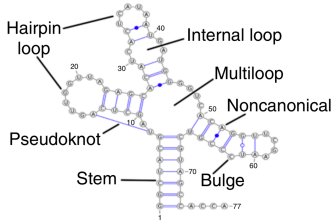
一、 标注 RNA 基本二级结构单元

如下图所示：一共有 hairpin loop, internal loop, multiloop, stem, bulge 5 种基本二级结构单元(这里没有 pseudoknot, noncanonical)。



用下列字母表示基本二级结构单元。

```
L : paired, 5' end  (
R : paired, 3' end  )
H : hairpin loop
T : internal loop
B : bulge loop
M : multiloop
E : external region(unpaired) RNA
```

parse.py 脚本中 toPairs 函数获取配对碱基对，parse 函数标注出该 RNA 序列中每个碱基所对应的基本二级结构单元，手动将 rRNA_1.txt 分为 rRNA_1_2.txt rRNA_1_1.txt 运行命令 python parse.py 得到 seq1 和 seq2 的结果。

```
>seq1
ELLLLLLLLTTTTTTTTTTTTTTTTTTTTTTTTTTLLBB
BLLLLLLLLLLLLLLLLLHHRRRRBBRRRRRRRRBRRMMM
MMMMLLLLLLLTLLLLTTTTTTTLLLLLLLLLLLLTTTTTT
TTLLTTLLLLLLLLLLHHHHRRRRRRRRRRTTTRRTTTTT
TTTTTTTTTTTRRRRRRRRRRTTTTTTTTTTTTRRRRTT
TRRLLLLLHHHHHRRRRRLLLLLHHHHRRRRRRRRRRMRR
RRRTTTTTRRRRRBRRREEEEE
>seq2
ELLLLLLLLTTTTTTTTTTTTTTTTTTTTTTTTTTTTTL
LBBBBLLLLLLLLLLLLLLLLHHHRRRBBBRRRRRRRRBR
RMMMMMMMLLLLLLLTLLLLTTTTTTTTLLLLLLLLLLLT
TTTTTTLLTTTLLLLLLLLLHHHHRRRRRRRRRTTTTTRR
TTTTTTTTTTTTTTTRRRRRRRRRRTTTTTTTTTTTTTR
RRRTTTRRLLLLLHHHHHRRRRRLLLLLHHHHRRRRRRRR
RRMRRRRRTTTTTRRRRRBRRREEEEE
```

二、 SCFG 计算 RNA 二级结构

1) 构建 CYK 模型。
非终止状态：**S, F, L**
终止状态：**A U C G (用 a 表示)**
规则：

$$L->aFa' \qquad S->L \qquad L->a \qquad S->LS \qquad F->aFa' \qquad F->LS$$

2) 初始化。根据初始化条件公式：

For $i = 1$ to n, W = S, F, L

$\gamma(i, i-1, \mathrm{W}) = -\infty$,

$\gamma(i, i, W) = \log P(W \to x_i)$ if the rule $W \to x_i$ exists,

otherwise, $\gamma(i, i-1, W) = -\infty$

得到本实验 CYK 模型初始化条件：

由于 1.非终止状态 L 能发射终止状态 a；2.非终止状态 S 能发射到 L 再到终止状态 a；因此可得：

$$\gamma(i, i, \mathrm{W}) = \begin{cases} -\infty, & W = F \\ \log P(L \to x_i) + \log P(x_i), & W = L \\ \gamma(i, i-1, \mathrm{L}) + \log P(S \to L) = \log P(L \to x_i) + \log P(x_i) + \log P(S \to L), & W = S \end{cases}$$

3) 动态规划迭代过程：

根据迭代公式：

For $i = 1$ to $n, i = i+1$ to n, W = S, F, L

$$\gamma(i, j, \mathrm{W}) = max \begin{cases} \gamma(i, j, \mathrm{Y}) + logP(W \to Y) \\ \max\limits_{i \le k \le j-1} \gamma(i, k, \mathrm{Y}) + \gamma(k+1, j, \mathrm{Z}) + \log P(W \to YZ) \\ \gamma(i+1, j, \mathrm{Y}) + logP(W \to x_i Y) \\ \gamma(i, j-1, \mathrm{Y}) + logP(W \to Y x_i) \\ \gamma(i+1, j-1, \mathrm{Y}) + logP(W \to x_i Y x_i) \end{cases}$$

根据 S，F，L 的发射规则，无$W \to Yx_i$和$W \to x_i Yx_i$整理可得：

$$\gamma(i,j,\mathrm{W}) = \max\begin{cases} \max\limits_{i\le k\le j-1} \gamma(i,k,\mathrm{Y}) + \gamma(k+1,j,\mathrm{Z}) + \log P(W \to YZ) & W = S,F \\ \gamma(i+1,j-1,\mathrm{Y}) + logP(W \to x_i Y x_i), & W = L,F \\ \gamma(i,j,\mathrm{Y}) + logP(W \to Y), & W = S \end{cases}$$

4) 终止条件：

$$\log P(x,\hat{\pi} \mid \theta) = \gamma(1,L,1)$$

5) 回溯算法：

(来源AN EFFICIENT ALGORITHM FOR ALIGNING SEQUENCES TO RNA SECONDARY STRUCTURE，KAN LIU)

TraceBack(New Linked List l , $\tau(W,i,j)=(r,k)$)

If (r is $W \to a$)

Add 'T' to the end of l

return

If (r is $W \to Y$)

TraceBack(l , $\tau(Y,i,j)$)

If (r is $W \to YZ$)

TraceBack(New Linked List l_1, $\tau(Y,i,k)$)

TraceBack(New Linked List l_2 , $\tau(Z,k+1,j)$)

L=Concatenation of l_1 and l_2

If (r is $W \to aY$)

TraceBack(l , $\tau(Y,i+1,j)$)

Add 'L' to the head of l

If (r is $W \to Ya$)

TraceBack(l , $\tau(Y,i,j-1)$)

Add 'R' to the end of l

If (r is $W \to aYb$)

TraceBack(l , $\tau(Y,i+1,j-1)$)

Add '>' to the head of l

Add '<' to the end of l

6) 运行 python main.py，得到结果：

```
>seq 1:
))))))))).)))....(((.........)))))))).)))))).....)))))...)))))).....(((((((((((
).............)))))))))....))))........((((...  .........((((((((((.........((((
((((((((((((((.........)))))).........(((((..........(((((((((.(((((((((
>seq 2:
```