Integrated Annotation of Biomedical Text: Creating the PennBioIE Corpus

Mark A. Mandel
Institute for Research in Cognitive Science
Linguistic Data Consortium
University of Pennsylvania
PennBioIE: Mining the Bibliome

http://bioie.ldc.upenn.edu/

• qualitatively better methods for automatically extracting information from the biomedical literature

• new general methods for information extraction from text
Goal:
automated extraction of relations

INPUT (text)
Amiodarone weakly inhibited CYP2C9, CYP2D6, and CYP3A4-mediated activities with Ki values of 45.1--271.6 μM.

OUTPUT (database entries)

<table>
<thead>
<tr>
<th>Substance</th>
<th>Enzyme inhibited</th>
<th>Measure of inhibition</th>
</tr>
</thead>
<tbody>
<tr>
<td>amiodarone</td>
<td>CYP2C9</td>
<td>Ki=45.1--271.6</td>
</tr>
<tr>
<td>amiodarone</td>
<td>CYP2D6</td>
<td>Ki=45.1--271.6</td>
</tr>
<tr>
<td>amiodarone</td>
<td>CYP3A4</td>
<td>Ki=45.1--271.6</td>
</tr>
</tbody>
</table>

[from the PennBioIE website, http://bioie.ldc.upenn.edu/ "Mining the Bibliome"]
(From the Call for Papers)

• What hurdles must be overcome in performing linguistic analysis of biological text?

• Can you build a generic system and just “drop in” a biological lexicon?
Types of annotation

- paragraphs, sentences, and tokens (pretagging)
- part of speech
- named entities
- syntactic structure (treebanking)
  - about 28%
- relations between entities
  - testing and planning
- database of propositions (propbank)
  - future
Domains

- **CYP**
  - inhibition of cytochrome P-450 enzymes
  - 1100 abstracts
  - GSK

- **Oncology ("onco")**
  - molecular genetics of cancer
  - 1157 abstracts
  - eGenome group, Children's Hospital of Philadelphia
K-ras mutations in sinonasal adenocarcinomas in patients occupationally exposed to wood or leather dust.

Saber AT, Nielsen LR, Dictor M, Hagmar L, Mikoczy Z, Wallin H.

National Institute of Occupational Health, Copenhagen, Denmark.

Of 39 males diagnosed with sinonasal adenocarcinomas over 30 years in the Lund University Hospital catchment area (1.5 million inhabitants), archival tumor tissue was available from 29. Of these, 16 had been exposed to wood dust and three had been exposed to leather dust. The intestinal-type and papillary adenocarcinomas were more common in the exposed patients (P = 0.0002, Fisher's exact test). The tumors from all but one of the 29 sinonasal adenocarcinomas could be analyzed for point mutations at codons 12, 13 and 61 of the K-ras gene. Four mutations were detected in the 28 tumors. The three mutations in the patients exposed to wood and leather dust were all G:C --> A:T transitions, with two at position 2 of codon 12 and one at position 2 of codon 13. The high proportion of G:C --> A:T mutations in this rare tumor may reflect a genotoxic agent in wood and leather dust.

PMID: 9563649 [PubMed - indexed for MEDLINE]
K-ras mutations in sinonasal adenocarcinomas in patients occupationally exposed to wood or leather dust.

Saber AT, Nielsen LR, Dictor M, Hagmar L, Mikoczy Z, Wallin H.

National Institute of Occupational Health, Copenhagen, Denmark.

Of 39 males diagnosed with sinonasal adenocarcinomas over 30 years in the Lund University Hospital catchment area (1.5 million inhabitants), archival tumor tissue was available from 29. Of these, 16 had been exposed to wood dust and three had been exposed to leather dust. The intestinal-type and papillary adenocarcinomas were more common in the exposed patients (P = 0.0002, Fisher's exact test). The tumors from all but one of the 29 sinonasal adenocarcinomas could be analyzed for point mutations at codons 12, 13 and 61 of the K-ras gene. Four mutations were detected in the 28 tumors. The three mutations in the patients exposed to wood and leather dust were all G:C --> A:T transitions, with two at position 2 of codon 12 and one at position 2 of codon 13. The high proportion of G:C --> A:T mutations in this rare tumor may reflect a genotoxic agent in wood and leather dust.

PMID: 9563649 [PubMed - indexed for MEDLINE]
K-ras mutations in sinonasal adenocarcinomas in patients occupationally exposed to wood or leather dust.

Saber AT, Nielsen LR, Dictor M, Hagmar L, Mikoczy Z, Wallin H.

National Institute of Occupational Health, Copenhagen, Denmark.

Of 39 males diagnosed with sinonasal adenocarcinomas over 30 years in the Lund University Hospital catchment area (1.5 million inhabitants), archival tumor tissue was available from 29. Of these, 16 had been exposed to wood dust and three had been exposed to leather dust. The intestinal-type and papillary adenocarcinomas were more common in the exposed patients (P = 0.0002, Fisher's exact test). The tumors from all but one of the 29 sinonasal adenocarcinomas could be analyzed for point mutations at codons 12, 13 and 61 of the K-ras gene. Four mutations were detected in the 28 tumors. The three mutations in the patients exposed to wood and leather dust were all G:C --> A:T transitions, with two at position 2 of codon 12 and one at position 2 of codon 13. The high proportion of G:C --> A:T mutations in this rare tumor may reflect a genotoxic agent in wood and leather dust.

PMID: 9563649 [PubMed - indexed for MEDLINE]
K-ras mutations in sinonasal adenocarcinomas in patients occupationally exposed to wood or leather dust.

Saber AT, Nielsen LR, Dictor M, Hagmar L, Mikoczy Z, Wallin H.

National Institute of Occupational Health, Copenhagen, Denmark.

Of 39 males diagnosed with sinonasal adenocarcinomas over 30 years in the Lund University Hospital catchment area (1.5 million inhabitants), archival tumor tissue was available from 29. Of these, 16 had been exposed to wood dust and three had been exposed to leather dust. The intestinal-type and papillary adenocarcinomas were more common in the exposed patients (P = 0.0002, Fisher's exact test). The tumors from all but one of the 29 sinonasal adenocarcinomas could be analyzed for point mutations at codons 12, 13 and 61 of the K-ras gene. Four mutations were detected in the 28 tumors. The three mutations in the patients exposed to wood and leather dust were all G:C --> A:T transitions, with two at position 2 of codon 12 and one at position 2 of codon 13. The high proportion of G:C --> A:T mutations in this rare tumor may reflect a genotoxic agent in wood and leather dust.

PMID: 9563649 [PubMed - indexed for MEDLINE]
Sequence of annotations
(original)

1. pretagging
2. POS
3. named entities
4. treebanking
5. relations
Sequence of annotations (present)

1. pretagging
   POS
2. named entities
3. POS
4. treebanking
5. relations
Embedded tags (1)
Avoided by default

Default: No embedded tags.
Tag the outermost mention only.
Embedded tags (2)

Avoidance

**Ewing's sarcoma** gene

Gene named for malignancy.

*Ewing's sarcoma* is usually tagged as Malignancy-type, but here it is included in the Gene/RNA mention and not tagged separately.

**ras** signal transduction mediators

Class of proteins whose name includes the name of a Gene/RNA class (**ras**).

Entire mention tagged only as Gene-Protein.
Embedded tags (3)
Our favorite horrible example

\begin{center}
\begin{tabular}{|c|c|c|c|c|}
\hline
\texttt{tyrosine kinase} & \texttt{kinase} & \texttt{kinase} & \texttt{kinase} & \ldots \\
\hline
\end{tabular}
\end{center}
Embedded tags (4)
Use in specific situations

Malignancy-histology within Malignancy-type

Malignancy-type

Malignancy-histology

chronic myelogenous leukemia

This is fairly common in Malignancy-type (names of malignancies).
Discontinuous mentions (1)
Existing practice (context of guideline)

(from current Penn “Simple Named Entity Guidelines”)

When a phrase refers to multiple named entities, mark each entity separately. For instance, this sentence contains two entities:

• [China] and [South Korea] signed the agreement.
Discontinuous mentions (2)
Existing practice (example)

When a phrase refers to multiple named entities, mark each entity separately. For instance, this sentence contains two entities:

- [China] and [South Korea] signed the agreement.

Similarly,

- [Jimmy] and [Rosalyn Carter]
- [North] and [South America]

\[
\begin{align*}
\text{Jimmy} & = \text{“Jimmy Carter”} \\
\text{North} & = \text{“North America”}
\end{align*}
\]
Discontinuous mentions (3)  
Existing practice (apparent consequence)

1. [North] and [South America]  
   North = “North America”

2. [North] and [South Korea]  
   North = “North Korea”

3. Therefore...  
   “North America” = “North Korea” ????
Discontinuous mentions (4)

PennBioIE Chaining

androgen and estrogen receptors

- “androgen receptors”
- “estrogen receptors”

codons 12, 13, and 61

- “codon 12”
- “codon 13”
- “codon 61”

*(We ignore the singular/plural distinction in order to capture the mention.)*
Discontinuous mentions (5)
Chaining (CYP domain)

**androgen** and **estrogen receptors**

---

Substance (chain)

**androgen** and **estrogen receptors**
Discontinuous mentions (6)
Chaining (onco domain)

codons 12, 13, and 61

Var-location

Var-location (chain) Var-location (chain)

Var-location

codons 12, 13, and 61
Discontinuous mentions (7)
Limitations on chaining

To stay compatible with treebanking:

- Conjunctive constructions with shared component
  - androgen and estrogen receptors
  - P450IA1 or IA2
  - CYP1A1/2
  - organic and inorganic acids and salts
    - organic acids inorganic acids
    - organic salts inorganic salts

- A few other specific notational constructions
“Fruit salad” (1)  
(terms containing characters 
not normally found in lexical items)

Examples

(Na+ + K+)ATPase

2,3,7, 8-tetrachlorodibenzo-p-dioxin

2-(4-acetoxyphenyl)-2-chloro N-methyl-ethylammonium

2-amino-6-methylidipyrido(1,2-a:3',2'-d)imidazole

2,2',4,5,5'-Cl5

1, 2-bis (o-aminophenoxy) ethane N, N, N', N'-
tetraacetic acid tetra(acetomethoxyl) ester
“Fruit salad” (2)

POS treatment

Noun (the default for fruit salad)

- EC(50)
- \((\text{Na}^+ + \text{K}^+)\text{ATPase}\)
- 2,3,7,8-tetrachlorodibenzo-p-dioxin
- 2-(4-acetoxyphenyl)-2-chloro N-methyl-ethylanmonium
- 2-amino-6-methyl[dipyrirdo(1,2-a:3',2'-d)imidazole
- 2,2',4,5,5'-Cl5
Institute for Research in Cognitive Science
University of Pennsylvania

"Fruit salad" (3)

POS treatment

Adjective

1, 2-bis (o-aminophenoxy) ethane N, N, N', N'-tetraacetic acid tetra(acetomethoxyl) ester

• 1, 2-bis (o-aminophenoxy) ethane N, N, N', N'-tetraacetic  JJ (adjective)
• acid  noun
• tetra(acetomethoxyl)  noun
• ester  noun
Symbols 1: Penn Treebank SYM

Penn Treebank definition (Santorini 1990):

This tag should be used for mathematical, scientific and technical symbols or expressions that aren’t words of English. It should not used for any and all technical expressions. For instance, the names of chemicals, units of measurements (including abbreviations thereof) and the like should be tagged as nouns.
Symbols 2:
SYM

• Individual keyboard symbols (*same as Santorini*):
  + * / < > =

• Composite symbols (*undefined*):
  +/−
  −→ ←− −−− −−−− etc.

• Names of Greek letters (*undefined*):
  alpha beta ...
Symbols 3: SYM

Mangled arrows (examples represent $\text{gly} \rightarrow \text{val}$):

- $\text{gly - greater than} \text{ val}$
- $\text{gly - &gt; val}$
- $\text{gly - gt val}$
- $\text{gly - &gt;#62; val}$
- ... 

(Different from Santorini)
Symbols 4: not SYM

- % "percent" N (noun) (same?)
- & "and" CC (coord. conjunction) (same?)
- pH "pH" N (noun) (different?)
- DNA "DNA" N (noun) (different?)
- plus CC (coord. conjunction) (same)
- multiplied by (same)
  - multiplied VBN (verb, past participle)
  - by IN (preposition)
Symbols 5:
±SYM, context-dependent

• – minus sign SYM (same)
  $-4^\circ \text{C}$
  $12 - 4 = 8$

• – hyphen HYPH (different)
  concentration-response
  Ki-, Ha-, and N-ras

• – range indicator HYPH (different)
  5–7.5 ml
AFX and HYPH (1)

Text

anti-CYP2E1-IgG

mono- and diglycerides
AFX and HYPH (2)

Text → entities

\[\text{Substance} \rightarrow \text{anti-CYP2E1-IgG} \]

\[\text{Substance (chain)} \rightarrow \text{mono- and diglycerides} \]
Entities → tokens

- **anti-CYP2E1-IgG**
  - Substance
  - ENTITY
  - text
  - TOKEN

- **mono- and diglycerides**
  - Substance (chain)
  - Substance
  - ENTITY
  - text
  - TOKEN
AFX and HYPH (4)
Tokens

anti-CYP2E1-IgG

mono- and diglycerides
AFX and HYPH (5)

Tokens → new tags

anti-CYP2E1-IgG

HYPH

AFX NN(noun)

mono- and diglycerides

HYPH CC

AFX AFX NNS(plural noun)

POS
False breaks in text (1)
As downloaded


[from source_file_2264_35216.src PMID: 2541801]

(The strings in grey -- RIII, RIV, RV -- are identifiers used for these compounds, not parts of the terms themselves.)
False breaks in text (2)
Artificial white space identified

- : Artificial space character
◆ : Artificial line break


Entity references (all tagged as Substance):
- 2-octyl-4-(3-iodine-2-oxopropylidene)-2,3,5,5-tetramethylimidazolidine-1-oxyl
- 2-nonyl-4-(3-iodine-2-oxopropylidene)-2,3,5,5-tetramethylimidazolidine-1-oxyl
- 2-heptadecyl-4-(3-iodine-2-oxopropylidene)-2,3,5,5-tetramethylimidazolidine-1-oxyl
Entity Types (1)
CYP

Original: 3
- CYP enzyme
- Chemical
- Process

Present: 5
- CYP enzyme
- (Other) substance
- Quantity
  - Q-name
  - Q-value
  - Q-unit
Entity Types (2)

Onco

Original: 3
- Gene
- Variation
- Malignancy

Present: 24
- Gene: 3
- Variation: 6
- Malignancy: 9
- Quantity: 6
Entity Types (3)
Onco present

Gene
• gene or RNA (genomic)
• protein (proteomic)
• generic (vague or inclusive)

Variation
• type (*deletion*)
• location (*codon 15*)
• states (*Glycine; Gly; G*)
  – initial state
  – altered state
  – generic state
• event [for use in relations]
Entity Types (4)
Onco present
Malignancy

- type (*neuroblastoma*)
- developmental state (*pediatric*)
- clinical stage (*Stage 3*)
- histology (*cardiac*)
- site (*colon*)
- differentiation (*well-differentiated*)
- heredity status (*familial*)
- survival status (*event-free survival*)
- survival status modifier (*without progression or relapse*)
Entity Types (5)

Onco present

Quantity

- count \((15; \text{two})\)
- proportion \((17 \text{ of } 20; \ 82\%)\)
- time \((3.8 \text{ sec; October 1994})\)
- measurement (any other use of a number)
- quantitative classifier \((\text{one year old}, \text{one year of age})\)
- statistical modifier \((\text{average size, median follow-up})\)
• **Fable**

Fast Automated Biomedical Literature Extraction

• FABLE allows a biomedical researcher to query a version of MEDLINE that has been annotated with our text-mining tools. Type a human gene or protein name into the search box, choose search options, and click submit. The result will list MEDLINE articles mentioning this gene.

• Release target April 3

• In betatest *this week* at [http://fable.chop.edu/](http://fable.chop.edu/)

**TRY IT!**
(From the Call for Papers)

• What hurdles must be overcome in performing linguistic analysis of biological text?

• Can you build a generic system and just “drop in” a biological lexicon?
Lessons learned

War

No battle plan survives first contact with the enemy.

Helmuth von Moltke (1800-1891)
Lessons learned
Annotation

No annotation plan survives first contact with the data.
Partial list of contributors to creating the corpus
And see http://bioie.ldc.upenn.edu/index.jsp?page=aboutus.html

Children’s Hospital of Philadelphia
Yang Jin
Jessica Kim
Peter White
Scott Winters

University of Pennsylvania
Students, Postdocs, Staff
Ann Bies
Hubert Jin
Seth Kulick
Dalal Zakhary
Ramez Zakhary

GSK
James Butler
Harry Gottlieb
Paula Matuszek

Programmer Analysts
Jeremy LaCivita
Tom Morton
Eric Pancoast

All the annotators, past and present
Contributors to PennBioIE

Institute for Research in Cognitive Science
University of Pennsylvania

CHOP
Yang Jin
Jessica Kim
Pete White
Scott Winters

UNIVERSITY OF PENNSYLVANIA

FACULTY AND STAFF

Ann Bies
Susan Davidson
Hubert Jin
Ara Vind Joshi
Seth Kulick
Jeremy LaCivita
Mark Liberman
Mitch Marcus
Ryan McDonald
Tom Morton
Martha Palmer
Eric Pancoast
Michael Patek

Fernando Pereira
Ted Sandler
Andrew Schein
Mike Schultz
Rishi Taileja
Partha Talukdar
Val Tannen
Ryan Tracy
Lyle Ungar
Dalal Zakzaky
Ramez Zakzaky

UNIVERSITY OF PENNSYLVANIA

ANNOTATORS
italic: active annotators

Sanipa Arnold
Rachel Barreto
Jee Bang
Avik Basu
Christine Brisson
Dan Caroff
Hareesh Chandrupatla
Dhinakaran Chinappen
Melissa Demian
Jacqueline Ewing
Amy Felix
Nadeene Francisco
Benjamin George
Ari Goldberg
Brian Golden
Robin Golden
Kaira Gui
Karen Jablonski
Justin Lacasse
Matt Leger
Alexis Lerro
Mark Manocchio
Marty McComick
Brett Merves

Brad Moatz
Grace Mrowicki
Sina Neshatian
Ben Newman
Michael Noda
Jesse Palma
Anita Patel
Ariel Richmond
Karen Rudo
Jonathan Schwartz
Amanda van Scoyoc
Nilay Shah
Rabiya Sheikh
Sarah Stippich
Sabrina Sumner
Rachel Swetz
Sophia Varghese
Julie Wang
Peng Wang
Colin Warner
Lakiya Wimbish
Christopher Wright
Johanna Wright

Text Mining, Ontologies and Natural Language Processing in Biomedicine, Manchester, UK 2006.03.20-21
Integrated Annotation of Biomedical Text: Creating the PennBioIE Corpus -- Mark A. Mandel
http://bioie.ldc.upenn.edu/
Thank you