

Text Mining for Knowledge Discovery and Ontology Extension



Jong C. Park

KAIST, Korea

park@nlp.kaist.ac.kr
nlp.kaist.ac.kr/~park



Workshop on Text Mining, Ontology and
Natural Language Processing in Biomedicine
March 20-21, 2006

Overview

1

Introduction

2

Applications for Biologists

Information Extraction

Knowledge Induction

Convenient Interface

3

Conclusion

Introduction: Techniques for Text Mining in Biomedicine

- Many techniques of NLP are utilized.

Grammar-based parsing

CFG, HPSG, CCG ...

Statistical approach

Co-occurrence ...



Pattern matching

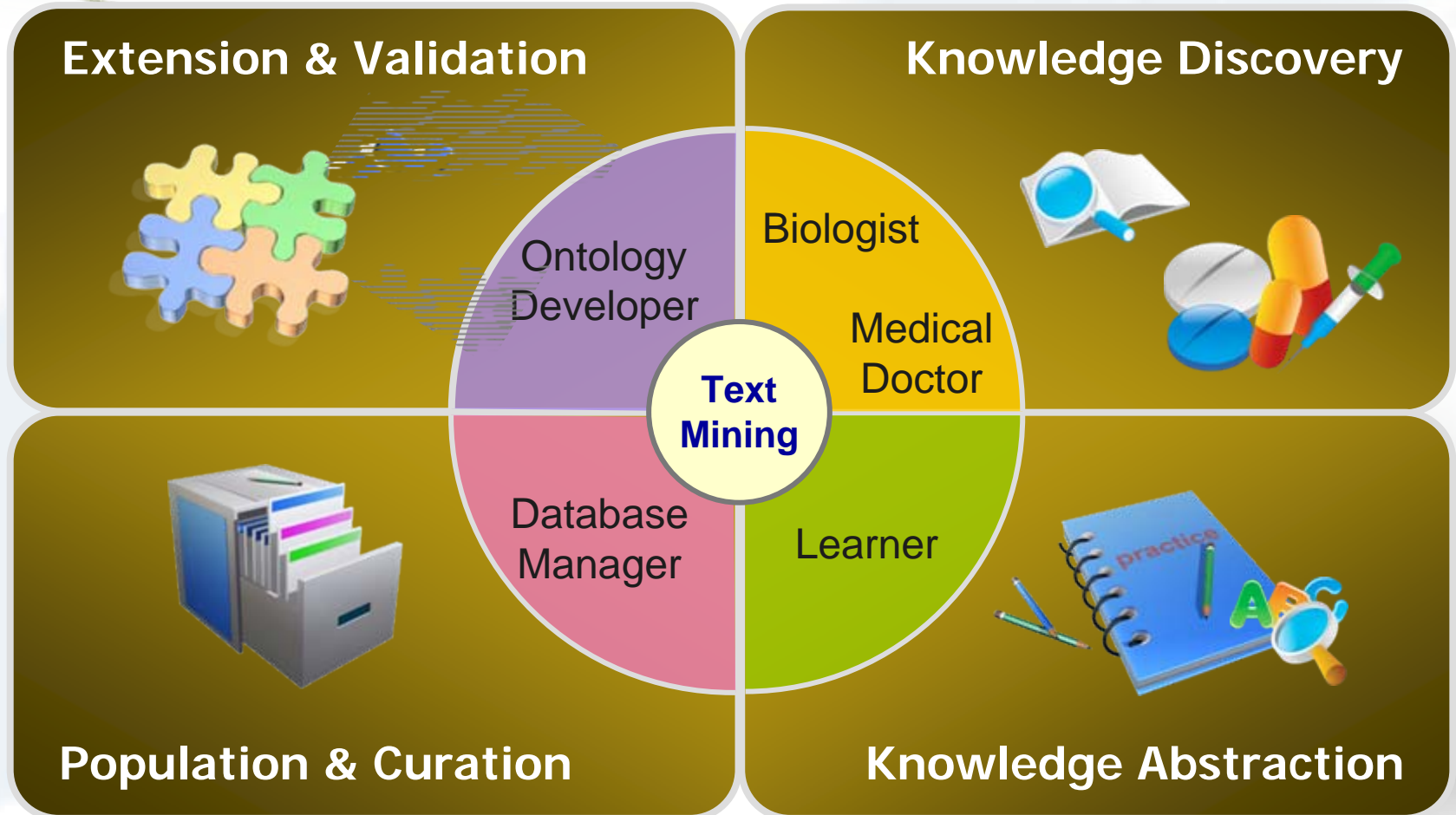
Regular expression

Machine learning

HMM, SVM, MEM ...

Introduction: Applications of Text Mining

- Text mining is helpful in many areas.



Introduction: Towards Better Performance of Text Mining

- Text mining may require sophisticated analyses.

Term variations in the literature



Word level

- Morphological variations
- Semantic variations:
Synonyms, Hyponyms



Syntax level

- The left-to-right order of component words may not always be the same as that of a corresponding term.



Discourse level

- Component words of a term may be distributed across multiple sentences.

Research Directions of Our Group at KAIST

Convenient Interface

BiopathwayBuilder

Visualization of
molecular interaction

BioNLQ

NL query for
heterogeneous DB

NLP techniques for Bioinformatics

Summarization

Generation of
gene summary

BioIE

Automatic extraction
of p-p interaction

Knowledge Induction

AutoGO

Automatic extension
of Gene Ontology

Information Extraction

BioContrasts

Automatic extraction of
contrastive information

Information Extraction

BioIE

BioContrasts



Jung-jae Kim

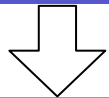
jjkim@nlp.kaist.ac.kr
nlp.kaist.ac.kr/~jjkim

- Information extraction of biological relation
 - Extraction of general biological interactions of arbitrary types, including protein-protein interaction

- Characteristics of BioIE
 - Unknown word handling
 - Semantic class identification
 - Analysis of important linguistic constructions such as acronyms, appositive structures, and anaphoric expressions

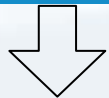
BioIE: Example Procedure

MEDLINE abstract



Keyword detection + Pattern matching (A inhibited B)

... **inhibited** ...



Selection of multiple pairs of candidate arguments (A, B)

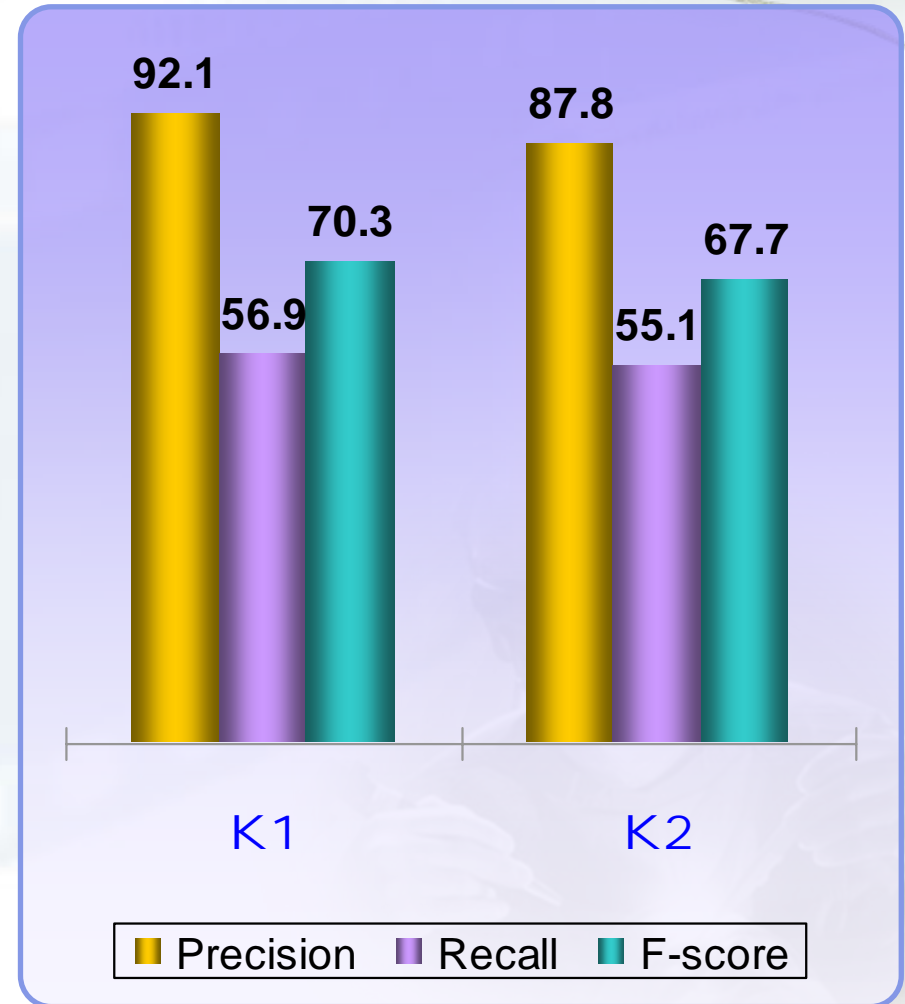
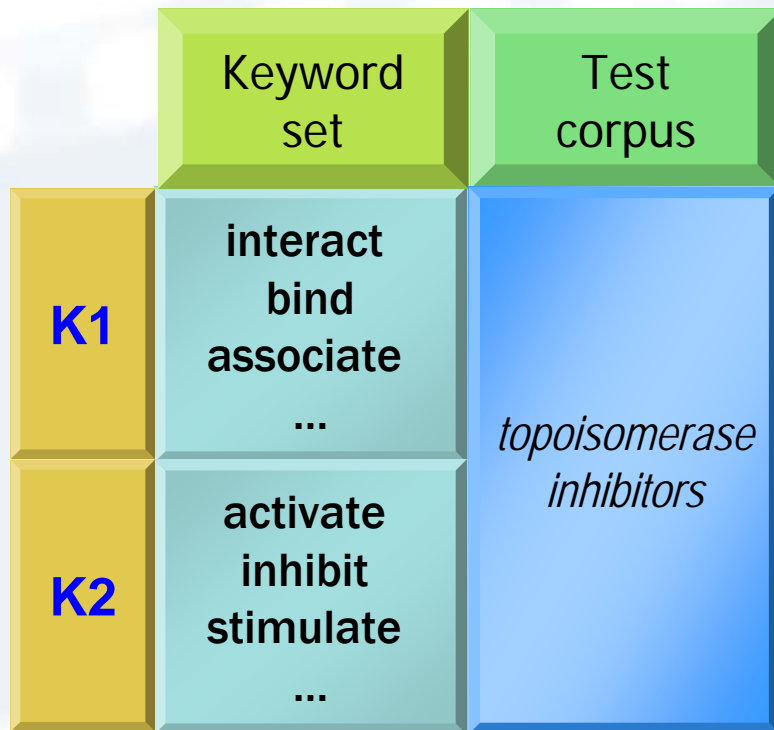
As assessed by [a genetic assay that measures [AAI-dependent DNA binding, [TraM]]] **inhibited** [[[TraR function] before and after the transcription factor] had bound to its DNA recognition site].



Parsing for grammaticality inspection with CCG

As assessed by a genetic assay that measures AAI-dependent DNA binding, **[TraM] inhibited [TraR function]** before and after the transcription factor had bound to its DNA recognition site.

BioIE: Experimental Results



BioContrasts

BioContrasts: extracting and exploiting protein-protein contrastive relations from biomedical literature

Bioinformatics, 22(5): 597-605, 2006

Jung-jae Kim, Zhou Zhang, Jong C. Park, and See-Kiong Ng

Contrasts

- Richly informative units of linguistic expressions from the biomedical literature
- Express explicit difference and implicit similarity in functions

BioContrasts

- Existing protein-protein interaction databases capture mostly positive and individual relations
- Current text mining work has also focused on extracting positive and individual relationships
- Our database enables biologists to exploit such a rich resource of contrastive information already available in the literature

BioContrasts: Example Procedure

“both eIF-4B and eIF-4F, but not eIF-4A, interact with ribosomes in the presence of specific factors and ATP”



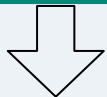
Identify Coordination

Expression for positive object: “both eIF-4B and eIF-4F”
Expression for negative object: “eIF-4A”



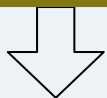
Extract Contrastive Protein Names

Positive protein names: “eIF-4B”, “eIF-4F”
Negative protein names: “eIF-4A”



Ground Protein Names

Positive Swiss-Prot entries: “IF4B_HUMAN”, “IF4B_YEAST”,
“IF4F1_YEAST”, “IF4F2_YEAST”
Negative Swiss-Prot entries: “IF4A_SCHPO”, “IF4A_YEAST”



Extract Presupposed Property

“X interact with ribosomes”, where
X = “IF4B_YEAST” or “IF4F1_YEAST” or “IF4F2_YEAST” and X ≠ “IF4A_YEAST”

BioContrasts: Results

- System implemented in Python

 - Enhanced performance: 0.038 sec/abstract

- Extraction corpus

 - 2.5 million 'not'-containing MEDLINE abstracts

- Contrast extraction

 - 799,169 pairs of contrastive expressions

 - 11,284 pairs of contrastive protein names

 - 41,471 contrast btw Swiss-Prot entries

- A web-portal for public access

BioContrasts: Evaluation

Evaluation set

- 100 randomly selected protein-protein contrasts

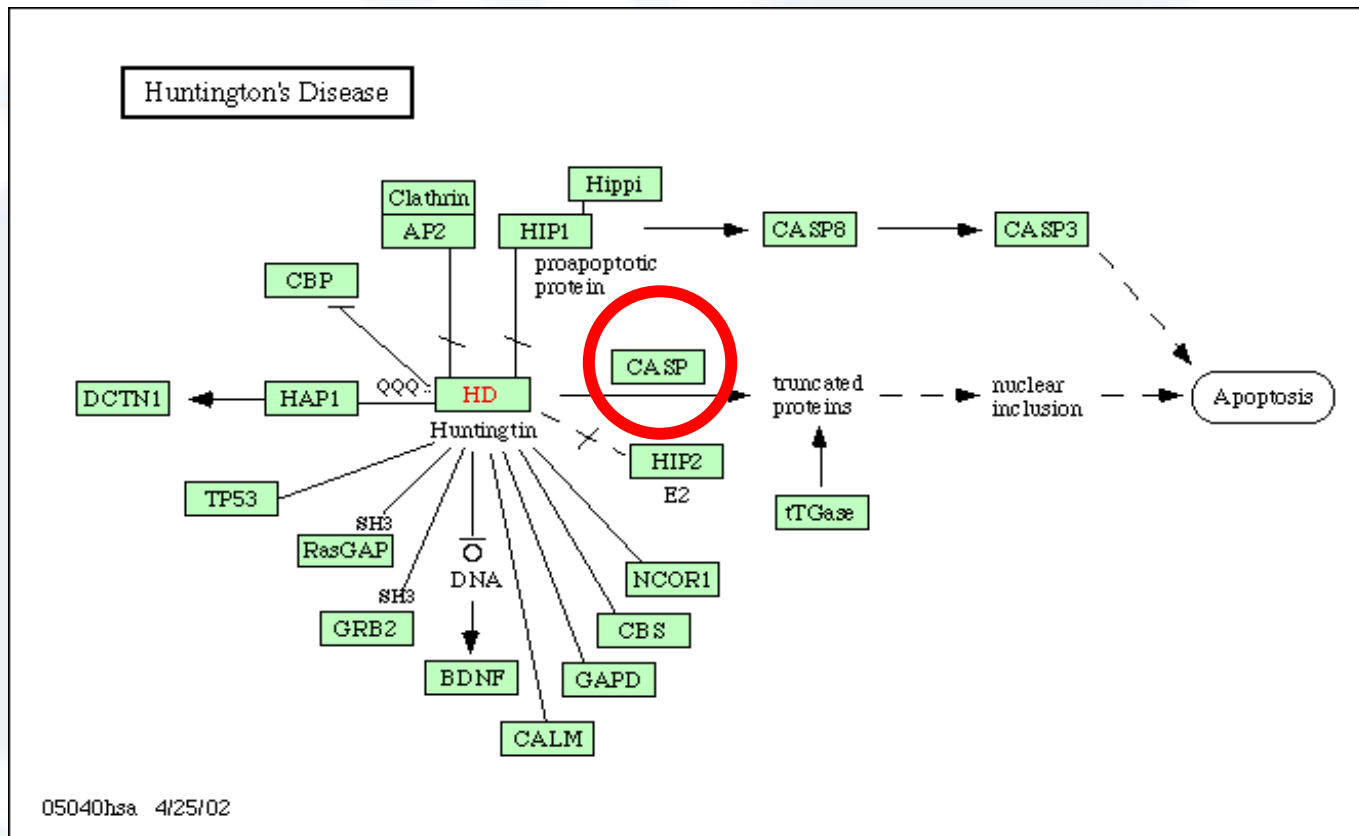
Indicative results

- 97/100 (97.0% precision)
on extraction of contrastive protein names
- 164/182 (90.1% precision)
on protein name grounding (18 names without clarifying phrases)

BioContrasts: Case Study

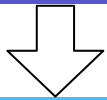
Huntington's disease

- Propose non-obvious (e.g. non-homologous) candidate proteins in incomplete pathways



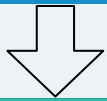
BioContrasts: Case Study

The pathway of **Huntington's disease (HD)** involves the neurotrophic factor (*BDNF*)



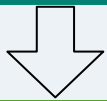
Find any contrast involving a protein member

BioContrasts includes a contrast between *BDNF* and *CNTF*, another neurotrophic factor, with the evidence of PMID:8584263



MEDLINE search

A MEDLINE search with "*BDNF CNTF Huntington's disease*" leads us to find an abstract (PMID:12062094)



Identify relation of the candidate with the pathway

For example, we have shown using an *in vitro* neuronal model of HD that *CNTF* and *BDNF* block polyQ-huntingtin-induced cell death. *In vivo*, *CNTF* has also been shown to be neuro-protective in rats and monkeys following excitotoxic lesions that reproduce HD. (PMID:12062094)

BioContrasts: Web-Portal for Public Access

<http://biocontrasts.biopathway.org>

<http://biocontrasts.i2r.a-star.edu.sg>

BioContrasts Database
Database of protein-protein contrastive information

HOME SEARCH APPLY FAQ CONTACT

KEGG pathway: [hsa05040](#) (Huntington's disease)

Application 1-1. Identify contrastive pathway members that may have different roles

Original KEGG Member	Contrastive Member 1	Contrastive Member 2	Evidence	Relevant Literature Search
CASP	CASP3_HUMAN (Caspase-3 precursor)	CASP1_HUMAN (Caspase-1 precursor)	1	<input type="button" value="PUBMED"/>
CASP	CASP6_HUMAN (Caspase-6 precursor)	CASP3_HUMAN (Caspase-3 precursor)	1	<input type="button" value="PUBMED"/>
Clathrin	CLCA_HUMAN (Clathrin light chain A)	CLCB_HUMAN (Clathrin light chain B)	1	<input type="button" value="PUBMED"/>

Application 1-2. Suggest candidate members of KEGG pathway

Candidate Member	Related KEGG Member	Relevant Literature Search
NT5_HUMAN (Neurotrophin-5 precursor)	BDNF_HUMAN (Brain-derived neurotrophic factor precursor)	<input type="button" value="PUBMED"/>
CNTF_HUMAN (Ciliary neurotrophic factor)	BDNF_HUMAN (Brain-derived neurotrophic factor precursor)	<input type="button" value="PUBMED"/>
FGF2_HUMAN (Heparin-binding growth factor 2 precursor)	BDNF_HUMAN (Brain-derived neurotrophic factor precursor)	<input type="button" value="PUBMED"/>
CASP3_HUMAN	BDNF_HUMAN	

Jin-Bok Lee

jblee@nlp.kaist.ac.kr
nlp.kaist.ac.kr/~jblee



**Knowledge
Induction**

AutoGO

Summarization

Extension of Gene Ontology

- Prediction of more detailed terms of GO
- Induction of the terms with context-sensitive rules from syntactic relations among the existing terms

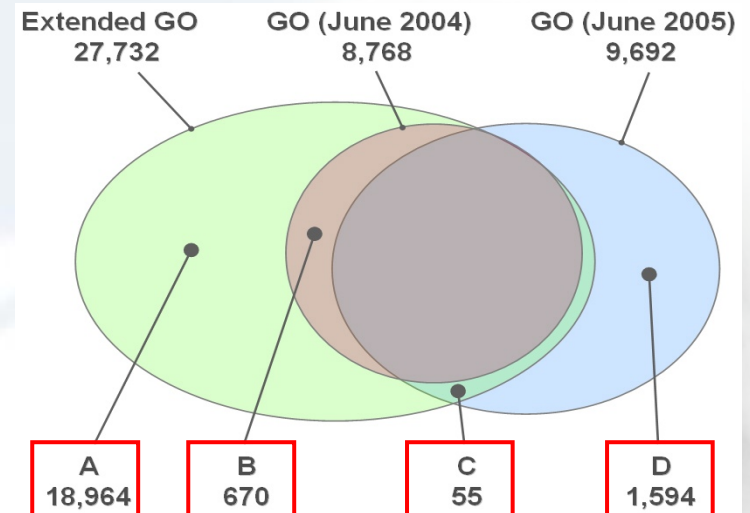
Validation of the extended GO

- Validation of the candidate terms with identification of the terms
- Flexible identification of terms using syntactic dependencies in the literature

AutoGO: Extension of GO

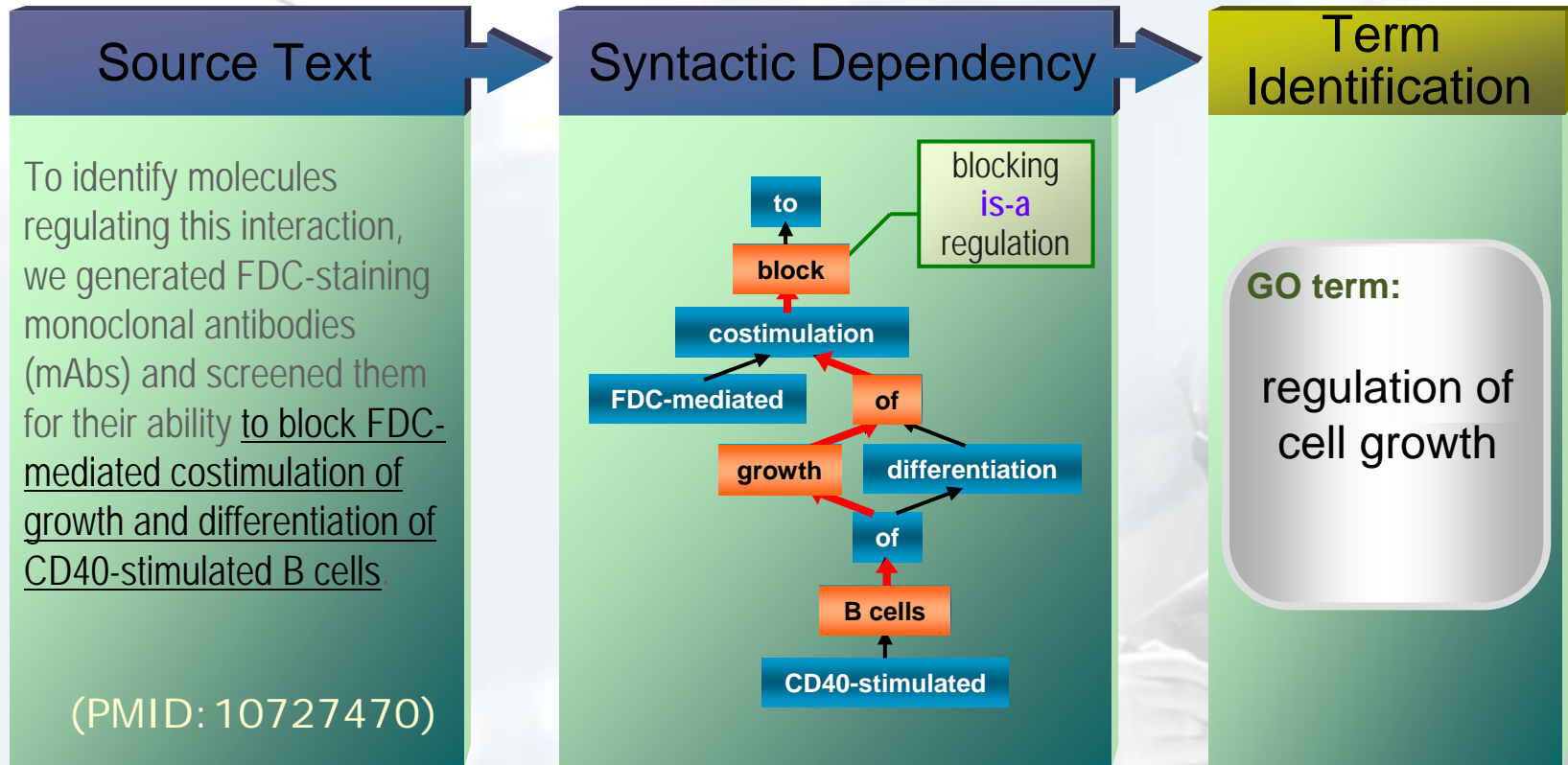
Comparison among 3 versions of GO

- Automatically predicted concepts that are newly introduced by domain experts in the more recent version of GO.
- Automatically predicted concepts that are NOT introduced by domain experts yet.
- Modification by domain experts
- Discarded concepts
- Not predicted by our system



AutoGO: Validation of Newly Introduced Term

Example of validation process



AutoGO: Evaluation of Validation

Experimental results of candidate term validation

		Sentence validation	Abstract validation	Verified terms
Confirmed	Precision	71.9 %	62.7 %	58.2 % (69.1 %)
	Recall	82.1 %	61.0 %	
Not confirmed	Precision	52.4 %	51.7 %	30.0 % (50.0 %)
	Recall	84.6 %	55.6 %	

AutoGO: Results of GO Extension

● Sample part of the extended GO

imaginal disc morphogenesis (549)
clypeo-labral disc morphogenesis (0)
eye-antennal disc morphogenesis (353)
genital disc morphogenesis (13)
halter disc morphogenesis (1)
imaginal disc eversion (4)
imaginal disc fusion (9)
labial disc morphogenesis (0)
leg disc morphogenesis (39)
morphogenesis of larval imaginal disc epithelium (22)
prothoracic disc morphogenesis (0)
regulation of imaginal disc morphogenesis
wing disc morphogenesis (192)

Summarization

● Gene summary

- An effective way to grasp new biological concepts

● Informative summary

- Concept ranking and contrastive information acquisition

● Coherent summary

- Discourse planning and sentence linking

Hodong Lee

hdlee@nlp.kaist.ac.kr
nlp.kaist.ac.kr/~hdlee



**Convenient
Interface**

BioNLQ

BiopathwayBuilder

BioNLQ

● Data search with a natural language query

- Conceptual data search
- Multiple search paths for relevant data
- Unified access to multiple databases

● Explorative search for biological data

- Diverse information access with ranking
- Accurate results and broad range of results
- Guidance over data retrieval process

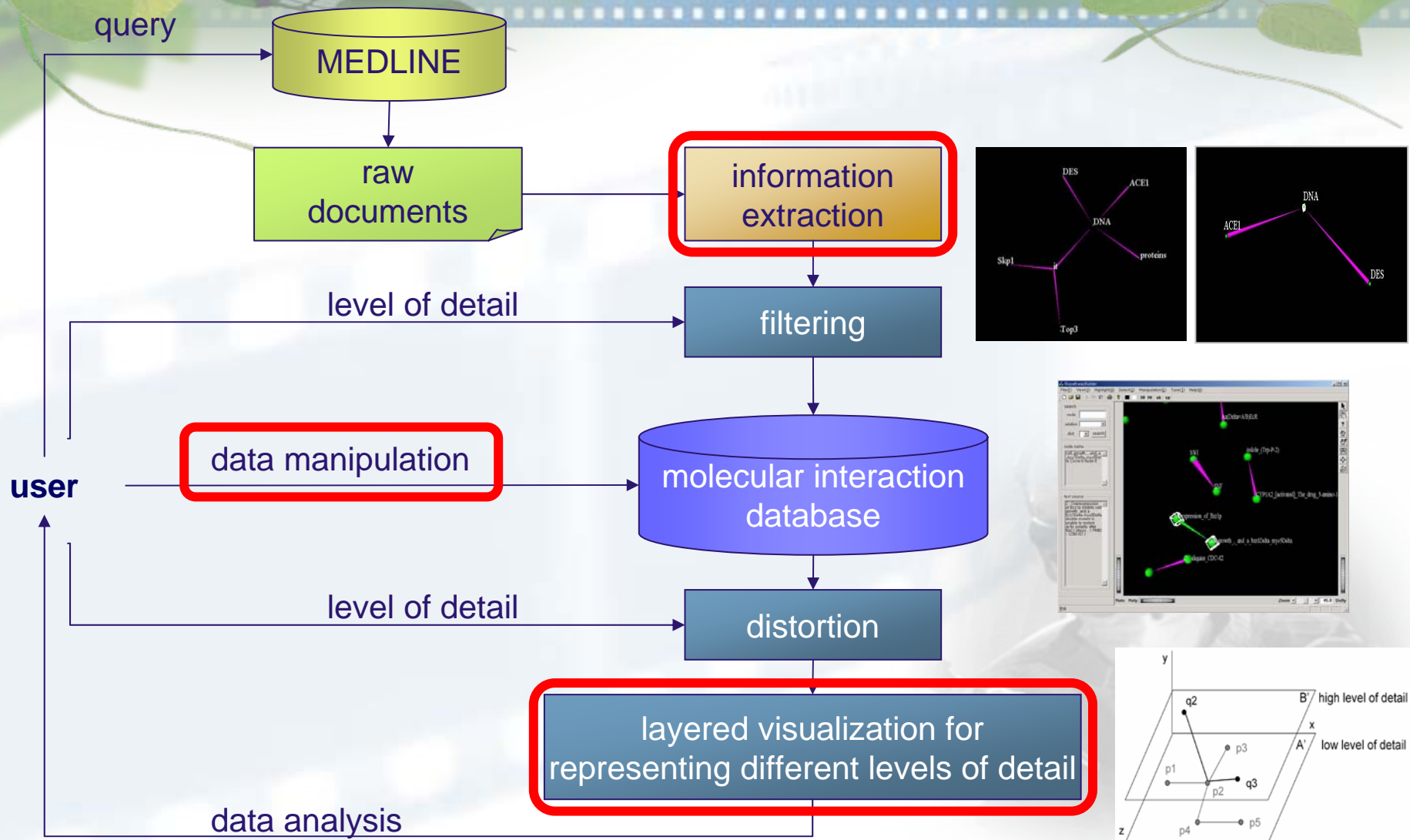
Roles of a visualization system

- To complement the functions of an IE system by guiding the user in his/her inference process over the extracted facts`
- To understand data of a high complexity and to lead to consequent knowledge discovery

Customized visualization

- Customized view using semantic classification
- Layered visualization for representing different levels of detail

BiopathwayBuilder

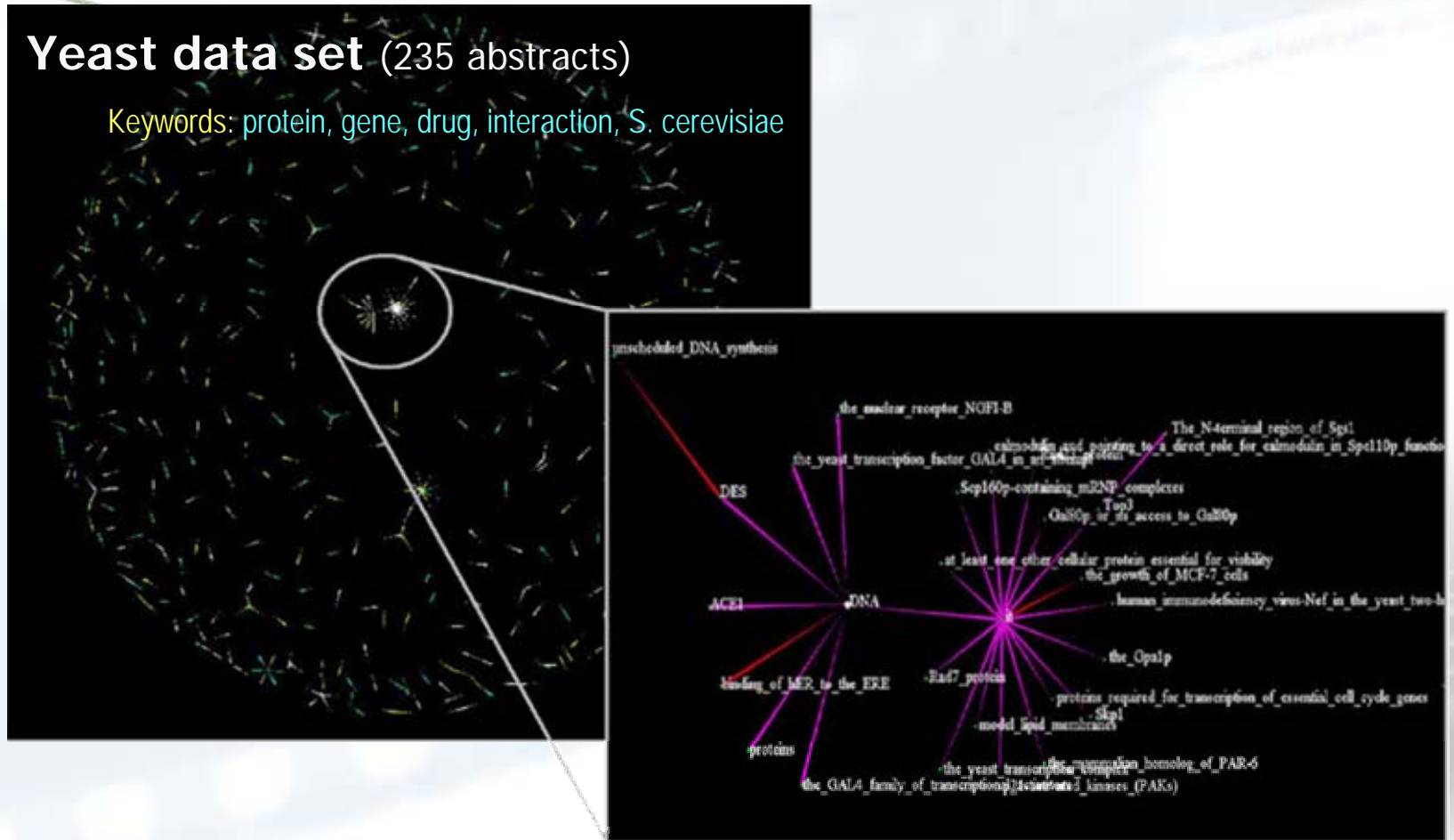


BiopathwayBuilder: Visualization Tool

- Intuitive interface for molecular interactions

Yeast data set (235 abstracts)

Keywords: protein, gene, drug, interaction, *S. cerevisiae*



Summary

● Text Mining

- Techniques for Text Mining in Biomedicine
- Applications of Text Mining
- Towards Better Performance of Text Mining

● Applications for Biologists

- Information Extraction
- Knowledge Induction
- Convenient Interface

Thank You !



Biopathway.org

NLP  .kaist.ac.kr