Automatically Generating Gene Summaries from Biomedical Literature

(Ling et al. PSB 2006)

CS 498 SS
Saurabh Sinha
(Slides courtesy of Xu Ling, UIUC)

Outline

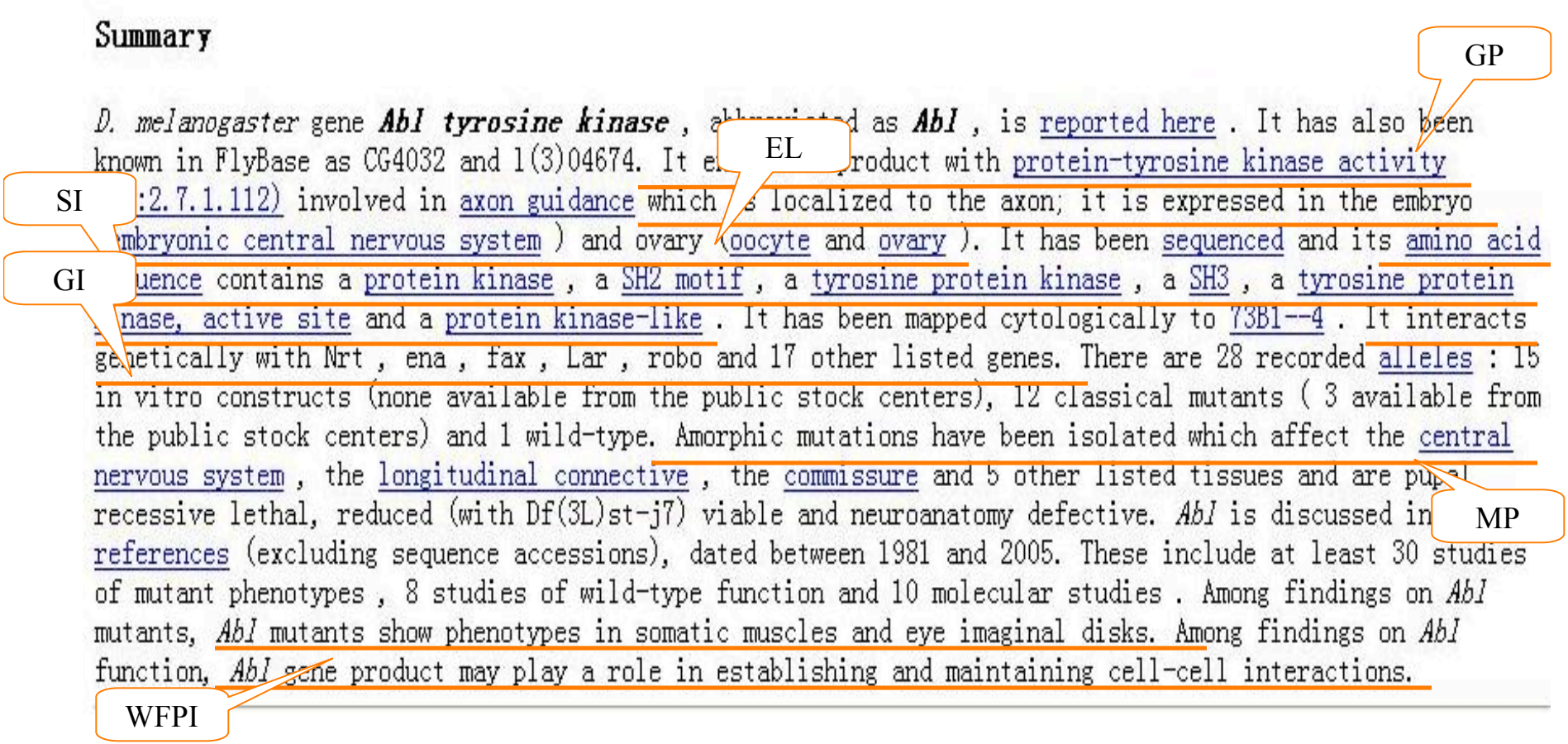
- Introduction
  - Motivation
- System
  - Keyword Retrieval Module
  - Information Extraction Module
- Experiments and Evaluations
- Conclusion and Future Work

Motivation

- Finding all the information we know about a gene from the literature is a critical task in biology research
- Reading all the relevant articles about a gene is time consuming
- A summary of what we know about a gene would help biologists to access the already-discovered knowledge

An Ideal Gene Summary

- http://flybase.bio.indiana.edu/.bin/fbidq.html?FBgn0000017



Problem with Current Situation?

- Manually generated
- Labor-intensive
- Hard to keep updated with the rapid growth of the literature information

#	Symbol	Name NEW	Map	Alleles	Stocks	Refs	DNA acc.
1	Abl	Abl tyrosine kinase	73B1--4	28	7	208	25
2	Abl UPD	-	-	6	-	13	-

FlyBase .. Aberrations .. Anatomy .. BLAST .. Genes .. Annotation/Sequences .. Gene Products .. Maps .. People .. References .. Stocks .. Transgenes/Transposons .|. Help .. Searches .. News .. Site

FlyBase Report - Recent update

Gene *Abl*

This recently updated data is new and has not yet been integrated with the full data set.

How can we generate such summaries automatically?

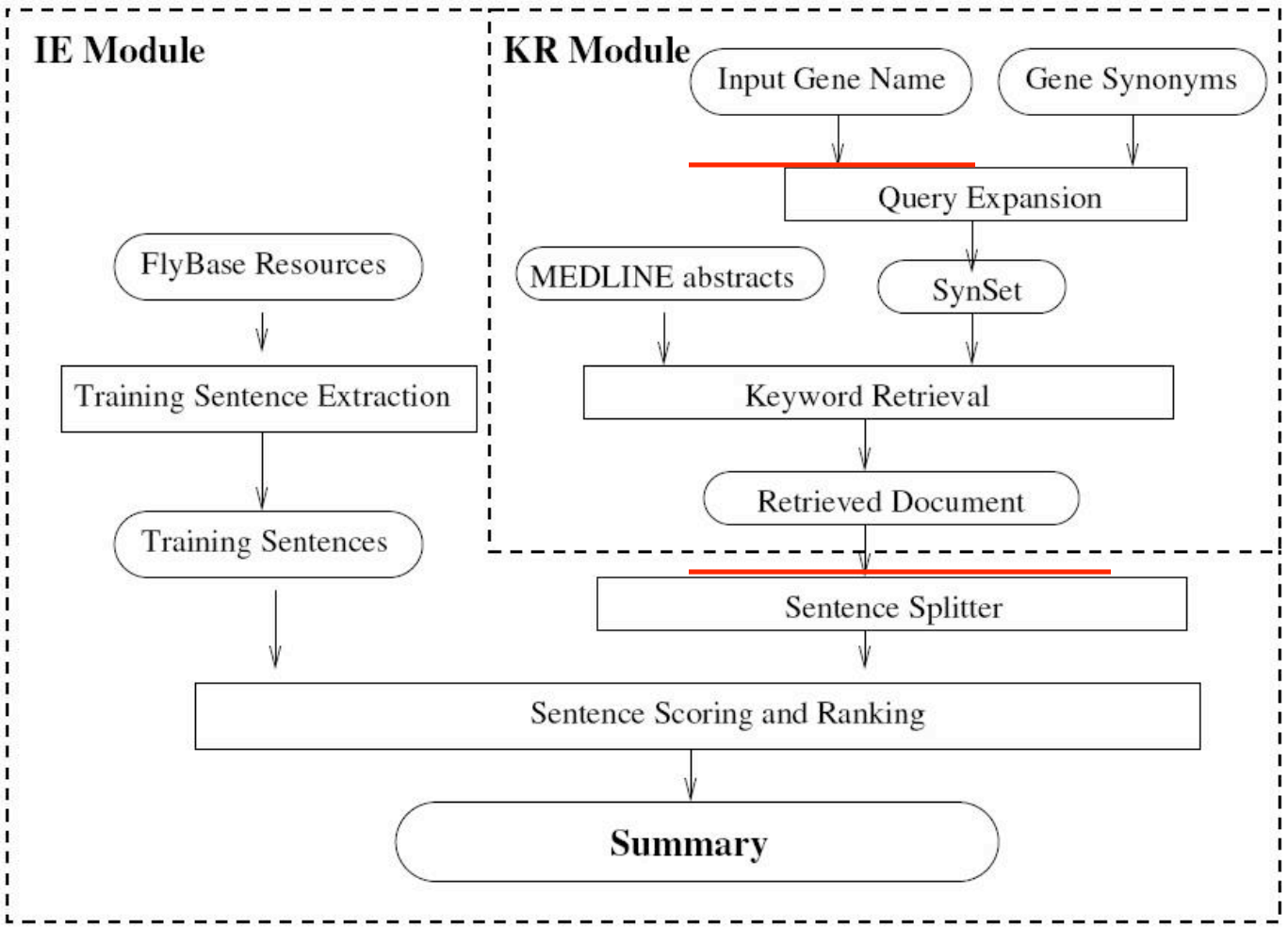
Their solution

- Structured summary on 6 aspects
  1. Gene products (GP)
  2. Expression location (EL)
  3. Sequence information (SI)
  4. Wild-type function and phenotypic information (WFPI)
  5. Mutant phenotype (MP)
  6. Genetic interaction (GI)
- 2-stage summarization
  - Retrieve relevant articles by keyword match
  - Extract most informative and relevant sentences for 6 aspects.

Outline

- Introduction
  - Motivation
- System
  - Keyword Retrieval Module
  - Information Extraction Module
- Experiments and Evaluations
- Conclusion and Future Work

System Overview: 2-stage

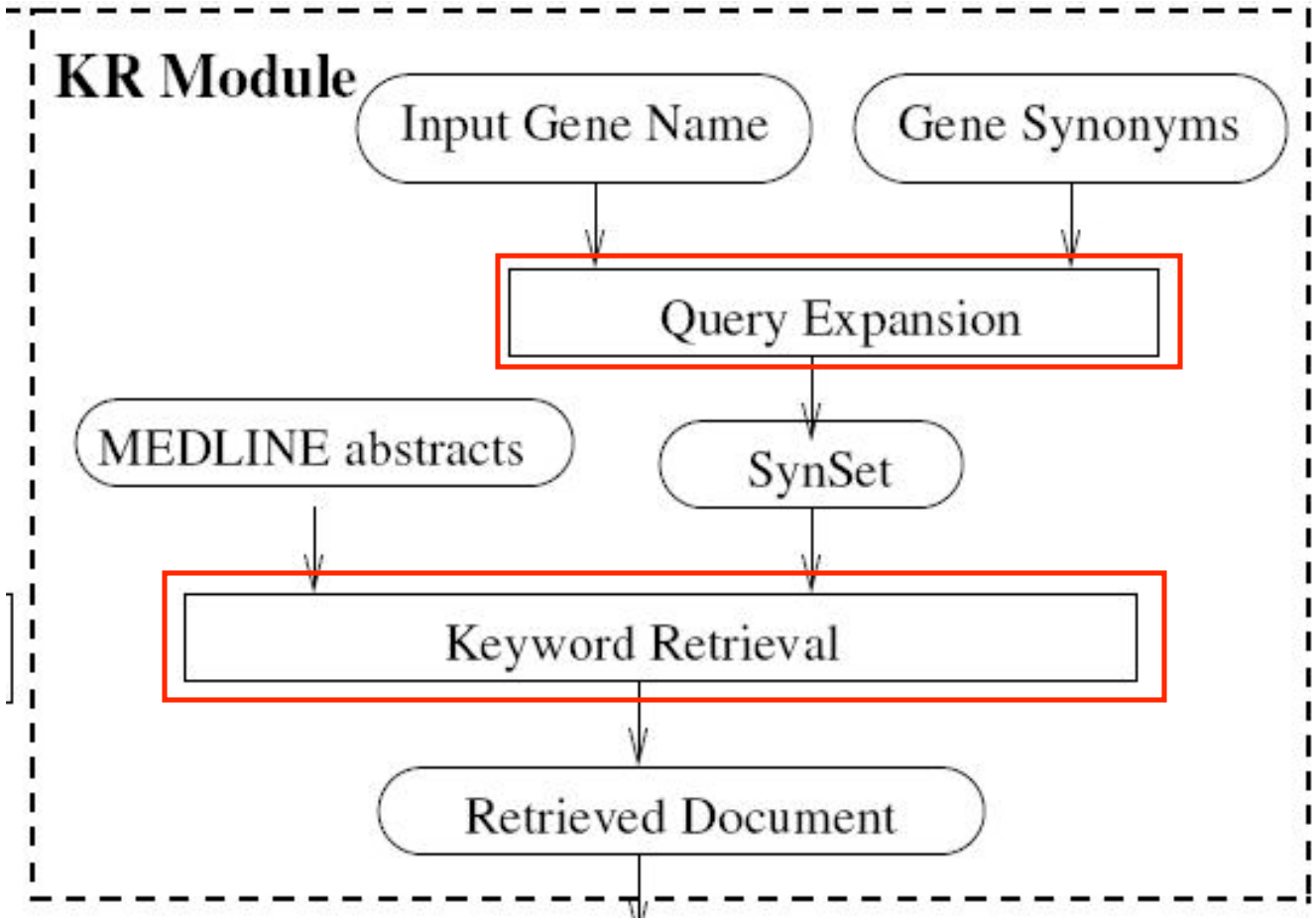


IE = Information Extraction; KR = Keyword Retrieval

Keyword Retrieval Module

- Dictionary-based keyword retrieval: to retrieve all documents containing any synonyms of the target gene.
  - Input: gene name
  - Output: relevant documents for that gene
  1. Gene ***SynSet*** Construction
  2. Keyword-based retrieval

KR module


KR Module
Input Gene Name
Gene Synonyms
Query Expansion
MEDLINE abstracts
SynSet
Keyword Retrieval
Retrieved Document

Gene ***SynSet*** Construction & Keyword Retrieval

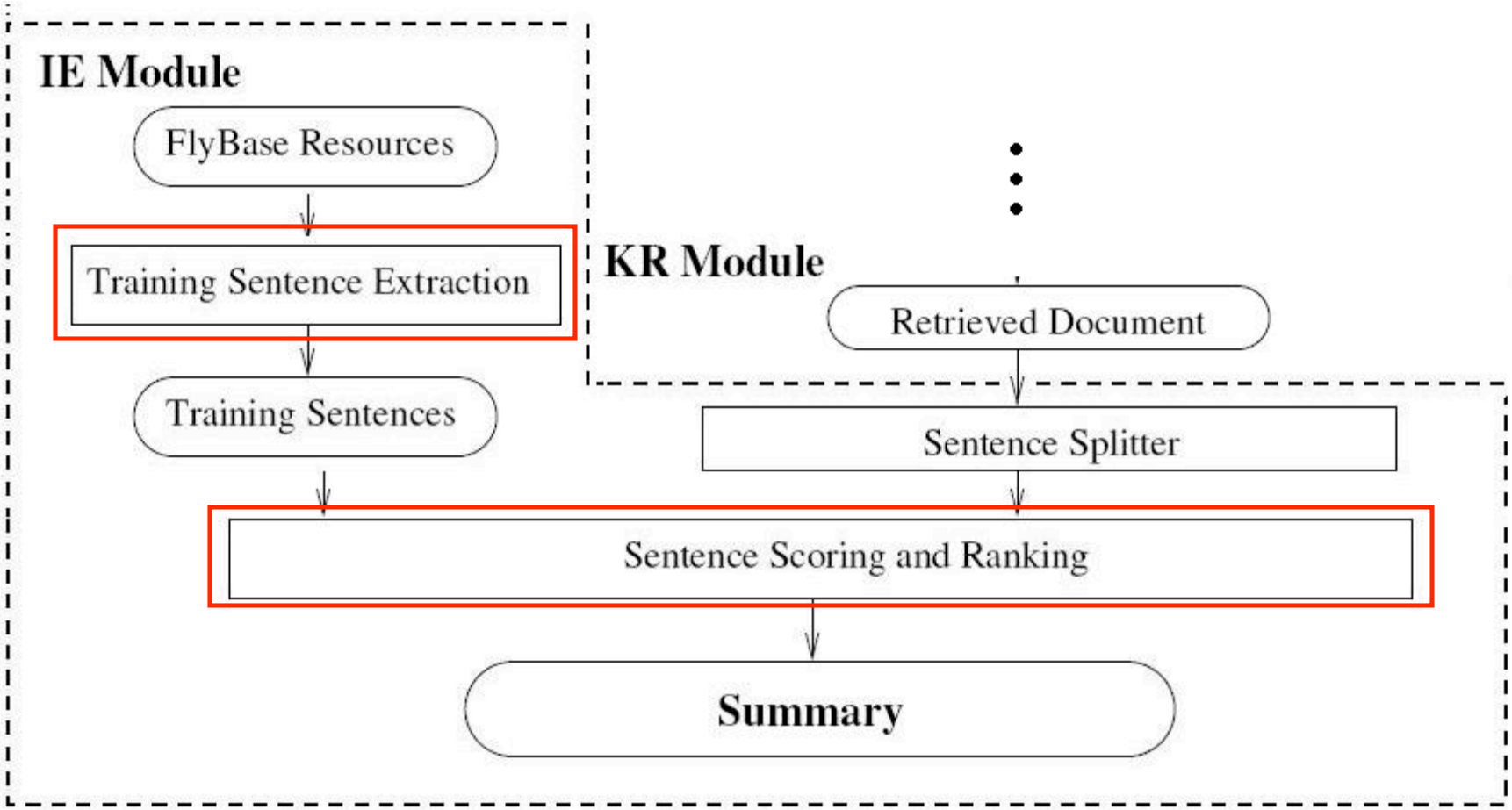
- Gene ***SynSet***: a set of synonyms of the target gene
- Issues in constructing ***SynSet***
  - Variation in gene name spelling
    - gene ***cAMP dependent protein kinase 2***:

      PKA C2, Pka C2, Pka-C2,…
    - normalized to "**pka c 2**"
  - Short names are sometimes ambiguous, e.g., gene name "PKA" is also a chemical term
  - Require retrieved document to have at least one synonym that is >= 5 characters long
- Retrieving documents based on keywords: Enforce the exact match of the token sequence

Information Extraction Module

- Takes a set of documents returned from the KR module, and extracts sentences that contain useful factual information about the target gene.
  - Input: relevant documents
  - Output: gene summary
  1. Training data generation
  2. Sentence extraction

IE module


IE Module
FlyBase Resources
Training Sentence Extraction
Training Sentences
KR Module
Retrieved Document
Sentence Splitter
Sentence Scoring and Ranking
Summary

Training Data Generation

- Construct a training data set consisting of "typical" sentences for describing a category (e.g., sequence information)
- Training data is not about the gene to be summarized. It is about a "type" of information in general.
- These sentences come from a manually curated database
  - e.g., Flybase has separate sections for each category.

Sentence Extraction

- Extract sentences from the documents related to our gene
- Then try to identify key sentences talking about a certain aspect of the gene ("category")
- In determining the importance of a sentence, consider 3 factors
  - Relevance to the specified category (aspect)
  - Relevance to its source document
  - Sentence location in its source abstract

Scoring strategies

- Category relevance score (*Sc*):
  - "Vector space model"
  - Construct "category term vector" V_c for each category c
  - Weight of term t_i in this vector is $w_{ij}=TF_{ij}*IDF_i$
    - TF_{ij} is frequency of t_i in all training sentences of category j
    - IDF_i is "inverse document frequency" = $1+\log(N/n_i)$, N = total # documents, n_i = number of documents containing t_i.
    - TF measures how relevant the term is, IDF measures how rare it is
  - Similarly, vector V_s for each sentence s
  - Category relevant score $Sc = \text{cosine}(V_c, V_s)$

Scoring strategies

- Document relevance score (S_d):
  - Sentence should also be related to this document.
  - V_d for each document, $S_d = \cos(Vd,\ Vs\ )$
- Location score (Sl):
  - News: early sentences are more useful for summarization
  - Scientific literature: last sentence of abstract
  - $S_l = 1$ for the last sentence of an abstract, 0 otherwise.
- Sentence Ranking: $S=0.5Sc+0.3Sd+0.2Sl$

Summary generation

- Keep only 2 top-ranked categories for each sentence.
- Generate a paragraph-long summary by combining the top sentence of each category

Outline

- Introduction
  - Motivation
  - Related Work
- System
  - Keyword Retrieval Module
  - Information Extraction Module
- Experiments and Evaluations
- Conclusion and Future Work

Experiments

- 22092 PubMed abstracts on “Drosophila”
- Implementation on top of Lemur Toolkit
  - Variety of information retrieval functions
- 10 genes are randomly selected from Flybase for evaluation

Evaluation

- Precision of the top k sentences for a category evaluated
- Three different methods evaluated:
  - Baseline run (BL): randomly select *k* sentences
  - *CatRel*: use Category Relevance Score to rank sentences and select the top-*k*
  - *Comb*: Combine three scores to rank sentences
- Ask two annotators with domain knowledge to judge the relevance for each category
- Criterion: A sentence is considered to be relevant to a category if and only if it contains information on this aspect, regardless of its extra information, if any.

		Abl (0000017)	Amy-d (0000078)	Dll (0000157)	eag (0000535)	Gld (0001112)	mam (0002643)	ss (0003513)	DNApol-α73 (0005696)	wts (0011739)	Camo\Sod (0012052)	Avg. Precision
EL	top-1	0/1/1	1/1/1	0/1/0.5	0/1/1	0/1/1	0/1/1	0/1/1	0/1/1	0/0/0	0/1/1	0.1/**0.9**/0.85
	top-2	0/2/2	1.5/2/2	0/2/1.5	0/1/1	0/2/2	0/1/1	0.5/2/1	0/2/2	0/0/0	0/2/2	0.1/**0.8**/0.73
	top-5	1/4/4	1.5/4/3.5	1/4.5/4.5	0/1/1	1/4/4.5	1/2/2	0.5/2/2	1/4.5/3.5	0/0/1	0/3/3	0.14/**0.58**/0.58
	top-10	2.5/5.5/5.5	2/8/8	2/5.5/7	0/2/2.5	4/7.5/7.5	3/5/5	1/3/3	2/6.5/6	0/0/2	1.5/4.5/4	0.18/0.48/**0.51**
GP	top-1	0.5/1/1	0/0/0	0/1/1	0.5/1/1	0.5/1/1	0.5/1/0.5	0.5/1/1	0/1/1	1/1/1	1/0/0	0.45/**0.8**/0.75
	top-2	1/2/2	0/1/1	0.5/2/2	1/2/2	0.5/1.5/1.5	1.5/2/1.5	1/1.5/2	0/1/1.5	1.5/1.5/1.5	1.5/1/1	0.43/0.78/**0.8**
	top-5	2.5/5/4.5	0.5/3.5/3	2/5/5	2/4.5/4.5	1/2.5/2.5	2/4/4.5	3/2/3.5	1.5/3/2	3.5/4.5/4.5	3/2.5/3.5	0.42/0.73/**0.75**
	top-10	4.5/6/8.5	2/4.5/4.5	3.5/8.5/8.5	5.5/9.5/9.5	1.5/5/5	4/8/7	5/2/4.5	2.5/4/4.5	6.5/4.5/9	4.5/5/5.5	0.4/0.57/**0.67**
SI	top-1	0/1/1	0/1/1	0/1/1	0/1/1	0/1/1	0/1/1	0/0/0	1/0.5/0.5	0/1/1	0/1/1	0.1/**0.85**/0.85
	top-2	0/2/2	0/2/2	0/1/1.5	0/2/2	0/1/1	0/2/2	0/0/0	1/1.5/1.5	0/2/2	0/2/2	0.05/0.78/**0.8**
	top-5	1/4/4	2.5/4.5/4.5	0/1/1.5	0/4/4.5	0/3/3.5	1.5/4/3.5	0/0/0	1/4/4.5	0/3/3	0/4/4	0.12/0.63/**0.66**
	top-10	4/6.5/6.5	4.5/7/7	0.5/1/1.5	0/4/5	1/6/7	2.5/6.5/6.5	0/0/0	1.5/8.5/9	0/3/5	1/6.5/6.5	0.15/0.49/**0.54**
MP	top-1	0/0.5/0.5	0/0/0	0/1/1	0/1/0.5	0/1/1	0/0/0	1/1/1	0/0/0	0/1/1	0/0.5/0.5	0.1/**0.6**/0.55
	top-2	0/1.5/1.5	0/0/0	0/1/1.5	0/1.5/1.5	0/2/2	0.5/0.5/0.5	2/1.5/1.5	0/0/0	0/2/2	0/0.5/0.5	0.13/0.53/**0.55**
	top-5	1.5/3.5/3.5	0/0/0	0/1.5/3.5	0/3.5/3.5	0/2/2	1.5/1/1.5	3/3.5/3.5	0/0/0	0/2/3	0.5/1/1	0.13/0.36/**0.43**
	top-10	4/6/7	0/0/0.5	0/1.5/4.5	4/8.5/8	0/2/5	2.5/2/5	4.5/7/8	0/0/0	0/2/4	1.5/3.5/2.5	0.17/0.33/**0.45**
GI	top-1	0/1/1	0/0/0	0/1/1	0/0/0	0/0.5/0.5	0.5/1/1	0/1/1	0/1/1	0.5/1/1	0/0.5/0.5	0.1/**0.7**/0.7
	top-2	0.5/2/2	0/0/0	0/1.5/1.5	0/1/0	0/1/1.5	0.5/1.5/1.5	0/2/2	0/2/2	1.5/2/2	0/0.5/0.5	0.13/**0.68**/0.65
	top-5	1.5/4/4.5	1/0/0.5	0/4/4	1.5/2/3	0/1.5/2	1/4/4	2/4.5/5	0/3.5/3.5	2.5/5/5	1/2.5/2	0.21/0.62/**0.67**
	top-10	3.5/7/7	1/0.5/1.5	0.5/7.5/7	3/6.5/5	0/2.5/3.5	3/8.5/8.5	4.5/7/8	0.5/7/7.5	4.5/7/6.5	2/2.5/3	0.23/0.56/**0.58**
WFPI	top-1	0.5/0.5/0.5	0.5/0.5/0	0.5/1/1	1/1/1	0/0/0	0/0.5/0.5	1/1/1	0/0.5/0.5	0.5/1/1	0.5/0/0	0.45/**0.6**/0.55
	top-2	1.5/1.5/1.5	0.5/1.5/0.5	1/2/2	1.5/2/2	0/0.5/0.5	1/1.5/1.5	2/2/2	1/1.5/1.5	1.5/2/2	1.5/1/1	0.58/**0.78**/0.73
	top-5	3/4/4.5	1.5/3.5/3	4/5/5	3/3.5/3.5	1/2.5/2.5	3.5/4/4	4.5/5/5	4/4.5/4.5	3.5/4.5/4	2/2.5/2.5	0.6/**0.78**/0.77
	top-10	6.5/8/8	2.5/5/5	7/8/8	7.5/8.5/8.5	3.5/5.5/6	6.5/9/9	8.5/9/9	6/7/7	8/8/9	3.5/5/5	0.6/0.73/**0.75**

Note: The FlyBase ID for each testing gene is indicated below the gene symbol, *e.g.*, (0000017) under gene *Abl*. Results for the baseline and our different heuristics are slash-delimited in form of baseline/Category Relevant Score/Weighted Combination Scoring. For instance, 0/1/3 means there are 0, 1, 3 sentences relevant for the results of baseline, Category Relevant Score (CRS), Weighted Combination Scoring (WCS) respectively. The last column of average precision is calculated from the above 10 genes.

Precision of the top-k sentences

cat.	top-k	Avg. Precision *BL*	*CatRel*	*Comb*	cat.	top-k	Avg. Precision *BL*	*CatRel*	*Comb*
EL	1	0.1	**0.9**	0.85	MP	1	0.1	**0.6**	0.55
	2	0.1	**0.8**	0.73		2	0.13	0.53	**0.55**
	5	0.14	**0.58**	0.58		5	0.13	0.36	**0.43**
	10	0.18	0.48	**0.51**		10	0.17	0.33	**0.45**
GP	1	0.45	**0.8**	0.75	GI	1	0.1	**0.7**	0.7
	2	0.43	0.78	**0.8**		2	0.13	**0.68**	0.65
	5	0.42	0.73	**0.75**		5	0.21	0.62	**0.67**
	10	0.4	0.57	**0.67**		10	0.23	0.56	**0.58**
SI	1	0.1	**0.85**	0.85	WFPI	1	0.45	**0.6**	0.55
	2	0.05	0.78	**0.8**		2	0.58	**0.78**	0.73
	5	0.12	0.63	**0.66**		5	0.6	**0.78**	0.77
	10	0.15	0.49	**0.54**		10	0.6	0.73	**0.75**

Discussion

- Improvements over the baseline are most pronounced for EL, SI, MP, GI categories.
  - These four categories are more specific and thus easier to detect than the other two GP, WFPI.
- Problem of predefined categories
  - Not all genes fit into this framework. *E.g.*, gene *Amy-d,* as an enzyme involved in carbohydrate metabolism, is not typically studied by genetic means, thus low precision of MP, GI.
  - Not a major problem: low precision in some occasions is probably caused by the fact that there is little research on this aspect.

Summary example (*Abl*)

GP	The Drosophila melanogaster abl and the murine v-abl genes encode tyrosine protein kinases (TPKs) whose amino acid sequences are highly conserved.
EL	In later larval and pupal stages, abl protein levels are also highest in differentiating muscle and neural tissue including the photoreceptor cells of the eye. abl protein is localized subcellularly to the axons of the central nervous system, the embryonic somatic muscle attachment sites and the apical cell junctions of the imaginal disk epithelium.
SI	The DNA sequence encodes a protein of 1520 amino acids with sequence homology to the human c-abl proto-oncogene product, beginning at the amino terminus and extending 656 amino acids through the region essential for tyrosine kinase activity.
MP	The mutations are recessive embryonic lethal mutations but act as dominant mutations to compensate for the neural defects of abl mutants.
GI	Mutations in the Abelson tyrosine kinase gene show dominant interactions with fasII mutations, suggesting that Abl and Fas II function in a signaling pathway that controls proneural gene expression.
WFPI	We have examined the expression of the abl protein throughout embryonic and pupal development and analyzed mutant phenotypes in some of the tissues expressing abl. abl protein, present in all cells of the early embryo as the product of maternally contributed mRNA, transiently localizes to the region below the plasma membrane cleavage furrows as cellularization initiates.

Summary example (*Camo|Sod*)

GP	Superoxide production by Drosophila mitochondria was measured fluorometrically as hydrogen peroxide, using its dependence on substrates, inhibitors, and added superoxide dismutase to determine sites of production and their topology.
EL	The aim of this study was to ascertain the status of CuZn superoxide dismutase (CuZn-SOD) expression in the central nervous system of Drosophila melanogaster.
SI	Comparison of the Drosophila Cu,Zn SOD amino acid sequences with the Cu,Zn SOD of Bos taurus and Xenopus laevis (whose three-dimensional structure has been elucidated) reveals conservation of all the protein's functionally important amino acids and no substitutions that dramatically change the charge or the polarity of the amino acids.
MP	The gene for cytoplasmic superoxide dismutase (cSOD) maps within this interval, as does low xanthine dehydrogenase (lxd).–Recessive lethal mutations were generated within the region by ethyl methanesulfonate mutagenesis and by hybrid dysgenesis.
GI	Drosophila orthologues of the mammalian Cu chaperones, ATOX1 (a human orthologue of yeast ATX1), CCS (copper chaperone for superoxide dismutase), COX17 (a human orthologue of yeast COX17), and SCO1 and SCO2, did not significantly respond transcriptionally to increased Cu levels, whereas MtnA, MtnB and MtnD (Drosophila orthologues of human metallothioneins) were up-regulated by Cu in a time- and dose-dependent manner.
WFPI	The 2.5 kb clone consists of a wild-type 1.84 kb EcoRI fragment containing the Cu,Zn SOD gene previously isolated in our laboratory, with an insertion of 0.68 kb derived (by an internal deletion) from an autonomous, 2.9 kb P element.

Camo\Sod encodes the protein, CuZn superoxide dismutase, involved in superoxide production. In Drosophila, it is suggested that this gene is expressed in central nervous system. All the protein's important amino acids are conserved in related organisms. The mutation of this gene is known to be lethal.

Summary

Superoxide dismutase, abbreviated as ***Camo\Sod***, is reported here. It has been sequenced. There is one recorded allele, which is wild-type. *Camo\Sod* is discussed in 4 references (excluding sequence accessions), dated between 1992 and 2001.

Outline

- Introduction
  - Motivation
  - Related work
- System
  - Keyword Retrieval Module
  - Information Extraction Module
- Experiments and evaluations
- Conclusion and future work

Conclusion and future work

- Proposed a novel problem in biomedical text mining: automatic structured gene summarization
- Developed a system using IR techniques to automatically summarize information about genes from PubMed abstracts
- Dependency on the high-quality training data in FlyBase
  - Incorporate more training data from other model organisms database and resources such as GeneRIF in Entrez Gene
  - Mixture of data from different resources will reduce the domain bias and help to build a general tool for gene summarization.

References

1. L. Hirschman, J. C. Park, J. Tsujii, L. Wong, C. H. Wu, (2002) Accomplishments and challenges in literature data mining for biology. *Bioinformatics* 18(12):1553-1561.
2. H. Shatkay, R. Feldman, (2003) Mining the Biomedical Literature in the Genomic Era: An Overview. *JCB*, 10(6):821-856.
3. D. Marcu, (2003) Automatic Abstracting. *Encyclopedia of Library and Information Science*, 245-256.

Vector Space Model

- Term vector: reflects the use of different words
- $w_{i,j}$: weight of term t_i in vactor j

$$w_{i,j} = \mathrm{TF}_{i,j} * \mathrm{IDF}_i,$$
$$\mathrm{TF}_{i,j} = c_{i,j},$$
$$\mathrm{IDF}_i = 1 + \log \frac{N}{n_i},$$