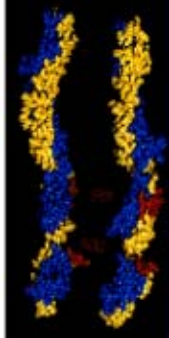
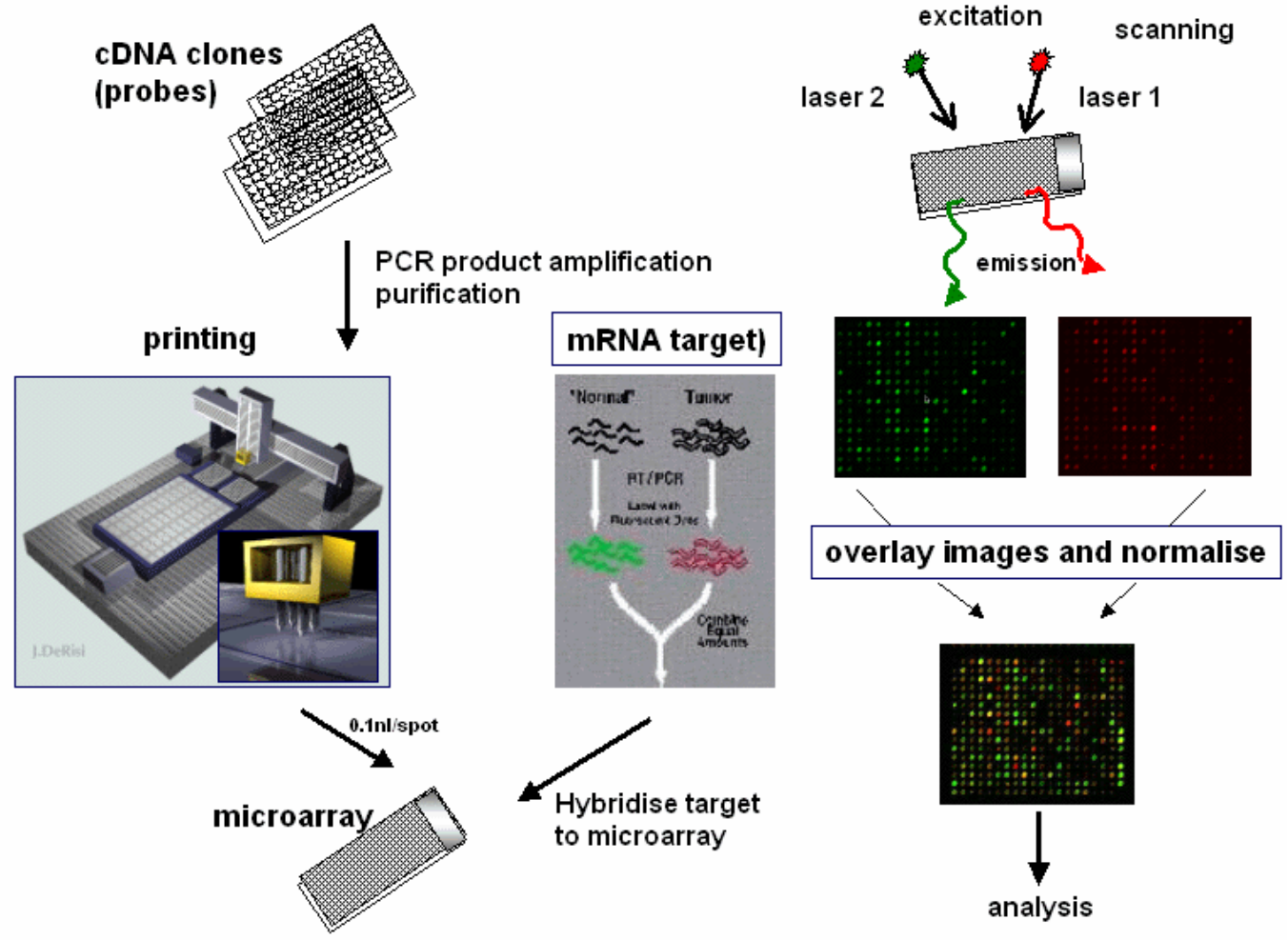
Mining Phenotypes and Informative Genes from Gene Expression Data



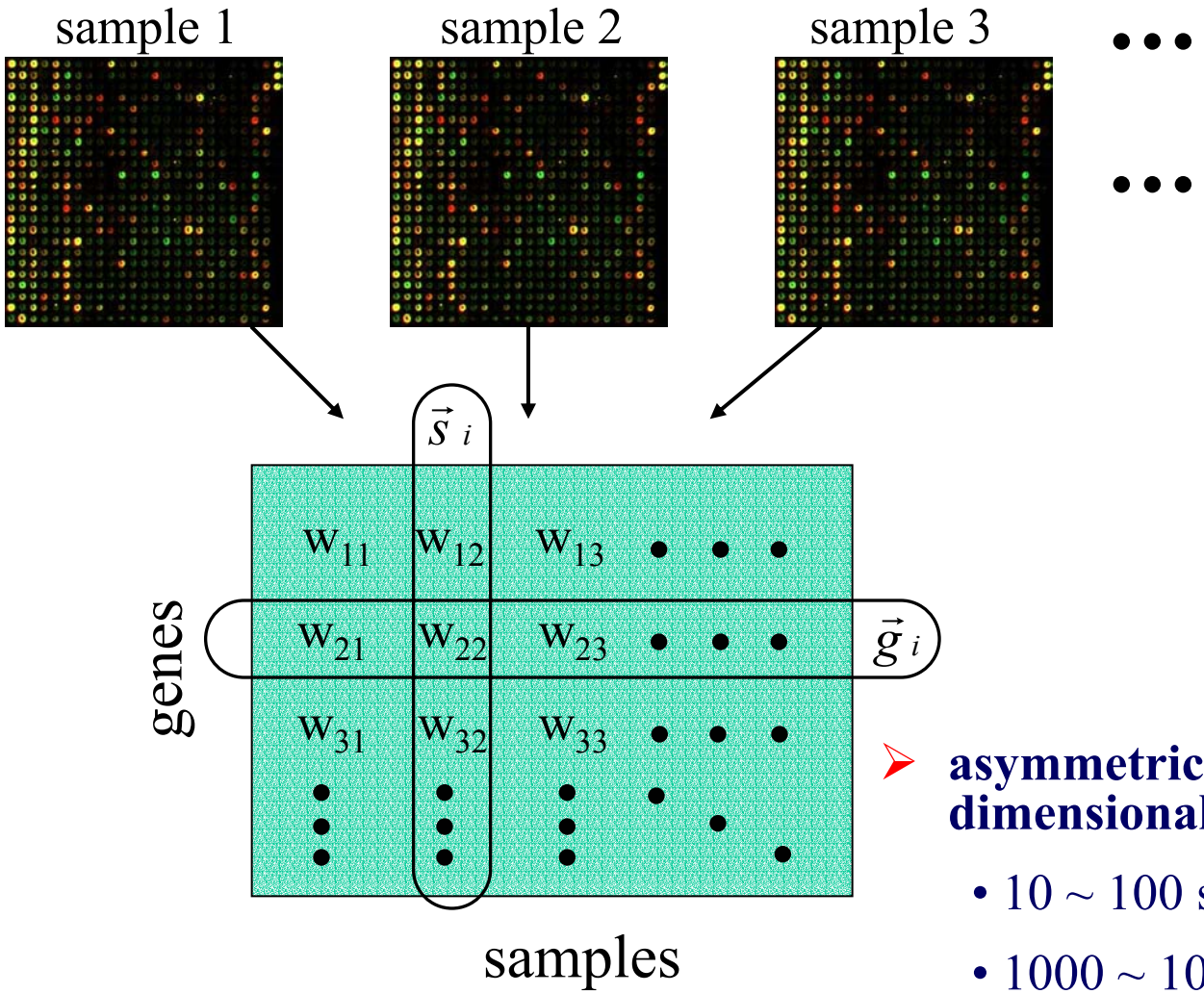
Chun Tang, Aidong Zhang and Jian Pei
Department of Computer Science and Engineering
State University of New York at Buffalo

cDNA Microarray Experiment



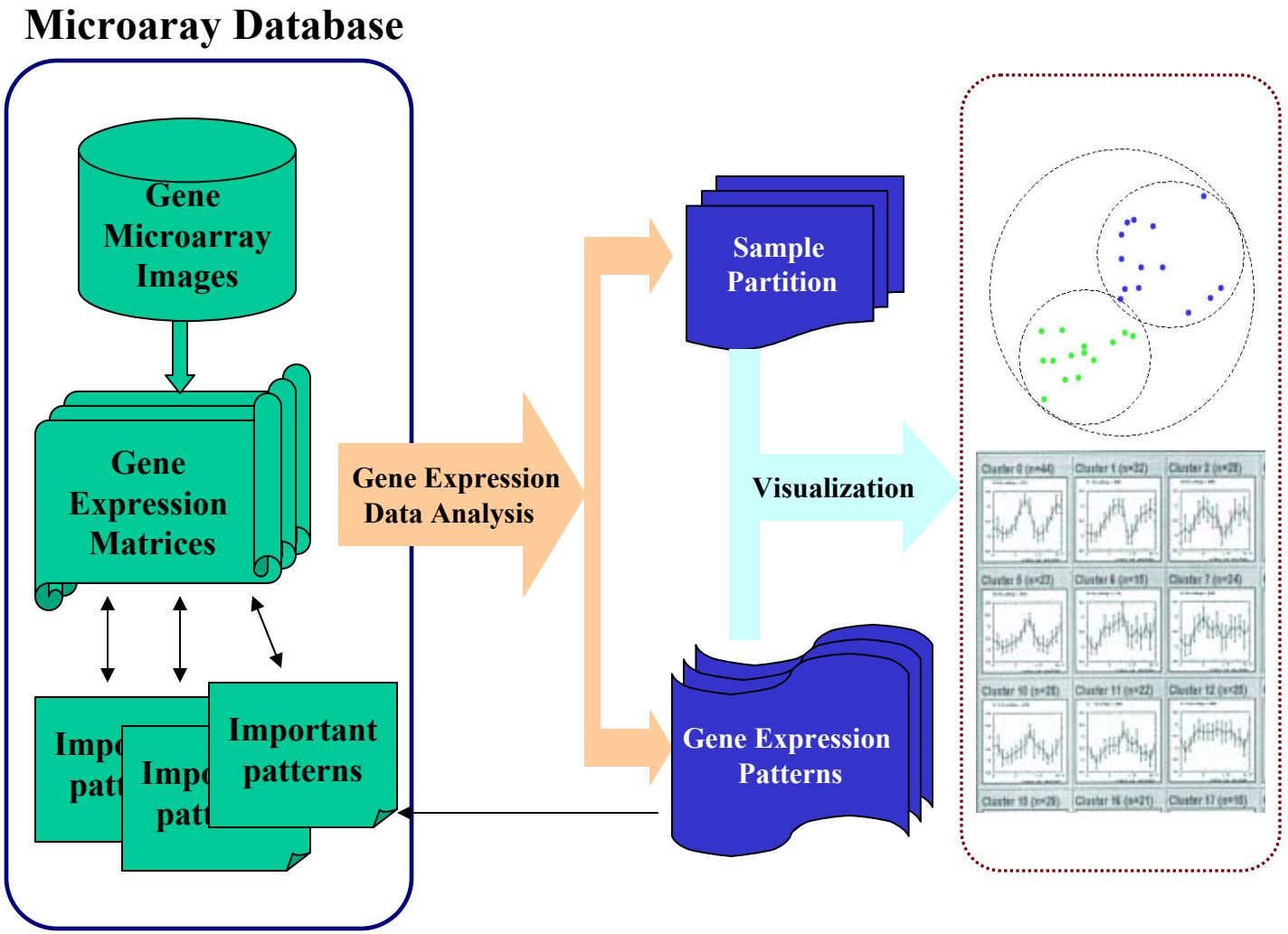
http://www.ipam.ucla.edu/programs/fg2000/fgt_speed7.ppt

Microarray Data



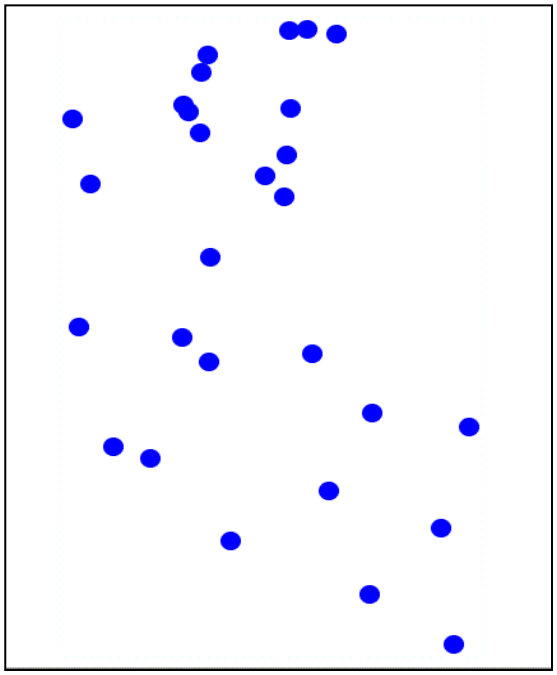
- **asymmetric dimensionality**
  - 10 ~ 100 samples
  - 1000 ~ 10000 genes

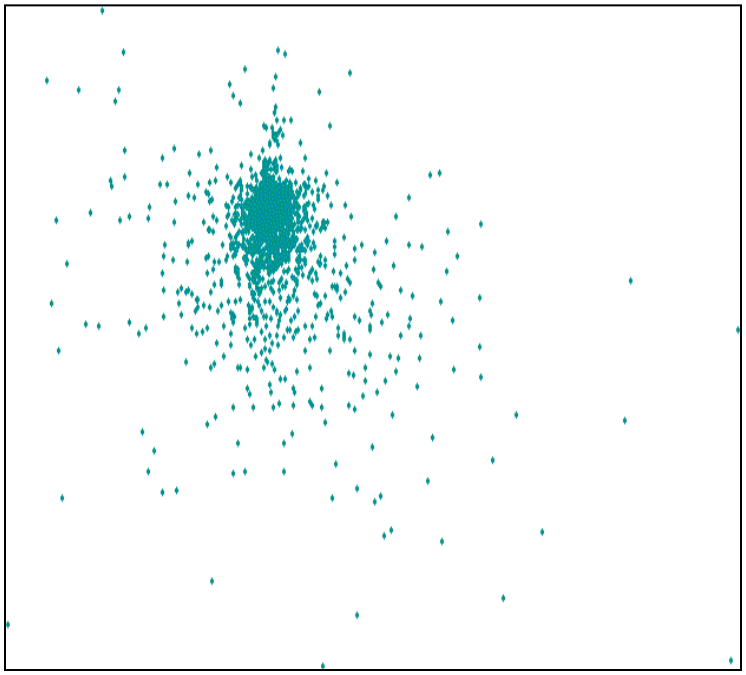
Scope and Goal



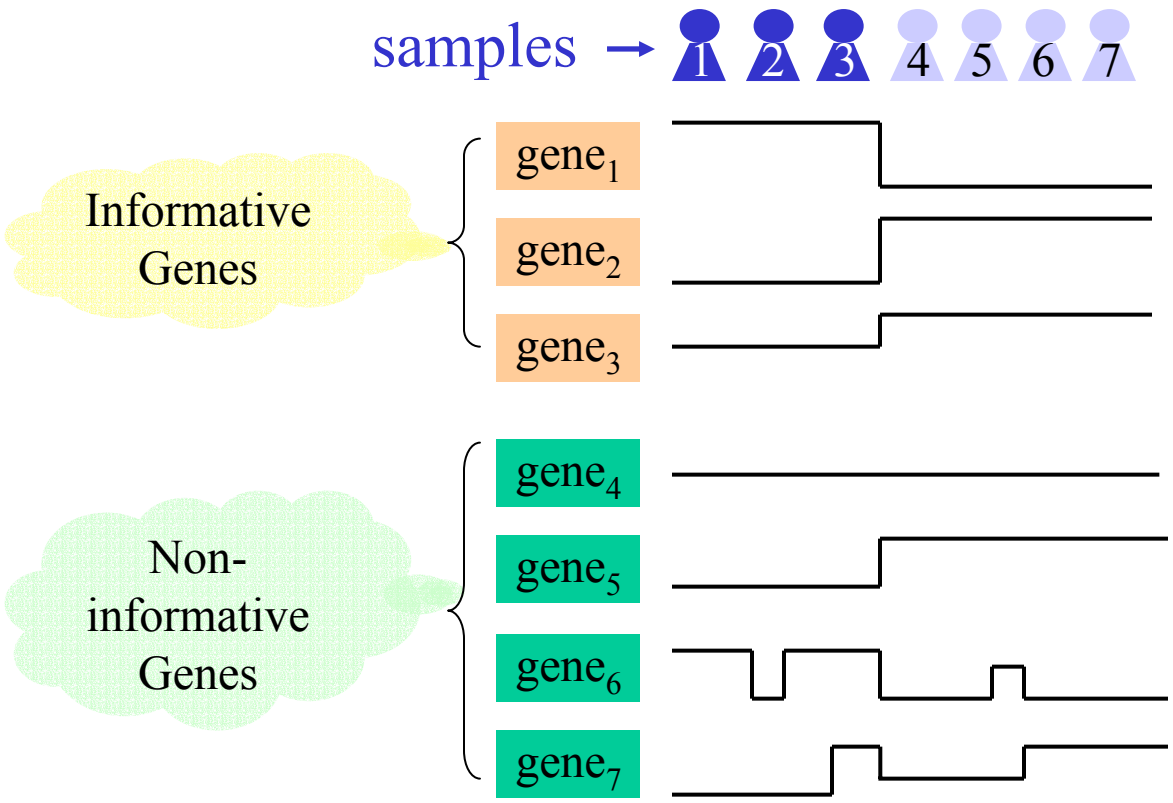
Microarray Data Analysis

- **Analysis from two angles**
  - **sample** as object, gene as attribute
  - gene as object, sample/condition as attribute





Sample-based Analysis



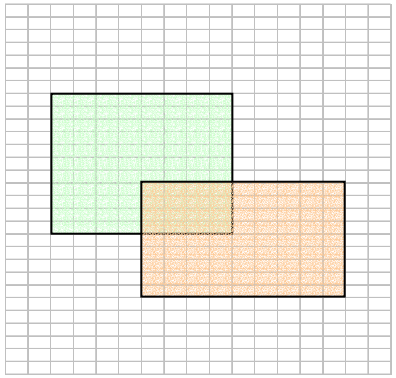
Related Work

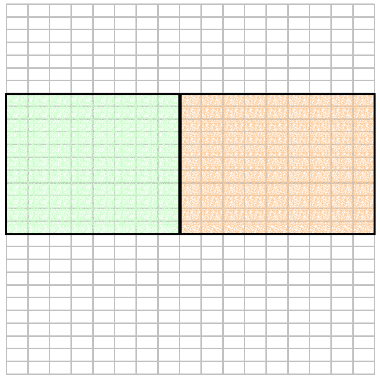
- **New tools using traditional methods :**

TreeView
CLUTO
CIT
SOTA
GeneSpring
J-Express
CLUSFAVOR

  - SOM
  - K-means
  - Hierarchical clustering
  - Graph based clustering
  - PCA

- **Clustering with feature selection:**
- **Subspace clustering**





Quality Measurement

- **Intra-phenotype consistency:**

$$Con\ (G', S') = \frac{1}{|G'| \bullet (|S'| - 1)} \sum_{\vec{g}_i \in G'} \sum_{\vec{s}_j \in S'} (w_{i,j} - \overline{w}_{i,S'})^2$$

- **Inter-phenotype divergency:**

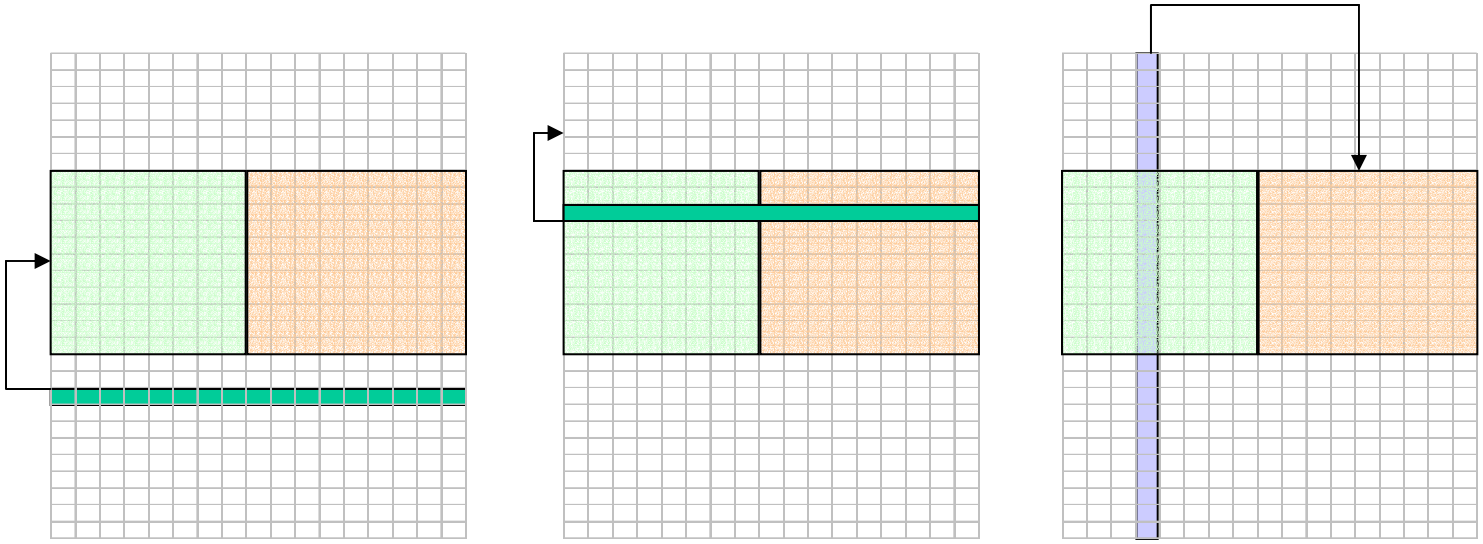
$$Div\ (G', S_1, S_2)) = \frac{\sum_{\vec{g}_i \in G'} \left| \overline{w}_{i,S_1} - \overline{w}_{i,S_2} \right|}{|G'|}$$

- **The quality of phenotype and informative genes:**

$$\Omega = \frac{1}{\sum_{S_i, S_j (1 \le i, j \le K; i \ne j)} \frac{\sqrt{Con\ (G', S_i) + Con\ (G', S_j)}}{Div\ (G', S_i, S_j)}}$$

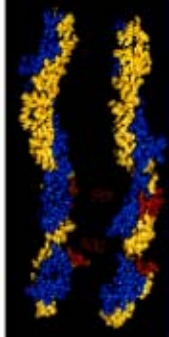
Heuristic Searching

- Starts with a random K-partition of samples and a subset of genes as the candidate of the informative space.
- Iteratively adjust the partition and the gene set toward the optimal solution.
  - for each gene, try possible insert/remove
  - for each sample, try best movement.



Mutual Reinforcing Adjustment

- Divide the original matrix into a series of exclusive sub-matrices based on partitioning both the samples and genes.
- Post a partial or approximate phenotype structure called a *reference partition* of samples.
  - compute *reference degree* for each sample groups;
  - select *k* groups of samples;
  - do partition adjustment.
- Adjust the candidate informative genes.
  - compute W for *reference partition* on *G*
  - perform possible adjustment of each genes
- Refinement Phase



Reference Partition Detection

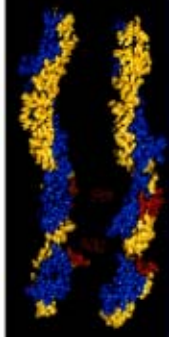
- *Reference degree*: measurement of a sample group over all gene groups

$$ref\ (S_j) = \log |S_j| \sum_{G_i \in G'} \frac{1}{Con\ (G_i, S_j)}$$

- The sample group having the highest reference degree – S_{p0} , S_{p1} , S_{p2} … S_{px} ,…

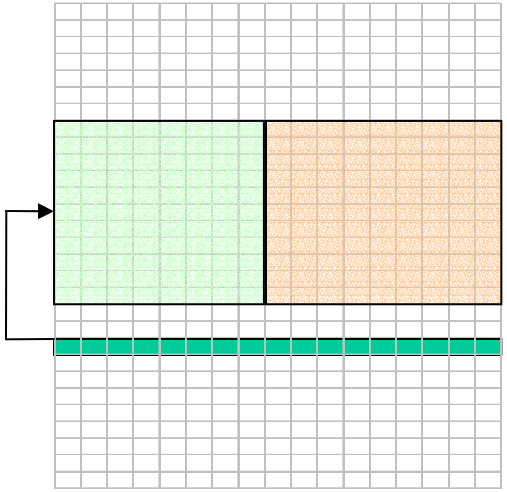
$$Ran\ (S_{px}) = \log |S_{px}| \sum_{G_i \in G'} \frac{\sum_{t=0}^{x-1} Div\ (G_i, S_{px}, S_{pt})}{Con\ (G_i, S_{px})}$$

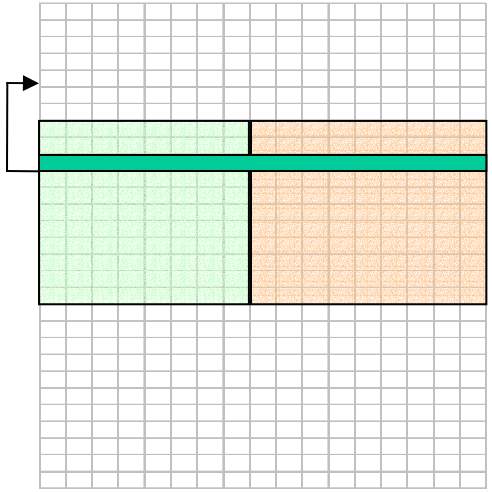
- Partition adjustment: check the missing samples



Gene Adjustment

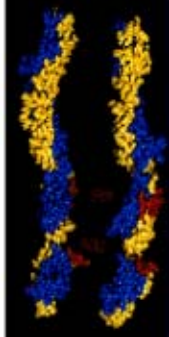
- For each gene, try possible insert/remove





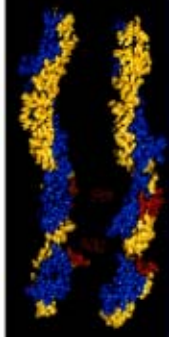
Refinement Phase

- The partition corresponding to the best state may not cover all the samples.
- Add every sample not covered by the reference partition into its *matching* group – the phenotypes of the samples.
- Then, a gene adjustment phase is conducted. We execute all adjustments with a positive quality gain – informative space.
- Time complexity $O(n*m^2*I)$

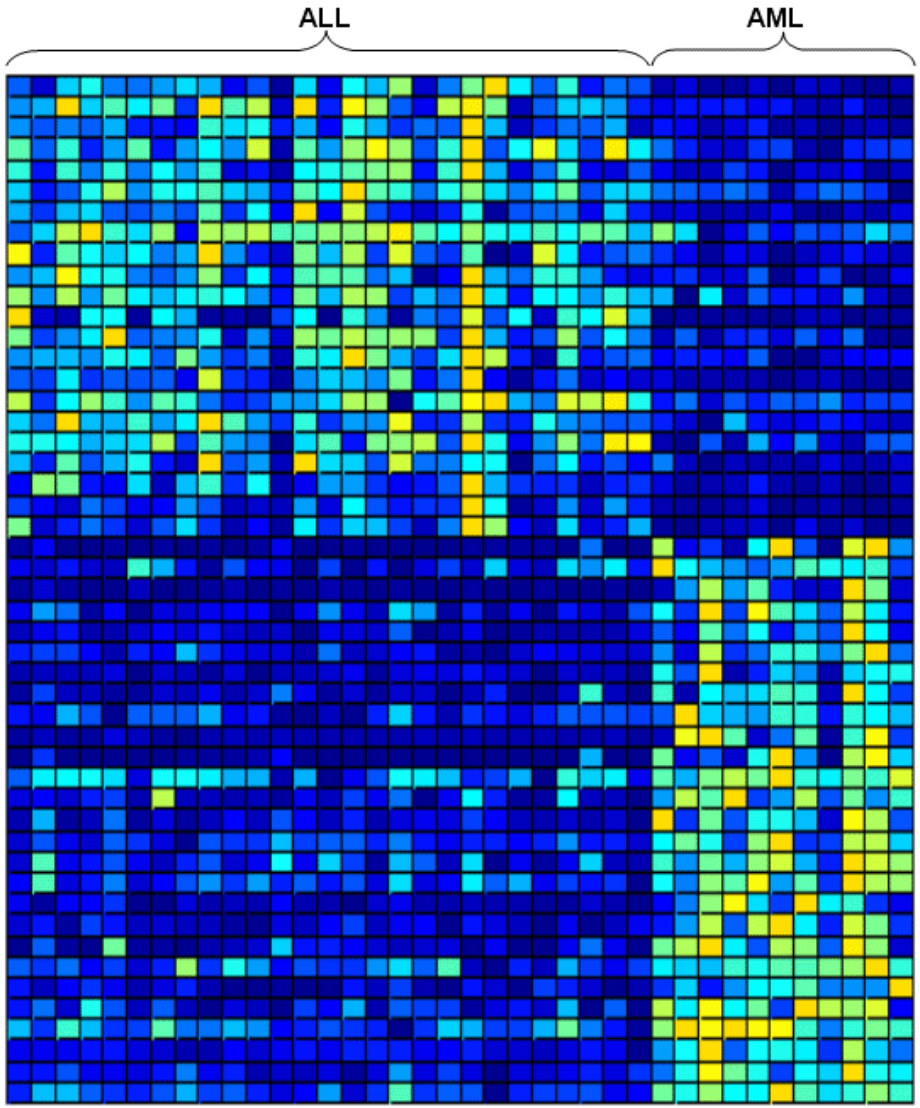


Phenotype Detection

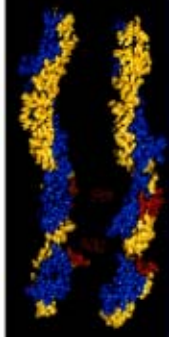
Data Set	MS-IFN	MS-CON	Leukemia-G1	Leukemia-G2	Colon	Breast
Data Size	4132*28	4132*30	7129*38	7129*34	2000*62	3226*22
J-Express	0.4815	0.4851	0.5092	0.4965	0.4939	0.4112
SOTA	0.4815	0.4920	0.6017	0.4920	0.4939	0.4112
CLUTO	0.4815	0.4828	0.5775	0.4866	0.4966	0.6364
Kmeans/PCA	0.4841	0.4851	0.6586	0.4920	0.4966	0.5844
SOM / PCA	0.5238	0.5402	0.5092	0.4920	0.4939	0.5844
δ-cluster	0.4894	0.4851	0.5007	0.4538	0.4796	0.4719
Heuristic	0.8052	0.6230	0.9761	0.7086	0.6293	0.8638
Mutual	0.8387	0.6513	0.9778	0.7558	0.6827	0.8749



Informative Gene Selection



Gene description	Probe	Match
C-myb	F009426	2
PROTEASOME IOTA	X59417	2
Myosin light chain	M31211	2
SNF2	D26156	2
Crystallin zeta	L13278	1
CTP synthetase	X52142	0
Cyclin D3	M92287	2
Actinin alpha 2	M86406	0
Thymopoietin beta	U09087	1
HKR-T1	S50223	2
Aldehyde reductase 1	X15414	1
MB-1	U05259	2
TFIIE beta	X63469	1
Clone 22	AF009426	0
RbAp48	X74262	2
CD19 CD19 antigen	M28170	0
Op18	M31303	2
ADH4	X56411	0
MCM3	D38073	2
Inducible protein	L47738	2
E2A	M31523	2
Topoisomerase II β	Z15115	2
IL8	M28130	1
M-CSF	M37435	0
Cystatin C	M27891	1
HU-K5	U67963	0
Cathepsin D	M63138	1
CD11C	M81695	1
p62	U46751	1
MAD3	M69043	1
HOX 2.2	X58431	0
adipsin	M84526	1
IL-8 PRECURSOR	Y00787	1
Neuromedin B mRNA	L08246	1
Catalase	X04085	1
MCL1	L08246	1
CyP3	M80254	2
Ferritin, light polypeptide	M11147	1
ATPase	M62762	2
CD33	M23197	2
LYN	M16038	2
Proteoglycan 1	X17042	2
Thrombospondin-1	U12471	1
Leptin receptor	Y12670	2
IGIF	D49950	1
HoxA9	U82759	2
Zyxin	X95735	2
LTC4 synthase	U50136	2
Fumarylacetoacetate	M55150	2



References

- Agrawal, Rakesh, Gehrke, Johannes, Gunopulos, Dimitrios and Raghavan, Prabhakar. Automatic subspace clustering of high dimensional data for data mining applications. In SIGMOD 1998, Proceedings ACM SIGMOD International Conference on Management of Data, pages 94–105, 1998.

- Ben-Dor A., Friedman N. and Yakhini Z. Class discovery in gene expression data. In Proc. Fifth Annual Inter. Conf. on Computational Molecular Biology (RECOMB 2001), pages 31–38. ACM Press, 2001.

- Cheng Y., Church GM. Biclustering of expression data. Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology (ISMB), 8:93–103, 2000.

- Golub T.R., Slonim D.K., Tamayo P., Huard C., Gassenbeek M., Mesirov J.P., Coller H., Loh M.L., Downing J.R., Caligiuri M.A., Bloomfield D.D. and Lander E.S. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science, Vol. 286(15):531–537, October 1999.

- Xing E.P. and Karp R.M. Cliff: Clustering of high-dimensional microarray data via iterative feature filtering using normalized cuts. Bioinformatics, Vol. 17(1):306–315, 2001.