

Task 1B: Normalized Gene List Extraction*

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

MITRE

March 2004

*This work has been supported in part under NSF Grant EIA 0326404

Outline

- What is Normalized Gene List Extraction?
- Why is this interesting?
- What are the results?
- What have we learned?

What is the Gene List Extraction Task?

- Given a set of abstracts:

A locus has been found, an allele of which causes a modification of some allozymes of the enzyme esterase 6 in *Drosophila melanogaster*. There are two alleles of this locus, one of which is dominant to the other and results in increased electrophoretic mobility of affected allozymes. The locus responsible has been mapped to 3-56.7 on the standard genetic map (Est-6 is at 3-36.8). Of 13 other enzyme systems analyzed, only leucine aminopeptidase is affected by the modifier locus. Neuraminidase incubations of homogenates altered the electrophoretic mobility of esterase 6 allozymes, but the mobility differences found are not large enough to conclude that esterase 6 is sialylated.

What is the Gene List Extraction Task?

- Return unique gene IDs mentioned in the abstract

Abstract ID

fly_00035_training

fly_00035_training

Organism Gene ID

FBgn0000592

FBgn0026412

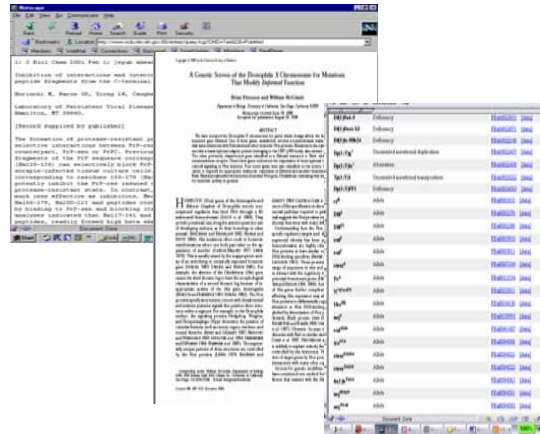
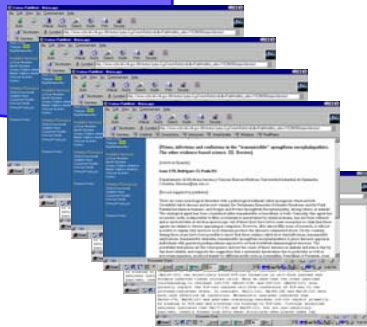
A locus has been found, an allele of which causes a modification of some allozymes of the enzyme **esterase 6** in *Drosophila melanogaster*. There are two alleles of this locus, one of which is dominant to the other and results in increased electrophoretic mobility of affected allozymes. The locus responsible has been mapped to 3-56.7 on the standard genetic map (**Est-6** is at 3-36.8). Of 13 other enzyme systems analyzed, only **leucine aminopeptidase** is affected by the modifier locus. Neuraminidase incubations of homogenates altered the electrophoretic mobility of esterase 6 allozymes, but the mobility differences found are not large enough to conclude that **esterase 6** is sialylated.

Why Gene List Extraction? A (Simplified) Curation Pipeline

3. Curate genes from paper

2. List curatable genes

1. Select papers



Why Are Normalized Gene Lists Interesting?

- Constitutes a step in the curation process
- Builds on Task 1A: gene mentions in text, but using vocabularies from model organism databases
- Focuses on normalization
 - Mapping of mentions to unique identifier
 - Biologically significant step, especially where terminology is ambiguous
 - Normalization of nomenclature critical for finding genes or proteins in the literature
- Builds towards Task 2: association of evidence for gene function

Data from Fly, Mouse Yeast Databases

- Text: abstracts from PubMed
 - Full articles better, but hard to obtain
- Synonym list: developed from each database
- Training data (noisy)
 - Start with gene lists for each paper from model organism DBs
 - Adjust DB gene list automatically to reflect information contained in abstract
- Gold Standard
 - Dev test data: hand-checked, to add in “missing” genes
 - Test data: hand-corrected - see talk by Marc Colosimo

Entity Extraction Task 1B: Data Set Sizes

	Fly	Mouse	Yeast
Training	5000	5000	5000
Development Test	108	250	110
Test	250	250	250

***Each abstract is around 250 words**

Example of Noisy Training Data

A locus has been found, an allele of which causes a modification of some allozymes of the enzyme **esterase 6** in *Drosophila melanogaster*. There are two alleles of this locus, one of which is dominant to the other and results in increased electrophoretic mobility of affected allozymes. The locus responsible has been mapped to 3-56.7 on the standard genetic map (**Est-6** is at 3-36.8). Of 13 other enzyme systems analyzed, only **leucine aminopeptidase** is affected by the modifier locus. Neuraminidase incubations of homogenates altered the electrophoretic mobility of esterase 6 allozymes, but the mobility differences found are not large enough to conclude that **esterase 6** is sialylated.

Original DB Gene List

fly_00035_training	FBgn0000592
fly_00035_training	FBgn0002722

Auto
Y
N

Removed:
Not found in
abstract

Gene ID and synonyms:

FBgn0000592: **Est-6**, **Esterase 6**, CG6917, Est-D, EST6, est-6, Est6, Est, EST-6, Esterase-6, est6, Est-5, Carboxyl ester hydrolase

FBgn0002722: m(Est-6), modifier of Esterase 6, M-est, m-est

Creating the Gold Standard

- The automatically cleaned data is carefully checked by a biologist
 - For genes mentioned in the abstract but not on the list (FBgn0026412: **leucine aminopeptidase**)
 - For gene mentions missed in the matching procedure (e.g., "Esterase-6" vs "esterase 6")
- Final hand-checked answer key

FILE	GeneID	Auto	Hand
fly_00035_training	FBgn0000592	Y	Y
fly_00035_training	FBgn0002722	N	N
fly_00035_training	FBgn0026412	X	Y

Added:
Not in curated
gene list

Why Are Normalized Gene Lists Hard?

- Step 1: Finding gene mentions
 - This is not a solved problem: best Task 1A system performed at 83% F-measure*
- Step 2: Normalization to unique ID
 - No match: name may not match anything in synonym list (typographical variants)
 - Ambiguity: name may match 2 or more names from synonym list
- Step 3: Duplicate removal
 - Depends on correct mapping to Gene ID
- Results for "best of breed" F-measure
 - Yeast: 92% Fly: 82% Mouse: 79%

*F-measure: harmonic mean of precision, recall:
$$2 * P * R / (P + R)$$

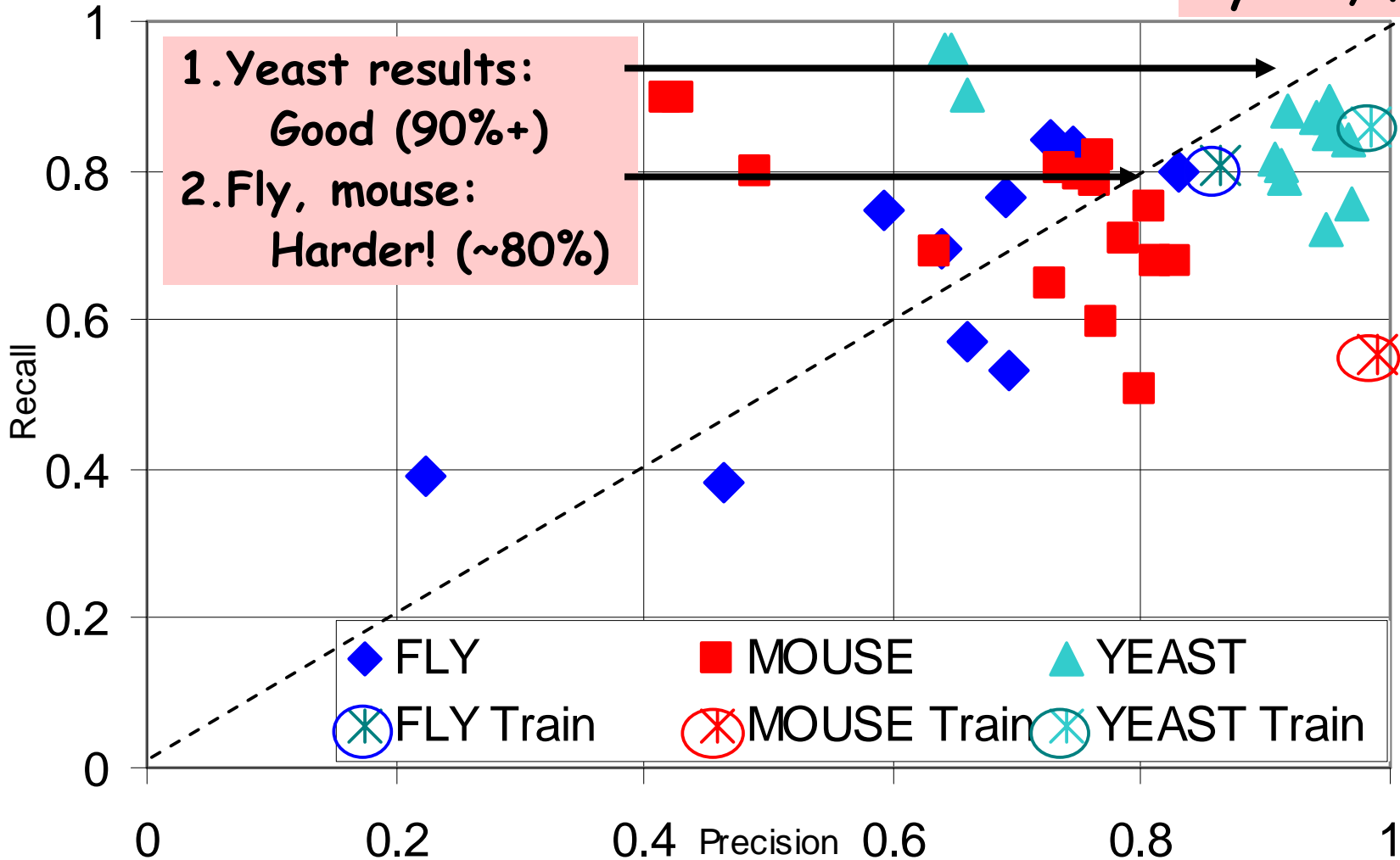
Submissions

- 8 Teams
 - 4 groups also participated in Task 1A
 - 1 group also participated in Task 2
- Varied approaches
 - Hybrid
 - 3-stage approach of gene name tagging, match to synonym, disambiguation
 - Machine learning
 - Conditional HMM for gene names, Max Entropy classifier for matching gene IDs
 - Lexicon-based
 - Cleaning and enrichment of synonym lists
 - Enriched the lexicon for high recall

Results: Precision vs Recall

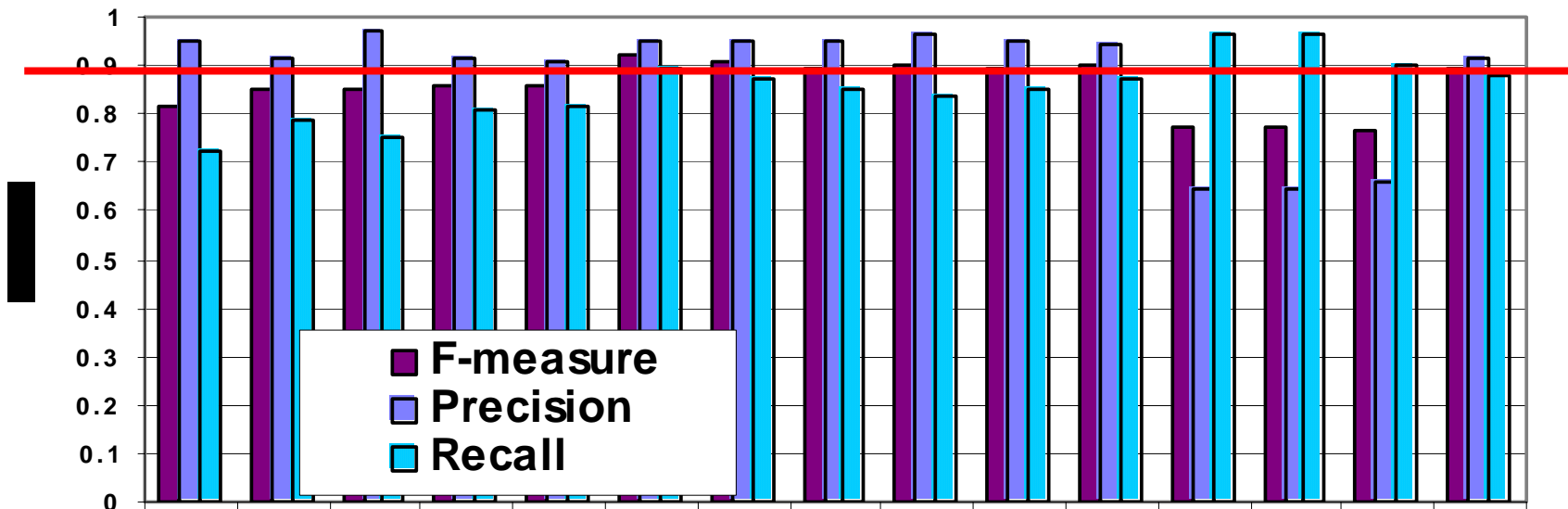
Training data:
Ceiling for
yeast, fly?

1. Yeast results:
Good (90%+)
2. Fly, mouse:
Harder! (~80%)



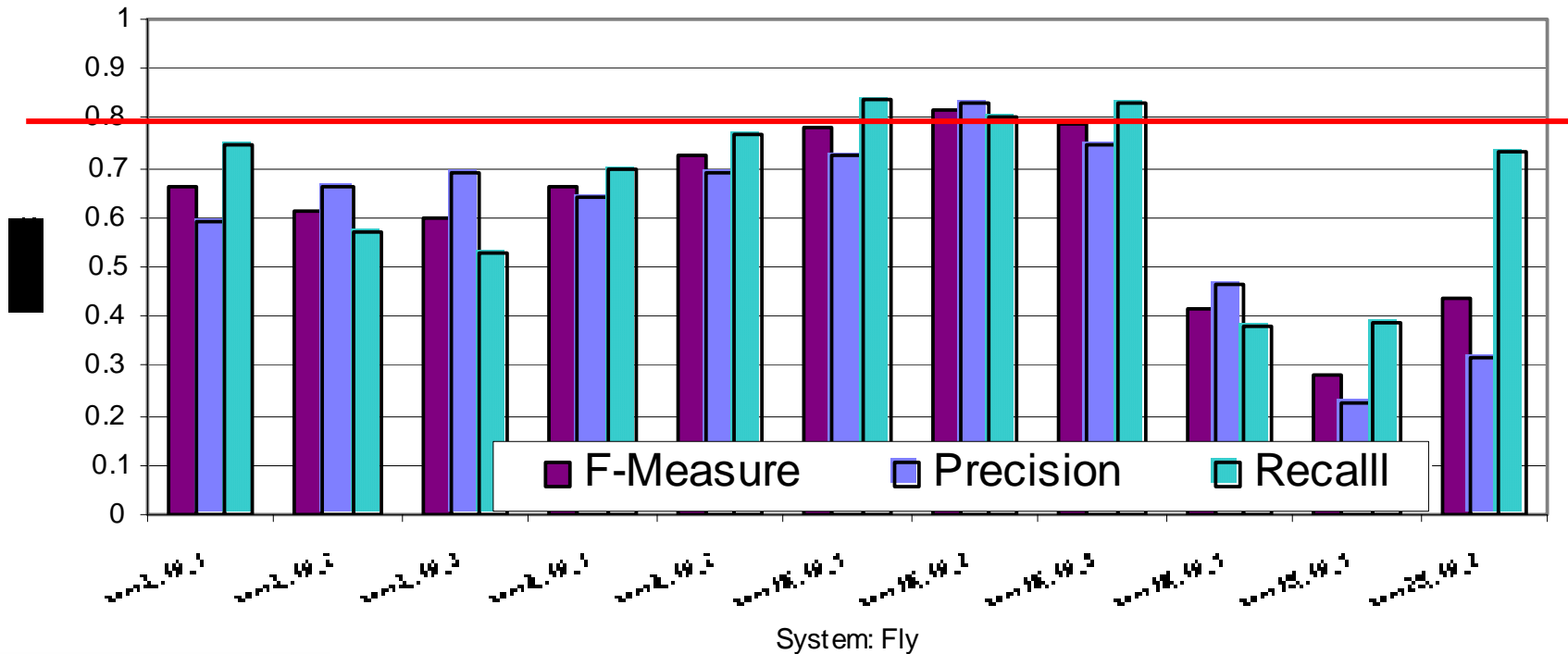
System Performance: Yeast

YEAST	High	Median	AutoTrain
F-measure	0.921	0.858	0.918
Precision	0.969	0.940	0.985
Recall	0.962	0.848	0.860



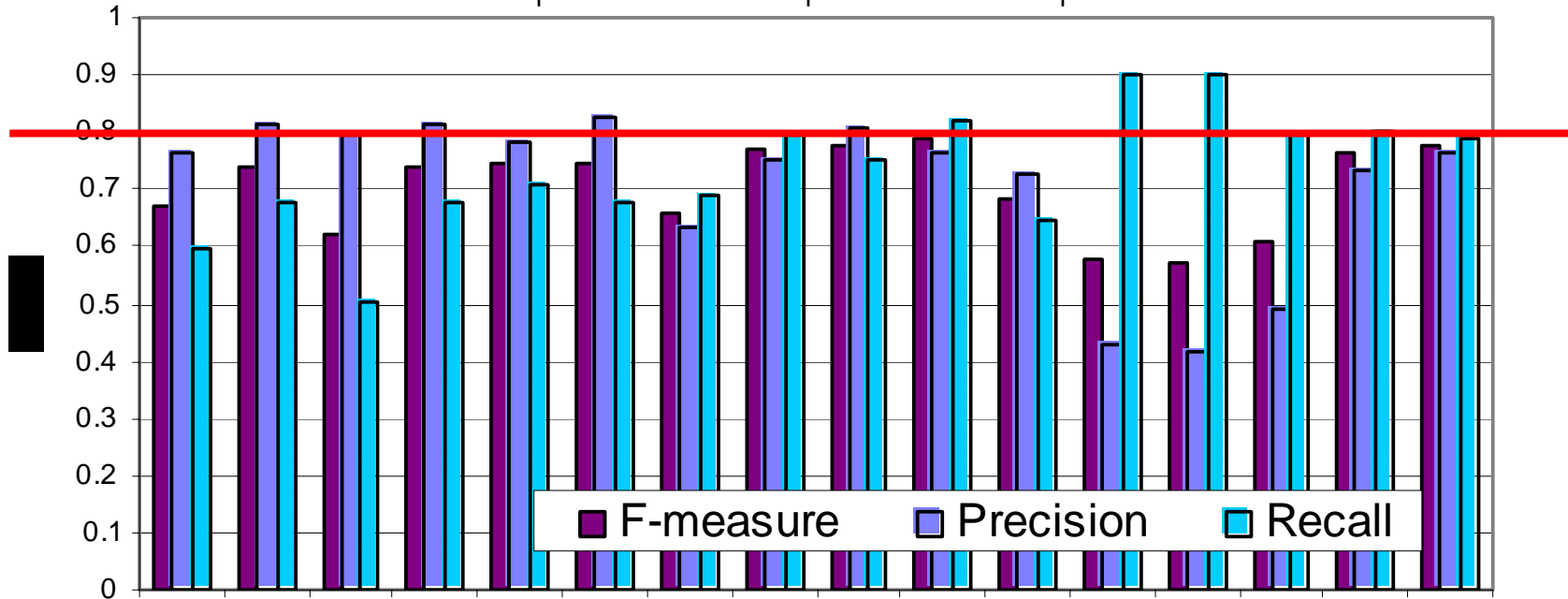
System Performance: Fly

FLY	High	Median	AutoTrain
F-measure	0.815	0.661	0.834
Precision	0.831	0.659	0.863
Recall	0.841	0.732	0.807



System Performance: Mouse

MOUSE	High	Median	AutoTrain
F-measure	0.791	0.738	0.709
Precision	0.828	0.765	0.990
Recall	0.898	0.730	0.552



Summary of Results (1)

- Different systems did well on different aspects
 - Overall, 7 of the 8 groups scored in the top 3 for at least one measure!
- Good correlation with Task 1A performance
 - Does Task 1A performance set an implicit ceiling for Task 1B?
- Range of successful techniques
 - Heuristic system, machine learning, hybrid, pure lexical based system all did well

Summary of Results (2)

- Results are dependent on the organism and the quality of the training data
- Yeast is "easy" (92% top F-measure)
 - Clean nomenclature, little ambiguity
 - Training data was "clean" (92% F-measure)
- Fly is harder: (82% top F-measure)
 - Extensive overlap w English words (e.g., "not", "per")
 - Training data somewhat noisy (83% F)
- Mouse is harder (79% top F-measure)
 - Training data was very noisy (71% F)
recall = 55%, precision = 99%

Lessons Learned on Task 1B

- Rapid adaptation possible to different domains
 - Should we do more organisms next time?
- Biologically motivated
 - Part of the curation pipe-line
 - Normalization important for retrieval, functional annotation
 - But still simplified (because no full text); should we use full text and real gene lists?
- Performance
 - Systems may be usable now on some real tasks (e.g., parts of yeast curation)
 - What performance is needed for mouse, fly?

Back-Up

