SEQUENCE ALIGNMENT: INVESTIGATING AN INFLUENZA OUTBREAK

BIO 300/CMPSC 300
Dr. Kristen Webb
Spring 2016

Descent with Modification

- DNA replication ensures a mostly faithful passing of the genome to progeny

- What would be the consequence of 100% accurate replication?

Descent with Modification

How does descent with modification happen?

Descent with Modification

How does descent with modification happen?

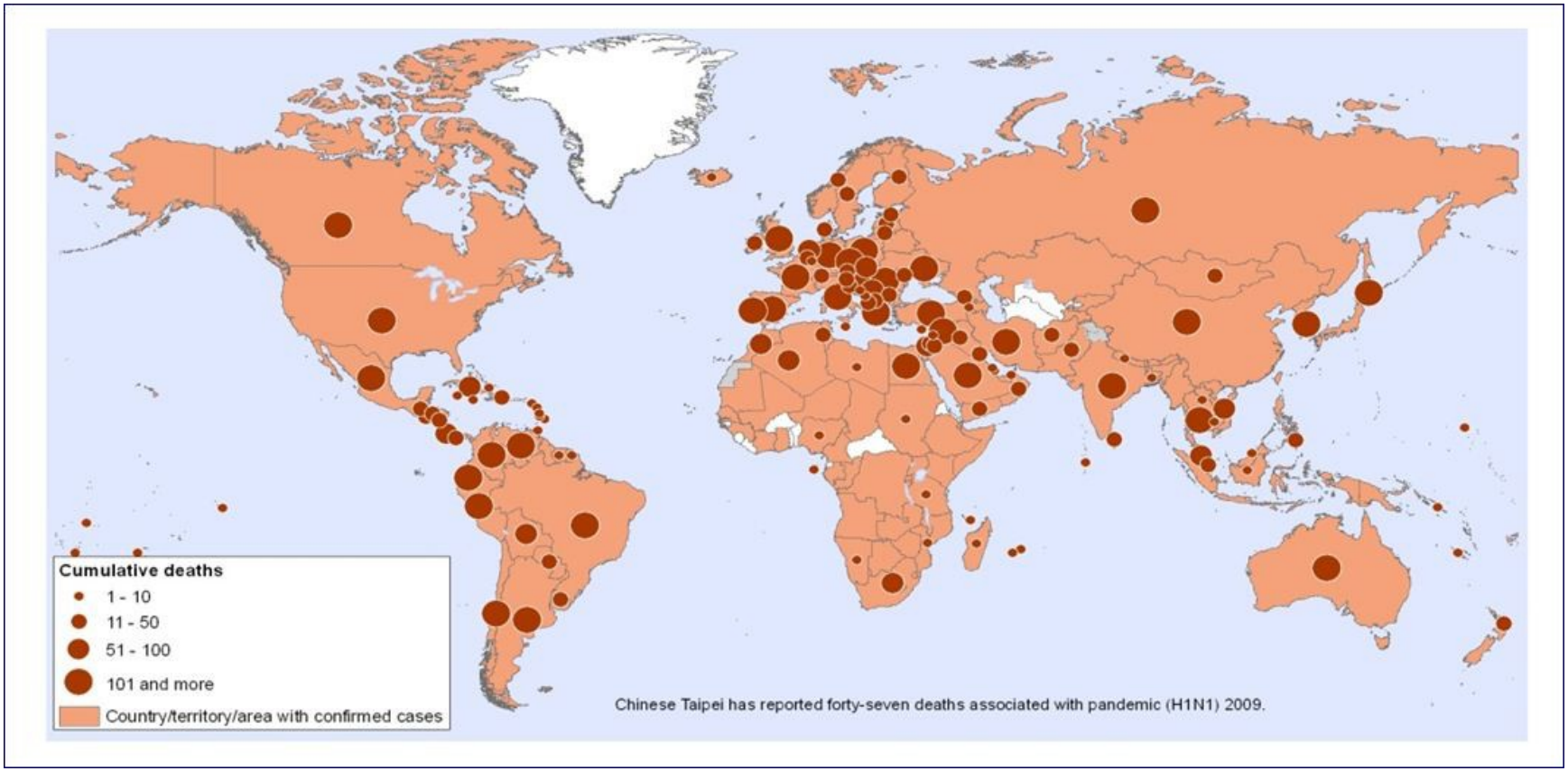
- **Mutation**
  - A change in a DNA sequence
    - Results from errors in replication or repair
  - Mutation is the ultimate source of genetic variation

Domestic Dog – *Canis lupus familiaris*
All descendent from the grey wolf
All the same species (sub-species)
Breeds – variation within the species

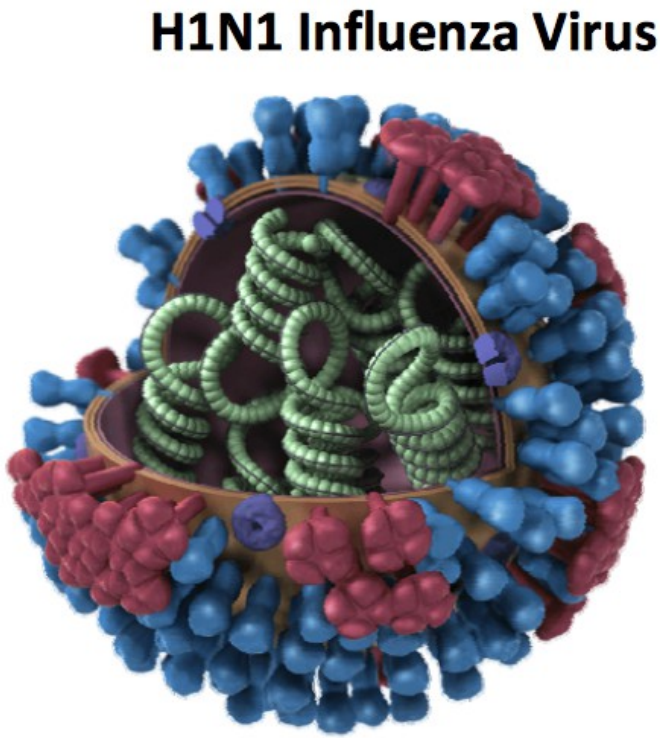
Sequence Variations

- sequences may have diverged from a common ancestor through various types of mutations:
  - substitutions (A**C**GA → A**G**GA)
  - insertions (ACGA → AC**CGGA**GA)
  - deletions (A**CGGA**GA → AGA)

Tracking Infectious Disease – 2009 H_1N_1 Influenza Pandemic



H_1N_1 Influenza Virus



Hemagglutinin- protein the virus uses to attach to the host cells

Neuraminidase- enables the virus to be released from the host cell

Viral Evolution

- Viruses evolve very quickly
  - Some of the highest mutation rates known
    - Arms race with immune system
  - Viruses – mutation rate 0.0001 - 0.000001 mutations per base per generation
    - One mutation every 10,000 – 1,000,000 nucleotides
    - Influenza genome size = ~14,000 nucleotides
    - Entire genome is coding regions (genes)
  - Humans – 0.00000001 mutations per base per generation
    - One mutation every 100,000,000 nucleotides
    - Human genome size 3 billion nucleotides
    - Only 1.5% of genome is coding regions (genes)

Pairwise Alignment
Similarity and Relatedness

Alignment of a gene from two closely related viruses

```
Hemagglutinin gene from virus A:  ATGAACGCAATACTCGTAGTT...
                                  ||||| |||||||| ||||||
Hemagglutinin gene from virus B:  ATGAAGGCAATACTAGTAGTT...
```

Alignment of a gene from two distantly related viruses

```
Hemagglutinin gene from virus A:  ATGAACGCAATACTCGTAGTT...
                                  ||| ||| ||| |||| |  |
Hemagglutinin gene from virus C:  ATGCACGAAATGCTCGGACCT...
```

Tracing an Infection to a Source - HIV

- 1990 – CDC report that a women in Florida had contracted HIV from her dentist
  - Dentist diagnosed with HIV in 1986, developed AIDS in 1987
  - Patient had no other risk factors and had not been in contact with other HIV-positive persons
  - Patient had had an invasive dental procedure

- Tested dentist's other patients - 10 tested positive for HIV
- Did they contract HIV from the dentist?

Molecular Epidemiology of HIV Transmission in a Dental Practice

Table 23.10 HIV-positive persons included in study of HIV isolates from a Florida dental practice

Person	Sex	Known risk factors	*Average differences in DNA sequences (%)* From HIV from dentist	From HIV from controls
Dentist	M	Yes		11.0
Patient A	F	No	3.4	10.9
Patient B	F	No	4.4	11.2
Patient C	M	No	3.4	11.1
Patient E	F	No	3.4	10.8
Patient G	M	No	4.9	11.8
Patient D	M	Yes	13.6	13.1
Patient F	**M**	**Yes**	**10.7**	**11.9**

Source: After C. Ou et al., *Science* 256(1992):1165–1171, Table 1.

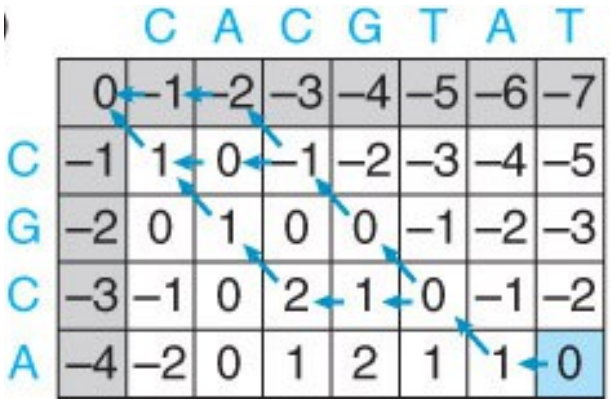
Sequence Aligment

- substitutions (A**C**GA → A**G**GA)
- insertions (ACGA → AC**CGGA**GA)
- deletions (A**CGGA**GA → AGA)

Global Pairwise Alignment

- Dynamic Programming
  - Divide a problem into a series of smaller subproblems
  - Solve each subproblem
  - Use the solutions to build the solution to the original problem

- Needleman-Wunsch Algorithm
  - Creates a matrix of partial alignment scores
  - Backtracks along a path to the best possible alignment



Needleman-Wunsch Algorithm

- Create N x M matrix
- Place each sequence along one axis
- Place score 0 at the up-left corner
- Fill in 1st row & column with gap penalty multiples
- Fill in the matrix with max value of 3 possible moves:
  - Vertical move: Score + gap penalty
  - Horizontal move: Score + gap penalty
  - Diagonal move: Score + match/mismatch score
- The optimal alignment score is in the lower-right corner
- To reconstruct the optimal alignment, trace back where the max at each step came from, stop when hit the origin.

Example

Alignment score = 0

Let:

Match = +1

Mismatch = 0

Gap = -1

		C	A	C	G	T	A	T
	0	-1	-2	-3	-4	-5	-6	-7
C	-1	1	0	-1	-2	-3	-4	-5
G	-2	0	1	0	0	-1	-2	-3
C	-3	-1	0	2	1	0	-1	-2
A	-4	-2	0	1	2	1	1	0

Example

Alignment score = 0

Let:
Match = +1
Mismatch = 0
Gap = -1

```
CACGTAT
--CGCA-
```

		C	A	C	G	T	A	T
	0	-1	-2	-3	-4	-5	-6	-7
C	-1	1	0	-1	-2	-3	-4	-5
G	-2	0	1	0	0	-1	-2	-3
C	-3	-1	0	2	1	0	-1	-2
A	-4	-2	0	1	2	1	1	0

Example

Alignment score = 0

Let:
Match = +1
Mismatch = 0
Gap = -1

CACGTAT
C--GCA-

		C	A	C	G	T	A	T
	0	-1	-2	-3	-4	-5	-6	-7
C	-1	1	0	-1	-2	-3	-4	-5
G	-2	0	1	0	0	-1	-2	-3
C	-3	-1	0	2	1	0	-1	-2
A	-4	-2	0	1	2	1	1	0

Example

Alignment score = 0

Let:
Match = +1
Mismatch = 0
Gap = -1

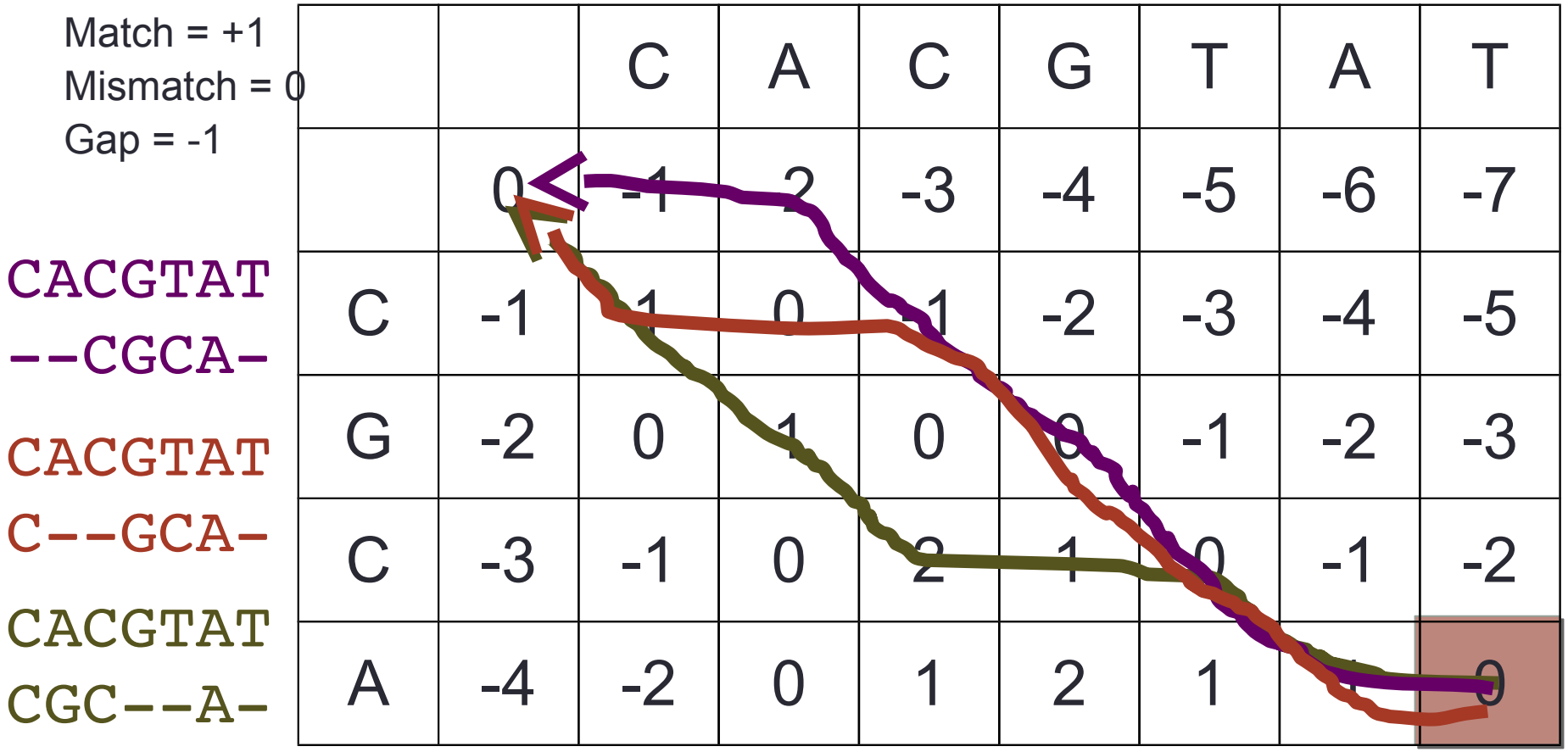
```
CACGTAT
CGC--A-
```

		C	A	C	G	T	A	T
	0	-1	-2	-3	-4	-5	-6	-7
C	-1	1	0	-1	-2	-3	-4	-5
G	-2	0	1	0	0	-1	-2	-3
C	-3	-1	0	2	1	0	-1	-2
A	-4	-2	0	1	2	1	1	0

Example

Alignment score = 0

Let:
Match = +1
Mismatch = 0
Gap = -1



		C	A	C	G	T	A	T
	0	-1	-2	-3	-4	-5	-6	-7
C	-1	1	0	-1	-2	-3	-4	-5
G	-2	0	1	0	0	-1	-2	-3
C	-3	-1	0	2	1	0	-1	-2
A	-4	-2	0	1	2	1	0	0

CACGTAT
--CGCA-

CACGTAT
C--GCA-

CACGTAT
CGC--A-