一种基于特征点匹配的生物大分子装配方法*1)

路建波
(国家卫生计生委科学技术研究所, 人类遗传资源中心, 北京 100081)
张世华
(中国科学院数学与系统科学研究院, 应用数学研究所, 北京 100190)
马 旭
(国家卫生计生委科学技术研究所, 人类遗传资源中心, 北京 100081)
卢本卓
(LSEC, 中国科学院数学与系统科学研究院, 计算数学与科学工程计算研究所, 北京 100190)

摘 要

研究高分辨率的大分子三维结构, 对于分析研究其功能和生物代谢路径具有很重要作用. 许多大分子结构可以由低温电镜得到, 然而低温电镜三维重构得到的三维结构仍然需要进一步提高其分辨率, 所以这些结构仍然需要具体信息. 大分子的单独组成部分可以通过结晶、NMR、或者蛋白质结构比较等方法得到. 如何将这些高分辨率的组成部分正确装配到大分子结构中是一个很复杂的问题. 该问题的实质可以理解为图像处理. 在生物大分子的装配过程中, 大部分研究者都会用到特征点的方法. 现有的研究方法几乎都是基于几何体内的特征点, 本文提出一种新的基于蛋白质和低温电镜密度图的表面高斯曲率的方法结合优化模型, 对该问题进行研究. 数值实验表明, 我们的方法是有效的.

关键词: 低温电镜; 高斯曲率; 特征点; 生物大分子装配; 优化模型
MR (2000) 主题分类: 65D17

A REPRESENTATION OF FEATURE POINTS IN MACROMOLECULE ASSEMBLY PROBLEM

Lu Jianbo
(*National Center for Human Genetics, National Research Institute for Family Planning, Beijing 100081, China*)
Zhang Shihua
(*Institute of Applied Mathematics, Academy of Mathematics and System Sciences, Chinese Academy of Sciences, Beijing 100190, China*)

* 2017 年 3 月 10 日收到.
1) 基金项目: 国家重点研究发展计划 2016YFC1000307 和 2016YFB0201304, 国家自然科学基金 (21573274), 国家重点研究发展计划子课题 2016YFC1000307-10. 国家卫生计生委科学技术研究所科技创新基金面上项目 (2017GJM04).

Ma Xu
(*National Center for Human Genetics, National Research Institute for Family Planning, Beijing 100081, China*)
Lu Benzhuo
(*LSEC, Institute of Computational Mathematics, Academy of Mathematics and System Sciences, Chinese Academy of Sciences, Beijing 100190, China*)

Abstract

The study of high-resolution three-dimensional macromolecule structure plays an important role in the analysis of the function and metabolic pathways. Many macromolecule structures can be obtained by cryo-EM density maps. But high-resolution three-dimensional structures are still difficult to achieve. On the other hand, The individual components of macromolecules can be obtained by crystallization, NMR, or protein structure comparison. How to properly fit these high-resolution components into macromolecules is a very complicated problem. This problem can be described as image processing problem. In the process of assembly of biological macromolecules, The feature points method is used by most researchers. The existing research methods are always based on the feature points within the geometry, In this paper, we propose a new method which uses the Gaussian curvature of the geometric surface and the optimizing model. Numerical experiments show that our method is effective.

Keywords: cryo-EM; Gaussian curvature; feature points; biomacromolecules assemly; Optimization model

2000 Mathematics Subject Classification: 65D17

1. 引 言

高分辨率结构的大分子在研究细胞和分子的功能和进化上具有重要意义. 最早的研究方法可以追溯到 1895 年伦琴发现 X 射线. 目前生物成像技术包括: X 射线计算机断层成像、放射性核素成像、超声成像、磁共振成像和冷冻电镜三维显微成像等. De Rosier 和 Klug 于 1968 年提出电镜显微三维重构的思想, 随后在 1974 年, Taylor 和 Glaeser 创建了冷冻电镜技术. 1984 年, Dubochet 等发表了第一张病毒的低温电子显微镜照片, 开创了低温电镜的研究时代.

虽然许多大分子结构可以由低温电镜确定, 然而低温电镜三维重构得到的三维结构仍然需要进一步提高其分辨率, 所以这些结构还需要具体信息 [1,2]. 幸运的是原子分辨率的单独部分是可以通过结晶、NMR、或者蛋白质结构比较等方法 [3] 得到, 这些单独的结构部分包括: 结构域, 蛋白质, 蛋白质集合的复合物, 核糖体等等. 把这些高分辨率的部分结构填充到低温电镜的整体密度图中 [4−9], 就可以得到一个完整的生物结构 [10,2] 和更加准确的精细结构, 并给出更深刻的功能解释. 该过程可以通过手动人机交互的方法实现. 然而, 通过自动的计算可以降提高计算准确率和效率 [11,2]. 最近几年, 将高分辨率晶体结构、低温电镜三维重构密度图和计算生物手段相结合的建模方法被发展出来 [12−19].

2. 研究现状和预备知识

Wriggers[20] 等人于 1999 年发表了将高分辨率晶体结构拟合到较低分辨率电镜密度图的 Situs 程序包, Wriggers 的方法是基于向量量化的方法, 即他们用向量量化的方法只用几个顶点代替低温电镜密度图的图像, 同时也用该方法代替高分辨率的蛋白质或者亚基. 之后用特征点代替原来的图像. Grigore 等人 [21] 于 2010 年采用分水岭和尺度空间滤波的方法 (watershed and scale-space filtering). 他们的具体方法如下: 首先用分水岭方法对密度图像进行分割, 然后用尺度空间滤波的方法进行分组. 然后把每一部分填充到密度图中. Hugo 等人 [22] 采用了表面的方法, 但是他们要考虑到各个蛋白质 (或者亚结构) 之间的连接关系. 2015 年, 施一公教授研究组在 science 上发表文章 [23,24], 该论文报道了通过低温电镜技术解析的酵母剪接体近原子分辨率的三维结构, 并且对此结构进行了详细分析, 阐述了剪接体对前体信使 RNA 执行剪接的基本机理. 2016 年, 研究人员首次解析出 HIV 包膜糖蛋白三聚体处于自然状态下的高分辨率结构图 [25]. 同样是在 2016 年, 研究人员解析出转录起始前复合物的高分辨率结构 [26].

Takeshi 于 2008 年提出用高斯混合模型做分子装配问题 [27]. 他的算法如下:1. 将来自低温电子显微镜的密度图通过高斯混合模型转化为点集的高斯分布. 2. 把蛋白质或者亚基结构也通过高斯混合模型转化为点集的高斯分布. 这样两部分都得到点集的高斯分布, 就可以做匹配. 但是他们只能做对称的大分子 – 即要求大分子是对称的结构. Topf 等 [28] 于 2008 年提出一种综合应用蒙特卡洛搜索、共轭梯度最小化、模拟退火分子动力学的方法提出一种新的嵌入和细化 (refine) 方法. Zhang 等人 [29] 在 2010 年用二次规划的方法给出了一种同时将各部分嵌入的方法. 他们的方法首先用向量量化的办法分别把密度图和蛋白质 (或者亚基) 转化为几个特征点, 然后利用二次规划模型. 2015 年毛有东等 [30] 利用低温电镜技术解析了近原子分辨率的炎症复合体的三维结构, 首次阐释了其复合物在免疫信号转导过程中的单向多聚活化的分子结构机理.

在大分子细胞器的分子装配过程中, 大部分研究者都会用到特征点的方法. 本文提出一种基于蛋白质表面高斯曲率的方法结合优化模型定义特征点. 特征点是图像处理中一个关键的概念, 对于图像的三维重构等有很大作用. 这里的特征点是图像沿不同方向上变化较大的局部极值点, 角点之类的 [31]. 在各种变换下特征点会保持良好的几何稳定特征不变. 现有的特征点主要基于三个方面: 图像的几何特征, 图像的灰度变化以及变换域等. 基于图像亮度的变化的主要方法有: Harris 算子、Moravec 算子和 Susan 算子. SIFT 特征点是由 Lowe 提出来的 [32].

下面介绍曲率的几个重要概念: 曲率是刻画曲面弯曲程度的重要概念之一. 曲率的定义是, 在曲面 S 的任意一点 $x(u,v)$ 的切平面上定义一个线性变换 W, 称之为 Weingarten 变换, 使得

$$W(x_u) = -n_u, W(x_v) = -n_v,$$

那么对于切平面上的任何一个向量 $t = ax_u + bx_v$, 有

$$W(t) = -an_u - bn_v.$$

这样定义的变换与曲面参数的选择无关, 并且 Weingarten 变换是自共轭的线性变换, 该变换的两个实特征值 k_1, k_2 称为曲面的主曲率. 曲面的高斯曲率定义为

$$K = k_1 k_2.$$

曲面的平均曲率定义为

$$K = \frac{k_1 + k_2}{2}.$$

3. 研究方法和结果

3.1. 数据的预处理

算法数据的处理均是在 linux 系统下完成的. 首先把高分辨率的蛋白质结构转化为低分辨率的密度图结构. 这一步用的是 Situs 软件包中的 pdb2vol 这个命令, 在转化的时候体素的取值均是 $3\mathring{A}$, 用的核函数均是高斯光滑核函数

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}. \tag{3.1}$$

这里的 $\sigma = 8.826\mathring{A}$. 这里原子转化为正方体网格用的是三线插值. 接下来用 Chimera 软件 [33] 生成结构体等值面的网格. Chimera 软件是一款比较常用的面绘制的可视化软件. 生成网格后保存为 obj 格式. 这个格式的文件包括三部分内容:1 网格的顶点, 2 网格顶点的法向量, 3 三角形网格的信息. 接下来用我们自己写的 shell 脚本语言提取顶点和三角形网格的信息, 分别存放在两个 txt 文件中. 图 1 表示 1CS4 蛋白质的 “表面网格” , 这里的表面网格指的是蛋白质用 Situs 软件转化为低温电镜的密度图, 在转化的时候为了计算方便取体素为 3, 为了与实际蛋白质保持大小一致所选择的等值面由多次试验确定.

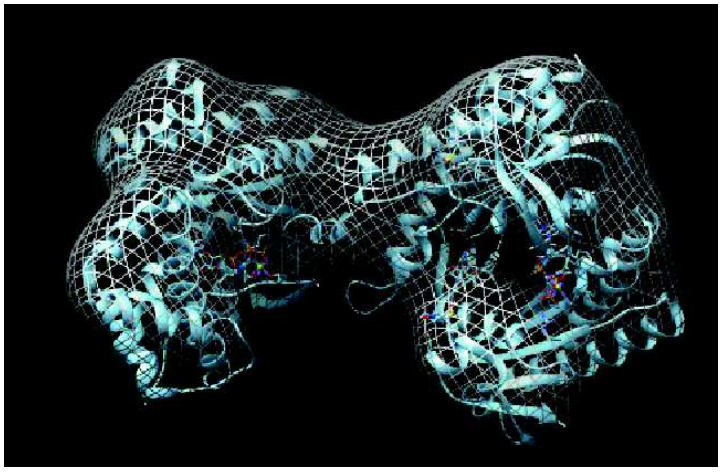


图 1 蛋白质对应的低温电镜密度图经过光滑以后的表面等值面网格

3.2. 不规则几何体表面的曲率计算

高斯曲率的离散化方法源于 Gauss-Bonnet 定理, 三角形网格上的高斯曲率离散化如下:

不规则曲面在点 P 处的高斯曲率可以由公式 [34,35,36]

$$K_G(P) = \frac{3(2\pi - \sum \theta_i)}{A(x_i)} \tag{3.2}$$

计算得到.

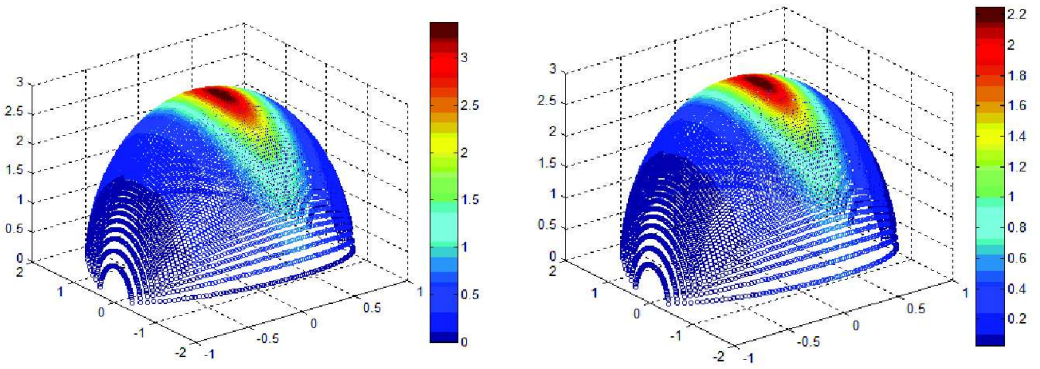
这里的 θ_i 是角度 $(\angle Q_i P Q_{i+1})$, $A(x_i)$ 是围绕点 P 的表面的三角形网格的面积总和, Q_i, Q_{i+1} 是三角形网格的两个顶点.

对于曲面的特殊参数表示 $z = z(x, y)$, 高斯曲率的准确计算公式是:

$$K = \frac{rt - s^2}{(1 + p^2 + q^2)^2}. \tag{3.3}$$

其中 $p = \frac{\partial z}{\partial x}, q = \frac{\partial z}{\partial y}, r = \frac{\partial^2 z}{\partial x^2}, s = \frac{\partial^2 z}{\partial x \partial y}, t = \frac{\partial^2 z}{\partial y^2}$.

本文中曲率如果没有说到高斯曲率还是平均曲率, 则曲率指的是高斯曲率.



a b

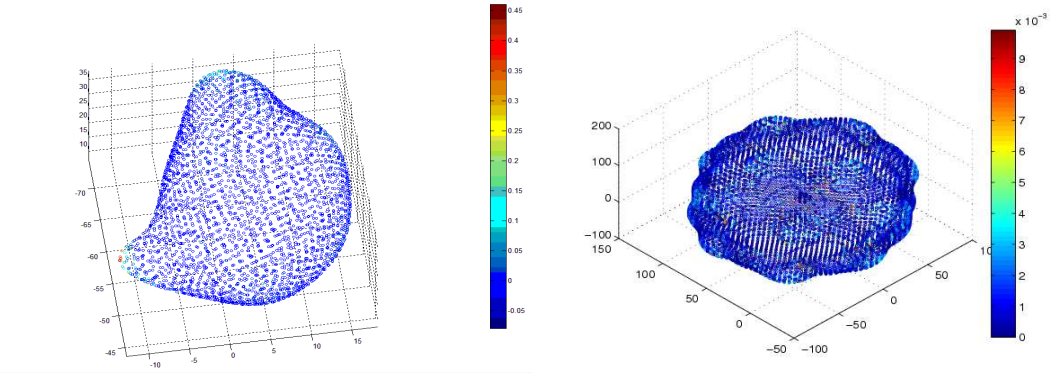
图 2 我们的算法求出的曲率和解析曲率的比较图. (a) 用算法公式 (3.2) 算的椭球体 $x^2 + \frac{y^2}{4} + \frac{z^2}{9} = 1$ 的离散高斯曲率分布图 (b) 是椭球体 $x^2 + \frac{y^2}{4} + \frac{z^2}{9} = 1$ 的解析 (公式 (3.3)) 高斯曲率分布图

图 2a 和图 2b, 说明了用数值计算的曲率和解析的曲率的差别. 图 2a 是用数值的方法计算的椭球体 $x^2 + \frac{y^2}{4} + \frac{z^2}{9} = 1$ 的高斯曲率分布, 图 2b 是用解析的方法计算的椭球体 $x^2 + \frac{y^2}{4} + \frac{z^2}{9} = 1$ 的高斯曲率的分布图. 从这两幅图可以看到: 数值计算和解析计算的曲率的分布规律是一样的. 但是高斯曲率的具体数值有一定差别, 用数值的计算方法得出来的结果偏大. 这可能和误差的存在有关系. 经过分析我们发现用数值计算的方法得到的高斯曲率和准确高斯曲率的差别在两倍以内.

3.3. 低温电镜得到的密度图的表面的曲率计算

由于用 Chimera 得到的网格的噪音很大, 例如: 有的三角形单元内的三条边的差距很大. 这使得高斯曲率将会非常大, 这里当出现这种情况就令其曲率为 0. 这样处理曲率的计算仍然有噪音, 为了减小噪音可以采用如下方法:

首先计算几何体表面的每一个点的曲率, 然后按照递减顺序排序. 取出前 rn 个顶点. 然后搜出每一个顶点的球形领域的点, 半径一般取 $r = 5$, 这个数据来源于多次试验. 这时得到一个矩阵 C, $C_{i,j}$ 是顶点在搜索范围内的顶点编号. 同时, 得到与之对应的高斯曲率的矩阵不妨记做 G, $G_{i,j}$ 是对应顶点的曲率. 在这一步为了避免重复计算, 把同一个顶点出现在不同的邻域的情况保留一种情况, 其他的都从矩阵中删掉. 把每个顶点的邻域的所有顶点的曲率求平均值然后赋给点 P. 图 3 是计算的低温电镜下的几何体密度图的曲率分布.



a b

图 3 生物大分子低温电镜密度图的高斯曲率分布图. (a) 蛋白质 1d1r 的低温电镜密度图的曲率分布 (b) 蛋白质 1xck 复合体等值面低温电镜密度图的曲率分布

3.4. 特征点的选取

特征点的选择以及特征点的准确定位非常重要. 这里的特征点是高斯曲率比较大的能代表明显几何结构的点. 例如: 侧链比较长的氨基酸大多数时候可以代表特征点. 由于曲面光滑化以后, 高斯曲率算出来都在 0 附近, 一般高斯曲率的绝对值都小于 1, 这样导致如果特征点仅仅用曲率去度量是不准确的. 如果再考虑几何中心以及以几何中心为球心的球的最外面的点, 那么就更加准确一些. 综上, 特征点的选择步骤如下几步:

步骤一, 计算几何中心.

步骤二, 以几何中心为球心, 以适当值 (具体实现是试出来的) 为半径做球. 在球的外面并且曲率又比较大的点就是特征点.

步骤三, 两个几何体都取前 5 个距离比较大而且又在曲率的大的范围内的顶点.

3.5. 特征点的匹配

为了量化两个数据集合 – 高分辨率的蛋白质数据和由低温电镜得到的密度图, 得到两个集合 w_i^{pr} , 其中 $i \in \{1, \ldots, N\}$ 和 w_j^{em}, 其中 $j \in \{1, \ldots, M\}$, 这里特征点的匹配实际上是找一个最优指标映射 $I : i \to j$ 和一个使 RMSD 最小的变换. 对于一个给定的集合 I, 最优的变换可以由旋转矩阵 $R(I)$ 和平移向量 $t(I)$ 通过最小二乘法得到, 然而这样会有 $\frac{M!}{(M-N)!}$ 多种指

标映射 I, 即使是很小的点云集合, 计算量也是很大的, 不可接受. 实际上三对特征点匹配就足够确定刚体运动的六个自由度. 接下来给出这三对特征点的匹配的方法.

这里要同时考虑曲率和距离, 具体方法如下: 首先在集合 w_i^{pr} 上选出特征点, 几何中心 O_1^{pr} 就是一个特征点, 接着选出曲率比较大前 n 个 (这里的 n 一般取 5), 并计算每一个特征点到几何中心 O_1^{pr} 的距离, 然后存储下来记为 M_i. 对这个距离按照从大到小的顺序排序, 最后从这 n 个特征点中选出 2 个点 (前两个), 不妨记做 s_1 和 t_1. 下面是在集合 w_j^{em} 上找对应的特征点, 第一个特征点是几何中心 O_2^{em}. 然后计算集合 w_j^{em} 中每一个特征点到几何中心 O_2^{em} 的距离, 并保存 N_j. 对应于点 M_1 的点在如下的公式里:

$$\min \quad d_{1j} = |m_1 - n_j| + |g(m_1) - g(n_j)| \tag{3.4}$$

$$s.t. \qquad |m_1 - n_j| \le 2 \tag{3.5}$$

$$|g(m_1) - g(n_j)| < 0.05 \tag{3.6}$$

$$n_i \in N \tag{3.7}$$

$$g(n_j) \in G \tag{3.8}$$

这里 m_1 是蛋白质上的特征点, n_j 是低温电镜密度图的特征点, $g(m_1)$ 是蛋白质上的特征点对应的高斯曲率, $g(n_1)$ 是低温电镜密度图上的特征点对应的高斯曲率, N 是正整数集合, G 是特征点对应的曲率的集合. 同时满足上面的条件求出来的与 M_1 匹配的点一般就有两三个, 记满足这些条件的集合为 S_1^{em}, 选择距离几何中心比较远的顶点作为特征点. 同样与 M_2 对应的顶点也是这么求的, 我们记满足这些条件的集合为 T_1^{em}. 这时候 S_1^{em} 和 T_1^{em} 中的顶点不唯一. 我们计算 S_1^{em} 中每一个顶点到 T_1^{em} 中每一个顶点的距离, 这样我们得到一个有关 S_1^{em} 和 T_1^{em} 的距离矩阵 ST, 这个矩阵的行和列分别为 $|S_1^{em}|$ 和 $|T_1^{em}|$. 此时, 计算几何中心 O_1^{pr} 到 O_2^{em} 的距离记为 $d_{o_1o_2}$. 接下来在矩阵 ST 中找与 $d_{o_1o_2}$ 最接近的元素.

$$\min \qquad |d_{o_1o_2} - ST_{i,j}| \tag{3.9}$$

$$s.t. \quad i \in \{1, 2, \ldots, |S_1^{em}|\} \tag{3.10}$$

$$j \in \{1, 2, \ldots, |T_1^{em}|\} \tag{3.11}$$

$d_{o_1o_2}$ 是计算几何中心 O_1^{pr} 到 O_2^{em} 的距离, $ST_{i,j}$ 是矩阵 ST 的第 i 行第 j 列的元素, $|S_1^{em}|$ 是集合 S_1^{em} 的元素个数, $|T_1^{em}|$ 是集合 T_1^{em} 的元素个数. 我们把找到的与 s_1 和 t_2 匹配点记做 s_2 和 t_2. 这样两组对应顶点就找到了, 它们分别是 $\{O_1, s_1, t_1\}$ 和 $\{O_2, s_2, t_2\}$.

3.6. 特征点的变换和几何体的变换

由上一节我们得到的 $w_i^{pr}, i \in \{1, \ldots, N\}$ 和 $w_j^{em}, j \in \{1, \ldots, M\}$ 的两组对应特征点, 它们分别是 $\{O_1, s_1, t_1\}$ 和 $\{O_2, s_2, t_2\}$, 我们利用下面的方法 (3.7), 求出对应特征点的旋转矩阵 $R(I)$ 和平移向量 $t(I)$. 然后再将 $R(I)$ 和 $t(I)$ 作用到原来的几何体.

3.7. 旋转矩阵和平移向量的求解

设 $X_1, X_2, \ldots, X_n$ 和 $Y_1, Y_2, \ldots, Y_n$ 分别是 X 和 Y 对应蛋白质的原子的三维坐标. 在下

面的目标函数中, R 是一个 3×3 的合适的旋转矩阵, T 是平移向量 [37]. $s_{i,j}$ 分配变量.

$$\min \quad \sum_{i=1}^{n_x}\sum_{j=1}^{n_y} s_{i,j}|X_i - R(Y_i - T)|^2 \tag{3.12}$$

$$s.t. \quad \sum_{i=1}^{n_x} s_{i,j} \le 1 \quad for \quad j = 1,\ldots,n_y, \tag{3.13}$$

$$\sum_{i=1}^{n_y} s_{i,j} \le 1 \quad for \quad j = 1,\ldots,n_x, \tag{3.14}$$

$$s_{i,j} \in \{0,1\}. \tag{3.15}$$

定义 $\overline{X}$ 和 $\overline{Y}$ 为:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{3.16}$$

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i. \tag{3.17}$$

在公式 (3.12) 中, $T = \overline{Y} - R^t\overline{X}$. 旋转矩阵 R 可以用欧拉角 ϕ,θ 和 ψ[38], 表示如下:

$$R = \begin{pmatrix} \cos\psi\cos\phi - \cos\theta\sin\phi\sin\psi & \cos\psi\sin\phi + \cos\theta\cos\phi\sin\psi & \sin\psi\sin\theta \\ -\sin\psi\cos\phi - \cos\theta\sin\phi\cos\psi & -\sin\psi\sin\phi + \cos\theta\cos\phi\cos\psi & \cos\psi\sin\theta \\ \sin\theta\sin\phi & -\sin\theta\cos\phi & \cos\theta \end{pmatrix}. \tag{3.18}$$

由于旋转矩阵是非线性和非对称的, 由文献 [39] 我们可以通过如下表达式定义四元数参数 ξ,η,ζ,χ :

$$\xi = \sin\frac{\theta}{2}\sin\frac{\psi-\phi}{2}, \tag{3.19}$$

$$\eta = \sin\frac{\theta}{2}\cos\frac{\psi-\phi}{2}, \tag{3.20}$$

$$\zeta = \cos\frac{\theta}{2}\sin\frac{\psi+\phi}{2}, \tag{3.21}$$

$$\chi = \cos\frac{\theta}{2}\cos\frac{\psi+\phi}{2}. \tag{3.22}$$

旋转矩阵 R 可以用四元数表示如下:

$$R = \begin{pmatrix} -\xi^2+\eta^2-\zeta^2+\chi^2 & 2(\xi\chi-\zeta\eta) & 2(\eta\zeta+\xi\chi) \\ -2(\xi\eta+\zeta\chi) & \xi^2-\eta^2-\zeta^2+\chi^2 & 2(\eta\chi-\xi\zeta) \\ 2(\eta\zeta-\xi\chi) & -2(\xi\zeta+\eta\chi) & -\xi^2-\eta^2+\zeta^2+\chi^2 \end{pmatrix}. \tag{3.23}$$

实际上, 四元数并不是独立的它们有如下的关系:

$$\xi^2+\eta^2+\zeta^2+\chi^2 = 1. \tag{3.24}$$

定义如下协方差矩阵 C:

$$C = \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})^t. \tag{3.25}$$

目标是最小化目标函数. 这就需要计算合适的 R 和 t 使目标函数最小. 通过合适的计算可以得到 $(\xi, \eta, \zeta, \chi)^t$ 是矩阵 P 的最大特征值的特征向量. 这里 p 是 4×4 的对称矩阵如下:

$$P = \begin{pmatrix} -C_{11} + C_{22} - C_{33} & -C_{12} - C_{21} & -C_{23} - C_{32} & C_{13} - C_{31} \\ -C_{12} - C_{21} & C_{11} - C_{22} - C_{33} & C_{13} + C_{31} & C_{23} - C_{32} \\ -C_{23} - C_{32} & C_{13} + C_{31} & -C_{11} - C_{22} + C_{33} & C_{12} - C_{21} \\ C_{13} - C_{31} & C_{23} - C_{32} & C_{12} - C_{21} & C_{11} + C_{22} + C_{33} \end{pmatrix}. \tag{3.26}$$

这里的 $C_{i,j}$ 是矩阵 C 的第 i 行, 第 j 列的元素. 我们可以用矩阵 SVD 分解求最大的特征值和特征向量.

3.8. 模拟结果和真实结果

本文用模拟数据 (2REC 的一条链) 和真实数据 (emd-1005) 验证方法的准确性. 模拟数据结果如图 4 所示: 黄色点和蓝色点表示的分别是 2REC 一条链的蛋白质和对应的低温电镜密度图, 红色点表示的是经过旋转平移后的图像. 通过比较发现旋转和平移后的图像与目标图像非常吻合.

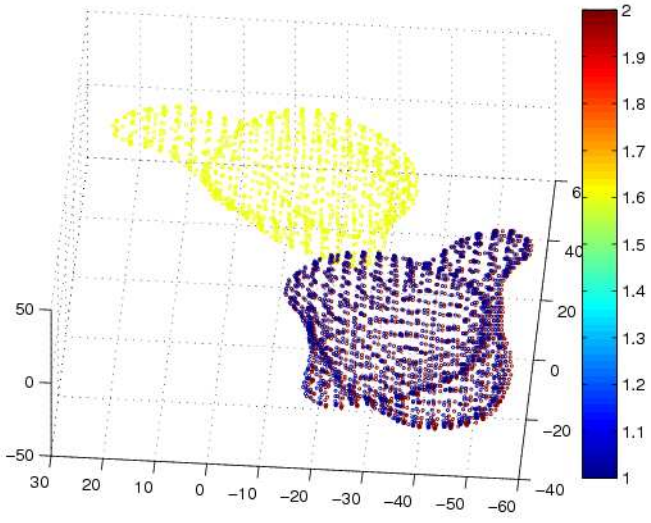
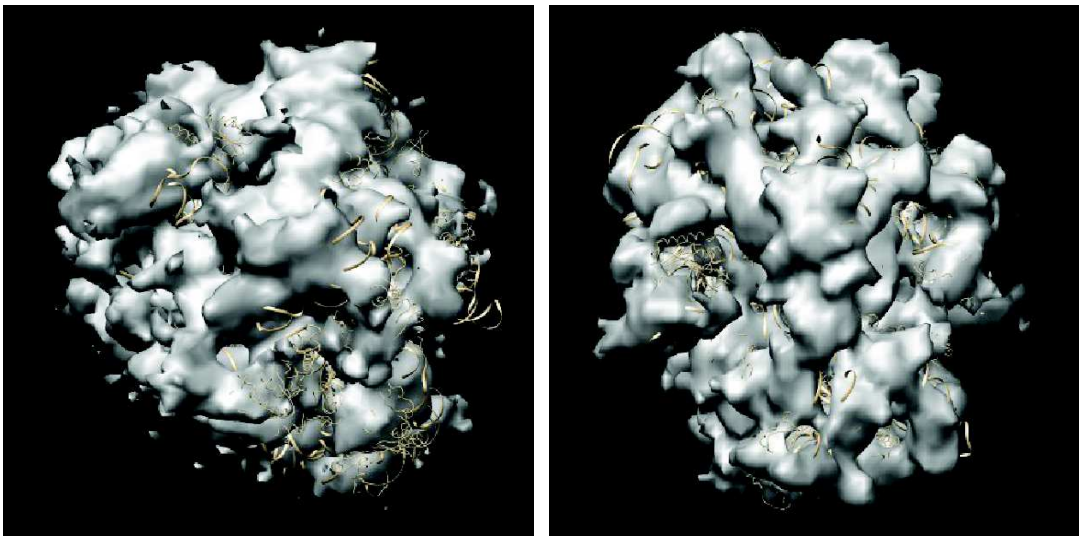


图 4 用该方法得到的数据结果: 黄色点和蓝色点表示的分别是蛋白质 2REC 的一条链和对应的低温电镜得到的密度图, 红色点表示的是经过旋转平移后的图像

真实数据是: 我们从 EMDB 数据库中挑选了一种研究常用的低温电镜密度图 emd-1005. emd-1005 是 $E.coli70S$ 分辨率为 $14\mathring{A}$ 的核糖体. 这个结构在 PDB 数据库中的蛋白质编号为 1ML5.pdb. 图 5 是用我们的方法计算的结果和真实结果比较的可视化显示. 其中图 a 是用我们的方法计算结果, 图 b 是真实的结果. 从图上可以看出蛋白质结构并没有完全嵌入到其相应的低温电镜密度图中, 存在一定误差. 由于生物大分子嵌入到低温电镜密度图是一个很复杂的问题, 涉及到一系列的计算和模型, 尤其是特征点的选择和匹配问题. 基于几何体表面曲率的选择定义的特征点, 为该问题打开了基于几何体表面嵌入的研究方法. 我们相信经过不断努力和改进我们的方法, 精度一定会提高.



a b

图 5 这两幅图是我们的计算结果和真实结果的比较. 其中图 a 和图 b 是从 EMDB 数据库中中提取的 emd-1005, 里面的蛋白质是 1ML5.pdb. 图 a 是我们的计算结果, b 是准确结果. 可以看出结果相差不大.

4. 结　　论

本文我们提出一种新的基于特征点的生物大分子装配方法, 该方法基于蛋白质或者低温电镜密度图的表面曲率以及特征点之间的距离等几何体的固有特征. 生物大分子的三维重构问题本身比较困难, 过程复杂. 我们通过计算低温电镜密度图的表面高斯曲率, 用曲率和距离来表示刻画蛋白质和低温电镜密度图的特征点. 在分别找到特征点之后, 利用建立的优化模型来匹配蛋白质和低温电镜密度图的特征点. 得到匹配的特征点以后, 用基于最小二乘的方法来旋转和平移特征点. 进而得到旋转矩阵和平移向量, 将该矩阵和向量作用到原来的蛋白质或者其对应的低温电镜密度图, 最后得到完美的大分子装配结果. 数值实验表明该方法的整个过程基本准确有效.

参 考 文 献

[1] Jiang W, Ludtke S J. Electron cryomicroscopy of single particles at subnanometer resolution[J]. Curr Opin Struct Biol, 2005, 15(5), 571-577.

[2] Chiu W, Baker M L, Jiang W, et al. Electron cryomicroscopy of biological machines at subnanometer resolution[J]. Structure, 2005, 13, 363-372.

[3] Eswar N, Webb B, Marti-Renom, M A, et al. Comparative protein structure modeling with modeller[J]. Curr. Protoc. Protein Sci, 2007, Nov; Chapter 2: Unit 2.9. doi: 10.1002/0471140864.ps0209s50.

[4] Zhang K, Zhang Y, Hu Z J, et al. Development and frontier of electron microscopy 3d reconstruction[J]. Acta Biophysica Sinica, 2010, 26(7), 533-559.

[5] Gabashvili I S, Agrawal R K, Spahn C M, et al. Solution structure of the e. coli 70s ribosome at 11.5×A resolution[J]. Cell, 2000, 100(5), 537-549.

[6] Fabian K, Mareike K, Markus G G, et al. Projection structure of the secondary citrate/sodium symporter cits at 6×A resolution by electron crystallography[J]. Journal of Molecular Biology, 2012, 418, 117-126.

[7] Henderson R. The potential and limitations of neutrons, electrons and x-rays for atomic resolution microscopy of unstained biological molecules[J]. Q Rev Biophys, 1995, 28(2), 171-193.

[8] Wang D N, Werner K. High-resolution electron crystallography of light-harvesting chlorophyll a/b-protein complex in three different media[J]. Journal of Molecular Biology, 1991, 217(4), 691-699.

[9] Stefan B, Willy W. Multi-resolution anchor-point registration of biomolecular assemblies and their components[J]. Journal of Structural Biology, 2007, 157, 271-280.

[10] Rossmann M G, Morais M C, Leiman P G, et al. Combining x-ray crystallography and electron microscopy[J]. Structure, 2005, 13, 355-362.

[11] Fabiola F, Chapman M S. Fitting of high-resolution structures into electron microscopy reconstruction images[J]. Structure, 2005, 13, 389-400.

[12] Sali A, Glaeser R, Earnest T, et al. From words to literature in structural proteomics[J]. Nature, 2003, 422(6928), 216-225.

[13] Frank J. Single-particle imaging of macromolecules by cryo-electron microscopy[J]. Annu Rev Biophys Biomol Struct, 2002, 31, 303-319.

[14] Wriggers W, Birmanns S. Using situs for flexible and rigid-body fitting of multiresolution single-molecule data[J]. J Struct Biol, 2001, 133(3), 193-202.

[15] Topf M, Sali A. Combining electron microscopy and comparative protein structure modeling[J]. Curr. Opin. Struct. Biol, 2005, 15, 578-585.

[16] Alber F, Dokudovskaya S, Veenhoff L M, et al. Determining the architectures of macromolecular assemblies[J]. Nature, 2007, 450(7170), 683-694.

[17] Robert M G, Kenneth H D. High-resolution electron crystallography of protein molecules[J]. Ultramicroscopy, 1993, 52, 478-486.

[18] Dempster A P, Laird N M, Rubin DB. Sur la division des corps materiels en parties[J]. Bull. Acad. Polon. Sci, 1957, 4(12), 801-804.

[19] MacKay, David. Chapter 20. an example inference task: Clustering. In: Information Theory, Inference and Learning Algorithms

[20] Wriggers W, Milligan R A, McCammon J A. Situs: A package for docking crystal structures into

low-resolution maps from electron microscopy[J]. J Struct Biol, 1999, 125(2), 185-195.

[21] Pintilie G D, Zhang J, Goddard T D, et al. Quantitative analysis of cryo-em density map segmentation by watershed and scale-space filtering, and fitting of structures by alignment to regions[J]. Journal of Structural Biology, 2010, 170, 427-438.

[22] Hugo C, Robert B R. Fast fitting of atomic structures to low-resolution electron density maps by surface overlap maximization[J]. J. Mol. Biol. 2004, 338, 783-793.

[23] Yan C, Hang J, Wan R, et al. Structure of a yeast spliceosome at 3.6-angstrom resolution[J]. Science, 2015, 349(6253): 1182-1191.

[24] Hang J, Wan R, Yan C, et al. Structural basis of pre-mRNA splicing[J]. Science, 2015, 349(6253):1191-1198.

[25] Lee J H, Ozorowski G, Ward A B. Cryo-EM structure of a native, fully glycosylated, cleaved HIV-1 envelope trimer. Science, 2016, 351(6277):1043-1048.

[26] Louder R K, He Y, L ó pez-Blanco J R, et al. Structure of promoter-bound TFIID and model of human pre-initiation complex assembly. Nature, 2016, 531(7596):604-609.

[27] Kawabata T. Multiple subunit fitting into a low-resolution density map of a macromolecular complex using a gaussian mixture model[J]. Biophysical Journal, 2008, 95, 4643-4658.

[28] Topf M, Lasker K, Webb B, et al. Protein structure fitting and refinement guided by cryo-em density[J]. Structure, 2008, 16, 295-307.

[29] Zhang S H, Daveb V, Min X, et al. A fast mathematical programming programming procedure for simultaneous fitting of assembly components into cryoem density maps[J]. Bioinformatics, 2010, 26, 261-268.

[30] Zhang L, Chen S, Ruan J, et al. Cryo-EM structure of the activated NAIP2-NLRC4 inflammasome reveals nucleated polymerization. Science, 2015, 350(6259):404-409.

[31] Lowe D G. Distinctive Image Features from Scale-Invariant Key points[J]. International Journal of Computer Vision, 2004, 60(2),91-110.

[32] Ji Z, shi J. A robust algorithm for feature point matching[J]. Computers and Graphics, 2002, 26, 429-436.

[33] Goddard TD, Huang CC, and Ferrin TE. Visualizing density maps with ucsf chimera[J]. J. Struct. Biol., 2007, 157, 281–287.

[34] Zhang W, Kaufmann B, Chipman P R, et al. Membrane curvature in flaviviruses[J]. Journal of Structural Biology, 2013, 183, 86-94.

[35] 徐国良, 计算几何中的几何偏微分方程方法 [M], 科学出版社, 北京, 2008.

[36] Zhang W, Barbel K, Paul R C, et al. Membrane curvature if flaviviruses. Journal of Structural Biology[J], 2013, 183, 86-94.

[37] Lu J, Xu G, Zhang S, et al. An effective sequence-alignment-free superpositioning of pairwise or multiple structures with missing data[J]. Algorithms Mol Biol, 2016, 11(18), 1-10.

[38] Goldstein H, Classical Mechanics[M], Addison-Wesley, 1965.

[39] Evans DJ. On the representation of orientation space[J]. Molec.Phys., 1977, 34, 317–325.