



Building Interaction Networks from MEDLINE with Deep Parsing

Yoshimasa Tsuruoka
GENiA, University of Tokyo

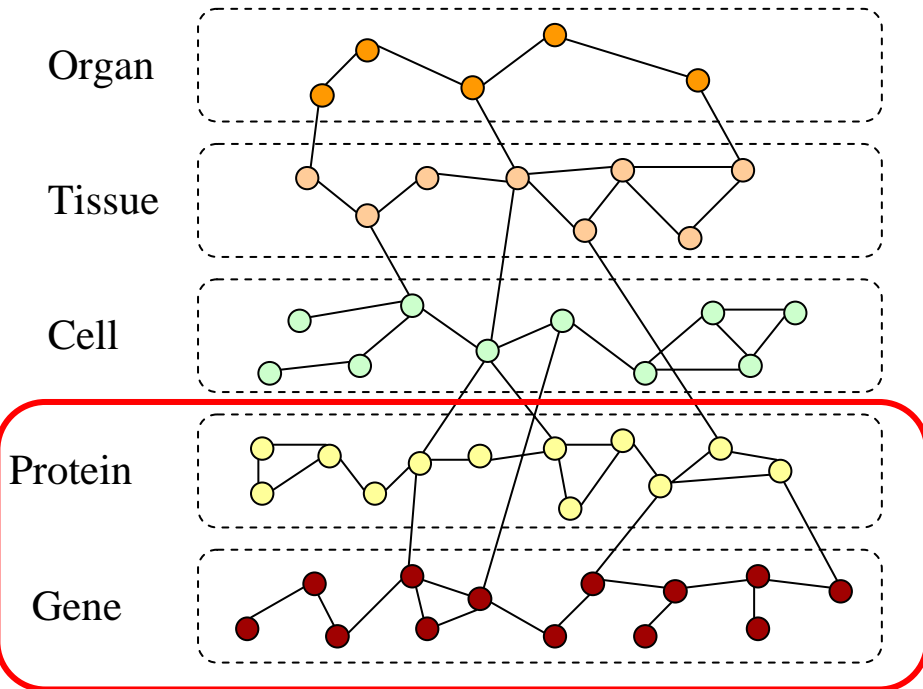
◆ CytoSailing

- ⊕ Interactive IE system that help biologists build protein-protein interaction networks from MEDLINE

◆ Hiiragi

- ⊕ IR system that allows for searches based on predicate-argument relations

◆ Domain Knowledge



◆ Text

The screenshot shows a Microsoft Internet Explorer browser window displaying the PubMed website. The search results for the query "Immunobiology" are shown. The first result is highlighted:

1: Immunobiology 2006;211(1-2):65-74. Epub 2005 Dec 27

RELEVANCE

Abstract

Evaluation of Brucella abortus DNA vaccine by expression of Cu-Zn superoxide dismutase antigen fused to IL-2.

Gonzalez-Smith A, Yeamanspalli R, Andrews F, Onate A

Faculty of Biological Sciences, Molecular Immunology Laboratory, Department of Microbiology, Universidad de Concepcion, P.O. Box 152-C, Concepcion, Chile.

The Cu-Zn superoxide dismutase (SOD) antigen of Brucella abortus was previously identified to be a T cell antigen which induces both proliferation of and gamma interferon (IFN-gamma) secretion by T cells from infected mice. In an earlier study, we demonstrated that intramuscular injection of mice with a plasmid DNA carrying the gene for SOD leads to the development of significant protection against B. abortus challenge. It has been reported that the antigen-specific immune responses generated by a DNA vaccine can be enhanced by co-delivery of certain cytokine genes. In this study, we evaluated the effect of delivering IL-2 on the efficacy of SOD DNA vaccine by generating a plasmid (pSecTag-SOD-IL2) that codes for a secretory fusion protein of SOD and IL-2. Another plasmid (pSecTag-SOD) that codes for only SOD as a secretory protein was used for comparison. BALB/c mice injected intramuscularly with pSecTag-SOD or pSecTag-SOD-IL2, but not the control plasmid pSecTag-SOD, developed SOD-specific antibody and T cell immune responses. Upon in vitro stimulation with recombinant SOD (rSOD) antigen, T cells from mice immunized with pSecTag-SOD-IL2, in comparison with those from mice immunized with pSecTag-SOD, exhibited a lower proliferation response but produced significantly higher concentrations of IFN-gamma. Both DNA vaccines, however, induced similar levels of SOD-specific antibodies and cytotoxic T cell response. Although mice immunized with pSecTag-SOD-IL2 showed increased resistance to challenge with B. abortus virulent strain 2308, this increase was not statistically significant from that of pSecTag-SOD vaccinated mice. These results suggest that a SOD DNA vaccine fused to IL2 did not improve protection efficacy.

PMID: 16446171 [PubMed - in process]



CytoSailing

◆ NLP technologies

⊕ Part-of-speech tagging

- ⊕ Tuned to biomedical text – 97-98% precision

⊕ Dictionary-based named-entity recognition

- ⊕ All the recognized entities have IDs.

⊕ Deep parsing

- ⊕ Predicate argument relations

⊕ PPI extraction with machine learning

- ⊕ SVM + predicate-argument features



Information Source of CytoSailing

◆ MEDLINE

- ⊕ Literature database covering most of the papers in the biomedical domain

Number of entries	15 million
Number of abstracts	7.5 million
Number of sentences	70 million
Number of words	1.4 billion



CytoSailing

Info-PubMed - Microsoft Internet Explorer

ファイル(E) 編集(E) 表示(V) お気に入り(A) ツール(O) ヘルプ(H)

戻る 検索 お気に入り

アドレス http://www.tsujii.is.u-tokyo.ac.jp/info-pubmed/

移動 リンク Norton AntiVirus

Trash Folder

Gene Searcher

Search for genes or gene product names

rafl Search

>> Organism >> Field

Content Viewer

INTER ACTION 4 / 2 / 27 NE RAF1 NE Mapk1

NE RAF1 : NE Mapk1 :

Sentences 1 -- 27 Next

- INTER ACTION** **SENTENCE** PMID15349122 **NE** RAF1 **NE** Mapk1
Studies using asRKIP and ssRKIP demonstrated that **RKIP** blocked activation of **MEK** and **ERK** by **Raf-1** in beta cells.
- INTER ACTION** **SENTENCE** PMID15349122 **NE** RAF1 **NE** Mapk1
Studies using asRKIP and ssRKIP demonstrated that **RKIP** blocked activation of **MEK** and **ERK** by **Raf-1** in beta cells.
- INTER ACTION** **SENTENCE** PMID15349122 **NE** RAF1 **NE** Mapk1
Studies using asRKIP and ssRKIP demonstrated that **RKIP** blocked activation of **MEK** and **ERK** by **Raf-1** in beta cells.
- INTER ACTION** **SENTENCE** PMID15349122 **NE** RAF1 **NE** Mapk1
Studies using asRKIP and ssRKIP demonstrated that **RKIP** blocked activation of **MEK** and **ERK** by **Raf-1** in beta cells.
- CONDICION PREVENTION** **SENTENCE** PMID15208680 **NE** RAF1 **NE** Mapk1
B-RAE depletion inhibits DNA synthesis and induces apoptosis in three melanoma cell lines and we show that the **RAE** inhibitor **BAY43-9006** also blocks **ERK** activity, inhibits DNA synthesis and induces cell death in these cells.
- CONDICION AFFECTION** **SENTENCE** PMID15208680 **NE** RAF1 **NE** Mapk1
B-RAE depletion by siRNA blocks **ERK** activity, whereas **A-RAE** and **C-RAE** depletion do not affect **ERK** signalling.
- CONDICION REVERSE** **SENTENCE** PMID15313890 **NE** RAF1 **NE** Mapk1
We found that **SPRY2**, an inhibitor homologous to **SPRY1**, which was previously shown to suppress Ras/**ERK** signaling via direct binding to **Raf-1**, had reduced expression in WT BRAF cells.

Product

Oncogene **RAF1**
Raf-1
v-**raf-1** murine leukemia viral oncogene homolog 1

A-raf-1

Raf-1
protein kinase **raf 1**
Raf-1 kinase inhibitor protein

Raf-1
v-**raf-1** murine leukemia viral oncogene homolog 1

ページが表示されました

インターネット



Part-Of-Speech Tagging

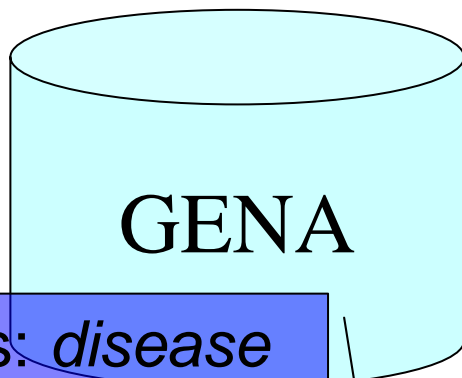
The peri-kappa B site mediates human immunodeficiency
DT NN NN NN VBZ JJ NN
virus type 2 enhancer activation in monocytes ...
NN NN CD NN NN IN NNS

- ◆ GENIA tagger assigns POS tags to tokens.
- ◆ Specifically tuned to biomedical text
 - ◆ 97-98% precision



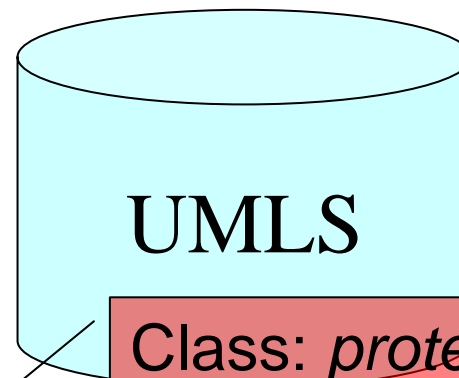
Entity Recognition

Gene dictionary



Class: *disease*
ID: *C0340708*
Name: *deep vein thrombosis*

Disease dictionary

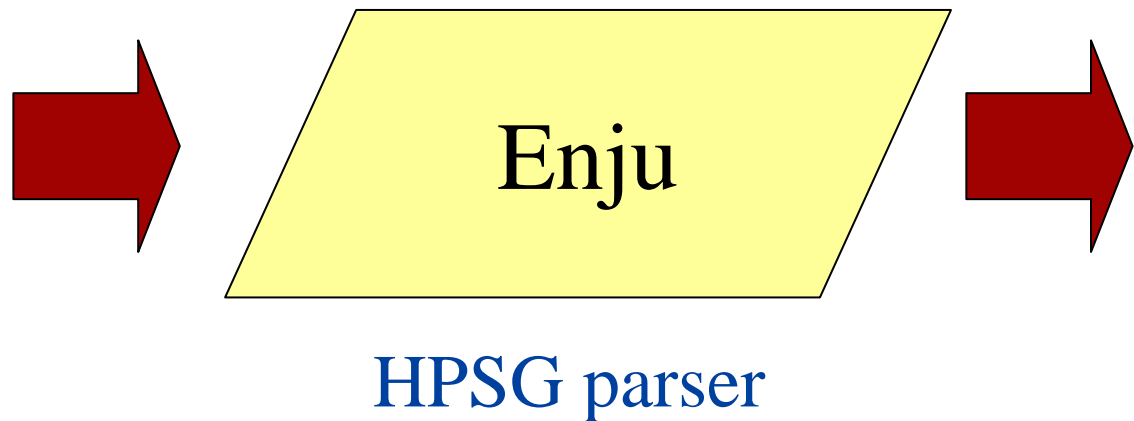


Class: *protein*
ID: *GHS003134*
Name: *C-reactive protein*

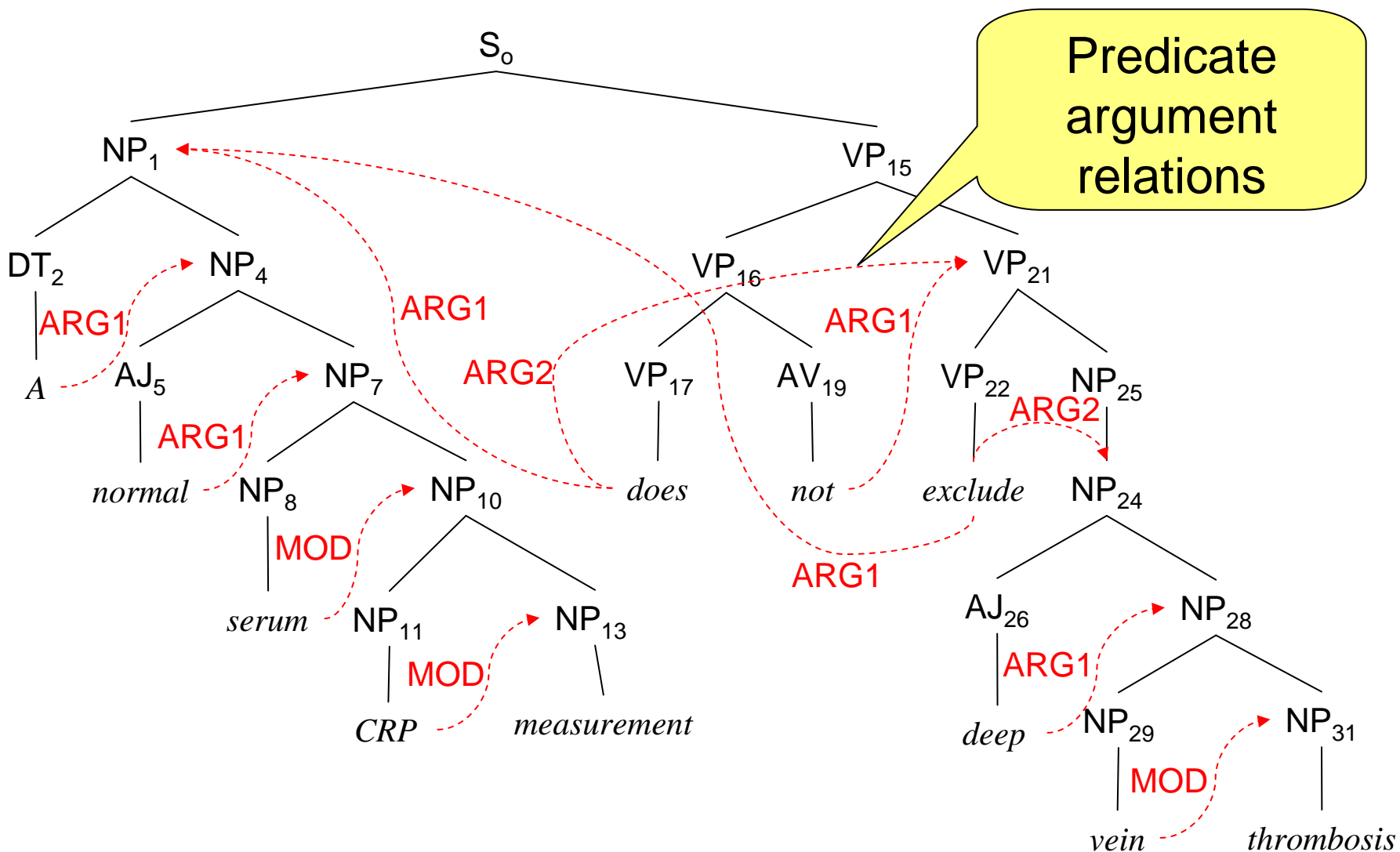
A normal serum CRP measurement does not exclude deep vein thrombosis

Deep Parsing

“A normal serum CRP measurement does not exclude deep vein thrombosis.”



Predicate-argument relations





Normalization with predicate-argument structure

Adenovirus-mediated high dose **p53** overexpression induced **Peg3 / Pw1** mRNA expression .

One of the mechanisms for **p53** to induce mitochondria-mediated **cell** death events is to activate genes that are directly involved in the initiation of mitochondria-induced apoptosis .

The **p53 gene** suppresses tumor **cell** growth by inducing **cell cycle arrest** or apoptosis .

Concomitant up-regulation of **p21** (**WAF1 / Cip1**) but not **p53** , especially in nodular hyperplasia , can be considered to induce **cell cycle arrest** of the parathyroid cells , but not cytotoxic effect of **OCT** .

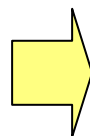
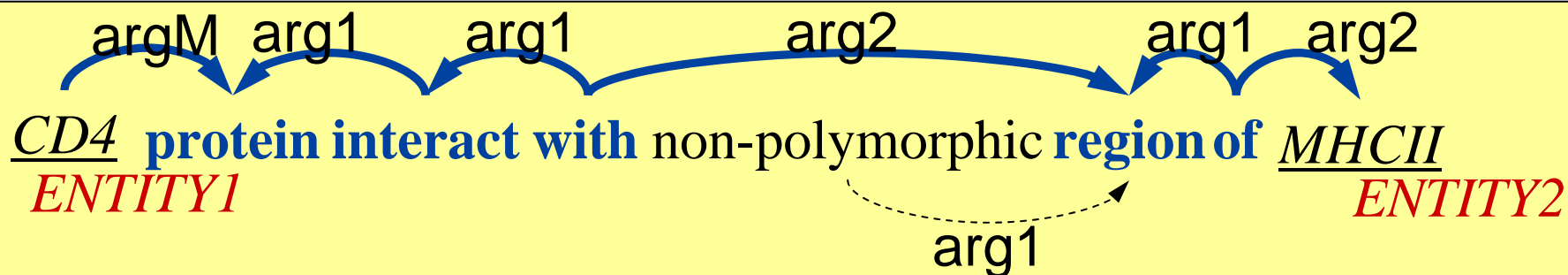
Resistin overexpression is induced by a **beta3** adrenergic agonist in diet-related overweightness .

Extracting protein protein interactions

◆ (Yakushiji, 2005)

CD4 protein interacts with non-polymorphic regions of MHCII .
ENTITY1 *ENTITY2*

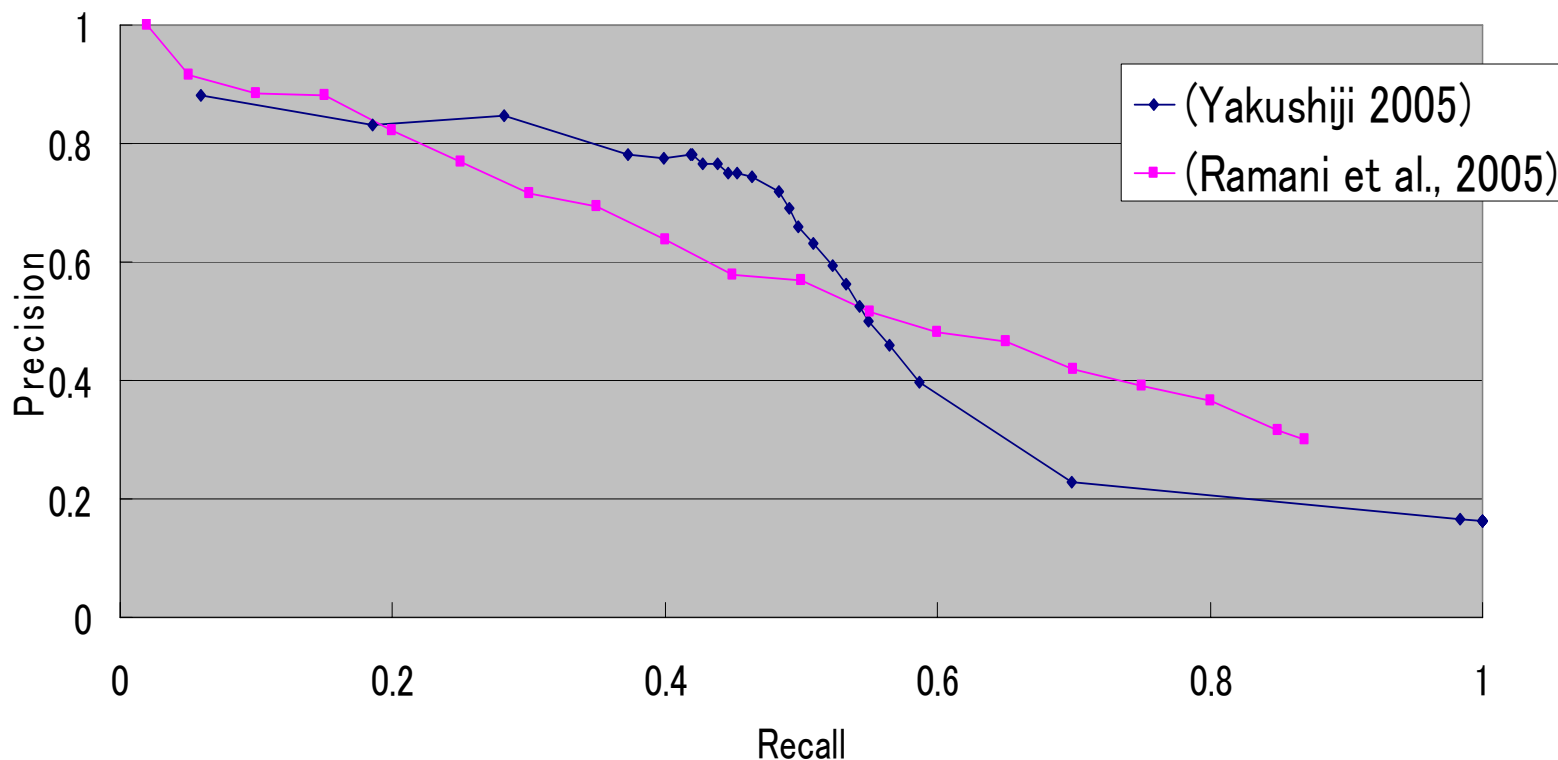
Extraction patterns based on predicate-argument relations



Machine learning with SVM

Extracting protein protein interactions

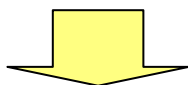
◆ Evaluation with the AImed corpus (Bunescu et al., 2004)



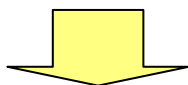


Parsing the MELDINE corpus

◆ 70 million sentences in total



◆ 2 years to finish with a parser that can parse one sentence per second



◆ Parallel processing

⊕ GXP (Taura, 2004)

⊕ PC cluster: 340CPUs



HPSG Parsing in Parallel with GXP

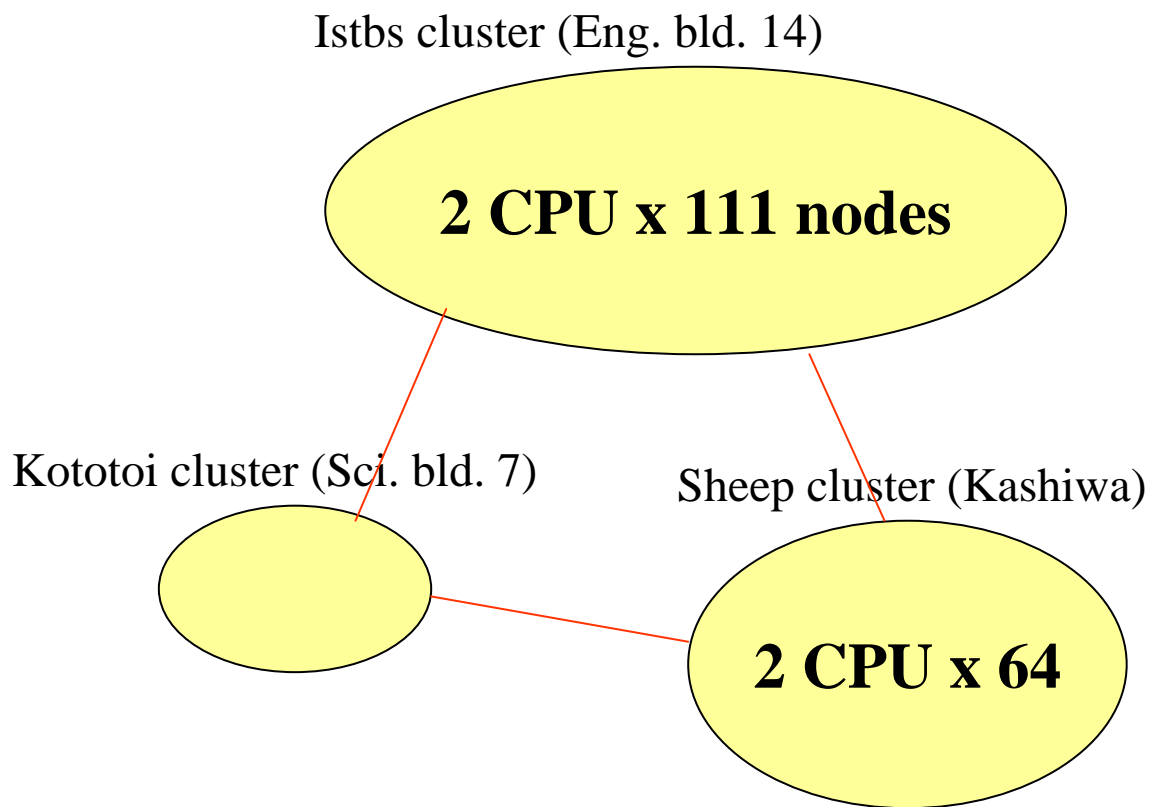
- ◆ Divide the files
 - ⊕ 10,000 files
 - ⊕ 7000 sentences/file



- ◆ Parse in Parallel
 - ⊕ Round robin

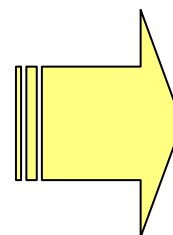
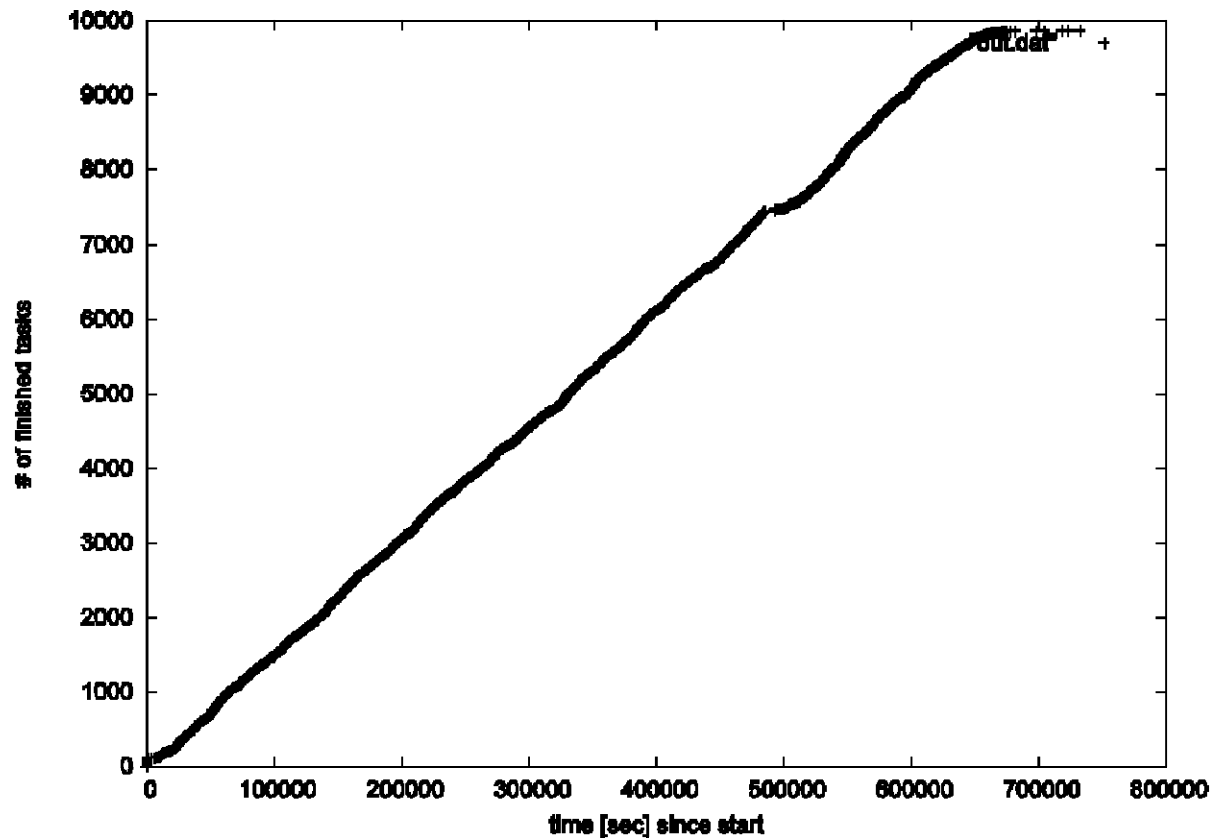


- ◆ Concatenate files





Task Progress



Finished in
9 days



Overview

◆ CytoSailing

- ⊕ Interactive IE system that help biologists build protein-protein interaction networks from MEDLINE

◆ Hiiragi

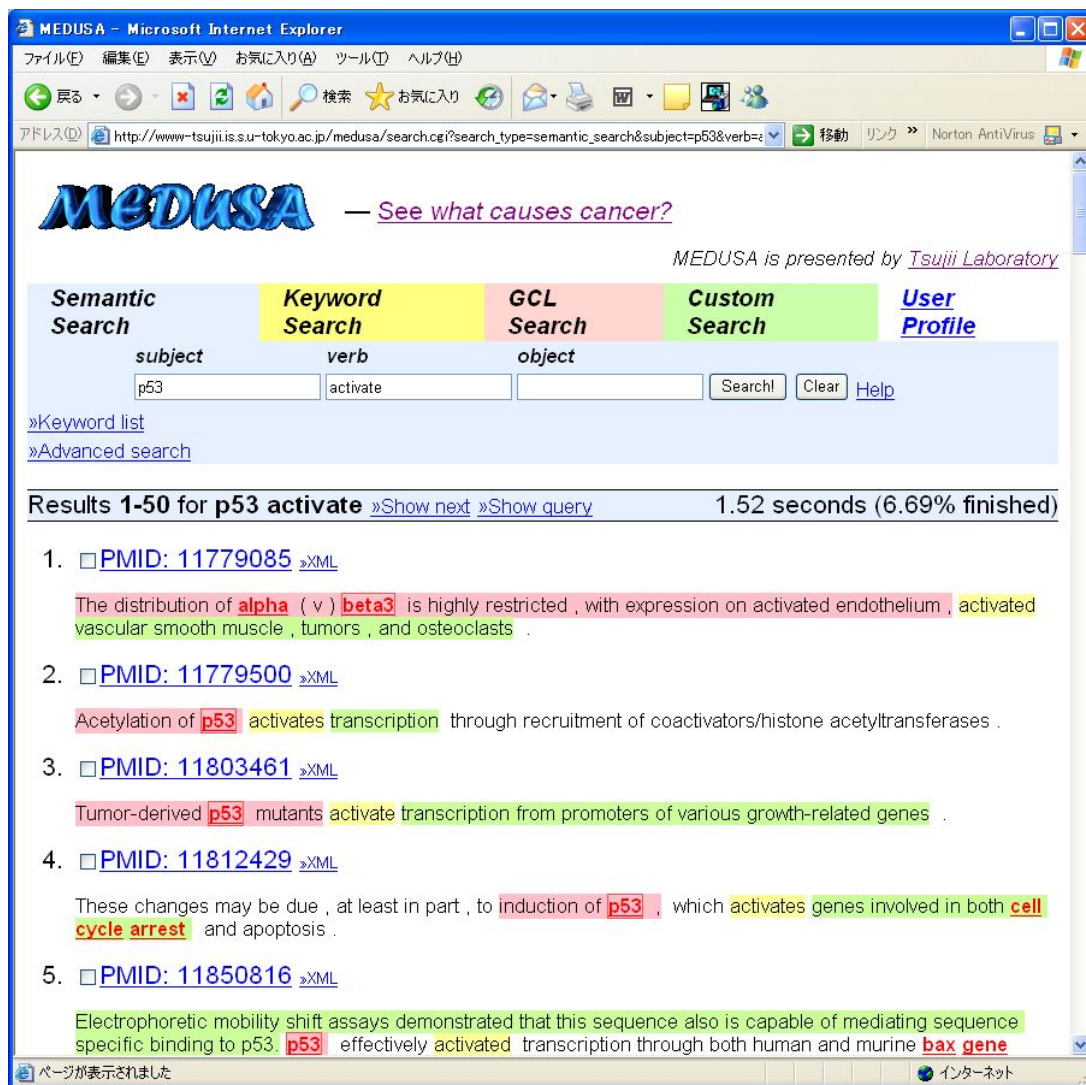
- ⊕ IR system that allows for searches based on predicate-argument relations

Hiiragi

◆ Subject:
p53

◆ Predicate:
activate

◆ Object:
any



MEDUSA — See what causes cancer?
MEDUSA is presented by [Tsuji Laboratory](#)

Semantic Search Keyword Search GCL Search Custom Search User Profile

subject verb object

p53 activate Search! Clear Help

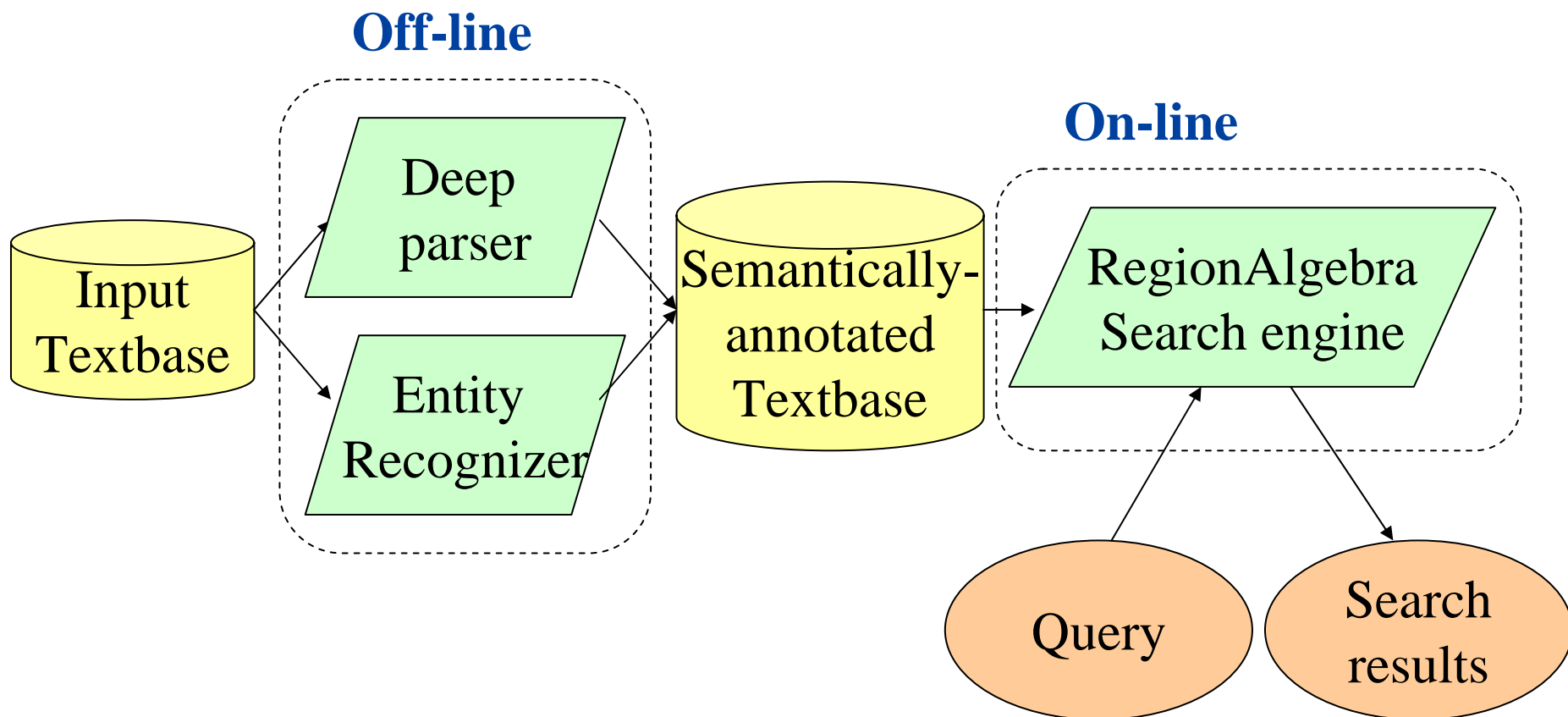
»Keyword list
»Advanced search

Results 1-50 for **p53 activate** »Show next »Show query 1.52 seconds (6.69% finished)

- PMID: 11779085 »XML
The distribution of **alpha** (v) **beta3** is highly restricted , with expression on activated endothelium , activated vascular smooth muscle , tumors , and osteoclasts .
- PMID: 11779500 »XML
Acetylation of **p53** activates transcription through recruitment of coactivators/histone acetyltransferases .
- PMID: 11803461 »XML
Tumor-derived **p53** mutants activate transcription from promoters of various growth-related genes .
- PMID: 11812429 »XML
These changes may be due , at least in part , to induction of **p53** , which activates genes involved in both **cell cycle arrest** and apoptosis .
- PMID: 11850816 »XML
Electrophoretic mobility shift assays demonstrated that this sequence also is capable of mediating sequence specific binding to p53. **p53** effectively activated transcription through both human and murine **bax gene**

ページが表示されました インターネット

Hiiragi system overview





Ontology of Event Verbs

Event type	Expressions
influence	effect, affect, role, response, ...
regulation	mediate, regulate, regulation, ...
activation	induce, activate, activation, ...
:	:



Evaluation of IR performance

- ◆ Queries
 - ⊕ 10 queries
- ◆ Answers
 - ⊕ 100 sentences for each query
- ◆ Judgment
 - ⊕ One biologist
- ◆ Compared with keyword search



Queries used in the experiment

Query No.	Users' input
1	<something> inhibit ERK2
2	<something> trigger diabetes
3	adiponectin increase <something>
4	TNF activate IL6
5	dystrophin cause <disease>
6	macrophage induce <something>
7	<something> suppress MAP phosphorylation
8	<something> enhance p53 (negative)



Performance

Query No.	Keyword search	Semantic search	
		Precision	Recall
1	71%	97%	48%
2	48%	87%	38%
3	23%	79%	79%
4	12%	98%	50%
5	73%	98%	51%
6	27%	72%	27%
7	41%	80%	17%
8	11%	88%	36%



Errors (false positive)

Type of error	Freq.
Term recognition error	28
Parsing error	15
Need to analyze noun phrase structure	12
Other	8

Examples of errors

◆ Term recognition error

“Reduced TNF-alpha secretion **cause** resistance of **db** / db mice to endotoxin”

↙ **Diabetes? No.**

◆ Noun phrase structure

“Nitric oxide, released by activated **macrophages** and endothelial cells, **induces** another type of apoptosis. ”

“**macrophages**” is not the head of the phrase



Errors (false negative)

Type of error	Freq.
Nominal expression	26
Compound verbs	14
Coreference resolution required	12
Parsing error	9
Shortage of event verbs	8
Other	14



Examples of errors

◆ Nominal expression

“The **inhibition** of ERK1 and **ERK2** had no effect on the induction of Prx I expression .”

◆ Compound verb

“The gluconeogenic pathway **contributes** to the fasting hyperglycemia of type II **diabetes**. ”



Summary

◆ CytoSailing

⊕ Interactive IE system for PPI

<http://www-tsujii.is.s.u-tokyo.ac.jp/CytoSailing/>

◆ Hiiragi

⊕ IR system based on predicate-argument relations

<http://www-tsuiji.is.s.u-tokyo.ac.jp/Hiiragi/>



GENiA

Thank you



CytoSailing

- ◆ Interactive information extraction system for building gene-gene interaction networks
- ◆ System components
 - ⊕ Hiiragi
 - ⊕ PPI extraction by SVM
 - ⊕ Multi-window system
- ◆ (demo)



Hiiragi

- ◆ Information retrieval based on predicate-argument structure
- ◆ System components
 - ⊕ GENIA tagger
 - ⊕ Enju (HPSG parser)
 - ⊕ Dictionary-based named-entity recognition
 - ⊕ Retrieval engine based on region algebra
- ◆ (demo)



HPSG parser

◆ Enju

- ⊕ Wide coverage HPSG parser

- ⊕ Accuracy on the Penn Treebank

 - ⊕ precision 87.12% recall 85.45%

- ⊕ Publicly available

 - ⊕ <http://www-tsuji.is.s.u-tokyo.ac.jp/enju/>

◆ Performance on the GENIA corpus

- ⊕ original: 85.1 (f-score)

- ⊕ tuned: 86.9 (f-score)