Unsupervised gene/protein entity normalization using dictionaries automatically extracted from online databases

Aaron M. Cohen

Oregon Health & Science University

June 24, 2005

Outline of Talk

- Background
- Methods
- Evaluation
- Results
- Discussion
- Conclusions

Gene/Protein NER and Normalization

- Named Entity Recognition (NER)
  - Identify the terms/strings within text that refer to genes and proteins
- Normalization
  - Mapping of recognized terms to unique identifiers
- Important fundamental task in biomedical text mining, basis for further text mining:
  - Association detection, Relationship extraction, etc.
- Proteins/genes that code them often treated as equivalent in biomedical text mining

NER Plus Normalization (NER+N)

- NER simply identifies terms of a given type
- Additional normalization step required to:
  - Ensure identical handling of identical concepts
  - Increase sample size by combining synonyms
- NER can be dictionary or machine-learning based (SVM, CRF, HMM, etc.)
- "+N" requires dictionaries in some form
- Evaluated in BioCreative Task 1B:

Hirschman L, Colosimo M, Morgan A, Yeh A: **Overview of BioCreAtIvE task 1B: normalized gene lists**. *BMC Bioinformatics* 2005;**6**(Suppl 1):S11.

Dictionary NER+N Basics

- Online genome databases
  - Great information source (single & multi-organism)
  - Represent many years of curator effort
- How well can we do with a dictionary-based NER+N approach for genes and proteins?
- Potential Benefits:
  - Simple, leverage genome DB information
  - Avoid machine learning, training corpus issues
  - Automatically update to include new genes
  - Good performance on long multi-word names
  - Fast (most computation is constructing dictionary)

Practical Questions:

- Which curated databases are the best source of gene symbols and names?
- Can simple rules be used for generating sufficient orthographic variants?
- How can common English word false positives be reduced?
- Overall, how well can a purely dictionary-based, untrained approach perform?

Method

Compile Time

Pre-processing

1. Building the initial dictionary
2. Generate orthographic variants
3. Separate common English words
4. Screen out most common words

Run Time

5. Search text
6. Perform disambiguation

1. Build the initial dictionary

- Extract gene names, symbols, aliases from downloadable extracts of online databases
- Simple process, but each database has its own format
- Used several databases for this work:
  - MGI (mouse specific)
  - Saccharomyces (yeast specific)
  - UniProt (multi-organism)
  - LocusLink (multi-organism)
  - Entrez Gene (multi-organism)
- Organize names by unique gene identifier

2. Generate variants

- Expand names in dictionary using orthographic generator rules:
  1. Replace spaces with hyphens
  2. Replaces hyphens with spaces
  3. Remove internal spaces & hyphens
  4. Insert hyphen before trailing digits
  5. Replace trailing a/b following digit with alpha/beta
  6. Replace trailing “–1” or “–2” with Roman numeral
  7. Append “p” to single word names (yeast only)
- Apply singly and iteratively until no new names are generated

3 & 4. Separate English words

- Many genes have names easily confused with common English words
- Reduce false positives by handling these specially
- Separate into two dictionaries using Moby (75K) into main and confusion dictionaries
  - Example: “DARK”
- Remove names that match stop words (300 most common English words + a few others)

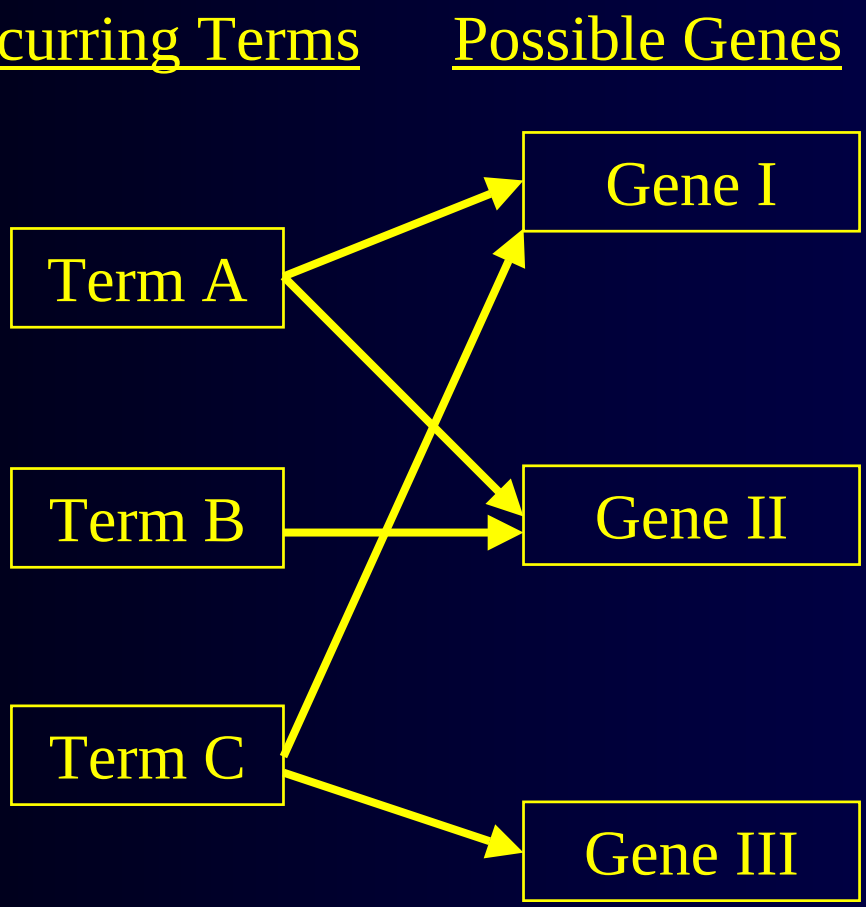
5. Search text

- Search text using the two dictionaries
  - Case-insensitive match for main dictionary
  - Strict case-sensitive match for confusion dictionary
- No need to tokenize, first search for matching names, then check for bordering delimiters
- Collect matching (string, unique identifier) pairs for entire text sample
- Text sample can be any size, but Abstract/Paragraph good for next step

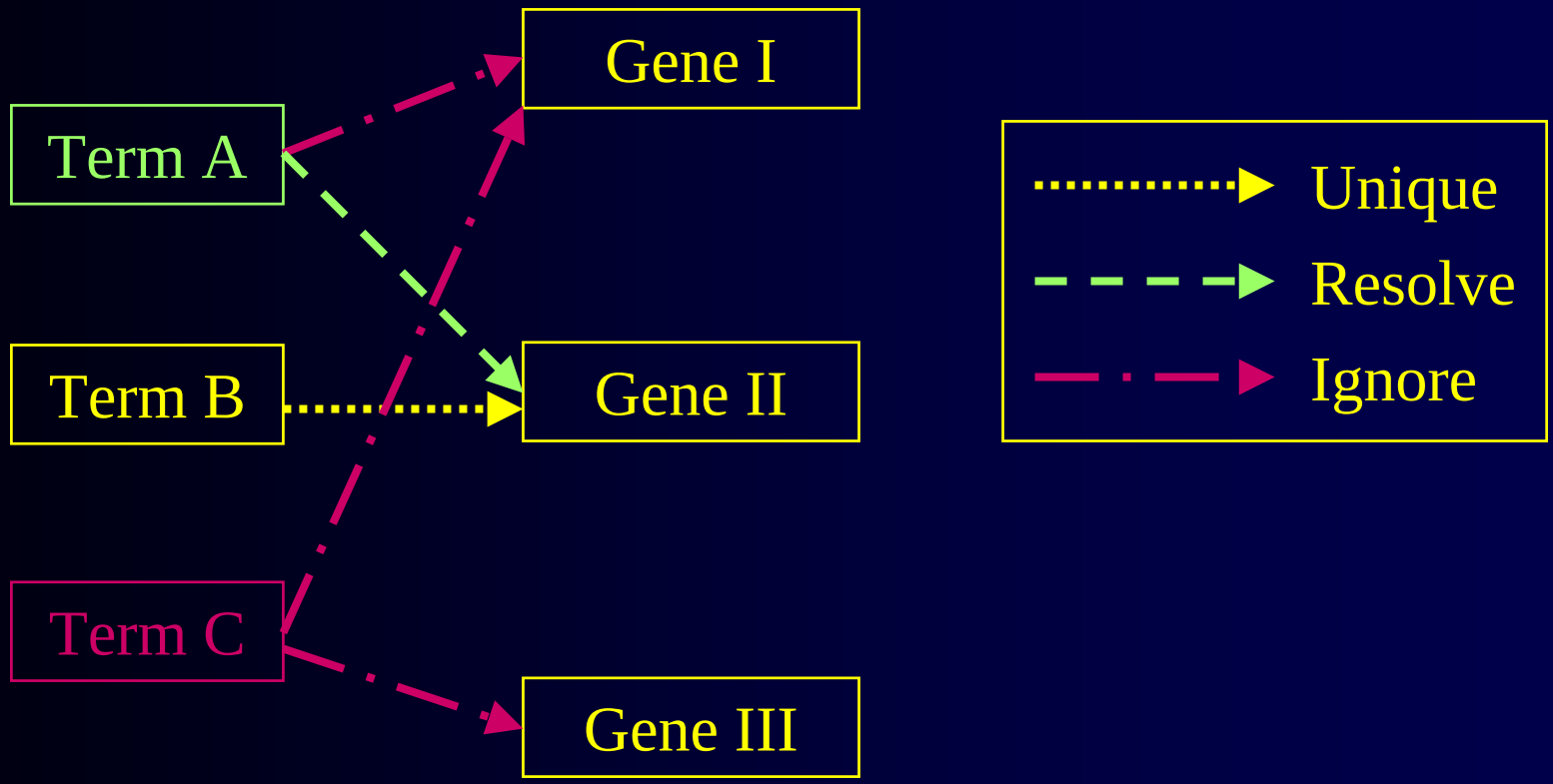
6. Disambiguation

- Large number of terms refer to more than one gene/protein (5% intra-, 85% inter-species)
- Need to disambiguate these cases to perform normalization, extract maximum accurate information
- Previous algorithms based on simply ignoring ambiguous terms or concept frequency
- Our approach based on using complementary unambiguous information in the text sample

Disambiguation Process 1



Disambiguation Process 2



Evaluation

1. Select best general DB source of gene names using GENIA corpus as test collection
2. Evaluate NER+N performance on mouse and yeast using BioCreative Task 1B test collections and scoring method

Results: NER only

- NER-only using names from a single database:

Dictionary	Precision	Recall	F-measure
Entrez	0.735	0.776	0.755
LocusLink	0.723	0.773	0.747
UniProt	0.785	0.474	0.591

- NER-only using two databases:

Dictionaries	Precision	Recall	F-measure
Entrez	0.735	0.776	0.755
Entrez+UniProt	0.707	0.792	0.747
Entrez+LocusLink	0.734	0.780	0.756

Results: NER+N on Mouse

- With Organism-specific database:

Dictionary	Precision	Recall	F-measure
Entrez/MGI	0.775	0.726	0.750
MGI	0.710	0.535	0.610

- Effect of individual features:

System	Precision	Recall	F-measure	Difference
full system	0.775	0.726	0.750	-
- case	0.493	0.746	0.594	-15.6%
- stop	0.643	0.726	0.682	-6.8%
- variant	0.771	0.693	0.730	-2.0%
- ambiguity	0.697	0.748	0.722	-2.8%
- all	0.301	0.713	0.423	-32.7%

Mouse & Yeast Results Comparison

Organism	System	Precision	Recall	F-measure
Mouse	biocreative-highest	0.765	0.819	0.791
	cohen	0.775	0.726	0.750
	biocreative-median	0.765	0.730	0.738
	biocreative-lowest	0.418	0.898	0.571
Yeast	biocreative-highest	0.950	0.894	0.921
	cohen	0.950	0.837	0.890
	biocreative-median	0.940	0.848	0.858
	biocreative-lowest	0.661	0.902	0.763

Discussion

- Entrez Gene best single source of names
- Entrez + species-specific database for NER+N with unique id sufficient
- All dictionary pre-processing methods and ambiguity resolution improved performance
- Likely not to work well on Fly
  - Too many genes polysemous with common words
  - Cannot rely on differences in capitalization
- Easily extended to human, but how to test?

Conclusions

- State of the art performance on mouse and yeast using simple dictionary-based methods
- Makes good use of curated information
- Dictionaries processed without human intervention or review
- Easy to incorporate additions to genomic databases, can be done automatically
- Training corpus and system training not needed