Tropin Nikolay

tropinnikolay@gmail.com

24 January 2021

1 Матрица замен

Найдем стоимости замен, пользуясь матрицей замен BLOSUM:
$D \to E = 2,\ I \to V = 3,\ D \to V = -3$. Видим, что замены внутри класса более вероятны, так например D (аспарагиновая кислота) и E (глутаминовая кислота) относятся к заряженным аминокислотам, а I (изолейцин) и V (валин) являются гидрофобными аминокислотами.
Рассчитаем вероятность: как и в матрицах PAM в матрицах BLOSUM результат представлен в виде логрифмов частот замен. Во избежание дробных чисел все значения в матрицах умножены на 10 и округлены. Поэтому искомые вероятности: $D \to E : 10^{0.2} = 1.6, I \to V : 10^{0.3} = 2, D \to V : 10^{-0.3} = 0.5$ - данные значения показывают во сколько раз конкретная замена происходит чаще, чем при случайной мутации.

2 Количество выравниваний

Мы изначально предполагали, что нет ситуации, когда в выравнивании стоит gap над gap'ом. Также предположим, что если есть два выравнивания следующего вида:

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

то такие два выравнивая отождествляются. Пусть количество всех возможных выравниваний последовательностей $a_1, \dots, a_n$ и $b_1, \dots, b_m$, которое удовлетворяет указанным выше условиям, обозначается как $g(n,m)$. Тогда для величины g верна рекуррентная формула $g(n,m) = g(n-1,m) + g(n,m-1)$. Начальные условия: $g(1,1) = 2,\ g(0,1) = g(1,0) = 1$.
Для доказательства формулы рассмотрим множество всех выравниваний V, тогда $|V| = g(n,m)$. Произвольные выравнивания относятся к одному из трёх типов, в зависимости от того, как выглядит последний столбец выравнивания:

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

Множества выравниваний, относящихся к данным трём типам, обозначаются V_1, V_2, 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 $|V| = g(n,m),\ |V_1| = g(n-1,m-1),\ |V_2| = g(n,m-1)$,
$|V_3| = g(n-1,m),\ |V_4| = g(n-1,m-1)$, то:
$g(n,m) = g(n-1,m-1) + g(n,m-1) + g(n-1,m) - g(n-1,m-1) = g(n,m-1) + g(n-1,m)$, ч.т.д

Из данной формулы следует, что

$$g(n,m) = \binom{n+m}{n} = \binom{n+m}{m}$$

Докажем по индукции:
$g(1,0) = \binom{1+0}{0} = 1$ и $g(0,1) = \binom{1+0}{1} = 1$ - начальные условия выполнены. Пусть соотношение выполнено для $g(n-1,m) = g(n,m-1)$, тогда:

$$g(n,m) = g(n-1,m) + g(n,m-1) = \binom{n-1+m}{n-1} + \binom{n+m-1}{n} = \binom{n+m}{n}.$$

Теперь воспользуемся формулой Стирлинга $n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$, предположив также, что $n \approx m$, тогда:

$$\binom{n+m}{m} = \frac{(n+m)!}{m!\,n!} \approx \frac{(2m)!}{(m!)^2} \approx \frac{\sqrt{2\cdot\pi\cdot 2m}\cdot\left(\frac{2m}{e}\right)^{2m}}{\left(\sqrt{2\cdot\pi\cdot m}\cdot\left(\frac{m}{e}\right)^m\right)^2} = \frac{2\cdot\sqrt{\pi m}\left(\frac{2m}{e}\right)^{2m}}{2\cdot\pi m\cdot\left(\frac{m}{e}\right)^{2m}} = \frac{2^{2m}}{\sqrt{\pi m}}$$