CMSC 423: Sequence Alignment

Part 1

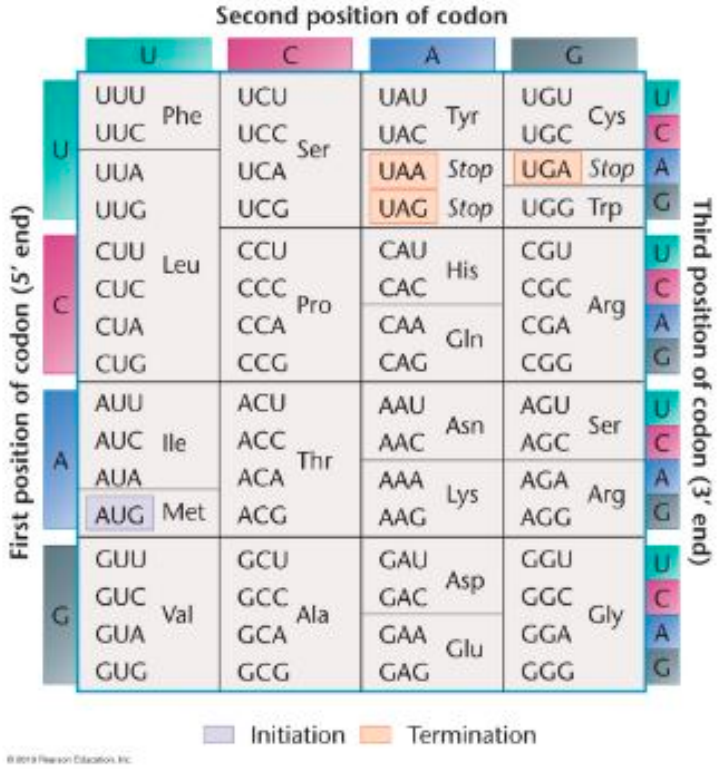
Inexact matching: why?

- Redundancy in the genetic code: nucleotide sequence may differ, but proteins are the same

S Y P T D

TCTTATCCTACTGAT

TCATACCCCACAGAC



Inexact matching: why?

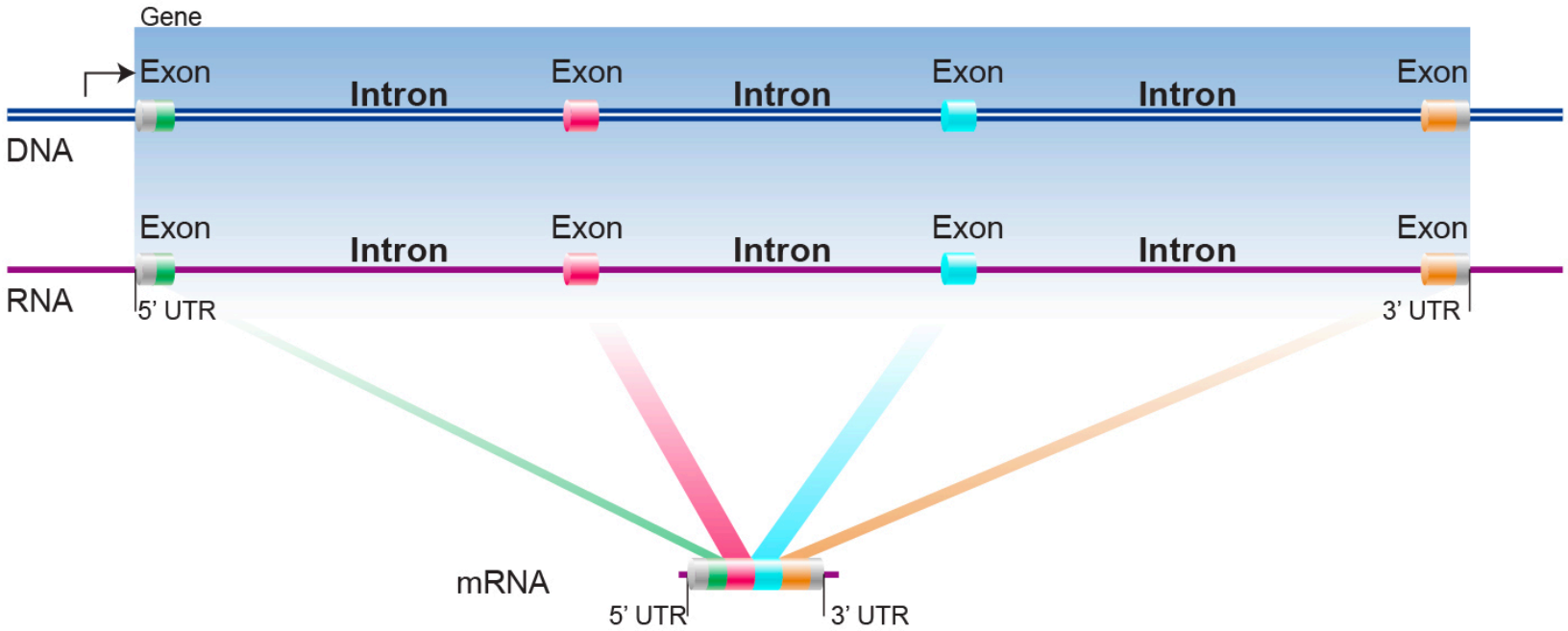
- Redundancy in the genetic code: nucleotide sequence may differ, but proteins are the same
- Different amino-acid sequences can still fold the same way: function is unchanged

Inexact matching: why?

- Redundancy in the genetic code: nucleotide sequence may differ, but proteins are the same
- Different amino-acid sequences can still fold the same way: function is unchanged
- Aligning RNA sequences to DNA- need to account for gaps corresponding to exons

Splicing

- Genes contain exons (portions that code for amino acids) and introns (portions that do not code for amino acids)
- During splicing, introns are removed and exons are joined together



https://www.genome.gov/genetics-glossary/Intron

Inexact matching: why?

- Redundancy in the genetic code: nucleotide sequence may differ, but proteins are the same
- Different amino-acid sequences can still fold the same way: function is unchanged
- Aligning RNA sequences to DNA- need to account for gaps corresponding to exons
- Sequencing errors

How can we compare two sequences?

- Hamming distance: the number of mismatches in a string

```
ATGCATGC
TGCATGCA
```

Hamming distance = 8

How can we compare two sequences?

- What if we align the sequences differently?

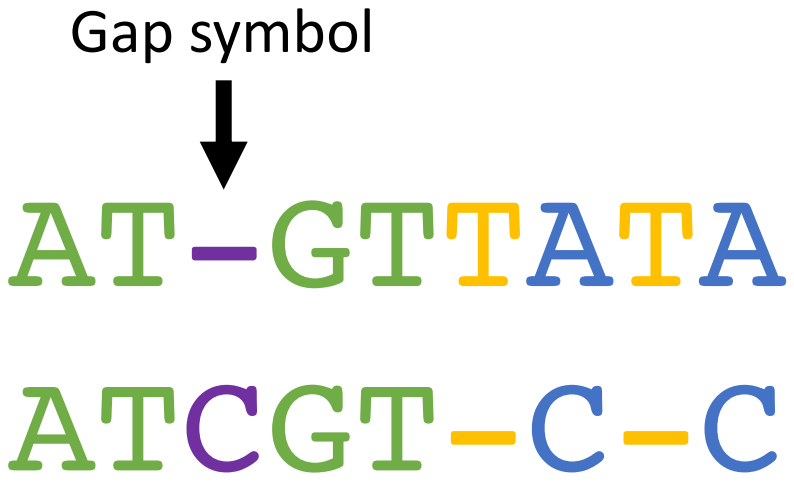
```
ATGCATGC-
-TGCATGCA
```

We have much fewer mismatches!

Alignment of sequences *v* and *w*

v = ATGTTATA

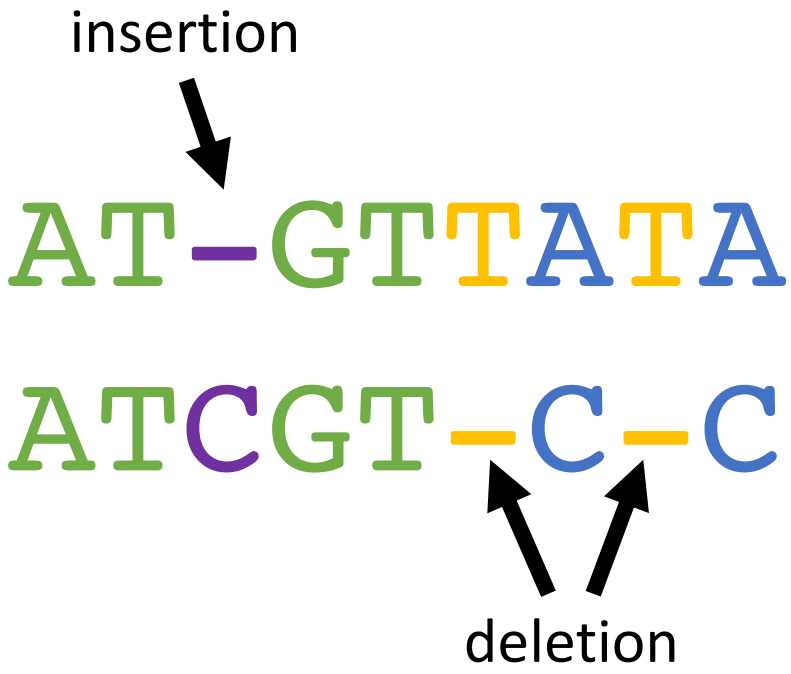
w = ATCGTCC



Alignment of sequences *v* and *w*

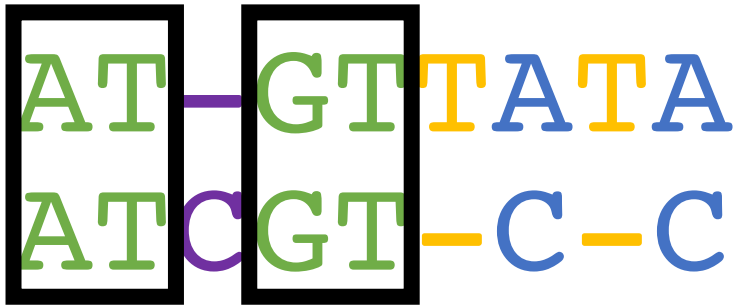
v = ATGTTATA

w = ATCGTCC



Alignment of sequences *v* and *w*

v = ATGTTATA
w = ATCGTCC



Common subsequence: `ATGT`

Longest Common Subsequence (LCS)

- An alignment of two string maximizing the number of matches corresponds to the longest common subsequence
- Two strings can have more than one longest common subsequences
- How do we solve this?