Что такое биоинформатика

Под биоинформатикой обычно понимают использование компьютеров для решения биологических задач. В настоящее время это почти исключительно задачи **молекулярной** биологии. Причина этого в том, что за последние 20–25 лет накоплен поистине колоссальный экспериментальный материал именно о строении и функционировании биологических молекул (белков и нуклеиновых кислот), в качестве примера достаточно привести геном человека. Этот материал требует развитых компьютерных методов для своего анализа. Поэтому биоинформатика как на нашем факультете, так и в подавляющем большинстве мировых научных центров понимается как синоним **вычислительной молекулярной биологии**.

Есть несколько основных направлений этого раздела науки, в зависимости от исследуемых объектов:

- Биоинформатика последовательностей.
- Структурная биоинформатика.
- Компьютерная геномика

С другой стороны биоинформатику можно условно разделить на несколько направлений в зависимости от типа решаемых задач:

- Применение известных методов анализа для получения новых биологических знаний.
- Разработка новых методов анализа биологических данных
- Разработка новых баз данных.

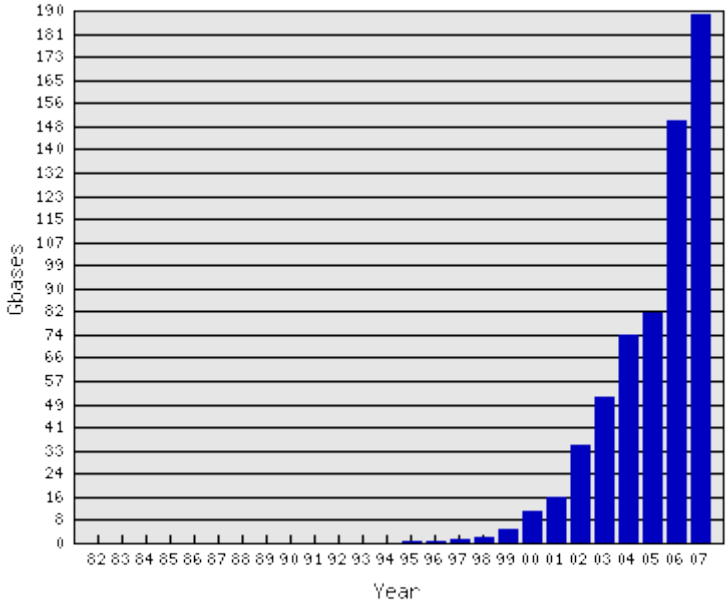
Наиболее известной и наиболее эффективной областью применения биоинформатики в настоящее время является анализ геномов, тесно связанный с анализом последовательностей.

Биоинформатика последовательностей

Этот раздел биоинформатики занимается анализом нуклеотидных и белковых последовательностей. В настоящее время разработаны эффективные экспериментальные методы определения нуклеотидных последовательностей. Определение нуклеотидных последовательностей стало рутинной хорошо автоматизированной процедурой.

В результате рутинной хорошо автоматизированной процедуры уже получено огромное количество генетических текстов. Так, в базе данных EMBL на 15.02.2007 г. хранится 87 000 493 документов с описанием нуклеотидных последовательностей, содержащих в целом 157545686001 символов (нуклеотидов), что соответствует примерно библиотеке в 105 толстых томов с убористым шрифтом. Найти нужный ген в EMBL, это все равно, что

найти цитату в такой библиотеке. Без помощи компьютера сделать это , мягко говоря, очень трудно. А число данных экспоненциально растет.



Представим себе геном небольшой бактерии – это непрерывная строка длиной в 1-10 миллионов символов символов, и далеко не вся ДНК кодирует белки. Первый тип биоинформатической задачи – это задачи поиска в нуклеотидных последовательностях особых участков, участков, кодирующих белки, участков, кодирующих РНК (например, тРНК), участков связывания с регуляторными белками и др. И это не всегда простые задачи, например, гены эукариотических организмов состоят из чередующихся «осмысленных» и «бессмысленных» фрагментов (экзонов и интронов), и расстояние между «осмысленными» фрагментами может достигать тысяч нуклеотидов.

Пусть ген найден. Что он кодирует? Зачем он нужен?

Если речь идет об участке ДНК, кодирующем белок, то с помощью весьма простой операции – трансляции с использованием известного генетического кода можно получить. аминокислотные (белковые) последовательности. Из известных на сегодня 4 273 512 белков около 94% последовательностей - это именно такие гипотетические трансляты, и больше о них ничего не известно. Скорость поступления информации с автоматических секвенаторов превышает скорость нашего понимания ее смысла!

Но биологические объекты – это объекты, возникшие в процессе эволюции. Сравнительно-эволюционный подход – один из мощнейших подходов в биологии.

Например, функция белка из одного организма хорошо экспериментально изучена, в другом организме нашли белок с похожей аминокислотной последовательностью. Можно предположить, что второй (неизвестный) белок выполняет ту же или схожую функцию. И здесь сразу возникает несколько вопросов. Во-первых, что значит похожая последовательность? Как сравнивать последовательности? При какой степени сходства последовательностей можно предполагать, что белки выполняют сходные функции?

Сравнение последовательностей (выравнивание) является важнейшей задачей биоинформатики. Трудно найти современного биолога, ни разу не использовавшего программы Blastp и ClustalX, появление этих программ – уже крупный успех

биоинформатики. Но современные биоинформатики недовольны и постоянно совершенствуют методы выравниваний.

Можно привести много примеров того, как сравнительно-эволюционный подход в сочетании с биоинформатическими методами порождает новое биологическое знание.

Генетические тексты – тексты с большой долей шума, сравнивая родственные последовательности, в ряде случаев удается отфильтровать шум и выявить сигнал, например, короткую последовательность нуклеотидов, способную связываться с белком-регулятором, или аминокислотные остатки в ферменте, отвечающие за связывание субстрата. Чтобы быть уверенными в результате, биоинформатики используют теорию вероятности и математическую статистику.

Подводя итог, можно сказать, что основные задачи биоинформатики, связанные с анализом отдельных последовательностей, состоят в следующем:

- Выравнивание и определение сходства двух последовательностей
- Построение множественных выравниваний
- Распознавание генов
- Предсказание сайтов связывания регуляторных белков
- Предсказание вторичной структуры РНК

Создание новых экспериментальных технологий ставит перед биоинформатикой целый ряд новых задач. Например, развитие масс-спектрометрии позволяет (пока в принципе) в одном эксперименте проанализировать весь набор белков, присутствующий в клетке. Для решения этой задачи необходим совместный анализ спектров масс и геномов. Открытие новых биологических явлений и механизмов также приводит к появлению новых задач. Хорошим примером служит открытие РНК интерференции, за которую в 2006 году дали Нобелевскую премию по физиологии. Это открытие породило целый вал биоинформатических работ, посвященных поиску участков связывания микроРНК и новых микроРНК. Многие находки были затем подтверждены экспериментально.

Про каждую задачу можно написать если не книгу, то большую главу в книге. Многие задачи биоинформатики стали рутинными методами анализа и применяются многими, в том числе и экспериментальными, биологами.

Структурная биоинформатика

Каждый белок, помимо своей уникальной последовательности аминокислот, из цепочки которых состоит его молекула, обладает ещё и уникальным способом укладки этой цепочки в пространстве. Задачу предсказания укладки по последовательности можно, в принципе, тоже считать задачей биоинформатики, но это задача в своём общем виде ещё слишком далека от своего решения. Поэтому структурная биоинформатика занимается анализом пространственных структур, уже определённых экспериментально.

Структур белков известно намного меньше, чем последовательностей белков. Это связано с тем, что экспериментальные процедуры для определения структуры намного сложнее, дороже, и к тому же (в отличие от секвенирования) не являются "рутинными", то есть их

результат вовсе не гарантирован. Тем не менее на начало 2007 года для анализа доступны более 30000 структур, что тоже немало (доступных белковых последовательностей — несколько миллионов). Среди них как структуры отдельных белковых молекул, так и структуры комплексов белков с ДНК, РНК, другими химическими веществами. Например, большинство лекарств представляют собой химические вещества, чьи молекулы способны связываться — образовывать комплексы — с молекулами тех или иных белков (как правило, в результате такого связывания белок оказывается неспособен выполнять свою природную функцию, что и обеспечивает эффект лекарства). Исследование механизма действия лекарств имеет большое практическое значение, поэтому определением структуры комплексов молекул белков с молекулами лекарств занимаются многие экспериментальные группы. Как результат — большое количество доступных для компьютерного анализа структур комплексов.

Примеры задач структурной биоинформатики:

- определение участков белковой молекулы, важных для той или иной функции данного белка (в биоинформатике часто вместо "определение" говорят "предсказание", поскольку компьютерный анализ не может иметь результатом научный факт, а лишь более или менее достоверное предсказание, которое должно быть затем проверено экспериментами);
- сравнительный анализ структур родственных белков, классификация белков на основе их пространственной структуры;
- анализ структур комплексов двух или нескольких молекул белка, комплексов молекул белка с другими молекулами; предсказание воздействия молекул химических веществ (в частности, потенциальных лекарств) на молекулы белков;
- предсказание структуры белка по структуре белка с похожей последовательностью (в такой ситуации задача предсказания укладки часто разрешима!).

Компьютерная геномика

В настоящее время определены полные или почти полные последовательности геномов многих организмов. Прочтение полной нуклеотидной последовательности какого-либо генома не является самоцелью. На самом деле это является первым шагом для исследования того, как функционирует та или иная клетка. Исследование геномов бактерий проводится для того, чтобы исследовать метаболизм бактерий и, в случае патогенных организмов, найти потенциальные мишени для лекарств. С другой стороны. Изучение геномов может позволить найти новые метаболические пути или ферменты, которые будут применены в биотехнологическом производстве (например, витаминов). В течение как минимум полувека сотни лабораторий исследовали кишечную палочку (E.coli). Но даже такой весьма изученный организм имеет как минимум 25% абсолютно не охарактеризованных генов. Значительное число секвенированных геномов принадлежат организмам, о которых вообще нет каких-либо других экспериментальных данных. Экспериментальное определение функции только одного гена требует интенсивной работы одной лаборатории как минимум в течение нескольких месяцев. Компьютерный же анализ позволяет с известной степенью точности охарактеризовать несколько тысяч генов силами небольшой группы примерно за неделю. Разумеется, компьютерный анализ не исключает экспериментальную проверку, однако в этом случае экспериментальная работа существенно упрощается.

Компьютерный анализ геномов состоит из следующих основных элементов:

- Предсказание генов в последовательностях. При этом в некоторых случаях удается даже найти ошибки в последовательности.
- Предварительная аннотация по сходству и другим особенностям белковых последовательностей.
- Сравнительный анализ геномов.
- Исследование регуляции работы генов.
- Поиск «пропущенных» генов. Представим себе, что в клетке есть цепочка реакций, преобразующих вещество А в вещество Б, а затем вещество Б в вещество В. При этом ген, ответственный за первую реакцию известен и в клетке присутствует, а для второй реакции гена не нашли. Это и есть пропущенный ген. На самом деле вторая реакция осуществляется и проблема заключается в том, чтобы найти в геноме подходящую кандидатуру.
- Исследование транспортеров (генов, обеспечивающих перенос питательных веществ в клетку и выброс вредных веществ из клетки)

Сравнительная геномика принесла уже несколько значительных открытий и «закрытий». В качесте «закрытия» можно привести, триклозан, который считался универсальным антибактериальным препаратом. Он входит в состав широко разрекламированного мыла "Safeguard". Его мишенью является белок, закодированный в гене fabI. Этот белок катализирует одну из реакций синтеза жирных кислот – необходимого компонента любой клетки. При этом у животных нет аналога этого белка, поэтому такой препарат безопасен для человека. Компьютерный анализ бактериальных геномов показал, что стрептококки не имеют белка fabI, а его функцию выполняет совсем другой белок fabR. Поэтому триклозан не действует на стрептококки. Одним из ярких открытий геномики является открытие принципиально новой системы регуляции – рибопереключателей. Это специфическая структура РНК, которая стабилизируется при непосредственном связывании с низкомолекулярным веществом и блокирует синтез матричной РНК. Предсказание структуры и механизма действия было блестяще подтверждено экспериментально.

Другой класс исследований, проводимых компьютерной геномикой – полногеномный анализ и исследование эволюции. В частности с помощью массового анализа было обнаружено, что альтернативный сплайсинг в генах человека является скорее правилом, чем исключением. Эволюционный взгляд на проблему позволяет выдвинуть гипотезу о том, что сплайсинг, в частности альтернативный сплайсинг, является эффективным механизмом для эволюции, позволяющем без значительного риска для генома перебирать варианты последовательностей.

Массовый анализ большого количества геномов показал, что, по крайней мере у безъядерных организмов (бактерий и архебактерий), явление горизонтального переноса генов между видами является весьма распространенным явлением – от10 до 30% генов в этих геномах горизонтально перенесены из других видов.

Применение известных методов анализа для получения новых биологических знаний

Существует широкий спектр методов и инструментов для компьютерного анализа биологических данных. Здесь можно упомянуть и BLAST — наиболее популярный сервис для поиска похожих последовательностей в базах данных, и программы множественного выравнивания аминокислотных последовательностей, и программы предсказания вторичных структур РНК, программы визуализации пространственных структур, программы моделирования динамики пространственных структур и многое другое. Большинство этих программ представлены в Интернете и имеют весьма удобный пользовательский интерфейс. Однако компьютер обладает одним свойством о котором нельзя забывать. На дурацкий вопрос компьютер всегда дает дурацкий ответ. Поэтому очень важно понимать границы применимости тех или иных методов. Любой компьютерный анализ биологических данных является экспериментом (только сделанным не в пробирке) и к нему предъявляются те же требования – важна четкость постановки и необходимы соответствующие контроли. Значительная часть биоинформатических работ сделана именно с применением уже существующих средств. Для проведения такого рода работ, как правило, нет необходимости уметь программировать. Достаточно только внимательно анализировать результаты работы уже готовых программ. При этом часто биоинформатический анализ предшествует постановке эксперимента. С другой стороны массовый (например, геномный) анализ требует использования простейших программ собственного исполнения (попробуйте глазами проанализировать 100 тыс. выравниваний).

Разработка новых методов анализа биологических данных

Иногда существующие программы недостаточны для решения поставленных задач, или существующие программы имеют не достаточную точность, или для интересующей исследователя биологической задачи нет подходящих средств, или появился новый тип данных. В этом случае приходится разрабатывать новые алгоритмы и программы. Таких примеров множество. Даже алгоритмы для такой классической задачи, как построение множественного выравнивания имеет достаточно сильные ограничения. Современная биоинформатика при разработке новых алгоритмов широко использует достижения теории вероятностей, математической статистики, информатики.

Разработка новых баз данных

Значительная (если не вся) биологическая информация поступает в различные банки данных. Эти банки содержат первичную, зачастую грязную, информацию. Далее эта информация перерабатывается, в том числе с привлечением научной литературы. В результате возникают литературные, курируемые и вторичные банки данных. Информация в них, как правило, заслуживает большего доверия. Однако, создание новых курируемых банков данных – весьма трудоемкая работа.