

Database Curation and Access for Bioinformatics

Lynette Hirschman
Alexander Yeh

781-271-7789 • lynette@mitre.org

MITRE Sponsored Research

The logo for the MITRE Technology Program, featuring a stylized graphic of stacked blocks in yellow, orange, and blue to the left of the text.

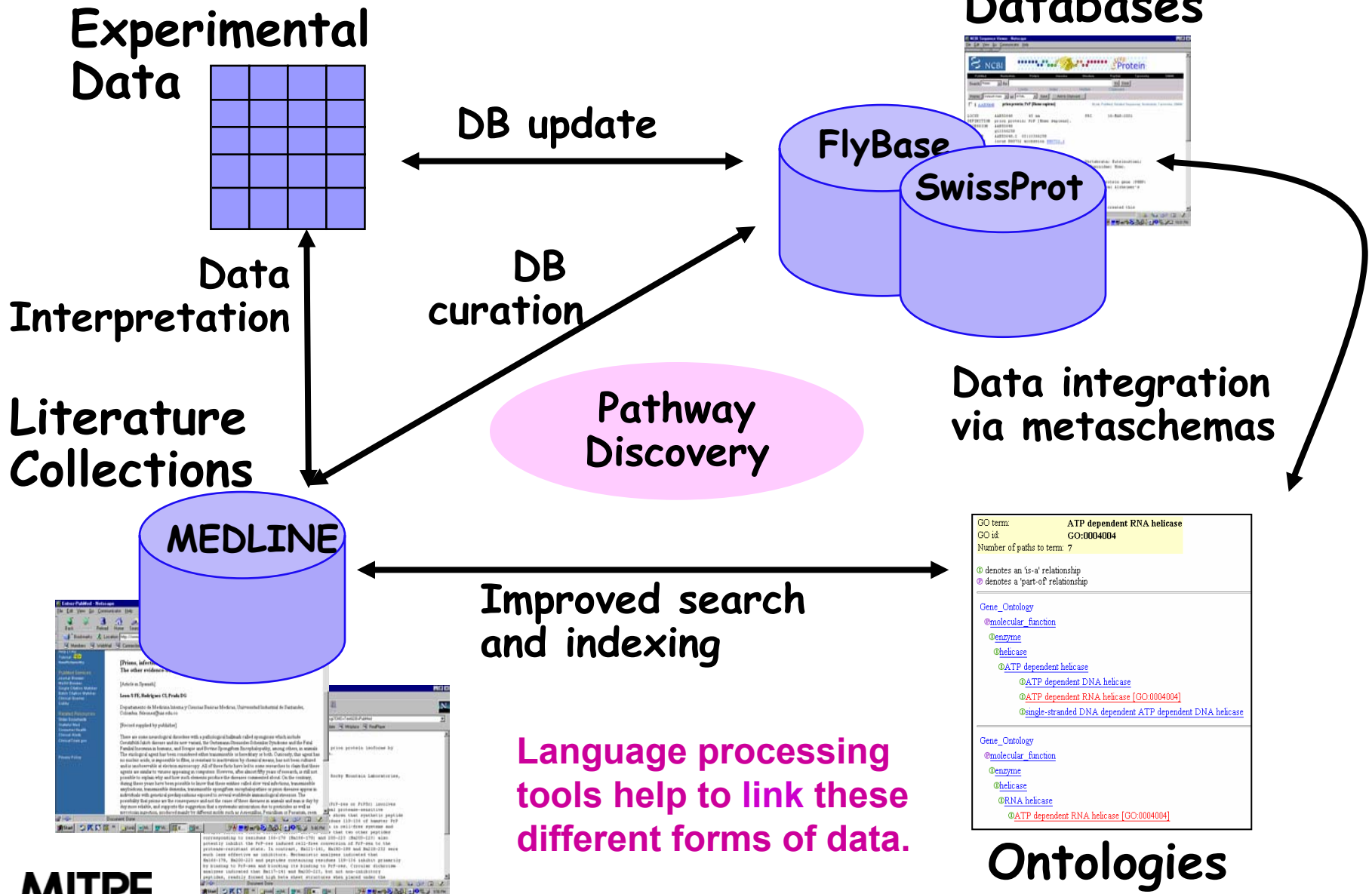
MITRE
Technology
Program

Problem

- **Biological information is exploding.**
 - **Estimates are around 1 terabyte/week!**
 - **New knowledge is constantly being discovered.**
- **Biology needs information technology to:**
 - **Maintain DB correctness, consistency and currency as new facts emerge and terminology evolves,**
 - **Access relevant information across many sources, and**
 - **Discover new relations from known information.**

Background

Biological data exists in unstructured (text) and structured (DBs) form.



Language processing tools help to link these different forms of data.

Ontologies

Objective

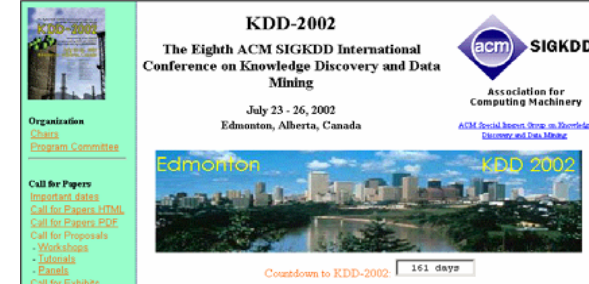
- **Develop interactive information extraction techniques for on-line biomedical literature**
 - **To support interactive, semi-automated curation* of biology databases**
 - **Focused on maintaining the currency and consistency of existing databases**
 - **In the face of exponential growth of research in genomics and proteomics**
- **Create “biology challenge problems” to promote progress in these areas**

* Curation refers to expert selection of entries for inclusion in database (e.g, gene name, function,...)

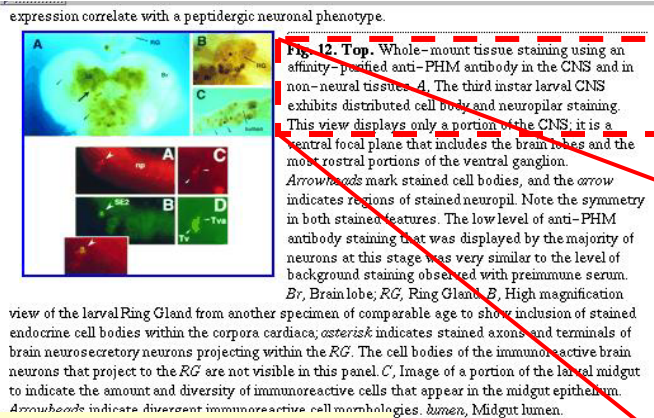
Activities

- Worked with FlyBase* to create KDD Cup 2002 challenge evaluation task
 - Evaluated 32 systems from 18 groups (8 countries) for ability to rank papers for curation based on presence of experimental evidence for gene products
- Building entity extraction for biological texts
 - Creating large amounts of training data for use with a trainable entity tagger
- 3 papers published and 1 submitted

*With many thanks to our FlyBase collaborators: William Gelbart, Beverly Matthews, Leyla Bayraktaroglu, David Emmert, Don Gilbert



Highlight: Sample DB Entry



Passage in text supporting that entry

Fig. 12. Top. Whole-mount tissue staining using an affinity-purified anti-PHM antibody in the CNS and in non-neural tissues. *A*, The third instar larval CNS exhibits distributed cell body and neuropilar staining. This view displays only a portion of the CNS; it is a

FlyBase: segment

Expression pattern	Publication	Stage	Tissue/Position
	Kolhekar et al., 1997	larva	embryonic/larval endocrine system
		larva	embryonic/larval digestive system
		larva	larval central nervous system
		larva	SE2 neuron
Expression info	Kolhekar et al., 1997 <i>Phm</i> protein is detected throughout all levels of the larval CNS as well as in other tissues, including the endocrine glands and the gut. Staining is observed in the cell bodies and in the neuropil of the brain. Staining is also prevalent in secretory cells of the ring gland, salivary gland, and in diverse cells in all levels of the midgut. In the CNS, several strongly staining cells were identified as neuroendocrine neurons. Many <i>Phm</i> shown to be peptidergic cells.		
Assay mode	Kolhekar et al., 1997	immunolocalization	
Antibodies generated	Kolhekar et al., 1997 polyclonal		

Challenge: can text mining extract this automatically?

Bio database entry: Immunolocalization

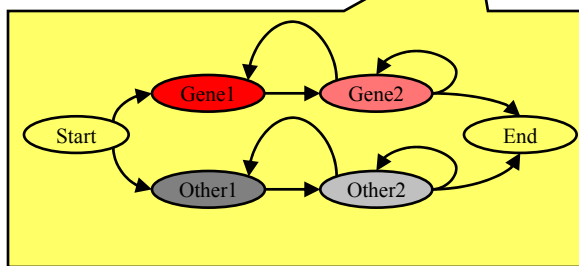
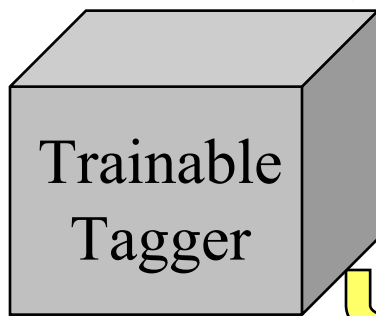
Highlight: Bio Name Tagger

Train a Genome Entity Tagger Using Cheap Noisy Training Data

Text automatically tagged using lexicon, filtered by FlyBase annotations



Noisy Training Data



MITRE Phrag HMM Tagger

Plain Text

dishabituation by a novel stimulus, attributable to pl brain. Mutations of rutabaga that diminish cAMP syn of habituation, whereas dunce mutations that increa: detectable but moderate increase in habituation rate habituation was extremely rapid in dunce rutabaga d corresponds to the extreme defects seen in double n learning tasks, and demonstrates that defects of the

dishabituation by a novel stimulus, attributable to pl brain. Mutations of **rutabaga** that diminish cAMP syn of habituation, whereas **dunce** mutations that increas detectable but moderate increase in habituation rate: habituation was extremely rapid in **dunce** rutabaga d corresponds to the extreme defects seen in double n learning tasks, and demonstrates that defects of the

Approx. Comparison of F-measure

Our System	.76
Krauthammer et al.	.75
Gaizauskas et al.	.83

Impacts

- **Biological entity extraction systems support automated curation and knowledge discovery from the biological literature.**
- **Challenge evaluations will drive progress and allow text mining to achieve accuracy comparable to that achieved for news.**
- **MITRE is being recognized as a leader in the application of text mining and natural language techniques to biology.**
 - **This will be a major investment area for IT, given its commercial importance and its importance related to bioterrorism.**

Future Plans



ISCBS INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY

HOME

Welcome!

Key Dates

REGISTRATION

Registration

SUBMISSIONS

Call For Papers (closed)

Submission Guidelines for Accepted Papers

Call For Tutorials (closed)

Call For Posters

Industry Demos

Non-profit Demos

Travel Fellowships

GENERAL INFO

About Brisbane

Brisbane Time

Housing

Transportation/Visas

Maps

Mailing List

FAQ

Contact Us

THANKS TO

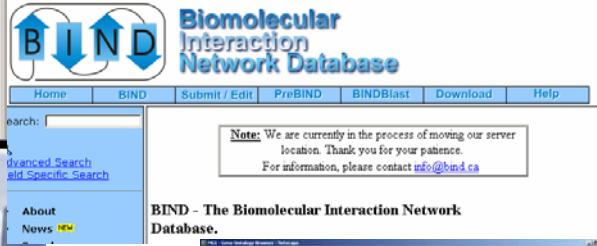
SIGs

Please check back in March 2003 for updated information.

Over the past 11 years a number of smaller, more specialized meetings have become regularly associated with the ISMB annual meetings. This year ISMB will feature several special interest group meetings associated with this year's conference.

Bioinformatics Open Source Conference (BOSC)		
Friday, June 27 09:00 - 21:00	Location TBD	Prices TBD
Saturday, June 28 09:00 - 21:00		
Biopathways		
Friday, June 27 09:00 - 21:00	Location TBD	Prices TBD
Saturday, June 28 09:00 - 21:00		
Text-Mining		
Friday, June 27 09:00 - 17:30	Location TBD	Prices TBD
Bio Ontologies		
Saturday, June 28 09:00 - 17:30	Location TBD	Prices TBD

Brisbane, Australia • June 29 - July 3, 2003



Second Meeting of the Special Interest Group on Text Data Mining

ISMB02 Satellite meeting on August 2, 2002. Westin Hotel

Aims

This meeting will have two purposes. The first is to bring together researchers developing text data mining tools and related language processing methods to manage the information explosion in the biomedical field. This part of the meeting will include invited and contributed papers, with a focus on developing shared infrastructure (tools, corpora, ontologies, see for example [Natural Language Processing in Biology](#)) and challenge evaluations, in the style of the [KDD Challenge Cups](#). It will include talks from two related SIGs, BioPathways and BioOntologies.

The second part of the meeting will discuss the establishment of a formal SIG for text data mining. We therefore strongly encourage interested people to attend, to participate in the definition and organization of this SIG.

Information

For further information contact [Lynette Hirschman](#) or [Christian Blaschke](#)

Text Mining SIG and Challenge Evaluation organized by MITRE and collaborators