Fishing Expedition

A Supervised Approach to Extract Patterns from a Compendium of Expression Profiles

Zhen Zhang, Grier Page, Hong Zhang

Johns Hopkins School of Medicine

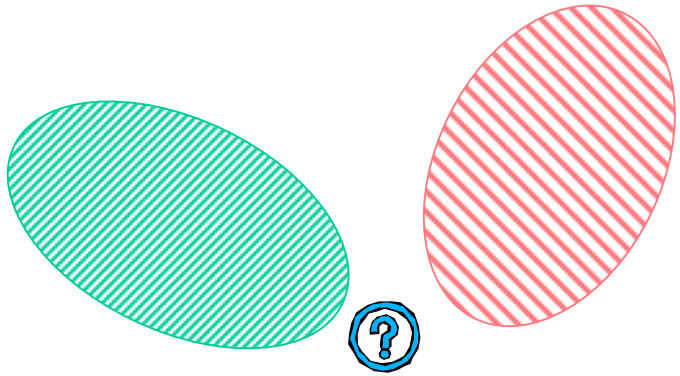
3Z Informatics, LLC

Medical University of South Carolina

BIOWulf Technologies

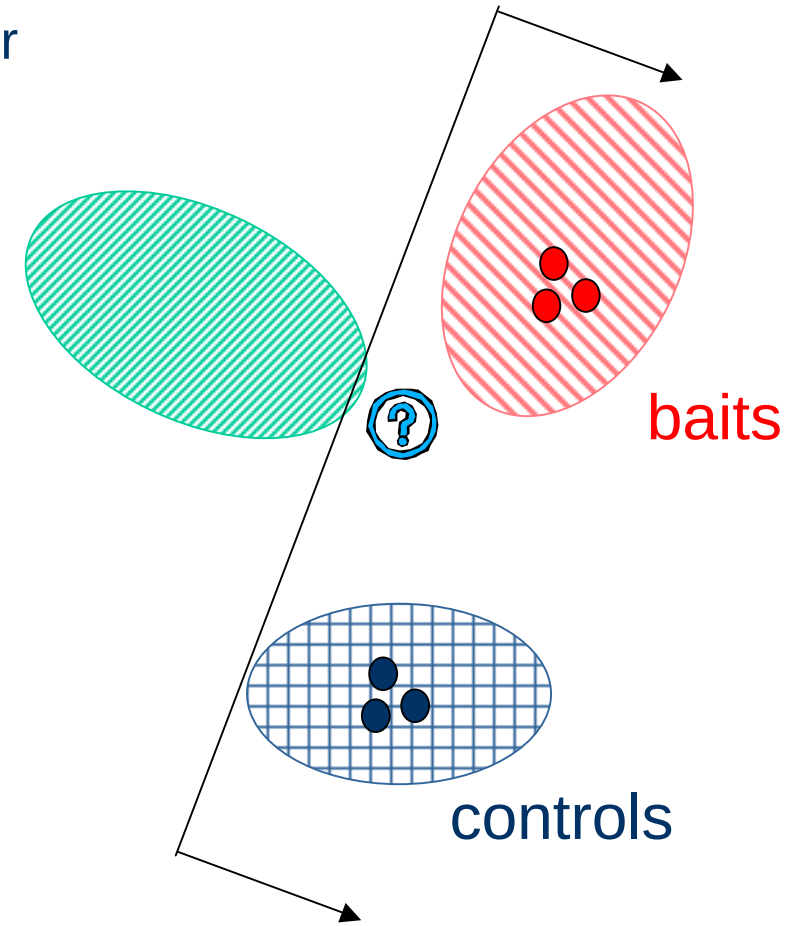
Motivation

- Many genes have multiple molecular functions and are involved in different biological processes;
- Direct application of 2D hierarchical cluster forces a gene to cluster to one of the clusters;
- May result in noisy and scattered patterns for large dataset.



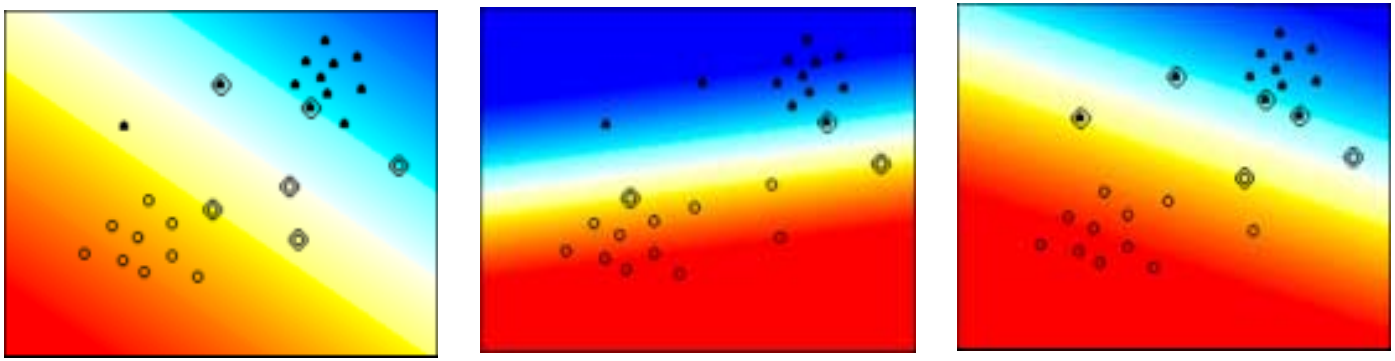
The Idea: Fishing with “Baits”

- Class 1 - the baits: a small number of profiles (or genes) with conditions associated with the molecular functions or biological processes of interest.
- Class 2: control profiles, or the unselected large number of profiles.
- Supervised component analysis methods to find a subset of relevant genes and profiles.
- 2D hierarchical cluster analysis and view to further identify target genes and/or profiles.



Unified Maximum Separability Analysis (UMSA)

- Incorporating data distribution information into the empirical risk minimization algorithm of support vector machine (SVM).
- More efficient use of information from a limited number of samples.
- Adjustable parameters controls the influence of distribution information.



LDA SVM UMSA

A Little Detail of UMSA

Minimize $\frac{1}{2}\nu \cdot \nu + \sum_{i=1}^{m} p_i \xi_i$

Subject to

$$c_i(\nu \cdot x_i + b) \geq 1 - \xi_i, \ i = 1, 2, \ldots, m,$$

$$p_i = K\phi(\delta_i),$$

A typical choice for the function $\phi(\cdot)$ would be

$$\phi(x) = e^{-x^2/\sigma^2}$$

UMSA Component Analysis

- Find a projection vector *d* along which two classes of data are optimally separated for a given set of UMSA parameters.
- Project the data onto a subspace perpendicular to *d*.
- Iteratively, apply UMSA to compute a new projection vector within this subspace, until a desired number of components have been reached.

UMSA Component Analysis vs. PCA/SVD

- Both reduce data dimension.
- PCA/SDV components represent directions along which the data have maximum variations
- UMSA components correspond to directions along which classes of data achieve maximum separation
- PCA/SVD: unsupervised, for data representation;
- UMSA component analysis: supervised, for data classification.


An Example: Extracting Patterns from a Compendium of Expression Profiles

- Reference database of expression profiles of yeast mutants and chemical treatments*.
- Experiments with ≥2 genes up- or down-regulated at ≥3 fold, and p≤ 0.01; and genes up- or down-regulated at ≥3 fold, and p≤ 0.01 in ≥2 experiments.
- 136 profiles and 551 ORFs selected from the original data of 300 experiment profiles and 6298 ORFs
- plus Profiles of 63 negative controls.

* Hughes T.R. et. al. Functional Discovery via a Compendium of Expression Profiles. Cell, 102 (July 2000), 109-126.

An Example: UMSA Component Analysis

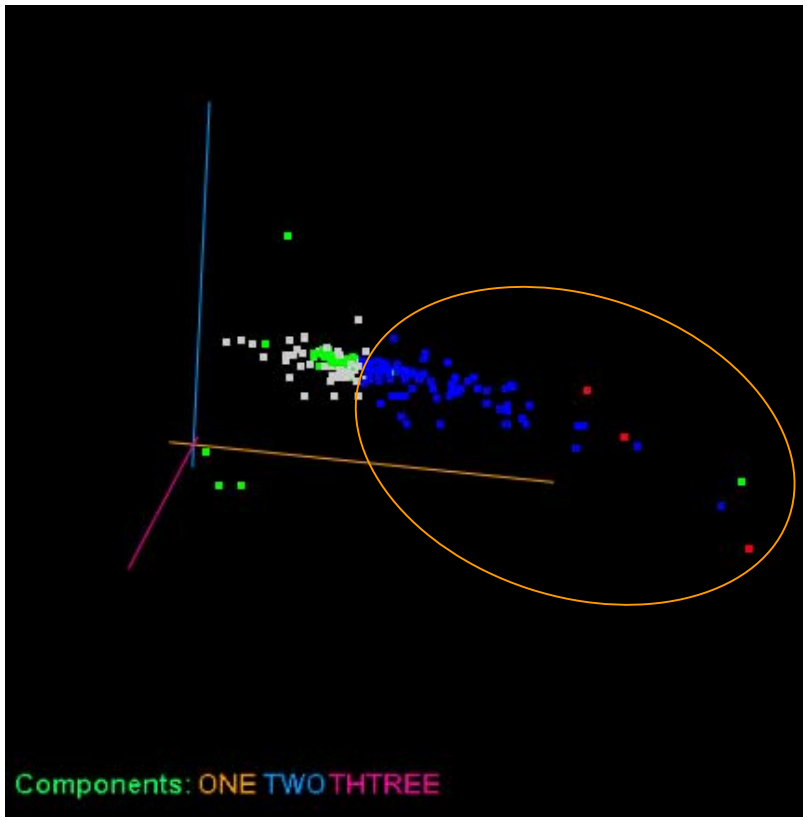
- Class 1 (“baits”) :Mutants erg2Δ, and erg3Δ, and tet-ERG11;
- Class 2: 63 negative controls.
- UMSA component analysis parameter s=10.0 and K=5.0.
- Results: 78 profiles and 200 genes were selected.

CrossView: a Software Package Implements UMSA

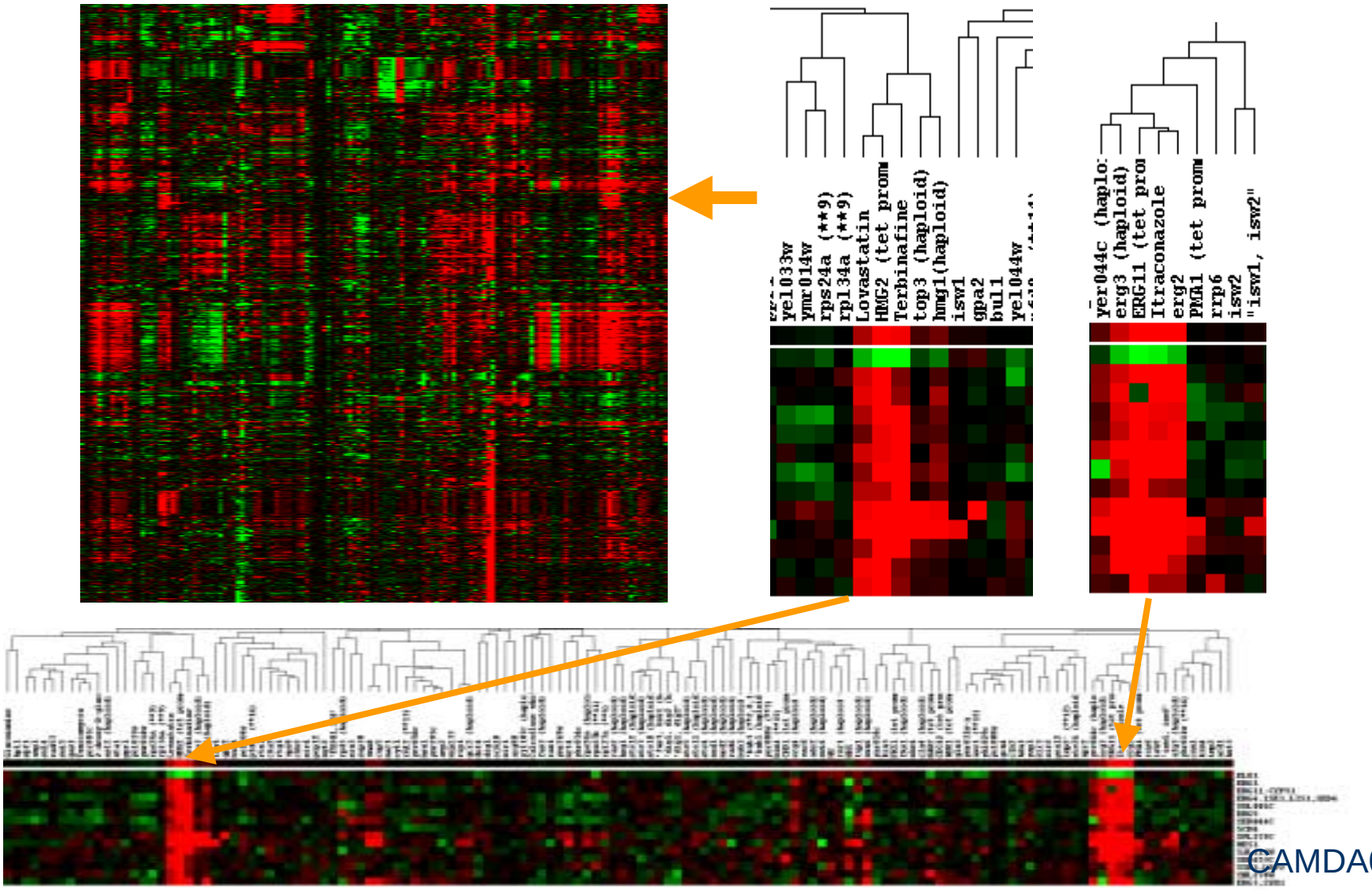
CAMDA01

Selection of Genes and Profiles

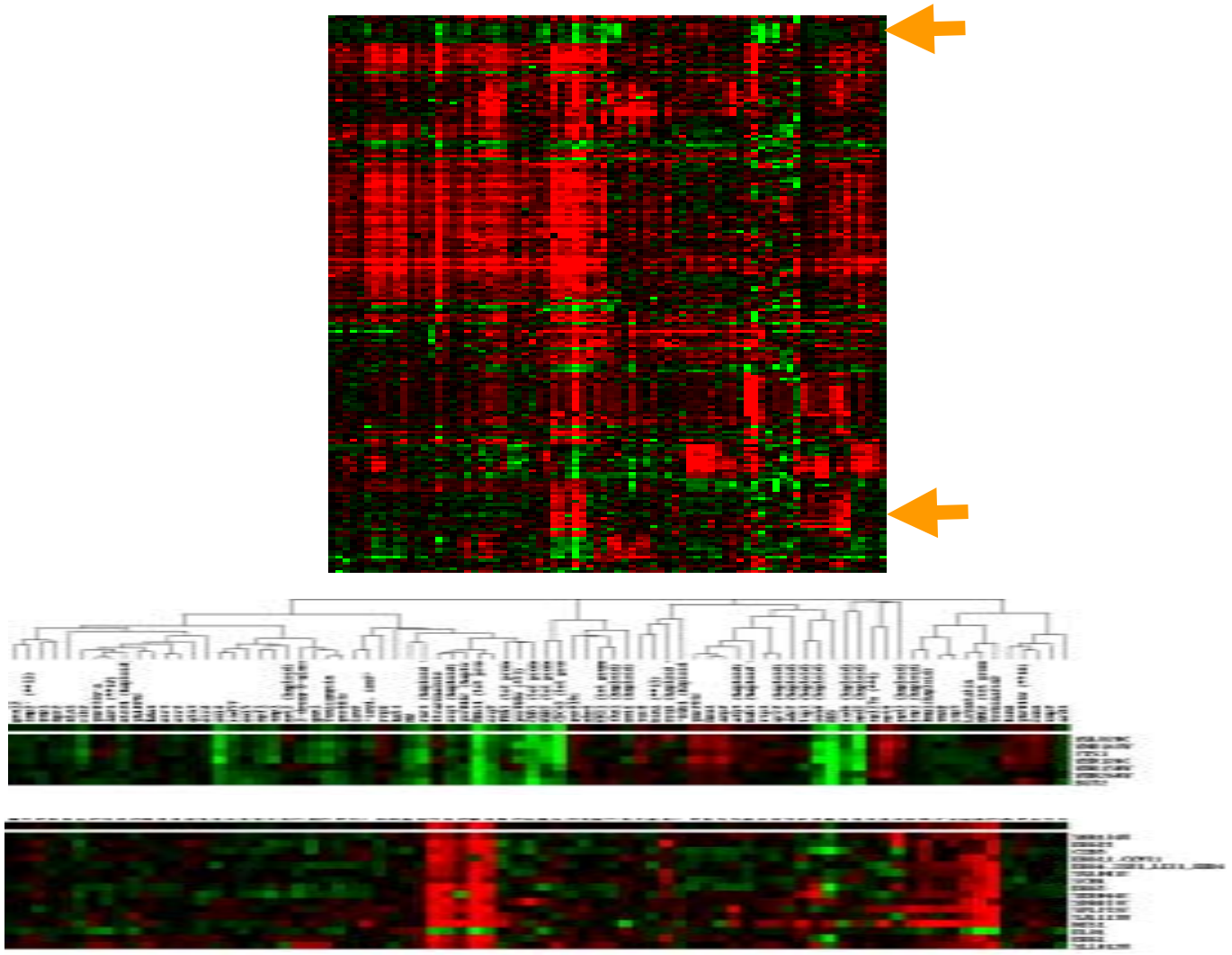
seq	Name	C1	C2	C3	ster)	CDC42 (tet promoter)	ERG11 (tet promoter)
505	YOR237W	21.44	14.28	17.21		0.037	1.397
211	YGR290W	18.75	11.96	24.7		-0.041	0.02
402	YMR096W	14.65	-0.2	-1.31		0.12	1.456
550	YPR192W	14.49	27.23	0.2		0.079	0.044
401	YMR095C	14.2	0.04	-2.14		-0.085	1.384
537	YPL272C	13.44	-2.38	4.21		0.088	1.627
67	YDL085W	13.37	23.83	6.19		-0.175	0.295
154	YFL014W	12.71	0.91	6.41		0.32	0.855
352	YLR042C	12.6	19.73	9.67		-0.078	-0.221
329	YKL178C	11.36	22.48	-0.21		-0.085	0.034
169	YGL117W	11.35	-0.74	-0.71		-0.152	1.15
340	YKR091W	11.14	11.41	13.03		0.585	0.476
152	YER175C	10.22	0.58	5.2		0.049	0.797
421	YMR251W	10.13	16.11	8.84		0.099	0.536
224	YHR029C	10.0	0.18	-1.15		0.059	0.83
400	YMR094W	9.89	0.51	-2.86		-0.037	0.916
303	YJR109C	9.52	0.64	-0.81		-0.04	0.92
38	YBR296C	9.47	-1.8	6.38		0.02	1.255
455	YNR044W	9.43	11.63	13.91		-0.274	-0.665
406	YMR107W	9.29	7.86	12.37		0.186	0.521
17	YBR047W	9.16	0.14	-3.13		-0.027	0.766
222	YHR018C	9.08	-1.25	-1.31		0.072	0.943
523	YOR394W	8.99	12.29	15.91		0.125	0.644
391	YMR015C	8.98	-0.8	-3.16		-0.122	0.891
431	YMR322C	8.83	36.53	-4.92		-0.07	0.472
136	YER069W	8.34	-2.22	-2.12		0.402	0.853
286	YJL213W	8.29	-1.01	-0.8		0.117	0.89
477	YOL118C	8.27	-0.76	4.27		-0.233	0.986
231	YHR137W	8.15	-1.57	-0.33		0.02	0.948



2D Hierarchical Cluster of All Data



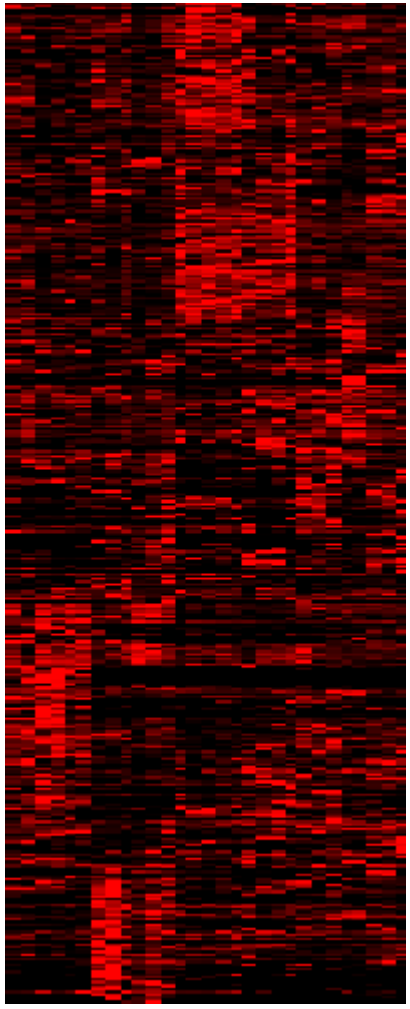
2D Hierarchical Cluster of Selected Data



Comparison of ORFs Identified

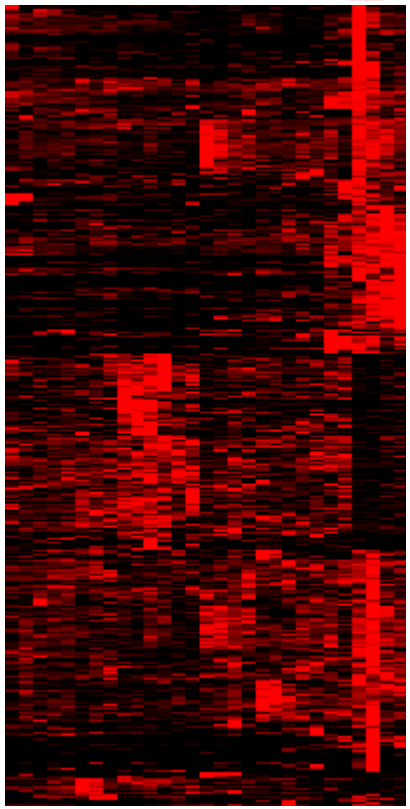
ORFs	Large Set	Reduced Set
YDR453C	*	*
YER044C	*	*
YGL001C	*	*
SCM4/YGR049W	*	*
ERG25/YGR060W		*
ERG1/YGR175C	*	*
ERG11/YHR007C	*	*
YJL113W	*	*
ELO1/YJL196C	*	*
YSR3/YKR053C	*	
ERG3,SYR1/YLR056W	*	
YLL012W		*
ERG6//YML008C	*	*
ERG5/YMR015C	*	*
YNL278W	*	
YMR134W		*
CYB5/YNL111C		*
HES1/YOR237W	*	*
YPL272C	*	*
* ORF identified.		

A Different Example



4000+ genes
After clustering

Tissue Specific Tumor



400 Selected Genes (Tissue Specific) after clustering

Conclusions

- Analysis of large database requires careful balance between efficiency through data reduction and minimizing the risk of losing useful information.
- Using a supervised method, known properties of experiments and genes are incorporated into the selection process to improve the effectiveness and efficiency of pattern matching and detection.
- Most useful for "fishing out" unknown relationships amongst genes and profiles that have something in common with the pre-selected "bait" profiles or genes.