Extracting Protein-Protein Interactions

Simon Tannert & Elias Zaied

IMS, Universität Stuttgart

Thursday 11th January, 2018

Outline

Background

What are proteins?

Proteins are macromolecules, consisting of one or more long chains of amino acid residues.

They function like tools in organisms:

- catalyze metabolic reactions
- replicate DNA
- respond to cell stimuli
- transport molecules

Myoglobin – binds iron and oxygen in the muscle tissue of vertebrates

Background

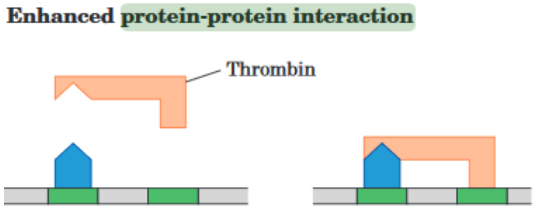
What are protein-protein interactions?

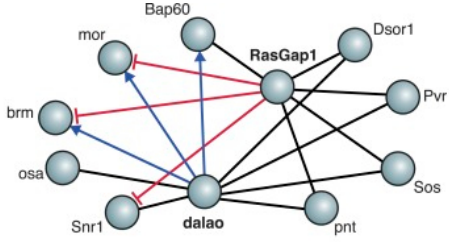
Protein-protein interactions (PPI) are physical contacts between proteins

- Vital to the protein functions mentioned before
- Enable higher protein function
- Help in determining the structure of unknown proteins
  - "Guilt by association": known with unknown protein functions

Background

Protein-protein interaction examples



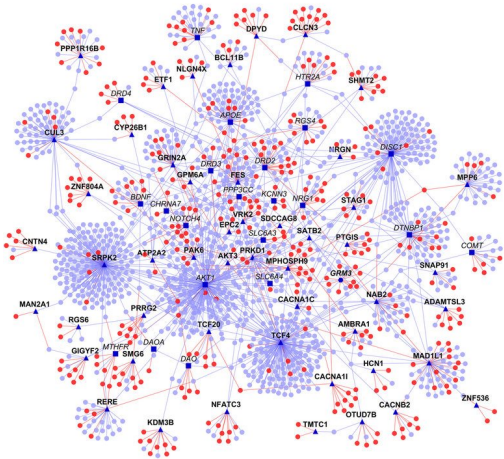


Motivation

Extracting protein-protein interactions from biomedical text

E.g. studying diseases

→ Here: Schizophrenia



Motivation

Extracting protein-protein interactions from biomedical text

Use cases:

- Scientists
- Molecular biologists
  - Running experiments for gene/interaction research
- Demand:
  - Easy access to state-of-the-art methods and publications

Motivation

Challenges of protein-protein interaction extraction

Increase of NLP method application to biomedical text

- Increase in number of corpora and methods for PPI extraction

No general consensus for PPI annotation

- Resources largely incompatible
- Methods are difficult to evaluate

Motivation

Challenges of protein-protein interaction extraction

Example sentence from the BioInfer corpus

The sentence

The inhibitory action of p27, a cyclin-dependent kinase inhibitor (CDKI), arises from its binding with the cyclin E/Cdk2 complex that results in G(1)-S arrest.

Entities and types

- *p27—individual protein*
- *action of p27—function property*
- *cyclin-dependent kinase inhibitor—protein family or group*
- *CDKI—protein family or group*
- *cyclin E—individual protein*
- *Cdk2—individual protein*
- *cyclin E/Cdk2 complex—protein complex*
- *its—individual protein*

Ontology knowledge about entity types

- *individual protein* $\prec$ *protein* $\prec$ *amino acid* $\prec$ *organic* $\prec$ *compound* $\prec$ *substance* $\prec$ *physical entity* $\prec$ *entity*
- *protein family or group* $\prec$ *protein* $\prec$ *amino acid* $\prec$ *organic* $\prec$ *compound* $\prec$ *substance* $\prec$ *physical entity* $\prec$ *entity*
- *protein complex* $\prec$ *individual protein* $\prec$ *protein* $\prec$ *amino acid* $\prec$ *organic* $\prec$ *compound* $\prec$ *substance* $\prec$ *physical entity* $\prec$ *entity*
- *function property* $\prec$ *property* $\prec$ *entity*

Relationships

- *EQUAL(cyclin-dependent kinase inhibitor, CDKI)*
- *MEMBER(cyclin-dependent kinase inhibitor, p27)*
- *ANAPHORA(its, p27)*
- *CONTAIN(cyclin E/Cdk2 complex, cyclin E)*
- *CONTAIN(cyclin E/Cdk2 complex, Cdk2)*
- *CAUSE(BIND(its, cyclin E/Cdk2 complex), action of p27)*

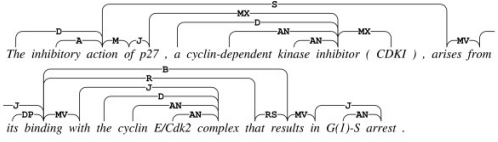
Ontology knowledge about relationships

- *EQUAL* $\prec$ *equality* $\prec$ *is_a* $\prec$ *relationship*
- *MEMBER* $\prec$ *collection:member* $\prec$ *part_of* $\prec$ *relationship*
- *CONTAIN* $\prec$ *object:component* $\prec$ *part_of* $\prec$ *relationship*
- *CAUSE* $\prec$ *causal* $\prec$ *relationship*
- *BIND* $\prec$ *assembly* $\prec$ *physical* $\prec$ *change* $\prec$ *causal* $\prec$ *relationship*

Text binding, i.e. text used to infer the relationships

- *CONTAIN: complex*
- *CAUSE: arises from*
- *BIND: binding with*

Syntactic annotation



Motivation

Problem is big enough in the BioNLP community

- Shared task

Some corpora were developed in the context of the shared task

Protein-protein interaction task

Searching for Proteins that are connected to each other via an interaction in order to extract

PPI is a relation extraction task

- Given pairs of entities from NER
- Identify if there is a relation

Corpora

Most popular PPI corpora:

- *AIMed* – PPI extraction method comparison
- *BioInfer* – corpus for training and testing PPI extraction programs
- *PRD50* – test set for the RelEx system
- *IEPA* – from sentences from PubMed abstracts
- *LLL* – shared dataset for the Learning language in Logic 2005 (LLL05) challenge

Corpora

Requirements for selection

- Freely available
- Specifically identified named entity
- Manually annotated interactions
- Negative exampled of PPI
  - Either: explicitly marked
  - Or: can be validly generated under the CWA

Corpora

Characteristics

Table 1: Corpora

		AIMed	BioInfer	HPRD50	IEPA	LLL
	size	1955	1100	145	486	77
Entity	*scope*	human P/G	P/G/R and related	human P/G	Chemicals	P/G
	coverage	all occurrences	all occurrences	NER system	list of 16 names	list of 116 names
	types	no	111 types (ontology)	no	no	P/G
PPI	*types*	no	no	68 types (ontology)	no	3 types
	binding	no	yes	no	yes	no
	directed	no	yes	no	yes	yes
	complex	no	yes	no	no	no
	negative	no	yes	no	no	no
	certainty	no	no	yes	no	no

Corpora

- Corpora are heterogeneous
- Some fields/relations are defined in some corpora but not in others

Unified format follows the *standoff annotation* principle:

- Original sentence text preserved
- Entities identified through character offsets

Corpora in the unified format

- Stored in XML files
- Very simple structure

Transformation is a very hard task

- Highly individual native format
- Complex transformation programs
- Sometimes manual intervention required to resolve ambiguities

BioInfer conversion was most challenging

- Complex interactions and broader scope

Corpora

Conversion

BioInfer transformed:

```
<document id="BioInfer.d363" origId="718">
  <sentence id="BioInfer.d363.s363" origId="718" text="In adherens junctions, the cytoplasmic domain of cadherins bind to
  beta-catenin, which in turn binds to the actin-associated protein alpha-catenin.">
    <entity id="BioInfer.d363.s363.e0" origId="e.718.6" charOffset="49-57" type="Protein_family_or_group" text="cadherins"/>
    <entity id="BioInfer.d363.s363.e1" origId="e.718.7" charOffset="67-78" type="Individual_protein" text="beta-catenin"/>
    <entity id="BioInfer.d363.s363.e2" origId="e.718.2" charOffset="108-112" type="Individual_protein" text="actin"/>
    <entity id="BioInfer.d363.s363.e3" origId="e.718.8" charOffset="133-145" type="Individual_protein" text="alpha-catenin"/>
    <interaction id="BioInfer.d363.s363.i0" e1="BioInfer.d363.s363.e0" e2="BioInfer.d363.s363.e1" type="BIND" directed=
    "Maybe"/>
    <interaction id="BioInfer.d363.s363.i1" e1="BioInfer.d363.s363.e1" e2="BioInfer.d363.s363.e3" type="BIND" directed=
    "Maybe"/>
  </sentence>
</document>
```

Two methods

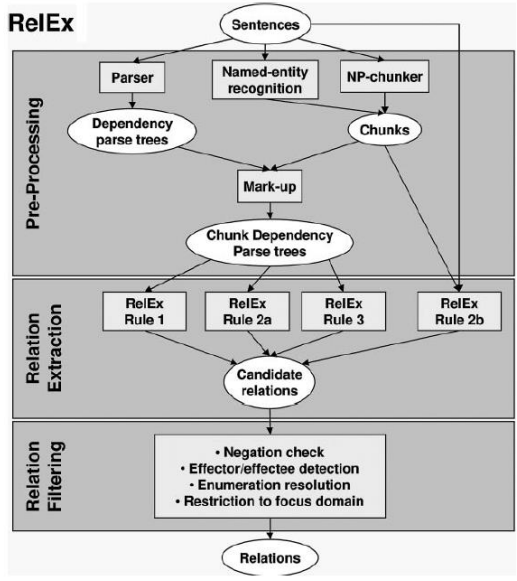
- Co-occurrence
- RelEx (full parsing-based PPI extraction method)
  - Both methods are rule-based

Corpora

PPI extraction - RelEx

Based on NLP preprocessing

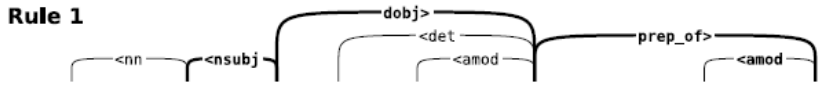
- Producing dependency parse trees
- Applying small number of simple rules to them



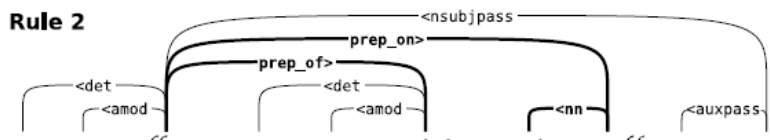
Corpora

PPI extraction - RelEx Rules

Rule 1



Rule 2



Rule 3

Corpora

PPI extraction - Performance

Table 2: PPI extraction performance

Corpus	Co-occurrence			RelEx		
	P	R	F	P	R	F
AIMed	0.17	0.95	0.29	0.40	0.50	0.44
BioInfer	0.13	0.99	0.23	0.39	0.45	0.41
HPRD50	0.38	1.0	0.55	0.76	0.64	0.69
IEPA	0.41	1.0	0.58	0.74	0.61	0.67
LLL	0.50	1.0	0.66	0.82	0.72	0.77

(P)recision, (R)ecall, and (F)-score for the co-occurrence and RelEx methods on the various corpora.

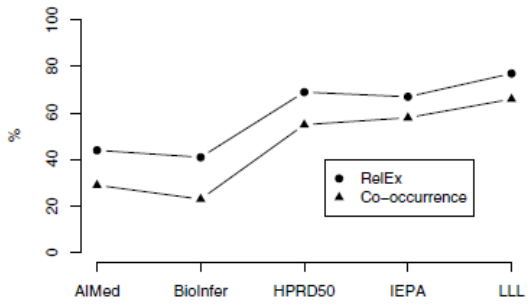


Figure 1
RelEx and co-occurrence F-scores for the five corpora

Corpora

PPI extraction - Filtering

Table 4: PPI extraction performance on filtered corpora

Corpus	Co-occurrence				RelEx			
	P	ΔP	F	ΔF	P	ΔP	F	ΔF
AIMed	0.53	0.36	0.68	0.39	0.85	0.45	0.63	0.19
BioInfer	0.53	0.40	0.70	0.47	0.78	0.39	0.57	0.16
HPRD50	0.64	0.26	0.78	0.23	0.93	0.17	0.76	0.07
IEPA	0.88	0.47	0.94	0.36	1.00	0.26	0.75	0.08
LLL	0.50	0.00	0.66	0.00	0.82	0.00	0.77	0.00

Precision and F-score for the co-occurrence and RelEx methods on the corpora with only entities that participate in an interaction preserved. Recall is not shown as it is unaffected by this modification. The Δ columns show absolute difference to results without filtering.

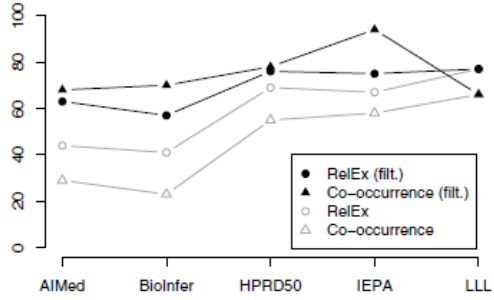


Figure 2
RelEx and co-occurrence F-scores for the filtered version of the corpora. Unfiltered results given in Figure 1 shown in gray for reference.

- The choice of corpus has a larger effect on the result than the choice between a naïve PPI extraction method and an advanced one.
- Annotation of proteins for which there is no annotated interaction
  - determines almost half of the performance difference between corpora. (Filtering)
- Notable differences in the distribution of interaction types
- Identifying often-unstated points of agreement and disagreement in annotation regarding negation, speculative statements, the explicitness of annotated statements and the directness of the corpus interactions

Corpora

Takeaway

- Understanding the challenges of PPI on the basis of this corpus study presenting simple baselines (the two rule-based systems)
- Criticism:
  - No experiment section in paper
  - They didn't try to develop or train a good ML system on unified corpora

Outline

Methods

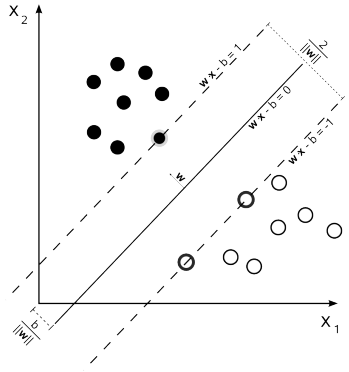
Two methods

- SVM classifier with kernels
  Tikk, D., Thomas, P., Palaga, P., Hakenberg, J., & Leser, U. (2010). A comprehensive benchmark of kernel methods to extract protein–protein interactions from literature. PLoS computational biology, 6(7), e1000837.
- Probabilistic Graphical Model
  Mallory, E. K., Zhang, C., Ré, C., & Altman, R. B. (2015). Large-scale extraction of gene interactions from full-text literature using DeepDive. Bioinformatics, 32(1), 106-113.

SVM classifier with kernels

A support vector machine (SVM) is a linear classifier that finds a hyperplane

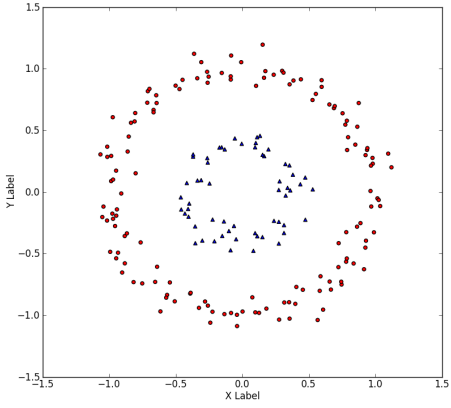
- separating the data in the space described by the feature dimensions
- that represents the largest separation between the two classes (max-margin)
- that is specified by the support vectors

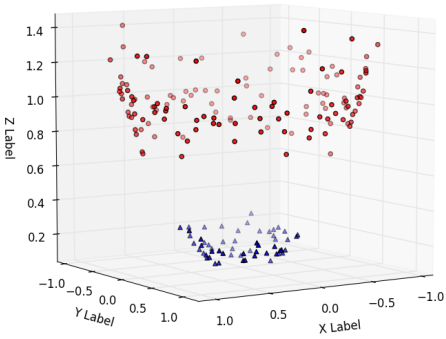


SVM classifier with kernels

Modifications needed for data that is not linearly separable

- Transform features to a space in which they are linearly separable



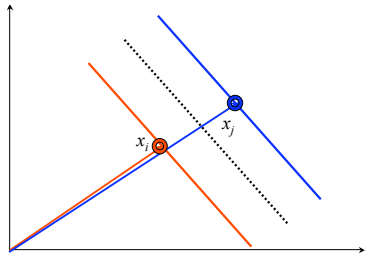


$$\phi(\vec{a}) = \phi((x, y)) = (x, y, x^2 + y^2)$$

SVM classifier with kernels

Maximizing margin can be formulated as Lagrangian dual problem

- Uses dot product of training vector pairs to measure their similiarity
- Similar vectors belonging to different classes might be support vectors



Kernels:

- Calculate distance between training pairs during training/prediction
  - We can replace the dot product with a kernel function: $k(\vec{a}, \vec{b}) = (\phi(\vec{a}), \phi(\vec{b}))$
- Implicitly transform features into a different feature space
- Can calculate similarity between more than only numbers (words, parse trees, ...)

SVM classifier with kernels

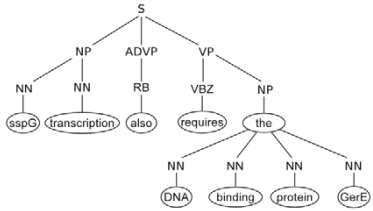
Shallow linguistic kernel

Sum of *global context kernel* and *local context-kernel*

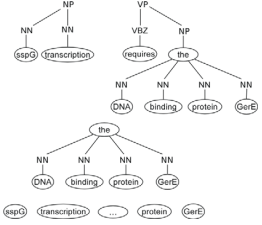
- *global context kernel*
  - Sum of three kernels capturing token frequencies
    - *fore-between*, e.g. binding of $[P_1]$ to $[P_2]$
    - *between*, e.g. $[P_1]$ inhibitor of $[P_2]$
    - *between-after*, e.g. $[P_1]$ and $[P_2]$ interact
- *local context kernel*
  - Dot product of features from words left and right of protein mention pair
    - capitalization, punctuation, numerals, ...
    - POS-tag, lemma

SVM classifier with kernels

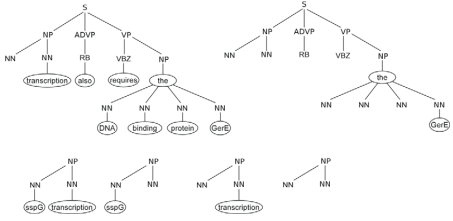
Constituent tree-based kernels



Subtree kernel

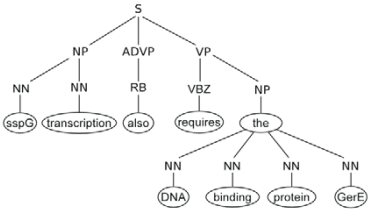


Subset tree kernel

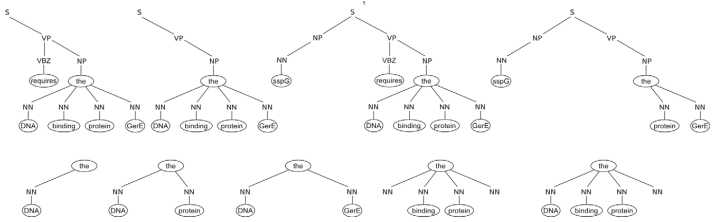


SVM classifier with kernels

Constituent tree-based kernels



Partial tree kernel

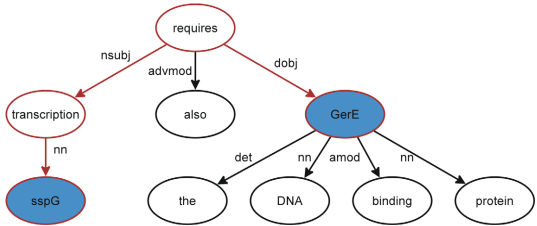


Spectrum tree kernel

- Compares vertex walk paths
- Adjustable with a path length parameter q

SVM classifier with kernels

Dependency parse-based kernels



k-band shortest path spectrum kernel

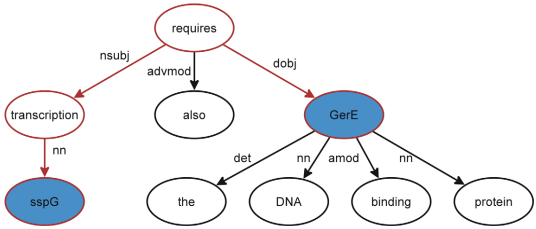
- Similar to Spectrum tree kernel (does vertex walks)
  - Works on dependency trees
  - Considers only shortest path between protein pairs
  - Additionally considers nodes within distance k of shortest path

$$\mathrm{SPS}_k(p_i, p_j) = \sum_{q=q_{min}}^{q_{max}} \max_{i \in p_i^q \, j \in p_j^q} (\text{t-score}_{L,E,D,l,e,d}(i,j)),$$

where L = token-, E = candidate entity-, D = dependency type-tolerance, l, e, d the scores

SVM classifier with kernels

Dependency parse-based kernels

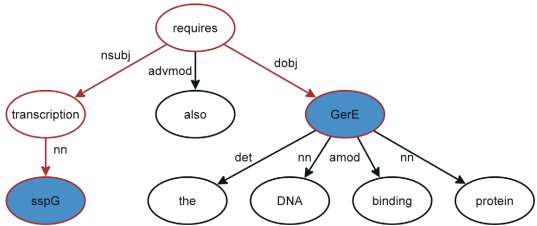


k-band shortest path spectrum kernel

- t-score examples:
  L = token-, E = candidate entity-, D = dependency type-tolerance, l, e, d the scores
  - t-score$_{0,0,-1,1,3,6}$(_ENT_1_ ← nn ← transcription, _ENT_1_ ← nn ← transcription) = 3+6+1
  - t-score$_{0,0,-1,1,3,6}$(_ENT_1_ ← nn ← transcription, _ENT_1_ ← nn ← expression) = 3+6+0
  - t-score$_{0,0,-1,1,3,6}$(_ENT_1_ ← nn ← transcription, protein ← nn ← _ENT_1_) = 0+6+0
  - t-score$_{0,0,-1,1,3,6}$(_ENT_1_ ← nn ← transcription, _ENT_1_ ← dobj ← requires) = 0

SVM classifier with kernels

Dependency parse-based kernels



Cosine-similarity

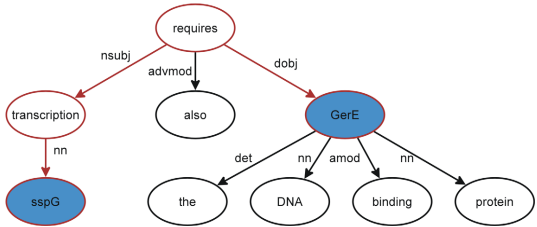
- Considers shortest path between protein mentions
- Represents word frequencies as vectors
- Compares candidate entities using cosine similarity

Edit distance

- Similarity score derived from edit distance
- Number of deletions, insertions and substitutions required to turn on into the other

SVM classifier with kernels

Dependency parse-based kernels



All-paths graph kernel

- Considers paths of all lengths
- Considers both dependency parse and sentence surface form
- Paths assigned higher weights the closer they are to shortest path between mentions

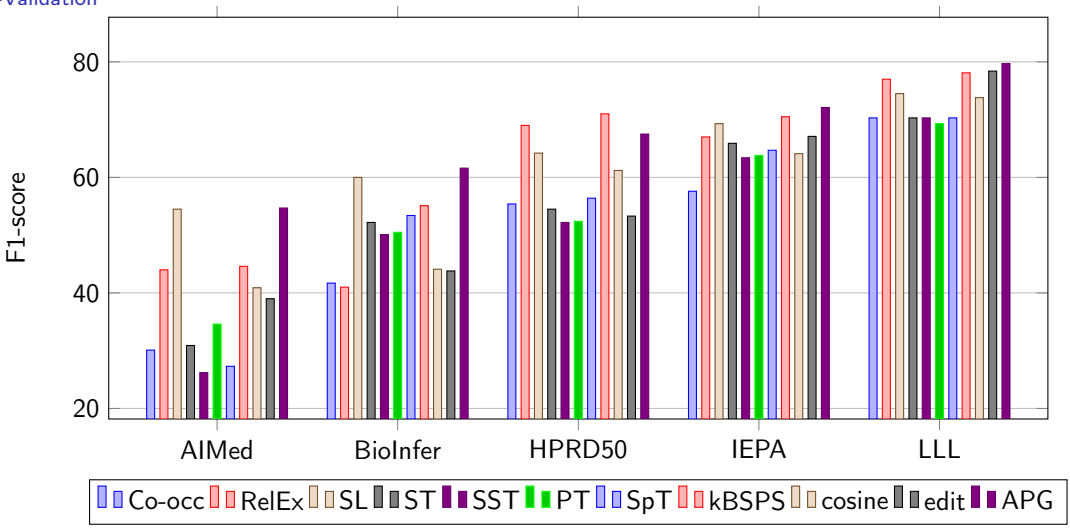
SVM classifier with kernels

Evaluation

- Uses the corpora introduced earlier (*AIMed*, *BioInfer*, *HPRD50*, *IEPA*, *LLL*)
- Three different evaluation methods
  - Cross-Validation (*document-level, 10-fold*)
  - Cross-Learning (*Train on four corpora, test on the fifth one*)
  - Cross-Corpus (*Train on one corpus, test on the other four*)

SVM classifier with kernels

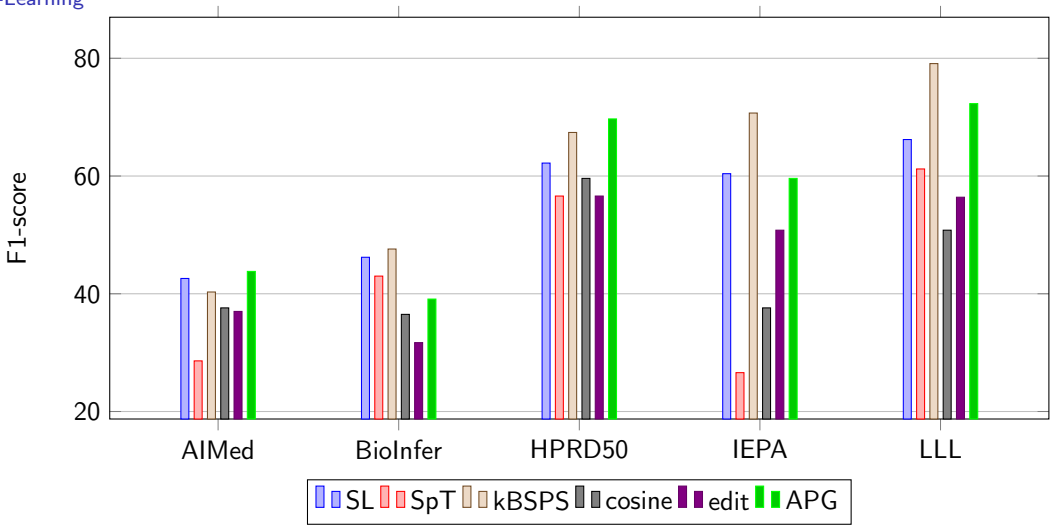
Cross-Validation



SL=Shallow linguistic, ST=Subtree, SST=Subset tree, PT=Partial tree, SpT=Spectrum tree, kBSPS=k-band shortest path spectrum, APG=all-paths graph

SVM classifier with kernels

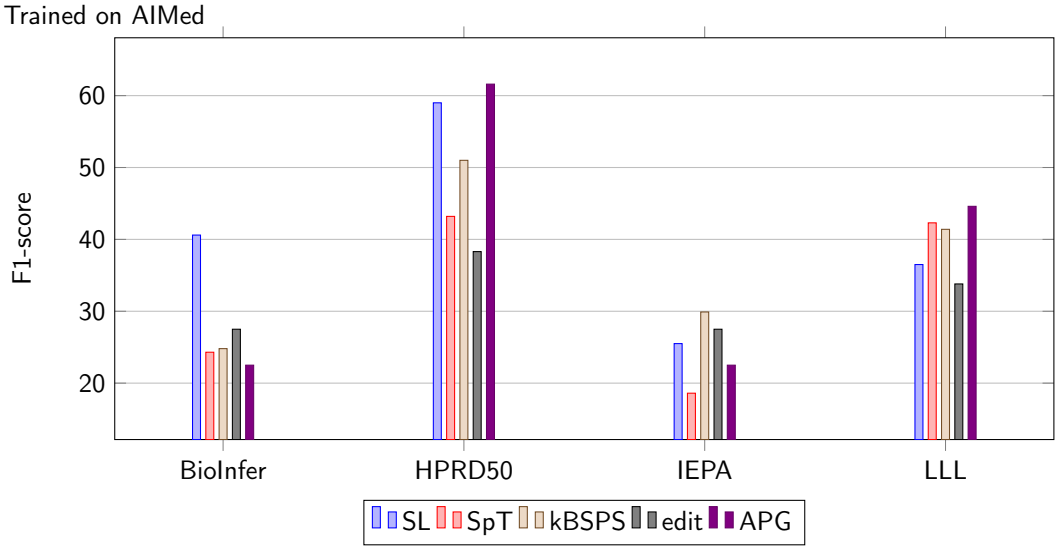
Cross-Learning



SL=Shallow linguistic, ST=Subtree, SST=Subset tree, PT=Partial tree, SpT=Spectrum tree, kBSPS=k-band shortest path spectrum, APG=all-paths graph

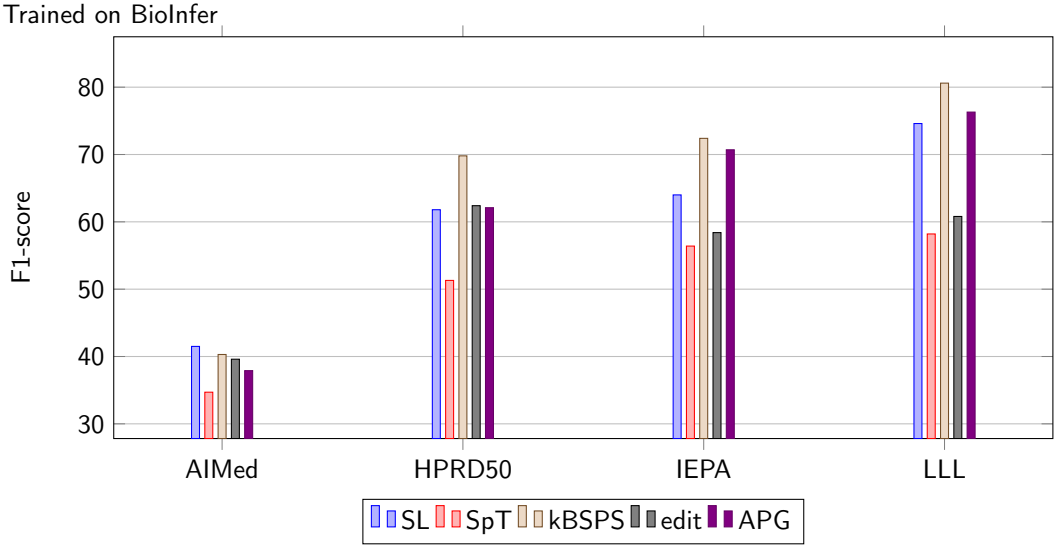
SVM classifier with kernels

Cross-Corpus



SVM classifier with kernels

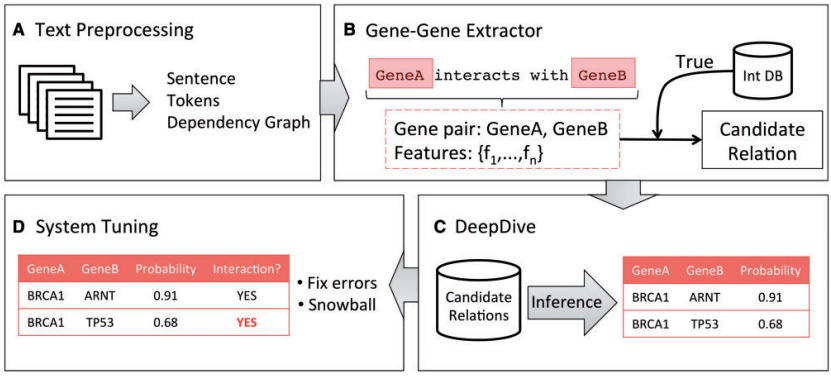
Cross-Corpus



Outline

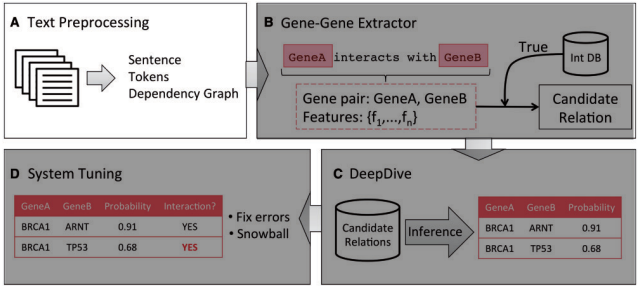
DeepDive

Implements whole processing pipeline



DeepDive

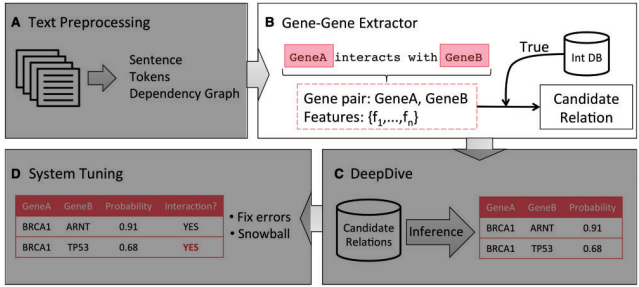
Preprocessing



- Takes PDF documents as input
  - Optical Character Recognition using Tesseract OCR
- Stanford CoreNLP for
  - Tokenization
  - POS-tagging, Named-entity recognition
  - Dependency parse

DeepDive

Gene-gene extractor



- Extracts set of candidate relations given sentence/document using dictionary
- Labels candidates with *is_correct*={true, false, unknown} using distant supervision
- Extracts features for each candidate relation

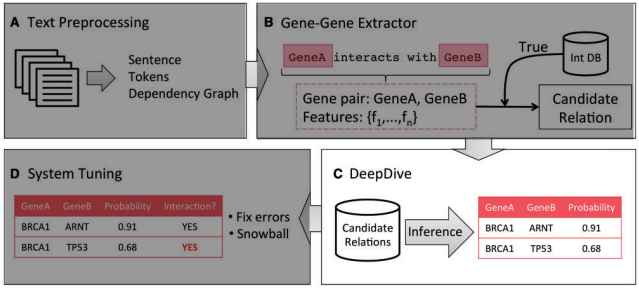
DeepDive

Gene-gene extractor

Feature Type	Example	System Representation
Dependency Path		nsubj(interacts, GeneA), prep(interacts, with), pobj(with, GeneB)
Prepositional Pattern	Interaction of GeneA and GeneB.	prep_pattern_[interaction_of]
Verb on Dependency Path		verb_on dep_path_[interact]
1-Word Window	GeneA interacts with GeneB in vitro.	g1_right_1gram_[interact]
2-Word Window	GeneA interacts with GeneB in vitro.	g1_right_2gram_[interact_with]
Word Sequence Window	GeneA interacts with GeneB in vitro.	word_seq_[interact_with]

DeepDive

Inference



- Models probability of candidates being PPI-interactions
- Implemented as probabilistic graphical model with single output *is_correct*

DeepDive

Evaluation

- Uses documents from PLOS (Public Library of Science) to train
  - PLOS One (multidisciplinary), PLOS Biology, PLOS Genetics
  - 102764, 3565, 4416 documents
  - candidate extraction yields 42,736 true, 65,606 false, 1,617,806 unknown
    - true and false candidates are added as unknown too
- Uses Gibbs Sampling to train
- Candidates with p(*is_correct=true*) > 0.9 are considered PPI interactions

Top 10 predictors for PPI interactions:

Feature	Weight
Single_Verb_Between_Genes_[bind]	1.25
Single_Verb_Between_Genes_[interact]	1.07
Verb_On_Dependency_Path_[bind]	0.91
Verb_On_Dependency_Path_[interact]	0.74
Single_Verb_Between_Genes_[regulate]	0.67
Verb_Between_Genes_[bind]	0.63
Verb_On_Dependency_Path_[regulate]	0.58
Window_Left_Gene1_Phrase_[GENE and]	0.57
Window_Right_Gene2_1gram_[protein]	0.57
Window_Left_Gene1_Phrase_[interaction between]	0.51

Evaluation against gold standard (DIP, Database of Interacting Proteins):

	Precision	Recall	F1-score
DIP	0.48	0.11	0.17
DIP-Rescue	0.68	0.14	0.23
DIP-Sentence	0.68	0.46	0.54
DIP-Indirect	0.76	0.49	0.59

Evaluation against manual annotated documents:

	Sentence-level precision	Document-level precision
Curation_Positive_Stringent	0.62	0.71
Curation_Positive_All	0.79	0.83

Conclusion

- PPI corpus resources are scarce
  - Existing resources are relative small and hard to compare due to differing annotations
  - Distant supervision can help create training data for machine learning classifiers
- Rule-based systems can still be competitive
- Pure PPI classifiers perform better than full pipeline systems

References

- Pyysalo, S., Airola, A., Heimonen, J., Björne, J., Ginter, F., & Salakoski, T. (2008). Comparative analysis of five protein-protein interaction corpora. BMC bioinformatics, 9(3), S6.
- Lehninger: Principles of Biochemistry, Fourth Edition 4th edition by David L. Nelson, Michael M. Cox (2004)
- Fischer B, Sandmann T, Horn T, Billmann M, Chaudhary V, Huber W, Boutros M (March 2015). "A map of directional genetic interactions in a metazoan cell“
- Ganapathiraju MK, Thahir M, Handen A, Sarkar SN, Sweet RA, Nimgaonkar VL, Loscher CE, Bauer EM, Chaparala S (April 2016). "Schizophrenia interactome with 504 novel protein-protein interactions“
- Pyysalo, Sampo, et al. "BioInfer: a corpus for information extraction in the biomedical domain." BMC bioinformatics 8.1 (2007): 50.
- Fundel, Katrin, Robert Küffner, and Ralf Zimmer. "RelEx—Relation extraction using dependency parse trees." Bioinformatics 23.3 (2006): 365-371.
- Tikk, D., Thomas, P., Palaga, P., Hakenberg, J., & Leser, U. (2010). A comprehensive benchmark of kernel methods to extract protein–protein interactions from literature. PLoS computational biology, 6(7), e1000837.
- Mallory, E. K., Zhang, C., Ré, C., & Altman, R. B. (2015). Large-scale extraction of gene interactions from full-text literature using DeepDive. Bioinformatics, 32(1), 106-113.

Discussion