электронный журнал
МОЛОДЕЖНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
Издатель ФГБОУ ВПО "МГТУ им. Н.Э. Баумана". Эл No. ФС77-51038.

УДК 004.942

Биоинформатика и биоинформационные системы: назначение, функции, обзор и перспективы развития

***Курникова А.О.**, студент*
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана,
кафедра «Системы обработки информации и управления»

Научный руководитель: Самохвалов Э.Н., к.т.н., профессор,
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана
Gapyu@bmstu.ru

Введение

Вторая половина XX века и начало XXI века ознаменовались возникновением и развитием новых наук и научных направлений. Зачастую эти направления образуются на стыке традиционных наук, обогащая их новыми знаниями и создавая новые направления для исследований. В ряде случаев такие новые науки сочетают в себе, казалось бы, не сочетаемые с точки зрения нормального человека вещи. Биоэнергетика, геополитика, политтехнология, криохирургия – все это науки или научные направления, возникшие на стыке традиционных наук с многолетней историей. И список можно продолжать.

Страсть человечества к познанию неизвестного является основной движущей силой при возникновении новых направлений. С другой стороны, бурное развитие компьютерной техники и огромные объемы накопленных экспериментальных данных позволяют подходить к изучению, казалось бы, уже давно изученных явлений и процессов по-новому.

Вот на стыке двух таких наук как биология и информатика возникла новая наука – биоинформатика.

Биоинформатика

В общем случае под биоинформатикой обычно понимают использование компьютеров для решения биологических задач. Когда-то главным орудием биологов были сачок и лупа. Потом — микроскоп и пробирки. Бурное развитие молекулярной биологии и генетики в конце XX – начале XXI веков привело к накоплению огромного массива экспериментальных данных, в первую очередь последовательностей ДНК, РНК и белков, цифровых биологических изображений и структур сигнальных сетей, хранение и

анализ которых не возможен без применения соответствующего программного обеспечения. Хотя и раньше информационные технологии использовались биологами, например, для статистической обработки полученных данных, именно бум молекулярной биологии вызвал у специалистов-биологов потребность в специализированных инструментах для решения конкретных задач по обработке биологической информации. Как раз с этим связано возникновение биоинформатики как самостоятельной области науки. В биоинформатике также используются знания, полученные из других наук – физики, химии, медицины, физиологии и др. Современная биоинформатика возникла в конце 70-х годов XX в. с появлением эффективных методов расшифровки последовательностей ДНК.

Основные задачи биоинформатики:

- описание генных сетей,

- разработка и внедрение новых лекарственных препаратов с заданными свойствами,

- разработка компьютерных моделей процессов, происходящих в организме человека.

Генные сети

Гены в клетках организма могут взаимодействовать друг с другом посредством своих продуктов - белков. Например, регуляторные белки способны связываться с определенными участками ДНК, и, таким образом, один ген может включить или выключить другой. Благодаря подобному взаимодействию образуется генная сеть, охватывающая значительное количество генов (от десятков до сотен), которые координируют свою деятельность и контролируют выполнение определенных функций в организме. Выяснение механизмов функционирования генных сетей представляет принципиально важную задачу, ведь именно они определяют внешние признаки организма и наследственные заболевания. Полная и ясная картина взаимодействия генов откроет новые возможности для генной диагностики и генной терапии.Огромную роль в развитии технологии чтения генетической информации сыграло развитие компьютерной техники и вычислительных методов. Неудивительно, что интенсивное развитие биоинформатики совпало по времени с победным шествием компьютерных технологий. Это лишний раз подтвердило, что глубина научного знания чрезвычайно сильно зависит от технических возможностей.

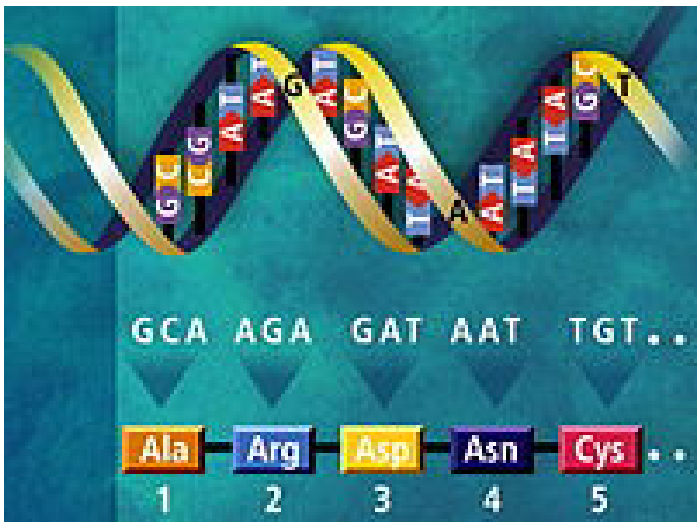

Рис.1. Генетический код

В 2002 году закончена расшифровка генома человека. Следующая важнейшая задача - расшифровать протеом. Этот термин образован от слова "протеин" (по аналогии с геномом) и означает совокупность белков, которые функционируют в организме. Конечно, получение "белкового портрета" организма потребует времени, но в принципе эта задача вполне решаема. Общий объем накопленной информации сейчас таков, что на первый план выходит системная биология, цель которой - не просто объединить достижения, полученные различными методами, но интегрировать знания и перевести их на качественно новый уровень.

Другой важнейшей вехой в развитии биоинформатики стало возникновение и повсеместное распространение технологий Всемирной сети - Интернета. Большое число разнообразных баз данных и программных инструментов теперь доступны через Интернет. Биоинформатика, пожалуй, является одной из тех областей науки, которые в очень большой степени зависимы от Интернета и успешно развиваются благодаря Интернету. Следует подчеркнуть, что очень важное для биологии и медицины политическое решение об открытости сложнейшего биологического текста современности - генома человека - сделало эту информацию по-настоящему доступной для ученых всего мира лишьблагодаря Интернету.

Новые лекарственные препараты

Одна из самых перспективных и быстро развивающихся областей биоинформатики - конструирование лекарств направленного действия. Действие таких препаратов нацелено на центры связывания конкретного белка в организме возбудителя болезни. При этом аналогичные белки человека не подвергаются изменениям, а значит, нет и побочных эффектов. Создание лекарства направленного действия требует знаний о трехмерной структуре белка-мишени, так как точное пространственное соответствие играет ключевую

роль. Таким образом, идеальное лекарственное средство должно связываться только с целевым белком и вызывать строго определенное изменение его свойств.

Не секрет, что процесс разработки нового лекарственного препарата достаточно долог и требует проведения множества испытаний перед тем, как быть внедренным в клиническую практику. При использовании средств биоинформатики возможно смоделировать процесс воздействия вещества на организм человека и получить информацию о том, насколько действенно то или иное лекарство и какие побочные явления оно может вызывать.

Биоинформационные системы

В настоящее время в мире существуют сотни биоинформационных систем. Они различаются по сложности, области применения, объему накопленной информации и др. Однако наряду с многообразием и специфическими особенностями можно выделить основные элементы структуры любой биоинформационной системы.

На рис.2 Представлена структура типичной БИС. Она состоит из метабазы, модели и программной оболочки. В зависимости от назначения БИС те или иные ее части могут отсутствовать, в то же время могут быть и другие составляющие, не показанные ниже.

База данных (Метабаза)

База данных является ядром, основой для всей биоинформационной системы (БИС). В ней содержится информация о генах, протеинах, метаболитах, веществах и т.д. Данная информация есть результат многолетних исследований и экспериментов. Как правило, такие базы данных формируются вручную целой командой специалистов и постоянно пополняется по мере появления новых данных. Этот этап является самым первым при создании БИС. От того, насколько качественно будут выполнены работы по сбору и систематизации информации, зависит то, насколько создаваемая БИС будет удовлетворять требованиям заказчика. Для формирования метабазы как правило используется программная платформа системы управления базами данных (СУБД), которая может работать с большими объемами информации с использованием дополнительной индексации для повышения скорости обработки. Такой СУБД может быть, например, СУБД SQL.

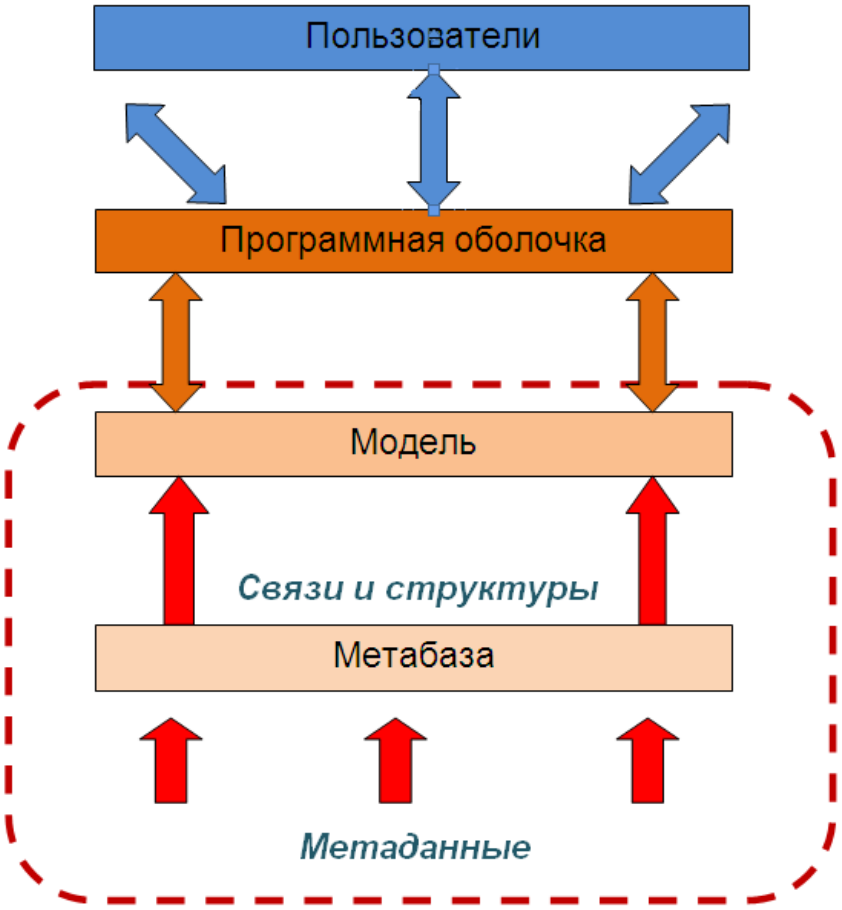

Рис. 2. Структура типичной БИС

Модель

Модель представляет собой упорядоченные и систематизированные данные, собранные на этапе наполнения базы. Модель определяет взаимосвязи между отдельными элементами метаданных, объединяет данные в классы и формирует правила обработки этих классов. Например, при изучении цепочек генов или протеинов, модель определяет положение каждого элемента в цепочке, его свойства и взаимодействие с другими элементами. Эти связи и взаимодействия могут быть заданы как средствами СУБД, так и с использованием языков программирования C++, C#, SQL и др. На этапе построения модели основной задачей разработчиков является наиболее полное и правильное с точки зрения современной науки описание известных связей и взаимодействий. Данная модель является преимущественно имитационной, так как связи между ее отдельными элементами не могут быть описаны строгими математическими выражениями или системами уравнений. Модель имитирует поведение биологического объекта (например, цепочек генов и протеинов) при их взаимодействии друг с другом или при внешнем воздействии. В отличие от строгих математических моделей, где выходной результат как

правило определяется математическими формулами и функциями, выходной результат имитационной модели нельзя предсказать заранее. И в этом ценность системы.

Технически метабаза и модель могут быть выполнены на единой платформе с использованием возможностей СУБД и языка программирования и управления СУБД.

Программная оболочка

Программная оболочка представляет собой верхний уровень БИС. На этом уровне с использованием языков программирования высокого уровня разрабатывается интерфейс конечного пользователя системы. В зависимости от назначения системы возможно наличие не одного, а нескольких интерфейсов для различных категорий пользователей. Например, различные интерфейсы и права доступа могут существовать для исследователей, для экспертов, для группы, занимающейся наполнением базы и т.д. Разграничение доступа позволяет пользователям сконцентрироваться именно на той задаче, которую они должны решить.

Что касается аппаратного обеспечения БИС – то здесь в основном не требуется каких-либо особых средств. Программная оболочка разрабатывается на стандартной платформе MSWindows, Linux или Unix. Основным требованием является наличие большого объема памяти для хранения данных и быстродействие памяти, что, в общем случае, не является критичным в связи с тем, что для БИС обычно не устанавливаются требования по времени отклика. Среди ученых были попытки использования специализированных ЭВМ для решения задач биоинформатики, но обычно все заканчивалось переходом на «стандартную» аппаратную платформу в связи со сложностью программирования для специальных ЭВМ и недостатком специалистов для их обслуживания и наладки.

Программная оболочка выполняется на языках программирования C++, C#, Perl, Java и др.

Таким образом, БИС представляет собой программно-аппаратный комплекс, призванный помогать специалистам в исследовании строения сложных биологических систем.

Обзор существующих БИС

В настоящее время в эксплуатации и в разработке находится большое колическтво различных БИС. Подавляющее большинство из них были разработаны иностранными фирмами (впрочем, с привлечением российских программистов и биологов). Остановимся лишь на некоторых из них.

БИС Mesquite

Страна разработки: США

Назначение: Mesquite - это программное обеспечение для сравнительной биологии, разработанное с целью помочь ученым-биологам в анализе сравнительных данных о живых организмах:

- имитация процесса зарождения, роста и смерти клетки,

- построение генных цепочек,

- анализ ДНК и протеинов

Язык программирования: Java

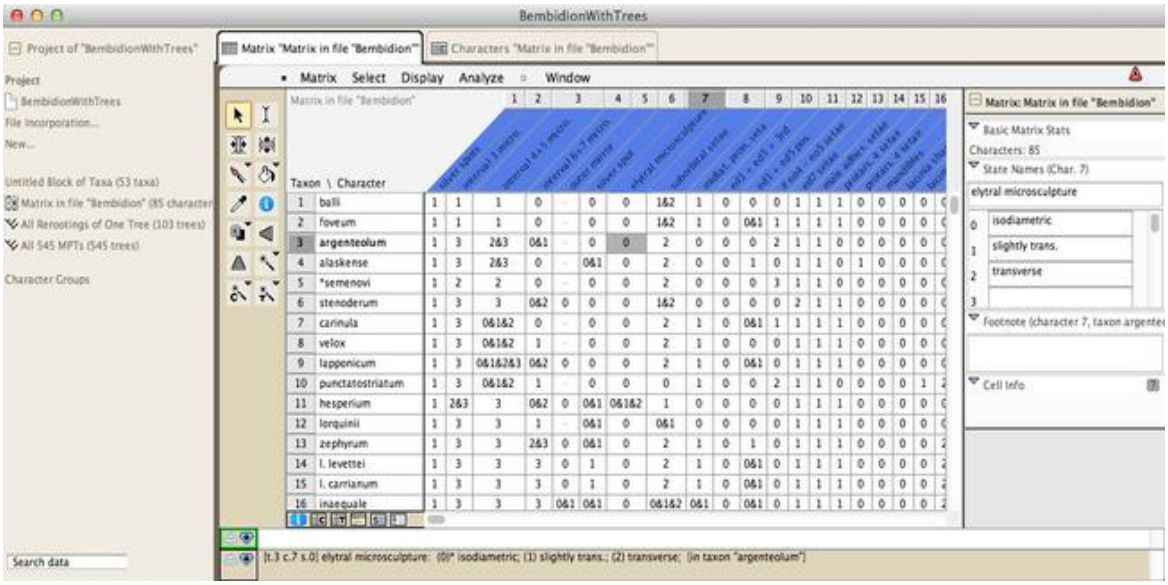

Рис. 3. Экран редактирования морфологических данных

PSI Protein Classifier

Страна разработки: США

Назначение: БИС, позволяющая обобщать результаты, полученные из других БИС. Определяет принадлежность найденных белков ранее известным семействам и разбивает оставшиеся белки на группы. Она позволяет количественно (числом итераций) оценить уровень родства между различными семействами белков-гомологов.

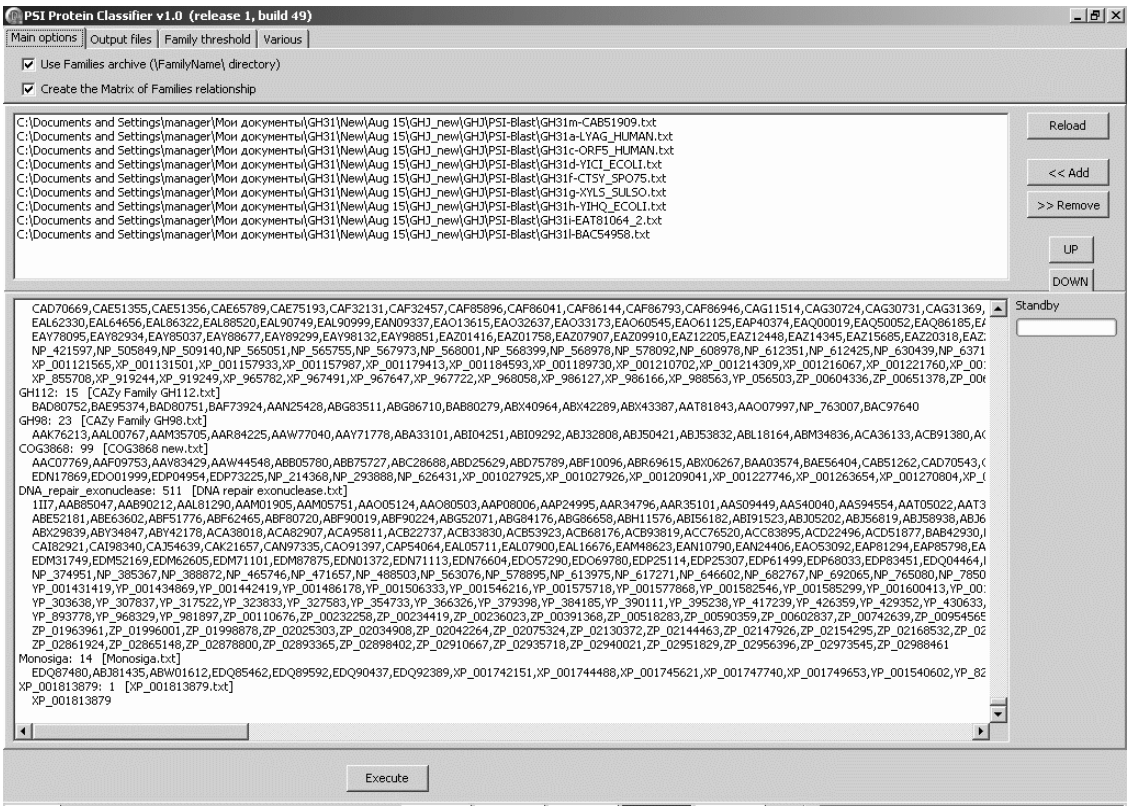

Рис. 4.Экран БИС PSIProteinClassifier. Описание структур протеинов

SPAdes Genome Assembler

QUAST — программа для оценки качества сборок

Страна разработки: Россия (СПБ)

SPAdes («Спейдз»), сборщик геномов, предназначенный для работы с данными как из одной клетки, так и из множества клонированных бактерий. В ассемблере применяется множество новых алгоритмических идей и улучшений по сравнению с существующими программами для сборки геномов.

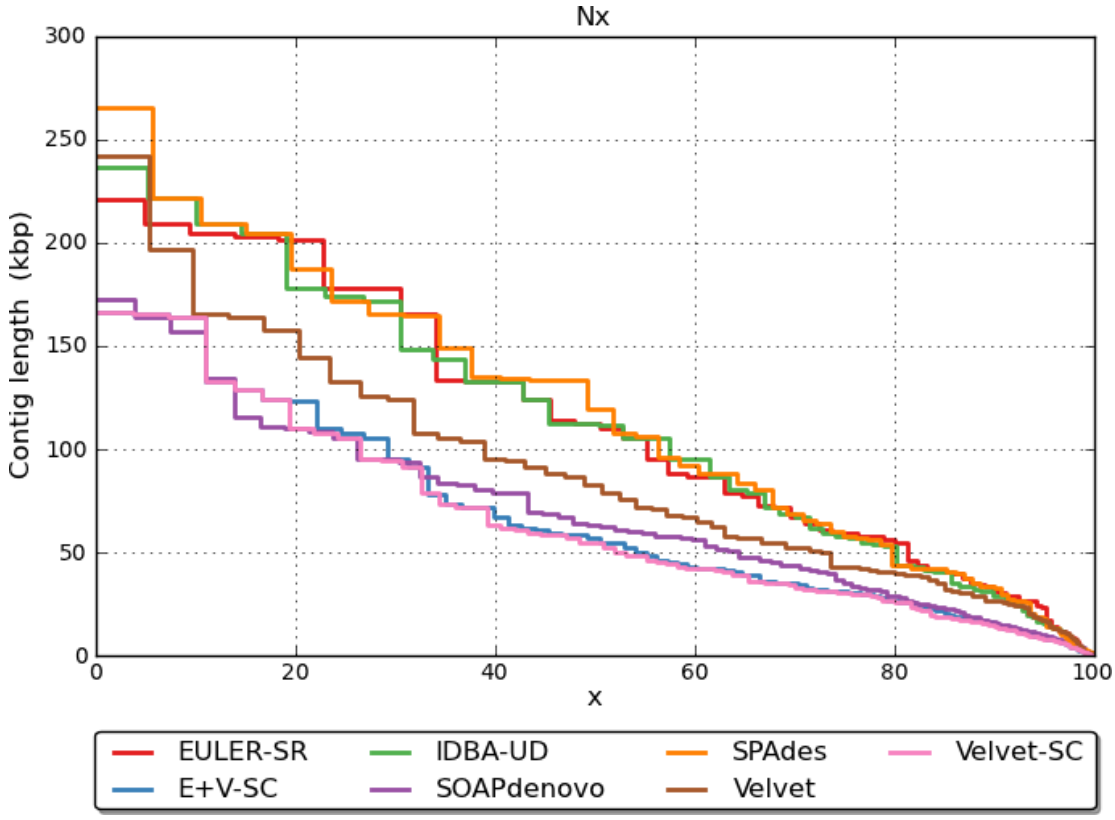

Рис. 5. Пример отчета, генерируемого программой КВАСТ

QUAST («Кваст») определяет качество геномных сборок

MetaDrug, MetaCore, Metabase

Страна разработки: США, компания GeneGo Inc.

Программный комплекс, состоящий из одной базы данных (**Metabase**), содержащей максимальное количество информации и нескольких программных надстроек, использующих данные из метабазы.

MetaDrug – БИС для разработчиков лекарственных препаратов. Содержит информацию о тысячах простых веществ, метаболических процессов, а также карты метаболических процессов. Позволяет с большой точностью предсказывать влияние того или иного препарата на метаболические процессы в организме человека.

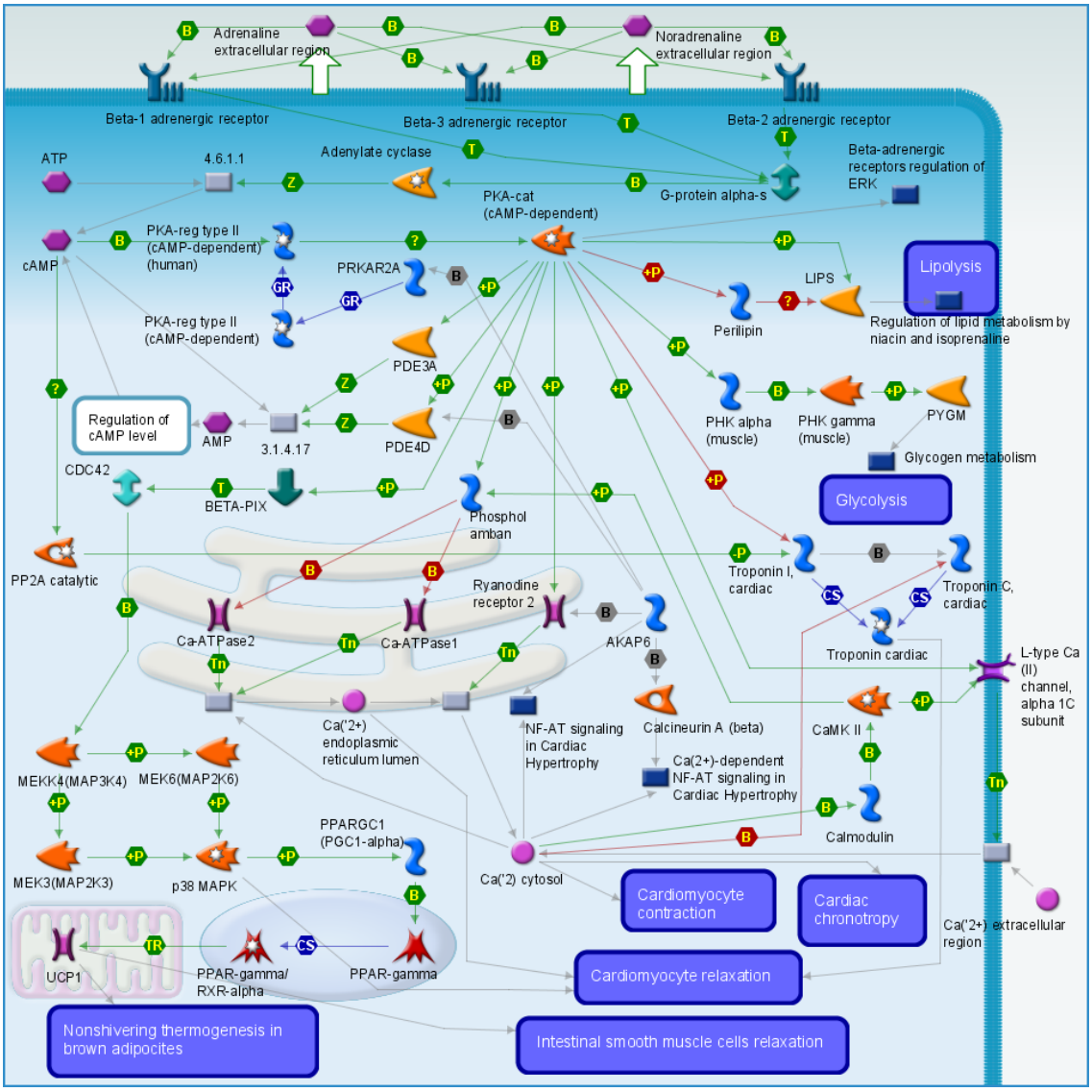

Рис.6. Карта метаболических процессов из системы MetaDrug

Система **MetaCore** предназначена для анализа генов и геномов. Также анализируется структура ДНК и РНК и пути прохождения метаболических сигналов. Используя возможности системы **MetaCore,** американские ученые добились значительных успехов в изучении механизмов возникновения болезни Паркинсона.

Обе системы ориентированы на биологов, биохимиков, фармакологов и генетиков.

Перспективы развития биоинформационных технологий

На рубеже XX-XXI веков биоинформатика превратилась в бурно развивающуюся область мировой биомедицинской науки. Наряду с исследователями, ведущими фундаментальные разработки, потребителями биоинформационных технологий являются медицинские, фармакологические, биотехнологичные и учебные учреждения. Эта область

науки определена в качестве приоритетной как в США, так и во всех других развитых странах.

Количество центров биоинформатики постоянно растет во всех странах Европы, Азии, США и Австралии. Наряду с государственными, академическими и образовательными центрами биоинформатики, в последние годы возникло значительное число организаций и проектов, ориентированных на коммерческое использование результатов исследований в области биоинформатики. Это прежде всего организации, деятельность которых ориентирована на структурный, функциональный и сравнительный анализ геномов, включая геном человека. Наряду с применением уже созданных методов биоинформатики интенсивно развивается техническая и программная база для решения прикладных задач, особенно в фармакологии. Быстрыми темпами совершенствуется также и индустрия программного обеспечения для решения таких задач.

Одной из задач биоинформатики как науки является выяснение функций генов. Выяснить функции всех генов опытным путем достаточно трудоемко.В этом случае биоинформатика помогает предсказывать их, опираясь насравнение с теми генами, функции которых уже определены.Сравнительно недавно в науке появился термин "биология in silico", буквальный смысл которого - "биология на кремнии", говоря иными словами, проведение биологического эксперимента на компьютере. Сейчас это понятие стало вполне официальным и широко используется. Есть журнал, который так и называется - "In silico biology".

Одним из интересных примеров использования методов биоинформатики является реконструкция истории расселения человечества. Экспериментально были изучены ДНК отдельных групп людей различных рас и национальностей, проживающих по всему земному шару. Затем методами биоинформатики было построено «дерево» человеческой истории. Оказалось, что человек как вид появился в Африке и произошло это около 200 тыс. лет назад. Около 100 тыс. лет назад наши предки вышли из Африки и начали расселяться сначала по Европе, затем в Азии, а оттуда двинулись в Северную и Южную Америку и в Австралию, оставляя следы своего передвижения в генетических текстах. Сравнение ДНК современного человека с ДНК обнаруживаемой в ископаемых останках неандертальцев показало, что на протяжении некоторого времени человеческий вид жил параллельно с неандертальцами, которые впоследствии вымерли. Данные, полученные методами биоинформатики коррелировали с данными антропологов. Интересно, что популяцияи человека, давшей начало всем людям, которые живут сейчас на земле, пришлось в какой-то момент пройти через «бутылочное горлышко», когда вся ее

численность составляла всего только 10 тысяч человек. Хотя пока нет данных, почему так произошло, в какой-то мере это указывает на обоснованность беспокойства ученых другой области науки по поводу возникновения вирусной пандемии.

Персонализированная (индивидуализированная, персонифицированная) медицина (ПМ) — сравнительно новое направление современной медицины, получившее развитие благодаря использованию методов направленного пациентассоциированного лечебно-диагностического воздействия, на основе учета влияний генетических, внешнесредовых и региональных факторов. Другими словами, это целевая диагностика (геномно--протеомная, метаболомная, транскриптомная) и лечение (индивидуально ориентированные воздействия, в том числе лекарственная, клеточная терапия) больного в соответствии с исходными результатами исследования его генетического профиля.ПМ также является перспективной областью интеграции современных биотехнологических подходов в медицинской практике, оптимизирующих понимание патофизиологической основы развития заболеваний, а также особенностей их молекулярной диагностики и терапии. Кроме того, такие активно развивающиеся направления медико-биологических исследований, как фармакогеномика, фармакогенетика, метаболомика и другие вносят свой вклад в развитие ПМ. Не остается в стороне и биоинформатика. Использование БИС для моделирования процессов, протекающих в организме человека, оценка воздействия тех или иных лекарственных препаратов на причину заболевания – вот те области, где принципы биоинформатики незаменимы.В основе развития ПМ лежит анализ особенностей генома человека. Так, в клиническом плане в первую очередь речь идет о том, чтобы с помощью генного анализа установить, стоит ли вообще принимать тот или иной препарат. Это необходимо потому, что даже незначительные индивидуальные различия в ДНК у двух пациентов могут привести к тому, что одно и то же лекарственное соединение будет действовать на них совершенно по-разному.

Сегодня мы находимся на начальном этапе использования генетической информации о живой материи, однако развитие все более эффективных методов расшифровки биологических текстов и разработка методов биоинформатики позволяет надеяться на серьезный прогресс в понимании строения, механизмов функционирования и регуляции живых систем. Таким образом, биоинформатика является перспективным научным направлением, сочетающим в себе достижения традиционных наук (химии, биологии, медицины, фармакологии и др.) с бурным развитием вычислительной техники и информационных технологий.

Список литературы

1. Bioinformatics.ru. Режим доступа: http://www.bioinformatics.ru (дата обращения 21.11.2014).
2. Несговорова Г.П. «Биоинформатика: пути развития и перспективы». Сайт Института систем информатики им. А.П. Ершова Сибирского отделения Российской академии наук. Режим доступа: http://www.iis.nsk.su (дата обращения 11.12.2014).
3. «Биоинформатика: виртуальный эксперимент в шаге от реальности». Портал «Наука и жизнь».. Режим доступа: http://www.nkj.ru (дата обращения 11.11.2014).
4. Сайт Лаборатории алгоритмической биологии Санкт-Петербургского Академического университета РАН. Режим доступа: http://quast.bioinf.spbau.ru (дата обращения 01.12.2014).
5. Официальный сайт компании GeneGo Inc. Режим доступа: http://www.genego.com (дата обращения 10.11.2014).