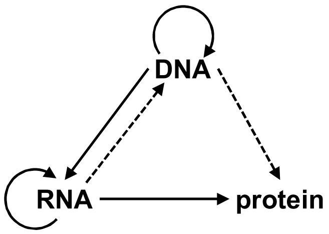
UCONN
UNIVERSITY OF CONNECTICUT

# Introduction to Computational Biology & Bioinformatics – Part II

ENGR 1166 Biomedical Engineering

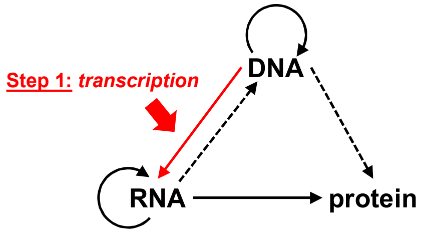
## Recap

- ❑ **Gene expression** is the process that produces proteins from DNA sequences



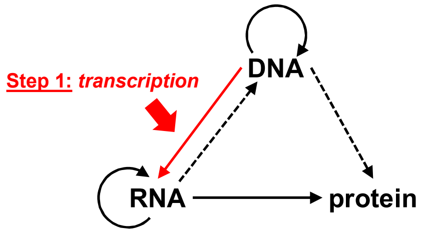
## Recap

- ❑ It consists of two steps
- ❑ First, a sequence of mRNA is generated from the DNA sequence



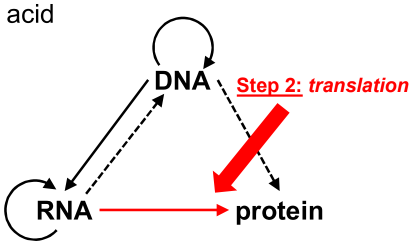
## Recap

□ The mRNA sequence is a **complement** of the DNA sequence, is read in the opposite direction, and replaces **T'**s with **U'**s



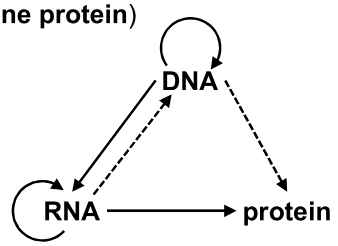
## Recap

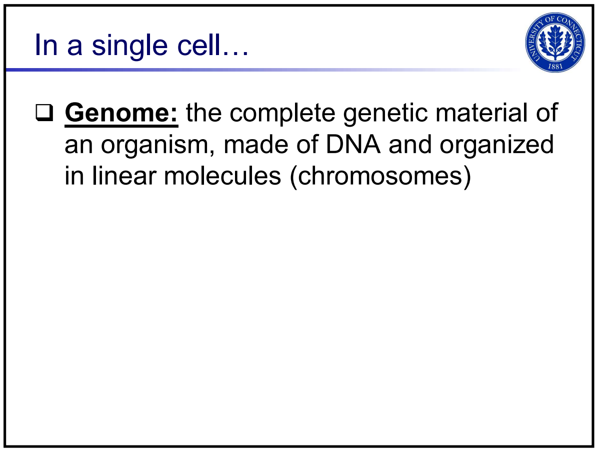
□ Second, the mRNA is divided in sub-sequences, each one including 3 bases, and every sub-sequence synthesizes one amino acid



## Recap

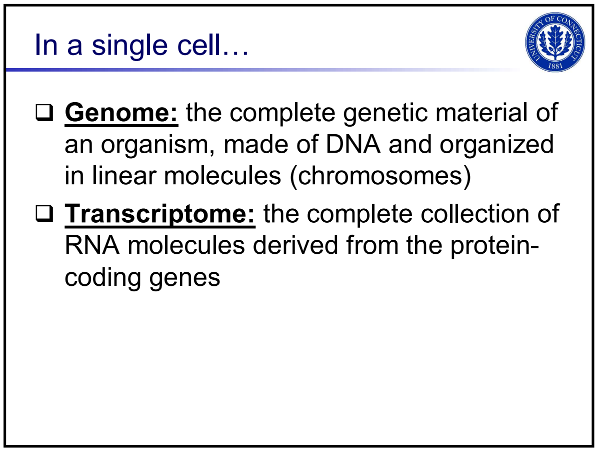
□ **Central Dogma:** the assumption that the genetic information flows from DNA to RNA to protein (i.e., **one gene** translates into **one protein**)





## In a single cell…

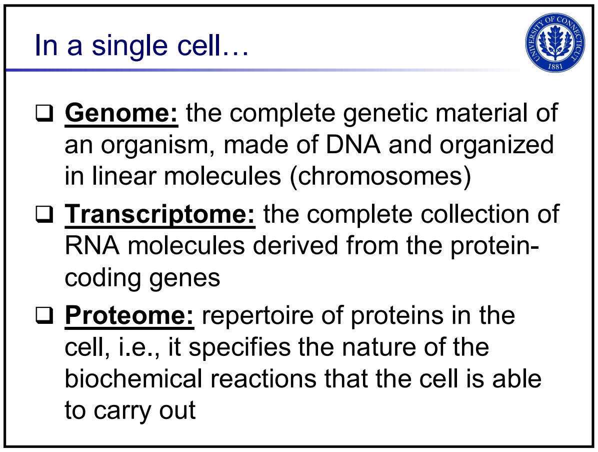
□ **Genome:** the complete genetic material of an organism, made of DNA and organized in linear molecules (chromosomes)



## In a single cell…

□ **Genome:** the complete genetic material of an organism, made of DNA and organized in linear molecules (chromosomes)

□ **Transcriptome:** the complete collection of RNA molecules derived from the protein-coding genes



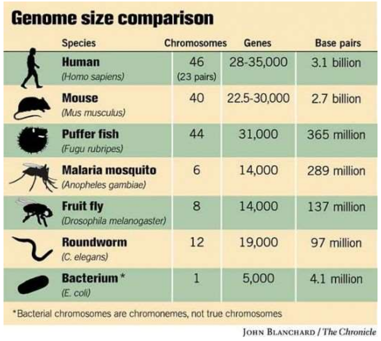
## In a single cell…

□ **Genome:** the complete genetic material of an organism, made of DNA and organized in linear molecules (chromosomes)

□ **Transcriptome:** the complete collection of RNA molecules derived from the protein-coding genes

□ **Proteome:** repertoire of proteins in the cell, i.e., it specifies the nature of the biochemical reactions that the cell is able to carry out

## Genomes vary dramatically in size



**Genome size comparison**

| Species | Chromosomes | Genes | Base pairs |
|---|---|---|---|
| **Human**<br>*(Homo sapiens)* | 46<br>(23 pairs) | 28-35,000 | 3.1 billion |
| **Mouse**<br>*(Mus musculus)* | 40 | 22.5-30,000 | 2.7 billion |
| **Puffer fish**<br>*(Fugu rubripes)* | 44 | 31,000 | 365 million |
| **Malaria mosquito**<br>*(Anopheles gambiae)* | 6 | 14,000 | 289 million |
| **Fruit fly**<br>*(Drosophila melanogaster)* | 8 | 14,000 | 137 million |
| **Roundworm**<br>*(C. elegans)* | 12 | 19,000 | 97 million |
| **Bacterium***<br>*(E. coli)* | 1 | 5,000 | 4.1 million |

*Bacterial chromosomes are chromonemes, not true chromosomes

JOHN BLANCHARD / *The Chronicle*

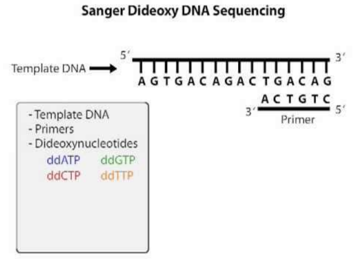
## DNA sequencing

- ❑ Biologists know how to access a DNA molecule but they need a way to precisely read the sequence of nucleotides (i.e., A, C, G, T) in it

## DNA sequencing

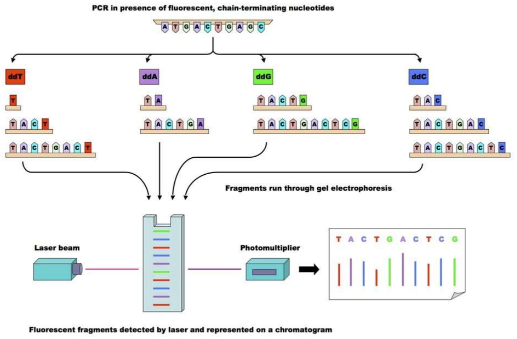
- ❑ Biologists know how to access a DNA molecule but they need a way to precisely read the sequence of nucleotides (i.e., A, C, G, T) in it
- ❑ **DNA sequencing** is the combination of methods and technologies used to read and store the sequence of nucleotides in an entire strand of DNA **<u>in the right order</u>**

## Method #1: Sanger sequencing

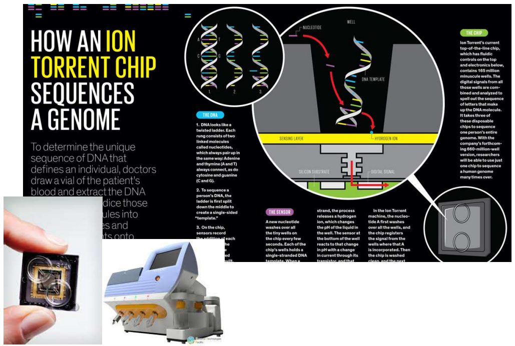


source: *http://www.youtube.com/watch?v=SRWvn1mUNMA*

## Method #1: Sanger sequencing


PCR in presence of fluorescent, chain-terminating nucleotides
A T G A C T G A G C
ddT
ddA
ddG
ddC
T
T A C T
T A C T G A C T
T A
T A C T G A
T A C T G
T A C T G A C T C G
T A C
T A C T G A C
T A C T G A C T C
Fragments run through gel electrophoresis
Laser beam
Photomultiplier
T A C T G A C T C G
Fluorescent fragments detected by laser and represented on a chromatogram

## Method #2: Ion-based sequencing



To determine the unique sequence of DNA that defines an individual, doctors draw a vial of the patient's blood and extract the DNA … dice those … ules into … es and … ts onto

THE DNA

1. DNA looks like a twisted ladder. Each rung consists of two linked molecules called nucleotides, which always pair up in the same way: Adenine and thymine (A and T) always connect, as do cytosine and guanine (C and G).

2. To sequence a person's DNA, the ladder is first split down the middle to create a single-sided "template."

3. On the chip, sensors record the addition of each …

THE SENSOR

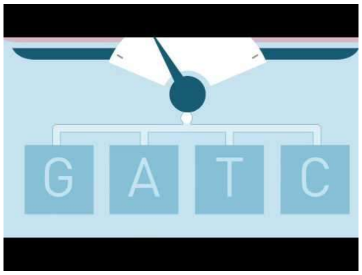
A new nucleotide washes over all the tiny wells on the chip every few seconds. Each of the chip's wells holds a single-stranded DNA template. When a … strand, the process releases a hydrogen ion, which changes the pH of the liquid in the well. The sensor at the bottom of the well reacts to that change in pH with a change in current through its transistor, and that …

In the Ion Torrent machine, the nucleotide A first washes over all the wells, and the chip registers the signal from the wells where that A is incorporated. Then the chip is washed clean, and the next …

THE CHIP

Ion Torrent's current top-of-the-line chip, which has fluidic controls on the top and electronics below, contains 165 million minuscule wells. The digital signals from all those wells are combined and analyzed to spell out the sequence of letters that make up the DNA molecule. It takes three of these disposable chips to sequence one person's entire genome. With the company's forthcoming 660-million-well version, researchers will be able to use just one chip to sequence a human genome many times over.

## Method #2: Ion-based sequencing



source: *http://www.youtube.com/watch?v=MxkYa9XCvBQ*

## Databases of biological data

Now that we can read a DNA strand, two questions occur:

□ **<u>Where</u>** do we store the outcomes of the DNA sequencing?

□ **<u>What</u>** do we do with the outcomes of the DNA sequencing?

## Databases of biological data

□ **<u>Where</u>** do we store the outcomes of the DNA sequencing?

## Databases of biological data

- ❑ **Where** do we store the outcomes of the DNA sequencing?

**Archival databases**

Online databases that provide access to repositories of DNA sequences, amino acid sequences, and protein 3-D structures

## Databases of biological data

- ❑ **Where** do we store the outcomes of the DNA sequencing?

**Archival databases**

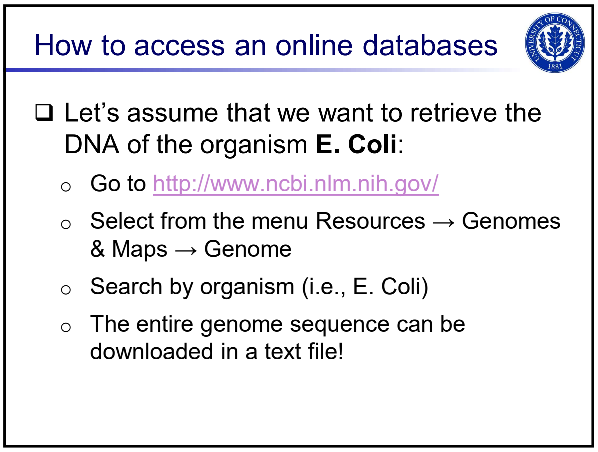
Online databases that provide access to repositories of DNA sequences, amino acid sequences, and protein 3-D structures

**Examples**

- ○ NCBI (National Center for Biotechnology Information): http://www.ncbi.nlm.nih.gov
- ○ EMBL (European Molecular Biology Laboratory): http://www.embl.org/
- ○ PDB (Protein Data Bank): http://www.rcsb.org/
- ○ Full list: http://en.wikipedia.org/wiki/List_of_biological_databases

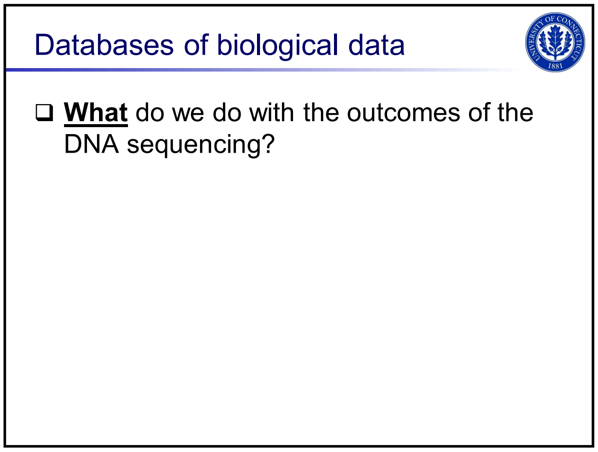
## How to access an online databases

- ❑ Let’s assume that we want to retrieve the 3D structure of the protein **hexokinase**:
  - ○ Go to http://www.rcsb.org/
  - ○ Search by molecule name (i.e., hexokinase)
  - ○ Select the structure from the organism in which you are interested
  - ○ View the 3D structure, download the atomic coordinates, etc.



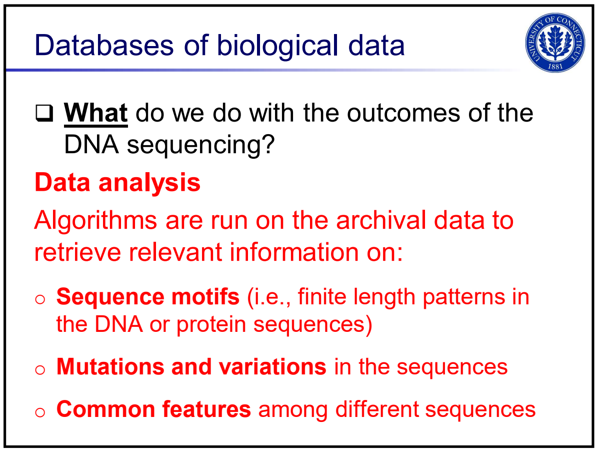
## How to access an online databases

- ❑ Let’s assume that we want to retrieve the DNA of the organism **E. Coli**:
  - Go to http://www.ncbi.nlm.nih.gov/
  - Select from the menu Resources → Genomes & Maps → Genome
  - Search by organism (i.e., E. Coli)
  - The entire genome sequence can be downloaded in a text file!



## Databases of biological data

- ❑ **What** do we do with the outcomes of the DNA sequencing?



## Databases of biological data

- ❑ **What** do we do with the outcomes of the DNA sequencing?

**Data analysis**

Algorithms are run on the archival data to retrieve relevant information on:

- **Sequence motifs** (i.e., finite length patterns in the DNA or protein sequences)
- **Mutations and variations** in the sequences
- **Common features** among different sequences

## Databases of biological data

- Sometimes the results of the data analysis need to be stored, i.e., **derived databases** are created
- Both archival and derived databases must be well-structured and organized to allow for user-friendly searches and multiple types of queries

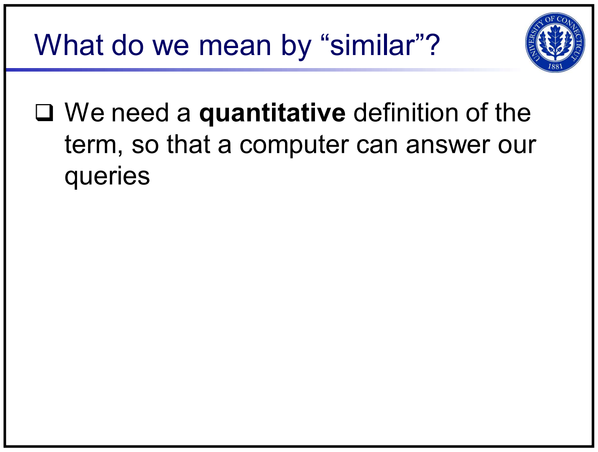
## Examples of queries

- Given a DNA or protein sequence $\boldsymbol{S}^*$, which sequences in the database are **similar** to $\boldsymbol{S}^*$?
- Given a protein 3D structure $\boldsymbol{X}^*$, which other proteins in the database have structure **similar** to $\boldsymbol{X}^*$?

## Examples of queries

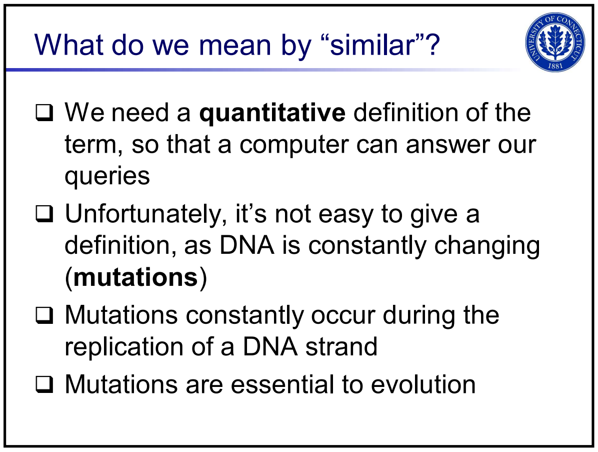
For instance, these queries are relevant if:

- *We have sequences from two different species and we want to know who is the last common ancestor*
- *We want to identify regions in a sequence that have been conserved (unchanged) throughout evolution*
- *We want to know what kind of structural and functional properties a certain protein has*



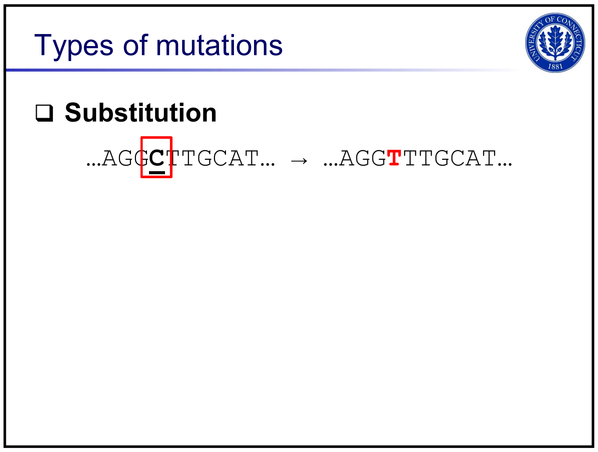
## What do we mean by "similar"?

- ❑ We need a **quantitative** definition of the term, so that a computer can answer our queries



## What do we mean by "similar"?

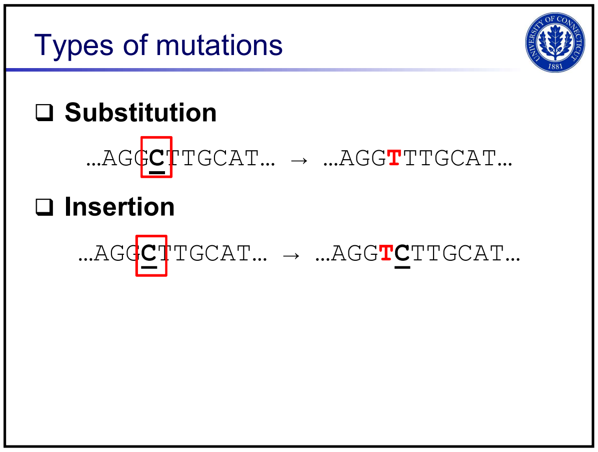
- ❑ We need a **quantitative** definition of the term, so that a computer can answer our queries
- ❑ Unfortunately, it's not easy to give a definition, as DNA is constantly changing (**mutations**)
- ❑ Mutations constantly occur during the replication of a DNA strand
- ❑ Mutations are essential to evolution



## Types of mutations

- ❑ **Substitution**

…AGG**C**TTGCAT… → …AGG**T**TTGCAT…



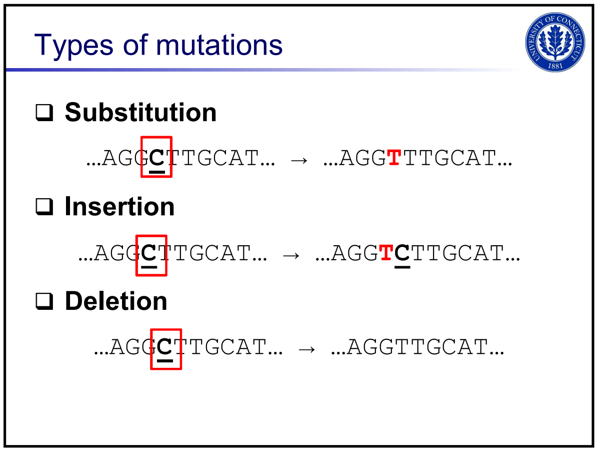
## Types of mutations

- ❑ **Substitution**

  …AGG**C**TTGCAT… → …AGG**T**TTGCAT…

- ❑ **Insertion**

  …AGG**C**TTGCAT… → …AGG**T**C̲TTGCAT…



## Types of mutations

- ❑ **Substitution**

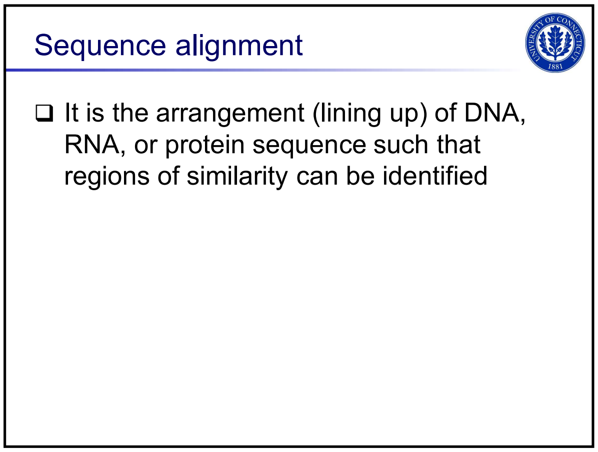
  …AGG**C**TTGCAT… → …AGG**T**TTGCAT…

- ❑ **Insertion**

  …AGG**C**TTGCAT… → …AGG**T**C̲TTGCAT…

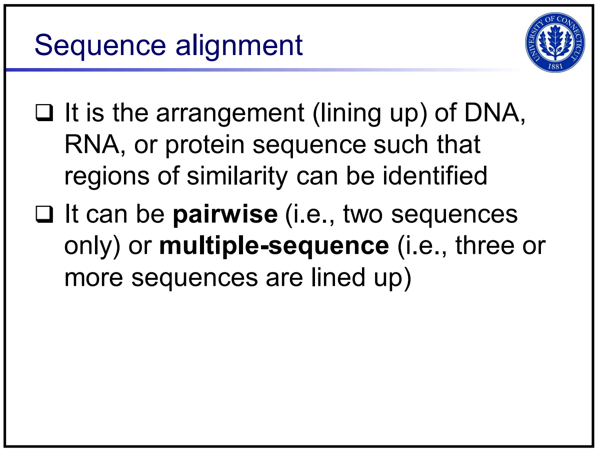
- ❑ **Deletion**

  …AGG**C**TTGCAT… → …AGGTTGCAT…



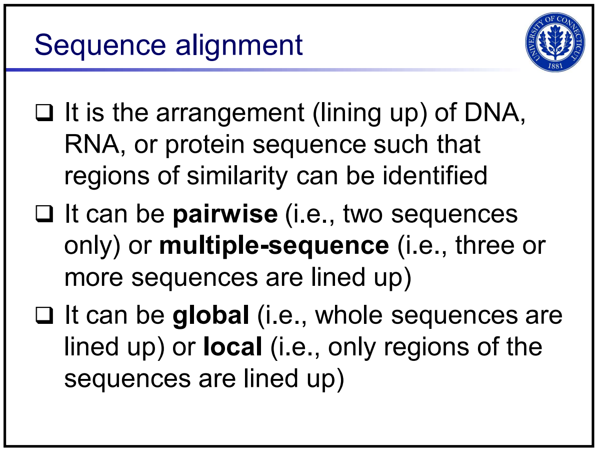
## Sequence alignment

- ❑ It is the arrangement (lining up) of DNA, RNA, or protein sequence such that regions of similarity can be identified



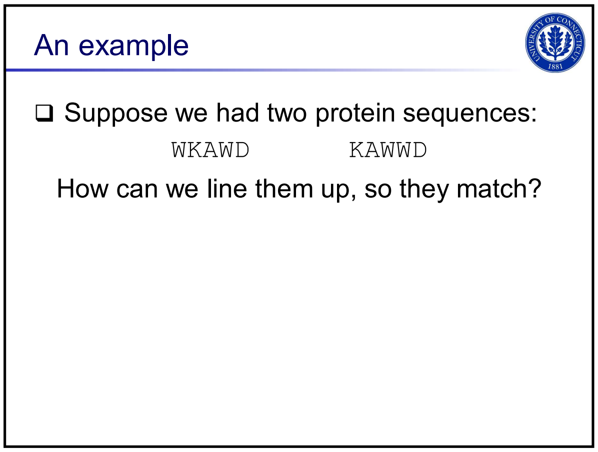
## Sequence alignment

- ❑ It is the arrangement (lining up) of DNA, RNA, or protein sequence such that regions of similarity can be identified
- ❑ It can be **pairwise** (i.e., two sequences only) or **multiple-sequence** (i.e., three or more sequences are lined up)



## Sequence alignment

- ❑ It is the arrangement (lining up) of DNA, RNA, or protein sequence such that regions of similarity can be identified
- ❑ It can be **pairwise** (i.e., two sequences only) or **multiple-sequence** (i.e., three or more sequences are lined up)
- ❑ It can be **global** (i.e., whole sequences are lined up) or **local** (i.e., only regions of the sequences are lined up)

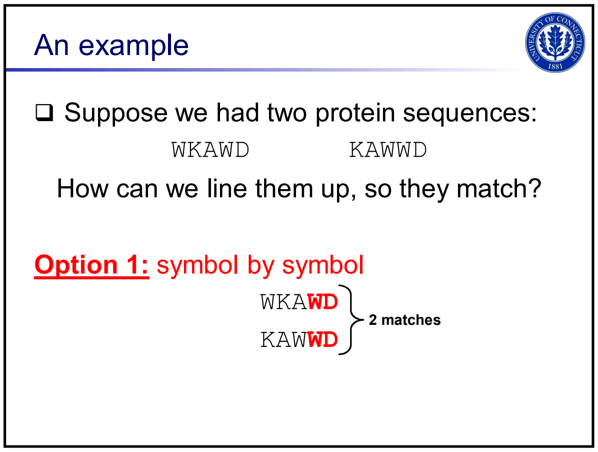


## An example

- ❑ Suppose we had two protein sequences:

```
WKAWD        KAWWD
```

How can we line them up, so they match?



## An example

- ❑ Suppose we had two protein sequences:

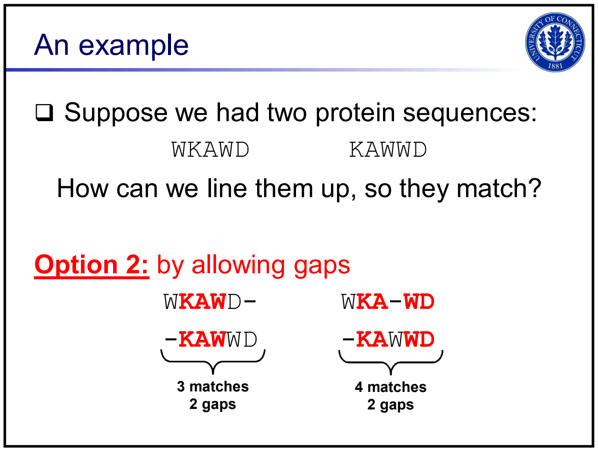
  WKAWD KAWWD

  How can we line them up, so they match?

**Option 1:** symbol by symbol

WKA**WD**<br>
KAW**WD**

2 matches



## An example

- ❑ Suppose we had two protein sequences:

  WKAWD KAWWD

  How can we line them up, so they match?

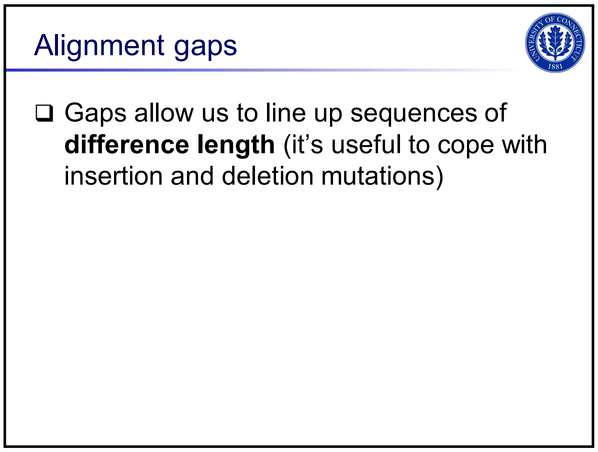
**Option 2:** by allowing gaps

W**KAW**D-<br>
-**KAW**WD

3 matches<br>
2 gaps

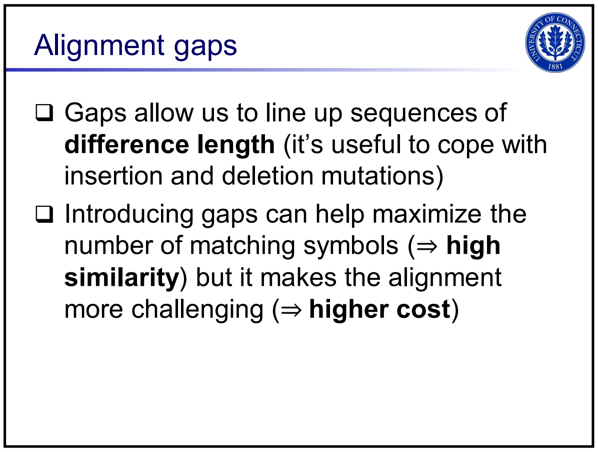
W**KA**-**WD**<br>
-**KA**W**WD**

4 matches<br>
2 gaps



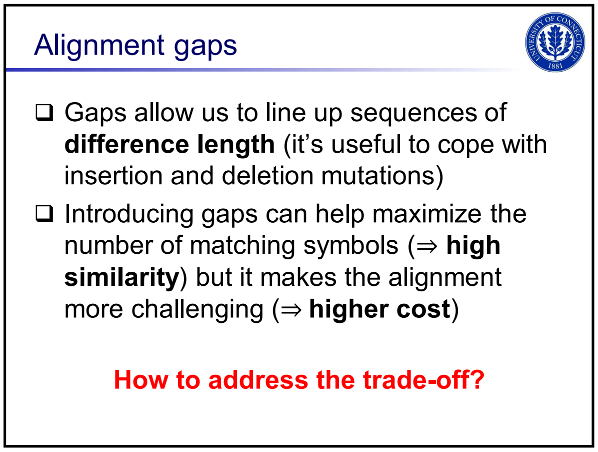
## Alignment gaps

- ❑ Gaps allow us to line up sequences of **difference length** (it's useful to cope with insertion and deletion mutations)



## Alignment gaps

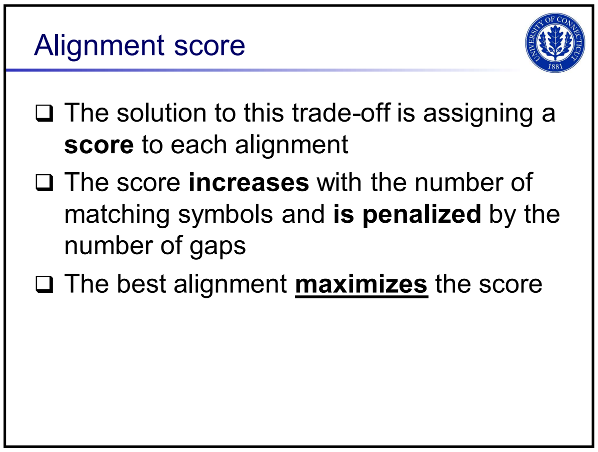
- Gaps allow us to line up sequences of **difference length** (it’s useful to cope with insertion and deletion mutations)
- Introducing gaps can help maximize the number of matching symbols (⇒ **high similarity**) but it makes the alignment more challenging (⇒ **higher cost**)



## Alignment gaps

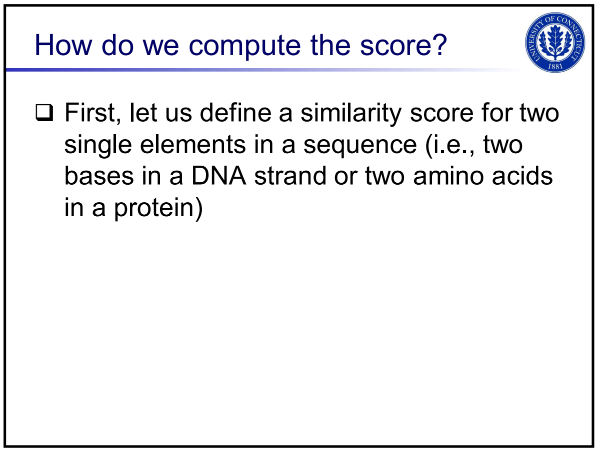
- Gaps allow us to line up sequences of **difference length** (it’s useful to cope with insertion and deletion mutations)
- Introducing gaps can help maximize the number of matching symbols (⇒ **high similarity**) but it makes the alignment more challenging (⇒ **higher cost**)

**How to address the trade-off?**



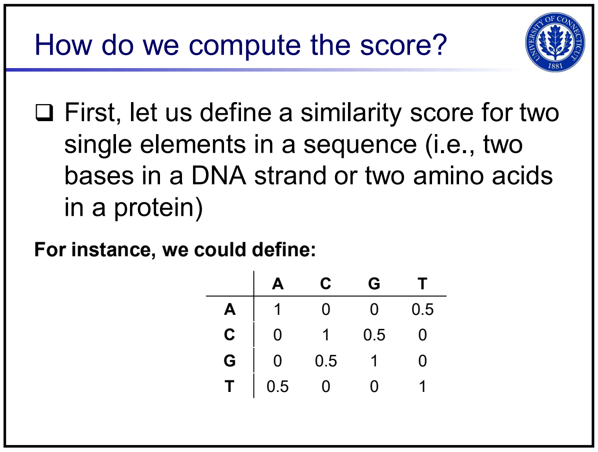
## Alignment score

- The solution to this trade-off is assigning a **score** to each alignment
- The score **increases** with the number of matching symbols and **is penalized** by the number of gaps
- The best alignment **maximizes** the score



## How do we compute the score?

- ❑ First, let us define a similarity score for two single elements in a sequence (i.e., two bases in a DNA strand or two amino acids in a protein)

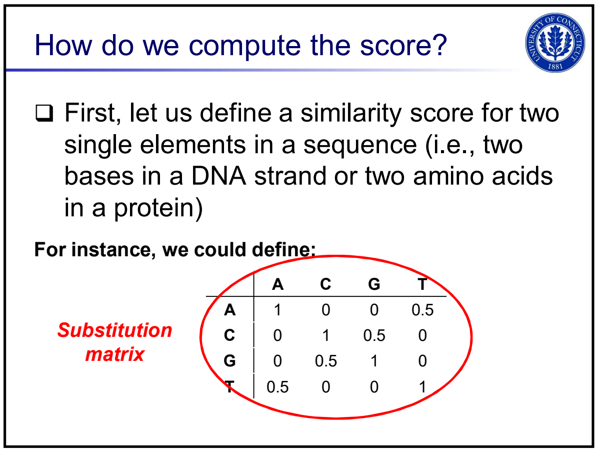


## How do we compute the score?

- ❑ First, let us define a similarity score for two single elements in a sequence (i.e., two bases in a DNA strand or two amino acids in a protein)

**For instance, we could define:**

| | A | C | G | T |
|---|---|---|---|---|
| **A** | 1 | 0 | 0 | 0.5 |
| **C** | 0 | 1 | 0.5 | 0 |
| **G** | 0 | 0.5 | 1 | 0 |
| **T** | 0.5 | 0 | 0 | 1 |



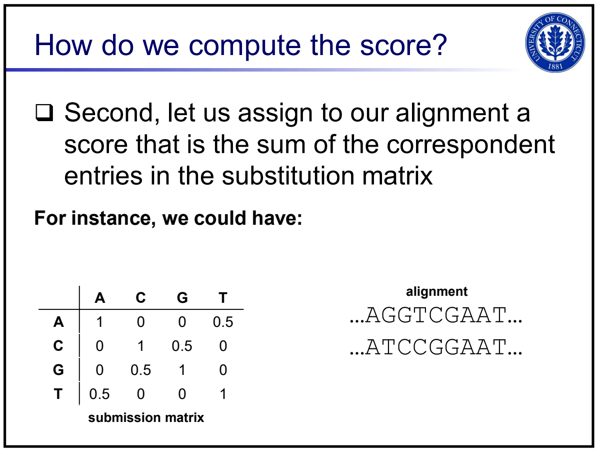
## How do we compute the score?

- ❑ First, let us define a similarity score for two single elements in a sequence (i.e., two bases in a DNA strand or two amino acids in a protein)

**For instance, we could define:**

***Substitution matrix***

| | A | C | G | T |
|---|---|---|---|---|
| **A** | 1 | 0 | 0 | 0.5 |
| **C** | 0 | 1 | 0.5 | 0 |
| **G** | 0 | 0.5 | 1 | 0 |
| **T** | 0.5 | 0 | 0 | 1 |



## How do we compute the score?

- ❑ Second, let us assign to our alignment a score that is the sum of the correspondent entries in the substitution matrix

**For instance, we could have:**

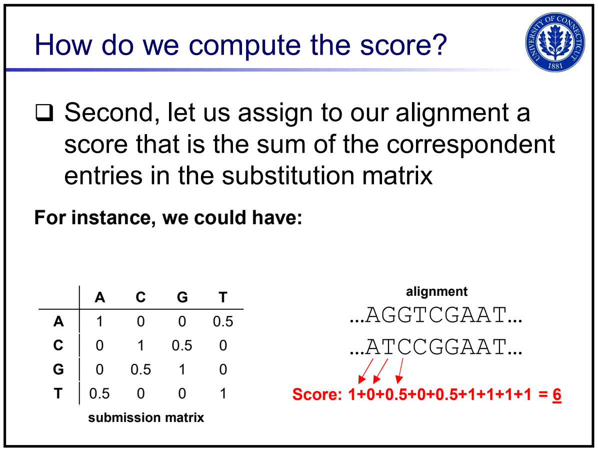
| | A | C | G | T |
|---|---|---|---|---|
| **A** | 1 | 0 | 0 | 0.5 |
| **C** | 0 | 1 | 0.5 | 0 |
| **G** | 0 | 0.5 | 1 | 0 |
| **T** | 0.5 | 0 | 0 | 1 |

**submission matrix**

**alignment**

…AGGTCGAAT…

…ATCCGGAAT…



## How do we compute the score?

- ❑ Second, let us assign to our alignment a score that is the sum of the correspondent entries in the substitution matrix

**For instance, we could have:**

| | A | C | G | T |
|---|---|---|---|---|
| **A** | 1 | 0 | 0 | 0.5 |
| **C** | 0 | 1 | 0.5 | 0 |
| **G** | 0 | 0.5 | 1 | 0 |
| **T** | 0.5 | 0 | 0 | 1 |

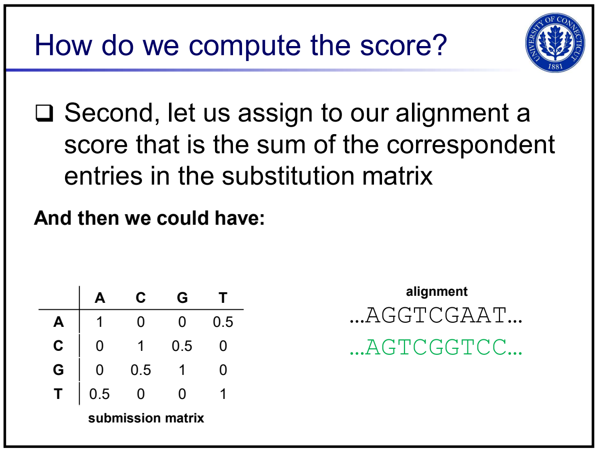
**submission matrix**

**alignment**

…AGGTCGAAT…

…ATCCGGAAT…

**Score: 1+0+0.5+0+0.5+1+1+1+1 = 6**



## How do we compute the score?

- ❑ Second, let us assign to our alignment a score that is the sum of the correspondent entries in the substitution matrix

**And then we could have:**

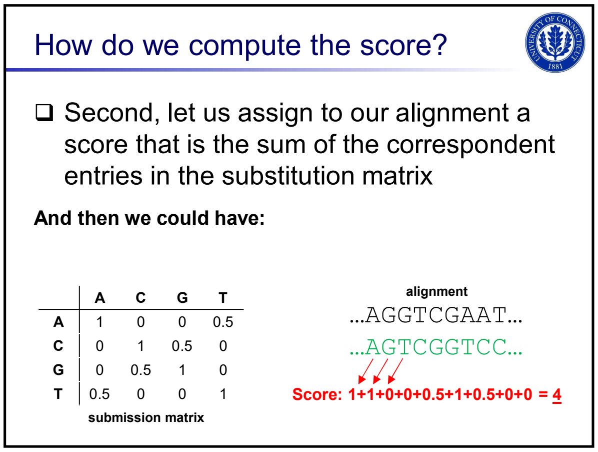
| | A | C | G | T |
|---|---|---|---|---|
| **A** | 1 | 0 | 0 | 0.5 |
| **C** | 0 | 1 | 0.5 | 0 |
| **G** | 0 | 0.5 | 1 | 0 |
| **T** | 0.5 | 0 | 0 | 1 |

**submission matrix**

**alignment**

…AGGTCGAAT…

…AGTCGGTCC…



## How do we compute the score?

❑ Second, let us assign to our alignment a score that is the sum of the correspondent entries in the substitution matrix

**And then we could have:**

| | A | C | G | T |
|---|---|---|---|---|
| **A** | 1 | 0 | 0 | 0.5 |
| **C** | 0 | 1 | 0.5 | 0 |
| **G** | 0 | 0.5 | 1 | 0 |
| **T** | 0.5 | 0 | 0 | 1 |

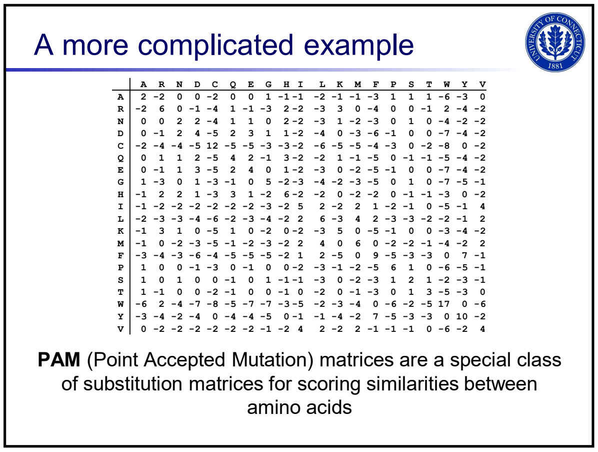
**submission matrix**

**alignment**

…AGGTCGAAT…

…AGTCGGTCC…

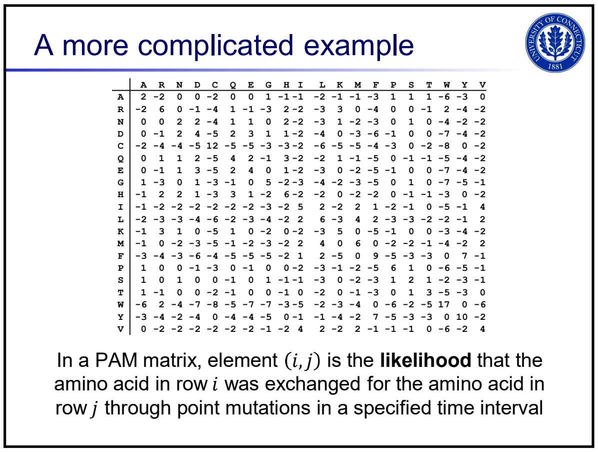
**Score: 1+1+0+0+0.5+1+0.5+0+0 = 4**



## A more complicated example

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 2 | -2 | 0 | 0 | -2 | 0 | 0 | 1 | -1 | -1 | -2 | -1 | -1 | -3 | 1 | 1 | 1 | -6 | -3 | 0 |
| R | -2 | 6 | 0 | -1 | -4 | 1 | -1 | -3 | 2 | -2 | -3 | 3 | 0 | -4 | 0 | 0 | -1 | 2 | -4 | -2 |
| N | 0 | 0 | 2 | 2 | -4 | 1 | 1 | 0 | 2 | -2 | -3 | 1 | -2 | -3 | 0 | 1 | 0 | -4 | -2 | -2 |
| D | 0 | -1 | 2 | 4 | -5 | 2 | 3 | 1 | 1 | -2 | -4 | 0 | -3 | -6 | -1 | 0 | 0 | -7 | -4 | -2 |
| C | -2 | -4 | -4 | -5 | 12 | -5 | -5 | -3 | -3 | -2 | -6 | -5 | -5 | -4 | -3 | 0 | -2 | -8 | 0 | -2 |
| Q | 0 | 1 | 1 | 2 | -5 | 4 | 2 | -1 | 3 | -2 | -2 | 1 | -1 | -5 | 0 | -1 | -1 | -5 | -4 | -2 |
| E | 0 | -1 | 1 | 3 | -5 | 2 | 4 | 0 | 1 | -2 | -3 | 0 | -2 | -5 | -1 | 0 | 0 | -7 | -4 | -2 |
| G | 1 | -3 | 0 | 1 | -3 | -1 | 0 | 5 | -2 | -3 | -4 | -2 | -3 | -5 | 0 | 1 | 0 | -7 | -5 | -1 |
| H | -1 | 2 | 2 | 1 | -3 | 3 | 1 | -2 | 6 | -2 | -2 | 0 | -2 | -2 | 0 | -1 | -1 | -3 | 0 | -2 |
| I | -1 | -2 | -2 | -2 | -2 | -2 | -2 | -3 | -2 | 5 | 2 | -2 | 2 | 1 | -2 | -1 | 0 | -5 | -1 | 4 |
| L | -2 | -3 | -3 | -4 | -6 | -2 | -3 | -4 | -2 | 2 | 6 | -3 | 4 | 2 | -3 | -3 | -2 | -2 | -1 | 2 |
| K | -1 | 3 | 1 | 0 | -5 | 1 | 0 | -2 | 0 | -2 | -3 | 5 | 0 | -5 | -1 | 0 | 0 | -3 | -4 | -2 |
| M | -1 | 0 | -2 | -3 | -5 | -1 | -2 | -3 | -2 | 2 | 4 | 0 | 6 | 0 | -2 | -2 | -1 | -4 | -2 | 2 |
| F | -3 | -4 | -3 | -6 | -4 | -5 | -5 | -5 | -2 | 1 | 2 | -5 | 0 | 9 | -5 | -3 | -3 | 0 | 7 | -1 |
| P | 1 | 0 | 0 | -1 | -3 | 0 | -1 | 0 | 0 | -2 | -3 | -1 | -2 | -5 | 6 | 1 | 0 | -6 | -5 | -1 |
| S | 1 | 0 | 1 | 0 | 0 | -1 | 0 | 1 | -1 | -1 | -3 | 0 | -2 | -3 | 1 | 2 | 1 | -2 | -3 | -1 |
| T | 1 | -1 | 0 | 0 | -2 | -1 | 0 | 0 | -1 | 0 | -2 | 0 | -1 | -3 | 0 | 1 | 3 | -5 | -3 | 0 |
| W | -6 | 2 | -4 | -7 | -8 | -5 | -7 | -7 | -3 | -5 | -2 | -3 | -4 | 0 | -6 | -2 | -5 | 17 | 0 | -6 |
| Y | -3 | -4 | -2 | -4 | 0 | -4 | -4 | -5 | 0 | -1 | -1 | -4 | -2 | 7 | -5 | -3 | -3 | 0 | 10 | -2 |
| V | 0 | -2 | -2 | -2 | -2 | -2 | -2 | -1 | -2 | 4 | 2 | -2 | 2 | -1 | -1 | -1 | 0 | -6 | -2 | 4 |

**PAM** (Point Accepted Mutation) matrices are a special class of substitution matrices for scoring similarities between amino acids



## A more complicated example

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 2 | -2 | 0 | 0 | -2 | 0 | 0 | 1 | -1 | -1 | -2 | -1 | -1 | -3 | 1 | 1 | 1 | -6 | -3 | 0 |
| R | -2 | 6 | 0 | -1 | -4 | 1 | -1 | -3 | 2 | -2 | -3 | 3 | 0 | -4 | 0 | 0 | -1 | 2 | -4 | -2 |
| N | 0 | 0 | 2 | 2 | -4 | 1 | 1 | 0 | 2 | -2 | -3 | 1 | -2 | -3 | 0 | 1 | 0 | -4 | -2 | -2 |
| D | 0 | -1 | 2 | 4 | -5 | 2 | 3 | 1 | 1 | -2 | -4 | 0 | -3 | -6 | -1 | 0 | 0 | -7 | -4 | -2 |
| C | -2 | -4 | -4 | -5 | 12 | -5 | -5 | -3 | -3 | -2 | -6 | -5 | -5 | -4 | -3 | 0 | -2 | -8 | 0 | -2 |
| Q | 0 | 1 | 1 | 2 | -5 | 4 | 2 | -1 | 3 | -2 | -2 | 1 | -1 | -5 | 0 | -1 | -1 | -5 | -4 | -2 |
| E | 0 | -1 | 1 | 3 | -5 | 2 | 4 | 0 | 1 | -2 | -3 | 0 | -2 | -5 | -1 | 0 | 0 | -7 | -4 | -2 |
| G | 1 | -3 | 0 | 1 | -3 | -1 | 0 | 5 | -2 | -3 | -4 | -2 | -3 | -5 | 0 | 1 | 0 | -7 | -5 | -1 |
| H | -1 | 2 | 2 | 1 | -3 | 3 | 1 | -2 | 6 | -2 | -2 | 0 | -2 | -2 | 0 | -1 | -1 | -3 | 0 | -2 |
| I | -1 | -2 | -2 | -2 | -2 | -2 | -2 | -3 | -2 | 5 | 2 | -2 | 2 | 1 | -2 | -1 | 0 | -5 | -1 | 4 |
| L | -2 | -3 | -3 | -4 | -6 | -2 | -3 | -4 | -2 | 2 | 6 | -3 | 4 | 2 | -3 | -3 | -2 | -2 | -1 | 2 |
| K | -1 | 3 | 1 | 0 | -5 | 1 | 0 | -2 | 0 | -2 | -3 | 5 | 0 | -5 | -1 | 0 | 0 | -3 | -4 | -2 |
| M | -1 | 0 | -2 | -3 | -5 | -1 | -2 | -3 | -2 | 2 | 4 | 0 | 6 | 0 | -2 | -2 | -1 | -4 | -2 | 2 |
| F | -3 | -4 | -3 | -6 | -4 | -5 | -5 | -5 | -2 | 1 | 2 | -5 | 0 | 9 | -5 | -3 | -3 | 0 | 7 | -1 |
| P | 1 | 0 | 0 | -1 | -3 | 0 | -1 | 0 | 0 | -2 | -3 | -1 | -2 | -5 | 6 | 1 | 0 | -6 | -5 | -1 |
| S | 1 | 0 | 1 | 0 | 0 | -1 | 0 | 1 | -1 | -1 | -3 | 0 | -2 | -3 | 1 | 2 | 1 | -2 | -3 | -1 |
| T | 1 | -1 | 0 | 0 | -2 | -1 | 0 | 0 | -1 | 0 | -2 | 0 | -1 | -3 | 0 | 1 | 3 | -5 | -3 | 0 |
| W | -6 | 2 | -4 | -7 | -8 | -5 | -7 | -7 | -3 | -5 | -2 | -3 | -4 | 0 | -6 | -2 | -5 | 17 | 0 | -6 |
| Y | -3 | -4 | -2 | -4 | 0 | -4 | -4 | -5 | 0 | -1 | -1 | -4 | -2 | 7 | -5 | -3 | -3 | 0 | 10 | -2 |
| V | 0 | -2 | -2 | -2 | -2 | -2 | -2 | -1 | -2 | 4 | 2 | -2 | 2 | -1 | -1 | -1 | 0 | -6 | -2 | 4 |

In a PAM matrix, element $(i, j)$ is the **likelihood** that the amino acid in row $i$ was exchanged for the amino acid in row $j$ through point mutations in a specified time interval

## A more complicated example

PAM-250

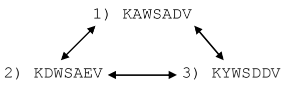
| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 2 | -2 | 0 | 0 | -2 | 0 | 0 | 1 | -1 | -1 | -2 | -1 | -1 | -3 | 1 | 1 | 1 | -6 | -3 | 0 |
| R | -2 | 6 | 0 | -1 | -4 | 1 | -1 | -3 | 2 | -2 | -3 | 3 | 0 | -4 | 0 | 0 | -1 | 2 | -4 | -2 |
| N | 0 | 0 | 2 | 2 | -4 | 1 | 1 | 0 | 2 | -2 | -3 | 1 | -2 | -3 | 0 | 1 | 0 | -4 | -2 | -2 |
| D | 0 | -1 | 2 | 4 | -5 | 2 | 3 | 1 | 1 | -2 | -4 | 0 | -3 | -6 | -1 | 0 | 0 | -7 | -4 | -2 |
| C | -2 | -4 | -4 | -5 | 12 | -5 | -5 | -3 | -3 | -2 | -6 | -5 | -5 | -4 | -3 | 0 | -2 | -8 | 0 | -2 |
| Q | 0 | 1 | 1 | 2 | -5 | 4 | 2 | -1 | 3 | -2 | -2 | 1 | -1 | -5 | 0 | -1 | -1 | -5 | -4 | -2 |
| E | 0 | -1 | 1 | 3 | -5 | 2 | 4 | 0 | 1 | -2 | -3 | 0 | -2 | -5 | -1 | 0 | 0 | -7 | -4 | -2 |
| G | 1 | -3 | 0 | 1 | -3 | -1 | 0 | 5 | -2 | -3 | -4 | -2 | -3 | -5 | 0 | 1 | 0 | -7 | -5 | -1 |
| H | -1 | 2 | 2 | 1 | -3 | 3 | 1 | -2 | 6 | -2 | -2 | 0 | -2 | -2 | 0 | -1 | -1 | -3 | 0 | -2 |
| I | -1 | -2 | -2 | -2 | -2 | -2 | -2 | -3 | -2 | 5 | 2 | -2 | 2 | 1 | -2 | -1 | 0 | -5 | -1 | 4 |
| L | -2 | -3 | -3 | -4 | -6 | -2 | -3 | -4 | -2 | 2 | 6 | -3 | 4 | 2 | -3 | -3 | -2 | -2 | -1 | 2 |
| K | -1 | 3 | 1 | 0 | -5 | 1 | 0 | -2 | 0 | -2 | -3 | 5 | 0 | -5 | -1 | 0 | 0 | -3 | -4 | -2 |
| M | -1 | 0 | -2 | -3 | -5 | -1 | -2 | -3 | -2 | 2 | 4 | 0 | 6 | 0 | -2 | -2 | -1 | -4 | -2 | 2 |
| F | -3 | -4 | -3 | -6 | -4 | -5 | -5 | -5 | -2 | 1 | 2 | -5 | 0 | 9 | -5 | -3 | -3 | 0 | 7 | -1 |
| P | 1 | 0 | 0 | -1 | -3 | 0 | -1 | 0 | 0 | -2 | -3 | -1 | -2 | -5 | 6 | 1 | 0 | -6 | -5 | -1 |
| S | 1 | 0 | 1 | 0 | 0 | -1 | 0 | 1 | -1 | -1 | -3 | 0 | -2 | -3 | 1 | 2 | 1 | -2 | -3 | -1 |
| T | 1 | -1 | 0 | 0 | -2 | -1 | 0 | 0 | -1 | 0 | -2 | 0 | -1 | -3 | 0 | 1 | 3 | -5 | -3 | 0 |
| W | -6 | 2 | -4 | -7 | -8 | -5 | -7 | -7 | -3 | -5 | -2 | -3 | -4 | 0 | -6 | -2 | -5 | 17 | 0 | -6 |
| Y | -3 | -4 | -2 | -4 | 0 | -4 | -4 | -5 | 0 | -1 | -1 | -4 | -2 | 7 | -5 | -3 | -3 | 0 | 10 | -2 |
| V | 0 | -2 | -2 | -2 | -2 | -2 | -2 | -1 | -2 | 4 | 2 | -2 | 2 | -1 | -1 | -1 | 0 | -6 | -2 | 4 |

PAM matrices come with a number (i.e., PAM-$n$, with $n$ =1,30,70, etc.), where the number means that the time interval used to compute the elements of the matrix is long enough for $n$ mutations to occur per 100 amino acids

## A PAM-based alignment score

How similar is each sequence to each other using PAM-250, assuming no gaps?

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 2 | -2 | 0 | 0 | -2 | 0 | 0 | 1 | -1 | -1 | -2 | -1 | -1 | -3 | 1 | 1 | 1 | -6 | -3 | 0 |
| R | -2 | 6 | 0 | -1 | -4 | 1 | -1 | -3 | 2 | -2 | -3 | 3 | 0 | -4 | 0 | 0 | -1 | 2 | -4 | -2 |
| N | 0 | 0 | 2 | 2 | -4 | 1 | 1 | 0 | 2 | -2 | -3 | 1 | -2 | -3 | 0 | 1 | 0 | -4 | -2 | -2 |
| D | 0 | -1 | 2 | 4 | -5 | 2 | 3 | 1 | 1 | -2 | -4 | 0 | -3 | -6 | -1 | 0 | 0 | -7 | -4 | -2 |
| C | -2 | -4 | -4 | -5 | 12 | -5 | -5 | -3 | -3 | -2 | -6 | -5 | -5 | -4 | -3 | 0 | -2 | -8 | 0 | -2 |
| Q | 0 | 1 | 1 | 2 | -5 | 4 | 2 | -1 | 3 | -2 | -2 | 1 | -1 | -5 | 0 | -1 | -1 | -5 | -4 | -2 |
| E | 0 | -1 | 1 | 3 | -5 | 2 | 4 | 0 | 1 | -2 | -3 | 0 | -2 | -5 | -1 | 0 | 0 | -7 | -4 | -2 |
| G | 1 | -3 | 0 | 1 | -3 | -1 | 0 | 5 | -2 | -3 | -4 | -2 | -3 | -5 | 0 | 1 | 0 | -7 | -5 | -1 |
| H | -1 | 2 | 2 | 1 | -3 | 3 | 1 | -2 | 6 | -2 | -2 | 0 | -2 | -2 | 0 | -1 | -1 | -3 | 0 | -2 |
| I | -1 | -2 | -2 | -2 | -2 | -2 | -2 | -3 | -2 | 5 | 2 | -2 | 2 | 1 | -2 | -1 | 0 | -5 | -1 | 4 |
| L | -2 | -3 | -3 | -4 | -6 | -2 | -3 | -4 | -2 | 2 | 6 | -3 | 4 | 2 | -3 | -3 | -2 | -2 | -1 | 2 |
| K | -1 | 3 | 1 | 0 | -5 | 1 | 0 | -2 | 0 | -2 | -3 | 5 | 0 | -5 | -1 | 0 | 0 | -3 | -4 | -2 |
| M | -1 | 0 | -2 | -3 | -5 | -1 | -2 | -3 | -2 | 2 | 4 | 0 | 6 | 0 | -2 | -2 | -1 | -4 | -2 | 2 |
| F | -3 | -4 | -3 | -6 | -4 | -5 | -5 | -5 | -2 | 1 | 2 | -5 | 0 | 9 | -5 | -3 | -3 | 0 | 7 | -1 |
| P | 1 | 0 | 0 | -1 | -3 | 0 | -1 | 0 | 0 | -2 | -3 | -1 | -2 | -5 | 6 | 1 | 0 | -6 | -5 | -1 |
| S | 1 | 0 | 1 | 0 | 0 | -1 | 0 | 1 | -1 | -1 | -3 | 0 | -2 | -3 | 1 | 2 | 1 | -2 | -3 | -1 |
| T | 1 | -1 | 0 | 0 | -2 | -1 | 0 | 0 | -1 | 0 | -2 | 0 | -1 | -3 | 0 | 1 | 3 | -5 | -3 | 0 |
| W | -6 | 2 | -4 | -7 | -8 | -5 | -7 | -7 | -3 | -5 | -2 | -3 | -4 | 0 | -6 | -2 | -5 | 17 | 0 | -6 |
| Y | -3 | -4 | -2 | -4 | 0 | -4 | -4 | -5 | 0 | -1 | -1 | -4 | -2 | 7 | -5 | -3 | -3 | 0 | 10 | -2 |
| V | 0 | -2 | -2 | -2 | -2 | -2 | -2 | -1 | -2 | 4 | 2 | -2 | 2 | -1 | -1 | -1 | 0 | -6 | -2 | 4 |



1) KAWSADV

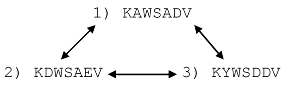
2) KDWSAEV

3) KYWSDDV

## A PAM-based alignment score

How similar is each sequence to each other using PAM-250, assuming no gaps?

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 2 | -2 | 0 | 0 | -2 | 0 | 0 | 1 | -1 | -1 | -2 | -1 | -1 | -3 | 1 | 1 | 1 | -6 | -3 | 0 |
| R | -2 | 6 | 0 | -1 | -4 | 1 | -1 | -3 | 2 | -2 | -3 | 3 | 0 | -4 | 0 | 0 | -1 | 2 | -4 | -2 |
| N | 0 | 0 | 2 | 2 | -4 | 1 | 1 | 0 | 2 | -2 | -3 | 1 | -2 | -3 | 0 | 1 | 0 | -4 | -2 | -2 |
| D | 0 | -1 | 2 | 4 | -5 | 2 | 3 | 1 | 1 | -2 | -4 | 0 | -3 | -6 | -1 | 0 | 0 | -7 | -4 | -2 |
| C | -2 | -4 | -4 | -5 | 12 | -5 | -5 | -3 | -3 | -2 | -6 | -5 | -5 | -4 | -3 | 0 | -2 | -8 | 0 | -2 |
| Q | 0 | 1 | 1 | 2 | -5 | 4 | 2 | -1 | 3 | -2 | -2 | 1 | -1 | -5 | 0 | -1 | -1 | -5 | -4 | -2 |
| E | 0 | -1 | 1 | 3 | -5 | 2 | 4 | 0 | 1 | -2 | -3 | 0 | -2 | -5 | -1 | 0 | 0 | -7 | -4 | -2 |
| G | 1 | -3 | 0 | 1 | -3 | -1 | 0 | 5 | -2 | -3 | -4 | -2 | -3 | -5 | 0 | 1 | 0 | -7 | -5 | -1 |
| H | -1 | 2 | 2 | 1 | -3 | 3 | 1 | -2 | 6 | -2 | -2 | 0 | -2 | -2 | 0 | -1 | -1 | -3 | 0 | -2 |
| I | -1 | -2 | -2 | -2 | -2 | -2 | -2 | -3 | -2 | 5 | 2 | -2 | 2 | 1 | -2 | -1 | 0 | -5 | -1 | 4 |
| L | -2 | -3 | -3 | -4 | -6 | -2 | -3 | -4 | -2 | 2 | 6 | -3 | 4 | 2 | -3 | -3 | -2 | -2 | -1 | 2 |
| K | -1 | 3 | 1 | 0 | -5 | 1 | 0 | -2 | 0 | -2 | -3 | 5 | 0 | -5 | -1 | 0 | 0 | -3 | -4 | -2 |
| M | -1 | 0 | -2 | -3 | -5 | -1 | -2 | -3 | -2 | 2 | 4 | 0 | 6 | 0 | -2 | -2 | -1 | -4 | -2 | 2 |
| F | -3 | -4 | -3 | -6 | -4 | -5 | -5 | -5 | -2 | 1 | 2 | -5 | 0 | 9 | -5 | -3 | -3 | 0 | 7 | -1 |
| P | 1 | 0 | 0 | -1 | -3 | 0 | -1 | 0 | 0 | -2 | -3 | -1 | -2 | -5 | 6 | 1 | 0 | -6 | -5 | -1 |
| S | 1 | 0 | 1 | 0 | 0 | -1 | 0 | 1 | -1 | -1 | -3 | 0 | -2 | -3 | 1 | 2 | 1 | -2 | -3 | -1 |
| T | 1 | -1 | 0 | 0 | -2 | -1 | 0 | 0 | -1 | 0 | -2 | 0 | -1 | -3 | 0 | 1 | 3 | -5 | -3 | 0 |
| W | -6 | 2 | -4 | -7 | -8 | -5 | -7 | -7 | -3 | -5 | -2 | -3 | -4 | 0 | -6 | -2 | -5 | 17 | 0 | -6 |
| Y | -3 | -4 | -2 | -4 | 0 | -4 | -4 | -5 | 0 | -1 | -1 | -4 | -2 | 7 | -5 | -3 | -3 | 0 | 10 | -2 |
| V | 0 | -2 | -2 | -2 | -2 | -2 | -2 | -1 | -2 | 4 | 2 | -2 | 2 | -1 | -1 | -1 | 0 | -6 | -2 | 4 |



1) KAWSADV

2) KDWSAEV

3) KYWSDDV

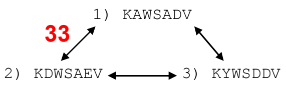
```
1) KAWSADV
2) KDWSAEV
```

**Score: 5+0+17+2+2+3+4 = 33**

## A PAM-based alignment score

How similar is each sequence to each other using PAM-250, assuming no gaps?

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 2 | -2 | 0 | 0 | -2 | 0 | 0 | 1 | -1 | -1 | -2 | -1 | -1 | -3 | 1 | 1 | 1 | -6 | -3 | 0 |
| R | -2 | 6 | 0 | -1 | -4 | 1 | -1 | -3 | 2 | -2 | -3 | 3 | 0 | -4 | 0 | 0 | -1 | 2 | -4 | -2 |
| N | 0 | 0 | 2 | 2 | -4 | 1 | 1 | 0 | 2 | -2 | -3 | 1 | -2 | -3 | 0 | 1 | 0 | -4 | -2 | -2 |
| D | 0 | -1 | 2 | 4 | -5 | 2 | 3 | 1 | 1 | -2 | -4 | 0 | -3 | -6 | -1 | 0 | 0 | -7 | -4 | -2 |
| C | -2 | -4 | -4 | -5 | 12 | -5 | -5 | -3 | -3 | -2 | -6 | -5 | -5 | -4 | -3 | 0 | -2 | -8 | 0 | -2 |
| Q | 0 | 1 | 1 | 2 | -5 | 4 | 2 | -1 | 3 | -2 | -2 | 1 | -1 | -5 | 0 | -1 | -1 | -5 | -4 | -2 |
| E | 0 | -1 | 1 | 3 | -5 | 2 | 4 | 0 | 1 | -2 | -3 | 0 | -2 | -5 | -1 | 0 | 0 | -7 | -4 | -2 |
| G | 1 | -3 | 0 | 1 | -3 | -1 | 0 | 5 | -2 | -3 | -4 | -2 | -3 | -5 | 0 | 1 | 0 | -7 | -5 | -1 |
| H | -1 | 2 | 2 | 1 | -3 | 3 | 1 | -2 | 6 | -2 | -2 | 0 | -2 | -2 | 0 | -1 | -1 | -3 | 0 | -2 |
| I | -1 | -2 | -2 | -2 | -2 | -2 | -2 | -3 | -2 | 5 | 2 | -2 | 2 | 1 | -2 | -1 | 0 | -5 | -1 | 4 |
| L | -2 | -3 | -3 | -4 | -6 | -2 | -3 | -4 | -2 | 2 | 6 | -3 | 4 | 2 | -3 | -3 | -2 | -2 | -1 | 2 |
| K | -1 | 3 | 1 | 0 | -5 | 1 | 0 | -2 | 0 | -2 | -3 | 5 | 0 | -5 | -1 | 0 | 0 | -3 | -4 | -2 |
| M | -1 | 0 | -2 | -3 | -5 | -1 | -2 | -3 | -2 | 2 | 4 | 0 | 6 | 0 | -2 | -2 | -1 | -4 | -2 | 2 |
| F | -3 | -4 | -3 | -6 | -4 | -5 | -5 | -5 | -2 | 1 | 2 | -5 | 0 | 9 | -5 | -3 | -3 | 0 | 7 | -1 |
| P | 1 | 0 | 0 | -1 | -3 | 0 | -1 | 0 | 0 | -2 | -3 | -1 | -2 | -5 | 6 | 1 | 0 | -6 | -5 | -1 |
| S | 1 | 0 | 1 | 0 | 0 | -1 | 0 | 1 | -1 | -1 | -3 | 0 | -2 | -3 | 1 | 2 | 1 | -2 | -3 | -1 |
| T | 1 | -1 | 0 | 0 | -2 | -1 | 0 | 0 | -1 | 0 | -2 | 0 | -1 | -3 | 0 | 1 | 3 | -5 | -3 | 0 |
| W | -6 | 2 | -4 | -7 | -8 | -5 | -7 | -7 | -3 | -5 | -2 | -3 | -4 | 0 | -6 | -2 | -5 | 17 | 0 | -6 |
| Y | -3 | -4 | -2 | -4 | 0 | -4 | -4 | -5 | 0 | -1 | -1 | -4 | -2 | 7 | -5 | -3 | -3 | 0 | 10 | -2 |
| V | 0 | -2 | -2 | -2 | -2 | -2 | -2 | -1 | -2 | 4 | 2 | -2 | 2 | -1 | -1 | -1 | 0 | -6 | -2 | 4 |



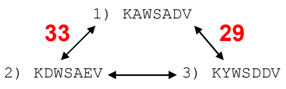
```
1) KAWSADV
3) KYWSDDV
```

Score: 5-3+17+2+0+4+4 = 29

## A PAM-based alignment score

How similar is each sequence to each other using PAM-250, assuming no gaps?

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 2 | -2 | 0 | 0 | -2 | 0 | 0 | 1 | -1 | -1 | -2 | -1 | -1 | -3 | 1 | 1 | 1 | -6 | -3 | 0 |
| R | -2 | 6 | 0 | -1 | -4 | 1 | -1 | -3 | 2 | -2 | -3 | 3 | 0 | -4 | 0 | 0 | -1 | 2 | -4 | -2 |
| N | 0 | 0 | 2 | 2 | -4 | 1 | 1 | 0 | 2 | -2 | -3 | 1 | -2 | -3 | 0 | 1 | 0 | -4 | -2 | -2 |
| D | 0 | -1 | 2 | 4 | -5 | 2 | 3 | 1 | 1 | -2 | -4 | 0 | -3 | -6 | -1 | 0 | 0 | -7 | -4 | -2 |
| C | -2 | -4 | -4 | -5 | 12 | -5 | -5 | -3 | -3 | -2 | -6 | -5 | -5 | -4 | -3 | 0 | -2 | -8 | 0 | -2 |
| Q | 0 | 1 | 1 | 2 | -5 | 4 | 2 | -1 | 3 | -2 | -2 | 1 | -1 | -5 | 0 | -1 | -1 | -5 | -4 | -2 |
| E | 0 | -1 | 1 | 3 | -5 | 2 | 4 | 0 | 1 | -2 | -3 | 0 | -2 | -5 | -1 | 0 | 0 | -7 | -4 | -2 |
| G | 1 | -3 | 0 | 1 | -3 | -1 | 0 | 5 | -2 | -3 | -4 | -2 | -3 | -5 | 0 | 1 | 0 | -7 | -5 | -1 |
| H | -1 | 2 | 2 | 1 | -3 | 3 | 1 | -2 | 6 | -2 | -2 | 0 | -2 | -2 | 0 | -1 | -1 | -3 | 0 | -2 |
| I | -1 | -2 | -2 | -2 | -2 | -2 | -2 | -3 | -2 | 5 | 2 | -2 | 2 | 1 | -2 | -1 | 0 | -5 | -1 | 4 |
| L | -2 | -3 | -3 | -4 | -6 | -2 | -3 | -4 | -2 | 2 | 6 | -3 | 4 | 2 | -3 | -3 | -2 | -2 | -1 | 2 |
| K | -1 | 3 | 1 | 0 | -5 | 1 | 0 | -2 | 0 | -2 | -3 | 5 | 0 | -5 | -1 | 0 | 0 | -3 | -4 | -2 |
| M | -1 | 0 | -2 | -3 | -5 | -1 | -2 | -3 | -2 | 2 | 4 | 0 | 6 | 0 | -2 | -2 | -1 | -4 | -2 | 2 |
| F | -3 | -4 | -3 | -6 | -4 | -5 | -5 | -5 | -2 | 1 | 2 | -5 | 0 | 9 | -5 | -3 | -3 | 0 | 7 | -1 |
| P | 1 | 0 | 0 | -1 | -3 | 0 | -1 | 0 | 0 | -2 | -3 | -1 | -2 | -5 | 6 | 1 | 0 | -6 | -5 | -1 |
| S | 1 | 0 | 1 | 0 | 0 | -1 | 0 | 1 | -1 | -1 | -3 | 0 | -2 | -3 | 1 | 2 | 1 | -2 | -3 | -1 |
| T | 1 | -1 | 0 | 0 | -2 | -1 | 0 | 0 | -1 | 0 | -2 | 0 | -1 | -3 | 0 | 1 | 3 | -5 | -3 | 0 |
| W | -6 | 2 | -4 | -7 | -8 | -5 | -7 | -7 | -3 | -5 | -2 | -3 | -4 | 0 | -6 | -2 | -5 | 17 | 0 | -6 |
| Y | -3 | -4 | -2 | -4 | 0 | -4 | -4 | -5 | 0 | -1 | -1 | -4 | -2 | 7 | -5 | -3 | -3 | 0 | 10 | -2 |
| V | 0 | -2 | -2 | -2 | -2 | -2 | -2 | -1 | -2 | 4 | 2 | -2 | 2 | -1 | -1 | -1 | 0 | -6 | -2 | 4 |



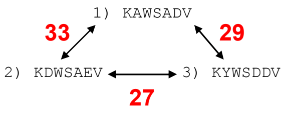
```
2) KDWSAEV
3) KYWSDDV
```

Score: 5-4+17+2+0+3+4 = 27

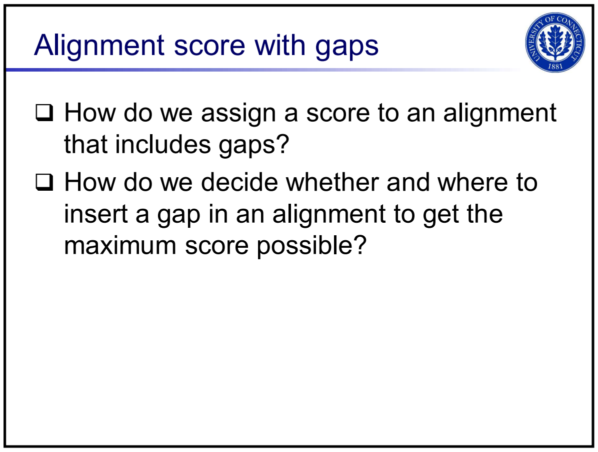
## A PAM-based alignment score

How similar is each sequence to each other using PAM-250, assuming no gaps?

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 2 | -2 | 0 | 0 | -2 | 0 | 0 | 1 | -1 | -1 | -2 | -1 | -1 | -3 | 1 | 1 | 1 | -6 | -3 | 0 |
| R | -2 | 6 | 0 | -1 | -4 | 1 | -1 | -3 | 2 | -2 | -3 | 3 | 0 | -4 | 0 | 0 | -1 | 2 | -4 | -2 |
| N | 0 | 0 | 2 | 2 | -4 | 1 | 1 | 0 | 2 | -2 | -3 | 1 | -2 | -3 | 0 | 1 | 0 | -4 | -2 | -2 |
| D | 0 | -1 | 2 | 4 | -5 | 2 | 3 | 1 | 1 | -2 | -4 | 0 | -3 | -6 | -1 | 0 | 0 | -7 | -4 | -2 |
| C | -2 | -4 | -4 | -5 | 12 | -5 | -5 | -3 | -3 | -2 | -6 | -5 | -5 | -4 | -3 | 0 | -2 | -8 | 0 | -2 |
| Q | 0 | 1 | 1 | 2 | -5 | 4 | 2 | -1 | 3 | -2 | -2 | 1 | -1 | -5 | 0 | -1 | -1 | -5 | -4 | -2 |
| E | 0 | -1 | 1 | 3 | -5 | 2 | 4 | 0 | 1 | -2 | -3 | 0 | -2 | -5 | -1 | 0 | 0 | -7 | -4 | -2 |
| G | 1 | -3 | 0 | 1 | -3 | -1 | 0 | 5 | -2 | -3 | -4 | -2 | -3 | -5 | 0 | 1 | 0 | -7 | -5 | -1 |
| H | -1 | 2 | 2 | 1 | -3 | 3 | 1 | -2 | 6 | -2 | -2 | 0 | -2 | -2 | 0 | -1 | -1 | -3 | 0 | -2 |
| I | -1 | -2 | -2 | -2 | -2 | -2 | -2 | -3 | -2 | 5 | 2 | -2 | 2 | 1 | -2 | -1 | 0 | -5 | -1 | 4 |
| L | -2 | -3 | -3 | -4 | -6 | -2 | -3 | -4 | -2 | 2 | 6 | -3 | 4 | 2 | -3 | -3 | -2 | -2 | -1 | 2 |
| K | -1 | 3 | 1 | 0 | -5 | 1 | 0 | -2 | 0 | -2 | -3 | 5 | 0 | -5 | -1 | 0 | 0 | -3 | -4 | -2 |
| M | -1 | 0 | -2 | -3 | -5 | -1 | -2 | -3 | -2 | 2 | 4 | 0 | 6 | 0 | -2 | -2 | -1 | -4 | -2 | 2 |
| F | -3 | -4 | -3 | -6 | -4 | -5 | -5 | -5 | -2 | 1 | 2 | -5 | 0 | 9 | -5 | -3 | -3 | 0 | 7 | -1 |
| P | 1 | 0 | 0 | -1 | -3 | 0 | -1 | 0 | 0 | -2 | -3 | -1 | -2 | -5 | 6 | 1 | 0 | -6 | -5 | -1 |
| S | 1 | 0 | 1 | 0 | 0 | -1 | 0 | 1 | -1 | -1 | -3 | 0 | -2 | -3 | 1 | 2 | 1 | -2 | -3 | -1 |
| T | 1 | -1 | 0 | 0 | -2 | -1 | 0 | 0 | -1 | 0 | -2 | 0 | -1 | -3 | 0 | 1 | 3 | -5 | -3 | 0 |
| W | -6 | 2 | -4 | -7 | -8 | -5 | -7 | -7 | -3 | -5 | -2 | -3 | -4 | 0 | -6 | -2 | -5 | 17 | 0 | -6 |
| Y | -3 | -4 | -2 | -4 | 0 | -4 | -4 | -5 | 0 | -1 | -1 | -4 | -2 | 7 | -5 | -3 | -3 | 0 | 10 | -2 |
| V | 0 | -2 | -2 | -2 | -2 | -2 | -2 | -1 | -2 | 4 | 2 | -2 | 2 | -1 | -1 | -1 | 0 | -6 | -2 | 4 |

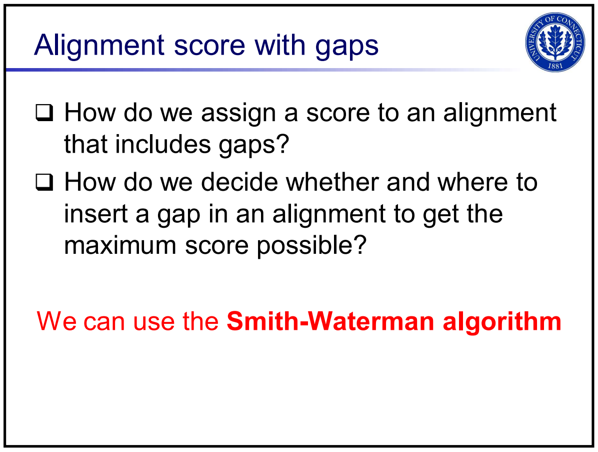


*The highest similarity is between sequences 1) and 2)*



## Alignment score with gaps

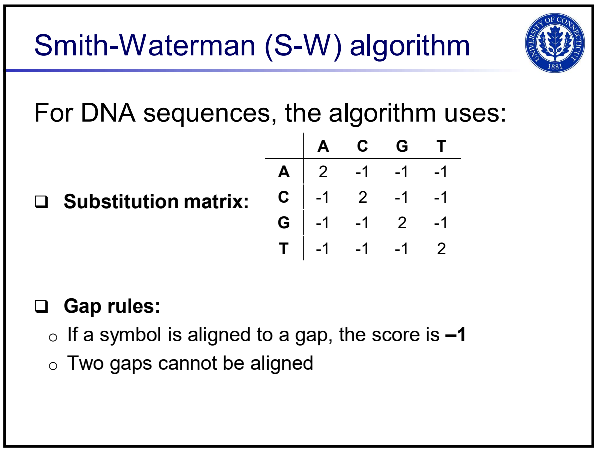
- ❑ How do we assign a score to an alignment that includes gaps?
- ❑ How do we decide whether and where to insert a gap in an alignment to get the maximum score possible?



## Alignment score with gaps

- ❑ How do we assign a score to an alignment that includes gaps?
- ❑ How do we decide whether and where to insert a gap in an alignment to get the maximum score possible?

We can use the **Smith-Waterman algorithm**



## Smith-Waterman (S-W) algorithm

For DNA sequences, the algorithm uses:

- ❑ **Substitution matrix:**

| | **A** | **C** | **G** | **T** |
|---|---|---|---|---|
| **A** | 2 | -1 | -1 | -1 |
| **C** | -1 | 2 | -1 | -1 |
| **G** | -1 | -1 | 2 | -1 |
| **T** | -1 | -1 | -1 | 2 |

- ❑ **Gap rules:**
  - ○ If a symbol is aligned to a gap, the score is **–1**
  - ○ Two gaps cannot be aligned

## S-W algorithm

Let us align these two sequences:

`ACAC        AGCA`

## S-W algorithm

Let us align these two sequences:

`ACAC        AGCA`

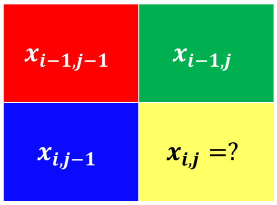
1) We build a table

|   | - | A | C | A | C |
|---|---|---|---|---|---|
| - | 0 | 0 | 0 | 0 | 0 |
| A | 0 |   |   |   |   |
| G | 0 |   |   |   |   |
| C | 0 |   |   |   |   |
| A | 0 |   |   |   |   |

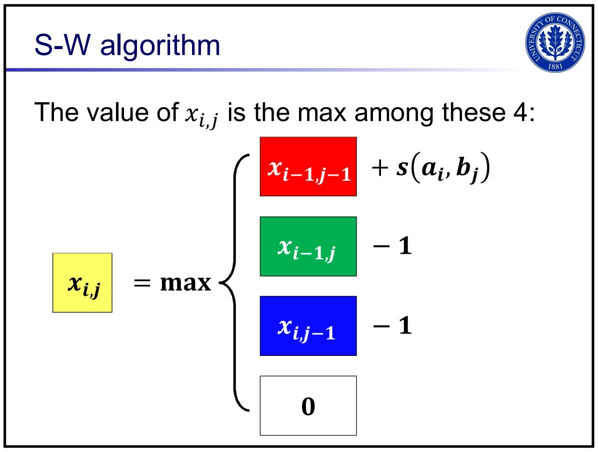
## S-W algorithm

2) We fill in the table recursively, starting at the top left and working our way down

**A graphical illustration of how to do it**



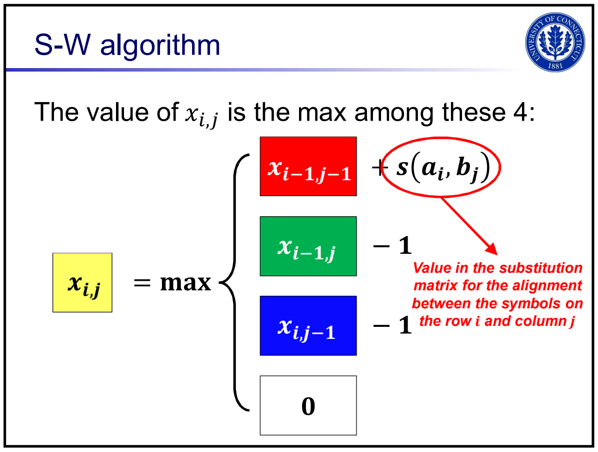
***If we already have numbers in red, green, and blue boxes, what do we put in the yellow box?***



## S-W algorithm

The value of $x_{i,j}$ is the max among these 4:

$$x_{i,j} = \max \begin{cases} x_{i-1,j-1} + s(a_i, b_j) \\ x_{i-1,j} - 1 \\ x_{i,j-1} - 1 \\ 0 \end{cases}$$

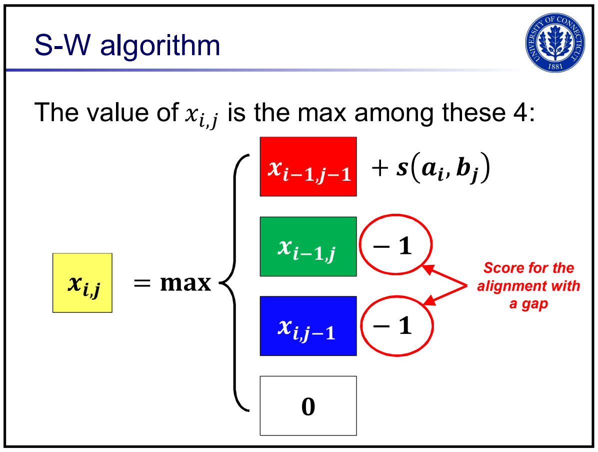


## S-W algorithm

The value of $x_{i,j}$ is the max among these 4:

$$x_{i,j} = \max \begin{cases} x_{i-1,j-1} + s(a_i, b_j) \\ x_{i-1,j} - 1 \\ x_{i,j-1} - 1 \\ 0 \end{cases}$$

*Value in the substitution matrix for the alignment between the symbols on the row i and column j*



## S-W algorithm

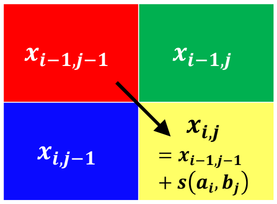
The value of $x_{i,j}$ is the max among these 4:

$$x_{i,j} = \max \begin{cases} x_{i-1,j-1} + s(a_i, b_j) \\ x_{i-1,j} - 1 \\ x_{i,j-1} - 1 \\ 0 \end{cases}$$

*Score for the alignment with a gap*

## S-W algorithm

**As we do this, we put little arrows in the table so we know where our numbers are coming from! For example if the yellow value came from the red square, we write:**



$x_{i-1,j-1}$

$x_{i-1,j}$

$x_{i,j-1}$

$x_{i,j} = x_{i-1,j-1} + s(a_i, b_j)$

***If ALL of the color numbers were negative and you just put zero, you don't need the arrow***

## S-W algorithm

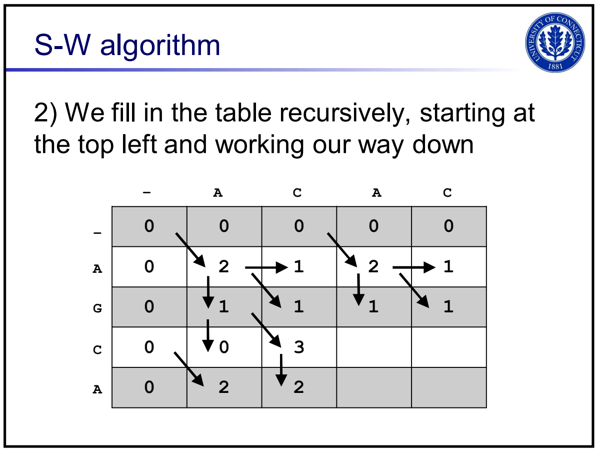
2) We fill in the table recursively, starting at the top left and working our way down

|   | - | A | C | A | C |
|---|---|---|---|---|---|
| - | 0 | 0 | 0 | 0 | 0 |
| A | 0 | ↘ 2 | → 1 | ↘ 2 | → 1 |
| G | 0 |  |  |  |  |
| C | 0 |  |  |  |  |
| A | 0 |  |  |  |  |

## S-W algorithm

2) We fill in the table recursively, starting at the top left and working our way down

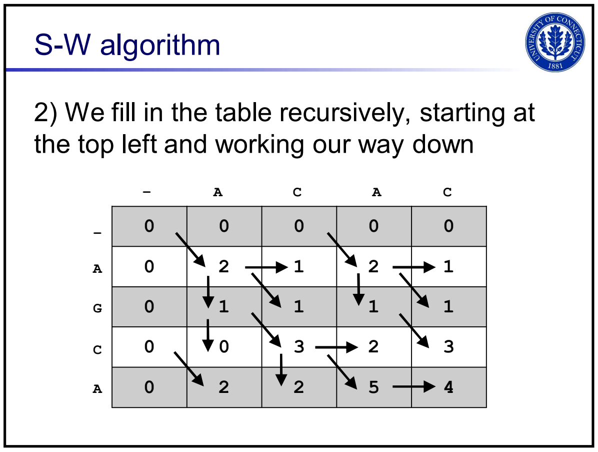
|   | - | A | C | A | C |
|---|---|---|---|---|---|
| - | 0 | 0 | 0 | 0 | 0 |
| A | 0 | ↘ 2 | → 1 | ↘ 2 | → 1 |
| G | 0 | ↓ 1 |  |  |  |
| C | 0 | ↓ 0 |  |  |  |
| A | 0 | ↘ 2 |  |  |  |



## S-W algorithm

2) We fill in the table recursively, starting at the top left and working our way down

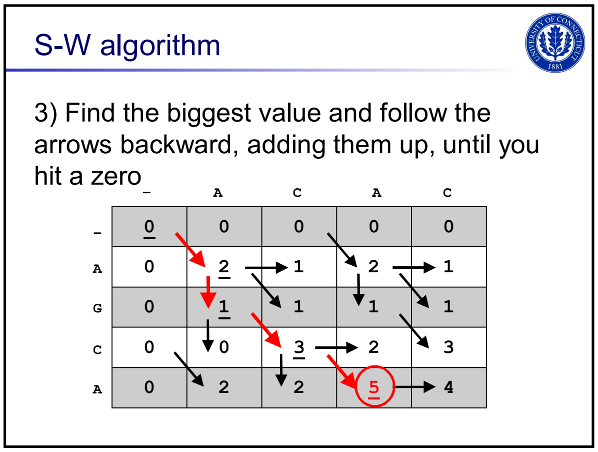
| | – | A | C | A | C |
|---|---|---|---|---|---|
| _ | 0 | 0 | 0 | 0 | 0 |
| A | 0 | 2 | 1 | 2 | 1 |
| G | 0 | 1 | 1 | 1 | 1 |
| C | 0 | 0 | 3 | | |
| A | 0 | 2 | 2 | | |



## S-W algorithm

2) We fill in the table recursively, starting at the top left and working our way down

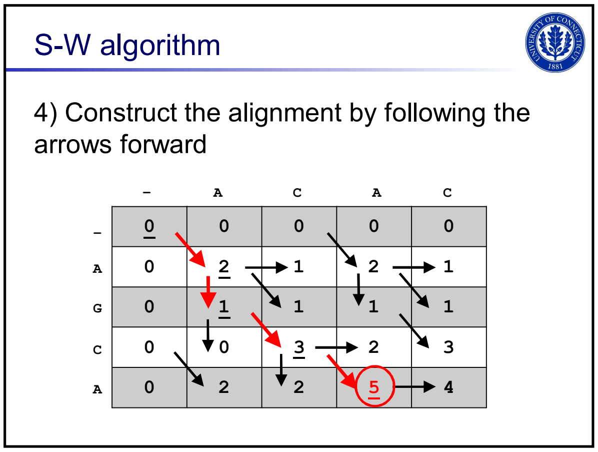
| | – | A | C | A | C |
|---|---|---|---|---|---|
| _ | 0 | 0 | 0 | 0 | 0 |
| A | 0 | 2 | 1 | 2 | 1 |
| G | 0 | 1 | 1 | 1 | 1 |
| C | 0 | 0 | 3 | 2 | 3 |
| A | 0 | 2 | 2 | 5 | 4 |



## S-W algorithm

3) Find the biggest value and follow the arrows backward, adding them up, until you hit a zero

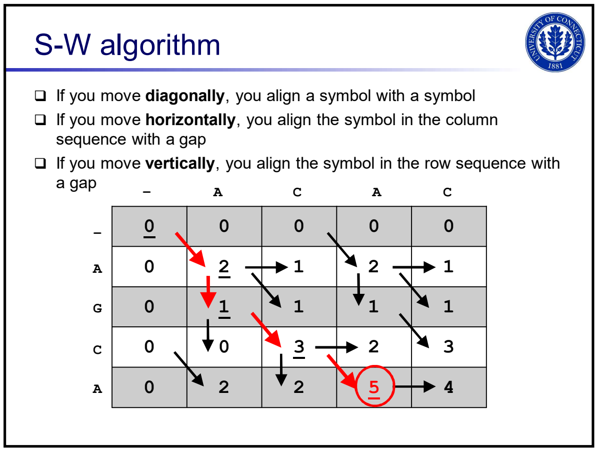
| | – | A | C | A | C |
|---|---|---|---|---|---|
| _ | 0 | 0 | 0 | 0 | 0 |
| A | 0 | 2 | 1 | 2 | 1 |
| G | 0 | 1 | 1 | 1 | 1 |
| C | 0 | 0 | 3 | 2 | 3 |
| A | 0 | 2 | 2 | 5 | 4 |



## S-W algorithm

4) Construct the alignment by following the arrows forward

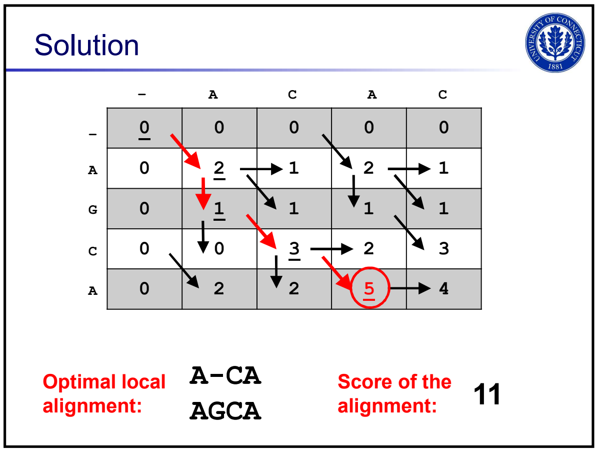
| | – | A | C | A | C |
|---|---|---|---|---|---|
| – | 0 | 0 | 0 | 0 | 0 |
| A | 0 | 2 | 1 | 2 | 1 |
| G | 0 | 1 | 1 | 1 | 1 |
| C | 0 | 0 | 3 | 2 | 3 |
| A | 0 | 2 | 2 | 5 | 4 |



## S-W algorithm

- If you move **diagonally**, you align a symbol with a symbol
- If you move **horizontally**, you align the symbol in the column sequence with a gap
- If you move **vertically**, you align the symbol in the row sequence with a gap

| | – | A | C | A | C |
|---|---|---|---|---|---|
| – | 0 | 0 | 0 | 0 | 0 |
| A | 0 | 2 | 1 | 2 | 1 |
| G | 0 | 1 | 1 | 1 | 1 |
| C | 0 | 0 | 3 | 2 | 3 |
| A | 0 | 2 | 2 | 5 | 4 |



## Solution

| | – | A | C | A | C |
|---|---|---|---|---|---|
| – | 0 | 0 | 0 | 0 | 0 |
| A | 0 | 2 | 1 | 2 | 1 |
| G | 0 | 1 | 1 | 1 | 1 |
| C | 0 | 0 | 3 | 2 | 3 |
| A | 0 | 2 | 2 | 5 | 4 |

**Optimal local alignment:**

```
A-CA
AGCA
```

**Score of the alignment:** 11

## Solution

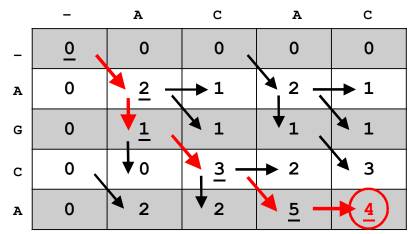
**NOTE:** The S-W algorithm finds an optimal *local* alignment and has left out two of the symbols (one per sequence)
To have a **complete alignment**, in which all symbols are paired, you have to start at the lower right of the table and use exactly the same process

**Optimal local alignment:**

```
A-CA
AGCA
```

**Score of the alignment:** **11**

## Complete alignment



|   | - | A | C | A | C |
|---|---|---|---|---|---|
| - | 0 | 0 | 0 | 0 | 0 |
| A | 0 | 2 | 1 | 2 | 1 |
| G | 0 | 1 | 1 | 1 | 1 |
| C | 0 | 0 | 3 | 2 | 3 |
| A | 0 | 2 | 2 | 5 | 4 |

**Optimal local alignment:**

```
A-CAC
AGCA-
```

**Score of the alignment:** **15**