

# Структура РНК

Миронов Андрей Александрович  
Факультет биоинженерии и биоинформатики МГУ

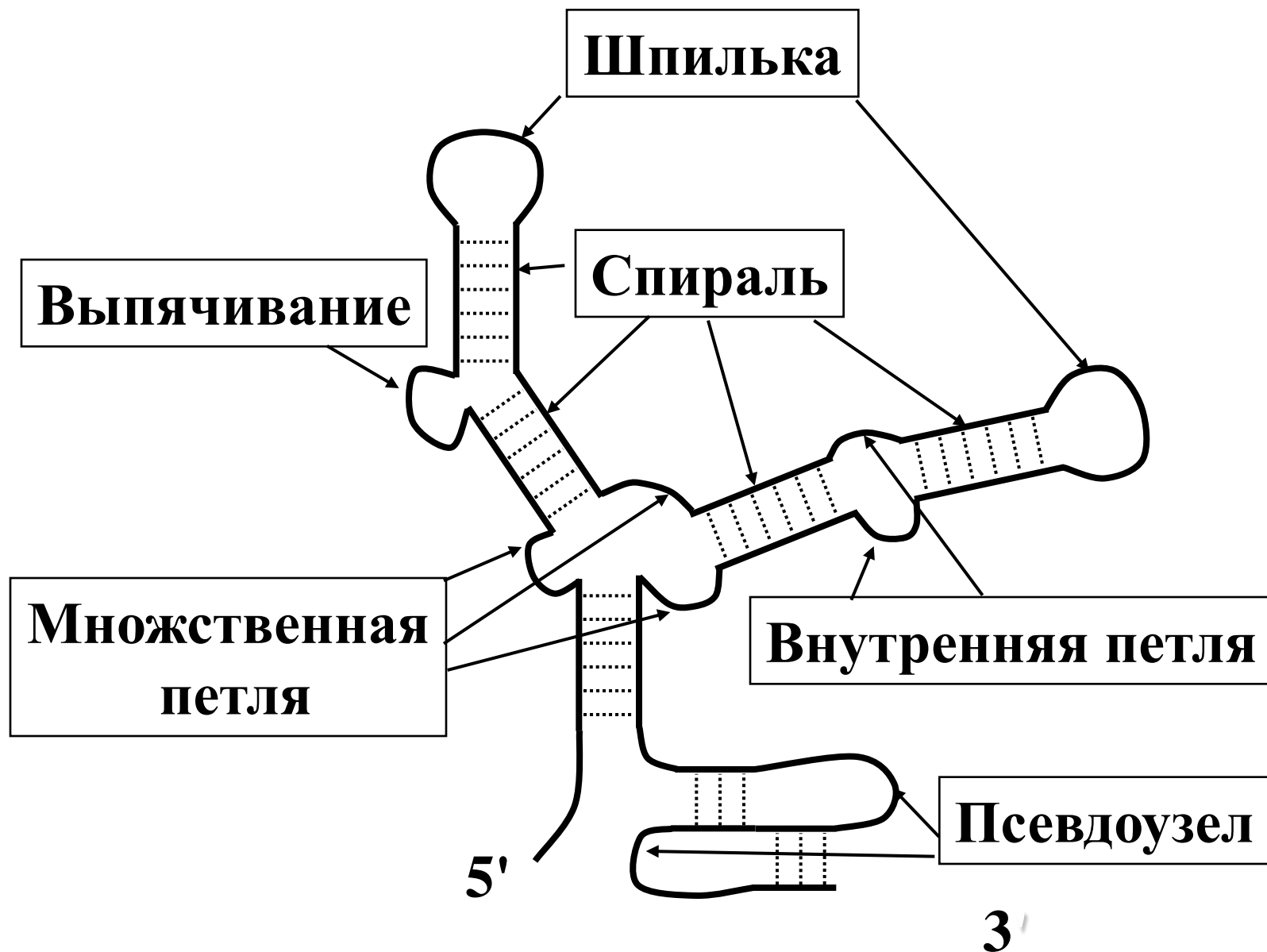
Рощино-2014

# Вторичная структура РНК

- ♦ Вторичной структурой называется совокупность спаренных оснований
- ♦ Биологическая роль вторичной структуры:
  - рибосомная
  - Модификации рРНК: sno
  - тРНК
  - Инициация трансляции: IRES
  - аттенюация
  - Доза гена: XIST / TSIX, rho1 / rho2
  - tmRNA
  - Копийность плазмид: RNA1 / RNA2
  - микроРНК
  - Теломеразная РНК
  - Рибозимы
  - Рибопереключатели

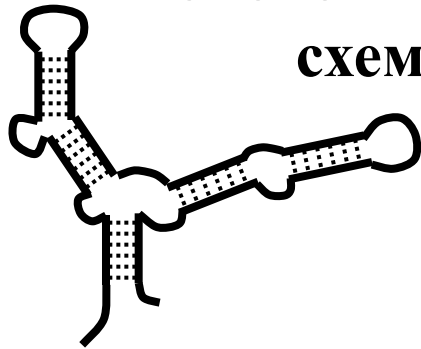
Много чего еще ...

# Элементы вторичной структуры

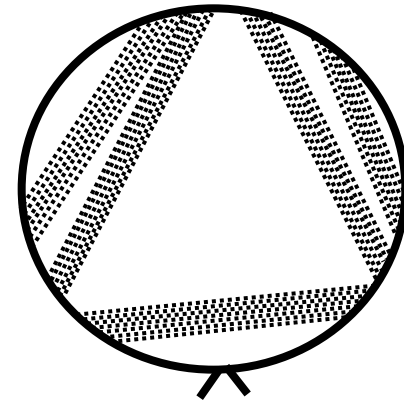


# Способы представления вторичных структур

**Топологическая  
схема**



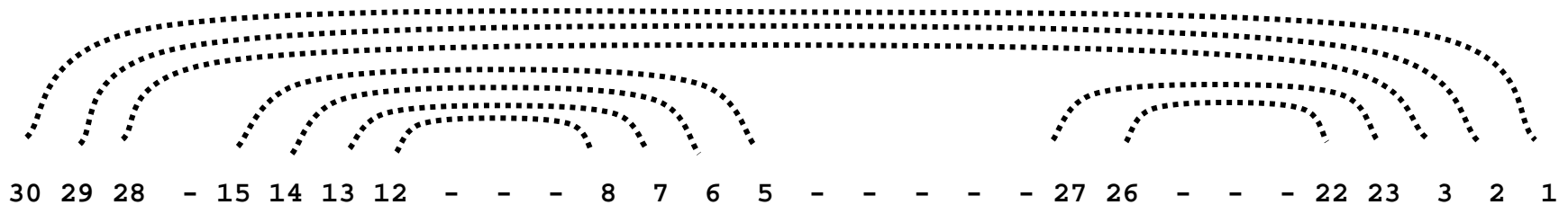
**Круговая диаграмма**



**Список  
спиралей**

		from <sub>1</sub>	to <sub>1</sub>	from <sub>2</sub>	to <sub>2</sub>
A		1	3	28	30
B		5	8	12	15
C		21	22	26	27

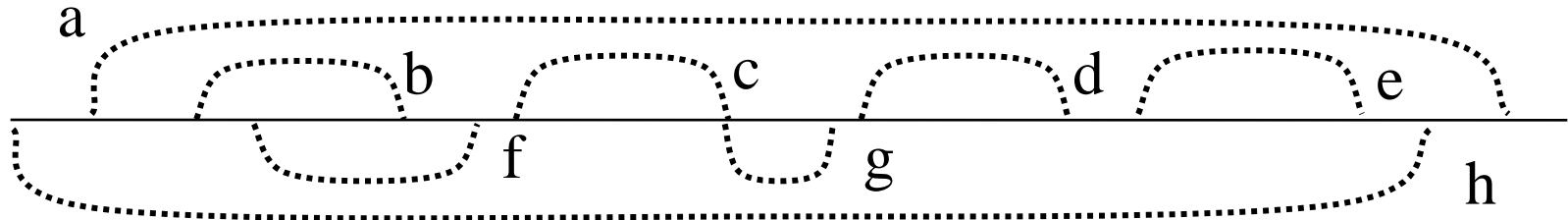
**Массив спаренных оснований**



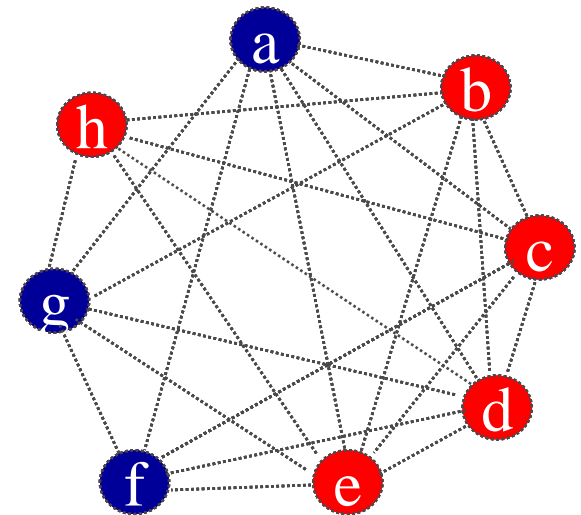
# Задача

- Дана последовательность.
- Найти правильную вторичную структуру.
- Золотой стандарт: тРНК, рРНК.
- Количество возможных вторичных структур очень велико.
- Дополнительные ограничения:
  - Нет псевдоузлов. (На самом деле они очень редки и энергетически невыгодны)
- Количество возможных структур все равно очень велико
- Надо найти *оптимальную* структуру. А что оптимизировать? Как оптимизировать?

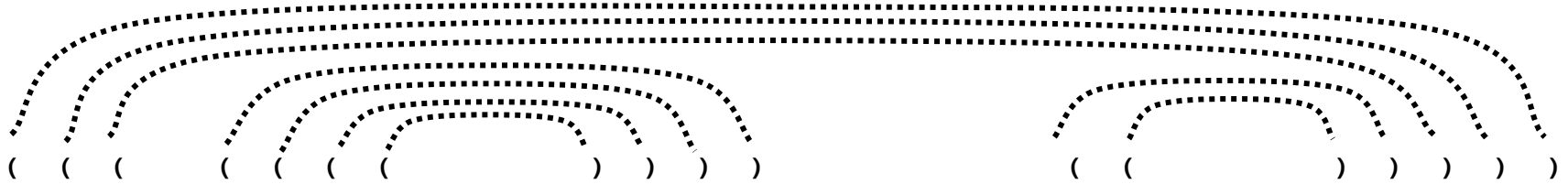
# Комбинаторный подход



- Построим граф:
  - вершины – потенциальные нуклеотидные пары (или потенциальные спирали)
  - Ребро проводится, если пары совместимы (не образуют псевдоузлов и не имеют общих оснований)
- Допустимая вторичная структура – клика в ЭТОМ графе



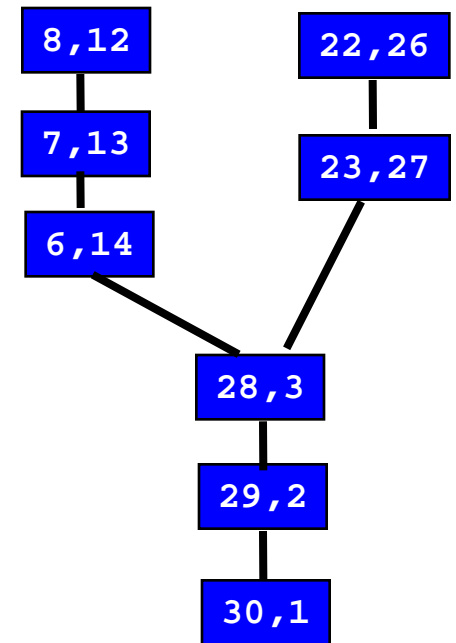
# Структуры без псевдоузлов



- Структура без псевдоузлов = правильное скобочное выражение
- Может быть представлено в виде дерева
- Оценка количества возможных структур:

$$T(L) \approx 1.8^L$$

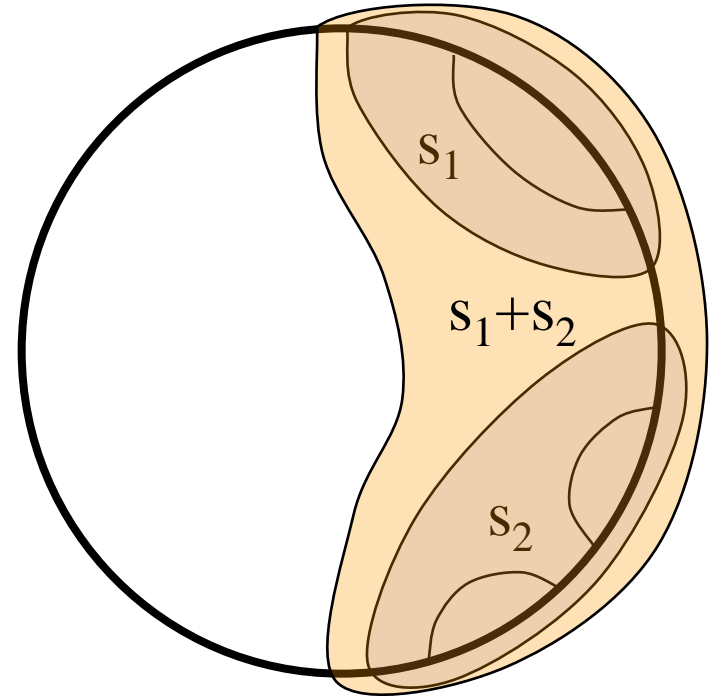
(очень много)



# Оптимизация количества спаренных оснований

- Обозначим  $|s|$  - мощность структуры (количество спаренных оснований)
- Пусть  $s_1$  и  $s_2$  две непересекающиеся структуры (структуры без общих оснований)
- Тогда

$$|s_1 + s_2| = |s_1| + |s_2|$$

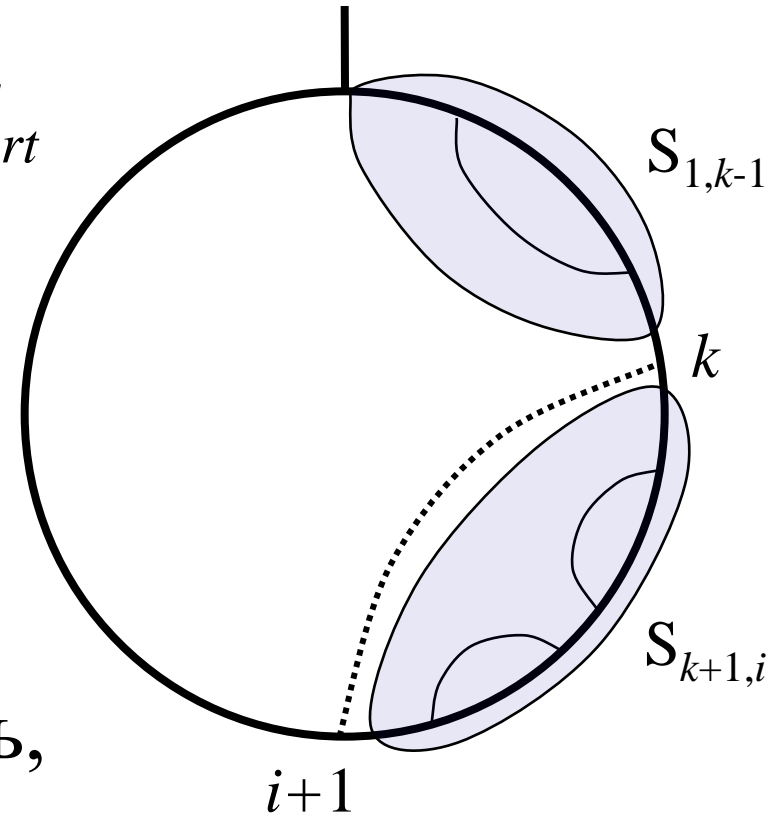




# Оптимизация количества спаренных оснований

## Nussinoff

- Пусть нам известны оптимальные структуры  $S_{rt}$  для всех фрагментов  
$$i \leq r \leq t \leq j$$
- Тогда можно найти оптимальную структуру для сегмента  $[i, j+1]$
- Для этого нам надо понять, спаривать ли основание  $j+1$ , и, если спаривать, то с кем



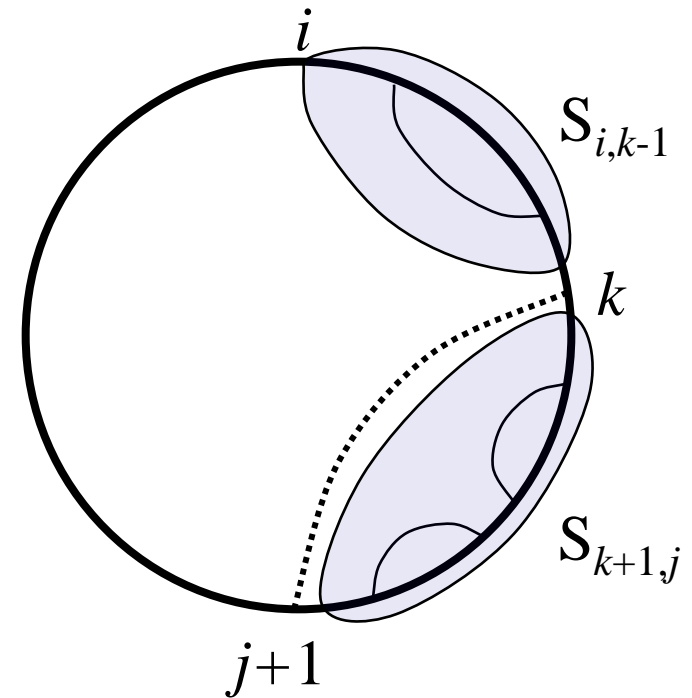
# Динамическое программирование для количества спаренных оснований

Количество спаренных оснований в оптимальной структуре  $S_{i,j+1}^*$  определяется как максимум:

$$S_{i,j+1}^* = \max \left\{ \begin{array}{l} S_{i,j}^*; \text{ (нет спаривания)} \\ \max_k (S_{i,k-1}^* + S_{k,j}^*) + 1; \\ \text{(} k \text{ спаривается с } j+1 \text{)} \end{array} \right\};$$

Время работы алгоритма:

$$T \approx O(L^3)$$



# Динамическое программирование для количества спаренных оснований

- При поиске оптимального количества спаренных оснований заполняется треугольная матрица весов  $S_{i,j}$ ,  $i < j$ .
- Обозначим  $\pi_{ij}$  – номер основания, с которым надо спарить основание  $j$  при анализе сегмента  $[i, j]$ , или 0, если не надо спаривать. При оптимизации запоминаем треугольную матрицу спаривания (аналог матрицы обратных переходов)



# Восстановление структуры по матрице спаривания

```
SearchStruct (int i, int j)
{
    int i0=i, j0=j;
    while(i < j){
        if( $\pi_{ij} == 0$ ) j--;
        if( $\pi_{ij} != i$ ) i++;
        if( $\pi_{ij} == i$ )
        {
            StorePair(i,j);
            SearchStruct (i0, i-1);
            SearchStruct (i+1, j0);
            return;
        }
    }
}
```

# Энергия структуры РНК

# Энергия вторичной структуры

- Энергия спиралей
- Энергия петель (энтропия)

Энергия спирали рассчитывается как сумма энергий стэкингов

	AU	CG			
AU	-2	-3.2			
CG	-3.2	-4.8			
GC	-3.7	-4.5			

A – U

C – G

A – U

G – C

C – G

$$\Delta G = -3.2 -3.2 -3.7 -4.5$$
$$= - 14.6$$

# Энергия петель

- Энергия свободной цепи

$$\Delta G = V + 3/2 kT \ln L$$

- Для шпилек при  $L=3..5$  кроме энтропии есть некоторое напряжение структуры.
- Для внутренних петель и для мультипетель  $L$  – суммарная длина петель + количество ветвей.
- Параметр  $V$  зависит от типа петли
- Для выпячивания сохраняется стэкинг.
- Обычно используют не формулу, а таблицы.



# Минимизация энергии

*Обычное динамическое программирование не проходит – нет аддитивности.*

## Определения

- нуклеотид  $h$  называется доступным для пары  $i \bullet j$ , если **НЕ** существует спаривания  $k \bullet l$ , такого, что

$$i < k < h < l < j$$

- Множество доступных нуклеотидов для пары  $i \bullet j$  называется петлей  $L_{ij}$ , а пара  $i \bullet j$  называется замыкающей парой. Частный случай петли – стэкинг.
- Энергия структуры рассчитывается как сумма энергий петель (в том числе и стекингов):

$$\Delta G = \sum e(L_{ij})$$

# Алгоритм Зукера

- Введем две переменные:
  - $W(i,j)$  – минимальная энергия для структуры на фрагменте последовательности  $[i, j]$ ;
  - $V(i,j)$  – минимальная энергия для структуры на фрагменте последовательности  $[i, j]$  при условии, что  $i$  и  $j$  спарены;
- Рекурсия:

$$V(i,j) = \min_{i < i_1 < j_1 < i_2 < j_2 < \dots < i_k < j_k < j} \left( \Delta G_{loop(i, i_1 \dots)} + \sum_l^k V(i_l, j_l) \right)$$

$$W(i,j) = \min \left\{ \begin{array}{l} W(i+1, j), i \text{ не спарено} \\ W(i, j-1), j \text{ не спарено} \\ V(i, j), i, j \text{ спарены} \\ \min_{i < k < j} (W(i, k) + W(k+1, j)), \\ i, j \text{ спарены с кем-то} \end{array} \right.$$

# Алгоритм Зукера

- Рекурсия для  $W$  требует времени

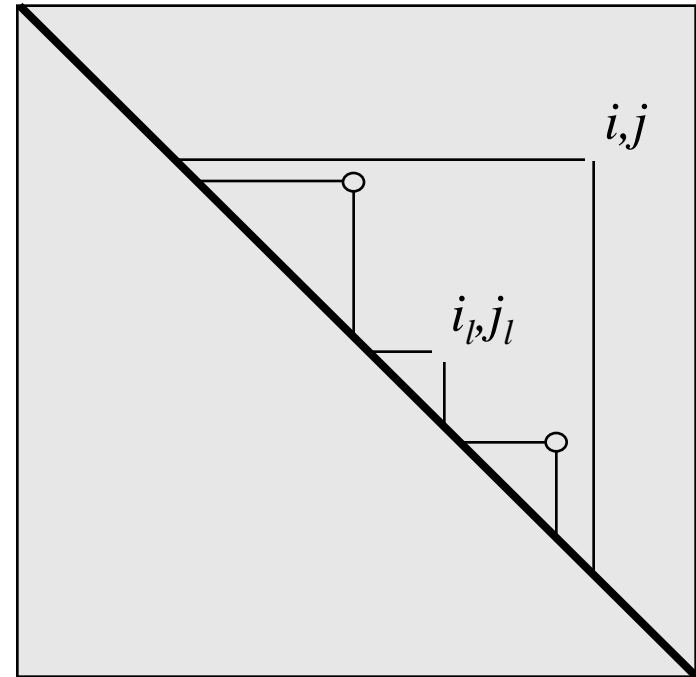
$$T \approx O(L^3)$$

- Рекурсия для  $V$  требует гораздо большего времени

$$T \approx O(2^L)$$

- Причина – мультипетли. Можно:

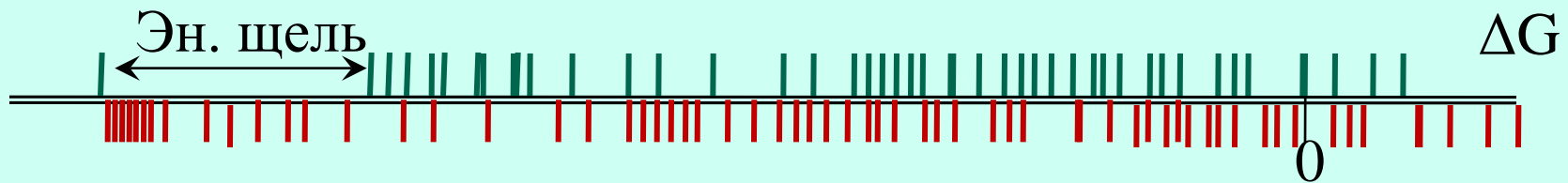
- Ограничить размер или индекс мультипетель
- Применить упрощенную формулу для их энергии
- Просматривать мультипетли только если  $i+1, j-1$  не спарены.
- Применить приближенную эвристику



# Проблемы минимизации энергии

- Только около 60% тРНК сворачиваются в правильную структуру
- Энергетические параметры определены не очень точно. Более того, в клетке бывают разные условия, и, соответственно, реализуются разные параметры.
- Находится единственная структура с минимальной энергией, в то время как обычно существует несколько структур с энергией, близкой к оптимальной.

## Физическая природа проблемы



# Решение проблем

- Использовать экспериментальные данные
- Искать субоптимальные структуры
- Искать кинетически доступные структуры
- Искать эволюционно консервативные структуры.
  - структуры тРНК и рРНК определены именно так

# Поиск субоптимальных структур и структурных элементов

- Статистическая сумма

$$Z = \sum \exp(-\Delta G_i / kT)$$

- Если мы просуммируем по всем структурам, содержащим данную пару, то мы можем оценить ее значимость (чем  $Z$  больше, тем более значимым является спаривание)
- Для подсчета  $Z$  алгоритм динамического программирования, заменив  $\min$  на суммирование, а сложение на умножение.
- Больцмановская вероятность того, что нуклеотиды  $i, j$  спарены равна:

$$P(i, j) = \frac{\sum_{\text{все струк.: } i \circ j} \exp(-\Delta G/kT)}{Z}$$

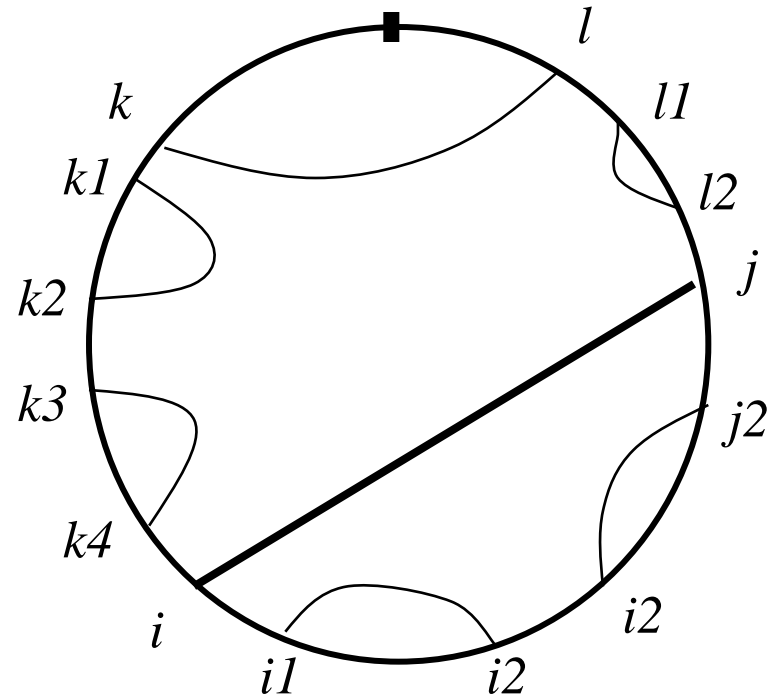
- Разыгрываем пары оснований в соответствии с этой вероятностью и восстанавливаем соответствующие субоптимальные вторичные структуры.

# Вычисление условных стат.сумм

$$Z(ij) = \sum_{\text{структуры}} \exp(-\Delta G_{\text{loop}}) =$$

$$\sum_{\text{внешн}} \exp(-\Delta G_{\text{out}}) \cdot \sum_{\text{внутр}} \exp(-\Delta G_{\text{in}}) =$$

$$I(ij) \cdot O(ij)$$

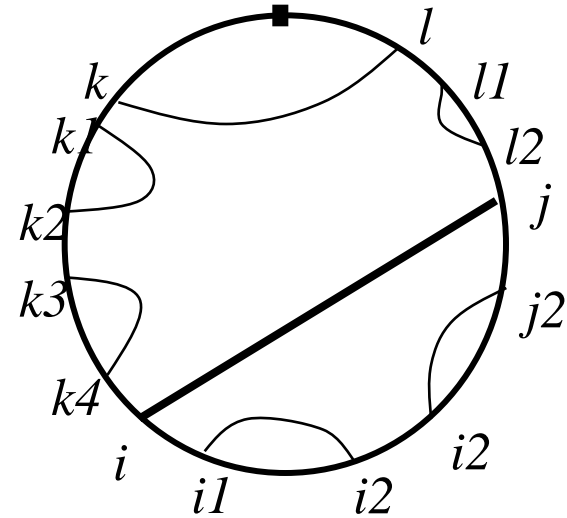


Полная сумма:

$$Z = I(0, L)$$

# Внутренняя и внешняя суммы

$$I(ij) = \sum_{j < j_1 < i_1 < j_2 < i_2 < \dots < i} \exp(-\Delta G_{j,j_1,i_1,\dots} - \Delta G_{j_1,i_1}^{in} - \Delta G_{j_2,i_2}^{in} - \dots) = \sum_{j < j_1 < i_1 < j_2 < i_2 < \dots < i} \exp(-\Delta G_{j,j_1,i_1,\dots}) \cdot \prod_s I(j_s i_s)$$



$$O(ij) = \sum_{l < l_1 < \dots < j < i < k_1 < k_2 < \dots < k} \exp(-\Delta G_{k,l_1,\dots,j,i,k_1,\dots,k} - \Delta G_{kl}^{out} - \Delta G_{l_1,l_2}^{in} - \dots) = \sum_{l < l_1 < \dots < j < i < k_1 < k_2 < \dots < k} \exp(-\Delta G_{j,j_1,i_1,\dots}) \cdot O(kl) \cdot \prod_s I(l_s l_{s+1}) \cdot \prod_s I(k_s k_{s+1})$$



# Поиск структур в геноме

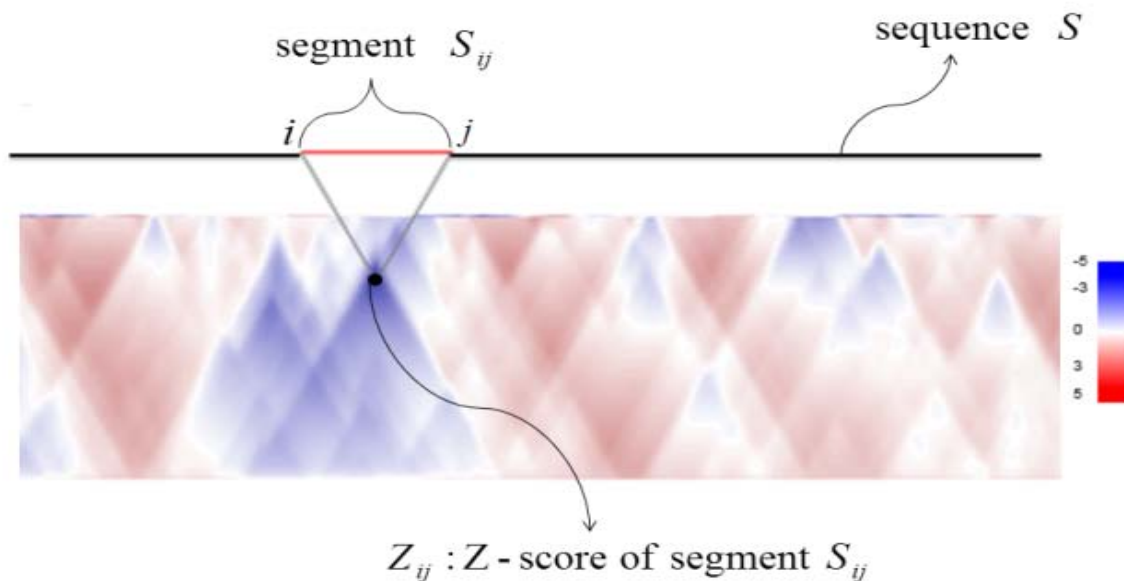
Программа rfold (Vienna package) идет скользящим окном (заданного размера)

- Проблема – размер окна

# RNA surface

$E(\Delta G) \sim l; D(\Delta G) \sim l;$

Можно вычислить Z-score:  $Z(i, j) = \frac{\Delta G(i, j) - E_{\Delta G}(l)}{\sigma_{\Delta G}(l)}$



<http://bioinf.fbb.msu.ru/RNASurface/>

**Консенсусные  
вторичные  
структуры РНК**

# Основные задачи

- **Построение консенсуса**

- Дано: набор последовательностей для которых известно, что они имеют общую вторичную структуру (например, тРНК или регуляторный элемент)
- Описать общую структуру

- **Поиск консенсуса**

- Дано: описание консенсуса.
- Найти в данной последовательности (например, в геноме) все случаи встречи консенсуса

# Метод ковариаций

- Пусть дано множественное выравнивание последовательностей

- Взаимная информация двух колонок:

$$I(A,B) = \sum_{\alpha\beta} f_{AB}(\alpha\beta) \log_2 \{f_{AB}(\alpha\beta) / (f_A(\alpha) \cdot f_B(\beta))\}$$

$f_{AB}(\alpha\beta)$  – частоты одновременной встречи буквы  $\alpha$  в колонке  $A$  и буквы  $\beta$  в колонке  $B$ .

$f_A(\alpha)$  – частота встречаемости буквы  $\alpha$  в колонке  $A$ .

$f_B(\beta)$  – частота встречаемости буквы  $\beta$  в колонке  $B$ .

- Пары колонок с высоким значением взаимной информации с большой степенью вероятности образуют комплементарную пару (если высоки совместные частоты для пар букв AT, CG)
- Для восстановления вторичной структуры можно использовать алгоритм Нуссинофф, приписывая в качестве весов пар значение взаимной информации.