УДК 004.031.2

**Д. А. Полунин** [1], **И. А. Штайгер** [1], **В. М Ефимов** [2, 3, 4]

[1] Новосибирский государственный университет
ул. Пирогова, 2, Новосибирск, 630090, Россия

[2] Институт цитологии и генетики СО РАН
пр. Акад. Лаврентьева, 10, Новосибирск, 630090, Россия

[3] Институт систематики и экологии животных СО РАН
ул. Фрунзе, 11, Новосибирск, 630091, Россия

[4] Томский государственный университет
пр. Ленина, 36, Томск, 634050, Россия

E-mail: polunin.denis@gmail.com; irina.shtaiger@gmail.com;
efimov@bionet.nsc.ru

# РАЗРАБОТКА ПРОГРАММНОГО КОМПЛЕКСА JACOBI 4 ДЛЯ МНОГОМЕРНОГО АНАЛИЗА МИКРОЧИПОВЫХ ДАННЫХ *

Программный комплекс JACOBI 4 является четвертой версией пакета JACOBI для многомерного анализа микрочиповых данных. Проект JACOBI развивается для поддержки новой технологии поиска генов-кандидатов в генные сети, разработанной в ИЦиГ СО РАН. Пакет представляет собой набор программ для многомерного анализа с открытым кодом, который может быть одинаково легко дополнен или изменен как пользователем, имеющим небольшой опыт работы с ПК, так и опытным пользователем.

*Ключевые слова*: программный пакет, многомерный анализ, скриптовый язык, скрипт, программа, подпрограмма.

## Введение

Технология микрочипов используется для исследования экспрессии тысяч генов в рамках одного эксперимента. С течением времени технология становится все более доступной, что приводит к ее применению в различных областях биологии. Микрочипы позволяют исследователям определить, какие гены экспрессируются в клетках заданного типа в определенное время при определенных условиях [1; 2].

Микрочиповые данные могут быть проинтерпретированы как таблица «объект – признак». В ИЦиГ СО РАН разработана новая технология поиска генов-кандидатов в генные сети, которая заключается в геометрическом представлении любого набора объектов множеством точек в евклидовых пространствах через вычисление матриц сходства-различия между их описаниями и сведение их к матрицам евклидовых расстояний. Несмотря на то что все звенья данной методологии хорошо проработаны в современной науке и поддержаны средствами многих пакетов, современные пакеты неудобны для ее применения.

* Разработка программного комплекса поддержана грантом РФФИ Проект № 13-07-00315 «Интеллектуальный анализ и комбинирование гетерогенных данных».

*Полунин Д. А., Штайгер И. А., Ефимов В. М* Разработка программного комплекса JACOBI 4 для многомерного анализа микрочиповых данных // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2014. Т. 12, вып. 2. С. 90–98.

Для поддержки этой технологии реализован программный комплекс с открытым кодом JACOBI 4, предназначенный в основном для потоковой обработки похожих данных одним и тем же алгоритмом. Целевую аудиторию JACOBI 4 можно разделить на две группы: пользователи, имеющие небольшой опыт работы с ПК, и продвинутые пользователи ПК. Основным принципом JACOBI 4 является разделение функциональности для этих двух групп пользователей таким образом, чтобы функциональность, которая считается простой и понятной для первой группы пользователей, была достаточной для полноценной работы и не смешивалась с дополнительными возможностями, которые требуют большего понимания. Такое разделение позволит избежать перегруженности интерфейса. JACOBI 4 базируется на разработанном ранее пакете JACOBI 2 [3; 4].

**Структура и схема функционирования JACOBI 4**

В JACOBI 4 разделение основной и дополнительной функциональности происходит на уровне архитектуры. Необходимый набор программ (ядро программного комплекса) включает головную программу-диспетчер в графическом и консольном вариантах и набор программ – директив комплекса. Всего в пакете предусмотрено 4 подсистемы: поддержки целостности, документации, тестирования встраиваемых модулей, синхронизации, а также головной модуль, реализующий взаимодействие с пользователем. Головной модуль вместе с документацией и директивами предоставляет функциональность первого уровня (для пользователей с небольшим опытом работы с ПК). Модуль синхронизации необходим только при условии, что пользователь использует дополнительные возможности: дополнения установятся корректно, если пользователь не изменит настройки по умолчанию (это возможно сделать только при помощи дополнительных функций).

Предполагается, что директивы вызываются программой-диспетчером, однако они являются самостоятельными программами с консольным интерфейсом, т. е. возможна работа непосредственно с директивами, минуя диспетчер.

JACOBI 4 ориентирован на работу с пользовательским скриптом, написанным на специальном языке, и предназначен для одинаковой обработки большого количества однотипных данных.

**Требования к программному комплексу**

1. Использование распространенного формата для хранения входных файлов.
2. Согласованность данных – каждая таблица имеет угловой ключ, ключи строк, ключи столбцов.
3. Согласованность аргументов – при вызове директив сначала указываются имена входных файлов, потом имена выходных файлов, затем параметры алгоритма.

Для хранения таблиц используется формат csv, который поддерживается многими пакетами для обработки данных, таких как R, STATISTICA и др.

Таблица представляет собой матрицу данных, дополненную ключами строк, столбцов и угловым ключом. При этом в качестве значения элемента матрицы может использоваться строка, число или отсутствие данных.

Заметим, что не каждая директива допускает на входе нечисловые значения, поэтому в комплекс добавлена возможность выполнения предварительной обработки данных.

Директивы программного комплекса можно разделить на две большие группы: директивы для подготовки данных и директивы для многомерного анализа данных [5–7].

**Директивы для подготовки данных**

При обработке данных постоянно возникает необходимость в преобразовании входных или промежуточных данных. С этой целью написаны следующие наборы директив.

1. Директивы для работы с таблицами, позволяющие производить удаление строк, содержащих нечисловые элементы, замену элементов, транспонирование, разделение / слияние таблиц, выборку строк / столбцов, сортировку строк / столбцов по заданному столбцу / строке, перестановку строк / столбцов и т. д.

2. Директивы для работы с матрицами, позволяющие производить их центрирование, нормирование, выполнять поэлементные операции, квантильное выравнивание, тест Мантеля и другие операции.

3. Директивы для вычисления мер сходства / различия – метрики Минковского, евклидовой метрики, расстояния Жаккара, расстояния Джукса – Кантора, расстояния Кимуры и т. д.

Часть вспомогательных директив требуют указания набора строк / столбцов. Для выполнения этой задачи разработан синтаксис, в котором можно выделить следующие важные особенности:

- возможность выбора среди неуникальных ключей;
- возможность использования как лексикографического диапазона, так и диапазона по абсолютным значениям;
- возможность указывать в качестве абсолютного номера как буквенный номер, используемый в MS Excel, так и десятичное число.

**Директивы для многомерного анализа данных**

Функционально из директив для многомерного анализа можно выделить директивы для понижения размерности данных с минимальными потерями значимой информации:

1) метод главных компонент;
2) метод главных координат;
3) неметрическое многомерное шкалирование.

Директивы для анализа взаимосвязи признаков:

1) дискриминантный анализ;
2) множественная линейная регрессия;
3) нейронные сети с обратным распространением ошибки.

Директивы для ПЛС-анализа [3]:

1) 2B-PLS-анализ;
2) PLS-регрессия.

Директивы для кластеризации:

1) алгоритм объединения;
2) алгоритм ближайшего соседа.

**Описание языка JACOBI 4**

Идея программного комплекса заключается в том, что пользователь на упрощенном скриптовом языке составляет программу, которая производит анализ данных.

Скриптовый язык не предполагает наличия у пользователя навыков программирования, поэтому реализованы только основные конструкции, такие как циклы, определения и вызов функций с параметром. Предполагается, что свой скрипт пользователь сможет писать в программе, аналогичной MicrosoftExcel. При этом каждая лексема записывается в отдельной ячейке. Предусмотрены лексемы следующих видов: переменная, присваивание, текстовая последовательность, название директивы, переменная цикла, границы цикла, элемент множества для цикла по множеству, конец цикла, слова, обозначающие начало и конец скрипта. Комментарии могут начинаться в любом месте. Пользователь должен сохранять скрипт в формате csv (разделитель – точка с запятой). Этот формат делает достаточно простым написание скрипта как в Excel, так и в любом текстовом редакторе, воспринимающем формат csv.

Для того чтобы большой скрипт можно было выполнять в несколько этапов, введено понятие блока, который ограничивается ключевыми словами «НАЧАЛО» и «КОНЕЦ». Если в скрипте имеет место неправильная последовательность этих двух команд, то программа выдаст сообщение об ошибке.

Реализованный язык предусматривает возможность определения переменной. Переменная может определяться только один раз. Исключение составляют переменные цикла. Это сделано для удобства пользователя, потому что переменные, в основном, используются для сокращения длинной строки.

Например, пользователь хочет проанализировать файл D:\myFolder\oneFolder\andMoreFolder\bigFileName.csv, но не хочет много раз писать такое длинное название. В этом случае он может до использования этого файла определить переменную:

Ирисы Фишера;=;D:\NSU\diploma\article\vestnik\iris.csv

После этого везде вместо слова «Ирисы Фишера» программа будет подставлять D:\NSU\diploma\article\vestnik\iris.csv.

Иначе говоря, синтаксис определения описывается так: «Имя_переменной ; = ; значение».

В языке представлена возможность для работы с циклами. Одним из способов задания цикла является цикл по переменной, пробегающей целые значения в указанном пользователем интервале, включая границы интервала (рис. 1).

| LOOP BEGIN | index | 1 | 7 |
|---|---|---|---|
| log | B_<<index>>.csv | B_<<index>>_LOG.csv | 2 |
| LOOP END | | | |

*Рис. 1.* Пример цикла переменной

Результатом работы такого цикла будет создание файлов B_1_LOG.csv,… B_7_LOG.csv, в которых записаны результаты работы алгоритма для файлов B_1.csv,… B_7.csv соответственно.

На этом же примере можно рассмотреть особенность использования индекса. Индекс может применяться и в качестве строки, в частности как название файла, и в качестве числа. Однако чаще всего требуется его использование именно в качестве строки.

В некоторых случаях возникает потребность реализовать некоторый набор действий определенное количество раз. Для этого предусмотрены циклы, в которых переменная не фигурирует, а задается лишь количество итераций. Пример такого цикла представлен на рис. 2.

В результате работы этого цикла в файле bootstrap_set.csv будет записано 20 бутстрепов выборки, содержащейся в input.csv.

| LOOP BEGIN | 20 | |
|---|---|---|
| bootstrep | input.csv | output.csv |
| addtofile | output.csv | bootstrep_set.csv |
| LOOP END | | |

*Рис. 2.* Пример цикла по количеству итераций

В реализованном языке также существует возможность задания списка строковых значений для переменной цикла. Такие циклы называются циклами по множеству (рис. 3).

| LOOP OVER LIST | index | file1.csv | u.csv | f.csv | file3.csv | f4.csv |
|---|---|---|---|---|---|---|
| log | index | LOG_<<index>> | 2 | | | |
| LOOP END | | | | | | |

*Рис. 3*. Пример цикла по множеству

Результатом работы этого цикла (рис. 3) будут файлы LOG_file1.csv, LOG_u.csv, LOG_f.csv, LOG_file3.csv, LOG_f4.csv, в которых записаны прологарифмированные по основанию 2 таблицы из файлов file1.csv, u.csv, f.csv, file3.csv, f4.csv соответственно.

Язык предусматривает возможность задания циклов любой глубины. При этом переменная внешнего цикла может использоваться как граница вложенного.

**Пользовательские скрипты**

Язык JACOBI 4 предусматривает возможность создания пользовательских подпрограмм с параметрами, которые представляют собой скрипты с описанием входных данных. Глубина вложенности скриптов не ограничена. Пакет JACOBI 4 планируется поставлять с различными наборами отлаженных скриптов-подпрограмм для включения в скрипты пользователей.

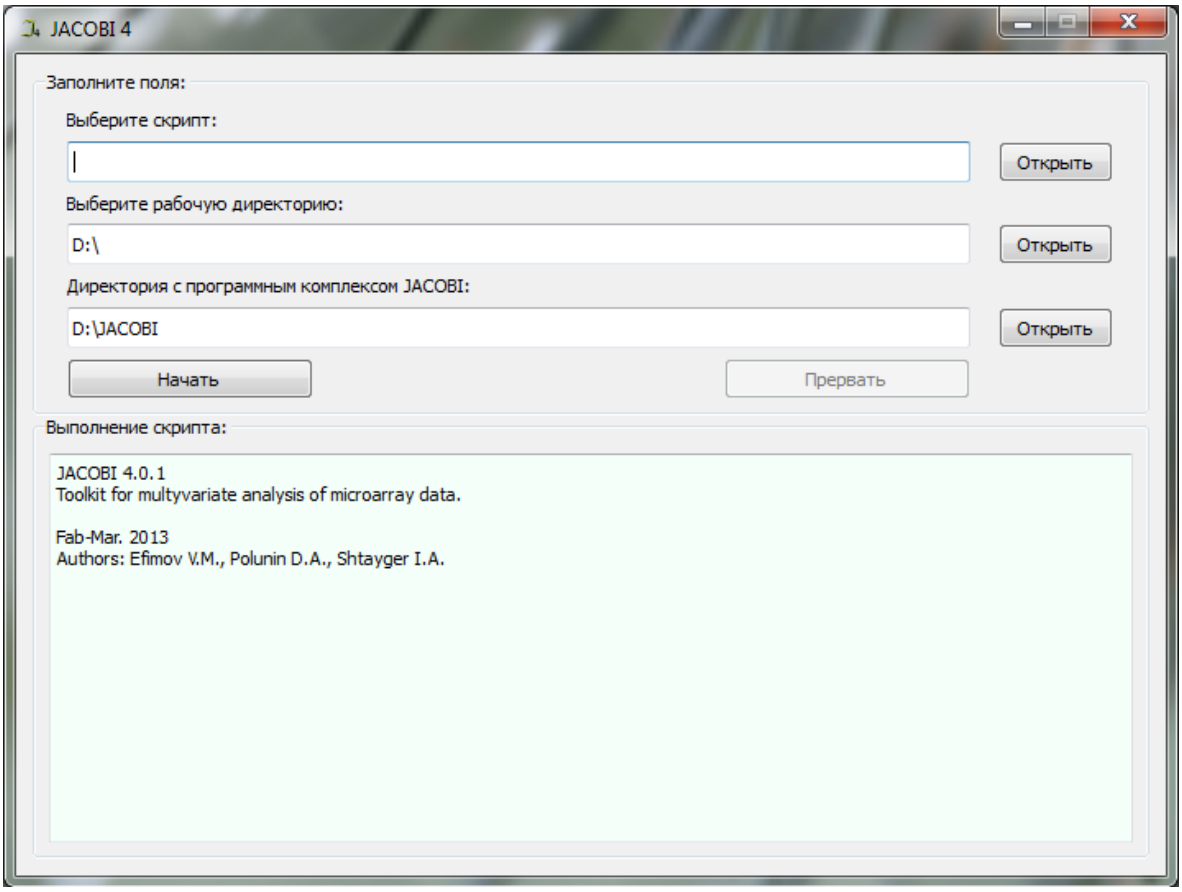


*Рис. 4.* Пользовательский интерфейс

*Пользовательский интерфейс.* Программный комплекс JACOBI 4 вплоть до версии 4.0.3 имеет графический интерфейс, ориентированный на пользователя, не обладающего большим опытом работы с ПК. Полноценный интерфейс на данный момент находится на стадии разработки. Тестовая бета-версия интерфейса будет доступна, начиная с версии 4.0.4 (рис. 4).

**Дополнительные возможности**

Одним из принципов пакета JACOBI 4 является возможность индивидуальных пользовательских настроек имен директив. Часто пользователи, работающие с распространенными пакетами для многомерного анализа, тратят много времени на запоминание в большинстве случаев неочевидных названий операций. Пакет JACOBI 4 предоставляет пользователю возможность самостоятельно изменять названия директив. Таким образом, пользователь будет работать с теми именами операций, к которым он привык, или с теми, которые ему кажутся простыми для запоминания. Такой принцип позволяет экономить много времени, которое обычно тратится на поиск информации в документации и запоминание.

Кроме того, пользователь может захотеть изменить набор директив в пакете. Как показывают результаты тестирования, пользователи предпочитают иметь в наборе только те директивы, которые им нужны. Это упрощает поиск, уменьшает время на написание скриптов и лучше воспринимается на подсознательном уровне.

Частным случаем изменения набора программ является добавление сторонних программ в пакет. Ни один из распространенных пакетов не предоставляет такой возможности, однако в ряде случаев ввиду специфичности задачи или новизны метода функция может быть не включена в функционал программы. В JACOBI 4 реализована так называемая обертка – программа, которая позволяет работать со сторонними программами как с собственными директивами пакета.

**Пример работы комплекса**

В качестве примера рассмотрим задачу определения направления максимальной изменчивости объектов по экспрессии генов и вычисление корреляции этих направлений со степенью выраженности болезни Хантингтона.

Для этих целей возьмем данные экспрессии более 22 000 генов по биочипу Affymetrix HG-U133B (B97) по трем отделам головного мозга caudate nucleus, frontal cortex, cerebellum у людей с разной степенью выраженности болезни Хантингтона [8].

Опишем последовательность действий, которую необходимо выполнить для достижения поставленной цели. Перед применением неметрического многомерного шкалирования необходимо произвести подготовку данных, затем посчитать корреляцию, таким образом, должны быть выполнены следующие действия:

1) оставить столбцы, соответствующие отделу мозга *caudate nucleus*;
2) удалить все строки, содержащие нечисловые значения;
3) прологарифмировать все значения матрицы данных;
4) произвести центрирование и нормирование;
5) вычислить матрицу расстояний между объектами с использованием метрики Минковского с параметром 5;
6) применить неметрическое многомерное шкалирование;
7) для каждого объекта сформировать числовые идентификаторы, показывающие степень выраженности болезни Хантингтона;
8) вычислить корреляцию между выраженностью болезни Хантингтона и полученными на шаге 6 переменными, выражающими экспрессию генов;
9) для осей с максимальной по модулю корреляцией построить график.

Пункты 1–8 выполнены в пакете, для этого составлен следующий скрипт:

| | | | | | |
|---|---|---|---|---|---|
| НАЧАЛО | | | | | |
| Копировать колонки | Brain-B97.csv | 1.CN.csv | ex.csv | [$1..$70] | |
| Удалить строки с нечисловыми значениями | 1.CN.csv | 2.numbers.only.csv | | | |
| Логарифмировать | 2.numbers.only.csv | 3.log2.csv | 2 | | |
| Центрировать | 3.log2.csv | 4.1.centre.csv | | | |
| Нормировать | 4.1.centre.csv | 4.2.normalize.csv | | | |
| Транспонировать | 4.2.normalize.csv | 4.3.transpose.csv | | | |
| Метрика минковского | 4.3.transpose.csv | 5.minkowski.csv | 5 | | |
| Неметрическое многомерное шкалирование | 5.minkowski.csv | 6.nmds.csv | 4 | 0.99 | 50 |
| Копировать колонки | Brain-N2.csv | 7.1.2columns.csv | ex.csv | [$9;$27] | |
| Копировать строки | 7.1.2columns.csv | 7.4.CN.rows.csv | ex.csv | [$1..$70] | |
| Подпрограмма | заменить ключи строк | 7.4.CN.rows.csv | 7.8.grades.csv | [$2] | |
| Транспонировать | 6.nmds.csv | 6.1.transposed.csv | | | |
| Транспонировать | 7.8.grades.csv | 7.9.transposed.csv | | | |
| Корреляция | 6.1.transposed.csv | 7.9.transposed.csv | 8.correlation.csv | | |
| КОНЕЦ | | | | | |

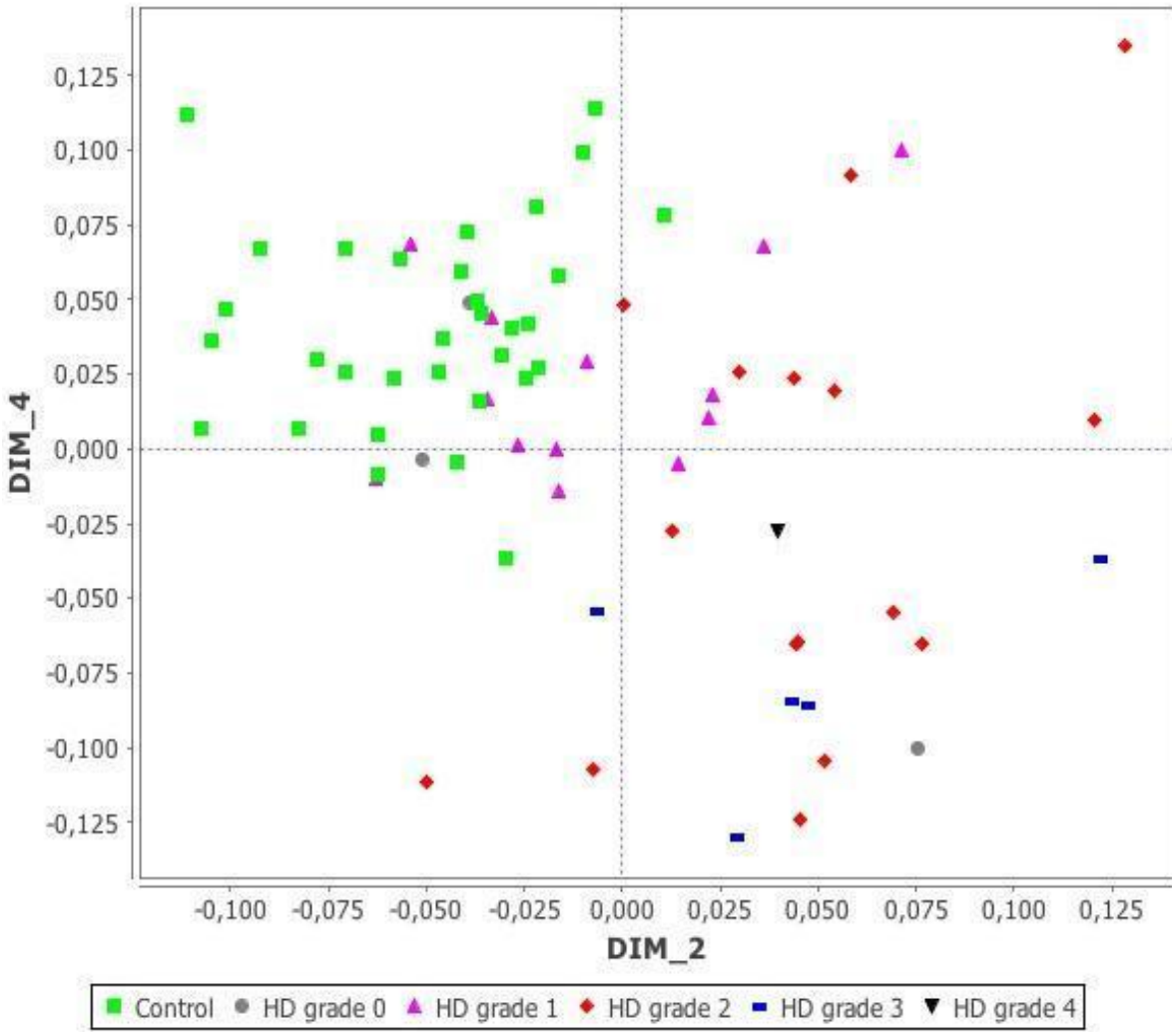


*Рис. 5.* Распределение проб, взятых у пациентов с разной степенью выраженности болезни Хантингтона

В результате работы скрипта был получен файл, содержащий оси и файл, включающий значение корреляции каждой оси со степенью выраженности болезни Хантингтона. Для построения графика были выбраны две оси с максимальной по модулю корреляцией: DIM_2 с коэффициентом корреляции, равным 0,71573, и DIM_4 с коэффициентом корреляции, равным минус 0,54624.

**Перспективы дальнейшего развития**

К сентябрю 2013 г. планируется реализация всех подсистем комплекса, механизма дельта-обновлений и создание базы данных для хранения документации. К 2014 г. – полный перенос пакета на кластер, расширение его функционала и дальнейшая адаптация пользовательского интерфейса.

**Заключение**

Реализован программный пакет JACOBI 4, позволяющий быстро и эффективно реализовывать однотипную обработку для множества входных данных. JACOBI 4 протестирован группой пользователей с различным опытом работы с ПК. В результате тестирования выявлен и устранен ряд ошибок. Кроме того, были учтены пожелания и предложения, которые составили новые требования для комплекса.

**Список литературы**

1. *Ефимов В. М.* О структуре пакета прикладных программ для обработки биологических данных // Науч.-техн. бюл. Новосибирск: ВАСХНИЛ, Сиб. отд-ние, СибНИИЗХим., 1983. Вып. 18. С. 34–38.

2. *Ефимов В. М., Речкин Д. В.* ЯКОБИ – входной язык пакетов прикладных программ статистической обработки биологических данных // Науч.-техн. бюл. Новосибирск: ВАСХНИЛ, Сиб. отд-ние, СибНИИЗХим., 1985. С. 12–17.

3. *Агбаш И. А., Полунин Д. А., Суслопаров Д. С., Штайгер И. А.* Программный комплекс для многомерного анализа микрочиповых данных. Многомерный PLS-анализ // Материалы L Междунар. науч. студ. конф. «Студент и научно-технический прогресс»: Информационные технологии / 13–19 апреля 2012 г., Новосиб. гос. ун-т. Новосибирск, 2012. 123 с.

4. *Штайгер И. А.* Программный комплекс для многомерного анализа микрочиповых данных: интерфейс, входной язык // Материалы L Междунар. науч. студ. конф. «Студент и научно-технический прогресс»: Информационные технологии / 13–19 апреля 2012 г., Новосиб. гос. ун-т. Новосибирск, 2012. 180 с.

5. *Агбаш И. А.* Программный комплекс для многомерного анализа микрочиповых данных: реализация функций базы данных и преобразования форматов данных // Материалы L Междунар. науч. студ. конф. «Студент и научно-технический прогресс»: Информационные технологии / 13–19 апреля 2012 г., Новосиб. гос. ун-т. Новосибирск, 2012. 64 с.

6. *Полунин Д. А.* Jacobi 4: базовые алгоритмы многомерного анализа, специальные биоинформатические методы // Материалы LI Междунар. науч. студ. конф. «Студент и научно-технический прогресс»: Информационные технологии / 12–18 преля 2013 г., Новосиб. гос. ун-т. Новосибирск, 2013 г. 54 с.

7. *Полунин Д. А.* «Программный комплекс для многомерного анализа микрочиповых данных: алгоритмы многомерного анализа» // Материалы L Междунар. науч. студ. конф. «Студент и научно-технический прогресс»: Информационные технологии / 13–19 апреля 2012 г., Новосиб. гос. ун-т. Новосибирск, 2012. 168 с.

8. *Borovecki F., Lovrecic L., Zhou J., Jeong H., Then F., Rosas H. D., Hersch S. M., Hogarth P., Bouzou B., Jensen R. V., Krainc D.* Genome-Wide Expression Profiling of Human Blood Reveals Biomarkers for Huntington's Disease // PNAS. 2005. № 102 (31). P. 23–28.

*Материал поступил в редколлегию 17.09.2013*

**D. A. Polunin, I. A. Shtayger, V. M. Efimov**

**JACOBI 4 SOFTWARE FOR MULTIVARIATE ANALYSIS OF MICROARRAY DATA**

JACOBI 4 software is a forth version of JACOBI package, which is designed for multivariate analysis of microarray data. JACOBI is intended to be an application which boosts a new technology created in Institute of Cytology and Genetics. It consists of a suite of opensource programs, which implement different algorithms for multivariate analysis. The suite of programs can be easily extended or changed by both user who has a lot of experience in PC and rather inexperienced user.

*Keywords*: software suite, multivariate analysis, script language, script, program, subprogram.

**References**

1. Efimov V. M. Structure of software suite for multivariate analysis of biological data. *Bulletin*, Novosibirsk, VASKhNIL, Siberian branch, SibRIAChem., 1983, issue 18, p. 34–38.

2. Efimov V. M., Rechkin D. V. JACOBI – input language for software suite for statistical processing of biological data. *Bulletin*. Novosibirsk, VASKhNIL, Siberian branch, SibRIAChem., 1985, no. 48, p. 12–17.

3. Agbash I. A., Polunin D. A., Susloparov D. S., Shtayger I. A. Software suite for multivariate analysis of microarray data. Multidimensional PLS analysis. *Proceedings of the 50th international students scientific conference ISSC-2012: Information technologie*s, April, 13–19, 2012, NSU, Novosibirsk, 2012, p. 123.

4. Shtayger I. A. Software suite for multivariate analysis of microarray data: interface, language. *Proceedings of the 50th international students scientific conference ISSC-2012: Information technologie*s, April, 13–19, 2012, NSU, Novosibirsk, 2012, p. 180.

6. Polunin D. A. Jacobi 4: core algorithms of multivariate analysis, special methods of bioinformatics. *Proceedings of the 51st international students scientific conference ISSC-2012: Information technologie*s, April, 12–18, 2013, NSU, Novosibirsk, 2013, p. 54.

7. Polunin D. A. Software suite for multivariate analysis of microarray data: algorithms of multivariate analysis. *Proceedings of the 50th international students scientific conference ISSC-2012: Information technologie*s, April, 13–19, 2012, NSU, Novosibirsk, 2012, p. 168.

8. Borovecki F., Lovrecic L., Zhou J., Jeong H., Then F., Rosas H. D., Hersch S. M., Hogarth P., Bouzou B., Jensen R. V., Krainc D. Genome-wide expression profiling of human blood reveals biomarkers for Huntington's disease. *PNAS*, 2005. no. 102 (31), p. 23–28.

5. Agbash I. A. JACOBI 4 software for multivariate analysis of microarray data: implementation of database functions and data format conversions. *Proceedings of the 50th international students scientific conference ISSC-2012: Information technologie*s, April, 13–19, 2012, NSU, Novosibirsk, 2012, p. 64.