We derive the paired-end multi-positional de Bruijn graph as a combination of the PE de Bruijn graph and the positional de Bruijn graph. Each k-mer of the PE multi-positional de Bruijn graph is composed of the k bases of the left read pair (called left k bases and the corresponding k bases of the right read pair called right k bases) and two left/right element pairs: the alignment position of each the left and the right k bases to the contigs, and the alignment position of each the left and the right k bases to the reference genome.

Two k-mers can be joined if they have similar k bases and close alignment positions within the constraints defined in the formulas below. Formally, let s be the left k bases from the left read pair and s' the corresponding k bases from the right read pair, then the k-mer of PE multi-positional de Bruijn graph is a 5-tuple \((s, c, g, c', g')\), where \(c\) is the alignment position of \(s\) to the contigs, \(g\) is the alignment position of \(s\) to the reference genome, \(c'\) is the alignment position of \(s'\) to the contigs, and \(g'\) is the alignment position of \(s'\) to the reference genome. Two k-mers \((s_i, c_i, g_i, c'_i, g'_i)\) and \((s_j, c_j, g_j, c'_j, g'_j)\) can be joined if the constraints (1)-(5) are met:

\[
\text{mismatch}(s_i, s_j) < \delta
\]

\[
|c_i - c_j| < \epsilon \text{ or } c_i = -1 \text{ or } c_j = -1
\] (1)

\[
|g_i - g_j| < \epsilon
\] (2)

\[
|c'_i - c'_j| < \epsilon + 2D \text{ or } c'_i = -1 \text{ or } c'_j = -1
\] (3)

\[
|g'_i - g'_j| < \epsilon + 2D
\] (4)

where \(\delta\) and \(\epsilon\) are small numbers with the default values 5 and 25, respectively, and \(D\) is the variability of the insert length \(I\) of the PE reads. The variability \(D\) is equal to \(\max\{I_u - I, I - I_l\}\) where \(I_u\) and \(I_l\) are the upper and lower limits of \(I\), respectively. The variables in the above formulas are explained below.

\(\delta\): To join two k-mers and tolerate sequencing errors, we allow a small number of mismatches \(\delta\) between \(s_i\) and \(s_j\) in (1).

\(\epsilon\): We allow a small shift \(\epsilon\) between each pair of alignment positions in (2)-(5), because the same k bases \(s_i\) and \(s_j\) from different reads may align to different but close positions in the contigs or genome as discussed in Ronen et al. (2012).

\(2D\): We allow a shift \(2D\) of \(s'_i\) and \(s'_j\)’s alignment positions to the contigs in (4) and to the reference genome in (5). The maximum and minimum alignment distances between a read pair are \(I - l + D\) and \(I - l - D\), respectively, where \(l\) is the read length, assuming the same read length for both members in a pair. Thus, the maximum alignment distance of two right reads with left reads aligned at the same position is \((I - l + D) - (I - l - D) = 2D\). This distance is equal to the distance between any two k-mers from the same position in the right read pairs, so the maximum distance between \(s'_i\) and \(s'_j\) will be \(2D\).

\(-1\): \(s_i\) and \(s_j\) can be joined if one or both of them are aligned directly to the reference genome rather than guided by the de novo contigs. In those cases, we assign -1 as alignment position to the contigs. This is important because we allow contig extensions only if the alignable and unalignable bases to contigs can be joined.