Supplementary material

March 22, 2018

Overlap information from the assembler is first corrected to account for mismatches between the overlapping sequences. If the nominally overlapping sections of the overlapping nodes are not identical, then the nodes are aligned to each other. If the alignment indicates a perfectly-matching overlap between the appropriate ends of the nodes, then that overlap is used; otherwise, the edge between the nodes is discarded.

After this preprocessing, the bluntification algorithm uses a data structure called a *pinch graph* to transitively resolve overlaps between nodes. Used in the Cactus aligner, a pinch graph is essentially a union-find over oriented positions in a set of DNA sequences. Alternately, it can be thought of as a dynamic representation of a multiple sequence alignment with rearrangements. Two base-pair positions in the sequences that the graph is built on can be *pinched* together, placing them into the same column of the multiple alignment. Bases can be pinched either in consistent orientations (in which case the forward strands of the two positions are aligned) or in opposing orientations (in which case the forward strand of one position is aligned to the reverse strand of the other position). If bases being pinched are already aligned to other bases, the relevant alignment columns are merged; as with a union-find data structure's union operation, pinching is transitive.

Internally, the pinch graph data structure is implemented in terms of contiguous ranges of bases, and so the fundamental operation is to pinch together two equal-length ranges of bases in either consistent or opposing orientations. After a series of pinches, the graph consists of a set of *blocks*, each of which contains a series of oriented *segments*. The segments in a block are contiguous ranges of input sequence bases, and each segment in a block is aligned to all of the other segments in the block.

The bluntification algorithm creates a pinch graph from the nodes that participate in overlaps, pinches together the overlapping regions of all pairs of overlapping nodes, and then converts the resulting pinch graph back into a sequence graph, creating a node for each block.