Improved Consensus Accuracy in DNA Fragment Assemblies

Carolyn F. Allex[1,2]
allex@cs.wisc.edu

Schuyler F. Baldwin[2]
schuy@dnastar.com

Jude W. Shavlik[1]
shavlik@cs.wisc.edu

Frederick R. Blattner[2,3,4]
fred@genetics.wisc.edu

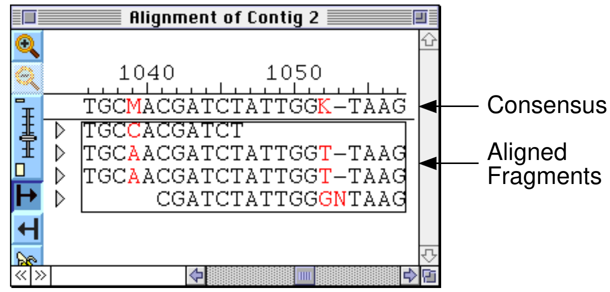
[1]University of Wisconsin Computer Sciences Department
[2]DNAStar Incorporated http://www.dnastar.com/
[3]University of Wisconsin Genetics Department
[4]University of Wisconsin *E. coli* Genome Project

We present a new method for determining the consensus sequence in DNA fragment assemblies that directly examines aligned fluorescent trace-data. The new method, Trace-Evidence Consensus, results in greater accuracy and fewer ambiguities with less coverage required.

Consensus Calling

The consensus calling problem is to determine the consensus for each column of aligned fragments. In an aligned column, there may be total agreement of base calls or some calls may conflict with the others. A decision must be made to call the consensus as a base (*A*, *C*, *G*, *T*), as a gap (indicating an insertion in one of the sequences), or as one of 11 ambiguities (combinations of *A*, *C*, *G*, and *T*).

Given: Aligned fragments.
Do: Determine the consensus sequence.

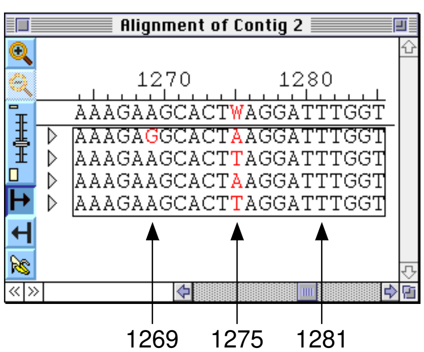


Possible Consensus Calls

A	*adenine*	**R**	*G or A*	**H**	*not G*
C	*cytosine*	**Y**	*T or C*	**B**	*not A*
G	*guanine*	**M**	*A or C*	**V**	*not T*
T	*thymine*	**S**	*C or G*	**D**	*not C*
		K	*G or T*		
X	*unknown*	**W**	*T or A*	**–**	*gap*

Previous Method: *Majority*

A simple method for determining the consensus chooses the majority base call in a column or an ambiguous combination if the majority is below a given threshold.

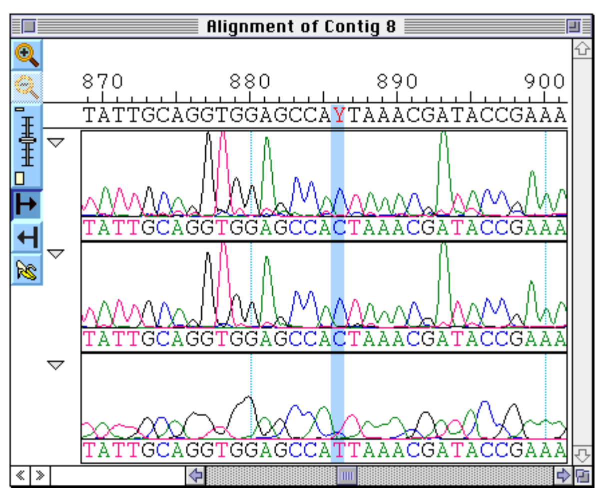


In this example, the threshold is 75%. Three consensus calls made with the *Majority* method are described below.

Column	Majority Base	%	Consensus Call
1269	A	75	A
1275	A, T	50	W (A or T)
1281	T	100	T

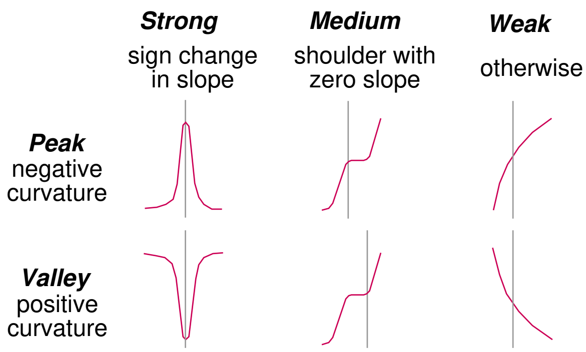
Motivation for New Method

Human editors examine fluorescent trace data to resolve ambiguities. In the example below, the ambiguity in the highlighted column must be resolved to *A*, *C*, *G*, or *T*. Editors examine the traces and observe that the the first two sequences are of good quality and exhibit sharp peaks in the *C* trace. The third sequence, although of poorer quality, also shows a peak in the *C* trace. Human editors determine that the consensus should be a *C*.

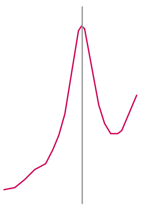


We improve the quality of automatic consensus calling by emulating human examination of trace data.

The new *Trace-Evidence* method we have developed directly incorporates trace-data information via *Trace-Data Classifications* (Allex et al. 1996). A *Trace-Data Classification* is a set of six scores between 0 and 100 that capture visual characteristics of the trace data associated with a single base call. The scores reflect the amount of *strong*, *medium*, and *weak peak* and/or *valley* shape that is exhibited by the trace.



Example



Peak		Valley	
Strong	57	*Strong*	19
Medium	43	*Medium*	7
Weak	0	*Weak*	0

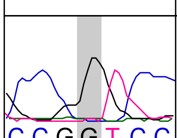
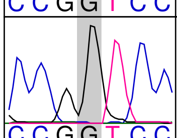
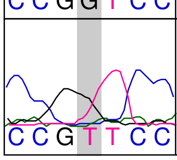
The *Trace-Evidence* method sums the evidence for each base that is reflected by the *Trace-Data Classification* scores. The method is based on the idea that each of the scores for the six classes supplies an amount of evidence that the base associated with the classified trace data should be assigned in the consensus. High *strong peak* (*SP*) scores supply the greatest amount of evidence, high *medium peak* (*MP*) scores supply the next greatest amount of evidence, followed by high *weak peak* (*WP*), *weak*, *medium*, and *strong valley* (*WV*, *MV*, and *SV*) scores in decreasing order. For example, if a *G* trace has a high *SP* score and the *A*, *C*, and *T* traces have high valley scores, there is confidence that the base should be called as a *G*.

For the *Trace-Evidence* method, we sum the evidence for each of the four bases and for gaps in an aligned column. The evidence for each base is weighted by the quality of the trace data in a local area. The quality is also determined by examining *Trace-Data Classification* scores. A threshold is specified that determines the allowed fraction of competing evidence. If the base with the highest evidence has no competing evidence greater than the threshold, the base with the highest evidence is assigned to the consensus. If competing evidence for one or more bases surpasses the threshold, then those bases are included in determining an ambiguous call.

Trace-Evidence

The consensus base for the center column of three aligned sequences must be called. For each sequence, the evidence for each base is multiplied by the corresponding quality weight. When these products are summed for the three sequences, the evidence for *A* and *C* is 0, for *G* is 91.9 and for *T* is 1.3. If the threshold is .50, *G* will be called unambiguously since no competing evidence surpasses 45.9 (91.9 x .50). In contrast, a majority method with a 75% threshold would make an ambiguous call of *K* (*T* or *G*).

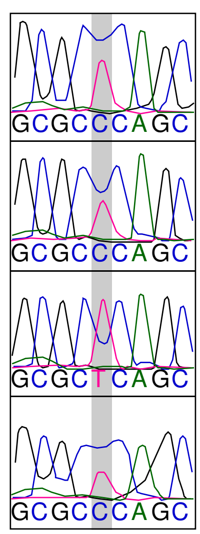
Consensus Call: G

Sequence	Quality	Evidence *A*	*C*	*G*	*T*	QualityxEvidence *A*	*C*	*G*	*T*
	.51	0	0	52	0	0	0	26.5	0
	.82	0	0	73	0	0	0	59.9	0
	.26	0	0	21	5	0	0	5.5	1.3
					Total	0	0	91.9	1.3

Majority vs. *Trace-Evidence*

We observe striking examples of the utility of the *Trace-Evidence* method when base calls in a column are systematically incorrect. In some instances, a well-defined peak is “hidden” below a high-intensity valley. In most or all cases, the base is incorrectly called as the one associated with the high-intensity valley. *Majority* methods incorrectly call the consensus as this base. Our new *Trace-Evidence* Consensus makes the correct consensus call even when all or most of the bases have been called incorrectly.

Majority: C
Trace-Evidence: T



In the center column, three bases have been incorrectly called as a *C* and one correctly as a *T*. With a 75% threshold, the *Majority* method incorrectly computes the consensus as a *C*. The *Trace-Evidence* method detects no evidence for a *C*, ample evidence for a *T*, and calls the correct consensus. With *Majority* this situation would be even more troublesome if the third sequence were not in the assembly. In that case, the call would have no conflicting base calls and would likely go unquestioned during hand-editing. In contrast, *Trace-Evidence* correctly computes a *T*, even in the absence of the third sequence.

Weight Vector

Define a weight vector, W, such that classes (such as *SP*) that imply better-quality data for a base are given higher values than those (such as *SV*) that imply lower-quality data:

$$W = |W_{SP} \;\; W_{MP} \;\; W_{WV} \;\; W_{SV} \;\; W_{MV} \;\; W_{SV}|$$

where

$$1 >= W_{SP} >= W_{MP} >= W_{WP} >= W_{WV} >= W_{MV} >= W_{SV} >= 0$$

We used:

$$W = |\,1 \;\; .66 \;\; .33 \;\; 0 \;\; 0 \;\; 0\,|$$

Quality

Each base in a fragment is assigned a value that reflects the quality of the local trace data surrounding the base. This value gives more weight to better-quality data.

1. For each base, i, in a window of size n, extract the vector of *Trace-Data Classification* scores, S_i, for the trace associated with the base that has been called.

   $$S_i = |SP_i \;\; MP_i \;\; WP_i \;\; WV_i \;\; MV_i \;\; SV_i|$$

2. The dot product of S_i and W produces a quality measure, Q_i, for base i:

   $$Q_i = S_i \bullet W$$

3. Average the measures to produce an overall quality score, Q, for the base at the center of the window:

   $$Q = (Q_1 + Q_2 + \ldots + Q_n) \,/\, n$$

Evidence

For each column, a value, E, for each possible base (*A*, *C*, *G*, and *T*) is computed that reflects the amount of evidence that the call should be that possible base.

1. Form a matrix of *trace-data classification* scores, S, by extracting the scores for each trace of a called base:

   $$S = \begin{vmatrix} SP_A & MP_A & WP_A & WV_A & MV_A & SV_A \\ SP_C & MP_C & WP_C & WV_C & MV_C & SV_C \\ SP_G & MP_G & WP_G & WV_G & MV_G & SV_G \\ SP_T & MP_T & WP_T & WV_T & MV_T & SV_T \end{vmatrix}$$

2. The matrix multiplication of S and W^t produces a vector of evidence values, E, for the possible bases:

   $$E = S \times W^t = |\,E_A \;\; E_C \;\; E_G \;\; E_T\,|^t$$

Trace-Evidence Sum

Find the sum of the evidence for each base in an aligned column so the consensus can be computed. (Gap sums are omitted for clarity.)

1. For each sequence, i, and each possible base, j, in a column of n fragments, multiply each E_{ij} by Q_i to produce values, E_{ij}', adjusted by quality of the data.

2. Sum corresponding E' values to produce the total evidence, T_j, for each possible base:

$$\begin{aligned} T_A &= E_{A1}' + E_{A2}' + \ldots + E_{An}' \\ T_C &= E_{C1}' + E_{C2}' + \ldots + E_{Cn}' \\ T_G &= E_{G1}' + E_{G2}' + \ldots + E_{Gn}' \\ T_T &= E_{T1}' + E_{T2}' + \ldots + E_{Tn}' \end{aligned}$$

Fragment assemblies for a 124 kb section of *E. coli* are used to compare *Majority* and *Trace-Evidence* consensus calls. Results shown for coverage (number of aligned fragments) from 2 to 10+ are each based on 20,000 - 68,000 consensus calls. Using *Trace-Evidence* with a coverage of only three, we see a leveling of the number of incorrect calls and a improvement over the *Majority* method in the number of correct and ambiguous calls.

Almost all of the incorrect calls for the *Trace-Evidence* method involve gaps in the column. A solution to calling the consensus when gaps are in the alignment would virtually eliminate incorrect calls with the *Trace-Evidence* method.