Microsoft Research Faculty Summit 2014

15TH ANNUAL

Microsoft Research
Faculty Summit
2014 15TH ANNUAL

Disease gene search engine (DigSee): Text mining for identifying disease-gene-biological events relationships

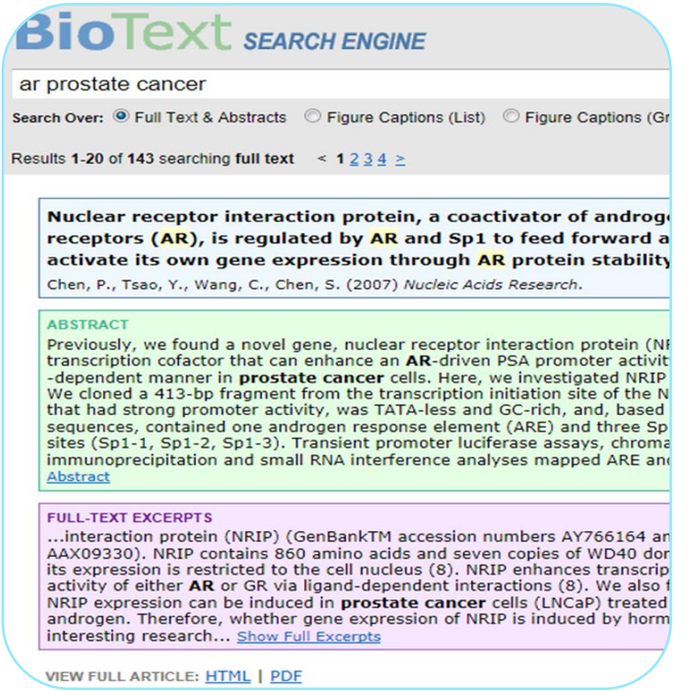
Hyunju Lee
Gwangju Institute of Science and Technology. Republic of Korea

Microsoft

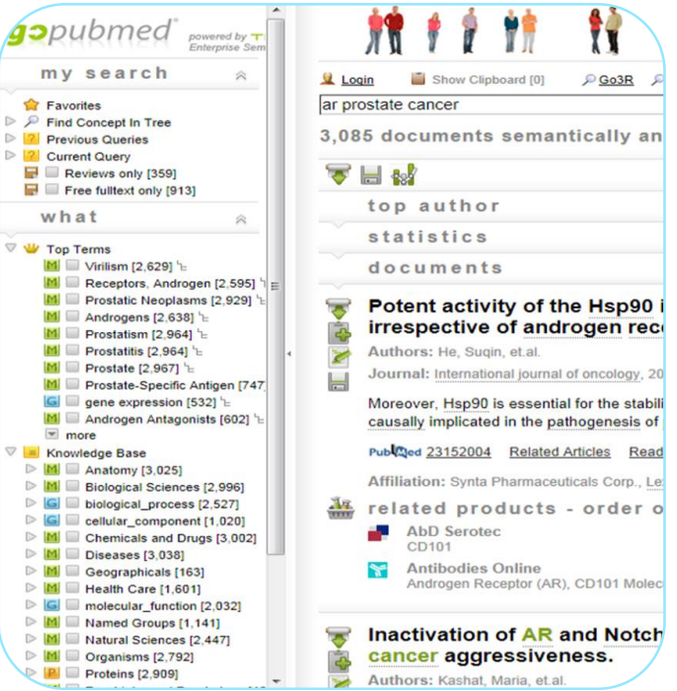
Search disease-related genes

- Simplify the relations between genes and diseases into typed binary relations

PubMed



BioText



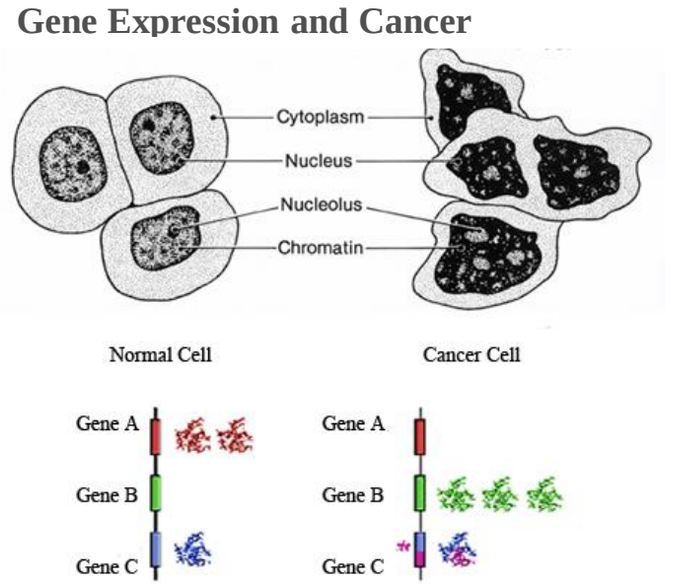
GoPubMed

Search disease-related genes with biological events

- Understanding the gene–disease relation can be further enhanced by identifying in which biological events the genetic effect is valid for the disease development.
- Biological events
  - Gene expression, transcription, protein catabolism
    - Related to protein production and breakdown

An example sentence

We conclude that MMP-7 over-expression correlates with breast cancer in vitro invasiveness and that MMP-7 may promote invasion by increasing the secretion and activation of proMMP-2 and proMMP-9. (PMID:16019136)

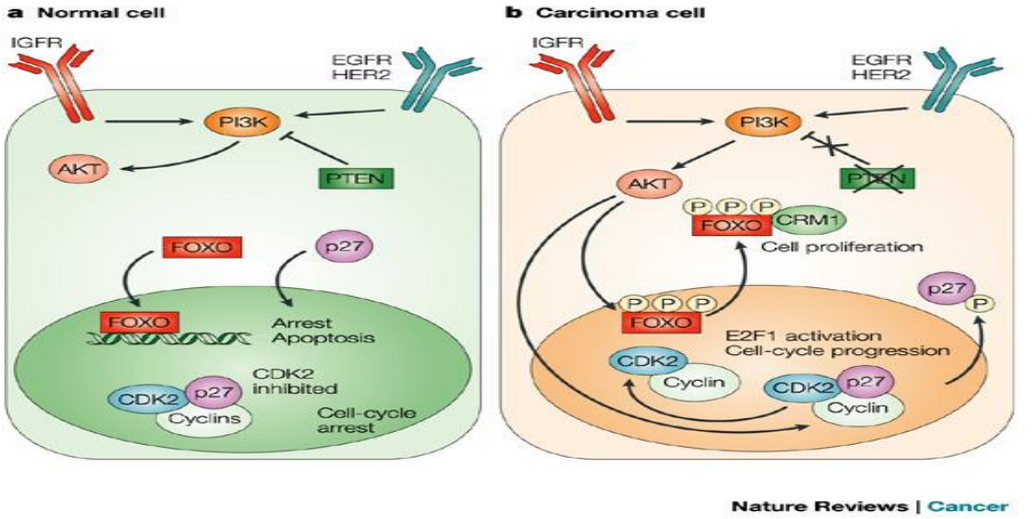


Search disease-related genes with biological events

- Understanding the gene–disease relation can be further enhanced by identifying in which biological events the genetic effect is valid for the disease development.
- Biological events
  - Phosphorylation

An example sentence

In carcinoma cells that lack PTEN activity, PI3K/AKT signalling is active, resulting in phosphorylation of FOXO and p27 by AKT. FOXO phosphorylation promotes its nuclear export in a CRM1-dependent manner, so that it can no longer activate transcription of its target genes. (PMID: 16450001)



From Nature Reviews Cancer 4, 106-117

Search disease-related genes with biological events

- Understanding the gene–disease relation can be further enhanced by identifying in which biological events the genetic effect is valid for the disease development.
  - **Localization:** A change of the location or presence of a protein

    > RRR-alpha-tocopheryl succinate-induced apoptosis of human breast cancer cells involves Bax translocation to mitochondria. (PMID:12750270)

  - **Binding:** The binding of two or more proteins, the binding of a protein and DNA

    > In breast cancer cells, NGF inhibits C2-induced apoptosis through binding of p75NTR and NF-kappaB activation. (PMID:17638883)

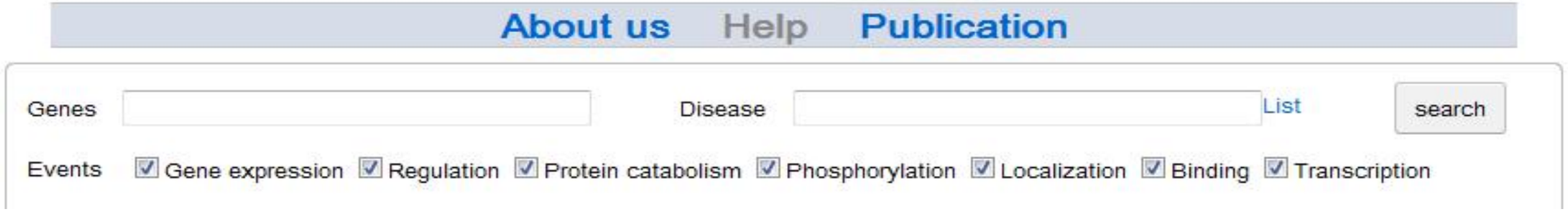
  - **Regulation:** A regulatory or causal relation between the above event classes or proteins

    > Since breast cancers that overexpress COX-2 are associated with a higher rate of metastasis to bone, we hypothesized that COX-2 expression in tumor cells would induce IL-11. (PMID:16457848)

DiGSeE

Disease Gene Search Engine
with Evidence Sentences
Ver. cancer

http://gcancer.org/digsee



For ~ 200 cancer names,
1,391,019 sentences were scored.

Breast Cancer	93,672
Cervical Cancer	11,308
Colon cancer	49,934
Gastric cancer	19,024
Glioblastoma	9,939
Hepatocellular cancer	18,079
Kidney cancer	10,998
Liver cancer	39,946
Lung cancer	38,205
Ovarian cancer	41,298
Prostate cancer	176,875
Retinoblastoma	43,955

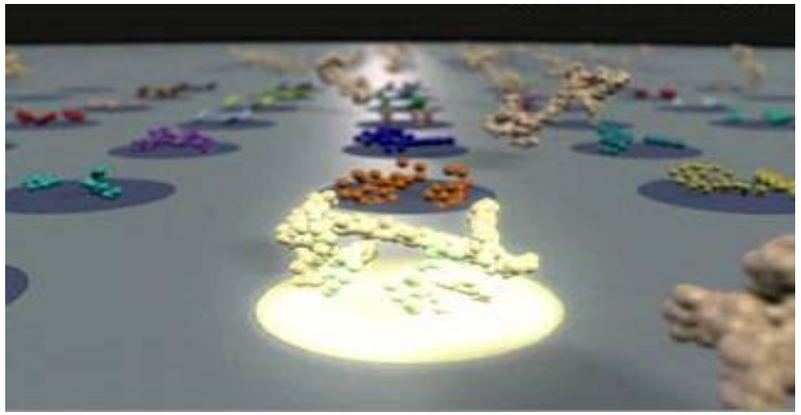
For ~ 700 nervous system diseases names,
130,451 sentences were scored.

Alzheimer Disease	64,986
Back Pain	11,139
CNS Diseases	15,129
Chronic Pain	11,139
Dementia	7,587
Hearing Disorders	3,341
Multiple Sclerosis	3,999
Parkinson Disease	6,433
Pituitary Diseases	15,129
Seizures	3,850
Stroke	4,001
Tics	88,565

Example query (1)

Microarray experiments often identify more than hundreds of candidate cancer-related genes.

Finding literature evidence of these genes is the first step to validate them.



Example)124 differentially expressed genes on prostate cancer using microarray experiment (Jia *et al.* 2011)

DigSee returns evidence sentences for 49 genes of the 124 genes, which are highlighted with the expression changes and regulations of the candidate genes in prostate cancer.

Search Results: 228 documents

Graph of gene co-occurrences | Send to file (Search result)

Total [228] | Gene expression [162] | Regulation [123]

Searched genes (Number of abstracts)

CAV1	60	ENO2	44	HSPB1	20	CDKN2C	9	BDNF	7	ITGA5	7	MCAM	7	S100B	6
PTRF	5	CDKN1C	4	ITGB3	4	MTMR11	4	PDLIM4	4	ACTC1	3	BCHE	3	CRTC3	3
PTPN11	3	ACO1	2	BIN1	2	CSPG4	2	FERMT2	2	PEA15	2	RXRG	2	TBXA2R	2
CADM3	1	DPYSL3	1	EFEMP1	1	FEZ1	1	HNRNPD	1	HPD	1	HSPB8	1	KPNA3	1

Show 92 more

Literature	Related Genes	Detail
RJ Jin, Y Lho, Y Wang, M Ao, MP Revelo, SW Hayward, ML Wills, SK Logan, P Zhang, RJ Matusik - Cancer research, 2008. (PMID : **18483241**) Title : Down-regulation of p57Kip2 induces prostate cancer in the mouse. Evidence sentence : **Here, we show that the expression of p57 (Kip2) is significantly decreased in human prostate cancer, and the overexpression of p57 (Kip2) in prostate cancer cells significantly suppressed cell proliferation and reduced invasive ability.** (score : 0.903)	CDKN1C	show
CS Aung, MM Hill, M Bastiani, RG Parton, MO Parat - European journal of cell biology, 2011. (PMID : **20732728**) Title : PTRF-cavin-1 expression decreases the migration of PC3 prostate cancer cells : role of matrix metalloprotease 9. Evidence sentence : **Caveolin-1 expression and secretion are increased in metastatic prostate cancer, and caveolin-1 seems to contribute to prostate cancer growth and metastasis.** (score : 0.889)	CAV1 PTRF ADAM9 (2 more)	show
L Li, G Yang, S Ebara, T Satoh, Y Nasu, TL Timme, C Ren, J Wang, SA Tahir, TC Thompson - Cancer research, 2001. (PMID : **11389065**) Title : Caveolin-1 mediates testosterone-stimulated survival\ growth and promotes metastatic activities in prostate cancer cells. Evidence sentence : **Previously, we demonstrated that up-regulation of caveolin-1 (cav-1) was associated with prostate cancer metastasis, biochemical recurrence after radical prostatectomy, and androgen insensitivity.** (score : 0.882)	CAV1	show
K Ge, F Minhas, J Duhadaway, NC Mao, D Wilson, R Buccafusca, D Sakamuro, P Nelson, SB Malkowicz, J Tomaszewski, GC Prendergast - International journal of cancer. Journal international du cancer, 2000. (PMID : **10738240**) Title : Loss of heterozygosity and tumor suppressor activity of Bin1 in prostate carcinoma. Evidence sentence : **Ectopic expression of Bin1 suppressed the growth of prostate cancer lines in vitro.** (score : 0.871)	BIN1 NME2	show

ch Engine
Search/
Google

Searched genes (Number of abstracts)

CAV1 52
S100B 4
PEA15 2
MFAP5 1
Show 92 more

Literature

Detail

Detail

PubMed ID	20732728
Title	PTRF-cavin-1 expression decreases the migration of PC3 prostate cancer cells : role of matrix metalloprotease 9 .
Related genes (HGNC link, GeneCards link)	CAV1 (H, G) PTRF (H, G) ADAM9 (H, G) MMP9 (H, G) BLZF1 (H, G)
Evidence sentences Max score	0.889
Related cancer	Prostate cancer

Abstract

Gene term, Event term, Cancer term

Caveolae are specialized plasma membrane subdomains with a distinct lipid and protein composition, which play an essential role in cell physiology by performing trafficking and signalling functions. The structure and functions of caveolae have been shown to require caveolin-1, a major protein component of caveolae. **(1) Caveolin-1 expression and secretion are increased in metastatic prostate cancer, and caveolin-1 seems to contribute to prostate cancer growth and metastasis.** Recently, a cytoplasmic protein named PTRF (Polymerase I and Transcript Release Factor) or cavin-1 was found to be required, in concert with caveolin-1, for the formation and functions of caveolae **(2) Genetic ablation of PTRF results in loss of caveolae while caveolin-1 is still expressed, albeit at reduced level, but associates with flat plasma membrane. (3) In metastatic PC3 prostate cancer cells that express abundant caveolin-1 but no PTRF, heterologous PTRF expression restores caveola formation and caveolin-1 distribution (Hill et al., 2008 ; Cell 132, 113-124).** We now show that PTRF\ -1 - expressing PC3 cells exhibit decreased migration, and that this effect is mediated by reduced MMP9 production PTRF\ -1, and to a lesser extent, cavin-2, -3, and -4 all decreased MMP9 We further show that the PTRF\ -1 - mediated reduction of MMP9 production is independent of caveola formation **(4) Taken together, our results suggest that PTRF\ -1 expression alters prostate cancer aggressiveness.**

Sentence No.	Gene Name	Gene Symbol	Event Text	Event Type	Score
(1)	Caveolin-1	CAV1	expression	Gene expression	0.889
(2)	caveolin-1	CAV1	expressed	Gene expression	0.148
(3)	caveolin-1	CAV1	express	Gene expression	0.516
(4)	PTRF\ -1	PTRF	expression	Gene expression	0.707

1-4 of 4

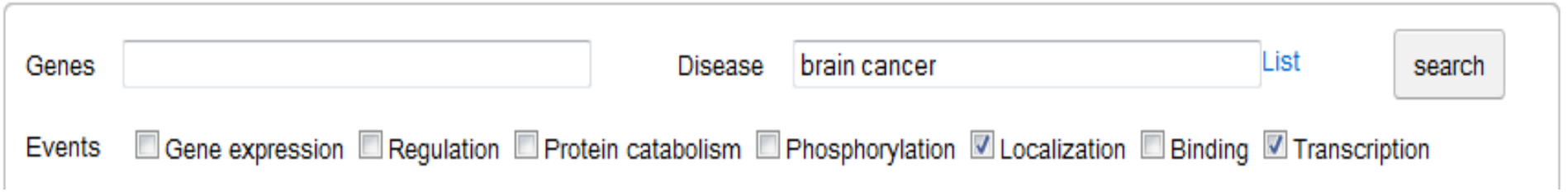
Close

Title : Expression of PTRF in PC-3 cells modulates cholesterol dynamics and the actin cytoskeleton impacting secretion

PTRF
CAV1

Example query (2)

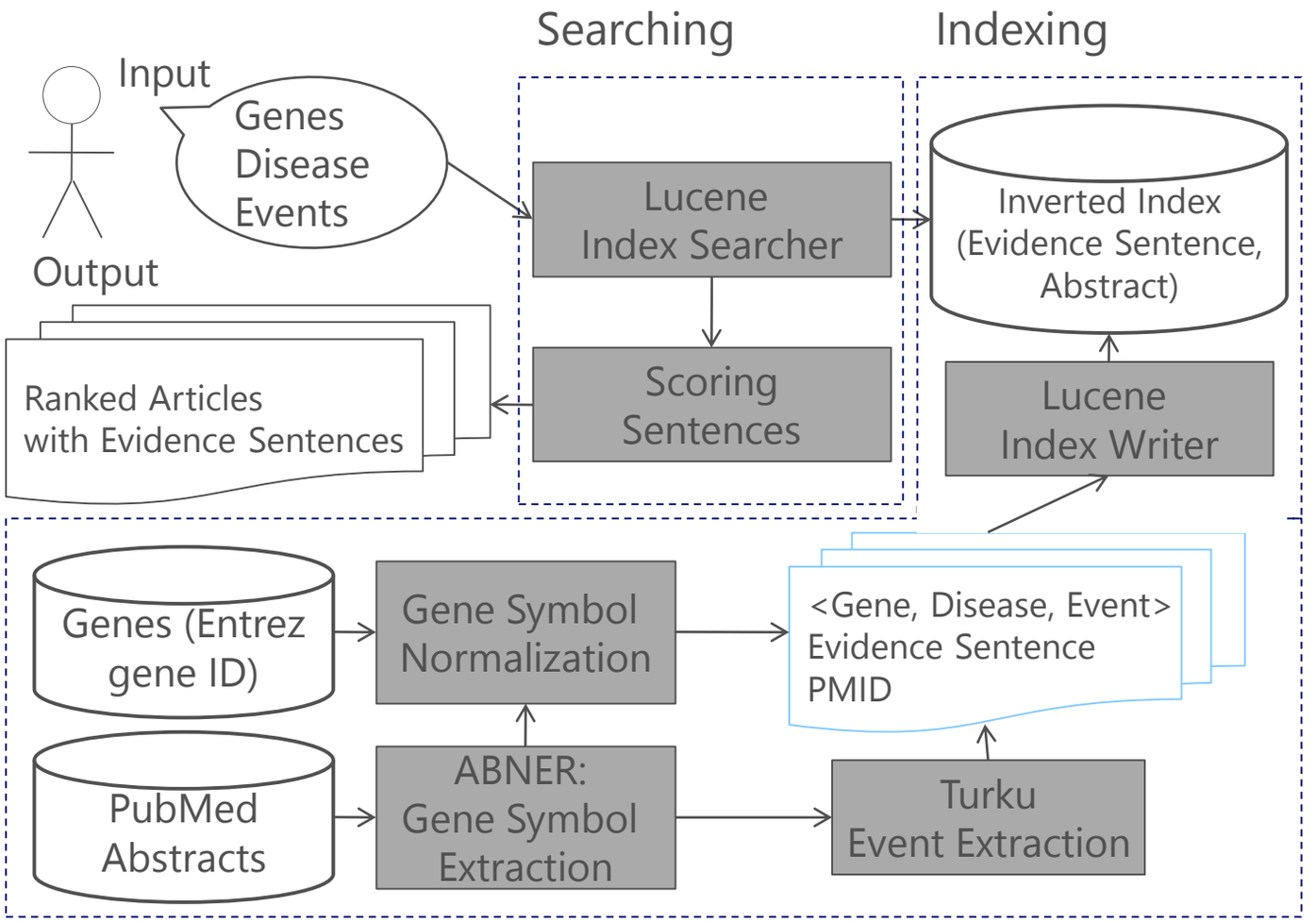
- Show a list of genes whose localization and transcription are related to brain cancer.



- DigSee system returns a list of 224 genes with evidence sentences in which localization and transcription events of these genes are related to brain cancer.

> In tumor samples, ING1 proteins aberrantly localized to the cytoplasm, and to a lesser extent, to the nucleus of glioma cells. [PMID:14676120]

Indexing and searching processes in the DigSee system



Gene symbol extraction and normalization

- Entity recognition system (NER), ABNER, was used to recognize gene/protein names
- Recognized genes were normalized using Moara
  - Moara uses gene synonyms from UniProt and the HUGO Gene Name Consortium

Gene mentions in documents	Gene Symbols
Alpha 2 macroglobulin	A2M
Alpha2 macroglobulin	
Alpha-2 macroglobulin	
Alpha-2-macroglobulin	
…	
Androgen receptor	AR
AR protein	
…	

Biological event extraction

- Turku event extraction system
  - Extract complex events among genes and proteins from biomedical literature
  - Require named entity recognition system and sentence parser

Epidemiological studies and prevention trials suggest selenium is a promising preventive agent for prostate cancer. Selenium-containing compounds inhibited the growth of prostate cancer cell lines including androgen sensitive LNCaP and androgen insensitive DU145 and PC3 cells in vitro. Previous study revealed a novel mechanism of selenium action in which selenium (methylseleninic acid (MSA)) markedly reduced androgen receptor (AR) signaling in prostate cancer cells , suggesting that selenium might act as an antiandrogen , which could serve as a therapeutic agent for prostate cancer. In this study , we tested whether selenium (methylselenocysteine (MSC)) affects tumor growth of human prostate cancer cells by targeting AR signaling in vivo. Prostate tumor xenografts were established in nude mice by co-inoculating LNCaP cells with Matrigel. The mice-bearing tumors were treated with or without MSC (100 microg\ ) via intraperitoneal injection for 2 weeks. The effect of MSC on tumor growth , AR , and prostate-specific antigen (PSA) expression was examined. Methylselenocysteine (MSC) significantly inhibited LNCaP tumor growth (P < 0.05). ARexpression in tumor tissues and serum PSA levels were considerably decreased in MSC-treated mice compared to the vehicle controls. Pharmacological dose of MSC inhibits the growth of LNCaP human prostate cancer in vivo accompanied by a decrease in the expression of AR and PSA. These findings suggest that selenium (MSC) can serve as a therapeutic agent aimed at disruption of AR signaling for prostate cancer.

The effect of MSC on tumor growth, AR, and prostate-specific antigen (PSA) expression was examined.

AR expression in tumor tissues and serum PSA levels were considerably decreased in MSC-treated mice compared to the vehicle controls.

Pharmacological dose of MSC inhibits the growth of LNCaP human prostate cancer in vivo accompanied by a decrease in the expression of AR and PSA.

Scoring sentences

- DigSee services the *evidence sentences* with information such that 'which genes' are involved in the development of 'which disease' through 'which biological events'.

> Significantly, down-regulation of SOX9 by siRNA in prostate cancer cells reduced endogenous AR protein levels, and cell growth indicating that SOX9 contributes to AR regulation and decreased cellular proliferation.[PMID: 17234760]

- DigSee demotes sentences that do not show relation between gene and disease (For example, explaining an experimental procedure)

> To determine the role of CD147 in the invasiveness properties of prostate cancer, we successfully down-regulated CD147 by RNA interference (RNAi) technology, in PC-3 cell line at high level of CD147 expression.[PMID: 16627983]

Scoring sentences: Corpus

- The ranking step measures the relevance of the candidate evidence sentences
- Gold standard data: 207 positive and 356 negative sentences from cancer-related articles

[Feature selection sentences]

Events	Binding	Gene expression	Localization	Phosphorylation	Protein catabolism	Transcription
Positive	11	20	9	19	6	5
Negative	26	20	23	24	24	45
Total	**37**	**40**	**32**	**43**	**30**	**50**

[Performance testing sentences]

Events	Binding	Gene expression	Localization	Phosphorylation	Protein catabolism	Transcription
Positive	18	52	19	18	8	22
Negative	29	46	38	38	17	26
Total	**47**	**98**	**57**	**56**	**25**	**48**

Gene expression event contains positive regulation, negative regulation, and regulation events.

Feature selection rules for ranking evidence sentences

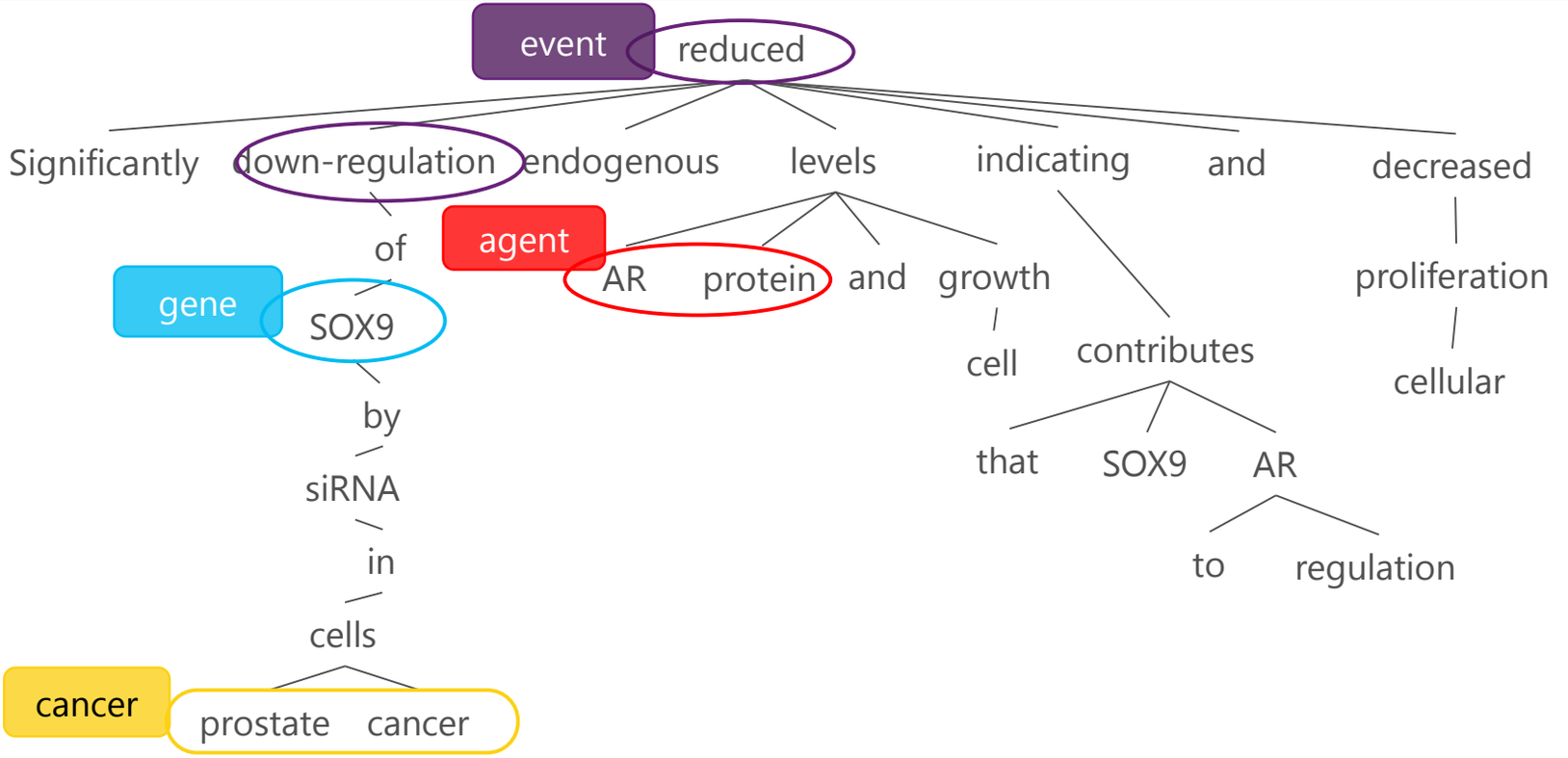
- Based on Turku event extraction system
  - Normalized event SVM score
  - Normalized edge SVM score
- Based on the dependency parser tree
  - Distance between gene and event
  - Distance between event and regulation
  - Distance between event and cancer
  - Event depth
  - Agent
- Based on terms in the sentence
  - Cancer link keyword count
  - Cancer hallmark keyword count: apoptosis, angiogenesis, growth, invasion, metastasis, and proliferation
  - Negative scores (to + infinitive, study-related keywords, negative keywords)

Feature selection rules for ranking evidence sentences

- Based on dependency parser tree

Gene-Event distance
Event-Regulation distance
Event-Cancer distance

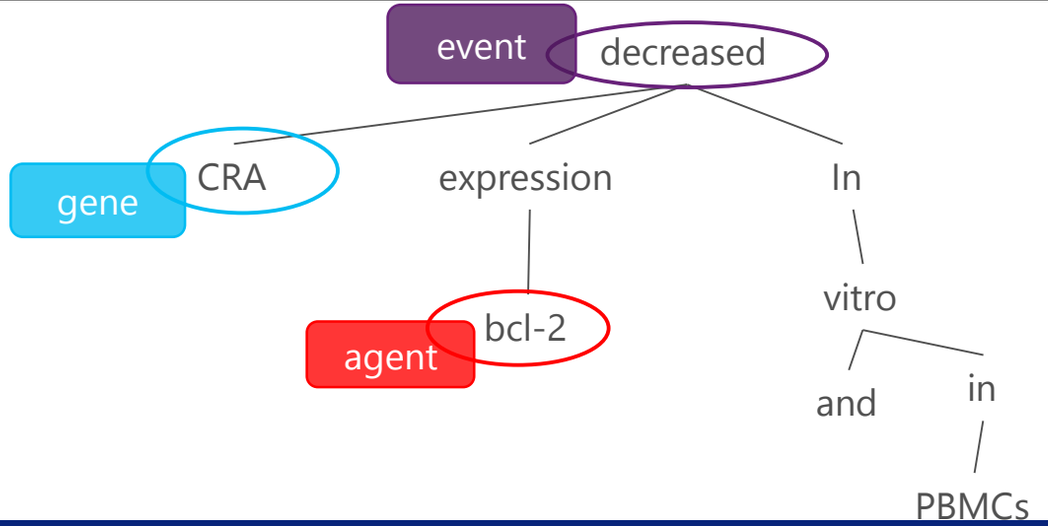
Significantly, down-regulation of SOX9 by siRNA in prostate cancer cells reduced endogenous AR protein levels, and cell growth indicating that SOX9 contributes to AR regulation and decreased cellular proliferation. [PMID: 17234760]



Feature selection rules for ranking evidence sentences

- Based on dependency parser tree
  - Agent: the genes that have relations with a query gene in a sentence
  - Since our goal is to find the disease-gene relationship, the sentence only containing events for the *gene-gene relationship* is more likely to be classified as a negative sentence.

CRA decreased bcl-2 expression in vitro and in PBMCs. [PMID: 10561278]



Feature selection rules for ranking evidence sentences

- Based on terms in the sentence
  - Cancer link keyword count
    - # of cancer terms or cancer-related keywords in the sentence
    - Hyponyms and hypernyms of cancer collected from WordNet
  - Cancer hallmark keyword count:
    - Keywords of known characteristic of cancer during its initiation, development, and progress
    - Apoptosis, angiogenesis, growth, invasion, metastasis, and proliferation

Cancer link keyword

Cancer hallmark keyword

Significantly, down-regulation of SOX9 by siRNA in prostate **cancer** cells reduced endogenous AR protein levels, and cell growth indicating that SOX9 contributes to AR regulation and decreased cellular **proliferation**. [PMID: 17234760]

Feature selection rules for ranking evidence sentences

- Based on terms in the sentence
  - Negative scores : phrases to detect negative sentences
    - to + infinitive: to determine, to find, to assess
    - Study-related keywords: troponym of study from Wordnet
    - Negative keywords: not, never

Negative scores

To **determine** the role of CD147 in the invasiveness properties of prostate cancer, we successfully down-regulated CD147 by RNA interference (RNAi) technology, in PC-3 cell line at high level of CD147 expression. [PMID: 16627983]

Bayesian classifier

- Bayesian model
  - A Bayesian classifier with these features was modeled to identify positive evidence sentences from negative sentences
  - Same prior for positive and negative evidence sentences
  - Likelihood ratio of features

$$\mathrm{L}(\text{features}) = \frac{p(\text{features} \mid \text{positive})}{p(\text{features} \mid \text{negative})}$$

  - Assuming independency between features except two cases

$$\mathrm{L}(\text{features}) = \prod_{\text{features}_i} \frac{p(\text{features}_i \mid \text{positive})}{p(\text{features}_i \mid \text{negative})} \cdot \frac{p(\text{cancer keywords} \mid \text{positive}, \text{event-cancer distance})}{p(\text{cancer keywords} \mid \text{negative}, \text{event-cancer distance})} \cdot \frac{p(\text{agent} \mid \text{positive}, \text{hallmark keywords})}{p(\text{agent} \mid \text{negative}, \text{hallmark keywords})}$$

Accuracy evaluation of identifying evidence sentences

- Accuracies of individual features from cancer-related documents

Features	F-measure	AUC
Normalized event SVM score	62.7	57.8
Normalized edge SVM score	60.3	42.5
Gene–event distance	62.1	52.0
Event–regulation distance	64.5	59.7
Event–cancer distance	71.5	74.1
Cancer keywords count (depending on event–cancer distance)	72.5	72.5
Hallmark keywords count	64.4	58.8
Event depth	60.2	47.7
Negative score	68.7	59.3
Agent (depending on hallmark keywords count)	64.1	62.6
Total	**72.7**	**80.5**

Accuracy evaluation of identifying evidence sentences

- Accuracies according to biological events from cancer-related documents

P: Precision, R: Recall, F: F-measure

Biological events	Bayesian classifier				SVM classifier				Random order			
	P	R	F	AUC	P	R	F	AUC	P	R	F	AUC
Binding	80.0	91.2	83.6	87.1	45.7	88.3	60.0	62.6	42.2	94.1	57.8	50.7
Gene expression	66.7	96.2	78.7	79.3	65.7	88.7	75.5	73.7	54.4	98.6	70.0	50.7
Localization	66.7	53.3	59.0	72.5	56.0	75.0	63.9	71.5	37.5	90.7	52.1	50.5
Phosphorylation	75.0	85.0	79.3	93.7	90.0	51.7	65.3	75.2	37.2	90.0	51.4	50.9
Protein catabolism	100	70.0	80.0	96.7	40.0	80.0	52.0	61.7	42.6	87.4	55.4	52.0
Transcription	63.3	87.0	73.1	74.7	54.3	86.0	66.4	71.0	48.5	97.2	64.5	48.6
Total	62.6	86.9	72.7	80.5	53.7	81.0	64.5	71.2	43.7	93.0	59.4	49.8

- SVM classifier (Bag of words model): 239 words were used.
- Random order (Baseline method): Sentences were randomly ordered 100 times

Accuracy evaluation of identifying evidence sentences

- Accuracies from Alzheimer's disease-related documents

✓ Training with gold standard sentences from cancer

Biological Events	AUC	Precision	Recall	F-score
Binding	77.9	80.0	80.0	80.0
Gene expression	78.1	68.2	83.3	75.0
Localization	77.4	72.4	100	84.0
Negative regulation	89.4	86.7	100	92.9
Phosphorylation	86.6	81.0	100	89.5
Positive regulation	74.6	77.8	93.3	84.8
Protein catabolism	87.0	83.3	93.7	88.2
Regulation	70.0	83.3	100	90.9
Transcription	60.7	42.9	100	59.9
Total	78.9	73.0	88.5	80.0

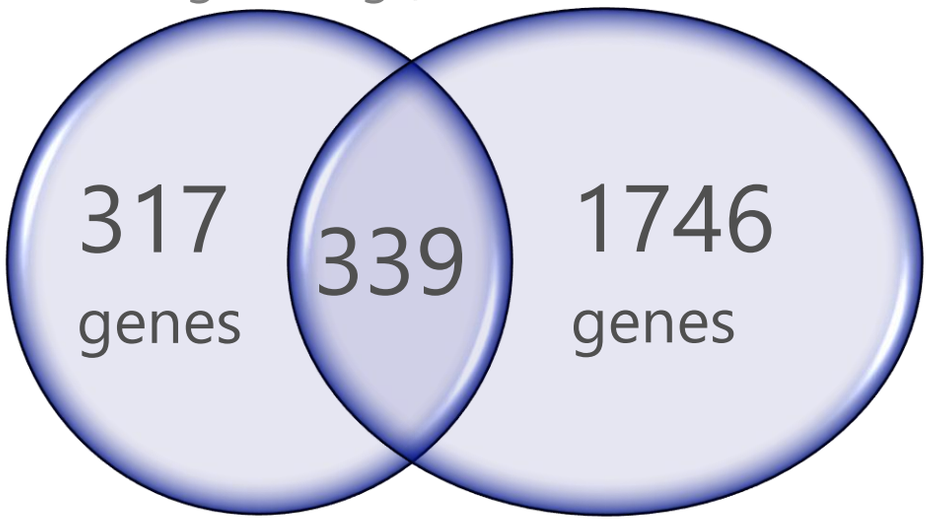
✓ Training with gold standard sentences from Alzheimer's disease

Biological Events	AUC	Precision	Recall	F-score
Binding	86.9	81.2	86.7	83.9
Gene expression	78.7	73.7	77.8	75.7
Localization	89.3	84.0	100	91.3
Negative regulation	95.2	89.7	100	94.5
Phosphorylation	89.9	80.9	100	89.5
Positive regulation	80.8	80.0	93.3	86.2
Protein catabolism	83.9	75.0	93.7	83.3
Regulation	98.3	90.9	100	95.2
Transcription	63.8	52.4	91.7	66.7
Total	84.2	70.8	82.7	80.3

Comparison to Alzheimer's disease DB

AlzGene DB (http://www.alzgene.org/)

DigSee search engine



A 8 to 16-fold **GFAP increase** in **Alzheimer** brain was established. [PMID:2514723]

Accordingly, **up-regulation** of **GSK-3** may contribute to cytoskeletal pathology within neurites in **AD**. [PMID: 12214113]

Future works

- We are currently incorporating all disease types and more biological events to DigSee thank to Windows Azure resource.
  - 32CPU cores and 10 TB for 1 year

		Crawling from PubMed	Gene Symbol Extraction	Gene Symbol Normalization	Biological Event Extraction	Scoring Evidence Sentences	Total
Cancer	Our previous work						
	# of abstracts	3,010,235	2,056,082	2,056,082	2,056,082	1,117,325	
	Processing time using a single CPU (days)	34	95	47	713	1	890
	Hard disk (GB)	1.5	2	2	120	20	145.5
All diseases	Current work						
	# of abstracts	25,923,285	17,282,190	17,282,190	17,282,190	8,641,095	
	Processing time using a single CPU (days)	300	800	400	6000	8	**7508**
	Hard disk (GB)	13	18	18	960	160	2951

Acknowledgement

- Data Mining and Computational Biology Laboratory at GIST
  - **Jeongkyun Kim, and Seongeun So**
- Nanyang Technological University, Singapore
  - **Prof. Jung-jae Kim**
- KAIST, Korea
  - **Prof. Jong C. Park, Hee-Jin Lee**

THANK YOU FOR YOUR ATTENTION

Save the planet and return
your name badge before you
leave (on Tuesday)

Microsoft Privacy Policy statement applies to all information collected. Read at research.microsoft.com

© 2014 Microsoft Corporation. All rights reserved. Microsoft, Windows, Windows Vista and other product names are or may be registered trademarks and/or trademarks in the U.S. and/or other countries.
The information herein is for informational purposes only and represents the current view of Microsoft Corporation as of the date of this presentation. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information provided after the date of this presentation. MICROSOFT MAKES NO WARRANTIES, EXPRESS, IMPLIED OR STATUTORY, AS TO THE INFORMATION IN THIS PRESENTATION.