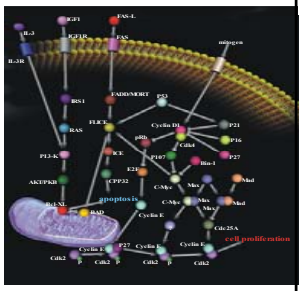


KDD Cup 2002 Task1: Information Extraction from Biomedical Articles

Co-Chair: Alexander Yeh, MITRE Corp.

Data: FlyBase (http://www.flybase.org)

July 2002

ITC

MITRE

Task Background

- Biomedical information exists in
  - Research literature
  - More than 280 semi-structured databases. Some examples:
    - FlyBase: fruit fly genes and proteins
    - Mouse Genome Database (MGB)
    - Protein Information Resource (PIR)
- Some databases act as distillations of subsets of the literature
- Currently, curators (people) read the literature to manually update (curate) the databases

ITC

MITRE

FlyBase: Example of Data Curation

expression correlate with a peptidergic neuronal phenotype.

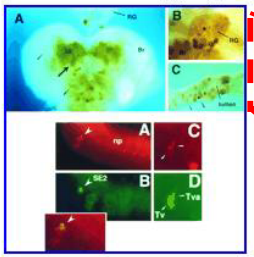


Fig. 12. Top. Whole-mount tissue staining using an affinity-purified anti-PHM antibody in the CNS and in non-neural tissues. *A*, The third instar larval CNS exhibits distributed cell body and neuropilar staining. This view displays only a portion of the CNS; it is a ventral focal plane that includes the brain lobes and the most rostral portions of the ventral ganglion. *Arrowheads* mark stained cell bodies, and the *arrow* indicates regions of stained neuropil. Note the symmetry in both stained features. The low level of anti-PHM antibody staining that was displayed by the majority of neurons at this stage was very similar to the level of background staining observed with preimmune serum. *Br*, Brain lobe; *RG*, Ring Gland. *B*, High magnification view of the larval Ring Gland from another specimen of comparable age to show inclusion of stained endocrine cell bodies within the corpora cardiaca; *asterisk* indicates stained axons and terminals of brain neurosecretory neurons projecting within the *RG*. The cell bodies of the immunoreactive brain neurons that project to the *RG* are not visible in this panel. *C*, Image of a portion of the larval midgut to indicate the amount and diversity of immunoreactive cells that appear in the midgut epithelium. *Arrowheads* indicate divergent immunoreactive cell morphologies. *lumen*, Midgut lumen.

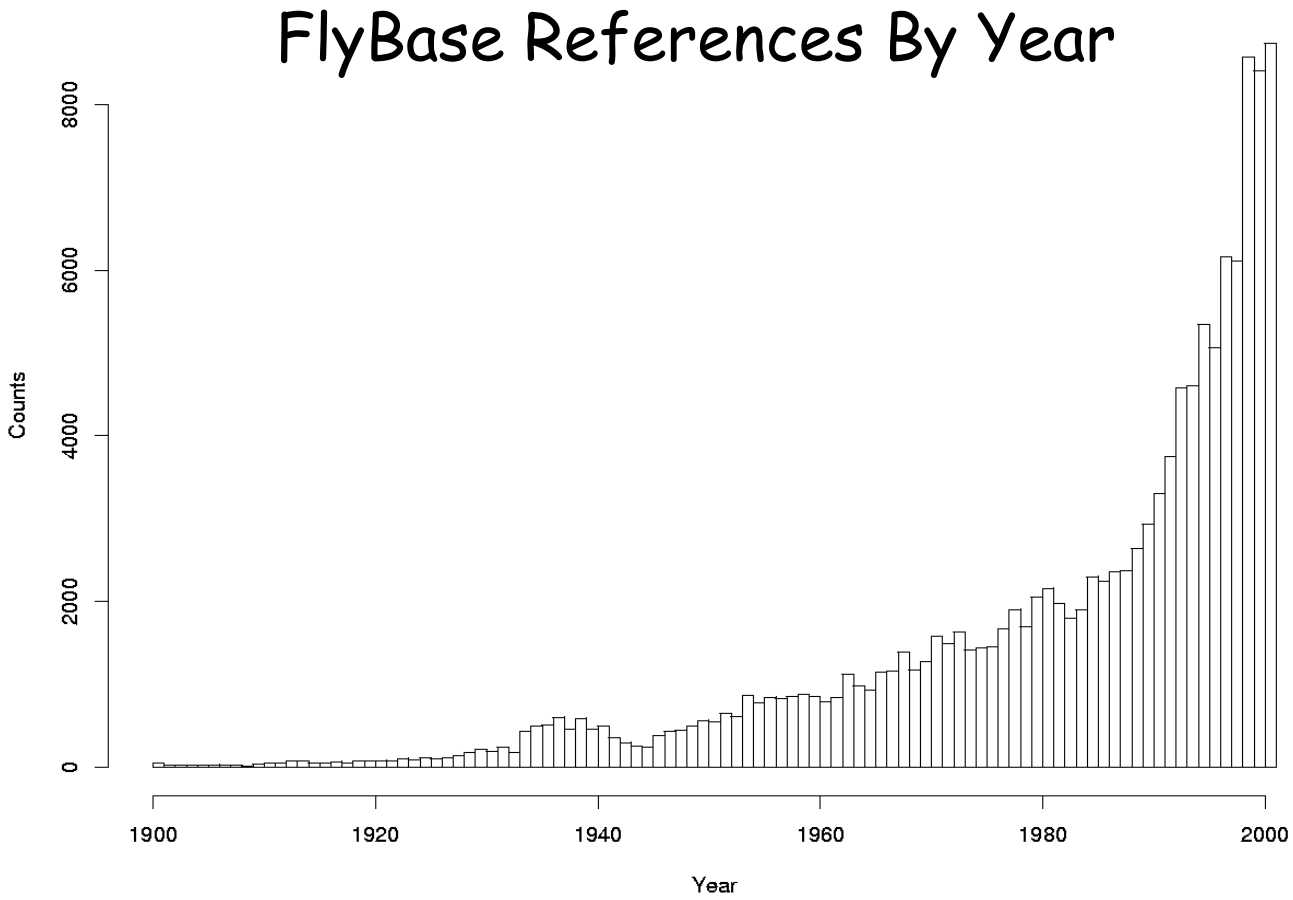
Fig. 12. Top. Whole-mount tissue staining using an affinity-purified anti-PHM antibody in the CNS and in non-neural tissues. *A*, The third instar larval CNS exhibits distributed cell body and neuropilar staining. This view displays only a portion of the CNS; it is a

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* Phm 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 Phm-positive neurons were shown to be peptidergic cells.		
Assay mode	*Kolhekar et al., 1997*	immunolocalization	
Antibodies generated	*Kolhekar et al., 1997* polyclonal		

ITC

MITRE

Curators Cannot Keep Up with the Literature!



Task Rationale and Description

Task Rationale and Description

- We want to see what can be done to help automate this curation process

Task Rationale and Description

- We want to see what can be done to help automate this curation process
- Start fairly simple. Try to help automate part of what one group of curators needs to do:
  - Determine which papers need to be curated for fruit fly gene expression information
    - "Curate" means read in detail to make entries in the database
  - Want to curate those papers containing **experimental results** on gene products (RNA transcripts and proteins)

ITCA

MITRE

Task Description: For a Set of Papers on Genetics or Molecular Biology

- Given for each paper
  - The full text of that paper
  - A list of the genes mentioned in that paper
- Determine for each paper
  - Does that paper contain any curatable gene product information (experimental results)?
  - For each gene mentioned in the paper, does that paper have experimental results for
    - Transcript(s) of that gene?
    - Protein(s) of that gene?
- Also produce a ranked list of the papers
  - Rank the curatable papers before the non-curatable papers

MITRE

Task is Harder Than It First Appears

- Interested in results applicable to "regular" (found in the wild) flies, not mutants
- Genes have multiple names (synonyms)
  - Given a list of the known synonyms
  - But list may be incomplete
  - Some names can refer to more than one gene
    - E.g., "Clk" is a symbol for one gene (**Clock**) and is also a synonym for another gene (**period**, symbol is "per")
- Contestants given evidence of experimental results found in the training data,
  - But only in the form that is recorded in the FlyBase database

ITG

MITRE

Evidence of Results Found in Training Data as Recorded in FlyBase

- Database (DB) records what evidence is found in a training paper, but not where in that paper
- The evidence is often recorded in a "normalized" form and domain knowledge is needed to find the corresponding text, e.g.,
  - DB: Assay mode: "immunolocalization"
    Text (PubMed ID#9006979):
    "*Figure 12.* …Whole-mount tissue staining using an affinity-purified anti-PHM antibody in the CNS … This view displays only a portion of the CNS"
    - Term "immunolocalization" is not in the text
    - Instead, text describes an instance of performing immunolocalization

ITC

MITRE

Task Details

- Task has 3 sub-tasks, that contribute equally to the overall score
  - 1. Ranked-list of papers (curatable before non-curatable)
    - Look at area under ROC curve

      ROC = Receiver Operating Characteristic

      Prob(detection) versus Prob(false alarm)
  - 2. Yes/No decisions on the papers being curatable (having any results of interest)
    - Use balanced F-score
  - 3. Yes/No decisions for having results for each type of product (transcript, protein) for each gene mentioned
    - Use balanced F-score

ITCX MITRE

Result Statistics

- 18 teams submitted 32 entries
- Entries from 7 "countries":
  - Japan, Taiwan, Singapore, India, UK, Portugal, USA
- Team type (lead member, some teams had multiple types):
  - Company: 7
  - University: 8
  - Other: 3

WINNER

- Combined team from ClearForest and Celera
  - Contacts: Yizhar Regev, Michal Finkelstein
  - Also had the best score in each of the 3 sub-tasks

	Best	Median
Ranked-list:	84%	69%
Yes/No curate paper:	78%	58%
Yes/No gene products:	67%	35%

  - The top entries for the ranked-list sub-task all had close scores for this sub-task

ITC

MITRE

Honorable Mentions (East to West):

- Combined team from the Design Technology Institute Ltd., the Mechanical Engineering Dept. at the National University of Singapore, and the Genome Institute of Singapore
  - Contact: Shi Min
- Combined team from the data mining group at Imperial College (UK) and Inforsense Limited
  - Contacts: Huma Lodhi, Yong Zhang
- Combined team from Verity, Inc. and Exelixis, Inc.
  - Contact: Bin Chen

Acknowledgements

- People at FlyBase, especially
  - William Gelbart
  - Beverly Matthews
  - Leyla Bayraktaroglu
  - David Emmert
  - Don Gilbert
- Project team at MITRE
  - Alexander Morgan
  - Lynette Hirschman

ITCR

MITRE