

Estimation of the Gumbel Scale Parameter for Local Alignment Score Distribution

Yonil Park¹, Sergey Sheetlin², John L. Spouge³

Keywords: Gapped sequence alignment, Gumbel scale parameter, Markov additive processes

1 Abstract.

In this poster, we present some non-rigorous numerical studies that use gapped global alignment simulations to estimate the Gumbel scale parameter for gapped local alignment statistics. For a given parameter accuracy, simulations of global alignment are more efficient than simulations of local alignment, because they require shorter sequence lengths (1).

Extensive numerical simulation studies (2-4) suggest that the distribution of local alignment score approximates a Gumbel extreme value distribution if the gap penalties are large enough. In the absence of an analytic formula for the Gumbel parameters, they must be estimated by computer simulations. In practice, estimates of scale parameter λ must have a 1% to 4% relative error and location parameter K must have less than 10% relative error. At present BLAST users are limited to particular scoring systems and gap penalties, because the Gumbel parameters must be computed off-line. If simulations were fast enough that λ and K could be calculated in less than about one second, however, BLAST users could employ arbitrary scoring systems and gap penalties.

For a pair of random sequences of length n , global alignment algorithm (5) calculates global alignment scores S_{ij} for $1 \leq i, j \leq n$. We now introduce an edge maximum $E_n = \max \left\{ \max_{1 \leq i \leq n} S_{in}, \max_{1 \leq j \leq n} S_{nj} \right\}$. A heuristic modeling approach to the global alignment based on Markov additive processes (6) suggests a regression model for simulating the Gumbel scale parameter λ as follows:

$$\ln \left(\mathbb{E} \left[\exp(\lambda E_n) \right] \right) = \beta_0 + \beta_1(\lambda)n + O(\varepsilon^n),$$

where β_0 is the constant and $\beta_1(\lambda)$ is the slope. Solving the equation $\beta_1(\lambda) = 0$ yields a new estimate of Gumbel scale parameter. We used Bundschuh's importance sampling method (1) to estimate $\mathbb{E} \left[\exp(\lambda E_n) \right]$.

2 Figure.

¹ National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD, USA. E-mail: park@ncbi.nlm.nih.gov

² National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD, USA. E-mail: sheetlin@ncbi.nlm.nih.gov

³ National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD, USA. E-mail: spouge@ncbi.nlm.nih.gov

Figure 1 Plots 20 independent estimates for λ for the default scoring system in the current protein-protein BLAST. Every point used 2000 realizations of a pair of sequences with length 40. The horizontal line $\lambda = 0.267$ represents the previous best estimate of the asymptotic constant λ (7). The average of absolute relative errors is 0.69%, which is less than our goal. Our numerical study shows that for the default scoring system in the current protein-protein BLAST, λ can be computed to less than 1% relative error in less than 2 sec on a PC.

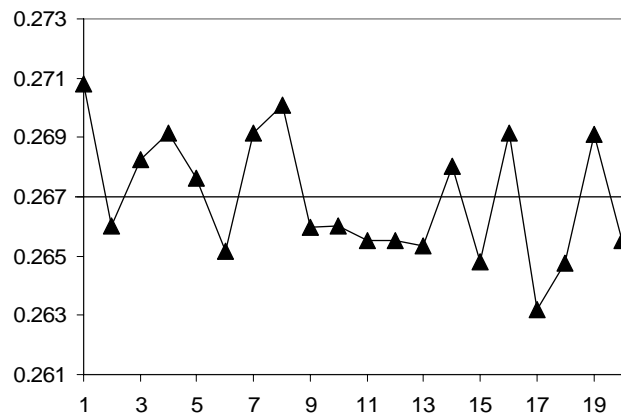


Figure 1: Estimation of lambda from global alignment simulation

References

1. Bundschuh, R. (2002) Rapid significance estimation in local sequence alignment with gaps *Journal of Computational Biology*, **9**, 243-260.
2. Olsen, R., Bundschuh, R. and Hwa, T. (1999) Rapid assessment of extremal statistics for gapped local alignment *Proc. 7th Int. Conf. Intelligent Systems for Molecular Biology*, 211-222.
3. Waterman, M.S. and Vingron, M. (1994) Rapid and accurate estimates of statistical significance for sequence data base searches *Proc Natl Acad Sci U S A*, **91**, 4625-4628.
4. Mott, R. (2000) Accurate formula for p-values of gapped local sequence and profile alignments *Journal of Molecular Biology*, **300**, 649-659.
5. Needleman, S.B. and Wunsch, C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins *Journal of Molecular Biology*, **48**, 443-453.
6. Asmussen, S. (2003) *Applied Probability and Queues*. Springer-Verlag, New York.
7. Altschul, S.F., Bundschuh, R., Olsen, R. and Hwa, T. (2001) The estimation of statistical parameters for local alignment score distributions *Nucleic Acids Research*, **29**, 351-361.