





# Mutation and semantic representations in natural language

K. Bretonnel Cohen

The MITRE Corporation, Human Language Technology  
Department

and

Center for Computational Pharmacology, U. Colorado  
School of Medicine

ECCB 2008 Mutations Workshop

# Why bother with text mining?

## Point mutation E959Q in MCR\_RAT

Point mutation:	E959Q	
Domain:	LBD HELIX 12	
General numbering (NucleaRDB):	1250	
Protein:	MCR_RAT ( <i>NR3C2</i> )	<a href="#">Swiss-Prot</a> <a href="#">Cross-reference table</a> <a href="#">Family page</a>
Other point mutations / same protein	<a href="#">List</a>	
Family alignments	<a href="#">3C2 Mineralocorticoid (MR)</a> <a href="#">3C Glucocorticoid-like (GR,MR,PR,AR)</a> <a href="#">3 Estrogen like (ER,ERR,GR,MR,PR,AR)</a>	
Other point mutations / same position	Position <a href="#">755</a> in 3C Glucocorticoid-like (GR,MR,PR,AR) family Position <a href="#">757</a> in 3 Estrogen like (ER,ERR,GR,MR,PR,AR) family	
Reference:	Characterization of transactivational property and coactivator mediation of rat mineralocorticoid receptor activation function-1 (AF-1). Fuse H Mol Endocrinol 2000 Jun;14(6):889-99.	<a href="#">Medline</a> 3/7
Other point mutations / same article	<a href="#">List</a>	
Text source	HTML and PDF full texts	

### Relevant sentences:

#### E959Q

- The MR mutant (E959Q) with a point mutation in helix 12, which causes a complete loss of MR AF-2 activity, still retained ligand-induced transactivation function, indicating a significant role for AF-1 in the full activity of the ligand-induced MR function
- Because the MR helix 12 contains only one negatively charged amino acid (Glu959) in the conserved amino acid sequence, we displaced this Glu959 into electrically neutral Gln by site-directed mutagenesis (E959Q-mutant as depicted in Fig. 2A(image))

Horn et al.  
(2004)

# Why bother with text mining?

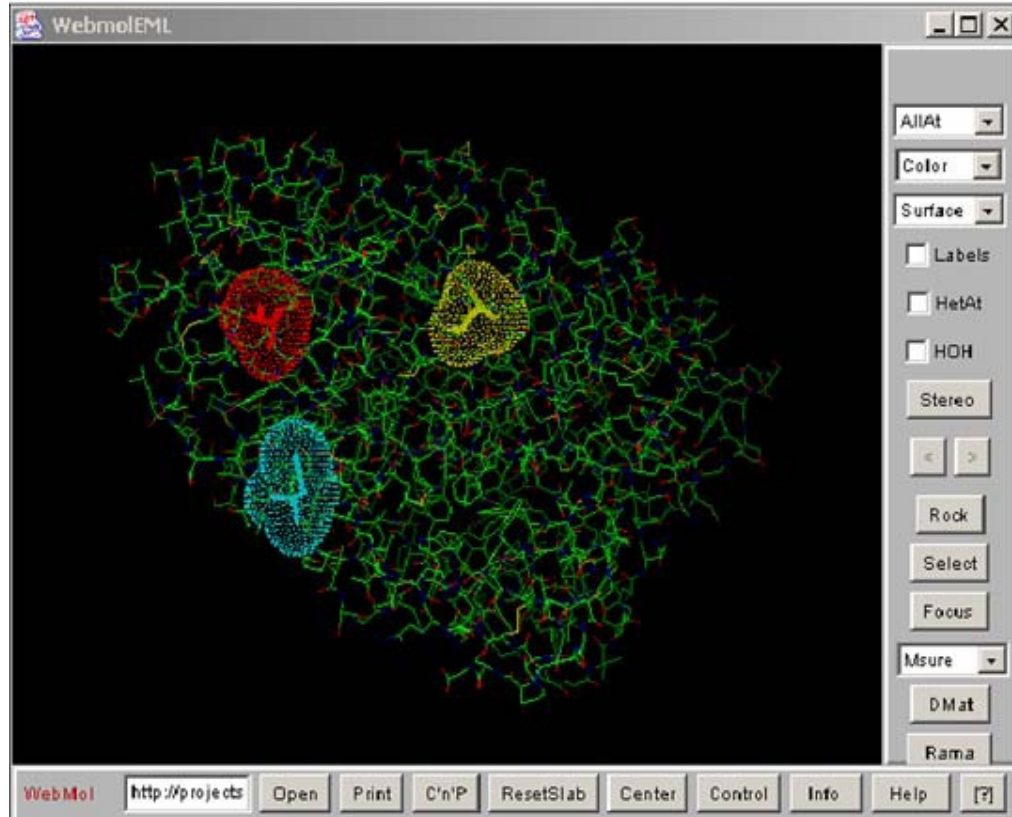
If only *one third* of the **MEMA-only genes** were relevant to **OMIM** (associated with human disease, not experimentally induced, not synonymous variant of **OMIM** pair), **MEMA** could *double* the number of mutation/gene pairs in **OMIM**!

	Mutation/gene pairs
OMIM	6,699
MEMA	20,503
Intersection	1,826
OMIM only	4,873
MEMA only	18,677

Rebholz-Schumann et al.  
(2004)

# Why bother with text mining?

**A**



**B**

```
<menu>
<status=on>
<label>Xylanase</label>
<item>
<range>11:A 11:A</range>
<color=orange>
<status=on>
<label>The mutations N11D and N38E did not have any significant effect:N11D increased the half life scarcely 1.5 times at 55°C, and N38E about 1.5 times at 57-60°C</label>
</item>
<item>
<range>162:A 162:A</range>
<color=purple>
<status=off>
<label>The mutations at the C-terminus of the -helix, Q162H, Q162Y, Q162L and Q162K, increased the half life of XYNII at 55°C (pH 5) to 36,39, 26 and 11 min, respectively. Q162H, Q162Y and Q162L did not show any stabilizing effect at 65°C (half-lives 1 min)
</label>
</item>
<item>
<range>210:A 210:A</range>
<color=yellow>
<status=on>
<label>The glutamic acid residue at position 210, which is part of the active center in this family of enzymes, was changed to either aspartic acid (E210D) or serine (E210S)</label>
</item></menu>
```

MOL\_ID: 1; MOLECULE: EXCELLOBIOHYDROLASE I; CHAIN: A; FRAGMENT: CATALYTIC DOMAIN 1-434;

✓ Xylanase

The mutations N11D and N38E did not have any significant effect N11D increased the half life scarcely 1.5 times at 55°C, (2006) about 1.5 times

Q162H, Q162Y, Q162L and Q162K, increased the half life of XYNII at 55°C (pH 5) to 36,39, 26 and 11 min, respectively. Q162H, Q162Y and Q162L

The glutamic acid residue at position 210, which is part of the active center in this family of enzymes, was changed to either aspartic acid (E210D) or

Baker and Witte

ECCB 2008 Mutations Workshop

# The talk in one slide

- Constructions like *L1014F mutation* are very odd from the point of view of linguistic theory, and theoretical and computational linguistics researchers should pay attention to the literature of mutations.

# Outline of the talk

- Theories of semantic representation in linguistics
- A representation for *mutate*
- The phenomenon of nominalization
- A phenomenon in the scientific literature about mutations and its implications for linguistic theory
- A set of possible conclusions

# Theories of semantic representation in linguistics

- Decompositional (features/events)
- Ontological/relational (between words)
- Argument-centric

# Theories of semantic representation in linguistics

- Decomposition into *features*
  - PATH
  - MANNER
  - PLACE
- E.g. *lose*:
  - GO
  - PATH: away
  - ORIGIN: possessor

# Theories of semantic representation in linguistics

- Decomposition into verbal primitives: BECOME, DO, CAUSE, CHANGE-STATE
- E.g. *kill*:
- CAUSE(BECOME(NOT(alive)))
- Nice fit to patterns in behavior of groups of verbs; language acquisition by children

# Theories of semantic representation in linguistics

## +’s of decompositional representations

- Explains some patterns of verb group behaviors, patterns in child language acquisition

## -’s of decompositional representations

- Don’t scale well across a large lexicon
- Difficult to evaluate

# Theories of semantic representation in linguistics

- Ontological/relational

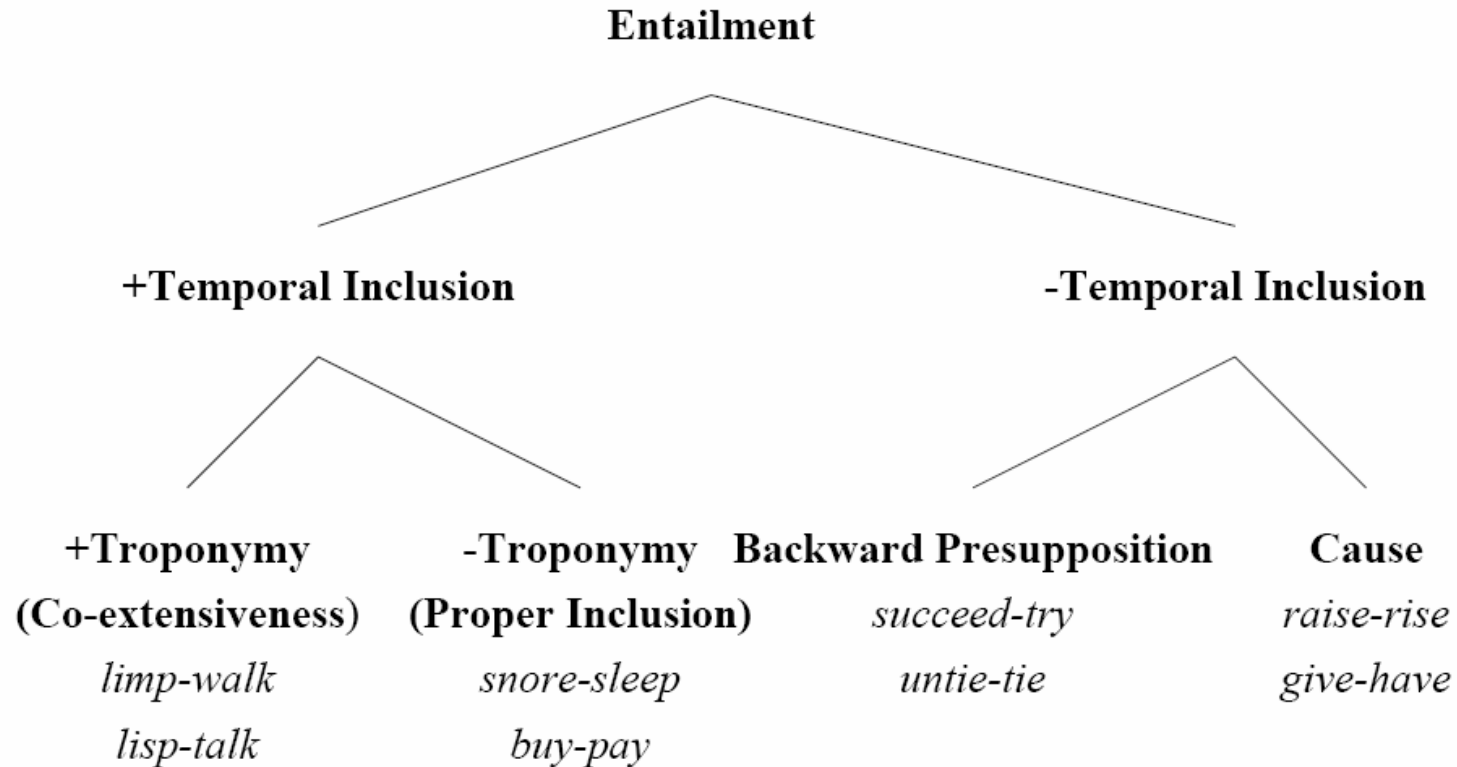


Figure 3. Four kinds of entailment relations among verbs

Fellbaum  
(1999)

# Theories of semantic representation in linguistics

## +’s of ontological/relational

- Psychological reality
- Entailment supports reasoning
- Many positive features of decompositional representations retained, e.g. causality
- Usable in some language processing applications (WordNet)

## -’s of ontological/relational

- Difficult to build on large scale (Poprat et al. 2008)
- Issues of granularity, sense lumping/splitting (Palmer and Dang 2001)

Predicate-argument structure: a  
scheme for representing the  
semantics of a verb

# A sample predicate-argument structure

- truncate.01
  - Arg0: truncator
  - Arg1: entity shortened (protein)
  - Arg2: break point

# A sample predicate-argument structure

- truncate.01 The predicate
  - Arg0: truncator
  - Arg1: entity shortened (protein)
  - Arg2: break point

A s...ment

Assumption:

Verbs are  
appropriate  
level of  
description

(vs. arbitrary  
frames)

d type

protein)

• truncate

-Arg

-Arg

-Arg

# A sample predicate-argument structure

Assumption:

- to  
—  
—  
—  
...and different senses  
have different PAS

# A sample predicate-argument structure

- truncat

- Arg0:

- Arg1:

- Arg2:

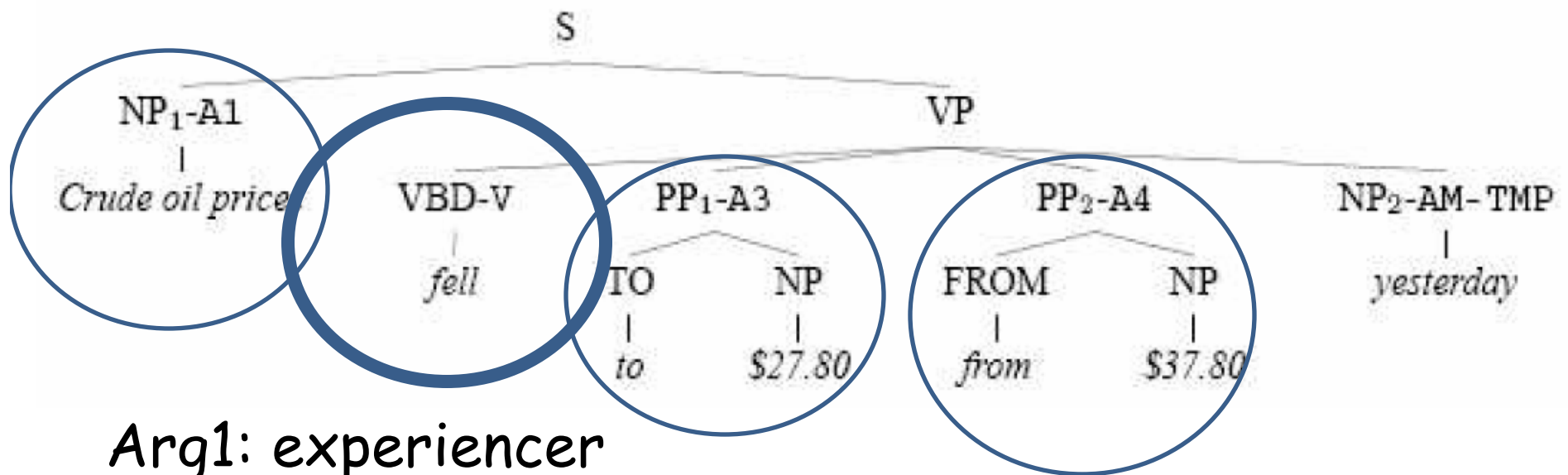
Assumption:

No constraint  
on arity of  
argument sets

nt set

rotein)

# Arguments must be syntactic constituents



- Arg1: experiencer
- Arg2: origin
- Arg3: distance
- Arg4: destination

Figure adapted from  
Haghighi et al. (2005)

# PAS representations can be evaluated

- truncate.01
  - Arg0: agent
  - Arg1: truncated entity
  - Arg2: entity removed
  - Arg3: break point

# PAS representations can be evaluated

- To evaluate the role of the N-terminus of *M.MspI*, 2-hydroxy-5-nitrobenzyl bromide (HNBB) was used to truncate *M.MspI* between residues 34 and 35.
- **Arg0:** 2-hydroxy-5-nitrobenzyl bromide (HNBB)  
**Arg1:** *M.MspI*  
**Arg2:** -  
**Arg3:** residues 34 and 35

# PAS representations can be evaluated

- The putative resulting protein retained the region encoding the structural and functional elements of the amine oxidase but **the second and fourth SRCR domains were truncated** and the potential BMP-1 cleavage site was not present.
- **Arg0:** -  
**Arg1:** -  
**Arg2:** the second and fourth SRCR domains  
**Arg3:** -

# PAS representations can be evaluated

- Truncate.01: *shorten* (examples with Arg1)
  - Arg0: agent
  - **Arg1: entity shortened (protein)**
  - Arg2: break point
- Truncate.02: *remove* (examples with Arg2)
  - Arg0: agent
  - **Arg1: entity removed (domain, terminus, etc.)**
  - Arg2: break point

# A PAS representation for *mutate*

- Arg0: mutagen
- Arg1: wild-type residue
- Arg2: mutant residue
- Arg3: location

# A PAS representation for *mutate*

- Arg0: mutagen
- Arg1: mutated gene/protein
- Arg2: wild-type residue
- Arg3: mutant residue
- Arg4 (or ARG-LOC): location

# *Mutation PAS example*

- *...mutation [of leucine] [at position 1014] [to phenylalanine]... (10196741)*
  - Arg0: absent
  - Arg1: *(of) leucine*
  - Arg2: *(to) phenylalanine*
  - Arg3 or ARG-LOC: *(at) position 1014*

# Predicate-argument structure

- Pred  
asse
  - Argu  
whos
  - Along  
seme
  - ...and
- Totally ignored:

Decompositional

  - Features
  - Events

Ontological

  - Entailment
  - Temporal inclusion

support inference
- f
- action

# ...but, they're useful!

- PropBank project: large treebanked corpus annotated with PAS (scalability)
- Birthed semantic role labelling (information extraction applications)
- Starting point for full set of frames (scalability)

...and, they can be evaluated

- Long history of prior work on semantic representation amazingly unevaluated
- Cohen and Hunter (2006): Evaluation of PASBio via distributional characteristics of arguments; theta-criterion violations

# Outline of the talk

- Theories of semantic representation in linguistics
- A representation for *mutate*
- The phenomenon of nominalization
- A phenomenon in the scientific literature about mutations and its implications for linguistic theory
- A set of possible conclusions

# Nominalization

- Nominalization: Conversion of other parts of speech to nouns
  - Adjective → Noun
    - *phosphorylatable* (PMID 17911107) → *phosphorylatability* (PMID 11158247)
  - Verb → Noun
    - *mutate* → *mutation*

# Nominalization

Far more common in academic writing than in other types of language

	CONV	FICT	NEWS	ACAD
<i>-tion</i>	500	1,500	4,500	<b>11,000</b>
<i>-ity</i>	1,000	1,000	2,500	<b>5,000</b>
<i>-ism</i>	<250	<250	500	<b>1,000</b>
<i>-ness</i>	< 250	<b>1,000</b>	500	500

Occurrences per million words of the four most common suffixes used to form abstract nouns.  
Biber et al. 1999:322.

# Nominalization

- Friedman et al. (2002): molecular biology publications tend to contain enormous amount of data in "complex nominalizations"
- Tateisi et al. (2004): *"...analysis of verb phrases is not sufficient because reactions and relations are often expressed in nominal phrases."*
- Nonetheless, mostly ignored in BioNLP (exceptions: Genescene, RLIMS-P, EDGAR)

# Nominalization & *mutate*

- Data set: development set from Caporaso et al. (2007a)
  - Abstracts/titles: 305
  - Tokens: 61,436
  - Nouns: 118
  - Mutate(s): 0
  - Mutated: 13
  - Mutating: 2
  - Total verbs: 15

Nominalization  
outnumbers verbs  
8:1

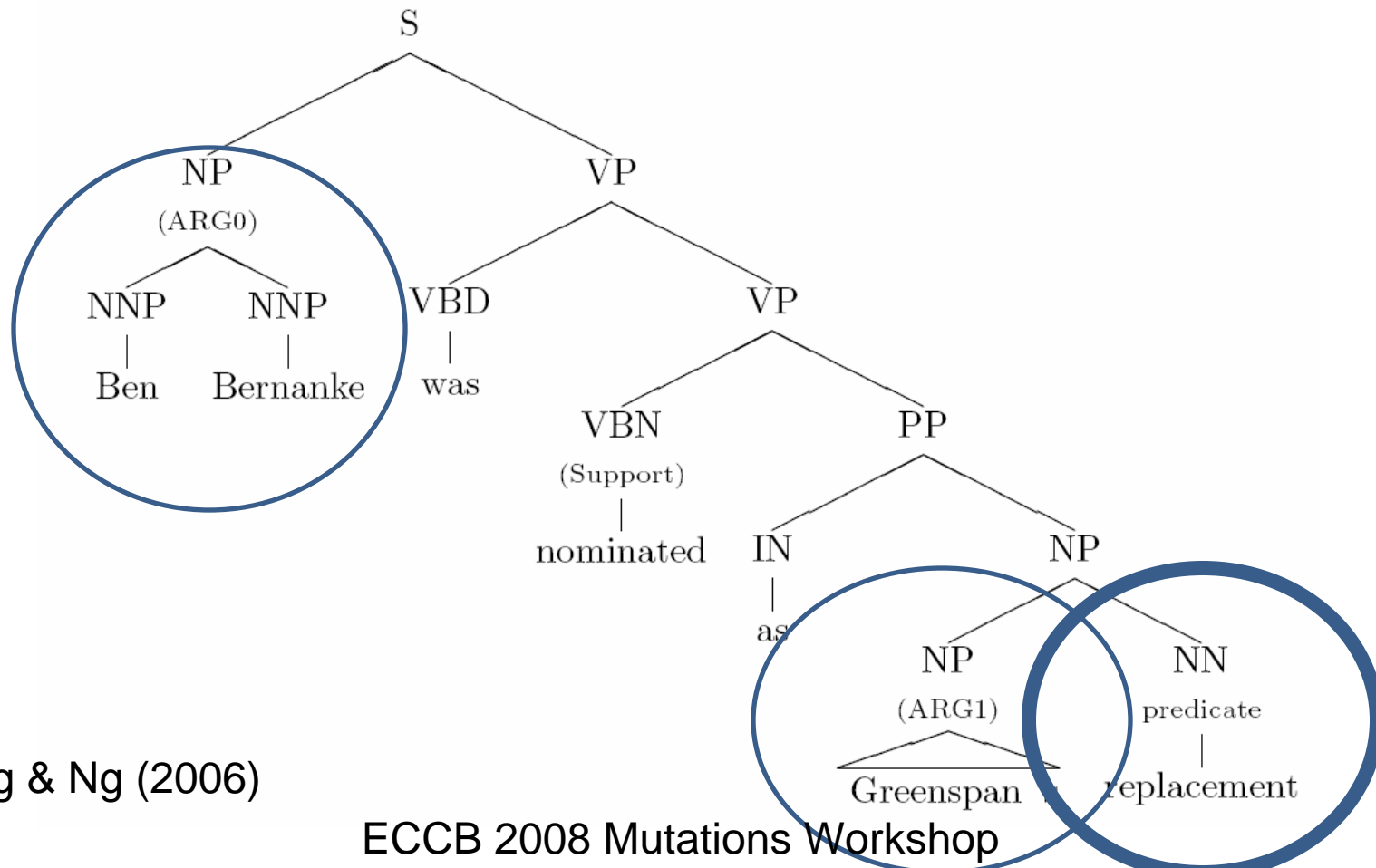
# Nominalization & *mutate*

- Data set: full-text articles from Lee et al. (2007) collection
- Nouns: 1473
- Mutate(s): 3
- Mutated: 70
- Mutating: 8
- Total verbs: 81

Nominalization  
outnumbers verbs  
18:1

# Nominalization

- Nouns can take arguments, too



Jiang & Ng (2006)

ECCB 2008 Mutations Workshop

# Nominalization

- Nominalizations take part in very subtle syntactic/semantic alternations
  - Active: *RK's phosphorylation of R* (PMID 10448166)
  - Passive (of a nominalization): *...the receptor's phosphorylation by the kinase* (PMID 6090944)
  - Passive (nominalization of): *...phosphorylability by cAMP-dependent protein kinase...* (PMID 9660676)

# Nominalization

- ...and the number of possible alternations (patterns) for nominalizations is huge (and much larger than for verbs)...
  - *Increase* and other “calibratable change-of-state verbs” can have  $4^5$  or 1,024 different patterns

# Nominalization

...and many of the  
the *possible*  
patterns tend to  
*actually* occur...

**Table 15.** Alternations for the five three-argument predicates.

	Alternations	Tokens	X	attested/ possible	type/token
<i>Inhibition</i>	24	95	5	0.375	0.253
<i>Induction</i>	19	92	8	0.297	0.21
<i>association.1</i>	5	8	0	0.078	0.625
<i>association.2</i>	10	78	1	0.156	0.128
<i>treatment.04</i>	9	58	7	0.141	0.155

The maximum number possible is  $4^3$ . Data is given for the full BioIE corpus. The column labelled *tokens* shows the number of tokens for which no argument was labelled "can't tell." The column labelled *X* shows the number of tokens with at least one argument labelled "can't tell."

doi:10.1371/journal.pone.0003158.t015

Cohen et al. (2008)

# Nominalization

- ...but, they can all be expected to respect one constraint: the theta-criterion.

# A fundamental concept in linguistic theory

- Question: what's the relationship between the syntax and the semantics of a sentence?
- Part of the answer: the  $\Theta$ -criterion or *Argument Realization Principle*

# A fundamental concept in linguistic theory

- "...the theta criterion...requires that every argument is assigned just one theta role and that every theta role is assigned to just one argument."  
(Crystal 2003:464)

# Outline of the talk

- Theories of semantic representation in linguistics
- A representation for *mutate*
- The phenomenon of nominalization
- A phenomenon in the scientific literature about mutations and its implications for linguistic theory
- A set of possible conclusions

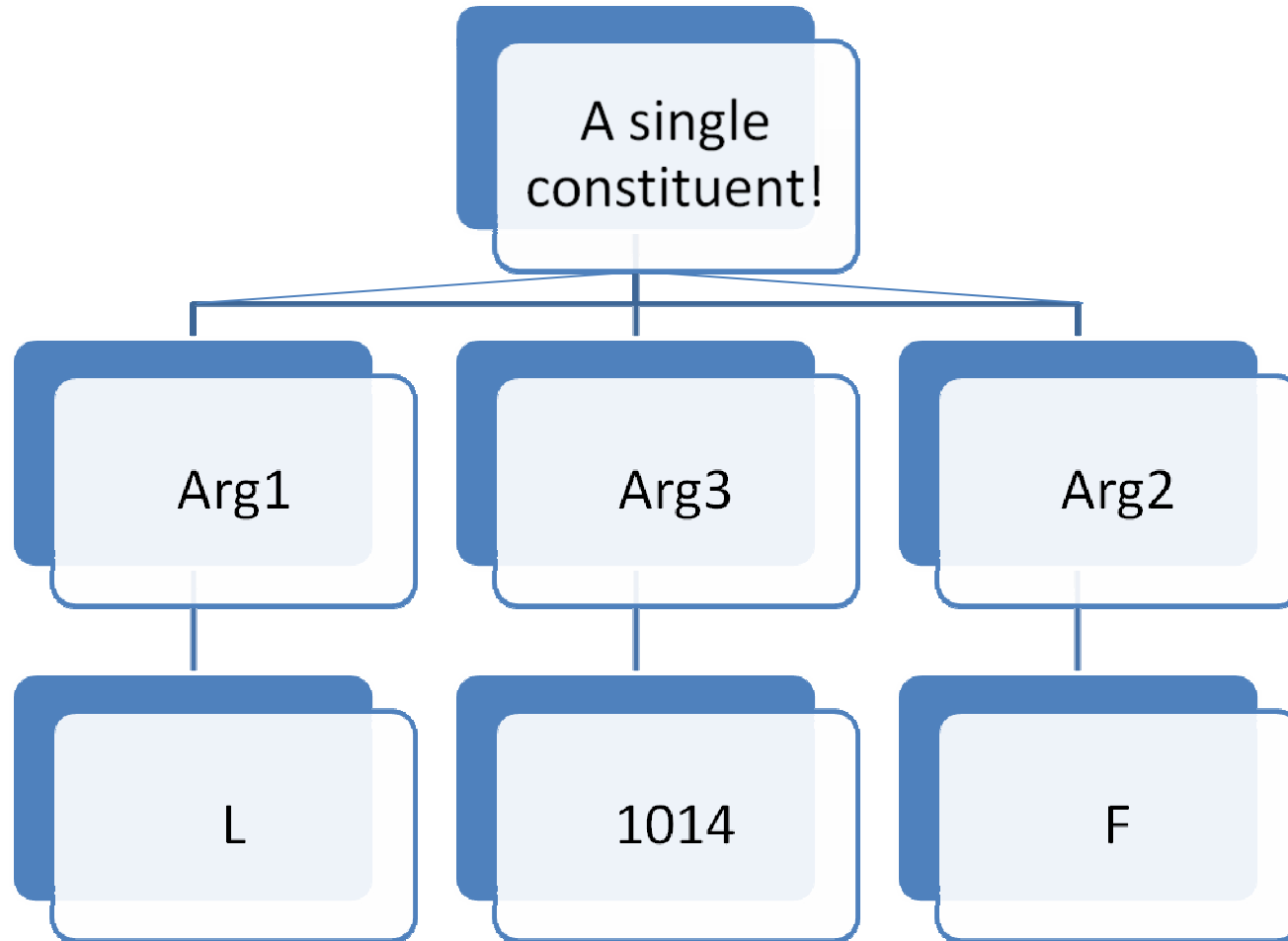
# An unusual pattern in molecular biology: the wNm mutation

- *L1014F mutation 250 hits on Google, Aug. 10 2008*
- *...mutation [of leucine] [at position 1014] [to phenylalanine]... (PMID 10196741)*
  - Arg0: absent
  - Arg1: *(of) leucine*
  - Arg2: *(to) phenylalanine*
  - Arg3 or ARG-LOC: *(at) position 1014*

# *Mutation PAS example*

- *...mutation [of leucine] [at position 1014] [to phenylalanine]... (10196741)*
  - Arg0: absent
  - Arg1: *(of) leucine*
  - Arg2: *(to) phenylalanine*
  - Arg3 or ARG-LOC: *(at) position 1014*

# The wNm construction is odd



# Many positional and elliptical variants possible

- *mutation of w to m*
  - *mutation of aspartate to asparagine*
- *mutation to m of w*
  - *Mutation to glutamate of each of the three aspartate residues...*
- *mutation to m*
  - *mutation to threonine (M918T) in Vssc1 (Lee et al.)*
- Note also that conjoined "objects" count as a single argument
  - *Mutation of Arg-309, Arg-315, or Arg-318...*

Data from Lee et al. (2007)'s

# Incidence of the *wNm* mutation phenomenon

- 3 tokens in the Caporaso (2007a) dev set (abstracts)
- 321 tokens in the Lee et al. (2007) corpus (full text, some duplicates?)

# Incidence of the wNm phenomenon

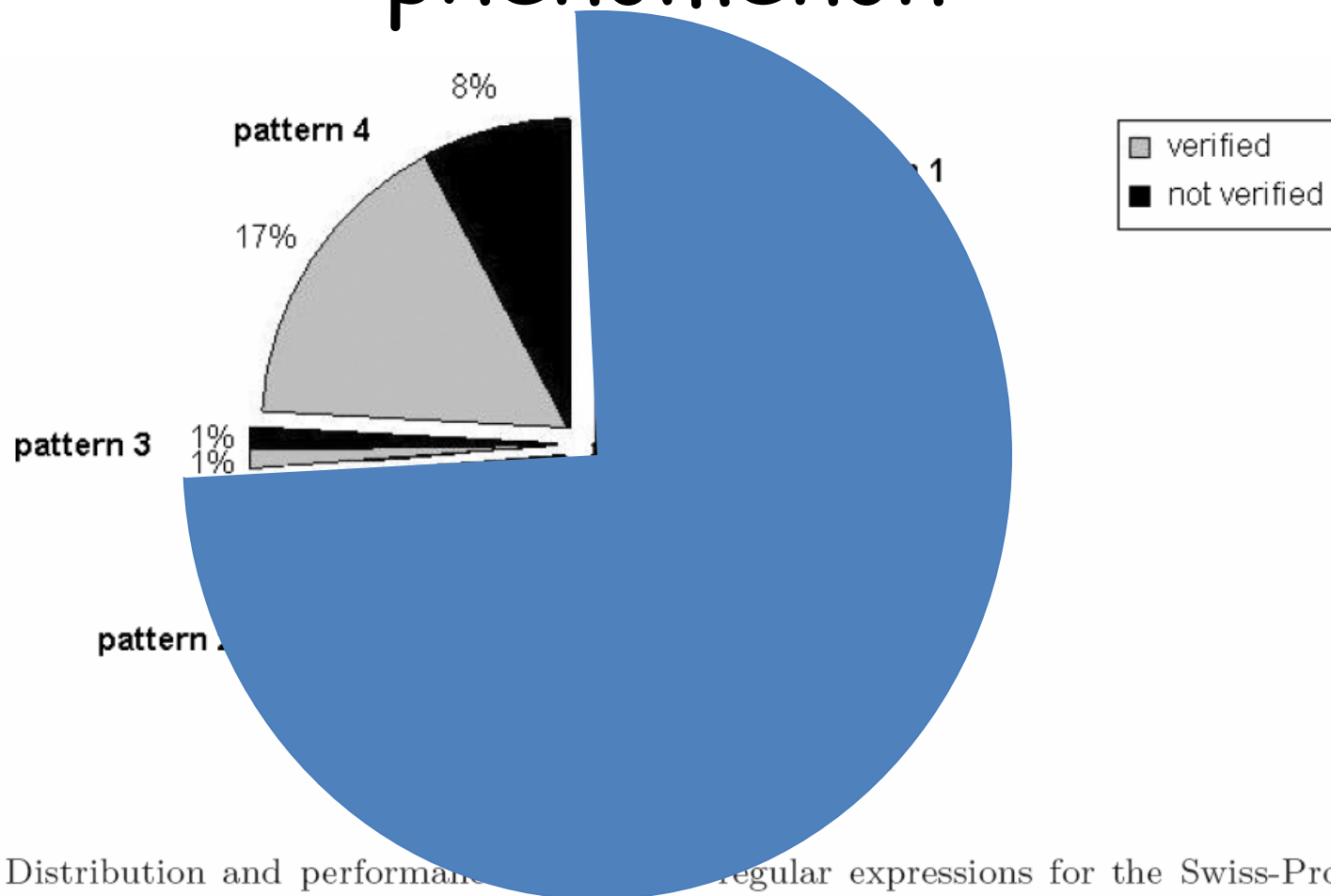


Fig. 2. Distribution and performance of regular expressions for the Swiss-Prot corpus after positional verification.

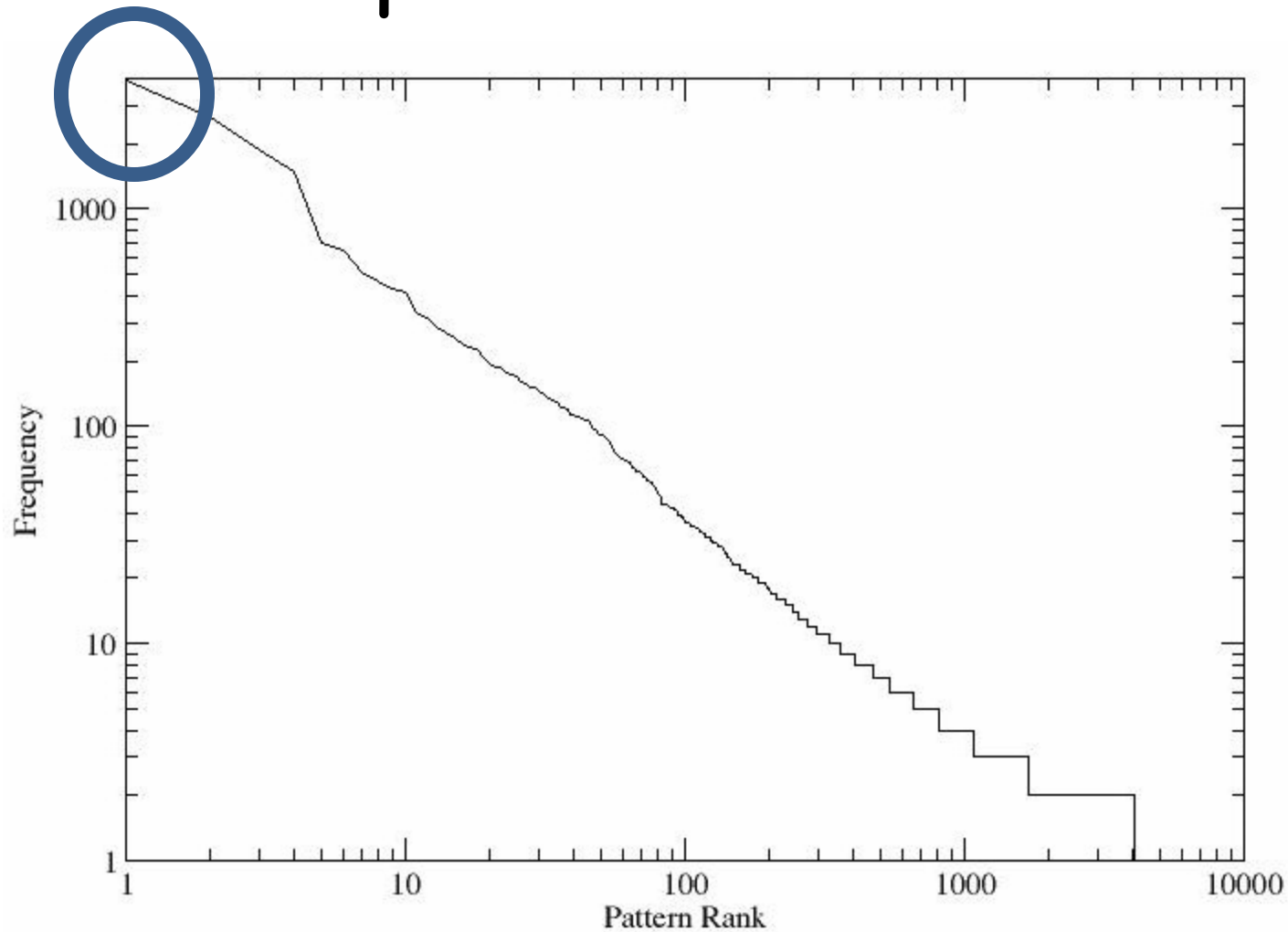
# Incidence of the wNm phenomenon

Table 6. Evaluation of the different patterns by manual checking of 500 documents. The recall and precision are calculated for verified patterns only.

	Verified	Nonverified	TP	FP	TN	FN	Precision (%)	Recall (%)
Pattern 1	283	24	283	0	18	6	100	97.9
Pattern 2	84	10	84	0	5	5	100	94.4
Pattern 3	1	0	1	0	0	0	100	100.0
Pattern 4	47	32	47	0	27	5	100	90.4
Total	415	66	415	0	50	16	100	96.3

TP; true positive; FP, false positive; TN, true negative; FN, false negative.

# Incidence of the wNm phenomenon



Unpublished data—see Caporaso et al.  
ECCB 2008 Mutations Workshop (2007b)

# Different (related?) phenomena

- *...mutation wNm...*
- *...mutation of Tyr 164 to Glu and Phe...  
(15449941)*
- *mutation of K356 to a cysteine... (FTC)*
- *Substitution, polymorphism, variant...*

# Outline of the talk

- Theories of semantic representation in linguistics
- A representation for *mutate*
- The phenomenon of nominalization
- A phenomenon in the scientific literature about mutations and its implications for linguistic theory
- **A set of possible conclusions**

# Take-home point (for linguists and NLPeople)

- Just as our understanding of mutations can benefit from linguistics (natural language processing), so can linguistics benefit from understanding the language of mutations.

# Some counter-arguments to my analysis (and their refutations)

- There is internal structure to the *wNm* construction, so it is > one “constituent,” so it does not violate the  $\Theta$ -criterion/Argument Realization Principle
  - Not on any principled definition of constituent!
    - What’s the phrasal category? Lexical category? Permitted modifiers? Permitted inflections?
  - CanNOT analyze as [NP [PP [NP]]], since we see *wN* but *not Nm*
- *wNm* is a name (“proper noun”) (see e.g. den Dunnen and Antonarakis nomenclature, (2000))
  - Names are constituents, too!
  - The *wNm* has internal structure—most parsimonious explanation for that structure is a reflection of PAS

# What are the implications of this? (4 hypotheses)

- A fundamental theorem is flawed.
  - Jibes with “constructional” alternative to generative (Chomskyan) theories—“the variety of constructions provide alternative ways of packaging information structure” (Goldberg 2005), and with Fillmore’s allowance of multiple Deep Cases per surface constituent
- A popular conception of semantic representations is flawed.
  - If so, implications are horrible, since less superficial conceptions of semantic representation are then even more flawed
- That popular conception of semantic representations is supported (and some less superficial representations are flawed).
  - See Dowty’s blurring of distinctions in the proto-agent/proto-role analysis (Dowty 1991)
- The *wNm mutation* phenomenon is an aspect of the *sublanguage* of molecular biology/genomics.
  - Jibes with notion that ARP must be specified as aspect of the grammar of (General) English; probably not relevant to focus-related arguments about licensing argument *omission*; suggests that the non-subset conception of sublanguages is correct

# Acknowledgements

- Greg Caporaso for extensive discussions of the language of mutations
- Bill Baumgartner and David Randolph for their work on their mutation corpus
- Karin Verspoor and Martha Palmer for discussions of linguistic implications
- Lee et al. (2007) for their work on mutation their corpus
- Anne-Lise Veuthey for help with her data
- Christopher J.O. Baker and Dietrich Rebholz-Schuhmann for organizing this workshop

# References

- Christopher J.O. Baker and René Witte (2006) Mutation mining—a prospector's tale. *Inf Syst Front* 8:47-57.
- J. Gregory Caporaso, William A. Baumgartner Jr., David A. Randolph, K. Bretonnel Cohen, and Lawrence Hunter (2007a) MutationFinder: A high-performance system for extracting point mutations from text. *Bioinformatics* 23(14):1862-1865.
- J. Gregory Caporaso, William A. Baumgartner Jr., David A. Randolph, K. Bretonnel Cohen, and Lawrence Hunter (2007b) Rapid pattern development for concept recognition systems: Application to point mutations. *Journal of Bioinformatics and Computational Biology* 5(6):1233-1259.
- K. Bretonnel Cohen, Martha Palmer, and Lawrence Hunter (2008) Nominalization and alternations in biomedical language. *PLoS ONE* 3(9).
- David Crystal (2003) *A dictionary of linguistics and phonetics*. Wiley-Blackwell.
- Johan T. den Dunnen and Stylianos Antonarakis (2000) Mutation nomenclature extensions and suggestions to describe complex mutations: A discussion. *Human Mutation* 15:7-12.
- Carol Friedman, Pauline Kra, and Andrey Rzhetsky (2002) Two biomedical sublanguages: a description based on the theories of Zellig Harris. *Journal of Biomedical Informatics* 35(4):222-235.
- Lawrence C. Lee, Florence Horn, and Fred Cohen (2007) Automatic extraction of protein point mutations using a graph bigram association. *PLoS Computational Biology* 3(2).
- Christiane Fellbaum (1998) *A semantic network of English verbs*.
- Adele E. Goldberg (2005) *Constructions, lexical semantics and the Correspondence Principle: Accounting for generalizations and subregularities in the realization of arguments*. In Erteschik-Shir and Rapoport (20005), eds.
- Aria Haghighi, Kristina Toutanova, and Christopher Manning (2005) A joint model for semantic role labeling. *CoNLL-2005*.
- Florence Horn, Anthony L. Lau, and Fred E. Cohen (2004) Automated extraction of mutation data from the literature: application of MuteXt to G protein-coupled receptors and nuclear hormone receptors. *Bioinformatics* 20(4):557-568.
- Zheng Ping Jiang and Hwee Tou Ng (2006) Semantic role labeling of NomBank: A maximum entropy approach. *Proc. EMNLP 2006*, pp. 138-145.
- Dietrich Rebholz-Schuhmann, Stephane Marcel, Sylvie Albert, Ralf Tolle, Georg Casari, and Harald Kirsch (2004) Automatic extraction of mutations from Medline and cross-validation with OMIM. *Nucleic Acids Research* 32:135-142.
- Yuka Tateisi, Tomoko Ohta, and Jun-ichi Tsujii (2004) Annotation of predicate-argument structure of molecular biology text. *IJCNLP04 workshop: Beyond Shallow Analyses*.
- Yum Lina Yip, Nathalie Lacheneval, Violaine Pillet, and Anne-Lisse Veuthey (2007) Retrieving mutation-specific information for human proteins in UniProt/Swiss-Prot knowledgebase. *Journal of Bioinformatics and Computational Biology* 5(6):1215-1231.