1. Выравнивание последовательностей.

Даже полученных за данный курс базовых знаний достаточно, чтобы понимать некоторые практически важные приложения комбинаторики.

Например, существует такая наука — биоинформатика, которая занимается расшифровкой ДНК. Последовательность ДНК можно рассматривать как последовательность символов, то есть совершенно комбинаторный объект. В задачи биоинформатики входит задача определения того, насколько два существа родственные. В частности, понять, родственные ли два существа, можно сравнивая их ДНК. При этом возникает задача сравнения двух последовательностей символов.

Математическая формулировка задачи следующая.

Пусть даны две последовательности $a_1, \dots, a_n$ и $b_1, \dots, b_m$. Длины двух последовательностей n и m, вообще говоря, не совпадают. Необходимо определить, что значит сравнить две последовательности.

Пример Пусть даны последовательности К,Р,О,Л,И,К и И,К,Р,А. У данных последовательностей есть два похожих участка: ИК и КР. Необходимо предоставить алгоритм поиска таких участков. Сначала эти две последовательности выравниваются, с помощью добавления символов $\emptyset$ (англ. gap) к последовательностям так, чтобы их длинны сравнялись.

Пример возможного (естественного) выравнивания.

$$\begin{array}{cccccc} К & Р & О & Л & И & К \\ И & К & Р & А & \emptyset & \emptyset \end{array}$$

Примеры других выравниваний: в которых уже можно заметить совпадения фрагментов последовательностей

$$\begin{array}{ccccccc} \emptyset & К & Р & О & Л & И & К \\ И & К & Р & А & \emptyset & \emptyset & \emptyset \end{array} \qquad \begin{array}{cccccccc} К & Р & О & Л & И & К & \emptyset & \emptyset \\ \emptyset & \emptyset & \emptyset & \emptyset & И & К & Р & А \end{array}$$

В связи с этим появляется задача нахождения числа всех возможных выравниваний для двух данных последовательностей: как содержательных, так и не очень.

Допустим, что ни в каком выравнивании не возникает ситуация, когда gap стоит над gap'ом. Это предположение естественно, поскольку не имеет смысла сравнивать два специально добавленных символа.

Также предположим, что если есть два выравнивания следующего вида

$$\begin{pmatrix} \dots & a & \emptyset & \dots \\ \dots & \emptyset & b & \dots \end{pmatrix}, \qquad \begin{pmatrix} \dots & \emptyset & a & \dots \\ \dots & b & \emptyset & \dots \end{pmatrix},$$

то такие два выравнивания отождествляются.

Пример Отождествляются следующие выравнивания:

$$\begin{pmatrix} a & \emptyset & \emptyset & a \\ \emptyset & b & b & \emptyset \end{pmatrix} = \begin{pmatrix} \emptyset & a & \emptyset & a \\ b & \emptyset & b & \emptyset \end{pmatrix} = \begin{pmatrix} \emptyset & \emptyset & a & a \\ b & b & \emptyset & \emptyset \end{pmatrix}.$$

Пусть количество всех возможных выравниваний последовательностей $a_1, \dots, a_n$ и $b_1, \dots, b_m$, которое удовлетворяют указанным выше условиям, обозначается как $g(n, m)$.

Теорема *Для величины g верна рекуррентная формула $g(n, m) = g(n-1, m) + g(n, m-1)$.*

Замечание В качестве начальных условий можно будет использовать следующие: $g(1, 1) = 2$, $g(0, 1) = g(1, 0) = 1$. После этого с помощью теоремы можно будет получить любое число $g(n, m)$ с помощью рекуррентного соотношения, например: $g(2, 2) = g(1, 2) + g(2, 1) = g(0, 2) + g(1, 1) + +g(2, 1) = \dots$

Доказательство Пусть V — множество всех интересующих выравниваний. $|V| = g(n, m)$. Произвольное выравнивание относится к одному из трех типов, в зависимости от того, как выглядит последний столбец выравнивания:

$$\begin{pmatrix} \dots & a_n \\ \dots & b_n \end{pmatrix} \quad \text{или} \quad \begin{pmatrix} \dots & \emptyset \\ \dots & b_n \end{pmatrix} \quad \text{или} \quad \begin{pmatrix} \dots & a_n \\ \dots & \emptyset \end{pmatrix}.$$

Множества выравниваний, относящихся к данным трем типам, обозначаются V_1, V_2 и V_3 соответственно. Множества V_2 и V_3 имеют непустое пересечение. Множество выравниваний следующего вида

$$\begin{pmatrix} \dots & a_n & \emptyset \\ \dots & \emptyset & b_n \end{pmatrix}$$

на самом деле есть множество равное пересечению множеств V_2 и V_3. Тогда

$$|V| = |V_1| + |V_2| + |V_3| - |V_2 \cap V_3| = |V_1| + |V_2| + |V_3| - |V_4|.$$

Поскольку $|V| = g(n,m)$, $|V_1| = g(n-1,m-1)$, $|V_2| = g(n,m-1)$, $|V_3| = g(n-1,m)$, $|V_4| = g(n-1,m-1)$:

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

Следствие $g(n,m) = C^m_{n+m} = C^n_{n+m}$

Доказательство Доказательство можно провести по индукции. Во-первых:

$$g(1,0) = C^0_{1+0} = 1 \qquad g(0,1) = C^1_{1+0} = 1,$$

то есть начальные условия согласуются. Во-вторых, если предположить, что соотношение выполняется для $g(n-1,m)$ и $g(n,m-1)$, тогда:

$$g(n,m) = g(n-1,m) + g(n,m-1) = C^{n-1}_{n-1+m} + C^n_{n+m-1} = C^n_{n+m}.$$

Пример В случае $n = m = 1000$ количество всевозможных выравниваний

$$g(1000,1000) = C^{1000}_{2000}.$$

Можно воспользоваться следующим соотношением, чтобы оценить полученное выражение.

$$C^0_{2000} + C^1_{2000} + ... + C^{1000}_{2000} + ... + C^{2000}_{2000} = 2^{2000} \quad \Longrightarrow \quad C^{1000}_{2000} < 2^{2000}$$

Чтобы оценить снизу C^{1000}_{2000} можно показать, что среди всех слагаемых C^k_{2000} слагаемое C^{1000}_{2000} самое большое. Это следует из того, что для $k \leq 999$:

$$\frac{C^k_{2000}}{C^{k+1}_{2000}} = \frac{2000!}{k!\,(2000-k)!} \cdot \frac{(k+1)(2000-k-1)!}{2000!} = \frac{k+1}{2000-k} \leq \frac{1000}{2000-k} \leq \frac{1000}{1001} < 1.$$

Тогда верна следующая оценка снизу:

$$C^{1000}_{2000} \geq \frac{2^{2000}}{2001} > \frac{2^{2000}}{2^{11}} = 2^{1989} = 10^{1989 \cdot \lg 2} < 10^{500}.$$

Такое количество выравнивании даже близко невозможно перебрать никакой вычислительной машиной.