

结构生物学实验技能课中应用可视化软件Pymol的教学体会

谢伟*

(中山大学生命科学学院生物化学系, 广州 510275)

摘要: 人们关于蛋白质结构预测的研究取得了突破性进展。从海量结构信息中对兴趣蛋白质进行功能分析及推断是生物学家未来需要探索的重点问题, 但前提是需要对已有蛋白质结构的模型具有充分的认识和理解。Pymol是蛋白质三维结构可视化最常用的软件之一, 其功能多样且强大, 深受广大结构生物学工作者的青睐。该文旨在分享作者在本科生结构生物学实验技能课教学中使用Pymol软件的经验, 深入浅出地讲解该软件的初步使用方法, 从而使初学者能够达到查看、展示并进行一般性结构绘图的目的。开设该课程对提高学生对于蛋白质结构信息的理解、激发其科研兴趣等方面均具有较大益处, 并且取得了良好的教学效果。

关键词: 结构生物学实验技能课; Pymol; 3D-结构可视化

Teaching experience of the visualization software Pymol in the Structural Biology Experimental Skills Course

XIE Wei*

(Department of Biochemistry, School of Life Sciences, Sun Yat-sen University, Guangzhou 510275, China)

Abstract: Major breakthroughs have been made on protein structure prediction in recent years. Analyzing and deducing the functions of proteins of interest from massive structural information is a problem that biologists need to focus on in the future, but the premise is that there is a full understanding and understanding of the existing protein structure models. Pymol is one of the most commonly used softwares for structural visualization. Its functions are diverse and powerful, so it is widely acclaimed by the majority of structural biology researchers. This paper aims to share the author's teaching experience in the introductory use of the Pymol software in a simplified manner. This teaching session is of great benefits to students in understanding protein structure information and stimulating their scientific interest, which achieved awesome teaching results.

Key Words: Structural Biology Experimental Skills Course; Pymol; 3D-structure visitation

生命科学在于揭示生命体内各种生命活动的奥秘, 而蛋白质作为生命活动的主要承担者在各项生命活动中起着不可或缺的作用。蛋白质的编码序列决定其结构, 而后者又决定其功能。近年来随着人工智能的兴起, 关于蛋白质结构预测的计算生物学获得了长足发展。DeepMind公司旗下

开发的AlphaFold/AlphaFold 2^[1,2]人工智能系统的强势推出, 极大地丰富了蛋白质的结构信息。2022年7月底, AlphaFold 2又在原有基础上, 释放了数以亿计的蛋白质的结构信息, 其预测结构超过2亿个, 几乎涵盖所有已知蛋白质。面对海量的结构信息, 我们该如何获得相应蛋白质的功能, 是十

分必要和值得研究的问题。作为结构生物学工作者,首先需要通过可视化软件来查看和分析蛋白质的结构,获得三维结构整体折叠和局部精细结构的信息,进行充分的认识和理解;并基于已有的物理化学知识,对其功能做出合理推断。目前蛋白质可视化软件较多,如Swiss-PDBViewer、VMD、chimera、Pymol^[3-5]等,其中最为主流的软件是由Schrodinger公司开发的Pymol。Pymol软件对学术型用户为免费,并且由于其用户界面友好、功能强大,备受科研工作者的喜爱。

1 Pymol使用的软硬件要求

Pymol名称由Py+mol命名,“Py”表示它是由计算机语言Python所衍生,“mol”表示它是用于显示分子(molecule)结构的软件。由于Pymol提供Windows/Mac/Linux多种安装版本,因此可在多个操作平台下运行,用户可根据需要下载。Pymol对电脑及操作系统要求不高,但为保证使用时的流畅性,电脑的配置(尤其是显卡)较高为优。不同操作系统下Pymol软件的用户界面大同小异,现以苹果电脑,macOS Catalina操作系统为例来作示范说明。现在Pymol最高版本已达到2.5,可在官网下载: <https://pymol.org/2/#download>。非商业(高校及研究机构)用户在提供自己的注册信息后,便可获得软件的安装版本及授权文件license。license的有效期是一年,必须每年向该网站申请新的license。安装过程较为简单,在此不作赘述。

本论文使用的软件及硬件情况如下。Mac操作系统: macOS Catalina(10.15.7版本); 机型: MacBook Pro(16英寸, 2019年出厂); CPU处理器: 2.3 GHz 8核Intel Core i9; 内存: 16 GB 2667 MHz DDR4; 显卡: AMD Radeon Pro 5500M 4 GB/Intel UHD Graphics 630 1536 MB; 软件版本: MacPymol 2.2.3(以下简称Pymol)。

2 pdb网站及其文件介绍

pdb文件是含有生物大分子结构信息(即三维坐标)的一个文本文件。因此,我们可以用文本编辑软件如Microsoft word、Pages等查看甚至编辑其内容。pdb文件可在蛋白质数据库(Protein data bank, <https://www.rcsb.org/>)下载。截至2022年8

月,该数据库已收藏约20万个由实验方法(包括晶体学、磁共振及冷冻电镜)所测定的生物大分子的三维结构。绿色荧光蛋白GFP是在生物学上具有广泛应用的一个明星分子,曾有多套结构被解析,我们在此以其最新结构7PNN.pdb为例进行演示。该结构是来自于水母*Aequorea victoria*的mVenus变体,暂无相关文献发表。我们可以在PDB网站主页右上方的搜索框内搜索“7PNN”,选择“PDB format”进行下载该pdb文件,之后用word打开。其最上方的信头区(Header)给出对该相关结构的基本描述,如生物大分子的名称、物种来源、衍射分辨率以及结构精修等一些信息。我们关注文档中下方由CRYST1关键词起始及后面的内容,即三维坐标信息(图1)。这些坐标信息由很多列组成,具有严格的格式,分别代表一定的物理含义(表1)。

```
CRYST1 61.064 91.048 118.556 90.00 90.00 90.00 C 2 2 21 8
ORIGX1 1.000000 0.000000 0.000000 0.000000
ORIGX2 0.000000 1.000000 0.000000 0.000000
ORIGX3 0.000000 0.000000 1.000000 0.000000
SCALE1 0.016376 0.000000 0.000000 0.000000
SCALE2 0.000000 0.016983 0.000000 0.000000
SCALE3 0.000000 0.000000 0.005435 0.000000
ATOM 1 N MET A 1 4.541 13.634 1.274 1.00 32.34 N
ANISOU 1 N MET A 1 3564 4644 4079 -1479 -763 -530 N
ATOM 2 CA MET A 1 5.766 14.417 1.177 1.00 29.71 C
ANISOU 2 CA MET A 1 3457 4194 3635 -1234 -790 -288 C
ATOM 3 C MET A 1 6.988 13.509 1.027 1.00 28.57 C
ANISOU 3 C MET A 1 3592 3798 3464 -1277 -748 -109 C
ATOM 4 O MET A 1 6.950 12.349 1.411 1.00 29.55 O
ANISOU 4 O MET A 1 3796 3804 3628 -1502 -670 -98 O
ATOM 5 CB MET A 1 5.908 15.308 2.418 1.00 28.68 C
ANISOU 5 CB MET A 1 3280 4128 3490 -1235 -670 -180 C
ATOM 6 CG MET A 1 4.885 16.419 2.453 1.00 29.65 C
ANISOU 6 CG MET A 1 3151 4492 3621 -1100 -747 -358 C
ATOM 7 SD MET A 1 5.013 17.405 3.959 1.00 29.08 S
ANISOU 7 SD MET A 1 3016 4497 3536 -1119 -597 -250 S
ATOM 8 CE MET A 1 3.744 16.615 4.950 1.00 45.83 C
ANISOU 8 CE MET A 1 4880 6772 5763 -1491 -423 -447 C
ATOM 9 H1 MET A 1 4.531 13.130 2.149 1.00 38.76 H
ATOM 10 H2 MET A 1 4.500 12.977 0.508 1.00 38.76 H
ATOM 11 H3 MET A 1 3.743 14.251 1.232 1.00 38.76 H
ATOM 12 HA MET A 1 5.721 14.993 0.397 1.00 35.60 H
ATOM 13 HB2 MET A 1 5.790 14.764 3.213 1.00 34.37 H
ATOM 14 HB3 MET A 1 6.790 15.710 2.420 1.00 34.37 H
ATOM 15 HG2 MET A 1 5.025 17.006 1.693 1.00 35.52 H
```

图1 pdb文件的内容

3 Pymol软件的使用

3.1 Pymol软件的图形界面及鼠标使用

Pymol的界面主要由两个部分组成:控制窗口(control window)和显示窗口(viewing window)(图2A)。前者主要包括下拉菜单(pull-down menus, MacPymol的菜单在窗口顶部)、常用控制按钮窗口(control buttons)、操作反馈窗口(feedback window)及命令行输入框(command line);而显示窗口则包括可视化窗口(display area)、动画控制(movie controls)、对象菜单(object menu)、鼠标提示模式(mouse hints)以及命令窗口(command line)。

鼠标建议使用三按钮鼠标,可以通过其中键实现这个软件的一些特殊功能。鼠标操作各键功

表1 pdb文件中各变量编号、所占据列数以及其物理含义

变量编号	列编号	描述
1	1-6	记录ID(如ATOM、HETATM)
2	7-11	原子编号
—	12	空白列
3	13-16	原子名称(如“CA” “ND1”)
4	17	其他位置编码(如果存在)
5	18-20	氨基酸残基的标准三字母缩写
—	21	空白列
6	22	多肽链编号
7	23-26	氨基酸编号
8	27	插入编码
—	28-30	空白列
9	31-38	原子的X坐标
10	39-46	原子的Y坐标
11	47-54	原子的Z坐标
12	55-60	该原子占有率
13	61-66	该原子的温度因子
—	67	空白列
14	68-70	脚注

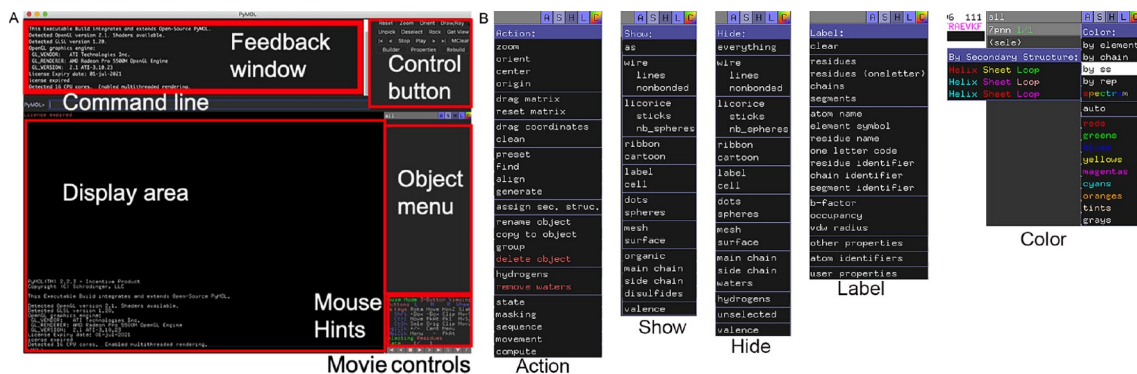
能如下。左键点击并按住三维结构不放进行移动；旋转三维结构；右键点击并按住三维结构不放进行上下移动；放大/缩小三维结构；中键点击并按住三维结构不放进行移动；移动三维结构；Shift+右键：移动剪切平面；鼠标上下移动：调整前剪切平面；鼠标左右移动：调整后剪切平面。

3.2 呈现方式、结构分析及结果保存

打开这个结构后，我们看到最直观的信息是：不对称单位里面的蛋白质分子个数，分子的大致形状，及其是否形成寡聚态、结构域及二级

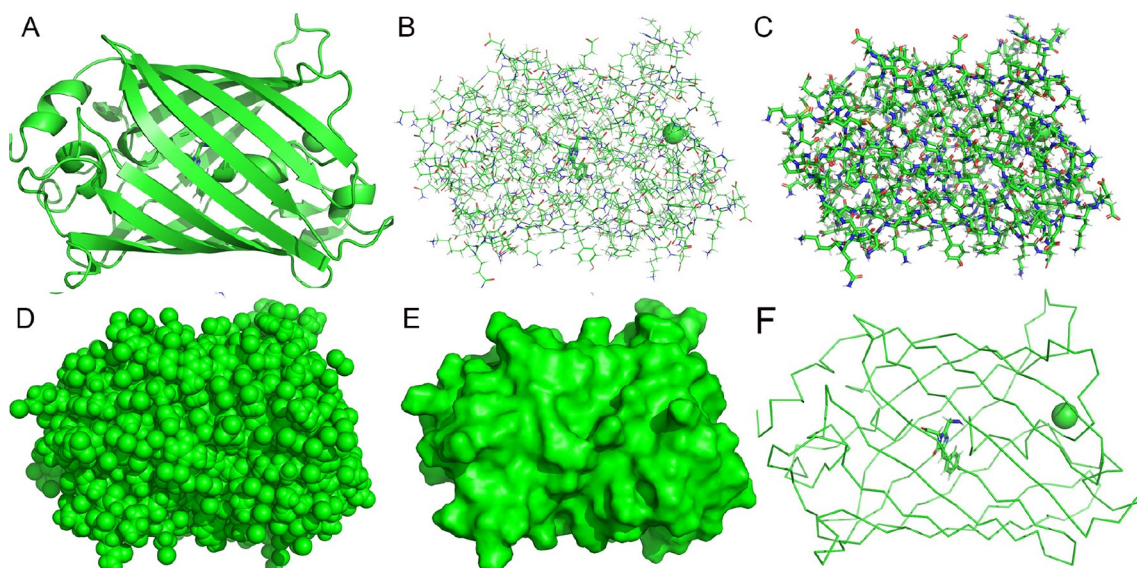
结构的基本情况。GFP由238个残基组成，相对分子质量约27 000。它具有反平行的 β -桶的结构，形成一个近乎完美的圆柱体。11个 β -折叠链形成筒的外周，筒底部被若干较短的 α -螺旋所封闭。 β -圆筒与短螺旋形成致密的结构域。我们下面进一步了解软件的使用情况。

对于初学者来说，最重要的是对象菜单(图2B)，用以显示或呈现生物大分子结构。其下又含有A、H、S、L、C五项下拉菜单，其意义分别解释如下。A(action)代表对这个对象的各种操作；S(show)代表显示这个对象的某种呈现方式；H(hide)代表隐藏该呈现方式；L(label)代表标记某个残基、原子等；C(color)代表着色。我们选择显示菜单，可以看到Pymol的常见的显示方式一共有六种，分别是lines、sticks、spheres、surface、ribbon、cartoon，其中最常用及默认的方式是cartoon，即显示该分子的主要二级结构。当我们双击7PNN.pdb文件时，Pymol被启动，该文件将被装载并在当前版本下呈现cartoon形式，以绿色表示。图3展示了各种不同呈现方式的差异。可以看到，其中lines的显示方式是将所有的原子显示出来，显示了该蛋白质在天然状态下每个原子的精细位置。由于该结构是一个高分辨率的结构，氢原子也被加在了最终模型里，以银色表示；氧、氮、碳、硫原子分别用红、蓝、绿、黄色表示。sticks跟lines差不多，以棍棒表示各原子间的键接，只是连接线较粗。spheres表示方式中则看不到原子之间的价键连接，原子均以密集堆积的小球表示。surface是表面呈现方式，重点突出蛋白质



A: 软件的主要窗口组成; B: 对象菜单及其子菜单: action、show、hide、label和color

图2 Pymol的图形用户界面



A~F分别是cartoon、lines、sticks、spheres、surface和ribbon的呈现方式

图3 mVenus-(7PNN.pdb)各种不同呈现方式的差异

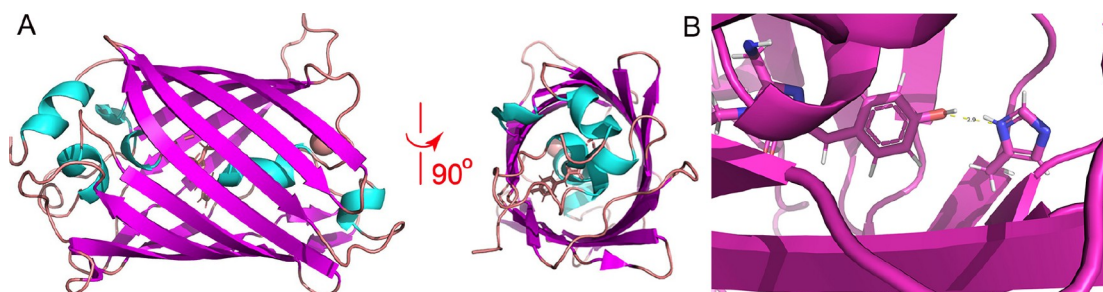
表面的凹凸面，这些不平整的表面对介导该蛋白质与其他分子互动可能起到关键性的作用。ribbon表示方式呈现C α 原子之间的连接，所以只看到主链C α 原子的位置，氧原子、氮原子以及侧链的碳原子均不被显示，此表现方式主要是强调蛋白质主链的走向及折叠方式。上述任何表现方式我们都可以用H选项隐藏，它与S操作相反。

第二个常用的菜单选项“C”的作用是给分子着色。我们可以以任意颜色来对整个分子、结构域或者某个特定区域进行着色。例如，我们可以选择C下面的magenta以及其子菜单选项magenta，将mVenus分子着色成紫红色，或者如图4A所示将其按照二级结构或者寡聚态各单体形式(要求能形成四级结构)来进行着色。Pymol具有复杂而细腻的颜色谱系，可满足多种着色需求。图4A即是按照二级结构对mVenus进行染色后的两种不同视角的卡通图。

在了解染色后，我们可以对GFP特定区域进行选择及简单的结构分析，这可以通过两种办法快速实现。其一是通过右下角的菜单，选择“S”在呈现区域上方显示蛋白质序列，这里显示最终的三维模型包含了结构中可见的229个氨基酸残基(以单字母缩写表示)，一个外源的离子(氯离子)，以及300多个水分子(编号从401到676)。我们可以点击序列中任意一残基对其进行操作，包括其主

链和侧链。在此情况下，最好采用sticks或lines模式，可以清晰地显示兴趣残基。进行上述操作时须按住左键，在序列上滑动就可以对该系列残基(选中的残基以高亮显示)进行选择并开展后续操作。另外一种方式是用命令行的方式，Pymol的命令行的语法通常是command arg1[, arg2[, ...]]。如在输入框输入以下命令：color red, resi 1-10，即可对该蛋白质的前十个残基进行着红色的操作。

当我们可以熟练的实现某一个原子选择的时候，就可以进行一些简单的结构分析。我们最希望分析两个原子之间的互动，例如通过查看其间距离，来推测是否形成氢键或者其他相互作用。选择wizard菜单，并依次点击measurement选项，然后分别点击两个原子，其间软件会在显示窗口区以“please click on the first/second atom”字样提示我们分别选择两原子。两个原子之间的连接将会以黄虚线形式表示，黄虚线上显示的数字即为距离(图4)。野生型组成生色团的三个残基为Gly65-Tyr66-Gly67。它们发生了分子内环化和氧化，并与 α -螺旋共价相连，位于圆筒中央螺旋中部。生色团被严格保护在筒内，性质稳定，不易被淬灭。在mVenus中，生色团被人为改成了Leu65-Tyr66-Leu67。图4B显示了Tyr66侧链羟基与His147侧链形成一对氢键(距离2.9埃)。如果我们对产生的模型图案满意，可以采取导出该展示结构



A: 按照二级结构对mVenus进行染色后的两种不同视角的卡通图; B: 兴趣残基Tyr66侧链与His147形成的一对氢键

图4 mVenus的整体及局部结构

的png格式进行保存。在保存之前点击ray选项,可使结构更具有光泽度和立体感。也可在下拉菜单“File”项中选择“export image as PNG”保存图像文件或“Save session”保存该过程.pse文件。下次如果想打开当前编辑过的状态,只须双击该文件即可。

3.3 软件高级方式

以上是关于Pymol软件简单功能的一些介绍。Pymol功能众多,例如可进行编程并运行一些较为复杂的脚本;在安装某些插件(如APBS^[6])还可以对蛋白质表面电荷分布情况进行精确分析;另外还具有动画制作等高级功能。这些功能可留待学有余力的学生去摸索和揣摩,在此不一一列举。

3.4 适用教学场景

本文作者给中山大学生命科学学院生物技术专业大三本科生开设了结构生物学专业选修课程,旨在向他们介绍结构生物学的入门知识。为加深他们对于该课程的理解,作者开设了配套实验课结构生物学实验技能课,主要进行蛋白质晶体学的相关实验操作的训练。该技能课又分为理论及实验操作部分,其最后一节理论课即如上所述讲解Pymol的初步使用。

3.5 教学经验分享

蛋白质的三维结构的查看及分析对于初学者来说,具有一定难度,需要诸多引导。作者在数次教学后,尝试一定的教改,取得了一定效果,目前发现以下教改措施可以达到显著效果。(1)鼓励学生自带电脑,在预装软件后进行同步软件学习和操作。自备电脑的好处在于学生离开课堂后,可以在课余时间进行自我练习,最终达到熟练使用的目的。(2)由于许多学生是首次接触此类

蛋白质结构分析的软件,建议教学内容不宜过多,难度也需适度控制,点到即止;教学目的以激发学生对蛋白质结构的兴趣为主旨,学有余力的学生可在课后做更深层次的挖掘。(3)建议配备3D眼镜,可提高学生对于结构的三维立体感,提高空间想象力。(4)建议小班教学,每班人数控制在10人以下,增强师生的互动性,也使教学更加个性化、生动化,从而寓教于学。(5)教学内容宜在两小时内完成。在一个半小时的内容讲述及软件演示、练习之后,配备半小时的提问环节,答疑解惑。对于学生在软件使用上的一些实操问题,鼓励学生直接在电脑上运行产生结果,以加深他们的记忆。(6)建议给学生留一个稍有难度的作业题,以保持他们的学习热情。如作者要求学生编写一个小的Pymol脚本并运行。根据结果,充分领悟各命令行的作用。(7)充分发挥教研室的作用,与其他课程形成互动。本课程结构生物学实验技能课为大三学生开设,学生已有生物化学专业课程基础,而生化中的难点之一即是蛋白质的层次结构,彼时学生只能通过课本知识以及教师提供的2D图形来理解蛋白质的三维结构,认识较片面。可在生化课程上即可初步引入对Pymol软件的介绍,打下基础,为本课程作出铺垫。

作者在课程最后给学生播放了两段由Pymol生成效果生动逼真的动画视频。学生在惊讶于软件强大的展示、编辑功能之余,又提出了诸多好奇的问题,如:默认结构显示中绿色的小球和红色的小叉代表什么?它们是怎样来的?为什么有的蛋白质显示氢原子而有些蛋白质不显示?该蛋白质结构中含何种生色团?具有何种作用?这些问题都与教学内容高度相关,证明学生确实动了脑筋,也

启发了他们的思维。课后学生纷纷反馈他们对这次教学内容印象尤深,难以忘怀。此教学内容的开设有力促进了学生对蛋白质三维结构的理解,极大提高了学生的学习兴趣以及加深其对于蛋白质生物大分子三维结构和层次结构的了解,激发他们努力去发掘和探索科学的真相,取得了较好的教学效果。

当然,教学过程中也遇到了一些问题。例如,在编写一些较为复杂的脚本时,学生表示理解上存在困难,可能需要降低难度;又如在运行小电影时,学生的笔记本电脑运行较慢,效果不佳,建议在具备条件时在专用机房进行该教学内容。

4 结语

结构生物学在未来生物医药领域将起到越来越重要的作用。AlphaFold 2的高准确率的结构预测能力带来的是生物学史无前例的一次技术革命,也是人类取得的最重要的科学突破之一。可以想象,随着这一学科的发展,生物学家可以深入理解未知蛋白质或发现蛋白质新的功能。例如,在酶工程领域,关键残基的空间结构信息对于酶的设计至关重要。天然酶经过蛋白质工程,会变为效率、稳定性及活性更高的工具酶。在理解病理及抗病临床应用方面,人们可以在了解功

能蛋白质突变引起的结构基础上,推测其功能改变和相关遗传性疾病的病理产生及发展过程,同时可以根据该蛋白质结构信息进行基于结构的理性设计,为疾病治疗提供新方法和思路。生物学工作者通过结构可视化软件Pymol等的熟练掌握及应用,可打下坚实的微观生物学基础,为我国生物技术及生物医药产业的腾飞贡献力量。

参考文献

- [1] Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 2021, 596(7873): 583-589
- [2] Varadi M, Anyango S, Deshpande M, et al. AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res*, 2022, 50(D1): D439-D444
- [3] Guex N. Swiss-PdbViewer: a new fast and easy to use PDB viewer for the Macintosh. *Experientia*, 1996, 52: A26
- [4] Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. *J Mol Graphics*, 1996, 14(1): 33-38
- [5] Pettersen EF, Goddard TD, Huang CC, et al. UCSF Chimera? A visualization system for exploratory research and analysis. *J Comput Chem*, 2004, 25(13): 1605-1612
- [6] Holst M, Saied F. Multigrid solution of the poisson? Boltzmann equation. *J Comput Chem*, 1993, 14(1): 105-113