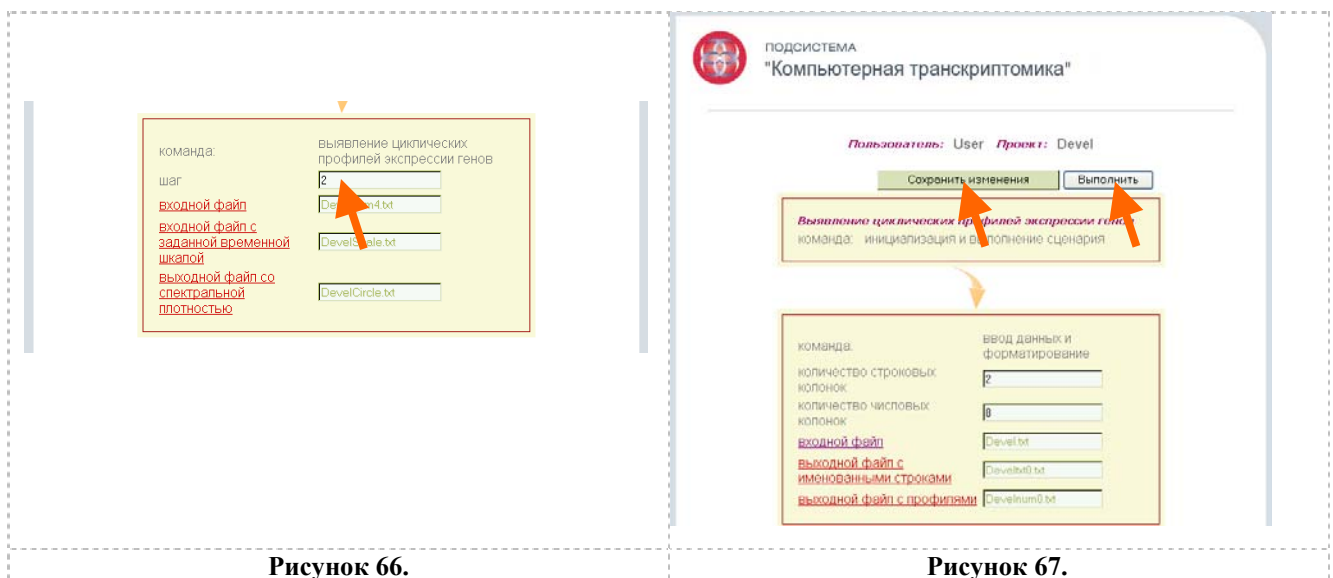
После того, как сервер выполнит эту команду, о чем можно проследить по окошку состояния и сообщениям в интернет-браузере, можно нажать кнопку «Выполнить» для реализации сценария расчетов (Рис. 63). Страница со сценарием обновляется, и теперь внизу страницы можно посмотреть результаты расчетов, нажав на ссылку «выходной файл со спектральной плотностью» (Рис. 64).



Выходной файл (с суффиксом \*Circle.txt) представляет собой таблицу со значениями g-статистики (доли спектра на максимально выраженной частоте), периода, соответствующего этой частоте, и спектральных плотностей всех Фурье-частот.

Параметр, который может изменять пользователь – это «шаг», т.е. значение дискретизации временной оси.

Если пользователь желает изменить этот параметр, он должен ввести новые значения параметров в соответствующие окошки (Рис. 66). Затем необходимо нажать на кнопку «Сохранить изменения» (Рис. 67).



После того, как сервер выполнит эту команду, о чем можно проследить по окошку состояния и сообщениям в интернет-браузере, можно нажать кнопку «Выполнить» (Рис. 67) для перерасчета и просмотреть новые результаты, используя указанные выше ссылки.

## 2.2.4. Кластеризация и классификация профилей транскрипционной экспрессии генов

Этот раздел подсистемы представляет собой комплекс программных компонент для кластеризации и многомерного анализа профилей экспрессии генов по сходству в разных

образцах/условиях и профилей образцов/условий по сходству профилей экспрессии генов. Выявление групп коэкспрессирующихся генов используется для нахождения генов, имеющих общие функции (Brazma, Vilo, 2000; Hand, Heard, 2005).

#### Иерархическая кластеризация.

Одними из распространенных методов кластеризации профилей экспрессии генов являются иерархические алгоритмы (Eisen et al., 1998; Hand, Heard, 2005). Эти алгоритмы или итеративно объединяют два наиболее похожих профиля, начиная с одного (агломеративные), или итеративно подразделяют кластеры, стартуя со всего набора профилей (дивизимные). Первый шаг состоит в вычислении матрицы расстояний между профилями генной экспрессии. После этого начинается кластеризация, т.е. поиск двух наименее «удаленных» профилей и их объединение в кластер. После каждого шага вычисляется новая матрица дистанций между образовавшимися кластерами и остальными (Pasanen et al., 2003). В подсистеме реализован аггломеративный иерархический алгоритм кластеризации по средневзвешенному расстоянию (pairwise average-linkage cluster analysis) (Eisen et al., 1998; de Hoon et al., 2004). Средневзвешенное расстояние рассчитывается как расстояние, среднее от всех попарных расстояний между двумя профилями  $x$  и  $y$ . Графическое изображение процесса объединения кластеров получается с помощью построения дендрограммы - дерева объединения кластеров. (Сценарий «Иерархическая кластеризация профилей экспрессии генов»)

#### Анализ главных компонент.

Часто для анализа профилей экспрессии генов применяют метод главных компонент (РСА). Этот метод относится к линейным методам понижения размерности массива данных, не приводящих к существенной потере информативности. При этом определяются попарно ортогональные направления максимальной вариации исходных данных, после чего данные проектируются на пространство меньшей размерности, порожденное компонентами с наибольшей вариацией (Кендалл, Стьюарт, 1976).

Пусть имеется матрица  $X$ , содержащая  $N$  строк (объектов) и  $M$  столбцов (признаков). Обозначим через  $X'$  транспонированную матрицу, в которой строки и столбцы меняются местами, и положим  $R=X'X$ . Пусть  $Q$  – матрица собственных векторов матрицы  $R$ ,  $\Lambda$  – диагональная матрица ее собственных значений и  $I$  – единичная матрица. Пусть  $U=XQ$ . Тогда (Кендалл, Стьюарт, 1976):

$$U'U = Q'X'XQ = Q'RQ = \Lambda \\ Q'Q = QQ' = I$$

Поскольку матрица  $Q$  ортогональна, то умножение на нее – это фактически поворот осей в многомерном пространстве, сохраняющий евклидово расстояние между объектами. Матрица  $U$  имеет те же размеры, что и  $X$ , однако ее столбцы не коррелируют между собой. Дисперсии  $\lambda_j$  столбцов матрицы  $U$  являются собственными значениями матрицы  $R$  и диагональными элементами матрицы  $\Lambda$ . Сами столбцы являются линейными комбинациями столбцов матрицы  $X$  с суммой квадратов коэффициентов, равной единице, и называются главными компонентами. Каждая компонента имеет дисперсию, максимально возможную из всех линейных комбинаций, ортогональных предыдущим компонентам. Обработка матрицы  $X$  методом главных компонент заключается в вычислении матриц  $U$ ,  $Q$  и  $\Lambda$ . Алгоритмы вычисления собственных векторов и собственных значений имеются в сборниках программного обеспечения современных компьютеров (Агеев и др., 1976; Гайдышев, 2001). Если собственные векторы матрицы  $Q$  умножить на квадратные корни из собственных чисел  $\lambda_j$ , то мы получим коэффициенты корреляции между компонентами и столбцами матрицы  $X$ , достоверность которых можно определять по стандартным статистическим таблицам.

Главные компоненты не коррелируют между собой. Каждая из них отвечает за свою долю изменчивости. Так как сумма дисперсий при поворотах не меняется, то смысл имеет

только доля каждой компоненты. Обычно она выражается в процентах. Главный смысл применения главных компонент заключается в том, что первые компоненты могут взять на себя значительную часть общей дисперсии и выявить реальную размерность данных.

Матрица  $R$  называется матрицей вторых или смешанных моментов. Если столбцы матрицы  $X$  центрированы ( $\sum_{ij} x_{ij}/N = 0$ ), то матрица  $R$  называется ковариационной, а если и нормированы на длину ( $\sum x_{ij}^2 = 1$ ), то корреляционной. Как правило, методу главных компонент предшествует центрирование и нормирование матрицы  $X$ .

Если поменять объекты и признаки местами, то получим транспонированную матрицу  $X'$ . Ее также можно обрабатывать методом главных компонент. Пусть  $U$ ,  $Q$  и  $\Lambda$  - матрицы, полученные в результате такой обработки и  $p = \min(N, M)$ . Тогда:

- а) ненулевые собственные значения матриц  $\Lambda$  и  $\Lambda'$  равны и их не более  $p$ ;
- б) первые  $p$  столбцов нормированной матрицы компонент  $U$  совпадают с первыми  $p$  столбцами (собственными векторами) матрицы  $Q$ ;
- в) первые  $p$  столбцов нормированной матрицы компонент  $U'$  совпадают с первыми  $p$  столбцами (собственными векторами) матрицы  $Q'$ .

В сценарии после центрирования и нормирования столбцов (колонок) реализовано центрирование и нормирование строк. Это приводит к тому, что признаками являются профили, а объектами – образцы. С вычислительной точки зрения подобная процедура эквивалентна транспонированию исходной матрицы. (Сценарий «Анализ профилей экспрессии генов методом главных компонент (PCA)»)

#### SOM-анализ.

Одним из широко используемых методов анализа микрочиповых данных являются самоорганизующиеся карты Кохонена (SOM-анализ) (Kohonen, 1997; Tamayo et al., 1999). Суть метода заключается в нелинейной трансформации множества точек, представляющих, например, профили экспрессии генов, из пространства большой размерности в визуализируемое пространство малой размерности с одновременной кластеризацией этих точек. Под центры будущих кластеров заранее выделяются узлы одномерной или двумерной прямоугольной или гексагональной решетки. Число узлов также выбирается априори. На первом шаге каждый узел заполняется следующим образом: с помощью датчика случайных чисел либо генерируются координаты центра, либо выбирается один из имеющихся профилей. Далее запускается итерационный процесс. Для каждого профиля координаты ближайшего к нему центра пересчитываются таким образом, чтобы приблизить его к этому профилю. Заодно приближаются, хотя и несколько меньше, соседи этого центра по решетке. Процесс повторяется заданное число раз. После этого каждый профиль относится к ближайшему центру кластера. Идея состоит в том, чтобы отобразить многомерное пространство в двумерное или даже одномерное с максимальным сохранением отношения соседства между точками. При реализации алгоритма предусмотрено, чтобы центры будущих кластеров не выходили за пределы множества профилей, которые после их стандартизации расположены на сфере. Операция приближения центра и его соседей является эвристической, приближение происходит по хордам, поэтому после каждой итерации все центры кластеров стандартизируются заново, возвращаясь на сферу. Очевидно, что полученные кластеры будут выпуклыми. Представляет самостоятельный интерес итоговое расположение центров кластеров среди исходного множества профилей с точки зрения выбора кластера с наиболее выраженной дифференциальной экспрессией, которая, в конечном счете, нас и интересует. Поиск отдельных генов с дифференциальной экспрессией малоперспективен вследствие большой «зашумленности» микрочиповых данных, тогда как нахождение кластера коэкспрессирующихся генов, по-разному экспрессирующихся на группах различающихся образцов, имеет гораздо более глубокий смысл. Для визуализации положения найденных центров кластеров среди исходного множества профилей можно использовать PCA



(Efimov et al., 2006). (Сценарий «Анализ профилей экспрессии генов с помощью самоорганизующихся карт признаков (SOM)»)

#### 2.2.4.1. Описание работы сценария «Иерархическая кластеризация профилей экспрессии генов»

Для того чтобы провести иерархическую кластеризацию профилей пользователю нужно нажать на знак «+» в строке с названием «Анализ профилей экспрессии генов», после чего на появившуюся ссылку «Иерархическая кластеризация профилей экспрессии генов» (Рис. 68), в результате чего появляется соответствующая страница (Рис. 69).

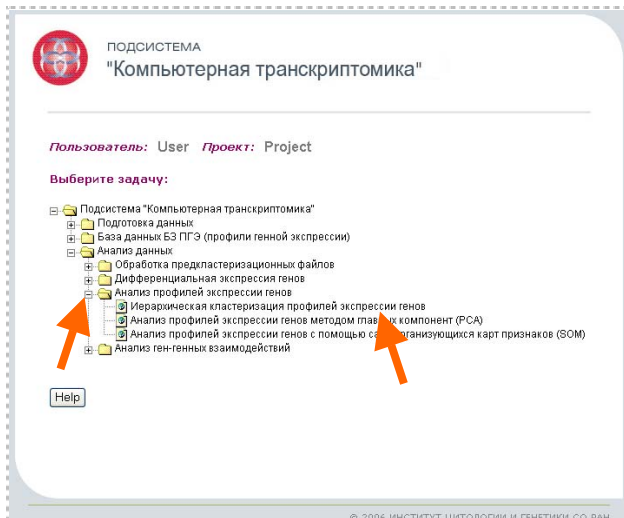


Рисунок 68.

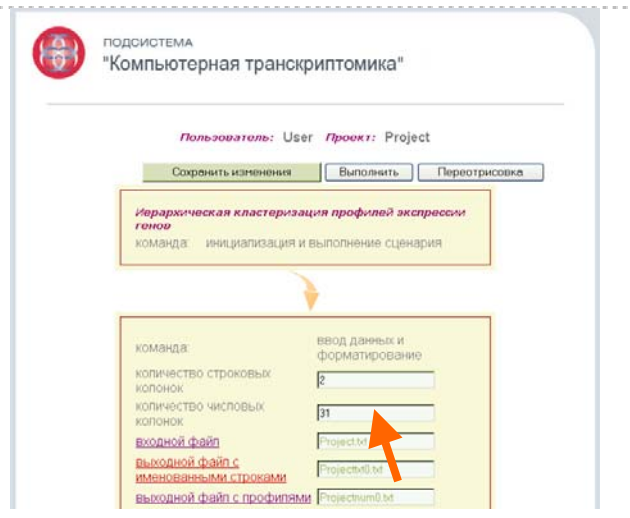
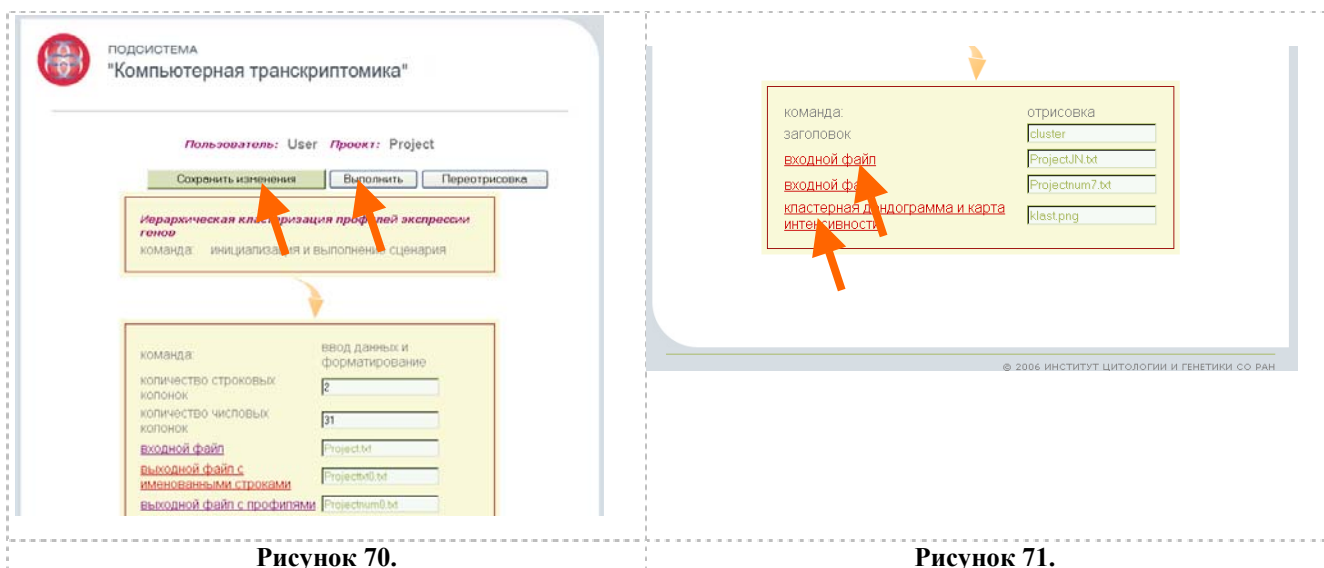


Рисунок 69.

Пользователь должен проверить соответствие значения для числа образцов, присутствующее в окошке «количество числовых колонок», значению, указанному или во вспомогательном файле «Файл группировки» (см. пункт 2.2.1.1.), если файлы были подготовлены и загружены пользователем со своего компьютера, или в документе описания биочип-исследования, если файлы были загружены пользователем из БЗ-ПГЭ.

Если числа совпадают, то можно нажать на кнопку «Выполнить» для реализации сценария расчетов (Рис. 70). В случае редактирования этого окошка необходимо затем нажать на кнопку «Сохранить изменения» (Рис. 70). После того, как сервер выполнит эту команду, о чем можно проследить по окошку состояния и сообщениям в интернет-браузере, можно нажать кнопку «Выполнить» для реализации сценария расчетов (Рис. 70). Страница с сценарием обновляется, и теперь внизу страницы можно просмотреть результаты расчетов, нажав на ссылки «входной файл» (входной для отрисовки) и «кластерная дендрограмма и карта интенсивности» (Рис. 71).



**Рисунок 70.**

**Рисунок 71.**

Результат иерархической кластеризации профилей генной экспрессии представлен в виде таблицы с числовыми значениями, а также графически в виде дендрограммы (Рис. 71).

Таблица с результатами иерархической кластеризации профилей генной экспрессии (файл \*JN.txt) (Рис. 72) является текстовым ASCII файлом с табуляторами в качестве разделителей. Первый столбец содержит числовые значения для узла, второй – числовые значения для подузла 1, третий - для подузла 2, четвертый - численные значения дистанции. Первая и вторая строка содержат информацию о количестве строк со значениями для узлов, третья и четвертая – о количестве столбцов со значениями для подузлов и дистанции, т.е. 3. Пятая строка содержит заголовки для столбцов.

```

Rows
17525
Cols
3
NODE          SUBNODE1          SUBNODE2          DISTANCE
0              0                  0                  0.0
-5785         NM_005109.1_PROBE1 NM_019595.1_PROBE1 0.201135
-1            1453873.5_PROBE1  BF056213_PROBE1  0.00590927
-2            -1                 NM_002645.1_PROBE1 0.00618454
-3            1447561.3_PROBE1 -2                 0.00814217
-4            1501738.6_PROBE1 -3                 0.00814945
-5            -4                 NM_030665.1_PROBE1 0.00847984
-6            NM_014113.1_PROBE1 -5                 0.00884507
-9            -6                 NM_033068.1_PROBE1 0.0101675

```

**Рисунок 72.**

Графический файл (Рис. 73) отображает матрицу расстояний в виде дерева, а справа от него показаны в виде карты интенсивности (чем выше интенсивность экспрессии гена, тем более красным показана ячейка, соответствующая значению экспрессии гена в определенном образце, а чем ниже – тем более синим). Сверху карты интенсивности выведены обозначения образцов, а справа – обозначения проб, представляющих гены.

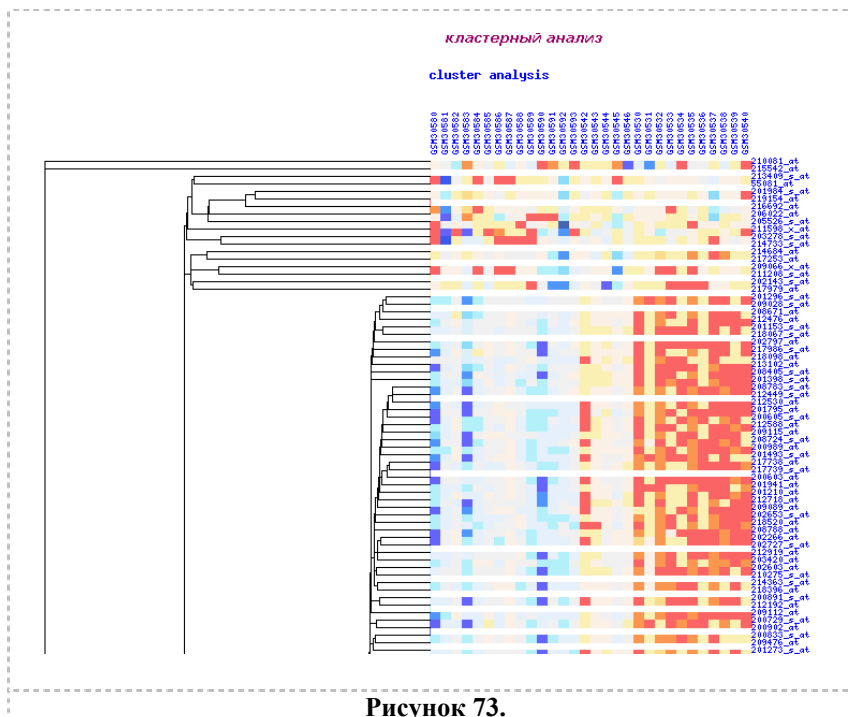


Рисунок 73.

#### 2.2.4.2. Описание работы сценария «Анализ профилей экспрессии генов методом главных компонент (PCA)»

Для того чтобы провести многомерный анализ профилей генной экспрессии методом анализа главных компонент пользователю нужно нажать на знак «+» в строке с названием «Анализ профилей экспрессии генов», после чего на появившуюся ссылку «Анализ профилей экспрессии генов методом главных компонент (PCA)» (Рис. 74), в результате чего появляется соответствующая страница (Рис. 75).

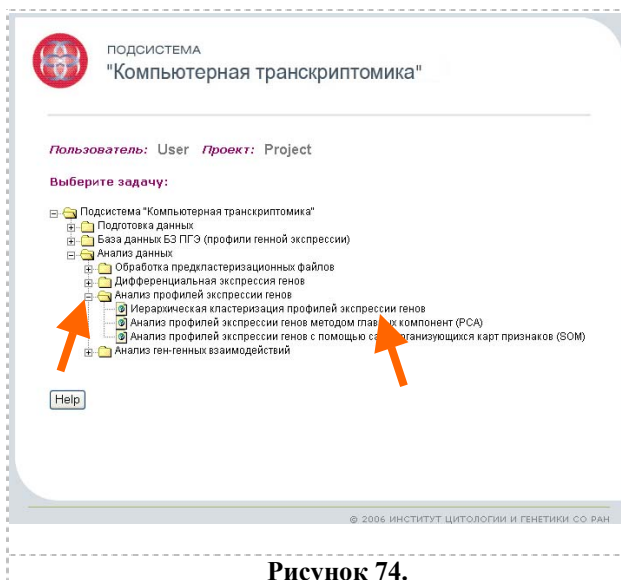


Рисунок 74.

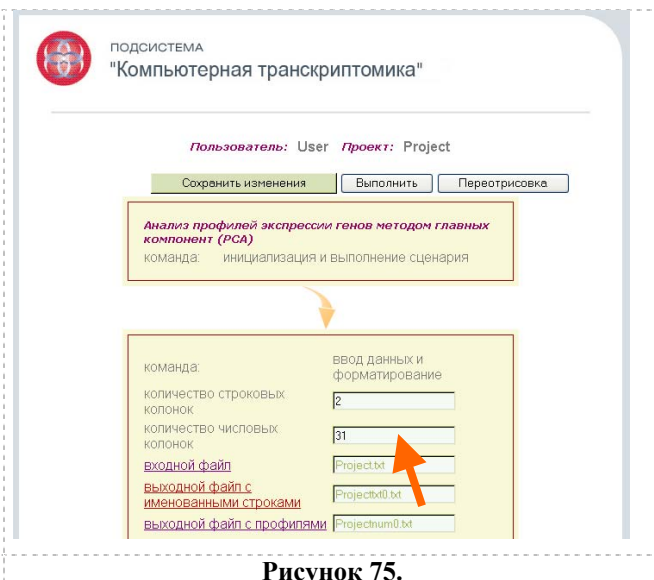


Рисунок 75.

Пользователь должен проверить соответствие значения для числа образцов, присутствующее в окошке «количество числовых колонок», значению, указанному или во вспомогательном файле «Файл группировки» (см. пункт 2.2.1.1), если файлы были

подготовлены и загружены пользователем со своего компьютера, или в документе описания биочип-исследования, если файлы были загружены пользователем из БЗ-ПГЭ.

Если числа совпадают, то можно нажать на кнопку «Выполнить» для реализации сценария расчетов (Рис. 76). В случае редактирования этого окошка необходимо затем нажать на кнопку «Сохранить изменения» (Рис. 76). После того, как сервер выполнит эту команду, о чем можно проследить по окошку состояния и сообщениям в интернет-браузере, можно нажать кнопку «Выполнить» для реализации сценария расчетов (Рис. 76). Страница с сценарием обновляется, и теперь внизу страницы можно просмотреть результаты расчетов, нажав на ссылку «расположение образцов на плоскости выбранных собственных векторов» («dataEV.png»), или на ссылку «расположение профилей экспрессии генов на плоскости выбранных главных компонент» («dataPC2.png») (Рис. 77).

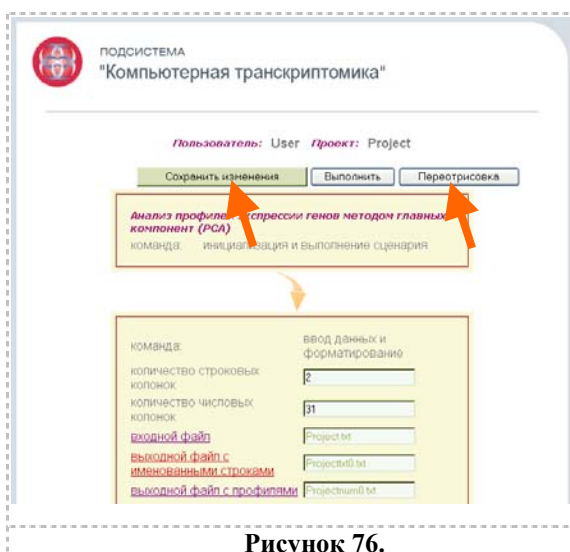


Рисунок 76.

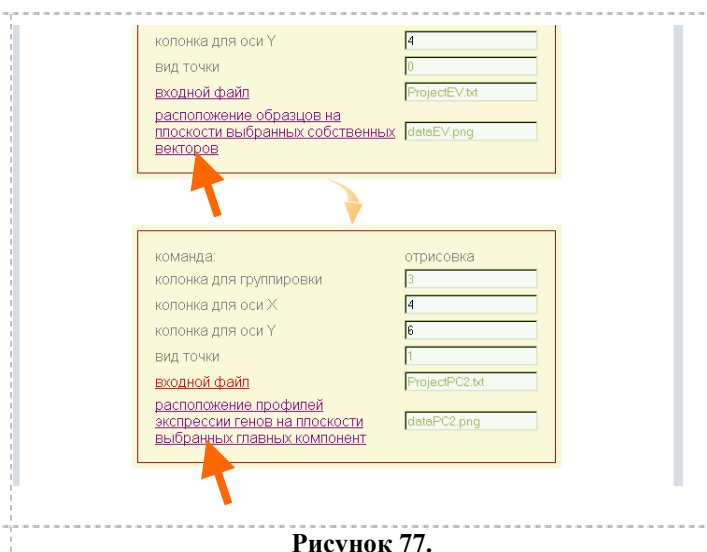


Рисунок 77.

Файл «расположение образцов на плоскости выбранных собственных векторов» («dataEV.png») демонстрирует различие в положении групп профилей образцов в пространстве после проецирования на выбранную плоскость (Рис. 78).

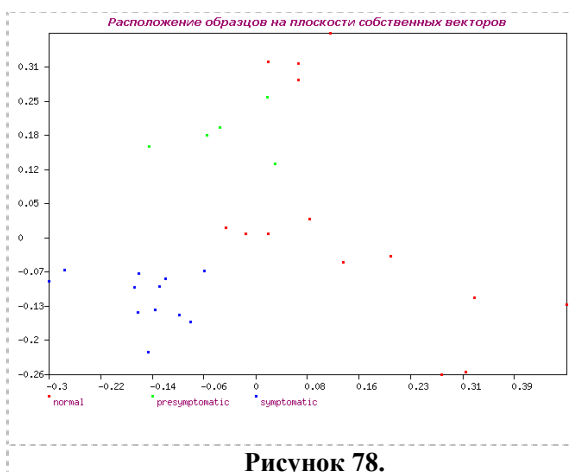


Рисунок 78.

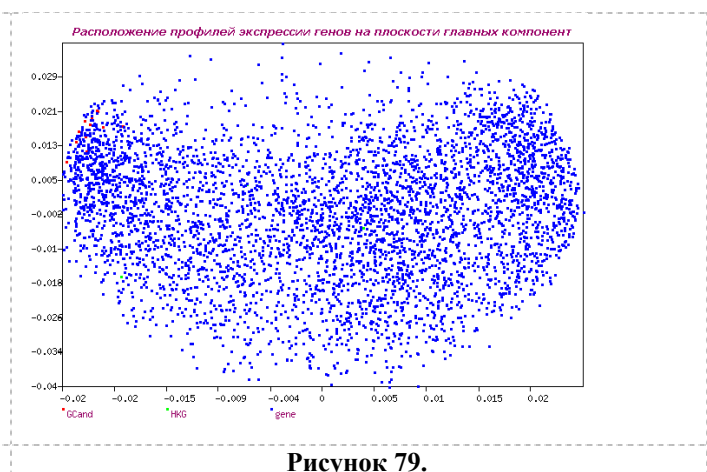
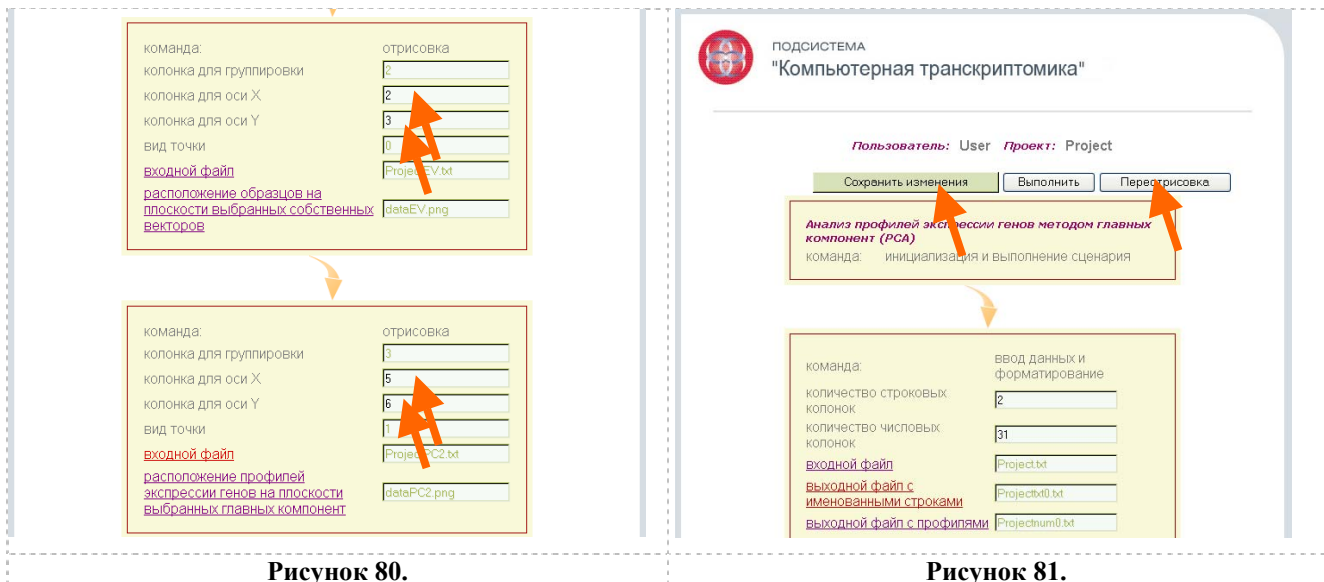


Рисунок 79.

Расположение генных профилей на плоскости выбранных главных компонент продемонстрировано в файле «dataPC2.png» (Рис. 79).

Параметры, которые может изменять пользователь – это «колонка для X-оси», номер столбца со значениями собственных векторов образцов (число должно быть больше значения

в соответствующем окошке «колонка для группировки»), которые будут откладываться по абсциссе, и «колонка для Y-оси », номер столбца со значениями собственных векторов образцов ( $n > 3$ ), которые будут откладываться по ординате.



Если пользователь желает посмотреть расположение профилей на плоскости других собственных векторов или главных компонент, он должен ввести новые значения параметров в соответствующие окошки (Рис. 80).

Затем необходимо нажать на кнопку «Сохранить изменения» (Рис. 81). После того, как сервер выполнит эту команду, о чем можно проследить по окошку состояния и сообщениям в интернет-браузере, можно нажать кнопку «Переотрисовка» (Рис. 81) для подготовки новых графиков и посмотреть их, используя указанные выше ссылки.

#### 2.2.4.3. Описание работы сценария «Анализ профилей экспрессии генов с помощью самоорганизующихся карт признаков (SOM)»

Для того чтобы провести кластерный анализ профилей генной экспрессии с помощью алгоритма SOM (самоорганизующихся карт) пользователю нужно нажать на знак «+» в строке с названием «Анализ профилей экспрессии генов», после чего на появившуюся ссылку «...» (Рис. 82), в результате чего появляется соответствующая страница (Рис. 83).

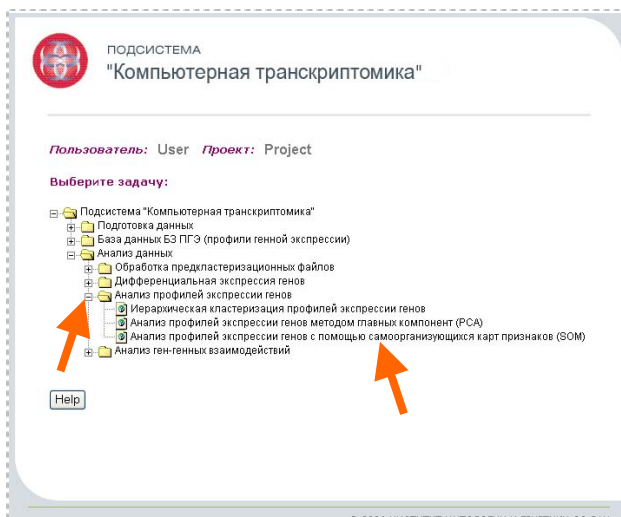


Рисунок 82.

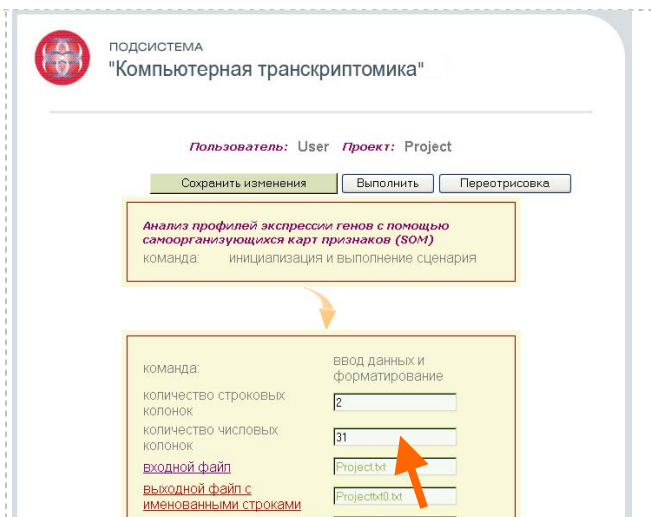


Рисунок 83.

Пользователь должен проверить соответствие значения для числа образцов, присутствующее в окошке «количество числовых колонок», значению, указанному или во вспомогательном файле «Файл группировки» (см. пункт 2.2.1.1), если файлы были подготовлены и загружены пользователем со своего компьютера, или в документе описания биочип-исследования, если файлы были загружены пользователем из БЗ-ПГЭ.

Если числа совпадают, то можно нажать на кнопку «Выполнить» для реализации сценария расчетов (Рис. 84). В случае редактирования этого окошка необходимо затем нажать на кнопку «Сохранить изменения» (Рис. 84). После того, как сервер выполнит эту команду, о чем можно проследить по окошку состояния и сообщениям в интернет-браузере, можно нажать кнопку «Выполнить» для реализации сценария расчетов (Рис. 84). Страница с сценарием обновляется, и теперь внизу страницы можно просмотреть результаты расчетов, нажав на ссылку «входной файл» (для отрисовки) и «графический файл со значениями центроидов SOM-кластеров» (Рис. 85).

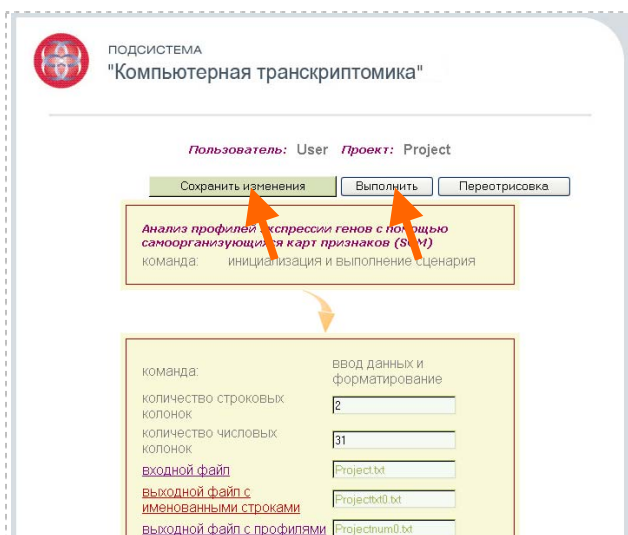


Рисунок 84.

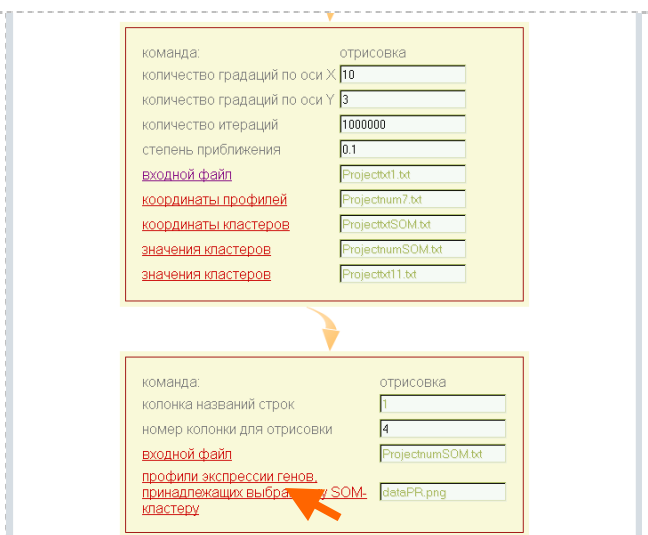


Рисунок 85.

Файл «dataPR.png» демонстрирует в графическом виде значения центроидов SOM-кластеров в разных образцах (Рис. 86). Файл с суффиксом «\*numSOM.txt» содержит результаты SOM-анализа в виде числовых данных (Рис. 87).

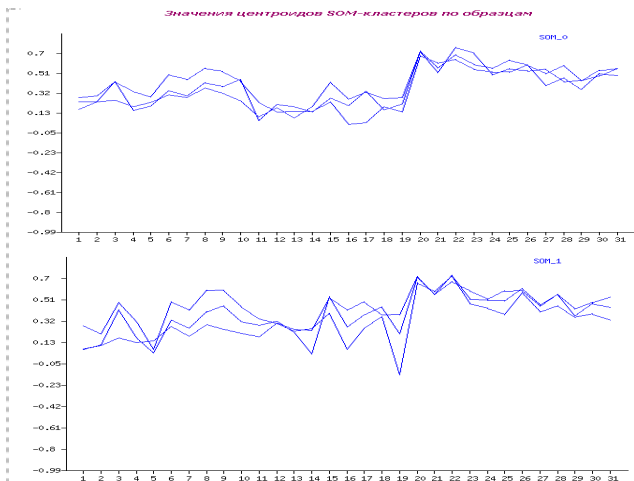


Рисунок 86.

**GDS1332numSOM.txt**

ID_REF	GSM30698	GSM30699	GSM30700	GSM30701
SOM_0_0	-0.0967424	-0.0831471	0.0541029	
SOM_0_1	-0.137601	-0.137896	0.0542873	
SOM_0_2	-0.208908	-0.141707	-0.124909	
SOM_1_0	-0.0984474	-0.172323	0.107585	
SOM_1_1	-0.304781	-0.26564	0.0409733	
SOM_1_2	-0.304028	-0.272582	-0.204094	
SOM_2_0	-0.167994	-0.23886	0.280508	
SOM_2_1	-0.274686	-0.391591	-0.0410842	
SOM_2_2	-0.311402	-0.315004	-0.19058	
SOM_3_0	-0.0929037	-0.277453	0.125241	
SOM_3_1	-0.296145	-0.46244	-0.198387	
SOM_3_2	-0.202411	-0.303673	-0.662952	

Рисунок 87.

Параметры работы сценария, которые может изменять пользователь – это «количество градаций по оси X», и «количество градаций по оси Y», «количество итераций», «степень приближения», параметры отрисовки результатов, которые может изменять пользователь - «номер колонки для отрисовки» (Рис. 88). Параметр «количество градаций по оси X» означает параметр сетки SOM-карты по абсциссе. Параметр «количество градаций по оси Y» означает параметр сетки SOM-карты по ординате. Параметр «количество итераций» означает число действий по приближению центроидов SOM-кластеров к профилям. Параметр «степень приближения» означает значение порога для приближения центроидов SOM-кластеров к профилям, принимает значения от 0 до 1 (обычно 0.05).

Рисунок 88.

Рисунок 89.

Если пользователь желает изменить значения параметров расчета по алгоритму, он должен ввести новые значения параметров в соответствующие окошки (Рис. 88). Затем

необходимо нажать на кнопку «Сохранить изменения» (Рис. 89). После того, как сервер выполнит эту команду, о чем можно проследить по окошку состояния и сообщениям в интернет-браузере, можно нажать кнопку «Выполнить» (Рис. 89) для перерасчета в соответствии с введенными новыми параметрами и просмотреть новые результаты, используя указанные выше ссылки.

Если пользователь желает изменить значения параметров для отрисовки результатов SOM-анализа, он должен ввести новое значение параметра в окошко «номер колонки для отрисовки», означающий номер SOM-кластера (Рис. 88). Затем необходимо нажать на кнопку «Сохранить изменения» (Рис. 89). После того, как сервер выполнит эту команду, о чем можно проследить по окошку состояния и сообщениям в интернет-браузере, можно нажать кнопку «Переотрисовка» (Рис. 89) для перерасчета в соответствии с введенными новыми параметрами и просмотреть новые результаты, используя указанные выше ссылки.

### **2.2.5. Выявление генов с коррелирующей экспрессией**

Этот раздел подсистемы представляет собой комплекс программных компонент для анализа корреляций в экспрессии генов по данным биочиповых экспериментов. Предполагается, что гены со сходными паттернами экспрессии и образующие в силу этого кластер, вероятнее всего, вовлечены в одни и те же биологические процессы и, в частности, регуляторные, сигнальные пути. Поэтому появляется возможность предсказывать функцию неизученных генов в кластере по аналогии с генами с известной функцией из того же кластера (D'haeseleer et al., 2000; Ideker, Lauffenburger, 2003).

Основной целью программных компонент для анализа корреляций в экспрессиях генов является поиск генов-представителей кластеров, полученных в сценарии «Анализ профилей экспрессии генов с помощью самоорганизующихся карт признаков (SOM)» (программная компонента «Кластеризация и классификация профилей транскрипционной экспрессии генов»), и выдача матрицы межгенных корреляций для дальнейшего использования при реконструкции ген-генных взаимодействий.

Начальная часть сценария «Поиск генов-представителей SOM-кластеров» совпадает со сценарием «Анализ профилей экспрессии генов с помощью самоорганизующихся карт признаков (SOM)», в котором вычисляются центры SOM-кластеров. После всех предварительных преобразований данные представлены в виде двумерной числовой таблицы «профиль экспрессии (ген) x образец» размером  $N \times M$  с текстовыми ключами строк и колонок. Существенно, что  $N \gg M$  (число образцов  $M$  находится в пределах нескольких десятков, число профилей  $N$  может достигать десятков и сотен тысяч). Далее массив исходных данных и массив центров SOM-кластеров объединяются в единый массив. Для каждого профиля ищется ближайший к нему центр SOM-кластера, к которому и относится данный профиль. Таким образом, все профили оказываются разнесенными по SOM-кластерам. Далее для каждого SOM-кластера ищется ближайший к его центру профиль, который объявляется представителем данного кластера. Гены-представители каждого SOM-кластера выводятся в отдельный файл. Гены в этом списке могут быть проассоциированы с разнообразной метаинформацией о функциях их белковых продуктов для более глубокой биологической интерпретации выявленных генных кластеров.

Одним из способов такого более глубокого анализа является вычисление матрицы межгенных корреляций. Начальная часть сценария «Вычисление матрицы межгенных корреляций» совпадает со сценарием «Поиск генов-представителей SOM-кластеров», в котором вычисляются представители каждого SOM-кластера. Для каждой пары представителей находится скалярное произведение профилей. Вследствие сделанных нормировок оно является коэффициентом корреляции между представителями. Матрица коэффициентов корреляции



между представителями SOM-кластеров выводится в отдельный файл. Эту матрицу можно анализировать средствами генной онтологии для возможного выявления паттерна ген-генных взаимодействий (D'haeseleer et al., 2000).

### 2.2.5.1. Описание работы сценария «Поиск генов-представителей SOM-кластеров»

Для того чтобы провести поиск профилей генов, объединенных в кластеры по SOM-алгоритму, пользователю нужно нажать на знак «+» в строке с названием «Анализ ген-генных взаимодействий», после чего на появившуюся ссылку «Поиск генов-представителей SOM-кластеров» (Рис. 90), в результате чего появляется соответствующая страница (Рис. 91).

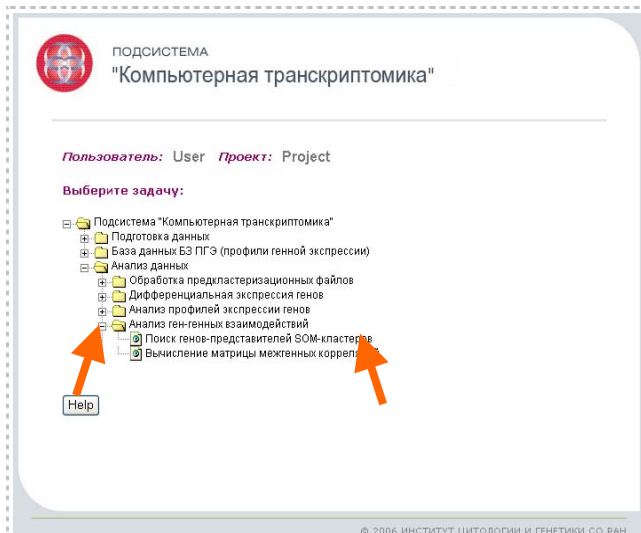


Рисунок 90.

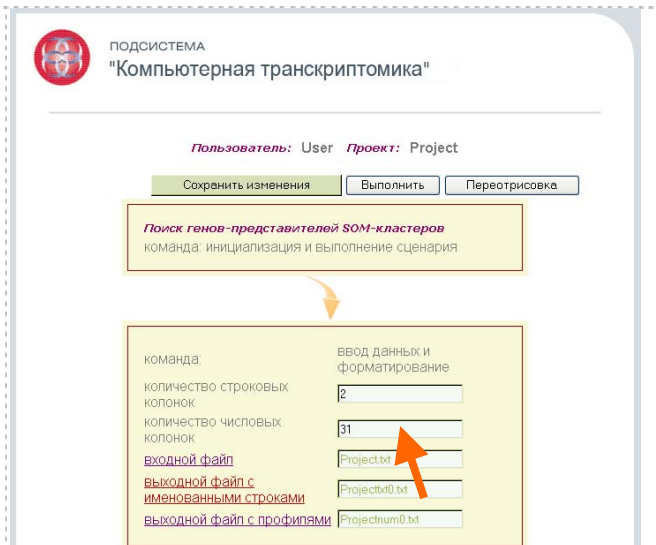


Рисунок 91.

Пользователь должен проверить соответствие значения для числа образцов, присутствующее в окошке «количество числовых колонок», значению, указанному или во вспомогательном файле «Файл группировки» (см. пункт 2.2.1.1.), если файлы были подготовлены и загружены пользователем со своего компьютера, или в документе описания биочип-исследования, если файлы были загружены пользователем из БЗ-ПГЭ.

Если числа совпадают, то можно нажать на кнопку «Выполнить» для реализации сценария расчетов (Рис. 92). В случае редактирования этого окошка необходимо затем нажать на кнопку «Сохранить изменения» (Рис. 92). После того, как сервер выполнит эту команду, о чем можно проследить по окошку состояния и сообщениям в интернет-браузере, можно нажать кнопку «Выполнить» для реализации сценария расчетов (Рис. 92). Страница с сценарием обновляется, и теперь внизу страницы можно просмотреть результаты расчетов, нажав на ссылки «графический файл со значениями центроидов SOM-кластеров», «выходной числовой файл» или «выходной текстовый файл» (Рис. 93).

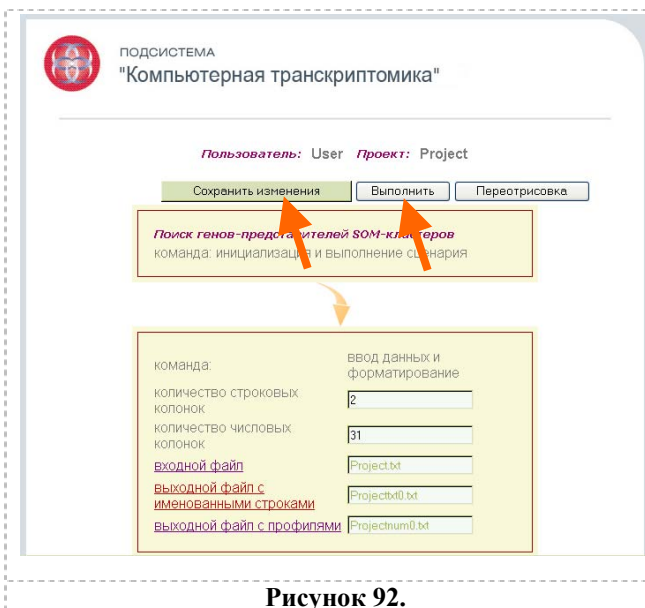


Рисунок 92.

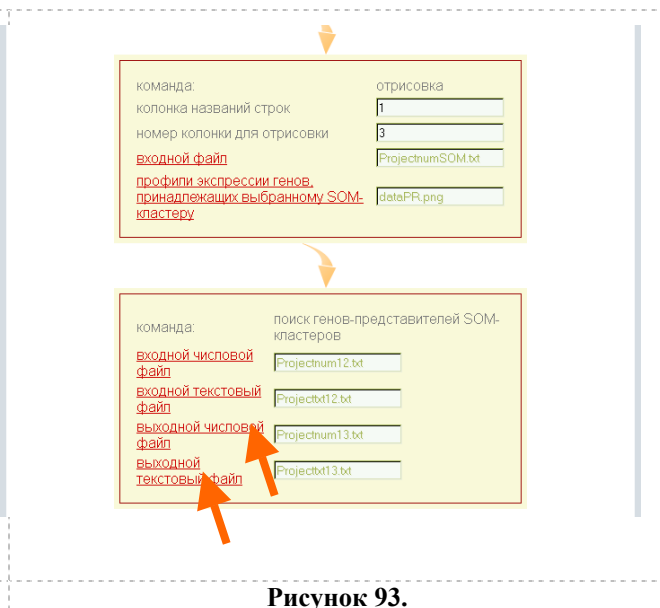


Рисунок 93.

Файл «dataPR.png» демонстрирует в графическом виде значения центроидов SOM-кластеров в разных образцах (Рис. 2.2.5.1.5). Выходные числовой файл (с суффиксом «\*num13.txt») и текстовый файл (с суффиксом «\*txt13.txt») (Рис. 2.2.5.1.6) дают возможность просмотреть результаты расчетов сценария и список генов-представителей каждого SOM-кластера.



Рисунок 94.

**Projectnum13.txt**

ID_REF	GSM30698	GSM30699	GSM30700	GSM30701
2767012CB1_PROBE1	-0.135226	-0.0510096	0.057427	
246037.33_PROBE1	-0.170977	-0.125217	0.101629	
1177116.1_PROBE1	-0.251713	-0.114283	-0.0515301	
1399108.1_PROBE1	-0.186346	-0.160427	0.169686	
1319716.8_PROBE1	-0.351411	-0.27695	-0.0270627	
253580.1_PROBE1	-0.321083	-0.209182	-0.22189	
1118800.1_PROBE1	-0.104952	-0.303266	0.27471	
1383366.34_PROBE1	-0.360249	-0.355543	-0.0389569	

**Projecttxt13.txt**

ID_REF	IDENTIFIER	TYPE	Xgrid	Ygrid	ClusterDist
2767012CB1_PROBE1	NH_001539	gene	0	0	0.116705
246037.33_PROBE1	NH_004242	gene	0	1	0.0745309
1177116.1_PROBE1	NH_152726	gene	0	2	0.0251368
1399108.1_PROBE1	AK024196	gene	1	0	0.227445
1319716.8_PROBE1	NH_017924	gene	1	1	0.108695
253580.1_PROBE1	AK074962	gene	1	2	0.0434715
1118800.1_PROBE1	1118800.1	gene	2	0	0.169648
1383366.34_PROBE1	NH_000983	gene	2	1	0.161928

Рисунок 95.

Параметры работы сценария, которые может изменять пользователь – это «количество градаций по оси X», и «количество градаций по оси Y», «количество итераций», «степень приближения», параметры отрисовки результатов, которые может изменять пользователь – «колонка названий строк», «номер колонки для отрисовки» (Рис. 96.). Параметр «количество градаций по оси X» означает параметр сетки SOM-карты по абсциссе. Параметр «количество градаций по оси Y» означает параметр сетки SOM-карты по ординате. Параметр

«количество итераций» означает число действий по приближению центроидов SOM-кластеров к профилям. Параметр «степень приближения» означает значение порога для приближения центроидов SOM-кластеров к профилям, принимает значения от 0 до 1 (обычно 0.05).

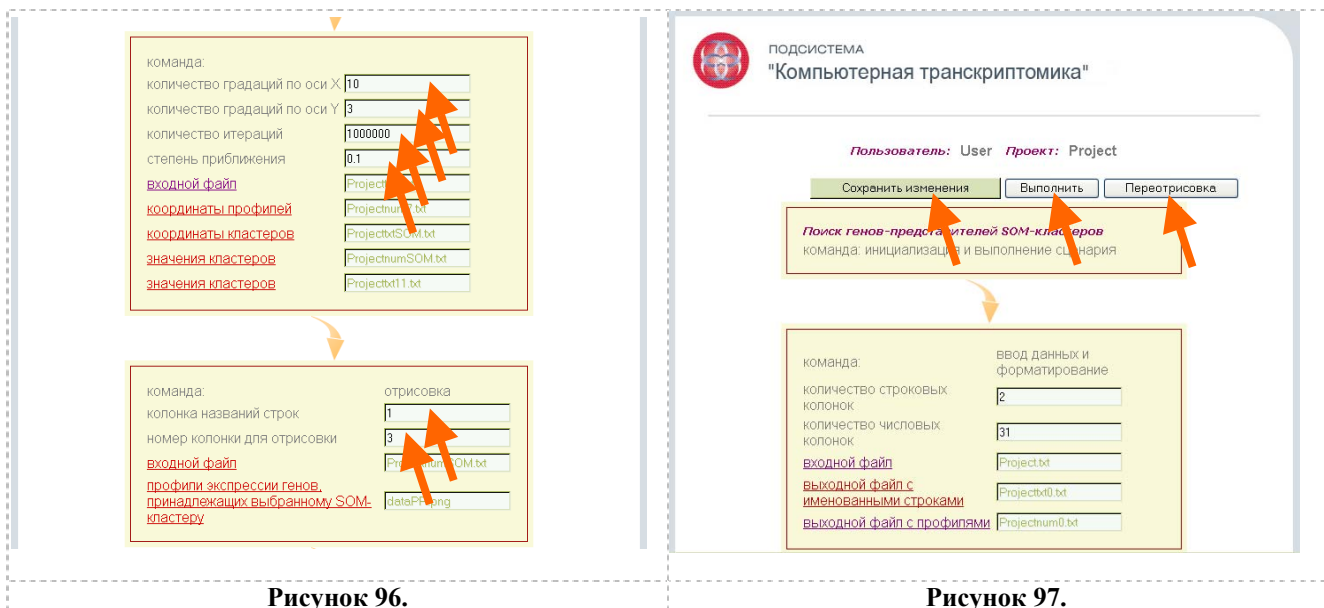


Рисунок 96.

Рисунок 97.

Если пользователь желает изменить значения параметров, он должен ввести новые значения параметров в соответствующие окошки (Рис. 96).

Затем необходимо нажать на кнопку «Сохранить изменения» (Рис. 97). После того, как сервер выполнит эту команду, о чем можно проследить по окошку состояния и сообщениям в интернет-браузере, можно нажать кнопку «Выполнить» (Рис. 97) для перерасчета в соответствии с введенными новыми параметрами, или кнопку «Переотрисовка» (Рис. 97) для изменения графического представления результатов, и просмотреть новые результаты, используя указанные выше ссылки.

#### 2.2.5.2. Описание работы сценария «Вычисление матрицы межгенных корреляций»

Для того чтобы выявить профили генов со коррелирующей экспрессией пользователю нужно нажать на знак «+» в строке с названием «Анализ ген-генных взаимодействий», после чего на появившуюся ссылку «Вычисление матрицы межгенных корреляций» (Рис. 98), в результате чего появляется соответствующая страница (Рис. 99).

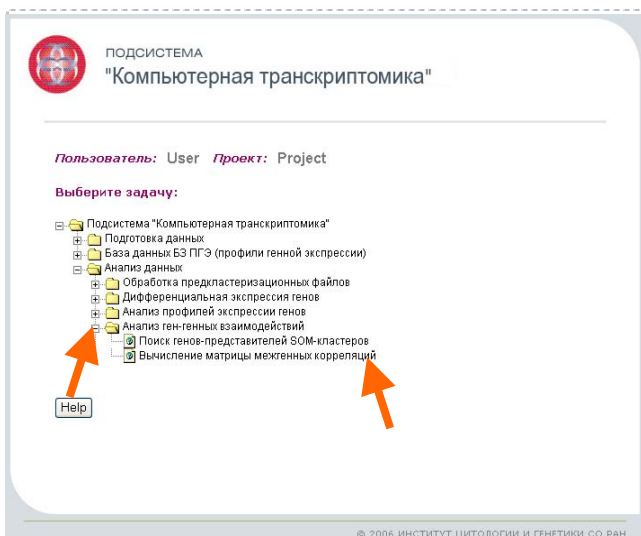


Рисунок 98.

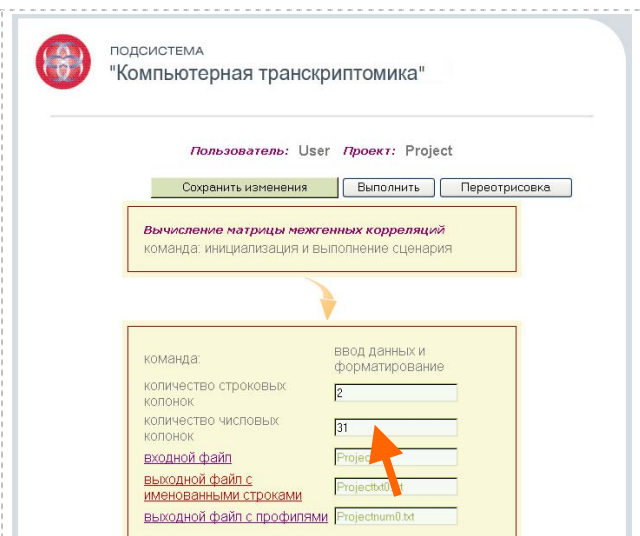


Рисунок 99.

Пользователь должен проверить соответствие значения для числа образцов, присутствующее в окошке «количество числовых колонок», значению, указанному или во вспомогательном файле «Файл группировки» (см. пункт 2.2.1.1.), если файлы были подготовлены и загружены пользователем со своего компьютера, или в документе описания биочип-исследования, если файлы были загружены пользователем из БЗ-ПГЭ.

Если числа совпадают, то можно нажать на кнопку «Выполнить» для реализации сценария расчетов (Рис. 100).

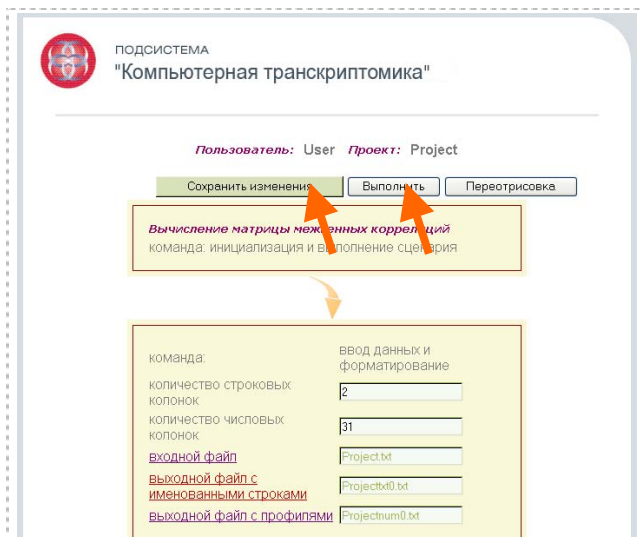


Рисунок 100.

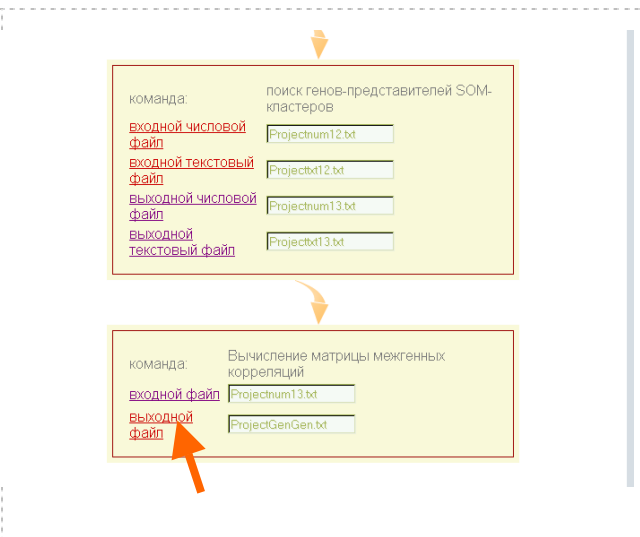


Рисунок 101.

В случае редактирования этого окошка необходимо затем нажать на кнопку «Сохранить изменения» (Рис. 100). После того, как сервер выполнит эту команду, о чем можно проследить по окошку состояния и сообщениям в интернет-браузере, можно нажать кнопку «Выполнить» для реализации сценария расчетов (Рис. 100). Страница с сценарием обновляется, и теперь внизу страницы можно просмотреть результаты расчетов, нажав на ссылки «графический файл со значениями центроидов SOM-кластеров» или «выходной файл» (Рис. 101).

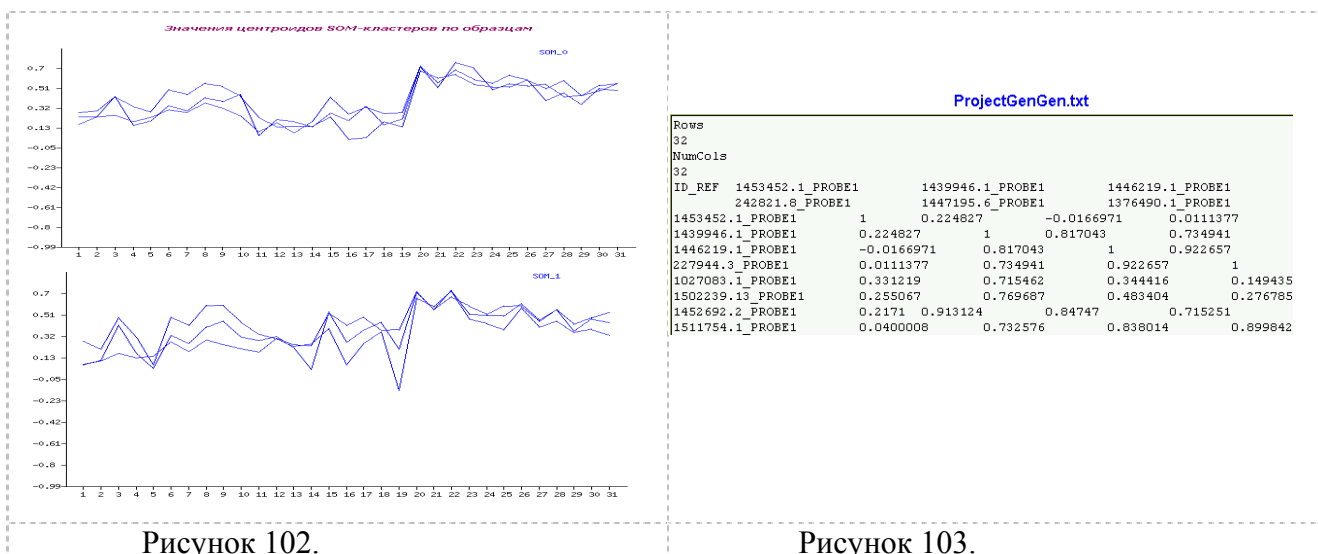


Рисунок 102.

Рисунок 103.

Файл «dataPR.png» демонстрирует в графическом виде значения центроидов SOM-кластеров в разных образцах (Рис. 102).

Выходной файл (с суффиксом «\*GenGen.txt») является матрицей коэффициентов корреляции между представителями SOM-кластеров (Рис. 103).

### 3. Полезные ссылки.

1. Агеев М.И., Алик В.П., Марков Ю.И. (1976) Библиотека алгоритмов 516-1006. М.: Сов. радио, 136с. (Справочное пособие; Вып.2).
2. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. (1985) Прикладная статистика: Исследование зависимостей. М.: Финансы и статистика, 487с.
3. Гайдышев И. (1991) Анализ и обработка данных: специальный справочник. СПб: Питер, 2001. 752с.
4. Животовский Л.А. (1991) Популяционная биометрия. М.: Наука, 267с.
5. Кендалл М., Стьюарт А. (1976) Многомерный статистический анализ и временные ряды. М.: Наука, 736с.
6. Крамер Г. (1975) Математические методы статистики. М.: Мир, 648с.
7. Мирзабеков А.Д. (2003) Биочипы в биологии и медицине XXI века // Вестник Российской Академии Наук. т. 73, № 5, с. 412-422.
8. Allison D.B., Cui X., Page G.P., Sabripour M. (2006) Microarray data analysis: from disarray to consolidation and consensus. Nat Rev Genet. 7(1):55-65.
9. Anderle P., Duval M., Draghici S., Kuklin A., Littlejohn T.G., Medrano J.F., Vilanova D., Roberts M.A. (2003) Gene expression databases and data mining. Biotechniques. Suppl:36-44.
10. Ball C.A., Ihab A.B.A., Demeter J., et al., (2005) The Stanford Microarray Database accommodates additional microarray platforms and data formats. Nucl. Acids Res. 33(Database issue):D580-D582.
11. Barrett T., Suzek T.O., Troup D.B., Wilhite S.E., Ngau W.C., Ledoux P., Rudnev D., Lash A.E., Fujibuchi W., Edgar R. (2005) NCBI GEO: mining millions of expression profiles--database and tools. Nucleic Acids Res. 33(Database issue):D562-566.
12. Brazma A., Vilo J. (2000) Gene expression data analysis. FEBS Lett. 480(1):17-24.

13. Brenner S., Johnson M., Bridgham J., *et al.*, (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat Biotechnol.* 18(6):630-634.
14. Chen J. (2005) Identification of significant periodic genes in microarray gene expression data. *BMC Bioinformatics.* 30;6:286.
15. Cheung V.G., Morley M., Aguilar F., Massimi A., Kucherlapati R., Childs G. (1999) Making and reading microarrays. *Nat Genet.* 21(1 Suppl):15-19.
16. Claverie J.M. (1999) Computational methods for the identification of differential and coordinated gene expression. *Hum Mol Genet.* 8:1821-1832.
17. de Hoon M.J., Imoto S., Nolan J., Miyano S. (2004) Open source clustering software. *Bioinformatics.* 20(9):1453-1454.
18. Deyholos M.K., Galbraith D.W. (2001) High-density microarrays for gene expression analysis. *Cytometry.* 43(4):229-238.
19. D'haeseleer P., Liang S., Somogyi R. (2000) Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics.* 16(8):707-726.
20. Duggan D.J., Bittner M., Chen Y., Meltzer P., Trent J.M. (1999) Expression profiling using cDNA microarrays. *Nat Genet.* 21(1 Suppl):10-14.
21. Eisen M.B., Spellman P.T., Brown P.O., Botstein D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A.* 95(25):14863-14868.
22. Efimov V.M., Badratinov M.S., Katokhin A.V. (2006) Interpretation of results of SOM-analysis of microarray data by principal components //Proceedings of the Fifth International conference on bioinformatics of genome regulation and structure (Novosibirsk, Russia, July 16-22 2006). –Novosibirsk: ICG SB RAS, V. 1. P. 44-47.
23. Greenfield A. (2000) Applications of DNA microarrays to the transcriptional analysis of mammalian genomes. *Mamm Genome.* 11(8):609-613.
24. Hack C.J. (2004) Integrated transcriptome and proteome data: the challenges ahead. *Brief Funct Genomic Proteomic.* 3(3):212-219.
25. Hand D.J., Heard N.A. (2005) Finding groups in gene expression data. *J Biomed Biotechnol.* 2005, 215-225.
26. Holloway A.J., van Laar R.K., Tothill R.W., Bowtell D.D. (2002) Options available--from start to finish--for obtaining data from DNA microarrays II. *Nat Genet.* 32 Suppl:481-489.
27. Hu Y.F., Kaplow J., He Y. (2005) From traditional biomarkers to transcriptome analysis in drug development. *Curr Mol Med.* 5(1):29-38.
28. Ideker T., Lauffenburger D. (2003) Building with a scaffold: emerging strategies for high- to low-level cellular modeling. *Trends Biotechnol.* 21(6):255-262.
29. Kohonen, T. (1982) Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43, 59-69.
30. Lindlof A. (2003) Gene identification through large-scale EST sequence processing. *Appl Bioinformatics.* 2(3):123-129.
31. Mendes Soares LM, Valcarcel J. (2006) The expanding transcriptome: the genome as the 'Book of Sand'. *EMBO J.* 25(5):923-931.
32. Moreau Y., Aerts S., De Moor B., De Strooper B., Dabrowski M. (2003) Comparison and meta-analysis of microarray data: from the bench to the computer desk. *Trends Genet.* 19(10):570-577.
33. Mutch D.M., Berger A., Mansourian R., Rytz A., Roberts M.A. (2002) The limit fold change model: a practical approach for selecting differentially expressed genes from microarray data. *BMC Bioinformatics.* 3(1):17.
34. Nadon R, Shoemaker J. (2002) Statistical issues with microarrays: processing and analysis. *Trends Genet.* 18(5):265-271.

35. Okubo K., Hori N., Matoba R., Niiyama T., Fukushima A., Kojima Y., Matsubara K. (1992) Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. *Nature Genetics*, 2:173-179.
36. Parkinson H., Sarkans U., Shojatalab M., Abeygunawardena N., Contrino S., Coulson R., Farne A., Lara G.G., Holloway E., Kapushesky M., Lilja P., Mukherjee G., Oezcimen A., Rayner T., Rocca-Serra P., Sharma A., Sansone S., Brazma A. (2005) ArrayExpress--a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* 33(Database issue):D553-555.
37. Pasanen T., Saarela J., Saarikko I., Toivanen T., Tolvanen M., Vihinen M., G. Wong. (2003) In: J. Tuimala and M. M. Laine (eds) *DNA Microarray Data Analysis*, 161 pages. CSC - Scientific Computing Ltd, Finland.
38. Quackenbush J. (2001) Computational analysis of microarray data. *Nat Rev Genet*, 2:418-427
39. Rhodes D.R., Chinnaiyan A.M. (2005) Integrative analysis of the cancer transcriptome. *Nat Genet.* 37 Suppl:S31-37.
40. Stears R.L., Martinsky T., Schena M. (2003) Trends in microarray analysis. *Nat Med.* 9(1):140-145.
41. Stoeckert C.J. Jr, Causton H.C., Ball C.A. (2002) Microarray databases: standards and ontologies. *Nat Genet.* 32 Suppl:469-473.
42. Tamayo P., Slonim D., Mesirov J., Zhu Q., Kitareewan S., Dmitrovsky E., Lander E., and Golub T. (1999) Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA*, 96, 2907-2912.
43. Velculescu V.E., Zhang L., Vogelstein B., Kinzler K.W. (1995) Serial analysis of gene expression. *Science*. 270(5235):484-487.
44. Velculescu V.E., Vogelstein B., Kinzler K.W. (2000) Analysing uncharted transcriptomes with SAGE. *Trends Genet.* 16(10):423-425.
45. Whitfield M.L., Sherlock G., Saldanha A.J., Murray J.I., Ball C.A., Alexander K.E., Matese J.C., Perou C.M., Hurt M.M., Brown P.O., Botstein D. (2002) Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol Biol Cell.* 13(6):1977-2000.
46. Wichert S., Fokianos K., Strimmer K. (2004) Identifying periodically expressed transcripts in microarray time series data. *Bioinformatics.* 20(1):5-20.
47. Zhu T. (2003) Global analysis of gene expression using GeneChip microarrays. *Curr Opin Plant Biol.* 6(5):418-425.