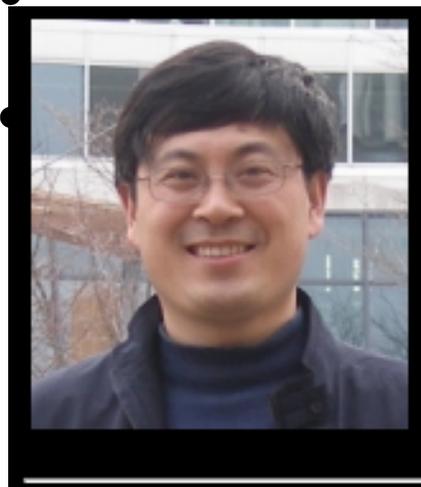


基于CAS@home的蛋白质结构预测网络服务器TreeThreader

报告人：张海仓

中科院计算技术研究所

介绍



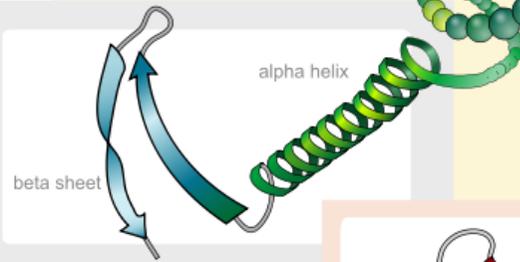
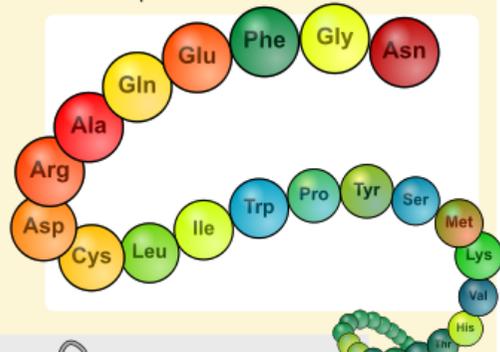
导师：卜东波
生物信息学算法研究，蛋白质结构预测，DNA拼接，
疾病传播网络研究

内容

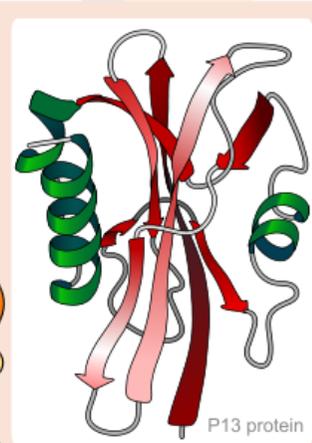
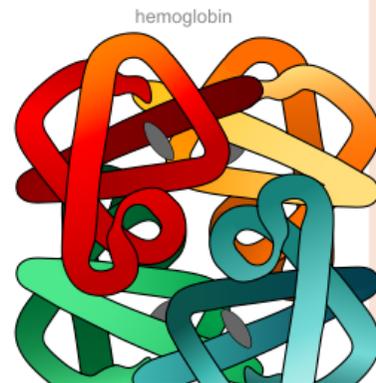
- 蛋白质结构预测
- CAS@home 志愿计算平台
- 基于CAS@home的蛋白质结构预测服务器
TreeThreader
- 结果

蛋白质结构

Primary structure
amino acid sequence



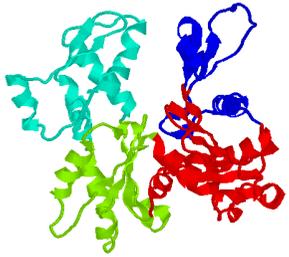
Secondary structure
regular sub-structures



Tertiary structure
three-dimensional structure

蛋白质结构

- 试验方法
 - NMR
 - x-Ray
- 计算方法
 - 从已知结构的蛋白中学习规律，用于预测未知结构的蛋白

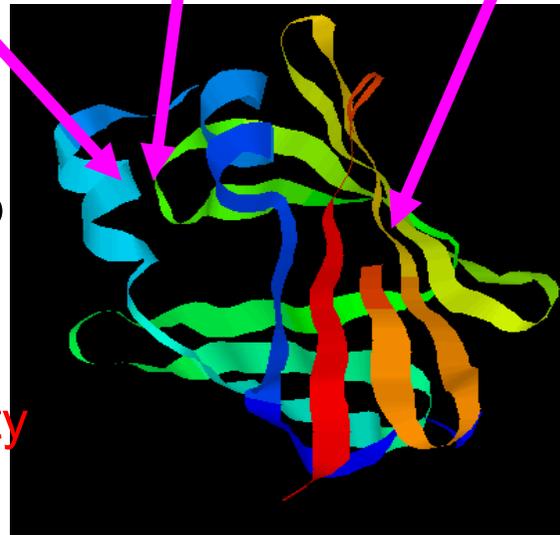


TreeThreader: 条件随机场

...YKLILNGKTKGETTTEAVDAATAEKVFQYANDNGVDGEW...

How preferable to
put two particular
residues nearby: E_p
(pairwise term)

Alignment gap penalty
 E_g



How well a residue
align to another residue
on sequence: E_m
(mutation term)

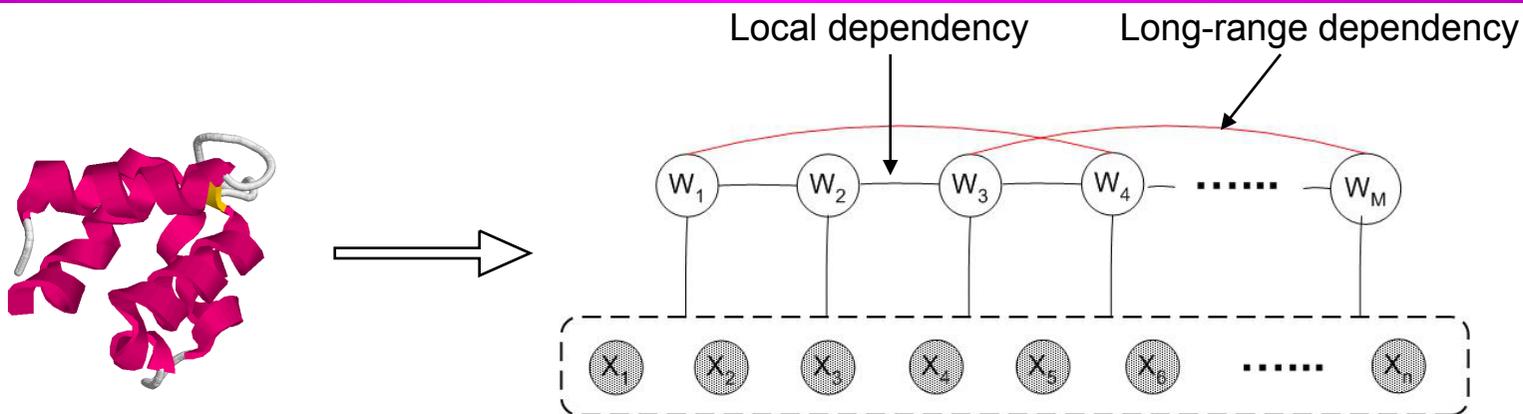
How well a residue
fits a structural
environment: E_s
(singleton term)

Total energy: $E_m + E_p + E_s + E_g$

Describe how sequence fit template



TreeThreader: 条件随机场



- Outputs $Y = \{M, \{W_i\}\}$, where $W_i = \{p_i, q_i, s_i\}$

- Feature definition

➤ 点特征

$$f_k(w_i, x) = f_k'(x, p_i, q_i)I(s_i = s', q_i - p_i + 1 = d')$$

➤ 局部链接特征

$$f_k(w_{i-1}, w_i, x) = I(s_i = s, s_{i-1} = s', p_i = q_{i-1} + 1)$$

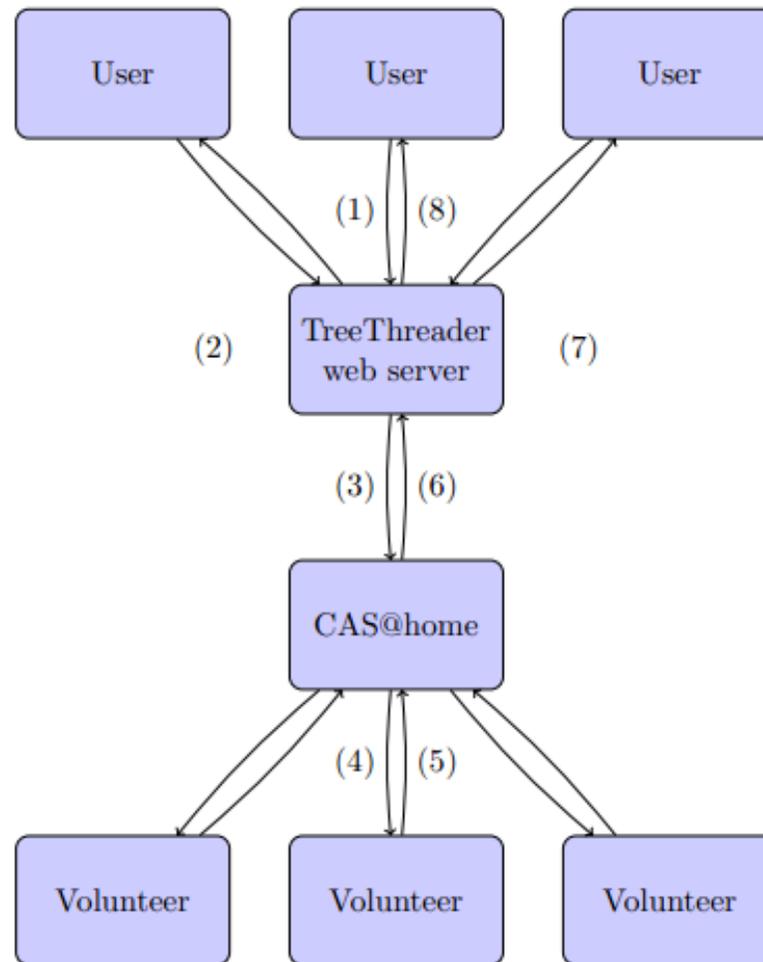
➤ 长程链接特征

$$f_k(w_i, w_j, x) = g_k'(x, p_i, q_i, p_j, q_j)I(s_i = s, s_{i-1} = s')$$

CAS@home 志愿计算平台

- CAS@home志愿计算平台是由中国科学院高能物理研究所的计算中心搭建，致力于为中国的科学家提供志愿计算服务的志愿平台。
- 计算任务提交给CAS@home平台
- CAS@home将计算任务分发给志愿者
- 志愿者从CAS@home得到相应的可执行程序 and 计算数据，完成计算任务后将计算结果返回。

基于CAS@home的蛋白质结构 预测服务器



基于CAS@home的蛋白质结构 预测服务器

- <http://protein.ict.ac.cn/TreeThreader/>

Input Job:

Paste a primary sequence:

or upload a file:

Job ID:

Email Address (Optional):

ThreeThreader服务器预测结果

Top 10 hits

Query:

Job ID : TR720

Length : 202

Sequence: MITEFLKKKLEEHL SHVKEENT IYVTDLVRCPRRVRYESEYKELAI SQVYAPSAILGDILHLGLESVLKGNFNAETEVE
TLREINVGGRVYKIKGRADAIIRNDNGKSI VIEIKTSRSDKGLPLIHHKMQQLQIYLWLFSAERGILVYITPDRIA EYEIN
EPLDEATI VRLAEDTIMLQNSPRFNWECKYCIFS VICPAKLT

No	Hit	MAC Score	Identity	Cols	Query HMM	Template HMM
1	1VE3_B	48.3233	0.111732	179	1-200	19-227
2	3M84_A	46.375	0.116751	197	1-202	20-350
3	2F3D_A	45.4394	0.126263	198	1-202	12-320
4	1OMX_A	44.6366	0.103093	194	1-202	34-292
5	3P4E_A	44.5126	0.12	200	1-200	20-349
6	3TQS_D	44.2244	0.0952381	189	1-202	12-238
7	1YXI_A	44.0103	0.128713	202	1-202	7-306
8	3T5G_A	43.9951	0.13986	143	15-202	1-180
9	1FJ6_A	43.6234	0.111111	198	1-202	11-306
10	1FPD_A	43.0036	0.121212	198	1-202	11-313

图 4.10: TreeThreader网络服务器预测结果Hit列表

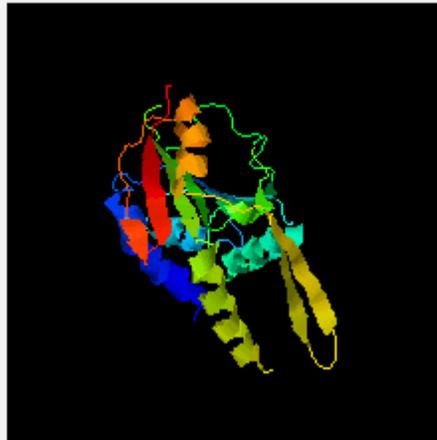
ThreeThreader服务器预测结果

Pairwise Query-Template Alignment

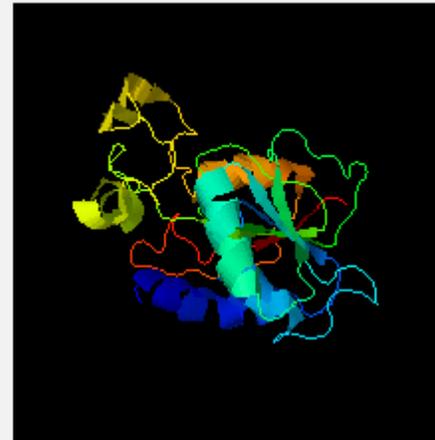
No 1

> 1VE3_B Alignment in PIR format

MAC score = 48.3233 Identity = 0.111732



[Download Partial Structure](#)



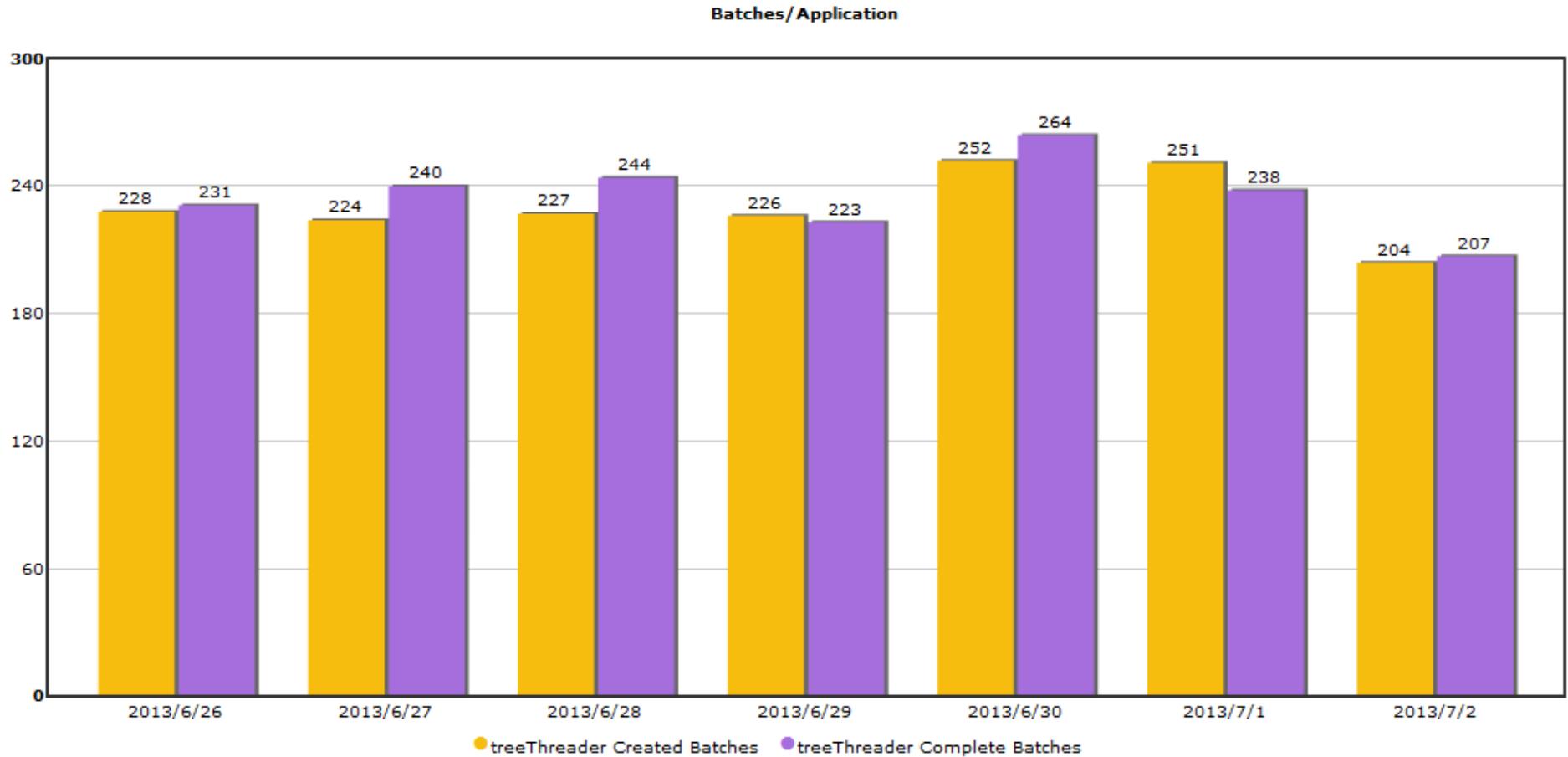
[Download Full Structure](#)

```
Q TR720 1 MITEFLLKKKLEEHLSHVKEENTIIYVIDLVRCPPRRVRYESEYKELAISQVYAPSAIILGDIHLGLGLESVLKG-NFNAETEV 80 (230)
T 1VE3_B 1 QEYRSRIETLEPLLKMYMKRGRKVLDLACGCVGGFSFLE-DYG-F-----EUVGVDISEDMIIRKAREYAKSRESNVEFIV 80 (230)

Q TR720 81 ETLREINVGGRVYKIKGRADAIIRNDN-----GKSIVIEIKTSRSDKGLPLIHHKMLQIYLVWLFSAE-- 160 (230)
T 1VE3_B 81 CDARKLSFEDKTFDYVIFIDSIVHFEPLELNQVFKVRRVLKPSGKFIMYFTD-----LRELLPALK--ESLVVGQKY 160 (230)

Q TR720 161 -----KGIIVYITPDRIA EY EINEPLDEATIVRLAEDTIMLQNSPRFNFWECKYCI FSVICPAK 230 (230)
T 1VE3_B 161 WISKVIPDQEBRTVWVIEFKSEQDSFRVRFNV-WGKTGVELLAKLYFTKEAREKVCNY---SY-LTVYNPK 230 (230)
```

服务器当前状态-完成的任务量



服务器当前状态

- 中国科学院深圳先进技术研究院的生物医学信息技术研究中心已经提交了微生物蛋白质组的**50,000** 多条序列。
- 欢迎更多相关研究人员使用我们的蛋白质结构网络服务器**TreeThreader** 并提供宝贵意见
- <http://protein.ict.ac.cn/TreeThreader/>

感谢

- 所有为TreeThreader 贡献计算资源的志愿者
- 感谢中科院高能所提供的志愿计算平台
- 感谢所有为ThreeThreader作出努力的人

问题和议?

谢谢!