BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 21 2010, pages 2780–2781
doi:10.1093/bioinformatics/btq507

Sequence analysis

Advance Access publication September 13, 2010

A unified index of sequence quality and contig overlap for DNA barcoding

Damon P. Little

Cullman Program for Molecular Systematics, The New York Botanical Garden, Bronx, NY, USA

Associate Editor: John Quackenbush

ABSTRACT

Summary: Barcode quality index (*B*) is a novel, unified measure of sequence quality and contig overlap tailored to the needs of DNA barcoding. Re-analysis of published data demonstrates the utility of *B*.

Availability and Implementation: A GPL PERL script is available for download (http://www.nybg.org/files/scientists/dlittle/B.html).

Contact: dlittle@nybg.org

Received on July 10, 2010; revised on August 26, 2010; accepted on August 30, 2010

1 INTRODUCTION

DNA barcoding is an emerging enterprise that aims to build a reference database of high-quality DNA sequences generated from expert-identified vouchered specimens (Hebert *et al.*, 2003). Once populated, the reference database will allow non-specialists to easily identify specimens by sequencing a standard set of markers. Unfortunately, low-quality sequences may obscure subtle differences among specimens resulting in inaccurate identification.

The BARCODE data standard governs the quality and type of sequences archived as references (Hanner, 2009). Quality is assessed through a combination of base caller error probabilities, sequence coverage and contig size. For example, a reference *COI* sequence should contain at least 486 bp of contiguous sequence within the reference region, have 1.6× (bidirectional) coverage, and have error probabilities of 0.01 or smaller assigned to at least 60% of its bases ($q = 20$; $-10 \times \log_{10} p$; Ewing and Green, 1998). For (nearly) fixed-length markers, a threshold for each variable can be effectively employed. In contrast, the application of absolute thresholds—particularly contig size—to variable-length markers results in severe distortion. Some of this distortion can be ameliorated by the use of contig size for normalization, but distortion persists in extremely long or short sequences due to dependance among variables: there is a strong positive correlation between the number (or percent) of low-quality bases and contig size, while at the same there is a strong negative correlation between sequence coverage and contig size.

The criteria for selecting barcode markers must include a comparison of sequence quality among candidate markers. To accomplish this the CBOL Plant Working Group (2009) generated a set of sequences for each marker; a set of trimming and assembly criteria were applied; and the numbers of passing contigs for each marker were compared. Although markers lacking adequate sequence quality were recognized, statistically significant differences in quality among markers could not be identified (CBOL Plant Working Group, 2009).

DNA barcoding needs a distortion-free index that combines measures of sequence quality with contig size and overlap. Such an index must be amenable to statistical analysis and comparable among sequences, markers and editing protocols. This will enable uniform quality evaluation of newly generated sequences—particularly those of variable-length markers.

2 IMPLEMENTATION

The overall quality (S) of a given sequencing read (S_R) can be assessed by tallying the number of positions at, or above, a user defined quality threshold (q) using a Heaviside step function:

$$S_R = \sum_{i=1}^{j} \begin{cases} 0, R_i < q \\ 1, R_i \geq q \end{cases} \quad (1)$$

Where R_i is the quality score for the i-th position of the sequencing read (R) and j is the length of the sequence. Figure 1A illustrates the properties of this Heaviside step function.

Overall contig sequence quality can be assessed by summing the constituent read qualities (S_R). In order to make comparisons among contigs, that use the same quality threshold (q), normalization by contig length and coverage is required:

$$B_q = \frac{\sum_{R=1}^{k} S_R}{cx} \quad (2)$$

Where k is the total number of sequencing reads in the contig, c is the observed contig length and x is the expected coverage. Figure 1B illustrates the sensitivity of B to contig coverage and sequence quality. A PERL script to compute barcode contig quality index (B) from Common Assembly Format (CAF) files is freely available for download under the GNU General Public License.

3 KEY PROPERTIES

The barcode contig quality index equally discounts regions with inadequate coverage and regions of low-quality sequence. In addition, the quality portion of the index is unbiased with respect to base location—interspersed low-quality bases are treated the same as concentrated low-quality bases. This conservative approach assumes that low-quality bases cannot be trusted and therefore should not contribute toward measures of contig overlap. As a result, an identical index value may be given to contigs with different mixtures of overlap and quality (Fig. 1B; e.g. 100% high quality with 60% overlap and 80% high-quality with 100% overlap).

When applied to length-variable markers, specimens that bear sequences of unusual length are not penalized, or rewarded, relative

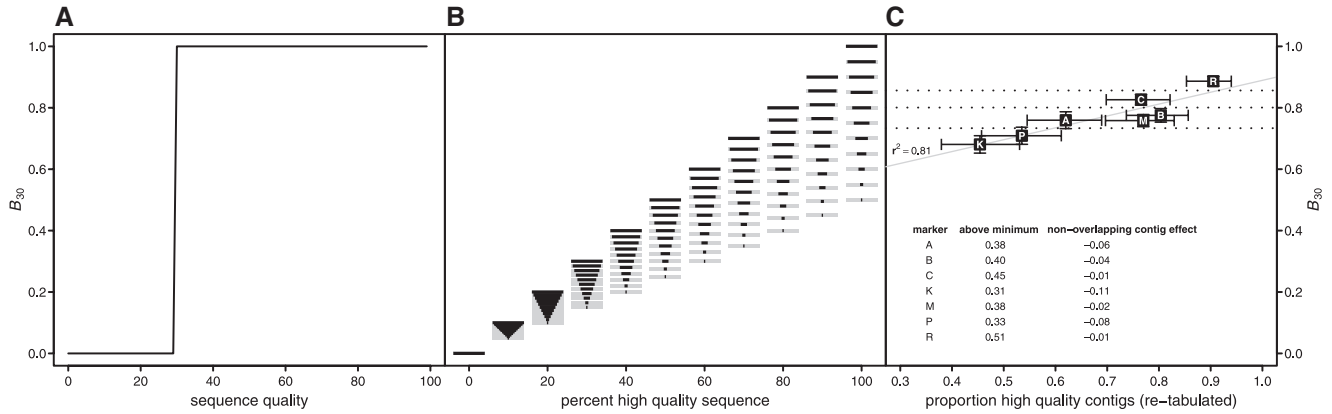

Fig. 1. (**A**) The Heaviside step function [Equation (1)] of B_{30} illustrated using contigs with maximum coverage and uniform quality. (**B**) B_{30} [Equation (2)] as a function of sequence quality and coverage (illustrated using uniform *cx*). Ideograms depict contigs with areas of half (gray) and full coverage (black). (**C**) CBOL Plant Working Group (2009) data: mean B_{30} versus re-tabulated quality determination. Marker names are abbreviated following the original publication. The 95% confidence intervals are indicated. Dashed lines indicate statistically significant groupings ($P = 0.05$). The solid gray line indicates linear correlation.

to those with 'normal' length sequences if observed contig length is used. For markers with little length variability, the user can penalize shorter than expected contigs by replacing observed contig length (*c*) with expected contig length (*C*). When *C* is used on a given set of reads, *B* will be highest following the application of an optimal (for a given *q*) trimming procedure. In contrast, higher *B* values may result from more conservative trimming procedures when *c* is used.

The magnitude of error in *B* is inversely proportional to the amount of error in *cx*. For markers with little length variability, misstatement of *cx* will rarely occur. For length-variable markers, users must accurately calculate *x* (given the sequencing technology's read length) and properly trim contigs so that *c* is not inflated.

An ensemble *B* can be calculated using the mean (or other measure of central tendency) of *B* values calculated for each contig.

4 APPLICATION

To avoid conflating PCR success with sequence quality, the *de novo* sequence traces generated by the CBOL Plant Working Group (2009) were re-tabulated excluding presumed PCR failures (i.e. read failure of both primers). In addition, the data were reanalyzed using B_{30}. The original base calls and quality values (KB 1.2) were extracted with TraceTuner (3.0.6; http://sourceforge.net/projects/tracetuner/) and trimmed and filtered as described in the original publication. Presumed PCR failures were excluded. Contigs were assembled with phrap (0.990329; http://www.phrap.org/). MUSCLE (3.8.31; Edgar, 2004) was used to align the contig to the trimmed reads in order to determine observed contig length (*c*) excluding phrap-induced trimming. Non-overlapping contigs were assumed for single, or paired reads, that could not be assembled into contigs (i.e. *c* = the combined length of the trimmed reads). Statistical differences in sequence quality among markers were examined using the method of Scheffé (1953) at $P = 0.05$. The binomial and Gaussian distributions were used for the re-tabulated and B_{30} analyses, respectively (R 2.11.1; MASS 7.3-6; agricolae 1.0-9; http://cran.r-project.org/).

The most striking differences between the published and re-tabulated analyses are a 5–22% increase in re-tabulated scores and a rank order rearrangement of B, C and M. Given the published trimming and assembly criteria, the minimum B_{30} score for a passing contig in the re-tabulated analysis is 0.375 (assuming a Gaussian quality distribution). Scores above the minimum reflect a greater number of high-quality bases or additional contig overlap. Although the re-tabulated and B_{30} analyses are very similar, they are not perfectly correlated (Fig. 1C). Mean B_{30} for non-coding markers—A, K and P—is 31–38% higher than the minimum passing score whereas coding markers are 38–51% higher. The inclusion of non-overlapping contigs depressed mean B_{30} 1–4% for coding and 6–11% for non-coding markers. Neither the published nor re-tabulated analyses could identify any statistically distinctive groups of markers, whereas the B_{30} analysis was able to identify four distinct groups. The increased statistical resolution of the B_{30} analysis can be attributed to the ability to use more powerful statistical distributions and a reduction in variation within markers.

Thus, *B* provides a more nuanced (resolved) and statistically amenable measure of sequencing success that considers both sequence quality and contig coverage.

Conflict of Interest: none declared.

REFERENCES

CBOL Plant Working Group (2009) A DNA barcode for land plants. *Proc. Natl Acad. Sci. USA*, **106**, 12794–12797.

Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

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

Hanner,R. (2009) Proposed standards for BARCODE records in INSDC (BRIs). *Technical report*, Database Working Group, Consortium for the Barcode of Life, http://barcoding.si.edu/PDF/DWG_data_standards-Final.pdf.

Hebert,P.D.N. *et al.* (2003) Biological identifications through DNA barcodes. *Proc. R. Soc. Lond., B*, **270**, 313–321.

Scheffé,H. (1953) A method for judging all contrasts in the analysis of variance. *Biometrika*, **40**, 87–104.