UCS2735 BIOINFORMATICS TECHNOLOGIES

Unit 5

Microarray Data Analysis

Microarray Data Analysis: Microarray technology for genome expression study - Image analysis for data extraction - Data Analysis for Pattern Discovery

* Microarray Data Analysis

→ Microarrays deal with DNA samples.

→ It is a device that contains an array of microscopic glass slides coated with DNA molecules.

→ Using this microarray, expressions of genes are measured using the laser spectroscopic technique.

→ In microarray analysis:

(i) DNA samples are placed on a high-density DNA chip.

(ii) Gene expression is measured by detecting how much of a specific gene is expressed in the sample. Laser spectroscopy helps identify the amount of fluorescent signal from each DNA spot.

(iii) Image array analysis is essential because it provides detailed data about each element on the microarray.

→ Varials analytical methods such as cluster analysis, temporal expression profile analysis and gene regulatory analysis are applied to interpret the data.

→ Microarray experiments generate a large volume of image data ~~from the microarray images~~. Analyzing this data requires the use of computer algorithms to extract meaningful information

* Analysis of Microarray Data

→ can be categorized into two:

(i) Image Analysis for Data Extraction - involves extracting data from the microarray images, identifying gene expression levels from each spot on the array.

(ii) Data Analysis on Gene Expression Ratios - involves comparing the expression levels of genes to understand their relative activity, which helps in identifying patterns of gene expression under different conditions

* Challenges in Microarray Image Analysis

→ poor contract between spots and background

→ many contamination / arhfacts such as:

(i) irregular spot shape and size

(ii) dust on the slide

(iii) large intensity variation within the spots and the background

(iv) non-specific hybridization

* Gene Expression Study

→ Gene expression profiling is a process used to determine when and where genes are expressed.

→ By understanding gene expression, scientists can gain insights into gene function, which can reveal how genes are turned on or off in response to various conditions.

→ Factors affecting gene expression are:

(i) external stimuli
(ii) cell type and development stage
(iii) cell cycle
(iv) mutations
(v) regulations by other genes

* Issues with Traditional Gene Expression Methods

→ Traditional gene expression methods are slow and may have good sensitivity, but aren't sufficient for studying the entire genome at once, as different genes are interdependent.

* Microarray Technology for Genome Expression Study

→ Microarray technology allows massively parallel, high-throughput analysis of gene expression in a single statement.

→ can measure the expressions of tens of thousands of genes simultaneously

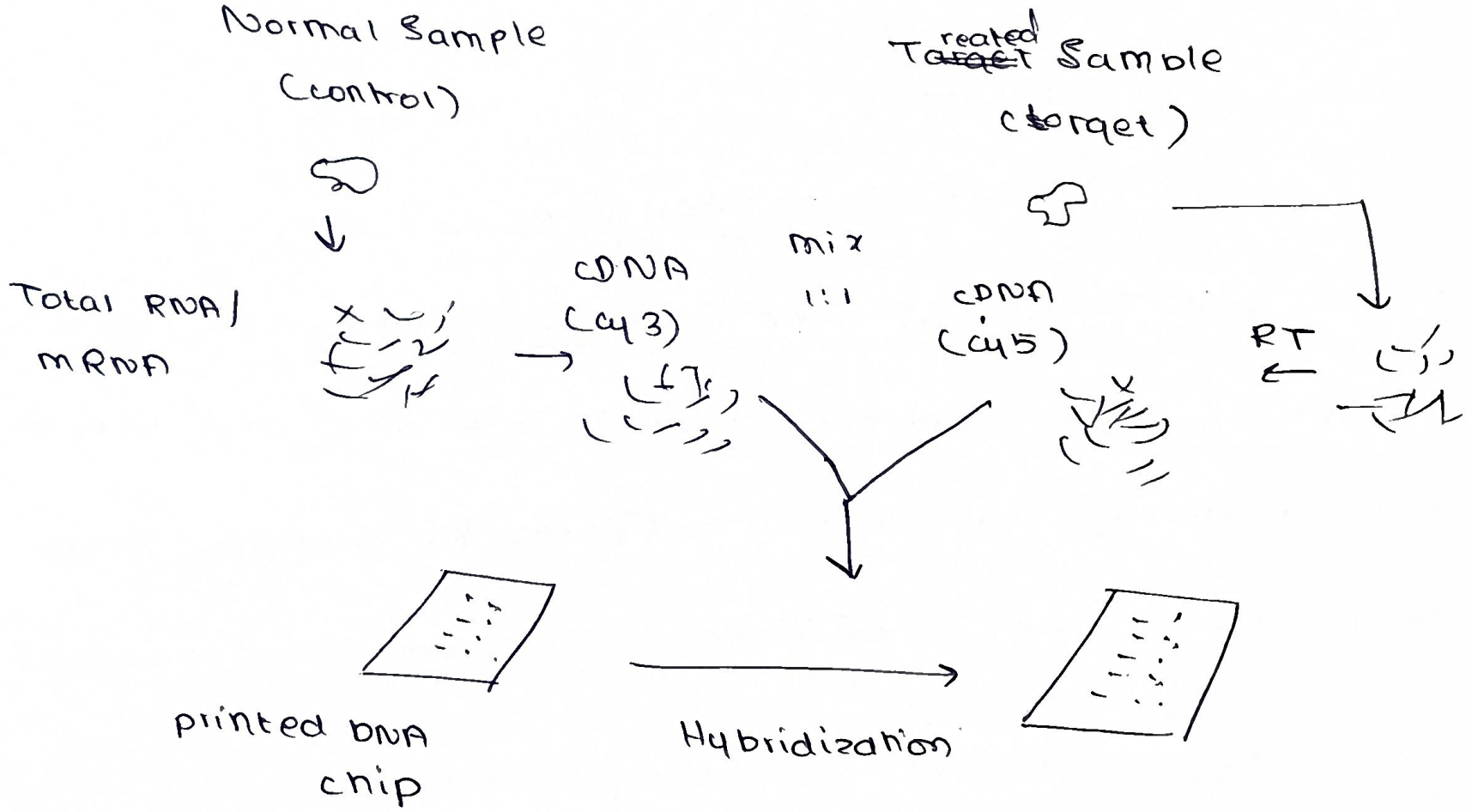
→ used in pharmaceutical and clinical research

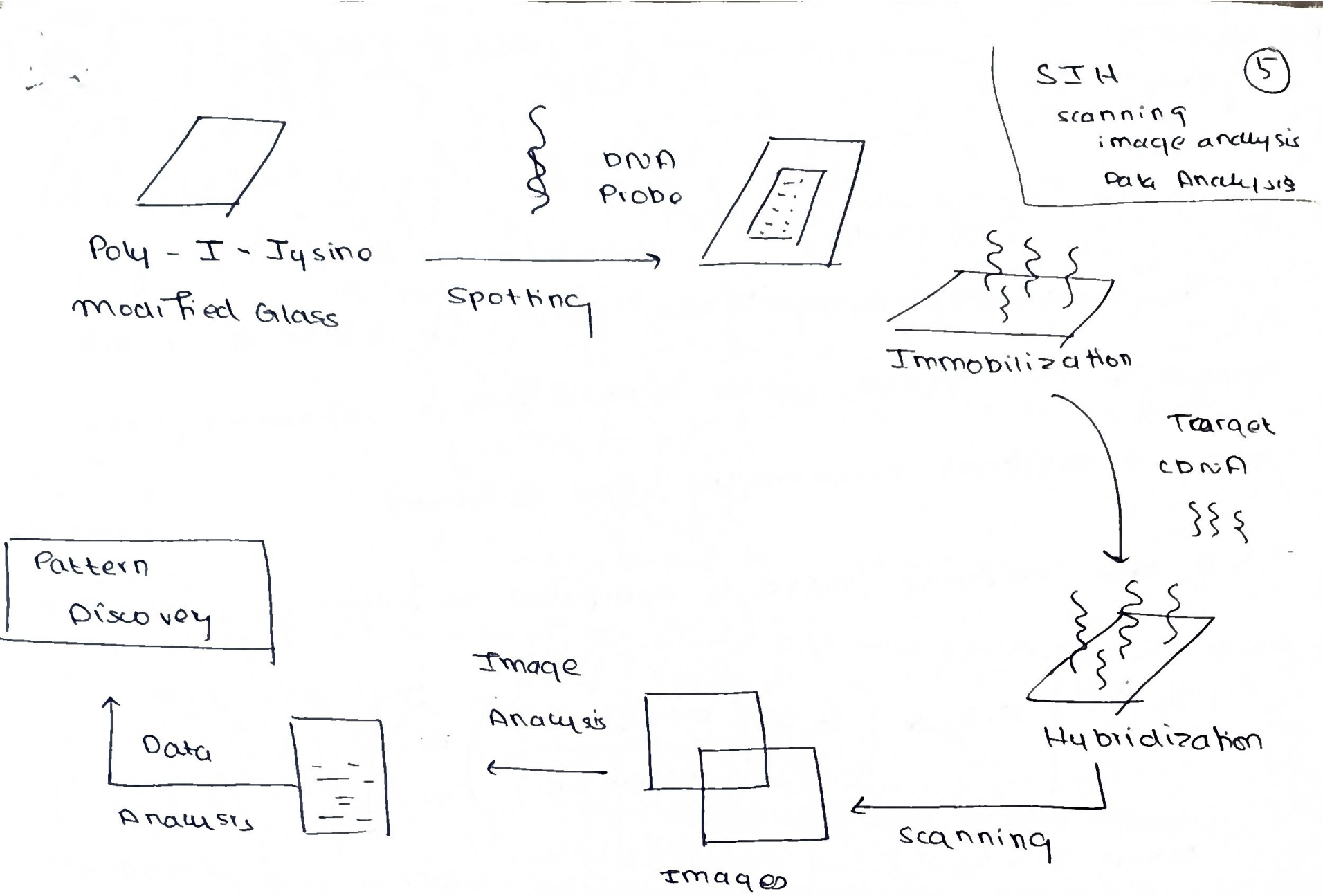
Process :-

1. Compare a normal (control) sample with a treated (target) sample.

2. Both samples' cDNA are labelled w/ fluoroscent dyes (Cy3 & Cy5 and then mixed and hybridized into a DNA chip

3. The expression ratio is calculated by comparing the fluoroscent intensities of the two samples, revealing gene expression differences between normal and treated states





* Microarray Gene Expression Data Analysis - Image Data Extraction

→ The spots on a microarray are printed in a regular manner.

→ An image array will typically contain N x M blocks, where each block has p x q spots.

→ Spots must be individually segmented from the background to compute the expression ratio.

→ Some software packages for microarray image analysis are: GenePix, ScanAlyze, Dapple, Spot etc.

Challenges: geometric distortion, blurring, intensity saturation, poor contrast

Steps

① Image Preprocessing

→ The input microarray images consist of a pair of 16-bit images in TIFF format, Laser scanned from a microarray slide, using 2 different wavelengths (Red & Green)

→ X, the composite image is computed as follows:

$$X = \left[0.5 * \left(G + \left(\frac{median(G')}{median(R')} \right) R' \right) \right]$$

where $G' = \sqrt{G}$

$R' = \sqrt{R}$

[ ] = rounding to the nearest integer in the range [0, 255]

② Block Segmentation

→ The blocks in a microarray image are arranged in a very rigid pattern due to the printing process.

→ Each block in a microarray image is surrounded by region voids of any spots.

→ An effective way for block segmentation is through an analysis of the vertical and horizontal image projection profiles.

→ The projection profiles are obtained from an adaptively binarized image.

→ By performing analysis on the projection profiles, accurate block

segmentation can be achieved.

(3) Automatic Gridding

→ Gridding strategy consists of first locating the good quality spots (called guide spots), and then inferring the geometry of the grid from these spots

→ In order to account for the variable background and spot intensity, a novel adaptive thresholding procedure and morphological processing are used to detect the guide spots

→ After the guide spots are found, global rotation of the image is compensated for, and the correct grid parameters are estimated based on the spatial arrangement of the guide spots.

(4) Spot Extraction

→ Spot segmentation is performed in each of the subregions defined by the grid.

→ The segmentation involves finding a circle that separates the spot from the background.

→ Spot segmentation has 3 steps:

(i) Background Equalization for intensity variation in the sub-region

(ii) Statistical intensity modeling and optimum thresholding of the sub-region

(iii) Finding the best fit circle that segments the spot

→ When a spot is present, the intensity distribution of the pixels within the sub-region is modelled using a 2-class Gaussian Mixture Model

(5) Background Correction, Data Normalization & Filtering and missing value Estimation

→ Once the spots in a microarray image are extracted, the intensity value of each spot can be obtained and the log ratio

i.e $M = \log_2 R/G$ indicates the differential expression of the 2 DNA samples

→ Due to contamination and experimental errors, preprocessing of the raw intensity value is needed. These are:

(i) background correction - the spot's measured intensity may include a contribution not due to the specific hybridization

(ii) Normalization - adjust for biases that may arise from variation in the microarray process

within-slide normalization : $M_{norm} = \log_2\left(\frac{R}{G}\right) - c$

(iii) Filtering - keep only useful information from a microarray experiment

(iv) Missing Value Estimation - arises from artifacts on the microarray image, insufficient resolution, image corruption etc.

- The unreliable spots on the microarray image are usually manually flagged, and excluded from subsequent analysis - resulting in missing data at those locations.

Handling Missing Values - impute with zeroes

- replace with average value
- KNN (KNNImpute)
- SVD (SVDimpute)

$$A_{m \times n} = U_{m \times m} \times \Sigma_{m \times n} \times V^{T}_{n \times n}$$

- projection onto convex sets algorithm (POCs)

* Microarray Gene Expression Data Analysis for Patterns

→ Once expression data is obtained from the microarray images, the information embedded in the data needs to be discovered and analyzed.

→ Methods of analysis include

(i) Cluster Analysis — BHC clustering, SSMCL clustering

(ii) Temporal Expression Profile Analysis and Gene Regulation

(iii) Gene Regulatory Network Analysis

A. **Cluster Analysis**

→ aims at finding groups in a given dataset such that objects in the same group are similar to each other, while object in different groups are dissimilar

→ Algorithms used include: K means

Self Organizing maps (SOMs)

Hierarchical clustering

Self-Organizing Tree Algorithm

Principal Component Analysis

Multidimensional Scaling

→ Applications — Clustering is used in: (i) the study of temporal expression of yeast genes in sporulation

(ii) the identification of gene regulatory networks

(iii) study of cancer.

→ Both hierarchical and partition clustering have been used extensively in the gene expression data study.

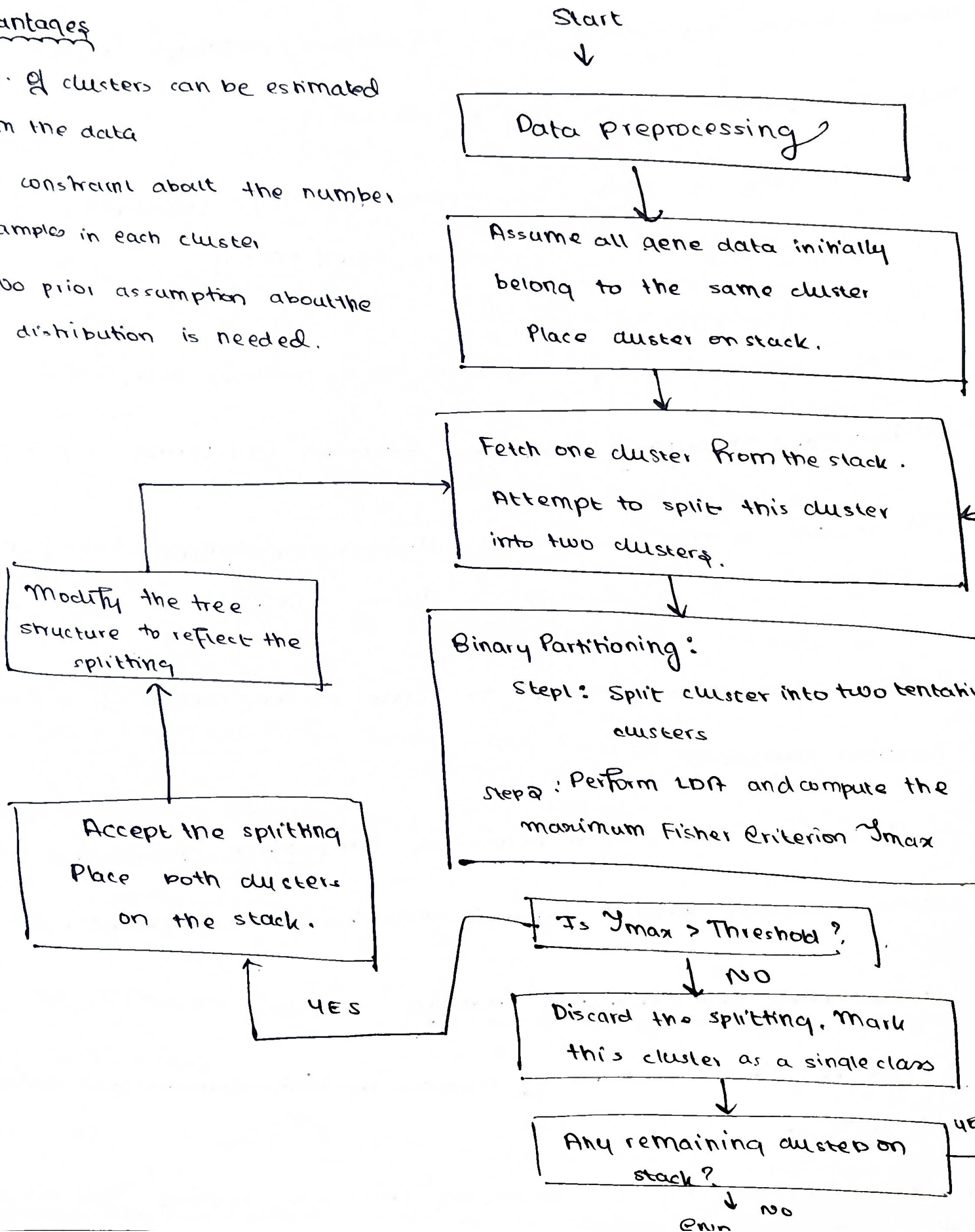
→ Binary Hierarchical Clustering (BHC) is a novel partitioning framework that combines the features of both categories of algorithms.

Avantages

① no. of clusters can be estimated from the data

② no constraint about the number of samples in each cluster

③ No prior assumption about the class distribution is needed.



SSMCL Clustering - Self-Splitting and Merging Competitive Learning Clustering (SSMCL)

→ In traditional clustering, few prototypes (cluster centers) than natural cluster can cause errors because a prototype might represent patterns from multiple clusters

→ This problem is called One-prototype-take-multiple-clusters (OPTMTC)

→ SSMCL solves this by introducing a framework called Self-Splitting and Merging Competitive learning Clustering.

→ It created a new paradigm called one-prototype-take-one-cluster (OPTOC)

→ OPTOC enables a prototype to settle at the center of a single natural cluster.

→ It achieves this by minimizing competition from other natural clusters through dynamic neighborhood adjustment.

Over Clustering and Merging Strategy

→ Estimating the correct no. of clusters in high-dimensional data is difficult

→ An over clustering and merging strategy is applied. This involves:

(i) Top-down clustering (Divisive Clustering) - Loose clusters are split until a larger than necessary number of clusters is created

(ii) Bottom-up clustering (Agglomerative Clustering) - similar clusters are merged back-to-back together systematically.

Steps: ① Loose clusters (high variance) are split

② Over-clustering ensures no natural clusters are missed.

③ Merging combines clusters that are close together - joint probability distribution is unimodal

Key Feature - a natural cluster should have a unimodal distribution

Outcome - By combining OPTOC and over-clustering and merging strategy, the correct no. of natural clusters is reliably estimated.

B.

Temporal Expression Profile Analysis and Gene Regulation

→ model dynamic biological processes as time series

↓

cell cycle, metabolic processes

→ can determine causal relationships between the expressions of diff. genes

→ use SVD and ICA → extract characteristics

↑ independent

→ use HMM / autoregressive model

c. Gene Regulatory Network Analysis

$G = (V, F)$ → model gene network as a directed graph

↙ ↘

elements — topology of edges between the nodes

→ The Boolean network models each gene as either being ON or OFF, and the state of each gene at the next time step is determined by a Boolean function of its input at the current time step.

→ can provide insight on the behaviour of gene interactions

can be used in the identification of drug targets for cancer therapy