Biomedical Text Analysis

BMI/CS 776
www.biostat.wisc.edu/bmi776/
Mark Craven
craven@biostat.wisc.edu
Spring 2009

Some Important Text-Mining Problems

- hypothesis generation
  Given: biomedical objects/classes of interest (e.g. dieases & dietary factors)
  Do: identify interesting, implied relationships among the objects

- experiment annotation
  Given: a set of genes/proteins exhibiting common behavior in an experiment
  Do: identify commonalities among genes/proteins in the set

- information extraction
  Given: classes, relations of interest
  Do: recognize and extract instances of the classes and relations from documents

Some Important Text-Mining Problems

- document classification
  Given: defined classes of interest
  Do: assign documents to the relevant classes

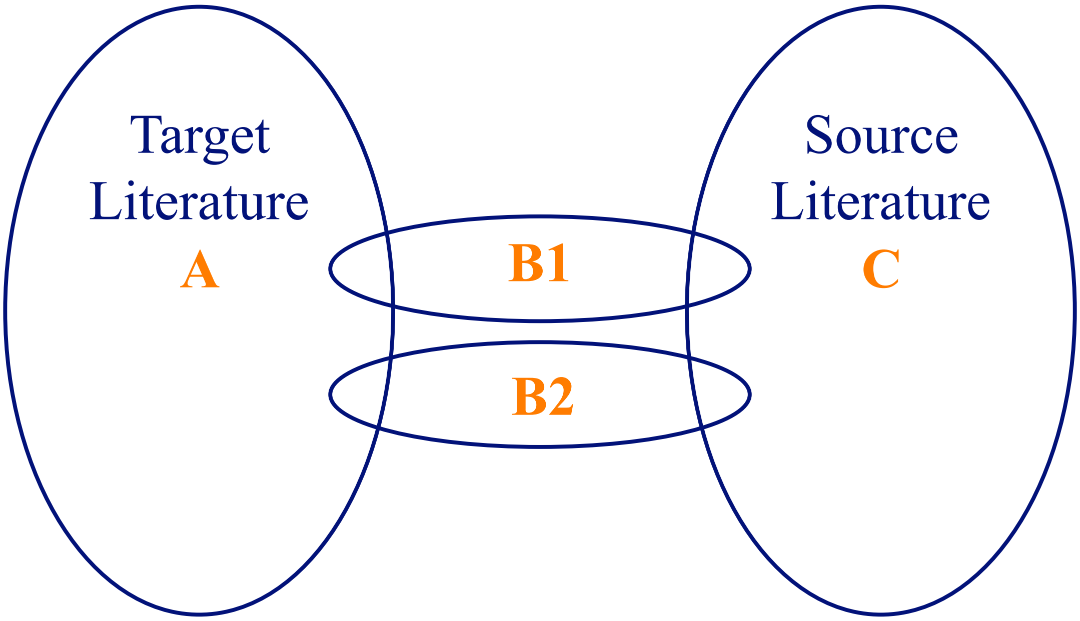
- ad-hoc retrieval
  Given: a query
  Do: return relevant documents/passages

- improving the accuracy of other inference tasks
  - querying with PSI-BLAST [Chang et al.]
  - predicting subcellular localization of proteins[Hoglund et al]
  - etc.

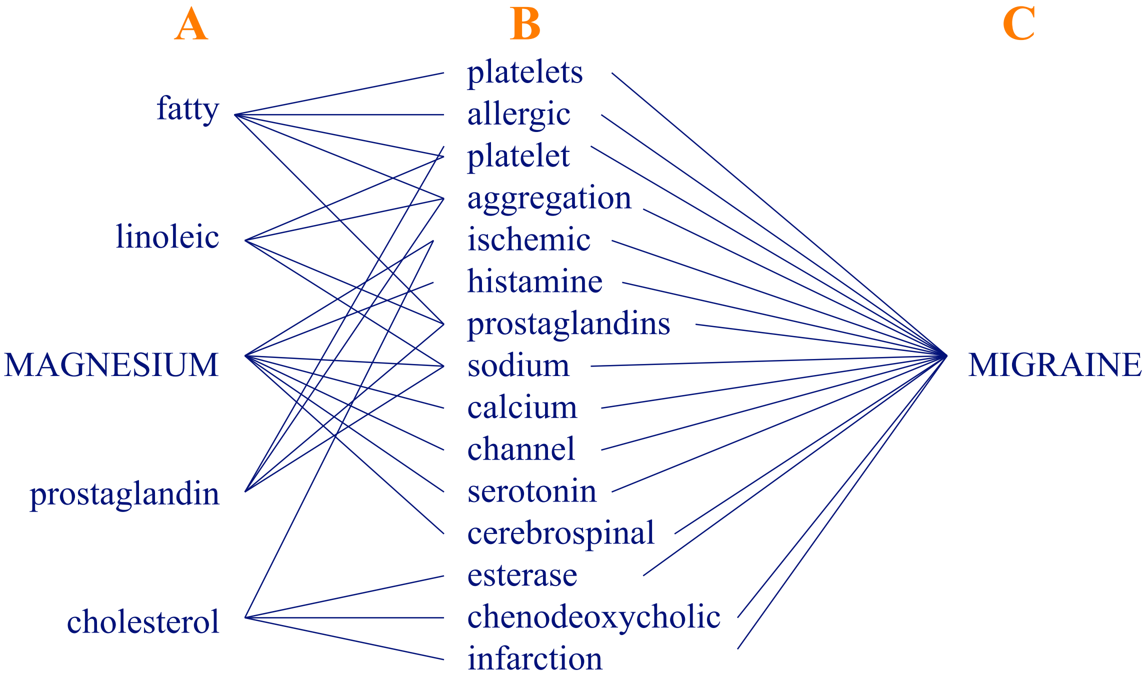
Hypothesis Generation by Finding Complementary Literatures

- Swanson & Smalheiser, *Artificial Intelligence* 91, 1997
- ARROWSMITH aids in identifying relationships that are implicit, but not explicitly described, in the literature
- http://arrowsmith.psych.uic.edu/

ARROWSMITH: Finding Complementary Literatures



ARROWSMITH Example: The Magnesium-Migraine Link



The ARROWSMITH Method

- given: query concept **C** (e.g. *migraine*)
- do:
  - run MEDLINE search on **C**
  - derive a set of words (**B**) from titles of returned articles
  - run MEDLINE search on each **B** word to assemble list of **A** words
  - rank **A-C** linkages by number of different intermediate **B** terms

Restricting the Search in ARROWSMITH

- prune **B** list by
  - using a predefined *stop-list* (“clinical”, “comparative”, “drugs”,…)
  - having a human expert filter terms
- prune **A** list using *category restrictions* (e.g. dietary factors, toxins, etc.)
- prune **C-B**, **B-A** linkages by requiring:

$$P(B \mid C) > P(B)$$

$$P(A \mid B) > P(A)$$

Given a document with word C, do we see B more often than we’d expect by chance?

ARROWSMITH Case Studies

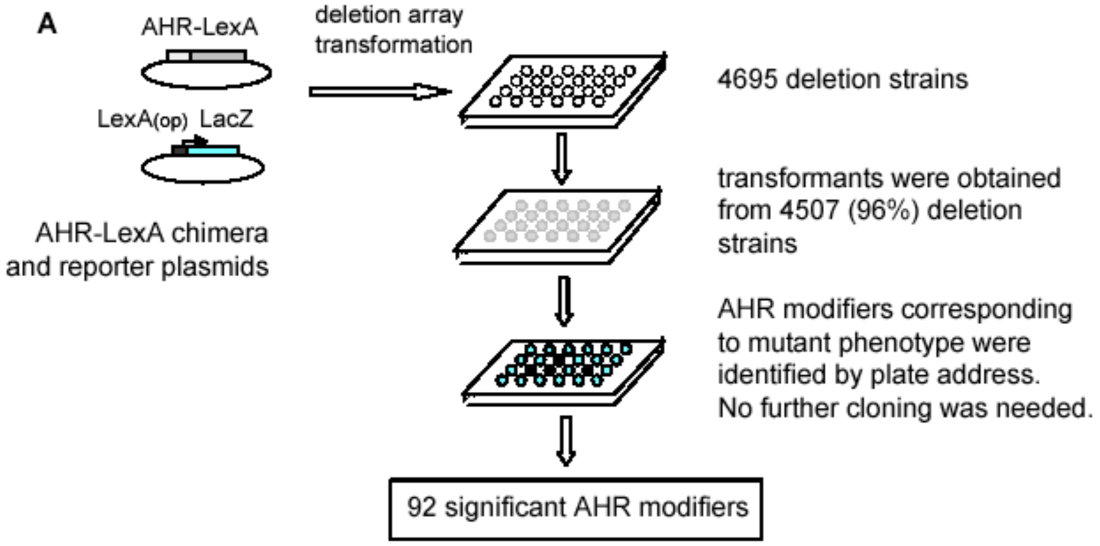
- *indomethacin* and *Alzheimer's disease*
- *estrogen* and *Alzheimer's disease*
- *phospholipases* and *sleep*
- etc.
- has led to hypotheses interesting enough to warrant further studies, peer-reviewed articles

Task: Automatic Annotation of Experiments

- Genes, Themes and Microarrays. Shatkay, Edwards, Wilbur & Boguski. *ISMB* 2000
- given: a set of genes with a "kernel" document for each
- return:
  - top-ranked words in theme for each gene
  - list of most similar genes, in terms of associated documents

High-throughput Experiment Example:

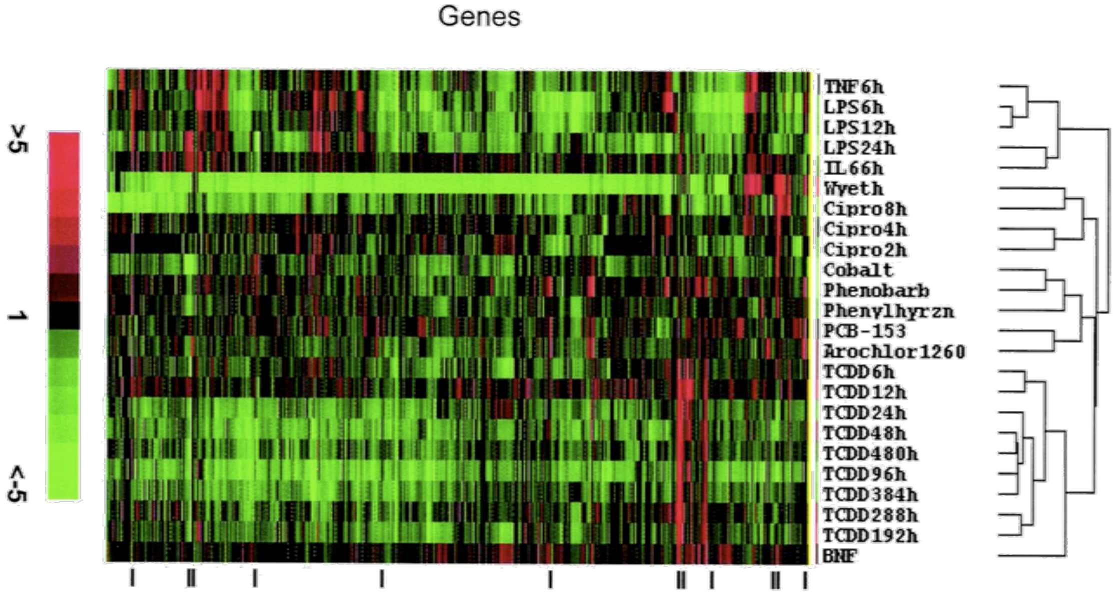
Yao et al., *PLoS Biology* 2004


A
AHR-LexA
LexA(op) LacZ
AHR-LexA chimera and reporter plasmids
deletion array transformation
4695 deletion strains
transformants were obtained from 4507 (96%) deletion strains
AHR modifiers corresponding to mutant phenotype were identified by plate address. No further cloning was needed.
92 significant AHR modifiers

- Experiment identified 92 genes that, when knocked out, modify AHR signaling. What do they have in common?

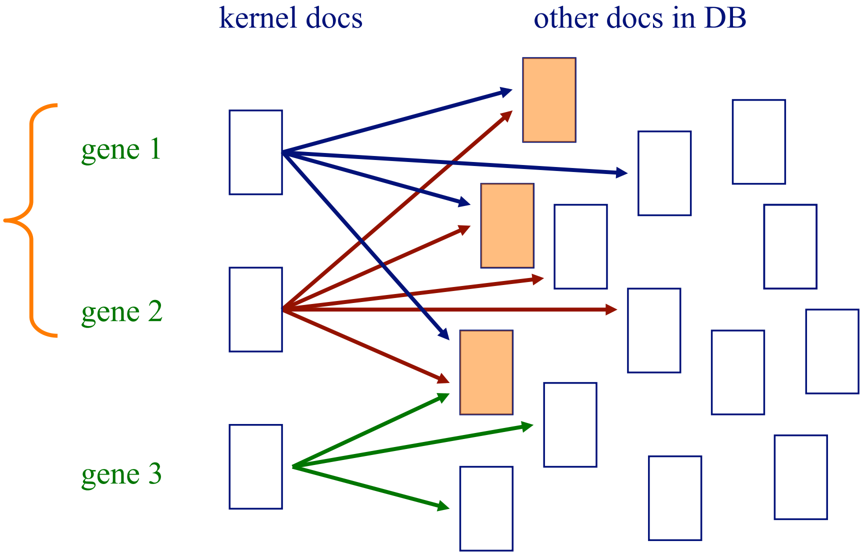
High-throughput Experiment Example:

Thomas et al., *Molecular Pharmacology* 2002



- In initial experiments, a mysterious set of genes that were upregulated in all treatments. What do they have in common?

Shatkay et al. Approach

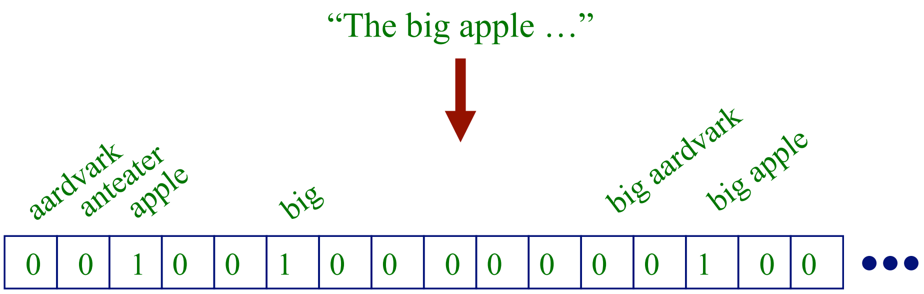


step 1: given kernel documents, find themes

step 2: given themes, find related genes

Representing Documents

- Shatkay et al. represent documents using fixed-length vectors
  - this is a common approach in many text processing systems (e.g. search engines)
- elements in vector represent occurrences of individual words (unigrams) and pairs of adjacent words (bigrams)



Themes

- a *theme*, *T*, is a set of documents discussing a common topic
- the occurrence of a given term t_i in a theme document *d* is represented by

$$p_i^T \equiv P(t_i \in d \mid d \in T)$$

- thus for every term in the vocabulary, we can characterize how likely it is to occur in a document on theme *T*

Theme Example

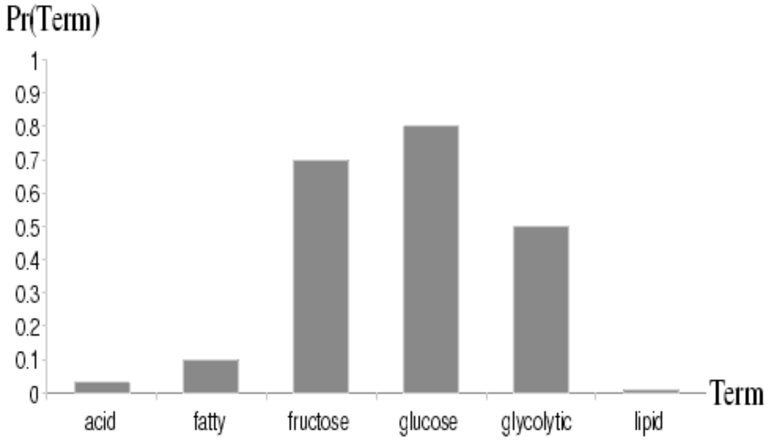


Figure from H. Shatkay et al., *ISMB* 2000

Other Parameters

- Shatkay et al. use similar parameters to represent
  - the occurrence of each term given that document *d* is not in the theme

$$q_i^T \equiv P(t_i \in d \mid d \notin T)$$

  - the occurrence of the term regardless of whether *d* is on-theme or off-theme

$$DB_i \equiv P(t_i \in d \mid d \in DB)$$

  - the prob that a term occurrence, t_i, is best explained by DB probability or by on-theme/off-theme probabilities

$$\lambda_i$$

Model for "Generating" Documents

- we can think of the document vectors as having been generated from a model with these parameters

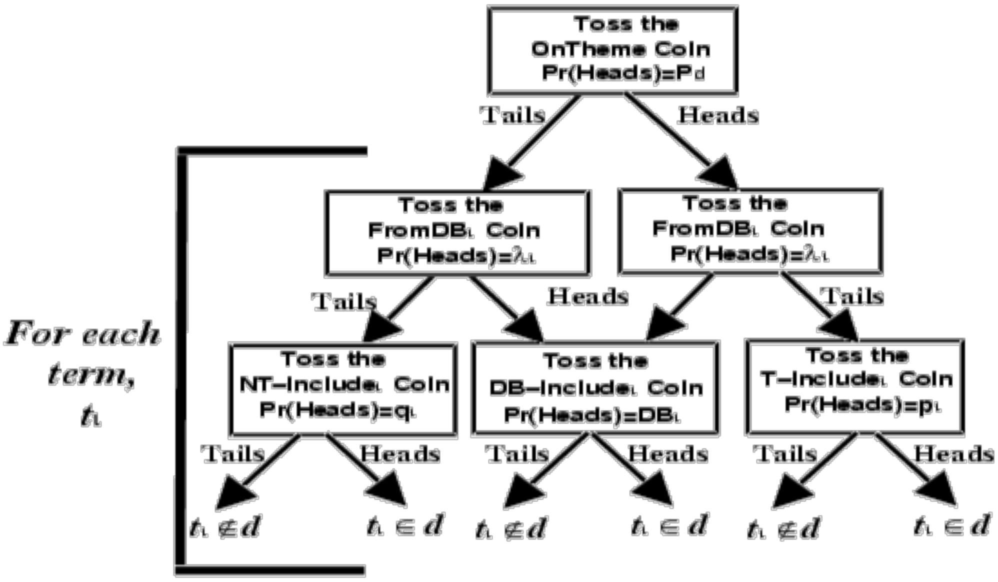

Toss the OnTheme Coin Pr(Heads)=P_d
Tails
Heads
Toss the FromDB_i Coin Pr(Heads)=λ_i
Toss the FromDB_i Coin Pr(Heads)=λ_i
Tails
Heads
Heads
Tails
Toss the NT-Include_i Coin Pr(Heads)=q_i
Toss the DB-Include_i Coin Pr(Heads)=DB_i
Toss the T-Include_i Coin Pr(Heads)=p_i
Tails
Heads
Tails
Heads
Tails
Heads
t_i ∉ d
t_i ∈ d
t_i ∉ d
t_i ∈ d
t_i ∉ d
t_i ∈ d
For each term, t_i

Figure from H. Shatkay et al., *Advances in Digital Libraries* 2000

Finding Themes

- given: a DB of documents and a "kernel" document
- do:
  - determine the parameters characterizing the theme *T*
  - determine the documents belonging to *T*

- if we knew the documents in *T*, it would be easy to determine the parameters
- if we knew the parameters, it would be easy to determine the documents in *T*
- but initially, we don't know either

Finding Themes

- Shatkay et al. solve this problem using EM

E-step: compute likelihood for each document that it's in same theme as kernel

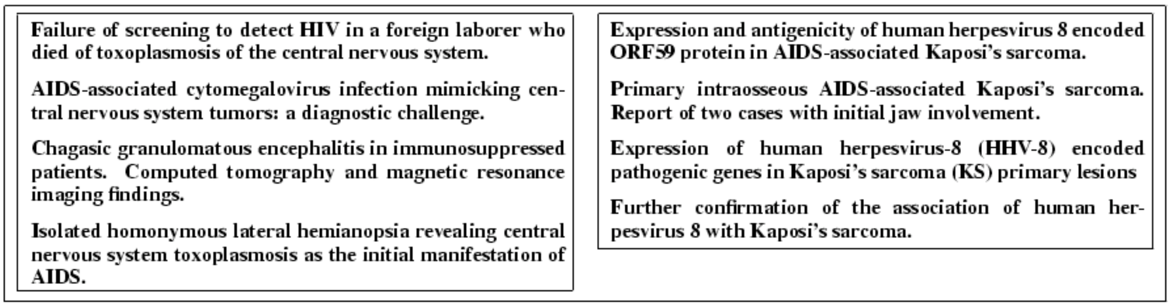
M-step: find new parameters that maximize the likelihood of this partition into theme/off-theme documents

Finding Themes: Output

- this EM process is run once for each gene/kernel document
- the results returned for each gene are
  - a list of the most highly weighted $\left(\frac{p_i^T}{q_i^T}\right)$ words in the associated theme
  - a list of the most on-theme documents

Finding Themes: Example

- Shatkay et al. have applied this method to find themes in the AIDS literature [*Advances in Digital Libraries*, 2000]



Failure of screening to detect HIV in a foreign laborer who died of toxoplasmosis of the central nervous system.

AIDS-associated cytomegalovirus infection mimicking central nervous system tumors: a diagnostic challenge.

Chagasic granulomatous encephalitis in immunosuppressed patients. Computed tomography and magnetic resonance imaging findings.

Isolated homonymous lateral hemianopsia revealing central nervous system toxoplasmosis as the initial manifestation of AIDS.

Expression and antigenicity of human herpesvirus 8 encoded ORF59 protein in AIDS-associated Kaposi's sarcoma.

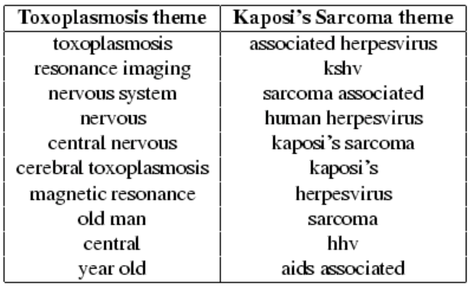
Primary intraosseous AIDS-associated Kaposi's sarcoma. Report of two cases with initial jaw involvement.

Expression of human herpesvirus-8 (HHV-8) encoded pathogenic genes in Kaposi's sarcoma (KS) primary lesions

Further confirmation of the association of human herpesvirus 8 with Kaposi's sarcoma.

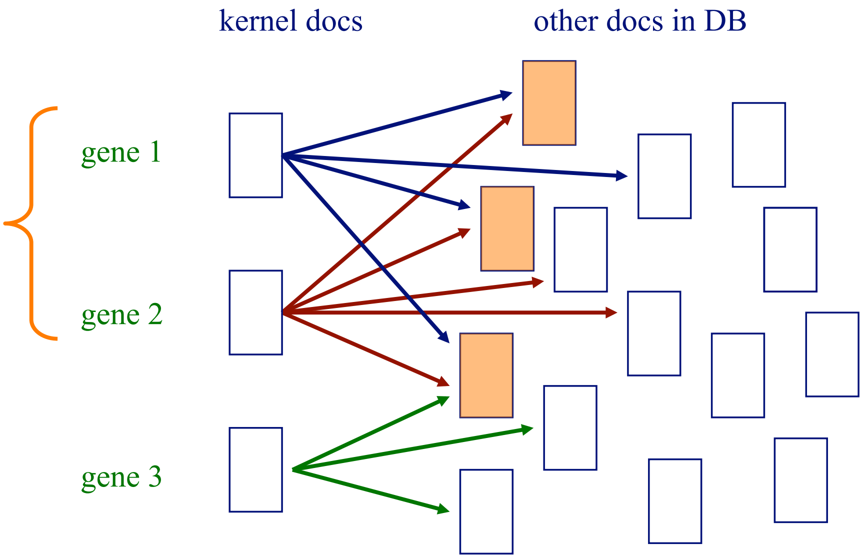
titles of top-4 documents for two themes

top-10 words for the themes



Toxoplasmosis theme	Kaposi's Sarcoma theme
toxoplasmosis	associated herpesvirus
resonance imaging	kshv
nervous system	sarcoma associated
nervous	human herpesvirus
central nervous	kaposi's sarcoma
cerebral toxoplasmosis	kaposi's
magnetic resonance	herpesvirus
old man	sarcoma
central	hhv
year old	aids associated

Finding Related Genes

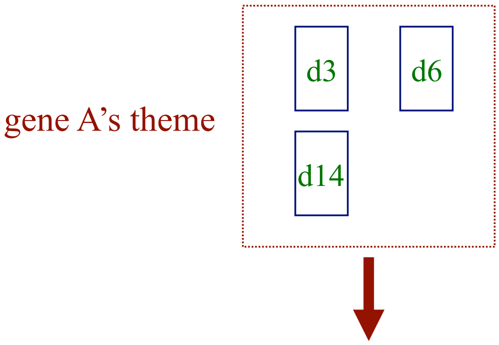


step 1: given kernel documents, find themes

step 2: given themes, find related genes

Representing Genes

- represent each gene using fixed-length vector in which each element corresponds to a document
- put a 1 in a given element if the associated document is strongly in the gene’s theme

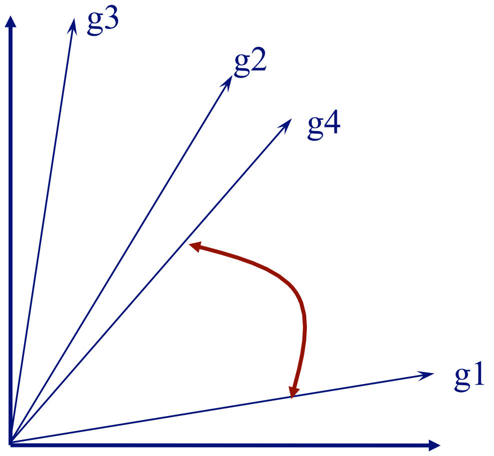


d1	d2	d3			d6								d14		
0	0	1	0	0	1	0	0	0	0	0	0	0	1	0	0

●●●

The Vector Space Model

- the similarity between two genes can be assessed by the similarity of their associated vectors
- this is a common method in information retrieval to assess document similarity; here we are assessing gene similarity



Vector Similarity

- one way to determine vector similarity is the cosine measure:

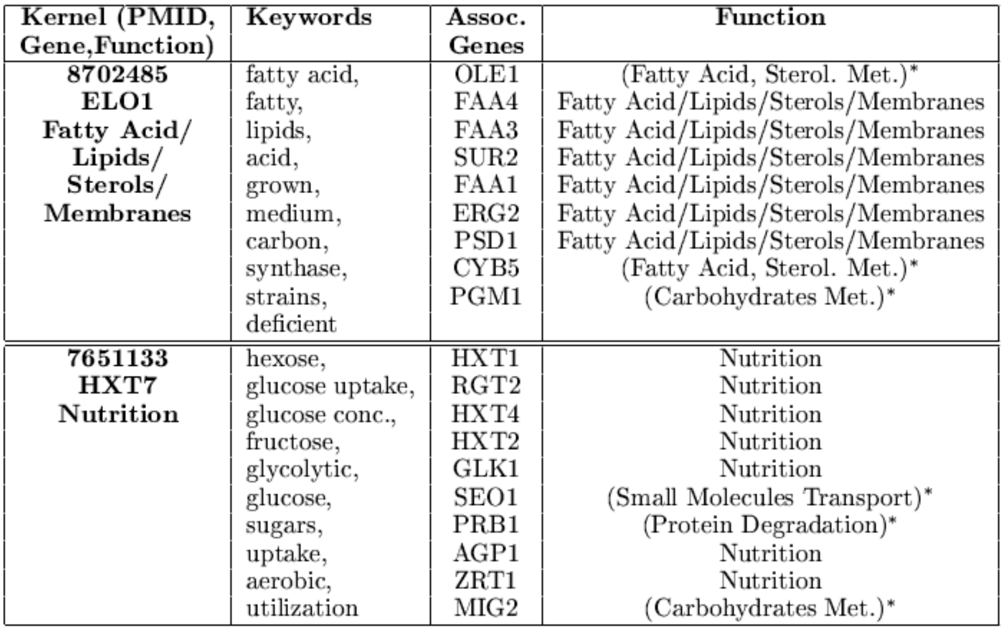
$$\cos(\vec{a},\vec{b}) = \frac{\sum_i a_i b_i}{\sqrt{\sum_i a_i^2}\sqrt{\sum_i b_i^2}}$$

- if the vectors are normalized, we can simply take their dot product

Shatkay et al. Experiment

- analyzed 408 yeast genes
- documents = abstracts
- kernel documents: oldest reference for each gene in SGD
- database: 33,700 yeast-related documents

Shatkay et al. Experimental Results



Kernel (PMID, Gene,Function)	Keywords	Assoc. Genes	Function
8702485 ELO1 Fatty Acid/ Lipids/ Sterols/ Membranes	fatty acid,	OLE1	(Fatty Acid, Sterol. Met.)*
	fatty,	FAA4	Fatty Acid/Lipids/Sterols/Membranes
	lipids,	FAA3	Fatty Acid/Lipids/Sterols/Membranes
	acid,	SUR2	Fatty Acid/Lipids/Sterols/Membranes
	grown,	FAA1	Fatty Acid/Lipids/Sterols/Membranes
	medium,	ERG2	Fatty Acid/Lipids/Sterols/Membranes
	carbon,	PSD1	Fatty Acid/Lipids/Sterols/Membranes
	synthase,	CYB5	(Fatty Acid, Sterol. Met.)*
	strains,	PGM1	(Carbohydrates Met.)*
	deficient		
7651133 HXT7 Nutrition	hexose,	HXT1	Nutrition
	glucose uptake,	RGT2	Nutrition
	glucose conc.,	HXT4	Nutrition
	fructose,	HXT2	Nutrition
	glycolytic,	GLK1	Nutrition
	glucose,	SEO1	(Small Molecules Transport)*
	sugars,	PRB1	(Protein Degradation)*
	uptake,	AGP1	Nutrition
	aerobic,	ZRT1	Nutrition
	utilization	MIG2	(Carbohydrates Met.)*

Figure from H. Shatkay et al., *ISMB* 2000