Количество выравниваний

Никита Вяткин

4 марта 2024 г.

Будем считать, что перестановка двух соседних gap-ов (расположенных в разных строках) в выравнивании не меняет выравнивание:

$$\begin{matrix} \cdots & A & - & \cdots \\ \cdots & - & C & \cdots \end{matrix} \longrightarrow \begin{matrix} \cdots & - & A & \cdots \\ \cdots & C & - & \cdots \end{matrix}$$

(здесь и везде далее A и C – произвольные символы выровненных строк)

С помощью таких перестановок мы можем добиться *канонического вида* для выравнивания, т.е. в нем не будет встречаться ситуация:

$$\begin{matrix} \cdots & A & - & \cdots \\ \cdots & - & C & \cdots \end{matrix}$$

Посчитаем, сколько канонических выравниваний.
Пусть $g(m, n)$ — количество канонических выравниваний двух строк длины m и n, у которых на последней позиции во второй строке стоит **gap** (а в первой строке, следовательно, ее последний символ, т.к. сопоставление gap-gap недопустимо).
Пусть $f(m, n)$ — количество канонических выравниваний двух строк длины m и n, у которых на последней позиции во второй строке стоит **последний символ этой строки** (не gap !).
Ясно, что $f(m, n) + g(m, n)$ суть общее число всех канонических выравниваний двух строк длины m и n. Напишем рекуррентные формулы для введенных функций:

$$g(m, n) = f(m - 1, n) + g(m - 1, n)$$

$$f(m, n) = f(m - 1, n - 1) + g(m - 1, n - 1) + f(m, n - 1)$$

В первом случае мы просто убираем последнюю позицию из выравнивания и получаем выравнивание строк длины $m - 1$ и n. Во втором случае действуем также, но на последней позиции могут быть выравнены символы строк (что дает вклад $f(m - 1, n - 1) + g(m - 1, n - 1)$) или gap с символом второй строки. В последнем случае мы получаем именно $f(m, n-1)$, поскольку ситуация $\begin{smallmatrix} \dots & A & - \\ \dots & - & C \end{smallmatrix}$ запрещена.

Теперь докажем индукцией по m и n, что $f(m, n) = \binom{m+n-1}{m}$, $g(m, n) = \binom{m+n-1}{m-1}$ – биномиальные коэффициенты. База: $m = 1$ или $n = 1$ — тривиальна. Порядок на остальных парах (m, n) можно устроить, например, следующим образом:

$$(2,2) \to (2,3) \to (2,4) \to (2,5) \to \dots$$
$$(3,2) \to (3,3) \to (3,4) \to (3,5) \to \dots$$
$$\dots$$

После улаженных формальностей, сделаем переход индукции:

$$\begin{aligned}
f(m,n) &= f(m-1,n-1) + g(m-1,n-1) + f(m,n-1) \\
&= \binom{m+n-3}{m-1} + \binom{m+n-3}{m-2} + \binom{m+n-2}{m} \\
&= \binom{m+n-1}{m} \\
g(m,n) &= f(m-1,n) + g(m-1,n) \\
&= \binom{m+n-2}{m-1} + \binom{m+n-2}{m-2} \\
&= \binom{m+n-1}{m-1}
\end{aligned}$$

по свойствам биномиальных коэффициентов.

Итак, количество всех каноничных выравниваний двух строк длины m и n равно $f(m, n) + g(m, n) = \binom{m+n-1}{m} + \binom{m+n-1}{m-1} = \binom{m+n}{m}$ – числу сочетаний из $m + n$ по m.

При $m, n \to \infty$ по формуле Стирлинга:

$$\binom{m+n}{m} = \frac{(m+n)!}{m!n!} \approx \frac{1}{\sqrt{2\pi}} \frac{\sqrt{m+n}(m+n)^{m+n}}{\sqrt{m}m^m\sqrt{n}n^n}$$

Если дополнительно предположить, что $m \approx n$, то

$$\binom{m+n}{m} \approx \frac{1}{\sqrt{2\pi}} \frac{\sqrt{2n}(2n)^{2n}}{n^{2n+1}} = \frac{2^{2n}}{\sqrt{\pi n}}$$