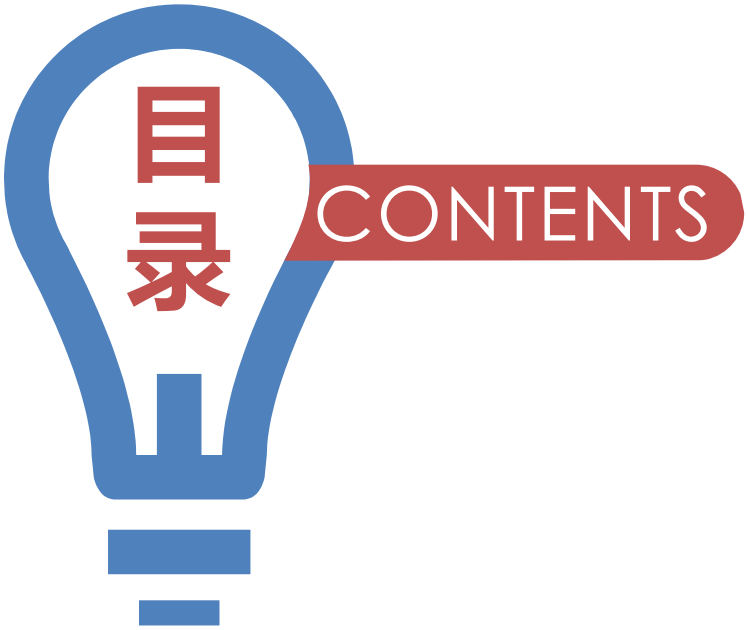
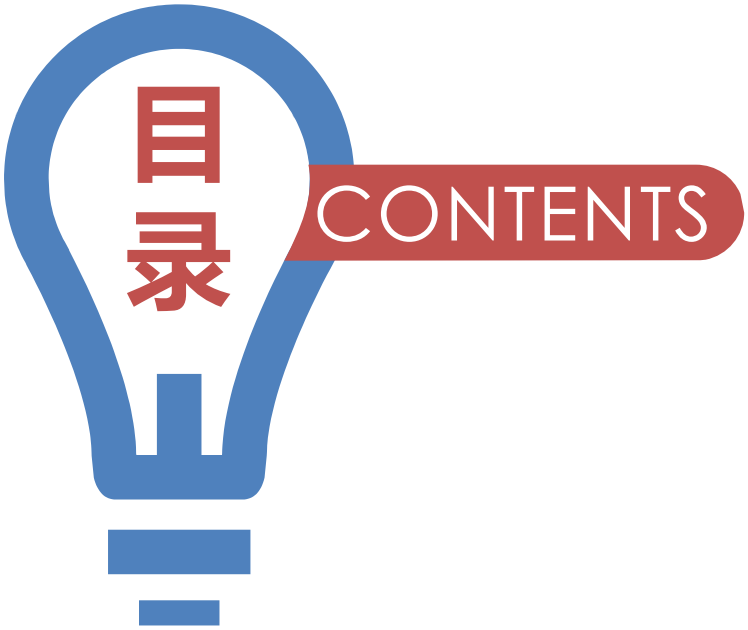
实用生物信息技术课程期末汇报

Structure predictions of WD40-repeat proteins based on homology modeling
基于同源建模的WD40重复蛋白三维结构预测

第14小组
组长：操茂武
组员：陈城杰 张栋 周靖波
汇报人：周靖波

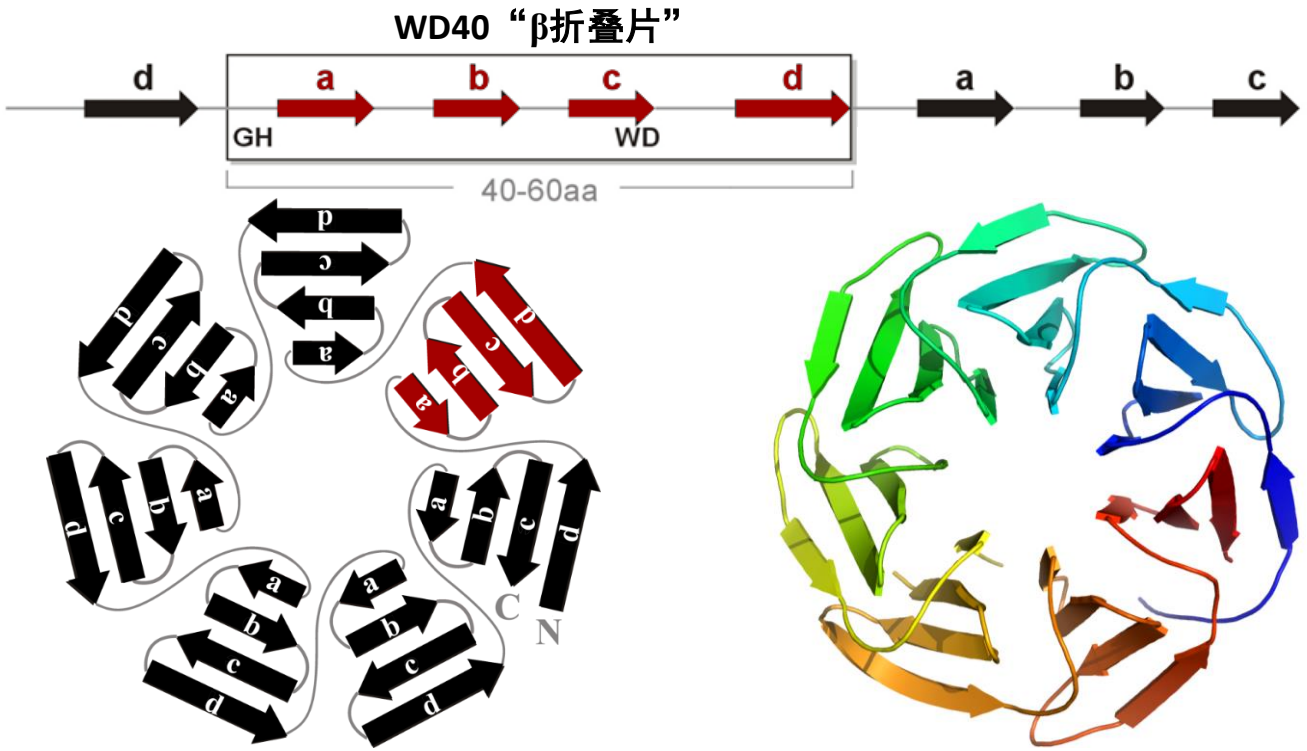
2016年1月7日





研究背景——WD40重复蛋白家族

- **Gβ**是最早发现的WD40蛋白（1995，Neer & Smith，利用序列特征识别出了29个WD40蛋白）
- **WD40重复单元**：保守的“**WD**”氨基酸，**40**-60长度的重复单元
- 1996，Sigler等人解出了第一个WD40晶体结构——β螺旋桨，尼龙搭扣结构



- 真核生物中最大的蛋白家族之一（占人类蛋白中的1%）
- 参与各种生物功能的调控和蛋白质复合体的组装

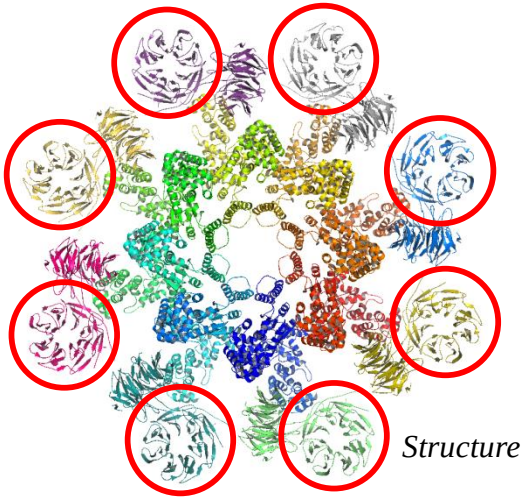
Stirnimann, C. U.; Petsalaki, E.; Russell, R. B.; Muller, C. W. *Trends Biochem Sci* **2010**, *35*, 565-74

研究背景——WD40重复蛋白家族

生物学功能	参与该功能的代表性WD40蛋白
信号转导	Gβ，RACK1，COP1，SPA1-4，PLAP
泛素化降解	DCAFs，DDB1，FBWX7，CDC4，β-TrCP
DNA复制	CDT2，WDHD1，CTF4
DNA损伤修复	DDB2，CSA
DNA转录	Tup1，TLE1，WDR61
大脑神经发育	WDR62，LIS1，LRRK2
组蛋白修饰与识别	WDR5，RbBP5，EED，MET50，RbBP7，TRM82
细胞周期调控	CDC4，CDC20，CDC40，P55，MAD2
RNA加工与剪接	TAFs，CstF，PRP19，RRP9
囊泡运输	SEC13，SEC31，COPA，COPB
程序性细胞凋亡	APAF1，CED-4
染色体组装	CAF1，HIR1，HIR2，MSI1
核孔组装	SEC13，SEH1，Nup43，Nup214
细胞骨架组装	MAP，dynein，Arp2，Arp3

Stirnimann, C. U.; Petsalaki, E.; Russell, R. B.; Muller, C. W. *Trends Biochem Sci* **2010**, *35*, 565-74

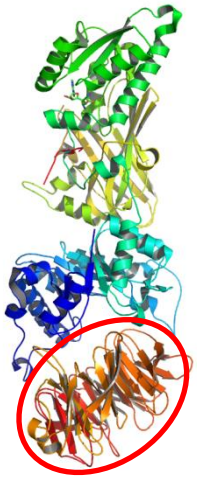
研究背景——WD40重复蛋白家族



Structure **2011**

Apoptosome: Apaf 1(WD40)

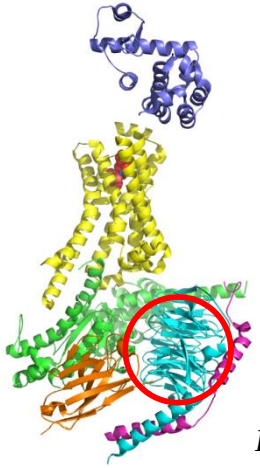
细胞凋亡



PNAS **2012**

PRMT5-MEP50: MEP50(WD40)

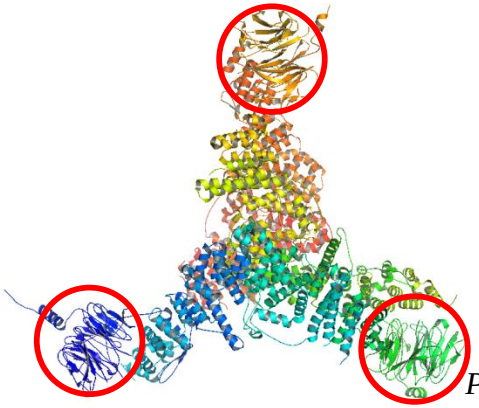
组蛋白修饰与识别



PNAS **2010**

Gα-Gβ-Gγ: Gβ(WD40)

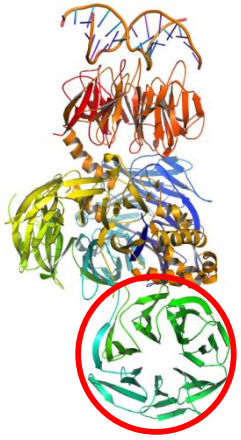
信号转导



PNAS **2009**

Nuclear pore: Sec13(WD40)

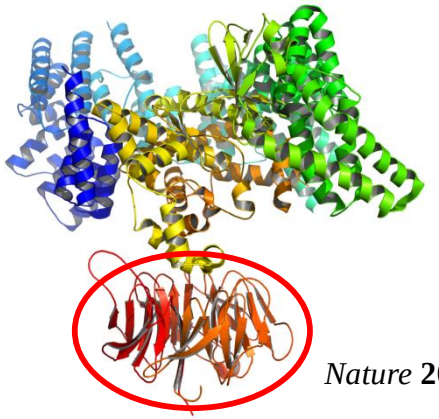
核孔组成与囊泡运输



Cell **2008**

CUL4-DDB1: DDB2(WD40)

DNA复制与修复



Nature **2013**

TOR-LST8:LST8(WD40)

细胞生长

研究背景——蛋白质结构是连接序列与功能的桥梁

蛋白质结构

- 理解蛋白质生物功能的分子机制
- 基于结构的药物设计
- 蛋白质设计与蛋白质工程
- 研究蛋白折叠
- 帮助点突变实验设计
- 理解疾病相关突变
- 理解蛋白质-蛋白质相互作用

获得蛋白质的结构尤为重要！

研究背景——WD40领域的困难

实验研究与理论研究都遇到困难

- 晶体结构难以获得（难表达、纯化）
- 生物功能研究体系难以建立（无催化活性）
- 序列识别困难（序列相似度低）
- 无针对WD40蛋白设计的药物（WDR5抑制剂，CDC4别构抑制剂）

如何获得晶体结构呢？

研究背景——获得蛋白质结构的方法

实验手段：X射线晶体学衍射、核磁共振（NMR）

缺点：1.繁琐而复杂　　2.需要昂贵的设备
　　3.技术流程繁杂　　4. 费时费力

WD40蛋白家族：3万多条蛋白质序列
　　<100个晶体结构

理论计算预测蛋白质结构的方法应运而生！

方法：1.同源建模（homology modeling）
　　2.折叠识别(fold recognition)
　　3.从头算法(de novo prediction）

研究背景——蛋白质结构预测及其应用范围

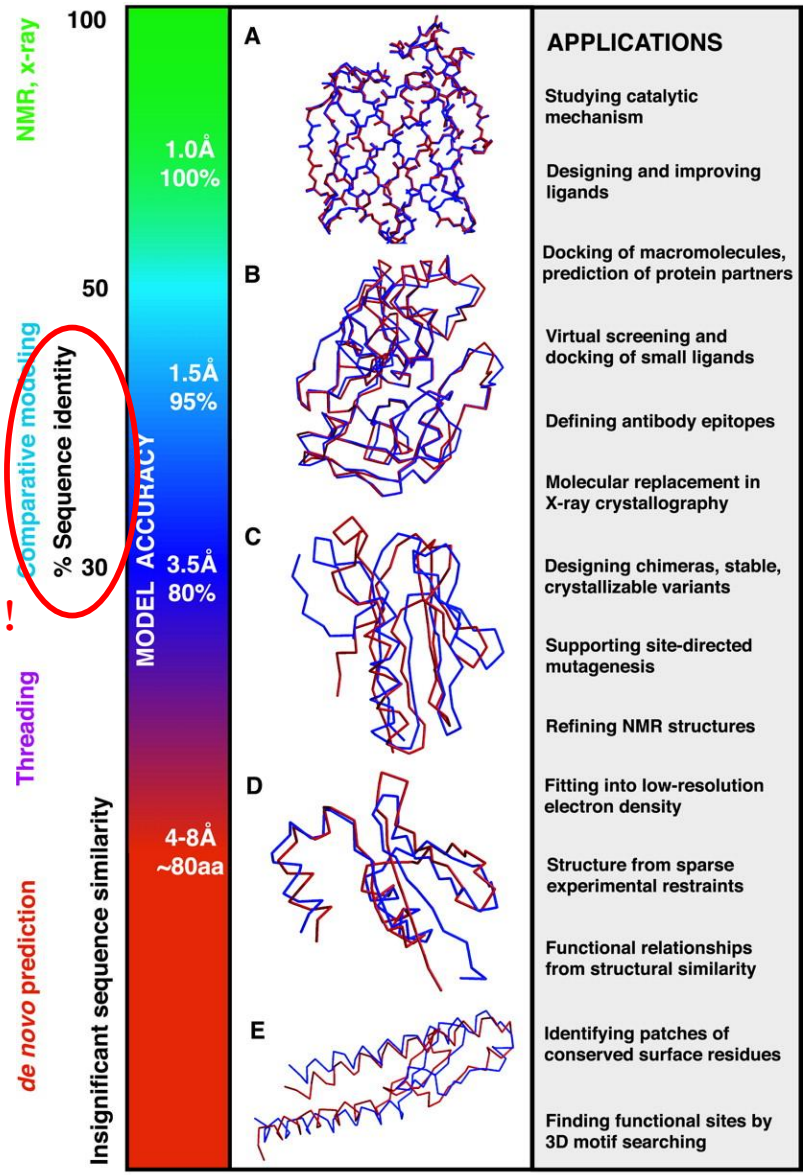
实验结构

比较模建
（同源模建）

最可靠的方法！

折叠识别
（穿线法）

从头算法
（自由模建）



催化机理研究

药物设计

药物筛选

分子置换

突变实验设计

酶活性位点识别

疾病相关突变识别

可变剪接识别

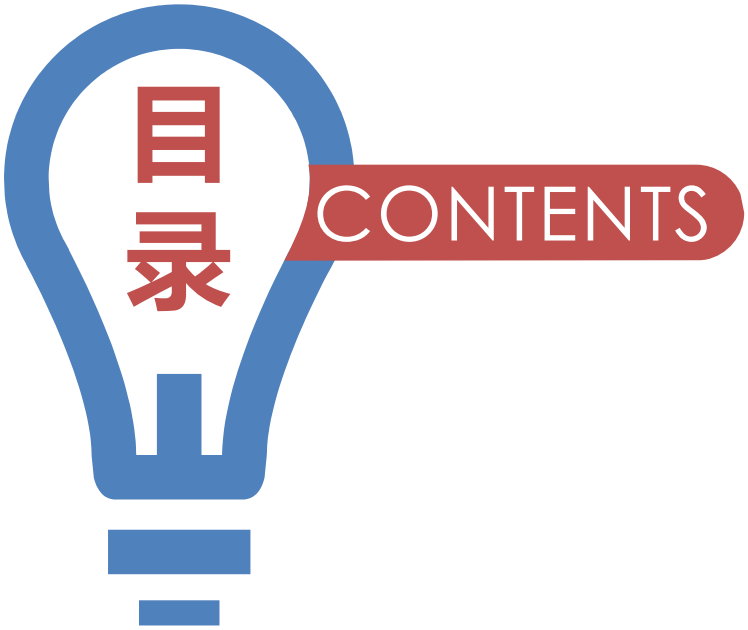
结构域边界

拓扑结构识别

家族指认

……

Baker, D.; Sali, A. *Science* **2001**, 294, 93-96



研究背景 1

同源建模原理 2

MODELLER建模 3

Phyre2 4

同源建模原理

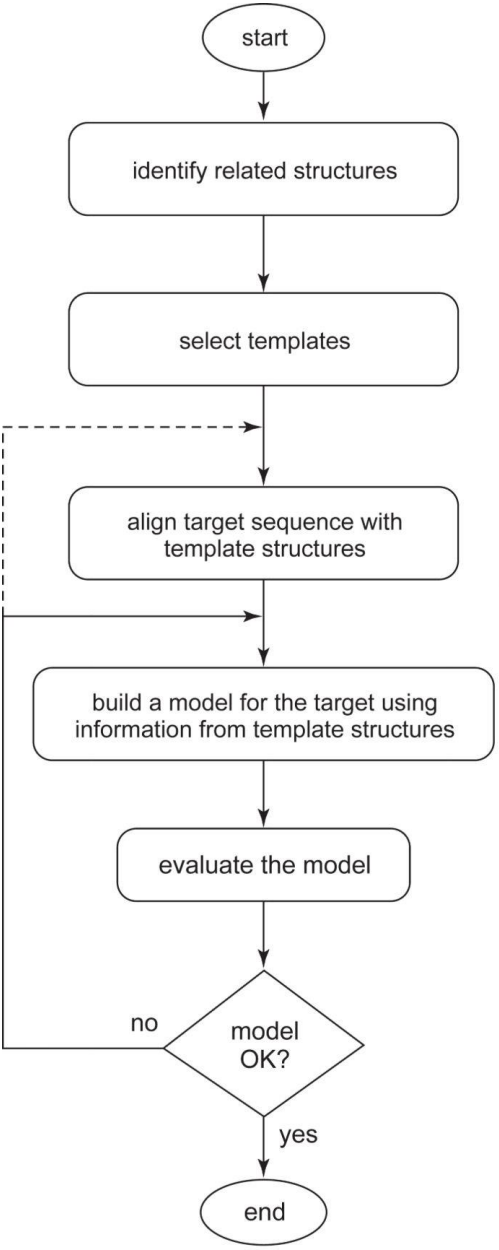
氨基酸序列决定三维结构，结构比序列更保守。

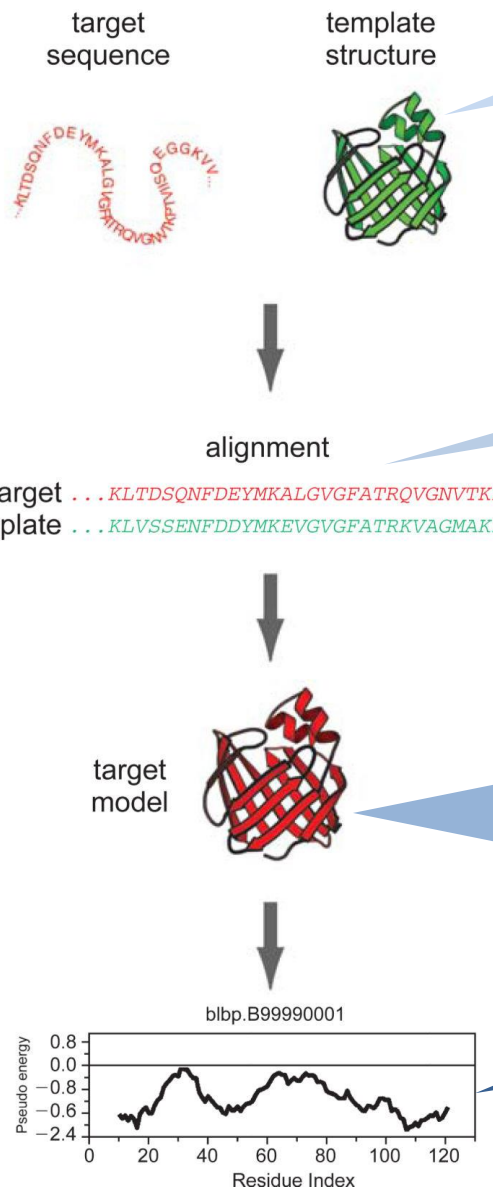
天然蛋白折叠方式是有限的

当两个序列具有很高的相似性时，则它们很可能具有相似的三维结构。

根据已知三维结构的蛋白质（template）来构建未知的蛋白质三维结构（target）

同源建模原理——基本步骤





1.挑选模版(Template Selection)

2. 序列比对（Sequence Alignment）

3. 基本骨架模型的建立(Backbone Model Building)；
4.连接环的建模(Loop Modeling)
5.侧链精修(Side Chain Refinement)
6.模型精修(Modeling Refinement)

7.模型评估（Model Evaluation）

同源建模原理——基本步骤

1.挑选模板（Template Selection）

在蛋白质结构数据库（ PDB数据库）中挑选合适的同源序列作为建模的模板

程序： BLAST、PSI-BLAST等等

2.序列比对（Sequence Alignment）——最关键的一步

目标蛋白和模板蛋白应该综合使用不同比对算法的程序来找到最佳比对结果

人工检查保证保守的关键残基进行了正确的比对

必要时可进行手动修改来提高比对的质量

同源建模原理——基本步骤

3.基本骨架模型的建立(Backbone Model Building)

待建模蛋白与模版蛋白比对上的残基可以认为其与模板的相应区域具有相似结构。

```
 _aln.pos          10        20        30        40        50        60
template   MKAPVRVAVTGAAGQIGYSLLFRIAAGEMLGKDQPVILQLLEIPQAMKALEGVVMELEDCAFPLLAGL
target     MSEAAHVLITGAAGQIGYILSHWIASGELYG-DRQVYLHLLDIPPAMNRLTALTMELEDCAFPHLAGF
 _consrvd  *     *  ********* *    ** **   * *   * * ** ** **   *     ********* ***
```

序列比对区域的残基相同（*位置）

模板蛋白残基的主链和侧链原子坐标同时复制给待建模蛋白

序列比对区域的残基不相同

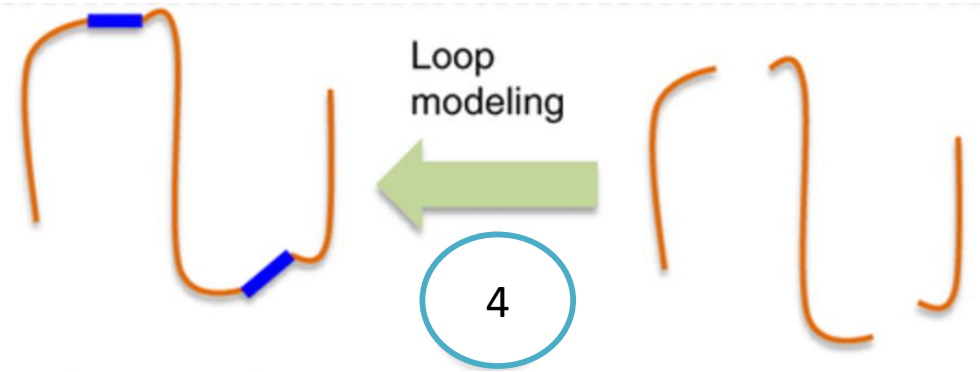
仅复制主链的原子坐标复制给待建模蛋白侧链原子坐标将在后面侧链精修的步骤中进行

同源建模原理——基本步骤

4.连接环的建模(Loop Modeling)————误差的主要来源

序列比对中，往往会出现由于插入或缺失造成的空位，这时需要使用连接环建模来填补结构上的空位。

```
 _aln.p    70        80        90        100       110       120       130
template   EATDDPDVAFKDADYALLVGAAPRL--QV-------NGKIFTEQGRALAEVAKKDVKVLVVGNPANTN
target     VATTDPKAAFKDIDCAFLVASMPLKPGQVRADLISSNSVIFKNTGEYLSKWAKPSVKVLVIGNPDNTN
 _consrvd   ** **   **** * * **    *     **          *   **    *   *    **   ***** *** ***
```



这一步也是同源结构建模中非常困难的一步，是同源建模误差的主要来源。不同软件有不同构建的方法。

同源建模原理——基本步骤

5.侧链精修(Side Chain Refinement)

此步骤需要确定第3步（基本骨架模型的建立）没有构建出的侧链原子的坐标

通过势能计算来构建出能量最低的侧链原子构象的坐标

6. 模型精修(Modeling Refinement)

对模型中出现的结构异常（不利的键角，键长或过密的原子接触等等）进行矫正，使得整个蛋白构象处于能量势处于最低状态。

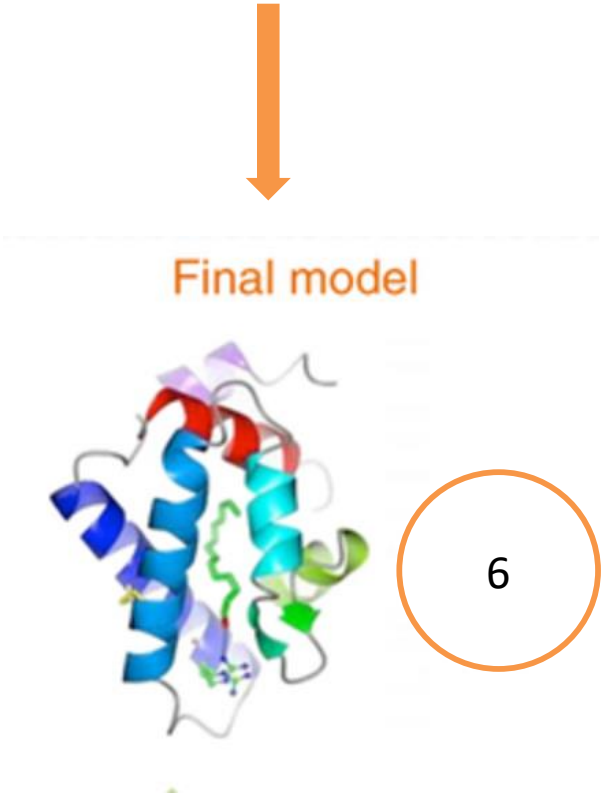
方法：

能量最小化：在不显著改变整体结构的前提下，缓解立体结构的不利碰撞和张力。

分子动态模拟：通过加热或冷却来模拟分子能量上升和下降的运动，可以克服一些能量最小化原理无法接近的能量垒，从而达到全局能量最小化

同源建模原理——基本步骤

1.挑选模板（Template Selection）
2.序列比对（Sequence Alignment）
3.基本骨架模型的建立(Backbone Model Building)
4.连接环的建模(Loop Modeling)
5.侧链精修(Side Chain Refinement)
6.模型精修(Modeling Refinement)



7.模型评估（Model Evaluation）

Procheck、WHAT IF、Verify3D

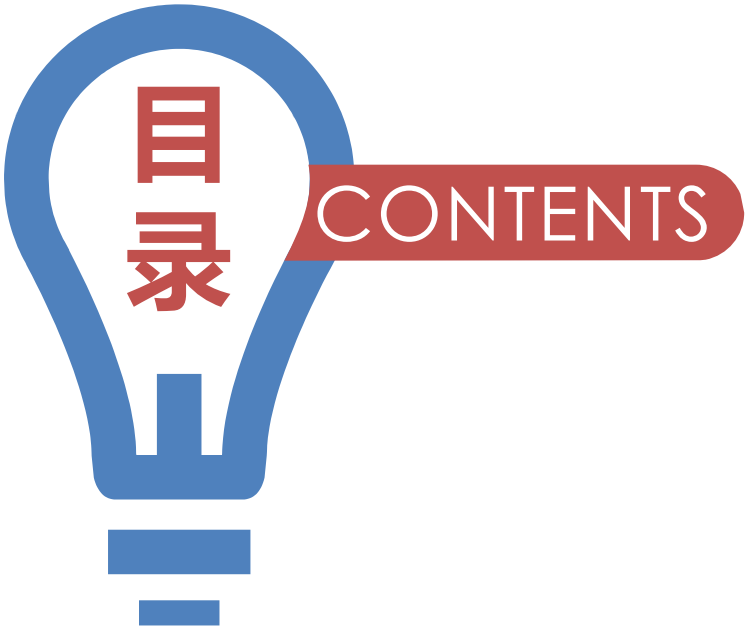
同源建模原理——同源建模的软件

理论计算预测蛋白质结构的方法应运而生！

方法：同源建模（homology modeling）
软件：MODELLER、Phyre2、i-TASSER
Swiss-Model、 HHpred等等

同源建模方法代表性的软件：MODELLER
优点：1.自动化建模程序
2.客观、迅速
3.编译的程序与代码

同源建模方法代表性的软件：Phyre2
优点：1.ease of use
2.user-friendly interface



MODELLER建模——挑选模板（Template Selection）

获得目标序列

蛋白质序列（FASTA格式）

获得模板数据库

PDB数据库  二进制文件

```
>P1;A6NED2
sequence:::::::::
MAEERPGAWEGEGPCGFGQELGSGRGRQVHSPSPLRAGVDICRVSASWSYTAFVTRGGRL
ELSGSASGAAGRCKDAWASEGLLAVLRAGPGPEALLQVWAAESALRGEPLWAQNVVPEAE
GEDDPAGEAQAGRLPLLPCARAYVSPRAPFYRPLAPELRARQLELGAEHALLLDAAGQVF
SWGGGRHGQLGHGTLEAELEPRLLEALQGLVMAEVAAGGWHSVCVSETGDIYIWGWNESG
QLALPTRNLAEDGETVAREATELNEDGSQVKRTGGAEDGAPAPFIAVQPFPALLDLPMGS
DAVKASCGSRHTAVVTRIGELYTWGWGKYGQLGHEDTTSLDRPRRVEYFVDKQLQVKAVT
CGPWNTYVYAVEKGKS*
```

蛋白质序列（PIR格式）

寻找合适的模板

可用BLAST代替

MODELLER建模——挑选模板（Template Selection）

选取最优蛋白结构

```
# Number of sequences:     12
# Length of profile  :    376
# N_PROF_ITERATIONS  :      1
# GAP_PENALTIES_1D   :   -500.0   -50.0
# MATRIX_OFFSET      : -450.0
# RR_FILE            : ${LIB}/blosum62.sim.mat
    1 A6NED2                                  S     0   376     1   376     0     0     0    0.     0.0
    2 1a12A                                   X     1   401   177   333    15   185   148   33.   0.35E-10
    3 4d9sA                                   X     1   383   158   347   107   258   152   43.    0.0
    4 4dnuA                                   X     1   372   158   347   111   262   152   43.    0.0
    5 1i2mB                                   X     1   388   177   372    12   229   187   31.   0.15E-10
    6 4jhnA                                   X     1   361   157   370   131   357   210   27.    0.0
    7 3kciA                                   X     1   366   166   354    95   299   181   34.    0.0
    8 4l1mA                                   X     1   358   169   354    80   285   176   30.    0.0
    9 3mvdK                                   X     1   391   160   340     2   191   171   30.   0.95E-08
   10 4o2wA                                   X     1   365   169   346    98   290   178   30.    0.0
   11 3of7A                                   X     1   438   147   362   165   412   210   26.   0.57E-09
   12 3qhyB                                   X     1   271   166   331    32   171   139   36.   0.79E-05
```

MODELLER建模——序列比对（Sequence Alignment）

序列比对

```
 _aln.pos         10        20        30        40        50        60
4d9sA     PRKVLIISAGASHSVALLSGDIVCSWGRGEDGQLGHGDAEDRPSPTQLSALDGHQIVSVTCGADHTVA
A6NED2    ---MAEERPGAWFGFGF------CGFGQE----LGSGRGRQVHSPSPLR--AGVDICRVSASWSYTAF
 _consrvd          **            *  *      ** *      **  *    *  *  *      *

 _aln.p   70        80        90       100       110       120       130
4d9sA     YSQSG-MEVYSWGWGDFGRLGHG-NSSDLFTPLPI----KALHGI-RIKQIACGD---SHCLAVTMEG
A6NED2    VTRGGRLELSGSASGAAGRCKDAWASEGLLAVLRAGPGPEALLQVWAAESALRGEPLWAQNVVPEAEG
 _consrvd     *  *      *  **      *  *   *       **           *            **

 _aln.pos  140       150       160       170       180       190       200
4d9sA     EVQSWGRNQNGQLGLGDTEDSLVPQKIQAFE----GIRIKMVAAGAEHTAAVTEDGDLYGWGWGRYGN
A6NED2    EDDPAGEAQAGRLPLLPCARAYVSPRAPFYRPLAPELRARQLELGAEHALLLDAAGQVFSWGGGRHGQ
 _consrvd *    *  * * * *       *             *      ****       *    ** ** *

 _aln.pos    210       220       230       240       250       260       270
4d9sA     LGLGDRTDRLVPERVTSTGGEKMSMVACGWRHTISVSYSGALYTYGWSKYGQLGHGDLEDHLIPHKLE
A6NED2    LGHGTLEAELEPRLLEALQGLVMAEVAAGGWHSVCVSETGDIYIWGWNESGQLALPT-RN-LAEDG-E
 _consrvd ** *     * *       *  *  ** *  *   **  *  *  **   ***        *     *

 _aln.pos      280       290       300       310       320       330       340
4d9sA     ALSNSFISQISGGWRHTMALTSDGKLYGWGWNKFGQVGVGNNLDQCSPVQVRFPDDQKVVQVSCGWRH
A6NED2    TVAREATELNEDG---SQVKRTGGAEDG--------APAPFIAVQPFPALLDLPMGSDAVKASCGSRH
 _consrvd             *          *   *                *  *     *     *  *** **

 _aln.pos        350       360       370       380       390       400
4d9sA     TLAVTERNNVFAWGRGTNGQLGIGE--SVDRN--FPKIIEA-LSVDGAS-GQHIESS-NIDPS-S
A6NED2    TAVVTRTGELYTWGWGKYGQLGHEDTTSLDRPRRVEYFVDKQLQVKAVTCGPWNTYVYAVEKGKS
 _consrvd *  **       ** *  ****     * **           * *     *            *
```

MODELLER建模——同源建模

基本骨架模型的建立(Backbone Model Building)
连接环的建模(Loop Modeling)
侧链精修(Side Chain Refinement)
模型精修(Modeling Refinement)

MODELLER自动化建模

Pymol、Swiss-pdb Viewer

同源建模三维结构图（红色为loop，绿色为β-strand）

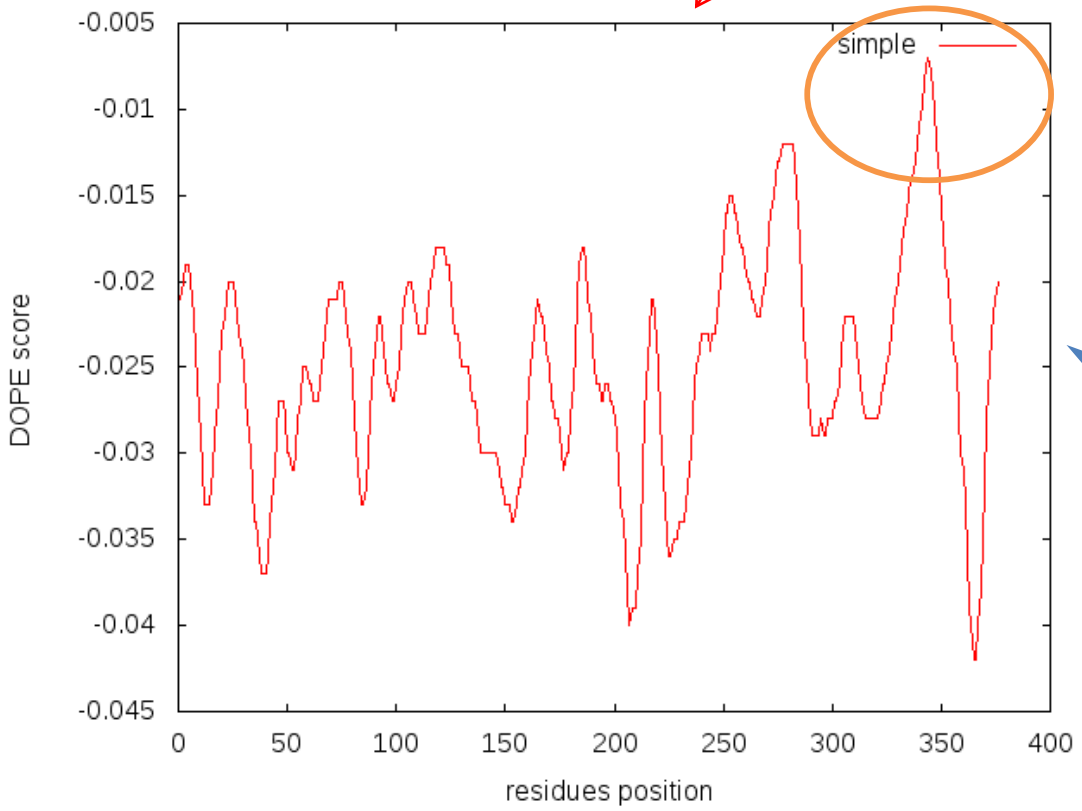
MODELLER建模——单个蛋白序列的三维结构同源建模

评估模型

```
>> Summary of successfully produced models:
Filename                          molpdf      DOPE score
----------------------------------------------------------
A6NED2.B99990001.pdb          3442.99487   -29768.96094
A6NED2.B99990002.pdb          3374.68530   -29802.24023
A6NED2.B99990003.pdb          3313.84790   -29519.46875
A6NED2.B99990004.pdb          3306.06274   -29274.86445
A6NED2.B99990005.pdb          3571.24023   -28923.56641
```

评估模型MODELLER中最常用的是DOPE（discrete optimized protein energy）方法

使用gnuplot作图



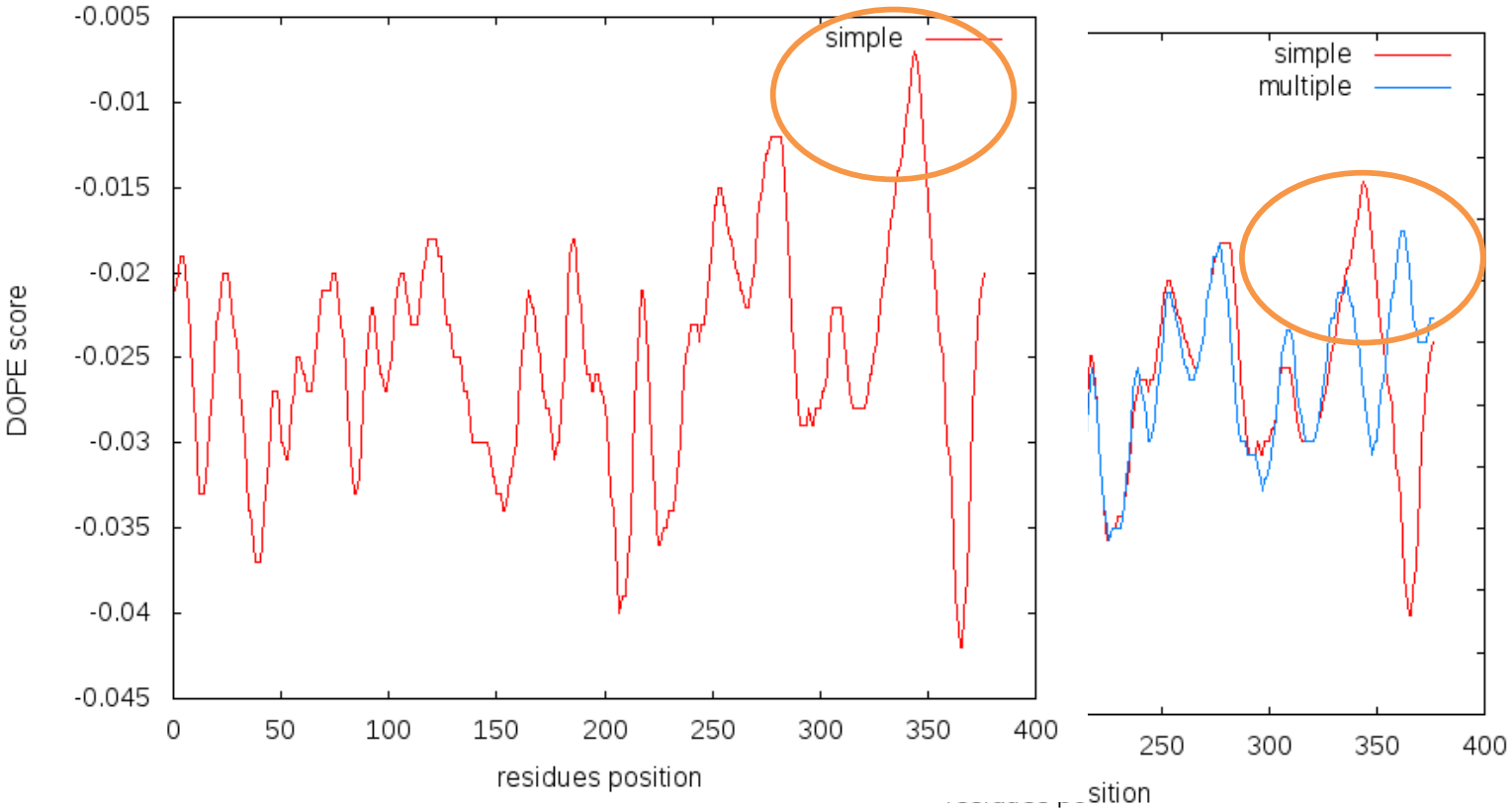
蛋白序列在320-350间电势较高，有一个峰，除这里的模拟效果不好以外其他整体DOPE电势都较低，模拟效果较好。

MODELLER建模——多模板同源建模

多模板建模

采用多模板比对来进行精确化建模以优化结构

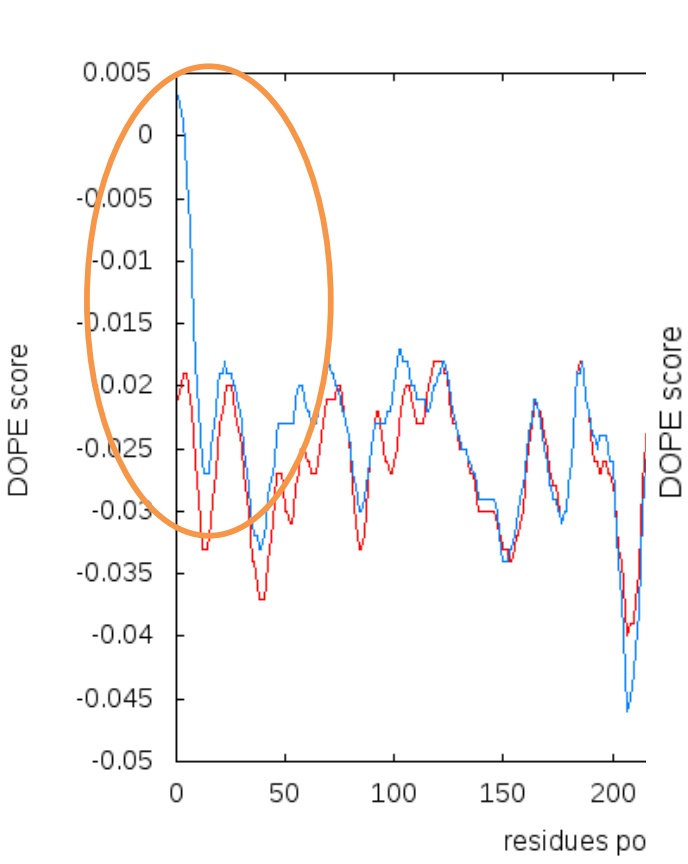
氨基酸序列在320-350间电势较高，有一个峰，这里的模拟效果不好，可以采用多模板比对来进行建模

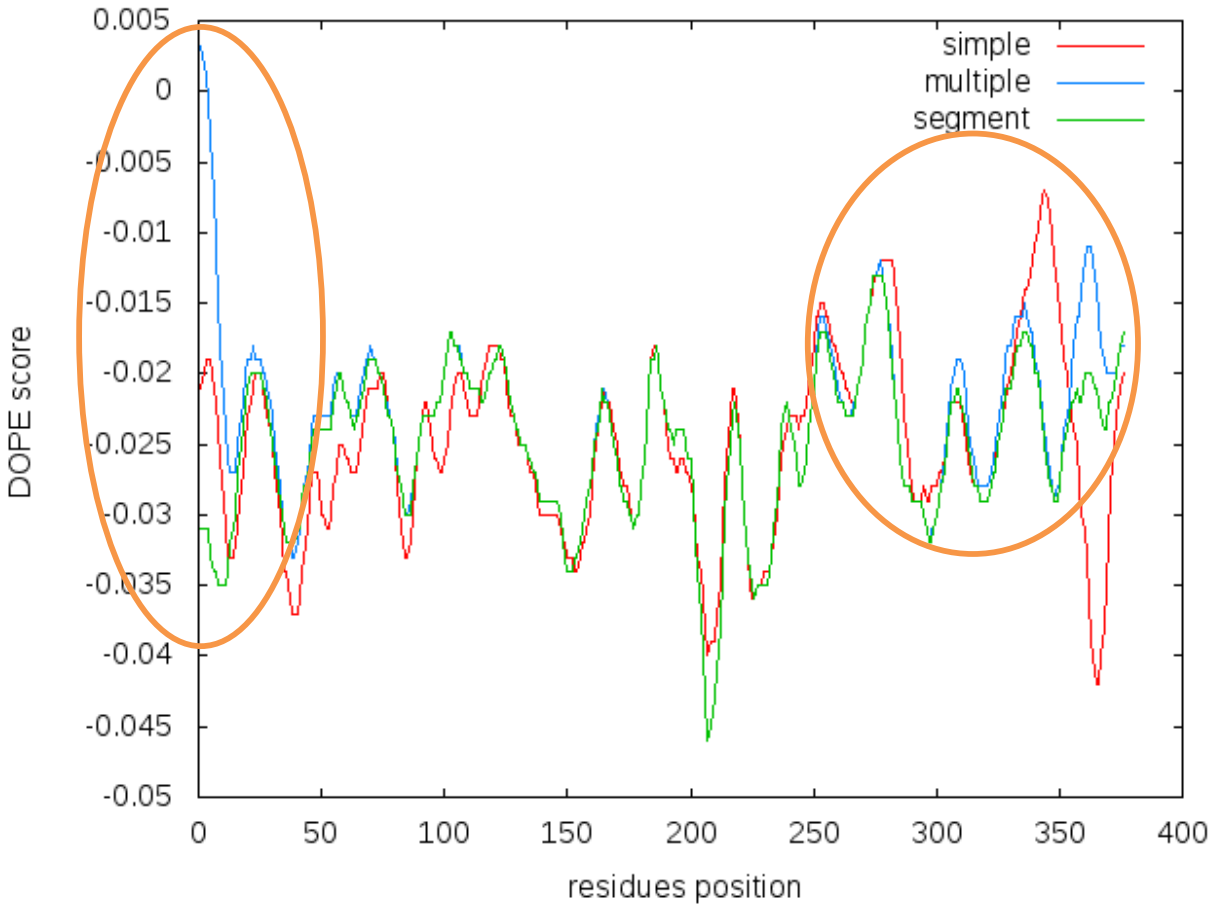


MODELLER建模——局部区域同源建模

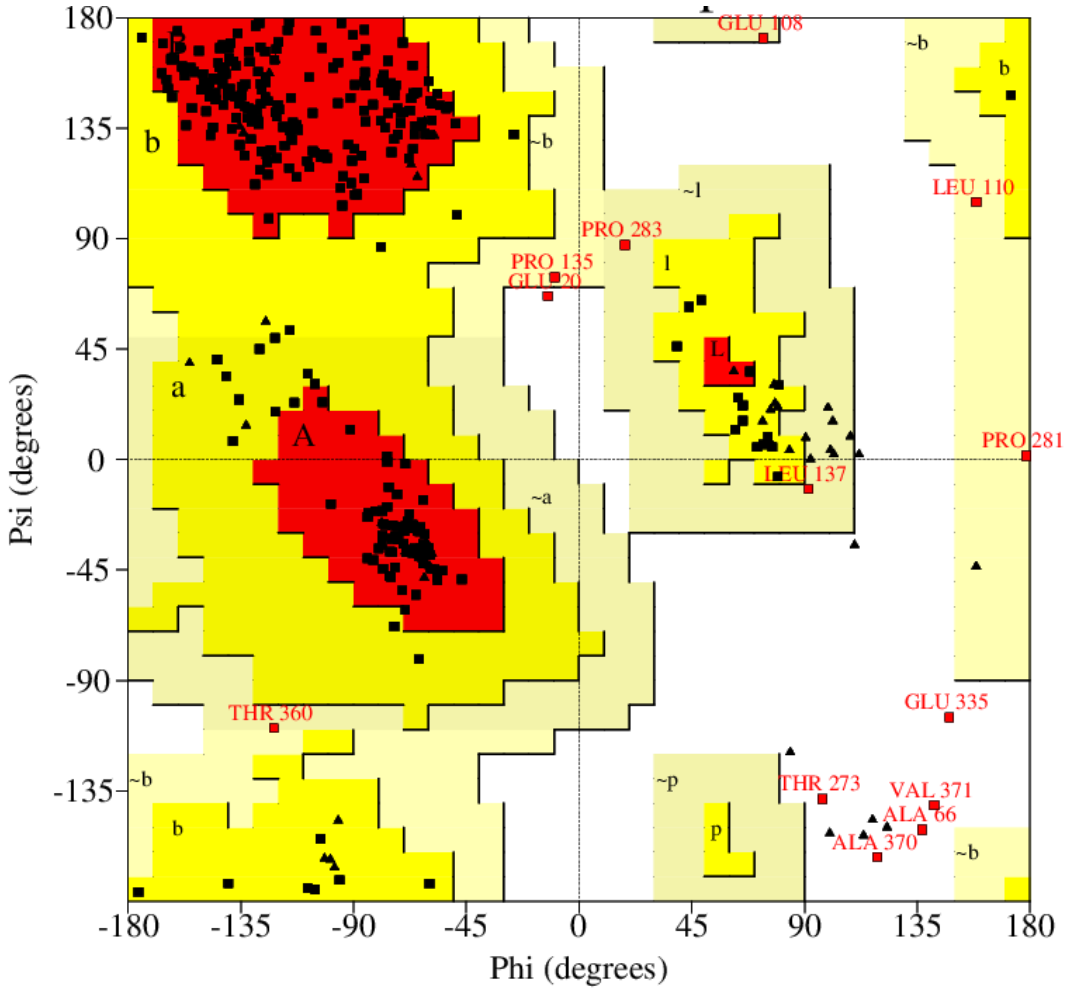
局部区域精确化建模

序列在1-10区域DOPE值仍然很高





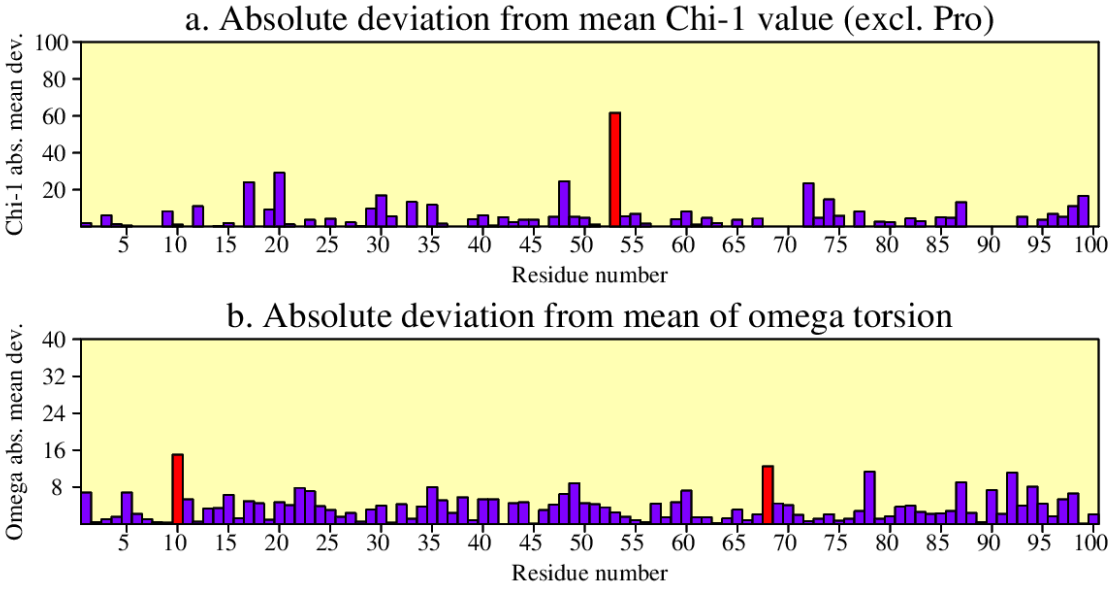
MODELLER建模——模型评估

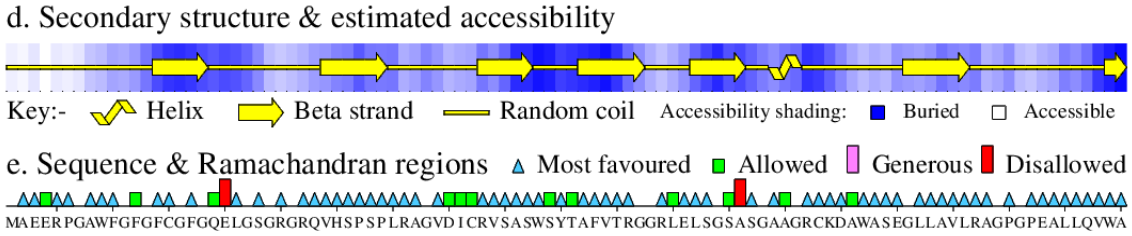


拉氏图不涉及能量，仅仅是检测空间构象是否在天然蛋白所允许的区域内。拉氏图表示天然蛋白中肽单位α碳的两面角（ψ和ϕ），并分别以ψ的角度为横坐标、以ϕ的角度为纵坐标作图，而且规定ϕ,ψ 角允许的构象范围。

红色域是核心区，深黄色为允许区，浅黄色区域是最大允许区，空白区是立体化学不允许区，三角形代表甘氨酸残基，方框代表其他19钟氨基酸残基。研究表明氨基酸残基在核心区、允许区和最大允许区的数量占整个蛋白质的比例高于90%，即可认为该蛋白质模型的构象符合立体化学的规则。

MODELLER建模——模型评估



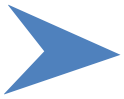


a显示出每个氨基酸残基phi扭转角的绝对偏差（排除脯氨酸），b显示出每个氨基酸残基的ω角的绝对偏差。在a和b中不合理的残基则被标红。

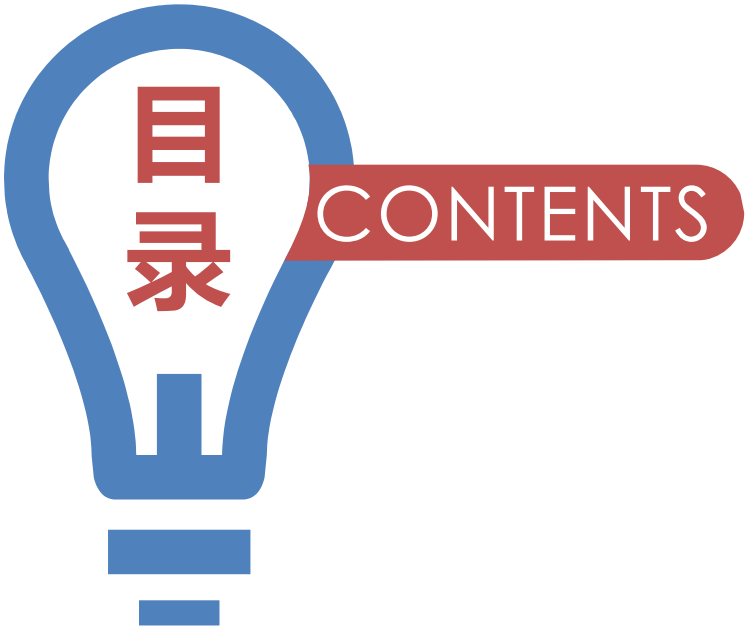
d显示出蛋白质的二级结构，并且也显示出每个残基的溶剂可及性（颜色越深则说明残基是包埋在三维结构中的，颜色越浅说明暴露在外面）。e显示氨基酸序列以及氨基酸残基的拉式区域。由氨基酸残基属性图可以看出绝大部分氨基酸残基在合理的区域内，也证明了此蛋白质模型是理想的。

MODELLER建模——批量化同源建模

单序列建模：
1.时间较长
2.只能单序列建模
3.需手动操作



批量化建模：
1.可处理大量数据
2.自动化建模
3.多线程工作
4.节约大量时间



研究背景 1

同源建模原理 2

MODELLER建模 3

Phyre2 4

Phyre2——简要介绍

理论计算预测蛋白质结构的方法应运而生！

方法：同源建模（homology modeling）
软件：MODELLER、Phyre2、i-TASSER
Swiss-Model、 HHpred等等

同源建模方法代表性的软件：MODELLER
优点：1.自动化建模程序
2.客观、迅速
3.编译的程序与代码

同源建模方法代表性的软件：Phyre2
优点：1.ease of use
2.user-friendly interface

Phyre2——简要介绍

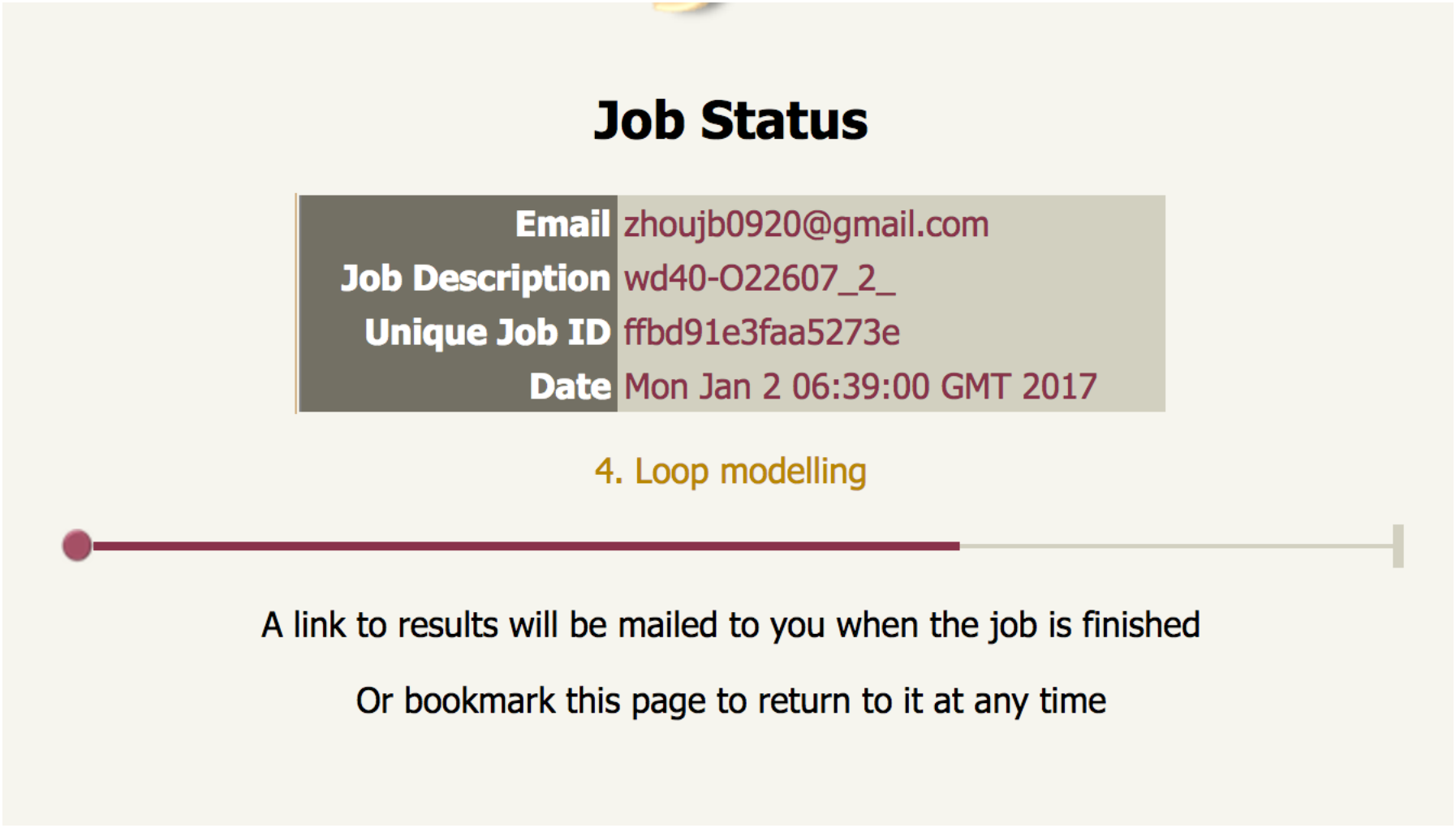
1.输入序列

EBI 2016 Workshop | Older Workshops | New Phyre2 paper | Fast structural search with PhyreStorm (beta-testing)

E-mail Address	zhoujb0920@gmail.com
Optional Job description	wd40-O22607(2)
Amino Acid Sequence	>sp|O22607|MSI4_ARATH WD-40 repeat-containing protein MSI4 OS=Arabidopsis thaliana GN=MSI4 PE=1 SV=3 MESDEAAAVSPQATTPSGGTGASGPKKRGRKPKTKEDSQTP SSQQQSDVKMKESGKKTQQSPSVDEKYSQWKGLVPILYDWL ANHNLVWPSLSCRWGPQLEQATYKNRQRLYLSEQTDGSVP NTLVIANCEVVKPRVAAAEHISQFNEEARSPFVKKYKTIIHPGE VNRIRELPQNSKIVATHTDSPDVLIWDVETQPNRHAVLGAANS RPDLILTGHQDNAEFALAMCPTEPFVLSGGKDKSVVLWSIQD HITTIGTDSKSSGSIIKQTGEGTDKNESPTVGPRGVYHGHEDT
	Or try the sequence finder
Modelling Mode	Normal ○ Intensive ●
	Phyre Search Reset

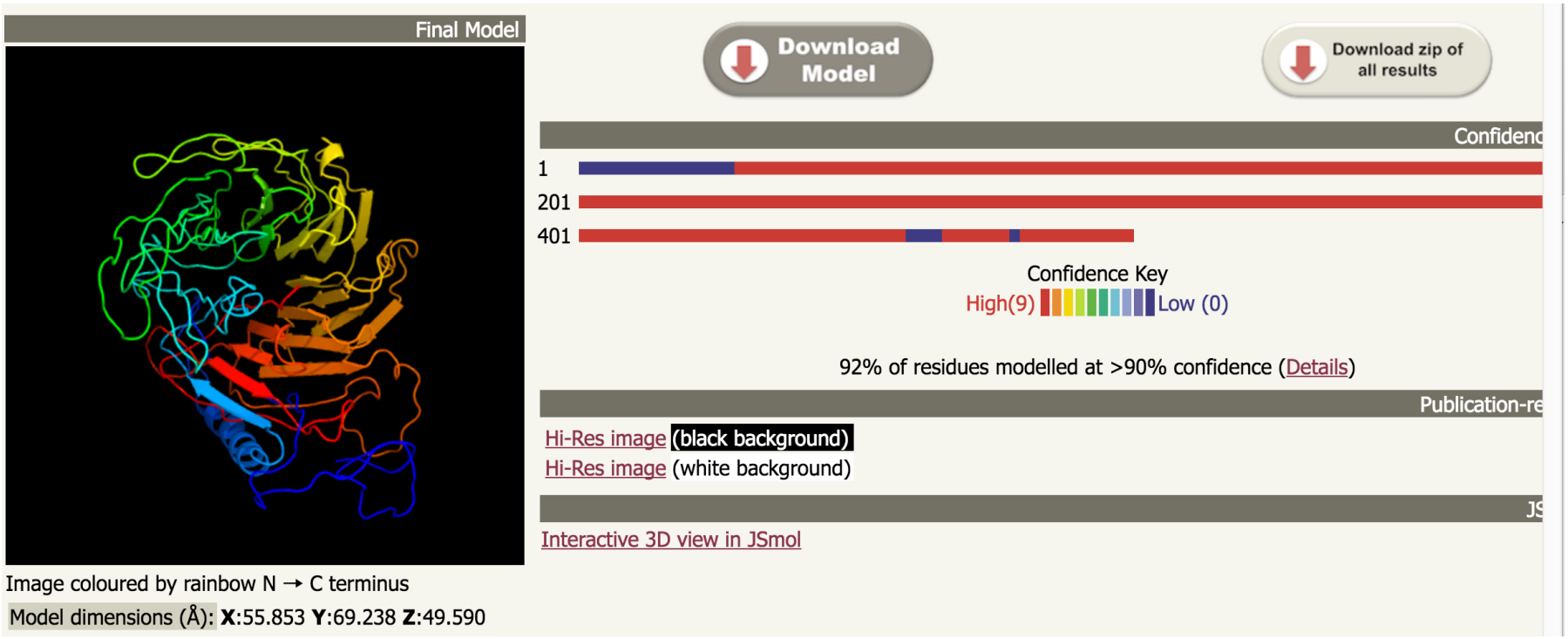
Phyre2——简要介绍

2.耐心等待



Phyre2——简要介绍

3.结果分析



Phyre2——简要介绍

Sequence analysis

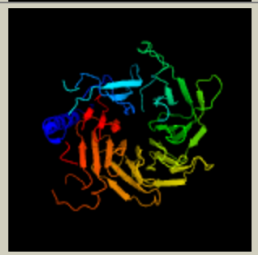
View PSI-Blast Pseudo-Multiple Sequence Alignment

Download FASTA version

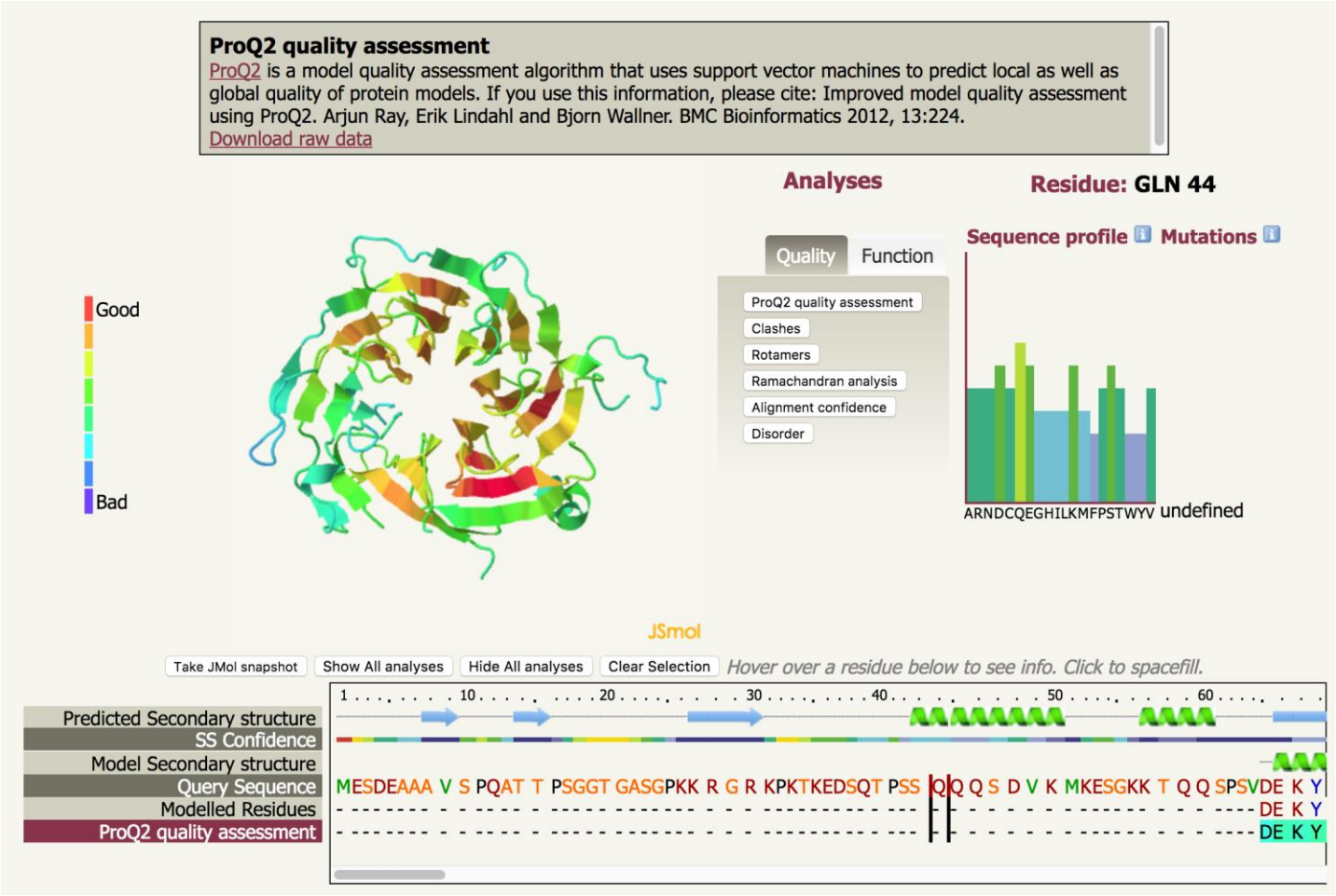
Secondary structure and disorder prediction [Hide]

```
                1 . . . . . 10 . . . . . 20 . . . . . 30 . . . . . 40 . . . . . 50 . . . . . 60
Sequence        M E S D E A A A V S P Q A T T P S G G T G A S G P K K R G R K P K T K E D S Q T P S S Q Q Q S D V K M K E S G K K T Q Q
Secondary structure
```

Detailed template information [Hide]

#	Template	Alignment Coverage	3D Model	Confidence	% i.d.	Template Information
1	c4xyhA_	Alignment		100.0	26	**PDB header:**chaperone **Chain:** A: **PDB Molecule:**kinetochore protein mis16; **PDBTitle:** wild-type full length mis16 in schizosaccharo Run Investigator
2	c3cfvA_	Alignment		100.0	28	**PDB header:**histone/chaperone **Chain:** A: **PDB Molecule:**histone-binding protein rbbp **PDBTitle:** structural basis of the interaction of rbap46/ h4 Run Investigator

Phyre2——简要介绍



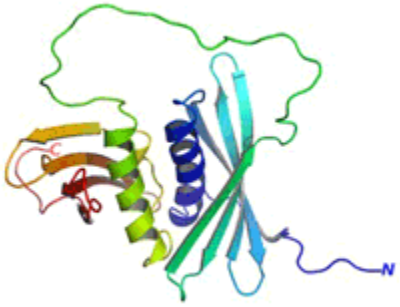
其他功能：1.Binding site prediction
2.Transmembrane helix prediction
3.Superposition of model
4.。。。

Phyre2——与MODELLER比较

同源建模方法代表性的软件：MODELLER
优点：1.自动化建模程序
2.客观、迅速
3.编译的程序与代码
缺点：1.需要使用python
2. 准确度略低于Phyre2
3.质量评估需借助其他软件

同源建模方法代表性的软件：Phyre2
优点：1.ease of use
2.user-friendly interface
缺点：1.只能模拟single-chain model
2.模拟过程较慢

同源建模的目的



蛋白质结构

- 理解蛋白质生物功能的分子机制
- 基于结构的药物设计
- 蛋白质设计与蛋白质工程
- 研究蛋白折叠
- 帮助点突变实验设计
- 理解疾病相关突变
- 理解蛋白质-蛋白质相互作用

获得蛋白质的结构后即可进一步分析，也可以辅助分析一些实验的结果

致谢

感谢罗老师这一学期的讲授
感谢叶老师的教导
感谢柯岚师姐的帮助
感谢小组成员的相互合作

Thank You