Introduction

Metagenomic sequencing data has increased the understanding of the role of the microbiome and given insight into a variety of (bacterial) communities. *Korem et al.* [SOURCE] used the pattern of metagenomic sequencing read coverage around the origin of replication (oriC) to gain insight into the growth dynamics of gut microbiota. Detecting the oriC in bacteria can hence be of great importance for metagenomic analysis. With the oriC being the starting point of bacterial replication, its detection can facilitate the analysis of certain motifs and features around this location or simply be used as a starting point for gene annotation.

A frequently used indicator of the oriC is the so called skew which is based on the strand asymmetry between the leading and lagging strand. The leadings strand usually is rich in guanine (G) and adenine (A) whereas a higher content of cytosine (C) and thymine (T) can be found in the lagging strand [SOURCE]. The putative location of the oriC is then indicated by a minimum of the GC skew (G-C)/(G+C) or a maximum in the AT skew respectively. Even though the GC skew is frequently used, the AT skew is able to cover for less pronounced regions in the GC skew [SOURCE, Grigoriev].

Another measure used for oriC prediction is the DnaA motif. DnaA is the key protein in the initiation of replication. It binds to clusters of DnaA boxes that accumulate around the oriC. The DnaA motif is nine base pairs long and is highly conserved with the consensus sequence “TT A/T TNCACA” [SOURCE, Blaesing et al]. Detecting the location of these clusters of DnaA boxes can therefore improve the prediction of a putative oriC.

This project tries to identify putative oriC locations of four different bacteria species (*Escherichia coli, Vibrio cholerae, Salmonella enterica, Thermotoga petrophila*). The approximate region of the oriC is identified using GC and AT skews. In this region, the DnaA motif clusters are then located to further specify the region.

//Last paragraph has to be changed depending on whether we compute a general DnaA motif or species specific

* Find DnaA motif with meme (for the corresponding strain?)
* Find positions of the DnaA motif near the max/min of the skew
* Predict the putative location of the ori
* Detect multiple origins(?)