序列两两比较之序列比对法：双序列全局比对

经典的全局比对算法是 Needleman-Wunsch 算法。1970 年，Needleman 和 Wunsch 首先将动态规划法应用于两条序列的全局比对，后来这个算法就称为 Needleman-Wunsch 算法。今天，所有比对软件使用的算法都是从这个经典算法衍生出来的。

我们用 Needleman-Wunsch 算法为序列 p 和序列 q 创建全局比对。输入值除了两条序列之外，还要有替换积分矩阵以确定不同字母间的相似度得分，以及空位罚分（图 1）。空位罚分就是当字母对空位的时候应该得几分。我们还是希望一致或相似的字母尽可能的对在一起，字母对空位的情况和不相似的字母对在一起的情况一样，都不是我们希望的，还是少出现为好，所以通常字母对空位会得到一个负分，这个负分就叫做空位罚分。这里我们让空位罚分，也就是 gap 分值为-5 分。在比对中没有空位对空位的情况。输入值就是这些。

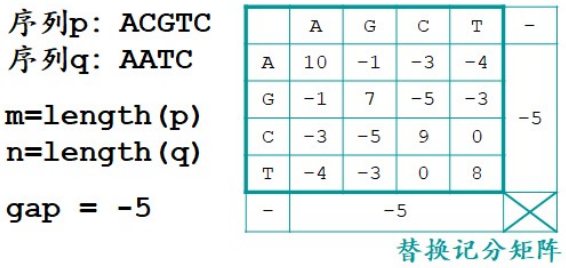

图 1. 全局比对输入值：序列 p 和序列 q，替换记分矩阵，空位罚分

接下来我们要创建一个得分矩阵（图 2-1），并根据公式（图 2-2）把得分矩阵填满。填满后全局比对就会跃然于纸上。得分矩阵的第一行是序列 p，第一列是序列 q，这一步和打点法很像。不过要注意，p 和 q 的前面各留一个空列和一个空行，也就是第 0 列和第 0 行。

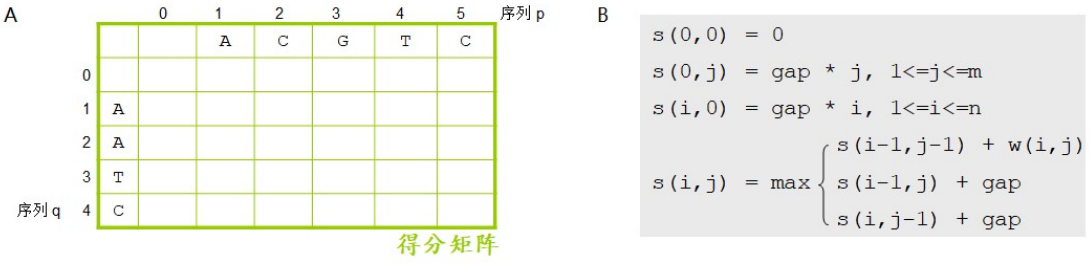

图 2. 全局比对计算公式及得分矩阵

现在开始给得分矩阵赋值（图 3）。根据公式：

s(0,0)是初始值 0。

第 0 行：s(0,j) = gap * j

j 从 1 到 m，m 是序列 p 的长度。也就是 s(0,1)=gap*1=-5，s(0,2)=gap*2=-10，依次类推。第 0 行实际是一种极端情况的假设。也就是当序列 p 全部对空位时的得分。A 对空位是-5 分，AC 都对空位就累计到了-10 分，ACG 都对空位就累积到了-15 分，如果序列 p 全部对空位，最终的累积得分就是-25 分。

第 0 列：s(i,0) = gap * i

第 0 列和第 0 行一样，也是反映了序列 q 如果全部对空位的累计得分。对一个空位累积 gap*1=-5 分，对两个空位累积 gap*2=-10 分，对三个空位累积 gap*3=-15 分，对四个空位累积 gap*4=-20 分。

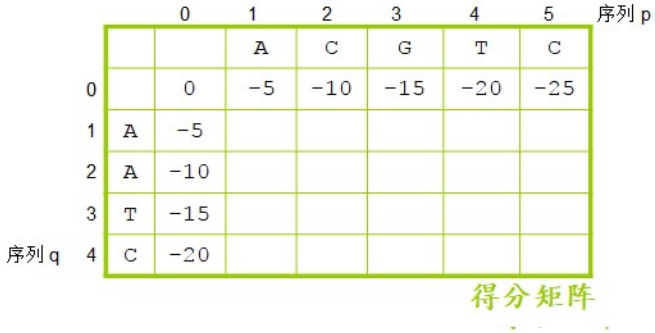

图 3. 得分矩阵中的第 0 行和第 0 列

第 0 行和第 0 列相对简单，其他的格就稍微复杂一点儿了。接下来填 s(1,1)（图 4）。这个格里的值来源于三个值中的最大值。哪那三个值呢，一个是上面格 s(0,1) 里的值加 gap，一个是左面格 s(1,0) 里的值加 gap，还有一个是斜上格 s(0,0) 里的值加当前这个位置字母对字母在替换记分矩阵里的分值 w(i,j)。什么意思呢？就是累积到这个位置时，是字母对字母得分高，还是序列 p 的字母对空位得分高，还是序列 q 的字母对空位得分高？有且只有这三种情况，我们要的是得分最高的那种情况。逐个看一下，上面格 s(0,1)+gap=-5+-5=-10。左面格 s(1,0) +gap=-5+-5=-10。斜上格 s(0,0)+w(1,1)=0+10=10。max(-10,-10,10)=10。所以当前这个格 s(1,1) 的分值就是 10。此外，我们还需要用箭头记录一下这个 10 是从哪里来的。它是从斜上这个格来的，所以我们画一个指向斜上的箭头。

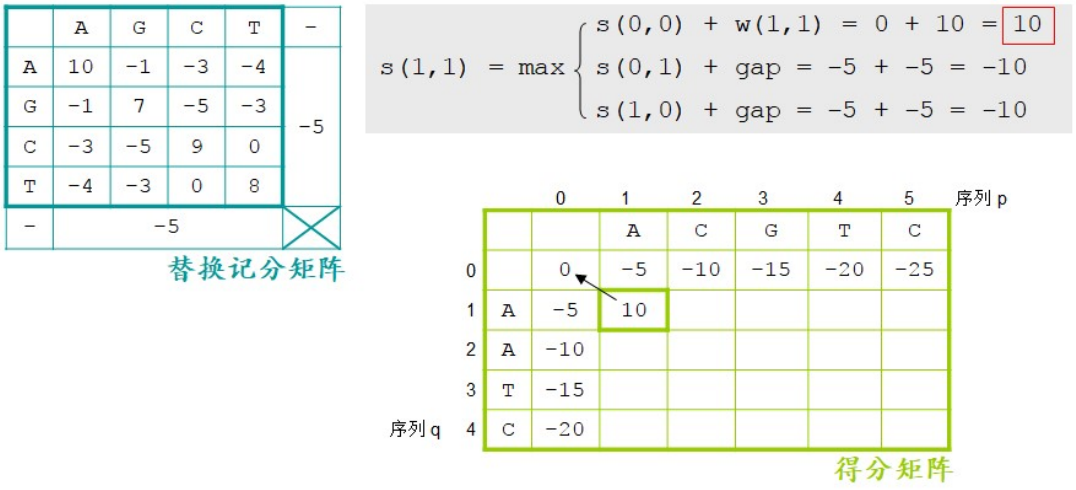

图 4. 得分矩阵中的 s(1,1)

接下这个格 s(1,2)值的计算（图 5），仍然是找三个值中的最大值。上面格 s(0,2)+gap=-10+-5=-15。左面格 s(1,1)+gap=10+-5=5。斜上格 s(0,1)+w(1,2)=-5+-3=-8。max(-15,5,-8)=5。大值是 5，来源于左面格 s(1,1)，画上向左的箭头。

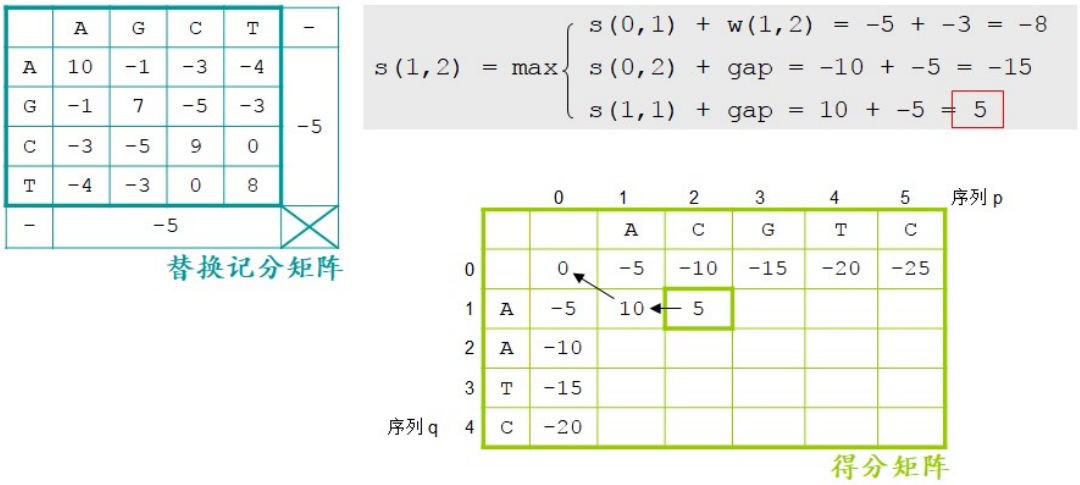

图 5. 得分矩阵中的 s(1,2)

按照上面的公式，将整个得分矩阵填满。这时，我们再回过头来看一下第一行和第一列（图 6）。其实，第一行的每一个值都是从左边的格加 gap 来的。所以我们给它们补上向左的箭头。第一列的每一个值都是从上边的格加 gap 来的。所以我们给它们补上向上的箭头。至此，所有的箭头和数值就都填好了。填满之后，右下角的分数就是整个全局比对最终的得分。然后从这个位置开始追溯箭头一直到左上角的零，并且把这些箭头标记出来。

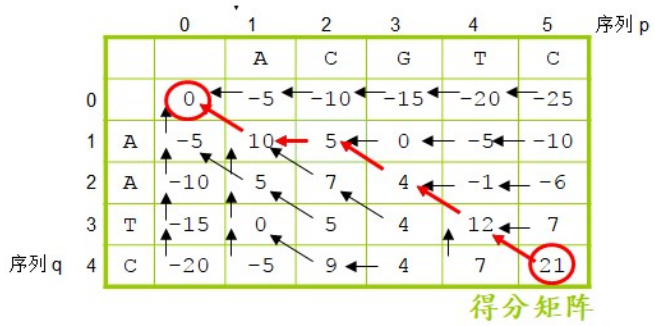

图 6 填满分值和箭头的得分矩阵

图 6 中标出的红色箭头是写出全局比对的唯一依据。追溯箭头是从右下角到左上角，但是写全局比对是从左上角开始，如果是斜箭头则是字符对字符，如果是水平箭头或垂直箭头则是字符对空位，箭头指着的序列为空位。我们看第一个是斜箭头，字母对字母，就是 A 对 A，第二个是水平箭头，字母对空位，箭头指着的序列是空位，也就是 C 对空位。然后斜箭头 G 对 A，斜箭头 T 对 T，斜箭头 C 对 C，一直写到右下角，全局比对就出现了（图 7）。唯一的一个空位插在序列 q 的 A 与 A 之间，这样最终的比对得分最高。不信的话可以试试，其他任何一种插入空位的比对结果，得分都不会超过 21 分。因为我们在得分矩阵的创建过程中，每一步都是在上一步最优的情况下得出的当前最优结果。

```
序列p:  A C G T C
序列q:  A - A T C
```

图 7. 序列 p 和序列 q 的全局比对