

CLAIMS

What is claimed is:

1. A method of automatically sequencing one or more deoxyribonucleic acid (DNA) molecules of a biological sample, comprising:
 - a. using a capillary electrophoresis (CE) genetic analyzer to measure the biological sample to obtain an input trace comprising digital data corresponding to fluorescent values comprising a plurality of scans of the biological sample;
 - b. using a trained artificial neural network comprising a plurality of layers including convolutional layers to generate scan labelling probabilities for the plurality of scans; and
 - c. determining a basecall sequence comprising a plurality of basecalls for the one or more DNA molecules based on the scan labelling probabilities for each of the plurality of scans.
2. The method of claim 1, further comprising:
 - a. determining a scan number position for each of the plurality of basecalls;
 - b. displaying, on an electronic display, the basecall sequence; and
 - c. using the scan number position to display, on the electronic display, a basecall position indication for each of the plurality of basecalls that

visually indicates a relative spacing between adjacent basecalls in the basecall sequence.

3. The method of claim 2, further comprising displaying, on the electronic display, the input trace such that an axis of the input trace corresponding to relative scan number positions of fluorescent values of the input trace is aligned with an axis for displaying the basecall position indications corresponding to relative scan number positions.
4. The method of claim 3 wherein the axis of the input trace is the same axis as the axis for displaying the basecall position indications.
5. The method of claim 1, wherein the plurality of layers comprise a plurality of residual blocks, wherein each residual block of the plurality of residual blocks comprises one or more non-causal convolutional layers.
6. The method of claim 5, wherein a residual block of the plurality of residual blocks further comprises a skip connection.
7. The method of claim 5, wherein a residual block of the plurality of residual blocks further comprises at least one spatial dropout layer following a non-causal convolution layer.

8. The method of claim 6, wherein a residual block of the plurality of residual blocks further comprises a 1x1 convolutional layer between an input and an output of the skip connection.
9. The method of claim 5, wherein a residual block of the plurality of residual blocks further comprises at least one normalization layer following a non-causal convolution layer.
10. The method of claim 5, wherein a residual block of the plurality of residual blocks further comprises at least one rectified linear activation function layer following a non-causal convolution layer.
11. The method of claim 5, wherein the plurality of residual blocks comprises at least a first residual block comprising one or more non-causal convolutional layers having a first dilation factor, and a second residual block comprising one or more non-causal convolutional layers having a second dilation factor different than the first dilation factor.
12. The method of claim 1, wherein the trained artificial neural network has been trained using a Connectionist Temporal Classification (CTC) loss function to minimize a loss between the scan labelling probabilities and a target sequence of bases.

13. The method of claim 1, wherein the trained artificial neural network further comprises a 1x1 convolutional reduction layer to reduce a number of extracted features to match a number of output labels.
14. The method of claim 1, wherein the trained artificial neural network further comprises a softmax layer to obtain the scan labelling probabilities.
15. The method of claim 1, wherein determining the basecall sequence further comprises decoding the scan labelling probabilities using a prefix beam search.
16. The method of claim 15, wherein decoding the scan labelling probabilities using the prefix beam search comprises:
 - a. initializing an empty basecall sequence as a prefix;
 - b. at each scan t of the plurality of scans,
 - i. extending the prefixes with each of a plurality of extended labels;
 - ii. scoring each prefix by incorporating a scan labelling probability of the extended label at the scan t ;
 - iii. saving an extended candidate subset comprising the K highest scoring prefixes wherein the subset does not exceed a beam width of size K ;

- iv. saving the highest scoring prefix at the scan in a candidate subset if the prefix is different from the highest scoring prefix at the previous scan $t-1$;
 - v. assigning the scan to the highest scoring prefix if the prefix is different from the highest scoring prefix at the previous scan $t-1$;
 - vi. returning the highest scoring prefix at the last scan as the final basecall sequence; and
 - vii. returning the candidate subset of the top candidates saved at each scan during the prefix beam search.
17. The method of claim 16, wherein the plurality of extended labels comprises pure base labels, mixed base labels, and a blank label.
18. The method of claim 16, further comprising finding a scan range for each basecall and then using the scan range to find a scan position having a peak labelling probability within the scan range.
19. The method of claim 18, wherein finding the scan range and the scan position with the peak labelling probability within the scan range for each basecall comprises:
- a. starting with a first basecall in a final basecall sequence y ;

- b. at each basecall y^i of the plurality of basecalls in the final basecall sequence y ,
 - i. searching the basecall sub-sequence $y^{l..i}$, with the first i basecalls in the basecall sequence y in the candidate subset;
 - ii. setting a begin scan of the scan range for the basecall y^i as the scan assigned to the found candidate;
 - iii. setting an end scan of the scan range for the basecall y^i by extending the begin scan with the prefixed scan number until the start scan of a next basecall y^{i+1} ; and
 - iv. selecting a scan position for the basecall y^i , between the begin scan and the end scan with the peak labelling probability; and
 - c. returning the begin and end scans and the scan positions for all basecalls in the final basecall sequence.
20. The method of claim 1, further comprising determining a quality value for each of the plurality of basecalls by, for a basecall of the plurality of basecalls, using a plurality of feature values derived from scan labelling probabilities corresponding to scans in scan range that includes a scan position of the basecall, the plurality of feature values comprising a peak scan labelling probability of the basecall, a noise-to-signal ratio, a basecall spacing ratio, and a resolution value.

21. The method of claim 20, further comprising using a machine learning algorithm to obtain the quality value using the plurality of feature values.
22. The method of claim 1, further comprising determining a quality value for each basecall, wherein predicting the quality value comprises:
 - a. determining a feature vector for the basecall, the feature vector comprising a plurality of feature values including: a scan labelling probability of the basecall at a basecall scan position, a noise-to-signal ratio, a basecall spacing ratio, and a resolution value;
 - b. finding a line having a smallest cut index in a quality value lookup table comprising a plurality of lines wherein each line has (1) a feature vector assigned to a cut comprising a plurality of basecalls, and (2) a quality value corresponding to an empirical error rate of the cut; and
 - c. traversing the quality value lookup table in order to assign a quality value corresponding to the line having the smallest cut index to the basecall, where the line having the smallest cut index comprises the line having a feature vector having all feature values greater than or equal to the feature vector for the basecall, or assigning a quality value of zero if no line having the smallest cut index is found.
23. The method of claim 22, wherein the quality value lookup table is constructed by:

- a. initializing a quality value lookup table;
 - b. computing a feature vector for each basecall in a quality value training dataset comprising a plurality of samples;
 - c. until all remaining cuts are added to the lookup table, grouping the basecalls into a plurality of cuts wherein each cut equalizes a histogram for the feature vector;
 - d. computing an empirical error rate for each of the one or more cuts;
 - e. adding a cut having the lowest empirical error rate to the quality value lookup table as a next new line comprising a feature vector assigned to the cut and a quality value corresponding to the empirical error rate of the cut;
 - f. removing the cut added to the quality value lookup table from the plurality of cuts;
 - g. removing all basecalls in the cut added to the quality value lookup table from the remaining cuts; and
 - h. repeating steps (c) through (g) until there are no more cuts remaining.
24. The method of any of claims 20 and 22, wherein the noise-to-signal ratio comprises a ratio of (1) a maximum scan labelling probability from one or more uncalled bases or noise scan labelling probabilities within the local scan window for the basecall, to (2) the scan labelling probability of the called base at a scan position for the basecall.

25. The method of any of claims 20 and 22 wherein the basecall spacing ratio comprises a ratio of a first base spacing value between the basecall and a first neighboring basecall and a second base spacing value between the basecall and a second neighboring basecall.
26. The method of any of claims 20 and 22, wherein the resolution value comprises a ratio of a local base spacing value to a width value of a scan labelling probability peak for the basecall.
27. The method of claim 1, further comprising displaying the basecall sequence and the input analyzed trace in an electropherogram on a computing device display.
28. The method of any of claims 2-4 further comprising displaying, on the electronic display, a visual indication of a quality value for each of the plurality of basecalls.
29. The method of any of claims 2-4 further comprising displaying a visual indication of a quality value for each of the plurality of basecalls wherein the visual indication of the quality value is also used as the basecall position indication by placing the visual indication of the quality value at a location on the electronic display that represents the basecall position.

30. A non-transitory computer readable medium comprising a memory storing one or more instructions which, when executed by a one or more processors of at least one computing device, perform automatically sequencing one or more deoxyribonucleic acid (DNA) molecules of a biological sample by:
- a. obtaining an input trace comprising digital data corresponding to fluorescent values in a plurality of scans of the biological sample conducted by a capillary electrophoresis genetic analyzer;
 - b. using a trained artificial neural network comprising a plurality of layers including convolutional layers to generate scan labelling probabilities for the plurality of scans; and
 - c. determining a basecall sequence comprising a plurality of basecalls for the one or more DNA molecules based on the scan labelling probabilities for each of the plurality of scans.
31. A method of determining scan positions of basecalls in a basecall sequence obtained using scan labelling probabilities for each of a plurality of scans conducted by a capillary electrophoresis instrument on one or more deoxyribonucleic acid (DNA) molecules of a biological sample, comprising:
- a. determining a scan range of the scan range probabilities for each basecall, the scan range comprising scans from a range starting from a

- first scan corresponding to the basecall up to and including a last scan corresponding to the basecall;
- b. searching within the scan range to determine a peak scan labelling probability in the scan range; and
 - c. using a scan position of the peak scan labelling probability to display, on an electronic display, an indication of the basecall position.
32. The method of claim 31, wherein determining the scan range and the scan position with the peak scan labelling probability within the scan range for each basecall comprises:
- a. starting with a first basecall in a final basecall sequence y ;
 - b. at each basecall y^i of the plurality of basecalls in the final basecall sequence y ,
 - i. searching the basecall sub-sequence $y^{1..i}$, with the first i basecalls in the basecall sequence y in the candidate subset;
 - ii. setting a begin scan of the scan range for the basecall y^i as the scan assigned to the found candidate;
 - iii. setting an end scan of the scan range for the basecall y^i by extending the begin scan with the prefixed scan number until the start scan of a next basecall y^{i+1} ; and

- iv. selecting a scan position for the basecall y^i , between the begin scan and the end scan with the peak scan labelling probability;
and
- c. returning the begin and end scans and the scan positions for all basecalls in the final basecall sequence.