Protein Structures Visualization

生物信息学导论

http://bis.zju.edu.cn/download/Protein_structure/

周银聪

foxandzhou@gmail.com

Dec 21 ,2015

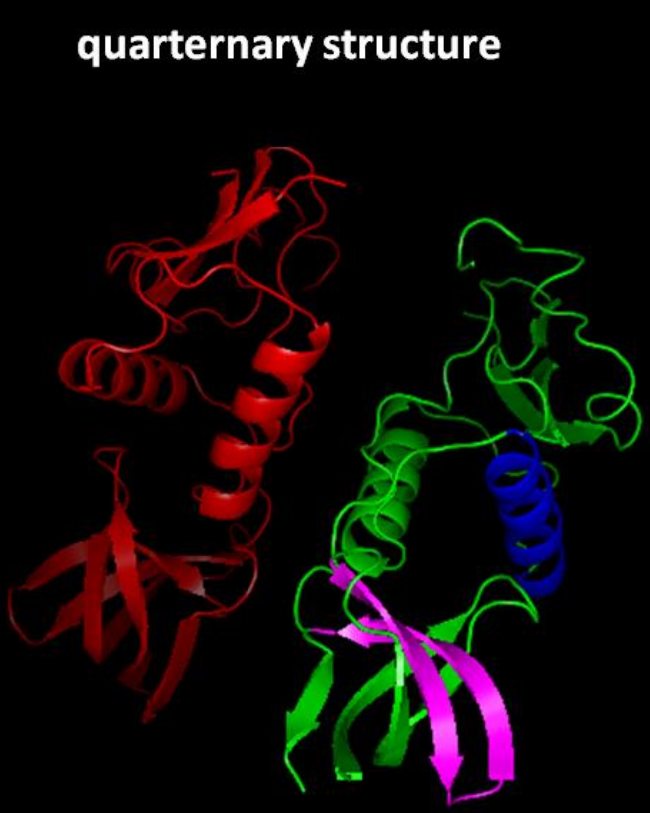
PDB—Protein Data Bank

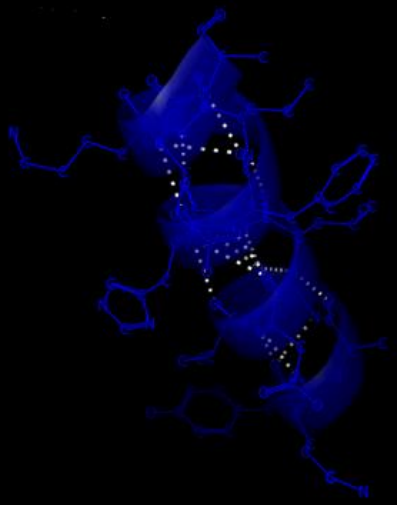
Pymol

secondary structure

quarternary structure

tertiary structure





α-helix

β-sheet

primary structure

Tyr-Lys- Ala-Ala-Val-Asp-Leu-Ser-His-Phe-Leu-Lys-Glu-Lys

Asp-Trp-Trp-Glu-Ala-Arg-Ser-Leu-Thr-Thr-Gly-Glu-Thr-Gly-Tyr-Pro-Ser

From wikipedia

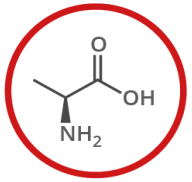
A GUIDE TO THE TWENTY COMMON AMINO ACIDS

AMINO ACIDS ARE THE BUILDING BLOCKS OF PROTEINS IN LIVING ORGANISMS. THERE ARE OVER 500 AMINO ACIDS FOUND IN NATURE - HOWEVER, THE HUMAN GENETIC CODE ONLY DIRECTLY ENCODES 20. 'ESSENTIAL' AMINO ACIDS MUST BE OBTAINED FROM THE DIET, WHILST NON-ESSENTIAL AMINO ACIDS CAN BE SYNTHESISED IN THE BODY.

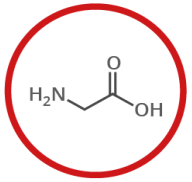
Chart Key: ALIPHATIC · AROMATIC · ACIDIC · BASIC · HYDROXYLIC · SULFUR-CONTAINING · AMIDIC · NON-ESSENTIAL · ESSENTIAL

Chemical Structure — single letter code

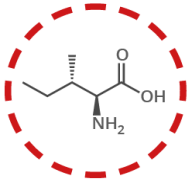
NAME (A) — *three letter code* — DNA codons



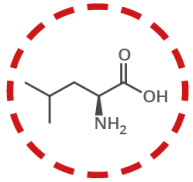
ALANINE (A) — *Ala* — GCT, GCC, GCA, GCG



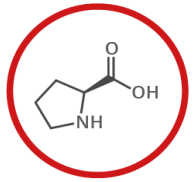
GLYCINE (G) — *Gly* — GGT, GGC, GGA, GGG



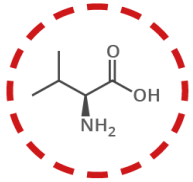
ISOLEUCINE (I) — *Ile* — ATT, ATC, ATA



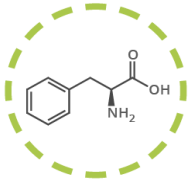
LEUCINE (L) — *Leu* — CTT, CTC, CTA, CTG, TTA, TTG



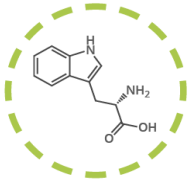
PROLINE (P) — *Pro* — CCT, CCC, CCA, CCG



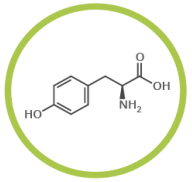
VALINE (V) — *Val* — GTT, GTC, GTA, GTG



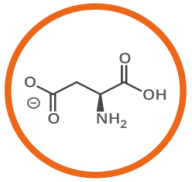
PHENYLALANINE (F) — *Phe* — TTT, TTC



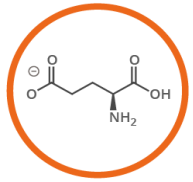
TRYPTOPHAN (W) — *Trp* — TGG



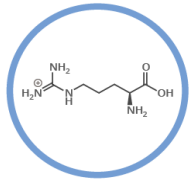
TYROSINE (Y) — *Tyr* — TAT, TAC



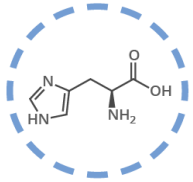
ASPARTIC ACID (D) — *Asp* — GAT, GAC



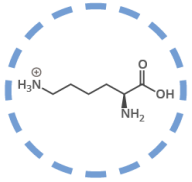
GLUTAMIC ACID (E) — *Glu* — GAA, GAG



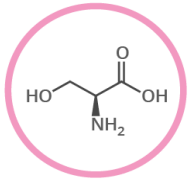
ARGININE (R) — *Arg* — CGT, CGC, CGA, CGG, AGA, AGG



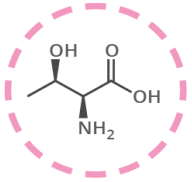
HISTIDINE (H) — *His* — CAT, CAC



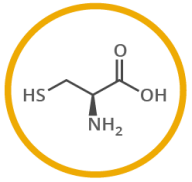
LYSINE (K) — *Lys* — AAA, AAG



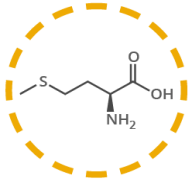
SERINE (S) — *Ser* — TCT, TCC, TCA, TCG, AGT, AGC



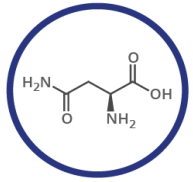
THREONINE (T) — *Thr* — ACT, ACC, ACA, ACG



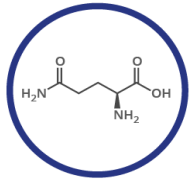
CYSTEINE (C) — *Cys* — TGT, TGC



METHIONINE (M) — *Met* — ATG



ASPARAGINE (N) — *Asn* — AAT, AAC

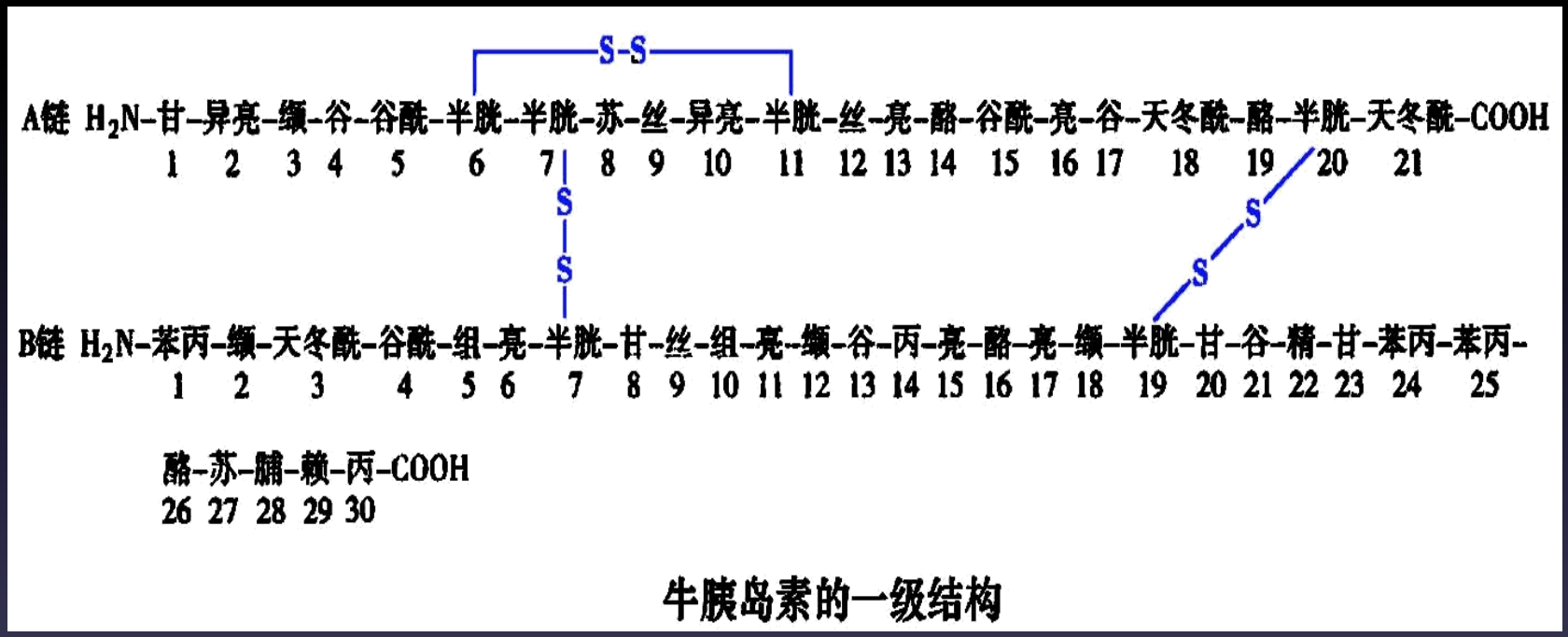


GLUTAMINE (Q) — *Gln* — CAA, CAG

Note: This chart only shows those amino acids for which the human genetic code directly codes for. Selenocysteine is often referred to as the 21st amino acid, but is encoded in a special manner. In some cases, distinguishing between asparagine/aspartic acid and glutamine/glutamic acid is difficult. In these cases, the codes asx (B) and glx (Z) are respectively used.

一级结构：

指多肽链中氨基酸的排列顺序，每个氨基酸由肽键连接，有些还包括二硫键



牛胰岛素的一级结构

一级结构是蛋白质空间构象和特异生物学功能的基础

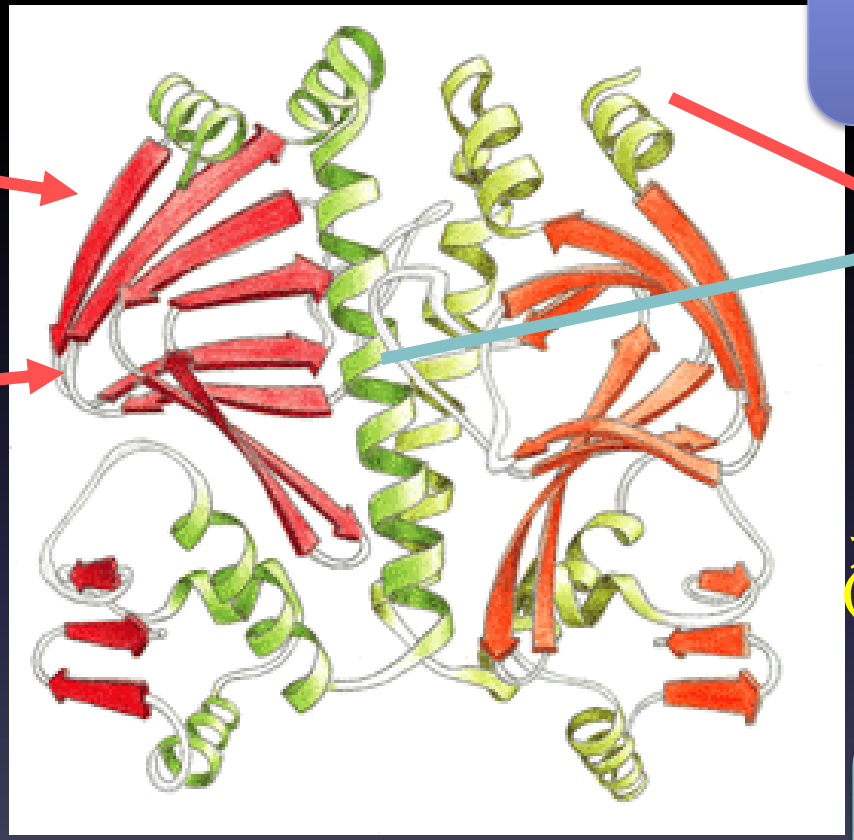
氨基酸二级结构：

存在于纤维状蛋白质（β角蛋白：蚕丝、蜘蛛丝中心丝蛋白）和球蛋白中

β-折叠
(β- pleated sheet)

β-转角(β- turn)

多肽链中180度回折形成的特定构象



存在于纤维状蛋白质（α角蛋白：皮肤、毛发、指甲）和球状蛋白质中

α-螺旋
(α- helix)

无规则卷曲
(random coil)

没有确定规律性的肽链结构，存在于球状蛋白质中（常常是酶功能部位）

结构域：

由二级结构组成的相对独立的结构单元，一般有100~200aa，能够承担特定的生物化学功能

三级结构:

由结构域在空间中按一定的方式排列形成，二级结构元件折叠成紧密地、近乎坚硬的物体，由极性基团和非极性集团都参与的弱相互作用所稳定.三级结构的一个效果是产生复杂的表面拓扑学结构，使得蛋白质能够与其它大小分子发生相互作用.

四级结构:

由几条具有三级结构的肽链组成一个蛋白质功能上的重要位点是在三级结构或四级结构层次上形成的

在结构生物学中，主要由3种确定蛋白质结构的方法

1. 实验方法：
   --X-射线晶体衍射法
   --核磁共振
   --冷冻电子显微镜

2. 同源建模法
   通过将目标蛋白与一个或几个已知结构的同源蛋白质进行比较
   从而预测其结构

3. 从头预测法：对于没有已知同源蛋白质结构可用的蛋白，采取通过物理学原理来预测结构

X射线衍射法是测定蛋白质结构最精确的方法，目前80％已知的结构是通过这种技术测定的

蛋白质溶液样品必须有足够高的浓度和纯度，并在合适条件下形成晶体

X-射线的波长为0.05~0.15nm，适合测量原子间距离，蛋白晶体将X光衍射到探测装置，晶体结构可以通过衍射图推导得到

核磁共振：Nuclear magnetic resonance，NMR。
将蛋白质溶液至于磁场中通过测量溶液中蛋白质指定原子共振之间的扰动来测定核间距离，从中推测蛋白结构.
测量的是原子核的相互作用，测定产生的时原子核间距离的数据集
不需要复杂的蛋白质结晶过程，但是蛋白质必须在接近于晶格中蛋白浓度的情况下可溶，适用于分子量小于50kDa的蛋白质.

同源建模：homology modeling
目标是通过和一个结构已知的同源蛋白质进行序列比较和结构分析，为未知蛋白产生一个合理的近似结构。

原理：如果两个蛋白质之间存在高度的顺序相似性（>40？），它们总体上的折叠方式往往是相似的。当序列一致性低于40%，结构可能有显著地差异

当存在多条对位排列的顺序时，就可以得到更高分辨率的模型

主要步骤：

--利用BLAST搜索同源蛋白的序列和结构
--根据搜索结果制定的折叠类型并选择模板
--将目标与模板的序列中相应位置对齐
--建立结构模型
--评估模型

当两个序列相似度超过50%，模型质量非常好；低30%，模型的错误会急剧上升

Protein Data Bank

收集了全世界利用核磁共振，X-ray衍射实验，理论模拟出来的蛋白质和DNA的三维立体结构.

PDB是全球最重要的蛋白质结构资料来源，主要提供的信息有：原子的空间坐标，引用文献，形成的α-helix和β-sheet的氨基酸序列，二硫键连接模式，与蛋白结合的ligand，参与生化功能的residue

RCSB PDB | Deposit | Search | Visualize | Analyze | Download | Learn | More | MyPDB Login

RCSB PDB PROTEIN DATA BANK — An Information Portal to 114217 Biological Macromolecular Structures

Search by PDB ID, author, macromolecule, sequence, or ligands — Go

Advanced Search | Browse by Annotations

PDB-101 | wwPDB | EMDataBank | NDB | StructuralBiology Knowledgebase

- Welcome
- Deposit
- Search
- Visualize
- Analyze
- Download
- Learn

A Structural View of Biology

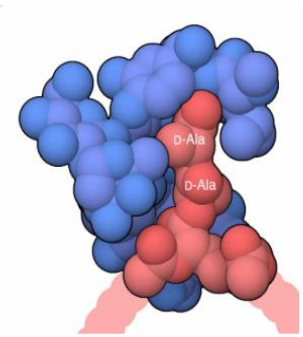
This resource is powered by the Protein Data Bank archive-information about the 3D shapes of proteins, nucleic acids, and complex assemblies that helps students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease.

As a member of the wwPDB, the RCSB PDB curates and annotates PDB data.

The RCSB PDB builds upon the data by creating tools and resources for research and education in molecular biology, structural biology, computational biology, and beyond.

Take an Interactive Tour of the PDB

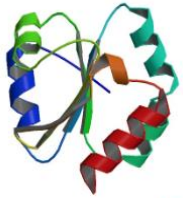
December Molecule of the Month



Vancomycin

Latest Entries

As of Tuesday Dec 08



4RUV — PDB Entry

Crystal structure of thioredoxin 2 from Staphylococcus aureus NCTC8325

View in 3D

New Features

October 2015 Release

Redesigned Structure Summary Page

New Organization. Improved Layout. Clean. Usable. Simple.

Improved Literature Tab

Better Support for Mobile Browsing

Redesigned Ligand Summary Page

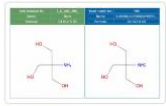
September 2015 Release

Validation Track on Protein Feature View

Mapping validation annotations to sequence

News

Publications



Tour of Ligand Deposition

Watch how to review and submit ligands using the wwPDB Deposition Tool.

» 12/08/15

Validate Before Depositing to Save Time » 12/01/15

Advanced Search: Multiple ID Search » 11/24/15

Comparison Tool for Exploring Sequence and Structure Alignments » 11/17/15

Phased PDB Release Process

» 08/24/15

Announcing the 2015 EMDataBank Map Challenge

» 08/11/15

PDB at a Glance | 37126 Distinct Protein Sequences | 30177 Structures of Human Sequences | 8112 Nucleic Acid Containing Structures | More Statistics

http://www.rcsb.org/pdb/home/home.do

PDB中的记录有唯一的PDB-ID，包括4个字符串，可由大写字母A~Z和数字0~9组合而成。

PDB和它的镜像站点提供每个PDB记录的查询，可按一些专门的查询项目（如提交数据、作者姓名、结构表达）进行检索。

练习：从PDB上下载抹香鲸肌红蛋白（myoglobin）1MBN的三维结构件文件，并用写字板打开

```
HEADER    OXIDOREDUCTASE                          25-MAR-09   3GRK
TITLE     CRYSTAL STRUCTURE OF SHORT CHAIN DEHYDROGENASE REDUCTASE
TITLE    2 SDR GLUCOSE-RIBITOL DEHYDROGENASE FROM BRUCELLA MELITENSIS
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: ENOYL-(ACYL-CARRIER-PROTEIN) REDUCTASE (NADH);
COMPND   3 CHAIN: A, B, C, D, E, F, G, H;
COMPND   4 EC: 1.3.1.9;
COMPND   5 ENGINEERED: YES
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: BRUCELLA MELITENSIS;
SOURCE   3 ORGANISM_TAXID: 29459;
SOURCE   4 STRAIN: BIOVAR ABORTUS 2308;
SOURCE   5 GENE: BMEI1512;
SOURCE   6 EXPRESSION_SYSTEM: ESCHERICHIA COLI;
SOURCE   7 EXPRESSION_SYSTEM_TAXID: 562;
SOURCE   8 EXPRESSION_SYSTEM_VECTOR_TYPE: AVA0421
KEYWDS    SSGCID, NIAID, STRUCTURAL GENOMICS, SEATTLE STRUCTURAL
KEYWDS   2 GENOMICS CENTER FOR INFECTIOUS DISEASE, OXIDOREDUCTASE
EXPDTA    X-RAY DIFFRACTION
AUTHOR    SEATTLE STRUCTURAL GENOMICS CENTER FOR INFECTIOUS DISEASE
AUTHOR   2 (SSGCID)
REVDAT   1   07-APR-09 3GRK    0
JRNL        AUTH   T.E.EDWARDS,B.L.STAKER,
JRNL        AUTH 2 SEATTLE STRUCTURAL GENOMICS CENTER FOR INFECTIOUS
JRNL        AUTH 3 DISEASE (SSGCID)
JRNL        TITL   CRYSTAL STRUCTURE OF SHORT CHAIN DEHYDROGENASE
JRNL        TITL 2 REDUCTASE SDR GLUCOSE-RIBITOL DEHYDROGENASE FROM
JRNL        TITL 3 BRUCELLA MELITENSIS
JRNL        REF    TO BE PUBLISHED
JRNL        REFN
REMARK   1
REMARK   2
REMARK   2 RESOLUTION.    2.35 ANGSTROMS.
REMARK   3
REMARK   3 REFINEMENT.
REMARK   3   PROGRAM     : REFMAC 5.5.0088
REMARK   3   AUTHORS     : MURSHUDOV,VAGIN,DODSON
REMARK   3
```

PDB对应的ID，存入的时间

介绍大分子

大分子的关键词

获得结构的方法

文献

实验细节，注释等

```
DBREF  3GRK A    1   272  UNP    Q8YFK8   Q8YFK8_BRUME     1    272
DBREF  3GRK B    1   272  UNP    Q8YFK8   Q8YFK8_BRUME     1    272
DBREF  3GRK C    1   272  UNP    Q8YFK8   Q8YFK8_BRUME     1    272
DBREF  3GRK D    1   272  UNP    Q8YFK8   Q8YFK8_BRUME     1    272
DBREF  3GRK E    1   272  UNP    Q8YFK8   Q8YFK8_BRUME     1    272
DBREF  3GRK F    1   272  UNP    Q8YFK8   Q8YFK8_BRUME     1    272
DBREF  3GRK G    1   272  UNP    Q8YFK8   Q8YFK8_BRUME     1    272
DBREF  3GRK H    1   272  UNP    Q8YFK8   Q8YFK8_BRUME     1    272
```

氨基酸序列来源

```
SEQRES  17 H  293  ASN ALA ILE SER ALA GLY PRO ILE LYS THR LEU ALA ALA
SEQRES  18 H  293  SER GLY ILE GLY ASP PHE ARG TYR ILE LEU LYS TRP ASN
SEQRES  19 H  293  GLU TYR ASN ALA PRO LEU ARG ARG THR VAL THR ILE ASP
SEQRES  20 H  293  GLU VAL GLY ASP VAL GLY LEU TYR PHE LEU SER ASP LEU
SEQRES  21 H  293  SER ARG SER VAL THR GLY GLU VAL HIS HIS ALA ASP SER
SEQRES  22 H  293  GLY TYR HIS VAL ILE GLY MET LYS ALA VAL ASP ALA PRO
```

各个α-helix, β-sheet等信息

```
HELIX   99  99 VAL H  124  MET H  138  1                                  15
HELIX  100 100 TYR H  149  GLU H  153  5                                   5
HELIX  101 101 TYR H  159  GLY H  181  1                                  23
HELIX  102 102 ASP H  205  ALA H  217  1                                  13
HELIX  103 103 THR H  224  SER H  237  1                                  14
HELIX  104 104 ASP H  238  ARG H  241  5                                   4
HELIX  105 105 GLY H  253  ILE H  257  5                                   5
SHEET    1   A 7 PHE A  62  HIS A  66  0
SHEET    2   A 7 GLU A  38  TYR A  43  1  N  PHE A  41   O  ALA A  64
SHEET    3   A 7 ARG A  12  LEU A  16  1  N  ILE A  15   O  ALA A  40
SHEET    4   A 7 PHE A  91  HIS A  94  1  O  VAL A  93   N  LEU A  14
SHEET    5   A 7 GLY A 142  THR A 148  1  O  LEU A 145   N  HIS A  94
SHEET    6   A 7 ILE A 185  ALA A 192  1  O  ARG A 186   N  ILE A 144
SHEET    7   A 7 VAL A 247  ALA A 250  1  O  HIS A 248   N  SER A 191
```

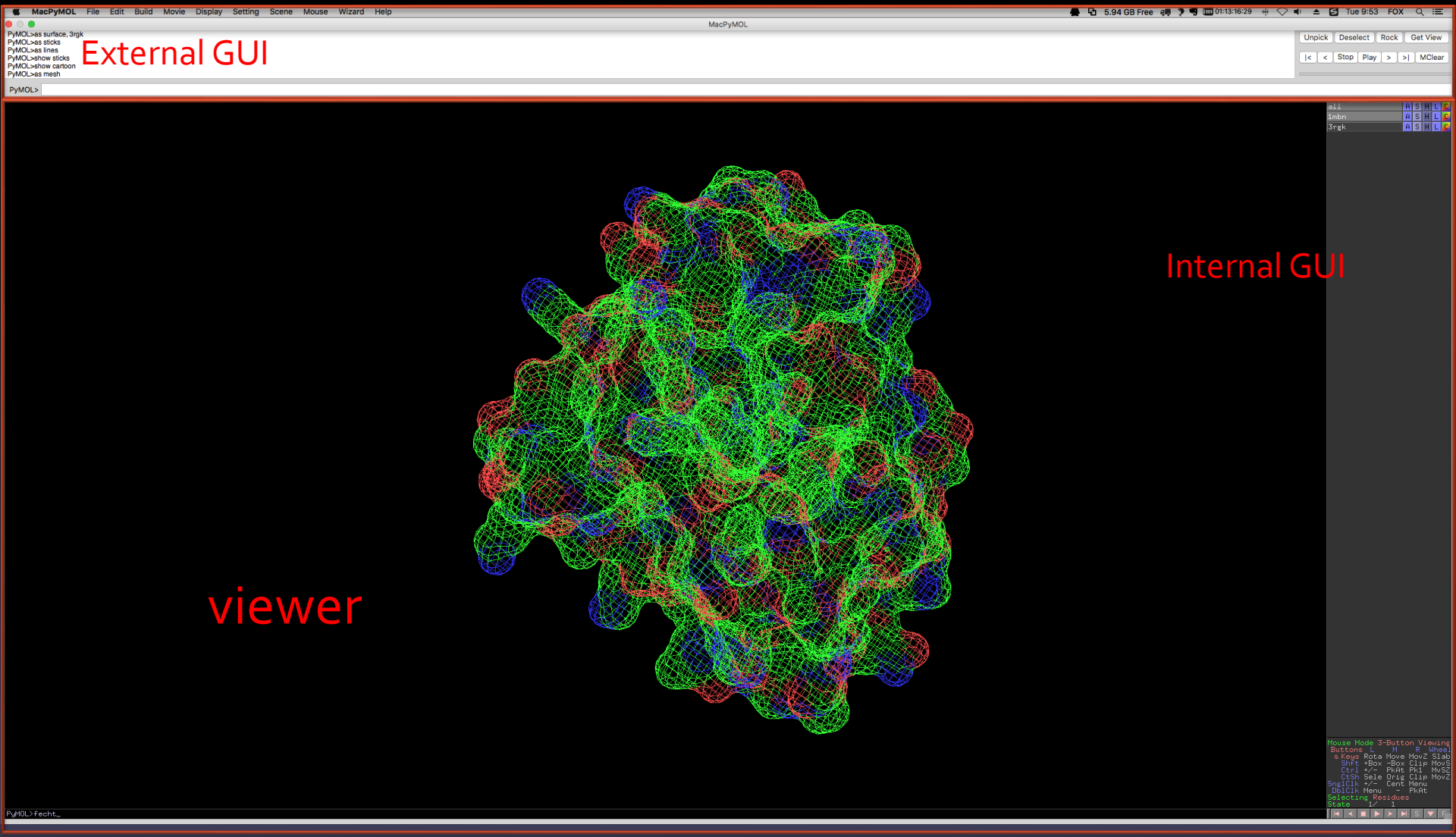
原子　　氨基酸　　X轴坐标　Y轴坐标　Z轴坐标

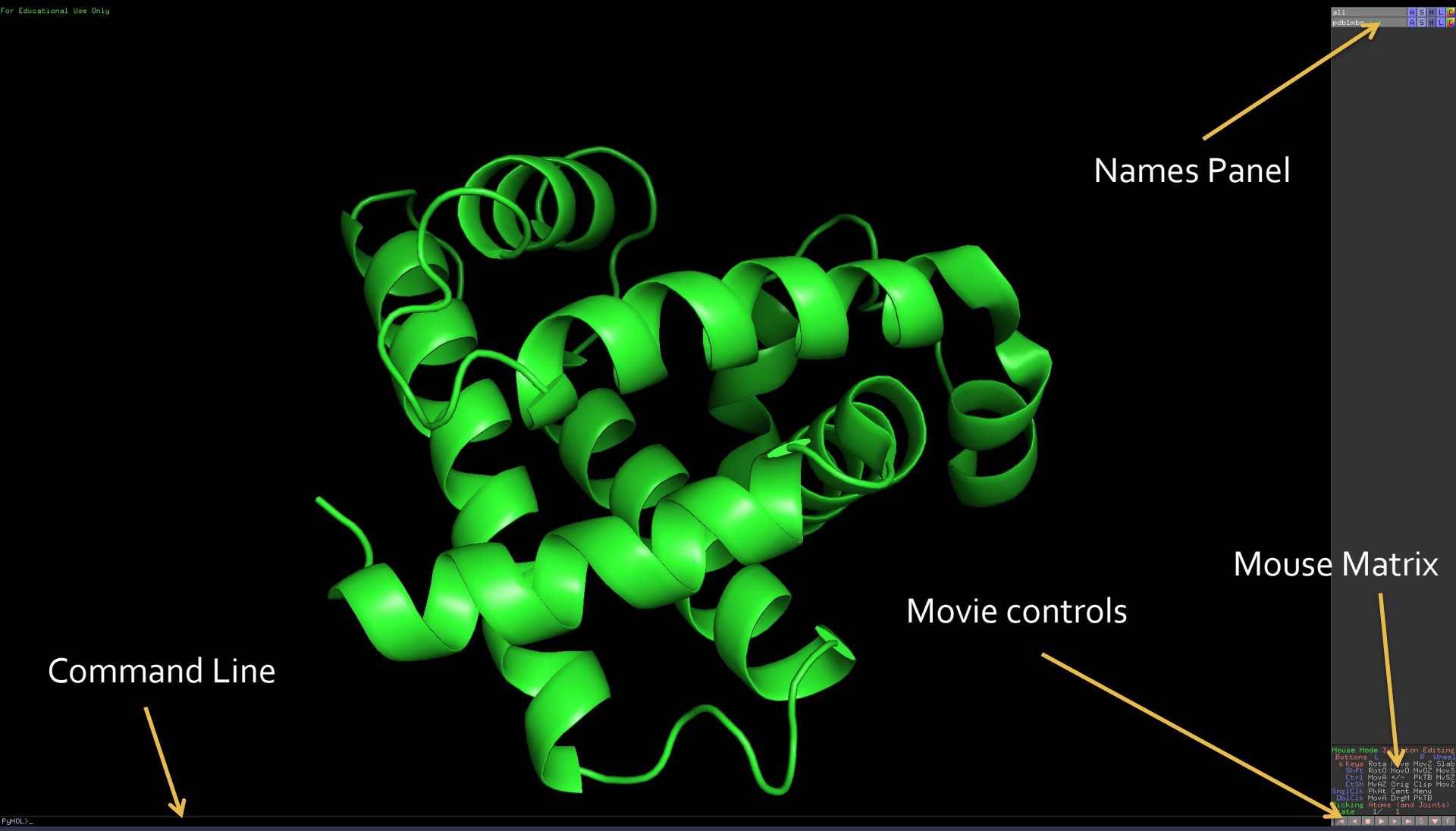
```
16361 ATOM  15108  N    MET H 259     -53.308 -53.688 -71.333  1.00 28.18           N
16362 ATOM  15109  CA   MET H 259     -54.753 -53.543 -71.270  1.00 29.08           C
16363 ATOM  15110  C    MET H 259     -55.426 -54.906 -71.479  1.00 29.61           C
16364 ATOM  15111  O    MET H 259     -56.535 -55.147 -70.983  1.00 30.91           O
16365 ATOM  15112  CB   MET H 259     -55.203 -52.581 -72.366  1.00 29.27           C
16366 ATOM  15113  CG   MET H 259     -56.593 -52.057 -72.196  1.00 29.78           C
16367 ATOM  15114  SD   MET H 259     -57.025 -50.897 -73.506  1.00 31.20           S
16368 ATOM  15115  CE   MET H 259     -58.779 -51.194 -73.508  1.00 32.74           C
16369 TER   15116       MET H 259
```

Pymol

- Pymol: python + molecule
- 开源
- 产生高质量三维结构图像

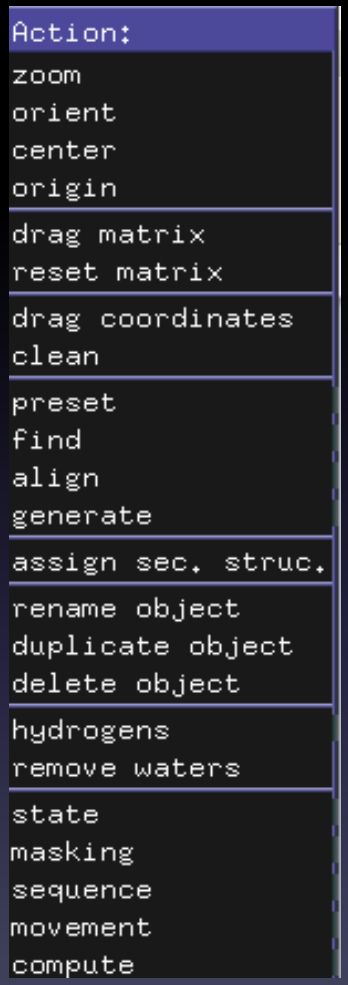
Pymol GUI


MacPyMOL File Edit Build Movie Display Setting Scene Mouse Wizard Help
MacPyMOL
PyMOL>as surface, 3rgk
PyMOL>as sticks
PyMOL>as lines
PyMOL>show sticks
PyMOL>show cartoon
PyMOL>as mesh
External GUI
Unpick Deselect Rock Get View
Stop Play MClear
PyMOL>
all
1mbn
3rgk
Internal GUI
viewer
Mouse Mode 3-Button Viewing
Selecting Residues
State
PyMOL>fecht_

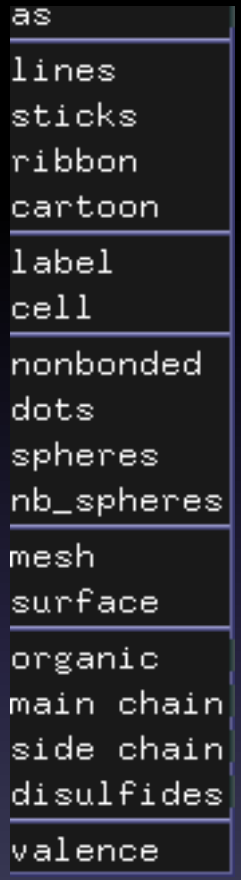


A: Action S: Show H: Hide L: Label C: Color

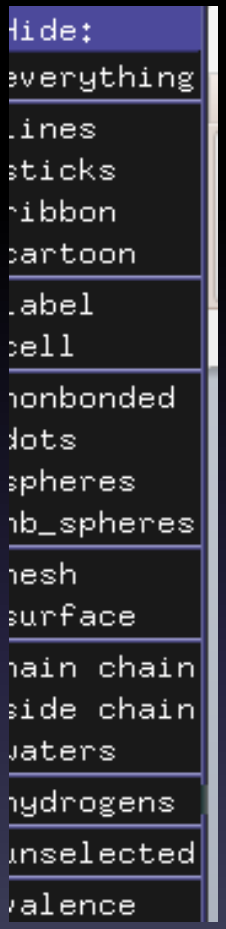
A: Action



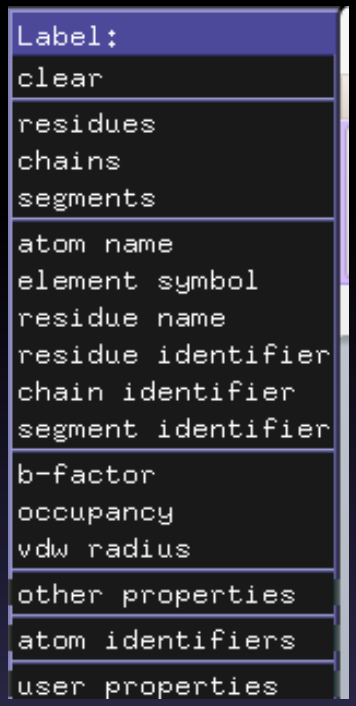
S: Show



H: Hide

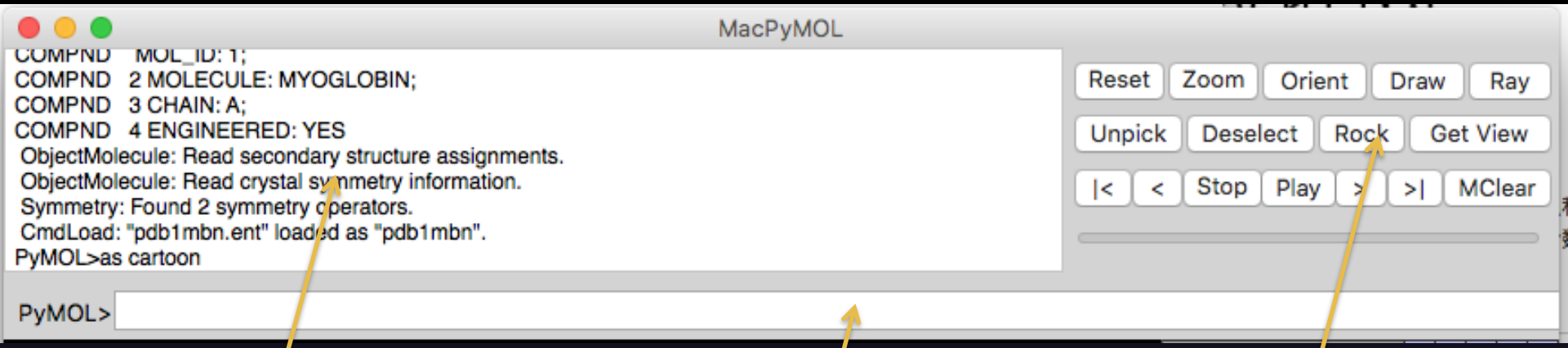


L: Label



C: Color

The External GUI Window



Output

Command line

Button

与Viewer window相比的优势:能通过Ctrl-X, Ctrl-C, and Ctrl-V使用剪切、复制、黏贴功能

几个基本的命令：

cd -> change directory(更改所在文件夹的位置)

pwd ->print work directory（显示当前目录）

ls -> list（列出当前目录下的文件）

练习：用pymol打开刚下载的肌红蛋白文件1MBN，熟悉简单操作。

打开方式：

窗口模式：file->open->your.file

命令行模式：load your.file

- File -> Open -> 1mbn.pdb

- hide everything, all
- show cartoon, all
- color purple, ss h

- as mesh
- as surface
- as sticks
- as lines

关于cartoon

Cartoon的命令格式如下

Cartoon的显示类型：
Automatic:默认的显示方式
Putty:按R-factor显示,值越大越粗
还有loop,Tube,Arrow等等

```
PyMOL> hide everything, all  --隐藏所有元件
PyMOL> show cartoon, all      --显示元件为cartoon元件
PyMOL> color purple, ss h       --给二级结构α螺旋上色
PyMOL> color yellow, ss s       --给β折叠上色
PyMOL> color green, ss l+       --给loop以及其他元素上色
PyMOL> select active, (resi 14-20 and chain A)  --选择A链上14到20的残基，命名
        为active
PyMOL> color yellow, active  --将名为active的区域改变为黄色
PyMOL> turn y, -60; turn x, -20   --以y轴逆时针旋转60度，X轴逆时针旋转20度
PyMOL> bg_color white       --将背景改为白色
PyMOL> ray                --渲染
File -> Save Image        --保存图片
```

练习2：从PDB上下载人肌红蛋白3RGK,与抹香鲸肌红蛋白比较1MBN进行比较

Pymol可直接访问PDB，命令行中输入： fetch 3rgk
可以直接下载并载入蛋白。

- delete all
- fetch 1mbn
- fetch 3rgk
- align 3rgk,1mbn

动画制作：Pymol有强大的分子动画制作功能

几个概念：

- States(状态):状态指对象(object)某一个时间点特定的原子坐标。

- Scenes (场景):场景存储镜头(camera)的位置和定向、对象的活 动信息、原子的可见性(visibility)、着色、表示形式和全局帧索引 (global frame index)

- Frames(帧):帧就像电影胶片中一个个单独的图片,在PyMOL中, 帧是由状态(states)而不是图片构成的,而且对帧可以进行相关操 作(如camera的选转)。帧存储状态信息和场景信息。

重要命令:

Mset命令

Mset命令用来指定那些状态作为动画的帧。

Mset命令后紧跟定义整个动画的状态列表。每个状态采用以下形式之 一:

1 # 一个数字:指定下一个放映的状态

X # 一个数字紧随小写“x”(无空格):指定状态总共该重复的次数

- # 一个数字紧随连字号(无空格):指定状态按载入的顺序的放映。

例子:

mset 1 x30 # 创建一个由状态1放映30遍组成的30帧的动画

Mdo命令

Mdo命令可以把一系列的PyMOL命令捆绑到帧上。

“util”组件为产生mdo命令有两个脚本命令,“util.mrock”和 “util.mroll”。这些功能还没入档,但源程序可在 modules/PyMOL/util.py找到。

util.mrock start, finish, angle, phase, loop-flag

util.mroll start, finish, loop-flag

```
mset  1x30
util.mrock 1,30,180,1,1
mplay
```

```
mset 1 x30                  创建一个三十帧动画
util.mrock(1,30,30,1,1)          创建动作
set ray_trace_frames=1          是否对每一个frame进行Ray处理（）
set cache_frames=1          是否将每个frame存入内存（）
mplay                       播放
mpng mov                    导出30张png图片
```

示例2

（1）输入“mset 1 x100”定义一个30帧的动画
（2）输入“frame 1”定义第一帧
（3）调整观察角度到一个合适的角度（比如从整体观察蛋白的视角）
（4）输入“mview store”存储该观察视角为第一帧
（5）输入“frame 50”定义第50帧
（6）调整观察角度到另一个合适的角度
（7）输入“mview store”存储该观察视角为第50帧
（8）输入“frame 100”定义第50帧
（9）调整观察角度到另一个合适的角度
（10）输入“mview store”存储该观察视角为第100帧
（11）输入“mivew reinterpolate”创建从第一帧到第100帧的平滑过渡
（12）输入“mplay”观看平滑过渡效果是否满意
（13）输入“set ray_trace_frames=1” 输出时渲染每一帧
（14）输入“set cache_frames=1”
（15）输入“mpng mov”输出100帧png图片

练习4：选择蛋白质结构(irisin)，创建蛋白质3D动画短片

Thank you