Exam of Bioinformatics

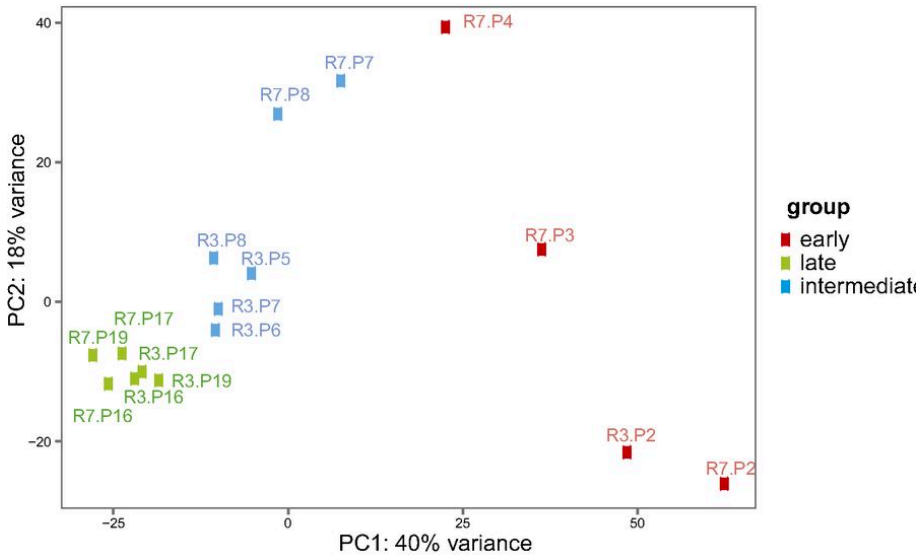
14/04/2019

Surname:
Name:
Identification number:

Question 1 — Score: 5 pts

Explain the main features of the Principal Component Analysis.
The following Figure shows a PCA analysis of 16 samples of two cell lines R3 and R7 at early, late and intermediate time points.

1) Is it possible to clearly group the samples belonging to R3 with respect to those belonging to R7 lines? If yes, which is the component that lead to this separation?
2) Is it possible to clearly group the samples considering the time points? If yes, which is the component that lead to this separation?
3) Explain the meaning of the percentage values associated with each component.



Question 2 — Score: 4 pts

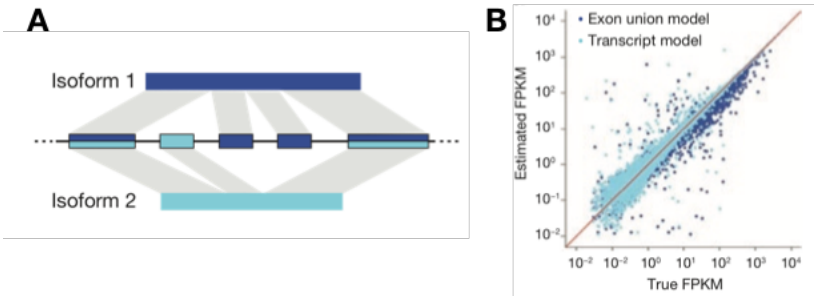
How is computed the *Enrichment Score* in the Gene Set Enrichment Analysis?

Question 3 — Score: 4 pts

In the gene fusion detection, how are the *spanning* and the *encompassing* reads used?

Question 4 **Score: 5 pts**

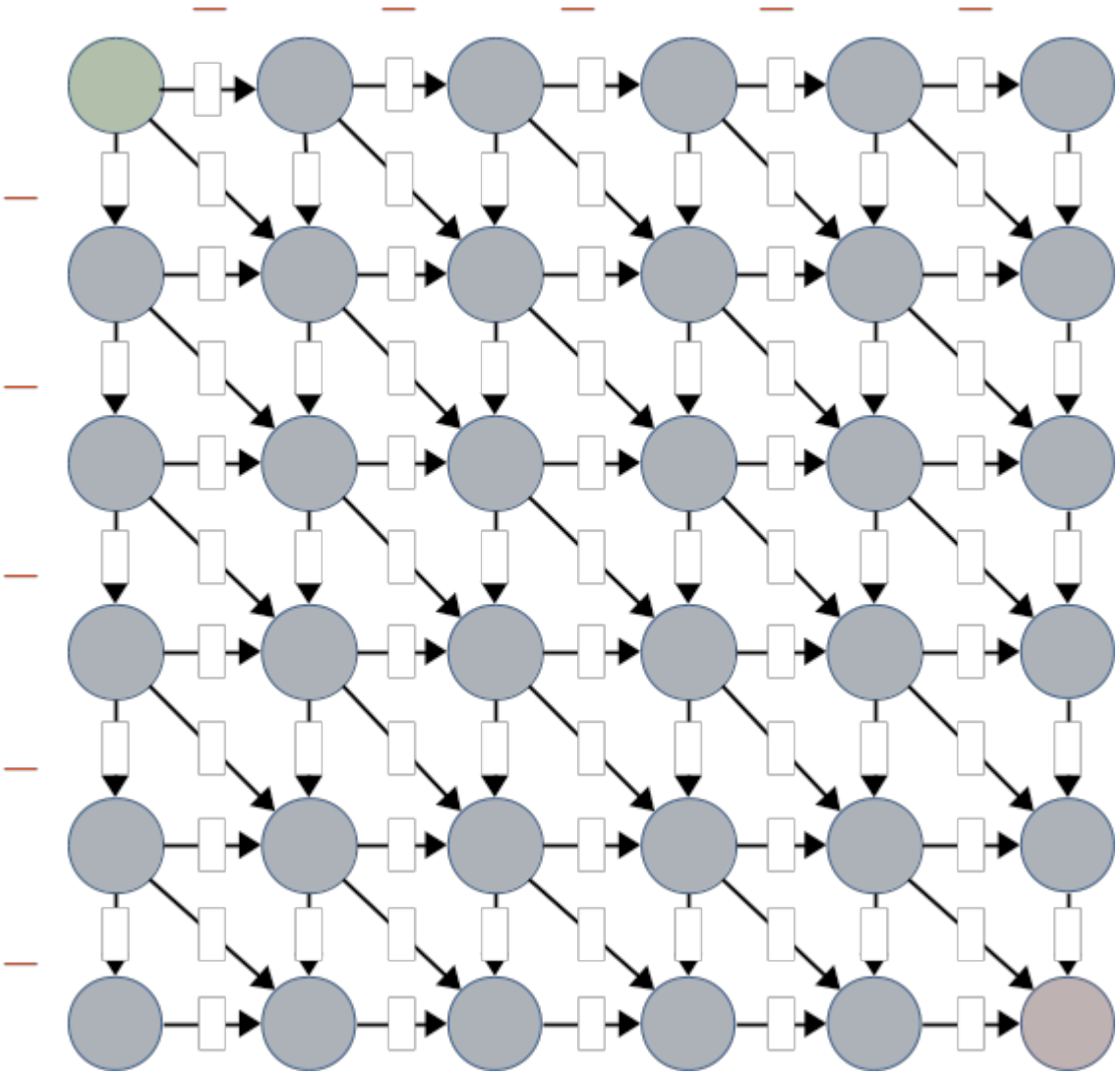
Figure A shows two gene isoforms. The quantification of these isoforms can be done by *exon intersection method* and by the *exon union method.* Explain how these quantification methods work and comment the plot in Figure B discussing the performance of the two methodologies.



Question 5 **Score: 3pts**

- Align locally the following sequences:
  - ACTGG
  - CGTAG

  assuming the following scores: Match = +3 Mismatch = -2 In/del= -1



Question 6 **Score: 4pts**

Align globally the following sequences:

- Apply the K-means algorithm on the following inputs:
  - k=2;
  - $P_1(1,3)$, $P_2(4,3)$, $P_3(1,4)$, $P_4(4,1)$; $P_5(2,4)$ P(1,1)
  - use **Manhattan distance**;
  - select as starting point $\mathbf{P_1}$ and $\mathbf{P_3}$

Question 7 **Score: 3pts**

- Describe the main difference between **K-means and Hierarchical clustering**

Question 8 **Score: 3pts**

- Describe how **complex scoring** for indel (i.e. penalty for opening a gap + penalty for extending a gap) are introduced in sequence alignment