

USING TEXTPRESSO FOR INFORMATION RETRIEVAL, FACT EXTRACTION AND DATABASE ENTRY




Textpresso



**WormBase: a collaborative effort for genetics,
genomics and biology of *C. elegans*
and some related nematodes**

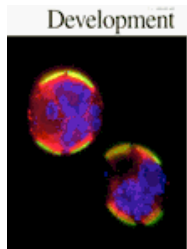
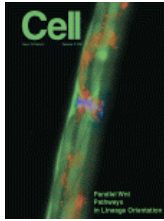


Literature
curation: 
Caltech,
Pasadena, CA

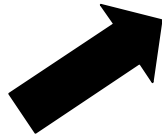


How does experimental data get into WormBase?

WormBase literature curation pipeline



A service of the [U.S. National Library of Medicine](#)
and the [National Institutes of Health](#)



Daily, automated search using keyword 'elegans'

Manually select papers for inclusion
into WormBase



Read all papers and 'flag'
various data types (first-pass)



email notification to individual data-type curators
paper-entity association
experiment curation



The screenshot shows the WormBase website interface. At the top, there is a navigation bar with links for Home, Genome, Synteny, Blast / Blast, WormMart, Batch Sequences, Markers, Genetic Maps, Submit, Searches, and Site Map. Below this is the 'WormBase Release WS200' banner and a search bar with the text 'Find: Any Gene' and a search button. There are also checkboxes for 'Exact match', 'Literature Search', and 'WormBase Suggest'. The main content area is titled 'News and Notes' and 'Recent Worm Community Forum Posts'. It features several news items, including 'October 08, 2008: Martin Chalfie wins Nobel Prize for GFP in C. elegans' and 'September 26, 2008: Award-winning C. elegans biologists'. At the bottom, there is a sidebar with links for 'About', 'Species', 'Community', 'Tools', 'Webiles', and 'Up-to-date'.



Why incorporate text mining into our curation pipeline?

WormBase curators curate >25 different types of data, but there are still data types we don't yet curate

For some data types we do curate, we have a backlog

older papers <2001

newer curation tasks

More genomes = more curation

If text mining tools can **improve curation efficiency, they'll help with all of this!**



Textpresso : Text Mining for Literature Curation

<http://www.textpresso.org>

Key features:

- **Search full text of articles, returns sentences**
- **Keyword (+synonyms) and/or category searches**
 - 'regulation': attenuate, downregulate, enhancing, inhibiting, misregulated, etc.**
- **Sort by score, year, journal, etc.**
 - sort by score = efficient prioritization**
- **Open source, with 19 different implementations**



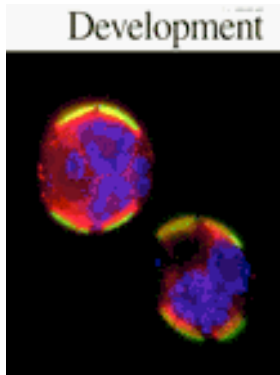
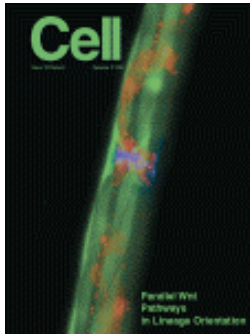
Textpresso Sentence Mark-Up Categorizes Terms

“...all wild-type embryos have foci of GFP:SAS-4
associated with both sperm centrioles...”

...all <phenotype_celegans> **wild-type** </phenotype_celegans>
<life_stages_celegans> **embryos** </life_stages_celegans> have
<assay_terms> **foci** </assay_terms> of <assay_terms> **GFP** </assay_terms>
: <protein_celegans> **SAS-4** </protein_celegans> <verbs> **associated** </verbs>
with both <anatomy_celegans> **sperm** </anatomy_celegans>
<cellular_components> **centrioles** </cellular_components>...



Textpresso for Paper-Entity Association: What objects are in a paper?



Gene names

Alleles

Transgenes

Life stages

Anatomy terms

Chemicals

Human disease



Paper-entity association: alleles and transgenes

Principle: standardized nomenclature allows for pattern matching in full text using regular expressions

Alleles: q71, lg6001, e2141ts, oz12oz75

[a-z]{1-3}\d+\w*\d*

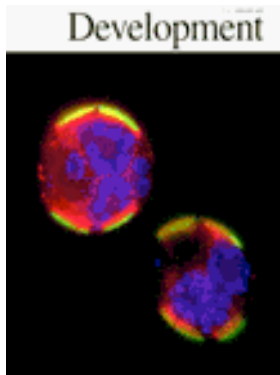
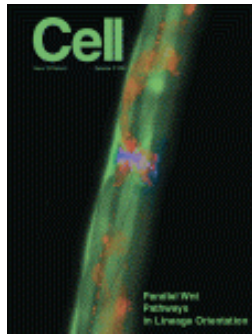
98% recall

Transgenes: syIs1, nIs2, bIn1

80% recall, >95% precision



Textpresso for Data Type Identification: What kinds of experiments?



Cloning

Mutant Phenotypes
RNAi
Alleles

Expression Patterns

Antibodies

Gene Regulation

Interactions

Genes

Gene Products

Site of Action

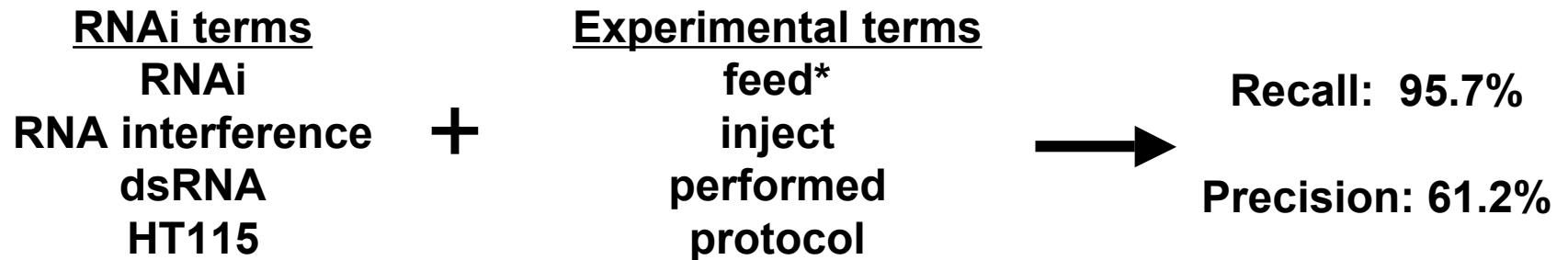


Data type identification: RNAi experiments

Principle: combinatorial category search to find evidence for experiments in full text of articles

“We performed RNAi feeding using lim-7_cDNA (see Section 2.2) on homozygous wild-type and heterozygous (tm674/hT2 [let GFP]) animals...”

“...we exposed ain-1 (ku322) animals to ain-2 (RNAi), starting with L1 larvae and using a published RNAi feeding construct and protocol...”



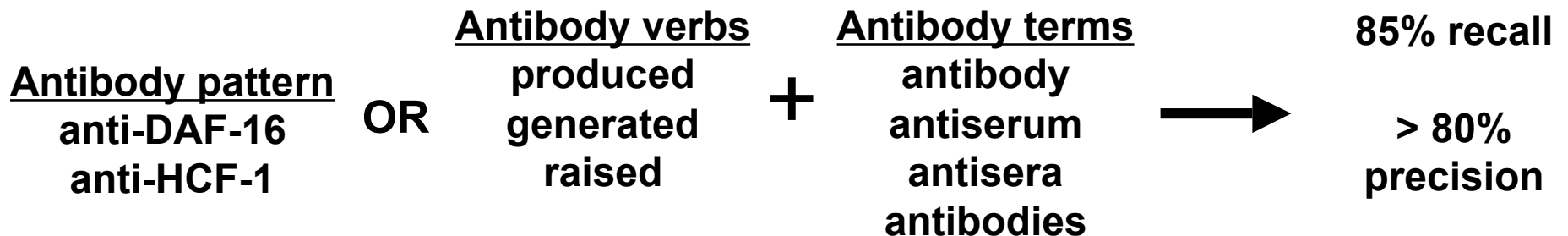


Data type identification: Antibodies and antibody production

Principle: pattern matching and combinatorial category search to find evidence for antibodies in full text

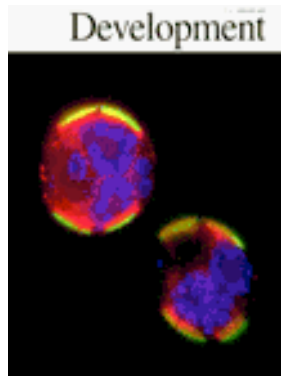
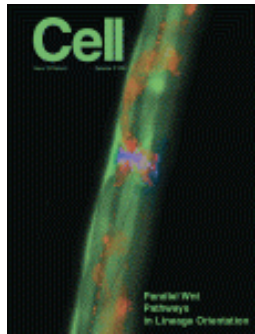
“The immunoprecipitated protein complexes were subsequently immunoblotted using anti-HCF-1, anti-DAF-16, or anti-GFP antibodies.”

“To determine the expression pattern of sepa-1, we raised antibodies against the KIX domain of SEPA-1.”





Textpresso for Fact Extraction: Finding Curatable Information



**loss of MAD1/MDF-1
suppresses the meiotic
defect of mat-3(or180)**

NXF-1 binds RNA directly

**GFP:: RAB-7 is visualized
on the membranes of late
endosomes**

**srsx-3 expression was
present in young odr-1
larvae but absent in odr-1
adults**



Fact Extraction: GO Cellular Component Curation

Principle: combinatorial category searches to retrieve experimental results from the full text of papers

Why GO Cellular Component curation?

-Subcellular localization is often expressed in a stereotypical way within a single sentence

GFP:: RAB-7 is visualized on the membranes of late endosomes

- Limited number of experimental strategies or assays
 - Microscopy**
 - Fractionation****
- Typically annotate using only one GO evidence code, IDA**



Fact Extraction: GO Cellular Component Curation

Needed to make new categories:

What words and phrases are diagnostic of subcellular localization experiments?

- 1 - Curators read ~240 papers containing localization data**
- 2 – Collected ~1,700 sentences that report experimental results**
- 3 – Examined word usage and frequency**

This approach took advantage of the work that curators already do and also allowed us to add annotations to WormBase while we created the new categories.



Fact Extraction: GO Cellular Component Curation

Three new categories:

Cellular Components

Verbs

Assay Terms

IDA-1 and IDA-1::GFP were not restricted to presynaptic sites but were abundantly localized throughout cell bodies, axons and dendrites of expressing neurons .



Fact Extraction: GO Cellular Component Curation

Can we annotate from the search results?

Test searches:

search with names of previously uncurated *C. elegans* proteins and three new categories

Criteria:

Returned sentences must contain a *C. elegans* protein plus one word from each of the three categories

Results:

Papers: 79.1% recall, 61.8% precision

Sentences: 30.3% recall, 80.1% precision

***Annotations: 66.2% recall, 97.3% precision**



Fact Extraction: GO Cellular Component Curation

Is Textpresso-based curation more efficient?

YES!

Tests comparing manual vs. Textpresso-based curation:

**Textpresso yields 8- to 15-fold improvement
in curation efficiency**



Incorporating Textpresso into our Curation Pipeline

Weekly searches:

Proteins	Category Term	Suggested GO terms based on previous annotation	Textpresso Sentences
WDR-23	nuclei	germ cell nucleus nucleus	<p>curate <input type="radio"/> already curated <input type="radio"/> scrambled sentence <input type="radio"/> false positive <input type="radio"/> not go curatable <input type="radio"/> Add To Go : <input type="checkbox"/> SentenceID 3 -- S 5 P WBPaper00032980 S s124 E As shown in Figure 5C , WDR-23 :: GFP is expressed in nuclei of the hypodermis , intestine and head neurons .</p>
SKN-1	nuclei	germ cell nucleus nucleus	<p>curate <input type="radio"/> already curated <input type="radio"/> scrambled sentence <input type="radio"/> false positive <input type="radio"/> not go curatable <input type="radio"/> Add To Go : <input type="checkbox"/> SentenceID 4 -- S 4 P WBPaper00032980 S s139 E Exposure to stressors or inhibition of the proteasome causes SKN-1 :: GFP to accumulate in the nuclei of intestinal cells (1 , 27) .</p>
SKN-1	nuclei	germ cell nucleus nucleus	<p>curate <input type="radio"/> already curated <input type="radio"/> scrambled sentence <input type="radio"/> false positive <input type="radio"/> not go curatable <input type="radio"/> Add To Go : <input type="checkbox"/> SentenceID 5 -- S 5 P WBPaper00032980 S s140 E As shown in Figures 6C-D , RNAi of wdr-23 , ddb-1 and cul4 caused accumulation of SKN-1 :: GFP in intestinal nuclei .</p>
SKN-1 WDR-23	nuclear	nucleus perinuclear region of cytoplasm	<p>curate <input type="radio"/> already curated <input type="radio"/> scrambled sentence <input type="radio"/> false positive <input type="radio"/> not go curatable <input type="radio"/> Add To Go : <input type="checkbox"/> SentenceID 6 -- S 8 P WBPaper00032980 S s148 E As shown in Figure 7A , skn-1 mRNA levels were not significantly (P>0 . 05) increased by knockdown of these genes , demonstrating that WDR-23 and the CUL-4/DDB-1 complex regulate nuclear accumulation of SKN-1 protein independently of skn-1 mRNA levels .</p>



Onward! Future Developments and Plans

Improve Search Results

False Positives

- Textpresso searches by paper section
- Word exclusion

False Negatives

- What do we miss and why?
 - Missing category terms
 - Non-standard nomenclature
 - Information in multiple sentences

More Category Development

- Use Textpresso searches to help make new categories
- Category Editor – interactive, iterative Textpresso searches

Incorporate Textpresso Searches into Existing Curation Tools

- Phenote for Phenotype and GO Curation



Acknowledgements

Textpresso – Caltech

Juancarlos Chan
Ruihua Fang
Joshua Jaffery*
Eimear Kenny*
Hans-Michael Müller
Arun Rangarajan
Tracy Teal*

WormBase – Caltech

Juancarlos Chan
Wen Chen
Jolene Fernandes
Ranjana Kishore
Raymond Lee
Cecilia Nakamura
Andrei Petcherski*

Slava Petcherski*
Gary Schindleman
Erich Schwarz
Kimberly Van Auken
Daniel Wang
Xiaodong Wang
Karen Yook

*alumni

Paul Sternberg, PI

**Funding: National Human Genome Research Institute, NIH
WormBase, Textpresso, Gene Ontology**

**Bioinformatics Group, Lewis Sigler Institute for Integrative Genomics,
Princeton University**