# Outline

- Overview of Gene Expression Clustering

- Graph Partitioning Model
  - Graph Partitioning Theory
  - Spectral Graph Partitioning

- Isoperimetric Graph Partitioning for Gene Analysis
  - Isoperimetric Graph Partitioning Model
  - Algorithm Derivation
  - Feature Selection via Two-way Ordering of Gene Expression

- Experiments and Results

# Gene Expression Data from Microarray

| | Sample 1…. | Sample J …. | Sample m |
|---|---|---|---|
| Gene 1 | $A_{11}$ | $A_{1J}$ | $A_{1M}$ |
| Gene …. | …. | …. | …. |
| Gene I | $A_{I1}$ | $A_{IJ}$ | $A_{IM}$ |
| Gene …. | …. | …. | …. |
| Gene N | $A_{N1}$ | $A_{NJ}$ | $A_{NM}$ |

Could be time, environmental, source (tissue /organ /cancerous) etc..

Usually log of relative expression with respect to Control. +/- tells whether over or under expressed.

# Gene Expression Data Clustering

- Gene Clustering : Grouping genes with similar expression patterns based on the samples
  - Unravel relations between genes
  - Deduce the function of genes
  - Reveal the underlying regulatory gene network

- Sample Clustering: Grouping samples corresponding to particular phenotypes (Normal vs. Tumor)
  - Classify different samples
  - Discover new subtypes of samples

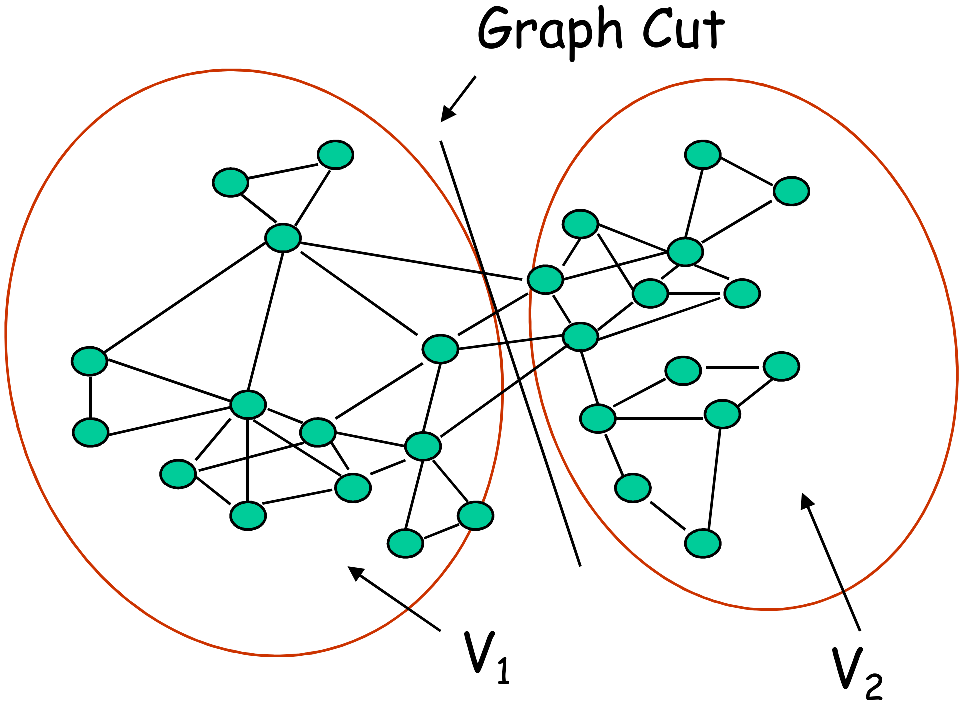
***NOTE: Few samples (~50) and large dimension (~10,000) of genes, samples clustering is more difficult than genes clustering.***

# Graph Partitioning Theory

- Clustering can be viewed as partitioning a weighted graph
- *Bi-partitioning* task:
  - Divide vertices into two disjoint groups $(V_1, V_2)$
- *Graph Cut:* find the minimal cut between groups

$$Cut(V_1, V_2) = \sum_{i \in v_1, j \in v_2} A_{ij}$$

where, A is adjacency (affinity matrix) represents edge weights



{Vertex: samples in gene expression data
Edge weights: similarity between samples

# Spectral Graph Partitioning Algorithm

- Identify an optimal partition is NP-hard
- Find Min-Cut between $V_1$ and $V_2$ with balance weights (Normalized Cuts) :

$$\min_{x \neq 0} \frac{x^T L x}{x^T D x}, s.t. x^T De = 0$$

where, x is clustering membership indicator vector, L is Laplacian matrix, D is degree matrix, e =[1,1, …1]$^T$

- Normalized cuts solution can be solved by a simple eigenvector x:

$$Lx = \lambda Dx$$

# Isoperimetric Graph Partitioning Model

- Optimization of an isoperimetric constant:

$$h = \inf_{S} \frac{|\partial S|}{Vol_{S}}$$

where, h is the infimum of the ratio over all possible, S is a region in the manifold, $Vol_S$ denotes the volume of region S, $|\partial S|$ is the area of the boundary of region S.

- Find minimum isoperimetric ratio ( isoperimetric constant) is NP-hard
- Minimum isoperimetric ratio can be solved by a parse system of linear equations

# Isoperimetric Graph Partitioning Algorithm

- Isoperimetric ratio can be written as:

$$h = \min_x \frac{x^T L x}{x^T d}$$

where, x is clustering membership indictor vector, L is Laplacian matrix, d is degree vector.

- Minimizing cost function:

$$Q(x) = x^T L x - \Lambda(x^T d)$$

where, $\Lambda$ is lagrange multiplier

- Linear system solution:

$$Lx = \Lambda d$$

# Isoperimetric Graph Partitioning Algorithm (Cont.)

- Challenge: matrix L is singular--all rows and columns sum to zero.

- Find a unique solution $x_0$ for a nonsingular system of equations:

$$L_0 x_0 = d_0$$

where, $L_0$ comes from L by removing the vertex of largest degree, $x_0$ and $d_0$ come from x and d by removing corresponding rows of x and d.

# Feature Selection via Two-way Ordering

- Why need unsupervised feature selection?

  Conditions:

  - high dimensionality of feature spaces (many genes are irrelevant or redundant)
  - without prior knowledge of cluster structure (some genes correspond to new phenotypes or subtypes)

  Objective: Improve performance

- Two-way ordering genes/samples
  - Bipartite graph to represent gene expression data
  - Using SVD (singular value decomposition) to re-ordering genes/samples
  - Discard irrelevant genes

# Two-way Ordering Algorithm

- Symmetric weighted adjacency matrix W for the bipartite graph:

$$W = \begin{pmatrix} 0 & A \\ A^T & 0 \end{pmatrix}$$

- Compute the second largest principle components $u_2$ and $v_2$ from:

$$\tilde{A} = D_g^{-1/2} A D_s^{-1/2}$$

where $D_g(i,i) = \sum_j W_{ij}, D_s(i,i) = \sum_i W_{ij}$ and g is an index permutation of

genes , and s is an index permutation of samples

# Two-way Ordering Algorithm (cont.)

- Get index permutation for genes $$g_2 = D_g^{-1/2} u_2$$

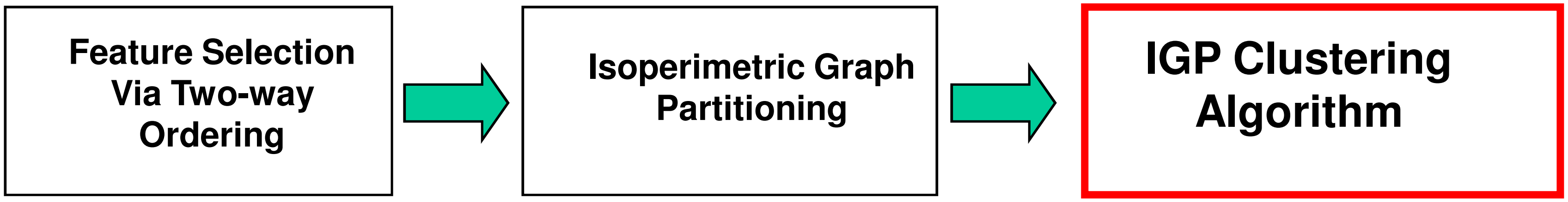
  and index permutation for samples $$s_2 = D_s^{-1/2} v_2$$

- Sort $g_2$ and $s_2$ to increasing order to reorder genes and samples to get reordering matrix $A'$

- Discard genes in the middle of matrix $A'$

# Isoperimetric Graph Partitioning Clustering Algorithm

- Combines two-way ordering for feature selection with isoperimetric graph partitioning to improve performance:
  - Feature selection via two-way ordering: eliminate irrelevant or redundant genes
  - Isoperimetric graph partitioning: group gene expression samples through a graph theoretical approach


Feature Selection Via Two-way Ordering
Isoperimetric Graph Partitioning
IGP Clustering Algorithm

# Experiments

- Colon tumor tissues:
  - 62 samples:

    40 tumor biopsies from tumors (negative)

    22 normal biopsies from healthy parts of colons (positive)
  - 2000 out of around 6500 genes are selected based on the confidence in the measured expression levels

- Leukemia Subtypes:
  - 38 bone marrow samples

    27 from acute lymphoblastic leukemia (ALL)

    11 from acute myeloid leukemia (AML)
  - 7129 probes from 6817 human genes

Data comes from Kent Ridge Bio-medical Data Set Repository (http://sdmc.lit.org.sg/GEDatasets/Datasets.html)

# Results I

- Clustering results are evaluated by
  - Q-accuracy (the higher, the better):

$$\sum_i t_{ii} / N$$

  - Isoperimetric ratio (the lower, the better):

$$h = \min_x \frac{x^T Lx}{x^T d}$$

IGP: Isoperimetric Graph Partitioning

SGP: Spectral Graph Partitioning

| | SGP | | IGP | |
|---|---|---|---|---|
| m genes | Q-accuracy | Iso. ratio | Q-accuracy | Iso. ratio |
| 2000 | 0.5806 | 0.9892 | 0.6290 | 0.5156 |
| 800 | 0.5968 | 0.9563 | 0.7258 | 0.4984 |
| 400 | 0.7258 | 0.9132 | 0.7419 | 0.4897 |
| 200 | 0.8065 | 0.8669 | 0.8226 | 0.4852 |

Table I. The comparison of Q-accuracy and Isoperimetric ratio of IGP and SGP for clustering colon cancer/normal samples based on selective genes through two-way ordering.

| | SGP | | IGP | |
|---|---|---|---|---|
| m genes | Q-accuracy | Iso. ratio | Q-accuracy | Iso. ratio |
| 7122 | 0.5263 | 1.0165 | 0.6842 | 0.5314 |
| 3000 | 0.5263 | 0.9995 | 0.7895 | 0.5144 |
| 2000 | 0.5526 | 0.9905 | 0.7895 | 0.5106 |
| 1000 | 0.6053 | 0.9730 | 0.7895 | 0.5023 |
| 400 | 0.7105 | 0.9423 | 0.7368 | 0.5040 |
| 200 | 0.7105 | 0.9027 | 0.7368 | 0.4696 |

Table II. The comparison of Q-accuracy and Isoperimetric ratio of IGP and SGP for clustering ALL/AML leukemia subtypes based on selective genes through two-way ordering.

# Results II

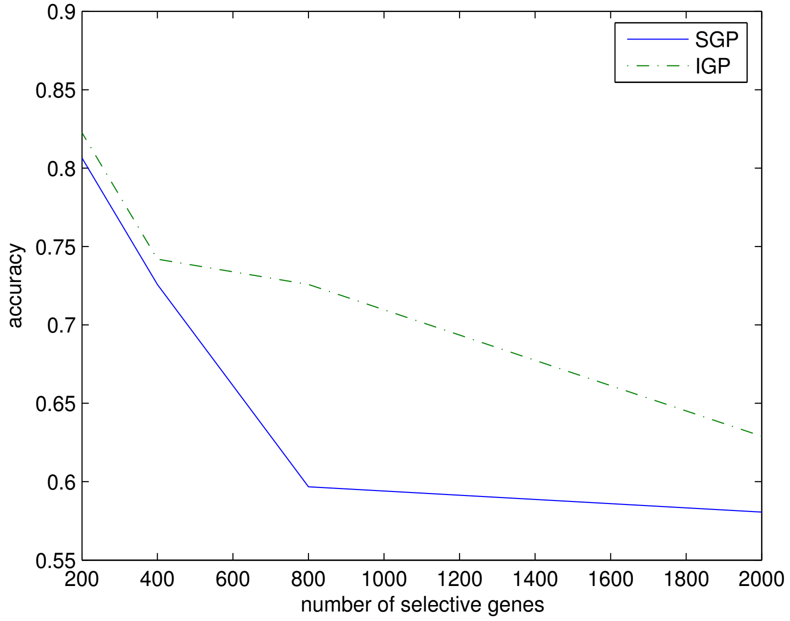


Fig. 1. Comparison of clustering accuracy between IGP (dot line) and SGP (dark line) on colon cancer/normal samples.

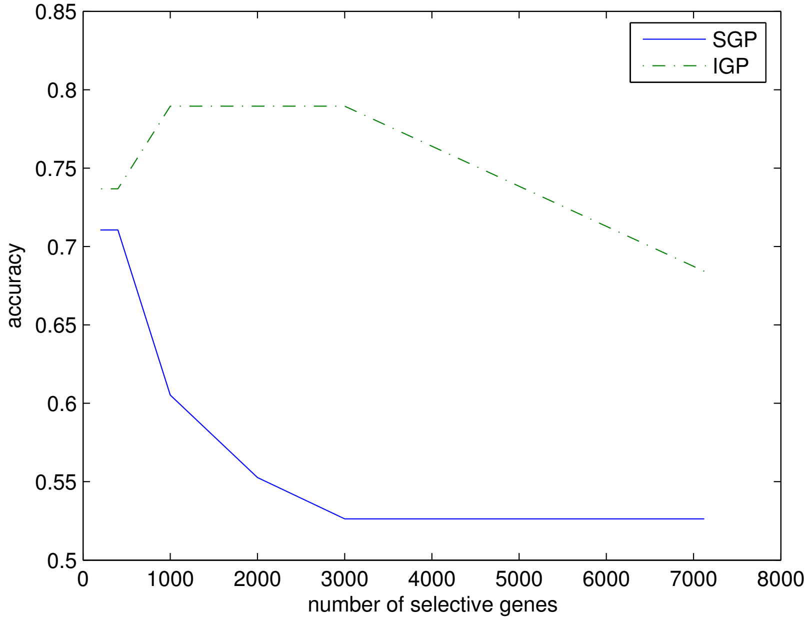


Fig. 2. Comparison of clustering accuracy between IGP (dot line) and SGP (dark line) on ALL/AML leukemia subtypes samples.

# Conclusion

- Isoperimetric graph partitioning model to group biological samples from gene expression data:

  - outperforms spectral graph partitioning with higher accuracies and lower isoperimetric ratios

- Integrated with unsupervised feature selection via two-way ordering of gene expression data, accuracies of clustering are improved significantly.