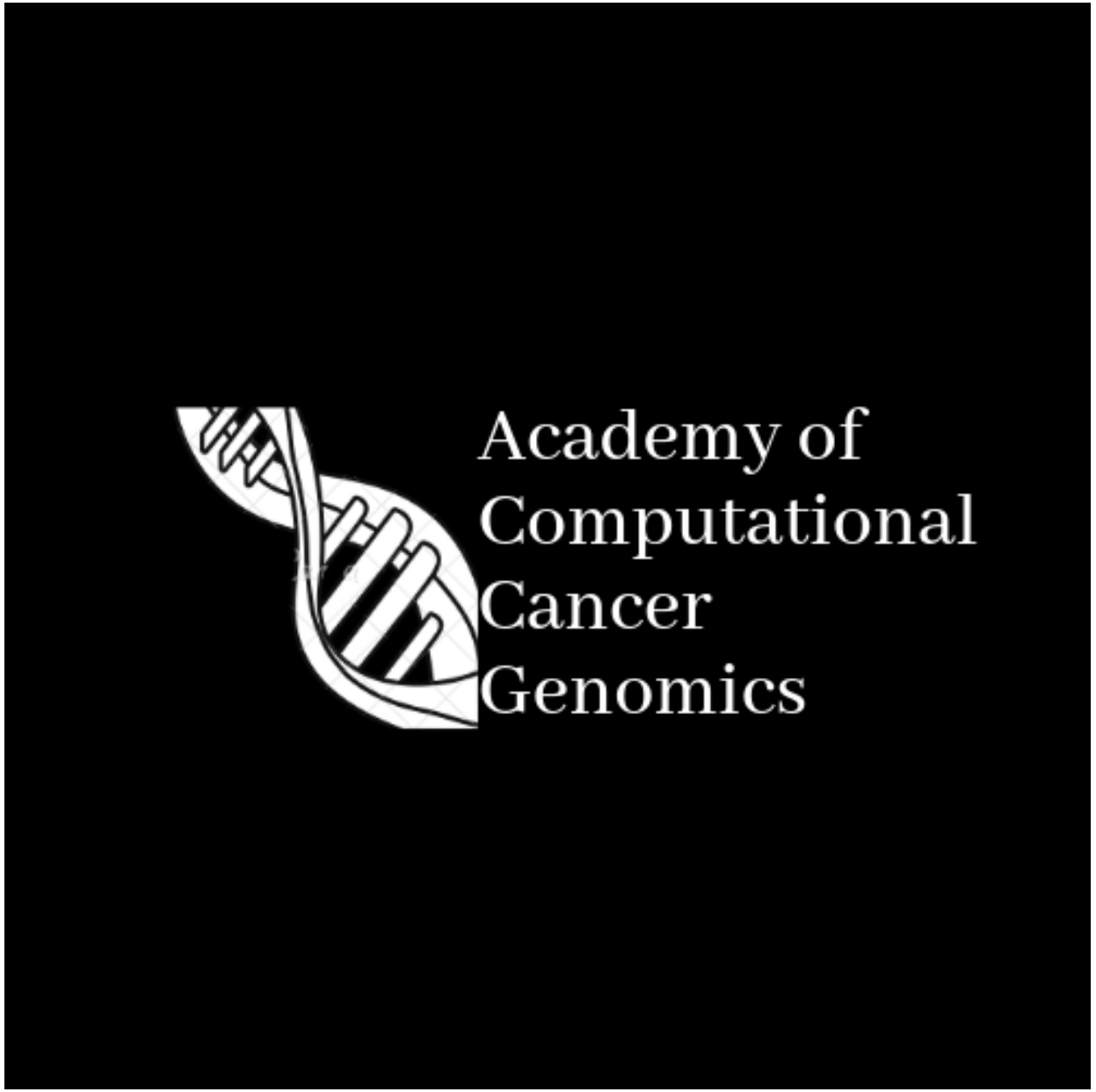
Computational Cancer Genomics



High-Throughput Technologies:

DNA Microarrays

Protein Mass Spectrometry

Comparative Genomic Hybridization

Chromatin Immunoprecipitation

Supervised Machine Learning

Filter out

- Filter out features and genes that show little variations across samples.

Transform

- Transform the data of each feature so that all have the same scale.

Select

- Select a distance or similarity measures.

Select

- Select Features to be used for the ML.

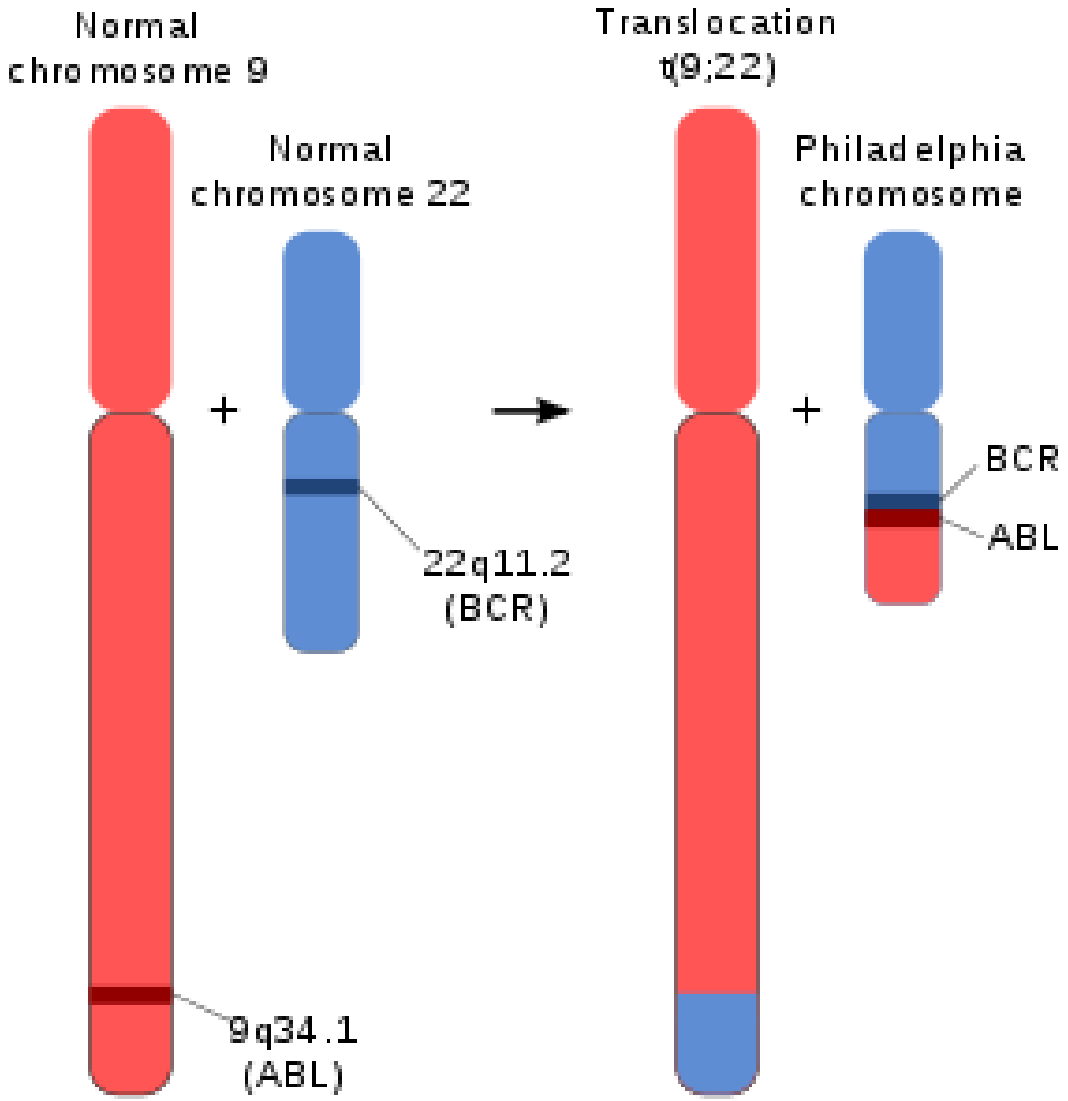
Choose

- Choose the algorithm

Assess

- Assess the performance of your analysis (e.g. Cross Validation)

Philadelphia Chromosome



Criteria for the Distance as a metric

Distance metric , Dissimilarities & similarities

Non-negativity $d(x, y) \geq 0$

Symmetry $d(x, y) = d(y, x)$

Identification mark $d(x, x) = 0$

Definitiveness $d(x, y) = 0 \rightarrow x = y$

Triangle inequality $d(x, y) + d(y, z) \geq d(x, z)$

Similarity Function

- Non-negativity $S(x, y) \geq 0$
- Symmetry $S(x, y) = S(y, x)$
- Monotonous Rise
  $S(x, y)$ *increses monotonously as* x *and* y *are more and more similar*

Distance from the higher perspective

Objects

Distribution

Distance Types

MINKOWSKI (GEOMETRIC)

CORRELATION

MAHALANOBIS

Gene Expression Data

Two values per features in in microarray data :

1. The estimated abundance of the mRNA for a particular genes

2. The standard error of the estimated abundance

The purpose of applying ML

- 1) Apply ML to samples:
- To identify patients with similar patterns of mRNA expression
- 2) Apply ML to features:
- To identify genes with similar patterns of expressions

Minkowski Distance Metrics

- Euclidean Distances

$$d(\,euclidean)(x,y) = \sqrt{\sum_{i=1}^{m}(xi\ - yi)^2}$$

- Manhattan Distance

$$d(man)(x,y) = \sum_{i=1}^{m}|xi\ - yi|$$

Correlation-Based Distance Metrics

- Pearson Sample Correlation Distance

$$d(x,y) = 1 - r(x,y) = 1 - \frac{\sum_{i=1}^{m}(xi - x^{\circ})(yi - y^{\circ})}{\sqrt[2]{\sum_{i=1}^{m}(xi - x^{\circ})^2 \sum_{i=1}^{m}(yi - y^{\circ})^2}}$$

$$x^{\circ} = \frac{1}{n}\sum_{i=1}^{m} xi$$

$$y^{\circ} = \frac{1}{n}\sum_{i=1}^{m} yi$$

Correlation-Based Distance Metrics

- Cosine Correlation Metrics

$$d(x,y) = 1 - \frac{|\sum_{i=1}^{m} xiyi|}{\sqrt[2]{\sum_{i=1}^{m} xi^2 \sum_{i=1}^{m} yi^2}}$$

Special case for Pearson correlation with the averages of x and y equals to zero.

Correlation-Based Distance Metrics

Spearman sample correlation distance,

$$d(x,y) = 1 - \frac{\sum_{i=1}^{m}(xi' - x^{\circ})(yi' - y^{\circ})}{\sqrt[2]{\sum_{i=1}^{m}(xi' - x^{\circ\prime})^2 \sum_{i=1}^{m}(yi' - y^{\circ})^2}}$$

$$xi' = rank(xi)$$

$$yi' = rank(yi)$$

Correlation-Based Distance Metrics

- Kendall`s τ sample correlation

$$d(x,y) = 1 - \ \tau(x,y) = 1 - \frac{\sum_{i=1}^{m}\sum_{j=1}^{m} C(xij)C(yij)}{m(m-1)}$$

$$C(xij) = sign(xi - xj)$$

$$C(yij) = sign(yi - yj)$$

Correlation-Based Distance Metrics

- Properties :

1) Invariant to location and scale transformation
2) Adversely affected by outliers
3) Kendal (TAU) and SPEARMAN is preferred in case of the outliers

Mahalonobis Distance Metric

1)The distance *between a point* ,***P***, and a *distribution* ,***D***

2) How many SD , P is away from the mean of the D

3) Widely used in cluster analysis and classification

4)In the Linear Discriminant Analysis (LDA)

5) Unitless and scale invariant

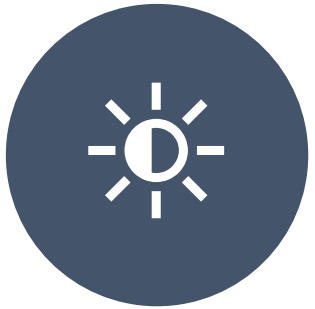
$$The\ Mahalanobis\ distance\ of\ the\ observation\ vector\ (x1, x2, \dots, xn)^T\ from\ the\ set\ of\ observation\ with\ mean\ vector = (\mu 1, \mu 2, \dots, \mu n)\ with\ covariance\ matrix\ S$$

$$d(D,P) = \sqrt[2]{(x-\mu)^T\ S^{-1}(x-\mu)}$$

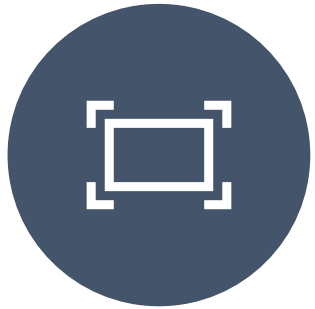
If the S is a Identity matrix , the d(D,P) reduces to Euclidean Distance

If the S is a diagonal matrix , the d(D,P) reduces to standardized Euclidean Distance.

Scaling Importance for microarray data



1) THE ORIGINAL SCANNED FLUORESCENCE INTENSITIES



2) LOGARITHMICALLY TRANSFORMED SCALE

3) VARIANCE STABILIZED SCALE

Distance Metrics Computation

Package	Function	Criteria	Class	Version
bioDist	Spearman.dist	Correlation Based	dist	BioConductor
bioDist	Tau.dist	Correlation Based	dist	BioConductor
Stats	dist	Euclidean & Manhattan	dist	R
cluster	daisy	Euclidean & Manhattan	dist	R

Distance from a distribution

- Is the shape of the distribution of features is similar between two genes ?
- Expression genes follow a multivariate normal distribution with diagonal variance covariance matrix
- Used when both expression levels and their associated standard errors are available.

Kullback_Leibler Distribution

Kullback_Leibler Information

$$KLI(f,g) = \int \log[\frac{f(x)}{g(x)}]\mathrm{f(x)}\ \mathrm{dx}$$

Hamming Mutual Information

$$MI(f,g) = \int \log[\frac{f(x,y)}{f(x)g(y)}]\ \mathrm{dx}\ \mathrm{dy}$$

Distance from a distribution

- Kullback_Leibler Information

How much the shape of one distribution resembles to the other?

- Hamming Mutual Information

A multivariate measure of the association

Distance Specification

1) Distance from the unstructured samples

2) Distance from the highly structured samples

Microarray Distance Metrics

- Factorial Experiments :

- Time_Course Experiment & Sobolev Space and metrics

Time-based Experiments

- Sobolev Metrics & Space :
- Standard wavelet decomposition
- Decompose expression profile
- Interpretable quantities
- Local frequency components
- New distances computed
- New Gene Clustering

Distance & Scaling of the features

- Standardization of features :

Impacts the distance between features

Makes the features comparable

Warning : Scaling can remove some potentially interesting features

Methods for Standardization of gene expression measures

$$x(i)^T = \frac{x(i) - center[x(i)]}{scale[x(i)]}$$

$$scale[x(i)]: Standard\ Deviation\ (\sigma), Interquartile\ range\ (IQR),\ Median\ Absolute\ Deviation\ about\ Median\ (\ MAD)$$

$$center[x(i)]: Center\ of\ Distribution\ , Mean\ , Median$$

cDNA microarray data standardization vs. Affymetrix data standardization

Invariance of gene distances or sample distances for relative and absolute expression measurements

Linear Discriminant Analysis

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}.$$

- Classify a new point according to which density is highest

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right) \qquad p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}$$

Discriminant Score

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

KNN
Classification