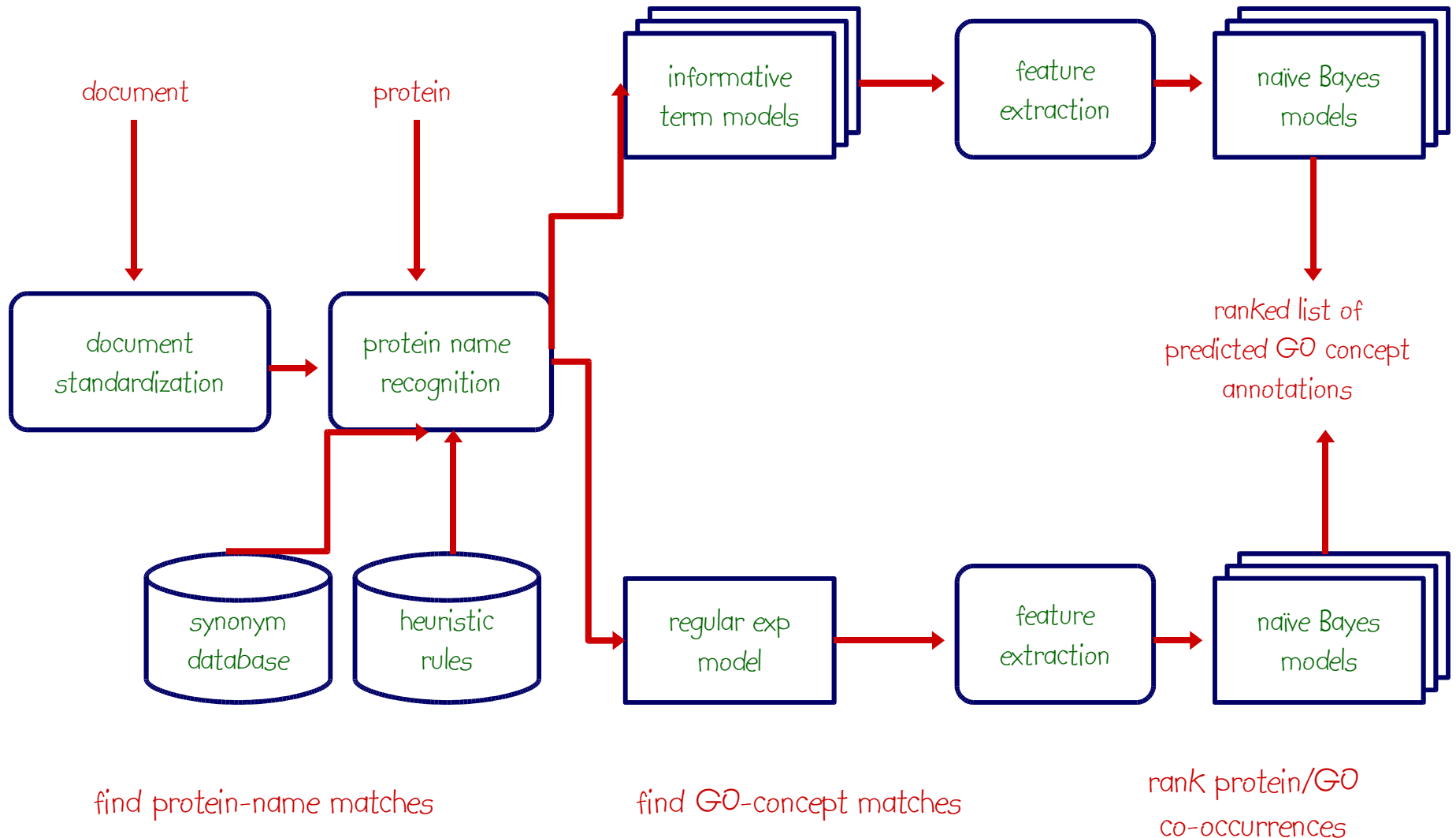


# A Simple Statistical Learning Approach for BioCreative Task 2

**Mark Craven and Soumya Ray**

Department of Biostatistics & Medical Informatics  
Department of Computer Sciences  
University of Wisconsin.  
craven@biostat.wisc.edu  
[www.biostat.wisc.edu/~craven](http://www.biostat.wisc.edu/~craven)

# System Overview



# Key Issues

1. How to recognize “occurrences” of GO concepts in text?
3. How to train models for thousands of GO concepts?  
Where do we get training data?
5. How to recognize occurrences of protein names in text?
7. How do we decide if protein occurrence is related to GO term occurrence?

# Recognizing Protein Name Occurrences

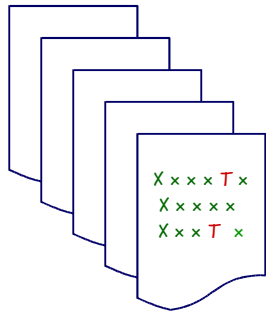
1. construct a regular expression for each protein
  - incorporates aliases from SwissProt, HuGO
  - allows small variations in punctuation, special characters
2. if no matches found, use heuristics to generate generalized aliases
  - drop words like *fragment* from end of given name
  - drop single-character tokens from end of given name
  - etc.

# Recognizing GO Concept Occurrences

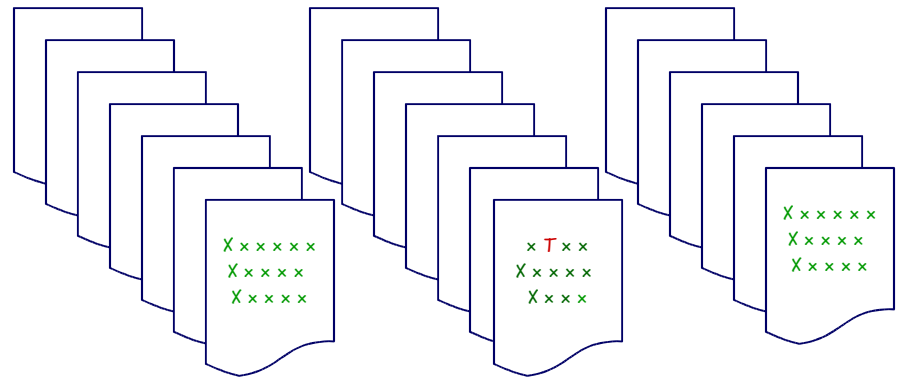
1. assemble a set of  $n$ -grams statistically associated with each GO code
  - since training data is extremely sparse, also use abstracts referenced by GO annotations in SGD, MGI, RGD and TAIR
2. construct regular expressions for matching other GO concepts

# Recognizing GO Concept Occurrences

documents associated  
with GO concept C



documents NOT associated  
with GO concept C



|                                 |                                 |
|---------------------------------|---------------------------------|
| # occurrences<br>of term T      | # occurrences<br>of term T      |
| # occurrences<br>of other terms | # occurrences<br>of other terms |

compute  $\chi^2$  value indicating association between T and C

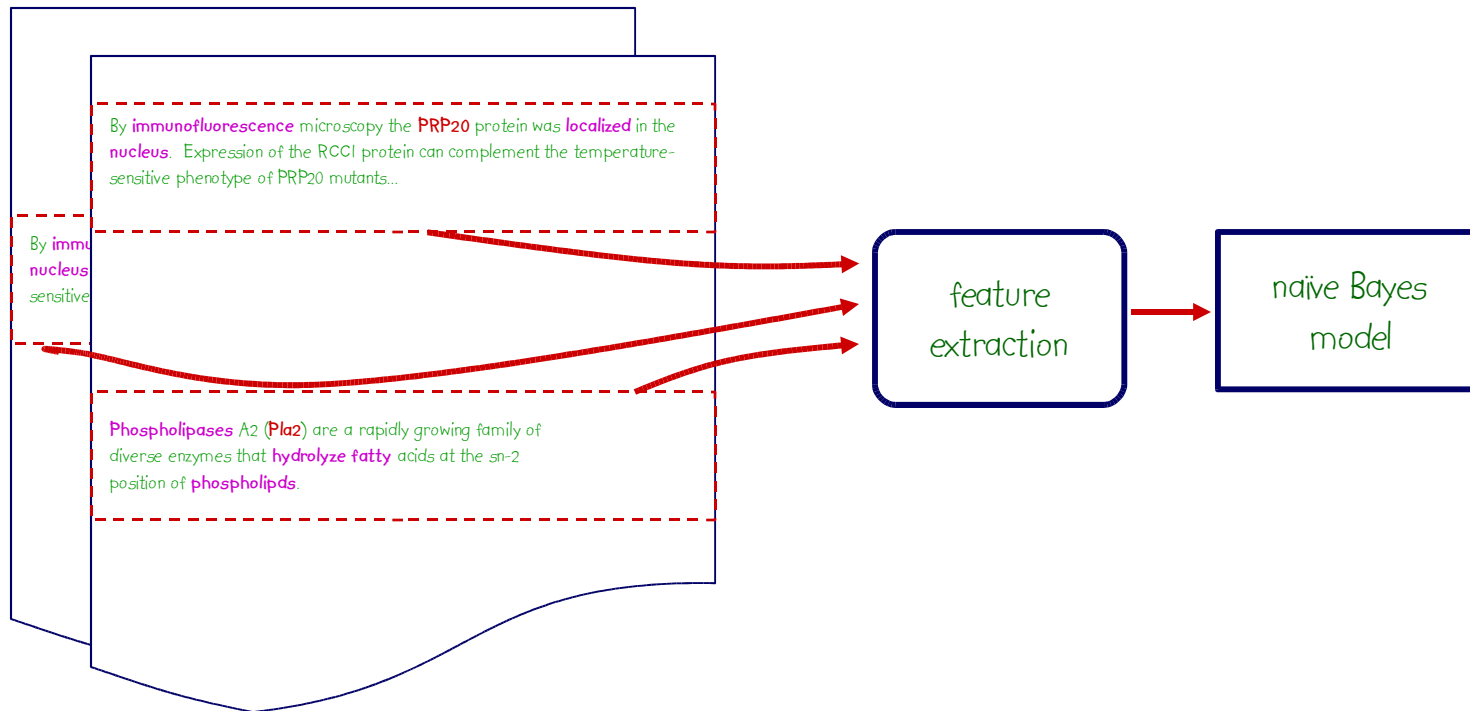
# Recognizing GO Concept Occurrences

- after scoring  $n$ -grams by  $\chi^2$  test, retain those that have significant association with GO concept  
e.g. unigrams associated with *sodium symporter activity*
  - pantothenate
  - biotin
  - transporter
  - lipoate
  - smvt
  - uptake
  - sodium-dependent
- documents linked to GO concept **C** are also linked to **C**'s ancestors (but with reduced weight)

# Linking Protein Names and GO Concepts

**Given:** passages of text with protein-name and GO-concept occurrences

**Do:** filter and rank these putative protein/GO annotations



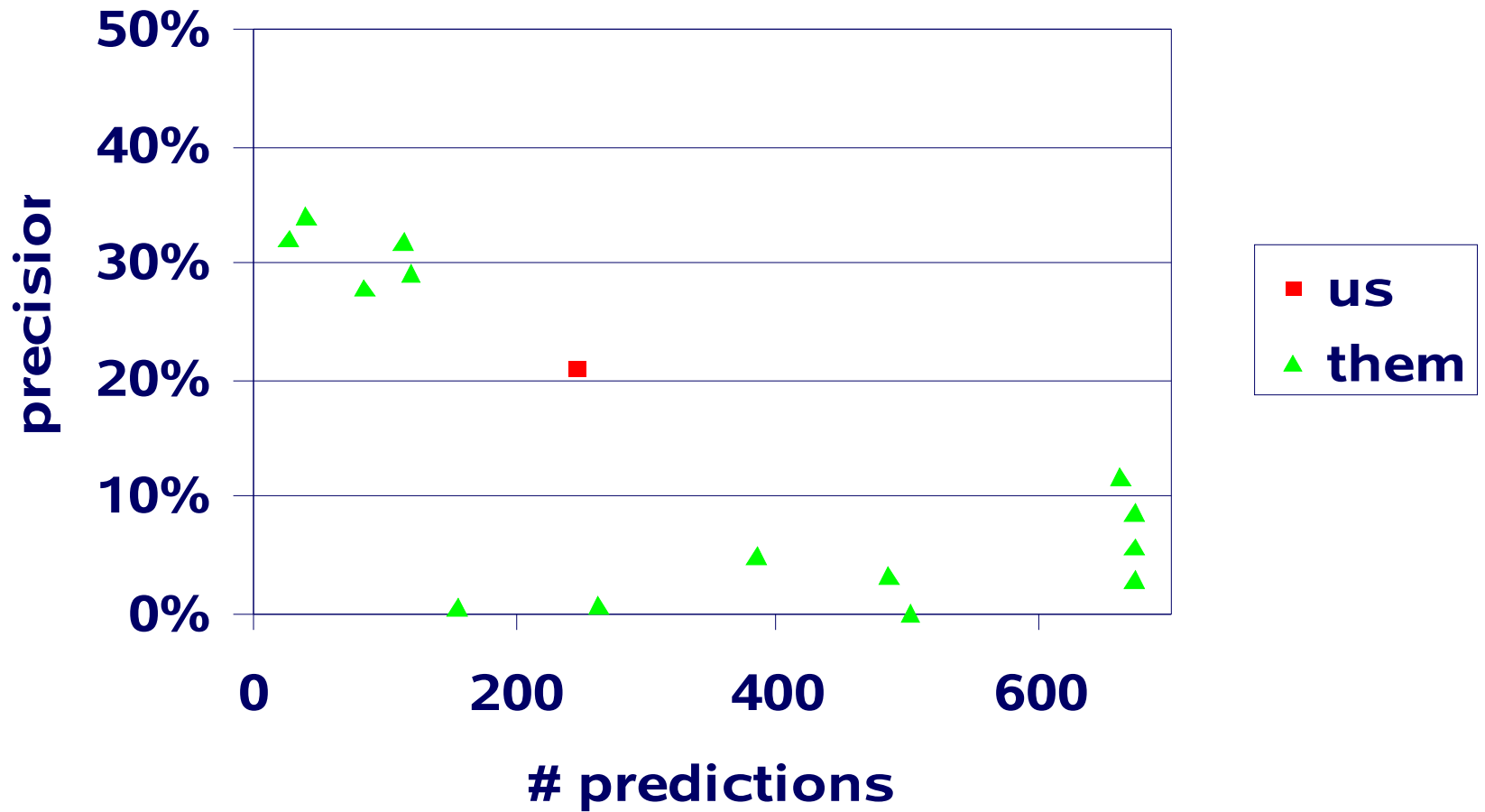
# Linking Protein Names and GO Concepts

- our naïve Bayes models use
  - bag-of-words representation of text passage
  - numeric features representing: score of GO match, avg. distance between protein-GO term pairs, etc.

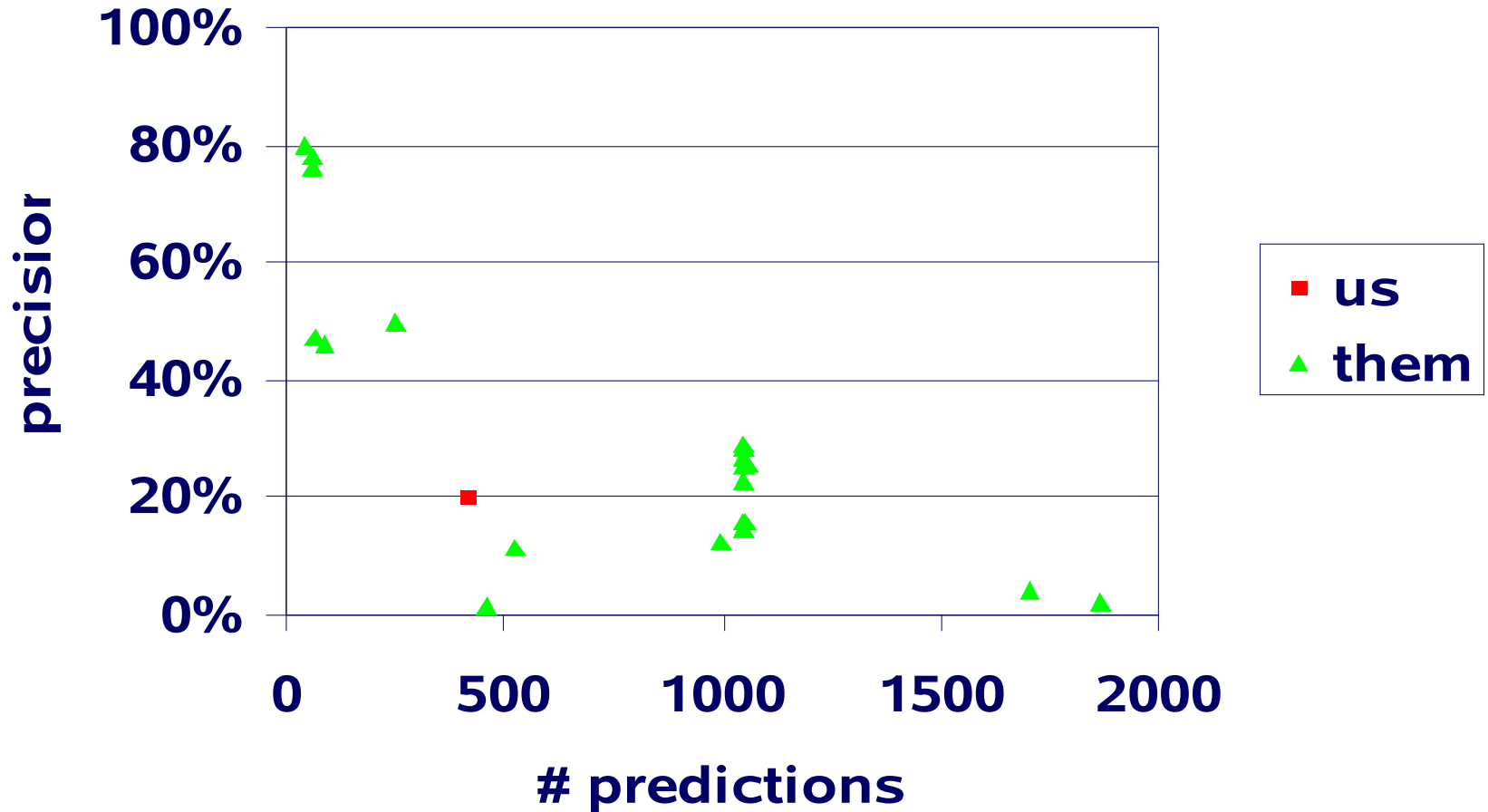
$$\Pr(\text{data} \mid \text{class}) = \prod_{i \in \text{vocab}(D)} \Pr(w_i \mid \text{class})^{n_i} \prod_{j \in \text{numeric}(D)} \left[ \frac{1}{\sqrt{2\pi}} e^{-\left(\frac{x_j - \mu_j}{\sigma_j}\right)^2} \right]$$

- *evidence text* bounded by paragraph w/highest-ranking score for a given protein-GO pair

# 2.2 Test Set Results



# 2.1 Test Set Results



# Discussion

- we also tried *logistic regression* and *multiple instance learning* – no consistent gain
- I wish we wouldn't have
  - predicted obsolete GO terms
  - predicted GO terms that were too high level
  - predicted plant and bacterial GO terms
  - processed *Materials & Methods* sections
  - processed the first paragraph
  - highlighted so much text
- but lots of technical things we would like to try too

# Acknowledgments

NSF CAREER grant IIS-0093016

NIH/NLM grant 1R01 LM07050-01