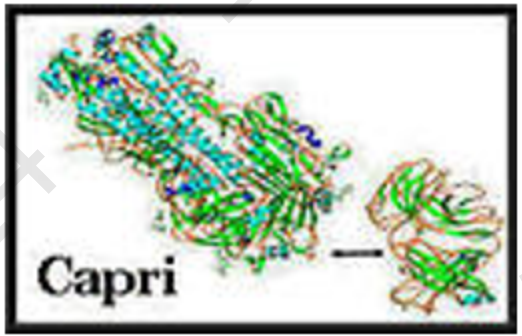
生物统计学：
生物信息中的概率统计模型

2024年秋



CASP14

有关信息

- 授课教师：宁康
  - Email: ningkang@hust.edu.cn
  - Office: 华中科技大学东十一楼606室
  - Phone: 18627968927

- 课程网页
  - http://www.microbioinformatics.org/Biostatistics.html
  - QQ群：717914581

课程安排

- 生物背景和课程简介
- 传统生物统计学及其应用
- 生物统计学和生物大数据挖掘
  - Hidden Markov Model (HMM)及其应用
    - Markov Chain
    - HMM理论
    - HMM和基因识别 (Topic I)
    - HMM和序列比对 (Topic II)
  - 进化树的概率模型 (Topic III)
  - Motif finding中的概率模型 (Topic IV)
    - EM algorithm
    - Markov Chain Monte Carlo (MCMC)
  - 基因表达数据分析 (Topic V)
    - 聚类分析-Mixture model
    - Classification-Lasso Based variable selection
  - 基因网络推断 (Topic VI)
    - Bayesian网络
    - Gaussian Graphical Model
  - 基因网络分析 (Topic VII)
    - Network clustering
    - Network Motif
    - Markov random field (MRF)
  - Dimension reduction及其应用 (Topic VIII)
- 面向生物大数据挖掘的深度学习

研究对象:
生物序列,
进化树,
生物网络,
基因表达
...

方法:
生物计算与生物统计

第8-1章:蛋白质相互作用的实验方法和预测

- Experimental methods
- Prediction of protein-protein interactions

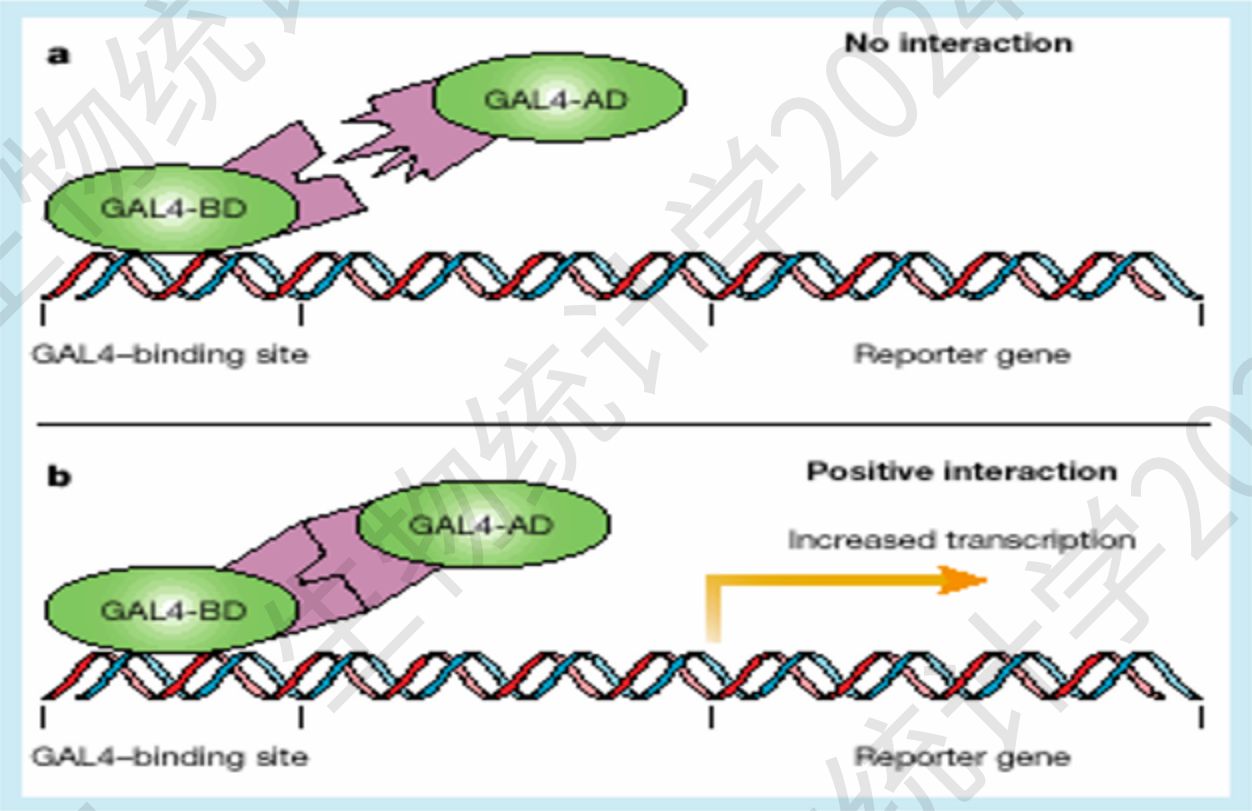
Part I: Experimental Methods

- Physical interaction
  - Yeast two hybrid system
  - TAP-mass spectrometry
- Genetic interaction
  - SGA
  - EMAP

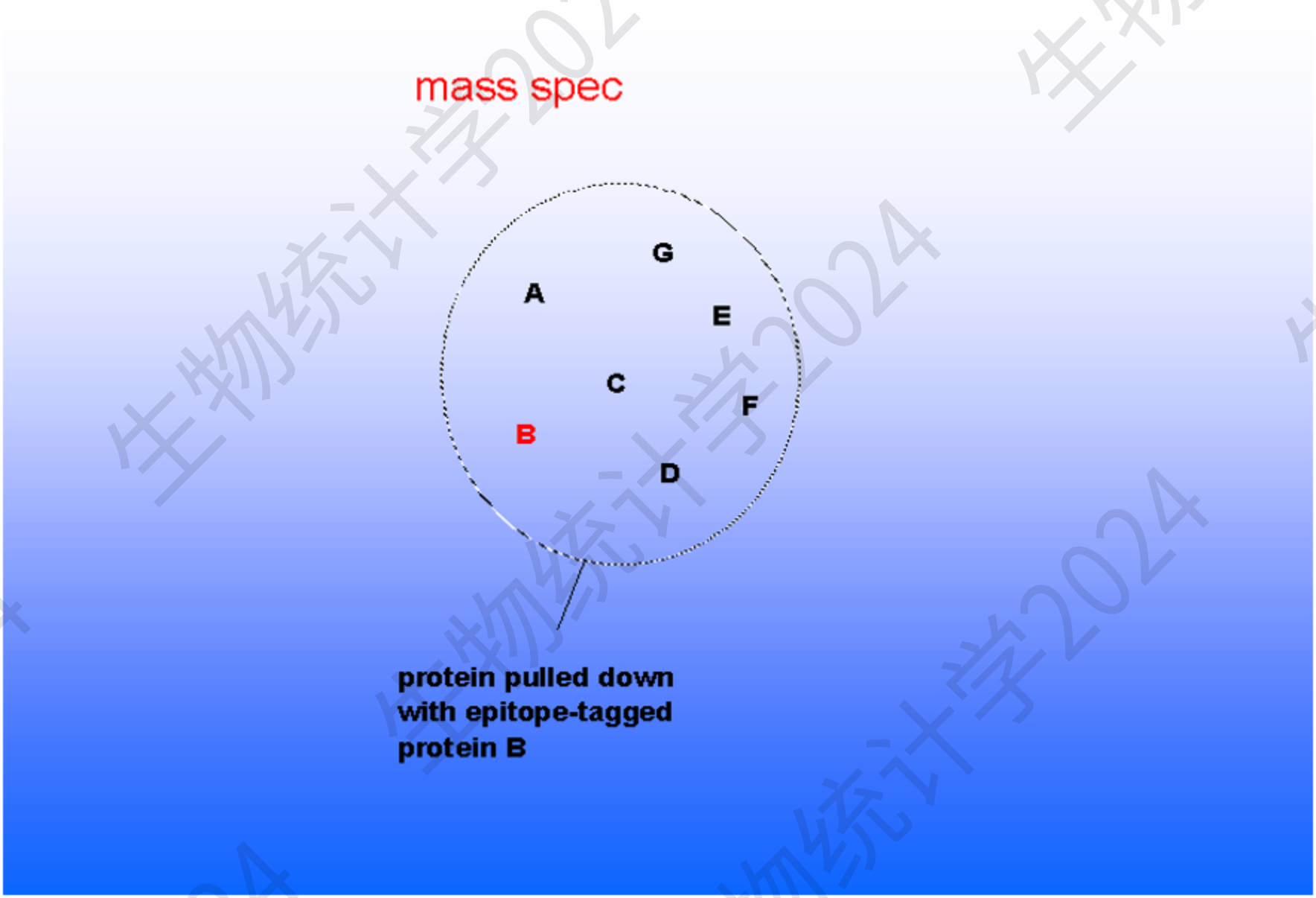
Protein-protein interactions (Experimental methods)

- Co-immunoprecipitation.
- Two-hybrid system (Uetz et al. 2000, Ito et al. 2000, 2001).
- Purified Complex by mass spectrometry
  - TAP: Tandem affinity purification (Gavin et al. 2002).
  - HMS-PCI: high-throughput mass spectrometric protein complex identification (Ho et al. 2002).

Mechanism of two-hybrid system



From: Nature 405, June 15, 2000, 837-846.



Gavin et al. (2002) Nature 415:141

Mass spectrometry

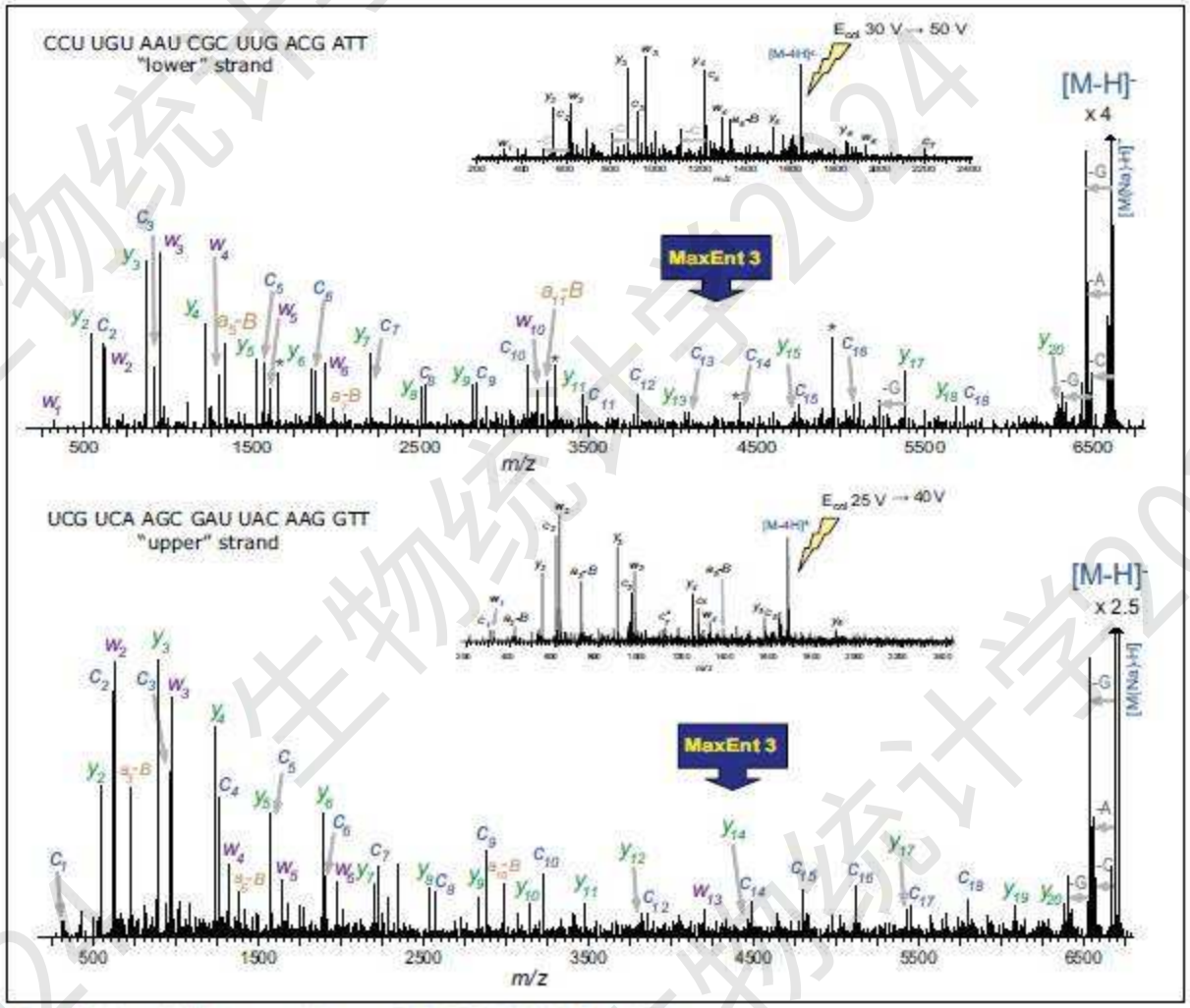
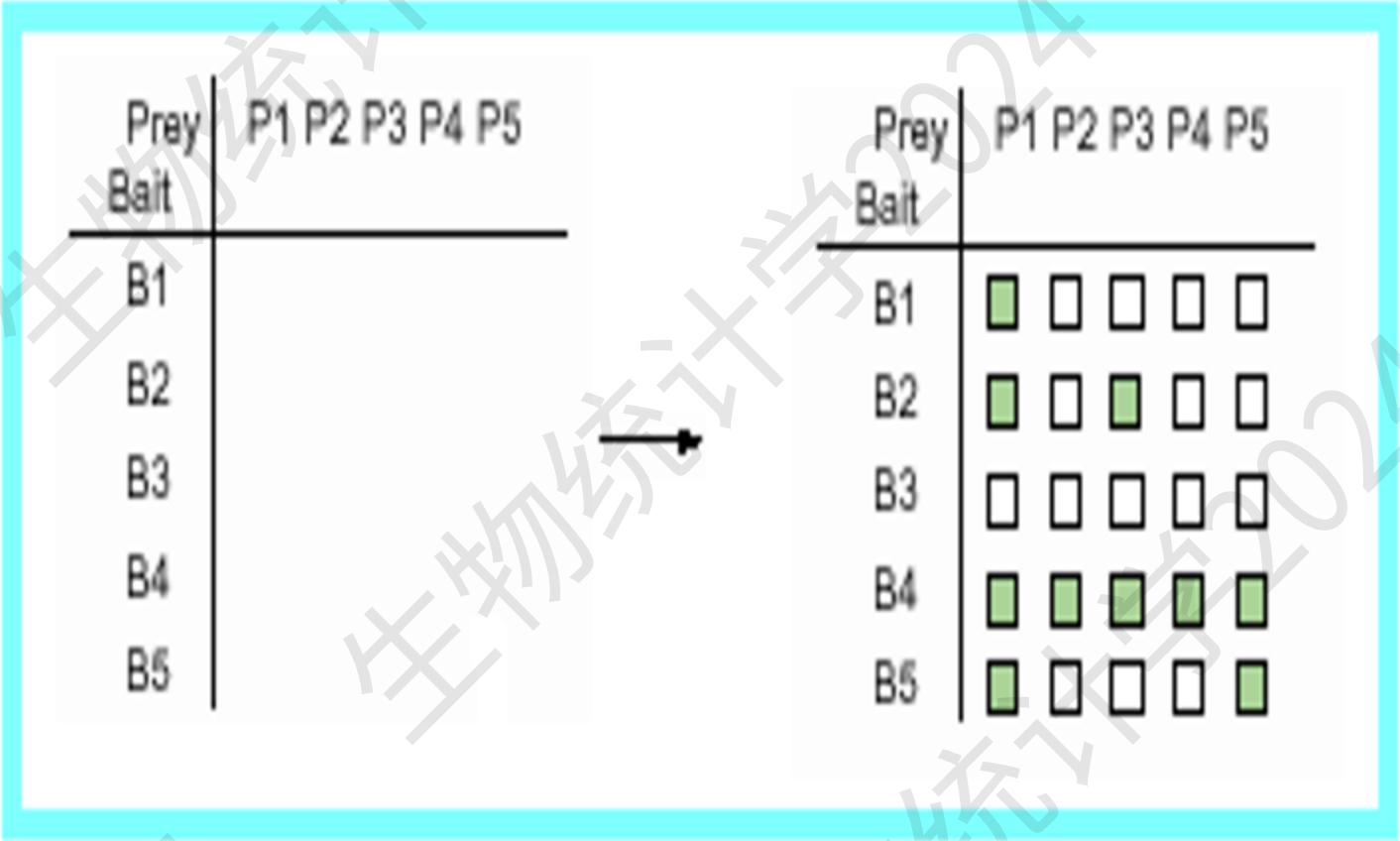


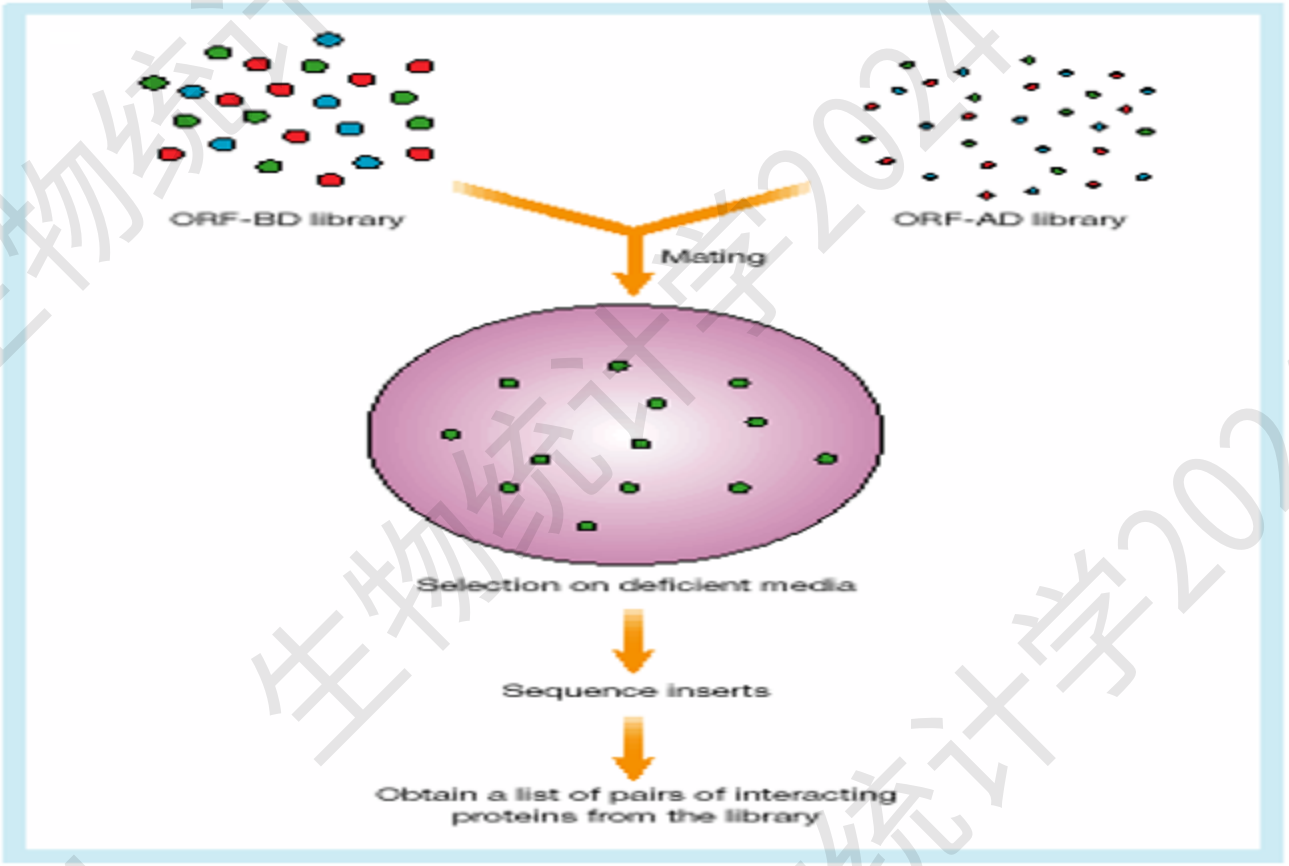
图2 21 核苷酸RNA寡核苷酸的MaxEnt3去卷积MS/MS图谱。星号为谐波峰(去卷积的结果)

Matrix method (two hybrid)



From: TRENDS in Genetics Vol.17, No.6, June 2001.

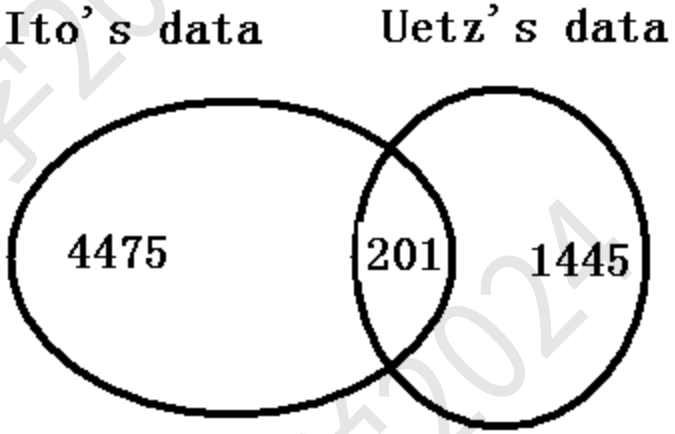
Interaction Sequence Tags (ISTs)



From: Nature 405, June 15, 2000, 837-846.

Two data sets from yeast two hybrid system

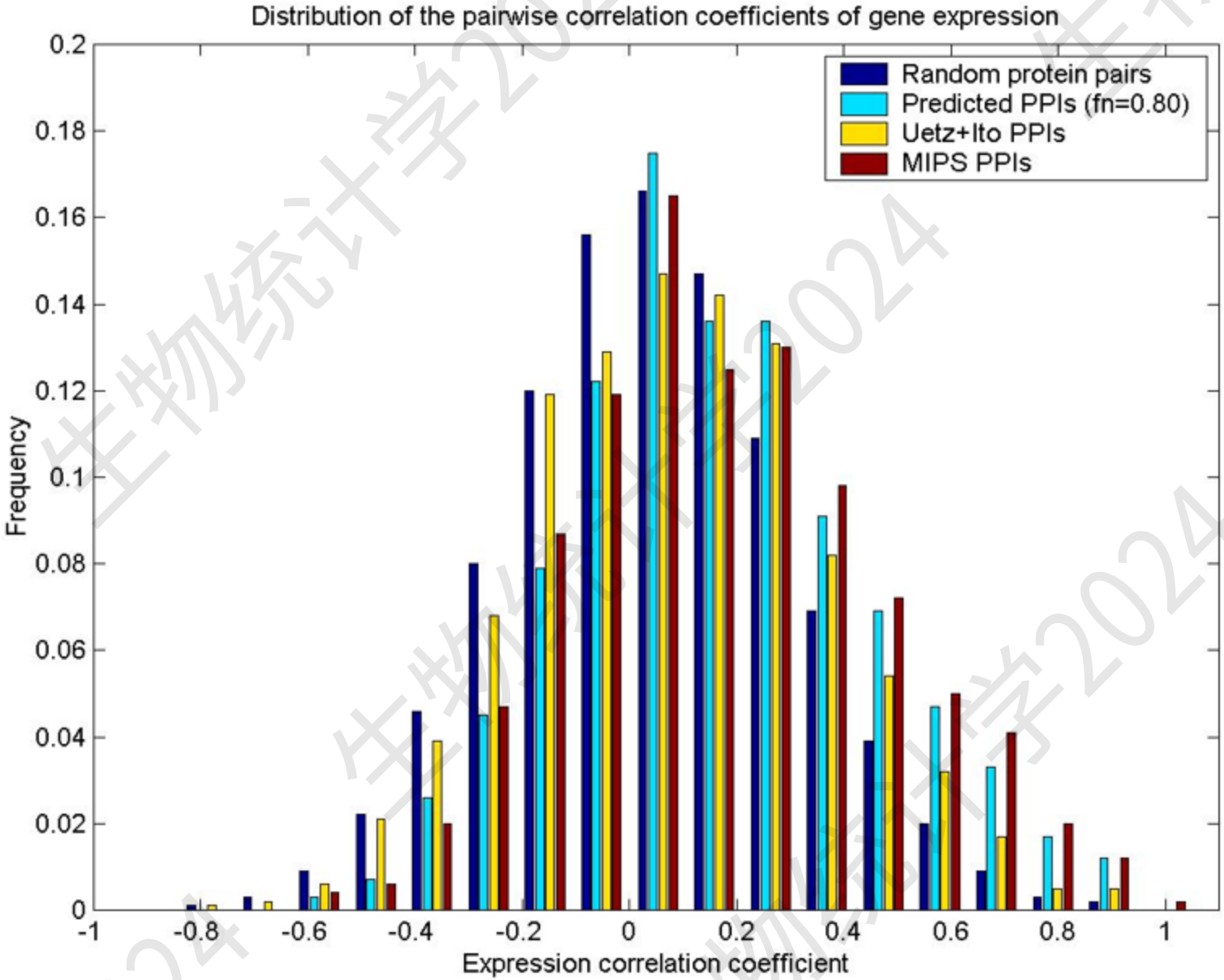
- Uetz's data (Uetz et al. 2000).
- Ito's data (Ito et al. 2000, 2001).

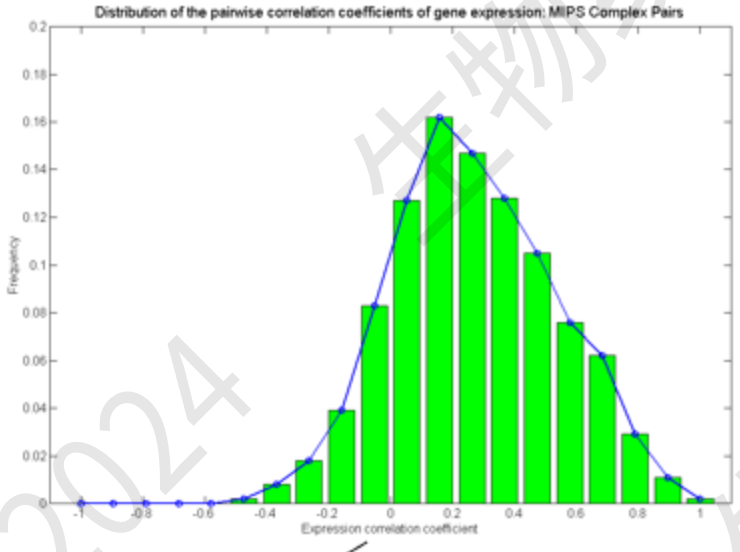


Possible Errors in 2-hybrid system

- False positive.
  - Possible mutation during PCR-amplifying.
  - Stochastic activation of reporter gene.
- False negative.
  - Membrane protein, post-translational modification protein, those self-activating reporter genes (Removed in experiment).
  - Weak interactions.

The size of interactome for yeast (5-50/protein)


Distribution of the pairwise correlation coefficients of gene expression
Random protein pairs
Predicted PPIs (fn=0.80)
Uetz+Ito PPIs
MIPS PPIs
Frequency
0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1
Expression correlation coefficient

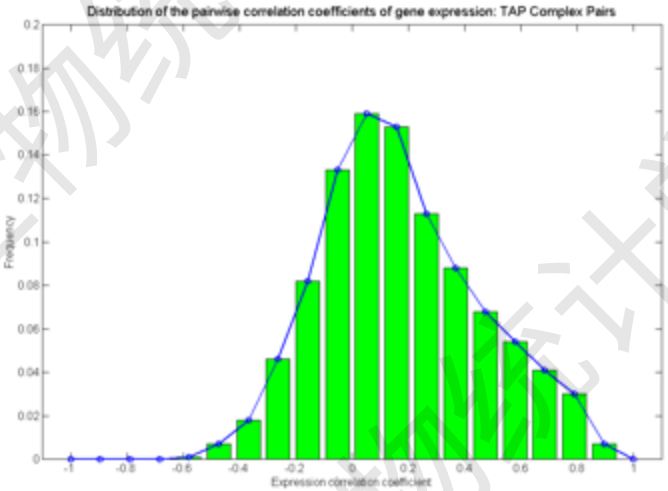


Non-interaction

$1-\alpha$

α

Real interaction



Observed interaction

MLE of the reliability

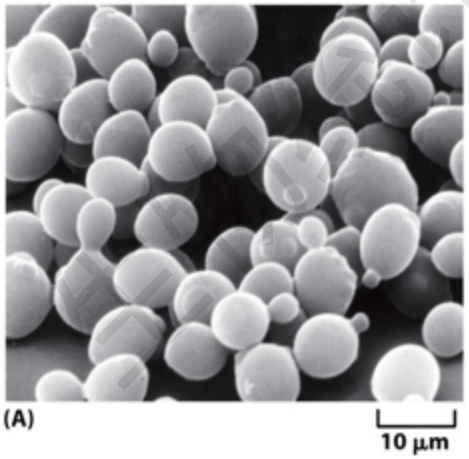
- Likelihood function

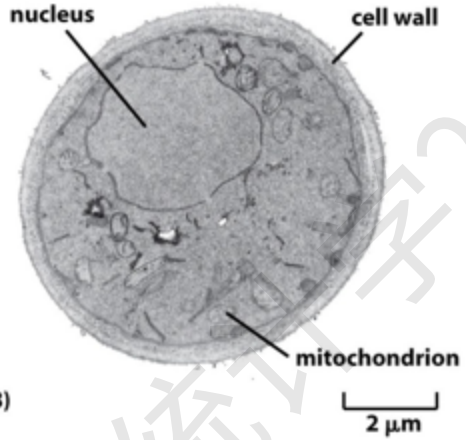
$$L(\alpha) = \prod_{k=1}^{K} (\alpha p_k + (1-\alpha) q_k)^{n_k}$$

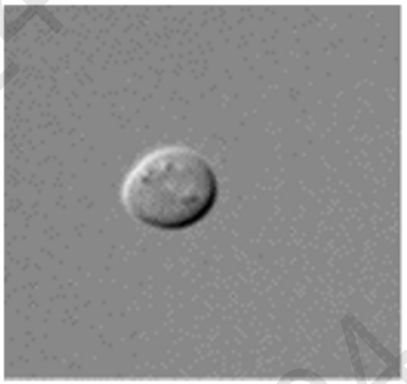
- Precision of the estimation

$$Var(\widehat{\alpha}) = \frac{1}{\sum_{k=1}^{K} n_k \frac{(p_k - q_k)^2}{(\widehat{\alpha} p_k + (1-\widehat{\alpha}) q_k)^2}}$$

Budding Yeast
Saccharomyces Cerevisiae

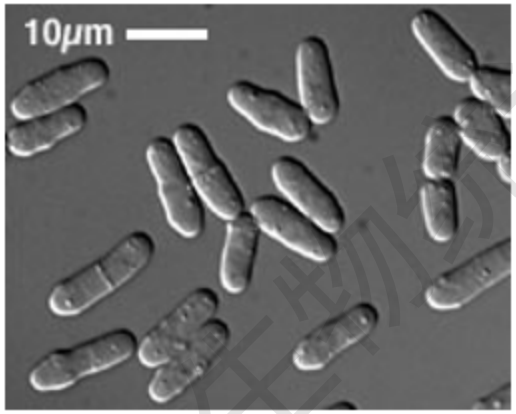


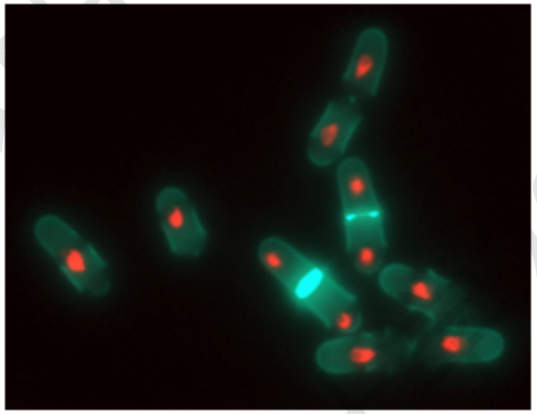


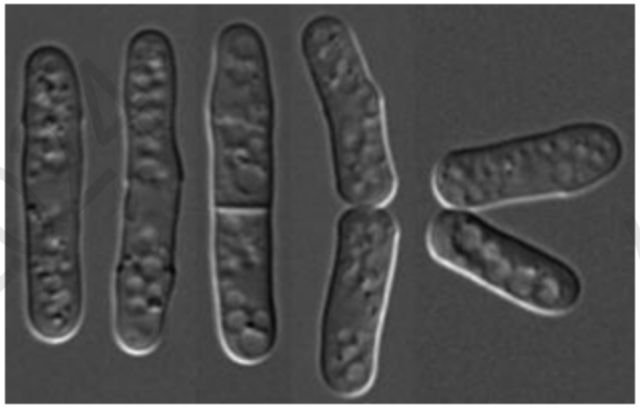


- a and α mating type, cell cycle
- 6300 genes (1997)
- Genome-wide single mutants analysis (2000~)

Fission yeast
Schizosaccharomyces Pombe







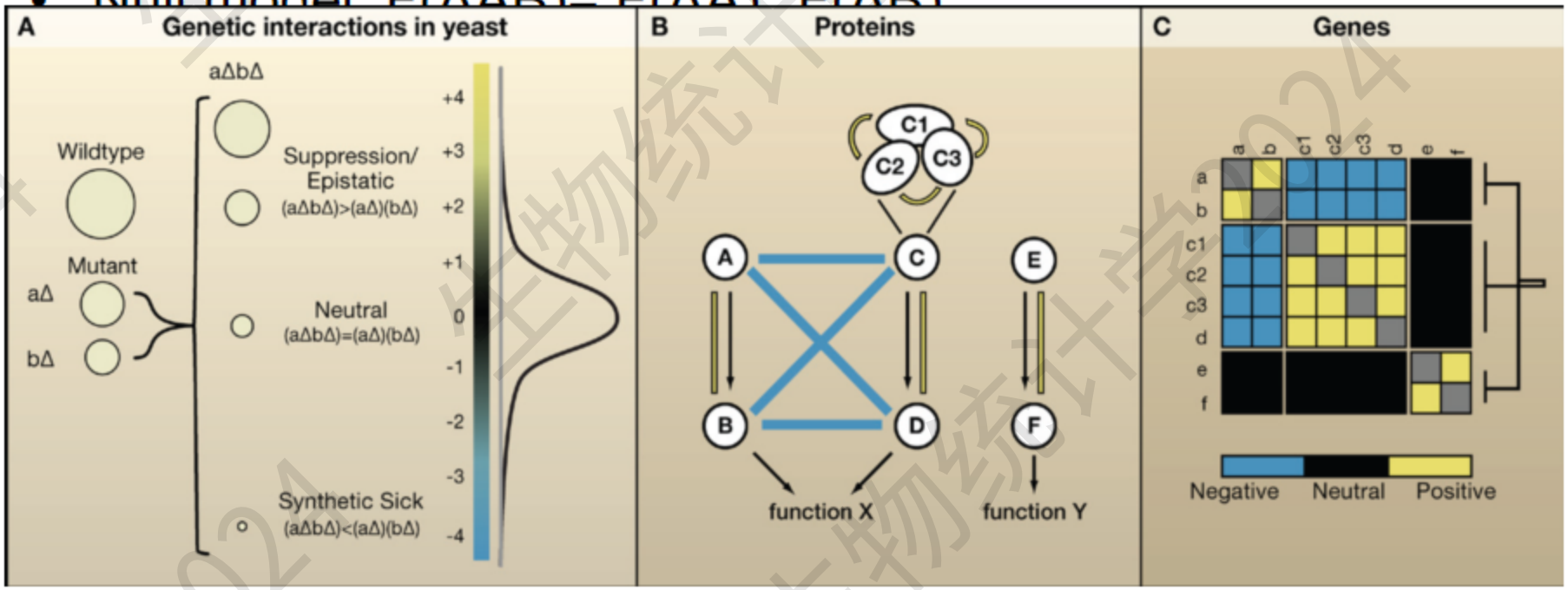
- 1000 million years separation from budding yeast;
- 13.8 Mb genome size, 4824 genes (open reading frames, OPF);
- 3 chromosomes, no genome-wide duplications; h+ and h- mating types;
- Cell cycle: 10% G1, 10% S, 70% G2 and 10% M phases.
- Genome-wide single mutants analysis (2010~)

more similar to metazoans than *S. cerevisiae*

- *cell cycle* regulation in G2/ M phase,
- gene regulation by the RNAi pathway
- the widespread presence of introns in genes

What's Genetic Interaction

- Genetic interactions between two loci can be mapped by measuring how the phenotype of an organism lacking both genes (double mutant) differs from that expected when the phenotypes of the single mutations are combined
- Null model: F(ΔAB)= F(ΔA)* F(ΔB)



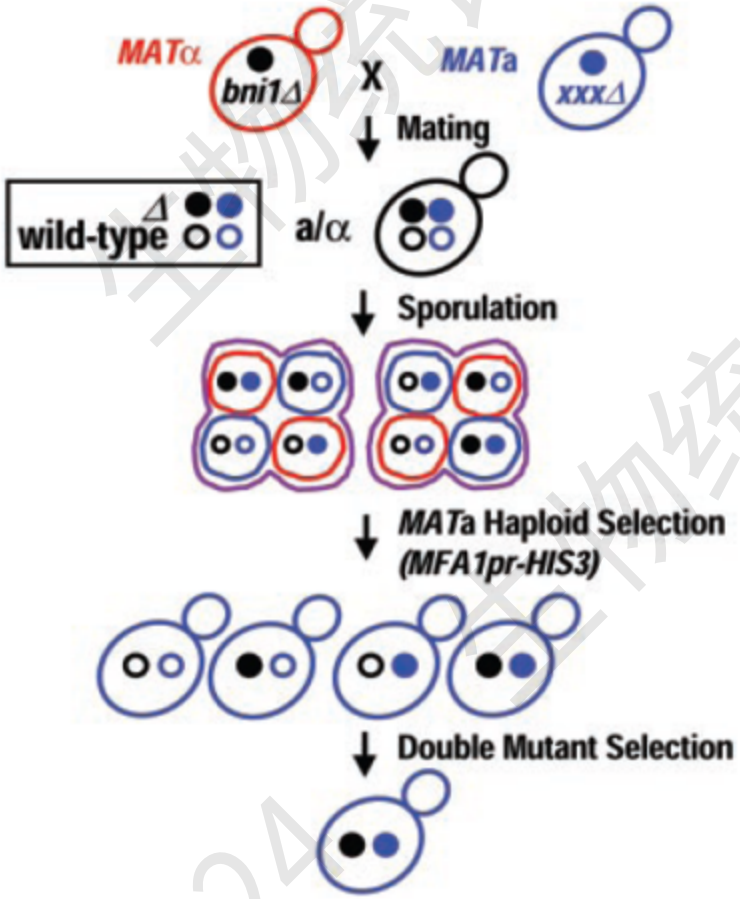
Beltrao et al. Cell 141: 739-745,

Identification of Genetic Interactions

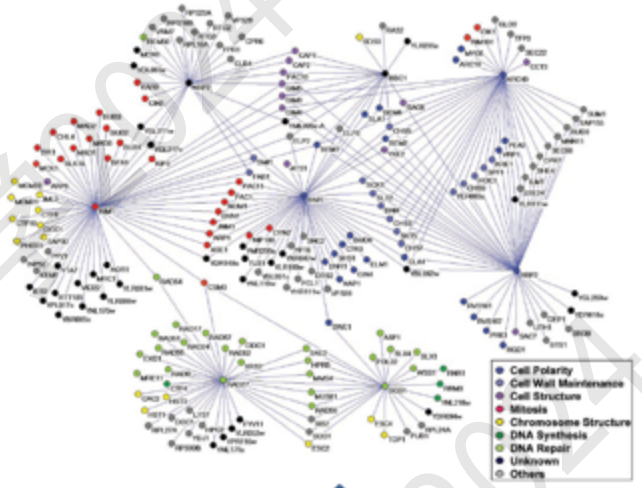
- Synthetic Gene Array (SGA) (Tong, et al. 2001)
- Diploid based Synthetic Lethality Analysis on Microarrays (dSLAM) (Pan, X., et al. 2004)

Synthetic Gene Array (SGA)

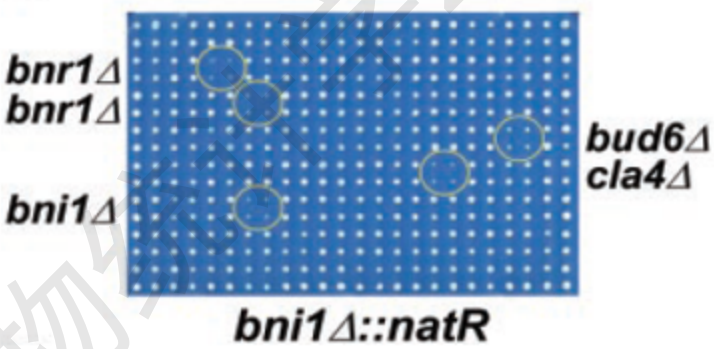
Synthetic genetic array methodology



Genetic Interaction Network



Final Screen



Amy Hin Yan Tong, et al. *Science*, 2001.

EMAP is the Extension of SGA

- EMAP: Epistatic Miniarray Profiles (Maya Schuldiner, et al. 2005. *Cell*)
- Quantitative measurement of phenotype (colony size)
  - Measure both positive and negative interactions.
  - Genome-wide (DAMP for lethal genes).

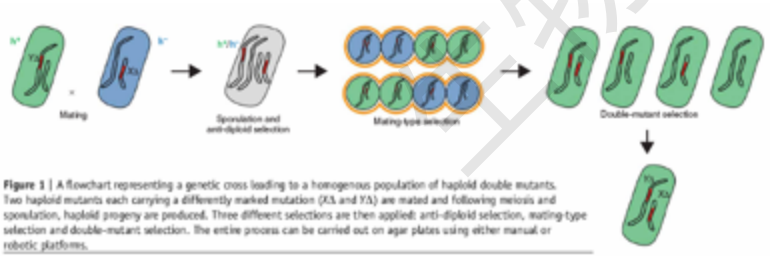
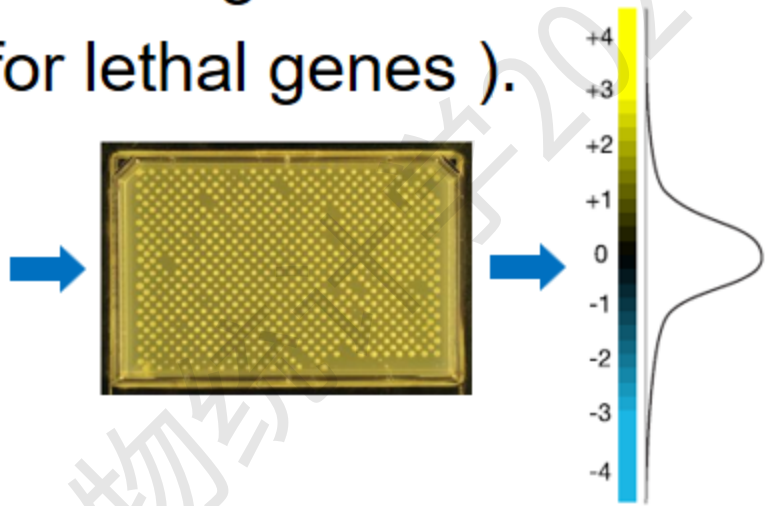


Figure 1 | A flowchart representing a genetic cross leading to a homogenous population of haploid double mutants. Two haploid mutants each carrying a differently marked mutation (XΔ and YΔ) are mated and following meiosis and sporulation, haploid progeny are produced. Three different selections are then applied: anti-diploid selection, mating-type selection and double-mutant selection. The entire process can be carried out on agar plates using either manual or robotic platforms.



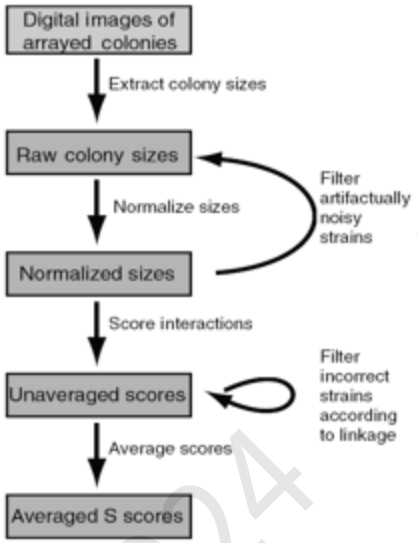
Dorothea Fiedler, et al. *Cell, 2009*

EMAP S-score

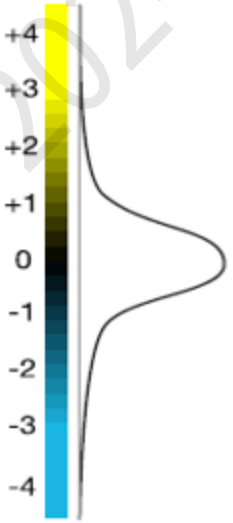
- Quantitative measure: $\epsilon = W_{ab} - W_a W_b,\ W_a = w/w_{wild}.$

No interaction	Synthetic sick/lethality	Synthetic alleviating
$\epsilon = 0$	$\epsilon < 0$	$\epsilon > 0$

  - T-Test with null hypothesis: $\epsilon = 0$



$$S_0 = \frac{\mu_{ab} - w_a \mu_b}{\sqrt{n_1 S_{ab}^2 + n_2 w_a^2 S_b^2}}$$



Sean R Collins, et al. *Genome Biology 7: R63,*

PPI databases

- MIPS: Munich Information center for Protein Sequences (http://mips.gsf.de)
- DIP: Database of Interacting Proteins (http://dip.doe-mbi.ucla.edu)
- BIND: Biomolecular Interaction Network Database (http://www.bind.ca)
- GRID: General Repository for Interaction Datasets (http://biodata.mshri.on.ca/grid)
- MINT: Molecular Interaction Database (http://cbm.bio.uniroma2.it/mint/)

Further Reading

- For more experimental methods and databases, please read the following review paper
  - Shoemaker BA, Panchenko AR (2007) Deciphering Protein–Protein Interactions. Part I. Experimental Techniques and Databases. PLoS Comput Biol 3(3): e42. doi:10.1371/journal.pcbi.0030042.

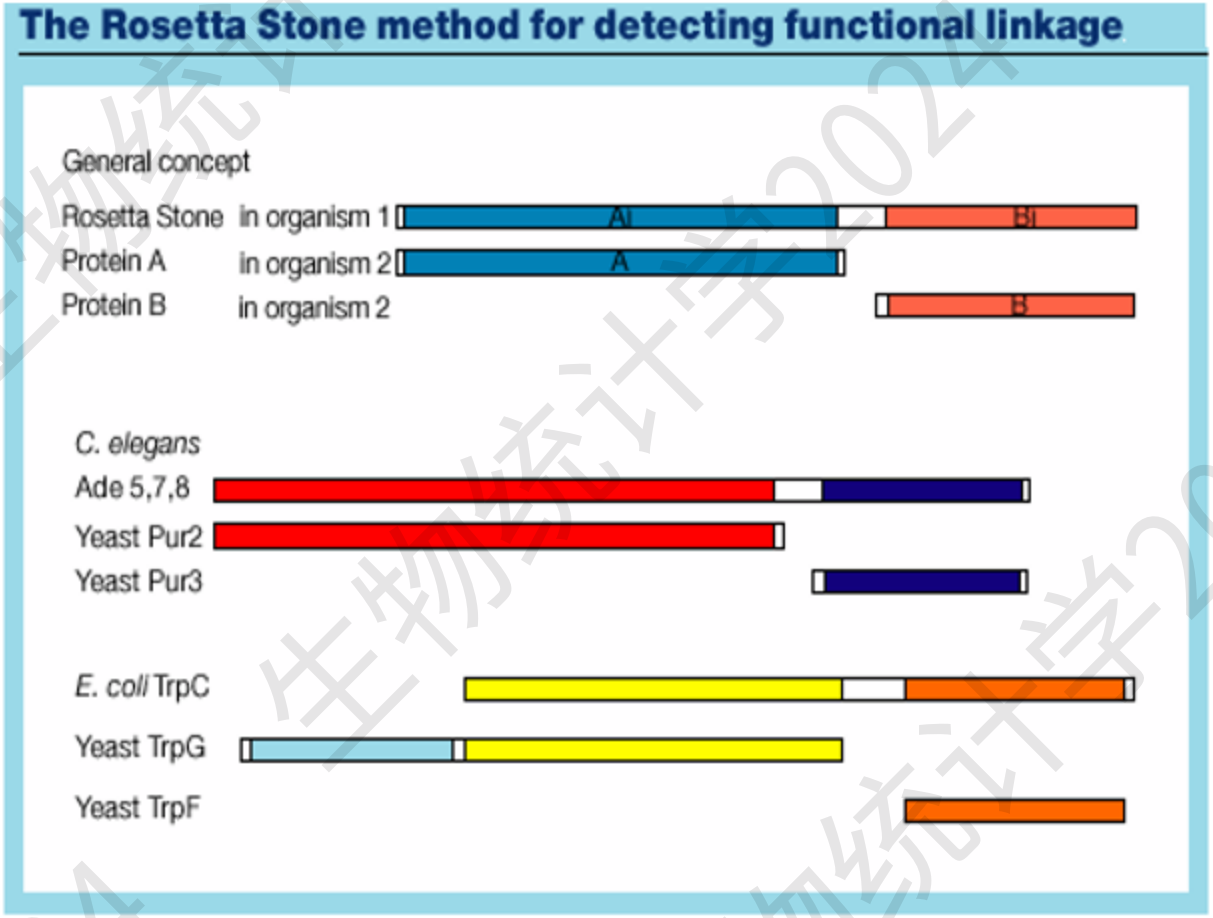
Protein-protein interactions (Computational Methods)

- Gene fusion method (A.Enright 1999. E.Maccote 1999)
- Phylogenetic profile method (M.Pellegrini 1999, D.Eisenberg, 1999).
- Gene cluster method (R.Overbeek, 1999).
- Highly co-expressed gene pairs.

Part II: Predicting Protein-protein Interactions

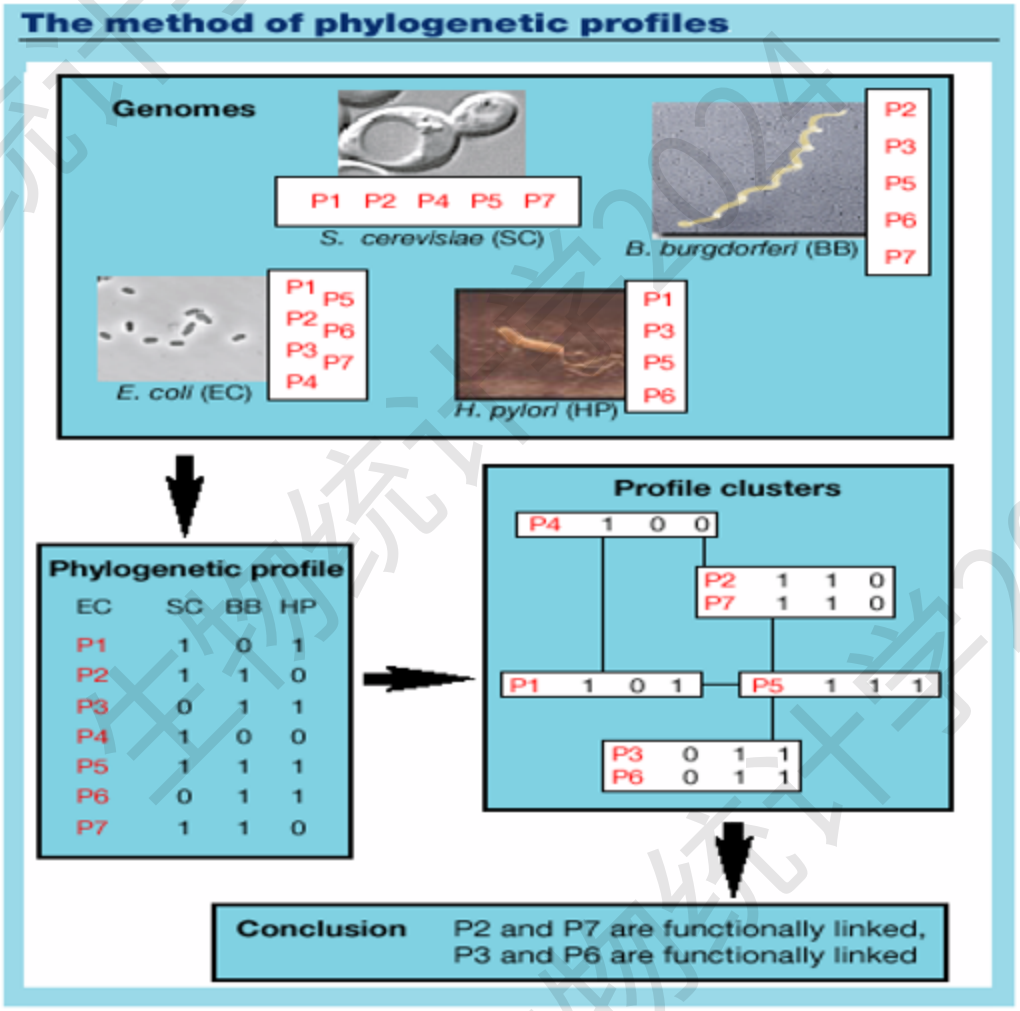
- Some computational methods
- Predicting protein-protein interaction from domains
  - Association method
  - MLE method

Rosetta Stone Method



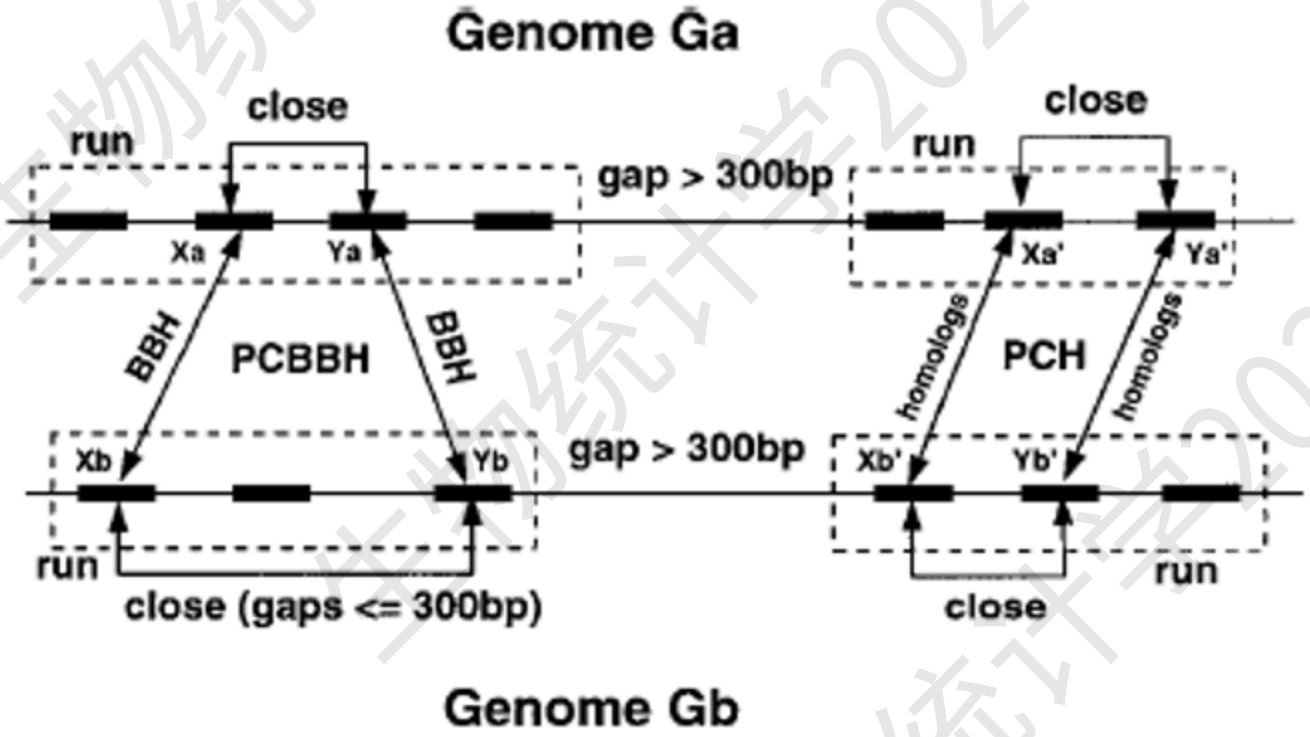
From: Nature Vol. 405, June.15, 2000, 823-826

Phylogenetic Profiles Method



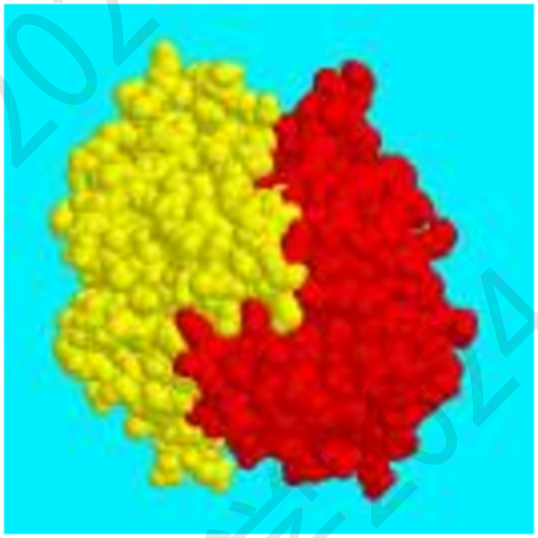
From: Nature Vol. 405, June, 15, 2000, 823-826

Using Gene Clusters to Infer Functional Coupling



From: R.Overbeek, PNAS 96, 2896-2901, 1999.

Structure of Proteins



Predicting PPIs from Domains

- Domains are treated as elementary unit of function.
- Domains are responsible for the generation of interactions.
- Understanding protein-protein interaction at the domain level.

Domain Databases

- Pfam, domain classification by HMM.
- Prodom.
- PRINTS, fingerprint information of protein sequences.
- SMART, mobile domain.
- BLOCKs, multiple alignment blocks.
- Interpro.

SwissPfam entry for PIG1_BOVIN

Description from Swissprot for PIG1_BOVIN :

1-phosphatidylinositol-4,5-bisphosphate phosphodiesterase gamma 1(ec 3.1.4.11) (plc-gamma-1) (phospholipase c-gamma-1) (plc-ii)(plc-148)

PH 33-142
PI-PLC-X 321-465
SH2 550-639
SH2 668-741
SH3 794-849
PH 864-931
PI-PLC-Y 952-1070
C2 1090-1177

Key

signal peptide: > pfamA: > Context: > smart: > transmembrane: > low complexity: > coiled coil: > pfamB:

Source	Domain	Start	End
Pfam	PH	33	142
Pfam	PI-PLC-X	321	465

Overlapping Domains: Change the domain order using the ^ and v buttons. View the changes by clicking the 'Change order' button.

high priority

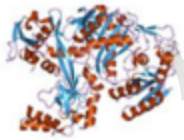
Piwi

Edit Wikipedia article

Piwi (or **PIWI**) genes were identified as regulatory proteins responsible for stem cell and germ cell differentiation.[4] Piwi is an abbreviation of P-element Induced WImpy testis in Drosophila.[5] Piwi proteins are highly conserved RNA-binding proteins and are present in both plants and animals.[6] Piwi proteins belong to the Argonaute/Piwi family and have been classified as nuclear proteins. Studies on Drosophila have also indicated that Piwi proteins have slicer activity conferred by the presence of the Piwi domain.[7] In addition, Piwi associates with Heterochromatin protein 1, an epigenetic modifier, and piRNA-complementary sequences. These are indications of the role Piwi plays in epigenetic regulation. Piwi proteins are also thought to control the biogenesis of piRNA as many Piwi-like proteins contain slicer activity which would allow Piwi proteins to process precursor piRNA into mature piRNA.

Contents [hide]

Piwi domain



Structure of the Pyrococcus furiosus Argonaute protein.[1]

Identifiers	
Symbol	Piwi
Pfam	PF02171
InterPro	IPR003165
PROSITE	PS50822
CDD	cd02826
Available protein structures:	[show]

Protein structure and function

The structure of several Piwi and Argonaute proteins (Ago) have been solved. Piwi proteins are RNA-binding proteins with 2 or 3 **domains**: The N-terminal **PAZ domain** binds the 3'-end of the guide RNA; the middle **MID domain** binds the 5'-phosphate of RNA; and the C-terminal **PIWI domain** acts as an RNase H endonuclease that can cleave RNA.[8][9] The small RNA partners of Ago proteins are microRNAs (miRNAs). Ago proteins utilize miRNAs to silence genes post-transcriptionally or use small-interfering RNAs (siRNAs) in both transcription and post-transcription silencing mechanisms. Piwi proteins interact with piRNAs (28–33 nucleotides) that are longer than miRNAs and siRNAs (~20 nucleotides), suggesting that their functions are distinct from those of Ago proteins.[8]

The piwi domain of an argonaute protein with bound siRNA, components of the RNA-induced silencing complex that mediates gene silencing by RNA interference.

Human Piwi proteins

Presently there are four known human Piwi proteins—PIWI-like protein 1, PIWI-like protein 2, PIWI-like protein 3 and PIWI-like protein 4. Human Piwi proteins all contain two RNA binding domains, PAZ and Piwi. The four PIWI-like proteins have a spacious binding site within the PAZ domain which allows them to bind the bulky 2′-OCH3 at the 3′ end of piwi-interacting RNA.[10]

One of the major human homologues, whose upregulation is implicated in the formation of tumours such as seminomas, is called hiwi (for human piwi).[11]

Homologous proteins in mice have been called miwi (for mouse piwi).[12]

Role in germline cells

PIWI proteins play a crucial role in fertility and germline development across animals and ciliates. Recently identified as a polar granule component, PIWI proteins appear to control germ cell formation so much so that in the absence of PIWI proteins there is a significant decrease in germ cell formation. Similar observations were made with the mouse homologs of PIWI, MILI, MIWI and MIWI2. These homologs are known to be present in spermatogenesis. Miwi is expressed in various stages of spermatocyte formation and spermatid elongation where Miwi2 is expressed in Sertoli cells. Mice deficient in either Mili or Miwi-2 have experienced spermatogenic stem cell arrest and those lacking Miwi-2 underwent a degradation of spermatogonia.[13] The effects of piwi proteins in human and mouse germlines seems to stem from their involvement in translation control as Piwi and the small noncoding RNA, piwi-interacting RNA (piRNA), have been known to co-fractionate polysomes. The piwi-piRNA pathway also induces heterochromatin formation at centromeres,[14] thus affecting transcription. The piwi-piRNA pathway also appears to protect the genome. First observed in Drosophila, mutant piwi-piRNA pathways led to a direct increase in dsDNA breaks in ovarian germ cells. The role of the piwi-piRNA pathway in transposon silencing may be responsible for the reduction in dsDNA breaks in germ cells.

Role in RNA interference

The piwi domain[15] is a protein domain found in piwi proteins and a large number of related nucleic acid-binding proteins, especially those that bind and cleave RNA. The function of the domain is double stranded-RNA-guided hydrolysis of single stranded-RNA that has been determined in the argonaute family of related proteins.[1] Argonautes, the most well-studied family of nucleic-acid binding proteins, are RNase H-like enzymes that carry out the catalytic functions of the RNA-induced silencing complex (RISC). In the well-known cellular process of RNA interference, the argonaute protein in the RISC complex can bind both small interfering RNA (siRNA) generated from exogenous double-stranded RNA and microRNA (miRNA)

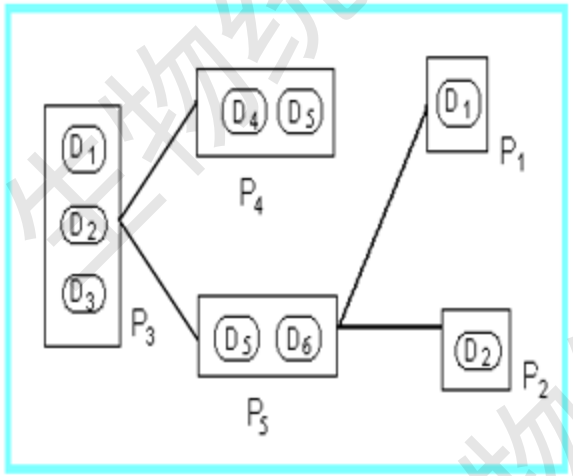
All human Piwi proteins and argonaute proteins have the same RNA binding domains, PAZ and Piwi.[2]

Association-A simple method

$$V(D_{ij}) = \frac{\#\{\text{Interacted protein pairs contain } D_{ij}\}}{\#\{\text{All protein pairs contain } D_{ij}\}}$$

More observed PPIs for one domain pair will give higher probability of interaction for that domain pair.

Simple Example



By assoiation method:

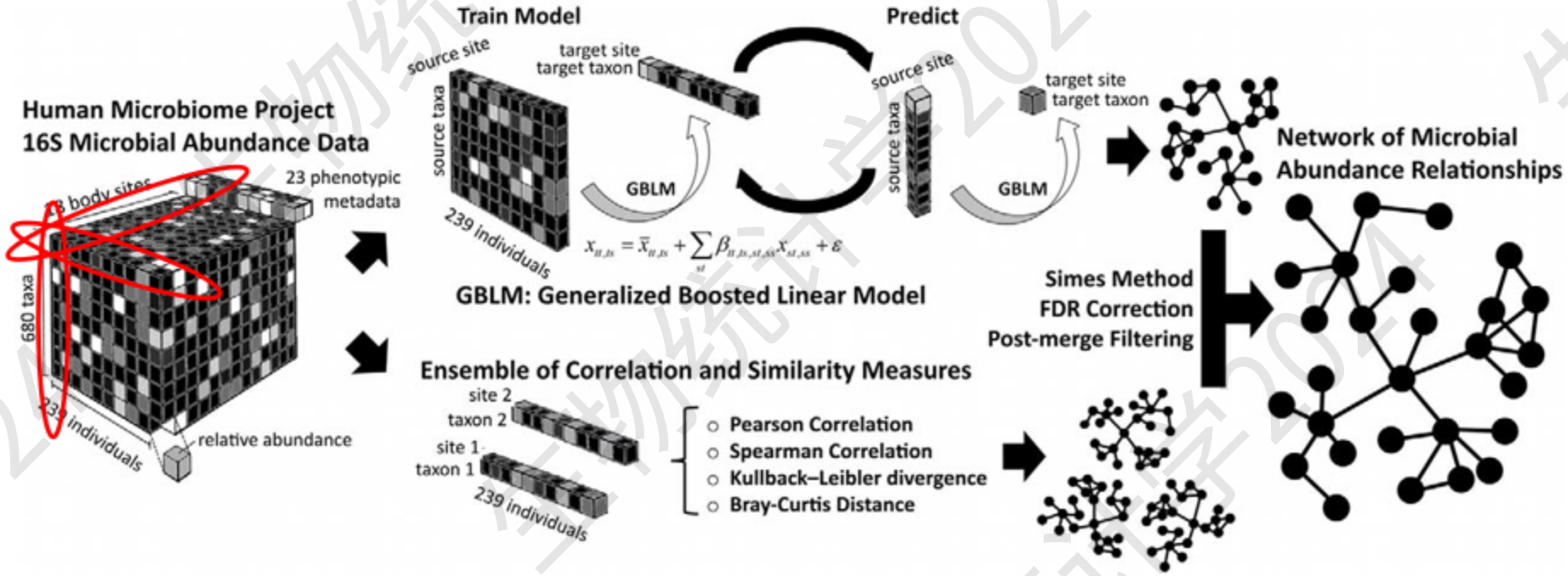
$$D_{34} = D_{35} = D_{36} = D_{26} = D_{16} = 1.0$$
$$D_{15} = D_{25} = 0.75, D_{14} = D_{24} = 0.5$$

Others are 0.0.

D_{15}: {P34, P35, P15}/{P_{34}, P_{35}, P_{15}, P_{14}}=0.75

More complicated example

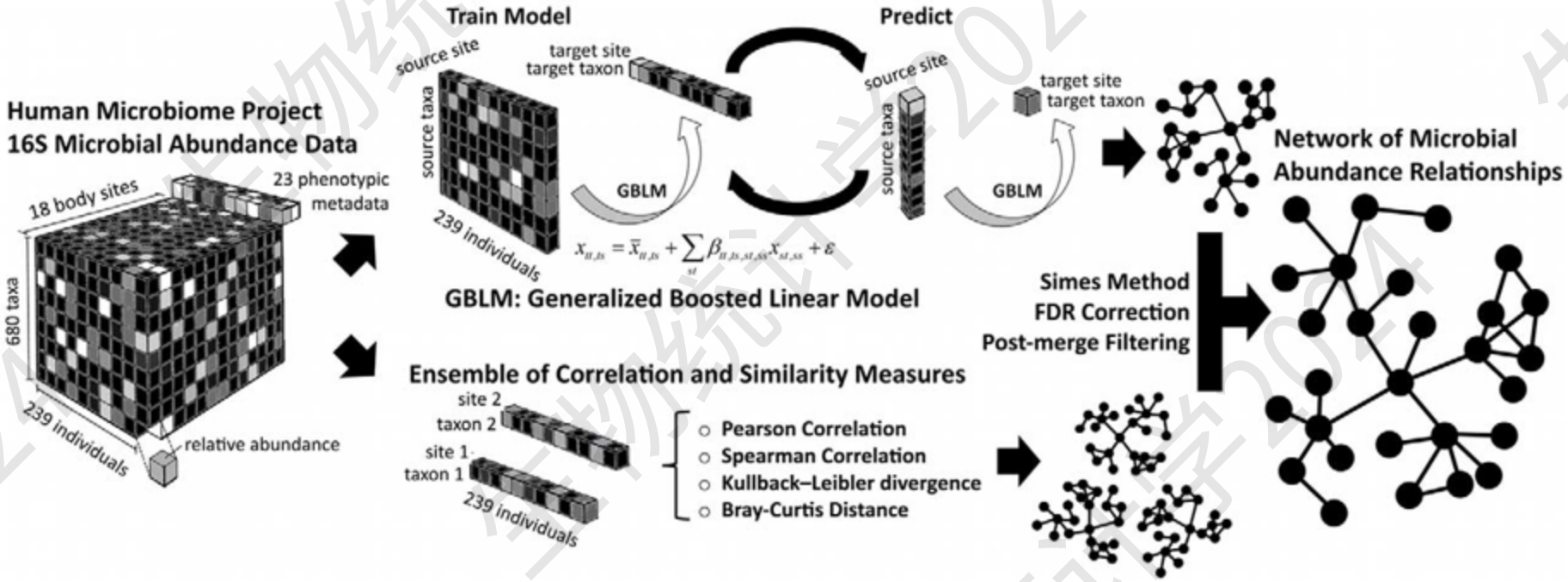
Generalized Boosted Linear Models (GBLM)：广义线性模型



Huttenhower, et al., *PLoS Computational Biology*, 201

More complicated example

Generalized Boosted Linear Models (GBLM)：广义线性模型



Ensemble scoring

Huttenhower, et al., *PLoS Computational Biology*, 201

More complicated example

Generalized Boosted Linear Models (GBLM)：广义线性模型

$$x_{tt,ts} = \bar{x}_{tt,ts} + \sum_{st} \beta_{tt,ts,st,ss} x_{st,ss}$$

$$\text{logit}(x_{tt,ts}) = \bar{x}_{tt,ts} + \sum_{st} \beta_{tt,ts,st,ss} x_{st,ss}$$

Huttenhower, et al., *PLoS Computational Biology*, 201

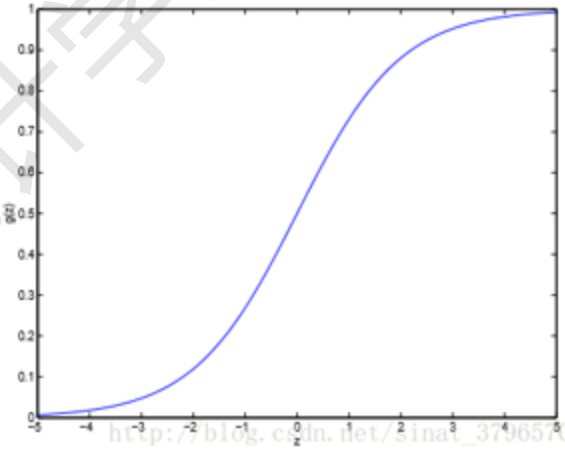
More complicated example

Generalized Linear Models (GLM)：广义线性模型

广义线性模型的核心体现在：

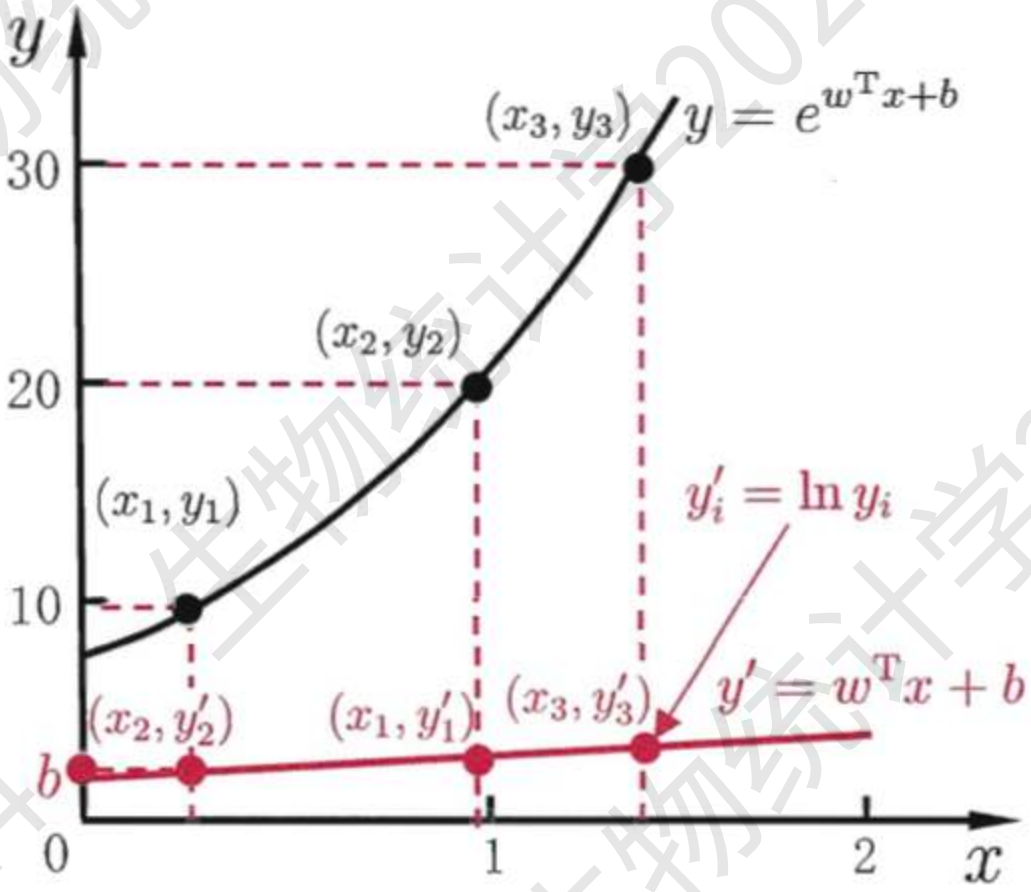
- y服从指数族分布(包括高斯分布，伯努利分布，多项式分布，泊松分布，beta分布……)，且同个样本的y必须服从同个分布

- 接着在具体分布中比较与指数分布族之间的参数关系，最重要的就是具体分布的参数(Φ)和指数分布参数(η)之间的关系



More complicated example

Generalized Linear Models (GLM)：广义线性模型



Limitation of Association Method

- For multiple-domain proteins, this method computes the value for a certain domain pair ignoring the value of other domain-domain pairs. So it's a local one.
- This method cannot deal with possible error of the data.

Probabilistic Model

- Domain-domain interactions are independent, which means that the event that two domains interact or not does not depend on other domains.
- Two proteins interact if and only if at least one pair of domains from the two proteins interact.

Yeast Data

- Interactions (Uetz's and Ito's interaction data).
- Domain: Pfam (Pfam-A, Pfam-B).
- Proteins: SGD, N=6359.

Protein Interaction Data Sources

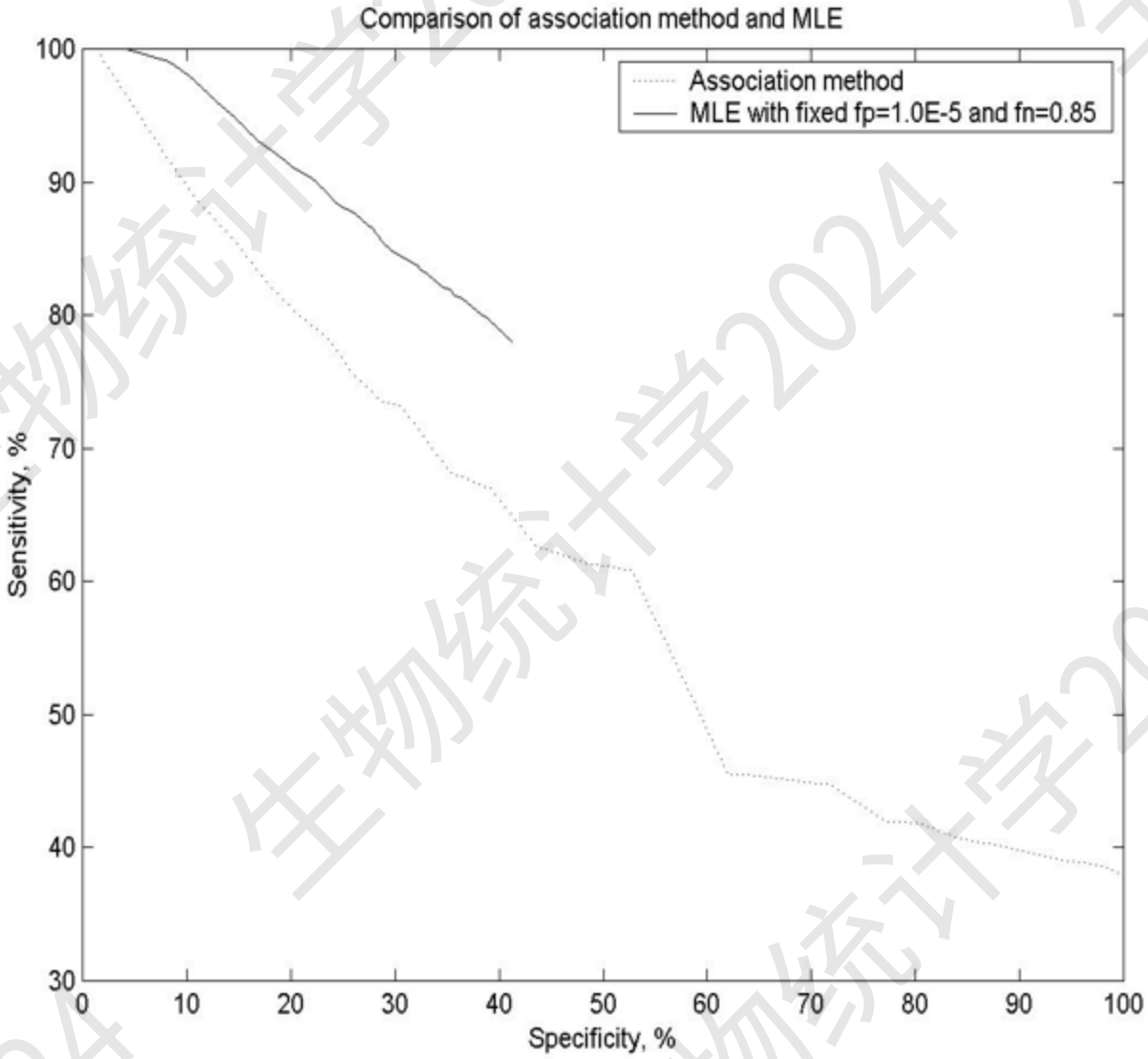
	Proteins	Pfam domains	Super - domains	PPI
Uetz	1337	1330	313	1445
Ito	3277	2776	909	4475
Uetz+Ito	3729	3124	1007	5719
Overlap	855	964	215	201

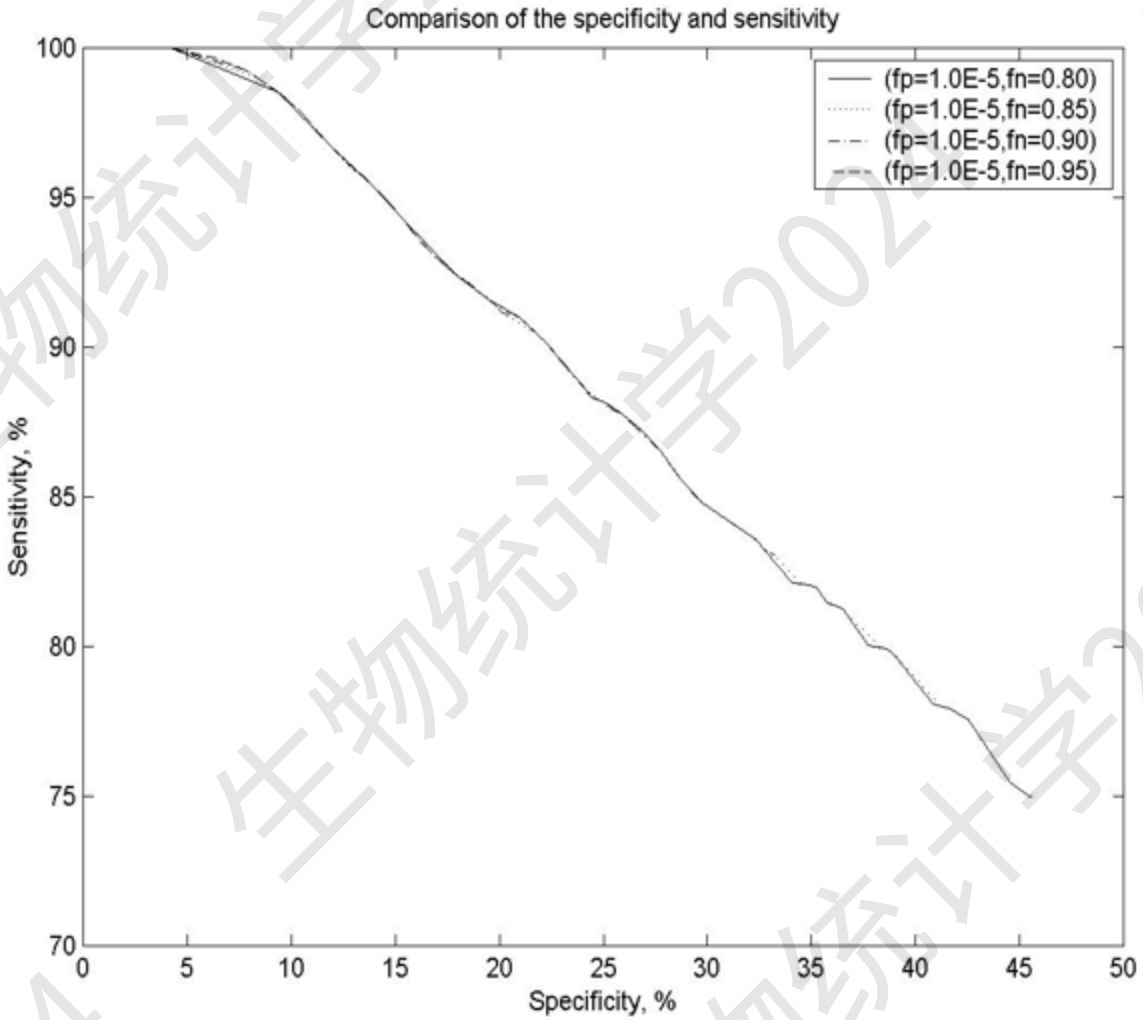
Measure the Accuracy

- Specificity and sensitivity.
- Verification by MIPS physical interactions (as TRUE interactions).
- Relationship between protein-protein interactions and expression data.

$$SP = \frac{\text{number of matches with observation}}{\text{number of prediction}}$$

$$SN = \frac{\text{number of match with observation}}{\text{number of observation}}$$


Comparison of association method and MLE
Association method
MLE with fixed fp=1.0E-5 and fn=0.85
Sensitivity, %
Specificity, %
100
90
80
70
60
50
40
30
0
10
20
30
40
50
60
70
80
90
100


Comparison of the specificity and sensitivity
(fp=1.0E-5,fn=0.80)
(fp=1.0E-5,fn=0.85)
(fp=1.0E-5,fn=0.90)
(fp=1.0E-5,fn=0.95)
Sensitivity, %
Specificity, %
100
95
90
85
80
75
70
0
5
10
15
20
25
30
35
40
45
50

Verification by Known PPIs

- MIPS physical interaction. (Totally 2570 PPIs, 1414 PPIs not overlapping with our training set).
- Compare with random matching.
  - Fold number
  - Larger fold number imply more reliable prediction

$$\#\text{Fold} = \frac{\#\{\text{Our prediction matched to MIPS}\}}{\#\{\text{Expectation of random pairs matched to MIPS}\}}$$

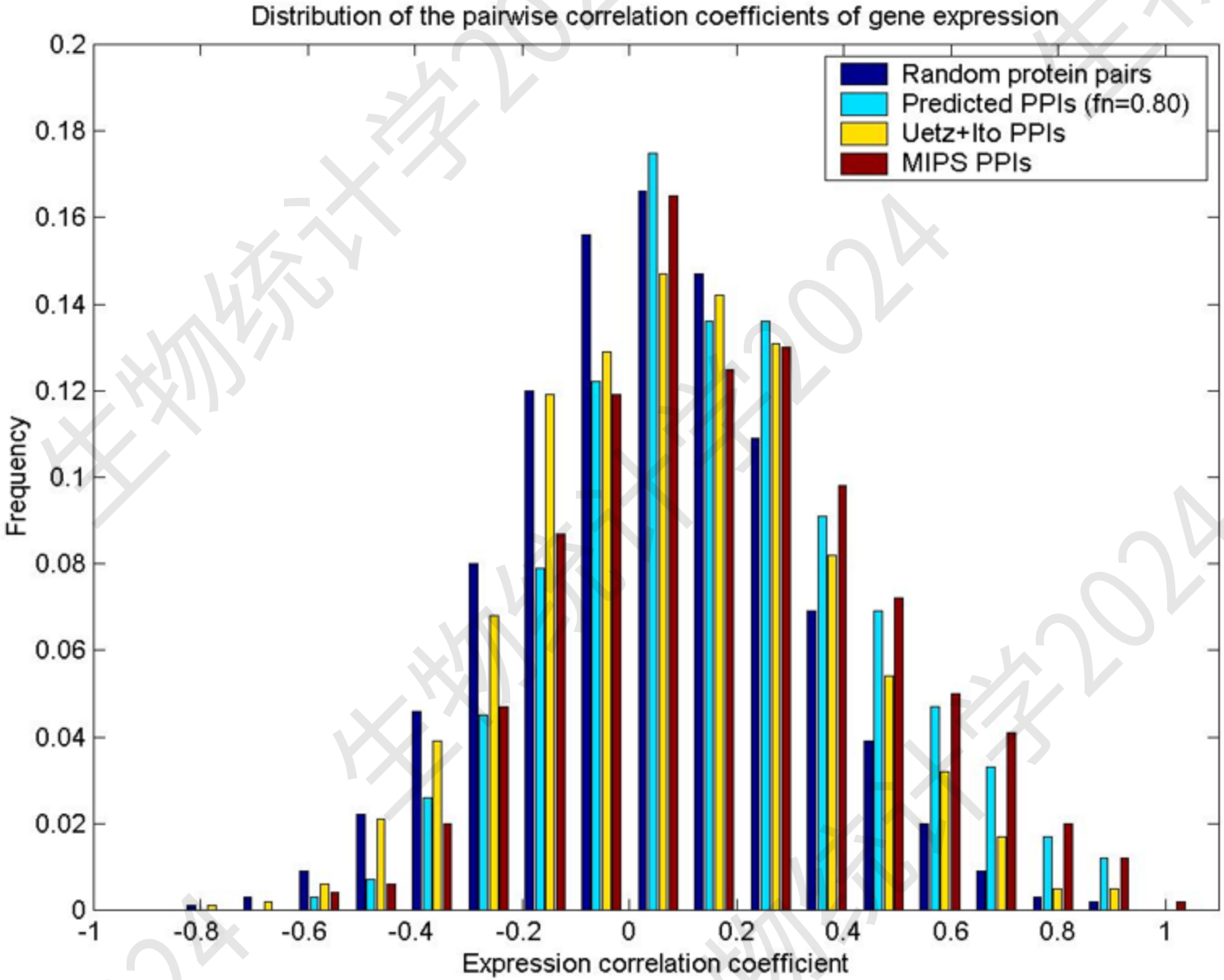
Think about FDR again...

Matching with MIPS PPIs

Prob	#Predict	#Train	#MIPS		#Fold
All	20221620	5719	2570	1414	1.00
>0.00	136463	5719	1265	109	11.92
>=0.20	26908	5238	1093	53	34.97
>=0.40	19360	5018	1035	48	47.85
>=0.60	14725	4775	971	47	67.53
>=0.80	12734	4647	932	43	76.02
>=0.975	10824	4461	886	40	89.88

Interaction Data Correlated With Gene Expression Data

- Interacted proteins seems to have high expression correlation
  - A.Grigorieve *Nucleic Acid Res.* 29, 2001;
  - H. Ge et al. *Nature Genetics* 29, 2001;
  - R. Jansen et al. *Genome Res.*12, 2002.
- Expression data (M.Eisen, 1998); 2465 Yeast ORFs with 79 data points/ORF.
- Pearson correlation coefficient.


Distribution of the pairwise correlation coefficients of gene expression
Random protein pairs
Predicted PPIs (fn=0.80)
Uetz+Ito PPIs
MIPS PPIs
Frequency
0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1
Expression correlation coefficient

Statistics of Pairwise Correlation of Gene Expression

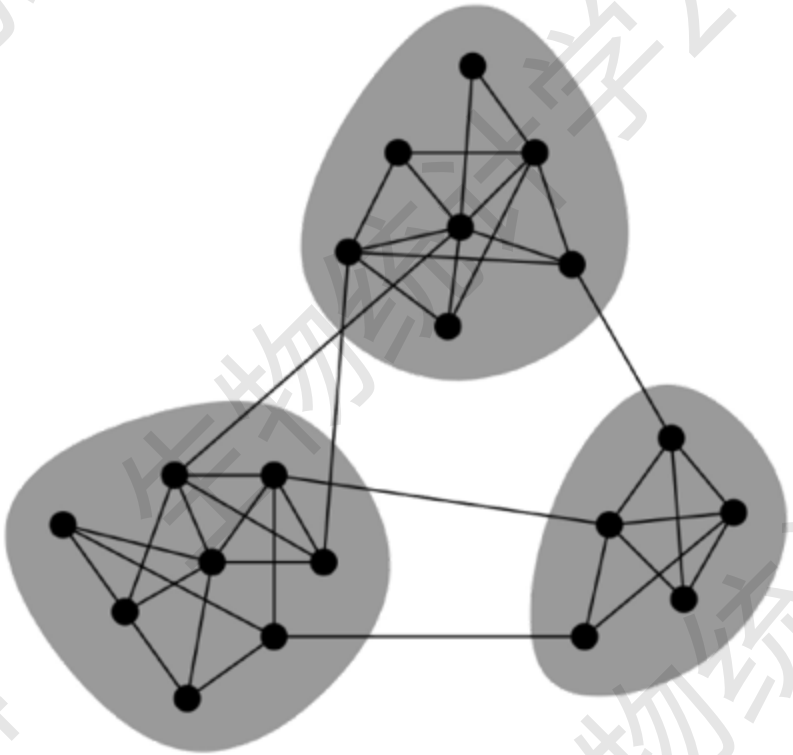
pairs	# pairs	mean	std	T-score	p-value	$R* > 0.5$
All ORFs	3036880	0.0428	0.2473	0.0000	5.000e-01	3.84%
$\geq$0.20	6392	0.0514	0.2550	2.7984	2.575e-03	4.79%
$\geq$0.40	4433	0.0510	0.2538	2.2232	1.311e-02	4.96%
$\geq$0.60	3318	0.0598	0.2579	3.9644	3.715e-05	5.42%
$\geq$0.80	2756	0.0626	0.2622	4.2196	1.238e-05	5.88%
$\geq$0.975	2266	0.0628	0.2637	3.8482	6.002e-05	5.87%
Uetz+Ito	1307	0.0586	0.2587	2.3213	1.015e-02	5.20%
MIPS	1106	0.1109	0.2767	9.1619	2.706e-20	8.23%

$$T = \frac{\mu_1 - \mu_2}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}}$$

第8-3章：Network Module

- Definition
- Module detection
- Bayesian approach
- Markov clustering algorithm

Network Modular



Modularity

- Suppose we are given a candidate division of the vertices into some number of groups. The modularity of this division is defined to be the fraction of the edges that fall within the given groups minus the expected such fraction if edges were distributed at random.

http://en.wikipedia.org/wiki/Modularity_(networks)

Modularity

- A_{ij}: adjacency matrix
- k_i: degree
- m: total number of edges

$$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - \frac{k_i k_j}{2m}) \delta(c_i, c_j)$$

Modularity

- For two class problem, let s_i=1 of node I belongs to group1 and s_i=-1 if it belongs to group 2,

$$\delta(c_i, c_j) = \frac{1}{2}(s_i s_j + 1)$$

$$Q = \frac{1}{4m}\sum_{ij} S^T B S$$

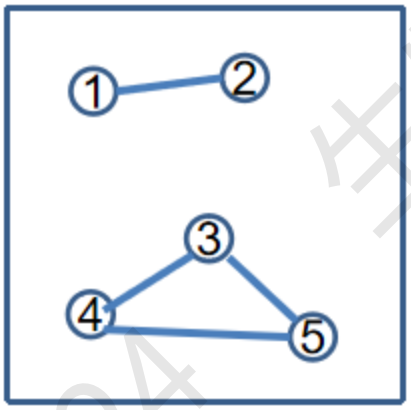
$$B = (B_{ij}), B_{ij} = A_{ij} - \frac{k_i k_j}{2m}$$

$$S = (s_1, \cdots, s_n)^T$$

Example

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{pmatrix}$$

$$m = 4, k_1 = k_2 = 1$$
$$k_3 = k_4 = k_5 = 2$$



$$B = \frac{1}{8} \begin{pmatrix} -1 & 7 & -2 & -2 & -2 \\ 7 & -1 & -2 & -2 & -2 \\ -2 & -2 & -4 & 4 & 4 \\ -2 & -2 & 4 & -4 & 4 \\ -2 & -2 & 4 & 4 & -4 \end{pmatrix}$$

Spectrum Method

- The largest eigenvectors will gives the best grouping, positive entries corresponding to one class, and negative ones corresponding to another class.

- This can be achieved by power method

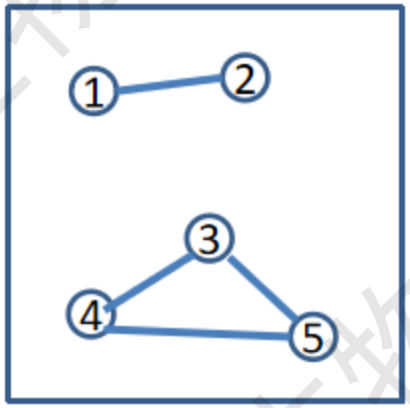
where e=(1,...,1)', $\lim_{k \to +\infty} \frac{A^k e}{e^T A^k e} = w$

Example

- 对于上述矩阵B, 可以计算出最大特征值为10, 对应的特征向量

$$v = (-0.55, -0.55, 0.37, 0.37, 0.37)$$

- 于是我们对节点的划分为{1,2}; {3,4,5}



优化方法

- 既然现在有一个衡量划分“好坏”的量Q, 那么一般的优化方法都可以使用;
  - 1. 给定初始划分
  - 2. 对于划分的某种修正, 计算Q的改变量
  - 3. 依据一定的原则考虑是否接受这种修正，重复步骤2, 直到某种收敛条件满足。
- Greedy方法
- 模拟退火方法

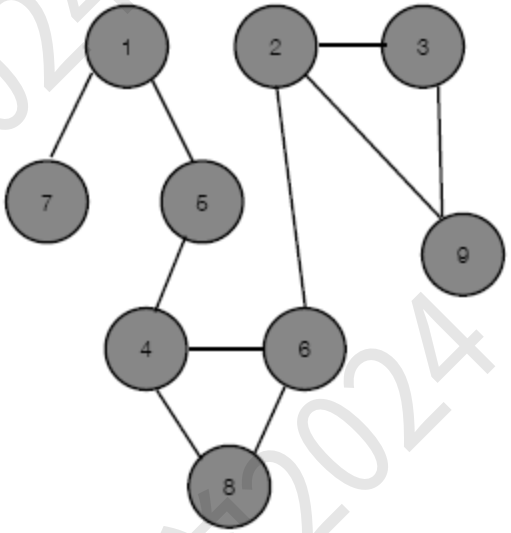
A Bayesian Approach to Network Modularity

Slides for this part are mainly from Hofman's talk

*www.jakehofman.com/talks/**apam_20071019**.pdf*

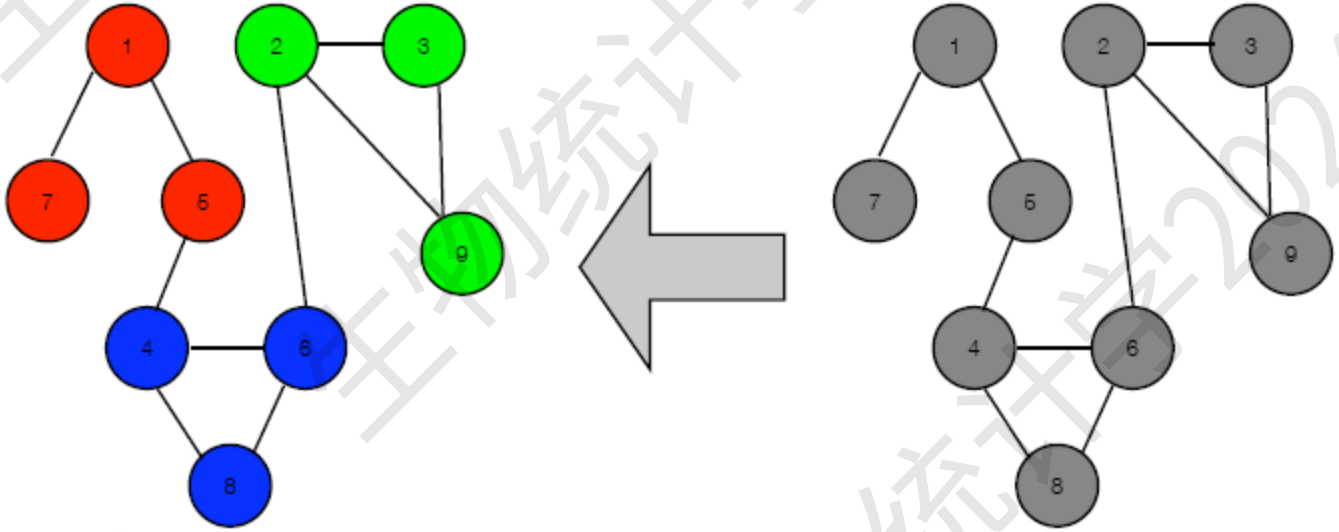
Overview: Modular Networks

- Given a network
  - Assign nodes to modules?
  - Determine number of modules(scale/complexity)?

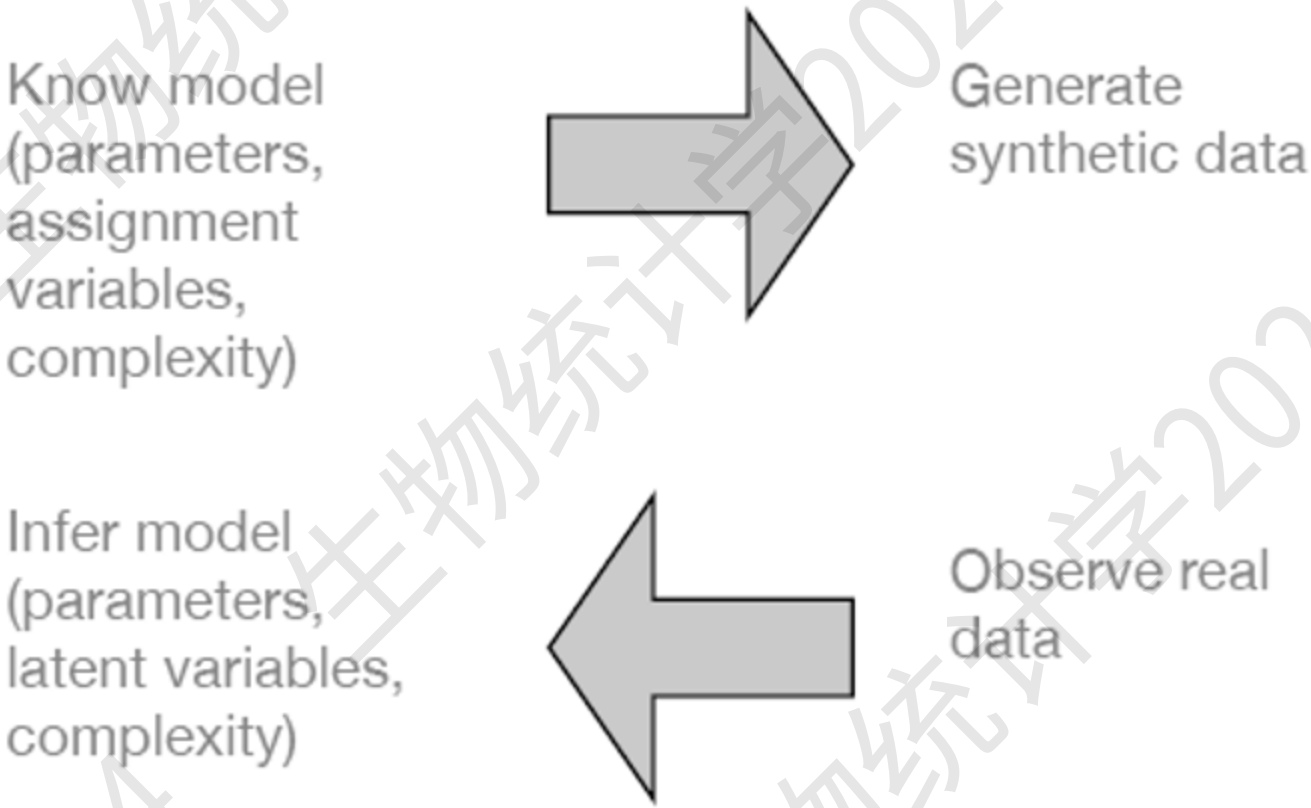


Overview: Modular Networks

- With a generative model of modular networks, rules of probability tell us how to calculate model parameters (e.g. number of modules & ass



Generative Models



Markov Clustering Algorithm

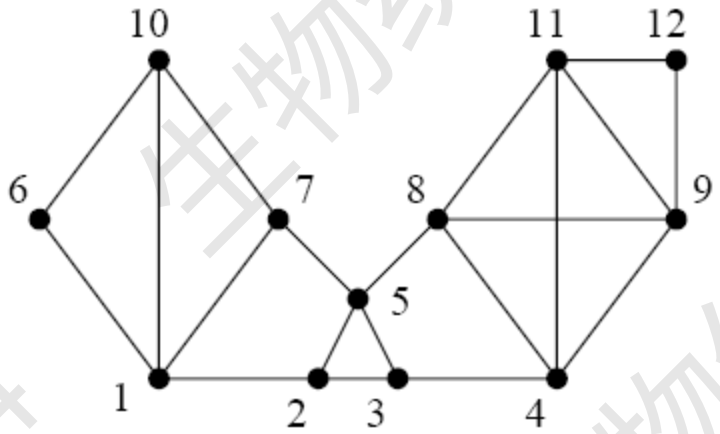
van Dongen. A cluster algorithm for graphs. Information Systems, 2000

K-length Path

- Basic idea: dense regions in sparse graphs corresponding with regions in which the number of k-length path is relatively large.

- Random walks can also be used to detect clusters in graphs, the idea is that the more closed is a subgraph, the largest the time a random walker need to escape from it.

K-path Clustering



$$\begin{pmatrix}
5 & 2 & 1 & 0 & 2 & \mathbf{3} & \mathbf{3} & 0 & 0 & \mathbf{4} & 0 & 0 \\
2 & 4 & \mathbf{3} & 1 & \mathbf{3} & 1 & 2 & 1 & 0 & 1 & 0 & 0 \\
1 & \mathbf{3} & 4 & 2 & \mathbf{3} & 0 & 1 & 2 & 1 & 0 & 1 & 0 \\
0 & 1 & 2 & 5 & 2 & 0 & 0 & \mathbf{4} & \mathbf{4} & 0 & \mathbf{4} & 2 \\
2 & \mathbf{3} & \mathbf{3} & 2 & 5 & 0 & 2 & 2 & 1 & 1 & 1 & 0 \\
\mathbf{3} & 1 & 0 & 0 & 0 & 3 & 2 & 0 & 0 & \mathbf{3} & 0 & 0 \\
\mathbf{3} & 2 & 1 & 0 & 2 & 2 & 4 & 1 & 0 & \mathbf{3} & 0 & 0 \\
0 & 1 & 2 & \mathbf{4} & 2 & 0 & 1 & 5 & \mathbf{4} & 0 & \mathbf{4} & 2 \\
0 & 0 & 1 & \mathbf{4} & 1 & 0 & 0 & \mathbf{4} & 5 & 0 & \mathbf{5} & \mathbf{3} \\
\mathbf{4} & 1 & 0 & 0 & 1 & \mathbf{3} & \mathbf{3} & 0 & 0 & 4 & 0 & 0 \\
0 & 0 & 1 & \mathbf{4} & 1 & 0 & 0 & \mathbf{4} & \mathbf{5} & 0 & 5 & \mathbf{3} \\
0 & 0 & 0 & 2 & 0 & 0 & 0 & 2 & \mathbf{3} & 0 & \mathbf{3} & 3
\end{pmatrix}$$

Matrix manipulation: $(N+I)^2$

Markov Clustering

- Expansion: Through matrix manipulation (power), one obtains a matrix for a n-steps connection.

- Inflation: Enhance intercluster passages by raising the elements to a certain power and then normalize

Markov Clustering Algorithm

- Iteratively running two operators
  - Inflation:

$$(T_r M)_{ij} = \frac{M_{ij}^r}{\sum_i M_{ij}^r}$$

Column normalization

  - Expansion:

$$\text{Expand}(M) = M^k$$

MCL Running

$$\begin{pmatrix}
0.380 & 0.087 & 0.027 & -- & 0.077 & 0.295 & 0.201 & -- & -- & 0.320 & -- & -- \\
0.047 & 0.347 & 0.210 & 0.017 & 0.150 & 0.019 & 0.066 & 0.012 & -- & 0.012 & -- & -- \\
0.014 & 0.210 & 0.347 & 0.056 & 0.150 & -- & 0.016 & 0.046 & 0.009 & -- & 0.009 & -- \\
-- & 0.027 & 0.087 & 0.302 & 0.062 & -- & -- & 0.184 & 0.143 & -- & 0.143 & 0.083 \\
0.058 & 0.210 & 0.210 & 0.056 & 0.406 & -- & 0.083 & 0.046 & 0.009 & 0.019 & 0.009 & -- \\
0.142 & 0.017 & -- & -- & -- & 0.295 & 0.083 & -- & -- & 0.184 & -- & -- \\
0.113 & 0.069 & 0.017 & -- & 0.062 & 0.097 & 0.333 & 0.012 & -- & 0.147 & -- & -- \\
-- & 0.017 & 0.069 & 0.175 & 0.049 & -- & 0.016 & 0.287 & 0.143 & -- & 0.143 & 0.083 \\
-- & -- & 0.017 & 0.175 & 0.012 & -- & -- & 0.184 & 0.288 & -- & 0.288 & 0.278 \\
0.246 & 0.017 & -- & -- & 0.019 & 0.295 & 0.201 & -- & -- & 0.320 & -- & -- \\
-- & -- & 0.017 & 0.175 & 0.012 & -- & -- & 0.184 & 0.288 & -- & 0.288 & 0.278 \\
-- & -- & -- & 0.044 & -- & -- & -- & 0.046 & 0.120 & -- & 0.120 & 0.278
\end{pmatrix}$$

$\Gamma_2 M^2$, M defined in Figure 8

MCL Running

$$
\begin{pmatrix}
0.448 & 0.080 & 0.023 & -- & 0.068 & 0.426 & 0.359 & -- & -- & 0.432 & -- & -- \\
0.018 & 0.285 & 0.228 & 0.007 & 0.176 & 0.006 & 0.033 & 0.005 & -- & 0.007 & -- & -- \\
0.005 & 0.223 & 0.290 & 0.022 & 0.173 & -- & 0.010 & 0.017 & 0.003 & 0.001 & 0.003 & 0.001 \\
-- & 0.018 & 0.059 & 0.222 & 0.040 & -- & 0.001 & 0.187 & 0.139 & -- & 0.139 & 0.099 \\
0.027 & 0.312 & 0.314 & 0.028 & 0.439 & 0.005 & 0.054 & 0.022 & 0.003 & 0.010 & 0.003 & 0.001 \\
0.116 & 0.007 & 0.001 & -- & 0.004 & 0.157 & 0.085 & -- & -- & 0.131 & -- & -- \\
0.096 & 0.040 & 0.013 & -- & 0.037 & 0.083 & 0.197 & 0.001 & -- & 0.104 & -- & -- \\
-- & 0.012 & 0.042 & 0.172 & 0.029 & -- & 0.002 & 0.198 & 0.133 & -- & 0.133 & 0.096 \\
-- & 0.001 & 0.015 & 0.256 & 0.009 & -- & -- & 0.266 & 0.326 & -- & 0.326 & 0.346 \\
0.290 & 0.021 & 0.002 & -- & 0.017 & 0.323 & 0.260 & -- & -- & 0.316 & -- & -- \\
-- & 0.001 & 0.015 & 0.256 & 0.009 & -- & -- & 0.266 & 0.326 & -- & 0.326 & 0.346 \\
-- & -- & 0.001 & 0.037 & 0.001 & -- & -- & 0.039 & 0.069 & -- & 0.069 & 0.112
\end{pmatrix}
$$

$$\Gamma_2(\Gamma_2 M^2 \cdot \Gamma_2 M^2)$$

MCL Running

$$\begin{pmatrix}
0.807 & 0.040 & 0.015 & -- & 0.034 & 0.807 & 0.807 & -- & -- & 0.807 & -- & -- \\
-- & 0.090 & 0.092 & -- & 0.088 & -- & -- & -- & -- & -- & -- & -- \\
-- & 0.085 & 0.088 & -- & 0.084 & -- & -- & -- & -- & -- & -- & -- \\
-- & 0.001 & 0.001 & 0.032 & 0.001 & -- & -- & 0.032 & 0.031 & -- & 0.031 & 0.031 \\
-- & 0.777 & 0.798 & -- & 0.786 & -- & 0.001 & -- & -- & -- & -- & -- \\
0.005 & -- & -- & -- & -- & 0.005 & 0.005 & -- & -- & 0.005 & -- & -- \\
0.003 & 0.001 & -- & -- & 0.001 & 0.003 & 0.003 & -- & -- & 0.003 & -- & -- \\
-- & -- & 0.001 & 0.024 & -- & -- & -- & 0.024 & 0.024 & -- & 0.024 & 0.024 \\
-- & -- & 0.002 & 0.472 & 0.001 & -- & -- & 0.472 & 0.472 & -- & 0.472 & 0.472 \\
0.185 & 0.005 & 0.001 & -- & 0.004 & 0.185 & 0.184 & -- & -- & 0.185 & -- & -- \\
-- & -- & 0.002 & 0.472 & 0.001 & -- & -- & 0.472 & 0.472 & -- & 0.472 & 0.472 \\
-- & -- & -- & 0.001 & -- & -- & -- & 0.001 & 0.001 & -- & 0.001 & --
\end{pmatrix}$$

$(\Gamma_2 \circ Squaring)$ iterated four times on M

MCL Running

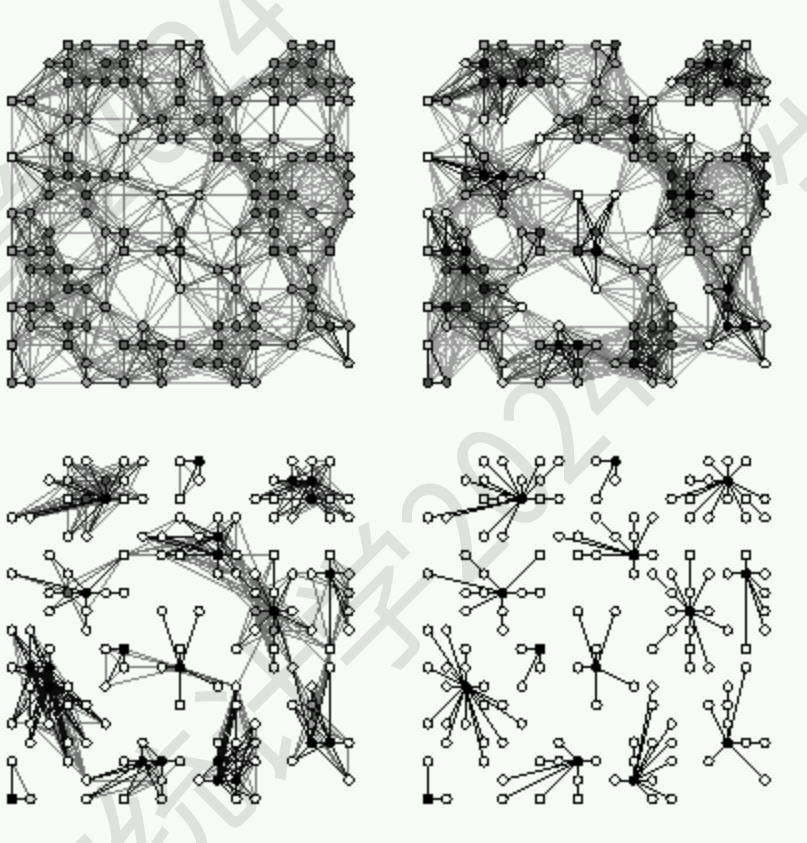
$$
\begin{pmatrix}
1.000 & -- & -- & -- & -- & 1.000 & 1.000 & -- & -- & 1.000 & -- & -- \\
-- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- \\
-- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- \\
-- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- \\
-- & 1.000 & 1.000 & -- & 1.000 & -- & -- & -- & -- & -- & -- & -- \\
-- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- \\
-- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- \\
-- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- \\
-- & -- & -- & 0.500 & -- & -- & -- & 0.500 & 0.500 & -- & 0.500 & 0.500 \\
-- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- \\
-- & -- & -- & 0.500 & -- & -- & -- & 0.500 & 0.500 & -- & 0.500 & 0.500 \\
-- & -- & -- & -- & -- & -- & -- & -- & -- & -- & -- & --
\end{pmatrix}
$$

M_{mcl}^{∞}

A Heuristic for MCL

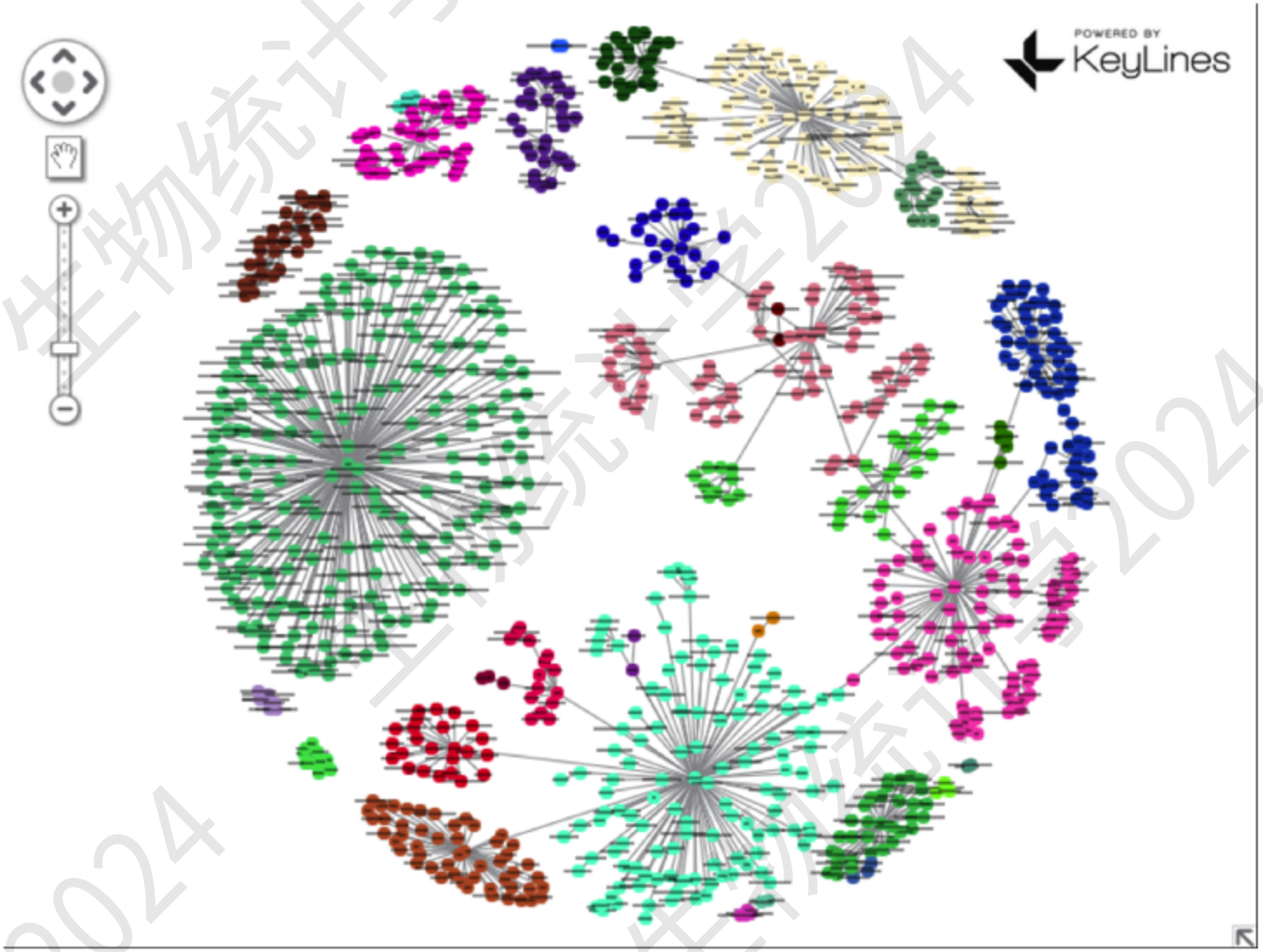
We take a random walk on the graph described by the similarity matrix

After each step we weaken the links between distant nodes and strengthen the links between nearby nodes

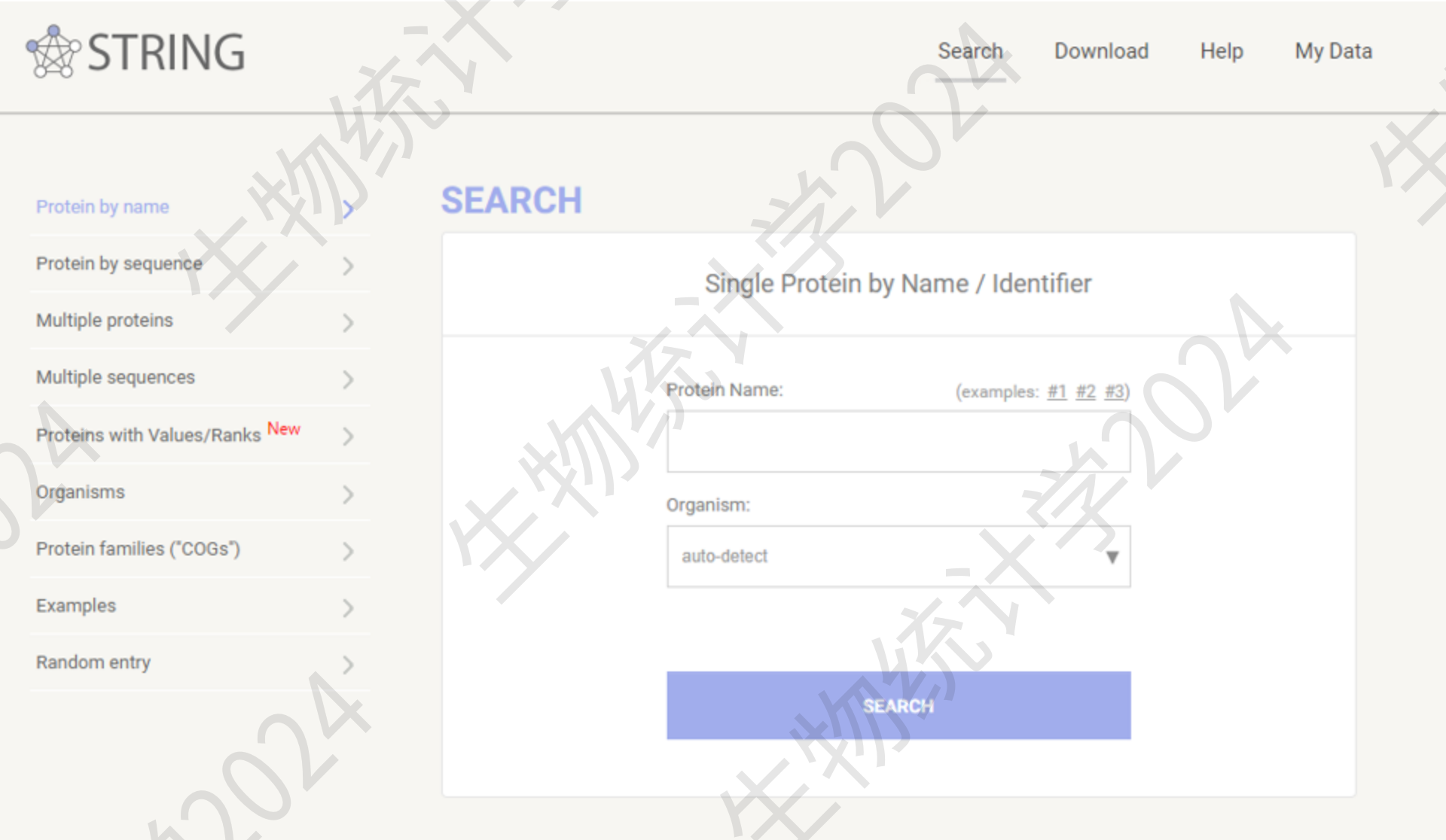


Graphic from van Dongen, 2000

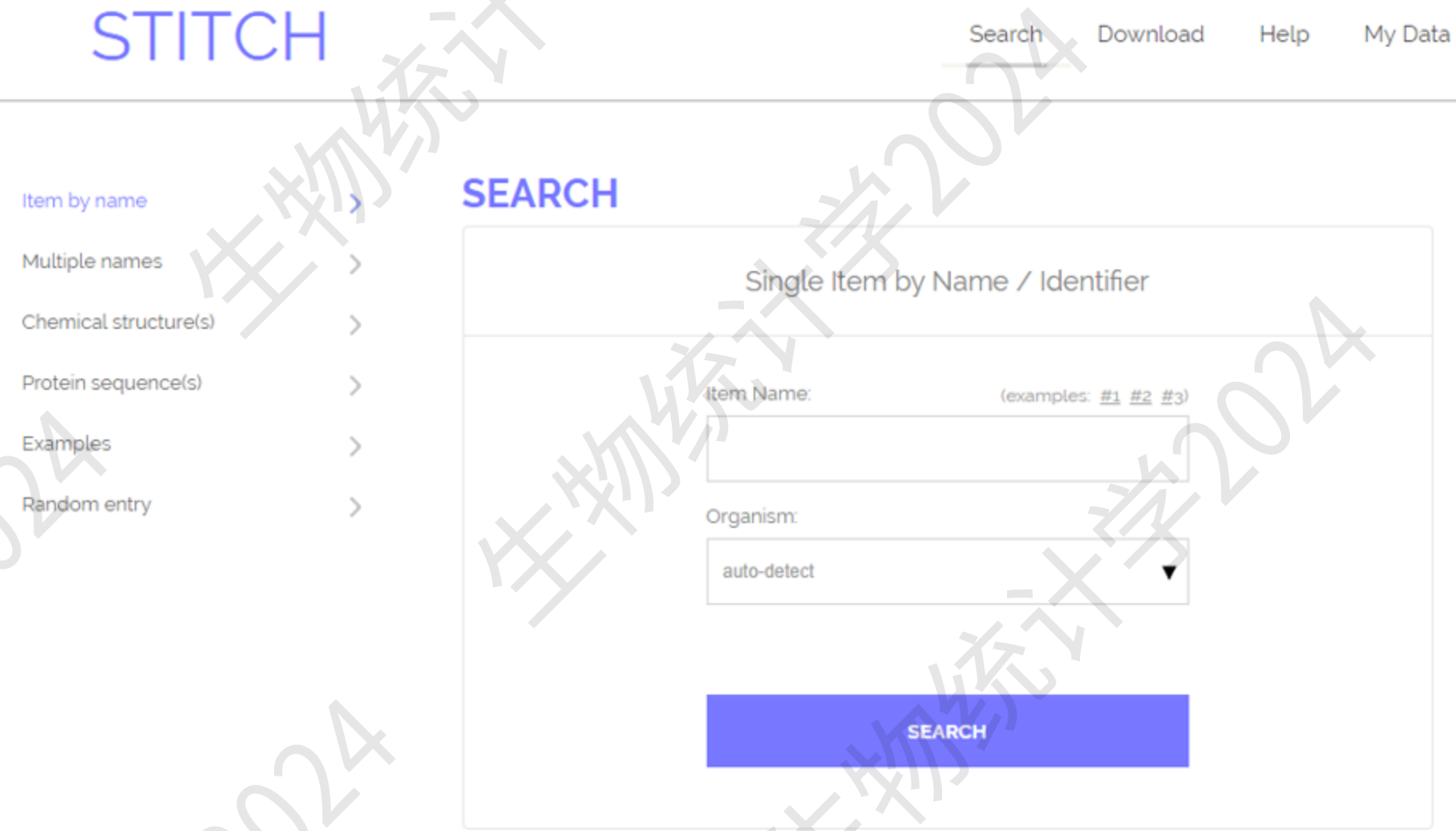
Clustering examples



STRING



STITCH



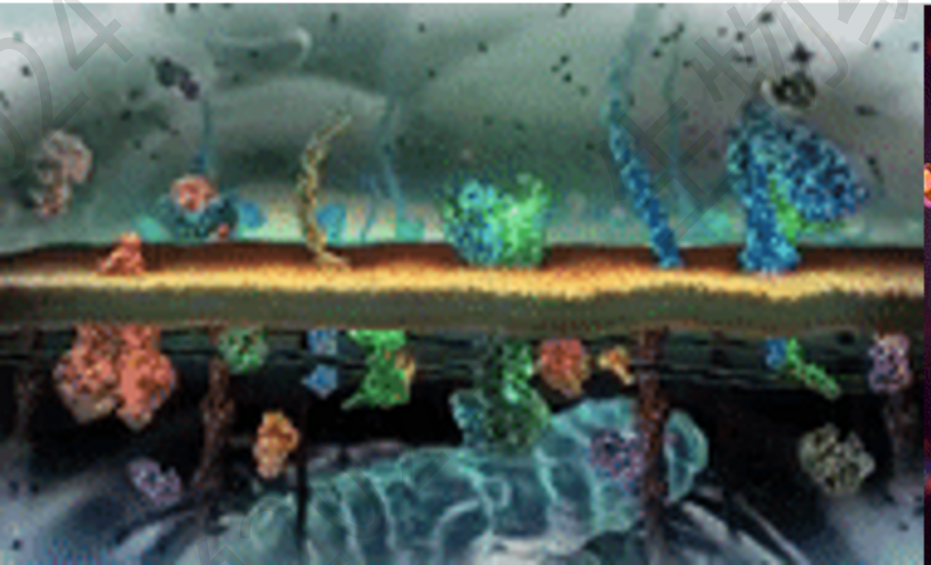
蛋白质3D结构

- 蛋白质3D结构预测
- AlphaFold2（利用AI攻克蛋白质3D结构预测问题）

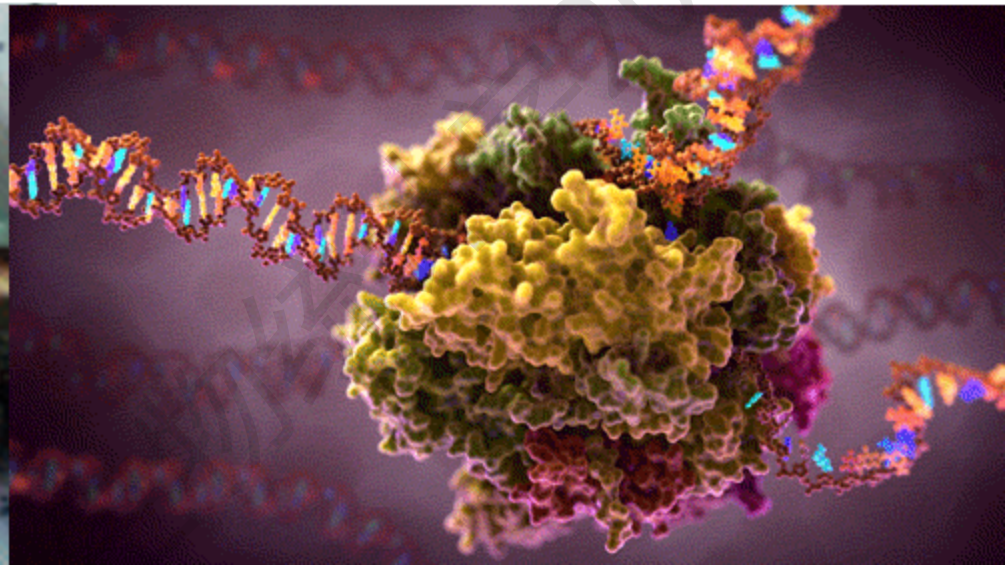
Protein 3D structure

眼见为实

Overview



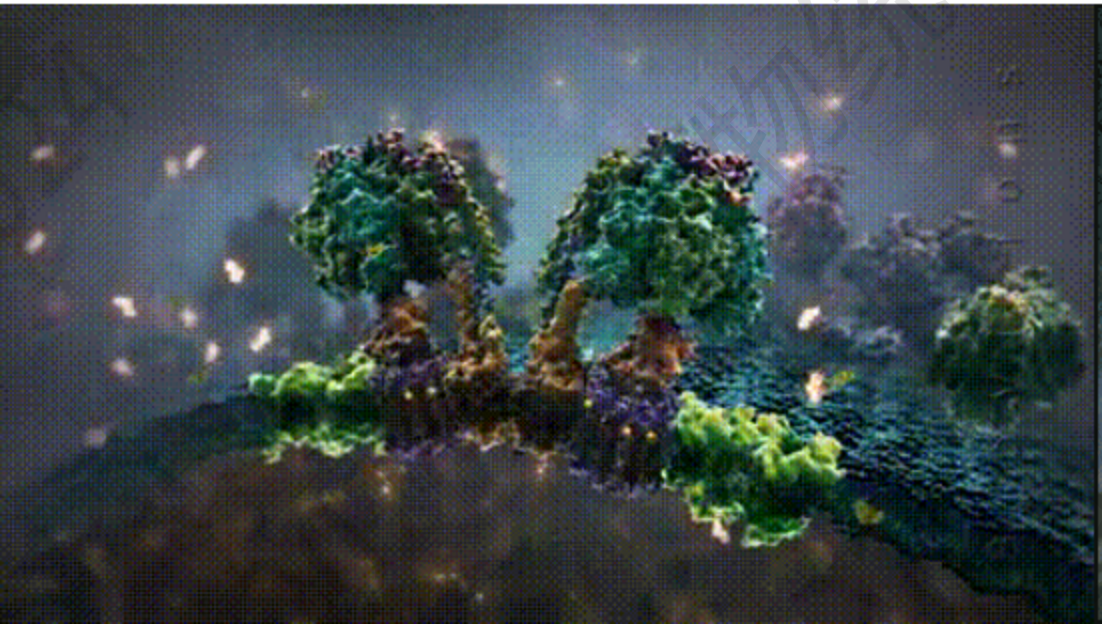
Ribosome



Protein 3D structure

眼见为实

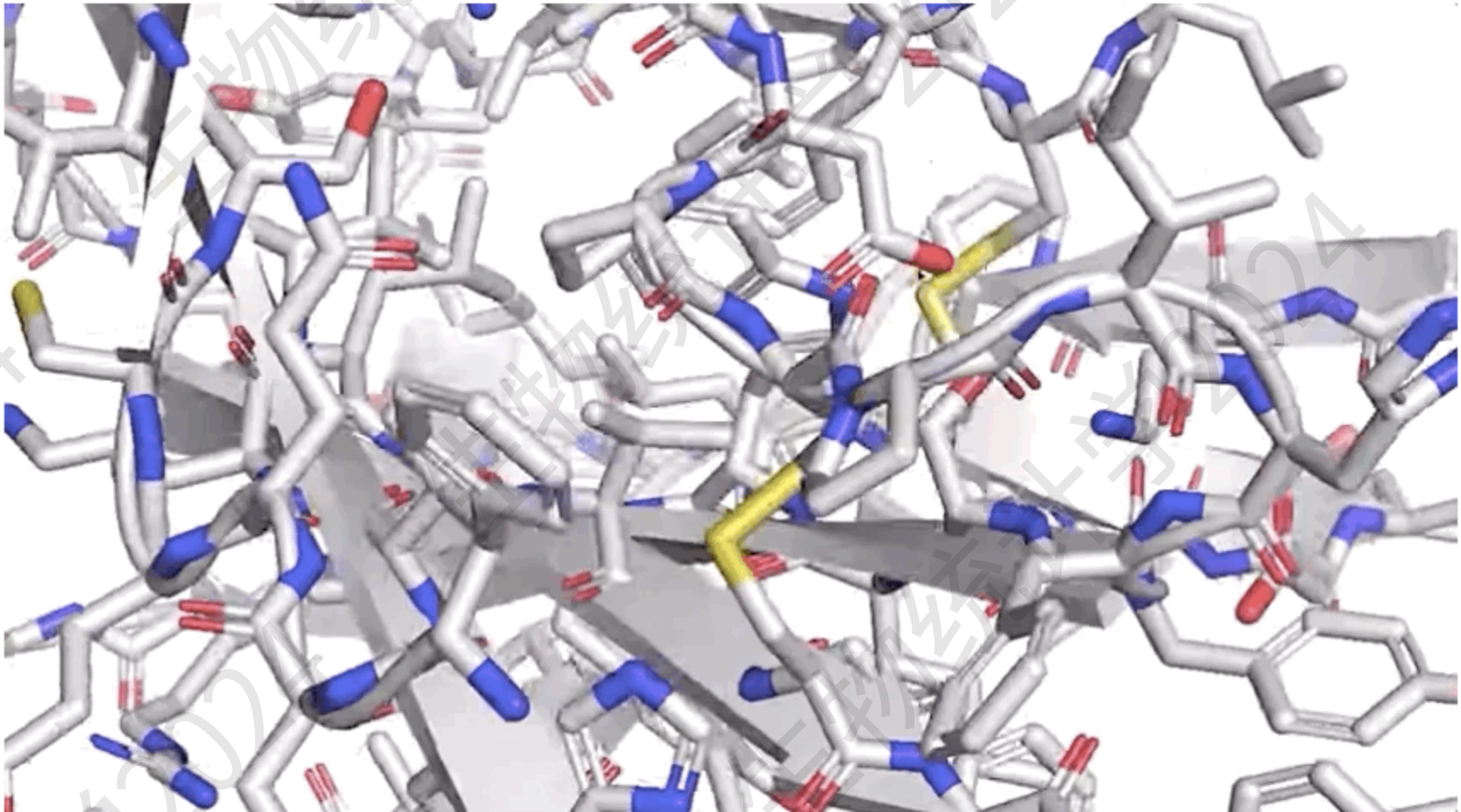
ATP



Transportation

Protein 3D structure

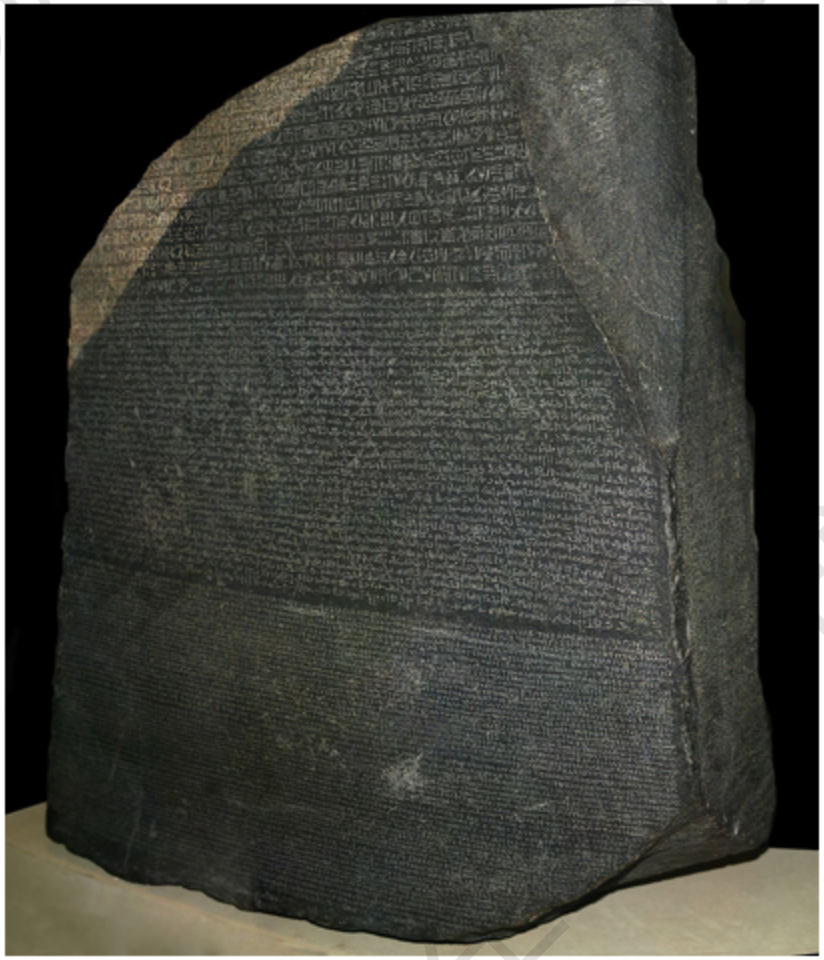
眼见为实



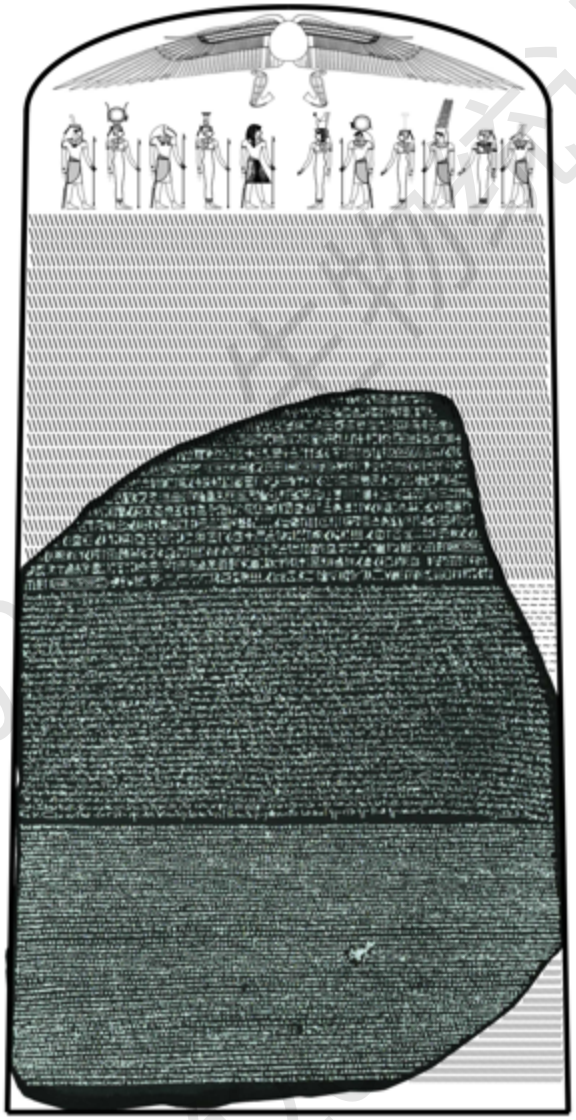
Protein 3D structure

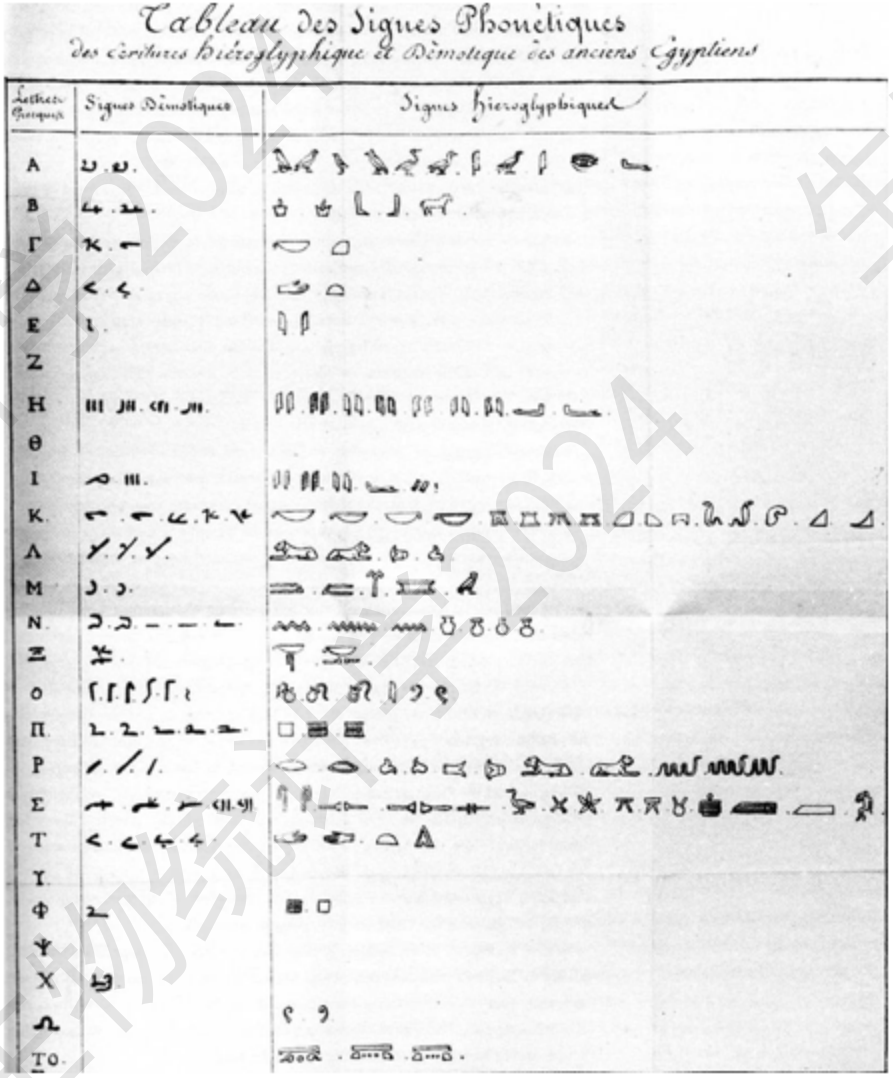
眼见为实

Rosetta stone

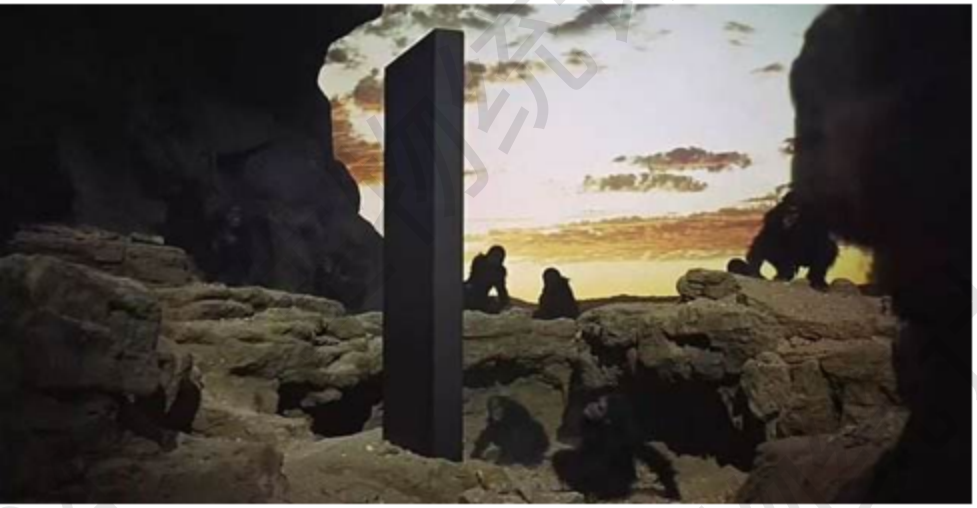


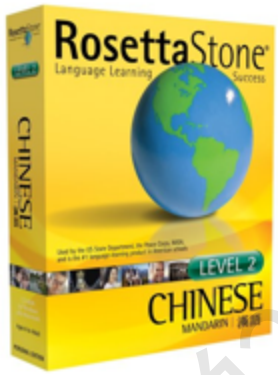
Rosetta stone

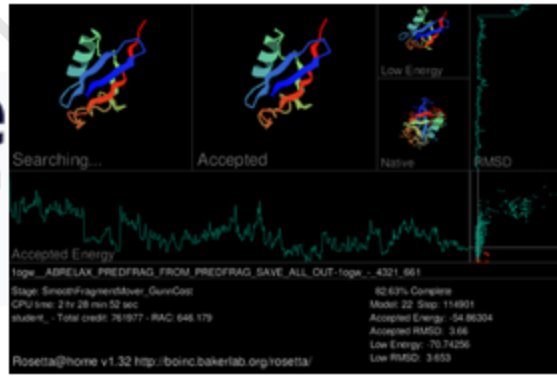




Rosetta stone







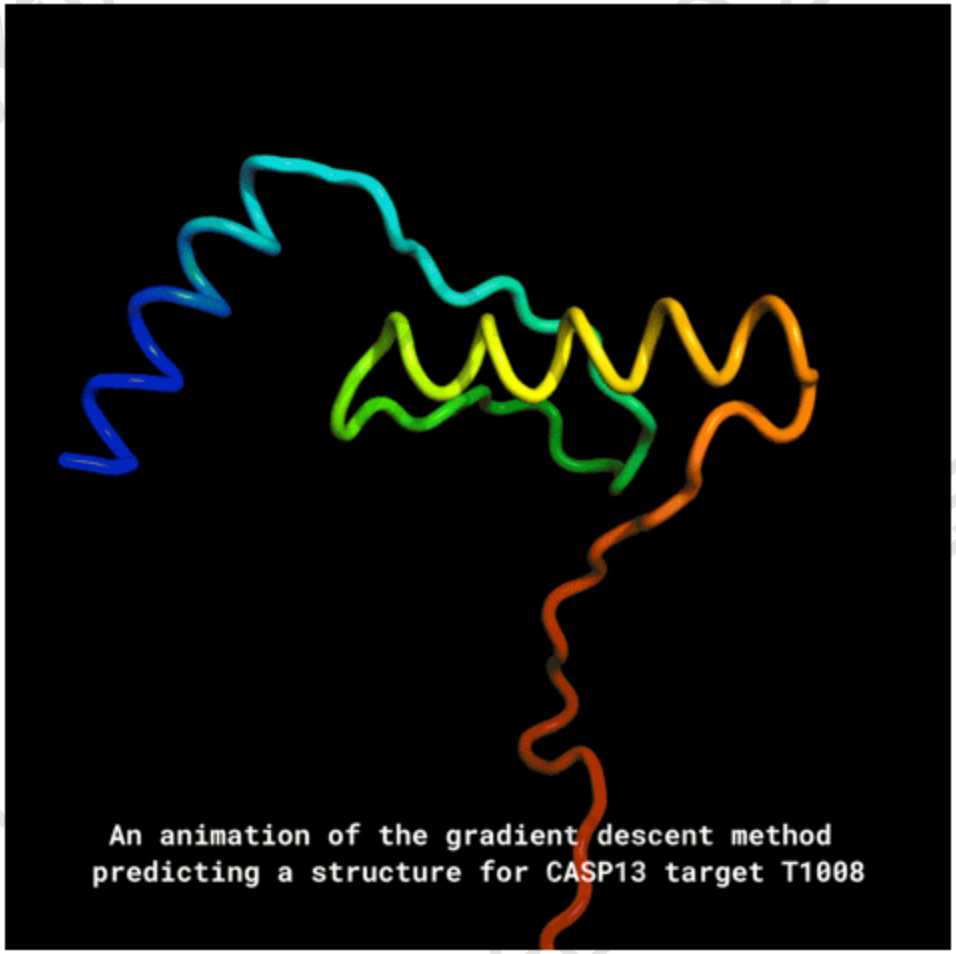
Protein 3D structure

1972年，克里斯蒂安·安芬森（Christian Anfinsen）在诺贝尔化学奖的获奖感言中，提出了一个著名的假设：

“理论上来说，蛋白质的氨基酸序列应该完全决定其结构。”

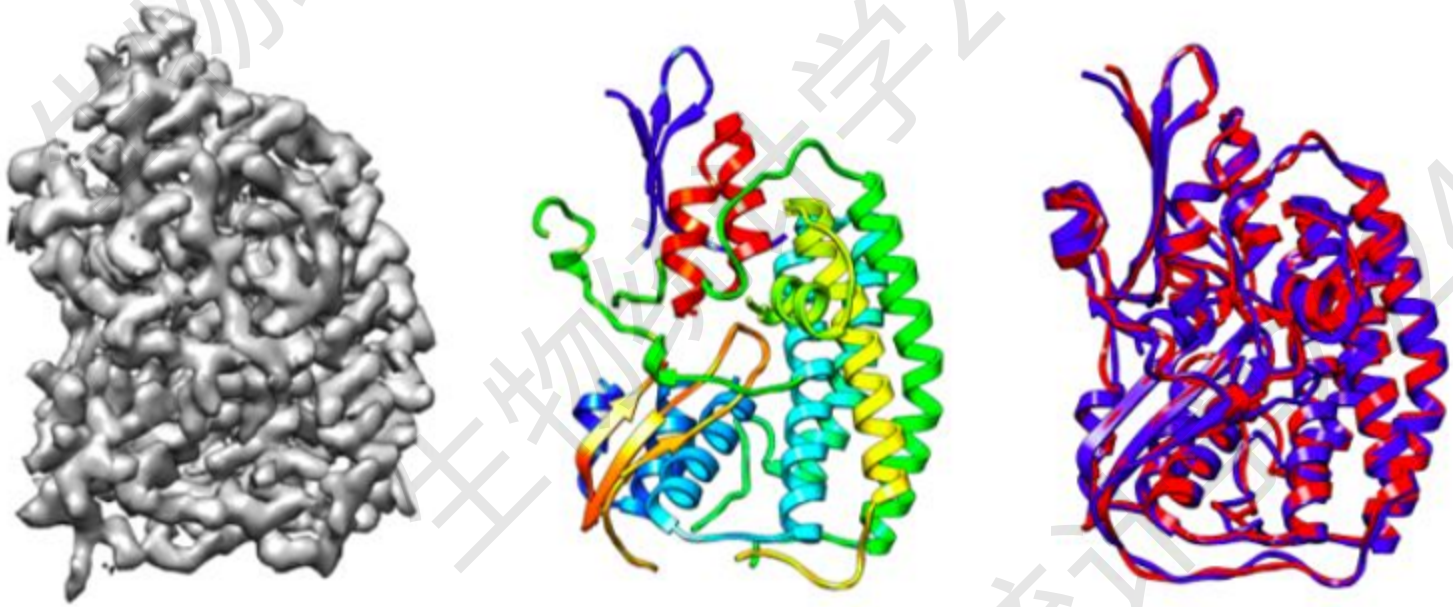
CASP

(Critical Assessment of Techniques for Protein Structure Prediction)



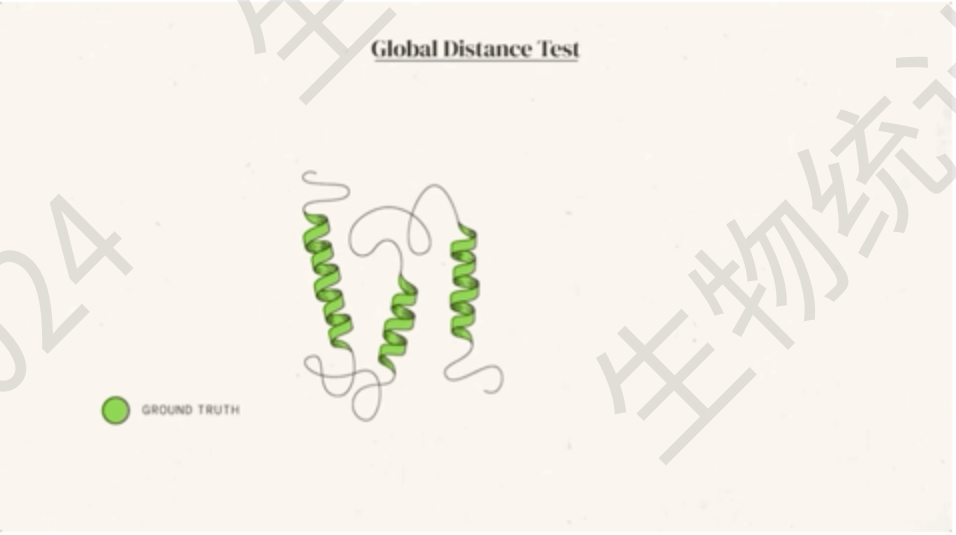
CASP

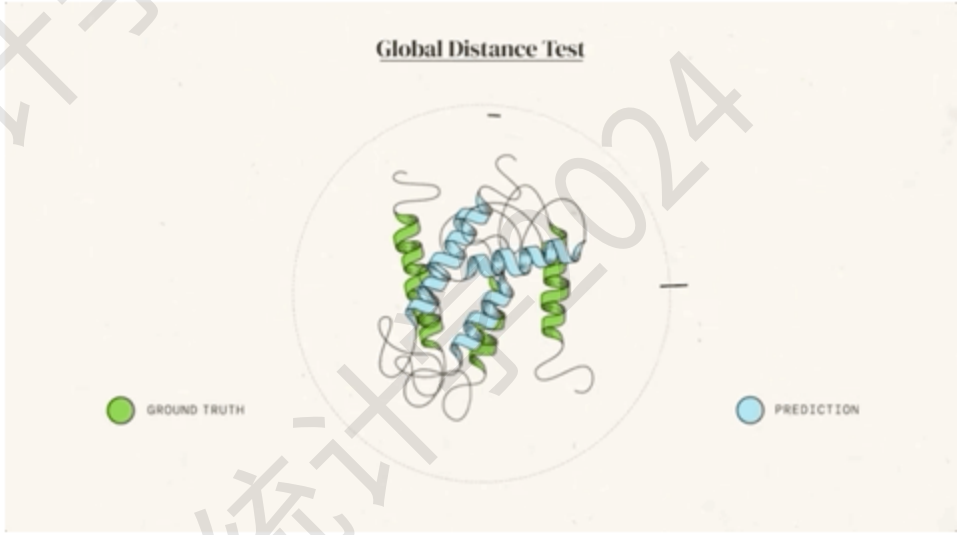
(Critical Assessment of Techniques for Protein Structure Prediction)



CASP

(Critical Assessment of Techniques for Protein Structure Prediction)





CASP

(Critical Assessment of Techniques for Protein Structure Prediction)

Rosetta Commons

The hub for Rosetta modeling software

I-TASSER

Protein Structure & Function Predictions

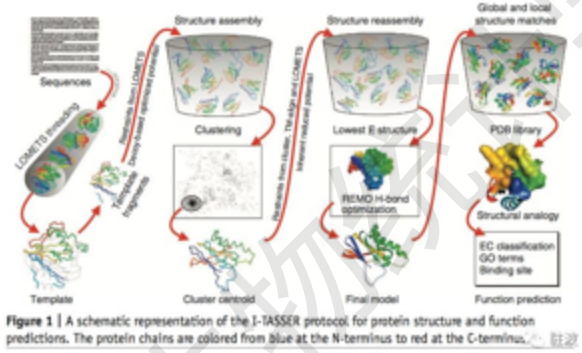
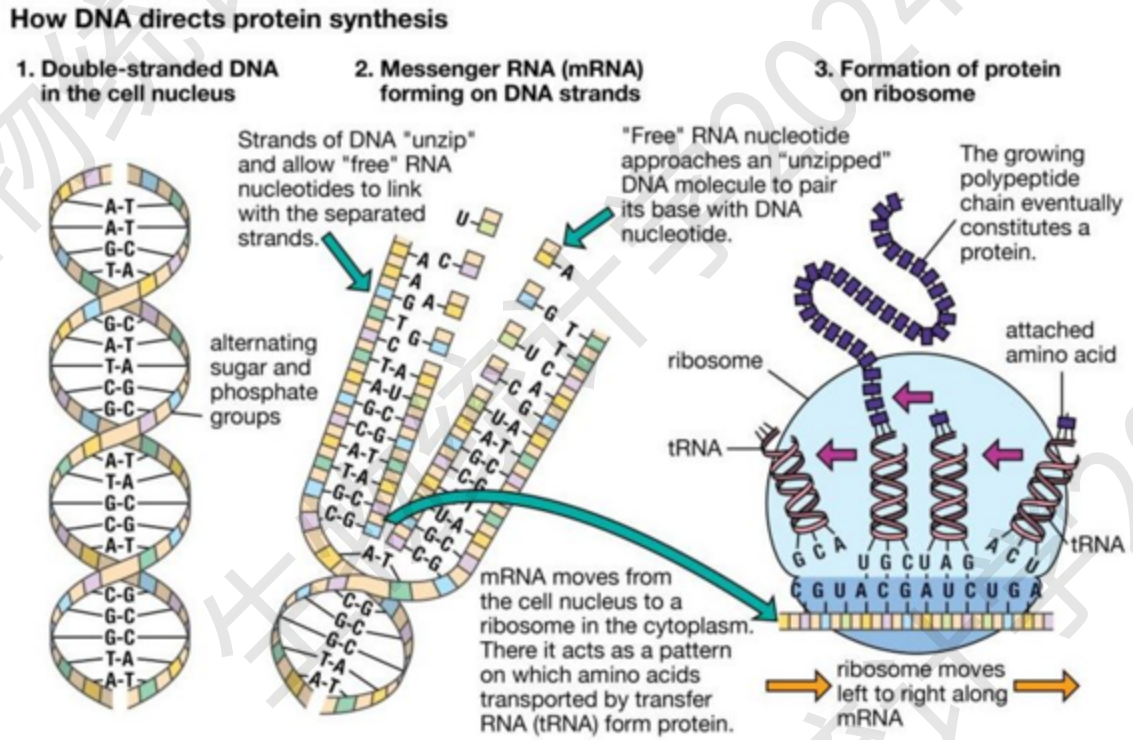


Figure 1 | A schematic representation of the I-TASSER protocol for protein structure and function predictions. The protein chains are colored from blue at the N-terminus to red at the C-terminus.

Protein 3D structure

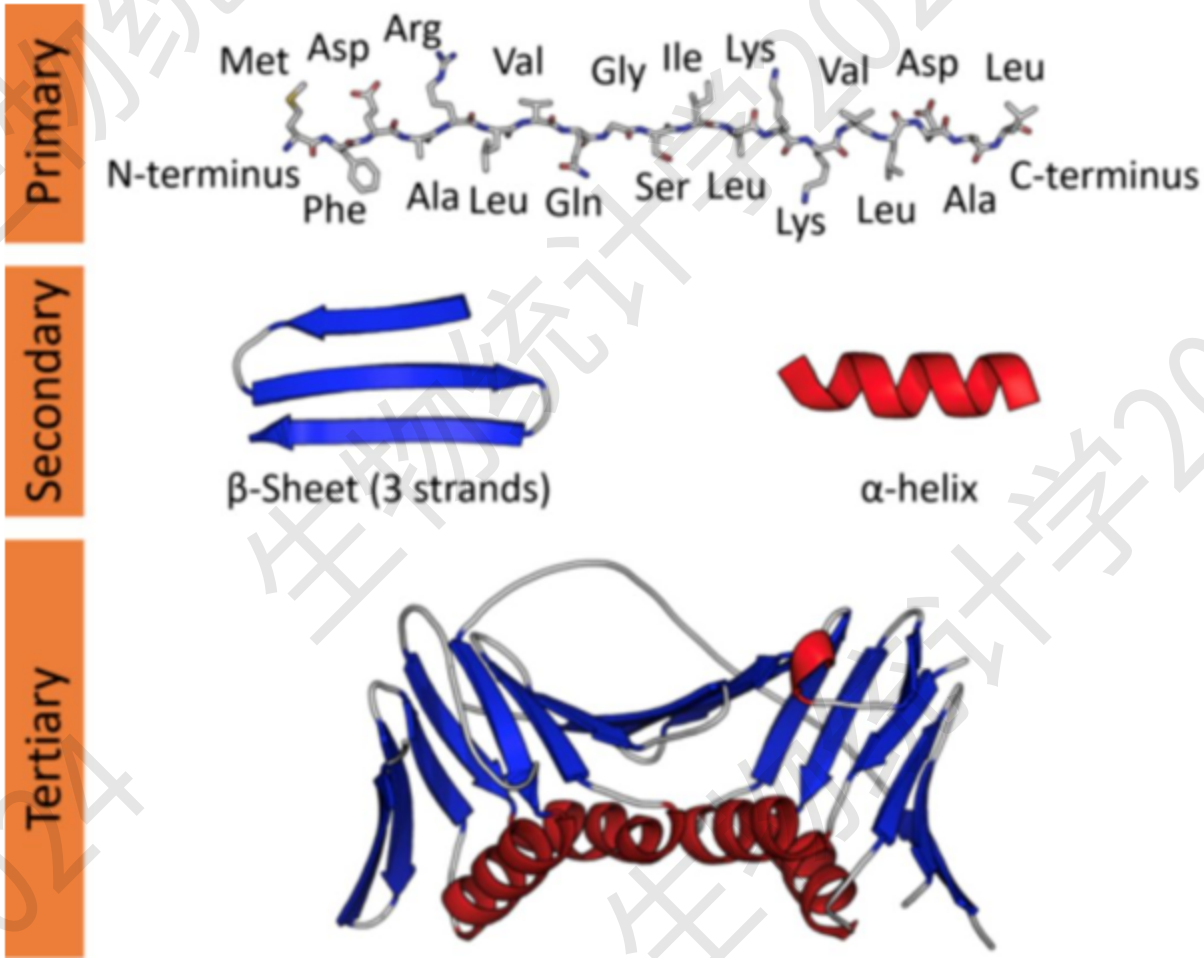
这一假设引发了长达五十年的探索，即仅仅基于蛋白质的一维氨基酸序列计算出其三维结构。



Levinthal 悖论：一种蛋白质大约存在 10^300 种可能构象。但在自然界中，蛋白质会自发折叠，有些只需几毫秒。

Protein 3D structure

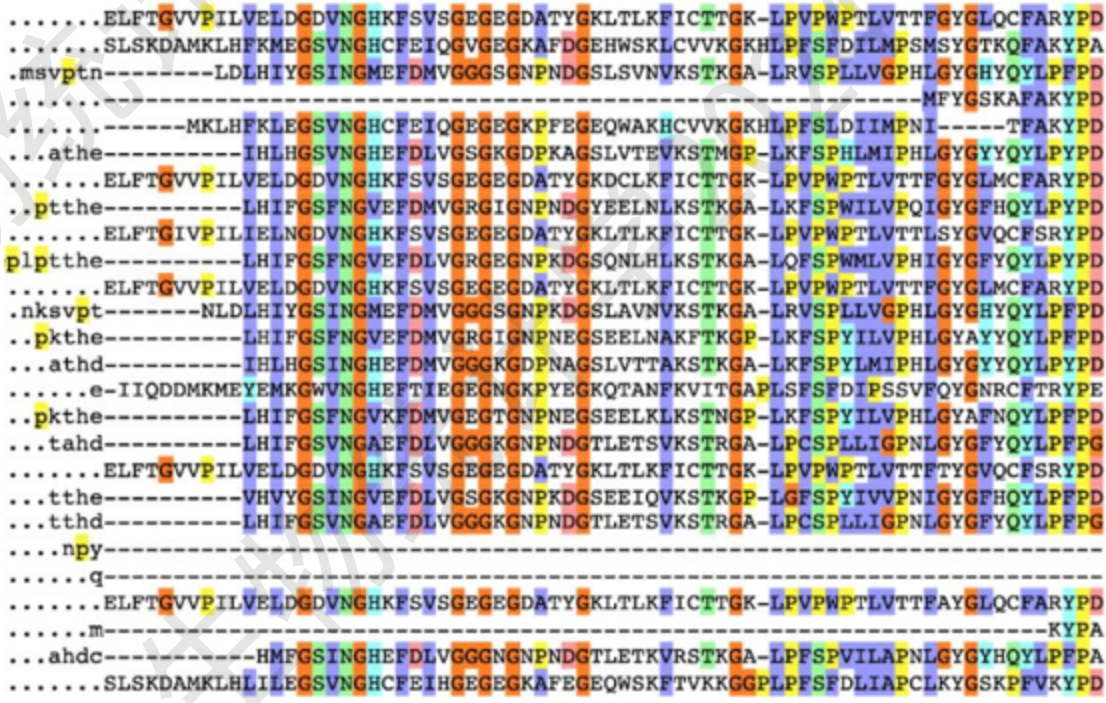
基本原理



Protein 3D structure

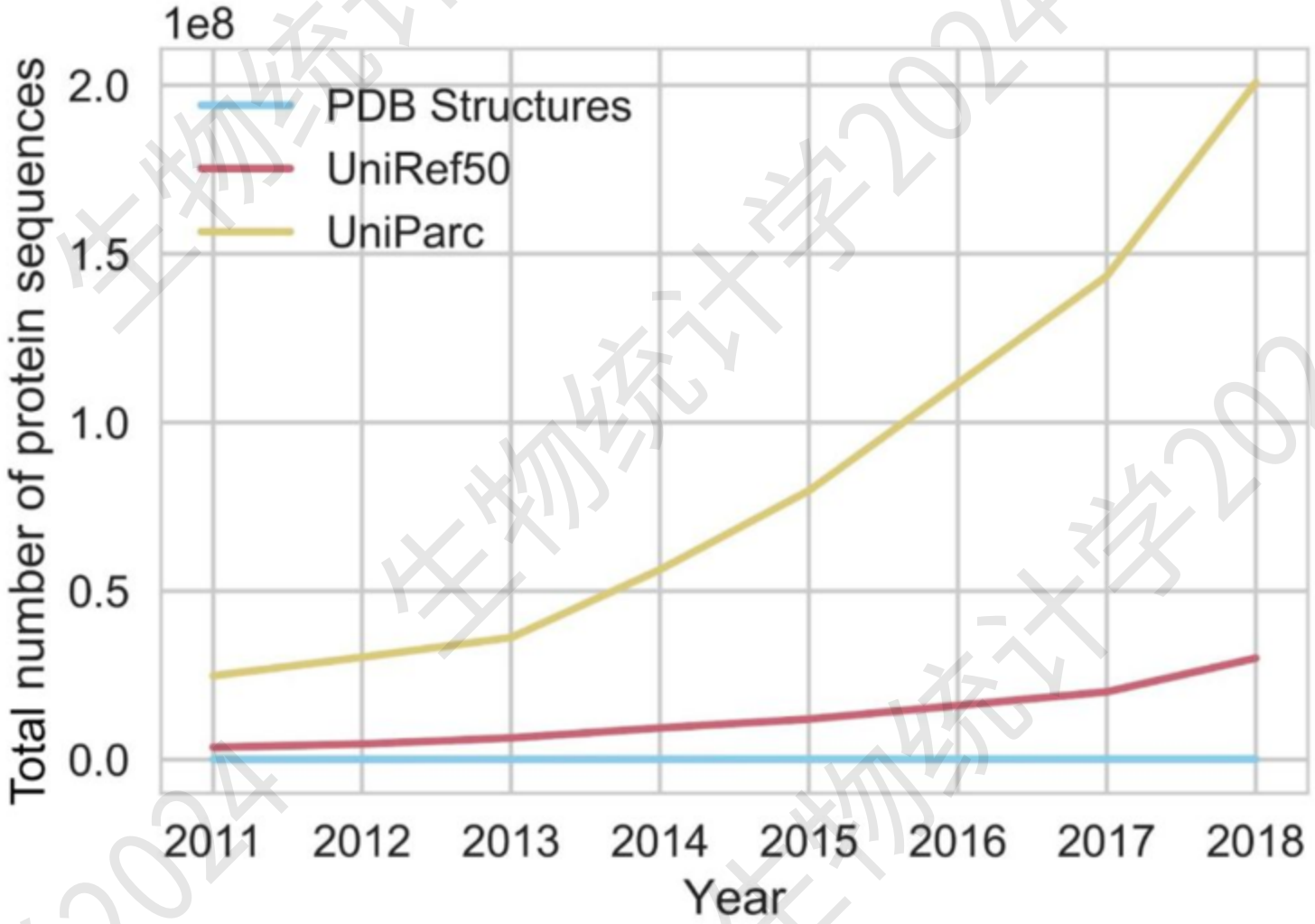
结构和序列

其中颜色渐变表示序列中从头到尾的索引



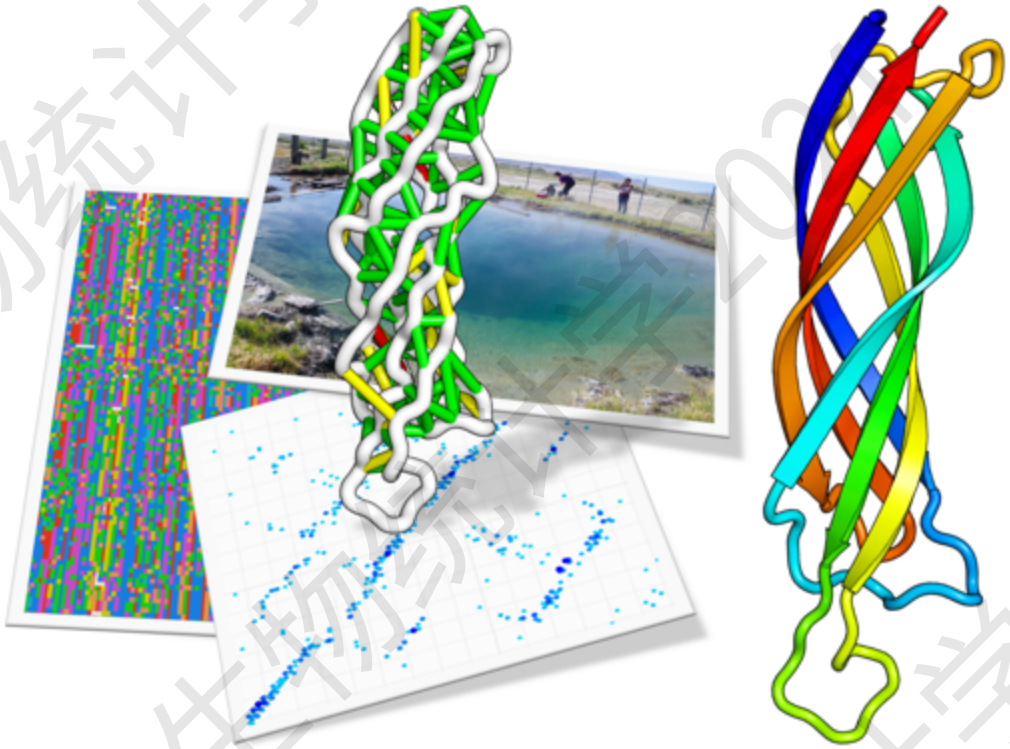
Protein 3D structure

大数据问题

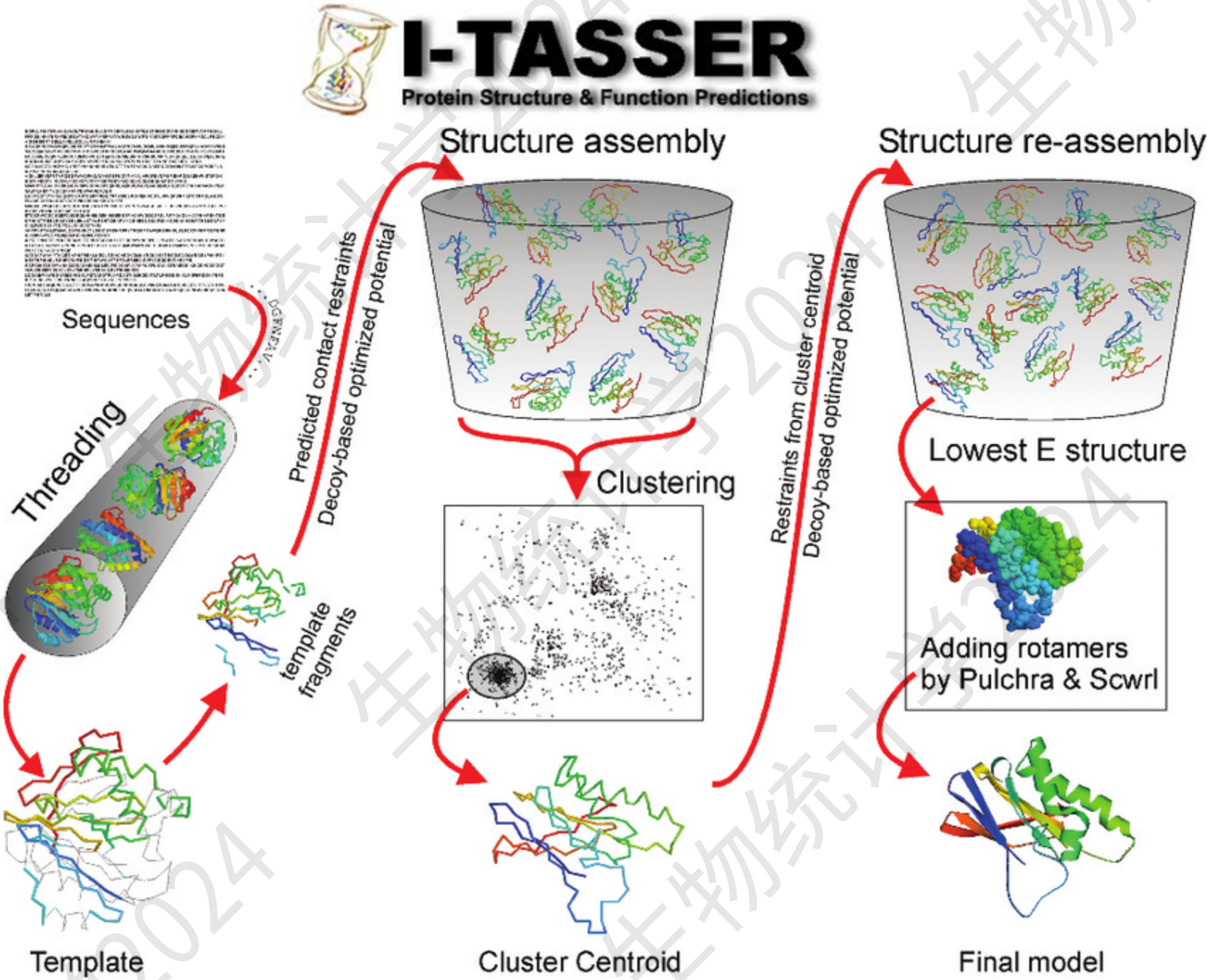


Rosetta Commons

The hub for Rosetta modeling software



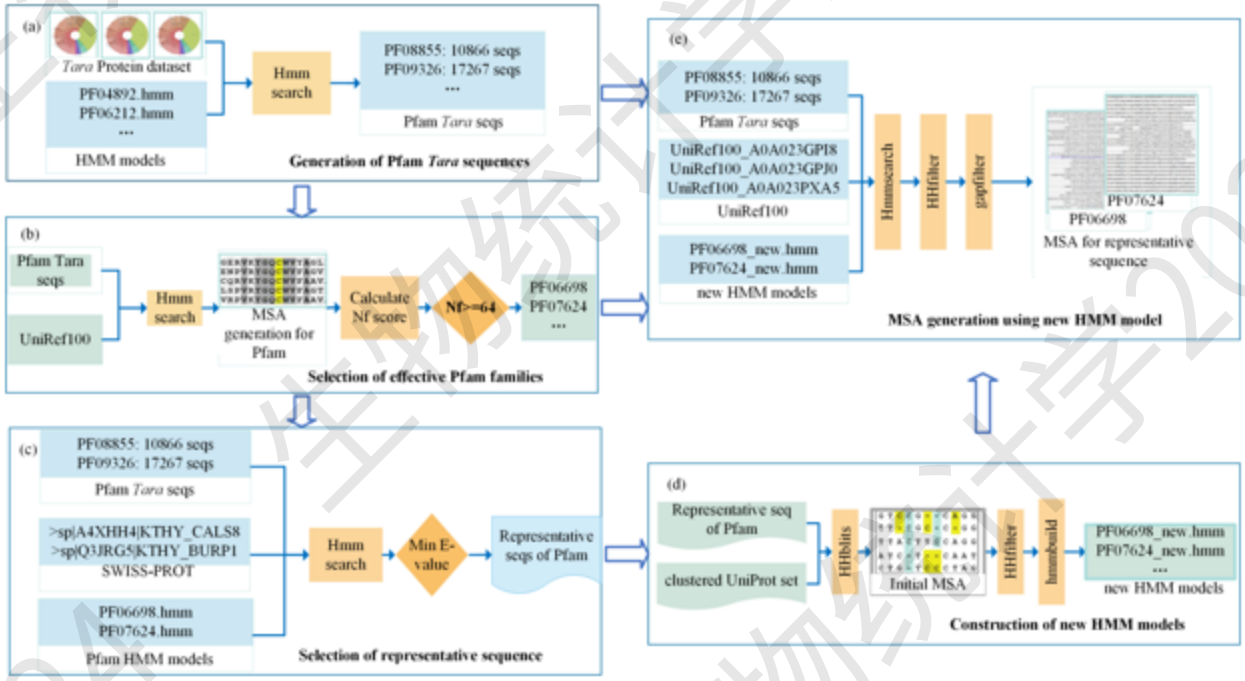
Top: Researchers gathering samples from Great Boiling Spring in Nevada. Left: a snapshot of aligned metagenomic sequences. Each row is a different sequence (the different colors are the different amino acid groups). Each position (or column) is compared to all other positions to detect patterns of co-evolution. Bottom: the strength of the top co-evolving residues is shown as blue dots, these are also shown as colored lines on the structure above. The goal is to make a structure that makes as many of these contacts as possible. Right: a cartoon of the protein structure predicted. The protein domain shown is from Pfam DUF3794, this domain is part of a Spore coat assembly protein SafA. (Image of Great Boiling Spring by Brian Hedlund, UNLV. Protein structure and composite image by Sergey Ovchinnikov, UW)


I-TASSER
Protein Structure & Function Predictions
Structure assembly
Structure re-assembly
Sequences
Threading
Predicted contact restraints
Decoy-based optimized potential
Clustering
Restraints from cluster centroid
Decoy-based optimized potential
Lowest E structure
template fragments
Adding rotamers by Pulchra & Scwrl
Template
Cluster Centroid
Final model

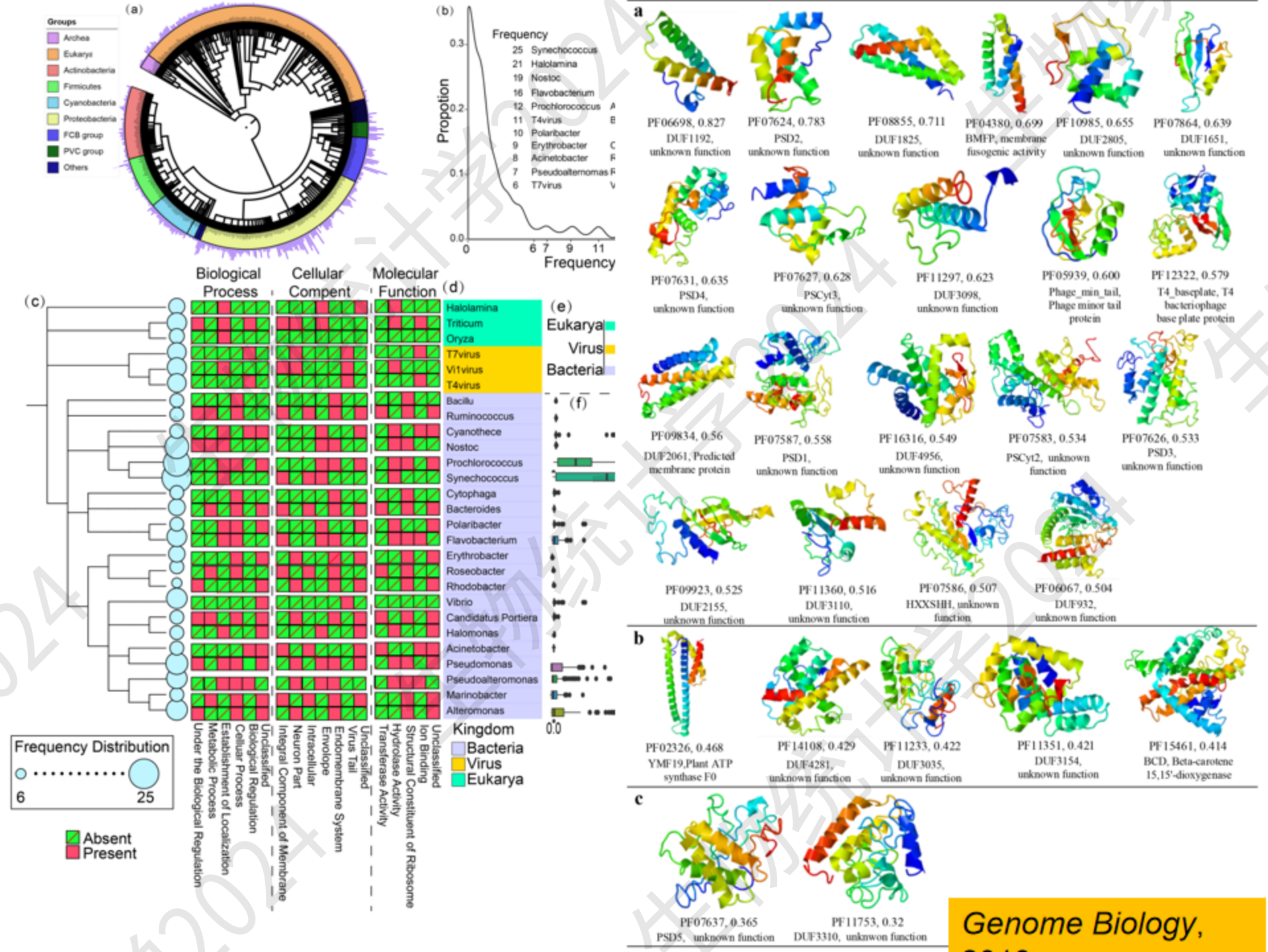
微生物组大数据 + 蛋白质3D结构

基于大数据挖掘的蛋白质结构预测和功能解析

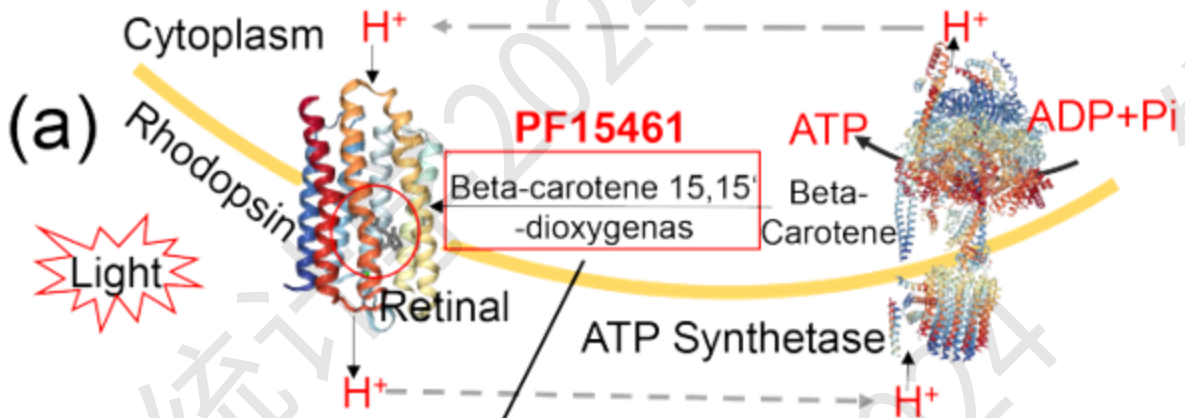
- 2TB的海洋微生物组大数据-->鉴定出了超过9千万的非冗余基因和超过3万个微生物物种
- 预测出了之前没有任何结构信息的27个蛋白质结构
- 利用人工神经网络挖掘蛋白结构和生境的关系



Genome Biology,



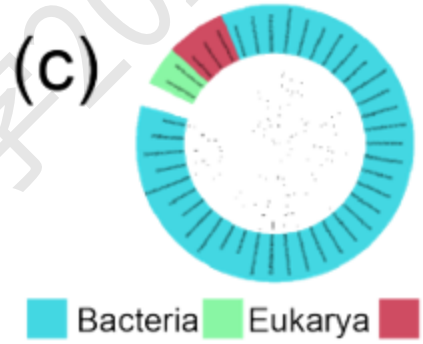
Genome Biology,

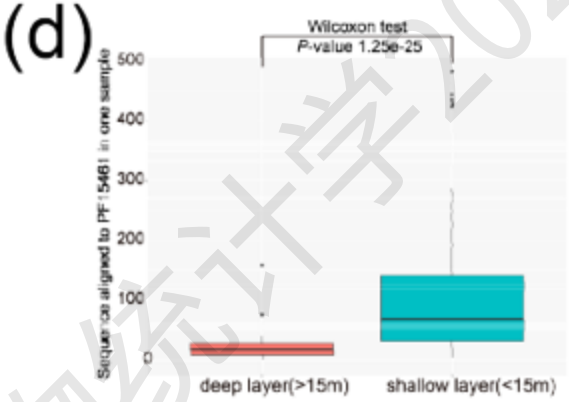


Predicted structure of PF15461

318 Amino Acid
Beta-carotene 15,15'-dioxygenase
369 sequence→14,353 sequence

Predicted Function:
Cellular Component: Respiratory Chain
Biological Process:
Single-organism Metabolic Process
Molecular Function:Oxidoreductase Activity





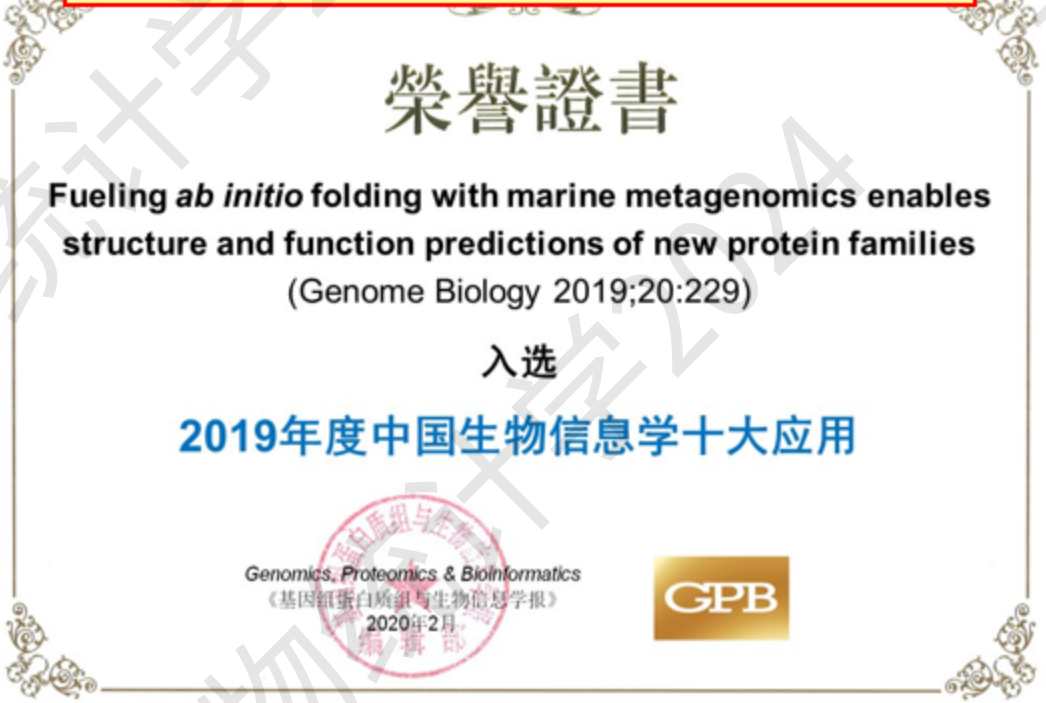
Genome Biology,

微生物组大数据 + 蛋白质3D结构

潜在应用途径

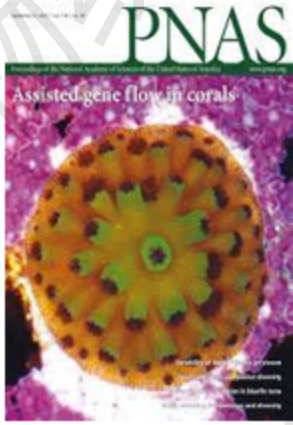
- 非定向/定向方法可以发掘大量未知功能基因
- 为合成生物学提供功能模块
- 药物发掘和药物设计
  - 环境菌群功能基因
  - 环境菌群代谢物

2019年度中国在生物信息学十大应用



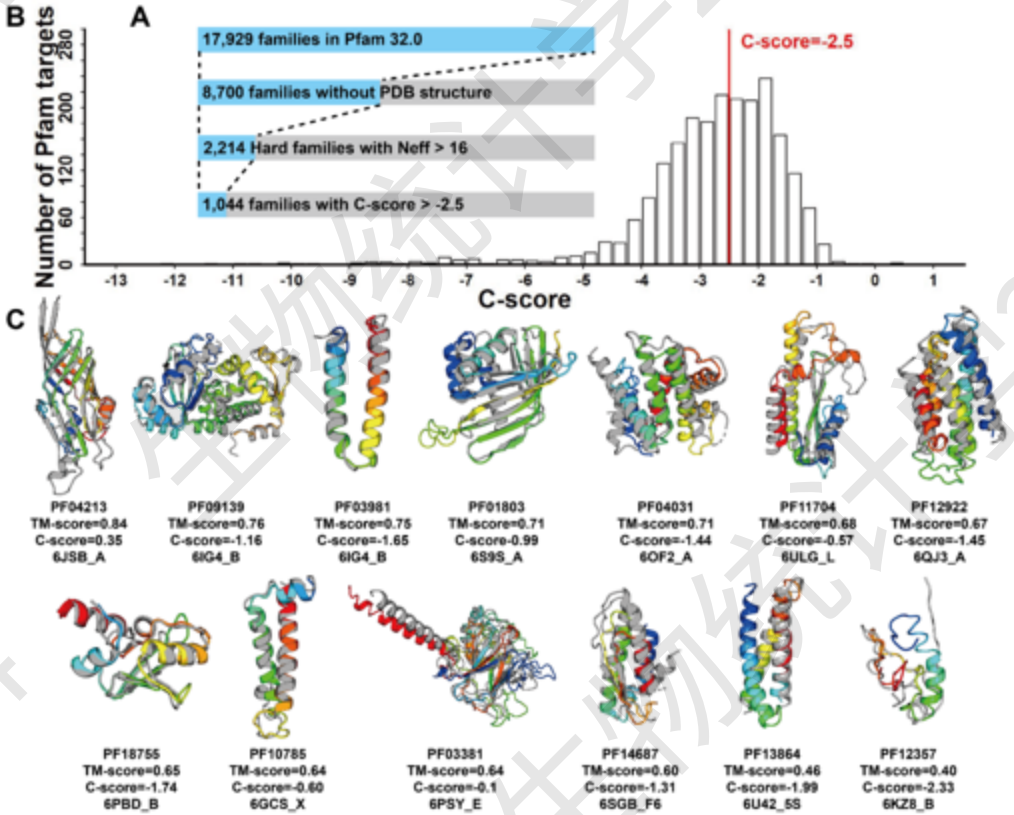
Genome Biology,

微生物组大数据 + 蛋白质3D结构



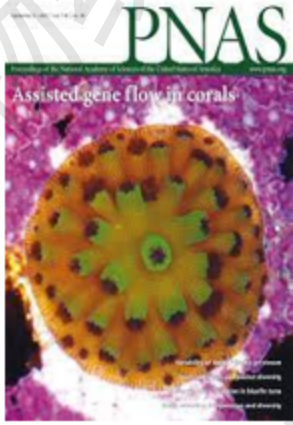
基于大数据挖掘的蛋白质结构预测和功能解析

- 42.5亿微生物组来源序列大数据
- 预测出了之前没有任何结构信息的1,044个蛋白质结构
- 利用人工神经网络挖掘蛋白结构和生境以及进化的关系



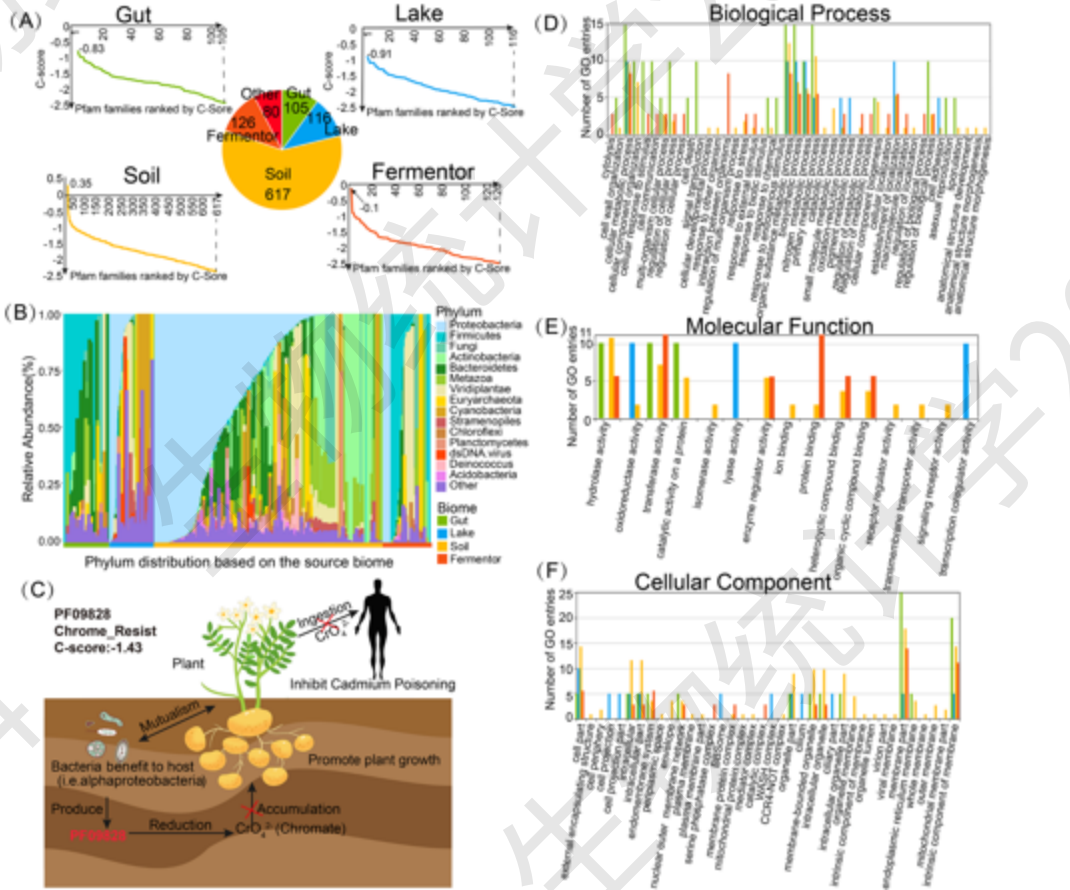
PNAS,

微生物组大数据 + 蛋白质3D结构



基于大数据挖掘的蛋白质结构预测和功能解析

- 42.5亿微生物组来源序列大数据
- 预测出了之前没有任何结构信息的1,044个蛋白质结构
- 利用人工神经网络挖掘蛋白结构和生境以及进化的关系

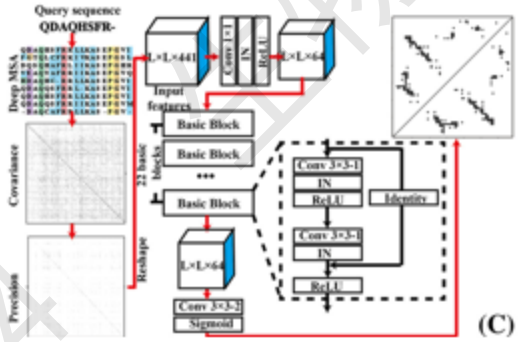


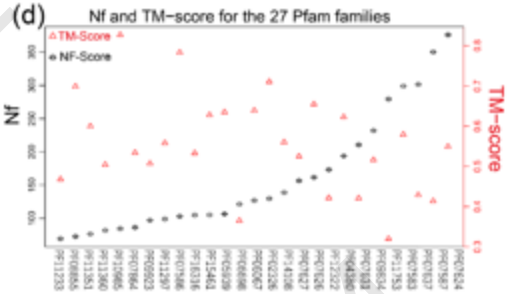
PNAS,

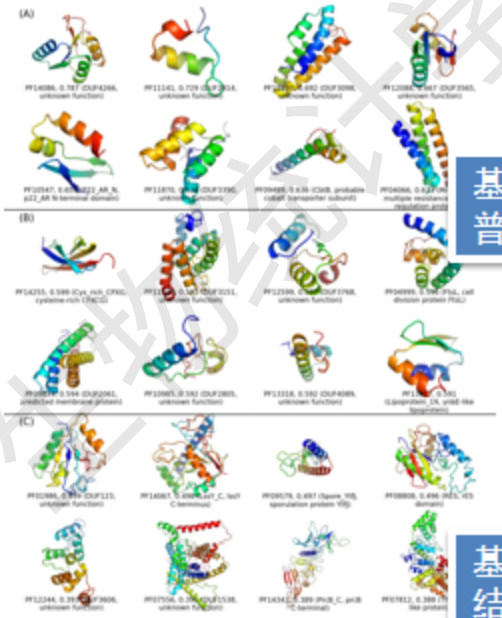
基于深度学习的微生物组大数据挖掘

- 整合异质性微生物组大数据（>10TB），预测数百个全新蛋白质家族3D结构
- 提出了靶向性微生物组大数据挖掘法，提高了蛋白结构预测效率
- 针对微生物组样本溯源、聚类提出了基于本体神经网络的全新方法

挖掘方法开发:
Bioinformatics, 2018;
Genome Biology, 2019;
PNAS, 2021;
BIB, 2021;
Genome Medicine, 2022;
BIB, 2022a; BIB, 2022b







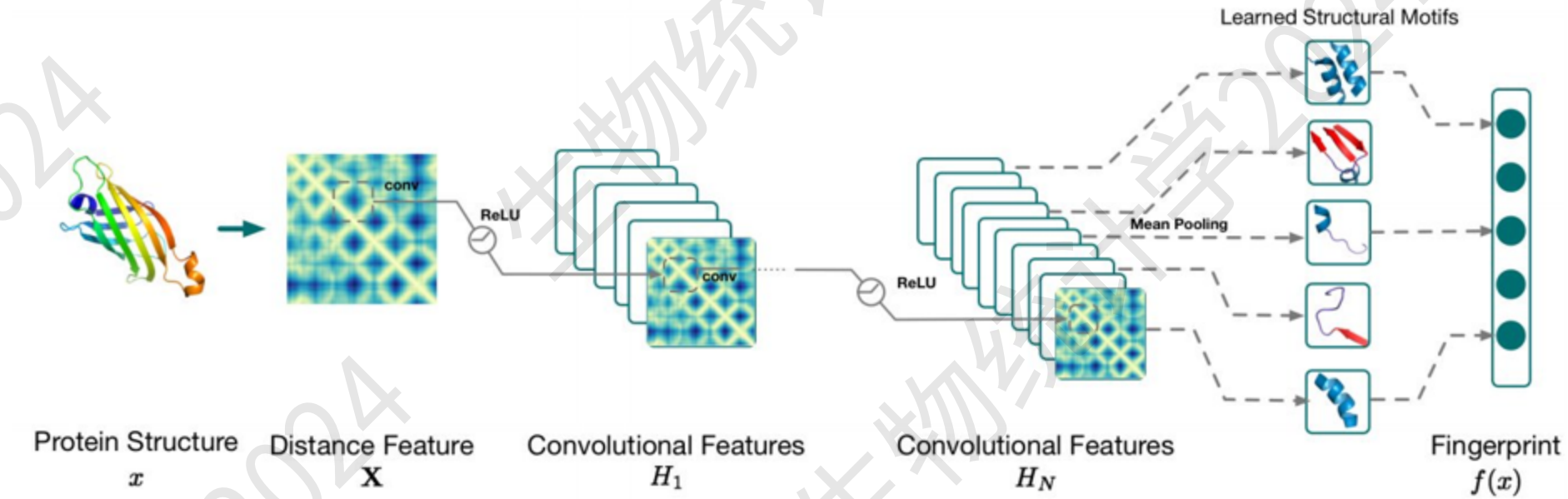
基于本体神经网络的普适性预测模型

基于卷积神经网络的结构感知模型

CASP13

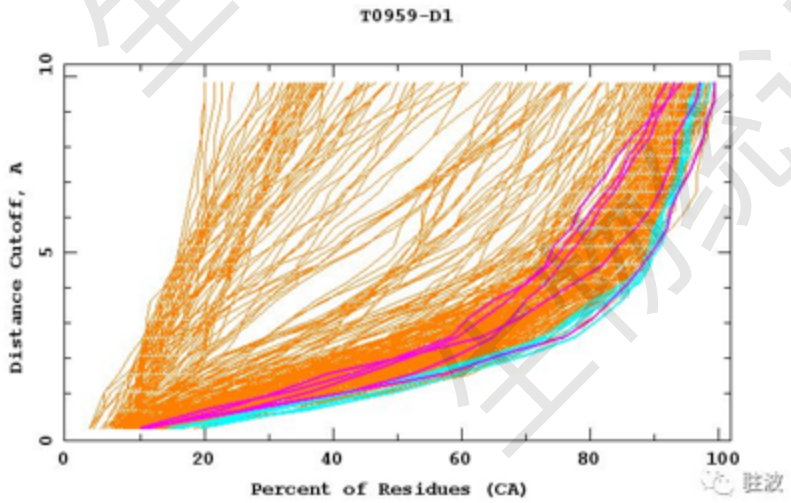
(Critical Assessment of Techniques for Protein Structure Prediction)

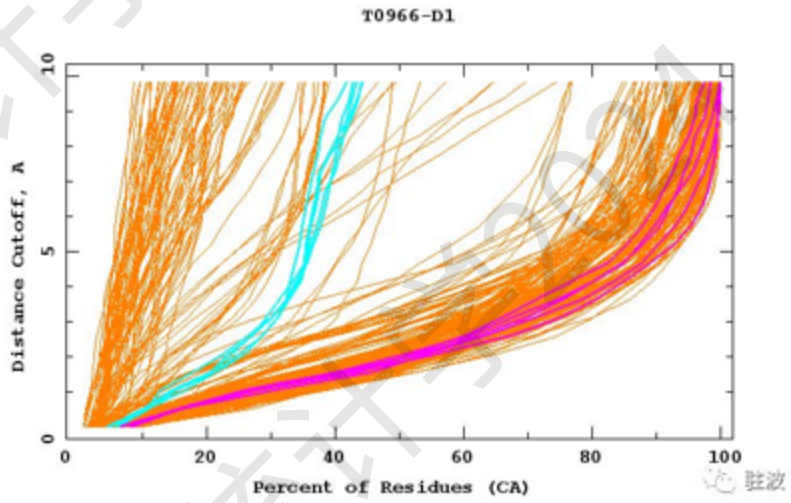
DeepFold



CASP13

(Critical Assessment of Techniques for Protein Structure Prediction)



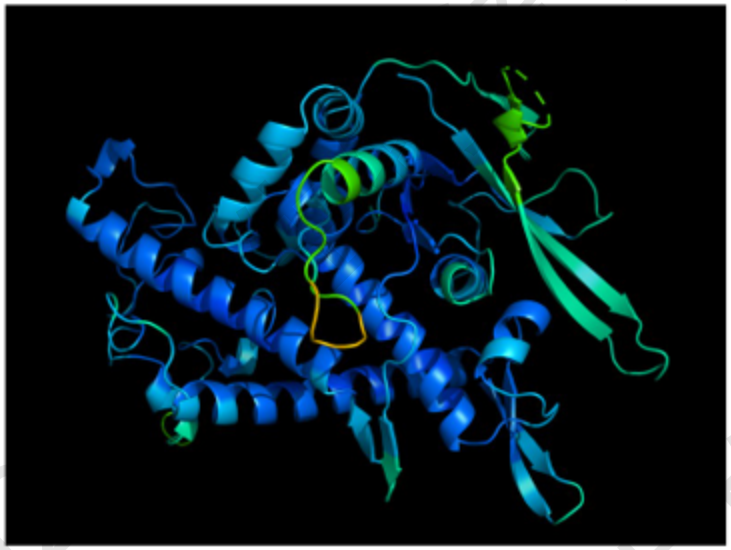


NEWS · 30 NOVEMBER 2020

'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures

Google's deep-learning program for determining the 3D shapes of proteins stands to transform biology, say scientists.

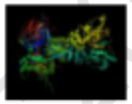
Ewen Callaway



A protein's function is determined by its 3D shape. Credit: DeepMind

RELATED ARTICLES

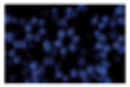
AI protein-folding algorithms solve structures faster than ever



The revolution will not be crystallized: a new method sweeps through structural biology

The computational protein designers

Revolutionary microscopy technique sees individual atoms for first time



SUBJECTS

CASP14

(Critical Assessment of Techniques for Protein Structure Prediction)

AlphaFold: a solution to a 50-year-old grand challenge in biology

DeepMind宣布，其新一代AlphaFold人工智能系统，在国际蛋白质结构预测竞赛（CASP）上击败了其余的参会选手，能够精确地基于氨基酸序列，预测蛋白质的3D结构。其准确性可以与使用冷冻电子显微镜（CryoEM）、核磁共振或 X 射线晶体学等实验技术解析的3D结构相媲美。

CASP14

(Critical Assessment of Techniques for Protein Structure Prediction)

Sundar Pichai @sundarpichai · 10h

.@DeepMind's incredible AI-powered protein folding breakthrough will help us better understand one of life's fundamental building blocks + enable researchers to tackle new and hard problems, from fighting diseases to environmental sustainability.

AlphaFold: a solution to a 50-year-old grand challenge in biology

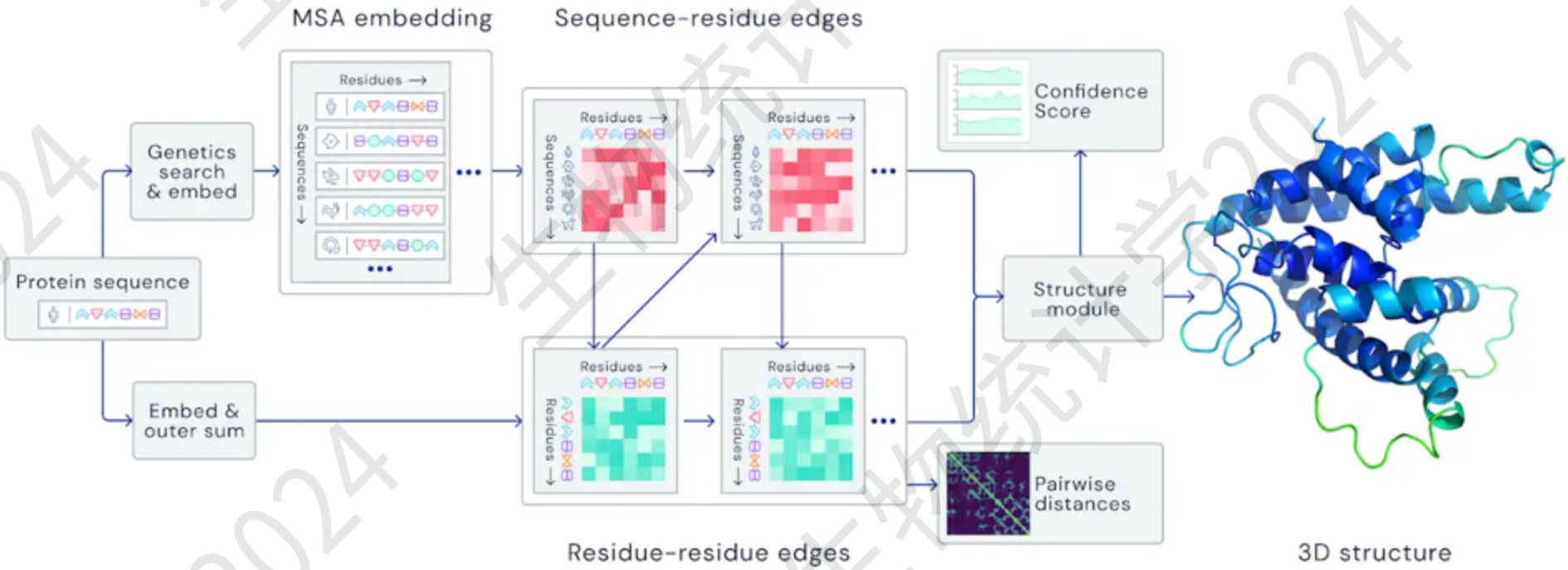
In a major scientific advance, the latest version of our AI system AlphaFold has been recognised as a solution to this grand challenge ...

deepmind.com

CASP14

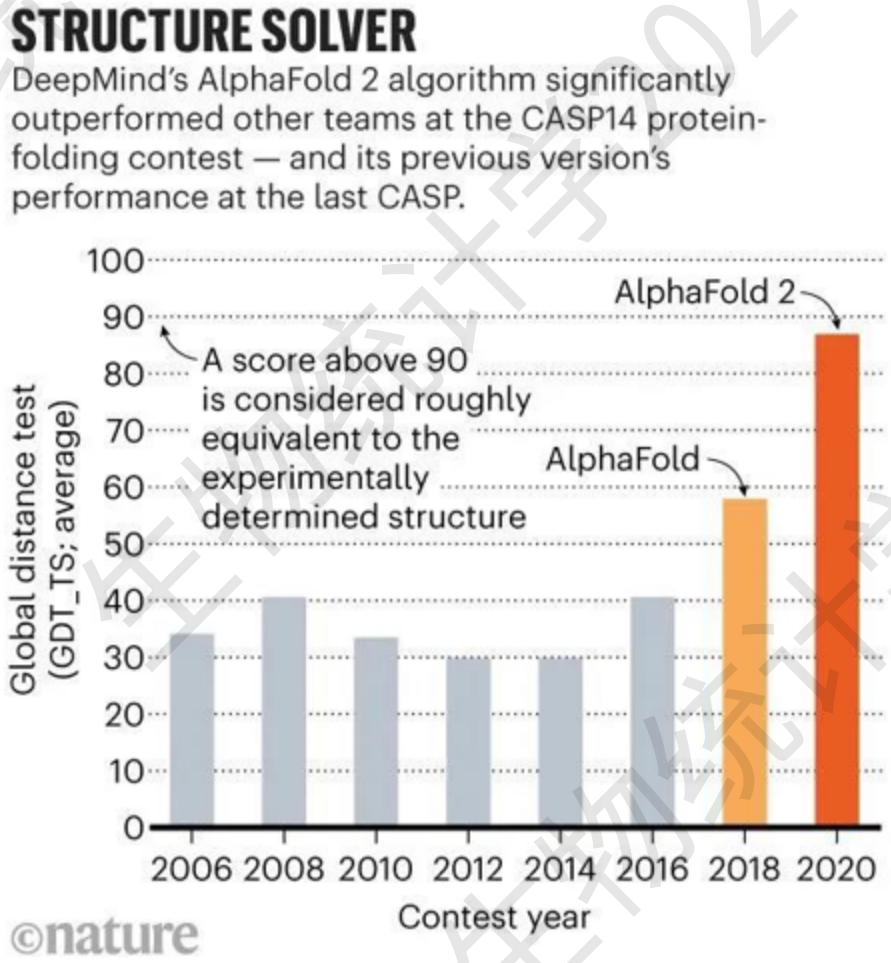
(Critical Assessment of Techniques for Protein Structure Prediction)

AlphaFold2



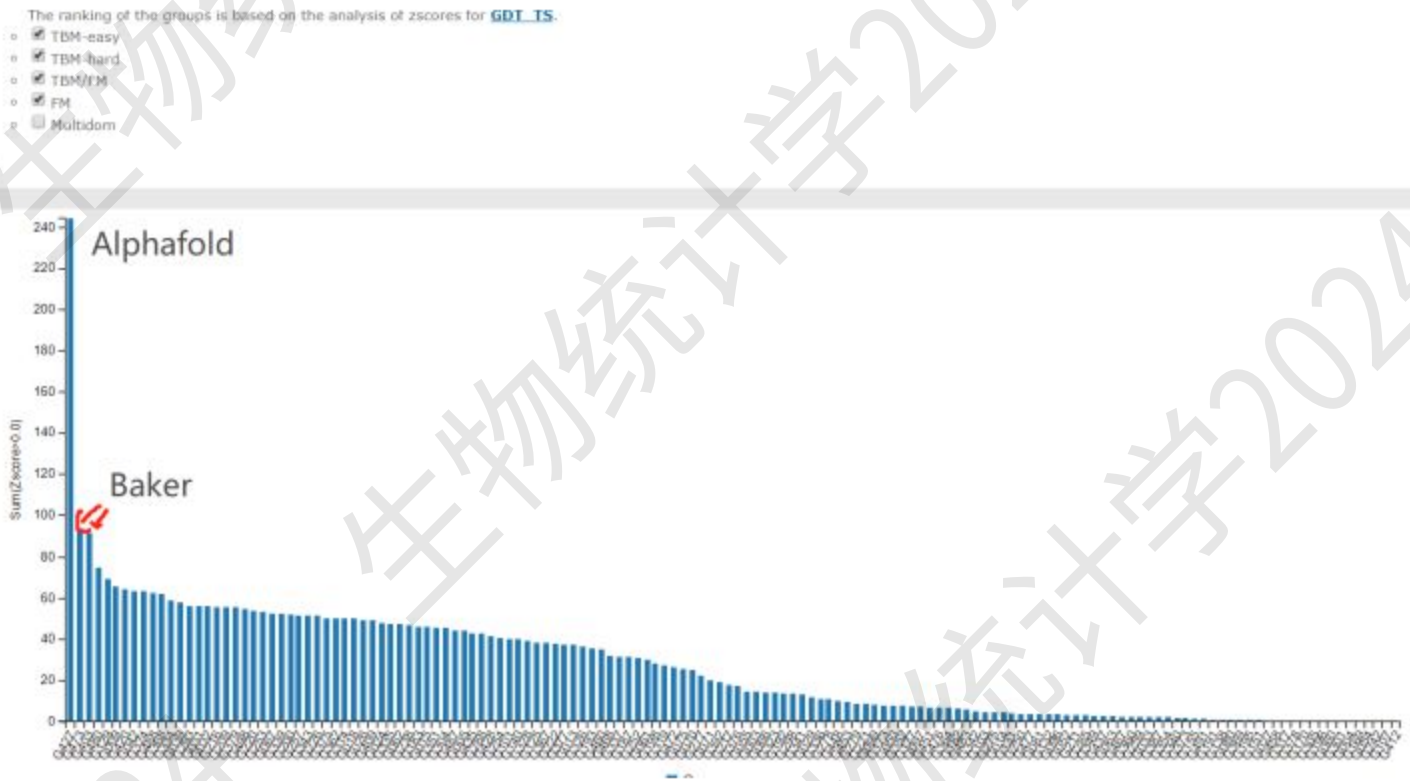
CASP14

(Critical Assessment of Techniques for Protein Structure Prediction)



CASP14

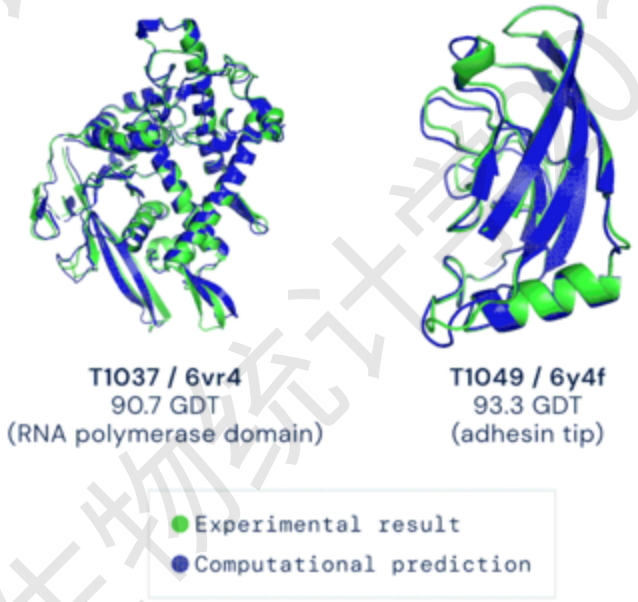
(Critical Assessment of Techniques for Protein Structure Prediction)



https://predictioncenter.org/casp14/zscores_final.cgi

CASP14

(Critical Assessment of Techniques for Protein Structure Prediction)



在2020的CASP中，AlphaFold系统对所有蛋白靶点3D结构预测的中位GDT评分为92.4分。即便是针对最难解析的蛋白靶点，AlphaFold的中位GDT评分也达到了87.0分。在接受检验的近100个蛋白靶点中，AlphaFold对三分之二的蛋白靶点给出的预测结构与实验手段获得的结构相差无几。CASP创始人Moult教授表示，在有些情况下，已经无法区分两者之间的区别是由于AlphaFold的预测出现错误，还是实验手段产生的假象。

CASP14
(Critical Assessment of Techniques for Protein Structure Prediction)

- 我们可以把蛋白质折叠看作一个「空间图」，节点表示残基（residue），边则将残基紧密连接起来。这个空间图对于理解蛋白质内部的物理交互及其演化史至关重要。对于在 CASP14 比赛中使用的最新版 AlphaFold，DeepMind 团队创建了一个基于注意力的神经网络系统，并用端到端的方式进行训练，以理解图结构，同时基于其构建的隐式图执行推理。该方法使用进化相关序列、多序列比对（MSA）和氨基酸残基对的表示来细化该图。

- DeepMind 团队在公开数据上训练这一系统，这些数据来自蛋白质结构数据库（PDB）和包含未知结构蛋白质序列的大型数据库，共包括约 170,000 个蛋白质结构。该系统使用约 128 个 TPUv3 内核（相当于 100-200 个 GPU）运行数周，与现今机器学习领域出现的大型 SOTA 模型相比，该系统所用算力相对较少。

- AlphaFold 还具备很多令人兴奋的技术潜力：探索数亿个目前还没有模型的数亿蛋白质，以及未知生物的广阔领域。由于 DNA 指定了构成蛋白质结构的氨基酸序列，基因组学革命使大规模阅读自然界的蛋白质序列成为可能——在通用蛋白质数据库（UniProt）中有 1.8 亿个蛋白质序列。相比之下，考虑到从序列到结构所需的实验工作，蛋白质数据库（PDB）中只有大约 170000 个蛋白质结构。在未确定的蛋白质中可能有一些新的和未确定的功能——就像望远镜帮助人类更深入的观察未知宇宙一样，像 AlphaFold 这样的技术可以帮助找到未确定的蛋白质结构。

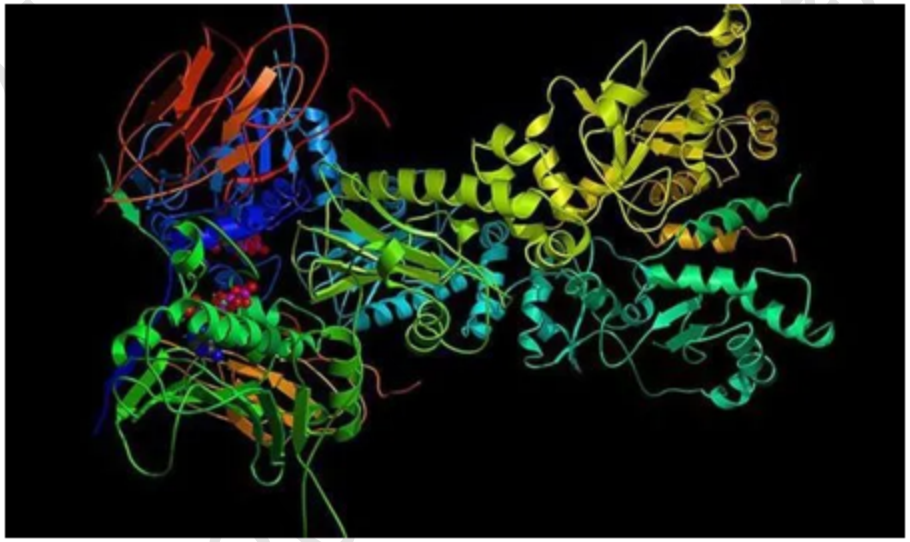
CASP14
(Critical Assessment of Techniques for Protein Structure Prediction)

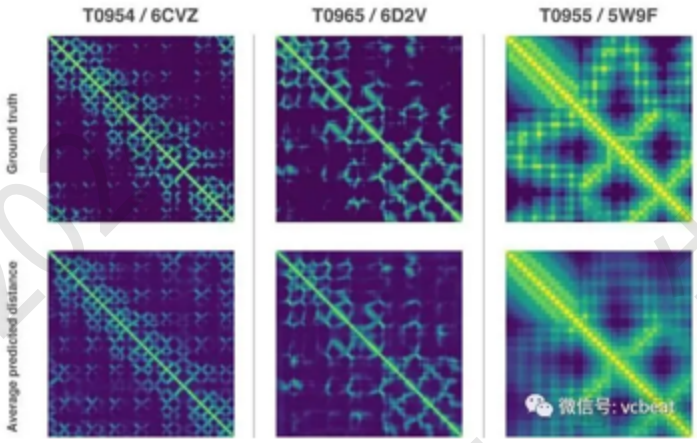
- 在今年早些时候，DeepMind已经利用这一系统预测了多种新冠病毒蛋白的结构。后续的实验显示，AlphaFold预测的新冠病毒Orf3a蛋白结构与冷冻电镜解析的结构非常相似。

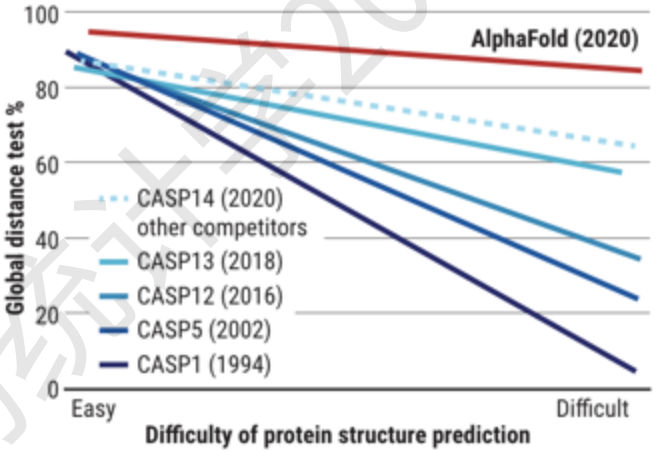
- 虽然，AlphaFold不见得会取代冷冻电子显微镜等其它实验手段，但是DeepMind的研究人员表示，这一令人兴奋的结果表明，生物学家们可以使用计算结构预测作为科学研究的核心工具之一。这一手段对于特定类型的蛋白来说可能尤为便利，例如膜蛋白一直非常难于结晶，因此很难用实验手段获得它们的结构。

- 而对于从事计算和机器学习研究的DeepMind团队来说，AlphaFold的表现证明了AI在辅助基础科学发现方面惊人的潜力。该团队在公司发布的博文中表示，他们相信，AI将成为人类拓展科学知识前沿最有力的工具之一！

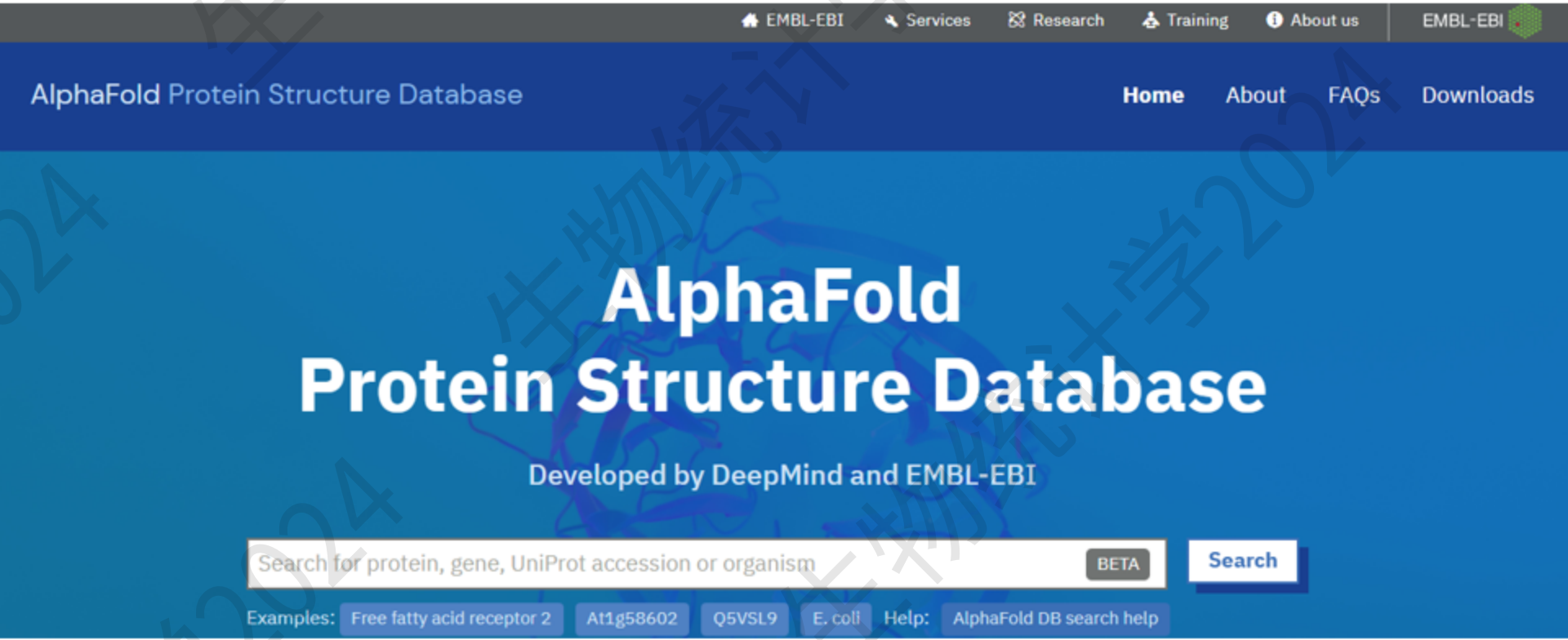
AI + 蛋白质3D结构




T0954 / 6CVZ
T0965 / 6D2V
T0955 / 5W9F
Ground truth
Average predicted distance
微信号: vcbeat


AlphaFold (2020)
CASP14 (2020)
other competitors
CASP13 (2018)
CASP12 (2016)
CASP5 (2002)
CASP1 (1994)
Global distance test %
100
80
60
40
20
0
Easy
Difficult
Difficulty of protein structure prediction

AI + 蛋白质3D结构



google-deepmind/ **alphafold**

Open source code for AlphaFold.

16 Contributors | 7 Used by | 11k Stars | 2k Forks

https://github.com/sokrypton/ColabFold

nature methods

BRIEF COMMUNICATION

https://doi.org/10.1038/s41592-022-01488-1

Check for updates

OPEN

ColabFold: making protein folding accessible to all

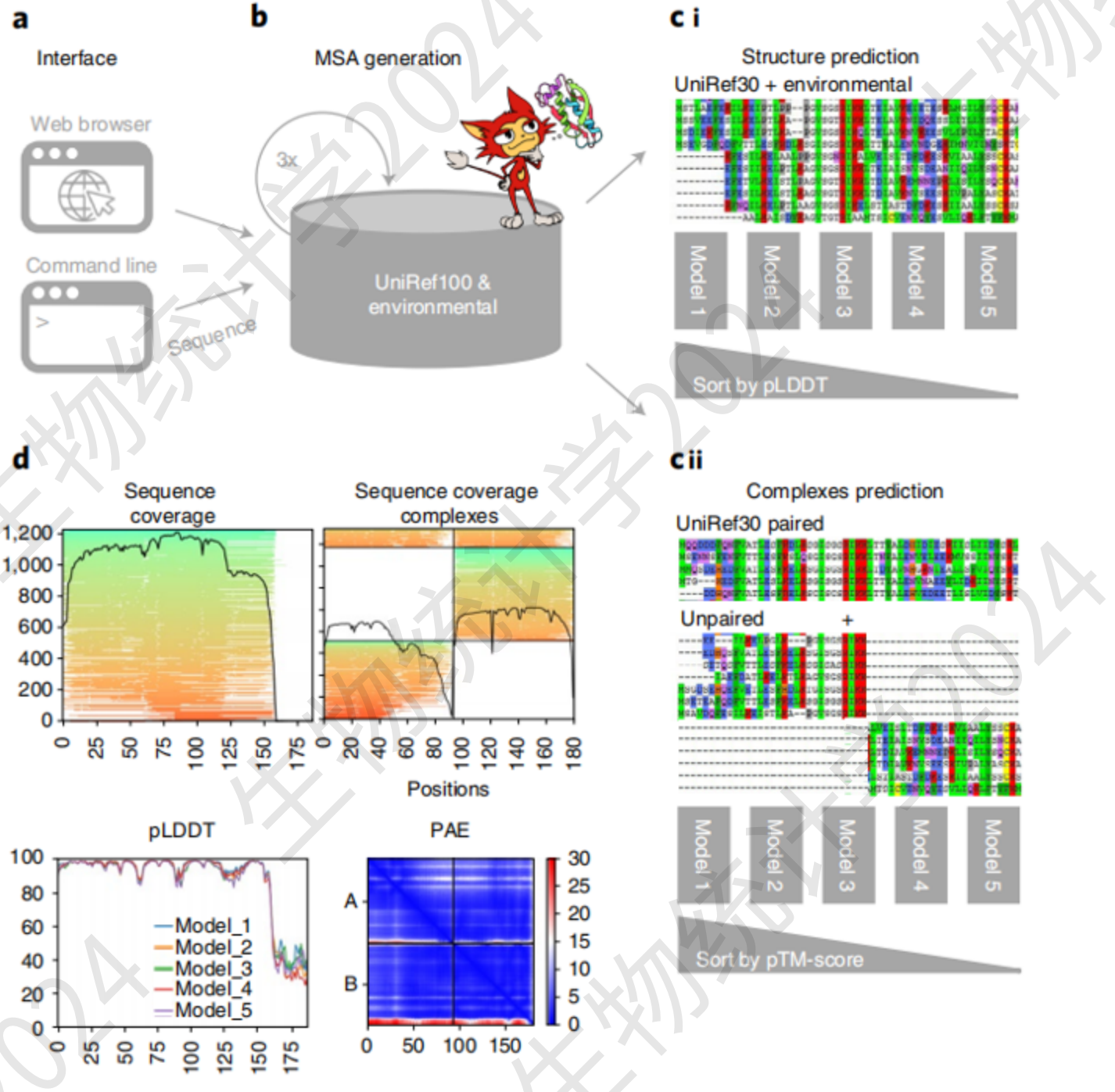
Milot Mirdita[1,10] ✉, Konstantin Schütze[2], Yoshitaka Moriwaki[3,4], Lim Heo[5], Sergey Ovchinnikov[6,7,10] ✉ and Martin Steinegger[2,8,9,10] ✉

ColabFold offers accelerated prediction of protein structures and complexes by combining the fast homology search of MMseqs2 with AlphaFold2 or RoseTTAFold. ColabFold's 40–60-fold faster search and optimized model utilization enables prediction of close to 1,000 structures per day on a server with one graphics processing unit. Coupled with Google Colaboratory, ColabFold becomes a free and accessible platform for protein folding. ColabFold is open-source software available at https://github.com/sokrypton/ColabFold and its novel environmental databases are available at https://colabfold.mmseqs.com.

protein sizes of ~1,000 residues. For these, however, the MSA generation dominates the overall run time.

To enable researchers without these resources to use AlphaFold2, independent solutions based on Google Colaboratory were developed. Colaboratory is a proprietary version of Jupyter Notebook hosted by Google. It is accessible for free to logged-in users and includes access to powerful GPUs. Concurrently, Tunyasuvunakool et al.[9] developed an AlphaFold2 Jupyter Notebook for Google Colaboratory (referred to as AlphaFold-Colab), for which the input MSA is built by searching with HMMer against the UniProt Reference Clusters (UniRef90) and an eightfold-reduced environ-

https://github.com/sokrypton/ColabFold


a
Interface
Web browser
Command line
Sequence
b
MSA generation
3x
UniRef100 &
environmental
c i
Structure prediction
UniRef30 + environmental
Model 1
Model 2
Model 3
Model 4
Model 5
Sort by pLDDT
c ii
Complexes prediction
UniRef30 paired
Unpaired
+
Model 1
Model 2
Model 3
Model 4
Model 5
Sort by pTM-score
d
Sequence
coverage
Sequence coverage
complexes
1,200
1,000
800
600
400
200
0
0 25 50 75 100 125 150 175
0 20 40 60 80 100 120 140 160 180
Positions
pLDDT
100
80
60
40
20
0
Model_1
Model_2
Model_3
Model_4
Model_5
0 25 50 75 100 125 150 175
PAE
A
B
0 50 100 150
30
25
20
15
10
5
0

AI + 蛋白质3D结构

Google DeepMind

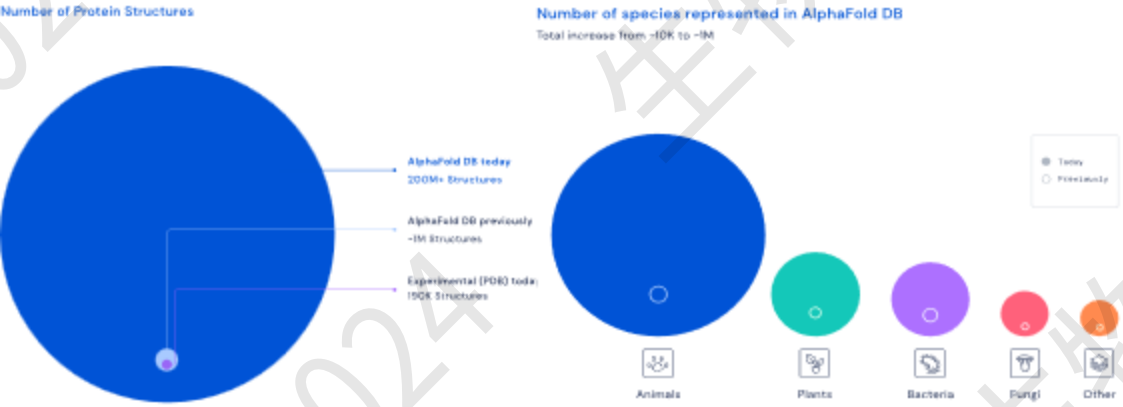
AlphaFold2
200M+

Meta

ESMFold
600M+

Tencent 腾讯

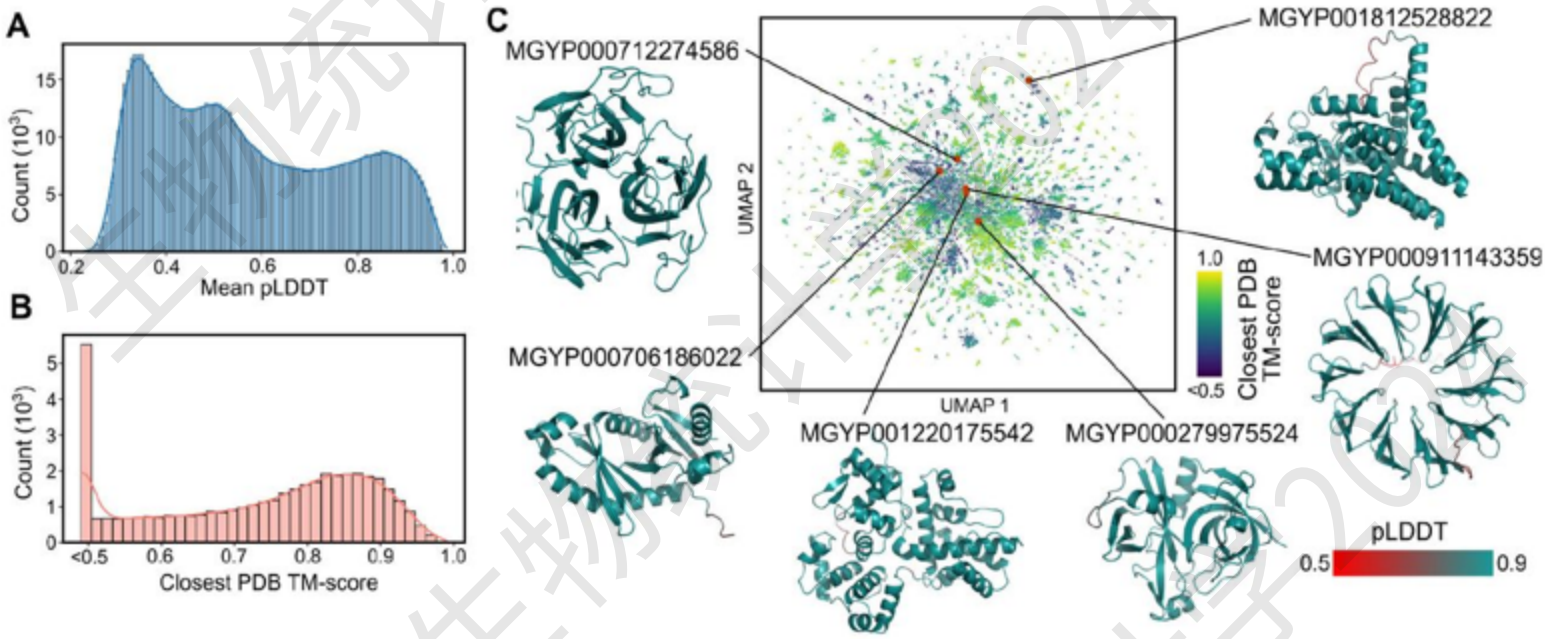
tFold
“SRD5A2”

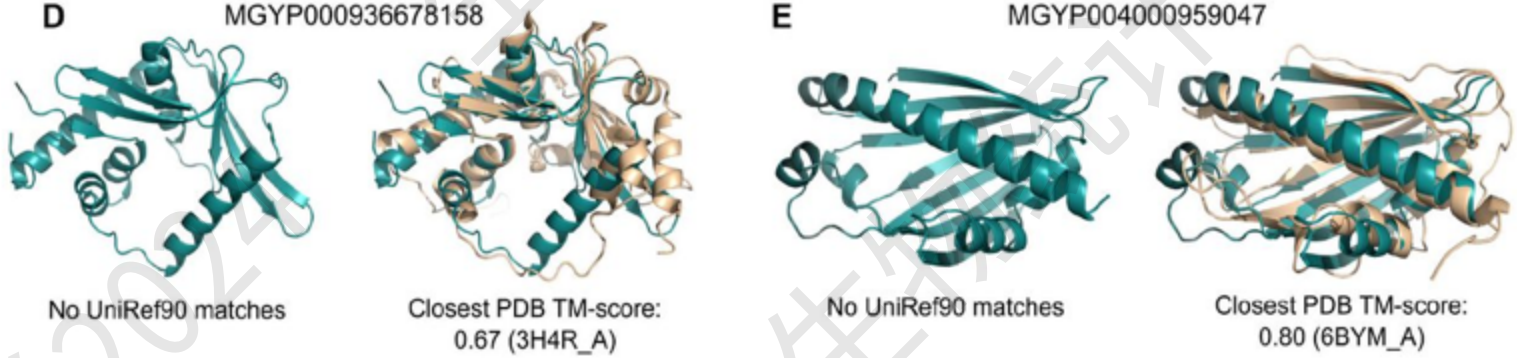


AI + 蛋白质3D结构

Meta

ESMFold
600M+

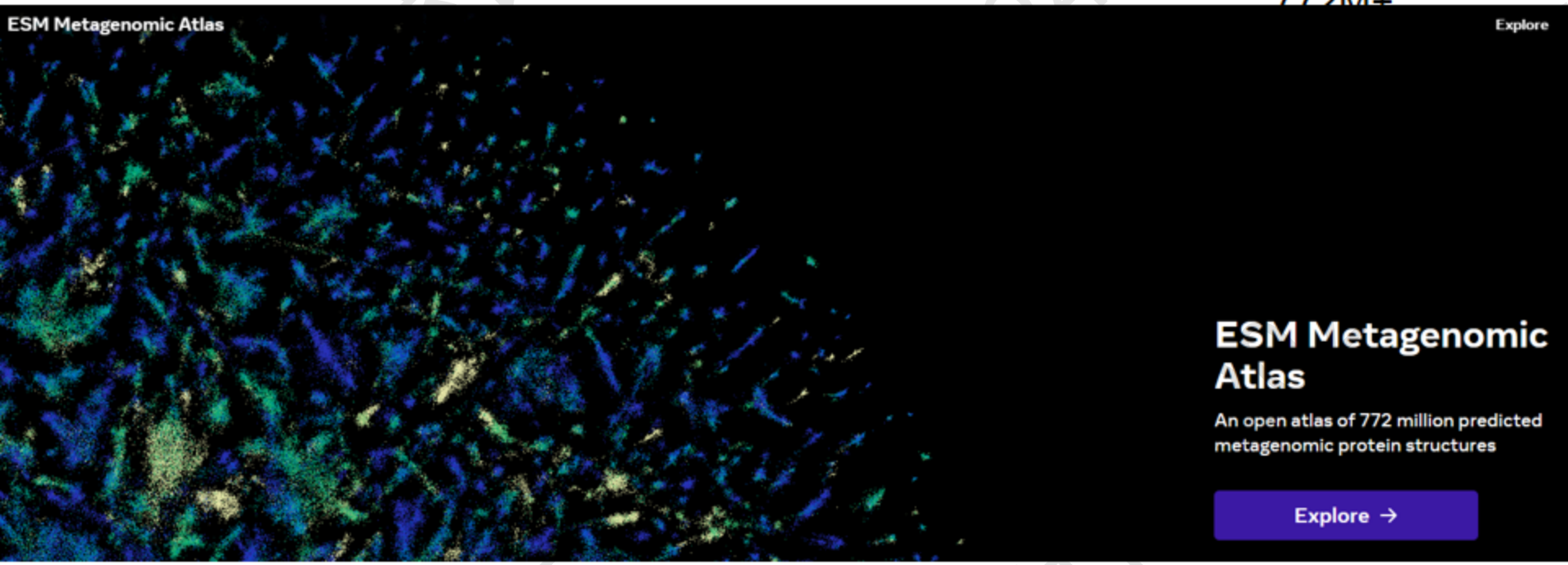




AI + 蛋白质3D结构

Meta

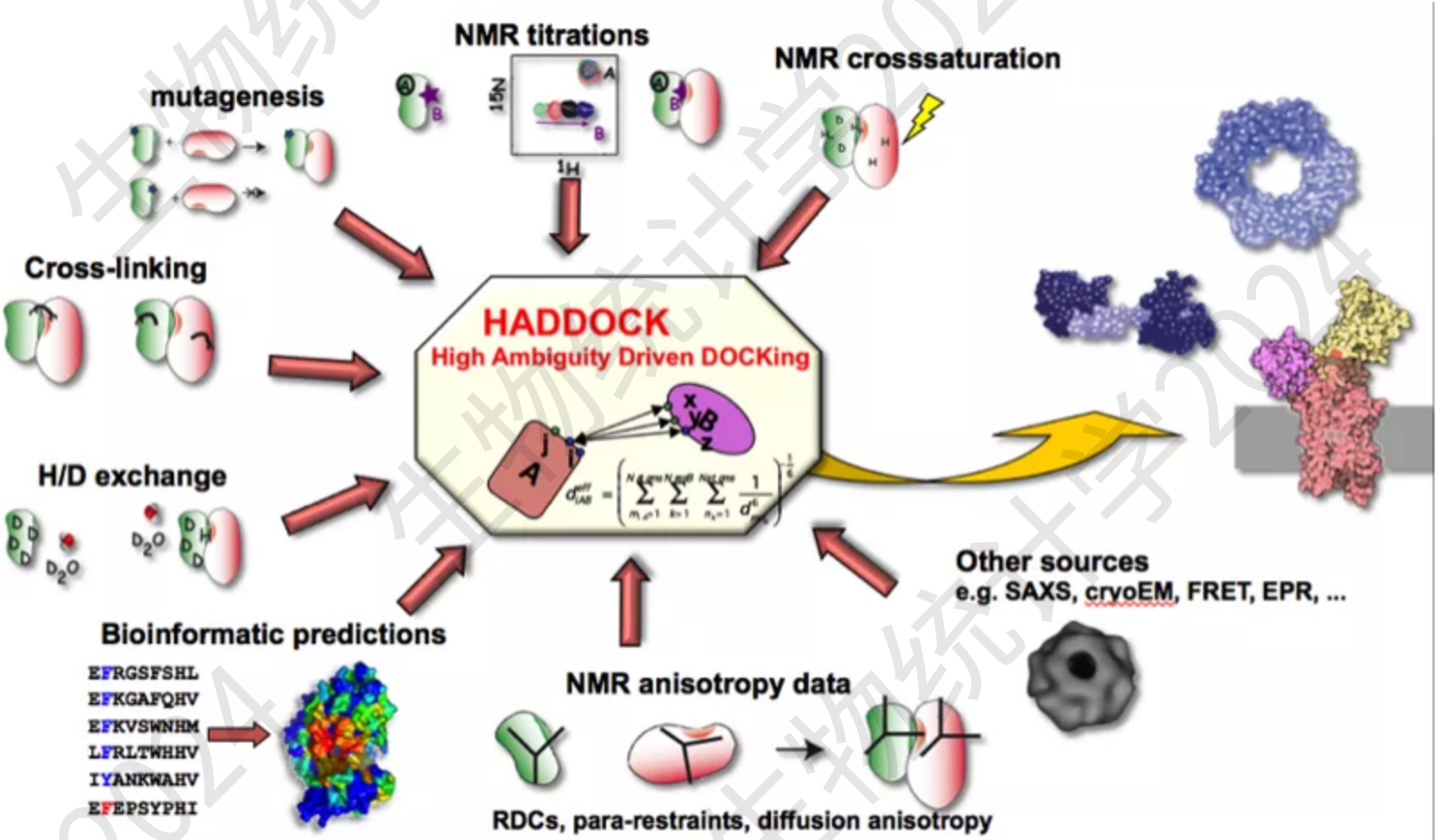
ESMFold
2
772M+



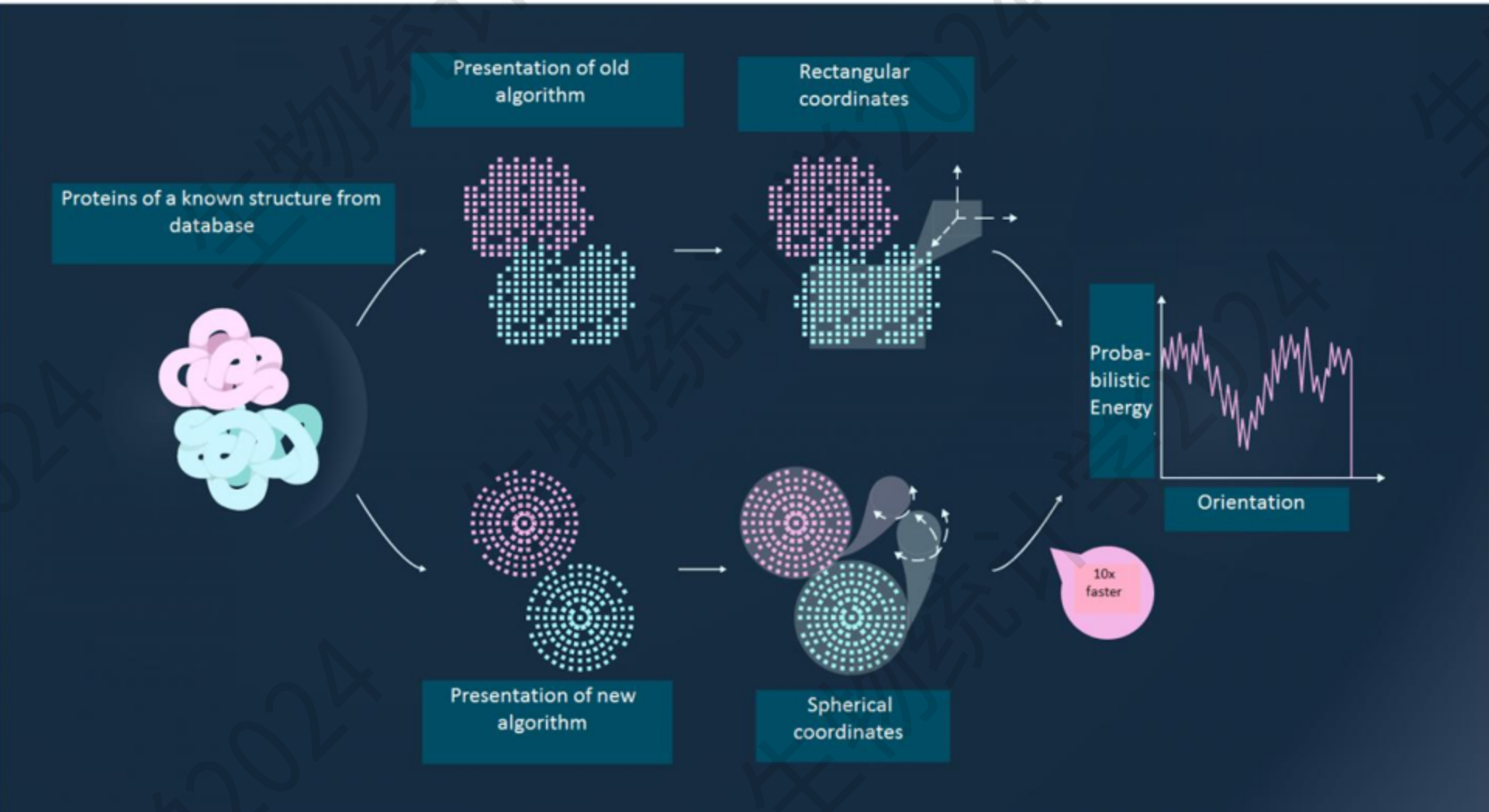
https://esmatlas.com/

CAPRI

(Critical Assessment of PRediction of Interactions)



CAPRI


Presentation of old algorithm
Rectangular coordinates
Proteins of a known structure from database
Proba-bilistic Energy
Orientation
10x faster
Presentation of new algorithm
Spherical coordinates

Protein-portein interaction by 3D modeling

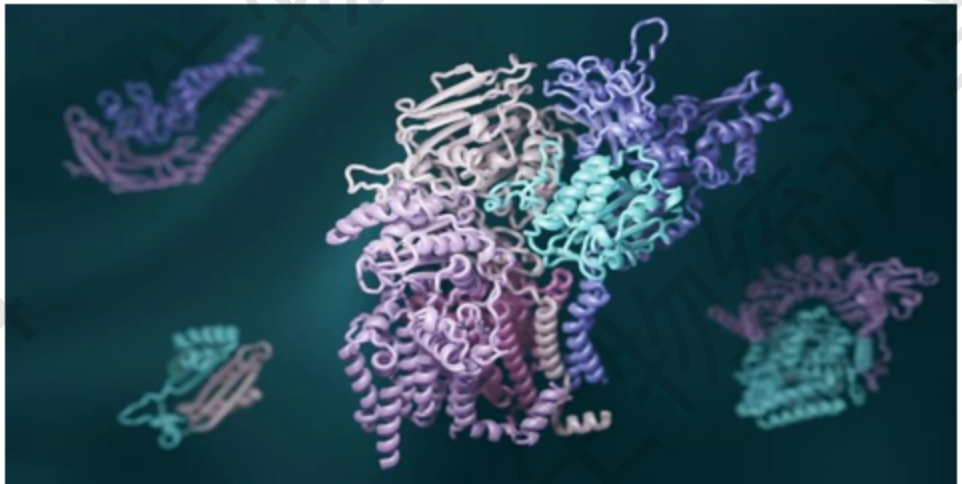
Science

RESEARCH ARTICLES

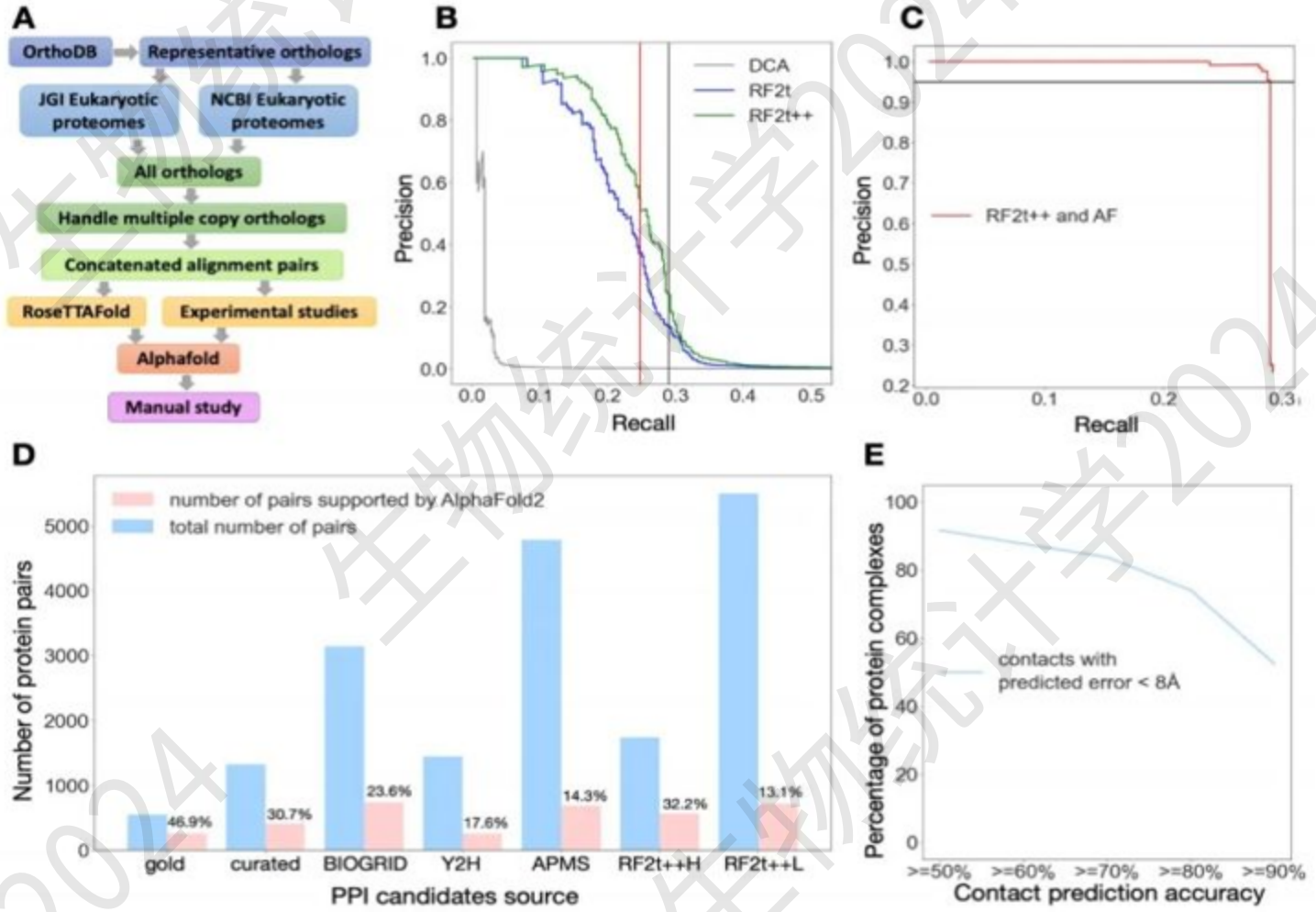
Cite as: I. R. Humphreys *et al.*, *Science*
10.1126/science.abm4805

Computed structures of core eukaryotic protein complexes

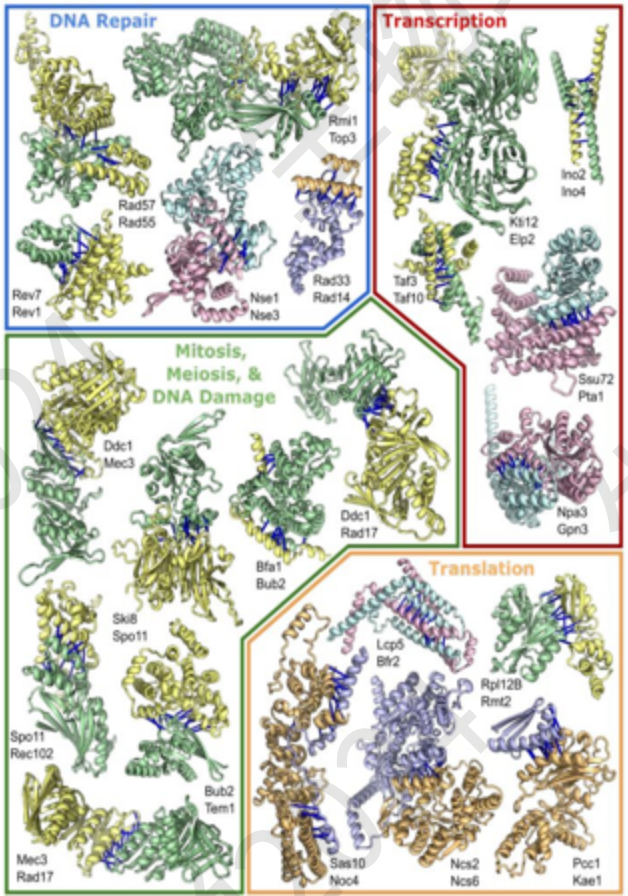
Ian R. Humphreys[1,2]†, Jimin Pei[3,4]†, Minkyung Baek[1,2]†, Aditya Krishnakumar[1,2]†, Ivan Anishchenko[1,2], Sergey Ovchinnikov[5,6], Jing Zhang[3,4], Travis J. Ness[7]‡, Sudeep Banjade[8], Saket R. Bagde[8], Viktoriya G. Stancheva[9], Xiao-Han Li[9], Kaixian Liu[10], Zhi Zheng[10,11], Daniel J. Barrero[12], Upasana Roy[13], Jochen Kuper[14], Israel S. Fernández[15], Barnabas Szakal[16], Dana Branzei[16,17], Josep Rizo[4,18,19], Caroline Kisker[14], Eric C. Greene[13], Sue Biggins[12], Scott Keeney[10,11,20], Elizabeth A. Miller[9], J. Christopher Fromme[8], Tamara L. Hendrickson[7], Qian Cong[3,4]§*, David Baker[1,2,21]§*

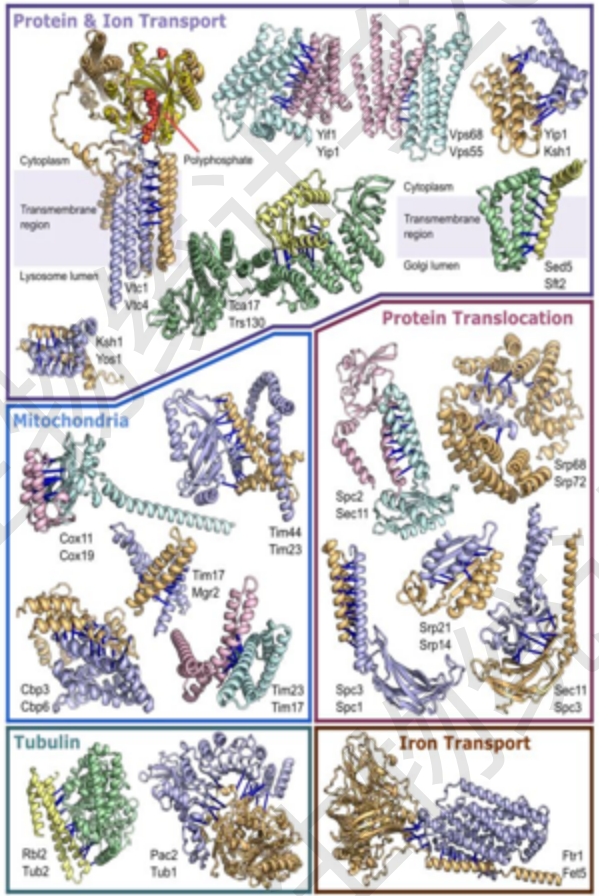


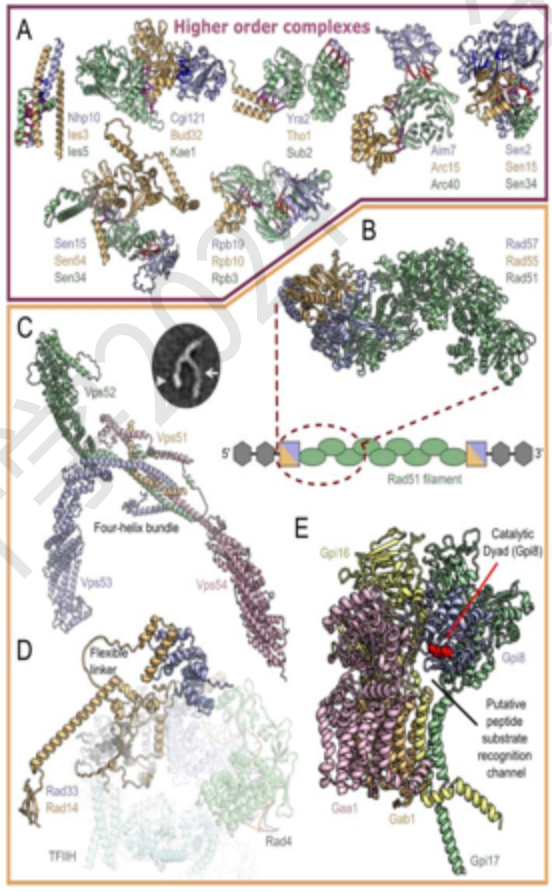
Protein-portein interaction by 3D modeling



Protein-portein interaction by 3D modeling







AI + 蛋白质3D结构

David Baker

"for computational protein design"

David Baker. Ill. Niklas Elmehed © Nobel Prize Outreach

Demis Hassabis

"for protein structure prediction"

Demis Hassabis. Ill. Niklas Elmehed © Nobel Prize Outreach

John Jumper

"for protein structure prediction"

John Jumper. Ill. Niklas Elmehed © Nobel Prize Outreach

- 2024年诺贝尔化学奖颁给人工智能模型AlphaFold，和蛋白质结构预测，是对人工智能与科学研究结合的一次里程碑式的认可。
- 今年的诺贝尔化学奖一半授予了David Baker，以表彰他在计算蛋白质设计方面的杰出贡献；另一半则共同授予了Demis Hassabis和John M. Jumper，以表彰他们在蛋白质结构预测方面的成就，特别是他们开发的AlphaFold人工智能模型。

AI + 蛋白质3D结构

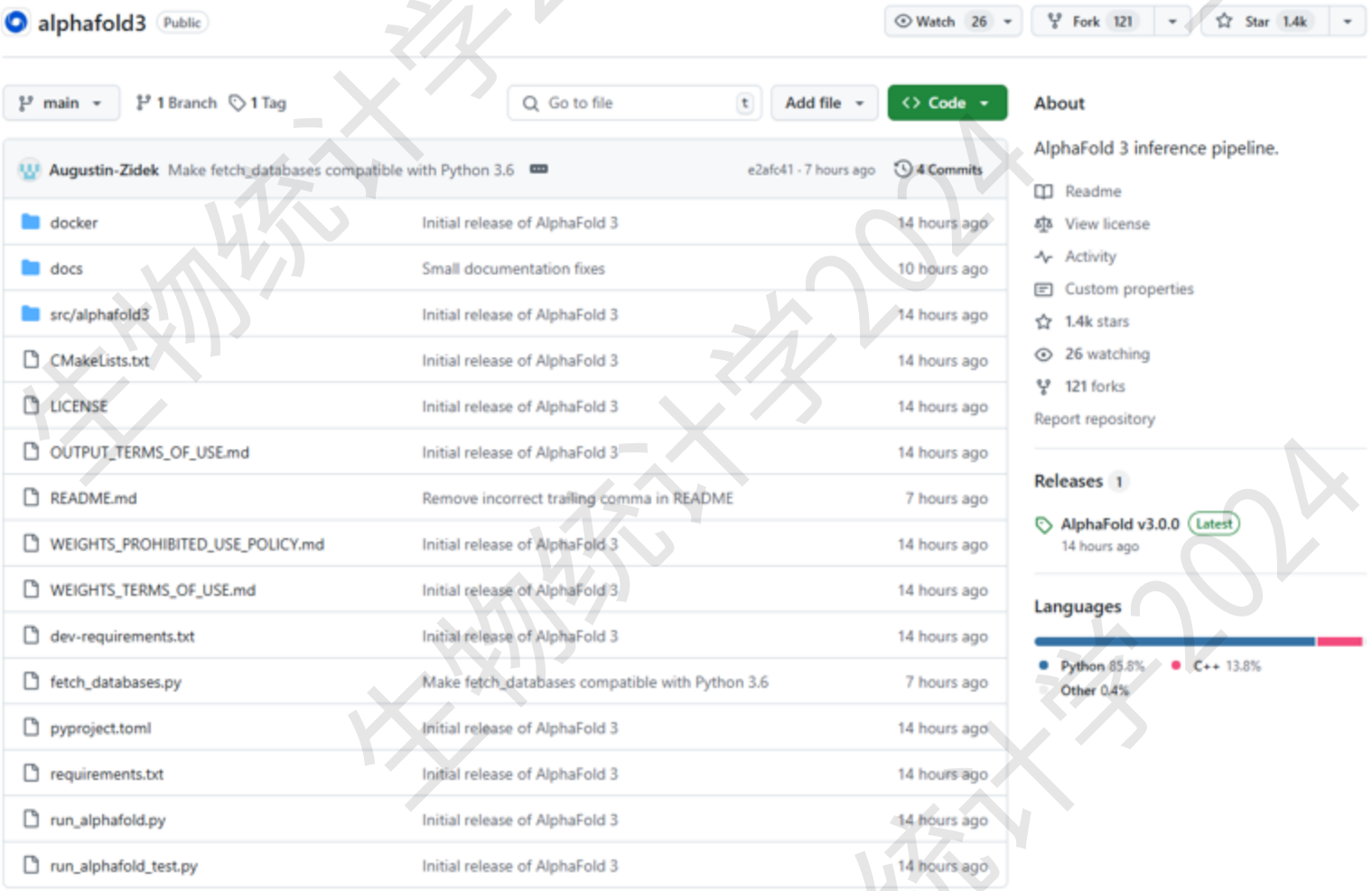
nature

Article | Published: 08 May 2024

Accurate structure prediction of biomolecular interactions with AlphaFold 3

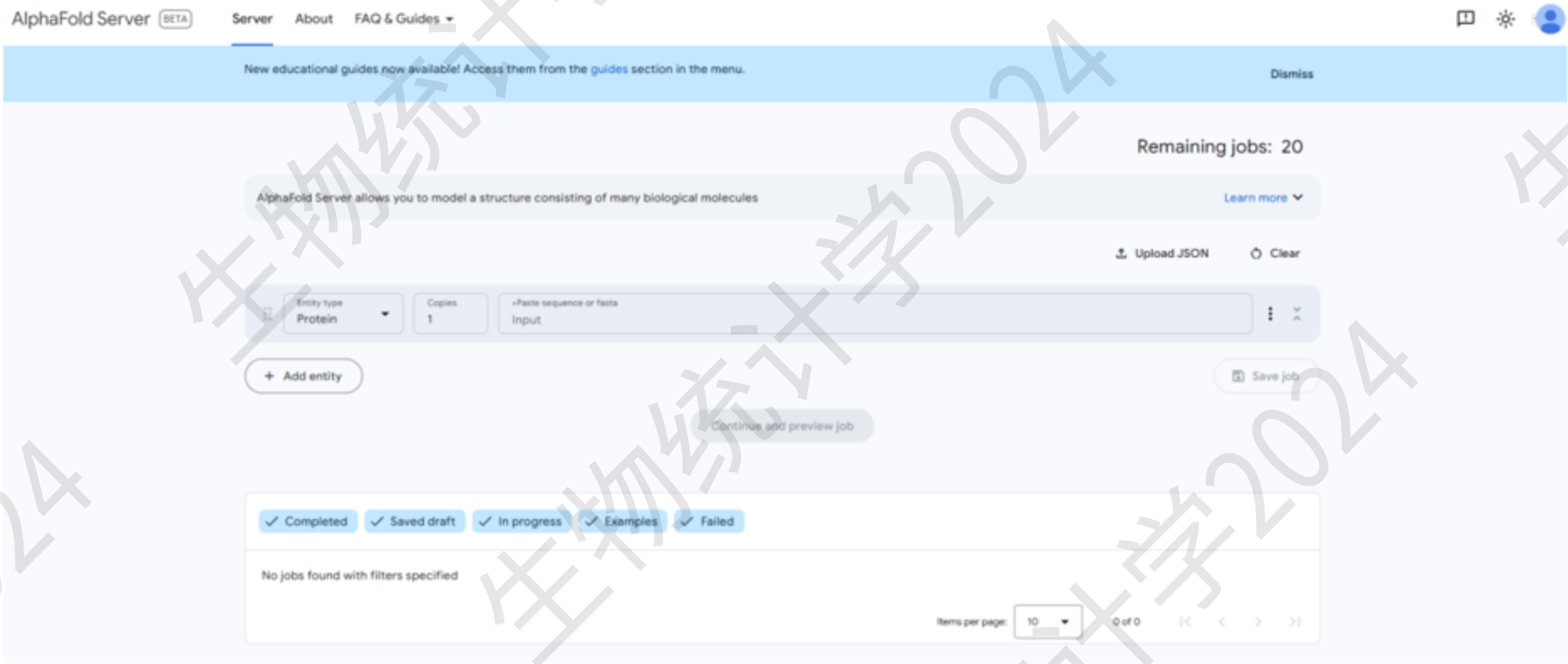
Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg ✉, Demis Hassabis ✉ & John M. Jumper ✉ — Show fewer authors

AI + 蛋白质3D结构



https://github.com/google-deepmind/alphafold3

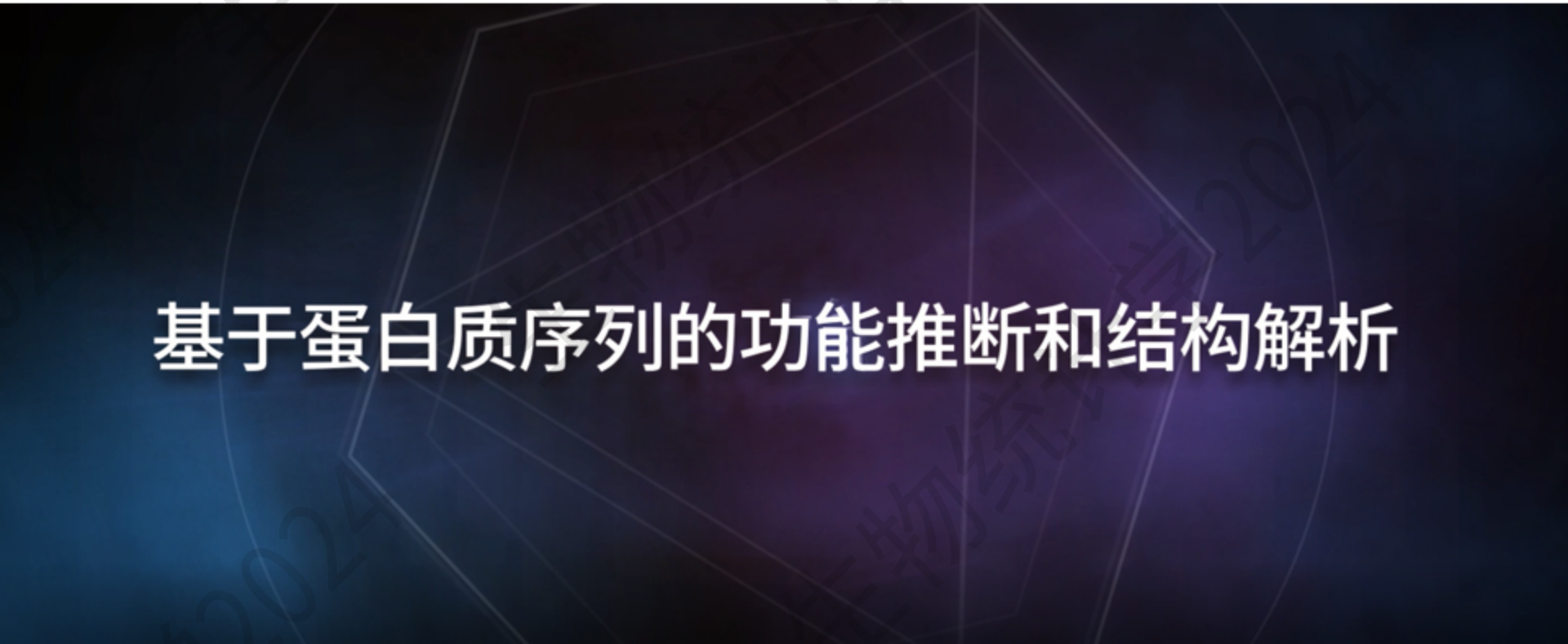
AI + 蛋白质3D结构



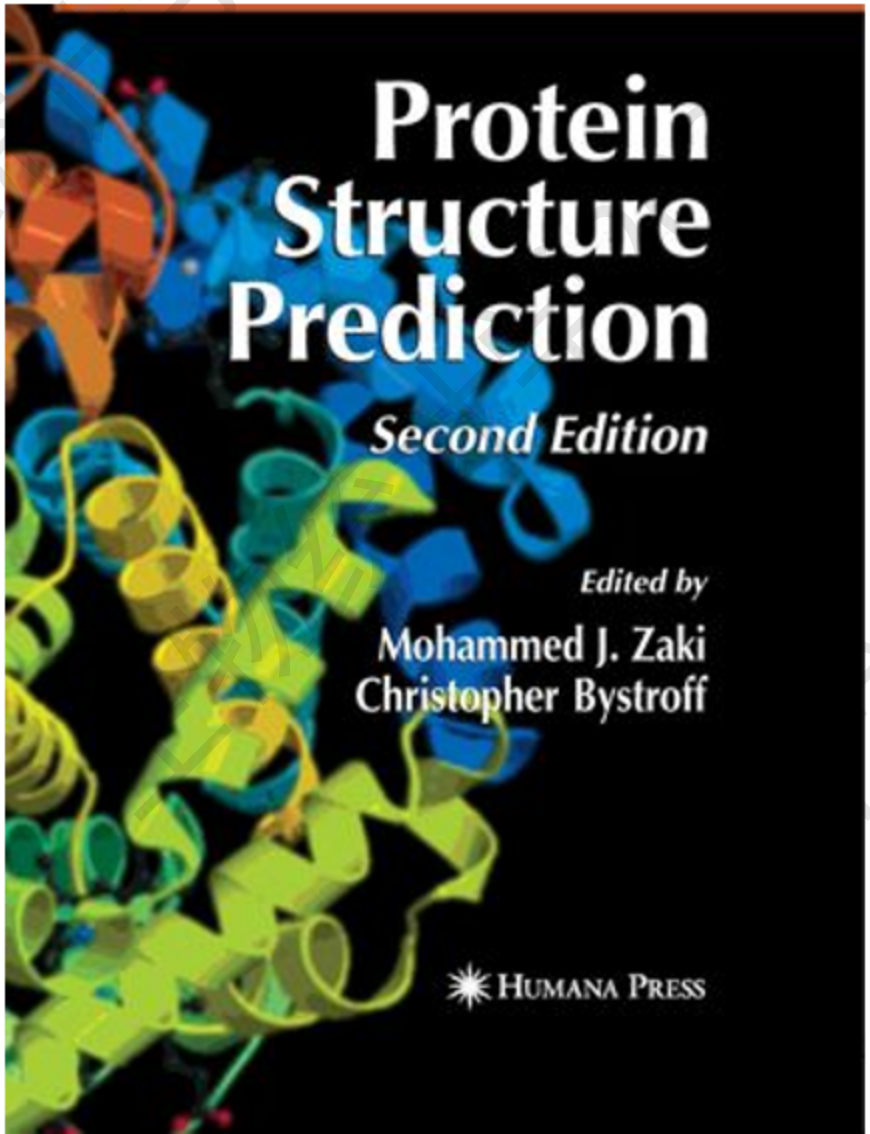
https://alphafoldserver.com/

Protein 3D structure

小教程101



References


Protein
Structure
Prediction
Second Edition
Edited by
Mohammed J. Zaki
Christopher Bystroff
HUMANA PRESS

Slides credits

——生物信息学研究方法概述：北京大学生物信息中心

——生物统计学：卜东波@中国科学院计算技术研究所，邓明华@北京大学

——神经网络与深度学习: 邱锡鹏@复旦大学

——Introduction to Computational Biology and Bioinformatics: Xiaole Shirley Liu Lab@Havard University

——Combinatorial Methods in Computation Biology: Ken Sung Lab@NUS

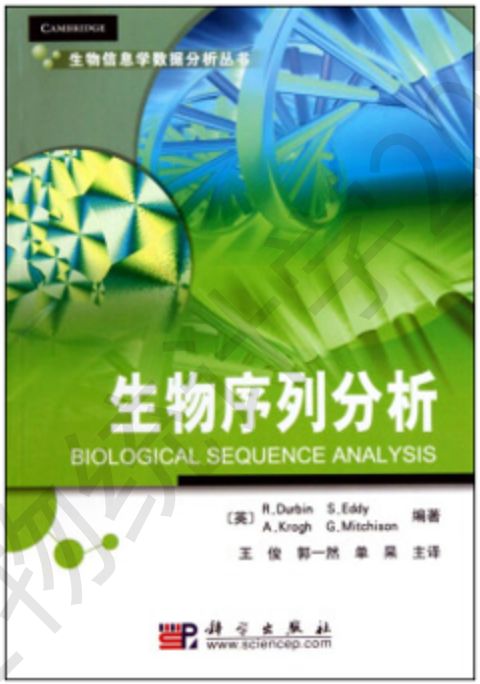
——Deep Learning in the Life Sciences: MIT

——Probabilistic Graphical Models: Eric Xing@CMU

——Numerous other leading researchers and leading labs.....

References


2016
基因组学
GENOMICS


CAMBRIDGE
生物信息学数据分析丛书
生物序列分析
BIOLOGICAL SEQUENCE ANALYSIS
R.Durbin S.Eddy
A.Krogh G.Mitchison
编著
王俊 郭一然 单梁 主译
www.sciencep.com


生物统计学
(第五版)