Knowledge-Base-Enriched Relation Extraction

BioCreative VI, Task 5

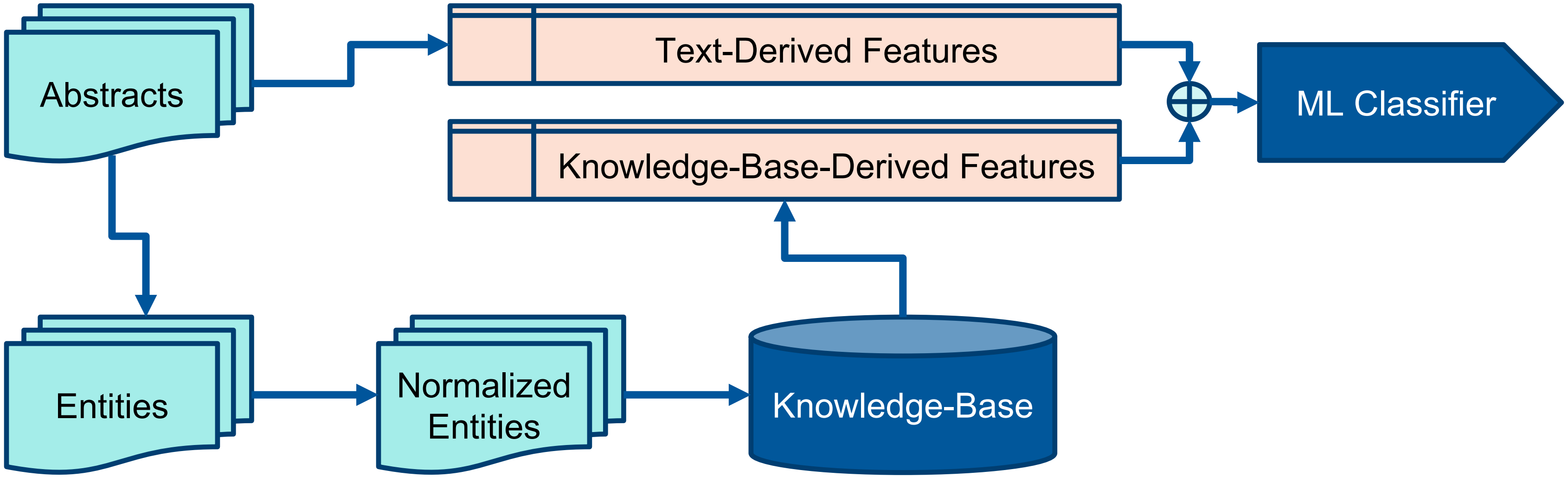
Ignacio Tripodi[1], M. Boguslav[2], N. Hailu[2], L. E. Hunter[2]

(1) University of Colorado, Boulder, BioFrontiers Institute

(2) University of Colorado, Denver, Computational Bioscience, Anschutz Medical Campus

Bethesda, MD 10/18/17

Mixed Approach



Identified potential relations

- Assumed any chemicals and proteins/genes in the same sentence as potentially related
- Chemical/protein combinations not addressed by given relations labeled as "NONE"
- Relations outside of task scope pooled as "OTHER"

Identified potential relations

… cetuximab (IMC-225, Erbitux). Agents that have only begun to undergo clinical evaluation include CI-1033, an irreversible pan-erbB tyrosine kinase inhibitor, and PKI166 and GW572016, both examples of dual kinase inhibitors (inhibiting epidermal growth factor receptor and Her2). Preclinical models have demonstrated...

PKI166	erbB	NONE
PKI166	tyrosine kinase	NONE
PKI166	kinase	CPR:4
PKI166	epidermal growth factor receptor	CPR:4
PKI166	Her2	CPR:4
GW572016	erbB	NONE
GW572016	tyrosine kinase	NONE
GW572016	kinase	CPR:4
GW572016	epidermal growth factor receptor	CPR:4
GW572016	Her2	CPR:4

Extracted features from text

- Entity 1
- Entity 2
- Entity 1 + Entity 2
- Text in between entities
- Text along dependency path between entities (unigrams up to trigrams)
- Labels of the dependency path between entities (unigrams up to trigrams)
- Words from entire sentence

* Unigrams and bigrams, otherwise

Knowledge-base Of Biomedicine (KaBOB*)

- AllegroGraph RDF triple store (~280 million triples) based on the Open Biomedical Ontologies
- Integrates 17 public biomedical ontologies
- Data from 13 public databases
- Semantically consistent
- SPARQL Endpoint

* Livingston, et al, *BMC Bioinformatics* , vol. 16, p. 126, Apr. 2015.

Extracted attributes from KaBOB

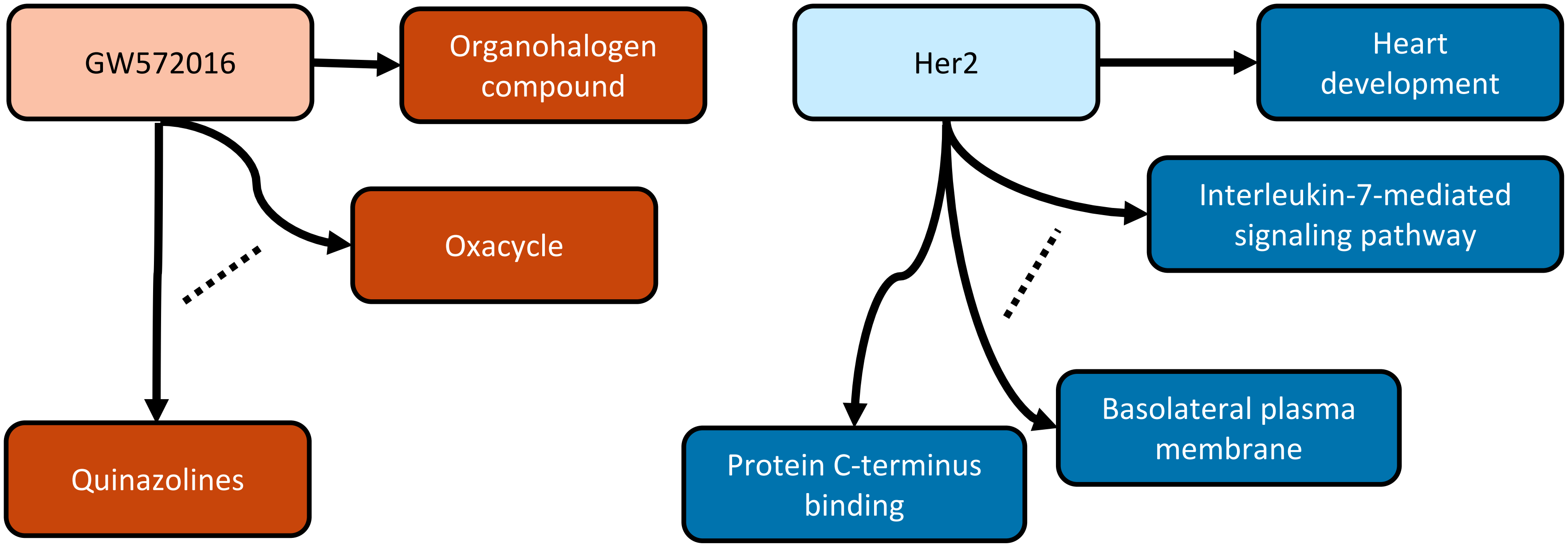
Mapped chemicals to ChEBI IDs

- Queried relevant parent classes

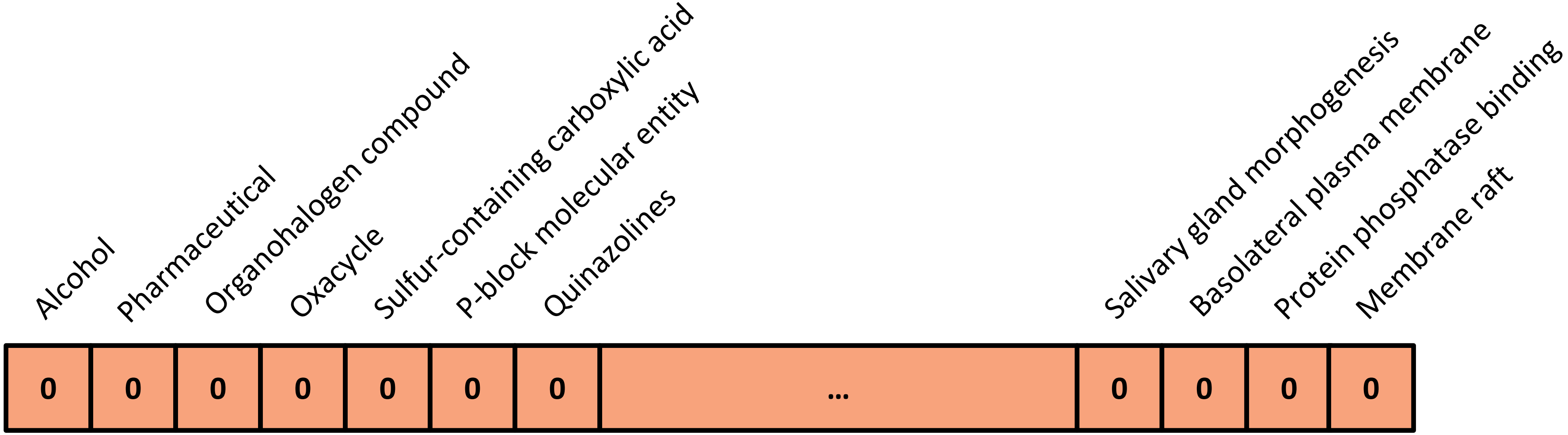
Mapped proteins/genes to PRO IDs

- Queried related biological processes, molecular functions and cellular components

Extracted attributes from KaBOB



Extracted attributes from KaBOB



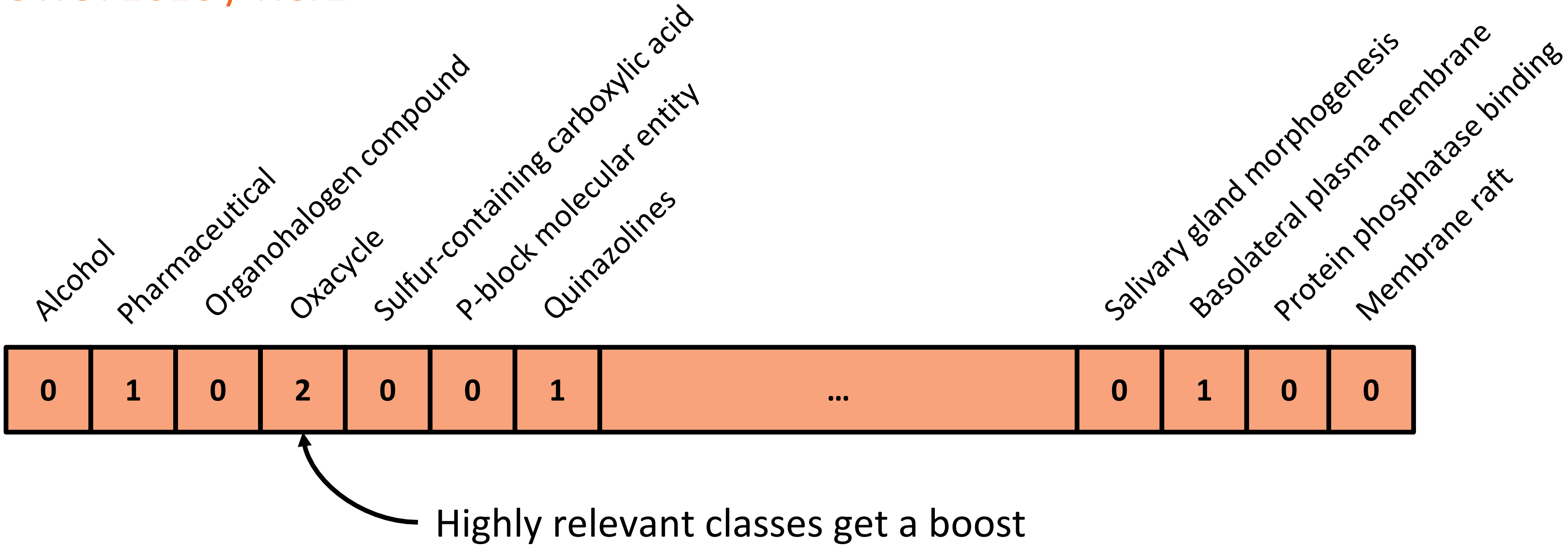
Extracted attributes from KaBOB

GW572016 / Her2

Alcohol	Pharmaceutical	Organohalogen compound	Oxacycle	Sulfur-containing carboxylic acid	P-block molecular entity	Quinazolines		Salivary gland morphogenesis	Basolateral plasma membrane	Protein phosphatase binding	Membrane raft
0	1	0	1	0	0	1	…	0	1	0	0

Extracted attributes from KaBOB

GW572016 / Her2



Tested multiple classifier configurations

- KaBOB features
- Tokens from the entire sentence
- Feature selection of 10k, 20k and 30k top features
- Include trigrams from words in dependency path
- Multiple machine learning classifiers: Random Forests, Perceptron, Naïve Bayes, Neural Networks

Results

Relation	Training Data Performance Metrics		
	Precision	*Recall*	*F-Score*
CPR:3	0.76	0.74	0.75
CPR:4	0.79	0.80	0.80
CPR:5	0.60	0.60	0.60
CPR:6	0.61	0.74	0.67
CPR:9	0.75	0.83	0.79
NONE	0.92	0.91	0.91
OTHER	0.72	0.75	0.73
AVG / TOTAL	0.86	0.86	0.86

Run	Evaluation Performance Metrics		
	Precision	*Recall*	*F-Score*
Run 1	0.3460	0.3913	0.3673
Run 2	0.3387	0.4078	0.3700
Run 3	0.3305	0.1666	0.2215
Run 4	0.3307	0.3641	0.3466
Run 5	0.3058	0.3603	0.3309

Results

- KaBOB features alone not yet enough to classify relations
- Tokens from the entire sentence introduced too much noise
- Using all features was worth the performance improvement over feature selection
- Best training performance accomplished by Neural Networks, Perceptron and Naïve Bayes, in that order

Future steps

- Improve KaBOB queries to extract more features
- Explore most relevant KaBOB features to use as prior probabilities in other applications
- Find top shortest paths between a chemical and drug node in KaBOB
- Better tuning of classifiers
- Discriminate across all relation types (no “OTHER” class)

Acknowledgements

- Mayla Boguslav
- Negacy Hailu
- Mike Bada
- Anis Karimpour-Fard
- Tiffany Callahan
- Larry Hunter
- BioCreative Organizers

Questions?