

Data Preparation and Interannotator Agreement

BioCreAtIvE Task 1B

Marc Colosimo, Lynette Hirschman
Alexander Morgan, Alexander Yeh
Jeff Colombe

MITRE

March 2004

<http://www.mitre.org/public/biocreative>



This work was supported in part by NSF grant #EIA-0326404

© 2004 The MITRE Corporation. ALL RIGHTS RESERVED.

MITRE

Overview

- What and how we annotated?
- How well did we annotate?
- Are there differences in the data between the organisms?
- Do the abstracts (and Our Task) contain the information of interest?

Resources

- Guidelines on how and what to curate
 - <http://www.mitre.org/public/biocreative>
- Gene lists from model organism databases (fly, mouse, yeast)
 - A list of genes (unique database identifiers).
 - This list is usually for the full paper.
- Synonym lists provided by each database to map alternate gene names, as mentioned in text, to their unique database identifier
- Abstracts from Medline
- Gold Standard lists genes found in the abstracts

Guidelines

- **Define the Task: list the unique gene identifiers for the genes mentioned in the abstract**
 - Genes must come from appropriate organism.
 - Must be listed using unique identifier (in synonym list).
- **What counts**
 - Explicit mentions of genes, gene mutants, alleles, and products.
 - Genes mentioned "in passing" even though these are often not included by the databases.
 - Explicit enumerations of genes.
- **What doesn't count**
 - Families of genes and non-specific mentions.

Steps in the Generation of the Gold Standard

- **Pruning of Full Text Gene Lists for Abstracts**
 - Automatic pruning using pattern matching.
- **Curation of Abstracts**
 - Become familiar with the Guidelines
 - Augmentation of pruned lists.
 - Adjustments to guidelines.
- **Interannotator Agreement Study**
 - Demonstrated need for further clean up.
- **Refinement Using Answer Pooling**
- **Release of "Final" Gold Standard**
 - Available on request.

Genes Found In Abstracts of Test Data: First Pass

Organism	# of Genes on Database List	Manually Found in Abstracts		
		Found on List	Additional Genes	Total Genes
Fly	1571	399	32	431
Mouse	795	290	205	495
Yeast	737	540	75	615

Genes that were on the database list

Genes that we added to the list

Interannotator Analysis of our "Gold Standard"

- Picked ~30 abstracts from each organism.
- Annotated by two additional curators (total of three curators per abstract).
- Compared all three gene lists for conflicting genes.

Interannotator Results

- Found a large number of conflicts for mouse
- Common mistakes
 - Missing a mention
 - Assigning a mention to a gene from a different species

Organism	Total # Genes Reviewed	# Conflicts	% Conflicts
Fly	129	17	13%
Mouse	89	28	31%
Yeast	64	6	9%

Answer Pooling and Selection Analysis

- Used participants results to check the Gold Standard.
- Looked at genes that $\geq 75\%$ of the participants returned as false positives.
- Looked at genes that ALL the participants failed to return (misses).

Answer Pooling and Selection Results

- Found a large number of conflicts for mouse
- Common mistake by participants was assigning a mention from a different species
- Common mistake by us was missing a mention

Organism	Total # of Genes	# Conflicts	% Conflicts
Fly	431	20	5%
Mouse	495	113	23%
Yeast	615	33	5%

Changes in the Gene Lists after Answer Pooling

Organism	Original	False Positives	Missed	Final	% Changed
Fly	431	4	2	429	1.4
Mouse	495	17	66	544	15.3
Yeast	615	7	9	613	2.6

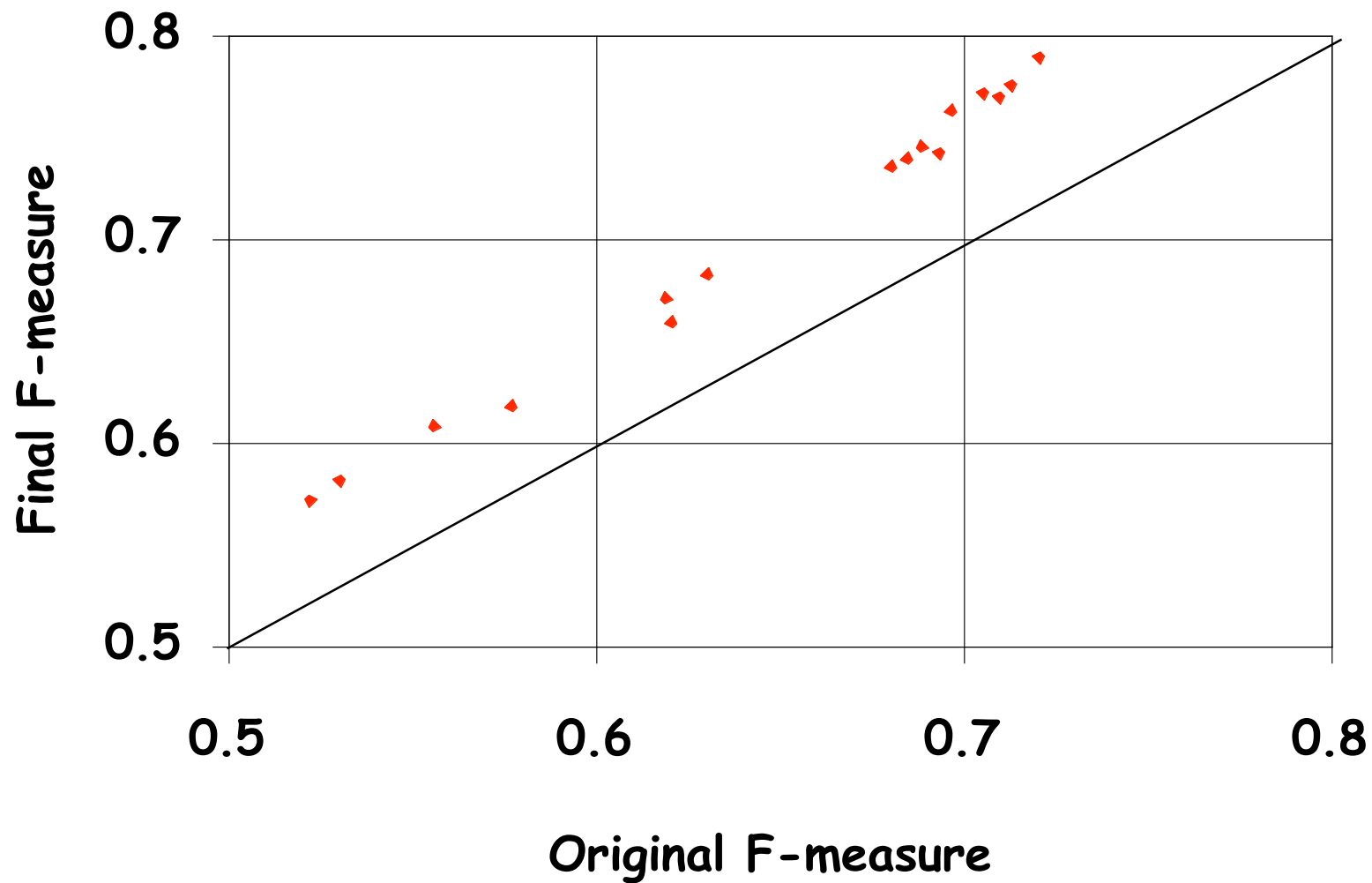
Genes that we
incorrectly assigned

Genes that we
missed

Effects of Changes in the Gold Standard

Metric	Scores of the Original Gold Standards		
	Fly	Mouse	Yeast
F-Measure	0.993	0.920	0.987
Precision	0.991	0.966	0.989
Recall	0.995	0.879	0.985

Changes in Participant's Mouse Scores



Why are Mouse Abstracts Harder to Curate than Fly?

● Number of Genes and Synonyms

- Mouse has ~25,000 predicted genes and ~53,000 gene entries versus Fly which has ~14,000 genes and ~36,000 entries.
- Mouse and Fly have about the same number of synonyms, with ~100,000 each.
 - Fly has 2.8 synonyms per gene.
 - Mouse has 2.1 synonyms per gene.
- Fly has more typographical variations.

● Gene Names

- Fly gene names are often common single words, such as white (w), hedgehog (hh), etc...
- Mouse gene names are usually multiple words, such as *mesoderm specific transcript (Mest)*.

Why are Mouse Abstracts Harder to Curate than Fly (Continued)?

Organism	Total # Additional Genes Added	% New Genes (out of Total)	Total Genes
Fly	34	7.9%	429
Mouse	271	49.8%	544
Yeast	84	13.7%	613

Why are Yeast Abstracts Easy to Curate?

- **Small Number of Genes**

- Yeast has ~6000 predicted genes and has ~8000 gene entries, with ~ 14,000 synonyms.

- **Enforced Nomenclature**

- Gene name abbreviations consist of three letters (the gene symbol) followed by an integer, such as PEP12 or pep12.
- Protein name abbreviations are similar to the gene name abbreviations, but usually have a "p", such as Pep12p.

- **Database Gene List Corresponded to Abstract**

- Curators often curated from abstracts, so the gene list needed less "clean-up".

Why are Yeast Abstracts Easy to Curate (Continued) ?

Organism	# Genes on Database List	# Genes Hand Found in Abstract	% Overlap With Database List
Fly	1571	399	25.4
Mouse	795	290	37.1
Yeast	737	540	73.3

Summary

- Well defined guidelines are needed.
- Annotators need to be checked against each other (interannotator analysis).
- Some organisms were harder to annotate than others (fly and yeast easy, mouse hard).
- In general, abstracts are a poor resource for identifying gene names in papers.
- **If we do this again...**
 - Should it be the "real" task, using full text and reproducing gene lists in model organism DB?
Lots of training data
But task will be harder for participants
 - Or should we do the same thing as this time?
Annotation is costly and error prone

The End

