

Higher Bandwidths through the Bottleneck

Semi-automated Approaches to the Interpretation of Gene Expression Arrays

Steve Edwards • Hagit Shatkay
Mark Boguski • John Wilbur

Agilent
2 August 2000

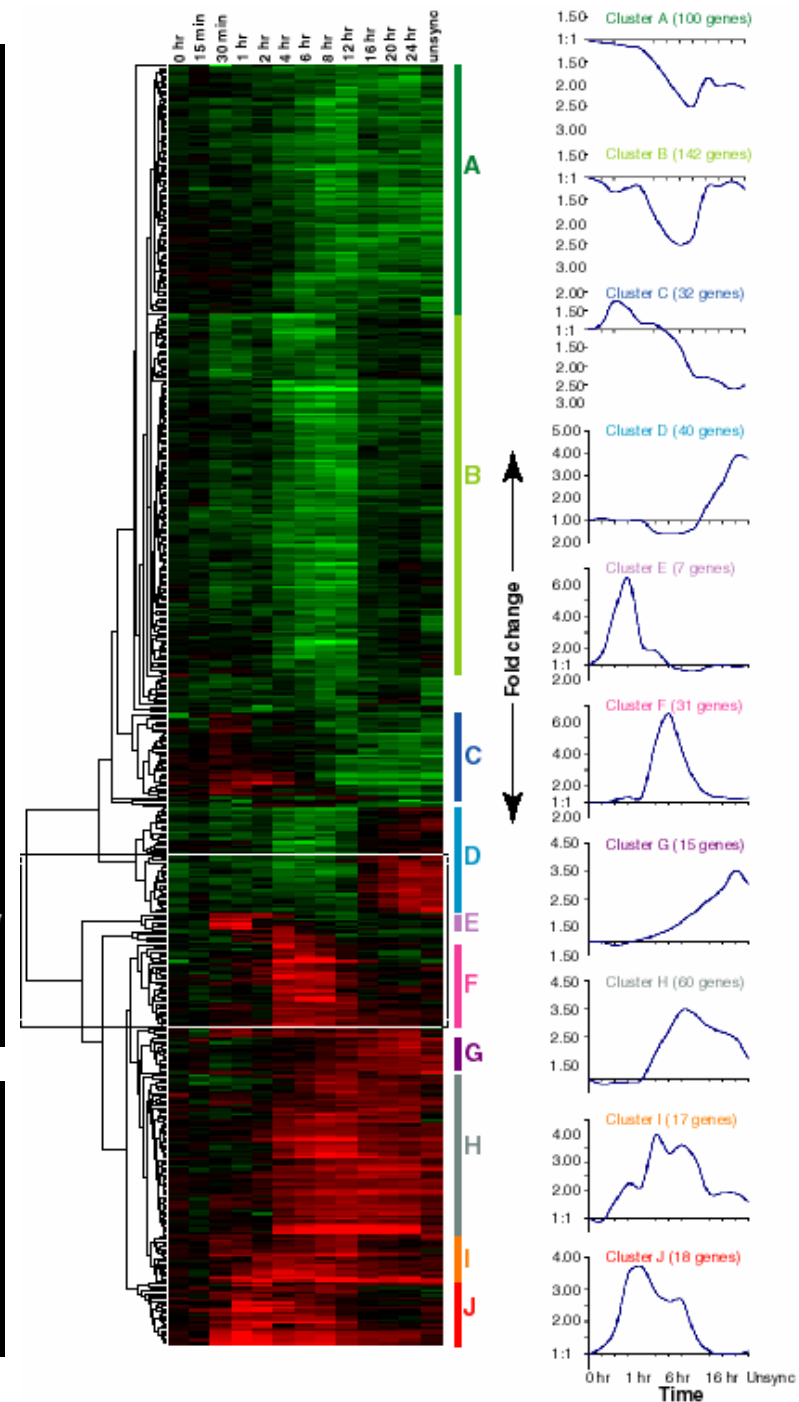
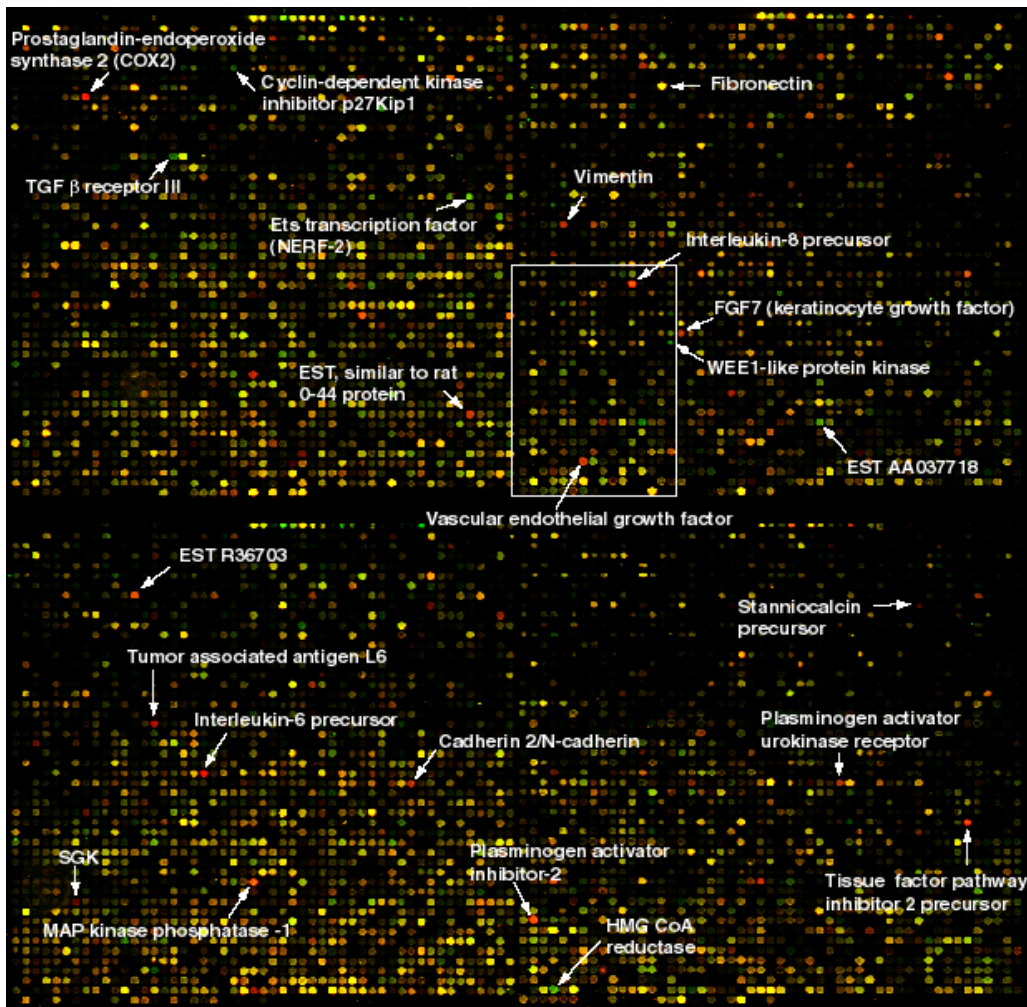
What is currently rate-limiting in gene expression profile analysis?

NCBI

- One can do an experiment (hybridization) overnight to generate data
- One can cluster the data in a few hours to generate information
- However it is still a manual, time-consuming process to transform information into knowledge and insights

Array Interpretation

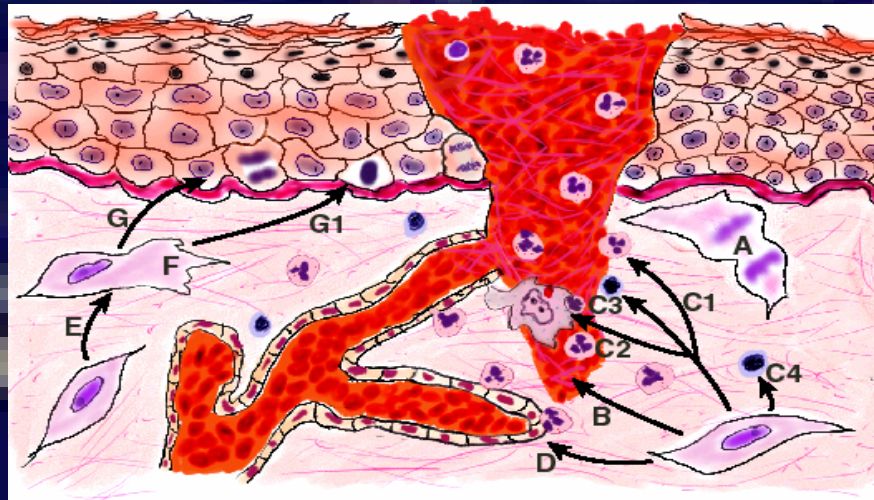
- **The “Materials & Methods” for doing this are rarely documented**
- **Procedures mainly consist of manual explorations of the literature and annotation in genome databases, one gene at a time**
- **At some point in the process, the biologist has enough information to tell a story**



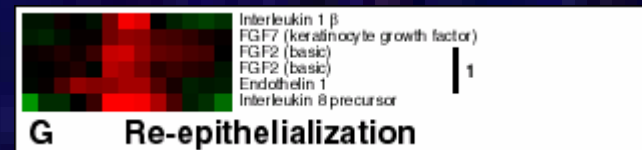
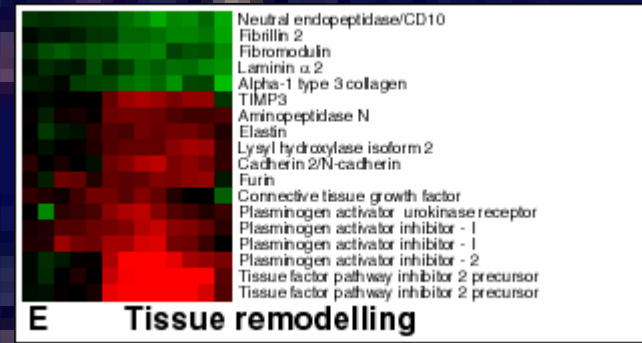
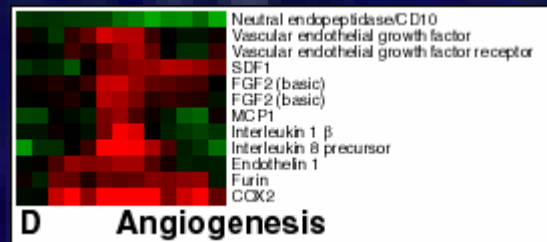
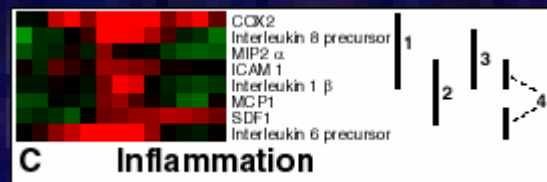
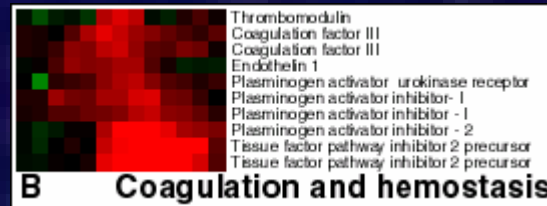
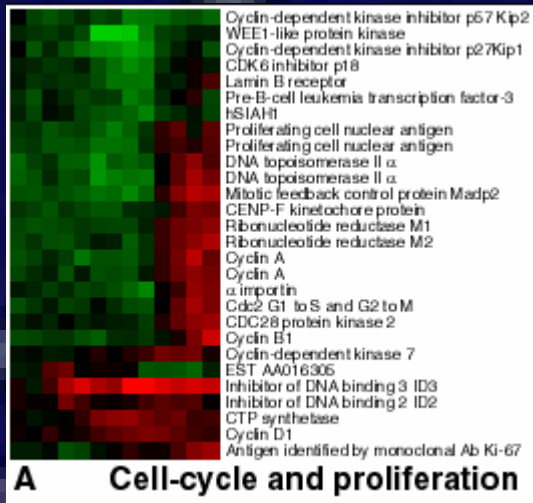
V. R. Iyer et al. (1999) The Transcriptional Program in the Response of Human Fibroblasts to Serum. *Science* 283:83-7

In vitro simulation of wound healing

Schematic representation of the wound repair pathway



Microarray analysis was used to monitor wound repair genes in cultured fibroblasts



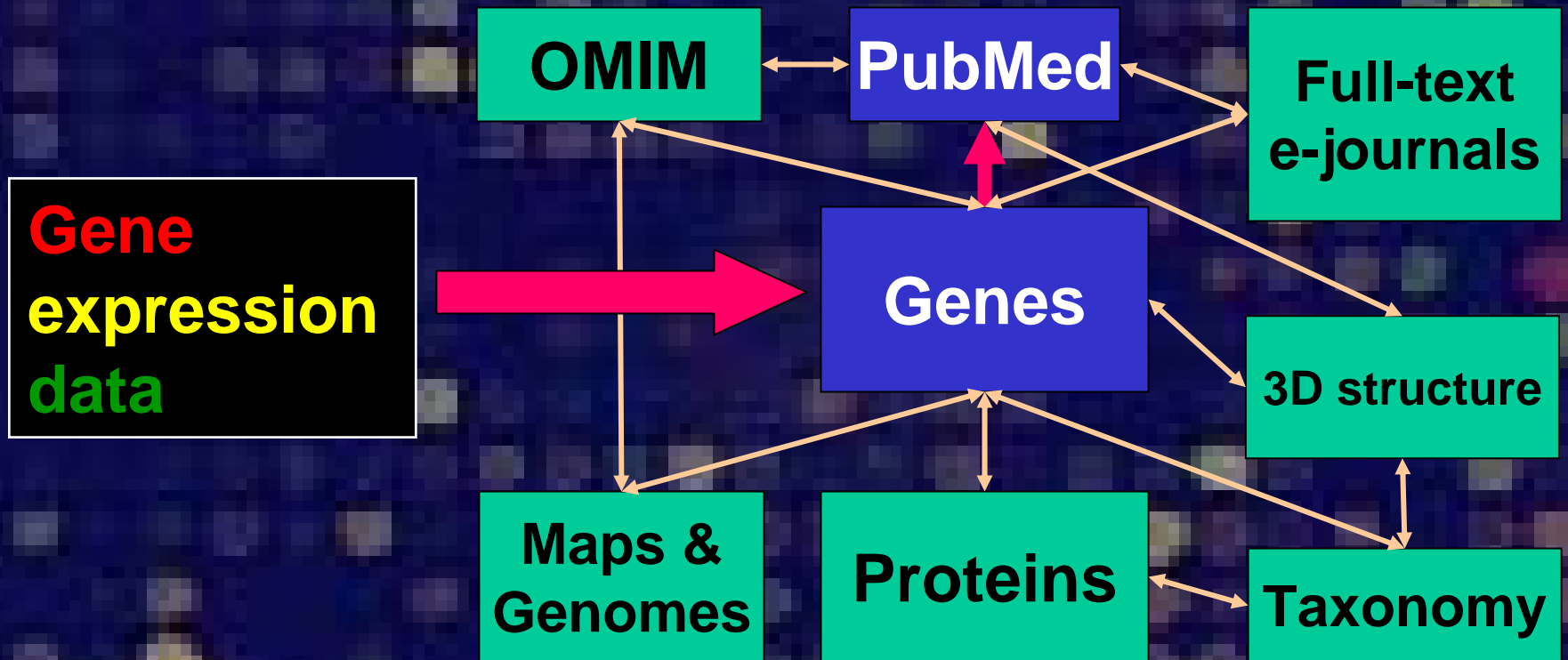
The Automated “Executive Summary”

NCBI

- Initially based on NCBI’s *Entrez* information retrieval system
- Automated pre-processing of expression data to produce first-pass summaries of significant or recurring themes

Connecting Gene Expression Data with *Entrez*

NCBI



Bookmarks Location: http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=m_s What's Related

Details Search Clear

Docs Per Page: 20 Entrez Date limit: No Limit

5 citations found

Display for the articles selected (default all).

Order documents on this page through Loansome Doc

- [McBride HJ, et al.](#) [\[See Related Articles\]](#)
Distinct regions of the Swi5 and Ace2 transcription factors are required for specific gene activation.
J Biol Chem. 1999 Jul 23;274(30):21029-36.
PMID: 10409653; UI: 99340035.
- [Lee J, et al.](#) [\[See Related Articles\]](#)
Interaction of yeast Rvs167 and Pho85 cyclin-dependent kinase complexes may link the cell cycle to the actin cytoskeleton.
Curr Biol. 1998 Dec 3;8(24):1310-21.
PMID: 9843683; UI: 99062011.
- [Tennyson CN, et al.](#) [\[See Related Articles\]](#)
A role for the Pcl9-Pho85 cyclin-cdk complex at the M/G1 boundary in *Saccharomyces cerevisiae*.
Mol Microbiol. 1998 Apr;28(1):69-79.
PMID: 9593297; UI: 98254129.
- [Aerne BL, et al.](#) [\[See Related Articles\]](#)
Swi5 controls a novel wave of cyclin synthesis in late mitosis.
Mol Biol Cell. 1998 Apr;9(4):945-56.
PMID: 9529390; UI: 98198430.

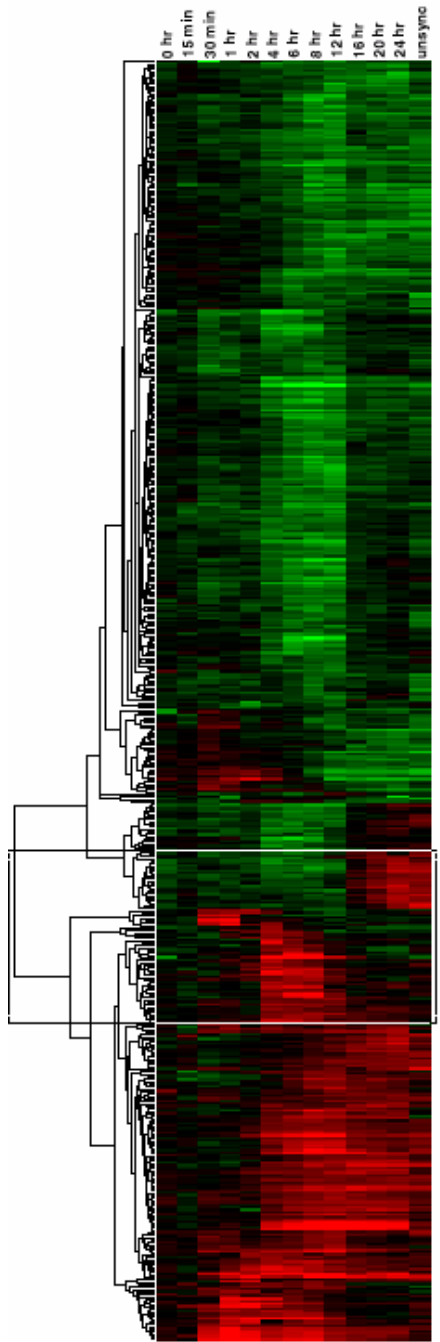
Document: Done



Query term: **PCL9**

PMIDs retrieved
initial search:

10409653
9843683
9593297
9529390
9032248



Iterative searches of PubMed to convergence



Iterative searches of PubMed to convergence



AutoSummary of key features of cluster A

AutoSummary of key features of cluster B

Summary C

Summary D

Summary E

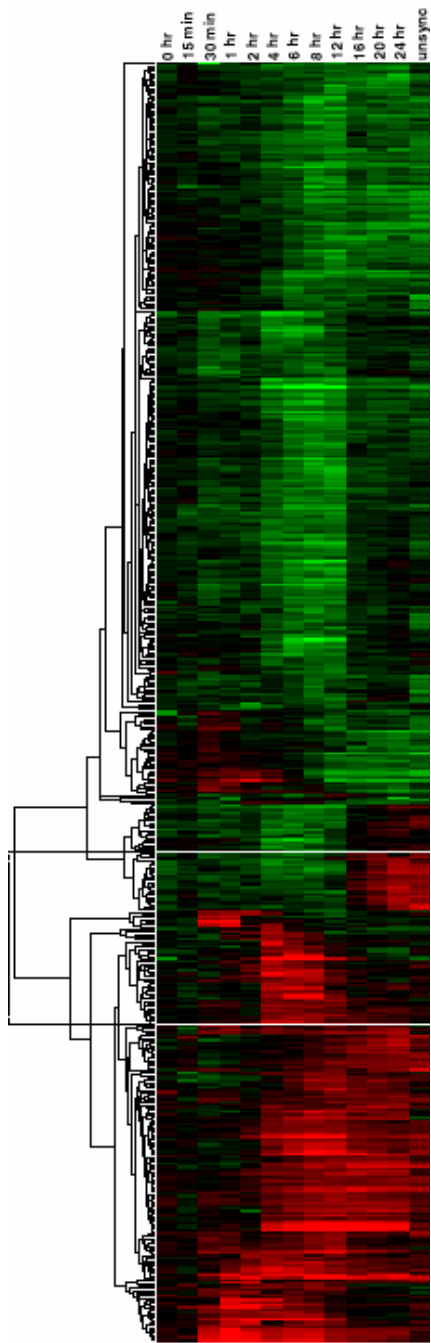
Summary G

Summary H

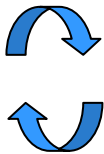
Summary I

Summary J

Executive Summary



**PMIDs for all over-
and under-
expressed genes**



**Iterative searches of
PubMed to convergence**

Then cluster documents
rather than gene
expression levels.

Summary D

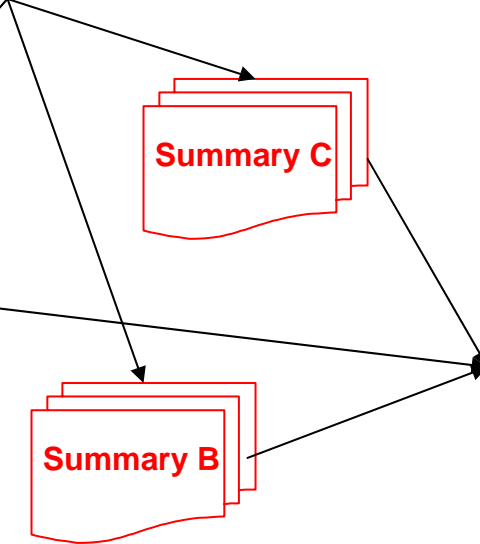
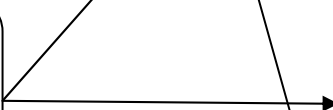
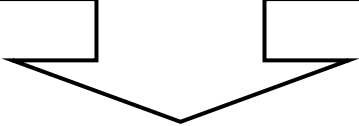
Summary E

Summary C

Summary A

Summary B

**Executive
Summary**



Entrez Text Neighboring

NCBI

Genetic Analysis
of Cancer in
Families

The Genetic
Predisposition to
Cancer

- Vector cosine method
- Common terms could indicate similar subject matter
- Weights based on term frequencies within document and within the database as a whole
- Some terms are better than others

Informative Document Features

NCBI

- **MeSH terms**
- **Title terms**
- **Abstract terms**
- **Full-text terms (not yet available)**
- **Bibliography (not yet available)**
- ****Other fields probably not useful***

Challenges with the Analysis of “Document Space”

NCBI

- Language, even in technical literature, is imprecise and full of ambiguity, nuance and shades of meaning
- Does professional indexing help?
 - Studies of MEDLINE have shown that indexer consistency is only ~34%

Indexer (in)consistency

- Funk & Reid (1983) *Bull. Med. Libr. Assoc.* 71(2):176-183

$$CP\% = \frac{100A}{A+M+N}$$

A= number of terms in agreement

M= number of terms used by indexer M, but not N

N= number of terms used by indexer N, but not M

What is a good metric to judge how well we are doing?

NCBI

- Compare results of automated process with published human interpretation
- Compare results of automated process with curated annotation in a genome database

Yeast Model System

NCBI

- **Advantages**

- Many published, available data sets
- Curated references and descriptions of biology for each gene in the **SGD**

- **Disadvantages**

- Results may not extrapolate well to mammalian systems
- Many yeast genes are contained in a single GenBank record making the associated reference paper uninformative for neighboring

P.T. Spellman et al. (1999) Comprehensive Identification of Cell Cycle-regulated Gene of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization. *Mol. Cell. Biol.* 9:3273-3297

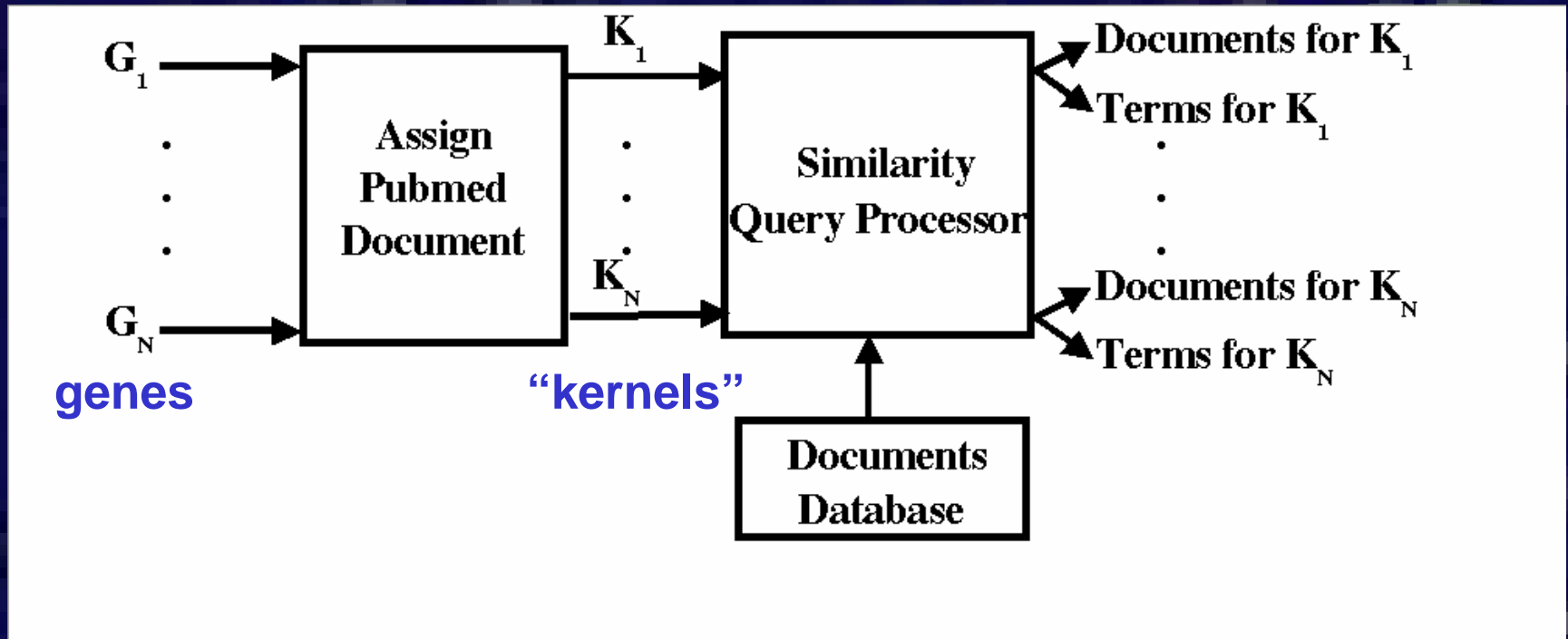
Figure 7		G1			S	G2		M		M/G1	
DNA repair		DHS1	PMS1	RAD54	HPR5	MEC3		ALK1			
		DUN1	RAD27	RDH54							
		MSH2	RAD5	RHC18							
		MSH6	RAD51	UNG1							
		OGG1	RAD53								
DNA Syn		CDC2	POL12	RFC4	TOF2						
		CDC9	POL30	RFC5							
		CTF18	HYS2	TEL2							
		CTF4	POL32	TOF1							
		DPB2	PRI2	TOP3							
		EST1	RFA1	YNK1							
		POL1	RFA2								
		POL2	RFA3								
Replication Init.		CDC45				ORC1		CDC47	MCM2	CDC6	MCM3
								CDC54	MCM6	CDC46	
Chromatin		ASF1	MIF2		ADA2	HTA1	RAP1	HST3		HST4	
		ASF2	RLF2		HHF1	HTA2	SAS3	WTM2		WTM1	
		CAC2	SPT16		HHF2	HTA3	TBF1				
		CBF2			HHO1	HTB1					
		ESC4			HHT1	HTB2					
		HIF1			HHT2						
Nucleotide Syn.		CDC21	RNR1	RNR3							
Budding	Site Selection/ Morphogenesis	BNI4	GIN4	SPH1	DFG5	BUD3		BEM1		RGA1	
		BUD9	MCD4	SRO4	GIC1	MSB4		BUD4			
		CDC10	MSB2		MSB1			BUD8			
		GIC2	RSR1								
	Glycosylation	MNN1	PMT3	QRI1	GDA1	ALG7					
	OCH1	PMT5	SVS1	GOG5							
	PMT1	PSA1	SSO1	PMI40							
Secretion		EMP24	SLY41		ERV25	SSO2				GYP6	
		SEC28	UFE1								
Cell Wall Synthesis		CWH41	GAS1		ECM17	KRE6	CHS6	CHS2	WSC4	CHS1	TIP1
		EXG1			ECM25	WSC2	CWP1	SED1		GFA1	YGP1

Summary of Spellman Data & Our Goals

NCBI

- 800 genes found to be cell-cycle regulated
- Only 408 had curated references in SGD
- Some genes shared the same reference
- We found 344 distinct documents (“kernels”) on which to test our methods
- The goal is to search a large document database with these kernels and, for each gene, return:
 - A set of related documents
 - A set of summarizing terms (a theme)

Similarity Queries over Document Space



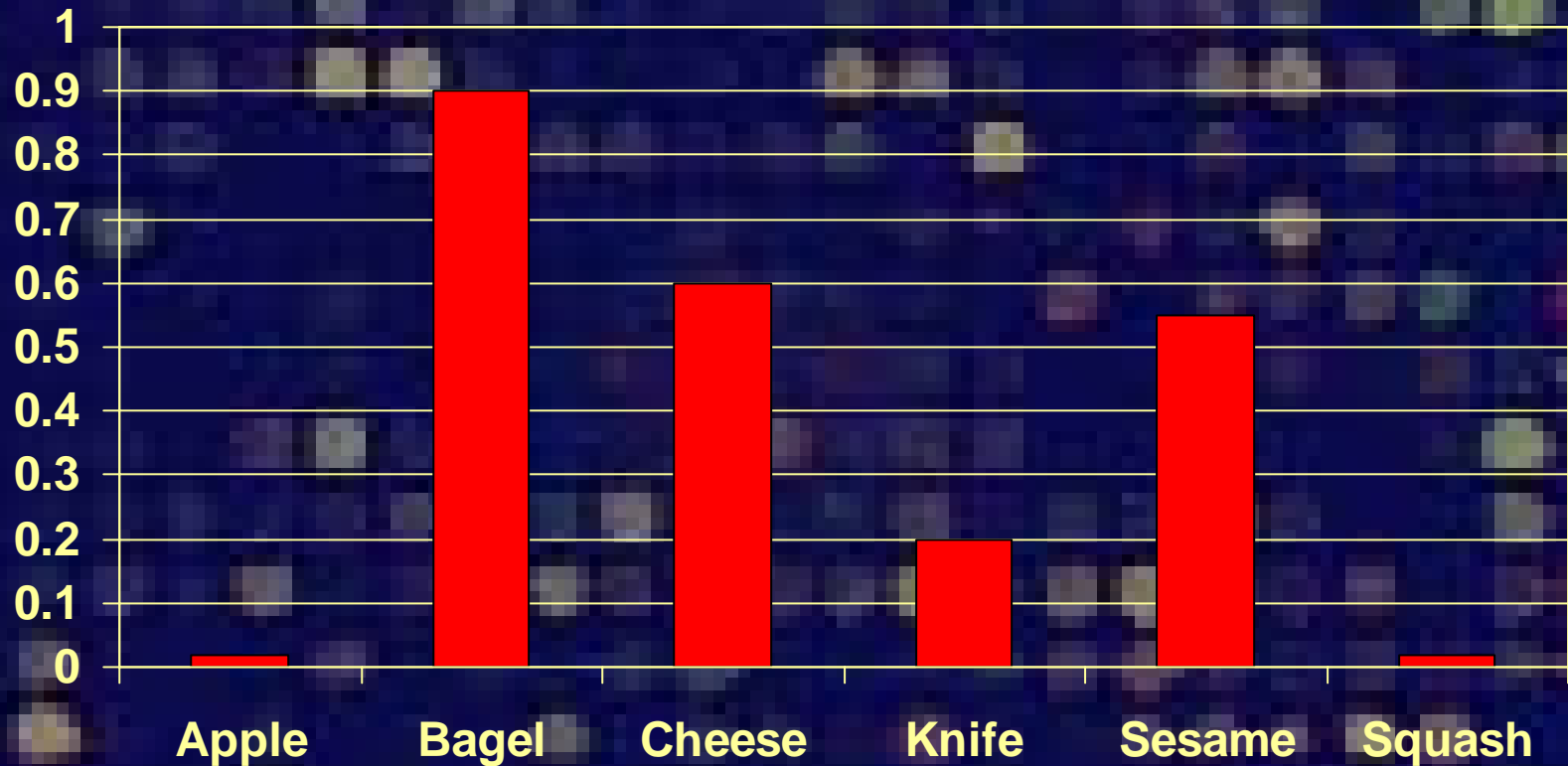
Similarity Query Processor (GeneTheme algorithm)

NCBI

- Starts by generating a rough approximation of the distribution based on the kernel document
- Uses an EM algorithm to iteratively rank the documents based on the current distribution
- Then generates a new distribution that maximizes the likelihood of the database partition into theme and off-theme documents

Term Distribution in “Bagel” Documents

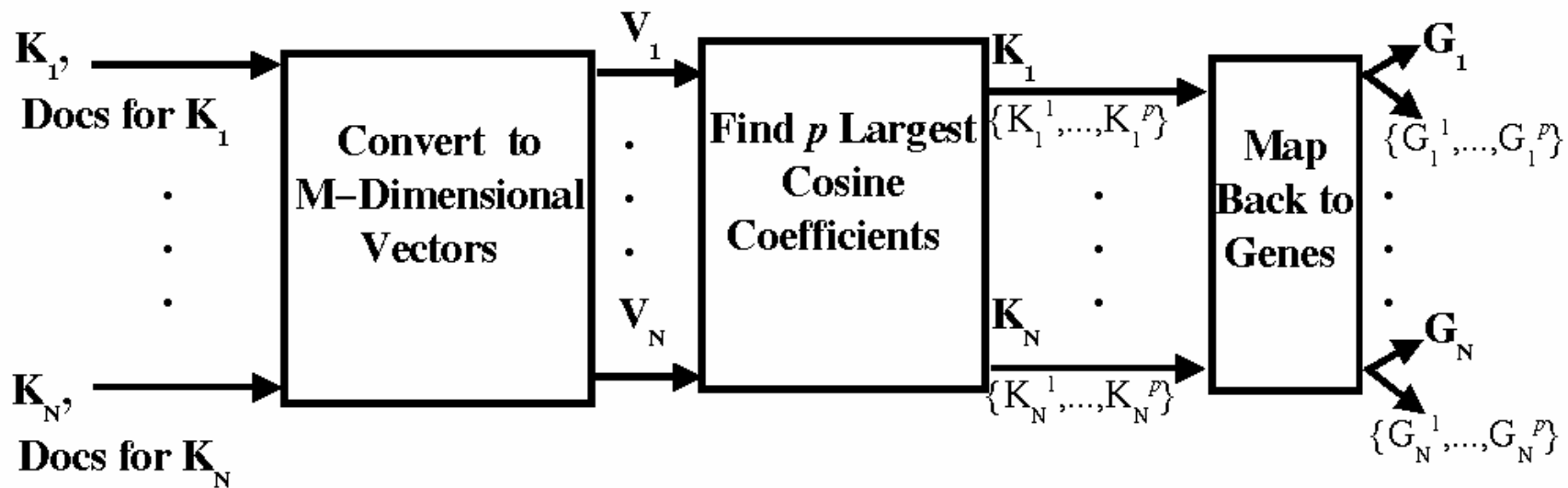
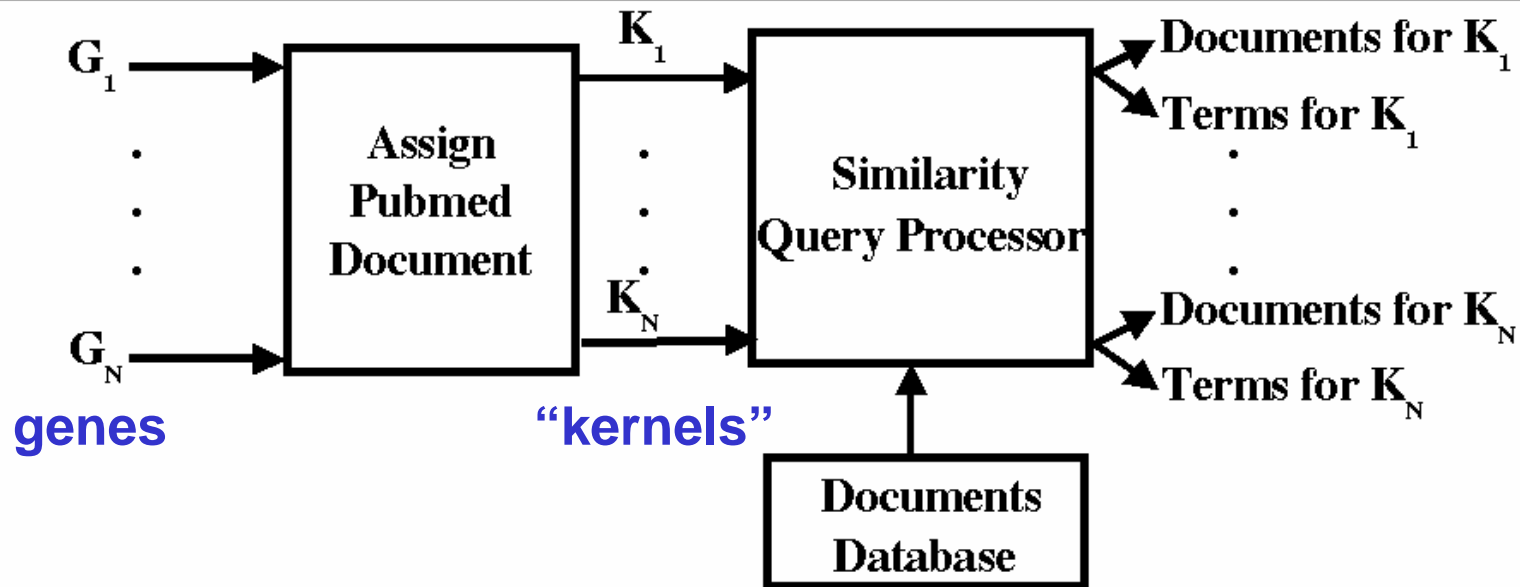
NCBI



The Output of GeneTheme

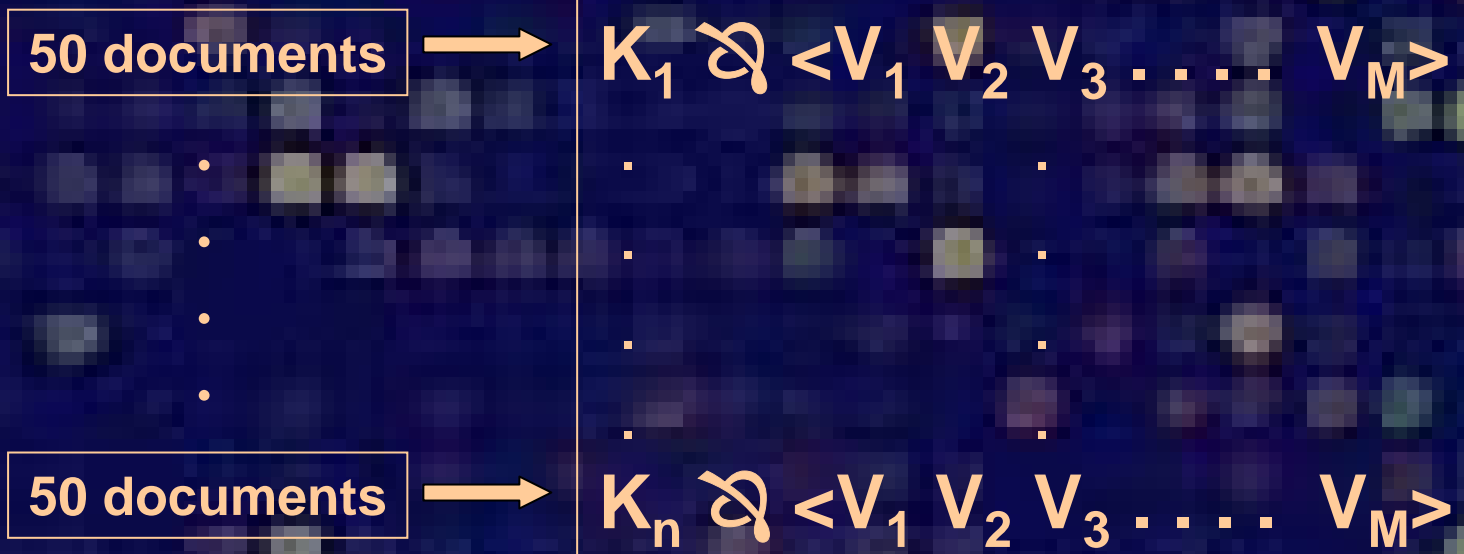
NCBI

- A list of the top 50 documents discussing the same theme as the kernel document, ordered by their degree of relevance to the theme
- A list of terms (“executive summary”) summarizing the theme
- Next: How do we find functional relationships among genes?



Finding Functional Relations among Genes

NCBI



1. "Unionize" to obtain dimension of vector

2. Convert each kernel to a vector of PMIDs

3. Use vector cosine method to find functional relationships

Kernal = PMID 8702485 (Gene ELO1)

Keywords	Genes	Function
Fatty acid,	OLE1	Fatty Acid, Sterol Metabolism*
Fatty,	FAA4	FA/Lipids/Sterols/Membranes
Lipids,	FAA3	FA/Lipids/Sterols/Membranes
Acid,	SUR2	FA/Lipids/Sterols/Membranes
Carbon,	ERG2	FA/Lipids/Sterols/Membranes
Growth,	FAA1	FA/Lipids/Sterols/Membranes
Medium,	ERG2	FA/Lipids/Sterols/Membranes
Synthase,	PSD1	FA/Lipids/Sterols/Membranes
Strains,	CYB5	Fatty Acid, Sterol Metabolism*
Deficient	PGM1	Carbohydrate Metabolism*

Kernal = PMID 7651133 (Gene HXT7)

Keywords	Genes	Function
Hexose, Glucose, Fructose, Uptake, Glycolytic, Sugars, Aerobic, Synthase	HXT1	Nutrition
	RGT2	Nutrition
	HXT4	Nutrition
	HXT2	Nutrition
	GLK1	Nutrition
	SEO1	Small molecule transport*
	PRB1	Protein degradation*
	AGP1	Nutrition
	ZRT1	Nutrition
	MIG2	Carbohydrate metabolism

Kernal = PMID 6323245 (Gene MCM2)

Keywords	Genes	Function
Ars,	CDC10	Site selection, Morphogenesis
Autonomous,	PHO3	Nutrition
Replicating,	EST1	DNA Synthesis
Minichrom.,	MIF2	Chromatin
Centromeric,	PHO12	Nutrition
leu2,	POL3	DNA Synthesis
Plasmids	DHS1	DNA Repair
ura3	SNQ2	DNA Repair
	SMC3	Chromatin
	EXG2	Cell Wall Synthesis

Summary of GeneTheme Results

- It is an efficient way for establishing putative functional relationships between genes
- It provides relevant literature needed by the researcher for further evaluation
- It generates a summary explaining the discovered relationships

Ongoing Experiments

NCBI

- **More robust selection of high-quality kernel documents**
 - machine learning techniques
 - human intervention at key points in process
- **Try different metrics**
- **Apply to mammalian data sets**
- **Extend to genes of unknown function**
- **Customized reports**
- **Establish a web server**

Interpreting Gene Expression Data using multiple information spaces

NCBI

