

INSERTION_PENALTY: -40

OPENING_PENALTY: -10

DELETION_PENALTY: -5

The relationship between the gap penalties (INSERTION_PENALTY, OPENING_PENALTY) above help limit the number of gap openings, i.e., favor grouping gaps together, by setting the gap insertion penalty higher than the gap opening cost. Of course, alternative relationships between MISMATCH_PENALTY, MATCH_BONUS, INSERTION_PENALTY, OPENING_PENALTY and DELETION_PENALTY are possible.

Once the alignment is complete, the aligned sequences can be assembled to produce a sequence that can be compared to a reference (i.e., a genetic standard) to identify variants. The variants can provide insight regarding diseases, stages of disease, recurrence and the like. In the case of amino acid alignments, the assembled amino acid sequences can be compared to a standard to determine evolutionary information about the protein, or functional information about the protein. This standard method of disease comparison is time consuming, however, because many of the variants are not necessarily correlated with a disease. For example, when the genetic standard is from a population having an ancestry different from the sample, many of the called variants are due to differences in things like hair color, skin color, etc.

SUMMARY

The disclosure describes algorithms and methods for their implementation that transform linear, local sequence alignment processes such as, for example, Smith-Waterman-Gotoh, into multi-dimensional alignment algorithms that provide increased parallelization, increased speed, increased accuracy, and the ability to align reads through an entire genome. Algorithms disclosed herein provide for a “look-back” type analysis of sequence information (as in Smith-Waterman), however, in contrast to known linear methods, the look back disclosed herein is conducted through a multi-dimensional space that includes multiple pathways and multiple nodes in order to provide more accurate alignment of complex and lengthy sequence reads, while achieving lower overall rates of mismatches, deletions, and insertions. The present invention refers to subject matter as defined in the attached set of claims. In one example, there is disclosed a system for aligning a plurality of sequence reads comprising a processor and non-transitory memory, wherein the memory comprises instructions that, when executed, cause the processor to obtain a plurality of

sequence reads as strings of symbols; compare each string of symbols corresponding to a sequence read to a plurality of positions in a reference sequence construct, wherein the construct comprises at least two different strings of symbols at multiple positions in the construct, score overlaps between each string of symbols corresponding to a sequence read and each of the plurality of positions in the reference sequence construct, wherein a higher score corresponds to a greater amount of overlap; identify the overlap corresponding to the highest score for each sequence read; assign each sequence read to a location on the construct corresponding to the highest score; and write a file to memory corresponding to the location of each aligned sequence read. In another example, there is disclosed a method of aligning a plurality of sequence reads, comprising obtaining a plurality of sequence reads; scoring sequence overlaps for each sequence read against a reference sequence construct, the construct comprising at least two alternative sequences per position at multiple positions in the construct, wherein greater overlap results in a higher score; and aligning each sequence read to a location in the construct such that the score for each sequence read is maximized. In yet another example, there is disclosed a method of aligning a plurality of nucleic acid reads, comprising obtaining a plurality of nucleic acid reads; aligning the nucleic acid reads against a reference sequence construct using an algorithm that scores overlaps between a base in the read and the construct, the construct comprising at least two alternative sequences per position at multiple positions in the construct. In a further example, there is disclosed a method of aligning a plurality of amino acid reads, comprising obtaining a plurality of amino acid reads aligning the amino acid reads against a reference sequence construct using an algorithm that scores overlaps between an amino acid in the read and the construct, the construct comprising at least two alternative sequences per position at multiple positions in the construct. In yet another example, there is disclosed a computer-implemented method of aligning a plurality of reads, comprising obtaining a plurality of reads as strings of symbols; comparing each string of symbols corresponding to a read to a plurality of positions in a reference sequence construct using a computer processor, wherein the construct comprises at least two different strings of symbols at multiple positions in the construct, scoring overlaps between the string of symbols and each of the plurality of positions in the reference sequence construct with a processor, wherein a higher score corresponds to a greater amount of overlap; identifying the overlap corresponding to the highest score for each read; and aligning the reads to a location on the construct corresponding to

the highest score. In a further example, there is disclosed a computer implemented method of aligning a plurality of reads, comprising obtaining a plurality of reads as strings of symbols; aligning the reads against a reference sequence construct using a processor executing an algorithm that scores overlaps between the read and the construct, the construct comprising at least two alternative sequences per position at multiple positions in the construct. In yet another example, there is disclosed a system for aligning a plurality of reads comprising a processor and memory, wherein the memory comprises instructions that, when executed, cause the processor to obtain a plurality of reads as strings of symbols; align the reads against a reference sequence construct using an algorithm that scores overlaps between the read and the construct, the construct comprising at least two alternative sequences per position at multiple positions in the construct; and write a file to memory corresponding to the aligned sequence. In one example, there is disclosed a method of aligning a plurality of sequence reads, comprising obtaining a plurality of sequence reads; scoring sequence overlaps for each sequence read against a reference sequence construct, the construct comprising at least two alternative sequences per position at multiple positions in the construct, wherein greater overlap results in a lower score; and aligning each sequence read to a location in the construct such that the score for each sequence read is minimized. In another example, there is disclosed a computer-implemented method of aligning a plurality of reads, comprising obtaining a plurality of reads as strings of symbols; comparing each string of symbols corresponding to a read to a plurality of positions in a reference sequence construct using a computer processor, wherein the construct comprises at least two different strings of symbols at multiple positions in the construct, scoring overlaps between the string of symbols and each of the plurality of positions in the reference sequence construct with a processor, wherein a lower score corresponds to a greater amount of overlap; identifying the overlap corresponding to the lowest score for each read; and aligning the reads to a location on the construct corresponding to the lowest score. In a further example, there is disclosed a system for aligning a plurality of reads comprising a processor and memory, wherein the memory comprises instructions that, when executed, cause the processor to obtain a plurality of reads as strings of symbols; compare each string of symbols corresponding to a read to a plurality of positions in a reference sequence construct, wherein the construct comprises at least two different strings of symbols at multiple positions in the construct, score overlaps between the string of symbols and each of the plurality of positions in the reference

CLAIMS

1. A system for aligning a plurality of sequence reads comprising a processor and non-transitory memory, wherein the memory comprises instructions that, when executed, cause the processor to:
 - obtain a plurality of sequence reads from a subject;
 - compare each sequence read to positions in a reference directed acyclic graph (DAG) comprising nodes, in which genetic variation in a sequence is represented as alternate nodes, wherein the reference DAG comprises an alternate node having a string comprising a plurality of symbols at multiple positions in the reference DAG,
 - score overlaps between each sequence read and the positions in the reference DAG, wherein the scoring comprises considering predecessor nodes if and only if a symbol is the first symbol of a string associated with its node;
 - assign each sequence read to a location on the reference DAG corresponding to a highest score corresponding to a greater amount of overlap; and
 - write a file to memory corresponding to the location of each aligned sequence read.
2. The system of claim 1, wherein the sequence reads are nucleic acid sequence reads.
3. The system of claim 1, wherein the sequence reads are amino acid sequence reads.
4. The system of claim 1, wherein the memory additionally comprises instructions that, when executed, cause the processor to:
 - assemble the plurality of aligned sequence reads into an assembled sequence based upon the assigned location of each sequence read; and
 - write a file to memory corresponding to the assembled sequence.
5. The system of claim 4, wherein the assembled sequence corresponds to a genetic sequence of an organism.

6. The system of claim 5, wherein the genetic sequence substantially covers a chromosome or a genome.
7. The system of claim 1, further comprising assigning a genotype to an organism based upon the location of one or more aligned sequence reads.
8. The system of claim 7, further comprising correlating the assigned genotype with a risk of disease.
9. The system of claim 8, wherein the disease is cancer.
10. The method of claim 1, wherein the reference DAG comprises a database of positions and symbols stored on a non-transitory computer-readable medium.
11. The system of claim 1, wherein the system comprises a plurality of processors, and wherein each processor is configured to compare and score a portion of the plurality of sequence reads against the reference DAG.
12. The system of claim 1, wherein the memory additionally comprises instructions that cause a first processor to instruct a second processor to obtain, compare, score, identify, and assign, and then send the assigned location to the first processor.
13. The system of claim 1, wherein each position on the reference DAG corresponds to a nucleic acid in a reference sequence.
14. The system of claim 1, wherein each position on the reference DAG corresponds to an amino acid in a reference sequence.
15. The system of claim 1, wherein each position on the reference DAG corresponds to a gene in a reference sequence.

16. The system of claim 1, wherein the reference DAG construct has a direction.
17. The system of claim 16, wherein a path through the reference DAG in the direction represents a genome of an organism.
18. The system of claim 16, wherein a path through the reference DAG in the direction represents a chromosome of an organism.
19. The system of claim 16, wherein a path through the reference DAG in the direction represents a protein.
20. The system of claim 16, wherein a path through the reference DAG in the direction is acyclic.
21. The system of claim 1, wherein at least two different strings of symbols in the DAG represent known sequence variation at the position.
22. The system of claim 1, wherein at least two different strings of symbols correspond to nucleic acid sequences differing from each other by a base insertion or a base deletion.
23. The system of claim 1, wherein at least two different strings of symbols correspond to polypeptides differing from each other by an amino acid insertion or an amino acid deletion.
24. The system of claim 1, wherein the reference DAG comprises greater than 1,000 symbols.
25. The system of claim 24, wherein the reference DAG comprises greater than 1,000,000 symbols.
26. The system of claim 1, wherein the plurality of sequence reads comprises greater than 1000 sequence reads.

27. The system of claim 1, wherein the majority of the plurality of sequence reads are greater than 100 symbols in length.

28. The system of claim 1, wherein the plurality of sequence reads are output files of a next-generation sequencing method.

29. A method of aligning a plurality of sequence reads, comprising:
obtaining a plurality of sequence reads;
scoring sequence overlaps for each sequence read against a reference directed acyclic graph (DAG) comprising nodes, in which genetic variation in a sequence is represented as alternate nodes, the reference DAG comprising an alternate node having a string comprising a plurality of symbols of an alternative sequence per position at multiple positions in the reference DAG, wherein greater overlap results in a higher score and wherein the scoring comprises considering predecessor nodes if and only if a symbol is the first symbol of a string associated with its node; and

aligning each sequence read to a location in the reference DAG such that the score for each sequence read is maximized.

30. The method of claim 29, further comprising assembling the sequence reads to each other based upon the alignment of the sequences with respect to the reference DAG.

31. The method of claim 29, wherein the sequence read is a nucleic acid sequence read.

32. The method of claim 29, wherein the sequence read is an amino acid sequence read.

33. The method of claim 29, wherein each position in the reference DAG corresponds to a base or an amino acid in a reference sequence.

34. The method of claim 29, wherein each position in the reference DAG corresponds to a gene in a reference sequence.

35. The method of claim 29, wherein the reference DAG has a direction.
36. The method of claim 35, wherein a path through the reference DAG in the direction represents a genome of an organism.
37. The method of claim 35, wherein a path through the reference DAG in the direction represents a chromosome of an organism.
38. The method of claim 35, wherein a path through the reference DAG in the direction represents a protein.
39. The method of claim 35, wherein a path through the reference sequence is acyclic.
40. The method of claim 29, wherein at least two alternative sequences in the reference DAG represent known genetic variation at the position.
41. The method of claim 40, further comprising assigning a genotype for an organism based upon the assigned location of the nucleic acid reads.
42. The method of claim 29, wherein at least two alternative sequences in the reference DAG represent known structural variation at the position.
43. The method of claim 29, wherein at least two alternative sequences differ from each other by a base insertion or a base deletion.
44. The method of claim 29, wherein at least two alternative sequences differ from each other by an amino acid insertion or an amino acid deletion.
45. The method of claim 29, wherein each read comprises a string of symbols representing a measured sequence of nucleic acid bases.

46. The method of claim 45, wherein the measured sequence of nucleic acids bases is determined with a sequencing method selected from Sanger sequencing, pyrosequencing, ion semiconductor sequencing, sequencing by synthesis, sequencing by ligation, and single-molecule real-time sequencing.

47. The method of claim 29, wherein each read comprises a string of symbols representing a measured sequence of amino acids.

48. The method of claim 47, wherein the measured sequence of amino acids is determined with mass spectrometry.

49. The method of claim 29, wherein the reference DAG comprises greater than 1,000 bases.

50. The method of claim 49, wherein the reference DAG comprises greater than 1,000,000 bases.

51. The method of claim 29, wherein the plurality of reads comprises greater than 1000 reads.

52. The method of claim 29, wherein the majority of the plurality of reads are greater than 100 base pairs in length.

53. The method of claim 29, wherein the majority of the plurality of reads are greater than 10 amino acids in length.

54. The method of claim 29, wherein the reference DAG comprises a database of positions and sequences stored on a computer-readable medium.

55. The method of claim 54, wherein the reads comprise strings of symbols stored on a computer-readable medium.

56. The method of claim 29, further comprising sequencing a plurality of nucleic acids from a sample.

57. The method of claim 56, further comprising the sample obtained from a patient.

58. The method of claim 56, wherein the sample is selected from blood, urine, saliva, sputum, stool, nipple aspirate, sweat, hair follicle, buccal swab, or tissue.

59. The method of claim 58, further comprising isolating a plurality of nucleic acids from the sample.

60. The method of claim 59, further comprising amplifying at least a portion of the isolated plurality of nucleic acids.

61. A method of aligning a plurality of nucleic acid reads, comprising:
obtaining a plurality of nucleic acid reads;
aligning the nucleic acid reads against a reference directed acyclic graph (DAG) comprising nodes, wherein the aligning comprises using an algorithm that scores overlaps between a base in the read and the reference DAG, the reference DAG comprising an alternate node having a string comprising a plurality of symbols of an alternative sequence per position at multiple positions in the reference DAG, wherein the scoring comprises considering predecessor nodes if and only if a symbol is the first symbol of a string associated with its node.

62. The method of claim 61, wherein the algorithm calculates a maximum overlap score depending upon a choice between at least two sequences.

63. The method of claim 61, wherein the reference DAG comprises at least 1,000 bases.

64. The method of claim 63, wherein the reference DAG comprises at least 1,000,000 bases.

65. A method of aligning a plurality of amino acid reads, comprising:

obtaining a plurality of amino acid reads;

aligning the amino acid reads against a directed acyclic graph (DAG) comprising nodes, wherein the aligning comprises construct using an algorithm that scores overlaps between an amino acid in the read and the reference DAG, the reference DAG comprising an alternate node having a string comprising a plurality of symbols of an alternative sequence per position at multiple positions in the reference DAG, wherein the scoring comprises considering predecessor nodes if and only if a symbol is the first symbol of a string associated with its node.

66. The method of claim 65, wherein the algorithm calculates a maximum overlap score depending upon a choice between at least two sequences.

67. The method of claim 65, wherein the reference DAG is a directed acyclic graph.

68. The method of claim 65, wherein the reference DAG comprises at least 100 amino acids.

69. The method of claim 68, wherein the reference DAG comprises at least 10,000 amino acids.

70. A computer-implemented method of aligning a plurality of reads, comprising:
obtaining a plurality of reads from a subject;
comparing each read to positions in a reference directed acyclic graph (DAG) using a computer processor, the reference DAG comprising nodes, in which genetic variation in a sequence is represented as alternate nodes, wherein the reference DAG comprises an alternate node having a string comprising a plurality of symbols at multiple positions in the reference DAG,
scoring overlaps between each read and the positions in the reference DAG with a processor, wherein the scoring comprises considering predecessor nodes if and only if a symbol is the first symbol of a string associated with its node;
and
aligning the reads to a location on the DAG corresponding to a highest score corresponding to a greater amount of overlap.

71. The method of claim 70, further comprising assembling the reads to each other based upon the aligned location on the reference DAG.

72. The method of claim 70, wherein the reads are nucleic acid reads.

73. The method of claim 70, wherein the reads are amino acid reads.

74. The method of claim 70, wherein each position on the reference DAG corresponds to a base in a reference sequence.

75. The method of claim 70, wherein each position on the reference DAG corresponds to an amino acid in a reference sequence.

76. The method of claim 70, wherein each position on the reference DAG corresponds to a gene in a reference sequence.

77. The method of claim 70, wherein the reference DAG has a direction.

78. The method of claim 77, wherein a path through the reference DAG in the direction represents a genome of an organism.

79. The method of claim 77, wherein a path through the reference DAG in the direction represents a chromosome of an organism.

80. The method of claim 77, wherein a path through the reference DAG in the direction represents a protein.

81. The method of claim 77, wherein a path through the reference DAG in the direction is acyclic.

82. The method of claim 70, wherein at least two alternative sequences in the DAG represent known genetic variation at the position.

83. The method of claim 82, further comprising assigning a genotype for an organism based upon the assigned location of the nucleic acid reads.

84. The method of claim 70, wherein at least two alternative sequences differ from each other by a base insertion or a base deletion.

85. The method of claim 70, wherein each nucleic acid read comprises a string of symbols representing a measured sequence of nucleic acid bases.

86. The method of claim 85, wherein the measured sequence of nucleic acids bases is determined with a sequencing method selected from Sanger sequencing, pyrosequencing, ion semiconductor sequencing, sequencing by synthesis, sequencing by ligation, and single-molecule real-time sequencing.

87. The method of claim 70, wherein the reference DAG comprises greater than 1,000 bases.

88. The method of clam 87, wherein the reference DAG comprises greater than 1,000,000 bases.

89. The method of claim 70, wherein the plurality of nucleic acid reads comprises greater than 1000 reads.

90. The method of claim 70, wherein the majority of the plurality of nucleic acid reads are greater than 100 base pairs in length.

91. A computer implemented method of aligning a plurality of reads, comprising:
obtaining a plurality of reads from a subject ;
aligning the reads against a reference directed acyclic graph (DAG) comprising nodes, wherein the aligning comprises using a processor executing an algorithm that scores overlaps between the read and the reference DAG, the reference DAG comprising two nodes, the nodes including an alternate node having a string comprising a plurality of symbols representing an alternative sequence per position at multiple positions in the reference DAG, wherein the scoring

comprises considering predecessor nodes if and only if a symbol is the first symbol of a string associated with its node.

92. The method of claim 91, wherein the algorithm calculates a maximum overlap score depending upon a choice between at least two alternative sequences.

93. The method of claim 91, wherein the plurality of nucleic acid reads are stored on a computer-readable medium.

94. The method of claim 91, wherein the reference DAG comprises a database stored on a computer-readable medium.

95. A system for aligning a plurality of reads comprising a processor and memory, wherein the memory comprises instructions that, when executed, cause the processor to:
obtain a plurality of reads from a subject; align the reads against a reference directed acyclic graph (DAG) comprising nodes, wherein the aligning comprises using an algorithm that scores overlaps between the read and the DAG, the reference DAG comprising an alternate node having a string comprising a plurality of symbols of an alternative sequence per position at multiple positions in the reference DAG, wherein the scoring comprises considering predecessor nodes if and only if a symbol is the first symbol of a string associated with its node; and write a file to memory corresponding to the aligned sequence.

96. The system of claim 95, wherein the memory additionally comprises instructions that, when executed, cause the processor to calculate a maximum overlap score depending upon a choice between at least two alternative sequences.

97. The system of claim 95, wherein the reference DAG is a directed acyclic graph.

98. The system of claim 95, wherein the plurality of reads are stored on a computer-readable medium.

99. The system of claim 95, wherein the reference DAG comprises a database stored on a computer-readable medium.

100. A method of aligning a plurality of sequence reads, comprising:

obtaining a plurality of sequence reads; scoring sequence overlaps for each sequence read against a reference directed acyclic graph (DAG) comprising nodes, in which genetic variation in a sequence is represented as alternate nodes, the reference DAG comprising an alternate node having a string comprising a plurality of symbols of an alternative sequence per position at multiple positions in the reference DAG, wherein greater overlap results in a lower score and wherein the scoring comprises considering predecessor nodes if and only if a symbol is the first symbol of a string associated with its node; and aligning each sequence read to a location in the reference DAG such that the score for each sequence read is minimized.

101. A computer-implemented method of aligning a plurality of reads, comprising:

obtaining a plurality of reads from a subject;

comparing each read to positions in a reference directed acyclic graph (DAG) using a computer processor, the reference DAG comprising nodes, in which genetic variation in a sequence is represented as alternate nodes, wherein the reference DAG comprises an alternate node having a string comprising a plurality of symbols at multiple positions in the reference DAG,

scoring overlaps between each read and the positions in the reference DAG with a processor, wherein the scoring comprises considering predecessor nodes if and only if a symbol is the first symbol of a string associated with its node; and

aligning the reads to a location on the reference DAG corresponding to the lowest score corresponding to a greater amount of overlap.

102. A system for aligning a plurality of reads comprising a processor and memory, wherein the memory comprises instructions that, when executed, cause the processor to:

obtain a plurality of reads from a subject;

compare each sequence read to positions in a reference directed acyclic graph (DAG) comprising nodes, in which genetic variation in a sequence is represented as alternate nodes, wherein the reference DAG comprises an alternate node having a string comprising a plurality of symbols at multiple positions in the reference DAG,

score overlaps between each sequence read and the positions in the reference DAG, wherein the scoring comprises considering predecessor nodes if and only if a symbol is the first symbol of a string associated with its node;

assign the nucleic acid reads to a location on the reference DAG corresponding to a lowest score corresponding to a greater amount of overlap; and

write a file to memory corresponding to the aligned sequence.