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

Summary of presentation. I attempted to present the Rubin et al paper in light of a recent paper from Allen Roses that appeared in *Drug Discovery Today* (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). The Rubin paper focuses on the authors' own machine-learning based approach to automatic identification of research papers that contain some textual information about drug-gene relationships.

The Roses *et al* paper suggests that perhaps *the* big problem of pharmacogenetics is how to optimize the identification of the *right* candidate drug target for a specific disease. It suggests that the explosion of human genomics data has not resulted in an increase of useful information but has rather produced more noise, thus implying the need for improvements in sifting through large collections of data for useful information. According to Roses et al, currently 95% of *candidates* fail to produce a drug, and that percentage is much higher targets identified at the start of the investigative pipeline; the failures are usually due to toxicity or inefficacy. A "quantal step" is needed in discovery, particularly a discovery approach that identifies highly specific relationships between, genes, drugs, and targets, wading through the noisy mass of genomic data.

In light of Roses et al, can the system Rubin et al propose overcome the information explosion by helping to identify efficacious (& nontoxic) drugs? Rubin et al envision their approach as part of a knowledge base development program: how do we populate a knowledge base with gene-drug relationships buried in the millions of MEDLINE documents? Rubin et al use a standard text mining approach (http://tinyurl.com/altpb) starting with a manually selected pharmacogenetics corpus; identification of drug-gene papers is performed using automatic classification techniques. Part of the paper focuses on comparing, in a black-box fashion, Naïve Bayes with regression analysis and log likelihood, as well as comparing different feature representations (MeSH terms, words, combination). The authors show that while recall is higher for word-based representations, precision is better for MeSH term-based ones (this should not be unsurprising, since terms are more informative than words, and MeSH terms are more carefully selected in assignment than words are selected in the process of writing—words are inherently noisier).

It was not altogether from the paper whether Rubin et al were careful in decomposing "drug" to the relationship between a *target/target class to specific disease* pairings. Decomposing "drug" as such would necessarily improve both precision and recall, especially if precision and recall are better defined as being limited to highly specific relationships. (For example, articles on drugs in general and genes in general would be useless for drug discovery.) Further, it is not altogether clear how the authors are capturing the crucial properties of *novelty* and *interestingness* in their system.

Discussion. I find it a highly dubious claim that "as the number of articles containing a particular co-occurrence increases, a true association becomes more likely" (128) yet Brad Hemminger defended the claim, stating it is a commonly accepted belief. It remains unclear what is better for drug discovery: a system that produces high precision and low recall, or vice-versa. My opinion is that, particularly in light not only of Rubin's intent to build a knowledge base but also Roses' desire to radically increase the *precision* of drug target identification at earlier stages, that a high precision, low recall system is preferable to a low precision, high recall system, yet Dr. Hemminger disagrees. Discussion of the last half hour focused on attempting to fill in the blanks of precision and recall values curiously omitted by the authors in their evaluation of their own system.