PathBinder—text empirics for automatic extraction of biomolecular interactions

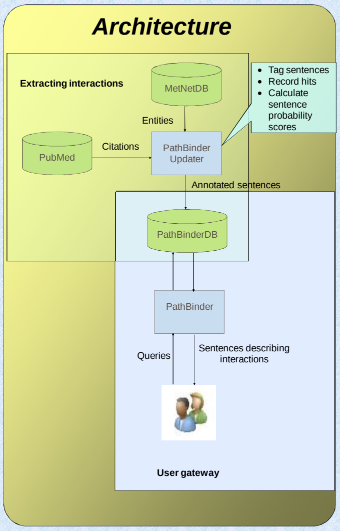
Lifeng Zhang, Department of Electrical & Computer Engineering, Iowa State University
Daniel Berleant, Department of Information Science, University of Arkansas at Little Rock (contact)
Jing Ding, Ohio State University Medical Center
Eve Wurtele, Department of Genetics, Development & Cell Biology, Iowa State University

Abstract

Motivation: Large amounts of free, online biological text makes automatic fact extraction attractive. We investigate text empirics to support mining of biomedical texts for biomolecular interactions.

Results: We analyzed readily computable sentence properties that are potentially relevant to extracting interactions between given biomolecules. The empirical result data was used to design an algorithm for the PathBinder system to identify these interactions from sentences in the literature. Given two biomolecules, it searches PubMed for sentences most likely to describe an interaction between them, and estimates the likelihood that each sentence describes an interaction. In addition, we designed and implemented a method to combine the evidence from multiple relevant sentences to get the likelihood of interaction between two given biomolecules. We then constructed a biomolecular interaction network.

Architecture



Approach

- Advance understanding about the **empirical properties** of biomedical texts. This is an alternative to machine learning based approaches. Apply this knowledge to automatic extraction of biomolecular interactions from the literature.

Some empirical properties of passages (Key: IIT means *interaction-indicating term*)	# / % that describe an interaction	Total number
Sentences where two biomolecules tri-occur with at least one IIT	331/55%	606
Sentences where two biomolecules co-occur without any IIT	3/7.9%	38
All sentences where two biomolecules co-occur	334/52%	644
Phrases where two biomolecules tri-occur with at least one IIT	236/71%	334
Phrases where two biomolecules co-occur without any IIT	0/0%	17
All phrase where two biomolecules co-occur	236/67%	351
Sentence co-occurrences not in phrases	98/33%	293

	IIT intervening	**IIT elsewhere in sentence**	**IIT either place**
Phrase co-occurrences (precision)	63%	24%	45%
Sentence (but not phrase) co-occurrences (precision)	30%	9.1%	21%
All co-occurrences (precision)	48%	17%	34%
Percent of all interactions (recall)	77%	23%	100%

Some empirical properties of interaction indicating terms:

IIT forms	**Sentences describing interactions**	**All sentences**	**Percentages**
noun	141	237	59%
adj	9	20	45%
adv	0	0	N/A
present tense	50	76	66%
-ing	35	69	51%
past/perfect	77	141	55%
IIT categories			
association	60	89	67%
modification	80	121	66%
negative regulation	33	84	39%
positive regulation	47	112	42%
transportation	14	21	67%
transcription	5	7	71%
create	63	96	66%
vague	41	76	54%

IIT forms	**phrases describing interactions**	**all phrases**	**percentage**
noun	97	148	66%
adj	3	7	43%
adv	0	0	0%
present	31	42	74%
-ing	16	29	55%
Past/perfect	56	86	65%
IIT categories			
association	41	55	75%
modification	60	77	78%
negative regulation	24	49	49%
positive regulation	30	58	52%
transportation	7	13	54%
transcription	2	2	100%
create	37	51	73%
vague	31	48	65%

- Use empirical properties to evaluate the probability that a given sentence describes an interaction between a given biomolecule pair.
- Scan each sentence in PubMed one by one, identify biomolecule pairs in the sentences, and record the probability scores that the sentences give to the pairs.
- Combine the evidence provided by multiple sentences containing a given pair of biomolecules to assess the probability that they interact. Try three ways as follows.

Finding $p_i = p$(*there is an interaction described between these two entities*):

a) ALL method: $p_i = 1 - (1 - p_1)(1 - p_2)(1 - p_3)\ldots(1 - p_n)$

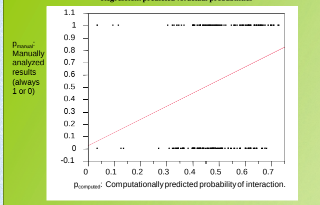
b) BEST2 method: average the two highest probability sentences… $p_i = (p_1+p_2) / 2$

c) BEST5 method: average the five best sentences… $p_i = (p_1+p_2+p_3+p_4+p_5) / 5$

Results

- ***Evaluating sentences as interaction descriptions***

Regression: predicted vs. actual probabilities

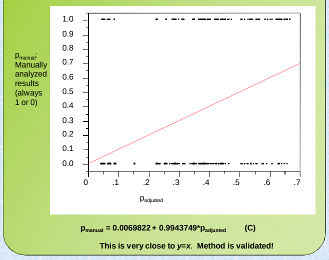


$$p_{manual} = 0.0288512 + 1.0660049 * p_{computed} \quad \text{(A)}$$

Adjusted so ***y=x***:

$$p_{adjusted} = (p_{manual} - 0.0288512) / 1.0660049 \quad \text{(B)}$$

Tested **$p_{adjusted}$** formula on new 600 sentence test set:

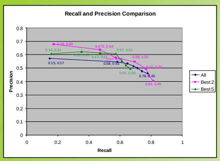


$$p_{manual} = 0.0069822 + 0.9943749 * p_{adjusted} \quad \text{(C)}$$

This is very close to *y=x*. Method is validated!

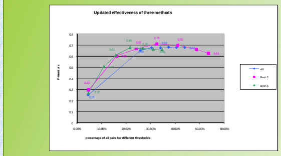
- ***Combining evidence from multiple sentences to create an interaction network***

Manually judged interactions of 400 random pairs from result interaction network. Compared to their scores calculated by ALL, BEST2 and BEST5.

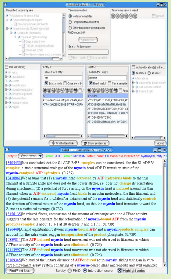


ALL		**BEST2**		**BEST5**	
Score Threshold	**Precision**	**Score Threshold**	**Precision**	**Score Threshold**	**Precision**
1	0.84	0.6	0.95	0.58	0.89
0.95	0.74	0.55	0.84	0.53	0.90
0.9	0.71	0.5	0.8	0.48	0.86
0.85	0.67	0.45	0.73	0.43	0.83
0.8	0.65	0.4	0.64	0.38	0.74
0.75	0.63	0.35	0.57	0.33	0.69
0.7	0.6	0.3	0.51	0.28	0.65

Precision improves significantly after deleting nonexistent biomolecule “names”



User Interface


PathBinder in MetNet - v1.0 alpha
Simplified taxonomy tree
Taxonomy option
○ No taxonomy filter
○ Simplified taxonomy tree
○ Other taxa under green plants
PMID count: N/A
Update filter
Search for taxonomy
☐ Exact match ☐ Case sensitive
Taxonomy search result
☐ Append
☐ Include verb(s)
Entity 1
search for Entity 1
atp
☒ Exact match ☐ Case sensitive
Entity 2
search for Entity 2
myosin
☒ Exact match ☐ Case sensitive
☐ Include location(s) in the ...
○ sentence ○ abstract