

Named Entity Recognition

Extracting Protein Names from Text



Ulf Leser
Knowledge Management in Bioinformatics
Humboldt-Universität Berlin



Case Report

- Patient with pneumonia and cough
- Normal dosage of codeine
- Patient not responding any more at day 4
- What's going on?
 - PubMed „Codeine intoxication“ -> 170 abstracts
 - Aren't there better ways?

Case report from Univ. Hospital Geneva, thanks to Christian Meisel, Roche

Ali Baba

File Preferences View Help

PubMed query: codeine intoxication

Go Max. results 50

Information

Objects Texts

- CYP2D6
 - Species (2)
 - Diseases (1)
- D(2)
- IMP
- Mat
- Monoamine oxidase
- OXY
- SRI
- TNS
- tryptase
- Drugs (22)
 - acetylcysteine
 - alcohol
 - amitriptyline
 - barbiturates

Proteins: [CYP2D6](#)

Textual Evidence

[PMID:15625333](#)

Codeine intoxication associated with ultrarapid **CYP2D6** metabolism.

Codeine is bioactivated by **CYP2D6** into **morphine**, which then undergoes further glucuronidation.

Tree: Feedback mode

Edge length: Confidence: Degree: Search: codeine

The network diagram illustrates the central role of codeine and its conversion to morphine. Codeine is linked to numerous clinical terms such as 'pain', 'overdose', 'pneumonia', 'Substance Abuse', 'death', 'body weight', 'abnormalities', 'injury', 'drowning', 'dextromethorp...', 'Serotonin syn...', 'fluvoxamine dermatitis', 'cough', 'propoxyphene', 'fever', 'toxic epiderm...', 'methadone', 'drunkenness', 'eczema', 'benzodiazepines', 'syndrome', 'barbiturates', 'cocaine abuse', 'pram', 'cocaine', 'Emergency', 'acetylparacetamol', 'lidocaine', 'pain', 'TNS', 'cabergoline', 'neurolo', 'patients', 'diphenhydramine', 'zolidem', 'OXY', 'jaundice', 'tryptase', 'D(2)', 'Mat', 'amitriptyline', 'morphine', 'CYP2D6', 'tryptase', 'D(2)', 'IMP', 'Mat', 'Monoamine oxidase', 'OXY', 'SRI', 'TNS', 'tryptase', 'Drugs (22)', 'acetylcysteine', 'alcohol', 'amitriptyline', 'barbiturates'.

Case Report

- Patient with pneumonia and cough
- Normal dosage of codeine
- Patient not responding any more at day 4
- What's going on?
 - PubMed „Codeine intoxication“ -> ~170 abstracts
 - Aren't there better ways?
- **Alibaba**
 - Potential cause is found quickly
 - Link to evidence provided
- Original article is at position 28 in PubMed result list

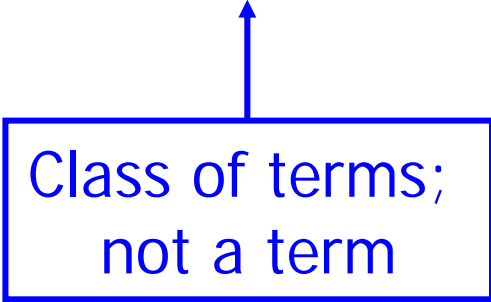
Finding Relevant Knowledge

- “Find information about ...”
- Much knowledge is in text (and only text)
- Find articles with information about ...
 - PubMed/Medline
 - Information Retrieval
- Find information ... **inside each article**
 - Reading many abstracts is tedious
 - What about a “**summarize results**” button?



Question

“Which proteins are associated to RAB5?”



Class of terms;
not a term

PubMed Results

Entrez PubMed - Mozilla Firefox

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=pubmed

Entrez PubMed (Unbenannt)

NCBI PubMed A service of the National Library of Medicine and the National Institutes of Health www.pubmed.gov

My NCBI Sign In Register

All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books

Search PubMed for rabf Go Clear Save Search

Limits Preview/Index History Clipboard Details

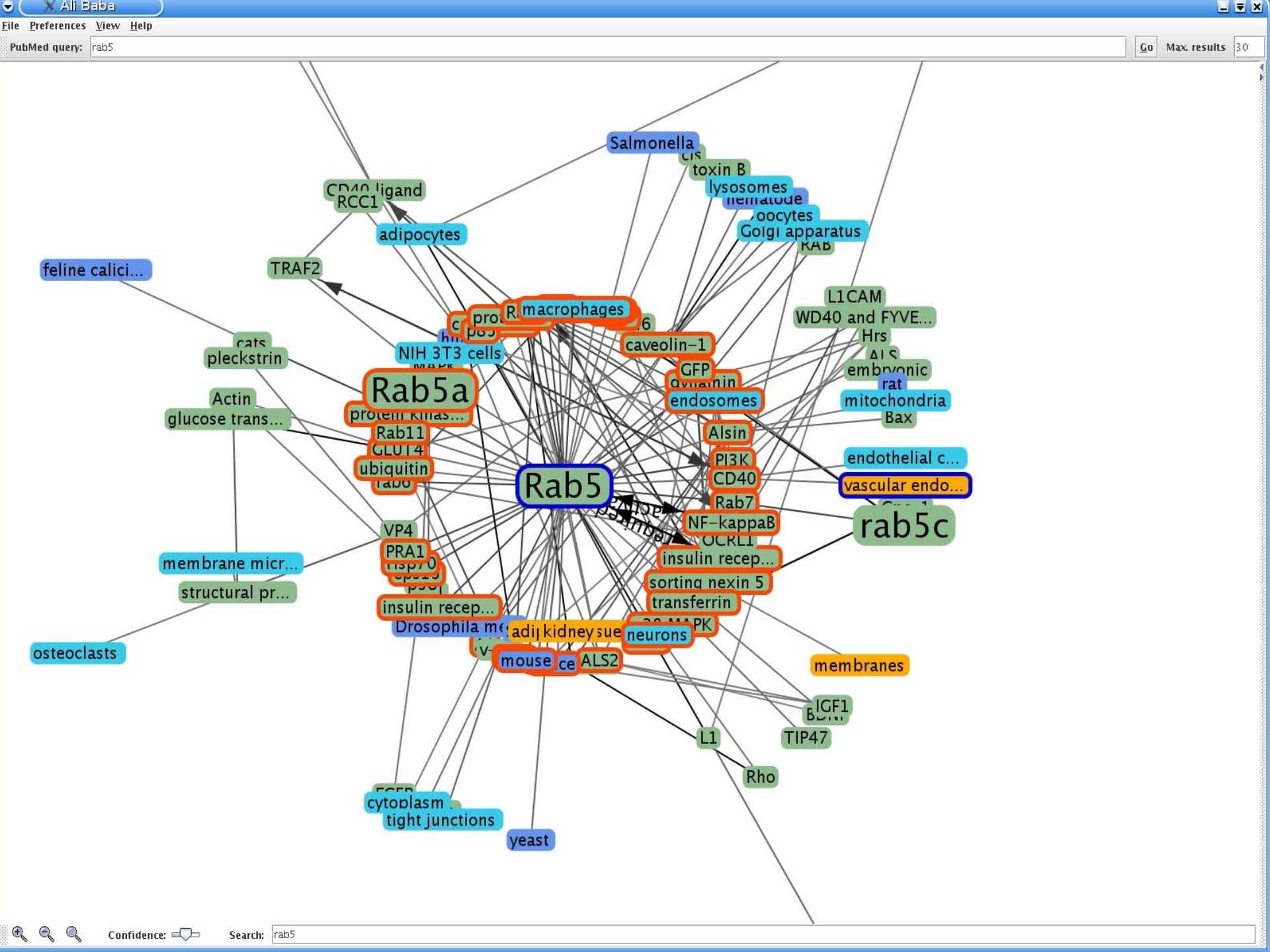
Display Summary 20 Sort by Send to

All: 630 Review: 3

Items 1 - 20 of 630 Page 1 of 32 Next

- 1: [Schneider-Brachert W, Torun M, Kunkel O, Jakob M, Hallas C, Kruse ML, Groitl P, Lehm A, Hildt E, Held-Feindt J, Dobner T, Kabelitz D, Kronke M, Schutze S.](#) Related Articles, Links
Inhibition of TNF receptor 1 internalization by adenovirus 14.7K as a novel immune escape mechanism. *J Clin Invest.* 2006 Oct 5; [Epub ahead of print]
PMID: 17024246 [PubMed - as supplied by publisher]
- 2: [Schroder A, Schroder B, Rospenser B, Linder S, Sinha B, Fassler R, Aepfelbacher M.](#) Related Articles, Links
Staphylococcus aureus Fibronectin Binding Protein-A Induces Motile Attachment Sites and Complex Actin Remodeling in Living Endothelial Cells. *Mol Biol Cell.* 2006 Oct 4; [Epub ahead of print]
PMID: 17021255 [PubMed - as supplied by publisher]
- 3: [Varsano T, Dong MQ, Niesman I, Gacula H, Lou X, Ma T, Testa JR 3rd, Faruqhar MG.](#) Related Articles, Links
GPCR is recruited by APPL to Peripheral TrkA Endosomes and Regulates TrkA Trafficking and Signaling. *Mol Cell Biol.* 2006 Oct 2; [Epub ahead of print]
PMID: 17015470 [PubMed - as supplied by publisher]
- 4: [Hawkes C, Kabogo D, Amritraj A, Kar S.](#) Related Articles, Links
Up-Regulation of Cation-Independent Mannose 6-Phosphate Receptor and Endosomal-Lysosomal Markers in Surviving Neurons after 192-IgG-Saporin Administrations into the Adult Rat Brain.

Fertig



Outline of This Talk

- Text Mining (Information Extraction)
- Named Entity Recognition
 - Using Machine Learning
- Evaluation
- Disambiguation
- AliBaba Walk Through
- Conclusions

What we need to do

Z-100 is an arabinomannan extracted from *Mycobacterium tuberculosis* that has various immunomodulatory activities, such as the induction of interleukin 12, interferon gamma (IFN-gamma) and beta-chemokines. The effects of Z-100 on human immunodeficiency virus type 1 (HIV-1) replication in human monocyte-derived macrophages (MDMs) are investigated in this paper. In MDMs, Z-100 markedly suppressed the replication of not only macrophage-tropic (M-tropic) HIV-1 strain (HIV-1JR-CSF), but also HIV-1 pseudotypes that possessed amphotropic Moloney murine leukemia virus or vesicular stomatitis virus G envelopes. Z-100 was found to inhibit HIV-1 expression, even when added 24 h after infection. In addition, it substantially inhibited the expression of the pNL43lucDeltaenv vector (in which the env gene is defective and the nef gene is replaced with the firefly luciferase gene) when this vector was transfected directly into MDMs. These findings suggest that Z-100 inhibits virus replication, mainly at HIV-1 transcription. However, Z-100 also downregulated expression of the cell surface receptors CD4 and CCR5 in MDMs, suggesting some inhibitory effect on HIV-1 entry. Further experiments revealed that Z-100 induced IFN-beta production in these cells, resulting in induction of the 16-kDa CCAAT/enhancer binding protein (C/EBP) beta transcription factor that represses HIV-1 long terminal repeat transcription. These effects were alleviated by SB 203580, a specific inhibitor of p38 mitogen-activated protein kinases (MAPK), indicating that the p38 MAPK signalling pathway was involved in Z-100-induced repression of HIV-1 replication in MDMs. These findings suggest that Z-100 might be a useful immunomodulator for control of HIV-1 infection.

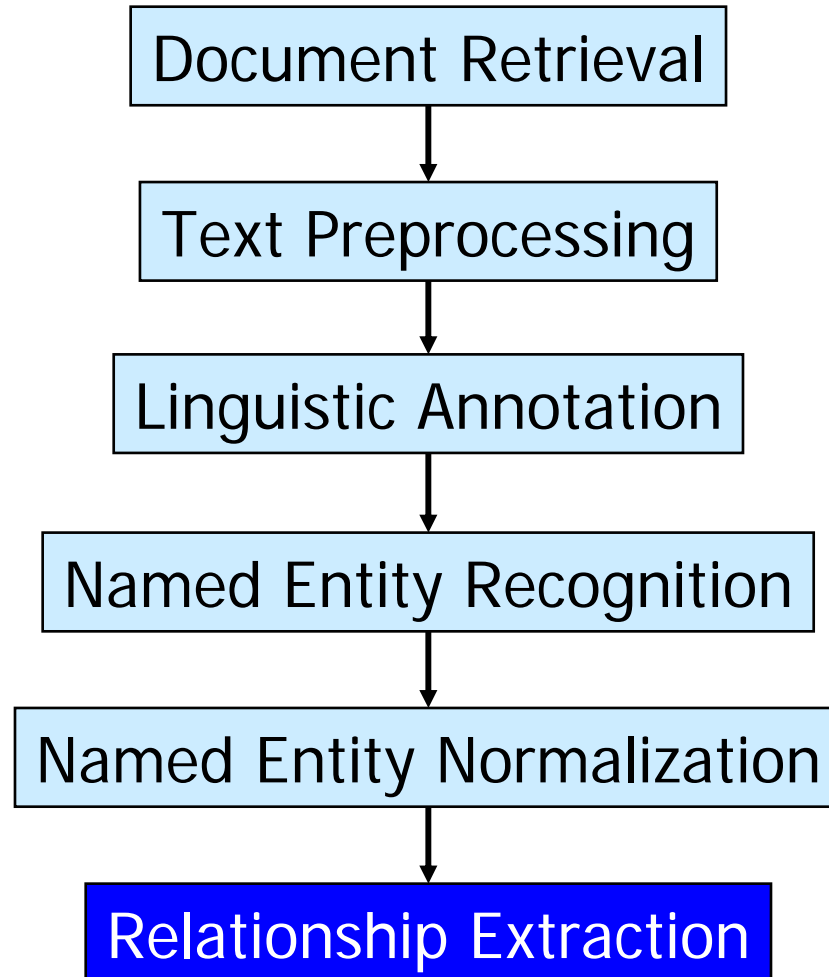
Find Names Entities (= Entity Names)

Z-100 is an *arabinomannan* extracted from *Mycobacterium tuberculosis* that has various immunomodulatory activities, such as the induction of *interleukin 12*, *interferon gamma* (*IFN-gamma*) and beta-chemokines. The effects of *Z-100* on *human immunodeficiency virus type 1* (*HIV-1*) replication in *human monocyte-derived macrophages* (*MDMs*) are investigated in this paper. In *MDMs*, *Z-100* markedly suppressed the replication of not only macrophage-tropic (M-tropic) *HIV-1* strain (*HIV-1JR-CSF*), but also *HIV-1* pseudotypes that possessed amphotropic *Moloney murine leukemia virus* or *vesicular stomatitis virus G* envelopes. *Z-100* was found to inhibit *HIV-1* expression, even when added 24 h after infection. In addition, it substantially inhibited the expression of the pNL43lucDeltaenv vector (in which the *env* gene is defective and the *nef* gene is replaced with the *firefly luciferase* gene) when this vector was transfected directly into *MDMs*. These findings suggest that *Z-100* inhibits virus replication, mainly at *HIV-1* transcription. However, *Z-100* also downregulated expression of the cell surface receptors *CD4* and *CCR5* in *MDMs*, suggesting some inhibitory effect on *HIV-1* entry. Further experiments revealed that *Z-100* induced *IFN-beta* production in these cells, resulting in induction of the 16-kDa *CCAAT/enhancer binding protein* (*C/EBP*) *beta transcription factor* that represses *HIV-1* long terminal repeat transcription. These effects were alleviated by SB 203580, a specific inhibitor of *p38 mitogen-activated protein kinases* (*MAPK*), indicating that the *p38 MAPK* signalling pathway was involved in *Z-100*-induced repression of *HIV-1* replication in *MDMs*. These findings suggest that *Z-100* might be a useful immunomodulator for control of *HIV-1* infection.

Find Relationships

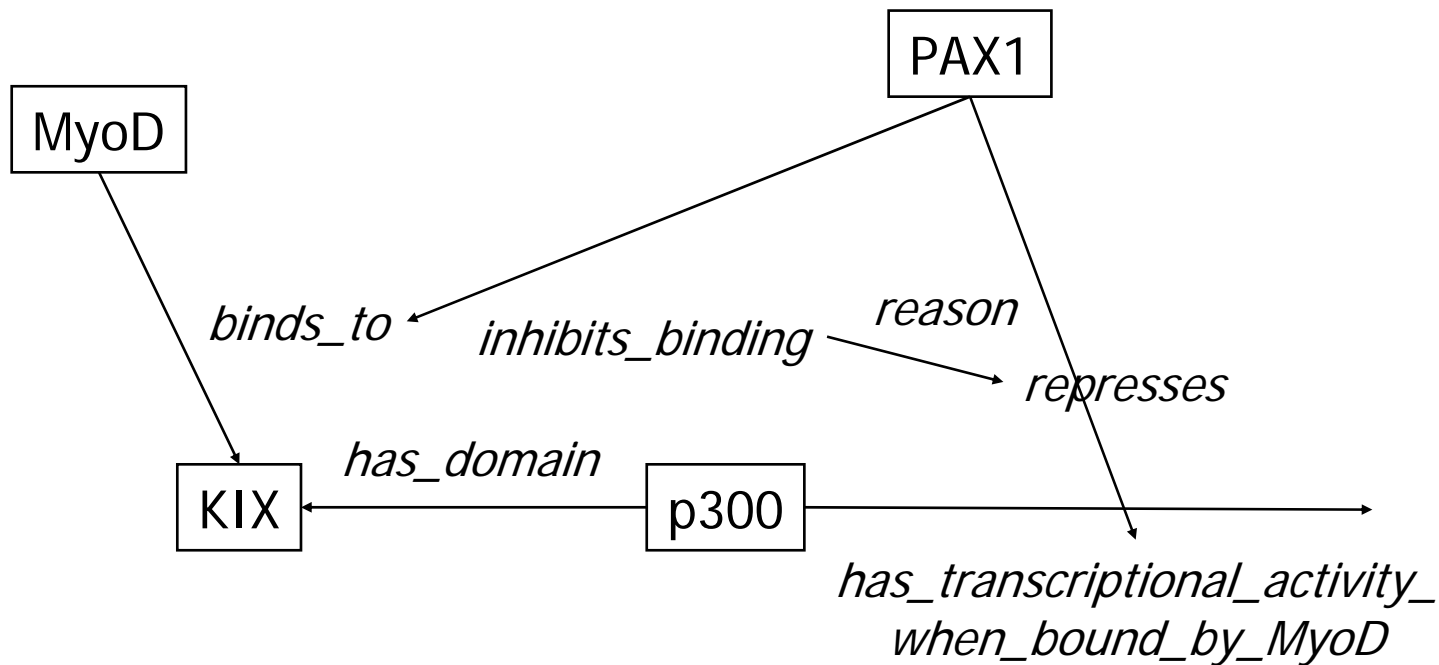
Z-100 is an **arabinomannan** extracted from **Mycobacterium tuberculosis** that has various immunomodulatory activities, such as the induction of **interleukin 12**, **interferon gamma** (**IFN-gamma**) and beta-chemokines. The effects of **Z-100** on **human immunodeficiency virus type 1** (**HIV-1**) replication in **human monocyte-derived macrophages** (**MDMs**) are investigated in this paper. In **MDMs**, **Z-100** markedly suppressed the replication of not only macrophage-tropic (M-tropic) **HIV-1** strain (**HIV-1JR-CSF**), but also **HIV-1** pseudotypes that possessed amphotropic **Moloney murine leukemia virus** or **vesicular stomatitis virus G** envelopes. **Z-100** was found to inhibit **HIV-1** expression, even when added 24 h after infection. In addition, it substantially inhibited the expression of the pNL43lucDeltaenv vector (in which the **env** gene is defective and the **nef** gene is replaced with the **firefly luciferase** gene) when this vector was transfected directly into **MDMs**. These findings suggest that **Z-100** inhibits virus replication, mainly at **HIV-1** transcription. However, **Z-100** also downregulated expression of the cell surface receptors **CD4** and **CCR5** in **MDMs**, suggesting some inhibitory effect on **HIV-1** entry. Further experiments revealed that **Z-100** induced **IFN-beta** production in macrophages, resulting in induction of the 16-kDa **CCAAT/enhancer binding protein (C/EBP) beta transcription factor** that represses **HIV-1** long terminal repeat transcription. These effects were alleviated by SB 203580, a specific inhibitor of **p38 mitogen-activated protein kinases (MAPK)**, indicating that the **p38 MAPK** signalling pathway was involved in **Z-100**-induced repression of **HIV-1** replication in **MDMs**. These findings suggest that **Z-100** might be a useful immunomodulator for control of **HIV-1** infection.

Typical IE-Workflow



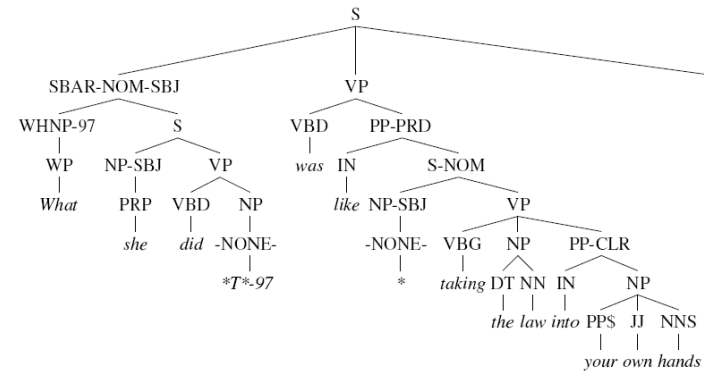
Understanding Text is Difficult (even for us)

„The PAX1 protein represses MyoD-dependent transcription by inhibiting MyoD-binding to the KIX domain of p300.“



Approaches to Text Understanding

- “AI” approach
 - Natural language processing
 - Full parsing, complete syntax tree
 - Aims at “understanding” the text
- Text mining
 - Simple NLP and machine learning
 - Stemming, part-of-speech (chunking)
 - Classification, pattern matching
 - Pragmatic approach
 - Usually not perfect
 - Needs careful *evaluation*



NOM	VRB	PRP	NOM
FLICE	bind	to	FADD
FLICE	binds	to	FADD

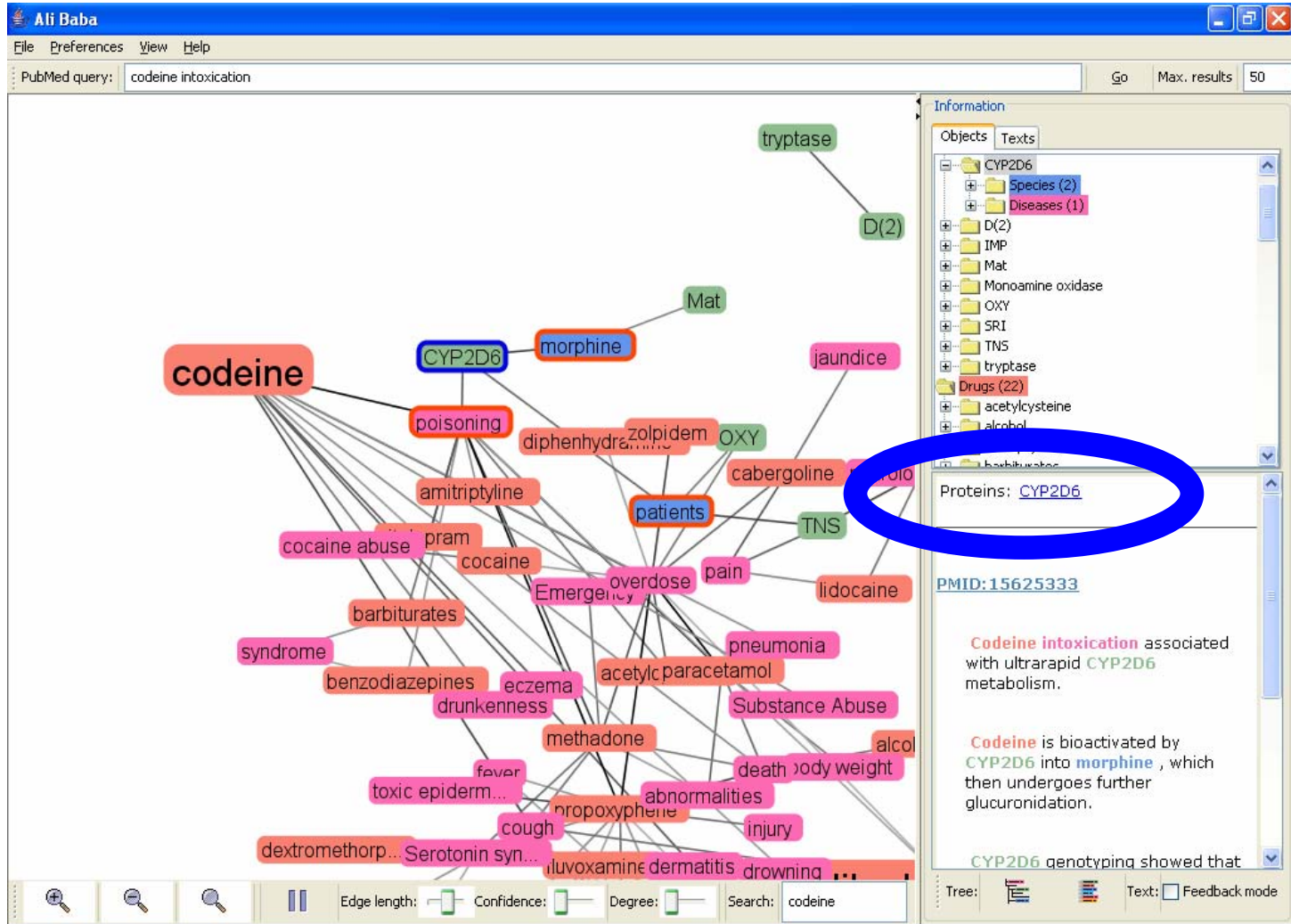
TOC

- Text Mining
- **Named Entity Recognition**
 - Using Machine Learning
- Evaluation
- Disambiguation
- AliBaba Walk Through
- Conclusions

Named Entity Recognition

- **Named Entity Recognition (NER)**: Find all mentionings of entity names in text
 - For one class (genes) or multiple classes (genes, diseases, species, ...)
 - Entity names may span more than one token
 - Does not identify the entity
- **Named Entity Normalization (NEN)**: Find all mentionings of entity names in text and assign each entity its canonical ID
 - „Canonical ID“ is highly application specific
 - Typical approach: Map mentioning into a set of known objects
 - A dictionary, a database, a list, ...
 - Alternative: Find equal entities in multiple text – duplicate detection“
 - What happens in case of incomplete dictionaries / new objects?
 - NEN is necessary to link entities to further information
- **NER is prerequisite for NEN**

Where is NEN?



Examples

“High plasma AVP levels observed in the two cases suggest that SSRIs stimulate AVP secretion, thereby causing SIADH .”

“A Drosophila shc gene product is implicated in signaling by the DER receptor tyrosine kinase.”

“The human T cell leukemia lymphotropic virus type 1 Tax protein represses MyoD-dependent transcription by inhibiting MyoD-binding to the KIX domain of p300.”

Examples

“High **plasma AVP** levels observed in the two cases suggest that SSRIs stimulate **AVP** secretion, thereby causing SIADH .”

- NER requires domain knowledge

“A *Drosophila* **shc** gene product is implicated in signaling by the DER receptor tyrosine kinase.”

“The human T cell leukemia lymphotropic virus type 1 Tax protein represses MyoD-dependent transcription by inhibiting MyoD-binding to the KIX domain of p300.”

Examples

“High **plasma AVP** levels observed in the two cases suggest that SSRIs stimulate **AVP** secretion, thereby causing SIADH .”

- NER requires domain knowledge

“A **Drosophila shc gene product** is implicated in signaling by the **DER** receptor **tyrosine kinase**.”

- Has to deal with ambiguities (context is important)

“The human T cell leukemia lymphotropic virus type 1 Tax protein represses MyoD-dependent transcription by inhibiting MyoD-binding to the KIX domain of p300.”

Examples

“High **plasma AVP** levels observed in the two cases suggest that SSRIs stimulate **AVP** secretion, thereby causing SIADH .”

- Requires domain knowledge

“A **Drosophila shc gene product** is implicated in signaling by the **DER** receptor **tyrosine kinase**.”

- Has to deal with ambiguities (context is important)

“The **human T cell leukemia lymphotropic virus type 1 Tax protein** represses MyoD-dependent transcription by inhibiting MyoD-binding to the **KIX domain of p300**.”

- Often has no clear answer

Evaluation

- Quality of a NER method is measured in terms of precision and recall

		Reality	
		+	-
Prediction	+	TruePositive (TP)	FalsePositive (FP)
	-	FalseNegative (FN)	TrueNegative (TN)

- Precision** = $TP / (TP + FP)$
 - Percentage of real hits among reported hits
- Recall** = $TP / (TP + FN)$
 - Percentage of real hits found
- Trade-Off**
 - Usually, NER algorithms compute a score per potential entity
 - Hopefully, the score correlates to the probability that a sequence of token is a mentioning
 - Use threshold on score
 - High threshold: Good precision, bas recall
 - Low threshold: Low precision, high recall

Example

High **plasma AVP** levels observed in the two cases suggest that SSRIs stimulate **AVP** secretion, thereby causing **SIADH**. A **Drosophila shc gene product** is implicated in signaling by the **DER** receptor **tyrosine kinase**

	Real: Positive	Real: Negative
Alg: Positive	TP = 2	FP = 2
Alg: Negative	FN = 3	?

- Precision = $TP / (TP + FP) = 2 / 4 = 50\%$
- Recall = $TP / (TP + FN) = 2 / 5 = 40\%$

NER: Two Main Approaches

- Dictionary-based
 - Build a dictionary of all known entities
 - Match every mentioning in the text against all dictionary entries
 - Usually: High precision, low recall (depending on dictionary)
 - Increase in recall though [fuzzy string matching](#) (or dictionary curation)
 - Advantages: Simple, fast, included NEN
 - Disadvantages: [Bad performance](#)
- Using classification / machine learning
 - Predict for each (seq. of) token whether it is an entity mentioning or not
 - Learn model based on [manually annotated training text](#)
 - Can be tuned towards balanced precision / recall
 - Advantages: [Good performance](#), also recognizes unseen entities
 - Disadvantages: Slow, complex, needs training data, assumes some communality among entity names, requires additional NEN step

TOC

- Text Mining
- Named Entity Recognition
 - [Using Machine Learning](#)
- Evaluation
- Disambiguation
- AliBaba Walk Through
- Conclusions

Classification-Based NER

- Various suggestions
- In the following
 - Convert each token into a feature vector
 - Binary classifier: Classify each token as ENT/NOENT based on vector
 - Classifier **learns model from training text**
 - Roughly: Learn typical feature values for tokens that are entity names
 - Recognize sequences of tokens forming one entity using a rule-based post-processing
- Properties
 - Simple approach with limited success
 - Main problem: Disregards context of tokens
 - Better models: Include sequence of types of tokens into model
 - Hidden Markov Models, Random Fields, ...

From Token to Features

- Feature selection and construction is the main challenge in classification
- Be creative! Find commonalities! Don't think binary!
- Find hints!

High **plasma AVP** levels observed in the two cases suggest that SSRIs stimulate **AVP** secretion, thereby causing SIADH.

A **Drosophila shc gene product** is implicated in signaling by the **DER** receptor **tyrosine kinase**.

The **human T cell leukemia lymphotropic virus type 1 Tax protein** represses MyoD-dependent transcription by inhibiting MyoD-binding to the **KIX domain of p300**.

Exemplary Features

Feature	Weight	Example
Word	tf * idf	kinase
n-grams		
N=1	tf * idf	k, i, n, a, s, e
N=2	tf * idf	ki, in, na, as, se
N=3	tf * idf	kin, ina, nas, ase
Special signs		
HasNumbers	[1 0]	p300
HasCapitals	[1 0]	abLIM
AllCaps	[1 0]	DMD
InitCap	[1 0]	Pax
HasNumbers & Letters	[1 0]	cMOAT2, EST90757
Context		
predecessing word	[1 0]	Gene
succeeding word	[1 0]	Product
distance to keywords	$1/(1+dist)$	(list of 15)
Dictionary		
Word match	[1 0]	
Phrase match	[1 0]	

Example: Bayes' Classification

- Simple method based on conditional probabilities
 - Works surprisingly well in practice
- Given
 - Set O of training objects, two classes c_1, c_2 (GENE or NOTGENE)
 - Objects are described as set F of binary features
 - $\text{hasNumbers}(p300) = \text{TRUE}$
- We seek $p(c_i|o)$, the probability of an object $o \in O$ being a member of class c_i
 - For every c_i and every o
- Each o is assigned to c where $p(c|o) = \text{argmax } p(c_i|o)$
- Rewrite o into feature representation

$$p(c | o) = p(c | f_1[o], \dots, f_n[o]) = p(c | f_1, \dots, f_n)$$

Probabilities

- What we learn from the training data
 - The **A-Priori probability** $p(f)$ of every feature f
 - For how many object f resolves to true?
 - The **A-Priori probability** $p(c)$ of every class $c \in C$
 - How many objects in the training set are of class c ?
 - The conditional probabilities $p(f|c)$ for feature f being true in class c
 - Proportion of objects in c with feature f among all objects in c
- Rephrase and use Bayes' theorem

$$p(c | f_1, \dots, f_n) = \frac{p(f_1, \dots, f_n | c) * p(c)}{p(f_1, \dots, f_n)} \approx p(f_1, \dots, f_n | c) * p(c)$$

↑
Term can be dropped; value is identical for all classes, and we only need the largest final value (not its actual value)

Naive Bayes

- We have $p(c | o) \approx p(f_1, \dots, f_n | c) * p(c)$
- The first term cannot be learned with any reasonably large training set
 - There are 2^n combinations of feature values
 - Every combination will be too sparse to obtain meaningful frequencies
- Solution: Be „naive“
 - Assume [statistical independence](#) of all features
- Then

$$p(f_1, \dots, f_n | c) = p(f_1 | c) * \dots * p(f_n | c)$$

$$p(c | o) \approx p(c) * \prod_{i=1}^n p(f_i | c)$$

Beispiel

- A-Priori

- $P(c=gene) = 3/7$
- $P(c=no) = 1 - p(c=gene) = 4/7$
- $p(f_1=1) = 1/7$
- $P(f_2=1) = 3/5$
- $P(f_1=1|gene) = 1/3, P(f_1=1|no) = 0$
- $P(f_2=1|gene) = 1/3, p(f_2=1|no) = 2/4$
- **Smoothing**: $P(f_1=1|no) = 0,01$

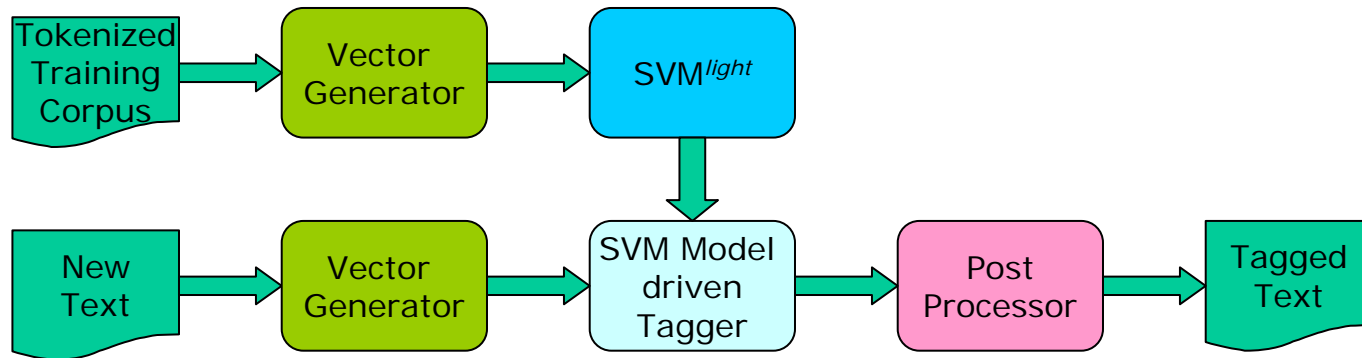
- Let's rate PAX5

- hasCap = hasNumber = 1
- Gene or not?

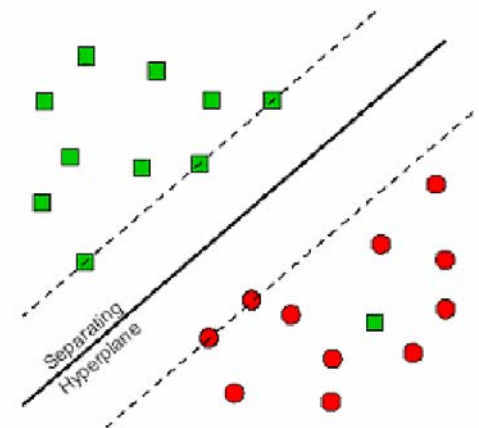
- $p(gene|f_1, f_2) = p(gene) * p(f_1|gene) * p(f_2|gene) = 3/7 * 1/3 * 1/3 = 3/63$
- $p(no|f_1, f_2) = p(no) * p(f_1|no) * p(f_2|no) = 4/7 * 0,01 * 2/4 = 8/2800$

Name	hasNumber	hasCap	class
AVP	0	1	gene
indicating	0	0	no
p300	1	0	gene
shc	0	0	gene
SSRIs	0	1	no
observed	0	0	no
MyoD	0	1	no
...

Complete Workflow



- Support Vector Machine (SVM) instead of Naive Bayes
 - More robust to errors in training data
 - Copes well with very high dimensional spaces
 - Copes better with unbiased training set



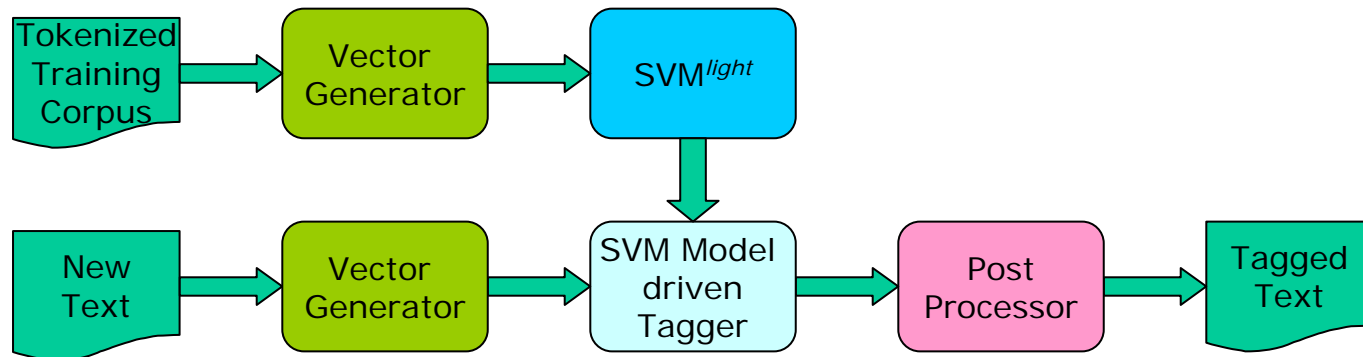
TOC

- Text Mining
- Named Entity Recognition
 - Using Machine Learning
- Evaluation
- Disambiguation
- AliBaba Walk Through
- Conclusions

Biocreative Cup 2004

- **Critical Assessment of Information Extraction Systems in Biology**
- International competition, three tasks
- Training data and evaluation script provided by organizers in cooperation with database curators (Swiss-Prot)
- Test data available for one week
- Objective evaluation of all submissions by (published) scripts
- **Major boost:** Top systems reached 84 F-measure
 - Previous: best systems around 60 F-Measure
 - Virtually no further improvements since then

Approach: SVM for NER



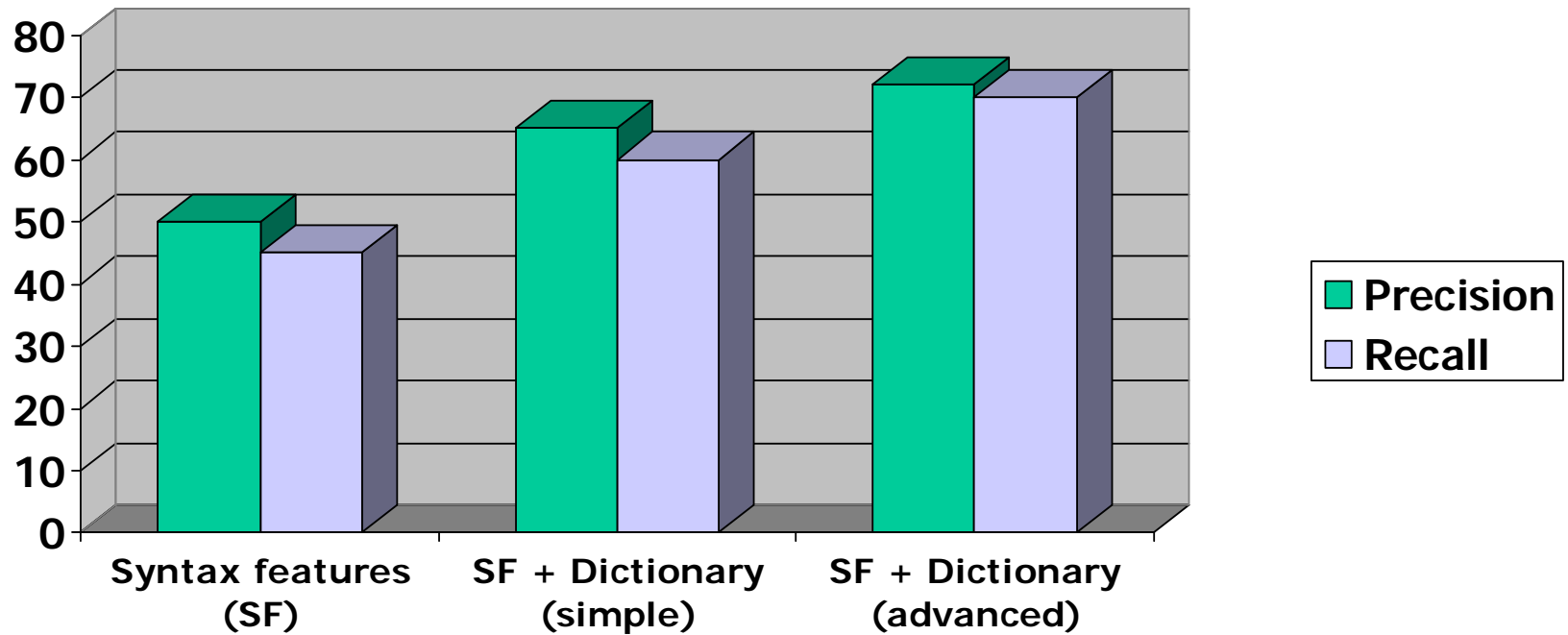
- Corpus of 7500 sentences
 - 140.000 non-gene words
 - 60.000 gene names
- Training SVM^{light} on different feature sets
- Dictionary compiled from Genbank, HUGO, MGD, YDB
- Post-processing for compound gene names

Post-processing

- SVM detects only single token candidates
- Most gene names are **multi-token names**
- Expand detected single-token genes based on set of heuristic rules

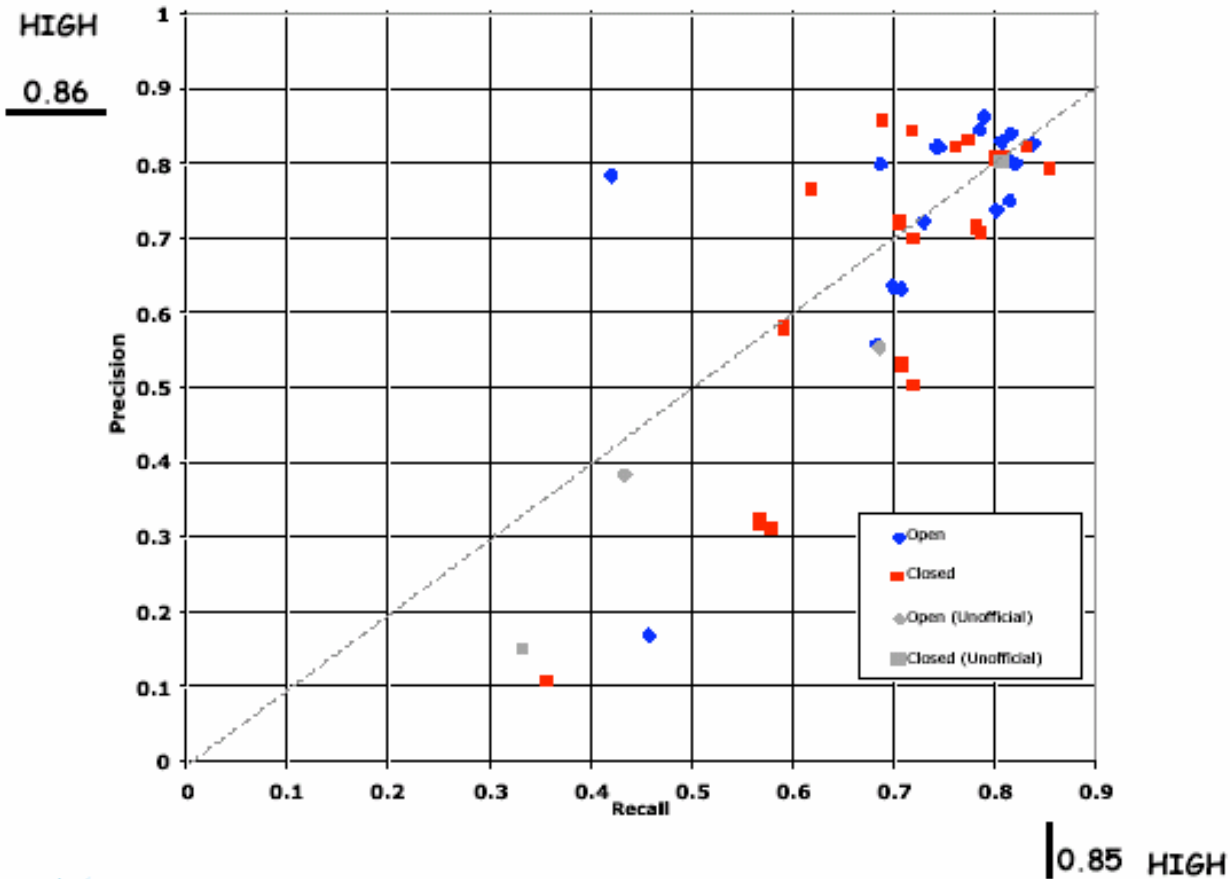
GENE NN*	→	GENE GENE
NN* GENE	→	GENE GENE GENE
GENE (NN)	→	GENE (GENE)
GENE protein	→	GENE GENE
GENE ADJ GENE	→	GENE GENE GENE

Performance



- Best result for BioCreative Cup: 73 F-measure
- Current feature set reaches 79 F-measure
- Raises from 73 to 83 for loose evaluation

BioCreative: All Results



NER – What's Left?

- Most successful features found by trial&error
 - Brute force approach (apparently true for all Biocreative participants)
 - We achieve the same performance when using only 10% of all features
 - But which 10% ?
- NER results depend on **type of object**
 - Gene or protein is hard
 - Gene and protein is much harder
 - Cell type: 81; virus strains: 67; disease: ?; drugs: ? ...
- What is left?
 - Entity names are not really defined (borders)
 - **Inter-Biologists agreement** on type (gene, protein, RNA) and exact borders around 70% (Krauthammer et al. 2000)
 - Overfitting to annotators likely; need for better and internationally accepted annotated corpora
 - Community-language: Species-specific NER is very promising

How far can we get without context?

- "DEAD" motif
- 12-kDa Ndk
- 12-tetradecanoate
- 15-kDa antigen of *Dirofilaria immitis*
- 16p13
- AAV promoter
- ACGCGT
- A-P diameter
- Bengal pink
- Bozozok / Dharma
- bZIP2
- C2H2 type ZFPs
- cAMP-regulated gene
- CAP SX1
- ...

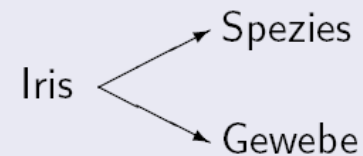
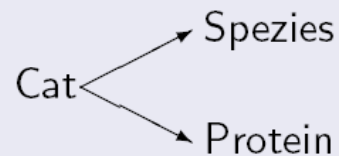
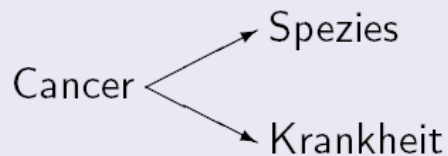
TOC

- Text Mining
- Named Entity Recognition
 - Using Machine Learning
- Evaluation
- Disambiguation
- AliBaba Walk Through
- Conclusions

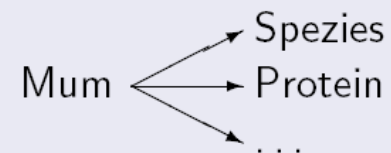
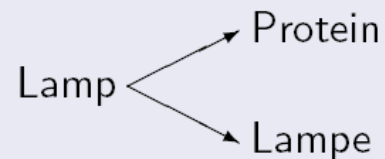
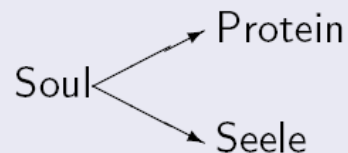
Ambiguity - Homonyms

- Often, entity names appear in **exactly the same** form elsewhere

Verwendung eines Terms in mehreren biologischen Fachgebieten



Verwendung ebenso im üblichen englischen Sprachgebrauch



Not a Rare Event

Entity class	Cell	Disease	Drug	Organism	Protein	Tissue	Common	Σ
Cell	–	5	2	0	0	123	0	130
Disease	5	–	5	18	4	3	3	37
Drug	2	5	–	9	67	1	15	99
Organism	0	18	9	–	175	4	38	239
Protein	0	4	67	175	–	2	55	300
Tissue	123	3	4	2	2	–	2	134
Common	0	3	15	38	55	2	–	108

Table 2: Amount of overlaps between terms of two classes.

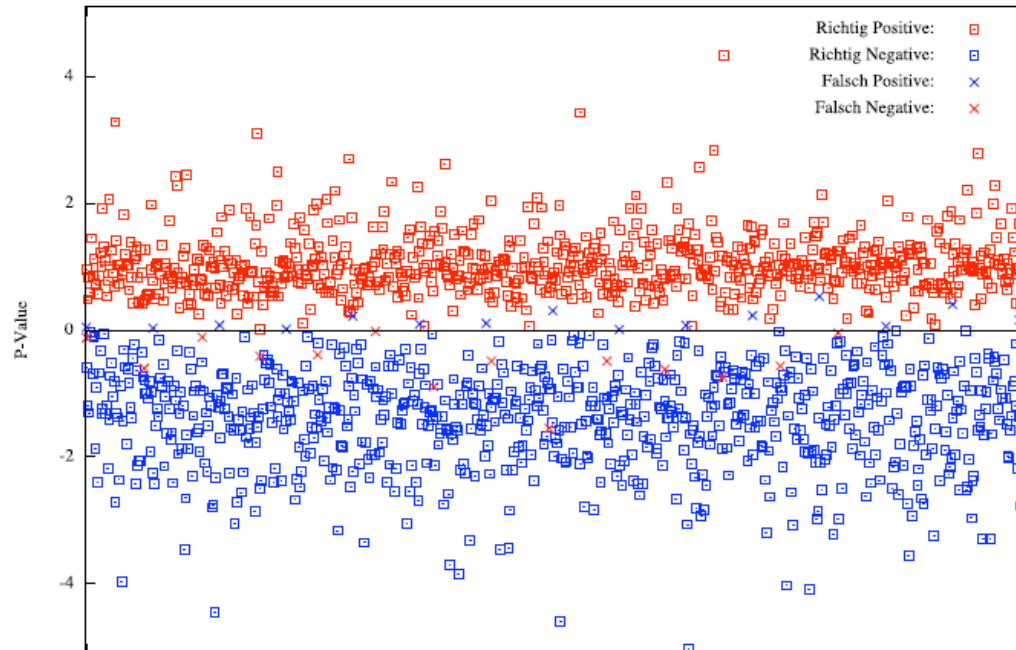
- Class-specific dictionaries compiled from many sources
 - MeSH
 - UMLS
 - UniProt, EntrezGene
 - OMIM
 - ...

Disambiguation

- Dictionaries don't help
 - Use English dictionary to find common names
 - Use class-specific dictionaries
 - Homonyms: Token appears in more than one class-specific dictionary
- Several approaches
 - [Sense dictionaries](#), alignment with examples, etc.
- Machine learning approach
 - Rely on “[one sense per discourse](#)” assumption
 - Build training data sets for each sense of a name
 - Cast problem in a multi-class classification problem
 - Build one model for each name

Works Well

Klasse v. <i>cat</i>	Protein	Species
Protein	747	15
Species	14	748



Training Data

- Problem: Need to find 100dreds of thousands of exemplary texts

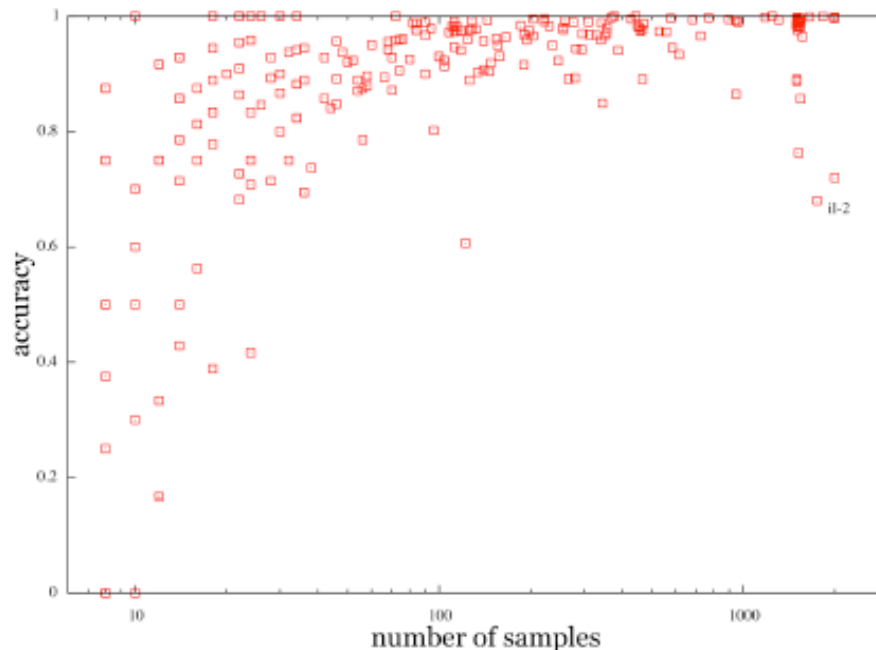
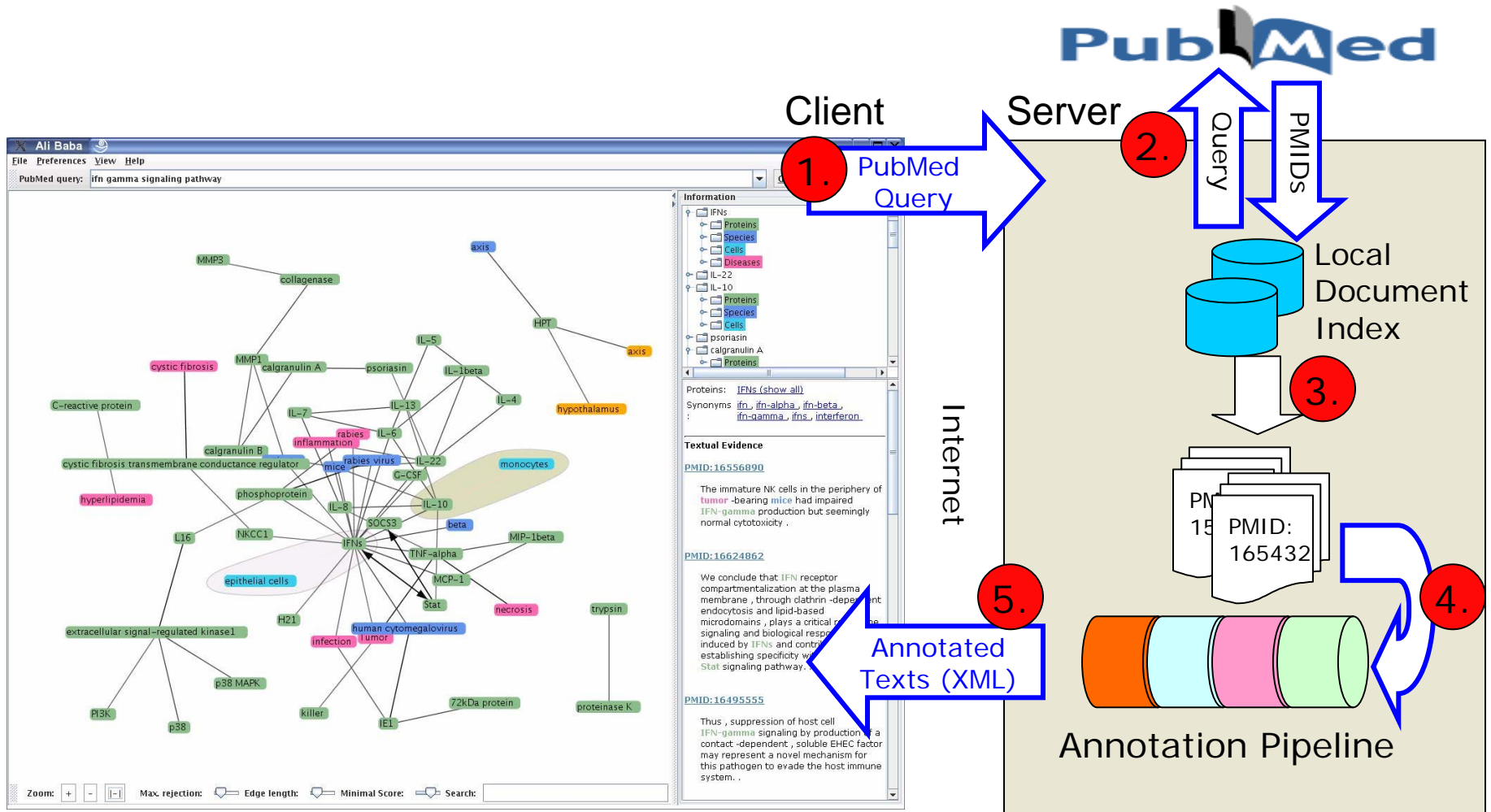


Figure 3: Dependency of accuracy on the number of training examples (logarithmic scale).

TOC

- Text Mining
- Named Entity Recognition
 - Using Machine Learning
- Evaluation
- Disambiguation
- [AliBaba Walk Through](#)
- Conclusions

Workflow



Query

Ali Ba
File Preferences View Help
PubMed query: ifn gamma signaling pathway

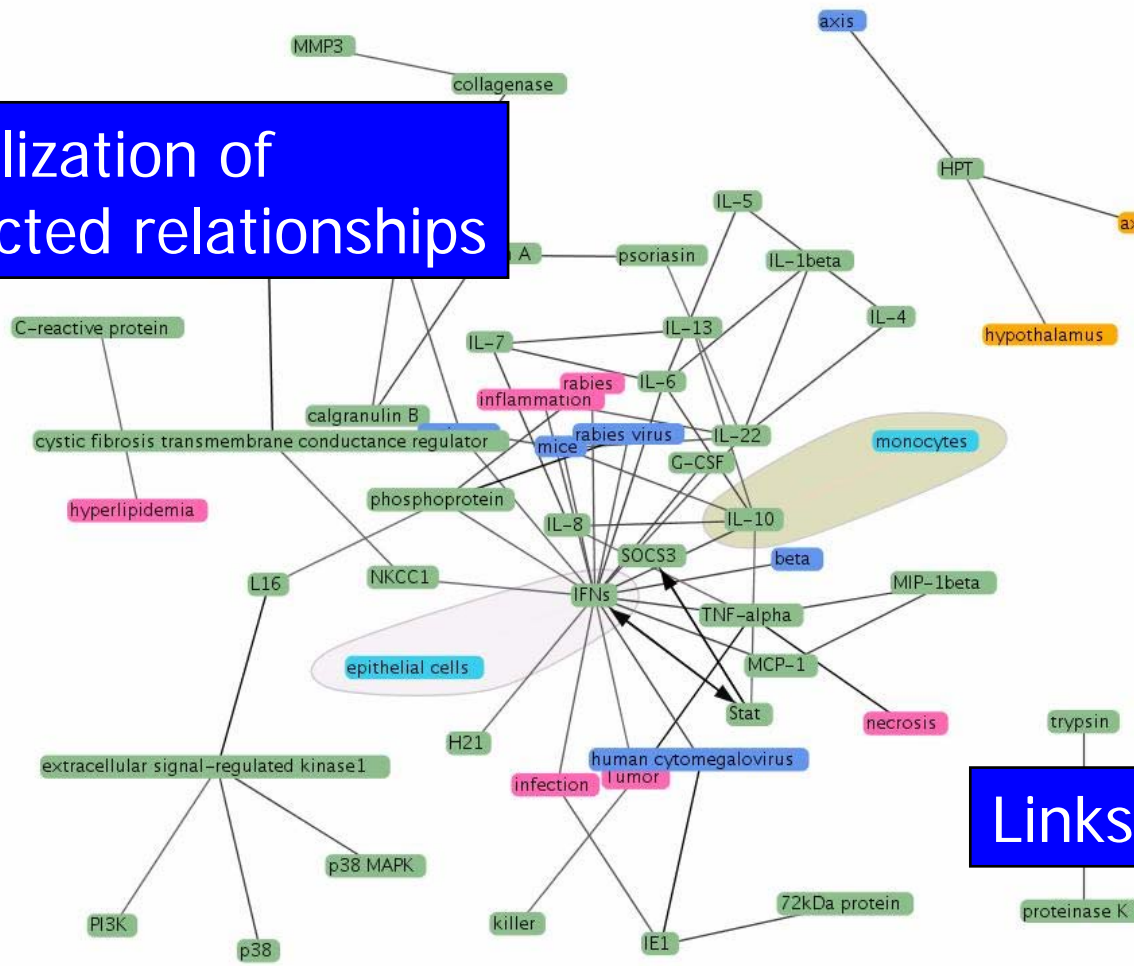
Extracted infos

Information

- IFNs
 - Proteins
 - Species
 - Cells
 - Diseases
- IL-22
 - Proteins
 - Species
 - Cells
- IL-10
 - Proteins
 - Species
 - Cells
- psoriasis
- calgranulin A
- Proteins

Proteins: [IFNs \(show all\)](#)
Synonyms: [ifn](#), [ifn-alpha](#), [ifn-beta](#), [ifn-gamma](#), [ifns](#), [interferon](#)

Visualization of extracted relationships



Links to databases

The immature NK cells in the periphery of **tumor**-bearing **mice** had impaired **IFN-gamma** production but seemingly normal cytotoxicity .

PMID:16624862

We conclude that **IFN** receptor compartmentalization at the plasma membrane , through clathrin -dependent endocytosis and lipid-based microdomains , plays a critical role in the signaling and biological processes .

Links to textual evidence

PMID:16495555

Thus , suppression of host cell **IFN-gamma** signaling by production of a contact -dependent , soluble EHEC factor may represent a novel mechanism for this pathogen to evade the host immune system . .

Zoom: + - | Search: Max. rejection: Edge length: Minimal Score:



Walk-through

- „Which proteins are associated with the TNF-alpha associated death domain (TRADD)?“

Many

The screenshot shows a network visualization application window titled "Ali Baba". The main area displays a dense network of nodes and edges. The node "TRADD" is highlighted with a blue circle. The interface includes a search bar at the top with the query "tradd" and a "Go" button. On the right side, there is an "Information" panel with a list of objects, a "Textual Evidence" section with three PubMed references, and a "Feedback mode" checkbox at the bottom.

Information

Objects: Texts

- STAT6
- TAB2
- TAK1
- THP
- TIMP 1
- TIMPs
- TLR2
- TLR3
- TLR4
- TLRs
- TNF
- TNF receptor-associated factor 2
- TNFR1
- TOLLIP
- TRADD

Proteins: [TRADD \(show all\)](#)

Textual Evidence

[PMID:17024246](#)

TNF -mediated apoptosis is initiated by ligand -induced recruitment of TNF receptor-associated death domain (TRADD), Fas-associated death domain (FADD), and caspase-8 to the death domain of TNF receptor 1 (TNFR1), thereby establishing the death -inducing signaling complex (DISC).

[PMID:16982613](#)

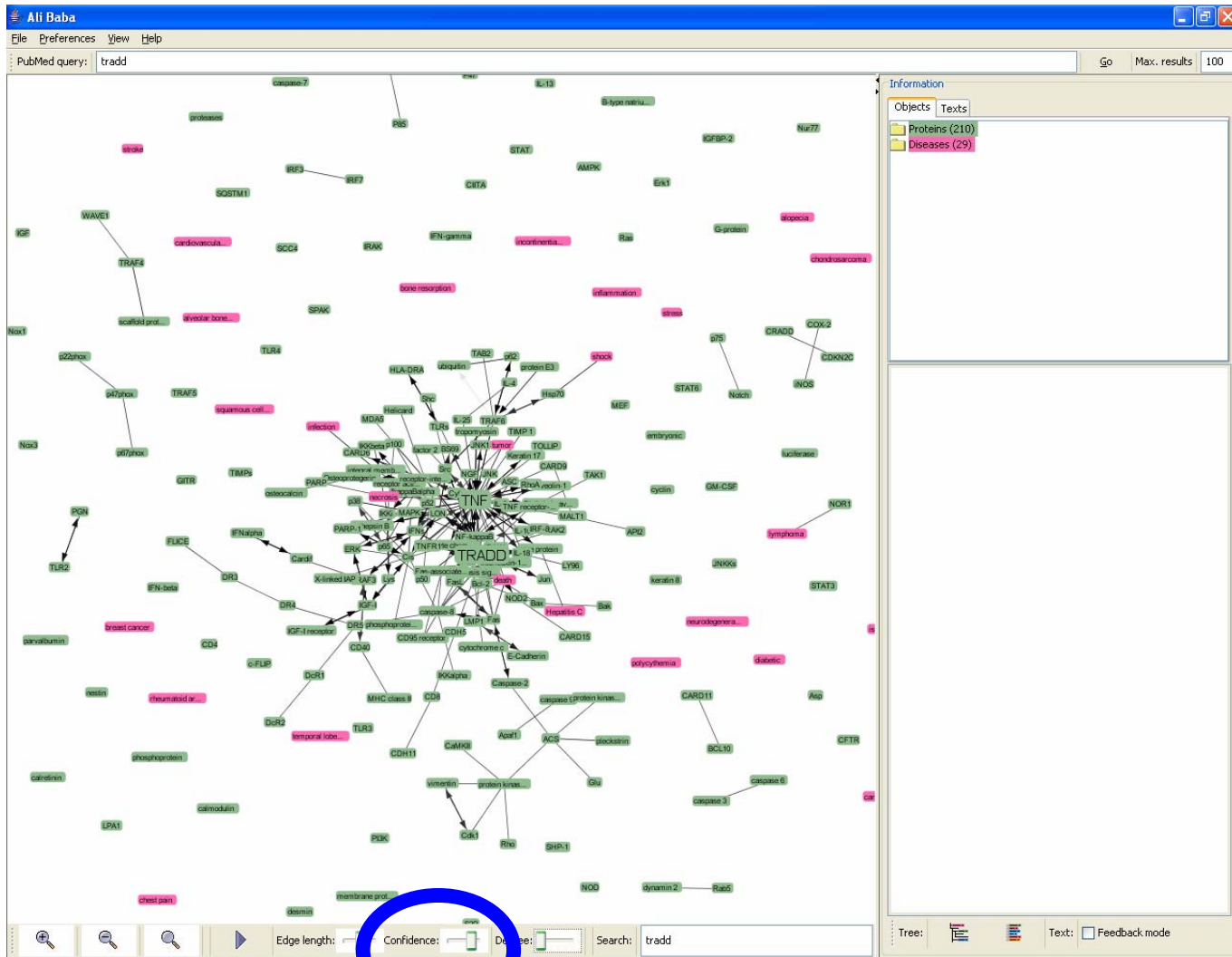
TNF-alpha engagement of TNFR1 recruited the adaptor proteins TRADD , TRAF-2 and RIP into lipid rafts and activated RhoA , NF-kappaB and MAPK pathways.

The signaling adaptors TRADD and RIP were also found to be necessary for ligand -induced RhoA activation.

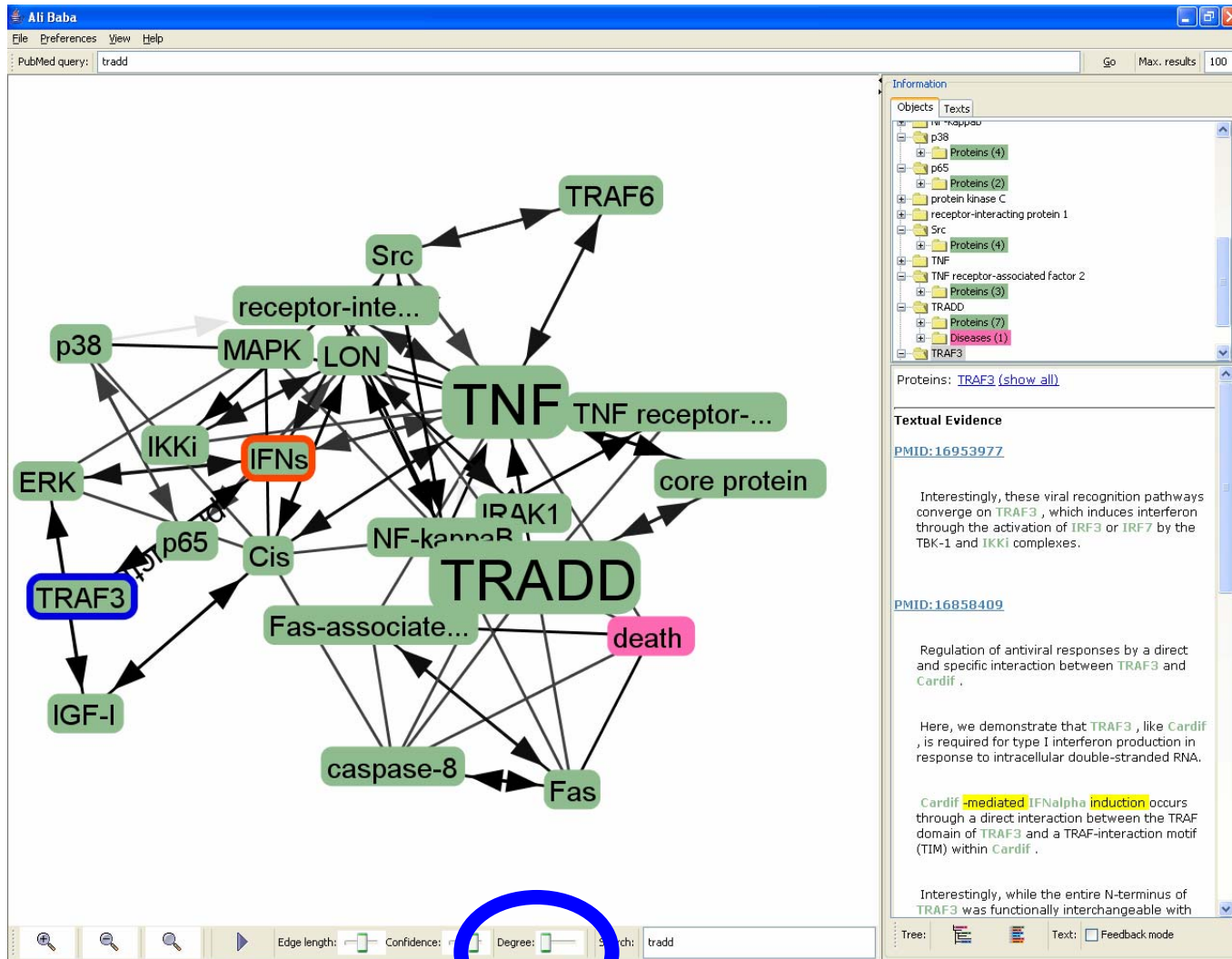
[PMID:16978814](#)

Tree: Feedback mode

Filter by Object Type and Confidence



Show only Connected Objects



Show Type of Interaction

The screenshot shows the Ali Baba software interface. The main window displays a network diagram with nodes representing proteins and edges representing interactions. The nodes are: MAPK, receptor-inte..., Cis, LON, TNF, TRAF6, TRADD, IFNs, and NF-kappaB. The interactions are: MAPK to receptor-inte... (binding), MAPK to Cis (induced), Cis to LON (induced), Cis to TNF (induced), LON to TNF (induced), TNF to TRAF6 (induced), TNF to TRADD (inhibits), TNF to IFNs (enhances), and TNF to NF-kappaB (induced). The right-hand panel shows the 'Information' section for TNF, including a list of 14 proteins, synonyms, and textual evidence from a PubMed entry (PMID:17024246) regarding TNF receptor internalization inhibition by adenovirus 14.7K.

Information

Objects Texts

Proteins (14)

- Cis
- Fas
- Fas-associated death domain
- IFNs
- IGF-1
- IKK
- LON
- MAPK
- NF-kappaB
- receptor-interacting protein 1
- Src
- TNF

Proteins (6)

- TRADD
- TRAF6

Proteins: [TNF \(show all\)](#)

[ntrk1](#), [tab1](#), [tab2](#), [tak1](#), [tak1-binding protein 2](#), [tca](#), [tccc](#), [tes](#), [tes1](#), [tfo](#), [tnf-beta-activated kinase 1](#), [tnf-beta1](#), [thymus and activation-regulated chemokine](#), [timp](#), [timps](#), [tio](#), [tirap](#), [tissue inhibitors of metalloproteinases](#), [tnf](#), [tnf a](#), [tnf receptor-associated factor 2](#), [tnf receptor-associated factor 6](#), [tnf receptor-associated factors 1](#), [tnf-receptor-associated factors 1](#), [tnf-alpha](#), [tnf-related apoptosis-inducing ligand](#), [tnfalpa](#), [tnfr-1](#), [tnfr1](#), [toa](#), [tradd](#), [traf1](#), [traf2](#), [traf6](#), [transforming growth factor-beta-activated kinase 1](#), [trk](#), [trka](#), [trks](#), [tumor necrosis factor](#), [tumor necrosis factor alpha](#), [tumor necrosis factor-alpha](#)

Synonyms:

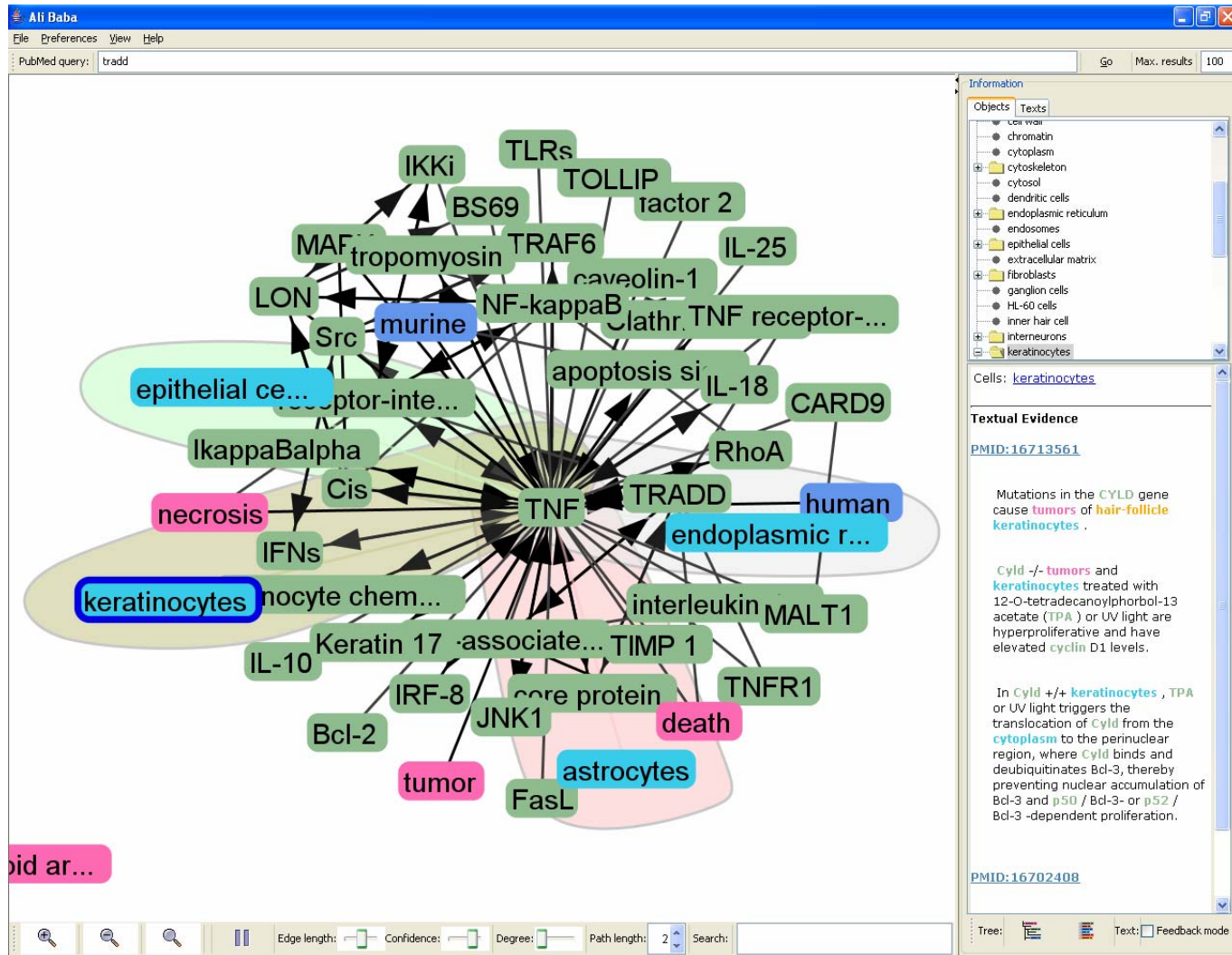
Textual Evidence

[PMID:17024246](#)

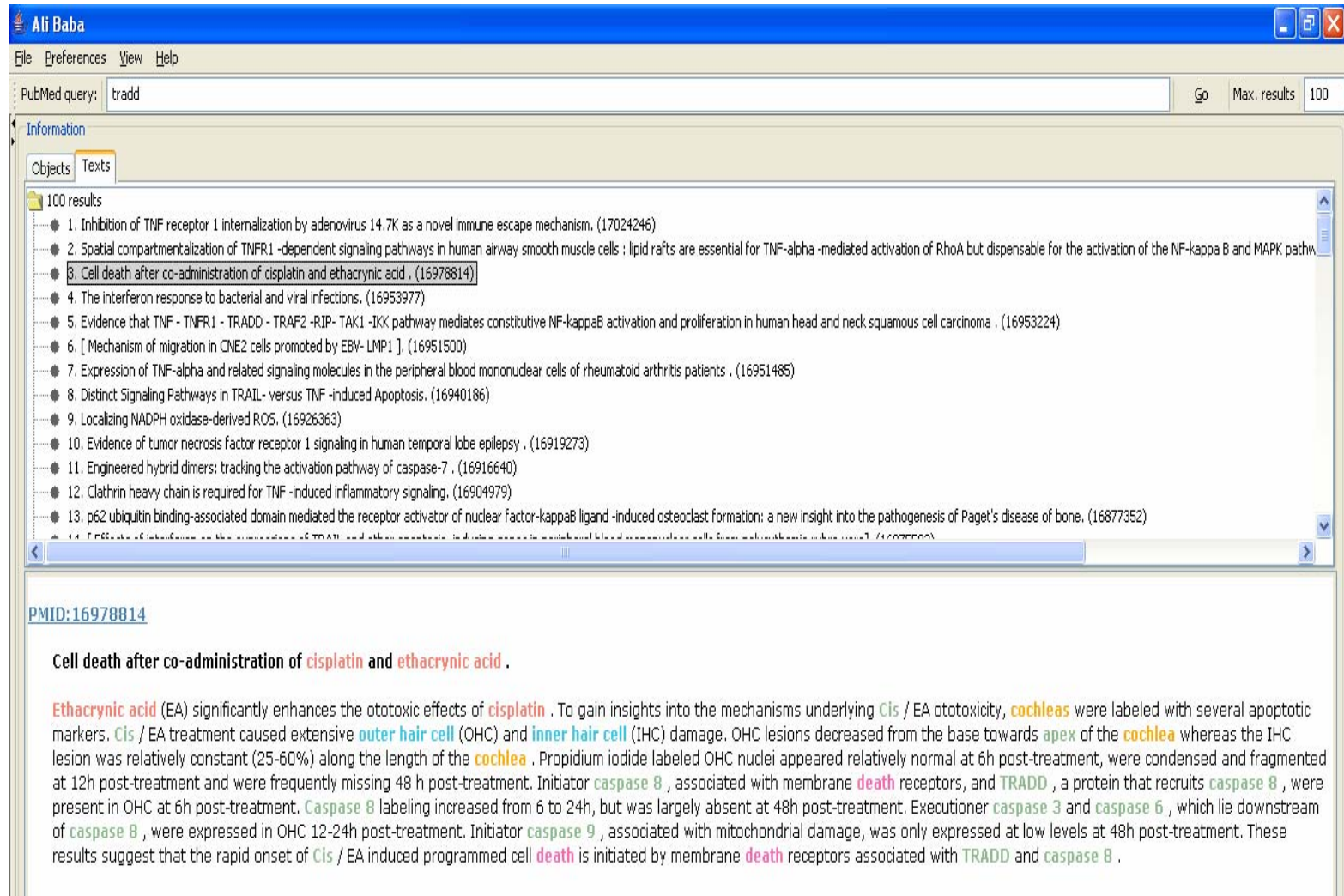
Inhibition of TNF receptor 1 internalization by adenovirus 14.7K as a novel immune escape mechanism.

The adenoviral protein E3 -14.7K (14.7K) is an inhibitor of TNF-induced apoptosis, but the molecular mechanism underlying this protective effect has not yet been explained exhaustively.

Location of Interaction



View Annotated Abstracts



Ali Baba

File Preferences View Help

PubMed query: tradd Go Max. results 100

Information

Objects Texts

100 results

1. Inhibition of TNF receptor 1 internalization by adenovirus 14.7k as a novel immune escape mechanism. (17024246)
2. Spatial compartmentalization of TNFR1 -dependent signaling pathways in human airway smooth muscle cells : lipid rafts are essential for TNF-alpha -mediated activation of RhoA but dispensable for the activation of the NF-kappa B and MAPK pathways. (16953224)
3. Cell death after co-administration of cisplatin and ethacrynic acid . (16978814)
4. The interferon response to bacterial and viral infections. (16953977)
5. Evidence that TNF - TNFR1 - TRADD - TRAF2 - RIP - TAK1 -IKK pathway mediates constitutive NF-kappaB activation and proliferation in human head and neck squamous cell carcinoma . (16953224)
6. [Mechanism of migration in CNE2 cells promoted by EBV- LMP1]. (16951500)
7. Expression of TNF-alpha and related signaling molecules in the peripheral blood mononuclear cells of rheumatoid arthritis patients . (16951485)
8. Distinct Signaling Pathways in TRAIL- versus TNF -induced Apoptosis. (16940186)
9. Localizing NADPH oxidase-derived ROS. (16926363)
10. Evidence of tumor necrosis factor receptor 1 signaling in human temporal lobe epilepsy . (16919273)
11. Engineered hybrid dimers: tracking the activation pathway of caspase-7 . (16916640)
12. Clathrin heavy chain is required for TNF -induced inflammatory signaling. (16904979)
13. p62 ubiquitin binding-associated domain mediated the receptor activator of nuclear factor-kappaB ligand -induced osteoclast formation: a new insight into the pathogenesis of Paget's disease of bone. (16877352)
14. Effects of interferon on the expression of TRAF1 and other proteins in peripheral blood mononuclear cells from patients with hepatitis B virus. (16875500)

PMID:16978814

Cell death after co-administration of cisplatin and ethacrynic acid .

Ethacrynic acid (EA) significantly enhances the ototoxic effects of **cisplatin** . To gain insights into the mechanisms underlying **Cis** / EA ototoxicity, **cochleas** were labeled with several apoptotic markers. **Cis** / EA treatment caused extensive **outer hair cell** (OHC) and **inner hair cell** (IHC) damage. OHC lesions decreased from the base towards **apex** of the **cochlea** whereas the IHC lesion was relatively constant (25-60%) along the length of the **cochlea** . Propidium iodide labeled OHC nuclei appeared relatively normal at 6h post-treatment, were condensed and fragmented at 12h post-treatment and were frequently missing 48 h post-treatment. Initiator **caspase 8** , associated with membrane **death** receptors, and **TRADD** , a protein that recruits **caspase 8** , were present in OHC at 6h post-treatment. **Caspase 8** labeling increased from 6 to 24h, but was largely absent at 48h post-treatment. Executioner **caspase 3** and **caspase 6** , which lie downstream of **caspase 8** , were expressed in OHC 12-24h post-treatment. Initiator **caspase 9** , associated with mitochondrial damage, was only expressed at low levels at 48h post-treatment. These results suggest that the rapid onset of **Cis** / EA induced programmed cell **death** is initiated by membrane **death** receptors associated with **TRADD** and **caspase 8** .

AliBaba and KEGG

The image displays the Ali Baba software interface, which is used for named entity recognition in biological text. The main window shows a complex metabolic pathway map with various chemical entities and reactions. A dialog box titled "KEGG File Chooser" is open, showing a list of metabolic pathways and a search field containing the text "glycol".

The KEGG File Chooser dialog box includes the following elements:

- Title: KEGG File Chooser
- Search field: glycol
- Buttons: Find, Cancel, Go
- Checkbox: specify organism
- List of pathways:
 - Glycolysis / Gluconeogenesis
 - Glycosaminoglycan degradation
 - Glycosphingolipid biosynthesis - ganglioseries
 - Glycosphingolipid biosynthesis - globoseries
 - Glycosphingolipid biosynthesis - lactoseries
 - Glycosphingolipid biosynthesis - neo-lactoseries
 - Glycosylphosphatidylinositol(GPI)-anchor biosynt

The main window shows a metabolic pathway map with various chemical entities and reactions. The entities are represented by colored boxes (pink, green, grey) and are connected by arrows representing reactions. The map is titled "Ali Baba" and includes a menu bar (File, Preferences, View, Help) and a PubMed query field.

Overlays (experimental)

Ali Baba

File Preferences View Help

PubMed query: hexokinase Go Max. results 5

Information

Objects Texts

- hexokinase
 - Proteins (6)
 - IGF-1
 - Strawberry
 - Invertase
 - cat
 - glucokinase
 - Akt
 - Species (2)
 - yeast
 - Kluyveromyces lactis
 - Reactions (2)
 - ATP + beta-D-Glucose => ADP + beta-D-Glucose
 - ATP + alpha-D-Glucose => ADP + alpha-D-Glucose

Proteins: [hexokinase](#)
[atp-dependent hexokinase](#), [glucose atp phosphotransferase](#), [hexokinase](#), [hexokinase \(phosphorylating\)](#), [hexokinase d](#), [hexokinase type iv](#), [hexokinase type iv glucokinase](#)

Synonyms:

Textual Evidence

PMID: 17581767

Effects of osmolytes on **hexokinase** kinetics combined with macromolecular crowding Test of the osmolyte compatibility hypothesis towards crowded systems. [\[more\]](#)

We investigated the effect of compatible and non-compatible osmolytes in combination with

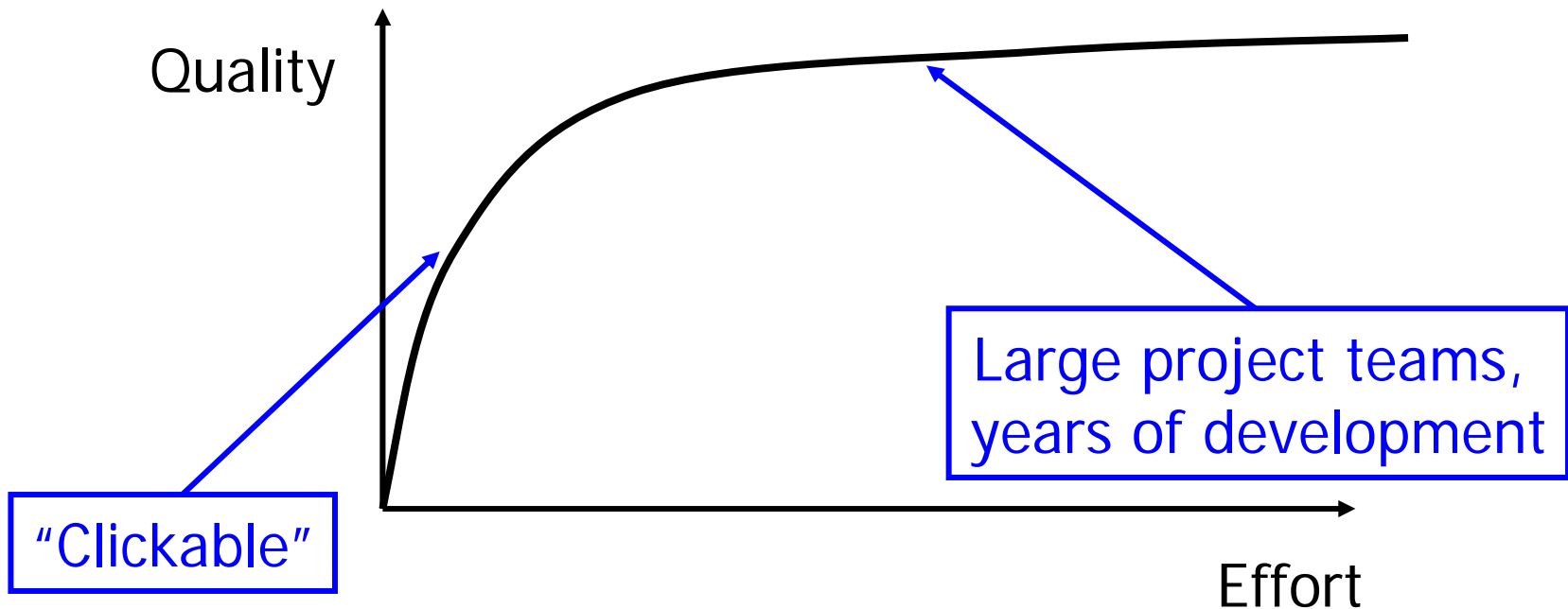
Tree: Feedback mode

Edge length: Confidence: Search:

TOC

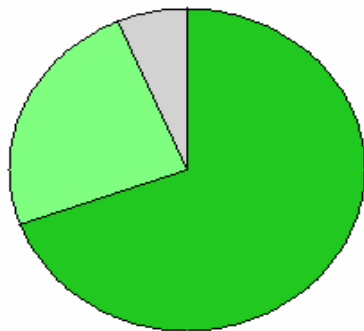
- Text Mining
- Named Entity Recognition
 - Using Machine Learning
- Evaluation
- Disambiguation
- AliBaba Walk Through
- Conclusions

Text Mining Performance Curve



Next Problem: Full Text (eprints.org)

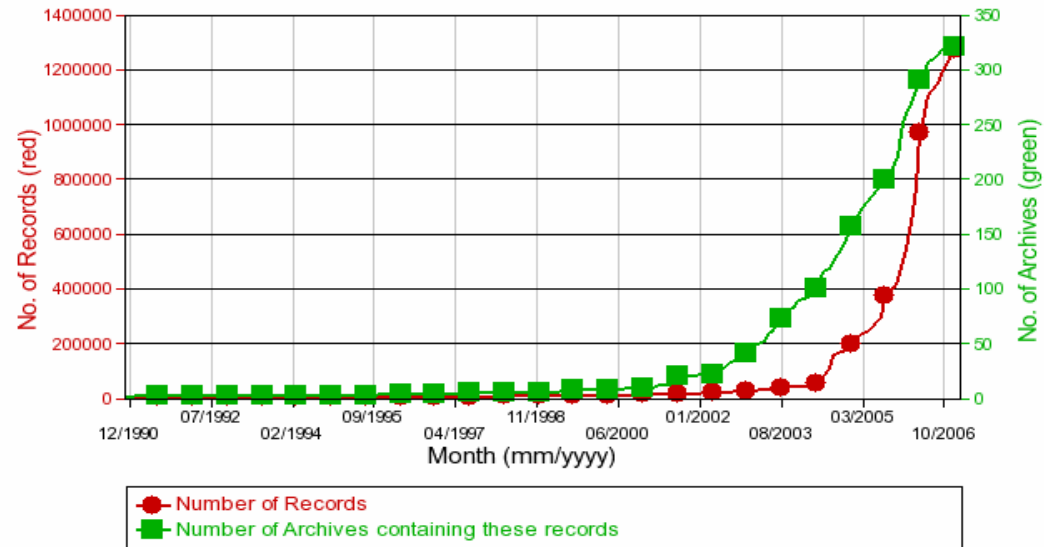
Journal Policy Chart



69.53% 6491 GREEN journals
24.07% 2247 PALE-GREEN journals
6.4% 597 GRAY journals

Growth of Institutional Archives and Contents

Generated by <http://archives.eprints.org/>



- Scalability is becoming really important

Acknowledgements

- Humboldt-Universität, Informatics
 - Jörg Hakenberg (DD)
 - Conrad Plake (DD)
 - Lukas Faulstich
 - Peter Palaga
 - Emre Kutbay
 - Torsten Schiemann
 - Markus Pankalla
 - Long Nguyen
 - Kevin Arnould
- Max-Planck-Institute for Molecular Genetics
 - Edda Klipp, Sebastian Schmeier, Axel Kowald
- European Bioinformatics Institute
 - Harald Kirsch, Dietrich Rebholz-Schumann

