Metric for ranking perturbations

Notation Explanation:

Notation	**Description**	**Vector** Dimension
Po	Empirical gene expression distribution of un-perturbed cells	(15077,)
Pi	Empirical gene expression distribution obtained by knocking out gene “i”	(15077,)
Q	Desired cell state proportion vector	(5,)
Qo	Cell state proportion vector obtained under un-perturbed condition	(5,)
Q-hat	Predicted cell state proportion vector obtained	(5,)
S(.)	Statistic that summarizes a vector	-

Metric Expression:

2-step Metric derivation

Assumption:

Q Vector = [‘progenitor’, ‘effector’, ‘terminal exhausted’, ‘cycling’, ‘other’]

- **1st step** chooses a statistic that summarizes the 15,077-dimensional gene expression vector.
- For this particular metric we are going forward with the already defined Q vector i.e. the 5-dimensional vector which describes the cell state proportion (sums to 1).
- Hence,
  - S(Po) = Qo
  - S(Pi) = Q-hat
  - Q is the ideal cell state proportion vector that could be most efficient in killing cancer cells.
- S(.) can be any complex mathematical function or any **dense Neural Network** that maps 15,077 -dimensional gene expression vector to 5- dimensional cell state vector.

- **2nd step** is a scoring function that takes into picture the resultant Qs and ranks the perturbations based on its value.
- **Higher** the Scoring function value **better** the perturbation.
- Below is the scoring function which is basically the ratio of Euclidean distance between ideal cell state proportion vector, Q and predicted Q-hat to a constant benchmark.

Final Expression:

$$M = 3 - {E(Q, Q - hat)} \Big/ {E(Q, Qo)}$$

Where, $E(Q, Q - hat)$ denotes the Euclidean distance between Q & Q-hat vector and $E(Q, Qo)$ denotes Euclidean distance between Q & Qo vector.

Intuition

- **The numerator** explains how similar (or how far), is the predicted cell state vector from the ideal cell state vector in a 5-dimensional space.
- **The denominator** will be constant benchmark. The benchmark here is defined as the distance (reverse of similarity) between the ideal cell state vector and the obtained cell state vector without any perturbations.
- The entire fraction is subtracted from a constant (The Euclidean distance ratio in this case can’t exceed the constant) to **bound** the metric from value **[0,3]**.
- The metric “M” signifies how “similar” the predicted the Q-hat is to the ideal Q. The higher the **M** better rank the perturbation will receive.
- The key in this approach lies in choosing the ideal **Q.**

Distance Metrics

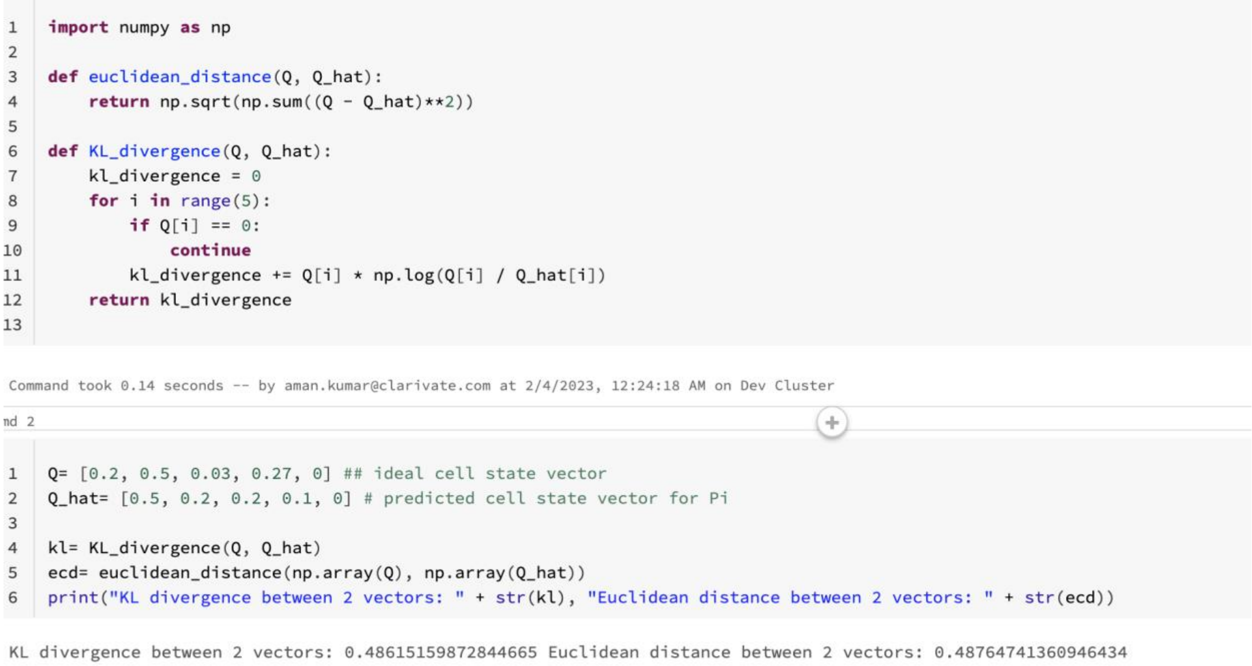
- The metric can be tweaked by choosing a different distance metric that calculates distance between two 5-dimensional probability vectors.
- A more general formula is below:

$$M = \ C - \ ^{D(Q, Q - hat)}\big/_{D(Q, Qo)}$$

, where **C** is a constant (the value should be equal or more the maximum value D.

D is distance metric.

Below is a simulation of comparison between 2 distance metrics: Euclidean distance & KL-divergence.



```
import numpy as np

def euclidean_distance(Q, Q_hat):
    return np.sqrt(np.sum((Q - Q_hat)**2))

def KL_divergence(Q, Q_hat):
    kl_divergence = 0
    for i in range(5):
        if Q[i] == 0:
            continue
        kl_divergence += Q[i] * np.log(Q[i] / Q_hat[i])
    return kl_divergence

```

Command took 0.14 seconds -- by aman.kumar@clarivate.com at 2/4/2023, 12:24:18 AM on Dev Cluster

nd 2

```
Q= [0.2, 0.5, 0.03, 0.27, 0] ## ideal cell state vector
Q_hat= [0.5, 0.2, 0.2, 0.1, 0] # predicted cell state vector for Pi

kl= KL_divergence(Q, Q_hat)
ecd= euclidean_distance(np.array(Q), np.array(Q_hat))
print("KL divergence between 2 vectors: " + str(kl), "Euclidean distance between 2 vectors: " + str(ecd))
```

```
KL divergence between 2 vectors: 0.48615159872844665 Euclidean distance between 2 vectors: 0.48764741360946434
```

Proposal for Ideal Q (optional)

Q= [0.2, 0.5, 0.03, 0.27, 0]

The above Q is chosen to keep **effector** state at its maximum, **terminal exhausted** at its lowest and **cycling state** being subsequently high so that quite a high number of cells are produced.

Limitation:

- The limitation of above approach (Metric **M)** is that it restricts Ideal Q to a specific number. There are many possibilities that the Q can be something different yet as strong in killing cancer cells.

- The future lies in improving the metric where Q is a probability density map & not a fixed 5-dim vector.