

# 生物信息学

## 蛋白质三维结构预测 2

# 蛋白质结构 --- 三级结构

PyMOL <http://www.pymol.org>

## 分子三维结构查看及分析软件PyMOL

PyMOL by Schrödinger

[DOWNLOAD](#) [SCREENSHOTS](#) [PRODUCTS](#) [SUPPORT](#) [CONTACT](#)

# Download PyMOL 2.3

Version 2.3.0 - Updated February 11th 2019 ([installation instructions](#))

For previous versions and Python 2.7 bundles, [see here](#).

**These bundles include Python 3.7**



Windows

EXE Installer



Windows

ZIP Archive



macOS

DMG Disk  
Image



Linux

TAR.BZ2  
Archive

# 蛋白质结构 --- 三级结构

## 分子三维结构查看及分析软件**VMD**

下载: <http://www.ks.uiuc.edu/Research/vmd/>

从PDB数据库下载任一个.pdb文件，用写字板打开，对照VMD显示出来的东东，熟悉一下.pdb文件格式。

# 蛋白质结构 --- 三级结构

THEORETICAL *and* COMPUTATIONAL  
BIOPHYSICS GROUP

Home

Research

Publications

Software

Instruction

News

Galleries

Facilities

About Us

Home

Overview

Publications

Research

Software

▶ VMD Molecular Graphics Viewer

▶ NAMD Molecular Dynamics Simulator

▶ BioCoRE Collaboratory Environment

▶ MD Service Suite

▶ Structural Biology Software Database

▶ Computational Facility

Outreach

VMD Mailing List

Download VMD

VMD Tutorials

VMD

Visual Molecular Dynamics

VMD is a molecular visualization program for displaying, animating, and analyzing large biomolecular systems using 3-D graphics and built-in scripting. VMD supports computers running MacOS X, Unix, or Windows, is distributed free of charge, and includes source code.  
([more details...](#))

Spotlight

VMD can now make movies easier than ever before, with the use of a [movie plugin](#) that takes care of the entire movie making process. The vmdmovie plugin generates one of several built-in movie types, according to user selectable options. Once preferences and selections are made, the movie generator takes control of VMD and takes care of the entire process, from the generation of individual movie frames using on-screen snapshots or ray tracers, image format conversion staging of the image data for compression, invocation of movie compressor programs, and final disk space cleanup and temporary file deletion. This makes the whole process of making movies much simpler for inexperienced users.

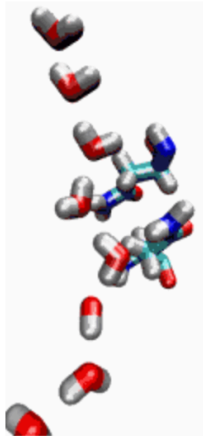
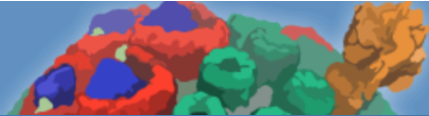
Other Spotlights

Overview

Molecular representations  
VMD plugin library  
Molecular file formats  
GPU-accelerated computing  
Interactive molecular dynamics

News and Announcements

Immersive Molecular Visualization with Omnidirectional Stereoscopic Ray Tracing and Remote Rendering, HPDAV 2016 **NEW**  
High Performance Molecular Visualization: In-Situ and Parallel Rendering with EGL, HPDAV 2016 **NEW**  
Evaluation of Emerging Energy-Efficient Heterogeneous Computing Platforms for Biomolecular



# 蛋白质结构分类数据库



**CATH**是一个按等级分类**PDB**中蛋白质结构的数据库。只有分辨率在**4埃**以内的晶体结构以及**NMR**结构才被分类。蛋白质结构分类结合了自动与手动两个过程。总共有四个水平上的分类：

**Class** → **Architecture** → **Topology** → **Homologous superfamily**

Go to: <http://www.cathdb.info/>

# 蛋白质结构分类数据库 SCOP

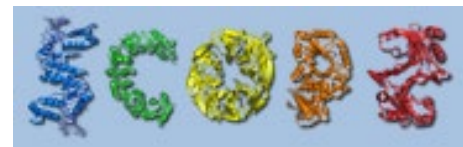
<http://scop2.mrc-lmb.cam.ac.uk/>

英国医学研究委员会（Medical Research Council, MRC）的分子生物学实验室和蛋白质工程研究中心于2014年2月正式发布了蛋白质结构分类数据库SCOP（structural classification of proteins）的全面升级版SCOP2。该数据库在搜集、整理、分析PDB数据中已知的蛋白质三维结构的基础上，详细描述了已知结构的蛋白质在结构、进化事件与功能类型三个方面的关系。数据库的构建除了使用计算机程序外，主要依赖于人工验证。SCOP2把SCOP中仅基于蛋白质结构的树状等级分类系统发展成为单向非循环网状分类系统。

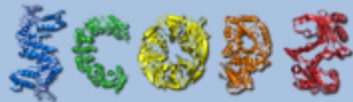
SCOP2分类基于四个层次，从顶部到底部分别为：

类（Class）、折叠（Fold）、超家族（Super family）、家族（Family）

All alpha proteins  
All beta proteins  
Alpha and beta proteins (a+b)  
Alpha and beta proteins (a/b)  
Small proteins



# 蛋白质结构分类数据库 SCOP



## Structural Classification of Proteins 2

[About](#) | [Browser](#) | [Graph](#) | [Download](#) | [Support](#)

MRC

Laboratory of  
Molecular Biology

### News

#### November, 2013

During the development of SCOP2, we have identified a new, previously unrecognised type of alpha-alpha superhelix. Unlike other alpha-alpha superhelices..  
[More...](#)

#### January, 2014

SCOP2 article in NAR is published  
[More...](#)

#### January, 2014

The structure of the month  
[More...](#)

## Welcome to SCOP2!

### Citation

Antonina Andreeva, Dave Howorth, Cyrus Chothia, Eugene Kulesha, Alexey Murzin, SCOP2 prototype: a new approach to protein structure mining (2014) Nucl. Acid Res., 42 (D1): D310-D314.  
[\[PDF\]](#)

### Description of the SCOP2 database

SCOP2 is a successor of Structural classification of proteins (SCOP). Similarly to SCOP, the main focus of SCOP2 is on proteins that are structurally characterized and deposited in the PDB. Proteins are organized according to their structural and evolutionary relationships, but, in contrast to SCOP, instead of a simple tree-like hierarchy these relationships form a complex network of nodes. Each node represents a relationship of a particular type and is exemplified by a region of protein structure and sequence.

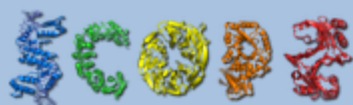
### Search Browser

Add an asterisk to search free text (e.g. serine\*)

### Search Graph

Add an asterisk to search free text (e.g. protein\*domain)

# 蛋白质结构分类数据库 SCOP



## Structural Classification of Proteins 2

[About](#) | [Browser](#) | [Graph](#) | [Download](#) | [Support](#)

MRC

Laboratory of  
Molecular Biology

### 2BOP A:0-0 (8007622)

Add an asterisk to search free text (e.g. serine\*)

[Protein Relationships](#) | [Structural Classes](#) | [Protein Types](#) | [Evolutionary Events](#) | [Keywords](#)



#### Parents

- species [Bovine papillomavirus type 1](#) (6004296) (10559)

Derived from [SCOP 54960](#)

#### Domain

[Go to Graph](#)

  [SP domain](#) (8007622)

[PDB](#) [2BOP A:0-0](#)

[UniProt](#)  [P03122 0-0](#)

protein/DNA complex, complexed with yb

Derived from [SCOP domain 39217](#)



Jmol



# 蛋白质结构分类数据库 SCOP

Structural Classification of Proteins





Protein: Papillomavirus-1 E2 protein from Bovine papillomavirus type 1 [[TaxId: 10559](#)]

## Lineage:

1. Root: [scop](#)
2. Class: [Alpha and beta proteins \(atb\)](#) [53931]  
*Mainly antiparallel beta sheets (segregated alpha and beta regions)*
3. Fold: [Ferrodoxin-like](#) [54861]  
*alpha+beta sandwich with antiparallel beta-sheet; (beta-alpha-beta)<sub>x2</sub>*
4. Superfamily: [Viral DNA-binding domain](#) [54957]  
*Superfamily*
5. Family: [Viral DNA-binding domain](#) [54958]
6. Protein: Papillomavirus-1 E2 protein [54959]  
*forms dimers with subunit beta-sheets making (8,12) barrel*
7. Species: [Bovine papillomavirus type 1](#) [[TaxId: 10559](#)] [54960]

## PDB Entry Domains:

1. [2bop](#)   
*protein/DNA complex; complexed with yb*
  1. [chain a](#) [39217] 

# 途径数据库



**KEGG**途径数据库是存储的是人工绘制的途径图谱，包括了目前已知的所有分子相互作用网络和生物反应网络：总途径图，代谢途径图，遗传信息加工图，环境信息加工图，细胞过程图，机体系统图，人类疾病相关途径图，药物开发图。

Go to: <http://www.genome.jp/kegg/pathway.html>

# 蛋白质结构 --- 三级结构测定

如何获得蛋白质的三级结构

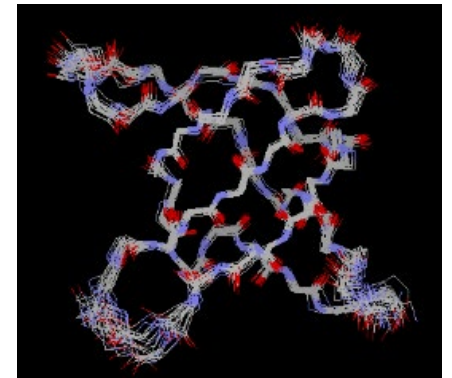
## 1. 实验方法



X-ray Crystallography



Nuclear Magnetic Resonance (NMR) ~200AA



# 蛋白质结构 --- 三级结构测定

如何获得蛋白质的三级结构

## 1. 实验方法

### SDU Experts in X-ray Crystallography



Prof. SUN Jinpeng Ph.D.

[sunjinpeng@sdu.edu.cn](mailto:sunjinpeng@sdu.edu.cn)

Institute of Biochemistry and  
Molecular Biology

School of Medicine, SDU



Prof. GU Lichuan Ph.D.

[lcgu@sdu.edu.cn](mailto:lcgu@sdu.edu.cn)

State Key Laboratory of  
Microbial Technology

School of Life Sciences, SDU

# 蛋白质结构 --- 三级结构预测

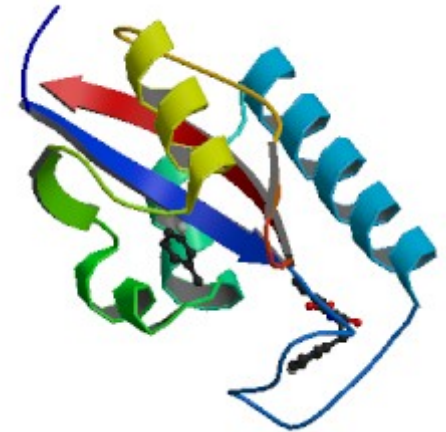
如何获得蛋白质的三级结构

## 2. 计算方法

MEAKIVKVLDSRCEDGFGKKRKRAASYAAYVTGV  
SCAKLQNVPPPNGQCQIPDKRRRLEGENKLSAYE  
NRSGKALVRYTYFYFKKTGIKRVMMYENGEWNDL  
PEHVICAIQNELEEKSAIEFKLCGHSFILDFLHMQR  
LDMETGAKTPLAWIDNAGKCFPEIYESDERTNYC  
HHKCVEDPKQNAPHDIKRLIEDVNGGETPRLNLE  
ECSDSGDNMMDDVPLAQRSSNEHYDEATEDSC  
SRKLEAAVSKWDETDAIVVSGAKLTGSEVLDKDAV  
KKMFAVG TASLGHPVLDVGRFSSEIAEARLALFQ  
KQVEITKKHRGDANVRYAWLPKREVLSAVMMQG  
LGVGGAFIRKSIYGVGIHLTAADCPYFSARYCDVDE  
NGVRYMVLRCRIMGNMELLRGDKAQFFSGGEEYD  
NGVDDIESPKNYIVWNINMNTIFPEFVVRFKLSNL  
PNAEGNLIKRDNSGVTLEGPKDLPPQLESNQGAR  
GSGSANSVGSSTTRPKSPWMPFPTLFAAISHKVAE  
NDMLLINADYQQLRDKKMTAEFVRKLRVIVGDDL  
LRSTITTLQNQPKSKEIPGSIRDHEEGAGGL



**Zero Cost**



input

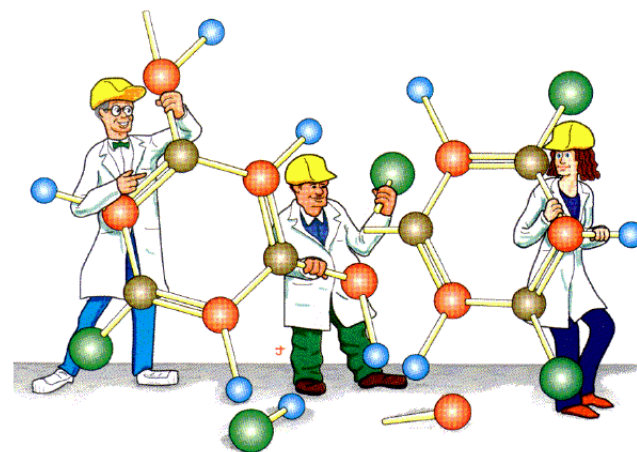
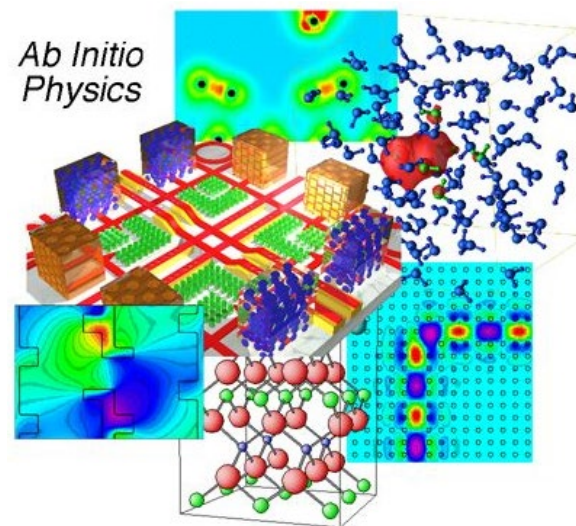
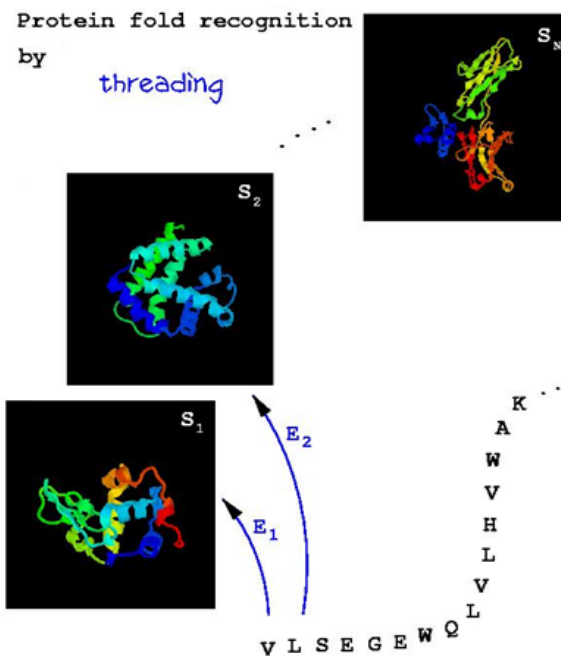
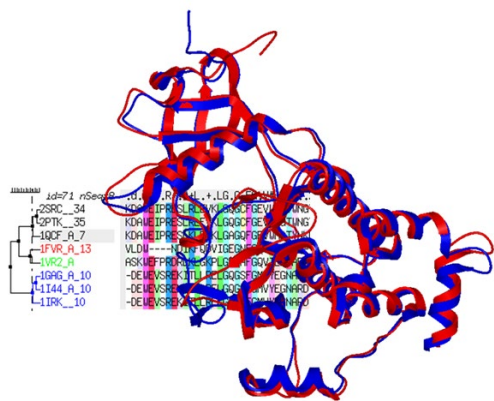
output

# 蛋白质结构 --- 三级结构预测

## 如何获得蛋白质的三级结构

## 2. 计算方法

- 从头计算法
- 同源建模法
- 穿线法
- 综合法



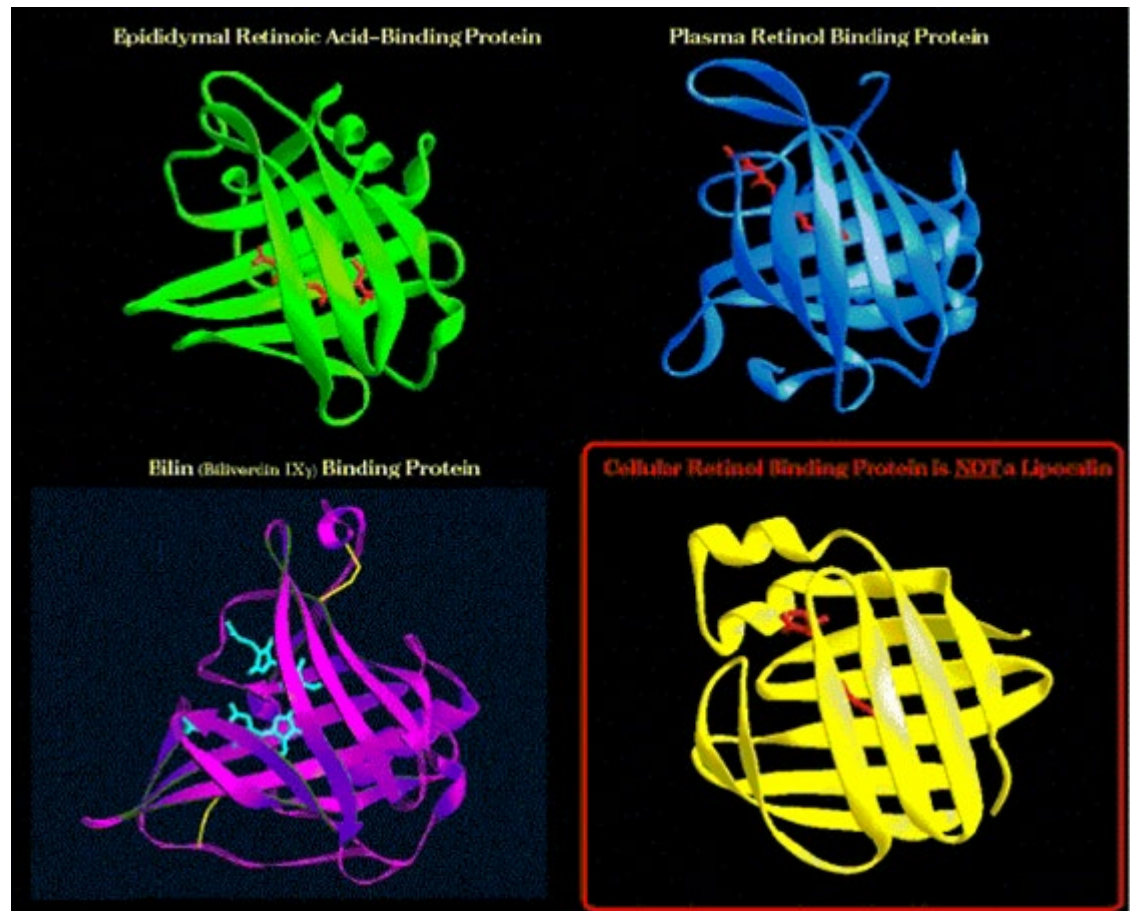


# 蛋白质结构 --- 三级结构预测

如何获得蛋白质的三级结构

## 2. 计算方法 --- 同源建模法

同源建模法：相似的氨基酸序列对应着相似的蛋白质结构。



# 自动同源建模法: SWISS-MODEL

SWISS-MODEL

<http://swissmodel.expasy.org>



SWISS-MODEL

Welcome to SWISS-MODEL

SWISS-MODEL is a fully automated protein structure homology-modelling server, accessible via the ExPASy web server, or from the program DeepView (Swiss Pdb-Viewer). The purpose of this server is to make Protein Modelling accessible to all biochemists and molecular biologists worldwide.

Start Modelling

SWISS-MODEL做出来的结果可以在SCI刊物论文上作为一项结果来发表，但前提条件是模板与目标的序列identity要足够高，至少高于30%，因为SWISS-MODEL是单纯的同源建模法，基本属于结构预测软件里的“傻瓜机”。



# 自动同源建模法: SWISS-MODEL

SWISS-MODEL

<http://swissmodel.expasy.org>

**BIOZENTRUM**  
Universität Basel  
The Center for Molecular Life Sciences

SWISS-MODEL

ModellingToolsRepos

---

## Start a New Modelling Project

Target Sequence:  
*(Format must be Fasta, Clustal, Promod, plain string, or a valid UniProtKB AC)*

+ Upload Target Sequence File...

Project Title:

Email:

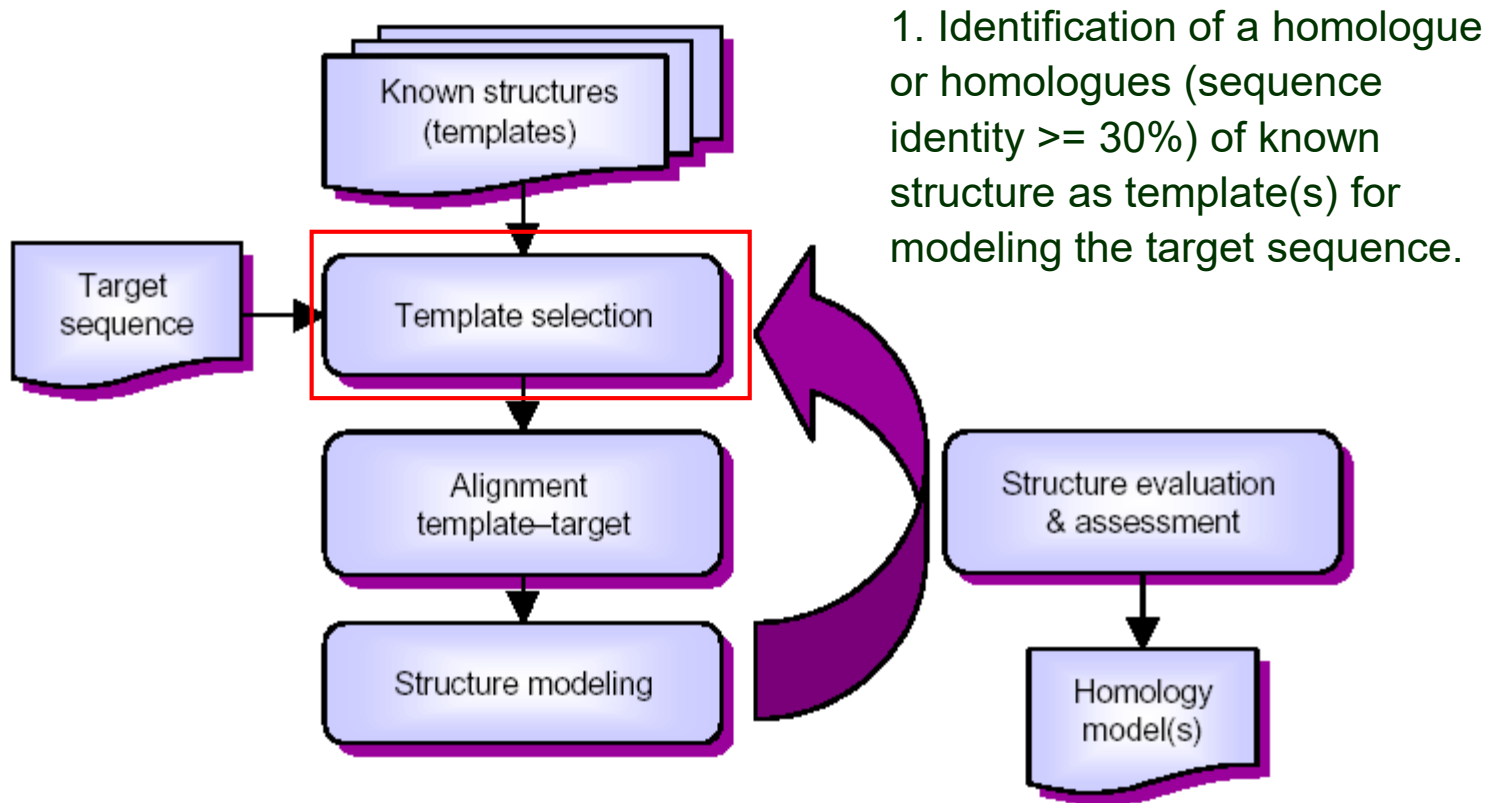
Search For TemplatesBuild Model

*By using the SWISS-MODEL server, you agree to comply with the following [terms of use](#) and to cite the corresponding [articles](#).*

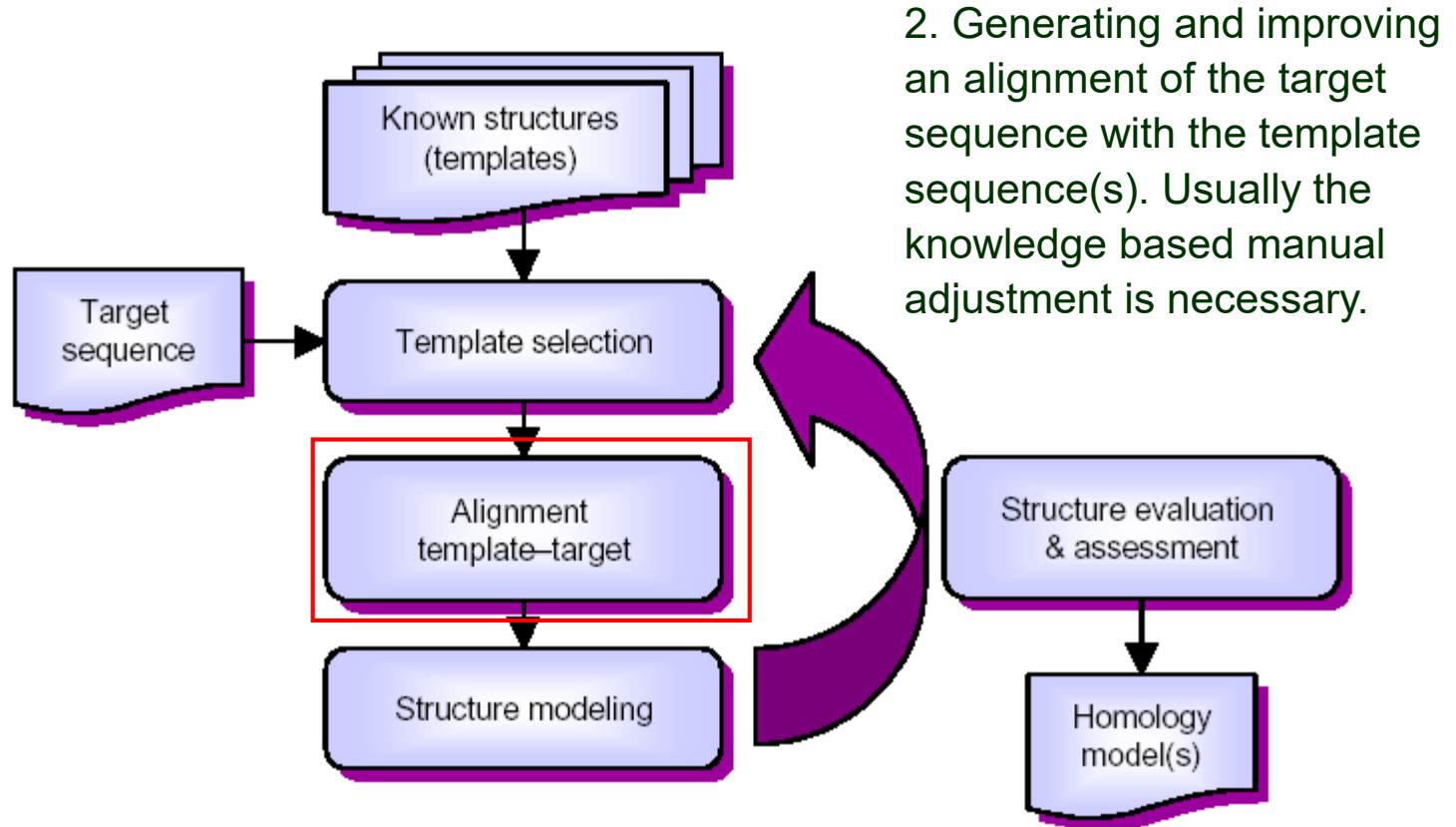
Seq.txt

结果在线等3-5分钟

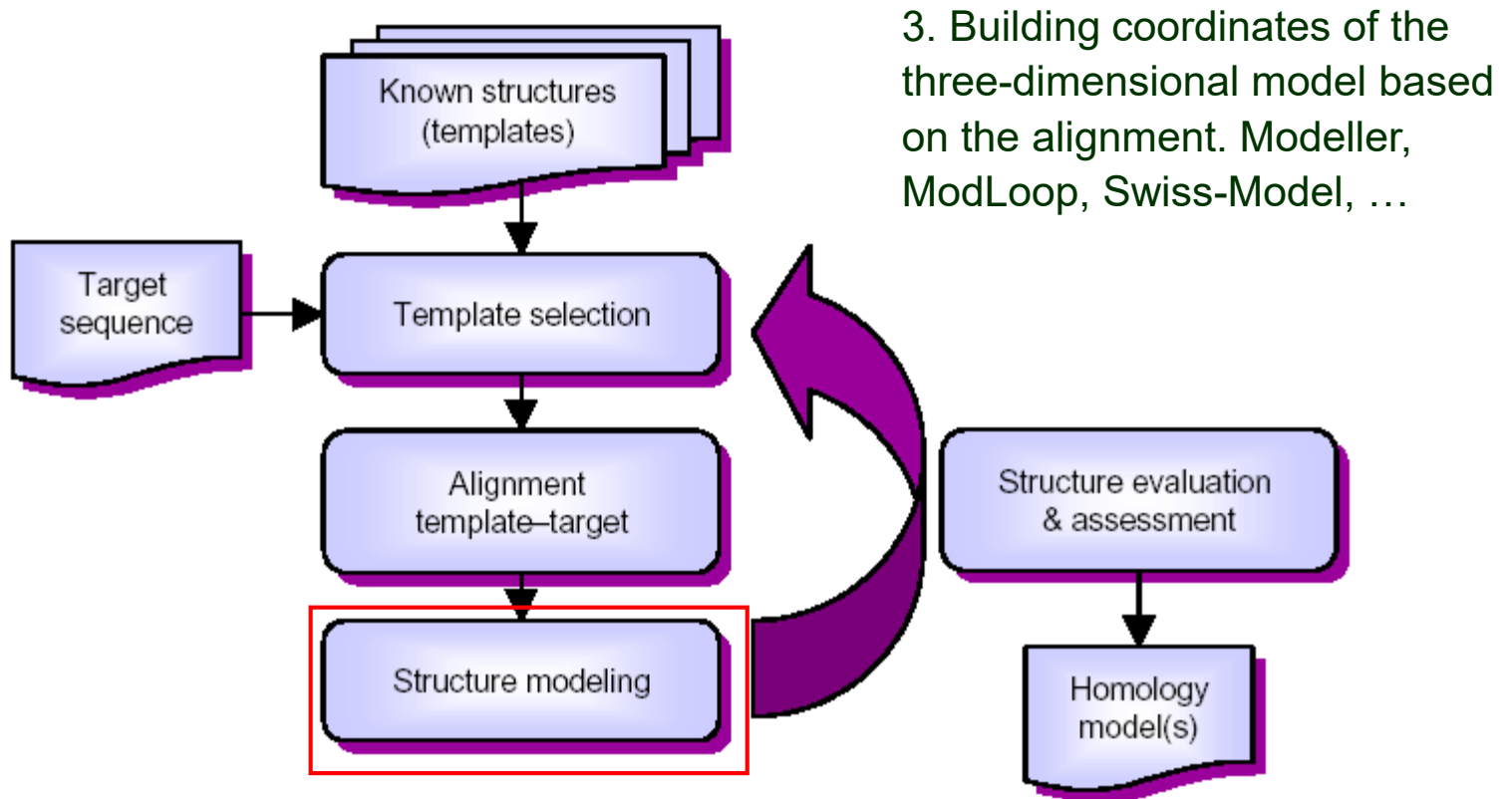
# 半手动同源建模法



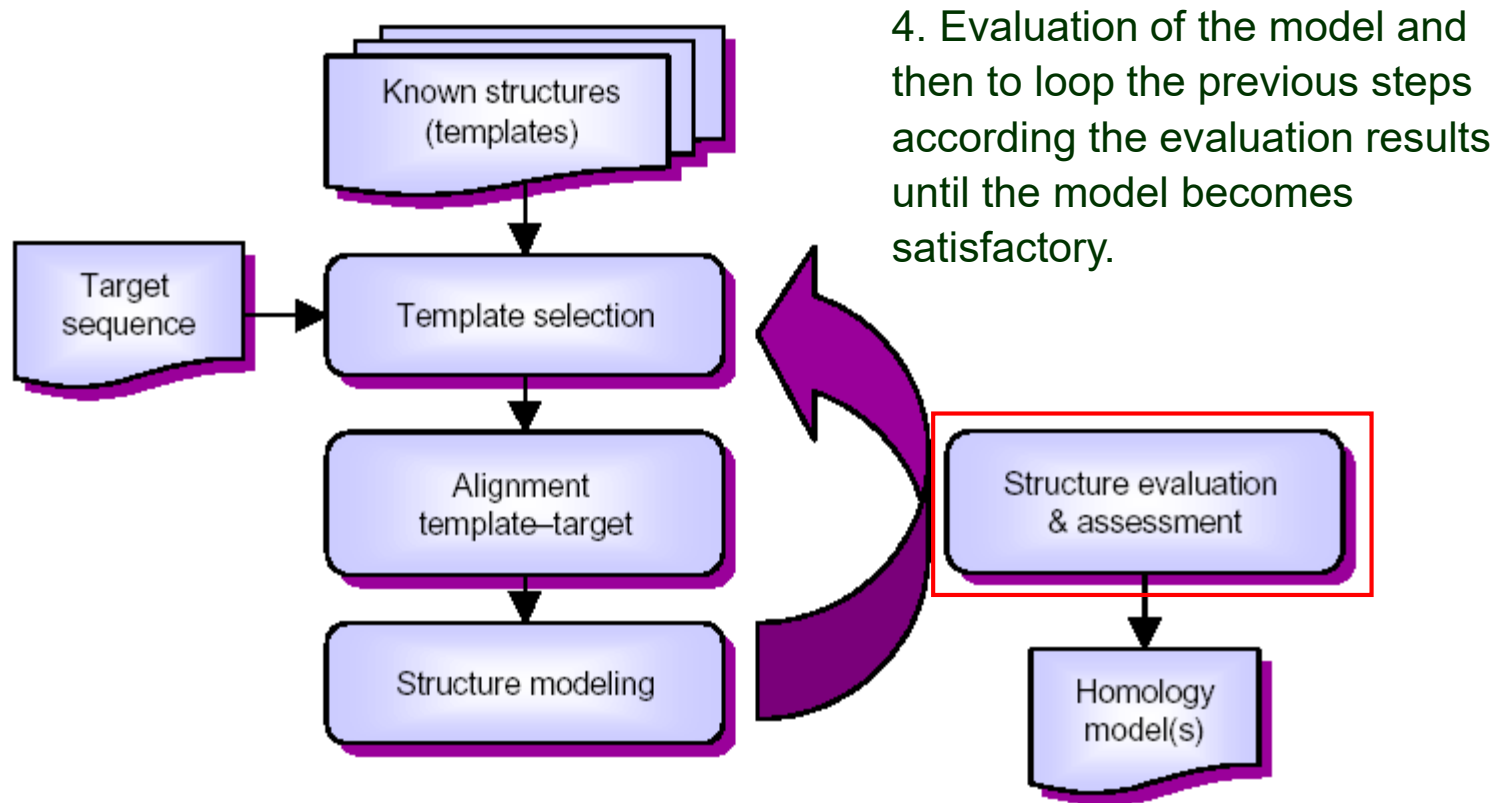
# 半手动同源建模法



# 半手动同源建模法




# 半手动同源建模法



# 半手动同源建模法: SWISS-MODEL

SWISS-MODEL

<http://swissmodel.expasy.org>

**BIOZENTRUM**  
University of Basel  
The Center for Molecular Life Sciences

**SWISS-MODEL**

Modelling Repository Tools Documentation Log in Create Account

### Start a New Modelling Project

**Target Sequence(s):**  
*(Format must be FASTA, Clustal, plain string, or a valid UniProtKB AC)*

Paste your target sequence(s) or UniProtKB AC here

+ Upload Target Sequence File... Validate

**Template File:** + Add Template File...

**Project Title:** Untitled Project

**Email:** Optional

Build Model

### Supported Inputs

Sequence(s) ▼

Target-Template Alignment ▼

User Template ▼

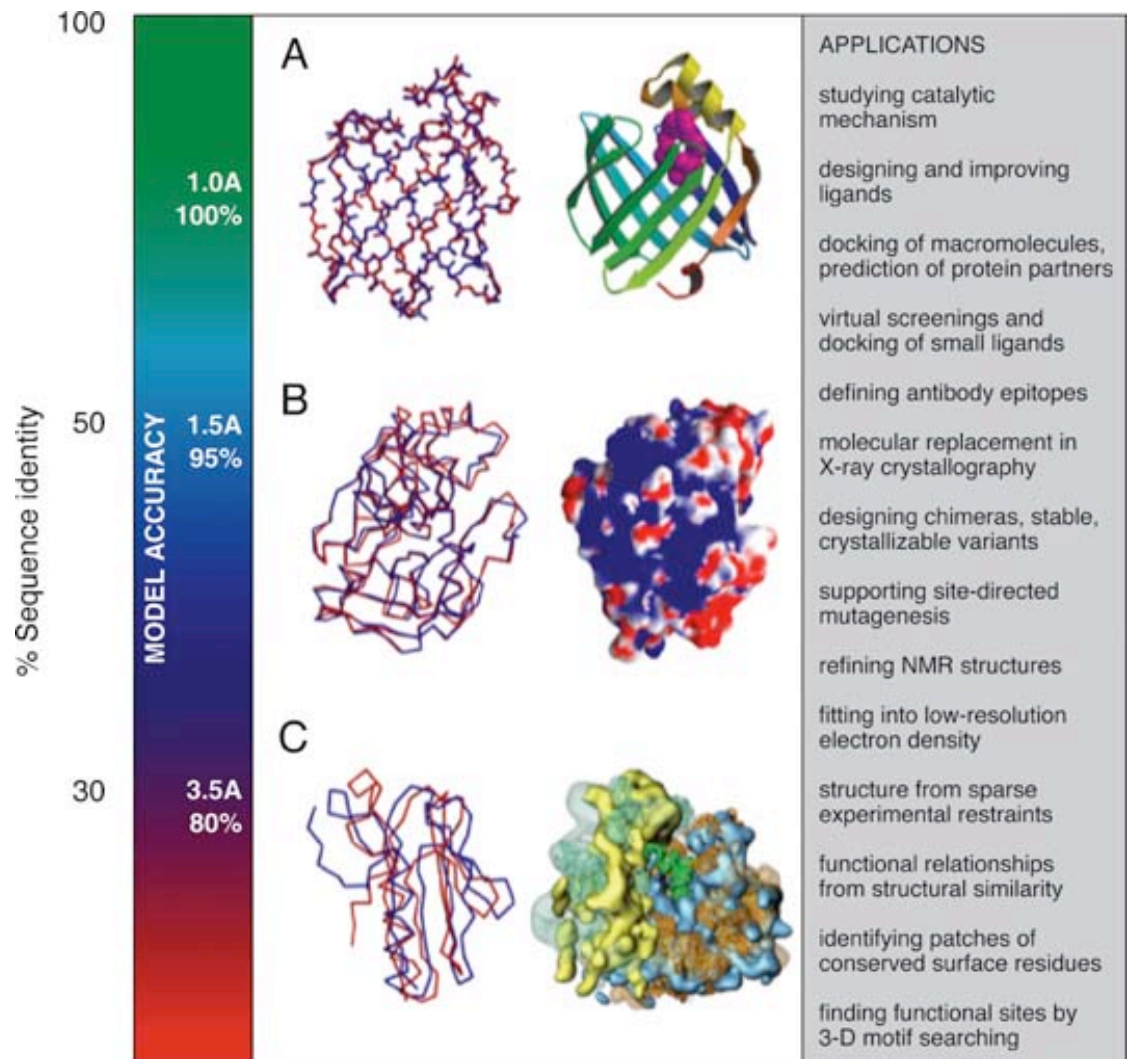
DeepView Project ▼

# 蛋白质结构 --- 三级结构预测

如何获得蛋白质的三级结构

## 2. 计算方法 --- 同源建模法

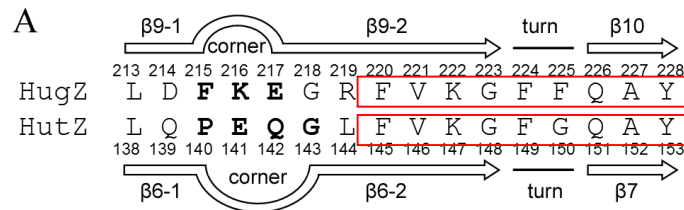
如果目标序列与模板序列相似度极高，那么同源建模法是最准确的方法。



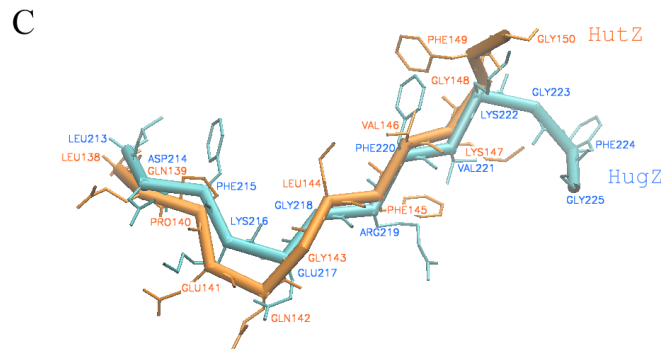
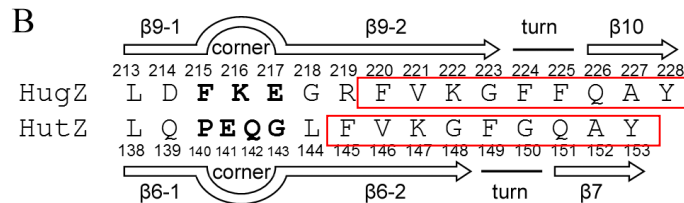
# 蛋白质结构 --- 三级结构预测

如何获得蛋白质的三级结构

## 2. 计算方法 --- 同源建模法



特例情况，虽然序列一致度达到很高水平，但是结构却并不相同。



D



[BMC Struct Biol, 2012, 12:23]

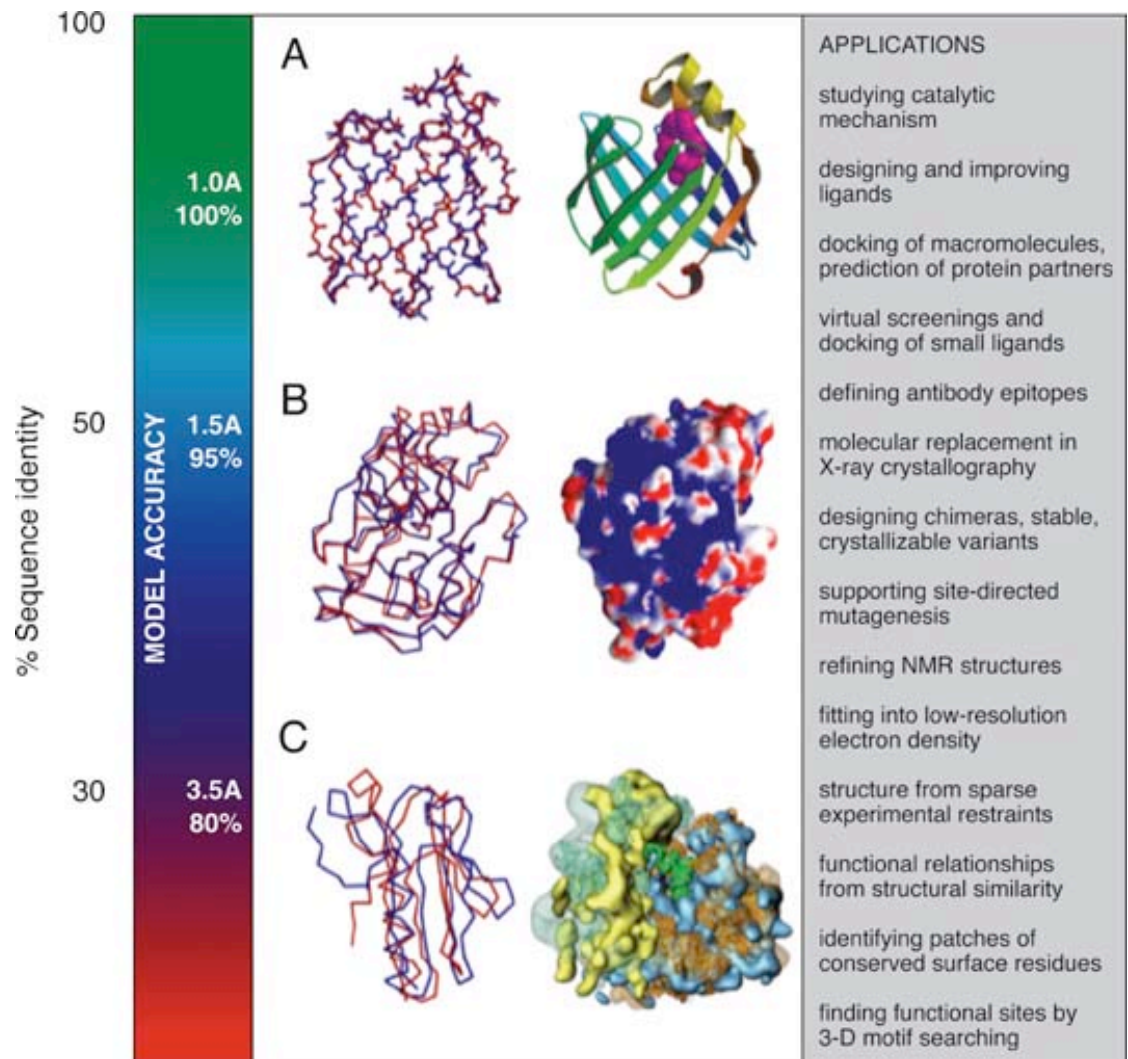


# 蛋白质结构 --- 三级结构预测

如何获得蛋白质的三级结构

## 2. 计算方法 --- 同源建模法

如果目标序列与模板序列之间的一致度  $< 30\%$ ，那么同源建模法是不适用的。

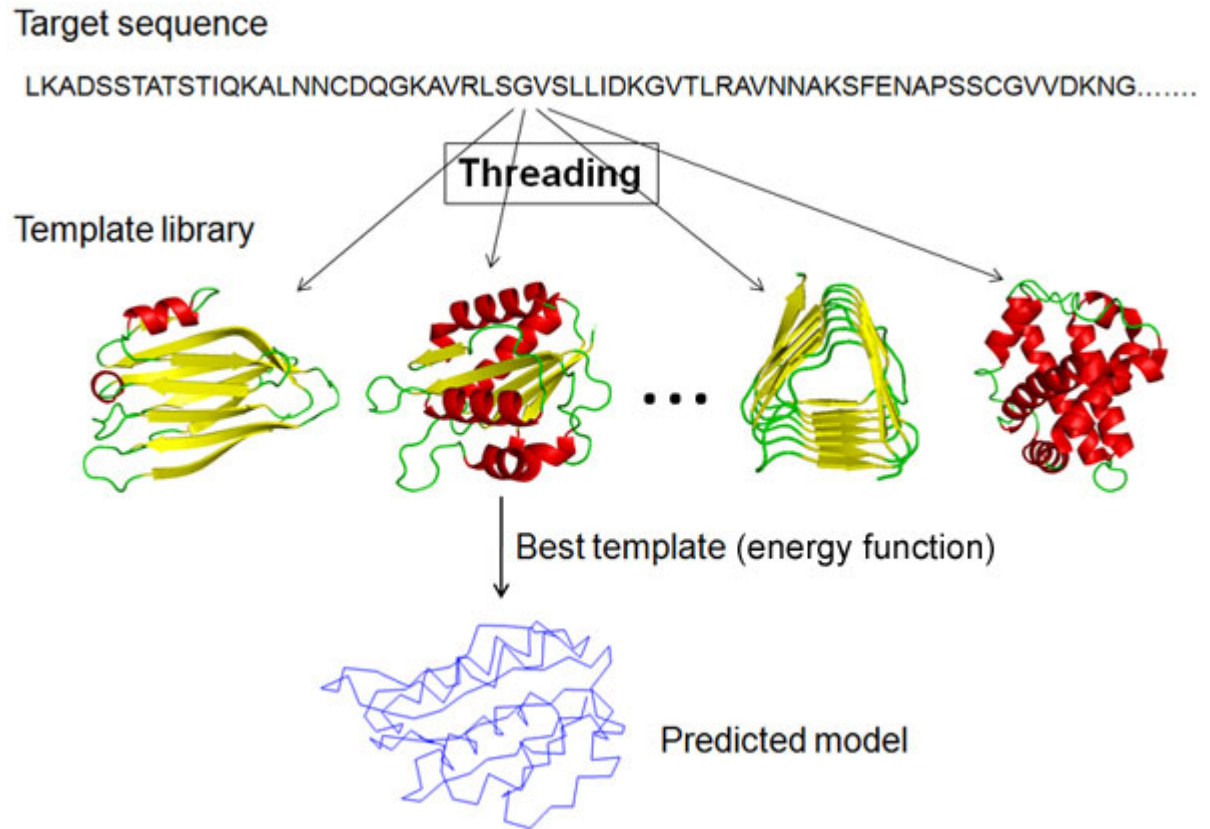


# 蛋白质结构 --- 三级结构预测

如何获得蛋白质的三级结构

## 2. 计算方法 --- 穿线法

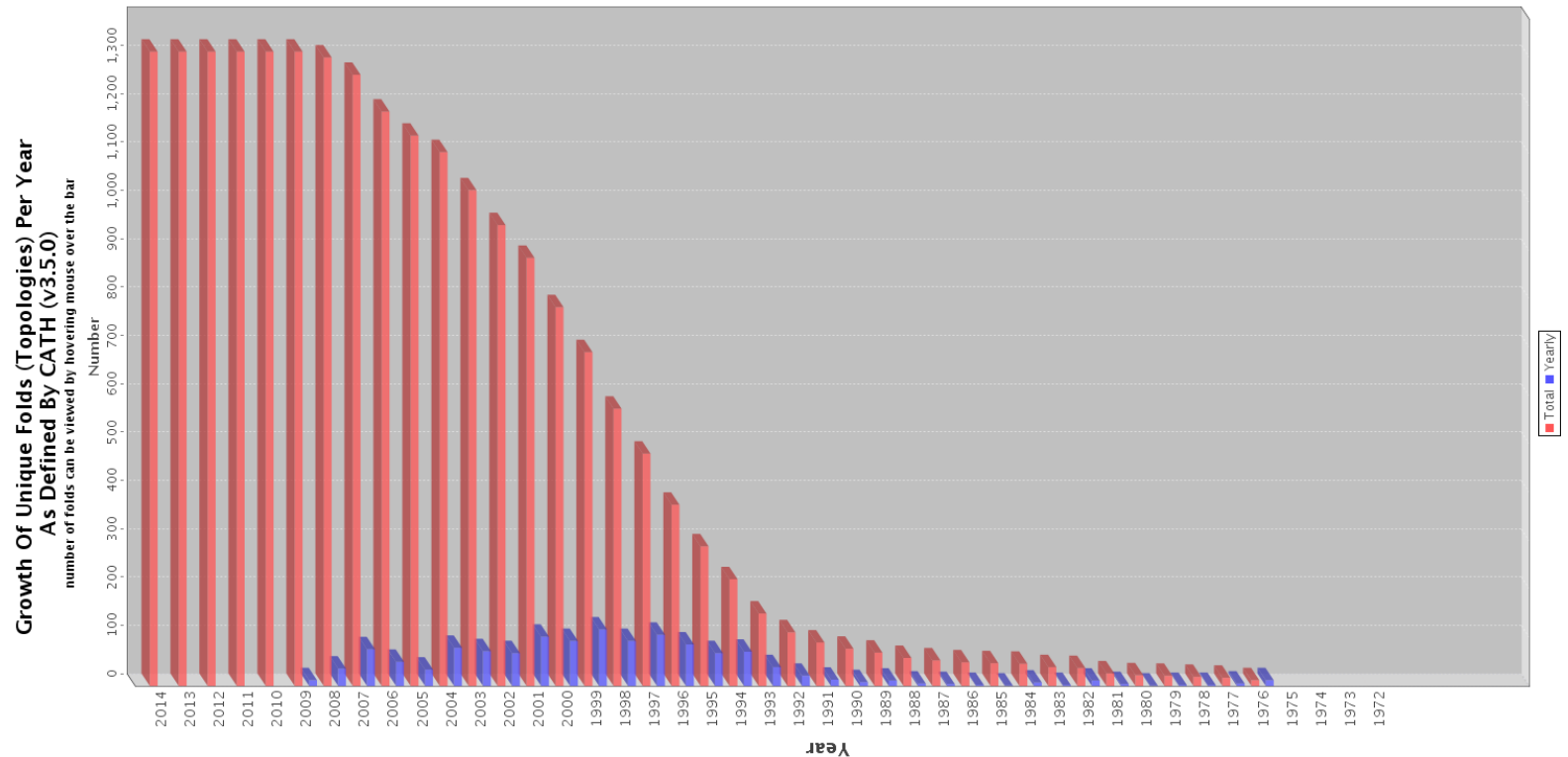
穿线法：不相似的氨基酸序列也可能对应着相似的蛋白质结构。



# 蛋白质结构 --- 三级结构预测

如何获得蛋白质的三级结构

## 2. 计算方法 --- 穿线法



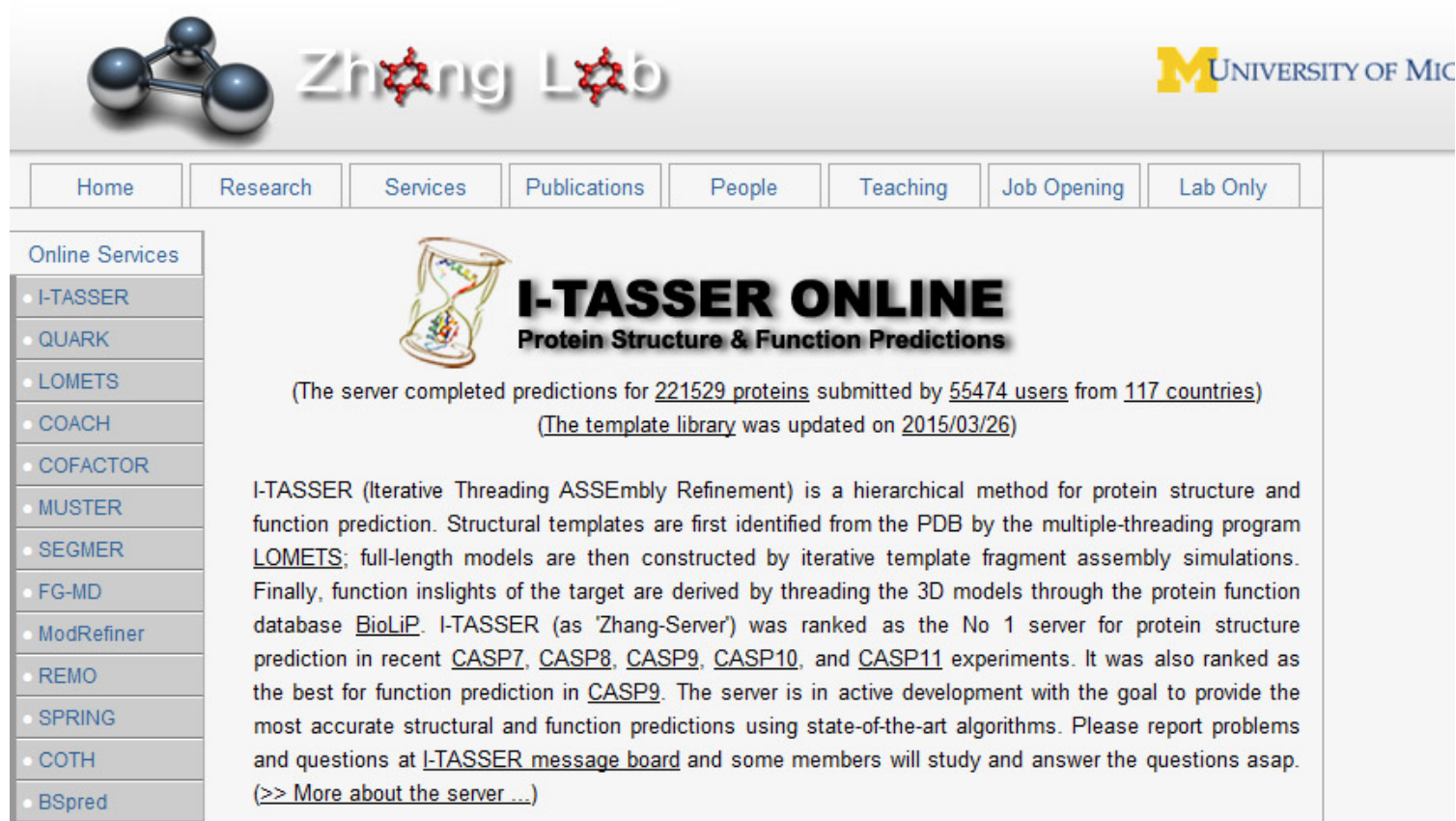
已知结构的蛋白质十几万，不同的结构拓扑 1313。

# 自动穿线法: I-TASSER

## I-TASSER

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

I-TASSER: 在线的蛋白质结构预测服务器，在近7届蛋白质结构预测比赛（CASP7/8/9/10/11/12/13）中皆排名第一。作者为美国密歇根大学的张阳教授。什么都不用自己操心，输入EMAIL和目标序列点RUN。



The screenshot shows the I-TASSER ONLINE website. At the top, there is a header with a logo on the left, the text 'Zhang Lab' in the center, and the 'UNIVERSITY OF MICHIGAN' logo on the right. Below the header is a navigation bar with links: Home, Research, Services, Publications, People, Teaching, Job Opening, and Lab Only. On the left side, there is a sidebar titled 'Online Services' with a list of links: I-TASSER, QUARK, LOMETS, COACH, COFACTOR, MUSTER, SEGMER, FG-MD, ModRefiner, REMO, SPRING, COTH, and BSpred. The main content area features the 'I-TASSER ONLINE Protein Structure & Function Predictions' logo, which includes an hourglass icon. Below the logo, it states: '(The server completed predictions for 221529 proteins submitted by 55474 users from 117 countries) (The template library was updated on 2015/03/26)'. A detailed paragraph describes the I-TASSER method, its ranking in CASP experiments, and provides a link to the I-TASSER message board. At the bottom of the paragraph, it says '( >> More about the server ... )'.

Home Research Services Publications People Teaching Job Opening Lab Only

Online Services

- I-TASSER
- QUARK
- LOMETS
- COACH
- COFACTOR
- MUSTER
- SEGMER
- FG-MD
- ModRefiner
- REMO
- SPRING
- COTH
- BSpred

**I-TASSER ONLINE**  
Protein Structure & Function Predictions

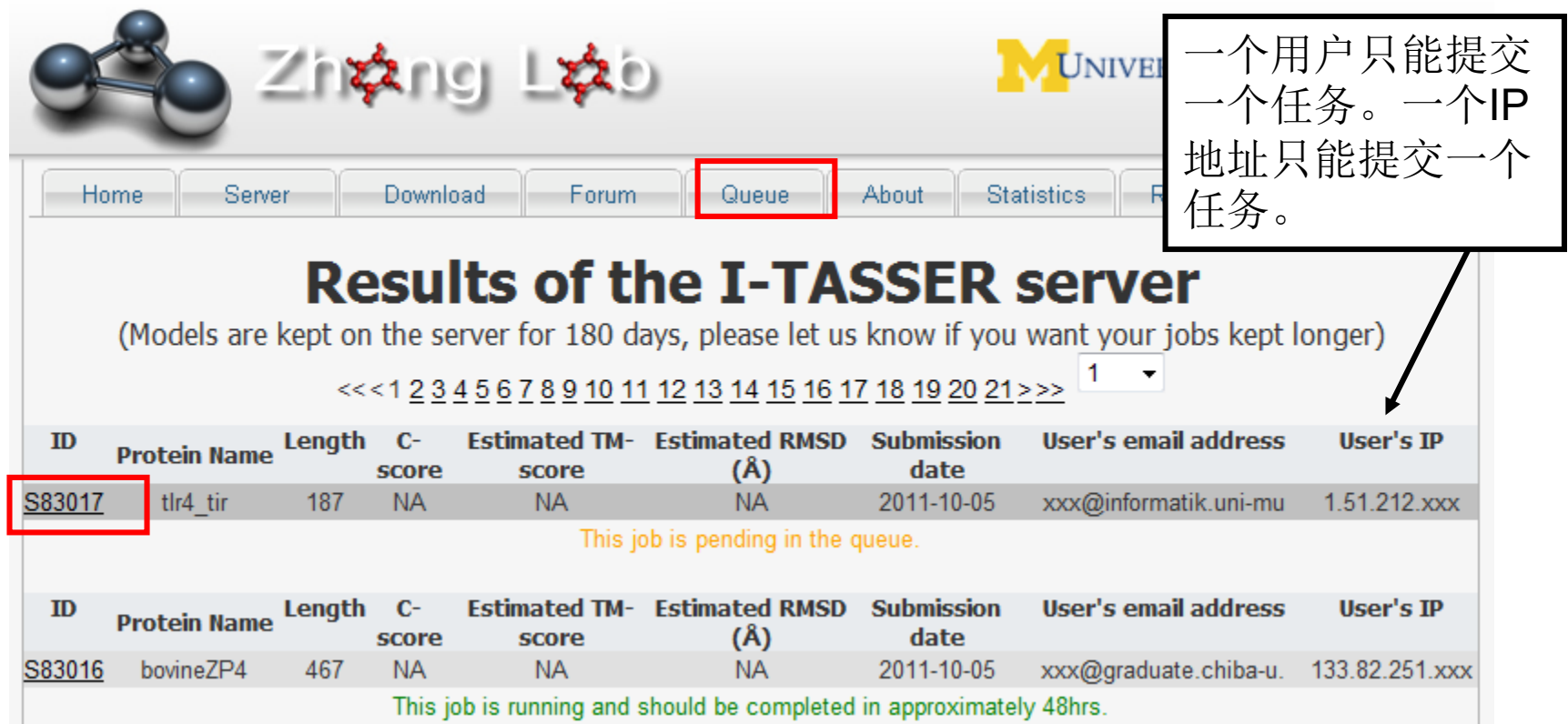
(The server completed predictions for 221529 proteins submitted by 55474 users from 117 countries)  
(The template library was updated on 2015/03/26)

I-TASSER (Iterative Threading ASSEMBly Refinement) is a hierarchical method for protein structure and function prediction. Structural templates are first identified from the PDB by the multiple-threading program LOMETS; full-length models are then constructed by iterative template fragment assembly simulations. Finally, function insights of the target are derived by threading the 3D models through the protein function database BioLiP. I-TASSER (as 'Zhang-Server') was ranked as the No 1 server for protein structure prediction in recent CASP7, CASP8, CASP9, CASP10, and CASP11 experiments. It was also ranked as the best for function prediction in CASP9. The server is in active development with the goal to provide the most accurate structural and function predictions using state-of-the-art algorithms. Please report problems and questions at I-TASSER message board and some members will study and answer the questions asap.  
( >> More about the server ... )

# 自动穿线法: I-TASSER

I-TASSER

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



The screenshot shows the I-TASSER server interface. At the top, there is a logo for 'Zhang Lab' and a navigation bar with links: Home, Server, Download, Forum, Queue, About, Statistics. The 'Queue' link is highlighted with a red box. Below the navigation bar, the title 'Results of the I-TASSER server' is displayed, followed by a note: '(Models are kept on the server for 180 days, please let us know if you want your jobs kept longer)'. A pagination control shows '<<< 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 >>>' with a dropdown menu set to '1'. Below this is a table of job results. The first row, with ID 'S83017', is highlighted with a red box. Below it, a message states 'This job is pending in the queue.' The second row, with ID 'S83016', is shown below, with a message stating 'This job is running and should be completed in approximately 48hrs.' A callout box on the right contains the text: '一个用户只能提交一个任务。一个IP地址只能提交一个任务。' with an arrow pointing to the 'User's IP' column header.

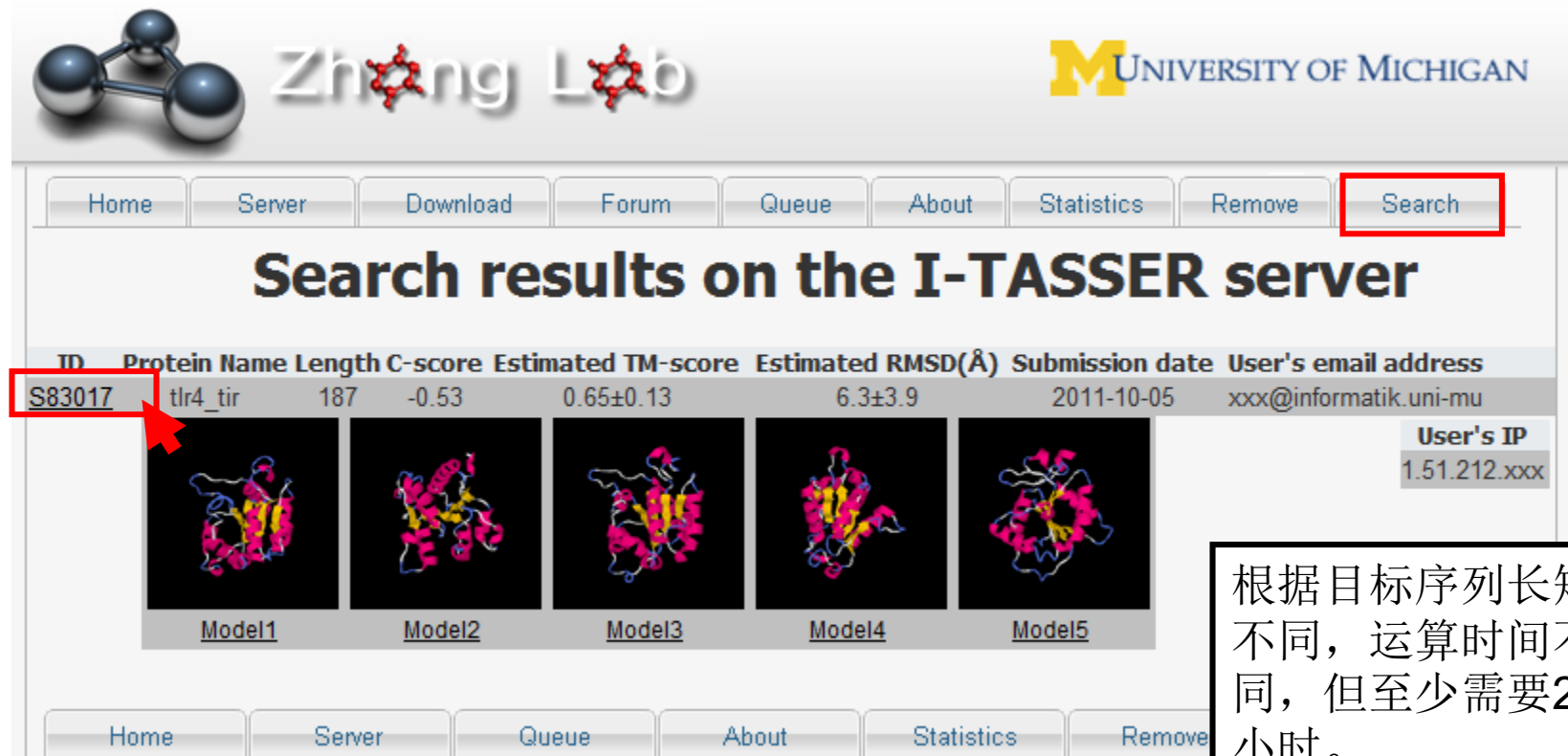
一个用户只能提交一个任务。一个IP地址只能提交一个任务。

ID	Protein Name	Length	C-score	Estimated TM-score	Estimated RMSD (Å)	Submission date	User's email address	User's IP
S83017	tlr4_tir	187	NA	NA	NA	2011-10-05	xxx@informatik.uni-mu	1.51.212.xxx
This job is pending in the queue.								
S83016	bovineZP4	467	NA	NA	NA	2011-10-05	xxx@graduate.chiba-u.	133.82.251.xxx
This job is running and should be completed in approximately 48hrs.								

# 自动穿线法: I-TASSER

I-TASSER

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



The image shows the I-TASSER web interface. At the top, there are logos for 'Zhang Lab' and 'UNIVERSITY OF MICHIGAN'. Below the logos is a navigation bar with buttons: Home, Server, Download, Forum, Queue, About, Statistics, Remove, and Search (highlighted with a red box). The main heading is 'Search results on the I-TASSER server'. Below this is a table of search results. The first row is highlighted with a red box and an arrow pointing to the ID 'S83017'. The table columns are: ID, Protein Name, Length, C-score, Estimated TM-score, Estimated RMSD(Å), Submission date, and User's email address. Below the table, there are five protein structure models labeled Model1 to Model5, each showing a different conformation of the protein. At the bottom right, there is a box containing the text: '根据目标序列长短不同，运算时间不同，但至少需要24小时。' (Depending on the length of the target sequence, the calculation time varies, but it takes at least 24 hours.)

ID	Protein Name	Length	C-score	Estimated TM-score	Estimated RMSD(Å)	Submission date	User's email address
S83017	tlr4_tir	187	-0.53	0.65±0.13	6.3±3.9	2011-10-05	xxx@informatik.uni-mu

Model1 Model2 Model3 Model4 Model5

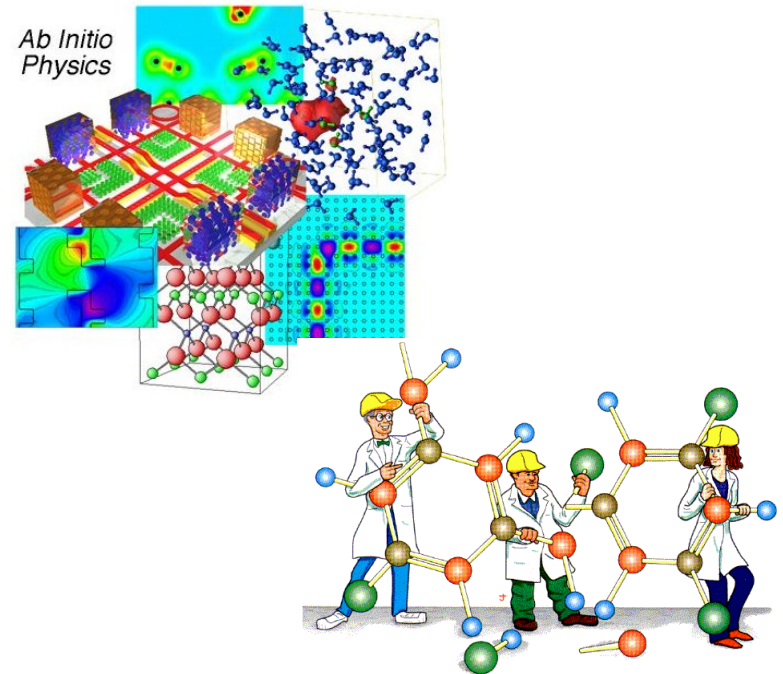
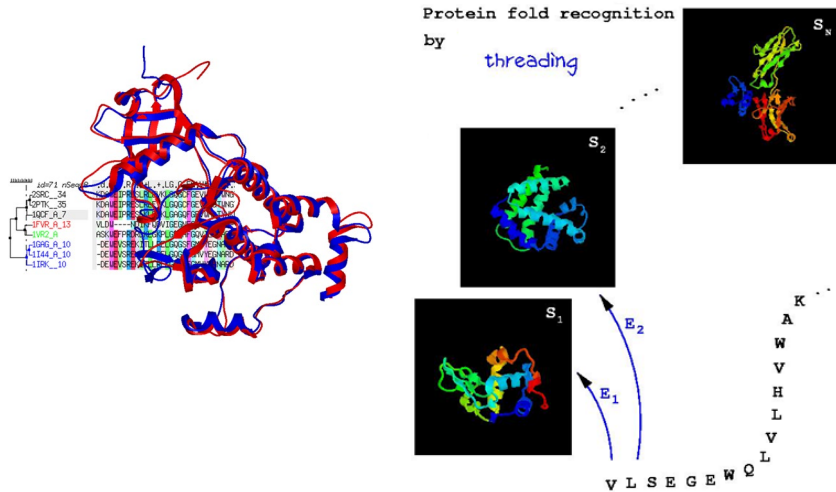
User's IP: 1.51.212.xxx

Home Server Queue About Statistics Remove



# 蛋白质结构预测

- 结构生物学中结合先验知识的计算方法
- 科学家们数十年的努力，只覆盖了人类蛋白质序列中**17%**的氨基酸残基。



# AlphaFold 2

nature

[Explore content](#) [About the journal](#) [Publish with us](#) [Subscribe](#)

[nature](#) > [news](#) > [article](#)

NEWS | 22 July 2021

## DeepMind's AI predicts structures for a vast trove of proteins

AlphaFold neural network produced a 'totally transformative' database of more than 350,000 structures from *Homo sapiens* and 20 model organisms.

2021年7月，98.5%的人类蛋白质结构被AlphaFold2预测出来。除了人类蛋白质组，数据集中还包括大肠杆菌、果蝇、小鼠等20个具有科研常用生物的蛋白质组数据，总计超过35万个蛋白质的结构。最重要的是，这些全都免费开放，交给欧洲生物信息学研究所托管。

DeepMind创始人哈撒比斯在官网发布题为《把AlphaFold的力量交到全世界手中》的文章，同时也在推特上表达了他抑制不住地兴奋：这是我一生中梦寐以求的日子，也是创办Deepmind的初衷：用AI推进科学发展并造福人类。

nature

[Explore content](#) [About the journal](#) [Publish with us](#)

[nature](#) > [articles](#) > [article](#)

Article | [Open Access](#) | Published: 22 July 2021

## Highly accurate protein structure prediction for the human proteome

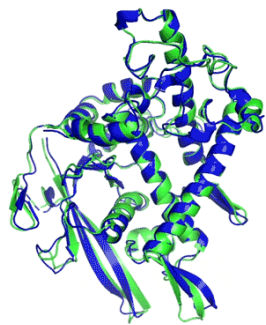
Kathryn Tunyasuvunakool , Jonas Adler, [...]Demis Hassabis 

*Nature* 596, 590–596 (2021) | [Cite this article](#)

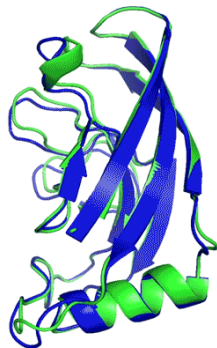
146k Accesses | 8 Citations | 1354 Altmetric | [Metrics](#)



# AlphaFold 2



T1037 / 6vr4  
90.7 GDT  
(RNA polymerase domain)



T1049 / 6y4f  
93.3 GDT  
(adhesin tip)

● Experimental result  
● Computational prediction

借鉴了AI研究中最近新兴起的Transformer架构。Transformer使用注意力机制兴起于NLP领域，用于处理一连串的文本序列。而氨基酸序列正是和文本类似的数据结构，AlphaFold2利用多序列比对，把蛋白质的结构和生物信息整合到了深度学习算法中。

**AlphaFold2准确性：**预测结构和蛋白质真实结构之间只差一个原子的宽度，真正解决了蛋白质折叠的问题。

# AlphaFold 2

欧洲分子生物学实验室（EMBL）的负责人Edith Heard说：  
“我们相信这对理解生命体是如何运作有着变革性的影响。”

哥伦比亚大学的计算生物学家Mohammed AlQuraishi表示，此前蛋白质结构预测领域总是要花费大量时间在一些基础工作上，浪费了学者的很多精力，现在他们可以更加专注于对蛋白质结构的研究了。

一些与DeepMind展开合作的研究团队，已经通过AlphaFold加速了研究进程：

DNDi（被忽视疾病药物开发组织）就表示，AlphaFold2推动了他们在热带疾病药物开发方面的研究。朴茨茅斯大学酶创新中心（CEI）也表示，他们正在利用AlphaFold2开发一些新的酶，可以用来降解污染环境的一次性塑料。

科罗拉多大学波尔德分校的生化学家Marcelo Sousa则利用AlphaFold来制作蛋白质结构模型，开展一项关于抗生素的研究。

加州大学旧金山分校的一个团队则表示，AlphaFold2可以帮助他们更好理解SARS-CoV-2的生物学机制。

AlphaFold2论文地址：<https://www.nature.com/articles/s41586-021-03828-1>