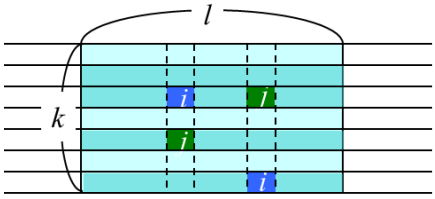
W3.3: BLOSUM の求め方

同一ファミリータンパク質のギャップなしでアラインされた領域（ブロック）に対し、アミノ酸の置換の頻度を調べて作成する。

```
-------NVYHDGACPEVKPVDNFDWSNYHGKWWEVAKYPNSVEKYGKCGWAEYTPEG---
------GDIFYPGYCPDVKPVNDFDLSAFAGAWHEIAKLPLENENQGKCTIAEYKYDG---
---------ERDCRVSSFRVKENFDKARFAGTWYAMAKKDPEGLFLQDNIVAEFSVDENGH
CLLLALALTCGAQALIVTQTMKGLDIQKVAGTWYSLAMAASDIS-LLDAQSAPLRVYVEEL
LLLCLGLTLVCVHAEEASSTGRNFNVEKINGEWHTIILASDKREKIEDNGNFRLFLEQ---
```

ブロック



ブロックの 1 つの図。実際には全ブロックをまとめて計算。

全ブロックの内容から、アミノ酸 i の出現確率 q_i を求める。

各ブロックから x %以上一致している配列群をクラスタリングしてまとめる。この際、クラスタリングされた配列を 1 本の配列として扱う。

その中で、アミノ酸 i,j が同じ位置（カラム）に現れる確率を求める。

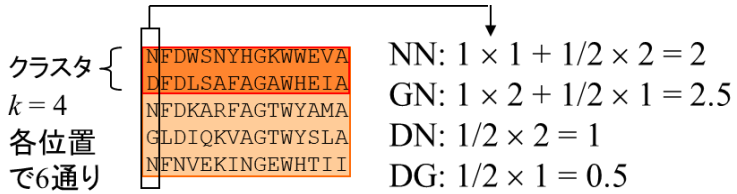
そのため、可能性のあるアミノ酸ペアの数 N をカウント。

長さ l のブロックで、 k 本の配列が取り出されたとき、アミノ酸のペアの数は、$l \times {}_k\mathrm{C}_2$

すべての b 個のブロックに対して $N = \sum_b l_b \times {}_{k_b}\mathrm{C}_2$

その中で、アミノ酸 i,j が同じ位置に現れる個数 n_{ij} を求める $\rightarrow p_{ij} = n_{ij} / N$

クラスタリングされた m 本の配列では、各アミノ酸は $1/m$ の重みをもつものとして計算。



BLOSUMx マトリックスのスコアは以下の式で与えられる。

$$s(i,j) = \log\left(\frac{p_{ij}}{q_i q_j}\right)$$

ここで、$p_{ij} = n_{ij} / N$