УДК 577.21:004.02:004.94

ICGenonics: ПРОГРАММНЫЙ КОМПЛЕКС АНАЛИЗА СИМВОЛЬНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ ГЕНОМИКИ

© 2012 г. **Ю.Л. Орлов[1, 2], А.О. Брагин[1], И.В. Медведева[1], К.В. Гунбин[1], П.С. Деменков[1], О.В. Вишневский[1], В.Г. Левицкий[1], Д.Ю. Ощепков[1], Н.Л. Подколодный[1], Д.А. Афонников[1, 2], И. Гроссе[3], Н.А. Колчанов[1, 2, 4]**

[1] Федеральное государственное бюджетное учреждение науки Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия, e-mail: orlov@bionet.nsc.ru;
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия;
[3] Институт информатики, Университет Мартина Лютера, Халле, Германия;
[4] НИЦ «Курчатовский институт», Москва, Россия

Поступила в редакцию 10 июля 2012 г. Принята к публикации 10 августа 2012 г.

Экспериментальный образец программного комплекса анализа символьных последовательностей геномики (ЭОПК АСПГ) ICGenomics предназначен для хранения, передачи, обработки и анализа данных о символьных последовательностях, полученных в рамках теоретической и прикладной геномики с целью повышения качества вычислительной обработки биологических данных, используемых в биомедицине и биотехнологии. В комплексе реализованы новые оригинальные методы обработки первичных данных высокопроизводительного секвенирования, в том числе данных ChIP-seq, предсказания регуляторных участков генов в нуклеотидных последовательностях, модели расположения нуклеосом, структурно-функциональной аннотации белков, включая их аллергенные свойства и особенности эволюции. Рассмотрено применение комплекса к анализу последовательностей паразитического червя *O. felineus*, данным ChIP-seq по профилям связывания транскрипционных факторов в геномах мыши и человека. Комплекс доступен по адресу: http://www-bionet.sscc.ru/icgenomics.

Ключевые слова: геномика, программный комплекс, высокопроизводительное секвенирование, последовательности ДНК, анализ данных, ChIP-seq.

ВВЕДЕНИЕ

Программный комплекс ICGenomics предназначен для компьютерной поддержки исследований в геномике, молекулярной биологии, биотехнологии и биомедицине. Основное назначение – функциональная аннотация геномных последовательностей, получаемых в результате массового высокопроизводительного секвенирования на уровне нуклеотидных и аминокислотных последовательностей. Рабочее название – экспериментальный образец программного комплекса анализа символьных последовательностей геномики (ЭОПК АСПГ).

Важная технологическая проблема обработки и анализа данных высокопроизводительного геномного секвенирования требует разработки специализированных компьютерных средств. Развитие новых экспериментальных методов геномики, прежде всего, секвенирования, привело к стремительному росту объемов экспериментальных данных, «информационному взрыву».

Основная задача компьютерного анализа геномных данных состоит в их функциональной аннотации, интеграции результатов с молекулярно-биологическими информационными ресурсами. В связи с этим большую актуальность приобретает разработка информационно-компьютерных технологий автоматического анализа и функциональной аннотации геномных последовательностей. Для решения

задачи разработан ряд программ для извлечения и интеграции данных, а также визуального представления накопленной информации в форме геномных профилей, представленных на серверах крупнейших международных научных центров NCBI (http://www.ncbi.nlm.nih.gov/), UCSC Genome Browser (http://genome.ucsc.edu/), EBI (http://www.ebi.ac.uk/).

Важнейшим объектом анализа являются молекулярно-генетические системы, координирующие функцию геномов, генов, РНК, белков, генных и метаболических путей на различных иерархических уровнях жизни: клеточном, тканевом, органном, организменном, популяционном. Основным источником данных являются нуклеотидные последовательности, получаемые в результате массовых экспериментов высокопроизводительного секвенирования (Ivanisenko *et al.*, 2012). Несмотря на доступность компьютерных программ биоинформатики, в связи со все возрастающими объемами данных остается ряд направлений, важных для более детальной разработки. Можно выделить следующие направления исследования геномных последовательностей:

1. Разработка конвейерного подхода (pipeline) для первичной обработки, процессинга, картирования на референсный геном последовательностей, полученных в ходе масштабного параллельного секвенирования.

2. Функциональная аннотация геномных последовательностей (генома человека и модельных организмов) с целью разметки регуляторных районов, сайтов формирования нуклеосом и определения структуры хроматина. Сюда входят аннотация потенциальных микроРНК и анализ промоторных последовательностей генов.

3. Разработка программ для разметки функциональных сайтов белков, определения свойств белковых фрагментов, кодируемых в нуклеотидных последовательностях, оценки потенциальной аллергенности кодируемых белков с использованием оригинальных баз данных и методов.

4. Сравнение функциональных свойств вновь секвенированных генов различных организмов. Исследование адаптивного режима эволюции на уровне отдельных семейств генов и на геномном уровне.

Решение перечисленных задач необходимо для обеспечения технической поддержки геномных исследований. Соответствующие технические средства реализованы в разработанном программном комплексе. Особое внимание было уделено оригинальным методам, не повторяющим стандартные алгоритмы для уже достаточно рутинных задач, таких, как выделение кодирующей последовательности или предсказание сайтов связывания транскрипционных факторов (ССТФ) только по нуклеотидной последовательности (с помощью весовых матриц), стандартные решения для которых представлены на серверах NCBI, UCSC, EBI.

Конкретная задача компьютерного анализа геномных последовательностей включала реализацию следующих независимых конкретных процедур, объединяемых общими типами данных:

– Обработка последовательностей ДНК из геномных фрагментов, полученных с помощью установок геномного секвенирования нового поколения.

– Функциональная аннотация геномных нуклеотидных последовательностей с возможностями аннотации нуклеосом, поиска экзонов, поиска промоторов генов миРНК.

– Предсказание аллергенности белков по их структурным и функциональным свойствам на основе метода функциональной аннотации пространственных структур белков, предсказание функциональных сайтов в пространственных структурах белков и предсказание специфической активности белков по их первичной и пространственной структуре.

– Реализованная в виде конвейера обработки данных процедура анализа режимов эволюции белок-кодирующих генов с возможностями реконструкции эволюционной истории белков на основе предсказания ортологов в секвенированных геномах, филогенетического анализа, а также изучения режимов отбора.

Программный комплекс ICGenomics (http://www-bionet.sscc.ru/icgenomics) был реализован и протестирован на вычислительном оборудовании ЦКП «Биоинформатика» СО РАН.

МАТЕРИАЛЫ И МЕТОДЫ

Программный комплекс ICGenomics позволяет выполнять следующие логически различные функции:

– процессинг (обработку) протяженных последовательностей нуклеотидов из данных секвенирования, полученных с помощью установок секвенирования нового поколения, в том числе: процессинг данных секвенирования платформ 454 и Illumina, процессинг данных секвенирования платформы SOLiD и обработку полногеномных профилей ChIP-seq, включая выделение пиков и предсказание ССТФ;

– аннотацию геномных нуклеотидных последовательностей, включая разметку положения нуклеосом на основе вейвлет-преобразования полногеномных профилей предсказания и распознавание сайтов формирования нуклеосом с помощью данных полногеномного секвенирования линкерной ДНК; поиск экзонов во вновь секвенированных последовательностях; поиск промоторов генов миРНК в нуклеотидных последовательностях на основе специфичных структурных мотивов;

– предсказание аллергенности белков по их структурным и функциональным свойствам на основе метода функциональной аннотации пространственных структур белков, в том числе предсказания функциональных сайтов в пространственных структурах белков;

– исследование режимов эволюции белок-кодирующих генов, включая реконструкцию эволюционной истории белков на основе предсказания ортологов в секвенированных геномах, филогенетический анализ и исследование режимов эволюционного отбора.

Каждая из перечисленных выше функций реализована в соответствующем программном компоненте (рис. 1).

Программный комплекс состоит из модуля управления (программной компоненты ICGenomics-Web и управляющей программы ICGenomics-start) и 4 программных компонент: ICGenomics-Processing, ICGenomics-Genome Annotation, ICGenomics-Allergen и ICGenomics-Evolution (рис. 1).

Общий интерфейс представлен на рис. 2. Рассмотрим компоненты более подробно:

1. ICGenomics-Processing – программный компонент, осуществляющий обработку последовательностей ДНК из фрагментов, полученных с помощью установок геномного секвенирования нового поколения, обладающий функционалом процессинга исходных («сырых») данных секвенирования, обработки полногеномных профилей ChIP-seq, выделения пиков и предсказания ССТФ.

2. ICGenomics-GenomeAnnotation – программный компонент функциональной аннотации геномных нуклеотидных последовательностей, обладающий возможностями:

– функциональной аннотации нуклеосом;
– поиска экзонов;
– поиска промоторов генов миРНК (рис. 3).

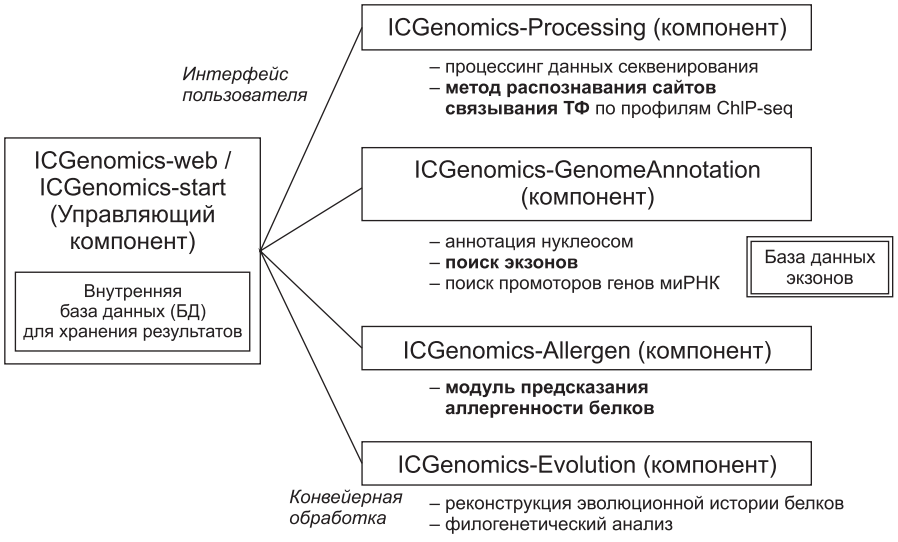

Рис. 1. Структура программного комплекса ICGenomics.

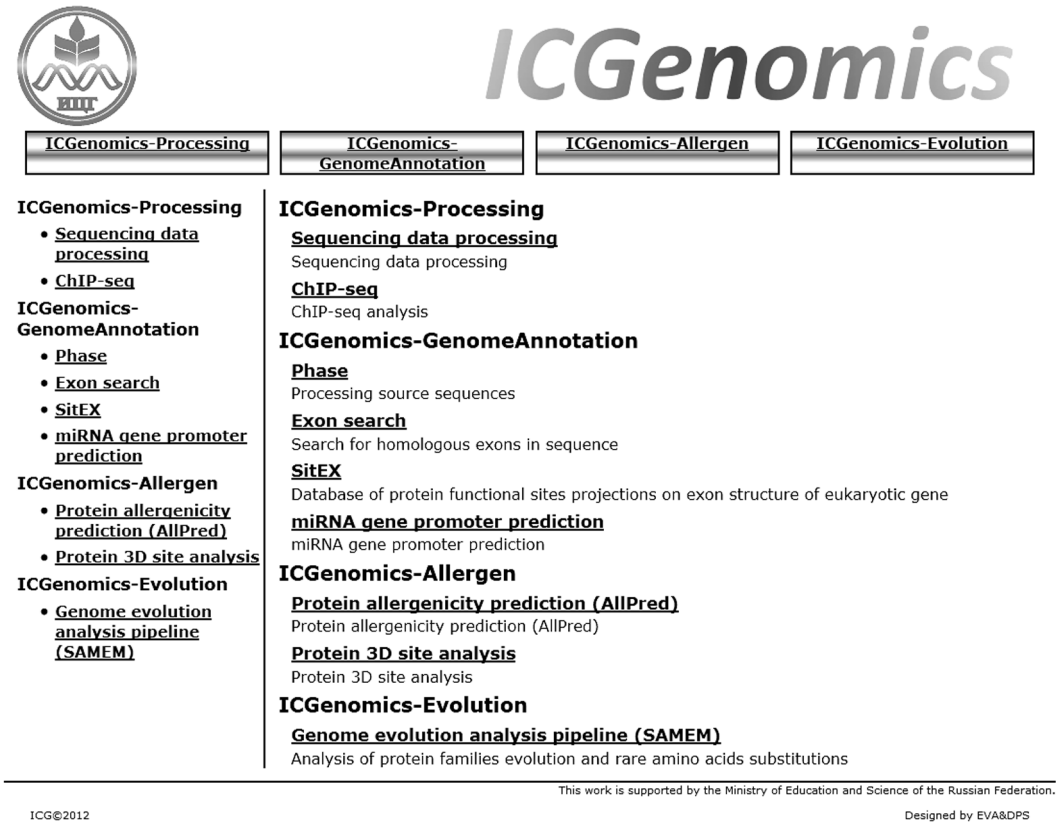

Рис. 2. Пример интерфейса управляющего модуля, содержащего функциональные компоненты.

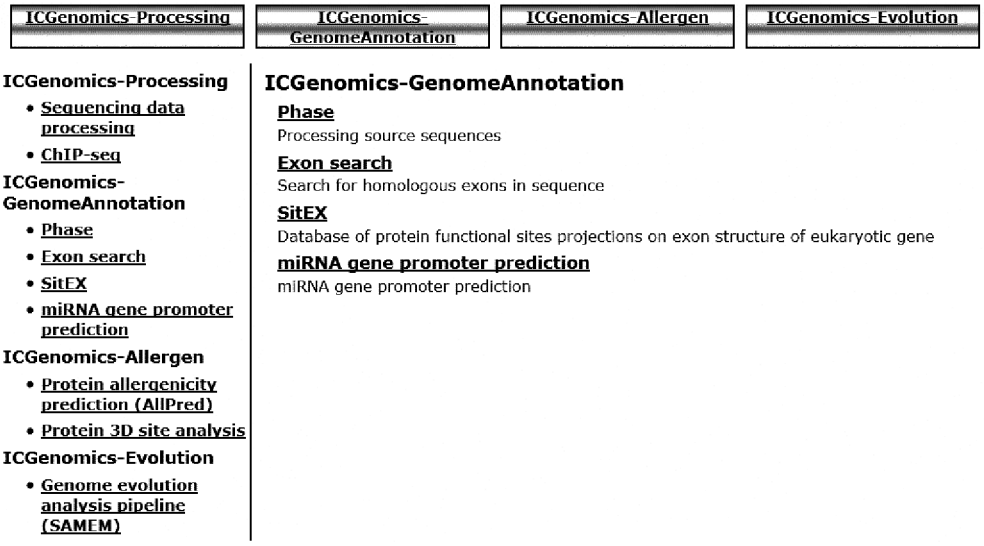

Рис. 3. Пример интерфейса ICGenomics-GenomeAnnotation.

3. ICGenomics-Allergen – программный компонент предсказания аллергенности белков по их структурным и функциональным свойствам.

4. ICGenomics-Evolution – программный компонент исследования режимов эволюции белок-кодирующих генов, обладающий функционалом: реконструкции эволюционной истории белков на основе предсказаний ортологов в секвенированных геномах; филогенетического анализа и исследования режимов отбора. Компонент реализован в виде конвейера обработки данных.

Входными данными для системы служат файлы нуклеотидных и аминокислотных последовательностей в формате FASTA, а также данные секвенирования в форматах платформ секвенирования Illumina, SOLiD. Возможно использование форматов геномных профилей bed (геномные координаты), wig (численный профиль). В комплексе используются базы данных SiteEx (Medvedeva *et al*., 2012), и PDBSite (Ivanisenko *et al*., 2005), содержащие скомпилированную ранее информацию об экзонах и пространственных сайтах белков.

Компонент ICGenomics-Processing включает в себя модули процессинга данных, в том числе конвертации форматов и фильтрации сигнала секвенирования ДНК, процессинга данных секвенирования платформ 454 и Illumina (исходные форматы fastq, qseq), процессинга данных секвенирования платформы SOLiD (в цветовой кодировке color-space – исходный формат csfasta) и конвейерной обработки задач картирования данных SOLiD. Этот компонент (модуль ChIP-seq pipeline) также выполняет обработку полногеномных профилей ChIP-seq, выделение пиков профиля и предсказание ССТФ в геноме.

Типичные задачи, которые решаются на этапе предобработки – преобразование данных, полученных в результате эксперимента, в стандартные форматы; анализ качества последовательностей и фильтрация по качеству; подготовка результата по проведенным операциям. Метод распознавания реализован в программе ChIP-seq pipeline и предназначен для конвейерной обработки выходных данных эксперимента по массовому секвенированию функциональных сайтов. Массовость означает полногеномный характер анализа и большие объемы данных. В качестве функциональных сайтов исследуются ССТФ различных типов. Программный комплекс преследует две основные задачи: а) обработку данных и привязку их к геномным картам; б) верификацию обнаруженных геномных локусов с помощью различных биоинформатических средств (программ распознавания ССТФ). Подобный подход позволяет: а) исключить из рассмотрения ошибки и артефакты, присутствующие в данных эксперимента ChIP-seq; б) правильно настроить параметры на различных этапах обработки данных (картирование на полный геном, выбор минимального числа прочтений сайта в геномном локусе и т. д.); в) получить в итоге исчерпывающий список генов-мишеней исследуемого ССТФ для полного генома (Lee *et al*., 2011).

В качестве программы для картирования прочтений (первичных данных эксперимента ChIP-seq) использовался рекомендованный производителем SOLiD™ BioScope™ Software с настройками по умолчанию. Далее в соответствии с рекомендациями производителя с помощью этого же программного обеспечения (SOLiD™ BioScope™ Software) производилась конвертация выходного формата файла с результатами картирования (формат «.ma» в формат «.bam») для последующей подачи на вход программы MACS (Zhang *et al*., 2008). MACS предназначена для проведения процедуры поиска пиков ChIP-seq (ChIP-seq peak calling) и является одной из самых широко используемых программ, кроме того, обладает наибольшей точностью в определении локализации сайта связывания (Malone *et al*., 2011).

Результатом работы программы является полногеномный профиль в формате wig, который представляет собой список пар «позиция»–«покрытие». «Позиция» – хромосомная локализация, включает в себя номер хромосомы и собственно позицию от начала хромосомы. «Покрытие» – число прочтений – это число зафиксированных взаимодействий исследуемого белка (ТФ) с ДНК в рассматриваемой хромосомной локализации. Далее могут быть определены нуклеотидные последовательности, содержащие ССТФ, проанализированы частоты олигонуклеотидов с помощью разработанных ранее программ (Putta *et al*., 2011).

Используемые для секвенирования фрагментов ДНК технологии компаний Illumina и ABI SOLiD характеризуются особенностями, связанными с проведением экспериментальных процедур, что отражено в форматах входных данных используемых ICGenomics. Технология компании Illumina (Solexa) (http://www.illumina.com) использует оптическое сканирование флюоресценции меченых нуклеотидов в клонированных колониях молекул ДНК на твердой поверхности, в то время как технология секвенирования ABI (Applied Biosystems) SOLiD (Sequencing by Oligonucleotide Ligation) использует лигирование и, соответственно, кодировку

по двум нуклеотидам. Для процессинга данных секвенирования реализована возможность использования следующих форматов геномных данных: FASTA, fastq, clustal.

Используя тот же формат FASTA, компонент ICGenomics-GenomeAnnotation функциональной аннотации геномных нуклеотидных последовательностей (рис. 3) решает задачи:

– функциональной аннотации нуклеосом (включая применение вейвлет-преобразования для анализа полногеномных профилей предсказания сайтов формирования нуклеосом и распознавания сайтов формирования нуклеосом с помощью данных полногеномного секвенирования линкерной ДНК);

– поиска экзонов во вновь секвенированных последовательностях для более подробной аннотации генов и кодируемых ими белков, а также входящих в их состав доменов на основе базы данных последовательностей экзонов и структур полипептидов, кодируемых единственным экзоном;

– поиска промоторов генов миРНК в нуклеотидных последовательностях на основе специфичных структурных мотивов.

Вызов отдельных модулей выполняется из общего интерфейса пошагово. На рис. 4 приведен пример вызова программы Phase предсказания положения нуклеосом в нуклеотидной последовательности. Программа Phase успешно применялась для анализа генома дрожжей и сравнения эффективности транскрипции генов в зависимости от предсказанной локализации нуклеосом в промоторах генов (Matushkin *et al*., 2012).

Таким же образом могут вызываться модуль анализа экзонов, в том числе база данных SiteEx (Medvedeva *et al*., 2012) и модуль предсказания промоторов генов миРНК (Vishnevsky *et al*., 2010).

Программный компонент ICGenomics-Allergen предсказания аллергенности белков по их структурным и функциональным свойствам выполняет предсказание аллергенности белков (пептидов) с использованием конформационных пептидов (Bragin *et al*., 2012). Кроме того, модуль может передавать данные функциональных сайтов в пространственных структурах белков (рис. 5). Программа вычисляет значения аллергенности по заданной последовательности

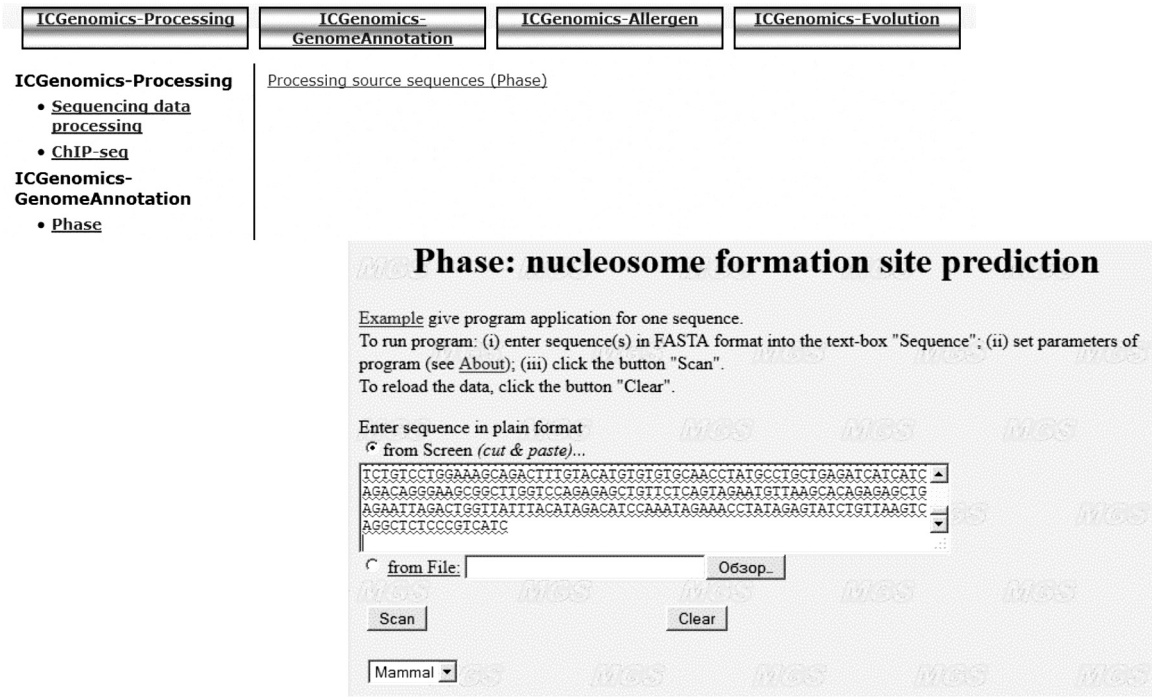

Рис. 4. Пример вызова модуля предсказания положения нуклеосом из компонента ICGenomics-GenomeAnnotation (верхняя панель) и интерфейс вызванной программы Phase предсказания положения нуклеосом по нуклеотидной последовательности (нижняя панель).

пептида. Результатом работы является числовое значение аллергенности и текстовое описание. Те же последовательности могут быть переданы на анализ гомологии с последовательностями экзонов в соответствующем модуле и на сравнительный анализ семейства белков, приводящих к появлению свойств аллергенности.

Точность предсказания аллергенности белков разработанным модулем сравнивалась с точностью предсказания стандартных программ (Bragin *et al*., 2012). Точность метода была оценена на выборке белков-аллергенов из работы H.C. Muh и соавт. (Muh *et al*., 2009), создавших программу AllerHunter (http://tiger.dbs.nus.edu.sg/AllerHunter/). Использование одной только программы BLAST (белок считался аллергеном, если значение E-value сходства его последовательности с известными аллергенами было ниже 10^{-21}) позволяет точно предсказать аллергенность только у 84 % белков. В то время как метод, применяющий поиск гомологов при помощи BLAST и поиск пептидов в анализируемых белках, правильно предсказывает 92 % белков-аллергенов из этой же выборки.

Программный компонент ICGenomics-Evolution исследования режимов эволюции белок-кодирующих генов выполняет задачи реконструкции эволюционной истории белков на основе предсказания ортологов в секвенированных геномах и филогенетического анализа и исследования режимов отбора. Компонент реализован в виде конвейера обработки данных (рис. 6).

Методы анализа режимов эволюции, входящие в данный программный компонент, были успешно использованы в работах Gunbin с соавт. (2010, 2011); Гунбин и др. (2011).

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Разработан комплекс ICGenomics, использующий ряд уникальных модулей. Программа позволяет выполнять следующие функции обработки и анализа геномных последовательностей:

– процессинг (обработку) протяженных последовательностей нуклеотидов из данных секвенирования, полученных с помощью установок секвенирования нового поколения, в том числе обработку полногеномных профилей ChIP-seq;

– аннотацию геномных нуклеотидных последовательностей, включая: разметку положения нуклеосом на основе вейвлет-преобразования полногеномных профилей предсказания, сайтов формирования нуклеосом; поиск экзонов во вновь секвенированных последовательностях; поиск промоторов генов миРНК в нуклеотидных последовательностях;

– предсказание аллергенности белков по их структурным и функциональным свойствам;

– исследование режимов эволюции белок-кодирующих генов, включая реконструкцию эволюционной истории белков на основе на предсказания ортологов в секвенированных геномах, филогенетический анализ и исследование режимов эволюционного отбора.

Реализованные в проекте авторские методы уникальны, что подтверждено регистрационными свидетельствами на программы, входящие в компоненты предсказания аллергенности и

Рис. 5. Интерфейс программы ICGenomics-Allergen предсказания аллергенности белков.

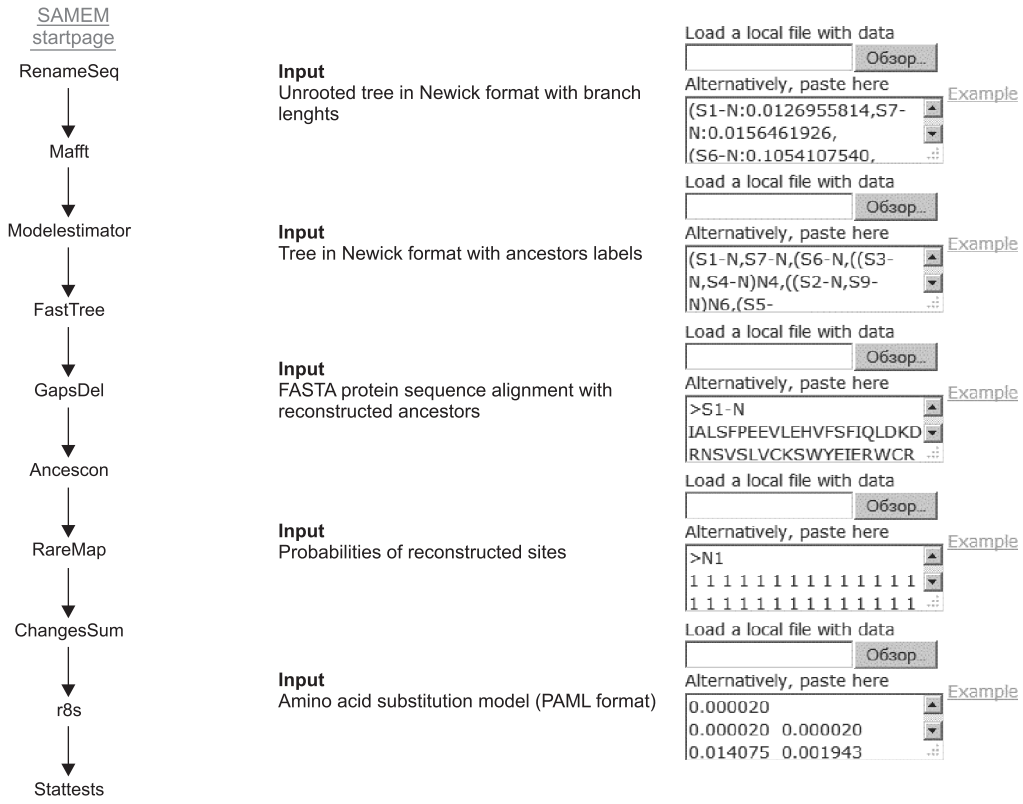

Рис. 6. Схема конвейера (левая панель) и выбор основных параметров (правая панель) в модуле исследования режимов эволюции ICGenomics-Evolution (SAMEM).

анализа режимов эволюции белков. Комплекс применялся к анализу геномных последовательностей паразитического червя *O. felineus* и к данным ChIP-seq по профилям связывания транскрипционных факторов в геномах мыши и человека (для фактора FoxA) (Левицкий и др., 2011).

Были исследованы три образца ткани *O. felineus* (стадии: марита без яиц, марита с яйцами, метацеркарий), а также препараты тканей *O. viverini* и *C. sinensis* на стадии марит. Картирование осуществляли на геном паразитического плоского червя шистосомы (*Schistosoma japonicum*). Шистосома – паразитический червь, который поражает кровеносную систему организма. Это ближайший родственный вид, геномная последовательность которого расшифрована практически полностью.

Для генома *S. japonicum* ранее была проведена функциональная аннотация генома и идентифицированы 55 последовательностей микроРНК. Эти гены принимают участие в регуляции стадий развития организма червя. Анализ локализации нуклеотидных фрагментов трех организмов позволил нам установить, что из этих 55 микроРНК 17 представлены и в геномах *O. felineus*, *O. viverini* и *C. sinensis*. При этом число картированных последовательностей для этих генов зависит как от стадии развития организма, так и от вида.

В ЭОПК АСПГ использовались разработанные в ИЦиГ СО РАН методы предсказания аллергенности белков по аминокислотным последовательностям (конформационным пептидам), предсказания позиций нуклеосом, предсказания сайтов. Конструктивными характеристиками разработанных методов являются возможности обрабатывать большие объемы данных секвенирования и возможность обмена данными в FASTA формате.

БЛАГОДАРНОСТИ

Авторы выражают благодарность В.А. Иванисенко и М.П. Пономаренко за научную дискуссию по данному проекту.

Разработка программного комплекса поддержана госконтрактом Минобрнауки РФ № 07.514.11.4003. Тестирование выполнялось на суперкомпьютерном кластере ССКЦ СО РАН, ЦКП «Биоинформатика».

ЛИТЕРАТУРА

Гунбин К.В., Суслов В.В., Афонников Д.А. Генетическая основа макроэволюционных преобразований: исследование режимов молекулярной эволюции ортологичных белков позвоночных и беспозвоночных // Тр. Междунар. конф. «Современные проблемы математики, информатики и биоинформатики», посвященной 100-летию со дня рождения чл.-корр. А.А.Ляпунова. 11–14 октября 2011 г. Новосибирск, Россия. 2011. ПП. 4.7. С. 52–53.

Левицкий В.Г., Ощепков Д.Ю., Ершов Н.И. и др. Разработка методов распознавания сайтов связывания транскрипционных факторов FoxA, их экспериментальная верификация и использование для анализа данных массовой иммунопреципитации хроматина // Докл. АН. 2011. Т. 436. № 3. С. 417–421.

Bragin A.O., Demenkov P.S., Kolchanov N.A., Ivanisenko V.A. Accuracy of protein allergenicity prediction can be improved by taking into account data on allergenic protein discontinuous peptides // J. Biomol. Struct. Dyn. 2012. Jul. 18. [Epub ahead of print]

Gunbin K.V., Genaev M. A., Afonnikov D. A., Kolchanov N.A. A computer system for the analysis of molecular evolution modes of protein-encoding genes (SAMEM): The relationship between molecular evolution and phenotypic traits // Mosc. Univ. Biol. Sci. Bull. 2010. V. 65. No. 4. P. 142–144.

Gunbin K.V., Suslov V.V., Turnaev I.I. *et al*. Molecular evolution of cyclin proteins in animals and fungi // BMC Evol. Biol. 2011. V. 11. P. 224.

Ivanisenko V.A., Demenkov P.S., Pintus S.S. *et al.* Computer analysis of metagenomic data-prediction of quantitative value of specific activity of proteins // Dokl. Biochem. Biophys. 2012. V. 443. P. 76–80.

Ivanisenko V.A., Pintus S.S., Grigorovich D.A., Kolchanov N.A. PDBSite: a database of the 3D structure of protein functional sites // Nucl. Acids Res. 2005. V. 33. Database, P. 183–187.

Lee K.L., Lim S.K., Orlov Y.L. *et al*. Graded Nodal/Activin signaling titrates conversion of quantitative phospho-Smad2 levels into qualitative embryonic stem cell fate decisions // PLoS Genet. 2011. V. 7. No. 6. e1002130.

Malone B.M., Tan F., Bridges S.M., Peng Z. Comparison of four ChIP-Seq analytical algorithms using rice endosperm H3K27 trimethylation profiling data // PLoS One. 2011. V. 6. No. 9. e25260.

Matushkin Y.G., Levitsky V.G., Orlov Y.L. *et al*. Translation efficiency in yeasts correlates with nucleosome formation in promoters // J. Biomol. Struct. Dyn. 2012. Jul. 18. [Epub ahead of print].

Medvedeva I., Demenkov P., Kolchanov N., Ivanisenko V. SitEx: a computer system for analysis of projections of protein functional sites on eukaryotic genes // Nucl. Acids Res. 2012. V. 40 (Database issue). P. 278–83.

Muh H.C., Tong J.C., Tammi M.T. AllerHunter: a SVM-pairwise system for assessment of allergenicity and allergic cross-reactivity in proteins // PLoS One. 2009. V. 4. No. 6. e5861.

Putta P., Orlov Yu.L., Podkolodnyy N.L., Mitra C.K. Relatively conserved common short sequences in transcription factor binding sites and miRNA // Вавилов. журн. генет. и селекции. 2011. Т. 15. № 4. С. 750–756.

Vishnevsky O.V., Gunbin K.V., Bocharnikov A.V., Berezikov E.V. Analysis of degenerate motifs in the promoters of miRNA genes expressed in different mammalian tissues // Mosc. Univ. Biol. Sci. Bull. 2010. V. 65. No. 4. P. 193–195.

Zhang Y., Liu T., Meyer C.A. *et al*. Model-based Analysis of ChIP-Seq (MACS) // Genome Biol. 2008. V. 9. No. 9. R137.

ICGenomics: A PROGRAM COMPLEX FOR ANALYSIS OF SYMBOL SEQUENCES IN GENOMICS

Y.L. Orlov[1, 2], A.O. Bragin[1], I.V. Medvedeva[1], K.V. Gunbin[1], P.S. Demenkov[1], O.V. Vishnevsky[1], V.G. Levitsky[1], D.Y. Oshchepkov[1], N.L. Podkolodnyy[1], D.A. Afonnikov[1, 2], I. Grosse[3], N.A. Kolchanov[1, 2, 4]

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia,
e-mail: orlov@bionet.nsc.ru;
[2] Novosibirsk National Research State University, Novosibirsk, Russia;
[3] Institute of Computer Science, Martin Luther University, Halle, Germany;
[4] National Research Centre «Kurchatov Institute», Moscow, Russia

Summary

The pilot program complex for analysis of symbol sequences in genomics, ICGenomics, has been designed for storage, mining, and analysis of sequences related to theoretical and applied genomics. ICGenomics enables wet-lab biologists to perform high-quality processing of data in the fields of genomics, biomedicine, and biotechnology. ICGenomics implements both conventional and modern methods for processing, analyzing, and visualizing sequence data. They include novel methods of the processing of initial high-throughput sequencing data. Examples are: ChIP-seq analysis; functional annotation of gene regulatory regions in nucleotide and amino acid sequences; prediction of nucleosome positioning; and structural and functional annotation of proteins, including their allergenicity and evolution features. Application of ICGenomics to the analysis of genomic sequences of the parasite *Opisthorchis felineus* and to ChIP-seq data on the mouse and human is considered. The system is available at http://www-bionet.sscc.ru/icgenomics.

Key words: genomics, program complex, high-throughput sequencing, nucleotide sequences, data analysis, ChIP-seq.