Khan Inan
Professor Truong
BI-GY 7683

Mystery organism

When an organism does not have a reference genome it is much more difficult to classify what it may have connections with. The three sequence comparison methods of alignment, assembly, and phylogeny may be used in cases where there is no reference genome. Valuable information can be obtained in regards to the organism's origin, similarities and differences to the genome it is being compared to. If we have a file of sequence reads that we need to analyze, it is important to understand that the information may be fragmented and difficulties may arise. For example, one sequence read may only be mappable to a small section of the chosen reference read, while another may be mappable to multiple locations on the reference read. So this is why choosing a correct reference read to compare our sequence reads with is as important as analyzing the alignment itself.

In observing the alignment of a read and reference genome, there are several kinds of differences which are important to be able to identify. Three differences that one may see when observing alignment are substitution, deletion and insertion. Substitution most often affects protein synthesis of an organism and can be as simple as a change in a single nucleotide. If a substitution is spotted during alignment comparison, then we can at the very least conclude that the mystery organism and the reference organism can have similar origins, since the majority of the sequences are still identical. If several complex differences like deletion and insertion are found during alignment, however, then the data may lead towards the organisms being unrelated and a different reference genome may need to be used.

Since it may be very difficult or impossible to find a compatible reference read, assembly may be the more preferred method of genome mapping in this case, especially since we have a file of sequence reads that may be scattered with duplicates, incomplete and ambiguous fragments. depending on the size of the contigs within the file of sequence reads, we may or may not be able to have a reliable assembly of the genome. One method of assembly known as De Bruijn Graph construction, for example, has a direct correlation of having a higher confidence value (trust or accuracy of the assembly) with larger size contigs.

Lastly, Phylogenetics is what is going to tie together all our findings from alignment analysis and assembly. The best way to describe the entire process of mapping this unknown organism is that alignment gives us a general idea of the similarities and differences between our sequence reads and the reference reads, assembly allows us to begin mapping the genome of this organism and phylogeny is what gives us information about this organism’s origins and how it relates to other biological entities.