A Statistical Approach to Scanning the Biomedical Literature for Pharmacogenetics Knowledge.

DL Rubin, CF Thorn, TE Klein, RB Altman.
JAMIA Vol 12, No 2, pp121-129.

Patrick Herron

INLS 279

19 April 2005

A Statistical Approach to Scanning the Biomedical Literature for Pharmacogenetics Knowledge

- What's the problem?
- Genetic basis of drug response
- Predict individual drug responses
- What genes produce or alter a drug effect?
- How do we capture gene-drug relationships?

The stakes

- Big problem of identifying the right candidate drug target for a specific disease
- Currently 95% of candidates fail to produce a drug – even smaller percentage of targets
- Sequencing & analysis has failed b/c it has generated too much information, w/decrease in signal-to-noise ratio
- Failure usually due to toxicity or inefficacy
- “Quantal step” needed in discovery

Roses AD, et al. Disease-specific target selection: a critical first step down the right road. *Drug Discovery Today*. Vol 10, No 3, February 2005, 176-189.

Can Rubin *et al* help us?

- Can the system the authors propose overcome the information explosion by helping to identify efficacious (& nontoxic) drugs?
- Can we use the literature to perform in silico validation?
- Can their system increase the signal?

Gene-target-disease specificity

- The drug-gene relationship is really better thought of as a triune relationship between a target molecule, its associated/potential disease impacts, and genes related to the target and/or the disease
- Best relationships for discovery are highly specific
- Genome-level data is highly specific, but highly noisy

Narrowing the relevant literature

- How do we identify Medline citations that contain data about SPECIFIC drug-gene relationships?
- No comprehensive knowledge base that contains all drug-gene relationships data exists
- Manual task of identifying literature/db support for gene-drug relationships too time consuming

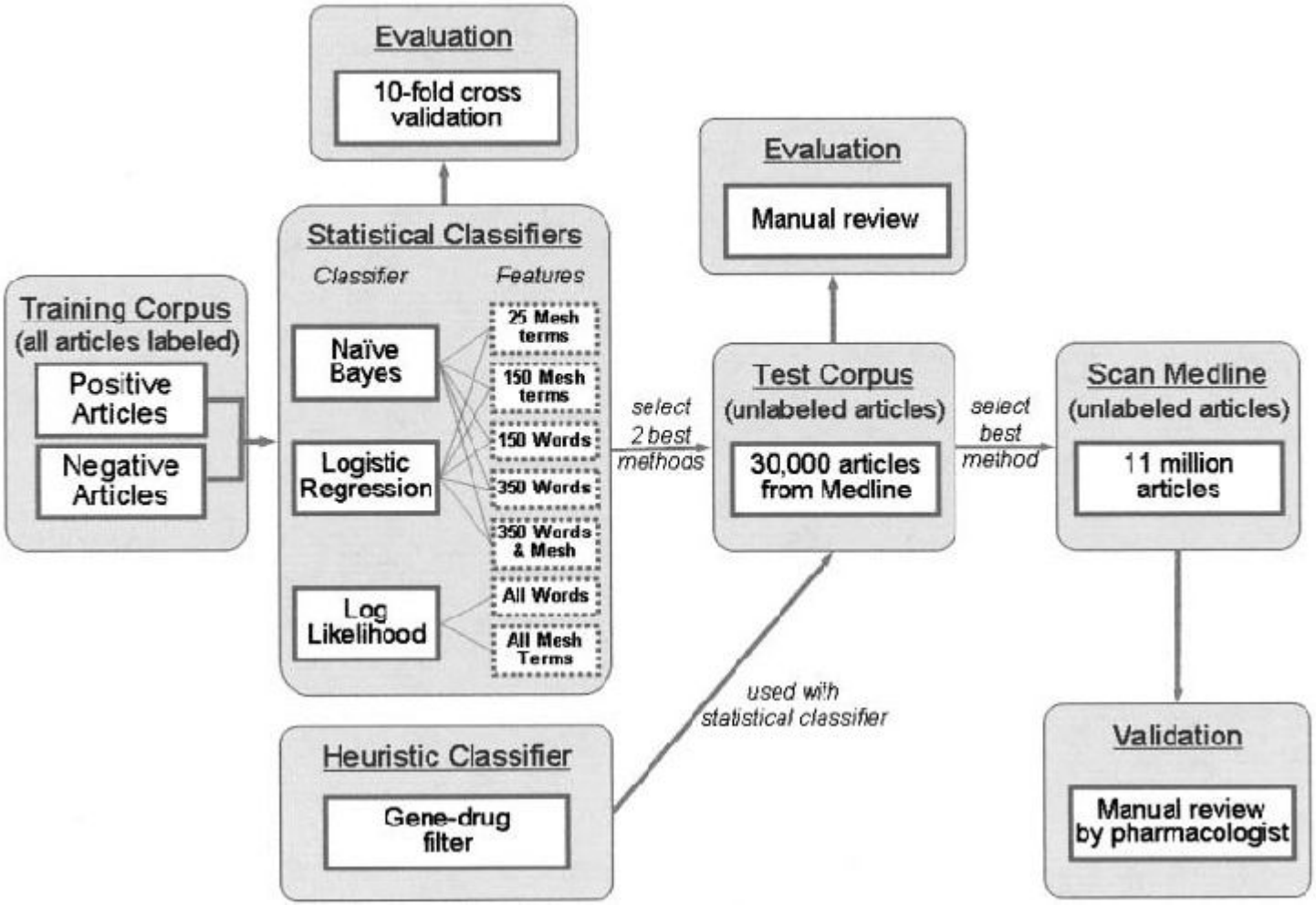
Method

- Pharmacogenetics corpus – manually selected drug-gene articles (standards?)
- Article Preprocessing
- Features describing Pharmacogenetics articles
- Classification methods
- Scanning Medline
- Manual validation

Factors

- Classification methods
  - Naïve Bayes
  - Regression
  - Log likelihood
- Feature representations
  - 25 best MeSH terms
  - 150 best MeSH terms
  - All MeSH terms
  - All MeSH terms with filtering
  - 150 best words
  - 350 best words
  - All words

Experimental Flow



Results

- Model performance – precision, recall, F measure
- MeSH terms generally showed higher precision
- Words yielded better recall
- Log likelihood on all MeSH terms performed best overall (by F measure)

Discussion

- MeSH terms show high precision and low recall—better precision than words alone
- What do you think is the heuristic drug-gene filter they're talking about?

Questions

- Is their method biased against ML approaches? Too few features? Training set too small?
- How much is a literature search going to get us?
- Do Rubin et al understand that a drug is embodied in the literature as a *target/target class to specific disease* pairing?
- Are we getting better information or just getting more information?
- How specific is the information identified by the system described in Rubin *et al*?
- Is it strength of association (figure 3) or just merely frequently written about (re: fashionable)? Authors claim that “as the number of articles containing a particular co-occurrence increases, a true association becomes more likely” (128)