

Class-10 Discussion Section

Genome 540

Chengxiang Qiu

HW 5: find multiple alignment for three sequences

- Create an edit graph for 3 sequences using the BLOSUM62 score matrix
- Run HW4 WDAG program on the edit graph to find the highest scoring path (local alignment)
- Report *in the specified format*:
 - Maximum path score for the multiple alignment
 - List of all edge weights (alphabetically sorted)
 - List of all edge counts (alphabetically sorted)
 - Highest scoring alignment

Inputs: Multiple Sequence Alignment (MSA)

```
>sp|P01308|INS_HUMAN Insulin OS=Homo sapiens GN=INS PE=1 SV=1  
MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN
```

```
>sp|P01317|INS_BOVIN Insulin OS=Bos taurus GN=INS PE=1 SV=2  
MALWTRLRPLLALLALWPPPPARAFVNQHLCGSHLVEALYLVCGERGFFYTPKARREVEGPPQVGALELAGGPGAGGLEGPPQKRGIVEQCCASVCSLYQLENYCN
```

```
>sp|P01315|INS_PIG Insulin OS=Sus scrofa GN=INS PE=1 SV=2  
MALWTRLLPLLALLALWAPAPAQAFVNQHLCGSHLVEALYLVCGERGFFYTPKARREAENPQAGAVELGGGLGGLQALALEGPPQKRGIVEQCCTSICSLYQLENYCN
```

Inputs: Multiple Sequence Alignment (MSA)

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V | B | Z | X | * |
|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| A | 4 | -1 | -2 | -2 | 0 | -1 | -1 | 0 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 0 | -3 | -2 | 0 | -2 | -1 | 0 | -4 |
| R | -1 | 5 | 0 | -2 | -3 | 1 | 0 | -2 | 0 | -3 | -2 | 2 | -1 | -3 | -2 | -1 | -1 | -3 | -2 | -3 | -1 | 0 | -1 | -4 |
| N | -2 | 0 | 6 | 1 | -3 | 0 | 0 | 0 | 1 | -3 | -3 | 0 | -2 | -3 | -2 | 1 | 0 | -4 | -2 | -3 | 3 | 0 | -1 | -4 |
| D | -2 | -2 | 1 | 6 | -3 | 0 | 2 | -1 | -1 | -3 | -4 | -1 | -3 | -3 | -1 | 0 | -1 | -4 | -3 | -3 | 4 | 1 | -1 | -4 |
| C | 0 | -3 | -3 | -3 | 9 | -3 | -4 | -3 | -3 | -1 | -1 | -3 | -1 | -2 | -3 | -1 | -1 | -2 | -2 | -1 | -3 | -3 | -2 | -4 |
| Q | -1 | 1 | 0 | 0 | -3 | 5 | 2 | -2 | 0 | -3 | -2 | 1 | 0 | -3 | -1 | 0 | -1 | -2 | -1 | -2 | 0 | 3 | -1 | -4 |
| E | -1 | 0 | 0 | 2 | -4 | 2 | 5 | -2 | 0 | -3 | -3 | 1 | -2 | -3 | -1 | 0 | -1 | -3 | -2 | -2 | 1 | 4 | -1 | -4 |
| G | 0 | -2 | 0 | -1 | -3 | -2 | -2 | 6 | -2 | -4 | -4 | -2 | -3 | -3 | -2 | 0 | -2 | -2 | -3 | -3 | -1 | -2 | -1 | -4 |
| H | -2 | 0 | 1 | -1 | -3 | 0 | 0 | -2 | 8 | -3 | -3 | -1 | -2 | -1 | -2 | -1 | -2 | -2 | 2 | -3 | 0 | 0 | -1 | -4 |
| I | -1 | -3 | -3 | -3 | -1 | -3 | -3 | -4 | -3 | 4 | 2 | -3 | 1 | 0 | -3 | -2 | -1 | -3 | -1 | 3 | -3 | -3 | -1 | -4 |
| L | -1 | -2 | -3 | -4 | -1 | -2 | -3 | -4 | -3 | 2 | 4 | -2 | 2 | 0 | -3 | -2 | -1 | -2 | -1 | 1 | -4 | -3 | -1 | -4 |
| K | -1 | 2 | 0 | -1 | -3 | 1 | 1 | -2 | -1 | -3 | -2 | 5 | -1 | -3 | -1 | 0 | -1 | -3 | -2 | -2 | 0 | 1 | -1 | -4 |
| M | -1 | -1 | -2 | -3 | -1 | 0 | -2 | -3 | -2 | 1 | 2 | -1 | 5 | 0 | -2 | -1 | -1 | -1 | -1 | 1 | -3 | -1 | -1 | -4 |
| F | -2 | -3 | -3 | -3 | -2 | -3 | -3 | -3 | -1 | 0 | 0 | -3 | 0 | 6 | -4 | -2 | -2 | 1 | 3 | -1 | -3 | -3 | -1 | -4 |
| P | -1 | -2 | -2 | -1 | -3 | -1 | -1 | -2 | -2 | -3 | -3 | -1 | -2 | -4 | 7 | -1 | -1 | -4 | -3 | -2 | -2 | -1 | -2 | -4 |
| S | 1 | -1 | 1 | 0 | -1 | 0 | 0 | 0 | -1 | -2 | -2 | 0 | -1 | -2 | -1 | 4 | 1 | -3 | -2 | -2 | 0 | 0 | 0 | -4 |
| T | 0 | -1 | 0 | -1 | -1 | -1 | -1 | -2 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 5 | -2 | -2 | 0 | -1 | -1 | 0 | -4 |
| W | -3 | -3 | -4 | -4 | -2 | -2 | -3 | -2 | -2 | -3 | -2 | -3 | -1 | 1 | -4 | -3 | -2 | 11 | 2 | -3 | -4 | -3 | -2 | -4 |
| Y | -2 | -2 | -2 | -3 | -2 | -1 | -2 | -3 | 2 | -1 | -1 | -2 | -1 | 3 | -3 | -2 | -2 | 2 | 7 | -1 | -3 | -2 | -1 | -4 |
| V | 0 | -3 | -3 | -3 | -1 | -2 | -2 | -3 | -3 | 3 | 1 | -2 | 1 | -1 | -2 | -2 | 0 | -3 | -1 | 4 | -3 | -2 | -1 | -4 |
| B | -2 | -1 | 3 | 4 | -3 | 0 | 1 | -1 | 0 | -3 | -4 | 0 | -3 | -3 | -2 | 0 | -1 | -4 | -3 | -3 | 4 | 1 | -1 | -4 |
| Z | -1 | 0 | 0 | 1 | -3 | 3 | 4 | -2 | 0 | -3 | -3 | 1 | -1 | -3 | -1 | 0 | -1 | -3 | -2 | -2 | 1 | 4 | -1 | -4 |
| X | 0 | -1 | -1 | -1 | -2 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -2 | 0 | 0 | -2 | -1 | -1 | -1 | -1 | -1 | -4 |
| * | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | 1 |

Edge:

x_1, x_2, x_3

Edge weight:

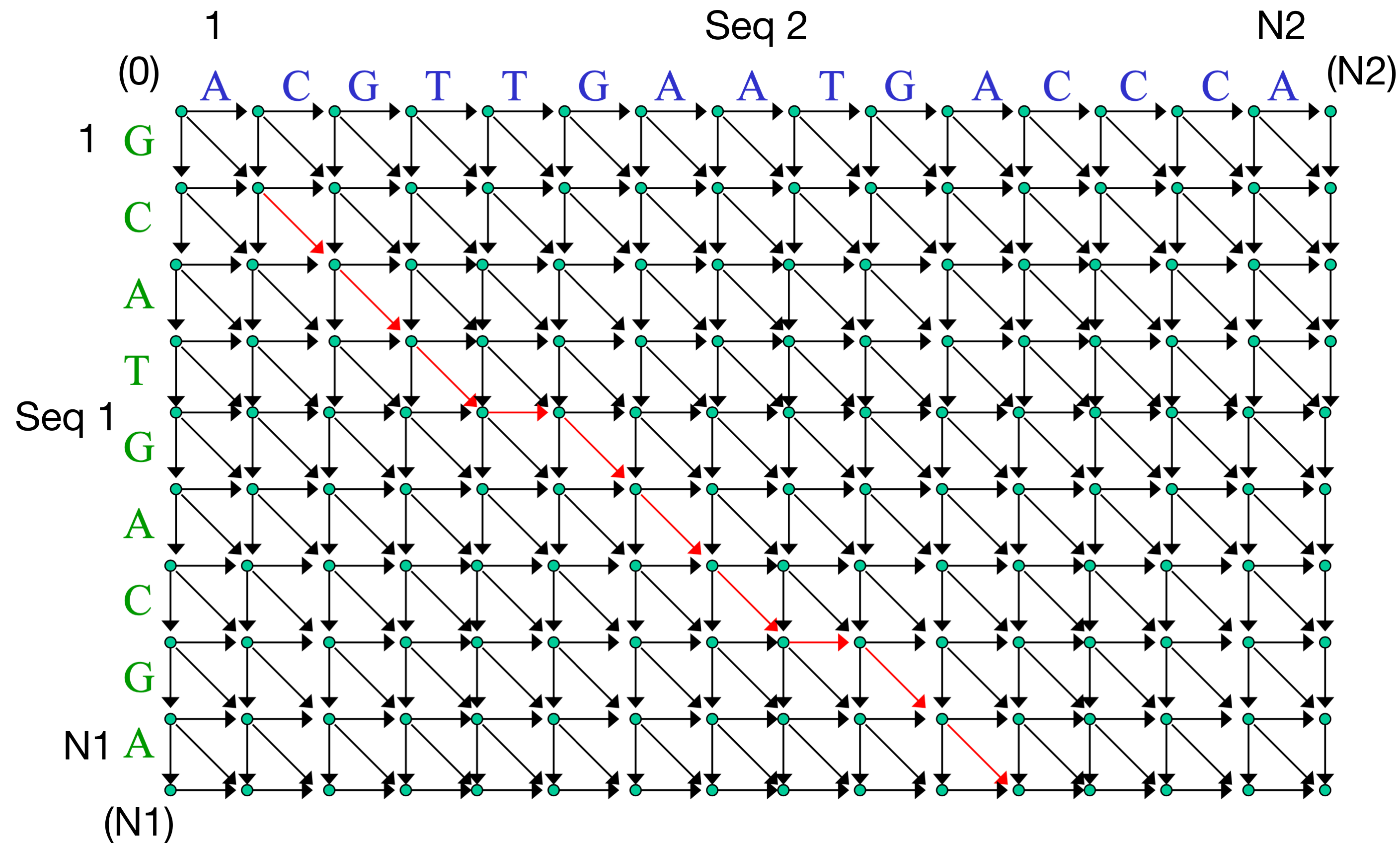
$\text{sum}((x_1, x_2), (x_1, x_3), (x_2, x_3))$

- the corresponding score matrix entry if x_i and x_j are both residues
- the gap penalty if one of x_i and x_j is a residue, and the other is a gap character
- 0 if both x_i and x_j are gap characters

Gap penalty: -6

If we only align two sequences

Sequence 1: from 1 to N1
 Sequence 2: from 1 to N2



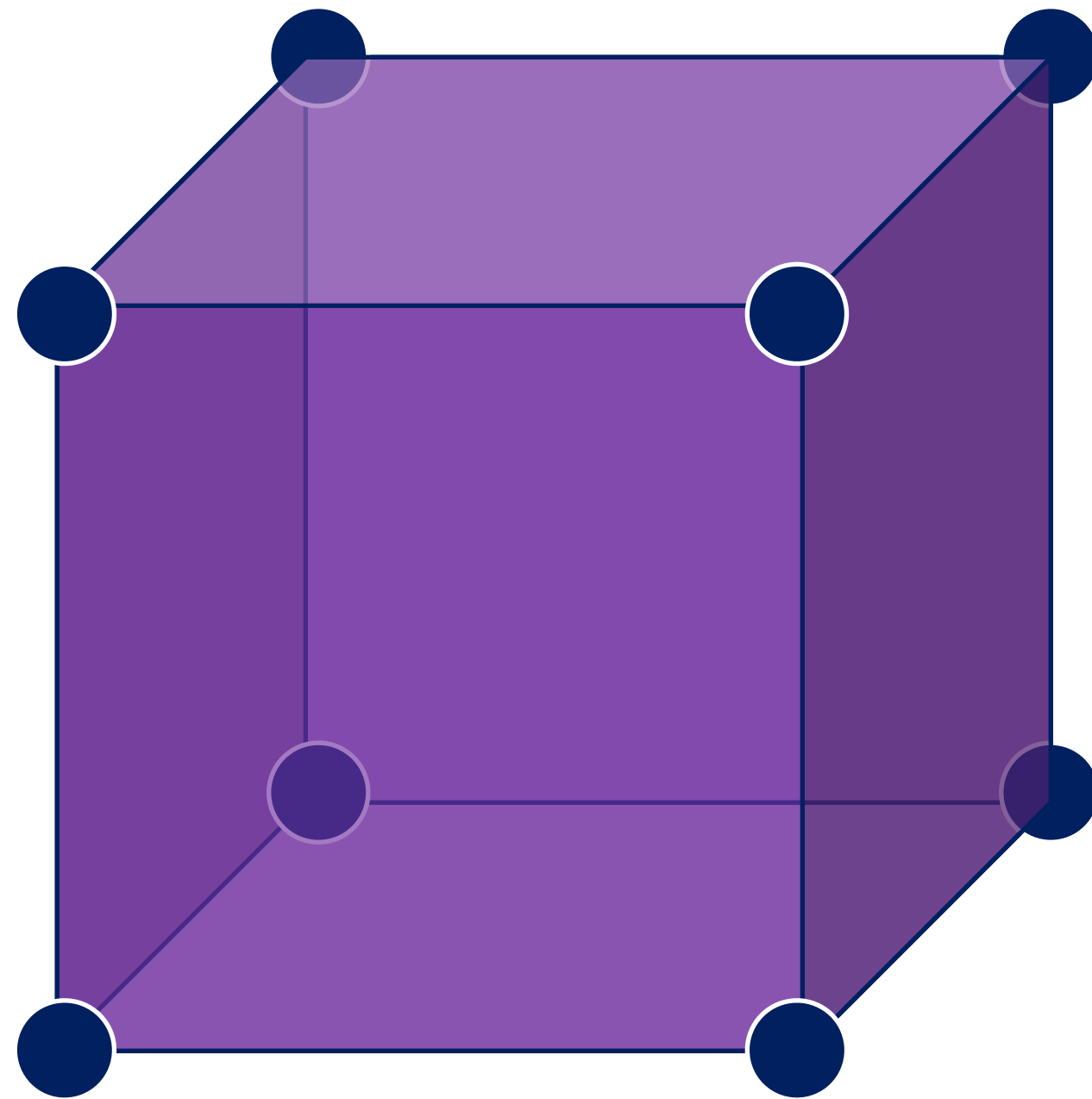
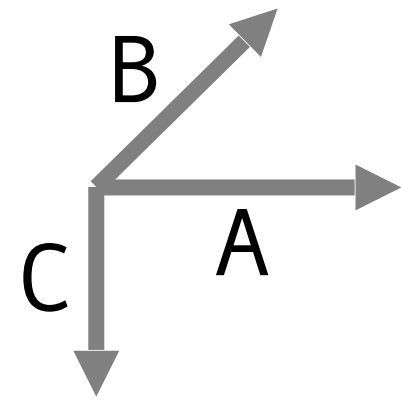
Vertice: (0,0) (0,1) (0,2) ... (0,N2)
 (1,0) (1,1)
 (2,0) ...
 ...
 (N1,0) (N1,N2)

Vertice: two *for* loops

Edges: (0,0) (0,1) weight ($_A$)
 (0,0) (1,0) weight ($G_$)
 (0,0) (1,1) weight (GA)
 ...

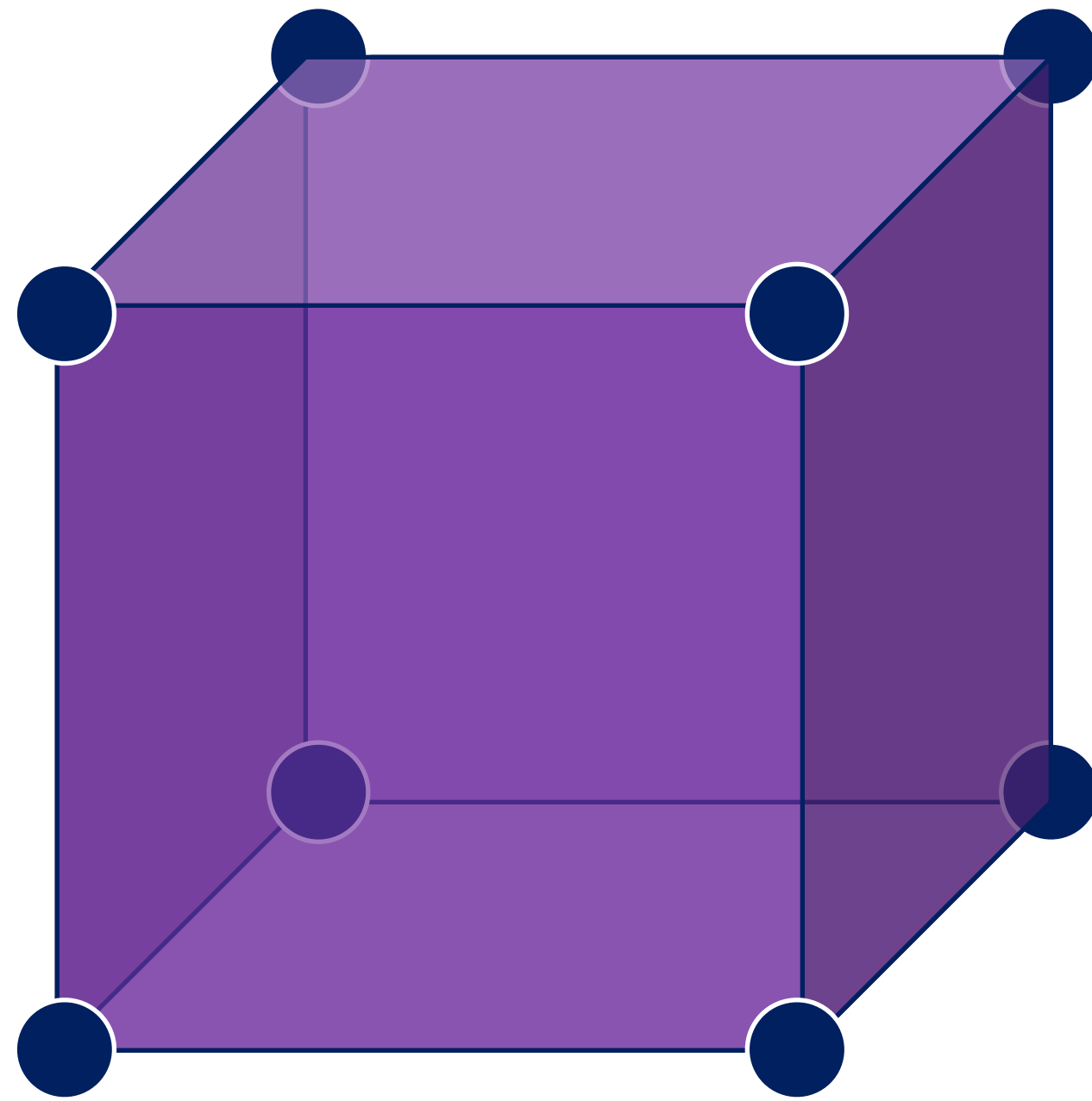
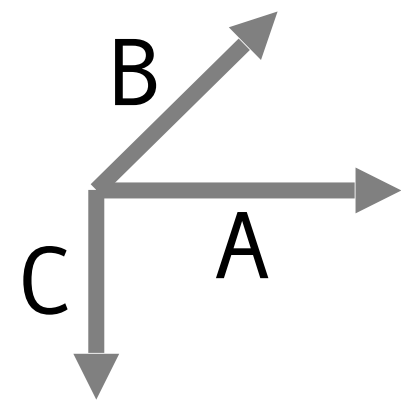
Edges: for any node (i, j)
 (i, j) -> (i+1, j)
 (i, j) -> (i, j+1)
 (i, j) -> (i+1, j+1)

Now we are aligning three sequences



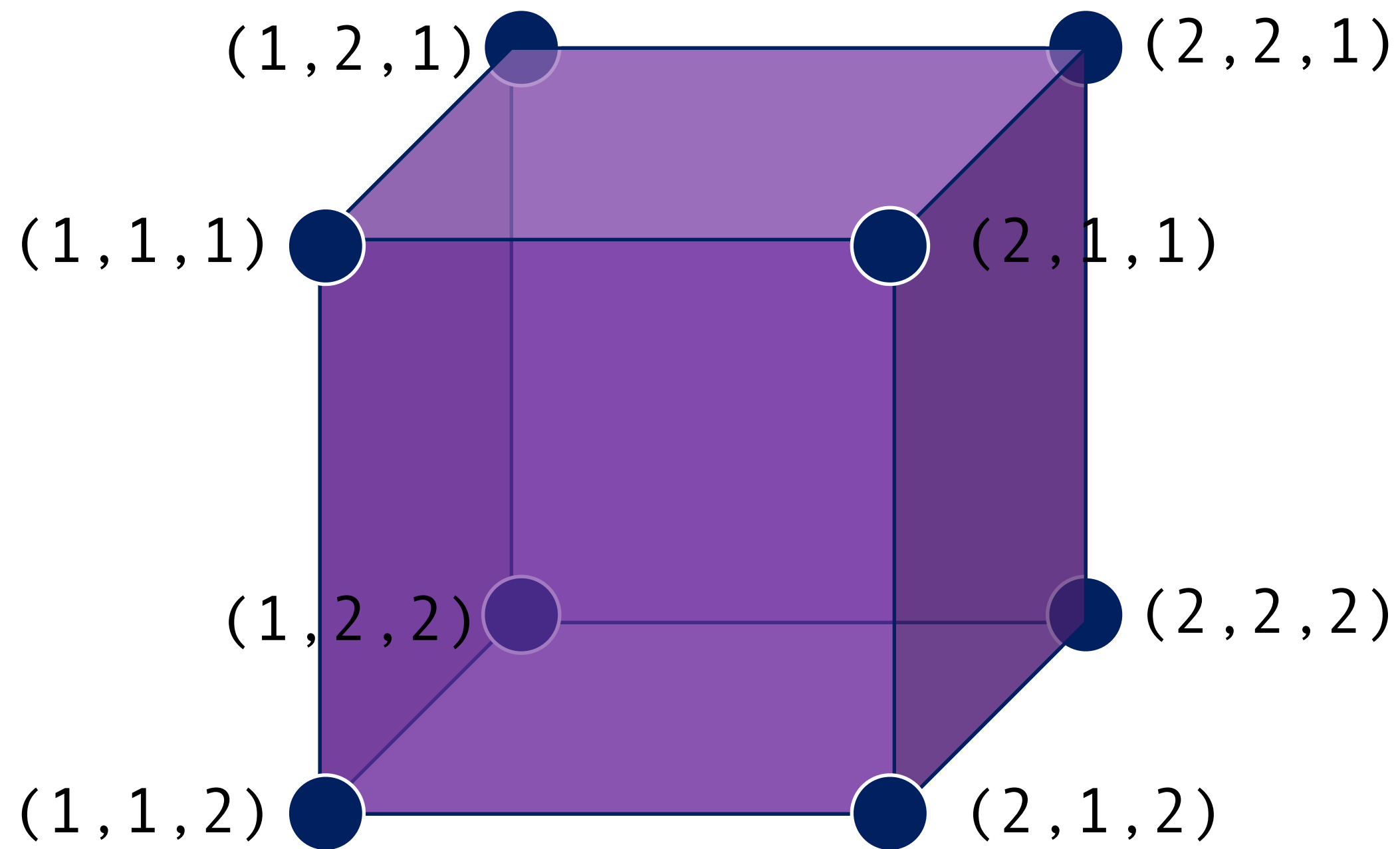
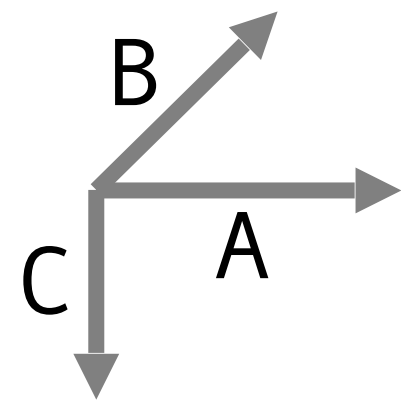
| | | | | | | |
|---|---|---|---|---|---|-----|
| A | = | M | C | D | R | ... |
| B | = | M | S | D | E | ... |
| C | = | M | V | D | R | ... |

HW5: Multiple sequence alignment



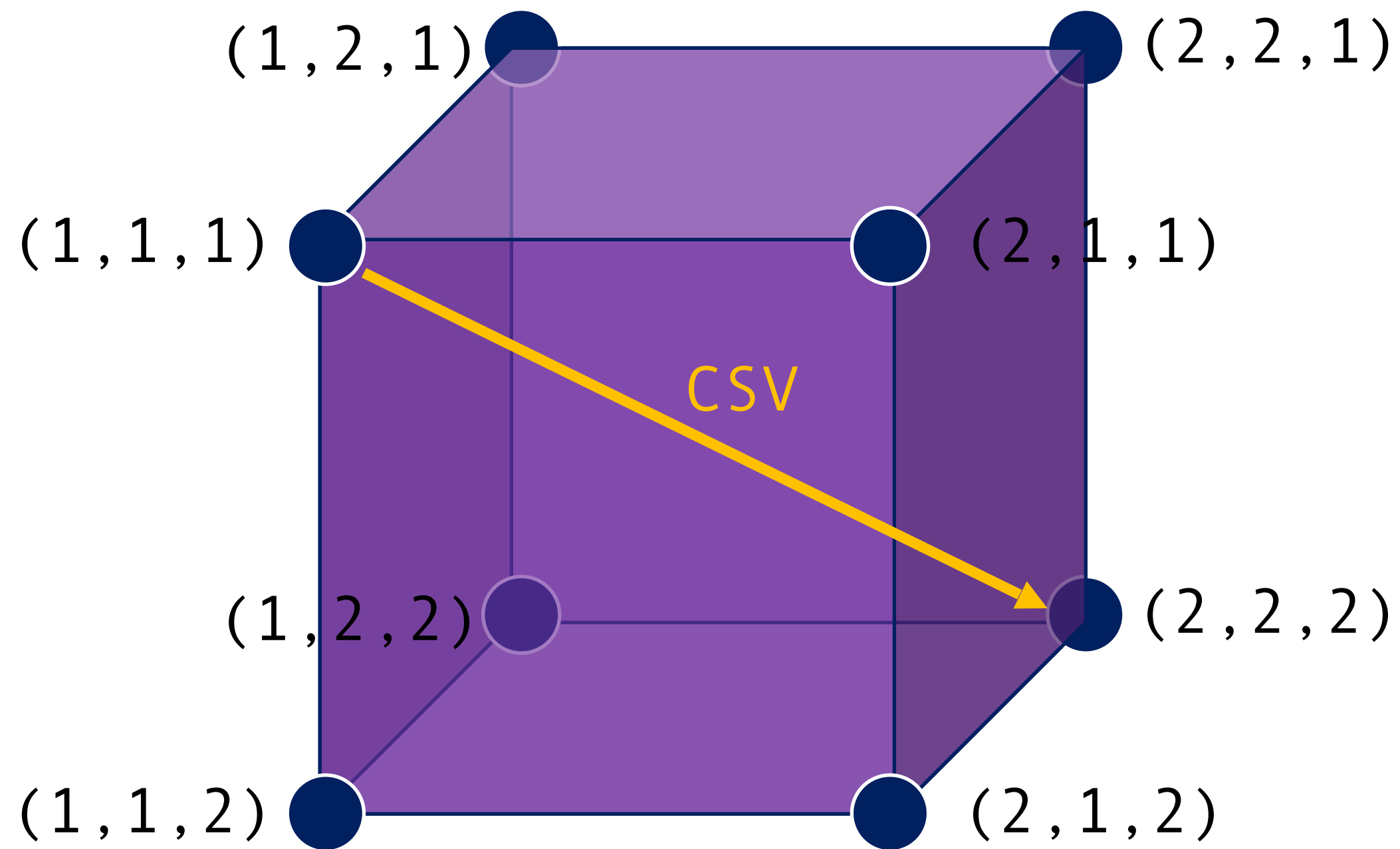
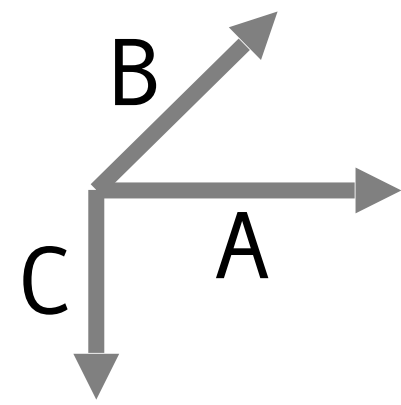
A = •M•C•D•R•...
B = •M•S•D•E•...
C = •M•V•D•R•...

HW5: Multiple sequence alignment



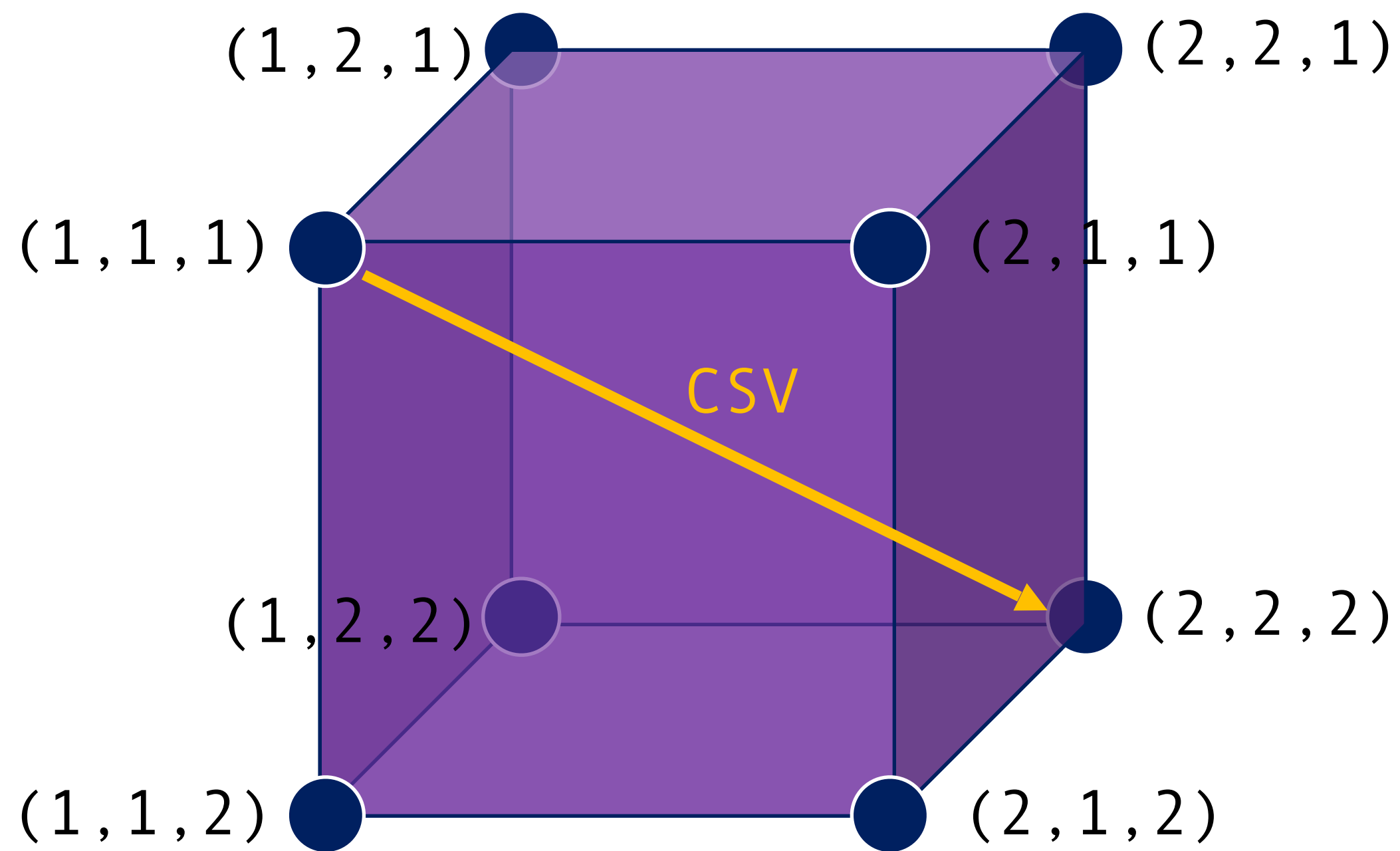
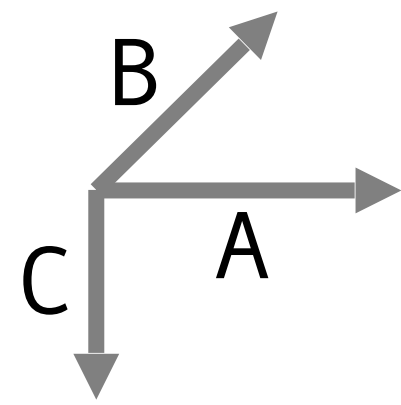
A = •M•C•D•R•...
B = •M•S•D•E•...
C = •M•V•D•R•...

HW5: Multiple sequence alignment



.
A = •M•C•D•R•...
B = •M•S•D•E•...
C = •M•V•D•R•...

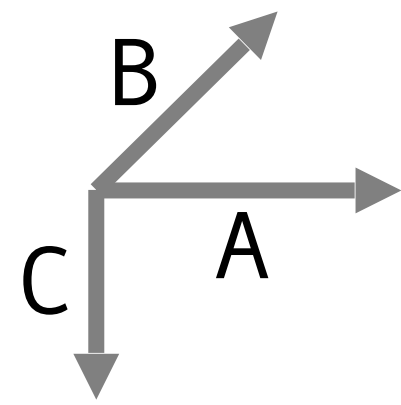
HW5: Multiple sequence alignment



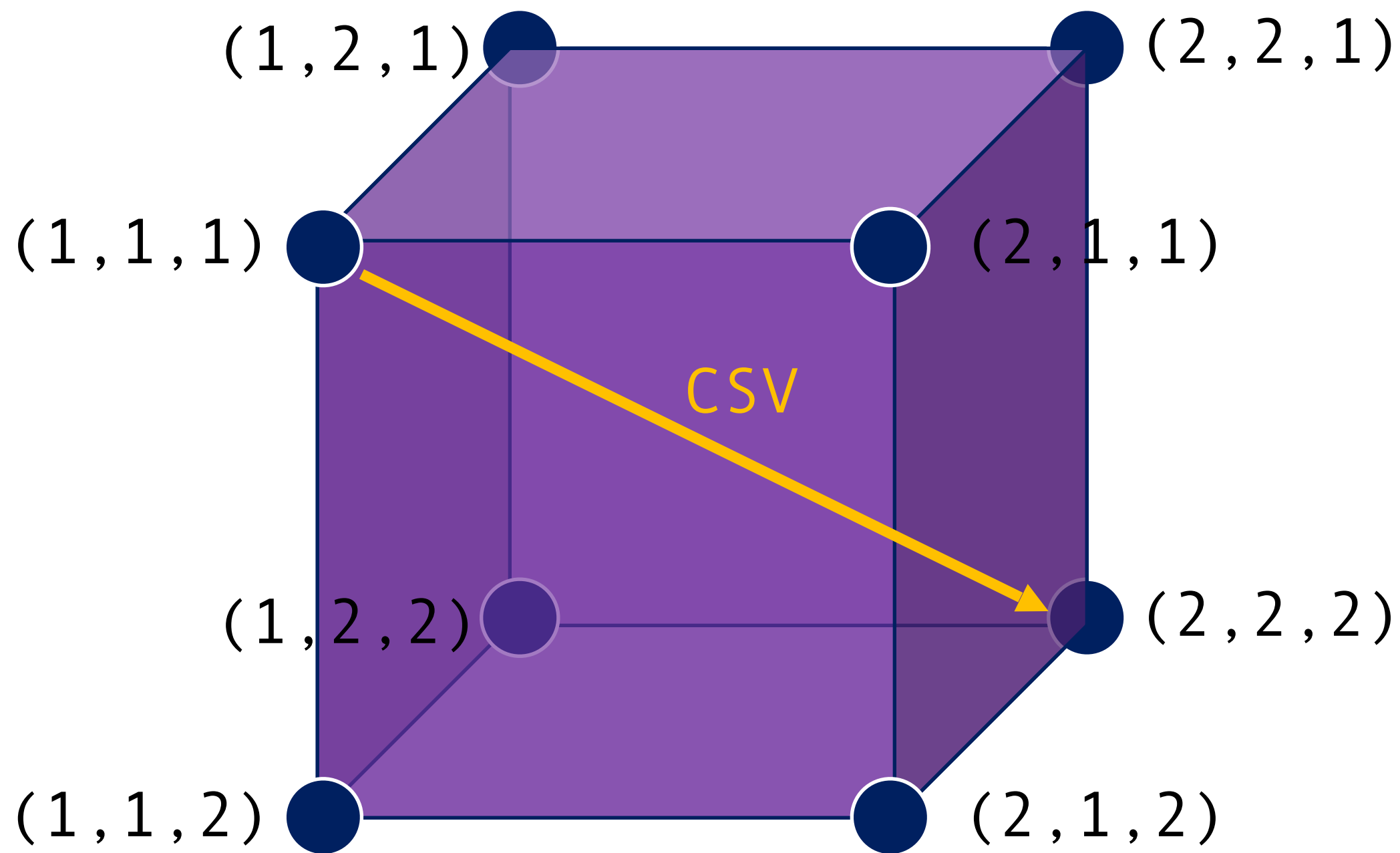
.
A = •M•C•D•R•...
B = •M•S•D•E•...
C = •M•V•D•R•...

$$\text{weight}(\text{CSV}) = \text{score}(\text{CS}) + \text{score}(\text{CV}) + \text{score}(\text{SV})$$

HW5: Multiple sequence alignment



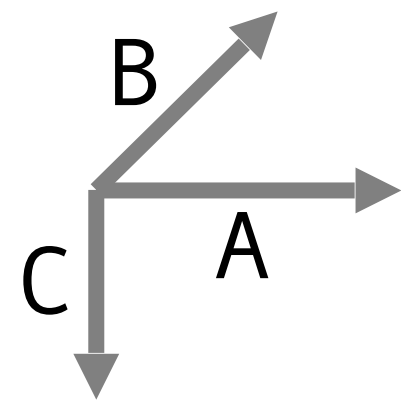
| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V | B | Z | X | * |
|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| A | 4 | -1 | -2 | -2 | 0 | -1 | -1 | 0 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 0 | -3 | -2 | 0 | -2 | -1 | 0 | -4 |
| R | -1 | 5 | 0 | -2 | -3 | 1 | 0 | -2 | 0 | -3 | -2 | 2 | -1 | -3 | -2 | -1 | -1 | -3 | -2 | -3 | -1 | 0 | -1 | -4 |
| N | -2 | 0 | 6 | 1 | -3 | 0 | 0 | 0 | 1 | -3 | -3 | 0 | -2 | -3 | -2 | 1 | 0 | -4 | -2 | -3 | 3 | 0 | -1 | -4 |
| D | -2 | -2 | 1 | 6 | -3 | 0 | 2 | -1 | -1 | -3 | -4 | -1 | -3 | -3 | -1 | 0 | -1 | -4 | -3 | -3 | 4 | 1 | -1 | -4 |
| C | 0 | -3 | -3 | -3 | 9 | -3 | -4 | -3 | -3 | -1 | -1 | -3 | -1 | -2 | -3 | -1 | -1 | -2 | -2 | -1 | -3 | -3 | -2 | -4 |
| Q | -1 | 1 | 0 | 0 | -3 | 5 | 2 | -2 | 0 | -3 | -2 | 1 | 0 | -3 | -1 | 0 | -1 | -2 | -1 | -2 | 0 | 3 | -1 | -4 |
| E | -1 | 0 | 0 | 2 | -4 | 2 | 5 | -2 | 0 | -3 | -3 | 1 | -2 | -3 | -1 | 0 | -1 | -3 | -2 | -2 | 1 | 4 | -1 | -4 |
| G | 0 | -2 | 0 | -1 | -3 | -2 | -2 | 6 | -2 | -4 | -4 | -2 | -3 | -3 | -2 | 0 | -2 | -2 | -3 | -3 | -1 | -2 | -1 | -4 |
| H | -2 | 0 | 1 | -1 | -3 | 0 | 0 | -2 | 8 | -3 | -3 | -1 | -2 | -1 | -2 | -1 | -2 | -2 | 2 | -3 | 0 | 0 | -1 | -4 |
| I | -1 | -3 | -3 | -3 | -1 | -3 | -3 | -4 | -3 | 4 | 2 | -3 | 1 | 0 | -3 | -2 | -1 | -3 | -1 | 3 | -3 | -3 | -1 | -4 |
| L | -1 | -2 | -3 | -4 | -1 | -2 | -3 | -4 | -3 | 2 | 4 | -2 | 2 | 0 | -3 | -2 | -1 | -2 | -1 | 1 | -4 | -3 | -1 | -4 |
| K | -1 | 2 | 0 | -1 | -3 | 1 | 1 | -2 | -1 | -3 | -2 | 5 | -1 | -3 | -1 | 0 | -1 | -3 | -2 | -2 | 0 | 1 | -1 | -4 |
| M | -1 | -1 | -2 | -3 | -1 | 0 | -2 | -3 | -2 | 1 | 2 | -1 | 5 | 0 | -2 | -1 | -1 | -1 | -1 | 1 | -3 | -1 | -1 | -4 |
| F | -2 | -3 | -3 | -3 | -2 | -3 | -3 | -3 | -1 | 0 | 0 | -3 | 0 | 6 | -4 | -2 | 2 | 1 | 3 | -1 | -3 | -3 | -1 | -4 |
| P | -1 | -2 | -2 | -1 | -3 | -1 | -1 | -2 | -2 | -3 | -3 | -1 | -2 | -4 | 7 | -1 | -1 | -4 | -3 | -2 | -2 | -1 | -2 | -4 |
| S | 1 | -1 | 1 | 0 | -1 | 0 | 0 | 0 | -1 | -2 | -2 | 0 | -1 | -2 | -1 | 4 | 1 | -3 | -2 | -2 | 0 | 0 | 0 | -4 |
| T | 0 | -1 | 0 | -1 | -1 | -1 | -1 | -2 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 5 | -2 | 2 | 0 | -1 | -1 | 0 | -4 |
| W | -3 | -3 | -4 | -4 | -2 | -2 | -3 | -2 | -2 | -3 | -2 | -3 | -1 | 1 | -4 | -3 | -2 | 11 | 2 | -3 | -4 | -3 | -2 | -4 |
| Y | -2 | -2 | -2 | -3 | -2 | -1 | -2 | -3 | 2 | -1 | -1 | -2 | -1 | 3 | -3 | -2 | -2 | 2 | 7 | -1 | -3 | -2 | -1 | -4 |
| V | 0 | -3 | -3 | -3 | -1 | -2 | -2 | -3 | -3 | 3 | 1 | -2 | 1 | -1 | -2 | 0 | -3 | -1 | 4 | -3 | -2 | -1 | -4 | |
| B | -2 | -1 | 3 | 4 | -3 | 0 | 1 | -1 | 0 | -3 | -4 | 0 | -3 | -3 | -2 | 0 | -1 | -4 | -3 | -3 | 4 | 1 | -1 | -4 |
| Z | -1 | 0 | 0 | 1 | -3 | 3 | 4 | -2 | 0 | -3 | -3 | 1 | -1 | -3 | -1 | 0 | -1 | -3 | -2 | -2 | 1 | 4 | -1 | -4 |
| X | 0 | -1 | -1 | -1 | -2 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -2 | 0 | 0 | -2 | -1 | -1 | -1 | -1 | -1 | -4 |
| * | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | 1 |



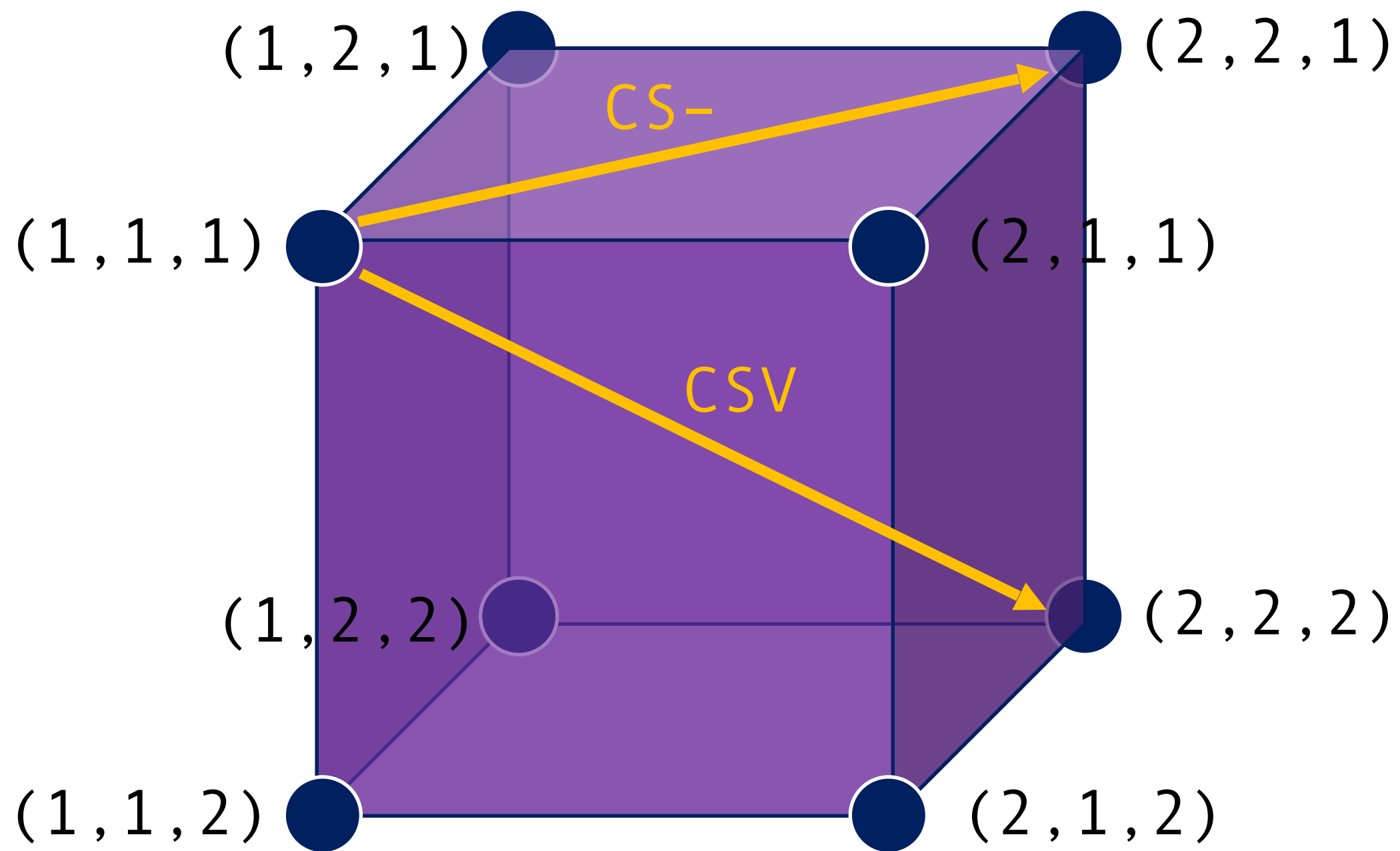
•
 A = •M•C•D•R•...
 B = •M•S•D•E•...
 C = •M•V•D•R•...

$$\text{weight}(\text{CSV}) = \text{score}(\text{CS}) + \text{score}(\text{CV}) + \text{score}(\text{SV})$$

HW5: Multiple sequence alignment



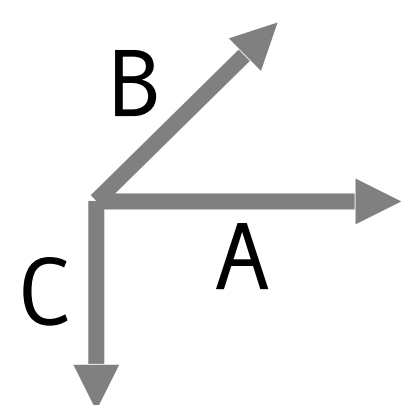
| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V | B | Z | X | * |
|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| A | 4 | -1 | -2 | -2 | 0 | -1 | -1 | 0 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 0 | -3 | -2 | 0 | -2 | -1 | 0 | -4 |
| R | -1 | 5 | 0 | -2 | -3 | 1 | 0 | -2 | 0 | -3 | -2 | 2 | -1 | -3 | -2 | -1 | -1 | -3 | -2 | -3 | -1 | 0 | -1 | -4 |
| N | -2 | 0 | 6 | 1 | -3 | 0 | 0 | 0 | 1 | -3 | -3 | 0 | -2 | -3 | -2 | 1 | 0 | -4 | -2 | -3 | 3 | 0 | -1 | -4 |
| D | -2 | -2 | 1 | 6 | -3 | 0 | 2 | -1 | -1 | -3 | -4 | -1 | -3 | -3 | -1 | 0 | -1 | -4 | -3 | -3 | 4 | 1 | -1 | -4 |
| C | 0 | -3 | -3 | -3 | 9 | -3 | -4 | -3 | -3 | -1 | -1 | -3 | -1 | -2 | -3 | -1 | -1 | -2 | -2 | -1 | -3 | -3 | -2 | -4 |
| Q | -1 | 1 | 0 | 0 | -3 | 5 | 2 | -2 | 0 | -3 | -2 | 1 | 0 | -3 | -1 | 0 | -1 | -2 | -1 | -2 | 0 | 3 | -1 | -4 |
| E | -1 | 0 | 0 | 2 | -4 | 2 | 5 | -2 | 0 | -3 | -3 | 1 | -2 | -3 | -1 | 0 | -1 | -3 | -2 | -2 | 1 | 4 | -1 | -4 |
| G | 0 | -2 | 0 | -1 | -3 | -2 | -2 | 6 | -2 | -4 | -4 | -2 | -3 | -3 | -2 | 0 | -2 | -2 | -3 | -3 | -1 | -2 | -1 | -4 |
| H | -2 | 0 | 1 | -1 | -3 | 0 | 0 | -2 | 8 | -3 | -3 | -1 | -2 | -1 | -2 | -1 | -2 | -2 | 2 | -3 | 0 | 0 | -1 | -4 |
| I | -1 | -3 | -3 | -3 | -1 | -3 | -3 | -4 | -3 | 4 | 2 | -3 | 1 | 0 | -3 | -2 | -1 | -3 | -1 | 3 | -3 | -3 | -1 | -4 |
| L | -1 | -2 | -3 | -4 | -1 | -2 | -3 | -4 | -3 | 2 | 4 | -2 | 2 | 0 | -3 | -2 | -1 | -2 | -1 | 1 | -4 | -3 | -1 | -4 |
| K | -1 | 2 | 0 | -1 | -3 | 1 | 1 | -2 | -1 | -3 | -2 | 5 | -1 | -3 | -1 | 0 | -1 | -3 | -2 | -2 | 0 | 1 | -1 | -4 |
| M | -1 | -1 | -2 | -3 | -1 | 0 | -2 | -3 | -2 | 1 | 2 | -1 | 5 | 0 | -2 | -1 | -1 | -1 | -1 | 1 | -3 | -1 | -1 | -4 |
| F | -2 | -3 | -3 | -3 | -2 | -3 | -3 | -3 | -1 | 0 | 0 | -3 | 0 | 6 | -4 | -2 | 2 | 1 | 3 | -1 | -3 | -3 | -1 | -4 |
| P | -1 | -2 | -2 | -1 | -3 | -1 | -1 | -2 | -2 | -3 | -3 | -1 | -2 | -4 | 7 | -1 | -1 | -4 | -3 | -2 | -2 | -1 | -2 | -4 |
| S | 1 | -1 | 1 | 0 | -1 | 0 | 0 | 0 | -1 | -2 | -2 | 0 | -1 | -2 | -1 | 4 | 1 | -3 | -2 | -2 | 0 | 0 | 0 | -4 |
| T | 0 | -1 | 0 | -1 | -1 | -1 | -1 | -2 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 5 | -2 | 2 | 0 | -1 | -1 | 0 | -4 |
| W | -3 | -3 | -4 | -4 | -2 | -2 | -3 | -2 | -2 | -3 | -2 | -3 | -1 | 1 | -4 | -3 | -2 | 11 | 2 | -3 | -4 | -3 | -2 | -4 |
| Y | -2 | -2 | -2 | -3 | -2 | -1 | -2 | -3 | 2 | -1 | -1 | -2 | -1 | 3 | -3 | -2 | -2 | 2 | 7 | -1 | -3 | -2 | -1 | -4 |
| V | 0 | -3 | -3 | -3 | -1 | -2 | -2 | -3 | -3 | 3 | 1 | -2 | 1 | -1 | -2 | 0 | -3 | -1 | 4 | -3 | -2 | -1 | -4 | |
| B | -2 | -1 | 3 | 4 | -3 | 0 | 1 | -1 | 0 | -3 | -4 | 0 | -3 | -3 | -2 | 0 | -1 | -4 | -3 | -3 | 4 | 1 | -1 | -4 |
| Z | -1 | 0 | 0 | 1 | -3 | 3 | 4 | -2 | 0 | -3 | -3 | 1 | -1 | -3 | -1 | 0 | -1 | -3 | -2 | -2 | 1 | 4 | -1 | -4 |
| X | 0 | -1 | -1 | -1 | -2 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -2 | 0 | 0 | -2 | -1 | -1 | -1 | -1 | -1 | -4 |
| * | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | 1 |



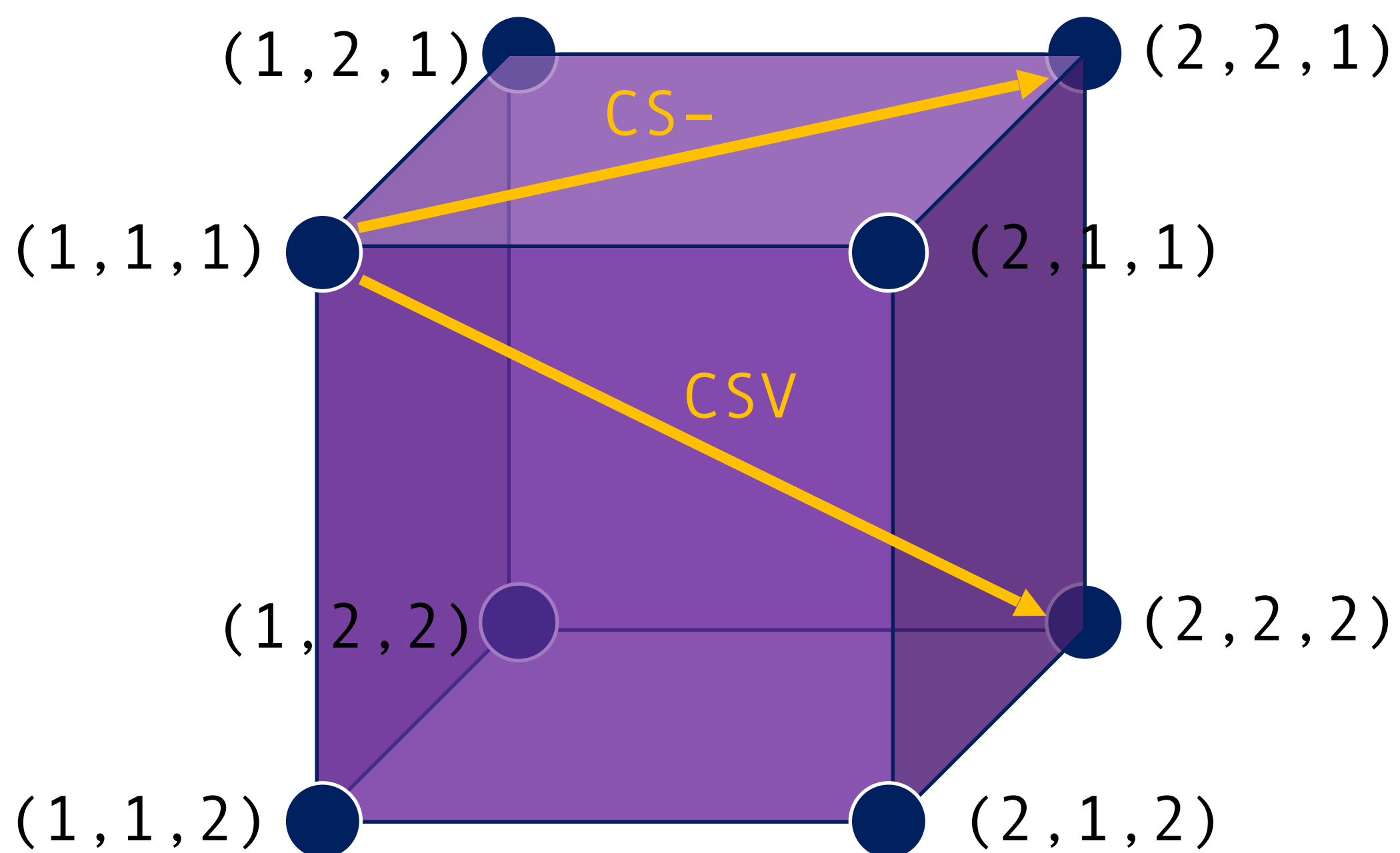
•
 A = •M•C•D•R•...
 B = •M•S•D•E•...
 C = •M•V•D•R•...

$$\text{weight}(\text{CSV}) = \text{score}(\text{CS}) + \text{score}(\text{CV}) + \text{score}(\text{SV})$$

HW5: Multiple sequence alignment



| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V | B | Z | X | * |
|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| A | 4 | -1 | -2 | -2 | 0 | -1 | -1 | 0 | -2 | -1 | -1 | -1 | -2 | -1 | 1 | 0 | -3 | -2 | 0 | -2 | -1 | 0 | -4 | |
| R | -1 | 5 | 0 | -2 | -3 | 1 | 0 | -2 | 0 | -3 | -2 | 2 | -1 | -3 | -2 | -1 | -1 | -3 | -2 | -3 | -1 | 0 | -1 | -4 |
| N | -2 | 0 | 6 | 1 | -3 | 0 | 0 | 0 | 1 | -3 | -3 | 0 | -2 | -3 | -2 | 1 | 0 | -4 | -2 | -3 | 3 | 0 | -1 | -4 |
| D | -2 | -2 | 1 | 6 | -3 | 0 | 2 | -1 | -1 | -3 | -4 | -1 | -3 | -3 | -1 | 0 | -1 | -4 | -3 | -3 | 4 | 1 | -1 | -4 |
| C | 0 | -3 | -3 | -3 | 9 | -3 | -4 | -3 | -3 | -1 | -1 | -3 | -1 | -2 | -3 | -1 | -1 | -2 | -2 | -1 | -3 | -3 | -2 | -4 |
| Q | -1 | 1 | 0 | 0 | -3 | 5 | 2 | -2 | 0 | -3 | -2 | 1 | 0 | -3 | -1 | 0 | -1 | -2 | -1 | -2 | 0 | 3 | -1 | -4 |
| E | -1 | 0 | 0 | 2 | -4 | 2 | 5 | -2 | 0 | -3 | -3 | 1 | -2 | -3 | -1 | 0 | -1 | -3 | -2 | -2 | 1 | 4 | -1 | -4 |
| G | 0 | -2 | 0 | -1 | -3 | -2 | 6 | -2 | -4 | -4 | -2 | -3 | -3 | -2 | 0 | -2 | -2 | -3 | -3 | -1 | -2 | -1 | -4 | |
| H | -2 | 0 | 1 | -1 | -3 | 0 | 0 | -2 | 8 | -3 | -3 | -1 | -2 | -1 | -2 | -1 | -2 | -2 | 2 | -3 | 0 | 0 | -1 | -4 |
| I | -1 | -3 | -3 | -3 | -1 | -3 | -3 | -4 | -3 | 4 | 2 | -3 | 1 | 0 | -3 | -2 | -1 | -3 | -1 | 3 | -3 | -3 | -1 | -4 |
| L | -1 | -2 | -3 | -4 | -1 | -2 | -3 | -4 | -3 | 2 | 4 | -2 | 2 | 0 | -3 | -2 | -1 | -2 | -1 | 1 | -4 | -3 | -1 | -4 |
| K | -1 | 2 | 0 | -1 | -3 | 1 | 1 | -2 | -1 | -3 | -2 | 5 | -1 | -3 | -1 | 0 | -1 | -3 | -2 | -2 | 0 | 1 | -1 | -4 |
| M | -1 | -1 | -2 | -3 | -1 | 0 | -2 | -3 | -2 | 1 | 2 | -1 | 5 | 0 | -2 | -1 | -1 | -1 | -1 | 1 | -3 | -1 | -1 | -4 |
| F | -2 | -3 | -3 | -3 | -2 | -3 | -3 | -1 | 0 | 0 | -3 | 0 | 6 | -4 | -2 | -2 | 1 | 3 | -1 | -3 | -3 | -1 | -4 | |
| P | -1 | -2 | -2 | -1 | -3 | -1 | -1 | -2 | -2 | -3 | -3 | -1 | -2 | -4 | 7 | -1 | -1 | -4 | -3 | -2 | -2 | -1 | -2 | -4 |
| S | 1 | -1 | 1 | 0 | -1 | 0 | 0 | 0 | -1 | -2 | -2 | 0 | -1 | -2 | -1 | 4 | 1 | -3 | -2 | -2 | 0 | 0 | 0 | -4 |
| T | 0 | -1 | 0 | -1 | -1 | -1 | -1 | -2 | -2 | -1 | -1 | -1 | -2 | -1 | 1 | 5 | -2 | -2 | 0 | -1 | -1 | 0 | -4 | |
| W | -3 | -3 | -4 | -4 | -2 | -2 | -3 | -2 | -2 | -3 | -2 | -3 | -1 | 1 | -4 | -3 | -2 | 11 | 2 | -3 | -4 | -3 | -2 | -4 |
| Y | -2 | -2 | -2 | -3 | -2 | -1 | -2 | -3 | 2 | -1 | -1 | -2 | -1 | 3 | -3 | -2 | -2 | 2 | 7 | -1 | -3 | -2 | -1 | -4 |
| V | 0 | -3 | -3 | -3 | -1 | -2 | -2 | -3 | -3 | 3 | 1 | -2 | 1 | -1 | -2 | -2 | 0 | -3 | -1 | 4 | -3 | -2 | -1 | -4 |
| B | -2 | -1 | 3 | 4 | -3 | 0 | 1 | -1 | 0 | -3 | -4 | 0 | -3 | -2 | 0 | -1 | -4 | -3 | -3 | 4 | 1 | -1 | -4 | |
| Z | -1 | 0 | 0 | 1 | -3 | 3 | 4 | -2 | 0 | -3 | -3 | 1 | -1 | -3 | -1 | 0 | -1 | -3 | -2 | -2 | 1 | 4 | -1 | -4 |
| X | 0 | -1 | -1 | -1 | -2 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -2 | 0 | 0 | -2 | -1 | -1 | -1 | -1 | -1 | -1 | -4 |
| * | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | 1 |

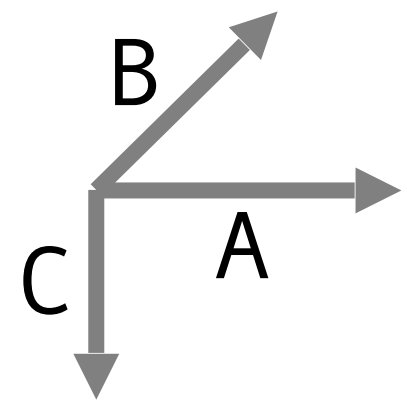


•
 A = •M•C•D•R•...
 B = •M•S•D•E•...
 C = •M•V•D•R•...

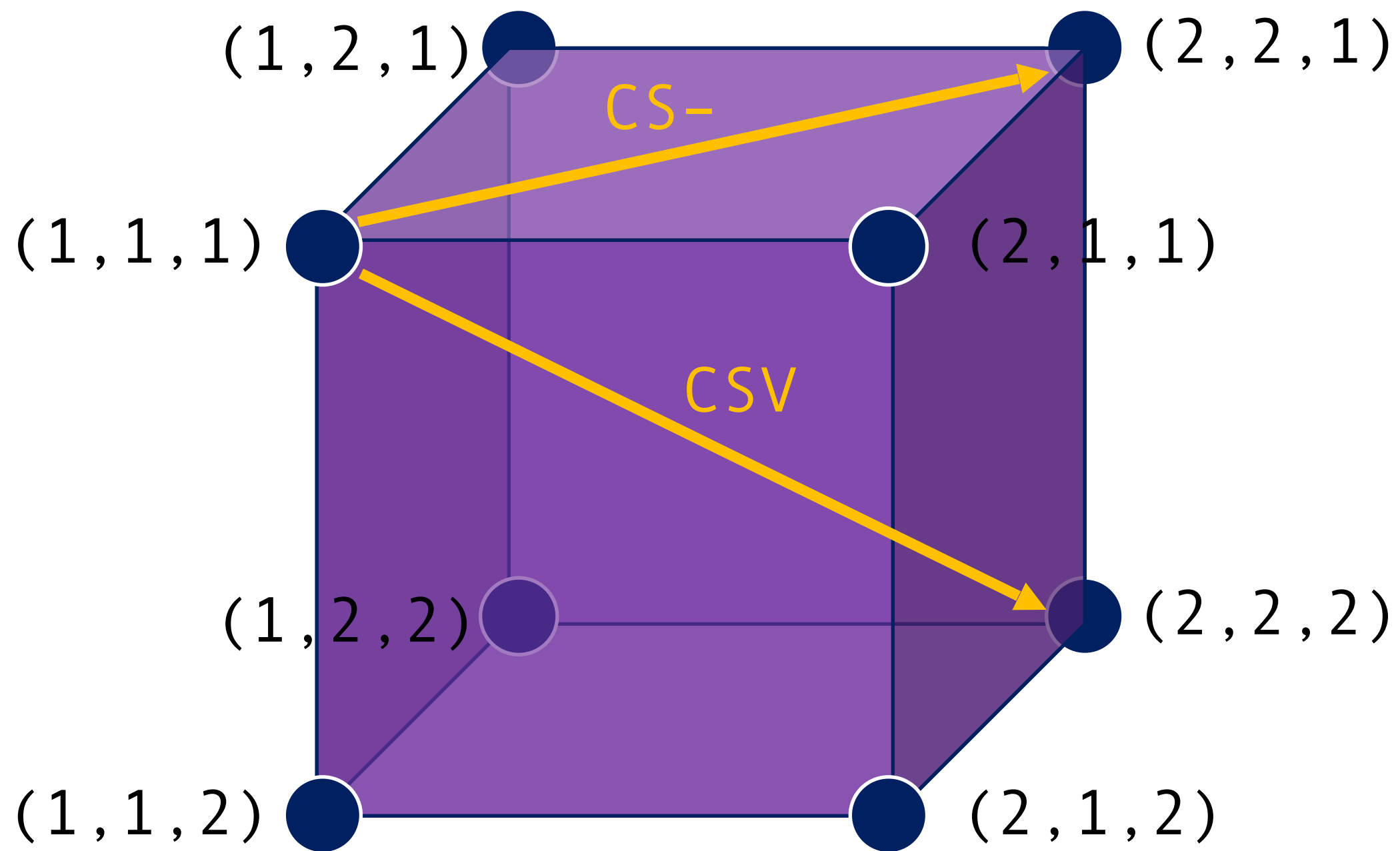
$$\text{weight}(\text{CSV}) = \text{score}(\text{CS}) + \text{score}(\text{CV}) + \text{score}(\text{SV})$$

$$\text{weight}(\text{CS-}) = \text{score}(\text{CS}) + \text{score}(\text{C-}) + \text{score}(\text{S-})$$

HW5: Multiple sequence alignment



| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V | B | Z | X | * |
|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| A | 4 | -1 | -2 | -2 | 0 | -1 | -1 | 0 | -2 | -1 | -1 | -1 | -2 | -1 | 1 | 0 | -3 | -2 | 0 | -2 | -1 | 0 | -4 | |
| R | -1 | 5 | 0 | -2 | -3 | 1 | 0 | -2 | 0 | -3 | -2 | 2 | -1 | -3 | -2 | -1 | -1 | -3 | -2 | -3 | -1 | 0 | -1 | -4 |
| N | -2 | 0 | 6 | 1 | -3 | 0 | 0 | 0 | 1 | -3 | -3 | 0 | -2 | -3 | -2 | 1 | 0 | -4 | -2 | -3 | 3 | 0 | -1 | -4 |
| D | -2 | -2 | 1 | 6 | -3 | 0 | 2 | -1 | -1 | -3 | -4 | -1 | -3 | -3 | -1 | 0 | -1 | -4 | -3 | -3 | 4 | 1 | -1 | -4 |
| C | 0 | -3 | -3 | -3 | 9 | -3 | -4 | -3 | -3 | -1 | -1 | -3 | -1 | -2 | -3 | -1 | -1 | -2 | -2 | -1 | -3 | -3 | -2 | -4 |
| Q | -1 | 1 | 0 | 0 | -3 | 5 | 2 | -2 | 0 | -3 | -2 | 1 | 0 | -3 | -1 | 0 | -1 | -2 | -1 | -2 | 0 | 3 | -1 | -4 |
| E | -1 | 0 | 0 | 2 | -4 | 2 | 5 | -2 | 0 | -3 | -3 | 1 | -2 | -3 | -1 | 0 | -1 | -3 | -2 | -2 | 1 | 4 | -1 | -4 |
| G | 0 | -2 | 0 | -1 | -3 | -2 | 6 | -2 | -4 | -4 | -2 | -3 | -3 | -2 | 0 | -2 | -2 | -3 | -3 | -1 | -2 | -1 | -4 | |
| H | -2 | 0 | 1 | -1 | -3 | 0 | 0 | -2 | 8 | -3 | -3 | -1 | -2 | -1 | -2 | -1 | -2 | -2 | 2 | -3 | 0 | 0 | -1 | -4 |
| I | -1 | -3 | -3 | -3 | -1 | -3 | -3 | -4 | -3 | 4 | 2 | -3 | 1 | 0 | -3 | -2 | -1 | -3 | -1 | 3 | -3 | -3 | -1 | -4 |
| L | -1 | -2 | -3 | -4 | -1 | -2 | -3 | -4 | -3 | 2 | 4 | -2 | 2 | 0 | -3 | -2 | -1 | -2 | -1 | 1 | -4 | -3 | -1 | -4 |
| K | -1 | 2 | 0 | -1 | -3 | 1 | 1 | -2 | -1 | -3 | -2 | 5 | -1 | -3 | -1 | 0 | -1 | -3 | -2 | -2 | 0 | 1 | -1 | -4 |
| M | -1 | -1 | -2 | -3 | -1 | 0 | -2 | -3 | -2 | 1 | 2 | -1 | 5 | 0 | -2 | -1 | -1 | -1 | -1 | 1 | -3 | -1 | -1 | -4 |
| F | -2 | -3 | -3 | -3 | -2 | -3 | -3 | -3 | -1 | 0 | 0 | -3 | 0 | 6 | -4 | -2 | 2 | 1 | 3 | -1 | -3 | -3 | -1 | -4 |
| P | -1 | -2 | -2 | -1 | -3 | -1 | -1 | -2 | -2 | -3 | -3 | -1 | -2 | -4 | 7 | -1 | -1 | -4 | -3 | -2 | -2 | -1 | -2 | -4 |
| S | 1 | -1 | 1 | 0 | -1 | 0 | 0 | 0 | -1 | -2 | -2 | 0 | -1 | -2 | -1 | 4 | 1 | -3 | -2 | -2 | 0 | 0 | 0 | -4 |
| T | 0 | -1 | 0 | -1 | -1 | -1 | -1 | -2 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 5 | -2 | 2 | 0 | -1 | -1 | 0 | -4 |
| W | -3 | -3 | -4 | -4 | -2 | -2 | -3 | -2 | -2 | -3 | -2 | -3 | -1 | 1 | -4 | -3 | -2 | 11 | 2 | -3 | -4 | -3 | -2 | -4 |
| Y | -2 | -2 | -2 | -3 | -2 | -1 | -2 | -3 | 2 | -1 | -1 | -2 | -1 | 3 | -3 | -2 | -2 | 2 | 7 | -1 | -3 | -2 | -1 | -4 |
| V | 0 | -3 | -3 | -3 | -1 | -2 | -2 | -3 | -3 | 3 | 1 | -2 | 1 | -1 | -2 | 0 | -3 | -1 | 4 | -3 | -2 | -1 | -4 | |
| B | -2 | -1 | 3 | 4 | -3 | 0 | 1 | -1 | 0 | -3 | -4 | 0 | -3 | -2 | 0 | -1 | -4 | -3 | -3 | 4 | 1 | -1 | -4 | |
| Z | -1 | 0 | 0 | 1 | -3 | 3 | 4 | -2 | 0 | -3 | -3 | 1 | -1 | -3 | -1 | 0 | -1 | -3 | -2 | -2 | 1 | 4 | -1 | -4 |
| X | 0 | -1 | -1 | -1 | -2 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -2 | 0 | 0 | -2 | -1 | -1 | -1 | -1 | -1 | -4 |
| * | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | 1 |

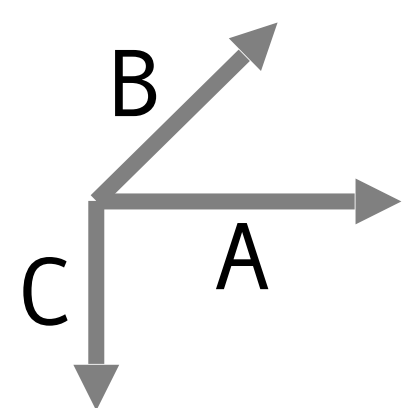


•
 A = •M•C•D•R•...
 B = •M•S•D•E•...
 C = •M•V•D•R•...

$$\text{weight(CSV)} = \text{score(CS)} + \text{score(CV)} + \text{score(SV)}$$

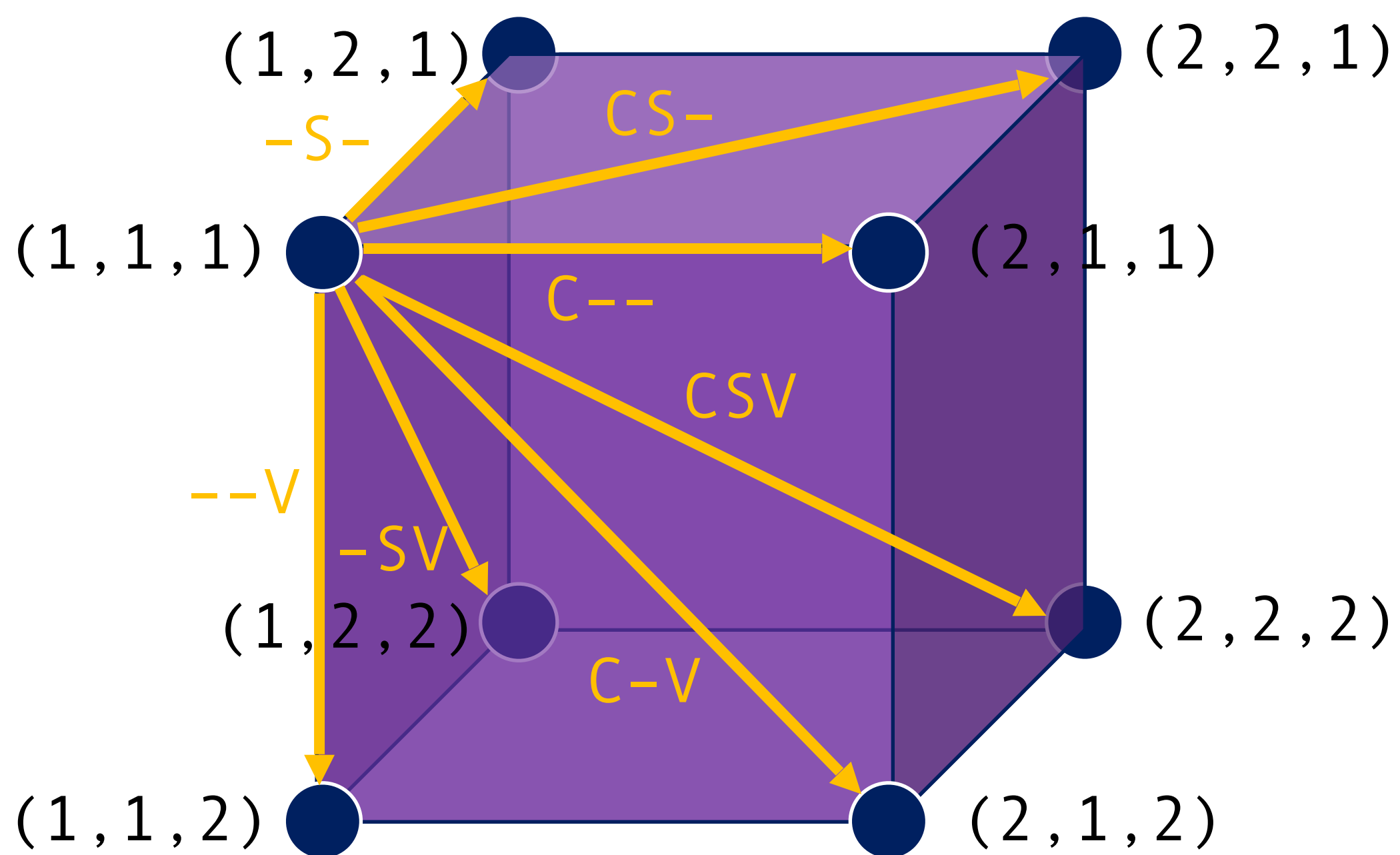
$$\text{weight(CS-)} = \text{score(CS)} + \text{gap_penalty} + \text{gap_penalty}$$

HW5: Multiple sequence alignment



$$2^3 - 1$$

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V | B | Z | X | * |
|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| A | 4 | -1 | -2 | -2 | 0 | -1 | -1 | 0 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 0 | -3 | -2 | 0 | -2 | -1 | 0 | -4 |
| R | -1 | 5 | 0 | -2 | -3 | 1 | 0 | -2 | 0 | -3 | -2 | 2 | -1 | -3 | -2 | -1 | -1 | -3 | -2 | -3 | -1 | 0 | -1 | -4 |
| N | -2 | 0 | 6 | 1 | -3 | 0 | 0 | 0 | 1 | -3 | -3 | 0 | -2 | -3 | -2 | 1 | 0 | -4 | -2 | -3 | 3 | 0 | -1 | -4 |
| D | -2 | -2 | 1 | 6 | -3 | 0 | 2 | -1 | -1 | -3 | -4 | -1 | -3 | -3 | -1 | 0 | -1 | -4 | -3 | -3 | 4 | 1 | -1 | -4 |
| C | 0 | -3 | -3 | -3 | 9 | -3 | -4 | -3 | -3 | -1 | -1 | -3 | -1 | -2 | -3 | -1 | -1 | -2 | -2 | -1 | -3 | -3 | -2 | -4 |
| Q | -1 | 1 | 0 | 0 | -3 | 5 | 2 | -2 | 0 | -3 | -2 | 1 | 0 | -3 | -1 | 0 | -1 | -2 | -1 | -2 | 0 | 3 | -1 | -4 |
| E | -1 | 0 | 0 | 2 | -4 | 2 | 5 | -2 | 0 | -3 | -3 | 1 | -2 | -3 | -1 | 0 | -1 | -3 | -2 | -2 | 1 | 4 | -1 | -4 |
| G | 0 | -2 | 0 | -1 | -3 | -2 | -2 | 6 | -2 | -4 | -4 | -2 | -3 | -3 | -2 | 0 | -2 | -2 | -3 | -3 | -1 | -2 | -1 | -4 |
| H | -2 | 0 | 1 | -1 | -3 | 0 | 0 | -2 | 8 | -3 | -3 | -1 | -2 | -1 | -2 | -1 | -2 | -2 | 2 | -3 | 0 | 0 | -1 | -4 |
| I | -1 | -3 | -3 | -3 | -1 | -3 | -3 | -4 | -3 | 4 | 2 | -3 | 1 | 0 | -3 | -2 | -1 | -3 | -1 | 3 | -3 | -3 | -1 | -4 |
| L | -1 | -2 | -3 | -4 | -1 | -2 | -3 | -4 | -3 | 2 | 4 | -2 | 2 | 0 | -3 | -2 | -1 | -2 | -1 | 1 | -4 | -3 | -1 | -4 |
| K | -1 | 2 | 0 | -1 | -3 | 1 | 1 | -2 | -1 | -3 | -2 | 5 | -1 | -3 | -1 | 0 | -1 | -3 | -2 | -2 | 0 | 1 | -1 | -4 |
| M | -1 | -1 | -2 | -3 | -1 | 0 | -2 | -3 | -2 | 1 | 2 | -1 | 5 | 0 | -2 | -1 | -1 | -1 | -1 | 1 | -3 | -1 | -1 | -4 |
| F | -2 | -3 | -3 | -3 | -2 | -3 | -3 | -3 | -1 | 0 | 0 | -3 | 0 | 6 | -4 | -2 | 2 | 1 | 3 | -1 | -3 | -3 | -1 | -4 |
| P | -1 | -2 | -2 | -1 | -3 | -1 | -1 | -2 | -2 | -3 | -3 | -1 | -2 | -4 | 7 | -1 | -1 | -4 | -3 | -2 | -2 | -1 | -2 | -4 |
| S | 1 | -1 | 1 | 0 | -1 | 0 | 0 | 0 | -1 | -2 | -2 | 0 | -1 | -2 | -1 | 4 | 1 | -3 | -2 | -2 | 0 | 0 | 0 | -4 |
| T | 0 | -1 | 0 | -1 | -1 | -1 | -1 | -2 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 5 | -2 | -2 | 0 | -1 | -1 | 0 | -4 |
| W | -3 | -3 | -4 | -4 | -2 | -2 | -3 | -2 | -2 | -3 | -2 | -3 | -1 | 1 | -4 | -3 | -2 | 11 | 2 | -3 | -4 | -3 | -2 | -4 |
| Y | -2 | -2 | -2 | -3 | -2 | -1 | -2 | -3 | 2 | -1 | -1 | -2 | -1 | 3 | -3 | -2 | -2 | 2 | 7 | -1 | -3 | -2 | -1 | -4 |
| V | 0 | -3 | -3 | -3 | -1 | -2 | -2 | -3 | -3 | 3 | 1 | -2 | 1 | -1 | -2 | -2 | 0 | -3 | -1 | 4 | -3 | -2 | -1 | -4 |
| B | -2 | -1 | 3 | 4 | -3 | 0 | 1 | -1 | 0 | -3 | -4 | 0 | -3 | -2 | 0 | -1 | -4 | -3 | -3 | 4 | 1 | -1 | -4 | |
| Z | -1 | 0 | 0 | 1 | -3 | 3 | 4 | -2 | 0 | -3 | -3 | 1 | -1 | -3 | -1 | 0 | -1 | -3 | -2 | -2 | 1 | 4 | -1 | -4 |
| X | 0 | -1 | -1 | -1 | -2 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -2 | 0 | 0 | -2 | -1 | -1 | -1 | -1 | -1 | -4 |
| * | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | 1 |

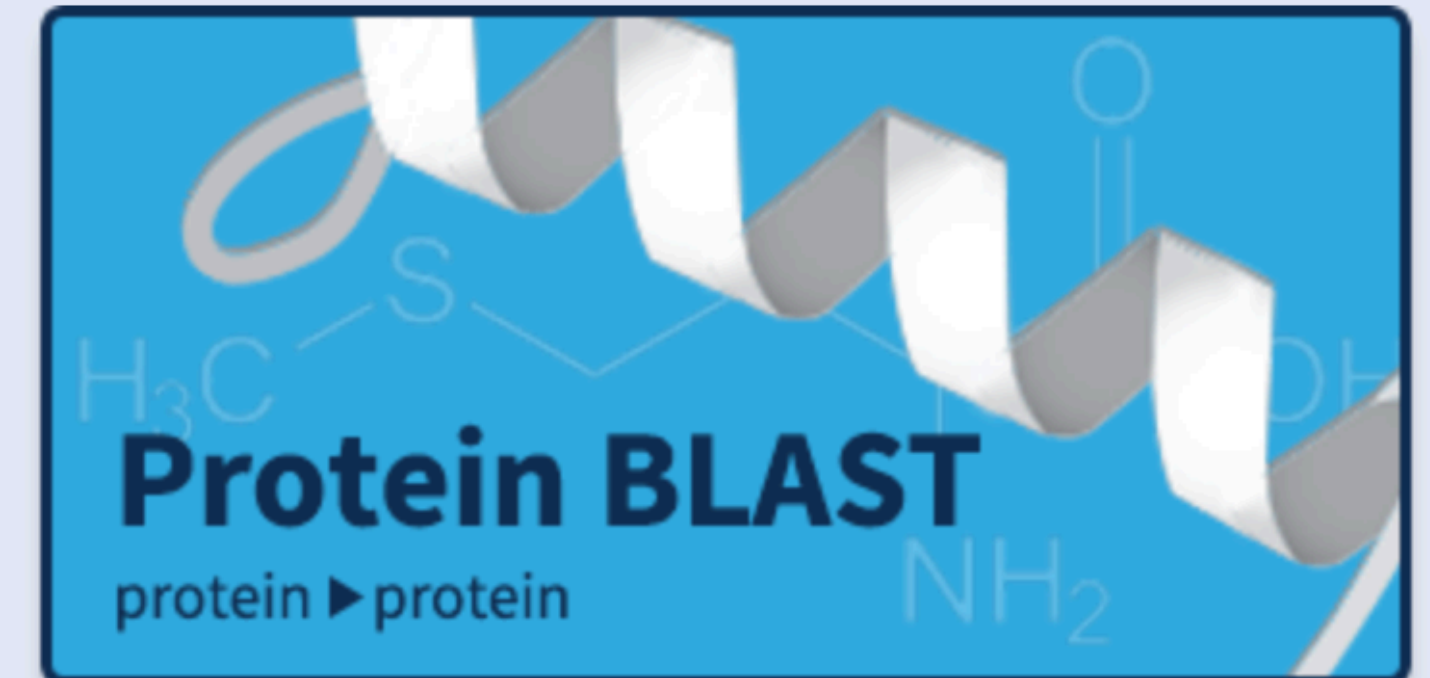
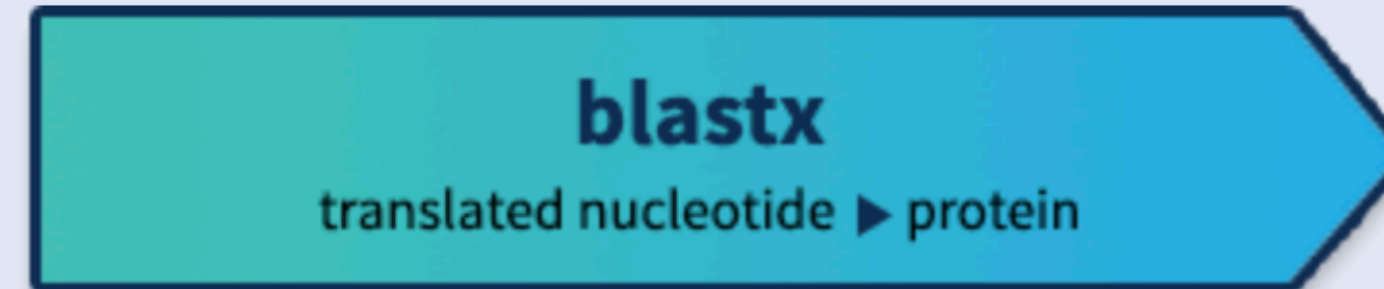


•
 A = •M•C•D•R•...
 B = •M•S•D•E•...
 C = •M•V•D•R•...

$$\text{weight}(\text{CSV}) = \text{score}(\text{CS}) + \text{score}(\text{CV}) + \text{score}(\text{SV})$$

$$\text{weight}(\text{CS-}) = \text{score}(\text{CS}) + \text{gap_penalty} + \text{gap_penalty}$$

Web BLAST



Web BLAST

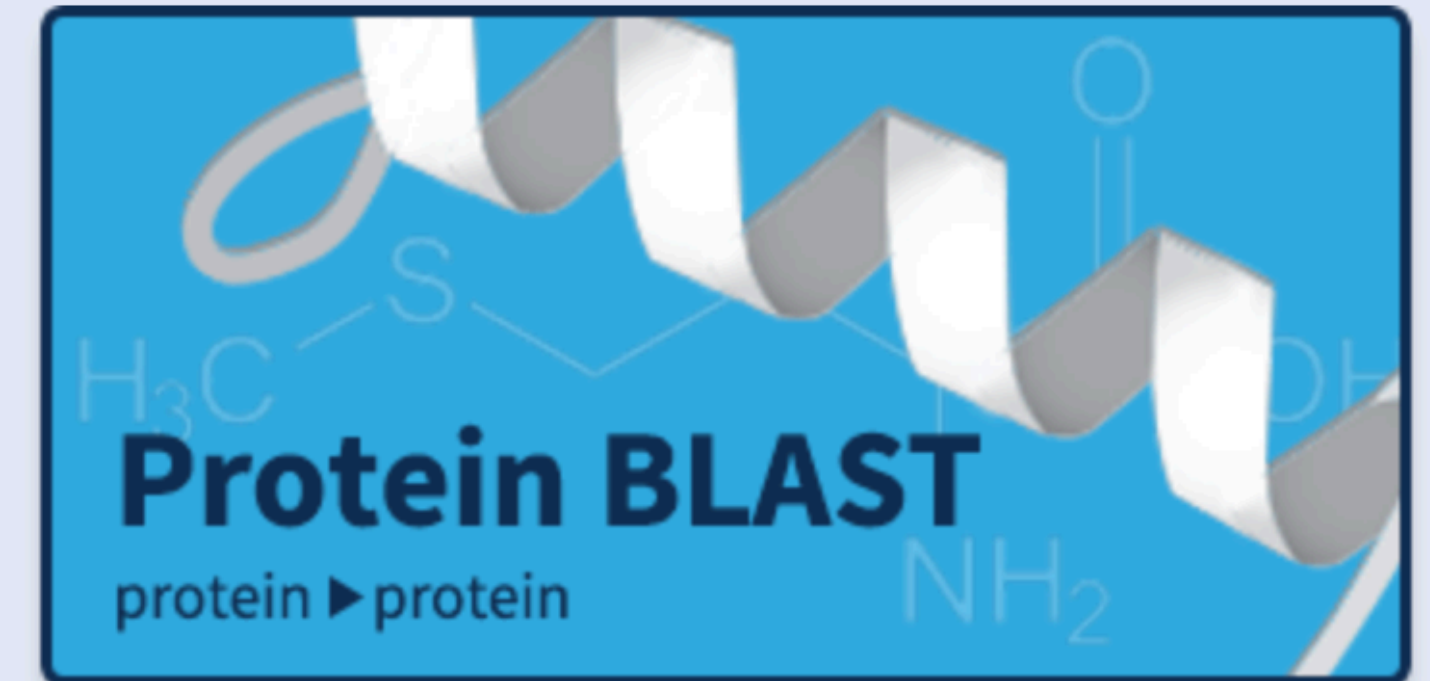
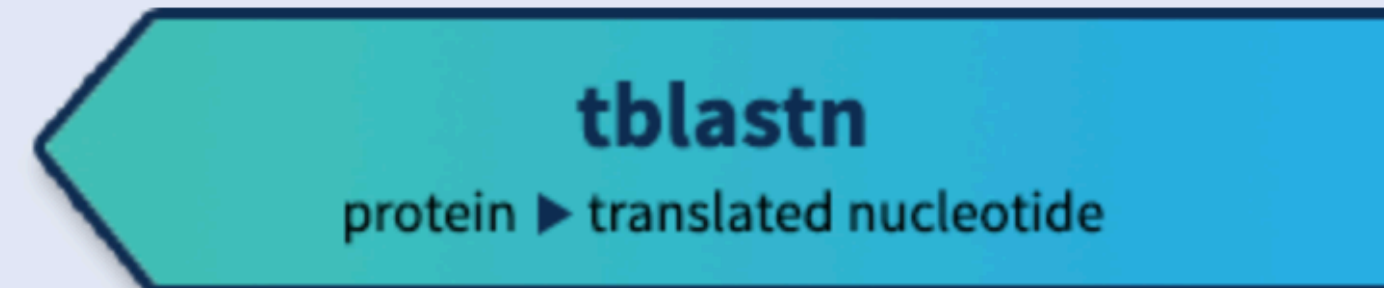
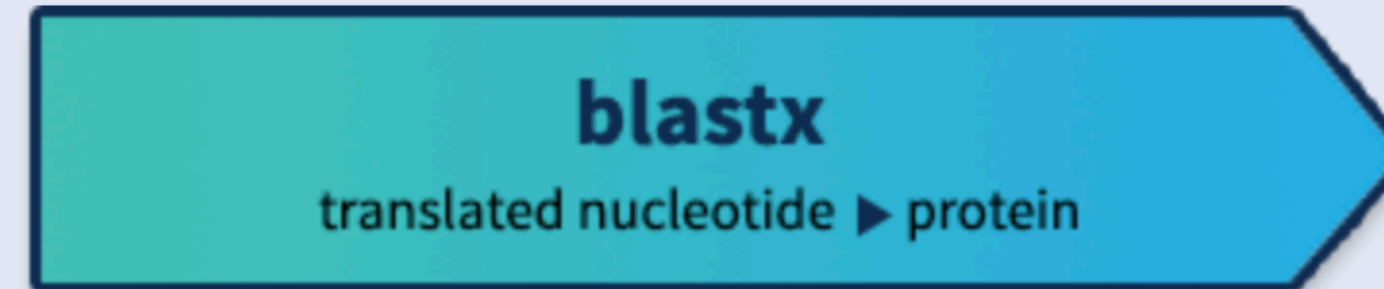


Table 1. Key features of the BLAST search pages in the “Basic BLAST” category

| Search page | Query & database combination | Alignment type | Programs & functions (default program in bold) |
|------------------|--|--------------------------|---|
| nucleotide blast | nucleotide vs nucleotide | nucleotide vs nucleotide | megablast : for sequence identification, intra-species comparison <u>discontiguous megablast</u> : for cross-species comparison, searching with coding sequences <u>blastn</u> : for searching with shorter queries, cross-species comparison |
| protein blast | Protein vs protein | protein vs protein | blastp : general sequence identification and similarity searches <u>DELTA-BLAST</u> [2] : protein similarity search with higher sensitivity than blastp <u>PSI-BLAST</u> : iterative search for position-specific score matrix (PSSM) construction or identification of distant relatives for a protein family <u>PHI-BLAST</u> : protein alignment with input pattern as anchor/constraint |
| blastx | nucleotide (translated) vs protein | protein vs protein | <u>blastx</u> : for identifying potential protein products encoded by a nucleotide query |
| tblastn | protein vs nucleotide (translated) | protein vs protein | <u>tblastn</u> : for identifying database sequences encoding proteins similar to the query |
| tblastx | nucleotide (translated) vs nucleotide (translated) | protein vs protein | <u>tblastx</u> : for identifying nucleotide sequences similar to the query based on their coding potential |

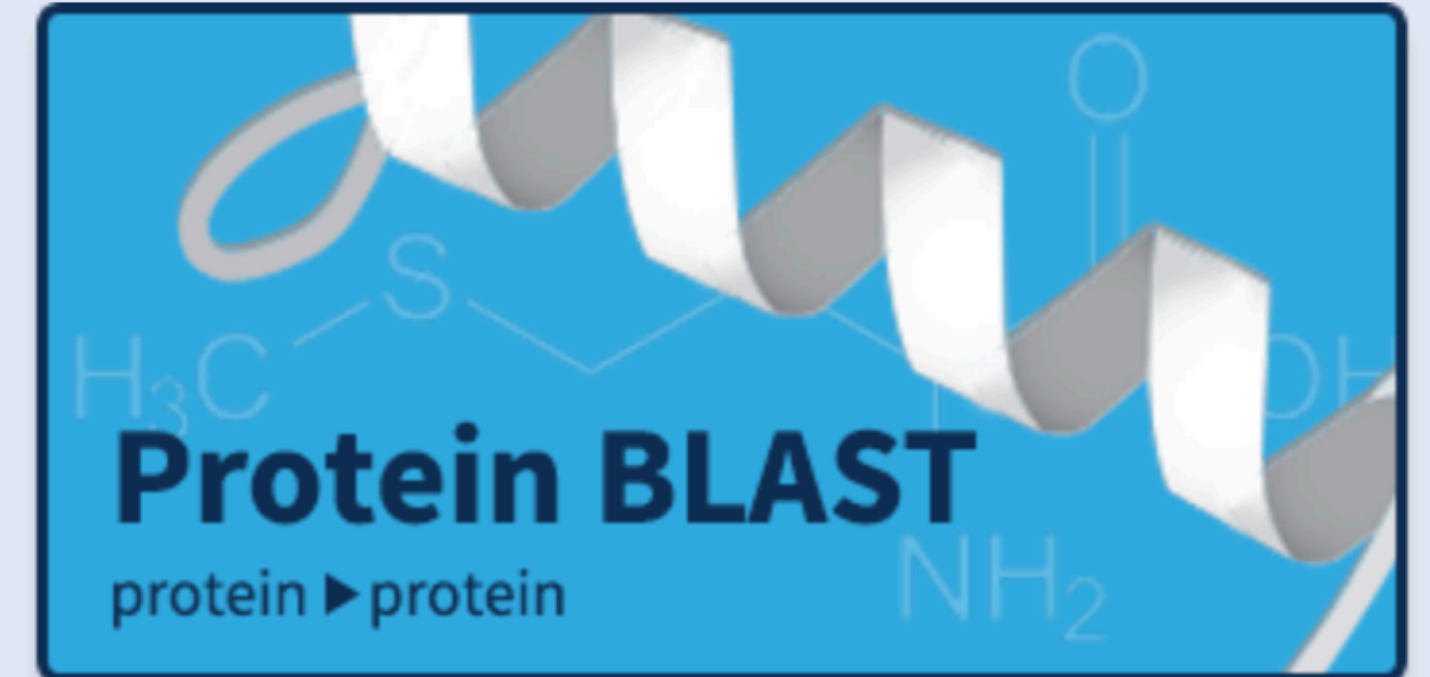
Web BLAST



Nucleotide BLAST
nucleotide ► nucleotide

blastx
translated nucleotide ► protein

tblastn
protein ► translated nucleotide



Protein BLAST
protein ► protein

Scoring Parameters

Match/Mismatch Scores
Gap Costs

1,-2

Linear

Scoring Parameters

Matrix

BLOSUM62

Gap Costs

Existence: 11 Extension: 1

✓ 1,-2

1,-3

1,-4

2,-3

4,-5

1,-1

✓ Linear

Existence: 5 Extension: 2

Existence: 2 Extension: 2

Existence: 1 Extension: 2

Existence: 0 Extension: 2

Existence: 3 Extension: 1

Existence: 2 Extension: 1

Existence: 1 Extension: 1

PAM30

PAM70

PAM250

BLOSUM80

✓ BLOSUM62

BLOSUM45

BLOSUM50

BLOSUM90

Existence: 11 Extension: 2

Existence: 10 Extension: 2

Existence: 9 Extension: 2

Existence: 8 Extension: 2

Existence: 7 Extension: 2

Existence: 6 Extension: 2

Existence: 13 Extension: 1

Existence: 12 Extension: 1

✓ Existence: 11 Extension: 1

Existence: 10 Extension: 1

Existence: 9 Extension: 1

What can I search against?

- Nucleotide databases

Genomic plus Transcript

Human genomic plus transcript (Human G+T)

Mouse genomic plus transcript (Mouse G+T)

Other Databases

- ✓ Nucleotide collection (nr/nt)
- 16S ribosomal RNA sequences (Bacteria and Archaea)
- Reference RNA sequences (refseq_rna)
- RefSeq Representative genomes (refseq_representative_genomes)
- RefSeq Genome Database (refseq_genomes)
- Whole-genome shotgun contigs (wgs)
- Expressed sequence tags (est)
- Sequence Read Archive (SRA)
- Transcriptome Shotgun Assembly (TSA)
- High throughput genomic sequences (HTGS)
- Patent sequences(pat)
- Protein Data Bank (pdb)
- Reference genomic sequences (refseq_genomic)
- Human RefSeqGene sequences(RefSeq_Gene)
- Genomic survey sequences (gss)
- Sequence tagged sites (dbsts)

What can I search against?

- Nucleotide databases

Genomic plus Transcript

Human genomic plus transcript (Human G+T)

Mouse genomic plus transcript (Mouse G+T)

Other Databases

- ✓ Nucleotide collection (nr/nt)
- 16S ribosomal RNA sequences (Bacteria and Archaea)
- Reference RNA sequences (refseq_rna)
- RefSeq Representative genomes (refseq_representative_genomes)
- RefSeq Genome Database (refseq_genomes)
- Whole-genome shotgun contigs (wgs)
- Expressed sequence tags (est)
- Sequence Read Archive (SRA)
- Transcriptome Shotgun Assembly (TSA)
- High throughput genomic sequences (HTGS)
- Patent sequences(pat)
- Protein Data Bank (pdb)
- Reference genomic sequences (refseq_genomic)
- Human RefSeqGene sequences(RefSeq_Gene)
- Genomic survey sequences (gss)
- Sequence tagged sites (dbsts)

Title: Nucleotide collection (nt)

Description: The nucleotide collection consists of GenBank+EMBL+DDBJ+PDB+RefSeq sequences, but excludes EST, STS, GSS, WGS, TSA, patent sequences as well as phase 0, 1, and 2 HTGS sequences and sequences longer than 100Mb. The database is non-redundant. Identical sequences have been merged into one entry, while preserving the accession, GI, title and taxonomy information for each entry.

Molecule Type: mixed DNA

Update date: 2019/01/30

Number of sequences: 50392116

What can I search against?

- Nucleotide databases

◆ Patent sequences(pat) ⓘ

Title: Nucleotide sequences derived from the Patent division of GenBank
Molecule Type: mixed DNA
Update date: 2019/01/30
Number of sequences: 35937288

◆ 16S ribosomal RNA sequences (Bacteria and Archaea) ⓘ

Title: 16S ribosomal RNA (Bacteria and Archaea)
Description: 16S ribosomal RNA sequences from bacteria and archaea
Molecule Type: Ribosomal RNA
Update date: 2019/01/30
Number of sequences: 20797

Genomic plus Transcript
Human genomic plus transcript (Human G+T)
Mouse genomic plus transcript (Mouse G+T)

Other Databases

- ✓ Nucleotide collection (nr/nt)
16S ribosomal RNA sequences (Bacteria and Archaea)
Reference RNA sequences (refseq_rna)
RefSeq Representative genomes (refseq_representative_genomes)
RefSeq Genome Database (refseq_genomes)
Whole-genome shotgun contigs (wgs)
Expressed sequence tags (est)
Sequence Read Archive (SRA)
Transcriptome Shotgun Assembly (TSA)
High throughput genomic sequences (HTGS)
Patent sequences(pat)
Protein Data Bank (pdb)
Reference genomic sequences (refseq_genomic)
Human RefSeqGene sequences(RefSeq_Gene)
Genomic survey sequences (gss)
Sequence tagged sites (dbsts)

◆ RefSeq Representative genomes (refseq_representative_genomes) ⓘ

Title: RefSeq Representative Genome Database
Description: This database contains the Reference and Representative genomes selected from the NCBI Refseq Genomes database. As a result, the genomes in this database are among the best quality genomes available at NCBI. It is also constructed with minimum redundancy in genome representation. For the eukaryotes, only one genome is included per organism. For other organisms, however, multiple genomes from diverse isolates of the same organism (such as E. coli) may be included.
Molecule Type: Genomic
Update date: 2015/09/30
Number of sequences: 14788425

What can I search against?

- Protein databases

- ✓ Non-redundant protein sequences (nr)
 - Reference proteins (refseq_protein)
 - Model Organisms (landmark)
 - UniProtKB/Swiss-Prot (swissprot)
 - Patented protein sequences (pat)
 - Protein Data Bank proteins (pdb)
 - Metagenomic proteins (env_nr)
 - Transcriptome Shotgun Assembly proteins (tsa_nr)

Non-redundant protein sequences (nr)



Title: All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects

Molecule Type: Protein

Update date: 2019/01/30

Number of sequences: 187087715

What can I search against?

- Protein databases

- ✓ Non-redundant protein sequences (nr)
 - Reference proteins (refseq_protein)
 - Model Organisms (landmark)
 - UniProtKB/Swiss-Prot (swissprot)
 - Patented protein sequences (pat)
 - Protein Data Bank proteins (pdb)
 - Metagenomic proteins (env_nr)
 - Transcriptome Shotgun Assembly proteins (tsa_nr)

◆ Metagenomic proteins(env_nr) ⓘ

Title: Proteins from WGS metagenomic projects (env_nr).
Description: Proteins from WGS metagenomic projects (env_nr).
Molecule Type: Protein
Update date: 2018/11/20
Number of sequences: 7023997

◆ UniProtKB/Swiss-Prot (swissprot) ⓘ

Title: Non-redundant UniProtKB/SwissProt sequences.
Molecule Type: Protein
Update date: 2019/01/30
Number of sequences: 471372

◆ Patented protein sequences(pat) ⓘ

Title: Protein sequences derived from the Patent division of GenBank
Molecule Type: Protein
Update date: 2019/01/30
Number of sequences: 2320648

◆ Protein Data Bank proteins(pdb) ⓘ

Title: PDB protein database
Description: This database consists of sequences from the Protein Data Bank (PDB), which contains information about experimentally-determined structures of proteins, nucleic acids, and complex assemblies.
Molecule Type: Protein
Update date: 2019/01/30
Number of sequences: 104504

What can I search against?

Specialized searches

SmartBLAST

Find proteins highly similar to your query

Primer-BLAST

Design primers specific to your PCR template

Global Align

Compare two sequences across their entire span (Needleman-Wunsch)

CD-search

Find conserved domains in your sequence

GEO

Find matches to gene expression profiles

IgBLAST

Search immunoglobulins and T cell receptor sequences

VecScreen

Search sequences for vector contamination

CDART

Find sequences with similar conserved domain architecture

Targeted Loci

Search markers for phylogenetic analysis

Multiple Alignment

Align sequences using domain and protein constraints

BioAssay

Search protein or nucleotide targets in PubChem BioAssay

MOLE-BLAST

Establish taxonomy for uncultured or environmental sequences

Using BLAST: Input

BLASTN programs search nucleotide databases using a nucleotide query. [more...](#)

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#)

```
AUGGCGGCGCUUCAAGAUGUUGCCUGACCACGCCUGCGCUGCUGCGGGCCUGGCCCA
GGCUGCACGUGCAGGACCUCCUUGGUGGCCGGAGCCUCCACAGCAGUGCGGUGGCAGC
CUUCCAAGUAUCGUGAACAUGCAGGAUCCCGAGAUGGACCGACAUGAAGUCAGUGACUG
ACCGGGCAGCCCGCACCCUGCUGUGGACUGAAUACCGAGGCCUGGGCAUGACCCUGAG
CUACCUGUUCGCGGAACCGGCCACCAUCAACUCCCGUUUCGAGAAGGCCCGCUGAGCC
```

Or, upload file No file chosen [?](#)

Job Title

Enter a descriptive title for your BLAST search [?](#)

Align two or more sequences [?](#)

Query subrange [?](#)

From

To

Seq Length: 755

Using BLAST: Output

Sequences producing significant alignments:

Select: [All](#) [None](#) Selected:0

[Alignments](#) [Download](#) [GenBank](#) [Graphics](#) [Distance tree of results](#)

| | Description | Max score | Total score | Query cover | E value | Ident | Accession |
|--------------------------|---|-----------|-------------|-------------|---------|-------|--------------------------------|
| <input type="checkbox"/> | Homo sapiens NADH:ubiquinone oxidoreductase core subunit S8 (NDUFS8), mRNA | 830 | 830 | 90% | 0.0 | 89% | NM_002496.4 |
| <input type="checkbox"/> | Human mitochondrial NADH dehydrogenase-ubiquinone Fe-S protein 8, 23 kDa subunit precursor (NDUFS8) nuclear mRNA encoding mitochondrial protein, complete cds | 830 | 830 | 90% | 0.0 | 89% | U65579.1 |
| <input type="checkbox"/> | Homo sapiens mRNA for NADH dehydrogenase (ubiquinone) Fe-S protein 8, 23kDa (NADH-coenzyme Q reductase) variant, clone: KAT09377 | 828 | 828 | 90% | 0.0 | 89% | AK223114.1 |
| <input type="checkbox"/> | Homo sapiens NADH dehydrogenase (ubiquinone) Fe-S protein 8, 23kDa (NADH-coenzyme Q reductase), mRNA (cDNA clone MGC:149876 IMAGE:40119290), complete cds | 824 | 824 | 90% | 0.0 | 89% | BC119754.2 |
| <input type="checkbox"/> | PREDICTED: Gorilla gorilla gorilla NADH:ubiquinone oxidoreductase core subunit S8 (NDUFS8), transcript variant X1, mRNA | 819 | 819 | 90% | 0.0 | 89% | XM_019035670.1 |
| <input type="checkbox"/> | PREDICTED: Pan troglodytes NADH:ubiquinone oxidoreductase core subunit S8 (NDUFS8), transcript variant X1, mRNA | 813 | 813 | 90% | 0.0 | 88% | XM_016919821.1 |

Using BLAST: Output

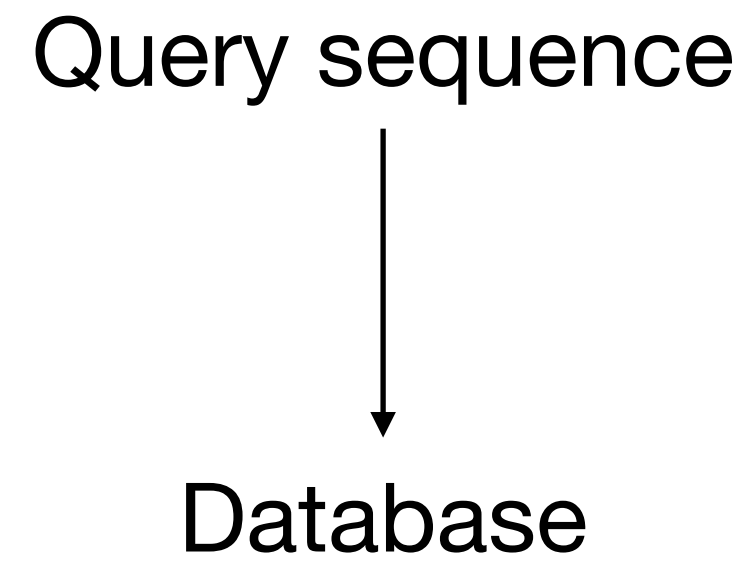
Sequences producing significant alignments:

Select: [All](#) [None](#) Selected:0

[Alignments](#) [Download](#) [GenBank](#) [Graphics](#) [Distance tree of results](#)

| | Description | Max score | Total score | Query cover | E value | Ident | Accession |
|--------------------------|---|-----------|-------------|-------------|---------|-------|--------------------------------|
| <input type="checkbox"/> | Homo sapiens NADH:ubiquinone oxidoreductase core subunit S8 (NDUFS8), mRNA | 830 | 830 | 90% | 0.0 | 89% | NM_002496.4 |
| <input type="checkbox"/> | Human mitochondrial NADH dehydrogenase-ubiquinone Fe-S protein 8, 23 kDa subunit precursor (NDUFS8) nuclear mRNA encoding mitochondrial protein, complete cds | 830 | 830 | 90% | 0.0 | 89% | U65579.1 |
| <input type="checkbox"/> | Homo sapiens mRNA for NADH dehydrogenase (ubiquinone) Fe-S protein 8, 23kDa (NADH-coenzyme Q reductase) variant, clone: KAT09377 | 828 | 828 | 90% | 0.0 | 89% | AK223114.1 |
| <input type="checkbox"/> | Homo sapiens NADH dehydrogenase (ubiquinone) Fe-S protein 8, 23kDa (NADH-coenzyme Q reductase), mRNA (cDNA clone MGC:149876 IMAGE:40119290), complete cds | 824 | 824 | 90% | 0.0 | 89% | BC119754.2 |
| <input type="checkbox"/> | PREDICTED: Gorilla gorilla gorilla NADH:ubiquinone oxidoreductase core subunit S8 (NDUFS8), transcript variant X1, mRNA | 819 | 819 | 90% | 0.0 | 89% | XM_019035670.1 |
| <input type="checkbox"/> | PREDICTED: Pan troglodytes NADH:ubiquinone oxidoreductase core subunit S8 (NDUFS8), transcript variant X1, mRNA | 813 | 813 | 90% | 0.0 | 88% | XM_016919821.1 |

Algorithm



How BLAST works

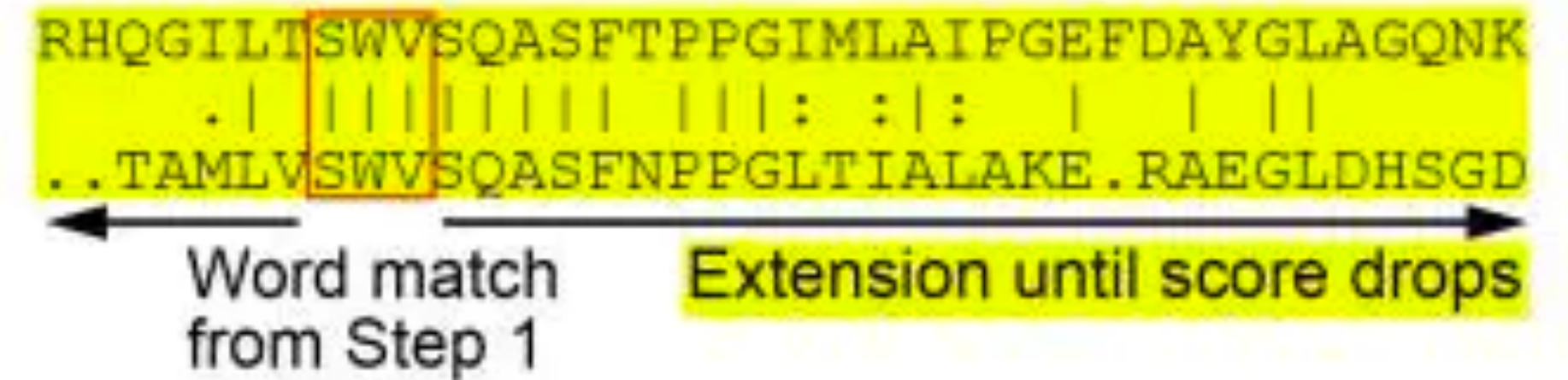


Extend the exact matches to high-scoring segment pair (HSP)

Fragmentation into words:



Extend matching hits in both directions



Selection of words scoring above threshold (for word SWV):

Substitution Matrix*

| | R | G | I | K | F | S | T | W | V |
|---|---|---|----|----|----|----|----|----|----|
| R | 5 | 0 | -1 | -1 | -2 | 1 | 0 | -3 | 0 |
| G | | 6 | -4 | -2 | -3 | 0 | -2 | -2 | -3 |
| I | | | 4 | -3 | 0 | -2 | -1 | -3 | 3 |
| K | | | | 5 | -3 | 0 | -1 | -3 | -2 |
| F | | | | | 6 | -2 | -2 | 1 | -1 |
| S | | | | | | 4 | 1 | -3 | -2 |
| T | | | | | | | 5 | -2 | 0 |
| W | | | | | | | | 11 | -3 |
| V | | | | | | | | | 4 |

- SWV (4+11+4 = 19)
- SWI (4+11+3 = 18)
- TWV (1+11+4 = 16)
- GWV (0+11+4 = 15)
- KWV (0+11+4 = 15)
- SWS (4+11-2 = 13)
- SFV (4+1+4 = 9)
- SRV (4-3+4 = 5)

Synonyms above threshold 11... (others not shown)

Synonyms below threshold 11... (others not shown)

*A portion of the BLOSUM 62 matrix

Query sequence: R P P Q G L F

Database sequence: D P P E G V V

↳ Exact match is scanned.

Score: -2 7 7 2 6 1 -1

↳ HSP

Optimal accumulated score = 7+7+2+6+1 = 23

$$S = \left(\sum M_{ij} \right) - cO - dG$$

E-Value: How a match is likely to arise **by chance**

- **The expected number** of alignments with a given score that would be expected to occur **at random** in the database that has been searched
 - e.g. if $E=10$, 10 matches with scores this high are expected to be found **by chance**

$$E = kmne^{-\lambda S}$$