DNASTAR®
Software for life scientists

Sanger Data Assembly in SeqMan Pro

DNASTAR provides two applications for assembling DNA sequence fragments: SeqMan NGen and SeqMan Pro. SeqMan NGen is primarily used to assemble Next Generation Sequencing (NGS) data, while SeqMan Pro is used to assemble Sanger ABI trace data. This white paper describes how quality scores are calculated and how they are used in SeqMan Pro's Sanger assembly algorithm.

Introduction

In 1999, DNASTAR software developer Carolyn Allex published a doctoral thesis (Allex CF, 1999), in which she compared several algorithmic methods for consensus calling, including the "Majority" method and her own "Trace Evidence" method (Allex CF *et al.* 1997). The latter method was a novel approach for generating quality scores and consensus calling based on geometry and quality of peaks in the trace data.

Allex's analysis indicated that Trace Evidence had significantly better consensus calling accuracy than Majority, even when many of the individual bases had been called incorrectly. Trace Evidence was also more likely than Majority to make the correct call when the base of the well-defined (true) peak was hidden below a high-intensity valley. By contrast, Majority methods often incorrectly called the base that was associated with the valley.

DNASTAR implemented the Trace Evidence algorithm into SeqMan Pro, where it remains the preferred consensus calling method today. The Majority method is now recommended only when data consist of text sequences rather than fluorescence trace data.

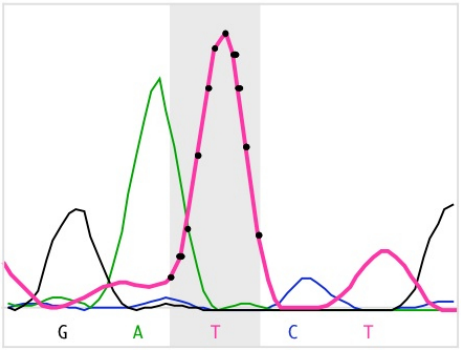
In addition, the quality scoring algorithm that was developed for use with the Trace Evidence method is now also used in SeqMan Pro for SNP calling and quality-based sequence end trimming.

Quality Score Calculations

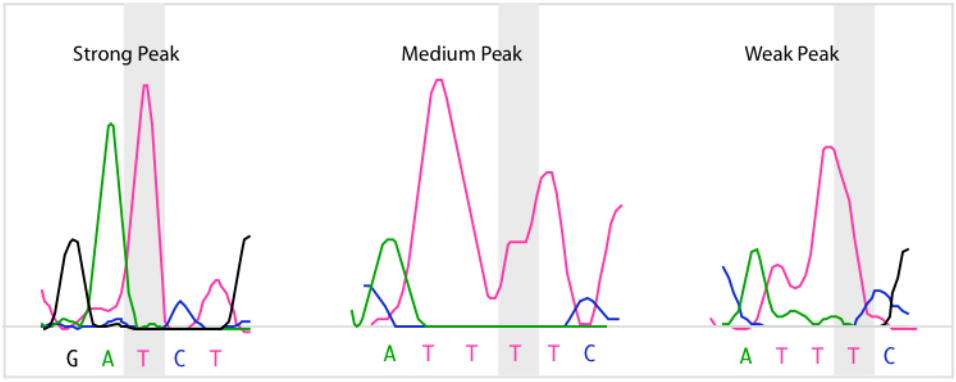
When ABI trace data are used in an assembly, SeqMan Pro analyzes the shape and intensity of peaks to calculate quality scores (Q), and averaged quality scores (Q/n). In quality score calculations:

- Taller, sharper peaks receive higher scores than less distinct peaks. The heights of any underlying peaks are subtracted from the highest peak's score during the calculation.
- The further a peak is from the location at which the base was called, the lower the quality score.

The trace data for a DNA sequence comprises four sets of traces—one each for *A*, *C*, *G*, and *T*. Each trace contains a sequence of intensity values that can be plotted to form a graphical display of trace data. The portions of the four traces associated with a single base call each contain about ten to twelve data points. Only the trace from which the base call is derived is used to calculate a quality score (e.g. if the base call is a *T*, only the *T* trace is analyzed to calculate a quality score). The figure below shows data plotted for five base calls. The data points associated with the center base—a *T*—are marked with black dots.

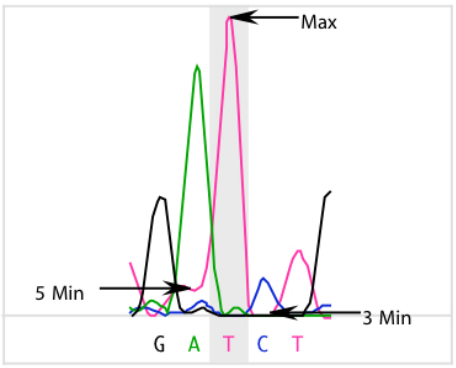


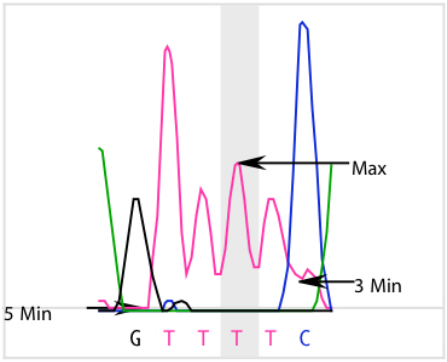
SeqMan Pro calculates each of the peaks in the trace data. A *peak* is defined as trace data that exhibits negative curvature. Slope is used to differentiate between three kinds of peaks: *strong*, *medium*, and *weak*. *Strong* peaks exhibit a change in the sign of the slope, *medium* peaks contain a shoulder with a slope of zero, and *weak* peaks have neither a change in sign nor a shoulder. If the trace data for a base call do not contain a peak, its quality score is zero. The figure below contains examples of the three kinds of peaks for the highlighted *T* base.

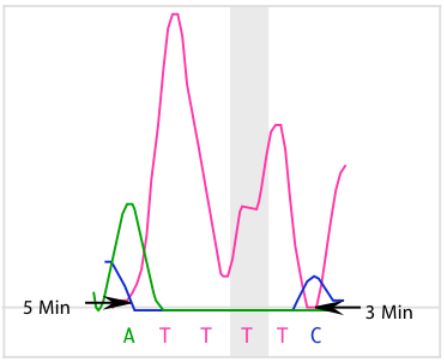


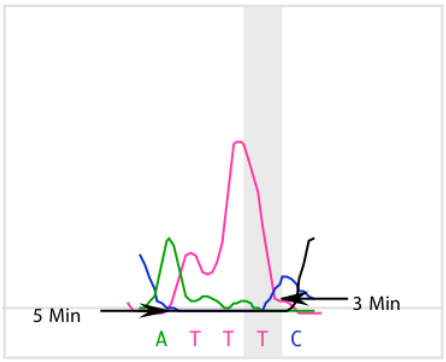
Quality score calculations take into account several parameters:

- Three extreme intensity points: *5Min* (5′ minimum), *3Min* (3′ minimum), and *Max*. *5Min* and *3Min* are the intensity values to either side of the base call that are the minimum values of the data for that base. If a run of identical base calls occurs, then the minimums are taken from either side of the homopolymeric run. *Max* is the intensity value of the peak.

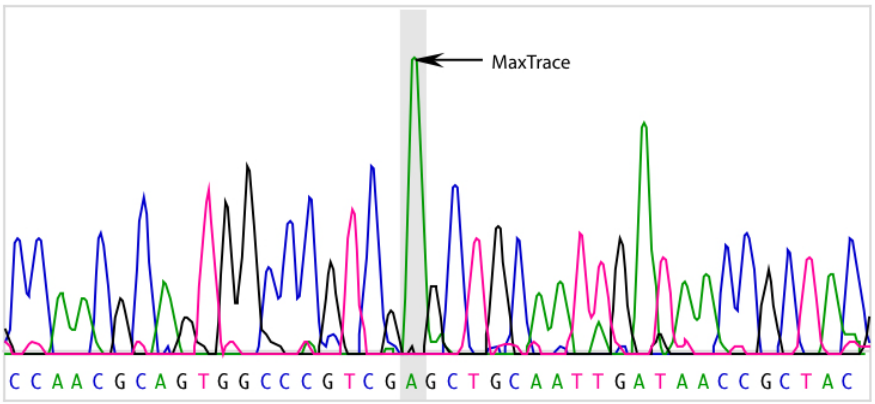




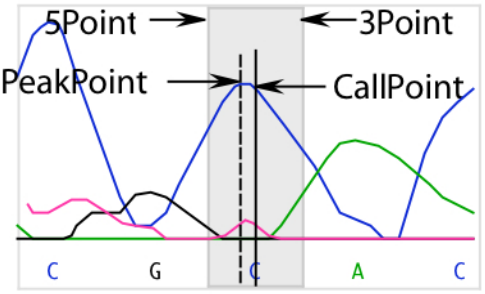




- Each quality score calculation includes division by the maximum intensity of all four traces for an entire sequence. This assigns higher scores to higher peaks. In this example, an *A* peak has the highest intensity value. Its intensity value, *MaxTrace*, is used in the quality score calculation for all bases in the sequence.



- Trace data files identify the point in the trace data where the base was called, or "distance weight." In high quality data, this usually coincides with the point where SeqMan Pro detects a peak. In poorer quality data, the peak can be offset significantly. Each quality score is adjusted to reflect the distance from the detected peak to the point where the base was called.



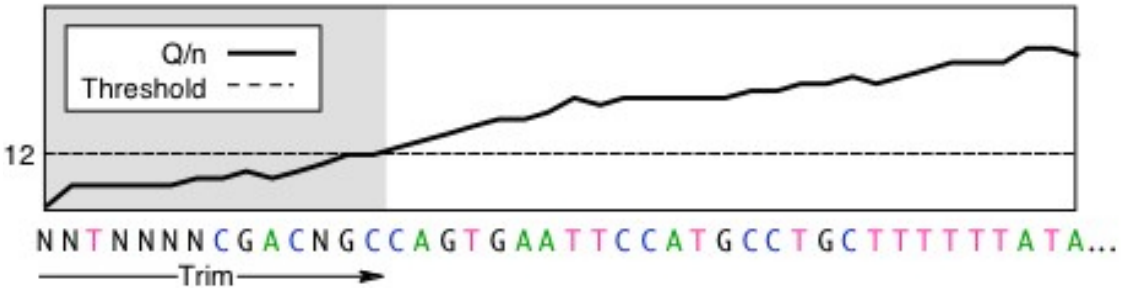
The fraction of the number of points in the offset to the total number of points is the *Dist* weight used in the quality score calculation. It is calculated as follows:

$$Dist = \frac{\text{abs}(PeakPoint - CallPoint) + 1}{(3Point - 5Point) + 1}$$

End-Trimming Based on Averaged Quality Scores

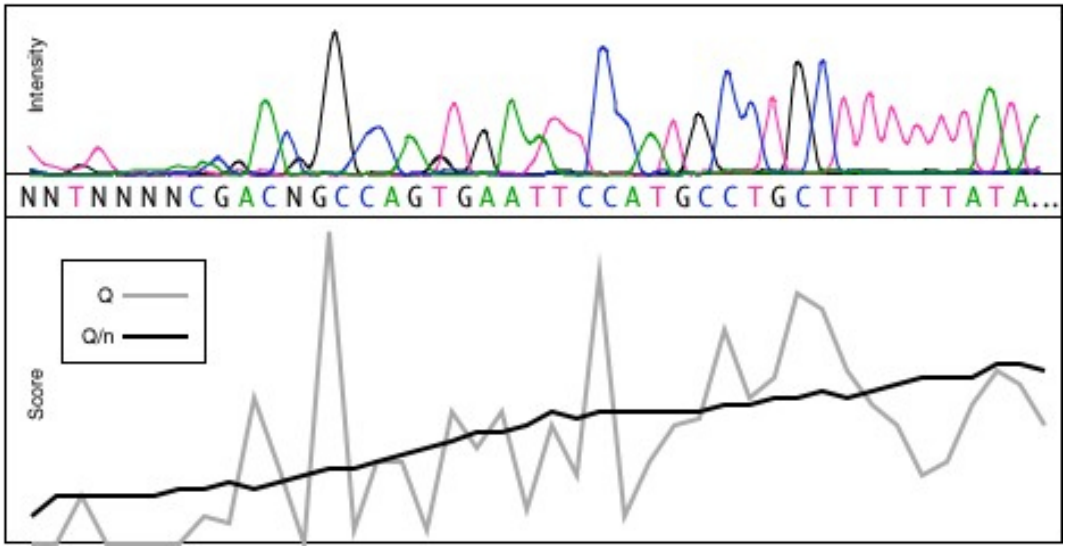
SeqMan Pro uses averaged quality scores (*Q/n*) to identify regions of poor quality data at the end of sequences. Averaged quality scores are calculated as the average of the quality scores, *Q*, over a window of 21 bases. The average score is assigned to the base in the center of the window. Averaging the scores smooths out the quality scores and quantifies the general quality of data in a region. To perform quality end-trimming, a threshold is set and the longest sequence of bases with all *Q/n* meeting the threshold is identified. Below-threshold ends to either side of the high-quality region of the sequence are trimmed off before assembly.

In the example below, the quality scores, *Q*, and averaged quality scores, *Q/n*, are graphed for the 5′ end of a 794 base pair sequence. A dashed horizontal line marks the quality end-trimming threshold. The average scores, *Q/n*, are compared to the threshold and the first 14 bases are trimmed from the 5′ end of this sequence.



Poor quality data on the ends of sequences often contain miscalled bases that produce mismatches in alignments with other sequences. If the number of mismatches is high enough that SeqMan Pro's Minimum Match Percentage threshold is not met, sequences will not be assembled in the same contig. Trimming the poor quality data from the ends of sequences allows better and more complete assembly.

The image below compares *Q* to *Q/n* scores for the 5′ end of a sequence.



SNP Calling Using Neighborhood Quality Scores

SeqMan Pro provides the option to use a neighborhood quality score threshold when identifying SNPs. This threshold can be changed by opening SeqMan Pro's SNP Discovery Parameters dialog and defining a new 'Neighborhood Window' value. By default, the Neighborhood Window value is zero, meaning that the neighborhood quality score threshold is not used.

A neighborhood quality score is equal to the lowest quality score of any of the bases in the defined window surrounding a SNP base. The size of the window can be adjusted by editing the Neighborhood Window value. For example, if the Neighborhood Window value is set to 5, then the 5 bases upstream and the 5 bases downstream from the SNP base will be considered. If the specified window contains one or more mismatches to the reference sequence, the putative SNP will be rejected.

Unless you have specific thresholds you would like to use, you may wish to start with the Q-Score Threshold values from Altshuler *et al.* (2000):

- Minimum Score at SNP: 20
- Minimum Neighborhood Score: 15
- Neighborhood Window: 5

Benchmark Testing

In October 2016, we used SeqMan Pro version 14.0 to perform a benchmark test comparing errors resulting from the Majority and Trace Evidence methods of consensus calling.

SeqMan Pro was used to automatically remove Janus vector and then *de novo* assemble 498 .abi trace files from *E. coli*: once using SeqMan Pro's "Pro" assembler and once using its heritage algorithm, the "Classic" assembler. Both algorithms assembled the data in a few seconds. For each assembler type, the consensus was calculated twice, once using the Majority method, and once using Trace Evidence. The four resulting consensus sequences were then exported and saved.

Next, SeqMan Pro was used to assemble all four consensus sequences against a 39,774 bp reference genome fragment from *E. coli* K12 MG1655 (Blattner FR *et al.*,1997). A SNP (variant) report was generated automatically. Since the reference genome should theoretically be an exact match to all four consensus sequences, any observed differences were presumed to be errors, rather than true SNPs. Table 1 lists the number of errors found in each of the four consensus sequences.

Table 1: Number of consensus calling errors using combinations of two assembly methods and two consensus calling algorithms in SeqMan Pro

SeqMan Pro Algorithm	Number of Errors	
	Majority Method	Trace Evidence Method
Classic	174	21
Pro	153	11

These data show that for both the Pro and Classic assemblers, the Trace Evidence method produced far fewer consensus calling errors than the Majority method, corroborating the findings of Allex CF (1999).

Resources and Free Trial Software

To try SeqMan Pro and see the results of its quality-score based algorithms, download and install a free trial of Lasergene 14. When the Navigator opens, click on its SeqMan Pro bar to launch the fully-functional application.

To learn more about the software and how it can help with your research goals, please refer to the resource list below:

- Comprehensive, easy-to-understand SeqMan Pro help and written tutorials.
- Almost 40 short training videos on SeqMan Pro-related topics.
- Friendly, helpful support from fellow scientists via support@dnastar.com or by calling one of the numbers in the footer of this document.

References

1) Allex, C.F., Baldwin, S.F., Shavlik, J.W., and Blattner, F.R. (1996). Improving the quality of automatic DNA sequence assembly using fluorescent trace-data classifications. *Proceedings, Fourth International Conference on Intelligent Systems for Molecular Biology*, 3-14. St. Louis, MO. Menlo Park, CA. *ISMB-96 Proceedings*, AAAI Press.

2) Allex, C.F., Baldwin, S.F., Shavlik, J.W., and Blattner, F.R. (1997). Increasing consensus accuracy in DNA fragment assemblies by incorporating fluorescent trace representations. *Proceedings, Fifth International Conference on Intelligent Systems for Molecular Biology*, 3-14. Halkidiki, Greece. Menlo Park, CA. *ISMB-97 Proceedings*, AAAI Press.

3) Allex, C.F., Shavlik, J.W., and Blattner, F.R. (1999). Neural network input representations that produce accurate consensus sequences from DNA fragment assemblies. *BioInformatics* 15(9):723-728.

4) Allex CF (1999). Computational Methods for Fast and Accurate DNA Fragment Assembly (Doctoral thesis). Department of Computer Sciences, University of Wisconsin-Madison.

5) Altshuler *et al.* (2000). A SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* 407, 513-516.

6) Blattner FR *et al.* (1997). The Complete Genome Sequence of Escherichia coli K-12. *Science* 05 Sep 1997: Vol. 277, Issue 5331, pp. 1453-1462. DOI: 10.1126/science.277.5331.1453.