Лекция пятая «Вторичная структура РНК и методы её расчета»

Версия для печати

Лектор – Воробьев Денис Геннадьевич

РНК

РНК (рибонуклеиновая кислота) – биологический гетерополимер, состоящий из сахарофосфатного остова и остатков нуклеотидов: аденина (A), урацила (U), гуанина (G) и цитозина (C). В отличие от ДНК, которая обычно встречается в форме спирали, образуемой двумя нитями, РНК в подавляющем большинстве случаев является *одноцепочечной молекулой*.

РНК является «направленной» молекулой, в том смысле, что её концы химически неидентичны, и не все равно, с какой стороны читается ее последовательность. Один конец РНК называется 5’-концом, другой 3’-концом. Последовательность РНК (как и ДНК) слева направо принято записывать в направлении 5’ —> 3’.

Как правило, молекулы РНК считываются с геномной ДНК в ходе транскрипции. РНК повторяют последовательность ДНК, с которой они были считаны (при этом, T последовательности ДНК в последовательности РНК заменяется на U).

В организме РНК выполняют массу функций, в зависимости от которых они подразделяются на классы:

Виды РНК	Размер в нуклеотидах	Функция
gРНК – геномные РНК некоторых вирусов	10^4-10^5	Несут наследственную информацию
mРНК – информационные (матричные) РНК	10^2-10^5	Являются матрицами для синтеза белка
tРНК – транспортные РНК	70-90	Участвуют в синтезе белка (поставляют аминокислоты для включения в белок)
rРНК – рибосомные РНК	10^2-10^5	Участвуют в синтезе белка (являются строительными блоками рибосом)
sРНК – малые РНК	20-300	Участвуют в упаковке рибопротеиновых частиц, сплайсинге, некоторых регуляторных событиях и проч.

Вторичная структура (ВС) РНК

При определенных (нефизиологических) условиях, таких как повышенная температура (выше 60°C) или концентрация NaCl, молекула РНК ведет себя как свободно-сочлененная цепь, которая способна принимать самые разные конформации. В этом случае нельзя говорить о какой-либо устойчивой трехмерной структуре РНК. Однако, в физиологических условиях некоторые участки одной молекулы РНК могут «прилипать» друг к другу, образуя двойную правозакрученную спираль на подобие той, что известна для A-формы ДНК (см. Рис.1).

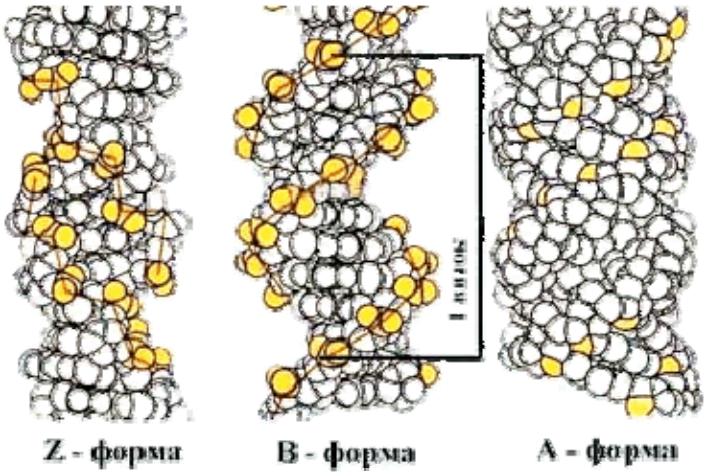


Рис.1- взято из лекций проф. Г.М. Дымшица по молекулярной биологии

Контакты, приводящие к формированию спирали, образуются за счет водородных связей: 3-х в паре G-C, 2-х в паре A-U, и 2-х в паре G-U (Рис.2). Пары G-C, A-U и G-U называются *комплементарными*. Кроме того, спирали стабилизируются т.н. стэкинг-взаимодействиями («стэк» - стопка), которые для простоты можно представить как притягивание между плоскостями соседних пар оснований в спирали.

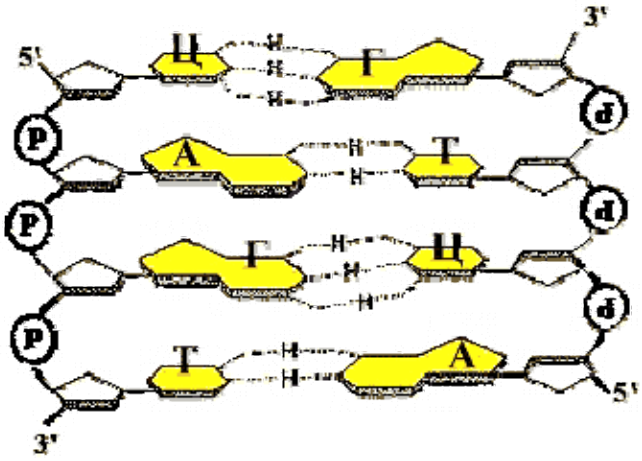


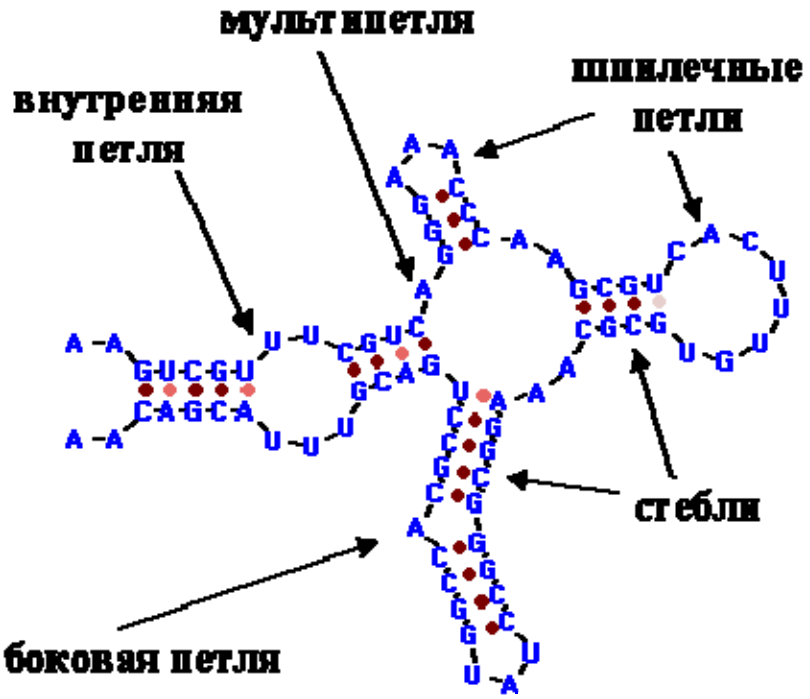
Рис.2- взято из лекций проф. Г.М. Дымшица по молекулярной биологии

В «канонической» спирали взаимодействующие цепи располагаются только в противоположных ориентациях. Пример:

```
5′ ...AUUGGCUUAG... 3′
3′ ...UAGCUGAGAC... 5′
```

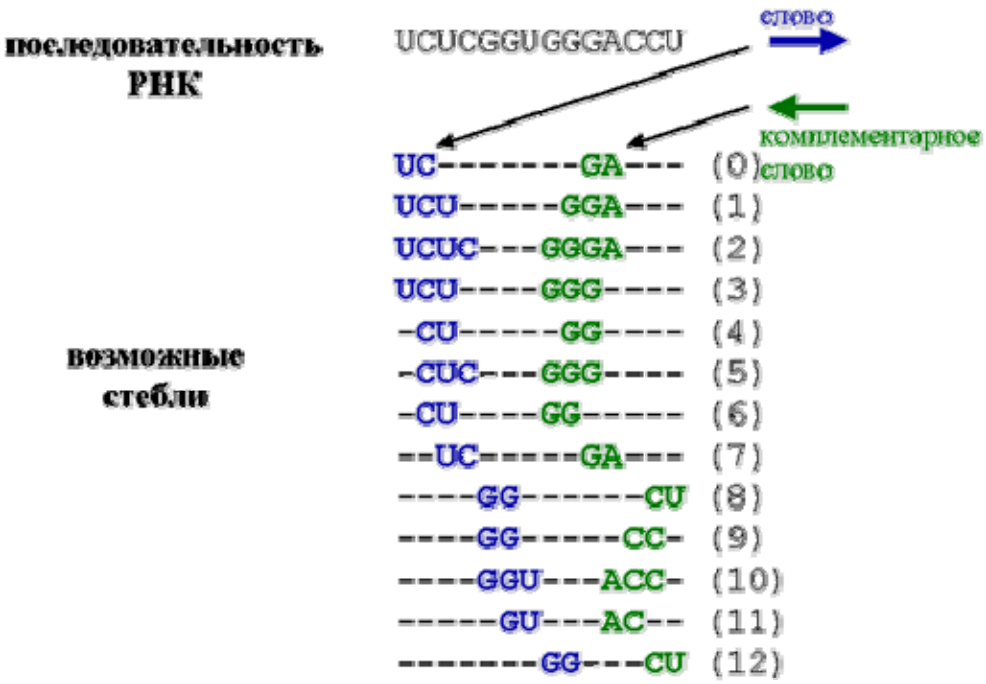
Совокупность всех непрерывных *“стеблей”* (участков спаривания) в пределах одной молекулы РНК называют *вторичной стуктурой (ВС)* этой РНК (в контексте ВС РНК термины “стебли”, “участки спаривания”, ”спирали”, ”дуплексы” часто используются как синонимы).

Одноцепочечные (неспаренные) участки молекулы называют *“петлями*”. Петли бывают трех типов: шпилечные петли, внутренние петли и мультипетли (см. рисунок).

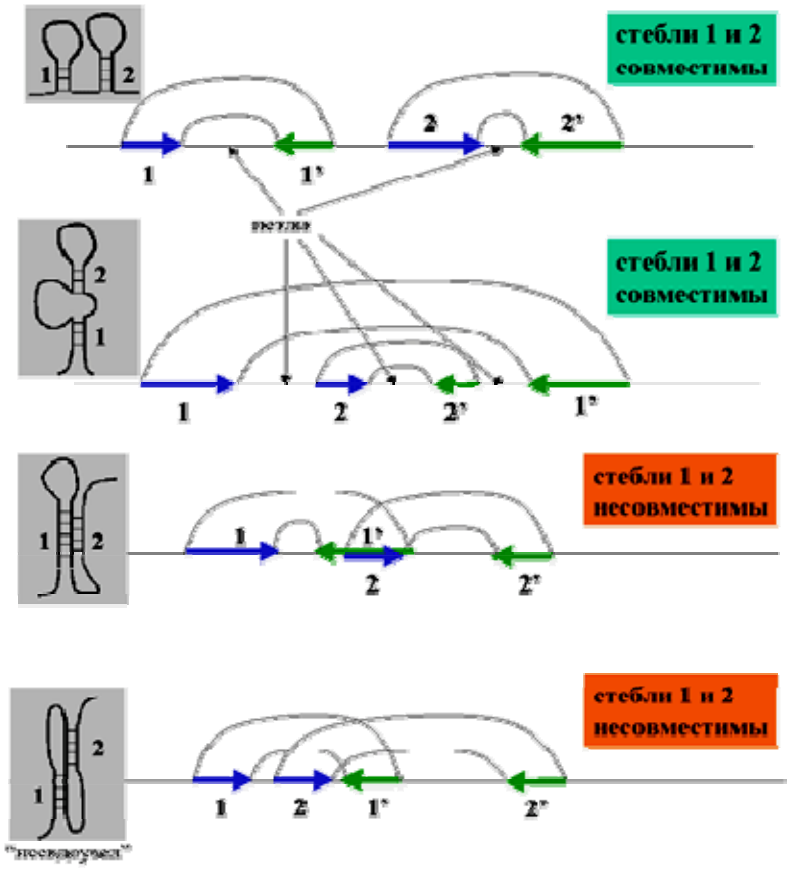


Шпилечная петля образуется одним стеблем, внутренние петли – двумя соседними стеблями, мультипетли – тремя и более соседними стеблями. Боковая петля является частным случаем внутренней петли, когда в одной из цепей образующие петлю стебли соседствуют.

Таким образом, элементом ВС является стебель. Стебель может формироваться из двух взаимно комплементарных слов, находящихся в последовательности на расстоянии не меньше 3-х нуклеотидов. Пример.



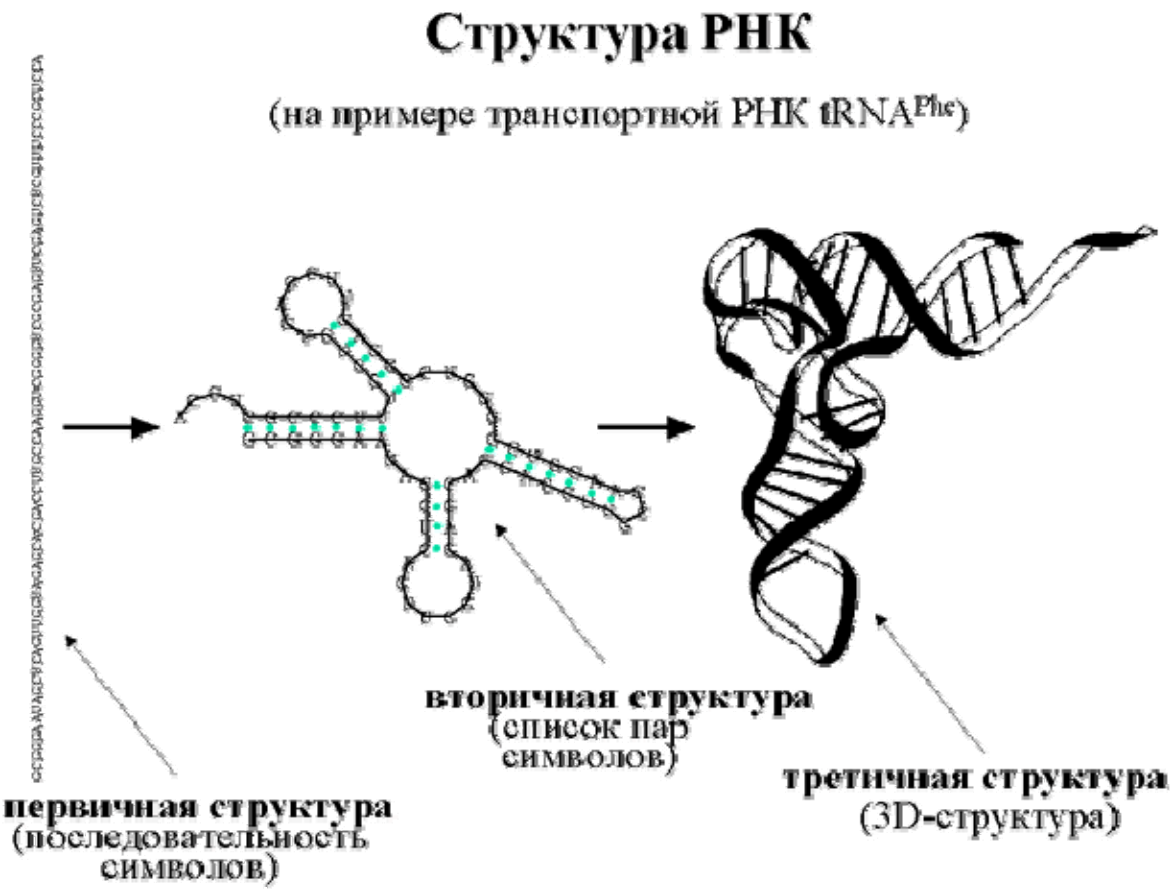
ВС однозначно описывается набором стеблей. Если рассматривать два произвольно выбранных стебля, то не все из них могут одновременно присутствовать в ВС. Правила топологической совместимости стеблей даются в следующем рисунке:



ВС, состоящая из N стеблей может существовать только тогда, когда любые 2 из этих стеблей совместимы.

Вторичная и трехмерная структура РНК

Как видно, модель ВС является серьезным упрощением модели трехмерной (3D) структры. Однако понятие ВС гораздо популярнее среди биологов, чем понятие 3D структуры, в силу следующих причин. Во-первых модель ВС подчеркивает фундаментальное для биологии понятие комплементарности. Во-вторых, для правильного протекания многих процессов в клетке важно именно то, какие нуклеотиды образуют комплементарные пары, а не точная 3D структура молекулы РНК (часто такие РНК вообще не имеют фиксированной 3D структуры). В случаях же, когда для функционирования РНК важна жестко фиксированная трехмерная конформация (например, tРНК), знание вторичной структуры облегчает задачу определения 3D структуры (поскольку на все разнообразие трехмерных конформаций наложено сильное ограничение в виде списка пар нуклеотидов, которые в пространстве должны располагаться рядом друг с другом).

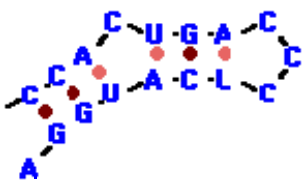


И наконец, биологи гораздо чаще имеют дело со вторичной, нежели с 3D структурой РНК, поскольку в экспериментах имеют возможность определять именно пары взаимодействующих оснований. Большинство РНК являются гибкими лабильными молекулами, вследствие чего они плохо кристаллизуется, и определить их трехмерную структуру экспериментально практически не возможно.

Способы представления ВС РНК.

Существует масса способов представления ВС РНК. Приведем некоторые из них на двух примерах: (1) на ВС короткой РНК с последовательностью CCACUGACCCUCAUGGA, и (2) ВС более длинной молекулы:

1. **Плоский рисунок.**



2. **Символьное представление.**

   Для символьного представления используются 3 и более символов. Для обозначения спаренных участков используются скобки, для петель – точки (если петли разных типов считаются неразличимыми), либо какие-то другие символы.

CCACUGACCCUCAUGGA

(((.(((...)))))). либо

(((b(((hhh)))))f

Здесь b (“bulge”) – боковая петля,

h (“hairpin”) – шпилечная петля,

f (“free end”) – свободный конец)

3. **Квадратная матрица взаимодействий.**

```
C C A C U G A C C C U C A U G G A
C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0
C 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0
A 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0
C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
U 0 0 0 0 0 0 0 0 1 0 0 0 0
G 0 0 0 0 0 0 0 1 0 0 0 0 0
A 0 0 0 0 1 0 0 0 0 0 0
C 0 0 0 0 0 0 0 0 0 0 0
C 0 0 0 0 0 0 0 0 0 0
C 0 0 0 0 0 0 0 0 0
U 0 0 0 0 0 0 0 0
C 0 0 0 0 0 0 0
A 0 0 0 0 0 0
U 0 0 0 0 0
G 0 0 0
G 0 0
A 0
```

4. **Список пар номеров взаимодействующих нуклеотидов:**

   1-16, 2-15, 3-14, 5-13, 6-12, 7-11

5. **Список стеблей.**

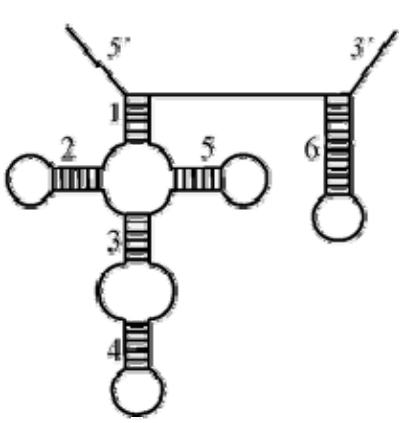
   Координатами стебля мы будем называть тройку чисел <*x,y,l*>, где *x* - координата левого конца левого плеча стебля, *y* – правого конца правого плеча и *l* – длина стебля, такую, что основания, находящиеся в позициях последовательности *x+a* и *y-a* (для всех *0<=a<l*) образуют комплементарные пары (AU, GC или GU).

   В рассматриваемом примере ВС состоит из двух стеблей:

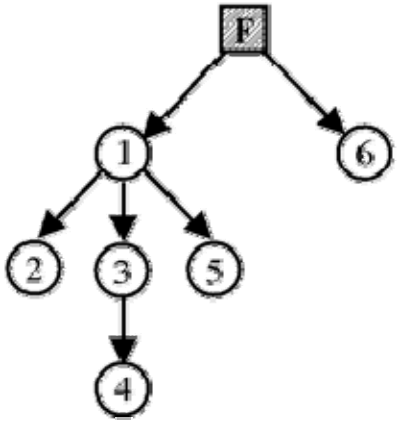
   <1,16,3>, <5,13,3>

6. **Плоский граф (дерево) (на примере длинной молекулы с разветвленной ВС).**

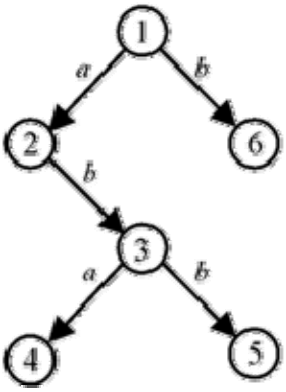
   Представление ВС (а) в виде плоского дерева произвольной степени ветвления (б), и бинарного плоского дерева (в).



а)



б)



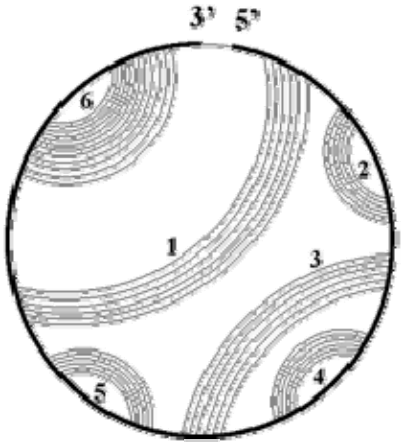
в)

Пояснение к представлению 6б: Узлам дерева соответствуют стебли, а соединяющим узлы ребрам - все петли, кроме шпилечных. Корнем дерева является фиктивный узел F, необходимый для обозначения внешних по отношению ко всем стеблям участков РНК (то есть концов молекулы и участка между стеблями 1 и 6 данного примера). Шпилькам (терминальным стеблям, замыкаемым шпилечной петлей) соответствуют терминальные вершины дерева. Каждый узел имеет произвольное количество исходящих дуг (“указателей”), указывающих на подструктуры, внутренние по отношению к данному узлу. Внутренней по отношению к стеблю с координатами *(x,y,l)* мы называем в данном случае любую позицию *a*, такую что *x+l<=a<=y-l*.

Пояснение к представлению 6в: Узлам дерева соответствуют стебли, а соединяющим узлы ребрам - все петли, кроме шпилечных. Корнем дерева является первая с 5'-конца последовательности стебель (на рисунке - стебель 1), в то время как шпилькам по-прежнему соответствуют терминальные вершины дерева. Каждый узел имеет не более двух указатей (поэтому дерево называется бинарным): первый указатель *a* (всегда левый) на поддерево (подструктуру), которое замыкается соответствующим узлу стеблем, и второй указатель *b* (всегда правый) на подструктуру, находящуюся в 3' положении по отношению к этому стеблю.

В обоих случаях, 6*а* и 6*б*, узлы и ребра должны быть помечены (указано количество нуклеотидов, формирующих соответсвующий стебель или петлю).

7. **Круговая диаграмма** (для BC из предыдущего примера).



Последовательность располагается по окружности, а взаимодействующие нуклеотиды соединяются непересекающимися линиями.

Число возможных ВС

Легко понять, что при больших длинах L последовательности РНК количество стеблей N ведет себя следующим образом:

$$N_{стеблей} \sim L^2,$$

Число же принципиально возможных вторичных структур для данной последовательности РНК растет с длиной последовательности гораздо быстрее:

$$N_{структур} \sim exp(L),$$

Приведем оценочные данные для РНК длиной 100 и 300 нт., с равным содержанием нуклеотидов разных типов: $(N_A = N_U = N_G = N_C)$:

Длина, нт.	Число возможных стеблей	Число возможных ВС
100	~ 1 000	~ 10^{25}
300	~ 10 000	~ 10^{70}

Мы видим, что число возможных для данной последовательности РНК вторичных структур астрономически велико. То, какую форму “предпочитает” принимать РНК в растворе, определяется множеством факторов, и прежде всего, свободной энергией F (или энергия Гиббса), характерной для этой формы. Вопрос об энергетике ВС и других факторах, влияющих на ВС в клетке будет рассмотрен ниже.

Термодинамика ВС РНК. Ансамбль вторичных структур.

Напомним несколько фактов из химической термодинамики.

Пусть у нас есть химическая реакция

$$A \rightleftarrows B$$

Значек ⇄ обозначает, что система находится в равновесии, то есть что скорость превращения A в B равна скорости превращения B в A. Концентрации [A] и [B], таким образом, в состоянии равновесия не зависят от времени. В этом случае их соотношение описывается равенством (статистика Больцмана):

$$\frac{[B]}{[A]} = K_{A \rightleftarrows B} = \exp\left(-\frac{\Delta F_{A \rightleftarrows B}}{RT}\right) = \exp\left(-\frac{F_B - F_A}{RT}\right),$$

где $K_{A \rightleftarrows B}$ – константа равновесия реакции A⇄B,

$\Delta F_{A \rightleftarrows B}$ – разность свободной энергии между состояниями A и B.

При $\Delta F_{A \rightleftarrows B} > 0$ равновесие смещено в сторону состояния A,

при $\Delta F_{A \rightleftarrows B} < 0$ – в сторону состояния B,

при $\Delta F_{A \rightleftarrows B} = 0$ концентрации A и B равны.

Подобные соотношения верны и для ВС РНК. Свободной энергией ВС принято называть энергию, описываемую соотношением

$$F_{BC_i} = -RT \ln\left(\frac{[BC_i]}{[BC_0]}\right),$$

где $[BC_i]$ – равновесная концентрация данной ВС, $[BC_0]$ - равновесная концентрация развернутого состояния РНК (то есть ВС, состоящей из 0 пар нуклеотидов). Энергия развернутой РНК традиционно принимается равной нулю.

Если для данной РНК возможны N структур, то в случае равновесия теоретически все они присутствуют в растворе в концентрациях, определяемых равенством

$$[BC_i] = c \exp\left(-\frac{F_{BC_i}}{RT}\right),$$

где *c* – нормировочный множитель. Таким образом, речь идет о том, что в растворе присутствует *равновесный ансамбль* вторичных струткрур. Чем ниже свободная энергия определенной ВС, тем больше вероятность встретить РНК в форме именно этой ВС.

Итак, для каждой из возможных ВС для данной РНК характерна какая-то свободная энергия *F* (в дальнейшем для краткости будет использоваться термин “энергия”). Для расчета *F* сейчас используется упрощенная модель, которая учитывает следующие обстоятельства.

(1)Энергия аддитивна. На какие бы две части мы не разбили ВС, её энергия есть сумма энергий двух этих частей. Соответсвенно, полная энергия ВС есть:

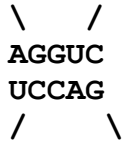
$$F_{BC} = \sum_{i=1}^{n} F_{\text{нуклеотид}}(i) + \sum_{j=1}^{m} F_{\text{пара}}(j)$$

(2) Энергия стеблей определяется в основном водородными связями и стэкинг-взаимодействиями. Считается, что на энергию образования одной комплементарной пары влияют только соседние по стеблю комплементарные пары. Такой подход

называют “моделью ближайших соседей”. Энергию стебля можно, таким образом, вычислить, зная энергии образования на конце уже существующего стебля новой комплементарной пары. Эти энергии для всех вариантов даны следующей таблицей (здесь не приводятся случаи с GU парой):

Реакция	ΔH^0, kcal/mol	ΔS^0, cal/(mol· K)	ΔF^0, kcal/mol
\ A- \ / XXXA -> XXXAA XXXU <- XXXUU / U- / \	-6.6	-18.4	-0.9
\ U- \ / XXXA -> XXXAU XXXU <- XXXUA / A- / \	-5.7	-15.5	-0.9
\ A- \ / XXXU -> XXXUA XXXA <- XXXAU / U- / \	-8.1	-22.6	-1.1
\ A- \ / XXXC -> XXXCA XXXG <- XXXGU / U- / \	-10.5	-27.8	-1.8
\ U- \ / XXXC -> XXXCU XXXG <- XXXGA / A- / \	-7.6	-19.2	-1.7
\ A- \ / XXXG -> XXXGA XXXC <- XXXCU / U- / \	-13.3	-35.5	-2.3
\ U- \ / XXXG -> XXXGU XXXC <- XXXCA / A- / \	-10.2	-26.2	-2.1
\ G- \ / XXXC -> XXXCG XXXG <- XXXGC / C- / \	-8.0	-19.4	-2.0
\ C- \ / XXXG -> XXXGC XXXC <- XXXCG / G- / \	-14.2	-34.9	-3.4
\ G- \ / XXXG -> XXXGG XXXC <- XXXCC / C- / \	-12.2	-29.7	-2.9

Пример вычисления энергии стебля:

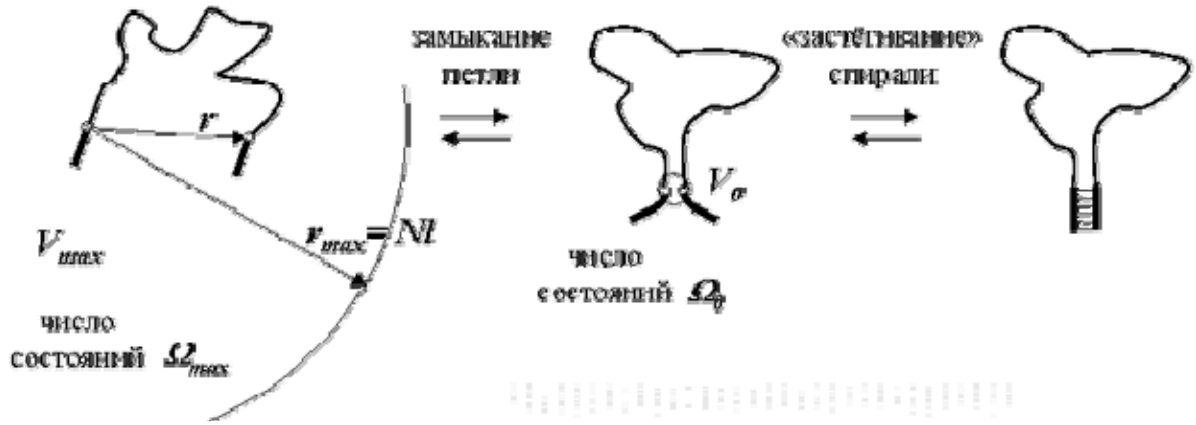


$$F_{стебля} = -1.7 - 2.9 - 2.1 - 1.7 \text{ kcal/mol.}$$

(3) Энергия петель определяется в основном энтропийной составляющей:

$$\Delta F_{петли} = \Delta H_{петли} - T\Delta S_{петли} \approx -T\Delta S_{петли}$$

Процесс образования петли качественно можно рассмотреть на следующем примере. Рассмотрим свободно-сочлененную цепочку длиной N звеньев, с длиной звена l, находящуюся в растворе. Допустим, что концы такой цепочки могут "прилипать" друг к другу. Будем считать, что остальные части молекулы не взаимодействуют.



Образование петли можно представить в виде двухстадийного процесса. На первой (медленной) стадии происходит сближение концов молекулы до расстояния, на котором возможно образование комплементарной пары. Медленным процесс является потому, что, грубо говоря, нужно дождаться, пока по случайным причинам концы цепочки попадут внутрь малого объёма V_0. Второй стадией образование продукта реакции является "застегивание" спирали по типу застежки-молнии, которое из-за высокой кооперативности этого процесса происходит сравнительно быстро. Энтропию образования петли целиком определяет энтропия первой стадии. Попытаемся оценить ее величину.

Энтропия системы, по определению, равна

$$S = R \ln \Omega,$$

где Ω - число возможных конформаций. Изменение этропии системы при замыкании петли равно

$$\Delta S_{петли} = R \ln \Omega_0 / \Omega_{max},$$

где Ω_0 – число конформаций, которые может принимать цепочка при условии, что ее концы находятся в объеме V_0, Ω_{max} – общее число возможных конформаций цепочки (концы находятся в объеме V_{max}).

Нормированная функция распределения расстояния r между концами свободно-сочлененной цепи приблизительно (кроме области, очень близкой к r_{max}) равна

$$W_N(r)dr = 4\pi r^2 (3/2\pi N l^2)^{3/2} \exp(-3r^2/2Nl^2)dr.$$

Доля конфигураций цепи, при которых концы оказываются на расстояниях, меньших a:

$$\Omega_a / \Omega_{all} = \int_0^a W_N(r)dr$$

При больших N

$$\exp(-3r^2 / 2Nl^2) \approx 1$$

Тогда, помня, что $V_0 = 4\pi a^3/3$, имеем

$$\Omega_a / \Omega_{all} = (3/2\pi Nl^2)^{3/2} \int_0^a 4\pi r^2 dr = (3/2\pi Nl^2)^{3/2} V_0$$

И, окончательно, получаем соотношение для изменения энтропии системы при образовании петли:

$$\Delta S_{loop} = -(3/2)R \ln N + R \ln[(3/2\pi l^2)^{3/2} V_0]$$

Вся зависимость от длины цепи определяется первым членом этого соотношения, второй же член остается постоянным для всех петель.

Таким образом, при фиксированной температуре свободная энергия петли равна

$$\Delta F_{loop} = \alpha \ln N + const > 0$$

Это соотношение, не смотря на приближенность использованной для вычисления модели, дает вид зависимости энергии петли от ее длины. Очевидно, что при таком рассмотрении энергия петли не зависит от последовательности нуклеотидов в ней.

Существенные отклонения от полученной зависимости есть только в области малых N, и вызваны они тем, что нуклеотиды имеют ненулевой объем. При уменьшении размера петель нуклеотиды начинают “мешать” друг другу, в следствие чего короткие шпилечные петли имеют большую энергию, чем средние (7-8 нуклеотидов), а шпилечные петли короче 3-х нуклеотидов из-за топологических затруднений вообще не образуются.

Еще раз подчеркнем интересующие нас соображения. Свободная энергия ВС аддитивна. Стебли имеют отрицательную энергию и *стабилизируют* ВС. Петли имеют положительную энергию и *дестабилизируют* ВС. Энергия стебля зависит от последовательности нуклеотидов в нем, энергия петли - нет. Баланс двух больших по модулю и противоположных по знаку величин – энергии стеблей и петель – определяет устойчивость ВС.

Задача расчета ВС РНК. Алгоритмы расчета.

Мы видим, что задача расчета ВС РНК сводится (по крайней мере, в большинстве постановок) к поиску варианта ВС с наинизшей свободной энергией. Таким образом,

она относится к классу задач оптимизации.

За более чем 30 тридцать лет с момента постановки задачи определения ВС для ее решения было предложено несколько десятков алгоритмов. Объясняется такое большое количество тем, что одни из них работают лучше на одних объектах, другие – на других. До сих пор нет универсального метода, который давал бы приемлимо малые ошибки при предсказании ВС всех РНК. Причины этого будут пояснены ниже.

Итак, основные алгоритмы можно отнести к следующим группам:

Ранние методы:

1) Полный перебор всех возможных структур.
2) Метод градиентного спуска.

Современные методы:

1) Методы динамического программирования (самые быстрые)
2) Генетические алгоритмы
3) Кинетический метод
4) Филогенетические методы

Переборные методы.

Методы полного перебора всех возможных ВС не получили широкого распространения из-за их экстенсивности. Как уже указывалось, число возможных ВС растет с длиной РНК в среднем экспоненциально, и для молекул длиной более 70-80 нт. полный перебор всех структур становится невозможным даже на очень производительных компьютерах.

Метод градиентного спуска.

Идея метода в следующем. Пусть имеется изначально развернутое состояние молекулы РНК. Множество стеблей, возможных для данной молекулы опишем как $\{h_0\}$. Выберем стебель s_0, формирование которого наиболее выгодно с энергетической точки зрения, то есть стебель, образующий наиболее стабильную ВС (при условии, что ВС состоит из одного стебля). Далее вычислим все стебли, совместимые со стеблем s_0. Они составляют множество $\{h_1\} \subseteq \{h_0\}$. Будем повторять процесс до тех пор, пока он приводит к понижению энергии ВС и пока множество стеблей, совместимых с уже использованными, $\{h_i\}$ не пусто. В результате получим ВС, состоящую из стеблей s_0, …, s_n.

Очевидно, что полученная ВС соответсвтует минимуму энергетического ландшафта данной РНК, в который можно спуститься из изначально развернутого состояния.

За исходное состояние описанного алгоритма можно принимать не только развернутое состояние, но и любой случайно построенный набор взаимно совместимых стеблей. Таким образом можно зондировать пространство состояний, но никогда нельзя гарантировать нахождение глобального минимума энергии. Именно поэтому применительно к ВС РНК метод градиентного спуска не получил широкого

распространения.

Близкая группа методов: алгоритмы симулированного отжига (simulated annealing).

Генетические алгоритмы.

Принципиально генетические алгоритмы (ГА) отличаются от методов градиентного спуска наличием рекомбинаций (см. ниже).

Приведем подробное описание алгоритма, разработанного в ИЦиГ СО РАН. Этот алгоритм является на данный момент самым быстрым в классе генетических алгоритмов, и единственным в этом классе, установленным в Internet (http://wwwmgs.bionet.nsc.ru/mgs/programs/2dstructrna).

Вообще, ГА начали применяться не в биологии, а в экономике и физике. Приведем несколько тезисов, касающихся общих принципов работы генетических алгоритмов.

- ГА способен эффективно решать задачу выбора оптимального решения из большого числа решений.

- ГА моделирует эволюцию популяции искусственных особей.
- Каждая особь характеризуется своим генотипом.
- Генотип особи состоит из генов. Гены определяют признаки организма, на основе которых оценивается приспособленность особи и осуществляется отбор.
- Отбор в популяции ведется в направлении оптимизации заданного интегрального признака (скалярной величины), определяемого совокупностью генов.
- Эволюция популяции осуществляется в результате циклического действия трех *генетических операторов*: (1) *отбора* особей по приспособленности; (2) *рекомбинаций*(скрещивания) решений, осуществляющих крупномасштабное зондирование пространства решений; (3) *мутаций*, осуществляющих локальный поиск решения.

Задача предсказания вторичной структуры РНК сводится к отысканию конфигурации молекулы с наинизшей свободной энергией, и является типичной задачей оптимизации. Поэтому ГА применим и к этой задаче. При такой постановке особь генетического алгоритма соответствует определенному варианту ВС. ВС РНК однозначно задана набором составляющих её стеблей. Поэтому практически все ГА рассматривают стебли в качестве генов.

Полный протокол алгоритма предсказания ВС РНК даётся следующей последовательностью процедур:

1. Для данной последовательности РНК построить список {h} всех возможных стеблей (ГЕНОВ);
2. Создать начальную популяцию вторичных структур РНК (особей) численностью N. Каждая вторичная структура состоит из некоторого поднабора стеблей {g} $\subseteq$ {h};
3. Вычислить энергию каждой вторичной структуры РНК в рассматриваемой популяции;
4. Подвергнуть популяцию отбору, уменьшая ее численность до заданного уровня. Отбор ведется в соответствии с приспособленностью особи (вторичной структуры), определяемой ее энергией.
5. Провести рекомбинации среди оставшихся особей и заполнить образовавшиеся

при отборе вакансии особями-потомками.
6. Провести множественные мутации (локальные изменения ВС).
7. Возвращаться к шагу 3, пока не удовлетворено одно из двух условий: (1) достигнуто заданное вырождение популяции, то есть она содержит набор сходных вторичных структур (особей - родственников); (2) проведено заданное число оптимизационных циклов (эволюционных поколений).
8. Выбрать из последнего поколения особь с наинизшей свободной энергией - в качестве результата вычислений.

Перейдем к описанию генетических операторов алгоритма.

Отбор.

Приспособленность особи (вторичной структуры) в популяции вычисляется следующим образом:

$$f_i = \exp\left(-\frac{F_i}{\Delta F}\right),$$

где F_i - свободная энергия её ВС, $\Delta F>0$ - эффективное энергетическое разрешение, то есть различие по энергии, при котором две структуры отличаются по приспособленности в e раз.

На каждом этапе селекции популяции решение об элиминации или выживании конкретной особи принимается на основе стохастической процедуры. Вероятность удаления структуры i из популяции вычисляется в соответствии с формулой:

$$p_i = \begin{cases} M\dfrac{1/f_i}{\sum_{k=1}^{N} 1/f_k}, & M\dfrac{1/f_i}{\sum_{k=1}^{N} 1/f_k} \le 1 \\ 1, & M\dfrac{1/f_i}{\sum_{k=1}^{N} 1/f_k} > 1 \end{cases}$$

Здесь параметр M ($0<M<N$) есть ожидаемое число погибающих при отборе особей. Видно, что вероятность удаления структуры обратно пропорциональна её приспособленности.

Мутации.

Общепринято, что в ГА мутациям соответствуют одиночные замены генов, обеспечивающие стохастическое движение в ближайшей окрестности В такой схеме слепого поиска мутации рассматриваются как минорный элемент ГА, так что на практике рекомендуют строго ограничивать их вклад с целью подавления вносимого ими шума.

Напротив, мутации, реализованные в нашем алгоритме, являются его важным элементом. Это обусловлено тем, что мы используем управляемый процесс мутирования. Для каждой выбранной для мутирования особи мутационный процесс

включается (или выключается) в зависимости от наличия (отсутствия) в заданной окрестности рассматриваемой вторичной структуры другой более стабильной ВС. На этой основе осуществляется управляемая (“зрячая”) локальная минимизация вторичной структуры РНК.

Выберем случайным образом из популяции заданную часть особей *N* (*N* - параметр алгоритма, в текущей версии *N*=10%), и выполним для них процедуру мутирования. Рассмотрим процедуру мутирования (множественной замены генов) у конкретной особи, состоящую из двух этапов: удаления и добавления генов (стеблей).

1. из структуры удаляется фиксированное число стеблей *S* (*S* - параметр алгоритма, в текущей версии *S*=3)
2. в структуру последовательно добавляются стебли, дающие наибольший выигрыш по энергии. Добавление осуществляется до тех пор, пока происходит понижение свободной энергиии структуры

Этап (2) описанной процедуры является операцией наискорейшего спуска, и идентичен ранним алгоритмам моделирования укладки РНК. Если полученная структура уступает по энергии начальной структуре, то результат мутаций отвергается. Таким образом, слепой поиск обычными мутациями в нашем алгоритме заменен движением между отдельными локальными минимумами, которое иногда находит глобальный минимум само по себе. То, что вместо фиксации случайных мутаций (как это делается в других ГА) в нашем алгоритме фиксируются только адаптивные мутации, сообщает ему значительное преимущество в эффективности поиска и скорости счета.

Рекомбинации.

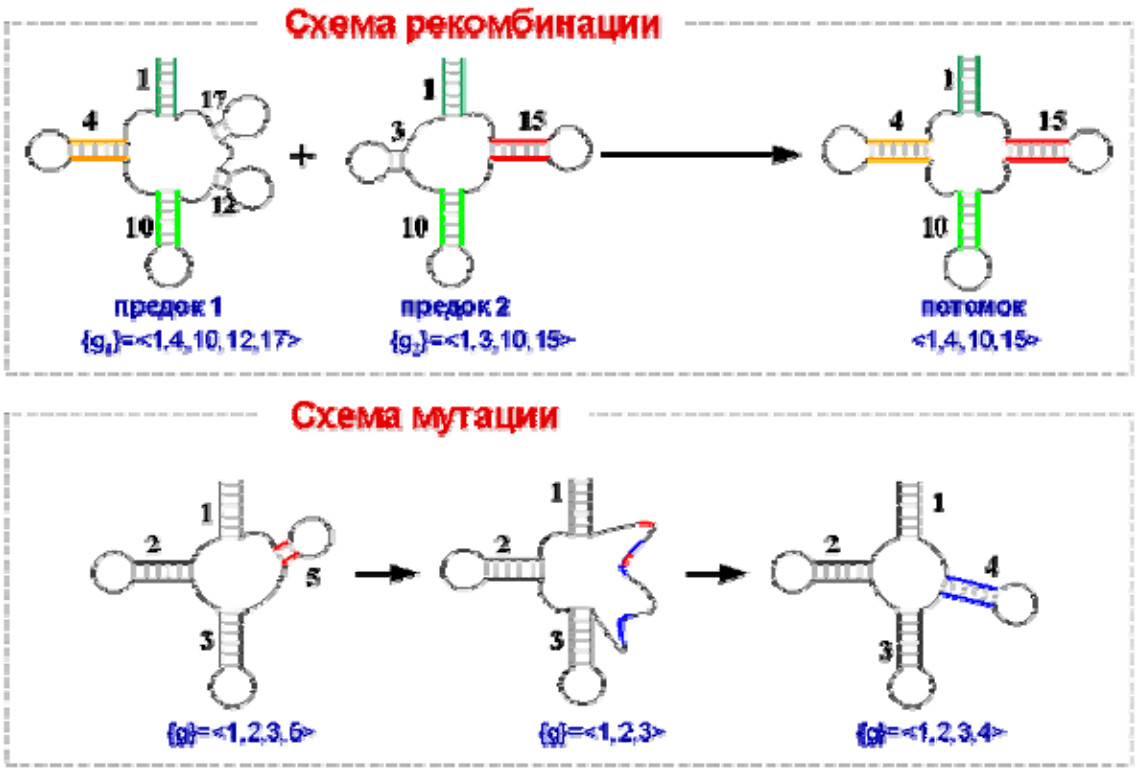
Производя крупномасштабный поиск решения, рекомбинации являются тем уникальным элементом, который отличают ГА от других алгоритмов стохастической оптимизации. Фактически, ГА представляет из себя совокупность этапов локальной минимизации за счет мутаций, и комбинирования решения из блоков счет рекомбинаций.

В нашем алгоритме рекомбинации реализованы таким образом, чтобы обеспечить равное отличие потомков от обоих родителей. Напомним, что продуктами рекомбинации заполняются вакансии, образовавшиеся при отборе. Для заполнения одной вакансии:

1. Выбираются случайным образом две структуры-родители.
2. Они сравниваются между собой, и их общие стебли образуют дочернюю структуру. Эта структура достраивается по одному случайному стеблю поочередным добавлением из родительских структур.
3. Когда все оставшиеся родительские стебли оказываются несовместимыми с уже добавленными, структура достраивается стеблями из общего списка {h} – до тех пор, пока это приводит к понижению свободной энергии структуры.

Обычно выбор точки кроссовера и размер обменивающихся фрагментов равномерно случайны, поэтому нет четкой границы между результатом мутаций и рекомбинаций.

Рекомбинации и мутации



Критерий останова.

Сильное вырождение популяции (то есть усиление сходства составляющих популяцию ВС) является критерием того, что процесс оптимизации находится в области глобального или глубокого локального минимума и в дальнейшем не выходит из него. В этой ситуации дальнейшие вычисления приводят только к накоплению копий оптимальной структуры, и поэтому процесс вычисления может быть здесь остановлен. Мы использовали следующий критерий для оценки вырожденности популяции:

$$D = \frac{2}{N(N-1)\max_i K_i} \sum_{i=1}^{N} \sum_{j=i+1}^{N} k_{ij},$$

где K_i - число стеблей в структуре i, k_{ij} - число стеблей, одинаковых в структурах i и j; N - размер популяции. D меняется в пределах от 0 до 1. При $D=0$ все особи (вторичные структуры) в популяции различны. При $D=1$популяция представлена копиями одной особи. Вычисления прекращаются после того как параметр D превысит критическое значение.

Методы динамического программирования.

Методы динамического программирования в настоящий момент наиболее распространены и популярны благодаря высокой скорости вычислений, фактическому отсутствию ограничения по длине молекулы (т.к. редко приходится вычислять ВС для РНК длинее 10000 нт.), гарантированному нахождению глобального минимума энергии ВС, и массе других достоинств. Методы основаны на применении индуктивного

(рекурсивного) подхода.

Рассмотрим принципы этого подхода на простой задаче поиска ВС с максимальным количеством спаренных оснований (вместо минимальной свободной энергии). Будем сейчас считать, что возможны шпилечные петли короче 3-х нуклеотидов (0, 1, 2, …).

Идея подхода проста. Пусть у нас имеется последовательность x длины L, состоящая из символов $x_1, \ldots, x_L$. Можно вычислить лучшую структуру $S_{i,j}$ для данного её фрагмента $x_i, \ldots, x_j$, зная лучшие структуры $S_{k,m}$ для всех её субфрагментов $x_k, \ldots, x_m$ ($i <= k < m <= j$, $m-k < j-i$).

Формально, это выглядит следующим образом. Пусть $\delta(i,j) = 1$, если x_i и x_j комплементарны, и 0 в противоположном случае. Вычислим рекурсивно матрицу $\gamma(i,j)$, $i <= j$, которая содержит в позиции i,j максимальное возможное для вторичной структуры, образуемой последовательностью $x_i, \ldots, x_j$, количество комплементарных пар.

База индукции:

$\gamma(i,j) = 0$ для любых $j <= i$

Индукция:

$$\gamma(i,j) = \max \begin{cases} \gamma(i,j-1) \\ \gamma(i+1,j-1) + \delta(i,j) \\ \max_{i<k<j} [\gamma(i,k-1) + \gamma(k,j)] \end{cases} \quad (***)$$

Первая строка в (***) отвечает ситуации, когда добавленный к $x_i, \ldots, x_{j-1}$ нуклеотид x_j не имеет возможности образовать пару ни с одним из свободных нуклеотидов с левого конца последовательности $x_i, \ldots, x_{j-1}$. Вторая строка в (***) отвечает ситуации, когда нуклеотид x_j может образовать пару с нуклеотидом x_i. Третье выражение в (***) объясняется требованием незаузленности ВС в случае образования пары $x_k x_j$ (то есть невозможности образования пар между нуклеотидами двух субфрагментов: $x_i, \ldots, x_{k-1}$ и $x_k, \ldots, x_j$).

После заполнения матрицы $\gamma(i,j)$, её элемент $\gamma(1,L)$ будет содержать максимально возможное для вторичной структуры, образуемой последовательностью x, количество комплементарных пар. Получить саму ВС с максимальным числом пар достаточно просто. Для этого нужно отследить путь по матрице γ, по которому было получено значение элемента $\gamma(1,L)$. Этим занимается алгоритм обратного просмотра. Как и заполняющий матрицу γ алгоритм, он использует выражение (***), но в обратном направлении. Структура с максимальным числом пар на $x_i, \ldots, x_j$ или содержит пару $x_i x_j$, если $\gamma(i,j) = \gamma(i+1,j-1)+1$, или распадается на две укладки: укладку на $x_i, \ldots, x_{k-1}$ и укладку на $x_k, \ldots, x_j$. Алгоритм находит это значение k и продолжает работу уже на двух субфрагментах. Рекурсия, таким образом, ведется от более длинных фрагментов к более коротким.

Легко понять, что алгоритм заполнения имеет сложность $O(L^3)$, алгоритм обратного просмотра - $O(L)$. Требуемая память растет как $O(L^2)$.

Совершенно аналогично описанному должен выглядеть алгоритм, оптимизирующий не

число, а *суммарную энергию спаренных оснований*, но игнорирующий (приравнивающий нулю) энергии петель. В этом случае max заменяется на min, $\gamma(i,j)$ заменяется на $F_{стеблей}(i,j)$, а $\delta(i,j)$ на $f(i,j)$ – изменение энергии, вызванное добавлением к структуре новой пары $x_i x_j$.

Если оптимизировать *свободную энергию, зависимую от петель*, то придется использовать двойную рекурсию. Это связано с тем, что стабилизирующая энергия образуемой пары $x_i x_j$, может не компенсировать проигрыш в энергии, создаваемый возможно образуемой петлей, хотя в дальнейшем, при добавлении пар $x_{i-1}x_{j+1}$, $x_{i-2}x_{j+2}$, … или других суммарное понижение свободной энергии все-таки может быть достигнуто.

Это ограничение обходится построением в дополнение к матрице $F(i,j)$, содержащей энергию наилучшей структуры, матрицы $F_c(i,j)$, которая содержит энергию наилучшей структуры с замкнутой парой $x_i x_j$. Очевидно, всегда выполняется $F_c(i,j) >= F(i,j)$. Если пара $x_i x_j$ невозможна по правилам комплементарности, то полагается $F_c(i,j) = \infty$. Полагается также $F_c(i,i) = F(i,i) = \infty$.

Правила заполнения $F(i,j)$ становятся следующими:

$$F(i,j) = \min \begin{cases} F(i+1,j) \\ F(i,j-1) \\ F_c(i,j) \\ \min_{i<k<j} [F(i,k-1) + F(k,j)] \end{cases}$$

Построение матрицы $F_c(i,j)$ несколько более сложно, но также подчиняется рекурсивным соотношениям, что делает и этот алгоритм достаточно быстрым. При этом, правда, необходимо использовать такие упрощения, как независимость энергии петли от контекста или линейную (вместо логарифмической) зависимость энергии петли от длины.

Недостатки методов, минимизирующих энергию.

Как показывает практика, даже с введением в самые последние (1999 год) энергетические правила зависимости энергии петель от их последовательности, зависимости энергии стэкинга не только от соседних, но и от более дальних нуклеотидов, и других сложных уточнений, общая точность (определяемая, как процент правильно предсказанных пар в ВС) минимизационных алгоритмов до сих пор не превышает 70%.

Ошибки термодинамических алгоритмов обычно вызваны одной из трех основных проблем: (1) неточностью энергетических правил, (2) невозможностью учесть энергетику третичных или РНК-белковых взаимодействий и (3) игнорированием особенностей кинетики фолдинга (процесса образования ВС). Поясним сказанное.

Плотность уровней энергии различных вариантов ВС крайне высока. Искажение энергетического ландшафта, вызванное неточностью энергетических правил, или какими-то внешними факторами, например, влиянием клеточных белков на ВС, приводит к тому, что реальный глобальный минимум при расчете таковым не оказывается, а в качестве решения выбирается другой минимум, кажущийся в

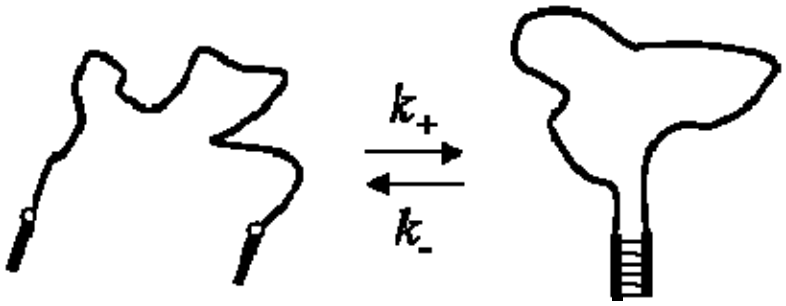
результате искажения ландшафта глобальным.

Что касается кинетики процесса укладки ВС, то дело здесь в следующем. Термодинамические подходы основаны на редко выполняющемся в живых клетках требовании равновесности ансамбля ВС (хорошим примером является система аттенюации аминокислотных оперонов; и таких примеров масса). Многие реальные РНК просто не достигают глобального оптимума энергии в течение жизни в клетке (от долей секунд до часов). Это можно пояснить на примере. Вообще говоря, процесс укладки проходит через последовательное формирование и распад многих стеблей. Расчетное характерное время распада 5-тинуклеотидной GC-спирали составляет порядка пяти часов. Время распада 8-тинуклеотидной GC-спирали составляет уже приблизительно 1500 лет (!). Ожидать, что клеточный ансамбль ВС РНК будет являться равновесным, при таких величинах не приходится.

Кинетический алгоритм.

Для процесса формирования ВС в реальном времени используются кинетические алгоритмы. Приведем здесь несколько соображений, на которых основывается этот класс алгоритмов.

Рассмотрим опять свободно-сочлененную цепочку, способную формировать простейшую ВС, состоящую из нескольких комплементарных пар на конце:



A **B**

Как мы помним из химической кинетики, термодинамическая константа равновесия реакции может быть выражена через константы скорости прямой и обратной реакции:

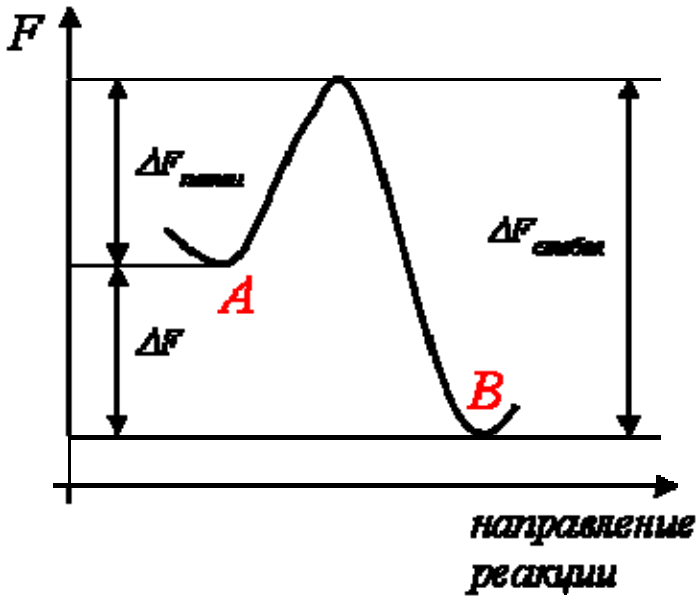
$$\frac{[B]}{[A]} = K_{равн} = \frac{k_+}{k_-}$$

Напомним формулу Аррениуса из теории переходного комплекса. Скорость преодоления системой энергетического барьера зависит от величины этого барьера $E^{\neq}$ следующим образом:

$$k = A_0 \exp\left(-\frac{E^{\neq}}{RT}\right),$$

где A_0 – некая константа.

Если мы посмотрим на диаграмму состояний нашей химической системы:



то поймем, что энергетическим барьером образования вторичной структуры, обозначенной как состояние B, является свободная энергия петли. Барьером же обратной реакции перехода из состояния B в A является свободная энергия стебля. Таким образом, формулы Аррениуса принимают вид:

$$k_+ = A_0 \exp\left(-\frac{\Delta F_{\text{образующейся\_петли}}}{RT}\right) \quad k_- = A_0 \exp\left(-\frac{\Delta F_{\text{образующегося\_стебля}}}{RT}\right)$$

Здесь подразумевается, что $\Delta F_{\text{образующейся\_петли}} > 0$, $\Delta F_{\text{образующегося\_стебля}} < 0$.

С помощью этих соотношений можно моделировать кинетику фолдинга РНК. Для этого, осуществляется такая последовательность действий:

(1) Для имеющейся вторичной структуры состоящей из N стеблей

$$BC_0 = \{h_1, h_2, h_3, \ldots, h_N\}$$

находятся все возможные подсписки из N-1 стебля. Они с необходимостью являются потенциальными BC. Для каждой такой BC_i в соответствии с приведенными формулами расчитываются константы k_i скоростей реакций

$$BC_0 \text{ -> } BC_i$$

Аналогично строятся все возможные BCj, состоящие из N+1 стебля, из которых N присутствуют в BC0 и совместимы с вновь добавляемым N+1-м стеблем. Расчитываем константы скорости *kj* всех возможных переходов

$$BC_0 \text{ -> } BC_j$$

(2) На основании констант скорости расчитываются вероятности всех указанных переходов, и методом Монте-Карло случайно разыгрывается то, какой переход к

соседней с $ВС_0$ по пространству состояний структуре произойдет.

Время перехода определяется как реализация случайной величины *t*, распределенной по закону Пуассона:

$$f(t)=k\cdot exp(-kt), \qquad k=\sum k_i+\sum k_j$$

(3) Осуществляется переход в новое состояние и сдвиг времени на величину *t*.

(4) Моделирование осуществляется до тех пор, пока суммарное время не превысит заранее заданную величину *T*.

(5) Моделирование по пунктам 1-4 повторяется N раз, в зависимости от объёма статистики, который необходимо набрать.

В качестве исходного состояния для п.(1) можно брать развернутое состояние. Тогда процесс отвечает моделированию отжига РНК. Если же включить в модель синтез (удлинение со временем) молекулы РНК, можно просчитывать процесс формирования ВС в ходе транскрипции.

В результате применения этого алгоритма получается *кинетический ансамбль* ВС РНК – набор вторичных структур с приписанными вероятностями, зависящими от времени.

Эволюционные подходы.

Остановимся кратко на группе наиболее мощных, однако редко применимых подходов – алгоритмах реконструкции эволюционно устойчивых ВС РНК по выборке нескольких последовательностей родственных РНК.

В их основу положено предположение, что функционально сходные РНК, хотя и имеют разную нуклеотидную последовательность, формируют сходные ВС. Например, все транспортные РНК должны укладываться одинаково. То же касается 5S, 16S, 18S и других рибосомных РНК, малых ядерных РНК, самосплайсирующихся интронов, лидеров вирусных мРНК, и многих других классов РНК.

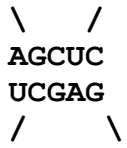
Одним из основных преимуществ анализа выборки родственных РНК является возможность поиска так называемых коадаптивных замен. Поясним это на примере. Допустим, в биологически активной форме ВС РНК-предка присутствовал стебель

```
\    /
AGGUC
UCCAG
/    \
```

В ходе эволюции, допустим, в РНК могла произойти замена одного нуклеотида другим (мутация). Например, после мутации стебель стал выглядеть так:

```
\ C  /
AG UC
UCCAG
/    \
```

Эта замены привела к локальному нарушению стебля ВС. Если этот стебель не важен для функции данной РНК, то такая мутация могла сохраниться в ряду поколений. Если же он важен, то в ряде случаев особи-носители этой мутации могли погибать, а мутация имела мало шансов зафиксироваться в ряду поколений. Единственными выжившими особями, несущими такую мутацию, могли бы стать те из них, у которых произошла еще одна мутация, восстанавливающая целостность нарушенного стебля:



Такие парные замены в РНК называют *коадаптивными*. Их поиск и учет значительно облегчают задачу предсказания ВС, общей для выборки РНК.

Второе преимущество анализа выборки вместо единичной последовательности – взаимная верификация стеблей, присутствующих одновременно в разных РНК выборки. Это значит, что ошибочно предсказанные (допустим на основе минимизации энергии) в одной из РНК стебли не могут существовать в соответствующих местах других РНК. Если стебель предсказан в одной последовательности но невозможен по правилам комплементарности в других последовательностях, или же обладает в них слишком низкой стабильностью, это первый сигнал о том, что этот стебель, скорее всего не имеет отношения к биологической действительности.

Главной сложностью филогенетических подходов является выяснение столбцов позиций в последовательностях, которые соответствуют друг другу в ВС. Универсальных методов решения этой проблемы пока не предложено.