Курс: биоинформатика
Теоретическая домашняя работа

Задача 1. Выведите рекуррентную формулу количества всех возможных выравниваний последовательностей длины n и m пользуясь разбиением всех выравниваний на непересекающиеся блоки. (1.5 балл)

Решение:

Рассмотрим выравнивание двух последовательностей A и B, длины n и m. Под выравниваем подразумеваются операции вставка, замена и удаление. Замена и удаление подразумевают вставку гэпа в последовательность.

Базовые случаи:

$$W(0, m) = 1, \qquad W(n, 0) = 1$$

Общий случай: При $n > 0$ и $m > 0$ возможны три варианта:
1. **Гэп в A (вставка символа в B):** $\quad W(n, m-1)$
2. **Гэп в B (вставка символа в A):** $\quad W(n-1, m)$
3. **Заменаопоставление символов:** $\quad W(n-1, m-1)$

Тогда общее количество возможных выравниваний задаётся рекуррентным соотношением:

$$W(n, m) = W(n-1, m) + W(n, m-1) + W(n-1, m-1)$$

Задача 2. Получите точную формулу, основываясь на начальные условия и рекуррентную формулу. (1.5 балла)

Решение:

Рекуррентное соотношение:

$$W(n, m) = W(n-1, m) + W(n, m-1) + W(n-1, m-1)$$

с начальными условиями:

$$W(0, m) = 1, \quad W(n, 0) = 1.$$

Рассмотрим теперь точное выражение для $W(n, m)$. Пусть k — количество позиций соответсвующие заменеопоставлению, не учитыаающие гэп. Тогда:
- из n символов строки A выбрано k позиций для выравнивания (остальные $n - k$ будут выровнены с гэпами),
- из m символов строки B выбрано k позиций для выравнивания (остальные $m - k$ будут выровнены с гэпами),
- на каждой из k выровненных позиций возможны 2 варианта: либо символы равны, либо происходит замена.

Таким образом, общее количество выравниваний при фиксированном k равно:

$$\binom{n}{k} \cdot \binom{m}{k} \cdot 2^k$$

Просуммировав по всем возможным k от 0 до $\min(n, m)$ (min, так как мы не можем заменить и сопоставить больше символов, чем их есть в самой короткой строке.), получим ответ:

$$W(n, m) = \sum_{k=0}^{\min(n,m)} \binom{n}{k} \binom{m}{k} 2^k$$

Задача 3. Воспользуйтесь приближением Стирлинга чтобы получить приближенную формулу количества выравниваний. (1)

Подставим

$$n! \approx \sqrt{2\pi n} \left(\frac{n}{e}\right)^n$$

в выражение

$$W(n, m) = \sum_{k=0}^{\min(n,m)} \binom{n}{k} \binom{m}{k} 2^k.$$

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}, \quad \binom{m}{k} = \frac{m!}{k!(m-k)!}$$

Тогда слагаемое суммы выглядит так:

$$S(k) = \binom{n}{k} \binom{m}{k} 2^k = \frac{n! \cdot m! \cdot 2^k}{(k!)^2 (n-k)!(m-k)!}$$

Используем приближение Стирлинга

$$S(k) \approx \frac{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n \cdot \sqrt{2\pi m}\left(\frac{m}{e}\right)^m \cdot 2^k}{(2\pi k)\left(\frac{k}{e}\right)^{2k} \cdot \sqrt{2\pi(n-k)}\left(\frac{n-k}{e}\right)^{n-k} \cdot \sqrt{2\pi(m-k)}\left(\frac{m-k}{e}\right)^{m-k}}$$

Приближённая формула для слагаемого суммы:

$$S(k) \approx \frac{\sqrt{nm}}{k\sqrt{(n-k)(m-k)}} \cdot \frac{n^n \cdot m^m \cdot 2^k}{k^{2k}(n-k)^{n-k}(m-k)^{m-k}}$$

$$W(n, m) = \sum_{k=0}^{\min(n,m)} S(k)$$