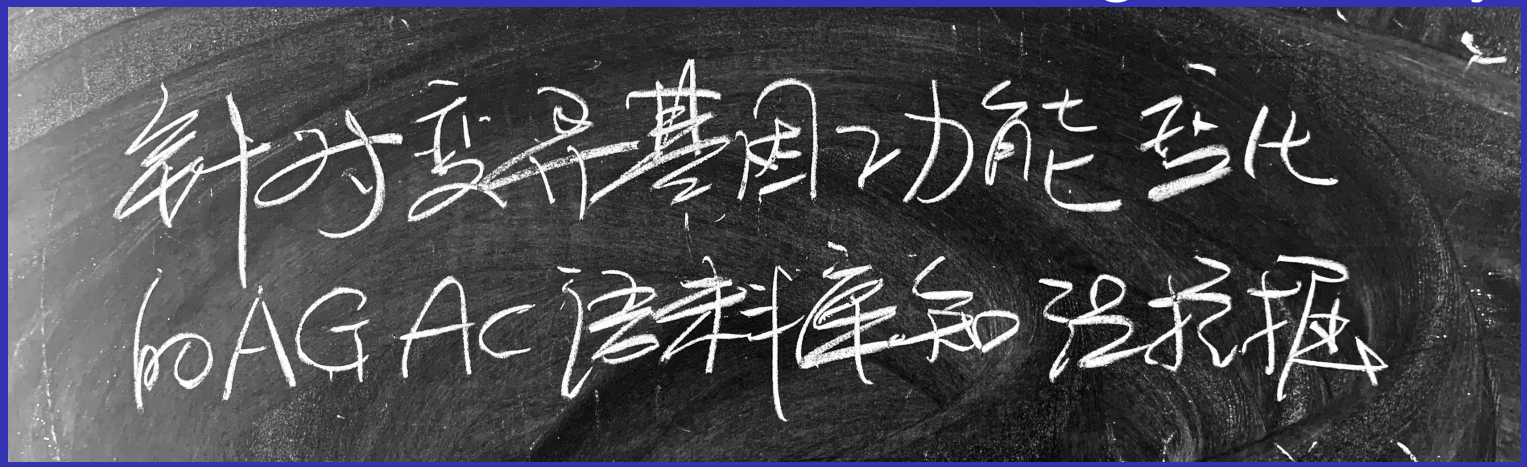
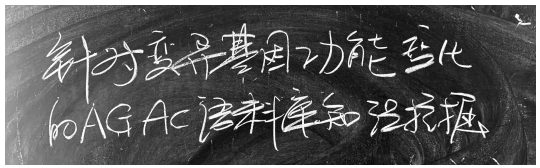
AGAC and AGAC-based Knowledge Discovery



Jingbo Xia

Huazhong Agricultural University

xiajingbo.math@gmail.com

The full slides is available in Google doc 

2019-3-29

Table of contents I

1. Fundamentals 3
2. AGAC corpus design and basic evaluation 8
3. AGAC and tensor decomposition for novel link discovery 15
4. Mathematical efforts to improve the current knowledge discovery scheme 23
5. Conclusion and AGAC track in BioNLP OST 2019 26

Outline

1. Fundamentals 3
2. AGAC corpus design and basic evaluation 8
3. AGAC and tensor decomposition for novel link discovery 15
4. Mathematical efforts to improve the current knowledge discovery scheme 23
5. Conclusion and AGAC track in BioNLP OST 2019 26

AGAC:

*A*ctive *G*ene *A*nnotation *C*orpus (AGAC.V1.0)[1][2]

or

*A*nnotation of *G*enes with *A*ctive mutation-*C*entric function changes (AGAC.V1.1)

[1] Yuxing Wang, et. al. Guideline Design of an Active Gene Annotation Corpus for the Purpose of Drug Repurposing. 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics(CISP-BMEI 2018), Oct, 2018, Beijing.

[2] Kaiyin Zhou, et al. GOF/LOF Knowledge Inference with Tensor Decomposition in Support of High order Link Discovery for Gene, Mutation and Disease. Mathematical Biosciences and Engineering, 2019, 16(3):1376-1391

Drug repurposing:

Drug repositioning (also known as drug repurposing, re-profiling, re-tasking or therapeutic switching) is the application of known drugs and compounds to treat a different disease[3].

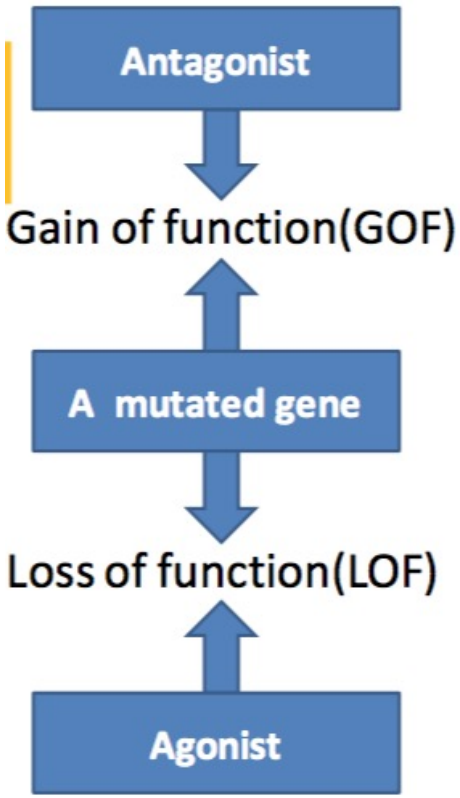
[3] Drug repositioning - Wikipedia https://en.wikipedia.org/wiki/Drug_repositioning

Fundamentals

Why AGAC and drug repurposing?

Several Facts:

1. 10 to 15 years from concept to marketing
2. Total cost to develop a drug ($0.5B to $0.9B)
3. 1 drug is FDA approved for every 5000 compounds tested
4. 1 out of 100 drugs succeeds to market



Pharmacological hypothesis: Focusing on Mutation-LOF/GOF

The hypothesis goes with two directions:

1. A drug act as an *agonist (stimulator)* is a possible cure for disease which is associated with a gene playing a *Loss-Of-Function (LOF)* role after mutation.

2. ... *antagonist (inhibitor)*... *Gain-Of-Function (GOF)* ...

The hypothesis[4] make it informative to find active genes, related to certain drug, and infer its GOF/LOF, and find corresponding repurposed drugs.

[4]Wang, Z. Y., and Zhang, H. Y. (2013). Rational drug repositioning by medical genetics. Nature biotechnology, 31(12), 1080-1082.

Outline

1 Fundamentals 3

2 AGAC corpus design and basic evaluation 8

3 AGAC and tensor decomposition for novel link discovery 15

4 Mathematical efforts to improve the current knowledge discovery scheme 23

5 Conclusion and AGAC track in BioNLP OST 2019 26

Guideline Design of AGAC

Raw text

An example of DAX1 with LOF mutation, C-to-A transversion is shown here:

```
DAX1 http://omim.org/entry/300473, an example of LOF mutation in
OMIM
0027 ADRENAL HYPOPLASIA, CONGENITAL
NR0B1, TYR399TER [dbSNP:rs104894906]
In the proband of a 5-generation Scottish kindred, 3 members of
which had adrenal hypoplasia (300200),Brown et al.  (2003)
identified a C-to-A transversion in the second exon of the DAX1
gene that resulted in the change of tyr399 to a premature stop
codon (Y399X), which truncates the DAX1 protein by 71 amino acid.
...  The mutation was associated with loss of Leydig cell
responsiveness to human chorionic gonadotropin...  the mutation
resulted in a severe loss of DAX1 repressor activity.
```

Guideline Design of AGAC

Raw text with highlighting

An example of DAX1 with LOF mutation, C-to-A transversion is shown here:

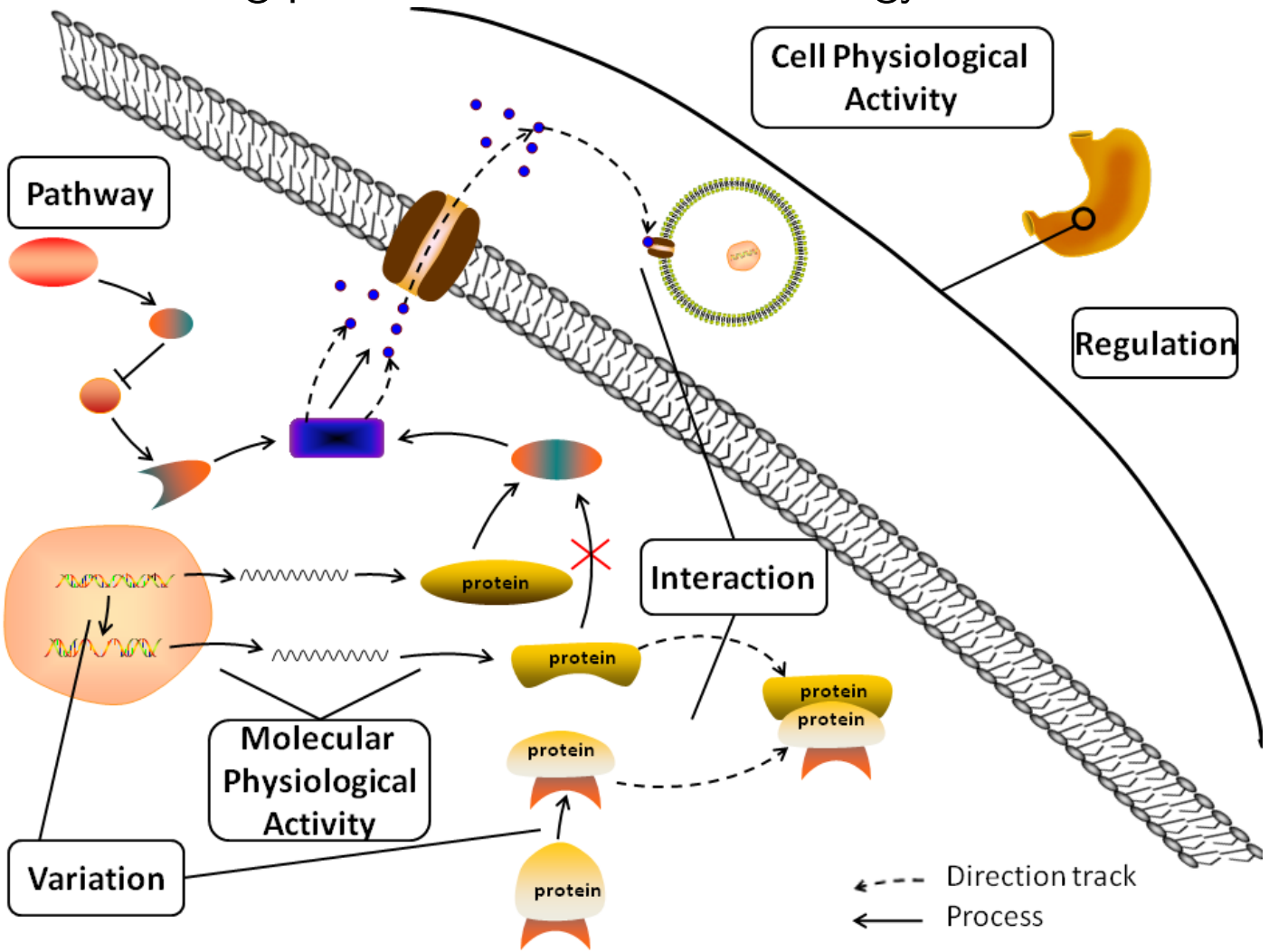
```
DAX1 http://omim.org/entry/300473, an example of LOF mutation in
OMIM
0027 ADRENAL HYPOPLASIA, CONGENITAL
NROB1, TYR399TER [dbSNP:rs104894906]
In the proband of a 5-generation Scottish kindred, 3 members of
which had adrenal hypoplasia (300200),Brown et al.  (2003)
identified a C-to-A transversion in the second exon of the DAX1
gene that resulted in the change of tyr399 to a premature stop
codon (Y399X), which truncates the DAX1 protein by 71 amino acid.
...  The mutation was associated with loss of Leydig cell
responsiveness to human chorionic gonadotropin...  the mutation
resulted in a severe loss of DAX1 repressor activity.
```

Here, the mutation, dbSNP:rs104894906, is a C(Cytosine) to A (Adenine) transversion, which resulted in a severe loss of DAX1 repressor activity. Therefore, the mutated DAX1 plays a LOF role.

Guideline Design of AGAC

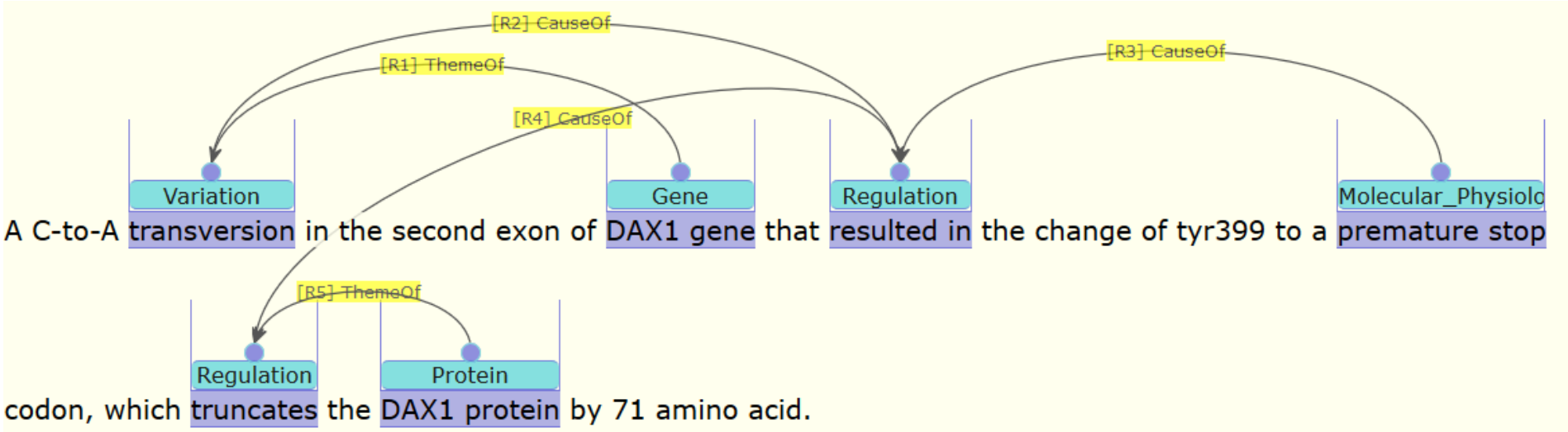
Domain setting for labels design

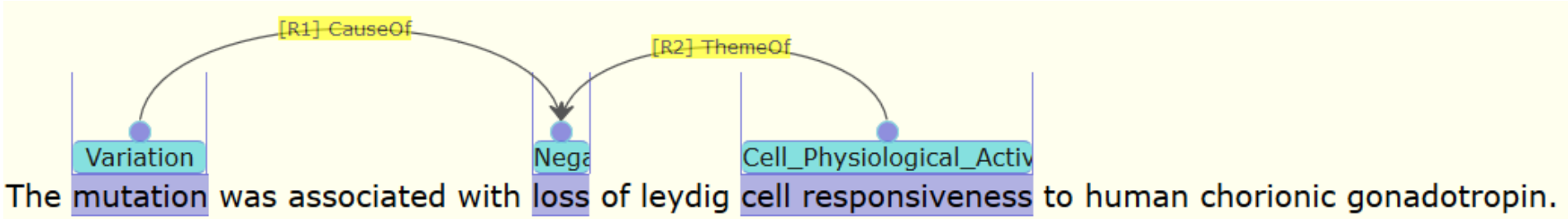
The logic of AGAC trigger words design follows the central dogma and fundamental functioning procedure of molecular biology

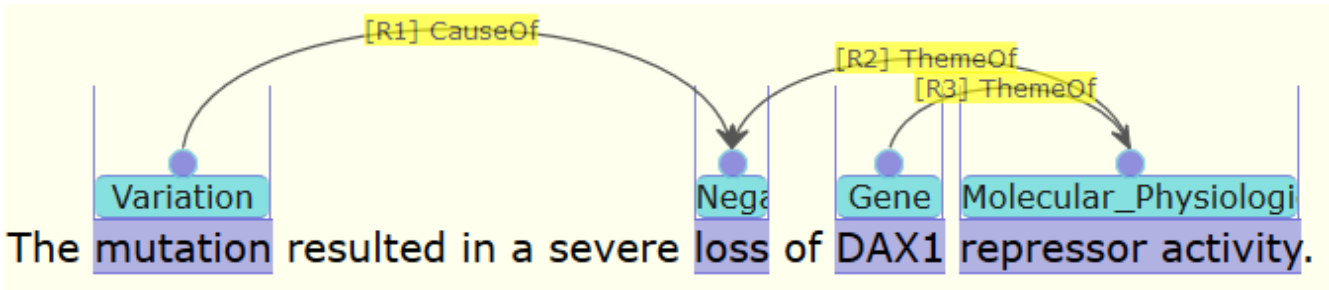


Guideline Design of AGAC

Annotation example







Inter Annotator Agreement

The IAA results were listed in Table 1.

	Annotator1	**Annotator2**	**Annotator3**
Recall	0.66	0.75	0.65
Precision	0.70	0.81	0.76
F-score	0.68	0.78	0.70

Table 1: Inter-annotator agreement among the principle annotator and fellow annotators

Three annotators were involved in the IAA evaluation, where two of them (annotator1, annotator2) major in biology and the other one (annotator3) majors in information engineering. Twenty texts, which contained 147 sentences, 3,120 words, were randomly chosen from AGAC. The three annotators annotated the same 20 texts, and compared with a standard annotation by PubAnnotation comparing tool.

LOF/GOF Topic classification

Performance for LOF/GOF/Unknown topic classifier

Method	Features	Precision	Recall	MacroF-score
Bi-LSTM	Word2Vec	0.387	0.349	0.317
Bi-LSTM-tags	Word2Vec,AGAC labels	0.571	0.534	0.576
Naïve Bayes	AGAC labels	0.669	0.665	0.639
Logistic Regression	AGAC labels	0.681	0.639	0.646
Adaboost	AGAC labels	0.746	0.728	0.736
Random forests	AGAC labels	0.797	0.783	0.789
XGboost	AGAC labels	0.826	0.837	0.831
Bagging	AGAC labels	0.838	0.842	0.840
SVM	AGAC labels	0.832	**0.851**	0.841
Decision Tree	AGAC labels	**0.852**	0.846	**0.848**

Table 2: Experimental results different models.

The purpose of "topic classifier" experiments is to test the effectiveness of AGAC class.

The results of the simulation experiments showed that AGAC annotation labels contributed substantially in the GOF/LOF/Unknown recognition for given gene-related short texts.

Outline

1 Fundamentals 3

2 AGAC corpus design and basic evaluation 8

3 AGAC and tensor decomposition for novel link discovery 15

4 Mathematical efforts to improve the current knowledge discovery scheme 23

5 Conclusion and AGAC track in BioNLP OST 2019 26

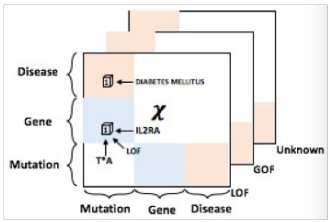
AGAC and tensor decomposition for novel link discovery

Comparison of n-way tensors

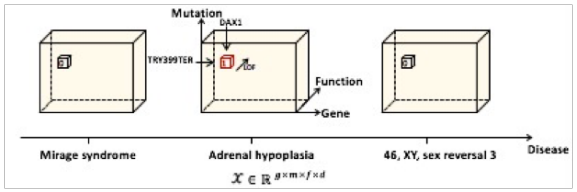
Entity: mutation

Entity: gene **Entity: disease**

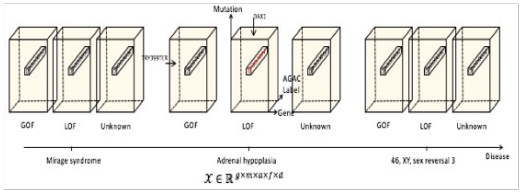
Entity: drug **Entity: function_change**



Three way tensor



Four way tensor



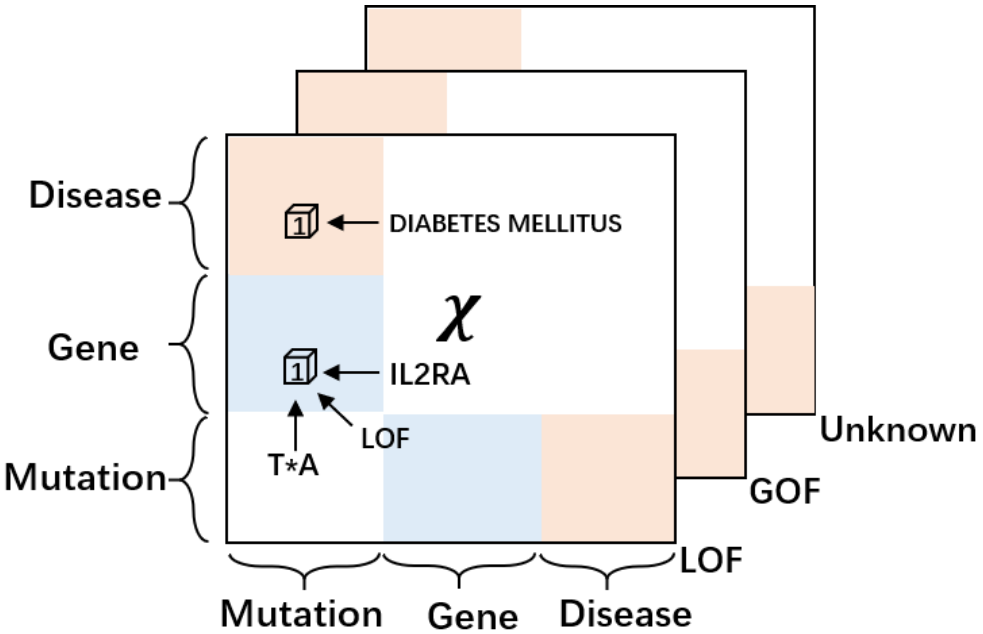
Five way tensor

- **Faster in computation**
- **Less connected info**
- **Directly use RESCAL**

- **Slower in computation**
- **Rich connected info**
- **CP or Tucker considered**

AGAC and tensor decomposition for novel link discovery

A three-way tensor, RESCAL-type tensor

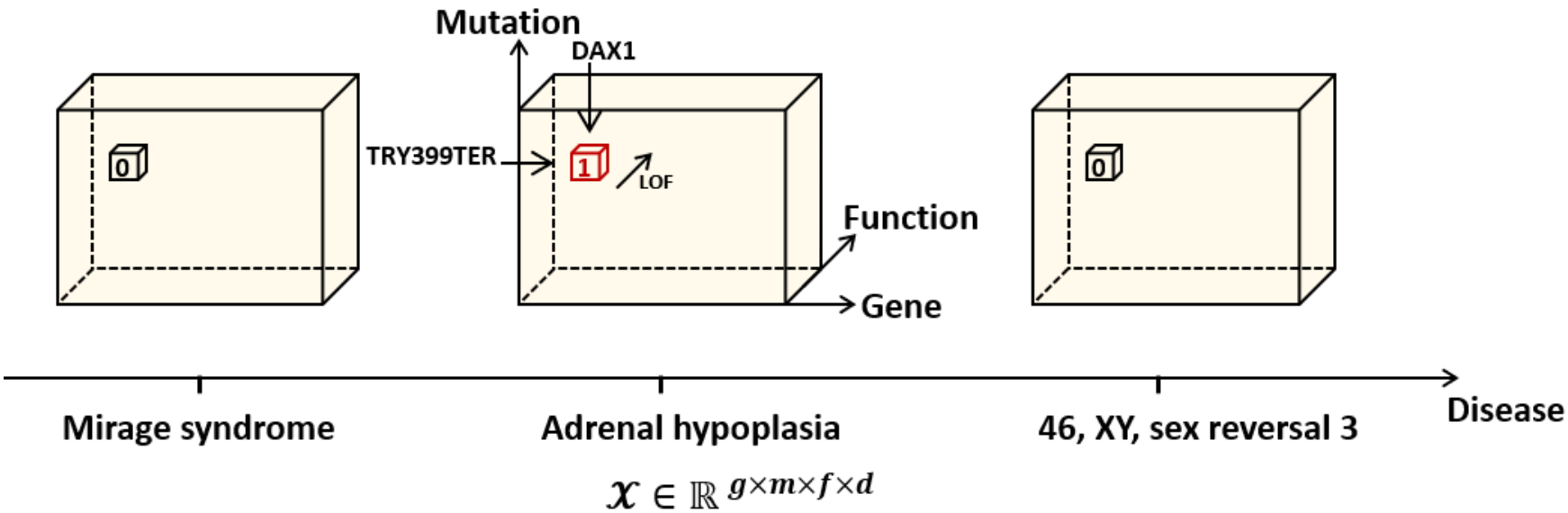


Assuming there are G genes, M mutations, $F(=4)$ types of function_changes, and D kinds of diseases, a three-way tensor $\mathcal{X}^{(3)} \in \mathbb{R}^{n \times n \times 3}$ was defined, where $n(=G+M+D)$ is the amount of the entities. Here,

$$\mathcal{X}^{(3)}_{ijf} = \begin{cases} 1, \text{ if the 3-tuple } (entity_i, function\_change_f, entity_j) \text{ existed} \\ 0, \text{ otherwize} \end{cases}, \quad (1)$$

AGAC and tensor decomposition for novel link discovery

A four-way tensor, Gene-Mutation-Functionchange-Disease (GMFD) tensor

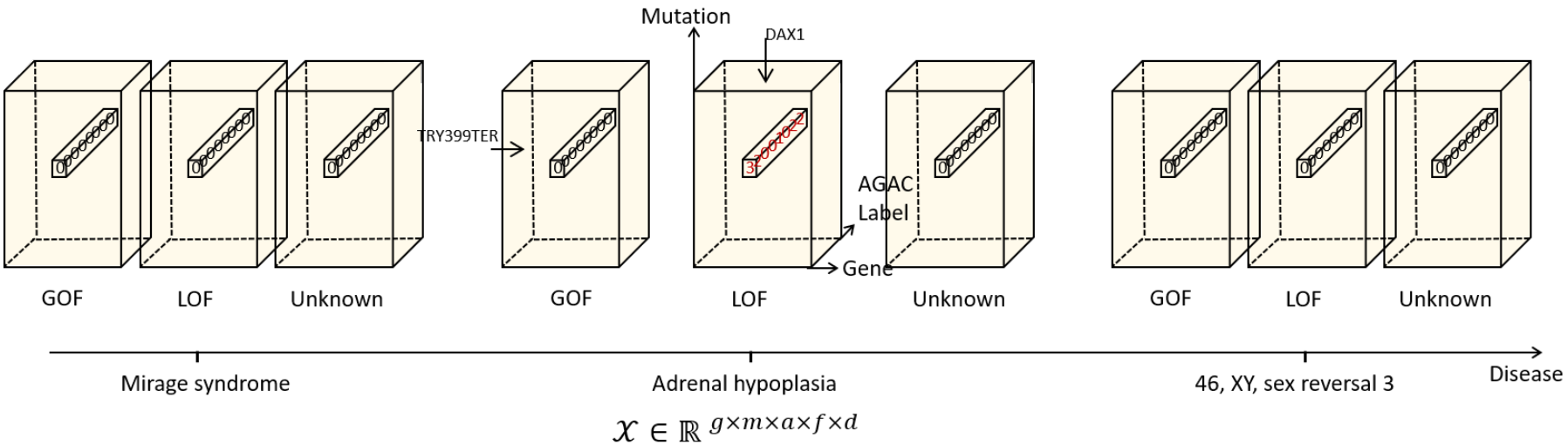


$$\mathcal{X}^{(4)}_{gmfd} = \begin{cases} 1, \text{ if high order } (gene, mutation, function\_change, disease) \text{ existed} \\ 0, \text{ otherwize} \end{cases} \tag{2}$$

when the g-th gene played f-th function_change after the m-th mutation events based on the retrieved texts of d-th disease.

AGAC and tensor decomposition for novel link discovery

A five-way tensor, Gene-Mutation-Annotation-Function_change-Disease (GMAFD) tensor



Assume there are A kinds of annotations in AGAC corpus for the related texts, $\mathcal{X}^{(5)} \in \mathbb{R}^{G \times M \times A \times F \times D}$. Here,

$$\mathcal{X}^{(5)}_{gmafd} = \begin{cases} \#\{a - annotation\}, \text{ for a given tensor } \mathcal{X}^{(4)} \in \mathbb{R}^{G \times M \times F \times D} \\ 0, \text{ otherwize} \end{cases},$$

if the g-th gene played f-th function_change after the m-th mutation events based on the retrieved texts of d-th disease, and there are $\#\{$a-annotation$\}$ anotation for each a-th annotation labels.

AGAC and tensor decomposition for novel link discovery

Performance of tensor decomposition with 3,4, and 5 way tensors

Original tensor contains 1,322 nonzero cells, which corresponded to various entities including 197 genes, 199 mutations, and 313 diseases.

Table 3: Comparison of three tensor decomposition methods.

	Precision	Recall	F-value	AE	MRR(%)	JR(%)
RESCAL for three way tensor	0.299	0.998	0.460	0.972	3.8	0
CPD for GMFD-tensor	0.508	0.298	0.376	0.436	5.1	61.5
CPD for GMAFD-tensor	0.0˜	0.0˜	0.0˜	0.043	0.0	99.4

Evaluation Metric:

(i) Recall, precision and F-score.

(ii) Approximation evaluation (AE). $AE = 1 - \frac{||\tilde{\mathcal{X}} - \mathcal{X}||_F^2}{||\mathcal{X}||_F^2}$.

(iii) Mask recall rate (MRR). In this evaluation, 20% cells with nonzero values in $\mathcal{X}$ were masked.

(iv) Jumping rate (JR). $JR = \frac{\#\{(gene_{novel}, function\_change_{fixed}, disease_{fixed})\}}{\#\{(gene_{all}, function\_change_{fixed}, disease_{fixed})\}}$. The higher the value, the more capable the new tensor produces novel link.

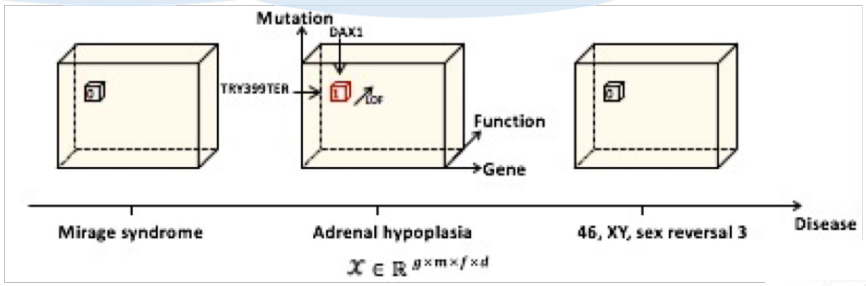
AGAC and tensor decomposition for novel link discovery

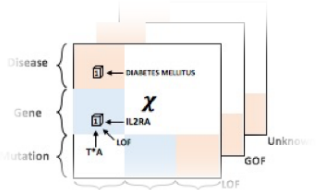
Comparison result of n-way tensors

Entity: mutation

Entity: gene **Entity: disease**

Entity: drug **Entity: function_change**

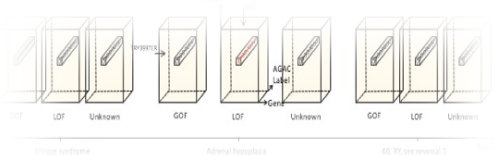




Three way tensor

- Faster in computation
- Less connected info
- Directly use RESCAL

Four way tensor



Five way tensor

- Slower in computation
- Rich connected info
- CP or Tucker considered

GMFD tensor, trade-off in computation efficiency and result reliability.

AGAC and tensor decomposition for novel link discovery

Conclusion of the "AGAC and tensor decomposition for novel link discovery" pipeline:

1. Gene-Mutation-GOF/LOF/-disease information is stored in a four way tensor.
2. AGAC is used when valuing the cell in this tensor.
3. New knowledge exists in $\tilde{\mathcal{X}}$.
4. New nonzero value in $\tilde{\mathcal{X}}$ leads to a novel link for a "Gene-Mutation-GOF/LOF/-disease" higher order link.

Outline

1 Fundamentals 3

2 AGAC corpus design and basic evaluation 8

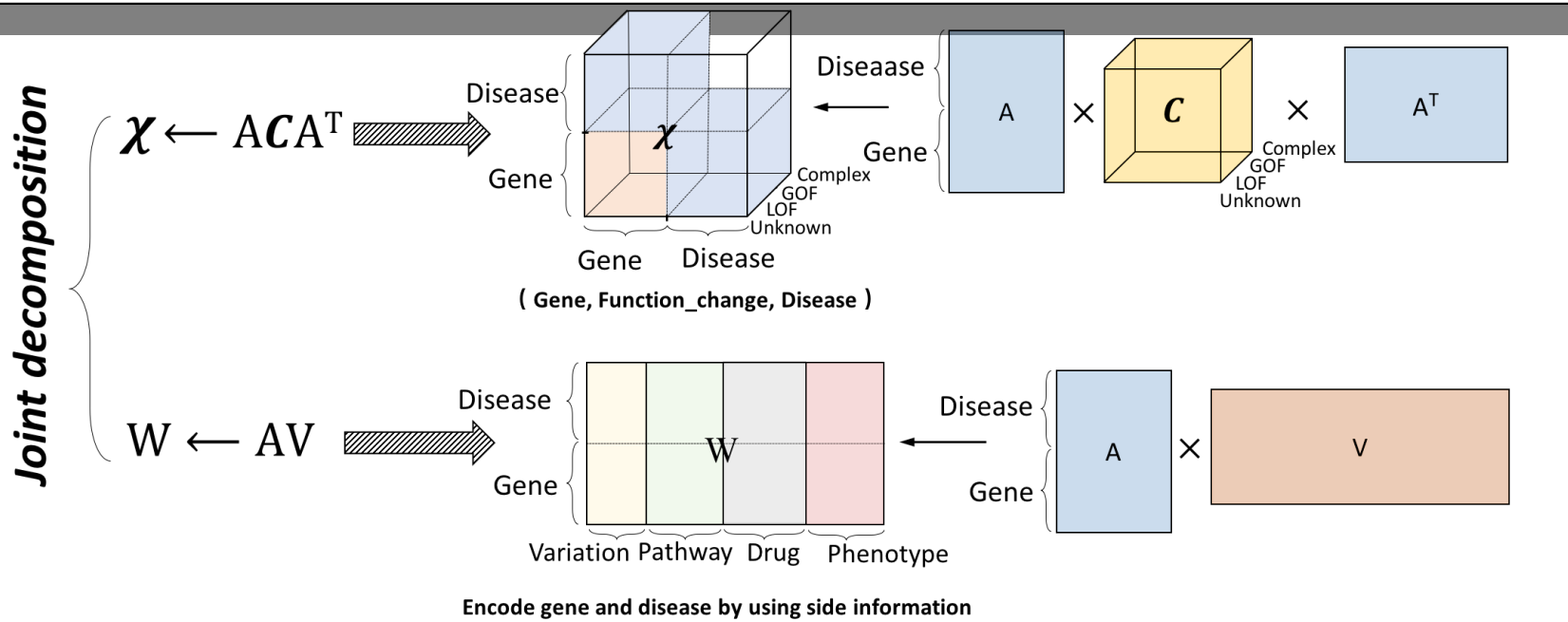
3 AGAC and tensor decomposition for novel link discovery 15

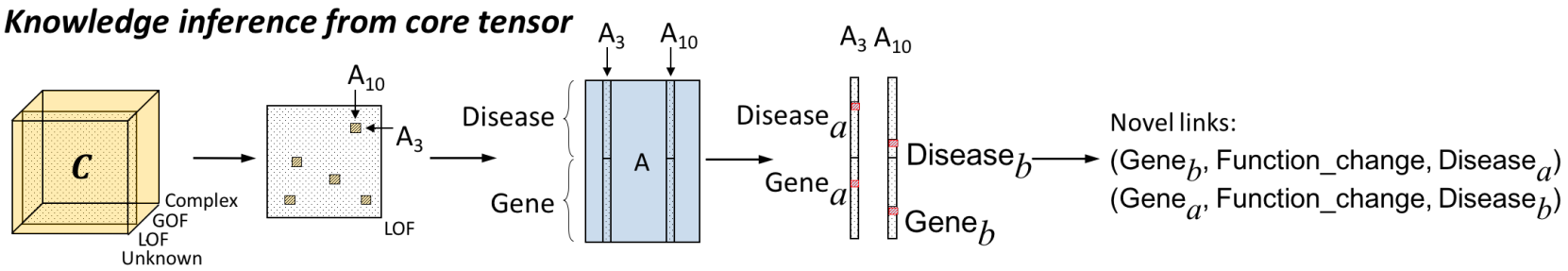
4 Mathematical efforts to improve the current knowledge discovery scheme 23

5 Conclusion and AGAC track in BioNLP OST 2019 26

Mathematical efforts to improve knowledge discovery

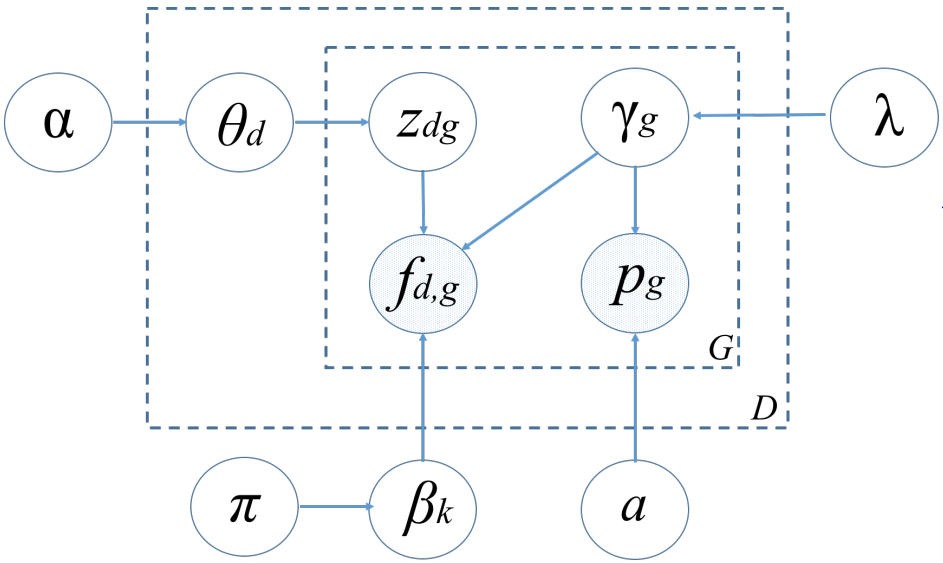
Effort 1: Joint decomposition on integrating Omics-data as side information to strongly infer gene-disease pair





Mathematical efforts to improve knowledge discovery

Effort 2: Variational inference on modeling function change for mutated gene



$$f_{dg} \sim \begin{cases} Multi(\beta_{z_{dg}}), & \text{if } \gamma_{dg} = 1 \\ N(0, \delta_e^2 I), & \text{if } \gamma_{dg} = 0 \end{cases}$$

$$p_{dg} \sim \begin{cases} Beta(a, 1), & \text{if } \gamma_{dg} = 1 \\ U(0, 1), & \text{if } \gamma_{dg} = 0 \end{cases}$$

$$\gamma_{dg} \sim Bernoulli(\lambda)$$

$$z_{dg} \sim Categorical(\theta_d), \quad z_{dg} \in \{1, 2, 3, 4\}$$

$$\theta_d \sim Dir(\alpha)$$

$$\beta_k \sim Dir(\pi), \quad k \in \{1, 2, 3, 4\}$$

Outline

1 Fundamentals 3

2 AGAC corpus design and basic evaluation 8

3 AGAC and tensor decomposition for novel link discovery 15

4 Mathematical efforts to improve the current knowledge discovery scheme 23

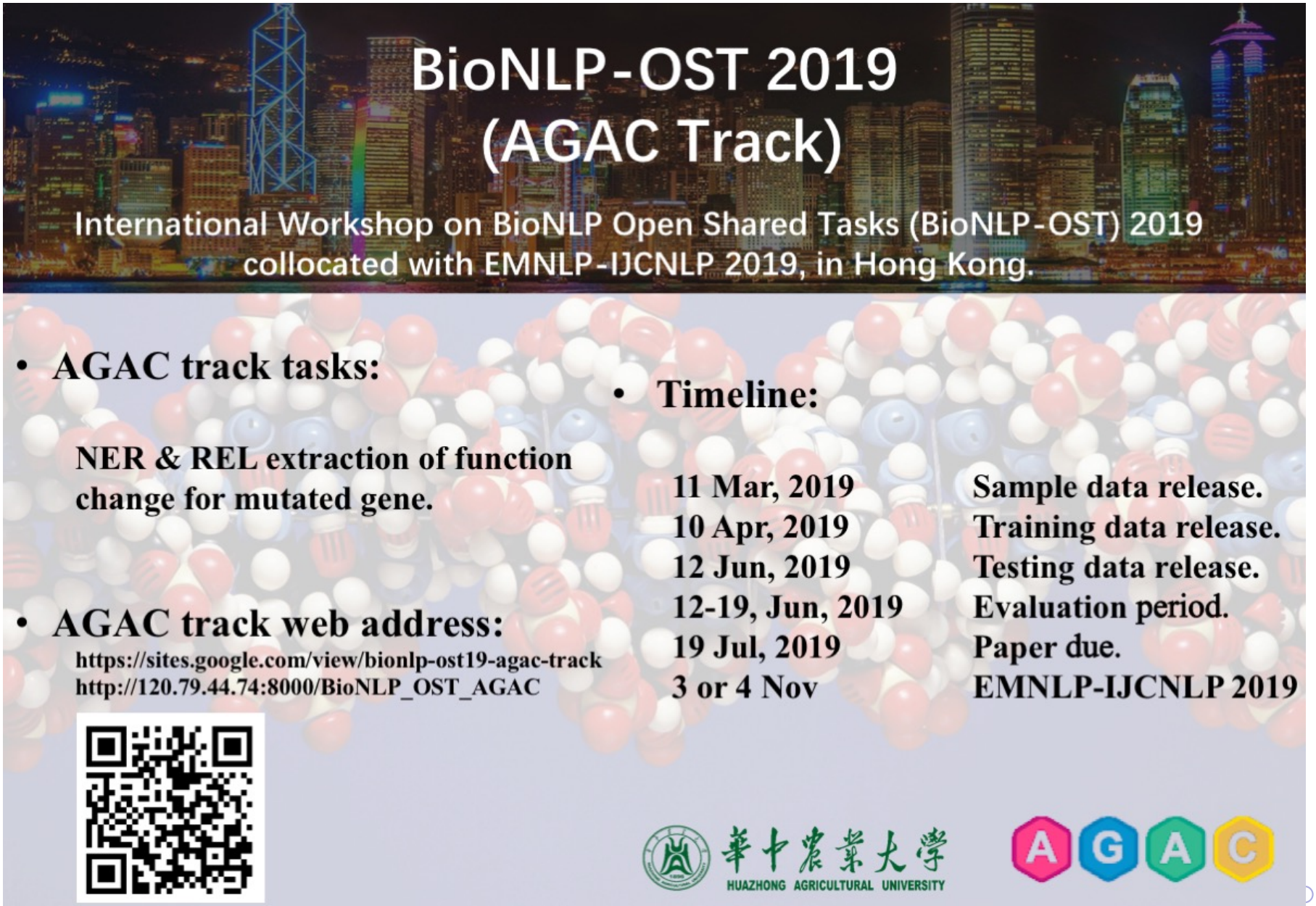
5 Conclusion and AGAC track in BioNLP OST 2019 26

Conclusion

Conclusion of this research:

1. Integration of AGAC labels into feature engineering is meaningful, i.e., to improve GOF/LOF recognition for OMIM text
2. Tensor decomposition works for hidden link discovery between mutated gene and disease
3. A large-scale "Agonist/LOF" or "Antagonist/GOF" pairs searching is possible
4. AGAC corpus is a potential addition to propel discovery of new drug

AGAC track in BioNLP OST 2019



Acknowledgements

AGAC annotation:
Yuxing Wang, Mina Gachloo, Yuxing Ren, Shangzhou Nie, Xinzhi Yao, Shuguang Wang (HZAU)

Corpus design and discussion:
Kevin Bretonnel Cohen (UCDAMC)
Jin-Dong Kim (DBCLS)

Annotatiion platform: PubAnnotation, PubDictionary (DBCLS)

Knowledge inference:
Kaiyin Zhou, Sheng Zhang, Qi Luo (HZAU)

Acknowledgements

Thank you!