

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

		Second position of codon				
		U	C	A	G	
First position of codon (5' end)	U	UUU Phe	UCU	UAU Tyr	UGU Cys	U
		UUC	UCC Ser	UAC	UGC	C
		UUA	UCA	UAA Stop	UGA Stop	A
		UUG	UCG	UAG Stop	UGG Trp	G
	C	CUU Leu	CCU	CAU His	CGU	U
		CUC	CCC Pro	CAC	CGC Arg	C
		CUA	CCA	CAA Gln	CGA	A
		CUG	CCG	CAG	CGG	G
	A	AUU Ile	ACU	AAU Asn	AGU Ser	U
		AUC	ACC Thr	AAC	AGC	C
		AUA	ACA	AAA Lys	AGA Arg	A
		AUG Met	ACG	AAG	AGG	G
G	GUU Val	GCU	GAU Asp	GGU Gly	U	
	GUC	GCC Ala	GAC	GCC	C	
	GUA	GCA	GAA Glu	GGA	A	
	GUG	GCG	GAG	GGG	G	

■ Initiation ■ Termination

Inexact matching: why?

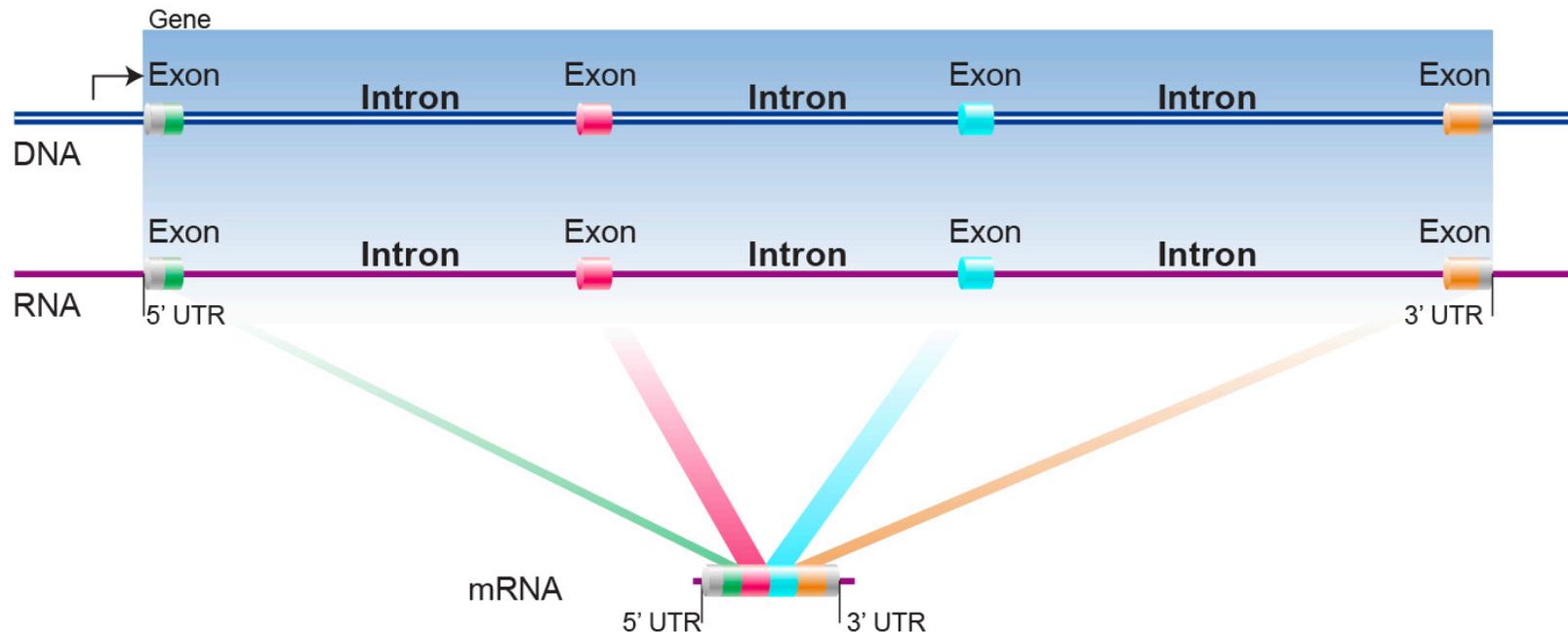
- Redundancy in the genetic code: nucleotide sequence may differ, but proteins are the same
- Different amino-acid sequences can still fold the same way: function is unchanged

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



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?

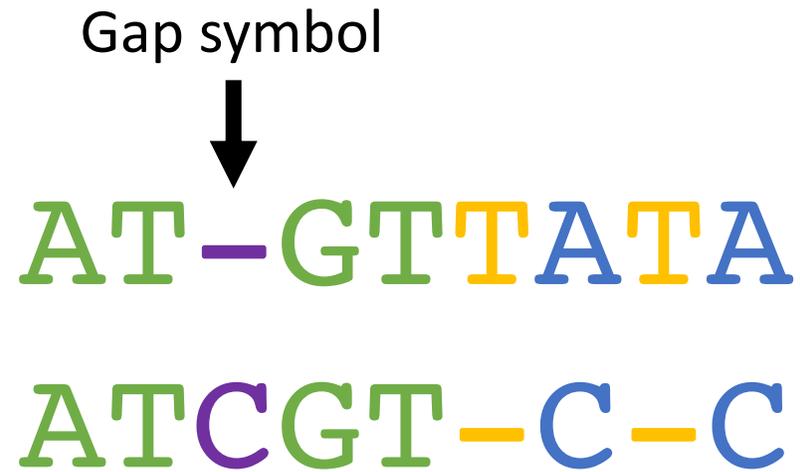
```
ATGCATGC-  
-TGCATGCA
```

We have much fewer mismatches!

Alignment of sequences v and w

$v = \text{ATGTTATA}$

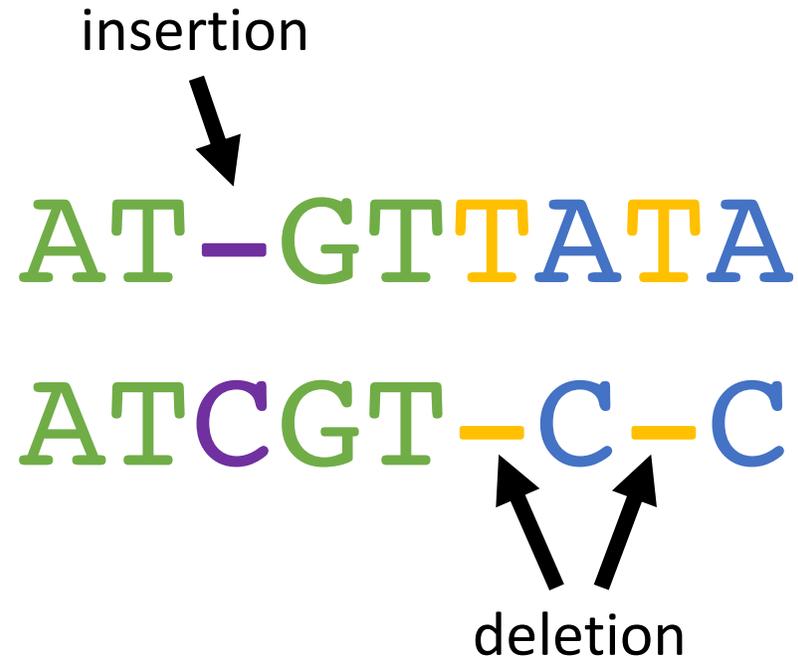
$w = \text{ATCGTCC}$



Alignment of sequences v and w

$v = \text{ATGTTATA}$

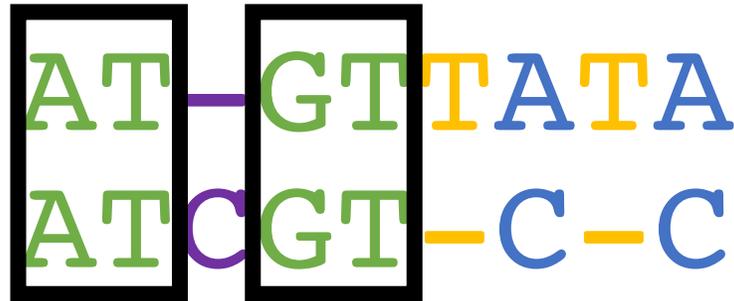
$w = \text{ATCGTCC}$



Alignment of sequences v and w

$v =$ ATGTTATA

$w =$ ATCGTCC



Common subsequence: ATGT

Longest Common Subsequence (LCS)

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