CS4054
Bioinformatics

Spring 2025

Rushda Muneer

Introduction to Sequence Alignment

- Comparing Genes is a Fundamental Problem in Biology
- **Comparing Genes Problem:**
  - **Input:** Two genes.
  - **Output:** How “similar” these genes are.
- **Goal:** Convert this important biological question into a well-defined computational problem.

Try 1: Hamming Distance

- **Hamming Distance Problem:**
  - **Input:** Two strings.
  - **Output:** The number of “mismatched” symbols in the two strings.
- Example:
  - ATGCATGC
  - TGCATGCA
  - Hamming distance = 8
- STOP & Think: What are the issues with this approach?
  - ATGCATGC
  - TGCATGCA
- These strings have a long shared substring, it just doesn’t line up perfectly.

Try 2: Longest Substring

- **Longest Shared Substring Problem:**
  - **Input:** Two strings.
  - **Output:** The longest substring shared by both strings.
- **STOP & Think:** What are the weaknesses of using the length of a longest shared substring to represent the similarity between two strings?
- Consider the strings
- AAACAAACAAACAAACAAACAAA
- AAAGAAAGAAAGAAAGAAAGAAAGAAA
- These strings are very similar, but they don't have a long shared substring in common.

Try 3: Counting Shared k-Mers

- Instead of finding a longest shared substring of two strings, we will count the number of shared substrings.
- For simplicity, we restrict to substrings of the same length; recall that a k-mer is the term we use in comp bio for a string of length k.

Try 3: Counting Shared k-Mers

s1 = ACGTATACACGTAT

String	Count
ACA	1
ACG	2
ATA	1
CAC	1
CGT	2
GTA	2
TAC	1
TAT	2

s2 = TATCGGTATATCCTAC

String	Count
ATA	1
ATC	2
CCT	1
CGG	1
CTA	1
GGT	1
GTA	1
TAC	1
TAT	3
TCC	1
TCG	1

STOP & Think: How should we count the # of shared 3-mers of two strings?

Try 3: Counting Shared k-Mers

s1 = ACGTATACACGTAT

String	Count
ACA	1
ACG	2
ATA	1
CAC	1
CGT	2
GTA	2
TAC	1
TAT	2

Take minimum counts for each shared k-mer:
1 + 1 + 1 + 2 = 5

s2 = TATCGGTATATCCTAC

String	Count
ATA	1
ATC	2
CCT	1
CGG	1
CTA	1
GGT	1
GTA	1
TAC	1
TAT	3
TCC	1
TCG	1

STOP & Think: What remaining weakness do you see with counting k-mers?
Answer: We lose info about the order of the shared strings.

Toward a Better Approach

- What similarities do you see in these strings?

```
ATGCTTA
TGCATTAA
```

- We can find similarities if we "*slide*" the strings, letting symbols shift (but stay in same order).

```
ATGC-TTA-
-TGCATTAA
```

Toward a More Accurate Problem

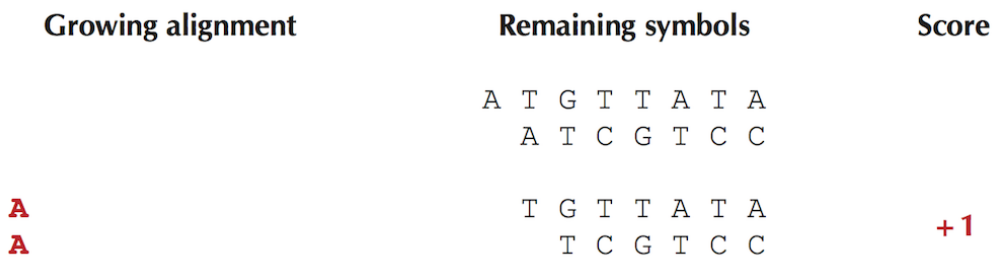
- **Symbol Matching Problem:**
  - **Input:** Two strings.
  - **Output:** The greatest number of matched symbols in any “alignment” of the two strings.
- **Exercise:** How many matches can you find if the strings are ATGTTATA and ATCGTCC? What algorithm did you use?

Growing alignment	Remaining symbols	Score
	A T G T T A T A A T C G T C C	

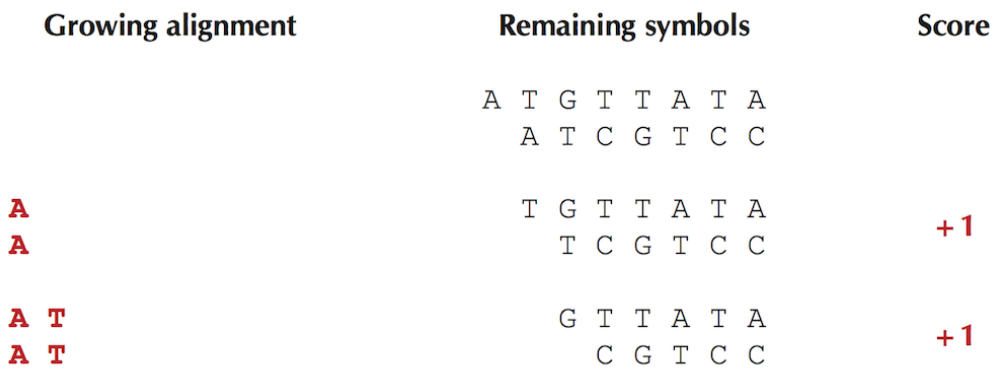
Matching Symbols as a Game

1. Remove the first symbol from each sequence, earn a point if the symbols match
2. Remove the first symbol from either of the two sequences, earn no points
3. Remove the first symbol from each sequence, earn no points if the symbols don’t match
4. Eventually try to maximize the number of points

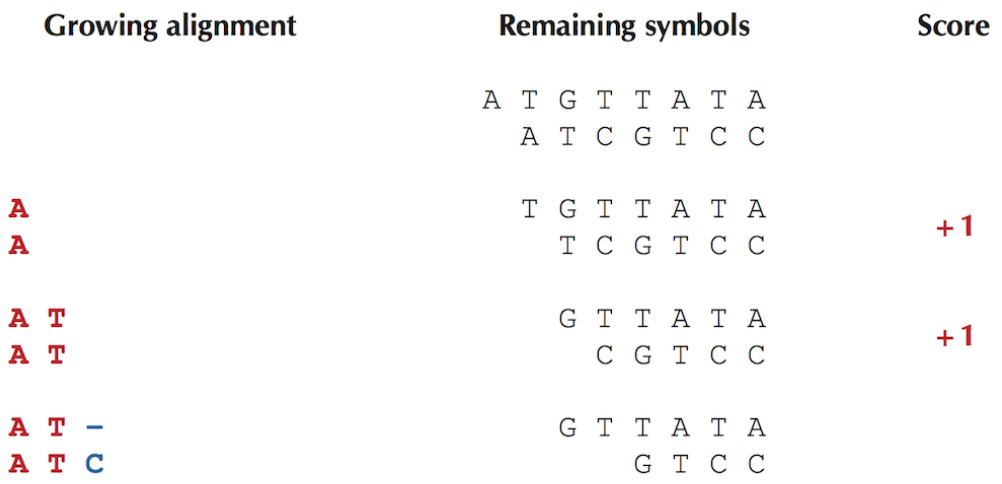
Matching Symbols as a Game



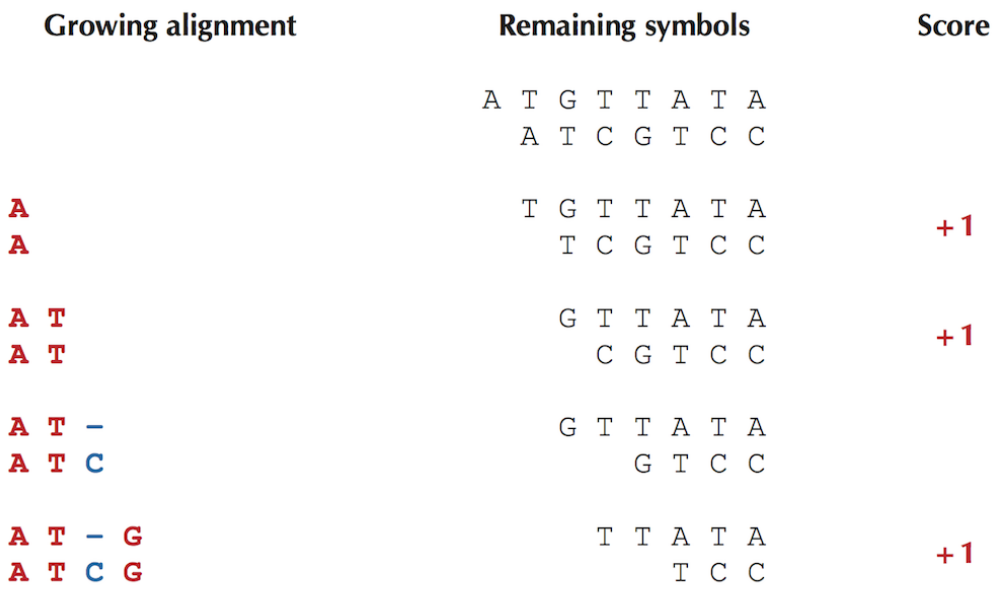
Matching Symbols as a Game



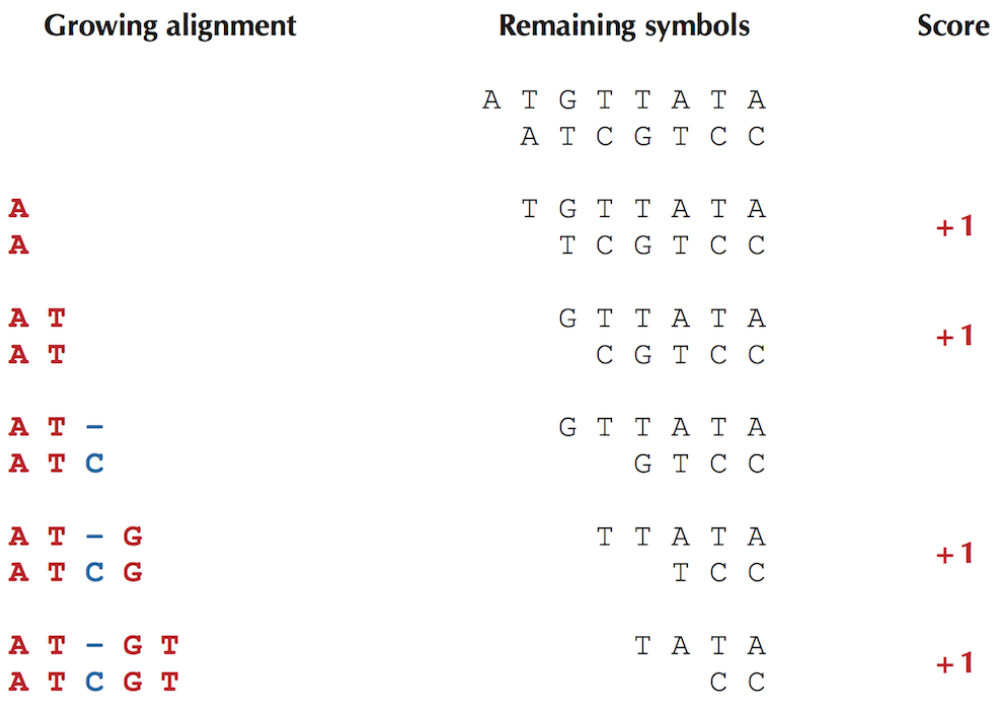
Matching Symbols as a Game



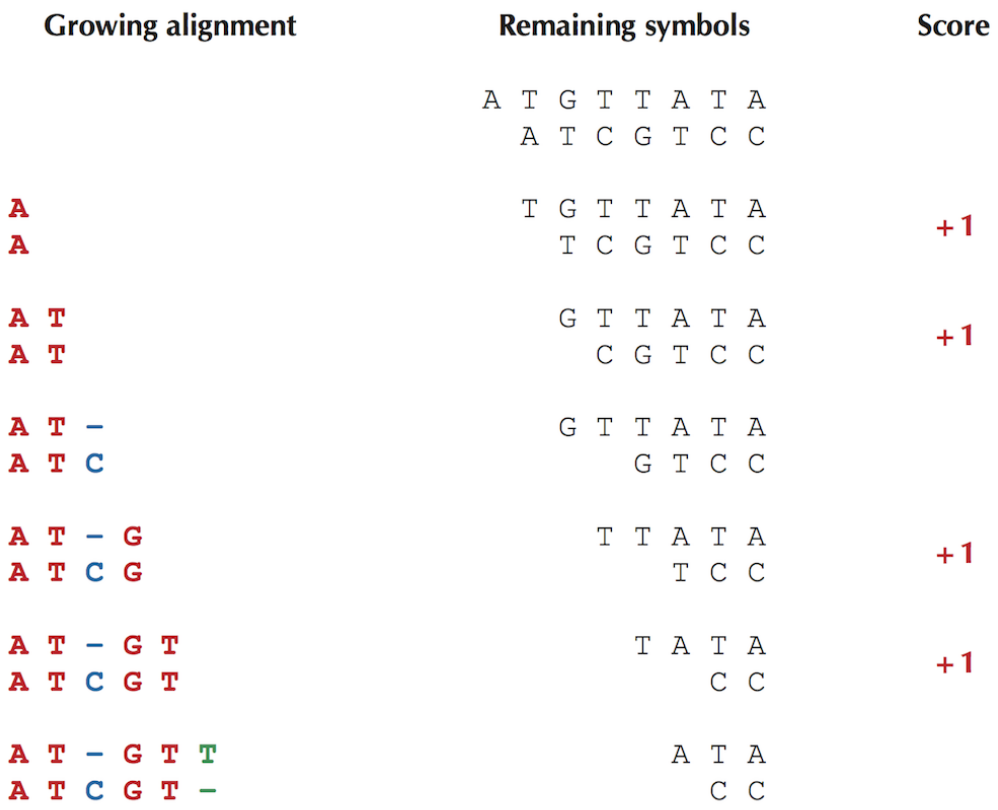
Matching Symbols as a Game



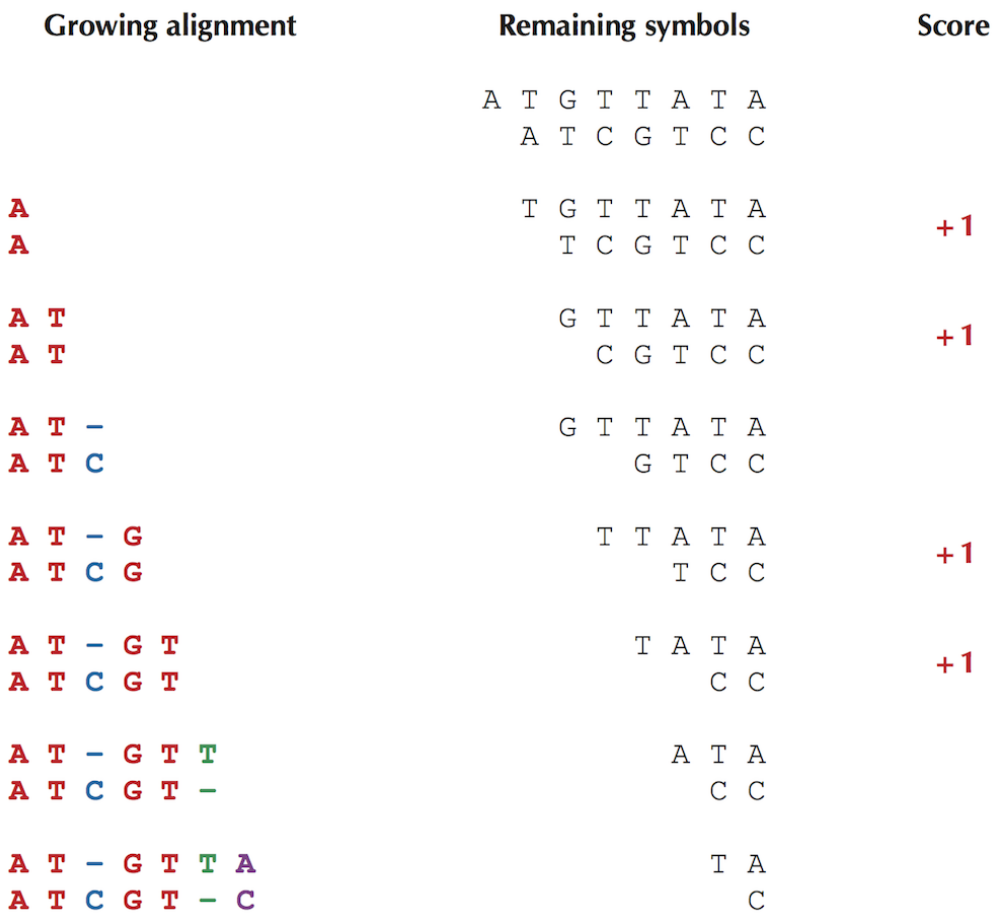
Matching Symbols as a Game



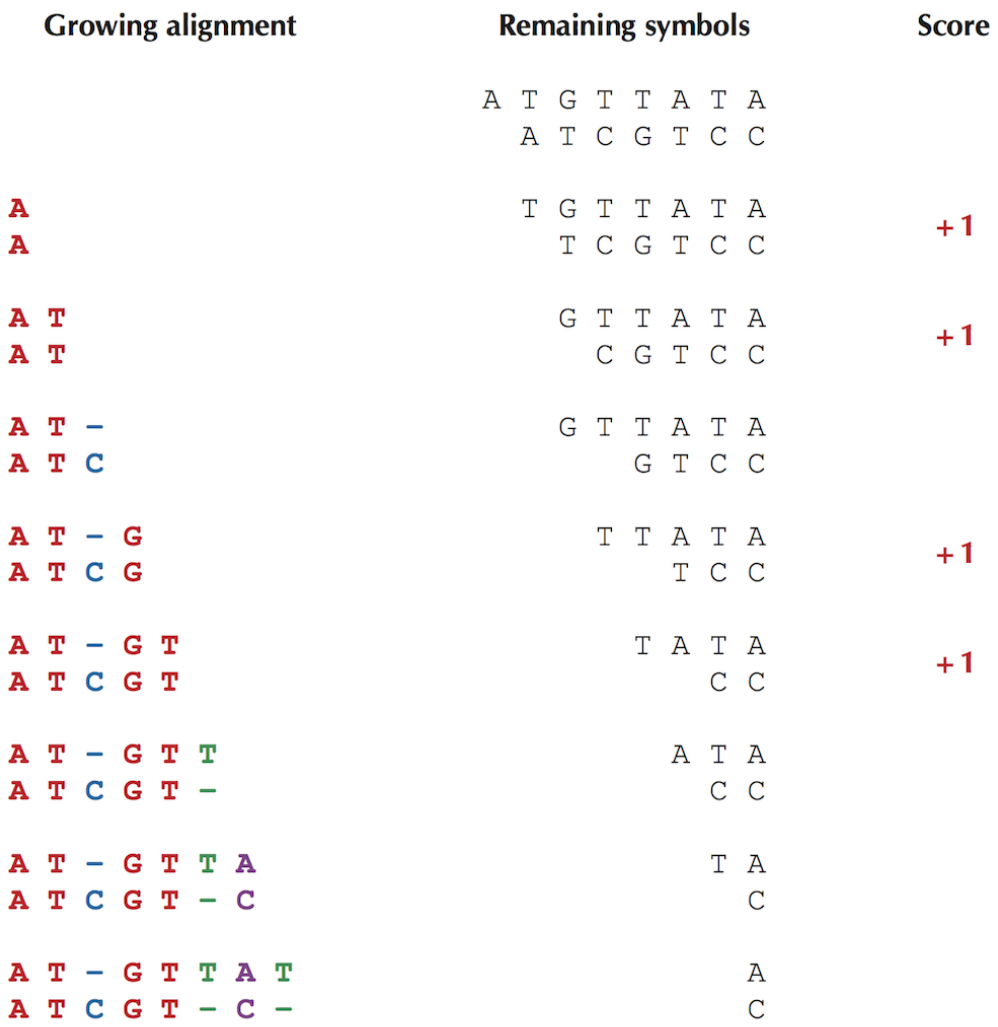
Matching Symbols as a Game



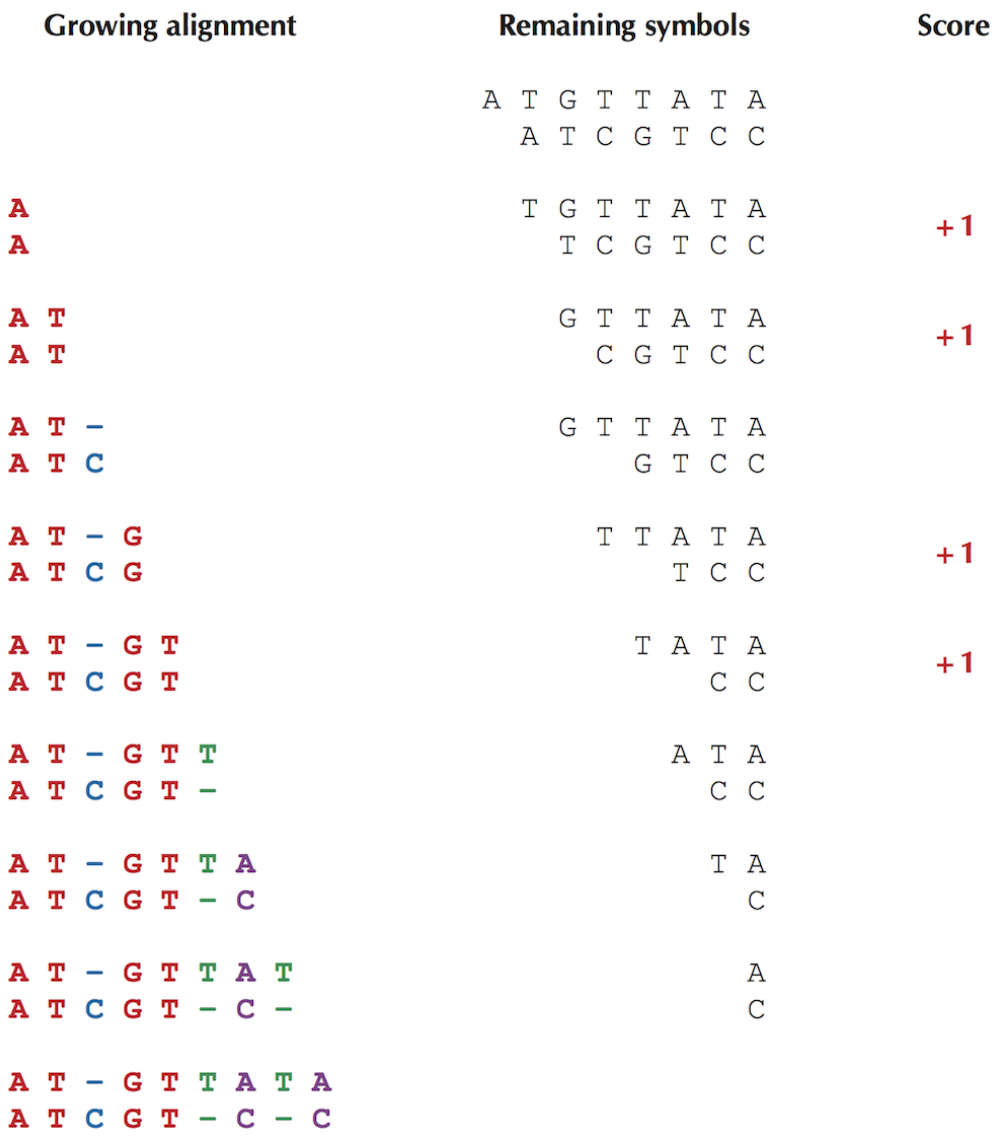
Matching Symbols as a Game



Matching Symbols as a Game



Matching Symbols as a Game



From a Game to a Definition

- Given two strings v and w, an alignment of v and w is a two-row matrix such that:
  - the first row contains symbols of v in order
  - the second row contains symbols of w in order
  - each row may also contain gap symbols (“-”)
  - no column has two gap symbols

```
AT-GTTATA
ATCGT-C-C
```

Definitions

V: AT-GTTATA
W: ATCGT-C-C

AT-GTTATA
ATCGT-C-C

Matches

Columns containing the same letter in both rows are called **matches** and represent conserved nucleotides

AT-GTTATA
ATCGT-C-C

Mismatches

Columns containing different letters are called **mismatches** and represent single-nucleotide substitutions

AT-GTTATA
ATCGT-C-C

Insertions

Column containing a space symbol in the top row of the alignment is called an **insertion**, as it implies the insertion of a symbol when transforming *v* into *w*

AT-GTTATA
ATCGT-C-C

Deletions

Column containing a space symbol in the bottom row of the alignment is called a **deletion**, as it indicates the deletion of a symbol when transforming *v* into *w*

Columns containing a space symbol are called **indels**

Finding a Longest Common Subsequence

- A common subsequence of v and w is a sequence of symbols occurring (not necessarily contiguously) in both v and w.
- The matches in an alignment of v and w form a common subsequence of v and w.

```
AT-GTTATA
ATCGT-C-C
```

The Problems are the Same!

- **Longest Common Subsequence Length Problem:**
  - **Input:** Two strings.
  - **Output:** The length of a longest common subsequence of these strings.
- **Symbol Matching Problem:**
  - **Input:** Two strings.
  - **Output:** The greatest number of matched symbols in any “alignment” of the two strings.