

# Recognition of Bio-molecular Events in Text: The BioNLP Shared Task

Lucy Vanderwende  
Senior Researcher, **NLP** group  
Microsoft Research

Joint work with  
Chris Quirk, Pallavi Choudhury, and Michael Gamon

May 5, 2011

A CONVERSATION OVERHEARD

- Who? SIG-BioMed

- Biomedical language processing poses specific technical challenges that make it of interest to general NLP practitioners and make it of compelling importance to the larger field of biomedical informatics. The purpose of the BioMed SIG is to bring together researchers in NLP, bioinformatics, medical informatics, and computational biology, ...

- What? Genia Corpus

- ... a corpus of annotated abstracts taken from National Library of Medicine's MEDLINE database. In GENIA Corpus we annotate a subset of the substances and the biological locations involved in reactions of proteins, based on a data model ([GENIA ontology](#)) of the biological domain, in XML format ([GPML](#)).
- GENIA Corpus Version 3.0x consists of 2000 abstracts. The base abstracts are selected from the search results with keywords (MeSH terms) *Human, Blood Cells, and Transcription Factors*.
- annotators who are biologists, in order to get qualified interpretations from a biological perspective. These annotators are not systematically aware of linguistic phenomena.

# Why? Information Overload

- Search engines dominate information gathering
  - PubMed
  - Web as corpus
- Search engines – today – are limited ... require a lot of human effort to make sense of the information served up
- Can you find the information you really need to gather?
  - Are key words or key phrases enough?
    - Some spelling, some alternate terms (domain-specific at best only for domain-specific search engines)

# Talk Overview

- Goal for this talk
  - Update non-NLP researchers on state of the art for Information Extraction
- The BioNLP shared task
  - <http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/SharedTask>
  - Challenges
- Description of the MSR system for BioNLP
- System results
- Translating these results into action
  - Human-machine collaboration
  - Information visualization

# Example

In this study we hypothesized that the phosphorylation of TRAF2 inhibits binding to the CD40 cytoplasmic domain.

T1 Protein 57 62 TRAF2  
T2 Protein 88 92 CD40



- 1) Phosphorylation of TRAF2
- 2) Binding of TRAF2 to CD40
- 3) Instance of negative regulation
- 4) It's the phosphorylation event that neg-regulates the binding event

# Example

In this study we hypothesized that the phosphorylation of TRAF2 inhibits binding to the CD40 cytoplasmic domain.

T1 Protein 57 62 TRAF2  
T2 Protein 88 92 CD40



T4 Phosphorylation 39 54 phosphorylation

E1 Phosphorylation:T4 Theme:T1

T5 Binding 73 80 binding

E2 Binding:T5 Theme1:T1 Theme2:T2

T6 Negative\_regulation 64 72 inhibits

E3 Negative\_regulation:T6 Theme:E2 Cause:E1

# Bio-Event:

## State change of bio-molecules

- Genia Corpus
  - Gene expression
  - Transcription
  - Protein catabolism
  - Localization
  - Phosphorylation
  - Binding
  - Regulation
  - Positive regulation
  - Negative regulation
- Epigenetics and Post-translational Modifications (EPI)
  - (De)Hydroxylation
  - (De)Phosphorylation
  - (De)Ubiquitination
  - (De) DNA methylation
  - (De)Glycosylation
  - (De)Acetylation
  - (De)Methylation
  - Catalysis



# Why is this challenging?

- Many ways to refer to one event
  - Negative\_regulation
    - 532 inhibited, 252 inhibition, 218 inhibit, 207 blocked, 175 inhibits, 157 decreased, 156 reduced, 112 suppressed, 108 decrease, 86 inhibitor, 81 Inhibition, 68 inhibitors, 67 abolished, 66 suppress, 65 block, 63 prevented, 48 suppression, 47 blocks, 44 inhibiting, 42 loss, 39 impaired, 38 reduction, 32 down-regulated, 29 abrogated, 27 prevents, 27 attenuated, 26 repression, 26 decreases, ...
- One word can refer to many events
  - “detected”
    - Gene\_expression(0.38)   Positive\_regulation(0.17)   Transcription(0.38)   Binding(0.03)  
Negative\_regulation(0.03)
- Complex nested event-argument structures
  - (phosphorylation of TRAF2) inhibits (binding ...

# **MICROSOFT RESEARCH SYSTEM FOR BIONLP**

## Read Data

- Parse Input: Split sentences, tokenization, mark tokens as proteins and triggers

## Data Preparation

- Obtain constituency parses (50 best or 1 best); McClosky-Charniak 2008, McClosky trained on Genia 2010, ENJU (U Tokyo)
- Optionally compute posterior probabilities for all parse edges
- Transform to labeled dependency parses using Stanford Dependency parser
- Optionally apply dependency conversion rules

## Feature Extraction

- Word-based
- Frequency-based
- Dependency parse-based
- Cluster-based

## Trigger Detection

- Train SVM models

## Edge Detection

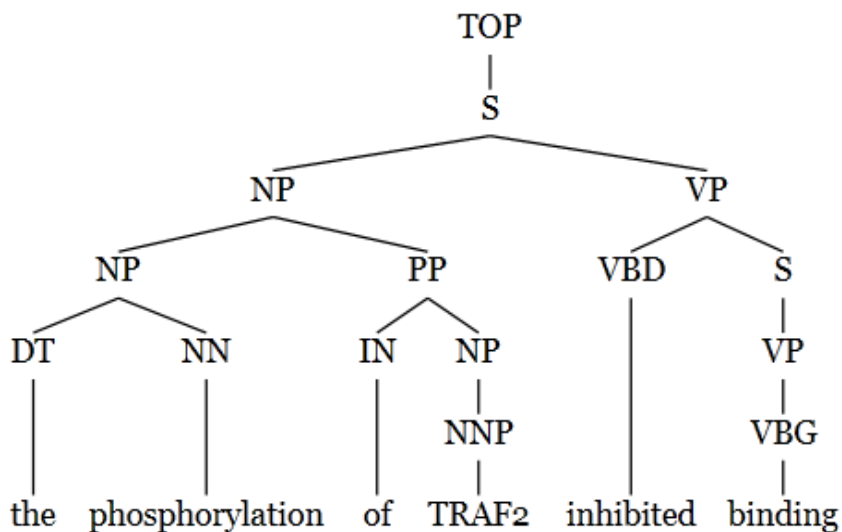
- Train MaxEnt models

## Post Processing

- Remove Cycles
- Remove unwarranted edges

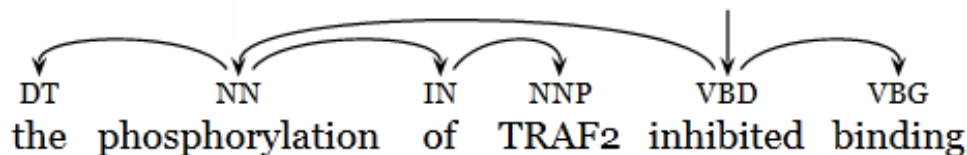
# Data preparation - parsing

- Constituency parse



- McClosky-Charniak 2008 - general
- McClosky 2010 – trained on Genia
- EnJu, ...
- One-best or n-best
- Posterior probabilities or none

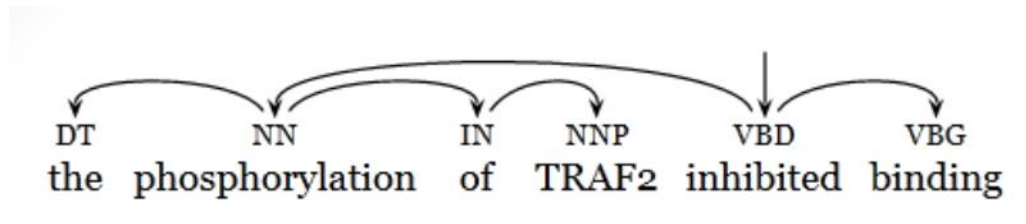
- Stanford Dependency parser, labeled



- Additional Conversion Rules
- Posterior probabilities or none

# Data preparation

- Create features from the following abstraction:



det(phosphorylation-2, the-1)

```
nsubj(inhibited-5, phosphorylation-2)
```

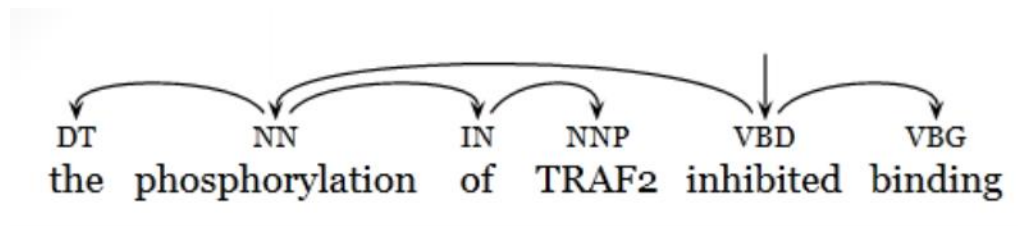
dobj(inhibited-5, binding-6)

```
prep_of(phosphorylation-2, TRAF-2-4)
```

dobj(phosphorylation-2, TRAF-2-4) -- optional conversion rule

# Trigger Detection

- Is the word a possible trigger?



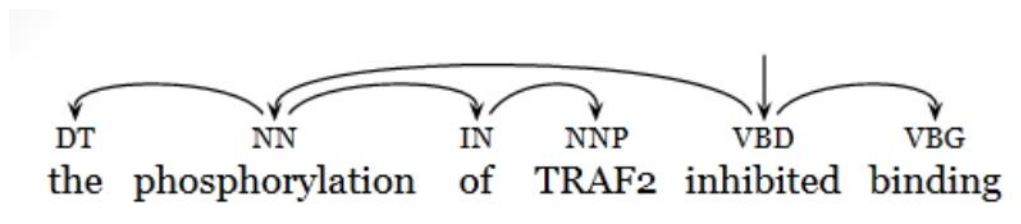
? The	no
? Phosphorylation	yes
? Of	no
? TRAF-2	protein
? Inhibited	yes
? Binding	yes

# Trigger Feature Extraction

- For each word, e.g., “phosphorylation”
  - Stem = phosphoryl
  - Bi-/trigrams “ph” “pho” “ho” “hos” “os” “osp” ...
  - Uppercase Y/N, has\_number Y/N
  - Was the word a trigger before? Y/N
- Frequency-based features
  - what are co-occurring entities in the sentence?
  - Co-occurring words in the sentence?

# Trigger Feature Extraction

- Dependency-based features
  - 1-, 2-, and 3-hop dependency paths types
  - 1-, 2-, and 3-hop dependency paths lexicalized
  - Shortest path from word to protein

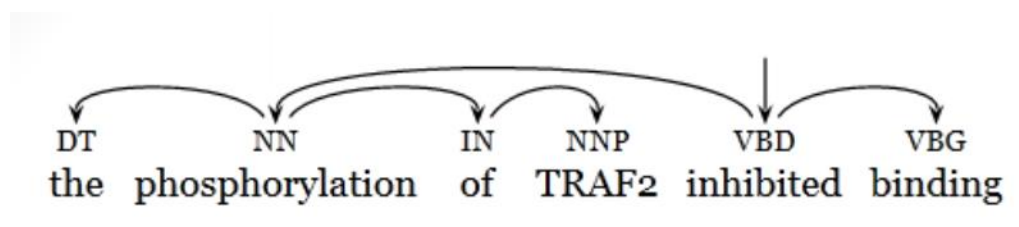


- Cluster-based features
  - To which automatically-generated cluster does word belong?
    - {activation, localization, ... lymphoma, malaria}



# Edge Detection

- Only 2 possible edge types, trigger-trigger and trigger-protein



?	Phosphorylation - TRAF-2	yes	trigger – protein	theme
?	Phosphorylation – inhibited	yes	trigger – trigger	cause
?	Phosphorylation – binding	no		
?	Inhibited – TRAF-2	no		
?	Inhibited – binding	yes	trigger – trigger	theme
?	Binding – TRAF-2	no		

# Edge Detection Features

- All of the features for trigger detection, plus
- For a trigger, is the possible edge part of the path to the nearest protein
- For a trigger, what protein/trigger type is the second node
  - Phosph. – Theme – binding – low probability
  - Phosph. – Theme – PROT – high probability

# Training

- In conjunction with previously published data on c-Raf-induced **phosphorylation** of GABP factors ...
- dt\_ti:1 bow\_.:1 t1HOut\_JJ:1 dist\_1POS\_NN:1 dep\_dist\_dist\_2nn:1 dist\_2POS\_NN:1 dist\_2isName:1 dist\_2annType\_Protein:1 dist\_2txt\_NAMED\_ENT:1 dist\_3POS\_JJ:1 POS\_NN:1 linear\_-1\_POS\_JJ:1 linear\_2\_POS\_NN:1 dep\_dist\_dist\_2dep:1 dist\_2POS\_JJ:1 dt\_at:1 linear\_-2\_POS\_JJ:1 linear\_-3\_POS\_JJ:1 dep\_dist\_dist\_2amod:1 nonstem\_ation:1 tt\_ati:1 tt\_tio:1 tt\_ion:1 dt\_io:1 dt\_on:1 linear\_1\_POS\_IN:1 t1HOut\_amod:1 t1HOut\_amod\_JJ:1 dep\_dist\_dist\_3amod:1 chain\_dist\_dist\_3-rev\_amod:1 dt\_or:1 linear\_-3\_isName:1 linear\_-3\_annType\_Protein:1 linear\_-3\_txt\_NAMED\_ENT:1 dt\_ph:1 **bow\_previously:1** bow\_of:1 bow\_(.:1 dist\_1POS\_IN:1 dep\_dist\_dist\_1advmod:1 dist\_1txt\_previously:1 dist\_1POS\_RB:1 dep\_dist\_dist\_1appos:1 linear\_1\_txt\_of:1 t1HOut\_prep\_of:1 t1HOut\_NNS:1 t1HOut\_prep\_of\_NNS:1 dep\_dist\_dist\_3prep\_of:1 dist\_3POS\_NNS:1 chain\_dist\_dist\_3-rev\_prep\_of:1 chain\_dist\_dist\_2-rev\_prep\_of-rev\_nn:1 linear\_3\_POS\_NNS:1 t1HIn\_NNS:1 dt\_la:1 tt\_ory:1 dt\_ry:1 dt\_sp:1 dt\_ho:1 bow\_.:1 dist\_2POS\_VBN:1 tt\_pho:1 tt\_lat:1 bow\_with:1 bow\_-:1 linear\_-2\_txt\_-:1 dep\_dist\_dist\_2hyphen:1 chain\_dist\_dist\_2-rev\_amod-rev\_hyphen:1 bow\_in:1 bow\_on:1 t1HIn\_prep\_on:1 dep\_dist\_3prep\_on:1 chain\_dist\_dist\_3-frw\_prep\_on:1 dt\_os:1 nameCount\_1:1 tt\_hor:1 dep\_dist\_2prep\_with:1 bow\_induced:1 dist\_3txt\_induced:1 linear\_-1\_txt\_induced:1 bow\_factors:1 t1HOut\_factors:1 dist\_3txt\_factors:1 linear\_3\_txt\_factors:1 **stem\_phosphoryl:1** tt\_hos:1 tt\_osp:1 tt\_sph:1 tt\_ryl:1 tt\_yla:1 dt\_yl:1 **stem\_gaz\_Phosphorylation:1** bow\_phosphorylation:1 t1HOut\_induced:1 t1HOut\_amod\_induced:1 txt\_phosphorylation:1 t1HOut\_prep\_of\_factors:1 t1HIn\_prep\_on\_NNS:1 chain\_dist\_dist\_2-frw\_prep\_on-rev\_amod:1 bow\_data:1 t1HIn\_data:1 dist\_3txt\_data:1 dep\_dist\_1pobj:1 dist\_1txt\_in:1 bow\_A:1 dist\_1txt\_A:1 dist\_2POS\_NNP:1 chain\_dist\_dist\_2-frw\_prep\_on-rev\_dep:1 chain\_dist\_dist\_2-frw\_prep\_on-frw\_prep\_with:1 bow\_conjunction:1 dist\_2txt\_conjunction:1 bow\_published:1 dist\_2txt\_published:1 chain\_dist\_dist\_1-frw\_prep\_on-rev\_amod-rev\_advmod:1 bow\_GABP:1 dist\_2txt\_GABP:1 **t1HIn\_prep\_on\_data:1** chain\_dist\_dist\_1-frw\_prep\_on-rev\_dep-rev\_appos:1 # phosphorylation

# Testing

- In conjunction with previously published data on c-Raf-induced **phosphorylation** of GABP factors ...

stem_gaz_Phosphorylation	(0.41)
dt_or	(0.35)
dt_os	(0.33)
stem_phosphoryl	(0.33)
tt_sph	(0.30)
tt_osp	(0.30)
tt_hos	(0.30)
tt_pho	(0.29)
tt_ryl	(0.29)
tt_hor	(0.29)

## Class Scores:

Phosphorylation:	0.602
None:	0.150
Binding:	0.005
Regulation:	0.004
Transcription:	0.003
Localization:	0.002
Phosphorylation/Positive_regulation:	0.002
Negative_regulation:	0.001
Positive_regulation:	0.001
Phosphorylation/Negative_regulation:	0.001

# System Results on Genia

Development Set				Test Set			
Event Class	Recall	Precision	F1		Recall	Precision	F1
Gene_expression	76.37	81.46	78.83		73.95	73.22	73.58
Transcription	49.37	73.58	59.09		41.95	65.18	51.05
Protein_catabolism	69.57	80.00	74.42		46.67	87.50	60.87
Phosphorylation	73.87	84.54	78.85		87.57	81.41	84.37
Localization	74.63	75.76	75.19		51.31	79.03	62.22
=[SVT-TOTAL]=	72.02	80.51	76.03		68.99	74.03	71.54
Binding	47.99	50.85	49.38		42.36	40.47	41.39
=[EVT-TOTAL]=	65.97	72.73	69.18		62.63	65.46	64.02
Regulation	32.53	47.05	38.62		24.42	42.92	31.13
Positive_Regulation	38.74	51.67	44.28		37.98	44.92	41.16
Negative_Regulation	35.88	54.87	43.39		41.51	42.70	42.10
=[REG-TOTAL]=	36.95	51.79	43.13		36.64	44.08	40.02
MSR-Total	50.20	62.60	55.72		48.64	54.71	51.50
FAUST-Total					49.41	64.75	56.04
UMass-Total					48.49	64.08	55.20
Uturku-Total					49.56	57.65	53.30

# System Results on Epigenetics

Development Set				Test Set			
Event Class	Recall	Precision	F1		Recall	Precision	F1
Hydroxylation	25.81	61.54	36.36		30.43	84.00	44.68
Dehydroxylation	100.00	100.00	100.00		100.00	100.00	100.00
Phosphorylation	71.88	85.19	77.97		72.31	85.45	78.33
Dephosphorylation	0.00	0.00	0.00		0.00	0.00	0.00
Ubiquitination	63.16	75.00	68.57		67.78	81.88	74.16
Deubiquitination	0.00	0.00	0.00		0.00	0.00	0.00
DNA_methylation	72.73	72.18	72.45		71.43	73.86	72.63
DNA_demethylation	0.00	0.00	0.00		0.00	0.00	0.00
Glycosylation	61.43	67.19	64.18		39.05	69.47	50.00
Deglycosylation	0.00	0.00	0.00		0.00	0.00	0.00
Acetylation	89.23	75.32	81.69		87.42	85.28	86.34
Deacetylation	68.42	92.86	78.79		62.50	93.75	75.00
Methylation	64.62	75.00	69.42		62.18	73.62	67.42
Demethylation	0.00	0.00	0.00		0.00	0.00	0.00
Catalysis	3.33	15.38	5.48		4.50	33.33	7.94
====[MSR TOTAL]====	57.22	72.23	63.85		55.70	77.60	64.85
UTurku Total					68.51	69.208	68.86
FAUST Total					59.88	80.25	68.59
UMASS Total					57.04	73.30	64.15

A CONVERSATION JOINED  
QUESTIONS ASKED

# Translating these results into ...

- Given input text:
  - The B cells were found to express BMP type I and type II receptors and BMP-6 rapidly induced phosphorylation of Smad1/5/8.
    - T6 Protein 561 566 BMP-6
    - T7 Protein 602 607 Smad1
    - T8 Protein 608 609 Smad5
    - T9 Protein 610 611 Smad8
- System Output
  - T31 Positive\_regulation 575 582 induced
  - T32 Phosphorylation 583 598 phosphorylation
  - E2 Positive\_regulation:T31 Theme:E3 Cause:T6
  - E4 Positive\_regulation:T31 Theme:E5 Cause:T6
  - E6 Positive\_regulation:T31 Theme:E7 Cause:T6
  - E3 Phosphorylation:T32 Theme:T7
  - E5 Phosphorylation:T32 Theme:T8
  - E7 Phosphorylation:T32 Theme:T9



# Translating these results into ...

- Given input text:
  - The B cells were found to express BMP type I and type II receptors and BMP-6 rapidly induced phosphorylation of Smad1/5/8.
    - T6 Protein 561 566      BMP-6
    - T7 Protein 602 607      Smad1
    - T8 Protein 608 609      Smad5
    - T9 Protein 610 611      Smad8
- “Friendly facts”
  - Positive\_regulation(BMP-6, phosphorylation of Smad1)
  - Positive\_regulation(BMP-6, phosphorylation of Smad5)
  - Positive\_regulation(BMP-6, phosphorylation of Smad8)
  - Phosphorylation( Smad1)
  - Phosphorylation( Smad5)
  - Phosphorylation( Smad8)

# “friendly facts”

- A better search engine
  - Solves the time-consuming nature of search
  - Solves the memory recall problem
  - Solves the problem of not knowing where to look, if all categories can be anticipated
- Entity focused ... who is? What is?
- You have to know what you're looking for



# Using graphs

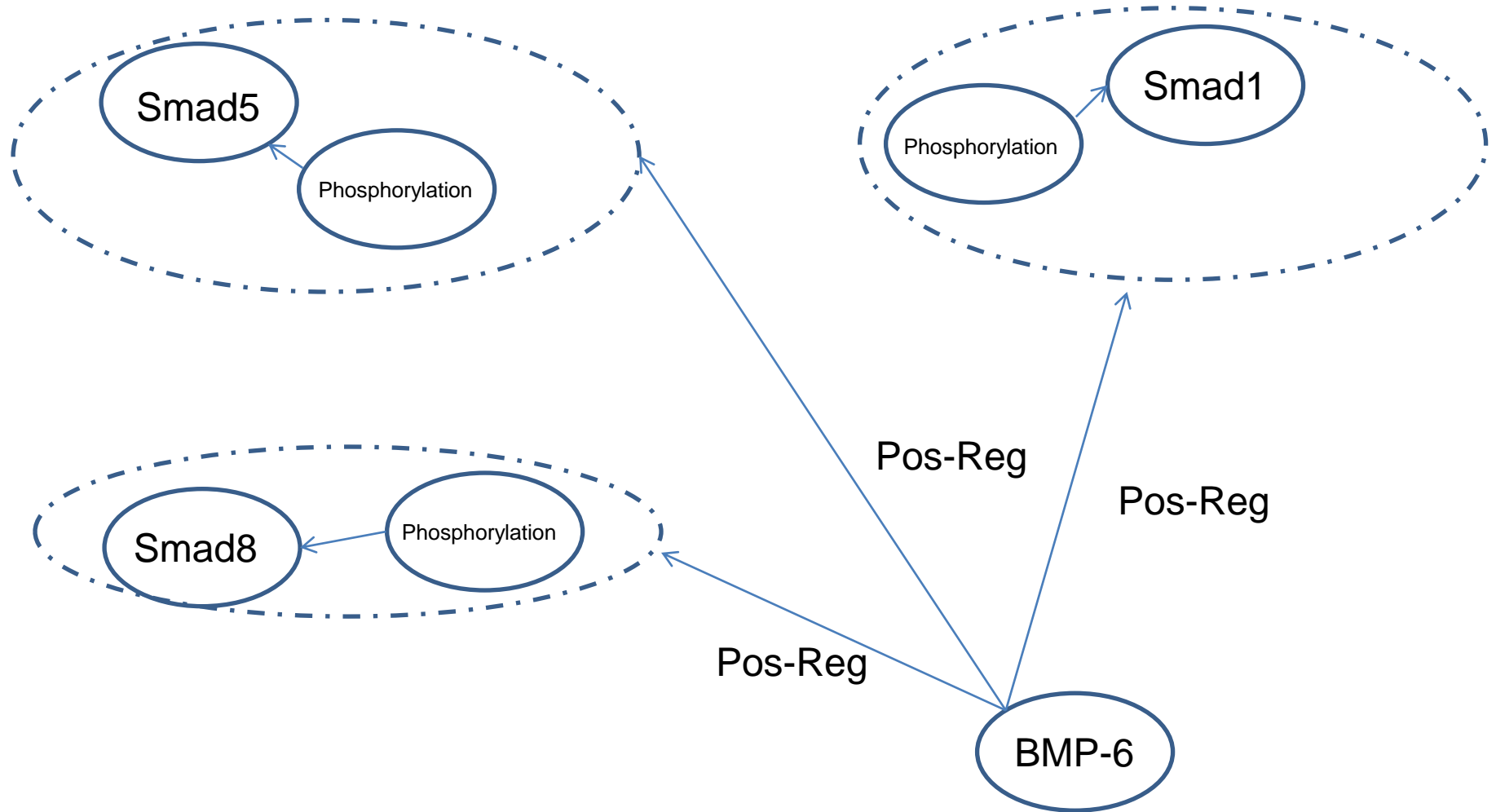
Natarajan, J. D. Berrar, W. Dubitzky, C. Hack, Y. Zhang, C. DeSesa, J. Van Brocklyn, E. Bremer.

Text mining of full-text journal articles combined with gene expression analysis reveals a relationship between sphingosine-1-phosphate and invasiveness of a glioblastoma cell line.

In BMC Bioinformatics, 2006, 7:373.

- They mined full-text articles and inferred gene-gene interaction networks for these 72 genes, from which they identified at least one interesting network which they further pursued.
- The paper shows success in applying information extraction to improve the workflow of biomedical researchers by adding improved search tools for literature review.

# Translating these results into ... a graph

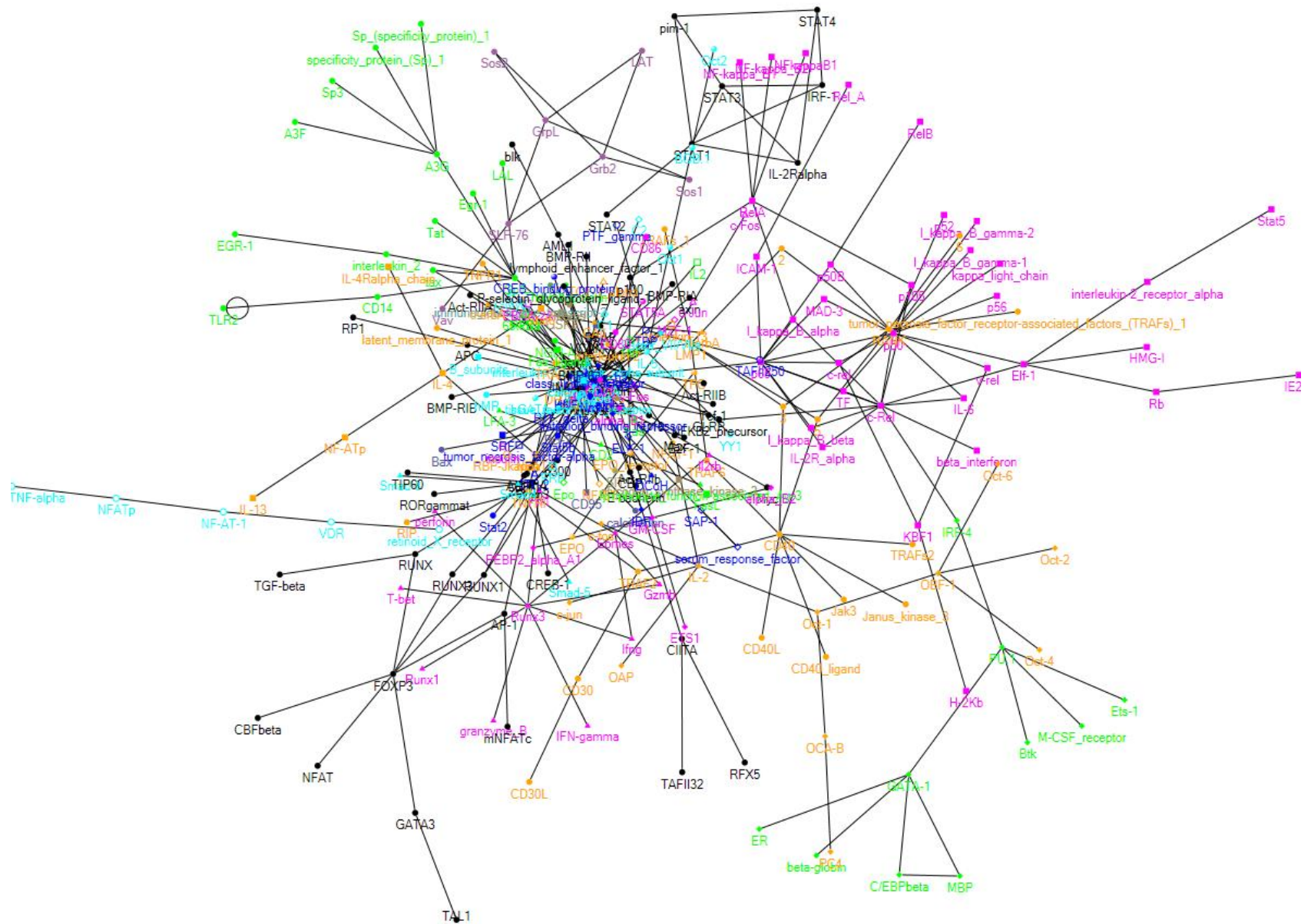


# NodeXL, an Excel template

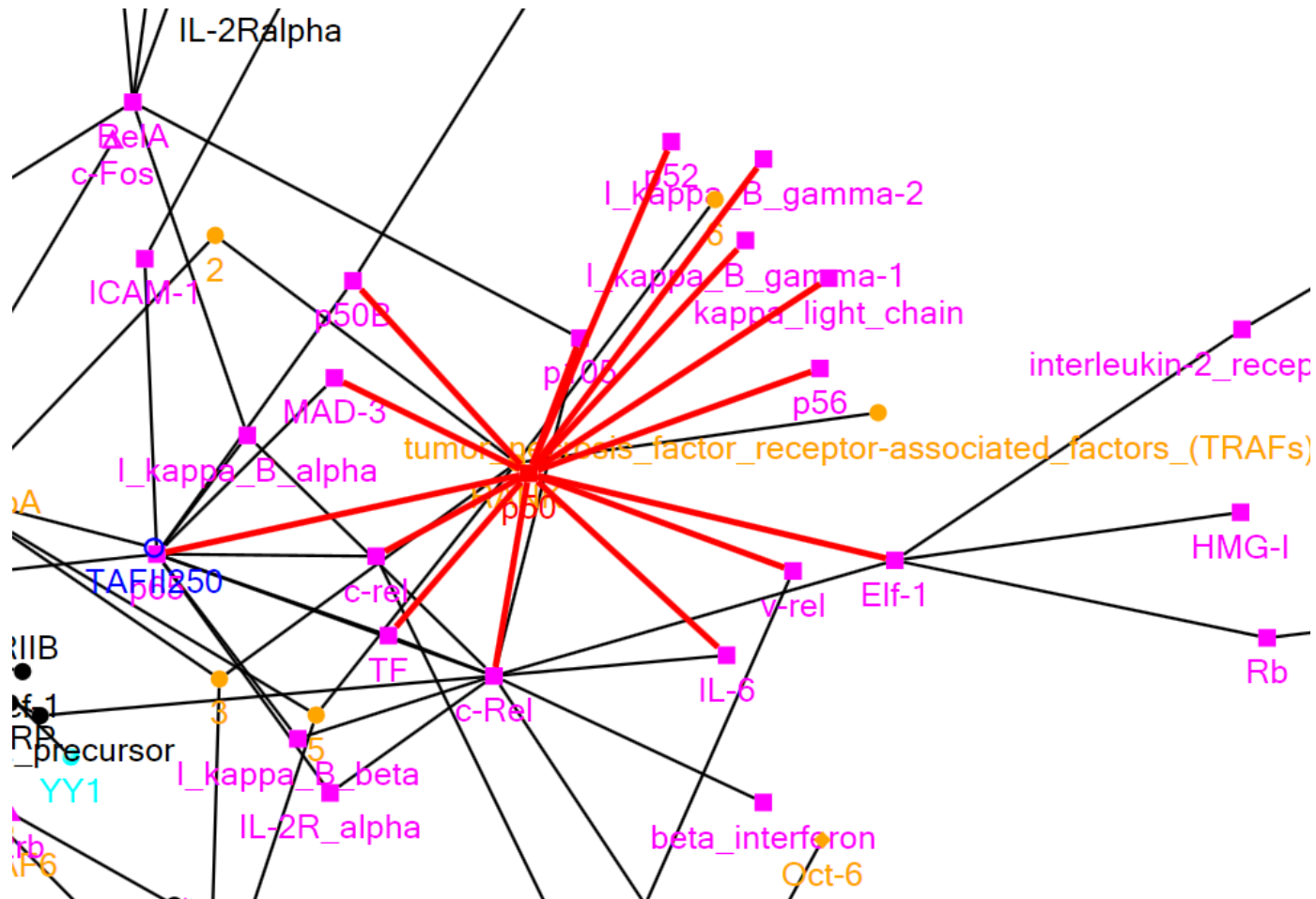
<http://nodexl.codeplex.com>

[illegible]

# A different view of Binding

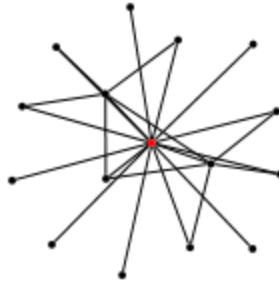


# Focus on binding relations for P50

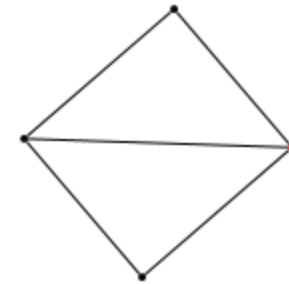


# An abstract view of binding properties using relations up to 2 edges away

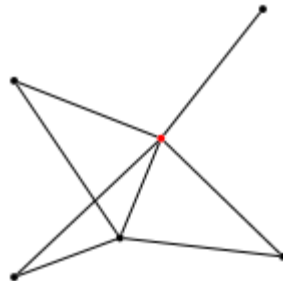
- P50



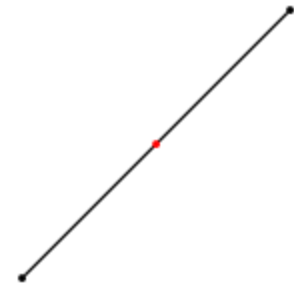
- P300



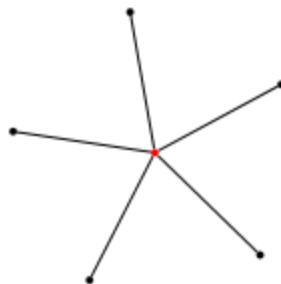
- STAT1



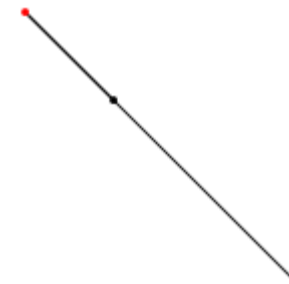
- C/EBPbeta



- PU.1



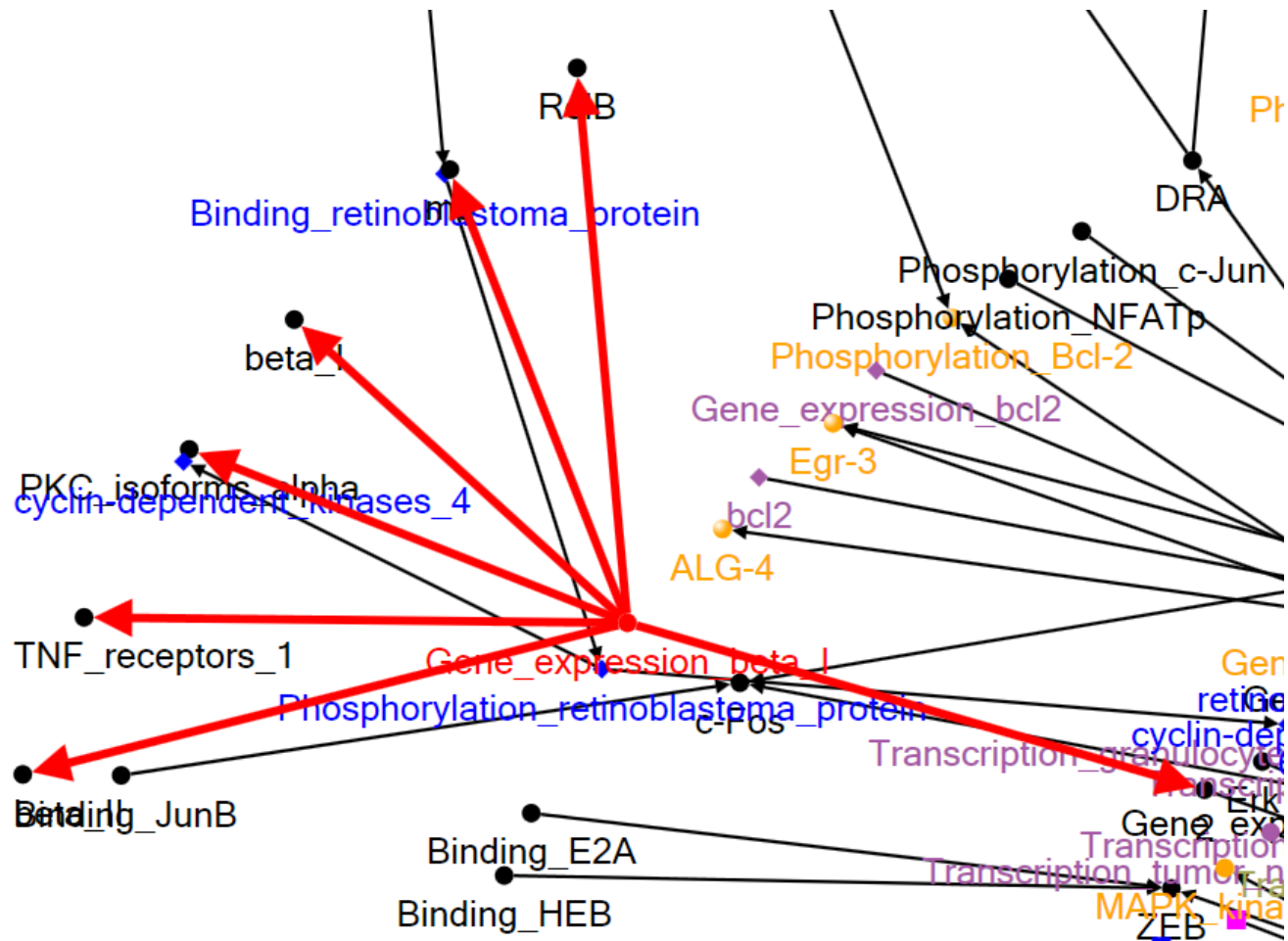
- TRAF6





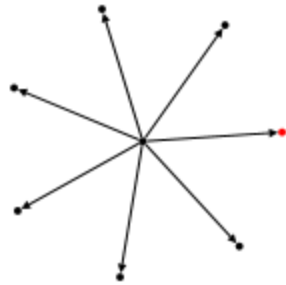


# Focus on regulations

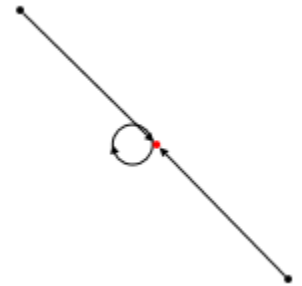


# An abstract view of regulation properties using relations up to 2 edges away

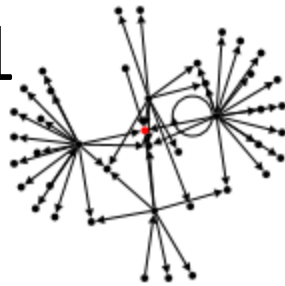
- Beta\_1



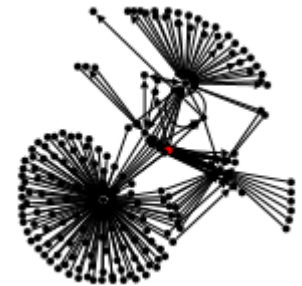
- Binding\_CD10



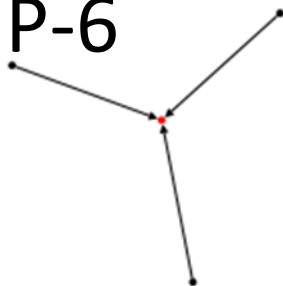
- Binding\_AP1



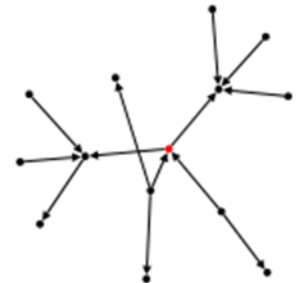
- Gene\_exp\_Foxp3



- Binding\_BMP-6



- Interleukin-6



# ActiveText, Microsoft External Research project

## Document Information

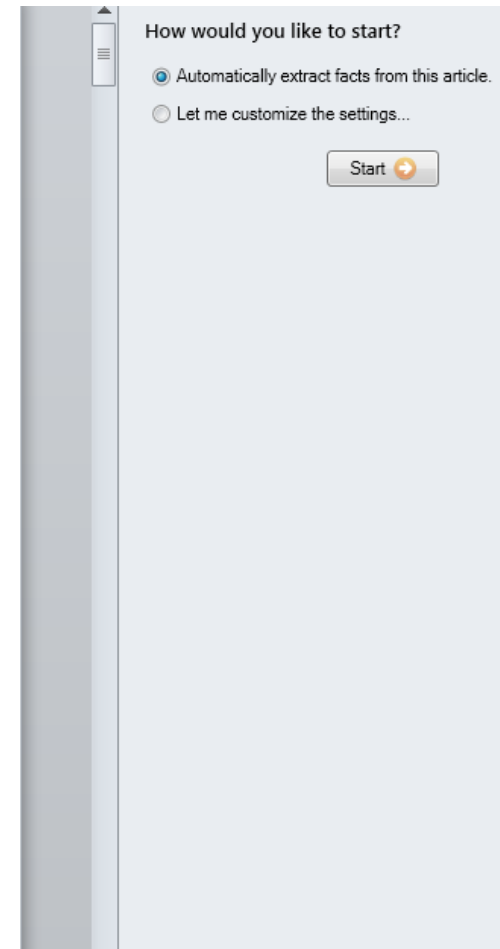
**Title:** Phenotypic characterization of a glucose transporter null mutant in *Leishmania mexicana*

**Authors:** Dayana Rodriguez-Contreras; Xuhong Feng; Kristie M. Keeney; H. G. Archie Bouwer; Scott M. Landfear

**Year:** 2007

## Abstract

Glucose is a major source of energy and carbon in promastigotes of *Leishmania mexicana*, and its uptake is mediated by three glucose transporters whose genes are encoded within a single cluster. A null mutant in which the glucose transporter gene cluster was deleted by homologous gene replacement was generated previously and shown to grow more slowly than wild type promastigotes but not to be viable as amastigotes in primary tissue culture macrophages or in axenic culture. Further phenotypic characterization demonstrates that the null mutant is unable to import glucose, mannose, fructose, or galactose and that each of the three glucose transporter isoforms, LmGT1, LmGT2, and LmGT3, is capable of transporting each of these hexoses. Complementation of the null mutant with each isoform is able to restore growth in each of the four hexoses to wild type parasites. Null mutant promastigotes are reduced in size to about 2/3 the volume of wild type parasites. In addition, the null mutants are significantly more sensitive to oxidative stress than their wild type counterparts. These results underscore the importance of glucose transporters in the parasite life cycle and suggest reasons for their non-viability in the disease-causing amastigote stage.



How would you like to start?

☒ Automatically extract facts from this article.

☐ Let me customize the settings...

Start →

# ActiveText, Microsoft External Research project

Microscopic examination of *Δlmg* null mutant promastigotes suggests that they are smaller than either wild type parasites or null mutants that have been complemented with the LmGT2 glucose transporter gene on an episomal expression vector (*Δlmg*[pGT2]) (Fig. 6). To quantify this reduction in size, we measured cell volume by monitoring partitioning of  $^3\text{H}_2\text{O}$  and [ $^{14}\text{C}$ ] carboxyl-inulin into the three cell lines, and we determined the protein content per cell. The ratio for wild type/*Δlmg* cells was  $1.57 \pm 0.10$  ( $n=7$ ) for the cell volume and  $1.53 \pm 0.10$  ( $n=13$ ) for the protein content measurement, confirming that the *Δlmg* null mutants are significantly smaller than wild type promastigotes. Episomal complementation with LmGT2 was able to restore the size of *Δlmg* parasites, as the ratio for wild type/*Δlmg* [pGT2] cells was  $1.01 \pm 0.07$  ( $n=3$ ) for the cell volume and  $1.00 \pm 0.02$  ( $n=3$ ) for protein content. These results complement the previous observation that *Δlmg* promastigotes grow slowly and to a lower cell density compared to wild type parasites (Burchmore RJ, 2003).

It is notable that restriction of nutrients such as glucose and amino acids leads to a pronounced reduction in cell size in many eukaryotes. This size reduction is thought to be controlled by a signal transduction pathway in which the TOR protein kinase plays a central role (Sarbasov dos D, 2005). This TOR pathway is highly conserved among eukaryotes, and potential homologs of the TOR, raptor, and GβL proteins, that constitute the functional protein kinase, are encoded within the L. major genome (systematic names LmjF36.6320 and LmjF34.4530 for TOR, LmjF25.0610 for raptor, and LmjF10.0780 for GβL proteins). These observations suggest that the reduced size of *Δlmg* null mutants might be mediated by a TOR-raptor-GβL complex.

3.5 Increased susceptibility of *Δlmg* promastigotes to oxidative stress

As intracellular parasites, Leishmania are exposed to reactive oxygen species (ROS) such as  $\text{H}_2\text{O}_2$ , superoxide anion ( $\text{O}_2^-$ ) and hydroxyl radical ( $\text{OH}^\bullet$ ), which arise from a number of different mechanisms including those generated from its own aerobic metabolism (Krauth-Siegel RL, 2005) and by the host immune response (Stafford JL, 2002). To evaluate whether *Δlmg* parasites are more sensitive to oxidative stress, we exposed both wild type and *Δlmg* promastigotes to increasing concentrations of a

Back Add Sentence... Delete Sentence...

**Sentence**

This size reduction is thought to be controlled by a signal transduction pathway in which the TOR protein kinase plays a central role (Sarbasov dos D, 2005).

**Fact** [\[+\] Add Fact](#) [\[-\] Delete Fact](#)

protein  
kinase plays  
role

**Sentence**

As intracellular parasites, Leishmania are exposed to reactive oxygen species (ROS) such as  $\text{H}_2\text{O}_2$ , superoxide anion ( $\text{O}_2^-$ ) and hydroxyl radical ( $\text{OH}^\bullet$ ), which arise from a number of different mechanisms including those generated from its own aerobic metabolism (Krauth-Siegel RL, 2005) and by the host immune response (Stafford JL, 2002).

**Fact** [\[+\] Add Fact](#) [\[-\] Delete Fact](#)

Leishmania  
are exposed to  
oxygen

**Sentence**

The increased sensitivity to oxidative stress and the decreased reducing capacity of the *Δlmg* null mutant may reflect the inability of the null mutants to take up and metabolize glucose via the pentose phosphate pathway (Maugen DA, 2003).

**Fact** [\[+\] Add Fact](#) [\[-\] Delete Fact](#)

sensitivity  
may reflect  
inability

ActiveText :: Edit Fact

Subject:

☒ Leishmania

☐ Write your own:

Verb:

☒ are exposed to

☐ Write your own:

Object:

☒ oxygen

☐ reactive oxygen

☐ reactive oxygen species

☐ reactive oxygen (ROS) such as  $\text{H}_2\text{O}_2$ , superoxide anion ( $\text{O}_2^-$ ) and hydroxyl radical ( $\text{OH}^\bullet$ ), which arise from a number of different mechanisms including

☐ Write your own:

References:

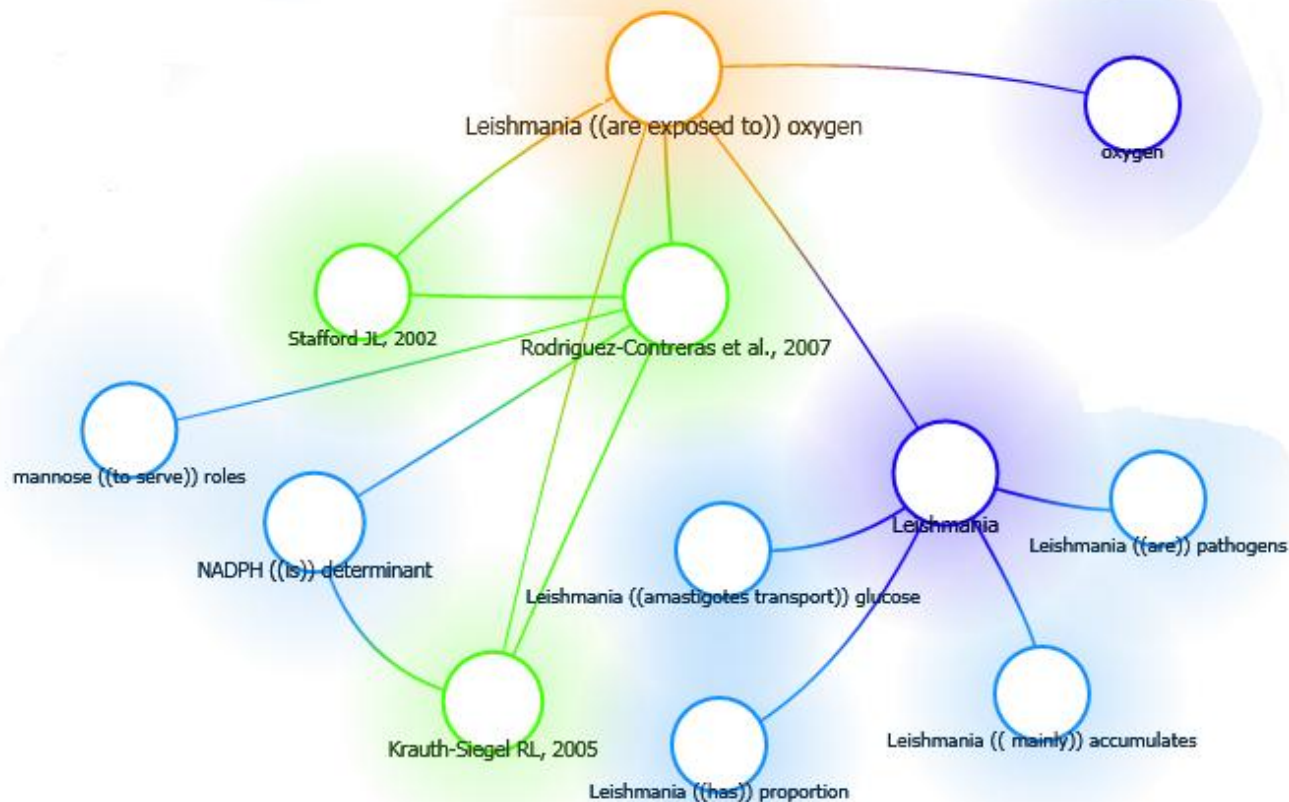
Krauth-Siegel RL, Bauer. Dithiol proteins as guardians of the intracellular redox milieu in parasites: old and new drug targets in trypanosomes and malaria-causing plasmodia. 2005

Stafford JL, Neumann. Macrophage-mediated innate host defense against protozoan parasites. 2002



# Visualizing the ActiveText data

Our hypothesis is that you might discover information you wouldn't have necessarily known to go looking for, so “not your old search engine”



# ActiveText

- Annotator can be either a reader or the author
- Annotator can accept/modify/reject the facts extracted
- Annotator can “write your own”, which hopefully will lead to high-level relation annotations, similar to “regulation” or “binding”

**A CONVERSATION CONTINUING**



# Summary

- The BioNLP shared task
- Machine Learning architecture with rich features derived from NLP (lexical, syntactic, semantic)
- Demonstrated feasibility of the task through system results
- Translating these results into action
  - Information visualization
  - Human-machine collaboration
- More communication between CS/NLP and Bioinformatics is essential to plan next steps