1. Summarization Statistic

> **Notes:**
>
> 1. As per the requirements : "For experimental design purposes, the statistic $S(.)$ should be predictable for unseen perturbations". Hence, the summarization statistics is designed to generalize over all the genes knock-out to predict the gene expression summary that best describes the T cell states it will induce. Hence the information for which gene was knocked out to obtain P_i is ignored.
> 2. In the gene expression matrix we **notice some of the genes are never expressed**. We have data from experiments on only 73 genes and zero can also occur from technical reasons , hence, we don't know whether those genes expression has any significance (*i.e. will be expressed and influence cell state for the remaining gene knockouts*).
> 3. In the gene expression matrix though there is evidence of group of genes being expressed together, this co-dependency is not used while summarizing. Because, for unseen perturbations the co-dependency might not be present in the gene expression matrix .

Given a cell distribution P_i , the summarization strategy is as below :

- It selects the all the genes
  1. whose mean expression is larger than $T_{Expression}$, and
  2. whose variance across cell states is larger than T_{State}) , and
  3. whose variance across cells in a population obtained from knocking out a gene is smaller than T_{Cell}
- And the genes which are never expressed in any of the gene expression (Ref. Note 2 above)

It produces $\hat{P}_i$ that has reduced dimensions, i.e. has N_p number of genes where $N_p < 15077$

---

2. Scoring Function For Perturbation

(Single Cell Perturbation Scoring)

The scorer for perturbation takes $\hat{P}_i$, P_0 and Q as input. And then outputs the $score$ of $\hat{s}(P_i)$, where higher score indicates the P_i (corresponding to $\hat{P}_i$) is more likely to induce the desired cell state Q .

$$score = \frac{\sum_{j=1}^{|\hat{P}_i|} \left(\frac{G_j}{\overline{P_{0j}}} \times V_j \times \sum_{x=1}^{5} (C_{jx} \times Q_x) \right)}{|\hat{P}_i|}$$

> - G_j : j -th gene expressed value from summarized gene expression $\hat{P}_i$
> - $\overline{P_{0j}}$: Corresponding j -th gene's **mean** expressed vlaue from unperturbed cells P_0 (all unperturbed cells).

Where,

Inputs:

- P_0 : Empirical gene expression distribution of the unperturbed cells
- $\hat{P}_i$: Summarized P_i for a *single cell* [1] as processed by summarization strategy. It is dimension reduced P_i .
- Q : Desired cell state proportion vector (5 dimensional vector which sums to 1)

Coefficients:

These coefficients are calculated from the perturbation experiments performed and resulting gene expression matrix.

- V_j : Variance factor of j -th gene in gene expression. It decides the cell state boundaries' confidence.

  > Note: **this variance does not depend on the knowledge of which gene was knocked out, hence can be used for unseen knockouts.**

  It is calculated as below.

  > $$V_j = \frac{\mathrm{Var}_{x=1}^{5}\left(\frac{\sum_{k=1}^{n_x} G_{jxk}}{n_x}\right)}{\frac{1}{5}\sum_{x=1}^{5}\left(\mathrm{Var}_{k=1}^{n_x}(G_{jkx})\right)}$$
  >
  > Where,
  >
  > n_x = count of cells for **x**-th cell state
  >
  > Var : Variance function
  >
  > **Numerator** is variance of mean expression of j -th gene observed across **different cell states**. it measures how much the expressed value of j -th gene (in the gene expression matrix) varies across cell states. This parameter governs the confidence of classification boundary , where larger value indicates **more** confidence.
  >
  > **Denominator**: is variance of jth gene expression observed across **n different cells for same cell state** and then meaned for 5 cell states. In other words, it measures how much the expressed value of j -th gene (in the gene expression matrix) varies across cells of a collection yielding same cell state. This parameter governs the confidence of classification boundary , where larger value indicates **less** confidence. [2]

- C_{jx}: Co-variance of gene j for x -th cell state. [x is between $[1,2,3,4,5]$ for $(\mathrm{progenitor}, \mathrm{effector}, \mathrm{terminal\ exhausted}, \mathrm{cycling}, \mathrm{other})$].

  This is added to favor the perturbation towards desired cell state proportions : Q.
- Q_x : Proportion of x -th cell state in desired cell state vector Q .

  > The factor $\sum_{x=1}^{5} C_x \times Q_x$ favors the perturbation towards desired cell state

Alternative Version:

(Cell Collection Perturbation Scoring)

There's suggestion in the challenge description, where it mentions : "*scoring function may also take into account the predicted number of cells resulting from a guide* and favor perturbations with a large growth rate".

Hence, an alternative version of this scorer is proposed, where , P_{ic} (collection of cells from a particular gene knockout) is used in stead of P_i.

This basically uses the scorer function described above to get score of individual cell's gene expression and then calculates mean score . This mean score is multiplied with $(1+\frac{n}{N})^f$ to favor perturbations with a large growth rate .

Given,

- P_{ic} : gene expression matrix for all the cells resulting from perturbation where gene i is knocked out. $[15077 \times n]$
- n : Count of cells in the collection where gene i is knocked out.
- N : Length of total cells in Perturb-Sequence dataset.
- f : favor strength for growth factor, describes the degree of favor to show for larger growth rate. *Default : 1*

$$\boxed{\widehat{score} = \frac{\sum_{w=1}^{n}\left(score\Big(summarize(P_{ciw}), Q, P_0\Big)\right)}{n} \times (1+\frac{n}{N})^f}$$

> P_{icw} : gene expression of w -th cell from P_{ic}

---

3. Scoring Function For Cell States

This scorer is used to measure the cell state proportion vector Q_i observed for knocking out gene i.

$$\boxed{score = \frac{\sum_{x=1}^{5} log(1+\frac{Q_{ix}}{Q_x+\epsilon})}{\sum_{x=1}^{5} log(1+\frac{Q_{0x}}{Q_x+\epsilon})} \times (1+\frac{n}{N})^f}$$

Given

- Q_i : Cell state proportion[^*] observed for knocking out gene i (5 dimensional vector which sums to 1).

> * : These proportion is derived from cell populations where same gene was knocked out : Co-variance of gene i for x -th cell state. [x is between $[1,2,3,4,5]$ for $(\text{progenitor,effector,terminal exhausted,cycling,other})$]

- ϵ : A very small value to prevent division by zero. (Example. $1e^{-12}$)
- Q_0 : Cell state proportion* observed for unperturbed T cells.

- Q : Desired cell state proportion vector (5 dimensional vector which sums to 1)
- n : Count of cells in the collection where gene i is knocked out.
- N : Length of total cells in Perturb-Sequence dataset.
- f : favor strength for growth factor, describes the degree of favor to show for larger growth rate. *Default : 1*

Proposal For Q (Desired cell state proportion vector)

While choosing A desired Q following things should be considered:

1. The T cell states are currently thought to be mutually exclusive. In other words for collection of cells, each state probability obtained is thought be independent of other states . But tis might not be case in reality. For example, If a collection yields higher number of progenitor states , it might also yields higher number of effector states (compared to other 3 states). Hence, simply minimizing effector while maximizing progenitor might not be good idea if it's true.
2. The absolute count of progenitor cells in collection is also should be considered. For example, Let's consider a *total 35 cells being perturbed*.

   Experiment A (Knocked out Gene_a)

Perturbation	progenitor	effector	terminal exhausted	cycling	other
Gene_a	7	0	0	1	1
Unperturbed	4	3	14	2	3

   Experiment B (Knocked out Gene_b)

Perturbation	progenitor	effector	terminal exhausted	cycling	other
Gene_b	20	3	2	2	1
Unperturbed	1	2	3	0	1

   (Each table cell represent count of T cell in the state)

   Although, Gene_a has higher probability of progenitor than Gene_b , but Gene_b knockout produces more progenitor cells in total.

References, Citations:

> *"the objective function in Part 3 in Challenge 2 is required to score over the 5 dimensional cell state proportion vectors"*
>
> *"Additionally we did leave the potential metric space in Challenge 3 intentionally open, so the metrics submitted for Challenge 3 can allow for distnace measurements over the full gene expression space."*
>
> — *YitongTseo* [Reference]

---

1. https://discussions.topcoder.com/discussion/25781/scorer-doubts ↩

2. *"scoring function should ideally take uncertainty into account given by the different sample sizes for the perturbations in the training dataset"* [Reference]. ↩