CS426: Introduction to Computational Biology

Section 12

Homework Six

- What you have to do:
  - Implement from scratch (de novo) the k-means clustering algorithm
  - Use it on the *Saccharomyces cerevisiae* (baker's yeast) expression data provided on the website.
  - Interpret the results

Homework Six

- What to give us:
  - The code resulted.
  - The best clusters obtained
  - You should answer the question:
    - ***Do these clusters have a biological meaning?*** (Use the gene description provided to establish functional relations between members of the same cluster.)

Saccharomyces cerevisiae Data

- Two files are provided:
- yeast.def :
  - the file contains the definitions of the genes in yeast for which we have expression data.
  - This file contains a number (1 – 6300) which was assigned to each gene, the Swissprot accesion number, and a short description including the name

```
38   (P39708) NADP-specific glutamate dehydrogenase 2 (EC 1.4.1.4)
(NADP-GDH 2).
```

Saccharomyces cerevisiae Data

- yeast:
  - Each line contains the expression data of a gene over 70 something experiments (yeast cell cycle data)
  - Each number represent the expression data in a specific experiment and the values on a line are separated by space.
  - Each line corresponds with the line with the same number form the *yeast.def* file
  - Eg: `-0.38 -666 -0.23 -0.17 -0.31 -0.16 0.07`
  - Missing data: the value –666 stands for missing data (dark humor? ☺)

K – Means Clustering Algorithm

- Input:
  - k - number of clusters;
  - The n data points

1. initialize the k centroids one for each cluster
2. repeat
   1. Classify each sample to the closest cluster (closest centroid)
   2. Recompute the centroid of each cluster as the average of the samples classified to that class
3. until means (centroids) converge

Problems with K - Means

- Initialization of the centroids
- What means closest:
  - Euclidian distance
  - Pearson correlation
  - Missing values
- Recomputing the centroids
- Termination:
  - When to stop?
  - What is the best cluster?
  - Repeat again? Why?

Initialization of the centroids

- The algorithm is based on an initial guess of the centroids.
- The quality of this guess often influence the speed of convergence and the quality of the final clustering.
- Ways of choosing the centroids:
  - At random – randomly pick points in the input space.
  - Randomly pick k data points from the data set to be the initial choice for centroids (why is this better ?)

Random Initialization

- The results can vary with the random choice.
- We are more confident that we reached a global optimum if we reach it several times from different positions.
- If we have also a function to evaluate the clusters we can pick also a best scoring clustering.
- We have to run the algorithm several times and pick the most frequent clustering or the best scoring clustering.

What means closest?

- In order to define closest we need a distance measure.
- We have defined two ways to compare expression data:
  - Normalized Euclidian distance;
  - Pearson Correlation
- There are other ways also.

Normalized Euclidian Distance

- Defined as:

$$Dist_{euc}(\mathbf{V},\mathbf{U}) = \sqrt{\frac{1}{d}\sum_{i=1}^{d}(V_i - U_i)^2}$$

- Why normalized?

Pearson Correlation

- Defined as:

$$Corr(\mathbf{V},\mathbf{U}) = \frac{1}{d}\sum_{i=1}^{d}\frac{(V_i - < V >)(U_i - < U >)}{\sigma_V \sigma_U}$$

- Where σ_V and σ_U are the sample standard deviation of the expression profile V and U respectively
- How to transform it into a distance metric?
  - $d_p(V, U) = 1 - Corr(V, U)$

Missing Values

- We will use available data.
- We have to pay attention to the normalization.
  - The d value takes the value of the number of common known features.

Recompute the centroids

- For each cluster we take the mean of each feature separately in the current cluster and they become the new coordinates of the new centroid.

Termination Criteria

- The centroids converge – the change in the centroid position (the distance between old centroids and new centroids) is lower then a predefined threshold (let's say 0.01).
- We can use the average change in distance of all centroids or we can enforce that all the changes should be less then the threshold
- The algorithm however is not guaranteed to converge so it is wise to also limit the number of iterations

What Is the Best Clustering?

- This is a very hard question to answer.
- Usually we don't know the underlying patterns we are trying to extract so we don't know what is the best clustering even if we see it. Also we don't know the possible number of clusters.
- We have to try multiple approaches and if we get roughly the same clustering we might consider it good enough.
- There are some methods to assess the clustering quality but they are not in the scope of this assignment

How to Assess the Quality of the Clusters

- Due to the time constraints we will focus on a different measure of the Cluster quality:
  - The average distance between the points in the same cluster

What to Do for the Homework

- Run the algorithm for k =100 number of clusters
- Use both Euclidian distance and Pearson correlation metrics. Compare the clusters obtained using both.
- For missing values use only the available data.

What to Do for the Homework

- Pick the five “best” nontrivial clusters (at least 5 members.
- Draw the expression profile of each gene in the cluster overlapping one another.
- Try to see if the clusters selected make sense based on the description of genes.

Implementation Details

- You are not allowed to use any library containing k-means the algorithm should be implemented from scratch.
- You are allowed to use only one of the following programming languages: Matlab. Java, C/C++. Please don’t use Python, Perl or SML.
- Take care that the experiments can be quite computational intensive leave some time to do them start well in advance.

Enzymatic reactions

- Most the functions of the cell are chemical reactions. Most of this reactions are catalyzed by enzymes.
- The substrates – the chemical components on which the operates (inputs)
- The products – the result of the reaction (output)
- The enzyme – the protein which helps the reaction to take place.

Enzymatic reactions

- Each enzymatic reaction can be catalysed by more then one protein.
- Enzymatic reactions are the same in different organisms but the enzymes are different.
- We classify the enzymes based on the reaction it catalyzes

Biochemical Pathway Databases

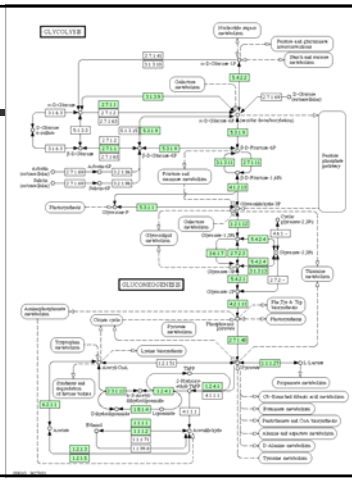
- The chemical reactions in cells do not occur in isolation but are organized in multistep sequences called ***pathways***.
- The product of one reaction in this chain is the substrate for the subsequent reaction
- The same (with small variations) pathway may appear in different organisms.

Pathway databases

- In the last decade databases have been developed for the pathway information:
  - The reactions it is composed of;
  - The compounds of the reactions;
  - Information on the reactions;
  - Enzymes which catalyzes the reactions;
  - Organisms in which they were observed;
  - Organisms in which they might exists (predicted)

Pathway databases

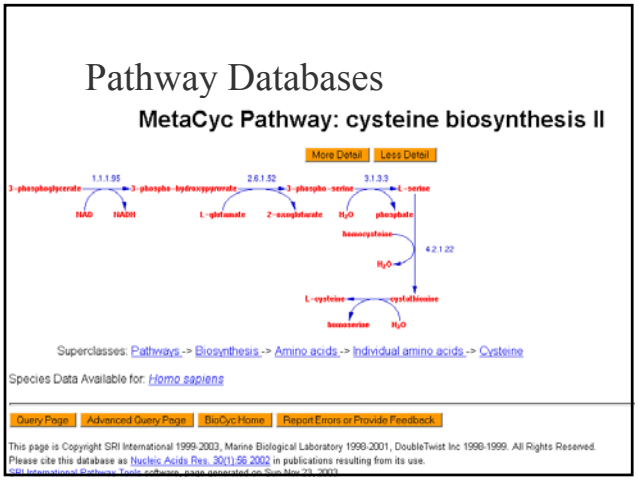
- Kegg (http://www.genome.ad.jp/kegg/)
  - Japanese site it is based on digitized biochemical information (huge charts developed by biochemists)
  - The database contains a lot of information about the genomes of the organisms in which the pathway are predicted.
  - The prediction of the pathways is done by filling in the blanks in the huge charts.



Pathway databases

- Biocyc (biocyc.org)
  - The database is centered on empirical proven Chemical pathways
  - It is centered around the concept **Pathway/Genome Database**
  - Prediction of pathway is also based on filling in the blanks in the template pathways (already known to exists

Pathway Databases



Pathway databases

- Other databases:
  - WIT

Pathway databases

- Why is important to study this?