Certified Specialist Programme in Next-Generation Sequencing

Genome Assembly and Alignment

Genome assembly and alignment are fundamental concepts in the field of next-generation sequencing (NGS). These techniques are used to determine the order of nucleotides in a DNA sequence and to compare sequences from different sources, respectively. In this explanation, we will discuss key terms and vocabulary related to genome assembly and alignment.

Genome Assembly:

* Sequencing reads: Short DNA sequences generated by NGS technologies.
* De Bruijn graph: A data structure used in genome assembly to represent overlapping sequencing reads.
* Contig: A contiguous stretch of DNA sequence that has been assembled from overlapping sequencing reads.
* Scaffold: A series of contigs that have been ordered and oriented using information from paired-end sequencing reads.
* Genome closure: The process of filling in gaps between contigs and scaffolds to produce a complete genome sequence.
* Reference-based assembly: The process of assembling sequencing reads using a reference genome as a guide.
* De novo assembly: The process of assembling sequencing reads without using a reference genome.
* Chimeric read: A sequencing read that contains sequences from two or more different regions of the genome.
* Base calling: The process of determining the sequence of nucleotides in a DNA molecule from the electrical signals generated during sequencing.
* Quality score: A measure of the confidence in the base call at each position in a sequencing read.

Genome Alignment:

* Alignment: The process of comparing two or more DNA sequences to identify regions of similarity or difference.
* Local alignment: The process of identifying the best-matching regions between two sequences without considering the order of the sequences as a whole.
* Global alignment: The process of identifying the best-matching regions between two sequences while considering the order of the sequences as a whole.
* Pairwise alignment: The process of comparing two sequences to identify regions of similarity or difference.
* Multiple sequence alignment: The process of comparing three or more sequences to identify regions of similarity or difference.
* Homology: The similarity between two DNA sequences due to their descent from a common ancestor.
* Identity: The proportion of nucleotides in two sequences that are identical.
* Similarity: The proportion of nucleotides in two sequences that are similar, taking into account chemical

properties such as base-pairing.
* Gap: A space introduced in an alignment to account for insertions or deletions in one of the sequences.
* Penalty: A score assigned to a gap or mismatch in an alignment, used to evaluate the quality of the alignment.
* Seed-and-extend algorithm: A common approach to genome alignment, which involves identifying short, high-similarity regions (seeds) and extending them to form the final alignment.
* Dynamic programming: A mathematical technique used in genome alignment to find the optimal alignment between two sequences.

Practical Applications:

Genome assembly and alignment have numerous applications in biology and medicine. Genome assembly can be used to generate reference genomes for model organisms, enabling researchers to study gene function and regulation in detail. De novo assembly can be used to study the genomes of non-model organisms, shedding light on their biology and evolution. Genome alignment can be used to compare the genomes of different species, helping to identify regions of conservation and variation. In medicine, genome alignment can be used to identify genetic variations associated with disease, enabling the development of targeted therapies.

Challenges:

Despite the many advances in genome assembly and alignment, these techniques still face significant challenges. Genome assembly is complicated by the presence of repetitive sequences, which can lead to errors in contig and scaffold construction. Genome alignment is complicated by the presence of insertions, deletions, and rearrangements in the sequences being compared. Additionally, both techniques require large amounts of computational resources, making them time-consuming and expensive.

Examples:

A common tool for genome assembly is SPAdes, which uses a de Bruijn graph approach to assemble sequencing reads into contigs and scaffolds. A common tool for genome alignment is BLAST, which uses a seed-and-extend algorithm to identify regions of similarity between sequences.

Conclusion:

Genome assembly and alignment are essential techniques in the field of next-generation sequencing, enabling researchers to study the structure and function of genomes in detail. While these techniques are powerful, they also face significant challenges, including the presence of repetitive sequences and genomic variation. Nevertheless, with continued advances in sequencing technology and computational methods, genome assembly and alignment will continue to be important tools for biological and medical research.