

Name: Chan Kei Lam     SID: 1155221419

### Data & Data types (20 pts)- done

1a. Text

1b. convert to data matrix

| Student | Gender | Height (cm) | Weight (kg) |
|---------|--------|-------------|-------------|
| 1       | Male   | 180         | 75          |
| 2       | Female | 165         | 50          |
| 3       | Female | 164         | 49          |

2.

Sequential data (associated with regression) – prediction of mutations in infectious microbiome

Networks- to show the correlation between gene sequence and survival rate under different types of antibiotics

---

### Basic concept of Python programming (20 pts)- done

1a. import numpy which is an additional plug-in for Python

1b. Calculate the mean of the list [1,2,3] using numpy and assign the result to variable x

1c. Print the string 'x is ' followed by the value of x

2a. import numpy

2b. `z = numpy.array([1, 2, 3])`

2c.

|      |                                     |
|------|-------------------------------------|
| mean | <code>mean_z = numpy.mean(z)</code> |
| max  | <code>max_z = numpy.max(z)</code>   |

|     |                                   |
|-----|-----------------------------------|
| min | <code>min_z = numpy.min(z)</code> |
| SD  | <code>std_z = numpy.std(z)</code> |

### Sequence alignment & Dynamic programming (20 pts)-done

1a. (A-C), (C-C), (C-T), (G-\_)  $\rightarrow (-7) + 2 + (-5) + (-10) = -20$

1b. (A-\_), (C-C), (C-C), (G-T)  $\rightarrow (-10) + 2 + 2 + (-7) = -13$

1c. Since  $-13 > -20$ ... Thus, sequence in ( a.) would not be the best alignment we are looking for.

2. The best alignment score (similarity) between the two sequences is -10.

Best alignment: F(ACCG,\_CCT) :  $-13 = (-10) + 2+2+(-7)$

|   | -                           | A         | C         | C          | G   |
|---|-----------------------------|-----------|-----------|------------|-----|
| - | <b>0</b> $\rightarrow$<br>v | -10       | -20       | -30        | -40 |
| C | -10                         | <b>-7</b> | -8        | -18        | -28 |
| C | -20                         | -17       | <b>-5</b> | -6         | -16 |
| T | -30                         | -27       | -15       | <b>-10</b> | -13 |

### Sequence assembly & Sequence mapping (20 pts) -un

1. CGA one repeated, cannot be read : GGA, ATCG, ATC

|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| T | A | A | T | G | C | G | A | T | G | G | C | T | G | G | G | T | T | A | A | T |
| T | A | A |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   | G | C | G |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   | C | G | A |   |   |   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   | C | G | A |   |   |   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   | G | A | T |   |   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   | C | G | A | T |   |   |   |   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   |   |   | G | G | G |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   |   |   |   | G | G | T |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   |   |   | G | G | G | T |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   | G | T | T |   |   |   |

## 2. gene expression matrix

| Sample | count number                |
|--------|-----------------------------|
| Red    | 4 ( or 5 with one repeated) |
| Blue   | 4                           |

Data cleaning & Data visualization (20 pts) done

1. Sample-wise, to remove difference due to biological and technological variations, e.g 10000000

merger

| Sample name | Gene- 1 | Gene-2 | Gene-3 |
|-------------|---------|--------|--------|
| Sample 1    | 679     | -      | 260    |
| Sample 2    | 448     | 515    | 211    |
| Sample 3    | 873     | 621    | -      |
| Sample 4    | 408     | 365    | 164    |
| Sample 5    | 401     | 499    | 202    |
| Sample 6    | 800     | 605    | 699    |

2. sample-wise, as we want to see the difference between gene expressions under similar conditions more than the difference induced by samples, e.g. biological and technological variations

For the missing data:

Mean of Gene-2 ( except Sample-1 ) :  $(515+621+365+499+605)/5 = 521$

Mean of Gene-3 :  $(260+211+164+202+699)/5 = 307.2$

\

| Sample name | Gene- 1 | Gene-2  | Gene-3 |
|-------------|---------|---------|--------|
| Sample 1    | 679     | - → 521 | 260    |
| Sample 2    | 448     | 515     | 211    |

|          |     |     |           |
|----------|-----|-----|-----------|
| Sample 3 | 873 | 621 | - → 307.2 |
| Sample 4 | 408 | 365 | 164       |
| Sample 5 | 401 | 499 | 202       |
| Sample 6 | 800 | 605 | 699       |

3. Use Min-max normalization :  $v - V_{\min} / v_{\max} - v_{\min}$

| Sample name | Calculation                                           |
|-------------|-------------------------------------------------------|
| Sample 1    | Gene-2: $(521 - 260) / (679 - 260) \approx 0.623$     |
| Sample 2    | Gene-1: $(448 - 211) / (515 - 211) \approx 0.780$     |
| Sample 3    | Gene-2: $(621 - 307.2) / (873 - 307.2) \approx 0.555$ |
| Sample 4    | Gene-2: $(365 - 164) / (408 - 164) \approx 0.824$     |
| Sample 5    | Gene-1: $(401 - 202) / (499 - 202) \approx 0.670$     |
| Sample 6    | Gene-2: $(699 - 605) / (800 - 605) \approx 0.482$     |

Last:

| Sample name | Gene- 1 | Gene-2 | Gene-3 |
|-------------|---------|--------|--------|
| Sample 1    | 1       | 0.623  | 0      |
| Sample 2    | 0.780   | 1      | 0      |
| Sample 3    | 1       | 0.555  | 0      |
| Sample 4    | 1       | 0.824  | 0      |
| Sample 5    | 0.670   | 1      | 0      |
| Sample 6    | 1       | 0      | 0. 482 |