New finite-size correction for local alignment score distributions

Yonil Park, Sergey Sheetlin, Ning Ma, Thomas L Madden and John L Spouge*

Abstract

**Background:** Local alignment programs often calculate the probability that a match occurred by chance. The calculation of this probability may require a “finite-size” correction to the lengths of the sequences, as an alignment that starts near the end of either sequence may run out of sequence before achieving a significant score.

**Findings:** We present an improved finite-size correction that considers the distribution of sequence lengths rather than simply the corresponding means. This approach improves sensitivity and avoids substituting an ad hoc length for short sequences that can underestimate the significance of a match. We use a test set derived from ASTRAL to show improved ROC scores, especially for shorter sequences.

**Conclusions:** The new finite-size correction improves the calculation of probabilities for a local alignment. It is now used in the BLAST+ package and at the NCBI BLAST web site (http://blast.ncbi.nlm.nih.gov).

Background

Local alignments are an essential tool for biologists and often provide the first information about the function of an unknown nucleotide or protein sequence. An important question concerns the relationship of the score of a local alignment with the probability that the alignment occurred by chance. Karlin and Altschul [1] developed an asymptotic theory for local alignments, assuming that no gaps are permitted. For two random sequences $I$ and $J$ of lengths $m$ and $n$, respectively, the resulting distribution of the optimal alignment score $\hat{M}$ approximates a Gumbel distribution [2]

$$P\{\hat{M} > y\} \approx 1 - \exp(-kmne^{-ly}). \quad (1)$$

The two statistical parameters in Equation (1) are $\lambda$, the scale parameter, and $k$, the pre-factor.

Several authors [3-12] extended this framework to local alignments with gaps and showed that the Gumbel distribution from Equation (1) is still valid, though different values for $\lambda$ and $k$ are required. Altschul [13] discussed the need for a “finite-size correction” to the lengths $m$ and $n$ to improve the accuracy of Equation (1). The resulting statistics are an integral part of the Basic Local Alignment Search Tool (BLAST) [14].

The following presentation emphasizes intuition over mathematical formality, to explain how the finite-size correction can account for the finite sequence lengths $m$ and $n$ to improve the accuracy of Equation (1). Let us begin with an optimal local alignment, which starts from score 0 and requires a non-zero sequence length within both $I$ and $J$, before it achieves score $y$. Let $L_I(y)$ ($L_J(y)$) be the required random lengths within both $I$ ($J$), and let $l_I(y) = E\{L_I(y)\}$ ($l_J(y) = E\{L_J(y)\}$) be the corresponding means. The main idea is that the optimal local alignment cannot start anywhere along the full length $m$ ($n$) of sequence $I$ ($J$), because there might be insufficient sequence to permit it to achieve the score $y$ (Figure 1). The finite-size correction described in [13] and used in BLAST therefore replaced the area $mn$ of the alignment matrix for Equation (1) by

$$|m - l_I(y)||n - l_J(y)|. \quad (2)$$

Equation (2) approximates the area within the alignment matrix where the optimal local alignment can start and on average still have enough space to exceed the score $y$. If $m < l_I(y)$ or $n < l_J(y)$, however, the resulting value in Equation (2) might become negative. The BLAST code for the old finite-size correction therefore set the corrected sequence length to an ad hoc value...
positions, expectations corresponding to symmetric scoring matrix and symmetric sequence compositions, however, e.g., \( l_I(y) \) and \( l_J(y) \) have different distributions, so the following retains the subscripts \( I \) and \( J \).

The new finite-size correction replaces \( mn \) in Equation (1) by

\[
\mathbb{E}\left\{ (m - L_I(y))^+ [n - L_J(y)]^+ \right\},
\]

where \( x^+ = \max\{x, 0\} \).

The practical computation of Equation (4) approximates the distribution of \( (L_I(y), L_J(y)) \) with a bivariate normal distribution, with means \( l_I(y) = \mathbb{E}L_I(y) \) and \( l_J(y) = \mathbb{E}L_J(y) \), variances \( \text{var} L_I(y) = v_I(y) \) and \( \text{var} L_J(y) = v_J(y) \), and covariance \( \text{cov} (L_I(y), L_J(y)) = c(y) \), all assumed to be linear in the score \( y \), i.e.,

\[
\begin{align*}
  l_I(y) &= a_I y + b_I, \\
  l_J(y) &= a_J y + b_J, \\
  v_I(y) &= a_I y + \beta_I, \\
  v_J(y) &= a_J y + \beta_J, \\
  c(y) &= \sigma y + \tau.
\end{align*}
\]

The estimation of the parameters \( a_I, a_J, \beta_I, \beta_J \) and \( \sigma \) has mathematical depth and involves many unproved speculations, but involves a heuristic modeling of a random sequence alignment with Markov additive processes [15], ultimately with use of the renewal-reward theorem.
The Appendix presents formulas for computing \( a_I, a_J, \alpha_I, \alpha_J \) and \( \sigma \).

BLAST \( p \)-values are relatively insensitive to the values of the intercepts \( b_I, b_J, \beta_I, \beta_J \) and \( \tau \), so the practical computation approximates them, as follows. Let \( a_u (\alpha_u) \) be the value of \( a_I (\alpha_I) \) for ungapped alignment. The mathematical theories for random walks and for renewals yield analytic formulas for \( a_u \) and \( \alpha_u \) [16]. For an ungapped optimal alignment, the alignment length required to exceed the score \( y \) is the same within the sequences \( I \) and \( J \), because it lacks gaps. Thus, \( a_u \) and \( \alpha_u \) do not depend on the sequence \( (I \) or \( J) \) under consideration, so they contain no subscripts \( I \) or \( J \). In a gapped alignment, let a gap of length 1 incur a penalty \( G \).

The following uncontrolled approximations hold [17]:

\[
\begin{align*}
    b_I &= 2G(a_u - a_I), \quad b_J = 2G(a_u - a_J) \\
    \beta_I &= 2G(\alpha_u - \alpha_I), \quad \beta_J = 2G(\alpha_u - \alpha_J) \\
    \tau &= 2G(\alpha_u - \sigma). 
\end{align*}
\]

Under the normal approximation, routine computation shows that Equation (4) is approximately

---

**Table 1** Retrieval accuracy for different subsets of SCOP database with the new and old finite-size correction

| Method          | 25\(^{th}\) percentile | 50\(^{th}\) percentile | Full database |
|-----------------|--------------------------|-------------------------|---------------|
| New correction  | 0.10373 ± 0.00022        | 0.10073 ± 0.00019       | 0.08535 ± 0.00013 |
| Old correction  | 0.09201 ± 0.00020        | 0.09282 ± 0.00017       | 0.08358 ± 0.00014 |

The three subsets contain proteins shorter than 91 residues (25\(^{th}\) percentile by length), shorter than 137 residues (50\(^{th}\) percentile by length), and the full database. ROC-4852 scores are presented with an error (one standard deviation). The 25\(^{th}\) percentile database contains 2533 sequences, the 50\(^{th}\) percentile database contains 5008 sequences, and the full database contains 10,569 sequences. There are 4852 queries.
We compared \( P \)-values for the new and old finite-size corrections with those for the old finite-size correction using the BLOSUM62 scoring matrix and affine gap penalty 11 + g. Hartmann used a rare-event simulation method to compute the local alignment score distribution for ranges that included small \( P \)-values like \( p = 10^{-50} \) [18], thereby producing a theoretical standard for small \( P \)-values.

Figure 2 plots relative errors in logarithmic scale against true \( P \)-values for equal sequence lengths \( m = n = 40, 100, 200, \) and 400. Using Hartmann’s theoretical standard, the new finite-size correction outperforms as the \( P \)-value decreases, sometimes by orders of magnitude.

Evaluation of accuracy
We evaluated the performance of the new finite-size correction using the ASTRAL SCOP 40 subset [19] of release 1.75 of the Structural Classification of Proteins (SCOP) [20] database. We sorted the SCOP domains by lexicographic order and used the even numbered sequences as our query set, but removed any query that was the sole member of the superfamily in ASTRAL 40. For a given query sequence, we considered any database sequence belonging to the same SCOP superfamily as a true positive, and any database sequence belonging to a different SCOP fold as a false positive. Following [21], in the retrieval list for each query, we censored all sequences belonging to the same fold but different superfamily, so those sequences contributed neither true or false positives to the retrieval.

We report the performance in terms of the Receiver Operator Characteristics (ROC). Specifically, we report the ROC\(_n\) score, which is obtained by pooling the results of all queries, ordering them by expect value, but only keeping results up the \( n \)-th false positive [21]. The expect value for the database search was obtained from the pairwise \( P \)-values using a length-proportional correction that takes the ratio of the database length to the target sequence length into account [13].

As discussed above, the new finite-size correction should show the greatest improvement for short sequences. Therefore, we also produced ROC\(_n\) scores for different subsets of the SCOP database. One database subset has sequences shorter than the 25th percentile length (95 residues), and another has sequences shorter than the 50th percentile length (137 residues).

Table 1 presents ROC\(_n\) scores for the full database as well as the two subsets described above. These scores have an average of one false positive per query (4852), a threshold found useful in other studies (Altschul SF, private communication). The ROC-4852 scores for the full database demonstrate a small improvement of the new finite-size correction over the older one. The subsets show a more impressive improvement. For the 50th percentile subset, the ROC-4852 score improves by 9%. For the 25th percentile subset, the ROC-4852 score shows a 13% improvement. In the 25th percentile subset, the new finite-size correction produces roughly 12% more true positives overall at 4852 false positives than the old finite-size correction (Figure 3). These results confirm our expectation that the new finite-size correction will display greatest improvement in retrieval for short sequences.

To assess the significance of this improvement on BLAST searches, one may look to the length distribution of sequences in a heavily used protein BLAST database. The non-redundant (“nr”) database is the default protein database at the NCBI BLAST web site. Of the sequences in the nr database, 11% are 95 residues or shorter; and 21%, 137 residues or shorter. The new finite-size correction improves the retrieval accuracy for a noticeable fraction of the proteins in the nr database.
Conclusion
We have described a new finite-size correction. The new correction has a more rigorous derivation than the current finite-size correction and avoids the use of an ad hoc value for short sequences. We have tested the retrieval accuracy of the new finite-size correction on the gold standard SCOP set, and have shown that the improvement is most important for short sequences. This correction has been made part of the BLAST+ protein-protein applications (e.g., BLASTP, BLASTX) as well as at the NCBI BLAST web site. In the future, we plan to implement this correction for nucleotide-nucleotide comparisons.

Availability and requirements
Project Name: BLAST Statistical Parameters
Project home page: http://www.ncbi.nlm.nih.gov/CCBresearch/Spouge/html_ncbi/html/blast/
Operating systems: Windows, MacOSX, LINUX, UNIX Programming language: C++
License: Public Domain (see http://www.ncbi.nlm.nih.gov/books/NBK22952/)

Appendix
Let $Z_+ = \{0,1,\ldots\}$. Consider two semi-infinite random sequences, yielding alignment scores $S_{i,j}$ at each vertex $(i,j) \in Z_+^2$, within their alignment graph. Define the edge maximum score $E_n = \max\{\max_{i,j} S_{i,j}, \max_{i,j} S_{i,j} \}$. Let $\kappa_0 = E_{\kappa_0} = 0$ and $\kappa_i = \inf\{n : n > \kappa_{i-1}, E_n > E_{\kappa_{i-1}}\}$ for $i \geq 1$. We call $\kappa_i$ the $i$th SALE (strict ascending ladder epoch) and $E_{\kappa_i}$ the $i$th SALE score. Let $\Delta E_i := E_{\kappa_i} - E_{\kappa_{i-1}}$, the increment between the $(i-1)$th and $i$th SALE scores.

Let $L_1(y) = \inf\{i : S_{i,j} \geq y, (i,j) \in Z_+^2\}$ and $L_2(y) = \inf\{i : S_{i,j} > E_{\kappa_i}, (i,j) \in Z_+^2\}$. We also define $L_n = \inf\{i : S_{i,j} = E_{\kappa_n}, (i,j) \in Z_+^2\}$ and $J_n = \inf\{i : S_{i,j} = E_{\kappa_n}, (i,j) \in Z_+^2\}$. Let $\Delta_1 := L_1 - L_1$, the incremental sequence length between $(i-1)$th and $i$th SALEs in sequence I, and $\Delta_j := J_j - J_{j-1}$, the incremental sequence length between $(j-1)$th and $j$th SALEs in sequence J. Last, we define $E^*\{\Delta_i\} := E[\Delta_i e^{\Delta_i}, \kappa_i < \infty]$, $E^{\ast}\{\Delta_j\} := E\left[\Delta_j e^{\Delta_j}, \kappa_j < \infty\right]$, and $E^{\ast}\{\Delta_i\} := E\left[\Delta_i e^{\Delta_i}, \kappa_i < \infty\right]$. The formulas for computing $a_i, b_i, a_j, \alpha_i$, $\alpha_j$ and $\sigma$ are:

\[
\begin{align*}
    a_i & = \lim_{i \to \infty} \frac{E^{\ast}\{\Delta_i\}}{E^{\ast}\{\Delta E_i\}}, \\
    b_i & = \lim_{i \to \infty} \frac{\var^\ast\{\Delta_i\}}{E^{\ast}\{\Delta E_i\}}, \\
    a_j & = \lim_{j \to \infty} \frac{\var^\ast\{\Delta_j\}}{E^{\ast}\{\Delta E_j\}}, \\
    \alpha_i & = \lim_{i \to \infty} \frac{\cov^\ast\{\Delta_i, \Delta E_i\}}{E^{\ast}\{\Delta E_i\}}, \\
    \alpha_j & = \lim_{j \to \infty} \frac{\cov^\ast\{\Delta_j, \Delta E_j\}}{E^{\ast}\{\Delta E_j\}}, \\
    \sigma & = \lim_{i \to \infty} \frac{\var^\ast\{\Delta_i, \Delta E_i\}}{E^{\ast}\{\Delta E_i\}},
\end{align*}
\]

where $\var^\ast$ and $\cov^\ast$ represent the variance and covariance associated with the probability measure underlying the expectation $E^\ast$. In practice, for computational efficiency, we use importance sampling to estimate the parameters above [15]. The parameters are estimated separately at each SALE. We then apply asymptotic regression to estimate the values of $\alpha_i, \alpha_j, \beta_i, \beta_j$ and $\sigma$ as $i \to \infty$ in the equation above [22].

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
YP, TM and JS drafted the manuscript. YP designed the $p$-value evaluation method. SS implemented the new finite-size correction. NM integrated the correction into the BLAST+ code, ran tests, and calculated the ROC scores. JS devised the new finite-size correction. YP and SS are equal contribution first authors for this article. TLM and JLS are equal contribution last authors for this article. All authors read and approved the final manuscript.

Acknowledgements
We thank Greg Boratyn for help in running the accuracy evaluations with the SCOP set. This research was supported by the intramural research program of the NIH, National Library of Medicine.

Received: 30 March 2012 Accepted: 16 May 2012
Published: 12 June 2012

References
1. Karlin S, Altschul SF: Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci U S A 1990, 87(6):2264–2268.
2. Galambos J: The asymptotic theory of extreme order statistics. New York: Wiley, 1978.
3. Mott R: Maximum likelihood-estimation of the statistical distribution of Smith-Waterman local sequence similarity scores. Bull Math Biol 1992, 54(1):59–75.
4. Waterman MS, Vingron M: Rapid and accurate estimates of statistical significance for sequence data base searches. Proc Natl Acad Sci U S A 1994, 91(1):4625–4628.
5. Altschul SF, Gish W: Local alignment statistics. Methods Enzymol 1996, 266:460–480.
6. Bundschuh R: Rapid significance estimation in local sequence alignment with gaps. J Comput Biol 2002, 9(2):243–260.
7. Chia N, Bundschuh R: A practical approach to significance assessment in alignment with gaps. J Comput Biol 2006, 13(2):429–441.
8. Newberg LA: Significance of gapped sequence alignments. J Comput Biol 2008, 15(9):1187–1194.
9. Agrawal A, Brendel VP, Huang X: Pairwise statistical significance and empirical determination of effective gap opening penalties for protein local sequence alignment. Int J Comput Biol Drug Des 2008, 1(4):347–367.
10. Poljakic A: Island method for estimating the statistical significance of profile-profile alignment scores. BMC Bioinformatics 2009, 10:112.
11. Oret P, Bastien O: Where does the alignment score distribution shape come from? Evol Bioinformatics 2010, 6:159–187.
12. Agrawal A, Huang X: Pairwise statistical significance of local sequence alignment using sequence-specific and position-specific substitution matrices. IEEE/ACM Trans Comput Biol Bioinformatics 2011, 8(1):194–205.
13. Altschul SF: Evaluating the statistical significance of multiple distinct local alignments. In Theoretical and computational methods in genome research. Edited by Suhai S. New York: Plenum Press; 1997:1–14.
14. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997, 25(17):3389–3402.
15. Park Y, Sheerlin S, Spouge J: Estimating the gumble scale parameter for local alignment of random sequences by importance sampling with stopping times. Ann Stat 2009, 37(6A):3697–3714.
16. Asmussen S: Applied probability and queues. New York: Springer; 2003.
17. Altschul SF, Bundschuh R, Olsen R, Hwa T: The estimation of statistical parameters for local alignment score distributions. Nucleic Acids Res 2001, 29(2):351–361.
18. Hartmann AK: Sampling rare events: statistics of local sequence alignments. Phys Rev E 2002, 65:056102. doi:10.1103/PhysRevE.65.056102.

19. Chandonia JM, Hon G, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE: The ASTRAL Compendium in 2004. Nucleic Acids Res 2004, 32:D189–D192.

20. Murzin AG, Brenner SE, Hubbard T, Chothia C: SCOP - a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995, 247(4):536–540.

21. Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF: Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res 2001, 29(14):2994–3005.

22. Sheetlin S, Park Y, Spouge JL: Objective method for estimating asymptotic parameters, with an application to sequence alignment. Phys Rev E 2011, 84(3). doi:10.1103/PhysRevE.84.031914.

doi:10.1186/1756-0500-5-286
Cite this article as: Park et al.: New finite-size correction for local alignment score distributions. BMC Research Notes 2012 5:286.