Компьютерные алгоритмы для предсказания вторичной структуры РНК, часть 2: термодинамический и сравнительный подходы.

к.ф.-м.н. Титов И.И.

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia

Рекурсивный алгоритм для расчета вторичной структуры РНК (Zuker, Stiegler, 1981 - алгоритм mfold)

- расчет минимальной энергии
- восстановление структуры, обладающей этой энергией

Упрощенно:

энергия вторичной структуры есть сумма энергий комплементарных взаимодействий

$$E(S) = \sum_{i,j \in S} e(r_i, r_j).$$

Расчет минимальной энергии вторичной структуры:

$$W = min\, W_s$$

$$\begin{aligned} W_{ij} &= 0 \text{ for } j - i < 4 \\ W_{ij} &= min\left\{W_{i+1,j}, W_{i,j-1}, e(i,j) + W(i+1, j-1),\ min_{k=i+1}^{j-1}\left(W_{i,k} + W_{k+1,j}\right)\right\} \end{aligned}$$

- i и j не вовлечены в комплементарное взаимодействие

либо - связаны с основаниями внутри подпоследовательности

либо - связаны друг с другом

Псевдокод для рекурсивного вычисления минимальной энергии вторичной структуры РНК

```
for ( d = 1 . . . n )
    for ( i = 1 . . . d )
        j = i + d
        C [ i , j ] = MIN (
            Hairpin (i , j ) ,
            MIN ( i < p < q < j : Interior ( i , j ; p , q ) + C [ p , q ] ) ,
            MIN ( i < k < j : FM [ i + 1 , k ] + FM [ k + 1 , j - 1 ] + cc ) )
        F [ i , j ] = MIN ( C [ i , j ] , MIN ( i < k < j : F [ i , k ] + F [ k +1 , j ] ) )
        FM [ i , j ] = MIN ( C [ i , j ] + ci , FM ( i + 1 , j ) + cu , FM [ i , j - 1 ] + cu ,
                             MIN ( i < k < j : FM [ i , k ] + FM [ k +1 , j ] ) )
free_energy = F [ 1 , n ]
```

F[i,j] - минимальная энергия для подпоследовательности от **i** до **j. C[i,j]** - энергия для **i** и **j** - пар. **FM** введен для описания мультипетель (энергетический вклад **F=cc+ci*I+cu*U**, где **I** - число внутренних комплементарных пар и **U** - число неспаренных оснований в петле). Энергетические параметры остальных петель вычисляются в функции **Interior (i , j ; p , q),** определяющей энергетический вклад петли образованной двумя парами оснований **i - j** и **p - q**. Временная сложность алгоритма - O(n^4), она снижается до O(n^3), если ввести ограничение на размер внутренних петель.

Алгоритм статистических сумм

(McCaskill, 1991 - Венский пакет)

Свободные энергии вторичной структуры РНК

$$\Delta G_{stack} = \Delta H_{37,stack} - T\Delta S_{37,stack}$$

$$\Delta G_{loop} = -T\Delta S_{37,loop}$$

Статистическая сумма

$$Q = \sum_{S \in \mathcal{S}} e^{-\frac{\Delta G_S}{RT}}.$$

Вероятность структуры S

$$\exp(-\Delta G_s / RT) / Q$$

Статсумма

$$Q = \sum_{all\_structures\_S} e^{-\frac{\Delta G(S)}{RT}}$$

Рекурсия для статсумм

$$\begin{aligned} Q_{ii} &= Q_{i,i+1} = 1, \\ Q^I_{ii} &= Q^I_{i,i+1} = 0. \\ Q_{ij} &= Q_{i+1,j} + \sum_{k=i+1}^{j-1} Q^I_{ik} Q_{k+1,j} + Q^I_{ij} \\ Q^I_{ij} &= e^{-\frac{\epsilon(i,j)}{RT}} Q_{i+1,j-1} \end{aligned}$$

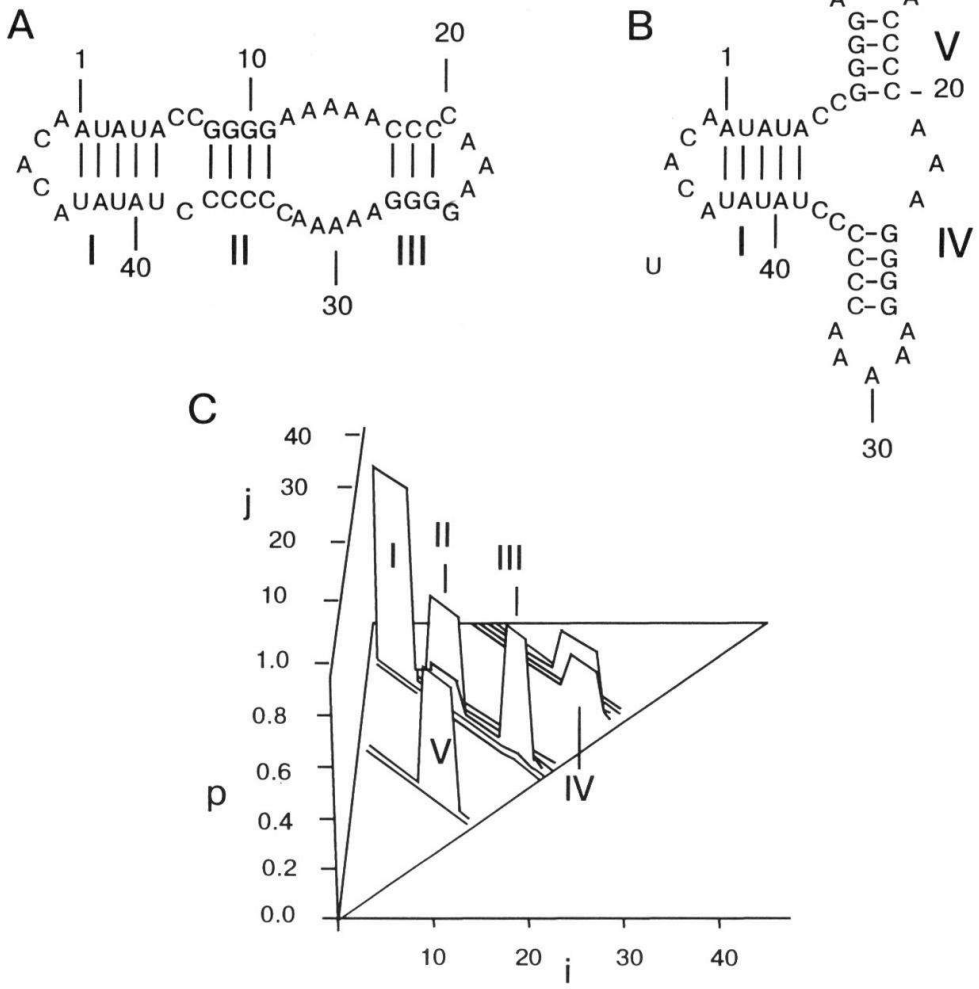
Вероятность наблюдения комплементарного взаимодействия ij

$$\Pr(r_i - r_j) = \frac{Q^I_{ij} Q^I_{j,i+n}}{e^{-\frac{\epsilon(i,j))}{RT}} Q_{1,n}}.$$

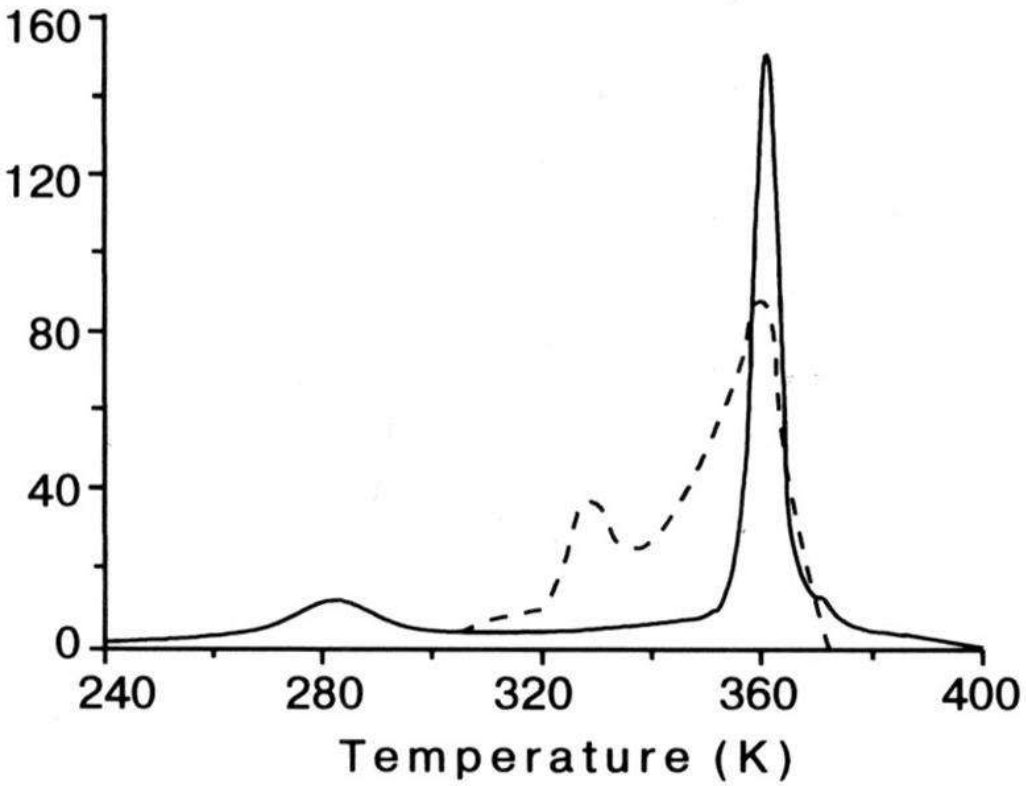
Псевдокод для рекурсивного вычисления статистической суммы вторичной структуры РНК

```
for ( d = 1 . . . n )
    for ( i = 1 . . . d )
        j = i + d
        QB [ i , j ]  =  EHairpin (i , j ) +
            SUM ( i < p < q < j  :  EInterior ( i , j ; p , q ) * QB [ p , q ]  ) +
            SUM ( i < k < j  :  QM [ i + 1 , k - 1 ] * QM1 [ k , j - 1 ] * Ecc )
        QM [ i , j ] =
            SUM (  i < k < j  : ( Ecu ^ ( k - i ) +QM [ i , k - 1 ] ) * QM1 [ k , j ] )
        QM [ i , j ] = SUM (  i < k <= j  : ( QB [ i , k ] * Ecu ^ ( j - k ) * Eci )
        Q [ i , j ] = 1 + QB [ i , j ] +
                        SUM ( i < p < q < j  : Q [ i , p - 1 ] * QB [ p , q ]  )
partition_function =  Q [ 1 , n ]
```

Ex:=exp(-x/RT) определяет веса распределения Больцмана. **Q[i,j]** - статсумма подпоследовательности от **i** до **j**. Массив **QM** - статсумма подпоследовательности при условии что **i** и **j** образуют комплементарную пару. QM и QM1 используются для вычисления вкладов мультипетель.

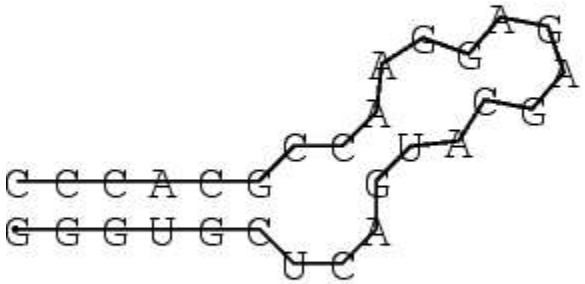


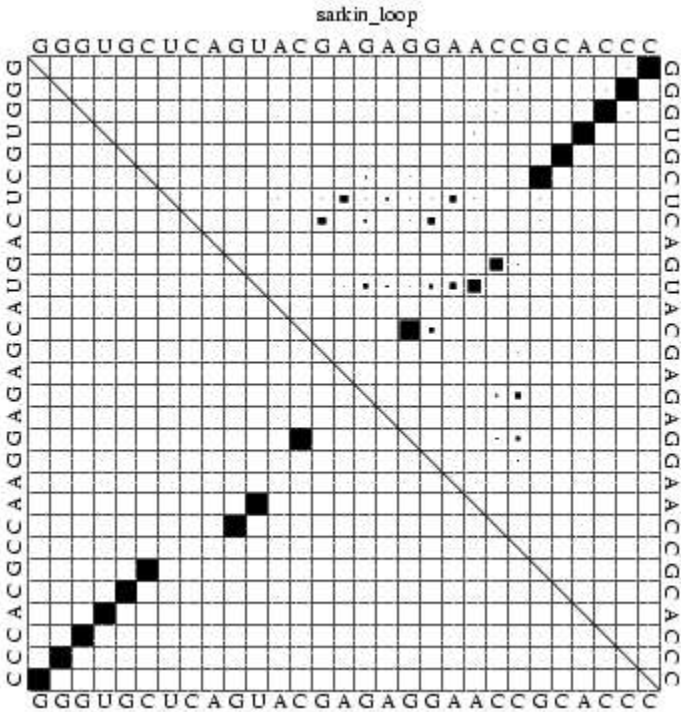
Представление равновесного распределения вторичных структур круговой РНК длиной 48 нуклеотидов. Римские цифры I - V обозначают спирали, идентичные для каждого представления. (A) Оптимальная вторичная структура при 35°C. (B) Первая субоптимальная вторичная структура. (C) Трехмерная диаграмма комплементарных пар при 35°C.



Удельная теплоемкость 5S РНК *E. coli* как функция температуры (сплошная линия - расчетная кривая, пунктир - экспериментальная кривая для А-формы 5S РНК)

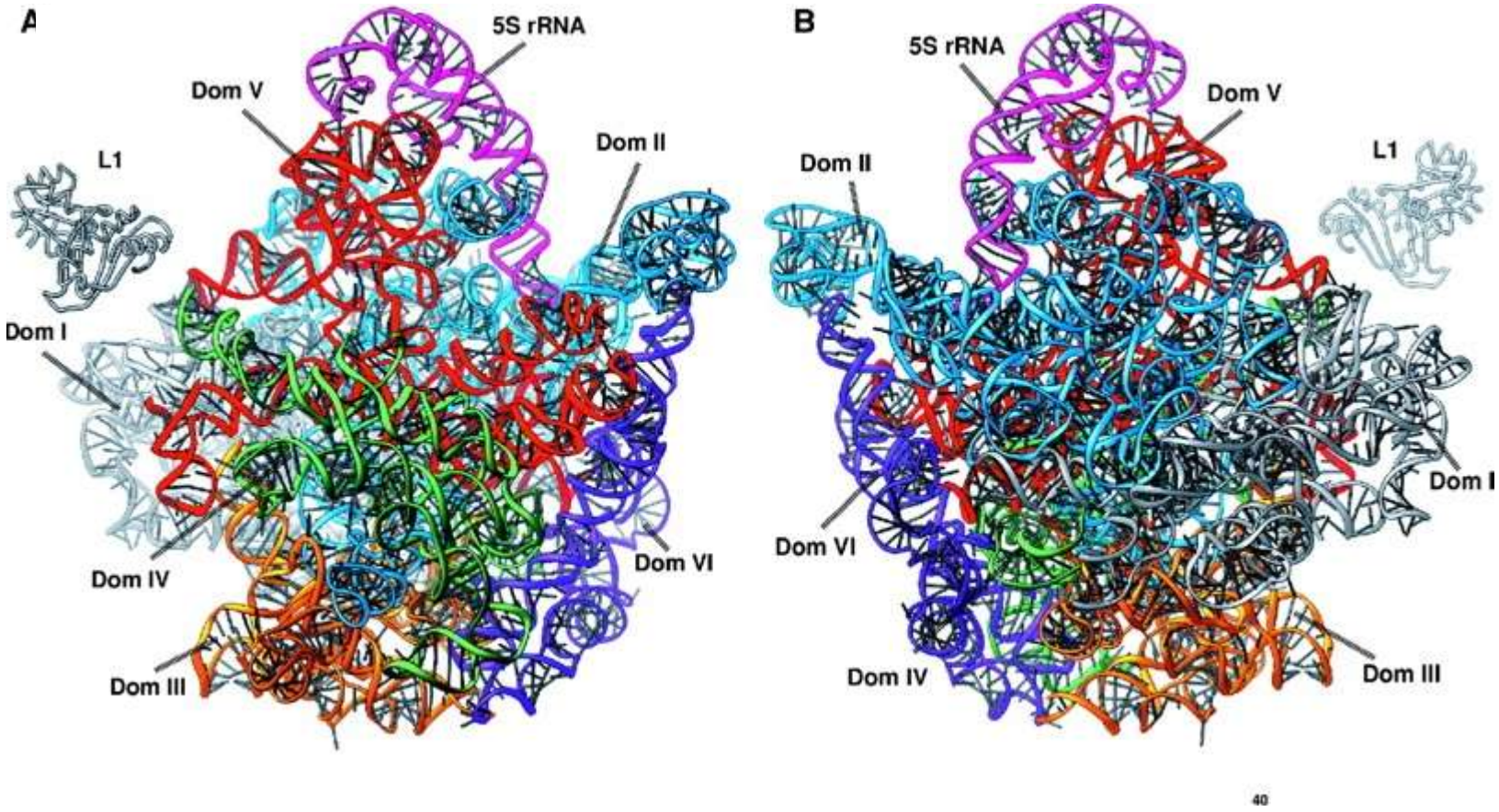
Сарциновая петля 23S рибосомальной РНК: Структура с минимальной энергией и вероятности спаривания оснований





Большая субъединица рибосомы

Haloarcula marismortui
Ban et. al., Science 289:905-920, 2000



Достоинства и недостатки рекурсивных алгоритмов

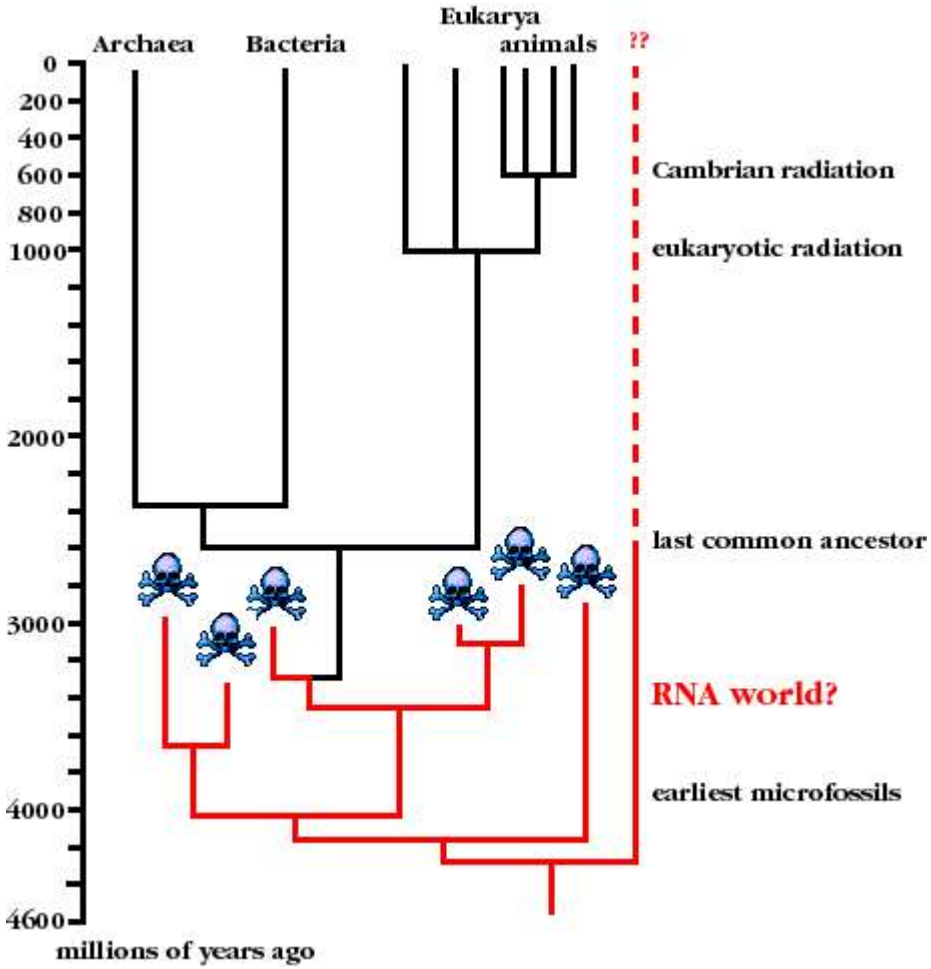
- обновление энергетических параметров (mfold)
- вычисление вторичной структуры с ограничениями (mfold)
- высокая скорость (mfold, Vienna package)
- вычисления через Интернет (mfold)
- вычисление физических свойств РНК (mfold, Vienna package)

Недостатки

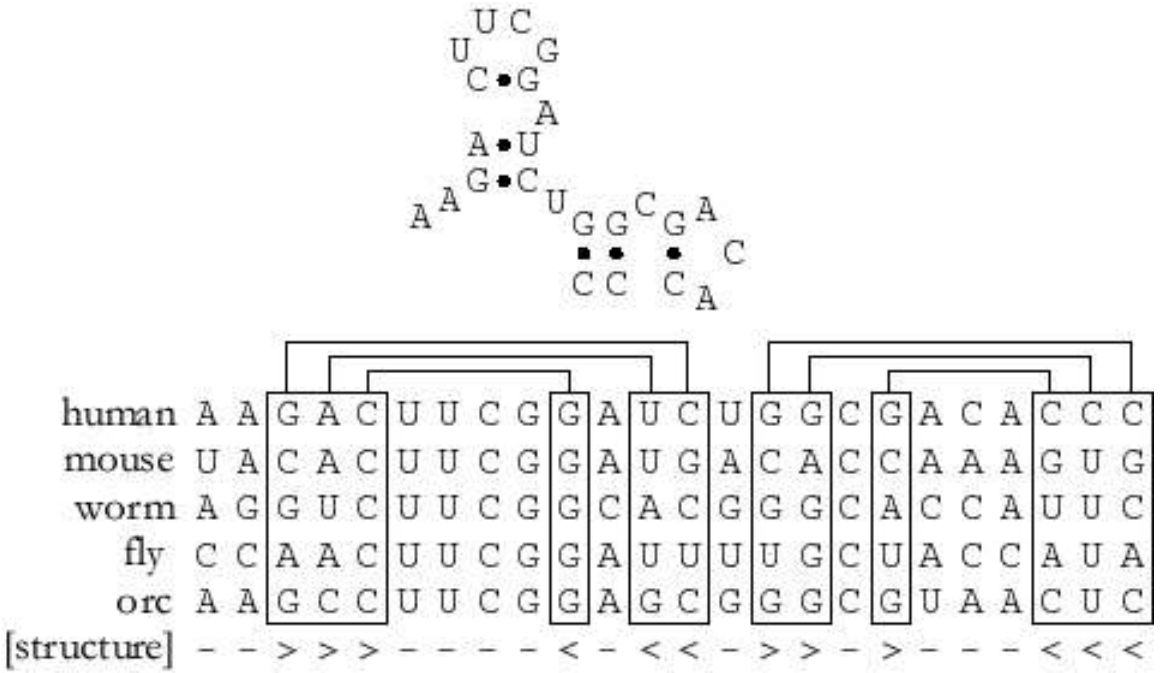
- точность зависит от знания энергетических параметров; пренебрегают сложными взаимодействиями, в том числе с белками
- основаны на предположении о тепловом равновесии РНК

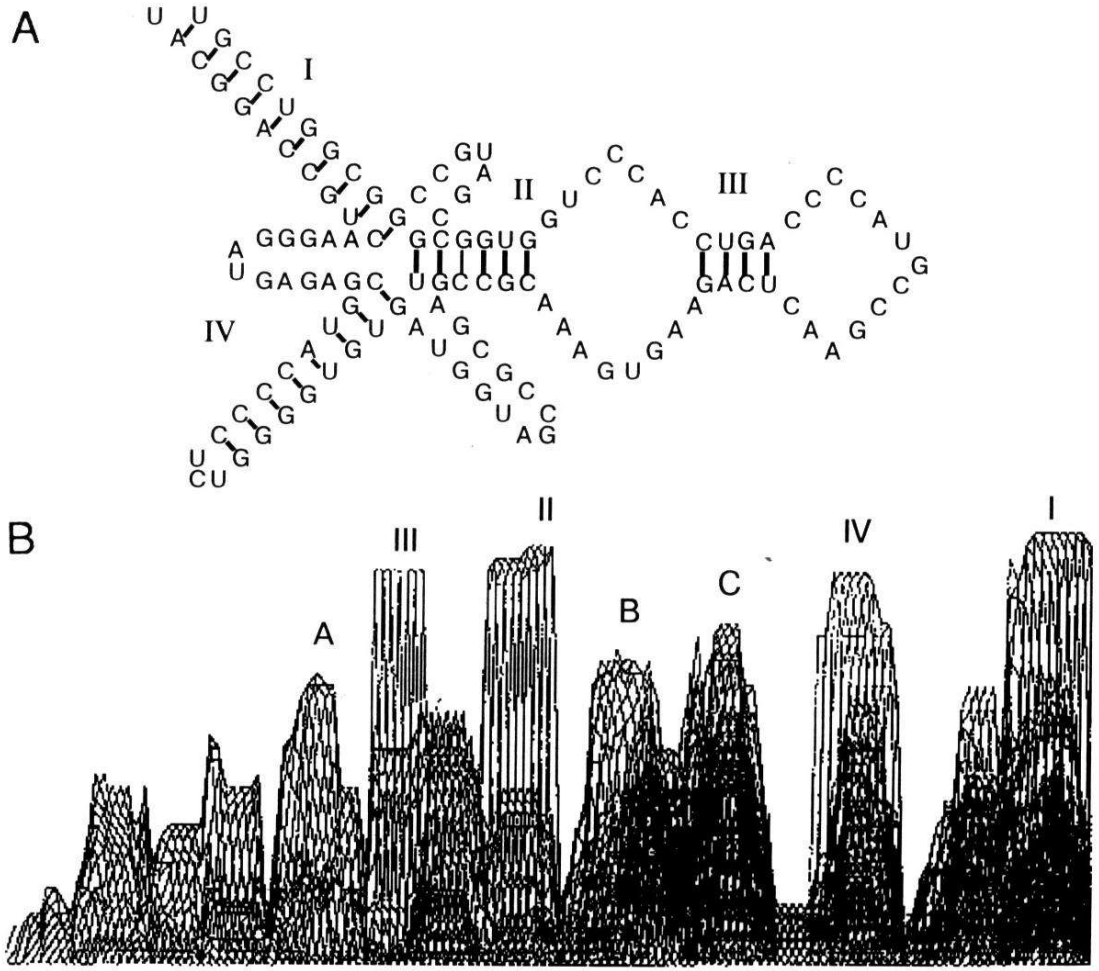
Древний Мир РНК

Gesteland & Atkins, The RNA World, *CSHL Press, 1999*



СТРУКТУРА РНК: блоки парных корреляций

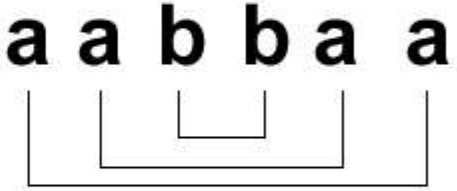




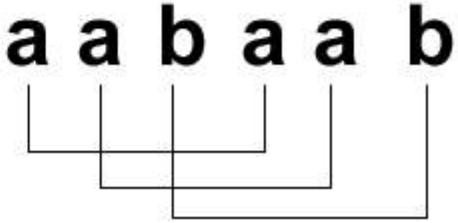
Реконструкция вторичной структуры 5S РНК с помощью сравнения последовательностей *E. coli* и др. родственных организмов. (A) Вторичная структура 5S РНК *E. coli* (B) Суперпозиция пиков спиральности для ряда выровненных последовательностей

Regular language: a b a a a b

Palindrome language:



Copy language:



В отличие от регулярных(обычных) языков, языки палиндромов и копий имеют корреляцию между далекими позициями, которая показана линиями на рисунках

Стохастические безконтекстные грамматики для укладки РНК

$S \rightarrow x\,S$

$S \rightarrow S\,x$

$S \rightarrow x\,S\,x'$

$S \rightarrow S\,S$

$$g\,S^{GC}\,c \rightarrow g\,a\,S^{AU}\,u\,c$$

"lexicalization" in natural language parsing

"stacking" in RNA structure - "Turner rules"

Basic CFG
"production rules"

$S \rightarrow aS$

$S \rightarrow Sa$

$S \rightarrow aSu$

$S \rightarrow SS$

a CFG "derivation"

$S \rightarrow \mathbf{a}\,\mathbf{S}$

$\rightarrow \mathbf{a}\,\mathbf{a}\,\mathbf{S}$

$\rightarrow \mathbf{a}\,\mathbf{a}\,\mathbf{S}\,\mathbf{S}$

$\rightarrow \mathbf{a}\,\mathbf{a}\,\mathbf{g}\,\mathbf{S}\,\mathbf{c}\,\mathbf{u}\,\mathbf{S}$

$\rightarrow \mathbf{a}\,\mathbf{a}\,\mathbf{g}\,\mathbf{a}\,\mathbf{S}\,\mathbf{u}\,\mathbf{c}\,\mathbf{u}\,\mathbf{g}\,\mathbf{S}\,\mathbf{c}$

$\rightarrow \mathbf{a}\,\mathbf{a}\,\mathbf{g}\,\mathbf{a}\,\mathbf{S}\,\mathbf{a}\,\mathbf{u}\,\mathbf{c}\,\mathbf{g}\,\mathbf{g}\,\mathbf{S}\,\mathbf{c}\,\mathbf{c}$

$\rightarrow \mathbf{a}\,\mathbf{a}\,\mathbf{g}\,\mathbf{a}\,\mathbf{c}\,\mathbf{S}\,\mathbf{g}\,\mathbf{a}\,\mathbf{u}\,\mathbf{c}\,\mathbf{u}\,\mathbf{g}\,\mathbf{g}\,\mathbf{c}\,\mathbf{S}\,\mathbf{c}\,\mathbf{c}$

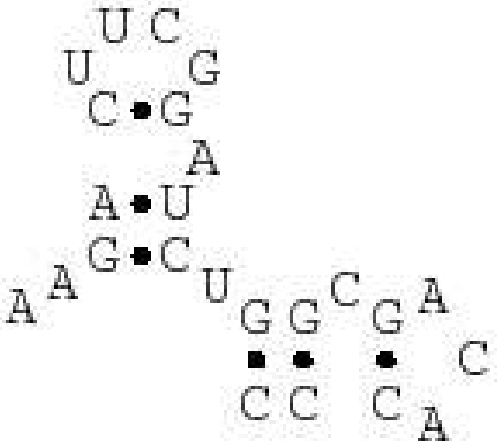
$\rightarrow \mathbf{a}\,\mathbf{a}\,\mathbf{g}\,\mathbf{a}\,\mathbf{c}\,\mathbf{u}\,\mathbf{S}\,\mathbf{g}\,\mathbf{a}\,\mathbf{u}\,\mathbf{c}\,\mathbf{u}\,\mathbf{g}\,\mathbf{g}\,\mathbf{c}\,\mathbf{g}\,\mathbf{S}\,\mathbf{c}\,\mathbf{c}\,\mathbf{c}$

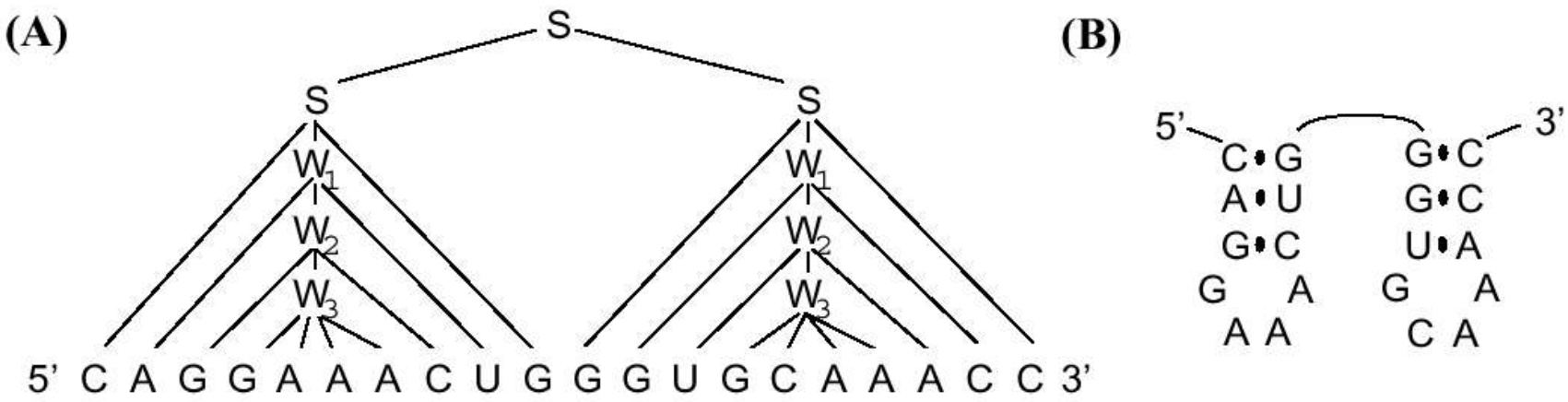
$\rightarrow \mathbf{a}\,\mathbf{a}\,\mathbf{g}\,\mathbf{a}\,\mathbf{c}\,\mathbf{u}\,\mathbf{u}\,\mathbf{S}\,\mathbf{g}\,\mathbf{a}\,\mathbf{u}\,\mathbf{c}\,\mathbf{u}\,\mathbf{g}\,\mathbf{g}\,\mathbf{c}\,\mathbf{g}\,\mathbf{a}\,\mathbf{S}\,\mathbf{c}\,\mathbf{c}\,\mathbf{c}$

$\rightarrow \mathbf{a}\,\mathbf{a}\,\mathbf{g}\,\mathbf{a}\,\mathbf{c}\,\mathbf{u}\,\mathbf{u}\,\mathbf{c}\,\mathbf{S}\,\mathbf{g}\,\mathbf{a}\,\mathbf{u}\,\mathbf{c}\,\mathbf{u}\,\mathbf{g}\,\mathbf{g}\,\mathbf{c}\,\mathbf{g}\,\mathbf{a}\,\mathbf{c}\,\mathbf{S}\,\mathbf{c}\,\mathbf{c}\,\mathbf{c}$

$\rightarrow \mathbf{a}\,\mathbf{a}\,\mathbf{g}\,\mathbf{a}\,\mathbf{c}\,\mathbf{u}\,\mathbf{u}\,\mathbf{c}\,\mathbf{g}\,\mathbf{S}\,\mathbf{g}\,\mathbf{a}\,\mathbf{u}\,\mathbf{c}\,\mathbf{u}\,\mathbf{g}\,\mathbf{g}\,\mathbf{c}\,\mathbf{g}\,\mathbf{a}\,\mathbf{c}\,\mathbf{a}\,\mathbf{S}\,\mathbf{c}\,\mathbf{c}\,\mathbf{c}$

$\rightarrow \mathbf{a}\,\mathbf{a}\,\mathbf{g}\,\mathbf{a}\,\mathbf{c}\,\mathbf{u}\,\mathbf{u}\,\mathbf{c}\,\mathbf{g}\,\mathbf{g}\,\mathbf{a}\,\mathbf{u}\,\mathbf{c}\,\mathbf{u}\,\mathbf{g}\,\mathbf{g}\,\mathbf{c}\,\mathbf{g}\,\mathbf{a}\,\mathbf{c}\,\mathbf{a}\,\mathbf{c}\,\mathbf{c}\,\mathbf{c}$





A) Представление в виде дерева грамматического разбора для «CAGGAAACUGGGUGCAAACC»

B) Вторичная структура РНК для этой же последовательности, которая соотносится с деревом грамматического разбора

Достоинства и недостатки эволюционных алгоритмов

- не требуют знания энергетических параметров и взаимодействий РНК с белками
- наиболее достоверное предсказание
- вычисления через Интернет

Недостатки

- низкая скорость
- точность зависит от объема и качества набора изофункциональных РНК, в том числе качества предварительной обработки (выравнивания)
- зависят от большого числа плохо контролируемых параметров