

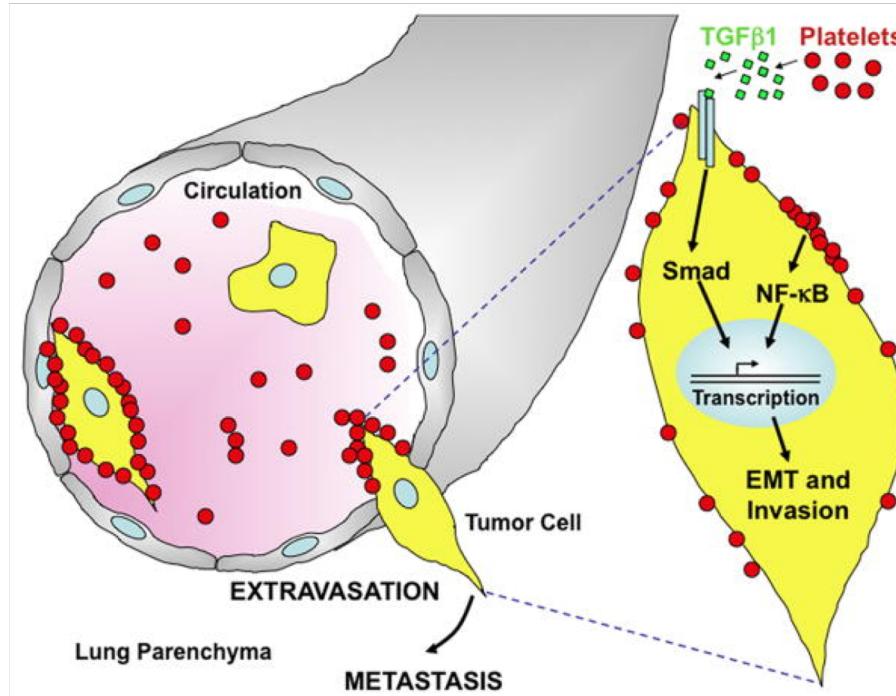
# CIE

## Modeling Cellular Response

Corey O'Connor and Saman Faramand

# Background

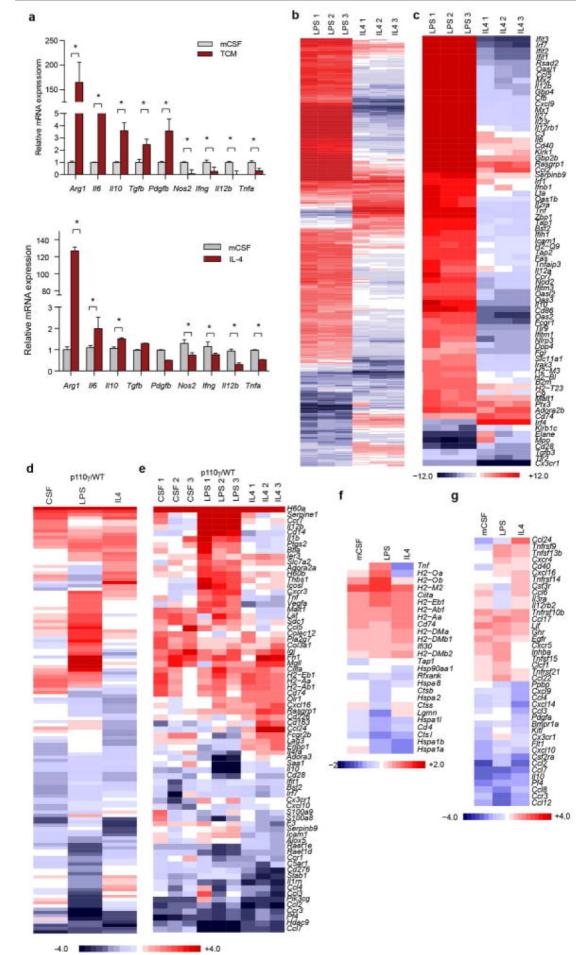
# Changes in Gene Expression



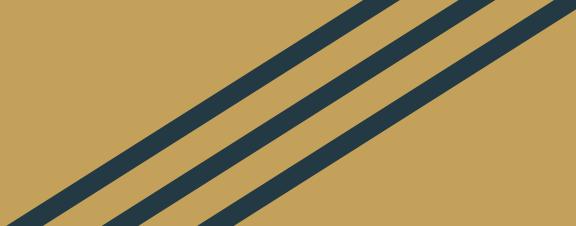
M. Labelle, S. Begum, and R. O. Hynes, "Direct Signaling Between Platelets and Cancer Cells Induces an Epithelial-Mesenchymal-Like Transition and Promotes Metastasis," *Cancer Cell*, vol. 20, no. 5, pp. 576–590, Nov. 2011.

# Evaluating Gene Expression

- Through comparing abundance of mRNA between cell conditions we can find genes which are repressed in one condition but activated in another to a significant extent.
- Based on these values, we can run further forms of data analysis to elucidate cell conditions, like those depicted to the right.



M. M. Kaneda *et al.*, “PI3K $\gamma$  is a molecular switch that controls immune suppression,” *Nature*, vol. 539, no. 7629, pp. 437–442, Nov. 2016.



# Approaches to Modeling Cellular Response



# Ingenuity Pathway Analysis

- Uses their own proprietary database of experimental data, Ingenuity Knowledge Base
- Has four modes of regulation: Upstream Regulator Analysis (URA), Mechanistic Networks (MN), Causal Network Analysis (CNA), and Downstream Effects Analysis (DEA).

# Fisher Test Statistic and p-Value

	Observed r	Observed 0	Total
Predicted r	$n_{rr}$	$n_{r0}$	$q_r$
Predicted 0	$n_{0r}$	$n_{00}$	$q_0$
Total	$n_r$	$n_0$	$N$

**2 × 2 Contingency table:** Tabulation of predictions vs. observations with no causal edges.

$$p = \frac{\binom{n_{rr}+n_{r0}}{n_{rr}} \binom{n_{0r}+n_{00}}{n_{0r}}}{\binom{N}{n_{rr}+n_{0r}}} = \frac{\binom{n_{rr}+n_{r0}}{n_{r0}} \binom{n_{0r}+n_{00}}{n_{00}}}{\binom{N}{n_{r0}+n_{00}}} = \frac{(n_{rr} + n_{r0})!(n_{0r} + (n_{00})!(n_{rr} + n_{0r})!(n_{r0} + n_{00})!}{n_{rr}!n_{r0}!n_{0r}!n_{00}!N!}$$

# Ternary Test Statistic and P-Value

$$p(S) = \sum_{(n_{++}+n_{+-})-(n_{-+}+n_{--})=S} \frac{D[n_{\pm\pm}]}{D_{\text{tot}}},$$

Where  $D[n_{++}, n_{+-}, n_{-+}, n_{--}] =$

$$\binom{q_+}{n_{++}, n_{+-}, n_{+0}} \binom{q_-}{n_{-+}, n_{--}, n_{-0}} \binom{q_0}{n_{0+}, n_{0-}, n_{00}}. \quad \&$$

$$D_{\text{tot}} := \sum_{n_{++}, n_{+-}, n_{-+}, n_{--}} D[n_{\pm\pm}] = \binom{|\mathcal{T}|}{n_+, n_-, n_0}.$$

In this method, the p-value can be calculated by

$$\sum_{((n_{++}+n_{+-})-(n_{-+}+n_{--})=S) \geq S_0} p(S)$$

L. Chindelevitch, P.-R. Loh, A. Enayetallah, B. Berger, and D. Ziemek, "Assessing statistical significance in causal graphs," *BMC Bioinformatics*, vol. 13, p. 35, Feb. 2012.

# Quaternary Test Statistic

	Observed +	Observed -	Observed 0	Total
Predicted +	$n_{++}$	$n_{+-}$	$n_{+0}$	$q_+$
Predicted -	$n_{-+}$	$n_{--}$	$n_{-0}$	$q_-$
Predicted $r$	$n_{r+}$	$n_{r-}$	$n_{r0}$	$q_r$
Predicted 0	$n_{0+}$	$n_{0-}$	$n_{00}$	$q_0$
Total	$n_+$	$n_-$	$n_0$	$N$

**$4 \times 3$  Contingency table:** Tabulation of predictions vs. observations with both causal and ambiguous edges.

Predicted / Actual	+	-	
+	x		5
-			5
	5	5	20

The p value of the contingency table can then be calculated by:

$$\sum_{i \leq x} Pr(x = i) = \frac{\binom{x}{5} \binom{5-x}{x}}{\binom{20}{5}}$$

# Publicly Available Databases

# Characteristics

- Can be experimental, literature, or combination-based.
- Often are a mix of signed and unsigned interactions or entirely the latter.
  - In contrast with proprietary databases where a higher percentage of interactions are signed.

# Databases Used in Causal Inference Engine

Database	Source of Data	Signed
BEL Large Corpus	Textmining [1]	Yes
ChIP Atlas	Experimental [2]	No
STRINGdb	Experimental, human curated databases, textmining, coexpression... [3]	Mix
TRED	Prediction, Databases [4]	No
TRRUST	Textmining [5]	Mix

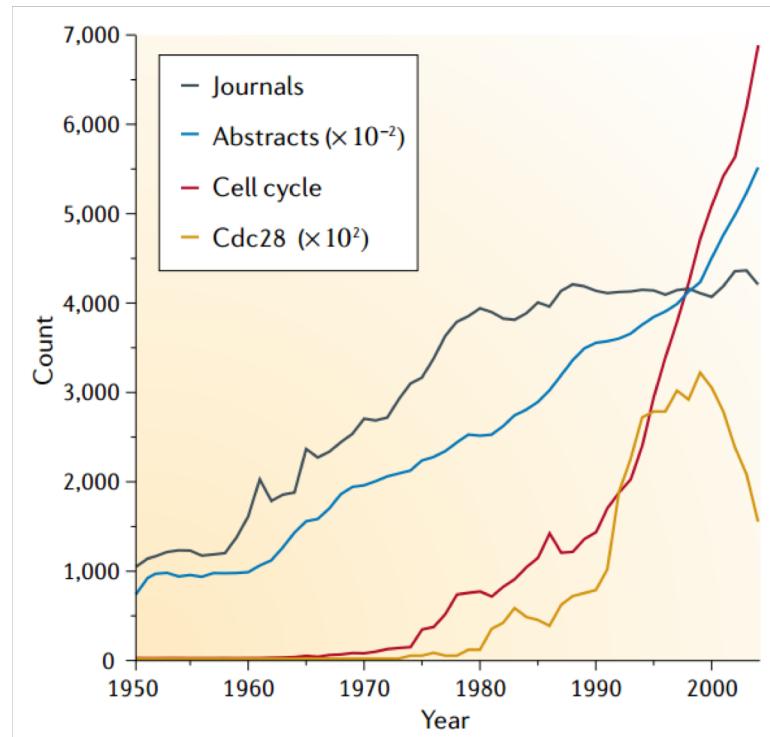
\*Citations at end of presentation

# ChIP Atlas

- Provides us with rich information about the conditions of the interaction (cell line in which it occurred, binding score)
- Is based on ChIP-seq experimental data and is therefore unsigned
- We propose the following to add directionality to ChIP Atlas so it may be used for more accurate methods of causal regulator prediction.

# ChIP-Seq Data Annotation

# Why Natural Language Processing (NLP)?

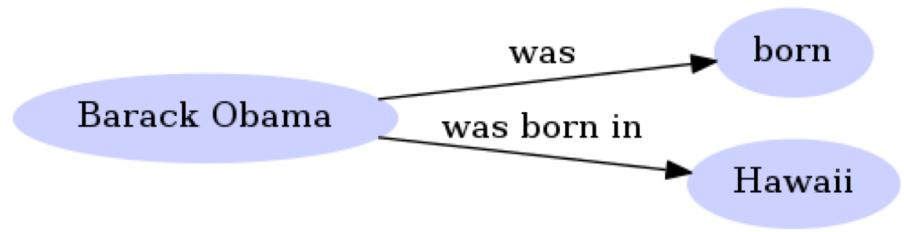


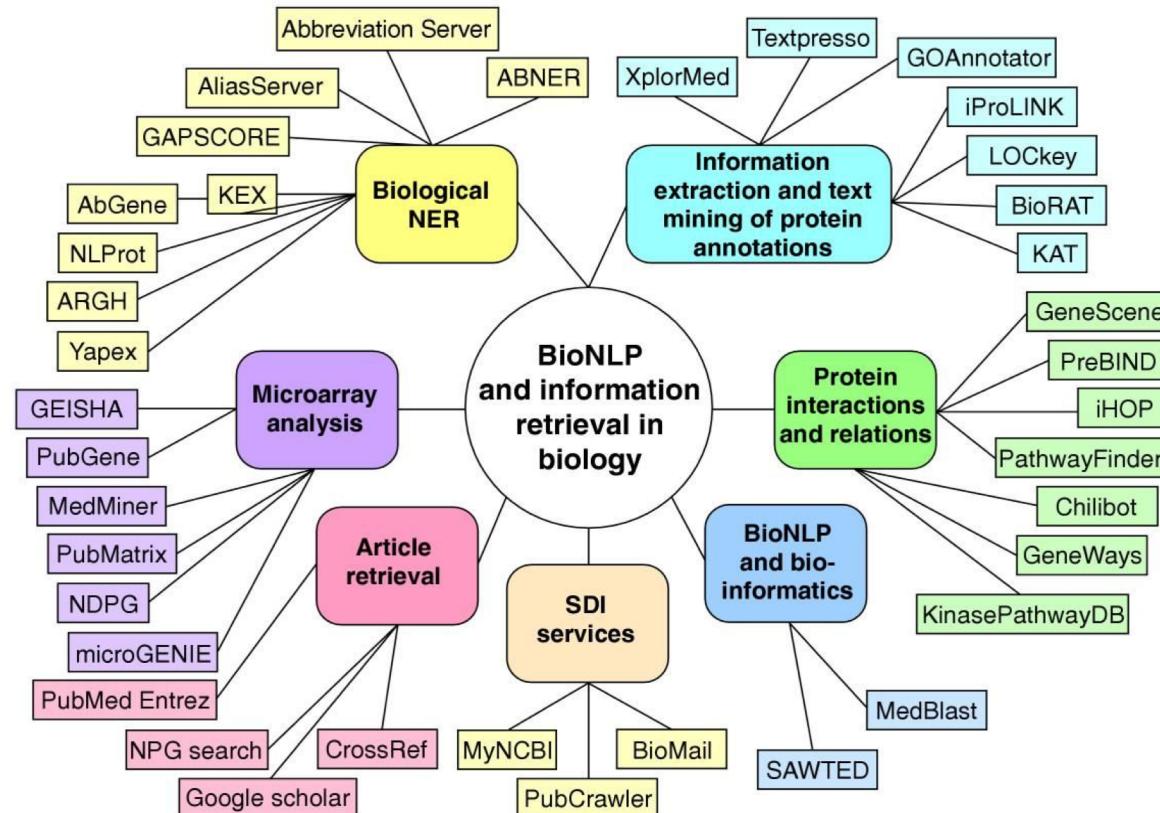
Jenden et al., Literature mining for the biologist: from information retrieval to biological discovery, *Nature Reviews Genetics* volume 7, pages 119–129 (2006)

# What is NLP?



- **Information Retrieval**
  - Representation, Storage and organization of information in databases and repositories and their retrieval according to an information need
  - Format: text, image, audio, video, ...
- **Name Entity Recognition**
  - The identification of entities in free text is known as named-entity recognition
  - Entities: gene, protein, drug, location, organization, person, ...
- **Information Extraction**
  - Identify relation between entities





# BELTracker: evidence sentence retrieval for BEL statements

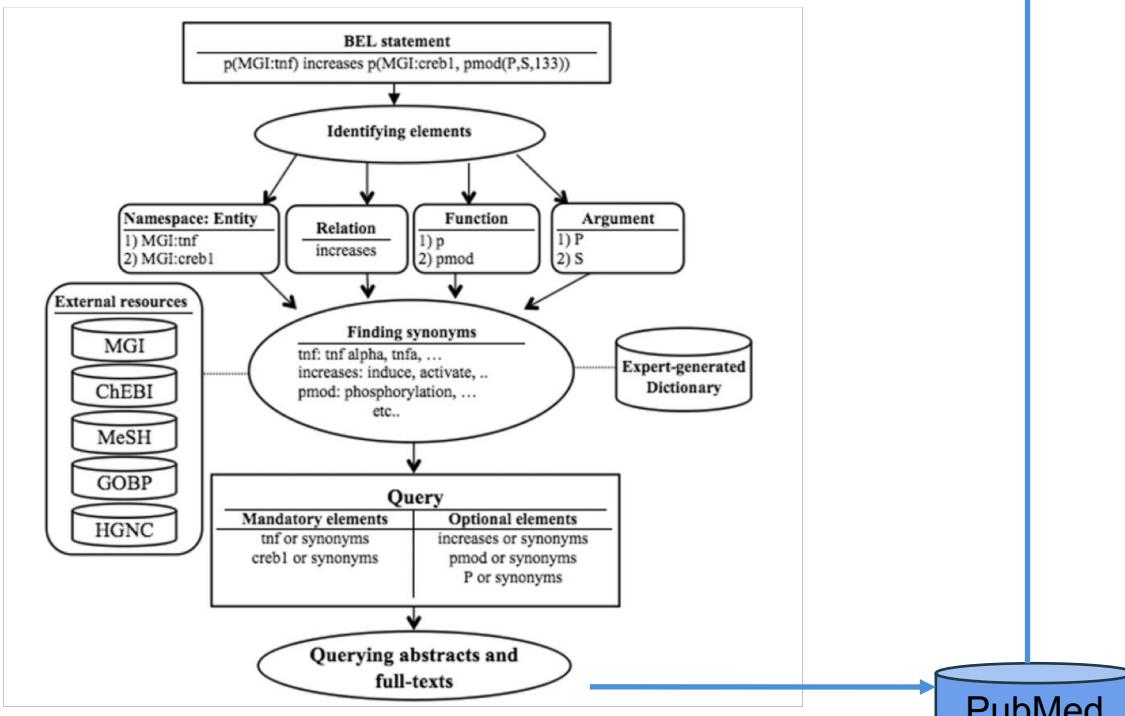
- Biological Expression Language (BEL)
- Representative method that effectively express the semantic of Biological pathway events
- Addressed the BioCreative V challenge
  - Given textual evidence for a BEL statement, generate the corresponding BEL statement (information extraction [IE] task)
  - Given a BEL statement, provide, at most, 10 evidence sentences (information retrieval [IR] task)

Sentences	BEL statement
We showed that HSF 1 is phosphorylated by the protein kinase RSK2 in vitro we demonstrate that RSK2 slightly represses activation of HSF1 in vivo	1: kin (p (HGNC: RPS6KA3)) increases p (HGNC: HSF1, pmod (P)) 2: kin (p (HGNC: RPS6KA3)) decreases tscript (p (HGNC: HSF1))
Whereas exposure of neutrophils to LPS or TNF- $\alpha$ resulted in increased levels of the transcriptionally active serine 133-phosphorylated form of CREB	p(MGI: TNF) increases p (MGI: CREB1, pmod (P, S, 133))
BEL Elements: Relationship, Function, Entity, Namespace, Sequence position	

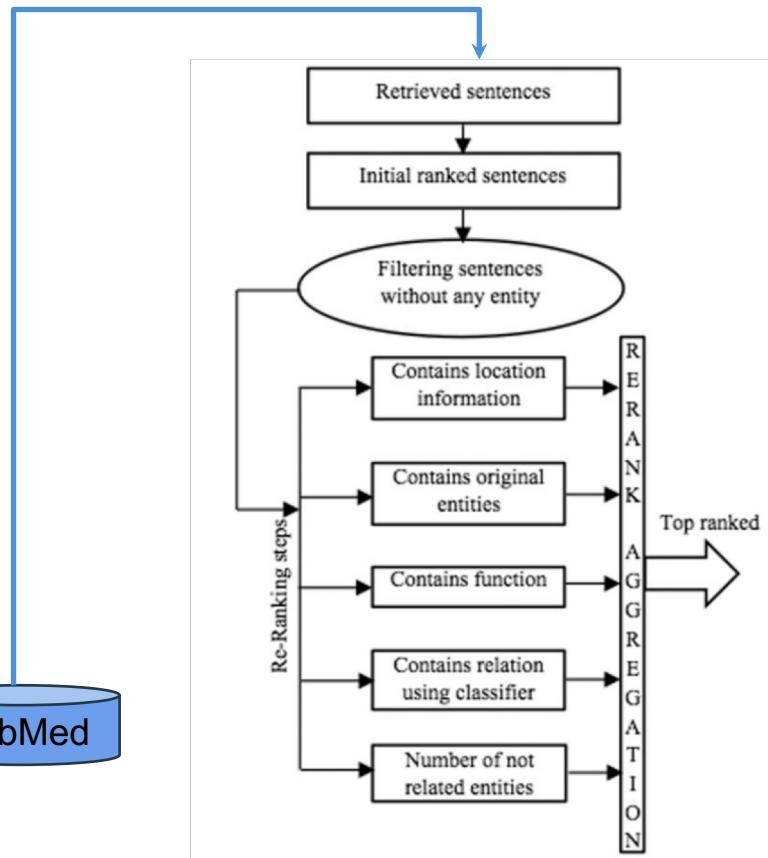
Table 1 shows example BEL statements curated from evidence sentences. Components of a BEL statement are highlighted using different colors.

```
graph LR; A[Input BEL statement] --> B[Query Translation]; B --> C[Retrieval]; C --> D[Ranking]; D --> E[Top ten ranked sentences]
```

# BELTracker: evidence sentence retrieval for BEL statements



Expanded query based on Boolean logic AND/OR (mandatory/optional)

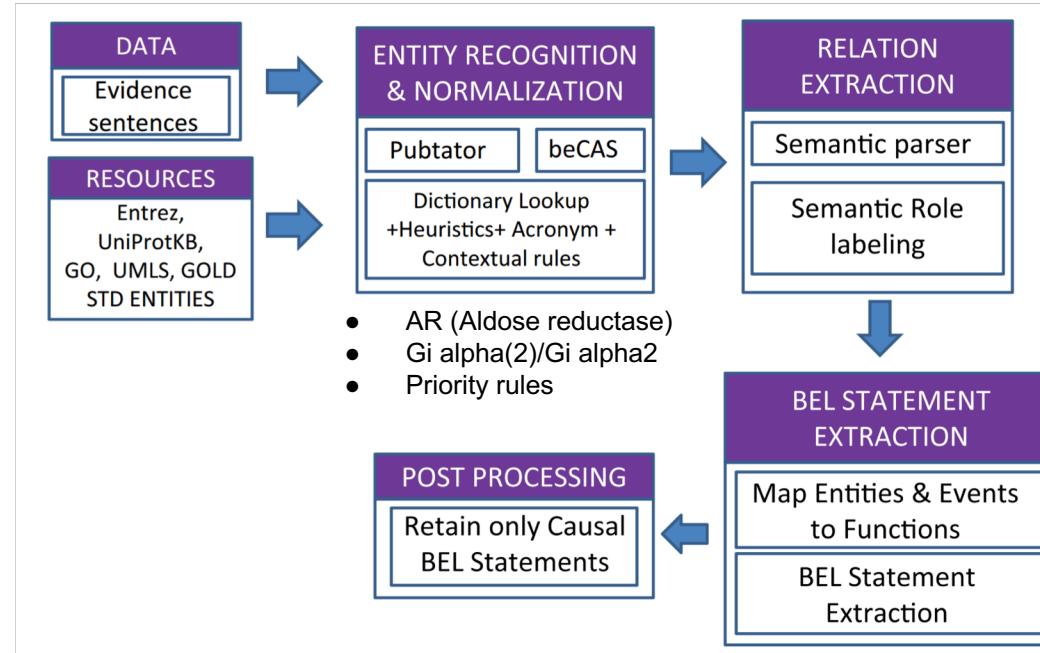


## BELTracker: evidence sentence retrieval for BEL statements

- Limitations
  - Considered lexical features to select sentences
  - Long response time
  - Ranking methods operate at the sentence level

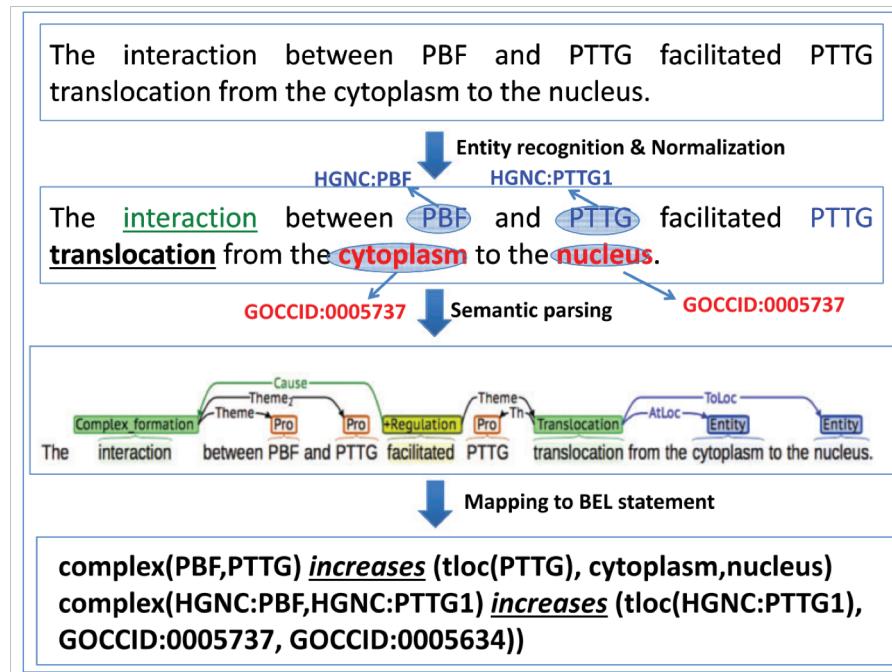
# BELMiner: adapting a rule-based relation extraction system to extract biological expression language statements from biomedical literature evidence sentences

- Addressed the BioCreative V challenge
  - Given textual evidence for a BEL statement, generate the corresponding BEL statement (information extraction [IE] task)



# BELMiner: adapting a rule-based relation extraction system to extract biological expression language statements from biomedical literature evidence sentences

- Addressed the BioCreative V challenge
  - Given textual evidence for a BEL statement, generate the corresponding BEL statement (information extraction [IE] task)



# BELMiner: adapting a rule-based relation extraction system to extract biological expression language statements from biomedical literature evidence sentences

**Table 2.** Verbs for causal relations

S. No	Verb categories	Verbs
1	decreases	reduce, decrease, suppress, block, down-regulate, decrease, down-regulation, inhibit
2	increases	increase, induce, activate, enhance, up-regulate, up-regulation
3	directlyIncreases	increase verbs preceded by an adjective “directly”
4	directlyDecreases	decrease verbs preceded by an adjective “directly”

**Table 3.** Performance of BELMiner on BioCreative BEL task (with and without gold standard entities)

Class		Entities from gold standard			Entities from NER		
		Pre (%)	Rec (%)	F-mes (%)	Pre (%)	Rec (%)	F-mes (%)
Term (T)	Run1	91.8	74.67	82.35	82.03	59.33	68.86
	Run2	92.51	70.00	79.70	83.33	50.00	62.5
FS	Run1	51.47	62.50	56.45	50.77	58.93	54.55
	Run2	51.61	57.14	54.24	54.72	51.79	53.21
Function	Run1	25.53	36.36	30.00	27.78	37.88	32.05
	Run2	27.06	34.85	30.46	30.67	34.85	32.62
Relation-Secondary (RS)	Run1	87.71	77.72	82.41	76.84	67.33	71.77
	Run2	94.38	74.75	83.43	92.37	59.9	72.67
Relation	Run1	77.93	55.94	65.13	69.37	38.12	49.20
	Run2	77.93	55.94	65.13	69.37	38.12	49.20
Statement	Run1	32.09	21.29	25.60	26.42	13.86	18.18
	Run2	32.09	21.29	25.60	26.42	13.86	18.18

Pre, precision; Rec, recall; F-mes, F-measure.

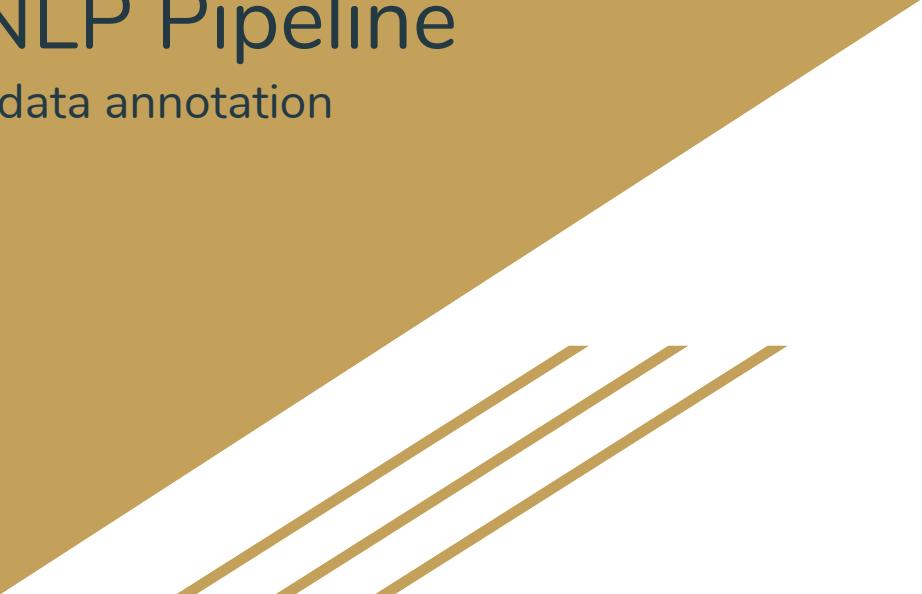
# BELMiner: adapting a rule-based relation extraction system to extract biological expression language statements from biomedical literature evidence sentences

- Limitations
  - Long response time
  - Ranking methods operate at the sentence level

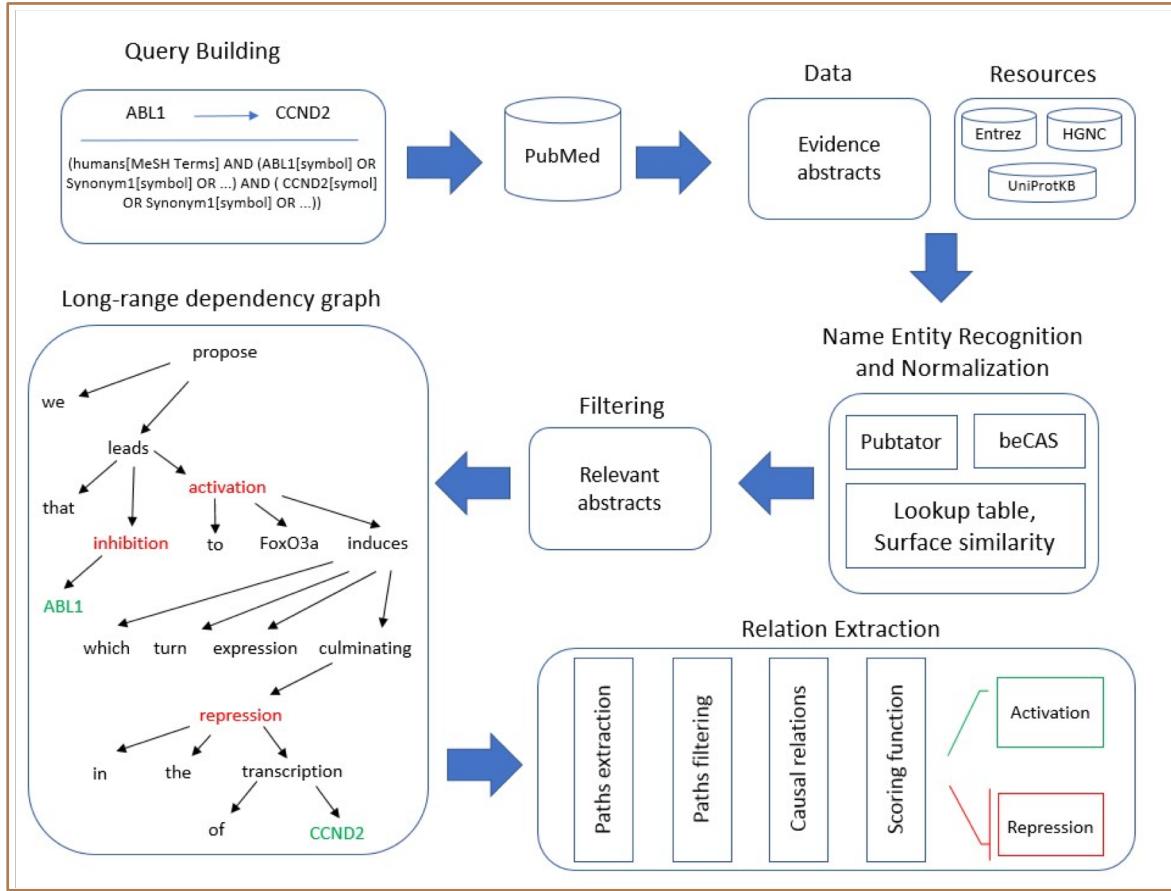


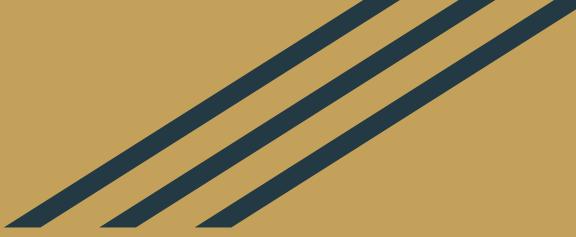
# Proposed NLP Pipeline

for ChIP-Seq data annotation



# Proposed NLP Pipeline



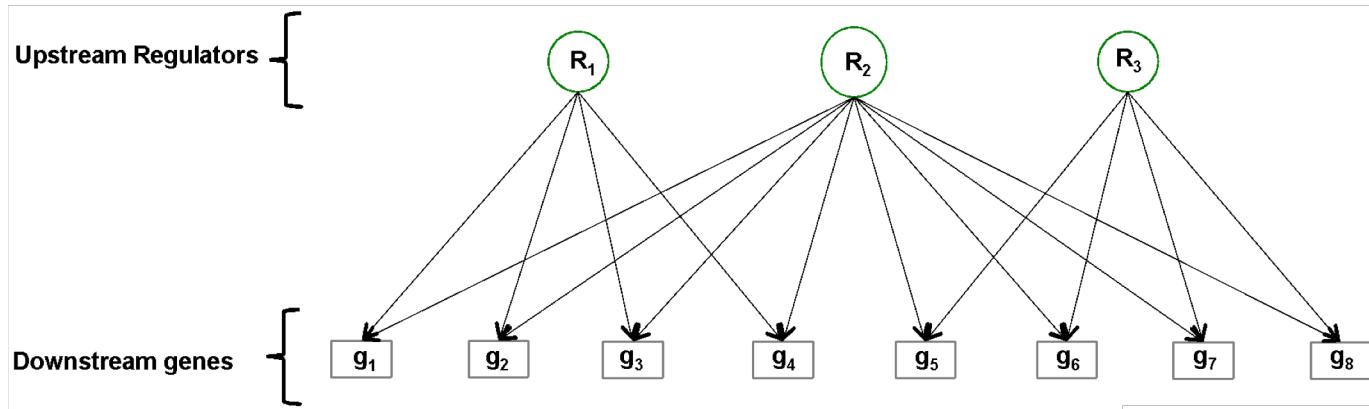


# Statistical Perspective

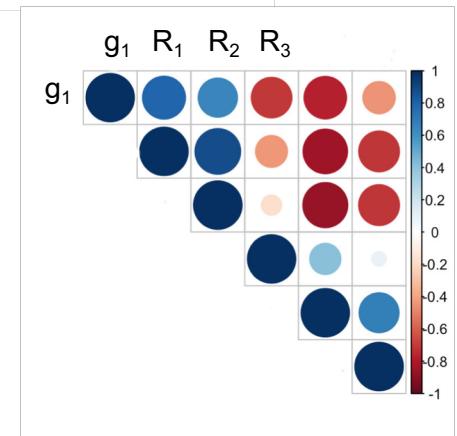
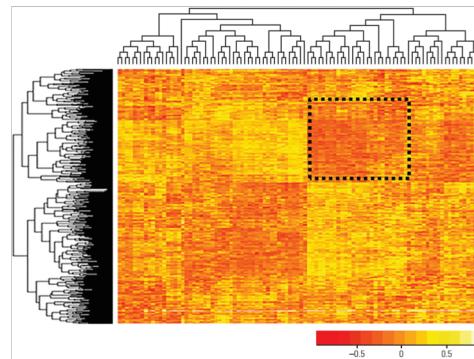
## Graphical Lasso



# Graphical Lasso

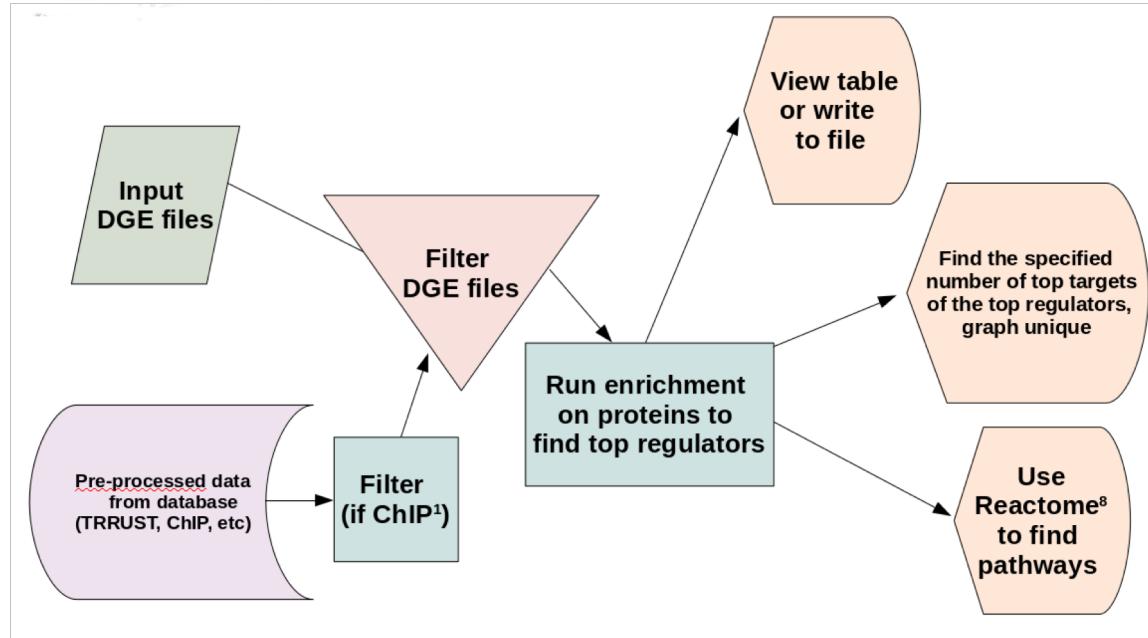


- Tissue-specific gene expression
- 11,688 samples
- 53 tissues
- 714 donors

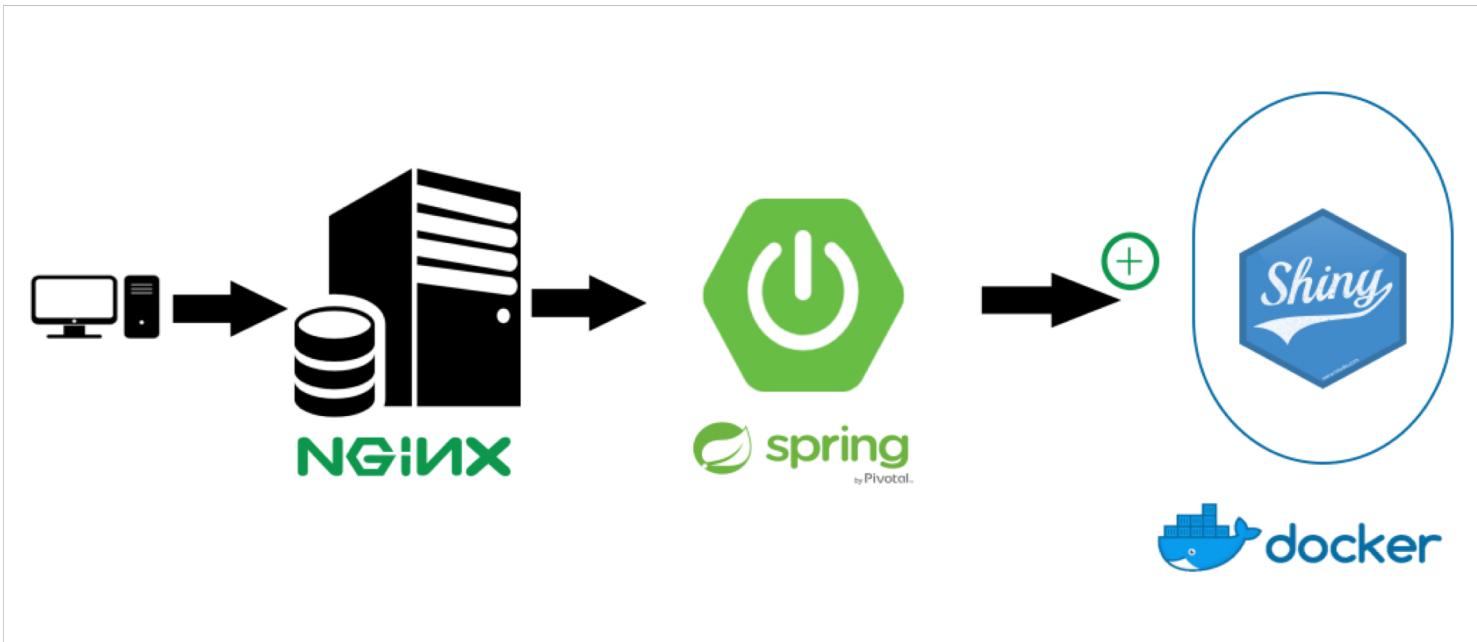


# CIE R Package and Software

# CIE R Package Workflow



# CIE App Workflow



# CIE App Screenshot of Enrichment

Causal Inference Engine

cie analysis Download files Press F11 to exit full screen

### Causal Inference and Directional Enrichment Methods on Biological Networks

Database Type: ChIP Atlas  
Distance from TSS: 1kb  
Cutoff Type: Automatic  
Limit ChIP Atlas results by cell line of experiment: NA  
Limit ChIP Atlas results by cell line tissue origin: NA  
Limit ChIP Atlas results by the diagnosis of the individual the cell line is from: NA  
Show if the Protein Targets Selected Genes: NA  
Upload your differentially expressed gene table, should be in .tsv format:  
Browse... cell\_line\_vehicle\_tgb; evidence\_edgeR.txt Upload complete  
p-Value Threshold: 0.05  
Select a value, the log of which will be the fold count threshold: 1.5  
Method of Enrichment Analysis: Fisher  
 Include whether a protein is a transcription factor in enrichment results  
 Include whether a protein is a BHHLH in enrichment results  
Run analysis

Graph of Regulators



Targets to Display: 100

Choose How Many Paths to Display: 100

#### Enrichment Results for Fisher analysis with data base ChIP

Download Full Enrichment Table Run pathway enrichment Download Full Pathway Table

name	total_targets	significant_targets	pval
351	7654	1364	3.5877922346924e-146
370	12331	1749	3.20082156352994e-112
529	13089	1805	3.25968666301058e-109
376	13689	1842	1.09325607767799e-103
136	11929	1695	2.53332347446923e-101
87	11425	1637	1.07423327508024e-94
654	12278	1712	1.78307068351446e-94
671	12236	1707	9.09615093093646e-94
629	13244	1790	2.9418855021937e-93
607	12354	1716	3.54432089091293e-93

Showing 1 to 10 of 793 entries

Previous 1 2 3 4 5 ... 73 Next

<https://markov.math.umb.edu/endpoint/ba87e091-7fd4-4439-982a-75462cd63c8b/#tab-8216-1>

# CIE App Screenshot of Pathway Enrichment

Causal Inference Engine

CIE Analysis Download files

### Causal Inference and Directional Enrichment Methods on Biological Networks

Database Type: ChIP Atlas  
Distance from TSS: 1kb  
Cutoff Type: Automatic  
Limit ChIP Atlas results by cell line of experiment: NA  
Limit ChIP Atlas results by cell line tissue origin: NA  
Limit ChIP Atlas results by the diagnosis of the individual the cell line is from: NA  
Show if the Protein Targets Selected Genes: NA  
Upload your differentially expressed gene table, should be in .tsv format:  
Browse... cell\_line\_vehicle\_tgb\_evidence\_edgeR.txt Upload complete  
p-Value Threshold: 0.05  
Select a value, the log of which will be the fold count threshold: 1.5  
Method of Enrichment Analysis: Fisher  
 Include whether a protein is a transcription factor in enrichment results  
 Include whether a protein is a BH4H in enrichment results  
Run analysis

Graph of Regulators

Targets to Display: 100  
Choose How Many Paths to Display: 100

### Enrichment Results for Fisher analysis with data base ChIP

Download Full Enrichment Table Run pathway enrichment Download Full Pathway Table

Show 10 entries Search:

id	name	pValue	fdr	proteinsFound
1	2559583 Cellular Senescence	0.0000478530174053357	0.000837078497277766	LMNB1; RELA; E2F1
2	68911 G2 Phase	0.0000900084405675017	0.000837078497277766	E2F1
3	68686 CDC6 association with the ORC-origin complex	0.000022196192640836	0.00137620899437318	E2F1
4	1362300 Transcription of E2F targets under negative control by p107 (RBL1) and p130 (RBL2) in complex with HDAC1	0.0000731581334113551	0.00270685093622014	E2F1
5	6804116 TP53 Regulates Transcription of Genes Involved in G1 Cell Cycle Arrest	0.0000731581334113551	0.00270685093622014	E2F1
6	1362277 Transcription of E2F targets under negative control by DREAM complex	0.000114118944210428	0.00353768727052328	E2F1
7	4532726 Mitotic G1/G1S phases	0.000150439752050646	0.0039114335533168	E2F1
8	2262752 Cellular responses to stress	0.0001952793128811991	0.00449124219628579	LMNB1; RELA; E2F1
9	1538133 G0 and Early G1	0.000262518545757762	0.00500963836493362	E2F1
10	69278 Cell Cycle, Mitotic	0.00029404425239495	0.00500963836493362	LMNB1; E2F1

Showing 1 to 10 of 10 entries Previous 1 Next

<https://markov.math.umb.edu/endpoint/ba87e091-7fd-4439-982a-75462cd63c8b/#tab-8216-2>

# Citations for Networks

- [1] "Summary of Large and Small BEL Corpuses - Home - OpenBEL Wiki." [Online]. Available: <https://wiki.openbel.org/display/home/Summary+of+Large+and+Small+BEL+Corpuses>. [Accessed: 05-Oct-2018].
- [2] S. Oki *et al.*, "Integrative analysis of transcription factor occupancy at enhancers and disease risk loci in noncoding genomic regions," *bioRxiv*, p. 262899, Apr. 2018.
- [3] D. Szklarczyk *et al.*, "The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible," *Nucleic Acids Res.*, vol. 45, no. Database issue, pp. D362–D368, Jan. 2017
- [4] C. Jiang, Z. Xuan, F. Zhao, and M. Q. Zhang, "TRED: a transcriptional regulatory element database, new entries and other development," *Nucleic Acids Res.*, vol. 35, no. Database issue, pp. D137–D140, Jan. 2007.
- [5] H. Han *et al.*, "TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions," *Nucleic Acids Res.*, vol. 46, no. Database issue, pp. D380–D386, Jan. 2018.