

Ontology-based Concept Recognition for Protein-Protein Interaction Extraction

Andrew Fox¹, William Baumgartner Jr.², Helen Johnson²,
Larry Hunter², Donna Slonim¹

¹ Department of Computer Science, Tufts University

² Center for Computational Pharmacology, University of Colorado School of Medicine

Motivation

- PubMed: 800K/yr, ~20M articles
 - LTP, high quality PPIs
 - Manual curation slow
- Automated extraction
 - Tag PPI partners in natural language biomedical text
 - BioCreAtIvE-2 Interaction Pair Subtask
 - Best F-measure < 0.3

Our Task

Infer PPIs from natural-language biomedical text

- **GeneRIFs** (references into function)

Why GeneRIFs?

- Short (< 256 characters)
- Constrained domain of discourse
- Linked to relevant EntrezGene IDs
 - *aid protein identification*

- **Goals**

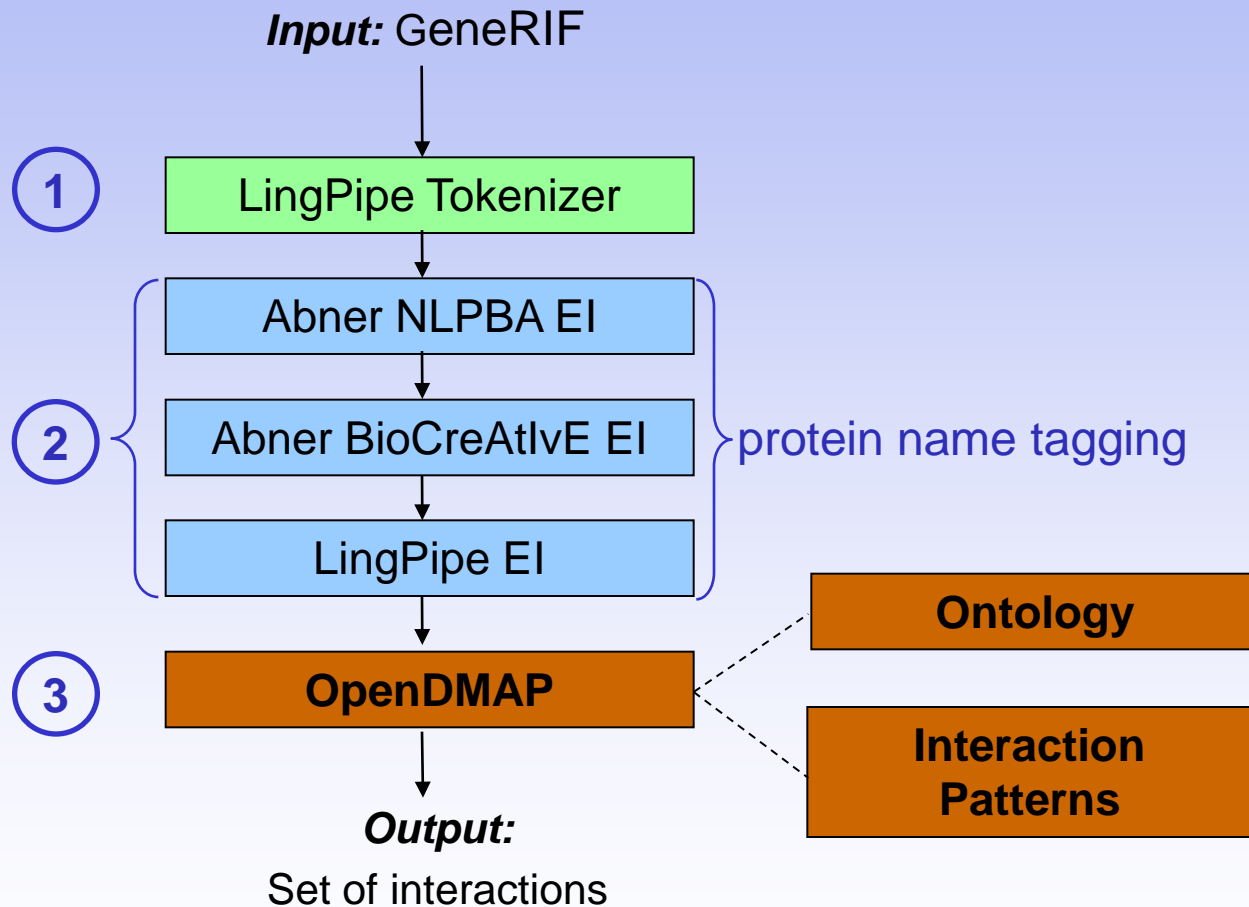
- *Evaluate OpenDMAP's performance on GeneRIFs*
- *Improve performance*

Train/Test Set

- Manually annotated geneRIFs
 - Yeast, Mouse, Fruit Fly
 - Inter-annotator agreement > 90%
 - available at bcb.cs.tufts.edu/GeneRIFs
- ~ 75% linked to 2 or more genes
~ 25% linked to only 1 gene

	# GeneRIFs	# Gene Pairs
Mouse Train	35	85
Mouse Test	37	96
Fly Train	67	137
Fly Test	68	141
Yeast Train	112	203
Yeast Test	111	165
TOTAL	430	827

Original OpenDMAP System



OpenDMAP Patterns

Base Concepts:

{interaction-verb} := interact, interacts, bind, ...

{interaction-noun} := **interaction**, phosphorylation, ...

{preposition} := among, between, by, of, with, ...

Protein Interaction Patterns:

{interaction} := {interaction-noun} involving the? [interactor1] and the? [interactor2]

Matches: “**Interaction** involving the **X protein** and **proteinY**”

(? = optional token)

Tagged
Proteins

System Modifications

- Metadata Tagger
- Protein Complex Recognition
- Ternary Interactions
- Pattern-set enhancements

Metadata Protein Tagger

- Tag aliases of metadata-linked genes

Example:

“Dysfusion dimerizes with the Tango bHLH-PAS protein....”



Gold standard protein tags

Metadata Protein Tagger

- Tag aliases of metadata-linked genes

Example:

“Dysfusion dimerizes with the Tango bHLH-PAS protein....”



Without metadata tagger

Metadata Protein Tagger

- Tag aliases of metadata-linked genes

Example: {Metadata: 43174; 41084}

“Dysfusion dimerizes with the Tango bHLH-PAS protein....”

Metadata Protein Tagger

- Tag aliases of metadata-linked genes

Example: {Metadata: 43174; 41084}

“Dysfusion dimerizes with the Tango ...”

| | | | |
[interactor1] {interact-verb} {prep} the? [interactor2]

Protein-Complex Recognition

{interaction} := [interactor1] [-/] [interactor2]

- ‘prot1-prot2’ single token → 3 tokens
 - split “prot1-prot2 complex” if one or both prots tagged
 - split “prot1-prot2” if both prots tagged

- Ternary support added

Pattern Enhancements

- Training set performance
 - Add, remove or modify patterns to reduce classification error
 - Added grammatical concepts to ontology
 - {article} := a, an, the, ...
 - {preposition} := by, of, with, to, ...

Test Set Results

	Recall	Precision	F-Measure
Mouse (orig)	0.14	0.56	0.22
Mouse (-meta)	0.38	0.50	0.43
Mouse (+meta)	0.38	0.54	0.44
Fly (orig)	0.07	0.40	0.12
Fly (-meta)	0.17	0.38	0.24
Fly (+meta)	0.29	0.46	0.36
Yeast (orig)	0.27	0.50	0.35
Yeast (-meta)	0.55	0.62	0.59
Yeast (+meta)	0.56	0.63	0.60
Combined(orig)	0.15	0.50	0.23
Combined(-meta)	0.36	0.53	0.43
Combined(+meta)	0.41	0.55	0.47

26 point improvement in Recall

5 point improvement in Precision

24 point improvement F-Measure

Test Set Results

	Recall	Precision	F-Measure
Mouse (orig)	0.14	0.56	0.22
Mouse (-meta)	0.38	0.50	0.43
Mouse (+meta)	0.38	0.54	0.44
Fly (orig)	0.07	0.40	0.12
Fly (-meta)	0.17	0.38	0.24
Fly (+meta)	0.29	0.46	0.36
Yeast (orig)	0.27	0.50	0.35
Yeast (-meta)	0.55	0.62	0.59
Yeast (+meta)	0.56	0.63	0.60
Combined(orig)	0.15	0.50	0.23
Combined(-meta)	0.36	0.53	0.43
Combined(+meta)	0.41	0.55	0.47

- Species-specific variation shows importance of entity identification
- Metadata tagger contributes primarily in fruit fly

Future Work

- Further exploit metadata
- More Protein Taggers
- Explore Deeper Parsing

Acknowledgements

- William Baumgartner
- Helen Johnson
- Kevin Cohen
- Larry Hunter
- Donna Slonim