大数据时代下基于人工智能算法的蛋白质功能预测研究

南京农业大学 人工智能学院

汇报人：朱一亨

2024年05月31日

目 录

01 研究背景

02 研究内容

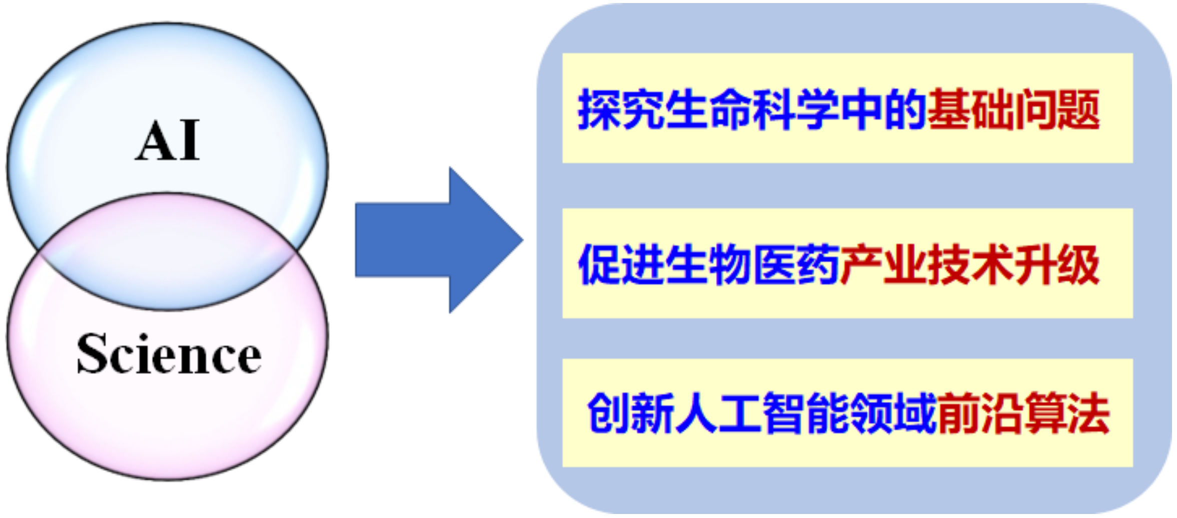
03 未来展望

01 Part one

研究背景

生物信息学

- 生物信息学：生物学 + 统计学 + 计算机科学 $\xrightarrow{\text{揭示}}$ 生物数据中所蕴含的生物学奥秘
- 国家发改委（2022）：首部“十四五” 生物经济发展规划
- 中国科协（2022）：重大前沿科学问题之一



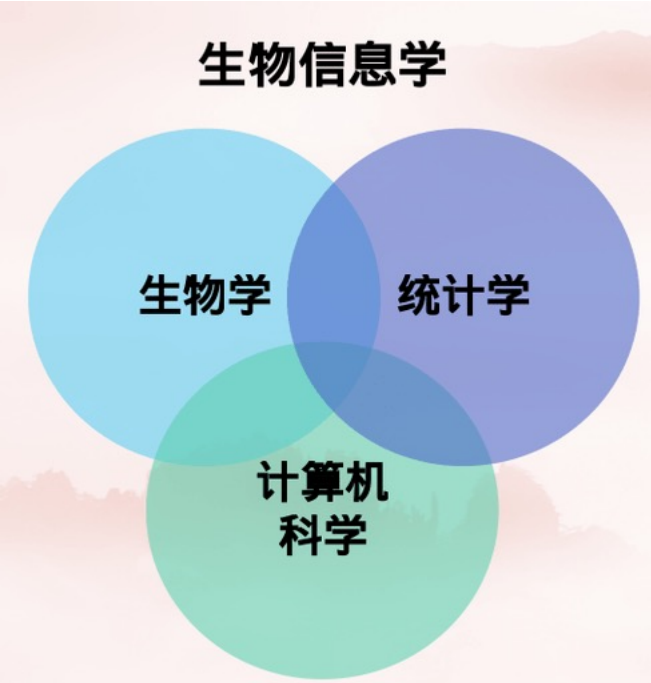
AI

Science

探究生命科学中的基础问题

促进生物医药产业技术升级

创新人工智能领域前沿算法



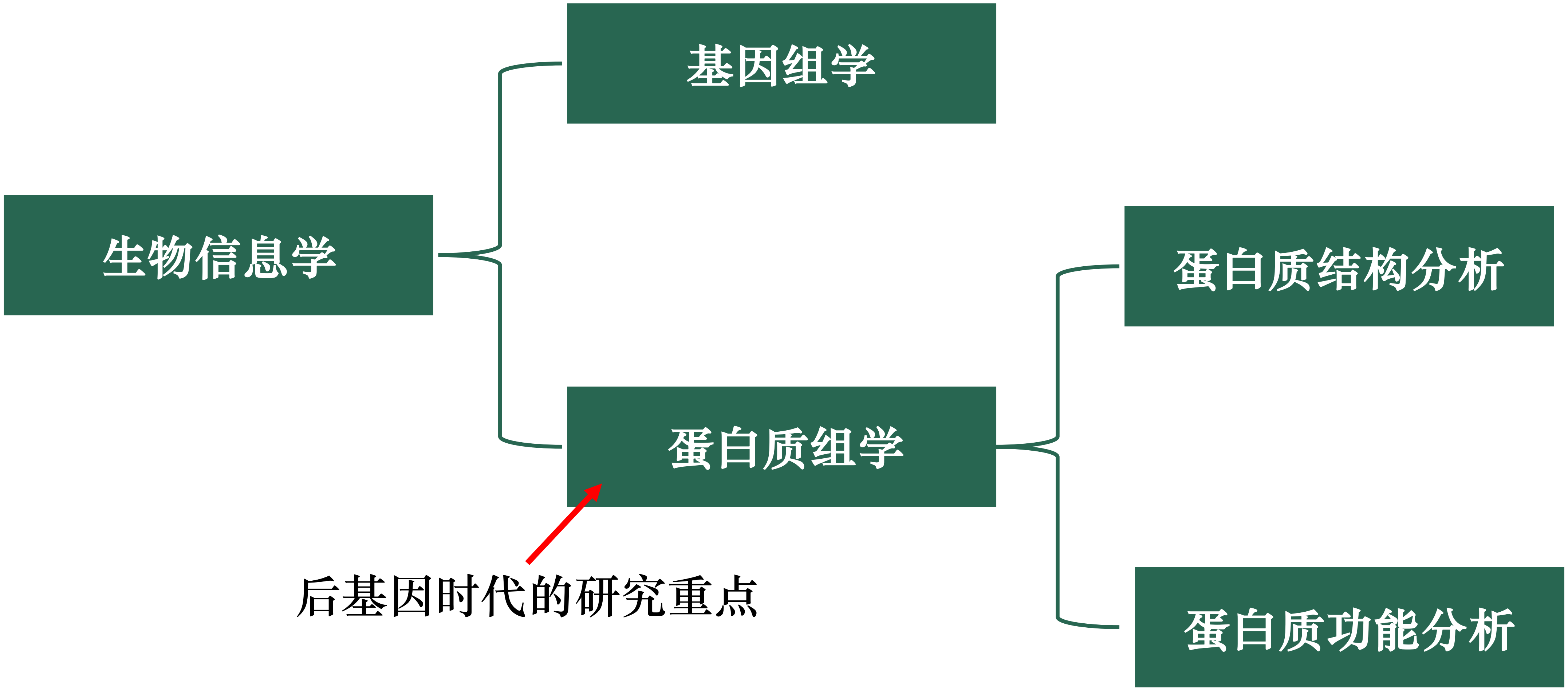
生物信息学

生物学

统计学

计算机科学

- 生物信息学的体系分类



后基因时代的研究重点

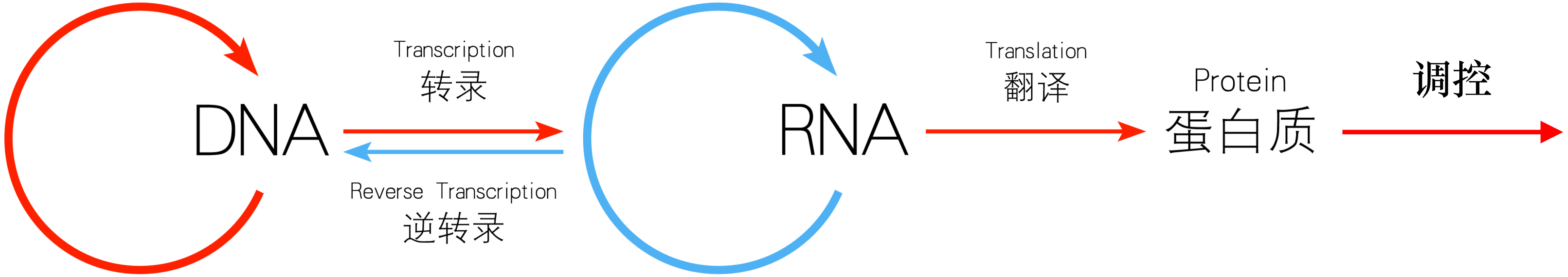
当前的研究课题：蛋白质功能预测

蛋白质基础知识

- 蛋白质是生命现象的物质基础之一

- 蛋白质参与了生物体内几乎全部的生命过程，并发挥着各种重要的功能

- 蛋白质是生命活动的主要承担者



DNA

Transcription
转录

Reverse Transcription
逆转录

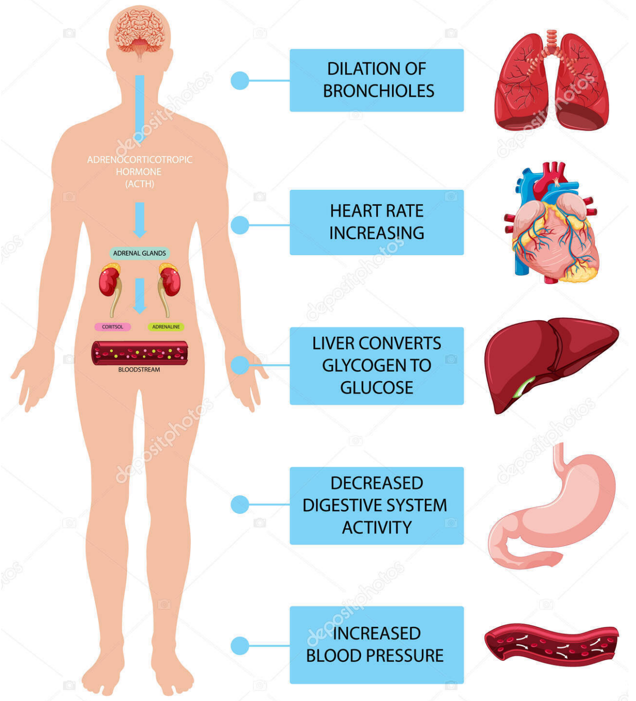
RNA

Translation
翻译

Protein
蛋白质

调控

生物中心法则

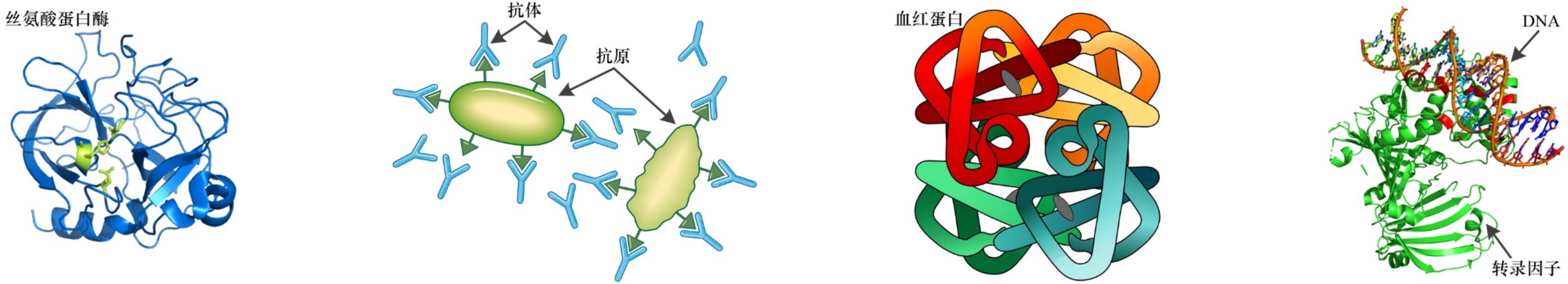


人体代谢活动

蛋白质基础知识

- 识别和分析蛋白质的功能有助于解释各种生命活动现象，并阐明相关疾病的发病机理，进而指导相应的药物设计，以期推动智能医疗的发展。

- 蛋白质功能注释是后基因时代的首要任务之一。

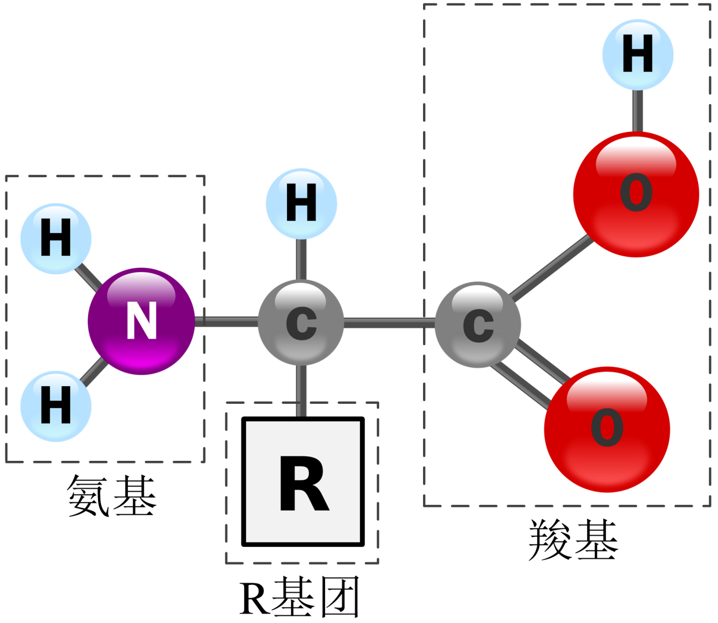


(a) 催化反应 (b) 免疫保护 (c) 运输载体 (d) 基因调控

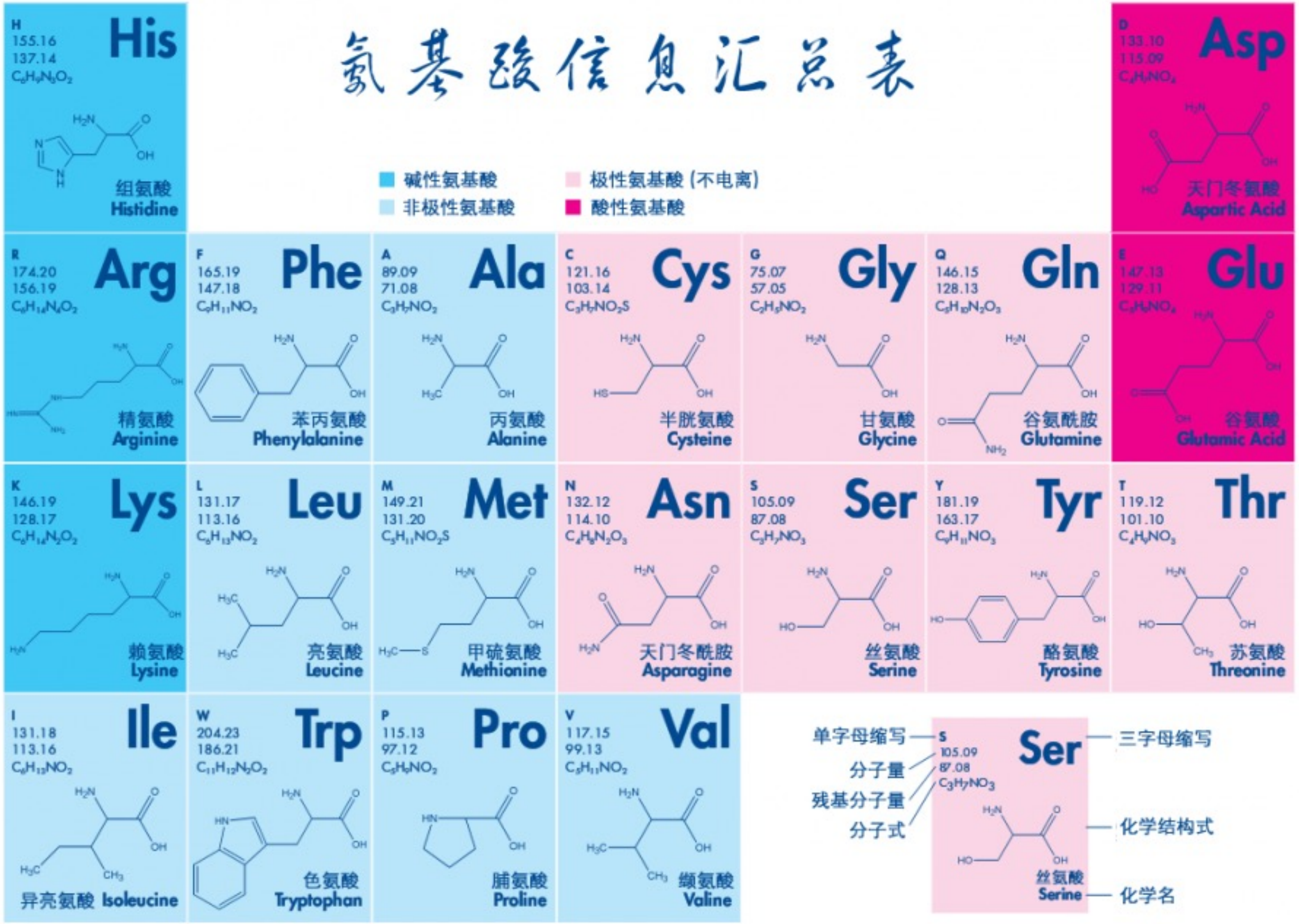
蛋白质的生物功能

蛋白质基础知识

➢ 蛋白质的基本组成单位是氨基酸



氨基酸的化学分子式



氨基酸信息汇总表

■ 碱性氨基酸　■ 非极性氨基酸　■ 极性氨基酸（不电离）　■ 酸性氨基酸

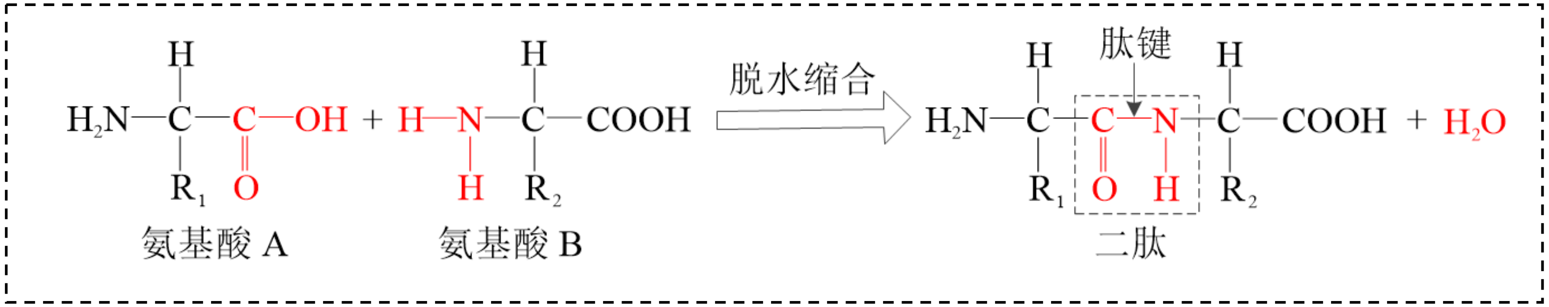
单字母缩写	三字母缩写	分子量	残基分子量	分子式	化学名
H	His	155.16	137.14	$C_6H_9N_3O_2$	组氨酸 Histidine
D	Asp	133.10	115.09	$C_4H_7NO_4$	天门冬氨酸 Aspartic Acid
R	Arg	174.20	156.19	$C_6H_{14}N_4O_2$	精氨酸 Arginine
F	Phe	165.19	147.18	$C_9H_{11}NO_2$	苯丙氨酸 Phenylalanine
A	Ala	89.09	71.08	$C_3H_7NO_2$	丙氨酸 Alanine
C	Cys	121.16	103.14	$C_3H_7NO_2S$	半胱氨酸 Cysteine
G	Gly	75.07	57.05	$C_2H_5NO_2$	甘氨酸 Glycine
Q	Gln	146.15	128.13	$C_5H_{10}N_2O_3$	谷氨酰胺 Glutamine
E	Glu	147.13	129.11	$C_5H_9NO_4$	谷氨酸 Glutamic Acid
K	Lys	146.19	128.17	$C_6H_{14}N_2O_2$	赖氨酸 Lysine
L	Leu	131.17	113.16	$C_6H_{13}NO_2$	亮氨酸 Leucine
M	Met	149.21	131.20	$C_5H_{11}NO_2S$	甲硫氨酸 Methionine
N	Asn	132.12	114.10	$C_4H_8N_2O_3$	天门冬酰胺 Asparagine
S	Ser	105.09	87.08	$C_3H_7NO_3$	丝氨酸 Serine
Y	Tyr	181.19	163.17	$C_9H_{11}NO_3$	酪氨酸 Tyrosine
T	Thr	119.12	101.10	$C_4H_9NO_3$	苏氨酸 Threonine
I	Ile	131.18	113.16	$C_6H_{13}NO_2$	异亮氨酸 Isoleucine
W	Trp	204.23	186.21	$C_{11}H_{12}N_2O_2$	色氨酸 Tryptophan
P	Pro	115.13	97.12	$C_5H_9NO_2$	脯氨酸 Proline
V	Val	117.15	99.13	$C_5H_{11}NO_2$	缬氨酸 Valine

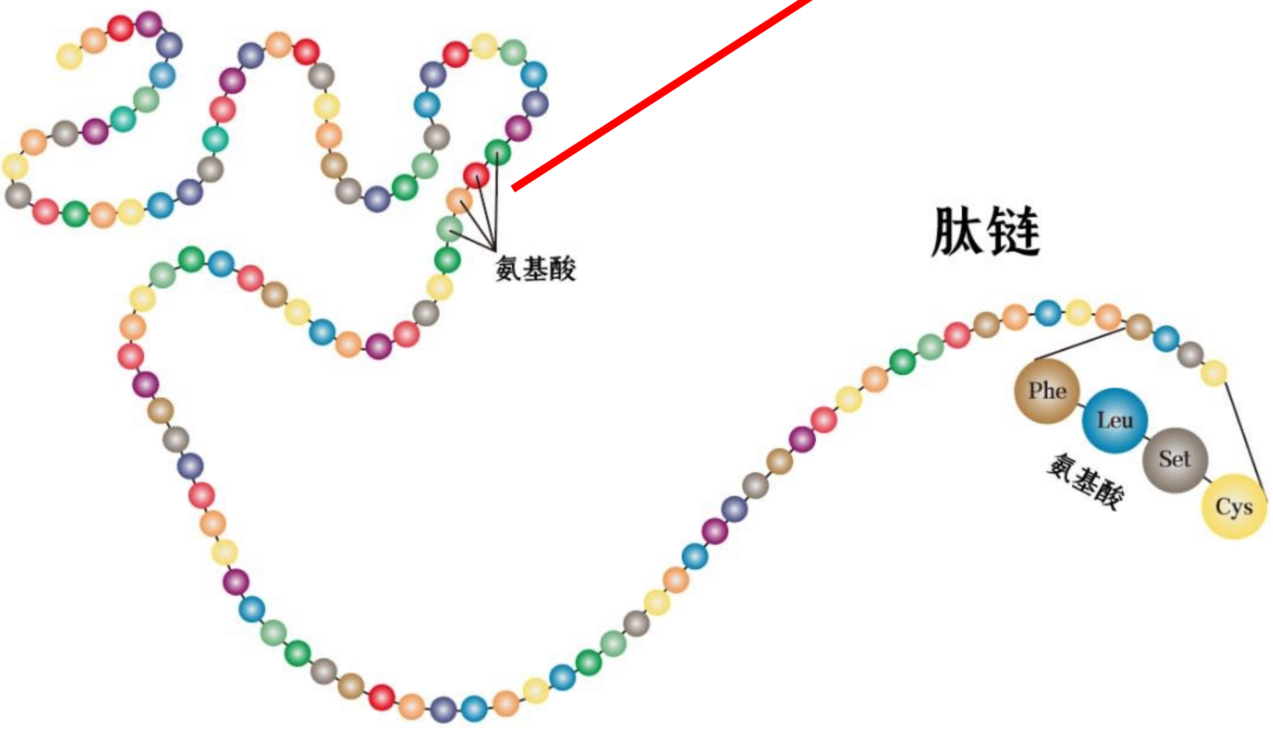
20种常见氨基酸

蛋白质基础知识

- 氨基酸脱水缩合组成肽链

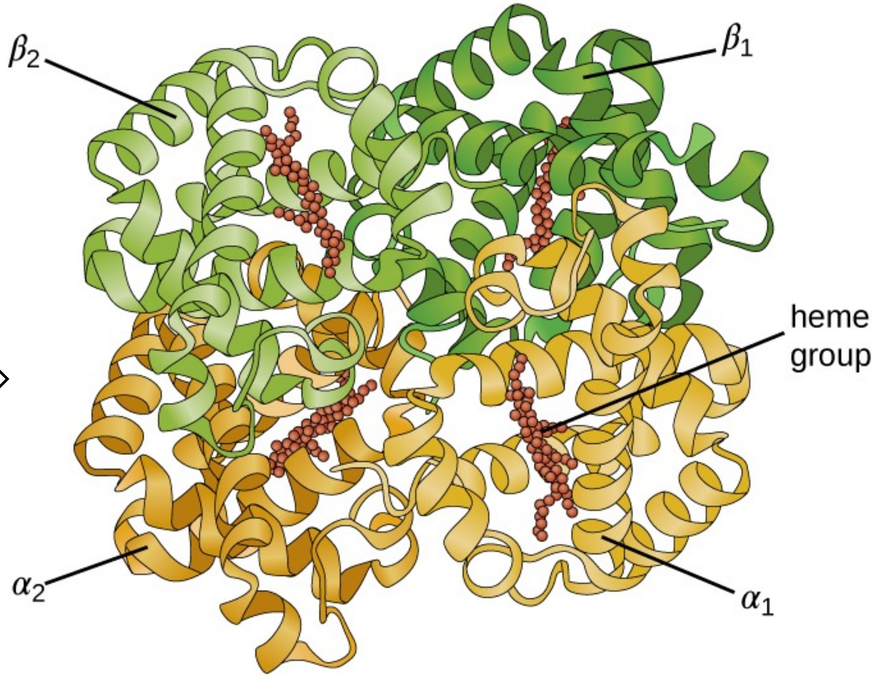
- 肽链相互缠绕组成蛋白质分子





肽链

相互缠绕



蛋白质分子

蛋白质基础知识

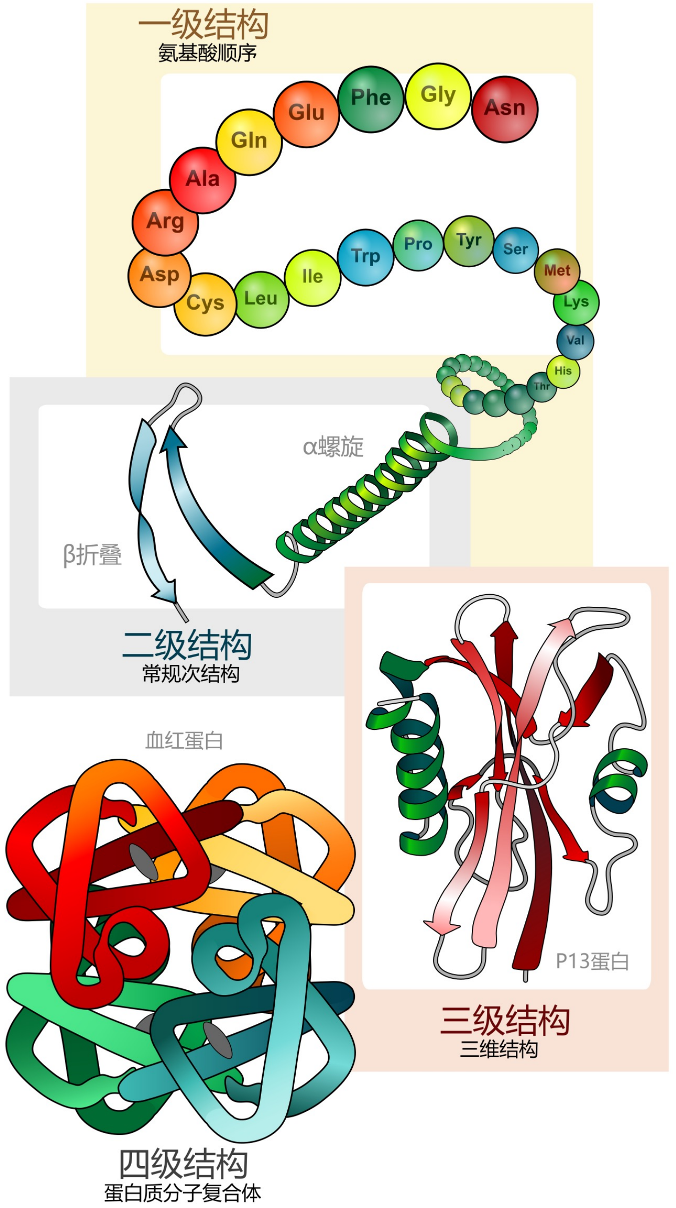
- 氨基酸四级结构

（1）一级结构：氨基酸的线性序列

（2）二级结构：肽链上的局部几何构象

（3）三级结构：肽链上所有原子的空间位置

（4）四级结构：肽链在空间上的相对位置



蛋白质基础知识

- ➢ 蛋白质功能注释的语义词汇标准

  ——基因本体论

  (1) 分子功能

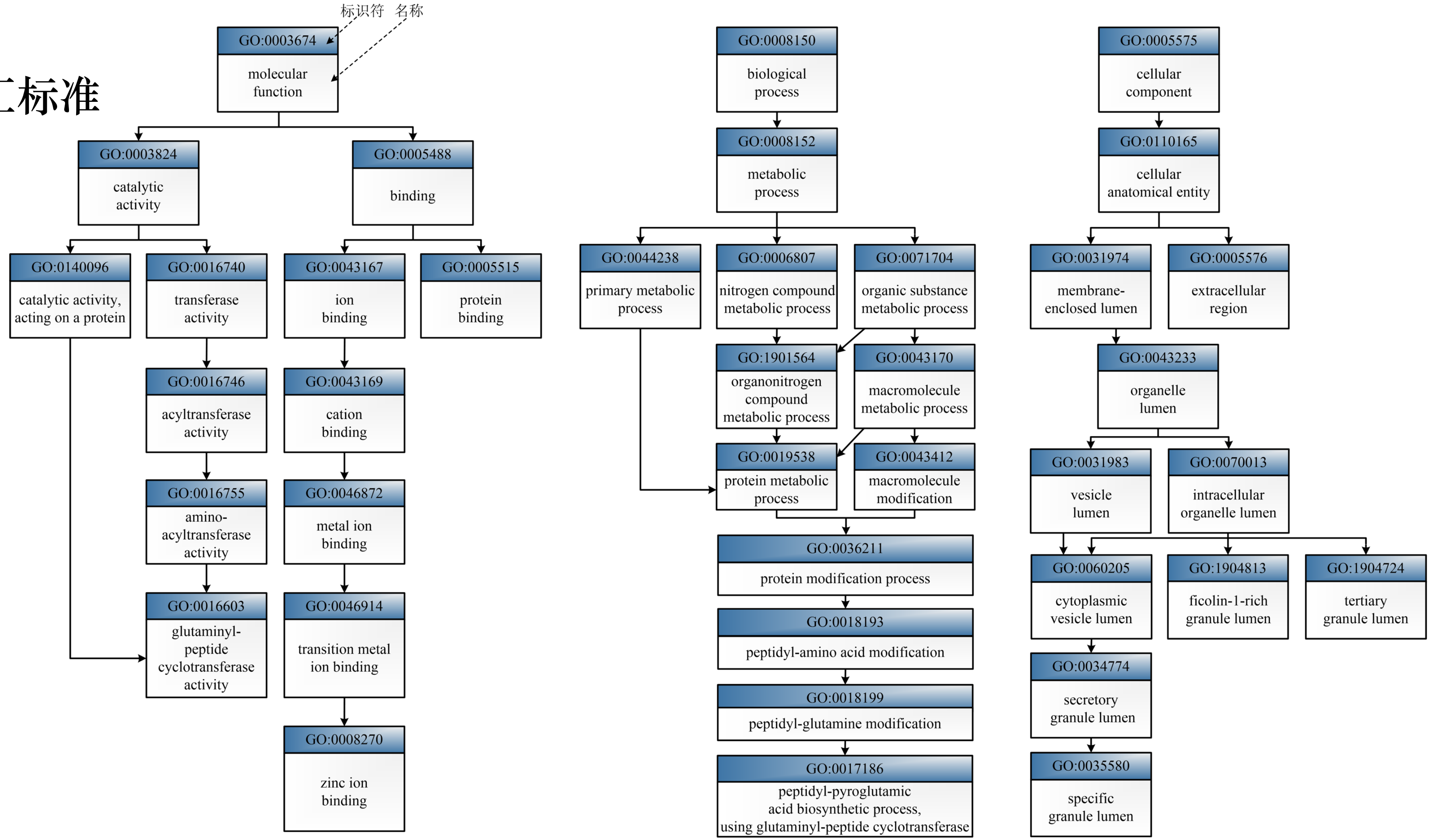
  Molecular Function

  (2) 生物过程

  Biological Process

  (3) 细胞组件

  Cellular Component



标识符 名称

GO:0003674 molecular function

GO:0003824 catalytic activity

GO:0005488 binding

GO:0140096 catalytic activity, acting on a protein

GO:0016740 transferase activity

GO:0043167 ion binding

GO:0005515 protein binding

GO:0016746 acyltransferase activity

GO:0043169 cation binding

GO:0016755 amino-acyltransferase activity

GO:0046872 metal ion binding

GO:0016603 glutaminyl-peptide cyclotransferase activity

GO:0046914 transition metal ion binding

GO:0008270 zinc ion binding

GO:0008150 biological process

GO:0008152 metabolic process

GO:0044238 primary metabolic process

GO:0006807 nitrogen compound metabolic process

GO:0071704 organic substance metabolic process

GO:1901564 organonitrogen compound metabolic process

GO:0043170 macromolecule metabolic process

GO:0019538 protein metabolic process

GO:0043412 macromolecule modification

GO:0036211 protein modification process

GO:0018193 peptidyl-amino acid modification

GO:0018199 peptidyl-glutamine modification

GO:0017186 peptidyl-pyroglutamic acid biosynthetic process, using glutaminyl-peptide cyclotransferase

GO:0005575 cellular component

GO:0110165 cellular anatomical entity

GO:0031974 membrane-enclosed lumen

GO:0005576 extracellular region

GO:0043233 organelle lumen

GO:0031983 vesicle lumen

GO:0070013 intracellular organelle lumen

GO:0060205 cytoplasmic vesicle lumen

GO:1904813 ficolin-1-rich granule lumen

GO:1904724 tertiary granule lumen

GO:0034774 secretory granule lumen

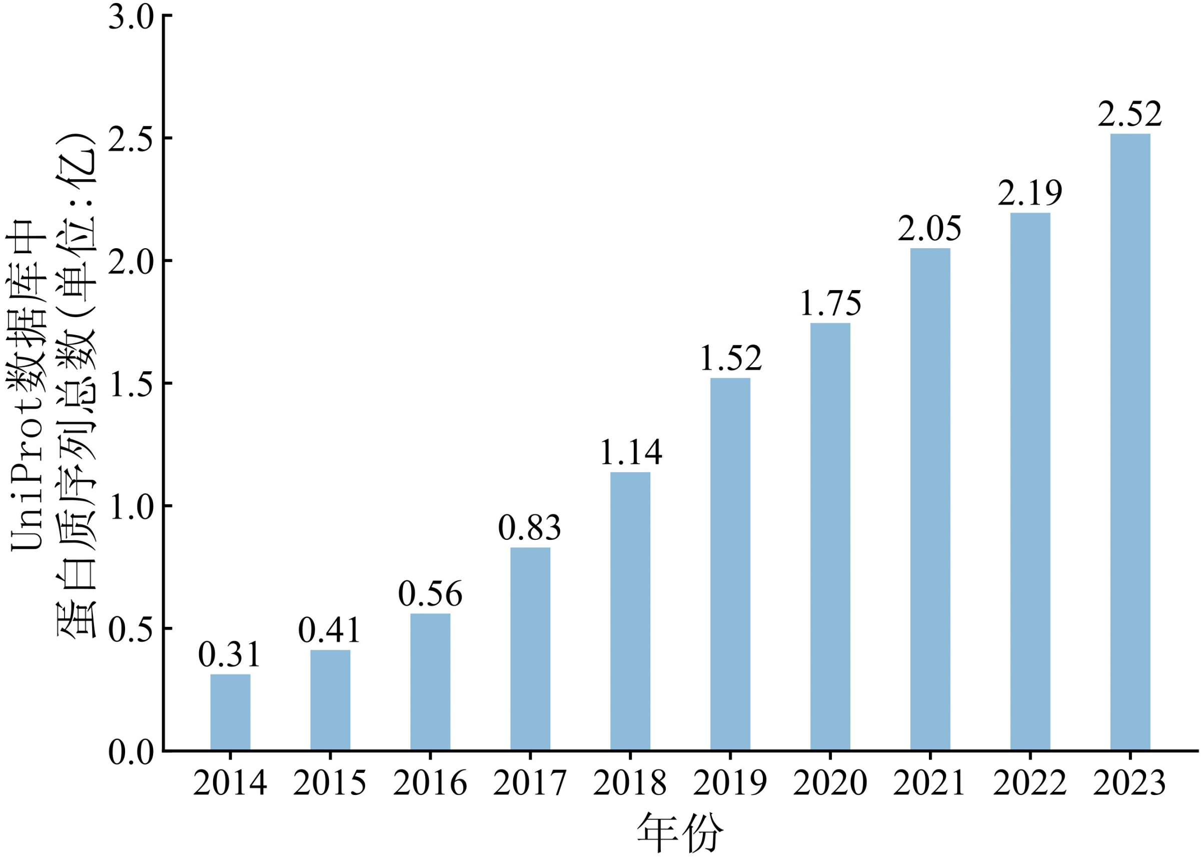
GO:0035580 specific granule lumen

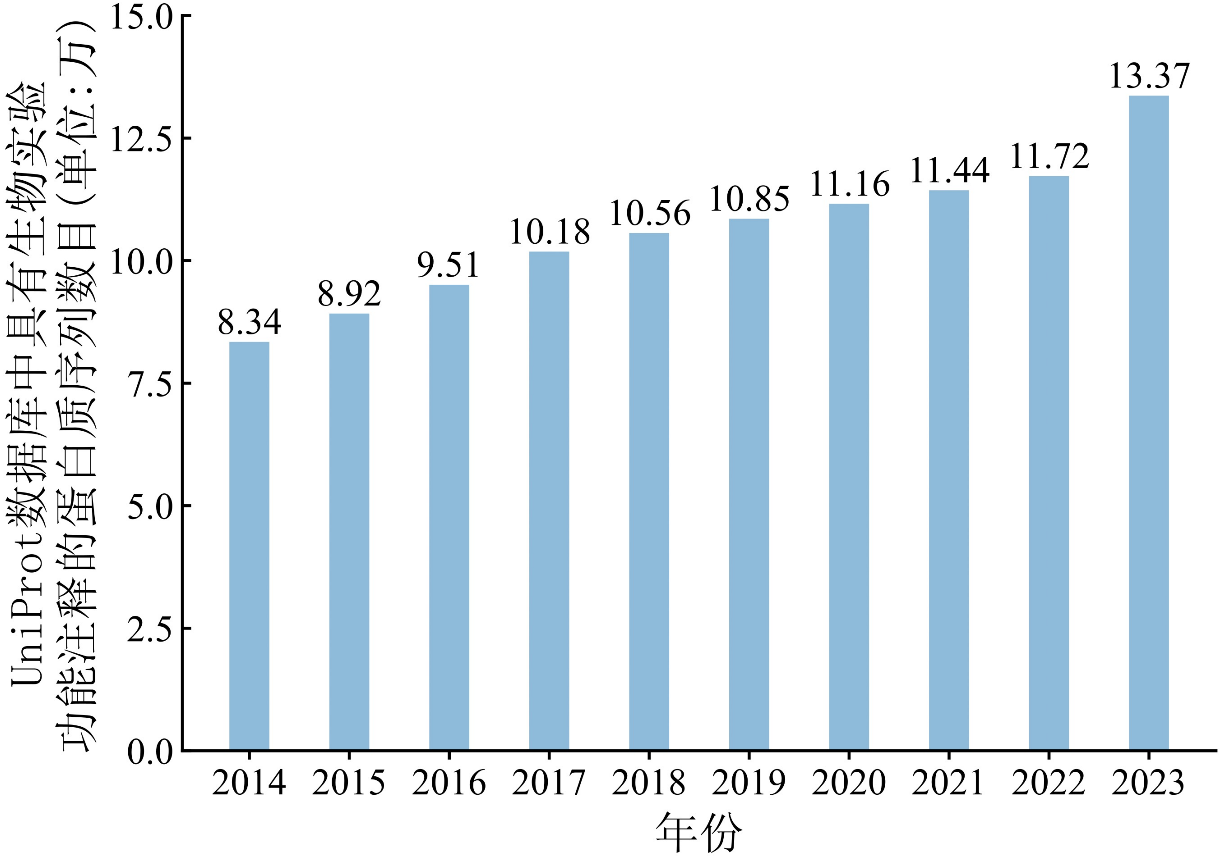
(A) 分子功能 (B) 生物过程 (C) 细胞组件

谷氨酰环化转移酶在分子功能、生物过程和细胞组件分支下的功能注释图

蛋白质功能注释面临的挑战

- 蛋白质功能注释最可靠的途径是生物实验，但它存在周期长、成本高等缺陷。
- 截止2024年1月，UniProt中已累积约2.52亿条序列，但具有生物实验功能注释的序列数目不足序列总数的0.1%。





基于人工智能算法的蛋白质功能预测研究

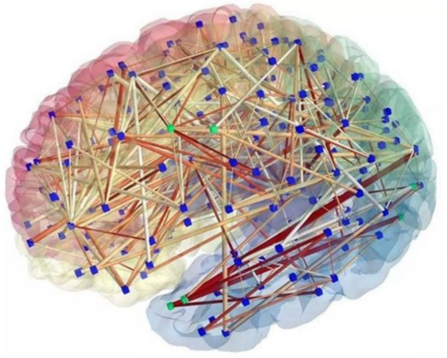
- ➢ 研发高效的生物计算方法来预测蛋白质功能已迫在眉睫。
- ➢ 蛋白质功能预测目标：利用生物计算方法准确地推断出查询蛋白质在基因本体论三个分支下的功能注释。

TACEVAEISYKKFRQLIQVN
P
D
VKESTVQLRRAMQASLRMLI
G
NLAFLDVTGRIAQTLLNLAKQ
P
VIQGIEQRTIKIQMGDPHTMAD
G
CSRETVGRILKMLEDQN

功能未知的蛋白质序列

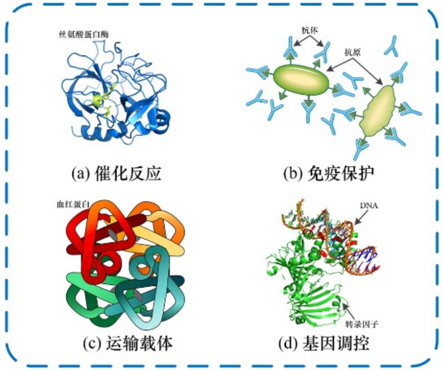
输入

蛋白质功能推断模型

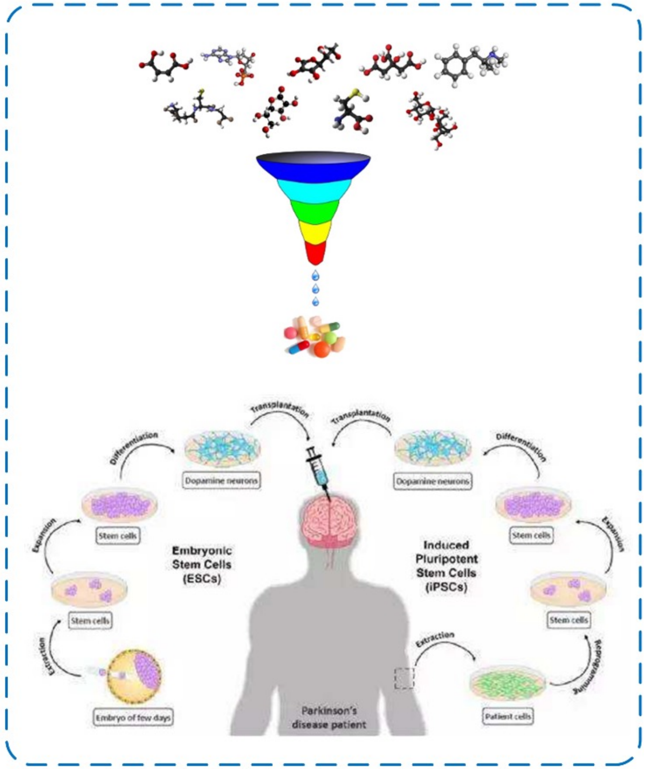


输出

预测的蛋白质功能

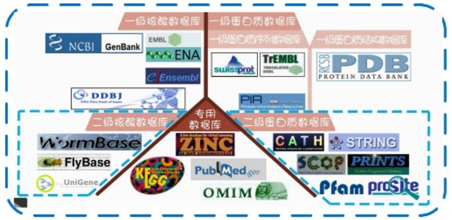


辅助



药物设计与临床治疗

人工智能算法



蛋白质大数据（序列、结构、功能等）

蛋白质功能预测研究所需的知识储备

- 数学

  高等数学、线性代数、概率论等

- 计算机科学

  编程语言(python、C++、R)、数据结构、操作系统等

- 人工智能算法

  机器学习、深度学习等

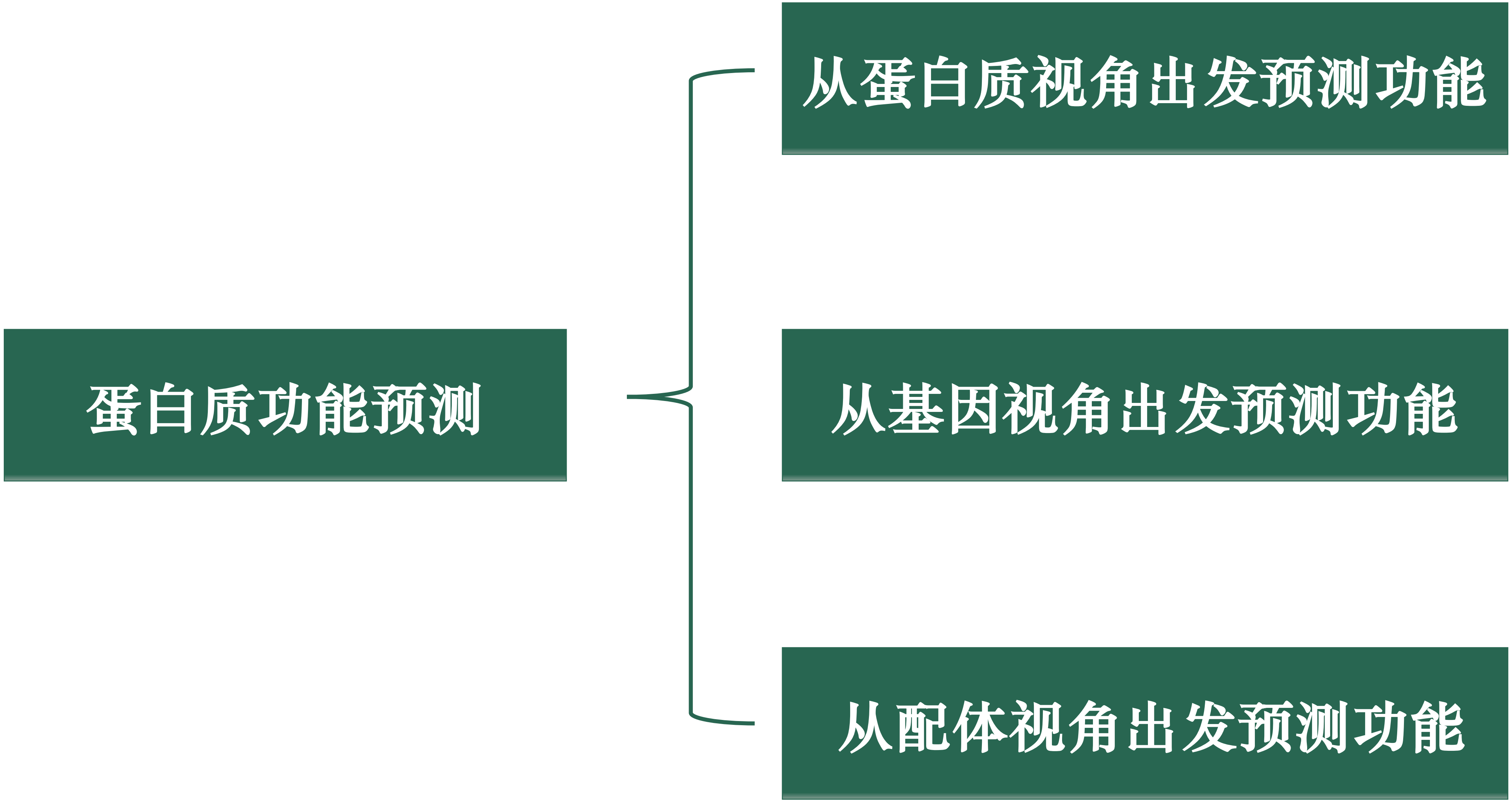
- 生物学

  蛋白质 (序列、结构和功能)、基因等基础知识

02 Part two

研究内容

研究内容



蛋白质功能预测

从蛋白质视角出发预测功能

从基因视角出发预测功能

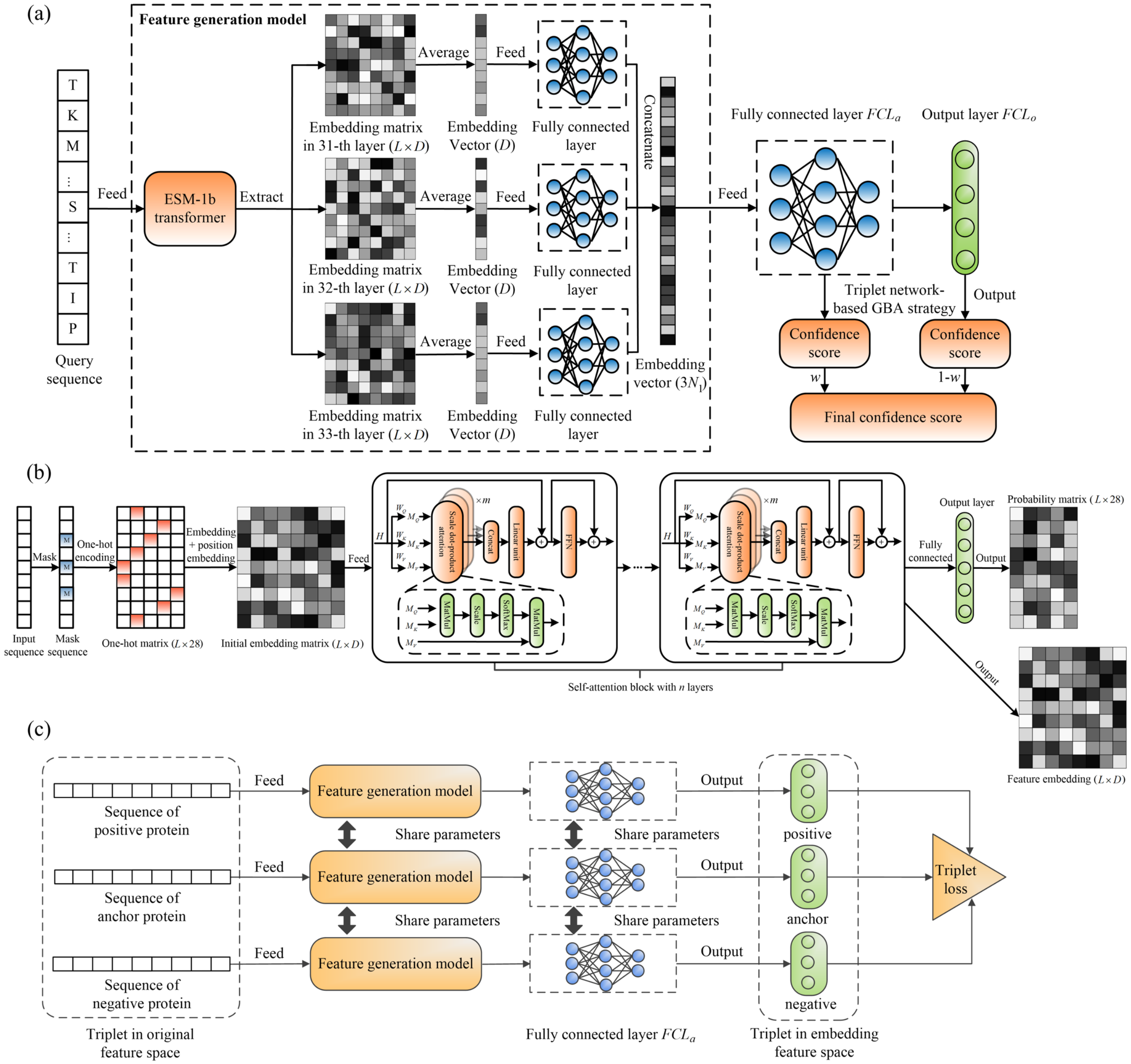
从配体视角出发预测功能

从蛋白质视角出发预测功能

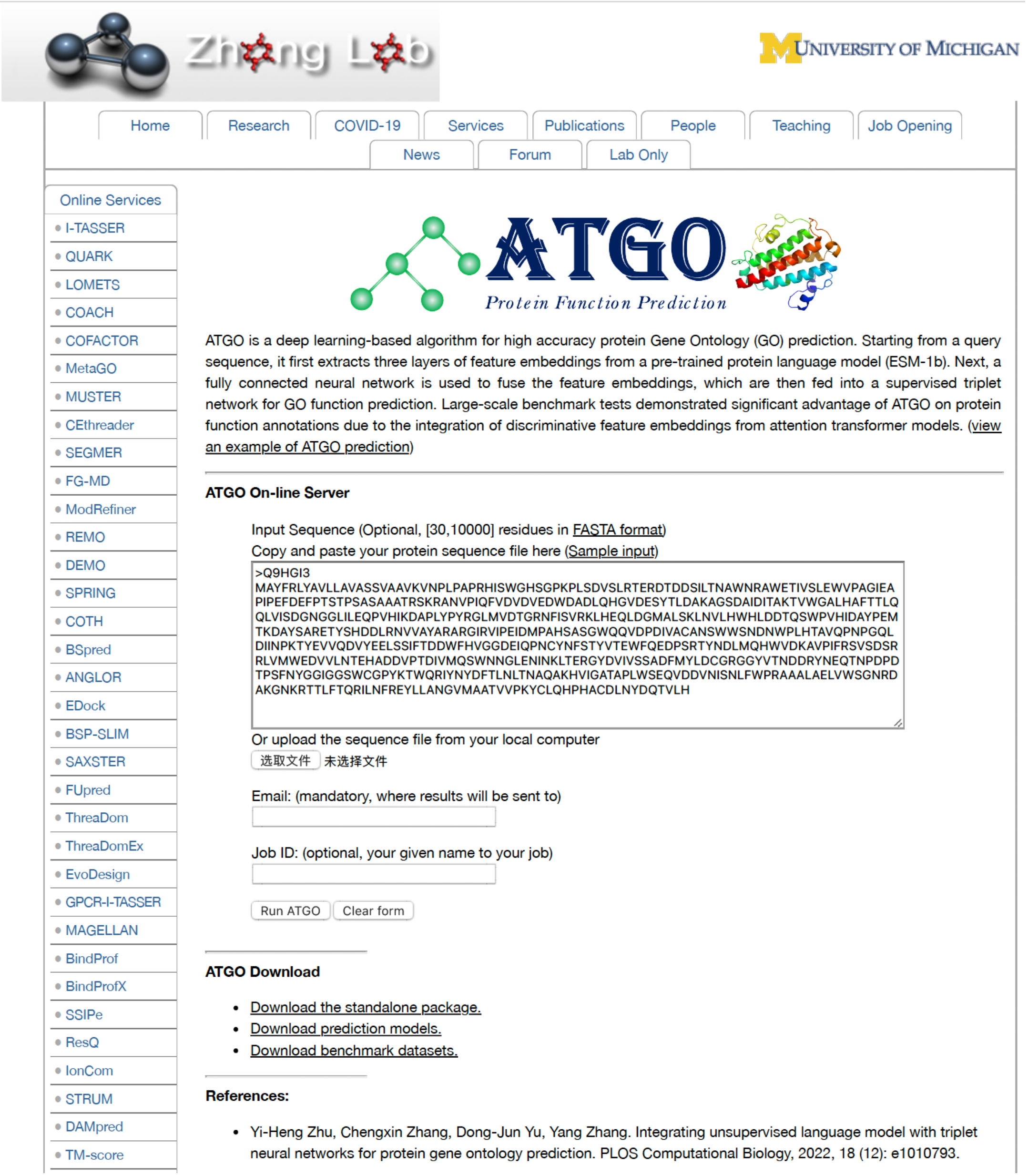
- 基于注意力机制与三元组神经网络的预测方法 ATGO

- 主要贡献：首次将计算机视觉领域的无监督语言模型迁移到蛋白质功能预测领域

Yi-Heng Zhu, Chengxin Zhang, Dong-Jun Yu, Yang Zhang. Integrating Unsupervised Language Model with Triplet Neural Networks for Protein Gene Ontology Prediction. **PLOS Computational Biology**. 2022, 18(12): e1010793.


(a)
Feature generation model
T
K
M
S
T
I
P
Query sequence
Feed
ESM-1b transformer
Extract
Average
Feed
Embedding matrix in 31-th layer ($L \times D$)
Embedding Vector (D)
Fully connected layer
Embedding matrix in 32-th layer ($L \times D$)
Embedding Vector (D)
Fully connected layer
Embedding matrix in 33-th layer ($L \times D$)
Embedding Vector (D)
Fully connected layer
Concatenate
Embedding vector ($3N_1$)
Feed
Fully connected layer FCL_a
Output layer FCL_o
Triplet network-based GBA strategy
Output
Confidence score
Confidence score
w
1-w
Final confidence score
(b)
Input sequence
Mask
Mask sequence
M
One-hot encoding
One-hot matrix ($L \times 28$)
Embedding + position embedding
Initial embedding matrix ($L \times D$)
Feed
H
W_Q M_Q
W_K M_K
W_V M_V
Scale dot-product attention
$\times m$
Concat
Linear unit
FFN
MatMul
Scale
SoftMax
MatMul
Self-attention block with n layers
Output layer
Fully connected
Output
Probability matrix ($L \times 28$)
Output
Feature embedding ($L \times D$)
(c)
Sequence of positive protein
Sequence of anchor protein
Sequence of negative protein
Triplet in original feature space
Feed
Feature generation model
Share parameters
Output
Fully connected layer FCL_a
positive
anchor
negative
Triplet in embedding feature space
Triplet loss

从蛋白质视角出发预测功能



Zhang Lab

UNIVERSITY OF MICHIGAN

Home | Research | COVID-19 | Services | Publications | People | Teaching | Job Opening
News | Forum | Lab Only

Online Services
- I-TASSER
- QUARK
- LOMETS
- COACH
- COFACTOR
- MetaGO
- MUSTER
- CEthreader
- SEGMER
- FG-MD
- ModRefiner
- REMO
- DEMO
- SPRING
- COTH
- BSpred
- ANGLOR
- EDock
- BSP-SLIM
- SAXSTER
- FUpred
- ThreaDom
- ThreaDomEx
- EvoDesign
- GPCR-I-TASSER
- MAGELLAN
- BindProf
- BindProfX
- SSIPe
- ResQ
- IonCom
- STRUM
- DAMpred
- TM-score

ATGO

Protein Function Prediction

ATGO is a deep learning-based algorithm for high accuracy protein Gene Ontology (GO) prediction. Starting from a query sequence, it first extracts three layers of feature embeddings from a pre-trained protein language model (ESM-1b). Next, a fully connected neural network is used to fuse the feature embeddings, which are then fed into a supervised triplet network for GO function prediction. Large-scale benchmark tests demonstrated significant advantage of ATGO on protein function annotations due to the integration of discriminative feature embeddings from attention transformer models. (view an example of ATGO prediction)

ATGO On-line Server

Input Sequence (Optional, [30,10000] residues in FASTA format)
Copy and paste your protein sequence file here (Sample input)

```
>Q9HGI3
MAYFRLYAVLLAVASSVAAVKVNPLPAPRHISWGHSGPKPLSDVSLRTERDTDDSILTNAWNRAWETIVSLEWVPAGIEA
PIPEFDEFPTSTPSASAAATRSKRANVPIQFVDVDVEDWDADLQHGVDESYTLDAKAGSDAIDITAKTVWGALHAFTTLQ
QLVISDGNGGLILEQPVHIKDAPLYPYRGLMVDTGRNFISVRKLHEQLDGMALSKLNVLHWHLDDTQSWPVHIDAYPEM
TKDAYSARETYSHDDLRNVVAYARARGIRVIPEIDMPAHSASGWQQVDPDIVACANSWWSNDNWPLHTAVQPNPGQL
DIINPKTYEVVQDVYEELSSIFTDDWFHVGGDEIQPNCYNFSTYVTEWFQEDPSRTYNDLMQHWVDKAVPIFRSVSDSR
RLVMWEDVVLNTEHADDVPTDIVMQSWNNGLENINKLTERGYDVIVSSADFMYLDCGRGGYVTNDDRYNEQTNPDPD
TPSFNYGGIGGSWCGPYKTWQRIYNYDFTLNLTNAQAKHVIGATAPLWSEQVDDVNISNLFWPRAAALAELVWSGNRD
AKGNKRTTLFTQRILNFREYLLANGVMAATVVPKYCLQHPHACDLNYDQTVLH
```

Or upload the sequence file from your local computer

Email: (mandatory, where results will be sent to)

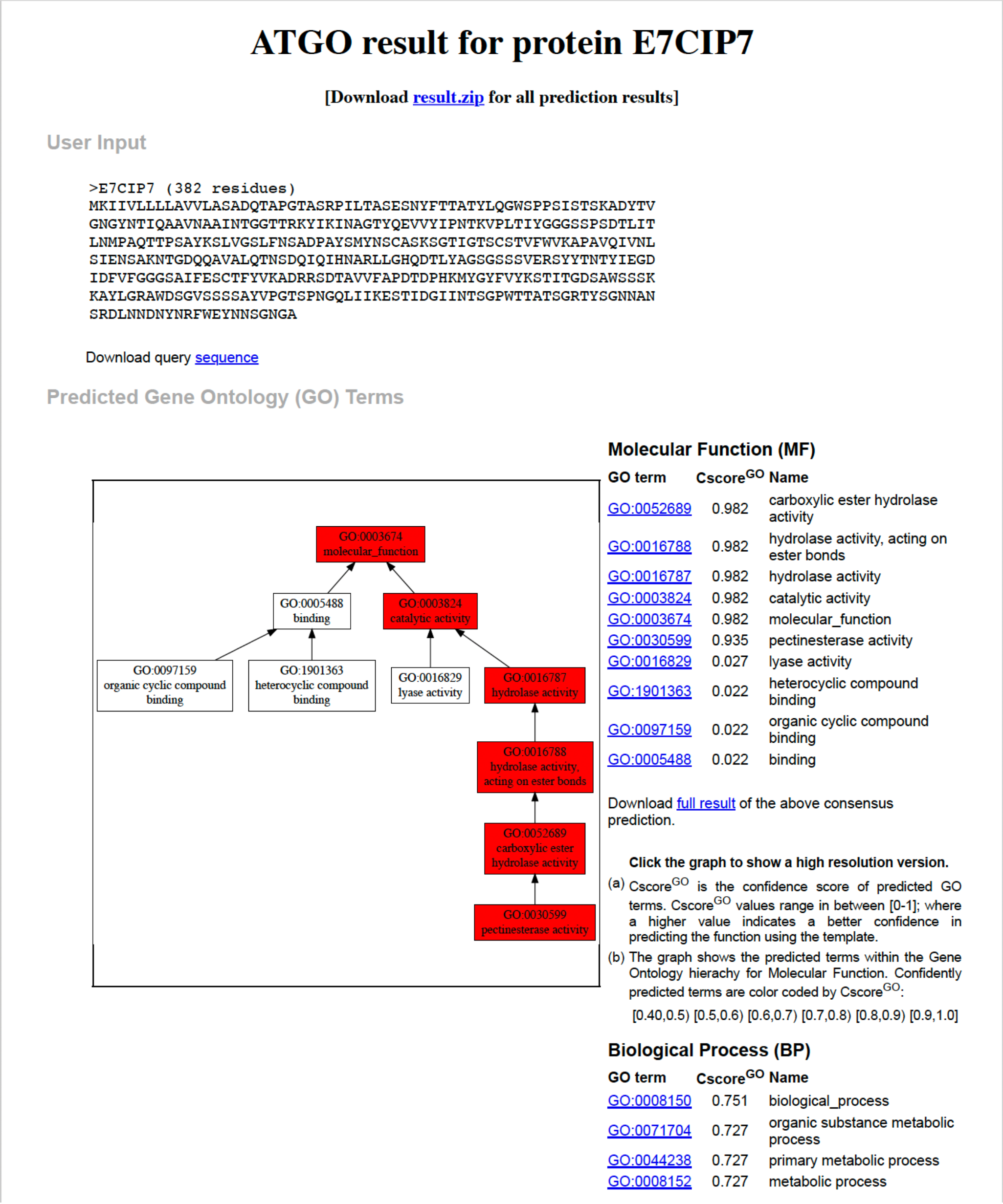
Job ID: (optional, your given name to your job)

ATGO Download

- Download the standalone package.
- Download prediction models.
- Download benchmark datasets.

References:

- Yi-Heng Zhu, Chengxin Zhang, Dong-Jun Yu, Yang Zhang. Integrating unsupervised language model with triplet neural networks for protein gene ontology prediction. PLOS Computational Biology, 2022, 18 (12): e1010793.



ATGO result for protein E7CIP7

[Download result.zip for all prediction results]

User Input

```
>E7CIP7 (382 residues)
MKIIVLLLLAVVLASADQTAPGTASRPILTASESNYFTTATYLQGWSPPSISTSKADYTV
GNGYNTIQAAVNAAINTGGTTRKYIKINAGTYQEVVYIPNTKVPLTIYGGGSSPSDTLIT
LNMPAQTTPSAYKSLVGSLFNSADPAYSMYNSCASKSGTIGTSCSTVFWVKAPAVQIVNL
SIENSAKNTGDQQAVALQTNSDQIQIHNARLLGHQDTLYAGSGSSSVERSYYTNTYIEGD
IDFVFGGGSAIFESCTFYVKADRRSDTAVVFAPDTDPHKMYGYFVYKSTITGDSAWSSSK
KAYLGRAWDSGVSSSSAYVPGTSPNGQLIIKESTIDGIINTSGPWTTATSGRTYSGNNAN
SRDLNNDNYNRFWEYNNSGNGA
```

Download query sequence

Predicted Gene Ontology (GO) Terms

Molecular Function (MF)

GO term	Cscore^{GO}	Name
GO:0052689	0.982	carboxylic ester hydrolase activity
GO:0016788	0.982	hydrolase activity, acting on ester bonds
GO:0016787	0.982	hydrolase activity
GO:0003824	0.982	catalytic activity
GO:0003674	0.982	molecular_function
GO:0030599	0.935	pectinesterase activity
GO:0016829	0.027	lyase activity
GO:1901363	0.022	heterocyclic compound binding
GO:0097159	0.022	organic cyclic compound binding
GO:0005488	0.022	binding

Download full result of the above consensus prediction.

Click the graph to show a high resolution version.

(a) Cscore^{GO} is the confidence score of predicted GO terms. Cscore^{GO} values range in between [0-1]; where a higher value indicates a better confidence in predicting the function using the template.

(b) The graph shows the predicted terms within the Gene Ontology hierachy for Molecular Function. Confidently predicted terms are color coded by Cscore^{GO}:

[0.40,0.5) [0.5,0.6) [0.6,0.7) [0.7,0.8) [0.8,0.9) [0.9,1.0]

Biological Process (BP)

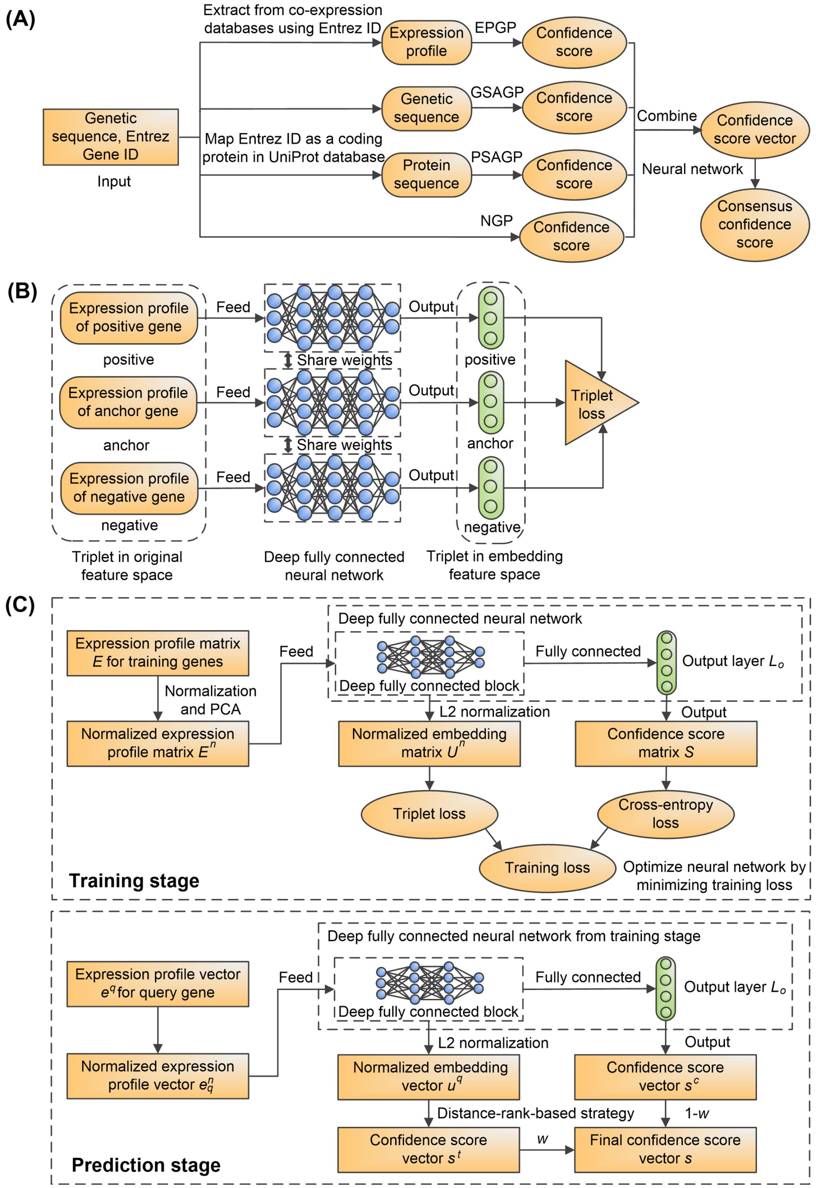
GO term	Cscore^{GO}	Name
GO:0008150	0.751	biological_process
GO:0071704	0.727	organic substance metabolic process
GO:0044238	0.727	primary metabolic process
GO:0008152	0.727	metabolic process

从基因视角出发预测功能

- 基于度量学习与多源信息融合的功能预测方法 TripletGO

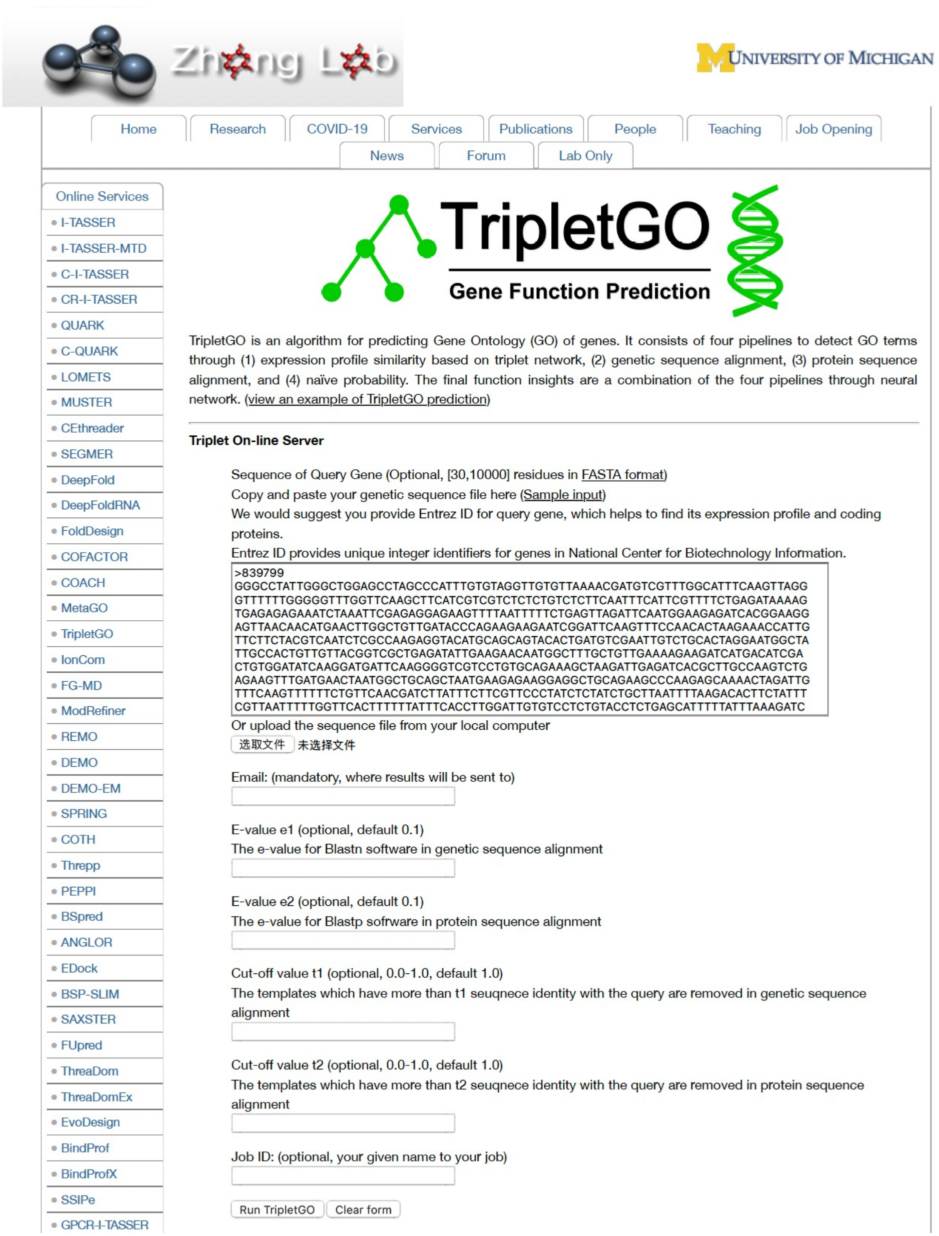
- 主要贡献：首次将基于基因视角的方法和基于蛋白质视角的相结合，为后续的研究开辟了新的思路

Yi-Heng Zhu, Chengxin Zhang, Yan Liu, Gilbert Omenn, Peter Freddolino, Dong-Jun Yu, Yang Zhang. TripletGO: Integrating Transcript Expression Profiles with Protein Homology Inferences for Gene Function Prediction. Genomics, **Proteomics & Bioinformatics**. 2022, 20(5): 1013-1027.



(A)
Genetic sequence, Entrez Gene ID
Input
Extract from co-expression databases using Entrez ID
Expression profile
EPGP
Confidence score
Genetic sequence
GSAGP
Confidence score
Map Entrez ID as a coding protein in UniProt database
Protein sequence
PSAGP
Confidence score
NGP
Confidence score
Combine
Confidence score vector
Neural network
Consensus confidence score
(B)
Expression profile of positive gene
positive
Expression profile of anchor gene
anchor
Expression profile of negative gene
negative
Triplet in original feature space
Feed
Feed
Feed
Share weights
Share weights
Deep fully connected neural network
Output
Output
Output
positive
anchor
negative
Triplet in embedding feature space
Triplet loss
(C)
Expression profile matrix E for training genes
Normalization and PCA
Normalized expression profile matrix E^n
Feed
Deep fully connected neural network
Deep fully connected block
Fully connected
Output layer L_o
L2 normalization
Output
Normalized embedding matrix U^n
Confidence score matrix S
Triplet loss
Cross-entropy loss
Training loss
Optimize neural network by minimizing training loss
Training stage
Expression profile vector e^q for query gene
Normalized expression profile vector e_q^n
Feed
Deep fully connected neural network from training stage
Deep fully connected block
Fully connected
Output layer L_o
L2 normalization
Output
Normalized embedding vector u^q
Confidence score vector s^c
Distance-rank-based strategy
1-w
Confidence score vector s^t
w
Final confidence score vector s
Prediction stage

从基因视角出发预测功能



Online Services

- I-TASSER
- I-TASSER-MTD
- C-I-TASSER
- CR-I-TASSER
- QUARK
- C-QUARK
- LOMETS
- MUSTER
- CEthreader
- SEGMER
- DeepFold
- DeepFoldRNA
- FoldDesign
- COFACTOR
- COACH
- MetaGO
- TripletGO
- IonCom
- FG-MD
- ModRefiner
- REMO
- DEMO
- DEMO-EM
- SPRING
- COTH
- Threpp
- PEPPI
- BSpred
- ANGLOR
- EDock
- BSP-SLIM
- SAXSTER
- FUpred
- ThreaDom
- ThreaDomEx
- EvoDesign
- BindProf
- BindProfX
- SSIPe
- GPCR-I-TASSER

TripletGO
Gene Function Prediction

TripletGO is an algorithm for predicting Gene Ontology (GO) of genes. It consists of four pipelines to detect GO terms through (1) expression profile similarity based on triplet network, (2) genetic sequence alignment, (3) protein sequence alignment, and (4) naïve probability. The final function insights are a combination of the four pipelines through neural network. (view an example of TripletGO prediction)

Triplet On-line Server

Sequence of Query Gene (Optional, [30,10000] residues in FASTA format)
Copy and paste your genetic sequence file here (Sample input)
We would suggest you provide Entrez ID for query gene, which helps to find its expression profile and coding proteins.
Entrez ID provides unique integer identifiers for genes in National Center for Biotechnology Information.

```
>839799
GGGCCTATTGGGCTGGAGCCTAGCCCATTTGTGTAGGTTGTGTTAAAACGATGTCGTTTGGCATTTCAAGTTAGG
GTTTTTTGGGGGTTTGGTTCAAGCTTCATCGTCGTCTCTCTGTCTCTTCAATTTCATTCGTTTTCTGAGATAAAAG
TGAGAGAGAAATCTAAATTCGAGAGGAGAAGTTTTAATTTTTCTGAGTTAGATTCAATGGAAGAGATCACGGAAGG
AGTTAACAACATGAACTTGGCTGTTGATACCCAGAAGAAGAATCGGATTCAAGTTTCCAACACTAAGAAACCATTG
TTCTTCTACGTCAATCTCGCCAAGAGGTACATGCAGCAGTACACTGATGTCGAATTGTCTGCACTAGGAATGGCTA
TTGCCACTGTTGTTACGGTCGCTGAGATATTGAAGAACAATGGCTTTGCTGTTGAAAAGAAGATCATGACATCGA
CTGTGGATATCAAGGATGATTCAAGGGGTCGTCCTGTGCAGAAAGCTAAGATTGAGATCACGCTTGCCAAGTCTG
AGAAGTTTGATGAACTAATGGCTGCAGCTAATGAAGAGAAGGAGGCTGCAGAAGCCCAAGAGCAAAACTAGATTG
TTTCAAGTTTTTTCTGTTCAACGATCTTATTTCTTCGTTCCCTATCTCTATCTGCTTAATTTTAAGACACTTCTATTT
CGTTAATTTTTGGTTCACTTTTTTATTTCACCTTGGATTGTGTCCTCTGTACCTCTGAGCATTTTTATTTAAAGATC
```

Or upload the sequence file from your local computer
选取文件 未选择文件

Email: (mandatory, where results will be sent to)

E-value e1 (optional, default 0.1)
The e-value for Blastn software in genetic sequence alignment

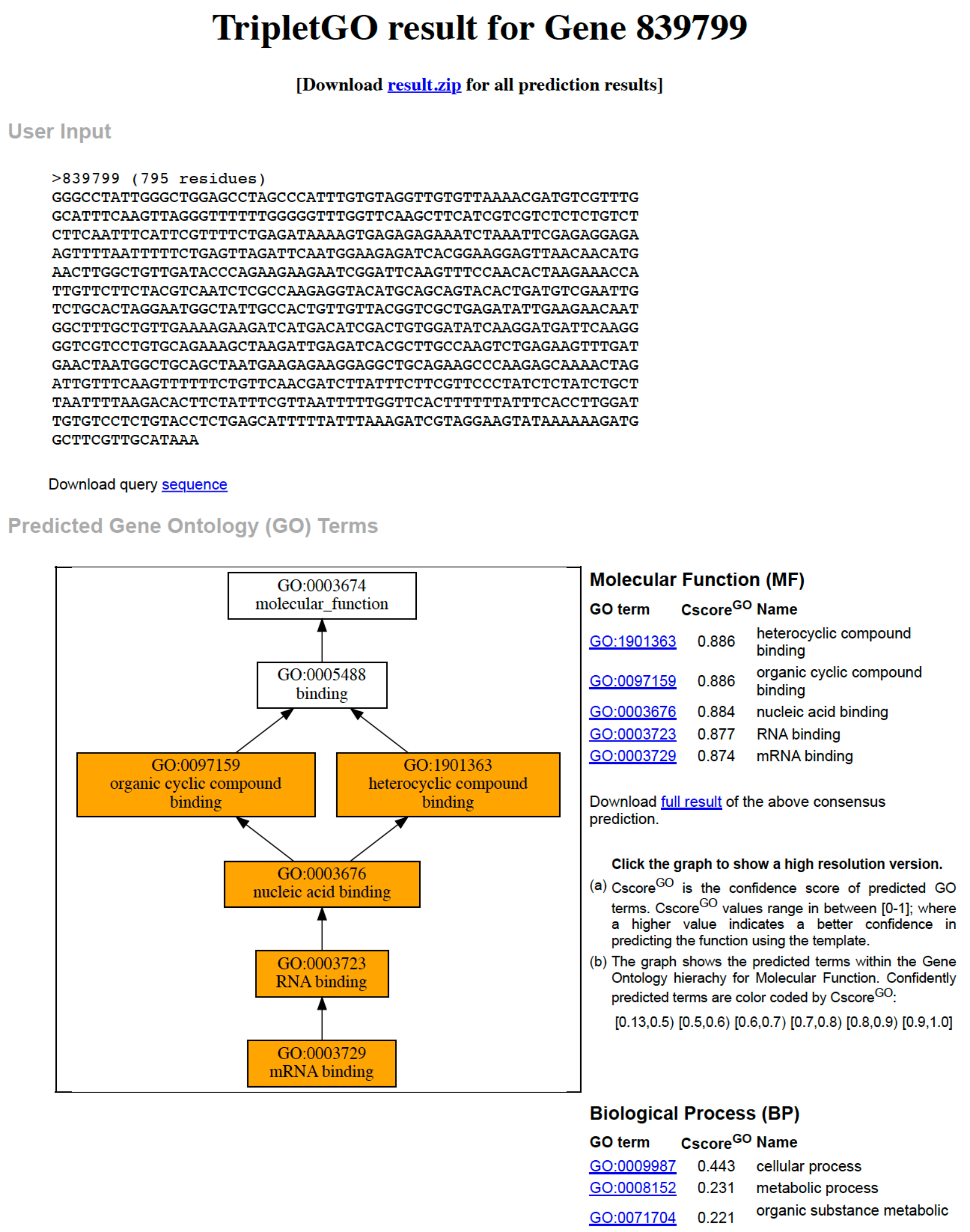
E-value e2 (optional, default 0.1)
The e-value for Blastp sofrware in protein sequence alignment

Cut-off value t1 (optional, 0.0-1.0, default 1.0)
The templates which have more than t1 seuqnece identity with the query are removed in genetic sequence alignment

Cut-off value t2 (optional, 0.0-1.0, default 1.0)
The templates which have more than t2 seuqnece identity with the query are removed in protein sequence alignment

Job ID: (optional, your given name to your job)

Run TripletGO Clear form



TripletGO result for Gene 839799

[Download result.zip for all prediction results]

User Input

```
>839799 (795 residues)
GGGCCTATTGGGCTGGAGCCTAGCCCATTTGTGTAGGTTGTGTTAAAACGATGTCGTTTG
GCATTTCAAGTTAGGGTTTTTTGGGGGTTTGGTTCAAGCTTCATCGTCGTCTCTCTGTCT
CTTCAATTTCATTCGTTTTCTGAGATAAAAGTGAGAGAGAAATCTAAATTCGAGAGGAGA
AGTTTTAATTTTTCTGAGTTAGATTCAATGGAAGAGATCACGGAAGGAGTTAACAACATG
AACTTGGCTGTTGATACCCAGAAGAAGAATCGGATTCAAGTTTCCAACACTAAGAAACCA
TTGTTCTTCTACGTCAATCTCGCCAAGAGGTACATGCAGCAGTACACTGATGTCGAATTG
TCTGCACTAGGAATGGCTATTGCCACTGTTGTTACGGTCGCTGAGATATTGAAGAACAAT
GGCTTTGCTGTTGAAAAGAAGATCATGACATCGACTGTGGATATCAAGGATGATTCAAGG
GGTCGTCCTGTGCAGAAAGCTAAGATTGAGATCACGCTTGCCAAGTCTGAGAAGTTTGAT
GAACTAATGGCTGCAGCTAATGAAGAGAAGGAGGCTGCAGAAGCCCAAGAGCAAAACTAG
ATTGTTTCAAGTTTTTTCTGTTCAACGATCTTATTTCTTCGTTCCCTATCTCTATCTGCT
TAATTTTAAGACACTTCTATTTCGTTAATTTTTGGTTCACTTTTTTATTTCACCTTGGAT
TGTGTCCTCTGTACCTCTGAGCATTTTTATTTAAAGATCGTAGGAAGTATAAAAAAGATG
GCTTCGTTGCATAAA
```

Download query sequence

Predicted Gene Ontology (GO) Terms

GO:0003674 molecular_function
GO:0005488 binding
GO:0097159 organic cyclic compound binding
GO:1901363 heterocyclic compound binding
GO:0003676 nucleic acid binding
GO:0003723 RNA binding
GO:0003729 mRNA binding

Molecular Function (MF)

GO term	Cscore^{GO}	Name
GO:1901363	0.886	heterocyclic compound binding
GO:0097159	0.886	organic cyclic compound binding
GO:0003676	0.884	nucleic acid binding
GO:0003723	0.877	RNA binding
GO:0003729	0.874	mRNA binding

Download full result of the above consensus prediction.

Click the graph to show a high resolution version.

(a) Cscore^{GO} is the confidence score of predicted GO terms. Cscore^{GO} values range in between [0-1]; where a higher value indicates a better confidence in predicting the function using the template.

(b) The graph shows the predicted terms within the Gene Ontology hierachy for Molecular Function. Confidently predicted terms are color coded by Cscore^{GO}:

[0.13,0.5) [0.5,0.6) [0.6,0.7) [0.7,0.8) [0.8,0.9) [0.9,1.0]

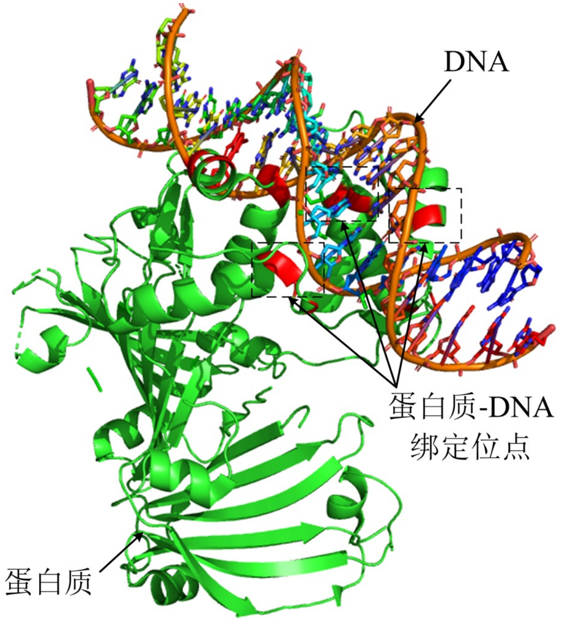
Biological Process (BP)

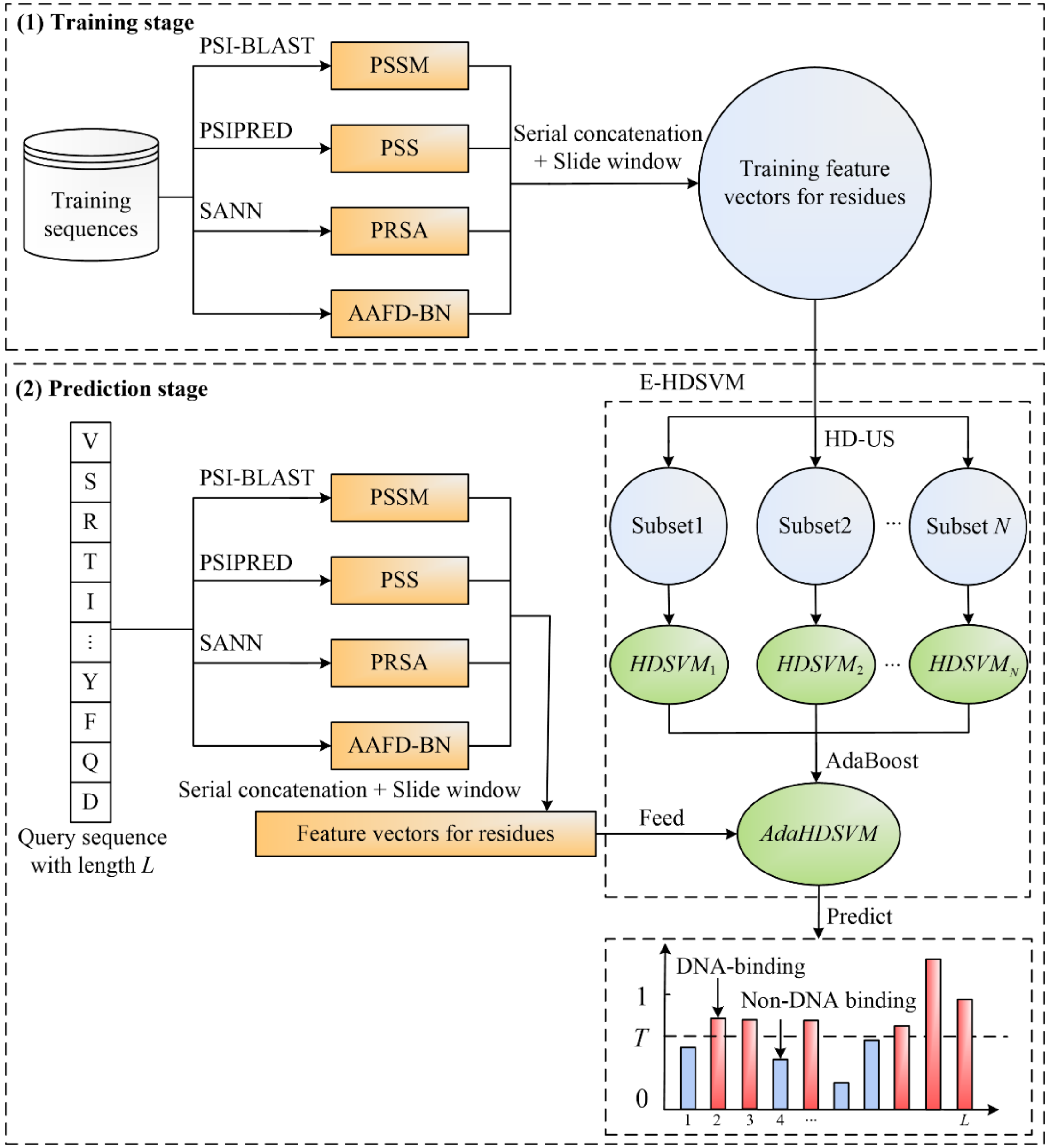
GO term	Cscore^{GO}	Name
GO:0009987	0.443	cellular process
GO:0008152	0.231	metabolic process
GO:0071704	0.221	organic substance metabolic

从配体视角出发预测功能

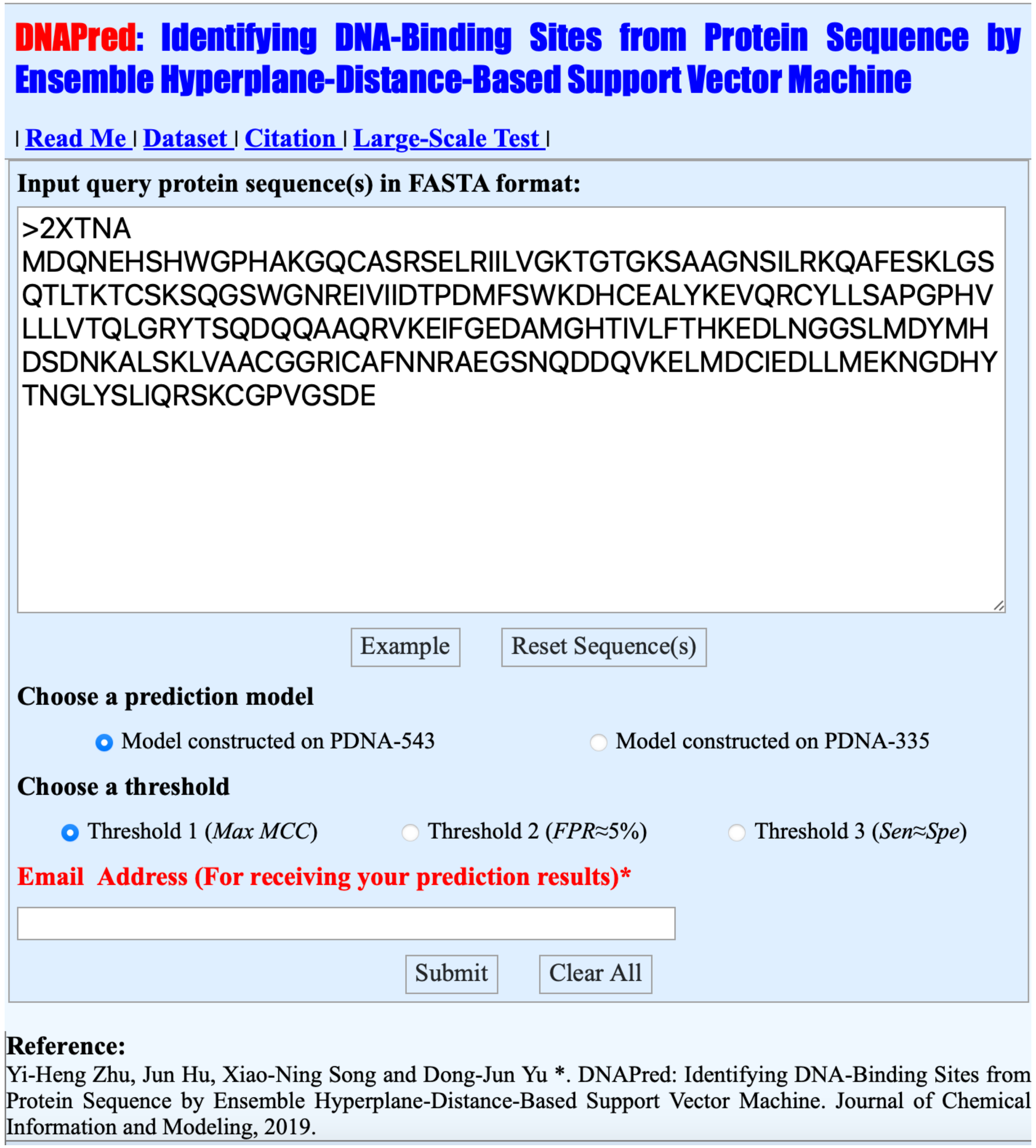
- 基于多粒度支持向量机集成与序列特征的蛋白质-DNA绑定位点预测方法 DNAPred
- 主要贡献：提出了新的类不平衡学习算法E-HDSVM，显著地提升了蛋白质-DNA绑定位点预测精度。

Yi-Heng Zhu, Jun Hu, Xiao-Ning Song, Dong-Jun Yu. DNAPred: Accurate Identification of DNA-binding Sites from Protein Sequence by Ensembled Hyperplane-Distance-Based Support Vector Machines. **Journal of Chemical Information and Modeling**. 2019, 59:3057-3071.





从配体视角出发预测功能



DNAPred: Identifying DNA-Binding Sites from Protein Sequence by Ensemble Hyperplane-Distance-Based Support Vector Machine

| Read Me | Dataset | Citation | Large-Scale Test |

Input query protein sequence(s) in FASTA format:

```
>2XTNA
MDQNEHSHWGPHAKGQCASRSELRIILVGKTGTGKSAAGNSILRKQAFESKLGS
QTLTKTCSKSQGSWGNREIVIIDTPDMFSWKDHCEALYKEVQRCYLLSAPGPHV
LLLVTQLGRYTSQDQQAAQRVKEIFGEDAMGHTIVLFTHKEDLNGGSLMDYMH
DSDNKALSKLVAACGGRICAFNNRAEGSNQDDQVKELMDCIEDLLMEKNGDHY
TNGLYSLIQRSKCGPVGSDE
```

Example | Reset Sequence(s)

Choose a prediction model

● Model constructed on PDNA-543 ○ Model constructed on PDNA-335

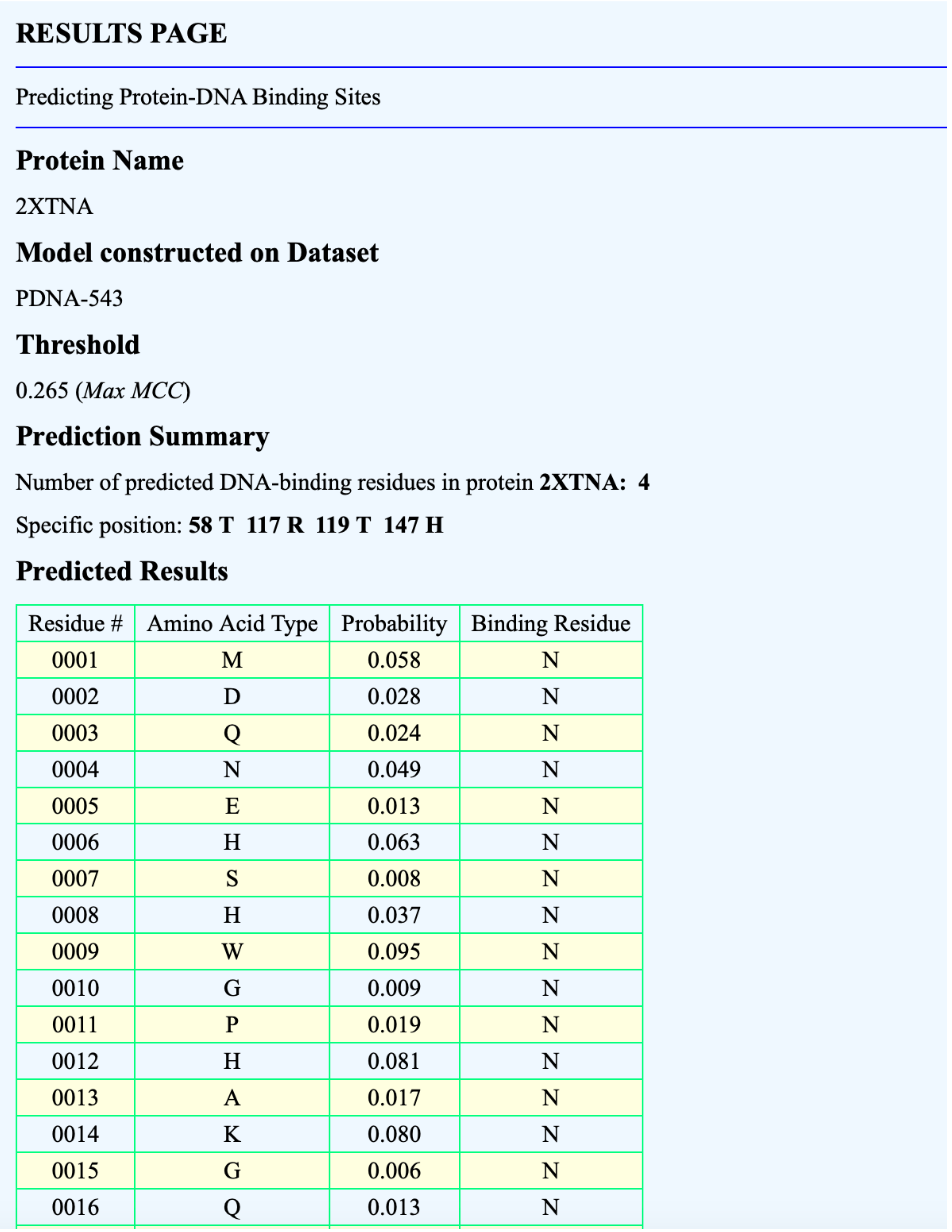
Choose a threshold

● Threshold 1 (*Max MCC*) ○ Threshold 2 (*FPR*≈5%) ○ Threshold 3 (*Sen*≈*Spe*)

Email Address (For receiving your prediction results)*

Submit | Clear All

Reference:
Yi-Heng Zhu, Jun Hu, Xiao-Ning Song and Dong-Jun Yu *. DNAPred: Identifying DNA-Binding Sites from Protein Sequence by Ensemble Hyperplane-Distance-Based Support Vector Machine. Journal of Chemical Information and Modeling, 2019.



RESULTS PAGE

Predicting Protein-DNA Binding Sites

Protein Name

2XTNA

Model constructed on Dataset

PDNA-543

Threshold

0.265 (*Max MCC*)

Prediction Summary

Number of predicted DNA-binding residues in protein **2XTNA: 4**

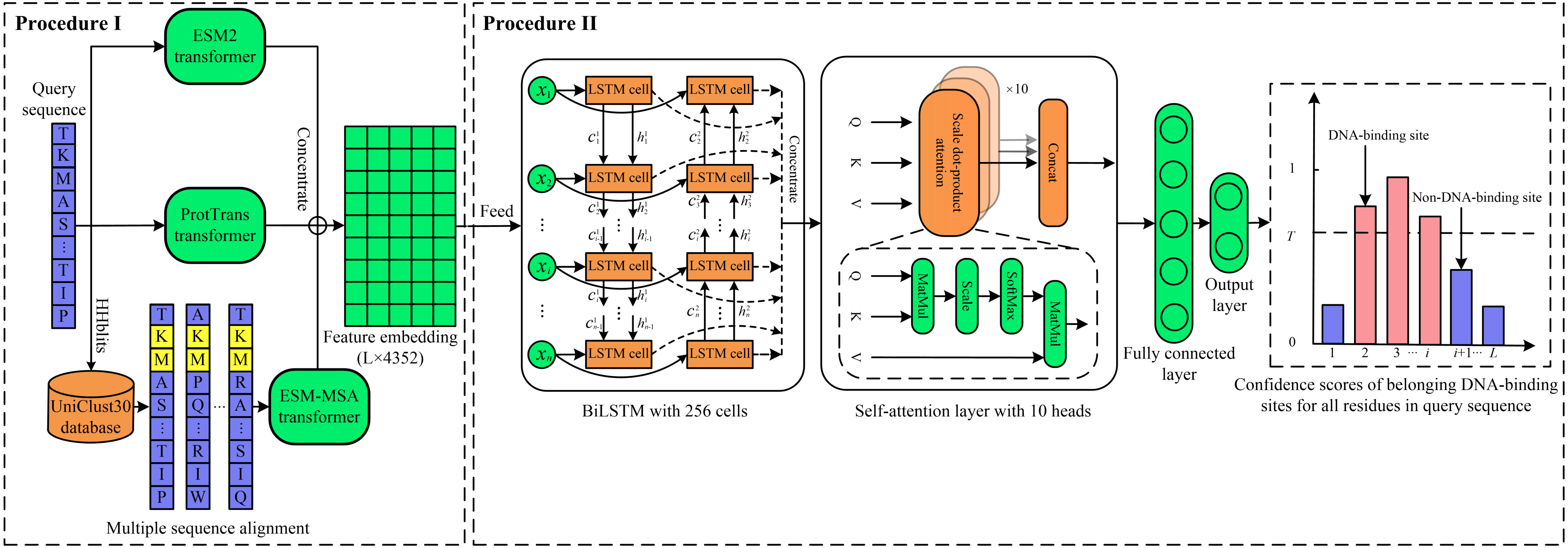
Specific position: **58 T 117 R 119 T 147 H**

Predicted Results

Residue #	Amino Acid Type	Probability	Binding Residue
0001	M	0.058	N
0002	D	0.028	N
0003	Q	0.024	N
0004	N	0.049	N
0005	E	0.013	N
0006	H	0.063	N
0007	S	0.008	N
0008	H	0.037	N
0009	W	0.095	N
0010	G	0.009	N
0011	P	0.019	N
0012	H	0.081	N
0013	A	0.017	N
0014	K	0.080	N
0015	G	0.006	N
0016	Q	0.013	N

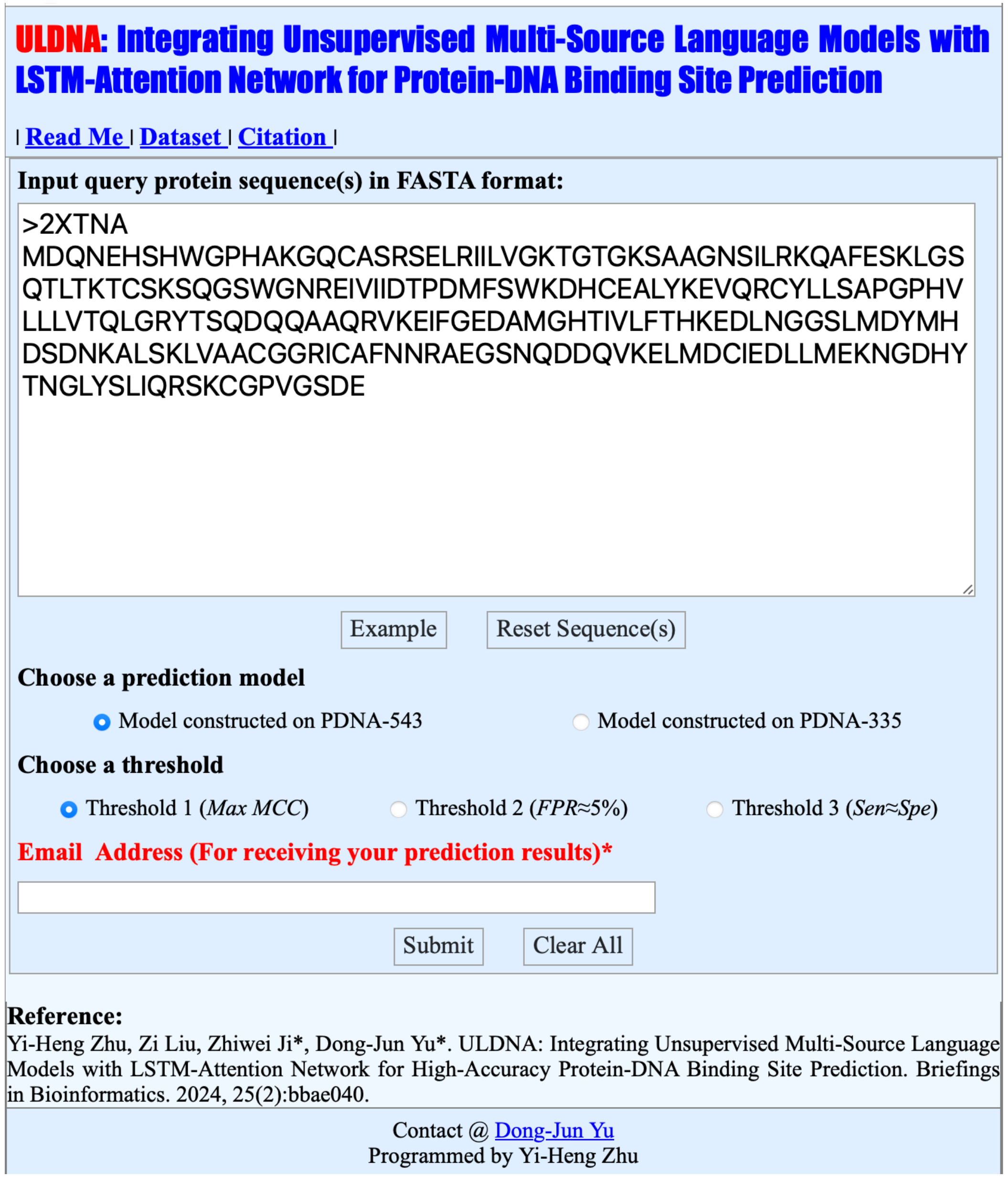
从配体视角出发预测功能

- 基于无监督语言模型与多源信息融合的蛋白质-DNA绑定位点预测方法 ULDNA
- 主要贡献：融合多种无监督蛋白质语言模型，显著地提升了蛋白质-DNA绑定位点预测精度。



Yi-Heng Zhu, Zi Liu, Zhiwei Ji, Dong-Jun Yu. ULDNA: Integrating Unsupervised Multi-Source Language Models with LSTM-Attention Network for High-Accuracy Protein-DNA Binding Site Prediction. **Briefings in Bioinformatics**. 2024, 25(2):bbae040.

从配体视角出发预测功能



ULDNA: Integrating Unsupervised Multi-Source Language Models with LSTM-Attention Network for Protein-DNA Binding Site Prediction

Input query protein sequence(s) in FASTA format:

```
>2XTNA
MDQNEHSHWGPHAKGQCASRSELRIILVGKTGTGKSAAGNSILRKQAFESKLGS
QTLTKTCSKSQGSWGNREIVIIDTPDMFSWKDHCEALYKEVQRCYLLSAPGPHV
LLLVTQLGRYTSQDQQAAQRVKEIFGEDAMGHTIVLFTHKEDLNGGSLMDYMH
DSDNKALSKLVAACGGRICAFNNRAEGSNQDDQVKELMDCIEDLLMEKNGDHY
TNGLYSLIQRSKCGPVGSDE
```

Choose a prediction model

● Model constructed on PDNA-543 ○ Model constructed on PDNA-335

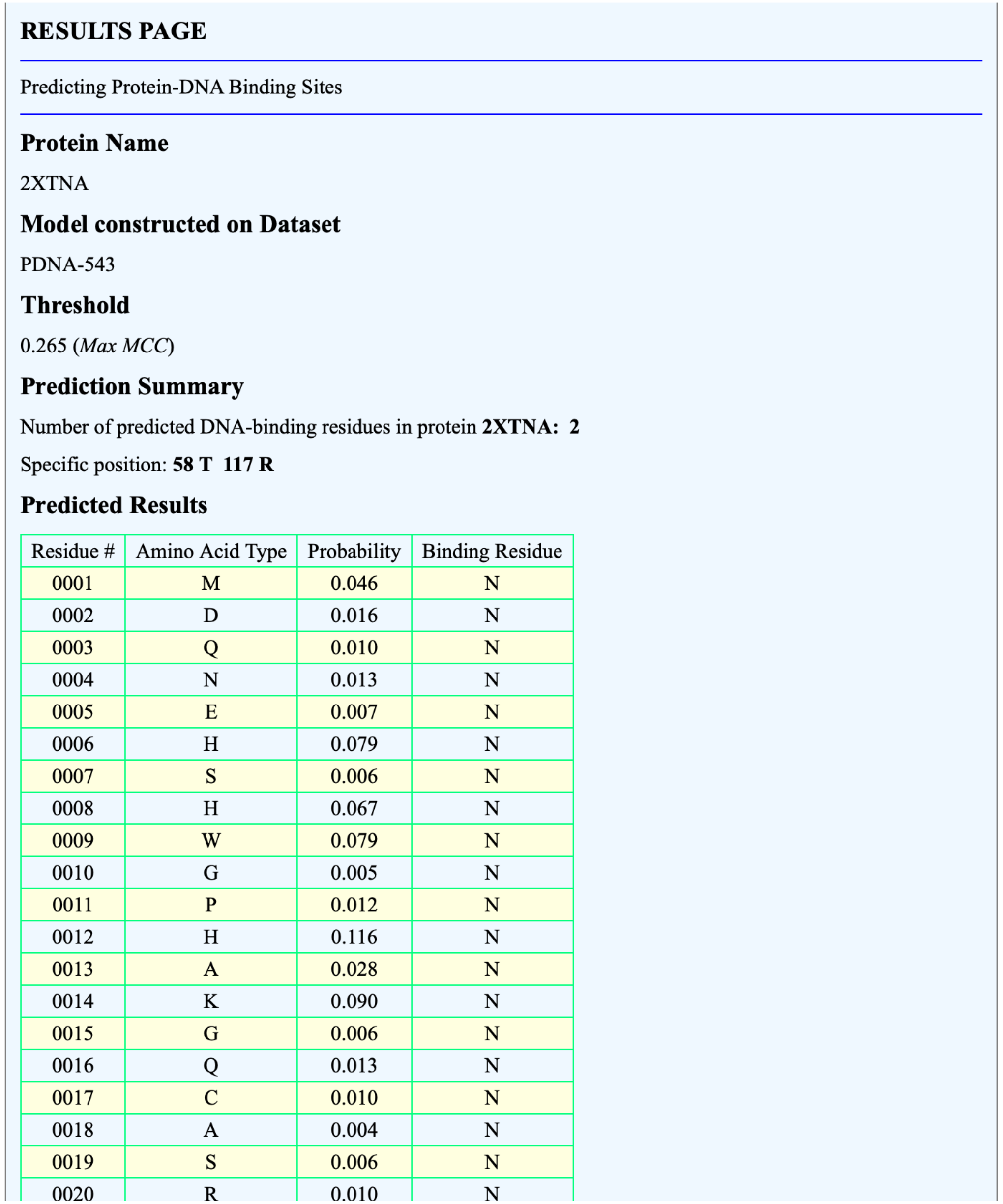
Choose a threshold

● Threshold 1 (*Max MCC*) ○ Threshold 2 (*FPR≈5%*) ○ Threshold 3 (*Sen≈Spe*)

Email Address (For receiving your prediction results)*

Reference:
Yi-Heng Zhu, Zi Liu, Zhiwei Ji*, Dong-Jun Yu*. ULDNA: Integrating Unsupervised Multi-Source Language Models with LSTM-Attention Network for High-Accuracy Protein-DNA Binding Site Prediction. Briefings in Bioinformatics. 2024, 25(2):bbae040.

Contact @ Dong-Jun Yu
Programmed by Yi-Heng Zhu



RESULTS PAGE

Predicting Protein-DNA Binding Sites

Protein Name

2XTNA

Model constructed on Dataset

PDNA-543

Threshold

0.265 (*Max MCC*)

Prediction Summary

Number of predicted DNA-binding residues in protein **2XTNA: 2**

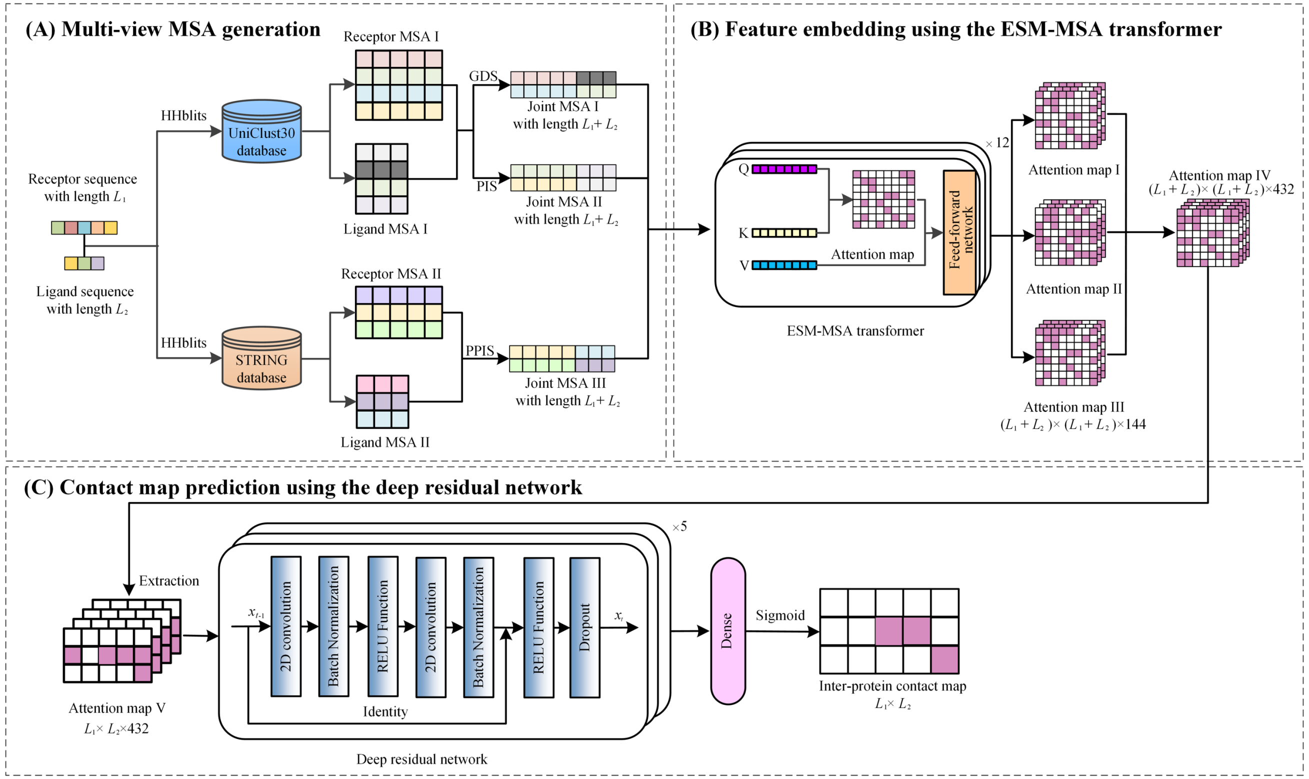
Specific position: **58 T 117 R**

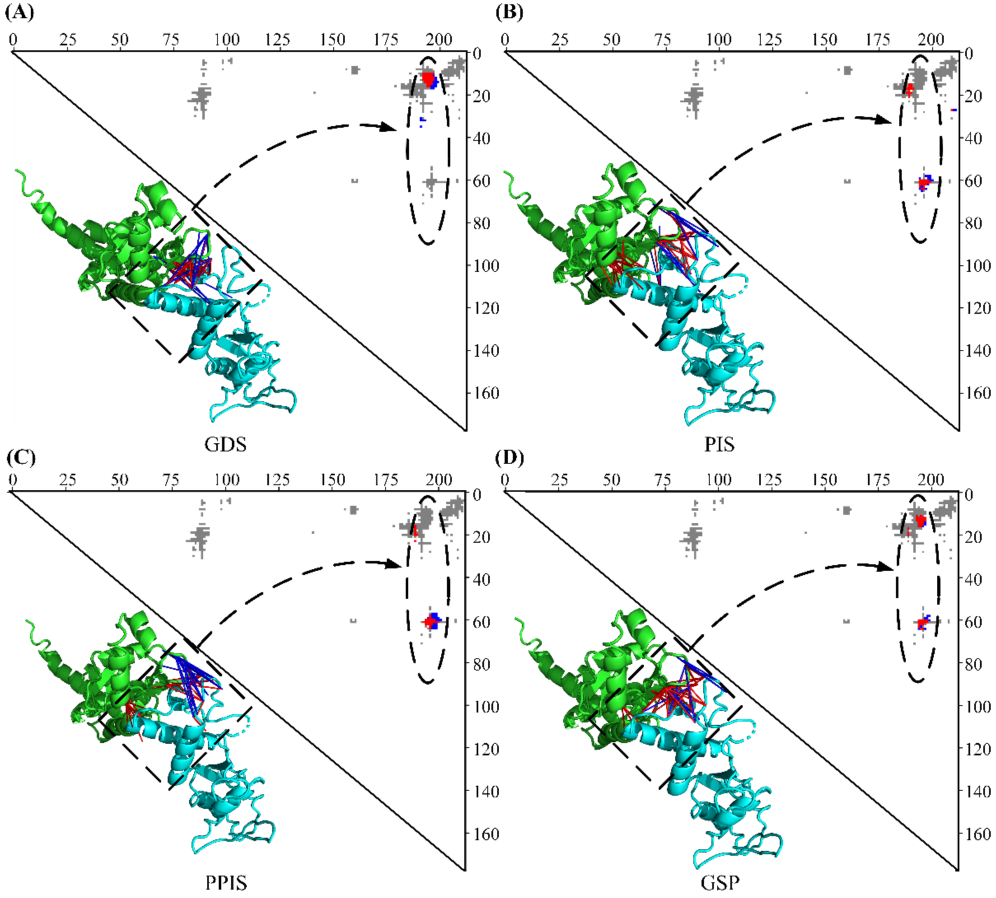
Predicted Results

Residue #	Amino Acid Type	Probability	Binding Residue
0001	M	0.046	N
0002	D	0.016	N
0003	Q	0.010	N
0004	N	0.013	N
0005	E	0.007	N
0006	H	0.079	N
0007	S	0.006	N
0008	H	0.067	N
0009	W	0.079	N
0010	G	0.005	N
0011	P	0.012	N
0012	H	0.116	N
0013	A	0.028	N
0014	K	0.090	N
0015	G	0.006	N
0016	Q	0.013	N
0017	C	0.010	N
0018	A	0.004	N
0019	S	0.006	N
0020	R	0.010	N

从配体视角出发预测功能

- ➢ 基于无监督语言模型与多视角多序列联配的蛋白质-蛋白质相互作用预测方法ICCPred





Zi Liu#, Yi-Heng Zhu#, Long-Chen Shen, Xuan Xiao, Wang-Ren Qiu, Dong-Jun Yu. Integrating Unsupervised Language Model with Multi-View Multiple Sequence Alignments for High-Accuracy Inter-Chain Contact Prediction. **Computers in Biology and Medicine**. 2023, 166: 107529

03 Part three

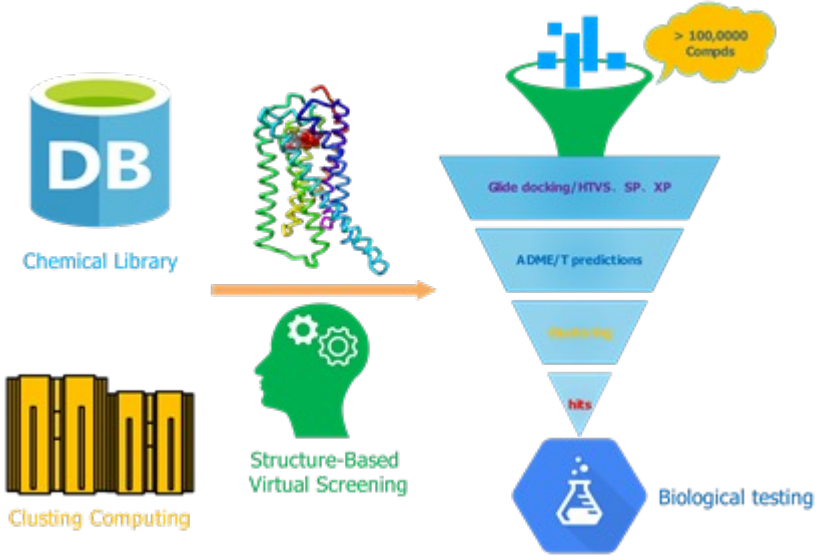
未来展望

蛋白质功能预测研究的应用前景

蛋白质功能预测研究有助于推动智能医疗的发展

(1) 辅助疾病分析和诊断 (推断关键致病蛋白质)

(2) 辅助药物设计（药物分子筛选）

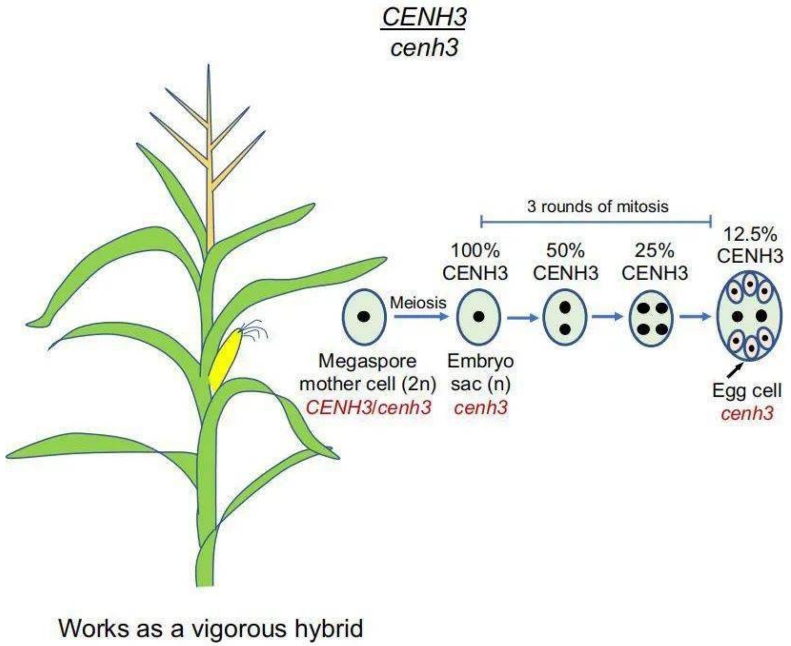


蛋白质功能预测在农业领域的应用前景

(1) 植物遗传育种

(2) 植物与微生物的相互作用

(3) 植物蛋白组学



团队介绍

团队组成：
教授1人
副教授1人
讲师3人
博士生3人
硕士生8人
已毕业研究生4人

研究方向

人工智能与模式识别

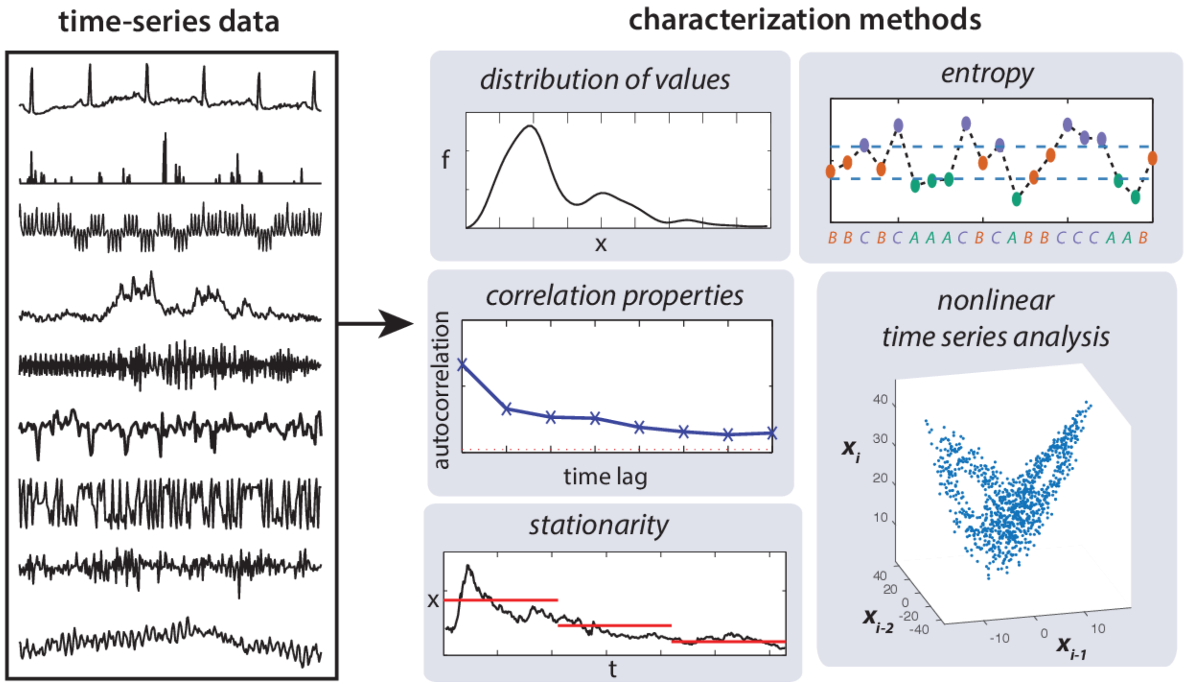
人工智能的理论及应用
大数据计算与模式识别

生物信息与系统生物学

多组学数据整合分析与计算
复杂生物系统的数学建模与预测

主要成果1：人工智能与模式识别领域

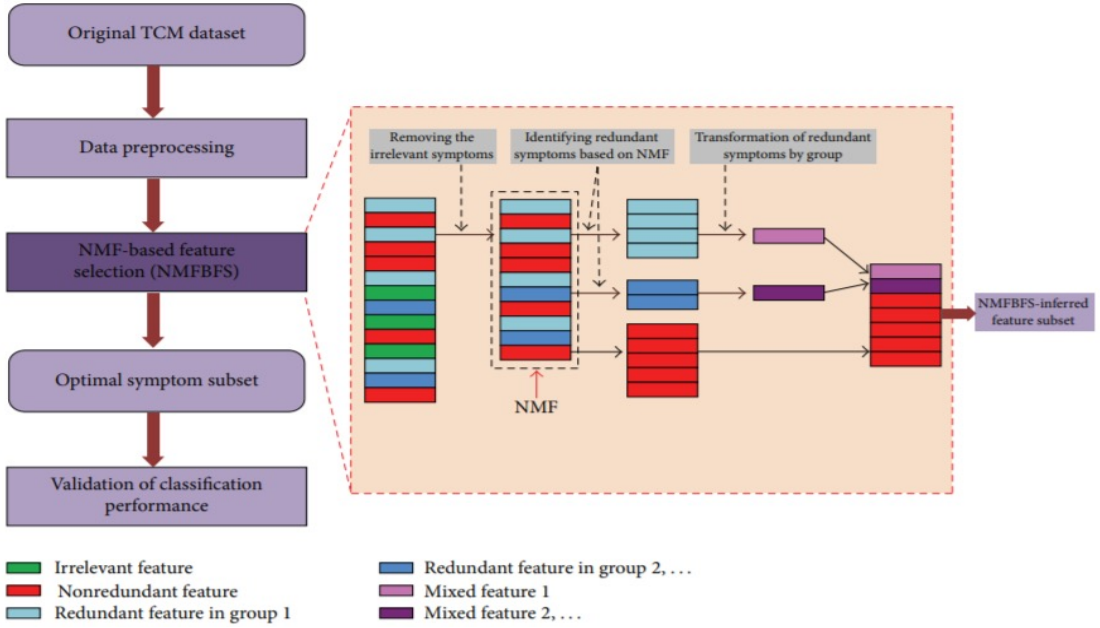
◆ 时间序列数据挖掘与异常模式发现



Expert Systems with Applications
TII
ies
INFORMATION SCIENCES
SYSTEMS, MAN, AND CYBERNETICS
NEUROCOMPUTING

Z Ji*, Y Wang, X Xie, et al., ***Expert Systems with Applications***, 2022.
N Jin, Y Zeng, K Yan, **Z Ji**, ***IEEE Transactions on Industrial Informatics***, 2021.
M Hu, X Feng, **Z Ji***, et al., ***Information Sciences***, 2019.
K Yan#, **Z Ji#**, et al., ***IEEE Transactions on Systems, Man, and Cybernetics: Systems***, 2019.
K Yan, **Z Ji***, et al., ***Neurocomputing***, 2017.

◆ 高维复杂数据的维度约简和模型优化


Original TCM dataset
Data preprocessing
NMF-based feature selection (NMFBFS)
Optimal symptom subset
Validation of classification performance
Removing the irrelevant symptoms
Identifying redundant symptoms based on NMF
Transformation of redundant symptoms by group
NMF
NMFBFS-inferred feature subset

🟩 Irrelevant feature
🟥 Nonredundant feature
🟦 Redundant feature in group 1
🟦 Redundant feature in group 2, . . .
🟪 Mixed feature 1
🟪 Mixed feature 2, . . .

Plant Phenomics
NANJING AGRICULTURAL UNIVERSITY
Frontiers in Plant Sciences
COMPUTATIONAL BIOLOGY AND BIOINFORMATICS
Knowledge-Based SYSTEMS
Computational and Mathematical Methods in Medicine

X Xie, F Xia, K Yan, H Xu, **Z Ji***, ***Plant Phenomics***, 2023.
F Xia, X Xie , S Jin, K Yan, **Z Ji***, ***Frontiers in Plant Science***, 2021.
K Yan, ..., **Z Ji.**, et al., ***IEEE/ACM Trans on Computational Biology and Bioinformatics,*** 2021.
X Xie, X Gu, Y Li, **Z Ji***, ***Knowledge-based Systems***, 2021.
Z Ji, ..., B Wang*, ***Computational and mathematical methods in medicine***, 2015.

主要成果2：生物信息与系统生物学领域

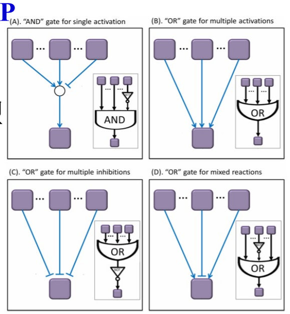
◆ 创立了一套独特的**生物分子网络**建模方法

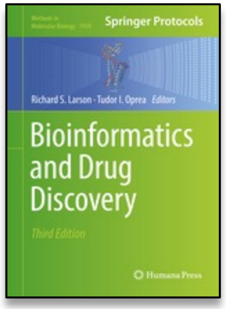
◆ 创立了**分子-细胞-组织**的3D多尺度建模方法

◆ 建立了**生物组学大数据挖掘**的计算框架

基于线性规划的离散时间建模方法BLP,DILP, TILP, MIP

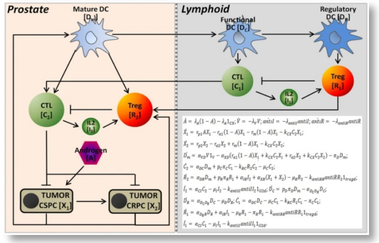
BLP模型


(A). "AND" gate for single activation
(B). "OR" gate for multiple activations
(C). "OR" gate for multiple inhibitions
(D). "OR" gate for mixed reactions
AND
OR

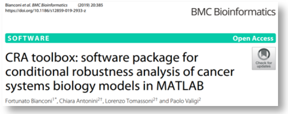

Springer Protocols
Methods in Molecular Biology 1939
Richard S. Larson · Tudor I. Oprea Editors
Bioinformatics and Drug Discovery
Third Edition
Humana Press

✓ **BLP作为经典模型**被写入了Springer教材Methods in Molecular Biology丛书之一《**Bioinformatics and Drug Discovery**》(第16章第287页)

基于微分方程的连续时间建模方法



Cell-cell interaction模型



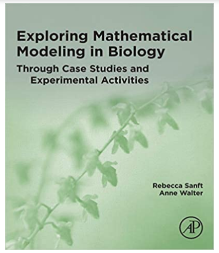
Bianconi et al. BMC Bioinformatics (2019) 20:385
https://doi.org/10.1186/s12859-019-2933-z

BMC Bioinformatics

SOFTWARE Open Access

CRA toolbox: software package for conditional robustness analysis of cancer systems biology models in MATLAB

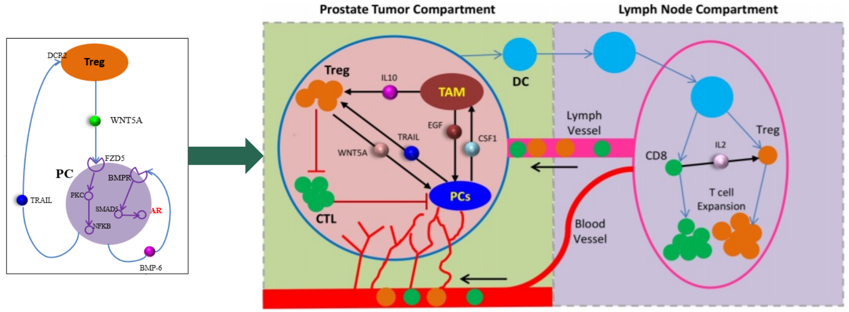
Fortunato Bianconi[1*], Chiara Antonini[2†], Lorenzo Tomassoni[2†] and Paolo Valigi[2]


Exploring Mathematical Modeling in Biology
Through Case Studies and Experimental Activities
Rebecca Sanft
Anne Walter

✓ 该模型被Matlab工具包**CRA**收录，并在BMC Bioinformatics进行长篇报道

✓ 作为**经典模型**被写入**Elsevier**教材(2020年)：《**Exploring Mathematical Modeling in Biology**》(第2章第54页)

发现了WNT5A调控CRPC（前列腺癌）进展的新机制
构建了面向分子-细胞-组织的多尺度3D模型HABM



"WNT5A在免疫调节中发挥重要作用"

International Immunopharmacology 81 (2020) 106274

Contents lists available at ScienceDirect

International Immunopharmacology

journal homepage: www.elsevier.com/locate/intimp

ELSEVIER

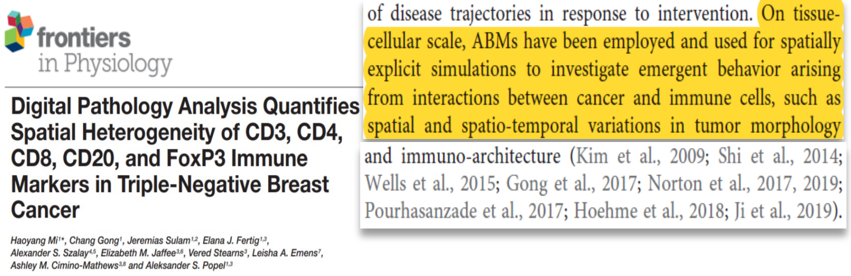
Development and validation of a prognostic immune-associated gen signature in clear cell renal cell carcinoma

Chengquan Shen[a,1], Jing Liu[b,1], Jirong Wang[a], Xiulong Zhong[a], Dahai Dong[a], Xiaokun Yonghua Wang[a,*]

Results: A total of 681 differentially expressed IAGs were identified and sev (*IFI30, WNT5A, IRF9, AGER, PLAUR, TEK, BID*) were finally selected in a IAGs signature. Survival analysis reve that high IAGs risk scores

disease-free survival in breast cancer patients [13]. *WNT5A* belongs to the large WNT family of cysteine-rich secreted glycoproteins, which is involved in multiple signaling pathways that regulate a variety of cellular processes [14]. Ji et al. revealed that *WNT5A* could mediate the activation of Treg and TAM cells, which induced the immunosuppression during castration-resistant prostate cancer progression [15]. Chen

✓ **HABM是十几年来首个对肿瘤生长-免疫反应-血管生成进行3D时空建模的数学模型**



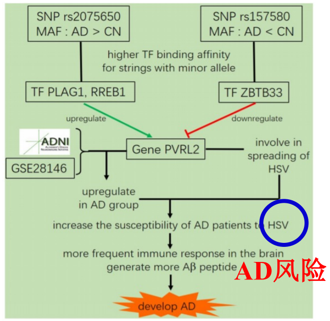
frontiers in Physiology

Digital Pathology Analysis Quantifies Spatial Heterogeneity of CD3, CD4, CD8, CD20, and FoxP3 Immune Markers in Triple-Negative Breast Cancer

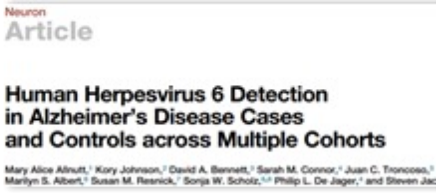
Haoyang Mi[1*], Chang Gong[1], Jeremias Sulam[1,2], Elana J. Fertig[1,3], Alexander S. Szalay[4,5], Elizabeth M. Jaffee[3,6], Vered Stearns[3], Leisha A. Emens[7], Ashley M. Cimino-Mathews[3,8] and Aleksander S. Popel[1,3]

of disease trajectories in response to intervention. On tissue-cellular scale, ABMs have been employed and used for spatially explicit simulations to investigate emergent behavior arising from interactions between cancer and immune cells, such as spatial and spatio-temporal variations in tumor morphology and immuno-architecture (Kim et al., 2009; Shi et al., 2014; Wells et al., 2015; Gong et al., 2017; Norton et al., 2017, 2019; Pourhasanzade et al., 2017; Hoehme et al., 2018; Ji et al., 2019).

组学大数据挖掘发现AD潜在风险因子 (TREM2, HSV-1)



AD风险因子



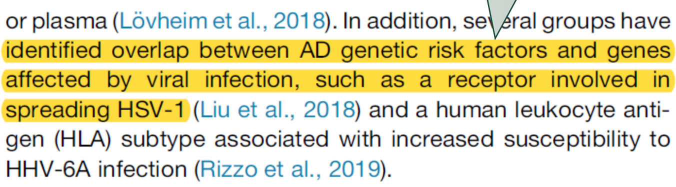
Neuron

Article

Human Herpesvirus 6 Detection in Alzheimer's Disease Cases and Controls across Multiple Cohorts

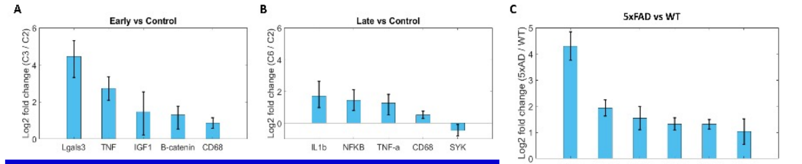
Mary Alice Allnutt,[1] Kory Johnson,[2] David A. Bennett,[3] Sarah M. Connor,[4] Juan C. Troncoso,[5] Marilyn S. Albert,[5] Susan M. Resnick,[7] Sonja W. Scholz,[6,7] Philip L. De Jager,[4] and Steven Jac

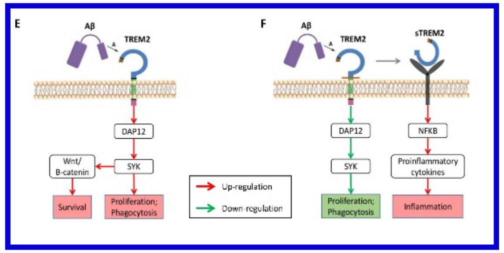
是关于"HSV-1是AD风险因子"的较早报道之一

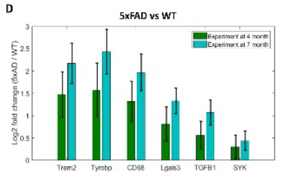


or plasma (Lövheim et al., 2018). In addition, several groups have identified overlap between AD genetic risk factors and genes affected by viral infection, such as a receptor involved in spreading HSV-1 (Liu et al., 2018) and a human leukocyte antigen (HLA) subtype associated with increased susceptibility to HHV-6A infection (Rizzo et al., 2019).

✓ **时序组学大数据挖掘，首次解析了TREM2调控Microglia表型转换的分子机制**







谢谢各位老师和同学观看

请批评指正！