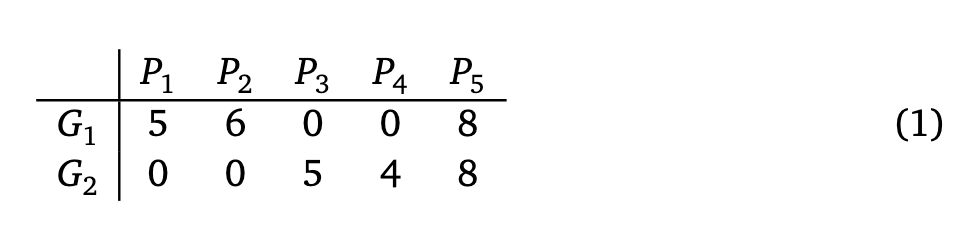
**Übungsblatt 7**  
Genexpressionsanalyse II  
Deadline: 29.6.2022 um 20:00 MEZ  
Bioinformatik für Biochemiestudierende  
Dr. Florian Klimm  
Sommersemester 2022

**Aufgabe 7.1 k-means Clustering (30 %)**  
*Sie haben die Expression zweier Gene in fünf verschiedenen Patienten bestimmt  
und erhalten*

*  
Führen Sie ein k-means Clustering durch. Als Initiale Zentren der Cluster nutzen  
Sie (1, 2), (4, 4), und (8, 8).  
Berechnen Sie die Within-cluster-sum-of-square (WSS) für das erhaltene Clustering  
nachdem der Algorithmus konvergiert ist.*

Idee des k-means Clustering:

Man legt erst die Anzahl an Clustern fest und ermittelt die Zentren der Cluster, bevor man die Werte den Zentren zuordnet. Dann werden die Zentren basierend auf den Werten aktualisiert und es werden wiederum die Werte neu den Zentren zugeordnet. Diese Prozedur wird bis zur Konvergenz wiederholt (dann ändern sich die Koordinaten der Zentren nicht mehr).

1. **K-means Clustering**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Zentrum 1 1 (1,2) | Zentrum 2 (4,4) | Zentrum 3 (8,8) | Zu welchem Cluster gehören die Werte? |
| P1(0,5) |  |  |  | Cluster 1 |
| P2(0,6) |  |  |  | Cluster 1 |
| P3(5,0) |  |  |  | Cluster 2 |
| P4(4,0) |  |  |  | Cluster 1 |
| P5(8,8) |  |  |  | Cluster 3 |

Berechnung euklidischer Distanzen:

Unter Benutzung der Formel

Weise die Werte den am nächsten liegendsten Clustern zu:

Aktualisiere die Zentren basierend auf den Werten:

🡪 bilde den Mittelwert aller x und y für jedes Cluster, um die neuen Zentren-Koordinaten zu erhalten

Ordne die Werte den Zentren neu zu:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Zentrum 1 (1.33,3.66) | Zentrum 2 (5,0) | Zentrum 3 (8,8) | Zu welchem Cluster gehören die Werte? |
| P1(0,5) |  |  |  | Cluster 1 |
| P2(0,6) |  |  |  | Cluster 1 |
| P3(5,0) |  |  |  | Cluster 2 |
| P4(4,0) |  |  |  | **🡪 Cluster 2** |
| P5(8,8) |  |  |  | Cluster 3 |

Aktualisiere erneut die Zentren basierend auf den Werten:

🡪 bilde den Mittelwert aller x und y für jedes Cluster, um die neuen Zentren-Koordinaten zu erhalten

Konvergenz ist erreicht, da sich die Koordinaten der Zentren nicht mehr ändern.

1. **Within-cluster-sum-of-square (WSS)**

* Euklidische Distanz zwischen Wert und Zentrum
* Bestimme diese für jeden Wert eines Clusters
* Summiere alle Werte des Clusters

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Zentren |  |  |  |  |  |
| G2 | G1 | G2 | G1 | G2 | G1 |
| 0 | 5,5 | 4,5 | 0 | 8 | 8 |
|  | | | | | |
| Werte |  |  |  |  |  |
| Cluster 1 |  | Cluster 2 |  | Cluster 3 |  |
| G2 | G1 | G2 | G1 | G2 | G1 |
| 0 | 5 | 5 | 0 | 8 | 8 |
| 0 | 6 | 4 | 0 |  |  |

**Aufgabe 7.2 Gene Ontologies (20 %)**  
*Sie haben die folgenden Gene in ihrem Experiment mit* Homo sapiens *Zellkulturen  
identifiziert:*  
APH1A ARRB1 CCND1 CUL1 DLL1 DTX1 DTX2 DTX4 FBXW11 FZD1 FZD5 FZD7  
HES1 HEYL JAG1 KAT2A LFNG MAML2 NOTCH1 NOTCH2 NOTCH3 PPARD PRKCA  
PSEN2 PSENEN RBX1 SAP30 SKP1 ST3GAL6 TCF7L2 WNT2 WNT5A  
*Nutzen Sie das Online Tool unter* http://geneontology.org/ *um ein Gene*  
*Ontology Enrichment durchzuführen.  
Welche Biologischen Funktionen können Sie dieser Gruppe an Genen zuordnen?  
Nennen Sie mindestens* ***fünf charakteristische biologische Prozesse****.*

* Negative regulation of inner ear auditory receptor cell differentiation
* Negative regulation of pro-B cell differentiation
* Cardiac right atrium morphogenesis
* Lateral inhibition
* Compartment pattern specification

*Recherchieren Sie warum nicht nur ein p-Wert sondern auch eine False Discovery  
Rate angegeben wird.*

Die False Discovery Rate (FDR) ist eine Form der Multiple Testing Correction, welche den erwarteten Anteil falsch positiver Ergebnisse angibt. Man kann in PANTHER17.0 2 verschiedene Arten der Multiple Testing Correction wählen – die FDR oder aber die Bonferroni-Korrektur, die wir bereits kennengelernt haben. Eine klassische Multiple Testing Correction kontrolliert die Wahrscheinlichkeit eines einzelnen falsch positiven Ergebnisses – ein relativ strenges Vorgehen, welches auch falsch negative mit sich bringen kann. Die sogenannten Benjamini-Hochburg False Discovery Rate hingegen ist weniger stringent im Vergleich zu der Bonferroni-Korrektur, und ermöglicht es einem somit, signifikante Ergebnisse für Analysen wie die GO Term Analysis zu erhalten (natürlich mit dem Nachteil, dass die Wahrscheinlichkeit für falsch Positive erhöht ist).

**Aufgabe 7.3 Gene-set Enrichment (20 %)**  
*Sie haben eine Liste von 50 Genen als differentiell expressiert identifiziert. Von  
diesen haben 40 die biologische Funktion Apoptose. Der Gesamte Datensatz besteht  
aus 500 Genen von denen 100 die biologische Funktion Apoptose haben.  
Verwenden Sie den χ2-Test um festzustellen ob in den von Ihnen identifizierten  
Gene signifikant solche mit der Funktion Apoptose vorhanden sind.  
Die kritischen Werte de χ2-Verteilung können Sie unter* https://davidmlane.  
com/hyperstat/chi\_square\_table.html *einsehen.*7 – 1

**Beobachtungen:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Differenziell exprimiert | Nicht differenziell exprimiert |  |
| Apoptose-assoziierte Gene | 40 | 60 | 100 |
| Nicht Apoptose-assoziierte Gene | 10 | 390 | 400 |
|  | 50 | 450 | 500 |

**Ermittlung der Erwartungswerte**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Differenziell exprimiert | Nicht differenziell exprimiert |  |
| Apoptose-assoziierte Gene | 40 | 60 | 100 |
| Nicht Apoptose-assoziierte Gene | 10 | 390 | 400 |
|  | 50 | 450 | 500 |

**Berechnung der Test-Statistik:**

Mit

Wo nr die Zeilensumme und nc die Spaltensumme ist.

**Aufgabe 7.4 Analyse eines Datensatzes mit R (30 %)**  
*In dieser Aufgabe werden wir einen Datensatz untersuchen der verschiedene  
Charakteristiken von Pinguinen beschreibt.  
1. Laden Sie die Daten mit dem PALMERPENGUINS Package  
2. Inspizieren Sie die Daten und beschreiben Sie die verschiedenen Spalten der  
Tabelle.  
3. Führen Sie eine PCA Analyse durch und visualisieren Sie in 2D  
4. Führen Sie ein k-means Clustering durch. Identifizieren Sie den optimalen  
Wert von k.  
Hinweis: Für diese Analyse können Sie verschiedene Packages benutzen. Recher-  
chieren Sie welche dafür brauchbar sind*