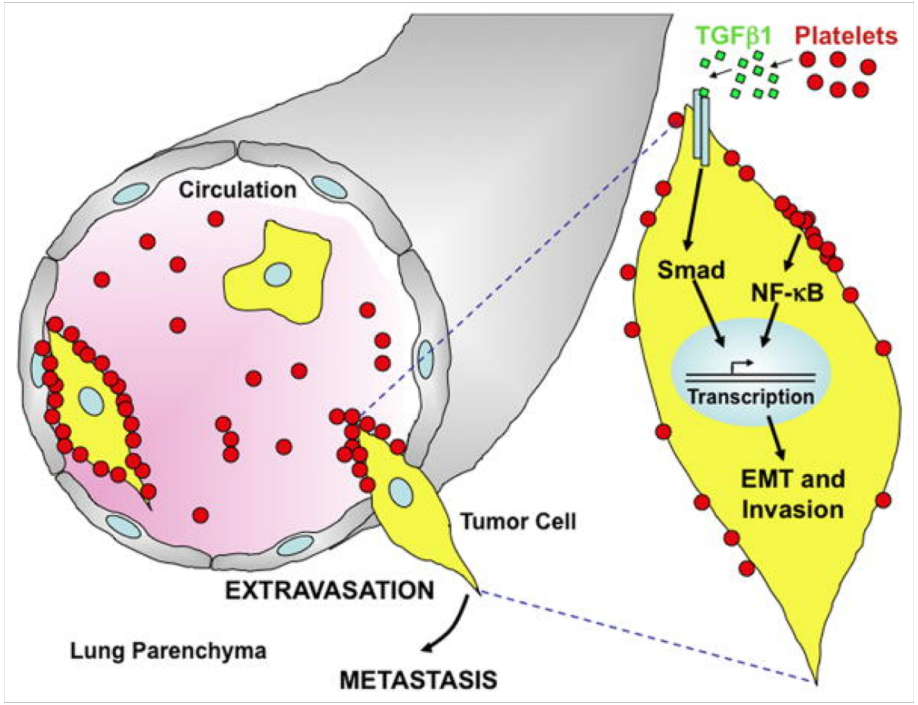
CIE
Modeling Cellular Response

Corey O’Connor and Saman Faramand

Background

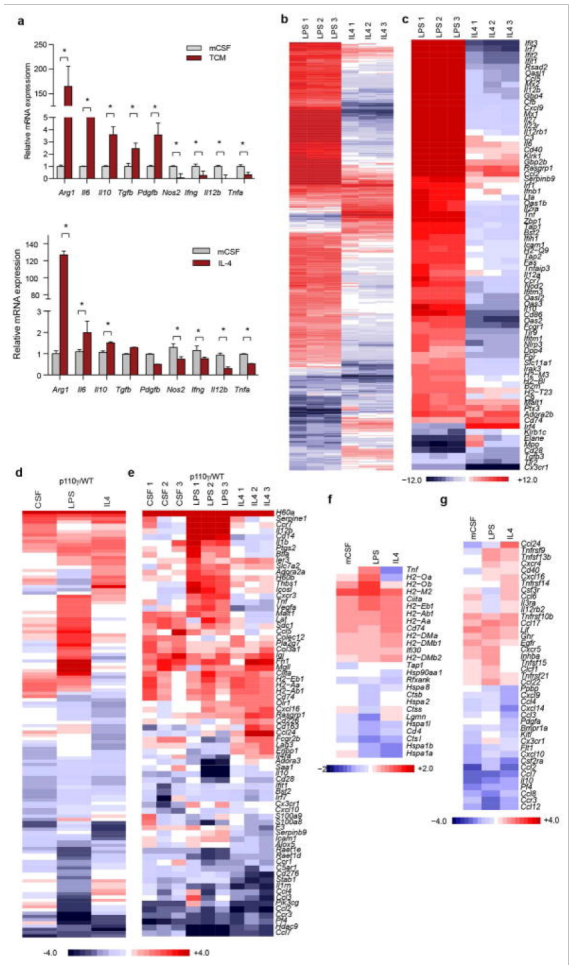
Changes in Gene Expression



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.*, “PI3Kγ is a molecular switch that controls immune suppression,” *Nature*, vol. 539, no. 7629, pp. 437–442, Nov. 2016.

Approaches to Modeling Cellular Repsonse

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 $D[n_{++}, n_{+-}, n_{-+}, n_{--}] = \binom{q_+}{n_{++}, n_{+-}, n_{+0}} \binom{q_-}{n_{-+}, n_{--}, n_{-0}} \binom{q_0}{n_{0+}, n_{0-}, n_{00}}$. &

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

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

4×3 **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	Predicttion, Databases [4]	No
TRRUST	Textmining [5]	Mix

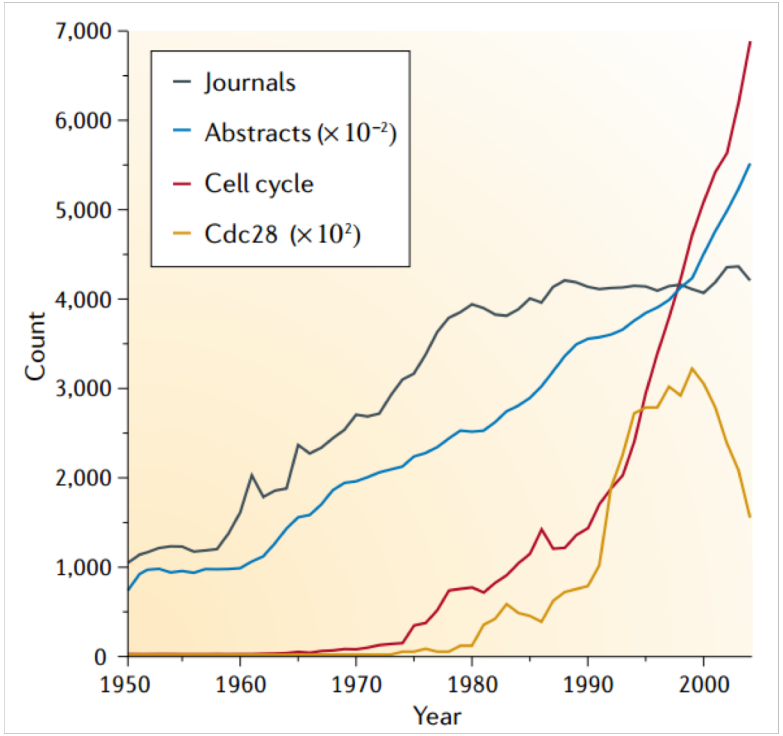
*Citations at end of presentation

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)?

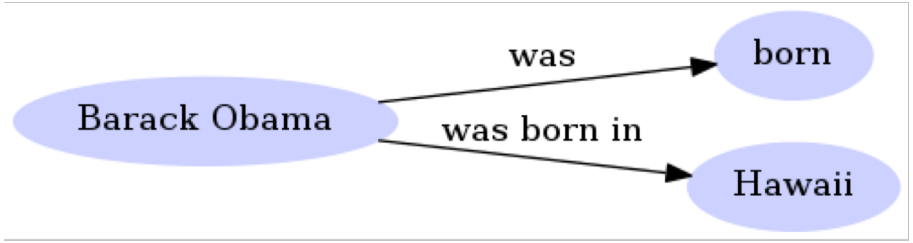


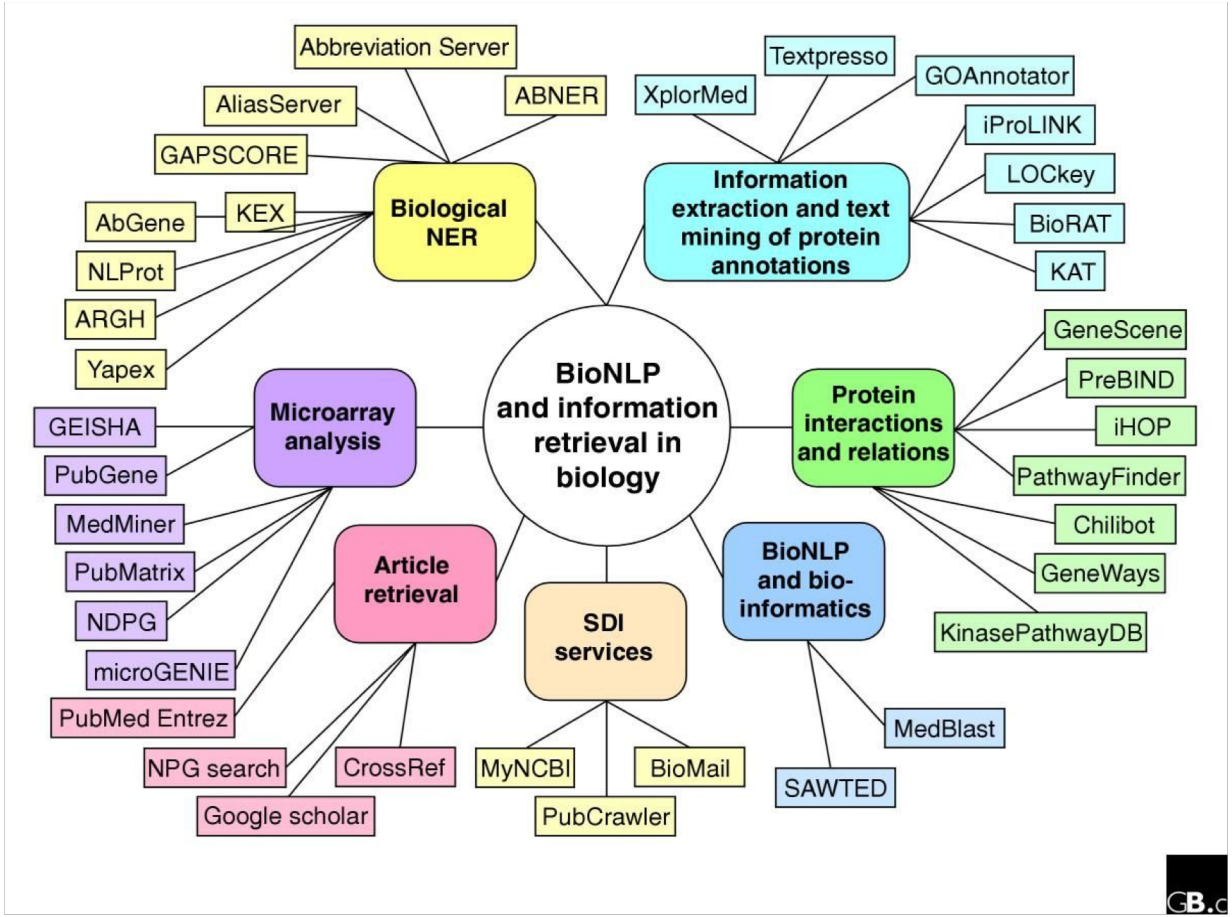
Jenden et al., Literature mining for the biologist: from information retrieval to biological discovery, *Nature Reviews Genetics* volume7, pages119–129 (2006)

What is NLP?

https://www.techsophy.com/chatbots-need-natural-language-processing/

- ## 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





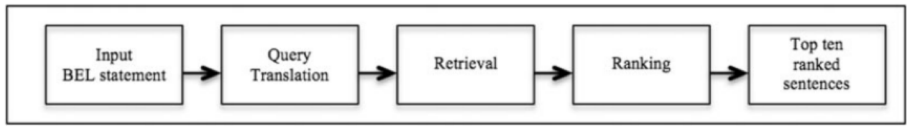
Krallinger et al, Text-mining and information-retrieval services for molecular biology, *Genome Biology* 2005 **6**:224

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)

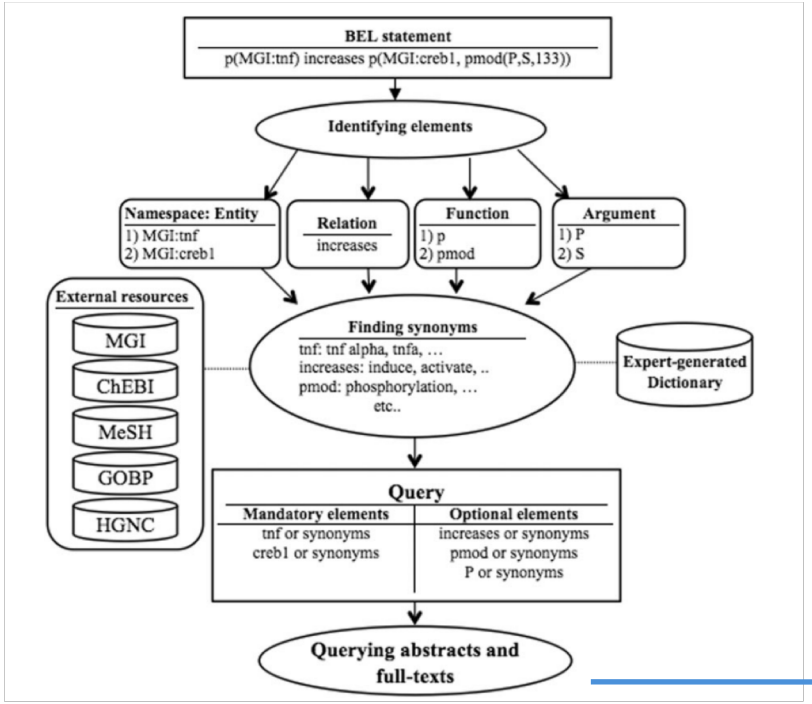
Sentences	BEL statement
We showed that HSF 1 is phosphorylated by the protein kinase RSK2 in vitro we demonstrate that RSK2 slightly represses activation of HSF1 in vivo	1: kin (p (HGNC: RPS6KA3)) increases p (HGNC: HSF1, pmod (P))2: kin (p (HGNC: RPS6KA3)) decreases tscript (p (HGNC: HSF1))
Whereas exposure of neutrophils to LPS or TNF-a 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.



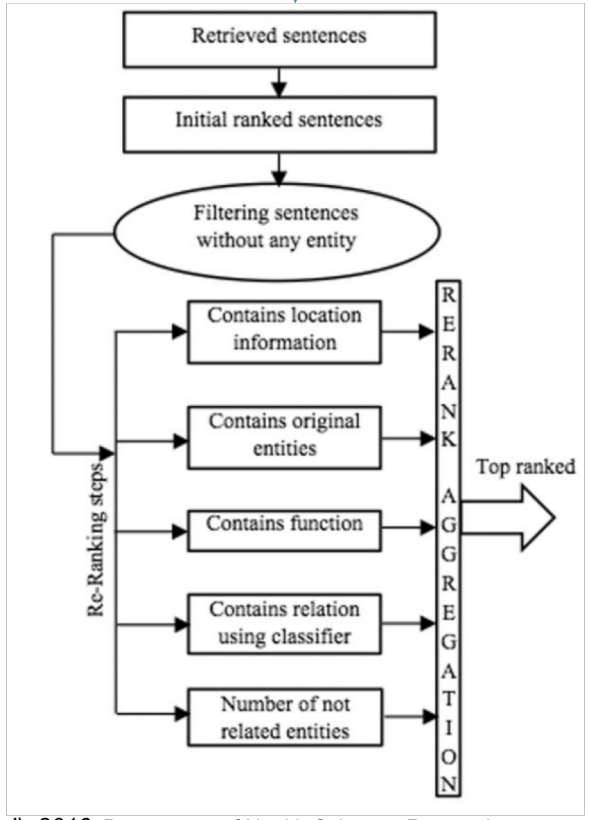
Rastegar-Mojarad et al, BELTracker: evidence sentence retrieval for BEL statements, *Database (Oxford). 2016*

BELTracker: evidence sentence retrieval for BEL statements



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

PubMed



Rastegar-Mojarad et al, BELTracker: evidence sentence retrieval for BEL statements, *Database (Oxford). 2016;* Department of Health Sciences Research, Mayo Clinic, USA

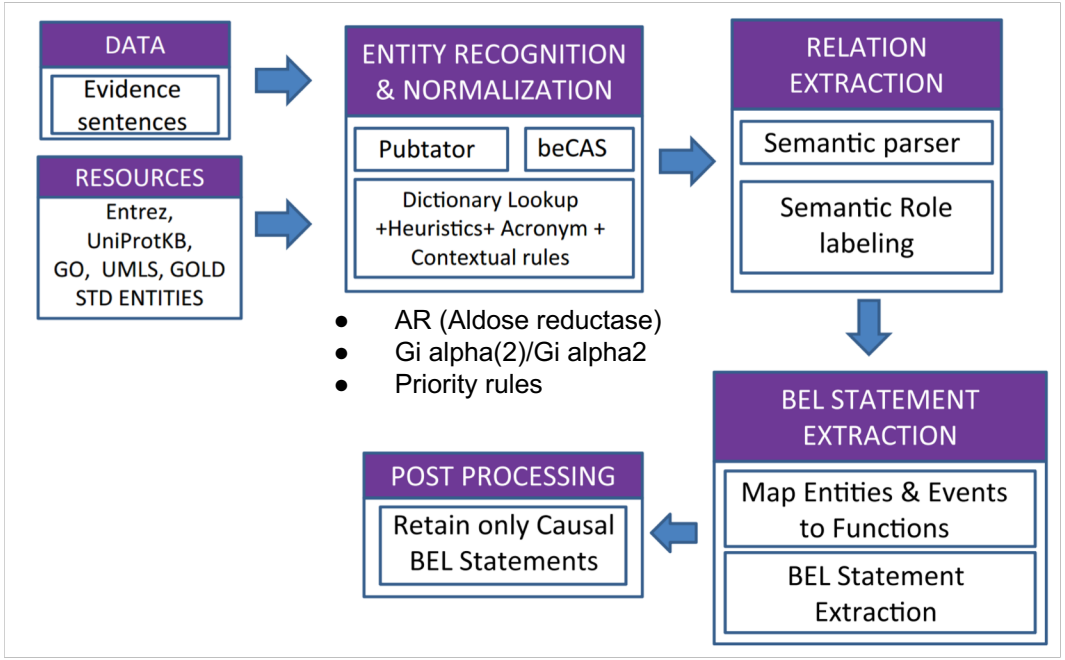
BELTracker: evidence sentence retrieval for BEL statements

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

Rastegar-Mojarad et al, BELTracker: evidence sentence retrieval for BEL statements, *Database (Oxford). 2016*

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

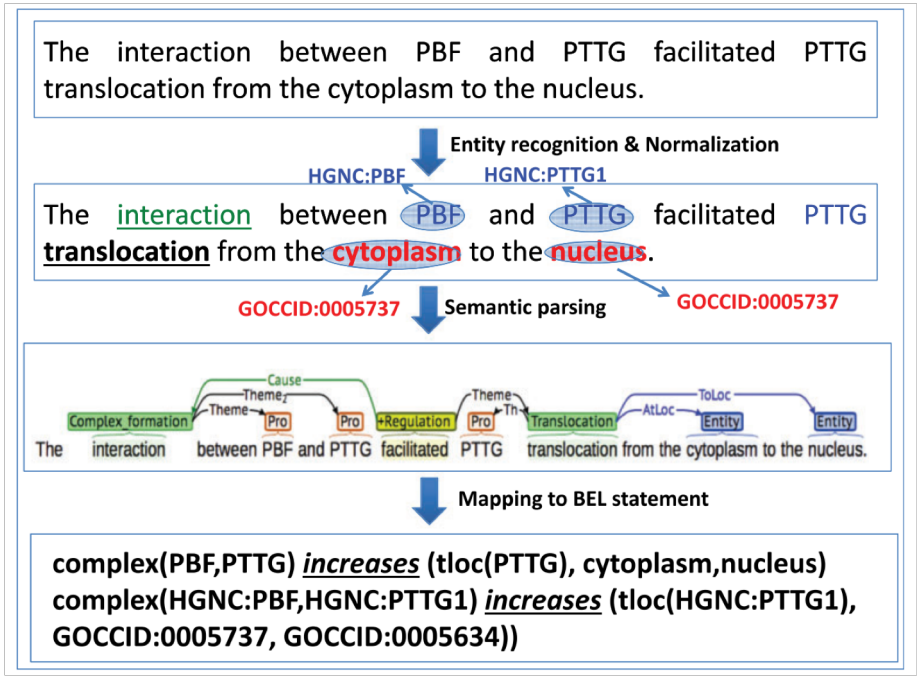
- Addressed the BioCreative V challenge
  - Given textual evidence for a BEL statement, generate the corresponding BEL statement (information extraction [IE] task)



Ravikumar et al, BELMiner, *Database*, Volume 2017, Department of Health Sciences Research, Mayo Clinic, USA

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)



Ravikumar et al, BELMiner, *Database*, Volume 2017, Department of Health Sciences Research, Mayo Clinic, USA

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.

Ravikumar et al, BELMiner, *Database*, Volume 2017, Department of Health Sciences Research, Mayo Clinic, USA

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

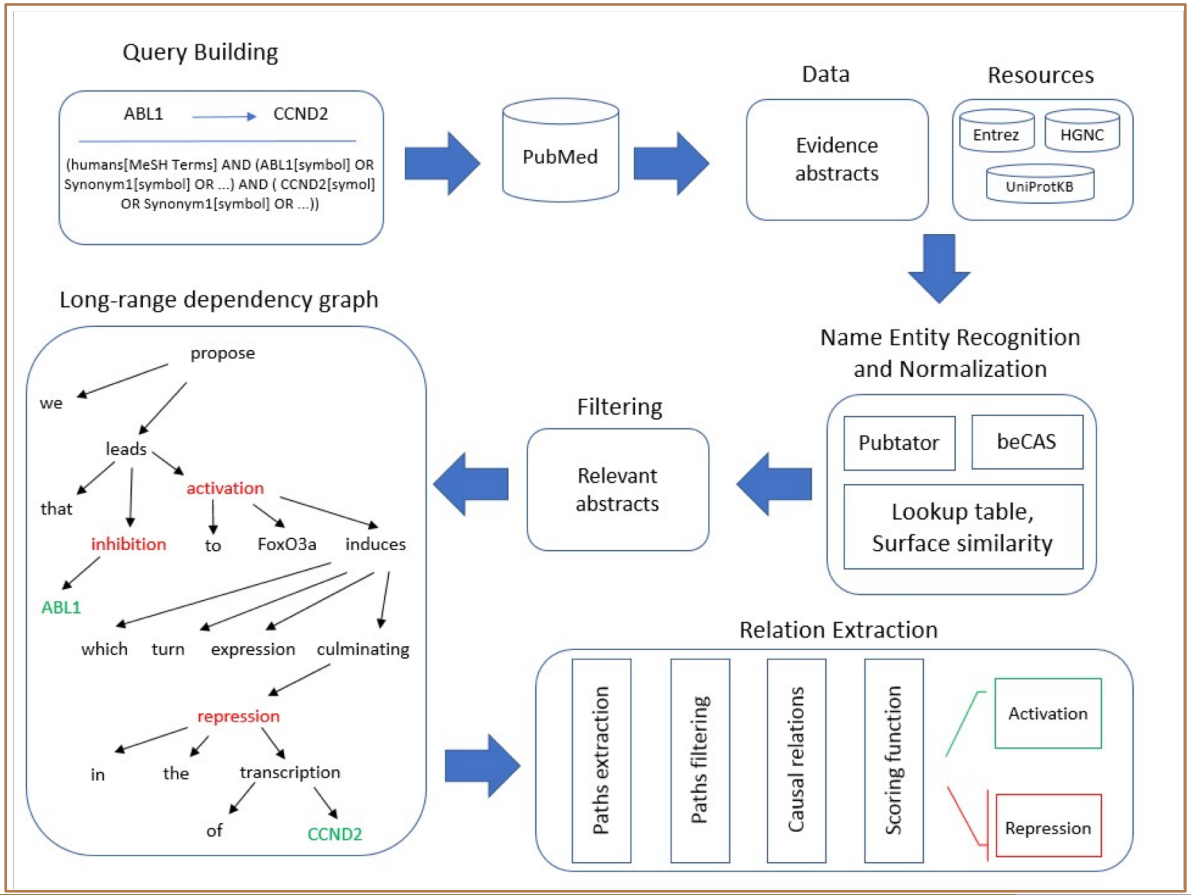
- Limitations
  - Long response time
  - Ranking methods operate at the sentence level

Ravikumar et al, BELMiner, *Database*, Volume 2017, Department of Health Sciences Research, Mayo Clinic, USA

Proposed NLP Pipeline

for ChIP-Seq data annotation

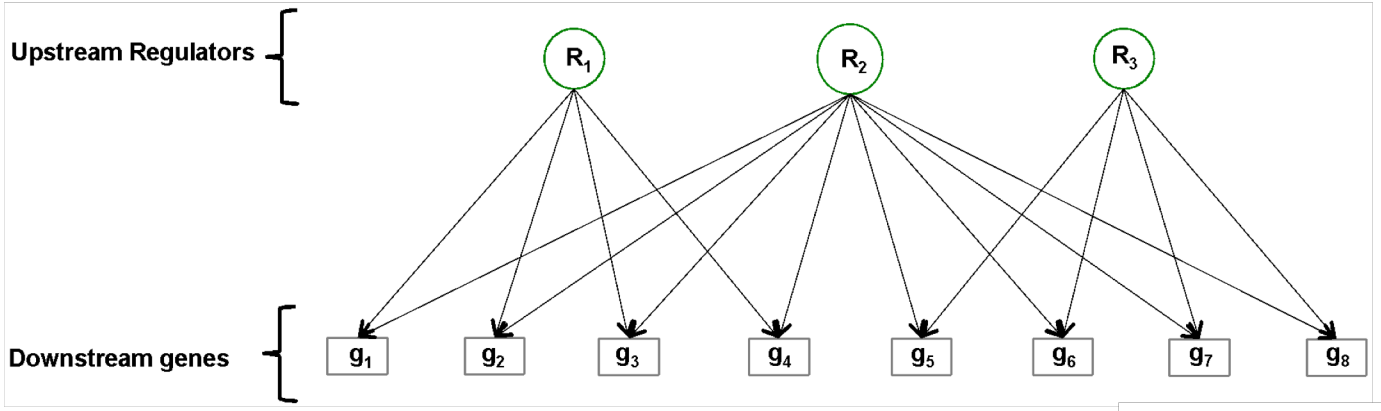
Proposed NLP Pipeline

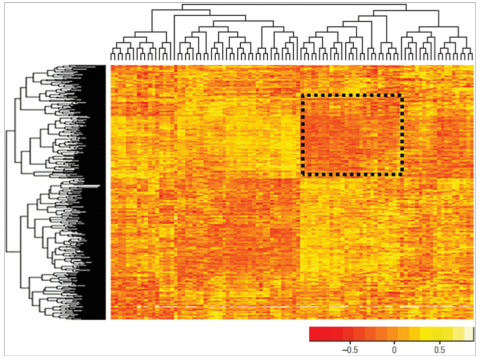


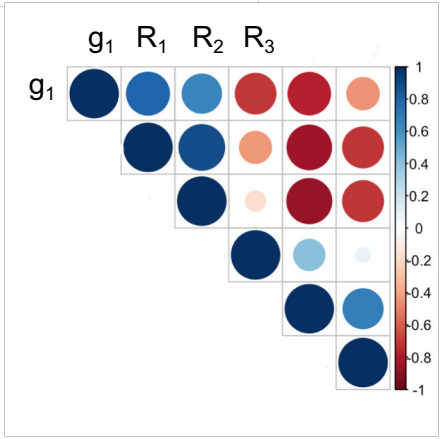
Statistical Perspective

Graphical Lasso

Graphical Lasso



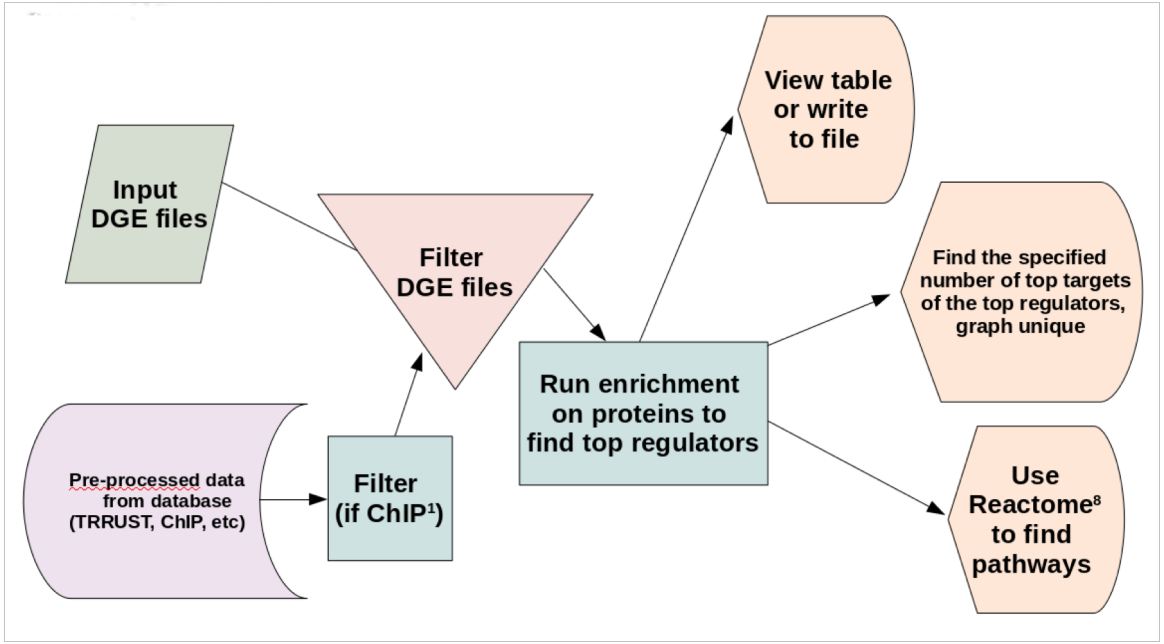




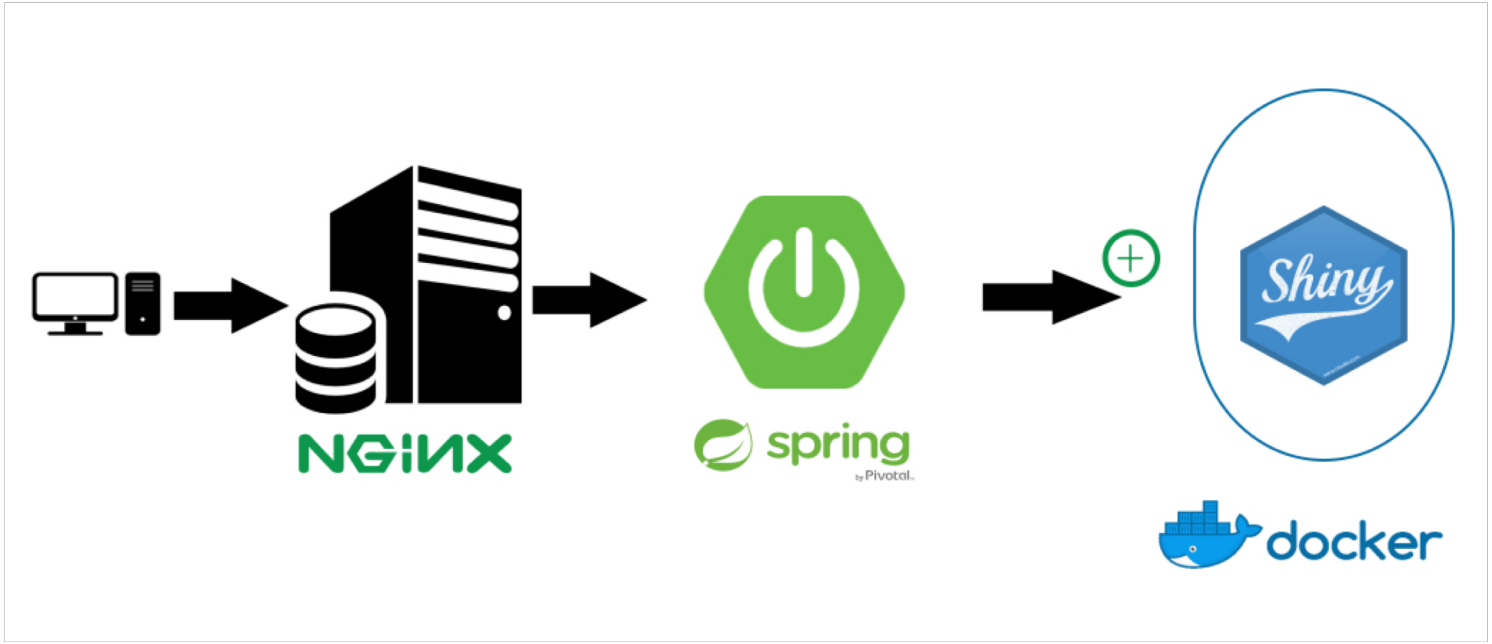
- Tissue-specific gene expression
- 11,688 samples
- 53 tissues
- 714 donors

CIE R Package and Software

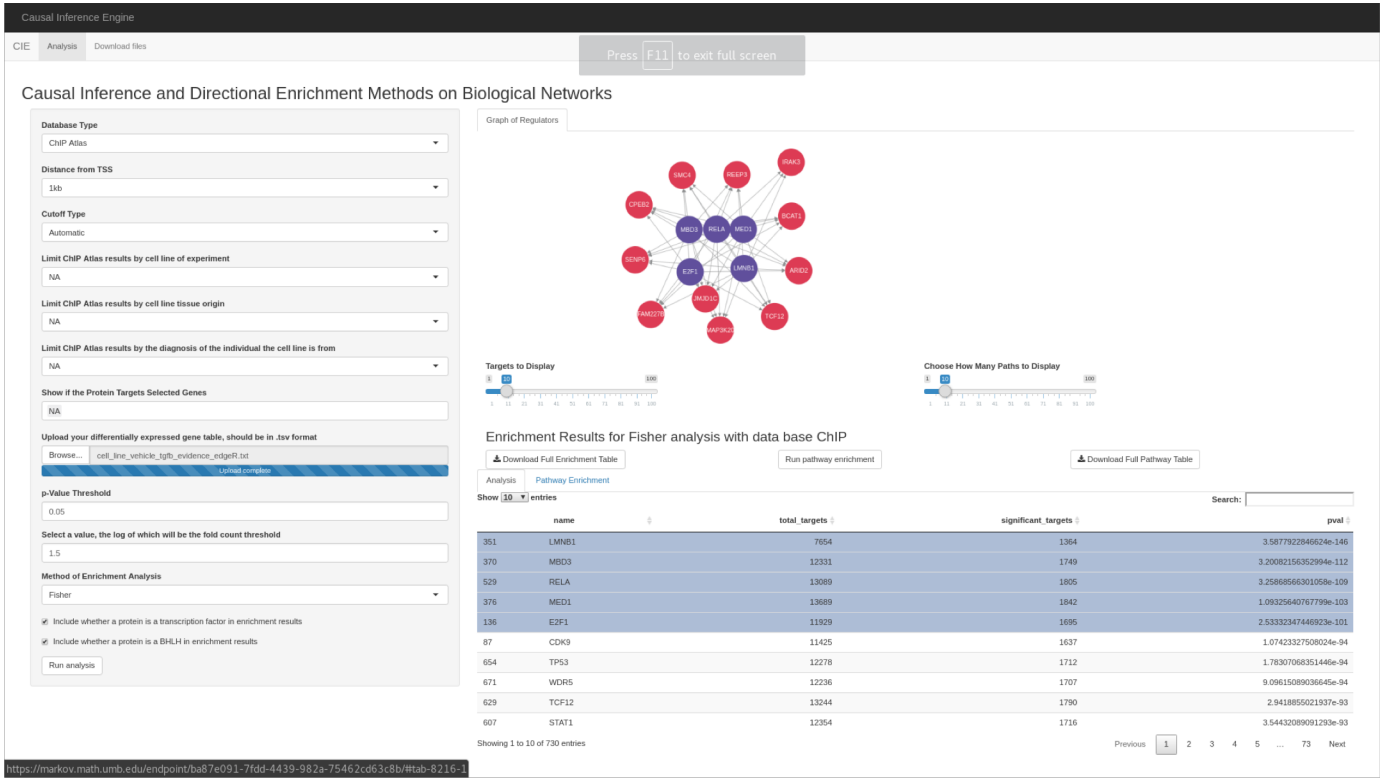
CIE R Package Workflow



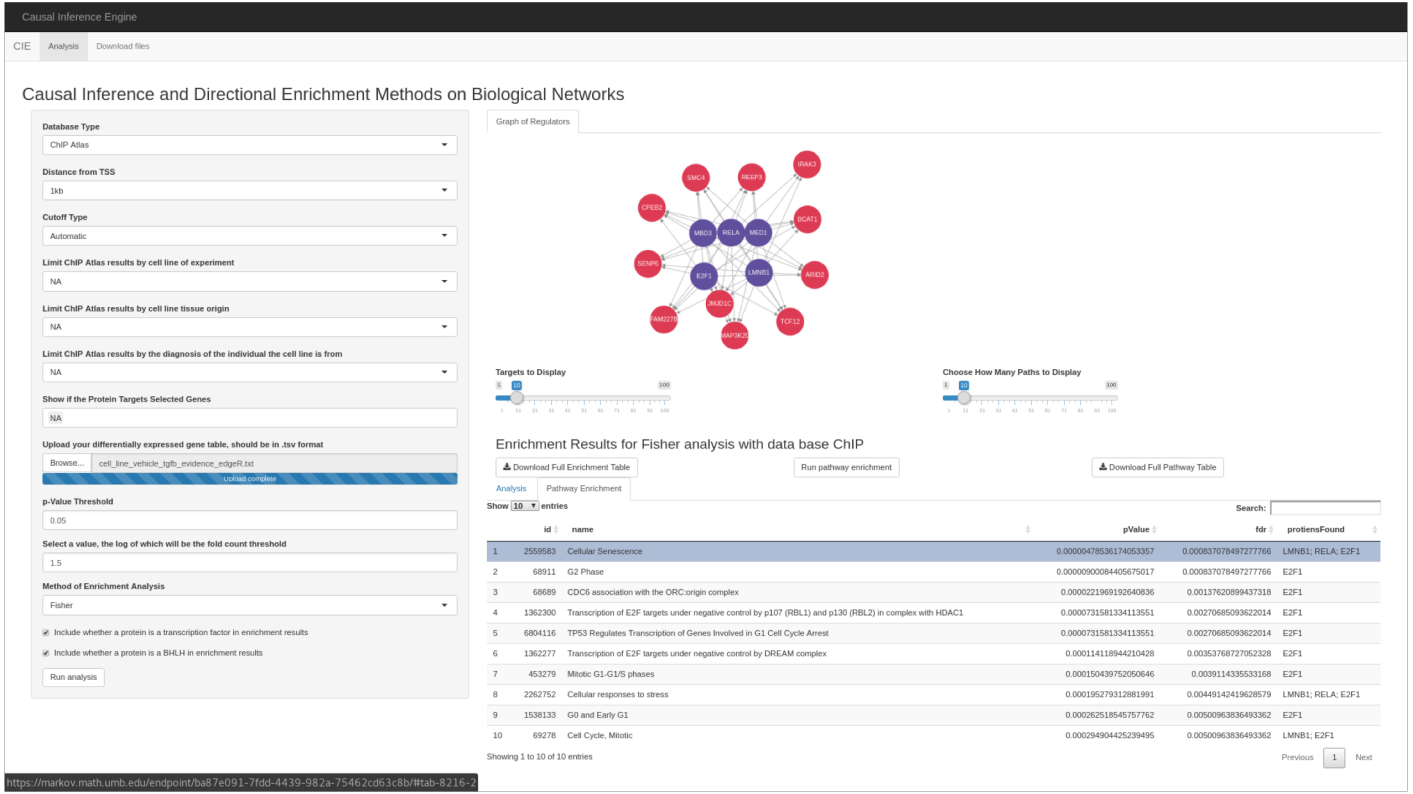
CIE App Workflow



CIE App Screenshot of Enrichment



CIE App Screenshot of Pathway Enrichment



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.