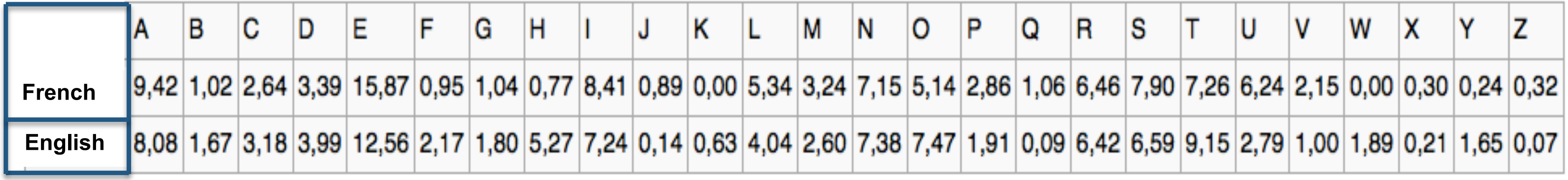
3. Gene prediction

- 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

Using letter frequencies

- Passages written in an unknown human-understandable language are hidden in a larger random sequence of letters
- How to identify these passages?

Letter frequencies in French and English



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
French	9,42	1,02	2,64	3,39	15,87	0,95	1,04	0,77	8,41	0,89	0,00	5,34	3,24	7,15	5,14	2,86	1,06	6,46	7,90	7,26	6,24	2,15	0,00	0,30	0,24	0,32
English	8,08	1,67	3,18	3,99	12,56	2,17	1,80	5,27	7,24	0,14	0,63	4,04	2,60	7,38	7,47	1,91	0,09	6,42	6,59	9,15	2,79	1,00	1,89	0,21	1,65	0,07

Using letter frequencies

- Passages written in an unknown human-understandable language are hidden in a larger random sequence of letters
- How to identify these passages?

- Compute letter frequencies in sliding windows
- Compare computed frequencies from expected frequencies
- Apply a statistical test (χ^2) to check the differences are meaningful

On genomic sequences

- Detection of biases in codon usage
- Markov chains
  - Compute probabilities to observe one specific nucleotide after k have just been encountered (conditional probabilities)
  - k is the order of the model

- Training phase: computation of a transition matrix
- Prediction phase: use the matrix to discriminate between coding and non-coding regions

The association of pattern searching
(stop and start codons, RBS)
with a well-trained Markov model
provides quite good results
on most bacterial genomes