

In the name of the most high

Introduction to Bioinformatics

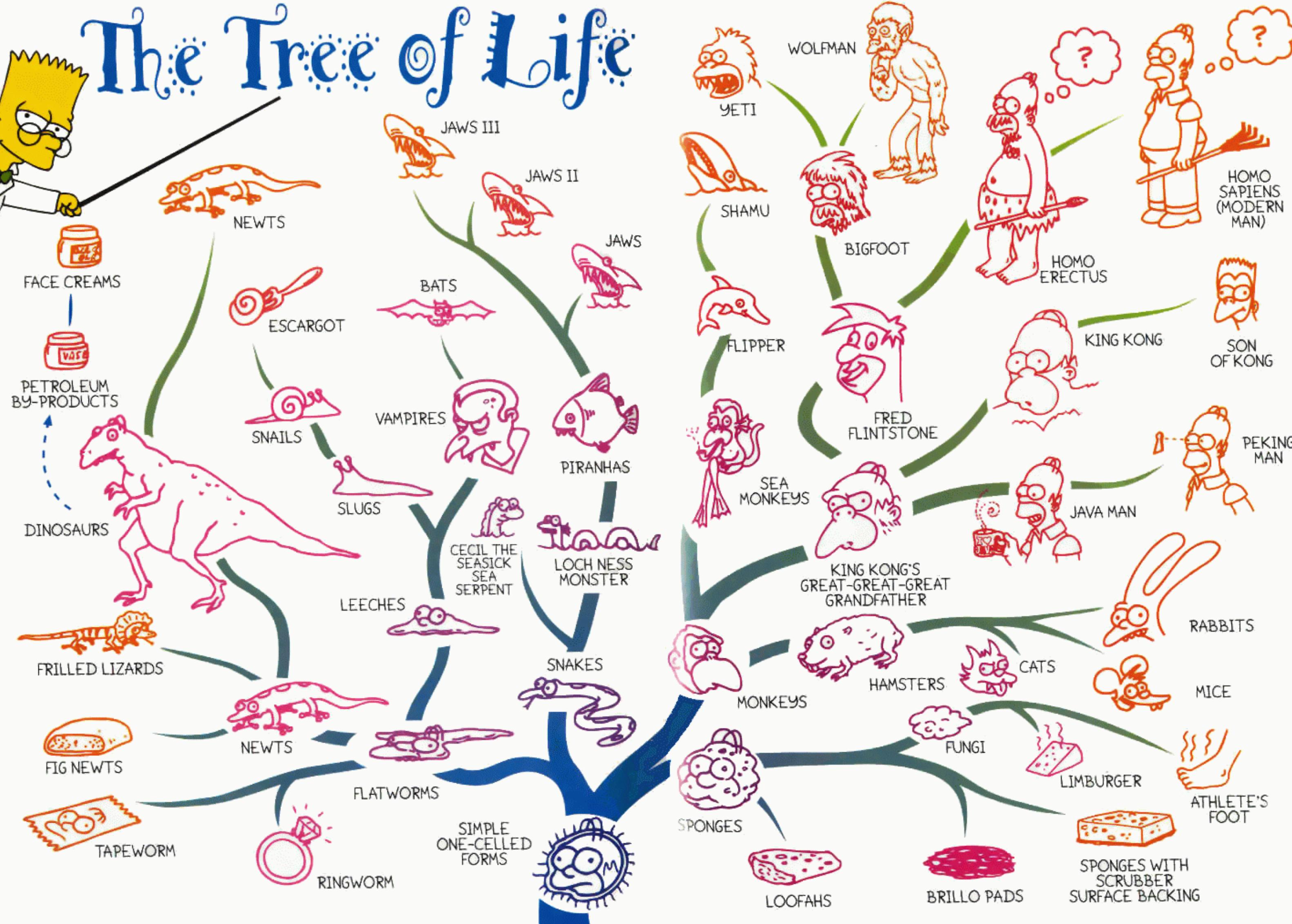
Pairwise Alignment

Ali Sharifi-Zarchi

Department of Computer Engineering, Sharif University of Technology

These slides are available under the Creative Commons Attribution License.

The Tree of Life



Pairwise Sequence Alignment

| | |
|---|--|
| BHB469831 gi_156991577 gb_EU12/1-457 2HU0_B PDBID CHAIN SEQUENCE/1-387 | 1 MNPNQKIIITIGSISIAIGIISLILQIGNIISIWASHSIQTGSQNHTGICNQR 52 |
| BHB469831 gi_156991577 gb_EU12/1-457 2HU0_B PDBID CHAIN SEQUENCE/1-387 | 53 IITYENSTWVNNTYVNINNTNVVAEKDKTSVTLAGNSSLCSISGWAIYTAKDN 104 1VKLAGNSSLCPINGWAVYSKDN 22 |
| BHB469831 gi_156991577 gb_EU12/1-457 2HU0_B PDBID CHAIN SEQUENCE/1-387 | 105 SIRIGSKGDVFVIREPFISCSHLECRTFFLTQGALLNDKHSNGTVKDRSPYR 156 23 SIRIGSKGDVFVIREPFISCSHLECRTFFLTQGALLNDKHSNGTVKDRSPHR 74 |
| BHB469831 gi_156991577 gb_EU12/1-457 2HU0_B PDBID CHAIN SEQUENCE/1-387 | 157 TLMSCPLGEAPSPYNSRFESVAWSASACHDGMGWLTIGISGPDNGAVAVLKY 208 75 TLMSCPVGEAPSPYNSRFESVAWSASACHDGTSWLTIGISGPDNGAVAVLKY 126 |
| BHB469831 gi_156991577 gb_EU12/1-457 2HU0_B PDBID CHAIN SEQUENCE/1-387 | 209 NGIITETIKSWKKRILRTQESECVCMNGSCFTIMTDGPSNGAASYKIFKIEK 260 127 NGIITDTIKSWRNNILRTQESECACVNGSCFTVMTDGPSNGQASYKIFKMEK 178 |
| BHB469831 gi_156991577 gb_EU12/1-457 2HU0_B PDBID CHAIN SEQUENCE/1-387 | 261 GKVTKTIELNAPNFHYEECSCYPDTGTVMCMVCRDNWHGSNRPWVSFNQNLDY 312 179 GKVVKSVELDAPNYHYEECSCYPNAGEITCVCRDNDWHSNRPWVSFNQNLEY 230 |
| BHB469831 gi_156991577 gb_EU12/1-457 2HU0_B PDBID CHAIN SEQUENCE/1-387 | 313 QIGYICSGGVFGDNPRPKDGEGLSCNPVTVVDGADGVKGFSYKYGNGVWIGRTKS 364 231 QIGYICSGGVFGDNPRPNDGTGSCGPVSSNGAYGVKGFSFKYGNGVWIGRTKS 282 |
| BHB469831 gi_156991577 gb_EU12/1-457 2HU0_B PDBID CHAIN SEQUENCE/1-387 | 365 NRLRKGFEMIWDPNGWTNTDSDFSVKQDVVAITDWSGYSGSFVQHPELTGLD 416 283 TNSRSGFEMIWDPNGWTETDSSFSVKQDIVAITDWSGYSGSFVQHPELTGLD 334 |
| BHB469831 gi_156991577 gb_EU12/1-457 2HU0_B PDBID CHAIN SEQUENCE/1-387 | 417 CIRPCFWVELVRGLPRENTTIWTSRSSISFCGVNSGTANWS----- 457 335 CIRPCFWVELIRGRPKESTIWTSGSSISFCGVNSDTVGWSWPDGAEELPFTI 385 |
| BHB469831 gi_156991577 gb_EU12/1-457 | .. |

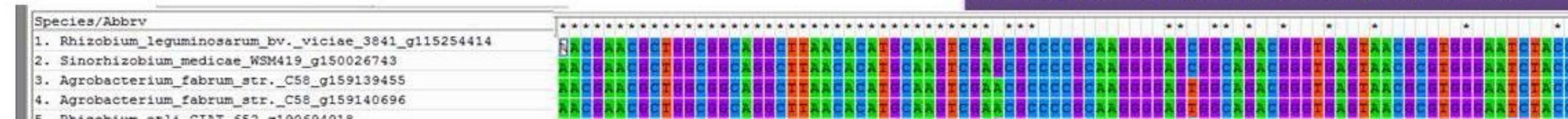
Global, Local, Multiple

| | |
|-----------------|---|
| Target Sequence | 5' ACTACTAGATTACTTACGGATCAGGTACTTAGAGGCTTGCAACCA 3' |
| | |
| Query Sequence | 5' TACTCACGGATGAGGTACTTAGAGGC 3' |

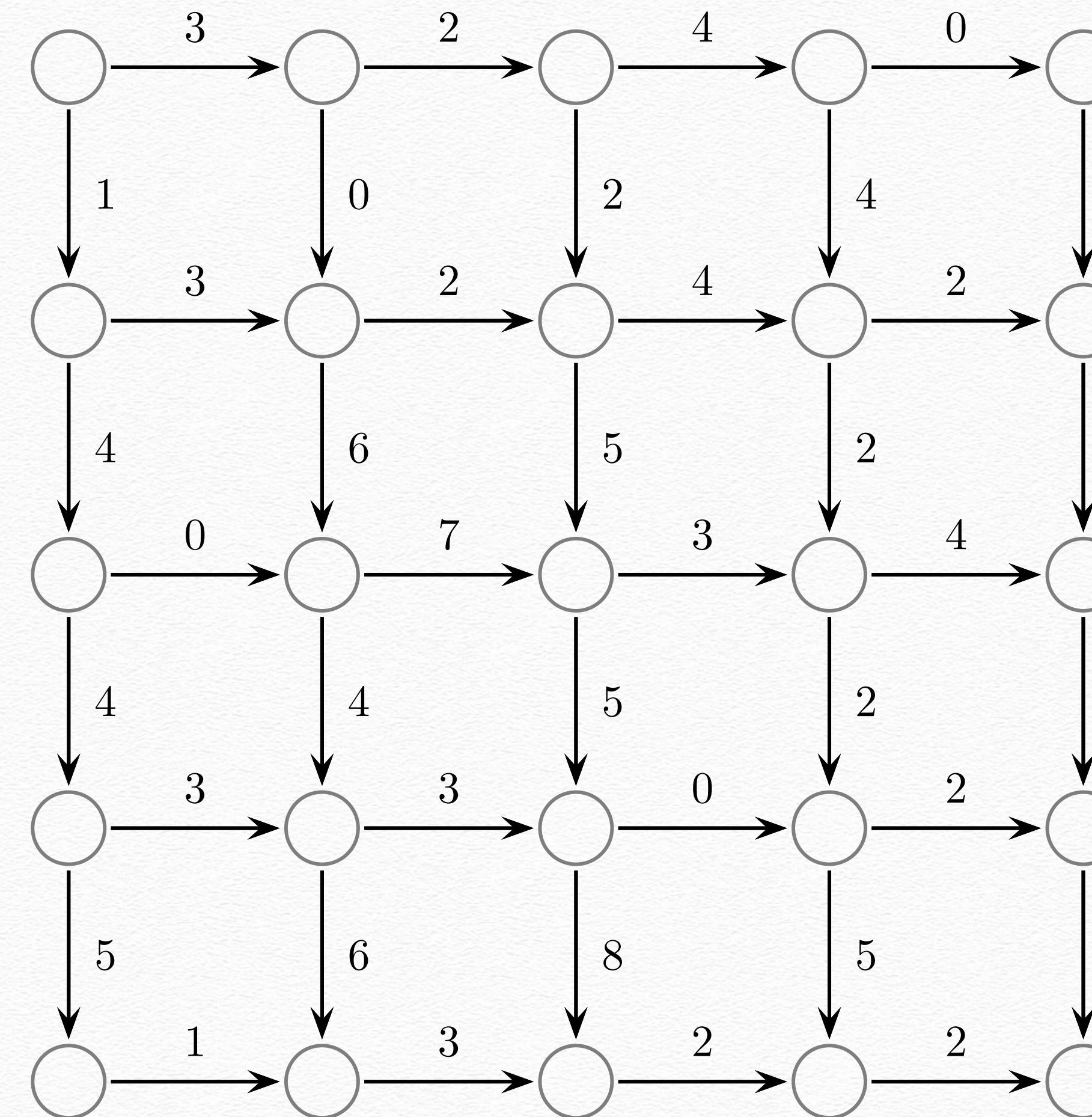
Global Alignment

| | |
|-----------------|---|
| Target Sequence | 5' ACTACTAGATTACTTACGGATCAGGTACTTAGAGGCTTGCAACCA 3' |
| | |
| Query Sequence | 5' ACTACTAGATT----ACGGATC--GTACTTAGAGGCTAGCAACCA 3' |

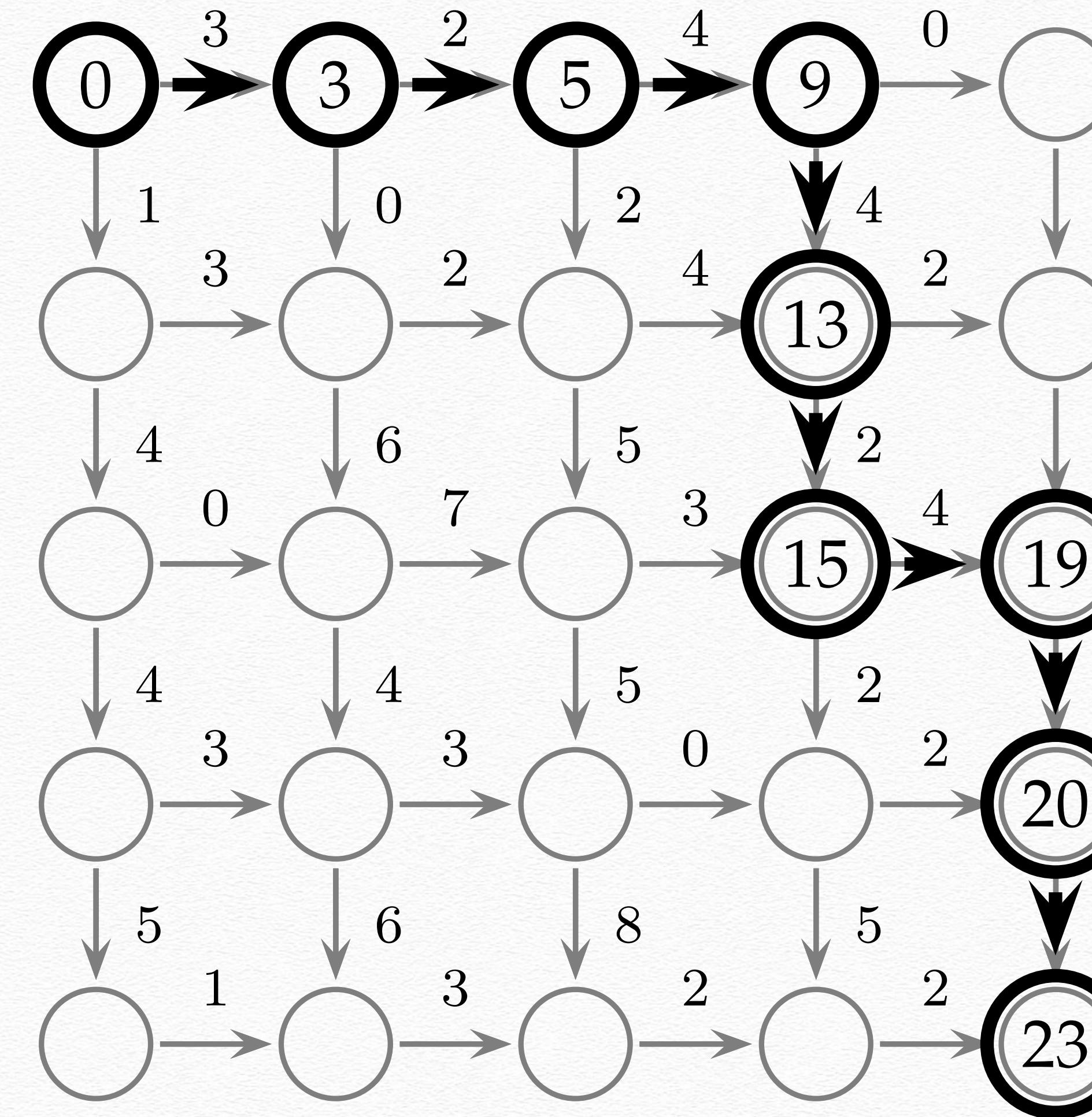
Multiple Sequence Alignment (MSA)



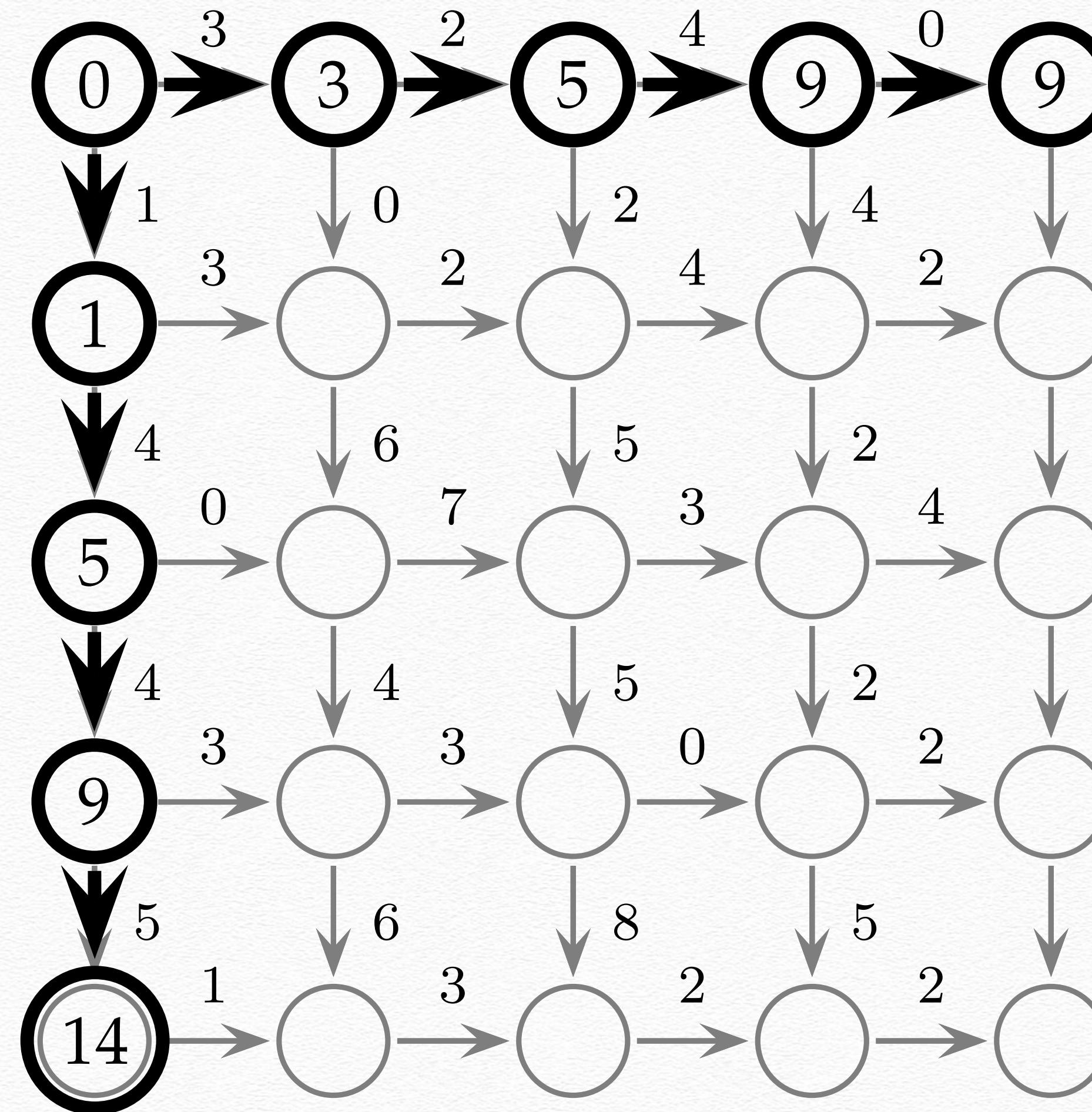
The Manhattan Tourist Problem



The Manhattan Tourist Problem



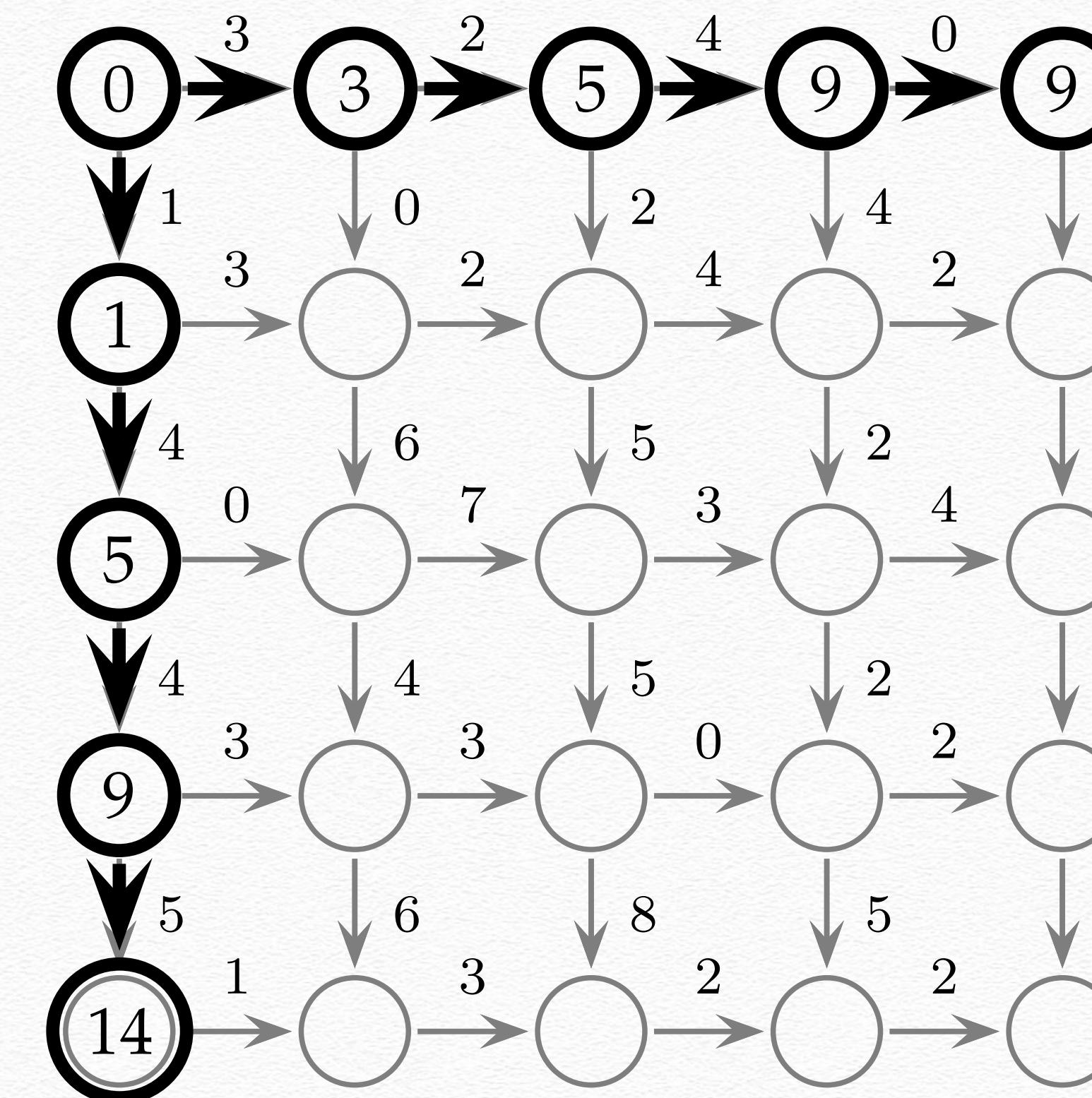
The Manhattan Tourist Problem



The Manhattan Tourist Problem

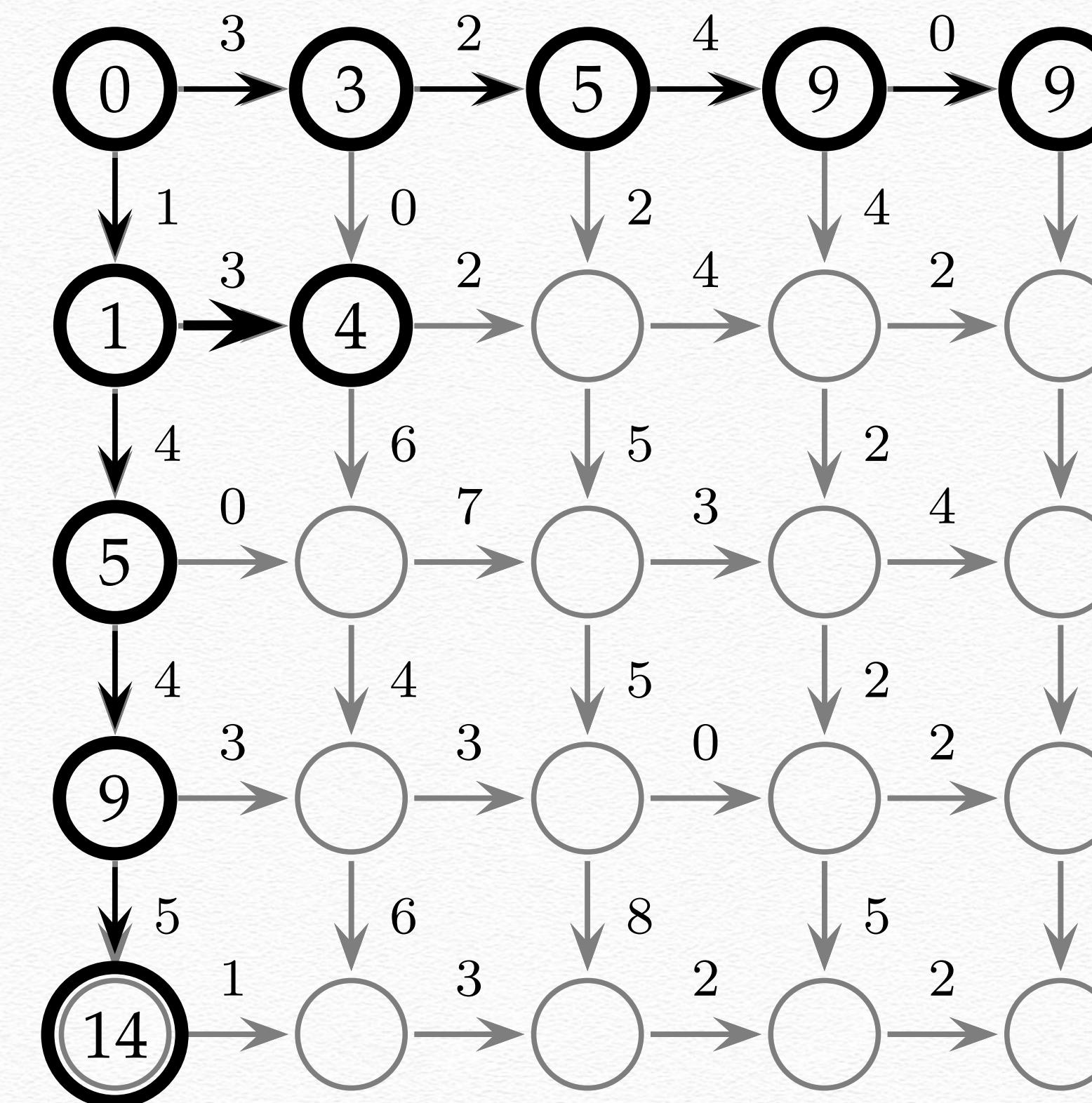
$s_{0,1} +$ weight of the edge (block) between (0,1) and (1,1);

$s_{1,0} +$ weight of the edge (block) between (1,0) and (1,1).

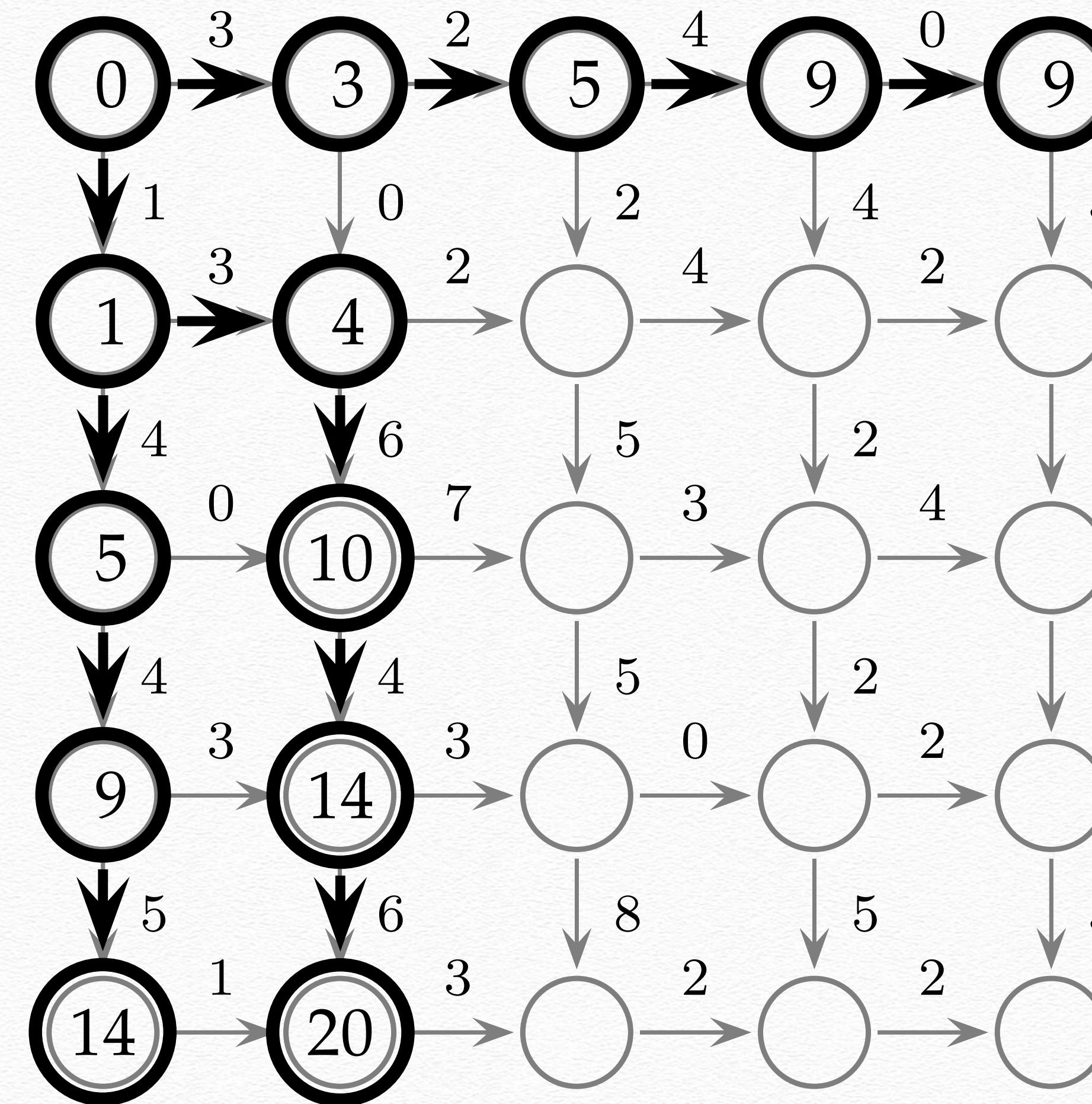


The Manhattan Tourist Problem

$$s_{1,2} = \max \left\{ \begin{array}{l} s_{1,1} + \text{ weight of the edge between } (1,1) \text{ and } (1,2) \\ s_{0,2} + \text{ weight of the edge between } (0,2) \text{ and } (1,2) \end{array} \right.$$

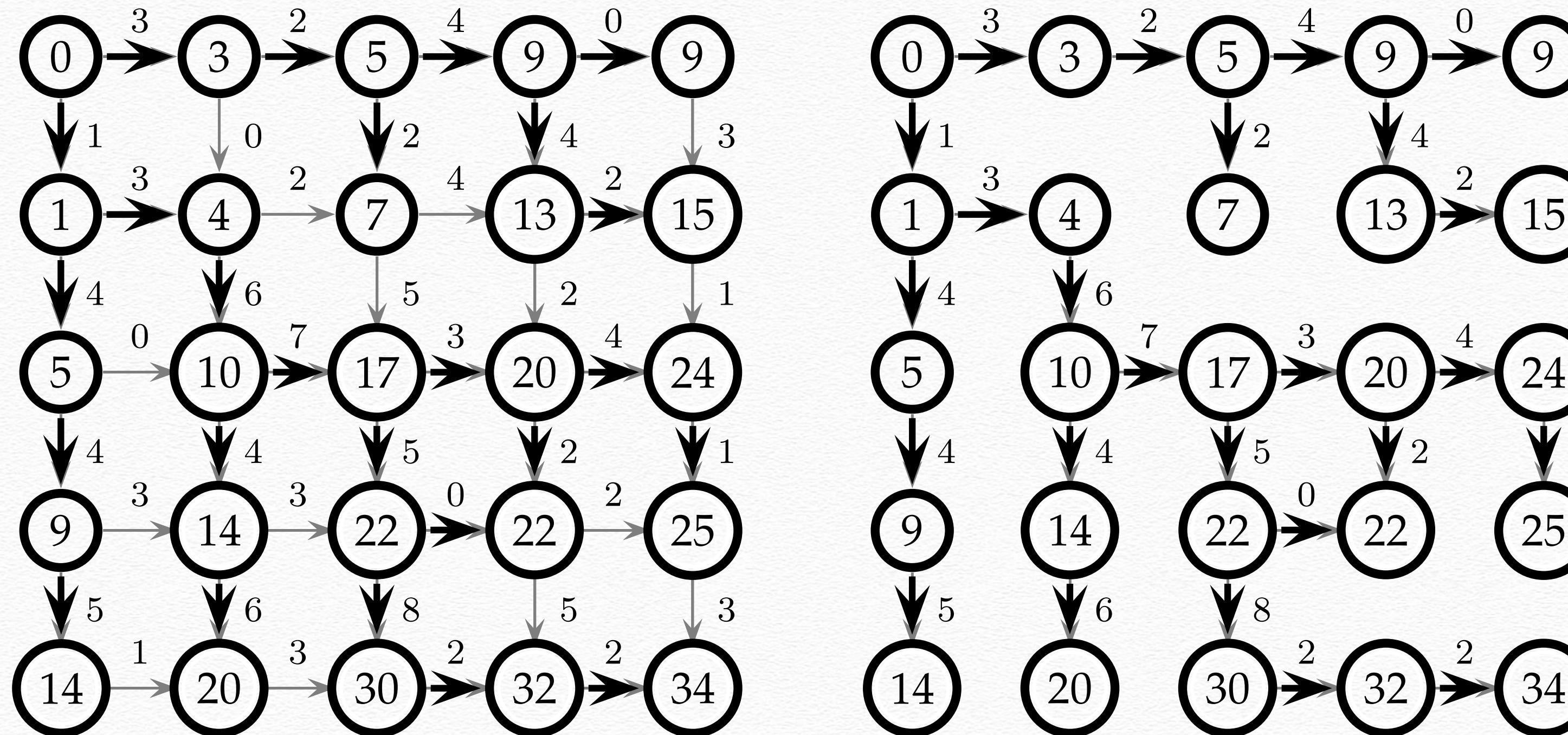


The Manhattan Tourist Problem



Dynamic Programming

$$s_{i,j} = \max \begin{cases} s_{i-1,j} + \text{ weight of the edge between } (i-1, j) \text{ and } (i, j) \\ s_{i,j-1} + \text{ weight of the edge between } (i, j-1) \text{ and } (i, j) \end{cases}$$

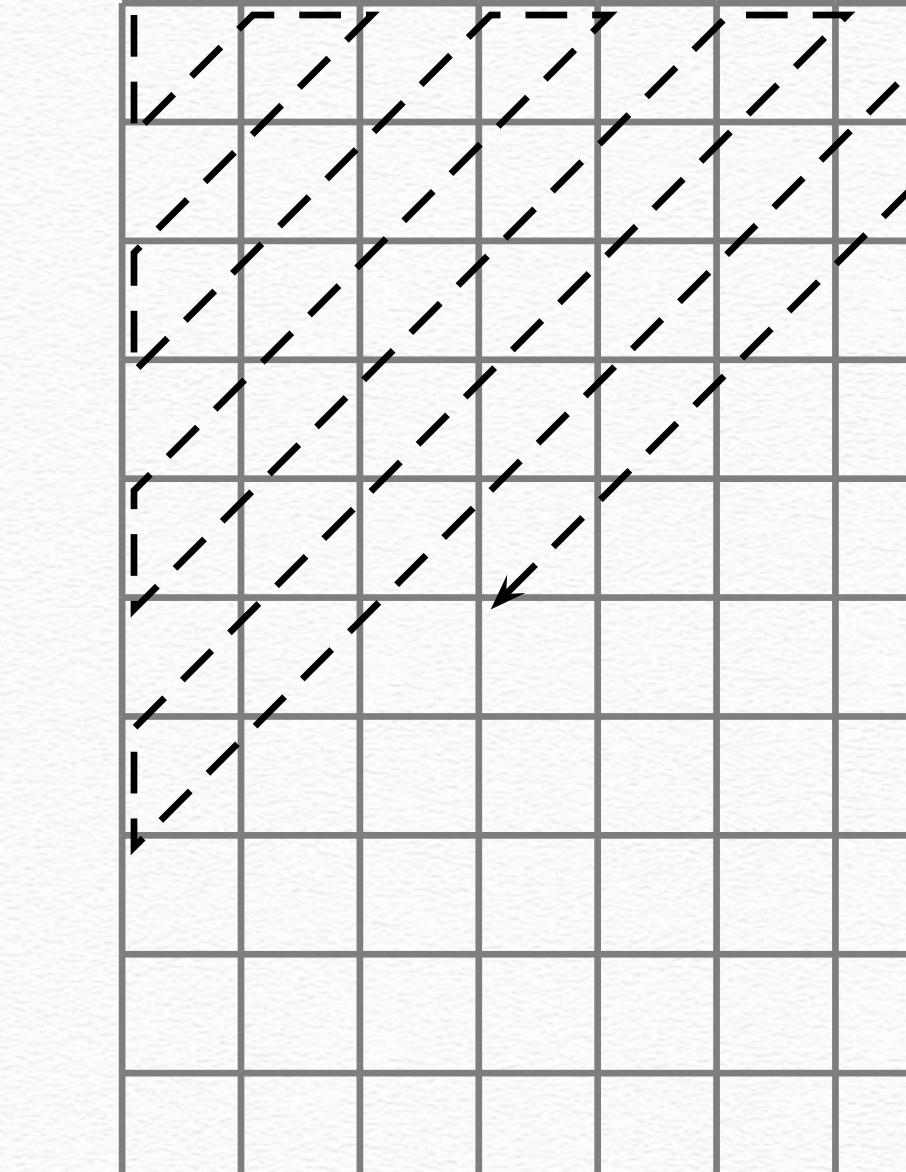
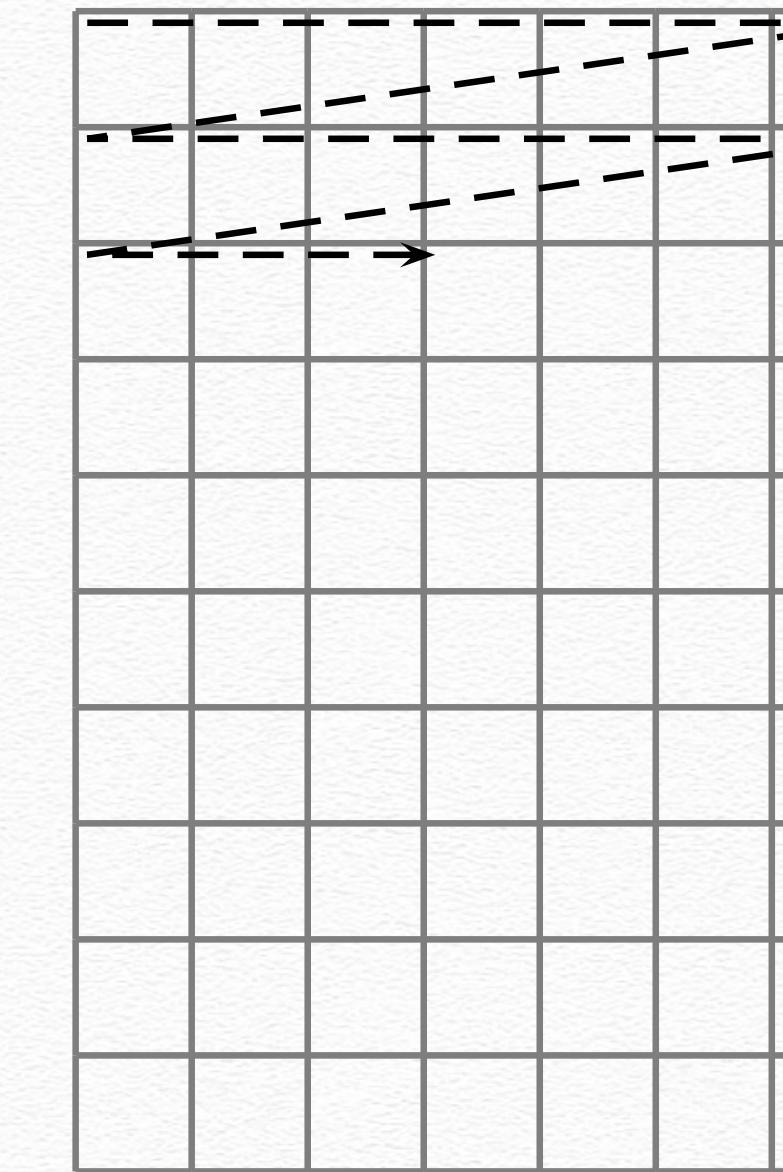
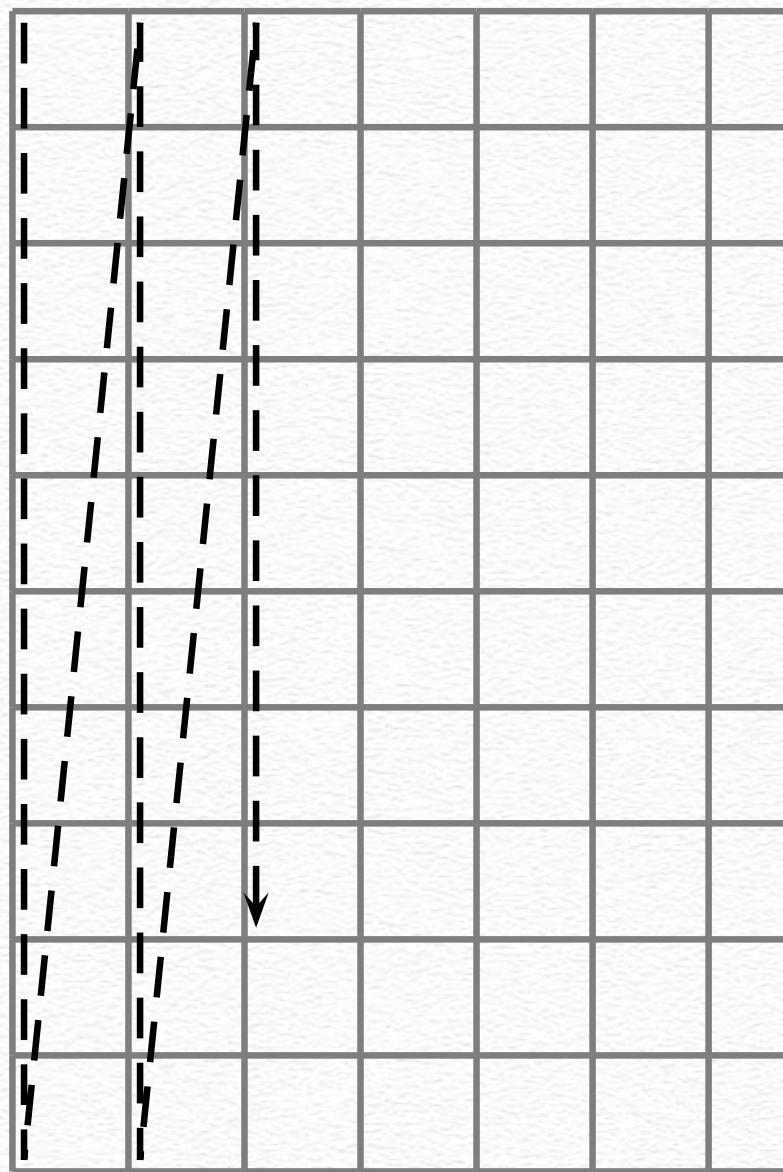


Dynamic Programming

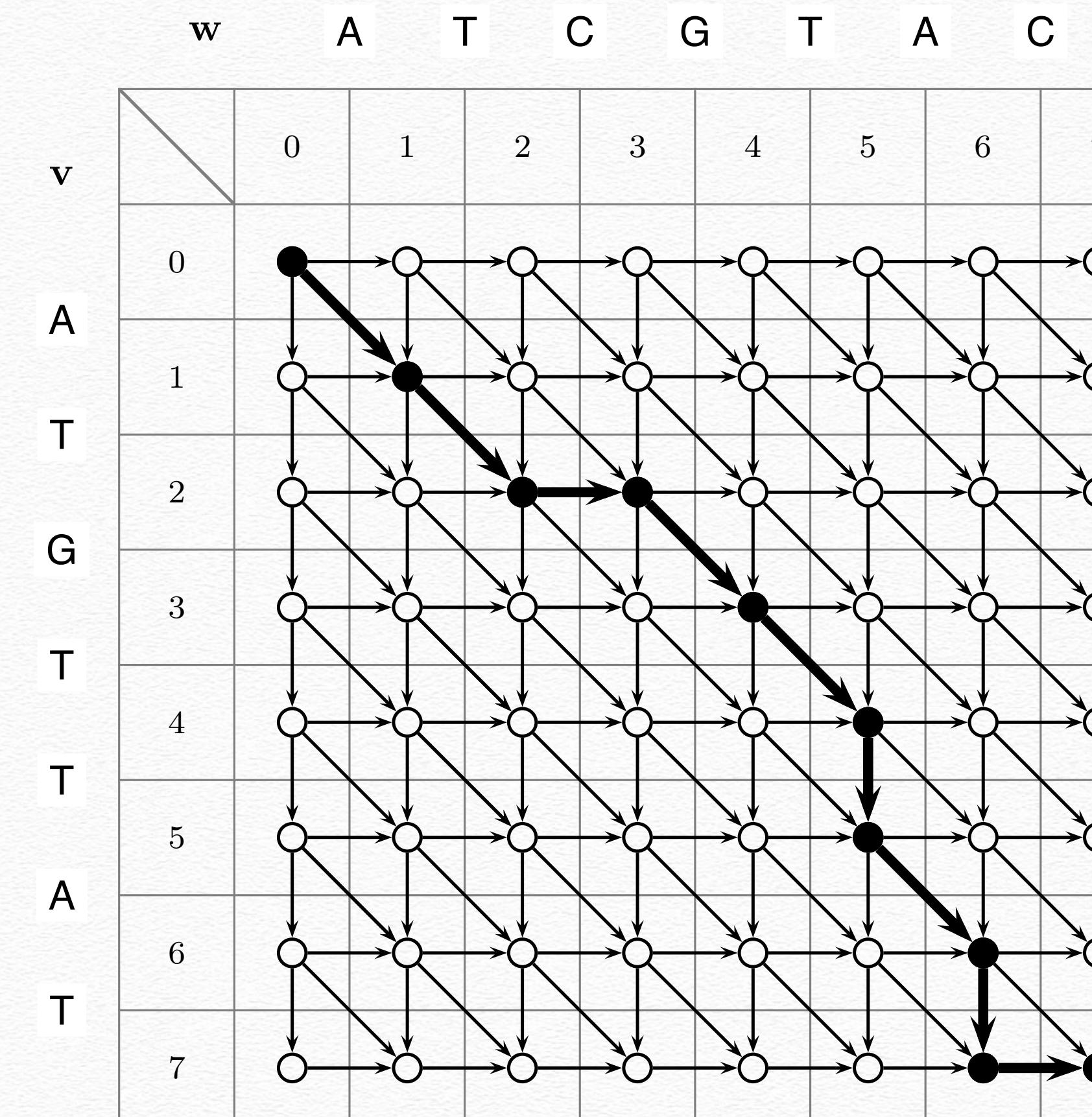
MANHATTANTOURIST($\downarrow \vec{\mathbf{w}}, \vec{\mathbf{w}}, n, m$)

```
1   $s_{0,0} \leftarrow 0$ 
2  for  $i \leftarrow 1$  to  $n$ 
3       $s_{i,0} \leftarrow s_{i-1,0} + \downarrow \vec{w}_{i,0}$ 
4  for  $j \leftarrow 1$  to  $m$ 
5       $s_{0,j} \leftarrow s_{0,j-1} + \vec{w}_{0,j}$ 
6  for  $i \leftarrow 1$  to  $n$ 
7      for  $j \leftarrow 1$  to  $m$ 
8           $s_{i,j} \leftarrow \max \begin{cases} s_{i-1,j} + \downarrow \vec{w}_{i,j} \\ s_{i,j-1} + \vec{w}_{i,j} \end{cases}$ 
9  return  $s_{n,m}$ 
```

Filling DP Table



Longest Common Subsequence



↓ → ↓ → ↓ →
A T - G T T A T -
A T C G T - A - C

LCS

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

LCS

$\text{LCS}(\mathbf{v}, \mathbf{w})$

```
1  for  $i \leftarrow 0$  to  $n$ 
2       $s_{i,0} \leftarrow 0$ 
3  for  $j \leftarrow 1$  to  $m$ 
4       $s_{0,j} \leftarrow 0$ 
5  for  $i \leftarrow 1$  to  $n$ 
6      for  $j \leftarrow 1$  to  $m$ 
```

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

```
8           $b_{i,j} \leftarrow \begin{cases} "\uparrow" & \text{if } s_{i,j} = s_{i-1,j} \\ "\leftarrow" & \text{if } s_{i,j} = s_{i,j-1} \\ "\nwarrow", & \text{if } s_{i,j} = s_{i-1,j-1} + 1 \end{cases}$ 
```

```
9  return  $(s_{n,m}, \mathbf{b})$ 
```

LCS

```
PRINTLCS(b, v, i, j)
1  if i = 0 or j = 0
2      return
3  if  $b_{i,j}$  = “↖”
4      PRINTLCS(b, v, i - 1, j - 1)
5      print  $v_i$ 
6  else
7      if  $b_{i,j}$  = “↑”
8          PRINTLCS(b, v, i - 1, j)
9      else
10         PRINTLCS(b, v, i, j - 1)
```

Global Sequence Alignment

Global Alignment Problem:

Find the best alignment between two strings under a given scoring matrix.

Input: Strings v , w and a scoring matrix δ .

Output: An alignment of v and w whose score (as defined by the matrix δ) is maximal among all possible alignments of v and w .

Needleman-Wunsch Global Alignment Algorithm

$$s_{i,j} = \max \left\{ \begin{array}{l} s_{i-1,j} + \delta(v_i, -) \\ s_{i,j-1} + \delta(-, w_j) \\ s_{i-1,j-1} + \delta(v_i, w_j) \end{array} \right.$$

Local Sequence Alignment

Local Alignment Problem:

Find the best local alignment between two strings.

Input: Strings v and w and a scoring matrix δ .

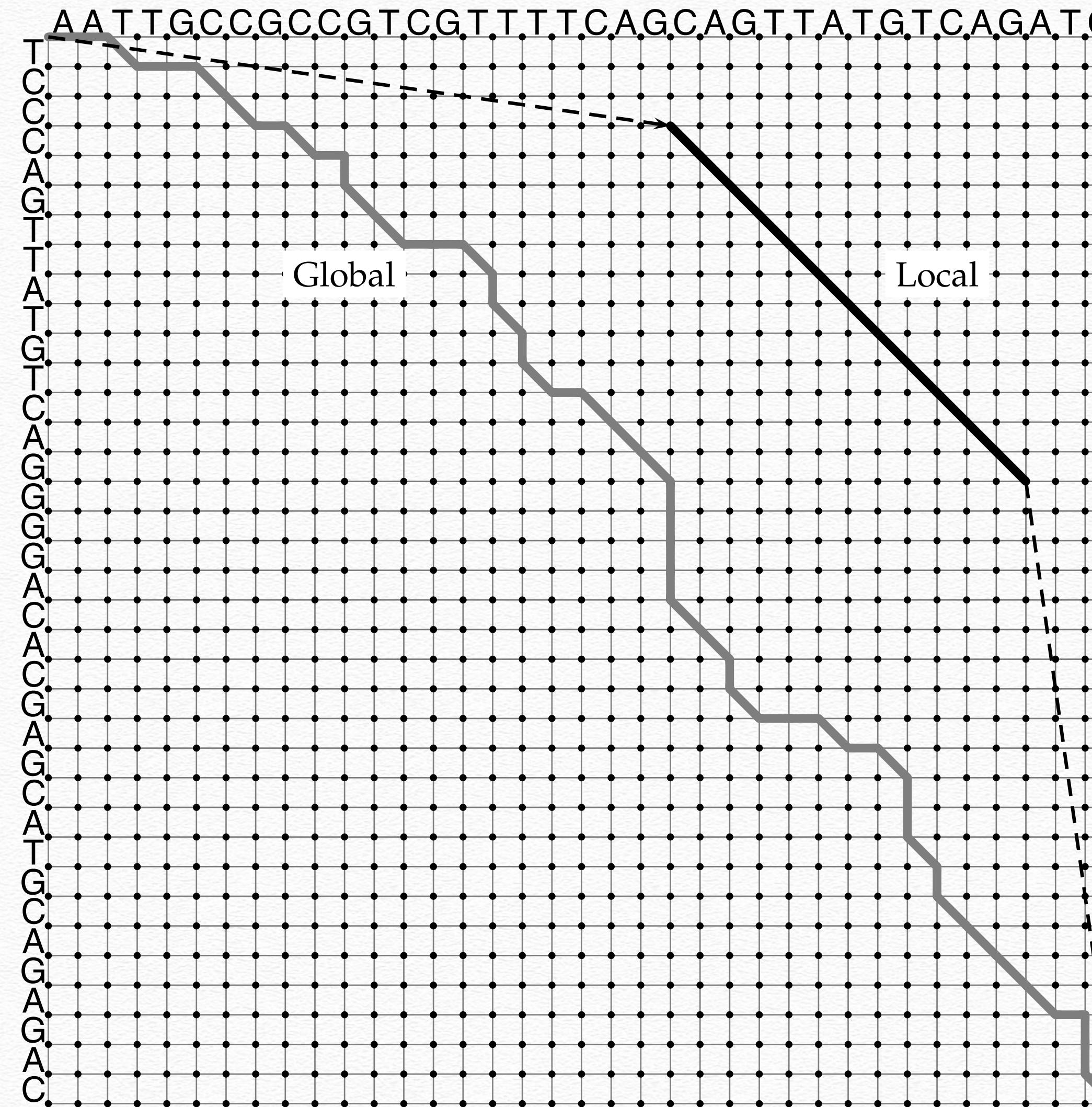
Output: Substrings of v and w whose global alignment, as defined by δ , is maximal among all global alignments of all substrings of v and w .

Global vs. Local

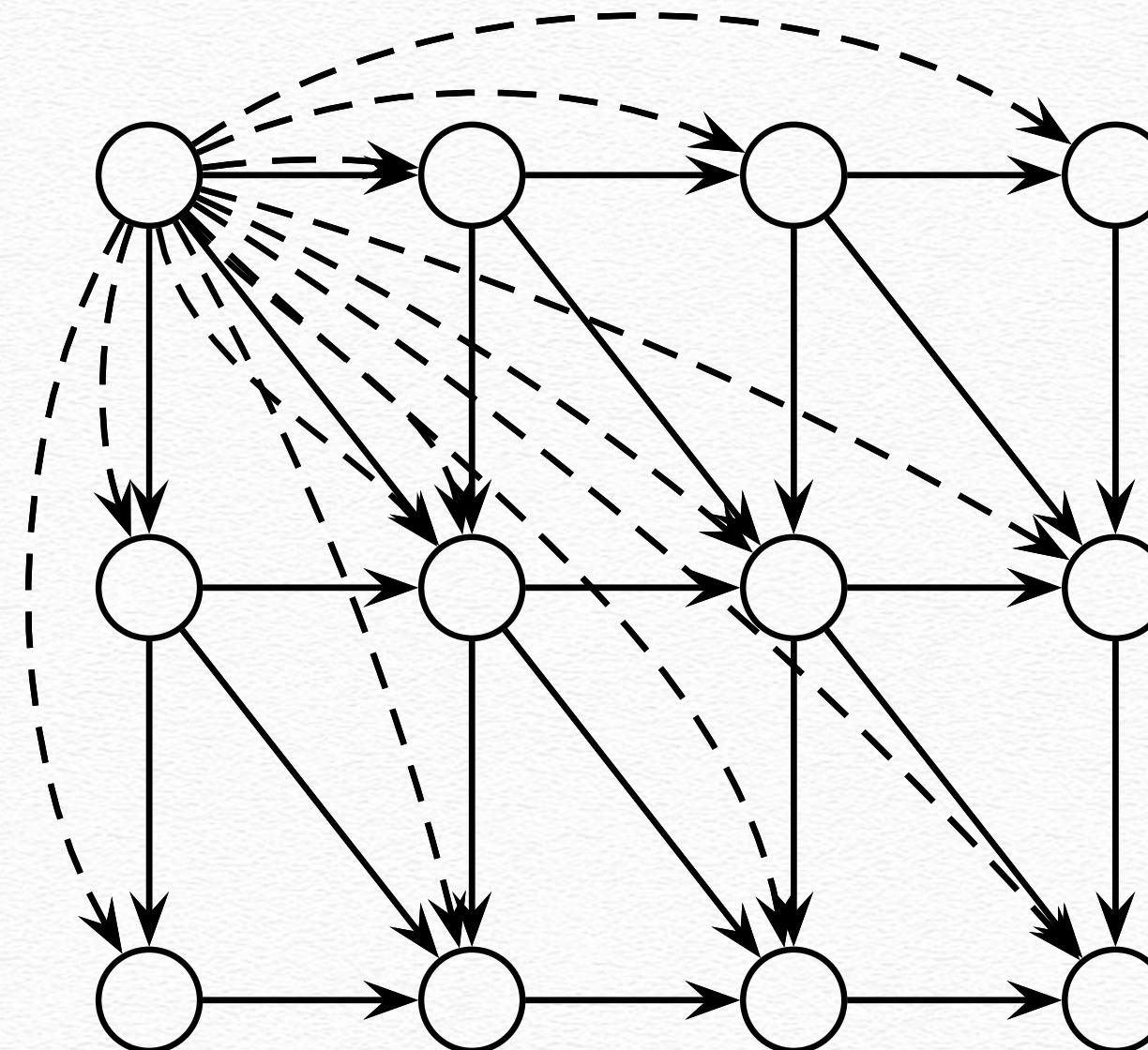
--T--CC-C-AGT--TATGT-CAGGGGACACG--A-GCATGCAGA-GAC
| | | | | | | | | | | | | | | | | | | | | | | | | | | |
AATTGCCGCC-GTCGT-T-TTCAG---CA-GTTATG--T-CAGAT--C

tccCAGTTATGTCAggacacgagcatgcagagac
| | | | | | | | |
aattgccgcgtcgtttcagCAGTTATGTCAgatc

Global vs. Local



Smith-Waterman Local Alignment Algorithm



$$s_{i,j} = \max \left\{ \begin{array}{l} 0 \\ s_{i-1,j} + \delta(v_i, -) \\ s_{i,j-1} + \delta(-, w_j) \\ s_{i-1,j-1} + \delta(v_i, w_j) \end{array} \right\}$$

Gap Penalties

$$\downarrow s_{i,j} = \max \left\{ \begin{array}{l} \downarrow s_{i-1,j} - \sigma \\ s_{i-1,j} - (\rho + \sigma) \end{array} \right.$$

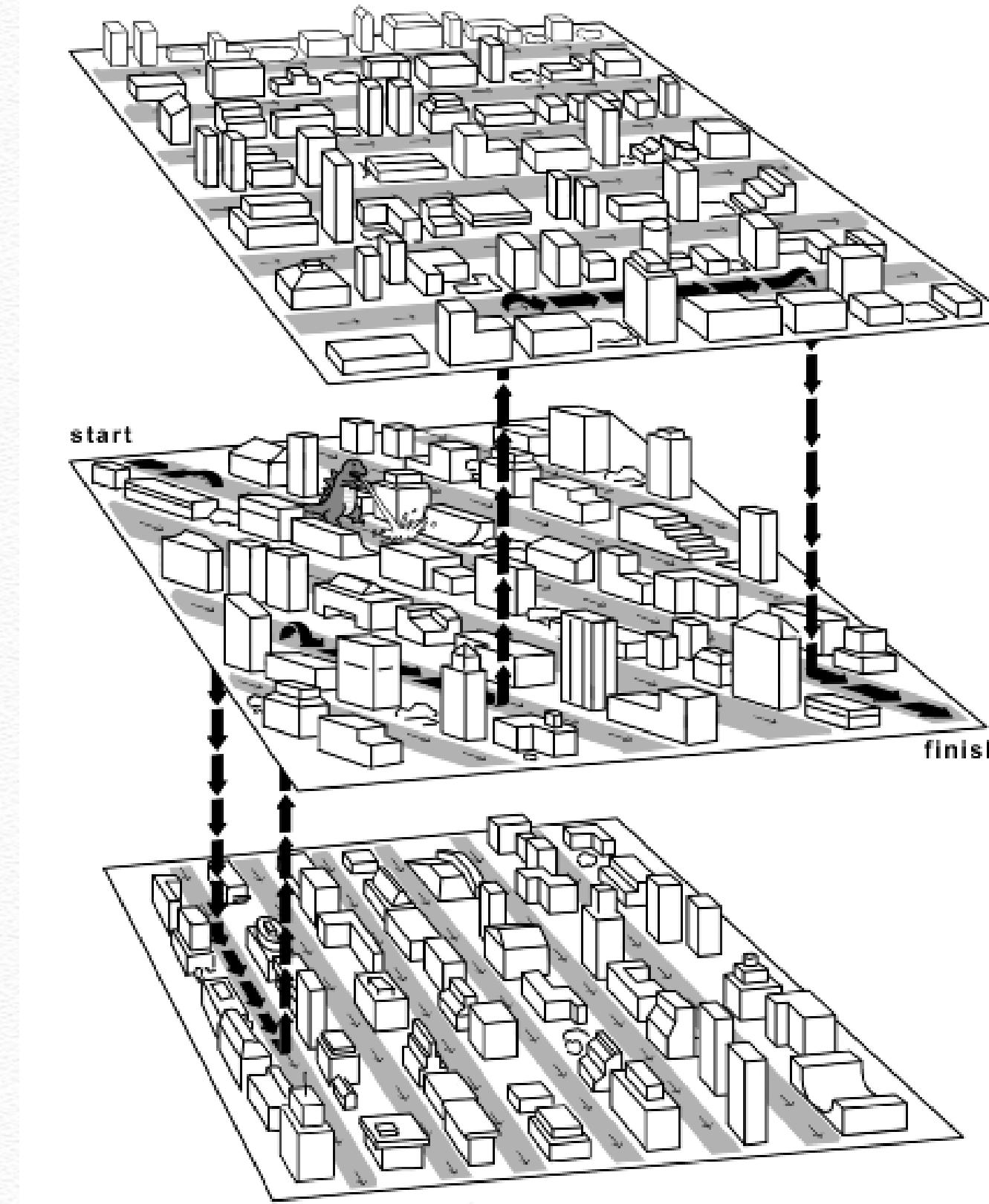
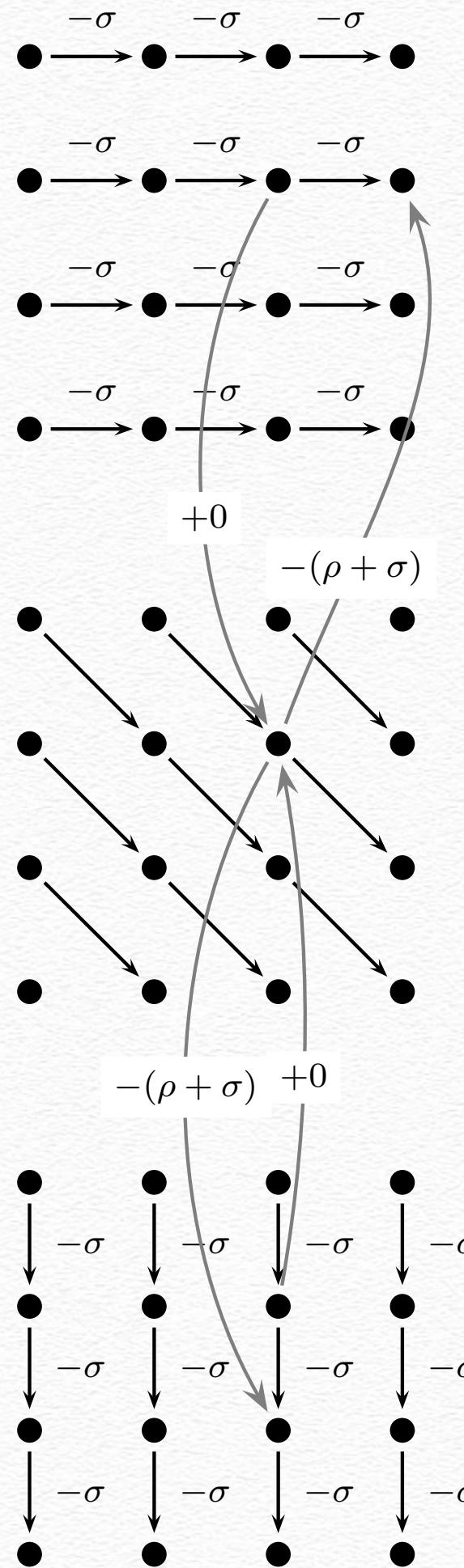
$$\rightarrow s_{i,j} = \max \left\{ \begin{array}{l} \rightarrow s_{i,j-1} - \sigma \\ s_{i,j-1} - (\rho + \sigma) \end{array} \right.$$

$$s_{i,j} = \max \left\{ \begin{array}{l} s_{i-1,j-1} + \delta(v_i, w_j) \\ \downarrow s_{i,j} \\ \rightarrow s_{i,j} \end{array} \right.$$

Triple Sequence Alignment

```
--T--CC-C-AGT--TATGT-CAGGGGACACG--A-GCATGCAGA-GAC
|   || |  || | | | | || | | | | | | | |
AATTGCCGCC-GTCGT-T-TTCAG----CA-GTTATG--T-CAGAT--C
||| | | | x | | | | | | | | | | | | |
-ATTGC-G--ATTCGTAT-----GGGACA-TGGATGCATGCAG-TGAC
```

Triple Sequence Alignment



Triple Sequence Alignment

$$s_{i,j,k} = \max \left\{ \begin{array}{ll} s_{i-1,j,k} & +\delta(v_i, -, -) \\ s_{i,j-1,k} & +\delta(-, w_j, -) \\ s_{i,j,k-1} & +\delta(-, -, u_k) \\ s_{i-1,j-1,k} & +\delta(v_i, w_j, -) \\ s_{i-1,j,k-1} & +\delta(v_i, -, u_k) \\ s_{i,j-1,k-1} & +\delta(-, w_j, u_k) \\ s_{i-1,j-1,k-1} & +\delta(v_i, w_j, u_k) \end{array} \right.$$

Triple Sequence Alignment

