CMSC 423: Sequence Alignment

Part5

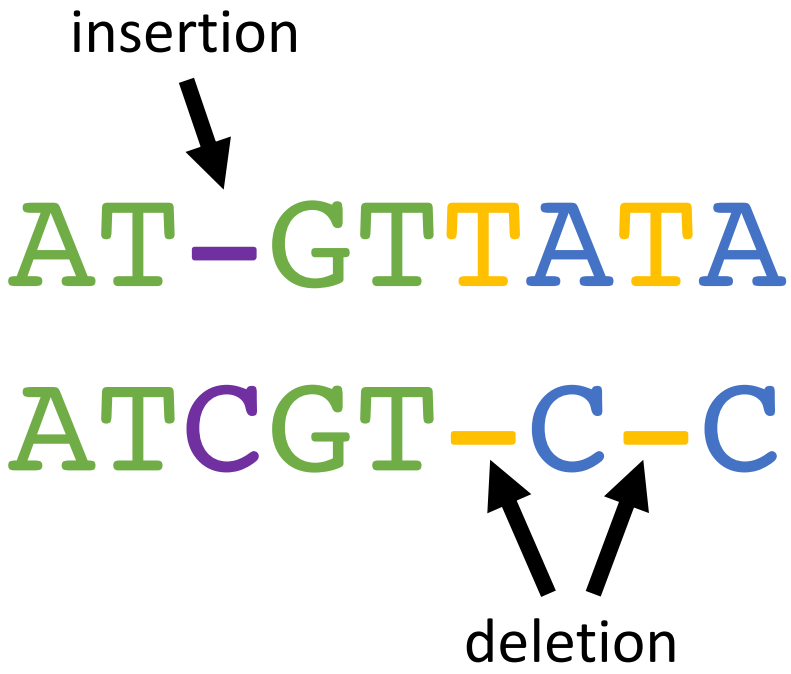
Inexact matching: why?

- 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

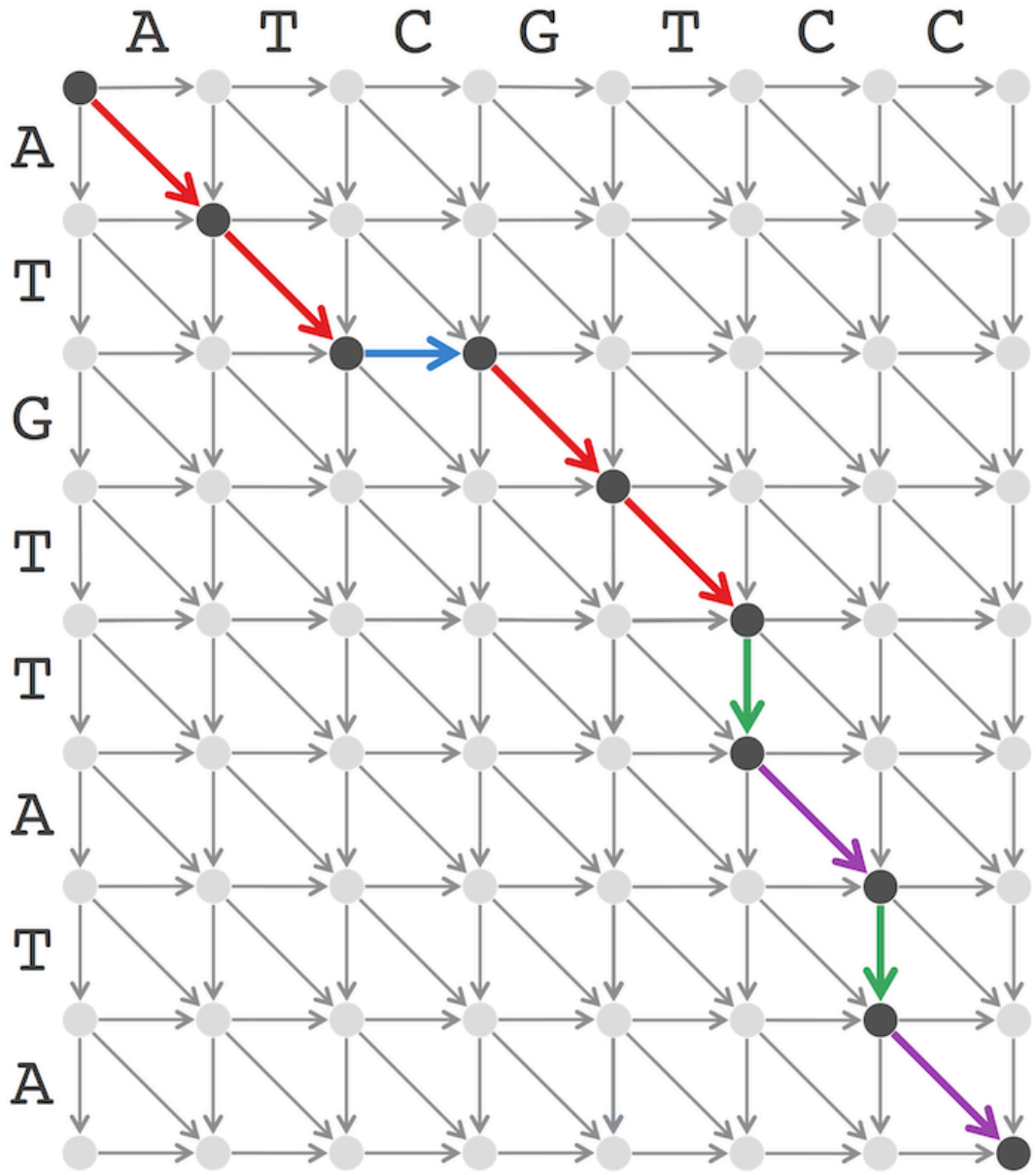
Alignment of sequences *v* and *w*

v = ATGTTATA

w = ATCGTCC



Sequence Alignment is the Manhattan Tourist Problem in Disguise



i

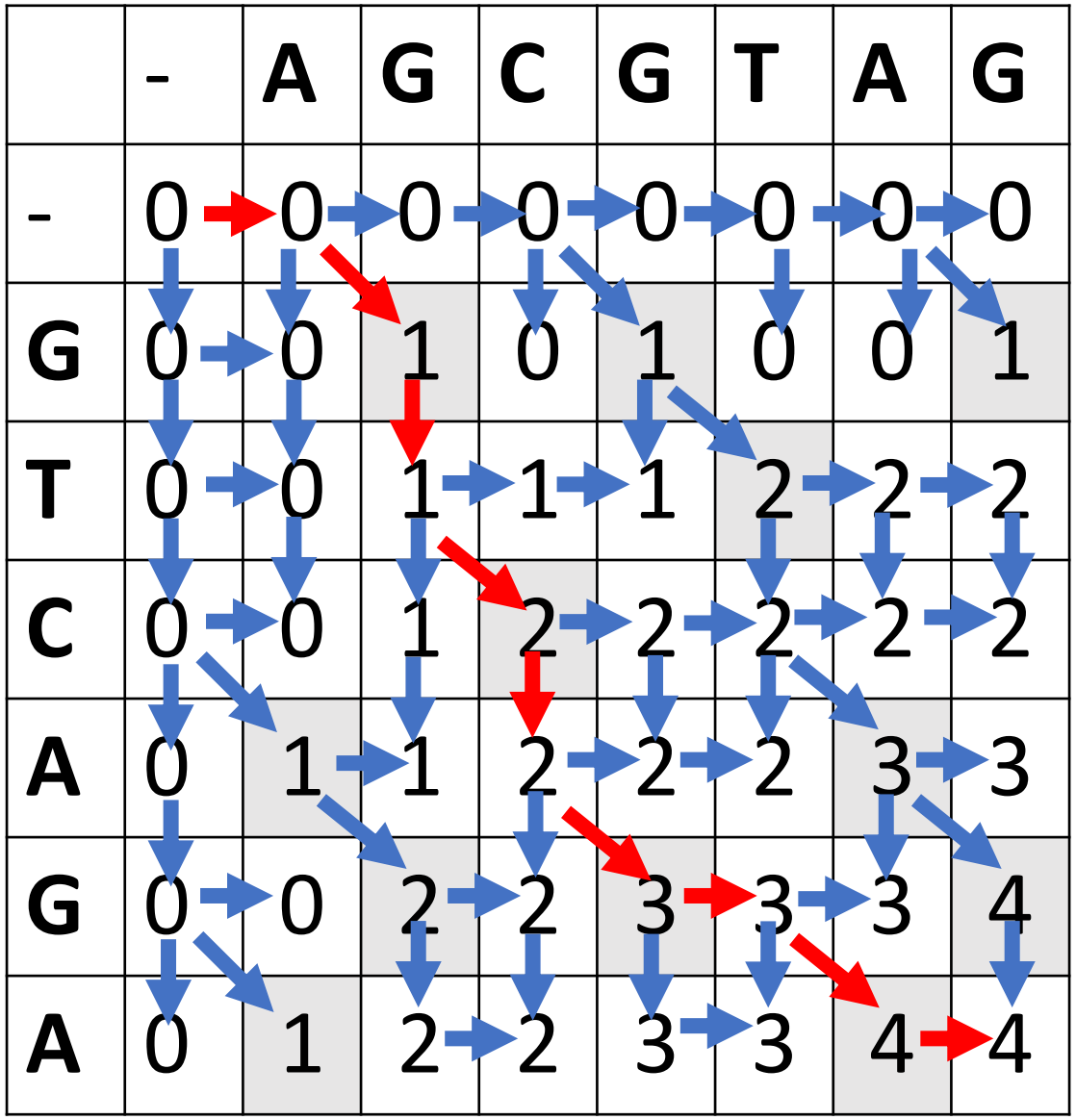
	-	**A**	**G**	**C**	**G**	**T**	**A**	**G**
-	0							
G								
T								
C								
A								
G								
A								

j

v = AGCGTAG

w = GTCAGA

$$s_{i,j} = \max \begin{cases} s_{i-1,\,j-1} + 1, & \text{if } v_i = w_j \\ s_{i-1,\,j} + 0 \\ s_{i,\,j-1} + 0 \end{cases}$$



v = AGCGTAG

w = GTCAGA

A G – C – G T A G

– G T C A G – A –

Local alignment

- What if we just want a region of similarity?
- Change the first row and column in the dynamic programming table to 0s
- Allow the alignment to start anywhere

  Score[i,j] = max{0, case 1, case 2, case 3}

- Answer is the location in the matrix with the highest score

Extending to sequence alignment

- When solving for the LCS, mis-alignments are free
- What happens if we pay for our "mistakes"? (this also allows us to account for "similar" amino acids)

Value[Match] = 10

Value[Mismatch] = -5

Value[Gap] = -2

Match = [A,A], [C,C], [G,G], [T,T]

Mismatch = [A,G], [A,C], [A,T], …

Gap = [A,-], [-,A], …

The same dynamic programming algorithm works!

Penalizing insertions and Deletions

- Linear scoring model
  - σ = penalty for insertion or deletion of a single simple
  - σ · *k* = penalty for insertion or deletion of *k* symbols

```
GATCCAG        GATCCAG
GA-C-AG        GA--CAG
```

Penalizing insertions and Deletions

Mutations are often caused by errors in DNA replication that insert or delete an entire interval of *k* nucleotides as a single event (instead of *k* independent insertions or deletions)

```
GATCCAG        GATCCAG
GA-C-AG        GA--CAG
```

Penalizing insertions and Deletions

Mutations are often caused by errors in DNA replication that insert or delete an entire interval of *k* nucleotides as a single event (instead of *k* independent insertions or deletions)

Affine gap penalties

$$\text{Cost}(k \text{ gaps in a row}) = \text{Cost}(\text{gap open}) + (k-1)^* \text{Cost}(\text{gap})$$

$$= \sigma + (k-1) \cdot \varepsilon$$

Gap opening penalty is high and gap extension penalty is low (once we start a gap we might as well pile more gaps on top)

and Think

How are the following alignments penalized using the new affine gap penalties?

```
GATCCAG
GA-C-AG
```

```
GATCCAG
GA--CAG
```

and Think

How are the following alignments penalized using the new affine gap penalties?

Cost(k gaps in a row) = Cost(gap open) + (k-1)* Cost(gap)

$= \sigma + (k-1) \cdot \varepsilon$

GATCCAG

GA–C–AG

$= \sigma + (1-1) \cdot \varepsilon + \sigma + (1-1) \cdot \varepsilon$

$= 2\sigma$

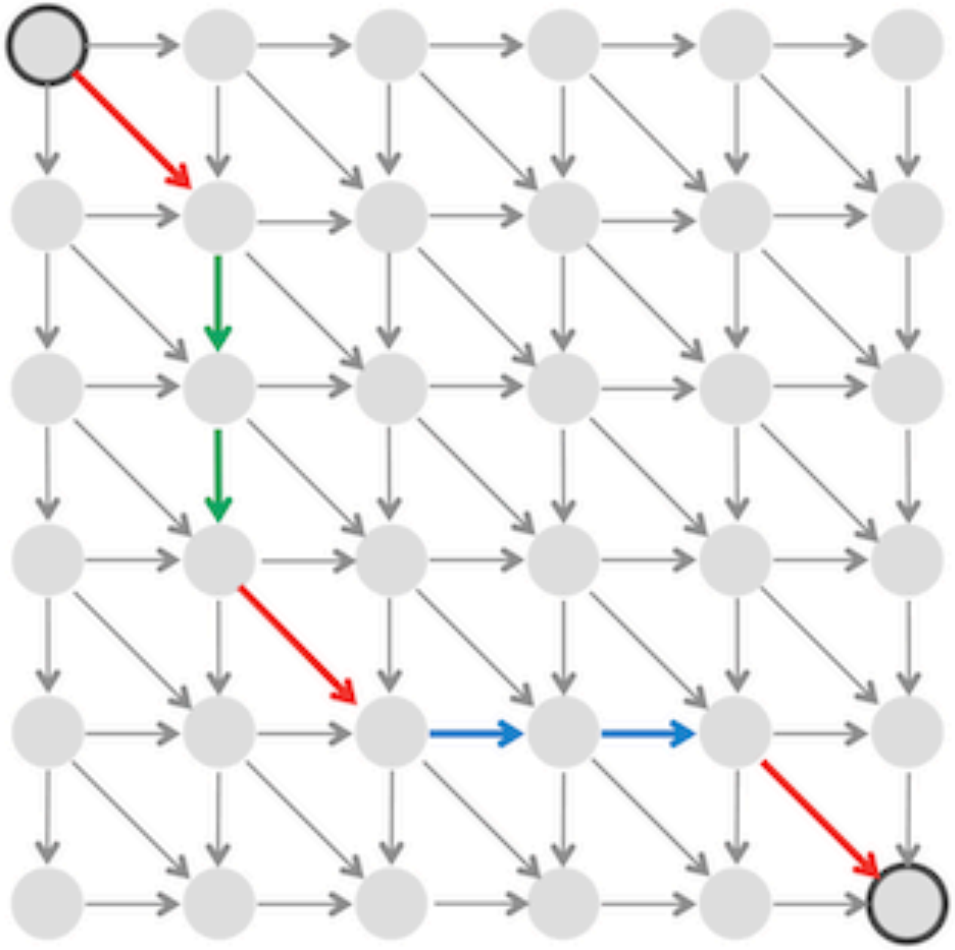
GATCCAG

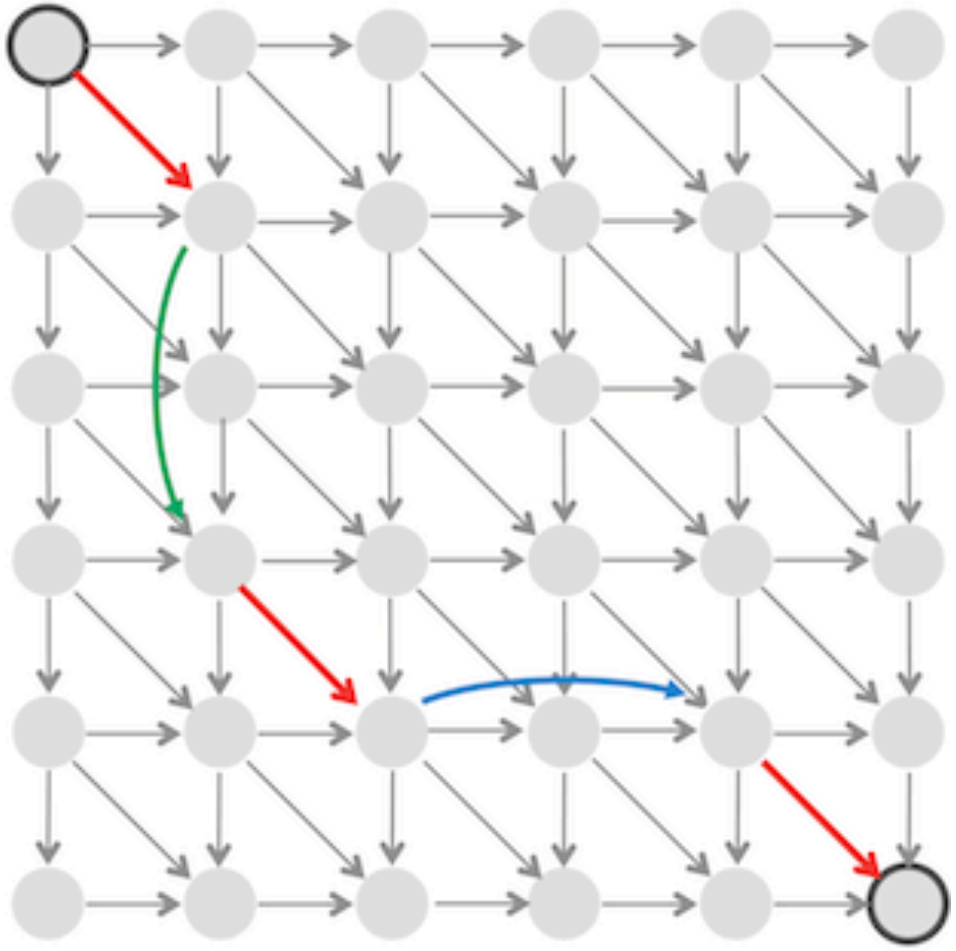
GA––CAG

$= \sigma + (2-1) \cdot \varepsilon$

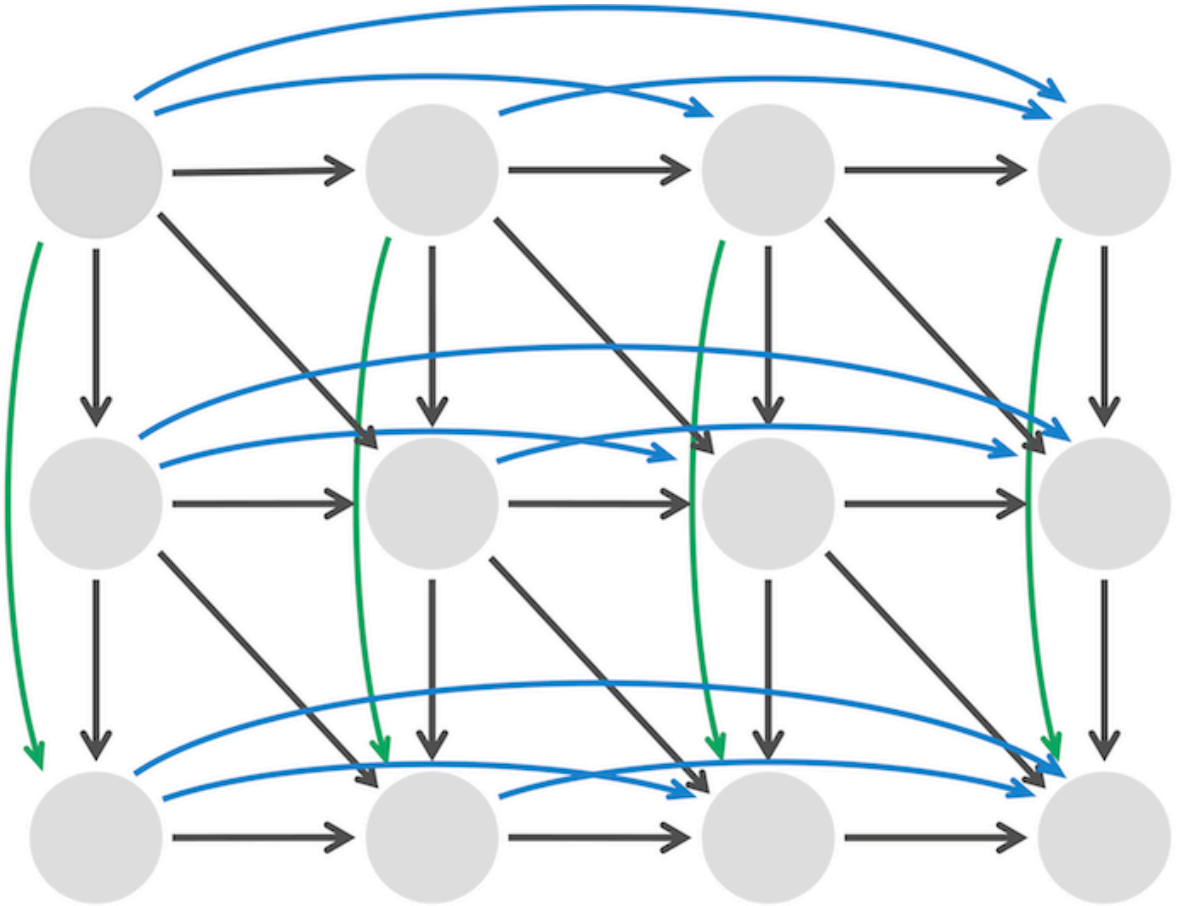
$= \sigma + \varepsilon$

Adding Affine Gap Penalties to the Alignment Graph

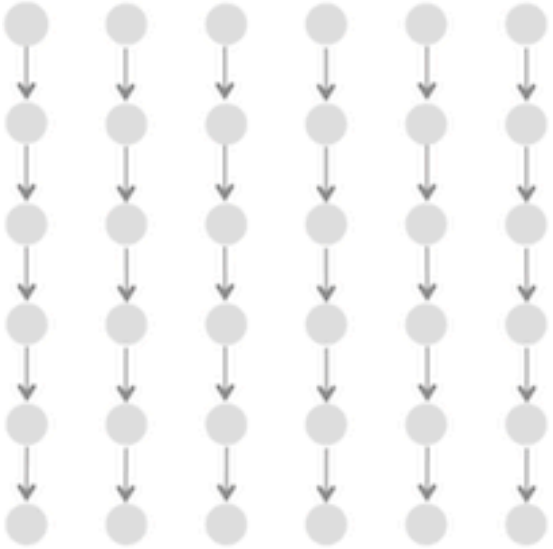


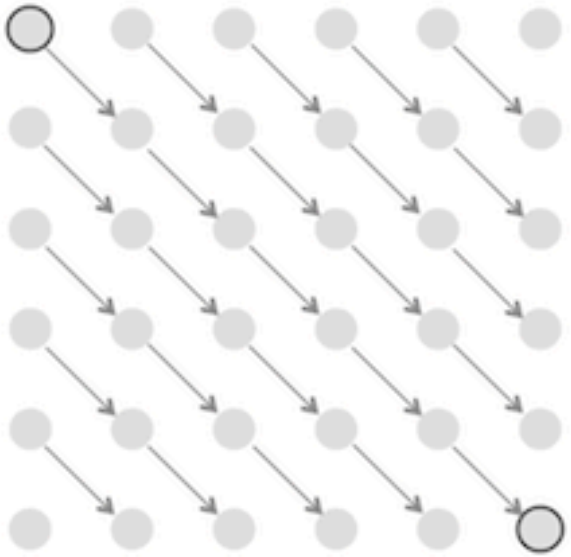


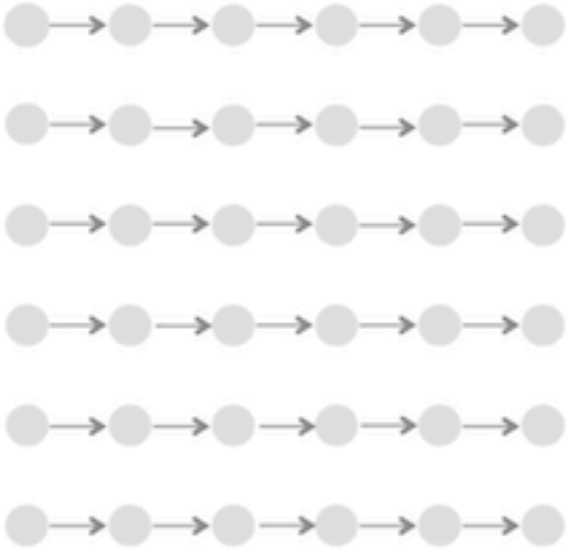
Adding Affine Gap Penalties to the Alignment Graph



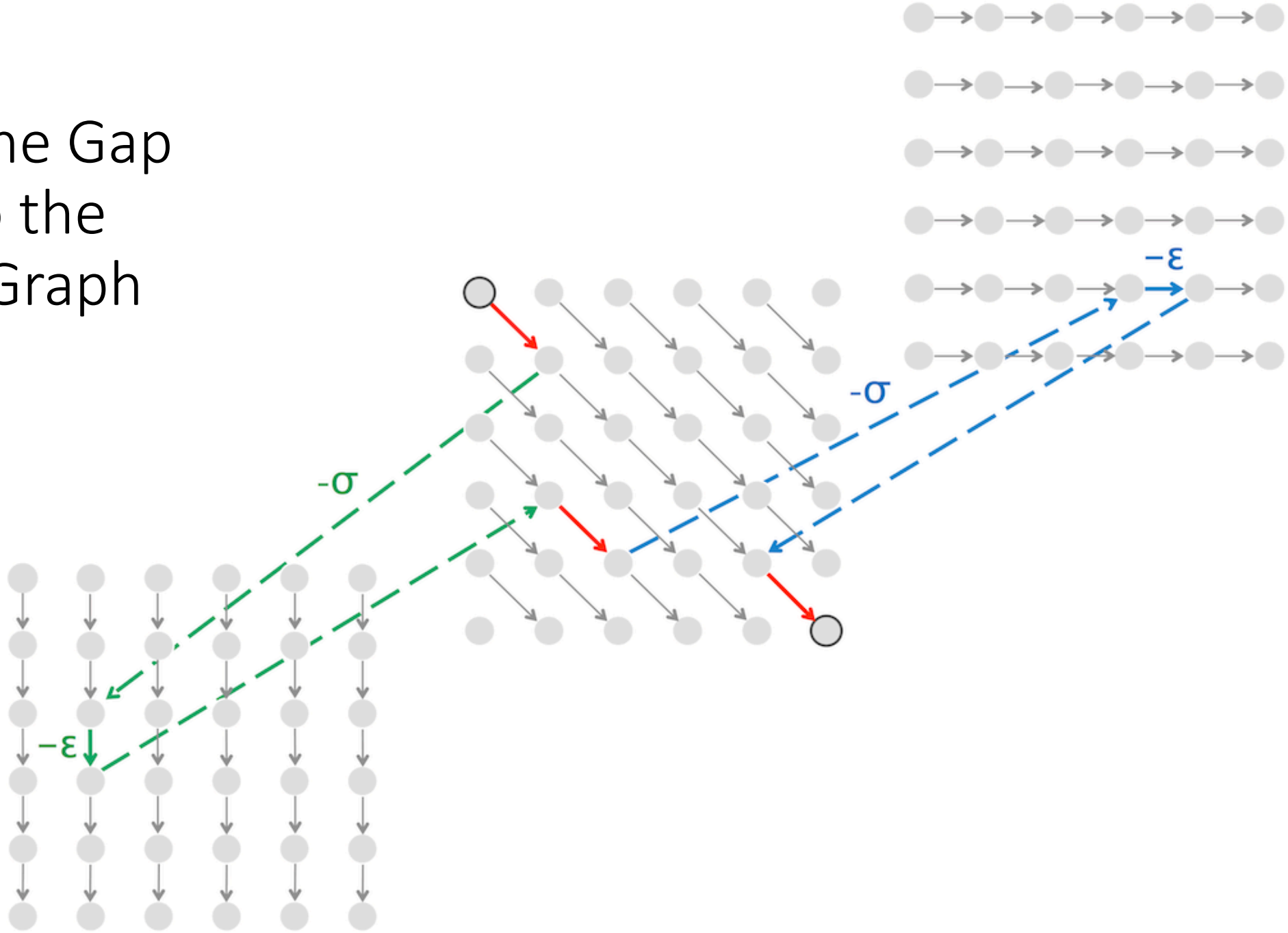
Adding Affine Gap Penalties to the Alignment Graph







Adding Affine Gap Penalties to the Alignment Graph



Where do the alignment scores come from?

- PAM matrices –
  - PAM1 – based on frequency of mutations between closely related proteins (within 1 "evolutionary step")
  - PAM 2 - ... within 2 evolutionary steps – ...
  - PAM 250 – commonly used
- BLOSUM matrices – Frequency of mutations between proteins that are x% similar
  - BLOSUM100 – based on proteins that are exactly the same (e.g. score(A,A) is defined but not score(A,G))
  - BLOSUM62 – commonly used
- gap scores usually determined empirically

BLOSUM62

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	−1	5																		
Asn	−2	0	6																	
Asp	−2	−2	1	6																
Cys	0	−3	−3	−3	9															
Gln	−1	1	0	0	−3	5														
Glu	−1	0	0	2	−4	2	5													
Gly	0	−2	0	−1	−3	−2	−2	6												
His	−2	0	1	−1	−3	0	0	−2	8											
Ile	−1	−3	−3	−3	−1	−3	−3	−4	−3	4										
Leu	−1	−2	−3	−4	−1	−2	−3	−4	−3	2	4									
Lys	−1	2	0	−1	−3	1	1	−2	−1	−3	−2	5								
Met	−1	−1	−2	−3	−1	0	−2	−3	−2	1	2	−1	5							
Phe	−2	−3	−3	−3	−2	−3	−3	−3	−1	0	0	−3	0	6						
Pro	−1	−2	−2	−1	−3	−1	−1	−2	−2	−3	−3	−1	−2	−4	7					
Ser	1	−1	1	0	−1	0	0	0	−1	−2	−2	0	−1	−2	−1	4				
Thr	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	1	5			
Trp	−3	−3	−4	−4	−2	−2	−3	−2	−2	−3	−2	−3	−1	1	−4	−3	−2	11		
Tyr	−2	−2	−2	−3	−2	−1	−2	−3	2	−1	−1	−2	−1	3	−3	−2	−2	2	7	
Val	0	−3	−3	−3	−1	−2	−2	−3	−3	3	1	−2	1	−1	−2	−2	0	−3	−1	4

https://en.wikipedia.org/wiki/BLOSUM#/media/File:BLOSUM62.png