Significance of Gapped Sequence Alignments

LEE A. NEWBERG

ABSTRACT

Measurement of the statistical significance of extreme sequence alignment scores is key to many important applications, but it is difficult. To precisely approximate alignment score significance, we draw random samples directly from a well chosen, importance-sampling probability distribution. We apply our technique to pairwise local sequence alignment of nucleic acid and amino acid sequences of length up to 1000. For instance, using a BLOSUM62 scoring system for local sequence alignment, we compute that the p -value of a score of 6000 for the alignment of two sequences of length 1000 is $(3.4 \pm 0.3) \times 10^{-1314}$. Further, we show that the extreme value significance statistic for the local alignment model that we examine does not follow a Gumbel distribution. A web server for this application is available at <http://bayesweb.wadsworth.org/alignmentSignificanceVI/>.

Key words: alignment, dynamic programming, genomics, phylogenetic trees, regulatory regions.

1. INTRODUCTION

THE ALIGNMENT OF A PAIR OF SEQUENCES is a fundamental task in genomics and proteomics, but methods for precisely gauging the significance of an alignment are not efficient (Mitrophanov and Borodovsky, 2006). Here we present an efficient method for precisely approximating the significance (also termed p -value) of an alignment score. We apply the approach to the alignment of sequences of length 1000 to precisely approximate p -values nearly as low as 10^{-4000} . Furthermore, we observe that, for local sequence alignments, the $\log(\text{significance})$ values approximately scale with the input sequence lengths; thus, the results of our method for select sequence length sizes can be scaled to arbitrary sequence lengths.

1.1. Notation and problem statement

Let (x, y) be a pair of sequences, composed of amino acid residues or nucleotides, with lengths (m, n) , and let A be an alignment of the sequences. We choose to allow an alignment to have an insertion into y immediately following an insertion into x , but not vice versa; however, alternative approaches should demonstrate similar results. We choose to focus on local sequence alignments, where initial and final unaligned regions are permitted without penalty, rather than on global or semi-local sequence alignments, although the general approach presented here should also work well for such alignments.

Let $s(x, y, A)$ be the score associated with x and y under the alignment A . It will be the sum of contributions from pairing scores, insertion start costs, and insertion extension costs. For instance, with SSEARCH (Pearson, 1991), the default parameters for sequences of nucleotides are: a pairing receives a score of +5 if the paired nucleotides match and a score of −4 if they are a mismatch; the first nucleotide of an insertion into a sequence incurs a score of −16; and extension of an insertion by a nucleotide incurs a score of −4.

We choose to define the alignment score for a pair of sequences (x, y) as

$$s(x, y) = s_{\max}(x, y) = \max_A \{s(x, y, A)\}, \quad (1)$$

the maximum achievable alignment score for the sequence pair, computable via the Viterbi algorithm (Viterbi, 1967) of Needleman and Wunsch (1970) or Smith and Waterman (1981), or a variant of either algorithm, depending upon the alignment type of interest. The general approach presented here should also work well with alternative choices for $s(x, y)$, such as the “forward sum” alignment score, its logarithm the “free” alignment score, or the “average” alignment score,

$$s_{\text{forward}}(x, y) = \sum_A e^{s(x, y, A)/T} \quad (2)$$

$$s_{\text{free}}(x, y) = T \ln \left(\sum_A e^{s(x, y, A)/T} \right) \quad (3)$$

$$s_{\text{average}}(x, y) = \left(\sum_A s(x, y, A) e^{s(x, y, A)/T} \right) / \left(\sum_A e^{s(x, y, A)/T} \right) \quad (4)$$

for some $T > 0$.

The definition of the statistical significance of a score s_0 is

$$p(s_0) = \sum_{\text{all } (x, y)} \Pr(x, y) \Theta(s(x, y) \geq s_0), \quad (5)$$

where the sum is over all sequence pairs (x, y) with lengths (m, n) , where $\Pr(x, y)$ is the null-model probability of the sequence pair (x, y) , as determined by GC-content or database residue frequency, and where $\Theta(\cdot)$ is a function that has value 1 if its argument is true and has value 0 otherwise. Because no one knows how to efficiently perform this sum exactly for x and y of moderate or large size, it is approximated via importance sampling (Hartmann, 2002; Wolfsheimer et al., 2007).

1.2. Importance sampling

We could sample N sequence pairs (x, y) directly from the probability distribution $\Pr(x, y)$, and then compute $s(x, y)$ for each sampled sequence pair. The fraction of the sampled sequence pairs for which $s(x, y) \geq s_0$ would thus be an estimate $\hat{p}(s_0)$ of the statistical significance of a score s_0 . However, the error in the estimate is

$$\sqrt{\text{Var}[\hat{p}(s_0)]} = \sqrt{\frac{p(s_0) - p(s_0)^2}{N}} \approx \sqrt{\frac{p(s_0)}{N}}, \quad (6)$$

which will be greater than $p(s_0)$ itself unless $N \geq 1/p(s_0)$. Because it is not unusual for $p(s_0)$ to be 10^{-10} or smaller, the cost of using $N \geq 1/p(s_0)$ samples is often prohibitive.

However, we can make the approximation more efficient by using importance sampling (Liu, 2001). We observe that

$$p(s_0) = \sum_{\text{all } (x,y)} \Pr(x, y) \Theta(s(x, y) \geq s_0) \quad (7)$$

$$= \sum_{\text{all } (x,y)} \Pr_*(x, y) \frac{\Pr(x, y) \Theta(s(x, y) \geq s_0)}{\Pr_*(x, y)} \quad (8)$$

$$= \sum_{\text{all } (x,y)} \Pr_*(x, y) f_*(x, y) \quad (9)$$

for any probability distribution $\Pr_*(x, y)$, where the last equality defines $f_*(x, y)$. Thus, if we draw samples according to $\Pr_*(x, y)$, we can estimate $p(s_0)$ as the average value over these samples of the function $f_*(x, y)$. This estimator $\widehat{p}_*(s_0)$ has an uncertainty of

$$\sqrt{\text{Var}[\widehat{p}_*(s_0)]} = \sqrt{\frac{\left[\sum_{\text{all } (x,y)} \Pr_*(x, y) f_*(x, y)^2 \right] - p(s_0)^2}{N}}, \quad (10)$$

which, we shall soon show, is significantly better than Equation (6) when $\Pr_*(x, y)$ is well chosen.

1.3. Previous work

Hartmann (2002) and Wolfsheimer et al. (2007) use $\Pr_*(x, y) \propto \Pr(x, y) e^{s_{\max}(x,y)/T}$ for some value T . To draw samples from this probability distribution, they employ a Metropolis-coupled Markov chain Monte Carlo technique (MCMCMC), wherein an (x, y) sequence pair at a given iteration is modified at a single sequence position, giving a sequence pair (x', y') ; the new sequence pair is then accepted with probability $\min\{1, \exp(\frac{s(x', y') - s(x, y)}{T})\}$, or is otherwise rejected. A collection of N nearly independent samples is derived from this process, by allowing sufficiently many steps to occur between consecutive samples, thus allowing proper mixing to occur. For computing p -values as extreme as 10^{-70} , they use this technique to approximate the significance of scores from the pairwise local alignment of two amino acid sequences, using the BLOSUM62 scoring matrix, SwissProt residue frequencies, and sequence of length up to 800.

Instead, we choose

$$\Pr_*(x, y) \propto \Pr(x, y) e^{s_{\text{free}}(x,y)/T}, \quad (11)$$

defining $\Pr_*(x, y)$ with the free alignment score of a sequence pair rather than the maximum alignment score of the pair. To proceed, we must choose a value for T ; we must demonstrate how to efficiently sample from this probability distribution; and we must demonstrate that the resulting estimate is statistically efficient.

2. METHODS

We explain the mathematics of the approach here, leaving some of the implementation details until Section 6 in the Supplementary Materials. (See online supplementary material at www.liebertonline.com.)

2.1. Efficient sampling

We define forward sum functions (also termed partition functions):

$$Z(x, y, T) = \sum_{\text{all } A} e^{s(x,y,A)/T} \quad (12)$$

$$Z(T) = \sum_{\text{all } (x,y,A)} \Pr(x, y) e^{s(x,y,A)/T}. \quad (13)$$

The value of $Z(x, y, T)$ is computed via a forward-sum adaptation of the relevant Viterbi algorithm for finding a maximum score alignment. In short, wherever a score s would be added to a value, instead a factor $e^{s/T}$ multiplies the value; whenever a maximum of alternative values would be chosen, instead the sum of these values is taken. For a discussion on forward algorithms for pair-HMM sequence alignment, see Durbin et al. (2006), or for a more general discussion on forward algorithms for hidden Markov models, see Rabiner and Juang (1986).

The value of $Z(T)$ is computed by use of the single mean value

$$\langle e^{s/T} \rangle = \sum_{\text{all } (a,b)} q_a q_b e^{s(a,b)/T} \quad (14)$$

in the forward-sum alignment algorithm in lieu of any individual value, $e^{s(a,b)/T}$, for a given nucleotide/residue pair (a, b) . Here q_a is the null-model probability of the nucleotide/residue a , as determined by GC-content or database residue frequency, and similarly for q_b .

We proceed by sampling an (x, y, A) triplet from the probability distribution

$$\Pr_T(x, y, A) = \frac{\Pr(x, y) e^{s(x,y,A)/T}}{Z(T)}, \quad (15)$$

via a backtrace through the forward-sum alignment algorithm that computes $Z(T)$. In that backtrace, we sample the alignment (i.e., the locations of pairings and insertion starts and extensions) in the usual way. See Durbin et al. (2006) for a discussion on probabilistic sampling for sequence alignment, and see Newberg (2008) for the memory efficient backtrace procedure that we employed. During the backtrace, we also sample the paired and unpaired nucleotides/residues that comprise the sequences x and y . An unpaired nucleotide/residue a is sampled with probability q_a , whereas a pair of nucleotides/residues (a, b) is sampled with probability

$$p_{ab} = \frac{q_a q_b e^{s(a,b)/T}}{\langle e^{s/T} \rangle}. \quad (16)$$

Then, by disregarding the sampled alignment A , we are, in effect, sampling a sequence pair (x, y) from the marginal probability distribution

$$\Pr_T(x, y) = \frac{\Pr(x, y) \sum_{\text{all } A} e^{s(x,y,A)/T}}{Z(T)}, \quad (17)$$

the promised importance sampling probability distribution of Equation (11). Thus, we estimate the significance of a score s_0 as indicated by Equation (8):

$$\widehat{p}_T(s_0) = \frac{1}{N} \sum_{\text{sampled } (x,y)} \frac{\Pr(x, y) \Theta(s(x, y) \geq s_0)}{\Pr_T(x, y)} \quad (18)$$

$$= \frac{1}{N} \sum_{\text{sampled } (x,y)} \frac{Z(T) \Theta(s(x, y) \geq s_0)}{Z(x, y, T)}, \quad (19)$$

where the sums are over the sequence pairs sampled from the $\Pr_T(x, y)$ probability distribution.

We estimate the uncertainty in the estimate $\widehat{p}_T(s_0)$ via the samples themselves:

$$\sqrt{\text{Var}[\widehat{p}_T(s_0)]} = \sqrt{\frac{\left[\frac{1}{N} \sum_{\text{sampled } (x,y)} \left(\frac{Z(T) \Theta(s(x,y) \geq s_0)}{Z(x,y,T)} \right)^2 \right] - \widehat{p}_T(s_0)^2}{N-1}}. \quad (20)$$

2.2. Choice of T

The choice for T depends upon the value of s_0 for which we wish to compute the significance. We find that, for s_0 sufficiently high, a good choice for T is one for which s_0 is near the average score for an (x, y, A) triplet drawn from the $\Pr_T(x, y, A)$ probability distribution. That is, we use T such that

$$s_0 \approx \sum_{\text{all } (x,y,A)} \Pr_T(x, y, A) s(x, y, A) = -T^2 \frac{d \ln Z(T)}{dT}. \quad (21)$$

We find a value for T by using the Newton method for zero-finding on the function $s_0 + T^2 \frac{d \ln(Z(T))}{dT}$, where the derivative is approximated through differencing of $Z(T)$ for two nearby values of T . This works well for high scores, which are usually the most interesting; unfortunately, when s_0 is less extreme, for example, when $p(s_0) \geq 10^{-6}$, the above procedure produces a value of T that is too low. For a description of how we handle this situation, see Section 6 in the Supplementary Materials. (See online supplementary material at www.liebertonline.com.)

We also have had some success with choosing the temperature so that the threshold s_0 is near the median $s_{\max}$ score of 11 sampled sequence pairs.

3. RESULTS

Figure 1 graphs the results of our significance calculations for five problem instances. We performed calculations on amino acid sequences using BLOSUM62 scores, SwissProt residue frequencies, an insertion start penalty of 12, and an insertion extension penalty of 1 for sequences of length $(m, n) \in \{(40, 40), (1000, 1000)\}$. We also performed calculations on nucleotide sequences using the default SSEARCH parameters (a score of +5 for a pairing of matching nucleotides, a score of -4 for a pairing mismatch, a score of -16 for an insertion start, and a score of -4 for an insertion extension) for sequences of length $(m, n) \in \{(100, 100), (100, 1000), (1000, 1000)\}$.

For example, the results for the $(m, n) = (1000, 1000)$ problem for amino acid sequences were computed to a tight precision using $N = 10,000$ samples for each of 76 temperatures (see Section 6 in the Supplementary Materials; see online supplementary material at www.liebertonline.com), with each temperature requiring 13–14 hours of run time. We were able to make these calculations in parallel, one Friday night, with 76 of the CPUs in our cluster. Typical relative error is approximately $\sigma = 10\%$ of the computed p -value; thus, the absolute error for $\log_{10}(p\text{-value})$ is approximately 0.05. The running time is approximately $\mathcal{O}(Nmn)$ for each temperature.

4. DISCUSSION

For the alignment of a pair of amino acid sequences of length 1000, we are able to precisely approximate the significance of alignment scores, even though the p -values can be nearly as low as 10^{-4000} (Fig. 1). Pre-computation of p -values for a variety of interesting alignment types and problem sizes is feasible; thus, lookup tables could fruitfully be employed by current alignment packages to report p -values to their users.

Furthermore, the similarity among the curves within each panel in the right half of Figure 1 indicates that significance scales nicely with the input sequence lengths (m, n) . We can approximate the significance of a score s_0 for input sequence lengths (m, n) from the significance values for input sequence lengths $(\hat{m}, \hat{n})$ as

$$\log_{10} p(s_0|m, n) \approx \log_{10} p(s_0 \hat{M} / M | \hat{m}, \hat{n}) \frac{\log_{10} p(M|m, n)}{\log_{10} p(\hat{M}|\hat{m}, \hat{n})}, \quad (22)$$

where M is the maximum achievable alignment score for input sequence lengths (m, n) , and $\hat{M}$ is the maximum achievable alignment score for input sequence lengths $(\hat{m}, \hat{n})$, and where these maximum achievable scores and their p -values are calculated exactly by inspection.

FIG. 1. Significance of sequence alignment scores. **(Left)** The panels give $\log_{10}(p\text{-value})$ as a function of alignment score. For instance, a score of 6000 for the alignment of two amino acid sequences of length 1000 residues is $(3.4 \pm 0.3) \times 10^{-1314}$. (The data used to make these graphs are available in Section 7 in the Supplementary Materials. See online supplementary material at www.liebertonline.com.) **(Right)** The panels compare problem instances of difference sizes on an equal footing, by a scaling of both axes of the graph so that the curves coincide at their endpoints; that the curves within each of these panels nearly coincide can be used to estimate p -values for one pair of sequence lengths using the data for another pair of sequence lengths (see Equation (22)).

In the limit, as the length of the sequences being aligned tends to infinity, the maximum scores for local *ungapped* alignments are expected to follow a Gumbel distribution

$$p(s_0) \approx G(s_0) = 1 - e^{-\alpha e^{-\beta s_0}}, \quad (23)$$

for some positive values α and β (Karlin and Altschul, 1990; Eddy, 2008). When s_0 is large,

$$G(s_0) \approx \alpha e^{-\beta s_0}, \text{ and} \quad (24)$$

$$\ln(G(s_0)) \approx \ln(\alpha) - \beta s_0. \quad (25)$$

Accordingly, the curves in Figure 1 might be expected to be straight lines, at least for high values of s_0 . For the pairwise alignment of amino acid sequences of length up to 800 and p -values down to 10^{-70} , Wolfsheimer et al. (2007) observes that these curves are not straight lines, and we now extend that observation to p -values nearly as extreme as 10^{-4000} .

Wolfsheimer et al. (2007) hypothesizes that the tails of these curves are parabolic. We find that the best-fit parabola of the $\log_{10}(p\text{-value})$ data for the BLOSUM62 tail scores 5,500–11,000 is $-3.66480 \times 10^{-5}s_0^2 + 0.120156s_0 - 720.633$. The best-fit parabola has a root mean square error of approximately 14 with respect to our estimates. That is fairly good, although well in excess of the typical 0.05 error estimated from our simulations. In comparison, for these BLOSUM62 tail scores, the root mean square error for the best-fit line is 84; the error for the best-fit cubic is 6.7; and the error for the best-fit quartic is 3.6.

Because the run time of the algorithm, for fixed sequence lengths m and n , is approximately $\mathcal{O}(\sigma^{-2})$, the calculation can be orders of magnitude faster if good precision is not required. For instance, the example of a score of 6000 for amino acid sequences of length 1000 that is given in Figure 1, can be computed as $\log_{10}(p\text{-value}) = -1314 \pm 1$ with $N = 100$ samples in 39 minutes using one of our processors, and likely even faster once our prototype software attains production quality. Using the $N = 100$ reduced number of samples, an alignment of two sequences of length 10,000 requires approximately 13–14 hours to compute the value $\log_{10}(p\text{-value}) \approx -13,100$ for $s_0 = 60,000$. This p -value might also have been estimated much more quickly, via Equation (22), from the -1314 value just mentioned.

We have focused on local sequence alignments with certain parameters, but the technique is easily applied to other sets of parameters and other alignment models. For instance, we conjecture that this technique will work well with profile-HMMs (Durbin et al., 2006), which permit the alignment model parameters (e.g., the nucleotide/residue frequencies and the indel penalties) to vary by position within a sequence.

Further, we conjecture that this technique will also be efficient when used to estimate extreme statistical significance values for a much broader set of applications, such as many hidden Markov models that emit a sequence of data (Durbin et al., 2006; Rabiner and Juang, 1986). For a given sequence of data, often it is the case that we wish to know a maximum-probability/Viterbi path through the model. The statistical significance of the log-probability of such a path can be estimated via our technique, as can the statistical significance of the sum of the probabilities over all paths emitting the given data sequence. We conjecture that, in many cases, such estimates will be quite efficient. Likewise, we conjecture efficiency for many of the even broader class of finite state automata, which are much like hidden Markov models, except that the scores are not necessarily derived from log probabilities; for instance, often scores are derived from (or interpreted as) log odds, the logarithm of the ratio of the probabilities from given alternative and null models (Karlin and Altschul, 1990).

5. CONCLUSION

We have presented a method for efficiently and accurately approximating the p -value for a pairwise alignment score. The approach can be used for a variety of definitions of local or global alignment and for a variety of definitions of alignment score. Via the pre-computation of p -values for a selection of interesting problem instances, algorithms that make pairwise alignments can employ scaling to quickly approximate the p -value for any given problem instance. A web server for this application is available at <http://bayesweb.wadsworth.org/alignmentSignificanceV1/>.

ACKNOWLEDGMENTS

We thank the Computational Molecular Biology and Statistics Core Facility at the Wadsworth Center for the computing resources to make these calculations. We thank Charles E. Lawrence for helpful ideas and suggestions. This research was supported by the National Institutes of Health/National Human Genome Research Institute (grant K25 HG003291 to L.A.N.).

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

- Durbin, R., Eddy, S., Krogh, A., et al. 2006. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK.
- Eddy, S.R. 2008. A probabilistic model of local sequence alignment that simplifies statistical significance estimation. *PLoS Comput. Biol.* 4, e1000069.
- Hartmann, A.K. 2002. Sampling rare events: statistics of local sequence alignments. *Phys. Rev. E* 65, 056102.
- Karlin, S., and Altschul, S.F. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. USA* 87, 2264–2268.
- Liu, J.S. 2001. *Monte Carlo strategies in scientific computing*. Springer Series in Statistics. Springer-Verlag, New York.
- Mitrophanov, A.Y., and Borodovsky, M. 2006. Statistical significance in biological sequence analysis. *Brief Bioinform.* 7, 2–24.
- 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. *J. Mol. Biol.* 48, 443–453.
- Newberg, L.A. 2008. Memory-efficient dynamic programming backtrace and pairwise local sequence alignment. *Bioinformatics* 24, 1772–1778.
- Pearson, W.R. 1991. Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics* 11, 635–650.
- Rabiner, L.R., and Juang, B.-H. 1986. An introduction to hidden Markov models. *IEEE ASSP Mag.* 3, 4–16.
- Smith, T.F., and Waterman, M.S. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* 147, 195–197.
- Viterbi, A.J. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Inform. Theory* 13, 260–269.
- Wolfsheimer, S., Burghardt, B., and Hartmann, A.K. 2007. Local sequence alignments statistics: deviations from gumbel statistics in the rare-event tail. *Algorithms Mol. Biol.* 2, 9.

Address reprint requests to:

Dr. Lee A. Newberg

Wadsworth Center

NYS Department of Health

P.O. Box 509

Albany, NY 12201-0509

E-mail: lee.newberg@wadsworth.org