

Normalization of Sequence Alignment Scores

Hilary S Booth¹²³, John H Maindonald¹³, Ole M Nielsen³⁴, Susan R Wilson¹³

Keywords: sequence analysis, composition bias

Abstract

We describe a normalization procedure for the score of a pairwise alignment of two biological sequences and an algorithm (POZITIV) that finds the normalized score efficiently. Designed to overcome the bias due to the composition of the alignment, the normalized score (which we call the POZ score) measures the distance (in standard deviations) between Smith-Waterman (S-W) score of the aligned letters and the mean value of all other S-W scores obtained by a permutation of either sequence. Unlike the usual Z-score calculation used in sequence alignment, the mean and standard deviation are taken without realigning the permuted sequences. We demonstrate that calculating the mean and standard deviation in this way enables us to calculate exactly and in two steps, the first being $O(N)$ time, where N is the length of the sequence, the second in a fixed number of calculations i.e. in $O(1)$ time.

1 Background on Z-scores

Z-scores can be used in the biological sciences to assess the significance of pair-wise matches between DNA, RNA or protein sequences. A high Z-score corresponds to an alignment that is less likely to occur by chance. Such an alignment is more likely to be biologically meaningful. Z-scores are usually calculated by taking the mean and standard deviation over either a random sample taken from the set of all permutations of the query sequence [1] [2], or a sample from a standard database [3]. The mean is calculated as the mean score of the query sequence aligned with all sequences from the sample set. The standard deviation is taken over the same set of realigned sequence pairs. The Z-score is defined as

$$Z(A, B) = \frac{S(A, B) - \text{mean}}{\text{standard deviation}} \quad (1)$$

where $S(A, B)$ is the Smith-Waterman (S-W) score between two sequences A and B and $Z(A, B)$ is the Z-score. The S-W score of a local alignment between two proteins is found using standard dynamic programming [4] [5] techniques.

2 The POZ Score and the POZITIV algorithm

Say we have aligned two sequences A and B of different lengths. Some gaps were left in the alignment and these resulted in gap penalties which contributed to the S-W score. Now let us consider the subsequences *consisting only of the matched letters in the above alignment*. In other words, we ignore all letters which were aligned to a gap. The subsequences will be of equal length N say. We will refer to the two subsequences as \bar{A} and \bar{B} .

$$\begin{aligned} \bar{A} &= (A_{i_1}, A_{i_2}, A_{i_3}, \dots, A_{i_N}) \\ \bar{B} &= (B_{j_1}, B_{j_2}, B_{j_3}, \dots, B_{j_N}) \end{aligned}$$

where $i_k, j_l \in (1, 2, \dots, 20)$. A_i and B_j are letters in the amino acid alphabet.

If we used a substitution matrix $[m_{ij}]$ to align sequences A and B , then the S-W score, $S(A, B)$ is essentially the sum of the score of each of the pairs (A_{i_k}, B_{j_k}) corresponding to the matrix entry $m_{i_k j_k}$ with the gap penalties subtracted:

$$S(A, B) = \sum_{k=1}^N m_{i_k j_k} - \text{gap penalties.} \quad (2)$$

The POZ-score normalizes the S-W score with the (ungapped) scores of all permutations of sequence \bar{A} . It is defined as

$$POZ(A, B) = \frac{S(A, B) - \mu_{perm}}{\sigma_{perm}} \quad (3)$$

where μ_{perm} is the mean and σ_{perm} is the standard deviation of the scores of all permutations. The POZITIV algorithm calculates μ_{perm} and σ_{perm} efficiently and returns $POZ(A, B)$.

¹Centre for Bioinformatics Science (CBiS), Australian National University (ANU) E-mail: Hilary.Booth@anu.edu.au

²John Curtin School of Medical Research (JCSMR), ANU

³Mathematical Sciences Institute (MSI), ANU

⁴Australian Partnership for Advanced Computing (APAC)

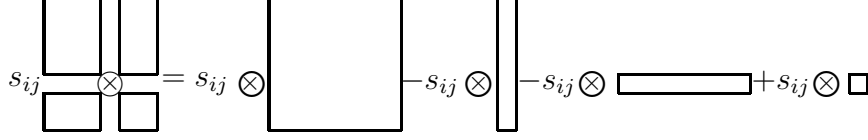


Figure 1: A Diagram showing an efficient calculation of the ij 'th cross term $s_{i,j} (\sum_{k \neq i, l \neq j} s_{k,l})$ in Equation (8) allowing the evaluation to be done on a 20×20 matrix.

3 Permutations are paths down the scoring matrix

In order to calculate the mean and standard deviation of the scores of all permutations of one of the sequences, we think of the permutations as all possible paths down the dynamic programming scoring matrix (with those letters removed which corresponded to gaps). Firstly, we want to find the mean over all *sums of scores* along each path. There are $(N-1)!$ possible paths passing through any particular element of the matrix. Each path is of length N . On average, the value along each path will be the average value in the matrix times N . That is, the mean score over all paths is simply the mean value of the matrix scores multiplied by N .

$$\mu_{perm} = \mu_{paths} = \mu_{matrix} \times N. \quad (4)$$

It is more difficult to calculate the standard deviation taken *down* (or *across*) all paths. Let us consider the standard deviation of all paths down the matrix.

Let the sum along a path be denoted by S_{path} so that

$$S_{path} = s_{i_1, j_1} + s_{i_2, j_2} + \dots + s_{i_N, j_N} \quad (5)$$

and the variance over the $N!$ paths is

$$\sigma_{perm}^2 = \sigma_{paths}^2 = \frac{1}{N!} \sum_{paths} S_{path}^2 - \mu_{paths}^2 \quad (6)$$

Now

$$\sum_{paths} S_{path}^2 = \sum_{paths} (s_{i_1, j_1} + s_{i_2, j_2} + \dots + s_{i_N, j_N})^2 \quad (7)$$

We can write this in terms of the $N \times N$ scoring matrix $[M_{ij}]$ as

$$\begin{aligned} \sum_{paths} S_{path}^2 &= (N-1)! (\sum_{i,j=1}^N s_{i,j}^2) + \\ &\quad (N-2)! (\sum_{i,j} s_{i,j} (\sum_{k \neq i, l \neq j} s_{k,l})) \end{aligned} \quad (8)$$

where the last term in (8) is the sum of the quadratic (cross) terms in (7). Note that $(N-2)!$ paths pass through any particular combination of two elements.

The calculation of the second term in Equation (8) can be efficiently handled by writing

$$\begin{aligned} C &= \sum_{i,j} s_{i,j} (\sum_{k \neq i, l \neq j} s_{k,l}) \\ &= (\sum s_{ij})^2 - \sum_i (\sum_j s_{ij})^2 \\ &\quad - \sum_j (\sum_i s_{ij})^2 + \sum s_{ij}^2 \end{aligned} \quad (9)$$

This calculation is illustrated in Figure 1.

4 Conclusion

Unlike the usual Z-score calculation used in sequence alignment [1] [2] we have developed a method (called POZITIV) in which both the mean and standard deviation are taken *without realigning the permuted sequences*.

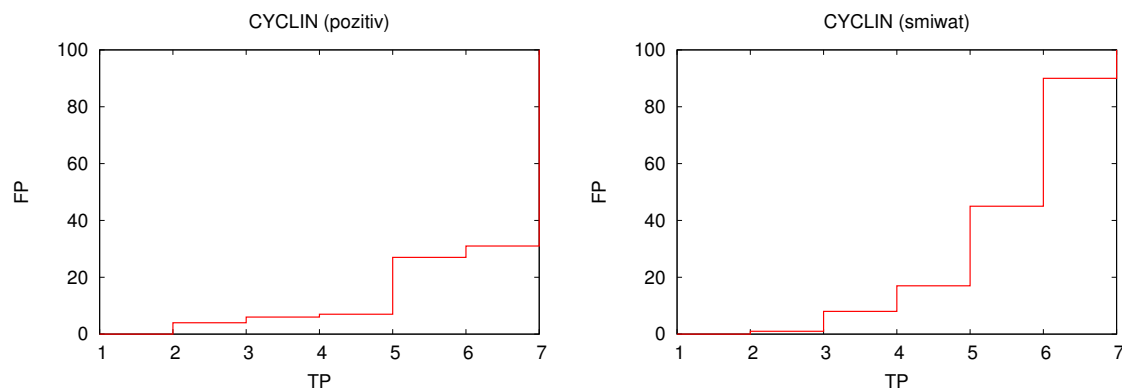


Figure 2: A ROC curve showing an example of one of the queries (CYCLIN) whose ROC100 value was improved by using POZITIV (left) compared with the S-W score (right).

Calculating the mean and standard deviation in this way enables us to calculate these exactly and in two steps, the first being $O(N)$ time, where N is the length of the sequence, the second in a fixed number of calculations i.e. in $O(1)$ time.

Some protein sequences in the aravinde dataset [6] benefit from this normalization procedure. Figure 2 gives an example of a protein query (CYCLIN) whose ROC100 value was improved by using POZITIV compared with the S-W score .

References

- [1] Comet J. P., Aude J. C., Glémet E., Risler J. L., Héneat A., Slonimski P. P. & Codani J. J. 1999. Significance of Z-value statistics of Smith-Waterman scores for protein alignments. *Computers and Chemistry* 23:317-331.
- [2] Comet J. P. and Bacro J. N. 2000. Sequence alignment: as approximation law for the Z-value with applications to databank scanning. *Computers and Chemistry* 25:401-410.
- [3] Pearson W. R. and Lipman D. J. 1988. Improved tools for biological sequence comparison *Proc. Natl. Acad. Sci. USA* 85:2444-2448.
- [4] Smith T. F. and Waterman M. S. 1981. Identification of common molecular subsequences *J. Mol.Biol.* 147:195-197.
- [5] Waterman M. S. and Vingron M. 1994. Sequence comparison significance and Poisson approximation. *Statistical Science* 9:367-381.
- [6] Schäffer, A. A., Aravind, L., Madden, T. L., Shavirin, S., Spouge, J. L., Wolf, Y. I. , Koonin, E. V., & Altschul, S. F. 2001. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucl. Acids Res.* 29:2994-3005.