BIOINFORMATICS

Vol. 00 no. 00 200x
Pages 1–4

The effect of sequence quality on sequence alignment

Ketil Malde

Institute of Marine Research, Bergen, Norway

ABSTRACT

Motivation: The nucleotide sequencing process produces not only the sequence of nucleotides, but also associated quality values. Quality values provide valuable information, but are primarily used only for trimming sequences and generally ignored in subsequent analyses.

Results: This paper describes how the scoring schemes of standard alignment algorithms can be modified to take into account quality values to produce improved alignments and statistically more accurate scores. A prototype implementation is also provided, and used to post-process a set of BLAST results. Quality-adjusted alignment is a natural extension of standard alignment methods, and can be implemented with only a small constant factor performance penalty. The method can also be applied to related methods including heuristic search algorithms like BLAST and FASTA.

Availability: http://malde.org/~ketil/qaa.

Contact: ketil.malde@imr.no

1 INTRODUCTION

Nucleotide sequencing is a process that is subject to errors, and when base calling software analyses chromatograms to determine the sequence, they commonly output quality values that estimates the error probability of each position. Although errors in the sequences are frequent, the quality values have been shown to accurately reflect the probability of incorrect base calls (Ewing and Green, 1998; Walther *et al.*, 2001). Quality values thus provides important information to subsequent stages in the analysis process.

Quality values are commonly used in sequence assembly (Huang and Madan, 1999; Green, 1999) to assist in determining the correct base for the consensus sequence. A similar use is in SNP analysis (e.g., Marth *et al.* 1999), where using quality information can help to differentiate between read errors and true polymorphism.

Aside from the sequence assembly and SNP analysis, quality information is rarely utilized directly, and in particular the ubiquitous database search tools like BLAST (Altschul *et al.*, 1990) and FASTA (Lipman and Pearson, 1988) make no particular provision for sequence quality. Consequently, it is sometimes suggested that low quality fragments of sequences be removed before searching, a process often referred to as *trimming*.

There exists a multitude of tools to perform quality based trimming — e.g. Pregap4 from the Staden package (Staden *et al.*, 1998) and Lucy and Lucy2 (Chou and Holmes, 2001; Li and Chou, 2004), and Phred also incorporates sequence trimming information in its output (Ewing *et al.*, 1998) — trimming puts the user in the unfortunate situation of having to choose between retaining dubious sequence parts or discarding a potentially large fraction of the data set. The former will reduce specificity of searches, while the latter will reduce sensitivity, and striking on optimal balance can be difficult.

This paper describes how standard sequence alignment algorithms can be extended to incorporate quality information in the alignments. The current implementation supports the standard dynamic programming algorithms for local and global alignment (Smith and Waterman, 1981; Needleman and Wunsch, 1970; Gotoh, 1982), but the method extends naturally to more commonly used heuristic methods like BLAST.

1.1 Substitution matrices

A substitution matrix defines the score for the substitution of each possible pair of letters, including the substitution of a letter with itself. For nucleotide sequences, often a uniform background distribution is assumed and the substitution matrix is simplified to two values: a positive score (or *reward*) in the case of identical letters, and a a negative score (*penalty*) for a mismatch.

Substitution matrices are normally calculated with log-odds scores. If the substitution $x \rightarrow y$ occurs with frequency q_{xy} in alignments, and x and y occur with respective frequencies f_x and f_y in the data set, the substitution is given the score (Altschul, 1991)

$$s_{xy} = \frac{1}{\lambda} \ln(q_{xy}/f_x f_y) \tag{1}$$

The scaling factor λ can be chosen freely, but a common choice is $\lambda = \ln 2$, which scales the alignment scores to bit units. All scores reported here will use this choice for λ. When the background distribution is uniform with each letter occurring with frequency f, and noticing that $q_{xy} = f_x P(x \rightarrow y|x)$, the scores can be simplified to $s_{xy} = \log_2(P(x \rightarrow y|x)/f)$.

For nucleotides, we assume uniform background distribution of $p = 0.25$, and probability ϵ for errors independent of base. Each of the three possible incorrect bases thus has probability $\epsilon/3$, and substituting in equation 1 gives

$$s_{xy} = \begin{cases} \log_2((1-\epsilon)/0.25) & \text{for } x = y \\ \log_2(\epsilon/(3*0.25)) & \text{for } x \neq y \end{cases} \tag{2}$$

A substitution matrix corresponds to a specific distance between the sequences being aligned. In most cases, this distance is evolutionary distance, but the principle applies equally well to distance incurred by random errors.

2 METHODS

For each position in the called sequence, the base calling software outputs a quality value that corresponds to the estimated probability of an incorrect base call. A given quality Q corresponds to the error probability ϵ where $\epsilon = 10^{-Q/10}$.

In order to determine the score of a substitution between two sequences at a position where the sequences have error rates of ϵ_1

and ϵ_2, we first calculate a combined error rate

$$\epsilon = \epsilon_1 + \epsilon_2 - \frac{4}{3}\epsilon_1\epsilon_2 \quad (3)$$

since there is a probability of $\frac{1}{3}\epsilon_1\epsilon_2$ that there is a match between two erroneous base calls.

The combined error in Equation 3 can then be applied to Equation 2. The resulting function generates the substitution scores dynamically for each pair of positions, effectively adjusting the substitution matrix for each position depending on the reliability of the base calls. Modifying the standard dynamic programming algorithms to use this function instead of static values is straightforward, and it should apply equally to similar heuristic methods.

2.1 Implementation

In the current implementation, quality adjustment is applied to the standard dynamic programming algorithms with affine gap penalties for local and global alignment. Quality adjustment is currently supported for nucleotide sequences only, and as discussed above, a uniform distribution of nucleotides is assumed.

Calculating substitution scores as a function of quality values is a computationally expensive operation, and the substitution scores for quality values from 0 to 99 are pre-calculated and stored in a table. Although the present implementation has not been optimized, this allows the quality adjusted algorithm to run with a speed close to the standard algorithm with fixed substitution scores.

BLASTN uses a default match reward of 1 and mismatch penalty of -3. This corresponds to comparing sequences with constant quality values of approximately 22. (This is equivalent to a single sequence error rate of 0.63%, and a combined error rate of 1.25%.) For sequences where no quality value is provided, the implementation therefore defaults to a quality value of 22.

Given that low quality sequence is likely to contain insertion or deletion errors, a reasonable argument can be made for reducing the severity of gap penalties according to quality. Unfortunately, the theory for choosing gap penalties is less developed than for substitution scores, and the current implementation makes the conservative choice of retaining fixed values for gap opening (-10) and gap extension (-4).

3 RESULTS

EST clustering is a common application that relies on comparing nucleotide sequences. ESTs are typically clustered using sequence similarity, attempting to group sequences that originate from the same gene. In practice, sequences that are found to have an adequate match are clustered together, and consequently, sequences that fail to have such matches remain as singletons.

Note that the comparison is usually not simply an alignment score, but often based on specific limits on e.g. sequence identity, alignment length, and unaligned overlap. In addition, chimeric sequences and splice variants can match other sequences partially. It is therefore not uncommon that singletons have good matches against other clustered sequences, and this does not necessarily reflect weaknesses in the clustering algorithm.

As an example usage of sequence alignment with quality adjustment, we will investigate the clustering of a set of 33 852 EST sequences from sea louse (*Lepeophtheirus salmonis*). Clustering these sequences, using TGICL Pertea *et al.* (2003) with default options resulted in 3 441 contigs and 14 185 singletons.

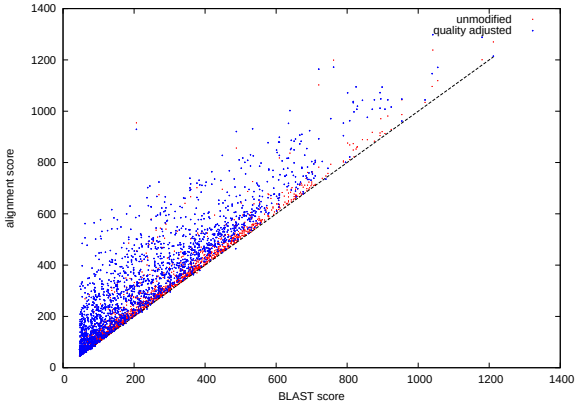

Fig. 1. Alignment scores as a function of BLASTN scores. The alignments are calculated using Smith-Waterman local alignment and BLASTN default scores (red) and alignment scores using quality-adjusted alignment scores (blue). The line is $f(x) = x$, indicating equal BLAST and Smith-Waterman alignment scores.

The sequences were classified according to average quality score, and sequences with an average quality score of less than 15 were selected for further study. Of 6 003 low quality sequences, only 384 were clustered with other sequences, while 5 619 low quality sequences were left as singletons by the clustering process.

The low quality sequences were aligned against the contigs using BLASTN with the e-value threshold set to 10^{-5}. Each match was then recalculated using Smith-Waterman local alignment, and using quality-adjusted Smith-Waterman.

Figure 1 shows the alignment scores for Smith-Waterman and quality-adjusted Smith-Waterman plotted against the BLAST scores for the same sequence pair. As expected, the standard Smith-Waterman alignment achieves scores close to the corresponding BLAST score, while the quality adjusted alignment tends to achieve higher scores.

Of the 1 285 distinct sequences that scored hits above the given threshold of 10^{-5}, 530 achieved BLAST scores higher than 300 bits. For quality adjusted alignment, 765 sequences scored higher than 300 bits..

Figure 2 shows the alignment score as a function of alignment length. For short alignments, scores close to the optimum of two bits per nucleotide can be seen – representing alignments involving a short but high-quality segment.

Figure 3 shows the difference in alignment score as a function of the difference in alignment length. The quality adjusted alignment tends to achieve somewhat longer alignments by being able to extend more easily across low-quality regions, but for the majority of alignments, the difference is small. Alignment score is generally substantially higher for quality adjusted alignment, reflecting that most sequencing errors occur in positions with low quality values, and that high-quality positions tend to match.

Figure 4 shows the quality of an example sequence. BLASTN reports a match between this sequence and a contig from position 202 to position 337, with an additional short, exact match from position 428 to 444. Local alignment with Smith-Waterman

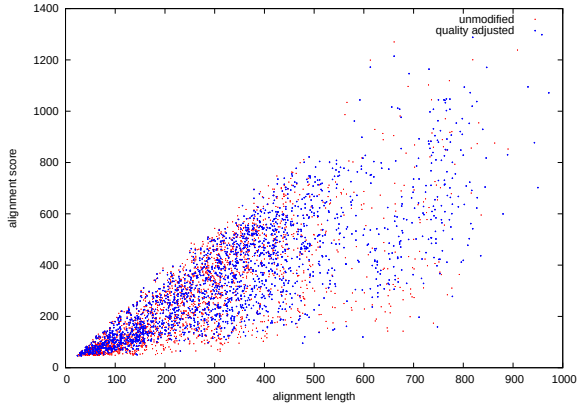

Fig. 2. Alignment scores as a function of alignment lengths for Smith-Waterman local alignments with default BLASTN parameters (red), and using quality-adjusted alignment scores (blue).

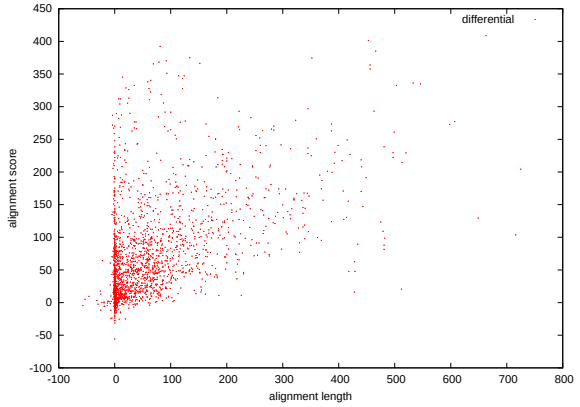

Fig. 3. Difference in alignment score and alignment length between Smith-Waterman local alignment, and the quality-adjusted local alignment. Positive values indicate that the quality-adjusted alignments are longer or higher scoring.

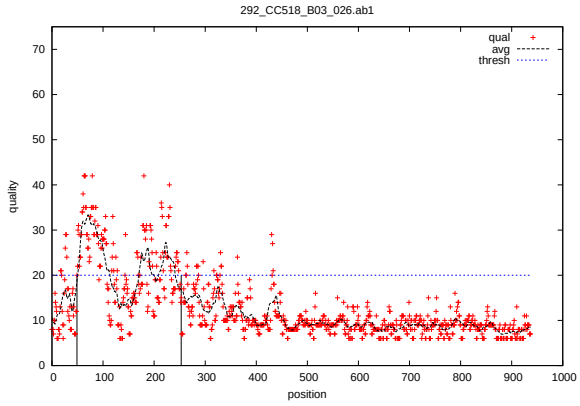

Fig. 4. The quality values of a sequence plotted against sequence position. Each position is represented by one point, the dashed line shows the sliding average over a 20-base window. The vertical bars show the trimming parameters suggested by Phred, indicating a high quality segment between positions 49 and 253.

identifies the former as the highest scoring match. Quality adjusted alignment identifies a match from position 192 to 549. This alignment scores 138.7 bits, compared to 48.1 bits with BLASTN and 48.9 bits with Smith-Waterman alignment.

4 DISCUSSION

As we have seen, quality information can be incorporated in standard sequence alignment algorithms. This generally improves the accuracy of score estimation, and in some cases, reveals long alignments, typically because they include regions of low quality which would otherwise terminate the alignment prematurely.

As noted by (States *et al.*, 1991), using scoring matrices with the correct PAM distance (Dayhoff *et al.*, 1978) is important to achieve correct results. As quality tend to vary substantially along the sequence length, a static substitution matrix will necessarily be suboptimal, at least for parts of the sequence. For instance, the average sequence quality of the sequence from Figure 4 is 12.7. Running BLASTN with the corresponding parameters `-q -2 -r 4` gives a match from position 192 to 376, which is longer than the default parameters, but shorter than the quality adjusted alignment.

Sequence assembly is one main area where quality information is routinely used, but only for calculating the final consensus. As trimming is often conservative in order to retain as much sequence as possible, sequences often have low quality regions, typically near the ends, which complicates the overlap alignments used in the assembly process. One way to address this is to have a configurable overhang parameter which allows matches to have an unaligned segment below a certain length at the end of the sequence (Pertea *et al.*, 2003). Using a quality based alignment method is likely to produce more accurate alignments, while eliminating this parameter.

A similar problem occurs when ESTs are matched against the genome to build clusters (Malde and Sczyrba, unpublished). ESTs occasionally match the genome with a short match between the end of the EST and a position genome sequence at considerable distance downstream from the rest of the match. In many cases, this is more likely to be caused by poor quality of the EST sequence than by a very long intron. Using a quality-adjusted alignment would help to classify this correctly.

Database searching with nucleotide sequences is an important tool for annotating unidentified sequences. When a poor quality sequence is aligned using static matching scores, matches will often be terminated prematurely. In addition to the incorrect significance estimate, this means that even if the sequence has a full length match in the database, the short alignments can easily be mistaken as matches against conserved domains, motifs, or binding sites. In contrast, quality adjusted alignments can bridge or incorporate the low quality segments in the alignment, and thus help to provide more accurate annotations.

Quality adjusted scores can be implemented as a table look-up, which incurs a constant time and space penalty compared to a static matrix. One optimization commonly used is scaling the scoring matrix (by careful selection of the λ parameter) in order to use integer calculations. The quality adjusted matrix requires a much wider range of values, and it is not clear whether an integer approximation is practical. Using floating point values can incur an additional constant time penalty.

The current implementation provides local (Smith-Waterman) and global (Needleman-Wunsch) alignments and scoring for nucleotide sequences. The general method is applicable to any algorithm with a similar scoring scheme, and it should be straightforward to extend BLAST or other search tools based on e.g. PSSMs or hidden Markov models with similar functionality.

Quality could also be incorporated in nucleotide-protein and protein-protein alignments, especially in the context of translated searches where the sequences may be less reliable. One possible option is to average qualities when translating ORFs, but as low quality sequences are likely to cause frame shifts, using an approach suggested by (Guan and Uberbacher, 1996) is likely to be more sensitive.

5 FUNDING

The Institute of Marine Research, Bergen, Norway.

REFERENCES

Altschul, S., Gish, W., Miller, W., Myers, E., and Lipman, D. (1990). A basic local alignment search tool. *Journal of Molecular Biology*, **215**(3), 403–410.

Altschul, S. F. (1991). Amino acid substitution matrices from an information theoretic perspective. *Journal of Molecular Biology*, **219**, 555–565.

Chou, H.-H. and Holmes, M. H. (2001). DNA sequence quality trimming and vector removal. *Bioinformatics*, **17**(12), 1093–1104.

Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. C. (1978). *A model of evolutionary change in proteins*, volume 5, pages 345–352. National Biomedical Research Foundation, Washington, DC.

Ewing, B. and Green, P. (1998). Base-calling of automated sequencer traces using Phred. II. error probabilities. *Genome Research*, **8**, 186–194.

Ewing, B., Hillier, L., Wendl, M. C., and Green, P. (1998). Base-calling of automated sequencer traces using Phred. I. Accuracy asessment. *Genome Research*, **8**, 175–185.

Gotoh, O. (1982). An improved algorithm for matching biological sequences. *Journal of Molecular Biology*, **162**, 705–708.

Green, P. (1999). Phrap documentation. `http://www.phrap.org/phredphrap/phrap.html`.

Guan, X. and Uberbacher, E. C. (1996). Alignments of DNA and protein sequences containing framshift errors. *CABIOS – Computer Applications in the Biosciences*, **12**(1), 31–40.

Huang, X. and Madan, A. (1999). CAP3: A DNA sequence assembly program. *Genome Research*, **9**, 868–877.

Li, S. and Chou, H.-H. (2004). LUCY2: an interactive DNA sequence quality trimming and vector removal tool. *Bioinformatics*, **20**(16), 2865–2866.

Lipman, D. J. and Pearson, W. R. (1988). Improved tools for biological sequence comparison. In *Proceedings of the National Academy of Science of the USA*, volume 85, pages 2444–2448.

Marth, G. T., Korf, I., Yandell, M. D., Yeh, R. T., Gu, Z., Zakeri, H., Stitziel, N. O., Hillier, L., Kwok, P.-Y., and Gish, W. R. (1999). A general approach to single-nucleotide polymorphism discovery. *Nature Genetics*, **23**, 452–456.

Needleman, S. and Wunsch, C. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, **48**(3), 443–53.

Pertea, G., Huang, X., Liang, F., Antonescu, V., Sultana, R., Karamycheva, S., Lee, Y., White, J., Cheung, F., Parvizi, B., Tsai, J., and Quackenbush, J. (2003). TIGR gene indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics*, **19**(5), 651–652.

Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, **147**, 195–197.

Staden, R., Beal, K. F., and Bonfield, J. K. (1998). *The Staden Package*. The Humana Press Inc., Totowa, NJ 07512.

States, D. J., Gish, W., and Altschul, S. F. (1991). Improved sensitivity of nucleic acid database searches using application-specific scoring matrices. *METHODS: A Companion to Methods in Enzymology*, **3**(1), 66–70.

Walther, D., Bartha, G., and Morris, M. (2001). Basecalling with LifeTrace. *Genome Research*, **11**, 875–888.