Information Extraction from Biomedical Text

BMI/CS 776

www.biostat.wisc.edu/bmi776/

Mark Craven

craven@biostat.wisc.edu

Spring 2009

The Information Extraction Task: Named Entity Recognition

Analysis of Yeast PRP20 Mutations and Functional Complementation by the Human Homologue RCC1, a Protein Involved in the Control of Chromosome Condensation

Fleischmann M, Clark M, Forrester W, Wickens M, Nishimoto T, Aebi M

Mutations in the **PRP20** gene of yeast show a pleitropic phenotype, in which both mRNA metabolism and nuclear structure are affected. **SRM1** mutants, defective in the same gene, influence the signal transduction pathway for the **pheromone** response . . .
By **immunofluorescence microscopy** the **PRP20** protein was localized in the **nucleus**. Expression of the **RCC1** protein can complement the temperature-sensitive phenotype of **PRP20** mutants, demonstrating the functional similarity of the yeast and mammalian proteins

- proteins
- small molecules
- methods
- cellular compartments

The Information Extraction Task: Relation Extraction

Analysis of Yeast PRP20 Mutations and Functional Complementation by the Human Homologue RCC1, a Protein Involved in the Control of Chromosome Condensation

Fleischmann M, Clark M, Forrester W, Wickens M, Nishimoto T, Aebi M

Mutations in the PRP20 gene of yeast show a pleitropic phenotype, in which both mRNA metabolism and nuclear structure are affected. SRM1 mutants, defective in the same gene, influence the signal transduction pathway for the pheromone response . . .
By immunofluorescence microscopy the **PRP20** protein was localized in the **nucleus**. Expression of the RCC1 protein can complement the temperature-sensitive phenotype of PRP20 mutants, demonstrating the functional similarity of the yeast and mammalian proteins

→ subcellular-localization(**PRP20**, **nucleus**)

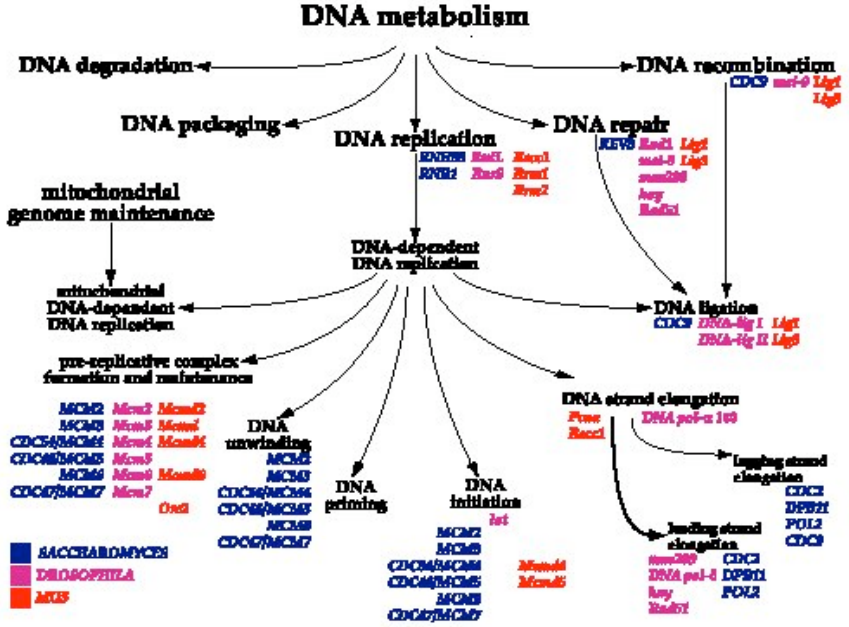
Motivation for Information Extraction

- motivation for named entity recognition
  - better indexing of biomedical articles
  - identifying relevant passages for curation
  - assisting in relation extraction
- motivation for relation extraction
  - assisting in the construction and updating of databases
  - providing structured summaries for queries

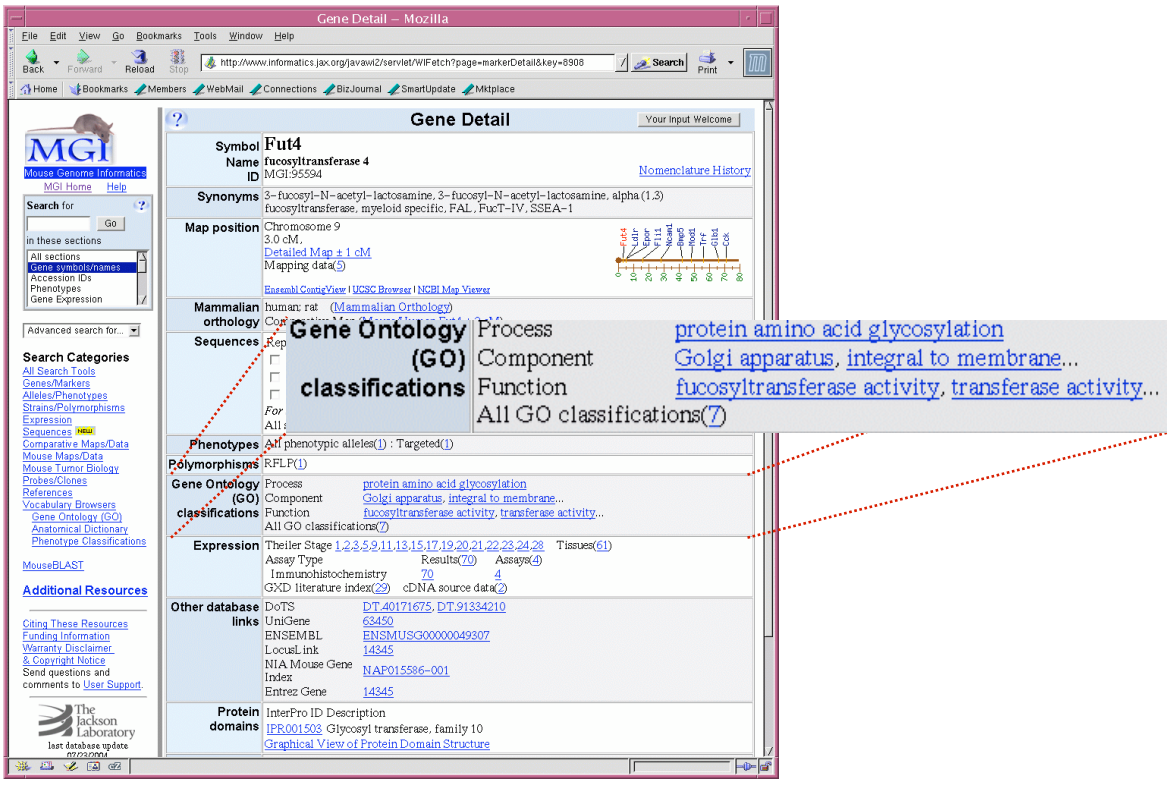
    What is known about protein X (subcellular & tissue localization, associations with diseases, interactions with drugs, …)?
  - assisting scientific discovery by detecting previously unknown relationships, annotating experimental data

The Gene Ontology

- a controlled vocabulary of more than 17,600 concepts describing molecular functions, biological processes, and cellular components



Aiding Annotation: MGI Example



Annotating Genomes: MGI Example

- the current method for this annotation process…

How Do We Get IE Models?

1. encode them by hand

2. learn them from training data

Some Biomedical Named Entity Types

- genes
- proteins
- RNAs
- cell lines/types
- cell components
- diseases/disorders
- drugs
- chromosomal locations

Why Named Entity Recognition is Hard

- these are all gene names

  CAT1
  lacZ
  3-fucosyl-N-acetyl-lactosamine
  MAP kinase
  mitogen activated protein kinase
  mitogen activated protein kinase kinase
  mitogen activated protein kinase kinase kinase
  hairless
  sonic hedgehog
  And

- in some contexts these names refer to the *gene*, in other contexts they refer to the *protein* product, in other contexts its ambiguous

Why Named Entity Recognition is Hard

- they may be referenced conjunctions and disjunctions
  human B- or T-cell lines ⇒
  human B-cell line human T-cell line

- these all refer to the same thing
  NF-kappaB
  NF KappaB
  NF-kappa B
  (NF)-kappaB

- there may be references to gene/protein families
  OLE1-4 ⇒
  OLE1 OLE2 OLE3 OLE4

Sources of Evidence for Biomedical NER

- *orthographic/morphological*: spelling, punctuation, capitalization
  e.g. alphanumeric? contains dashes? capitalized? ends in "ase"
  Src, SH3, p54, SAP, hexokinase

- *lexical*: specific words and word classes
  ___ kinase, ___ receptor, ___ factor

- *syntactic*: how words are composed into grammatical units
  binds to ___, regulated by ___, ___ phosphorylates

Recognizing Protein Names: A Rule-Based Approach

[Fukuda et al., *PSB* '98]

1. morphological and lexical analysis is used to identify "core terms" (e.g. Src, SH3, p54, SAP) and "feature terms" (e.g. receptor, protein)

   The focal adhesion kinase (FAK) is...

2. lexical and syntactic analysis is used to extend terms into protein names

   The focal adhesion kinase (FAK) is...

Recognizing Protein Names: Morphological Analysis in Fukuda Approach

- make list of candidate terms: words that include upper-case letters, digits, and non-alphanumeric characters
- exclude words with length > 9 consisting of lower-case letters and -'s (e.g. full-length)
- exclude words that indicate units (e.g. aa, bp, nM)
- exclude words that are composed mostly of non-alphanumeric characters (e.g. +/-)

Recognizing Protein Names: Lexical/Syntactic Analysis in Fukuda Approach

- merge adjacent terms

Src SH3 domain ⇒ Src SH3 domain

- merge non-adjacent terms separated only by nouns, adjectives and numerals

Ras guanine nucleotide exchange factor Sos

⇒

Ras guanine nucleotide exchange factor Sos

Recognizing Protein Names: Lexical/Syntactic Analysis in Fukuda Approach

- extend term to include a succeeding upper-case letter or a Greek-letter word

p85 alpha ⇒ p85 alpha

Another Approach: Construct Dictionaries of Protein Terms

[Bunescu et al., *AIM* '05]

Protein name (OD)	*Generalized name (GD)*	*Canonical form (CD)*
interleukin-1 beta	interleukin $\langle n \rangle$ $\langle g \rangle$	interleukin
interferon alpha-D	interferon $\langle g \rangle$ $\langle r \rangle$	interferon
NF-IL6-beta	NF IL $\langle n \rangle$ $\langle g \rangle$	NF IL
TR2	TR $\langle n \rangle$	TR
NF-kappa B	NF $\langle g \rangle$ $\langle r \rangle$	NF

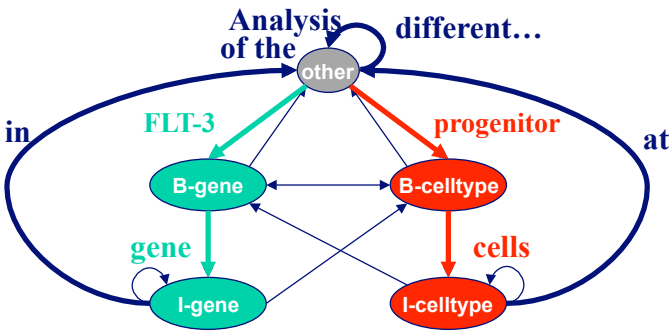
- *original dictionary*: extracted 42,172 gene/protein names from HPI and GO databases
- *generalized dictionary*: replaced numbers with <n> , Roman letters with <r> , Greek letters with <g>
- *canonical dictionary*: stripped generic tags from generalized dictionary entries

NER Results from Bunescu et al.

Table 1 Performance of protein taggers in various settings

IE methods and additional information used	Precision(%)	Recall(%)	*F*-measure(%)
Dictionary-based			
Original dictionary	56.70	27.24	36.80
Plus generalized dictionary	62.27	45.85	**52.81**
Plus canonical dictionary	41.88	54.42	47.33
RAPIER			
Words only	76.11	9.97	17.63
Part-of-speech	70.84	11.05	19.12
Dictionary-based tagger	74.49	12.22	**21.00**
BWI (300 iterations, 2 lookaheads, max. recall)			
Words only	70.67	11.52	19.81
Dictionary-based tagger	71.01	24.06	**35.94**
k-NN ($k = 1, N = 2$)			
Part-of-speech	34.66	40.66	37.42
Dictionary-based tagger	47.30	47.82	**47.56**
TBL			
Words only	47.08	36.65	41.22
Dictionary-based tagger	56.80	34.62	**43.02**
SVM ($N = 2$, full training set, max. recall)			
Preceding class labels	69.16	19.74	30.72
Preceding class labels and part-of-speech	70.18	19.72	30.79
Preceding class labels and dictionary-based tagger	65.00	45.43	53.48
with additional suffix features	70.38	44.49	**54.42**
MaxEnt ($N = 1$, Viterbi w/o greedy extraction, max. recall)			
W/o dictionary	71.10	42.31	53.05
With dictionary	73.37	47.76	**57.86**
With dictionary, two tags only (I,O)	66.41	44.74	53.46
KEX	14.68	31.83	**20.09**
ABGENE	32.39	45.87	**37.97**

NER with a Probabilistic Sequence Model



"Analysis of the FLT-3 gene in progenitor cells at different…"

Features for NER

- in addition to the words themselves, we may want to use other features to characterize the sequence

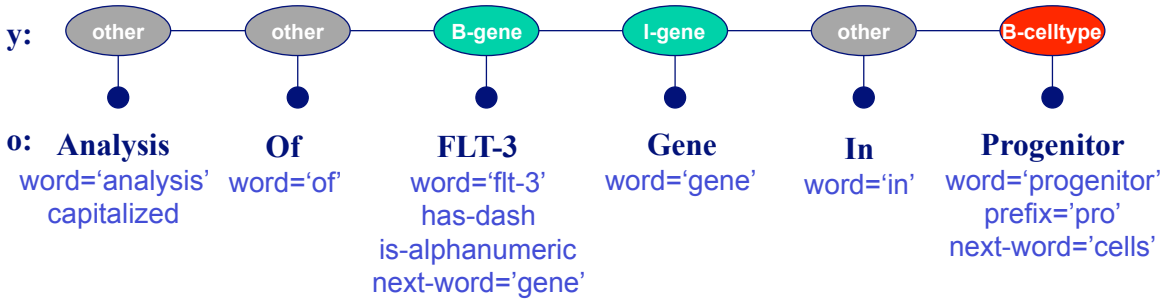
Table 4.3 Some features that have been used in learned models for the biomedical NER task. The left column lists various types of features, the middle column lists specific instances of each type, and the right column lists tokens that match each instance.

type	*example*	*example matching token*
word	word=mitogen?	mitogen
orthographic	is-alphanumeric?	SH3
	has-dash?	interleukin-1
shape	AA0	SH3
	A_aaaaa	F-actin
substring	suffix=ase?	kinase
lexical	is-amino-acid?	leucine
	is-Greek-letter?	alpha
	is-Roman-numeral?	II
part-of-speech	is-noun?	membrane

Conditional Random Fields for NER
[Lafferty et al., 2001]

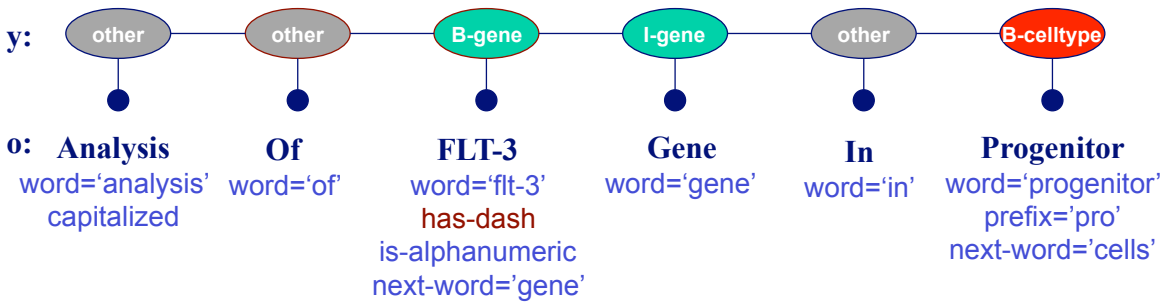
- first-order CRFs define conditional probability of label sequence **y** given input sequence **o** to be:

$$P(\mathbf{y} \mid \mathbf{o}) = \frac{1}{Z_{\mathbf{o}}} \exp\left(\sum_{i=1}^{L} \sum_{k=1}^{F} \lambda_k f_k(y_{i-1}, y_i, \mathbf{o}_i) \right)$$



Conditional Random Fields for NER

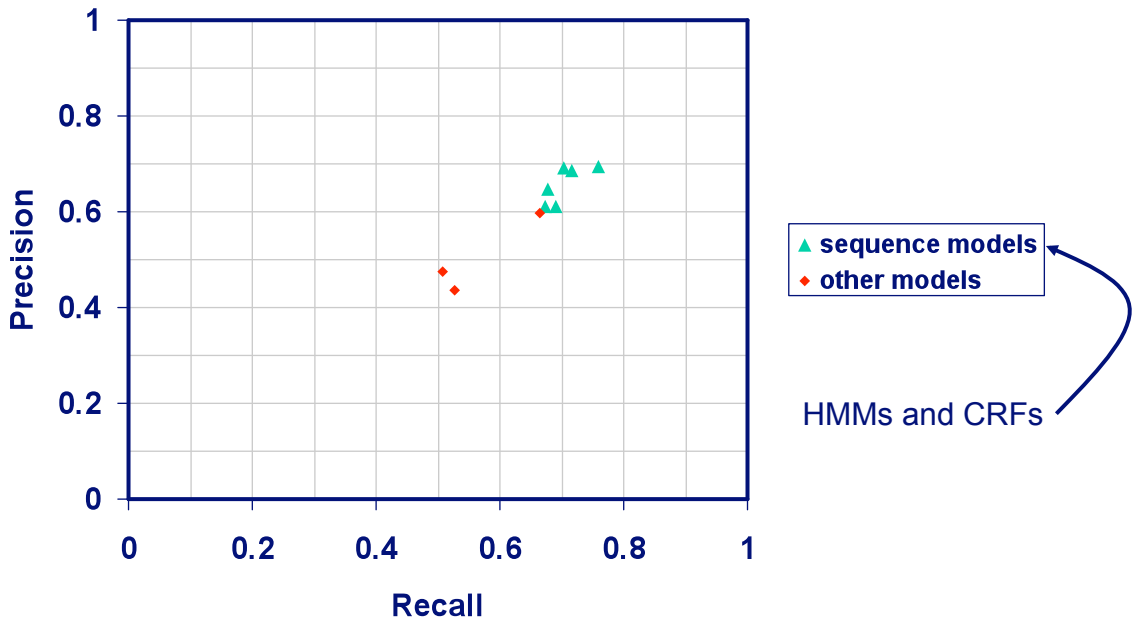
$$P(\mathbf{y} \mid \mathbf{o}) = \frac{1}{Z_{\mathbf{o}}} \exp\left(\sum_{i=1}^{L} \sum_{k=1}^{F} \lambda_k f_k(y_{i-1}, y_i, \mathbf{o}_i) \right)$$



- an example feature: $f_k(\text{other, B-DNA, has-dash})$

Comparison of NER Systems

NLPBA Workshop (COLING 2004)



Comparison of NER Systems

BioCreative Workshop (*BMC Bioinformatics* 2005)