

Лекция седьмая «Компьютерные алгоритмы для предсказания вторичной структуры. Часть 2. Термодинамический и сравнительный подходы»

Спецкурс "Информационная Биология"

Лектор к.ф.-м.н., с.н.с. И.И.Титов.

На первой части лекции вы получили представление о том, что такое вторичная структура РНК, и зачем собственно нужно искать, по каким правилам она образуется, какие физические взаимодействия лежат в ее основе, каковы механизмы ее формирования. Эта лекция будет более формальная математически, в основном будет посвящена анализу преобразования символьной последовательности РНК во вторичную структуру, точнее алгоритмам, которые такое преобразование производят.

Одним из наиболее используемых алгоритмов предсказания вторичной структуры РНК является термодинамический алгоритм, который основан на предположении о том, что молекулы РНК формируют свою вторичную структуру путем физических взаимодействий, причем за время формирования вторичной структуры она успевает прийти в равновесное состояние. Это означает, что среди большого числа потенциальных вторичных структур в растворе, или в клетке, реализуется малое количество, так что если, скажем, экспериментатор будет производить эксперимент, определять вторичную структуру с помощью своего экспериментального метода, он получит достаточно устойчивый результат, собственно говоря, это чаще всего и реализуется в природе.

Те РНК, которые рассматриваются как структурные, несут свои функции благодаря пространственной структуре, такие как, например tRNA, 5sRNA, другие рибосомальные RNA. Их экспериментальный анализ дает основания считать, что их структура уникальна и что в определенных условиях практически все они имеют определенную пространственную структуру. Тем не менее, поскольку РНК, как вам на предыдущей лекции было рассказано, имеет некоторые взаимодействия, во вторичной структуре они могут быть очень стабильными и долгоживущими, то РНК может не успеть перейти в это равновесное состояние и таким образом оказаться в каком-то метастабильном состоянии. Это одно из ограничений термодинамического подхода. Другим ограничением является то, что необходимо знать большое количество энергетических параметров для образования структуры, которые дают вклад в энергию, и погрешности в знании этих параметров неизбежны, они приводят к неточности расчета структуры РНК. Еще одним моментом может быть то, что мы а priori не знаем, в каком окружении находится РНК. Дело в том, что на нее могут влиять белки, ее укладка в пространственную структуру может управляться разными белками. В большинстве случаев РНК в природе не изолирована и действует в комплексе с другими молекулами. Все эти взаимодействия практически невозможно учесть, что опять же ограничивает гипотезу о достаточности термодинамических параметров для предсказания вторичной структуры РНК. Тем не менее, термодинамический подход является самым распространенным, экспериментаторы пользуются алгоритмами именно такого сорта, по той причине, что они быстры и их параметры в отличие от кинетических алгоритмов наиболее точно установлены, и экспериментаторы уделяют большое внимание уточнению этих параметров. Работа в этом направлении постоянно ведется, и за 40 лет с тех пор, когда впервые эта задача была поставлена, достигнут значительный прогресс в этой области.

Хотя в настоящее время нельзя назвать ни одного метода, который бы надежно предсказывал вторичную структуру РНК, но по сравнению с первыми алгоритмами, которые основывались на том, чтобы подсчитать ту структуру, которая максимизирует количество комплементарных пар, прогресс все-таки значительный. Наиболее распространенным алгоритмом является mfold (Слайд 2), который 20 лет назад примерно был предложен Цукером и Стиглером, и использующий для скорости счета рекурсивные соотношения. Это

означает, что вычисляется некая характеристика на фрагменте последовательности, и когда мы переходим к большему фрагменту, энергия большего фрагмента зависит от энергии меньшего фрагмента, и благодаря этой рекурсии, двигаясь, увеличивая размер фрагмента, можно быстро подсчитать вторичную структуру, энергию вторичной структуры. Чисто математически такая процедура возможна по той причине, что энергия вторичной структуры считается очень упрощенно в аддитивном виде, т. е. энергия вторичной структуры состоит из энергии петель и энергии спиралей. Чисто математически это свойство позволяет формулировать рекурсивную зависимость и в конечном итоге эта задача решается с временной сложностью, т. е. время счета зависит от длины с полиномиальной временной сложностью.

Самые быстрые алгоритмы имели кубическую временную сложность, то есть если мы увеличили последовательность в 10 раз, то время счета увеличится в 1000. Тем не менее, в сравнении с алгоритмами полного перебора рост не такой большой, т. к. полное число всех структур растет экспоненциально. Если для длины 10 существует считанное количество потенциальных вторичных структур, то для длины 50 типичное значение - 10 в 30-й степени, а для длины 150 уже 10 в 60-й степени, и прямым перебором просмотреть их становится невозможным. Именно здесь выходят на арену вот такие рекурсивные алгоритмы.

Проиллюстрирую на простейшем примере (Слайд 3): ищется только минимальная энергия структуры, причем пренебрегается энергиями петель, подсчитывается энергия только комплементарных связей. Рекурсивная, это значит сама постановка задачи, что нужно найти минимальную из всех возможных структур (минимальную по энергии) и здесь формальное рекурсивное соотношение. Это ограничение обозначает, что шпильчатые петли не могут быть длины меньше чем три. Строчкой ниже - три возможных варианта при удлинении последовательности на 1 основание, то есть мы рассматриваем последовательность, которая начинается с буквы i и кончается буквой j, и увеличиваем и на 1. При этом здесь проявляется несколько членов, которые соответствуют разным ситуациям. Два терминальных основания не вовлечены вообще в комплементарные взаимодействия, либо они связаны с основаниями внутри вот этой подпоследовательности i-j, вся структура разрывается на две части, либо они связаны друг с другом.

В результате здесь представлен псевдокод для рекурсивного вычисления минимальной энергии вторичной структуры РНК (Слайд 4), и если сюда ввести энергию петель (петли бывают нескольких типов: либо мультипетли, которые формируются несколькими спиралями одновременно, либо есть петли, которые формируются только двумя спиралями, это боковые или внутренние петли, есть петли, которые формируются только шпильками), то таким образом для вычисления энергии фрагмента, находящегося между вот этими двумя парами, нужно рассмотреть какая петля здесь формируется, шпильчатая, боковая или мультипетля, рассчитать энергию у этого фрагмента, увеличить его на 1, с использованием знания энергии на предыдущем шаге и так двигаться до того, как пройдет вся последовательность. То, что вычисленная в этой процедуре энергия не дает знаний о том, какая именно вторичная структура, энергию какой именно вторичной структуры вычислили, это дает только значение самой энергии, которое является минимальным.

Для того, чтобы определить саму структуру в этом алгоритме, он называется алгоритм mfold, производят движение назад так называемое, то есть вначале алгоритм движется вперед, пока последовательность полностью не будет покрыта, и затем, сохраняя информацию о траектории движения, производят движение в обратную сторону и таким образом устанавливают все комплементарные пары, которые формируются в этой структуре. Физически эта ситуация не очень удовлетворительна по причине того, что не факт, что все-таки вторичная структура будет находиться в минимуме своей энергии, во-первых, а во-вторых, тем не менее, хотя костяк структуры может быть неизменным, в нем могут быть достаточно лабильные участки, которые из-за теплового движения будут отрываться и снова

захлопываться либо вместо них могут образовываться другие конкурирующие спирали. Если взять все количество молекул, которые находятся в пробирке, часть из них будет регистрировать одни, часть другие взаимодействия, и какие именно взаимодействия будут проявляться наиболее ярко - простое вычисление одной только структуры об этом информации не дает.

Поэтому в дальнейшем, 10 лет назад, МакКаскиллом был предложен алгоритм статсумм (Слайд 5). Статсумма дает всю информацию об ансамбле молекул, которые находятся под наблюдением. Постановка задачи здесь такая - берется огромное число молекул, и каждая из них пребывает в каком-то состоянии ввиду теплового движения. Эти молекулы, двигаясь, могут менять свои состояния, и поэтому неизбежно этот ансамбль будет рассыпаться по набору состояний, а экспериментальная регистрация будет соответствовать некой характеристике, которая усреднена по этому набору. Например, если у вас имеется двухуровневая система, например (простейший случай) - цис и транс изомер, то под действием молекул воды, которые все время находятся в тепловом движении этот цис-транс изомер будет разогреваться и находиться при какой-то температуре. По случайным причинам барьер между цис и транс конфигурациями может быть преодолен, и значит молекула может переходить из одного состояния в другое. В равновесии любая характеристика системы будет зависеть от доли молекул в том или ином состоянии.

Формально вся информация о статистическом ансамбле записывается в статсумме. При этом свободная энергия для вторичной структуры РНК вычисляется таким образом, что энергия спирали это есть энтальпия взаимодействия, энергия, с которой притягиваются два фрагмента нити друг к другу и зависит так же от энтропийного вклада. Он определяется тем, насколько несвободными стали мономеры после того, как они образовали друг с другом комплементарные взаимодействия. Энергия же петли (обычно считается что энергия петли не имеет никакого энтальпийного вклада) зависит от того, насколько сильно нить РНК при образовании теряет в своей конформационной свободе при образовании спирали из-за того, что концы, которые образуют спираль, должны находиться рядом друг с другом и поэтому число таких конформаций меньше. Таким образом, энергия петли считается зависящей от энтропии. Предположим, что мы знаем энергию всех спиралей, всех петель, возможных в конкретной молекуле РНК.

После этого мы можем вычислить характеристику, называемую статистической суммой (Слайд 6), и вероятность реализации какой-то конкретной структуры, одной из многих. Этот подход оказывается идентичным подходу минимизации энергии, в той ситуации, когда свободная энергия одной из структур сильно отличается от остальных в сторону уменьшения, то есть между ней и остальными структурами большая так называемая термодинамическая щель. Когда эта термодинамическая щель большая, поскольку здесь стоит достаточно быстрая функция, экспонента, эта щель превышает тепловую энергию многократно, и соответственно вероятность этой структуры будет значительно выше вероятности другой заданной структуры. Это отношение может быть еще больше, десятки, тогда оказывается, что структура является доминирующей не относительно какой-то другой структуры, а относительно всех структур вместе. Чисто формально этот алгоритм напоминает рекурсивный алгоритм для минимизации энергии. Здесь тоже используется рекурсия, статистическую сумму большего фрагмента можно вычислить, зная статсумму меньшего фрагмента, таким образом, начиная с самых маленьких фрагментов и увеличивая размер фрагмента, можно прийти до статсуммы всей молекулы.

После того, как рекурсия проходит, вероятность наблюдения какой-то комплементарной пары, какого-то комплементарного взаимодействия, может быть вычислена. После вычисления мы потеряем знания о том, в составе какой именно структуры комплементарное взаимодействие реализовалось, но зато у нас будет информация о том, какова эта

вероятность в не зависимости от привязки к какой либо структуре, т. е. это соответствует такому эксперименту - мы поместили все РНК в пробирку и действуем ферментом, который специфичен к двунитиевому состоянию конкретной пары. Это очень близко к тому, как происходит биохимический эксперимент.

Обычно имеется несколько типов ферментов, одни из которых действуют на определенные нуклеотиды в определенном состоянии, разрезают ковалентную нить в тех условиях, когда, скажем, определенный нуклеотид находится в двунитевом или в одонитевом состоянии. Таких ферментов достаточно большое количество и ими действуют независимо друг от друга. Затем все РНК из пробирки подвергаются гелевому электрофорезу, и таким образом идентифицируют, в каком месте произошел разрез в полимерной цепи. Таким образом, реально имеется информация о том, с какой интенсивностью данный нуклеотид вступает в комплементарные взаимодействия, но не получается информации о том, с кем он вступает во взаимодействие. Поэтому эта характеристика более детальна, в ней содержится больше информации.

Псевдокод для этой рекурсии абсолютно аналогичен (Слайд 7), и на выходе, естественно, надо учитывать вероятность Больцмана при расчете вероятностей нахождения нуклеотидов в том или ином состоянии. Такой алгоритм полезен для вычисления всех физических характеристик РНК, то есть реально нас может интересовать не только структура, но и другие ее физико-химические свойства.

Например, в калориметрическом эксперименте измеряется теплоемкость РНК. Обычно постановка эксперимента такая, что изменяется температура и измеряется количество тепла, которое поглощает раствор при разных температурах, и в той ситуации когда происходит структурный переход, например, РНК денатурируется или расплавляется конкретная спираль, на кривой теплоемкости наблюдается пик и с помощью алгоритма статистических сумм такую кривую можно строить и сопоставлять с наблюдаемыми данными.

На рисунке (Слайд 8) показывается пример как паттерн, вовлеченный во вторичную структуру, может меняться в зависимости от температуры, т. е. при одной температуре наиболее стабильной является одна структура, при другой температуре с минимальной энергией может быть другая структура, и поэтому информация о субоптимальных структурах важна.

Здесь (Слайд 9) представлена удельная теплоемкость 5s-рибосомальной РНК кишечной палочки, и в принципе кривая предсказывает этот пик достаточно точно.

Одним из удобных представлений для самой вторичной структуры так называемая доп-матрица, строки и столбцы матрицы соответствуют нуклеотидам последовательностей РНК. В этой ситуации какое-то конкретное комплементарное взаимодействие будет сигнализировать на этой матрице о комплементарном взаимодействии двух нуклеотидов (Слайд 10).

Термодинамический алгоритм Зукера дает информацию только об одной структуре, дает одно предсказание без всяких альтернатив, а алгоритм статистических сумм показывает, что существуют вероятности того, что нестабильные взаимодействия могут распасться и образоваться новые. Хотя они маловероятны, тем не менее, они присутствуют. Когда же мы берем структуру достаточно длинной последовательности, таких альтернативных взаимодействий оказывается очень много. Но, тем не менее, в ситуации, когда имеется одна наиболее стабильная структура, эти два типа алгоритмов дают сходные предсказания.

Но на самом деле РНК практически всегда присутствует в сложных контактах с другими молекулами (Слайд 11), поэтому те расчеты, о которых я говорил, соответствуют моделированию молекулы в вакууме. Но они в принципе годны для расчета малых длин, малых фрагментов РНК.

Таким образом, если суммировать все достоинства и недостатки (Слайд 12) термодинамического подхода, можно сказать следующее - что, во-первых, достоинством является то, что они непрерывно поддерживаются, т. е. работа над ними ведется непрерывно. Экспериментаторы работают над тем, чтобы параметры, которые определяют структуру, становились все более точными, и их многообразие расширялось. Однако, этих параметров очень много и взаимодействий такое большое количество, что их практические абсолютные знания практически никогда не будут достигнуты. К тому же взаимодействия с другими молекулами, в том числе с белками, никак не отражаются в энергетических таблицах и можно достаточно пессимистично отнестись к тому, что когда-нибудь это знание у нас будет. Это означает, что если у нас есть некий набор вторичных структур РНК, реализуется одна из форм РНК, но какая именно, со всей определенностью сказать нельзя, то есть необходимо в реальности брать некое ограниченное число стабильных структур и затем уже выбирать одну из них в зависимости от того, как она накладывается на экспериментальные данные или на какие-то другие биологические данные, причем можно эти данные включать в расчет, и в этом смысле задача облегчается. Другим достоинством этой группы алгоритмов является то, что они полиномиальны по скорости, время их работы зависит от длины последовательности, как правило, кубичным образом. К примеру, для больших молекул, скажем, для молекул длиной 10.000 нуклеотидов требуется уже суперкомпьютер, но для малых молекул становится практически возможным рассчитывать их структуру и в том числе можно пользоваться интернетом в этом расчете, что немаловажно.

Во-вторых, здесь можно вычислять другие характеристики, то есть имеется дополнительная свобода в тестировании модели структуры РНК, скажем, можно проводить калориметрический эксперимент или измерять какое-то другое свойство и сопоставлять с тем, какое должно получаться при реализации конкретной структуры. Опять же, в настоящее время существует два типа пакетов программ, которыми можно пользоваться в этой задаче, один пакет - это mfold, он установлен в интернете в Америке, так же есть Windows-версии, которые можно скачивать и устанавливать, и еще есть так называемый Венский пакет, в котором реализован алгоритм статистических сумм. Через интернет он, к сожалению, не работает и требует некоторых компьютерных знаний для инсталляции.

Определение: так называемая *ab initio* задача - задача, когда есть последовательность РНК, и нужно определить ее структуру. На самом деле организмов живых очень много, и они очень сильно похожи друг на друга, например, человек от обезьяны отличается генетически на 1 процент, и большое количество генов оно идентично, что в человеке, что, скажем, в дрожжах. И, естественно, поскольку они живут независимо друг от друга, ДНК одного организма эволюционируют независимо от другого, то в процессе эволюции накапливаются мутации, последовательности начинают отличаться друг от друга, и эту информацию нужно учитывать при анализе вторичной структуры РНК (Слайд 13). Поэтому этот факт породил целое направление в предсказании вторичной структуры РНК, разработки алгоритмов для предсказания вторичной структуры РНК, так называемый сравнительный подход.

Вообще существует гипотеза, что жизнь состояла только из РНК, поскольку было установлено, что РНК могут осуществлять практически все функции, которые несут белки, поэтому можно представить себе простейший организм, который мог бы состоять только из РНК. В этой гипотезе предполагается, что в процессе предбиотической эволюции возникли РНК, которые были способны самореплицироваться и катализировать. Можно представить, что у них был стимул усилить эту функцию саморепликации, так как если были две разные РНК, одна из которых реплицировалась быстрее, чем другая, то в конечном итоге потомки первой должны вытеснить вторую в конкуренции из-за субстрата (субстрат в данном случае - мономеры, точнее химические элементы, из которых эти РНК состоят).

Сравнительный подход основывается на том, что берутся РНК, которые имеются у разных

организмов, функциональные РНК, имеющие одну и ту же функцию, и сравниваются их структуры. Если вторичная структура, которая поддерживается эволюцией, является консервативной, то эта инвариантность приводит к тому, что образуются пары, которые сохраняют комплементарность, то есть если в какой-то позиции произошла замена А на У, то, скорее всего нужно ожидать, что будет обратная замена. Если по случайным причинам произошла мутация, то организм либо очень быстро найдет так называемую коадаптивную замену, чтобы восстановить комплементарность, либо он окажется непригодным и отомрет. Так что если эти все реальные гипотетические организмы собрать все вместе, то можно провести анализ, который заключается в том, чтобы найти те столбцы, в которых комплементарность является инвариантной. Естественно, это обозначает, что эти последовательности должны быть предпроцессированы перед таким анализом, предпроцессинг называется выравниванием (Слайд 14), потому что в процессе эволюции опять же могут возникать делеции и вставки, у какой-то РНК петля может оказаться большего размера в одном из организмов, но если это не будет влиять на функцию РНК, то такое изменение будет нейтрально с точки зрения приспособленности организма, и он его легко перенесет.

Фактически это обозначает, что после того, как мы секвенировали все последовательности, мы должны их выровнять, то есть найти ядро, которое является эволюционным инвариантом, и сама эта процедура зависит от того, какова вторичная структура, если она имеется, то есть выравнивать на самом деле нужно иначе. С другой стороны, анализ вторичной структуры будет сильно зависеть от результатов выравнивания, то есть какие мы столбцы здесь оставим, какие выбросим, то есть в принципе эта ситуация не удовлетворительна, но до сих пор никакой единой процедуры, которая бы делала все разом, не предложено. Вот есть структурная РНК, когда их много, определять вторичную структуру очень легко. Взяв 5s рибосомальную РНК из многих организмов, построили профиль комплементарности (Слайд 15). И видно, что на самом деле существуют пики, которые принадлежат тем или иным спиральным участкам во вторичной структуре. И с помощью суперпозиции пиков можно построить модель вторичной структуры.

Фактически, эту задачу можно решать более формально, вычисляя ковариации между позициями, и получать матрицу комплементарности, которая будет похожа на ту, которая получается в алгоритме статистических сумм. И точно так же, если есть ярко выраженная вторичная структура, если выравнивание сделано очень хорошо, то вторичная структура на этой матрице комплементарности проявится. Чисто формально, если рассмотреть последовательность РНК как некий неизвестный текст, который нам нужно декодировать в структуру, то в этом тексте требование структуры, требование комплементарности вносит корреляцию между символами, причем корреляции специфического вида, так называемые полиндромы, причем не простой, а комплементарный полиндром. Полиндром - это зеркальная последовательность, если взять символьную последовательность и зеркально ее отобразить, то получится полиндром.

В последнее время начали развиваться подходы, которые зная последовательность, стараются восстановить правила, по которым последовательность образуется. Это называется грамматиками, и впервые задачи такого сорта поставил известный лингвист, Чомский. Чисто формально вопрос может быть поставлен так - принадлежит ли данное предложение какому-то конкретному языку, но математически эта постановка неправильная, правильно задать вопрос так - может ли грамматика произвести данное предложение, при этом предложение может быть абсолютно бессмысленным, но носитель языка тем не менее его идентифицирует его как предложение. Вообще Чомский - довольно интересная личность, он предложил в биологии интересную гипотезу о происхождении языка, его гипотеза была в том, что в человеке произошла мутация, которая изменила морфологию его звукового аппарата, и благодаря изменению этой морфологии он стал обладать возможностью издавать большое число различных звуков, достаточно членораздельных, так что другой сородич мог

эти комбинации звуков отличать от другой. Таким образом, возникла система коммуникаций между людьми, возникла речь, это стимулировало развитие мозга и социальной структуры, и происхождение человечества обязано вот такой мутации. Как я уже говорил, от обезьяны мы отличаемся на 1 процент по ДНК.

Так вот Чомский поставил вопрос именно таким образом - у нас есть много предложений на неизвестном языке, и нам нужно определить грамматические правила, по которым эти предложения написаны. С точки зрения РНК это символьные последовательности из разных организмов, и нам нужно восстановить структуру РНК. При этом Чомский классифицировал грамматики по принципам от простых к более сложным (Слайд 16), самая простая грамматика, когда следующий символ порождается предыдущим, или набором предыдущих. Ясно, что грамматики такого сорта не порождают текстов, которые обладают длинными корреляциями.

Для того, чтобы иметь длинные корреляции на произвольном расстоянии, нам нужно либо делать копии, либо вставлять палиндромы, значит нужна другая грамматика, в которой мы можем писать символы как справа так и слева, то есть стартуя с какого-то символа, можем писать новый символ справа, слева, можем одновременно справа-слева, и таким образом могут образовываться полиндромы и копии (Слайд 17, 18). Грамматика такого сорта называется бесконтекстной грамматикой, а стохастической она называется по той причине, что правило, по которому мы пишем каждый следующий символ, не является строго детерминированным, то есть мы пишем тот или иной символ в зависимости от некоторой вероятности, при этом каждый раз бросается монетка. То есть это опять же осмысленно с точки зрения РНК, так как опять же замены и полиндромные комплементарные участки могут отличаться друг от друга, но иметь один и тот же функциональный смысл. То есть есть несколько правил порождений, каждый раз правило может разыгрываться случайно.

Так же это можно представить в виде некоего дерева грамматического разбора (Слайд 19), и такие сравнительные алгоритмы могут быть использованы не только для построения, для восстановления структуры РНК, но и для поиска неизвестных РНК. Геном у нас большой, и вероятно, что не все РНК до сих пор идентифицированы (задача их поиска поставлена сравнительно недавно), и оказалось, что существуют десятки РНК, очень коротких, которые, тем не менее, помогают организму функционировать. Считается, что эволюционные алгоритмы являются наиболее точными, правда, как я уже говорил, они требуют большого количества последовательностей. То есть нельзя делать *ab initio* предсказание по той причине, что не с чем сравнивать, скажем, если ищется корреляция между столбцами, то есть если матрица состоит из одной строки, ни о какой достоверности не может идти и речь, к тому же там имеются свои проблемы, если скажем организмы-предки разошлись сравнительно недавно, то нужно это учитывать. В этой ситуации нужно восстановить некую предковую последовательность, или набор последовательностей, которые образуют предковую структуру. При этом мы не знаем, каким образом протекала эволюция, а результат такого анализа зависит от того, как она протекала.

Филогенетическое дерево, в вершине которого имеются предки, а дальше идет ветвление. От топологии этого дерева оказывается сильная зависимость в этих алгоритмах, то есть если у нас, например, дерево не сбалансированное, то это соответствует тому, что каждый вновь отделившийся потомок дальше производит следующих потомков, а скажем какой-нибудь древний организм - от него дальше никаких ветвлений не происходит. Это называется максимально несбалансированным деревом, или максимально сбалансированным когда был общий предок, потом был так называемый взрыв, то есть он породил очень много потомков, все быстро образовались, так что дерево стало напоминать звезду - граф, у которого есть вершина и все ребра из нее выходят.

Безусловно, стоит учитывать разницу между этими двумя случаями. Какой конкретно случай, заранее не известно, и к тому же этот алгоритм сильно зависит от выравнивания и от большого количества других параметров, которые практически никак не контролируются, так что здесь важна роль эксперта при практическом использовании этого алгоритма такого типа. Поскольку нужно процессировать большое количество информации (речь идет о грамматиках), то эти алгоритмы обладают низкой скоростью. Если же речь идет о ковариационном анализе, с чего я начал, там безусловно все просто, но возникают проблемы с точностью. С другой стороны, с помощью этого подхода можно находить тройные взаимодействия или какие-нибудь более высокого уровня, потому что, скажем, в природе РНК может быть достаточно компактной, например тРНК - клеверный лист, так что тройки нуклеотидов могут взаимодействовать друг с другом, причем это взаимодействие слабоспецифично.

Например, если два взаимодействия комплементарны, то третье основание, которое с ними взаимодействует, слабоспецифично. Обычно эти взаимодействия очень сложно искать, если у нас имеется только вторичная структура, непонятно, как они располагаются, то здесь как бы мы видим продукт эволюции и можем такие тройки находить, несколько таких троек было найдено в 5sRNA. С другой стороны, больших молекулах, скажем 16sRNA которые имеют в длину не десятки или сотни пар оснований, а скажем, полторы тысячи и больше, очень велик статистический шум и чтобы выяснять такие тройные взаимодействия, количества современных организмов может быть даже и недостаточно. Но с другой стороны, поскольку здесь имеется продукт эволюции, они никак от точности не зависят, зависят только от темпов секвенирования последовательностей. Именно с помощью этих алгоритмов были построены модели всех известных структурных РНК, правда, естественно, там применялись и методы минимизации, для того чтобы как-то локально уложить структуру РНК.

Таким образом, вторичная структура РНК - далеко не законченное исследование в биоинформатике, хотя математически очень просто сформулировать правила, по которым РНК укладывается хотя-бы во вторичную структуру, правила комплементарности, но тем не менее оказывается что простых правил недостаточно, или возникают другие осложняющие факторы вроде взаимодействий с другими молекулами.

Подводя итог, приведу кратко основные достоинства и недостатки эволюционных алгоритмов (Слайд 20).