

Gene Summarizer and InsectBase Experiments

Xu Ling

Computer Science Department
University of Illinois at Urbana-Champaign

BeeSpace Programmers' Group
Institute for Genomic Biology, UIUC

Oct. 10, 2007

Outline

- Quick review of GS
- Current features
- Existing problems
- Proposed solutions
- New features
- Discussion ...

A quick review of the Gene Summarizer (1)

- Pre-defined six generic aspects for summarizing genes:
 - **GP** (Gene Product): describing the product (protein, rRNA, etc.) of the target gene;
Ex. The eag gene encodes a polypeptide that shares sequence similarities with several different ionic channel proteins...
 - **EL** (Expression Location): describing where the target gene is mainly expressed;
Ex. Dll homologs are expressed in developing appendages in at least six coelomate phyla...
 - **SI** (Sequence Information): describing the sequence information of the target gene and its product;
Ex. The genomic DNA sequences from which these cDNAs are derived extend over 37.5 kb, providing a minimum estimate of the size of the eag transcription unit.

A quick review of the Gene Summarizer (2)

- **WFPI** (Wild-Type Function & Phenotypic Information): describing the wild-type functions and the phenotypic information about the target gene and its product;

Ex. Ultimately, sequence analysis of these cDNAs should enable us to identify the eag polypeptide and to elucidate its role in membrane excitability.

- **MP** (Mutant Phenotype): describing the information about the mutant phenotypes of the target gene;

Ex. The *Drosophila* ether-a-go-go (eag) mutant is responsible for altered potassium currents in excitable tissue.

- **GI** (Genetic interaction): describing the genetical interactions of the target gene with other molecules

Ex. Calcium/calmodulin-dependent protein kinase II phosphorylates and regulates the *Drosophila* eag potassium channel.

A quick review of the Gene Summarizer (3)

- Two-stage approach for generating semi-structured gene summary
 - Retrieving sentences about a gene
(keyword match/NER → precision, recall)
 - Extracting sentences for each specified semantic aspect
(categorization → training sentences, classification algorithm)
- [Demo](#)

Current features (1)

- Improved gene retrieval by NER + automated synonym expansion
 - Key word match: precision not high enough
 - Ex. Screening a cDNA library prepared from silk-producing glands of the black widow spider,...*
 - NER component: recall not high enough
 - Ex. ...beta-alanine biosynthesis is regulated by black.*
 - Using known synonym information from NCBI: build a gene name table/dictionary

Current features (2)

- Dynamical summarization: context-based gene summary
 - Only interested to see summarized information about a honeybee gene in a behavior-related context.
Ex. what are the discoveries of gene “for” about foraging behavior?
 - Solution: dynamic filtering.
Only summarize sentences within a certain collection.

Current features (3)

- Improved summary performance by pre-computing sentence relevance scores
 - Fixed categories: N
 - Fixed master collection/corpus: M
 - ⇒ At most $N \cdot M$ entries to store;
 - ⇒ Quickly retrieve scores by mysql query

Existing Problems (1)

- Gene retrieval
 - Performance is mostly effected by the retrieval step.
 - A simpler problem than NER as we already know the gene name and its synonyms (basically, a gene name ambiguity problem) => expect high recall
 - Irrelevant gene mentions: some popular gene names are frequently mentioned in abstracts for comparison/reference purpose.
- Ex.* It encodes a protein predicted to contain 688 amino acid residues, including 11 zinc finger motifs of the C-2H-2 type in the C-terminal region, that are Kruppel-like in the conservation of the H/C link sequence connecting them.

Existing Problems (2)

- Categorization:
 - Lack of high quality example sentences: training sentences are sentences written by the FlyBase curators to explain their database decisions.
 - Domain bias: only sentences about *Drosophila melanogaster* are used for training the GS. => have problems on summarizing other organisms' genes.

Existing Problems (3)

- Not using user's prior knowledge
- No user control of precision/recall
- Fixed categories => no application-specific categories allowed

Proposed solutions (1)

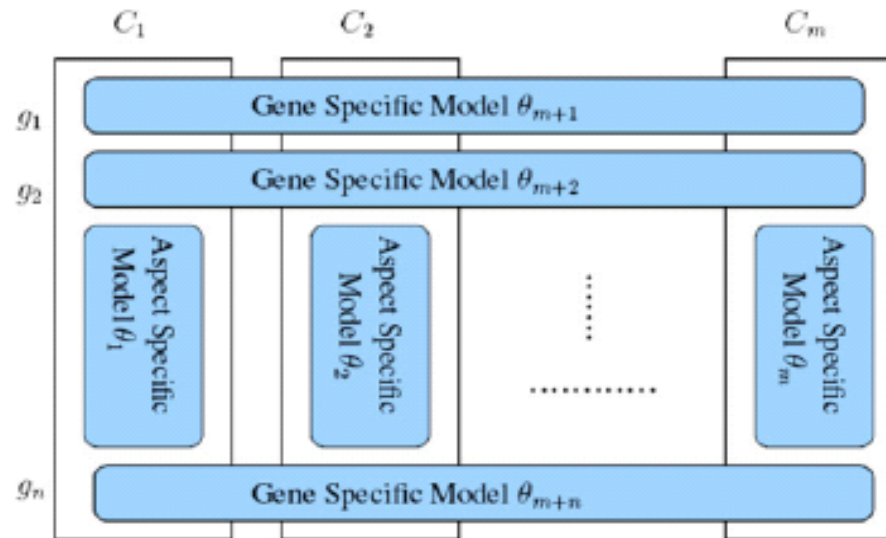
- Gene name disambiguation problem of the gene retrieval component
 - A ***simpler*** problem than NER as we already know the gene name and its synonyms
 - Build a classifier that focuses on contextual features to identify false gene mentions
 - *Ex.* The purpose of this study was to investigate the black gene, and protein...; Screening a cDNA library prepared from silk-producing glands of the black widow spider...
 - Only use contextual features because the term/phrase already matches a gene name
 - Can also solve the problem of popular gene names being frequently mentioned in abstracts for comparison/reference purpose.

Proposed solutions (2)

- Exploit other model organisms to eliminate species-specific bias
 - Collect training sentences from SGD, Wormbase
 - *Ex. Null mutant is viable; aat1 leu2 double mutant is inviable. (MP of yeast gene YKL106W)*
 - *Adenine deaminase (adenine aminohydrolase), involved in purine salvage and nitrogen catabolism (WFPI of yeast gene YNL141W)*
 - *Although AAP-1::GFP expression was weak, fluorescence was consistently observed in the intestine and in neurons and was occasionally observed in body wall muscles and the hypodermis. Expression was observed at all developmental stages beginning with early embryos and continued through adulthood. (EL of C. elegans gene aap-1)*
 - *aap-1 encodes the C. elegans ortholog of the phosphoinositide 3-kinase (PI3K) p50/p55 adaptor/regulatory subunit*

Generalizable Aspect Model (GAM)

- Basic idea: considering each sentence is associated with multiple contexts: aspect, gene, organism...
- Goal: to extract the language models associated with the aspect contexts but not the gene/organism contexts.



Proposed solutions (3)

- Construct “real” sentences for improving categorization
 - Collaboration with Beetle people
- New categories
 - GP + SI => PS (Gene product and protein domain/structure)
 - SI => HO (homologs/orthologs in other species)
 - EL
 - SI => RE (Regulatory elements),
 - WFPI + MP => PHP (Wild-type/mutant phenotype, function)
 - GI => IT (Genetic/Physical interaction)
 - => PG (Population genetics)?

Evaluation

- Randomly selected genes from fruit fly, honeybee, beetle for evaluation.
- Compare the generated summaries for each query gene using different gene retrieval and sentence extraction methods.
- Gold Standard
 - Biologists assign each candidate sentence to relevant aspects separately.
 - The sentences that are decided as relevant to a certain aspect are collected as the gold standard for this gene in the corresponding aspect.

New features (1)

- Allow user control of gene retrieval
 - Enable synonym selections by user: allow user choose which synonyms for which species to be searched, to control the precision.
 - Ex. “black” is recommended as a synonym (for *D. ananassae* “ebony” gene), user can decide whether want GS to extract all the sentences mentioning “black” for summarizing “ebony”.
 - Enable user control of precision/recall
 - A parameter in the gene name disambiguation step.

New features (2)

- Allow user-defined application-specific categories (eg., population genetics).
 - Cluster sentences with very low relevance score to pre-defined categories.
 - Allow user set priors for certain aspects.

New features (3)

- Build knowledge base upon GS
 - Extract key information for each aspect:
sentence → entity-relation table

Discussions ...

Thoughts ?

Questions?

Thanks

A quick review of the NER component

- Use two types of information to make a prediction
 - Word features and word surface features
 - E.g. p53, XXXless
 - Contextual features
 - E.g. XXX expression, XXX mutants
- Prediction of the same word/phrase is context-sensitive

Examples of Some Ambiguous Gene Names

- **foraging**

- We assayed response decrement for natural and mutant rover and sitter alleles of the foraging (for) gene that encodes a Drosophila PKG. (FN)
- Hybrid disadvantage in the larval foraging behaviour of the two neotropical species of Drosophila pavani and Drosophila gaucha... (TN)

Examples of Some Ambiguous Gene Names

- **SS**

- ...SmZF1 binds both ds and ss DNA oligonucleotides,... (TN)
- Coexpression of Ss and Tgo in Drosophila SL2 cells... (TP)
- The origin of germline-limited chromosomes (Ks) as descendants of somatic chromosomes (Ss) and their... (FP)

Examples of Some Ambiguous Gene Names

- **black**

- The purpose of this study was to investigate the black gene, and protein,... (FN)
- ...beta-alanine biosynthesis is regulated by black. (FN)
- Screening a cDNA library prepared from silk-producing glands of the black widow spider,... (TN)

Examples of Some Ambiguous Gene Names

- **clock**

- ...a novel fitness-related phenotype may be linked to noncircadian expression of clock genes in the ovaries. (TP)
- ...mPer1 could operate in the adaptation of the circadian clock of nocturnal mice to... (TN)

Examples of Some Ambiguous Gene Names

- **ERG**
 - To establish the predicted existence of a *Drosophila* gene in the erg subfamily and... (FN)
 - The ERG analysis of the *norpA* mutants suggests that... (TN)
 - Here we show that the electroretinogram (ERG), the extracellular recording...(FP)

Examples of Some Ambiguous Gene Names

- **pdf**

- PDF is coded in a precursor protein together with another neuropeptide... (TP)
- ...the Drosophila brain that express the period (per) and pigment dispersing factor (pdf) genes play... (TP)

Effects of NER on Gene Summarizer

- FP → TN (increase precision)
 - ...mPer1 could operate in the adaptation of the circadian clock of nocturnal mice to... (TN)
- TP → FN (decrease recall)
 - ...beta-alanine biosynthesis is regulated by black. (FN)
- FP → FP (no effect, but not what we want)
 - Here we show that the electroretinogram (ERG), the extracellular recording...(FP)

An Ideal Gene Summary

Summary

D. melanogaster gene ***Abi tyrosine kinase***, abbreviated as ***Abi***, is [reported here](#). It has also been known in FlyBase as CG4032 and 1(3)04674. It encodes a protein product with [protein-tyrosine kinase activity](#) (FlyBase ID: 2.7.1.112) involved in [axon guidance](#) which is localized to the axon; it is expressed in the embryo ([embryonic central nervous system](#)) and ovary ([oocyte](#) and [ovary](#)). It has been [sequenced](#) and its [amino acid sequence](#) contains a [protein kinase](#), a [SH2 motif](#), a [tyrosine protein kinase](#), a [SH3](#), a [tyrosine protein kinase, active site](#) and a [protein kinase-like](#). It has been mapped cytologically to [73B1--4](#). It interacts genetically with [Nrt](#), [ena](#), [fax](#), [Lar](#), [robo](#) and 17 other listed genes. There are 28 recorded [alleles](#): 15 in vitro constructs (none available from the public stock centers), 12 classical mutants (3 available from the public stock centers) and 1 wild-type. Amorphic mutations have been isolated which affect the [central nervous system](#), the [longitudinal connective](#), the [commissure](#) and 5 other listed tissues and are pupal lethal, reduced (with [Df\(3L\)st-j7](#)) viable and neuroanatomy defective. *Abi* is discussed in [references](#) (excluding sequence accessions), dated between 1981 and 2005. These include at least 30 studies of mutant phenotypes, 8 studies of wild-type function and 10 molecular studies. Among findings on *Abi* mutants, [Abi mutants show phenotypes in somatic muscles and eye imaginal disks](#). Among findings on *Abi* function, [Abi gene product may play a role in establishing and maintaining cell-cell interactions](#).

GP

EL

SI

GI

MP

WFPI

Gene Summary of "abl"

[Search Again](#)

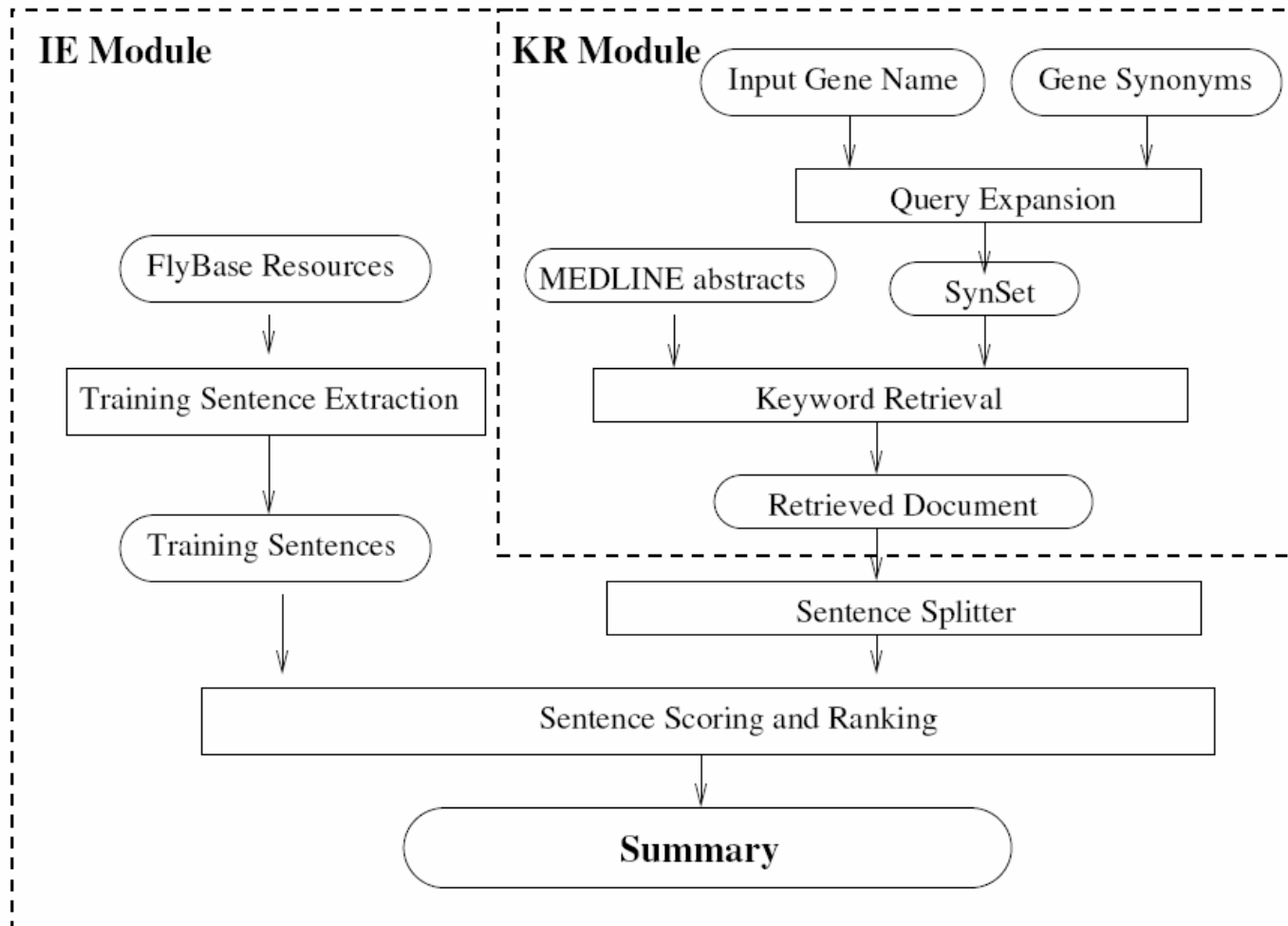
171 sentences found with Gene Summarizer V1

[Show All](#)

[[GP](#) | [EL](#) | [SI](#) | [WFPI](#) | [MP](#) | [GI](#)]

	DOCNO	Sentence
Gene Product (11 found)	3119227	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.
	2832740	The DNA sequence encodes a protein of 1,520 amino acids with strong sequence similarity to the human c-abl proto-oncogene beginning in the type 1b 5' exon and extending through the region essential for tyrosine kinase activity.
	2157882	The Drosophila melanogaster abl and the murine v-abl genes encode tyrosine protein kinases (TPKs) whose amino acid sequences are highly conserved.
	2832740	These results show that the abl gene is highly conserved through evolution and encodes a functional tyrosine protein kinase required for Drosophila development.
	2406026	The Drosophila Abelson (abl) proto-oncogene homolog encodes a cytoplasmic tyrosine kinase that is expressed during embryogenesis primarily in developing CNS axons; abl mutants show no gross defects in CNS morphogenesis.
	2188361	The Drosophila abelson (abl) gene encodes the homolog of the mammalian c-abl cytoplasmic tyrosine kinase and is an essential gene for the development of viable adult flies.
Expression Location (13 found)	1295746	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.
	1295746	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.
	8647396	The fax gene encodes a novel 47-kD protein expressed in a developmental pattern similar to that of Abl in the embryonic mesoderm and axons of the central nervous system.
	9926937	We targeted expression of human/fly chimeric Bcr-Abl proteins to the developing central nervous system (CNS) and eye imaginal disc of Drosophila melanogaster.
	2502313	Embryos that are homozygous mutant for both abl and dab fail to develop any axon bundles in the CNS, although the peripheral nervous system and the larval cuticle appear normal.

System Overview: 2-stage



Keyword Retrieval Module

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

Information Extraction Module

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

Sentence Extraction

- Steps
 1. Compute the relevance score S for each sentence with respect to each aspect
 2. Extract ranked list of sentences for each aspect of the query gene.
- Methods
 - Vector Space Model
 - Probabilistic Methods: Language Modeling Approaches

Vector Space Model (VSM)

- Construct a corresponding term vector V_c using the training sentences for the aspect
 - The weight of a term t_i in the aspect term vector for aspect j :
 $w_{ij} = TF_{ij} IDF_i$, where TF_{ij} = term frequency, $IDF_i = 1 + \log(N/n_i)$ is the inverse document frequency (N = total number of documents, n_i = number of documents containing term t_i).
- Construct a sentence term vector V_s for each sentence
 - with the same IDF and TF = number of times a term occurs in the sentence
- Aspect relevance score $S = \cos(V_c, V_s)$.

Probabilistic Methods: Language Modeling Approaches

- Estimate a language model for each aspect
- Use the negative KL-divergence function to measure the similarity between the retrieved sentence and the aspect language model.
 - i.e., $S = -D(\theta_s || \theta_m) = \sum_w p(w|\theta_s) \log p(w|\theta_m) - \sum_w p(w|\theta_s) \log p(w|\theta_s)$
 - where θ_s, θ_m represents the language model of the sentence and the aspect respectively.
 - $p(w|\theta_s)$ can be computed using relative frequency of words in sentence s . The main challenge is thus how to compute $p(w|\theta_m)$

Language Modeling Approaches

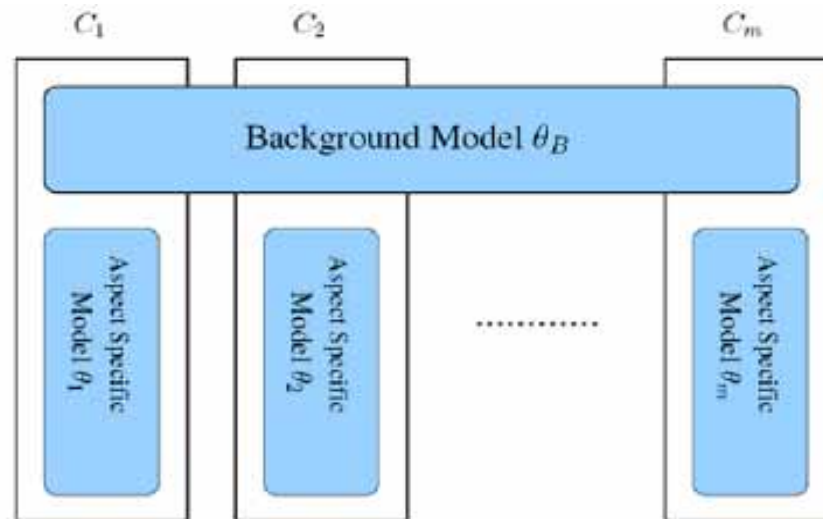
- **Baseline Language Model (baseLM)**
 - Based on the smoothed relative word frequency $p(w|\theta_i) = \lambda_B p(w|\theta_B) + (1 - \lambda_B) \frac{c(w, C_i)}{|C_i|}$ $p(w|\theta_B) = \frac{1+c(w, C)}{|C|+|V|}$
- **Advanced Aspect Models**
 - Explicitly distinguish a common background model from special topic models.
 - Distinguish between different topic models that characterize different information in different context.
 - Underlying basic idea: to treat the words as observations from a mixture model where the component models are the topic-specific word distributions and the background word distributions across different document collections.
 - Use EM to estimate the topic-specific models (i.e., the aspect-specific word distributions $p(w|\theta_m)$ in our task).

Advanced Aspect Models

- Discriminative Aspect Model (**DAM**)
 - To extract the discriminative aspect models by taking into consideration the hidden background model of the whole collection
- Generalizable Aspect Model (**GAM**)
 - To further factor out training-gene-specific language models embedded in the training sentences.

Discriminative Aspect Model

- Explicitly distinguishes common background model that characterizes common words across all aspects from special aspect models that characterize aspect-specific information



- A document d is regarded as a sample of the mixture model

$$p_d(w) = \lambda_B p(w|\theta_B) + (1 - \lambda_B) p(w|\theta_i)$$

Discriminative Aspect Model

- log-likelihood

$$\log p(\mathcal{C}) = \sum_{i=1}^m \sum_{d \in C_i} \sum_{w \in V} \{c(w, d) \log[\lambda_B p(w|\theta_B) + (1 - \lambda_B)p(w|\theta_i)]\}$$

- EM updating formulas

$$p(z_{d,w} = B) = \frac{\lambda_B p^{(n)}(w|\theta_B)}{\lambda_B p^{(n)}(w|\theta_B) + (1 - \lambda_B)p^{(n)}(w|\theta_i)}$$

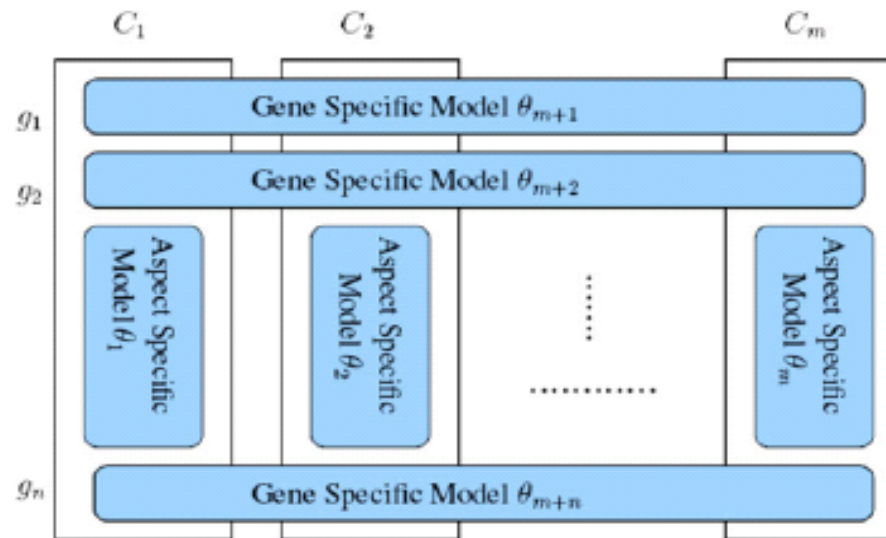
$$p(z_{d,w} = i) = \frac{(1 - \lambda_B)p^{(n)}(w|\theta_i)}{\lambda_B p^{(n)}(w|\theta_B) + (1 - \lambda_B)p^{(n)}(w|\theta_i)}$$

$$p^{(n+1)}(w|\theta_B) = \frac{\sum_{i=1}^m \sum_{d \in C_i} c(w, d)p(z_{d,w} = B)}{\sum_{w' \in V} \sum_{i=1}^m \sum_{d \in C_i} c(w', d)p(z_{d,w'} = B)}$$

$$p^{(n+1)}(w|\theta_i) = \frac{\sum_{d \in C_i} c(w, d)(1 - p(z_{d,w} = B))p(z_{d,w} = i)}{\sum_{w' \in V} \sum_{d \in C_i} c(w', d)(1 - p(z_{d,w'} = B))p(z_{d,w'} = i)}$$

Generalizable Aspect Model

- Each training sentence has two features: it is about a gene and it is associated with an aspect. => To further filter out the gene-specific word distributions from the aspect models.



- Document d is generated by generating each word as
$$p_d(w) = \sum_{i=1}^{m+n} p(v_i|d, C_d)p(w|\theta_i)$$
 - Choose a view v_i from a context according to the view distribution $p(v_i|d, C_d)$
 - Generate a word from the language model θ_i of the view v_i .

Generalizable Aspect Model

- log-likelihood

$$\log p(\mathcal{C}) = \sum_{d \in \mathcal{C}} \sum_{w \in V} c(w, d) \log \sum_{i=1}^{m+n} p(v_i | d, C_d) p(w | \theta_i)$$

- EM updating formulas

$$p(z_{w,i,d} = 1) = \frac{p^{(n)}(v_i | d, C_d) p^{(n)}(w | \theta_i)}{\sum_{i'=1}^{m+n} p^{(n)}(v_{i'} | d, C_d) p^{(n)}(w | \theta_{i'})}$$
$$p^{(n+1)}(v_i | d, C_d) = \frac{\sum_{w \in V} c(w, d) p(z_{w,i,d} = 1)}{\sum_{i'=1}^{m+n} \sum_{w \in V} c(w, d) p(z_{w,i',d} = 1)}$$
$$p^{(n+1)}(w | \theta_i) = \frac{\sum_{d \in \mathcal{C}} c(w, d) p(z_{w,i,d} = 1)}{\sum_{w' \in V} \sum_{d \in \mathcal{C}} c(w', d) p(z_{w',i,d} = 1)}$$

Evaluation

- Gold Standard
 - Two experts assign each candidate sentence to at most two most relevant aspects separately.
 - The sentences that are decided as relevant to a certain aspect are collected as the judgment for this gene in the corresponding aspect.
- Evaluation Metrics: ROUGE
 - ROUGE-1, ROUGE-2, ROUGE-3: models n-gram co-occurrence, $n=1, 2, 3$
 - ROUGE-W-1.2

Text Summary of Gene *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.
-