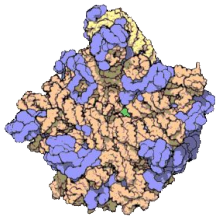# 4. Prediction of non coding genes
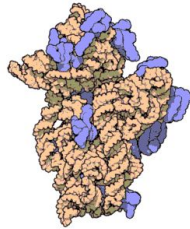
Guillaume GAUTREAU, 28/09/2022

Different kinds of non coding genes : rRNA

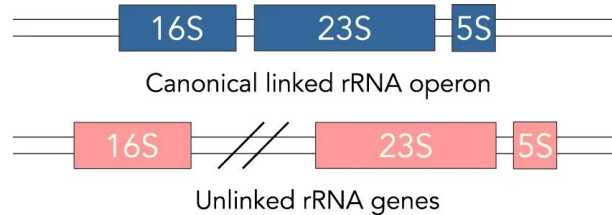- Several clusters composed of 3 rRNA genes can be detected in the genome

credit: Wikipedia

Large subunit 50S
(23S+5S+proteins)

small subunit 30S
(16S+proteins)

16S   23S   5S

Canonical linked rRNA operon

16S   23S   5S

Unlinked rRNA genes

Brewer *et al.,* 2020

- One or some tRNA genes often lie between 16S and et 23S genes (case of linked rRNA genes)

- 16S rRNA genes are conserved in some regions and variables in other ones (V1,V2…) :
  - Largely used for phylogeny (Carl Woese, 1977)
  - And for metabarcoding to describe the biological content of a sample ("metagenomic")

16S rRNA genes

V1   V2   V3   V4   V5   V6   V7   V8   V9

credit: https://www.repertoire.co.jp/en/research/technology/16srrnainfo/

Different kinds of non coding genes : rRNA

- Detection of rRNA can be performed using RNAmmer
  1. patterns of rRNA genes are stored as Hidden Markov Model profiles (HMM) :

An alignment
of rRNA genes:

AAATT
TTATC
CTACT
ATACC
CCTTC
TGATA
CGAAC
TGACC
CGACC
CGATC
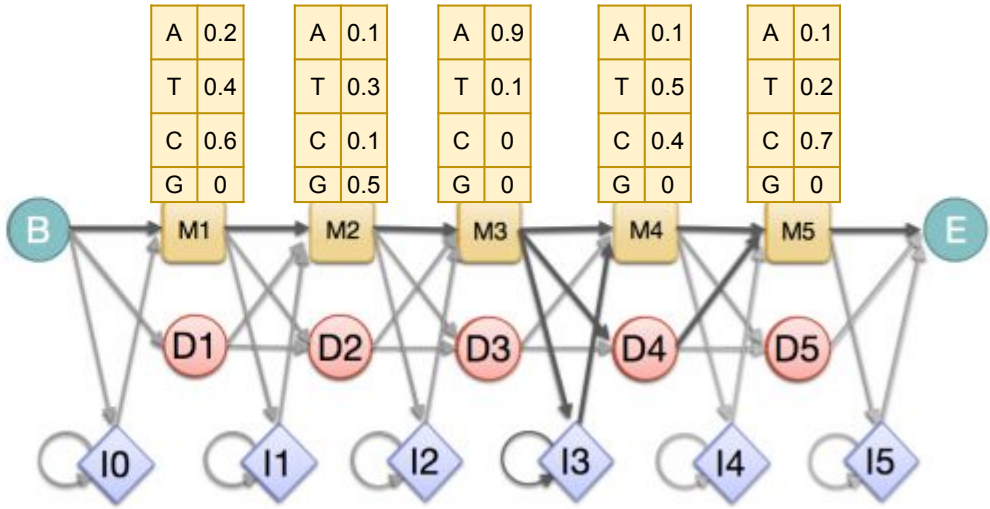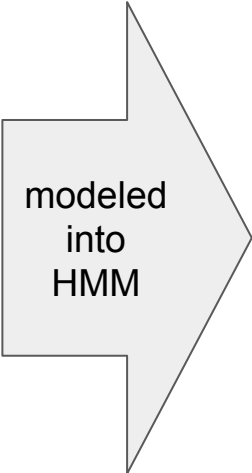
modeled
into
HMM

	M1		M2		M3		M4		M5
A	0.2	A	0.1	A	0.9	A	0.1	A	0.1
T	0.4	T	0.3	T	0.1	T	0.5	T	0.2
C	0.6	C	0.1	C	0	C	0.4	C	0.7
G	0	G	0.5	G	0	G	0	G	0

B M1 M2 M3 M4 M5 E

D1 D2 D3 D4 D5

I0 I1 I2 I3 I4 I5

modified from: Wikipedia

3

Different kinds of non coding genes : rRNA

- Detection of rRNA can be performed using RNAmmer
- 2. sequences of our genome are aligned on the HMM profiles :

sequences from the genome

…CTAGTC…

Alignment

probability
of matching

A	0.2
T	0.4
C	0.6
G	0

A	0.1
T	0.3
C	0.1
G	0.5

A	0.9
T	0.1
C	0
G	0

A	0.1
T	0.5
C	0.4
G	0

A	0.1
T	0.2
C	0.7
G	0

B M1 M2 M3 M4 M5 E

D1 D2 D3 D4 D5

I0 I1 I2 I3 I4 I5
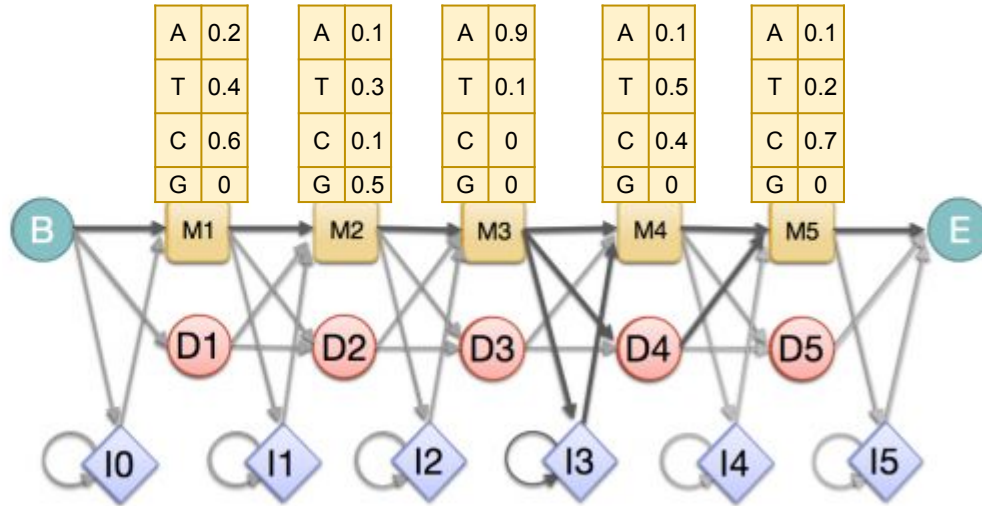
Different kinds of non coding genes : rRNA

- Detection of rRNA can be performed using RNAmmer
  3. the sequence having the best probability of matching corresponds to the genes encoded in the HMM profiles

...CTAGTC...

Alignment

probability
of matching

match   not a match

A	0.2
T	0.4
C	0.6
G	0

A	0.1
T	0.3
C	0.1
G	0.5

A	0.9
T	0.1
C	0
G	0

A	0.1
T	0.5
C	0.4
G	0

A	0.1
T	0.2
C	0.7
G	0

B M1 M2 M3 M4 M5 E

D1 D2 D3 D4 D5

I0 I1 I2 I3 I4 I5

modified from: Wikipedia

5
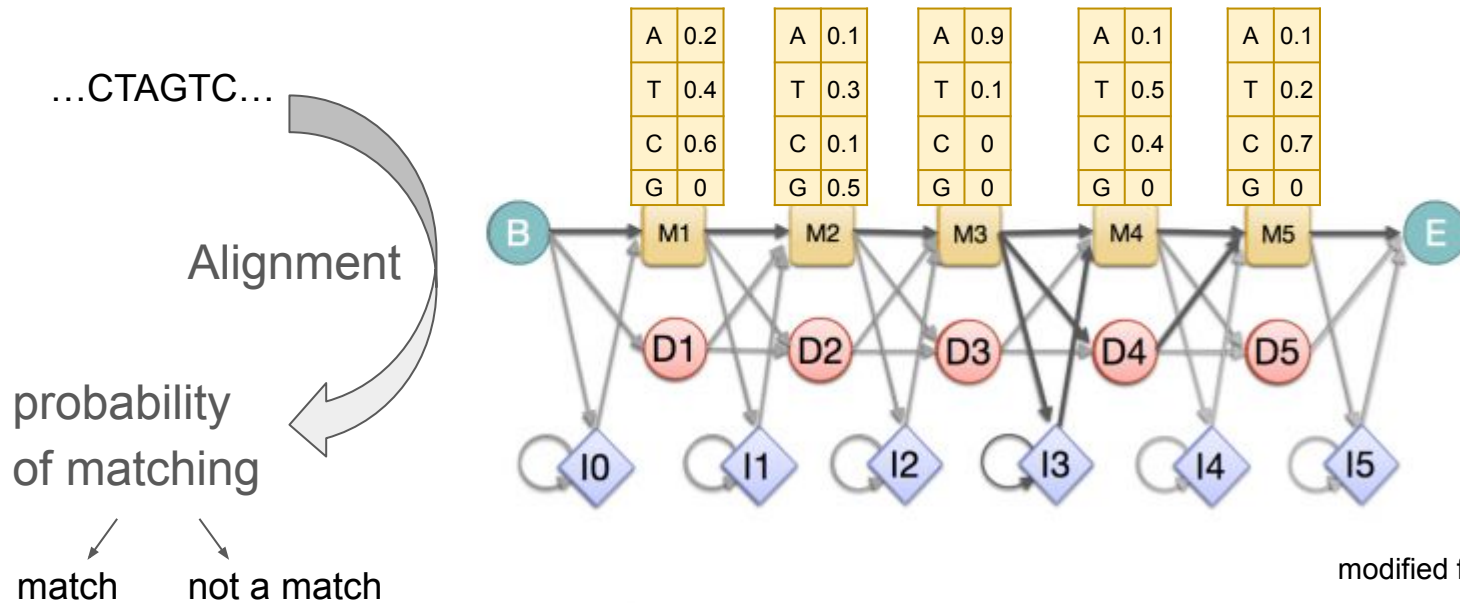
Different kinds of non coding genes : tRNA

- tRNA : several clusters of tRNA scattered all along the genome
  - One or several tRNAs for each codon
  - The pool of tRNA drives the codon usage
  - Very stable across bacterial genomes (except the variation loop and, of course, the anticodon)

credit: https://fr.dreamstime.com/arn-trna-code-g%C3%A9n%C3%A9tique-image113257763

Different kinds of non coding genes : tRNA

- Detection of tRNA can be performed using ARAGORN
- Here is the algorithm (an expert recipe):
  1. Searches for partially mismatched, non-gapped, occurrences of the sequence GTTC (subset of the B box consensus sequence)
  2. Around each hit, the algorithm attempts to construct a T-loop from five to nine bases long and a T-stem from 4 to 5 bp long.
  3. To detect tRNA genes, the sequence is searched from 28 to 85 bases upstream of this T-stem for the sequence motif TRGYNAA, a subset of the A box consensus sequence which allows for a D-stem from 3 to 4 bp long.
  4. Around the motif, a D-loop from five to 11 bases long and containing the sequence A- - - - -GG-R is constructed.
  5. A 7–9 bp long A-stem is constructed using the sequence from two to three bases upstream of the D-stem and immediately downstream of the T-stem.
  6. The length of the V-loop is allowed to vary from three to 25 bases, upstream of the T-stem.
  7. Finally, the 3′ end of the C-stem is constructed by searching between the D- and T-stems for a sequence that is complementary to the 5′ end of the C-stem, immediately downstream of the D-stem and spacer base.
  8. Limited use of tertiary structures contacts between the T-loop, V-loop and D-loop is made.
  9. If position 55 in the T-loop is a non-consensus G, then a non-consensus TT at positions 18 and 19 in the D-loop is given an improved score.
- Useless to have a complete understanding (to my mind)

tRNA

Arg-tRNA

Freire *et al.*, 2010

7

Different kinds of non coding genes : tmRNA

- tmRNA allows recycling stalled ribosomes during the translation of a truncated mRNA (so without any stop codon)

credit: wikipedia
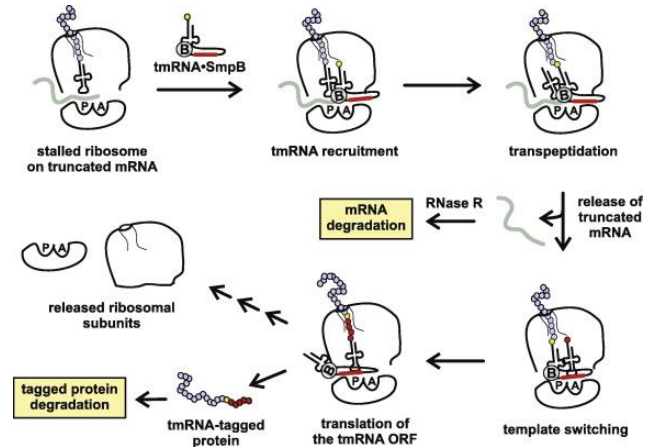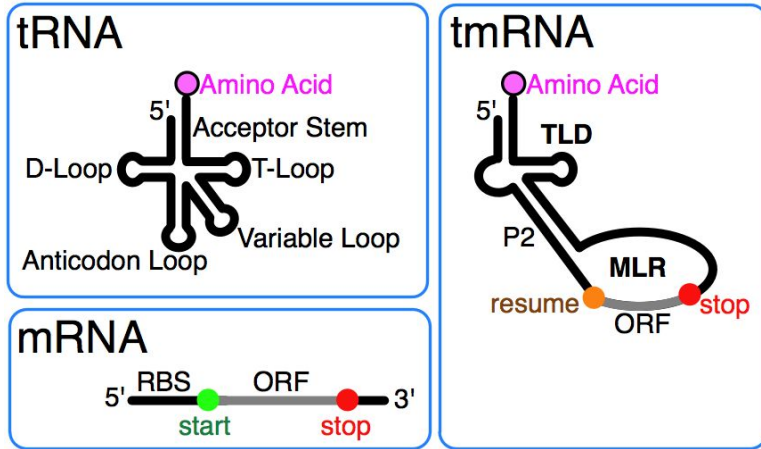
Hayes *et al.*, 2010

- Detection of tRNA can also be performed using ARAGORN using an expert recipe

Different kinds of non coding genes : ncRNA

- Non coding RNAs correspond to a set of RNA (having varieties of functions) which are not mRNA, rRNA, tRNA, and tmRNA
- Detection of tRNA can be performed using Infernal based on Rfam, which is database of HHM profiles of ncRNA

Conclusion

- Coding genes still correspond to the main part of genomes but non coding genes can't be ignored
- Detection of non coding genes can help annotation of coding genes (overlaps)
- The diversity of non coding genes is detected via 2 main kinds of approaches
  - Alignment on a database of HMM profiles
  - Expert recipes ruling a specific pattern to match