ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.17, Supplement, November 2006, pp.103−109

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

面向大分子的三维数据场特征分析与可视化*

韩 玮[1], 汪 莉[1], 陈 为[1+], 万华根[1], 彭群生[1], 吴 韬[2], 王 琦[2]

[1](浙江大学 CAD&CG 国家重点实验室,浙江 杭州 310027)

[2](浙江大学 化学系 分子设计与分子热力学研究所,浙江 杭州 310027)

Feature Analysis and Visualization of 3D Scalar Field with the Applications to the Macromolecule

HAN Wei[1], WANG Li[1], CHEN Wei[1+], WAN Hua-Gen[1], PENG Qun-Sheng[1], WU Tao[2], WANG Qi[2]

[1](State Key Laboratory of CAD&CG, Zhejiang University, Hangzhou 310027, China)

[2](Institute of Molecular Design & Thermodynamics, Zhejiang University, Hangzhou 310027, China)

+ Corresponding author: Phn: +86-571-88206681 ext 522, Fax: +86-571-88206680, E-mail: chenwei@cad.zju.edu.cn

Han W, Wang L, Chen W, Wan HG, Peng QS, WU T, WANG Q. Feature analysis and visualization of 3D scalar field with the applications to the macromolecule. *Journal of Software*, 2006,17(Suppl.):103−109.
http://www.jos.org.cn/1000-9825/17/s103.htm

Abstract: This paper introduces the primary attempts on the modeling, analysis and visualization of the 3D macromolecular scalar field. According to the quantum chemical theory, one protein molecular structure is transformed into a regularly sampled 3D scalar field, in which each node records the combined effect of different actions in protease. By applying the first order and the second order local differential operators on individual node, a set of critical points which potentially depicts the active region of protein molecule are found. Also the paper gives some results after computing a sequence of molecular potential energy in the data field and interactively exploring the potential "tunnel" region exhibiting biological sense. In addition, the point-based, surface and volume rendering techniques are exploited to find the macro-structure inside the data field. With all these techniques, the escape route of water molecules hidden in the HIV-1 protease is successfully detected, which is in accordance with the experimental results.

Key words: 3D scalar field; feature analysis; visualization; critical point; protein, macromolecule

摘 要: 介绍了在面向生物大分子结构和功能分析的三维数据场建模、特征分析与可视化方面的初步尝试.从蛋白质分子结构出发,采用量子化学理论计算得到一个规则采样的三维数据场,场的每个格点上记录蛋白酶分子内部各种力的综合作用.在每个格点上实施离散一阶、二阶局部微分计算,从而筛选出一系列数据场内的临界点,这些临界点潜在地揭示了蛋白质分子的功能区域所在.继而,计算数据场内各种型值的分子势能面,交互地探寻具有一定生物活性的"通道"区域.此外,探索运用多种点、面和体可视化技术,来寻找分子内部的宏观结构.通过上述多种特征分析与可视化手段,成功地寻找到 HIV-1 蛋白酶分子中隐藏的水分子排出通道.

* Supported by the National Natural Science Foundation of China under Grant Nos.60533050, 60503056, 60021201 (国家自然科学基金)

Received 2006-03-15; Accepted 2006-09-11

关键词: 三维数据场;特征分析;可视化;临界点;蛋白质;生物大分子

Anfinsen 指出:“理解细胞行为的最佳方式是研究蛋白质分子的结构与功能的关系”[13].而蛋白质结构除了通过生物实验进行验证外,还可以从序列相似性比较和几何拓扑的角度来进行预测[5,14,18].这两类方法都建立在对蛋白质分子结构合理建模的基础之上.从 20 世纪 70 年代开始,研究者们推出了一系列蛋白质分子的几何表示方法,典型的有线框表示、棍状表示、球棍表示、CPK 表示、带状表示、卡通表示、管片表示等[12].本质上,它们是基于实验数据对蛋白质中各原子间作用关系的一种抽象,在图形学中表达为一系列线段和面模型的集合.它们的优势在于能提供用户一种简单直观和交互的方式辅助观察分子的几何和拓扑结构.但是,现有的模型缺乏对分子的运动、分子静电势场的有力刻画.其一是,分子图和线面表示缺乏对力场的表示.力场是一个分布在三维空间中连续的空间量.在分子图表示中,原子之间的距离用连通关系表示,而线面表示中,力的作用是通过球面之间相连的棍状表示.两者仅仅反映了关系的存在,而无法表达力的大小、位置和相对方位.其二,蛋白质分子时刻都处于运动之中,三维空间中的力场是一个变化的量.仅用分子图和线面表示难以表达出动态的整体状态.因而在此基础上进行几何结构分析尚存在难以突破的局限性.

一个自然的想法是,可否用一种三维的空间场来表征蛋白质分子的空间功能结构?这样的一种模型应该直接建立在实验数据的基础上,且操作简单,可控性好,能描述分布在空间各点处的分子势能或其他物理化学量.我们将单个蛋白质分子表示为三维空间中一系列规则采样的离散三维格子的集合,每个格子点上记录该格子中心点处的属性数值.三维数据场的建模与分析在科学计算可视化、计算机辅助几何设计、飞行模拟与动画等领域得到了广泛的应用,其在宏观世界的建模与可视化技术日臻成熟,但至今尚很少有相关文献涉及到分子级别的几何处理.在分子设计和蛋白质分子结构预测研究领域,一般的工作集中在分子图和线面模型表示上,已有的体表示研究都不是建立在全局的三维场的表达上,它们或对整个空间计算一个场值[6],或将面模型体素转化为体模型并进行蛋白质分子三维相似性计算[1,2,8,10,11],或对分子的电子断层扫描(CT)图像序列进行特征提取和可视化[4].

从分子学的角度看,以力场能为基础来确定配体与蛋白质之间的相互作用和热力学构象所进行的简单自由能计算在配体结构设计方面是非常有用的工具.因此,建立一套针对蛋白质分子属性的三维数据场表示与分析方法,具有重要的研究价值.在三维数据场的语境下,特征是指数据场数据中蕴涵的某类特殊的信息、用户感兴趣的区域或能区别不同数据之间的标识.以 HIV-1 蛋白酶(即艾滋病病毒)为例,现已证实,其活性位点位于一条狭长的“通道”底部,具有二重对称性,当与抑制剂结合后,蛋白质的结构,特别是挡板的结构会发生很大变化[1].本文以 HIV-1 蛋白酶分子结构为实验对象,进行了三维数据场计算、基于局部微分算子的特征分析、基于分子势能面的 HIV-1 蛋白酶“通道”特征区域抽取和体可视化等一系列工作.初步实验表明,我们的方法计算出的特征区域具有重要的生物意义,与已知的生物学结论一致.

本文第 1 节介绍与三维数据场分析相关的背景工作.第 2 节以 HIV-1 蛋白酶分子为实验对象,详细描述了我们的方法和实验结果.第 3 节概括全文,并简述未来方向.

1 相关工作

有关蛋白质分子的表达、建模与分析的文献很多,本节我们简单描述最相关的代表性工作.

1.1 蛋白质分子的数据采集和几何表达模型

蛋白质分子实测方法除了利用X-射线衍射、核磁共振等实验手段以外,也借助于信息、自动化方法对分子结构予以预测,主要分为两类:一类依赖于序列数据,采用统计学方法来分析其结构和功能;另一类直接从实验测定已知的(或预测出的)三维结构出发,着重考虑结构与几何拓扑性质,进而分析其功能.这两类方法均建立在对生物大分子结构合理建模的基础之上.

目前,针对蛋白质分子表示的计算机模型有很多.这些模型的建立主要是依赖于由原子方位、排列顺序、连

接方式等决定的分子骨架形状、表面几何及拓扑性质.通过对蛋白质三维结构原子空间定位及连接关系、Cα链、二级结构、模体(motif)等进行合理抽象,构造一系列线/面模型,可以更直观地表示蛋白质分子的几何与结构.

1.2 比较分子场分析法(CoMFA)

在计算机辅助药物设计中,比较分子场分析法(comparative molecular field analysis,简称 CoMFA)[6,8]一直是研究的热点,经过十几年的发展,目前已成为最成熟且应用最广泛的三维定量构效方法(3D-QSAR).其基本原理是:首先在分子周围定义分子场空间并均匀划分,在每个格点上计算分子场特征(一般为静电场和立体场,有时也包含疏水场和氢键),然后采取偏最小二乘法进行回归分析,建立化合物生物活性和分子场特征之间的关系.

对于小分子(<1nm),CoMFA 从分子的拓扑、几何、结构、物理、化学属性出发,寻求结构与功能的关系,取得一定的成功.但是,对于蛋白质等大分子来说,一方面,结构的动态性对功能的意义重大;另一方面,缺乏有效的算法对蛋白质分子(大小一般在 1nm~100nm 之间)构建具有明确物理意义的数据.

1.3 分子拓扑学

自 Mezey 开展分子势能面拓扑性质的研究以来,微分和拓扑已经成为有效地分析分子体系化学结构以及与反应机理之间关系的工具.这些工具通常考虑某一邻域范围内关键点,并有效地抽取局部特征.例如,定义分子势能面为多维空间上的超曲面,在其上定义一个连续的势能函数 $U(X)$,其临界点即指梯度为零的点.由于临界点处蕴涵着某种特征,故须在临界点处对势能函数做二阶微分,计算曲率并分析其类别.基于数据来源的限制,分子势能函数 $U(X)$多以离散形式表示.

拓扑分析的方法也存在一定的局限性.例如,它缺乏定量描述,没有具体的感知和度量标准,需要和其他有效的分析方法相结合来描述分子势能面特性.我们以分子拓扑学中临界点理论为知识背景,在一个规则采样的数据场中考虑分子系统综合作用函数,计算并抽取临界点及判断三维空间中各种满秩临界点情况.

2 我们的工作

本节依次给出蛋白质分子的三维数据场计算方法、临界点抽取原理及可视化效果,并抽取蛋白质分子势能等值面,最后给出对蛋白质分子的三维数据场的直接体绘制结果.综合这些分析和可视化手段,我们成功地识别出 HIV-1 蛋白酶分子中的特征区域.

2.1 蛋白质分子的三维数据场计算

蛋白质分子的三维数据场是分布在三维空间的离散场.具体而言,蛋白质分子中各原子或亚结构的动力学特征可以用其哈密尔顿来表示.将蛋白质分子所处空间均匀剖分为网格,并在网格点上定义离散函数,即可将蛋白质分子的三维数据场哈密尔顿写为

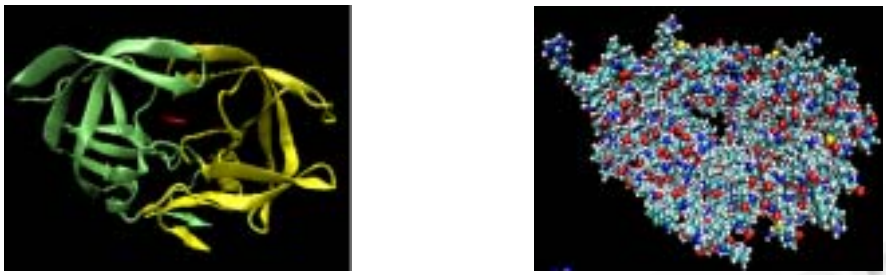
$$H_{field}=\sum_i\sum_j\sum_k H_{ijk}(t),$$

其中, H_{ijk} 是描述特定空间格点运动行为的哈密尔顿.可由该离散场出发描述体系的特征.

研究以 HIV-1 蛋白酶(PDB code: 1A30,Louis,J.M., *et al*., Biochemistry, 2105, 1998)为目标原型.由于实验上发现其可以作为抗 HIV 药物的有效靶点,目前针对该蛋白已有大量的理论和实验研究见诸报道.我们首先采用其 X-ray 衍射构象为出发点,构造同时含有 4 691 个水分子的体系.然后在 310K,1atm 条件下采用 Charmm 力场进行 1ns 的平衡计算,以模拟该蛋白在体内液体环境的柔性结构.而后做 20ps 的采样计算.这里展示的是其中的一个采样,根据以上思想,采用我们在 Gaussian03 基础上,自行设计的线性标度分子三维数据场计算程序 MolField 计算得到的数据场.这里分子三维数据场格点数据是采用 AM1 方法计算得到的量子化学静电势分布表示.这是由于此算法已被大量研究证实可有效地表述有机分子的结构信息.

HIV-1 蛋白酶是由一个小阻抗剂和两条含 99 个氨基酸的多肽链形成的 C2 对称的均二聚体,每个单体中包含有两个模体,都由反平行的β折叠组成.图 1(a)为基于二级结构的新卡通显示模型,绿色和黄色分别代表两条多肽链,红色为阻抗剂.图 1(b)为 HIV-1 蛋白酶的球棍显示模型,其中不同颜色的圆球表示不同原子,并以该原子

的范德华半径作为圆球半径,原子间以无向棍棒相连接.



(a) (b)

Fig.1 The NewCartoon (a) and CPK (b) models of HIV-1 protease

图 1 HIV-1 蛋白酶的新卡通模型(a)和球棍模型(b)

2.2 基于三维数据场表示的蛋白质分子临界点抽取

令 X 为蛋白质分子三维离散的规则数据场,(x,y,z)为某格点坐标,$U(X)$为给定的势能函数.在每一格点上,其梯度为

$$g(x,y,x)=\frac{U(x+1,y,z)-U(x-1,y,z)}{2}i_x+\frac{U(x,y+1,z)-U(x,y-1,z)}{2}i_y+\frac{U(x,y,z+1)-U(x,y,z-1)}{2}i_z \quad (1)$$

梯度方向代表正交于 $U(X)$的等值面,i_x,i_y,i_z 表示三个单位向量.临界点即满足 $g(x,y,z)=0$ 的点.我们取梯度 3 个分量的绝对值之和作为梯度的模绘制梯度标量场.图 2(a)中灰色亮片为临界点插值所得曲面.

根据 Morse 理论[9],"一个纯量场的二次微分能够表现出此纯量场的局部分布情况,对于电子密度纯量场而言,其二次微分可定义出局部电子密度之累积增加."对于分子势能函数亦如此.为进一步考虑蛋白质分子三维数据场的特性,我们计算其分子势能函数 U(X)的 Hessian 矩阵:

$$\begin{bmatrix} H_{11},H_{12},H_{13} \\ H_{21},H_{22},H_{23} \\ H_{31},H_{32},H_{33} \end{bmatrix},$$

其中, H_{ij} 为二阶微分或表示为

$$\nabla^2U(X)=\nabla(\nabla U(X))=\lambda_1+\lambda_2+\lambda_3,\quad \lambda_1<\lambda_2<\lambda_3 \quad (2)$$

$\lambda_1,\lambda_2,\lambda_3$ 分别为 Hessian 矩阵的 3 个特征值,其正负号的分布决定了临界点性质和分类.计算每个临界点的 Hessian 矩阵,判断其是否满秩.用符号(r,s)表示临界点,其中 r 指 Hessian 矩阵非零特征值个数,s 指 Hessian 矩阵特征值正负符号之和,并依据 r,s 的不同将临界点分类,逐一进行分析[3].以三维空间为例,共有 4 种满秩临界点:

(3,−3)原子核(nucleic attractor),其分子势能各个方向来看都是极大值.

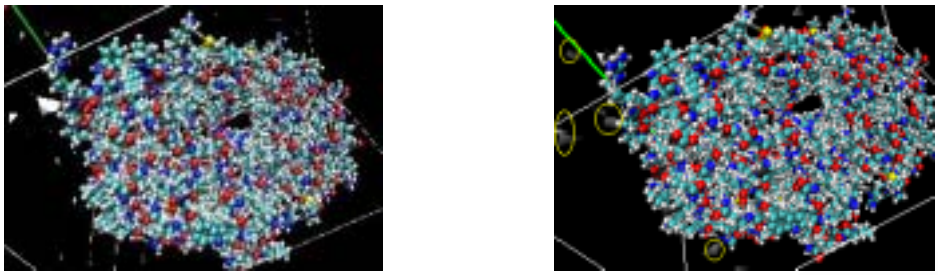
(3,−1)键临界点(bond critical point),是势能面上正马鞍型的鞍点.在任意两个具备键结性质的原子之间,沿键结方向看是极小值,从另外两个垂直键结方向观察是极大值.

(3,+1)环临界点(ring critical point),与 BCPs 观察极值效果相反,是势能面上反马鞍型的鞍点.

(3,+3)笼临界点(cage critical point),其分子势能各个方向来看都是极小值.

且由 Poincar é-Hopf 规则,临界点满足方程:$n-b+r-s=1$,其中,n 为原子核临界点(NAs)个数,b 为键临界点(BCPs)个数,r 为环临界点(RCPs)个数,s 为笼临界点(CCPs)个数.

参考分子势能面非退化临界点分类准则,依据(r,s)不同,将满秩点分为 4 类并全部绘制.如图 2(b)所示,灰色曲面为满秩临界点线性插值所得.



(a) (b)

Fig.2 The gray surfaces of (a) and (b) were constructed by the interpolation of critical points which satisfy equation1 and whose Hessian matrix was non-degenerate respectively

图 2 (a)灰色区域为临界点(即满足式(1))插值形成,(b)灰色区域为非退化临界点(Hessian 矩阵满秩)插值形成)

2.3 基于三维数据场表示的蛋白质分子势能面抽取

HIV-1 蛋白酶的活性位点位于“通道”底部,采用等值面抽取算法分为以下几步[17]:

(1) 给定阈值 C,将蛋白质三维数据场中每个体元的所有角点与阈值 C 相比较,根据比较结果,构造该体元的状态表.

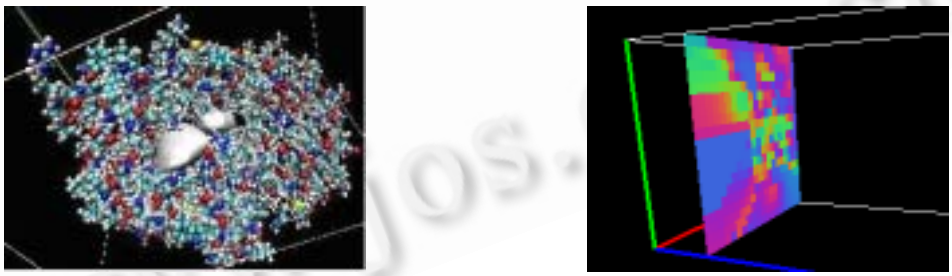
(2) 根据状态表,得出将与等值面有交点的体元边界.

(3) 通过线性插值方法,求出体元边界与等值面的交点.

(4) 利用中心差分方法,计算出体元各交点处的法向,再通过线性插值方法,求出三角形各顶点法向.

(5) 采用不同的阈值可得到一系列等值面根据各三角面片各顶点的坐标值及法向量绘制等值面图像.若取 $C=-0.042$,结果如图 3(a)所示.

观察得知,HIV-1 蛋白酶分子数据场值的范围约为–0.055~0.019,阈值 C 取 0.000 左右时数据场分布较大,且“通道”附近的分子数据场值范围约为–0.048~–0.033.用任意平面切 HIV-1 蛋白酶,该平面上任意点的颜色值可以定性地表示该点数据场值的大小.如图 3(b)所示,用平面 $z=0.23$(坐标归一化后)切 HIV-1 蛋白酶,做颜色映射并显示所得切平面,这样可以较直观地看到其数据分布及变化规律.



(a) The isosurface whose potential value is –0.042

(a) 显示 HIV-1 蛋白酶分子势能函数值为–0.042 时的等值面,此时较逼近通道区域

(b) The volume slice with z value 0.23

(b) 显示平面 $z=0.23$,并映射颜色定性表示其数值大小,如红色代表该点数值较小,蓝色代表该点数值较大

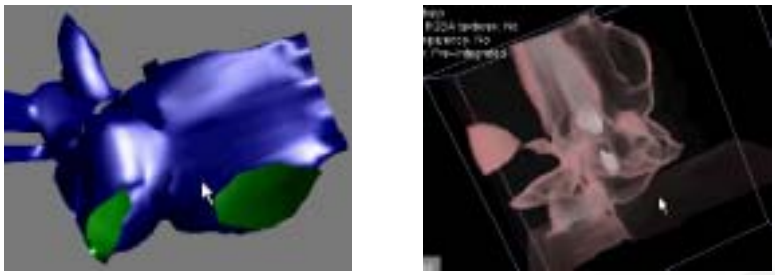
Fig.3

图 3

2.4 面向蛋白质分子的三维数据场体可视化

我们分别采用光线投射和三维纹理映射算法[17]实现了三维数据场的体可视化.其中,光线投射技术是一种较为成熟的以图像空间为序的体绘制方法.其基本思想是对于图像平面上的每一像素,从视点投射出一条穿过该像素的视线,直接利用视线穿过体数据空间时的采样值计算像素的光强.另一种直接体绘制是三维纹理映射算法.我们用两种算法实现了等值面绘制(图 4(a))和半透明绘制(图 4(b)).从可视化结果可以发现,图 4(b)的半透明绘制效果对理解蛋白质分子的空间能量分布非常有帮助.我们从中成功地观察到 HIV-1 蛋白酶分子中隐藏

的水分子排出通道.其生物化学功能还有待进一步的理论和实验工作揭示.



(a) Isosurface rendering (b) Semi-Transparent rendering

Fig.4 The volume rendering of 3D protein molecular field

图 4 蛋白质分子的三维数据场体绘制效果

3 结论与展望

本文以 HIV-1 蛋白酶为例,着重于蛋白质三维数量场的特征运算,详细描述数据场计算原理;借助 Gaussian 03,Visual C++ 6.0,VMD,POV-Ray 等工具,利用一阶、二阶局部微分算子,得到一系列有可能蕴涵某种生物特性的临界点(如图 2 所示);通过计算各种型值的等值面,成功抽取并可视化具有一定生物活性的通道区域(如图 3(a)所示),并判断出“通道”位置数据场数据的大概范围;采用多种可视化技术观测蛋白质分子的整体结构(如图 4 所示).为今后进一步分析 HIV-1 蛋白酶功能与结构的关系提供丰富的实验数据和理论依据.

蛋白质分子三维数据场的特征抽取和可视化研究是一项有价值且意义甚远的研究.对三维数据场的研究一直是国际前沿问题,但应用到生物大分子领域的少之又少.我们率先在国内实现了对生物大分子标量场的可视化,这只是第一步,今后拟实现对其向量场的可视化,可用多种图形描述手段如颜色、长度、角度、透明度、箭头、锥体、六面体等进行显示.考虑到蛋白质分子时刻处于运动当中,对时变数据场进行可视化能更加有效地揭露分子在完成其功能过程中的演化情况及多分子之间相互作用的机制.拟采用的可视化手段有基于颜色和光学属性的向量场映射、基于质子跟踪的显示方法和基于纹理的向量场动态可视化等.

致谢 在此,我们向对本文的工作给予支持和建议的老师和同学们表示感谢.

References:

[1] Ankerst M, Kastenmüller G, Kriegel HP, Seidl T. 3D Shape histograms for similarity search and classification in spatial databases. In: Güting RH, Papadias D, Lochovsky F, eds. Proc. of the 6th Int’l Symp. on Spatial Databases (SSD 1999). Heidelberg: Springer-Verlag, 1999. 207.

[2] Ankerst M, Kastenmüller G, Kriegel HP, Seidl1 T. Nearest neighbor classification in 3D protein databases. In: Proc. of the 7th Int’l Conf. on Intelligent Systems for Molecular Biology (ISMB’99). Heidelberg: AAAI Press, 1999. 34–43.

[3] Bader RFW, Austen MA. Properties of atoms in molecules: Atoms under pressure. Journal of Chemical Physics, 1997,107:4271–4285.

[4] Bajaj CL, Pascucci V, Shamir A, Holt RJ, Netravali AN. Multiresolution molecular shapes. TICAM Report, 1999. 99–42.

[5] Branden C, Tooze J. Introduction to Protein Structure. 2nd ed., New York: Garland Publishing, Inc, 1998.

[6] Cramer RD, Patterson DE, Bunce JD. Comparative molecular field analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. Journal of Am. Chem. Soc., 1988,110:5959–5967.

[7] Keim DA. Efficient geometry-based similarity search of 3D spatial databases. In: Proc. of the ACM SIGMOD Int’l Conf. on Management of Data (SIGMOD’99). 1999. 419–430.

[8] Kubinyi H. Comparative molecular field analysis (CoMFA). In: The Encyclopedia of Computational Chemistry. John Wiley & Sons Ltd., 1998. 448–460.

[9] Morse P, Feshbach H. Methods of Theoretical Physics, Part 1. New York: McGrawHill, 1953.

[10] Katchalski-Katzir E, Shariv I, Eisenstein M, Friesem AA, Aflalo C, Vakser IA. Molecular surface recognition: Determination of geometric fit between proteins and their ligands by correlation techniques. In: Proc. of the National Academy of Sciences. 1992, 89:2195−2199.

[11] Kriegel HP, Kroeger P, Mashael Z, Pfeifle M, Poetkey M, Seidlz T. Effective similarity search on voxelized CAD objects. In: Proc. of the 8th Int'l Conf. on Database Systems for Advanced Applications (DASFAA 2003). Los Alamitos: IEEE Computer Society Press, 2003.

[12] Leach AR. Molecular Modelling: Principles and Applications. 2nd ed., Pearson Education EMA, 2001.

[13] Anfinsen CB. The Molecular Basis of Evolution. New York: John Wiley & Sons, Inc. 1959.

[14] Xu XJ, Hou TJ, Qiao XB, Zhang W. Computer Aided Molecular Drug Design. Beijing: Chemical Industry Press, 2004 (in Chinese).

[15] Lai LH. Protein's Structure Prediction and Molecular Design. Beijing: Peking University Press, 1993 (in Chinese).

[16] Wang BN, Chen LJ, Wang J. The method to research the chemical molecule using the theory of electronic density topology. Physics Bimonthly, 2004,26(3):530−536 (in Chinese with English abstract).

[17] Tang ZS, *et al*. Scientific Visualization of 3D Data Set. Beijing: Tsinghua University Press, 1996 (in Chinese).

[18] Xin HW. Molecular Topology. Hefei: USTC Press, 1992 (in Chinese).

附中文参考文献:

[14] 徐筱杰,侯廷军,乔学斌,章威.计算机辅助药物分子设计.北京:化学工业出版社,2004.

[15] 来鲁华.蛋白质的结构预测与分子设计.北京:北京大学出版社,1993.

[16] 王本宁,陈立基,王瑜.电子密度拓扑学研究化学分子的方法.物理双月刊,2004,26(3):530−536.

[17] 唐泽圣,等.三维数据场可视化.北京:清华大学出版社,1996.

[18] 辛厚文.分子拓扑学.合肥:中国科学技术大学出版社,1992.

韩玮(1980－)女,河北辛集人,博士生,主要研究领域为分子图形学,计算几何,科学计算可视化.

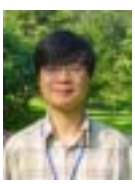
汪莉(1982－),女,博士生,主要研究领域为分子图形学,生物计算.

陈为(1976－),男,副研究员,主要研究领域为可视化,计算机图形学.

万华根(1968－),男,副研究员,主要研究领域为计算机动画,虚拟现实,科学计算可视化.

彭群生(1947－),男,博士,教授,博士生导师,主要研究领域为真实感图形,虚拟现实,红外成像仿真,分子图形学,科学计算可视化.

吴韬(1973－),男,副研究员,主要研究领域为计算化学,计算机在化学、化工、生物医药中的应用.



王琦(1963－),男,教授,主要研究领域为计算化学,分子模拟,分子设计.