3. Gene prediction

- All genes end on a stop codon
- A simple algorithm for gene prediction
- **Searching for start and stop codons**
- Predicting all the genes in a sequence
- Making the predictions more reliable
- Boyer-Moore algorithm
- Index and suffix trees
- Probabilistic methods
- Benchmarking the prediction methods
- Gene prediction in eukaryotic genomes

Searching for stop and start codons

Searching for triplets

```
TAG
ATTGCTTACTAGAATCGTACGGGTACGTAAATCGTATTCCGAT
```

Searching for triplets

```
   TAG
ATTGCTTACTAGAATCGTACGGGTACGTAAATCGTATTCCGAT
```

Searching for triplets

```
      TAG
ATTGCTTACTAGAATCGTACGGGTACGTAAATCGTATTCCGAT
```

Searching for triplets

```
          TAG
ATTGCTTACTAGAATCGTACGGGTACGTAAATCGTATTCCGAT
```

Searching for triplets

```
          TAG
ATTGCTTACTAGAATCGTACGGGTACGTAAATCGTATTCCGAT
```

Number of comparisons in the worst case for one of the three stop codons

= length of the text

For the three stop codons

= 3 * length of the text

Stop codons

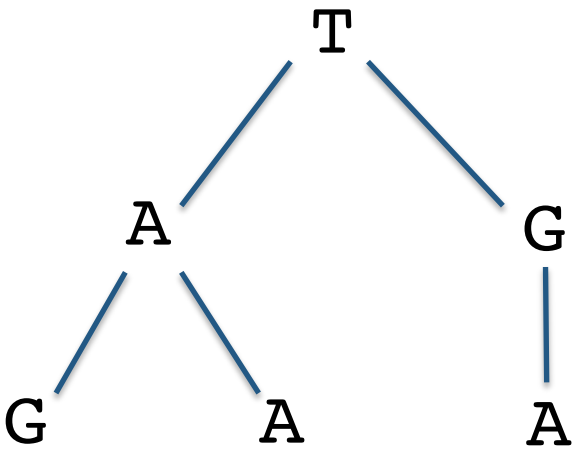
TGA

or

TAG

or

TAA



The NextStopCodon function

```
Function NextStopCodon (index: integer) returns integer
   if index + 2 ≤  lengthSequence
   then
     repeat
          if Sequence [index] = "T"
               then
                    if Sequence [index+1] = "A"
                         then if sequence [index+2] = "G" or sequence [index+2] = "A"
                                   then return index
                                   else index ← index + 2
                         else if  Sequence [index+1] = "G" and sequence [index+2] = "A"
                                   then return index
                                   else index ← index + 2
               else index ← index + 3
     until index ≥ lengthSequence - 2
     return 0
    else return 0
end NextStopCodon
```