SEPA: Approximate Non-Subjective Empirical p-Value Estimation for Nucleotide Sequence Alignment

Ofer Gill and Bud Mishra

Courant Institute of Mathematical Sciences,
New York University, 251 Mercer Street, New York NY 10012, USA,
gill@cs.nyu.edu,
http://bioinformatics.nyu.edu/~gill/index.shtml

Abstract. In the bioinformatics literature, pairwise sequence alignment methods appear with many variations and diverse applications. With this abundance, comes not only an emphasis on speed and memory efficiency, but also a need for assigning confidence to the computed alignments through p-value estimation, especially for important segment pairs within an alignment. This paper examines an empirical technique, called SEPA, for approximate p-value estimation based on statistically large number of observations over randomly generated sequences. Our empirical studies show that the technique remains effective in identifying biological correlations even in sequences of low similarities and large expected gaps, and the experimental results shown here point to many interesting insights and features.

1 Introduction

In the field of comparative genomics, an emphasis is placed on its functional genomics aspects. Most often we align two or more sequences, because we expect that the important areas selected from that alignment will point to a significant common biological function, even when we realize that there can be no absolute guarantee of this. In order to draw our attention very quickly to the most pertinent similar subsequences, it is necessary to compare the important areas of alignments and rank them in order of their relevance. For instance, by comparing alignments in related sequences to those of unrelated sequences with no common biological function, we may derive, for any alignment, the probability that its important areas occur by mere coincidence. This probability measure is also known as a p-value, and low p-values relate to high relevance rank.

Many p-value estimation techniques have been suggested and examined previously, for instance, Karlin-Altschul [4] and Siegmund-Yakir [7], but none have proven completely satisfactory. In this paper, we focus on using empirical results to improve the p-value approximation in case of alignments of noncoding nucleotide sequences of lengths varying from .5 Kb to 12 Kb, with expected large gaps and low similarities. These alignments are often computed with the complex but biologically faithful model involving piecewise-linear gap penalty functions as in PLAINS [2]; nonetheless, other techniques such as LAGAN and EMBOSS have also proven effective. We demonstrate the effectiveness of a p-value approximation technique called SEPA (Segment Evaluator for Pairwise Alignments) as it selects and scores important segments pairs. Furthermore, for random sequences, we also empirically characterize how various alignment statistics, such as the segment pair lengths, scores, and magnitudes, distribute as a function of sequence lengths. From this analysis, the parameters for a p-value approximation are estimated, and used to demonstrate the method of sensitivity in distinguishing important homologies from unimportant chance occurrences of subalignments within sequences. Furthermore, SEPA is non-subjective, since it can easily be applied to any alignment tool. We will illustrate this advantage by using it to compare the results of PLAINS with LAGAN and EMBOSS. Because of these strengths and despite its empirical foundation, SEPA fulfills a practical computational need by speeding up the core search processes in comparative genomics.

2 Overview

We introduce some notations as follows: Assume the sequences to be aligned are X and Y, and their respective lengths are m and n, where $m \geq n$. Let X_u and Y_v denote respectively the u^{th} character of X and the v^{th} character of Y, where $1 \leq u \leq m$ and $1 \leq v \leq n$.

Let us suppose that aligning X and Y with some arbitrary alignment tool produces an alignment A of length a, where $m \leq a \leq m + n$. We will represent an alignment A as follows: For each i, $A[i]$ denotes the i^{th} position in alignment A, and it is represented as a pair of index coordinates (u, v) taken from X and Y, and this corresponds to X_u and Y_v being aligned to each other at position i in A if $u > 0$ and $v > 0$, or one of X_u or Y_v being aligned against a gap if either $v \leq 0$ or $u \leq 0$.

Next, let $A[i : j]$ denote the portion of alignment $A[i], A[i+1], \ldots, A[j]$. We will refer to $A[i : j]$ as a *strip* or *segment pair* from position i to position j.

Let $ww(i)$ denote the penalty for a gap of length i. $ww(\cdot)$ can be any arbitrary function, but for this paper, we will assume it is a d-part piecewise-linear function where each successive slope is smaller than the previous one. A more specific version of this score-function is where $d = 1$, which is the affine function used in the Smith-Watermann algorithm.

Also, let $S(i, j)$ denote the score for strip $A[i : j]$ where the score is computed in the following way: m_a is a score for each match, m_s is the penalty for each mismatch, and $ww(\cdot)$ is used to penalize the gaps. To compute $S(i, j)$ from $A[i : j]$, each match and mismatch within it is added or deducted from the score individually, while each region of X against a gap and Y against a gap is penalized as a whole using $ww(\cdot)$ based on the length of that region.[1]

[1] For this paper, the m_a reward is 1, the m_s penalty is 0.346445, and the $ww(i)$ penalty is the piecewise-linear approximation of $1.01742 + 1.015724 \ln(i + 1)$. This

Suppose we have a scheme that marks r non-overlapping strips as important. Suppose that the endpoints for these strips are denoted as $(i_1, j_1), \ldots, (i_r, j_r)$. For each k, we wish to measure in some way how strip $A[i_k : j_k]$ provides a meaningful correlation between X and Y. One common mathematical approach is to, given a certain null hypothesis, compute the p-value of $Pr(x \geq s)$ where $s = S(i_k, j_k)$. This p-value is known as the coincidental probability of obtaining a strip with score at least s. For this paper, we will assume the null-hypothesis is the behavior of important strips taken from pairwise-aligning randomly generated DNA sequences. Also, if the total scores of all strips is $t = \Sigma_{k=1}^{r} S(i_k, j_k)$, then $\zeta = Pr(x \geq t, y \leq r)$, the probability of obtaining at least a total score of t using at most r strips.

One should note that coincidental probabilities of the segments (both p-values and ζ) are dictated by the scheme used to determine the segments as important. One scheme might deem strip $A[i : j]$ as important, but SEPA might not, and instead SEPA may consider a possibly overlapping strip $A[i' : j']$ as important. As a result, the formula for the p-values and ζ value could differ from one scheme to the other. For instance, in the method used to obtain important segments mentioned in Karlin-Altschul [4], $Pr(x \geq s) = 1 - exp(Kmne^{-\lambda s})$ holds. However, as argued later in this paper, the way SEPA obtains the segments from an alignment A leads us to approximate the p-value as $Pr(x \geq s) = \frac{K}{\lambda} e^{-\lambda s}$.

2.1 Obtaining High-Scoring Strips from an Alignment

Given an alignment A produced from sequences X and Y, we produce important strips as follows: Given fixed constants W and ω, and ρ (where W is an integer, and ω and ρ are real numbers in the range $[0, 1]$), let W denote the window size to be used, ω denote the value used to prevent portions of A of lowest match percentage from becoming considered as important strips, and ρ denote the value used to filter away areas of A that have too low of a p-value. We obtain our segment pairs in the following steps:

(1) For all i from 1 to $a - (W - 1)$, we compute $p_a(i)$, the percentage of entries in $A[i : i + W - 1]$ where a match has occurred. Let μ and σ denote the mean and standard deviation of our $p_a(\cdot)$ values. Next, for each i, we mark[2] $p_a(i)$ values as "special" if they exceed a threshold value of $\mu + \omega\sigma$. Hence, we filter away $A[i : i + W - 1]$ if it fails to meet this threshold value.

(2) For each u and u' (with $u \leq u'$), if $p_a(u)$, $p_a(u + 1)$, …, $p_a(u')$ are all marked as "special", but $p_a(u - 1)$ and $p_a(u' + 1)$ are not, then we consider the strip $A[u : u' + W - 1]$ as important (i.e., we consider as important the strip starting the leftmost entry repsented by $p_a(u)$, up till the rightmost entry represented by $p_a(u')$).

selection empirically provides a good numerical contrast in scores between strips of high homology, and strips of lower homology.

[2] The choice of using $\mu + \omega\sigma$ as the cutoff value instead of a fixed constant gives us the flexibility of catching important regions in the two sequences, regardless of how homologous they are to each other.

(3) For each strip $A[i:j]$ deemed important, we trim it so that it starts and ends at a position in the alignment where a match occurred. Thus, if i' is the smallest value such that $i' \geq i$ and $A[i']$ is a match position, and j' is the largest value such that $j' \leq j$ and $A[j']$ is a match position, then we trim strip $A[i:j]$ into strip $A[i':j']$.

(4) Next, we merge together any important strips that overlap. Namely, if we have two strips $A[i:j]$ and $A[k:l]$ such that $i \leq k \leq j$, then we merge these strips into one larger strip $A[i:\max(j,l)]$.

(5) With all strips now representing non-overlapping regions, we then proceed to give each strip $A[i:j]$ its corresponding score $S(i,j)$, as well as its p-value. We delete $A[i:j]$ if its p-value exceeds ρ, since that indicates that $A[i:j]$ may be coincidental. We can optionally also collect other information at this point, such as the length of each strip.

(6) The r strips kept at this step are considered the "good" ones. We now compute t, the sum of the scores of the these strips. Using this value, we can compute ζ, coincidental probability for all r strips obtained.

Note that these steps for SEPA are similar to that of [2], except that the calculation for each segment pair's coincidental probability differs. Based on empirical experimentation, setting $W = 50$, $\omega = 0.5$, and $\rho = 0.5$ yields segment pairs that are reasonably long, non-coincidental, and have significantly higher matches than the alignment "background". We reasoned that since our method of obtaining segment pairs differs from that of Karlin-Altschul, then the method for computing p-values for each segment pair cannot build upon their assumptions.

2.2 Methods: Analyzing Segment Pairs

In order to approximate an appropriate p-value estimation for SEPA, we analyzed segment pairs behavior over our assumed null hypothesis of alignments for randomly generated nucleotide sequences. For length values ranging from 1000 bp to 8000 bp, we generated 25 random sequences. We also generated 25 random sequences of length 500 bp. For each combination of these length pairs, we ran all 625 possible pairwise alignments using Plains, and analyzed results using SEPA where $\rho = 1$ (to avoid filtering any segments out due to low p-value). The results for mean length-to-score and mean mean segment scores are shown in Fig. 1. From this, we infer that both are constant in terms of m and n. We also analyzed the behavior of r (the number of segment pairs observed) and t (the total score of all the segments) over these random sequences, and found them to depend on m and n.[3] See the full version of this paper[4] for figures that elaborate further.

[3] The mean of r is $\approx 10^3 + \ln(7.95 \times 10^{-10} mn + 1.54 \times 10^{-7}(m+n) + 1.01)$, the variance of r is $\approx 10^3 + \ln(1.93 \times 10^{-10} mn + 1.97 \times 10^{-7}(m+n) + 1.00)$, the mean of t is $\approx 10^5 + \ln(4.29 \times 10^{-10} mn + 1.33 \times 10^{-8}(m+n) + 1.00)$, and the deviation for t is $\approx \max\{100, -5.54 \times 10^{-5} j \cdot d + 4.63 \times 10^{-1} j + 1.04 \times 10^{-2} d - 65.01\}$, where $j = \min(m,n)$, and $d = \|m-n\|$.

[4] `http://bioinformatics.nyu.edu/~gill/index.shtml`

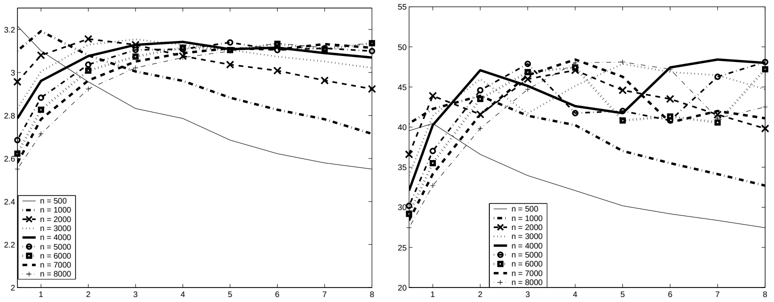

Fig. 1. Shown in the above graphs respectively are the mean length-to-score ratio and mean segment scores observed in the strips from aligning randomly generated DNA sequences. In these plots, a unique line is plotted corresponding to each value of n in the thousand lengths ranging from 1000 to 8000, and x represents the m value divided by 1000, and y represents the mean observed for that particular m and n. These plots indicate that, for small n values, the average length-to-score ratio and average score decrease with increasing m. However, asymptotically (for large n) the average length-to-score ratio and average segment scores stay roughly constant in terms of m (at 3.1 : 1 and 45 respectively), and attempts at using Gumbel distributions failed to provide better approximations than this.

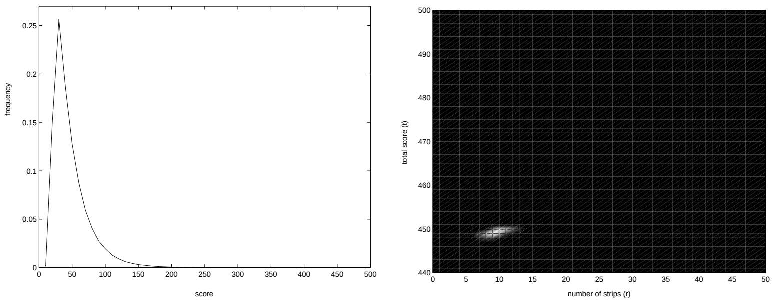

Fig. 2. On the left, is a plot of segment scores (x-axis) to frequency (y-axis) for randomly generated sequences using our assumption that segment score is length-independent. On the right, is a surface plot of observed frequency for number of segments r and total score t after adjusting both variables for average and variance behaviors, where lighter spots indicate higher frequencies. From the left graph, the tail of this plot for $x \geq 30$ indicates an exponential distribution of form $P(S = x) = Ke^{-\lambda x}$, where $K = 8.69 \times 10^{-2}$ and $\lambda = 3.26 \times 10^{-2}$. From the right graph, we observe that the majority of the data is concentrated in one area, and this area approximates to $e^c e^{-a_t T^2 + b_t T + c_t} e^{-a_r R^2 + b_r R + c_r}$, where $c = -183.90, a_t = 10.1, b_t = 9070, c_t = -2.04 \times 10^6, a_r = 0.241, b_r = 4.71, c_r = -27.5$.

Since the average ratio of segment lengths to score is almost uniform in these plots, it suggests that the gap penalty used to score the strips can be treated as if it is a differently-weighted mismatch. Also, note that the p-values computed with the model studied by Siegmund-Yakir[7] differs mildly from the model using the simplifying assumption that gaps are differently-weighted mismatches. For this reason, it is common for tools to ignore the effects of gaps in generating their p-values, much like BLAST, and in our case, SEPA as well. The left graph in fig. 2 plots segment scores to frequency, from which we derive our p-value approximation. Using it, we approximate that $P(x = s) = Ke^{-\lambda s}$, and our p-value of $P(x \geq s)$ is therefore:

$$P(x \geq s) = \int_s^\infty Ke^{-\lambda x}dx = \frac{K}{\lambda}e^{-\lambda s}$$

And notice that by this construction, $P(x \geq 30) = \frac{K}{\lambda}e^{-30\lambda} \approx 1$. We have designed our p-value estimation this way since strip scores below 30 are empirically observed to be unimportant.

Our next natural step, after obtaining p-values for each segment pair, is to provide a p-value estimate ζ for coincidental probability for the whole alignment, determined by the strips found. As mentioned earlier, we have learned that both r and t depend on sequence lengths m and n. Hence, if R and T are supposed to be the number of segment pairs and the total score of the segment pairs after adjusting for mean and variance based on sequence length, then the coincidental probability $\zeta = P(x \geq T, y \leq R)$. More specifically, ζ is the coincidental probability of seeing a total score of at least T using at most R segment pairs.

The right graph in figure 2 shows the distribution of r and t values observed from randomly generated sequences after adjusting for mean and variance. From it, our approximation of T and R for $P(x = T, y = R)$ gives us for ζ that[5]:

$$\zeta = P(x \geq T, y \leq R) = \int_T^\infty \int_0^R e^c e^{-a_t x^2 + b_t x + c_t} e^{-a_r y^2 + b_r y + c_r} dydx =$$

$$= \frac{\pi e^{c+c_t+c_r+\frac{b_t^2}{4a_t}+\frac{b_r^2}{4a_r}}}{4\sqrt{a_t a_r}} \left(1 - Erf\left(\frac{-b_t + 2a_t T}{2\sqrt{a_t}}\right)\right)\left(Erf\left(\frac{-b_r + 2a_r R}{2\sqrt{a_r}}\right) - Erf\left(\frac{-b_r}{2\sqrt{a_r}}\right)\right)$$

Table 1 shows a comparison of alignments for biologically related sequences in terms of unadjusted r and t values, and ζ' values, where $\rho = 0.5$, and $\zeta' = -\ln(\zeta)$. Note that Plains does not always yield the results of least coincidental probability in this table. This is because the nature of Plains is to aggressively align as many regions as possible. This produces higher r values than other tools, which hurts its ζ' values. In order to better understand how r and t vary for all the tools used, we chose to observe what happens when we vary ρ, and figure 3 elaborates in detail for one such experiment. From it, we see that, for any fixed r, the r best segment pairs generated by Plains have smaller coincidental probabilities than the best r segment pairs generated by other tools.

[5] Note that $Erf(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-x^2} dx$

Test Name	PLAINS			LAGAN			EMBOSS		
	t	r	ζ'	t	r	ζ'	t	r	ζ'
fugu2r	534.14	5	11.15	360.22	3	13.05	151.39	2	14.07
HFortho1	734.82	7	10.94	349.33	4	14.18	374.35	5	13.05
HFortho2	600.22	4	16.78	555.61	4	16.78	327.91	1	20.18
HFortho3	637.52	7	14.53	259.44	3	19.05	409.99	5	16.71
HFortho4	1004.97	10	21.74	529.16	5	0.00	367.86	4	0.00
HFortho5	739.71	7	11.07	450.93	5	13.07	453.61	5	13.07
human_mouse.1_1	676.29	10	8.46	52.36	1	18.29	186.98	2	17.00
human_mouse.1_3	552.55	6	15.14	406.79	6	15.14	429.51	6	15.14
human_mouse.3_9	1260.69	15	15.47	432.25	7	24.23	801.15	12	18.44
human_mouse.4_3	262.19	3	15.44	74.91	1	17.79	176.83	2	16.59
human_mouse.4_5	421.71	6	7.35	221.57	3	10.47	401.71	5	8.32
human_mouse.7_11	594.32	8	9.06	164.10	2	15.44	476.71	7	9.99
human_mouse.17_11	608.75	7	13.93	171.96	3	18.57	451.60	6	15.02
human_dog.6_12	1284.79	14	13.88	548.19	7	21.23	394.04	6	22.44
human_dog.7_16	1042.19	13	10.45	128.07	2	22.40	309.03	4	19.84

Table 1. Shown here for Plains, EMBOSS, and LAGAN are the r, t, and ζ' values obtained from aligning genomic DNA sequences of lengths between 0.5 Kb and 12 Kb within human, mouse, dog, and fugu, where the pairs are biologically related and mainly noncoding DNA with expected large gaps and low homology regions. The conversion from ζ to ζ' was carried out for convenience in comparing lab results, where higher ζ' indicates results that are less coincidental. Also, note the loss of precision involved in reporting ζ' values. Hence, if for a paricular alignment, Plains and LAGAN receive ζ' values that differ by less than 1×10^2, then their ζ' values would "appear" equal in this table. Further information regarding the sequences used can be found in the full version of this paper at site `http://bioinformatics.nyu.edu/~gill/index.shtml`.

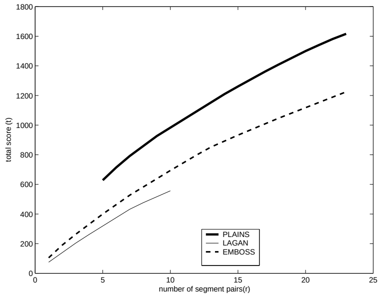

Fig. 3. In this figure, we observe the unadjusted r and t values produced by Plains, LAGAN, and EMBOSS from the human-mouse.3 – 9 experiment where we vary the ρ variable used to filter our segment pairs. On each curve, we observed the t and r values of each tool when varying ρ over various values from 0.1 till 0.9. Observe from table 1 that Plains performed poorly in terms of ζ' values for $\rho = 0.5$ for the human-mouse.3 – 9 experiments. The key to note here is that the r-to-t ratio is almost uniform for all tools, but the y-intercept differs from one tool to the next, with Plains having the highest y-intercept. This means that, for any fixed r, Plains yields higher t and hence better ζ' results. Many other experiments from table 1 have a similar plot to this one.

3 Conclusions and Future Work

Our empirical analysis leads us to the conclusion that the SEPA-based p-value technique models coincidental probabilities much more accurately than the earlier technique employed in [2]. Furthermore, we note that aggressively incorporating too many segments into an alignment can lower the ζ' coincidental probability value, in spite an apparent improvement in the total score, as illustrated by Plains. However, if we keep only the best r segments from an alignment, the strength of Plains becomes obvious, since its r segments are less coincidental than its competition.

However, in spite of the promising results from SEPA, there is still plenty of room for further improvements by using random portions of DNA from Human, Mouse, and Fugu instead of randomly generated DNA sequences. In that case, our concern shifts from the coincidental probability of a segment's score from aligning random DNA, to the coincidental probability of a segment's score from aligning unrelated random regions of organisms under comparison. Further extension includes development of better statistics that realistically captures the base-pair and coding/noncoding distributions within the sequences, as well as the effects of secondary and tertiary structures.

References

1. Brudno, M., Do, C., Cooper, G., Kim, M.F., Davydov, E., Green, E.D., Sidow, A., Batzoglou, S.: LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Research **13(4)** (2003) 721–731
2. Gill, O., Zhou, Y., Mishra, B.: Aligning Sequences with Non-Affine Gap Penalty: PLAINS Algorithm. Series in Mathematical Biology and Medicine **8** (2005). An unabridged version can be found at: `http://bioinformatics.nyu.edu/~gill/index.shtml`
3. Gu, X., Li, W.H.: The size distribution of insertions and deletions in human and rodent pseudogenes suggests the logarithmic gap penalty for sequence alignment. J. Mol. Evol. **40(4)** (1995) 464–473
4. Karlin, S., Altschul, S.F.: Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc. Natl. Acad. Sci. USA **87** (1990) 2264–2268
5. Karlin, S., Altschul, S.F.: Applications and statistics for multiple high-scoring segments in molecular sequences. Proc. Natl. Acad. Sci. USA **90** (1993) 5873–5877
6. Rice, P., Longden, I., Bleasby, A.: EMBOSS: the European Molecular Biology Open Software Suite. Trends Genetics **Jun 16(6)** (2000) 276–277
7. Siegmund, D., Yakir, B.: Approximate p-Values for Local Sequence Alignments. The Annals of Statistics **28 (3)** (2000) 657–680
8. Smith, T.F., Waterman, M.S.: Identification of Common Molecular Subsequences. Journal of Molecular Biology **147** (1981) 195–197
9. Shpaer, E., Robinson, M., Yee, D., Candlin, J., Mines, R., Hunkapiller, T.: Sensitivity and Selectivity in Protein Similarity Searches: A Comparison of Smith-Waterman in Hardware to BLAST and FASTA. Genomics **38** (1996) 179–191
10. States, D.J., Gish, W., Altschul, S.F.: Basic Local Alignment Search Tool. Journal of Molecular Biology **215** (1990) 403–410
11. Zhang, Z., Gerstein, M.: Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. Nucleic Acids Res. **31(18)** (2003) 5338-5348