Facilitating Discovery Through Data Integration and Analysis

David J. States, M.D., Ph.D.

National Center for Integrative
Biomedical Informatics

Assimilating the Information Deluge

The peer-reviewed biomedical literature is the primary medium for reporting biomedical results

- Research reports, review articles, conference proceedings, etc.
- Important information is often available only in free text

NLP is relevant to:

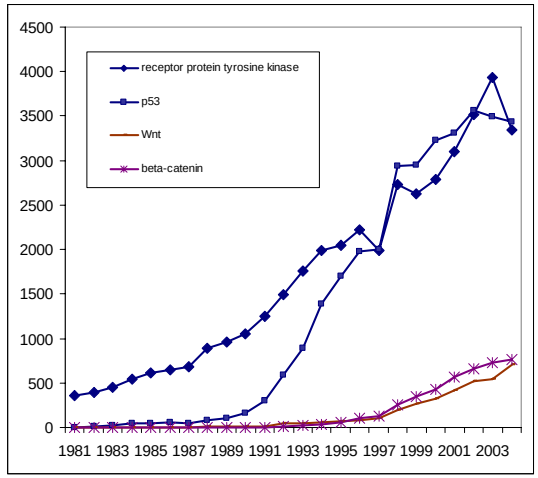
Information retrieval

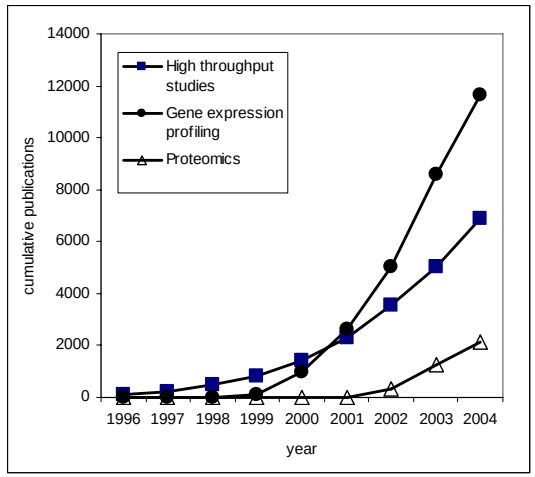
Data integration

- Defining context
- Provenance tracking and citation identification

Complement to high throughput biology

- Microarray analysis
- Proteomics
- Network analysis





Goals and Value Added for NLP

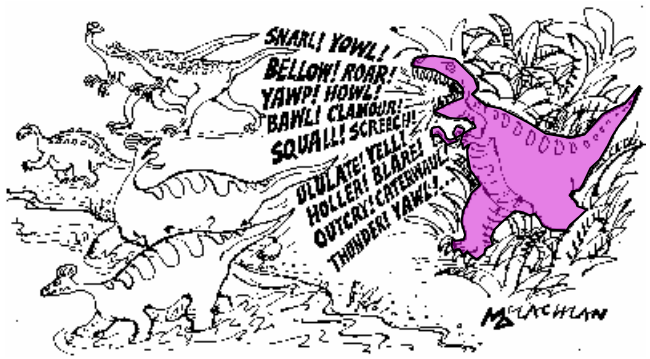
Capturing the complexity of biomedical science represented in the literature

- Pathways and interaction maps are a very simplistic view
- Accelerated access to full text
- Fine grained indexing and retrieval

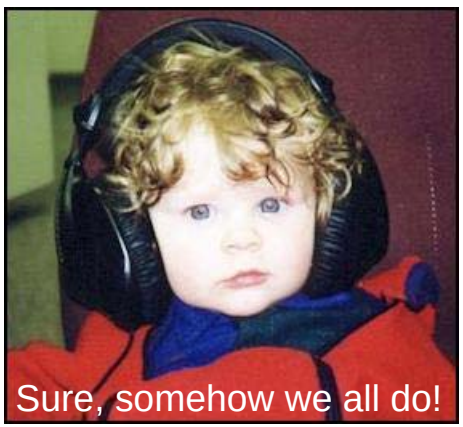
Analysis of complex processes

- WNT signaling mapping literature references ⇔ interactions
- Free text to machine readable and searchable formats
  - MarkerInfoFinder
  - SNP Annotaiton

NLP: Can Machines Read?



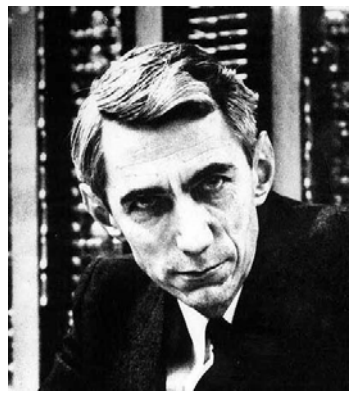
"Help! It's a thesaurus!"



Biomedical literature is complex
- Specialized vocabularies
- Complex sentence structures
- Complex and uncertain underlying knowledge

Literature is the knowledge repository
- Half century of failed attempts to get experts to use inputs that are convenient for the computer
- Tradition of peer reviewed literature is responsible for the scientific revolution
  - Did not begin with Gutenberg
  - Transaction of the Royal Society
    - Newton, Hooke, Raleigh, …

NCIBI goals
- Partnership and synergy with NCBO
- Take limited validated steps
  - Build increasingly sophisticated modes into information retrieval
  - Machine assisted knowledge extraction
- Build on successes in the NLP community
  - Named entity recognition and matching



Information Extraction vs Information Retrieval

Information Retrieval

- Article Retrieval (publishers)
- Term-Based Queries (e.g. Pubmed)
- Structured Databases (e.g. BIND)
- Canonical Resources (e.g. STKE)

Information Extraction and Analysis

- Database Integration
- Full and Partial Parsing
- Statistical Text Processing
- Assist Model Building (e.g. ODE)

Pilot Project: Wnt Signal Pathway Reconstruction

- full parse vs. human expert curation
- good performance, can we expand it?

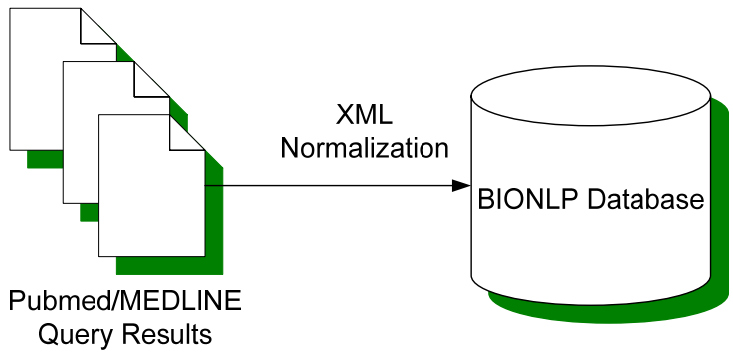
NLP Pipeline

NLP Pipeline Overview
- MS-SQL Server
- XML-based documents
- Tabular data for names and individual sentences

User-selected queries against Pubmed:
- “prostate cancer”
- “prostatic neoplasms”[MH]
- “androgen receptor”
- “wnt and beta-catenin”

Provenance on parse and terms
Sentence-level subject-verb-object tuples
Named entities resolved against databases

For each document:
- Perl, Minipar, TGrep2, tidy => normalized XML
- Split to document sections
- Split sections to paragraphs and sentences
- Named entity tagging and resolution
- Parse sentences with Minipar
- sentences individually assigned ID
- Extract with Tgrep2 to: SUBJECT-VERB-OBJECT tuples



Challenges in Biomedical NLP

Complex

- Very large vocabulary
  - More than a million gene names and synonyms
- Long sentences with complex structure
  - Many parsers literally fail

Bottom up

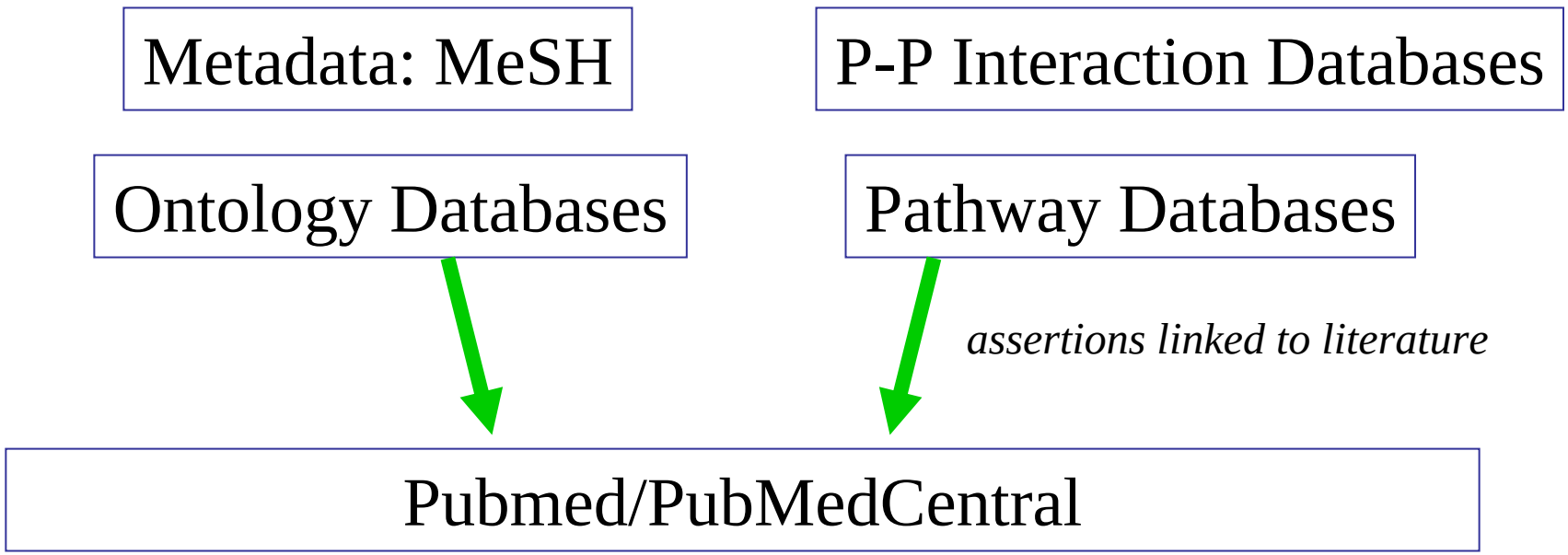
- Name collisions
  - PCR => phosphocreatine (and premature contraction)
- Inconsistent and domain specific definitions
- Too many ontologies

Available Resources

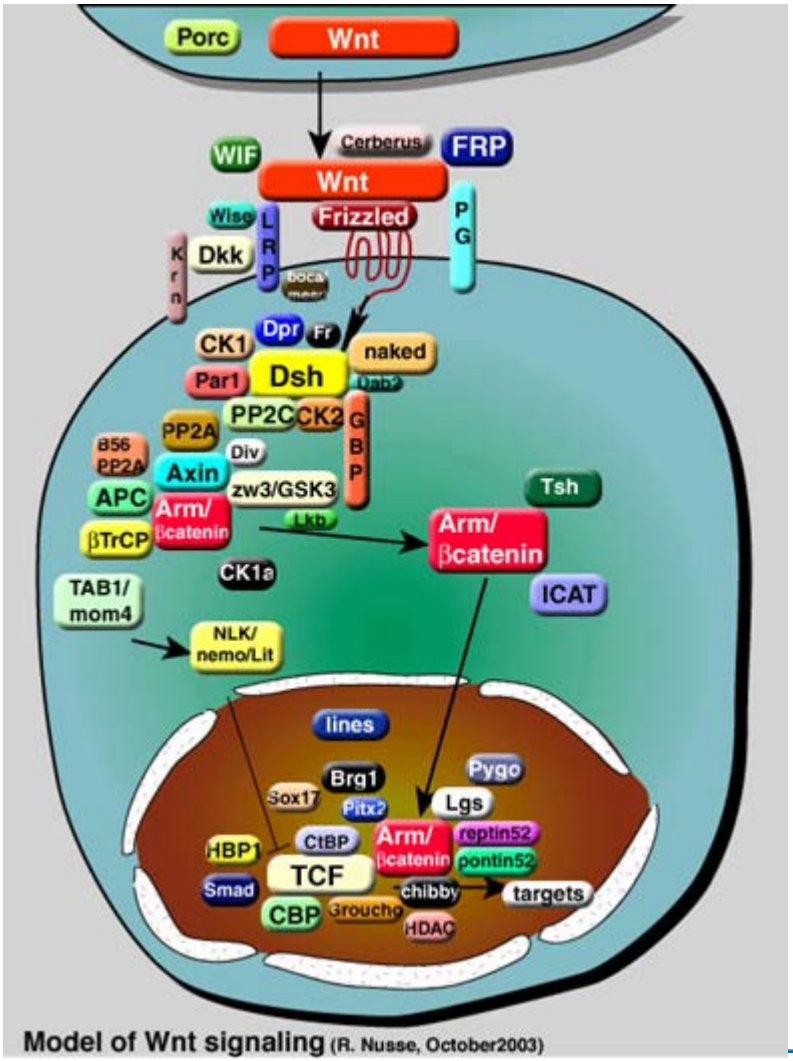
Focus on

protein-protein interactions

protein-gene interactions



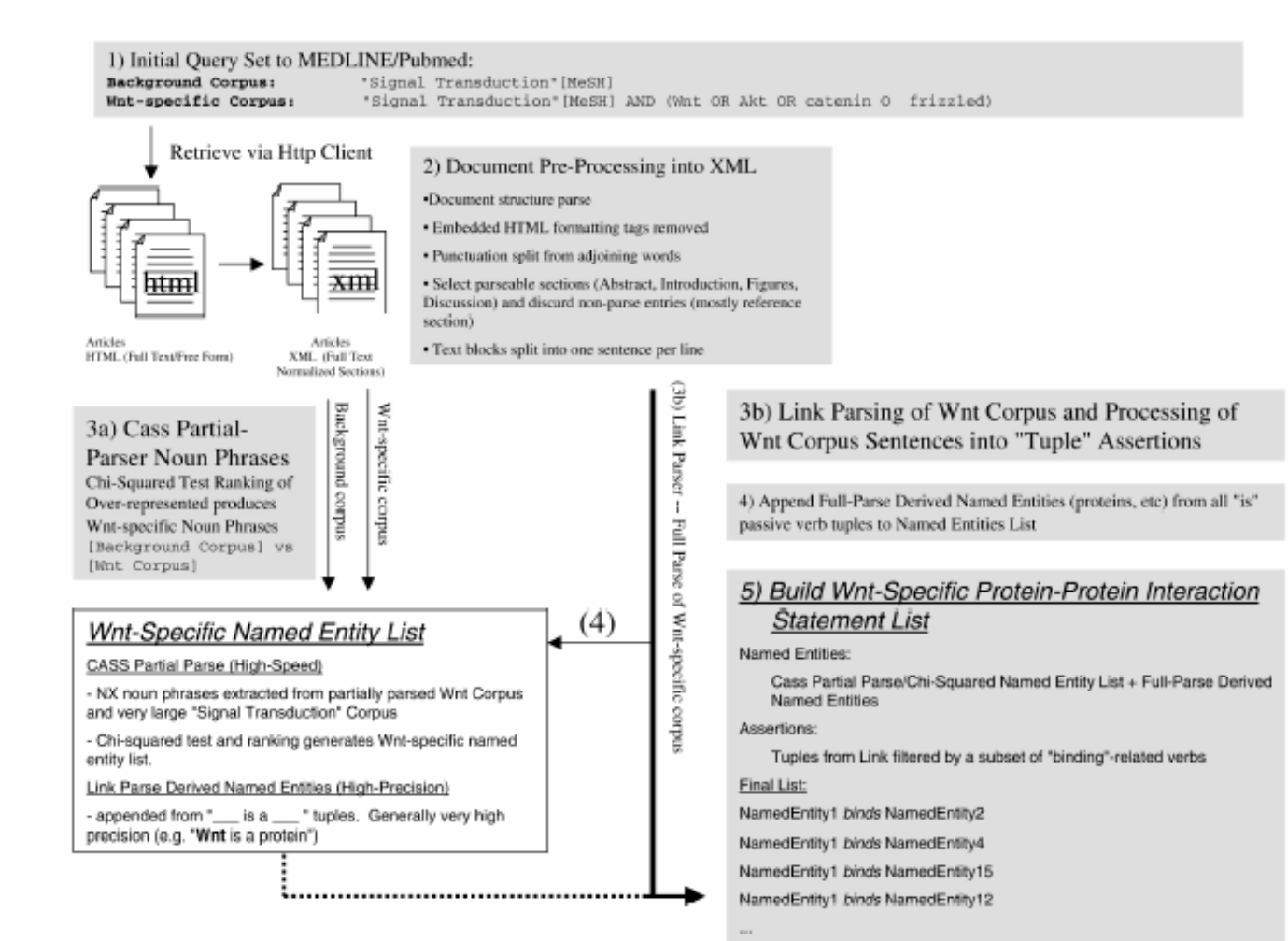
Wnt Pathway Project: Human Curation vs NLP



Model of Wnt signaling (R. Nusse, October2003)

Cerberus -> Wnt Wnts 10067895
WIF <-> Wnt 10201374
Dickkopf Dkk <-> LRP 11357136 11433302 11448771
Dickkopf Dkk <-> Kremen Krn 11357136
Wise <-> LRP 12900447
Wnt <-> Frizzled 8717036
Wnt <-> FRP Frp 8717036
LRP <-> Wnt Wnts 11029006 11029007 11029008
LRP <-> boca mesd 12581525 12581524
Proteoglycans PG<-> Wnt 2158444
Dishevelled Dishevelled dishevelled disheveled Dsh Dvl <-> CK1e CKI 105176 3210535959
Dishevelled Dishevelled dishevelled disheveled Dsh Dvl Dvl <-> CK2 CKII 9214626 12700239
Dishevelled Dishevelled dishevelled disheveled Dsh Dvl <-> GBP Frat1 Frat-1 10428961 10882137 10684251
Dishevelled Dishevelled dishevelled disheveled Dsh Dvl <-> Par-1 11433294
Dishevelled Dishevelled dishevelled disheveled Dsh Dvl <-> PP2C 10644691
PP2C <-> Axin 10644691
Dishevelled Dishevelled dishevelled disheveled Dsh Dvl <-> Frodo 11941372
Dishevelled Dishevelled dishevelled disheveled Dsh Dvl <-> naked cuticle gene naked 10693810 11274052
Dishevelled Dishevelled dishevelled disheveled Dsh Dvl <-> Axin 10329628 10882137 9920888
Dishevelled Dishevelled dishevelled disheveled Dsh Dvl <-> Dapper Dpr 11970895
Dishevelled Dishevelled dishevelled disheveled Dsh Dvl <-> Disabled-2Dab-2
Disabled2 Dab2 12805222
Disabled-2Dab-2 Dab2 Disabled2 <-> Axin 12805222
LKB1 XEEK1 <-> GSK 12973359
Armadillo beta-catenin <-> zw3 GSK-3b GSK3 GSK3beta 9554852 9601644 10073940
11927557 12000790
Armadillo beta-catenin <-> Casein Kinase 1 casein kinase 1 CK1a CKI CKIalpha 955485
2 9601644 10073940 11927557 12000790
Armadillo beta-catenin <-> APC 9554852 9601644 10073940
Armadillo beta-catenin <-> Axin 9554852 9601644 10073940
Armadillo beta-catenin <-> Slimb b-TrCP 9461217 9784611 10072378
Axin <-> PP2A 9920888
Axin <-> LRP 11336703
Axin <-> GSK-3b GSK3 GSK3beta 9482734 9501208 9601644
Axin <-> APC 9482734 9501208 9601644
PP2A <-> APC 10092233
Axin <-> Diversin 12183362
beta-catenin <-> TCF 0000000
TCF <-> Groucho 9783586
Groucho <-> HDAC 10485845
beta-catenin <-> Legless Bcl9 11955446 11967528 12015286
beta-catenin <-> Pygopus pygopus pygo 11955446 11967528 12015286
beta-catenin <-> Chibby 12712206
TCF <-> CBP P300 9774110 10775268 10769018
beta-catenin <-> Pitx2 12464179
beta-catenin <-> Brg-1 11532957
beta-catenin <-> Pontin52 Pontin pontin 11080158
beta-catenin <-> Reptin52 reptin Reptin 11080158
beta-catenin <-> XSox17 10549281
beta-catenin <-> Smad4 10693808
TCF <-> CtBP 10375506
TCF <-> HBP1 11500377
TCF <-> Lit1 NLK Nemo 10380924 10391247 10391246
Lit1 NLK Nemo <-> TAB1 TAK1 MOM-4 10380924 10391247 10391246
Teashirt Tsh <-> beta-catenin 10205174
beta-catenin <-> ICAT 10898789

Overview of NLP for Wnt Signaling



Link Interaction Detection

Total manually sample counted	370
Total Gold Standard Associations Detected	31 of 53 (58%)
Parse/Extract Precision Total correct (direct+indirect, ignoring name errors):	344 of 370 (92%)
Parse/Extract Recall with respect to Gold Standard Wnt Signaling Review Derived Set	31/53 (58%)
Separate Unique Interactions (overall)	1176
Separate Unique With Correct Name Recognition	1043

Variations on a Name: NFκB

Query: NF-kappa B
8000 abstracts
2000 full text

154361	nf- kappab
15507	nf-kappab
12744	nf kappab
8586	nf-kappa b
1904	nfkappab
871	nf- kappab

Solution: Regular expressions?
It works for PreBIND!

But…

Nuclear Factor kappa B
kappa B Enhancer Binding Protein
Immunoglobulin Enhancer-Binding Protein
Enhancer-Binding Protein, Immunoglobulin
Immunoglobulin Enhancer Binding Protein
Transcription Factor NF-kB
Factor NF-kB, Transcription
NF-kB, Transcription Factor
Transcription Factor NF kB
Ig-EBP-1
Ig EBP 1
NF-kB
NF kB
NFkB

Gene Name Tagging

Domain specific dictionary
- Identify species from MeSH annotations
- Build gene name table based on species

Efficient suffix tree algorithm
- Million gene names and synonyms
- Case dependent and independent matching

Resolving ambiguities
- Neighboring term frequency based classifier

Human “gene rifs” analyzed
95,214

Tagged
82,587

Correct gene
76,447

Precision/accuracy = 92.6%

Recall = 80.3%

And…

545,540 tags including multiple occurrences and many more genes

Text Tagging and Indexing

Text Processing Pipeline

1. Query
2. Pubmed search
3. Document retrieval
4. Conversion to XML
5. Document structure parse
6. Sentence splitting
7. Named entity tagging
8. Named entity resolution
9. Deep parsing

Documents	121,899
Sentences	2,167,762
Cell lines	216,504
Gene names	1,139,220
Mesh heading	7,450,689
Substance	496,518

Biomedical text is information rich!

Graphical Summarization of Complex Data in the Biomedical Literature

A single paper references dozens of genes and hundreds of gene to gene relationships

Graphical representation

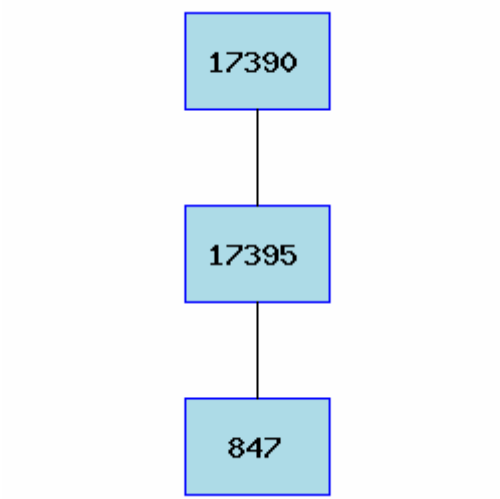
- Genes as nodes
- Gene to gene relationships as edges

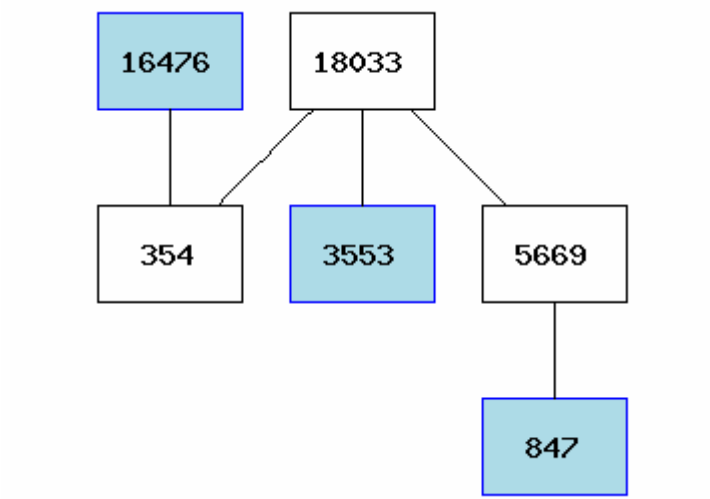
Applications

- Powerful visual interface
- Unification across the center (Concept maps,)
- Accessible to computational search (SAGA)

Graph Based Literature Search

Full text named entity tagging of the query and target
Nodes => genes, Edges => sentences referring to a pair of genes

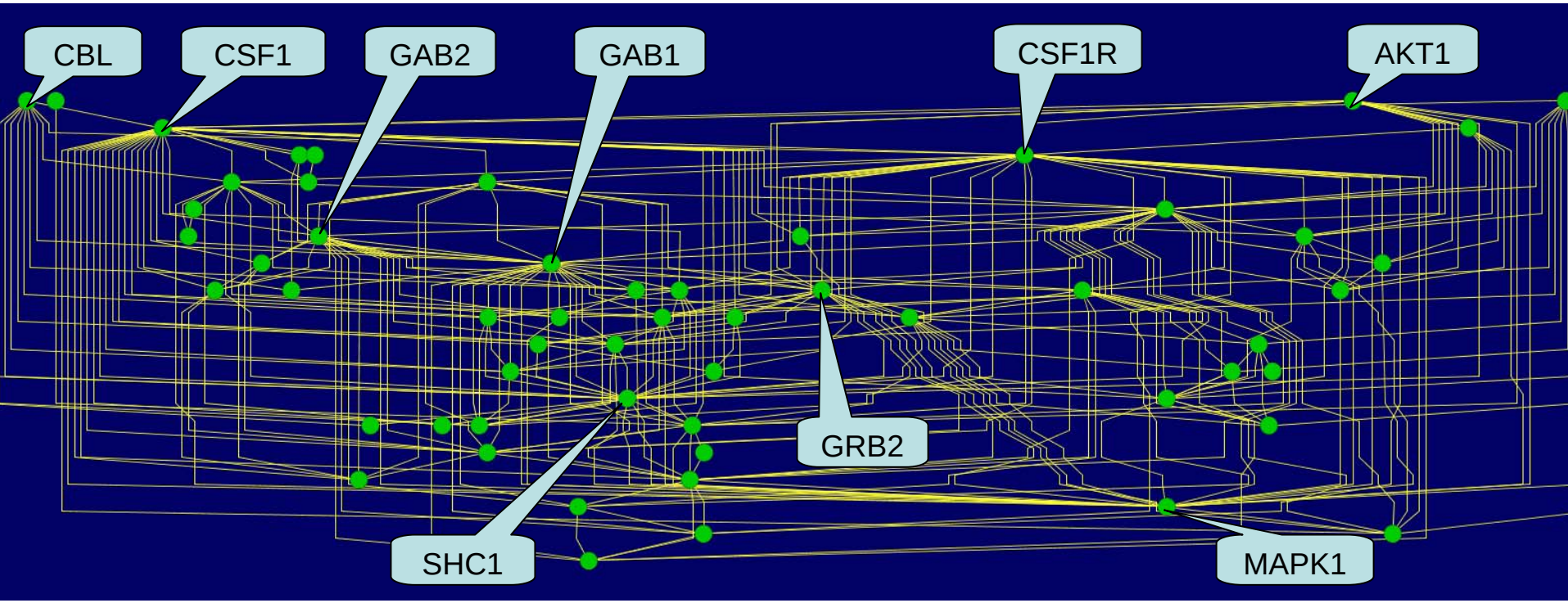




Subgraph matching 3 nodes and 2 edges

Subgraph matching 3 nodes but no edges

Graphical Text Summarization

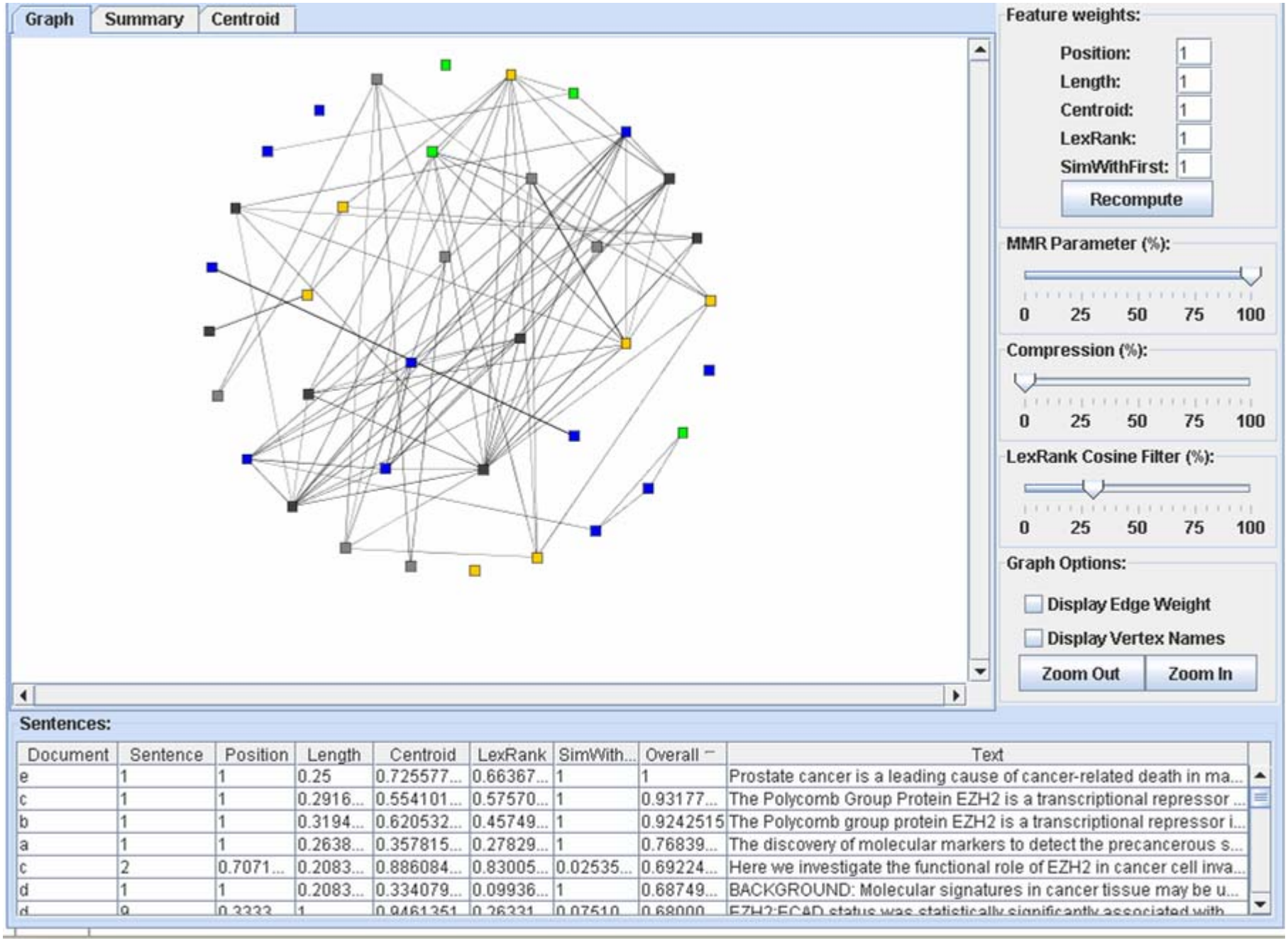


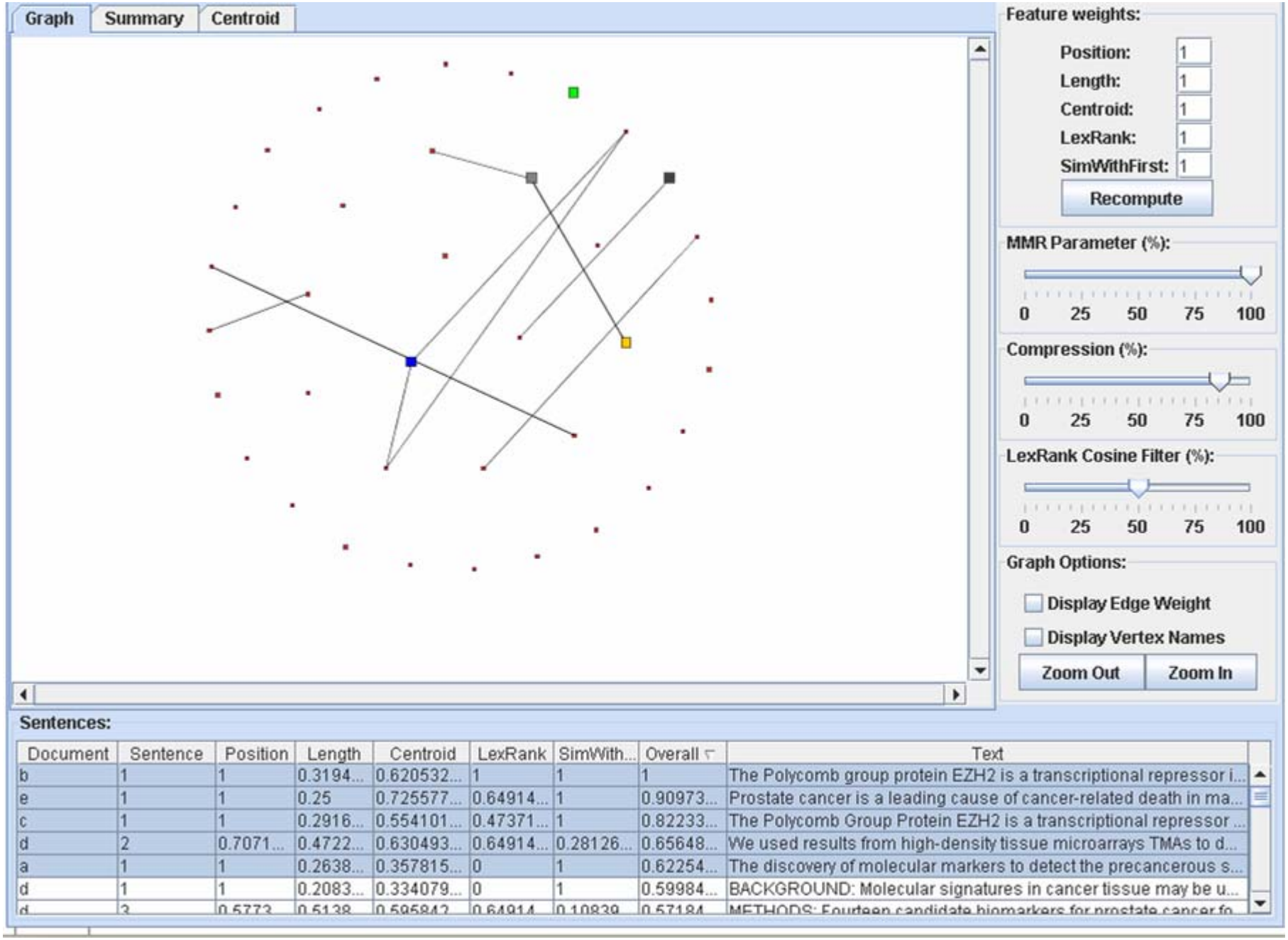
Nodes => genes Edges => sentences referring to multiple genes

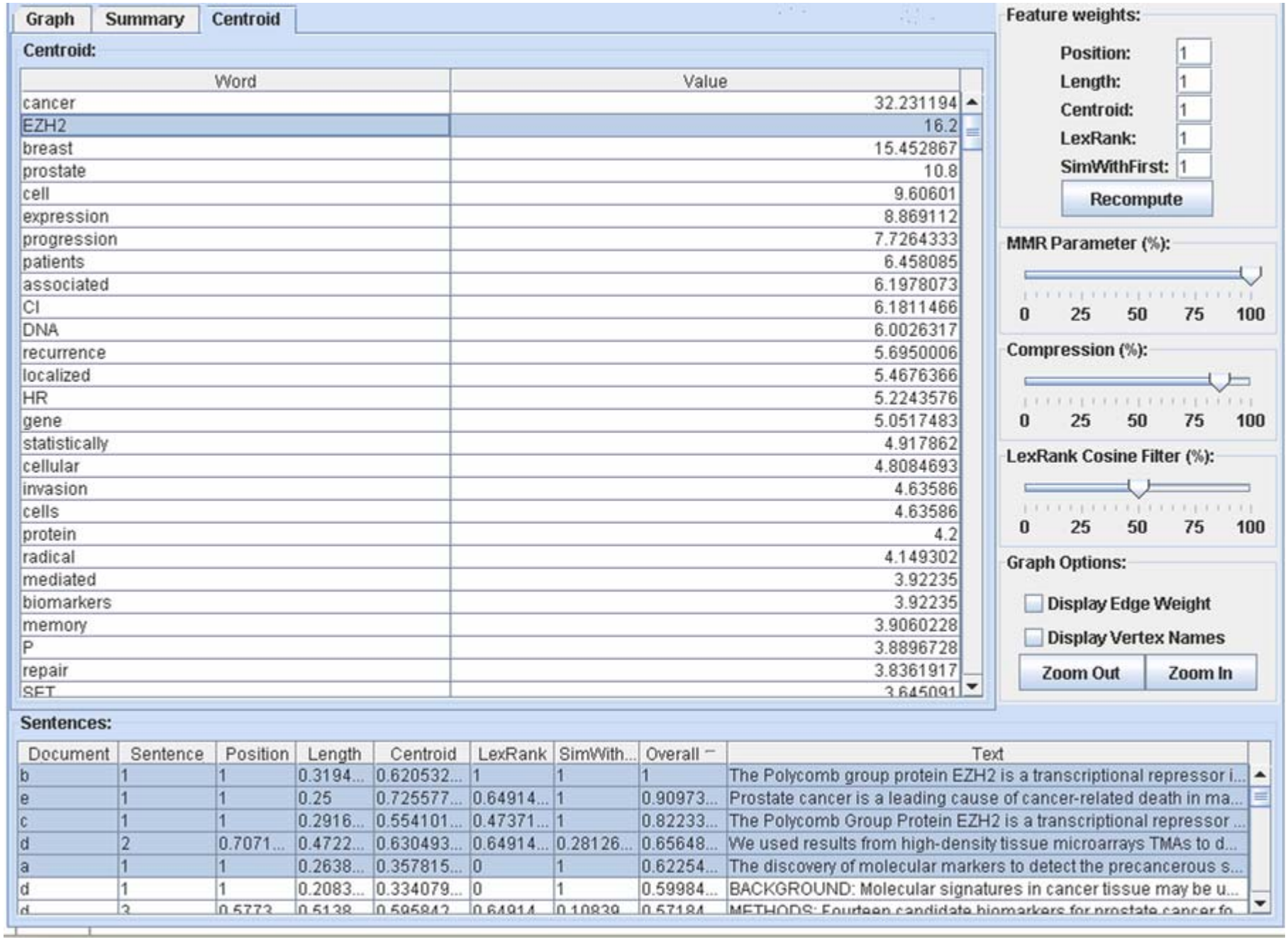
Genes and relationships in Lee AW, States DJ (2000) Mol Cell Biol. 2000 Sep;20(18):6779-98.

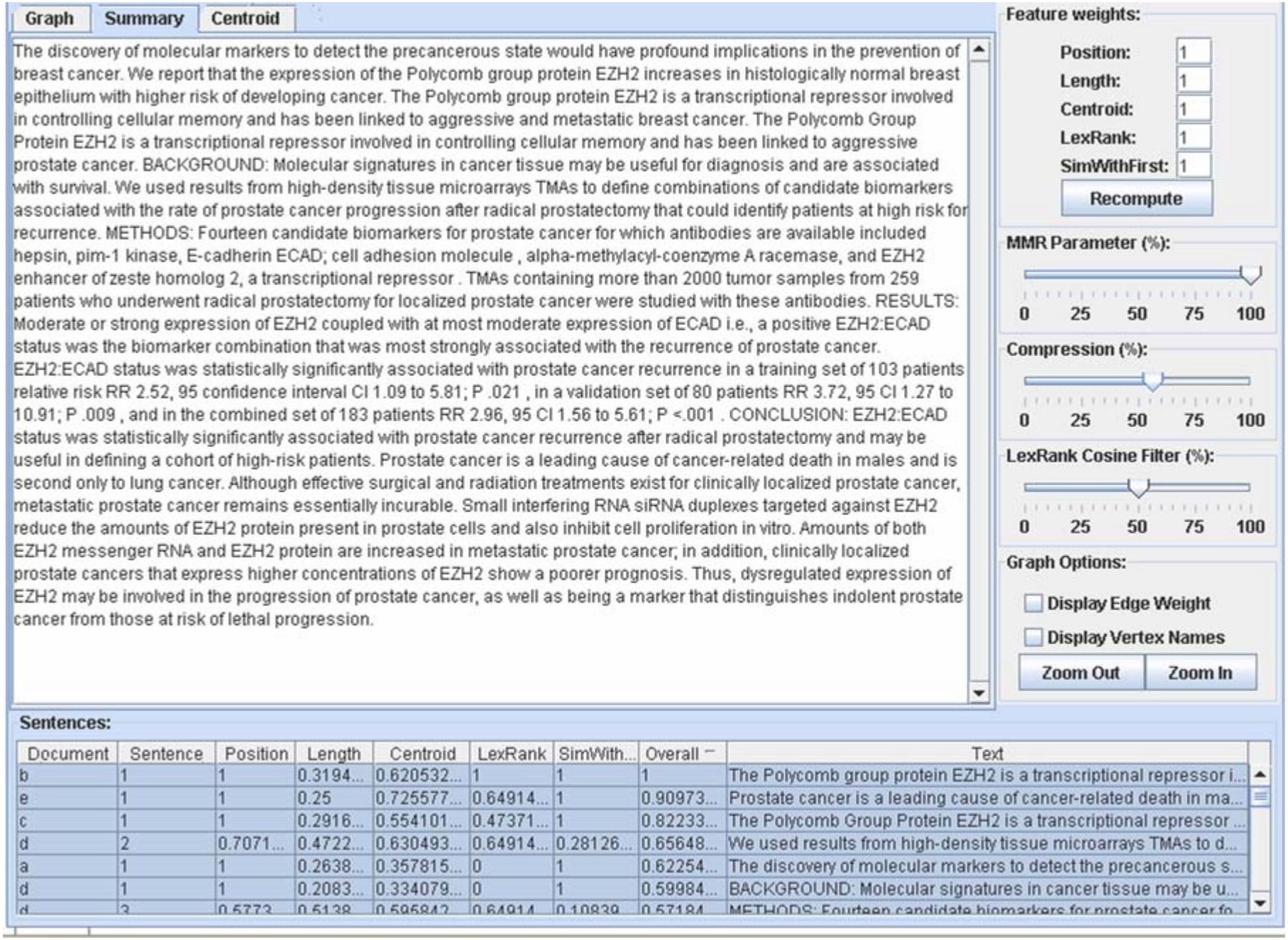
Document and Multi-document Summarization

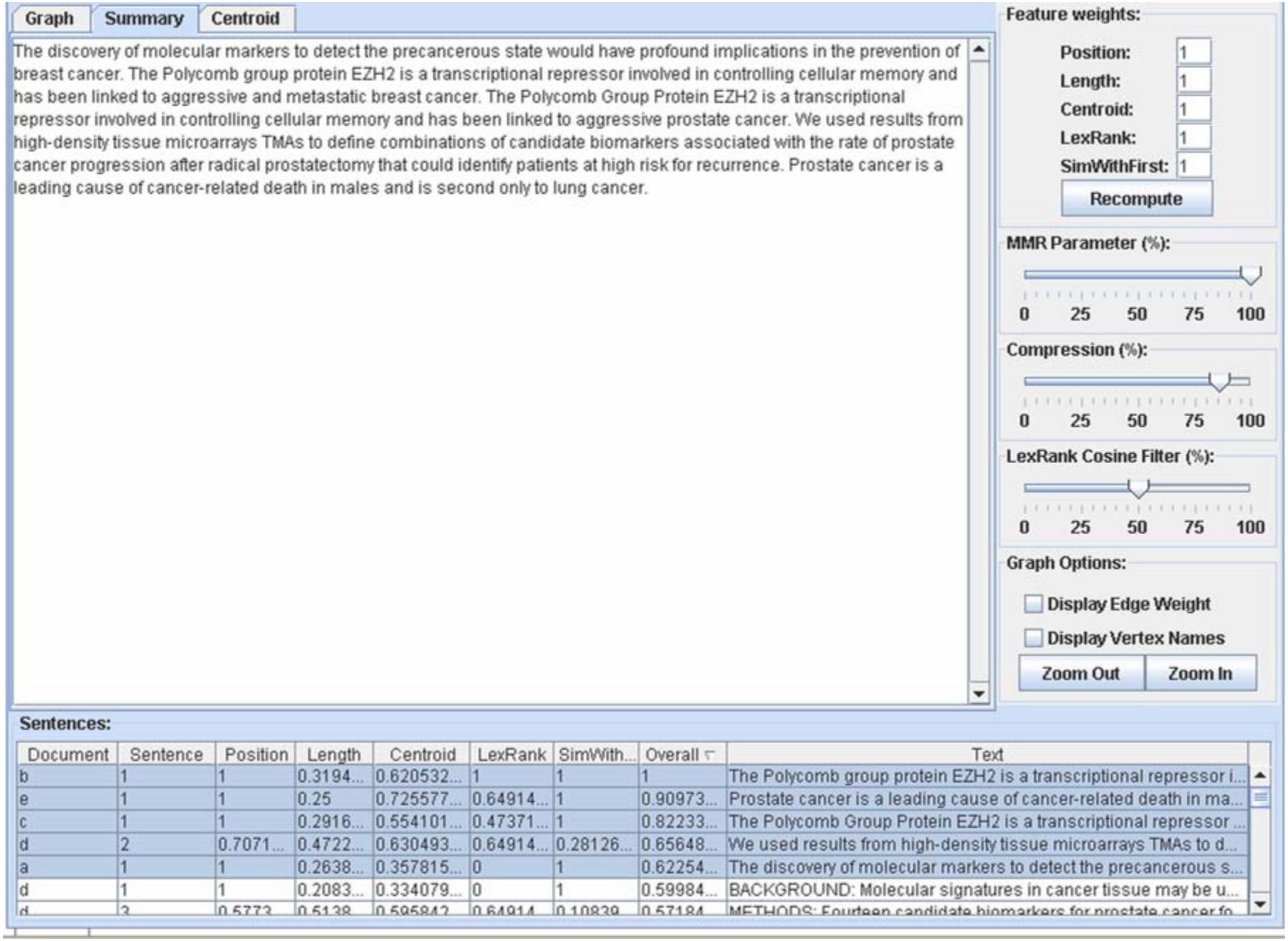
- Complex task with many applications
  - One shoe is not going to fit all feet
- Graphical description of information relationships within one or more document(s)
- Provide multiple measures of term and sentence similarity
  - Position in the document
  - Lexical similarity (LexRank)
  - Centroid in the graph of sentence relationships
- Allow the user to interactively weight differement measures


Graph
Summary
Centroid
Feature weights:
Position: 1
Length: 1
Centroid: 1
LexRank: 1
SimWithFirst: 1
Recompute
MMR Parameter (%):
0 25 50 75 100
Compression (%):
0 25 50 75 100
LexRank Cosine Filter (%):
0 25 50 75 100
Graph Options:
Display Edge Weight
Display Vertex Names
Zoom Out
Zoom In
Sentences:
Document | Sentence | Position | Length | Centroid | LexRank | SimWith... | Overall | Text
e | 1 | 1 | 0.25 | 0.725577... | 0.66367... | 1 | 1 | Prostate cancer is a leading cause of cancer-related death in ma...
c | 1 | 1 | 0.2916... | 0.554101... | 0.57570... | 1 | 0.93177... | The Polycomb Group Protein EZH2 is a transcriptional repressor ...
b | 1 | 1 | 0.3194... | 0.620532... | 0.45749... | 1 | 0.9242515 | The Polycomb group protein EZH2 is a transcriptional repressor i...
a | 1 | 1 | 0.2638... | 0.357815... | 0.27829... | 1 | 0.76839... | The discovery of molecular markers to detect the precancerous s...
c | 2 | 0.7071... | 0.2083... | 0.886084... | 0.83005... | 0.02535... | 0.69224... | Here we investigate the functional role of EZH2 in cancer cell inva...
d | 1 | 1 | 0.2083... | 0.334079... | 0.09936... | 1 | 0.68749... | BACKGROUND: Molecular signatures in cancer tissue may be u...
d | 9 | 0.3333... | 1 | 0.9461351 | 0.26331... | 0.07510... | 0.68000... | EZH2:ECAD status was statistically significantly associated with

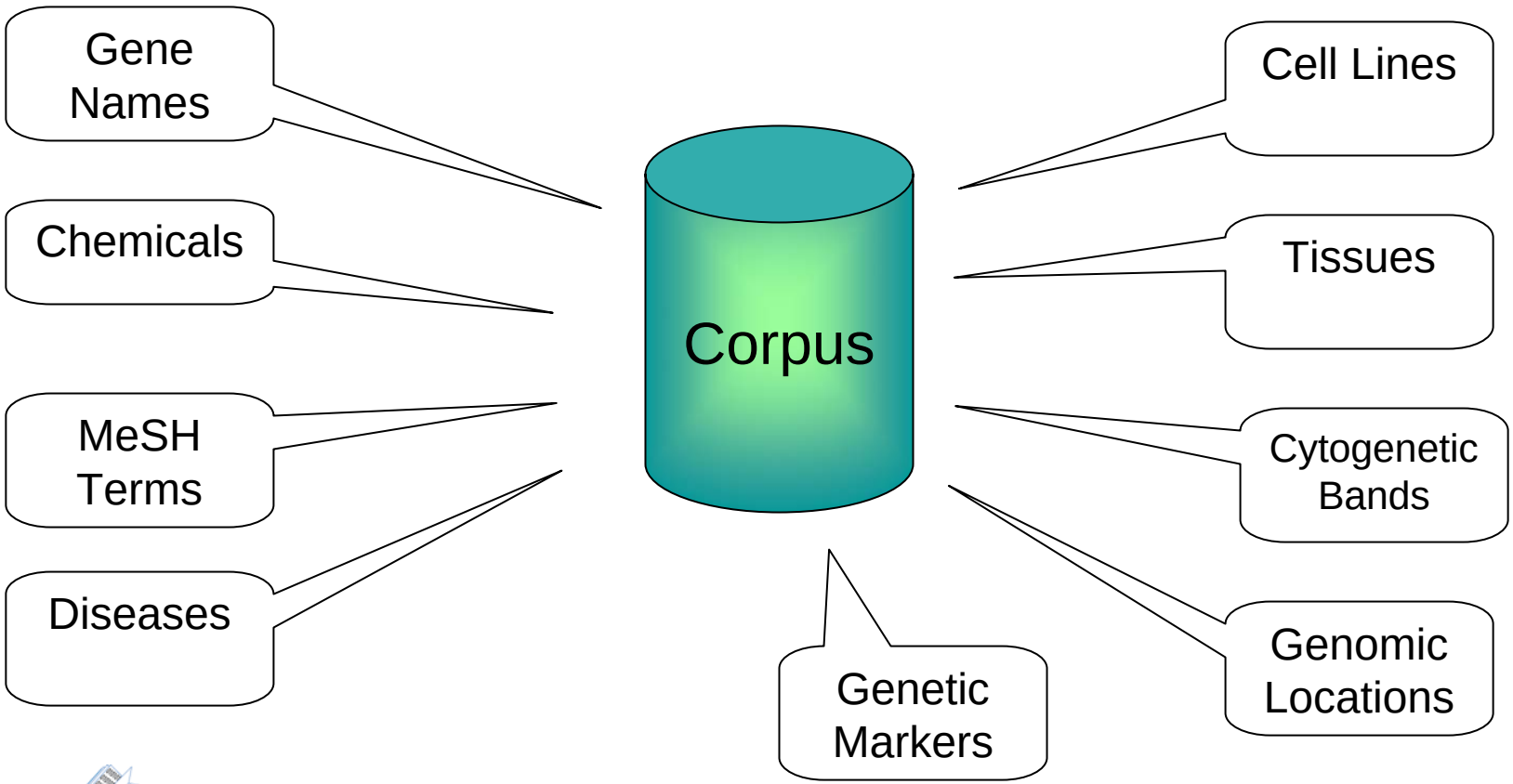

Graph
Summary
Centroid
Feature weights:
Position: 1
Length: 1
Centroid: 1
LexRank: 1
SimWithFirst: 1
Recompute
MMR Parameter (%):
0 25 50 75 100
Compression (%):
0 25 50 75 100
LexRank Cosine Filter (%):
0 25 50 75 100
Graph Options:
Display Edge Weight
Display Vertex Names
Zoom Out
Zoom In
Sentences:
Document | Sentence | Position | Length | Centroid | LexRank | SimWith... | Overall | Text
b | 1 | 1 | 0.3194... | 0.620532... | 1 | 1 | 1 | The Polycomb group protein EZH2 is a transcriptional repressor i...
e | 1 | 1 | 0.25 | 0.725577... | 0.64914... | 1 | 0.90973... | Prostate cancer is a leading cause of cancer-related death in ma...
c | 1 | 1 | 0.2916... | 0.554101... | 0.47371... | 1 | 0.82233... | The Polycomb Group Protein EZH2 is a transcriptional repressor ...
d | 2 | 0.7071... | 0.4722... | 0.630493... | 0.64914... | 0.28126... | 0.65648... | We used results from high-density tissue microarrays TMAs to d...
a | 1 | 1 | 0.2638... | 0.357815... | 0 | 1 | 0.62254... | The discovery of molecular markers to detect the precancerous s...
d | 1 | 1 | 0.2083... | 0.334079... | 0 | 1 | 0.59984... | BACKGROUND: Molecular signatures in cancer tissue may be u...


Graph
Summary
Centroid
Centroid:
Word
Value
cancer 32.231194
EZH2 16.2
breast 15.452867
prostate 10.8
cell 9.60601
expression 8.869112
progression 7.7264333
patients 6.458085
associated 6.1978073
CI 6.1811466
DNA 6.0026317
recurrence 5.6950006
localized 5.4676366
HR 5.2243576
gene 5.0517483
statistically 4.917862
cellular 4.8084693
invasion 4.63586
cells 4.63586
protein 4.2
radical 4.149302
mediated 3.92235
biomarkers 3.92235
memory 3.9060228
P 3.8896728
repair 3.8361917
SET 3.645091
Feature weights:
Position: 1
Length: 1
Centroid: 1
LexRank: 1
SimWithFirst: 1
Recompute
MMR Parameter (%):
0 25 50 75 100
Compression (%):
0 25 50 75 100
LexRank Cosine Filter (%):
0 25 50 75 100
Graph Options:
Display Edge Weight
Display Vertex Names
Zoom Out
Zoom In
Sentences:
Document Sentence Position Length Centroid LexRank SimWith... Overall Text
b 1 1 0.3194... 0.620532... 1 1 1 The Polycomb group protein EZH2 is a transcriptional repressor i...
e 1 1 0.25 0.725577... 0.64914... 1 0.90973... Prostate cancer is a leading cause of cancer-related death in ma...
c 1 1 0.2916... 0.554101... 0.47371... 1 0.82233... The Polycomb Group Protein EZH2 is a transcriptional repressor ...
d 2 0.7071... 0.4722... 0.630493... 0.64914... 0.28126... 0.65648... We used results from high-density tissue microarrays TMAs to d...
a 1 1 0.2638... 0.357815... 0 1 0.62254... The discovery of molecular markers to detect the precancerous s...
d 1 1 0.2083... 0.334079... 0 1 0.59984... BACKGROUND: Molecular signatures in cancer tissue may be u...
d 3 0.5773... 0.5138... 0.595842... 0.64914... 0.10839... 0.57184... METHODS: Fourteen candidate biomarkers for prostate cancer fo...


Graph
Summary
Centroid
The discovery of molecular markers to detect the precancerous state would have profound implications in the prevention of breast cancer. We report that the expression of the Polycomb group protein EZH2 increases in histologically normal breast epithelium with higher risk of developing cancer. The Polycomb group protein EZH2 is a transcriptional repressor involved in controlling cellular memory and has been linked to aggressive and metastatic breast cancer. The Polycomb Group Protein EZH2 is a transcriptional repressor involved in controlling cellular memory and has been linked to aggressive prostate cancer. BACKGROUND: Molecular signatures in cancer tissue may be useful for diagnosis and are associated with survival. We used results from high-density tissue microarrays TMAs to define combinations of candidate biomarkers associated with the rate of prostate cancer progression after radical prostatectomy that could identify patients at high risk for recurrence. METHODS: Fourteen candidate biomarkers for prostate cancer for which antibodies are available included hepsin, pim-1 kinase, E-cadherin ECAD; cell adhesion molecule , alpha-methylacyl-coenzyme A racemase, and EZH2 enhancer of zeste homolog 2, a transcriptional repressor . TMAs containing more than 2000 tumor samples from 259 patients who underwent radical prostatectomy for localized prostate cancer were studied with these antibodies. RESULTS: Moderate or strong expression of EZH2 coupled with at most moderate expression of ECAD i.e., a positive EZH2:ECAD status was the biomarker combination that was most strongly associated with the recurrence of prostate cancer. EZH2:ECAD status was statistically significantly associated with prostate cancer recurrence in a training set of 103 patients relative risk RR 2.52, 95 confidence interval CI 1.09 to 5.81; P .021 , in a validation set of 80 patients RR 3.72, 95 CI 1.27 to 10.91; P .009 , and in the combined set of 183 patients RR 2.96, 95 CI 1.56 to 5.61; P <.001 . CONCLUSION: EZH2:ECAD status was statistically significantly associated with prostate cancer recurrence after radical prostatectomy and may be useful in defining a cohort of high-risk patients. Prostate cancer is a leading cause of cancer-related death in males and is second only to lung cancer. Although effective surgical and radiation treatments exist for clinically localized prostate cancer, metastatic prostate cancer remains essentially incurable. Small interfering RNA siRNA duplexes targeted against EZH2 reduce the amounts of EZH2 protein present in prostate cells and also inhibit cell proliferation in vitro. Amounts of both EZH2 messenger RNA and EZH2 protein are increased in metastatic prostate cancer; in addition, clinically localized prostate cancers that express higher concentrations of EZH2 show a poorer prognosis. Thus, dysregulated expression of EZH2 may be involved in the progression of prostate cancer, as well as being a marker that distinguishes indolent prostate cancer from those at risk of lethal progression.
Feature weights:
Position: 1
Length: 1
Centroid: 1
LexRank: 1
SimWithFirst: 1
Recompute
MMR Parameter (%):
0 25 50 75 100
Compression (%):
0 25 50 75 100
LexRank Cosine Filter (%):
0 25 50 75 100
Graph Options:
Display Edge Weight
Display Vertex Names
Zoom Out
Zoom In
Sentences:
Document Sentence Position Length Centroid LexRank SimWith... Overall Text
b 1 1 0.3194... 0.620532... 1 1 1 The Polycomb group protein EZH2 is a transcriptional repressor i...
e 1 1 0.25 0.725577... 0.64914... 1 0.90973... Prostate cancer is a leading cause of cancer-related death in ma...
c 1 1 0.2916... 0.554101... 0.47371... 1 0.82233... The Polycomb Group Protein EZH2 is a transcriptional repressor ...
d 2 0.7071... 0.4722... 0.630493... 0.64914... 0.28126... 0.65648... We used results from high-density tissue microarrays TMAs to d...
a 1 1 0.2638... 0.357815... 0 1 0.62254... The discovery of molecular markers to detect the precancerous s...
d 1 1 0.2083... 0.334079... 0 1 0.59984... BACKGROUND: Molecular signatures in cancer tissue may be u...


Graph
Summary
Centroid
The discovery of molecular markers to detect the precancerous state would have profound implications in the prevention of breast cancer. The Polycomb group protein EZH2 is a transcriptional repressor involved in controlling cellular memory and has been linked to aggressive and metastatic breast cancer. The Polycomb Group Protein EZH2 is a transcriptional repressor involved in controlling cellular memory and has been linked to aggressive prostate cancer. We used results from high-density tissue microarrays TMAs to define combinations of candidate biomarkers associated with the rate of prostate cancer progression after radical prostatectomy that could identify patients at high risk for recurrence. Prostate cancer is a leading cause of cancer-related death in males and is second only to lung cancer.
Feature weights:
Position: 1
Length: 1
Centroid: 1
LexRank: 1
SimWithFirst: 1
Recompute
MMR Parameter (%):
0 25 50 75 100
Compression (%):
0 25 50 75 100
LexRank Cosine Filter (%):
0 25 50 75 100
Graph Options:
Display Edge Weight
Display Vertex Names
Zoom Out
Zoom In
Sentences:
Document Sentence Position Length Centroid LexRank SimWith... Overall Text
b 1 1 0.3194... 0.620532... 1 1 1 The Polycomb group protein EZH2 is a transcriptional repressor i...
e 1 1 0.25 0.725577... 0.64914... 1 0.90973... Prostate cancer is a leading cause of cancer-related death in ma...
c 1 1 0.2916... 0.554101... 0.47371... 1 0.82233... The Polycomb Group Protein EZH2 is a transcriptional repressor ...
d 2 0.7071... 0.4722... 0.630493... 0.64914... 0.28126... 0.65648... We used results from high-density tissue microarrays TMAs to d...
a 1 1 0.2638... 0.357815... 0 1 0.62254... The discovery of molecular markers to detect the precancerous s...
d 1 1 0.2083... 0.334079... 0 1 0.59984... BACKGROUND: Molecular signatures in cancer tissue may be u...
d 3 0.5773... 0.5138... 0.595842... 0.64914... 0.10839... 0.57184... METHODS: Fourteen candidate biomarkers for prostate cancer fo...

Named Entity Tagging



Overview: Data Resources

MarkerInfoFinder incorporates four major categories:

Genetic markers:

- SNP: dbSNP/SNP search web service
- STS/Microsatellite: UniSTS

Chromosome/Genomic Locations:

- Cytoband: extract from free text
- Genomic locations-based search

Gene/Probe:

- Batch sequence IDs: Gene, UniGene, GenBank, Affymetrix Probe.
- Gene/protein keyword search, to locate a set of genes.

Diseases:

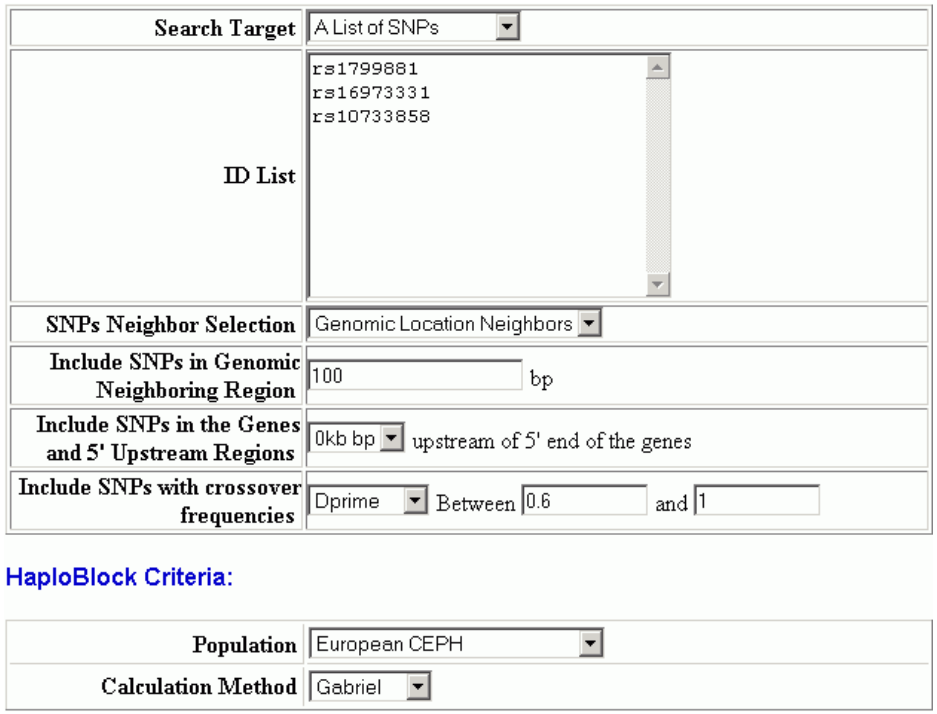
- Normalized names. OMIM, for human genetic inherited diseases.
- Supplement: UMLS (semantic: disease/syndrome), ICD

Genetic Marker Statistics

Gene/Protein initial list	881,089 unique terms 212,085 flexible patterns 576,286 strict named entity patterns
Word frequency statistics	556,974 filtered single words
STS name dictionary	924,302 STS, 454,439 unique
Medline citations	In our database: 15,572,691 citation, 8,018,148 abstracts
Detected STS	1,041,646 occurrences
Detected cytoband	248,048 occurrences
Identified gene/protein	Mapped to 22,257 unique Entrez Gene IDs

BrainArray: GeneInfoMiner

Search for SNP/STS/Microsatellite Rrelated Papers:



Genetic Marker to Literature Mapping

- Genomic Location View

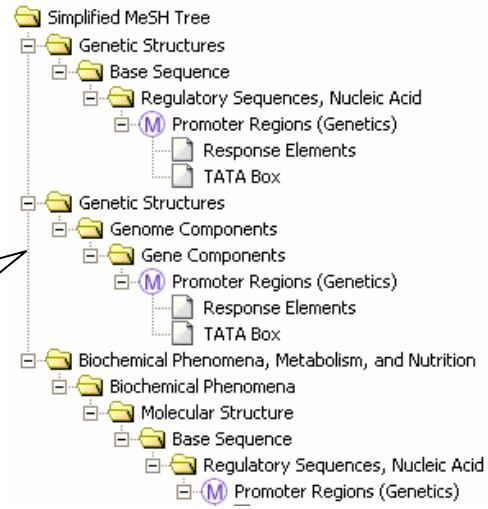
Chromosome	Location	SNP	Citaions	Publications
10	71656058	rs10733858	186	Ⓟ
16	70160166	rs16973331	147	Ⓟ

- MeSH group view

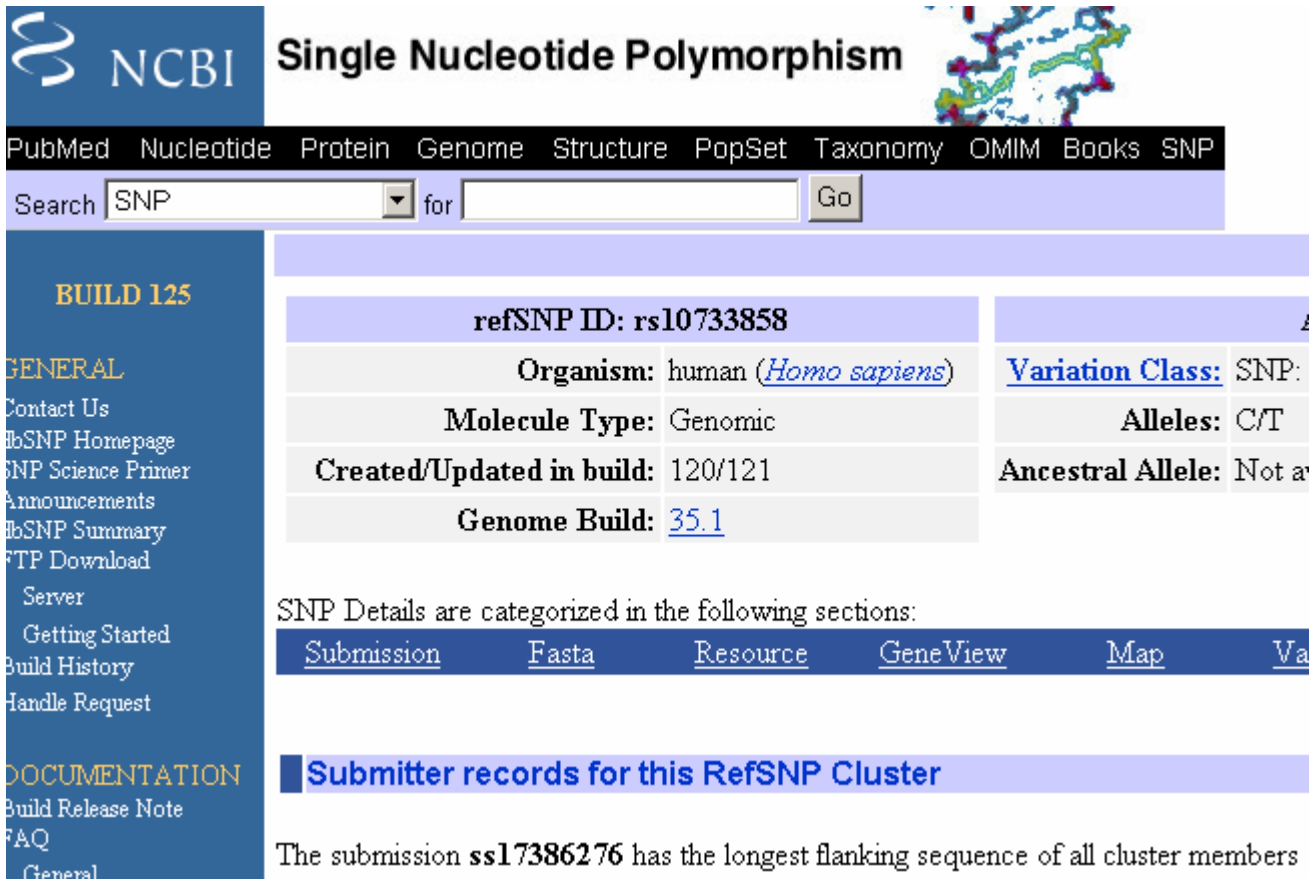
Genomic Location View | MeSH Group View

[Mapping results grouped by MeSH concepts.]

MeSH Heading	Papers	MeSH Tree	Publication
Transcription, Genetic	37	Ⓜ	Ⓟ
Gene Expression Regulation	28	Ⓜ	Ⓟ
Mutation	23	Ⓜ	Ⓟ
Promoter Regions (Genetics)	16	Ⓜ	Ⓟ



Links to External Database/Websites



Selecting Most Relevant Papers

Select Entity: rs2305110 Sort Abstracts By: Sort... Select Abstracts: Select

PMID	Year	Journal	# Marker	# Gene	Symbols	Title of Abstract	
3930961	1985	29.065	1	5	IGH@,IGHM, ..	Rearrangement of the T-cell receptor beta-chain gene in non-T-cell, non-B-cell acute lymphoblastic leukemia of childhood.	☑
9010145	1997	29.065	1	1	F5	The risk of recurrent venous thromboembolism in patients with an Arg506-->Gln mutation in the gene for factor V (factor V Leiden).	☐
11779510	2001	16.611	1	3	EEF2,GFM1,ERAL1 ..	The nucle(ol)ar Tif6p and Efl1p are required for a late cytoplasmic step of ribosome synthesis.	☑
10700252	2000	15.668	1	1	RHOA	Rho GTPases regulate distinct aspects of dendritic arbor growth in Xenopus central neurons in vivo.	☐
12426392	2002	12.459	1	17	XPOT,ERF,EEF2 ..	Exp5 exports eEF1A via tRNA from nuclei and synergizes with other transport pathways to confine translation to the cytoplasm.	☑
14673072	2003	10.896	1	4	ETF1,TrnM,EEF2 ..	Divergent tRNA-like element supports initiation, elongation, and termination of protein biosynthesis.	☐
9892677	1999	10.896	1	2	HPRT1,HPRT1	Gender-specific frequency of background somatic mutations at the hypoxanthine phosphoribosyltransferase locus in cord blood T lymphocytes from preterm newborns.	☑
8355680	1993	9.836	1	2	G6PD,NFKB1	Inducible transcriptional activation of the human immunodeficiency virus long terminal repeat by protein kinase inhibitors.	☐

Linking Through Literature

Examine genes in an abstract

Related Gene Information for Selected Citation

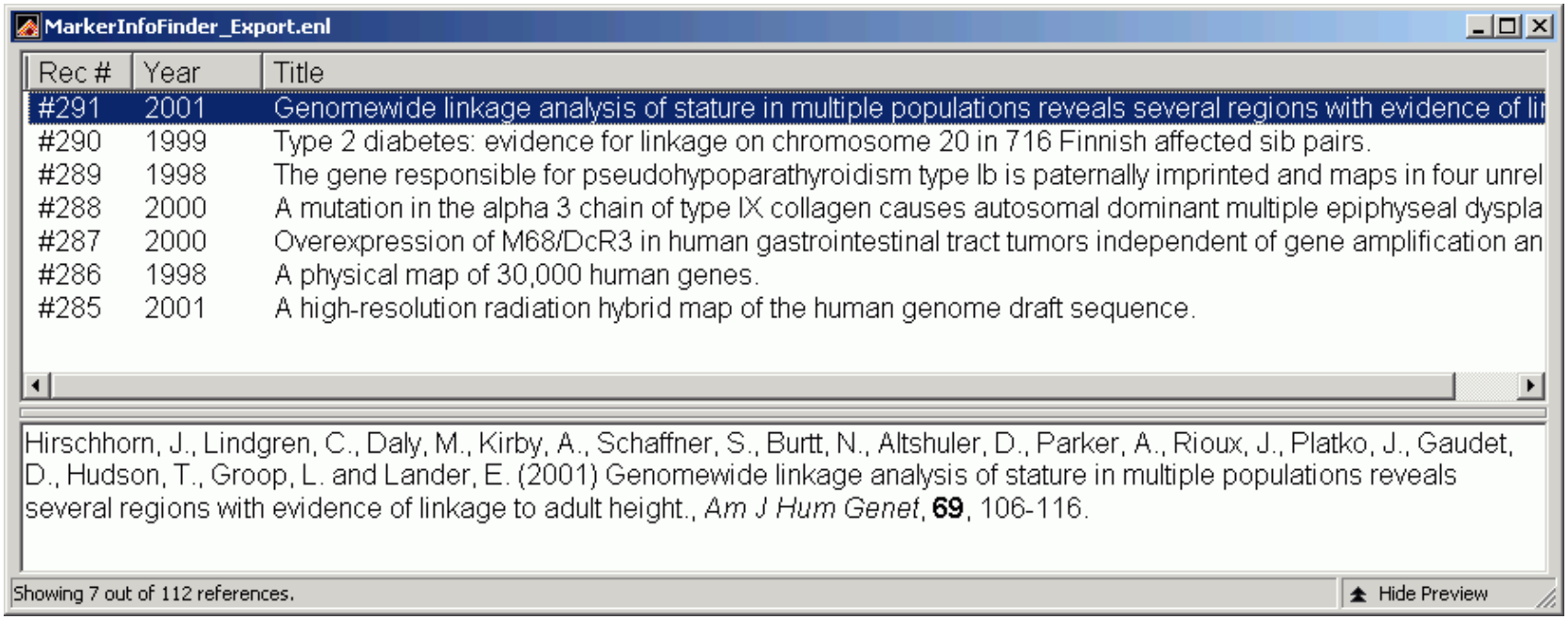
GeneID	InputID	Official Symbol	Official Name	EntrezGene	UniGene	GeneCards	Ontology
1938		EEF2	eukaryotic translation elongation factor 2	E	U	G	O
85476		GFM1	G elongation factor, mitochondrial 1	E	U	G	O
26284		ERAL1	Era G-protein-like 1 (E. coli)	E	U	G	O

Review journal information

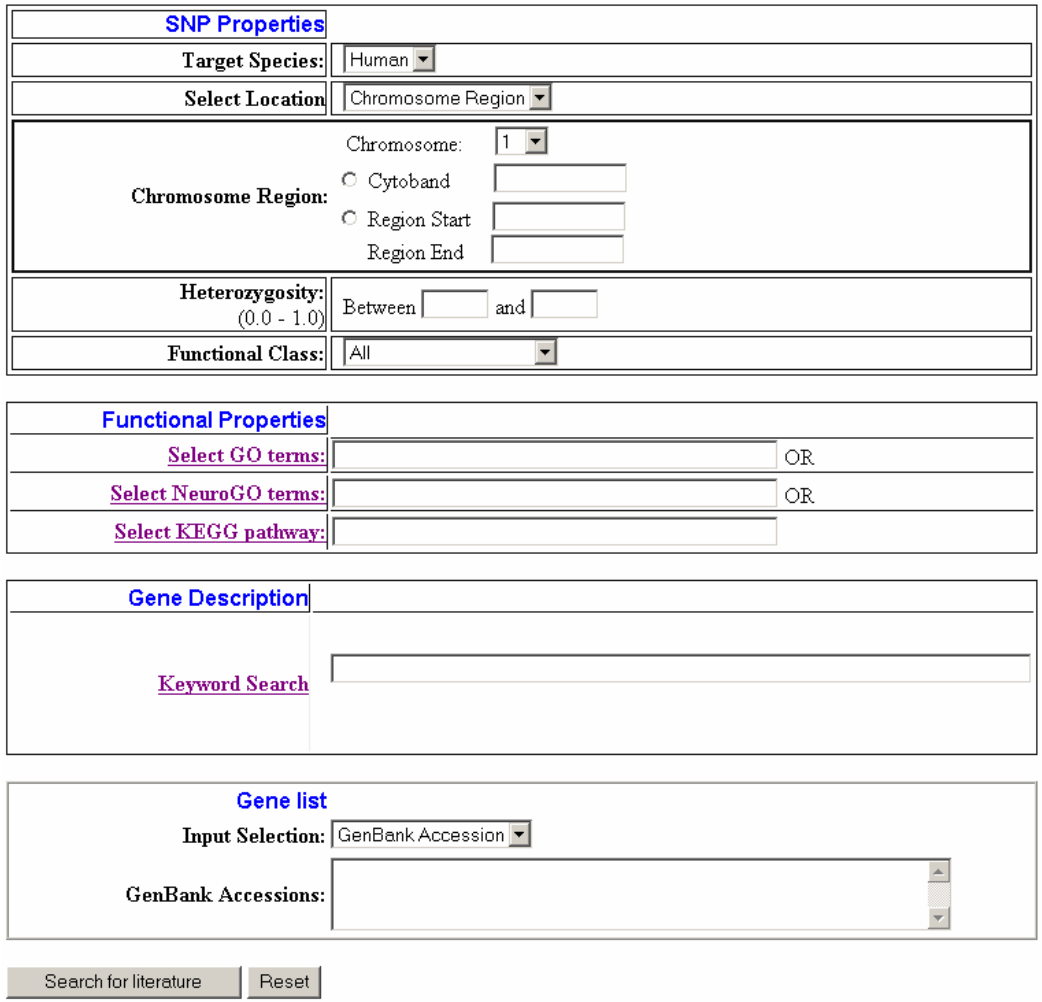
Information of Selected Journal

ISSN	1097-2765
Impact Factor	16.611
Medline Abbreviation	Mol Cell
ISO Abbreviation	Mol. Cell
ISI Abbreviation	MOL CELL
Full Title	Molecular cell.
Note	This journal is included in the Entrez molecular biology databases, including Nucleotide, Protein, and Genome.
PubMed Journal	Full Text (if accessible)
ISI Journal	ISI Master Journal Search (including Journal website link)
Further Information	Acess to additional information distributed by NLM.

Export Citations to Citation Managers

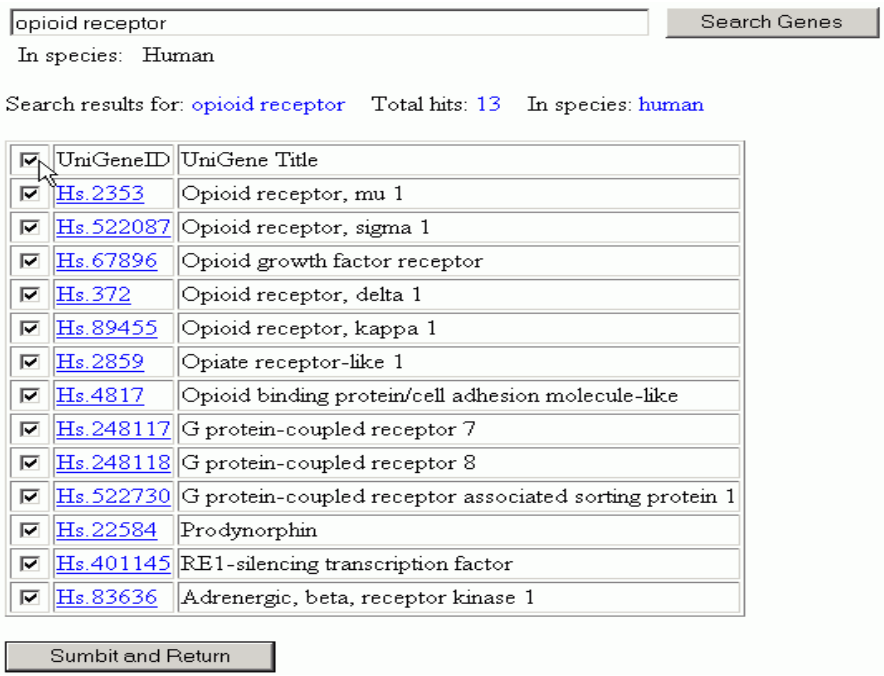


Search by Chromosome Regions (cytoband/location)



Gene/Protein Retrieval

Retrieve all gene/proteins containing search keywords

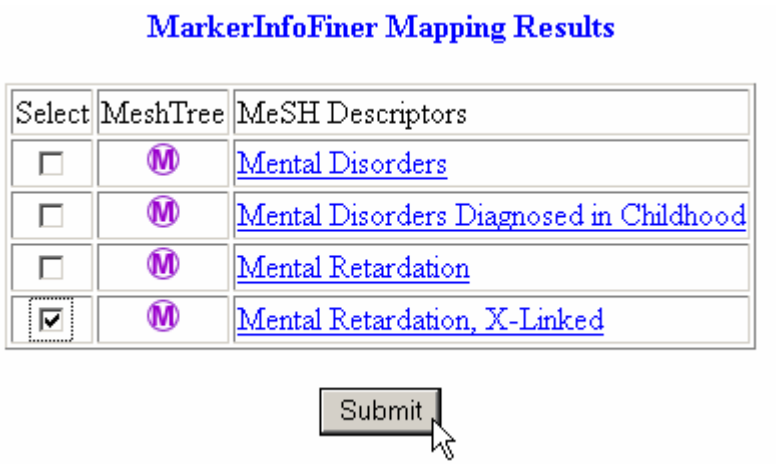


web service calls

Disease and Keyword Search

Return a list of disease names for given keywords

Users can select disorders of interest, our system will return a set of filtered citations.



Knowledge-Based Genome Wide Association Analysis

Knowledge-based analysis of gene expression data, such as GSEA and SigPathway is very useful in providing novel insights

Consider functionally related SNPs together in Genome Wide Association (GWA) analysis is important due to the nature of complex disorders.

Existing knowledge, in the form of pathways and function categories, provides many elementary hypotheses for statistical tests

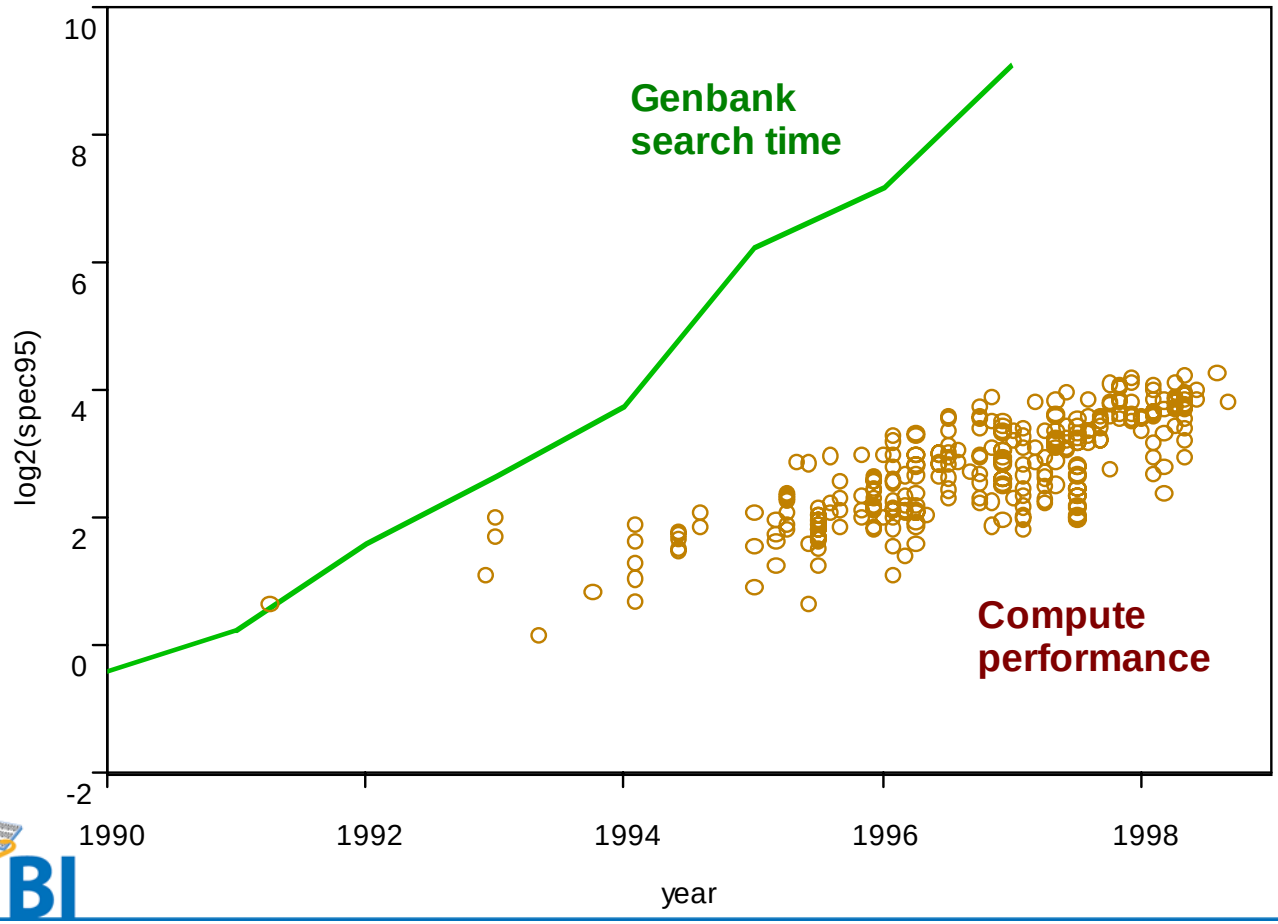
Develop methods for automated evaluation of analysis results

Re-analysis of the Perlegen/Mayo Clinic Parkinson's Disease Tier 1 Data

Method	GSEA p=3 set size>=10		Mayo Clinic-Perlegen Tier 1	
Top ranked gene cutoff	PD co-occurrence #	p-value	PD co-occurrence #	p-value
10	2	0.053	0	0.32
20	3	**0.039**	1	0.54
30	6	**0.001**	1	0.69
50	10	**0.000**	3	0.30
100	11	**0.002**	7	0.09
200	14	**0.021**	9	0.35
500	30	**0.010**	20	0.44

Moore's Law, Data Growth and the Need for Algorithms

Spec95 Integer Performance vs. Genbank Search



Acknowledgements

David States
Alex Ade
Jeremy Phillips
Jing Gao
Rajasree Menon
Carlos Santos
Sirarat Sarntivijai
Ji Chen
Yili Chen

Dragomir Radev
Anthony Fader
Jacob Balzer

Fan Meng
Weijian Xuan

Peter Woolf
Abhik Shah

Terry Weymouth

HV Jagadish
Glenn Tarcea
Aaron Elkiss

Jignesh Patel
Yuanyuan Tian

Jill Mesirov
Michael Reich

Robert Murphy

Mark Musen
Natasha Noy

Bruce Shatz

Peter Karp