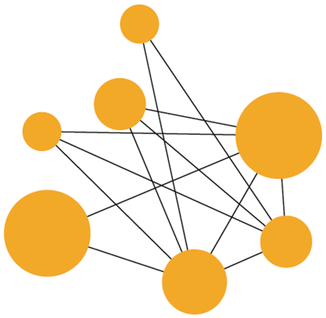
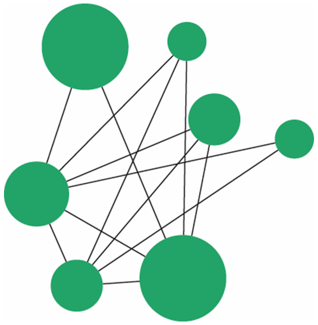
computational bioscience program
university of colorado school of medicine

Verbs in Biomedical Text

Karin Verspoor
Faculty, Computational Bioscience Program
University of Colorado School of Medicine

Kevin Bretonnel Cohen
Biomedical Text Mining Group Lead
University of Colorado School of Medicine





Karin.Verspoor@ucdenver.edu
http://compbio.ucdenver.edu/Hunter_lab/Verspoor

Kevin.Cohen@gmail.com
http://compbio.ucdenver.edu/Hunter_lab/Cohen

THE CONTEXT for biomedical natural language processing

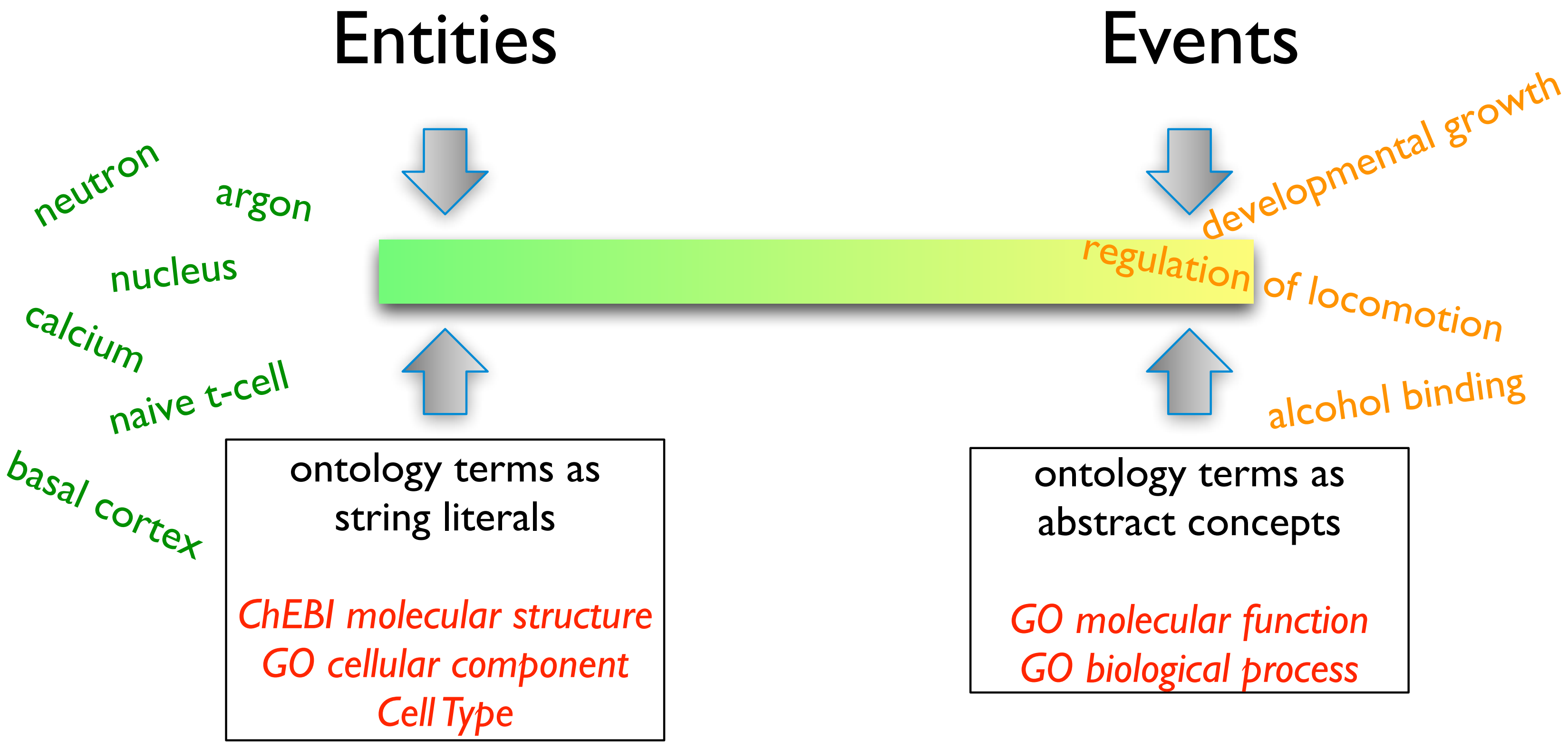
Exponential knowledge growth in biomedicine

- 1,330 peer-reviewed gene-related databases in 2011 NAR db issue
- Over 20 million PubMed entries (> 2,200/day)
- Breakdown of disciplinary boundaries makes more of it relevant to each of us
- “Like drinking from a firehose” – Jim Ostell

© 1998 The Expositor

Language processing of Biomedical texts

- Tools that support identification, indexing, and extraction of biological concepts



Entities

Events

neutron

argon

nucleus

calcium

naive t-cell

basal cortex

developmental growth

regulation of locomotion

alcohol binding

ontology terms as string literals

ChEBI molecular structure
GO cellular component
Cell Type

ontology terms as abstract concepts

GO molecular function
GO biological process

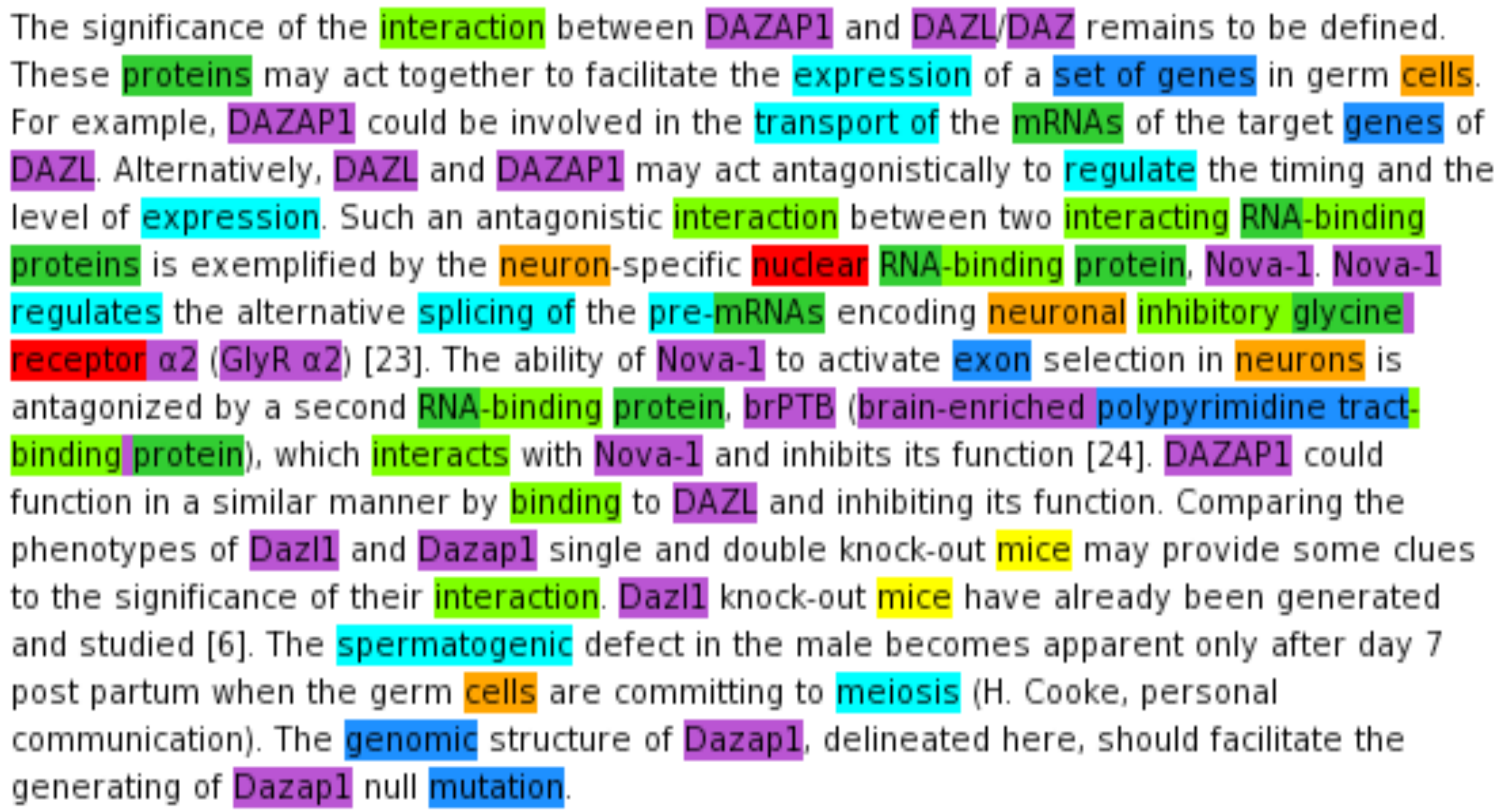
Scientific Publishing & Semantics

- Content enrichment
- Direct access to (relevant) external data
- Structured digital abstracts

- Enables
  - Interactivity
  - targeted searches
  - relevance linking
  - formalizing content; actionable data

Making BioNLP relevant

- Recognition of OBO terms, relations
- CRAFT corpus (first release very soon)

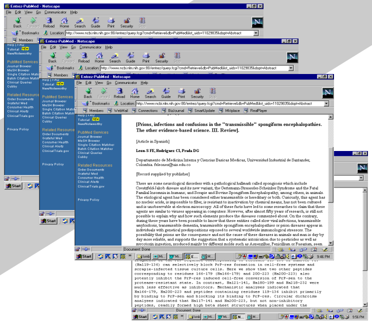


The significance of the interaction between DAZAP1 and DAZL/DAZ remains to be defined. These proteins may act together to facilitate the expression of a set of genes in germ cells. For example, DAZAP1 could be involved in the transport of the mRNAs of the target genes of DAZL. Alternatively, DAZL and DAZAP1 may act antagonistically to regulate the timing and the level of expression. Such an antagonistic interaction between two interacting RNA-binding proteins is exemplified by the neuron-specific nuclear RNA-binding protein, Nova-1. Nova-1 regulates the alternative splicing of the pre-mRNAs encoding neuronal inhibitory glycine receptor α2 (GlyR α2) [23]. The ability of Nova-1 to activate exon selection in neurons is antagonized by a second RNA-binding protein, brPTB (brain-enriched polypyrimidine tract-binding protein), which interacts with Nova-1 and inhibits its function [24]. DAZAP1 could function in a similar manner by binding to DAZL and inhibiting its function. Comparing the phenotypes of Dazl1 and Dazap1 single and double knock-out mice may provide some clues to the significance of their interaction. Dazl1 knock-out mice have already been generated and studied [6]. The spermatogenic defect in the male becomes apparent only after day 7 post partum when the germ cells are committing to meiosis (H. Cooke, personal communication). The genomic structure of Dazap1, delineated here, should facilitate the generating of Dazap1 null mutation.

Model Organism Curation Pipeline

1. Select papers

MEDLINE



[Prions, infections and confusions in the "transmissible" spongiform encephalopathies. The other evidence-based science. III. Review].

[Article in Spanish]

Leon-S FE, Rodriguez CI, Prada DG

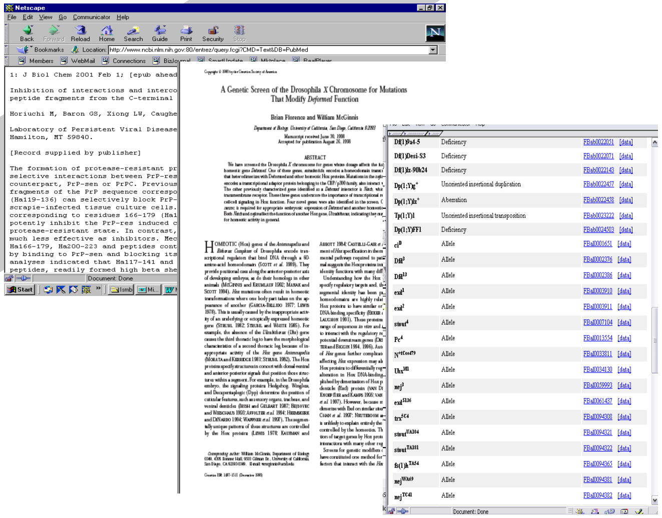
Departamento de Medicina Interna y Ciencias Basicas Medicas, Universidad Industrial de Santander, Colombia. feleones@uis.edu.co

[Record supplied by publisher]

There are some neurological disorders with a pathological hallmark called spongiosis which include Creutzfeld-Jakob disease and its new variant, the Gertsmann-Straussler-Scheinker Syndrome and the Fatal Familial Insomnia in humans; and Scrapie and Bovine Spongiform Encephalopathy, among others, in animals. The etiological agent has been considered either transmissible or hereditary or both. Curiously, this agent has no nucleic acids, is impossible to filter, is resistant to inactivation by chemical means, has not been cultured and is unobservable at electron microscopy. All of these facts have led to some researches to claim that these agents are similar to viruses appearing in computers. However, after almost fifty years of research, is still not possible to explain why and how such elements produce the diseases commented about. On the contrary, during these years have been possible to know that these entities called slow viral infections, transmissible amyloidosis, transmissible dementia, transmissible spongiform encephalopathies or prion diseases appear in individuals with genetical predispositions exposed to several worldwide immunological stressors. The possibility that prions are the consequence and not the cause of these diseases in animals and man is day by day more reliable, and supports the suggestion that a systematic intoxication due to pesticides as well as mycotoxin ingestion, produced mainly by different molds such as Aspergillus, Penicillium or Fusarium, seem

```
(Ha119-136) can selectively block PrP-res formation in cell-free systems and
scrapie-infected tissue culture cells. Here we show that two other peptides
corresponding to residues 166-179 (Ha166-179) and 200-223 (Ha200-223) also
potently inhibit the PrP-res induced cell-free conversion of PrP-sen to the
protease-resistant state. In contrast, Ha121-141, Ha180-199 and Ha218-232 were
much less effective as inhibitors. Mechanistic analyses indicated that
Ha166-179, Ha200-223 and peptides containing residues 119-136 inhibit primarily
by binding to PrP-sen and blocking its binding to PrP-res. Circular dichroism
analyses indicated that Ha117-141 and Ha200-223, but not non-inhibitory
peptides, readily formed high beta sheet structures when placed under the
```

2. List genes for curation



Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Guide Print Security Stop
Bookmarks Location: http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?CMD=Text&DB=PubMed
Members WebMail Connections BizJournal SmartUpdate Mktplace RealPlayer
Document: Done
Start
Ismb
Mi...

```
1: J Biol Chem 2001 Feb 1; [epub ahead

Inhibition of interactions and interco
peptide fragments from the C-terminal

Horiuchi M, Baron GS, Xiong LW, Caughe

Laboratory of Persistent Viral Disease
Hamilton, MT 59840.

[Record supplied by publisher]

The formation of protease-resistant pr
selective interactions between PrP-res
counterpart, PrP-sen or PrPC. Previous
fragments of the PrP sequence correspo
(Ha119-136) can selectively block PrP-
scrapie-infected tissue culture cells.
corresponding to residues 166-179 (Ha1
potently inhibit the PrP-res induced c
protease-resistant state. In contrast,
much less effective as inhibitors. Mec
Ha166-179, Ha200-223 and peptides cont
by binding to PrP-sen and blocking its
analyses indicated that Ha117-141 and
peptides, readily formed high beta she
```

Copyright © 1998 by the Genetics Society of America

A Genetic Screen of the Drosophila *X* Chromosome for Mutations That Modify *Deformed* Function

Brian Florence and William McGinnis

Department of Biology, University of California, San Diego, California 92093

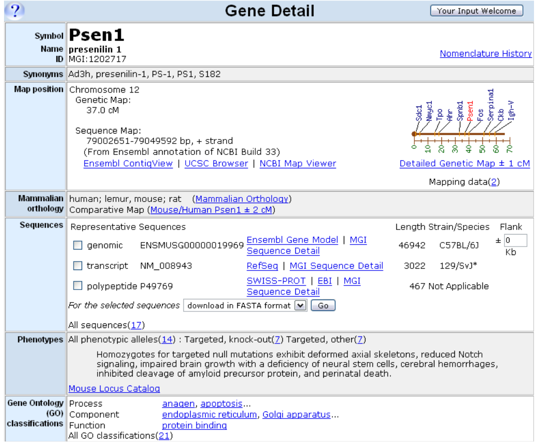
Manuscript received June 30, 1998
Accepted for publication August 26, 1998

ABSTRACT

Df(1)9a4-5	Deficiency	FBab0022051	[data]
Df(1)Desi-S3	Deficiency	FBab0022071	[data]
Df(1)lz-90b24	Deficiency	FBab0022143	[data]
Dp(1;Y)g^+	Unoriented insertional duplication	FBab0022457	[data]
Dp(1;Y)lz^+	Aberration	FBab0022458	[data]
Tp(1;Y)l	Unoriented insertional transposition	FBab0023222	[data]
Dp(1;Y)FF1	Deficiency	FBab0024503	[data]
ci^D	Allele	FBal0001651	[data]
Dfd^3	Allele	FBal0002376	[data]
Dfd^{13}	Allele	FBal0002386	[data]
exd^1	Allele	FBal0003910	[data]
exd^2	Allele	FBal0003911	[data]
$stout^4$	Allele	FBal0007104	[data]
Pc^4	Allele	FBal0013554	[data]
$N^{+tCos479}$	Allele	FBal0033811	[data]
Ubx^{M1}	Allele	FBal0034130	[data]
nej^3	Allele	FBal0059993	[data]
exd^{S136}	Allele	FBal0061437	[data]
trx^{5C4}	Allele	FBal0094308	[data]
$stout^{UA104}$	Allele	FBal0094321	[data]
$stout^{TA181}$	Allele	FBal0094322	[data]
$fs(1)h^{TA54}$	Allele	FBal0094365	[data]
nej^{WA69}	Allele	FBal0094381	[data]
nej^{TC41}	Allele	FBal0094382	[data]

Document: Done

3. Curate genes from paper



Gene Detail

Symbol Name ID	**Psen1** **presenilin 1** MGI:1202717 Nomenclature History
Synonyms	Ad3h, presenilin-1, PS-1, PS1, S182
Map position	Chromosome 12 Genetic Map: 37.0 cM Sequence Map: 79002651-79049592 bp, + strand (From Ensembl annotation of NCBI Build 33) Ensembl ContigView | UCSC Browser | NCBI Map Viewer Detailed Genetic Map ± 1 cM Mapping data(2)
Mammalian orthology	human; lemur, mouse; rat (Mammalian Orthology) Comparative Map (Mouse/Human Psen1 ± 2 cM)
Sequences	Representative Sequences — Length Strain/Species Flank □ genomic ENSMUSG00000019969 Ensembl Gene Model | MGI Sequence Detail 46942 C57BL/6J ± 0 Kb □ transcript NM_008943 RefSeq | MGI Sequence Detail 3022 129/SvJ* □ polypeptide P49769 SWISS-PROT | EBI | MGI Sequence Detail 467 Not Applicable *For the selected sequences* download in FASTA format All sequences(17)
Phenotypes	All phenotypic alleles(14) : Targeted, knock-out(7) Targeted, other(7) Homozygotes for targeted null mutations exhibit deformed axial skeletons, reduced Notch signaling, impaired brain growth with a deficiency of neural stem cells, cerebral hemorrhages, inhibited cleavage of amyloid precursor protein, and perinatal death. Mouse Locus Catalog
Gene Ontology (GO) classifications	Process anagen, apoptosis... Component endoplasmic reticulum, Golgi apparatus... Function protein binding All GO classifications(21)

From Hirschman et al. BMC Bioinformatics 2005 6 (Suppl 1):S1

THE LINGUISTIC CHALLENGES

of BioNLP

Verbs in Biomedical Text

Biomedical	BNC
show	do
suggest	say
use	make
indicate	go
contain	see
describe	take
express	get
bind	know
require	come
observe	give
find	think
determine	use
demonstrate	find
perform	look
induce	want

- Verb usage differs significantly from general English
- Domain-specific verbs: *phosphorylate, ubiquitinate*
- Verbs that have a domain-specific sense: *express, regulate, signal, transcribe*

Arity

K. Bretonnel Cohen and Lawrence Hunter (2006). A critical review of PASBio's argument structures for biomedical verbs. *BMC Bioinformatics* 7(Suppl. 3):S5.

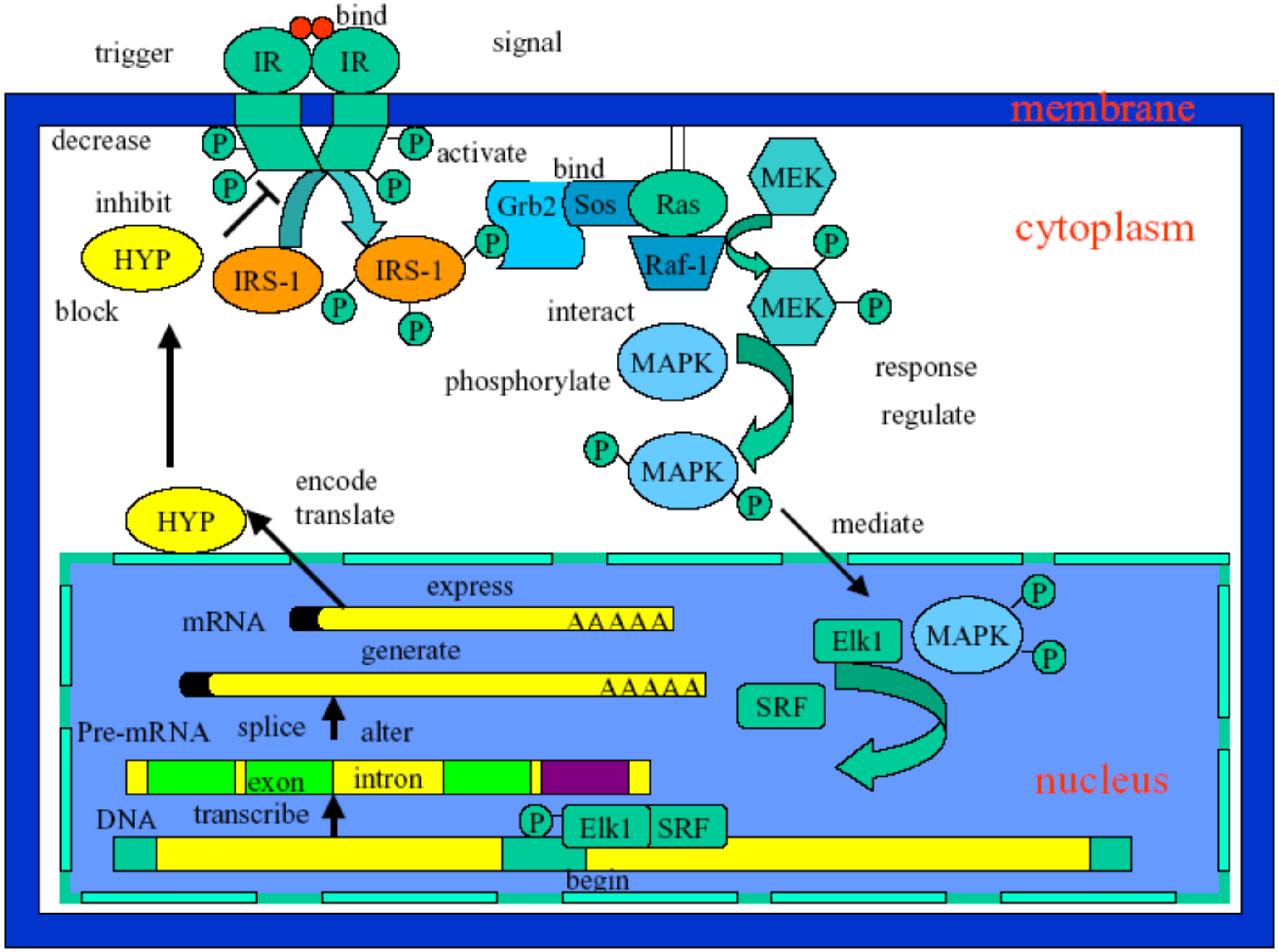
Most research on biomedical semantics

- All relationships binary
  - Protein/protein
  - Drug/gene
  - Drug/disease
  - Drug/effect

PASBio (Predicate-Argument Structures for Biology) Project

- Described in Wattarujeekrit, Shah, and Collier (2004)
- Set of 29 verbs, 34 predicates with associated argument structures
- 10 annotated examples each
- Publicly available

Predicate selection for PASBio:gene expression; regulation; signalling



PASBio findings

- Overlap between domain-specific and "General English" semantics is low
- Biological domain has more "core arguments," fewer "adjuncts"

PASBio findings

- 9/29 didn't occur in PropBank or had different sense
- 45% (9/20) had more arguments
- 25% (5/20) had fewer arguments
- 30% (6/20) had same number

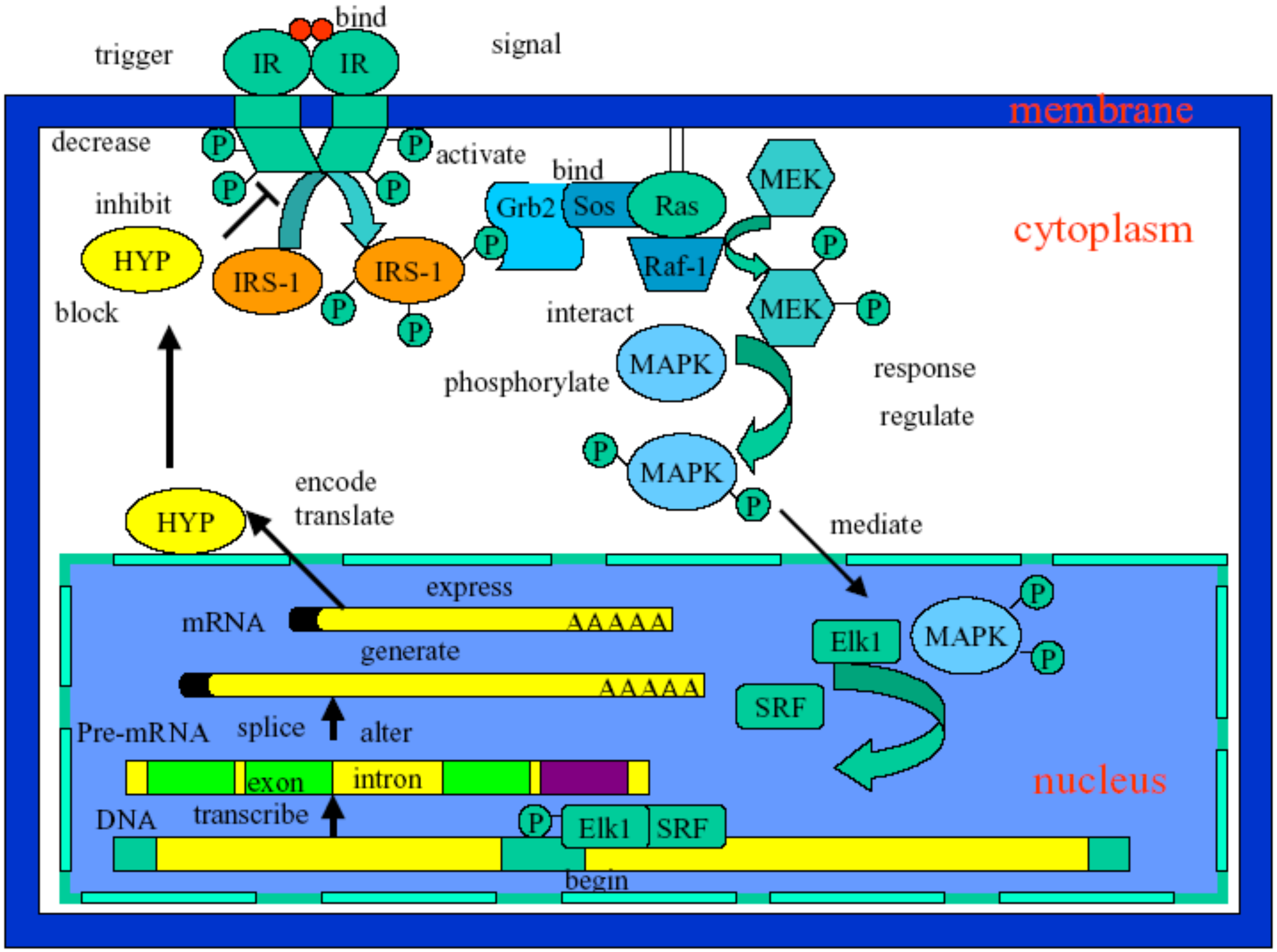
Native speaker intuition behind this

- I don’t believe what you tell me unless I know when, where, at what pH, at what temperature…
- Consequence: weak distinction between core arguments and adjuncts

Example representation: *express*

- Arg1: named entity being expressed (gene or gene product)
- Arg2: property of the existing name entity
- Arg3: location refering to organelle, cell or tissue

Expression


trigger
bind
IR
IR
signal
membrane
decrease
P
P
activate
bind
inhibit
P
P
Grb2
Sos
Ras
MEK
HYP
IRS-1
IRS-1
P
Raf-1
P
cytoplasm
block
P
P
interact
MEK
P
MAPK
phosphorylate
response
regulate
P
MAPK
encode
translate
P
HYP
mediate
express
P
mRNA
AAAAA
Elk1
MAPK
generate
P
AAAAA
SRF
Pre-mRNA
splice
alter
exon
intron
nucleus
DNA
transcribe
P
Elk1
SRF
begin

Are these arguments to a predicate…

- Example Arg2s:
  - two mRNA isoforms of 2.4 and 4.0 kb
  - 2.0 and 2.4 kilobases in length

Two equally abundant mRNAs for il8ra, 2.0 and 2.4 kilobases in length, are expressed in neutrophils and arise from usage of two alternative polyadenylation signals.

…or slots in a frame?

- PUNDIT:
  - Customer
  - Symptoms
  - Actions taken
  - Success or failure

Note: frame slots inferred from Palmer et al. (1986)

Subcategorization frames

- Help to derive the correct interpretation
- e.g. resolve attachment ambiguities:

assessing ubiquitin expression in infected mice brains

two poly(A)+ RNAs transcribed from the opposite strand of the upstream flanking regions lacked …

Subcategorization frame learning

- Goal: acquire subcategorization frames from a corpus
- utilize dependency relations derived from a statistical parser
- map from sets of dependency relations to a SCF via manually developed (unification) rules
- rules defined for COMLEX, ANLT, NOMLEX frames

Preiss, Briscoe, Korhonen. “A System for Large-Scale Acquisition of Verbal, Nomical, and Adjectival, Subcategorization Frames from Corpora”, ACL 2007.

Biomedical verb classification

- Induce lexical classes from corpora
  - Build an inventory of subcategorization frames for each verb
  - Cluster verbs according to shared subcat frames

	indicate		protein	activate	gene
It	suggests	that	p53	activates	WAF1
	demonstrates		TP53	up-regulates	CIP1
	indicates		DMP53	induces	p21
	implies		…	stimulates	…

Korhonen, Krymolowski, Collier. "Automatic Classification of Verbs in Biomedical Texts", ACL 2006, p. 345-352.

Nominalization and Alternations in Biomedical Text

K. Bretonnel Cohen, Martha Palmer, and Lawrence Hunter (2008). Nominalization and alternations in biomedical language. *PLoS ONE* 3(9).

Goals of the study

- Characterize behavior of nominalizations in biomedical text
- Determine implications for system-building

Definitions

- Nominalization: noun derived from a verb
  - Verbal nominalization: *activation, inhibition, induction*
  - Argument nominalization: *activator, inhibitor, inducer, mutant*

Nominalizations are dominant in biomedical texts

Predicate	Nominalization	All verb forms
Express	**2,909**	1,233
Develop	**1,408**	597
Analyze	**1,565**	364
Observe	185	**809**
Differentiate	**737**	166
Describe	10	**621**
Compare	185	**668**
Lose	**556**	74
Perform	86	**599**
Form	**533**	511

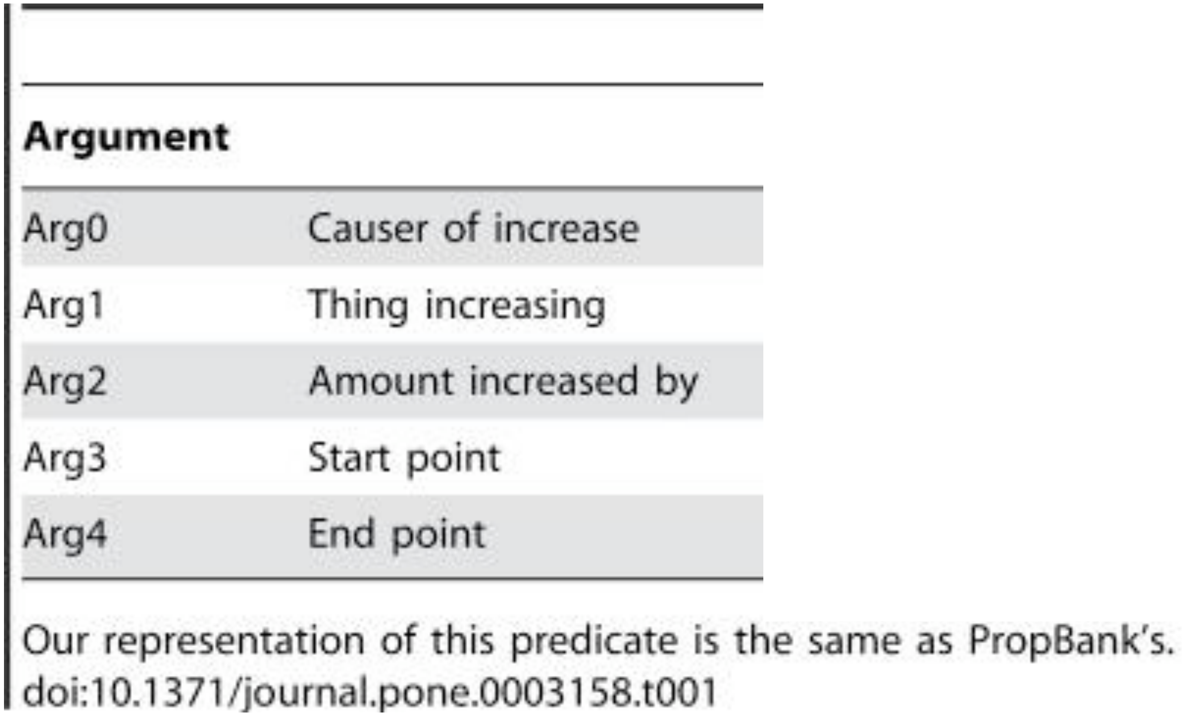
Data from CRAFT corpus

Relevant points for text mining

- Nominalizations are an obvious route for scaling up recall
- Nominalizations are more difficult to handle than verbs…
- …but can yield higher precision (Cohen et al. 2008)

Definitions

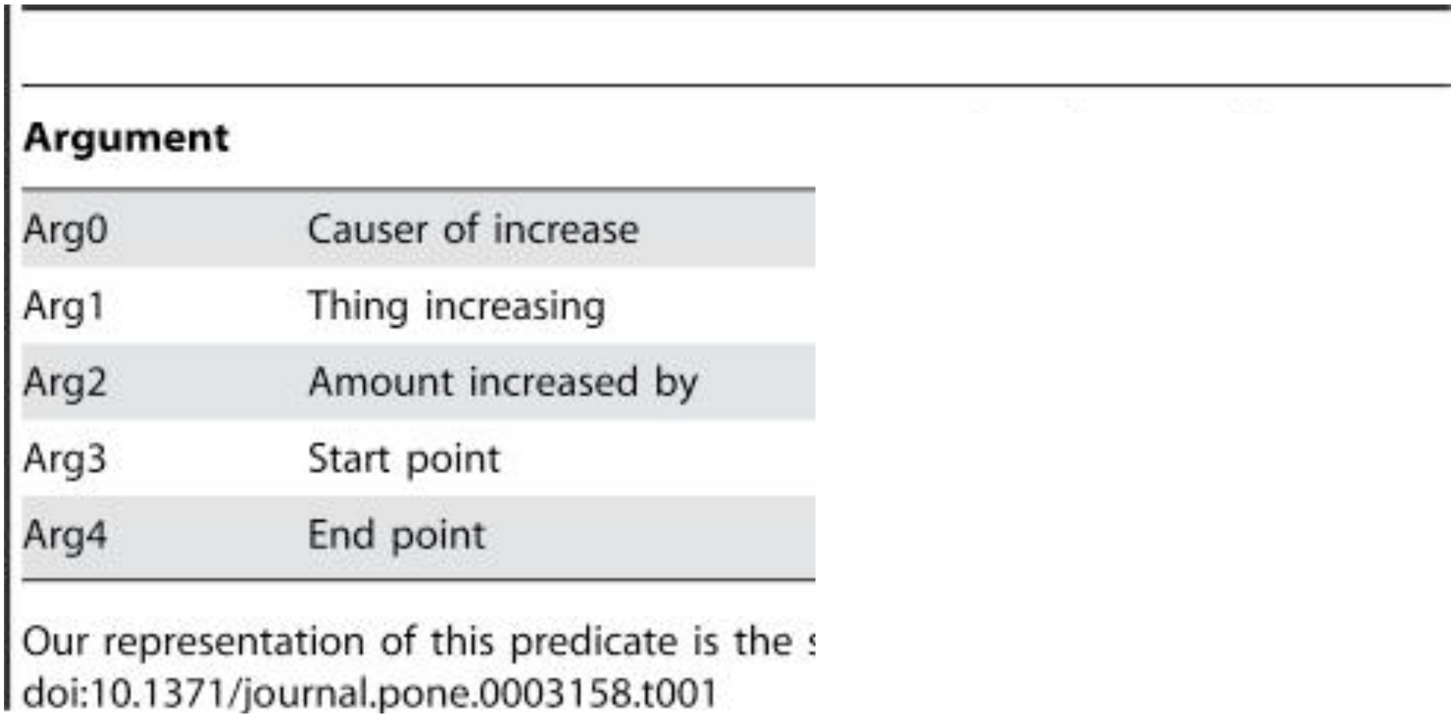
- Argument: participant in or qualifier of the action of the predicate



Argument	
Arg0	Causer of increase
Arg1	Thing increasing
Arg2	Amount increased by
Arg3	Start point
Arg4	End point

Our representation of this predicate is the same as PropBank's.
doi:10.1371/journal.pone.0003158.t001

Definitions



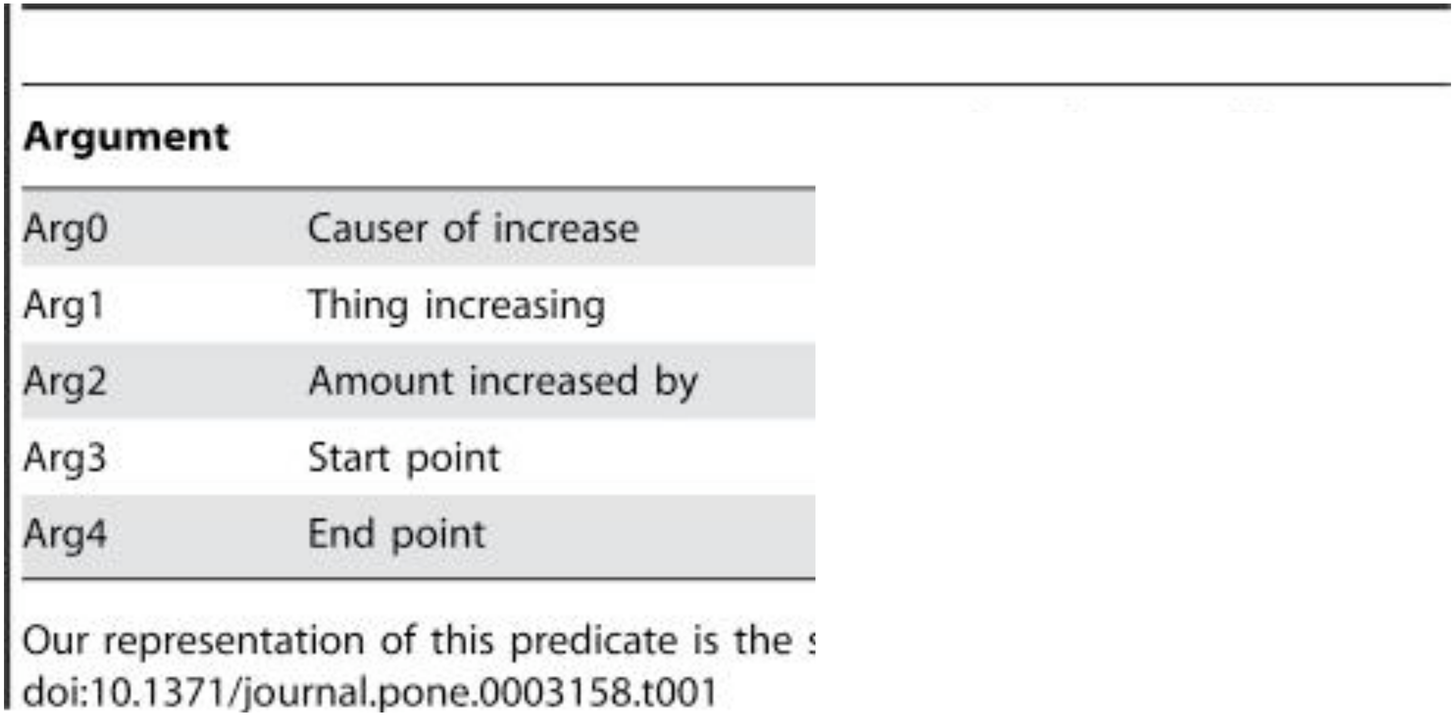
Argument	
Arg0	Causer of increase
Arg1	Thing increasing
Arg2	Amount increased by
Arg3	Start point
Arg4	End point

Our representation of this predicate is the

doi:10.1371/journal.pone.0003158.t001

D0870, an azole antifungal agent[Arg0], *produced dose-related* ***increases*** *in total cytochrome P450 and aldrin epoxidase*[Arg1]

Definitions



Argument	
Arg0	Causer of increase
Arg1	Thing increasing
Arg2	Amount increased by
Arg3	Start point
Arg4	End point

Our representation of this predicate is the s

doi:10.1371/journal.pone.0003158.t001

Increase *in phosphorylation of APP*[Arg1] *by overexpression of the nerve growth factor receptor Trk A*[Arg0]

Definitions

- 2 arguments: *Activate*
  - Arg0: Activator
  - Arg1: Activatee
- 3 arguments: *Inhibit*
  - Arg0: Inhibitor
  - Arg1: Inhibitee
  - Arg2: Amount of inhibition

Definitions

- Alternation: variations in the surface syntactic form of predicates and their arguments
  - Active/passive
    - X phosphorylates Y
    - Y is phosphorylated by X
  - Transitive/intransitive
    - X decreases Y
    - Y decreases

Alternations of nominalizations: positions of arguments

- Any combination of the set of positions for each argument of a nominalization
  - Pre-nominal: *phenobarbital* ***induction***, *trkA* ***expression***
  - Post-nominal: ***increases*** *of oxygen*
  - No argument present: ***Induction*** *followed a slower kinetic…*
  - Noun-phrase-external: *this enzyme can undergo* ***activation***

Pre-nominal arguments

- Agent (Arg0)
  - cytochrome(s) P-450 **mediation**
  - interferon-gamma **inhibition** of VSV replication
  - Phenobarbital **treatment**
- Patient (Arg1, ≅ logical object)
  - trkA **expression**
  - agonist **association**
  - cancer **treatment**

Noun-phrase-external arguments

- EWS/FLI-1 antagonists induce growth **inhibition** of Ewing tumor cells
  - Support verb links agent to noun phrase
- potency of sertraline for dopamine reuptake **inhibition**
  - Transparent noun
- Phenobarbital (PB) has long been known as an inducer of drug-metabolizing enzymes in liver, but the molecular mechanism underlying this **induction** is still poorly understood
  - Event coreference

Alternations of nominalizations

- **activation** of molecular oxygen by alkaline hemin
  - Arg0 post-nominal, Arg1 post-nominal
- K(ATP) **activation** by cromakalim
  - Arg0 post-nominal, Arg1 pre-nominal
- Mutational **activation** of the ras genes
  - Arg0 pre-nominal, Arg1 post-nominal

Previous work on nominalizations in the biomedical domain

- Ono et al. (2001): *interaction, association, complex,* and *binding*
- Pustejovsky et al. (2002): *inhibition* and *inhibitor*
- Hu et al. (2005), Narayanaswamy et al. (2005), Yuan et al. (2006): *phosphorylation*
- Lots of early work by Zellig Harris, the Linguistic String Project, other workers in sublanguage model

Prediction investigated

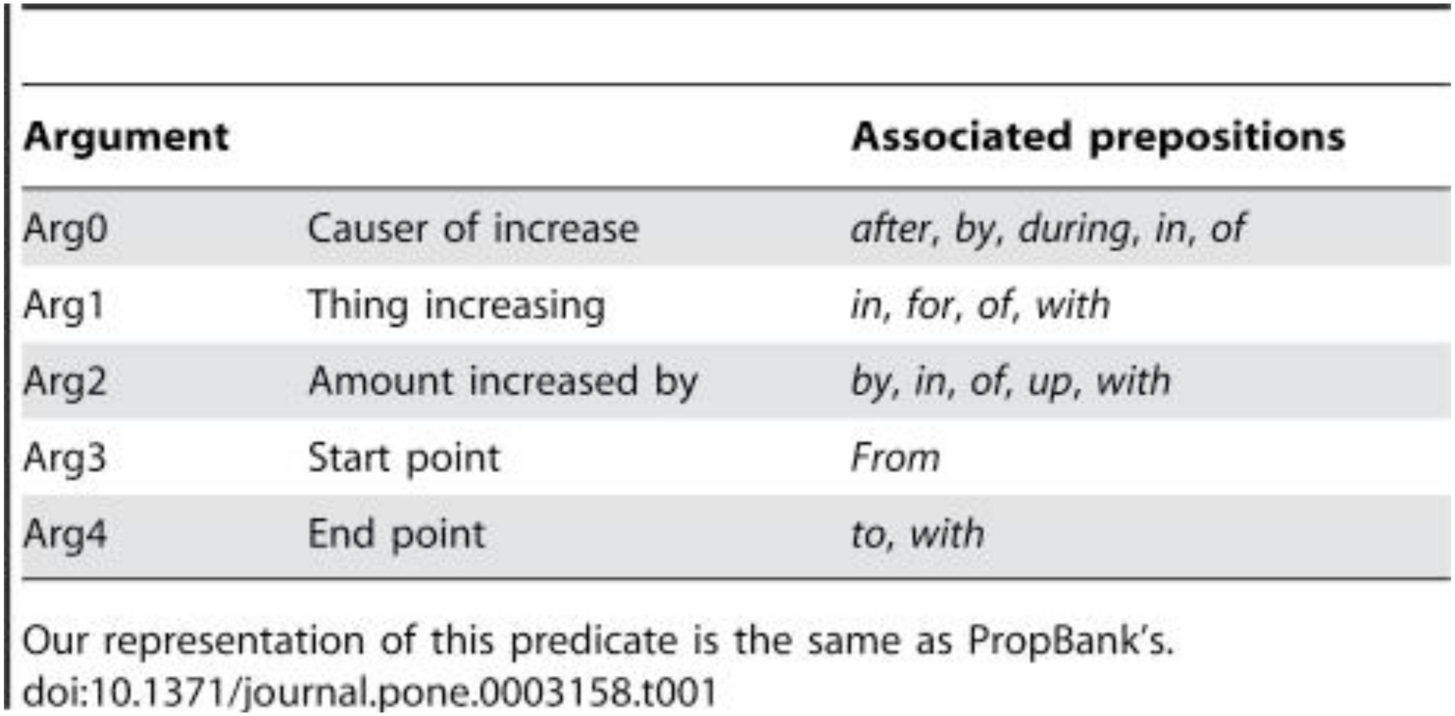
- Within scientific language, we should expect a limited variety of alternations

Previous work on nominalizations in the biomedical domain

- GENESCENE: tackles all verbal nominalizations
  - Arguments recognized only if following nominalization and preceded by *of*, *in*, or *by*

Leroy and Chen (2002), Leroy et al. (2003),
Leroy and Chen (2005)

- A sample predicate for which the three prepositions *of*, *in*, and *by* are insufficient for capturing all arguments.



Argument		Associated prepositions
Arg0	Causer of increase	*after, by, during, in, of*
Arg1	Thing increasing	*in, for, of, with*
Arg2	Amount increased by	*by, in, of, up, with*
Arg3	Start point	*From*
Arg4	End point	*to, with*

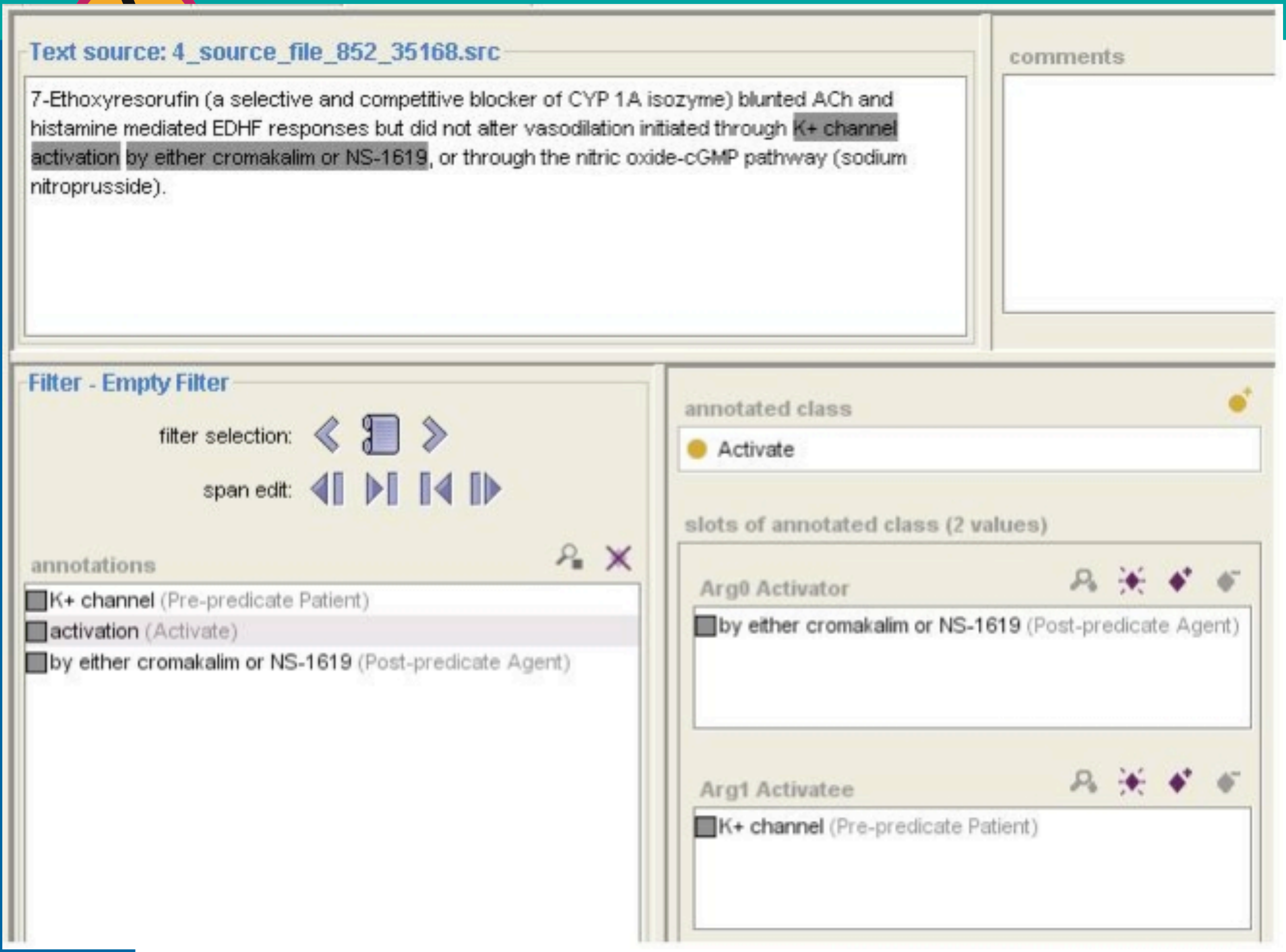
Our representation of this predicate is the same as PropBank's.
doi:10.1371/journal.pone.0003158.t001

Materials and methods

- Release 0.9 of the PennBioIE corpus (collection of abstracts of journal articles, annotated with parts of speech, syntactic structure, and entities)

Materials and methods

- Marked arguments for 746 tokens of nominalizations of the 10 most common verbs
- Second annotator marked 15% of these to calculate interannotator agreement (87.5%)



Text source: 4_source_file_852_35168.src

7-Ethoxyresorufin (a selective and competitive blocker of CYP 1A isozyme) blunted ACh and histamine mediated EDHF responses but did not alter vasodilation initiated through K+ channel activation by either cromakalim or NS-1619, or through the nitric oxide-cGMP pathway (sodium nitroprusside).

comments

Filter - Empty Filter

annotations

- K+ channel (Pre-predicate Patient)
- activation (Activate)
- by either cromakalim or NS-1619 (Post-predicate Agent)

annotated class

● Activate

slots of annotated class (2 values)

Arg0 Activator

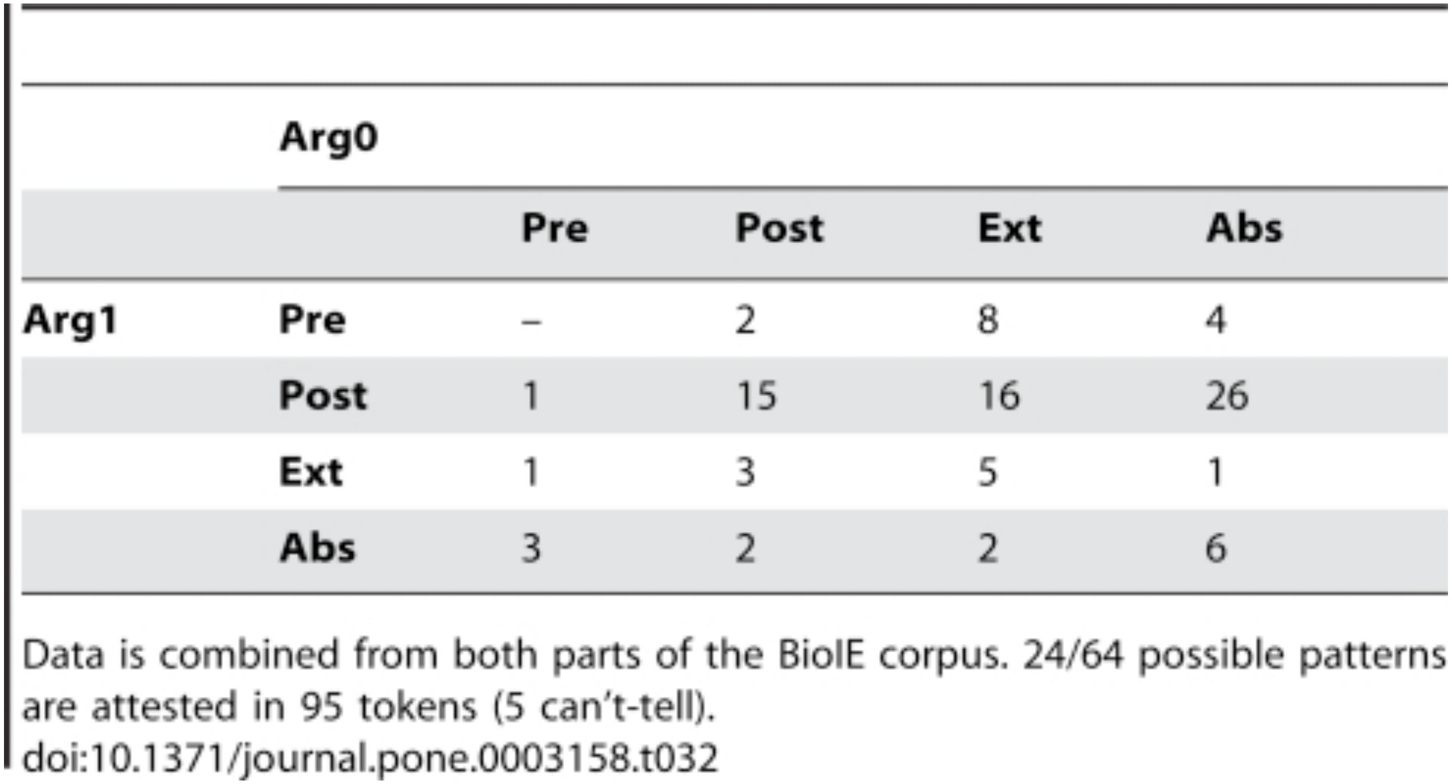
- by either cromakalim or NS-1619 (Post-predicate Agent)

Arg1 Activatee

- K+ channel (Pre-predicate Patient)

Result 1: attested alternations are extraordinarily diverse

- *Inhibition*, a 3-argument predicate—Arguments 0 and 1 only shown

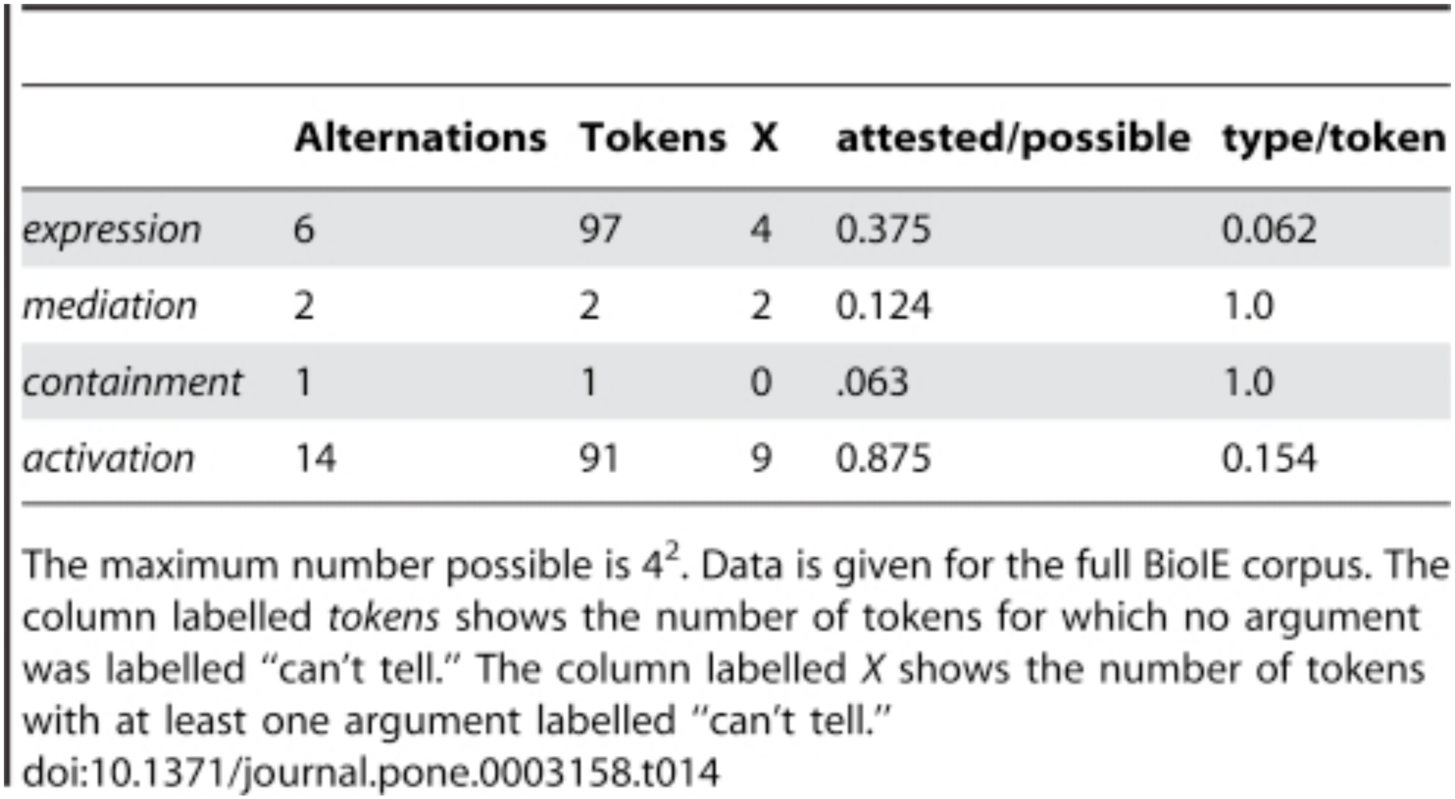


		Arg0			
		Pre	Post	Ext	Abs
Arg1	Pre	–	2	8	4
	Post	1	15	16	26
	Ext	1	3	5	1
	Abs	3	2	2	6

Data is combined from both parts of the BioIE corpus. 24/64 possible patterns are attested in 95 tokens (5 can't-tell).

doi:10.1371/journal.pone.0003158.t032

Results for 2-argument verbs

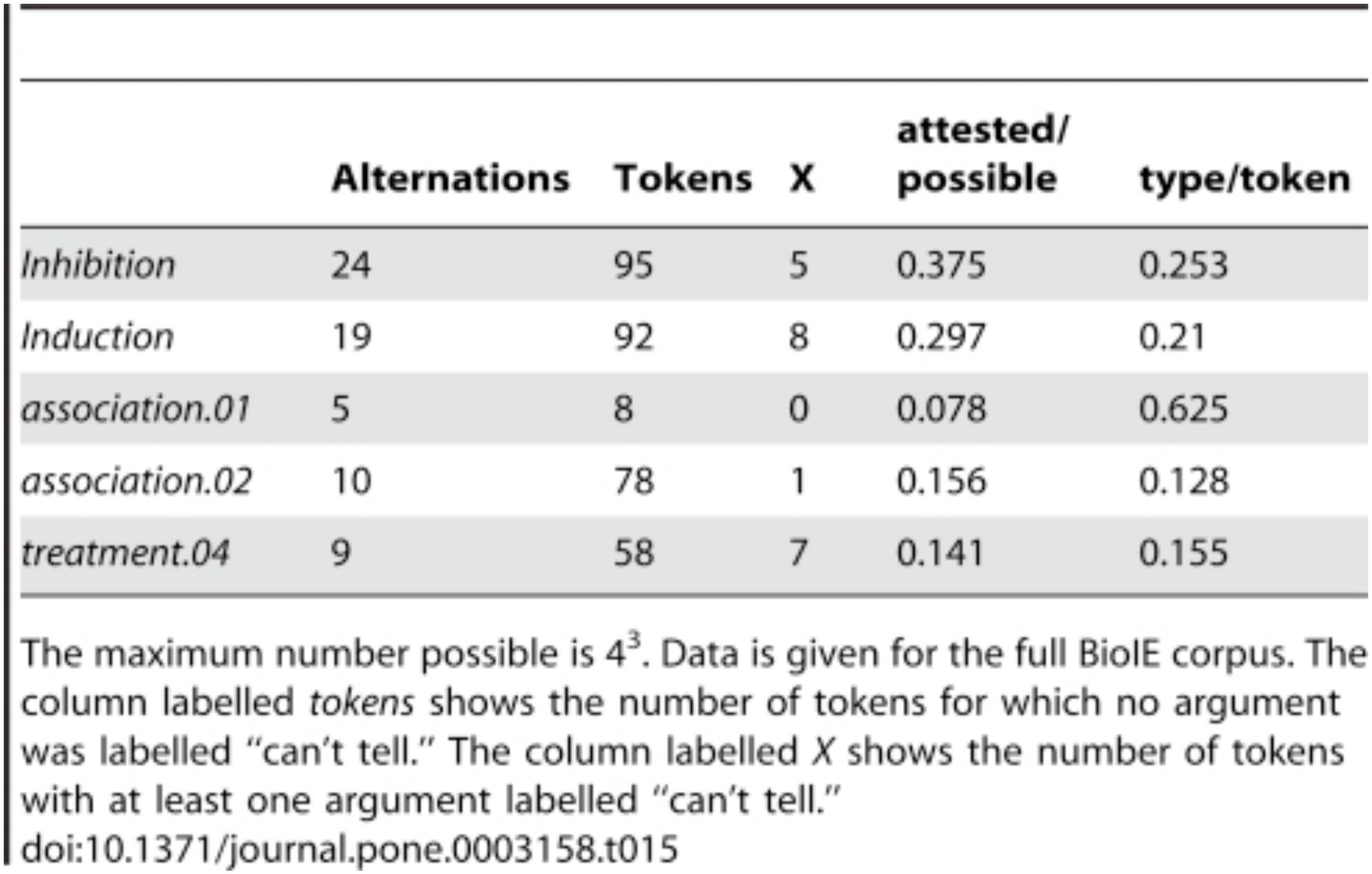


	Alternations	Tokens	X	attested/possible	type/token
expression	6	97	4	0.375	0.062
mediation	2	2	2	0.124	1.0
containment	1	1	0	.063	1.0
activation	14	91	9	0.875	0.154

The maximum number possible is 4^2. 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.t014

Results for 3-argument verbs



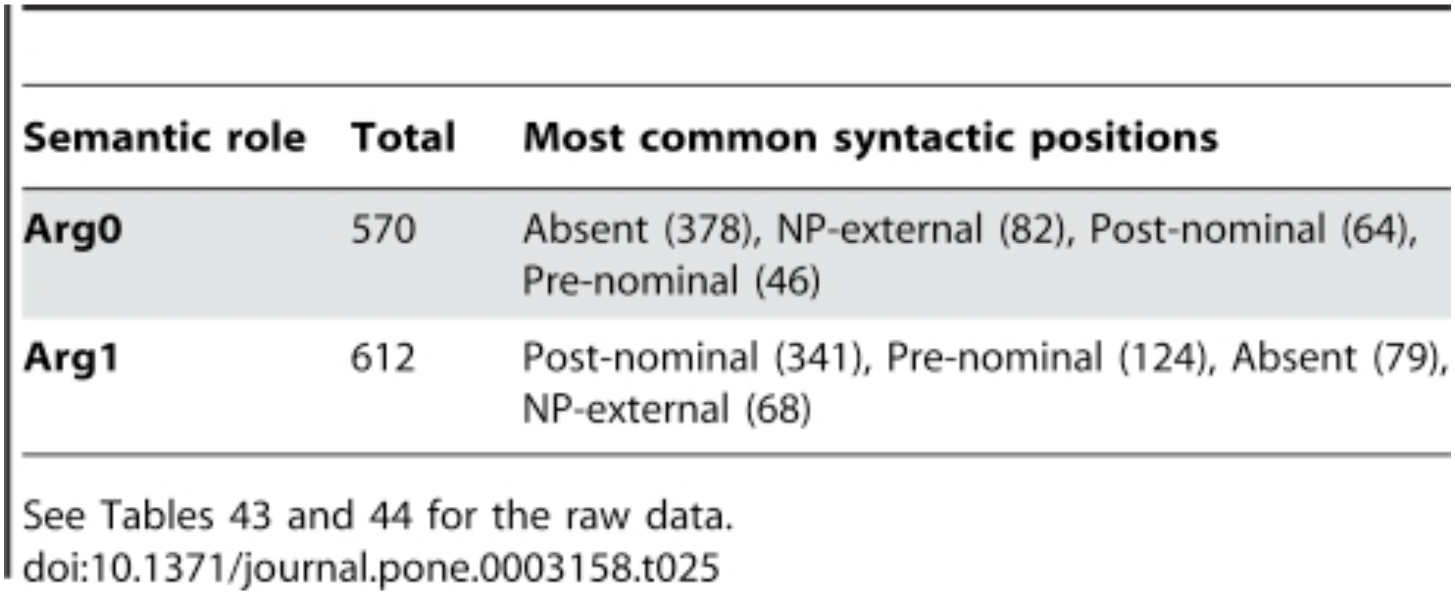
	Alternations	Tokens	X	attested/ possible	type/token
Inhibition	24	95	5	0.375	0.253
Induction	19	92	8	0.297	0.21
association.01	5	8	0	0.078	0.625
association.02	10	78	1	0.156	0.128
treatment.04	9	58	7	0.141	0.155

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

Result 2: syntactic positions

- Most common syntactic positions for each semantic role:

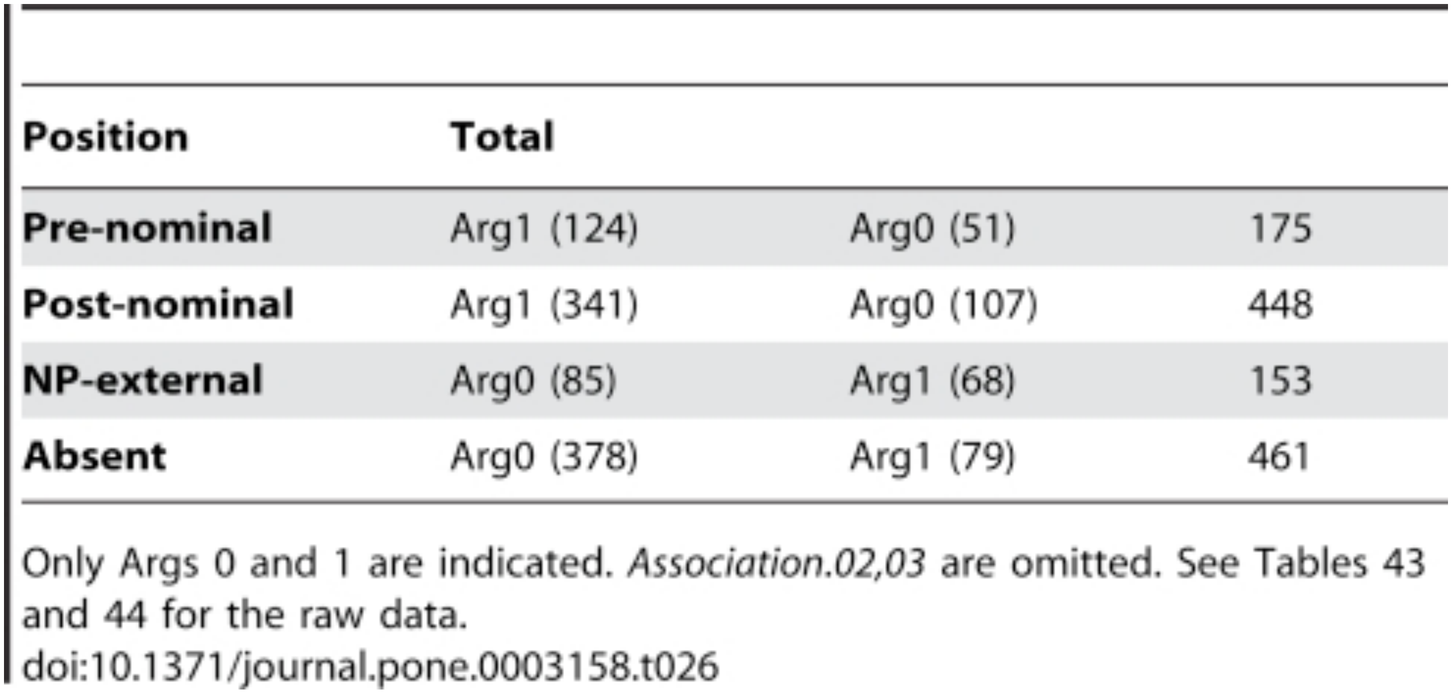


Semantic role	Total	Most common syntactic positions
Arg0	570	Absent (378), NP-external (82), Post-nominal (64), Pre-nominal (46)
Arg1	612	Post-nominal (341), Pre-nominal (124), Absent (79), NP-external (68)

See Tables 43 and 44 for the raw data.
doi:10.1371/journal.pone.0003158.t025

Result 3: semantic roles

- Most frequent semantic roles for each syntactic position:



Position	Total		
Pre-nominal	Arg1 (124)	Arg0 (51)	175
Post-nominal	Arg1 (341)	Arg0 (107)	448
NP-external	Arg0 (85)	Arg1 (68)	153
Absent	Arg0 (378)	Arg1 (79)	461

Only Args 0 and 1 are indicated. *Association.02,03* are omitted. See Tables 43 and 44 for the raw data.
doi:10.1371/journal.pone.0003158.t026

Implications for system-building

- Distinction between absent and noun-phrase-external arguments is crucial and difficult, and finite state approaches will not suffice; merging data from different clauses and sentences may be useful
- Pre-nominal arguments are undergoer by ratio of 2.5:1
- For predicates with agent and patient, post/post and pre/post patterns predominate, but others are common as well

What can be done?

- External arguments:
  - semantic role labelling approach
    - ...but, very important to recognize the absent/ external distinction, especially with machine learning
  - pattern-based approach
    - ...but, approaches to external arguments (RLIMS-P) are so far very predicate-specific

What can be done?

- Pre-nominal arguments:
  - apply heuristic that we have identified based on distributional characteristics
  - for most frequent nominalizations, manual encoding may be tractable

Future analysis

- Can identity of pre-predicate arguments be characterized on a per-predicate basis?
  - At minimum will require word sense disambiguation (phenobarbital treatment/ cancer treatment)
- Can pre-predicate arguments be characterized by semantic class?

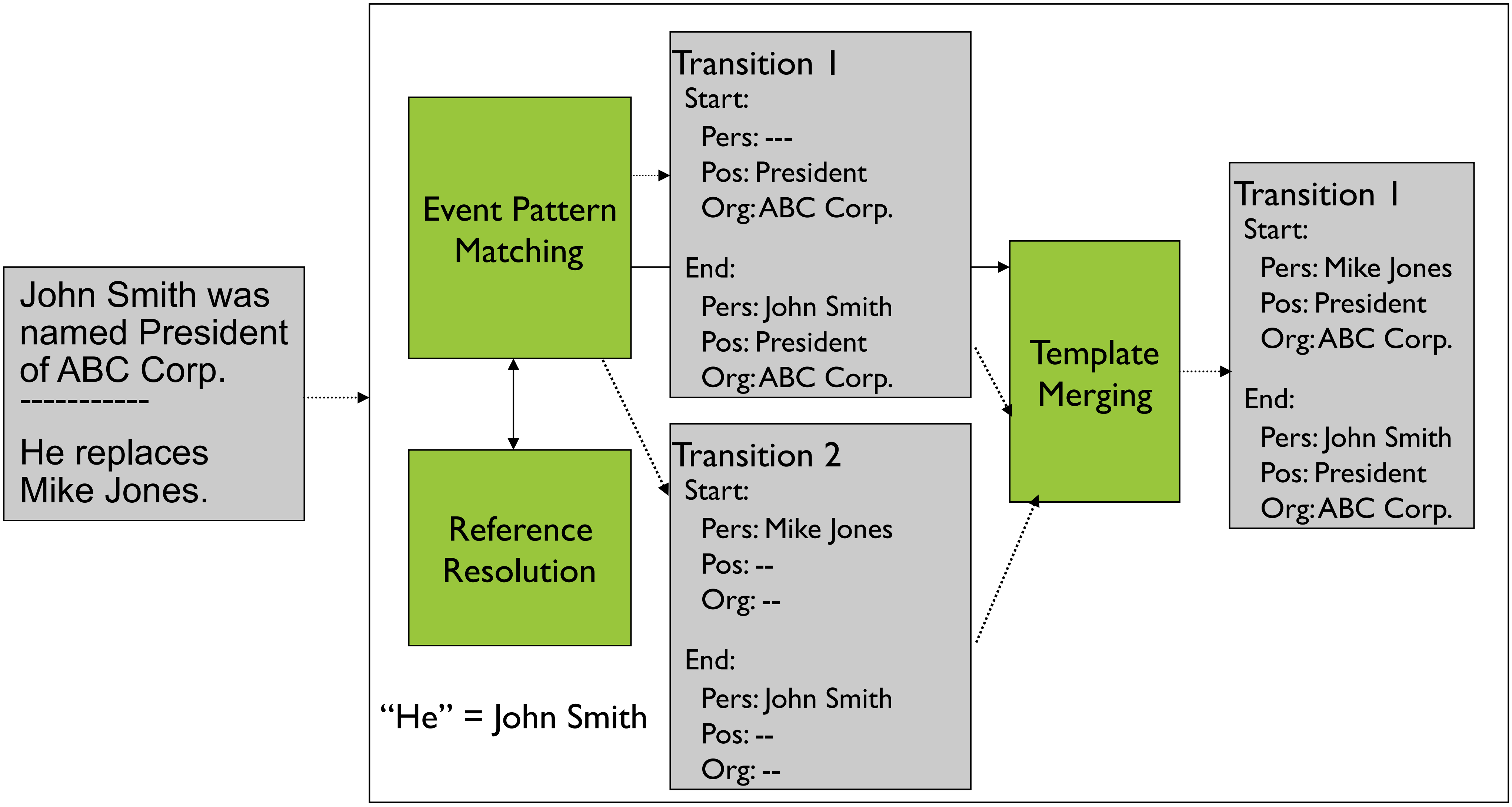
INFORMATION EXTRACTION

technology for BioNLP

Information Extraction

- Algorithms that
  - automatically extract structured information from unstructured (natural language) text
  - aim to identify entities and events of interest
  - utilize natural language processing
    - (linguistic) rule-based
    - machine learning

Information Extraction (MUC example)



John Smith was named President of ABC Corp.

He replaces Mike Jones.

Event Pattern Matching

Reference Resolution

"He" = John Smith

Transition 1
Start:
Pers: ---
Pos: President
Org: ABC Corp.
End:
Pers: John Smith
Pos: President
Org: ABC Corp.

Transition 2
Start:
Pers: Mike Jones
Pos: --
Org: --
End:
Pers: John Smith
Pos: --
Org: --

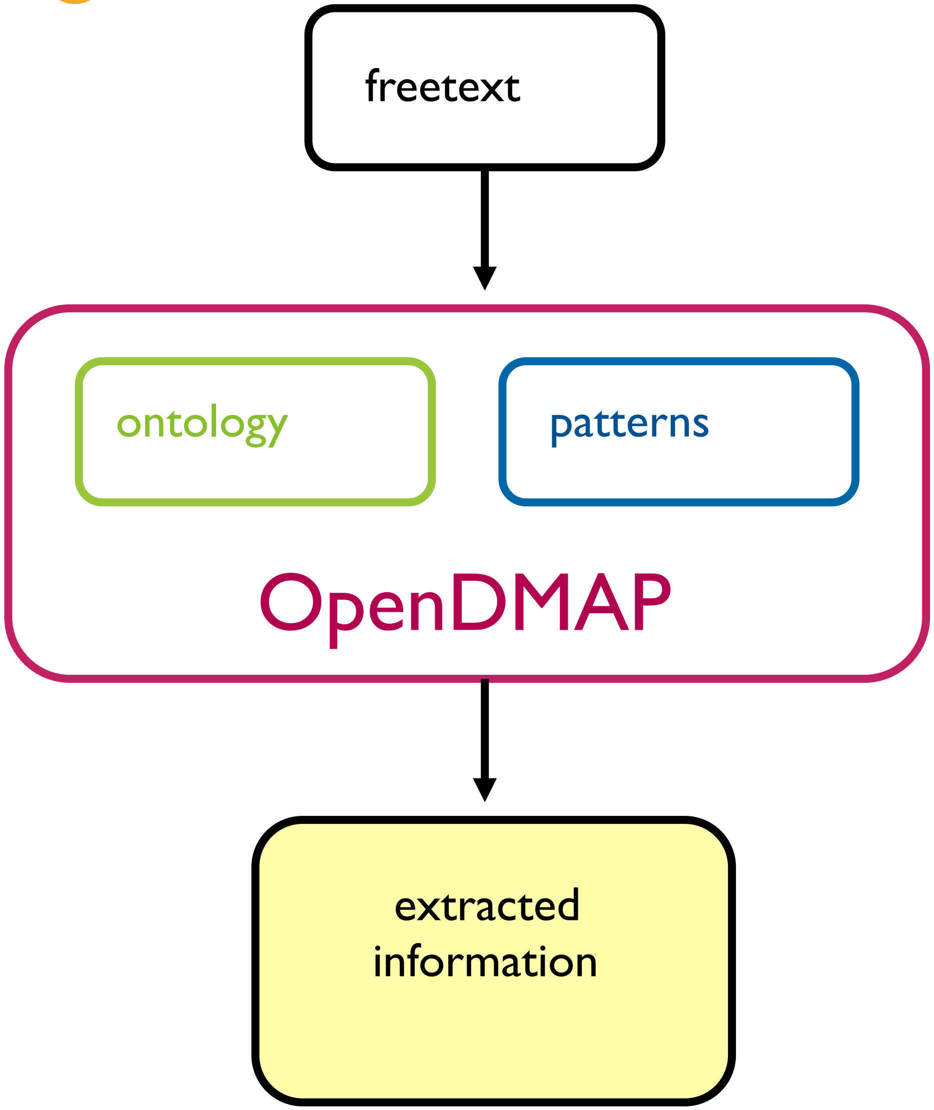
Template Merging

Transition 1
Start:
Pers: Mike Jones
Pos: President
Org: ABC Corp.
End:
Pers: John Smith
Pos: President
Org: ABC Corp.

OpenDMAP extracts typed relations from the literature

- Concept recognition tool
  - Connect ontological terms to literature instances
  - Built on Protégé knowledge representation system
- Language patterns associated with concepts and slots
  - Patterns can contain text literals, other concepts, constraints (conceptual or syntactic), ordering information, or outputs of other processing.
  - Linked to many text analysis engines via UIMA
- Best performance in BioCreative II IPS task
- >500,000 instances of three predicates (with arguments) extracted from Medline Abstracts
- [Hunter, et al., 2008] http://bionlp.sourceforge.net

OpenDMAP



freetext

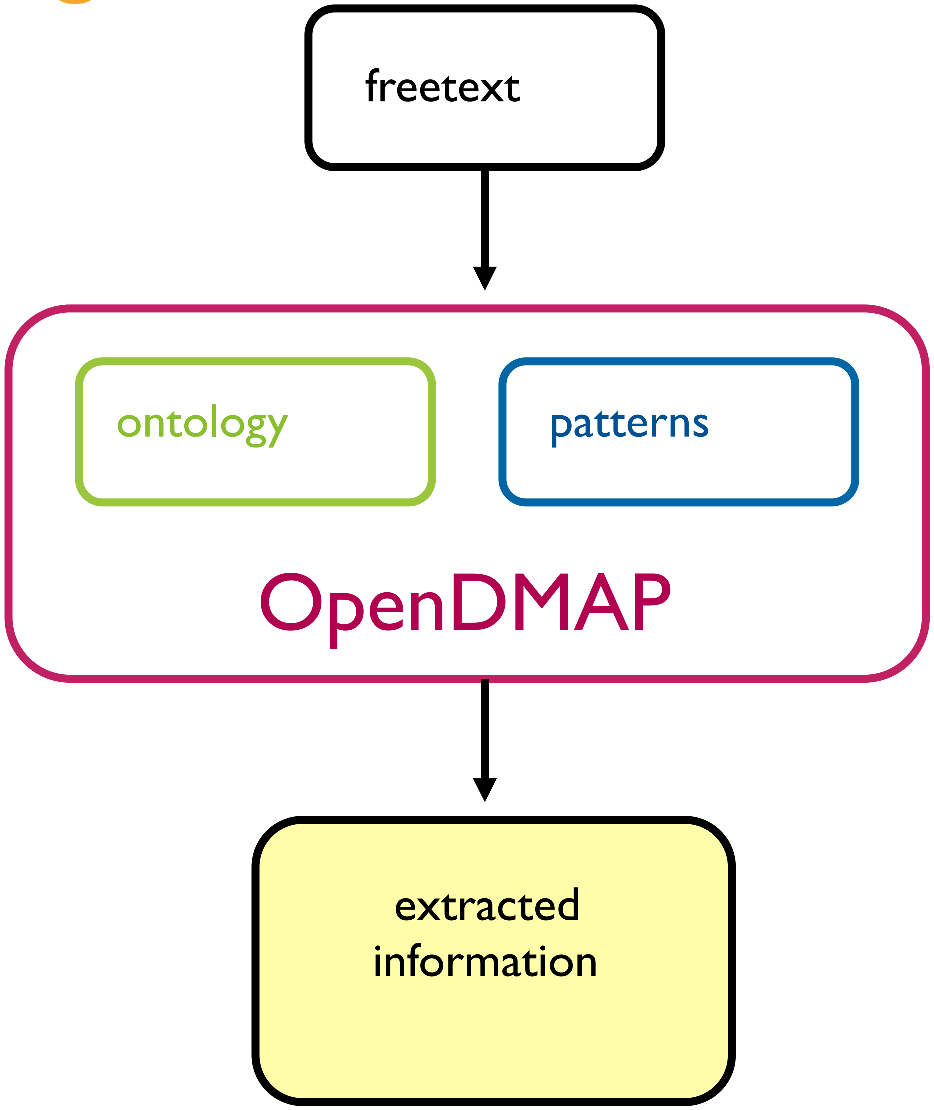
ontology

patterns

OpenDMAP

extracted information

OpenDMAP



freetext

ontology

patterns

OpenDMAP

extracted information

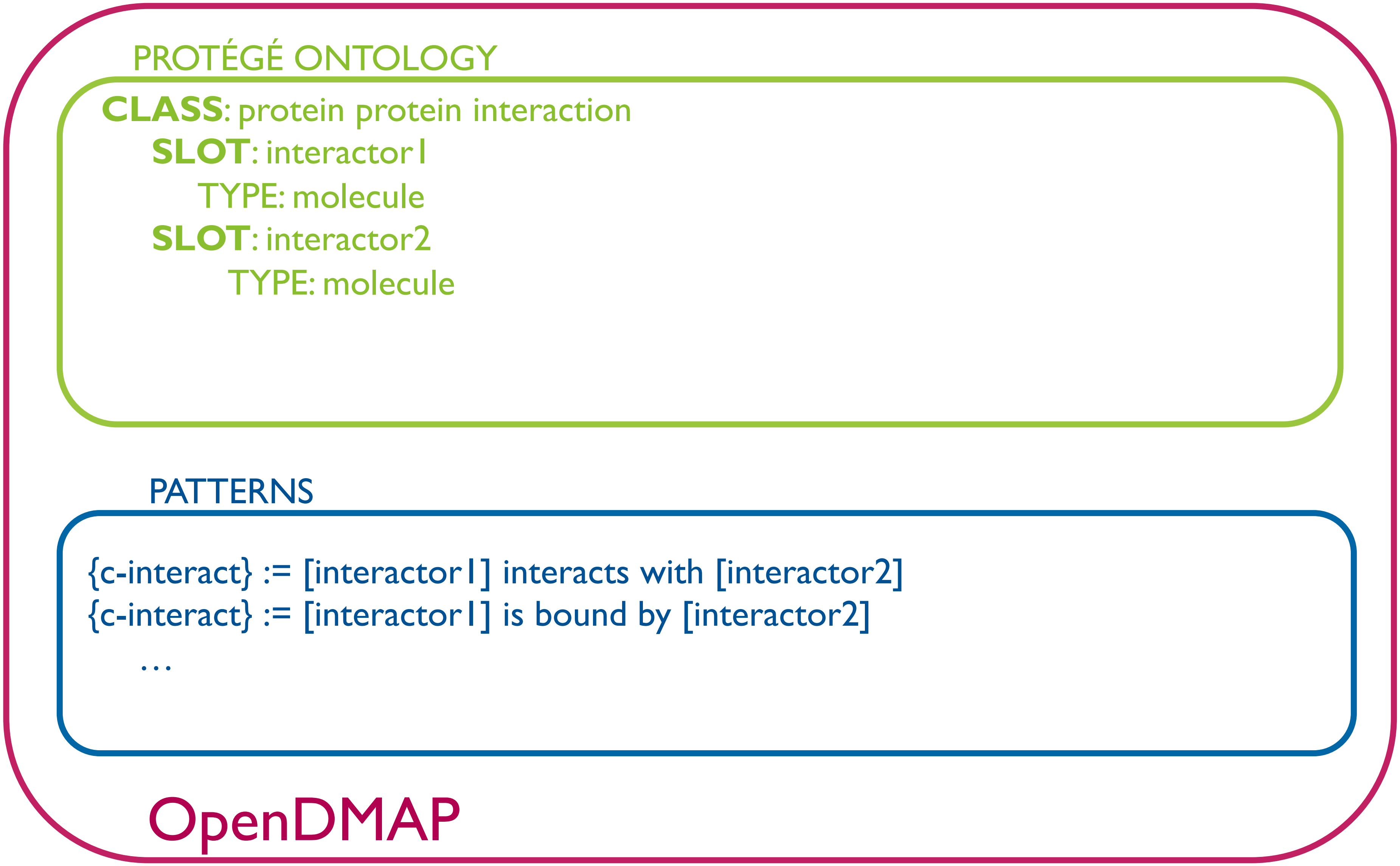
Cyclin E2 interacts with Cdk2 in a functional kinase complex.

<ontology>

Protein protein interaction := [int1] interacts with [int2]

protein protein interaction:
    interactor1: cyclin E2
    interactor2: cdk2

OpenDMAP



PROTÉGÉ ONTOLOGY

CLASS: protein protein interaction

SLOT: interactor1

TYPE: molecule

SLOT: interactor2

TYPE: molecule

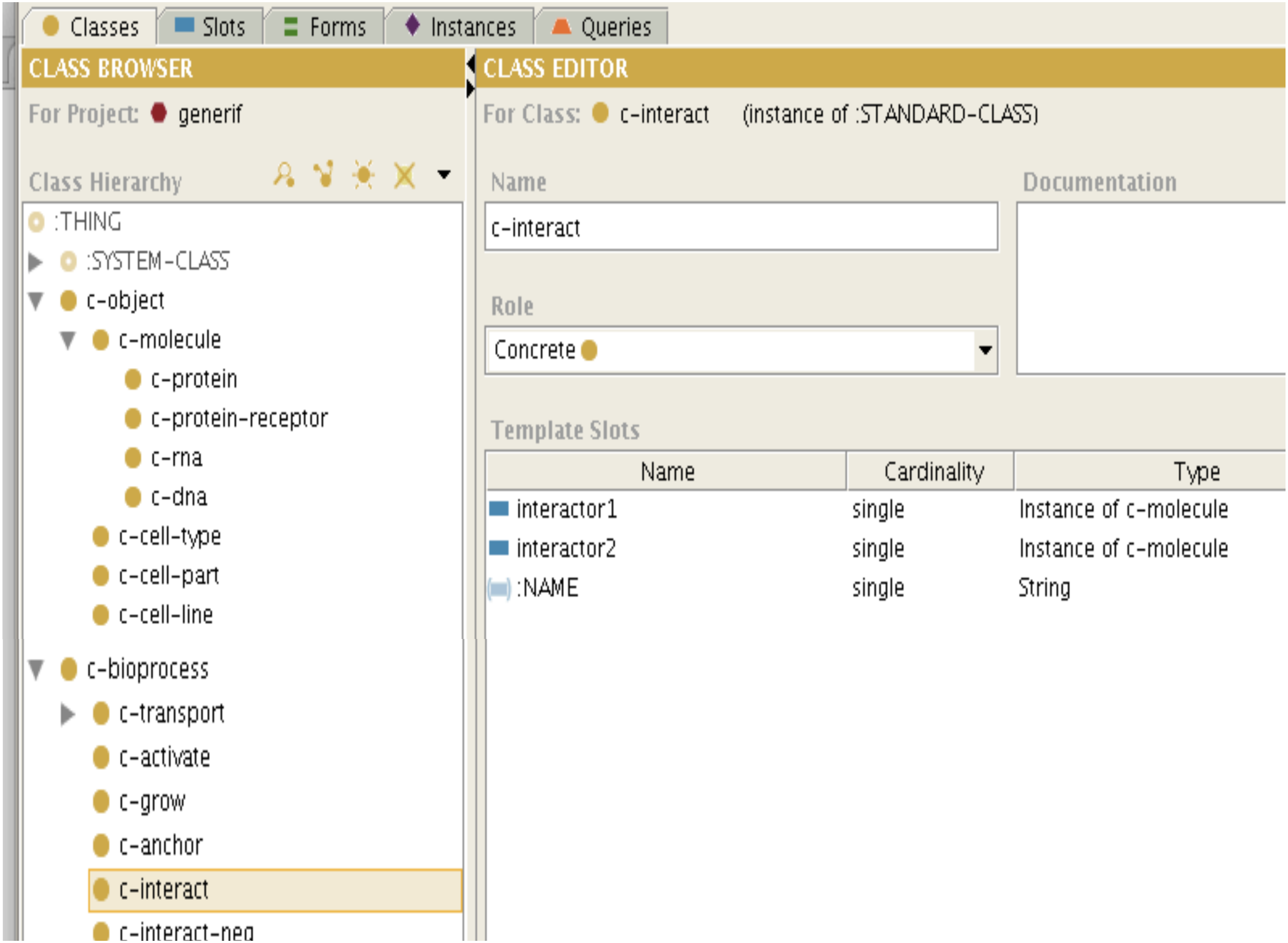
PATTERNS

{c-interact} := [interactor1] interacts with [interactor2]

{c-interact} := [interactor1] is bound by [interactor2]

. . .

OpenDMAP



For Project: generif

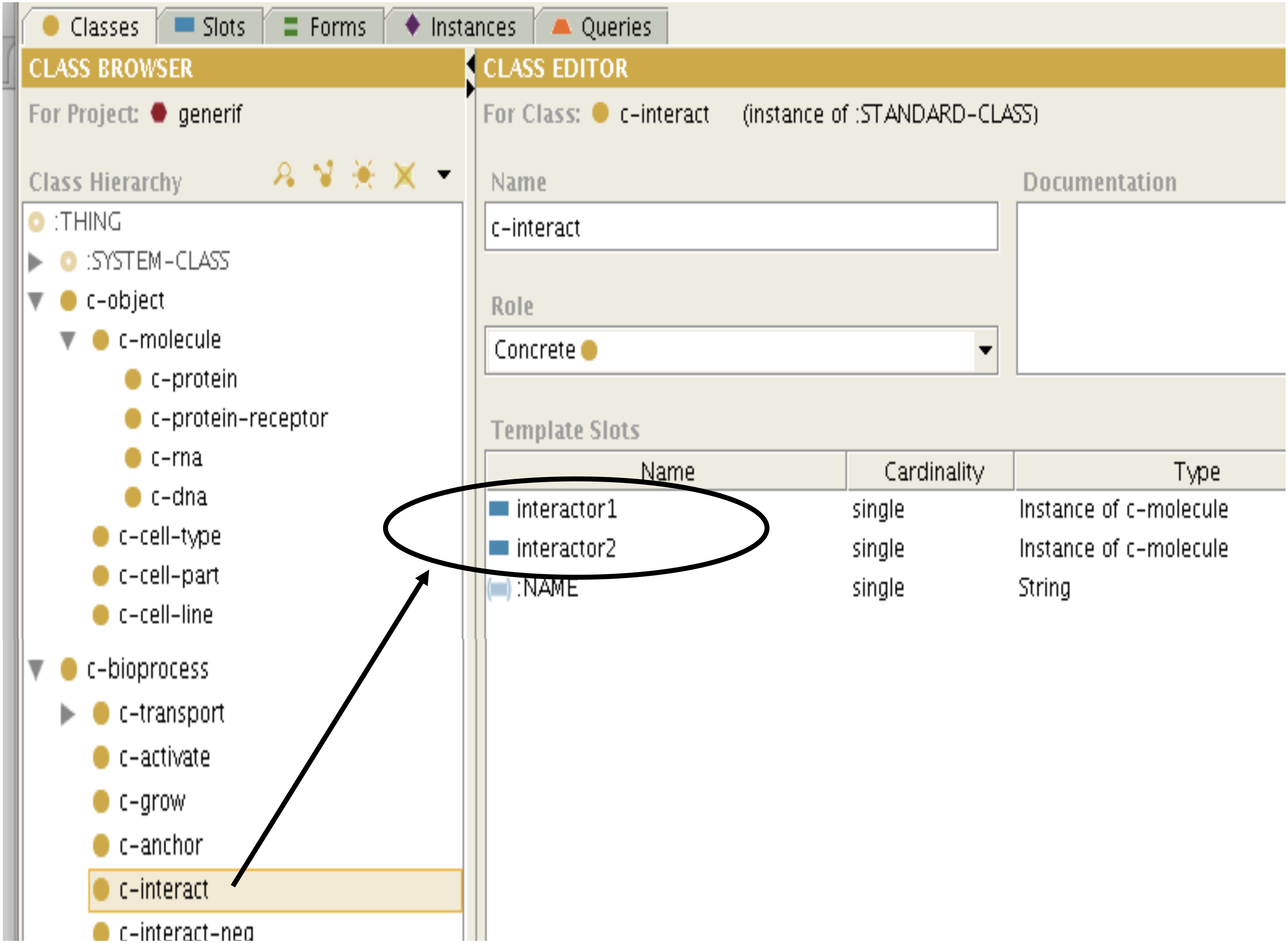
For Class: c-interact (instance of :STANDARD-CLASS)

- :THING
  - :SYSTEM-CLASS
  - c-object
    - c-molecule
      - c-protein
      - c-protein-receptor
      - c-rna
      - c-dna
    - c-cell-type
    - c-cell-part
    - c-cell-line
  - c-bioprocess
    - c-transport
    - c-activate
    - c-grow
    - c-anchor
    - c-interact
    - c-interact-neg

Name: c-interact

Role: Concrete

Name	Cardinality	Type
interactor1	single	Instance of c-molecule
interactor2	single	Instance of c-molecule
:NAME	single	String



CLASS BROWSER

For Project: ● generif

Class Hierarchy

- :THING
- ▶ :SYSTEM-CLASS
- ▼ c-object
  - ▼ c-molecule
    - c-protein
    - c-protein-receptor
    - c-rna
    - c-dna
  - c-cell-type
  - c-cell-part
  - c-cell-line
- ▼ c-bioprocess
  - ▶ c-transport
  - c-activate
  - c-grow
  - c-anchor
  - c-interact
  - c-interact-neg

CLASS EDITOR

For Class: ● c-interact (instance of :STANDARD-CLASS)

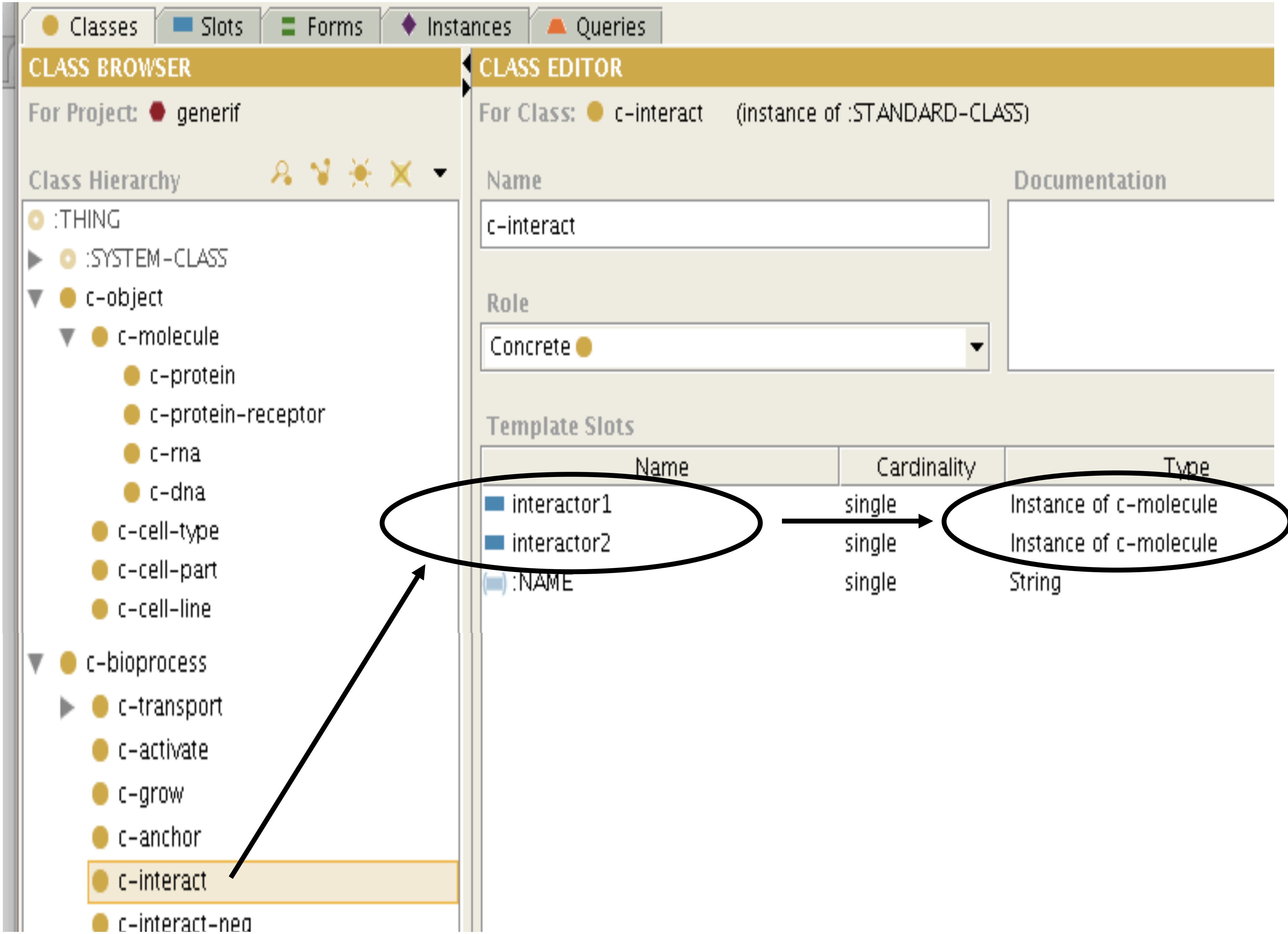
Name: c-interact

Documentation

Role: Concrete ●

Template Slots

Name	Cardinality	Type
interactor1	single	Instance of c-molecule
interactor2	single	Instance of c-molecule
:NAME	single	String



CLASS BROWSER

For Project: generif

Class Hierarchy

- :THING
- :SYSTEM-CLASS
- c-object
  - c-molecule
    - c-protein
    - c-protein-receptor
    - c-rna
    - c-dna
  - c-cell-type
  - c-cell-part
  - c-cell-line
- c-bioprocess
  - c-transport
  - c-activate
  - c-grow
  - c-anchor
  - c-interact
  - c-interact-neg

CLASS EDITOR

For Class: c-interact (instance of :STANDARD-CLASS)

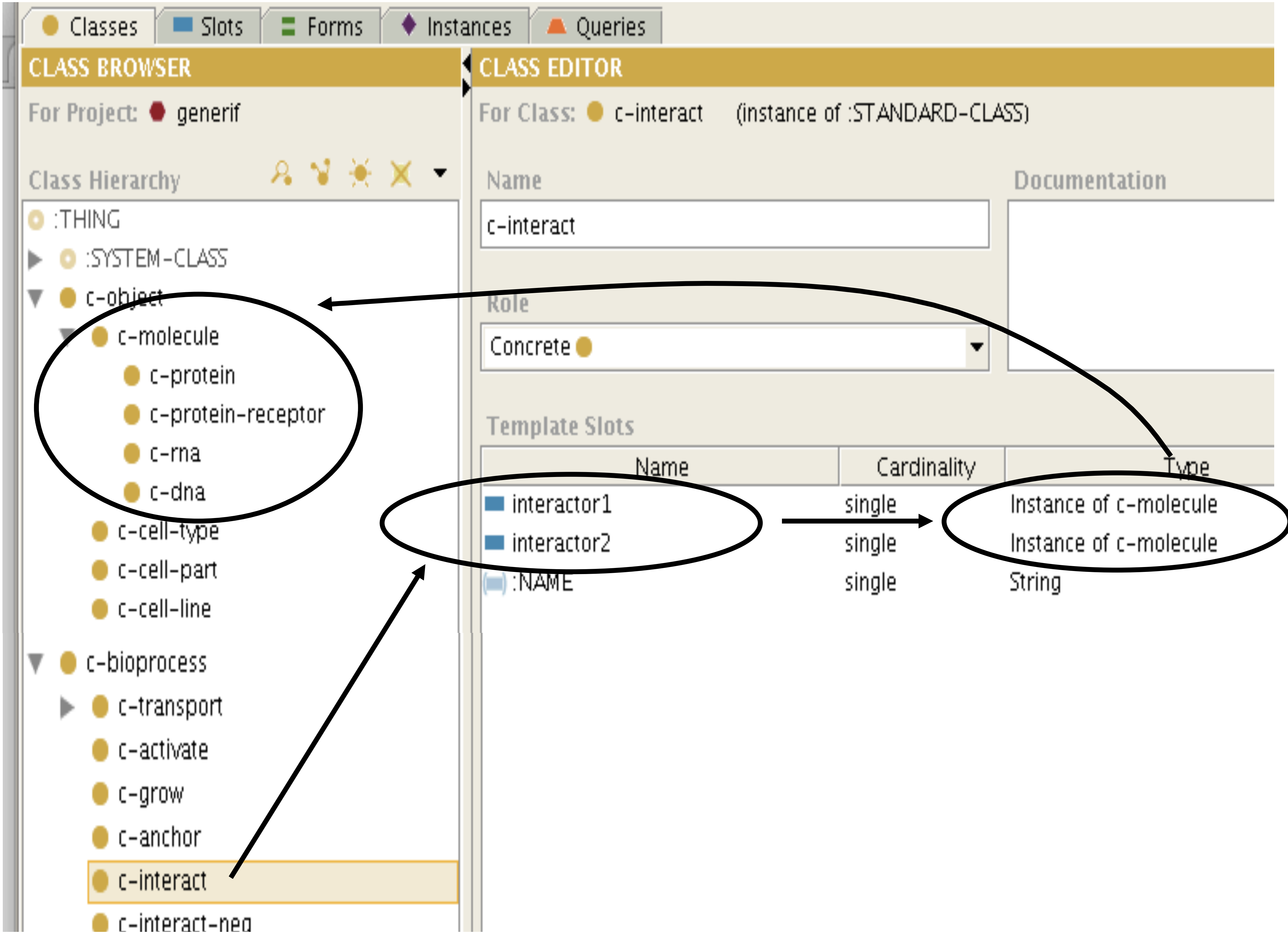
Name: c-interact

Documentation

Role: Concrete

Template Slots

Name	Cardinality	Type
interactor1	single	Instance of c-molecule
interactor2	single	Instance of c-molecule
:NAME	single	String



CLASS BROWSER

For Project: generif

Class Hierarchy

- :THING
  - :SYSTEM-CLASS
  - c-object
    - c-molecule
      - c-protein
      - c-protein-receptor
      - c-rna
      - c-dna
    - c-cell-type
    - c-cell-part
    - c-cell-line
  - c-bioprocess
    - c-transport
    - c-activate
    - c-grow
    - c-anchor
    - c-interact
    - c-interact-neg

CLASS EDITOR

For Class: c-interact (instance of :STANDARD-CLASS)

Name: c-interact

Documentation

Role: Concrete

Template Slots

Name	Cardinality	Type
interactor1	single	Instance of c-molecule
interactor2	single	Instance of c-molecule
:NAME	single	String

BioCreative Example

- Some BioCreative patterns for *interact*

  {c-interact} := **[interactor1]** {w-is} {w-interact-verb1} {w-preposition} the? **[interactor2]**;

  {w-is} := is, are, was, were;

  {w-interact-verb1} := co-immunoprecipitate, co-immunoprecipitates, co-immunoprecipitated, co-localize, co-localizes, co-localized;

  {w-preposition} := among, between, by, of, with, to;

- Matched text:

  PMID 16494873, SENT_ID 16494873_114

  Upon precipitation of the SOX10 protein with anti-HA antibody, Western blot detection revealed expression of UBC9-V5 (25 kDa) in the sample (Fig. 1, line 6), indicating that **{UBC9** was co-immunoprecipitated with **SOX10}**.

  INTERACTOR_1: UBC9 resolved to UniprotID: UBC9_RAT

  INTERACTOR_2: SOX10 resolved to UniProtID: SOX10_RAT

  {c-interact} := **[UBC9_RAT]**$_{\text{interactor\_1}}$, **[SOX10_RAT]**$_{\text{interactor\_2}}$

BioCreative Results

- 359 full-text articles in the test set
- 385 interaction assertions produced
- Performance averaged per article (to avoid dominance of a few assertion-heavy articles)

P = 0.39, R = 0.31, F = 0.29

- Best result in the evaluation!
  - F score 10% higher than next-scoring system
  - F score > 3 standard deviations above mean
  - Recall 20% higher than next-scoring system

BioCreative conclusions

- Information extraction in biomedical text is hard
  - Linguistic variability in how concepts are expressed
  - Complex concepts with multiple “slots”

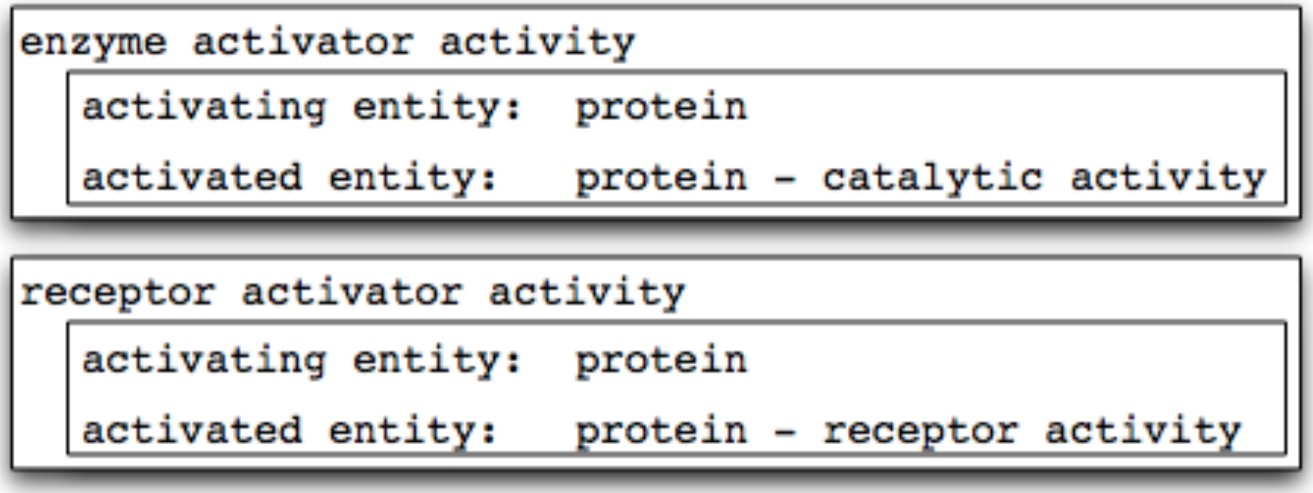
- OpenDMAP advances the state of the art
  - Use of an ontology grounds the search for information
  - Flexibility of the pattern language to incorporate constraints at different levels (conceptual, lexical, word order, linguistic)

Integrating background knowledge

- Can improve OpenDMAP precision with minimal cost to recall
  - Take advantage of background knowledge
  - Tighten constraints on slot fillers in the ontology
  - No change to existing patterns
- Proof of concept:
  - Distinguish among several types of protein activation (enzyme and receptor) in GeneRIFs
  - Utilize Gene Ontology annotations

Livingston, K., Johnson, H., Verspoor, K., Hunter, L. (submitted). "Leveraging Gene Ontology Annotations to Improve a Memory-Based Language Understanding System".

Refining selectional restrictions



```
enzyme activator activity
    activating entity:   protein
    activated entity:    protein - catalytic activity
```

```
receptor activator activity
    activating entity:   protein
    activated entity:    protein - receptor activity
```

TP: [GeneRIF 104155]

an ER stress induces the activation of [caspase-12_protein - catalytic activity]activated_entity via [caspase-3_protein]activator

prevented FP: [GeneRIF 105594]

factor Xa can induce mesangial cell proliferation through the activation of ERK_protein via PAR2_protein in mesangial cells

Results

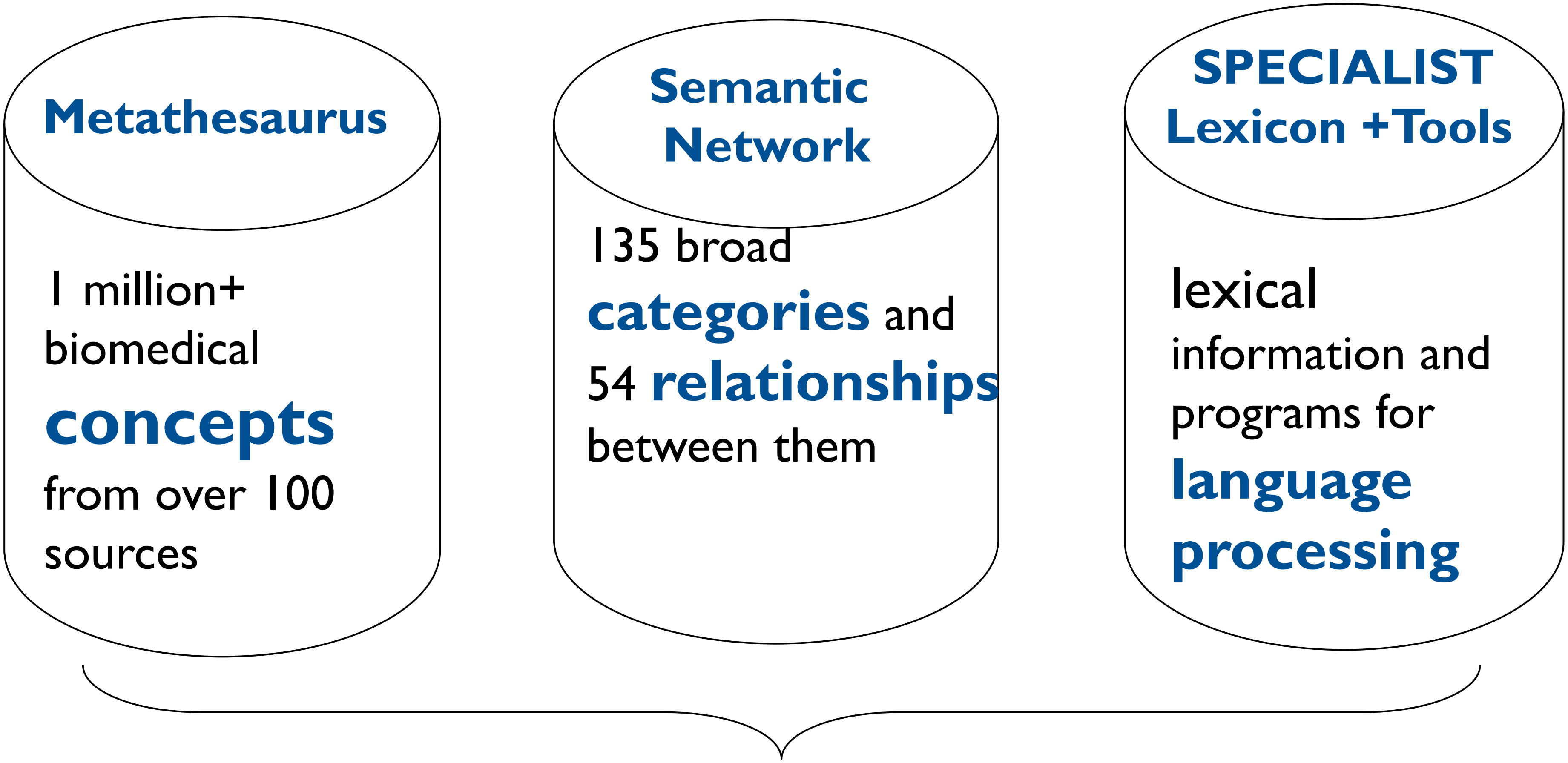
		Original	Additional Memory	Difference
Enzyme Events	Precision	0.24	0.37	0.13
	Recall	0.27	0.20	-0.07
	F-measure	0.26	0.26	0.00
Receptor Events	Precision	0.08	0.34	0.26
	Recall	0.17	0.12	-0.05
	F-measure	0.11	0.18	0.07
Total	Precision	0.16	0.36	**0.20**
	Recall	0.24	0.18	**-0.06**
	F-measure	0.19	0.24	**0.05**

LEXICAL RESOURCES for BioNLP

The importance of lexical resources

- Need to characterize the linguistic behavior of terms to establish word meaning in context
  - Morphosyntactic behavior
    - inflectional patterns
    - part of speech
    - argument structure
  - Semantic information
- Need to recognize different terms that express the same or closely related meanings
  - To support database integration, multi-database querying
  - To enable generalization of information extraction templates
  - To support general text understanding and meaning analysis (e.g. semantic reasoning over text or during text processing)

Unified Medical Language System



Metathesaurus

1 million+ biomedical **concepts** from over 100 sources

Semantic Network

135 broad **categories** and 54 **relationships** between them

SPECIALIST Lexicon +Tools

lexical information and programs for **language processing**

3 Knowledge Sources

used separately or together

Metathesaurus

- 100+ general and specialized biomedical vocabularies
- 17 languages (63% English)

- 1 million+ concepts; 6 million+ names
- 100K+ relationships (hierarchical, semantic, statistical and mapping relationships)

- Distributed in a common electronic format

Metathesaurus Concepts

- Synonymous terms clustered into a concept
- Unique identifier (CUI) is assigned
- Source information preserved

Addison's disease	**SNOMED CT**	PT	363732003
Addison's Disease	**MedlinePlus**	PT	T1233
Addison Disease	**MeSH**	PT	D000224
Primary Adrenal Insufficiency	**MeSH**	EN	D000224
Primary hypoadreanlism syndrome, Addison	**MedDRA**	LT	10036696

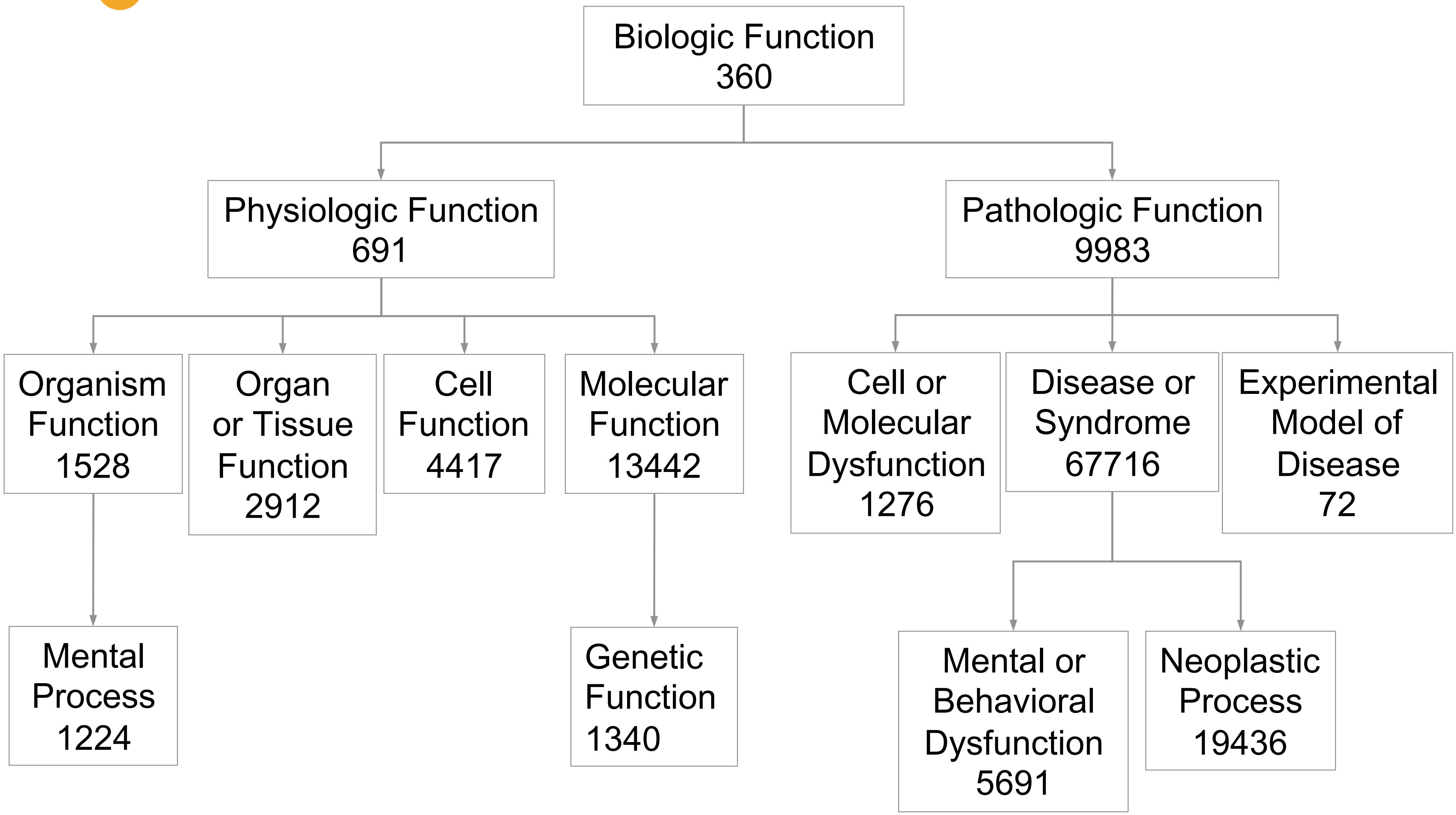
C0001403

Addison's disease

Semantic Network

- 135 Semantic Types
  - Broad subject categories in 2 hierarchies
  - Assigned to all Metathesaurus concepts
- 54 Semantic Relationships
  - Useful, important links between Types
  - Hierarchical "isa" and associative relations
- Categorize the Metathesaurus
- Enhance meaning of concepts

“Biologic Function” hierarchy



Biologic Function 360

Physiologic Function 691

Pathologic Function 9983

Organism Function 1528

Organ or Tissue Function 2912

Cell Function 4417

Molecular Function 13442

Cell or Molecular Dysfunction 1276

Disease or Syndrome 67716

Experimental Model of Disease 72

Mental Process 1224

Genetic Function 1340

Mental or Behavioral Dysfunction 5691

Neoplastic Process 19436

Semantic Relations

- Disease or Syndrome associated_with Finding

- Disease or Syndrome result_of Pathologic Function

- Body Part, Organ, or Organ Component location_of Disease or Syndrome

- Hormone affects Disease or Syndrome

  Hormone causes Disease or Syndrome

  Hormone complicates Disease or Syndrome

SPECIALIST Lexicon

- English lexicon of 300K+ common words and biomedical terms
- Lexical records encode information on:
  - Syntax
  - Morphology
  - Orthography
- Used with associated lexical tools
  - in Metathesaurus production
  - in natural language processing applications

SPECIALIST Lexical Entry

```
{base=disease
    entry=E0023270
       cat=noun
       variants=reg
       variants=uncount
       compl=pphr(of,np|bone|)
       compl=pphr(of,np|breast|)
       compl=pphr(of,np|liver|)
       compl=pphr(of,np|ovary|)}
```

Base form

Unique identifier

Part of speech

Lexical variants

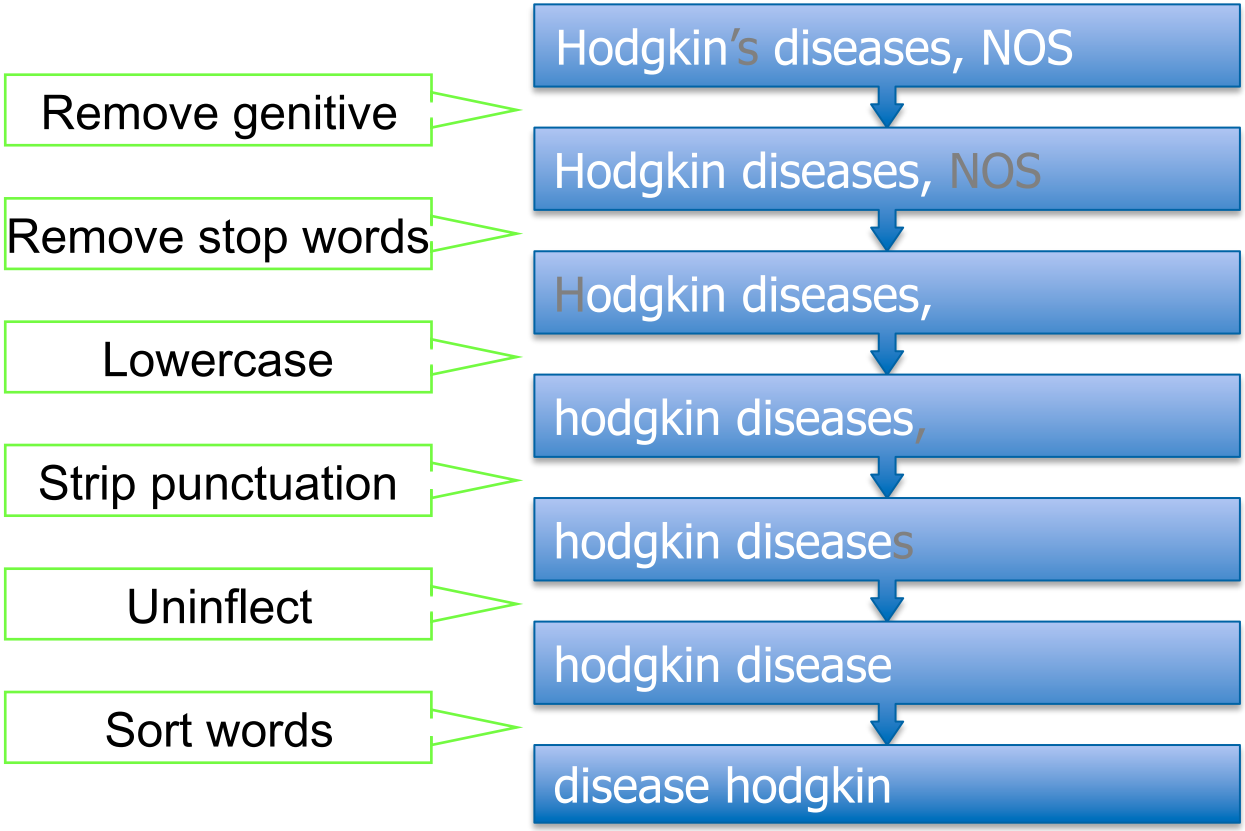
Prepositional phrase complements

Lexical Tools

- Manage lexical variation in biomedical terminologies and text
- Used separately or with SPECIALIST Lexicon
- Perform transformations selected and ordered by users
- 3 primary programs: normalizer, word index generator, lexical variant generator

http://umlslex.nlm.nih.gov/lvg/current/

Normalization I



Hodgkin’s diseases, NOS

Remove genitive

Hodgkin diseases, NOS

Remove stop words

Hodgkin diseases,

Lowercase

hodgkin diseases,

Strip punctuation

hodgkin diseases

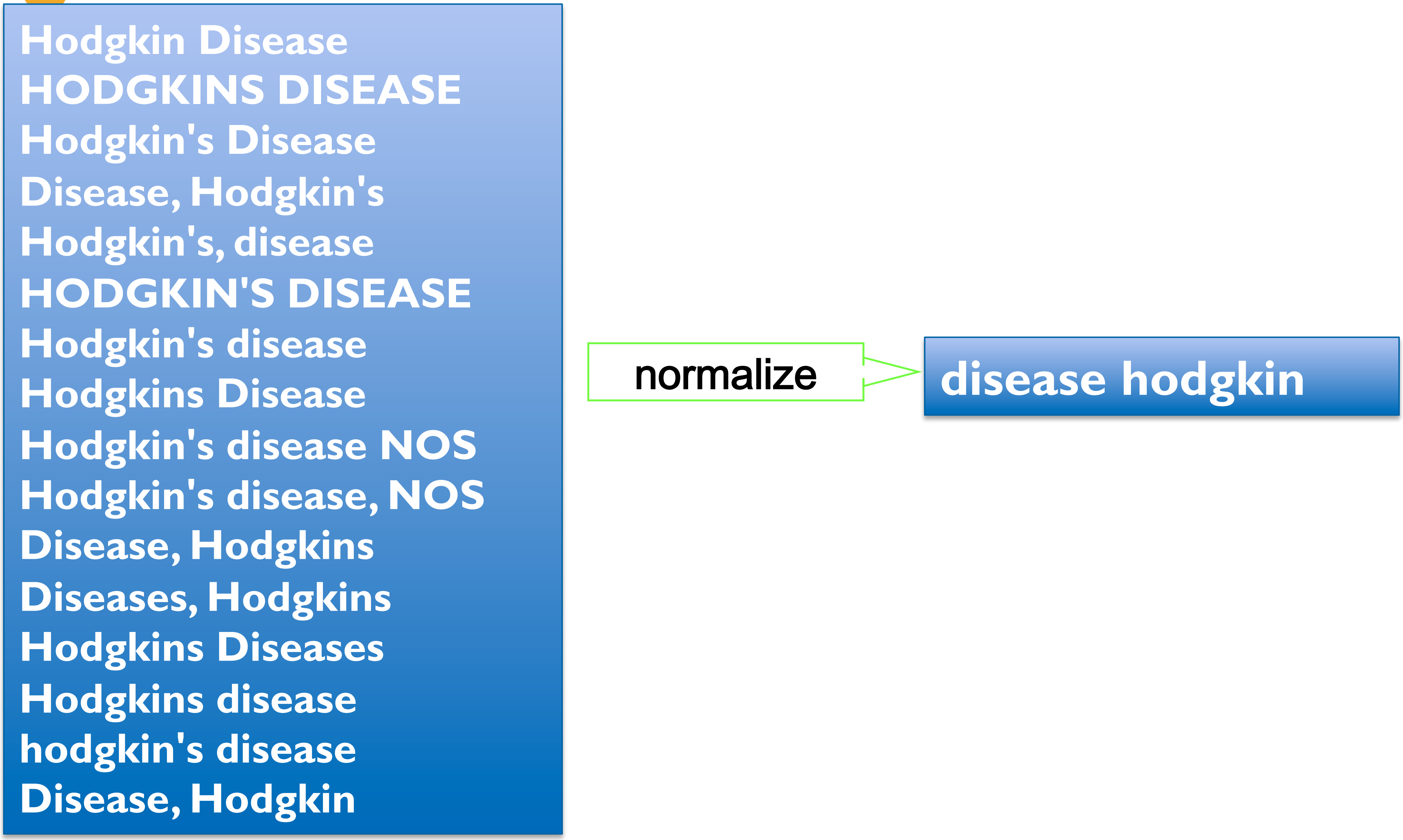
Uninflect

hodgkin disease

Sort words

disease hodgkin

Normalization 2



Hodgkin Disease
HODGKINS DISEASE
Hodgkin's Disease
Disease, Hodgkin's
Hodgkin's, disease
HODGKIN'S DISEASE
Hodgkin's disease
Hodgkins Disease
Hodgkin's disease NOS
Hodgkin's disease, NOS
Disease, Hodgkins
Diseases, Hodgkins
Hodgkins Diseases
Hodgkins disease
hodgkin's disease
Disease, Hodgkin

normalize

disease hodgkin

BioFrameNet

Andrew Dolbey, PhD dissertation, 2009, BioFrameNet: a FrameNet Extension to the Domain of Molecular Biology

Transport Valence Pattern		
Transport_destination	**Transported_Entity**	**Transporting_Entity**
PP[to]	NP	CNI
Dep	Obj	--

GRIF 82174 EntrezGene ID: 66013 [symbol: Arhgef9] PMID: 1521530

TRANSLOCATES gephyrin to submembrane microaggregates CNI

Transport Valence Pattern		
Transport_destination	**Transported_Entity**	**Transporting_Entity**
PP[to]	NP	CNI
Dep	Ext (subj)	--

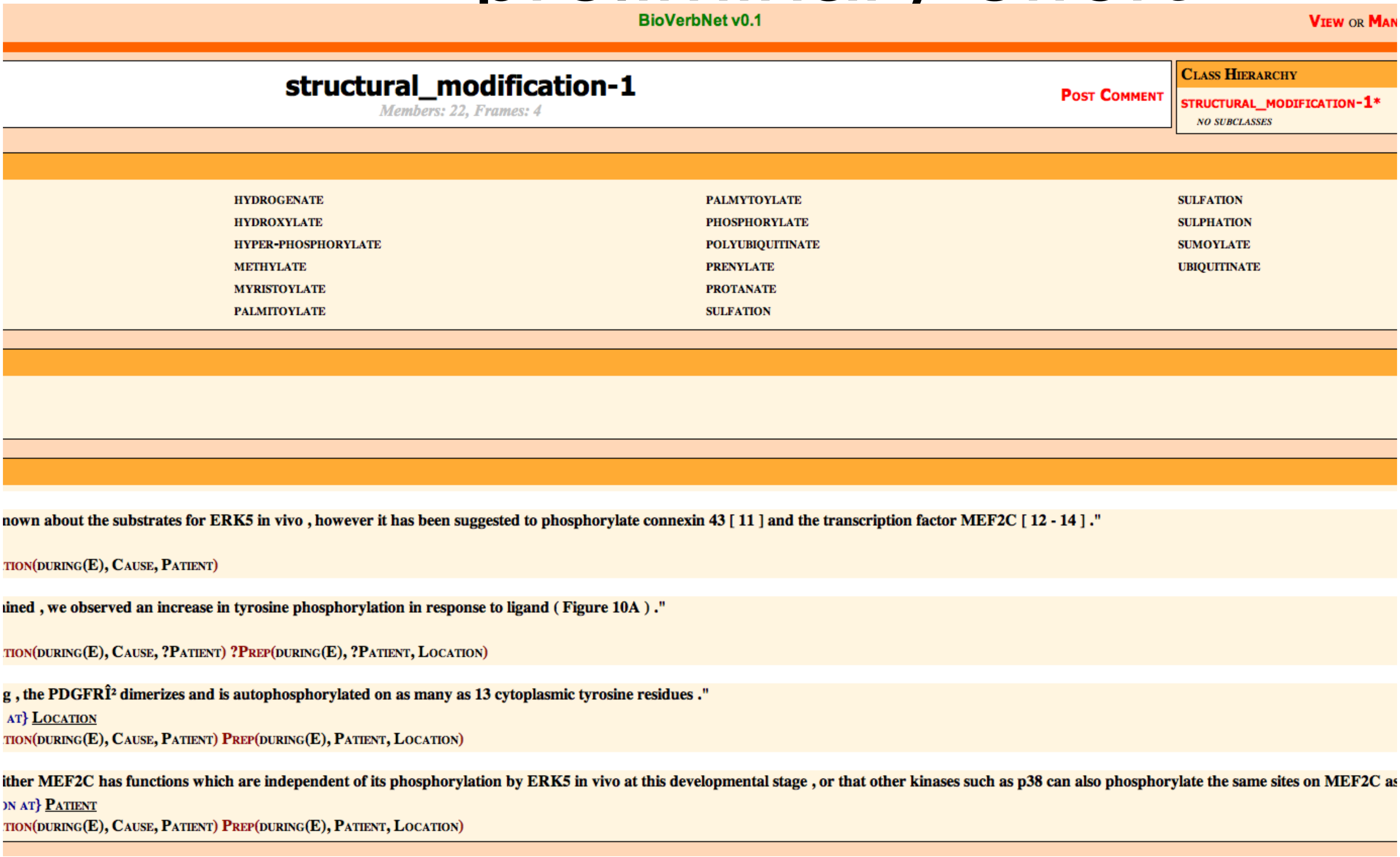
GRIF 72788 EntrezGene ID: 654817 [symbol: NCF1C] PMID: 1285569

p47phox is TRANSLOCATED to membrane ruffles through a VEGF-WAVE1 pathway CNI

ELDA BioLexicon

- Subcategorization frames + Event frames
  - Roles: agent, theme, manner, instrument, destination, condition, rate, descriptive agent, descriptive theme, purpose, *location*, *temporal*
  - Used annotations in Gene Regulation corpus and thematic hierarchies to guide linking
  - Result: 668 event frames for 168 verbs

BioVerbNet preliminary effort



BioVerbNet v0.1

structural_modification-1

Members: 22, Frames: 4

Class Hierarchy

structural_modification-1*

NO SUBCLASSES

HYDROGENATE
HYDROXYLATE
HYPER-PHOSPHORYLATE
METHYLATE
MYRISTOYLATE
PALMITOYLATE

PALMYTOYLATE
PHOSPHORYLATE
POLYUBIQUITINATE
PRENYLATE
PROTANATE
SULFATION

SULFATION
SULPHATION
SUMOYLATE
UBIQUITINATE

nown about the substrates for ERK5 in vivo , however it has been suggested to phosphorylate connexin 43 [11] and the transcription factor MEF2C [12 - 14] ."

TION(DURING(E), CAUSE, PATIENT)

ined , we observed an increase in tyrosine phosphorylation in response to ligand (Figure 10A) ."

TION(DURING(E), CAUSE, ?PATIENT) ?PREP(DURING(E), ?PATIENT, LOCATION)

g , the PDGFRÎ² dimerizes and is autophosphorylated on as many as 13 cytoplasmic tyrosine residues ."
AT} LOCATION
TION(DURING(E), CAUSE, PATIENT) PREP(DURING(E), PATIENT, LOCATION)

ither MEF2C has functions which are independent of its phosphorylation by ERK5 in vivo at this developmental stage , or that other kinases such as p38 can also phosphorylate the same sites on MEF2C as
ON AT} PATIENT
TION(DURING(E), CAUSE, PATIENT) PREP(DURING(E), PATIENT, LOCATION)

BioLemmatizer

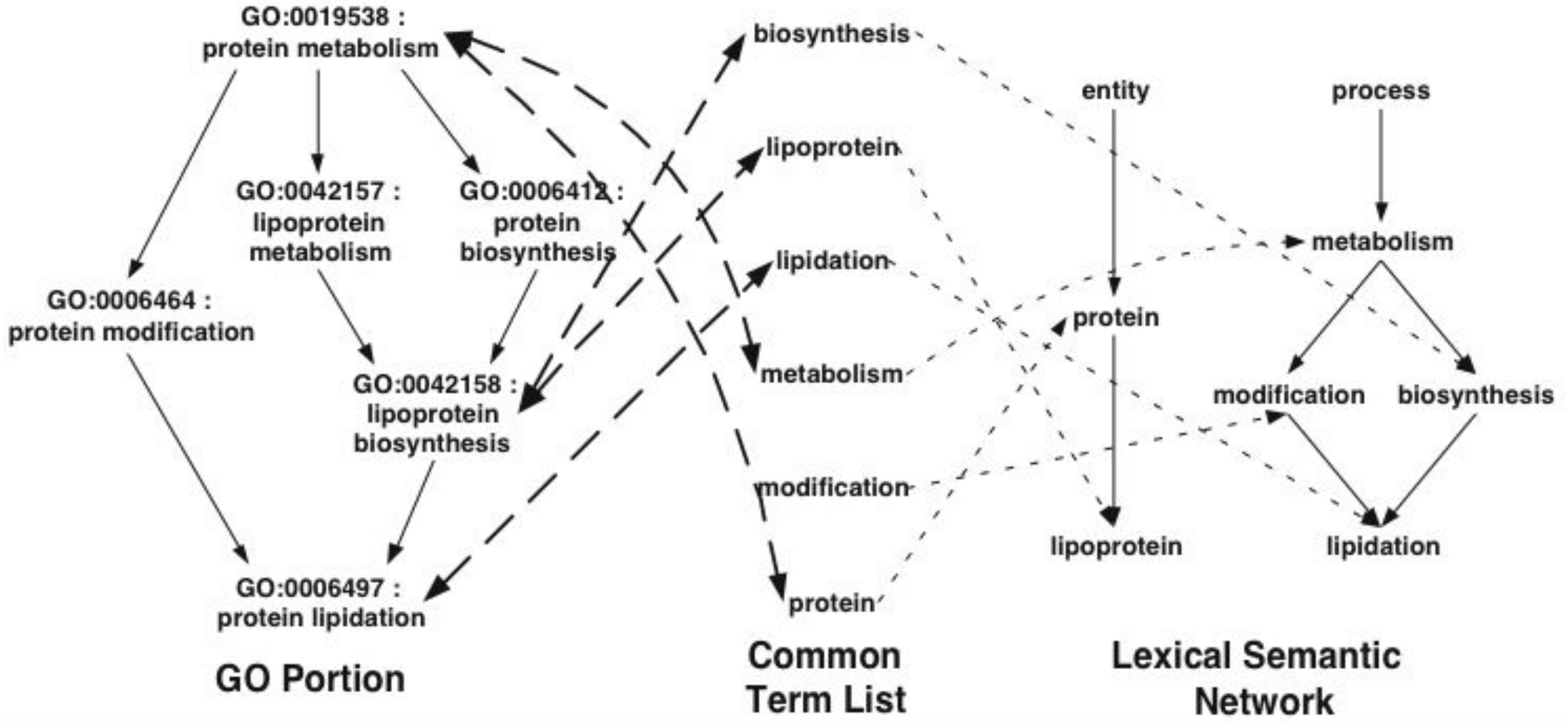
- New tool under development in our group
- Based on the MorphAdorner tool, using specialized lexicons

- Given {token, POS}
- Produce base form for token

- (default behavior for token w/out POS)

GO as a lexical semantic resource

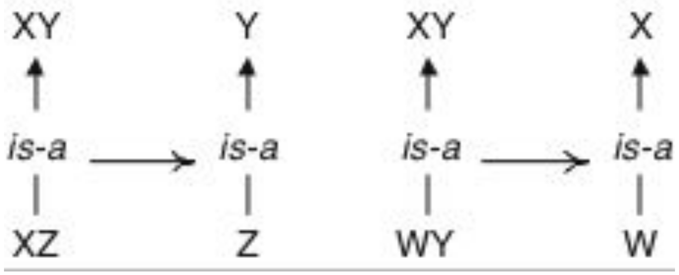
- The Gene Ontology represents semantic relationships (is_a, part_of) between biological phrases representing molecular functions/processes
- Utilize the structure of the GO and lexical correspondences to infer relationships at the term level from relationships between phrases



GO:0019538 : protein metabolism
GO:0042157 : lipoprotein metabolism
GO:0006412 : protein biosynthesis
GO:0006464 : protein modification
GO:0042158 : lipoprotein biosynthesis
GO:0006497 : protein lipidation
GO Portion
biosynthesis
lipoprotein
lipidation
metabolism
modification
protein
Common Term List
entity
process
metabolism
protein
modification
biosynthesis
lipoprotein
lipidation
Lexical Semantic Network

Verspoor, C., C. Joslyn and G. Papcun (2003). "The Gene Ontology as a Source of Lexical Semantic Knowledge for a Biological Natural Language Processing Application". In Proceedings of the SIGIR'03 Workshop on Text Analysis and Search for Bioinformatics.

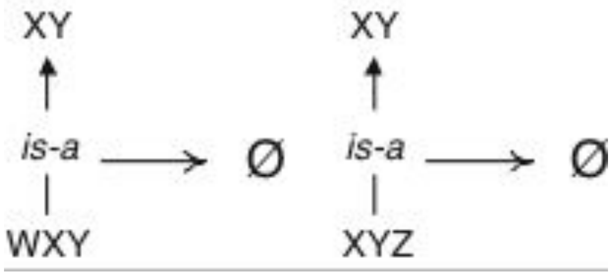
Inferring Lexical Relations from GO



Parallel rule:

vanillin metabolism *isa* aldehyde metabolism ⇒
vanillin isa aldehyde

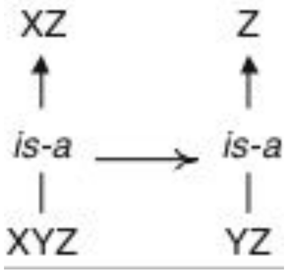
lipoprotein biosynthesis *isa* lipoprotein metabolism ⇒
biosynthesis isa metabolism



Modifier rule: blocking rule for modifiers

Positive gravitactic behavior *isa* gravitactic behavior ⇒ Ø

Larval feeding behavior (sensu insecta) *isa* Larval feeding behavior ⇒ Ø



Insertion rule: right-branching heuristic

adult feeding behavior *isa* adult behavior ⇒
feeding behavior *isa* behavior

chemosensory jump behavior *isa* chemosensory behavior
⇒jump behavior *isa* behavior

Verspoor et al. (2003)

Relations inferred (with counts)

581	biosynthesis isa metabolism
577	catabolism isa metabolism
44	receptor isa binding
38	deoxyribonucleoside isa nucleoside
35	ribonucleoside isa nucleoside
33	permease isa transporter
27	Saccharomyces isa Fungi
22	porter isa transporter
15	oxidation isa metabolism
14	tRNA isa RNA

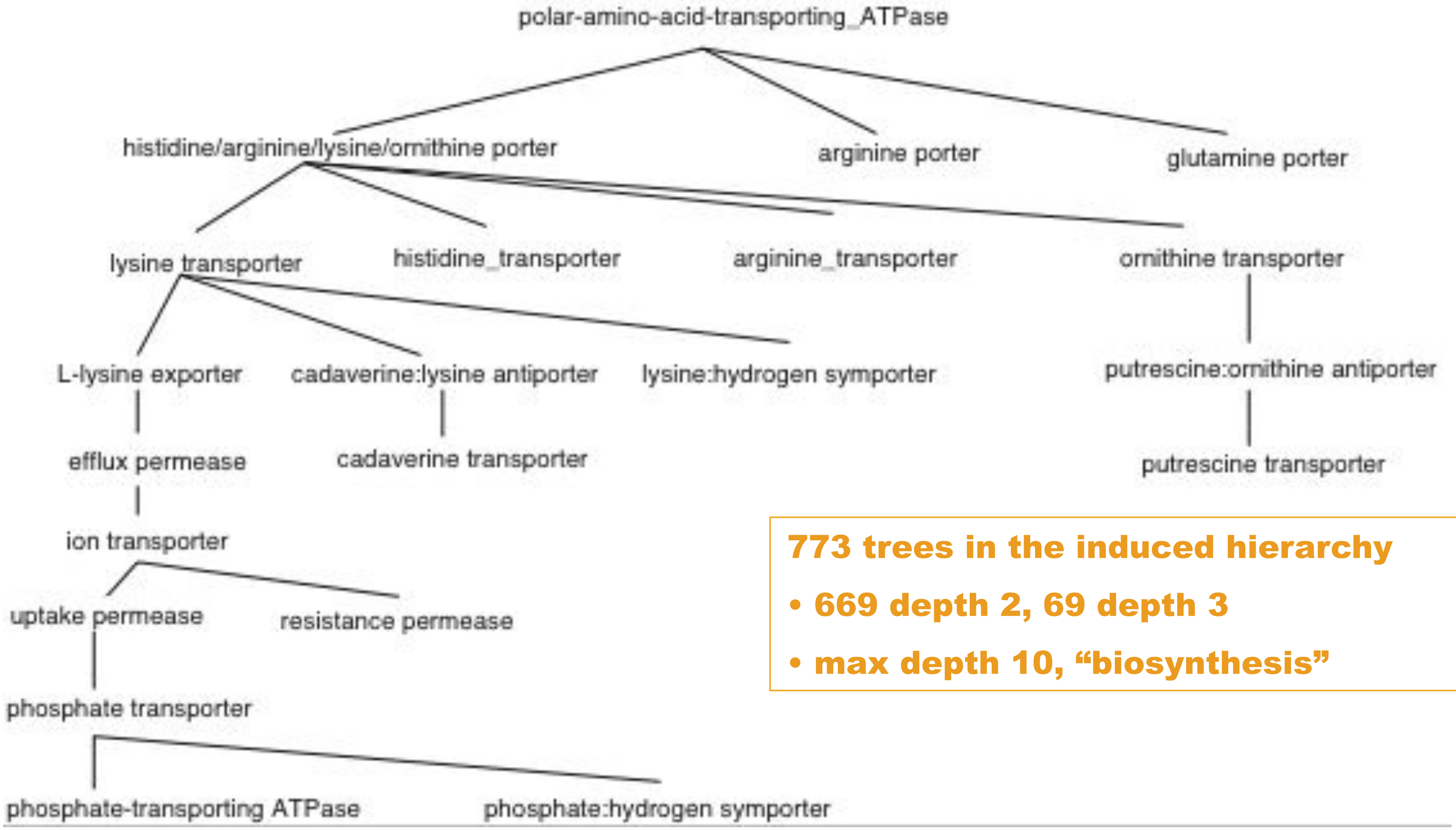
14	inhibitor isa regulator
13	ribonucleotide isa nucleotide
11	proliferation isa activation
11	differentiation isa activation
11	deoxyribonucleotide isa nucleotide
10	rRNA isa RNA
10	mRNA isa RNA
9	snRNA isa RNA
8	modification isa metabolism
8	methylation isa modification

6,364 unique relations inferred; only 70 already exist in the GO

3,270/6,589 unique labels inferred that do not occur in the GO as terms

Verspoor et al. (2003)

A portion of the induced network



773 trees in the induced hierarchy

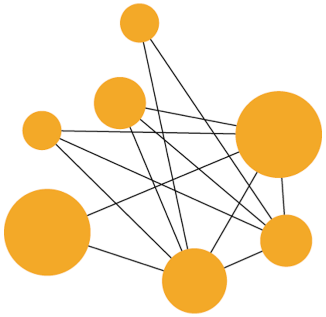
- **669 depth 2, 69 depth 3**
- **max depth 10, “biosynthesis”**

Verspoor et al. (2003)

Acknowledgements

- Kevin B. Cohen
- NIH/NLM and NSF grants to me and to Larry Hunter
- Other members of the Center for Computational Pharmacology at University of Colorado Anschutz Medical Campus (Denver)

- THANK YOU!



- disk drive (was) down (at) 11/16-2305.
- (has) select lock.
- spindle motor is bad.
- (is) awp spindle motor.
- (disk drive was) up (at) 11/17-1236.
- replaced spindle motor.

Biomedical verb semantics

- Semantic Network (NLM)
  - Can relate objects in an ontology
- Friedman et al. (2002)
  - Complex embedding
- McDonald et al. (2005)
  - Arity > binary
- LSAT
  - Applied PASBio PAS to information extraction
- Kogan et al. (2005)
  - General & medical domains require different PAS

Some recent history

- 2004: PASBio publication (BMC Bioinformatics)
- 2005: Extension of PASBio to medical predicates (Kogan et al., AMIA)