Machine Learning For Genomics

Final Presentation

Indrik Wijaya

National University of Singapore
Supervisors: Assoc Prof Vincent Tan and Dr Huili Guo

Apr 6, 2018

Outline

- Machine Learning and Genomics
- Objective
- Description of Data
- Different Clustering Methods
- Results
- Conclusion
- Possible Future Work

Machine Learning and Genomics

Machine Learning

- Self-learning algorithm to gain knowledge from data
- Unsupervised, supervised, semi-supervised

Genomics

- DNA→transcription→RNA →translation→proteins
- mitochondria

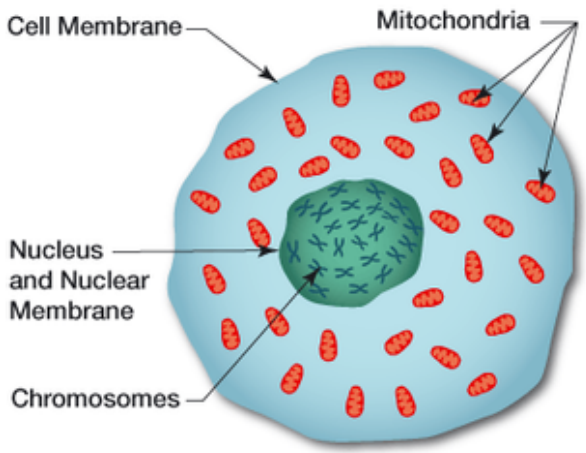


Figure: Cell with its organelles, retrieved from http://www.eastfhs.org/3-mitochondrial-dna-and-the-ancient-organisms-in-your-cells.html

Objective

Translational Coordination

- Mitochondrial biogenesis
- OXPHOS

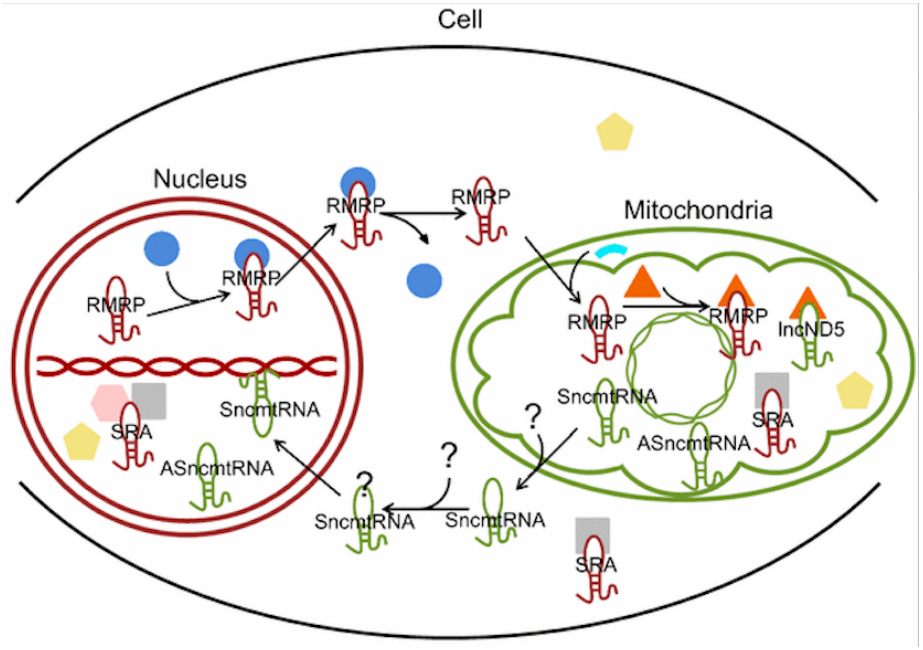


Figure: Coordination between nucleus and mitochondria, retrieved from http://www.lncrnablog.com/long-noncoding-rnas-coordinate-functions-between-mitochondria-and-the-nucleus/

Gene	t0	t1	t2	t3	t4	t5	t6
A	x_{A0}	x_{A1}	x_{A2}	x_{A3}	x_{A4}	x_{A5}	x_{A6}
B	x_{B0}	x_{B1}	x_{B2}	x_{B3}	x_{B4}	x_{B5}	x_{B6}

Table: Gene expression values at different time points

Clustering

- Subdivide a set of items
- Similarity measures: Euclidean distance, correlation coefficient
- Genes with similar biological properties will fall under the same cluster in terms of expression values
- Quality of clustering

Description of Data

Data

- Datasets: Bulk and Crude
- Unlabelled gene expression values $\rightarrow$ Unsupervised
- Short Time-series, 7 time points
- $N \approx 10k$ samples
- Special subset: Mitochondria genes ($\approx 1.1k$ samples)
- Transformed to log ratio with respect to first time point

Short Time-series

- Data not rich enough
- Time dependencies, sequential nature

Different Clustering Methods

- Short Time-series Expression Miner (STEM)
- Gaussian Mixture Model
- K-means
- Hierarchical Clustering

STEM

Steps

- Obtain model profiles (define m, c)
- Assign genes to a predefined set of model profiles
- *Dist metric: *Correlation Coefficient* $= 1 - \rho(x, y)$
- Determine significance of each of these profiles (compute enrichment of genes) using *Permutation Test*
- Grouping significant profiles (based on δ)

No of Genes In Significant Profiles, $m = 100, c = 3$

	Bulk	Bulk Mito	Crude	Crude Mito
Total	11078	1006	8675	911
STEM	7867	656	4894	420
STEM (%)	71.01	65.21	56.41	46.10

Table: Comparing no of genes for different datasets

- STEM excludes many genes (or remove noises)
- Decided to use k from STEM as initial no of clusters for other algorithms

STEM: Outputs for Crude Mito

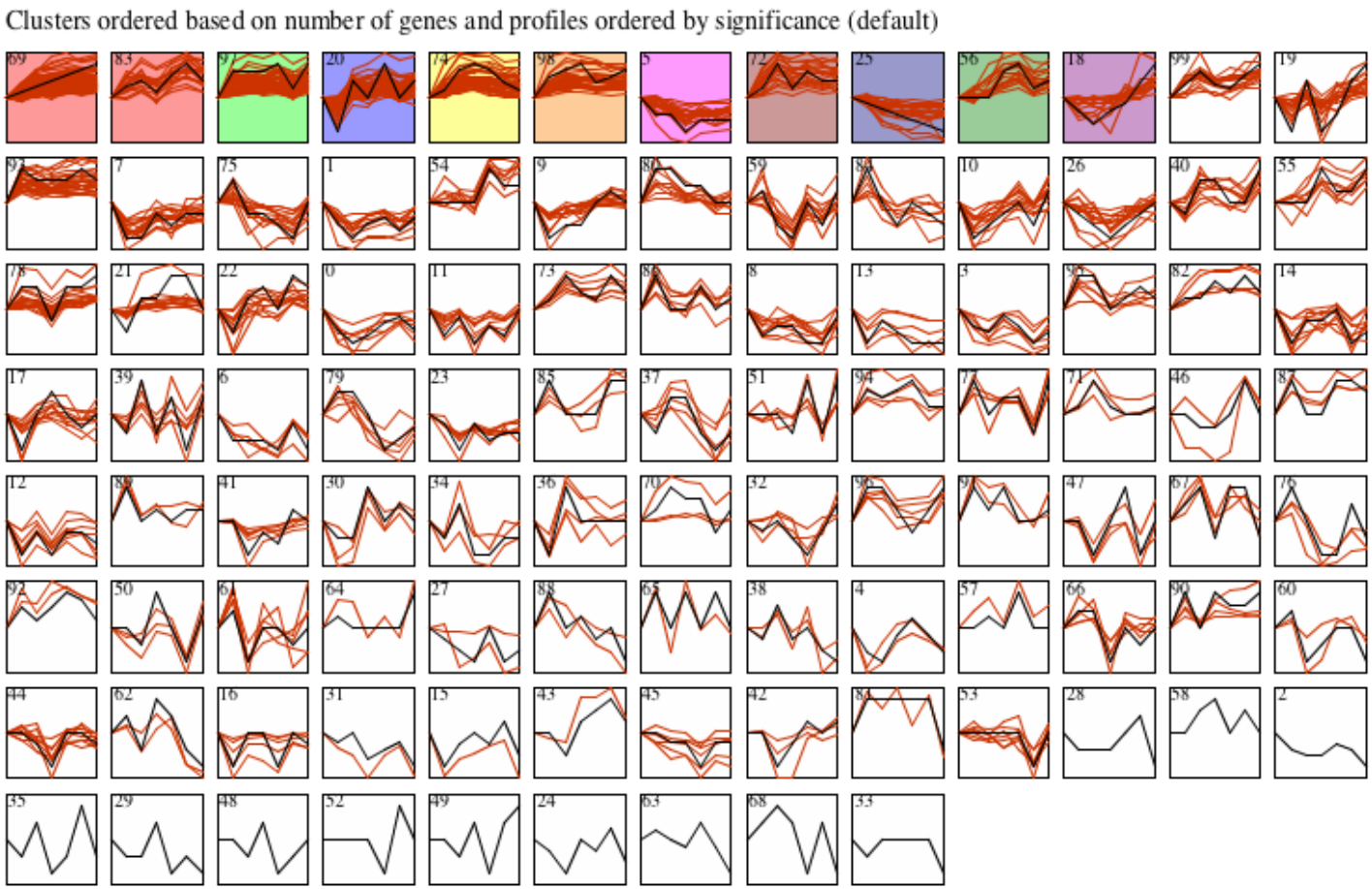
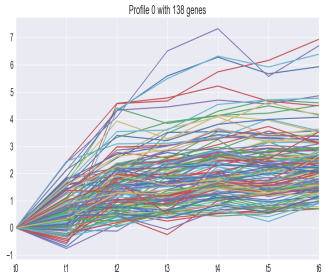


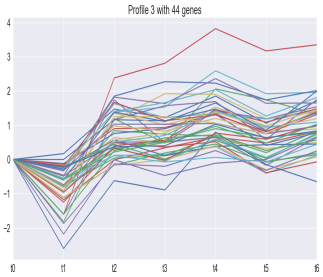
Figure: Clustering Results from STEM, $m = 100, c = 3, \delta = 0.3$

STEM: Outputs for Crude Mito

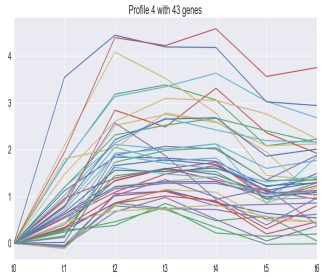


138

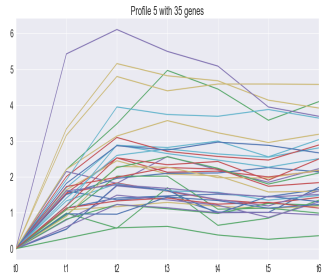
48



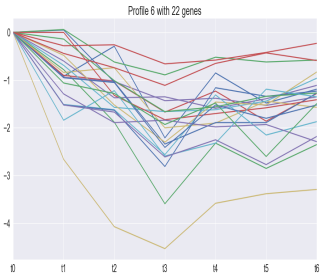
44



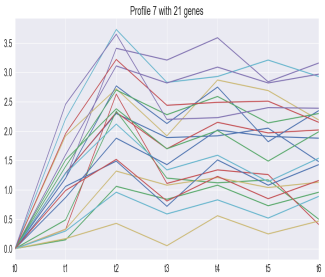
43



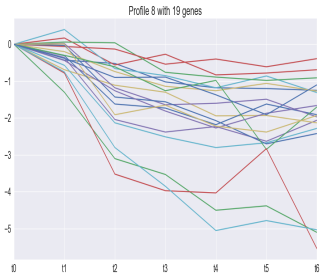
35



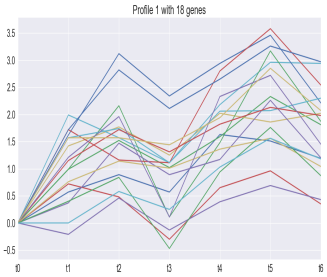
22



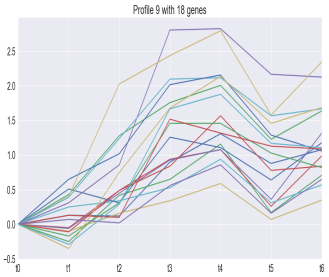
21



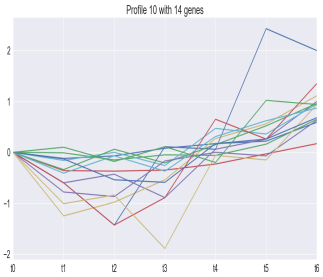
19



18



18



14

Mixture Model

Outline

$D = \{X_i\}_{1 \leq i \leq N}, X_i = \{x_{i1}, \cdots, x_{id}\}$
$x_{ij} = \alpha_k(t_j) + \beta_i + \epsilon_{ij}, \beta_i \sim N(0, \theta_k), \epsilon_{ij} \sim N(0, \theta).$

$$\Sigma_k = \theta_k I_T + \theta J_T = \begin{pmatrix} \theta_k + \theta & \theta & \cdots & \theta \\ \theta & \theta_k + \theta & \cdots & \theta \\ \cdots & \cdots & \cdots & \cdots \\ \theta & \theta & \cdots & \theta_k + \theta \end{pmatrix}$$

Clustering problem becomes a mixture model:

$$X_i \sim \sum_{k=1}^{K} \pi_k N(\alpha_k, \Sigma_k)$$

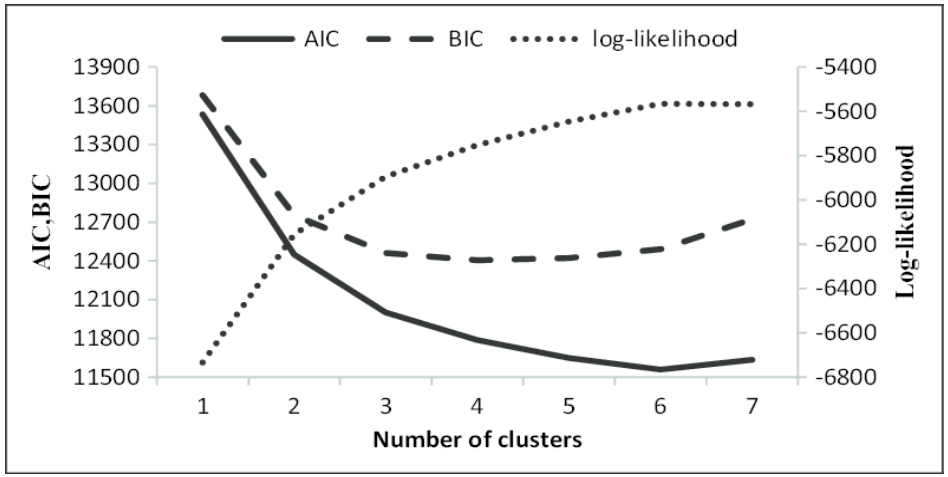
π_k are the mixture coefficients of the mixture model

Mixture Model: Model Selection (AIC & BIC)

- Estimate optimum parameters $\xi = \{\pi_k, \alpha_k, \Sigma_k\}$ using EM algorithm that maximizes

$$l(D) = logP(D|\xi) = log \prod_{i=1}^{N} P(X_i|\xi) = \sum_{i=1}^{N} log \sum_{k=1}^{K} \pi_k N(X_i|\alpha_k, \Sigma_k)$$

- Optimum k: $BIC = -2l(D) + log(N)Y; AIC = -2l(D) + 2Y$
$Y = Kd + K - 1 + P$: no of free parameters
$d = 7$ and P depending on the covariance type

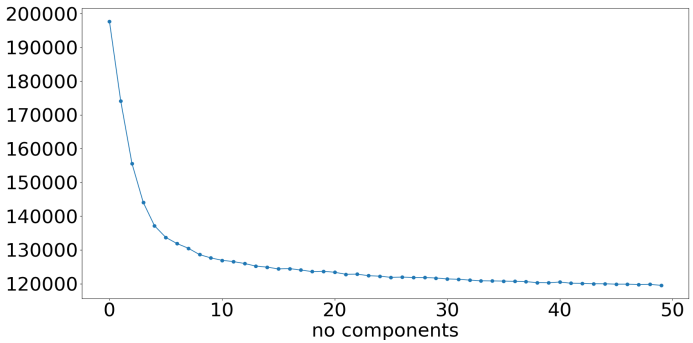


Mixture Model: Model Selection

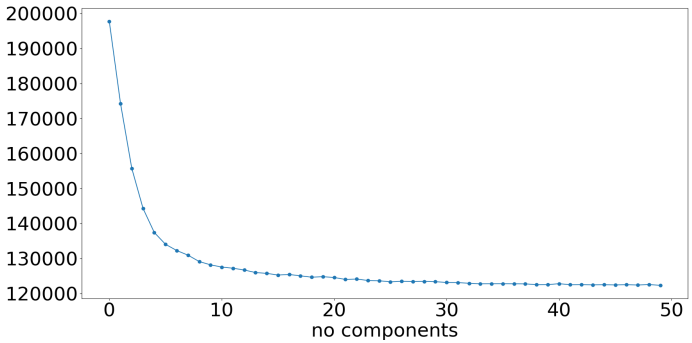
Different covariance matrices:

- full: each component has its own general covariance matrix ($P = kd(d+1)/2$)
- tied: all components share the same general covariance matrix ($P = d(d+1)/2$)
- diag: each component has its own diagonal covariance matrix ($P = kd$)
- **spherical**: each component has its own single variance ($P = k$)

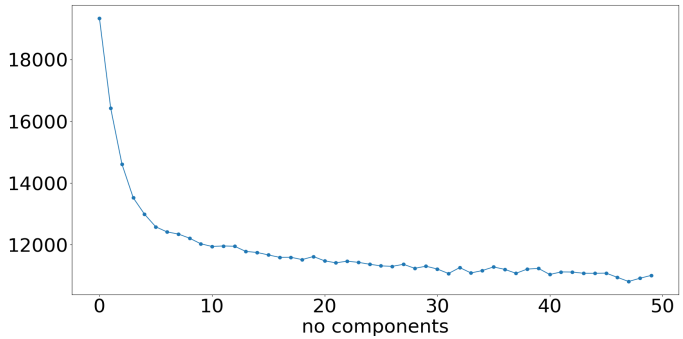
Covariance Type: Spherical



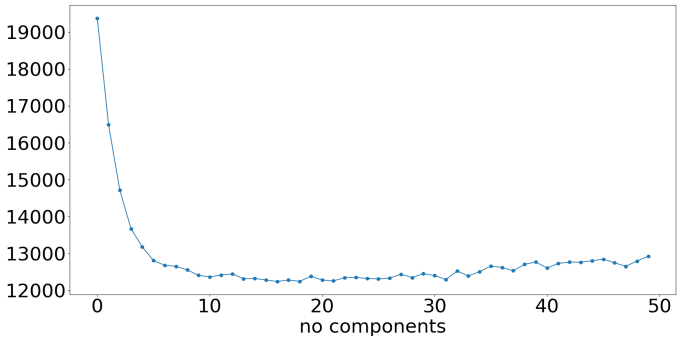
AIC (Crude)



BIC (Crude)



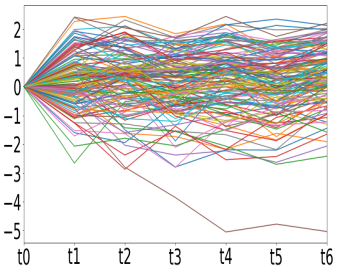
AIC (Crude Mito)



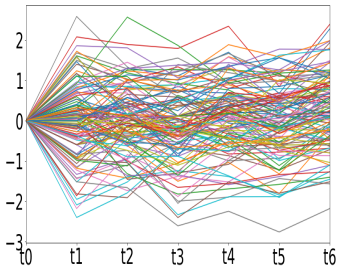
BIC (Crude Mito)

- Similar plots for bulk and bulk mito
- The data cannot be modelled as a mixture of Gaussians

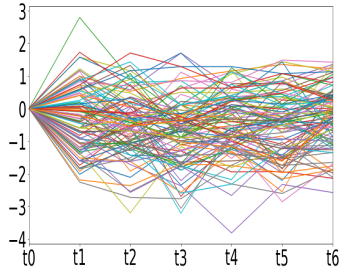
Mixture Model (Spherical): Outputs for Crude Mito, 11 components



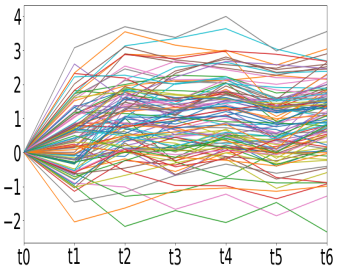
140



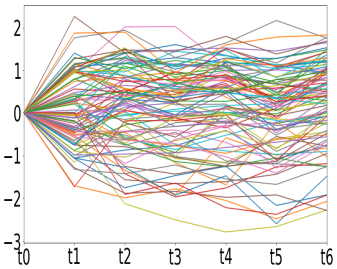
122



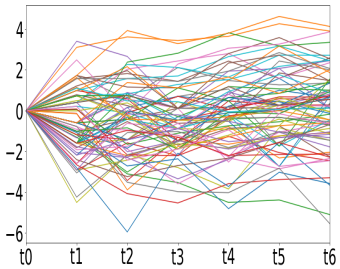
102



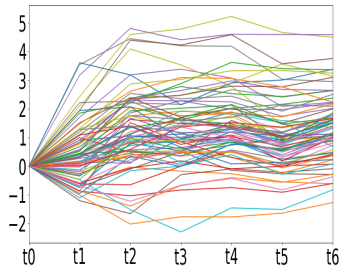
102



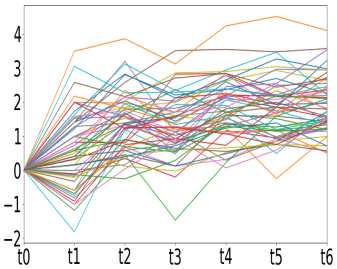
99



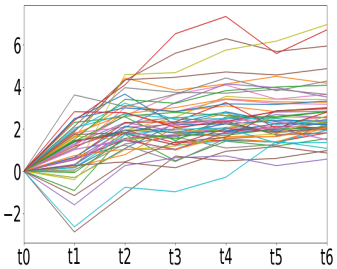
78



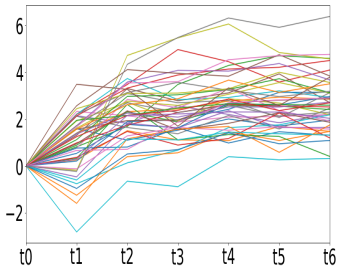
77



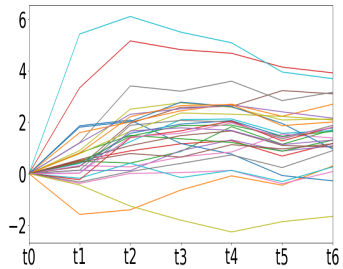
57



56



48



30

K-means

Outline

- Given $D = \{x_t\}_{t=1}^{n}$ and fix a number of clusters $1 \leq k \leq n$,

minimize the cost function $J = \sum_{j=1}^{k} \sum_{x \in D_j} \|x - \mu_j\|^2$

- K-means++ for faster convergence and more consistent results
- Model Selection: Elbow curve and gap statistic analysis

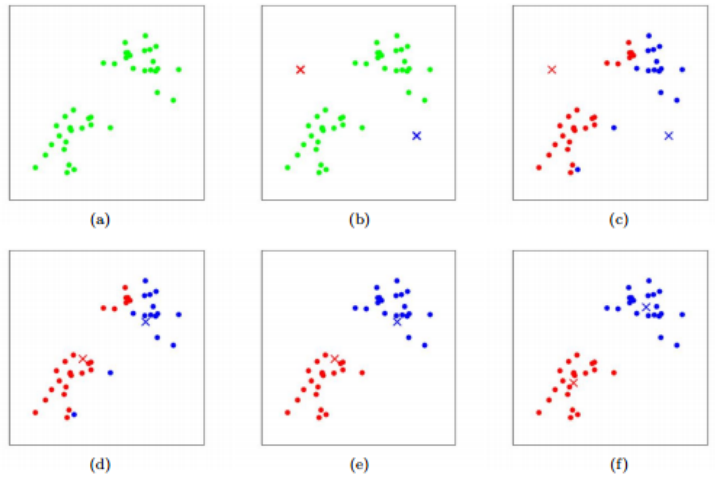
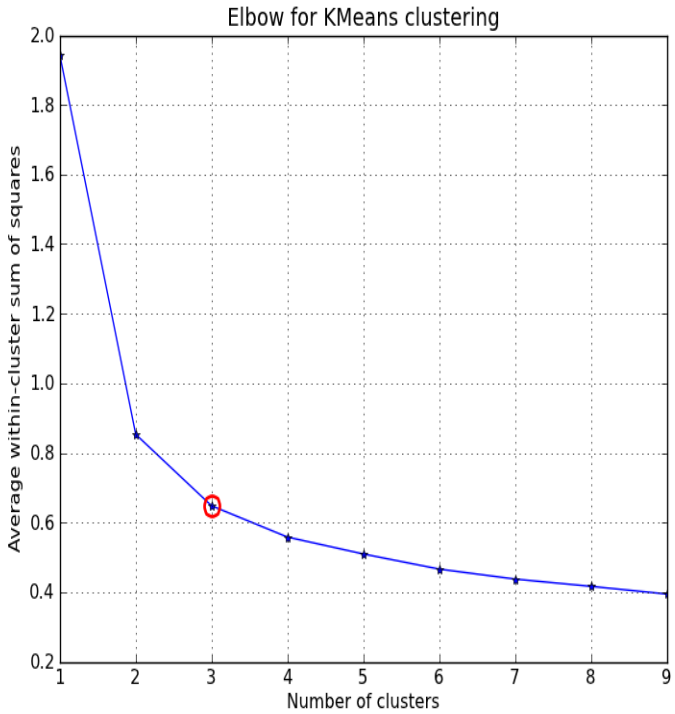
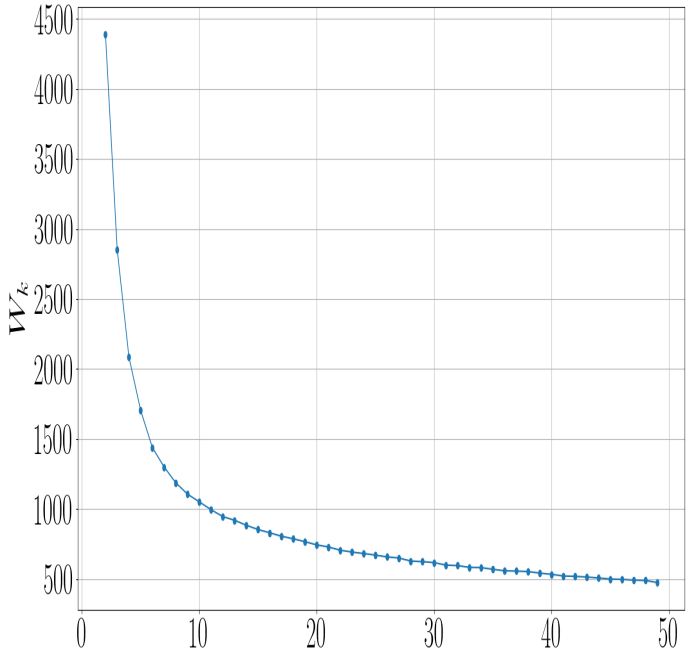


Figure: Iterative steps for K-means algorithm

K-means: Model Selection (Elbow Curve Output for Crude Mito)



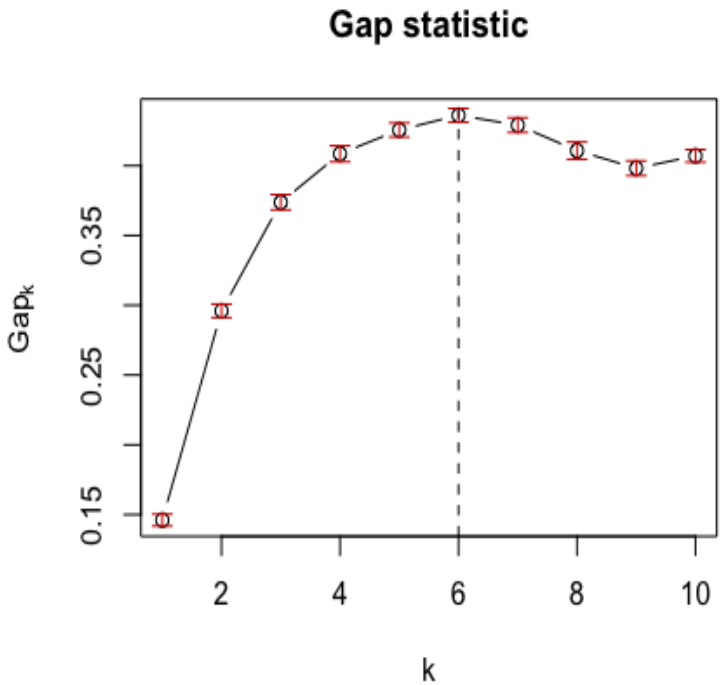
(a) Ideal Elbow Curve



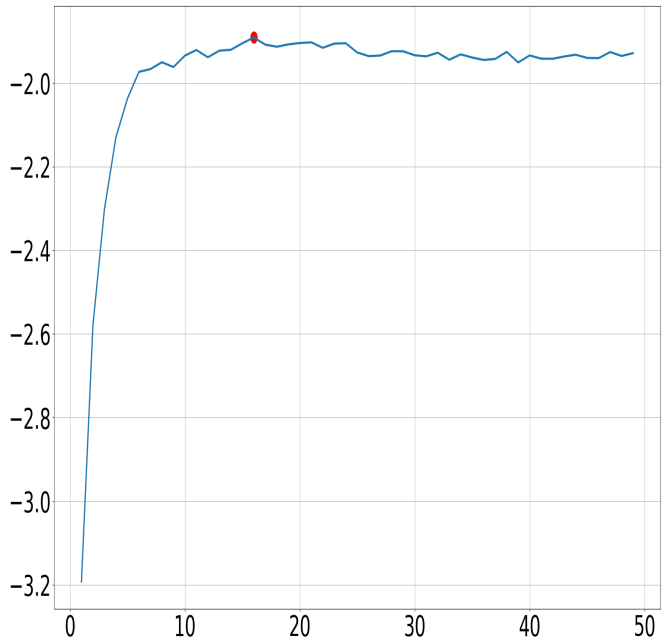
(b) Experimental Elbow Curve

Figure: Ideal Elbow Curve vs Experimental Elbow Curve

K-means: Model Selection (Gap Statistics Output for Crude Mito)



(a) Ideal Gap Stats



(b) Experimental Gap Stats

Figure: Comparing Ideal gap stats curve vs experimental gap stats curve

Hierarchical Clustering

Outline

Stepwise Algorithm which merges/splits 2 objects at each step with the least dissimilarity.
Two types: Agglomerative & Divisive

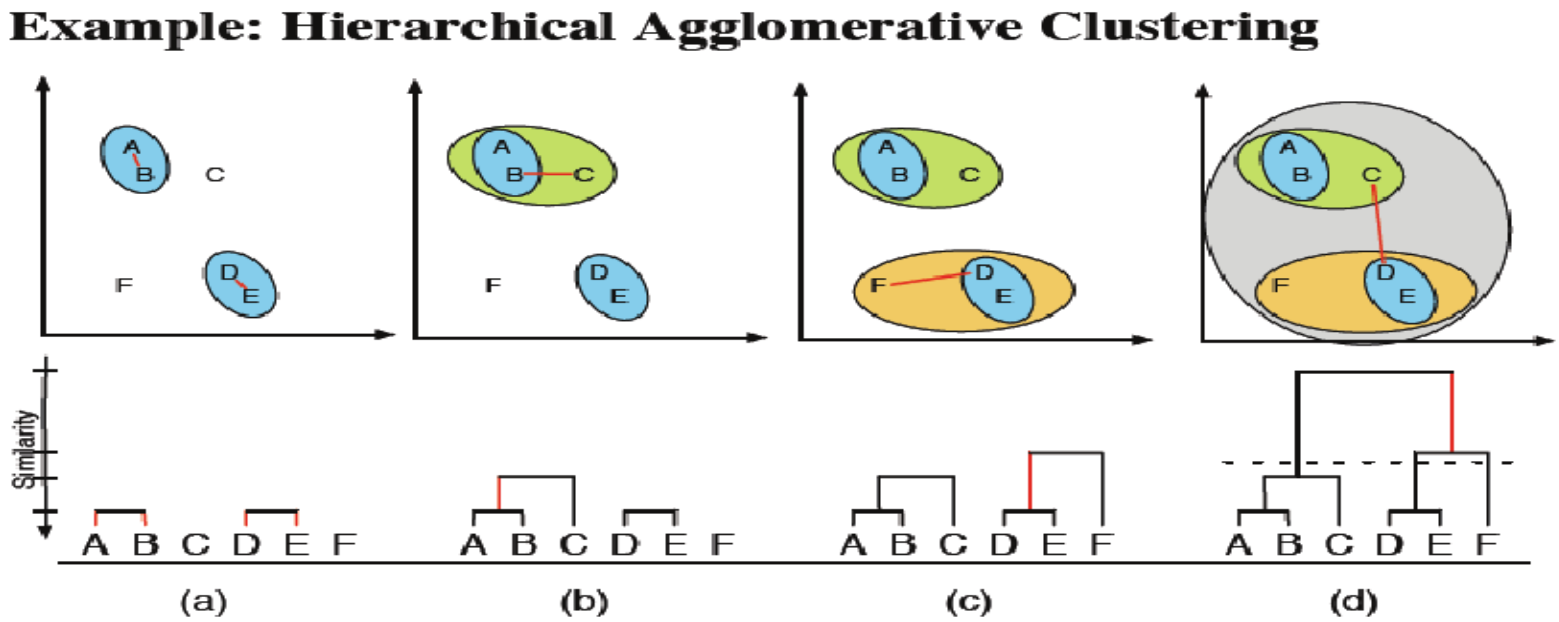


Figure: Iterative Steps of Hierarchical Agglomerative Clustering

Hierarchical Clustering: Linkage

Dissimilarity Measures

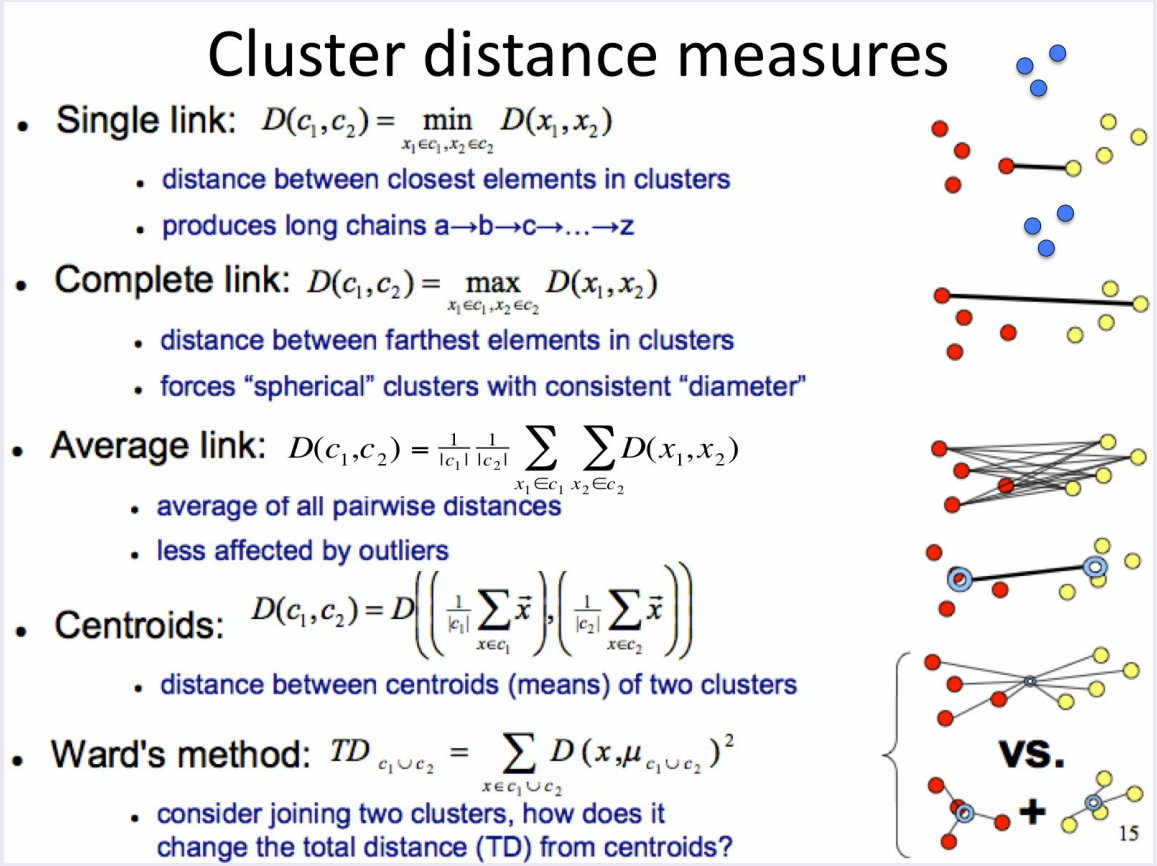


Figure: Different Dissimilarity Measures

Clustering Evaluation

$\{t_k\}$: true classes, $\{c_k\}$: clusters found by a cluster algorithm

Clustering Accuracy

- Label all the data in cluster c_j as t_i if they share the most data objects.
- Note: #clusters need not be $=$ #classes
- Clustering Accuracy, $\eta = \frac{\sum_x I(c_i(x)=t_j(x))}{N}$
  $I(.)$: indicator function
  $c_i(x)$: label of the cluster which x belongs to
  $t_j(x)$: true class of x
  N: size of dataset

Misclassification Error Distance

- ME dist, $d_{ME}(t_k, c_k) = 1 - \frac{1}{N} \max_{\pi \in P_K} \sum_{k=1}^{K} |t_k \cap c_{\pi(k)}|$
- P_K: set of all permutations of $[K]$

	RNA Bulk		RNA Bulk (Mito)		RNA Crude		RNA Crude (Mito)	
Algorithm	**CA**	**ME**	**CA**	**ME**	**CA**	**ME**	**CA**	**ME**
K-means	56.37	68.17	66.77	59.91	48.39	72.60	45.95	68.81
GMM (Full)	50.87	71.41	59.15	58.84	51.70	71.00	53.81	64.76
GMM (Diag)	52.79	72.31	68.90	59.15	44.34	76.24	42.86	71.67
GMM (Sph)	57.23	70.55	66.92	60.52	53.11	72.00	48.10	67.86
HAC (Ward)	53.17	70.75	63.41	60.52	47.10	73.03	43.10	71.43
HAC (Comp)	71.68	49.61	78.51	31.55	67.94	49.94	71.43	44.76
HAC (Ave)	66.40	38.88	73.48	33.84	64.73	44.79	61.43	44.76

Table: Performance evaluation for the different algorithms (in %), the underlined values are the best performing algorithm for that particular dataset and metric

Comparison of Algorithms (in terms of no of genes)

Cluster	K-means	GMM (Sph)	GMM (Diag)	HAC (Comp)
1	134	140	121	357
2	126	122	110	98
3	111	102	107	84
4	99	102	100	76
5	82	99	100	76
6	82	78	89	69
7	77	77	88	41
8	69	57	87	40
9	55	56	56	32
10	54	48	35	28
11	20	30	18	10

Table: Gene count in each cluster for the different algorithms ($k = 11$), total number of genes $= 911$

Clustering Evaluation

Gene Ontology(GO)

- Perform enrichment analysis on gene sets. e.g. given a set of genes that are up-regulated under certain conditions, an enrichment analysis will find which GO terms are over (or under)-represented using **annotations for that gene set**.

Types

- cellular component (cc)
- molecular function (mf)
- biological process (bp)
- KEGG pathway

GO for Mixture Model (Spherical)

- Cluster 3 (102 genes)
  - cc: (**organelle inner membrane, mitochondrial inner membrane**, mitochondrial membrane part)
  - KEGG: (**OXPHOS, carbon metabolism**)
- Cluster 6 (77 genes)
  - cc: (**mitochondrial inner membrane, organelle inner membrane**, mitochondrial protein complex)
  - KEGG: (**OXPHOS, Parkinson's disease, Alzheimer's diseases**)

GO for K-means

- Cluster 3 (110 genes)
  - cc: **mitochondrial inner membrane, organelle inner membrane**
  - KEGG: **carbon metabolism**, TCA cycle, **OXPHOS**
- Cluster 5 (81 genes)
  - bp: mitochondrial transport, mitochondrial gene expression
  - cc: **mitochondrial inner membrane**, **mitochondrial matrix**
  - KEGG: ribosome, **OXPHOS**, **Parkinson's disease**

GO for HAC-comp

- Cluster 1 (357 genes)
  - cc: **mitochondrial inner membrane, organelle inner membrane**
  - KEGG: **OXPHOS, thermogenesis, carbon metabolism**
- Cluster 2 (98 genes)
  - bp: purine ribonucleoside triphosphate metabolic process, ribonucleoside triphosphate metabolic process
  - cc: **organelle inner membrane, mitochondrial inner membrane**, mitochondrial matrix
  - KEGG: **Parkinson's disease, OXPHOS**

GO Summary

- Some enriched GO terms appear in multiple clusters
  - share mitochondrial-related GO terms
  - possibility of merging
  - interaction between clusters
- Common GO terms:
  - cc: mitochondrial inner membrane, organelle inner membrane, mitochondrial matrix
  - KEGG: OXPHOS, thermogenesis, Alzheimer's disease, Parkinson's disease
- There is no common GO terms for *molecular function* and *biological process*

Conclusion

STEM As Benchmark

- Gives optimal no of clusters
- Takes into account the sequential nature of time-series
- Generates profiles independent of data
- Allows for better study of dynamics for each cluster

Other Findings

- STEM may exclude many genes (or remove noises)
- Algorithms with correlation coefficient perform better (same similarity measure)
- Use STEM to initialize number of clusters
- Euclidean distance is not a good measure

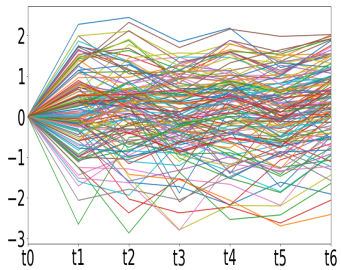
Possible Future Work

- Different algorithms: Fuzzy-clustering, Bayesian approach mixture model
- Interaction between clusters: Look deeper into results from hierarchical clustering, *fast optimal leaf ordering for hierarchical clustering*
- Other distance measures: dynamic time warping, short time-series Fuzzy-clustering distance, Manhattan distance
- Clustering Evaluation for Mixture model: normality test for each cluster

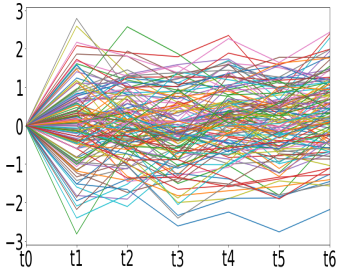
Thank you! Q&A

EXTRA MATERIAL

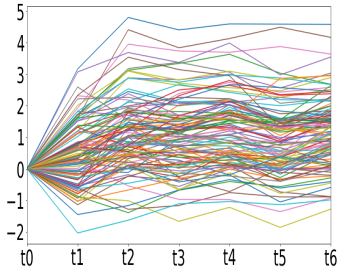
K-means: Outputs for Crude Mito, 11 clusters



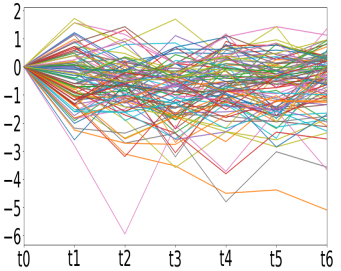
134



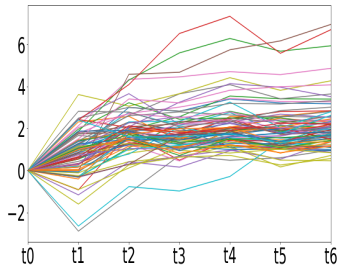
126



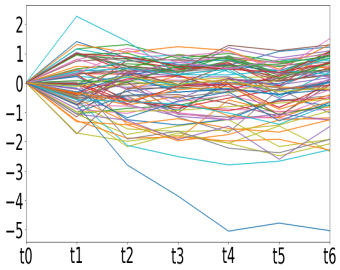
111



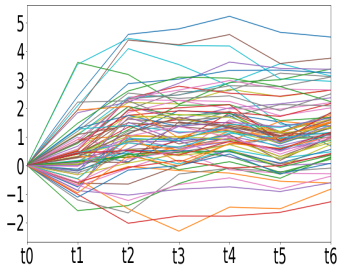
99



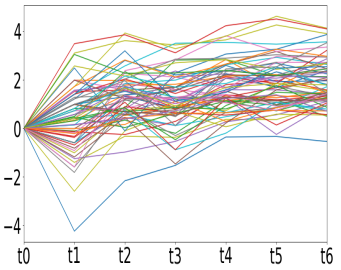
82



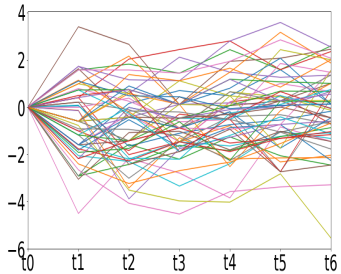
82



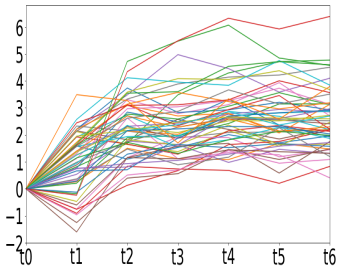
76



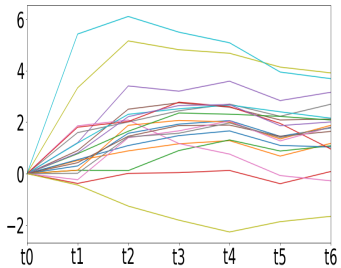
69



76

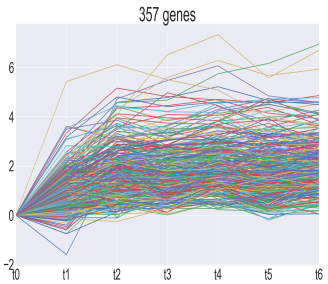


54

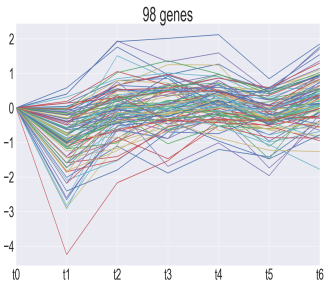


20

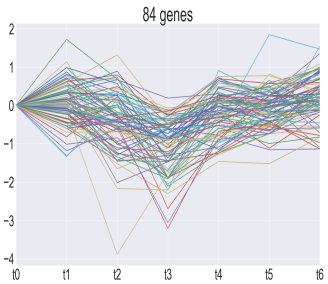
HAC-comp: Outputs for Crude Mito, 11 clusters



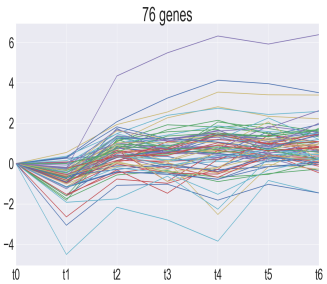
357



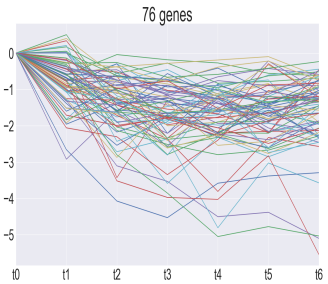
98



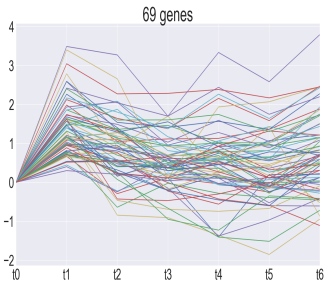
84



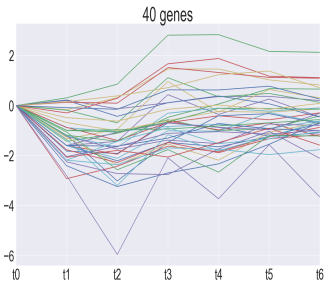
76



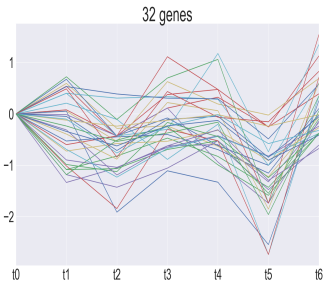
76



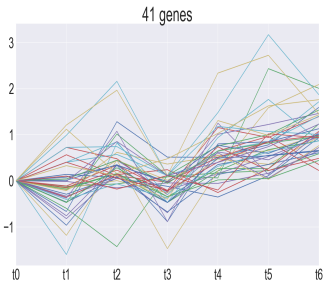
69



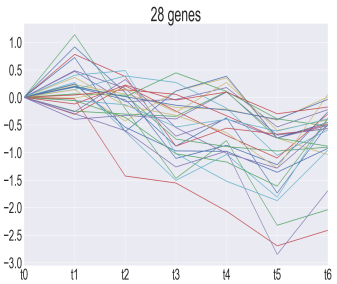
40



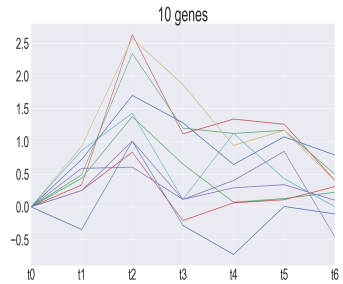
32



41

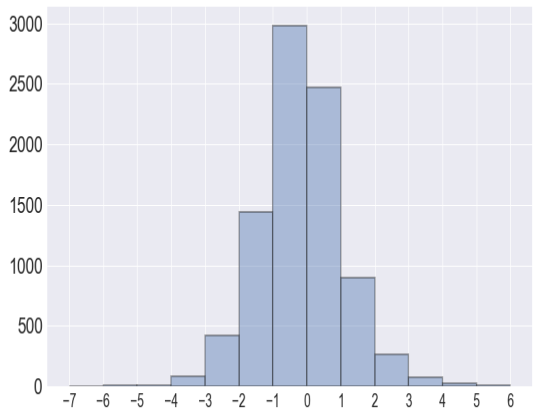


28

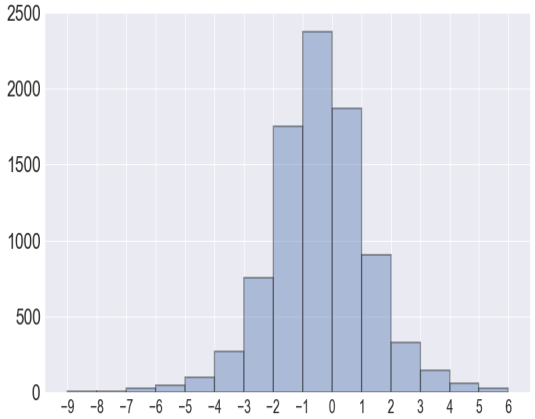


10

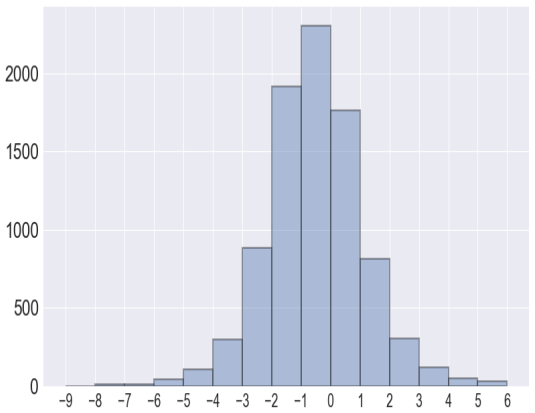
Crude



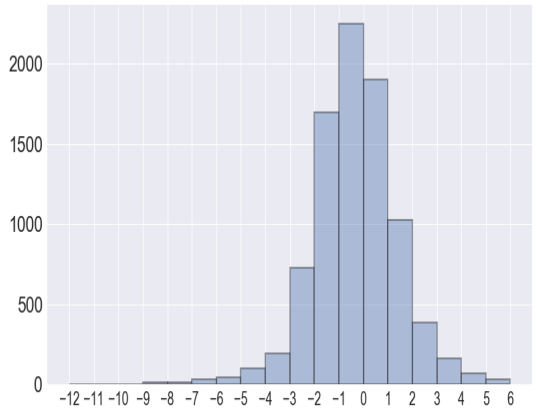
t_1



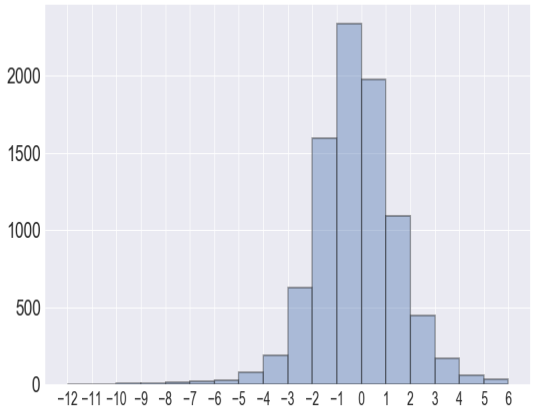
t_2



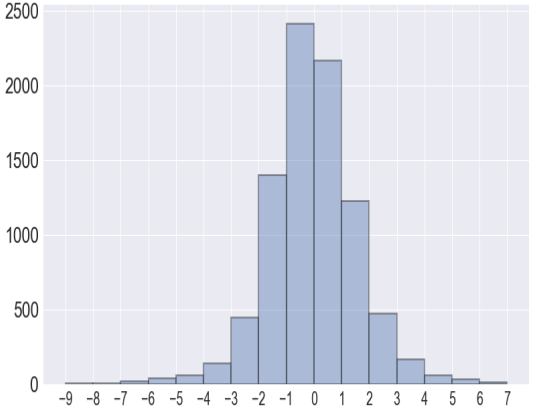
t_3



t_4

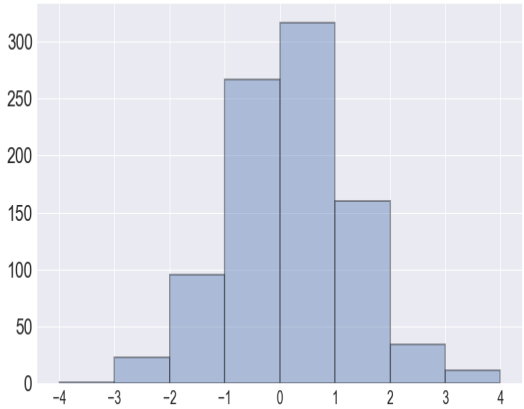


t_5

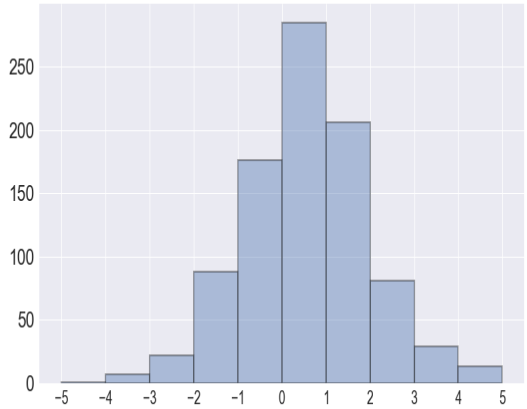


t_6

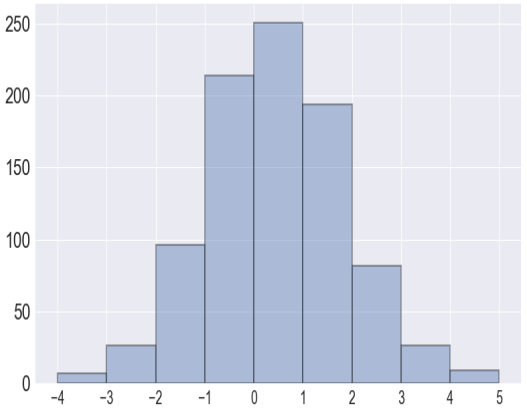
Crude(Mito)



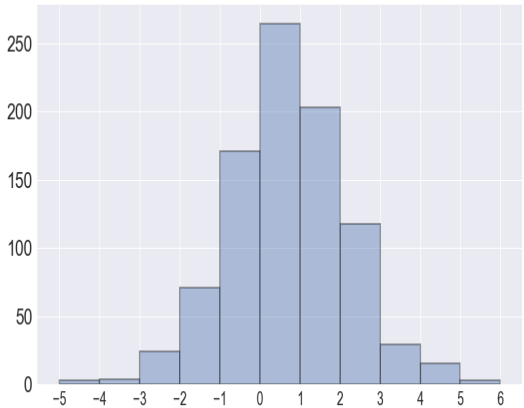
t_1



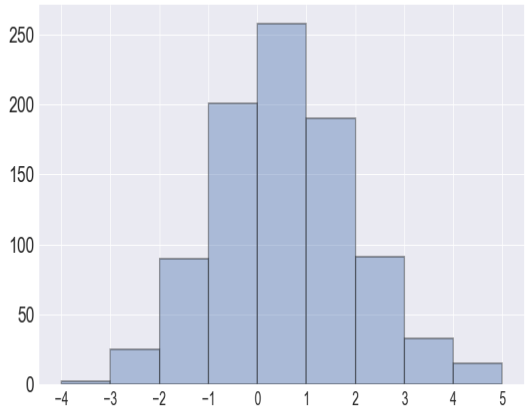
t_2



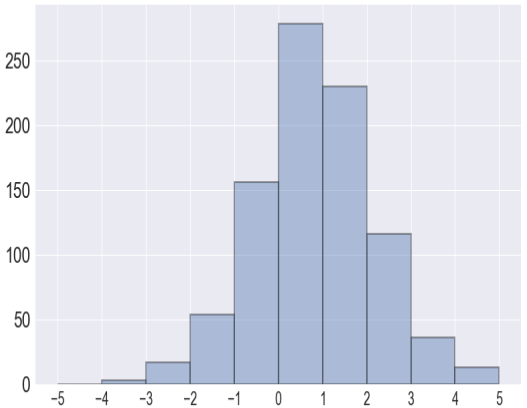
t_3



t_4

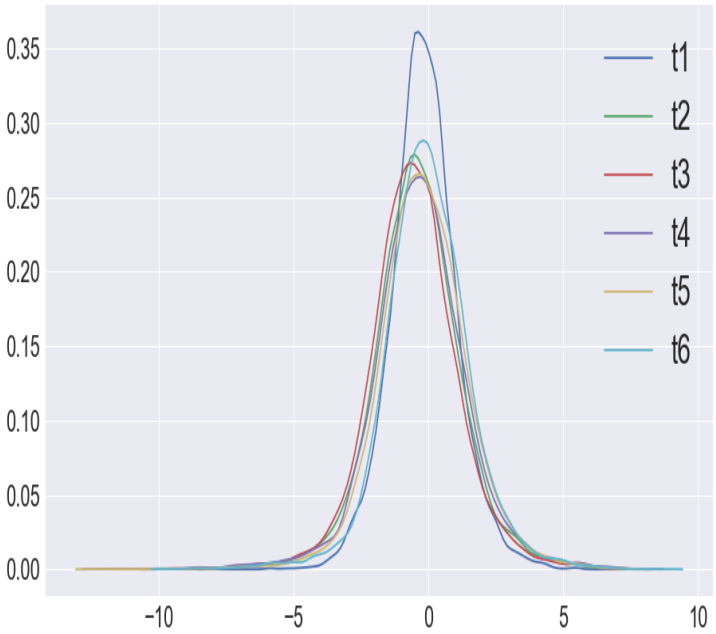


t_5

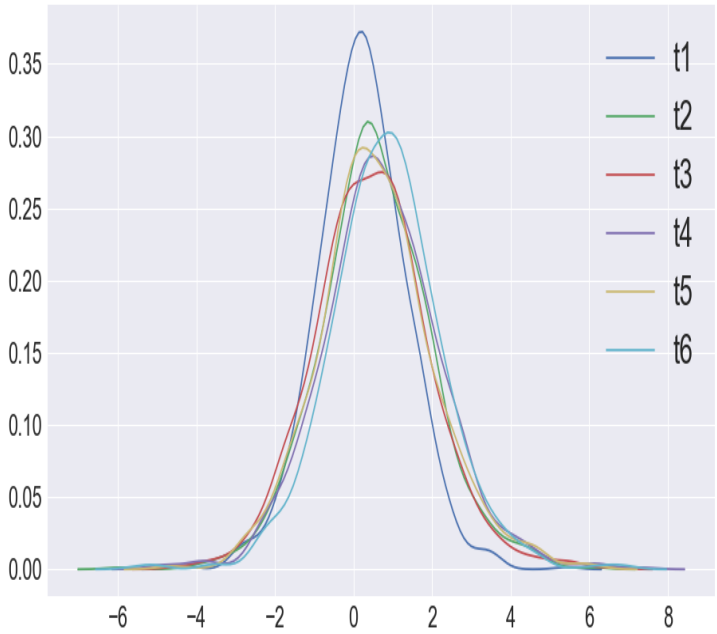


t_6

KDE Plots



(a) Crude



(b) Crude Mito

Figure: KDE plots Crude vs Crude Mito

Normality Test

- H_0: sample comes from a normal distribution
- p-value: small p-value means it is unlikely that data come from normal distribution,
  standard threshold $= 0.05$

	t1	t2	t3	t4	t5	t6
Cru Mito	e-04	e-04	e-05	e-07	e-06	e-09
Cru	e-064	e-087	e-087	e-148	e-108	e-120
Bulk Mito	e-13	e-17	e-18	e-19	e-25	e-28
Bulk	e-174	e-271	e-217	e-202	e-116	e-125

Table: Normality Test, p-value

Select profiles

Select a set $R \subset P$ with m profiles ($|R| = m$) such that the minimum distance between any two profiles in R is maximised

$\arg\max_{R \subset P, |R|=m} \arg\min_{p_1, p_2 \in R} d(p_1, p_2)$ (1)

For a set R, $b(R) = \arg\min_{p_1, p_2 \in R} d(p_1, p_2)$

Algorithm

Let p_1 be the profile that always goes down one unit between time points. $R = p_1; L = P \setminus p1;$

for $i = 2$ to m **do**

let $p \in L$ be the profile that maximises $min_{p_1 \in R^d} d(p, p_1)$

$R = R \bigcup p; L = L \setminus p;$

end for; **return** R

Theorem 1

Let d be a distance metric. Let $R' \subset P$ be the set of profiles that maximizes (1). Let $R \subset P$ be the set of profiles returned by our algorithm, then $b(R) \geq b(R')/2$.

Identifying Significant Model Profiles

- Given a set M of model profiles and a set of genes G, each gene $g \in G$ is assigned to a model expression profile $m_i \in M$ such that $d(e_g, m_i)$ is the minimum over all $m \in M$.
- Count the no of genes assigned to each model profile and denote this for profile m_i as $t(m_i)$.
- Perform hypothesis testing with H_0 : data are memoryless.
- Use a permutation based test:
  - For each possible permutation, assign genes to their closest model profile.
  - s_i^j: no of genes assigned to model profile i in permutation j
  - Set $S_i = \sum_j s_i^j$. Then $E_i = \frac{S_i}{n!}$ is the expected no of genes for each profile model if the data were indeed generated according to H_0.

Continue

- Since each gene is assigned to one of the profiles, we can assume that this $Bin(|G|, E_i/|G|)$.
- Consider the no of genes assigned to p_i to be statistically significant if $P(X \geq t(m_i)) < \alpha/m$.

Distance metric

We use *Correlation Coefficient* here as it can group together genes with similar expression profiles even if their units of change are different.
To take into account negative values, use $gm(x, y) = 1 - \rho(x, y)$.
Lemma 1 $gm(x, z) \leq 2(gm(x, y)) + gm(y, z))$
Shows that correlation coefficient is transitive meaning two highly dissimilar profiles cannot be very similar to a third profile.

Grouping Significant Profiles

Want to group similar model profiles by transforming this problem into a graph problem.

- $G(V, E)$, where V is the set of significant model profiles and E the set of edges.
- δ is the distance threshold to measure the similarity between two model profiles.
- Two profiles $v_1, v_2 \in V$ are connected with an edge iff $d(v_1, v_2) \leq \delta$.
- Cliques in this graph correspond to sets of significant profiles which are all similar to one another.
- Partition the graph into cliques and thus to group significant profiles.

Greedy Algorithm

- Initialize $C_i = p_i$
- Look for a profile p_j such that if $d(p_i, p_j) \leq \delta$ for all profiles $p_k \in C_i$, add p_j to C_i
- Otherwise, stop and declare C_i as the cluster for p_i
- Select cluster with the largest no of genes (by counting the no of genes in each of the profiles that are included in this cluster).
- Stop when all profiles have been assigned to clusters.

K-means++

Outline

- Initialize an empty set M to store the k centroids being selected
- Randomly choose 1st centorid $\mu^{(j)}$ from the input and assign it to M
- For each sample $x^{(i)} \notin M$, find $\min d(x^{(i)}, M)^2$ to any of the centroids in M
- Use a weight probability distribution $= \frac{d(\mu^{(p)}, M)^2}{\sum_i d(x^{(i)}, M)^2}$
- Repeat steps 2 and 3 until k centroids are chosen
- Proceed with the standard k-means algorithm

K-means: Model Selection (Elbow Curve)

Elbow Curve

- Sum of intra-cluster distances between points in a given cluster D_k

$$C_k = \sum_{x_i \in D_k} \sum_{x_j \in D_k} \|x_i - x_j\|^2 = 2n_k \sum_{x_i \in D_k} = \|x_i - \mu_k\|^2$$

- Normalized intra-cluster sums of squares: measure compactness of our clustering

$$W_k = \sum_{k=1}^{K} \frac{1}{2n_k} C_k$$

- Looks at the $\%$ of variance explained as a function of the no of clusters

K-means: Model Selection (Gap Statistics)

Gap Statistics

- Compare $logW_k$ vs null reference distribution of the data, i.e. a distribution with no obvious clustering.
- Optimal K : $\arg\max_k Gap_n(k) = E_n^*\{logW_k\} - logW_k$
- Estimate $E_n^*\{logW_k\} = \frac{1}{B}\sum_{b=1}^{B} log(W_{kb}^*)$, each of which is computed from a Monte Carlo sample $X_1^*, \cdots, X_n^*$ drawn from our reference distribution.
- Simulation error, $s_k = sd(k)\sqrt{1+\frac{1}{B}}$, where $sd(k)^2 = \frac{1}{B}\sum_b (logW_{kb}^* - \frac{1}{B}\sum_{b=1}^{B} log(W_{kb}^*))^2$
- OR Optimal K : $\arg\min_k Gap(k) \geq Gap(k+1) - s_{k+1}$