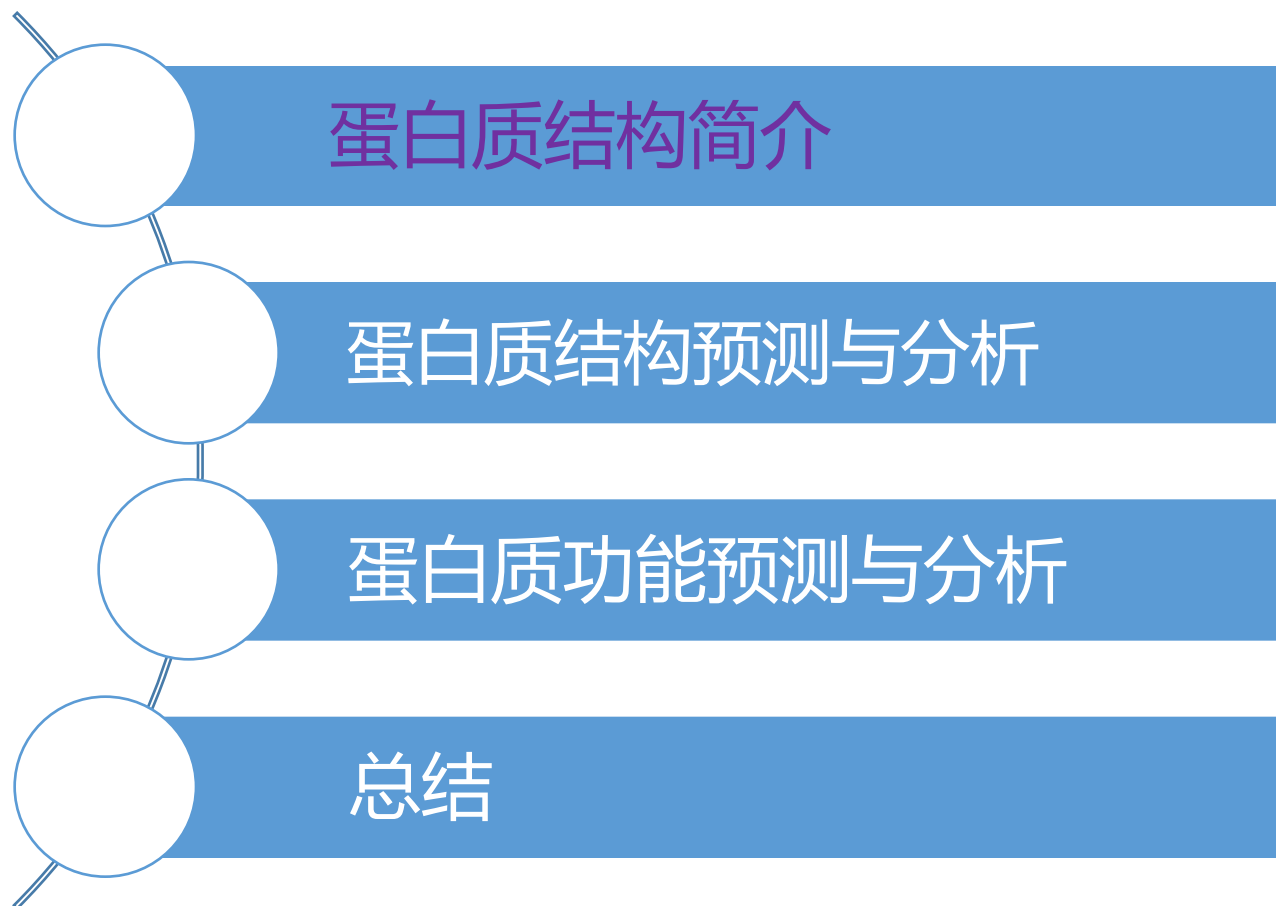
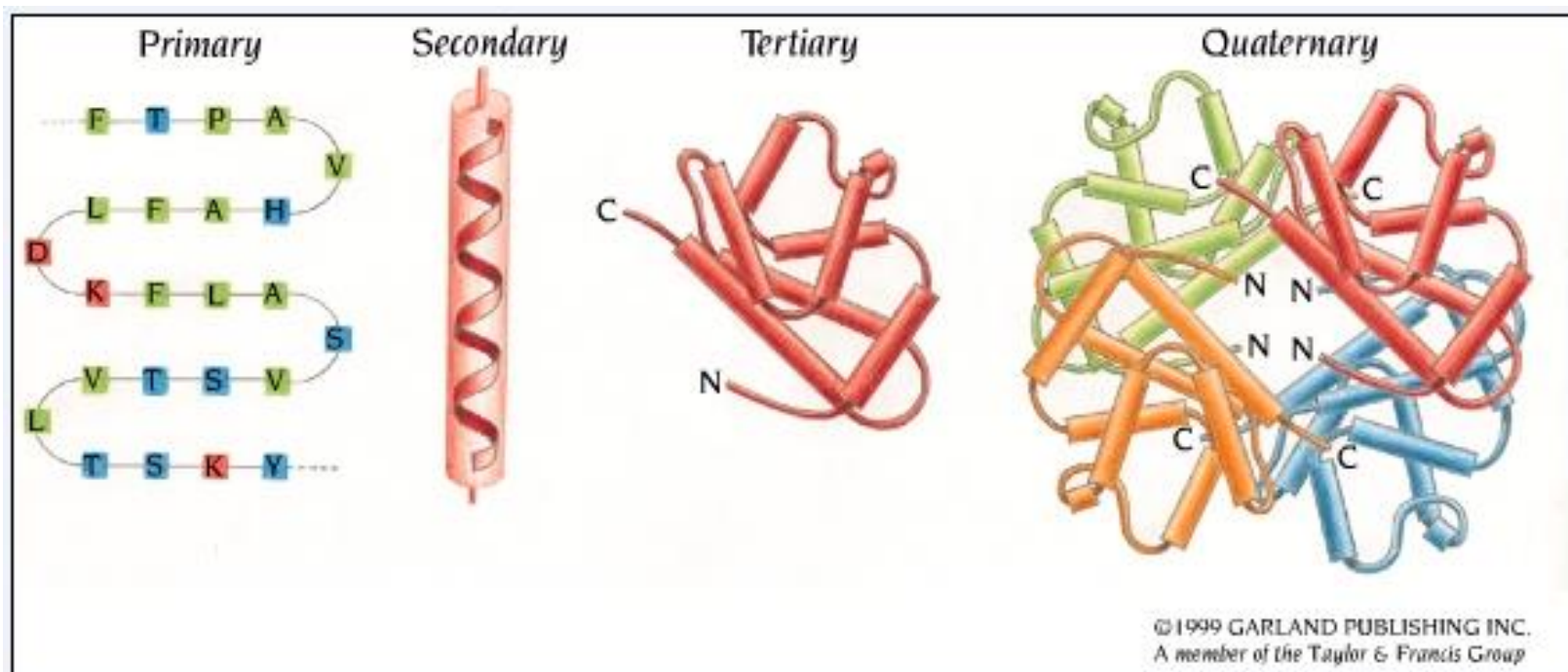


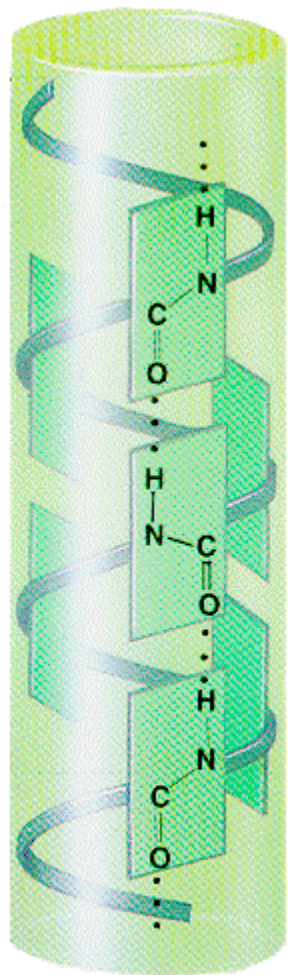
第九章 蛋白质与代谢组学 (第2&4节)

张子丁, 王明钰, 田卫东



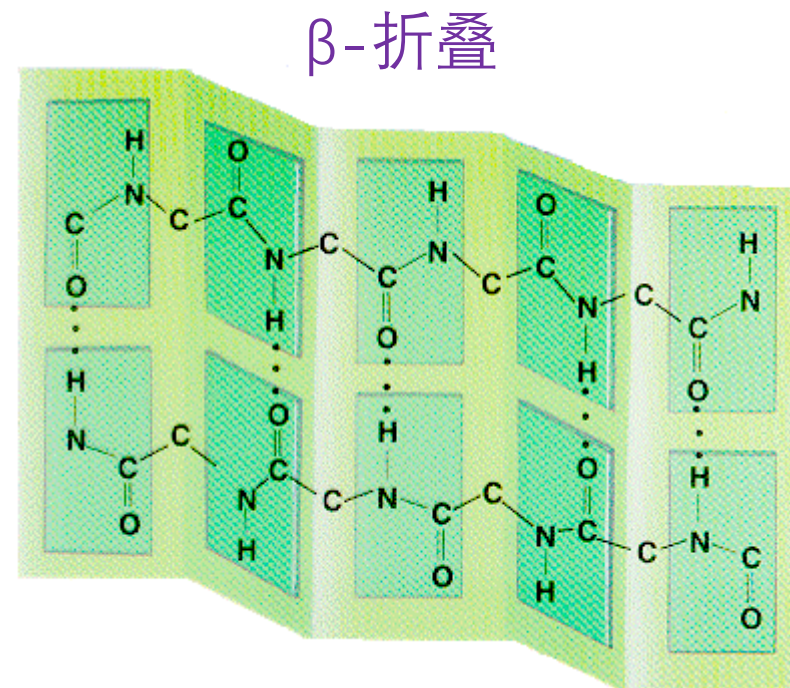
分为四个层次（一级，二级，三级和四级结构）





α -螺旋

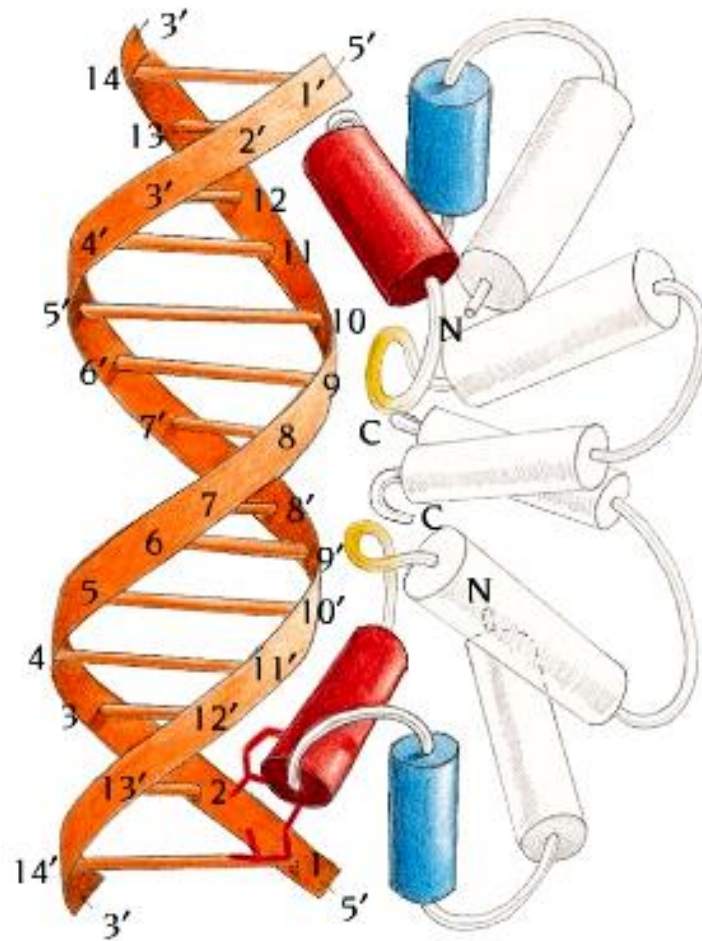
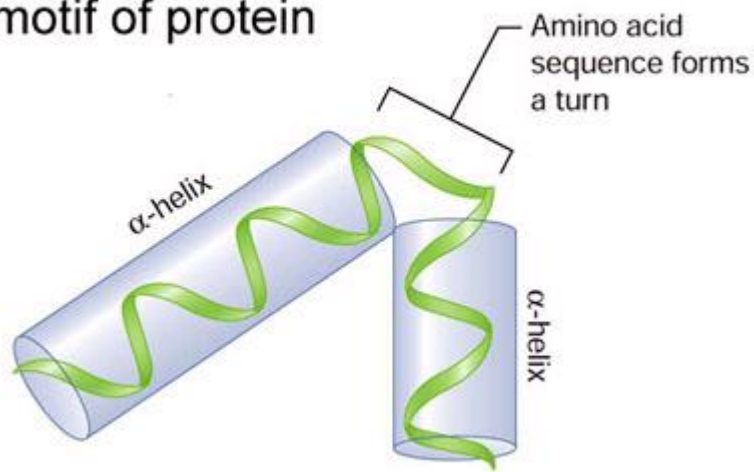
- ✓ 主链每隔4个残基形成一个氢键
- ✓ 最常见的二级结构
- ✓ 一个蛋白的平均 α -螺旋含量为30%
- ✓ 平均长度变化为5-40个残基



- ✓ 分为平行和反平行两种形式
- ✓ 一个蛋白的平均 β -折叠含量为20%
- ✓ 平均长度变化为5-10个残基
- ✓ 形成氢键的两个片段序列距离可以很远

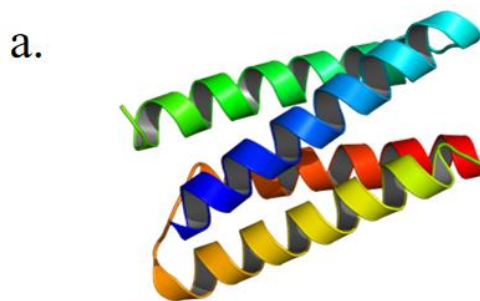
- 连接 α -螺旋和 β -折叠片的区域
- 通常位于蛋白结构的表面
- 对突变有高容忍度
- 构象具有多样性及更大柔性
- 带电荷和极性氨基酸居多
- 频繁构成活性区域
- 可以划分为一些不同的结构类型，例如转角结构

Helix-turn-helix
motif of protein

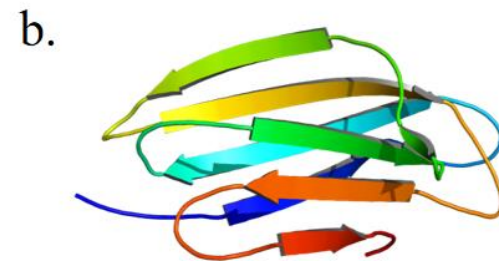


©1999 GARLAND PUBLISHING INC.
A member of the Taylor & Francis Group

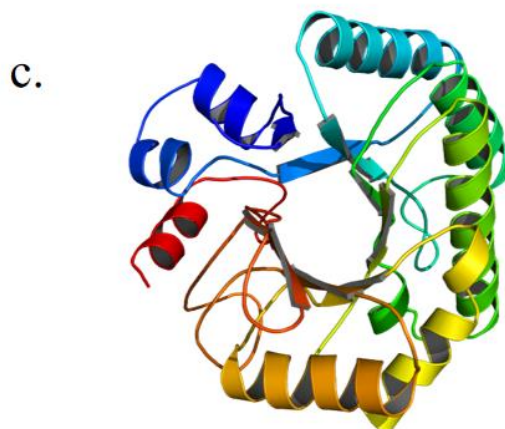
- ✓ 蛋白质结构域可分为四种主要类型： α 、 β 、 α/β 和 $\alpha+\beta$ 结构域。
- ✓ α 结构域的核心区主要由 α 螺旋构成。
- ✓ β 结构域的核心区主要由 β 片层构成。
- ✓ α/β 和 $\alpha+\beta$ 结构域的核心区既含有 α 螺旋也含有 β 片层，其中 α/β 型主要含有平行 β 片层， $\alpha+\beta$ 型主要含有反平行 β 片层。



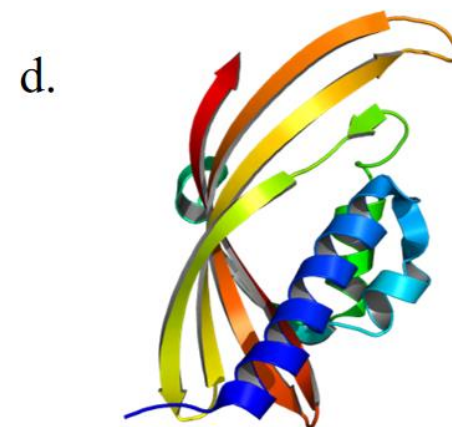
1EC5



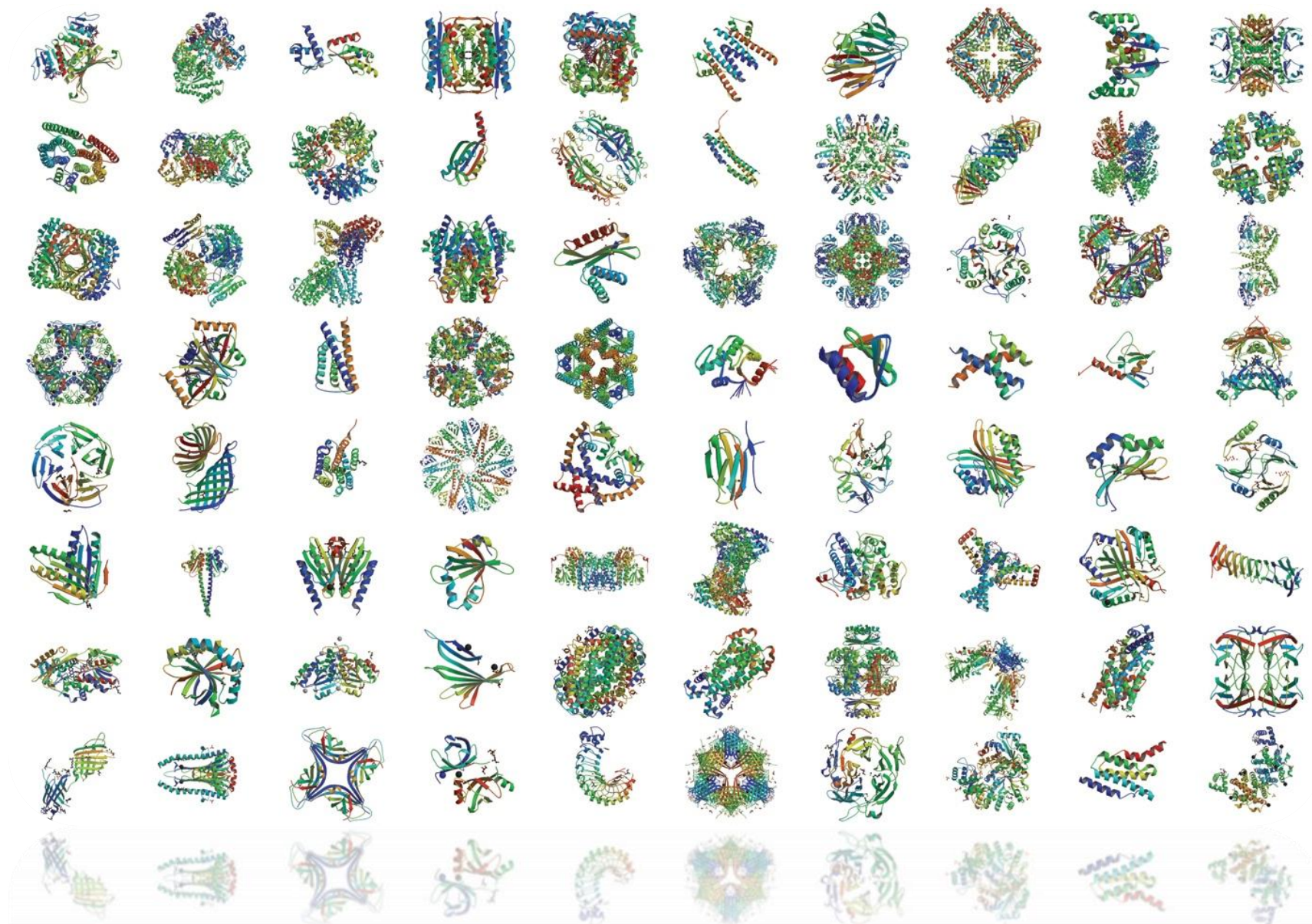
1LK2_B



1TML



1IDP_A



➤ 蛋白质结构实验测定方法

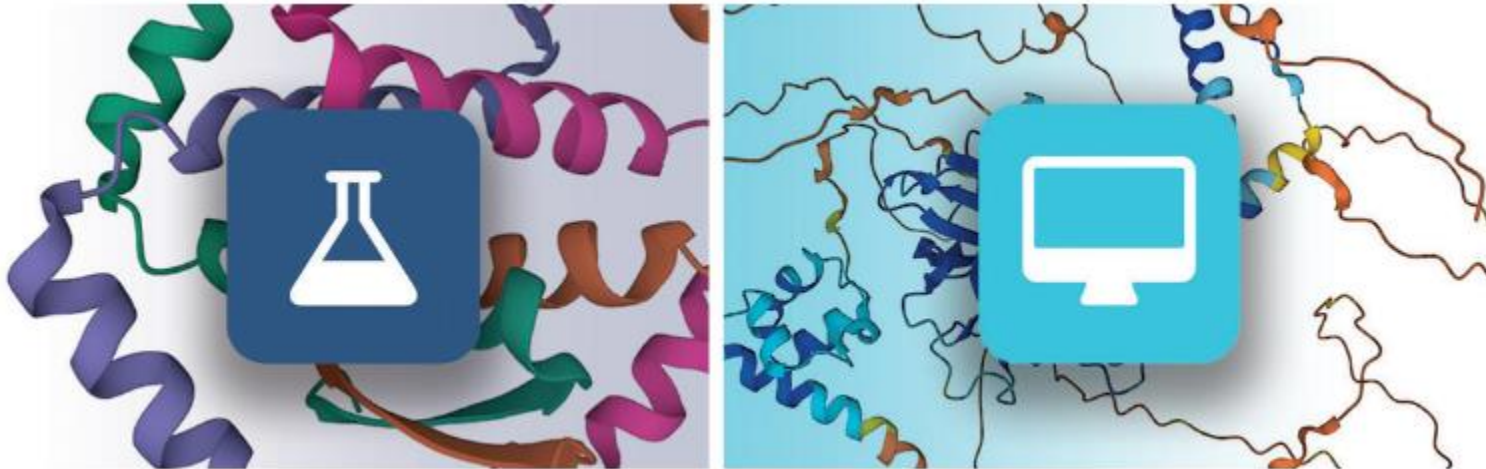
- X-射线衍射法
- 核磁共振技术(NMR)
- 低温冷冻电镜

➤ 实验获得的蛋白质序列数目与结构数目的差距很大

- 序列远超过2亿
- 结构仅超过22万，而且包含很多结构被重复测定的蛋白

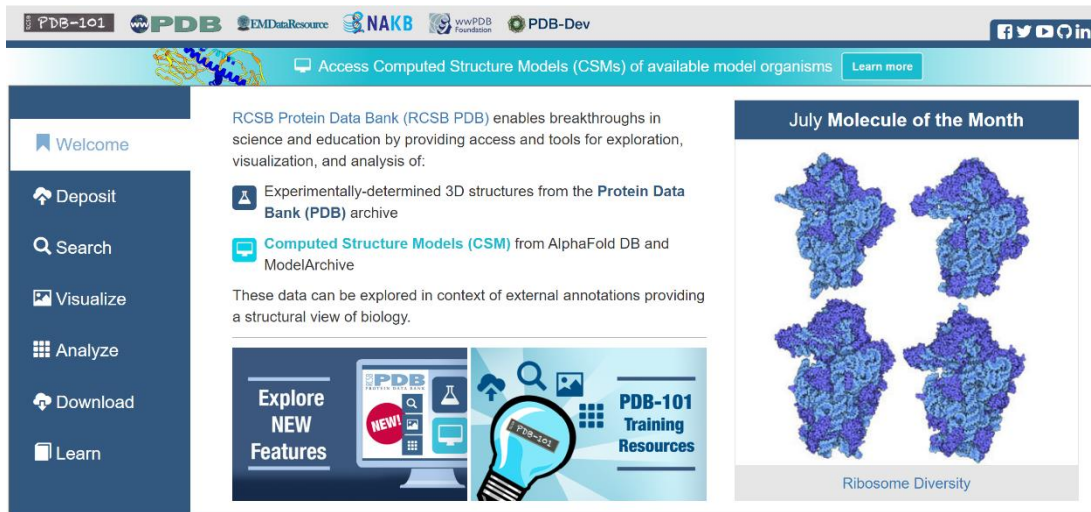
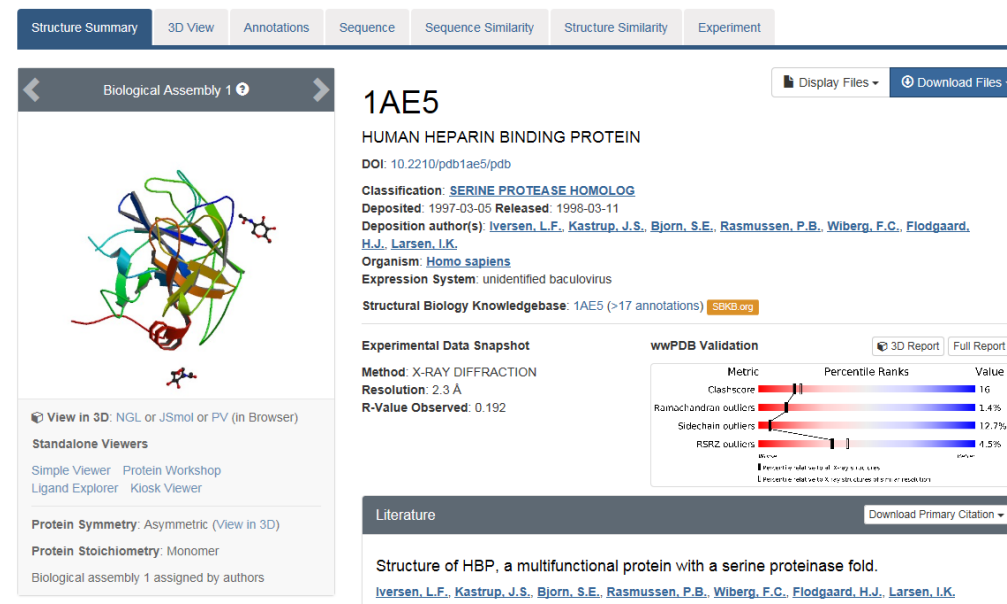
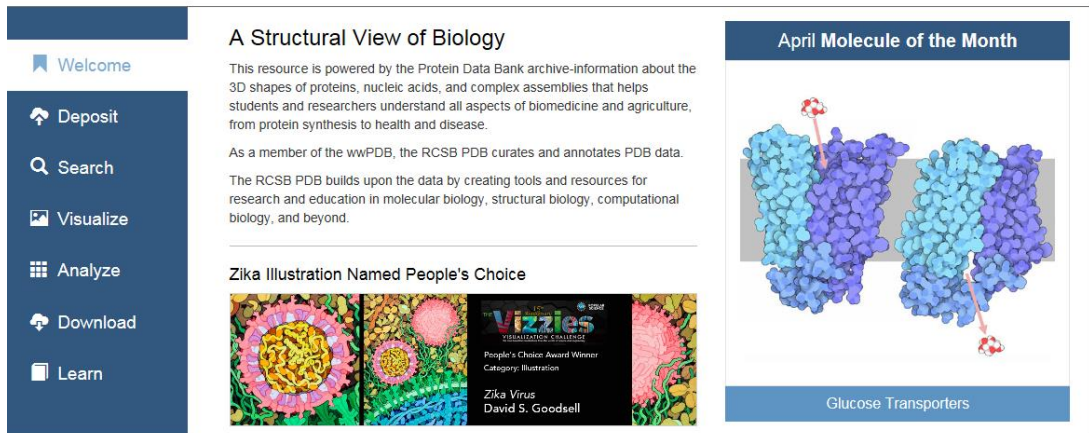


- www.rcsb.org
- 3D Macromolecular structural data
- Total n° of experimental structures **222,036 (2024/7/4)**
- Offers access to ~1 million Computed Structure Models (CSMs) from AlphaFoldDB and RoseTTAFold





e.g. 1AE5



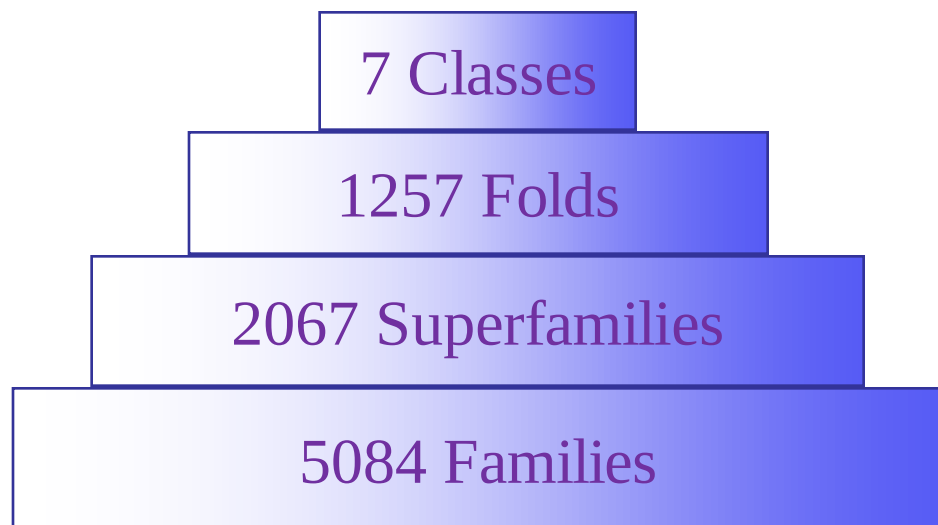
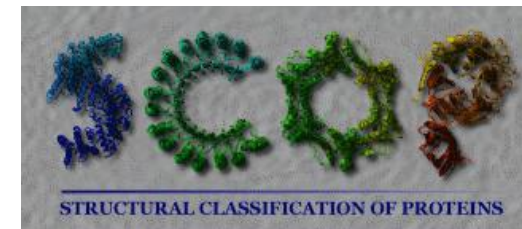
PDB数据库还含有许多重要的资源，已成为学习蛋白结构知识的重要社区。

PDB文件格式

ATOM	1	N	ILE	1	11.912	28.495	38.531
ATOM	2	CA	ILE	1	13.047	27.997	39.280
ATOM	3	C	ILE	1	13.469	28.944	40.382
ATOM	4	O	ILE	1	13.846	30.074	40.104
ATOM	5	CB	ILE	1	14.289	27.733	38.411
ATOM	6	CG1	ILE	1	14.063	26.608	37.402
ATOM	7	CG2	ILE	1	15.473	27.376	39.306
ATOM	8	CD1	ILE	1	14.104	25.229	38.071
ATOM	9	N	VAL	2	13.489	28.452	41.634
ATOM	10	CA	VAL	2	13.948	29.265	42.735
ATOM	11	C	VAL	2	15.380	28.972	43.175
ATOM	12	O	VAL	2	15.848	27.832	43.180
ATOM	13	CB	VAL	2	13.014	29.443	43.930
ATOM	14	CG1	VAL	2	11.764	28.582	43.936
ATOM	15	CG2	VAL	2	13.811	29.355	45.210
ATOM	16	N	GLY	3	16.095	30.037	43.539
ATOM	17	CA	GLY	3	17.454	29.903	44.026
ATOM	18	C	GLY	3	18.470	29.585	42.957
ATOM	19	O	GLY	3	19.588	29.155	43.254
ATOM	20	N	GLY	4	18.109	29.821	41.709
ATOM	21	CA	GLY	4	19.037	29.505	40.650
ATOM	22	C	GLY	4	19.767	30.689	40.023
ATOM	23	O	GLY	4	19.698	31.859	40.449
ATOM	24	N	ARG	5	20.457	30.311	38.976
ATOM	25	CA	ARG	5	21.224	31.156	38.162
ATOM	26	C	ARG	5	20.672	31.130	36.734
ATOM	27	O	ARG	5	20.023	30.202	36.256
ATOM	28	CB	ARG	5	22.631	30.628	38.231
ATOM	29	CG	ARG	5	23.640	31.657	37.840
ATOM	30	CD	ARG	5	24.935	30.980	37.455
ATOM	31	NE	ARG	5	25.749	30.674	38.619
ATOM	32	CZ	ARG	5	26.696	29.739	38.632
ATOM	33	NH1	ARG	5	26.989	28.973	37.565
ATOM	34	NH2	ARG	5	27.380	29.575	39.762
ATOM	35	N	LYS	6	20.931	32.187	36.040
ATOM	36	CA	LYS	6	20.505	32.318	34.687
ATOM	37	C	LYS	6	21.302	31.401	33.774
ATOM	38	O	LYS	6	22.525	31.377	33.820
ATOM	39	CB	LYS	6	20.798	33.749	34.258
ATOM	40	CG	LYS	6	19.973	34.230	33.081
ATOM	41	CD	LYS	6	20.789	35.115	32.168

单位是Å

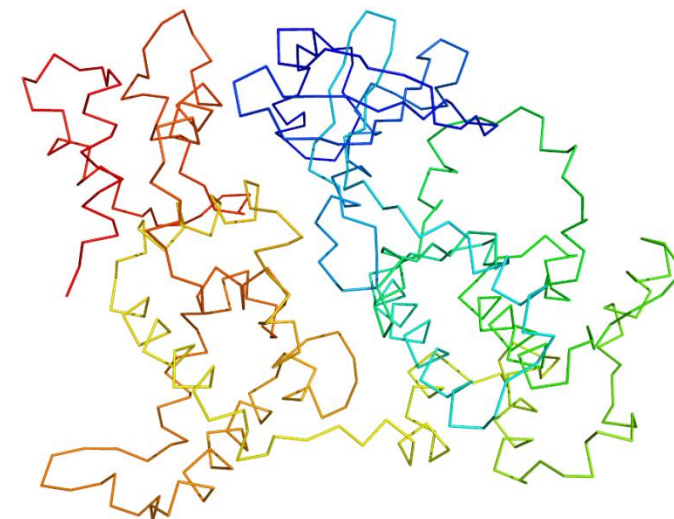
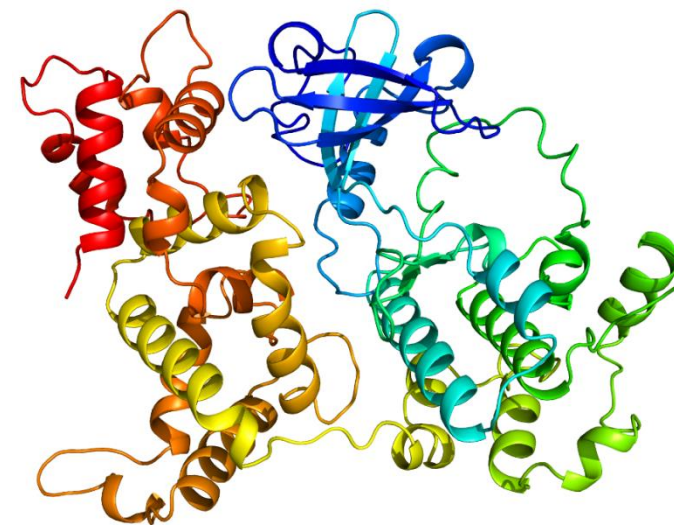
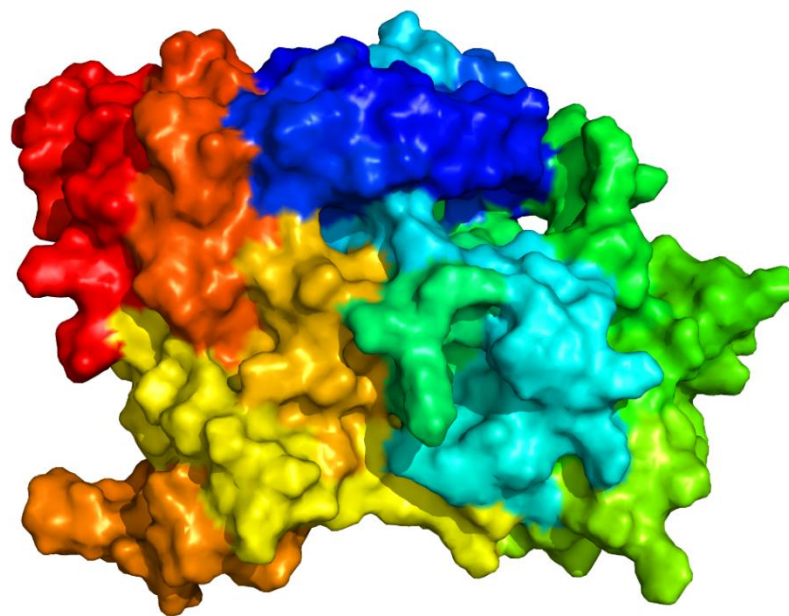
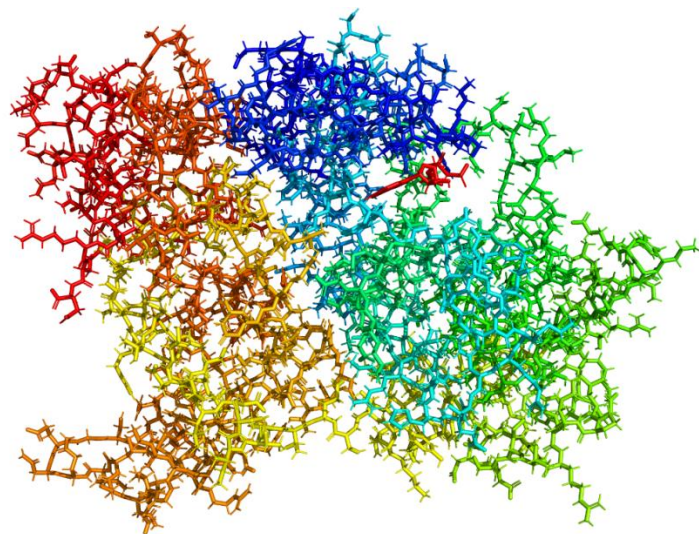
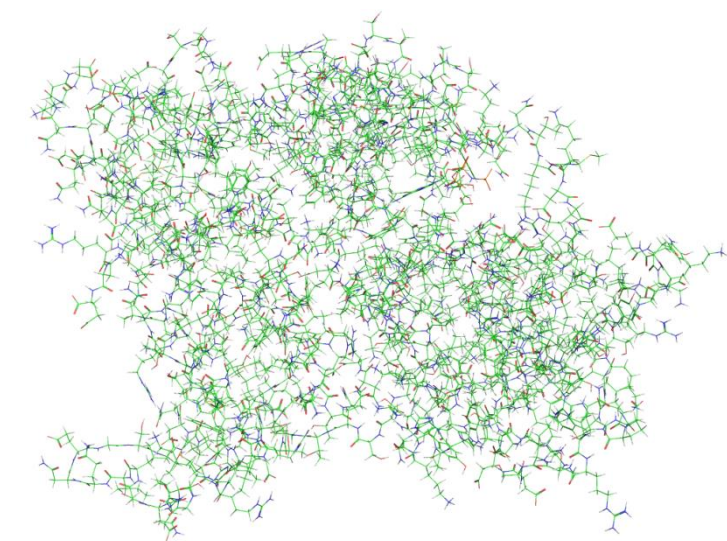
- SCOP (Structural Classification Of Proteins)
- 拓展版SCOPe (<http://scop.berkeley.edu>)
- 目前版本e2.08, 包含107643 个蛋白结构, 346905个结构域。



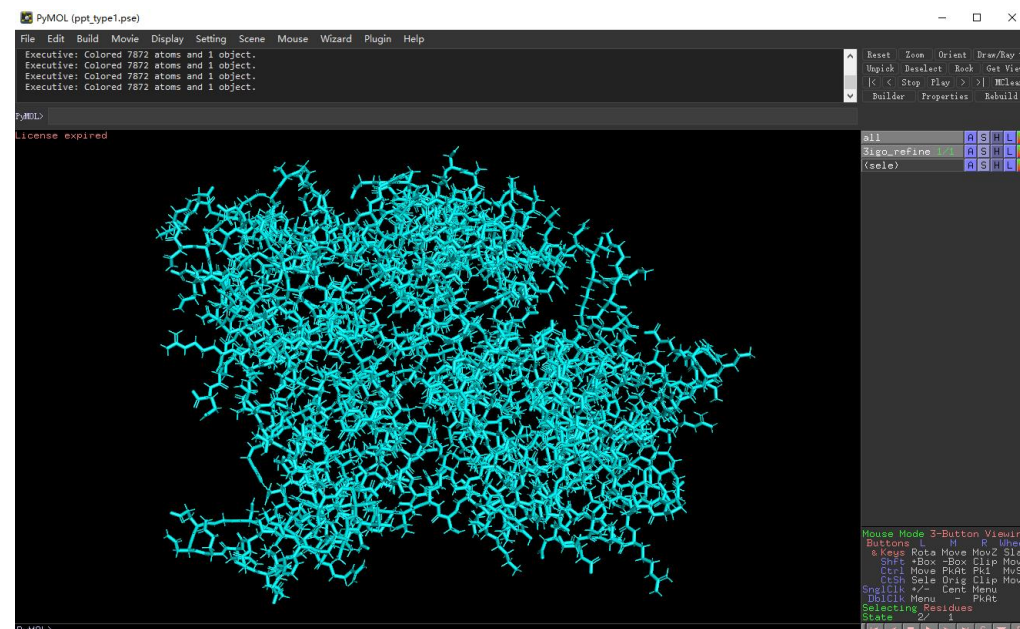
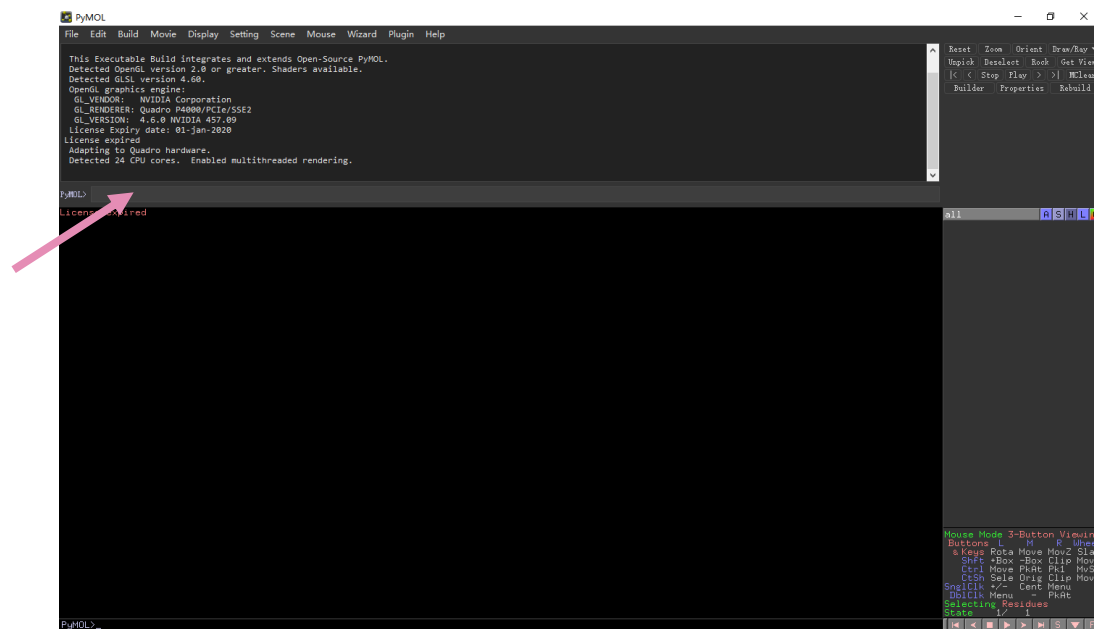
Classes in SCOPe 2.08:

1. a: All alpha proteins [46456] (290 folds)
2. b: All beta proteins [48724] (180 folds)
3. c: Alpha and beta proteins (a/b) [51349] (148 folds)
4. d: Alpha and beta proteins (a+b) [53931] (396 folds)
5. e: Multi-domain proteins (alpha and beta) [56572] (74 folds)
6. f: Membrane and cell surface proteins and peptides [56835] (69 folds)
7. g: Small proteins [56992] (100 folds)
8. h: Coiled coil proteins [57942] (7 folds)
9. i: Low resolution protein structures [58117] (25 folds)
10. j: Peptides [58231] (151 folds)
11. k: Designed proteins [58788] (44 folds)
12. l: Artifacts [310555] (1 fold)

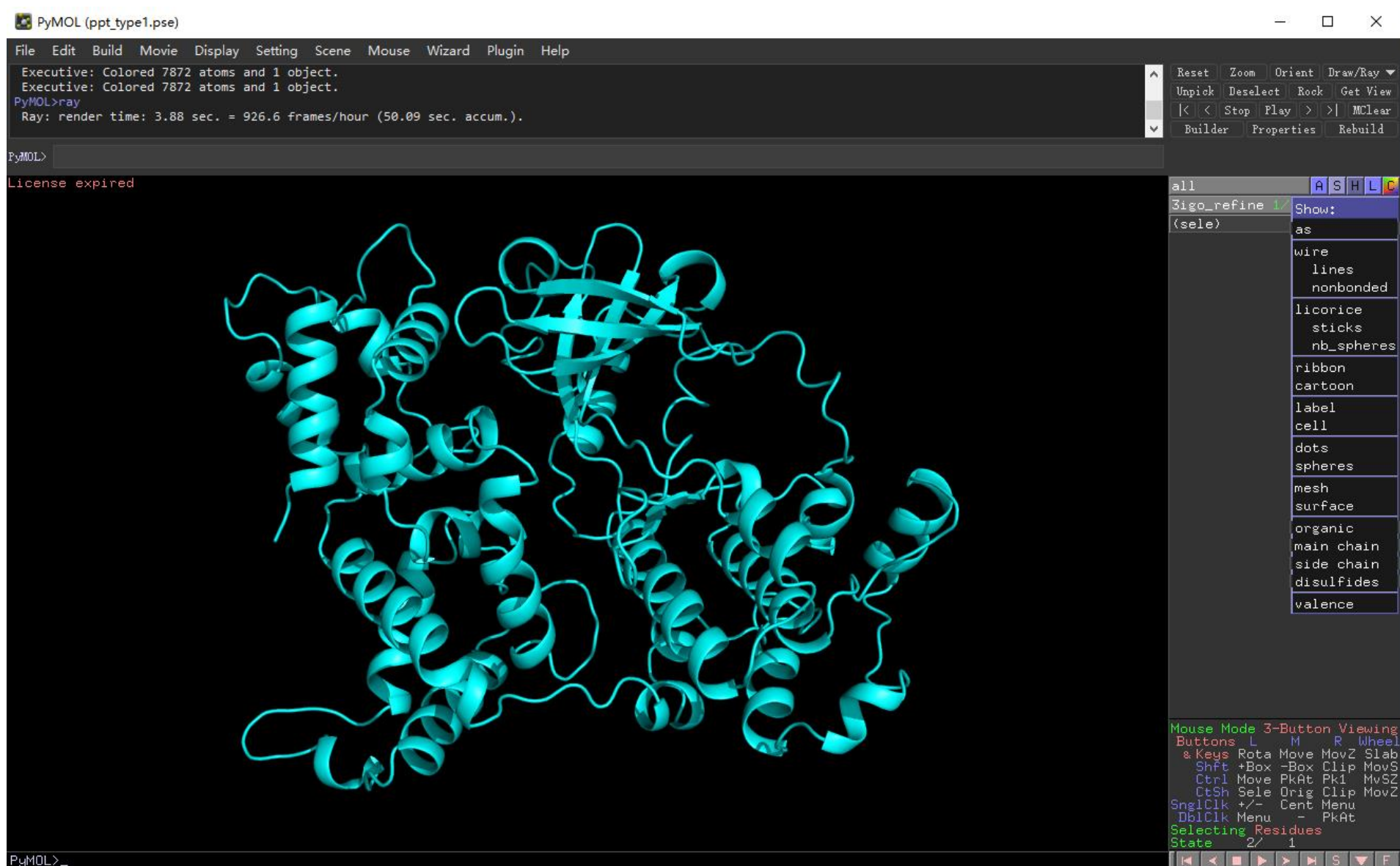
- 线（ Lines ）
- 棍子（ Sticks ）
- Cartoon (α 螺旋：螺旋或圆柱体； β 折叠片： 箭头)
- 只显示主链的走向
- 显示分子可及表面
- 根据需要标上不同颜色
 - 根据原子不同
 - 不同结构域
 - 不同二级结构
 - 彩虹着色（从N端到C端, 蓝到红）

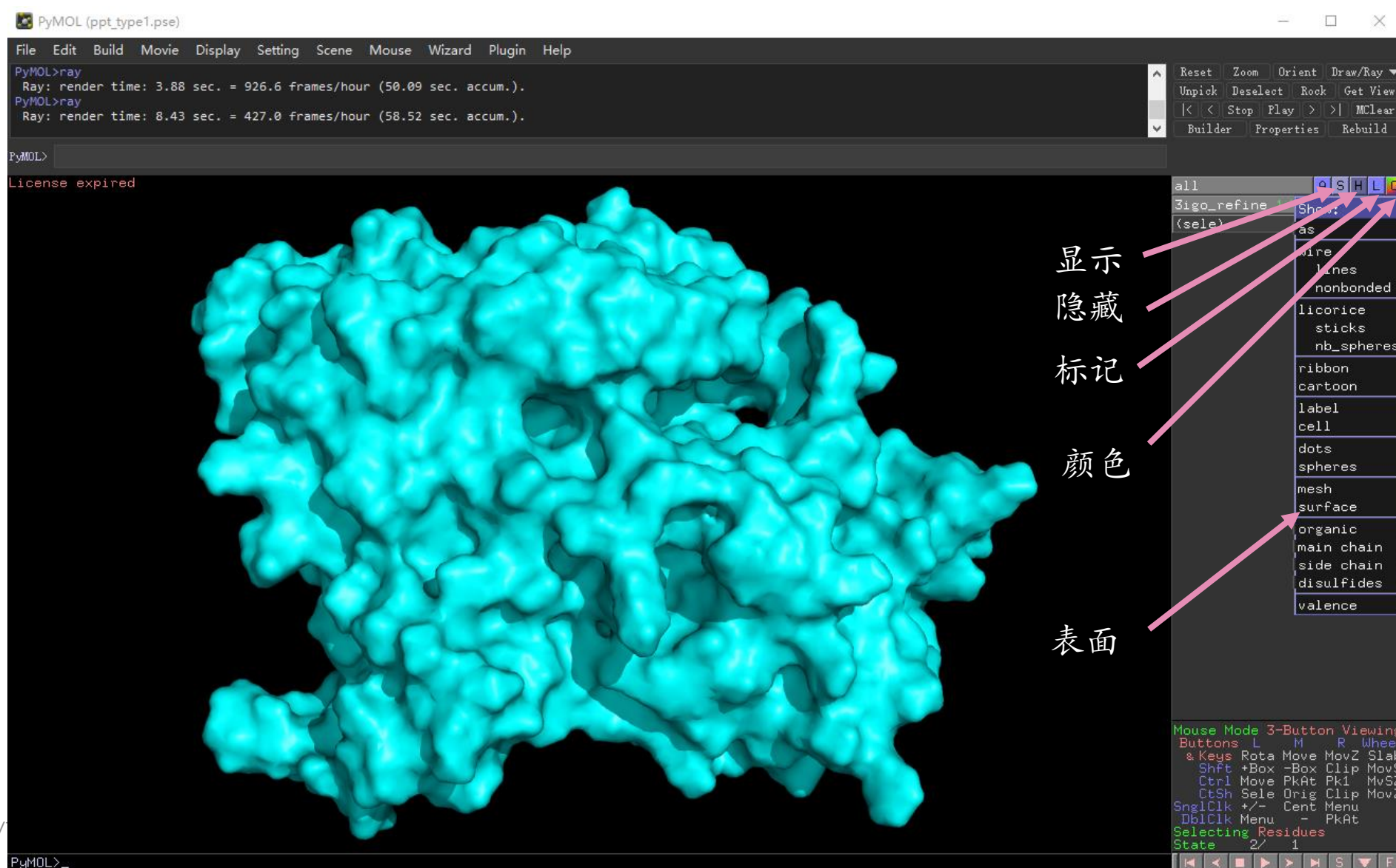


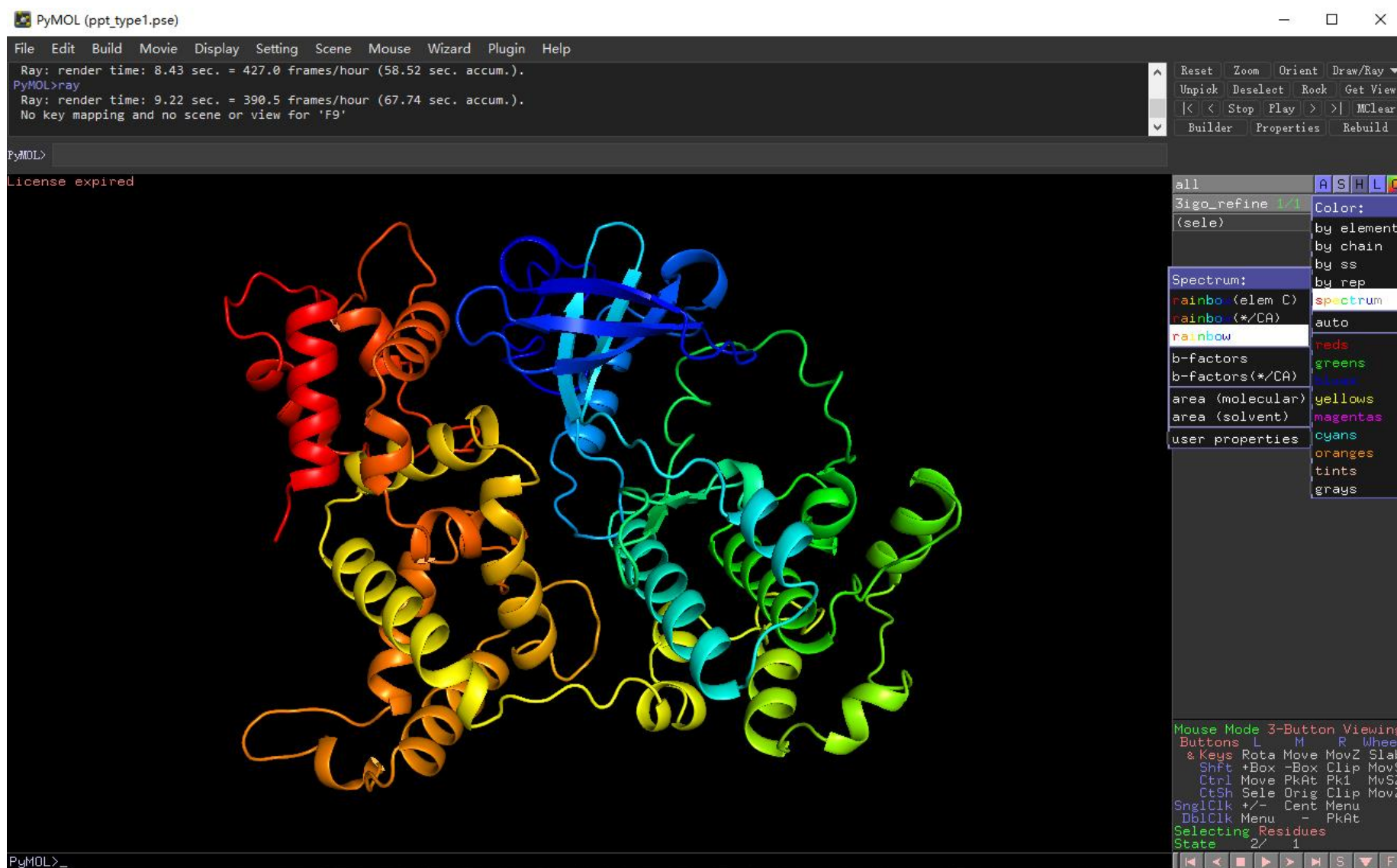
- PyMOL是一个开放源码, 由使用者赞助分子三维结构显示软件。
- 由Warren Lyford DeLano编写, 并且由DeLano Scientific LLC将它商业化。
- DeLano Scientific LLC是一个私人的软件公司, 它致力于创造让普遍的科学
与教育社群都能取得的好用软件工具。
- 在Warren Lyford DeLano过世后, PyMOL由薛定谔公司接手并负责进一步的开发。
- 目前网址: <https://pymol.org/>

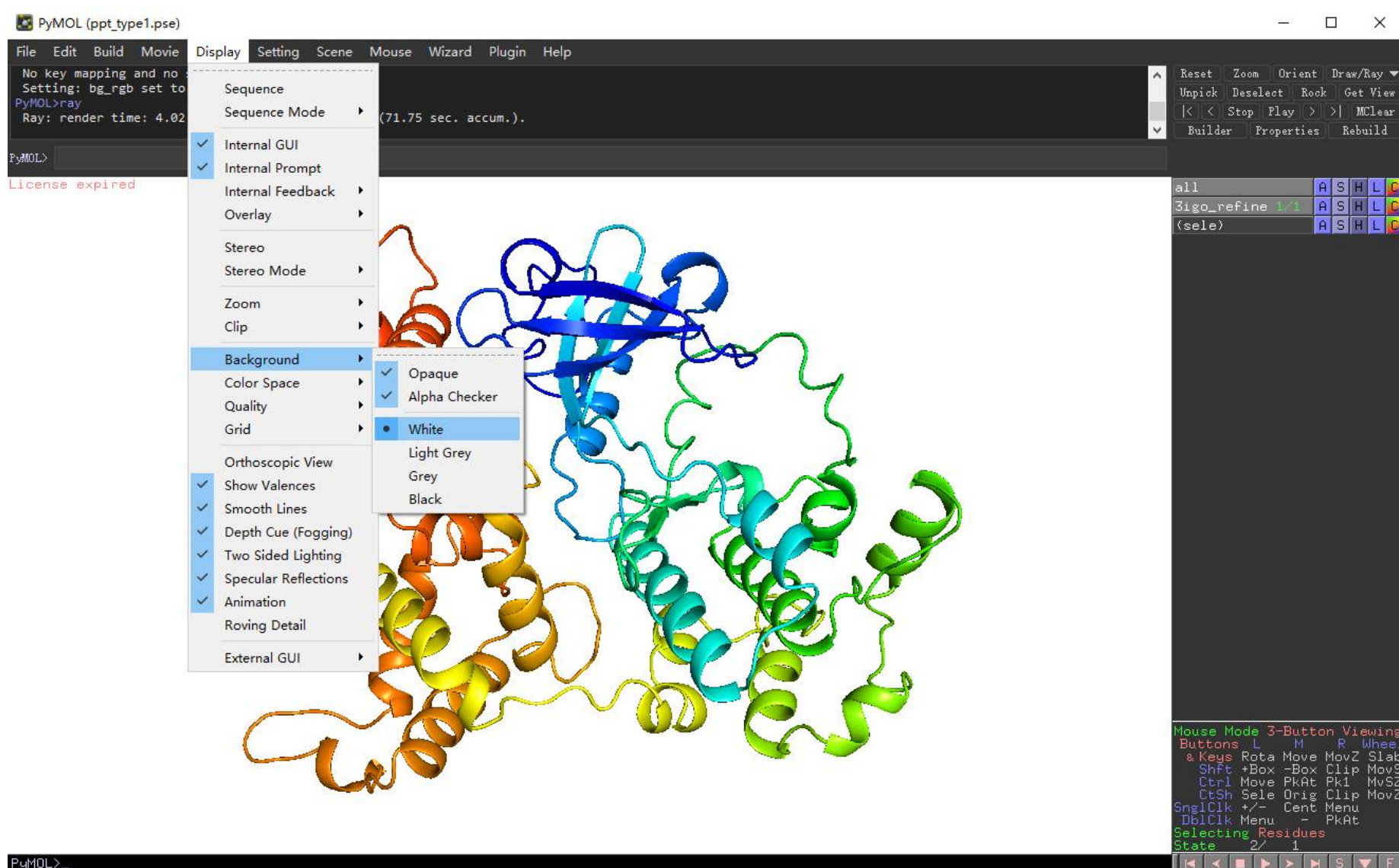


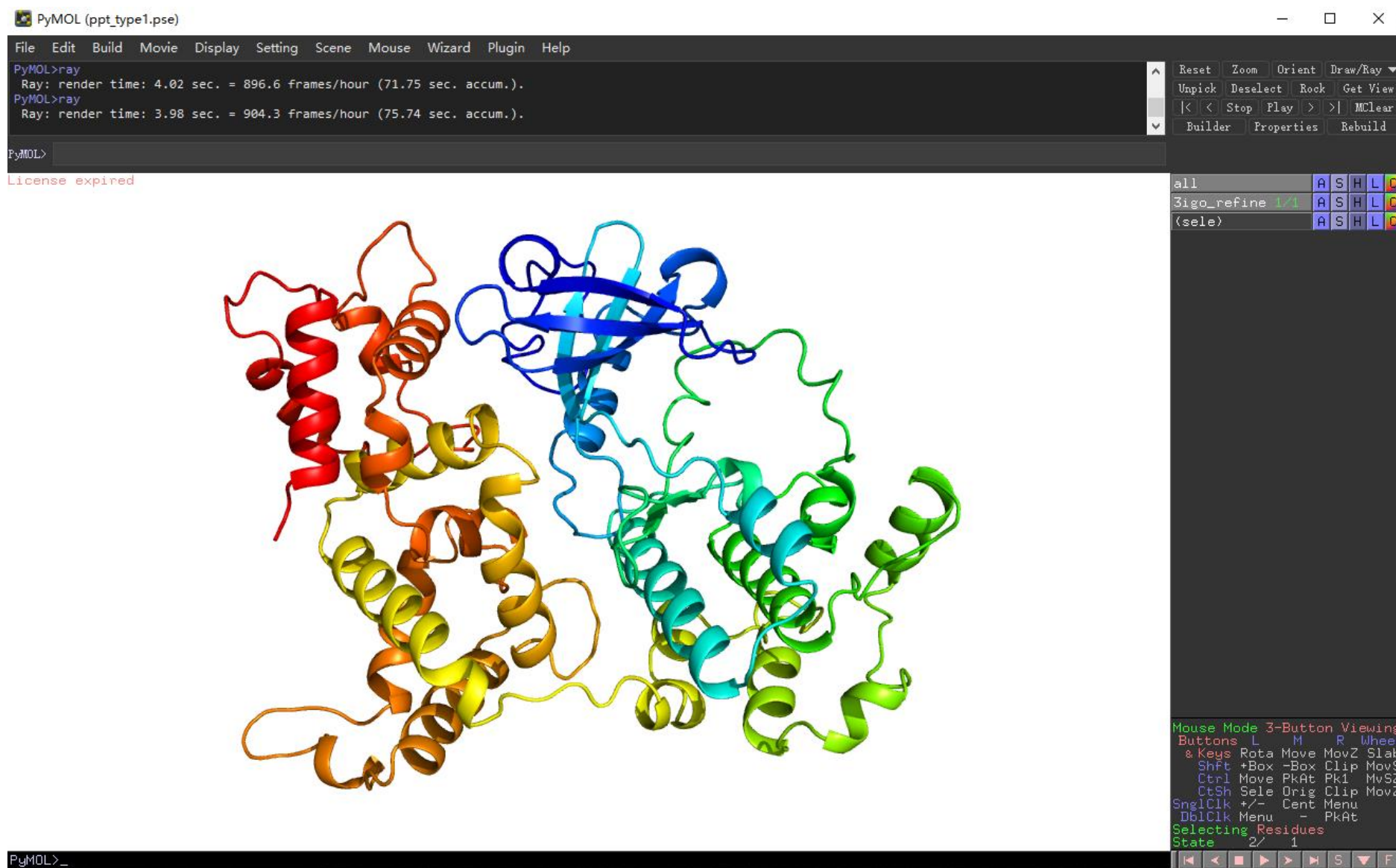
命令行和鼠标形式实现各种操作

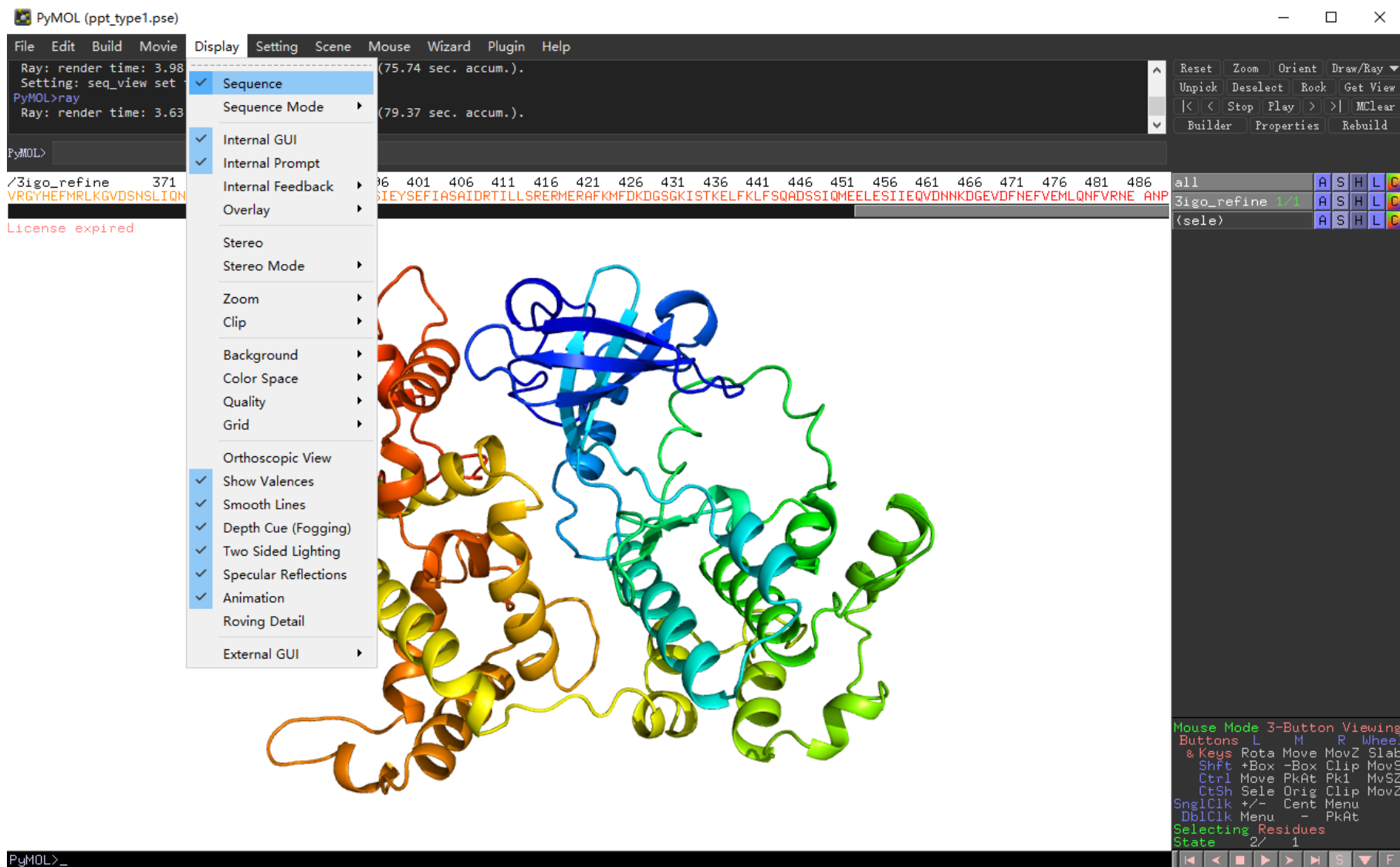


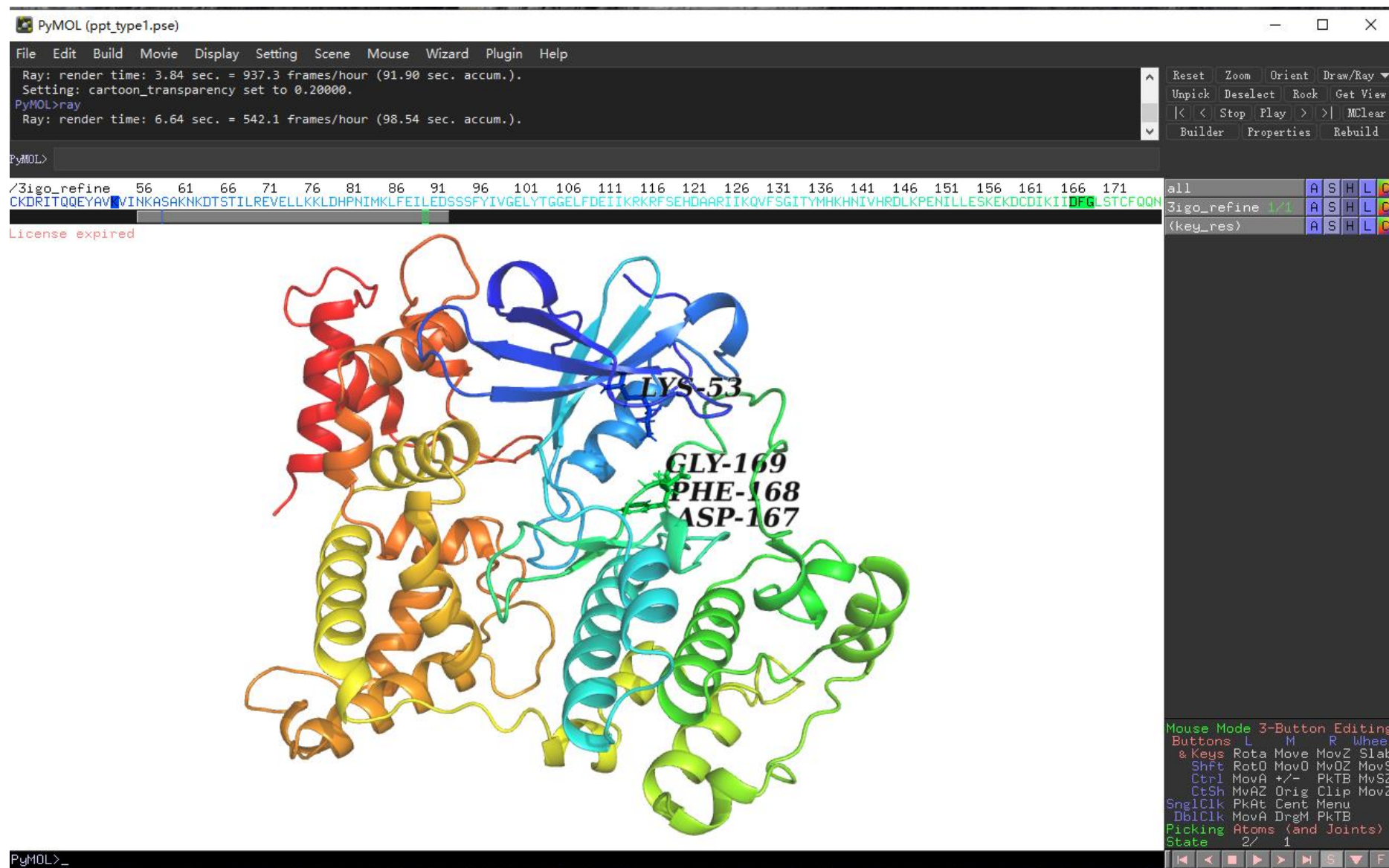


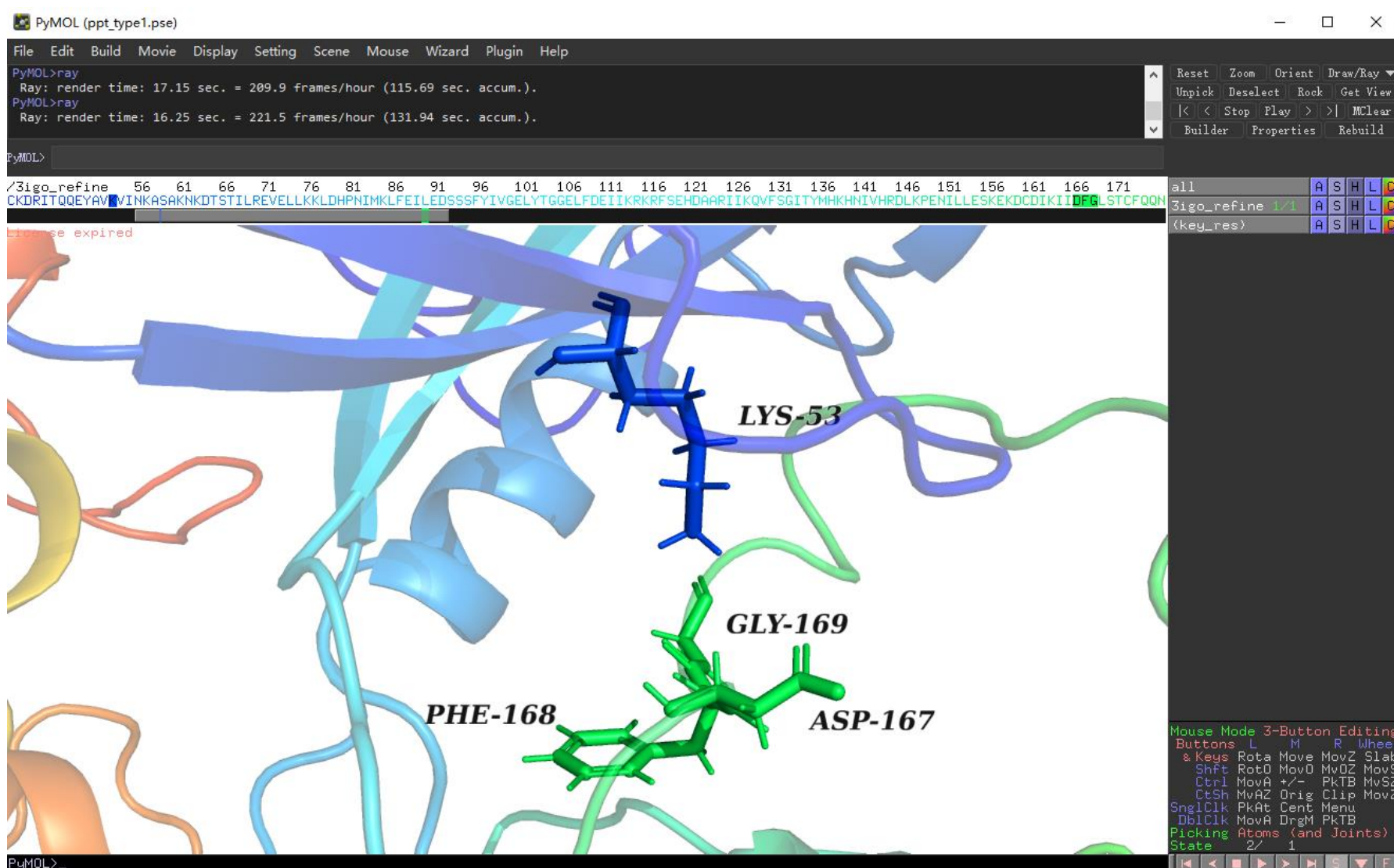


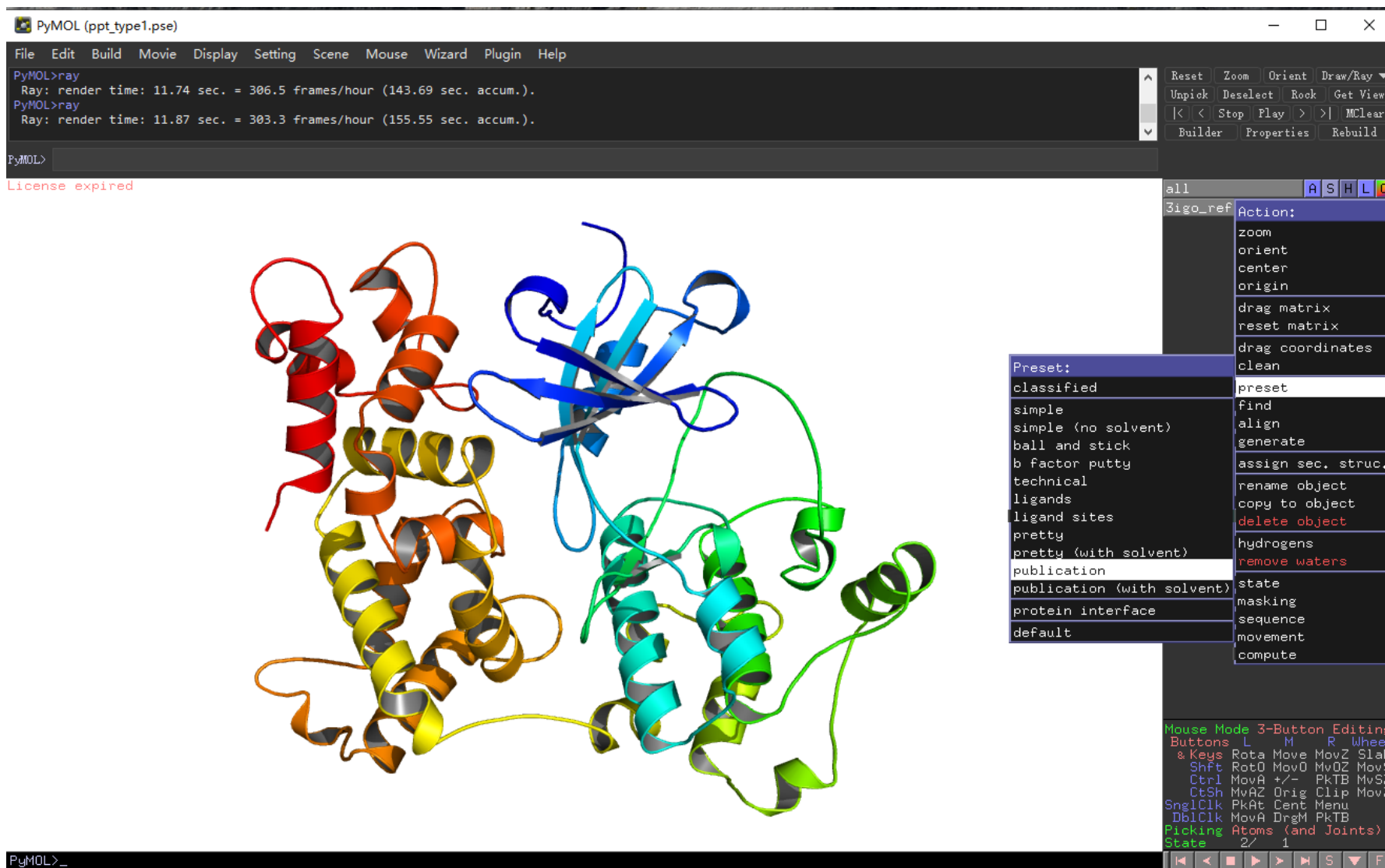


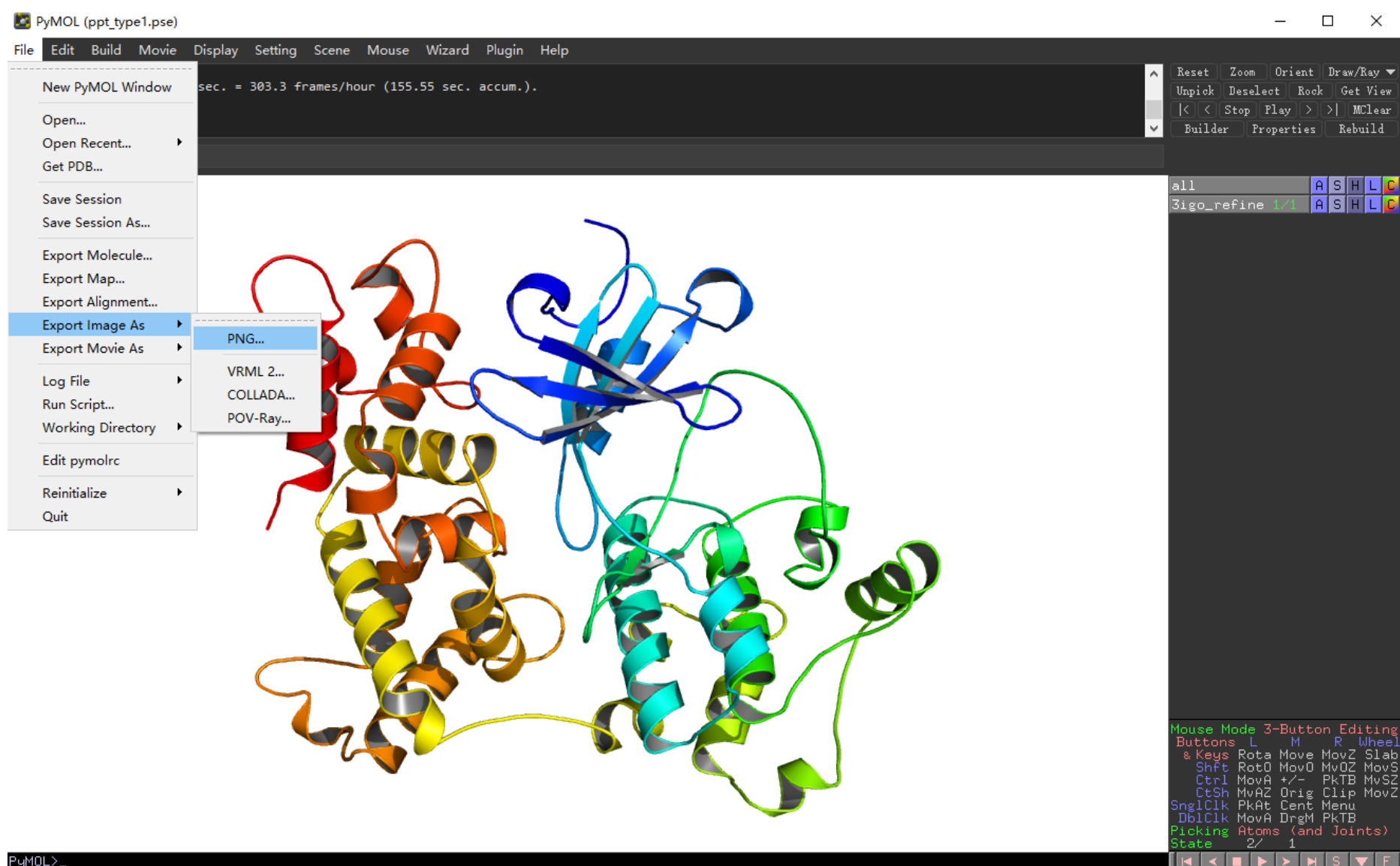








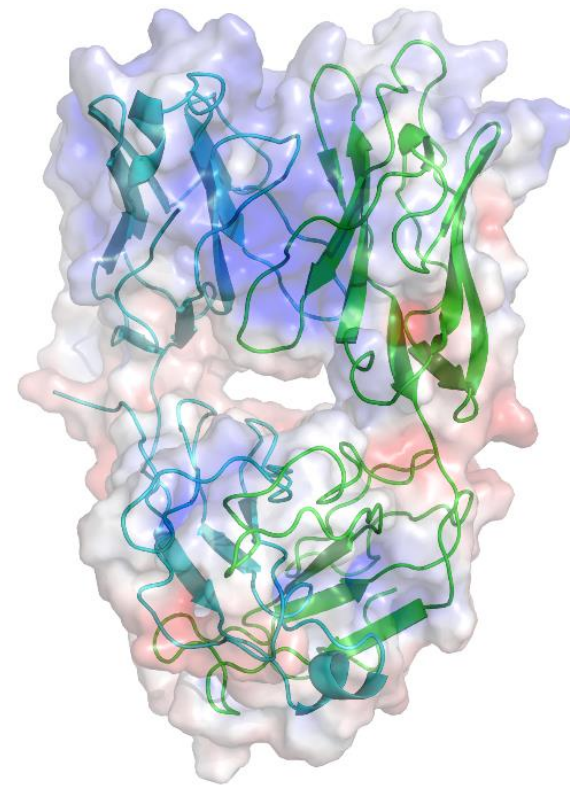


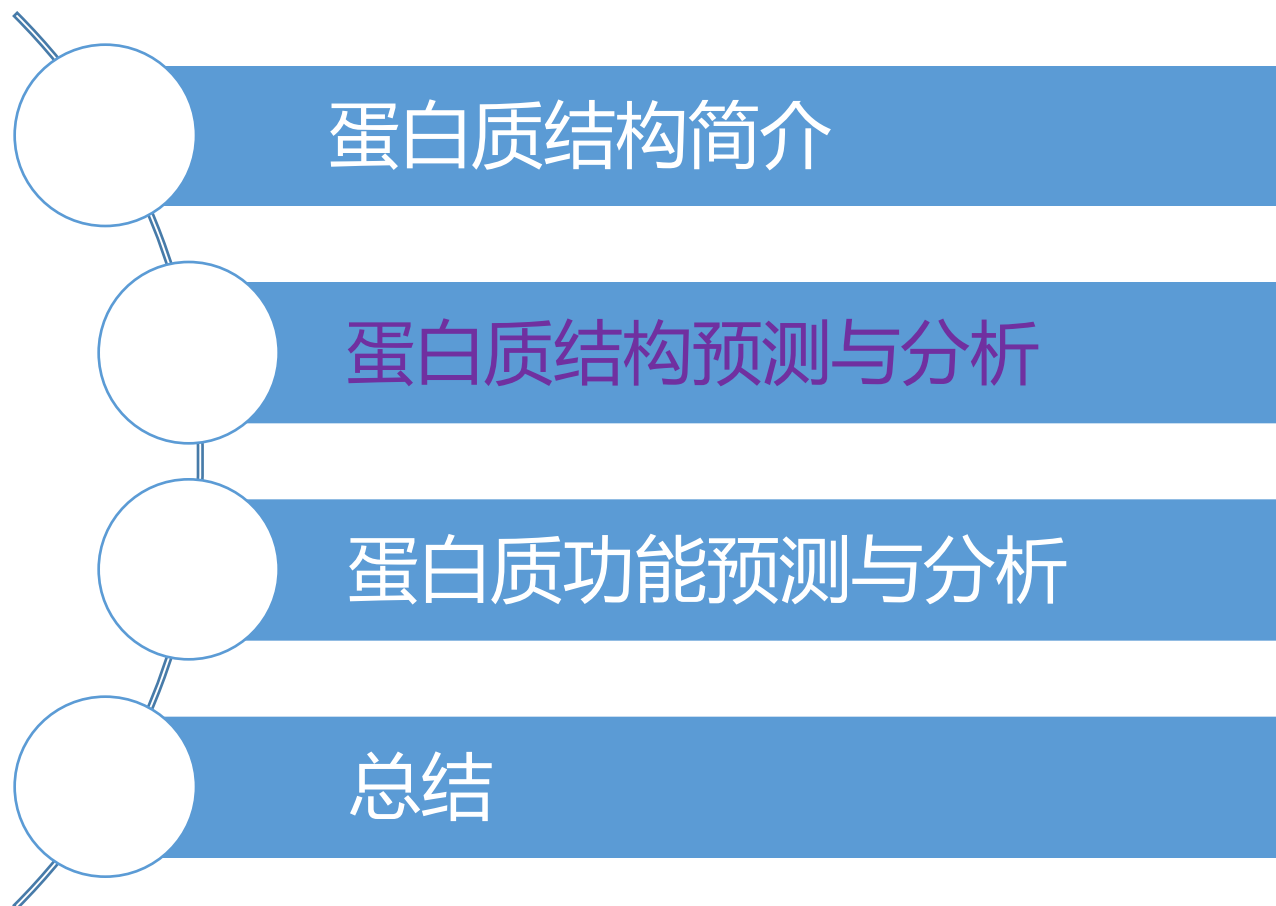


结构比对，通过命令行实现：
> align pdb1, pdb2



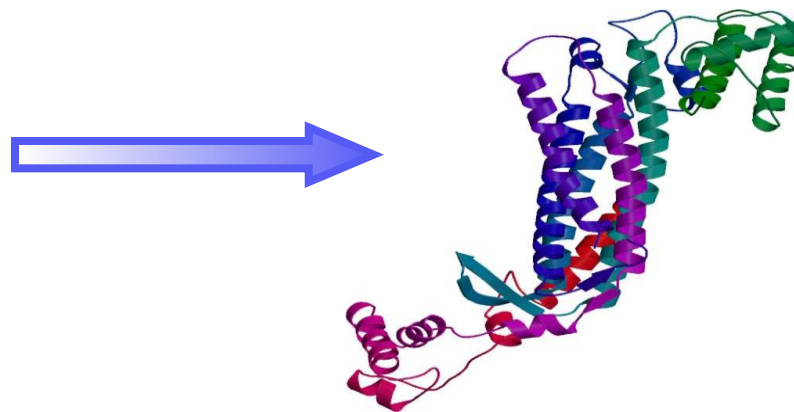
安装APBS插件进行表面
电荷分析



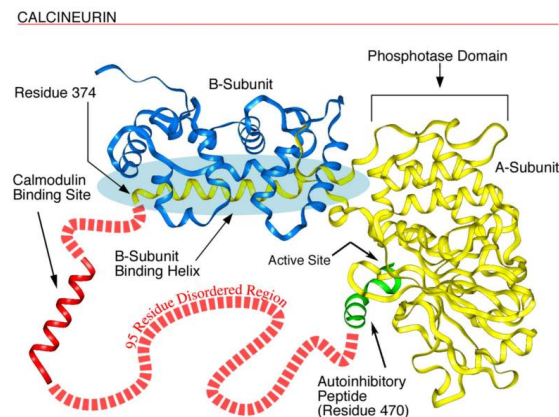
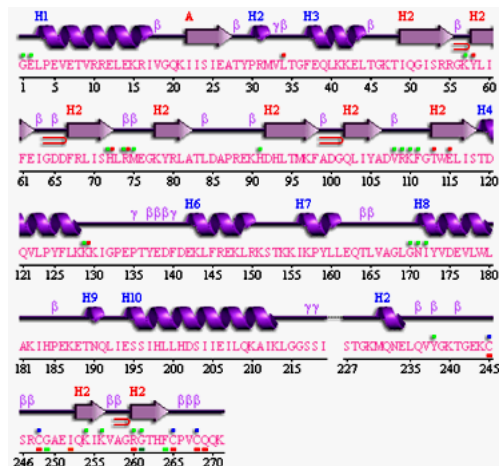


1. 蛋白质三级结构预测

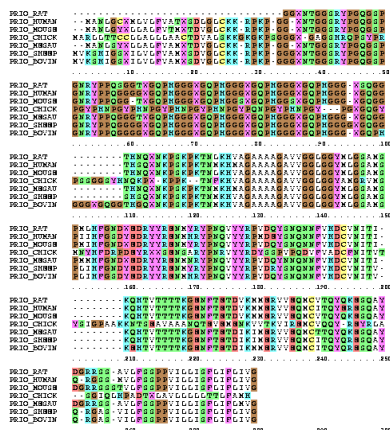
```
MNIFEMLRID EGLRLKIYKD TEGYYTIGIG
HLLTKSPSLN AAKSELDKAI GRNCNGVITK
DEAEKLFNQD VDAAVRGILR NAKLKPVYDS
LDAVRRCALI NMVFQMGETG VAGFTNSLRM
LQQKRWDEAA VNLAKSRYWYN QTPNRAKRVI
TTFRTGTWDA YKNL
```



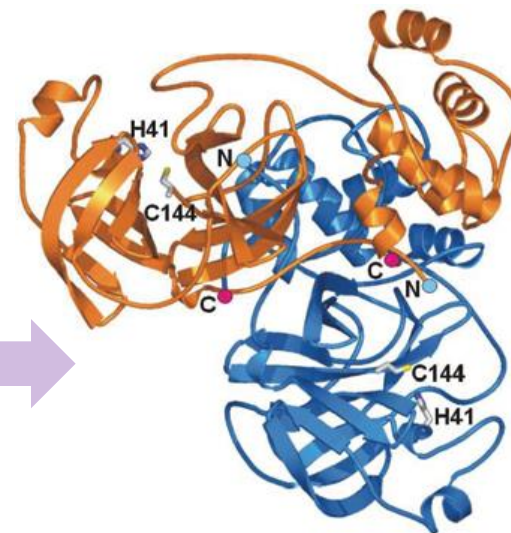
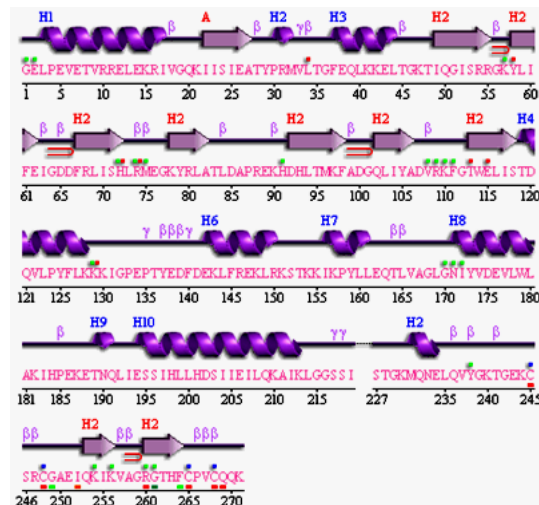
2. 蛋白质结构性性质预测（例如二级结构预测和无序区预测）



CINEMA alignment output



>80%



- 通过对已知蛋白的序列和其二级结构对应关系的统计分析，学习和归纳出一些预测规则，用于待测蛋白质的二级结构预测
- 为其它基于统计的生物信息学预测问题奠定方法学基础
- 二级结构预测可为三级结构预测提供一个好的初始点

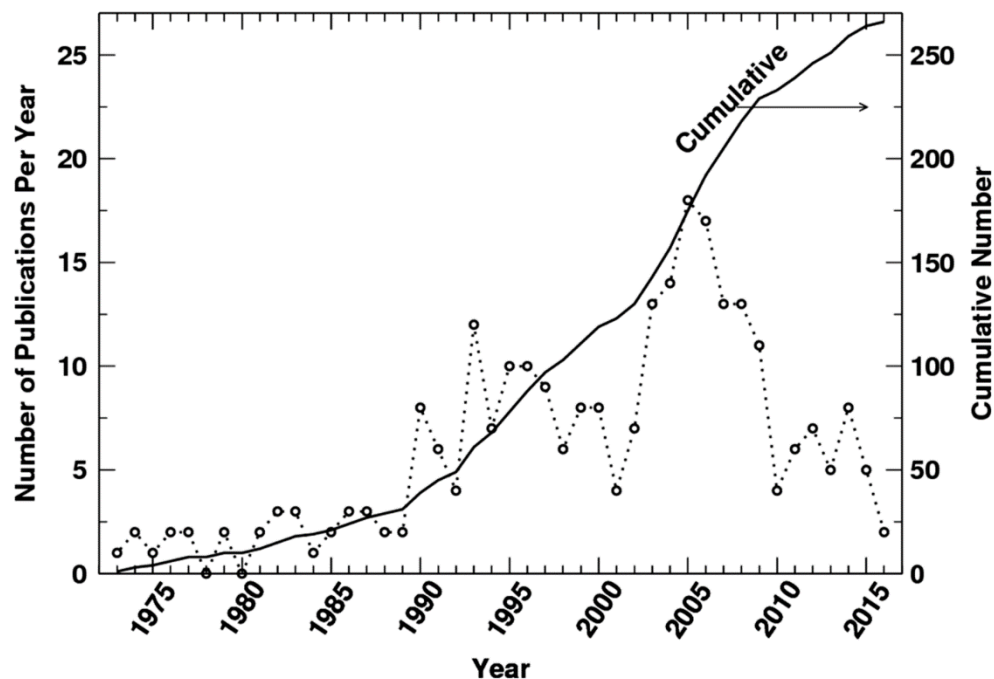


Table 1. Method comparison based on Q3 using newly released structures (TS115) and CASP12 targets (15 proteins) for secondary structure prediction

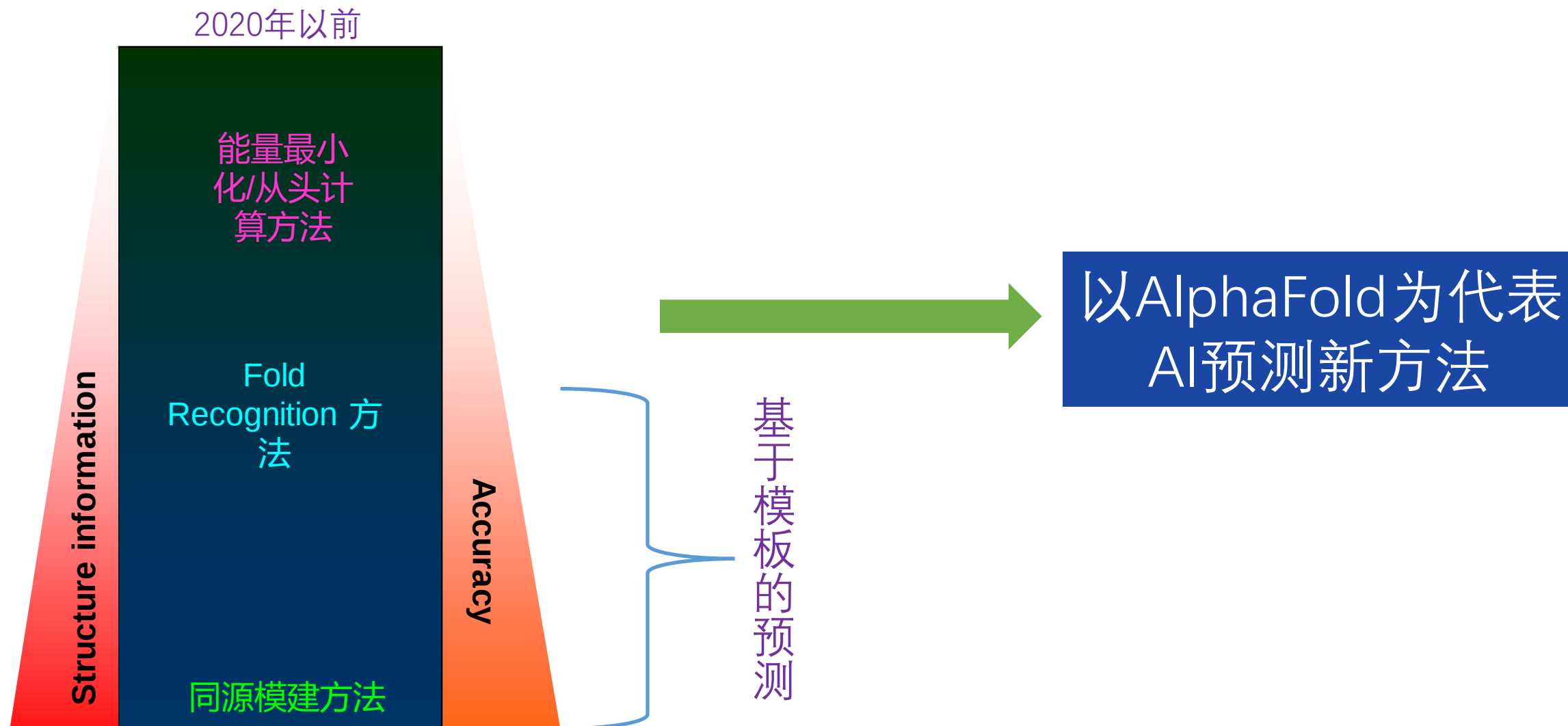
Data set Method	TS115		CASP12		Server location
	Q3	P-value ^a	Q3	P-value ^a	
Jpred4	0.771 ^b	0.0007	0.751	0.04	http://www.compbio.dundee.ac.uk/jpred4/index.html
SPINE X	0.801	0.0002	0.769	0.006	http://sparks-lab.org/SPINE-X/
PSIPRED 3.3	0.802	0.12	0.780	0.19	http://distillf.ucd.ie/porterpaleale/
SCORPION	0.817	0.45	0.805	0.44	Stand-alone version from http://hpcr.cs.odu.edu/c3scorpion/
SPIDER2	0.819	NA	0.798	NA	http://sparks-lab.org/server/SPIDER2/
PORTER 4.0	0.820	0.17	0.798	0.67	http://distillf.ucd.ie/porterpaleale/
DeepCNF	0.823	0.01	0.821	0.14	http://raptorx2.uchicago.edu/StructurePropertyPred/predict/

Note.

^aPaired t-test from SPIDER2.

^bJpred only predicts the sequence <800 residues. For TS115, there is one sequence (5hdtA) with 1085 residues. 5hdtA was divided into two chains with 800 residues and 285 residues, respectively.

Yang et al. (2016) *Briefings in Bioinformatics*, 19: 482-494.





原理：天然稳定构像对应自由能最低的构像

困难：折叠态与非折叠态能量差很小，要求特别准确的能量项，局部能量最小问题

代表人物：Scheraga HA, Cornell Univ, >1000 publications

目前流行的两种从头算法



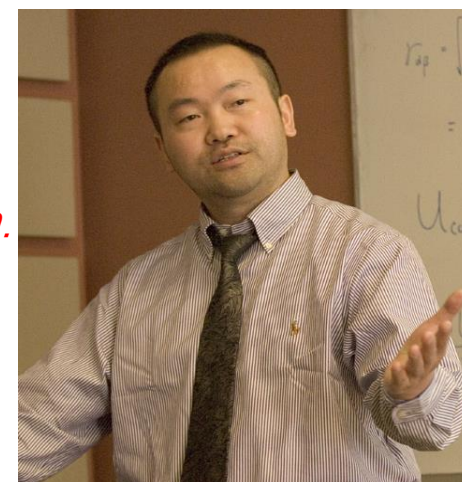
David Baker

Rosetta

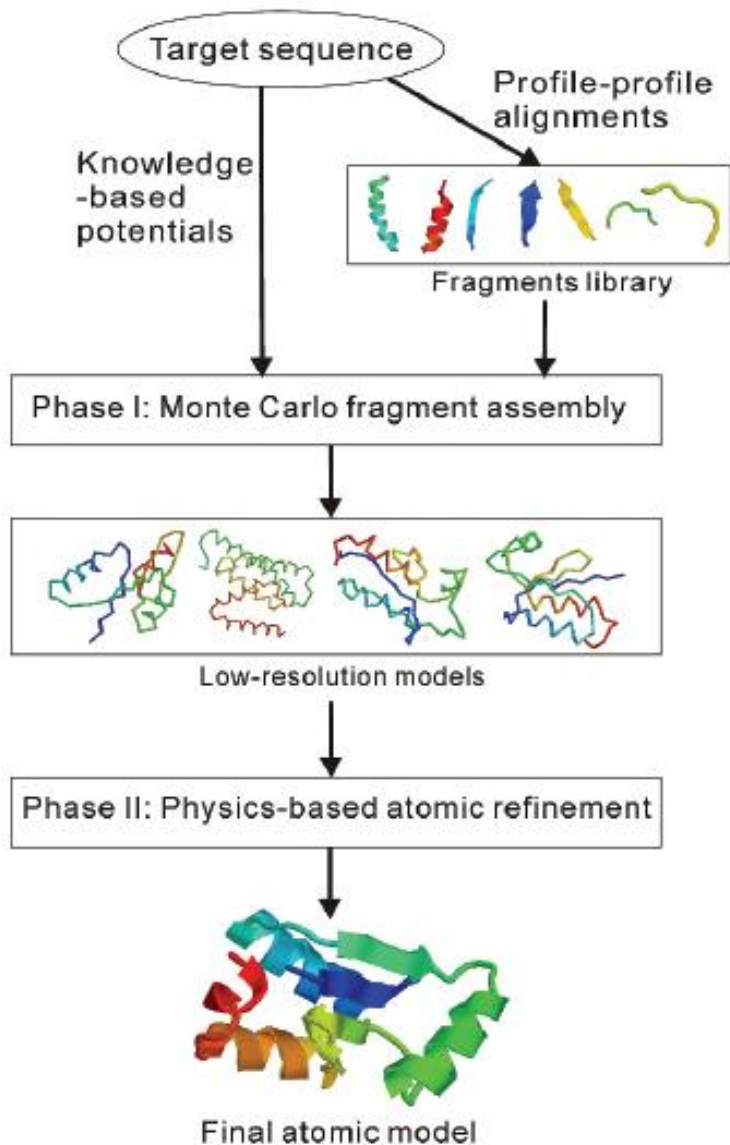
<http://rosetta.bakerlab.org/>



<http://zhanglab.ccmb.med.umich.edu/I-TASSER/>



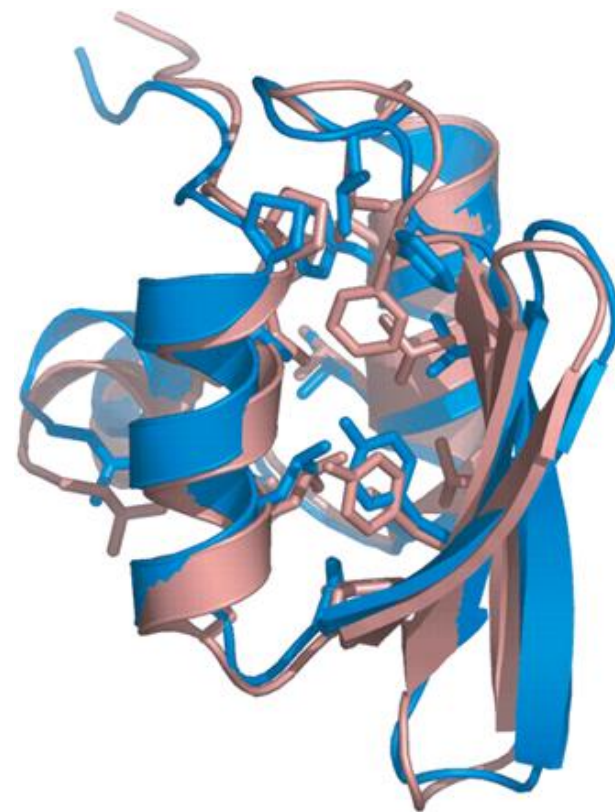
Yang Zhang



分两步:

- Knowledge-based potential, 粗粒化模型
- Physics-based potential, 全原子模型

成功预测例子:
RMSD=1.6Å 预测 (蓝) 晶体结构(红, PDB: 1whz)
Science, 2005,309:1868-1871



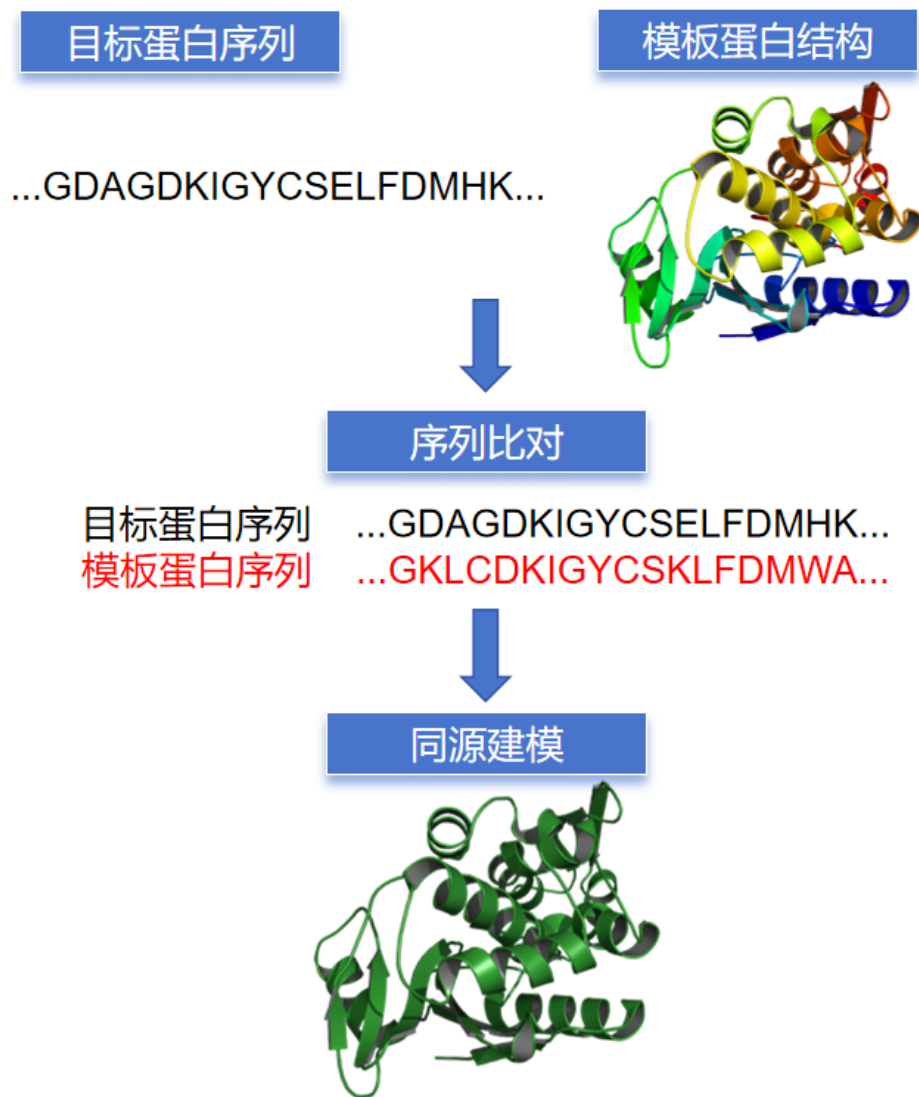
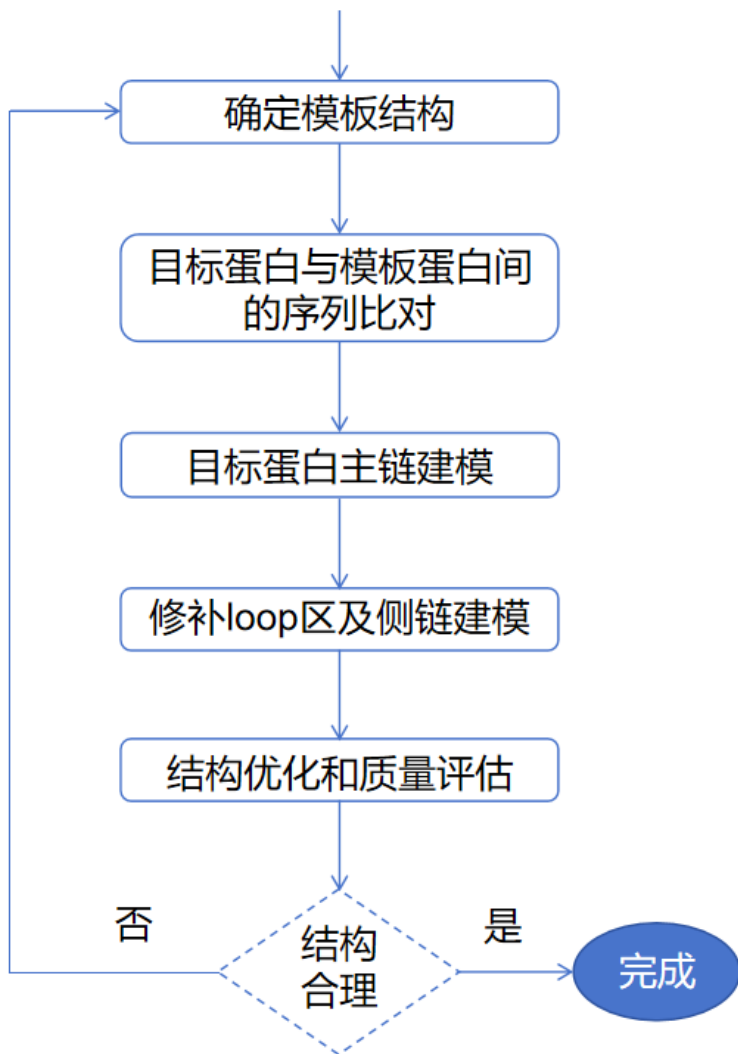
原理：蛋白质三维结构比一级结构更为保守。通过找出结构未知蛋白质与结构已知蛋白质的同源性，则可以用结构已知蛋白质为结构模板，预测未知蛋白质的结构。

Blundell TL, Sibanda TL, Sternberg MJE & Thornton JM. (1987) Knowledge-based prediction of protein structures and the design of novel molecules, *Nature* **326**, 347 - 352 .

可能最流行的同源模建程序：Modeller
<http://salilab.org/modeller/modeller.html>
Sali A & Blundell TL. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779-815.

可能最流行的同源模建在线服务器：Swiss-Model
<https://swissmodel.expasy.org/>
Waterhouse A et al (2018) *Nucleic Acids Res* **46**, W296-W303.



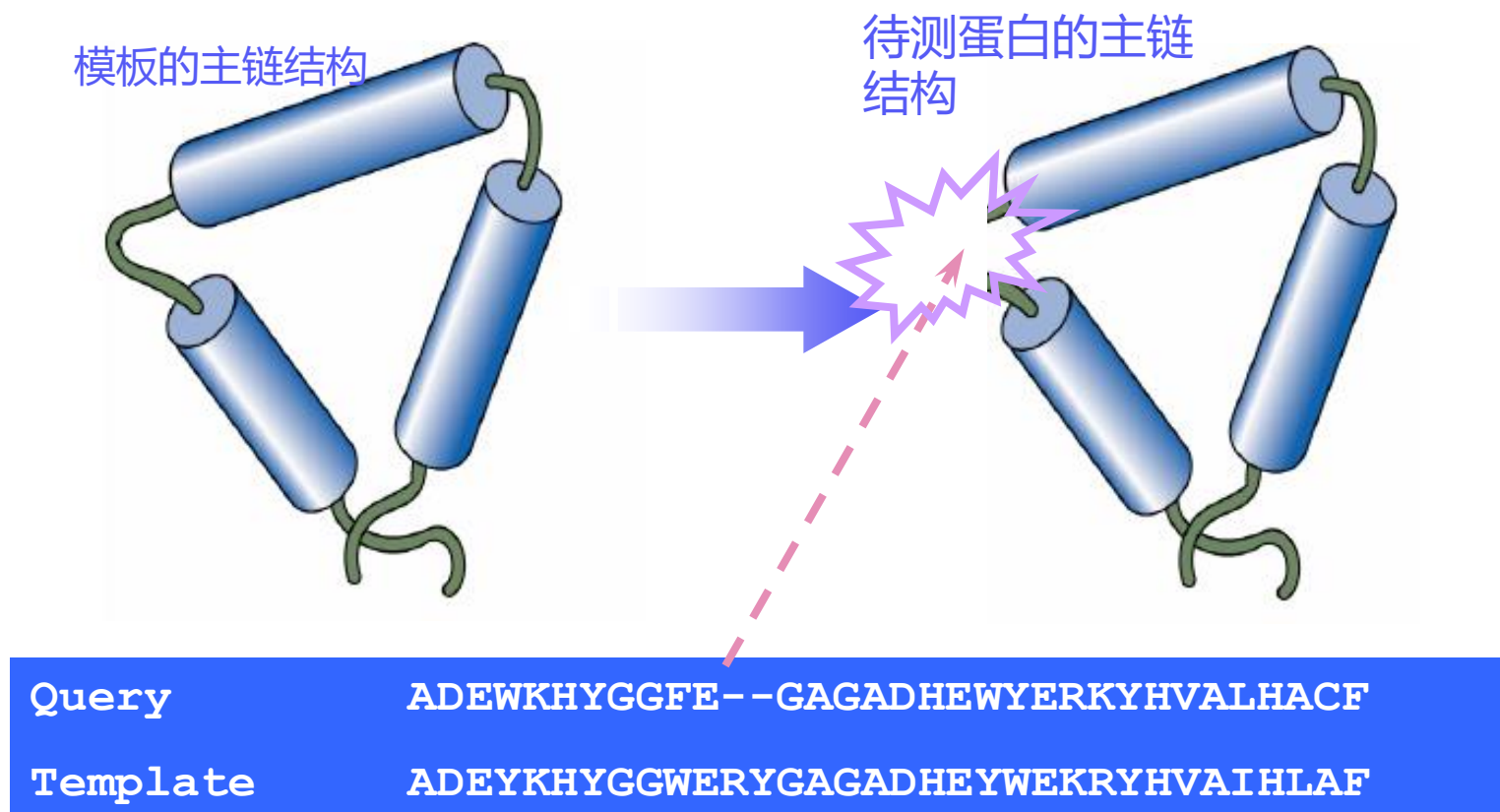


模板的选取及比对的获得

- 同源性搜索PDB或蛋白质结构分类数据库 (Sequence Identity或*E*-value)
- 目标序列和模板的序列比对(例如ClustalW)

主链结构预测

- 根据比对结果, 将已知结构中的片段拷贝到未知结构中去

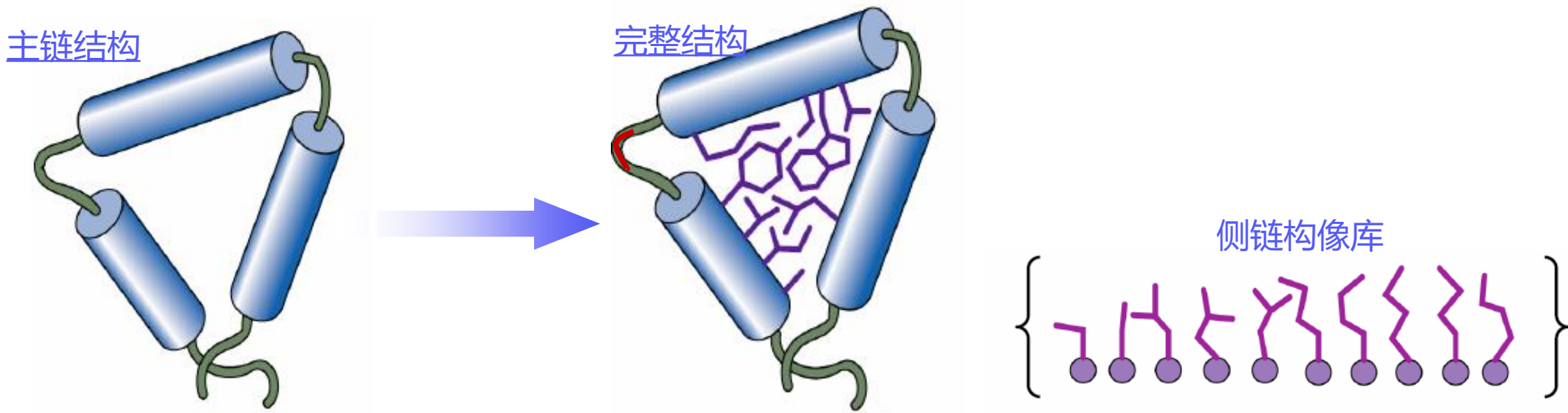


非保守区(loop区)主链结构预测

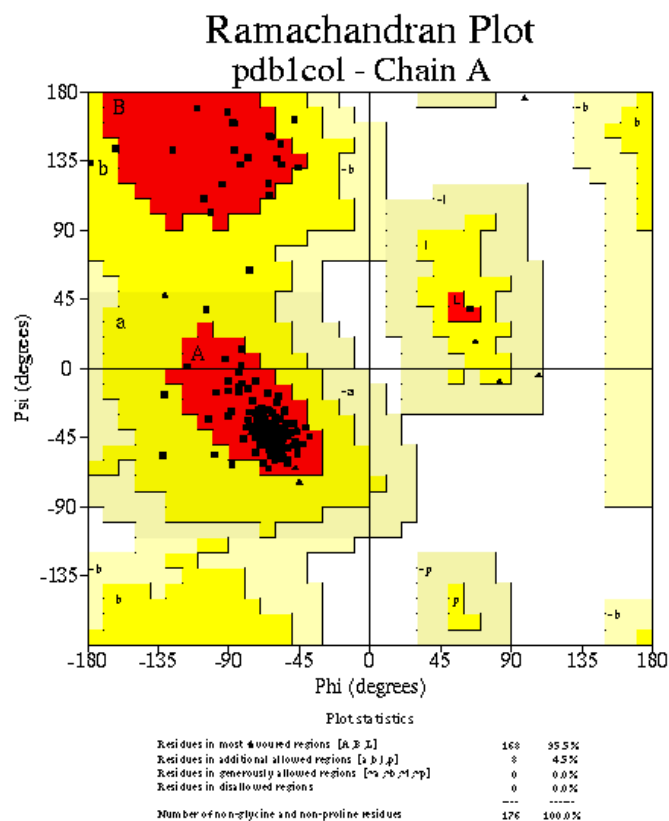
- 数据库查询办法 (经常采用)
- 系统构象搜索办法 (要求更多计算量)
- 对于较长氨基酸片段的loop, 很难保证模拟的结果是可靠的

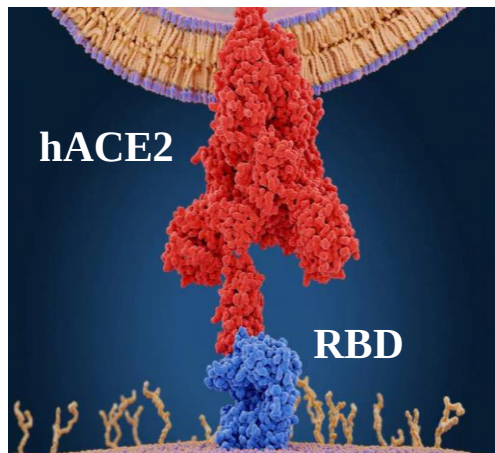
侧链安装

- 主要原则是借鉴已知结构蛋白的侧链构象, 或从能量角度和空间位阻角度出发预测目标蛋白中残基的侧链构象
- 目前侧链构象预测方法主要利用构象库 (rotamers) 的概念



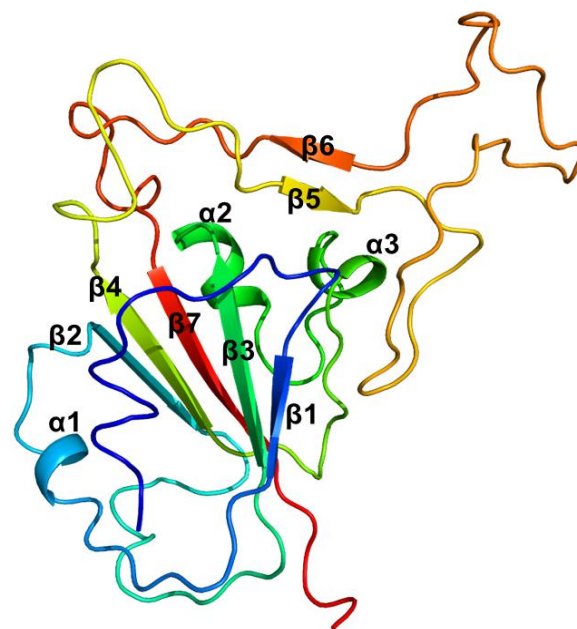
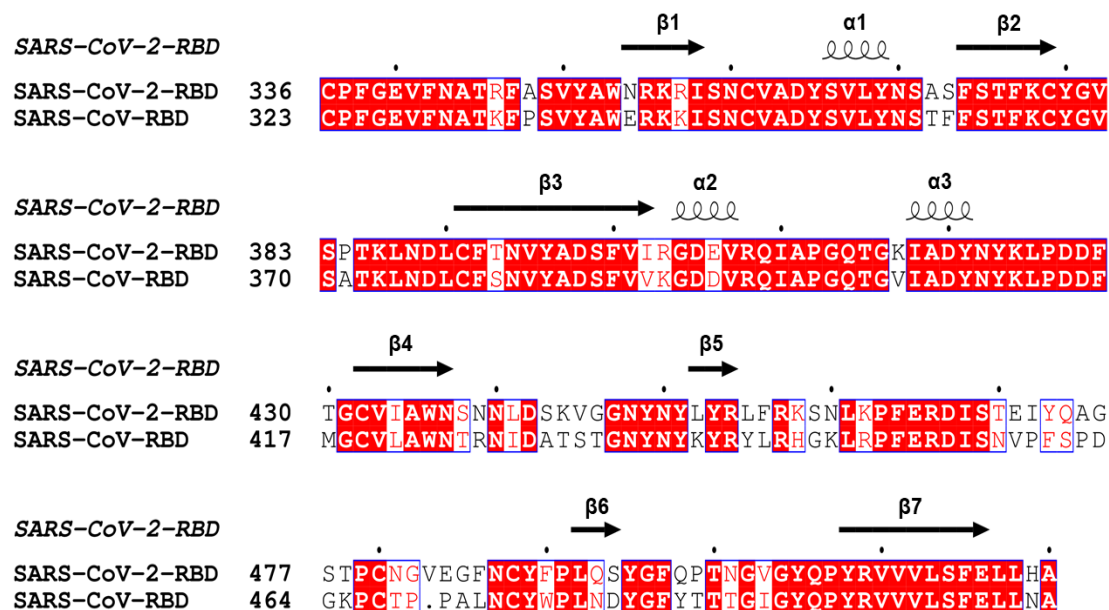
- 同源预测得到的蛋白质结构模型，通常含有较多不合理的原子间接触，因而必须进行能量优化
- 拉氏构象图可用于评估模型的立体化学性质



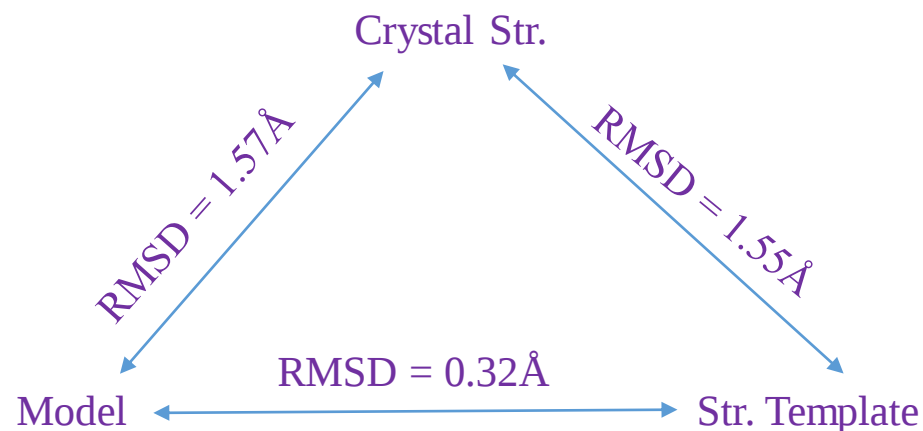
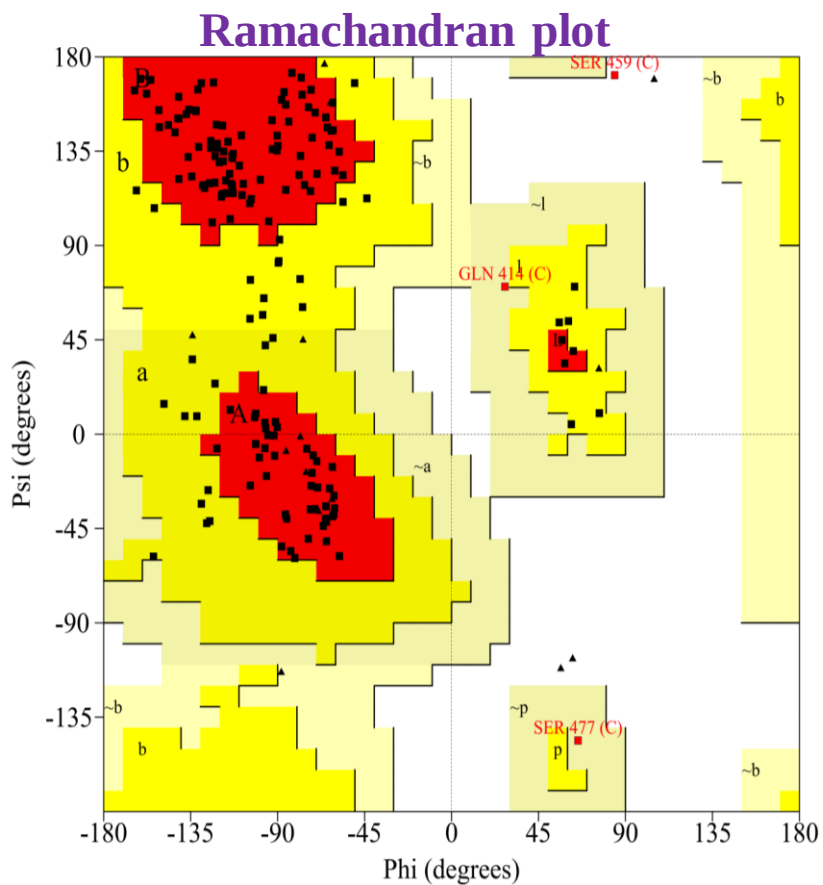


Spike protein (刺突蛋白)

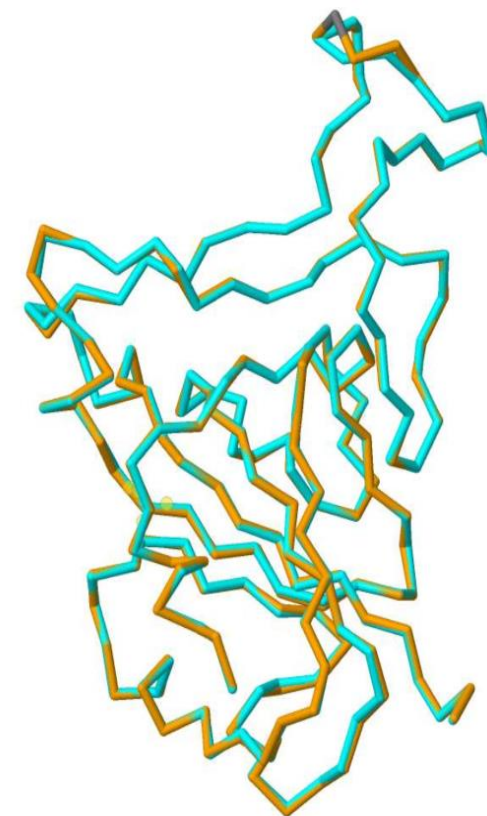
- A glycoprotein on the surface of SARS-CoV-2
- Can recognize human receptor protein hACE2 (Angiotensin-converting enzyme 2, 血管紧张素转化酶2 hACE2) to complete cell fusion through RBD (Receptor Binding Domain)
- The RBD of SARS-CoV was selected as the structure template (Seq. identity%=72.83%)
- Sequence alignment was generated from Clustal
- 3D model was generated from SWISS-MODEL



The 3D structure of SARS-CoV-2 RBD, containing 7-antiparallel β -sheets and 3-helices.



- The current model is highly accurate
 - Well superimposition in most area
 - Only slight difference in loop regions
- It can be used for virtual screening or further computational analysis



Structural superimposition
between predicted model (黄色)
& Crystal Str.(浅蓝色)

依据：研究表明蛋白质序列之间存在很低的序列相似性，但它们仍具有结构相似性。因此，弱同源模拟可拓宽同源模拟的使用范围。

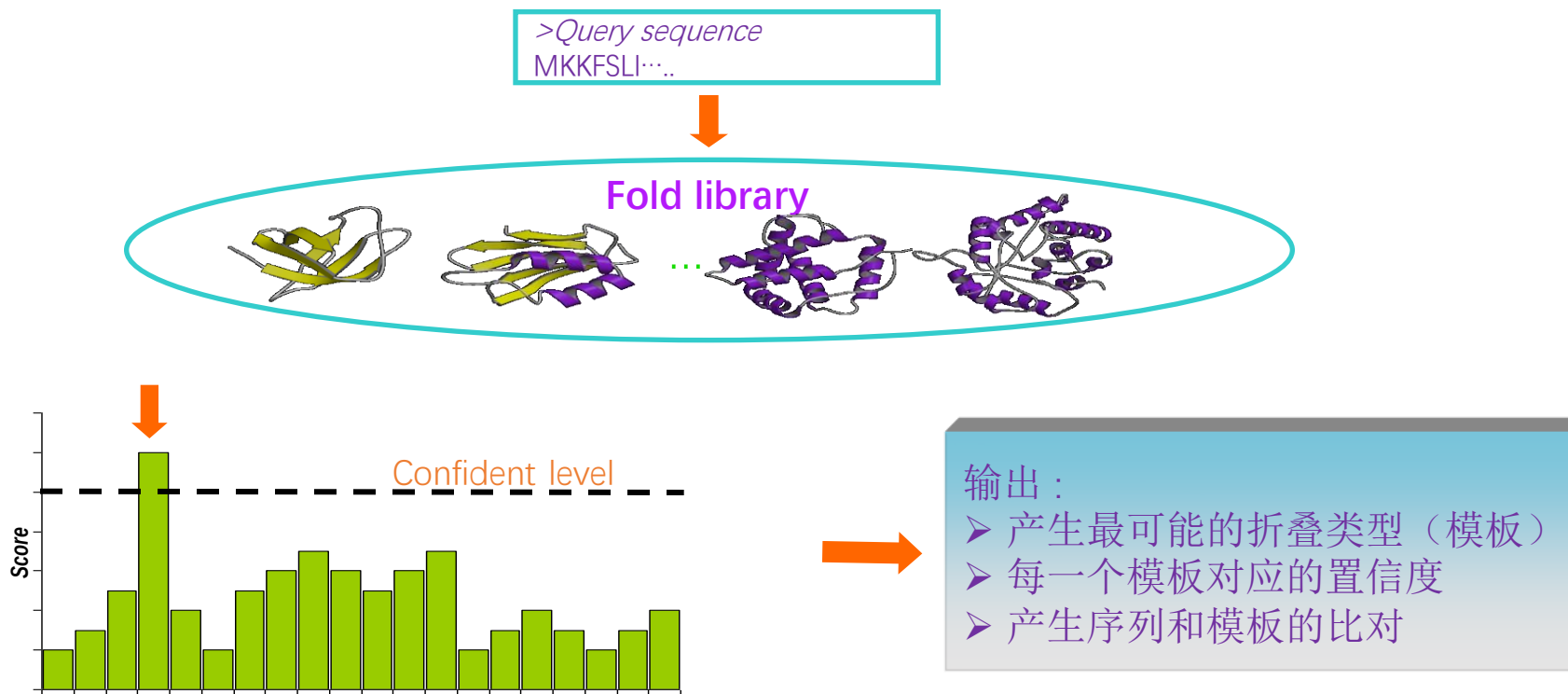
第一代方法 (又称为 Threading 和 Inverse folding):将待测的氨基酸序列与已知的蛋白质结构进行匹配，找出匹配最好的结构作为结构预测模板

最早的文献：

Bowie JU, Luthy R, Eisenberg D. (1991) "A method to identify protein sequences that fold into a known three-dimensional structure." *Science*, **253**:164-70.



第二代方法 (又称为 Fold Recognition):最显著特点是多信息耦合，特别是对PSI-BLAST的整合。其可靠性仅次于同源建模，是蛋白质结构预测一种重要方法。



最核心问题：

如何构造打分函数，即如何衡量Sequence-Structure的相容性？

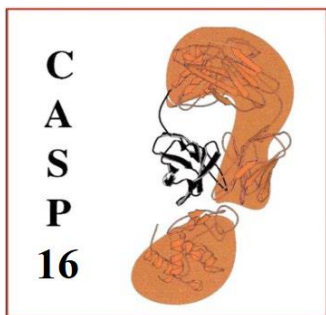
主要策略：

第一代方法：平均势能，基于环境的氨基酸替换矩阵

第二代方法：PSI-BLAST整合，结构信息的整合，机器学习方法等

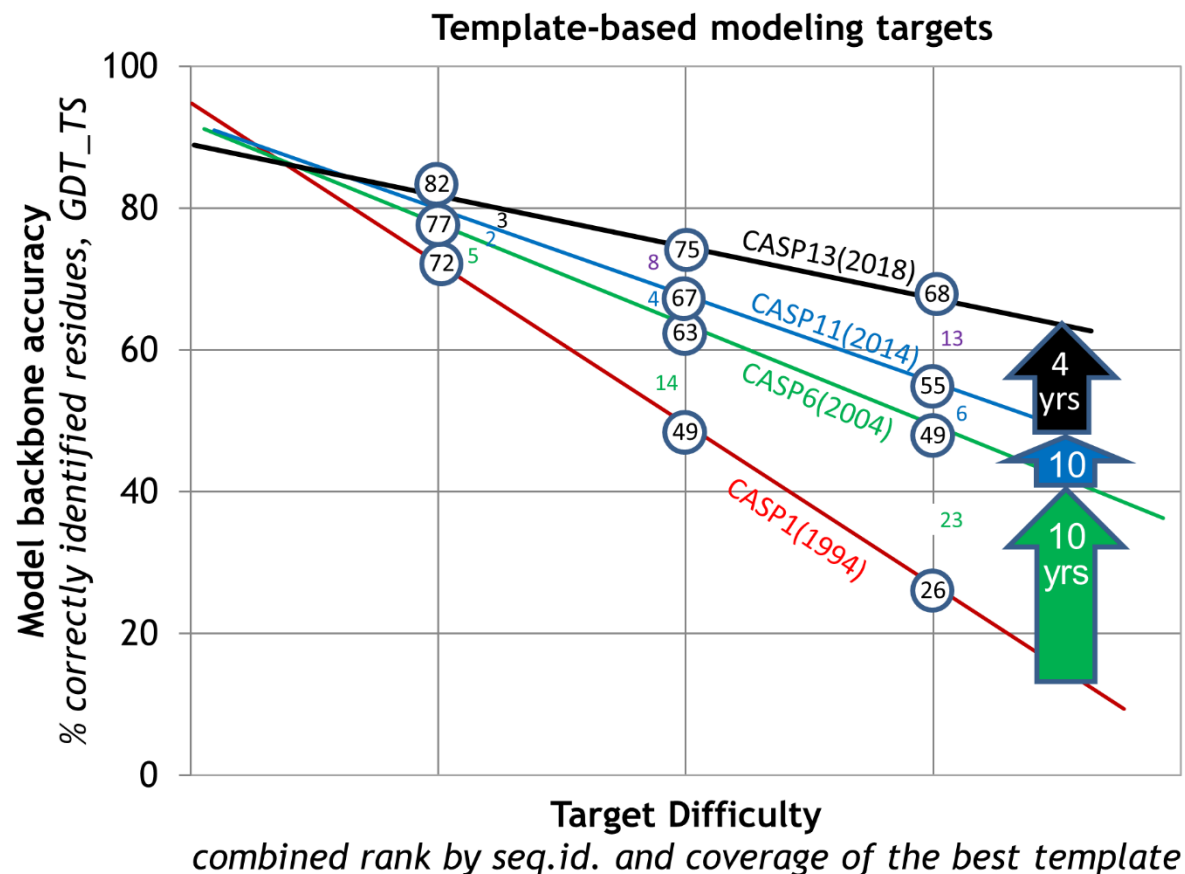
流行的方法：3D-PSSM、pGenThreader和FFAS03等

- 生物信息学家和结构生物学家合作，开展蛋白结构“双盲”预测竞赛。
- <http://predictioncenter.org/>
- 自1994年起，每两年举行一次，今年是第16届。

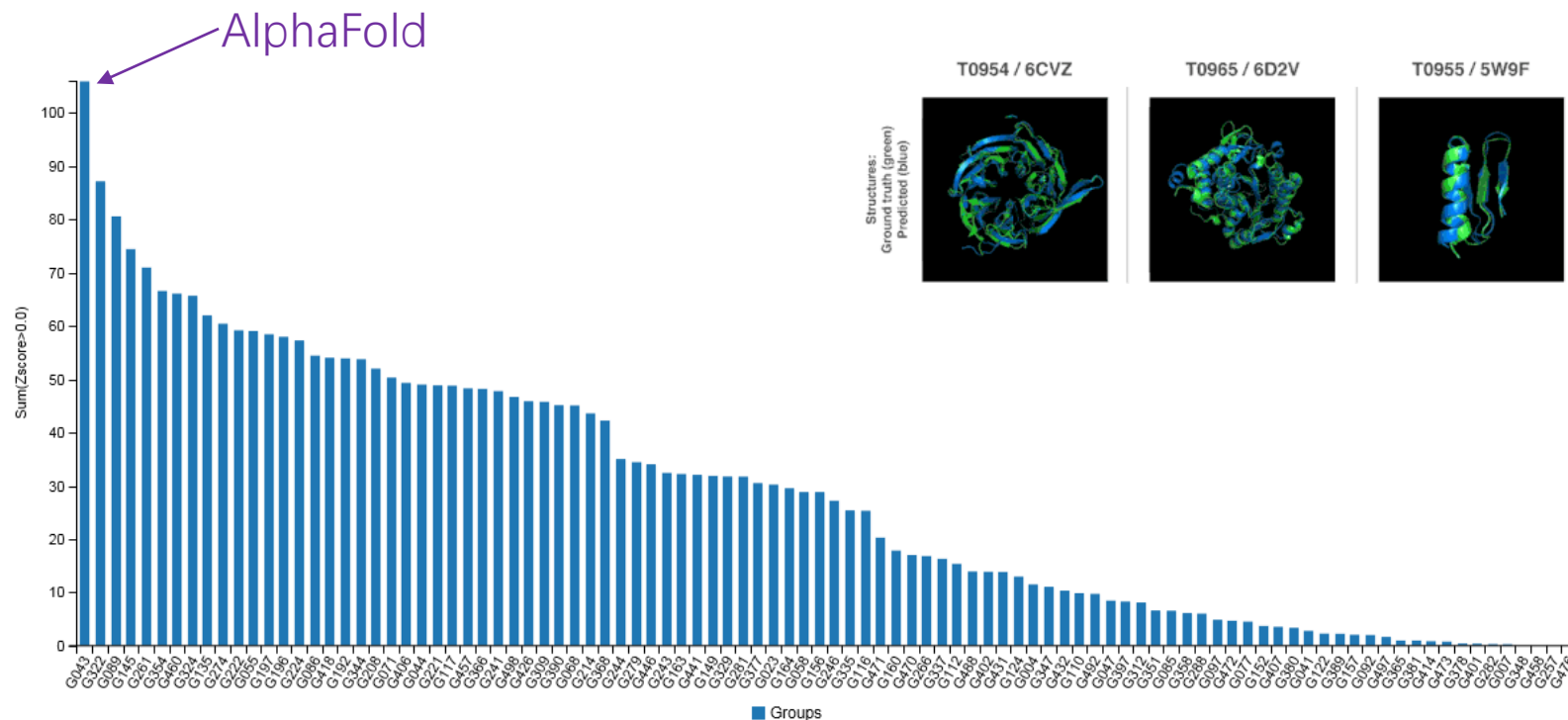
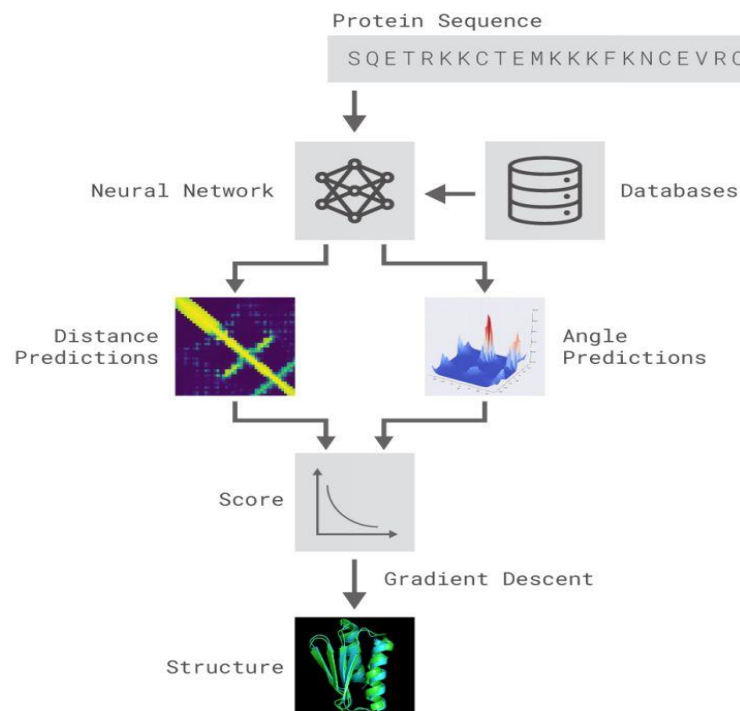


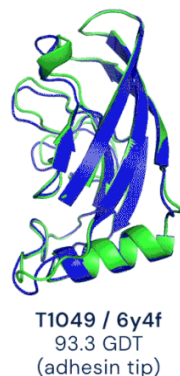
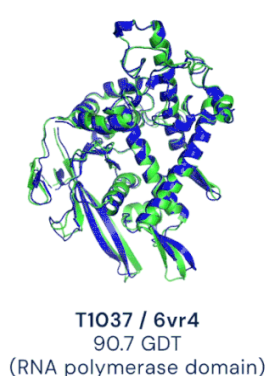
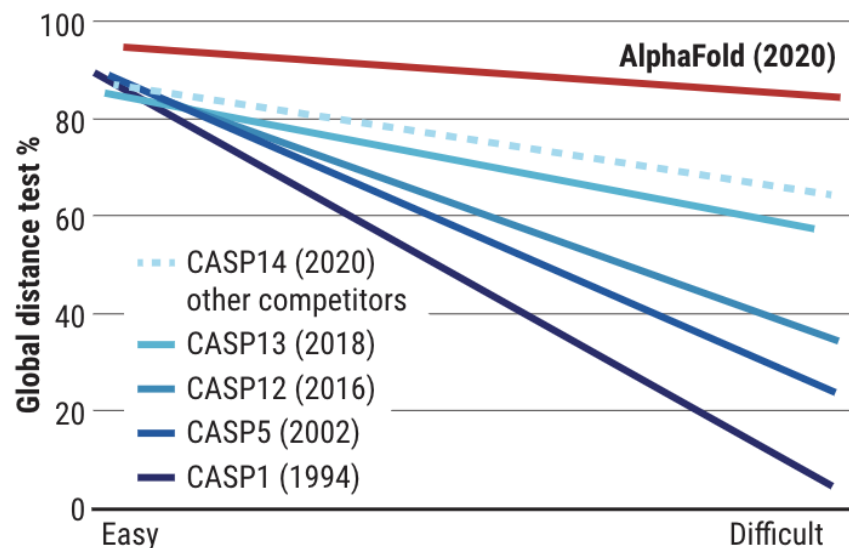
- CASP竞赛极大促进了蛋白质结构预测的发展。
- 从CASP11起，大数据和AI引领蛋白结构预测取得新进展。

历届CASP竞赛的表现



AlphaFold的模型框架





● Experimental result
● Computational prediction

“

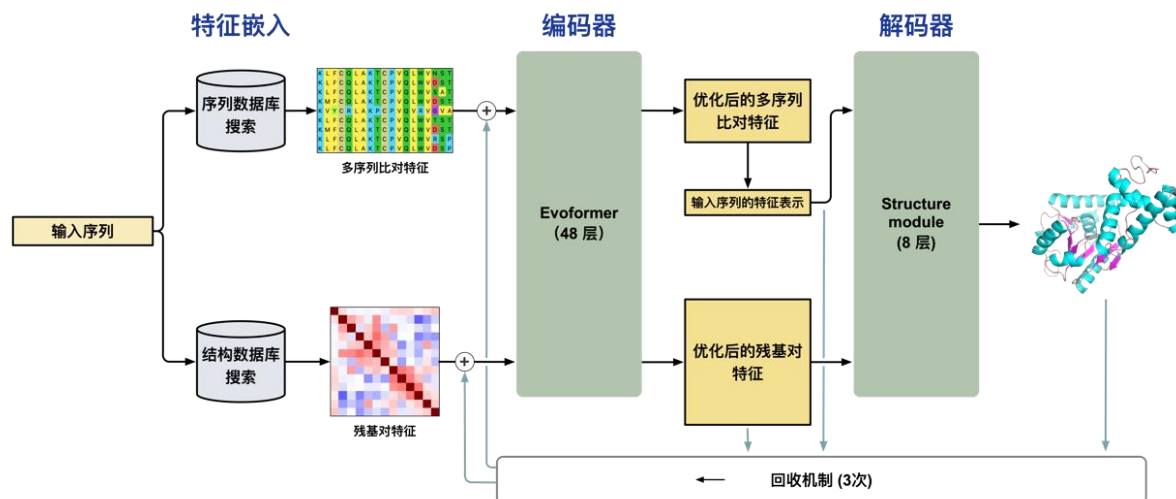
We have been stuck on this one problem – how do proteins fold up – for nearly 50 years. To see DeepMind produce a solution for this, having worked personally on this problem for so long and after so many stops and starts, wondering if we'd ever get there, is a very special moment.

PROFESSOR JOHN MOULT
CO-FOUNDER AND CHAIR OF CASP, UNIVERSITY OF MARYLAND



Jumper et al. (2021). *Nature* 596, 583–589.

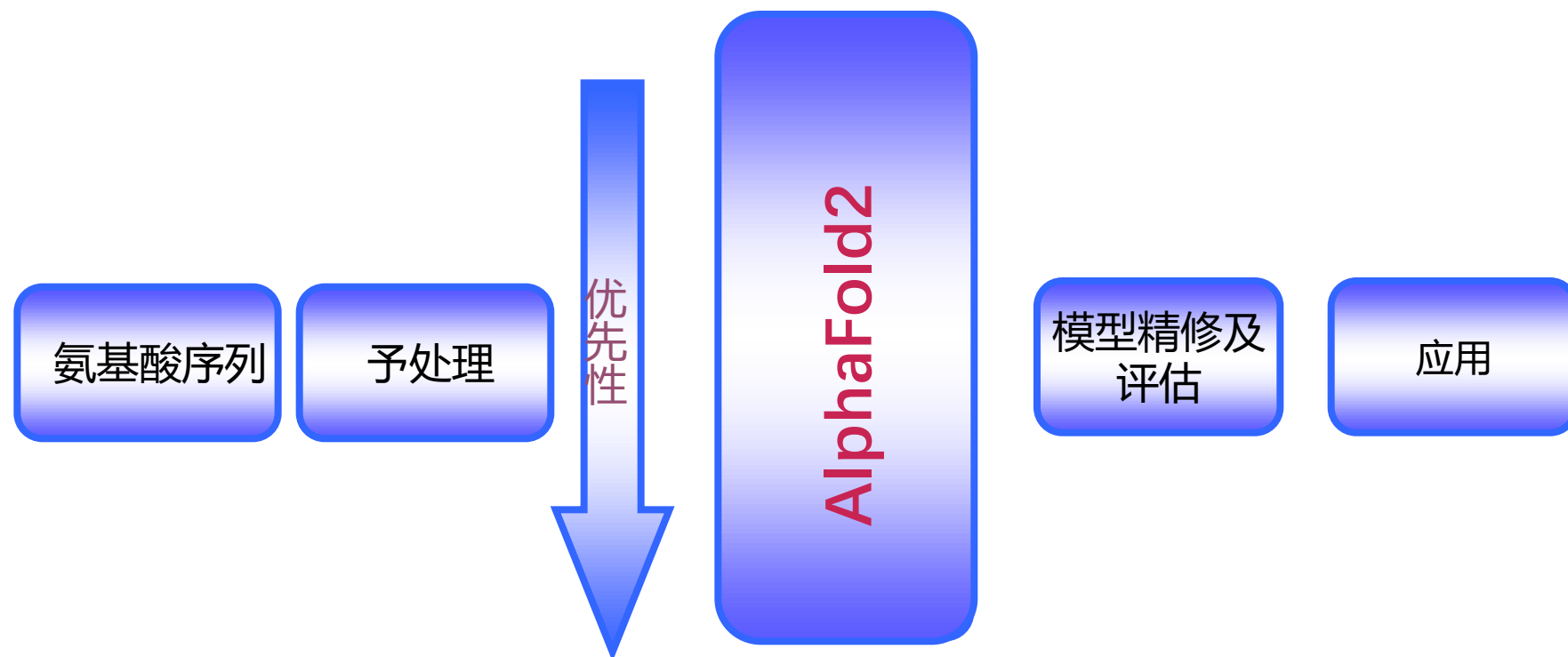
Tunyasuvunakool et al. (2021) *Nature*, 596, 591–596.



AlphaFold2 方法流程图

AlphaFold的获取方式

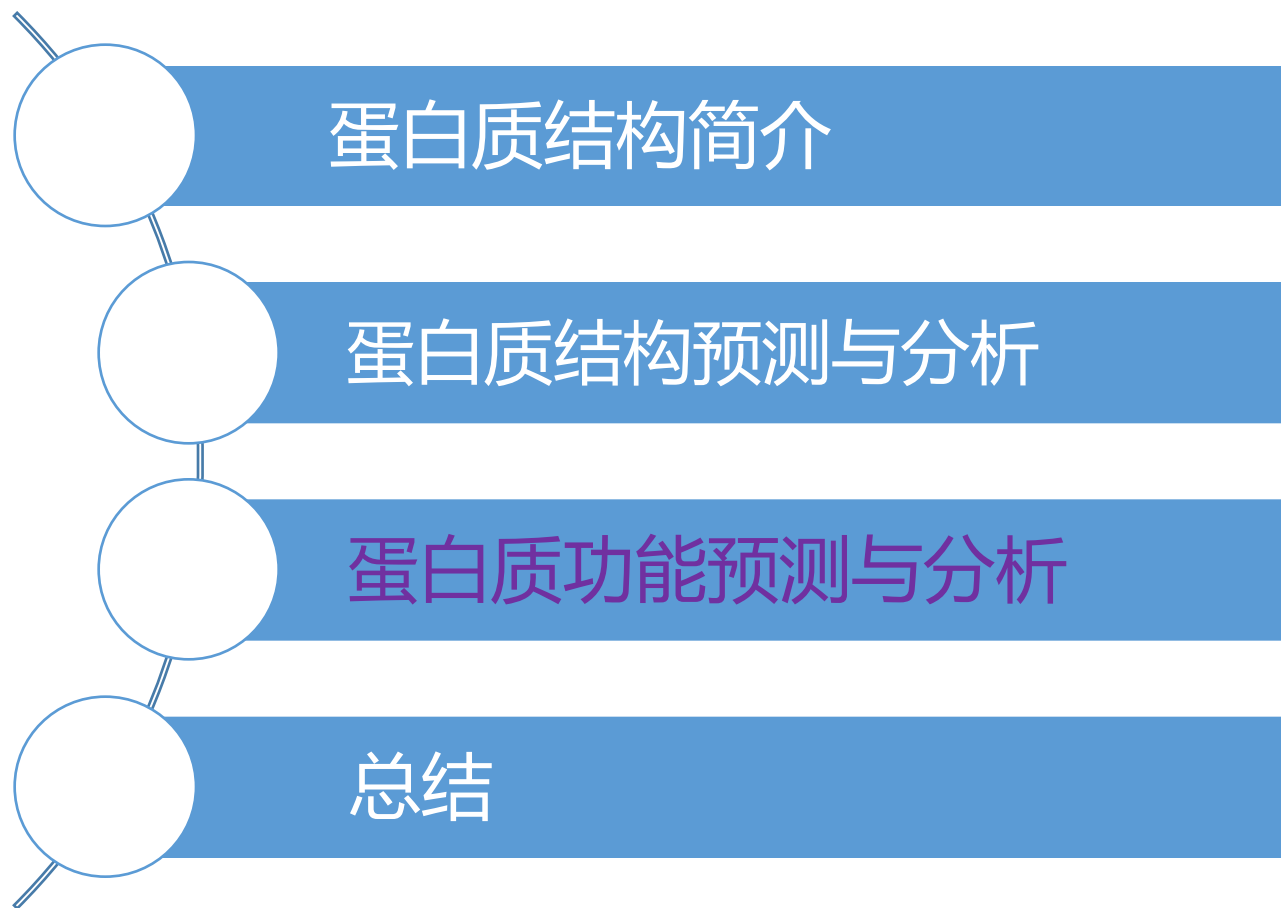
1. 数据库查询 (<https://alphafold.ebi.ac.uk/>)
包括模式生物与常见引起疾病的寄生虫、真菌和细菌等48个物种
2. 使用在线预测服务器
<https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb>
3. 搭建本地预测服务（需要大量计算资源）
<https://github.com/deepmind/alphafold>



蛋白质结构数据库和结构显示软件

蛋白质序列比对和结构比对

AI {进化信息 + 能量} !



P49597 - P2C56_ARATH

Protein	Protein phosphatase 2C 56
Gene	ABI1
Organism	Arabidopsis thaliana (Mouse-ear cress)
Status	Reviewed - ●●●●● - Experimental evidence at protein level ⁱ

Functionⁱ

Key component and repressor of the abscisic acid (ABA) signaling pathway that regulates numerous ABA responses, such as stomatal closure, osmotic water permeability of the plasma membrane (Pos), drought-induced resistance and rhizogenesis, response to glucose, high light stress, seed germination and inhibition of vegetative growth. During the stomatal closure regulation, modulates the inward calcium-channel activity as well as the actin reorganization in guard cells in response to ABA. Involved in the resistance to the bacterial pathogen Pseudomonas syringae pv. tomato. Controls negatively fibrillin expression that is involved in mediating ABA-induced photoprotection. May be involved in ABA regulation. Plays a role in the Pro accumulation in response to reduced water availability (low water potential). Required for the ABA mediated inhibition of the ethylene-induced hyponastic growth. Involved in acquired thermotolerance of root growth and seedling survival. Activates the GSK2E/OST1 in response to ABA-dependent stimuli, especially in stomata closure regulation involving SLAC1. 49 Publications

Catalytic activityⁱ

[a protein]-serine/threonine phosphate + H₂O = [a protein]-serine/threonine + phosphate. 3 Publications

Cofactorⁱ

Mg²⁺ 2 Publications, Mn²⁺ 2 Publications

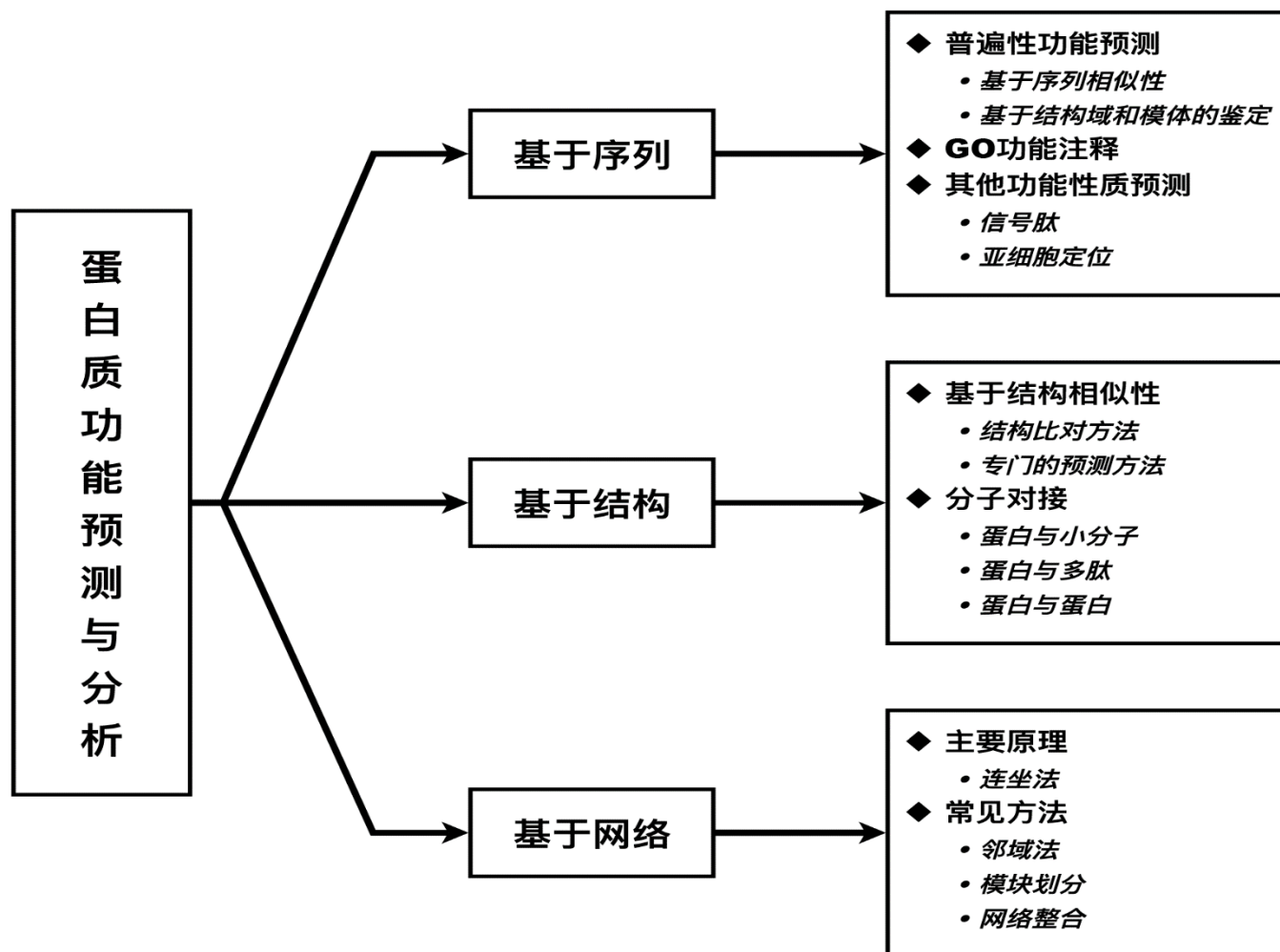
Note: Binds 2 magnesium or manganese ions per subunit. 2 Publications

信息量大，但难以归纳！

GO是由基因本体联合会建立的用来统一描述不同物种的基因及其产物的分子功能，生物学过程等的术语（term）集，其广泛应用于生物信息学领域。GO要包括三个分支：

- 生物学过程 (biological process), 描述基因产物可行使功能的生物学过程，如转录调控。
- 分子功能 (molecular function), 描述分子的生物化学功能，如转录因子活性。
- 细胞组件 (cellular component), 描述基因产物在细胞中的位置信息，如线粒体，叶绿体等。

- 广义上讲，与蛋白功能相关的生物信息学预测问题都属于蛋白质功能预测，包括蛋白质功能属性预测，功能家族预测，功能位点预测。
- 狭义上讲，功能预测通常指的是未知蛋白的GO注释预测。

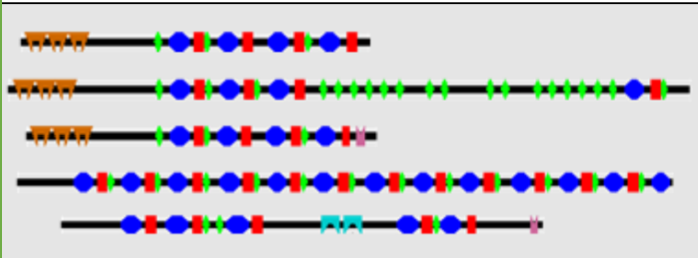


- 基本原理：蛋白质的序列相似，结构也相似，发挥的功能亦相似。
- 蛋白质行使功能往往和其含有的结构域和模体息息相关。
- 常用的序列分析工具（例如BLAST）就可以用来对蛋白进行功能预测。
- 可以通过一些常用的蛋白质数据库进行功能预测。

代表性的数据库及分析资源





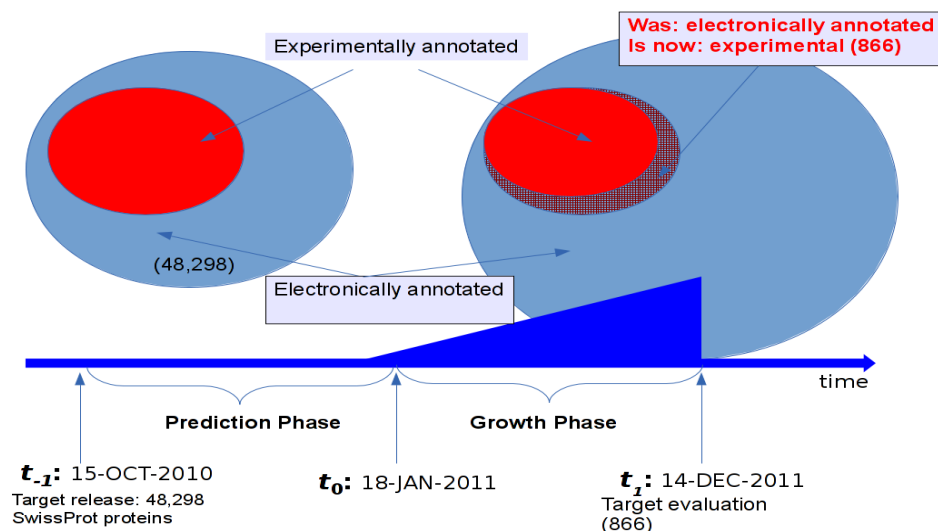


The Eukaryotic Linear Motif resource for Functional Sites in Proteins

CDD

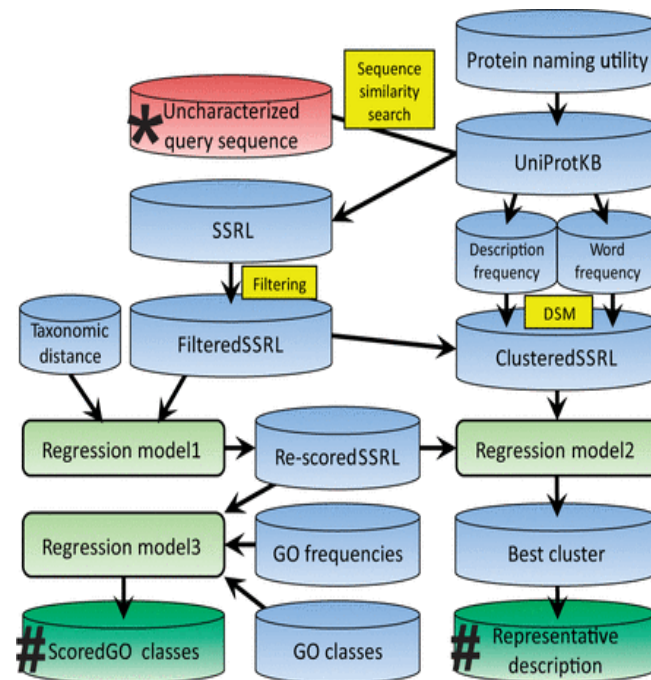
The Conserved Domain Database is a resource for the annotation of functional units in proteins. Its collection of domain models includes a set curated by NCBI, which utilizes 3D structure to provide insights into sequence/structure/function relationships.

CAFA蛋白质功能预测竞赛



以原有人类蛋白功能注释为训练集，以新获得的人类蛋白功能注释为测试集

- PANNZER (<http://ekhidna2.biocenter.helsinki.fi/sanspanz/>)
- 对于一个待测蛋白，通过UniProt进行同源性搜索，采用带权重的KNN算法找到若干同源蛋白，并对同源蛋白的GO词条进行富集和打分，最终实现对待测蛋白进行GO注释。
- PANNZER还基于同源蛋白的功能描述，通过聚类 and 统计，实现对待测蛋白的总体功能描述。



信号肽预测

SignalP 6.0 (<https://services.healthtech.dtu.dk/services/SignalP-6.0/>)

DTU Health Tech
Department of Health Technology
Research Education Collaboration Services and Products News About

Contact DTU

DTU HEALTH TECH > SERVICES AND PRODUCTS > BIOINFORMATIC SERVICES > SIGNALP-6.0

SHARE ON

SignalP - 6.0

Prediction of Signal Peptides and their cleavage sites in all domains of life

The SignalP 6.0 server predicts the presence of signal peptides and the location of their cleavage sites in proteins from Archaea, Gram-positive Bacteria, Gram-negative Bacteria and Eukarya. In Bacteria and Archaea, SignalP 6.0 can discriminate between five types of signal peptides:

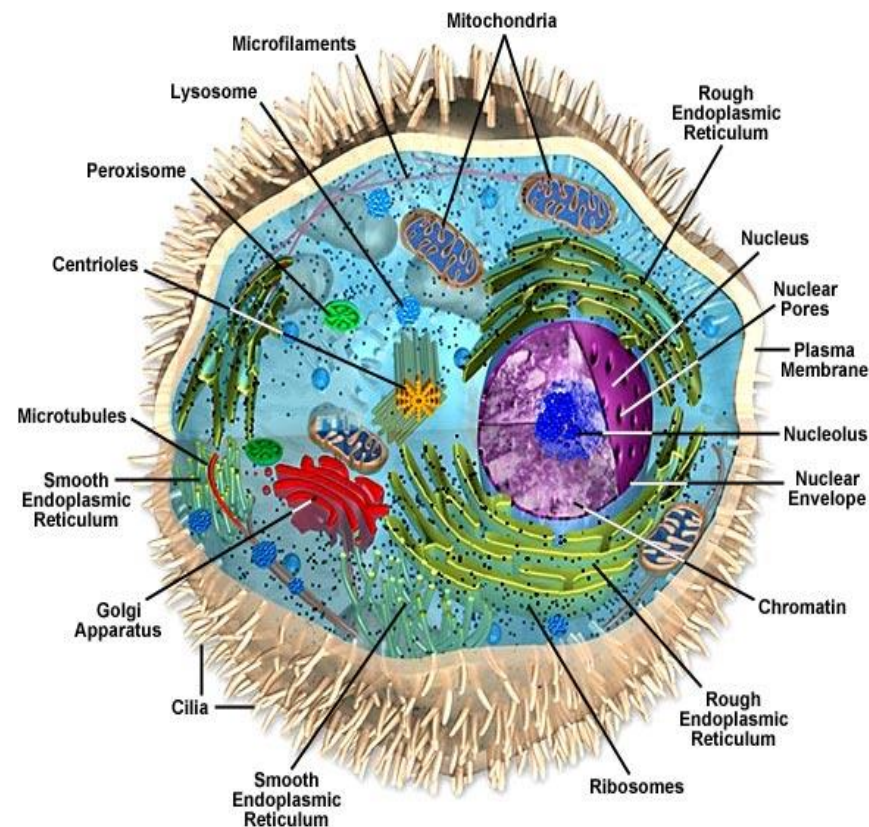
- Sec/SPI: "standard" secretory signal peptides transported by the Sec translocon and cleaved by Signal Peptidase I (*Lep*)
- Sec/SPII: lipoprotein signal peptides transported by the Sec translocon and cleaved by Signal Peptidase II (*Lsp*)
- Tat/SPI: Tat signal peptides transported by the Tat translocon and cleaved by Signal Peptidase I (*Lep*)
- Tat/SPII: Tat lipoprotein signal peptides transported by the Tat translocon and cleaved by Signal Peptidase II (*Lsp*)
- Sec/SPIII: Pilin and pilin-like signal peptides transported by the Sec translocon and cleaved by Signal Peptidase III (*PilD/PilB*)

Additionally, SignalP 6.0 predicts the regions of signal peptides. Depending on the type, the positions of *n*-, *h*- and *c*-regions as well as of other distinctive features are predicted. SignalP 6.0 is based on a [transformer protein language model](#) with a conditional random field for structured prediction.

Behind the Paper: Check out the [blog post about the SignalP 6.0 publication](#) in the Nature Portfolio Bioengineering Community.

History paper: Click here to read "[A Brief History of Protein Sorting Prediction](#)". The Protein Journal, 2019

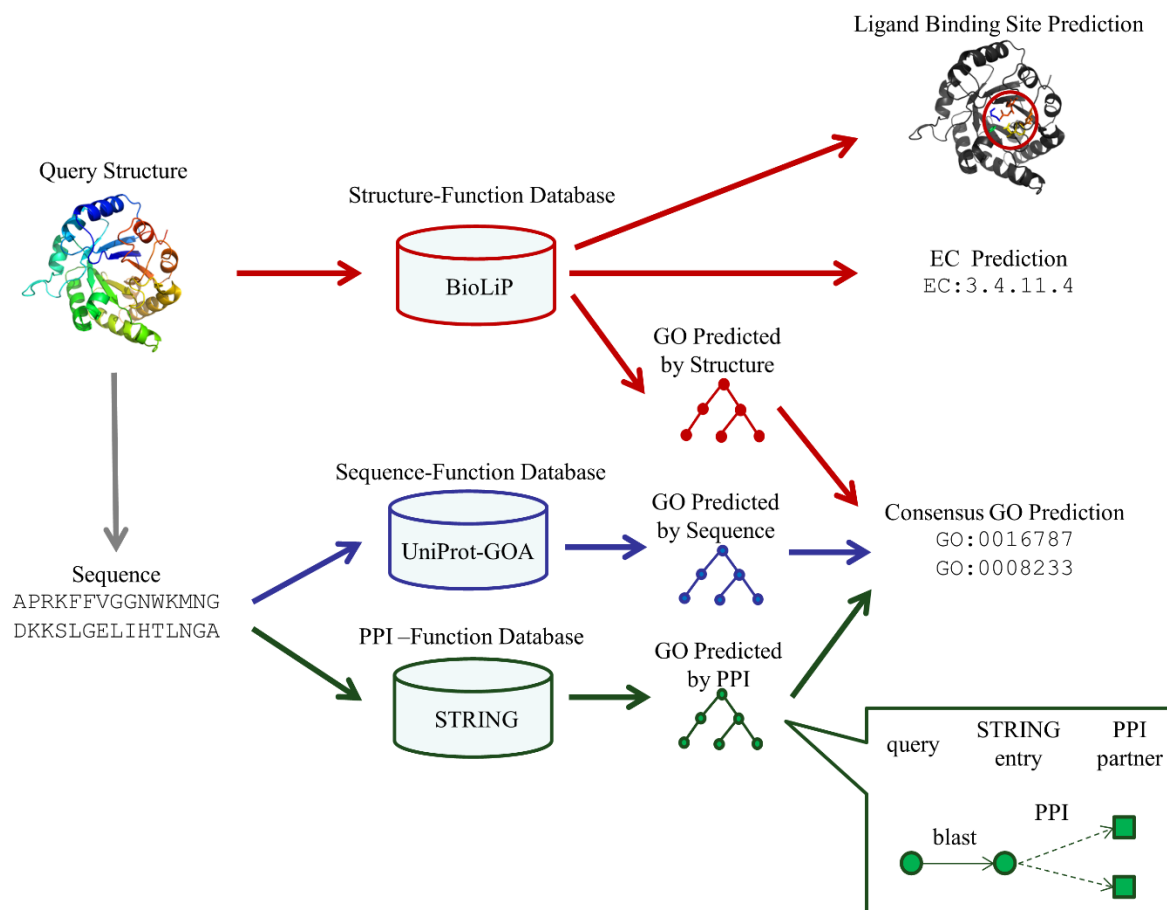
Eukaryotic proteins: Remember, the presence or absence of a signal peptide is not the whole story about the localization of a protein! If you want to find out more about the sorting of your eukaryotic proteins, try the protein subcellular localization predictor [DeepLoc](#). You may also want to check whether proteins with signal peptides have GPI anchors that keep them attached to the outer face of the plasma membrane using the predictor [NetGPi](#).

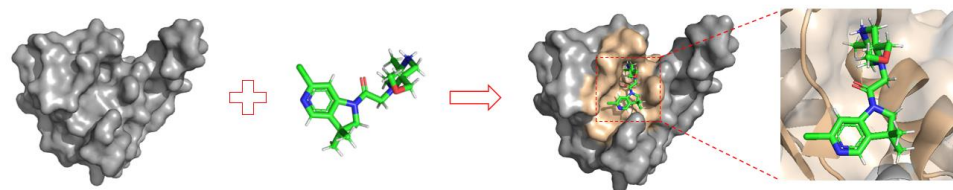


蛋白质亚细胞定位预测

代表性方法WoLF PSORT (<http://wolfpsort.org>)

- 两个蛋白（局部）结构相似，发挥的功能也有可能相似。
- 常用的结构比对方法可用来开展基于结构的功能注释。
- COFACTOR是一个代表性的基于结构的功能预测方法
- <http://zhanglab.ccmb.med.umich.edu/COFACTOR>
- 从蛋白空间结构出发，COFACTOR首先将待测蛋白结构比对（包括全局和局部结构比较）到功能已知的结构模板，实现待测蛋白的功能注释。





受体蛋白

配体小分子

分子对接结果

受体-配体分子对接的一般步骤：

1. 受体和配体分子处理
2. 寻找潜在的结合口袋
3. 旋转平移配体，生成一系列配体结合姿势 (pose); 对于柔性对接，配体本身的构象变化也要考虑
4. 对每个结合姿势的结合能进行估计
5. 对结合姿势聚类与排名

代表性的开源分子对接软件

AutoDock

home | downloads | resources | faqs & help | forum | contact

you are here: home

AutoDock
by [morris](#) — last modified 2013-02-27 12:43
Contributors: Sargis Dallakyan

Welcome!

- What is AutoDock?
- What is AutoDock Vina?
- How do I get started with AutoDock?
- What's new?
- What is AutoDockTools?
- Where is AutoDock used?
- Why use AutoDock?
- Run AutoDock on World Community Grid!
- How to cite AutoDock?
- More FAQs...

What is AutoDock?
AutoDock is a suite of automated docking tools. It is designed to predict how small molecules, known 3D structure.

[Garrett M. Morris](#)
[David S. Goodsell](#)
[Ruth Huey](#)
[William Lindstrom](#)
[William E. Hart](#)
[Scott Kurowski](#)
[Scott Halliday](#)
[Rik Belew](#)
[Arthur J. Olson](#)

OPEN ACCESS Freely available online

PLOS COMPUTATIONAL BIOLOGY

Education

Accessible High-Throughput Virtual Screening Molecular Docking Software for Students and Educators

Reed B. Jacob, Tim Andersen, Owen M. McDougal*

Boise State University, Boise, Idaho, United States of America

高通量虚拟筛选的基本原理

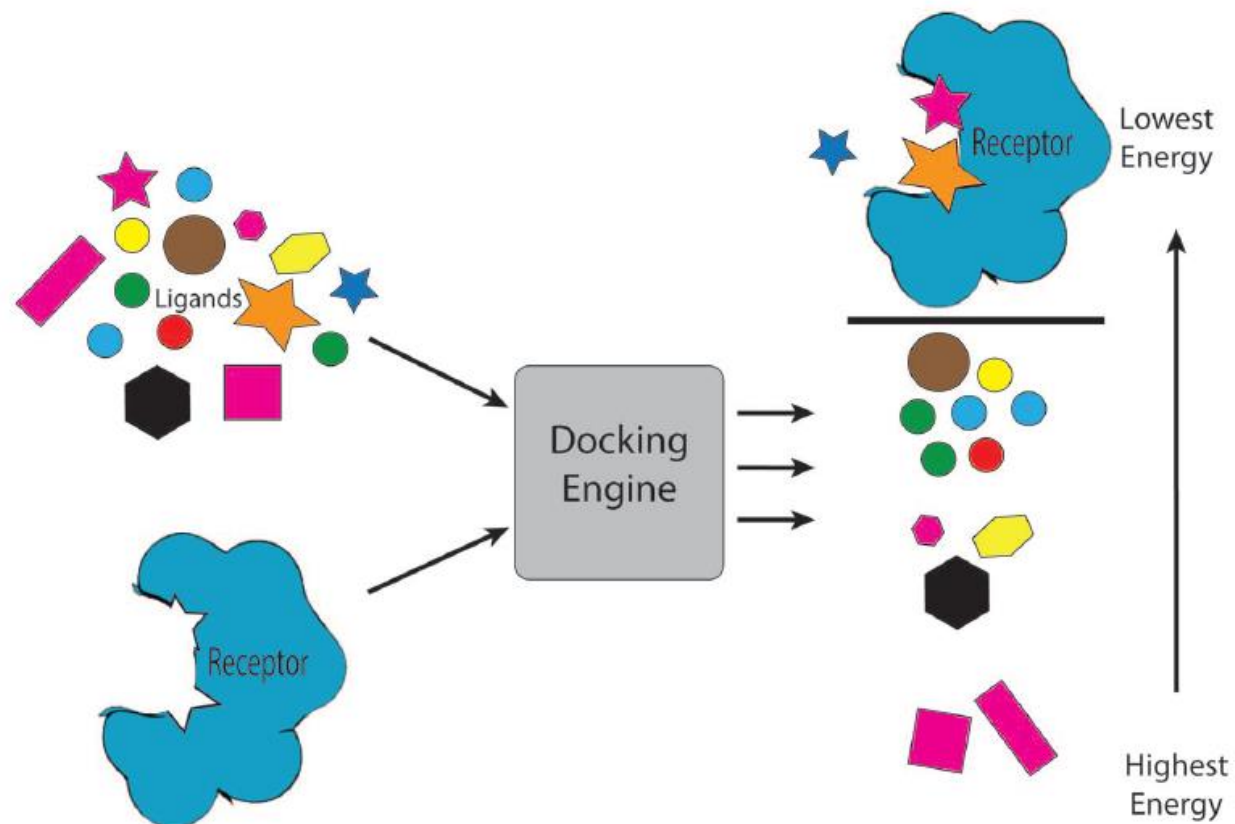
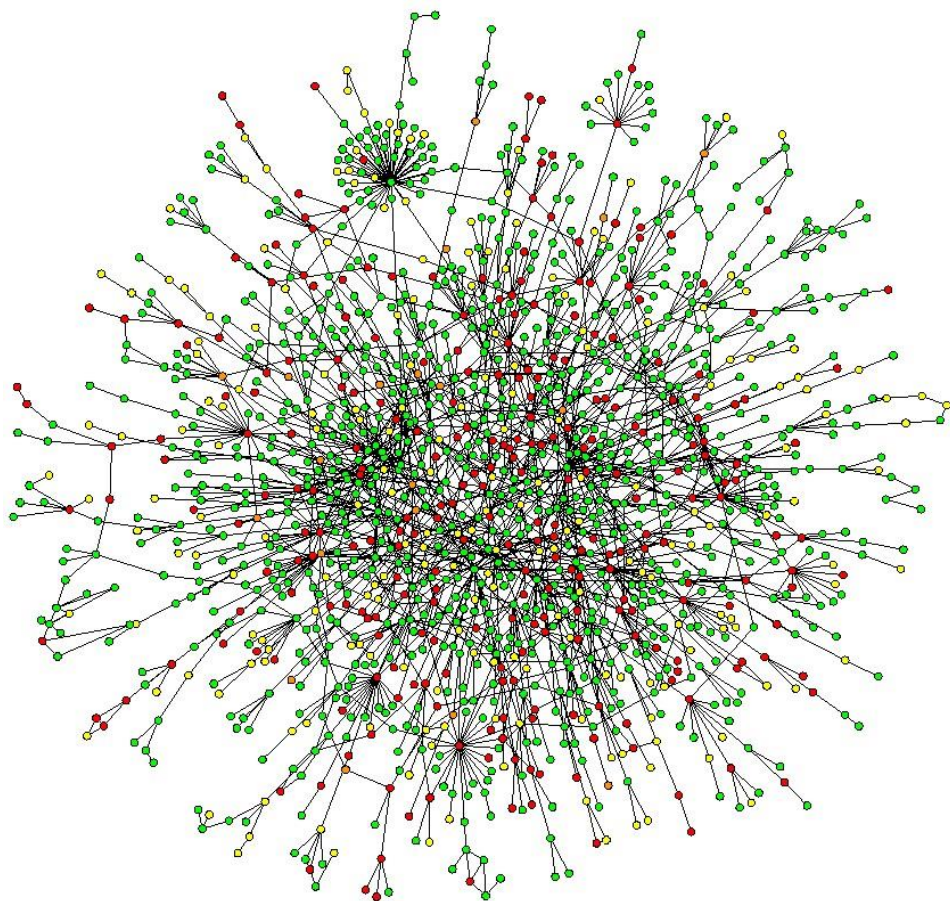


Figure 1. Depiction of high-throughput virtual screening: multiple ligands are docked to a receptor and ranked by energy estimate.
doi:10.1371/journal.pcbi.1002499.g001

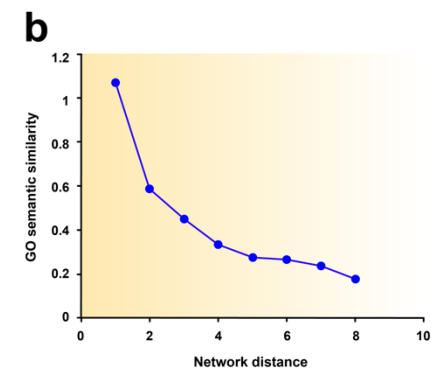
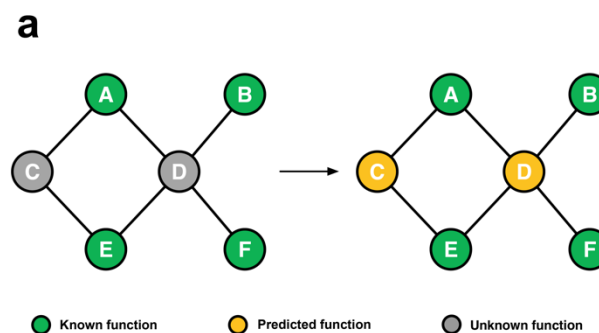


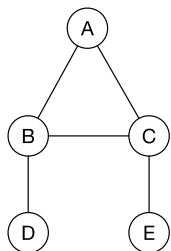
生物分子网络：酵母蛋白互作网络



- 高连接度的“枢纽”蛋白质通常在生物过程中起关键作用
- 互作的两个蛋白往往倾向于行使相似的生物学功能
- 网络模块化结构则有助于识别共同参与特定生物过程的蛋白质集合。

基于网络的蛋白功能预测主要原理是连坐法



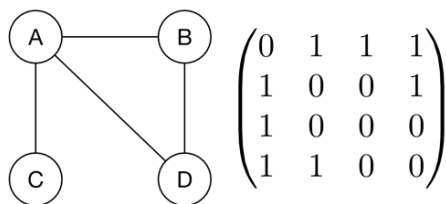


- 距离 (Distance)：两个节点间的最短路径上的边数。
- 直径 (Diameter)：网络中所有节点间的最大距离，用于评估网络的整体连接性。
- 邻域 (Neighborhood)：一个节点的邻域是与其直接相连的一组节点。
- 群 (clique)：一个完全连接的子图。
- 度 (degree)：一个节点的度指的是其相邻节点的数量。
- 密度 (density)：网络实际存在的边与可能存在的边之间的比例，用于评估网络的紧密程度。
- 聚类系数 (clustering coefficient)：一个节点的邻域的密度。

常见的预测方法

- 基于邻域的蛋白功能预测方法
- 基于网络模块划分的预测方法
 - ✓ 基于层次聚类的模块发现方法
 - ✓ 基于随机游走的模块发现方法
 - ✓ 基于模块度的模块发现方法
 - ✓ 基于图分割的方法
- 基于网络整合的蛋白功能预测方法

A. 一个简单的网络及其邻接矩阵



B. 第一步：增添自连接

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

C. 第二步：矩阵归一化（列）

$$\begin{pmatrix} 0.25 & 0.33 & 0.5 & 0.33 \\ 0.25 & 0.33 & 0 & 0.33 \\ 0.25 & 0 & 0.5 & 0 \\ 0.25 & 0.33 & 0 & 0.33 \end{pmatrix}$$

D. 第三步：矩阵扩展（矩阵平方）

$$\begin{pmatrix} 0.25 & 0.2175 & 0.25 & 0.2175 \\ 0.25 & 0.2175 & 0.125 & 0.2175 \\ 0.125 & 0.0825 & 0.25 & 0.0825 \\ 0.25 & 0.2175 & 0.125 & 0.2175 \end{pmatrix}$$

E. 第四步：矩阵膨胀（先对矩阵元素平方，再对矩阵每列进行归一化）

$$\begin{pmatrix} 0.0625 & 0.0473 & 0.0625 & 0.0473 \\ 0.0625 & 0.0473 & 0.0156 & 0.0473 \\ 0.0156 & 0.0068 & 0.0625 & 0.0068 \\ 0.0625 & 0.0473 & 0.0156 & 0.0473 \end{pmatrix}$$

$$\begin{pmatrix} 0.3077 & 0.3182 & 0.4 & 0.3182 \\ 0.3077 & 0.3182 & 0.1 & 0.3182 \\ 0.0769 & 0.0458 & 0.4 & 0.0458 \\ 0.3077 & 0.3182 & 0.1 & 0.3182 \end{pmatrix}$$

F. 重复第三、四步

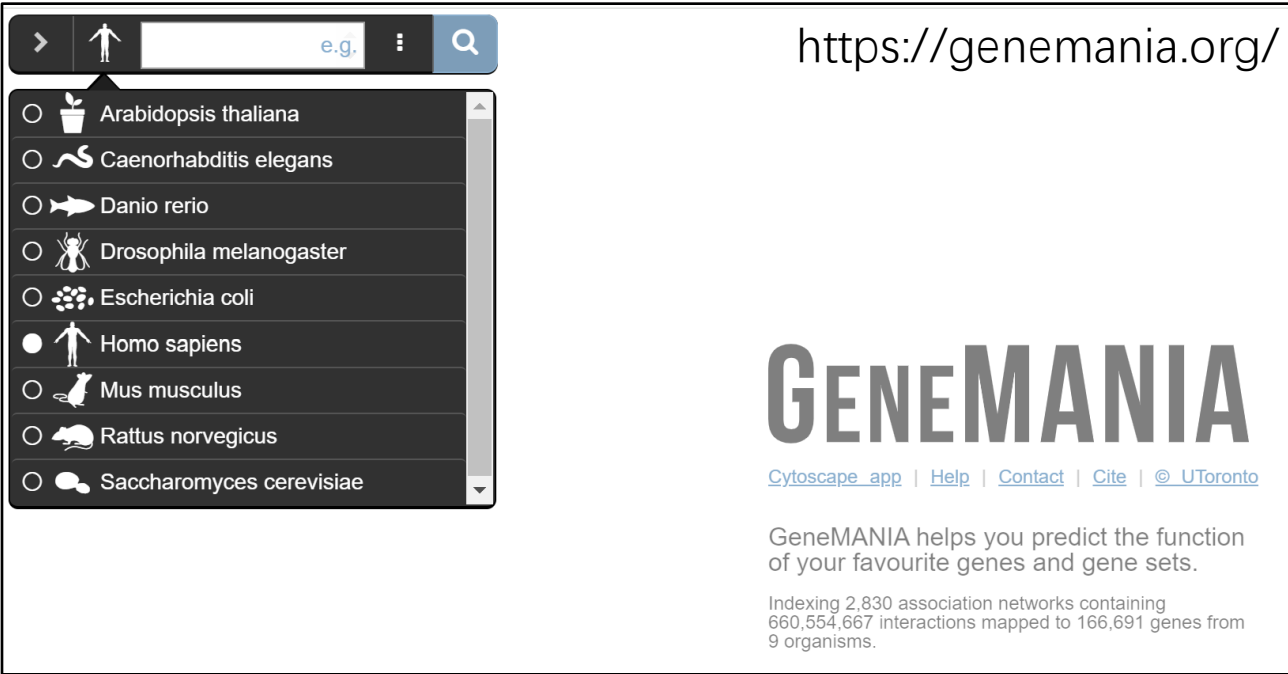
G. 最终收敛的矩阵

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

马尔可夫聚类算法（Markov Cluster Algorithm, MCL）是一种经典的基于随机游走的算法，其核心思想是模拟信息在网络中的流动，通过反复的扩展（Expansion）和膨胀（Inflation），加强模块内部的连接并削弱模块之间的连接，从而识别出网络中的模块。

- 将模块内已注释蛋白的功能分配给模块内的未注释蛋白。
- 进行富集分析，以确定模块内是否显著富集某些特定的功能或生物过程，再进行功能注释。

- 采用监督学习方法，通过线性回归学习每个网络的权重，创建加权整合网络。
- 该整合网络的邻接矩阵结合了多个网络的信息，更准确地反映了基因之间的功能关联性。
- GeneMANIA接着使用标签传播算法在整合网络上进行蛋白功能预测。



https://genemania.org/

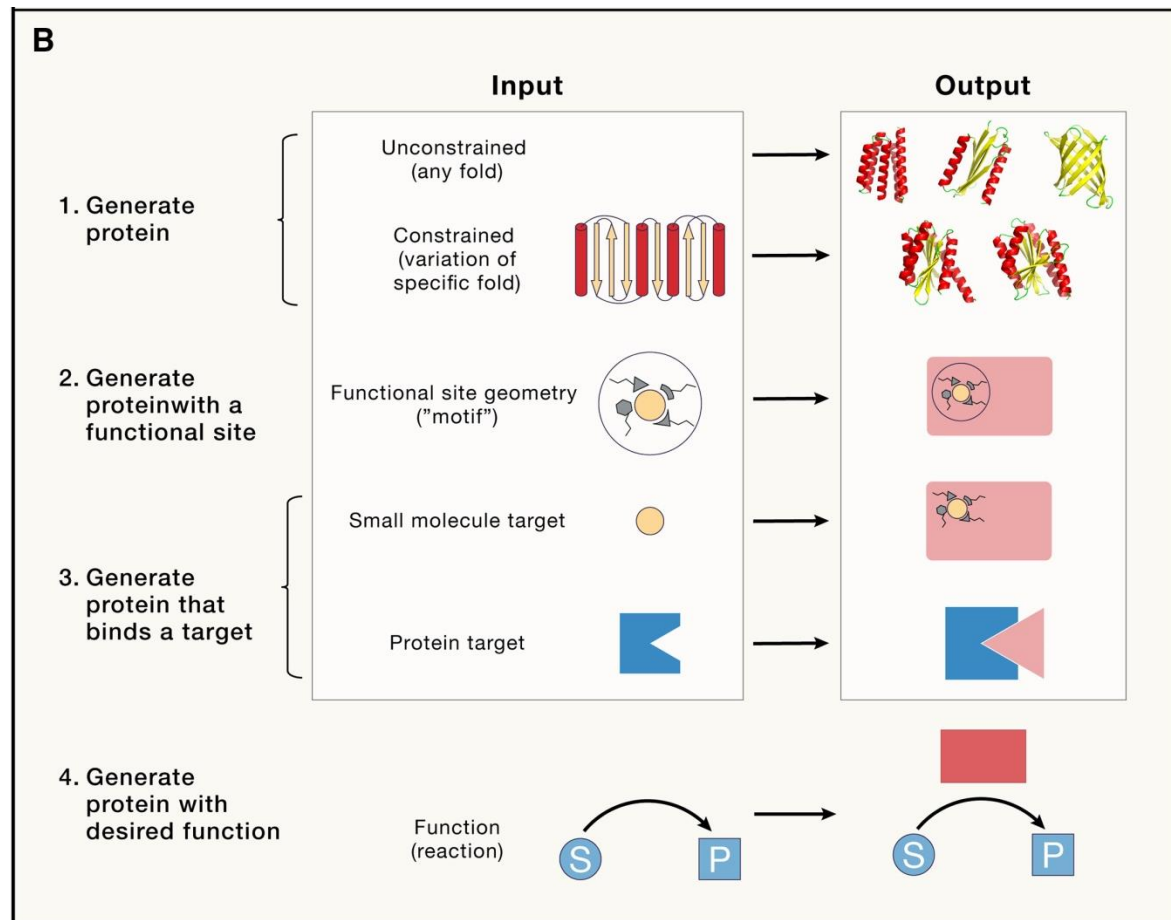
GENEMANIA

[Cytoscape app](#) | [Help](#) | [Contact](#) | [Cite](#) | [© UToronto](#)

GeneMANIA helps you predict the function of your favourite genes and gene sets.

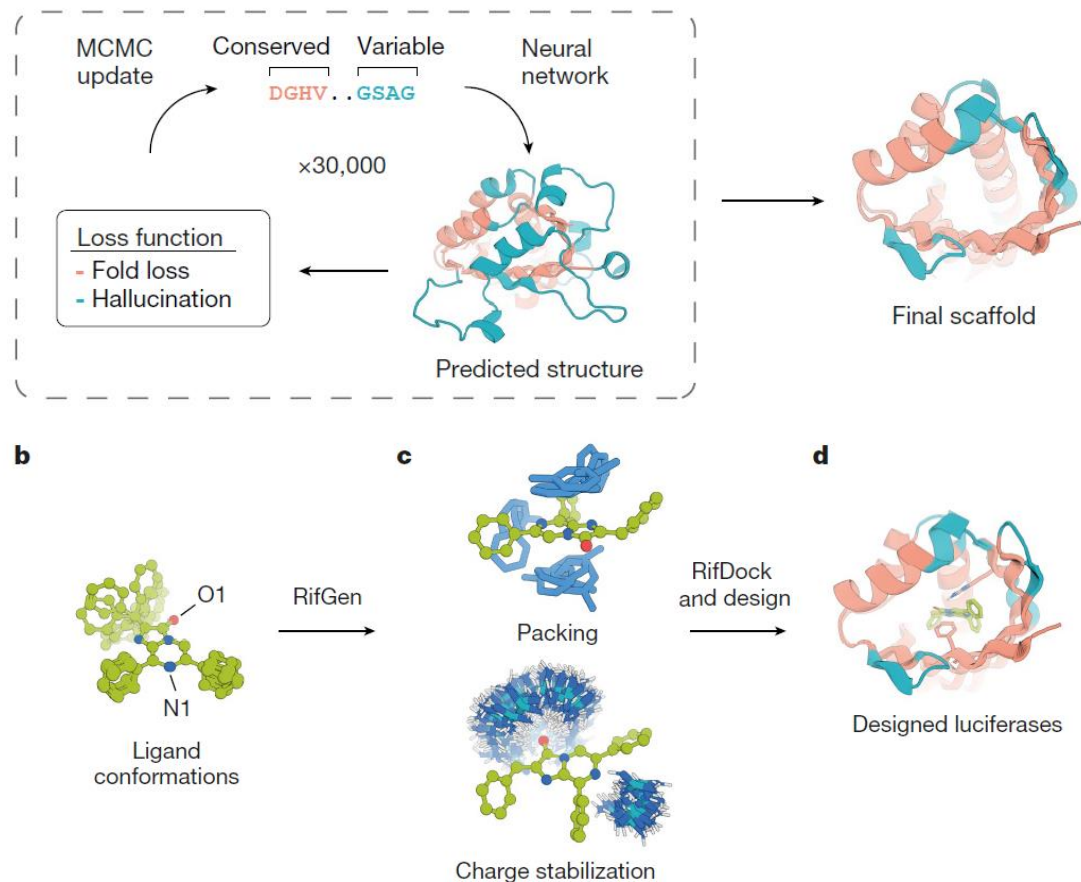
Indexing 2,830 association networks containing 660,554,667 interactions mapped to 166,691 genes from 9 organisms.

1. 基于机器学习的整合方法往往能提高蛋白质功能预测精度。
2. 蛋白质结构预测的突破性进展使得从结构出发预测蛋白质功能越来越重要。
3. 预训练蛋白质语言模型的应用使得许多蛋白质功能预测问题的精度得以提升。
4. 可用于蛋白质功能预测的大规模蛋白质语言模型ProteinChat已被开发。



Kortemme T. (2024) *De novo* protein design—From new structures to programmable functions. *Cell*, 187(3):526-44.

Baker团队成功利用全家族幻视（Family-wide Hallucination）设计出全新的荧光素酶LuxSit（Nature, 2023, 614(7949):774-80.）



- 选择荧光素发光基团-二苯基特拉嗪（DTZ）作为目标发光基团，力图设计能结合该基团并发光的蛋白质。
- 对接发现很多核转运因子2（NTF2）类似蛋白拥有和DTZ大小和形状相吻合的结合位点，因此选择NTF2超家族作为设计的目标拓扑结构。
- 通过全家族幻视方法生成上千个具有DTZ底物结合口袋样结构的、全新的、理想的蛋白质三维结构骨架及其相应的氨基酸序列。
- 随后，产生了DTZ的各种结构状态，并使用RifGen方法对每一种状态进行分析，计算其“旋转异构体相互作用场”（RIF）。使用RifDock方法将每一个DTZ状态及其相关的RIF同所设计的蛋白进行对接，最大化其蛋白-DTZ相互作用。序列优化后，筛选得到7648种候选蛋白。
- 最终实验获得3个有效的荧光素酶。其中，活性最强的荧光素酶命名为LuxSit（拉丁语“让光存在”）。

- 蛋白质结构组织层次
- PDB数据库及蛋白结构显示软件PyMOL的使用
- 蛋白质结构预测主要方法及原理
- 蛋白功能预测的主要方法及原理
- 强调了AI在蛋白质结构和功能预测中重要作用
- 强调了蛋白序列、结构与功能之间的关系

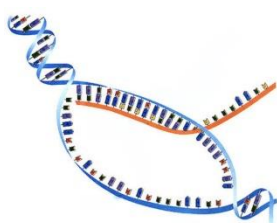
第五节 蛋白质与代谢组学（第5节）

朱峰

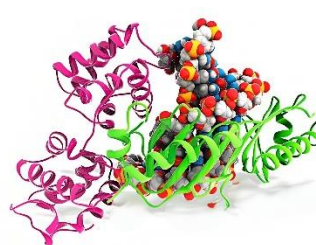
代谢组学 (metabolomics) 是研究处于特定时期（如不同发育状态等）或受到特殊扰动（如伤病、给药等）的生物体内所有代谢物组成及其变化的学科。其是对生物体内代谢产物的组成及其变化进行系统性研究的新兴组学领域，是系统生物学的重要组成部分，是最接近生物表型的组学。



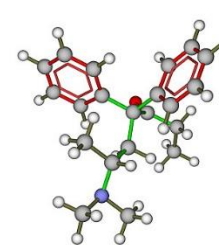
基因



转录本



蛋白

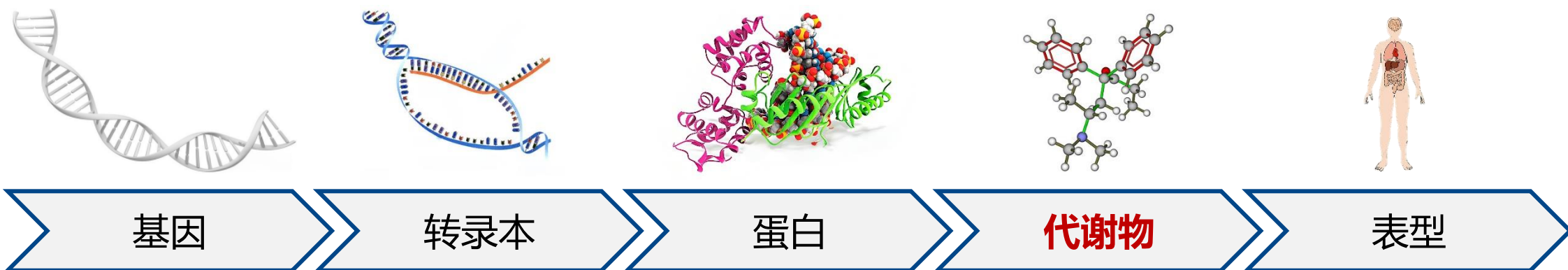


代谢物



表型

主要针对的是**相对分子质量小于1000 Dalton的内源性分子**，包括糖、脂类、氨基酸、芳香烃等。代谢物在不同物种中数量存在差异，一般动物约有2500种、微生物约有1500种，植物约有20万种。



作为基因表达的下游产物，**代谢物在生理过程中的作用不可忽视。**

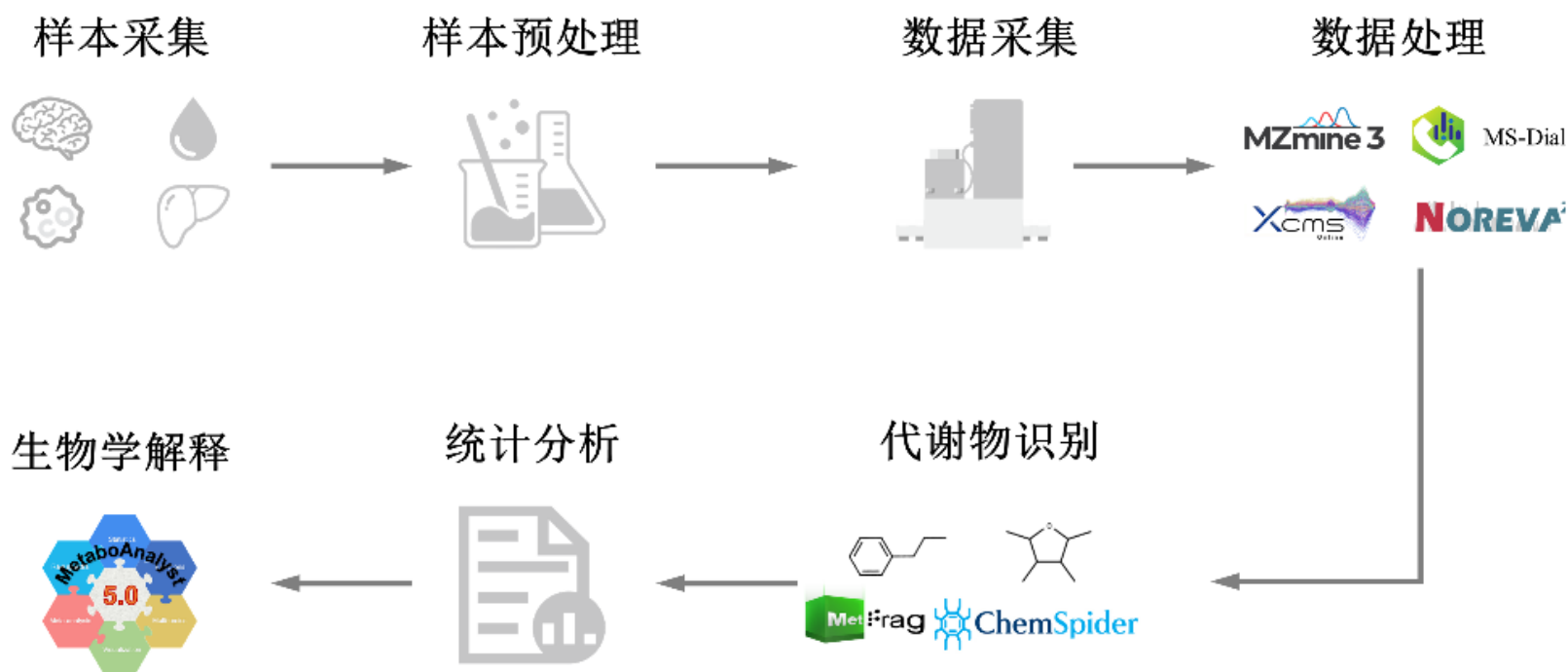
生物机体中的生理活动由基因、蛋白质以及代谢物共同调节，而**上游大分子（核酸、蛋白质等）的变化大多最终会体现在代谢物层面**，如细胞间通讯、激素调控、受体作用效应等。值得注意的是，虽然代谢物处于基因调控网络和蛋白互作网络的下游，但其也能够通过**与上游基因、蛋白进行相互作用**，实现对上游生命活动网络的反馈，参与众多生理过程的调控。

代谢组学相关的研究起源于二十世纪70年代的代谢谱分析，而**代谢组学的概念在20世纪九十年代首次被英国帝国理工大学Jeremy Nicholson教授提出**。那时代谢组学还被称为metabonomics，而不是metabolomics。



Jeremy Nicholson 教授

Metabonomics: the quantitative measurement of the dynamic multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification



代谢组学研究的主要环节

1. 样本采集
2. 样本预处理
3. 数据采集
4. 数据处理
5. 代谢物识别
6. 统计分析
7. 生物学解释

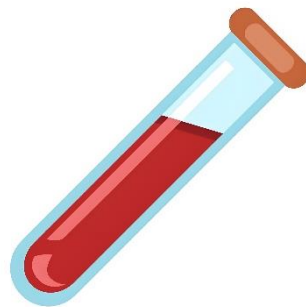
代谢组学研究的主要样本类型

由于代谢物能够随着体液被排出机体，因此**代谢组学样本的获取相较其他组学更加便捷**。**尿液、血液、动物粪便提取物、植物组织提取物**能够快速获取的生物体液是代谢组学研究中最常用的样本。正是因为代谢组学研究的生物样本易获取，其在医疗诊断领域具有极大的应用前景。此外，代谢组学研究的对象**也可以是各种临床上或动植物实验中取得的实体组织样本等较复杂的生物样本**。

最常用的
样本类型



尿液



血液



动物粪便提取物



植物组织提取物

样本制备注意事项

- ❑ 收集代谢组学样本的过程需要进行严格的把控。收集样本时应该根据实验设计综合考虑采样时间、样本保存条件以及保存时间等实验因素的一致性。在收集临床生物样本时，还有必要尽量保证实验组与对照组之间的年龄、性别、膳食、生活作息、病史、用药状况等多方面因素的一致性。这些与实验目的无关的外在因素均可能导致代谢组学数据的变化，影响最终的实验结果。
- ❑ 生物样本中的代谢物容易受到环境中众多因素的影响（如由于样本中残留酶活性而降解原有代谢物、产生新的代谢物），因此通常需要对样本尽快进行淬灭，并且进行合适的提取。
- ❑ 不同的代谢组学技术平台的样本提取方法不同，需根据实际使用的技术平台以及实验目的来确定使用的提取方法。

最常用的样品分析平台同分为两类别：

基于核磁共振光谱（NMR）的技术平台、基于质谱（MS）的技术平台

研究者应当根据样本性质、实验目的以及目标代谢物的理化性质，选择合适的实验技术平台。



基于 NMR 的技术平台

优点：

无代谢物相关的偏向性；
无复杂的前处理；
可实时动态检测；
不会损伤样本；
不会改变样本结构、性质。

缺点：

灵敏度低；
代谢物身份识别困难。



基于 MS 的技术平台

优点：

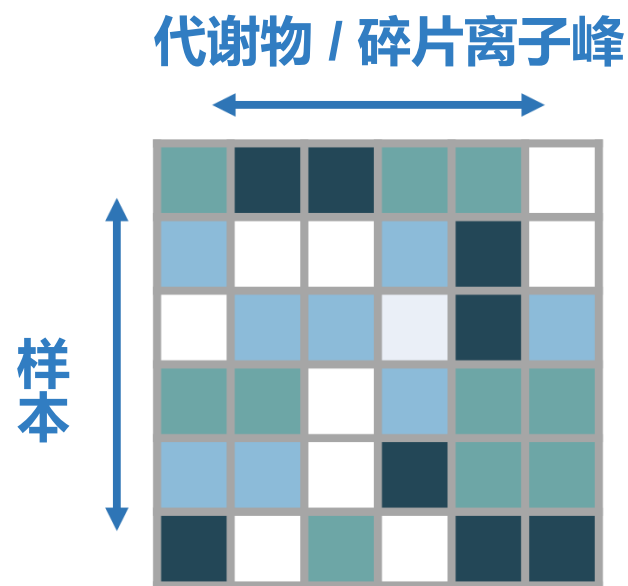
极高的灵敏度；
动态范围较宽；
代谢物身份识别较易。

缺点：

会损伤样本；
无法实时动态检测。

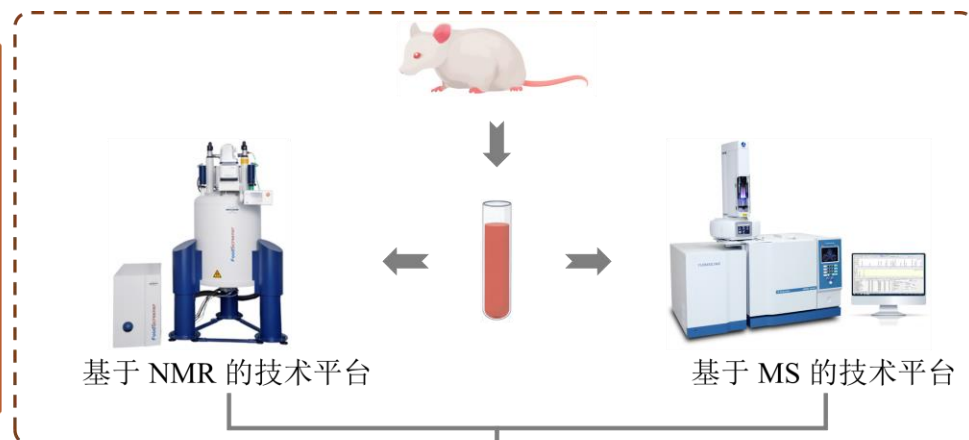
最常用的基于MS的技术平台包括**液相色谱-质谱联用** (liquid chromatography-mass spectroscopy, **LC-MS**) 技术平台以及**气相色谱-质谱联用** (gas chromatography-mass spectroscopy, **GC-MS**) 技术平台。

- 液质联用 (LC-MS)，是一种广泛使用的物质分离分析技术，具有灵敏度较高、动态范围较宽、无需衍生化等优点。LC-MS通常被用于GC-MS难以分析的强极性、难挥发、热不稳定性代谢物的检测分析。
- 气质联用 (GC-MS)，主要用于挥发性物质的代谢组学研究，得益于有NIST数据库提供相对完善的小分子EI-MS质谱信息，在代谢物身份识别方面较为便利，目前已经在植物样本代谢组学研究中得到广泛的应用。在动物样本代谢组学研究中，由于动物的内源性代谢物多为不挥发性代谢物，很大程度上限制了GC-MS的使用。但值得注意的是，随着衍生化方法的创新，GC-MS也逐渐在动物样本的代谢组学研究中得到广泛应用。与LC-MS相比，GC-MS所检测的化合物主要集中在极性程度低，分子量相对小的代谢物 (<300Da) 。

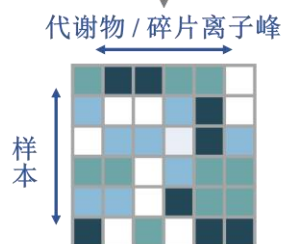


通常，基于NMR技术平台获取的原始数据通过代谢物注释（annotation）获得数据矩阵（行为样本，列为代谢物），而基于质谱技术平台获取的原始数据通过峰检测（peak detection）获取**数据矩阵**（行为样本，列为代谢物碎片离子峰）。

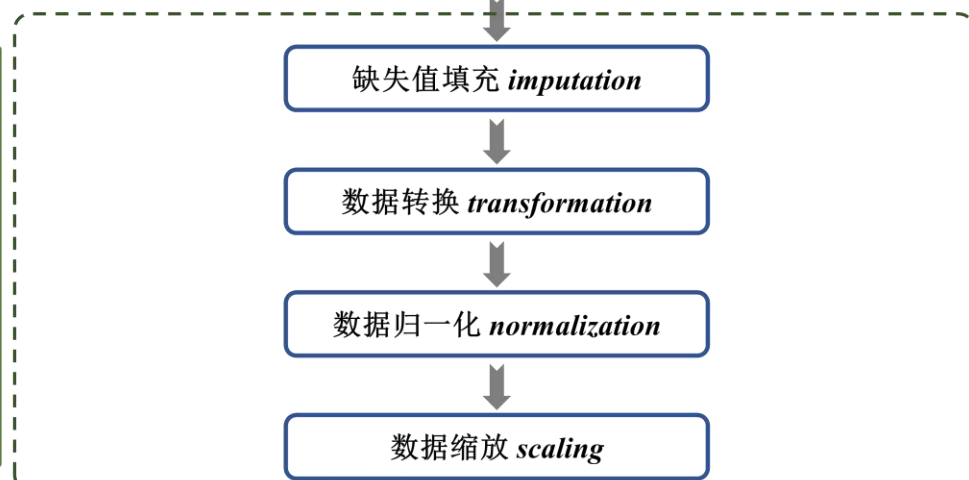
数据获取



数据矩阵



数据处理



代谢组学检测技术能够运行的样本数、检测到的代谢物数量飞速增加。代谢组学数据的复杂性也随之快速提高，有必要在后续分析之前进行完善的数据处理，以提供后续数据分析的可靠性。

针对所得数据矩阵的**常用的数据处理流程**包括**缺失值填充** (imputation)、**数据转换** (transformation)、**数据归一化** (normalization)、**数据缩放** (scaling)。

1. 缺失值填充 (imputation)

代谢组学实验中，部分代谢物产生的**信号峰质量较差或者代谢物含量较低而无法被检测到**，导致数据矩阵中出现相应的缺失值。

常用方法：均值替换、中值替换、最近邻替换等

2. 数据转换 (transformation)

实际数据矩阵中的特征分布可能并不是正态分布，而是偏态分布，需要在后续分析前使其呈现近似正态分布。此外，数据经常会受到来自各种来源的异方差噪音的影响，而分析时通常假设噪音是同方差的。**对数据矩阵进行数据转换就是为了纠正异方差和偏态分布**，为后续分析提供便利

常用方法：对数变换、方根变换等

3. 数据归一化 (normalization)

数据归一化是为了消除样本间不需要的变异（例如不同的样本液体浓度等），并允许对样本进行定量比较。

常用方法：总面积归一化法、概率商归一化法等。也有部分归一化方法不需要计算归一化系数，例如分位数归一化方法（Quantile normalization）等

4. 数据缩放 (scaling)

在数据矩阵中，丰度较高的代谢物表现出较高的数值，在样本间显示出更大的差异。后续分析中采用的部分方法基于方差进行分析，这可能会导致丰度较高的代谢物被认为是关键代谢物，而部分在生物学上有重要意义但丰度较低的代谢物则被忽视。因此，有必要对数据进行缩放，以平衡代谢物间的表达值差异。

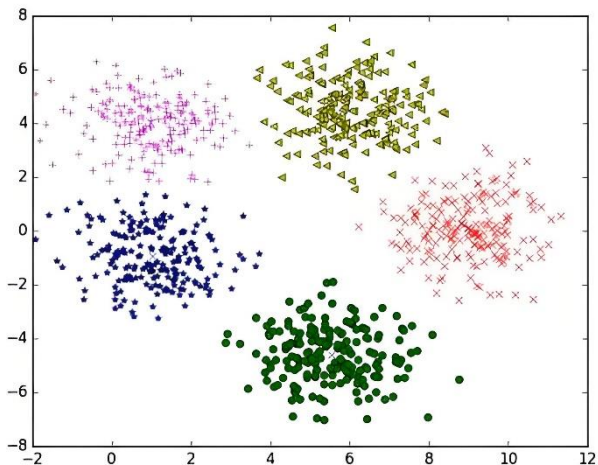
常用方法：自动缩放（auto-scaling）、范围缩放（range scaling）等

常见的数据处理工具：

- ❑ **XCMS**：XCMS是一个用于代谢组学数据处理和分析的流行开源软件包。它提供了用于峰识别、对齐、配准和归一化的功能，可以处理质谱和色谱数据。
- ❑ **MetaboAnalyst**：MetaboAnalyst是一个集成的在线平台，用于代谢组学数据处理、统计分析和功能解释。它提供了丰富的工具和算法，包括数据处理、多变量分析、通路富集分析等[15]。
- ❑ **MZmine**：MZmine是一个用于质谱数据处理和分析的开源软件。它支持从原始质谱数据中进行特征提取、峰识别和定量分析，还提供了工作流的创建和自定义功能。
- ❑ **MS-Dial**：MS-Dial是一个用于代谢组学数据处理和分析的开源软件工具。它专门设计用于处理液相色谱-质谱联用（LC-MS）和气相色谱-质谱联用（GC-MS）数据，旨在帮助研究人员从复杂的代谢组学数据中提取有关样品中代谢物的信息。
- ❑ **NOREVA**：NOREVA是一个专业用于代谢组学数据处理的工具。它支持数百种不同的代谢组学数据处理流程，并能够对这些数据处理流程的性能进行评估。

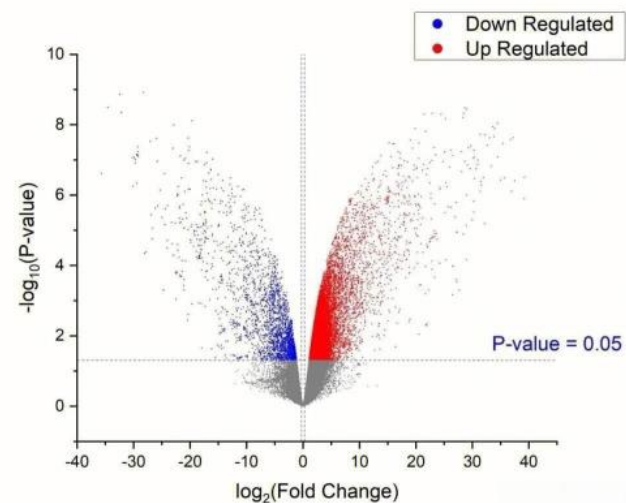
聚类分析:

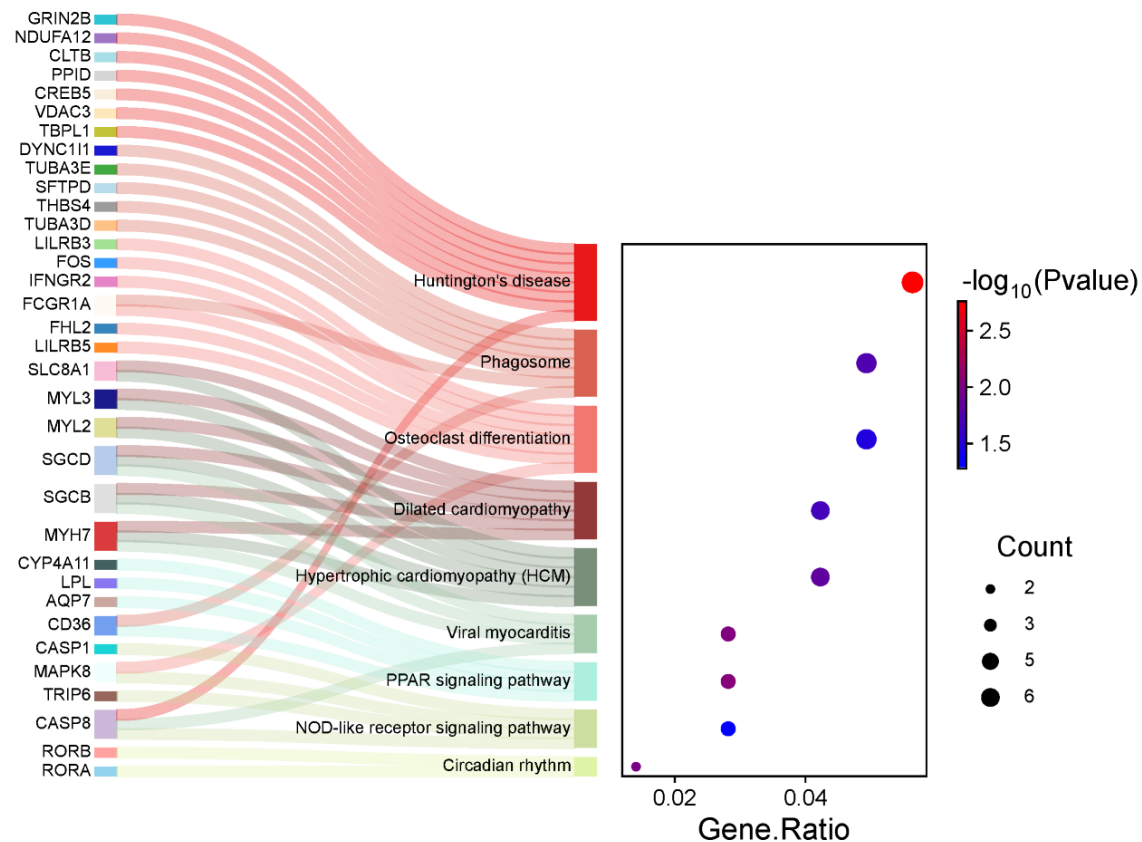
聚类分析常被用于对经过数据处理的代谢组数据进行数据挖掘。它可用于对一组对象进行分组，其中相似的对象比那些关系较远的对象分组得更紧密。



差异分析:

通过差异分析，研究者能够快速找到不同组别中的差异代谢物，例如疾病与健康状态下的关键生物标志物。通常使用t-test、ANOVA检验，识别出不同组别之间的差异代谢物，也可以结合多元统计分析的结果对关键代谢物进行识别。





通路分析：

代谢组学的通路分析是对代谢物进行功能注释和关联分析，以揭示代谢物之间的相互关系以及其参与的生物过程和代谢通路。代谢通路是一系列相互作用的化学反应，负责维持生物体内代谢平衡和能量转化。通过通路分析，可以识别出与特定生物过程或疾病相关的代谢通路，并深入了解其调控机制。通路分析有助于深入理解代谢物在生物体内的作用和调控机制。通常基于KEGG数据库进行代谢组数据的通路分析。

1. **版权声明：**本PPT及其所有内容（以下简称“本PPT”）仅用于教育和教学用途，版权归属于本PPT作者。
2. **使用要求：**任何使用本PPT的行为均须遵守以下条件：
 - 1) **致谢和标注：**若部分或全部使用本PPT的内容，请在使用内容的适当位置标注出处，并致谢本PPT作者。
 - 2) **修改和再分发：**未经作者书面许可，不得对本PPT进行修改或再分发。
3. **禁止商业化使用：**严禁将本PPT用于任何形式的商业化用途，包括但不限于：
 - 1) 通过网络或其他途径进行付费使用或分发；
 - 2) 在商业培训、广告或其他商业活动中使用本PPT的内容。
4. **法律责任：**任何违反上述条款的行为，作者保留追究法律责任的权利，包括但不限于：
 - 1) 要求停止侵权行为；
 - 2) 追究侵权使用者的经济赔偿责任。
5. **其他规定：**
 - 1) 本使用条款的解释权归本PPT作者所有。
 - 2) 作者保留随时更新本使用条款的权利，更新后的条款将即时生效。

谢谢大家!