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)

Substance	Enzyme inhibited	Measure of inhibition
amiodarone	CYP2C9	Ki=45.1--271.6
amiodarone	CYP2D6	Ki=45.1--271.6
amiodarone	CYP3A4	Ki=45.1--271.6

[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

Downloaded abstract (1)

1: Cancer Lett. 1998 Apr 10;126(1):59-65.

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]

Downloaded abstract (2)

1: Cancer Lett. 1998 Apr 10;126(1):59-65.

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]

Downloaded abstract (3)

1: Cancer Lett. 1998 Apr 10;126(1):59-65.

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]

Biomedical text: Title and Body only

Cancer Lett. 1998 Apr 10;126(1):59-65.

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.

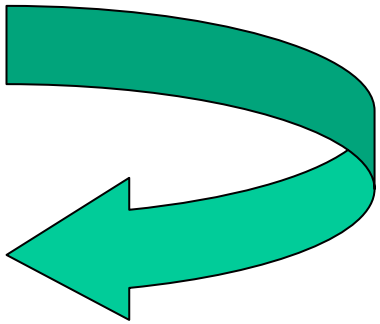
PMID: 9563649 [PubMed - indexed for MEDLINE]

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

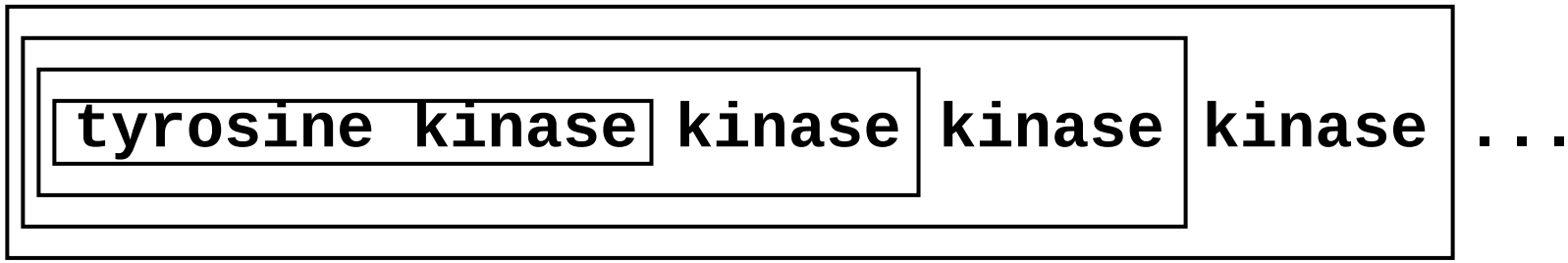
`Ewing's sarcoma` gene

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



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]**

Jimmy	=	"Jimmy Carter"
North	=	"North America"

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"

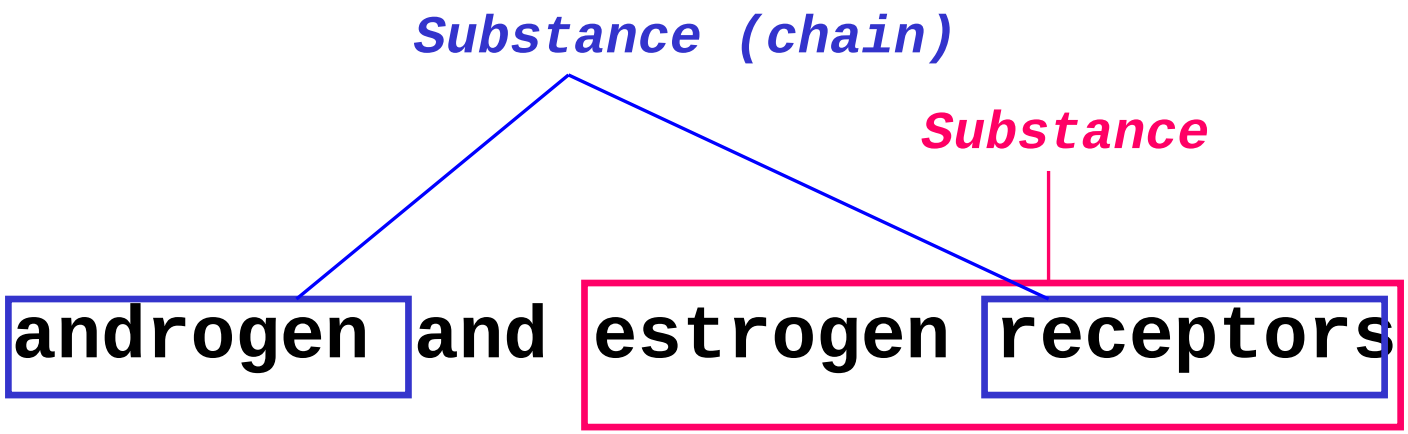
`codons 12, 13, and 61`

- "codon 12"
- "codon 13"
- "codon 61"

(We ignore the singular/plural distinction in order to capture the mention.)

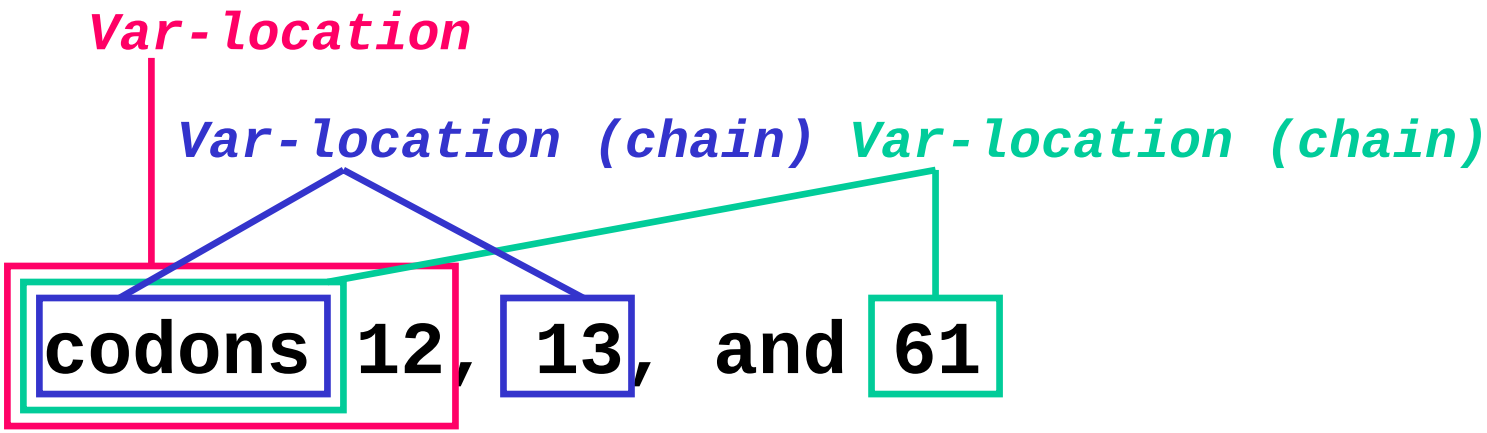
Discontinuous mentions (5) Chaining (CYP domain)

androgen and estrogen receptors



Discontinuous mentions (6) Chaining (onco domain)

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-methyldipyrido(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)**
- **(Na+ + K+)ATPase**
- **2,3,7, 8-tetrachlorodibenzo-p-dioxin**
- **2-(4-acetoxyphenyl)-2-chloro N-methyl-ethylammonium**
- **2-amino-6-methyldipyrido(1,2-a:3',2'-d)imidazole**
- **2,2',4,5,5'-Cl5**

"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 **`gly -> val`**):

- **`gly - greater than val`**
- **`gly - > val`**
- **`gly - gt val`**
- **`gly - > 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° C

  12 - 4 = 8

- **-** hyphen — HYPH *(different)*

  concentration-response

  Ki-, Ha-, and N-ras

- **-** range indicator — HYPH *(different)*

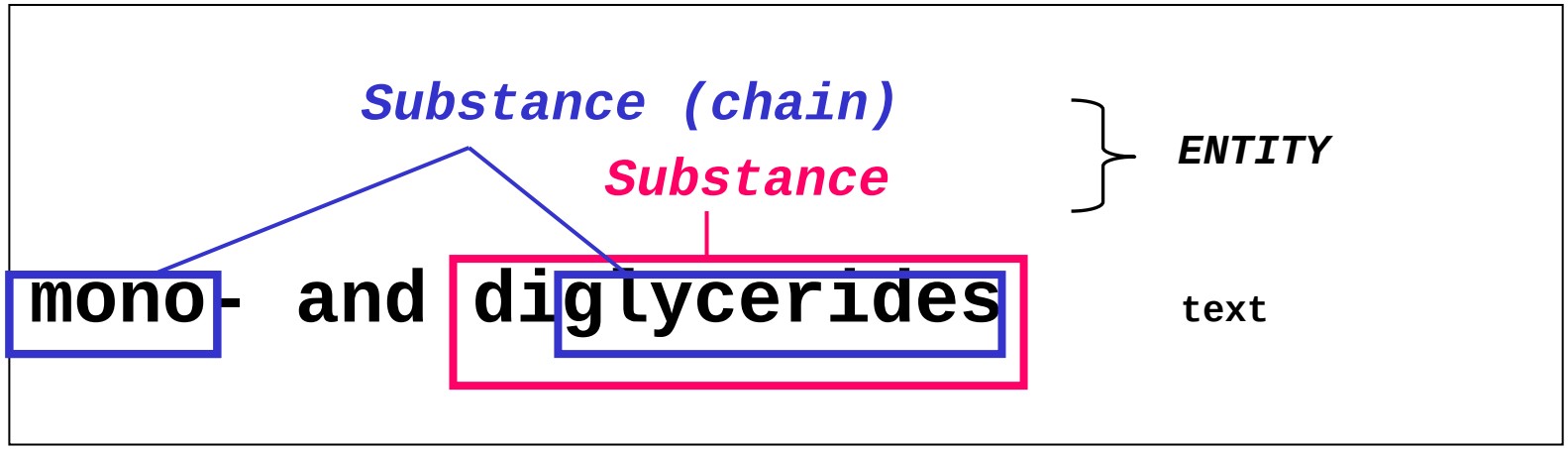
  5-7.5 ml

AFX and HYPH (1)
Text

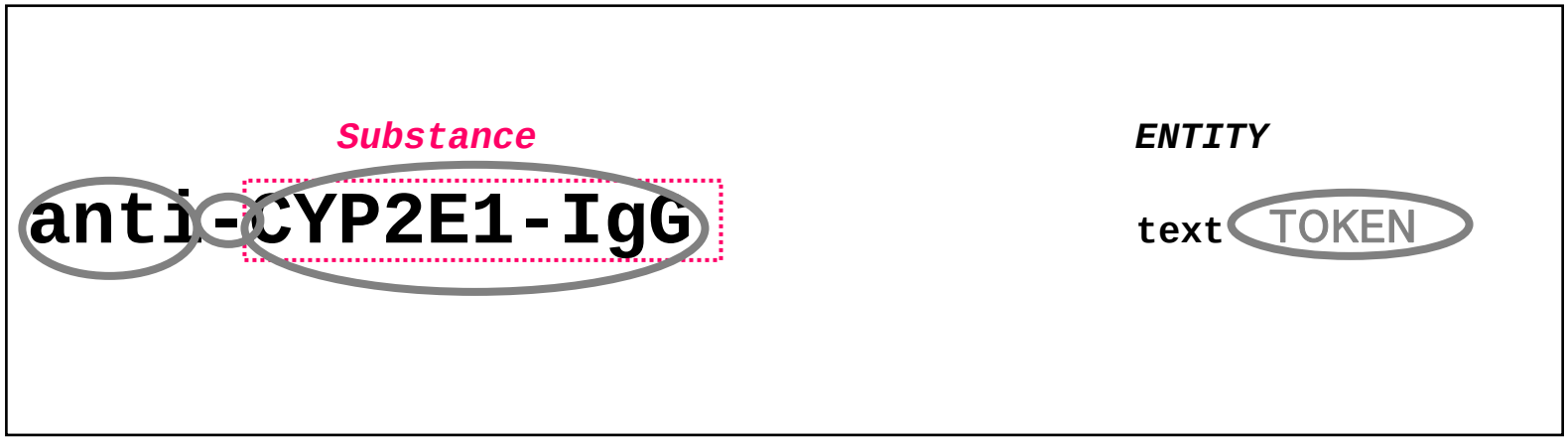
anti-CYP2E1-IgG **text**

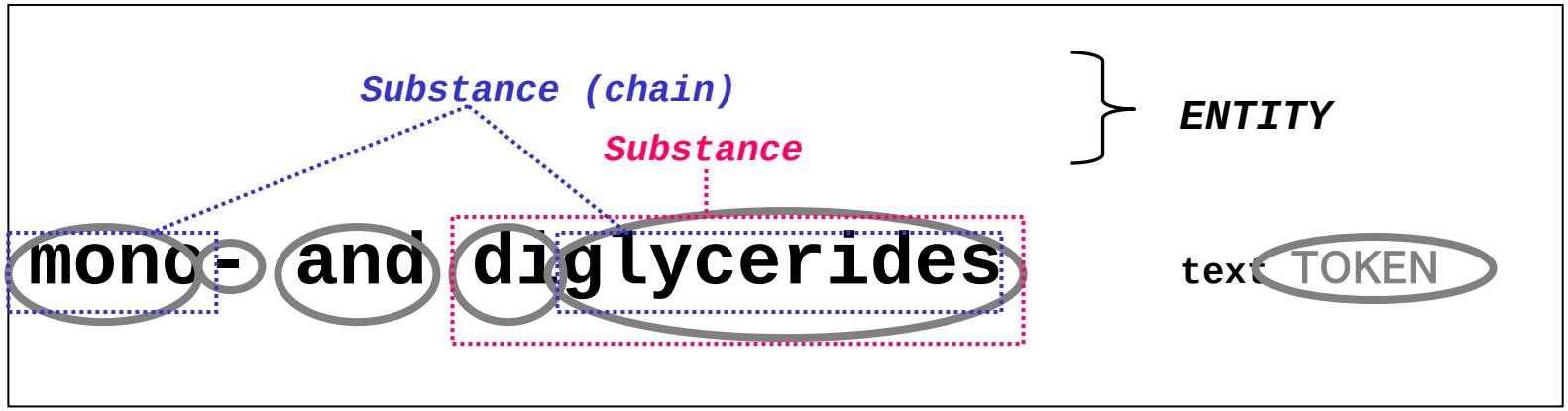
mono- and diglycerides **text**

AFX and HYPH (2)
Text → entities



AFX and HYPH (3)
Entities → tokens



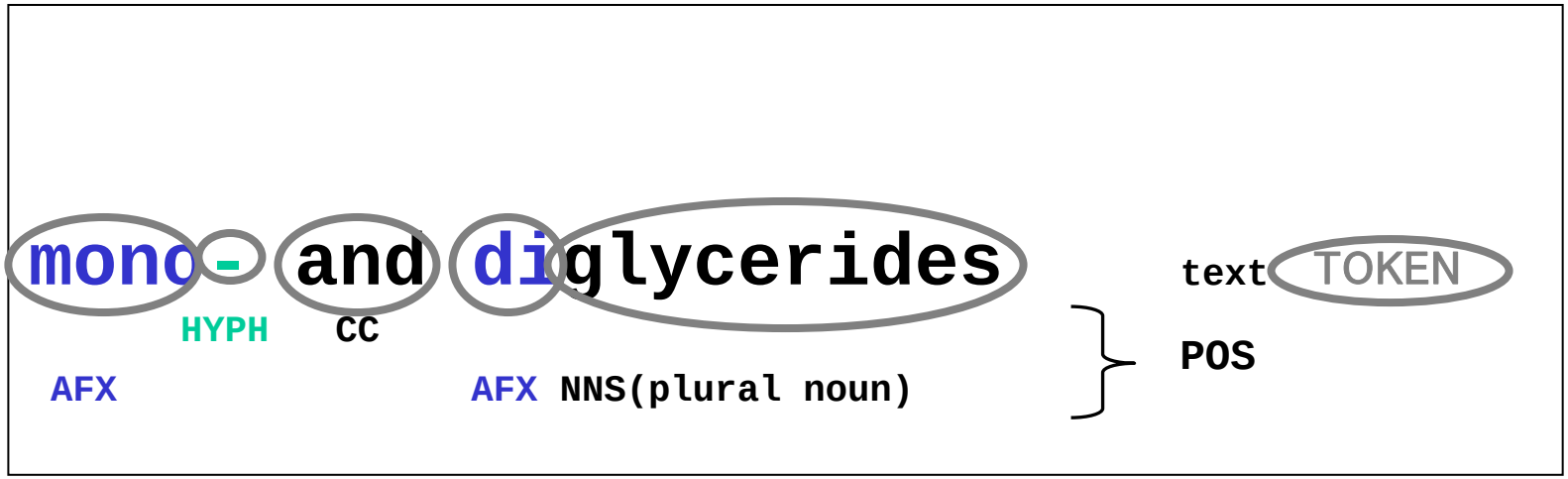


AFX and HYPH (4) Tokens

anti-CYP2E1-IgG text TOKEN

monc- and diglycerides text TOKEN

AFX and HYPH (5)
Tokens → new tags



False breaks in text (1)
As downloaded

```
2-octyl-4-(3-iodine-2-oxopropylidene)-2,3,5,5-tetramethylimidaz olidine-1-oxyl
(RIII), 2-nonyl-4-(3-iodine-2-oxopropylidene)-2,3,5,5-tetramethylimidaz
olidine-1-oxyl (RIV),
2-hepta-decyl-4-(3-iodine-2-oxopropylidene)-2,3,5,5-tetramethyl imidazolidine-1-
oxyl (RV).
```

[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)
Artifactual white space identified

- ◘ : Artifactual space character
- ◆ : Artifactual line break

2-octyl-4-(3-iodine-2-oxopropylidene)-2,3,5,5-tetramethylimidaz◘olidine-1-oxyl (RIII), 2-nonyl-4-(3-iodine-2-oxopropylidene)-2,3,5,5-tetramethylimidaz◆olidine-1-oxyl (RIV), 2-hepta-decyl-4-(3-iodine-2-oxopropylidene)-2,3,5,5-tetramethyl◘imidazolidine-1-◆oxyl (RV).

Entity references (all tagged as Substance):

- **2-octyl-4-(3-iodine-2-oxopropylidene)-2,3,5,5-tetramethylimidaz olidine-1-oxyl**
- **2-nonyl-4-(3-iodine-2-oxopropylidene)-2,3,5,5-tetramethylimidaz olidine-1-oxyl**
- **2-hepta-decyl-4-(3-iodine-2-oxopropylidene)-2,3,5,5-tetramethyl imidazolidine-1- oxyl**

Entity Types (1)
CYP

Original: 3		Present: 5
• CYP enzyme		• CYP enzyme
• Chemical	→	• (Other) substance
• ~~Process~~	+	• Quantity
		– Q-name
		– Q-value
		– Q-unit

Entity Types (2)
Onco

<u>Original: 3</u>

- Gene
- Variation
- Malignancy

<u>Present: 24</u>

- 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; two)*
- proportion *(17 of 20; 82%)*
- time *(3.8 sec; October 1994)*
- measurement (any other use of a number)
- quantitative classifier *(one year <u>old</u>, one year <u>of age</u>)*
- statistical modifier *(<u>average</u> size, <u>median</u> 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 ba, 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/`**

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

GSK

James Butler
Harry Gottlieb
Paula Matuszek

All the annotators, past and present

University of Pennsylvania

Students, Postdocs, Staff

Ann Bies
Hubert Jin
Seth Kulick
Dalal Zakhary
Ramez Zakhary

Programmer Analysts

Jeremy LaCivita
Tom Morton
Eric Pancoast

Contributors to PennBioIE

CHOP

Yang Jin
Jessica Kim
Pete White
Scott Winters

GSK

James Butler
Harry Gottlieb
Paula Matuszek

UNIVERSITY OF PENNSYLVANIA

FACULTY AND STAFF

Ann Bies
Susan Davidson
Hubert Jin
Aravind 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 Talreja
Partha Talukdar
Val Tannen
Ryan Tracy
Lyle Ungar
Dalal Zakhary
Ramez Zakhary

ANNOTATORS

italic: active annotators

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

Thank you