**Open Reading Frames**

In molecular genetics, an **open reading frame (ORF)** is the part of a reading frame that contains no stop codons. The transcription termination pause site is located after the ORF, beyond the translation stop codon, because if transcription were to cease before the stop codon, an incomplete protein would be made during translation.

**Problem summary**

One common use of open reading frames is as one piece of evidence to assist in gene prediction. Long ORFs are often used, along with other evidence, to initially identify candidate protein coding regions in a DNA sequence. The presence of an ORF does not necessarily mean that the region is ever translated. For example in a randomly generated DNA sequence with an equal percentage of each nucleotide, a stop-codon would be expected once every 21 codons. A simple gene prediction algorithm for prokaryotes might look for a start codon followed by an open reading frame that is long enough to encode a typical protein, where the codon usage of that region matches the frequency characteristic for the given organism's coding regions. By itself even a long open reading frame is not conclusive evidence for the presence of a gene.

**Example**

If a portion of a genome has been sequenced (e.g. 5'-ATCTAAAATGGGTGCC-3'), ORFs can be located by examining each of the three possible reading frames on each strand. In this sequence two out of three possible reading frames are entirely open, meaning that they do not contain a stop codon:

1. ...A TCT AAA ATG GGT GCC...
2. ...AT CTA AAA TGG GTG CC..
3. ...ATC TAA AAT GGG TGC C...

Possible stop codons in DNA are "TGA", "TAA" and "TAG". Thus, the last reading frame in this example contains a stop codon (TAA), unlike the first two.

**Pseudo code**

***# Use a set since we want to return distinct protein.***

***# Sets keep track of distinct elements without us needing to worry about adding duplicates.***

***# Find for the Start codon.***

***# Use a new index since we'll want to return to the ith position of the strand in case there are multiple start codons in a row.***

***# Continue, if necessary, until we hit the end of the DNA sequence.***

***# Add the protein and break if we hit a Stop codon.***

***# Otherwise, add to the current protein.***

***# Convert protein from a set to list of strings to allow output to be written in the correct form more efficiently.***