*CDS prediction and additional HMM refinement*

The genome annotation of public available phages are the product of gene prediction programs with different sensitivity. This results in genomes where some coding sequences (CDS) have not been annotated. To examine the value of HMMs to identify such missing phage CDS, the intergenic regions of each phage Genbank files used in this study was scanned using the profile HMMs. In total 234 nucleotide regions have been identified that encode gene products that align to one of the protein families modeled by the HHMs (Table Ign\_hmm\_scan.csv). In deed profile HMMs can be used to identify missing CDS.

To investigate whether these new proteins may improve the profile HMMs we generated refined HMMs the original proteins plus the new identified CDS as described in the material and methods section (ClassiPhage). An evaluation of the refined HMMs identified exactly the same proteins per HMM with slightly moderated hit scores. The test revealed that the refinement of the HMMs did not yield better performing HMMs. The sensitivity of HMMs is correlated much stronger to the diversity than to the number of the proteins used in the initial alignment step. We concluded that our original profile HMMs already contain sufficient diverse protein to model the protein families and thus the models predictive power is already close to saturation.