Text Analytics at the Bench

Brian Bartell, PhD

ICIC Nimes, Oct 2005

Text Analytics at the Bench

- Resolve the Disconnect between Literature Study and the Bench
  - Integrate Literature Search and Analysis with the daily research activity
  - Literature analysis embedded within scientific protocols
  - Increase diversity of literature scientist is exposed to
- Literature serves at the Bench as:
  - Hypothesis Generation
  - Result Annotation
  - Extra-modal Result Enrichment

Objectives

- Familiar Framework: Pipelining
  - Issue: converting raw scientific data in to qualified Literature queries
- Scientists, not Literature Scientists
- Encapsulate engine-specific details
  - Universal Query Language
  - Common document model
- Simultaneously:
  - Ready-to-run Protocols
  - Build-it-yourself Components
  - 3rd party integration

Text Analytics for Pipeline Pilot

Notiora

➢Method highlights

- One Query Language simultaneously searches

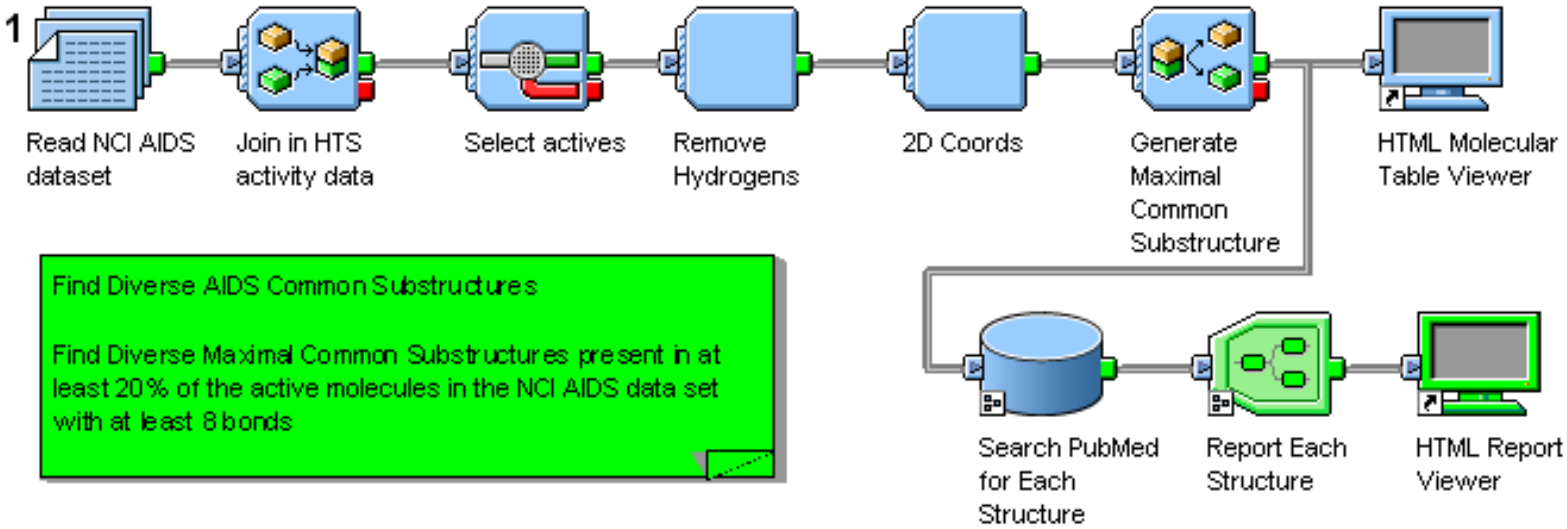
PubMed Google ?

- Discover trends and correlations between your experimental data and the scientific literature
- Extract key concepts (compounds, diseases) from text
- Both: Ready-to-run protocols and open platform

Example 1: AIDs MCSSs in PubMed

- GOAL
  - Enhance understanding of MCSS Analysis
- METHOD
  - Consider active compounds in the '97 NCI AIDS dataset
  - Find Diverse Maximal Common Substructures (MCSSs)
  - Find literature referencing compounds with substructure
  - Measure Antagonists & Inhibitors in literature
- RESULT
  - MCSSs are enriched with relevant literature context

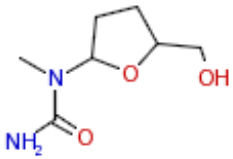
AIDs MCSSs in PubMed : Protocol



AIDs MCSSs in PubMed : Results

Notiora

MCS Substructure



Frequency: 75

Related Documents

92 Documents

Example Document

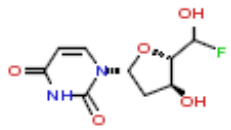
Thymidylate synthase inhibition triggers glucose-dependent apoptosis in p53-negative leukemic cells.

FEBS Lett 2004:

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=retrieve&db=pubmed&dopt=Citation&list_uids=15251465

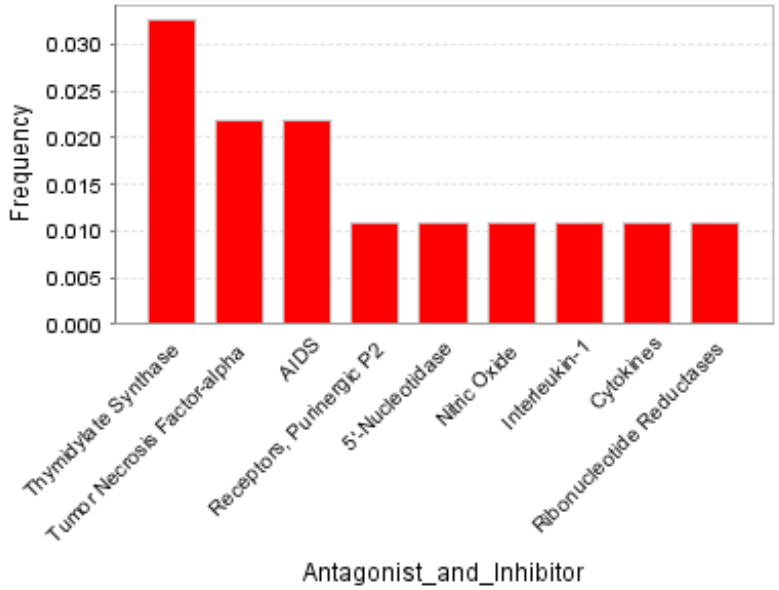
Chemotherapeutic drugs that inhibit the synthesis of DNA precursor thymidine triphosphate cause apoptosis, although the mechanism underlying this process remains rather unknown. Here, we describe thymineless death of human myeloid leukemia U937 cells treated with the thymidylate-synthase inhibitor 5'-fluoro- 2'-deoxyuridine (FUdR). This apoptotic process was shown to be independent of p53, reactive oxygen species generation and CD95 activation. Caspases were activated downstream of cytochrome c but upstream of mitochondrial depolarization. Furthermore, FUdR-induced apoptosis required the presence of glucose in the culture medium at a step upstream of the release of cytochrome c from mitochondria.

Example Molecule



5'-fluoro- 2'-deoxyuridine

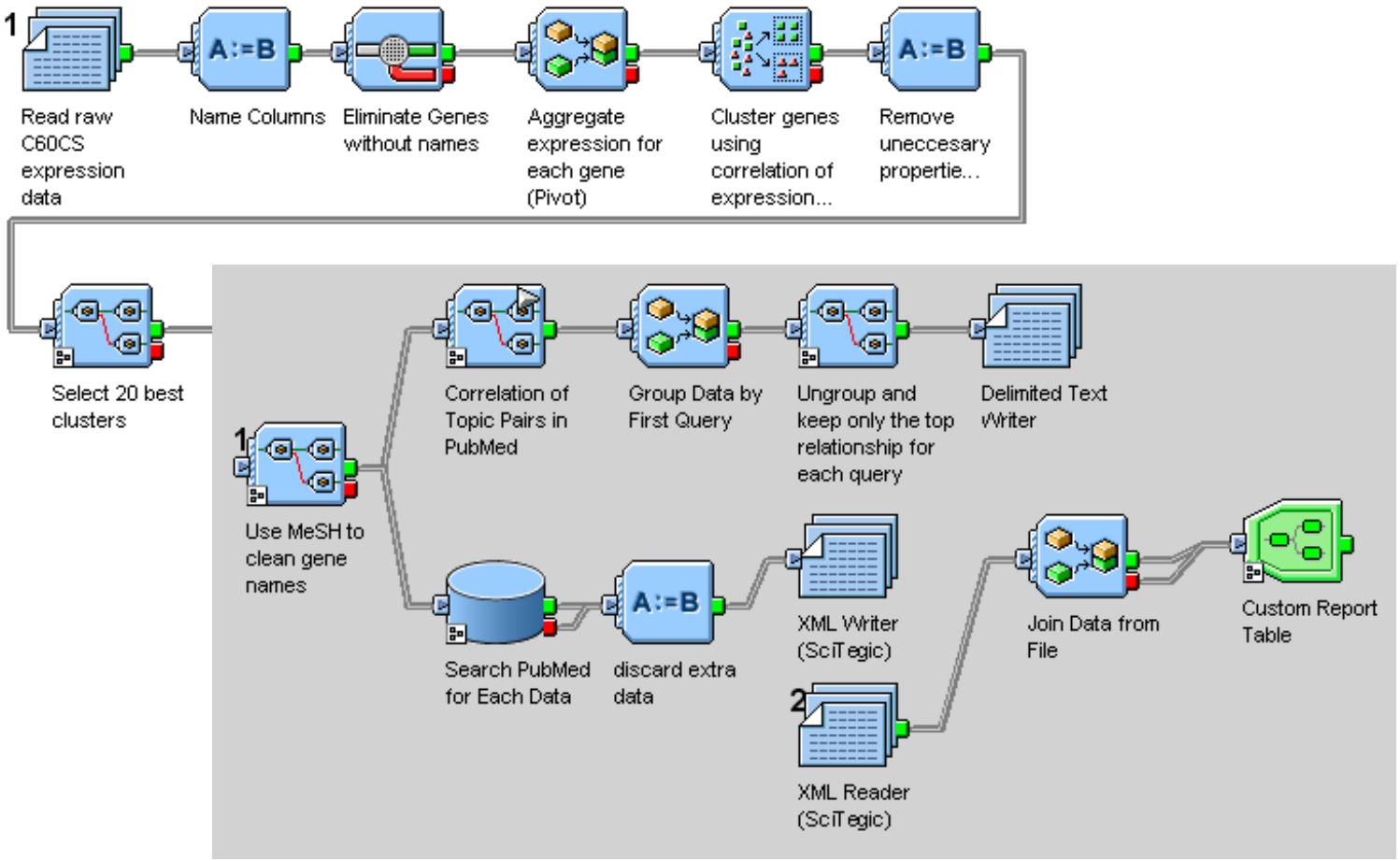
Antagonist & Inhibitor Activity



Example 2: Validate Gene Clusters

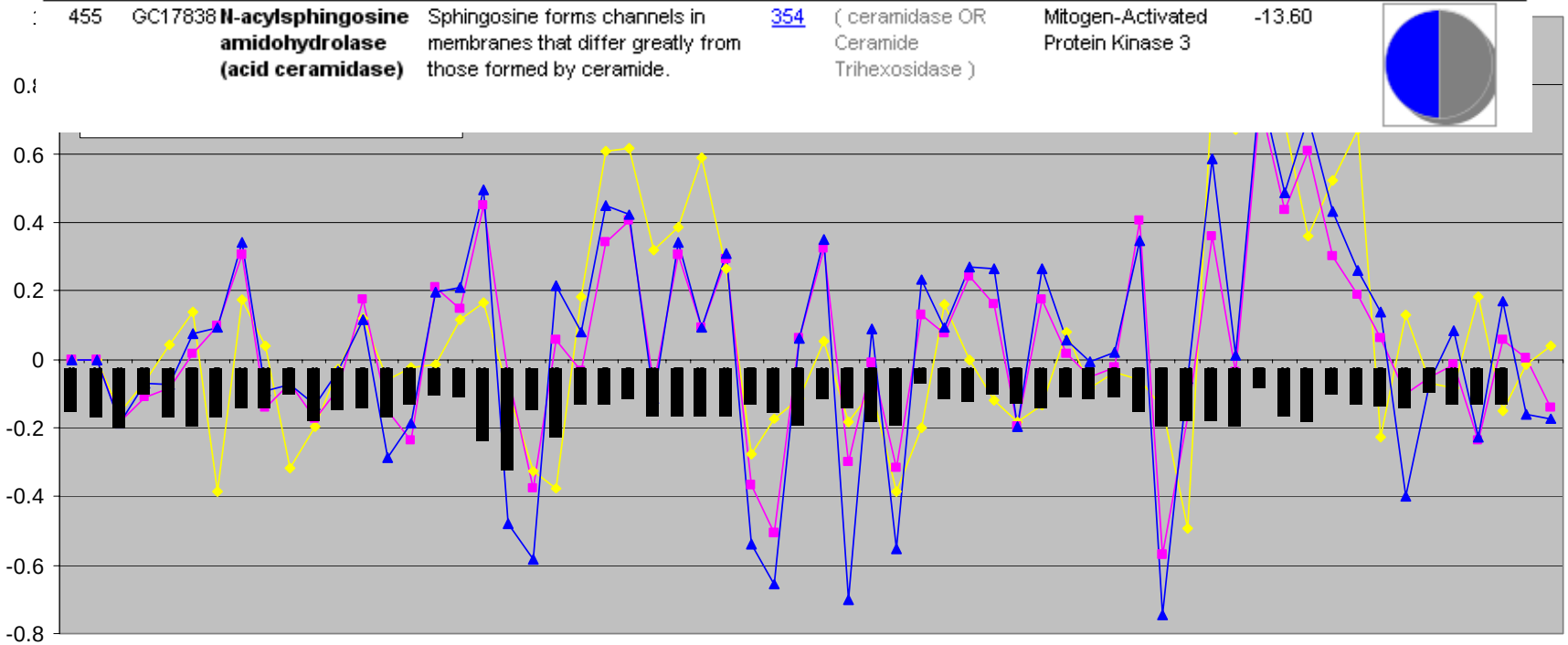
- GOAL
  - Validate gene clustering via gene correlation in PubMed
- METHOD
  - Cluster genes by expression patterns in 60 cancer cell lines
  - Validate names of genes using MeSH
  - Calculate RMI for all pairs of genes contained within each cluster
- RESULT
  - Identify some clusters as *valid*, *poor*, or potentially novel
  - For example, fatty-acid-Coenzyme A ligase, long chain 1

Validate Gene Clusters : Protocol



Validate Gene Clusters : Results

Cluster	Gene ID	Name	Title of First Article in PubMed	Articles	Validated Query	Most Related to	Best RMI	IntraCluster Hits
455	GC19166	**mitogen-activated protein kinase 3**	Inhibition of lipopolysaccharide-induced tissue factor expression in monocytes by urinary trypsin inhibitor in vitro and in vivo.	4538	Mitogen-Activated Protein Kinase 3	(ceramidase OR Ceramide Trihexosidase)	-13.60	
455	GC19189	**fatty-acid-Coenzyme A ligase, long-chain 1**	Blood cell gene expression profiling in rheumatoid arthritis. Discriminative genes and effect of rheumatoid factor.	8	fatty-acid-Coenzyme A ligase, long-chain 1	_	-	
455	GC17838	**N-acylsphingosine amidohydrolase (acid ceramidase)**	Sphingosine forms channels in membranes that differ greatly from those formed by ceramide.	354	(ceramidase OR Ceramide Trihexosidase)	Mitogen-Activated Protein Kinase 3	-13.60	



Validate Gene Clusters : Results

Notiora

Cluster	Gene ID	Name	Title of First Article in PubMed	Articles	Validated Query	Most Related to	Best RMI	IntraCluster Hits
727	GC18754	caveolin 2	Caveolin isoform expression during differentiation of C6 glioma cells.	161	caveolin-2	caveolin-1	-7.38	
727	GC9844	caveolin 1	Overexpression of PKCepsilon sensitizes LNCaP human prostate cancer cells to induction of apoptosis by bryostatin 1.	1411	caveolin-1	caveolin-2	-7.38	
727	GC19031	calpain 2, (m/II) large subunit	Genetic and Nongenetic Regulation of CAPN10 mRNA Expression in Skeletal Muscle.	4019	Calpain	Annexin A2	-13.88	

Caveolin-1 is not required for murine intestinal cholesterol transport.

Valasek MA, Weng J, Shaul PW, Anderson RG, Repa JJ.

Department of Physiology, The University of Texas Southwestern Medical Center at Dallas, Dallas, Texas 75390, USA.

Caveolin-1 (CAV1) is the structural protein of the filamentous coat that decorates the cytoplasmic surface of each caveola. Cell culture studies have implicated CAV1 in playing an important role in intracellular cholesterol trafficking. In addition, it has been reported that CAV1 forms a detergent-resistant protein complex with Annexin-2 in enterocytes that can be disrupted by the cholesterol absorption inhibitor ezetimibe, suggesting a possible role for CAV1 in cholesterol absorption. In this report, we have evaluated cholesterol homeostasis in Cav1 knock-out mice. Deletion of CAV1 does not result in either a compensatory increase of CAV2 or CAV3 in intestine. In addition, Cav1 knock-out mice display normal mRNA and protein levels of Annexin-2 or the putative cholesterol transport protein Niemann-Pick C1-like 1 (NPC1L1) in proximal intestinal mucosa. Fractional cholesterol absorption and fecal neutral sterol excretion are statistically similar in Cav1 knock-out mice and their wild-type littermates. Moreover, oral administration of ezetimibe is equally effective in decreasing cholesterol absorption in Cav1 null mice and wild-type controls. The mRNA expression levels of genes sensitive to intracellular cholesterol concentration (ATP-binding cassette transporters ABCA1 and ABCG5, hydroxymethylglutaryl-CoA synthase and the LDL receptor) are similarly altered in the proximal intestinal mucosa of Cav1 null and wild-type mice following ezetimibe treatment. These results demonstrate that CAV1 is not required for cholesterol absorption or ezetimibe sensitivity in the mouse.

PMID: 15919660 [PubMed - indexed for MEDLINE]

727	GC15931	regulator of G-protein signalling 12	Signalling mechanisms regulating lipolysis.	28849	GTP-Binding Proteins	caveolin-1	-13.43	
727	GC10437	annexin A2	CD44 and Annexin A2 Mediate the C5a Chemotactic Cofactor Function of the Vitamin D Binding Protein.	436	Annexin A2	caveolin-1	-10.93	

Summary

- Target Research Scientist in daily practice

- Examples illustrate how Literature serves as:
  - Result Annotation
  - Extra-modal Result Enrichment
  - Hypothesis Generation

Notiora

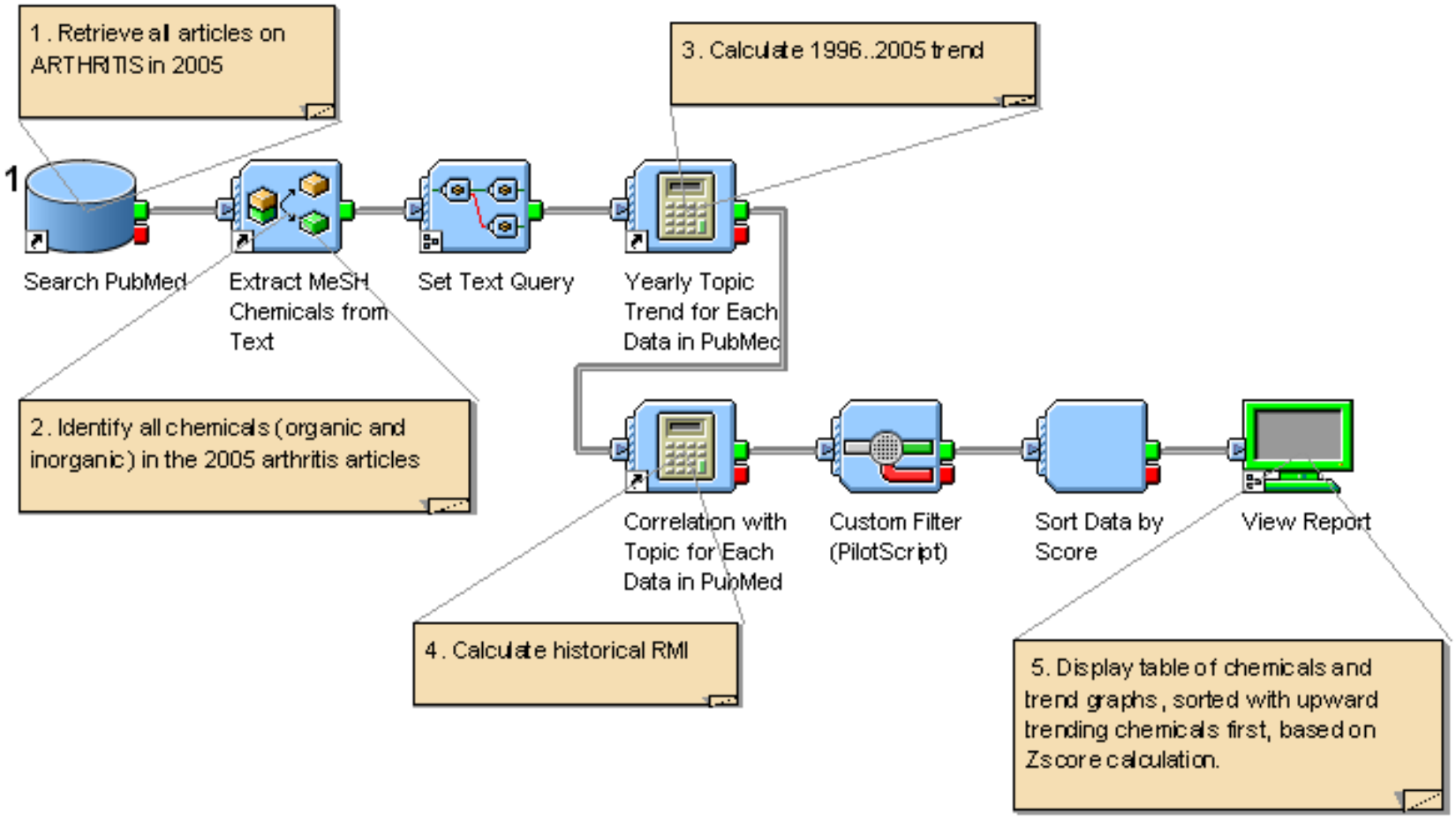
sci·tegic

*ask **more** of your **data***

Example 3: Chemical Trends in Arthritis

- GOAL
  - Identify molecules newly discussed in an Arthritis context
- METHOD
  - Scan 10-year Arthritis literature for all molecules
  - Determine correlation of molecules to Arthritis topic
  - Discover recently up-trending molecules
- RESULT
  - Metformin (oral hypoglycemic) merits further investigation

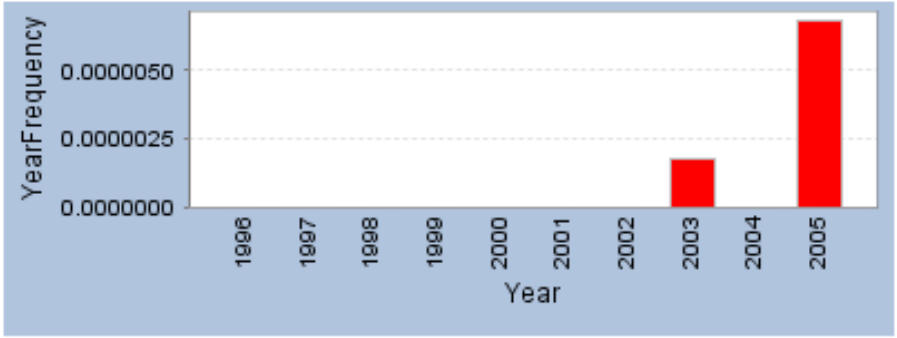
Chemical Trends in Arthritis : Protocol



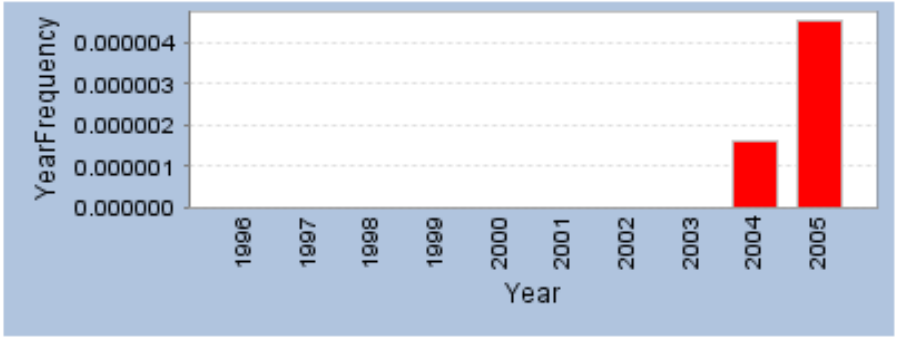
Chemical Trends in Arthritis: Results

Topic: arthritis chemicals

Trend	Chemical	
Yearly entity AND arthritis citation count normalized by year count	**Name**	Nordihydroguaiaretic Acid
	2005 Count (in arthritis)	3
	RMI (all years)	-17.92
	Zscore (most recent year)	2.76



Yearly entity AND arthritis citation count normalized by year count	**Name**	Metformin
	2005 Count (in arthritis)	2
	RMI (all years)	-18.595
	Zscore (most recent year)	2.67



Yearly entity AND arthritis citation count normalized by year count	**Name**	candesartan
	2005 Count (in arthritis)	2
	RMI (all years)	-18.405
	Zscore (most recent year)	2.67