reference 参考序列　　举例：GOO GOL

short-read / query 短序列 / 查询序列　　举例：GOL

定义：Σ为一个字母表集合 / 符号集合

符号 $ 不出现于 Σ 中，并且 $ 的字典序小于 Σ 中所有符号

定义：字符串 $X = a_0 a_1 a_2 a_3 a_4 \cdots\cdots a_{n-2} a_{n-1}$，并且 $a_{n-1}$ 始终为 $

$|X|$ 表示 X 的长度 $|X| = n$

$X[i]$ 表示 X 中的符号 $X[i] = a_i$，$i = 0, 1, 2, \cdots\cdots, n-1$

$X[i,j]$ 表示 X 的子串 $X[i,j] = a_i a_{i+1} \cdots a_{j-1} a_j$，$i \le j$

$X_i$ 表示 X 的后缀 $X_i = X[i, n-1] = a_i a_{i+1} \cdots a_{n-2} a_{n-1}$

对于 GOOGOL 来说，X = GOO GOL $　$|X| = 7$

经过 BWT 变换后 X 被为 B

BWT 变换的过程

① 循环移位　② 字典序排序　③ 排序后序列的最后一位

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 0 | googol$ | 0 | 6 | $googol | l | | |
| 1 | oogol$g | 1 | 3 | gol$goo | o | X = googol$ | |
| 2 | ogol$go | 2 | 0 | googol$ | $ | ⇓ BWT | |
| 3 | gol$goo | ⇒ 3 | 5 | l$googo | ⇒ o | ⇒ | B = lo$oogg |
| 4 | ol$goog | 4 | 2 | ogol$go | o | | |
| 5 | l$googo | 5 | 4 | ol$goog | g | | |
| 6 | $googol | 6 | 1 | oogol$g | g | | |

# 关于 BWT 变换的一些结论

取循环移位后的序列最后一位得到 B，也称为 L

取循环移位后的序列第一位得到 F

B / L = l o $ o o g g

F = $ g g l o o o

① L/B 序列的第一个元素是原文本的最后一个元素

② F 序列可由 L/B 序列按照字典序排列而来，即 B/L 可得到 F

③ 可以由 F，L/B 序列还原原文本

3.1 确定某个元素，与 L/B 列对应的 F 列元素为原串的后一个元素

3.2 确定某个元素，与 F 列对应的 L/B 列元素为原串的前一个元素

大致可认为有如下的对应关系

| 前 | 后 |
| --- | --- |
| L/B | F |

如果出现重复的字符

只需要知道该字符在它之前 L / F 列出现过几次，对应排名

F / L 列同来字符就是要找列的位置

⑩ ⑪ ⑫ ⑬ ⑭ ⑮

B / L = l o $ o o g g

F = $ g g l o o o

① ⑭ ⑫ ⑮ ⑬ ⑯

以 $ 为开始，按 3.1 向后恢复原文本

$ → g　g → o　o → o　o → g　g → o　o → l

得到原文本 $ googol ⇒ googol

以$为开始，按3,2向前恢复原文本

l ← $　　o ← l　　g ← o　　o ← g　　o ← o　　g ← o

得到原文本　googol$ ⇒ googol