BMI/CS 776
Lecture #30
Biomedical Text Mining

Colin Dewey
(adapted from slides by Mark Craven)
2007.05.10

Some Important Text-Mining Problems

- hypothesis generation

  Given: biomedical objects/classes of interest (e.g. diseases & 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

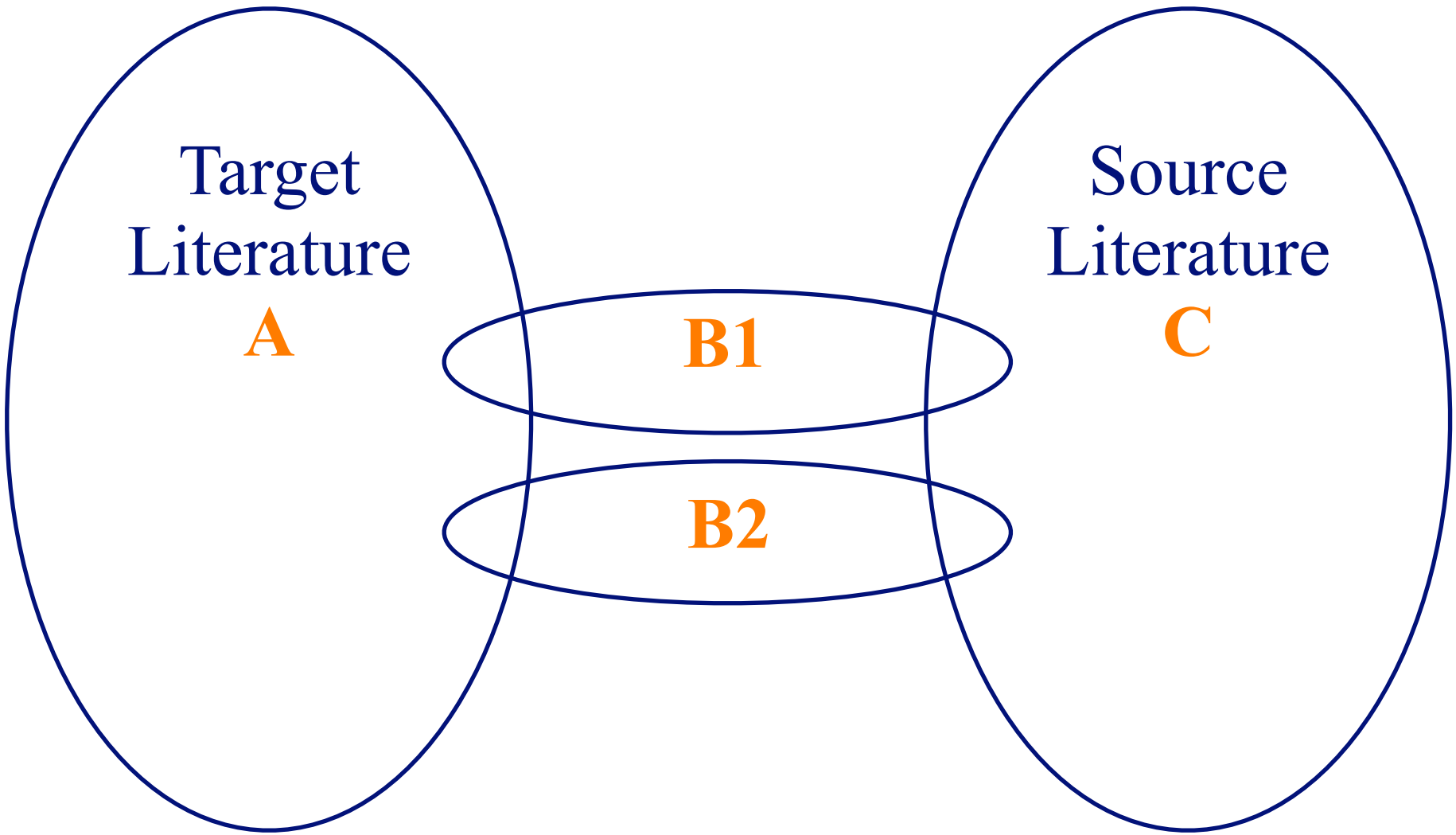
- ad-hoc retrieval
  Given: a query
  Do: return relevant documents/passages

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

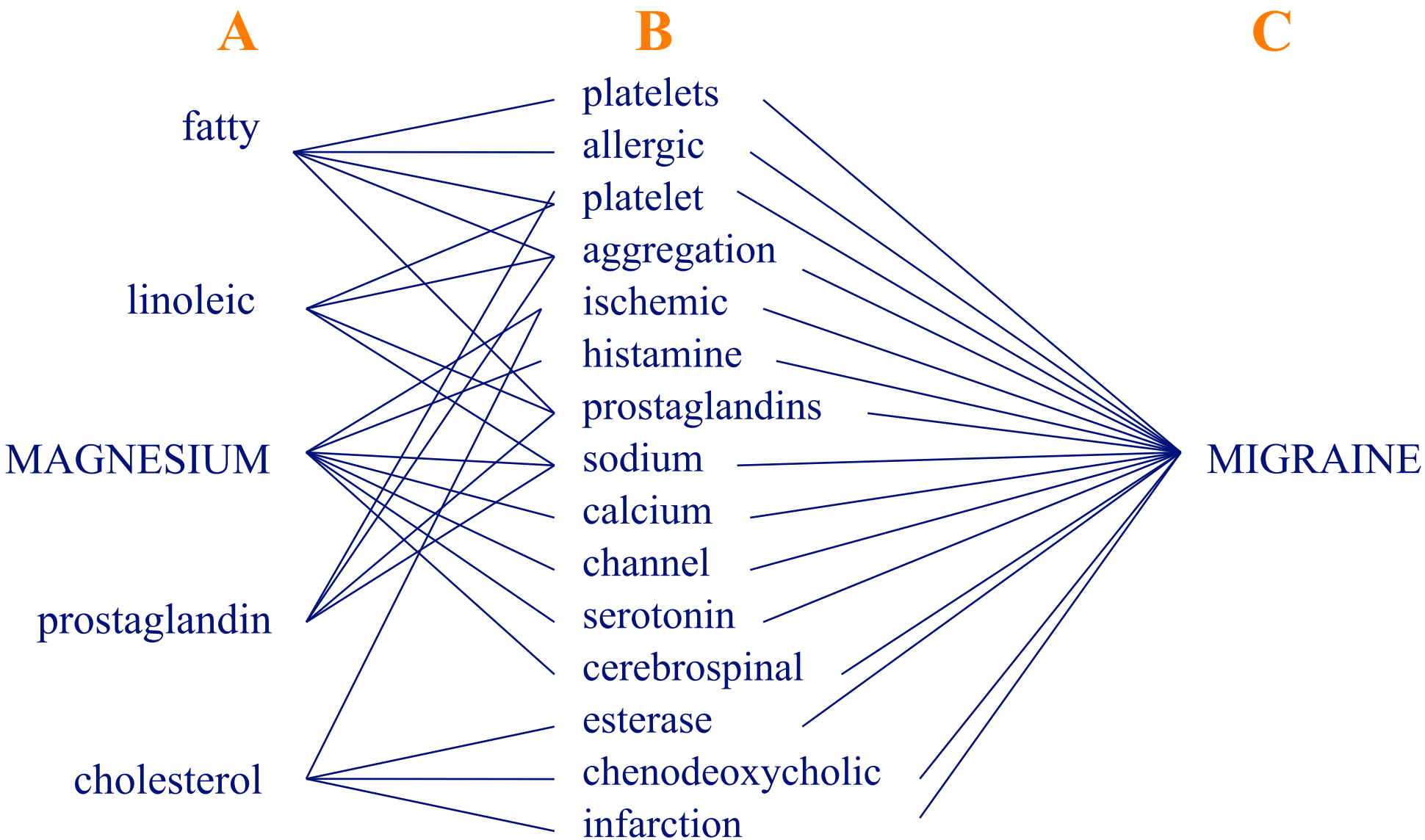
Hypothesis Generation by Finding Complementary Literatures

- [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; retain words
  - 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 <u>human</u> expert filter terms
- prune **A** list using *category restrictions* (e.g. dietary factors, toxins, etc.)
- prune **C-B**, **B-A** linkages by requiring:

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

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

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

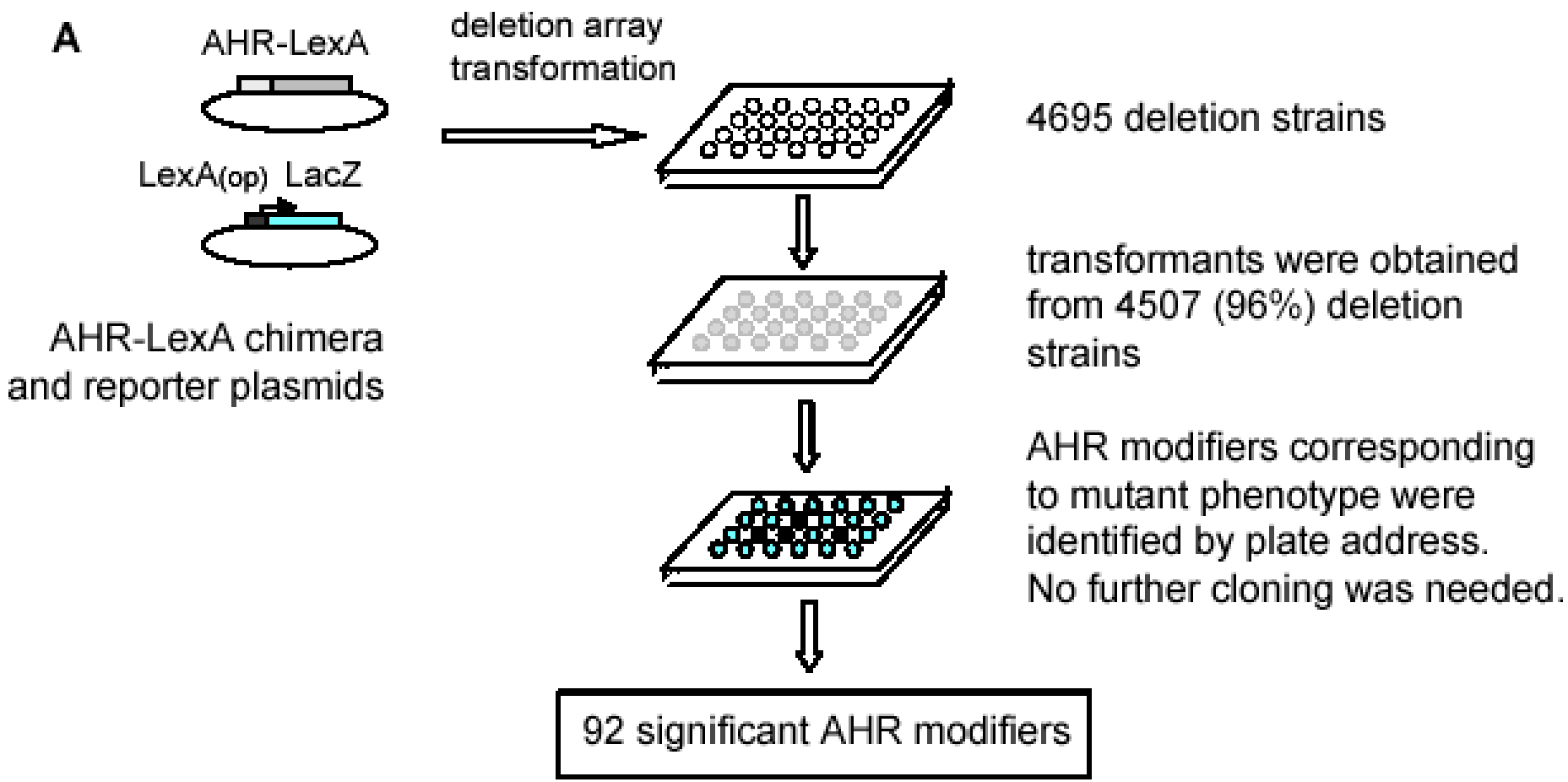
ARROWSMITH Case Studies

- *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

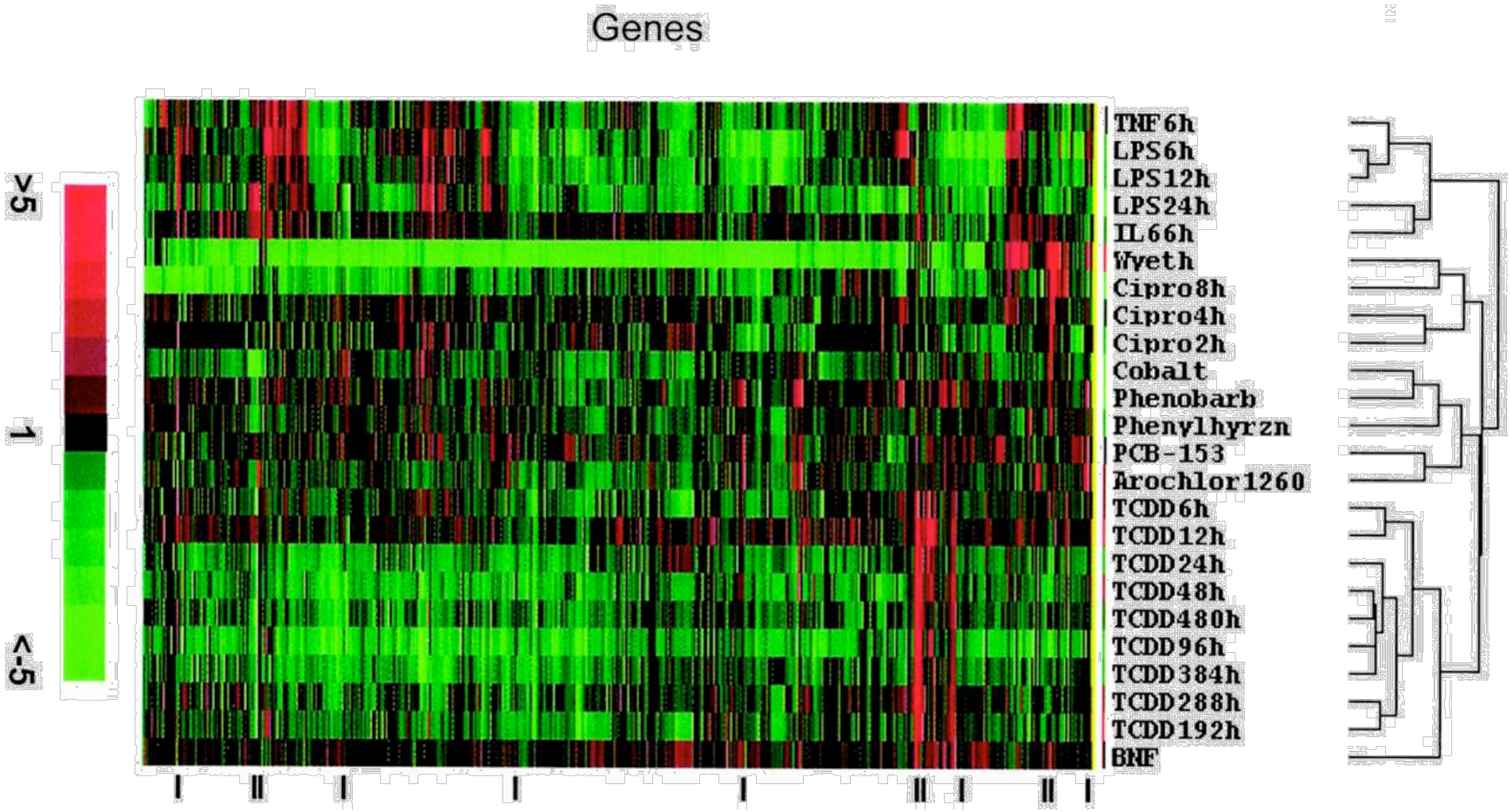
High-throughput Experiment Example:
Yao et al., *PLoS Biology* 2004



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

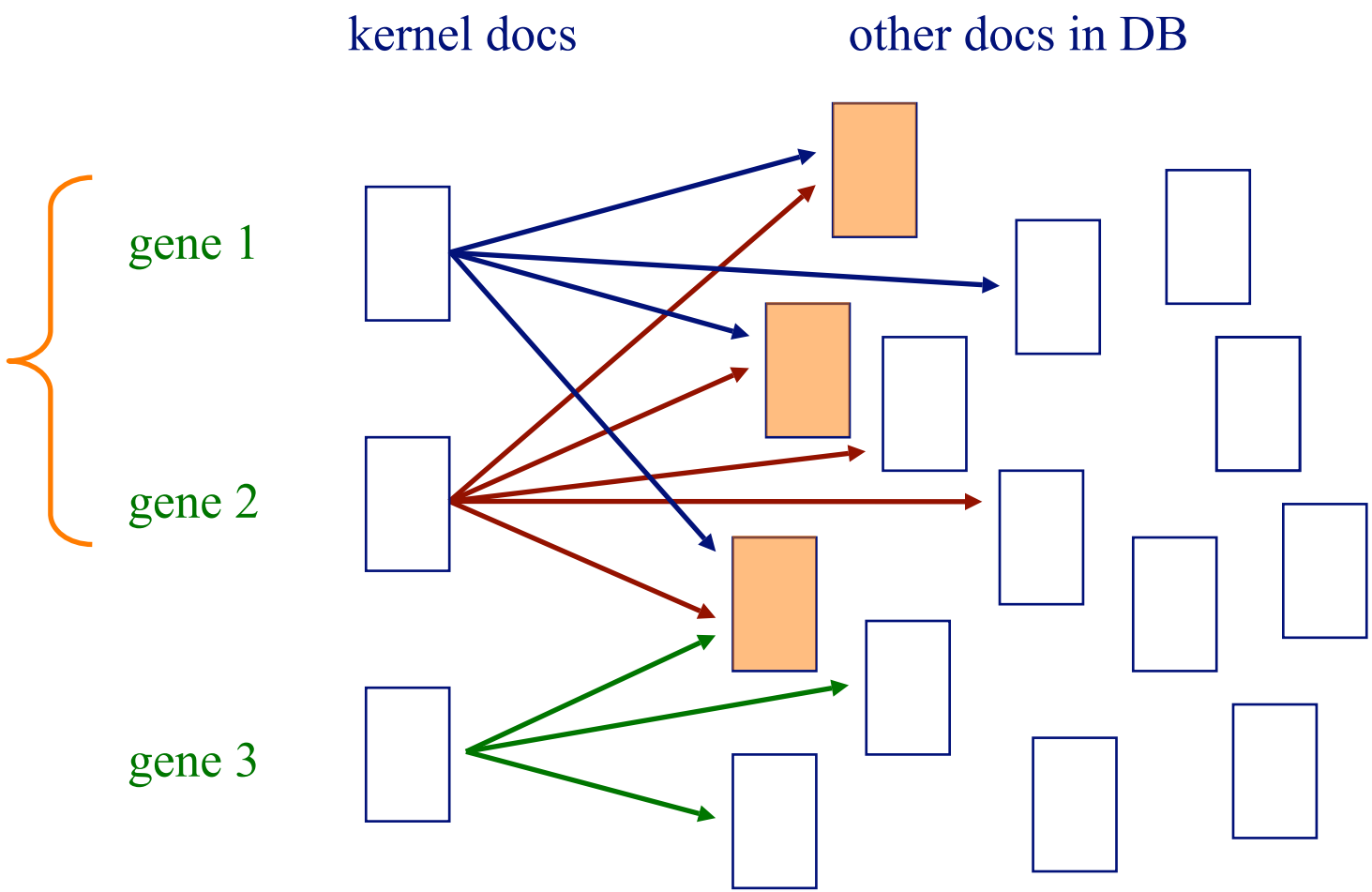
High-throughput Experiment Example:

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

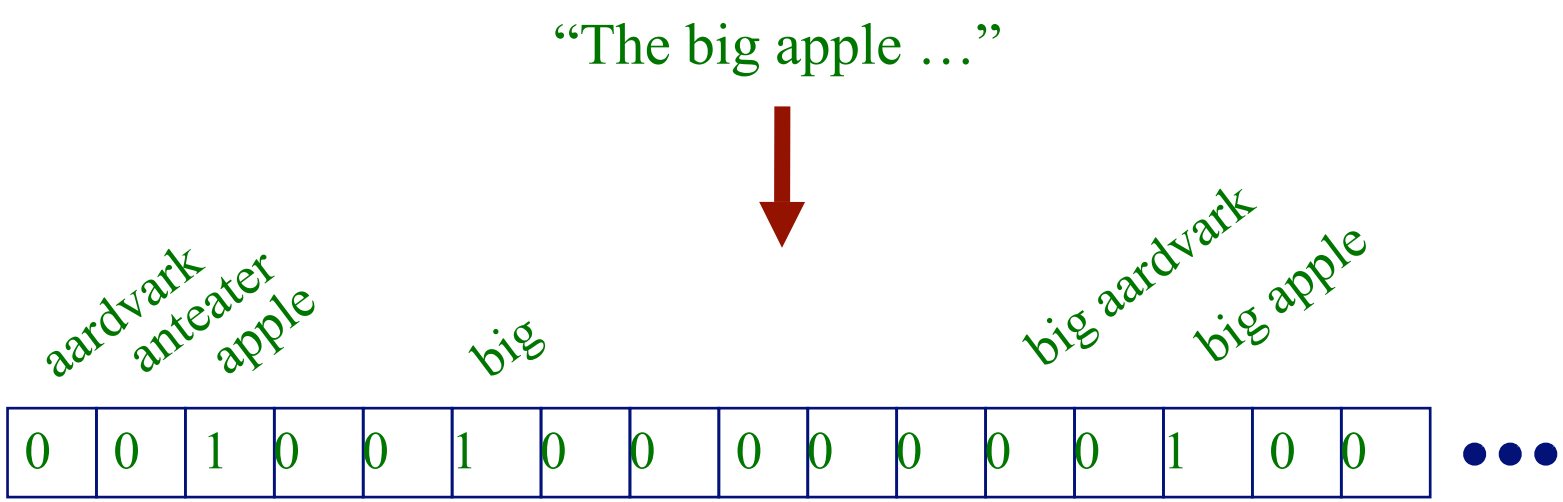


step 1: given kernel documents, find themes

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

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

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

Theme Example

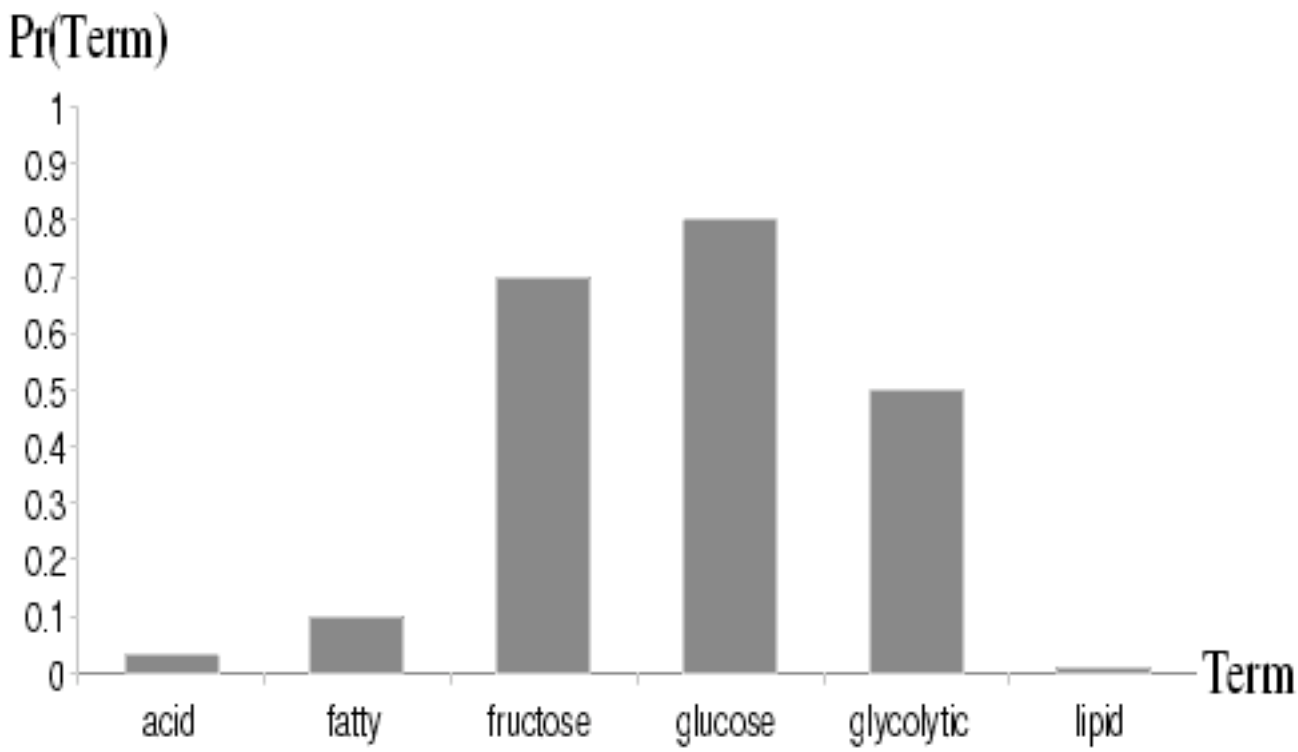


Figure from H. Shatkay et al., *ISMB* 2000

Theme = "Nutrition"

Other Parameters

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

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

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

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

  - 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

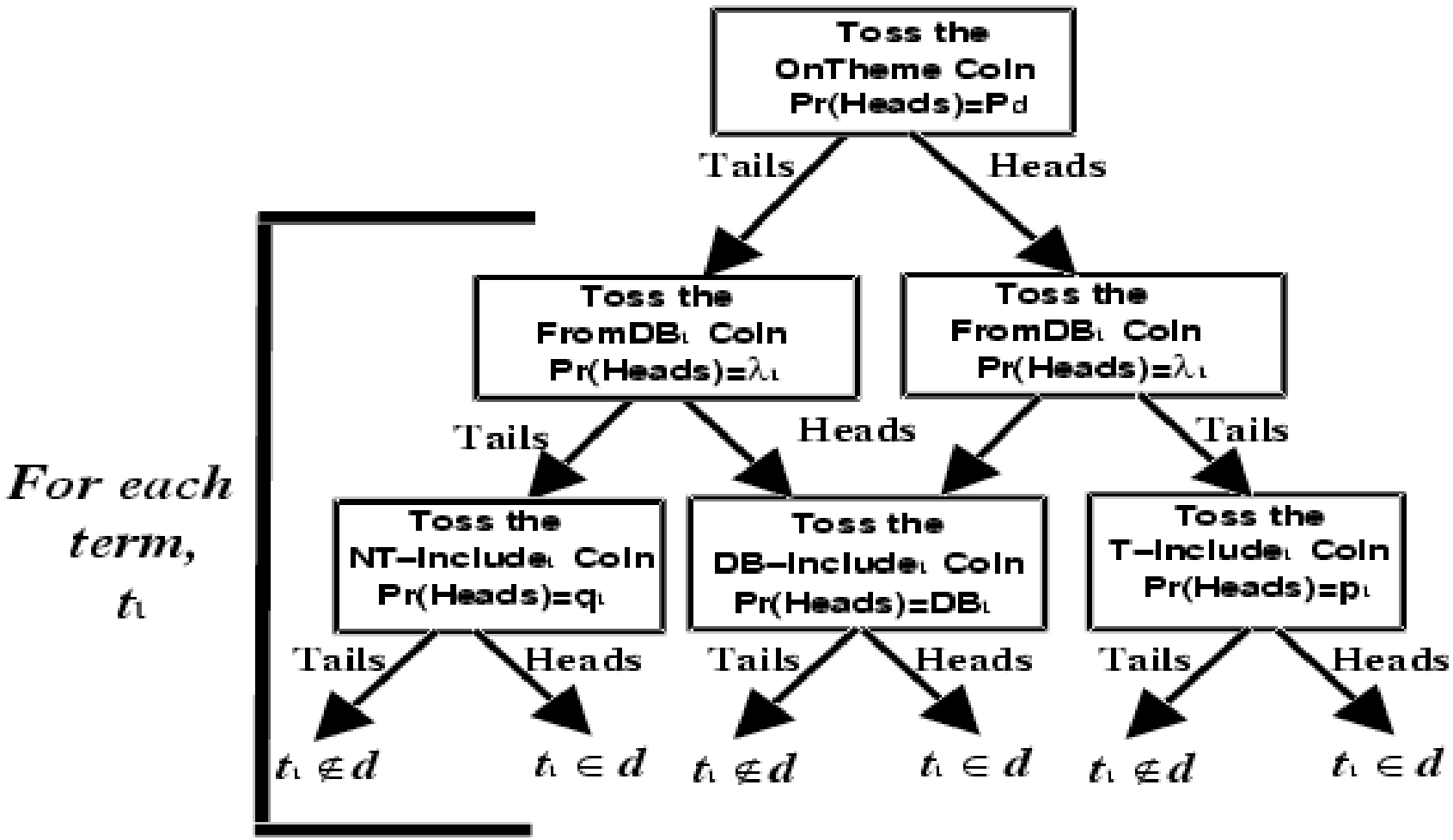


Figure from H. Shatkay et al., *Advances in Digital Libraries* 2000

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

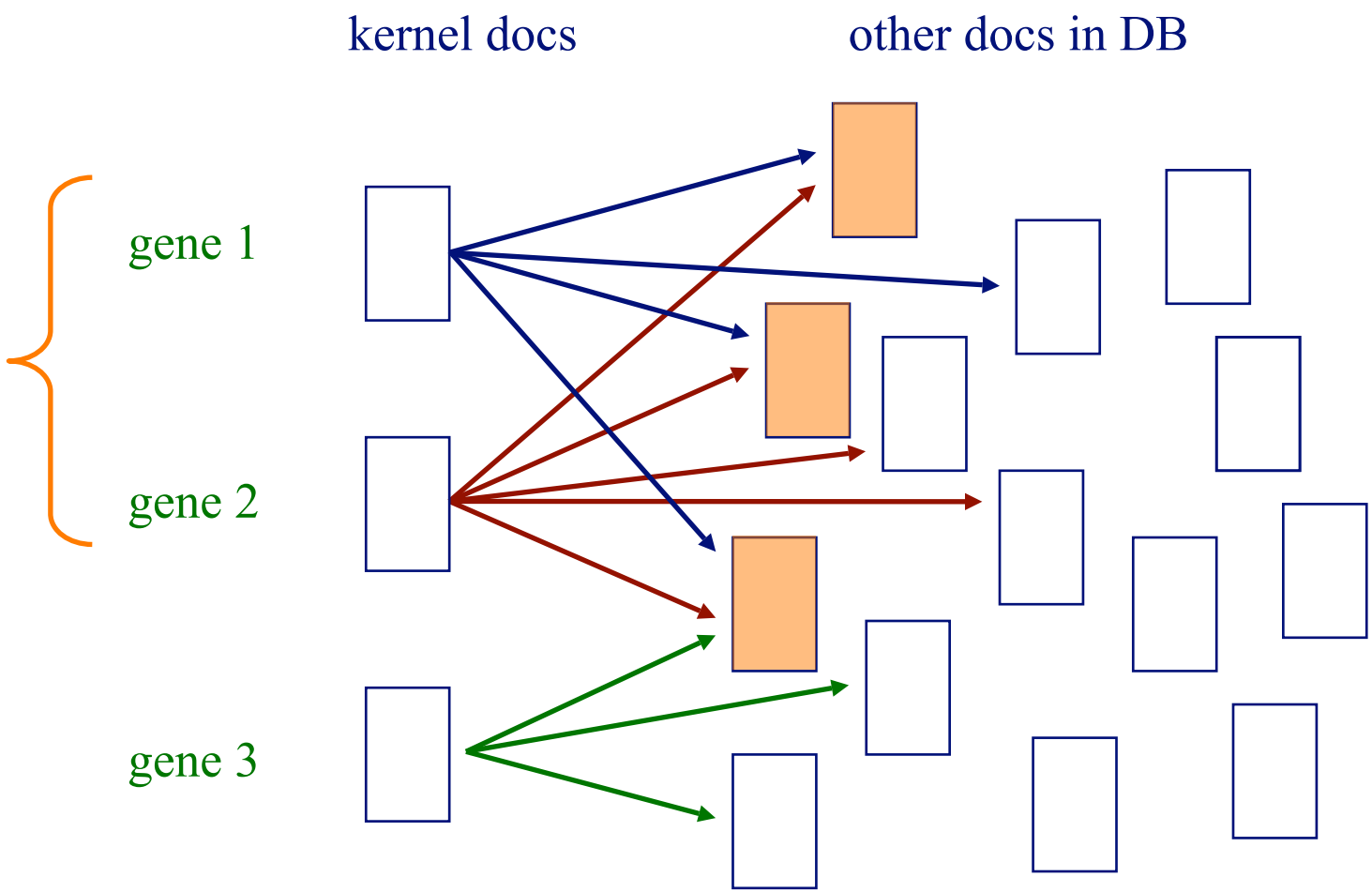
Further confirmation of the association of human herpesvirus 8 with Kaposi's sarcoma.

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

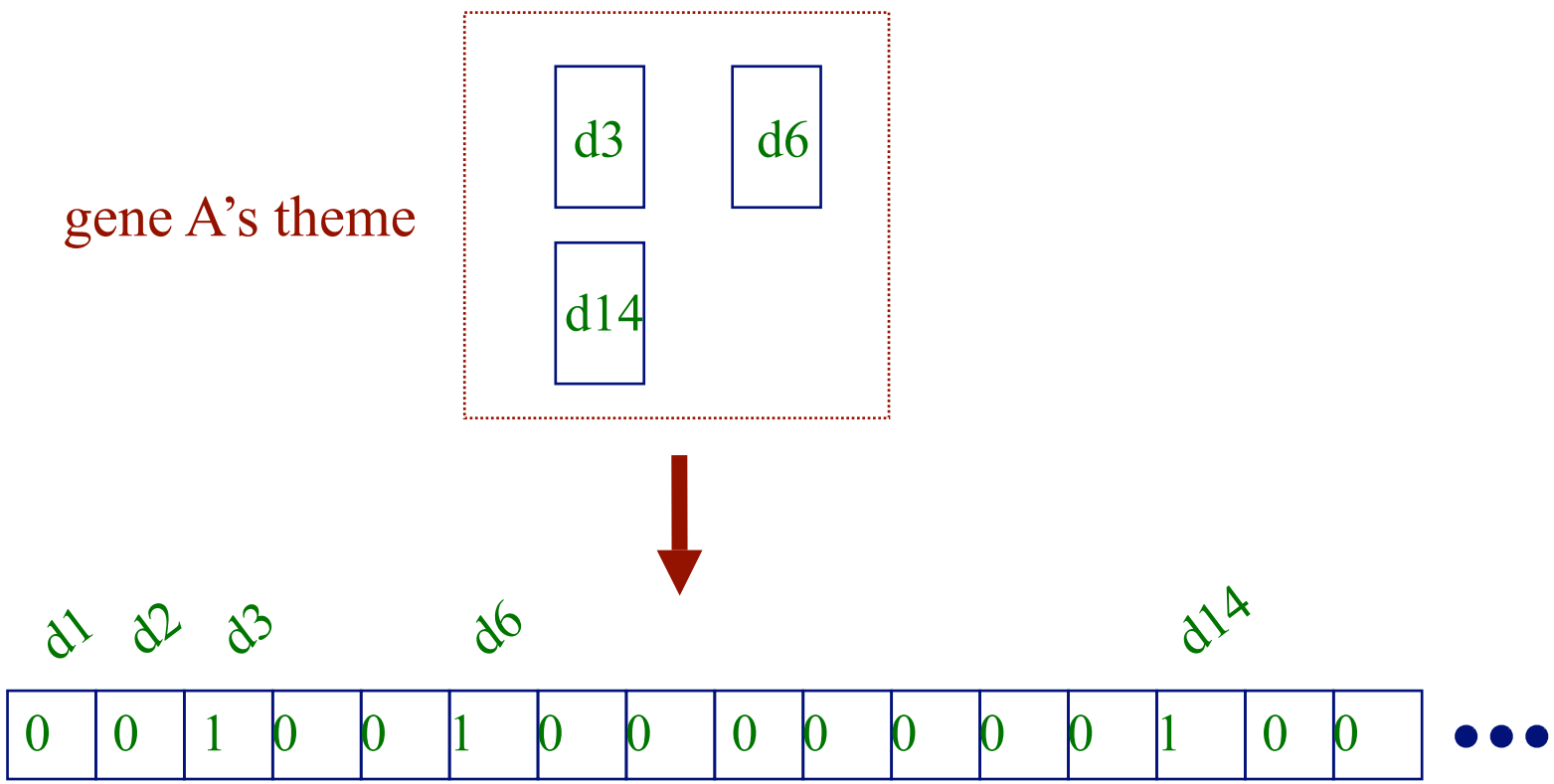


step 1: given kernel documents, find themes

step 2: given themes, find related genes

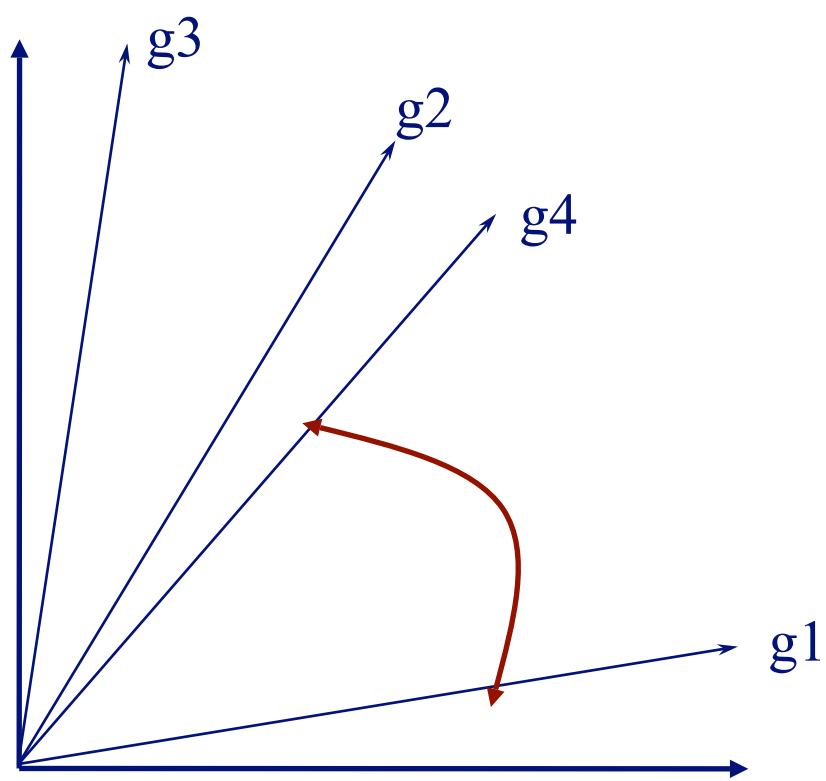
Representing Genes

- 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



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

The Information Extraction Task: Named Entity Recognition

Analysis of Yeast PRP20 Mutations and Functional Complementation by the Human Homologue RCC1, a Protein Involved in the Control of Chromosome Condensation

Fleischmann M, Clark M, Forrester W, Wickens M, Nishimoto T, Aebi M

Mutations in the **PRP20** gene of yeast show a pleitropic phenotype, in which both mRNA metabolism and nuclear structure are affected. **SRM1** mutants, defective in the same gene, influence the signal transduction pathway for the **pheromone** response . . .
By **immunofluorescence microscopy** the **PRP20** protein was localized in the **nucleus**. Expression of the **RCC1** protein can complement the temperature-sensitive phenotype of **PRP20** mutants, demonstrating the functional similarity of the yeast and mammalian proteins

- proteins
- small molecules
- methods
- cellular compartments

The Information Extraction Task: Relation Extraction

Analysis of Yeast PRP20 Mutations and Functional Complementation by the Human Homologue RCC1, a Protein Involved in the Control of Chromosome Condensation

Fleischmann M, Clark M, Forrester W, Wickens M, Nishimoto T, Aebi M

Mutations in the PRP20 gene of yeast show a pleitropic phenotype, in which both mRNA metabolism and nuclear structure are affected. SRM1 mutants, defective in the same gene, influence the signal transduction pathway for the pheromone response . . .
By immunofluorescence microscopy the **PRP20** protein was localized in the **nucleus**. Expression of the RCC1 protein can complement the temperature-sensitive phenotype of PRP20 mutants, demonstrating the functional similarity of the yeast and mammalian proteins

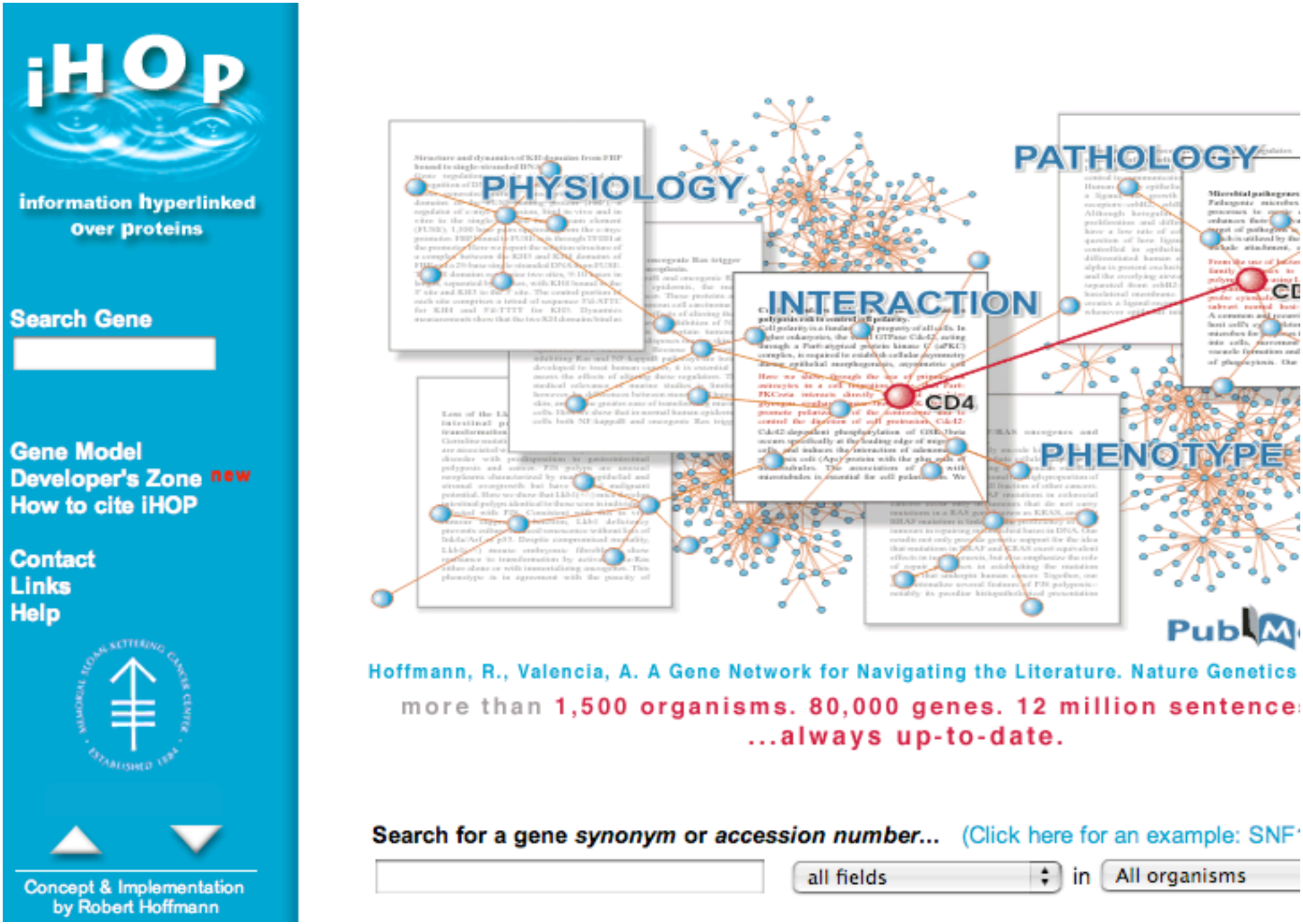
→ subcellular-localization(**PRP20**, **nucleus**)

Motivation for Information Extraction

- motivation for <u>named entity recognition</u>
  - better indexing of biomedical articles
  - assisting in relation extraction

- motivation for relation extraction
  - assisting in the construction and updating of databases
  - providing structured summaries for queries

    What is known about protein X (subcellular & tissue localization, associations with diseases, interactions with drugs, …)?
  - assisting scientific discovery by detecting previously unknown relationships, annotating experimental data



http://www.ihop-net.org/UniPub/iHOP/ 32

Task: Aiding Annotation of Gene/Protein Function

1. *document filtering/classification*

   Given: article

   Do: determine if it is relevant for curation of any genes in a particular category (GO, tumor biology, expression, etc.)

   (TREC Genomics Track in 2004, 2005)

2. *annotation assignment*

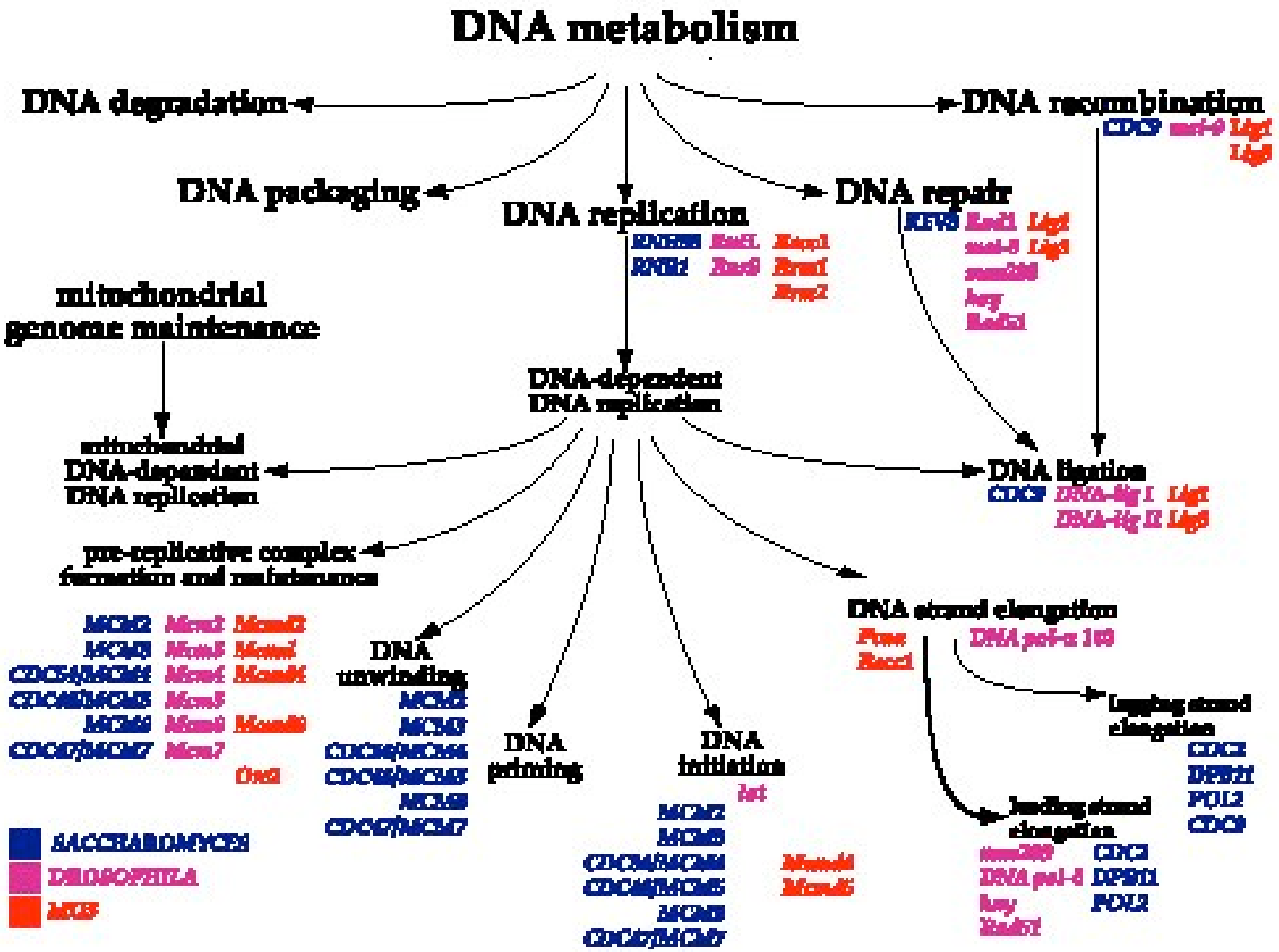
   Given: article, gene

   Do: return Gene Ontology concepts for the gene that are supported by the article

   (BioCreative evaluation in 2003)

The Gene Ontology

- a controlled vocabulary of more than 17,600 concepts describing molecular functions, biological processes, and cellular components



Aiding Annotation: MGI Example

Gene Detail – Mozilla

File Edit View Go Bookmarks Tools Window Help

Back Forward Reload Stop http://www.informatics.jax.org/javawi2/servlet/WIFetch?page=markerDetail&key=8908 Search Print

Home Bookmarks Members WebMail Connections BizJournal SmartUpdate Mktplace

MGI
Mouse Genome Informatics
MGI Home Help

Search for
Go
in these sections
All sections
Gene symbols/names
Accession IDs
Phenotypes
Gene Expression

Advanced search for...

Search Categories
All Search Tools
Genes/Markers
Alleles/Phenotypes
Strains/Polymorphisms
Expression
Sequences NEW
Comparative Maps/Data
Mouse Maps/Data
Mouse Tumor Biology
Probes/Clones
References
Vocabulary Browsers
Gene Ontology (GO)
Anatomical Dictionary
Phenotype Classifications

MouseBLAST

Additional Resources

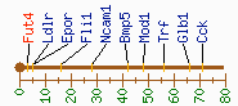
Citing These Resources
Funding Information
Warranty Disclaimer & Copyright Notice
Send questions and comments to User Support.

The Jackson Laboratory
last database update 07/23/2004

Gene Detail

Your Input Welcome

Symbol	Fut4
Name	fucosyltransferase 4
ID	MGI:95594 Nomenclature History
Synonyms	3-fucosyl-N-acetyl-lactosamine, 3-fucosyl-N-acetyl-lactosamine, alpha (1,3) fucosyltransferase, myeloid specific, FAL, FucT-IV, SSEA-1
Map position	Chromosome 9 3.0 cM, Detailed Map ± 1 cM Mapping data(5) Ensembl ContigView | UCSC Browser | NCBI Map Viewer
Mammalian orthology	human; rat (Mammalian Orthology)
Sequences	Rep For All
Phenotypes	All phenotypic alleles(1) : Targeted(1)
Polymorphisms	RFLP(1)
Gene Ontology (GO) classifications	Process: protein amino acid glycosylation Component: Golgi apparatus, integral to membrane... Function: fucosyltransferase activity, transferase activity... All GO classifications(7)
Expression	Theiler Stage 1,2,3,5,9,11,13,15,17,19,20,21,22,23,24,28 Tissues(61) Assay Type Results(70) Assays(4) Immunohistochemistry 70 4 GXD literature index(29) cDNA source data(2)
Other database links	DoTS DT.40171675, DT.91334210 UniGene 63450 ENSEMBL ENSMUSG00000049307 LocusLink 14345 NIA Mouse Gene Index NAP015586-001 Entrez Gene 14345
Protein domains	InterPro ID Description IPR001503 Glycosyl transferase, family 10 Graphical View of Protein Domain Structure



Gene Ontology (GO) classifications

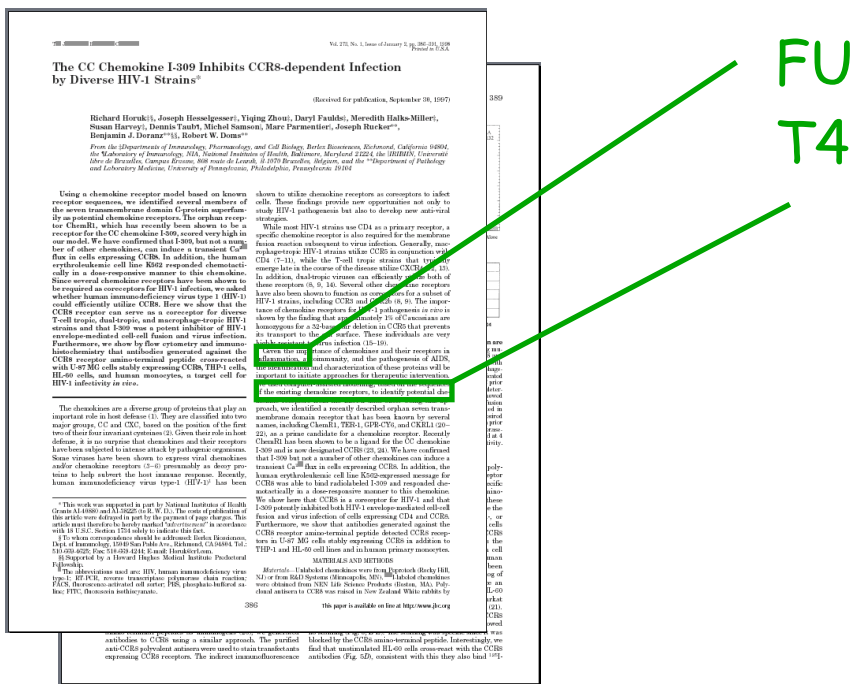
Process	protein amino acid glycosylation
Component	Golgi apparatus, integral to membrane...
Function	fucosyltransferase activity, transferase activity...
All GO classifications(7)	

Annotating Genomes: MGI Example

- the current method for this annotation process…

(Partially) Automating This Annotation Process

- first step: need to recognize references to gene names and GO concepts in text


The CC Chemokine I-309 Inhibits CCR8-dependent Infection by Diverse HIV-1 Strains*

FU

T4

GO:protein amino-acid glycosylation

- recall: more than 17,600 GO terms, tens of thousands of gene names plus variants of each

Recognizing References to GO Terms in Text Passages

- it's not trivial to identify relevant references to GO concepts in the literature
- two examples of gene annotations and supporting passages:

(GO:0008285) negative regulation of cell proliferation

"...inhibition of cell proliferation..."

(GO:0007186) G-protein coupled receptor protein signaling pathway

"...in the signaling pathway, by receptor phosphorylation at the level of receptor/G protein coupling..."

Recognizing References to GO Terms in Text Passages

- *normalization* using linguistic knowledge

  inhibition of cell proliferation ⇒

  cell proliferation inhibition

  inhibits cell proliferation

- matching on statistically associated terms; e.g. unigrams associated with sodium symporter activity

  pantothenate
  biotin
  transporter
  lipoate
  smvt
  uptake
  sodium-depdendent

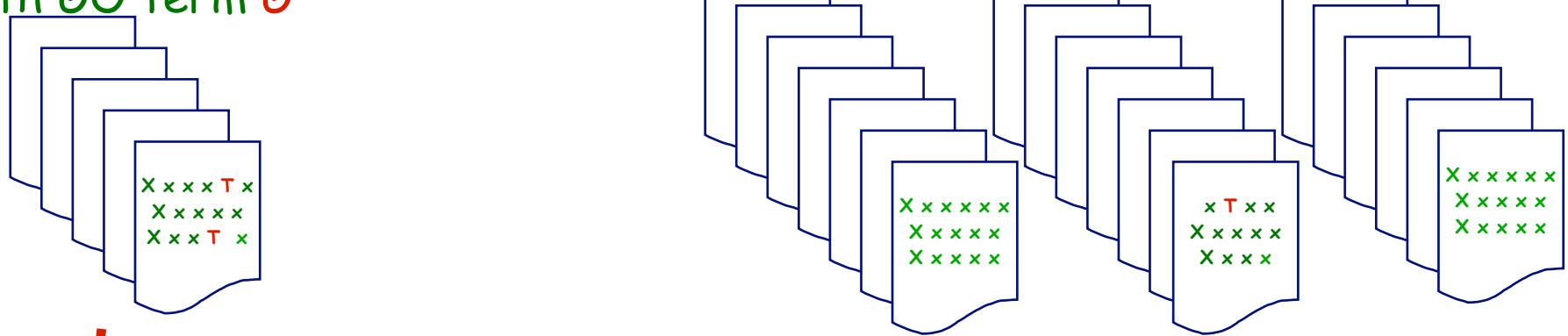
- approximate string matching

Identifying Terms Statistically Associated with GO Concepts

[Ray & Craven, *BMC Bioinformatics* 2005]

documents associated with GO term G

documents NOT associated with GO term G



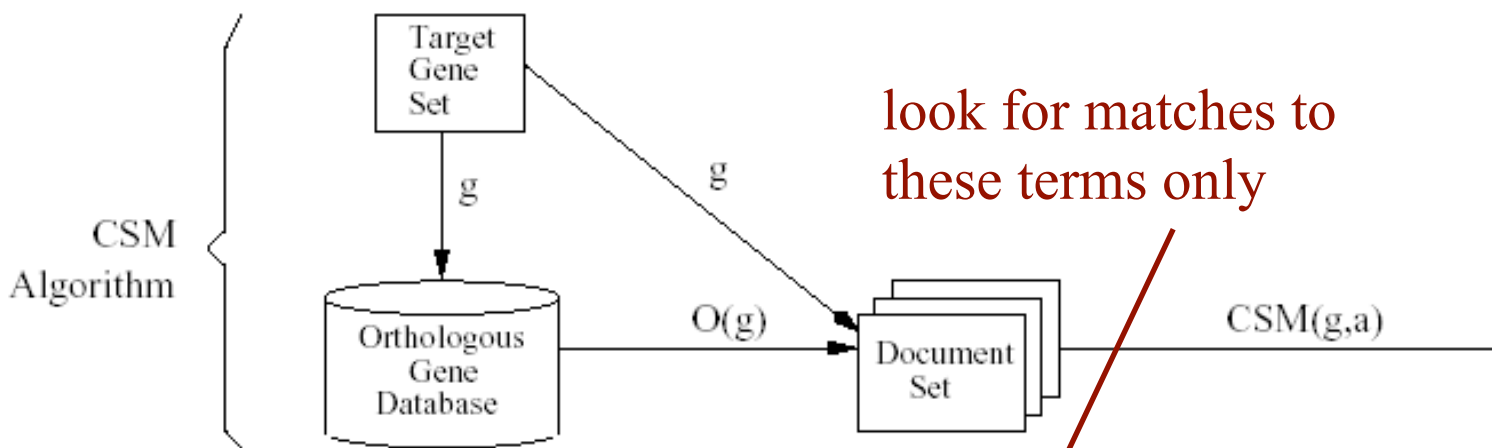
# occurrences of term T	# occurrences of term T
# occurrences of other terms	# occurrences of other terms

compute χ^2 value indicating association between T and G

Using Orthologous Genes in Annotation
[Stoica & Hearst, *PSB* '06]

- *cross-species match* (CSM) method: look for GO matches only for concepts that have been used to annotate orthologous genes

given human FUT4



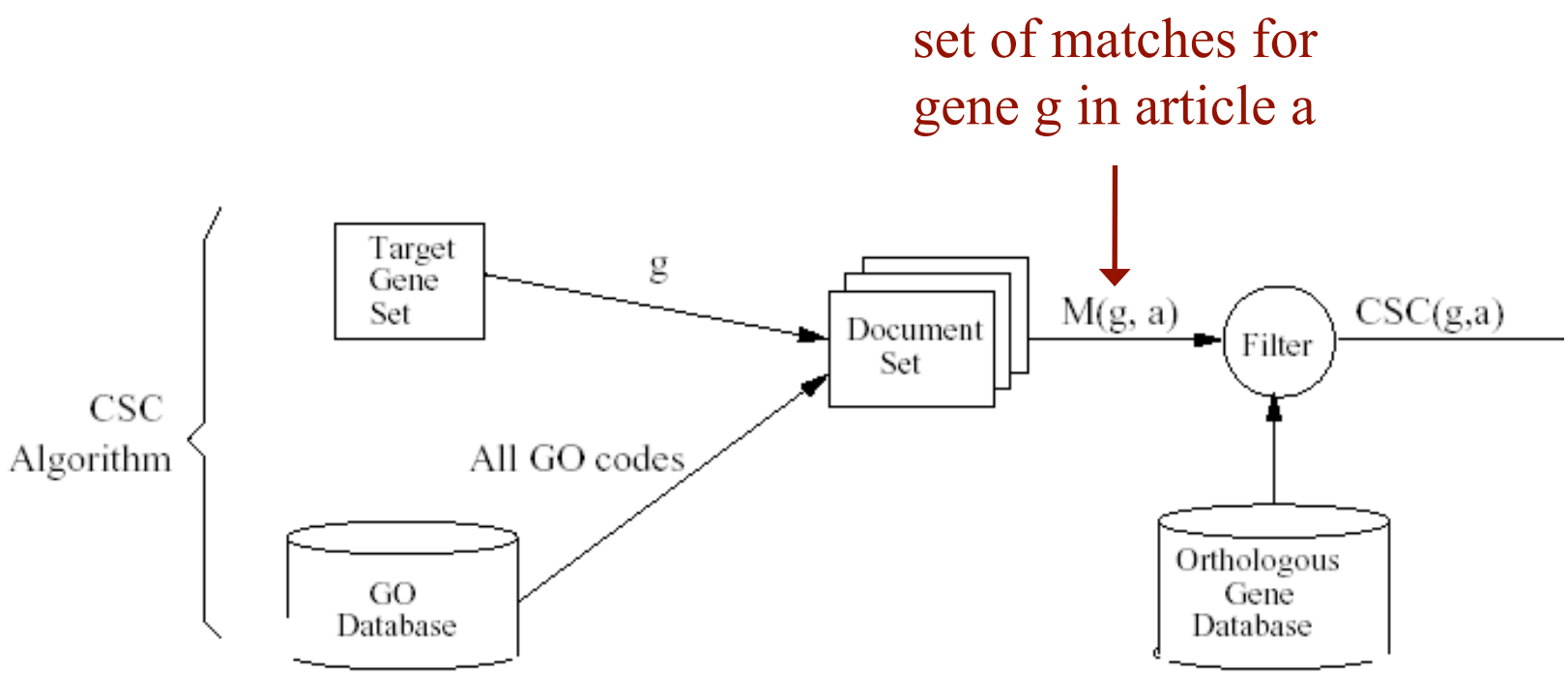
look for matches to these terms only

look up annotations for mouse FUT4

Gene Ontology (GO) classifications	Process	protein amino acid glycosylation
	Component	Golgi apparatus, integral to membrane...
	Function	fucosyltransferase activity, transferase activity...
	All GO classifications(7)	

Using Orthologous Genes in Annotation
[Stoica & Hearst, *PSB* '06]

- *cross-species correlation* (CSC) method: look for GO matches that <u>correlate</u> with concepts that have been used to annotate orthologous genes



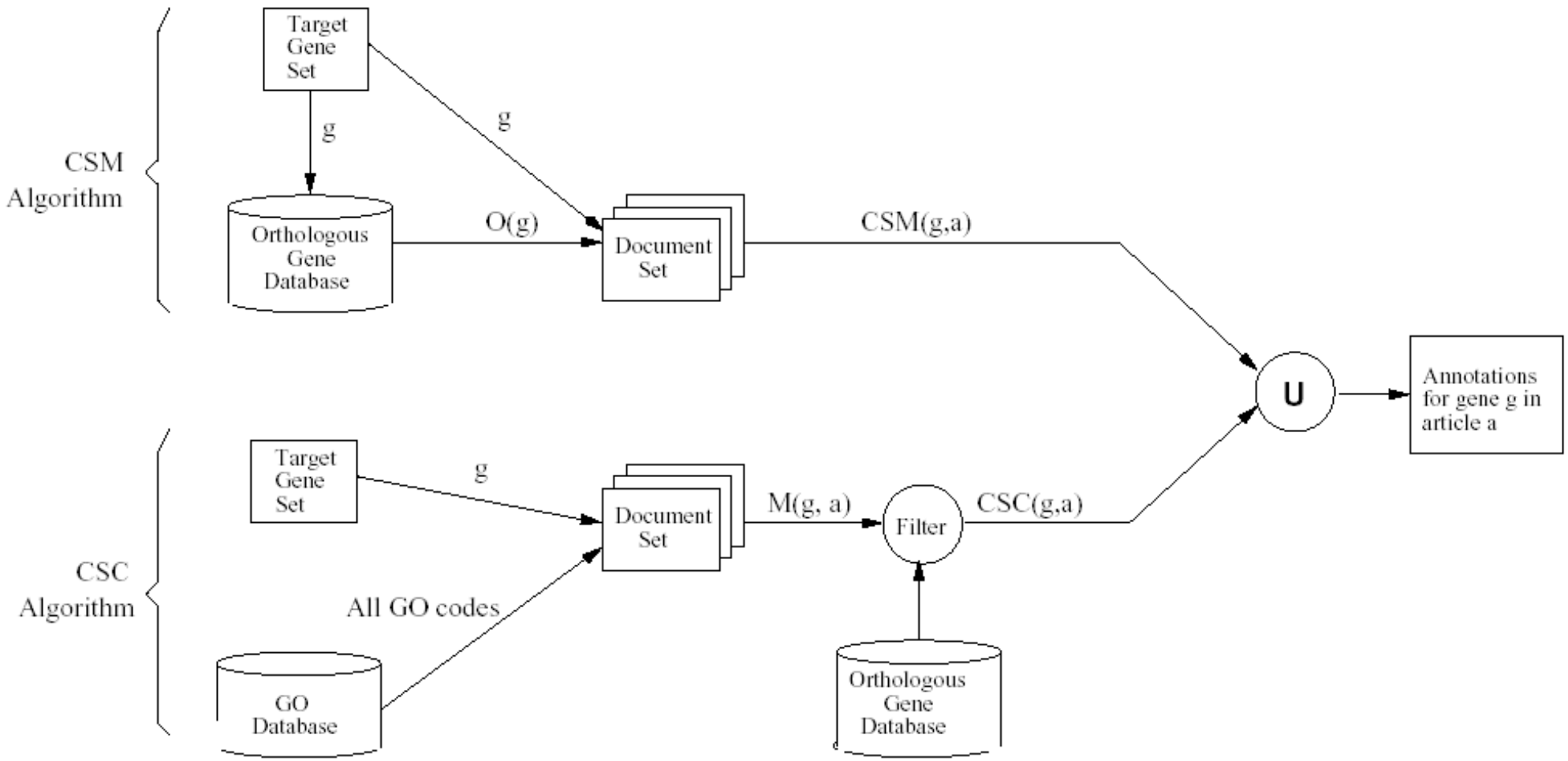
The CSC Algorithm

$CSC(g, a) = \{\}$;
for every GO_1 in $M(g, a)$ // set of matches for gene *g* in article *a*
    $count = 0$;
    for every GO_2 in $O(g)$ // concepts assigned to orthologs of *g*
        if $((\chi^2(GO_1, GO_2) > 3.84)\&\&(GO_1 \neq GO_2))$
            $count++$;
    if $(count > p * o)$ // *o* is size of *O*(*g*), *p* is a specified fraction
        add GO_1 to $CSC(g, a)$;

do concepts GO_1 and GO_2 tend to be associated with one another in annotations in the database?

Using Orthologous Genes in Annotation
[Stoica & Hearst, *PSB* '06]

- full system takes union of CSM and CSC predictions



Stoica & Hearst Results

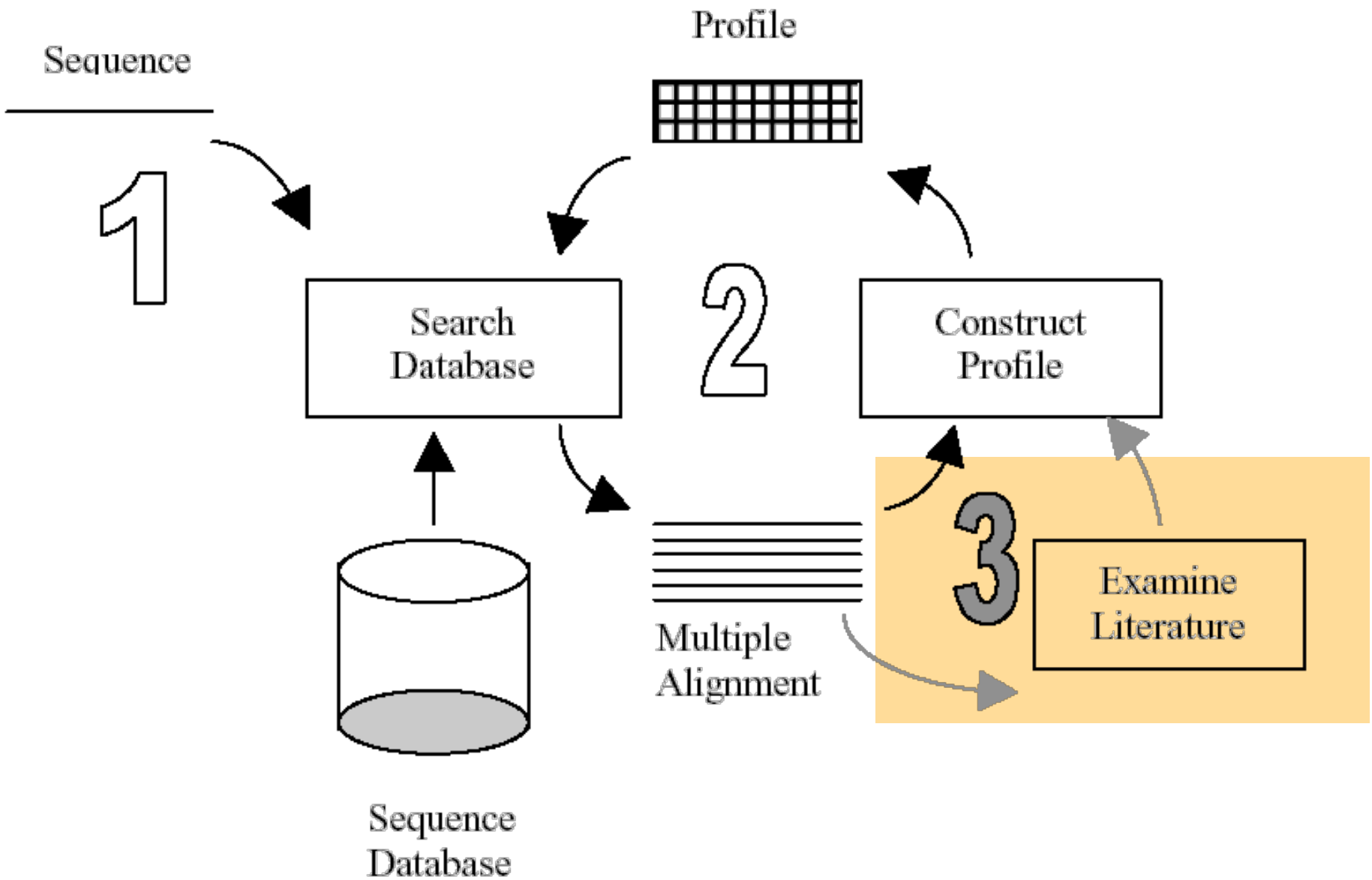
- annotate function on a test-set of 99 articles, 138 human genes
- stringent definition of a *correct* answer: passage references gene name and provides evidence for GO annotation at the right level of the hierarchy

from BioCreative evaluation in 2003

System	Precision	TP (Recall)	F-measure
CSM	0.364	16 (0.068)	0.114
CSC	0.182	44 (0.185)	0.178
CSM + CSC	0.241	51 (0.215)	**0.227**
Ray and Craven[23]	0.213	52 (0.219)	0.216
Chiang and Yu[7,8]	0.327	37 (0.156)	0.211
Ehler and Ruch[11]	0.123	78 (0.329)	0.179
Couto et al.[10]	0.089	58 (0.245)	0.131
Verspoor et al.[26]	0.055	19 (0.080)	0.065
Rice et al.[25]	0.035	16 (0.068)	0.046

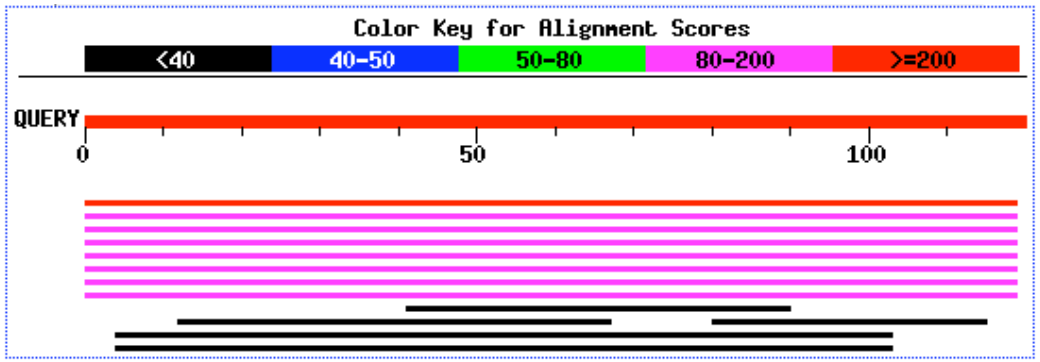
Task: Augmenting PSI-BLAST with Text

[Chang et al., *PSB* '01]

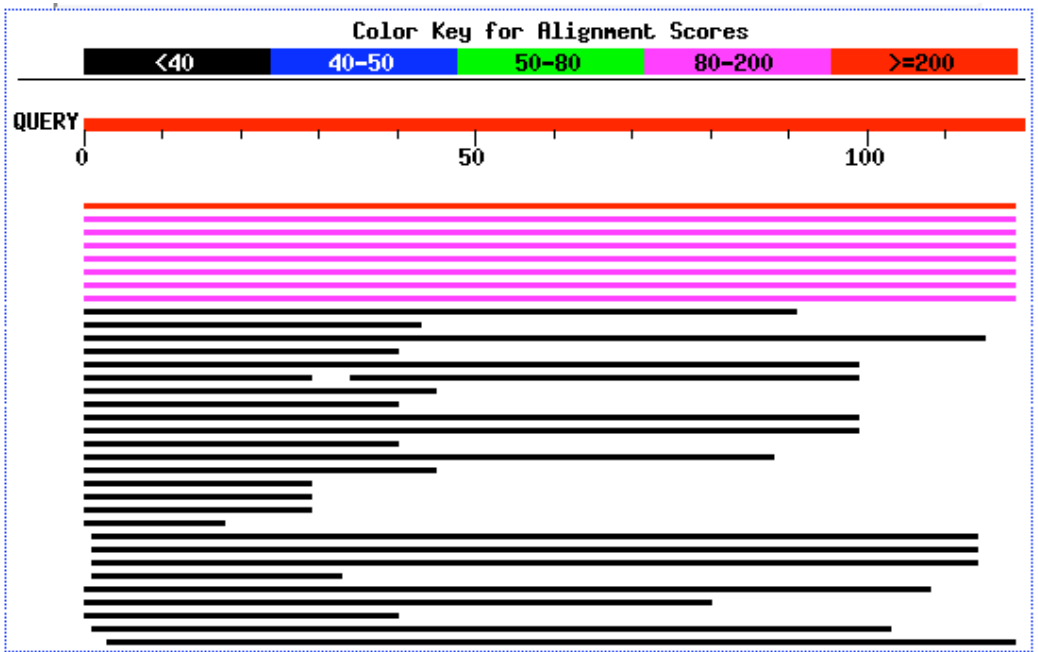


A PSI-BLAST Run Illustrated

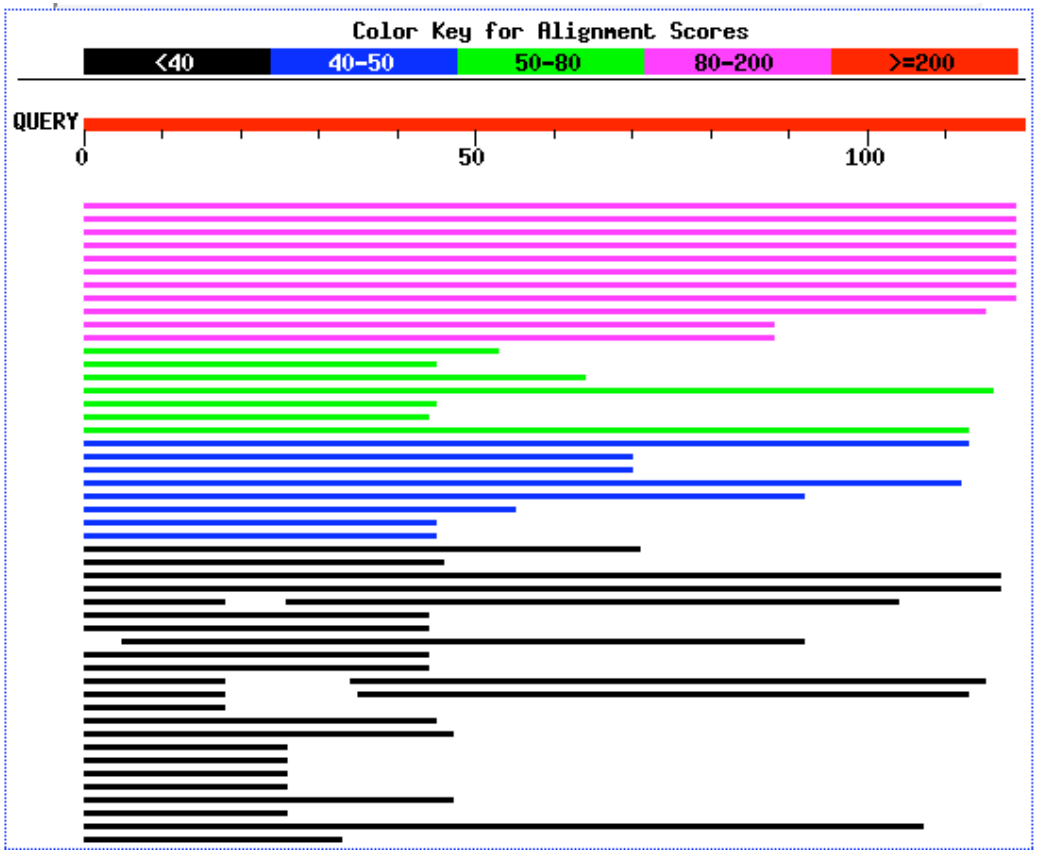
iteration 0



iteration 1



iteration 2



Augmenting PSI-BLAST with Text

- PSI-BLAST often has greater sensitivity than BLAST
- but profile "drift" can occur, as the query is generalized

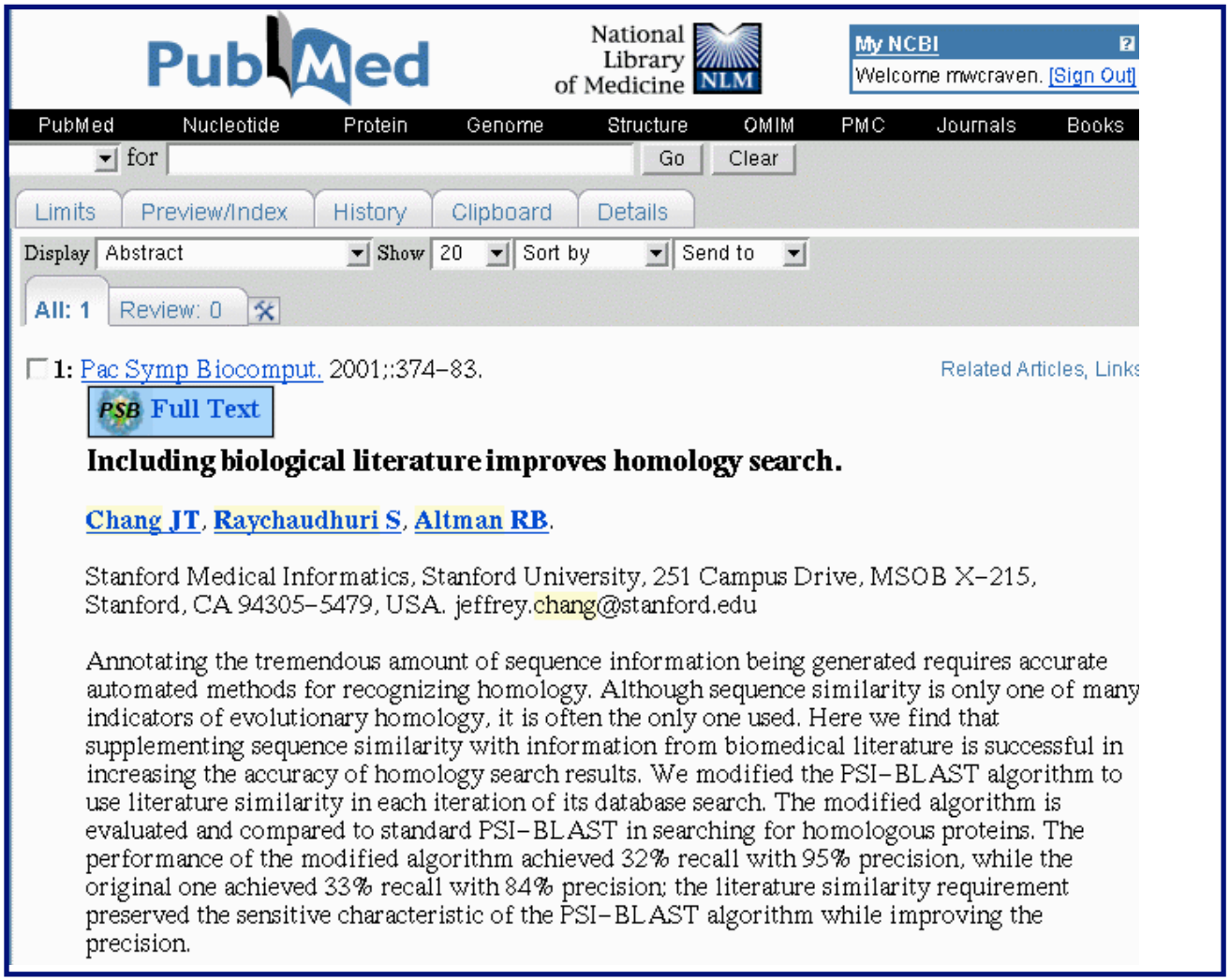
- key idea:
  - represent each protein with text, in addition to amino-acid sequence
  - discard proteins that have low text similarity to the query

Calculating Text Similarity of Proteins

- for each protein, collect MEDLINE entries referenced in Swiss-Prot database
  - collect text in abstracts/MESH headings
  - drop low/high frequency words (associated with fewer than 3 or more than 85,000 sequences)
  - represent protein by vector of word-occurrence counts

- use cosine similarity to assess similarity of proteins
- drop proteins from profile whose similarity to query is below a specified threshold

Calculating Text Similarity of Proteins



PubMed National Library of Medicine NLM

My NCBI
Welcome mwcraven. [Sign Out]

PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books

for Go Clear

Limits Preview/Index History Clipboard Details

Display Abstract Show 20 Sort by Send to

All: 1 Review: 0

1: Pac Symp Biocomput. 2001;:374-83. Related Articles, Links

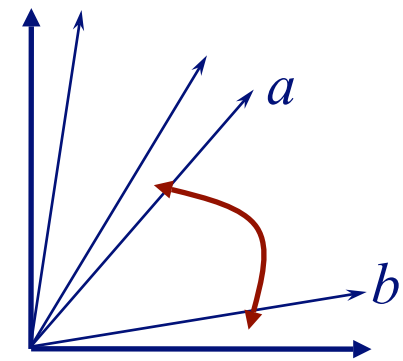
PSB Full Text

Including biological literature improves homology search.

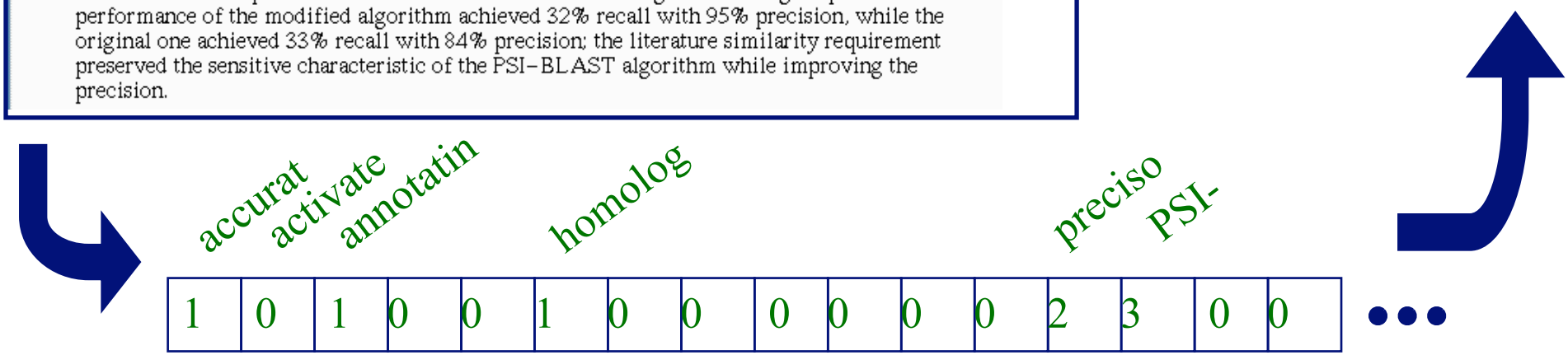
Chang JT, Raychaudhuri S, Altman RB.

Stanford Medical Informatics, Stanford University, 251 Campus Drive, MSOB X-215, Stanford, CA 94305-5479, USA. jeffrey.chang@stanford.edu

Annotating the tremendous amount of sequence information being generated requires accurate automated methods for recognizing homology. Although sequence similarity is only one of many indicators of evolutionary homology, it is often the only one used. Here we find that supplementing sequence similarity with information from biomedical literature is successful in increasing the accuracy of homology search results. We modified the PSI-BLAST algorithm to use literature similarity in each iteration of its database search. The modified algorithm is evaluated and compared to standard PSI-BLAST in searching for homologous proteins. The performance of the modified algorithm achieved 32% recall with 95% precision, while the original one achieved 33% recall with 84% precision; the literature similarity requirement preserved the sensitive characteristic of the PSI-BLAST algorithm while improving the precision.



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



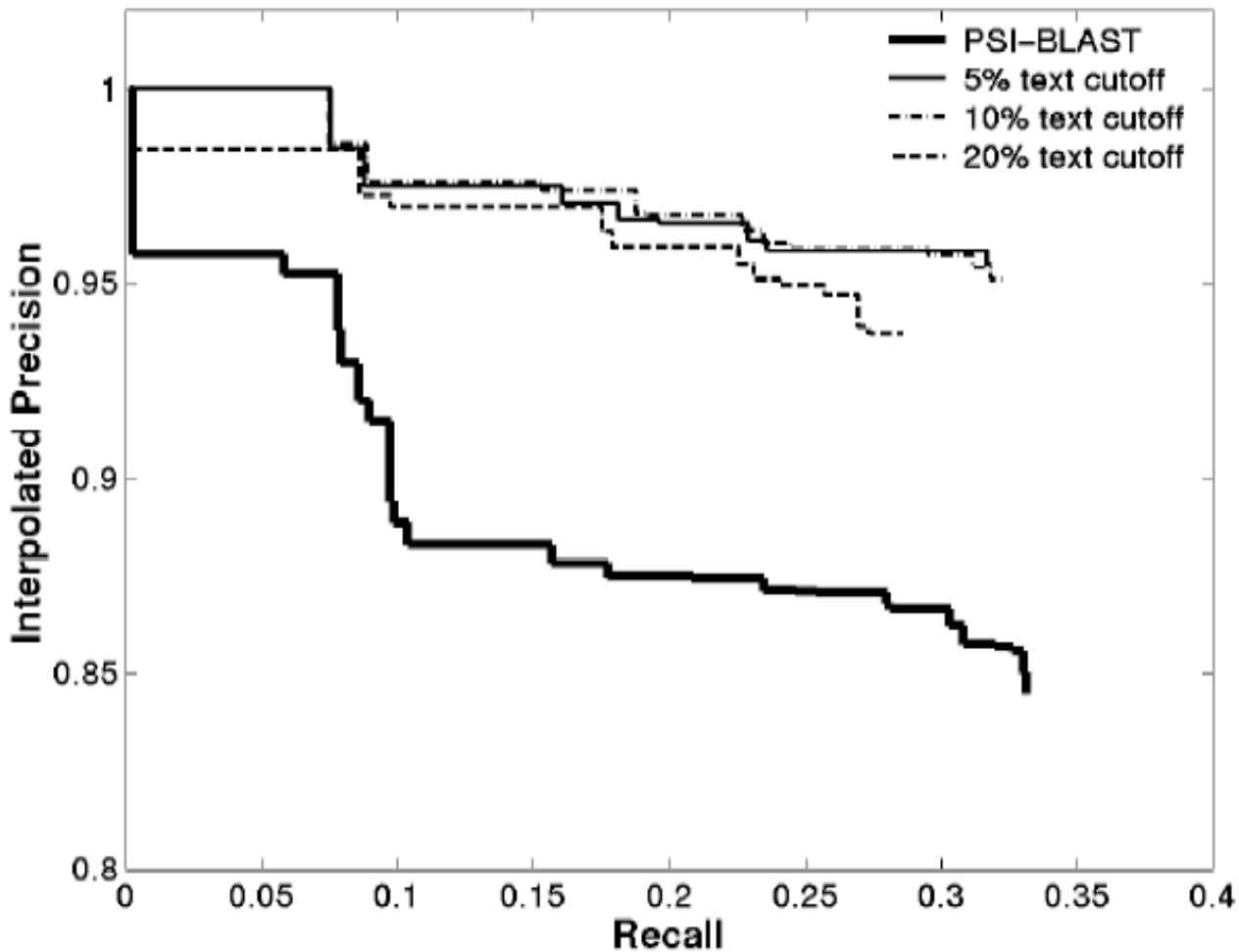
PSI-BLAST with Text Experiment

- assembled 54 families of homologous protein sequences
  - all proteins within a family have same structural (superfamily) classification in SCOP database
  - sequences have at least 4 associated abstracts in MEDLINE; linked from SwissProt
  - one query sequence per family – the most divergent member of each family

- measured precision/recall for these queries using
  - PSI-BLAST
  - **text**-PSI-BLAST with different literature similarity thresholds

Augmenting PSI-BLAST with Text

[Chang et al., *PSB* ’01]



Most of the difference is explained by a few families for which ordinary PSI-BLAST didn’t converge, but **text**-PSI-BLAST did