

# Enriched Protein-Protein Interactions from Biomedical Text

Barry Haddow, Michael Matthews

University of Edinburgh

13th March 2007



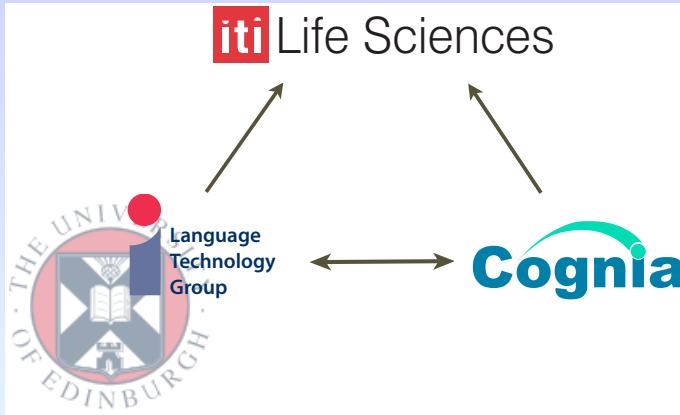
- The TXM Project
- Protein-Protein Interactions
- Enriched Protein-Protein Interactions
  - Properties
  - Attributes
  - Annotation
  - Methods and Evaluation



- Text Mining Programme funded (3 years from Feb 2005) by ITI Life Sciences
  - General goal of ITI is to encourage market-driven research that can be commercialised in Scotland.
  - Programme is developing technologies for finding, retrieving and storing structured data from unstructured text.
  - Approach is intended to be generic, but current focus is on biological data in research papers.



# Project Participants



Beatrice Alex, Claire Grover, Barry Haddow, Mijail Kabadjov, Ewan Klein, Michael Matthews, Leif Neilsen, Stuart Roebuck, Richard Tobin, Xinglong Wang



# Tools to Assist Biomedical Curation

- LTG is developing an NLP pipeline
  - Based on LT-XML2 tools
  - Linguistic preprocessing
  - Named Entity Recognition
  - Term Identification
  - Relation Extraction
- Cogna is developing a curation tool
  - Will use the output of the NLP pipeline
  - Aims to speed up biomedical curation



- Extraction of protein-protein interaction (PPI) mentions.

## PPI example

The interaction between **APC10** and **Smad3** was first observed in the yeast two-hybrid system.

- Annotation:
  - Entities: protein, prot\_frag, complex, fusion
  - Relations: protein-protein interaction, protein-protein association (PPA)
  - Normalisations: proteins normalised to Cognition lexicon (derived from RefSeq)



# Phase 1 (2005): Lessons Learnt

- The distinction between PPI and PPA was not clearly defined.
- Inter-annotator Agreement (IAA) was low in many categories
  - Even with PPI and PPA collapsed only  $\sim 50\%$   $F_1$  for RE.
- Component performances were 80-90% of IAA.
- Was used in the BioCreAtIvE II PPI task.
- Curation required more information about PPIs.
  - Direct, Experimentally-proven etc.
  - Cell-line, experimental method, drug treatment etc.



# Phase 2 (2006): Enriched protein-protein interactions

- ePPI
- Aim was to extract extra information of use to biologists
- Worked closely on data model with biologist from Cognition.
- Annotated new corpus
  - New entities, enriched PPIs.
  - Tightened up guidelines to improve IAA.
- Annotation:
  - Entities: Protein, Complex, Fusion, Fragment, Mutant, ExperimentalMethod, Modification, CellLine, DrugCompound
  - Relations: PPI, Frag (Fragment/Mutant to parent Protein)
  - Normalisations: proteins normalised to RefSeq
  - Properties: (Name, Value) pairs assigned to PPIs.
  - Attributes: Relation between PPI (or its participants) and Entity



# Properties

Name	Values	Explanation
IsPositive	Positive, Negative	The polarity of the statement about the PPI.
IsDirect	Direct, NotDirect	Whether the PPI is direct or not.
IsProven	Proven, Referenced, Unspecified	Whether the PPI is proven in the paper or not.

**Table:** The properties attached to each PPI and their possible values.



# Property Examples

## Positive, NotDirect, Proven

Mass spectroscopic analysis of proteins coprecipitating with **Sept6** identified the microtubule-associated protein **MAP4** as a septin binding partner.

## Positive, Direct, Referenced

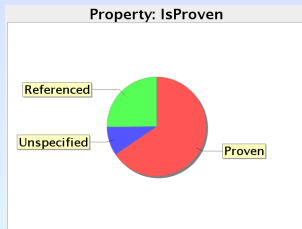
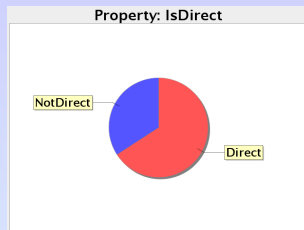
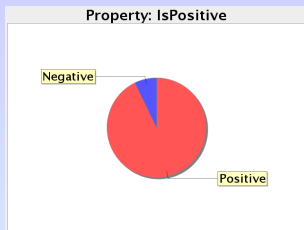
Indeed, Plk1 has been shown to bind **Cdc25C** by means of a **PBD-binding motif** in a CDK phosphorylation-dependent manner, and the phosphorylation of Cdc25C by Plk1 is important for the initiation of mitosis (6, 8).

## Positive, Negative, Direct, Proven

On the other hand, **KD-Cdk1** is able to bind and fully inhibit phosphorylated but not dephosphorylated **separase** (Figures 2B and 2C).



# Property Occurrence Counts



# Attributes

Name	Entity type	Explanation
InteractionDetectionMethod	ExperimentalMethod	The method used to detect the PPI.
ParticipantIdentificationMethod	ExperimentalMethod	The method used to detect the participant.
ModificationBefore	Modification	Modification of participant before interaction.
ModificationAfter	Modification	Modification of participant after interaction.
DrugTreatment	DrugCompound	Treatment applied to the participant.
CellLine	CellLine	Cell-line from which the participant was drawn.

**Table:** The attributes that could be attached to the PPIs, with their entity type.



# Attribute Examples

## ModificationAfter

**Tat** may also increase initiation of HIV-1 transcription by enhancing **phosphorylation** of **SP1**, a transcription factor involved in the basal HIV-1 transcription [14].

## InteractionDetectionMethod, CellLine

**Co-immunoprecipitation** of HIV-1 **Tat** with **LIS1** from **HeLa cells**.

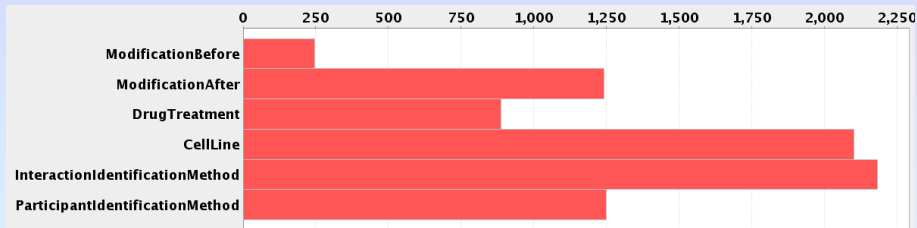
## InteractionDetectionMethod

Similar, LIS1 co-precipitated with Tat when Flag-Tat was **immunoprecipitated** with anti-Tat polyclonal. [...] These results indicate that **Tat** associates with **LIS1** in cultured cells.



# Attribute Occurrence Counts

Total number of PPIs: 12074



# Annotation

- Selected 217 papers from PubMed and PubMedCentral, as containing PPI.
- Annotated full-texts of papers (excluding materials and methods).
- XML form of papers was used when available, otherwise converted HTML.
- Nine annotators, all qualified biologists.
- 65 double annotated and 27 triple annotated - to measure IAA
- Data split into TRAIN (66%), DEVTEST (17%) and TEST (17%)
- IAA for PPI relations:  $F_1 = 64.77$



# Inter-Annotator Agreement: Properties

- Measured on corresponding pairs of annotated documents.
- Calculated  $F_1$  for each (Name, Value) pair.

Name	Value	$F_1$
IsPositive	Positive	99.57
	Negative	90.12
IsDirect	Direct	86.59
	NotDirect	61.38
IsProven	Proven	87.75
	Referenced	88.61
	Unspecified	34.38
All		87.17

- Low IAA for Unspecified - a “none-of-the-above” class



# Inter-Annotator Agreement: Attributes

Name	$F_1$
InteractionDetectionMethod	59.96
ParticipantIdentificationMethod	36.94
ModificationBefore	68.13
ModificationAfter	86.87
DrugTreatment	49.00
CellLine	64.38
All	61.58

- Scoring on only those attributes where both annotators had chosen to attach gives 95.10 overall.
- So, low IAA mainly caused by uncertainty whether to attach or not.



# Examples of Annotator Disagreement: 1

## Attachment of DrugTreatment

Additionally, Cry1Aa toxin binding capability was assayed by “slot blotting” all fractions and probing with biotin-Cry1Aa. The chromatogram in Fig. 1 displays the separation of cadherin and APN from BBMV proteins. APN isozymes of 100- and 110-kDa were detected that did not show Cry1Aa-binding in slot blot assays (Fig. 1; fractions 2425 and 3031).

- Proteins marked in red, ExperimentalMethods and DrugCompounds in blue.
- Both annotators marked (negative) PPI between “APN” and “Cry1Aa” in the third sentence.
- Both marked “slot blot assays” as ParticipantIdentificationMethod
- However only one attached “biotin” as a DrugTreatment



TXM

## Different InteractionDetectionMethod

When the peak fraction was immunoprecipitated with anti-geminin antibodies, Cdt171(193-447) was efficiently co-precipitated, confirming that this represents a complex between mini-geminin and Cdt172(193-447) (Figure 6B).

- Proteins and Fragments marked in red, ExperimentalMethods in blue
- Both annotators marked a PPI between “mini-geminin” and “193-447” .
- One attached “immunoprecipitated” as an InteractionDetectionMethod, the other marked “co-precipitated” .



# Relation Extraction: Methods and Evaluation

- Generate candidate relations from pairs of entities
- Convert each pair to a feature representations
  - Shallow linguistic features
  - Context, parts-of-speech, chunks, other entities, simple patterns, bigrams etc.
- Train a maximum entropy (ME) model using the feature representation.
- Performance:  $F_1 = 51.75$  (train on TRAIN , test on DEVTEST )
- Approximately 80% of IAA.
- Support Vector Machines (SVM) give similar performance



# Property Tagging: Methods

- Extract a set of features for each PPI.
- Simple, local features provides genericity.
  - Bigrams of context before, between and after participant entities in PPI.
  - Context around entities.
  - Text and type of entities.
  - Head words of chunks before, between and after entities.
- Train a model for each property name
- Used both ME and SVM



# Property Tagging: Evaluation

- Evaluated using  $F_1$  for each value, and overall micro-averaged.
- Tested on output of relation extractor.
- Used 5-fold cross-validation on TRAIN and DEVTEST .

Name	Value	Baseline	ME	SVM	IAA
IsPositive	Positive	97.54	97.87	98.15	99.57
	Negative	0.00	40.70	54.64	90.12
IsDirect	Direct	83.72	85.41	85.46	86.59
	NotDirect	0.00	54.96	57.05	61.38
IsProven	Proven	78.85	86.14	87.81	87.75
	Referenced	0.00	80.19	84.22	88.61
	Unspecified	0.00	24.04	21.15	34.38
All		77.33	84.68	85.71	87.17

- Performance stands up well against IAA
- Negative case is problematic
- Small class - linguistic complexity.



# Attribute Tagging: Methods

- Rule-based:

- Rules derived from training data
- Use words, POS tags, lemmas, chunks and entities
- Example rule for ModificationAfter attribute.

[NP containing Protein A] [Active voice VP containing Modification] [NP containing Protein B]

- 20 highest precision rules selected and applied in order of precision
- Machine Learning:
  - Attributes were treated as relations between PPIs and other entities
  - Candidates extracted from corpus and a feature representation created
  - Features similar to those used for properties
  - Additional features to indicate relative locations of entities



# Attribute Tagging: Evaluation

- Trained on TRAIN , tested on DEVTEST , with gold PPIs

Attribute	Baseline	Rule	ME	IAA
InteractionDetectionMethod	39.24	44.15	29.68	59.96
ParticipantIdentificationMethod	13.76	32.33	10.98	36.94
ModificationBefore	07.71	37.29	12.05	68.13
ModificationAfter	32.26	65.30	59.73	86.87
DrugTreatment	42.72	41.95	34.91	49.00
CellLine	44.74	44.74	45.17	64.38
All	26.75	43.61	24.10	61.58



# Conclusions and Future Work

- Some properties, and particularly attributes are difficult to annotate.
- Properties can be predicted using simple context features.
- Rule-based approaches work best for attributes (so far)
- Maximum entropy and Support Vector Machines offer similar performance, with SVM slightly better.
- Need to look at task-based evaluation
  - BioCreative II - PPI tasks
  - Curation speedup
- Investigate deeper linguistic models - parsing.
- Use non-local information.
- Try to bring in biological knowledge.



# The End

