基于三维数据场拓扑抽取的蛋白质结构分析*

张 涛, 陈 为+, 谢利广, 胡 敏, 彭群生

(浙江大学 CAD&CG 国家重点实验室,浙江 杭州 310027)

Molecular Structural Analysis Based on the Topological Extraction of 3D Scalar Fields

ZHANG Tao, CHEN Wei+, XIE Li-Guang, HU Min, PENG Qun-Sheng

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

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

Zhang T, Chen W, Xie LG, HU M, Peng QS. Molecular structural analysis based on the topological extraction of 3D scalar fields. *Journal of Software*, 2006,17(Suppl.):120−125. http://www.jos.org.cn/1000-9825/17/s120.htm

Abstract: In this paper, efforts on extracting local features and building the global topology of 3D macromolecular scalar field are introduced. Base on the topological analysis, the critical points which potentially indicate the active regions are calculated. This technique is employed to construct the topolog ical and geometrical structures of the scalar field. The preliminary experiments with the protein data sets verify the feasibility of the method.

Key words: 3D scalar field; critical point; contour; contour tree; Morse function

摘 要: 介绍了抽取蛋白质分子三维数据场的局部特征,并组织数据场的整体拓扑的初步工作.基于拓扑分析的方法,计算出数据场的局部特征点,这些特征点潜在地揭示了蛋白质的活性位点.在此基础上,利用这些局部特征点构造出表示数据场的几何拓扑结构.将实验结果与实际的蛋白质数据比较,验证了方法的可行性.

关键词: 三维数据场;奇点;轮廓;轮廓树;Morse 函数

蛋白质是生命中的重要分子,几乎在所有的生物学过程中都扮演着重要的角色,因此,关于其结构与功能的研究一直是人们关注的焦点.现有的许多蛋白质结构预测方法通过原子序列来预测它们的空间结构.近 40 年来,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 表示等,它们为用户提供了一种简单直观和交互的方式来观察分子的几何和拓扑结构.然而,这些方法仅能反映原子之间的连接关系,无法对分子的静电势场进行有力刻画.此外,蛋白质分子处于运动之中,它的力场是一个变化的量.现有的表示方法难以对这种动态的整体状态进行很好的表示,也无法在此基础上进行几何结构分析.

另一方面,现在已有一些实测方法来对蛋白质分子的数据进行采集,如 X-射线衍射、核磁共振等实验手段,这些实测数据表示为三维空间中规则采样的体数据,反映了蛋白质的空间作用力的分布情况.同时,人们也开发了多种用于计算分子静电势场的参数模型,如 MM3,MM4,DREIDING 等,它们同样表示为三维空间中规则采样

* Supported by the 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

© 中国科学院软件研究所 http://www.jos.org.cn

的体数据.与现有的蛋白质表示方法相比,这种表示在刻画蛋白质分子体系方面具有其鲜明特色,首先,它将蛋白质分子体系考虑为一个整体的场;其次,数据场遍布于三维空间.因此基于三维数据场的几何和拓扑描述将有利于预测分子结构,寻找分子的活性位点和功能区域.

三维数据场的特征分析在科学计算可视化、流体力学和动画领域的到了广泛应用,但是涉及到它在蛋白质功能预测和活性位点的定位方面的文献还不多见.本文以分子 pdb1a07 的 A 链骨架碳原子和分子 1crn 计算出来的电子密度场为实验对象,对它进行了局部微分特征抽取和整体拓扑建立的工作.初步结果表明,我们的结论与实际数据比较吻合,方法具有可行性,有望推广到实测数据和动态蛋白质分子数据.

1 相关工作

三维数据场的几何分析与拓扑分析一直是研究的热点,本节介绍最相关的代表性工作.

Masaki Hilaga 等人利用 Reeb Graph[1]对曲面上的标量场函数进行研究.他们通过 Reeb Graph 对曲面进行编码,非常有效的提取了曲面的拓扑结构,为三维检索提供了有效的工具.Carr 等人研究了轮廓树[2],提出了在任意维空间计算轮廓树的算法,Yi-Jen Chiang 改进了这种算法[3].

2 背景知识

2.1 Morse理论

设 M 是一个光滑的 3 维流形, $f:M\to R$ 是一个光滑函数,如果在 $p\in M$ 处有 $\nabla f(p)=0$,那么 p 称为 f 的一个奇点,否则称其为正则点.如果 f 在其奇点 p 处的海赛矩阵可逆,那么 p 称为非退化奇点,否则称为退化奇点.如果 f 的所有奇点都是非退化奇点,并且所有奇点上的函数值都不相等,那么 f 称为 Morse 函数.

Morse 理论给出了 f 在它的奇点附近的标准表达式,以此为基础将其奇点分成 4 类:极大点、极小点、2-鞍点和 1-鞍点.Morse 理论进一步指出:对于 $h\in R, f^{-1}(h)$ 连通分支的亏格只在 f 的奇点处发生改变[4],因此奇点反映了矢量场函数 f 的等值面的拓扑,刻画了它的局部微分属性.

2.2 轮廓树

f 的上水平集、下水平集和水平集分别定义为

$$M_{>h}:=\{x\in M\mid f(x)>h\},$$
$$M_{<h}:=\{x\in M\mid f(x)<h\},$$
$$M_{=h}:=\{x\in M\mid f(x)=h\}.$$

$M_{=h}$ 的连通分支称为轮廓,f 的轮廓树是记录其所有轮廓之间关系的数据结构,它的节点表示轮廓的产生,消失,分裂或者合并,它的边则表示了这些变化的连续过程.轮廓树的严格定义如下:

定义 1. Morse 函数 f 的轮廓树是满足下面条件的图:

1. 每个度为 1 的节点是一个局部极值点,表示轮廓的产生或者消失.
2. 每个度不为 1 的节点是鞍点,表示一个轮廓分裂成多轮廓,或者表示多个轮廓合并成为一个轮廓.
3. 每条边是所有轮廓构成的空间中的一个连通分支,表示了 1,2 中的轮廓(即等值面)的变化过程.

一个 Join 分支定义为 $M_{\le h}$ 的连通分支,一个 Split 分支定义为 $M_{\ge h}$ 的连通分支.Join 树(记为 JT)是一颗树,它的边表示一个 Join 分支,叶节点表示一个 Join 分支的产生.Split 树(记为 ST)是与 Join 树对应的树,它的边表示一个 Split 分支.Carr[2]证明了轮廓树可以通过合并 JT 和 ST 得到.

3 分子电子密度场的特征抽取与表示

轮廓树是表示三维数据场的整体几何和拓扑的数据结构,它表示了分子电子密度场的几何特征,可能表示了分子潜在的活性位点,对于寻找分子的活化功能区域和进行蛋白质的相似性对比有重要的意义.本节将阐述对分子三维数据场的局部特征点抽取、利用轮廓树建立整体拓扑并进行简化的方法.我们将计算出来的轮廓树

© 中国科学院软件研究所 http://www.jos.org.cn

与分子的实际结构数据进行了比较,实验结果表明轮廓树确实体现了分子的结构,证明了方法的可靠性.

3.1 局部特征搜索

奇点是三维数据场的局部特征,我们采用文献[3]的方法计算奇点.

顶点 v 的连接图 $N(v)$ 是所有包含 v 的单形[5]的所有顶点和边,除去 v 以及和 v 相邻的边后得到的图. $N_+(v)$ 和 $N_-(v)$ 都是 $N(v)$ 的子图,它们分别由 $\{w\in N(v)\mid f(w)>f(v)\}$ 和 $\{w\in N(v)\mid f(w)<f(v)\}$ 生成. $C_+(v)$ 和 $C_-(v)$ 分别表示 $N_+(v)$ 和 $N_-(v)$ 连通分支的个数.

定义在可三角化流形上分段线性函数的奇点只能是网格顶点[2],可按照下面的规则来计算并进行分类[3]: (1) $C_+(v)=0$,v 是极大点;(2) $C_-(v)=0$,v 是极小点;(3) $C_+(v)>1$ 并且 $C_-(v)>1$,v 是鞍点.

3.2 整体拓扑建立

奇点反映了三维数据场的局部微分特征,通过它们建立起来的轮廓树则刻画了三维数据场的整体拓扑和几何.我们采用基于 Carr 算法的改进算法[3]构造轮廓树,过程如下:

按照上一节的判据找出所有奇点并分类对这些奇点根据它们的函数值进行排序建立 JT:按函数值从小到大的规则遍历所有奇点,对每个当前访问到的奇点 v_i 进行以下操作:

(a) 对 $N_-(v_i)$ 的每个分支,从它的一条函数值递减路径进行搜索,直到遇上之前访问过的顶点 w;

(b) 如果 w 还没有与 v 连接,建立它们的连接关系.

建立 ST:按函数值从大到小的顺序遍历所有奇点,其余同第 3 步合并 JT 和 ST 得到轮廓树.

3.3 轮廓树的简化

我们计算了分子 pdb1a07 的 A 链骨架碳原子和分子 1crn 的电子密度场[6]的轮廓树,这两个数据场的分辨率都是 64×64×64 ,得到了以下结果:

蛋白质	顶点数	边数
Pdb1a07	207	206
1crn	2 179	2 178

可以看出来,计算出来的轮廓树很复杂,为了有效分析分子的结构,需要对其进行简化.通过分析,我们提出了 3 种简化方法.

3.3.1 基于数据场的简化

轮廓树过于复杂是三维数据场的数据量大导致的,通过简化三维数据场并重新计算轮廓树可以达到简化的目的.我们选择向下采样的方法来简化数据场,原来数据场的分辨率是 64×64×64 ,通过将其分别重新均匀采样成分辨率是 32×32×32 和 16×16×16 的数据场进行实验,得到如下结果:

(a) pdb1a07A 链骨架碳原子

分辨率	顶点数	边数
64×64×64	207	206
32×32×32	189	188
16×16×16	116	115

(b) 1crn

分辨率	顶点数	边数
64×64×64	2179	2178
32×32×32	1097	1096
16×16×16	398	397

这些结果表明,降低数据场的分辨率可以大幅度的对轮廓树进行简化,其代价是损失了数据场的精确度.

3.3.2 基于点的简化算法

由于误差,数据场中很多周围数据值很小的顶点被当成了奇点,针对这种情况我们提出基于点的简化策略:对轮廓树中的每个节点,如果其数据值小于某一个阈值,就在轮廓树中删除这个顶点和与这个顶点相邻的边,算法描述如下:

for each v in ContourTree:

© 中国科学院软件研究所 http://www.jos.org.cn

```
if val(v) < epsilon:
    for each u in neighbor(v):
        removeEdge(v,u)
    removeVertex(v)
```

我们分别选取所有数据中最大数值的 5%,10%和 15%为阈值进行了实验,实验结果如下:

(a) pdb1a07A 链骨架碳原子

阈值	顶点数	边数
64×64×64	207	206
32×32×32	189	188
16×16×16	116	115

(b) 1crn

阈值	顶点数	边数
64×64×64	2179	2178
32×32×32	1097	1096
16×16×16	398	397

实验结果表明,阈值选取为数据场的最大数据值的 10%比较合适,小于这个阈值的数据一般是误差导致,在不损失精度的情况下,选取超过这个数的阈值并不能取得更有效的简化结果.

3.3.3 基于边的简化算法

轮廓树的很多边长度很短,通过选取一个阈值,把长度小于这个阈值的边收缩为它的某一个端点可以达到简化的目的.数据场的局部极大点和极小点比其他类型的奇点更重要,因此在进行边收缩的时候,如果边的某个顶点是极大点或者极小点,就把它收缩为这个点,否则,把边收缩为它的任何一个顶点.以下是这种简化策略的算法描述:

```
for each edge in ContourTree:
    v1←firstVertex(edge)
    v2←secondVertex(edge)
    if length(edge) < epsilon:
        if isMin(v1) or isMax(v1):
            for each u in neighbor(v2):
                removeEdge(v2,u)
                addEdge(v1,u)
            removeVertex(v2)
            removeEdge(v1,v2)
        elif isMin(v2) or isMax(v2):
            for each u in neighbor(v1):
                removeEdge(v1,u)
                addEdge(v2,u)
            removeVertex(v1)
            removeEdge(v1,v2)
        else:
            for each u in neighbor(v2):
                removeEdge(v2,u)
                addEdge(v1,u)
            removeVertex(v2)
        removeEdge(v1,v2)
```

我们选取三维数据场的包围盒对角线长度的 5%,10%,15%和 20%作为阈值进行实验,实验结果如下:

(a) pdb1a07A 链骨架碳原子

阈值(%)	顶点数	边数
5	160	159
10	114	113
15	83	82
20	64	63

© 中国科学院软件研究所 http://www.jos.org.cn

(b) 1crn

阈值(%)	顶点数	边数
5	1393	1392
10	1175	1174
15	1108	1107
20	1063	1062

结果表明,选取对角线长度的比例为 15%作为阈值,既可以取得比较好的简化效果,又保证了精确度.

3.3.4 实验结果

我们对分子 pdb1a07 的 A 链骨架碳原子的原始数据场(分辨率 64×64×64)的轮廓树首先进行了基于边的简化(取包围盒对角线长度的 15%作为阈值),然后进行基于点的简化(取所有数据最大值的 15%作为阈值),简化后的轮廓树有 54 个顶点,53 条边.接着用 Opengl 绘制了轮廓树和 pdb1a07 的 A 链骨架碳原子的实际原子结构,并从正面,上面和侧面 3 个角度对它们进行观察对比(图 4~图 6).通过观察,我们发现它们在整体形状上非常相似.这是因为研究的数据场是电子密度场,数值越大的地方必然是原子越密集的地方,而这个电子密度场是 pdb1a07 的 A 链骨架碳原子的电子密度场,因此骨架碳原子的整体形状必然反映了它的电子密度场的整体几何信息.轮廓树与蛋白质分子的实际结构吻合得很好,说明轮廓树确实反映了蛋白质分子数据场的整体几何特征,因此我们的方法方面是可行的,可以将其应用到实测数据和动态数据.

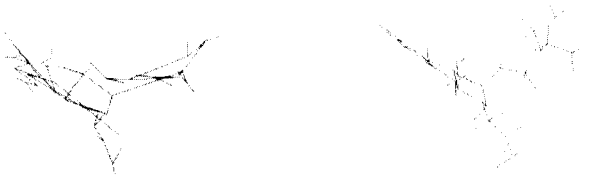

Fig.4 View from front, the left is the contour tree, the right is the protein structure

图 4 正面观察,左边是轮廓树,右边是实际结构

Fig.5 View from up, the left is the contour tree, the right is the protein structure

图 5 上面观察,左边是轮廓树,右边是实际结构

Fig.6 View from profile, the left is the contour tree, the right is the protein structure

图 6 侧面观察,左边是轮廓树,右边是实际结构

3.3.5 与其他工作的比较

本文的工作借鉴了 Masaki Hilaga 等人工作[2]的思想,他们对曲面模型进行了拓扑分析,用 Reeb Graph 来表示曲面的整体拓扑.我们把研究对象推广到三维蛋白质分子电子密度场,利用轮廓树研究它的整体拓扑结构,取得了初步成果,为研究蛋白质分子提供了一个工具.传统的蛋白质分子结构的研究方法侧重于研究蛋白质分子的三维结构,利用 RMSD 等方法[7]比较它们三维结构之间的差异,而实际上蛋白质分子的三维结构是由它所形成的电子密度场决定的,所以与传统方法相比,我们的研究对象对于探索蛋白质分子的结构和功能之间的联系

起着更加直接的作用.

4 结论与展望

本文以 pdb1a07 的 A 链骨架碳原子和 1crn 的电子密度场为例,描述了蛋白质三维数据场的局部特征计算、整体拓扑抽取以及简化的过程.我们以 VC 2003 和 Opengl 为工具,得到并显示可能存在的活性位点,并且建立了它们之间的拓扑关系.

本文利用轮廓树研究了蛋白质的电子密度场,今后我们会对其他实测数据进行研究,而且由于蛋白质分子时刻处于运动当中,对动态数据的特征抽取能更加有效地揭露分子在完成其功能过程中的演化情况,这是我们未来工作的一部分.

References:

[1] Hilaga M, Shinagawa Y. Topology matching for fully automatic similarity estimation of 3d shapes. In: ACM SIGGRAPH. 2001.

[2] Snoeyink H. Computing contour trees in all dimensions. In: Computational Geometry: Theory and Applications 24. 2003. 75–94.

[3] Chiang YJ, Lenz T. Simple and optimal output-sensitive construction of contour trees using monotone paths. Computational Geometry, 2004. 245–256.

[4] Milnor JW. Morse Theory. Princeton University Press, 1963. 4–30.

[5] Hatcher A. Algebraic Topology. Cambridge University Press, 2001. 5–15.

[6] http://ccvweb.csres.utexas.edu/cvs/

[7] Jewett AI, Huang CC, Ferrin TE. MINRMS: An efficient algorithm for determining protein structure similarity using root-mean-squared-distance. Bioinformatics, 2003,19(5):625–634.

张涛(1981－),男,广西柳州人,博士生,主要研究领域为计算机图形学.

陈为(1976－),男,副研究员,主要研究领域为可视化,计算机图形学.

谢立广(1983－),男,博士生,主要研究领域为生物信息学.

胡敏(1965－),女,高级工程师,主要研究领域为科学计算可视化,生物信息学.

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,计算机动画.