

A Brief Note on Computing a BLOSUM Matrix

ChengXiang Zhai
Department of Computer Science
University of Illinois at Urbana-Champaign

October 9, 2005

BLOSUM (for BLOcks SUBstitution Matrix) is a commonly used scoring matrix for sequence alignment. It gives a score for each pair of amino acids based on how likely we will observe such a pair in alignments of truly conserved blocks of amino acids. A higher score indicates that such a pair of amino acids are often seen to be aligned to each other when we align functionally similar proteins with each other.

There are three steps in computing a BLOSUM matrix, which we explain below.

1 Collecting sample blocks

We first collect a sample of blocks of amino acids that represent conserved regions, which can be obtained from proteins that are known to be in the same functional family. A block of amino acids is a set of equal-length strings of amino acids, such as the following:

```
A B C
B B C
A B C
A C B
```

The sample is typically weighted according to an identity threshold to model either long-time evolution or recent evolution. When modeling long-time evolution, we down-weight those amino acid sequences that are very similar to each other in a block so that distinct amino acids would have a higher chance of getting high scores, which is reasonable since after long-time evolution, an amino acid is more likely to be changed to a different one. The “identity threshold” (e.g., 62%) for defining which sequences to be taken as “very similar” is often used to label the variants of a BLOSUM matrix. For example, BLOSUM62 is computed using the threshold of 62%. The higher the threshold is, the more we would tolerate alignment of two distinct amino acids.

2 Computing probabilities

The basic idea of BLOSUM is, for each pair of amino acids, to compare the *actual* observed frequency of them being aligned together in our block sample with their *expected* frequency of being aligned together if they occur *independently*. Thus we will be interested in (1) the probability of observing each amino acid in the sample; and (2) the probability of observing a pair of amino acid aligned to each other in our sample.

To simplify the explanation of the essential ideas, we assume that each sequence has an equal weight; the methods can be easily generalized to deal with weighted sequences. In the homework, you do not need

to worry about sequence weighting.

Given a set of sequences $S = \{S_1, \dots, S_k\}$, each with n amino acids, i.e., $S_i = s_{i1} \dots s_{in}$, where s_{ij} is the j -th amino acid in sequence S_i . The probability of observing each amino acid X can be estimated as the relative frequency count of X in all the observed sequences, i.e.,

$$p(X) = \frac{\sum_{i=1}^k c(X, S_i)}{\sum_{X' \in \mathcal{A}} \sum_{i=1}^k c(X', S_i)}$$

where $c(X, S_i)$ is the count of amino acid X in sequence S_i , \mathcal{A} is the set of all the 20 amino acids.

Suppose we randomly sample a pair of amino acids according to $p(X)$ to form an alignment. Let $\{X, Y\}$ denote an alignment of amino acids X and Y . Note that we do not distinguish the order, so $\{X, Y\}$ and $\{Y, X\}$ would denote the same alignment. The chance of having an alignment $\{X, Y\}$, where $X \neq Y$, would be $2p(X)p(Y)$ because we can either first generate X then Y or first generate Y then X , while the chance of having an alignment $\{X, X\}$ would be $p(X)^2$. That is,

$$p(X, Y | \text{random}) = \begin{cases} p(X)^2 & \text{if } X = Y \\ 2p(X)p(Y) & \text{otherwise} \end{cases}$$

Now consider all the possible alignments we can obtain by doing pairwise alignment of S_1, \dots, S_k . We will have $\frac{k(k-1)}{2}$ pairwise sequence alignments. Since each sequence has n amino acid, we have a total of $m = \frac{nk(k-1)}{2}$ pairwise amino acid alignments, which we denote as $M = \{\{X_j, Y_j\}\}$, where $j = 1, \dots, m$. Using M as our sample, we can now count how many times we see a particular pair of amino acids, X and Y , which we denote by $c(\{X, Y\}, M)$. The probability of observing a pair of alignment $\{X, Y\}$ in our sample can thus be estimated as

$$p(X, Y | \text{sample}) = \frac{c(\{X, Y\}, M)}{m}$$

One potential problem is when $c(\{X, Y\}, M) = 0$, i.e., when we do not see $\{X, Y\}$ in our sample alignments. To avoid assigning zero probability to any pair, we can smooth the probability estimate by giving each pair an extra ‘‘pseudo count’’. As a result, we will have

$$p(X, Y | \text{sample}) = \frac{c(\{X, Y\}, M) + 1}{m + \mu}$$

where $\mu = (|\mathcal{A}|(|\mathcal{A}| - 1)/2) + |\mathcal{A}| = |\mathcal{A}|(|\mathcal{A}| + 1)/2$ is the total number of pseudo counts we added and is equal to the total number of pairs of amino acids. In the homework, since $\mathcal{A} = \{A, B, C\}$, we have $\mu = 6$. The six pairs are

$$\{\{A, A\}, \{B, B\}, \{C, C\}, \{A, B\}, \{A, C\}, \{B, C\}\}$$

3 Computing the BLOSUM matrix

Once we have these probabilities, the score of an amino acid pair $\{X, Y\}$ is obtained as

$$\text{score}(X, Y) = \text{score}(Y, X) = 2 \log_2 \frac{p(\{X, Y\} | \text{sample})}{p(\{X, Y\} | \text{random})}$$

In the homework, I defined the score similarly but without the constant 2.