TITLE OF THE INVENTION

[0001] String graph assembly for polyploid genomes

BACKGROUND OF THE INVENTION

[0002] Development in the biomolecule sequence determination has revolutionized the field of molecular and cellular biology. . However,the sequence information's quality must be carefully monitored, and has many factors that can be compromised including the biomolecule or the sequencing system used, including the composition of the biomolecule.

[0003] These factors can affect design the of a base-call true variant which might lead to miscall.

[0004] A string graph can be used to model a genome e.g. it can help us to assemble the genome from sequencing data. Modeling a genome has several advantages.

[0005] A string graph's vertex is the beginning of a sequence fragment, and an edge is the sequence fragment between two vertices. The crux of the string graph algorithm is to convert each "proper overlap" between two fragments into a string graph.

[0006] Some other features of a string graph are branching, knots and bubbles. Branching is caused because of repetitive sequences e.g. due to repeat regions in the genome. Knots, edges connected by the same node, caused because of many strings containing the same repeat. Simple bubbles are seen where structural variations are observed, and it is easy to resolve these.

[0007] Complex bubbles are observed that might be generally caused by more complicated repeats in haplotypes. .

[0008] It is important to distinguish between bubbles that are caused artificially due to structural variations in homologous sequences. Therefore there is a need for improving string

graph assembly for polyploid genomes.

BRIEF SUMMARY OF THE INVENTION

[0009] The invention is for processes to analyze sequence data from mixed populations of nucleic acids, to assign each string to a particular point of origin, and to ultimately identify sequences of biomolecular target sequences from the sequence information. The method provided here is not only sequence data having relatively high rates of insertions, deletions, and/or mismatch errors but also for relatively sequence have less errros. Therefore, the invention is also for systems that carry out these processes. Some of these methods are beneficial for sequencing polyploid organisms where the sequence string are assigned to a specific homolog.

[0010] In one embodiment, receiving step that includes matching the sequence reads; finding overlaps between the matched reads; determining consensus sequences from the matched reads, and building a string graph from these consensus sequences.

[0011] Another aspect, the exemplary embodiment might include doing some more steps in diploid assembly depending upon the primary and associated contigs.