Integrated network construction using event based text mining

Yvan Saeys, Sofie Van Landeghem and Yves Van de Peer

Bioinformatics and Evolutionary Genomics Group
Ghent University, Belgium
http://bioinformatics.psb.ugent.be

Saturday September 5th, 2009

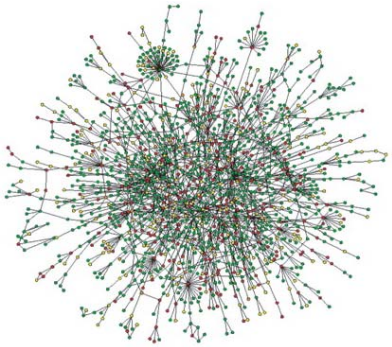
MLSB 2009

Integrated network construction

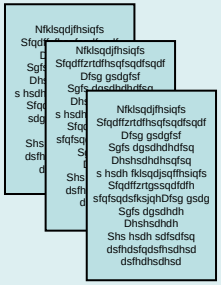
Databases **Text**

Integration

Construction of interaction network



Information retrieval



Named entity recognition

xxxxxx
xxxx xxxxxxxx
xx **CDC42**
xxxxx xxx xxx
xxxxx **PAK4**
KTN1 xxx xx
xxxxxxxx xx

Structured text

```
<sentence>
  <text>xxxxx</text>
  <prot>CDC42</prot>
  <prot>PAK4</prot>
   ….
</sentence>
```

Information extraction

interaction(CDC42, PAK4)
interaction(CDC42, KTN1)
…

Text mining in systems biology

- Up to last year, main focus on extracting protein-protein interactions (PPI)
  - Co-occurrence based approaches [Ding et al. 2002, Hoffmann and Valencia 2004]
  - Hand crafted rules [Fundel et al. 2007]
  - Machine learning approaches [Zelenko et al. 2003, Bunescu et al. 2005, Airola et al. 2008]
- 2009 (BioNLP ’09 Shared task):
  - Multiple types of interactions: “events”
  - Similar to ML challenge: given training and validation set, hidden test set

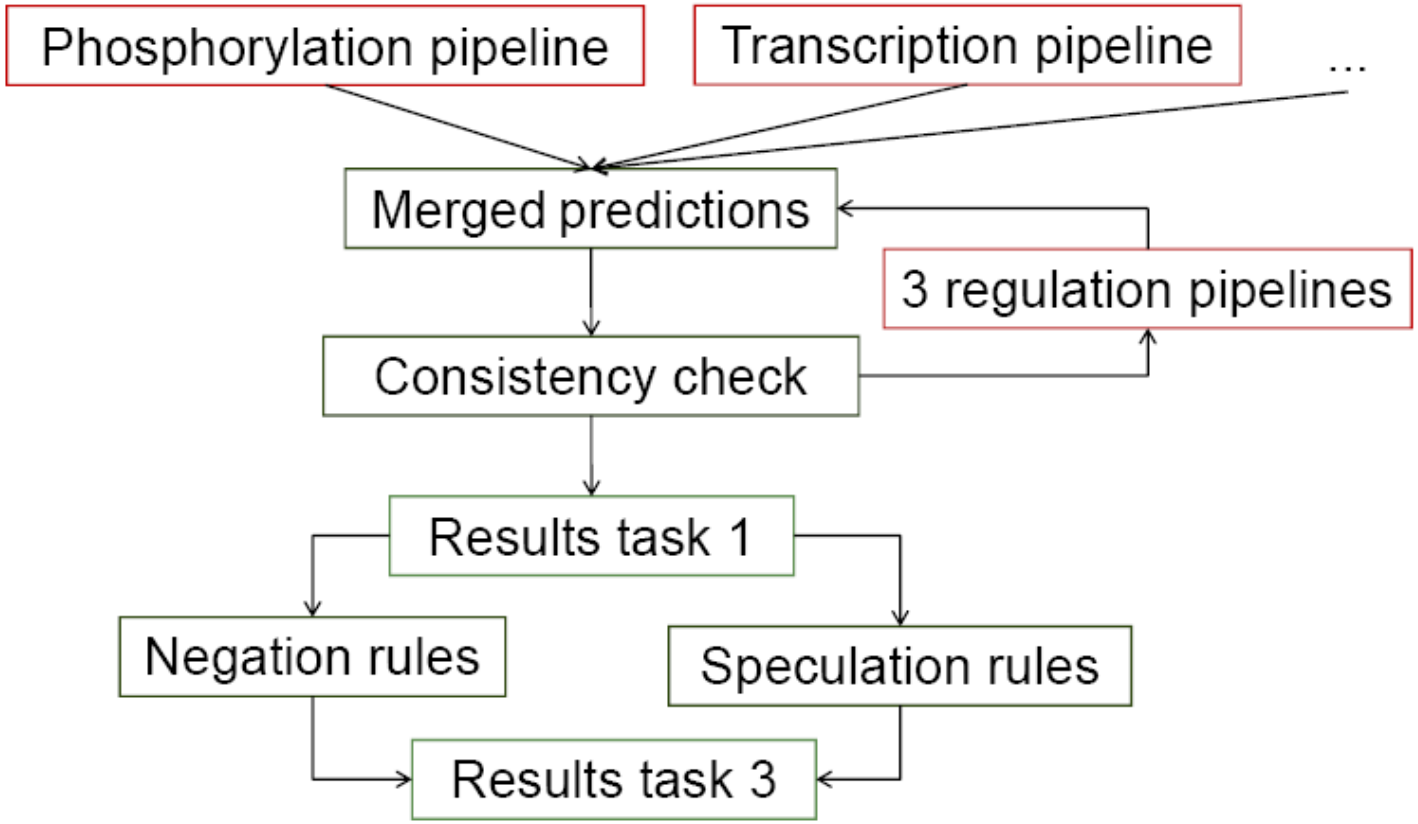
BioNLP'09 Shared Task: event extraction

- Task 1: Core event extraction (mandatory)
  - 6 different event types
    - gene expression, localization, transcription, binding, protein catabolism, phosphorylation
  - 3 regulation events : can take both proteins and other events as arguments
    - Positive regulation, Negative regulation, Regulation
  - Example: phosphorylation of TRAF2 -> (Type:Phosphorylation, Theme:TRAF2)
- Task 2: Event enrichment (optional)
  - Example: localization of beta-catenin into nucleus -> (Type:Localization, Theme:beta-catenin, ToLoc:nucleus)
- Task 3: Negation and speculation recognition (optional)
  - Example: TRADD did not interact with TES2 -> (Negation (Type:Binding, Theme:TRADD, Theme:TES2))

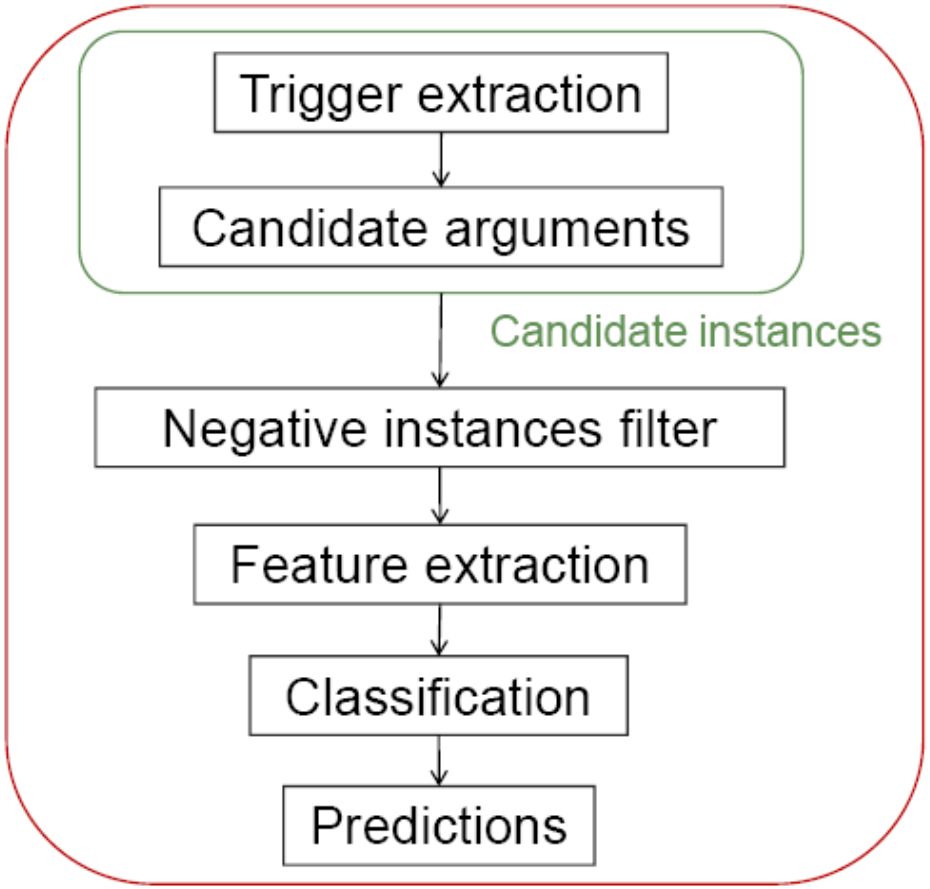
Data sets

Type	Training		Devel.		Test	
ABSTRACT	800		150		260	
	T1	T2	T1	T2	T1	T2
Gene_exp.	1738		356		722	
Transcription	576		81		137	
Prot._catabolism	110		21		14	
Phosphorylation	165	169	47	47	135	139
Localization	263	265	53	53	174	174
Binding	880	887	248	249	347	349
Regulation	960	961	174	178	291	292
Posi._regulation	2843	2847	632	633	983	987
Nega._regulation	1062	1062	197	197	379	379

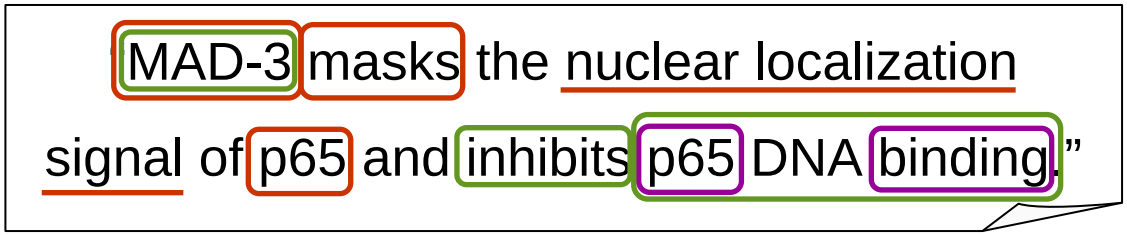
Framework overview



Event pipeline



Example



3 proteins

- T1 : Protein : “MAD-3”
- T2 : Protein : “p65” (first occurrence)
- T3 : Protein : “p65” (second occurrence)

3 triggers

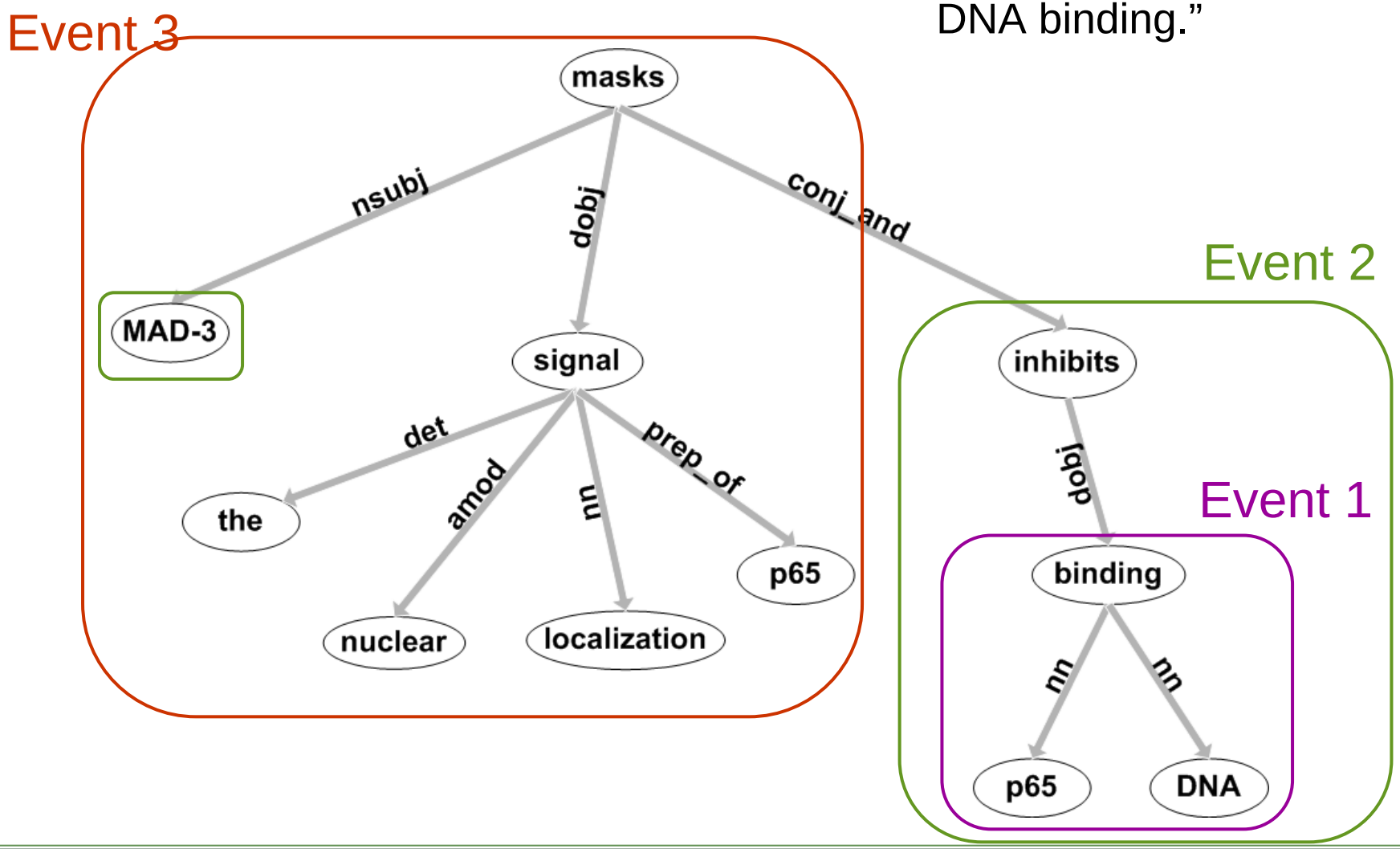
- T4 : Negative regulation : “masks” Event 3
- T5 : Negative regulation : “inhibits” Event 2
- T6 : Binding : “binding” Event 1

1 extra argument

- T7 : Entity : “nuclear localization signal”

Dependency graph

"MAD-3 masks the nuclear localization signal of p65 and inhibits p65 DNA binding."



Trigger dictionaries

1. Dictionary of words which can trigger an event
   - E.g. “secretion”, “phosphorylated”, “overexpression”
   - Stemmed words using Porter stemming algorithm
   - A seperate dictionary for each type of event
   - Compiled automatically from training data
   - Manually filtered to remove general words such as “are”, “via” or “through”
   - Binding : distinction between
     - “Single” (e.g. “homodimer”, “binding site”)
     - “Multiple” (e.g. “heterodimer”, “complex”)

Feature generation

- Stanford dependency parsing: smallest subgraph
- Vertex walks extracted from the dependency subgraph:
  - Vertex – edge – vertex
  - Lexical variant: trigger/protein blinded, e.g. "*trigger nsubj protx*" which expresses that the given protein is the subject of a trigger)
  - Syntactic variant : e.g. "*nn nsubj nn*"
- Bag-of-words: nodes of dependency graphs
  - Exclude uninformative words such as prepositions
  - Stemmed trigrams
- Lexical and syntactic information of the triggers
- Length of the sub-sentence & size of the subgraph
- Regulation: whether arguments are proteins or events

Feature generation

Event type	# Features	# neg. inst.	# pos. inst.	% pos. inst.
Localization	18 121	3415	249	7
Single binding	21 332	3548	522	13
Multiple binding	11 228	2180	185	8
Gene expression	31 332	5356	1542	22
Transcription	30 306	6930	489	7
Protein catabolism	1 883	175	96	35
Phosphorylation	2 185	163	153	48
Unspecified regulation (Unary)	27 915	6076	408	6
Positive regulation (Unary)	48 944	13834	1367	9
Negative regulation (Unary)	16 673	3233	489	13
Unspecified regulation (Binary)	4 239	778	81	9
Positive regulation (Binary)	19 468	5405	249	4
Negative regulation (Binary)	4 166	819	29	3

Classification

- High-dimensional and highly unbalanced datasets
- By processing all events in parallel, binary classifiers can be used (Event <-> No Event)
- Support vector machine (SVM)
- LibSVM implementation as provided by WEKA
- Kernel type : radial basis function (default)
- Internal 5-fold CV loop to tune parameters

Performance comparison

Team	Protein Events	Binding	Regulation	All
UTurku	70.21	44.41	40.11	51.95
JULIELab	68.38	41.20	34.60	46.66
ConcordU	61.76	27.20	35.43	44.62
UT+DBCLS	63.12	31.19	32.30	44.35
VIBGhent	**64.59**	**38.32**	**22.41**	**40.54**
UTokyo	55.96	41.10	20.09	36.88
UNSW	55.39	28.92	20.90	34.92
UZurich	53.66	33.75	19.89	34.78
ASU+HU+BU	56.82	27.49	09.01	32.09
Cam	51.79	18.14	15.79	30.80

- Task 1: 3rd position for protein events, 4th position for binding events, 5th position for regulation events (overall 5th position out of 24 teams)
  Last experiments: overall 44.04 F-measure
- Task 3: 2nd best performance out of 6 groups

Intermezzo: the power of ensembles

Ensemble	Equal	Averaged	Event Type
Top 3	53.19	53.19	54.08
Top 4	54.34	54.34	55.21
Top 5	54.77	55.03	55.10
Top 6	**55.13**	**55.77**	**55.96**
Top 7	54.33	55.45	55.73
Top 10	52.79	54.63	55.18

- Best single system obtained 51.95 overall F-measure
- 4% overall improvement by combining the best 6 systems

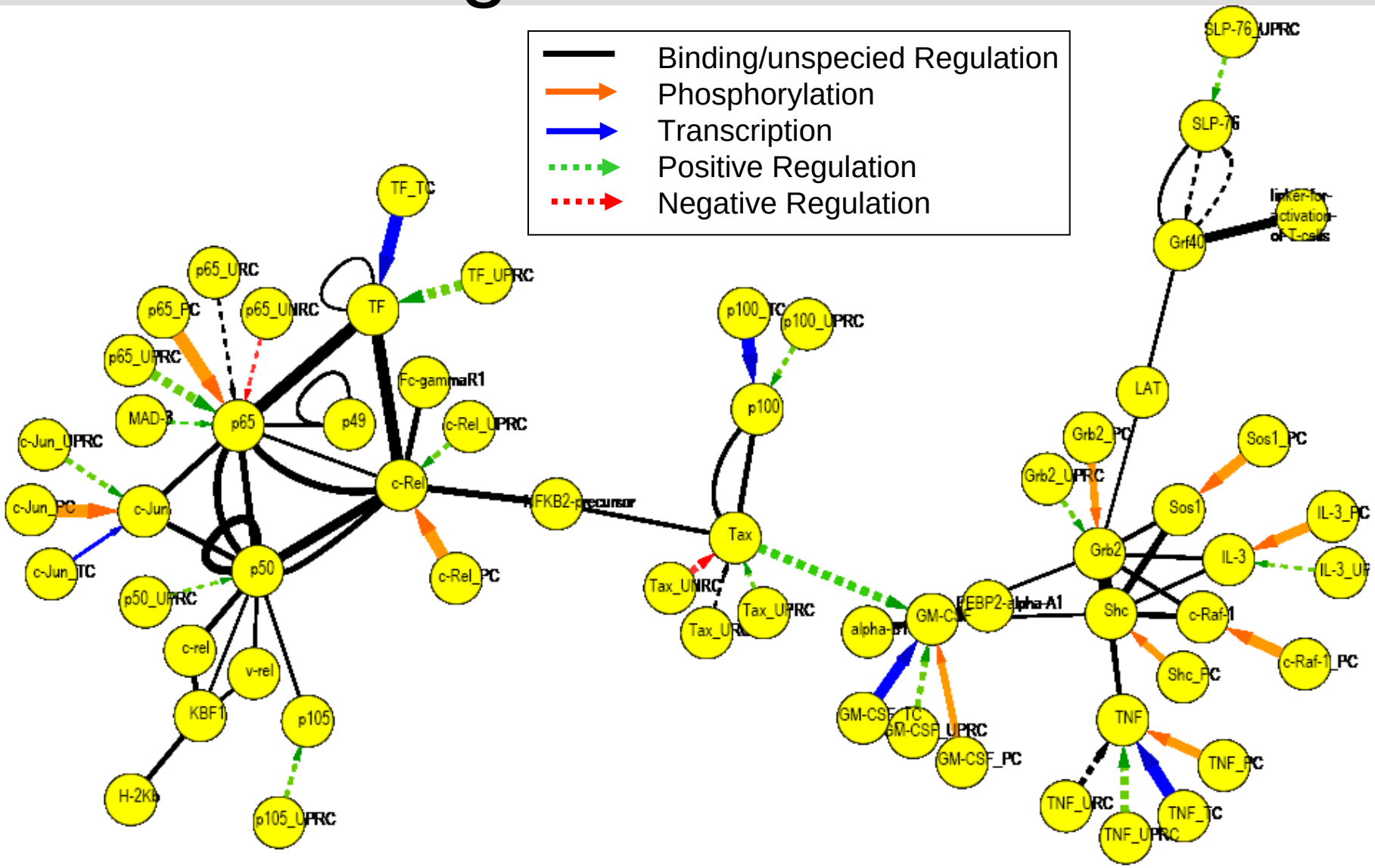
Integrated networks

- Start from a set of interaction events $\{I_1, I_2, \ldots, I_Q\}$
- With each event type I_i, a heterogenous graph J_i can be associated:
  - there might be multiple edges between two nodes in a graph
  - some of the edges may be directed (e.g. *A* regulates *B*)
  - others may be undirected (e.g binding of *C* and *D*)
  - edges are weighted by the confidence of the associated prediction (scaled SVM output)
- Each graph J_i can be represented by its associated matrix $J_i(m,n)$, where each entry in the matrix is a ***set*** of weighted connections between node m and node n.

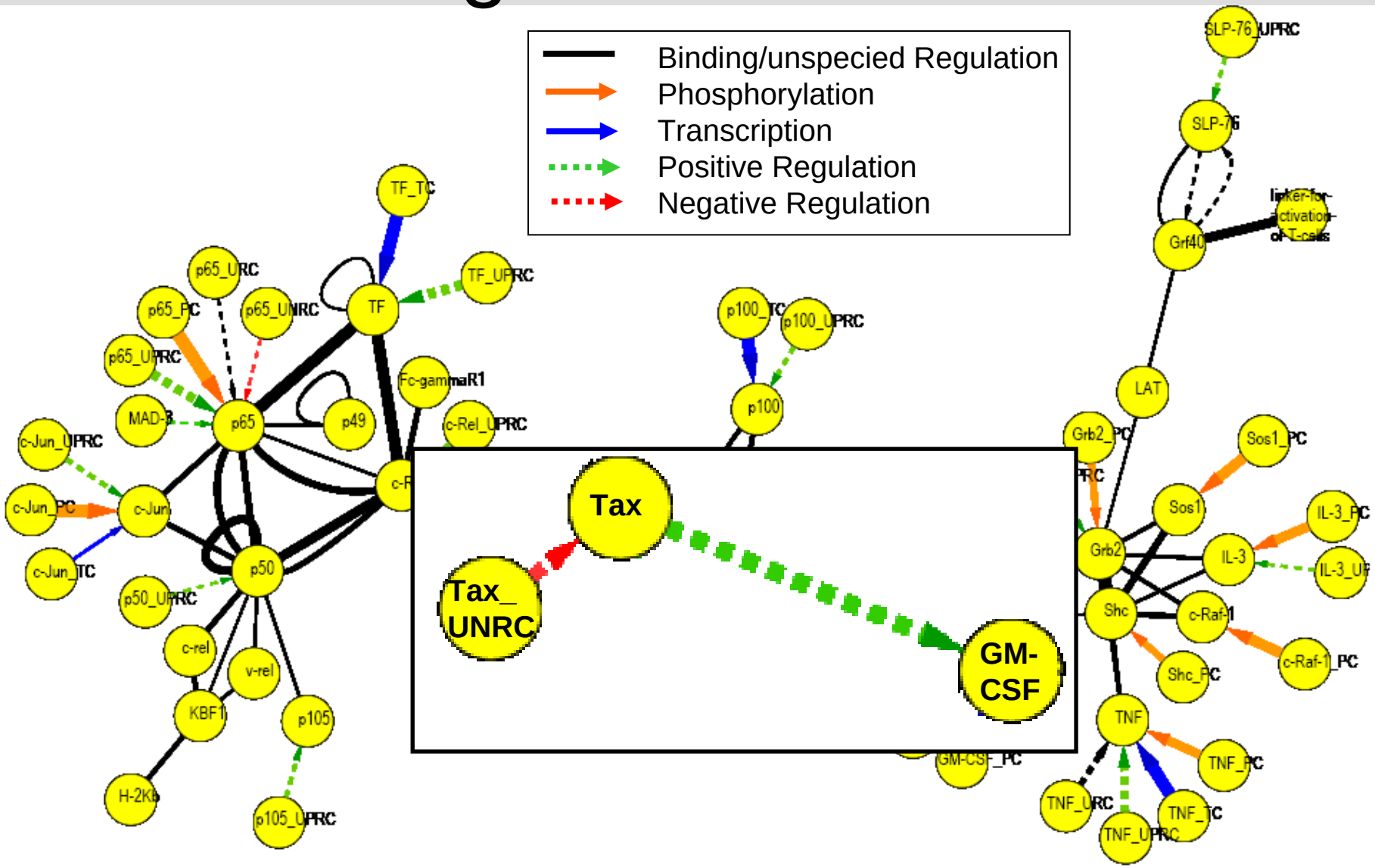
Integrated networks

- Integrate all matrices J_l into a 3D tensor W_{mno}
- Dim(W) = $P \times P \times Q$,
  - P is the the cardinality of the union of all nodes in all J_l
  - Q is the number of events to integrate
- The tensor entry W_{mno} represents a connection (set of predictions) from node m to node n for event type o.

Integrated networks



Integrated networks



Future work

- Still a lot of room to improve prediction performance
- Application of feature selection techniques
- Library for biomedical text mining support
- Increasing robustness of integrated networks:
  - Combining with module networks
  - Combining with existing databases
- Apply inference algorithms to derive potentially new biological knowledge