Predicting Gene Function From Patterns of Annotation

by Oliver D. King, Rebecca E. Foulger, Selina S. Dwight, James V. White, and Frederick P. Roth
Genome Research 13(5), 896-904, May 2003
http://www.genome.org/cgi/doi/10.1101/gr.440803

Presented by Christopher Maier
INLS 279: Bioinformatics Research Review
2006-02-22

Annotating with Gene Ontology

- Genes in bioinformatics databases are annotated with entries from GO according to their function or composition

- Annotations are incomplete for several reasons
  - We don’t yet know everything about all genes
  - Curators have not processed the latest literature

- Is it possible to predict the GO annotations a gene should have?

Previous Prediction Approaches

- Natural Language Processing (NLP) to extract gene-attribute information from article abstracts, full text

- Correlations based on microarray hybridization data

- Correlations based on protein domain databases

- All these approaches are well suited for previously un-annotated genes; can existing annotations somehow be utilized in the prediction task?

Current Approach

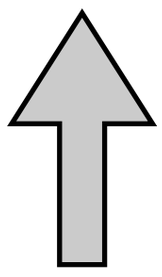
- Helpful analogy: customer preference prediction

    - However, we also have the ontological structure for the annotation prediction task

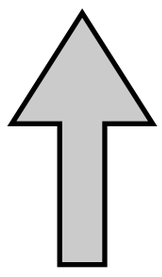
Customer Preference	Annotation Prediction
Customer	Gene
Product	GO Term
Purchase	Annotation

The Data

SGD	S000007287	15S_RRNA	GO:0005763	SGD_REF:S000073642|PMID:6261980	ISS		C	15S_rRNA|15S_RRNA_2	gene	taxon:4932	20040202	SGD
SGD	S000007287	15S_RRNA	GO:0003735	SGD_REF:S000073642|PMID:6261980	ISS		F	15S_rRNA|15S_RRNA_2	gene	taxon:4932	20030723	SGD
SGD	S000007287	15S_RRNA	GO:0006412	SGD_REF:S000073643|PMID:6280192	IGI		P	15S_rRNA|15S_RRNA_2	gene	taxon:4932	20030723	SGD
SGD	S000007287	15S_RRNA	GO:0042255	SGD_REF:S000051605|PMID:2167435	IGI		P	15S_rRNA|15S_RRNA_2	gene	taxon:4932	20030723	SGD
SGD	S000007288	21S_RRNA	GO:0005762	SGD_REF:S000073372|PMID:6759872	IDA		C	21S_rRNA|21S_rRNA_3|21S_rRNA_4	gene	taxon:4932	20040202	SGD
SGD	S000007288	21S_RRNA	GO:0003735	SGD_REF:S000073372|PMID:6759872	IMP		F	21S_rRNA|21S_rRNA_3|21S_rRNA_4	gene	taxon:4932	20030721	SGD
SGD	S000007288	21S_RRNA	GO:0003735	SGD_REF:S000073372|PMID:6759872	ISS		F	21S_rRNA|21S_rRNA_3|21S_rRNA_4	gene	taxon:4932	20030721	SGD
SGD	S000007288	21S_RRNA	GO:0006412	SGD_REF:S000073372|PMID:6759872	IMP		P	21S_rRNA|21S_rRNA_3|21S_rRNA_4	gene	taxon:4932	20030721	SGD
SGD	S000007288	21S_RRNA	GO:0006412	SGD_REF:S000073372|PMID:6759872	ISS		P	21S_rRNA|21S_rRNA_3|21S_rRNA_4	gene	taxon:4932	20030721	SGD
SGD	S000007288	21S_RRNA	GO:0042255	SGD_REF:S000073372|PMID:6759872	IMP		P	21S_rRNA|21S_rRNA_3|21S_rRNA_4	gene	taxon:4932	20030721	SGD
SGD	S000004660	AAC1	GO:0005743	SGD_REF:S000050955|PMID:2167309	TAS		C	ADP/ATP translocator	YMR056C gene	taxon:4932	20010118	SGD
SGD	S000004660	AAC1	GO:0005471	SGD_REF:S000050955|PMID:2167309	IDA		F	ADP/ATP translocator	YMR056C gene	taxon:4932	20010213	SGD
SGD	S000004660	AAC1	GO:0006839	SGD_REF:S000050955|PMID:2167309	IGI	SGD:S000000126	P	ADP/ATP translocator	YMR056C gene	taxon:4932	20040226	SGD
SGD	S000004660	AAC1	GO:0009060	SGD_REF:S000050955|PMID:2167309	IGI	SGD:S000000126	P	ADP/ATP translocator	YMR056C gene	taxon:4932	20040226	SGD



Gene



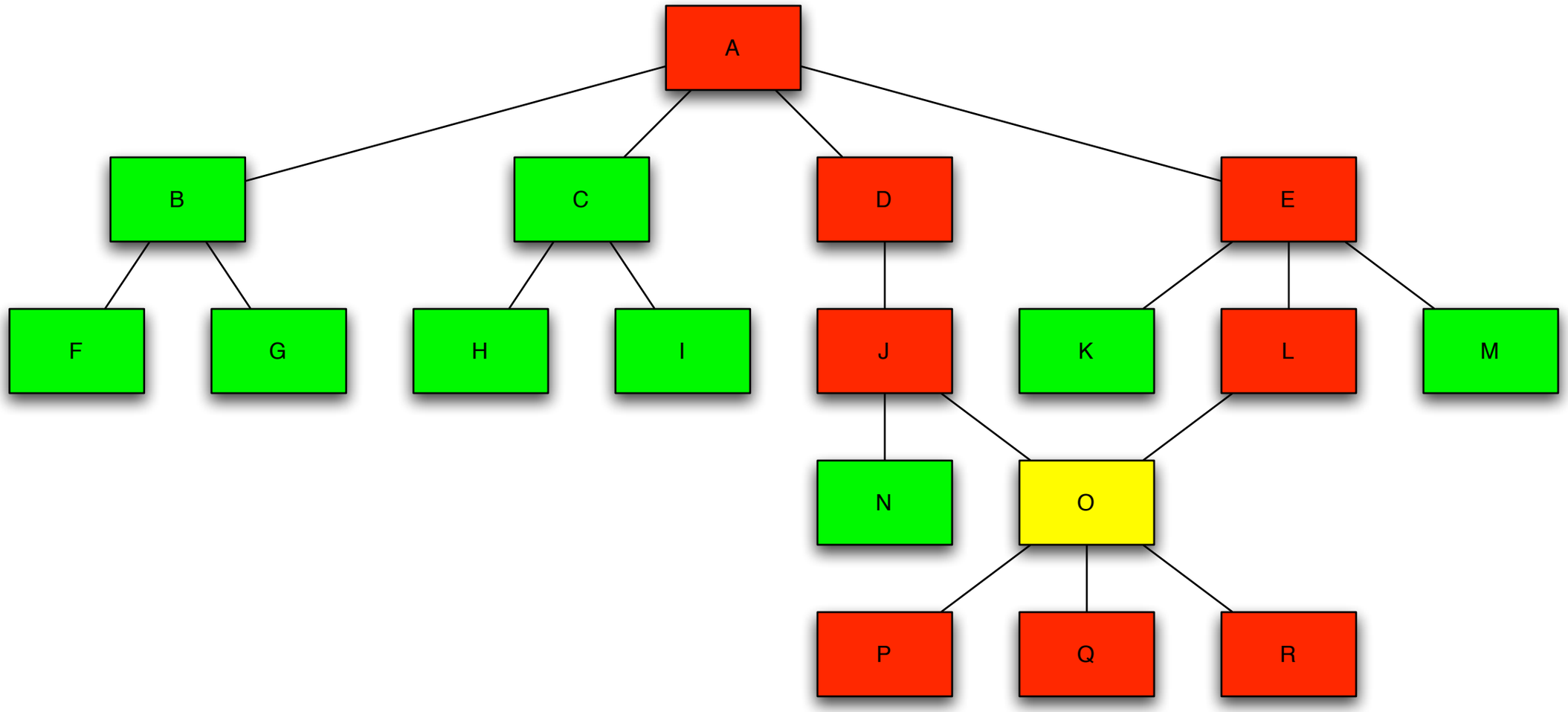
GO ID

Annotation Matrix *Z*(*i*,*j*)

	GO 1	GO 2	GO 3	GO 4	GO 5	…	…	GO *n*
Gene 1	1	0	0	0	1	…	…	1
Gene 2	0	1	0	0	0	…	…	0
Gene 3	0	0	0	1	0	…	…	0
Gene 4	1	0	0	0	0	…	…	0
Gene 5	1	1	0	0	0	…	…	0
…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…
Gene *m*	0	0	0	0	0	…	…	1

A10 A10 A10

nad Vector

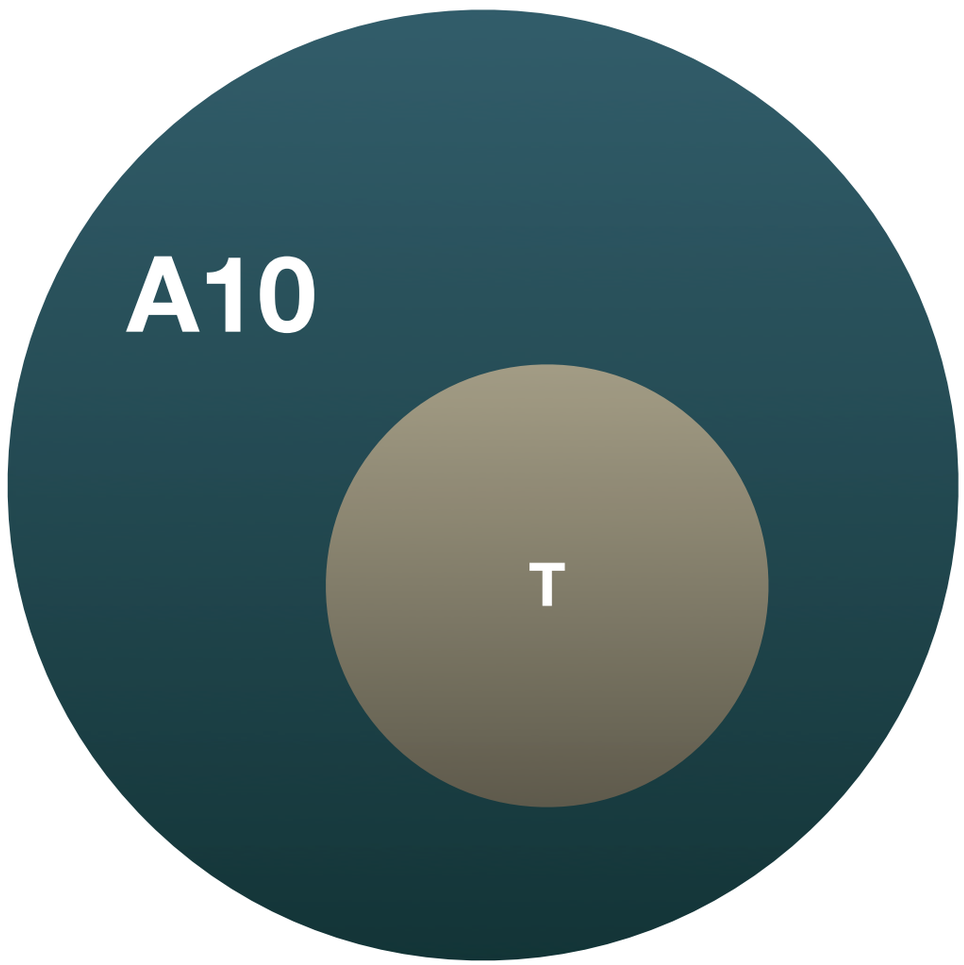


Model Creation

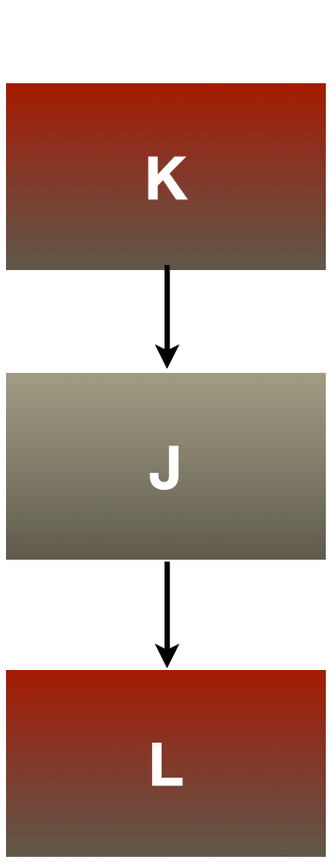
- Based on previously described statistics, generate two machine learning models
    - Decision Tree
    - Bayesian Network
- Also create “independent” model
    - ignore statistics; prediction is based on the fraction of genes in the database that have a given annotation
- 10-fold cross validation and ROC analysis

10-Fold Cross Validation

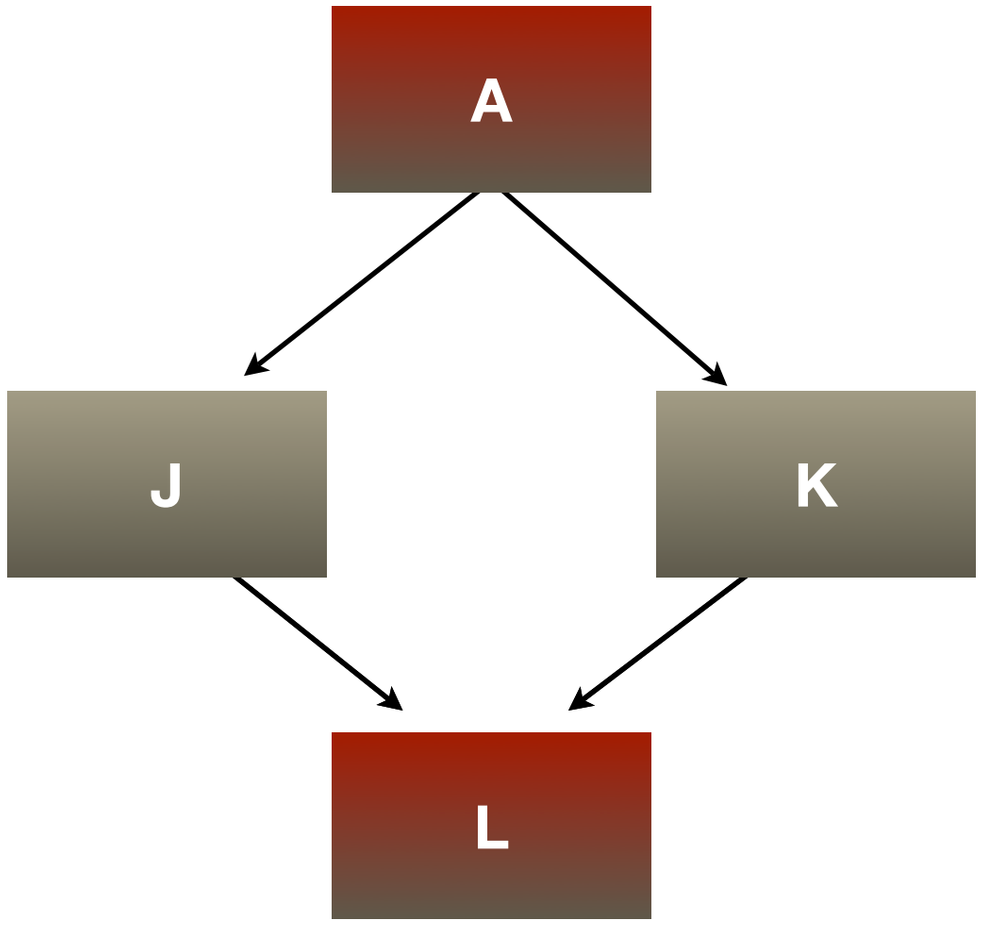
$$q(i,j) = Pr(X_j = 1 \mid \mathbf{nad}(X_j) = \mathbf{nad}(X_j)(i))$$



Explanation of “T” Group



OK

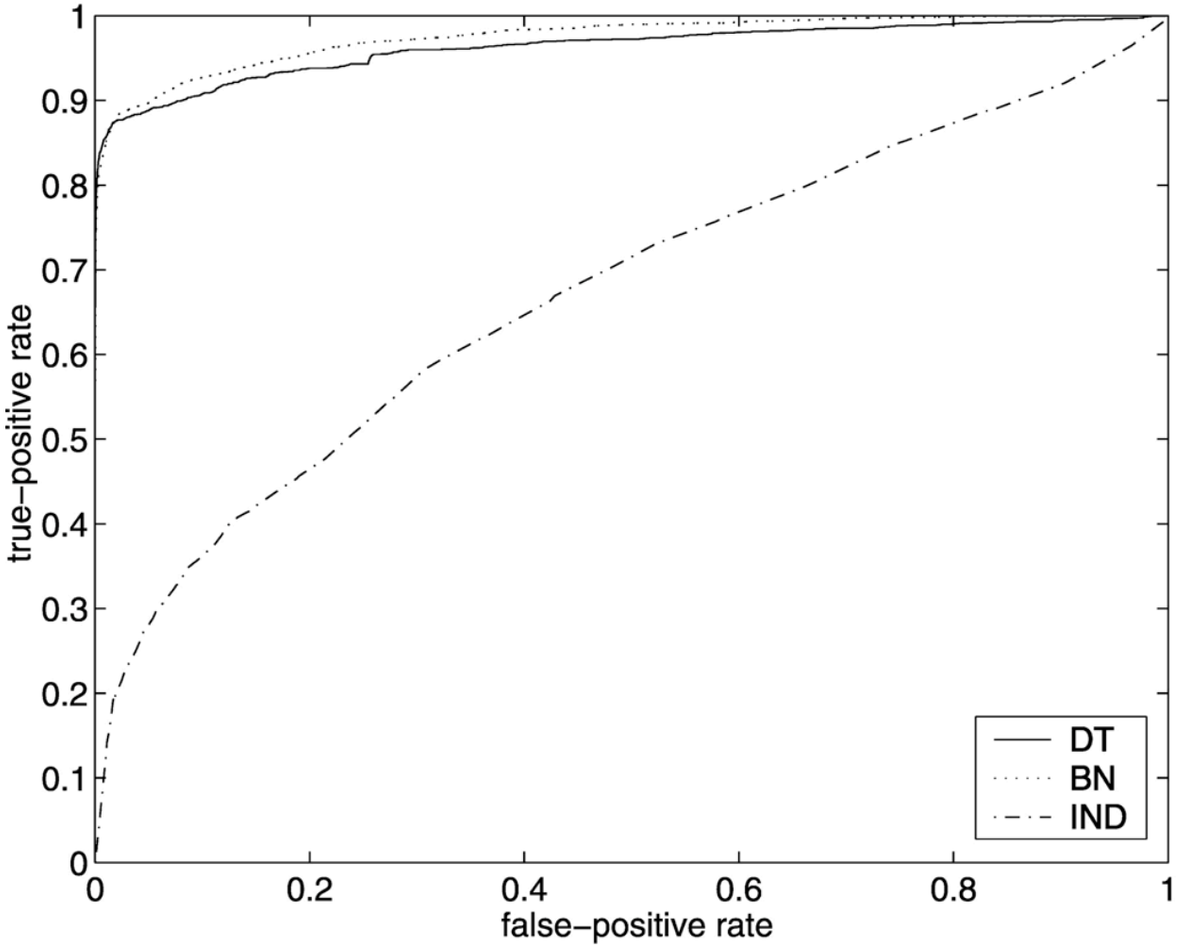


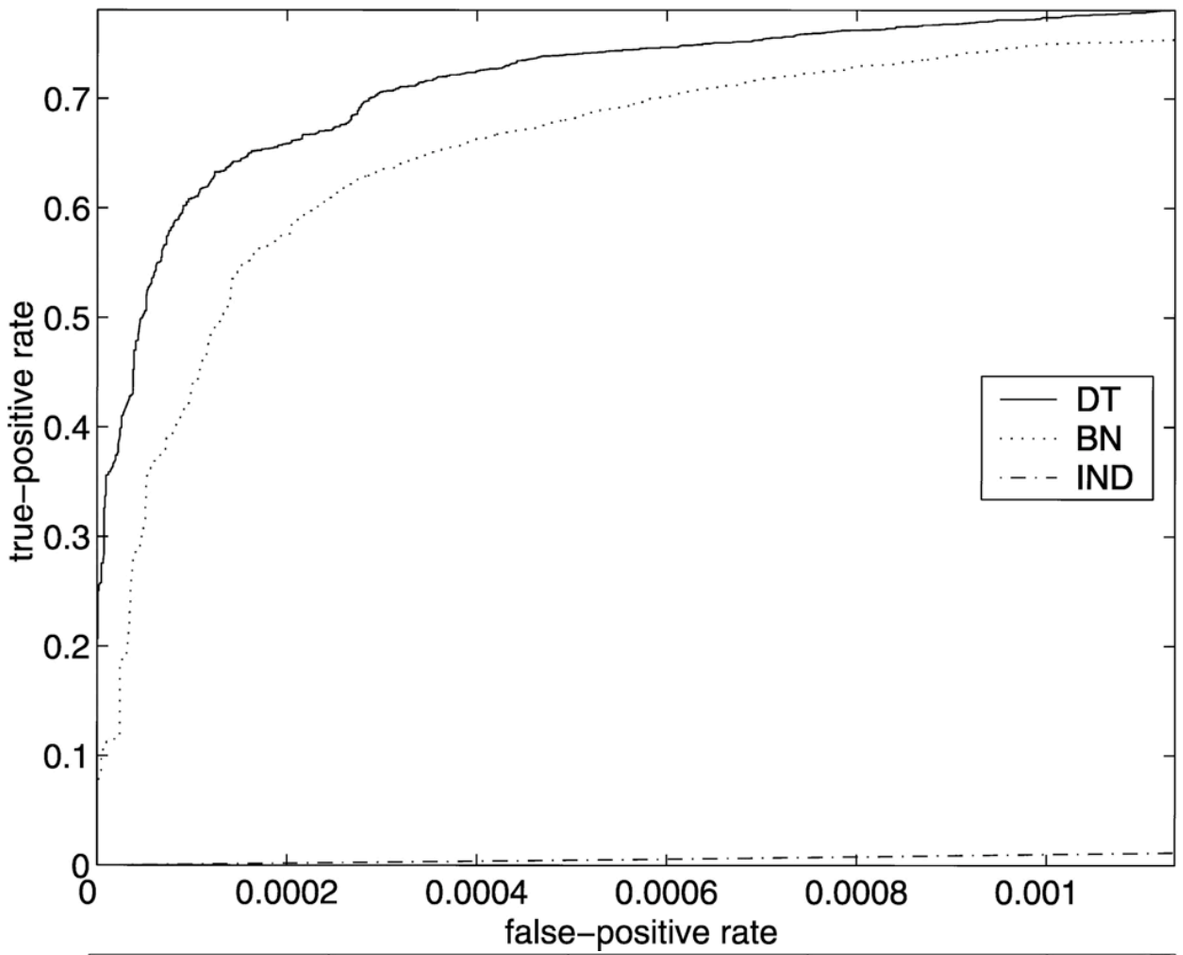
Wrong!

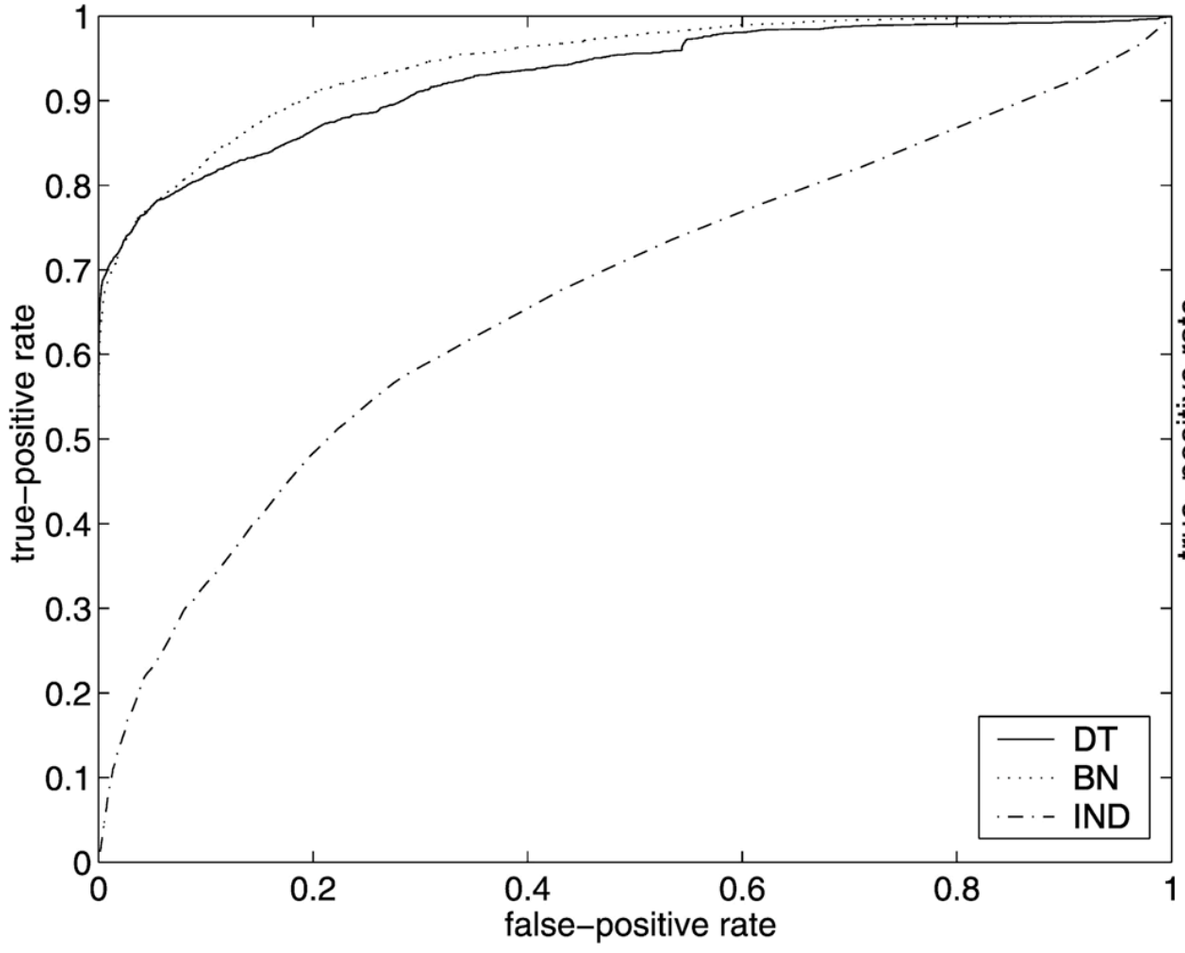
J⇒A

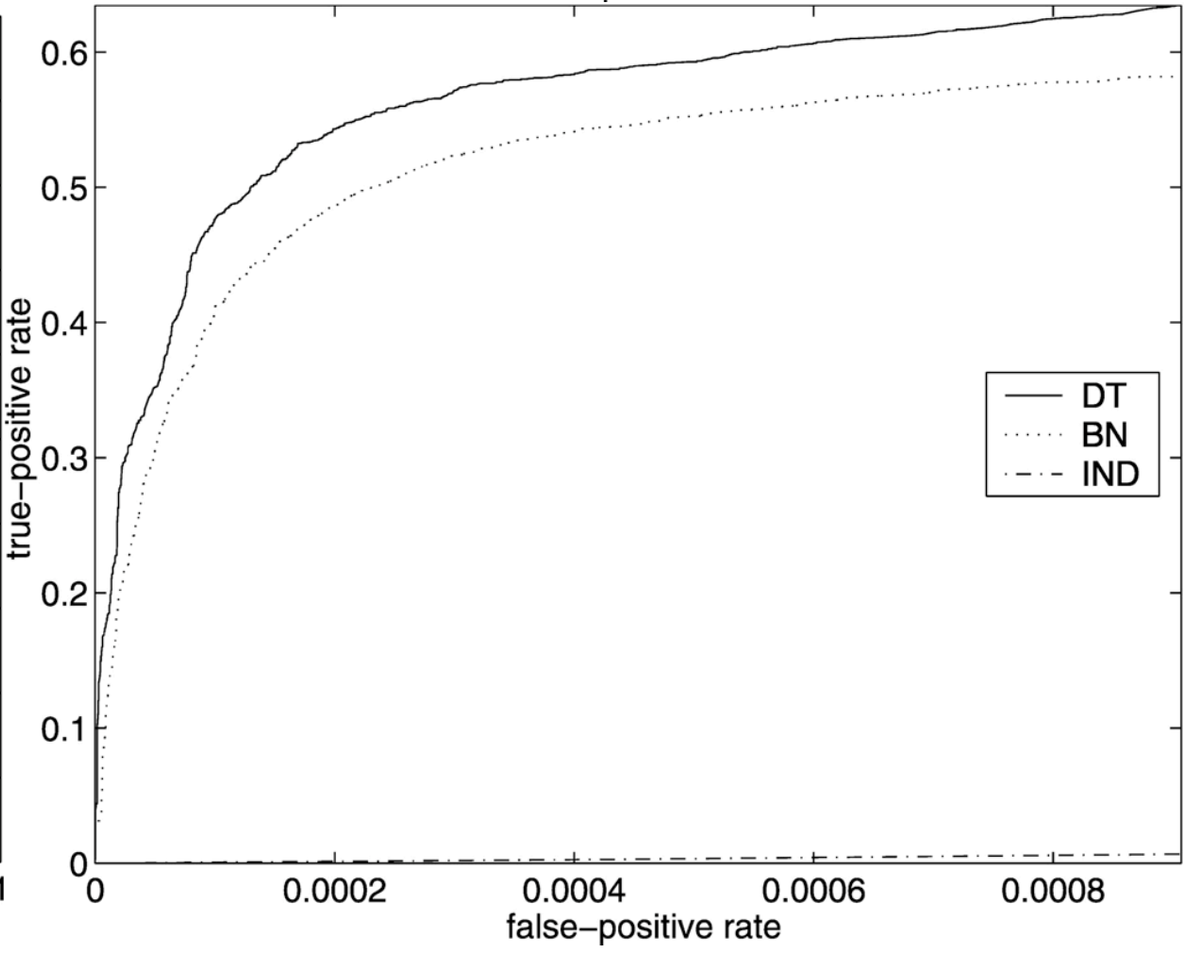
L⇒A,J,K

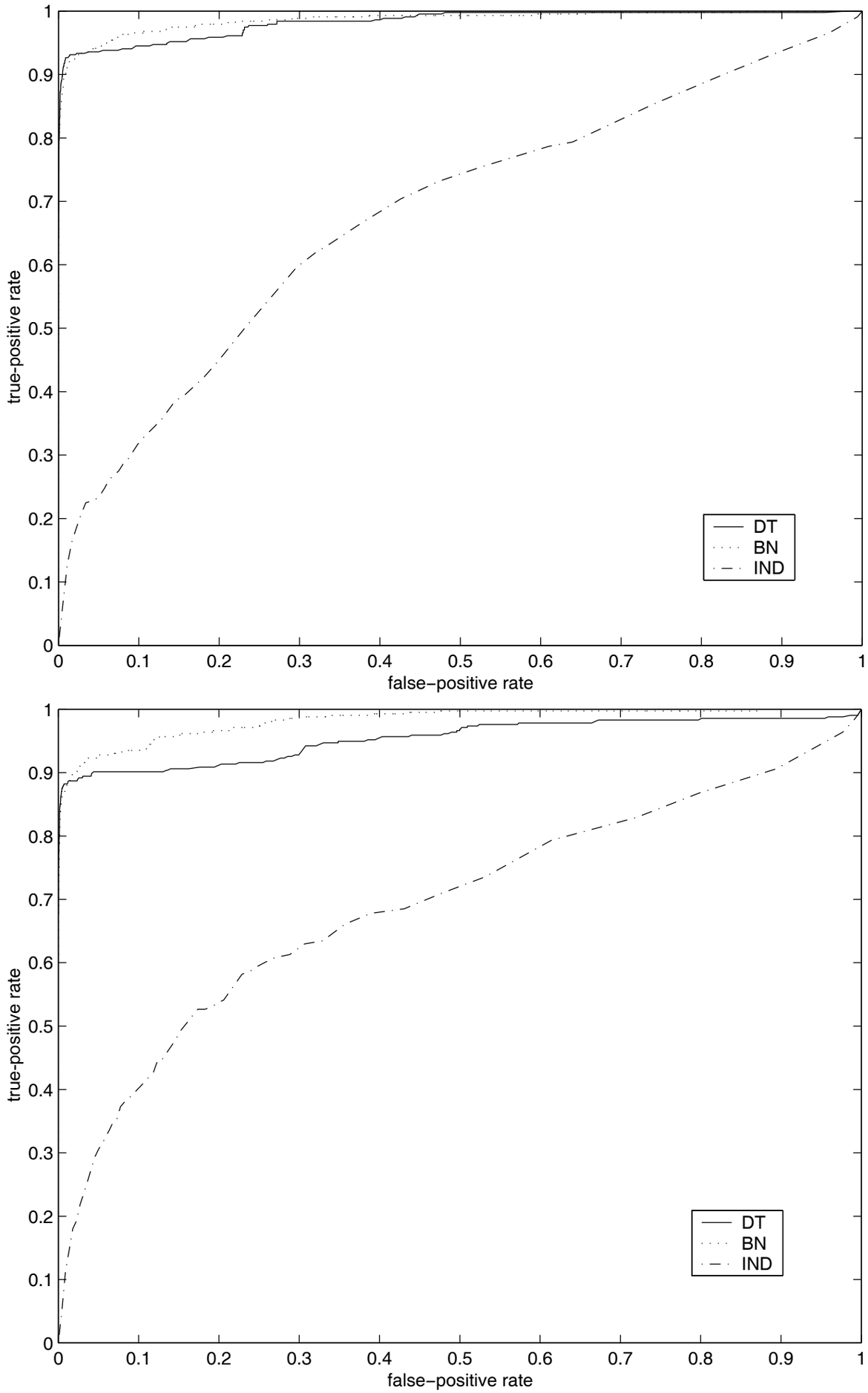
K⇒A


true−positive rate
false−positive rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
DT
BN
IND


true−positive rate
false−positive rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
0.0002
0.0004
0.0006
0.0008
0.001
DT
BN
IND


true−positive rate
false−positive rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
DT
BN
IND


true−positive rate
false−positive rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0
0.0002
0.0004
0.0006
0.0008
DT
BN
IND


true-positive rate
false-positive rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
DT
BN
IND

Errors of Omission: Manual Validation

- Situations where no annotation exists, but predicted probability of annotation is high

- Two authors, each curators from either FlyBase or SGD, manually inspected top 50 predicted omission errors on genes from their specific organism

- 1 = “known to be true”; 2 = “known to be false”; 3 = “neither of the above”

- Many hypotheses led the curators to officially add the predicted annotations to their respective datasets

Possible Applications of the Technique

- Help curators maintain inter-branch logical relationships
- Hypothesis suggestion
- Errors of Omission, Inclusion

Considerations

- Depends on quality of existing curation
    - This is a supervised learning method, after all

Questions

- Why A10? What's magic about "10"?
    - How do lower thresholds affect results?
- Why add genes to FlyBase training set? How did this affect the results for classifiers trained on this set?