INPUT:
Protein Sequence = ATWGRTG
$\ell = 7$
k = 3

DebruijnExtend

Dreycey Albin, Angela Folz

OUTPUT (most probable):
CCEEEEE Prob: 0.1215
TC: $O(3^{\ell+1} * (\ell - \mathrm{k} + 1))$

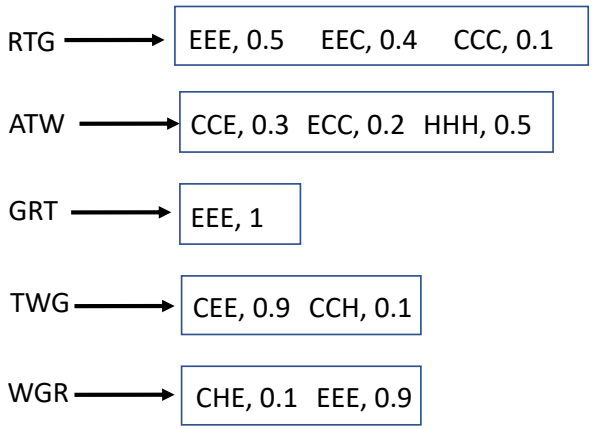
Example of using software (CMD line)

python DebruijnExtend.py gfp.fasta 4 gfp.ss3

```
CCCCCCCCCCCEEEEEEEEEEECCEEEEEEEEEE
ECCCCEEEEEEECCCCCCCCHHHHCCCCCCHHHC
-log(P)=231.1
```

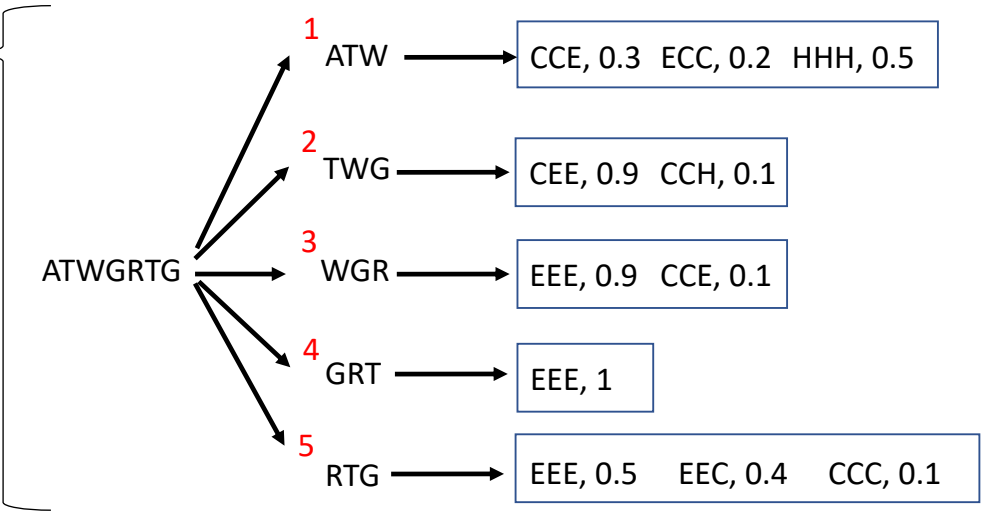
STEP 0 – Hash Table (Training)

For each k-mer in a training database, find every possible secondary structure and its probability.



STEP 1 – K-mer Mapping

Look up the corresponding set of secondary structures for each k-mer using the precomputed hash table.



STEP 2 – Stitch-Extend

Use BFS to traverse the Debruijn graph to find the highest weighted path. This is implemented using a dynamic programming method where the subproblem is matching the contracted nodes to the nodes of the next layer.