Gene Expression Data Analysis

- Preprocessing - normalization of data and calculation of pairwise similarity values between elements

- Clustering

- Assessment of results

Preprocessing

- Normalize data (e.g.transform each fingerprint to have mean 0 and variance 1 (Yeung et al, Tamayo et al), have a fixed norm)

- Define a similarity measure between elements (e.g dot product, Euclidean distance (Ben-Dor et al) ,Pearson correlation coefficient (Ben-Dor et al))

- Characterize mates (genes in the same cluster) and non-mates in terms of pairwise similarity values

Assumptions on data

- Assumption: Similarity values between mates (non-mates) are normally distributed with mean μ_T (μ_F) and variance σ_T (σ_F).

- Practical justification: Observed on similarity values between pairs of cDNA fingerprints from blood monocytes dataset whose clustering was known (obtained through hybridization procedures).

- Theoretical justification: Suppose the fingerprint values for each gene i are i.i.d random variables. Let $u = (u_1, ..., u_p)$, $v = (v_1, ..., v_p)$ be two fingerprints and $S_{uv} = \sum_{i=1}^{p} u_i v_j$ be their similarity value. If u and v are mates and $u_i v_i$ has mean μ_T/p with variance σ_T^2/p then by Central Limit Theorem $S_{uv} \sim N(\mu_T, \sigma_T^2)$

Parameter Estimation

- If a true partition of the subset of the data is available then the sample mean and variances are used as estimates

- Otherwise use EM

Outline of CLICK (CLuster Identification via Connectivity Kernels)

- Forming similarity graph
- Basic-CLICK : produces some clusters
- Theoretical guarantees
- Full Algorithm : merges clusters from Basic-CLICK together

Forming similarity graph

Let $f(x|\mu_T, \sigma_T)$ ($f(x|\mu_F, \sigma_F)$) be the mates (non-mates) probability density function. Let p_{mates} be the probability any two random elements are mates.

Let S be the pairwise similarity matrix for M (recall M)

Form similarity graph $G = (V, E)$ where $V = \{e_1, ..., e_n\}$ (set of genes) and

$$w_{ij} = \log \frac{p_{mates} f(S_{ij}|i,j \text{ are mates })}{(1-p_{mates}) f(S_{i}j|i,j \text{ are non-mates })}$$

where

$$f(S_{ij}|i, j \text{are mates}) = \frac{1}{\sqrt{2\pi}\sigma_T} e^{-\frac{(S_{ij}-\mu_T)^2}{2\sigma_T^2}}$$

Algorithm Outline

Basic-CLICK(G):
If $V(G) = \{v\}$ **then** move v to singleton set R
Else if G is a kernel **then**
    Output $V(G)$
Else
    $(H, \tilde{H}) \leftarrow$ MinWeightCut(G).
    Basic-CLICK(H)
    Basic-CLICK($\tilde{H}$)

Properties of complete similarity graph G

G is a *kernel* iff $V(G)$ is a subset of some true cluster.

For a cut C in G define the hypothesis:

- H_0^C : C contains only edges between non-mates
- H_1^C : C contains only edges beween mates

G is a kernel iff H_1^C is accepted for every cut C in G. Let $W(C)$ be the weight of a cut C. Then **Lemma 1** shows that

$$W(C) = \log \frac{Pr(H_1^C|C)}{Pr(H_0^C|C)}$$

Note: The minimum weight cut C has the minimum posterior probability ratio $\frac{Pr(H_1^C|C)}{Pr(H_0^C|C)}$

Sketch of proof of Lemma 1

1. Using Bayes Theorem we get:

$$\frac{Pr(H_1^C|C)}{Pr(H_0^C|C)} = \frac{Pr(H_1^C)f(C|H_1^C)}{Pr(H_0^C)f(C|H_0^C)} \qquad (1)$$

2. Using independence assumption we get:

$$f(C|H_1^C) = \frac{1}{\sqrt{2\pi}\sigma_T} e^{-\frac{(S_{ij}-\mu_T)^2}{2\sigma_T^2}}$$

3. and $Pr(H_1^C)$ is $p_{mates}^{|C|}$.

4. Similarly values are defined for $f(C|H_0^C)$ and $Pr(H_0^C)$

5. Substituting the above values in (1) and the definition of w_{ij} gives us the desired result.

How to decide if a complete G is a kernel

1. G is a kernel iff $\forall$ cuts C, H_1^C is accepted iff $\forall$ cuts C, $Pr(H_1^C|C) > Pr(H_0^C|C)$.

2. Let C be a minimum cut, hence by Lemma 1, $\forall$ cuts $C^{'} \neq C$

$$\log \frac{Pr(H_1^C|C)}{Pr(H_0^C|C)} = W(C) \leq W(C^{'} = \log \frac{Pr(H_1^{C^{'}}|C^{'})}{Pr(H_0^{C^{'}}|C^{'})}$$

3. H_1^C is accepted for C iff $H_1^{C^{'}}$ is accepted $\forall$ cuts $C^{'} \neq C$

4. Thus, H_1^C is accepted (G is a kernel) iff $W(C) > 0$

What about when G is incomplete

- Lemma 1 holds only for complete G. But the input may be a threshold matrix S_t. To apply the test criterion one needs to approximate the constribution of missing edges from C to the posterior proabability ratio $Pr(H_1^C|C)/Pr(H_0^C|C)$

- Let $F = (H \times \tilde{H}) \setminus C, r = |H||\tilde{H}|$. Define

$$W^*(C) = \log \frac{\prod_{(i,j)\in F} p_mates Pr(S_{ij} \leq t_S | H_1^C)}{\prod_{(i,j)\in F} (1-p_mates) Pr(S_{ij} \leq t_S | H_0^C)}$$

$$= (r - |C|) \log \frac{p_mates \Phi((t_S - \mu_T)/\sigma_T)}{(1-p_mates)\Phi((t_S-\mu_F)/\sigma_F)}$$

where t_S is the threshold and Φ is the cumulative standard normal distribution

- Now G is a kernel iff $W(C) + W^*(C)$ is a kernel

Recap

Basic-CLICK(G):
If $V(G) = \{v\}$ **then** move v to singleton set R
Else if G is a kernel **then**
    Output $V(G)$
Else
    $(H, \tilde{H}) \leftarrow$ MinWeightCut(G).
    Basic-CLICK(H)
    Basic-CLICK($\tilde{H}$)

How to merge kernels, singletons and post-processing

- Let $\mathcal{L}$ and R be the set of kernels and singletons respectively.
- **Adoption**
  - Search for singleton v and a kernel K whose pairwise fingerprint similarity is maximum (fingerprint of K is the mean fingerprint of all $e \in K$).
  - Let m be the maximum. If $m >$ some predefined threshold then v is added (adopted) to K and removed from R.
- When no change in adoption step, then recursively cluster R by discarding remaining vertices from initial graph.
- **Merge**: Merge pair of kernels whose similarity exceeds a threshold
- Finally perform a last adoption step

The Full Algorithm

$R \leftarrow N$
While some change occurs **do**:
    Execute Basic-CLICK(G_R)
    Let $\mathcal{L}$ be list of kernels produced
    Let R be the set of singletons produced
    Adoption($\mathcal{L}, R$)
Merge($\mathcal{L}$)
Adoption($\mathcal{L}, R$)

Assessment of results - When true results are known

- Let C be the matrix representing true solution - $C_{ij} = 1$ iff i and j belong to same clustert, and let T be the matrix representing the algorithm solution. Then n_{11} is number of mates detected by the solution $(i.e. T_{ij} = C_{ij} = 1)$, n_{00} is number of non-mates $(i.e. T_{ij} = C_{ij} = 0)$ detected, and n_{01} and n_{10} are the differences.

- **Minkowski measure** :

$$\frac{||T-C||}{||T||} = \sqrt{\frac{n_{01}+n_{10}}{n_{11}+n_{10}}}$$

  The lower the better

- **Jaccard coefficient**:

$$\frac{n_{11}}{n_{11}+n_{10}+n_{01}}$$

  Higher the better, optimal = 1

Assessment of results - When true results are NOT known

- Let $cl(u)$ be the cluster of u, $F(X)$ and $F(u)$ be the fingerprints of cluster X and element u respectively, and $S(x, y)$ be the similarity measure of x and y.
- Define:

$$H_{Ave} = \frac{1}{|N|} \sum_{u \in N} S(F(u), F(cl(u)))$$

$$H_{Min} = min_{u \in N} S(F(u), F(cl(u)))$$

$$S_{Ave} = \frac{1}{\sum_{i \neq j} |X_i||X_j|} \sum_{i \neq j} |X_i||X_j| S(F(X_i), F(X_j))$$

$$S_{Max} = max_{i \neq j} S(F(X_i), F(X_j))$$

- Note
  - H_{Ave} is similar to k-means objective function
  - Solution improves if H_{Ave} and H_{Min} increases and S_{Ave} and S_{Min} decreases.

- **Assessment of results given only similarity matrix S**

  Define:

$$H = \frac{\{(i,j)|cl(i)=cl(j) \text{ and } (i,j)\in E\}}{|\{(i,j)|cl(i)=cl(j)\}|}$$

$$S = \frac{\{(i,j)|cl(i)\neq cl(j) \text{ and } (i,j)\in E\}}{|\{(i,j)|cl(i)\neq cl(j)\}|}$$

Results on real data - Yeast Cell Cycle

- Expression levels of 6,218 gene transcripts (ORFs) measured at 10 minute intervals over 160 minutes. 826 genes were analyzed and compared to GENECLUSTER (Tamayo et al) which uses Self Organizing Maps.
- **Data Preprocessing**:
  - Removal of 90 minute time point
  - Normalization of expression levels to have mean 0 and variance 1.
    Note: Yeung et al have shown that transforming the data for this dataset brings it closer to the Gaussian mixture assumption.
  - Similarity function is Pearson correlation coefficient

Program	#Clusters	Homogeneity		Separation	
		H_{Ave}	H_{Min}	S_{Ave}	S_{Max}
CLICK	30	0.8	-0.19	-0.07	0.65
GENE-CLUSTER	30	0.74	-0.88	-0.02	0.97

- **Some more results**:
  - Used subset of 698 out of 800 as used in Spellman et al.
  - Removed 90 minutes timepoint
  - Genes were permuted in the expression matrix
  - True clustering is not accurate as reported by Eisen

Program	#Clusters	Hom	Sep
CLICK	6	0.66	-0.1
K-Means	49	0.63	0.09
GeneCluster (SOM)	6	0.62	-0.07
CAST	5	0.6	-0.15
True	5	0.57	-0.13

Human Fibroblasts response to Serum

- Human fibroblasts were deprived of serum for 48 hours and then stimulated by adding serum. Expression level were measured at 12 time points. 517 genes from the original 8,613 were analyzed and compared to hierarchical clustering results of Eisen et al.
- **Data Preprocessing**:
  - Fingerprints were transformed to have a fixed norm
  - Similarity function is dot product - same as Eisen et al.

Program	#Clusters	Homogeneity		Separation	
		H_{Ave}	H_{Min}	S_{Ave}	S_{Max}
CLICK	10	0.88	0.13	-0.34	0.65
Hierarchical	10	0.87	-0.75	-0.13	0.9