Abstract : On the use of quality scores with short read mapping

New DNA sequencing technologies produce huge amounts of short DNA sequence reads. Often the initial bioinformatics task is to map these reads to a reference genome. For this, several new methods have been developed, which are all based on mapping the reads exactly apart from a few mismatches (sometimes allowing for insertions and deletions). We propose to use quality scores in a probabilistic framework to improve the mapping accuracy of short reads, give a more precise view of uniqueness, and even speed up mapping under certain circumstances. Using simulated reads, we show that this approach can improve mapping performance, especially for very short reads, which are relevant in the sequencing of ancient (fragmented) DNA and for sequencing of small RNAs.