蛋白质结构比较与电镜图像装配问题的研究

路建波

摘要

蛋白质是生命活动的载体，计算蛋白质间的结构相似性是结构生物学中的一个重要问题，在蛋白质折叠结构分类、蛋白质折叠功能推断、蛋白质结构预测和模拟、结构与功能之间的进化关系等研究中有重要应用。另一方面，由于目前低温电镜所得到的生物大分子的三维结构还难以达到原子分辨率。因此，将高分辨率晶体结构、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，这些方法以高分辨原子结构为基础，同时以电镜密度图为约束，保证了所搭建的原子模型的可靠性，极大地拓展了人们从电镜密度图中所获得的信息。本论文主要内容有以下几个部分：

1. 首先简要介绍了本文所涉及到的一些基本知识。重点介绍了蛋白质结构比较叠加和生物大分子装配的研究现状。

2. 然后给出了一种缺失数据下的蛋白质结构比较的迭代算法，该算法对于缺失数据下的蛋白质结构比较是有效的。我们的算法比首先提出此问题的 THESEUS 方法快，后者的实质是 EM 算法。而 EM 算法收敛速度慢而且对初始值依赖比较大。

3. 接着提出了一种有效的两结构叠加和多结构叠加的算法，算法采用主成分分析 (PCA) 求初值，接着用最近点迭代 (ICP) 求准确的旋转与平移。数值试验表明我们的方法有如下的优点：(1) 不需要蛋白质序列一一对应。(2) 运算结果比较准确。(3) 大多数情况下运算的时间少。(4) 可以处理缺失数据的蛋白质叠加问题。(5) 给出了算法所能容忍的缺失数据量。据我们所知：这是第一个不需要序列比对的缺失数据下的多结构叠加方法。

4. 最后给出了一种有效的低温电镜下生物大分子装配的新方法。该方法是基于几何体表面的曲率计算的。我们分两种情况处理这个问题。一种是假设在装配前各个子部分已经结合好了，只需做一个整体嵌入。数值实验表明：我们对于整体的嵌入算法运行的时间比较短，结果嵌入准确。另一种情况是将蛋白质或者亚基部分地嵌入到低温电镜对称的密度图中。

由于我们的方法要用到 K-means 分类，所以我们目前能做的是对称的生物大分子。

数值结果显示我们的算法对于某些简单的蛋白质嵌入能得到准确的结果。