# GIN-IE: Interaction Extraction from the Literature

*Arzucan Ozgur[1], Dragomir R. Radev[1,2], and Alex Ade[3]*

*[1]CLAIR Lab, Department of EECS, [2]School of Information, [3]National Center for Integrative Biomedical Informatics, University of Michigan, Ann Arbor, MI*

## Motivation:

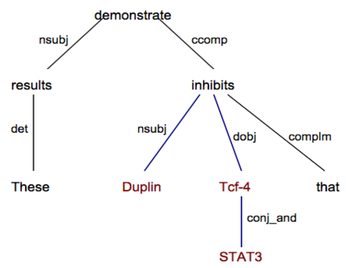
Challenge in extracting relevant information from vast amount of publications
- Biomedical literature is growing rapidly ( > 16 million articles in PubMed).
- Delay in including new discoveries to manually curated databases.
- Most information uncovered in unstructured text of biomedical publications.

## Approach:

Natural language processing, machine learning, and network analysis methods to extract biologically important information.

## Machine Learning and Dependency Parsing for Protein Interaction Extraction

"These results demonstrate that Duplin inhibits not only Tcf-4 but also STAT3."




Path1: **Duplin** – nsubj – inhibits – dobj – Tcf-4 – conj_and –**STAT3**
Path2: **Duplin** – nsubj – inhibits – dobj – **Tcf-4**
Path3: **Tcf-4** – conj_and – **STAT3**

Stanford Parser is used to generate the dependency parse trees (de Marneffe et al., 2006).

Define Path Edit Kernel:

$$edit\_sim(p_i, p_j) = e^{-\gamma(edit\_distance(p_i, p_j))}$$

Integrate path edit kernel and path cosine kernel with Support Vector Machines (SVM)

## Performance

### Data Sets

| Data Set | Sentences | + Sentences | - Sentences |
|---|---|---|---|
| **AIMED** | 4026 | 951 | 3075 |
| **CB** | 4056 | 2202 | 1854 |

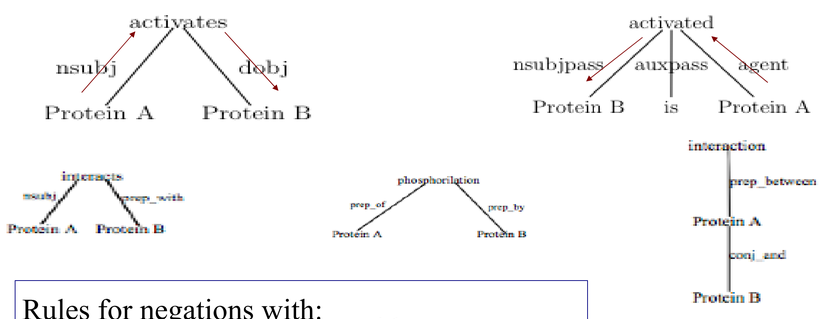
### Results: 10 fold cross-validation

| | Precision | Recall | F-measure |
|---|---|---|---|
| AIMED | 77.52 | 43.51 | 55.61 |
| CB | 85.15 | 84.79 | 84.96 |

## Dependency Tree Rules for Interaction Type and Directionality Extraction

- Type of relationship: Inhibition
- Directionality: Duplin ->Tcf-4; Duplin ->STAT3
- Real-life applications (integration to MiMI database): high precision in the expense of recall

Protein A activates Protein B.  Protein B is activated by Protein A.



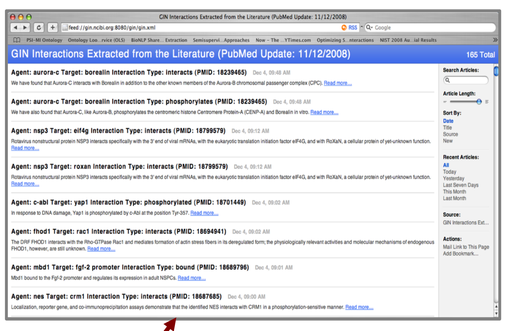

Rules for negations with:
not, no, fail, lack, and ...neither ...nor...

Precision: 94%, Recall: 18%

ex: The **lack** of cooperative interaction between E5 and the epidermal growth factor receptor...

## Integrating GIN-IE into the Pubmed NLP Pipeline

- Modified the GIN-IE scripts to work with the newest Pubmed 09 database schema
- Created a wrapper script to call each of the GIN-IE scripts in succession
- Created a script to merge the individual RSS feed files into a single file that's sent to the server with each batch update
- Currently, over 6,000 interactions have been added to the database and RSS feed
- The RSS feed is available at: http://gin.ncibi.org/rss/gin-ie/interactions.rss




Daily RSS Feed for the Extracted Interactions

## Detecting Speculations in Biomedical Articles

- We showed that the Roaz protein bound specifically to O/E-1 by using the yeast two-hybrid system. (Factual)
- We previously identified Ly6k as a candidate TEX101-associated protein, but as molecular probes are not currently available to detect Ly6k, **we do not have conclusive evidence** of the association between TEX101 and Ly6k. (Speculative)
- Like RAD9, RAD9B associates with HUS1, RAD1, and RAD17, **suggesting** that it is a RAD9 paralog that engages in similar biochemical reactions. (Speculative)

While speculative information might still be useful for biomedical scientists, it is important that it is distinguished from the factual information. (18% of sentences in Genia Abstracts are speculative)

### Speculation Keywords

- might, suggest, likely, hypothesize, could, predict, no evidence of, address the question of, remains to be elucidated, issue is raised, ...
- Not always used in speculative context : 1273 Genia Abstracts, 138 unique speculation keywords: Number of their occurrence is 6125. In only 2694 (less than 50%) of their occurrences used in speculative context.
  - Thus, it **appears** that the T-cell-specific activation of the proenkephalin promoter is mediated by NF-kappa B. (appears: speculative context)
  - Differentiation assays using water soluble phorbol esters reveal that differentiation becomes irreversible soon after AP-1 **appears**. (appears: non-speculative context, becoming visible)

**Approach:** Solving two sub-problems:
- Identifying speculation keywords – supervised classification task
- Resolving their linguistic scopes – syntactic structures of the sentences

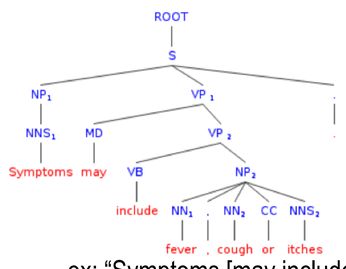
### Genia Abstracts - Results:

| Method | Recall | Precision | F-Measure |
|---|---|---|---|
| Baseline 1 | 52.84 | 92.71 | 67.25 |
| Baseline 2 | 97.54 | 43.66 | 60.30 |
| BOW 3 - stemmed | 81.47 | 92.36 | 86.51 |
| BOW 2 - stemmed | 81.56 | 93.29 | 86.97 |
| BOW 1 - stemmed | 83.08 | 93.83 | 88.05 |
| BOW 3 | 82.58 | 92.04 | 86.98 |
| BOW 2 | 82.77 | 92.74 | 87.41 |
| BOW 1 | 83.27 | 93.67 | 88.10 |
| KW: kw, kw-stem, kw-pos | 88.62 | 92.77 | 90.61 |
| KW, DEP | 88.77 | 92.67 | 90.64 |
| KW, DEP, BOW 1 | 88.46 | 94.71 | 91.43 |
| KW, DEP, BOW 1, POS | 88.16 | 95.21 | 91.50 |
| KW, DEP, BOW 1, POS, CO-KW | 88.22 | 95.56 | 91.69 |

Various Types of Features that represent the context of the keyword:
- KW: Keyword specific features (kw, stem, part-of-speech)
- BOW: Surrounding words (window sizes: 1, 2, or 3)
- DEP: Dependency relation features (used with a clausal complement, infinitival clause, negation, auxiliary)
- POS: Positional features (Title, Figure or Table Legend, Results and Discussion Section, Conclusion, Materials and Methods, Last or First sentence of abstracts)
- CO-KW: Other co-occurring keywords

### Resolving Speculation Scopes

Define rules to resolve the scopes of the keywords based on the part-of-speech of the keywords and the syntactic structures of the sentences.




ex: "Symptoms [may include {fever, cough or itches}]."

Conjunction: The phrase it is attached to
Modal verbs: from the keyword to the end of the sentence (clause)
Adjective or adverb: following noun phrase or whole sentence
Verb followed with an infinitival clause: whole sentence
Default rule: from keyword to the end of the sentence

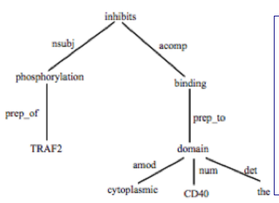
| Method | Accuracy |
|---|---|
| Baseline 1 | 4.82 |
| Baseline 2 | 67.60 |
| Rule-based method | 74.14 |

Genia Abstracts
10 fold cross-validation results

## Extracting Biomedical Events: BioNLP'09 Shared Task on Event Extraction

- Lexical and Part-of-Speech Features (trigger and its POS)
- Positional Features (relative position and distance)
- Dependency Relation Features

Class 1: Single Theme (e.g. Phosphorylation)
Class 2: Multiple Themes (e.g Binding)
Class 3: A theme and a cause (e.g. Regulation)




"The **phosphorylation of TRAF2** inhibits binding to the CD40 cytoplasmic domain."
dependency relationship type path from trigger to participant:
(phosphorylation, TRAF2): prep_of
(phosphorylation, CD40): nsubj acomp prep_to num

| Event Type | Recall | Precision | F-measure |
|---|---|---|---|
| Localization | 41.95 | 60.83 | 49.66 |
| Binding | 31.41 | 34.94 | 33.08 |
| Gene_expression | 61.36 | 69.00 | 64.96 |
| Transcription | 37.23 | 30.72 | 33.66 |
| Protein_catabolism | 64.29 | 64.29 | 64.29 |
| Phosphorylation | 68.15 | 80.70 | 73.90 |
| Event Total | 50.82 | 56.80 | 53.64 |
| Regulation | 15.12 | 19.82 | 17.15 |
| Positive_regulation | 24.21 | 33.33 | 28.05 |
| Negative_regulation | 21.64 | 32.93 | 26.11 |
| Regulation Total | 22.02 | 30.72 | 25.65 |
| All Total | 35.86 | 44.69 | 39.79 |
| Best System | 46.73 | 58.48 | 51.95 |
| Median System | 25.96 | 36.26 | 30.26 |

## References:

- A. Ozgur and D. R. Radev. Supervised classification for extracting biomedical events. Proceedings of the BioNLP'09 Workshop Shared Task on Event Extraction at NAACL-HLT, Boulder, Colorado, 2009. (To appear)
- A. Ozgur, T. Vu, G. Erkan, and D. R. Radev. "Identifying gene-disease associationsusing centrality on a literature mined gene interaction network", Bioinformatics, Vol. 24, Num. 13, pp. i277-i285, 2008.
- F. Leitner, M. Krallinger, C. Rodriguez-Penagos, J. Hakenberg, C. Plake, C-J. Kuo, C-N. Hsu, R. T-H. Tasi, H-C. Hung, W. W. Lau, C. A. Johnson, R. Saetre, K. Yoshida, Y. H. Chen, S. Kim, S-Y. Shin, B-T. Zhang, W. A. Baumgartner, Jr., L. Hunter, B. Haddow, M. Matthew, X. Wang, P. Ruch, F. Ehrler, A. Ozgur, G. Erkan, D. R. Radev, M. Krauthammer, T. Luong, R. Hoffmann, C. Sander, and A. Valencia. "Introducing meta-services for biomedical information extraction", Genome Biology, 9(S2):S6, 2008.
- G. Erkan, A. Ozgur, and D. R. Radev. "Semi-supervised classification for extracting protein interaction sentences using dependency parsing", In Proceedings of EMNLP, Prague, Czech Republic, June 28-30 2007.

## Acknowledgements

This work was supported by National Institutes of Health: Grant #U54 DA021519.

We would like to thank Glenn Tarcea, Zach Wright, Terry Weymouth,and H. V. Jagadish for their contributions to the project.