



✱ Critical Assessment for Information Extraction in Biology ✱

# The 2nd BioCreative Evaluation: Lessons Learned & Future Directions

Lynette Hirschman, MITRE

5th Fraunhofer Symposium on  
Text Mining

Sept. 24-25, 2007

# Outline

- What is BioCreative?
- What are the results
  - BioCreative I
  - BioCreative II
  - What did we learn?
- Where next?
  - New tasks
  - Open research questions

# Acknowledgements

- Thanks to the organizers for BioCreative II
  - Gene Mention: John Wilbur, Larry Smith, Lorrie Tanabe
  - Gene Normalization: Alex Morgan, Jeff Colombe, Marc Colosimo
  - Protein-Protein Interaction: Martin Krallinger, Florian Leitner, Alfonso Valencia
  - The IntAct (EBI) and MINT (U Rome) protein-protein interaction database groups
- Thanks to
  - The National Science Foundation for support of MITRE's activities on BioCreative I and II
- Thanks to Dr. Juliane Fluck for her invitation and her participation in BioCreative I and II

# Why a Challenge Evaluation?

- Many people reporting results on different problems
  - Can't compare!
  - Can't tell what works and what doesn't
  - Can't tell if the community is making progress
- Solution: Invite people to run their systems in a challenge evaluation
  - Long history of success in bioinformatics (CASP<sup>1</sup>) and text mining (MUC<sup>2</sup>, TREC<sup>3</sup>)
  - Design challenge to address central problem(s) and measure performance – over time

<sup>1</sup>Critical Assessment of Techniques for Protein Structure Prediction

<sup>2</sup>Message Understanding Conference

<sup>3</sup> Text Retrieval Conference

# Why BioCreative?

- For news, automated systems exist now that can:
  - Identify entities (85-95% F-measure\*)  
Entities include persons, places, time, money,...
  - Extract relations among entities (70-80% F), e.g.,  
*person lives\_in place*
  - Answer simple factual questions using large document collections at 75-85% accuracy  
(question answering)
- Goals of BioCreative:
  - How good is text mining applied to biology?
  - Is text mining is good enough to be applied to real biological problems?

F-measure is harmonic mean of precision and recall:  $2 * P * R / (P + R)$

Precision =  $TP / (TP + FP)$ ; Recall =  $TP / (TP + FN)$

- First BioCreative 2003-2004
  - Participation: 27 teams/10 countries
  - Tasks
    - Gene mention
    - Gene normalization for fly, mouse, yeast
    - Functional annotation: find evidence passages for Gene Ontology annotation for proteins in full text
- Second BioCreative 2006-2007
  - Participation: 44 teams/13 countries
  - Tasks
    - Gene mention
    - Gene normalization (human)
    - Protein-protein interaction

# BioCreative Tasks

- Foundational tasks
  - Gene mention: identify all gene or protein mentions in running text
    - Results: ~90% balanced precision/recall
  - Gene normalization: list unique identifiers (EntrezGene) of genes/proteins in a text
    - Results: ~ 80-90% balanced precision/recall
- Biological tasks
  - 1<sup>st</sup> BioCreative: functional annotation for proteins (identifying evidence passage supporting assignment of a GO term to a protein)
  - 2<sup>nd</sup> BioCreative: supporting the curation pipeline for protein-protein interaction, based on IntAct and MINT
  - Results: these are very hard tasks!

# Philosophy of BioCreative

- Address needs of working biologists and bioinformaticians
  - Identify real needs
  - Enlist users to provide data and help evaluate (e.g., biological database curators)
  - Choose tasks that make use of accepted standards (e.g., Gene Ontology<sup>1</sup>, PSI<sup>2</sup>)
  - Define evaluation metrics that make sense to the end user and to the developer
- Provide resources, e.g., by making (training and test) data available
  - Ideal: make tools and/or processed data available
- Encourage people to create new challenge evaluations for the community

<sup>1</sup>Gene Ontology; <sup>2</sup>Proteomics Standards Initiative



# BioCreative Gene Mention Task\*

- Identify all mentions of genes, proteins in sentences from PubMed abstracts

Mutation of TTF-1-binding sites (TBE) 1, 3, and 4 in combination markedly decreased transcriptional activity of SP-A promoter-chloramphenicol acetyltransferase constructs containing SP-A gene sequences from -256 to +45 .

- Serves as building block for more complex tasks
- Results based on 15,000 expert annotated sentences
- Many approaches based on statistical word co-occurrence (HMM, Conditional Random Fields)
- High score: F-measure (balanced precision/recall)
  - 1<sup>st</sup> BioCreative: 0.82
  - 2<sup>nd</sup> BioCreative: 0.87; 0.90 combined systems

BioCreative II Gene Mention task designed and evaluated by John Wilbur, Larry Smith, Lorrie Tanabe at NCBI

# Why is Gene Mention Hard?

- Not just another Named Entity Recognition task
  - Millions of gene names
  - New names constantly created
  - Standardized names not used
- Morphological and contextual similarity to other types *e.g.*, cells, drugs
  - Human annotators must have specific educational background
- Genes are not simple concepts, *e.g.* polymorphism, multiple alleles, translocation, cross-species analogues

# BioCreative II Gene Mention Results\*

<i>rank</i>	<i>p</i>	BioCreative		<i>signif</i>	<i>% alt</i>
		<i>r</i>	<i>F</i>		
1	88.48	85.97	87.21	4-21	32.48
2	89.30	84.49	86.83	6-21	14.02
3	84.93	88.28	86.57	6-21	14.08
4	87.27	85.41	86.33	7-21	31.77
5	85.77	86.80	86.28	7-21	16.67
6	82.71	89.32	85.89	7-21	16.02
7	86.97	82.55	84.70	8-21	14.83
8	84.35	81.39	82.85	10-21	14.57
9	86.28	79.66	82.84	10-21	14.55
10	85.22	78.44	81.69	12-21	33.02
11	85.54	76.83	80.95	12-21	19.76
12	72.95	88.49	79.97	14-21	16.82
13	92.67	68.91	79.05	15-21	19.73
14	88.83	69.70	78.11	16-21	37.05
15	80.46	73.61	76.88	17-21	20.43
16	82.28	71.08	76.27	18-21	16.80
17	84.32	68.57	75.63	18-21	34.02
18	71.68	62.33	66.68	19-21	28.23
19	65.83	61.55	63.62	20, 21	27.23
20	60.56	64.11	62.29	21	31.71
21	50.09	46.12	48.02	-	28.46

BioCreative I:

Top F-score = 82.2

BioCreative II:

**Top F-score ~ 0.87**

Top 3 systems not significantly different

Results from simulated “combined system”:

**Top F-score ~ 0.90**

\*Wilbur, Smith, Tanabe, “BioCreative 2 Gene Mention Task,” Proc 2<sup>nd</sup> BioCreative Challenge Evaluation Workshop, Madrid, 2007

# Gene Normalization Task Definition\*

- For all human genes and direct gene products mentioned in a given MEDLINE abstract
- Return
  - A list of the NCBI gene identifiers (Entrez Gene)
  - A text excerpt of the mention from the abstract, for each identifier
- Scoring
  - Compare list of gold standard identifiers to list generated by participating system
- Corpora for BioCreative II
  - Training: 281 abstracts
  - Test: 262 abstracts with 785 gene IDs

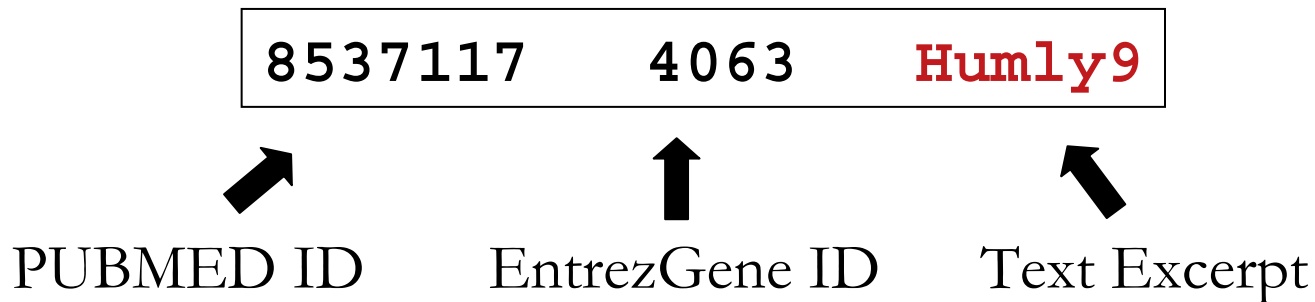
Morgan, A.A., et al. Evaluating the Automatic Mapping of Human Gene and Protein Mentions to Unique Identifiers. in Pacific Symposium for Biocomputing. 2007. Maui.

# Example: Gene Normalization

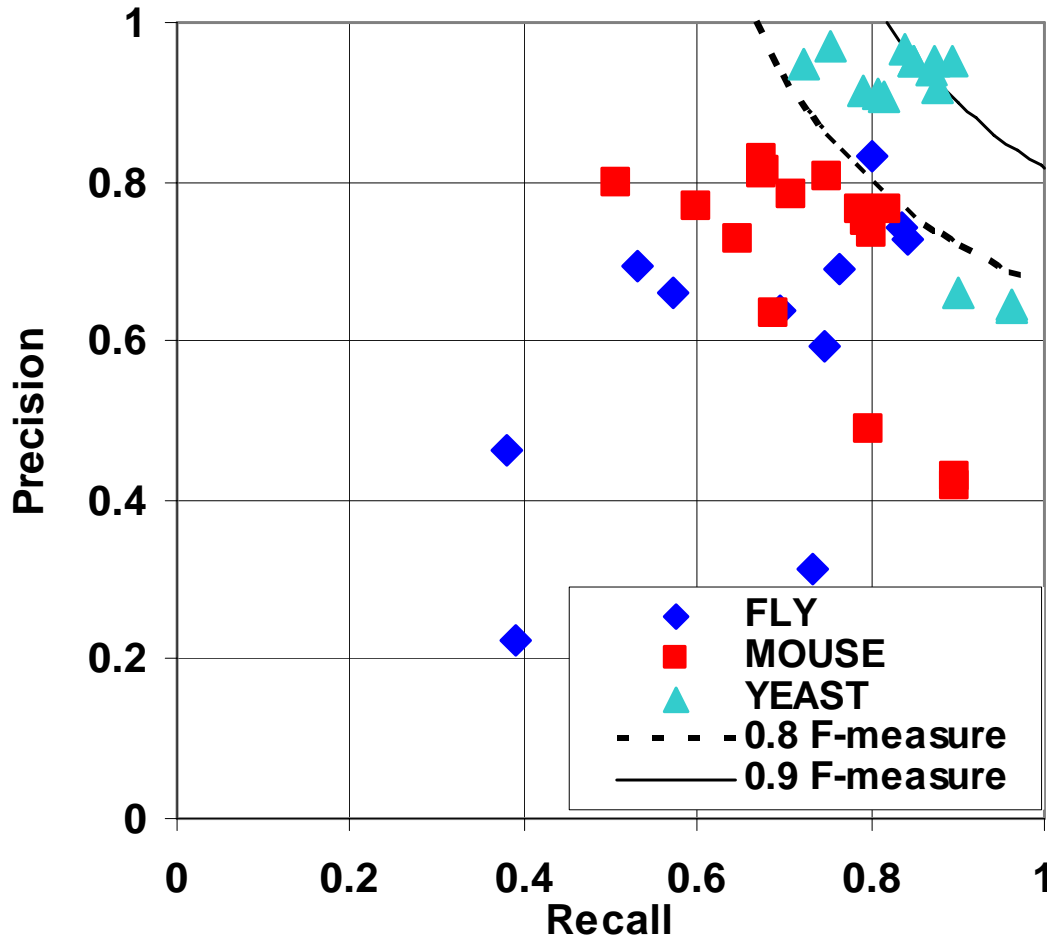
Input passage (PUBMED ID 8537117):

Ly9 is a mouse cell membrane antigen found on all lymphocytes and coded for by a gene that maps to chromosome 1. We previously described the isolation and characterization of a full-length cDNA clone for mouse Ly9. Using cross-species hybridization we isolated cDNA clones encoding the human homologue **Humly9**.

System produces:



# BioCreative I Results (Fly, Mouse, Yeast)



9 groups submitted  
16 runs

Yeast results:

Good (>90%)

Best: 0.93 F-score

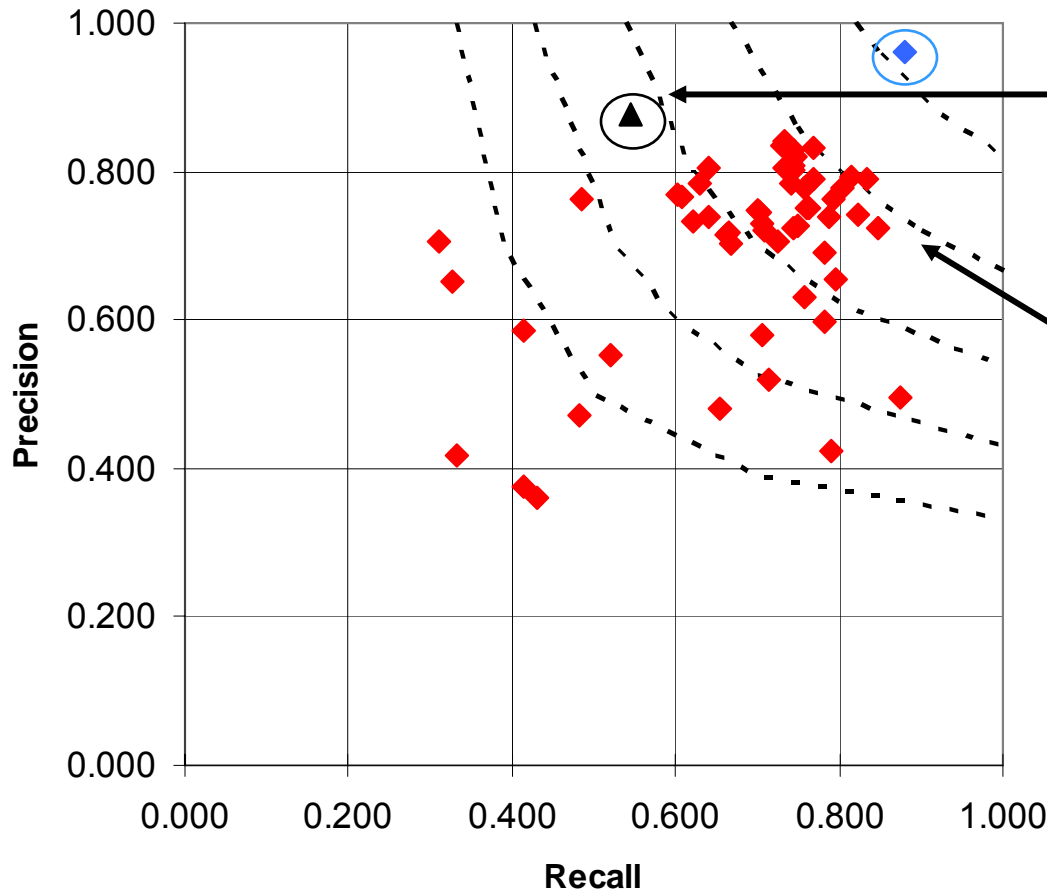
Fly, Mouse:

Harder! (<80%)

Best Fly: 0.81 F

Best Mouse 0.79 F

# BioCreative II Results



Combined results  
from pooling answers

$R = 0.88$   
 $P = 0.96$

Noisy training data:

$R = 0.544$   
 $P = 0.877$

Overall results:

3 teams  $\geq 0.80$  F  
9 teams  $> 0.75$  F

High recall:

0.88 @ 0.50 P

Morgan and Hirschman, Overview of BioCreative II Gene Normalization.  
Proc. 2<sup>nd</sup> BioCreative Challenge Evaluation Workshop, Madrid 2007.

# Results Table (Best Run per Team)

Team/ Run	Recall	Precision	F-measure Micro-Avg	Rank Micro	F-measure Macro-Avg	Rank Macro	Signif Range
T042_1	0.833	0.789	<b>0.810</b>	1	0.811	1	<b>3-20</b>
T034_1	0.815	0.792	<b>0.804</b>	2	0.782	2	8-20
T013_1	0.768	0.833	<b>0.799</b>	3	0.779	3	8-20
T004_1	0.734	0.841	0.784	4	<b>0.777</b>	4	8-20
T109_1	0.824	0.743	0.781	5	0.775	5	8-20
T104_1	0.743	0.807	0.774	6	0.773	6	9-20
T101_2	0.743	0.801	0.771	7	0.755	7	10-20
T107_1	0.74	0.784	0.761	8	<b>0.739</b>	<b>*9</b>	12-20
T113_2	0.761	0.752	0.756	9	<b>0.745</b>	<b>*8</b>	11-20
T108_3	0.749	0.726	0.737	10	0.724	10	13-20
T007_2	0.703	0.746	0.724	11	<b>0.694</b>	<b>*12</b>	16-20
T017_1	0.708	0.72	0.714	12	<b>0.71</b>	<b>*11</b>	15-20
T110_1	0.629	0.783	0.698	13	<b>0.685</b>	<b>*14</b>	16-20
T111_3	0.664	0.717	0.689	14	<b>0.664</b>	<b>*15</b>	17-20
T030_1	0.661	0.716	0.687	15	<b>0.649</b>	<b>*16</b>	17-20
T006_2	0.606	0.767	0.677	16	<b>0.686</b>	<b>*13</b>	19-20
T036_1	0.713	0.52	0.602	17	0.595	17	19-20
T014_1	0.485	0.762	0.593	18	0.584	18	20
T102_3	0.79	0.425	0.552	19	0.559	19	20
T058_2	0.415	0.375	0.394	20	0.398	20	

Rank changes micro- vs macro-average in **bold\***  
Significance at 10% level, one-sided t-test

# Why Gene Normalization is Hard

- Descriptions: *bile salt-stimulated milk lipase*
- Non-standard short forms: c10 for homeoprotein C10
- Mentions of families: “the MMP-19 subfamily”
- Ambiguity
  - Among species – does gene/protein name refer to human or model organism (e.g., rat)?
  - Distinguishing from other expressions, e.g., “activated by the Erk or p38 pathway”
- Range expressions  
freac1-freac7 must be expanded into all 7 genes
- What constitutes a mention of a gene/protein?
  - “another clone codes for a protein homologous to the product of the C. elegans gene lin-7”
  - “recently isolated huntingtin-interacting proteins”

# Protein Protein Interaction\*

- A series of four interrelated tasks supporting two protein-protein interaction databases:
  - MINT (University of Rome) led by Gianni Cesarini
  - IntAct (EBI) led by Henning Hermjakob
- Tasks assembled by Martin Krallinger, Florian Leitner, Alfonso Valencia at CNIO (Spanish National Cancer Research Center)

\*Krallinger and Valencia, Evaluating the Detection and Ranking of Protein Interaction Relevant Articles,” Proc. of the 2nd BioCreative Challenge Evaluation Workshop, Madrid, 2007

Krallinger, Leitner and Valencia, “Assessment of the Second BioCreative PPI Task: Automatic Extraction of Protein-Protein Interactions,” Proc. of the 2<sup>nd</sup> BioCreative Challenge Evaluation Workshop, Madrid, 2007

Use of Krallinger and Valencia’s slides is gratefully acknowledged

# PROTEIN-PROTEIN INTERACTION (PPI) TASK

**INTERACTION ARTICLE  
SUB-TASK (IAS)**

**INTERACTION PAIR  
SUB-TASK (IPS)**

**INTERACTION METHOD  
SUB-TASK (IMS)**

**INTERACTION SENTENCES  
SUB-TASK (ISS)**

## TASK GOALS

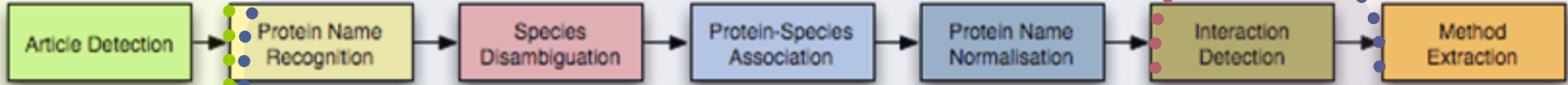
25 teams  
128 runs

**Detection of curation-relevant  
articles**

**Extraction of normalized  
protein interaction pairs**

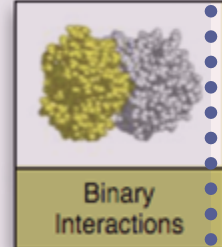
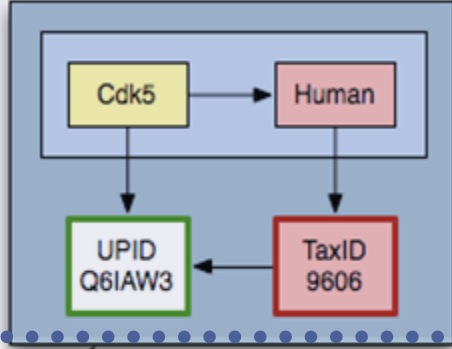
**Association of articles to  
interaction detection methods**

**Retrieval of best summarizing  
Interaction description passage**



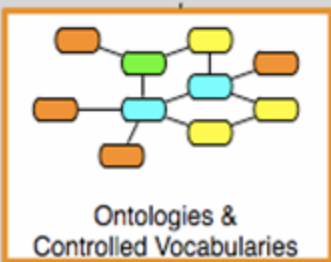
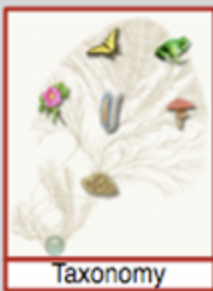
Cdk5 and its essential activator p25 play an important role in neuronal migration and proper development of the brain cortex. We show that p25 binds directly to *tin* and *tinman*. Microtubule polymerization has not been shown to be essential for p25 function. We show that p25 binds directly to *tin* and *tinman*. Microtubule polymerization has not been shown to be essential for p25 function. We show that p25 binds directly to *tin* and *tinman*. Microtubule polymerization has not been shown to be essential for p25 function.

**IPS: Protein Interactions**

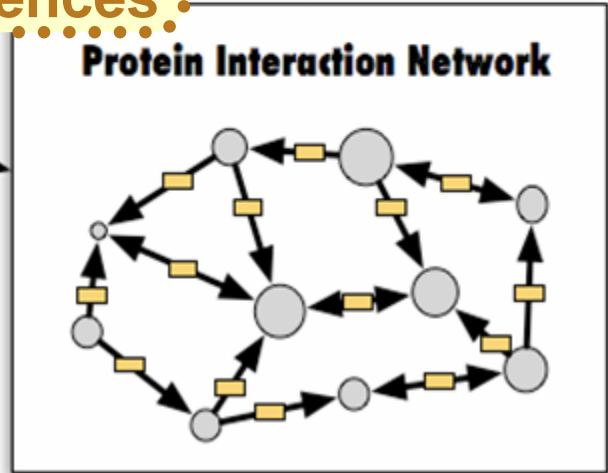


**IMS: Expt Methods**

**IAS: Article Selection**

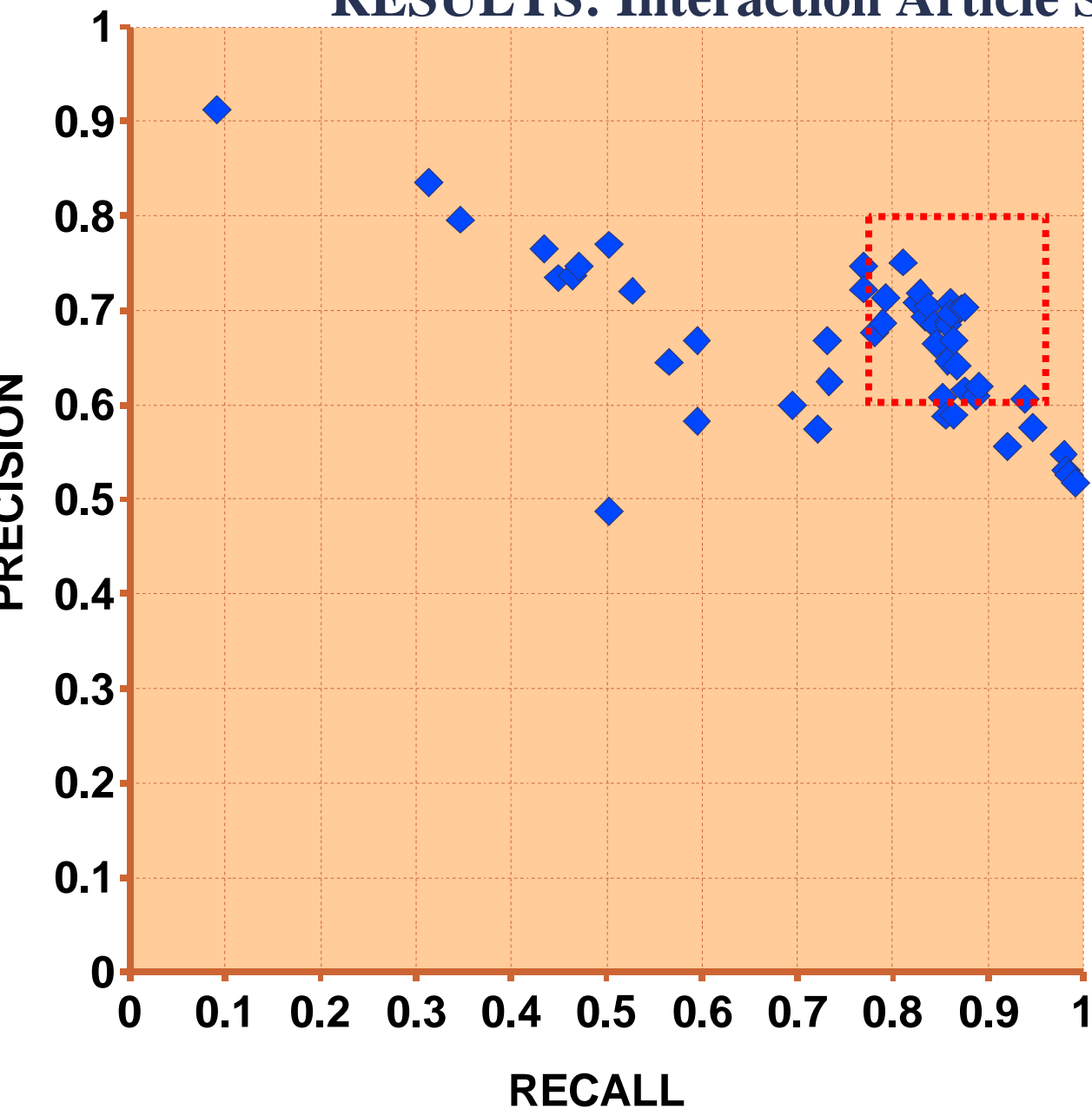


**ISS: Interaction Sentences**





## RESULTS: Interaction Article Subtask



**Document Triage for Curation**

**P: 0.6 - 0.8**

**R: 0.7 - 0.9**

**Most systems reached high recall (0.8-0.9) with precision 0.6-.07**

**Highest acc: 0.77**

**Best AUC: 0.86**

**Highest F-score 0.78**

Precision (P) =  $TP / (TP + FP)$

Recall (R) =  $TP / (TP + FN)$

F-score =  $2 * P * R / (P + R)$

Accuracy =  $TP + TN / (P + N)$

AUC = area under ROC curve

## Interaction Pairs Subtask

**Aim: Identify pairs of interacting proteins from full text  
Normalize the individual interactor proteins to SwissProt**

**Include only experimentally verified interactions**

**Several relevant aspects to be considered:**

**Full text processing (preprocessing): PDF / HTML?**

**Detection of protein mentions (NER): Protein and NOT genes!**

**Normalization of these protein mentions to SwissProt DB records**

**– May require species identification for correct mapping**

**Determination of whether it is an interactor protein**

**Correct identification of the interaction pair: both partners  
(relation extraction)**

**Determination if it was experimentally characterized in the article  
(qualitative information, annotation, experimental evidence).**

**Provide a ranked list of interaction pairs for each article**

## Interaction Pairs Prediction Example

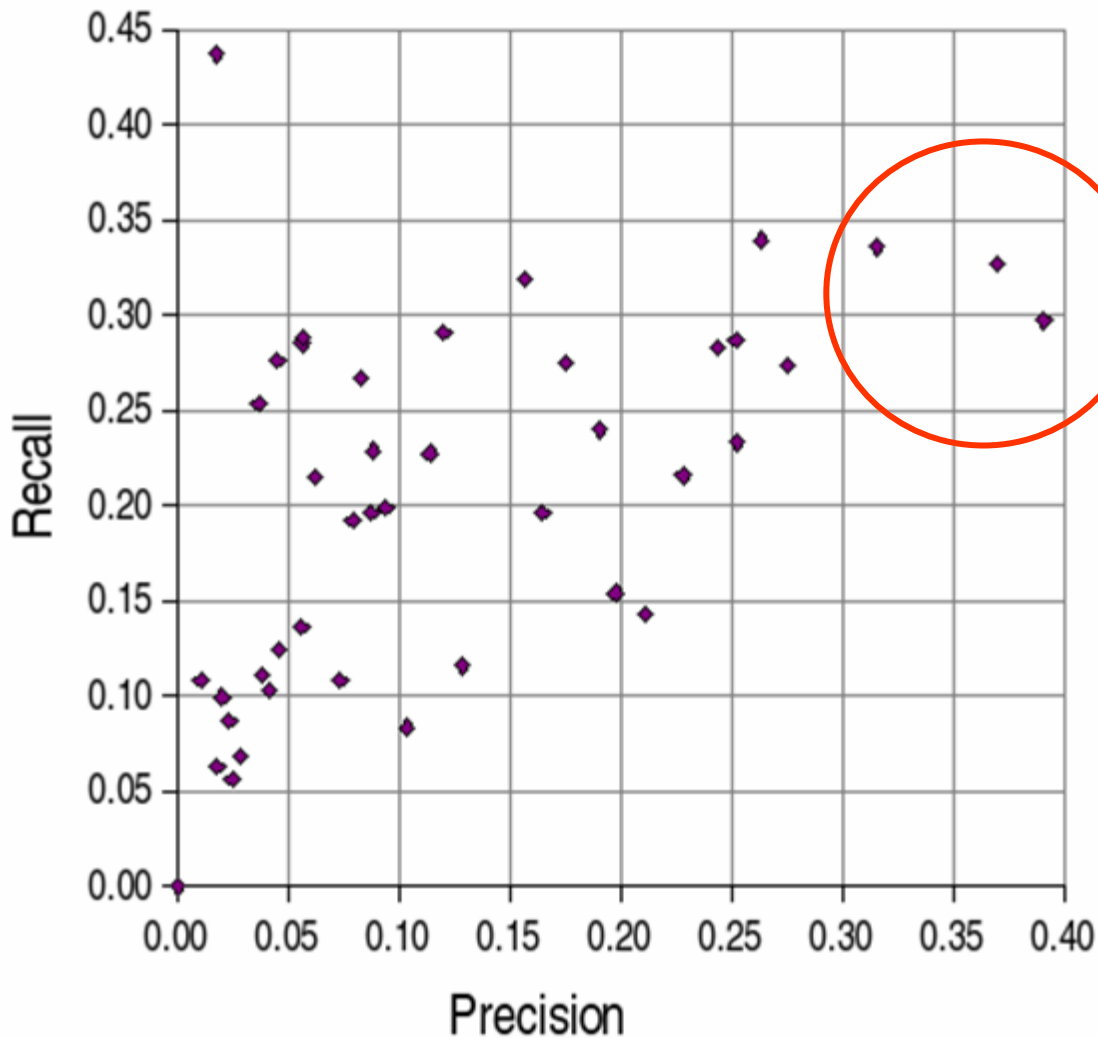
<ENTRY>	Single evidence passage prediction
<PPI_SUB_TASK_ID> BC2_PPI_ISS </>	Task ID, team ID, run, PMID
<TEAM_ID> T1_BC2_PPI </>	
<RUN_NR> 1 </>	
<PMID> <b>10924507</b> </>	
<INTERACTION_PAIR>	Interaction Pair (UniProt IDs)
<INTERACTOR_1> <b>DHX9_HUMAN</b> </>	ATP-dependent RNA helicase A DHX9
<INTERACTOR_2> <b>NXF1_HUMAN</b> </>	Tip-associating protein, TAP
</INTERACTION_PAIR>	
<SENTENCE_RANK> 1 </>	Sentence rank
<SENTENCE_PASSAGE>	Interaction Sentence (from TITLE)
<b>Specific Interaction between RNA Helicase A and Tap, Two Cellular Proteins That Bind to the Constitutive Transport Element of Type D Retrovirus</b>	
</SENTENCE_PASSAGE>	
</ENTRY>	

**NOTE:**

Interaction\_pair input is for the IPS (interaction pairs) subtask;

Sentence\_passage input (and sentence\_rank) are for the ISS (interaction sentence) subtask

## SwissProt-only set



SwissProt-only results:  
Articles where  
all interacting proteins  
had SwissProt IDs

Best results:

R: 30-40%

P: 30-35%

Results for whole data  
(including TrEMBL IDs)  
somewhat lower

This task requires  
getting 3 things right:

- 2 normalized  
protein IDs

- Experimentally  
verified interaction

## Why Protein Interaction Is Hard

- Proteins often have ambiguous names/symbols, or unclear species or isoform information
  - Some proteins are not in SwissProt, so TrEMBL IDs (less well curated) have to be used
  - Must normalize proteins over multiple species
  - ~10% of articles had to be discarded because of inability of humans to disambiguate the protein described (12 of 131 articles from IntAct)\*
  - Homodimer interaction types
- Some articles described in vitro interactions between proteins from different organisms
  - DAG1\_BOVIN - LAMA1\_MOUSE
  - IKKB\_MOUSE - RB\_HUMAN

\*Khadake, Aranda, Derow, Huntley, Kerrien, Leroy, Orchard, Apweiler, Hermjakob, "IntAct – Serving the text mining community with high quality molecular interaction data," Proc. 2<sup>nd</sup> BioCreative Challenge Evaluation Workshop, Madrid 2007.

## Analysis of Protein Normalization

- Protein normalization is more complex than BioCreative II gene normalization (F-score: 0.81)
  - Multiple species, proteins harder than genes
- To quantify performance for protein normalization
  - Collapse all interaction pairs into a list of unique interacting proteins
  - Compare gold standard list vs. system list
- Results (highest average F-score):
  - SwissProt only subset: 0.52 F-score
  - All proteins 0.48 F-score
- Conclusion
  - Protein normalization is harder than genes
  - Good candidate task for next BioCreative

## Interaction Method Subtask

### Articles

structural features that distinguish CRC from CaM and other typical EF-hand calcium sensor proteins. To test the proposal that it serves as a calcium sensor, titrations of CRC-N with the seventh centrin-binding repeat of Sfi1 were performed, using intrinsic tryptophan fluorescence and NMR spectroscopy to characterize the interaction.

**EXPERIMENTAL PROCEDURES**

Recombinant *C. reinhardtii* centrin N-terminal domain was expressed and purified as described elsewhere (19). The 96-residue construct used in this study consists of residues Met<sup>1</sup> through Met<sup>96</sup>, with an additional Gly-Ser sequence at the N terminus left after cleavage of the His<sub>6</sub> tag.

A 24-residue peptide (IVSLKEANLVKRIFHSWKKLLYID) including the seventh centrin-binding repeat of Sfi1 (underlined) was synthesized by Sigma Genosys and further purified by high performance liquid chromatography.

**Fluorescence Spectroscopy**—All fluorescence experiments were performed on a Spex Fluorolog 1681 fluorimeter (Spex Industries Inc., Edison, NJ) at 20 °C. The excitation wavelength was 285 nm, with slit width set to 2.0 nm. Small aliquots of appropriate dilutions of a 1 mM CRC-N stock solution containing 150 mM KCl and 25 mM Tris at pH 7.1 were added to a 5 μM (initial concentration) Sfi1 peptide solution under identical conditions, then incubated with 1 mM EDTA or 5 mM Ca<sup>2+</sup>. Corrections for background fluorescence were made by subtracting the spectra from identical solutions without peptide.

**NMR Spectroscopy**—NMR data were acquired on five different samples of CRC-N with the following isotopic compositions: unlabeled; U-<sup>15</sup>N; U-<sup>13</sup>C; U-<sup>15</sup>N, <sup>13</sup>C; and 10% <sup>13</sup>C. Buffers contained either 10% or 100% <sup>2</sup>H<sub>2</sub>O as appropriate. Each sample typically had a protein concen-

### PSI-MI 2.5 Ontology

- interaction detection method
  - experimental interaction detection
    - biophysical
      - circular dichroism
      - electron resonance
      - fluorescence technology
        - bioluminescence resonance energy transfer
        - classical fluorescence spectroscopy
        - fluorescence correlation spectroscopy
        - fluorescence polarization spectroscopy
        - fluorescence-activated cell sorting
        - fluorescent resonance energy transfer
        - homogeneous time resolved fluorescence
      - isothermal titration calorimetry
      - light scattering
      - mass spectrometry studies of complexes
      - molecular sieving
      - nuclear magnetic resonance
      - scintillation proximity assay
      - surface plasmon resonance
      - x-ray crystallography
    - protein complementation assay
    - genetic interference
    - post transcriptional interference
    - biochemical
    - imaging techniques

PMID:16317001

MI:0017

## Interaction Method Results

### Exact Matching

TEAM ID	RUN	Precision_Av	Recall_Av	F-Score_Av
T14_BC2_PPI	1	0.3628	0.2172	0.2513
T14_BC2_PPI	2	0.3186	0.1980	0.2249
T14_BC2_PPI	3	0.3348	0.1938	0.2265
T40_BC2_PPI	1	0.6679	0.3383	0.4207
T40_BC2_PPI	2	0.4028	0.5548	0.4363
T40_BC2_PPI	3	0.5068	0.5222	0.4836

Only 2 groups participated

Simplified task:  
map experimental method to  
Molecular Interaction  
Ontology at whole document level

### Parent Matching

TEAM ID	RUN	Precision_Av	Recall_Av	F-Score_Av
T14_BC2_PPI	1	0.4986	0.3078	0.3495
T14_BC2_PPI	2	0.4471	0.2847	0.3170
T14_BC2_PPI	3	0.4881	0.2953	0.3375
T40_BC2_PPI	1	0.6794	0.3472	0.4302
T40_BC2_PPI	2	0.5899	0.8548	0.6519
T40_BC2_PPI	3	0.6541	0.7093	0.6375

Results are promising!

Exact match  
Best F: 0.48

Allowing parent matching

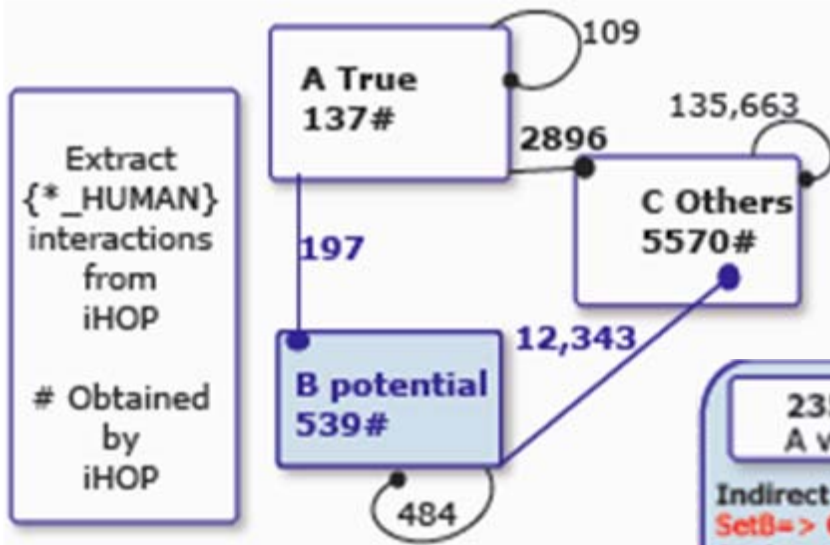
Best F: 0.65



# COMBINATION OF METHODS AS SPINDLE PREDICTORS: A USE CASE from EU ENFIN

1. Semantic Similarity calculations between the GO terms of known and putative spindle proteins
2. Fused Domain architectures between known and putative human spindle proteins.
3. Domain Over-Representation: characteristic spindle domains in target proteins
4. Inherited physical interaction data
- 5. iHOP: Using genes and proteins as hyperlinks between sentences and abstracts in PubMed  
(*iHOP protein-protein direct association & path association scores*)**
- 6. Vector Machine  
(*VM score for proteins; VM score for papers (pmid)*)**
7. Data driven Machine-Learning method based on artificial Neural Network.

# iHOP Based Hypothesis Exploration



99 B proteins with at least 1 direct interaction.

Direct interaction

SetB=> Gen P08670 = VIM  
SetA-> Gen Q96GD4 = AIM1, AURKB, STK12, ARK2, AIK2

PubmedID 14722118  
<Sentence>The activation of Aurora-B spatio-temporally **co-related** with the site-specific phosphorylation of its physiological substrates, histone H3 and vimentin.</Sentence>

235 B proteins with at least 1 indirect interaction with A via B or via C.

Indirect interaction via C

SetB=> Gen P08670 = VIM  
SetC=> Gen Q02962 = PAX2

12138259 <Sentence>The PAX2 signal **co-localized** more with cytokeratin staining than with vimentin.</Sentence>  
12138259 <Sentence>We have investigated PAX2 expression with cytokeratin and/or vimentin in 17 biopsies of juvenile nephronophthisis (NPH), an autosomal-recessive disease characterized by diffuse renal fibrosis and occasional cysts.</Sentence>  
<Sentence>However, podocytes in partially sclerotic glomeruli that still expressed WT1 at high levels showed **reduced** vimentin expression, **cell cycle re-entry**, and re-expressed desmin, cytokeratin and Pax-2.</Sentence>  
14613807 <Sentence>Purified cells were immunostained with astrocyte markers: GFAP, vimentin, Pax2, A2B5, nestin and NCAM.</Sentence>

*iHOP: Information Hyperlinked over Proteins*

<http://www.ihop-net.org/UniPub/iHOP/>

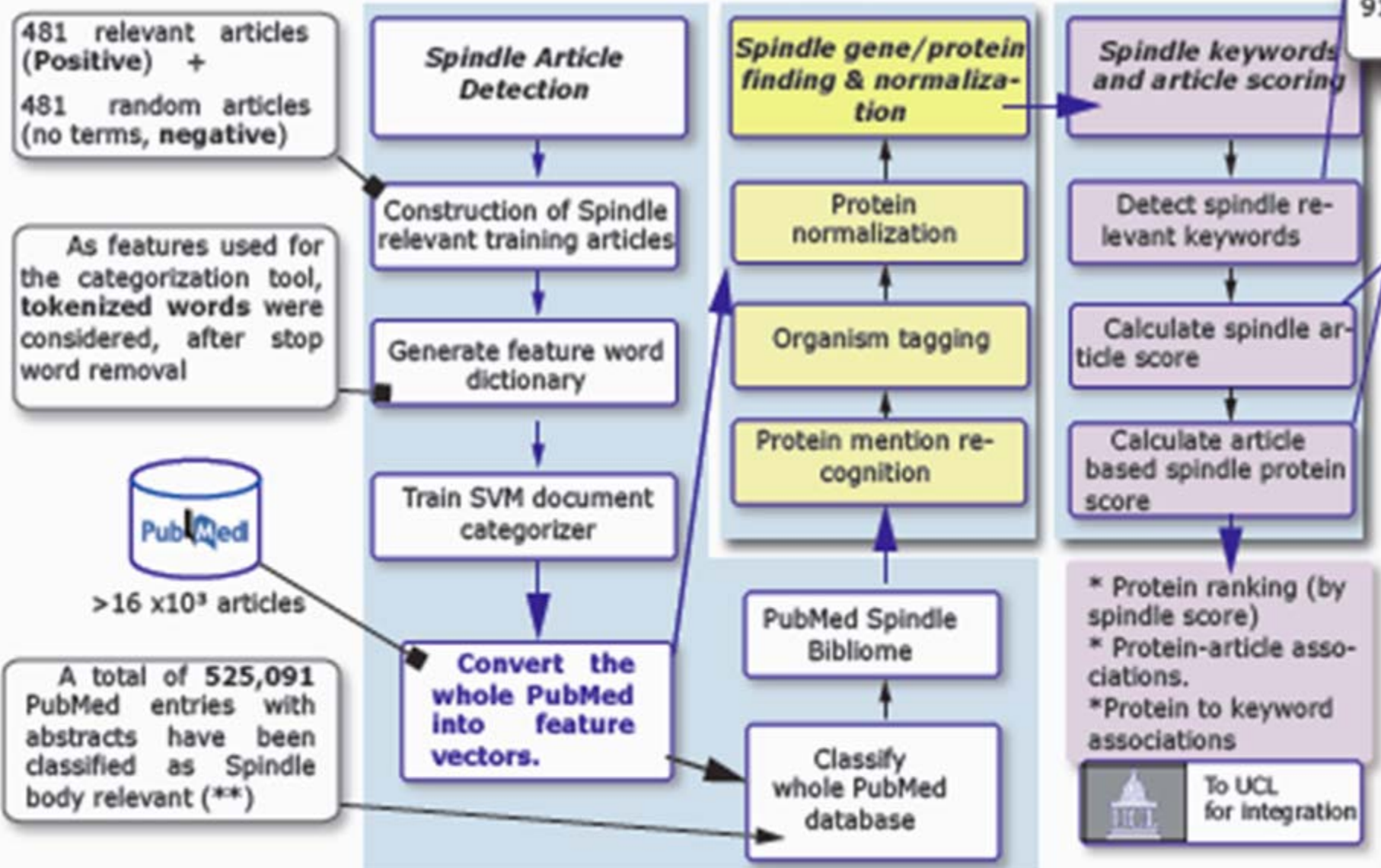
*Identified proteins that co-occur w known spindle proteins*

Hoffmann Valencia Nat Genet 2004

Hoffmann Valencia Bioinformatics 2005

Fernandez-Gonzalez et al., submitted





Identification of new candidate spindle proteins, based on classification of text for information on spindle proteins; Now being used to nominate proteins for further expts

Im  
-Hi  
-In  
-In  
-In  
-Re  
-C  
-Fold cross validation.

Some  
-Small  
-Missi  
-The S  
(Preci

# Conclusions

- BioCreative II a major success
  - More participants (44 teams vs. 27 for BioCreative I)
  - Active participation from potential users (IntAct, MINT database curation teams)
  - Special issue of Genome Biology is in preparation
  - Distributed Annotation Server web site at CNIO
- Field is progressing
  - Significant progress in Gene Mention task  
Combined system scores at over 0.90 F-score
  - Progress in single organism Gene Normalization  
Combined system score estimated ~0.9 F-score
  - Tools being incorporated into end-to-end applications (e.g., ENFIN spindle protein example)
  - Curators are eager to have text mining tools

# Open Questions

- Are text mining tools good enough to help biologists?
  - Tools/services will be made available using DAS model (Distributed Annotation Server)
  - Can future BioCreative measure utility to end user?
- What new application areas to focus on?
  - Species independent protein normalization
  - New entity types: tissues, organisms, drugs, chemicals
  - Text mining embedded in complex workflow

\*PSB 2008 has a text mining session: Translating Biology: Tools That Work to address these issues

# Choosing New Tasks: The 5 BioCreative Criteria

- Real users
- Real data
- Real uses
- Common standards for representation
- Availability of data, results, tools

If you have data,  
see me!

## The rationale:

- If we can connect to the right users,
- They are already preparing the data
- They are willing to share the data,
- In exchange for getting developers to help them!
- So the data and results must be made available
- Standards enable developers to reuse systems

# The Big Challenge: Make Text Mining Invisible

- Biologists want to easy-to-use tools
  - They don't want to learn computational linguistics
- Each application is unique
  - There is no “out of the box” solution
- This means text mining tools must be
  - Easy to tailor
  - Easy to embed
  - Intuitive to use
- So everyone knows how to use them without needing to understand how they work (like BLAST)

# The Bigger Challenge

## Avoid Text Mining!

- Right now, information is lost at each step
  - In laboratory
  - In writing the article
  - In publication
    - Articles not accessible
    - Failure to index key concepts
    - Loss of info in supplementary material
  - In deposition to biological databases
  - In searching databases
- Solution
  - Capture information at point of creation
  - In structured form, to support indexing, search

# The “Meta-Data Checker”

- Provide a free open-source *meta-data checker*
  - Let authors capture meta-data WHILE WRITING
  - Related to discussion about structured digital abstract (Gerstein et al<sup>1</sup>, and Hahn et al<sup>2</sup>)
- Meta-data checker can interact with authors to get
  - Normalized gene, protein, tissues,...
  - Experimental procedures, environment information
  - Other Minimal Information (MIBBI<sup>\*\*</sup>) checklists
- This will require large scale community buy-in
  - From scientists, publishers, curators

*\*New NSF grant: Mining Metadata for Metagenomics to Hirschman at MITRE*

*\*\*MIBBI: Minimal Information for Biological and Biomedical Investigations; see [mibbi.sourceforge.net](http://mibbi.sourceforge.net)*

<sup>1</sup>*Nature* **447**, 142 (10 May 2007); <sup>2</sup>*Nature* **448**, 130 (12 July 2007)

# Meta-Data Checking and Text Mining

- Hypothesis
  - Current text mining tools are good enough for an interactive “metadata checker”
  - To flag entity mentions and suggest mappings to unique ID
- Requirements
  - Accepted standard controlled vocabulary or ontology for terms exists
  - Checker is usable and useful
    - Costs minimal time to the author
    - Doesn't cost time to the publisher or reviewer
    - Speeds access to article
- A possible BioCreative challenge??