Large scale text mining challenges for systems biology

Yvan Saeys

Bioinformatics and Evolutionary Genomics (BEG)
Department of Plant Systems Biology, VIB/UGent

yvan.saeys@ugent.be

BIOMAGNET
Bioinformatics and Modelling: from Genomes to Networks

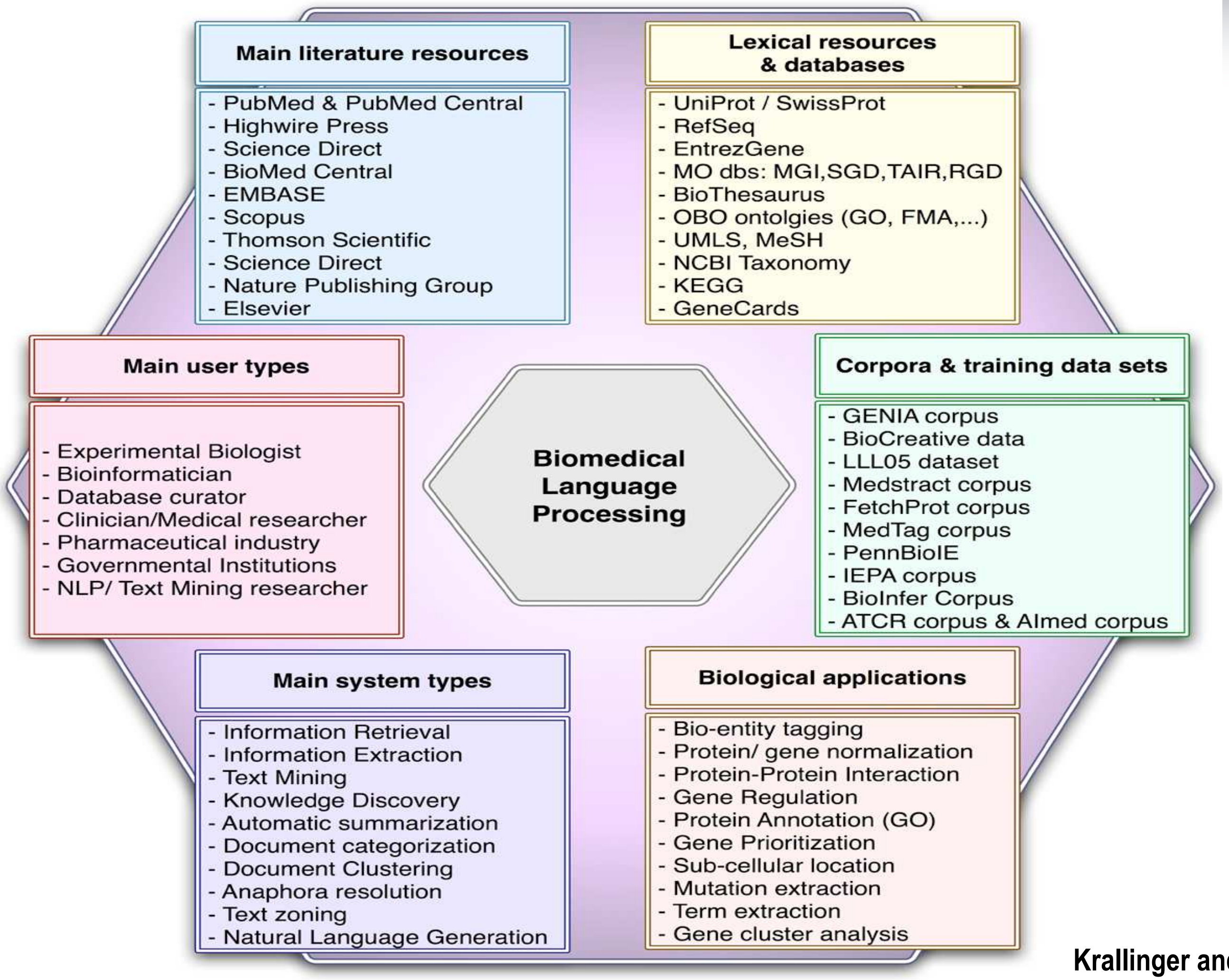
belspo

Motivations for text mining

- Keeping up with the pace at which articles are produced is hard
  - PubMed currently contains about 20 million citations
  - Citations in English with abstract: 11 million
  - PubMed Central: 2 million full-text articles
- Literature contains a wealth of information, a lot of which is not available in public databases
- Many motivations for automated text mining, a.o.:
  - Intelligent reading tools to assist researchers
  - *Large scale information extraction, looking for new relations and interesting patterns*

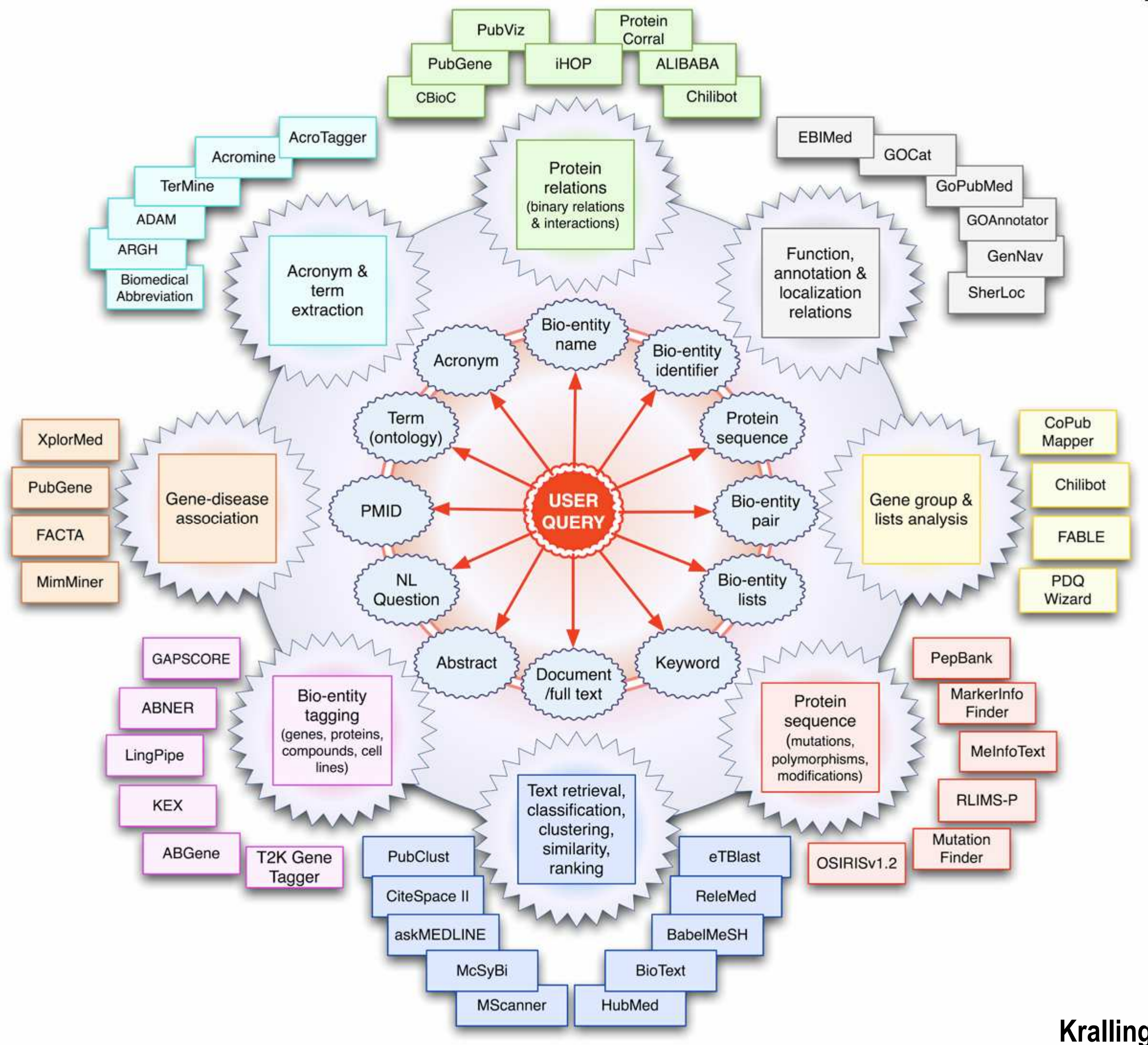
Systems Biology

Biological literature processing



Krallinger and Valencia (2009)

Literature mining tools

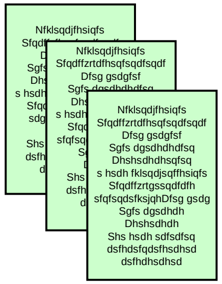


Krallinger and Valencia (2009)

Text mining pipeline

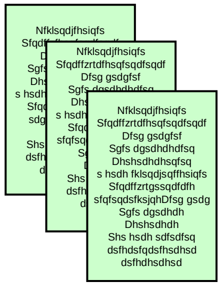
Step 1:

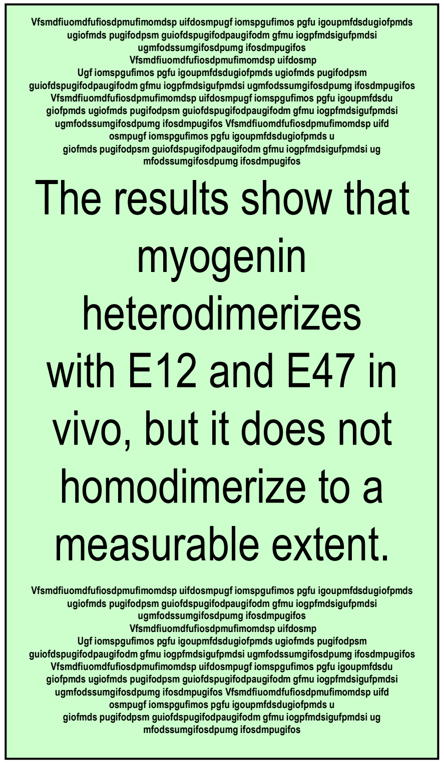
Information retrieval (IR)



Step 2:

Lexical analysis: tokenization, morphological analysis





Text mining pipeline

The/DT results/NNS show/VBP that/IN myogenin/NNP heterodimerizes/VBZ with/IN E12/NNP and/CC E47/NNP in/FW vivo/FW ,/,

but/CC it/PRP does/VBZ not/RB homodimerize/VB to/TO a/DT measurable/JJ extent/NN ./.

Step 3:

Syntactic analysis:

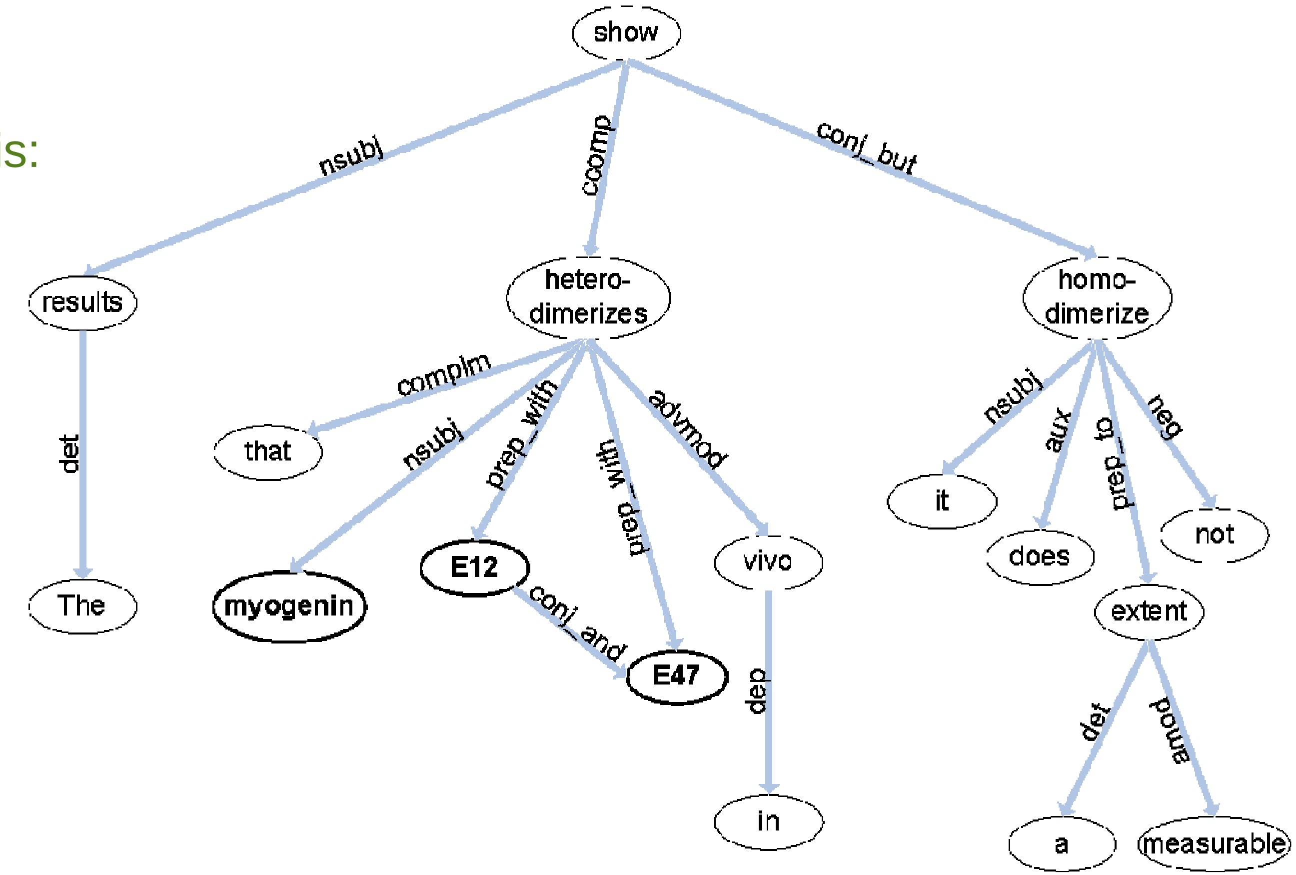
1. Part-of-speech-tagging
2. Phrase chunking

Text mining pipeline

Step 3:

Syntactic analysis:

1. Part-of-speech-tagging
2. Phrase chunking
3. Parsing

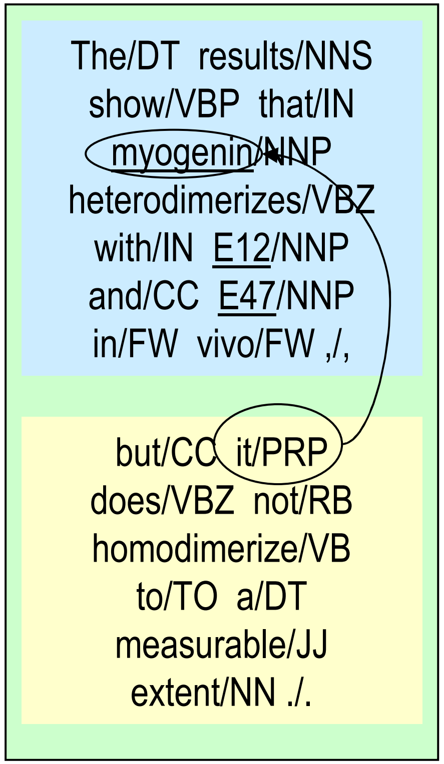


Text mining pipeline

Step 4:

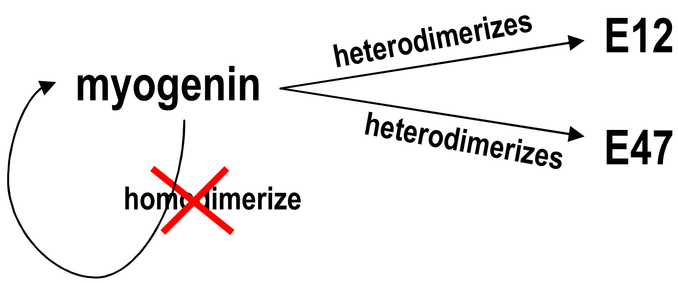
Semantic analysis:

1. Named entity recognition (NER)
2. Word sense disambiguation
3. Coreference resolution



Step 5:

Relation extraction



Limitations of text mining

- Deep understanding is an unsolved AI problem
  - Beyond the limit of what can be achieved by a computer
  - Possible only for very limited domains after extensive knowledge modelling

```
  Welcome to
                    EEEEEE  LL      IIII  ZZZZZZZ   AAAAA
                    EE      LL       II        ZZ  AA   AA
                    EEEEE   LL       II      ZZZ   AAAAAAA
                    EE      LL       II    ZZ      AA   AA
                    EEEEEE  LLLLLL  IIII  ZZZZZZZ  AA   AA

  Eliza is a mock Rogerian psychotherapist.
  The original program was described by Joseph Weizenbaum in 1966.
  This implementation by Norbert Landsteiner 2005.


ELIZA: How do you do.  Please tell me your problem.
YOU:   Do you really like me?
ELIZA: What makes you think I really like you ?
YOU:   Does that mean you don't really like me?
ELIZA: Really, I don't really like you ?
YOU:   I'm confused. Do you like me or not?
ELIZA: Do you know anyone else who is confused ?
YOU:   Don't change the subject! I need to know if you're my friend!
ELIZA: What would you do if I are your friend ?
YOU:   I would be happy - we could go on picnics!
ELIZA: Can you elaborate on that ?
YOU:
```

Limitations of text mining

- Context is very important, e.g. in WSD
  - *complex* relations, a protein *complex*, X can *complex* metals
  - A binds to DNA with an alternating G-C sequence
  - A binds to DNA with low affinity

Challenges for bio-text mining

- Tokenization
  - Problems with abbreviations, gene variants, …

  “There was no evidence that the 10.4K protein is glycosylated , and the 10.4K protein was not required for glycosylation of 14.5K .”
- Morphological analysis
  - Orthographic variants

    e.g. `amyloid beta-protein' versus `amyloid ß-protein’
  - Morphological variants

    e.g. `transcription intermediary factor-2' versus `transcriptional intermediate factor 2’
  - Lexical variants

    e.g. `hepatic leukaemia factor’ versus `liver leukemia factor’.

Challenges for bio-text mining

- Named entity recognition
  - Multiple names (many-to-one) e.g. CDC28, Cdc28p, cyclin-dependent kinase 1
  - Ambiguity (1-to-many)
    - Biomedical terms: RA
      - retinoic acid
      - retrograde amnesia
      - refractory anemia
      - rheumatoid arthritis
  - Common English words: e.g. hairy, hair loss, CAT, WHO

    "The gene cannonball is referred to in FlyBase by the symbol can (CG6577 , FBgn0011569) ."
- Current NER performance:
  - Around 80% for combined precision & recall (F-measure)

History of bio-text mining methods

- Before 2005
  - Using hand-crafted rules to find disease-related genes, protein-protein interactions
  - Co-occurrence based approaches
- From 2005
  - Machine learning approaches to improve protein-protein interaction recognition
- From 2009
  - BioNLP shared task
  - More specific types of interactions: “events”

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))

Example

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

3 proteins

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

3 triggers

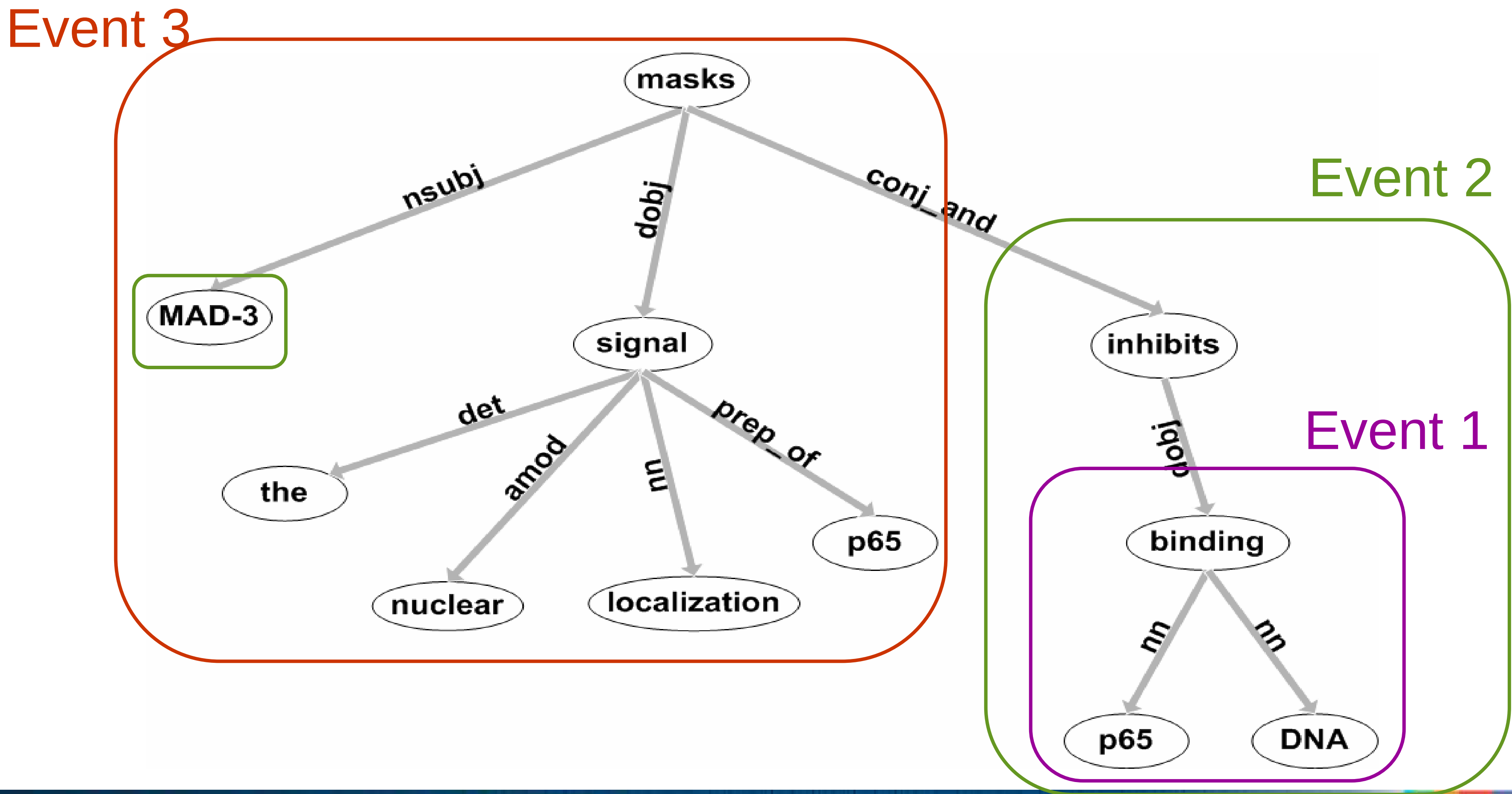
- T4 : Negative regulation : "masks" Event 3
- T5 : Negative regulation : "inhibits" Event 2
- T6 : Binding : "binding" Event 1

1 extra argument

- T7 : Entity : "nuclear localization signal"

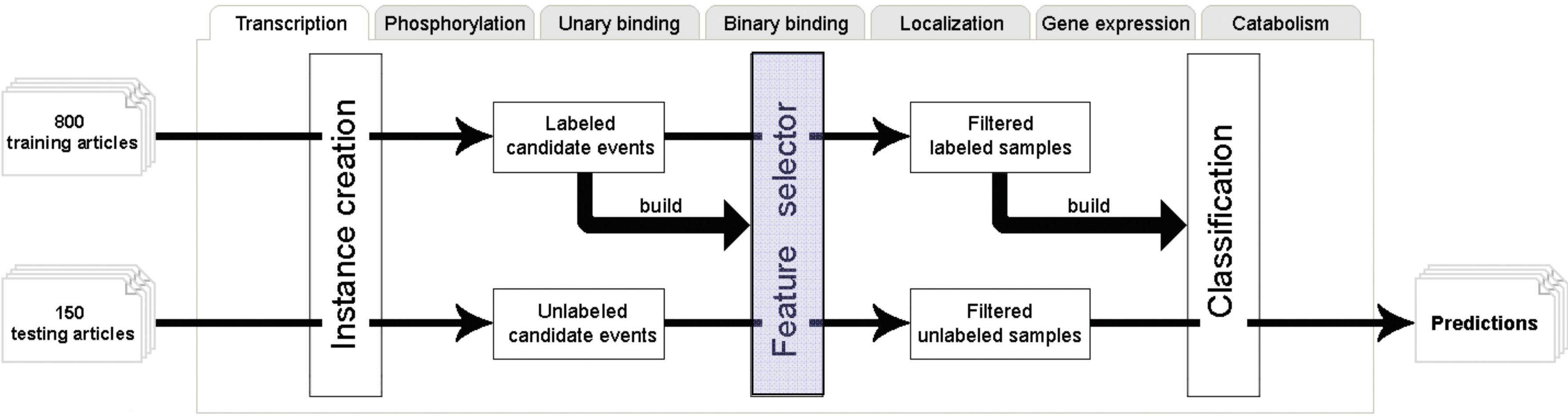
Dependency graph

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



Methodology

1. Training
   - Create candidate instances
   - Build feature selector
   - Build classifier (binary support vector machine)
2. Testing
   - Create candidate instances
   - Apply feature selector & classifier to generate predictions



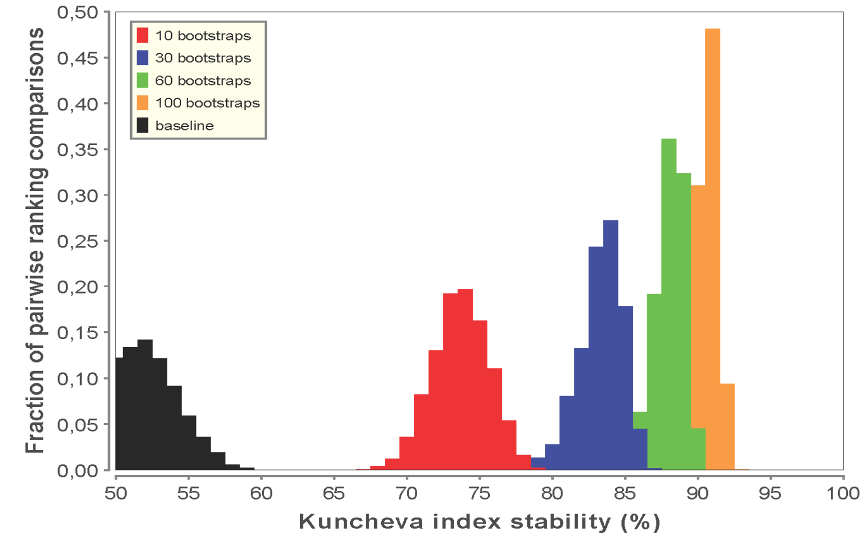
Baseline results (without FS)

- 6 different event types
  - Gene expression, Localization, Protein catabolism, Transcription, Binding, Phosphorylation
- Official results, on final test set (spring 2009)
  - 24 participating teams

Team	**Precision**	**Recall**	**F-score**
1. Finland	71.33	58.73	64.42
2. Germany	63.97	57.49	60.56
3. Ghent	67.24	50.75	57.85

[Van Landeghem, S., Saeys, Y., De Baets, B., Van de Peer, Y. (2009) Analyzing text in search of bio-molecular events: a high-precision machine learning framework. Proceedings of Natural Language Processing in Biomedicine (BioNLP) NAACL 2009 Workshop 128-136.]

Advanced results (validation set)



Feature space	Minimum F	Maximum F	Average F
100% (baseline)			65.02
75%	64.85	65.33	65.26
50%	65.60	66.43	65.88
30%	64.94	66.60	65.86
25%	65.51	66.82	**66.14**
20%	65.08	66.56	65.85
10%	61.75	64.90	63.59

- Ensemble feature selection has a beneficial effect on the stability of selected features
- System performance is slightly improved, while eliminating up to 75% of the features
  - More cost-effective models

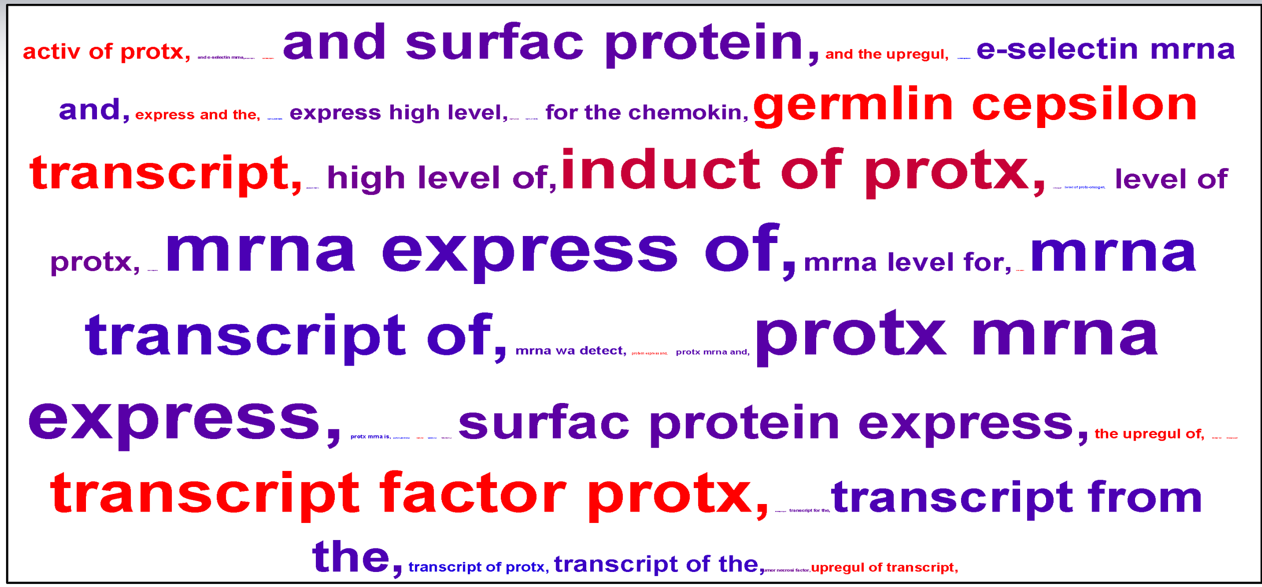
[Van Landeghem, S., Abeel, T., Saeys, Y., Van de Peer, Y. (2010) Discriminative and informative features for biomolecular text mining with ensemble feature selection. Bioinformatics 26, 554-560]

Advanced results (test set)

Team	Precision	Recall	F-score
1. Finland	71.33	58.73	64.42
2. Germany	63.97	57.49	60.56
3. Ghent	67.24	50.75	57.85

→ 60.49

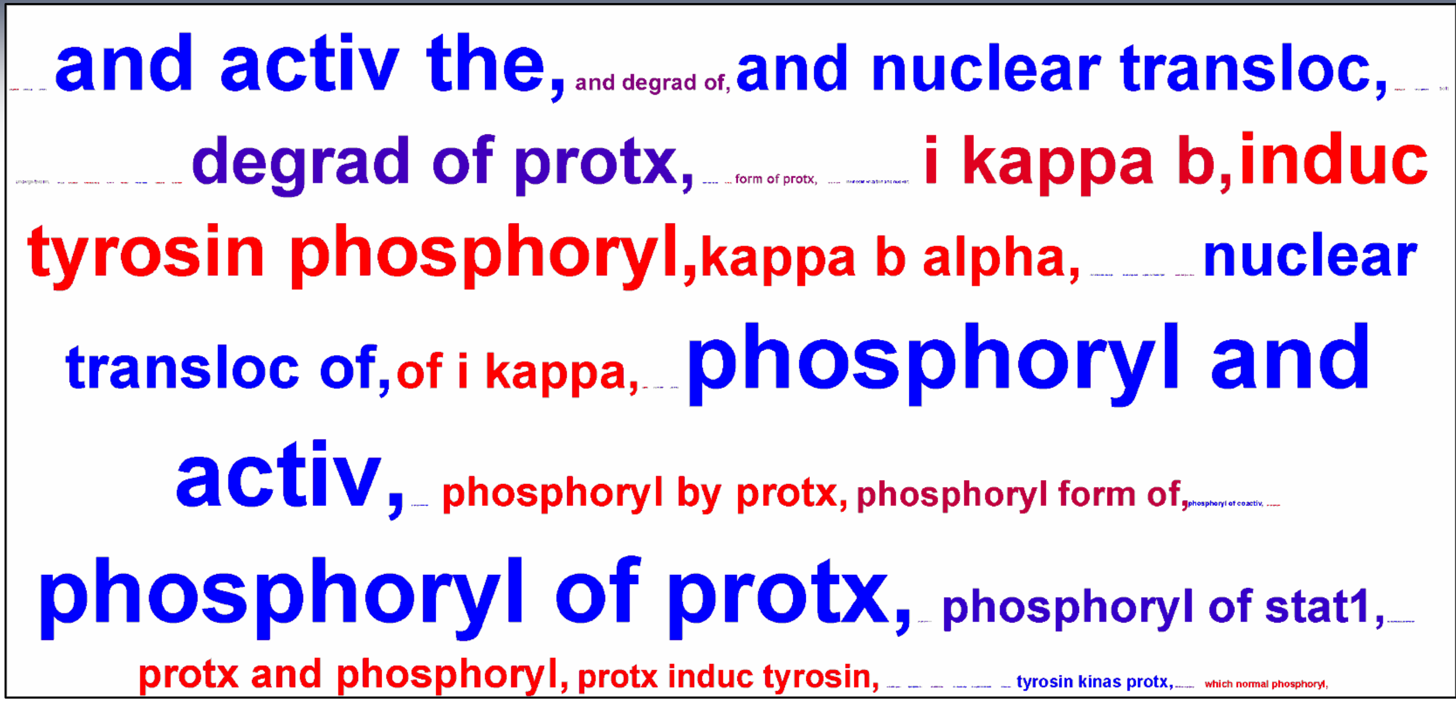
Informative patterns for systems designers



- Positive patterns: often include “mrna”, which is also the most informative bag-of-word feature
- “transcription factor protx” : strong negative
- The machine learning framework thus automatically deduces biological knowledge & general lexical patterns

[Van Landeghem, S., Abeel, T., Saeys, Y., Van de Peer, Y. (2010) Discriminative and informative features for biomolecular text mining with ensemble feature selection. Bioinformatics 26, 554-560]

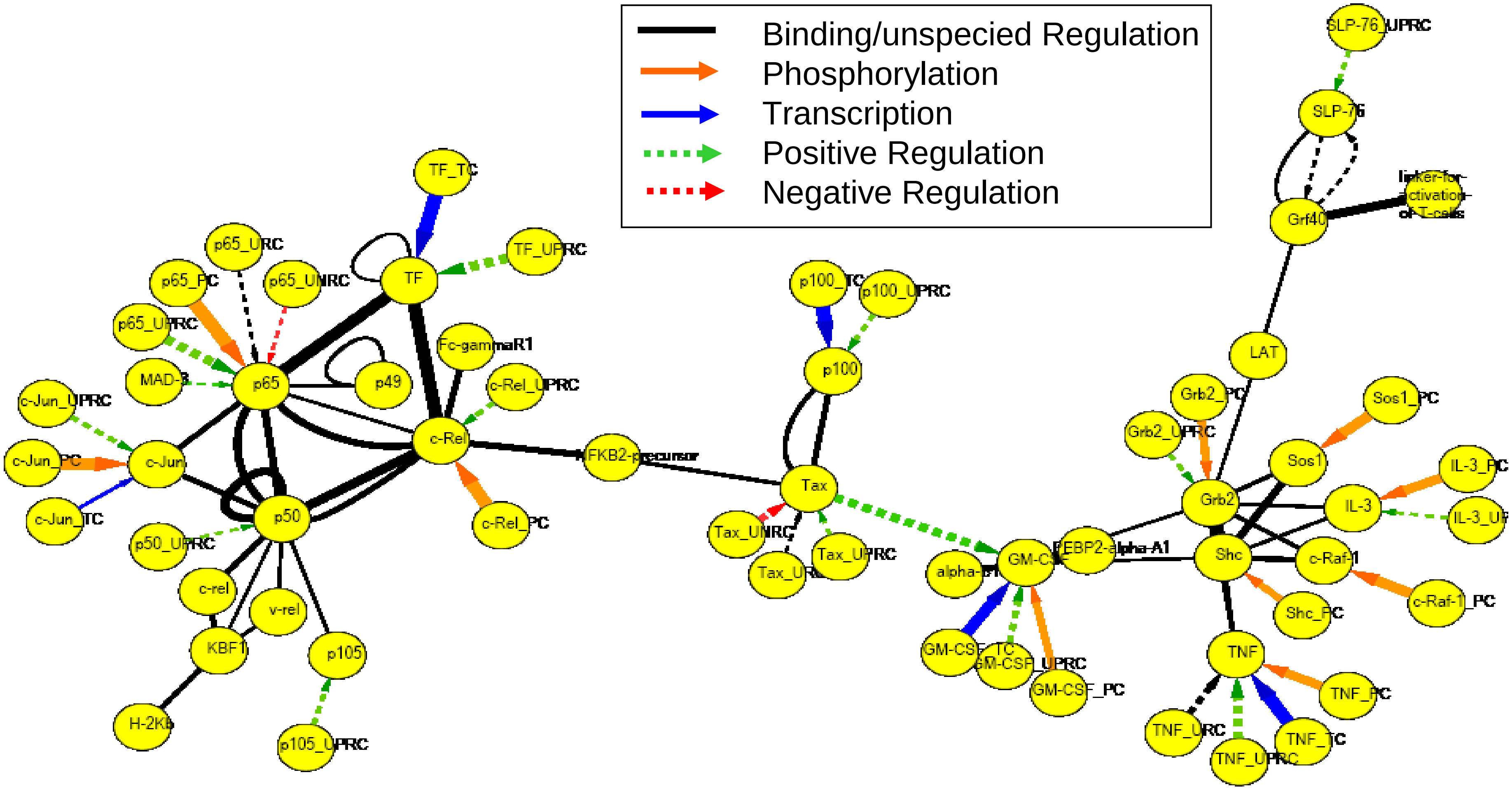
Informative patterns for systems designers



- "I kappa b alpha" : not fully captured by trigrams
  - Consider N-grams with N > 3
- Lexical variants : iKappaBAlpha, I kappa B-alpha
  - Dictionary look-up?

[Van Landeghem, S., Abeel, T., Saeys, Y., Van de Peer, Y. (2010) Discriminative and informative features for biomolecular text mining with ensemble feature selection. Bioinformatics 26, 554-560]

From text mining to integrated networks



[Saeys, Y., Van Landeghem, S., Van de Peer, Y. (2010) Event based text mining for integrated network construction. Journal of Machine Learning Research, Workshop and Conference proceedings 8, 112-121.]

Recent advances and applications to systems biology

- Going from abstracts to full text
- Mining figures, tables, …
- Text mining at PubMed scale
  - Requires high-performance computing environment
  - Currently only done on abstracts
  - Full text will certainly follow soon

PubMed scale text mining

- Going from 13,600 manually annotated events from 1,210 PubMed abstracts in the entire Shared Task data
- … to 11,000,000 abstracts + 17,800 000 PubMed citation titles
- Results:
  - Parsing 200,000,000 sentences
  - 19,200,000 event occurrences (4,500,000 unique ones)
  - 2,100,000 occurrences contain at least two different NEs (1,600,000 unique ones)
- Required time : 346 CPU days

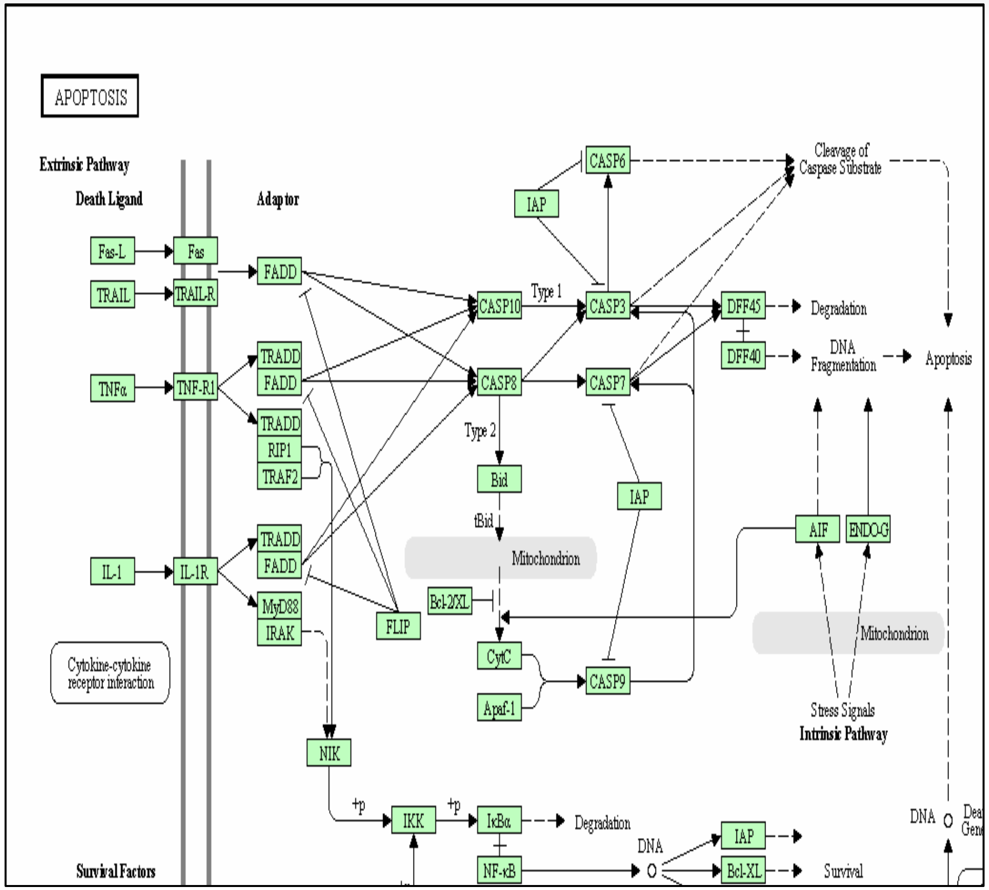
[Björne J, Ginter F, Pyysalo S, Tsujii J, Salakoski T. Complex event extraction at PubMed scale (2010) Bioinformatics 15;26(12):i382-90.]

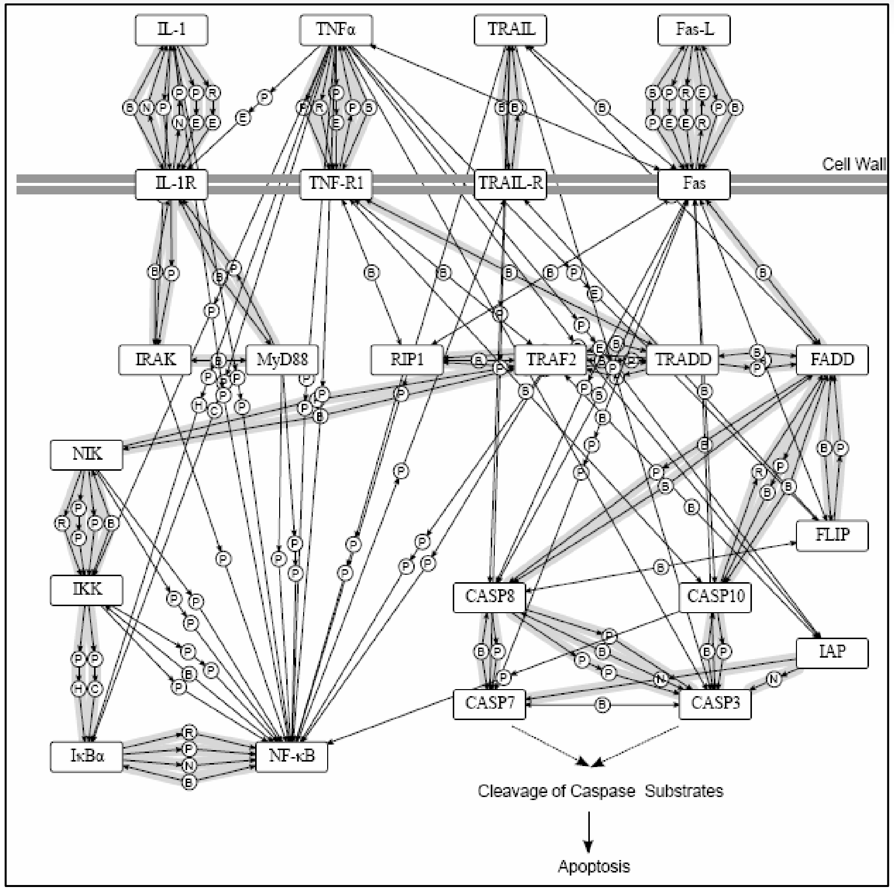
Estimated output quality

- Event extraction: 64% precision
- Site and location prediction: 53% precision
- Negation: 82% correctly predicted as negated
- Speculation: 88% correctly predicted as speculated

[Björne J, Ginter F, Pyysalo S, Tsujii J, Salakoski T. Complex event extraction at PubMed scale (2010) Bioinformatics 15;26(12):i382-90.]

Case study: apoptosis pathway





[Björne J, Ginter F, Pyysalo S, Tsujii J, Salakoski T. Scaling up Biomedical Event Extraction to the Entire PubMed(2010) In Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, pp. 28-36.

Future work

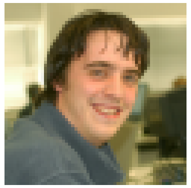
- Event extraction on full text
  - PDF/HTML to text conversion
  - Copyright issues
- Improve annotated corpora
- Design of new text mining corpora
  - Gene – mutant - phenotype associations for Arabidopsis
  - Requires careful annotation design

Acknowledgements

@UGENT:

Sofie Van Landeghem

Thomas Abeel

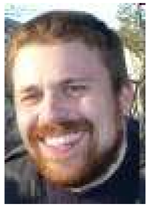


Yvan Saeys

@UTU:

Jari Björne

Filip Ginter



@UTokio:

Tomoko Ohta

Sampo Pyssalo