華中農業大學
HUAZHONG AGRICULTURAL UNIVERSITY

An Overview of the Active Gene Annotation Corpus (AGAC) and the BioNLP OST 2019 AGAC Track Tasks

Yuxing Wang
wang-yuxing@foxmail.com
College of Informatics, Huazhong Agricultural University
China

AGAC Corpus

- Design logic
- Statistics
- AGAC features

AGAC Track

- Task setting
- Challenges
- Participants performance
- Conclusion

AGAC Applications

- Pharmacological hypothesis
- DAX1 example
- Epilepsy example

華中農業大學
HUAZHONG AGRICULTURAL UNIVERSITY

AGAC Corpus

AGAC Corpus – Design logic

華中農業大學
HUAZHONG AGRICULTURAL UNIVERSITY

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... 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.

https://omim.org/entry/300473

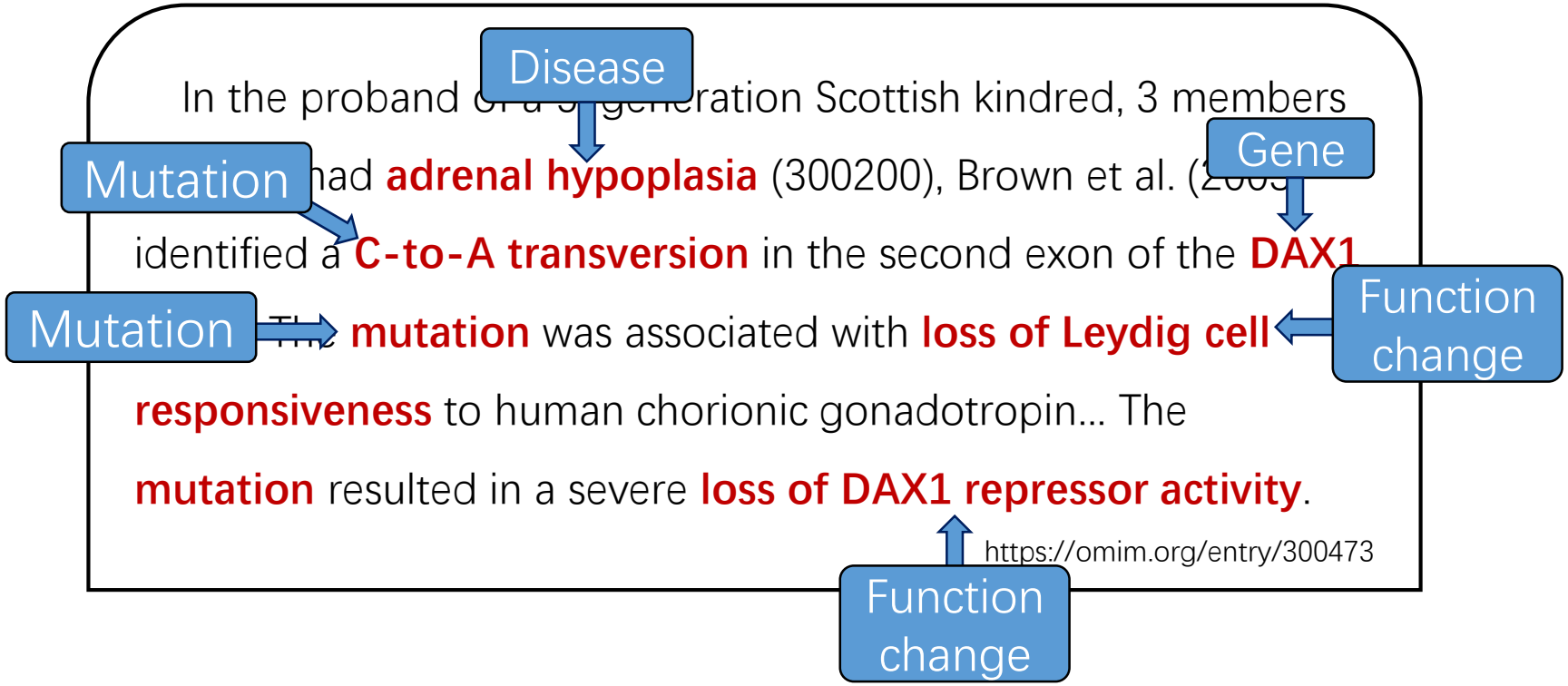
AGAC Corpus – Design logic

華中農業大學
HUAZHONG AGRICULTURAL UNIVERSITY

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... 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**.

https://omim.org/entry/300473

AGAC Corpus – Design logic



AGAC Corpus – Design logic

華中農業大學
HUAZHONG AGRICULTURAL UNIVERSITY

- Definition of Active Gene Annotation Corpus (AGAC) :

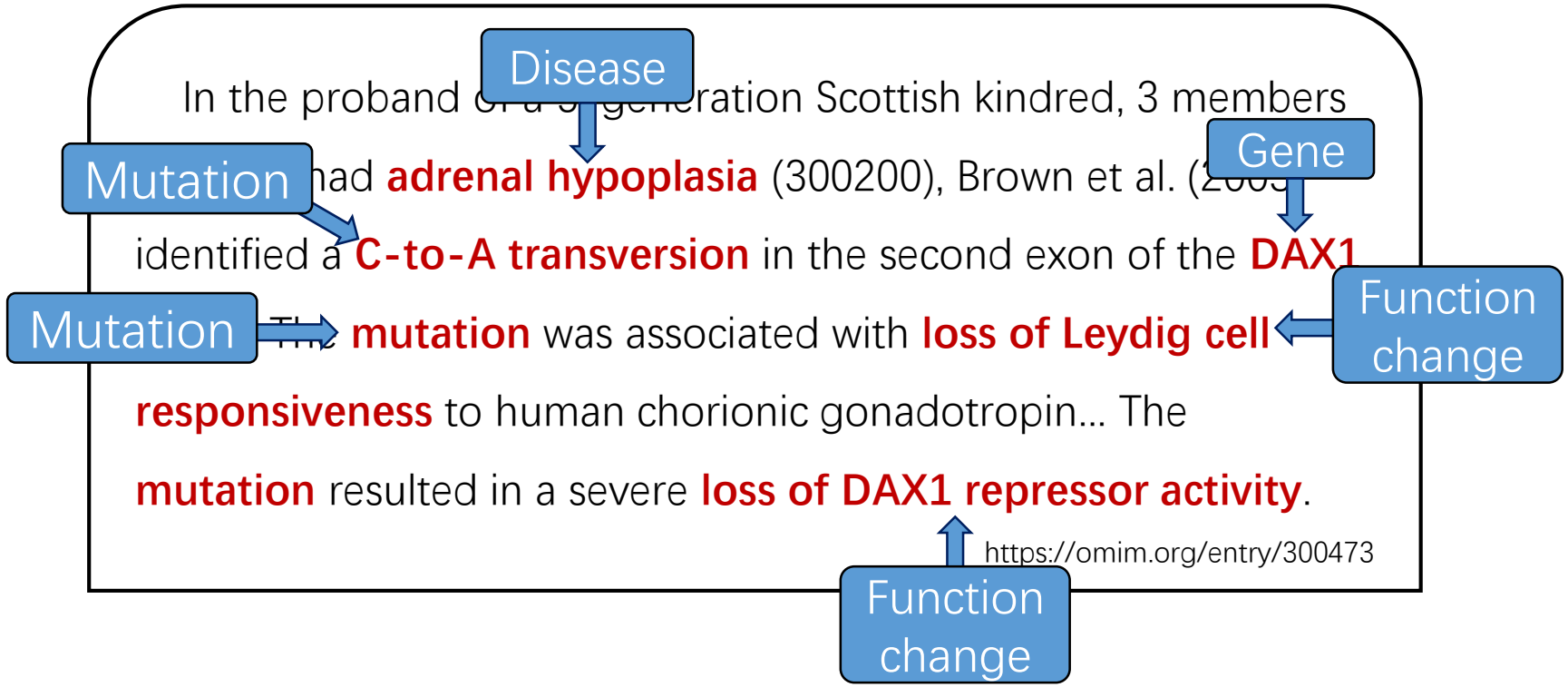
AGAC focuses on the mutations and the biological function changes raised by them. The mutations are classified as two types, Loss of Function (LOF) and Gain of Function (GOF), based on the effects of the function changes.

AGAC Corpus – Design logic

華中農業大學 HUAZHONG AGRICULTURAL UNIVERSITY

- Definition of Active Gene Annotation Corpus (AGAC) :

AGAC focuses on the mutations and the biological function changes raised by them. The mutations are classified as two types, Loss of Function (LOF) and Gain of Function (GOF), based on the effects of the function changes.

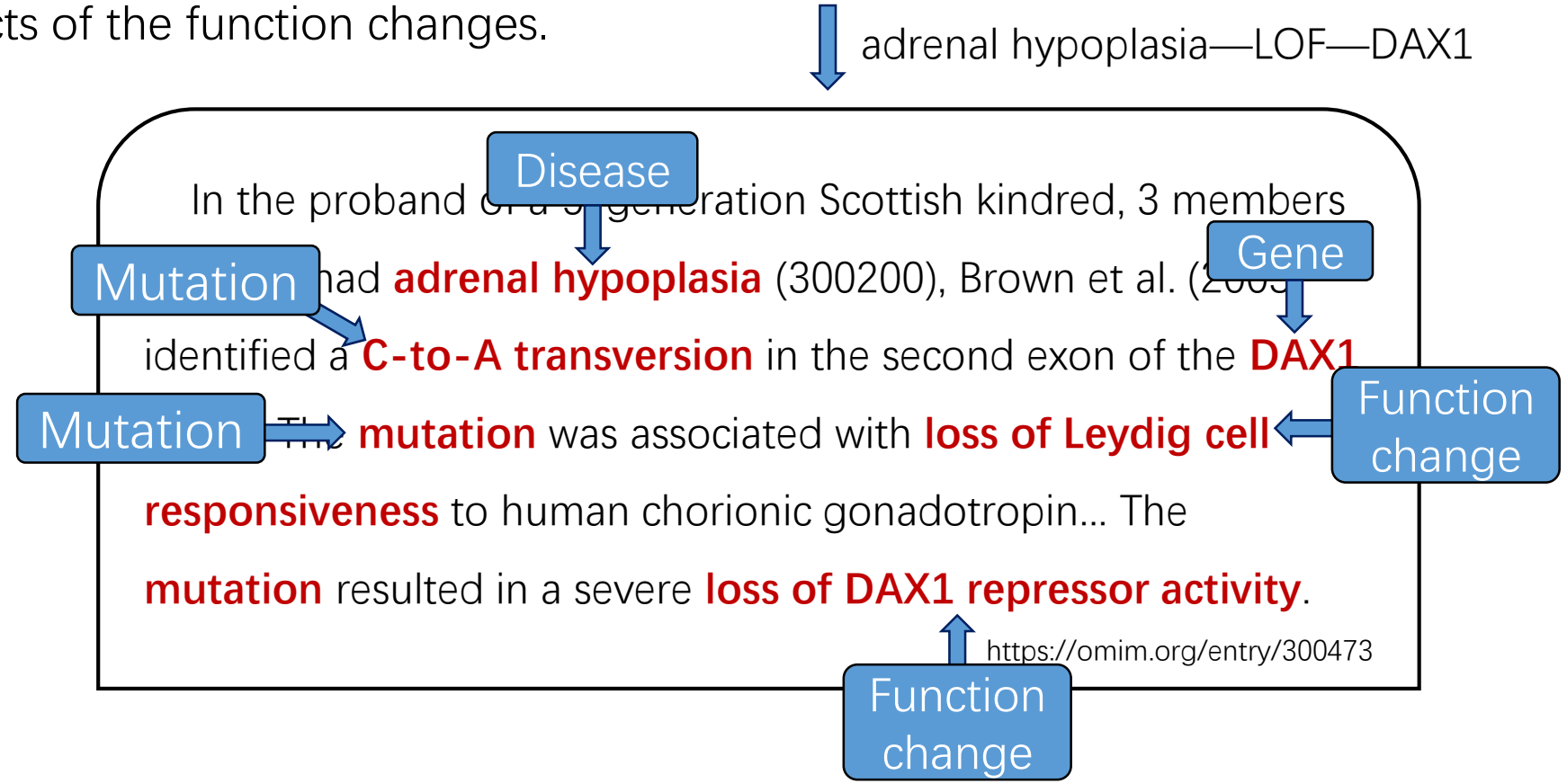


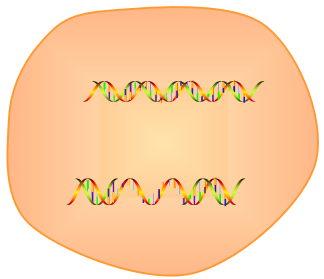
AGAC Corpus – Design logic

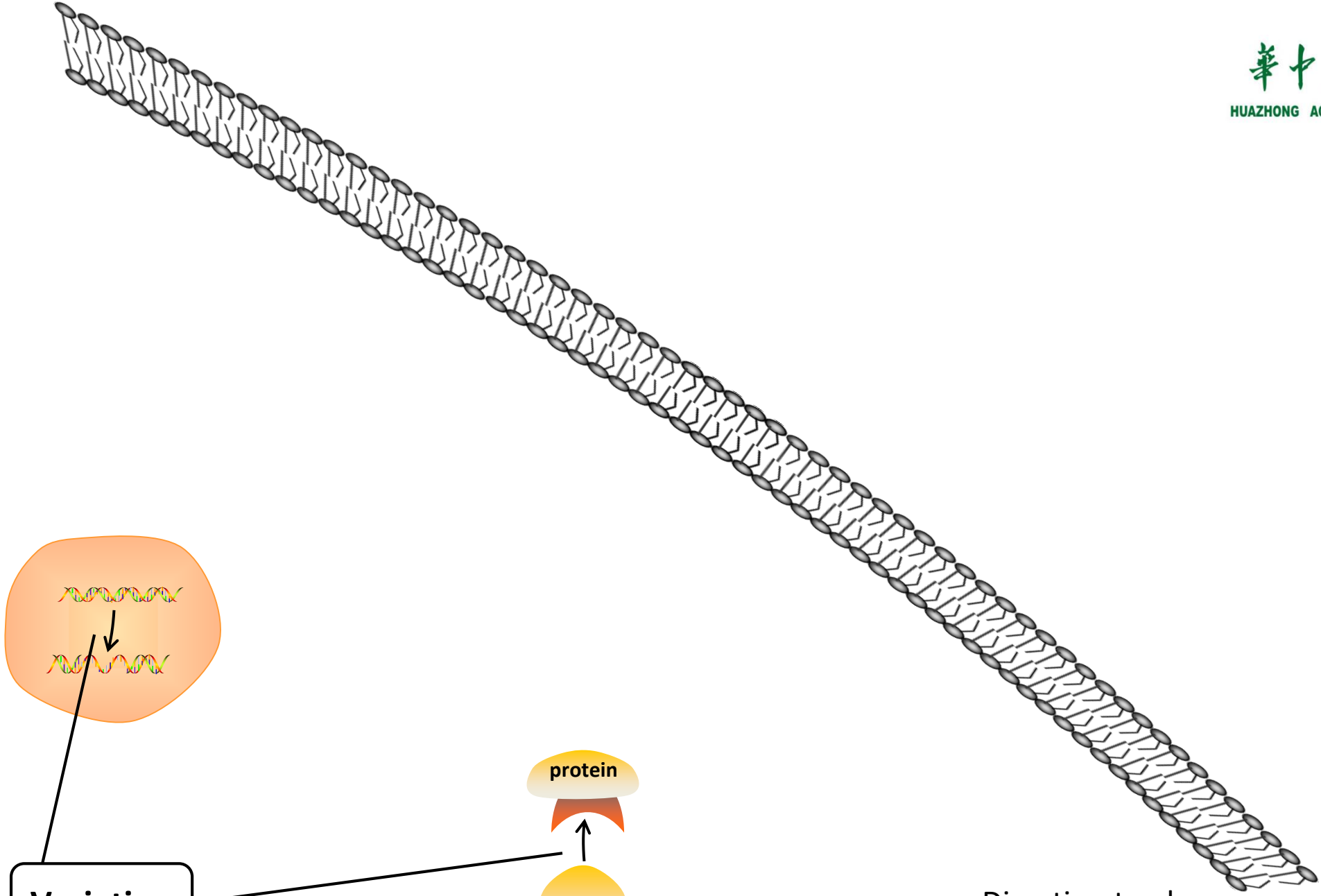
華中農業大學 HUAZHONG AGRICULTURAL UNIVERSITY

- Definition of Active Gene Annotation Corpus (AGAC) :

AGAC focuses on the mutations and the biological function changes raised by them. The mutations are classified as two types, Loss of Function (LOF) and Gain of Function (GOF), based on the effects of the function changes.



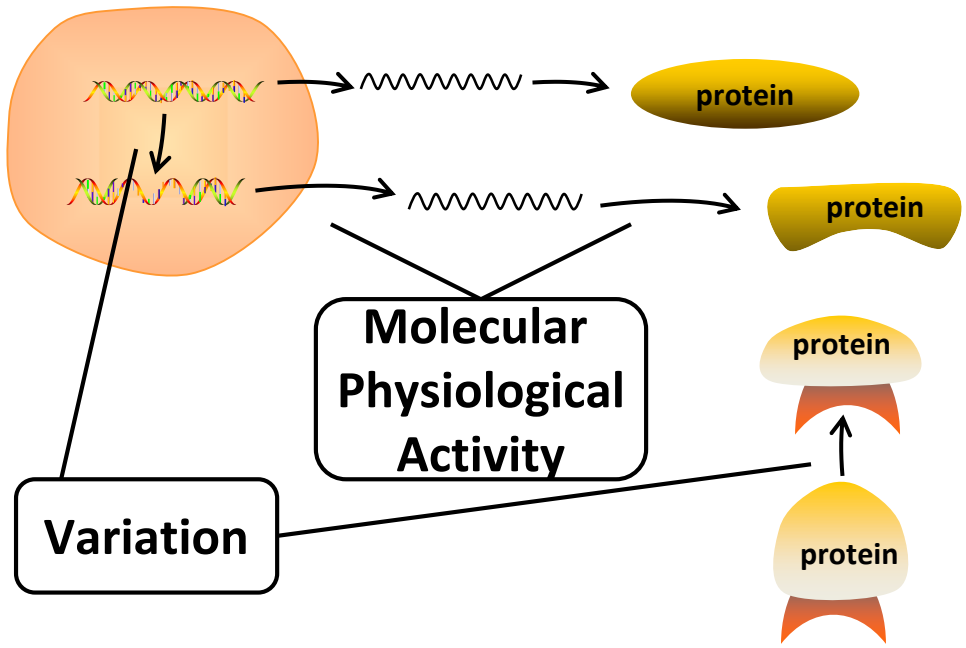




Variation

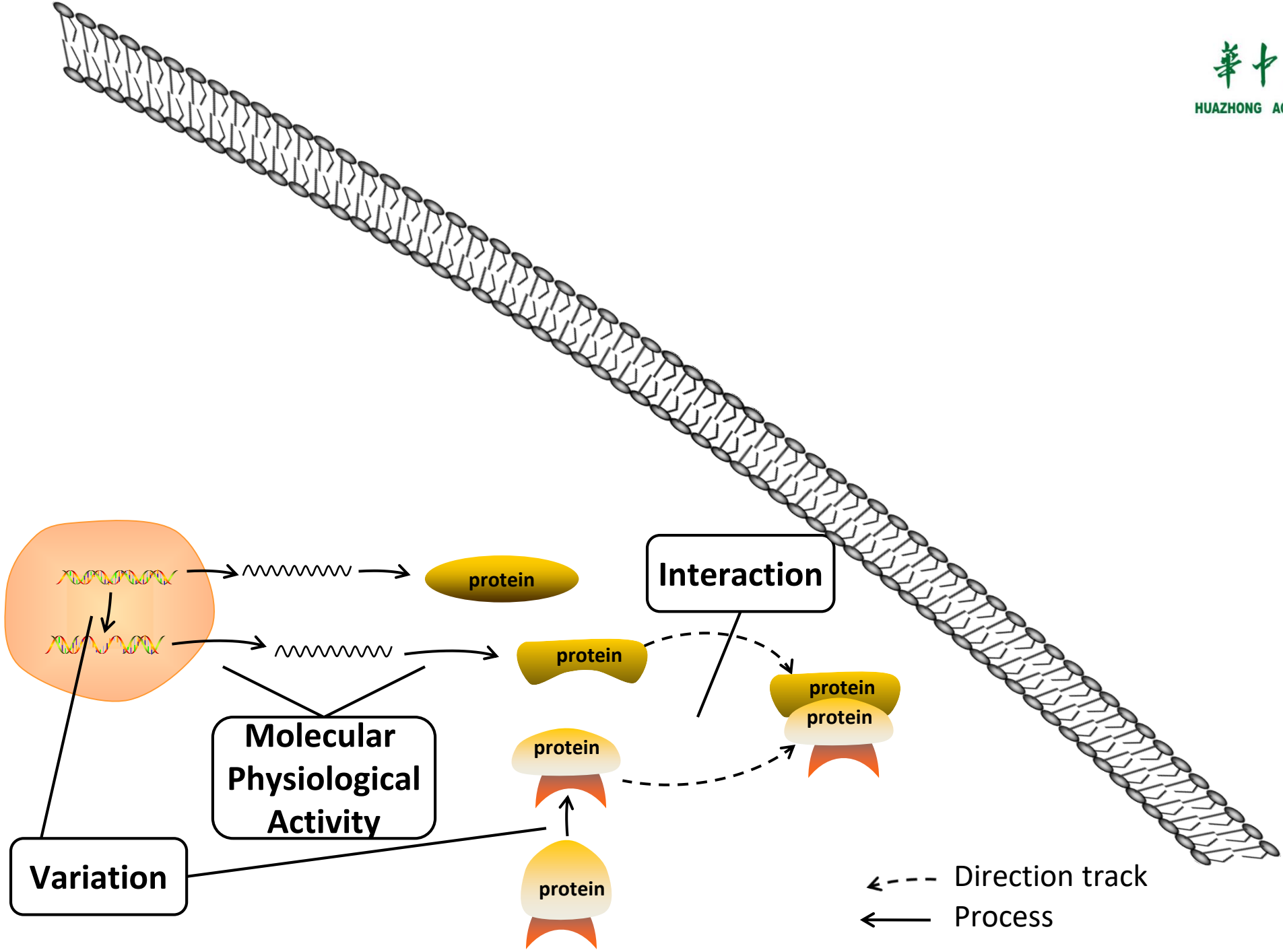
Direction track

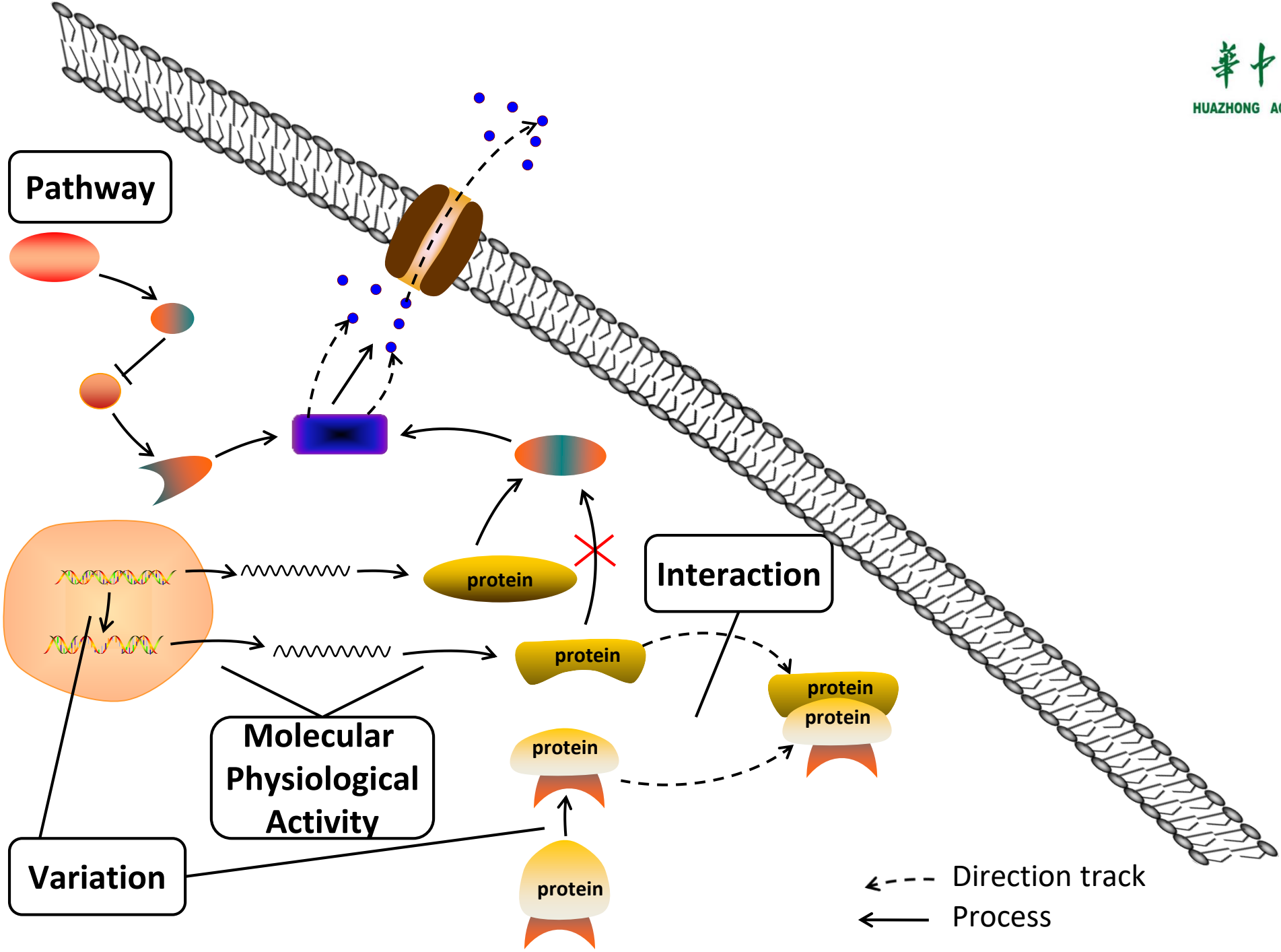
Process

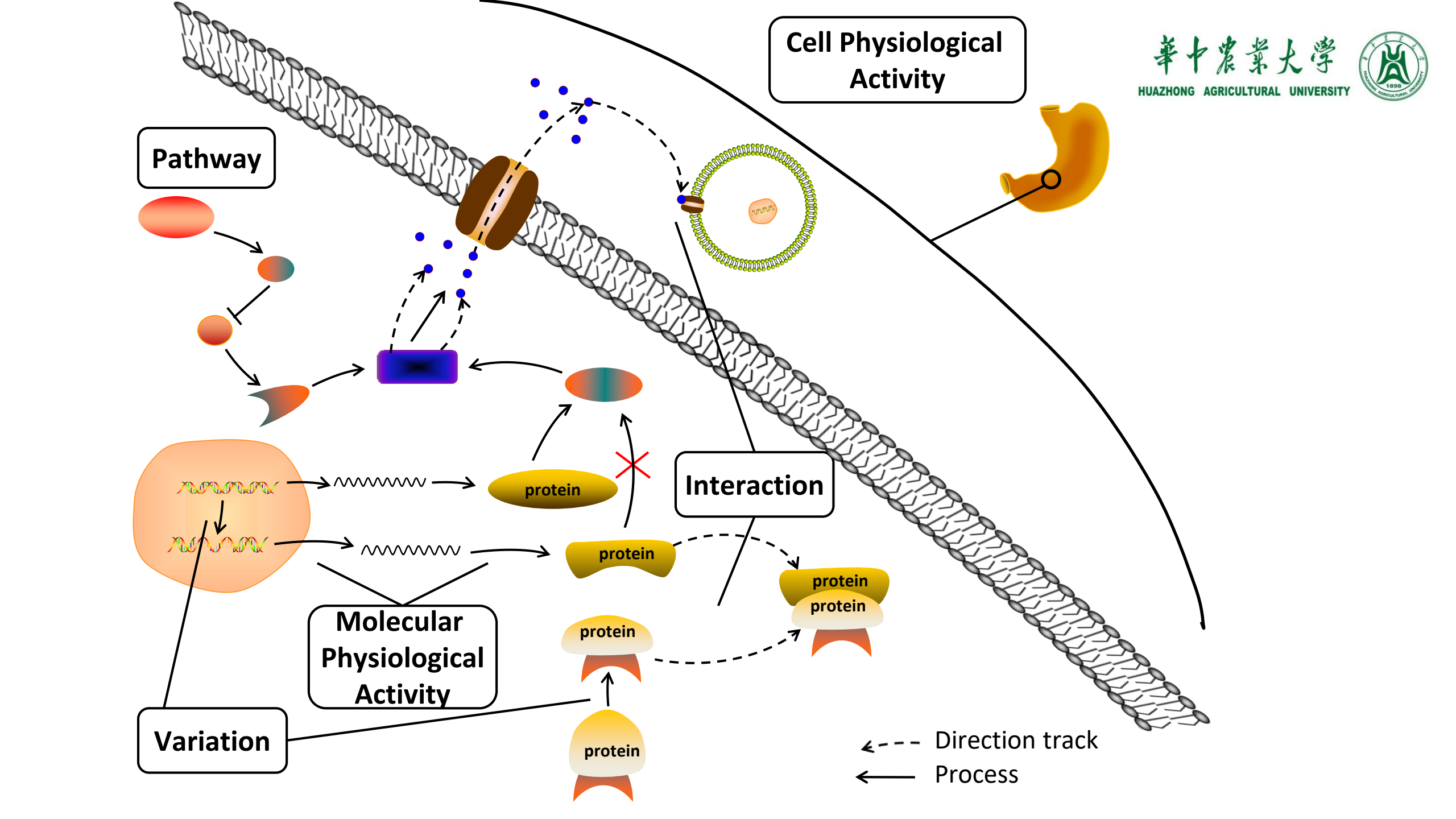


Direction track

Process


protein
protein
protein
protein
protein
protein
Interaction
Molecular
Physiological
Activity
Variation
Direction track
Process


Pathway
protein
Interaction
protein
protein
protein
protein
protein
Molecular
Physiological
Activity
Variation
Direction track
Process


Cell Physiological Activity
華中農業大學
HUAZHONG AGRICULTURAL UNIVERSITY
Pathway
protein
Interaction
protein
protein
protein
protein
protein
Molecular Physiological Activity
Variation
Direction track
Process

AGAC Corpus – Design logic

華中農業大學 HUAZHONG AGRICULTURAL UNIVERSITY

- **Bio-concept labels**
  - **Variation**: Abnormal in DNA, RNA, protein sequences or structures.
  - **Molecular physiological activity**: Activities at the molecule level including gene expression and other molecular activity.
  - **Interaction**: The associations among molecules or cells.
  - **Pathway**: Pathway like signal transduction pathway and metabolic pathway.
  - **Cell physiological activity**: Activities at cell level, including cell responsiveness and the development and growth of cells or organs.
- **Regulatory concept labels**
  - **Positive regulation**: clue word or phrase that meant gain of function.
  - **Negative regulation**: clue word or phrase that meant loss of function.
  - **Regulation**: neutral clue word or phrase which meant no loss or gain.
- **Other Entities**
  - **Disease, Gene, Protein, Enzyme**

AGAC Corpus – Design logic

華中農業大學 HUAZHONG AGRICULTURAL UNIVERSITY

- **Bio-concept labels**
  - **Variation**: Abnormal in DNA, RNA, protein sequences or structures.
  - **Molecular physiological activity**: Activities at the molecule level including gene expression and other molecular activity.
  - **Interaction**: The associations among molecules or cells.
  - **Pathway**: Pathway like signal transduction pathway and metabolic pathway.
  - **Cell physiological activity**: Activities at cell level, including cell responsiveness and the development and growth of cells or organs.
- **Regulatory concept labels**
  - **Positive regulation**: clue word or phrase that meant gain of function.
  - **Negative regulation**: clue word or phrase that meant loss of function.
  - **Regulation**: neutral clue word or phrase which meant no loss or gain.
- **Other Entities**
  - **Disease, Gene, Protein, Enzyme**

- **Thematic Relations**
  - **ThemeOf:** a theme of an event (or a regulatory named entities) is the object which undergoes a changes of its state due to the event.
  - **CauseOf:** a cause of an event (or a regulatory named entities) is the object which leads the event to happen.

AGAC Corpus – Statistics

華中農業大學 HUAZHONG AGRICULTURAL UNIVERSITY

	Total	Training set	Test set
# of Abstracts	**500**	**250**	**250**
# of Sentences	**5,080**	**2,534**	**2,546**
# of Named entities	**5,741**	**3,317**	**2,424**
.Bio-concept Named Entities	**2,274**	**1,428**	**846**
Var (Variation)	1,304	735	569
MPA (Molecular Physiological Activity)	618	418	200
Interaction	35	28	7
Pathway	38	24	14
CPA (Cell Physiological Activity)	279	223	56
.Regulatory Named Entities	**1,514**	**905**	**609**
Regulation	613	215	398
Positive Regulation	406	323	83
Negative Regulation	495	367	128
.Other Entities	**1,953**	**984**	**969**
Disease	751	336	415
Gene	1,004	529	475
Protein	150	90	60
Enzyme	48	29	19
# of Thematic roles	**4,677**	**2,729**	**1,948**
ThemeOf	2,986	1,698	1,288
ThemeOf (Intra/inter sentential)	(2910/76)	(1657/41)	(1253/35)
CauseOf	1,691	1,031	660
CauseOf (Intra/inter sentential)	(1581/110)	(961/70)	(620/40)

- We collected 500 abstracts from PubMed by using the MeSH terms “Mutation/physiopathology” and “Genetic Disease”
- The amount of the different concept labels vary from each other.

AGAC Corpus – AGAC features

- Imbalanced data
  - It is naturally because the mutation reports tends to describe the process that in molecule level.
  - It doesn't mean that the other labels are not important.

AGAC Corpus – AGAC features

華中農業大學 HUAZHONG AGRICULTURAL UNIVERSITY

- Imbalanced data
  - It is naturally because the mutation reports tends to describe the process that in molecule level.
  - It doesn't mean that the other labels are not important.
- Selective annotation
  - The annotation need the knowledge of domain experts.
  - The words will not be annotated if the sentence doesn't describe mutation.
  - "The expression of BRCA gene was decreased."
  - "The expression of mutated BRCA gene was decreased."

AGAC Corpus – AGAC features

華中農業大學 HUAZHONG AGRICULTURAL UNIVERSITY

- Imbalanced data
  - It is naturally because the mutation reports tends to describe the process that in molecule level.
  - It doesn't mean that the other labels are not important.
- Selective annotation
  - The annotation need the knowledge of domain experts.
  - The words will not be annotated if the sentence doesn't describe mutation.
  - "The expression of BRCA gene was decreased." ✗
  - "The expression of mutated BRCA gene was decreased." √

AGAC Corpus – AGAC features

華中農業大學 HUAZHONG AGRICULTURAL UNIVERSITY

- Imbalanced data
  - It is naturally because the mutation reports tends to describe the process that in molecule level.
  - It doesn't mean that the other labels are not important.
- Selective annotation
  - The annotation need the knowledge of domain experts.
  - The words will not be annotated if the sentence doesn't describe mutation.
  - "The expression of BRCA gene was decreased." ✗
  - "The expression of mutated BRCA gene was decreased." √
- Latent topic annotation
  - LOF and GOF context of a gene-disease association may not be directly visible from the text.
  - The LOF and GOF topic should be inferred from the named entity annotations and the thematic role annotations.

華中農業大學
HUAZHONG AGRICULTURAL UNIVERSITY

AGAC Track[1]

[1] Yuxing Wang, Kaiyin Zhou, Mina Gachloo, **Jingbo Xia*.** An Overview of the Active Gene Annotation Corpus and the BioNLP OST 2019 AGAC Track Tasks. BioNLP Open Shared Task 2019, Hong Kong.

AGAC Track – Task setting

華中農業大學 HUAZHONG AGRICULTURAL UNIVERSITY

- Task 1. NER

To recognize named entities appearing in given texts, and to assign them their entity class.

"denotations":

[…, {"id":"T4","span":{"begin":771,"end":778},"obj":"Protein"},

{"id":"T5","span":{"begin":779,"end":789},"obj":"NegReg"},

{"id":"T6","span":{"begin":790,"end":794},"obj":"Var"},

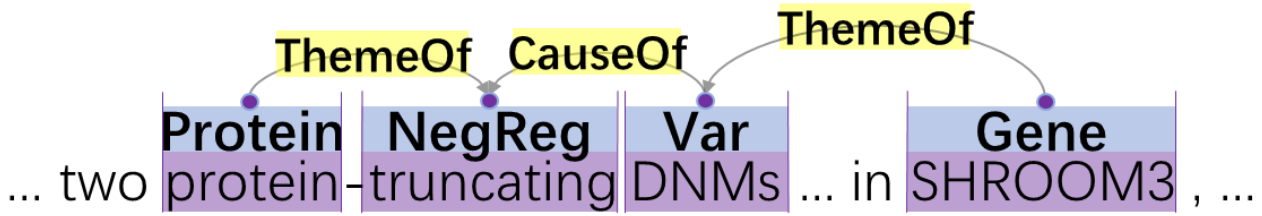
{"id":"T3","span":{"begin":823,"end":830},"obj":"Gene"}, …]

AGAC Track – Task setting

華中農業大學 HUAZHONG AGRICULTURAL UNIVERSITY

- Task 2. Thematic relation identification

To identify the thematic relation, ThemeOf, CauseOf, between named entities.



"relations":

[…, {"id":"R1","pred":"ThemeOf","subj":"T3","obj":"T6"},

{"id":"R2","pred":"CauseOf","subj":"T6","obj":"T5"},

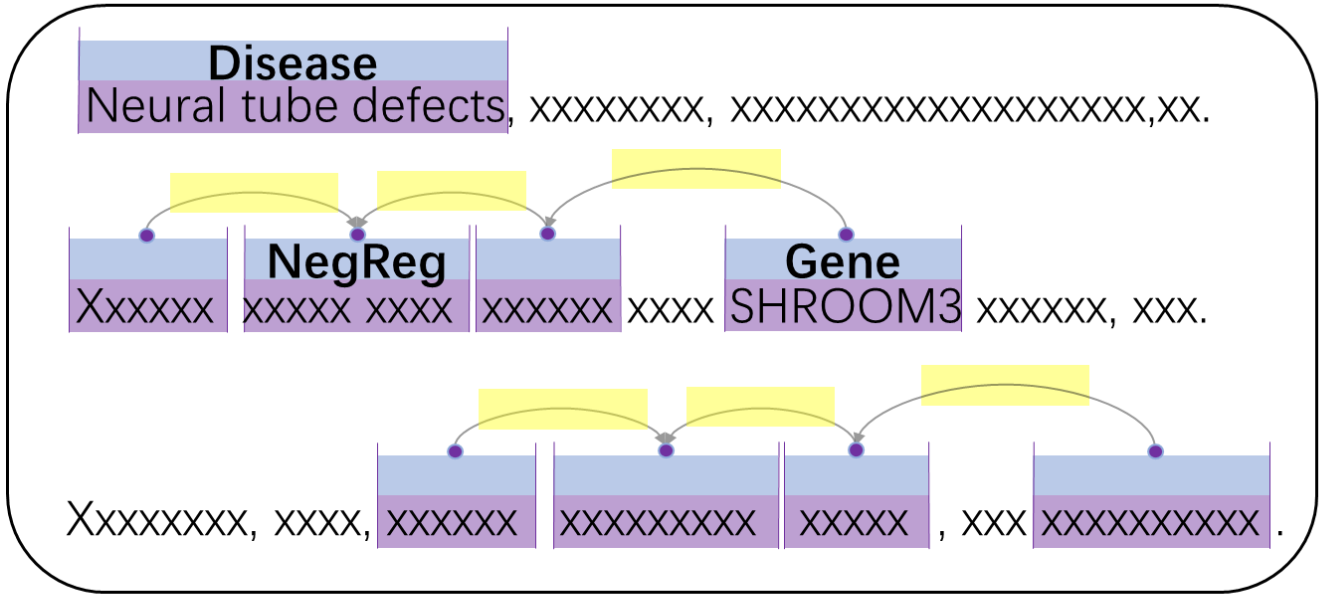
{"id":"R3","pred":"ThemeOf","subj":"T4","obj":"T5"}, …]

AGAC Track – Task setting

華中農業大學
HUAZHONG AGRICULTURAL UNIVERSITY

- Task 3. Mutation-disease knowledge discovery

To extract the triples of a gene, a function change, and a disease.



25805808; SHROOM3; LOF; Neural tube defects

AGAC Track – Participants Performance

華中農業大學 HUAZHONG AGRICULTURAL UNIVERSITY

➢ Task 1

	Participants	Precision	Recall	F-score	Main NLP techniques
1st	DX-HITSZ	0.63	0.56	0.60	Bert, joint learning
*	Baseline	0.50	0.51	0.50	Bert, joint learning
2nd	Zheng-UMASS	0.36	0.59	0.45	Bert, CNN, Bi-LSTM
3rd	YaXXX-SiXXX/LMX	0.55	0.28	0.37	CRF, Bi-LSTM
4th	DJDL-HZAU	0.16	0.25	0.20	CRF

*: Baseline.

➢ Task 2

	Participants	Precision	Recall	F-score	Main NLP techniques
1st	Zheng-UMASS	0.40	0.31	0.35	Bert, CNN, Bi-LSTM
2nd	DX-HITSZ	0.61	0.16	0.25	Bert, joint learning
3rd	YaXXX-SiXXX/LMX	0.05	0.02	0.03	SVM

➢ Task 3

	Participants	Precision	Recall	F-score	Main NLP techniques
*	Baseline	0.72	0.59	0.65	Bert, joint learning
L	Ashok-BenevolentAI	0.26	0.20	0.23	Bert

*: Baseline

L: Late submission.

AGAC Track – Conclusions

華中農業大學 HUAZHONG AGRICULTURAL UNIVERSITY

- Various NLP technologies are applied on AGAC.
  - Kernel-based linear classification model, SVM; modern neural network models, CNN and Bi-LSTM; pre-trained language representation model, Bert, which is very popular.
  - The result shows that the modern network methods are better than classical machine learning methods in AGAC track.
- The annotations in AGAC are helpful to do the LOF/GOF information extraction.
  - See the results in Task 3.

華中農業大學
HUAZHONG AGRICULTURAL UNIVERSITY

AGAC Applications[1]

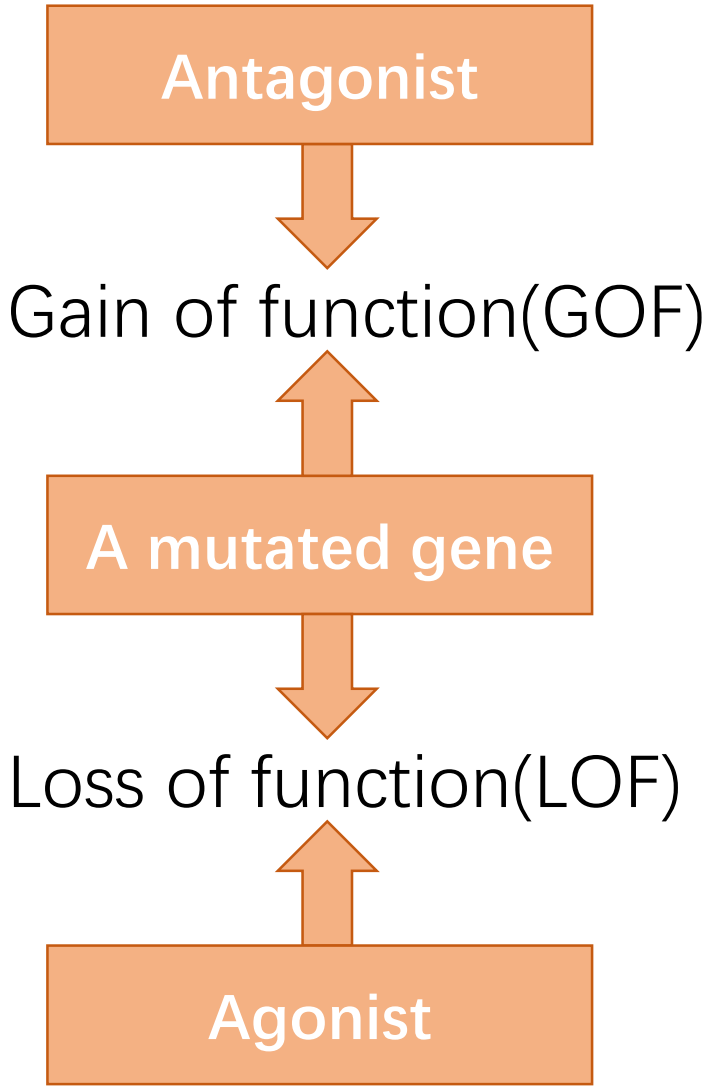
[1] Yuxing Wang, Kaiyin Zhou, Jin-Dong Kim, Kevin Cohen, Mina Gachloo, Yuxin Ren, Shanghui Nie, Xuan Qin, Panzhong Lu, **Jingbo Xia***. An Active Gene Annotation Corpus and Its Application on Anti-epilepsy Drug Discovery. BIBM 2019: International Conference on Bioinformatics & Biomedicine, San Diego, U.S, Nov, 2019.

AGAC Applications – pharmacological hypothesis

華中農業大學 HUAZHONG AGRICULTURAL UNIVERSITY

- LOF-agonist/GOF-antagonist hypothesis[1]:

For a given disease caused by driven gene with Loss of function (LOF) or Gain of function (GOF) mutation, an targeted antagonist/agonist drug is the candidate drug to this disease.



[1] Zhong-Yi Wang and Hong-Yu Zhang. 2013. Rational drug repositioning by medical genetics. *Nature biotechnology*, 31(12):1080.

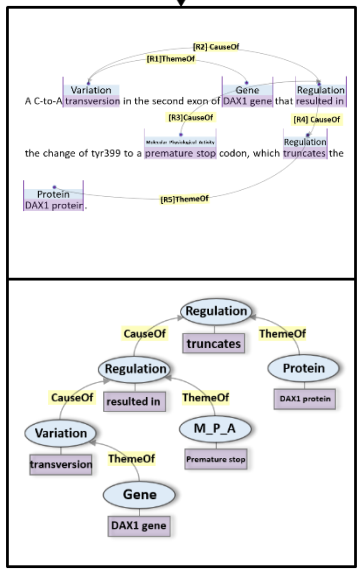
AGAC Applications – DAX1 example

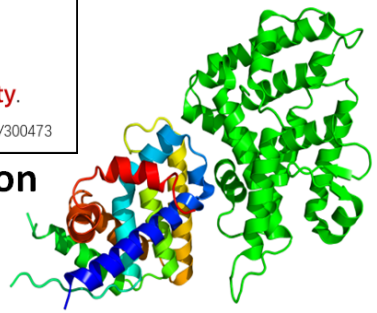
華中農業大學 HUAZHONG AGRICULTURAL UNIVERSITY

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... 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**.

https://omim.org/entry/300473

AGAC annotation





DAX1

LOF

Agonist

drug

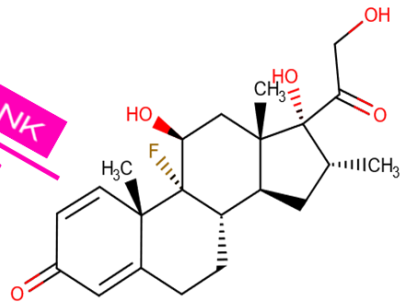
Inference

cure

Adrenal hypoplasia

DRUGBANK

Stimulator

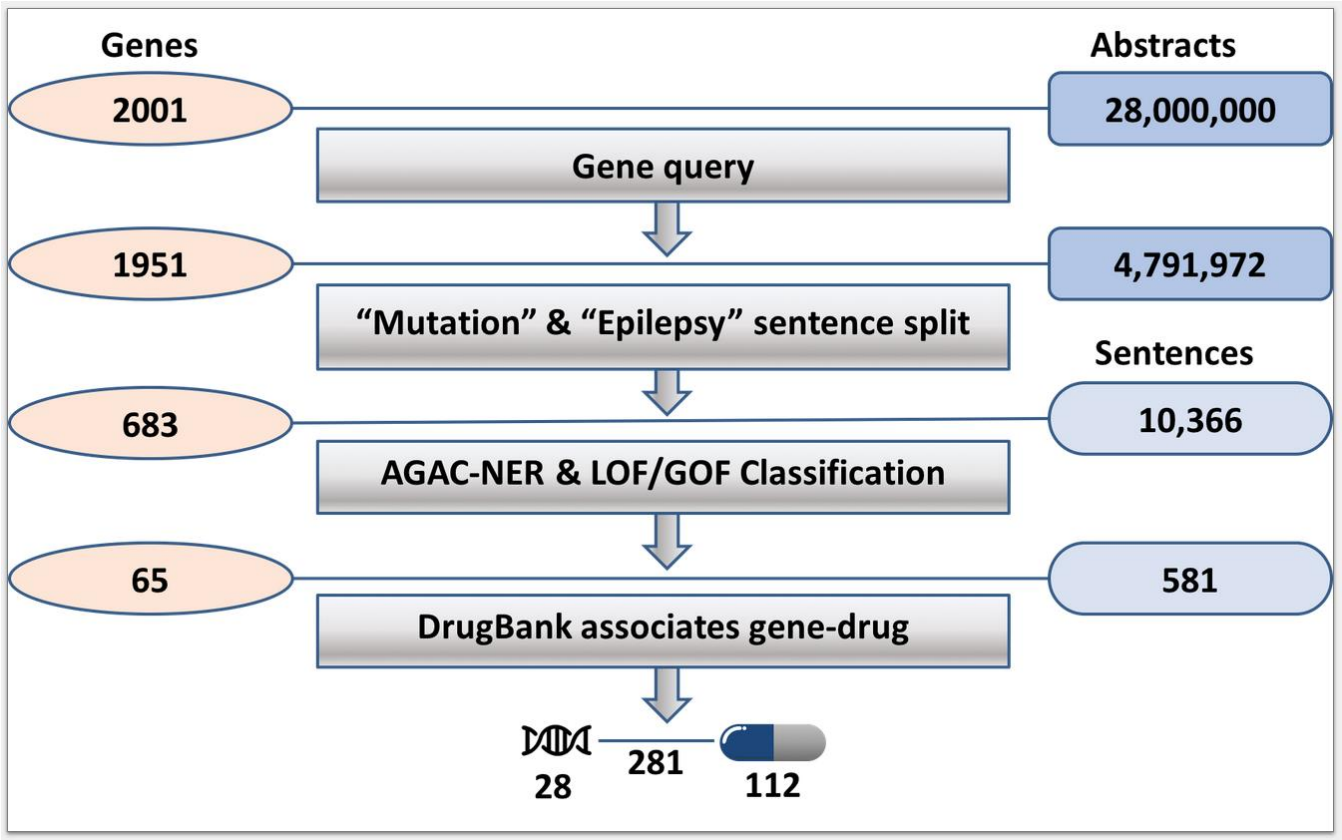


Dexamethone

- In the DAX1 example, we inferred that LOF mutation in DAX1 resulted in adrenal hypoplasia.
- Based on the pharmacological hypothesis, the agonist drug targeting to DAX1 is the candidate drug for adrenal hypoplasia.
- To verify this, we found a drug in DrugBank that targets to DAX1 and acts as an agonist to active DAX1.

AGAC Applications – Epilepsy example

華中農業大學
HUAZHONG AGRICULTURAL UNIVERSITY



① PubMed abstracts retrieval with epilepsy gene query.
② Abstracts filtering by keyword matching and sentence splitting.
③ Acquisition of gene-function change pairs by performing AGAC-based sequence labeling and LOF/GOF classification.
④ Prediction of novel anti-epilepsy drug by incorporating DrugBank gene-drug associations information.

AGAC Applications – Epilepsy example

華中農業大學 HUAZHONG AGRICULTURAL UNIVERSITY

- Among the 112 predicted drugs, 30 of them are recorded in Databases (DrugBank, TTD, Clinical Trails, Malacards) as anti-epilepsy drugs.
- The rest 82 drugs are considered as the potential drugs of epilepsy.
- Among the 82 drugs, 10 of them are multi-target drugs.

	Drug*	Action	Target	Function change	Hypoth**
Increase the open probability of ion channel.	**Oxazepam, Temazepam**	potentiator ↑ potentiator ↑ potentiator ↑	GABRA1 GABRG2 GABRB3	LOF [31] ↓ LOF [32] ↓ LOF [33] ↓	✓ ✓ ✓
	Halazepam, Prazepam, Zolpidem	potentiator ↑ potentiator ↑	GABRG2 GABRB3	LOF [34] ↓ LOF [35] ↓	✓ ✓
Increase the open time of ion channel.	**Thiamylal**	agonist ↑ inhibtor ↓	GABRA1 KCNJ11	LOF [31] ↓ GOF [36] ↑	✓ ✓
Candidate anti-epilepsy drugs but unreported.	**Fostamatinib**	inhibitor ↓ inhibitor ↓ inhibitor ↓	BRAF FGFR3 MTOR	GOF [37] ↑ GOF [38] ↑ GOF [39] ↑	✓ ✓ ✓
	Glimepiride	inducer ↑ inhibtor ↓	ABCC8 KCNJ11	LOF [40] ↓ GOF [41] ↓	✓ ✓
	Tenocyclidine, Meperidine	antagonist ↓ antagonist ↓	GRIN2A GRIN2B	GOF [42] ↑ GOF [43] ↑	✓ ✓

* Drugs that share same targets are list in one block. The first 6 AED drugs are supported by literature, while the rest 4 drugs are unreported.

** A refers to the action of drug and the function change of target gene matched the pharmacological hypothesis.

Acknowledgements

華中農業大學
HUAZHONG AGRICULTURAL UNIVERSITY

Director:
Jingbo Xia.

Annotators:
Yuxing Wang, Yuxin Ren, Shanghui Nie, Mina Gachloo.

Corpus design and discussion:
Jin-Dong Kim, Kevin B. Cohen.

Knowledge inference:
Kaiyin Zhou, Sheng Zhang, Qi Luo, Xiaohang Ma.