

# The ITI TXM Corpora: Tissue Expressions and Protein-Protein Interactions

**Bea Alex**, Claire Grover, Barry Haddow, Mijail Kabadjov,  
Ewan Klein, Michael Matthews, Stuart Roebuck, Richard Tobin  
and Xinglong Wang

Building and Evaluating Resources  
for Biomedical Text Mining  
Marrakesh, 26 May 2008

# Outline

- 1 Introduction and Related Work
- 2 Document Selection and Preparation
- 3 Markables
- 4 Annotation Process
- 5 Inter-Annotator Agreement
- 6 Summary

# Introduction: TXM Project

- Text Mining Programme funded (3 yrs from Feb 2005) by ITI Life Sciences Scotland
- Goals
  - Encourage market-driven commercialisable research
  - Tools to extract structured data from unstructured text
  - Intended to be generic, but current focus is on biology
- Motivation
  - Biological databases are in demand
  - Manual curation is too slow
  - Automatic curation using text mining is too inaccurate
  - Assisted curation is just right

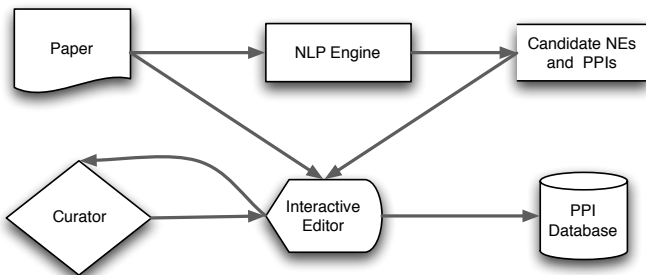
# Introduction: TXM Project

- Text Mining Programme funded (3 yrs from Feb 2005) by ITI Life Sciences Scotland
- Goals
  - Encourage market-driven commercialisable research
  - Tools to extract structured data from unstructured text
  - Intended to be generic, but current focus is on biology
- Motivation
  - Biological databases are in demand
  - Manual curation is too slow
  - Automatic curation using text mining is too inaccurate
  - Assisted curation is just right

# Introduction: TXM Project

- Text Mining Programme funded (3 yrs from Feb 2005) by ITI Life Sciences Scotland
- Goals
  - Encourage market-driven commercialisable research
  - Tools to extract structured data from unstructured text
  - Intended to be generic, but current focus is on biology
- Motivation
  - Biological databases are in demand
  - Manual curation is too slow
  - Automatic curation using text mining is too inaccurate
  - Assisted curation is just right

# Introduction: TXM Project



## Introduction: PPI and TE Corpora

Two large corpora of semantically annotated full-text biomedical research papers with the following characteristics:

- Size: large collection of documents to maximize performance of trained classifier
- Domains: protein-protein interactions and tissue expressions
- Text type/zone: full texts

## Introduction: PPI and TE Corpora

Two large corpora of semantically annotated full-text biomedical research papers with the following characteristics:

- Annotation guidelines: developed based on piloting
- Markables and levels of annotation: variety of semantic annotations (including normalisations)
- Inter-annotator agreement: measured throughout the annotation process
- Distributed data format: XML with annotations in standoff

# Document Selection and Preparation

## Document Selection

- Full papers selected from PubMedCentral OpenAccess and PubMed Central
- Filtering for PPI terms and manual selection by inspecting abstracts for mentions of presence/absence of mRNA or protein in any organism or tissue
- Annotators were allowed to reject papers for not being suitable for annotation
- Final annotated set: 217 PPI papers and 238 TE papers (not used during piloting)
- Documents split into train, devtest and test sets at ratio of 64:16:20

# Document Preparation

- Conversion to XML if XML version was not available  
LT-XML2 tools: <http://www.ltg.ed.ac.uk/software/xml>
- Tokenisation, sentence boundary detection, POS tagging, chunking, lemmatising
- Random set selected for double/triple annotation with multiple annotations left in the corpus
- In total, 74.6K sentences and 2.0M tokens for in the PPI corpus and 62.8K sentences and 1.9M tokens in the TE corpus

# Annotated Corpus Documents

Annotations	TRAIN	DEVTEST	TEST	All
PPI				
Single	65	25	35	125
Double	48	9	8	65
Triple	20	5	2	27
Total documents	133	39	45	<b>217</b>
Total annotations	221	58	57	<b>336</b>
TE				
Single	82	34	34	150
Double	68	7	11	86
Triple	1	0	1	2
Total documents	151	41	46	<b>238</b>
Total annotations	221	48	59	<b>328</b>

# Markables

## Named Entities

## Named Entities - PPI

Entity type	Count
<b>CellLine</b>	7,676
Complex	7,668
DrugCompound	11,886
ExperimentalMethod	15,311
Fragment	13,412
Fusion	4,344
Modification	6,706
Mutant	4,829
<b>Protein</b>	88,607

## Named Entities - TE

Entity type	Count
Complex	4,033
DevelopmentalStage	1,754
Disease	2,432
DrugCompound	16,131
ExperimentalMethod	9,803
Fragment	4,466
Fusion	1,459
GOMOP	4,647
Gene	12,059
<b>mRNAcDNA</b>	8,446
Mutant	1,607
<b>Protein</b>	60,782
<b>Tissue</b>	36,029

## Named Entities

- Additional entities: interaction and expression level words.
- Annotation of nested entities was allowed but not of crossing ones.
- Discontinuous coordinations were annotated as nesting entities.
- Annotators were able to override the tokenisation with entity boundaries stored as character offsets.
- XML representation allows retokenisation as proposed by Grover et al. (2006).

# Markables

## Normalisations

# Normalisations

Normalisation: linking entity mentions to unique ontology identifiers and thereby disambiguating them

- Full normalisation and species normalisation for proteins, genes and mRNAs
- Full normalisation for part of the data: assigning RefSeq identifier to protein and mRNA terms and EntrezGene identifier to gene terms
- Species normalisation for all of the data: assigning NCBI taxonomy identifier to all three types of terms (Wang and Grover, LREC 2008).

# Normalisations

Database	PPI	TE
NCBI Taxonomy	Protein	Gene, mRNACDNA, Protein, GOMOP
RefSeq	Protein	Protein, mRNACDNA
EntrezGene	Protein	Gene, mRNACDNA, Protein, GOMOP
ChEBI	—	DrugCompound
MeSH	—	Tissue

# Normalisations

## Special cases:

- Species mismatch: only species normalisation.
- Several host species: multiple normalisations with a unique species each time.
- Host species unclear: normalisation with default species *Homo sapiens* and keyword “gen” and *Homo sapiens* species normalisation or just “gen” if that is the incorrect species.

# Markables: Relations

## Relations

## Markables: Relations

Relations enriched with additional biomedical information enabling finer-grained classification

- Relation types: different types of binary relations between two entity mentions
- Relation properties: name-value pair assigned to a relation conveying additional information
- Relation attributes: named links between relations and other entities

## Markables: Relation Types

Corpus	Relation type	Count
PPI	PPI	11,523
PPI	FRAG	16,002
TE	TE	12,426
TE	CHILD-PARENT	4,735

## Markables: Relation Properties

Name	Value	PPI	TE
IsPositive	Positive	10,718	10,243
	Negative	836	2,067
IsDirect	Direct	7,599	—
	NotDirect	3,977	—
IsProven	Proven	7,562	9,694
	Referenced	2,894	1,837
	Unspecified	1,096	736

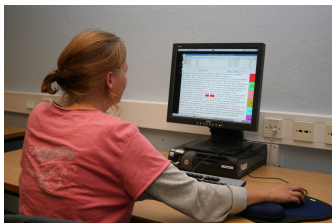
## Markables: Relation Attributes - PPI

Name	Entity type	Count
ModificationBeforeEntity	Modification	240
ModificationAfterEntity	Modification	1,198
DrugTreatmentEntity	DrugCompound	844
CellLineEntity	CellLine	2,000
ExperimentalMethodEntity	ExperimentalMethod	1,197
MethodEntity	ExperimentalMethod	2,085
InteractionWordEntity	InteractionWord	11,386

## Markables: Relation Attributes - TE

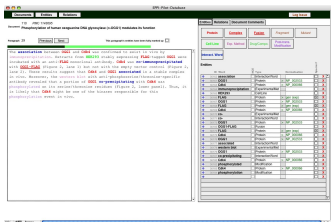
Name	Entity type	Count
te_rel_ent-drug-compound	DrugCompound	1,549
te_rel_ent-exp-method1	ExperimentalMethod	1,878
te_rel_ent-disease	DiseaseType	332
te_rel_ent-dev-stage	DevelopmentalStage	327
te_rel_ent-expr-word	ExpressionLevelWord	2,815

# Annotation Process



- Biologist annotators + annotation manager
- Close collaboration with NLP team
- Piloting phase to establish list of markables and create comprehensive annotation guidelines
- Main annotation phase using in-house annotation tool developed using FilemakerPro and customised version of Callisto

# Annotation Process



- Documents selected for double/triple annotation
- Weekly annotation meetings and latest IAA measurements
- Annotation of full-text papers (excluding contact details, reference and materials and methods sections)
- All annotated documents in in-house XML format with annotations in standoff

## Example

*Rrs1p has a two-hybrid interaction with L5.*

- Two proteins of species *Saccharomyces cerevisiae* (4932) normalised to the RefSeq identifiers NP\_014937 and NP\_015194
- One experimental method and an interaction word
- A direct, positive and proven relation between both proteins
- A relation attribute specifying that the interaction was detected using the experimental method

## Example

*Rrs1p* has a two-hybrid interaction with *L5*.

- Two proteins of species *Saccharomyces cerevisiae* (4932) normalised to the RefSeq identifiers NP\_014937 and NP\_015194
- One experimental method and an interaction word
- A direct, positive and proven relation between both proteins
- A relation attribute specifying that the interaction was detected using the experimental method

## Example

*Rrs1p* has a *two-hybrid interaction* with *L5*.

- Two proteins of species *Saccharomyces cerevisiae* (4932) normalised to the RefSeq identifiers NP\_014937 and NP\_015194
- One experimental method and an interaction word
- A direct, positive and proven relation between both proteins
- A relation attribute specifying that the interaction was detected using the experimental method

## Example

*Rrs1p* has a two-hybrid interaction with *L5*.

- Two proteins of species *Saccharomyces cerevisiae* (4932) normalised to the RefSeq identifiers NP\_014937 and NP\_015194
- One experimental method and an interaction word
- **A direct, positive and proven relation between both proteins**
- A relation attribute specifying that the interaction was detected using the experimental method

## Example

*Rrs1p* has a *two-hybrid interaction* with *L5*.

- Two proteins of species *Saccharomyces cerevisiae* (4932) normalised to the RefSeq identifiers NP\_014937 and NP\_015194
- One experimental method and an interaction word
- A direct, positive and proven relation between both proteins
- A relation attribute specifying that the interaction was detected using the experimental method

## Example

```
...  
<s><w id="A33864">Rrs1p</w> <w id="A33870">has</w>  
<w id="A33874">a</w> <w id="A33876">two</w>  
<w id="A33879">-</w> <w id="A33880">hybrid</w>  
<w id="A33887">interaction</w> <w id="A33899">with</w>  
<w id="A33904">L5</w> <w id="A33906">.</w></s>  
...  
<ent id="e933262" norm="refseq:NP_014937" type="Protein"  
  species="4932" sw="A33864" ew="A33864">Rrs1p</ent>  
<ent id="e933263" type="ExperimentalMethod" sw="A33876"  
  ew="A33880">two-hybrid</ent>  
<ent id="e933264" type="InteractionWord" sw="A33887"  
  ew="A33887">interaction</ent>  
<ent id="e933265" norm="refseq:NP_015194" conf="100" type="Protein"  
  species="4932" sw="A33904" ew="A33904">L5</ent>  
...  
<relation type="ppi" id="r903106" lsProven="Proven"  
  lsDirect="Direct" lsPositive="Positive" >  
<argument ref="e933262" />  
<argument ref="e933265" />  
<attribute name="MethodEntity" ref="e933263" />  
<attribute name="InteractionWordEntity" ref="e933264" /></relation>
```

...

## Example

```
<s><w id="A33864">Rrs1p</w> <w id="A33870">has</w>  
<w id="A33874">a</w> <w id="A33876">two</w>  
<w id="A33879">-</w><w id="A33880">hybrid</w>  
<w id="A33887">interaction</w> <w id="A33899">with</w>  
<w id="A33904">L5</w> <w id="A33906">.</w></s>
```

## Example

```
<ent id="e933262" norm="refseq:NP_014937" type="Protein"  
  species="4932" sw="A33864" ew="A33864">Rrs1p</ent>  
<ent id="e933263" type="ExperimentalMethod" sw="A33876"  
  ew="A33880">two-hybrid</ent>  
<ent id="e933264" type="InteractionWord" sw="A33887"  
  ew="A33887">interaction</ent>  
<ent id="e933265" norm="refseq:NP_015194" conf="100" type="Protein"  
  species="4932" sw="A33904" ew="A33904">L5</ent>
```

## Example

```
<relation type="ppi" id="r903106" IsProven="Proven"  
  IsDirect="Direct" IsPositive="Positive" >  
<argument ref="e933262" />  
<argument ref="e933265" />  
<attribute name="MethodEntity" ref="e933263" />  
<attribute name="InteractionWordEntity" ref="e933264" /></relation>
```

# Inter-Annotator Agreement

IAA

# Inter-Annotator Agreement: Named Entities

IAA calculated:

- For each pair of annotations on the same document
- As precision, recall and F1 using CoNLL-scoring
- As a micro-average giving equal weight to each example

## Inter-Annotator Agreement: Named Entities - PPI

Entity type	F1	TP
CellLine	81.6	(2,456)
Complex	76.4	(2,243)
DrugCompound	76.4	(3,705)
ExperimentalMethod	74.0	(4,673)
Fragment	75.3	(3,985)
Fusion	78.5	(1,270)
Modification	87.6	(1,900)
Mutant	60.4	(1,008)
Protein	91.6	(32,799)
All	84.9	(54,039)

## Inter-Annotator Agreement: Named Entities - TE

Entity type	F1	TP
Complex	82.6	(886)
DevelopmentalStage	72.7	(357)
Disease	74.3	(435)
DrugCompound	84.9	(4,453)
ExperimentalMethod	76.7	(2,013)
Fragment	77.7	(1,179)
Fusion	73.9	(359)
GOMOP	50.2	(655)
Gene	77.7	(1,911)
mRNAcDNA	78.1	(1,768)
Mutant	63.9	(310)
Protein	90.3	(16,329)
Tissue	84.1	(8,210)
All	83.8	(38,865)

# Inter-Annotator Agreement: Normalisations

IAA calculated:

- For each pair of normalised entities where the annotators agreed on the entity annotation (general normalisations removed)
- As precision, recall and F1
- As a micro-average giving equal weight to each example

## Inter-Annotator Agreement: Normalisations

Type	PPI		TE	
DrugCompound	—		97.7	(215)
GOMOP	—		77.3	(214)
Gene	—		95.1	(1,463)
mRNAcDNA	—		88.0	(892)
Protein	88.4	(7,595)	90.0	(5,979)
Tissue	—		82.9	(6,776)
All	88.4	(7,595)	83.8	(15,785)

# Inter-Annotator Agreement: Relations

IAA calculated:

- For each relation type/property/attribute where the annotators agreed on the entity annotation
- As precision, recall and F1
- As a micro-average giving equal weight to each example

## Inter-Annotator Agreement: Relations

Type	PPI		TE	
PPI	67.0	(2,729)	—	—
TE	—	—	70.1	(2,078)
FRAG	84.6	(3,661)	84.0	(1,012)
All	76.1	(6,390)	74.1	(3,090)

# Inter-Annotator Agreement: Relation Properties

Name	Value	PPI		TE	
IsPositive	Positive	99.6	(2,553)	97.2	(1,807)
	Negative	90.1	(155)	88.9	(280)
IsDirect	Direct	86.8	(1,746)	—	
	NotDirect	61.4	(449)	—	
IsProven	Proven	87.8	(1,543)	92.8	(1,547)
	Referenced	88.6	(626)	75.3	(204)
	Unspecified	34.4	(448)	29.3	(38)
	All	87.2	(7,165)	91.2	(3,779)

# Inter-Annotator Agreement: Relation Attributes - PPI

Name	IAA	
ModificationBeforeEntity	65.3	(31)
ModificationAfterEntity	86.7	(248)
DrugTreatmentEntity	45.4	(61)
CellLineEntity	64.0	(244)
ExperimentalMethodEntity	36.9	(94)
MethodEntity	55.4	(274)
All	59.6	(952)

## Inter-Annotator Agreement: Relation Attributes - TE

Name	IAA
te_rel_ent-drug-compound	77.9 (229)
te_rel_ent-exp-method1	81.3 (261)
te_rel_ent-disease	64.0 (16)
te_rel_ent-dev-stage	57.8 (13)
All	77.2 (521)

# Summary

- ITI TXM corpora: result of one of the largest biomedical corpus annotation projects attempted to date.
- Annotated data in two domains of crucial importance to biologists.
- Annotation of normalisations for multiple entity types and multiple species.
- Extensive annotation guidelines were developed as a result of several rounds of piloting.

# Summary

- Annotator consistency was continuously measured in terms of IAA.
- Interaction between annotators and NLP team was crucial at all stages.
- Useful for text mining research, e.g. comparing different systems.
- Both corpora include a selection of full-text papers and will be distributed in XML.

## Acknowledgements

- ITI Life Sciences, Scotland  
<http://www.itilifesciences.com>
- Annotation teams lead by Elizabeth Fairley and Lynn Morrice
- Cogna EU's software development team
- NLP team at the Language Technology Group at the School of Informatics, University of Edinburgh
- Malvina Nissim, Kirsten Lillie and Henk Harkema

## Questions?

Thank you! Questions?

# Discontinuous Coordinations and Tokenisation Override

Discontinuous coordination:

“A and B cells”

- “A and B cells” = entity 1 referring to “A cell”
- “B cells” = entity 2 referring to “B cell”

Tokenisation override:

“Cdt1(193-447)” (one token)

- Cdt1 = Protein (with end offset of -9)
- 193-447 = Fragment (with start offset of 5)

# Discontinuous Coordinations and Tokenisation Override

Discontinuous coordination:

“A and B cells”

- “A and B cells” = entity 1 referring to “A cell”
- “B cells” = entity 2 referring to “B cell”

Tokenisation override:

“Cdt1(193-447)” (one token)

- Cdt1 = Protein (with end offset of -9)
- 193-447 = Fragment (with start offset of 5)