Improved Consensus Accuracy in DNA Fragment Assemblies

Carolyn F. Allex¹,²
alex@cs.wisc.edu

Schuyler F. Baldwin²
schuy@dnastar.com

Jude W. Shavlik¹
shavlik@cs.wisc.edu

Frederick R. Blattner²,³,⁴
fred@genetics.wisc.edu

¹University of Wisconsin Computer Sciences Department
²DNAStar Incorporated
³University of Wisconsin Genetics Department
⁴University 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).

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.

<table>
<thead>
<tr>
<th>Column</th>
<th>Majority Base</th>
<th>%</th>
<th>Consensus Call</th>
</tr>
</thead>
<tbody>
<tr>
<td>1269</td>
<td>A</td>
<td>75</td>
<td>A</td>
</tr>
<tr>
<td>1275</td>
<td>A, T</td>
<td>50</td>
<td>W (A or T)</td>
</tr>
<tr>
<td>1281</td>
<td>T</td>
<td>100</td>
<td>T</td>
</tr>
</tbody>
</table>

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

Possible Consensus Calls

- A adenine
- C cytosine
- G guanine
- T thymine
- X unknown

- R G or A
- Y T or C
- M A or C
- S G or T
- V not T
- D not C
- H not G
- W T or A
- – gap

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 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.

<table>
<thead>
<tr>
<th>Peak</th>
<th>Quality</th>
<th>Valley</th>
<th>Quality</th>
<th>Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Strong</td>
<td></td>
<td>Strong</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td></td>
<td>Medium</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Weak</td>
<td></td>
<td>Weak</td>
<td></td>
</tr>
</tbody>
</table>

**Example**

**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 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.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Quality</th>
<th>Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>.51</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>.82</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>.26</td>
<td>0</td>
</tr>
</tbody>
</table>

**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.
**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

\[
1 \geq W_{SP} \geq W_{MP} \geq W_{WP} \geq W_{WV} \geq W_{MV} \geq W_{SV} \geq 0
\]

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.

\[
S_i = |S_{P_i} \; M_{P_i} \; W_{P_i} \; W_{V_i} \; M_{V_i} \; S_{V_i}|
\]

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

\[
Q_i = S_i \cdot W
\]

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:

\[
S = \begin{bmatrix}
S_{P_A} & M_{P_A} & W_{P_A} & W_{V_A} & M_{V_A} & S_{V_A} \\
S_{P_C} & M_{P_C} & W_{P_C} & W_{V_C} & M_{V_C} & S_{V_C} \\
S_{P_G} & M_{P_G} & W_{P_G} & W_{V_G} & M_{V_G} & S_{V_G} \\
S_{P_T} & M_{P_T} & W_{P_T} & W_{V_T} & M_{V_T} & S_{V_T}
\end{bmatrix}
\]

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

\[
E = S \times W = |E_A \; E_C \; E_G \; E_T|
\]

**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 \( E_{ij}' \) values to produce the total evidence, \( T_{ij} \), for each possible base:

\[
T_A = E_{A1}' + E_{A2}' + \ldots + E_{An}'
\]

\[
T_C = E_{C1}' + E_{C2}' + \ldots + E_{Cn}'
\]

\[
T_G = E_{G1}' + E_{G2}' + \ldots + E_{Gn}'
\]

\[
T_T = E_{T1}' + E_{T2}' + \ldots + E_{Tn}'
\]

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.