2013.6.21

PROTOCOL

I-TASSER: a unified platform for automated protein structure and function prediction

Ambrish Roy[1,2], Alper Kucukural[2] & Yang Zhang[1,2]

[1]Center for Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, USA. [2]Center for Bioinformatics and Department of Molecular Bioscience, University of Kansas, Lawrence, Kansas, USA. Correspondence should be addressed to Y.Z. (zhng@umich.edu).

- 主讲人：张潇潇
- 组员：姜明君，李文静，张春晓，杨倩

背景

进行蛋白质结构预测的动因：

蛋白质三维结构 → 生物学功能

基于结构的功能分配

计算方法分类：

- 比较建模：以进化相关的模板作为框架。
- **threading**：将所求序列与已知蛋白的三维结构进行匹配。
- ***ab initio***: 没有找到与结构相关的蛋白 。**< 120 AA.**

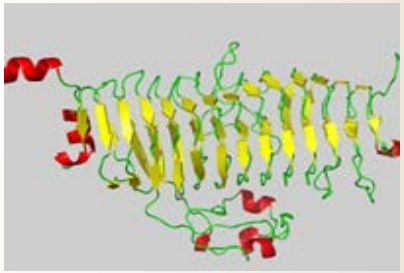
UniProt

蛋白序列

```
        10         20
MVLSPADKTN VKAAWGKVGA HAG
        70         80
KKVADALTNA VAHVDDMPNA LSA
       130        140
AVHASLDKFL ASVSTVLTSK YR
```

RCSB PDB PROTEIN DATA BANK

晶体结构



背景

I-TASSER (iterative threading assembly refinement)

在最近的四次CASP实验中排名最高：

- 序列-结构-功能 的范式。
- 结合了threading， *ab initio* 建模和原子级的结构细化。
- 始终使模板更接近于天然结构。
- 局部和全局的结构相似性比较。

→ 准确性！

开始：
- 序列
- 约束条件
- 实验数据

→

生成：
- 三维原子模型
- 二级和三级结构
- 置信区间

→

功能解释：
- 配体结合位点
- 酶功能
- 细胞定位等

简要介绍

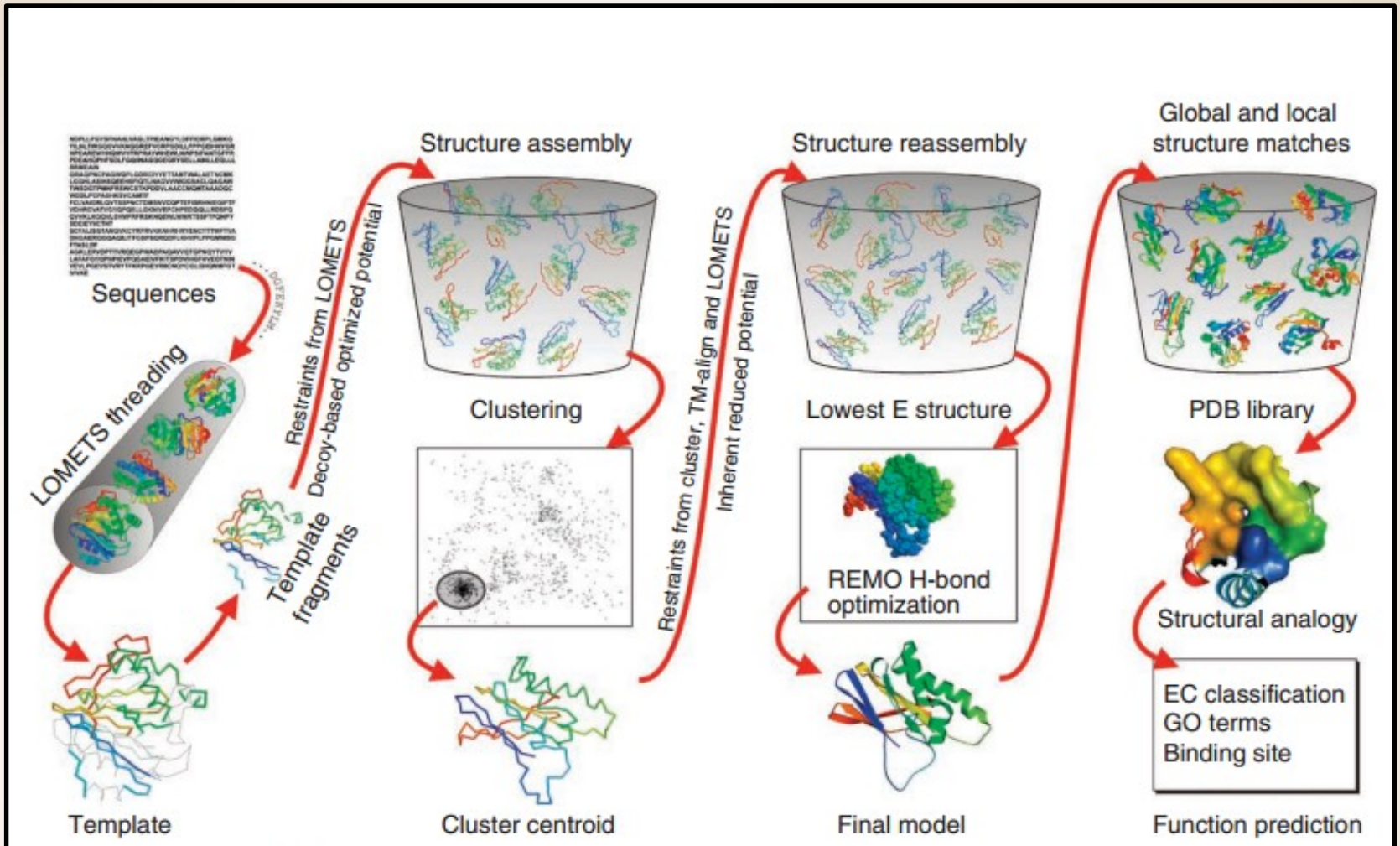
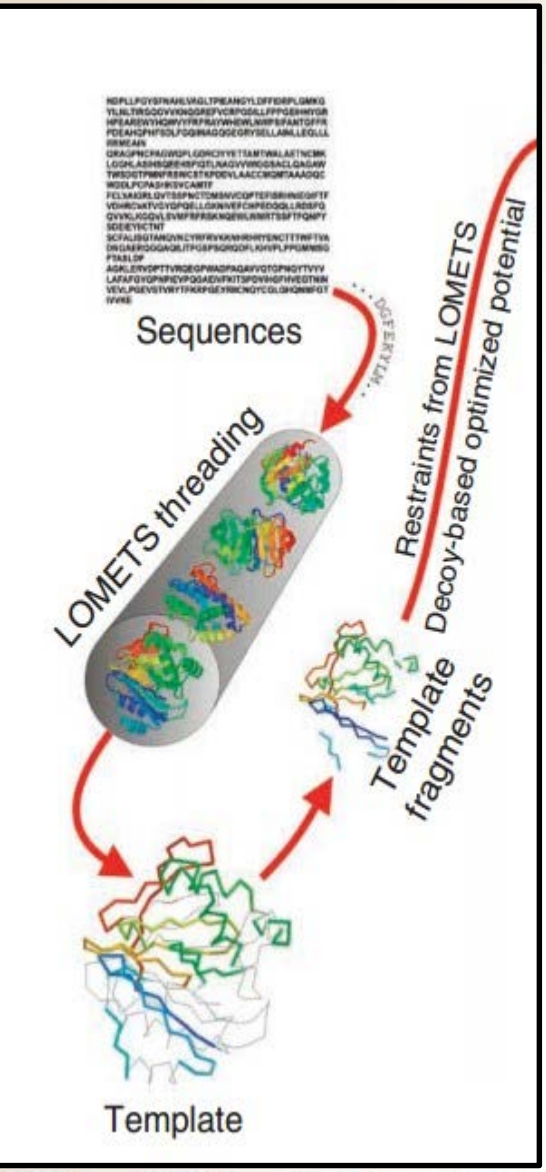


Figure 1 | A schematic representation of the I-TASSER protocol for protein structure and function predictions. The protein chains are colored from blue at the N-terminus to red at the C-terminus.

第一步：threading——构建模板



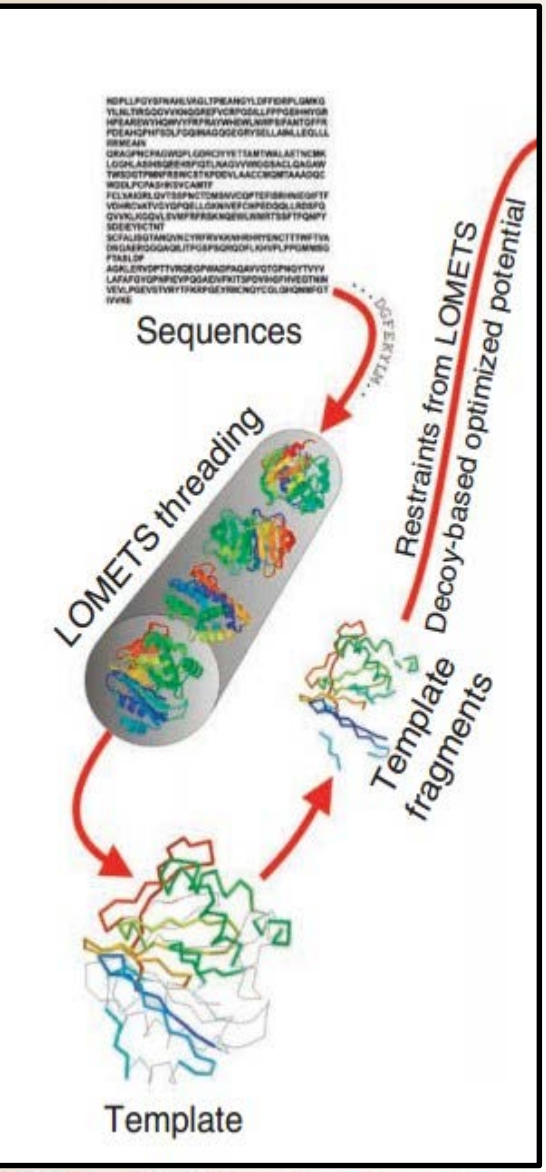
PSI-BLAST: 寻找进化亲缘关系相近的序列。

PSIPRED: 预测序列二级结构。

LOMETS (7)： 根据非冗余PDB数据库进行一致性threading。

```
                    20                  40
                    |                   |
CCCCCCCCCCCHHHHHHHHHHHHHHHHHHHCCCHHHHHHHHHC
MFGVGIDIIEIDRIRKSYQTYGDRFLKKIFTEGERVYCFSK
IYGIGLDITELKRIASMAGRQ-KRFAERILTRSELDQYYEL
IIGIGTDILCVNRIYKILEK-NINFIKKVLNPFELAEFETQ
IIGIGTDILCVNRIYKILEK-NINFIKKVLNPFELAEFETQ
IIGVGIDVAEVERFGAALER-TPALAGRLFLESELLLPGGE
IVGHGIDIEELASIESAVTRH-EGFAKRVLTALEMERFTSL
```

第一步：threading——构建模板



由归一化的Z-分数来决定模板结构的优劣。

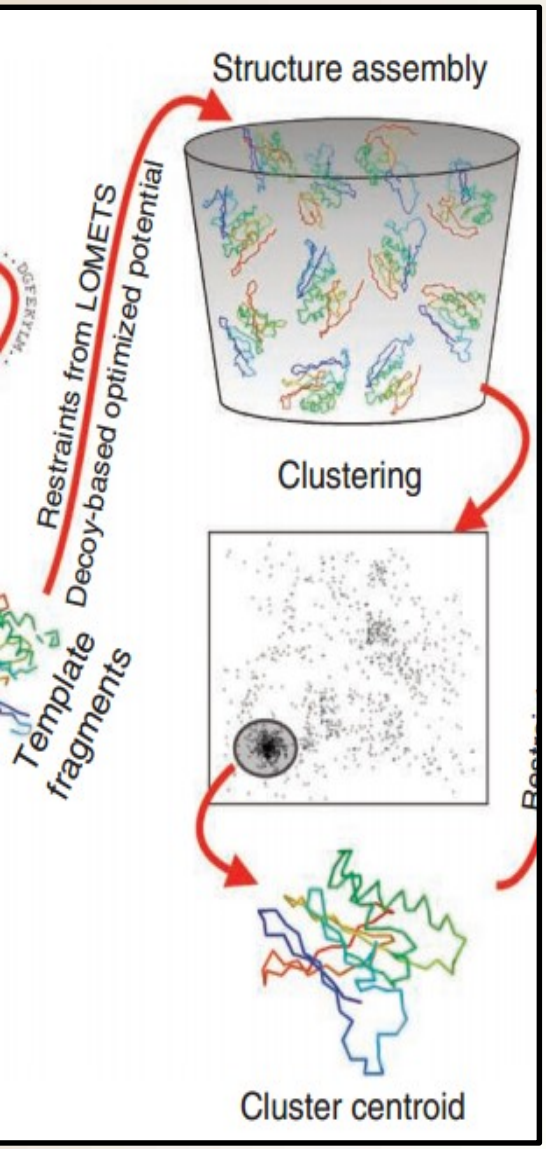
Z-分数：能量相对于平均值的标准差。

Top ten templates used by I-TASSER

Rank	PDB hit	Iden1	Iden2	Cov.	Norm. *Z*-score	Download align.
1	1f71A	0.37	0.34	0.94	3.44	Download
2	2bddA	0.28	0.26	0.93	4.49	Download
3	2qg8A	0.28	0.23	0.98	4.57	Download
4	2jcaA	0.35	0.33	0.97	2.89	Download
5	1fteA	0.30	0.27	0.92	2.32	Download
6	1qr0A	0.17	0.13	0.87	2.07	Download
7	2z0rA	0.21	0.23	0.91	0.82	Download
8	1qr0A	0.13	0.21	0.83	0.74	Download
9	1y2tA	0.13	0.23	1.00	0.43	Download
10	1qr0A	0.22	0.21	0.86	0.80	Download

>1, 优秀模板

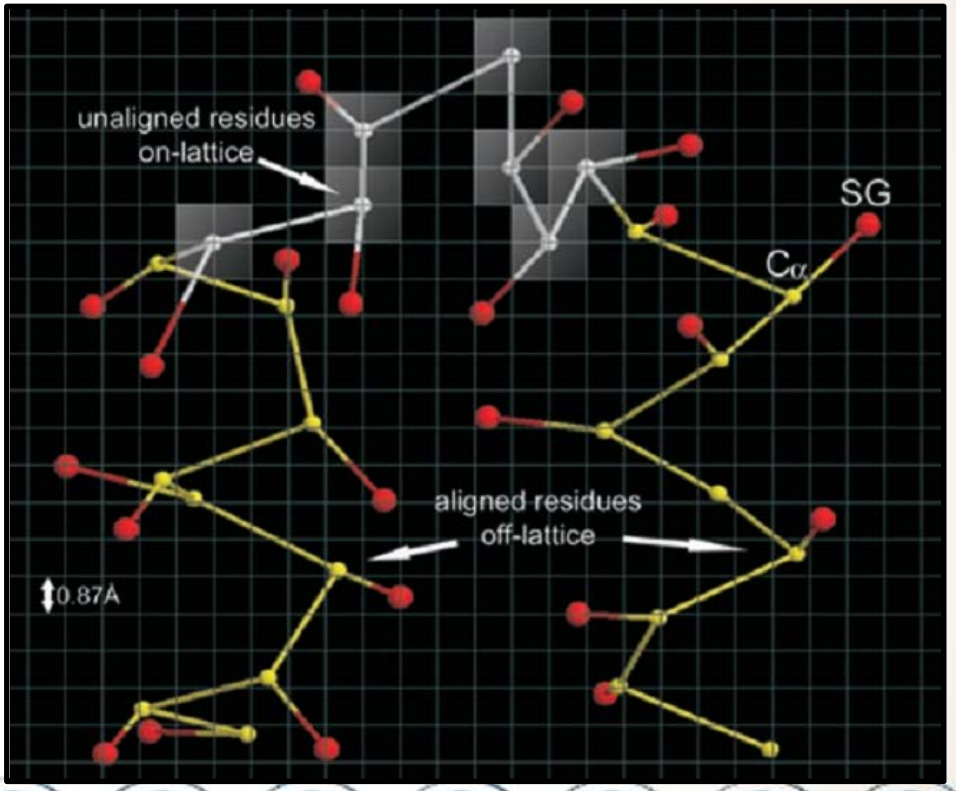
第二步：结构装配



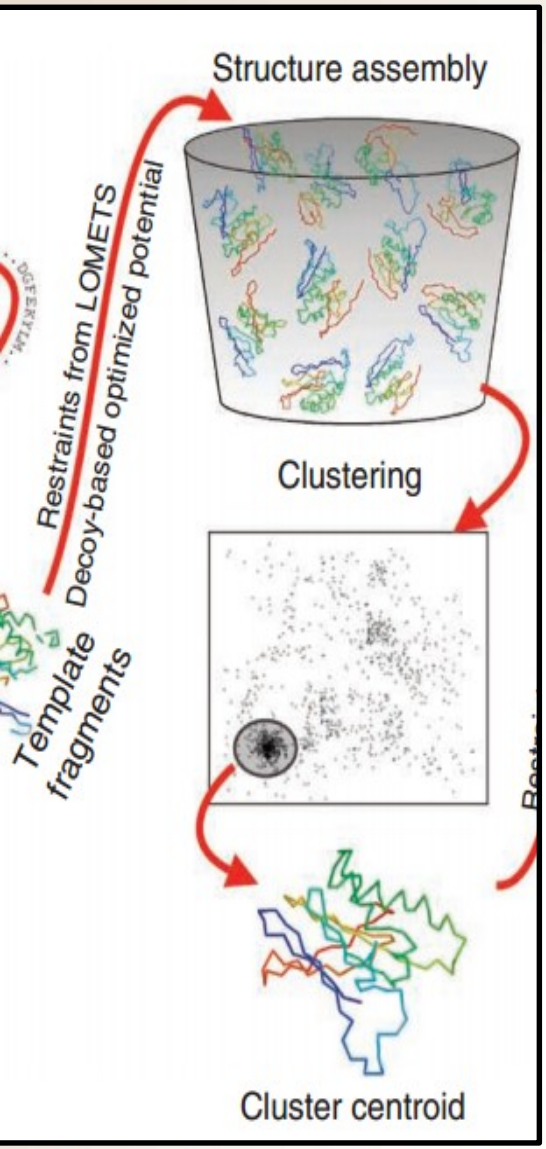
CAS model: 只有Cα和侧链的中心

模板片段：非格点，由3个欧拉角控制

未对齐的片段：从头建模，Cα被晶格限制，2-6个键移动的自由度。



第二步：结构装配



改良的 REMC/经验性力场/增强抽样方法。

SPICKER: 聚类分析，识别低自由能态 。

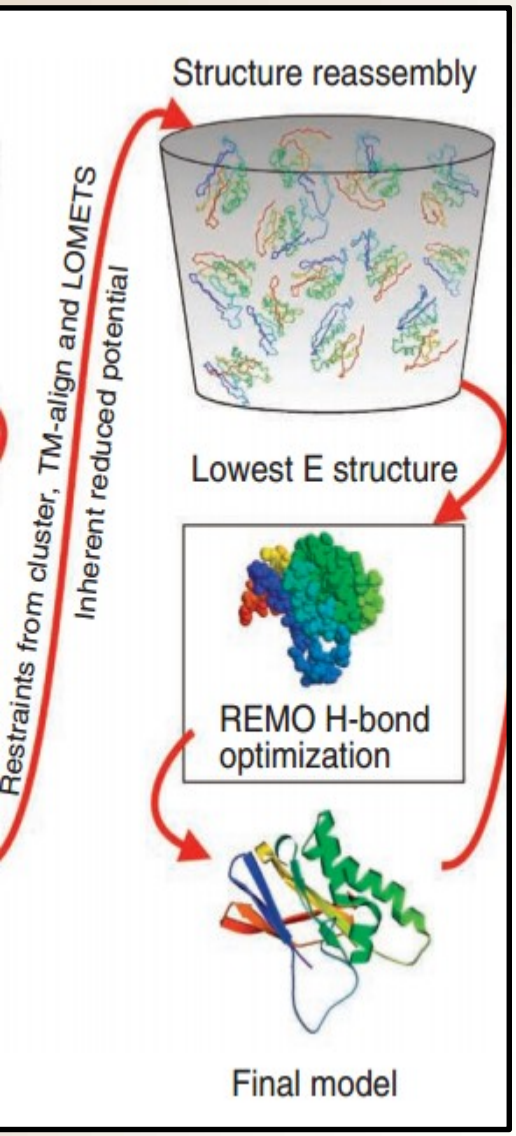
构象中心：聚类构象三维坐标的平均。

力场包括：

来源于PDB的一般统计值（氢键，Cα/侧链相关）。

来源于SVMseq的基于序列的接触预测。

第三步：模板优化



TM-比对：

确定PDB中与聚类中心结构的结构相似性。

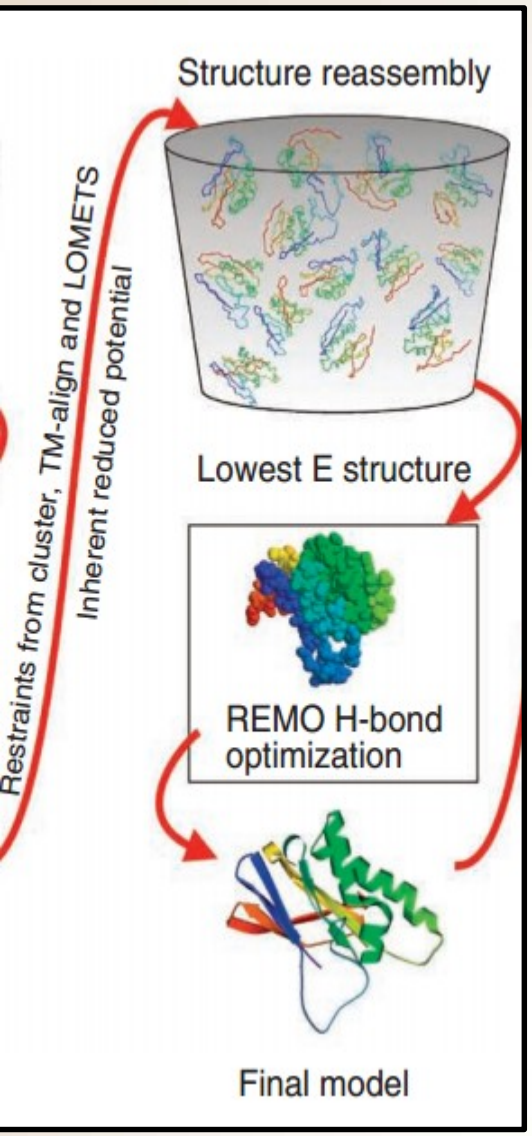
$$TM_{score} = \frac{1}{L}\sum_{i=1}^{L}\frac{1}{1+{d_i}^2 \Big/ {d_0}^2} \qquad d_0 = 1.24\sqrt[3]{L-15} - 1.8$$

TM分数~[0,1]，数值越高，表示模型更好。

Ten proteins in PDB, which are structurally closest to the first I-TASSER model

Rank	TM-score	RMSD[a]	IDEN[a]	Cov.	PDB hit	Structural alignment using TM-align
1	0.8950	1.53	0.36	0.97	2jbzA Model 1	HMSIIGVGIDVAEVERFGAALERT-PALAGR ---MFGVGIDIIEIDRIRKSYQTYGDRFLKK
2	0.8879	1.31	0.38	0.95	1f71A Model 1	GIYGIGLDITELKRIASMAGRQ-KRFAERIL -MFGVGIDIIEIDRIRKSYQTYGDRFLKKIF
3	0.8765	1.31	0.27	0.95	2bddA Model 1	QGHHIIGIGTDILCVNRIYKILEKNI-NFIK ----MFGVGIDIIEIDRIRKSYQTYGDRFLK

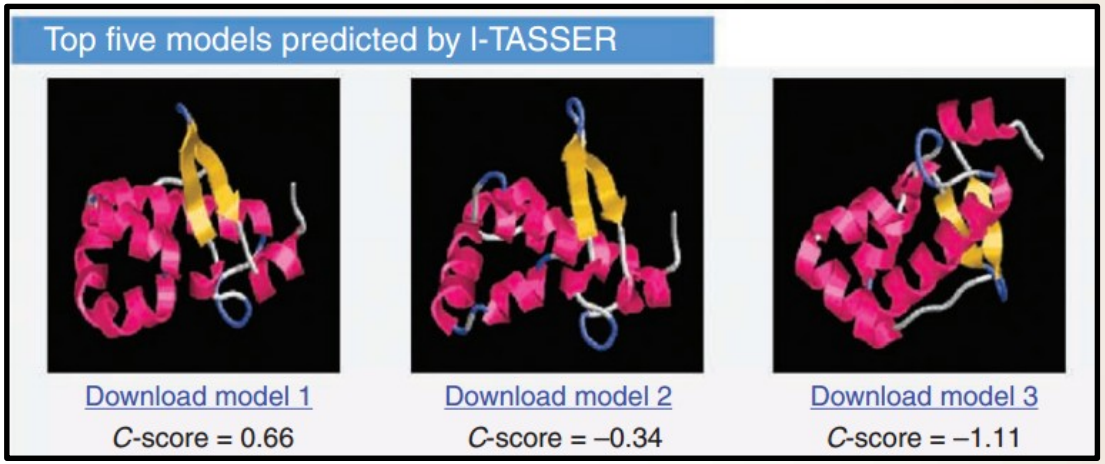
第三步：模板优化



第二轮模拟：确定具有最低自由能的结构。

REMO：建立全原子模型并优化氢键网络。

C-分数>-1.5，可以生成90%正确的拓扑结构。



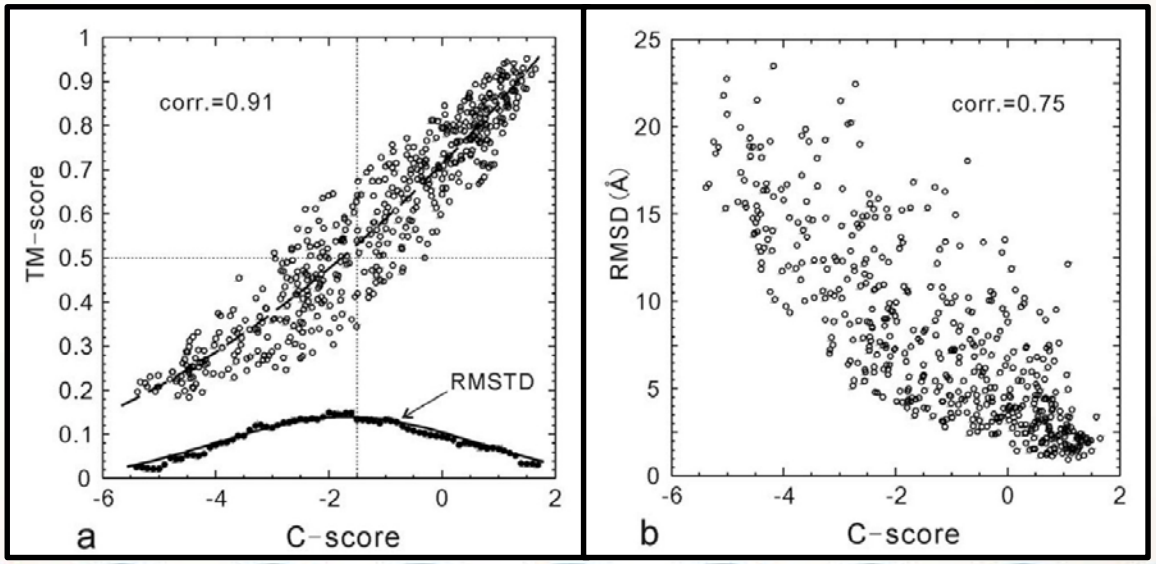
估计预测精度：C-分数

在Z-分数和Monte Carlo模拟收敛的基础。

结合C-分数和L，RMSD的平均误差可以2Å 。

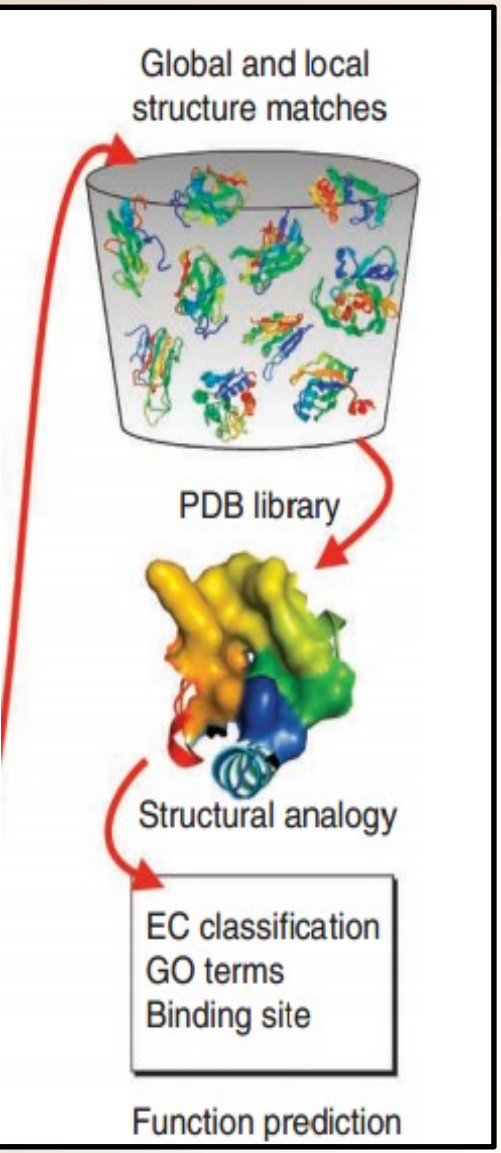
$$C\_score = \ln\left(\frac{M}{M_{tot}} \times \frac{1}{< RMSD >} \times \frac{1}{7}\sum_{i=1}^{7}\frac{Z(i)}{Z_0(i)}\right)$$

通过最大化最终模型与晶体结构的相关性和质量，进行因子和其系数的优化。



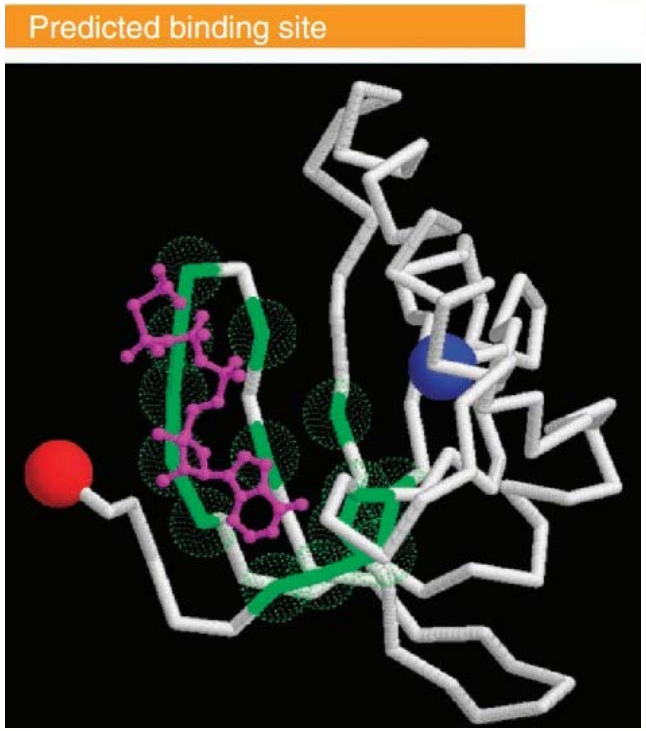
Pearson 相关

第四步：功能注释



局部相似性搜索寻找保守的序列和三维结构。

使用TM-分数评估拓扑相似性。



预测的结合位点残基具有最高的置信度。

使用不同的模型，不同的预测。

第四步：功能注释

Predicted EC numbers

Rank	TM-score	RMSD[a]	IDEN[a]	Cov.	EC-score	PDB hit	EC no.
1	0.8639	1.44	0.38	0.95	1.6627	1f80C	2.7.8.7
2	0.8753	1.70	0.37	0.97	1.7830	2jcaC	2.7.8.7
3	0.8288	1.54	0.32	0.90	1.7315	1fthC	2.7.8.7
4	0.7393	2.62	0.22	0.92	1.3020	2cq5A	1.1.1.100
5	0.5126	3.58	0.15	0.70	1.0700	1iv4F	4.6.1.2

酶功能预测：
酰基载体蛋白合酶

Consensus prediction of Gene Ontology terms

Molecular function		Biological process		Cellular location	
GO term	GO-score	GO term	GO score	GO term	GO score
GO:0000287	1.124	GO:0009059	1.124	GO:0005737	0.988
GO:0008897	1.124	GO:0006631	0.822	GO:0005829	0.302
		GO:0006633	0.822	GO:0043231	0.130
		GO:0046394	0.822	GO:0009532	0.066
		GO:0032787	0.822	GO:0009570	0.066
		GO:0017000	0.250		

分子功能：酰基载体蛋白合成酶，Mg^{2+}的结合。

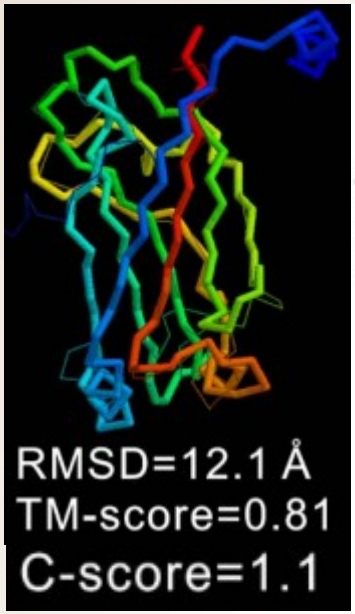
生物过程：高分子合成代谢，脂肪酸代谢。

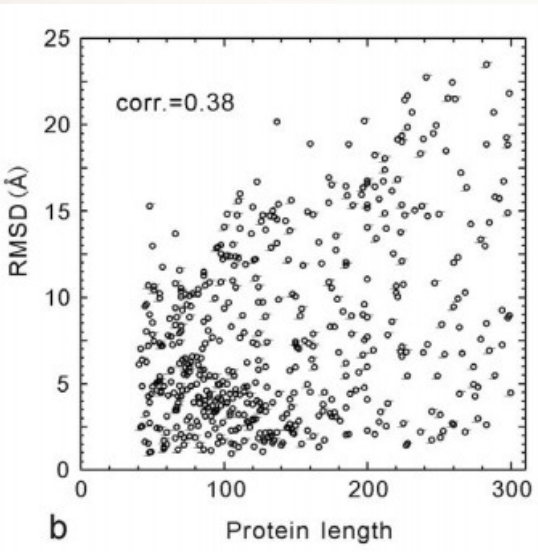
细胞定位：细胞质

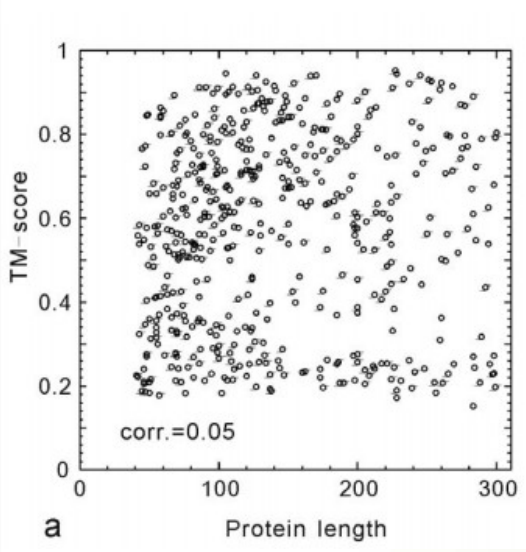
RMSD VS TM-score

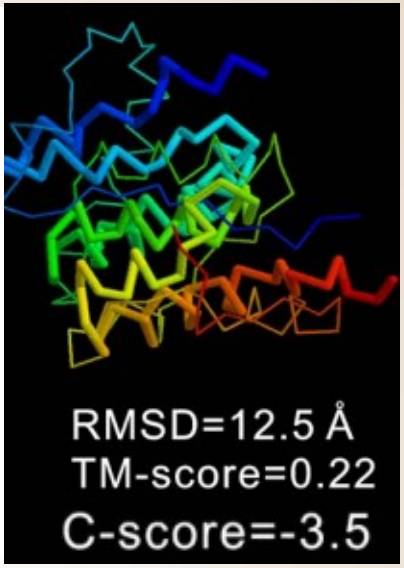
$$RMSD = \sqrt{R^2{}_A + R^2{}_B - 2\frac{\sum_i r_{Ai} \cdot r_{Bi}}{\sqrt{\sum_i r^2{}_{Ai} \sum_i r^2{}_{Bi}}} R_A R_B}$$

$$TM_{score} = \frac{1}{L}\sum_{i=1}^{L} \frac{1}{1 + {d_i}^2 / {d_0}^2}$$









RMSD平均地衡量所有的残基距离，取决于蛋白质的长度，因此对拓扑变化不敏感。

TM-打分在衡量小的距离时增加权重，因此能够更好地衡量结构的相似性。

结论

- 仔细的进行统计评估和使用数据。
- 非物理的，但是使用性强。
- 基于模板和实验数据的预测。
- 全局与局部的相似性检测。
- 一致性 threading和适当的估计函数。

准确性和可靠性！


Thank you!

实例

a

Submitted sequence

```
>your_protein
MFGVGIDIIEIDRIRKSYQTYGDRFLKKIFTEGERVYCFSKSNPYASLAARFAAKEAVAK
ALGTGIGKLLKWKEIEMCRDSRQPQVVVPEALLCSLGVKR
```

User-specified restraint

b

Predicted secondary structure

```
                    20        40        60        80        100
                    |         |         |         |         |
Sequence    MFGVGIDIIEIDRIRKSYQTYGDRFLKKIFTEGERVYCFSKSNPYASLAARFAAKEAVAKALGTGIGKLLKWKEIEMCRDSRQPQVVVPEALLCSLGVKR
Prediction  CCCCCCCCCCHHHHHHHHHHHHHHHHHCCCHHHHHHHHCCCHHHHHHHHHHHCCCCCCCCCSSSSSSSSSSSSCCCSSSSSSCCCHHHHCCCCC
Conf. score 97988748983899999987336999986999999998708778999999889999999985657657857045999857999709988899987299 71
```

c

Top five models predicted by I-TASSER

Download model 1

C-score = 0.66

Download model 2

C-score = –0.34

Download model 3

C-score = –1.11

Download model 4

C-score = –2.44

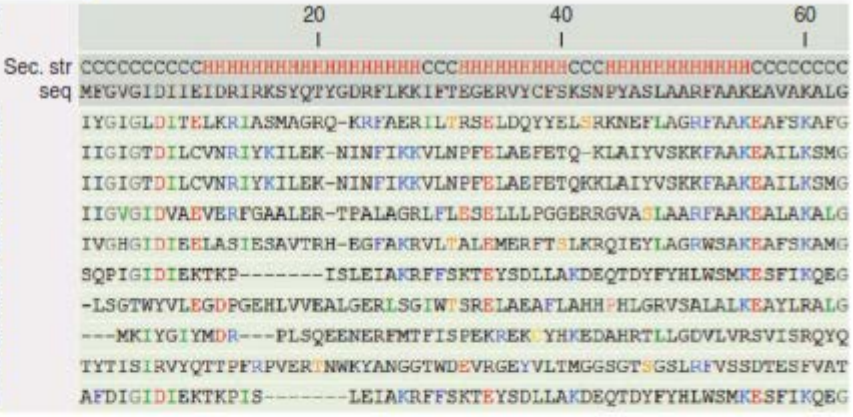
Download model 5

C-score = –3.34

a

Top ten templates used by I-TASSER

Rank	PDB hit	Iden1	Iden2	Cov.	Norm. Z-score	Download align.
1	1f71A	0.37	0.34	0.94	3.44	Download
2	2bddA	0.28	0.26	0.93	4.49	Download
3	2qg8A	0.28	0.23	0.98	4.57	Download
4	2jcaA	0.35	0.33	0.97	2.89	Download
5	1fteA	0.30	0.27	0.92	2.32	Download
6	1qr0A	0.17	0.13	0.87	2.07	Download
7	2z0rA	0.21	0.23	0.91	0.82	Download
8	1qr0A	0.13	0.21	0.83	0.74	Download
9	1y2tA	0.13	0.23	1.00	0.43	Download
10	1qr0A	0.22	0.21	0.86	0.80	Download



b

Ten proteins in PDB, which are structurally closest to the first I-TASSER model (identified by TM-align)

Rank	TM-score	RMSD[a]	IDEN[a]	Cov.	PDB hit	Structural alignment using TM-align
1	0.8950	1.53	0.36	0.97	2jbzA Model 1	
2	0.8879	1.31	0.38	0.95	1f71A Model 1	
3	0.8765	1.31	0.27	0.95	2bddA Model 1	
4	0.8504	1.51	0.32	0.93	1fteA Model 1	
5	0.8144	1.15	0.34	0.87	3f09A Model 1	
6	0.6456	3.03	0.14	0.89	2bydA Model 1	
7	0.6044	2.77	0.17	0.82	1qr0A Model 1	
8	0.5355	3.64	0.16	0.72	2uzhA Model 1	
9	0.5312	3.32	0.11	0.71	3f0dA Model 1	

a

Predicted EC numbers

Rank	TM-score	RMSD[a]	IDEN[a]	Cov.	EC-score	PDB hit	EC no.	
1	0.8639	1.44	0.38	0.95	1.6627	1f80C	2.7.8.7	Consensus prediction of acyl-carrier protein synthase
2	0.8753	1.70	0.37	0.97	1.7830	2jcaC	2.7.8.7	
3	0.8288	1.54	0.32	0.90	1.7315	1fthC	2.7.8.7	
4	0.7393	2.62	0.22	0.92	1.3020	2cq5A	1.1.1.100	Acyl-carrier protein reductase
5	0.5126	3.58	0.15	0.70	1.0700	1iv4F	4.6.1.2	

b

Predicted Go terms

Rank	TM-score	RMSD[a]	IDEN[a]	Cov.	PDB hit	Fh-score	Associated GO terms
1	0.8620	1.38	0.38	0.55	1f80C	1.50	GO:0000287 GO:0008897 GO:0006631 GO:0009059 GO:0032787 GO:0046394 GO:0005623 GO:0005737
2	0.8659	1.84	0.37	0.57	2jcaC	1.49	GO:0000287 GO:0008897 GO:0006631 GO:0009059 GO:0032787 GO:0046394 GO:0005623 GO:0005737
3	0.8907	1.44	0.28	0.57	2qq8A	1.39	GO:0000287 GO:0008897 GO:0006631 GO:0009059 GO:0032787 GO:0046394
4	0.8103	1.43	0.32	0.52	1fthC	1.33	GO:0000287 GO:0008897 GO:0006631 GO:0009059 GO:0032787 GO:0046394 GO:0005623 GO:0005737
5	0.7508	2.32	0.22	0.52	2cq5A	1.08	GO:0000287 GO:0008897 GO:0009059 GO:0005622 GO:0005623 GO:0005737 GO:0005829
6	0.7508	2.32	0.22	0.52	2cq5A	1.08	GO:0000287 GO:0008897 GO:0009059 GO:0005622 GO:0005623 GO:0005737 GO:0005829
7	0.6485	2.95	0.14	0.51	2cq5A	0.85	GO:0000287 GO:0008897 GO:0009059 GO:0005622 GO:0005623 GO:0005737 GO:0005829
8	0.6151	2.60	0.17	0.47	1qr0A	0.84	GO:0000287 GO:0008897 GO:0006631 GO:0009059 GO:0017000 GO:0032787 GO:0044249 GO:0046394
9	0.6151	2.60	0.17	0.47	1qr0A	0.84	GO:0000287 GO:0008897 GO:0006631 GO:0009059 GO:0017000 GO:0032787 GO:0044249 GO:0046394
10	0.5590	2.70	0.22	0.47	1qr0A	0.82	GO:0000287 GO:0008897 GO:0006631 GO:0009059 GO:0017000 GO:0032787 GO:0044249 GO:0046394

Consensus prediction of Gene Ontology terms

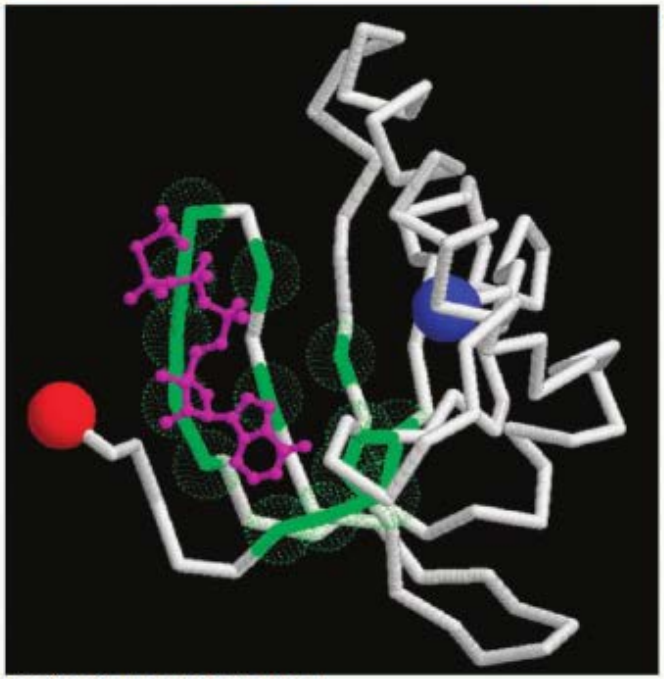
Molecular function		Biological process		Cellular location	
GO term	GO-score	GO term	GO score	GO term	GO score
GO:0000287	1.124	GO:0009059	1.124	GO:0005737	0.988
GO:0008897	1.124	GO:0006631	0.822	GO:0005829	0.302
		GO:0006633	0.822	GO:0043231	0.130
		GO:0046394	0.822	GO:0009532	0.066
		GO:0032787	0.822	GO:0009570	0.066
		GO:0017000	0.250		

High confidence consensus prediction of Gene Ontology terms:

1. Molecular function: acyl-carrier protein synthase, Mg^{2+} binding
2. Biological process: macromolecule anabolism, fatty acid metabolism, fatty acid formation, carboxylic acid biosynthesis
3. Cellular location: cytoplasm

C

Predicted binding site



Identified analogs with similar binding site:

Rank	PDB Hit	TM-score	RMSD[a]	IDEN[a]	Cov.	BS-score	Download binding site residues on the predicted complex model
1	1qaqA	0.7626	2.25	0.24	0.85	0.76	Download 21,22,23,24,25,27,29,42,44,74,96,97,98,99
2	2owbA	0.8161	1.74	0.25	0.84	0.71	Download 19,21,22,23,24,27,28,29,42,44,74,96,97,98,99
3	2hy81	0.7817	2.03	0.24	0.85	0.71	Download 21,22,23,24,29,42,44,74,96,97,98,99,100
4	1wvxA	0.7626	2.38	0.26	0.86	0.70	Download 21,22,29,42,44,74,96,97,98,99100,101
5	1yvjA	0.7252	2.49	0.28	0.85	0.68	Download 21,22,23,24,29,42,44,74,96,97,98,99,100
6	3fupB	0.7249	2.25	0.27	0.83	0.66	Download 21,22,23,24,25,27,28,29,42,74,96,97,98,99
7	3f5pC	0.7617	2.31	0.24	0.87	0.66	Download 1,21,22,29,42,43,44,58,61,65,74,94,95,96,97,98
8	1ql6A	0.8109	1.46	0.27	0.83	0.65	Download 21,22,23,24,25,26,27,29,42,44,74,96,97,98,99
9	2bpmA	0.7727	1.74	0.26	0.83	0.62	Download 19,20,21,22,29,31,42,44,74,96,97,98,99
10	2j0jA	0.7226	2.51	0.25	0.84	0.62	Download 21,22,23,24,29,42,44,74,96,97,98,99

Binding site residues in the model:

LEU: 21 GLY: 22 THR: 23 GLY: 24 GLY: 25 GLY: 27 VAL: 29
ALA: 42 LYS: 44 VAL: 74 MET: 96 GLU: 97 TYR: 98 CYS: 99
LYS: 106

} Predicted binding site residues with highest confidence