doi:10.3772/j.issn.1002-0470.2020.03.001

基于小批量梯度下降和 Spark 分布式方法的局部断层细化对齐①

吕永春②[*,**,***] 赵晓芳[*,**] 李 华[*,**,***] 曾祥睿[****] 徐 旻[****]

([*]中国科学院计算技术研究所 北京 100190)

([**]中国科学院大学 北京 100190)

([***]中国科学院智能信息处理重点实验室 北京 100190)

([****]美国卡内基梅隆大学计算机科学学院 匹兹堡 15213)

摘 要 生物样品在获取电子冷冻断层扫描(cryo-ET)图像时的辐射损伤,信息缺失和低信号噪声比(SNR),限制了从断层数据中恢复3维结构信息。仿照电子低温显微镜(cryo-EM)单颗粒3维重构技术,对局部断层数据进行对齐和平均,产生高精度大分子复合体的3维结构。现有的局部断层对齐技术都会涉及6个自由度(3个旋转参数、3个平移参数),因此,它们在每次迭代对齐中处理整个3维体积图像来计算这6个自由度,这是计算密集型的。针对上述问题,本文提出一种基于小批量梯度下降(MBGD)方法实现局部断层3维数据细化对齐,并首次利用 Spark 分布式框架实现局部断层对齐全局择优。通过对仿真数据和实验数据的对齐,基于小批量梯度下降细化对齐算法与基线方法相比,实现了对齐精度和速度的提高。

关键词 小批量梯度下降(MBGD);Spark;局部断层细化对齐;电子冷冻断层扫描技术(cryo-ET)

0 引 言

近年来,电子低温显微镜技术(electron cryomicroscopy,cryo-EM)在大分子复合体的3维结构恢复中发挥了越来越重要的作用。该技术促进了大分子和复合体的3维结构进一步被研究,揭示其功能可能在生物细胞机理、制药、疾病治疗等方面产生重大的突破。电子冷冻断层扫描(electron cryotomography, cryo-ET)技术是电子低温显微镜技术(cryo-EM)的一种应用。电子冷冻断层扫描技术可以实现生物大分子和细胞亚纳米分辨率(1 ~4 nm)的3维成像(3维密度图像)。Cryo-ET 技术类似人体的 CT 扫描,由于受机械臂的限制,对样品进行有角度采样时,通常只能旋转到 ±70 °,导致部分角度无法采样,使得重构的3维图像在傅里叶空间存在部分角度范围信息的缺失。另外,为了保证电子束不损伤样品,成像过程中会使用较低剂量的电子,这样就会造成噪声很大,使得信号和噪声的比率(signal-to-noise ratio,SNR)很小,产生的原始2维图像很模糊,也影响3维重构的分辨率。为了提高 cryo-ET 技术重构的分辨率,局部断层平均(subtomogram averaging, SA)技术被使用以提高3维重构的信噪比,从而得到高分辨的大分子复合体3维结构。

局部断层平均技术需要对3维颗粒进行旋转和平移,最小化3维颗粒与参考颗粒间不相似值,实现与参考颗粒的对齐。在局部断层3维对齐过程中,每个局部断层3维颗粒对齐都包含6个参数(3个旋转参数,3个平移参数),计算量很大。为减少局部断层3维对齐的计算量,Kovacs 等人[1,2]在傅里叶空间利用球谐函数实现快速旋转匹配,把3维数

① 国家重点研发计划(2018YFB0904503,2017YFB1002703)和国家自然科学基金(61379082 ,61672499)资助项目。

② 男,1984 年生,博士生;研究方向:分布式计算,机器学习,三维图像重构和生物图像处理等;联系人,E-mail: lvyongchun@ncic.ac.cn

(收稿日期:2019-03-21)

据转换到 2 维空间进行旋转匹配,得到旋转参数,利用傅里叶空间性质,得到平移参数。Xu 等人[3]改进了局部断层快速旋转对齐方法,使用傅里叶空间的 3 维数据进行旋转参数计算,然后计算平移参数,得到 6 个参数更加准确,但这样局部断层对齐是粗粒度的,为了得到全局择优的旋转和平移参数,需要对局部断层进行细化对齐。Xu 等人[4]提出利用莱文贝格-马夸特方法(Levenberg-Marquardt algorithm)实现上述 6 个参数的更新,实现局部断层的精对齐。虽然 Xu 的方法是较早提出的局部断层细化对齐方法,但该方法每次计算会处理整个 3 维体,导致计算量较大。为减少计算量,本文提出利用小批量梯度下降算法(mini-batch gradient descent, MBGD)实现局部断层 3 维颗粒细化对齐,得到局部最优解;同时引入 Spark 分布式框架进行局部断层细化对齐,利用 Spark 进行候选集的分布式并行计算,得到全局择优旋转和平移参数,实现局部断层的对齐。通过对仿真数据和实验数据进行对齐,本文提出的基于 MBGD 和 Spark 的细化对齐方法比现有的基线对齐方法,具有明显的优势,不仅计算速度快,同时对齐精度也有所提高。

本文的主要贡献如下:(1) 提出一种 3 维版本的小批量梯度下降算法,并应用到局部断层 3 维对齐过程中。(2) 利用小批量梯度下降算法进行局部断层 3 维颗粒对齐的优化,减少局部断层 3 维对齐的计算量。(3) 首次实现 Spark 分布式局部断层细化对齐,利用 Spark 进行候选参数的分布式计算,实现较快得到最佳对齐参数。

本文第 1 节对相关研究进行介绍。第 2 节对局部断层对齐的相关工作进行介绍,提出基于小批量梯度下降的局部断层对齐算法,并对算法进行详细介绍,利用小批量梯度下降细化对齐算法实现局部断层对齐,利用 Spark 框架实现对候选集的分布式并行计算,得到全局择优参数。第 3 节介绍实验数据,实验环境,并从对齐精度和速度两方面对 2 种对齐算法进行比较。最后总结全文。

1　相关研究

局部断层对齐是局部断层平均的前提和关键步骤,当所有 3 维颗粒都与参考颗粒进行对齐,才能对所有 3 维颗粒进行平均,从而提高局部断层平均结构的信噪比,得到高分辨率 3 维结构。

1.1　电子冷冻断层扫描技术和局部断层平均技术

电子冷冻断层扫描(cryo-ET)技术是从一个物体的投影图像重构获得物体内部结构的技术,通过获取同一物体的多个连续角度下的 2 维投影图来反向重构它的 3 维结构。与医院中使用的 CT 扫描类似,简单地说,电子断层扫描技术就是将一个样品(物体)沿着一个与电子束垂直的轴旋转,每旋转一个角度,采集这个物体在相对应方向上的 2 维投影图,通过对这些 2 维投影图的处理(如图像配准、滤波等),然后将处理过的不同角度 2 维投影图利用反向重构技术进行重构,如加权背投影(weighted back projection,WBP)等算法,获得样品整体 3 维结构的技术(见图 1)。

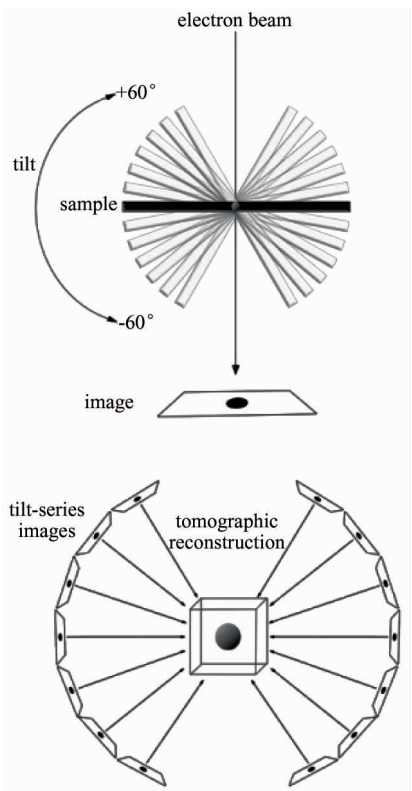

图 1　电子冷冻断层扫描技术成像示意图[5]

电子冷冻断层扫描技术研究对象包括不具有均一性的蛋白、病毒(如包膜病毒)、细胞器、细胞等,因此电子冷冻电镜断层扫描技术也是目前唯一能够研究原位生物信息的强有力工具。

电子冷冻断层扫描技术的优点在于旋转角度参数已知,不需要重新求解,但存在缺失锥(missing wedge)、单张信噪比低、不同投影角度间存在非均匀信噪比等问题,其中缺失锥是电子冷冻断层扫描技术面临的最大问题。为了解决上述问题,需要使用局部断层平均技术。

局部断层平均技术是将单颗粒 3 维重构与电子断层成像技术相结合的技术。它的基本思路是:先对同构象的分子颗粒做电子断层重构,再将这些同构象的分子颗粒从重构体中挑选出来,进行类似单颗粒分析技术的对齐、分类、平均等步骤,这样消除了缺失锥,提高了 3 维颗粒的信噪比,从而实现高分辨的 3 维结构(见图 2)。

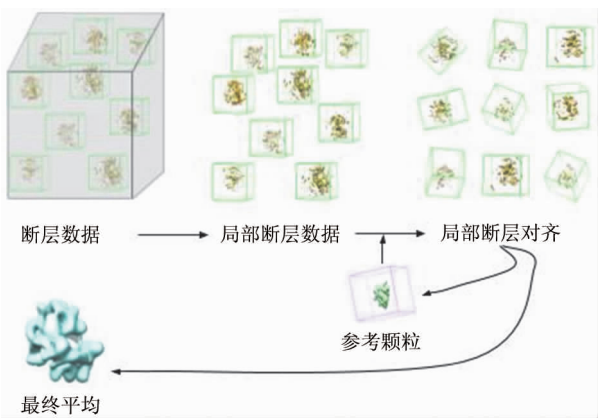

图 2　局部断层平均的简单描述

1.2　局部断层对齐

局部断层平均技术工作流程的核心是迭代对齐和平均过程。在迭代对齐过程,N 个局部断层 3 维颗粒(3 个旋转参数、3 个平移参数)共有 6^{N-1} 个维度,非凸优化问题需要解决。局部断层平均的前提就是局部断层 3 维图像颗粒都已经对齐,故局部断层 3 维颗粒对齐是局部断层平均中最核心的任务。由于局部断层 3 维颗粒数据存在缺失锥,故在对齐过程中应该考虑缺失锥问题,最普遍的校正缺失锥的方法是采用约束性相关系数[6](constrained correlation coefficient, CCC)。在实空间,缺失信息造成 3 维结构的变形,影响局部断层数据的分类和平均,所以局部断层对齐过程中通过约束性互相关解决缺失锥的影响。

为了加快局部断层对齐的速度,Bartesaghi 等人[2]将沿射线经过傅里叶空间原点的所有傅里叶系数大小投影到单位球面上相应的点上,然后对相应的 2 幅 2 维球面图像利用球谐函数进行 2 维匹配,得到旋转参数,并且利用傅里叶空间性质,得到平移参数。Xu 等人[3]改进了该方法,实现了 3 维的快速旋转匹配,并利用局部断层整体数据进行对齐。这些局部断层对齐方法都是粗粒度的。Xu 等人[4]利用了莱文贝格-马夸特方法,迭代计算 3 个旋转和 3 个平移参数,利用多线程进行并行计算,实现局部断层细化对齐。但这些细化对齐方法在每次迭代过程中需要计算整个局部断层 3 维体,非常耗时。

1.3　Spark 分布式计算框架

Apache Spark[7,8]是一个通用的开源分布式集群计算框架。Spark 提供了一个在集群上实现数据并行和编程的接口。Spark 是基于弹性分布式数据集(resilient distributed dataset, RDD)。RDD 是分布在一组计算机上并以容错方式维护的只读多集数据项。在 Spark 框架里,客户端任务转换成 RDD,然后 RDD 经过一系列转换算子(transformation)操作,通过利用行动算子(action)触发任务的执行。因为 RDD 具有容错机制的特性,Spark 框架在内存处理数据,适合快速数据处理和迭代处理。

在局部断层对齐过程中,本文使用 Spark 分布式并行计算各初始旋转和平移参数对各节点分别计算,得到对应的局部最优值,然后对各节点的计算结果进行比较,得到全局择优的旋转参数和平移参数,从而快速实现局部断层细化对齐。

2　基于小批量梯度下降和 Spark 分布式框架的局部断层细化对齐

2.1　基于实空间的局部断层对齐优化

局部断层 3 维图像定义为一个可积分函数,$V(\boldsymbol{x}):\mathbb{R}^3\to\mathbb{R}$,局部断层 3 维图像的平移操作 $\Lambda_{\boldsymbol{T}}$,且 $\boldsymbol{T}\in\mathbb{R}^3$,定义为[6]:

$$\Lambda_{\boldsymbol{T}}V(\boldsymbol{x}):=V(\boldsymbol{x}-\boldsymbol{T}) \tag{1}$$

局部断层 3 维图像旋转操作 $\Lambda_{\boldsymbol{R}}$ 可以表示为[6]:

$$\Lambda_{\boldsymbol{R}}V(\boldsymbol{x}):=V[\boldsymbol{R}^{-1}(\boldsymbol{x})] \tag{2}$$

其中旋转 $\boldsymbol{R}$ 是一个 3×3 旋转矩阵。

局部断层 3 维 $V(\boldsymbol{x})$ 的旋转和平移操作可以表示为[6]：

$$\Lambda_{\boldsymbol{T}}(\Lambda_{\boldsymbol{R}}V(\boldsymbol{x})) = V(\boldsymbol{R}^{-1}(\boldsymbol{x}) - \boldsymbol{T}) \tag{3}$$

局部断层 3 维变换参数 $\boldsymbol{\beta}$ 包括成对的旋转和平移操作，表示为 $\boldsymbol{\beta} = (\boldsymbol{R},\boldsymbol{T}) = (\phi,\theta,\psi,\tau_1,\tau_2,\tau_3)^{\mathrm{T}}$，其中旋转参数 $\boldsymbol{R} = (\phi,\theta,\psi)^{\mathrm{T}}$ 可视为使用 ZYZ 惯例的欧拉角[9]，平移参数 $\boldsymbol{T} = (\tau_1, \tau_2, \tau_3)^{\mathrm{T}}$。

局部断层 3 维图像 V_1 的缺失锥由傅里叶空间的 3 维模板 $\mathcal{M}$ 表示，其中在频率测量可用的区域中值为 1，在有限倾斜范围导致没有数据覆盖的区域中值为 0。2 个局部断层 3 维图像 V_1 和 V_2 在傅里叶空间对齐后的重叠区域 $\Omega := \mathcal{M}(V_1)\ \Lambda_R\mathcal{M}(V_2)$。当 2 个局部断层 3 维图像对齐时，约束性互相关只考虑傅里叶空间中旋转的最佳重叠区域，并根据傅里叶空间的平移不变性，消除平移的影响，即只考虑在傅里叶空间旋转操作。为了降低局部断层周围噪声的影响，本文在实空间定义一个二值化模板函数 M。

在实空间，局部断层 3 维图像 V_1 的正则化函数表示为[6]：

$$V_1^* := \frac{(\mathbb{F}^{-1}(\mathbb{F}(V_1)\cdot\Omega) - \overline{V_1^*})\cdot M(x,y,z)}{\sqrt{\sum_{x,y,z}((\mathbb{F}^{-1}(\mathbb{F}(V_1)\cdot\Omega) - \overline{V_1^*})\cdot(M(x,y,z)))^2}} \tag{4}$$

其中，$\mathbb{F}$ 定义为傅里叶变换，$\mathbb{F}^{-1}$ 定义为傅里叶逆变换。局部断层均值 $\overline{V_1^*}$ 受 M 和 Ω 限制，表示为[6]：

$$\overline{V_1^*} = \frac{1}{\sum_{x,y,z}M}\sum_{x,y,z}\mathbb{F}^{-1}(\mathbb{F}(V_1)\cdot\Omega) \tag{5}$$

局部断层 3 维图像 V_2 的约束性函数表示为[6]：

$$\Lambda_\beta V_2^* := \frac{(\mathbb{F}^{-1}(\mathbb{F}(\Lambda_T\Lambda_R V_2)\cdot\Omega) - \overline{\Lambda_\beta V_2^*})\cdot M(x,y,z)}{\sqrt{\sum_{x,y,z}((\mathbb{F}^{-1}(\mathbb{F}(\Lambda_T\Lambda_R V_2)\cdot\Omega) - \overline{\Lambda_\beta V_2^*})\cdot(M(x,y,z)))^2}} \tag{6}$$

其中，

$$\overline{\Lambda_\beta V_2^*} = \frac{1}{\sum_{x,y,z}M}\sum_{x,y,z}\mathbb{F}^{-1}(\mathbb{F}(\Lambda_T\Lambda_R V_2)\cdot\Omega)。$$

实际上局部断层体积是离散体素点，本文定义了归一化和对齐的局部断层 3 维图像 V_1^* 和 $\Lambda_\beta V_2^*$ 的约束性互相关函数[6]：

$$CCC := \max\sum_{x,y,z}V_1^*(x,y,z)\cdot\Lambda_\beta V_2^*(x,y,z) \tag{7}$$

在局部断层对齐过程中，最大化约束性互相关函数值，等同于最小化不相似值 d。给定一个标准化和对齐的局部断层 3 维图像 V_1^* 和 $\Lambda_\beta V_2^*$，d 在数学上可以表示为[4]：

$$d(\boldsymbol{R},\boldsymbol{T}) := (V_1^* - \Lambda_\beta V_2^*)^2 = 2 - 2\cdot CCC \tag{8}$$

在傅里叶空间的平移不变性，只需考虑旋转参数，通过快速旋转匹配算法[2,3]，得到一组初始旋转候选集合 $\{\boldsymbol{R}^1, \boldsymbol{R}^2,\cdots,\boldsymbol{R}^N\}$，然后通过快速平移匹配算法[10]，获取一组对应的平移候选集合 $\{\boldsymbol{T}^1, \boldsymbol{T}^2, \cdots,\boldsymbol{T}^N\}$。这样形成 N 对变换候选集合，$\{(\boldsymbol{R}^1, \boldsymbol{T}^1), (\boldsymbol{R}^2, \boldsymbol{T}^2),\cdots,(\boldsymbol{R}^N, \boldsymbol{T}^N)\}$，其中 N 是候选集合数，但这些初始成对候选集合不能实现局部断层的精对齐，需要进行局部断层细化对齐。

给定一组参数 $\{\boldsymbol{R}, \boldsymbol{T}\}$，本文在实空间通过小批量梯度下降算法[11]对局部断层 3 维图像进行局部细化对齐。利用该算法可以实现局部断层 3 维图像与参考图像进行局部精对齐。在局部断层体积 V 中，得到一组新的旋转参数值 $\boldsymbol{R}^k$ 和平移参数值 $\boldsymbol{T}^k$，使得标准化欧式距离变得越来越小：

$$d_{\boldsymbol{R}^k,\boldsymbol{T}^k} \geqslant d_{\boldsymbol{R}^{k+1},\boldsymbol{T}^{k+1}} \tag{9}$$

在局部断层 3 维图像对齐时，通常一个局部断层 3 维图像固定作为参考体积(无缺失锥)，另一个局部断层 3 维图像(有缺失锥)进行旋转和平移操作，与固定的参考体积进行对齐。然而不能直接对局部断层体积使用 MBGD 算法。针对局部断层体积的特点，本文设计沿局部断层 3 维图像的 x 轴进行小批量采样，这样就可以在局部断层 3 维图像上使用小批量梯度下降算法进行变换参数的计算，实现局部断层体积的细化对齐。

通过 MBGD 算法不断更新 $\boldsymbol{\beta} = (\boldsymbol{R}, \boldsymbol{T})$，定义 3 维局部断层体积对齐的损失函数 $\mathcal{L}$ 为：

$$\mathcal{L}(\boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{R}, \boldsymbol{T}) = \frac{1}{2n}\sum_{i=1}^{n}H_{(\boldsymbol{R},\boldsymbol{T})}(x_i) \tag{10}$$

其中 n 是 3 维局部断层体积沿 x 轴的长度，$H_\beta(x_i) := (V_1(x_i)^* - \Lambda_\beta V_2(x_i)^*)^2$。

MBGD 算法的迭代表达如下：

$$\beta^k := \beta^{k-1} - \frac{\alpha_k}{B}\sum_i^{i+B} H_{(\boldsymbol{R},\boldsymbol{T})}(\mathrm{x_i})',\ \mathrm{k} \geqslant 1 \quad (11)$$

其中在每次迭代中,局部断层 3 维体中沿 x 轴的截面坐标 i 在 $\{1,\cdots,n-B\}$ 被随机地选择,小批量的长度为 B,α_k 是步长。

通过式(10)和式(11),描述算法的流程(算法 1)。Xu 的对齐方法每次迭代需要全部数据参与计算,计算量大,但能实现较快收敛。而随机梯度下降算法[12](stochastic gradient descent, SGD)每次迭代只需一个样本,虽然计算速度快,但不是每次迭代都向着整体优化方向,收敛速度缓慢。基于 MBGD 细化对齐算法每次仅需要小批量的数据参与梯度计算,所以计算速度快,且能较快实现收敛。所以本文提出的基于 MBGD 细化对齐算法结合了 Xu 的对齐算法和 SGD 算法的优点,既能实现较快的计算时间,又能保证收敛的速度,适合在局部断层体积精对齐中应用。

算法 1 基于 MBGD 局部断层细化对齐方法最小化约束性不相似值 $\frac{1}{2n}\sum_{i=1}^{n} H_\beta(x_i)$

$d=0$ 且 $i=\{1,2,\cdots,n\}$
eps = 0.0001
old _ d = min _ d = 1000
for $k=0;\ k \leqslant maxIter;\ k++$ **do**
  随机选择截面 i, $i=\{1,2,\cdots,n-B\}$
  $d=\sum_i^{B+i} H_\beta(x_i)=\sum_i^{B+i}(V_1(x_i)^*-\Lambda_\beta V_2(x_i)^*)^2$
  $\beta=\beta-\frac{\alpha}{B}\sum_i^{i+B} H_\beta(x_i)'$
  if $d <$ min _ d **then**
    min _ d = d
  end if
  if |min _ d − old _ d| ≤ eps **then**
    break
  else
    old _ d = min _ d
  end if
end for

对于初始的旋转参数 $\boldsymbol{R}$ 和平移参数 $\boldsymbol{T}$,算法的最终迭代结果得到精确的局部断层对齐旋转参数 $\boldsymbol{R}^{k+1}=\boldsymbol{R}^k-\frac{\alpha_k}{B}\sum_{i_k}^{i_k+B} H_{\boldsymbol{R}}(x_i)'$ 和平移参数 $\boldsymbol{T}^{k+1}=\boldsymbol{T}^k-\frac{\alpha_k}{B}\sum_{i_k}^{i_k+B} H_{\boldsymbol{T}}(x_i)'$,其中 k 和 $k+1$ 分别是迭代次数。

2.2 基于 Spark 架构分布式并行局部断层细化对齐过程

为了实现全局择优的旋转和平移参数,本文首次提出基于 Spark 架构分布式并行局部断层细化对齐过程,首先需要对不同旋转候选参数和平移候选参数中执行多次局部断层细化对齐,实现每个候选参数(旋转和平移)得到局部最优,然后比较不同旋转参数和平移参数的局部最优值,得到全局择优的旋转和平移参数。为了实现不同候选集同步计算,通过 1.3 节的分析,本文选取 Spark 架构实现并行局部断层细化对齐,得到全局择优的旋转和平移参数。虽然 Spark 在内存中处理数据,具有容错机制和冗余机制,适合高频数据交换和大数据迭代处理,但 Spark 仍然是一个粗粒度的分布式框架,为了实现基于 Spark 架构分布式局部断层细化对齐,需要设计新颖的数据交换和转移代码。

本文设计基于 MBGD 和 Spark 框架分布式局部断层细化对齐,具体流程如图 3,通过快速旋转匹配和快速平移匹配,得到 N 对旋转和平移集合数组,把数组转换成 RDD,并对 RDD 进行分区,然后通过调用 mapPartitions 操作,执行各节点任务的分配,调用 collect 操作,每个节点并行执行基于 MBGD 局部断层细化对齐,返回各节点局部最优变换参数,通过比较,得到全局择优参数。

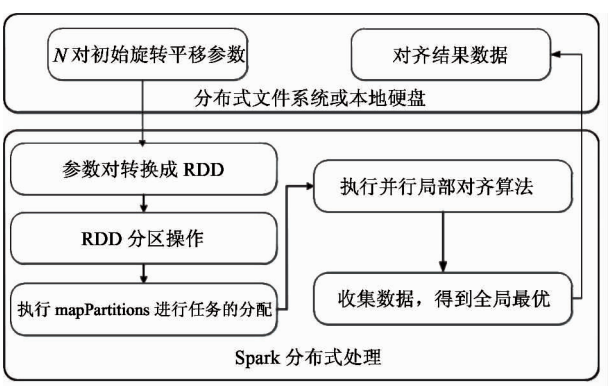

图 3 Spark 框架的分布式局部断层细化对齐算法流程

3 实验与分析

3.1 仿真数据

从 Protein Data Bank (PDB)数据库下载分子伴侣蛋白质(GroEL)原子模型,设定了分辨率和空间体素值,并对 GroEL 数据进行低通滤波。

本文使用 Situs PDB2VOL[13]程序获得 GroEL 电子密度图,对 GroEL 电子密度图进行随机旋转和平移操作,利用散焦值模拟对比度传递函数(contrast transfer function, CTF)。在指定的倾斜范围和角度增量条件下,对 GroEL 电子密度图进行投影,以模拟断层数据进行有角度采样过程。对投影图像添加高斯噪声和调制传递函数噪声(modulation transfer function noise, MTF)来模拟电子光学效应。投影后的 2 维图像利用加权反投影算法(weighted back projection, WBP)进行 3 维重构,这样就产生模拟的局部断层 3 维数据集。

GroEL 蛋白质原子模型(PDB ID:1KP8)用于产生网格点数为 64 × 64 × 64 的局部断层 3 维体积,其中体素尺寸为 0.6 nm × 0.6 nm × 0.6 nm, 散焦为 −6 μm。

在倾斜范围为 ±60 °、角度增量为 1 °的 3 种不同 SNR(0.01,0.03,0.003)条件下,分别模拟 20 个局部断层数据,这些模拟数据进行随机旋转和平移操作。在倾斜范围为 ±40 °、角度增量为 1 °的 3 种不同 SNR(0.01,0.03,0.003)条件下,也进行相同的操作。GroEL 局部断层仿真数据经常出现在局部断层分析方面的相关文献中[2-4]。

沿 *x-z* 平面得到的不同倾斜范围和 SNR 条件下的中心切片,如图 4 所示。在图 4 中,具有较小倾斜范围(如 ±40 °)和较低 SNR(0.003)的局部断层数据明显具有较大的变形,肉眼几乎无法分辨其中所包含的结构。

3.2 实验数据

实验的局部断层数据为分子伴侣蛋白质复合体(GroEL 和 GroEL/ES)数据集[6]。

为了收集这些 $GroEL_{14}GroES_7$ 复合体,Förster 等人[6]采用如下流程:1 μM $GroEL_{14}$ 和 5 μM $GroES_7$ 需要放在 5 mM $MgCl_2$,5 mM KCL, 5 mM ADP,1 mM DTT 和 12.5 mM Hepes (pH 7.5)的缓冲液中进行孵育,并在 30 ℃下培育 15 min。使用网格将 3.5 μl 蛋白质溶液和 0.5 μl 的 10 nm BSA 胶体金悬浮液进行混合。样品用插入式冷冻法进行玻璃化。

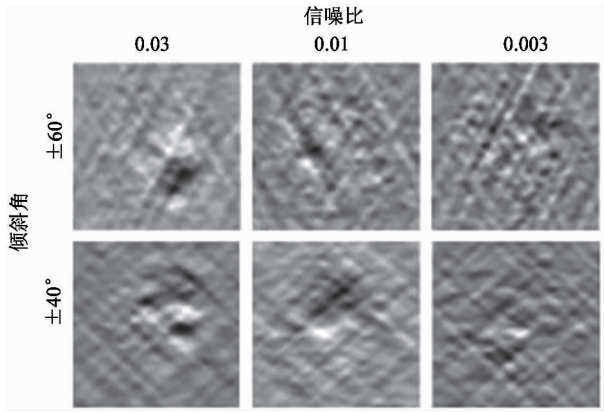

图 4 仿真局部断层数据切片(*x-z* 平面)

样品在倾斜角度 ±65 °,2 °或 2.5 °角度增量条件下,使用 Tecnai G2 Polara 显微镜(配备 2k × 2k FEI CCD 相机)进行低温采样和投影,并利用 UCSF tomography 软件在单轴倾斜下进行 2 维投影数据的获取。图像在 2k × 2k 像素的 CCD 相机和 7 ~ 4 μm 的散焦水平上进行记录。物体像素尺寸为 0.6 nm。

3.3 实验环境

本文实现的基于 MBGD 和 Spark 分布式方法的局部断层细化对齐算法在 4 台服务器上运行,每台服务器配置包括 2 个 1.7 GHz Intel Xeon Bronze 3104 CPUs,含有 12 个物理核,1 个千兆网卡,30 G 内存。

Spark 集群部署采用 stand alone 模式,1 个为管理节点,4 个为工作节点。代码开发是在 Spark 2.0 和 Python 2.7 环境下进行。

3.4 实验数据分类

数千个实验的分子伴侣蛋白质复合体(GroEL 和 GroEL/ES)数据集也包含假定粒子,因此需要手动挑选,并且与局部断层平均颗粒进行对齐,排除低的互相关系数(如 $CCC \leqslant 0.42$),剩下的颗粒被挑选进行局部断层对齐和分类。实验的 ~ 800 kDa $GroEL_{14}$ 和 $GroEL_{14}/GroES_7$ 局部断层复合体数据集作为局部断层对齐和分类研究的准标准[6,14,15]。在数据

集中 786 个局部断层颗粒以任意方向和非监督方式对所有局部断层颗粒进行平均而对齐。

本文使用 MCO-A[13] 分类算法对 GroEL 和 GroEL/ES 复合体分类,其中含有 10 个初始类和 7 倍对称。本文用 MCO-A 方法得到最终 3 个不同的类,其结果与先前在文献[6,14-16]中发布的结果一致。由 MCO-A 分类方法产生的每个分类平均的中心切片显示在图 5 中。

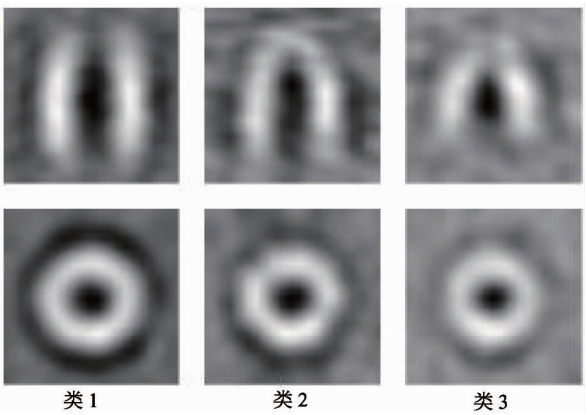

图 5　使用 MCO-A 方法分类 GroEL 和 GroEL/ES 复合体

3.5　不同局部断层对齐算法精度比较

为了比较基于 MBGD 局部断层细化对齐方法与 Xu 的对齐方法,本文使用无噪声的参考体积,在不同倾斜范围和不同信噪比条件下,20 个仿真局部断层数据与参考体积进行对齐,然后通过假设检验计算两种方法对应的互相关系数值,计算对应的 t 值和 P 值,比较 2 种算法的对齐精度。

无噪声参考体积通过 GroEL 结构(PDB ID:1KP8)产生的。参考体积低通滤波至 6 nm 分辨率,并用作对齐过程的初始参考。

在 Xu 的对齐方法使用约束性互相关方法评价其对齐精度,为了科学地评价 2 种对齐算法精度,在基于 MBGD 细化对齐方法中也使用约束性互相关方法评价其对齐算法的精度。

使用基于成对数据的 t 检验来比较 2 种不同对齐方法的精度。对 2 种对齐方法得到的一批成对的互相关值,利用基于 MBGD 对齐方法得到一系列互相关值减去 Xu 的对齐方法得到的一系列互相关值,这样就形成了一系列差值,然后通过计算差值的 P 值,来比较两者对齐方法是否存在显著性。

如表 1 和图 6 所示,在倾斜角范围为 ±60 ° 下,本文提出的基于 MBGD 细化对齐方法与 Xu 的对齐方法进行比较,各自使用模拟的信噪比为0.003的局部断层数据进行对齐,得到系列 CCC 差值的平均值为正,对应的 P 值为 1.01E-12(P <0.01),说明 2 种方法具有非常明显的差异,基于 MBGD 细化对齐方法的性能优于 Xu 的对齐方法。但在倾斜角范围为 ±60 °、信噪比为 0.01 和 0.03 条件下,基于 MBGD 细化对齐方法和 Xu 的对齐方法无显著的差异(P > 0.05)。

表 1　在倾斜范围 ±60 ° 下 2 种对齐算法间 P 值比较

信噪比	P 值
0.03	0.20
0.01	0.26
0.003	1.01E-12

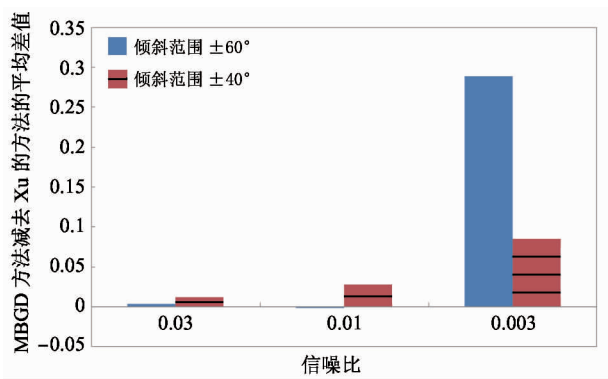

图 6　不同倾斜范围和信噪比下 2 种对齐方法互相关差值的平均值

如表 2 和图 6 所示,在倾斜角范围为 ±40 °、信噪比为 0.003 条件下,仿真的局部断层数据缺失信息更多,且信噪比也更低,基于 MBGD 细化对齐方法与 Xu 的对齐方法进行比较,得到系列 CCC 差值的均值为正,对应的 P 值为 2.15E-05(P <0.01),表明2种对齐方法有非常显著差异,基于MBGD细化

表 2　在倾斜范围 ±40 ° 下 2 种对齐算法间 P 值比较

信噪比	P 值
0.03	0.17
0.01	0.05
0.003	2.15E-05

对齐方法的对齐性能也优于 Xu 的对齐方法。而在倾斜角范围为 ±40 °、信噪比为 0.01 和 0.03 条件下,2 种对齐方法没有显著差异($P>0.05$)。

实际上,利用电子冷冻断层扫描技术重构的断层图像,因为电子剂量低,存在缺失锥,造成 3 维图像信噪比低,信噪比为 0.003 时更接近于实验条件下的局部断层 3 维图像。通过对 2 种对齐算法精度的比较,在倾斜范围为 ±60 °和 ±40 °下,信噪比为 0.003,通过对表 1、表 2 和图 6 的分析,基于 MBGD 细化对齐方法在对齐精度上优于 Xu 的对齐方法,这也表明基于 MBGD 细化对齐方法更加适合对实验环境下局部断层数据图像对齐。

3.6　不同局部断层对齐算法运算时间比较

为了客观公正地比较 2 种对齐算法的运行时间,2 种对齐算法都用 Python 语言进行实现。

本文使用随机方向上的局部断层平均体积作为两者对齐算法的初始参考,每次局部断层对齐算法收敛时,就会得到一个新的参考和对应的分辨率,通过在指定的最大迭代次数下,得到最佳的分辨率值对应的运算时间和迭代次数。这种初始参考称为无参考策略[16]是不需要外部参考,因为外部参考会导致参考体积的偏差。

首先比较每个对齐算法使用一次的运行时间。利用在倾斜角范围为 ±60 °、信噪比为 0.003 条件下仿真的 20 个局部断层数据进行实验。如图 7 所示,本文提出的基于 MBGD 细化对齐算法运行时间是 63 s,而 Xu 的方法耗时是 150 s。

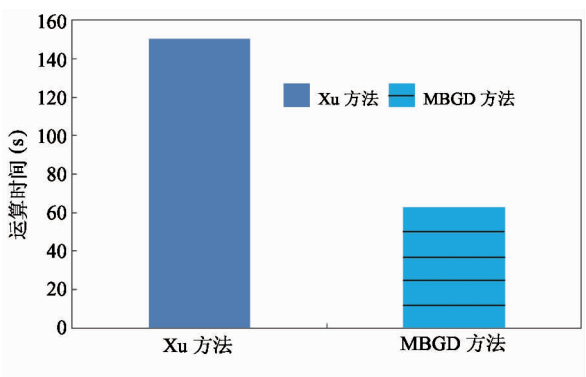

图 7　2 种对齐算法运算时间

通过 2 种对齐算法运行时间的比较,基于 MBGD 局部断层细化对齐算法的运行时间大体上是 Xu 对齐方法的一半,甚至更短。

然后利用 GroEL 局部断层实验数据比较 2 种对齐方法在得到最佳分辨率时的迭代次数。把实验数据分成相同的两部分,每部分独立进行对齐和迭代,然后对每次迭代后对齐的数据进行平均,使用金标准的 FSC(FSC =0.143)方法计算分辨率,平均全部数据为最新的参考,不断循环,直到循环结束。利用金标准 FSC 为 0.143 的规则,基于 MBGD 细化对齐算法和 Xu 的对齐方法分别对 GroEL 局部断层实验数据进行对齐和平均,记录得到最佳分辨率对应的迭代次数。

首先使用基于 MBGD 细化对齐算法对 GroEL 局部断层实验数据进行迭代对齐和平均,经过 5 次迭代,得到最佳分辨率为 30 Å 的 GroEL 局部断层平均(见图 8)。

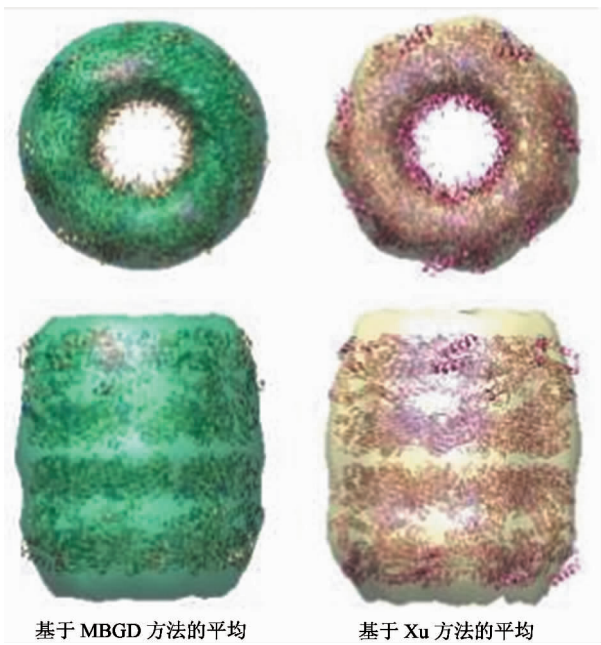

图 8　2 种对齐方法对 GroEL 局部断层实验数据平均

然后,使用 Xu 的对齐方法对 GroEL 局部断层实验数据进行迭代对齐和平均,经过 9 次迭代,最终 GroEL 局部断层数据平均的最佳分辨率为 32.5 Å(见图 8)。

通过对得到最佳分辨率时迭代次数的比较,基于 MBGD 局部断层细化对齐算法得到最佳分辨率的迭代次数也近似是 Xu 方法得到最佳分辨率迭代次数的一半。

基于上述 2 种运算时间的分析,在细化对齐得到的最佳分辨率基本不变的条件下,基于 MBGD 局部断层细化对齐算法的运算时间明显优于 Xu 的对齐方法,近似为 Xu 方法的一半,并且基于 MBGD 局部断层细化对齐算法的迭代次数基本上为 Xu 的对齐方法迭代次数的 1/2。

4 结 论

针对局部断层对齐过程计算量大,且需要细化对齐的问题,本文提出基于小批量梯度下降的细化对齐算法,在实空间实现优化约束性不相似值。通过对仿真局部断层数据进行测试,在倾斜范围为 ±60 °和 ±40 °、信噪比为 0.003 条件下,通过成对数据的 t 检验,P 值都小于 0.01,证明本文提出的基于 MBGD 细化对齐算法在对齐精度上明显胜过 Xu 的对齐方法。相对于 Xu 的对齐方法,基于 MBGD 细化对齐方法更加适合对实验环境下(如信噪比为 0.003)局部断层数据进行对齐。

通过对仿真局部断层数据和实验 GroEL 数据进行对齐,在对齐算法运行一次的时间比较上,本文提出的基于 MBGD 对齐方法的运算时间大体为 Xu 对齐方法运算时间的一半;在得到最佳分辨率的迭代次数比较上,本文提出的基于 MBGD 局部断层细化对齐算法迭代次数基本上为 Xu 的对齐方法迭代次数的 1/2,同时基于 MBGD 对齐方法得到的最佳分辨率(30 Å)也略优于 Xu 对齐方法得到的最佳分辨率(32.5 Å)。本文提出的基于 MBGD 局部断层细化对齐方法有利于局部断层对齐的结果优化和更改。

此外,本文在局部断层对齐过程中使用 Spark 框架实现局部断层细化对齐,该框架可以实现局部断层细化对齐的分布式计算,从而得到全局择优参数,相对于其他分布式框架,Spark 框架实现更简单。

参考文献

[1] Kovacs J, Wriggers W. Fast rotational matching[J]. *Acta Crystallographica Section D: Biological Crystallography*, 2002, 58(8):1282-1286

[2] Bartesaghi A, Subramaniam P, Liu J, et al. Classification and 3D averaging with missing wedge correction in biological electron tomography[J]. *Journal of structural biology*, 2008, 162(3):436-450

[3] Xu M, Beck M, Alber F. High-throughput subtomogram alignment and classification by Fourier space constrained fast volumetric matching[J]. *Journal of structural biology*, 2012, 178(2):152-164

[4] Xu M, Alber F. High precision alignment ofcryo-electron subtomograms through gradient-based parallel optimization [J]. *BMC systems biology*, 2012, 6(1):18

[5] WIKIPEDIA. Electron Cryotomography[EB/OL]. https://en.wikipedia.org/wiki/Electron_cryotomography: Wikipedia, 2019

[6] Förster F, Pruggnaller S, Seybert A, et al. Classification of cryo-electron sub-tomograms using constrained correlation[J]. *Journal of structural biology*, 2008, 161(3): 276-286

[7] Zaharia M, Chowdhury M, Franklin M J, et al. Spark: cluster computing with working sets[J]. *HotCloud*, 2010, 10(10):95

[8] 姚晓,邱强,肖茁健,等. Spark 框架下矢量多边形求交算法研究[J]. 高技术通讯, 2018, 28(6): 500-507

[9] Brink D M, Satchler G R. Angular Momentum[M]. 2nd ed. Oxford: Clarendon Press,1968

[10] Frangakis A S, Böhm J, Förster F, et al. Identification of macromolecular complexes in cryoelectron tomograms of phantom cells[J]. *Proceedings of the National Academy of Sciences*, 2002, 99(22):14153-14158

[11] Friedlander M P, Schmidt M. Hybrid deterministic-stochastic methods for data fitting[J]. *SIAM Journal on Scientific Computing*, 2012, 34(3): A1351-A1379

[12] Robinds H, Monro S. A stochastic approximation method [J]. *Annals of Mathematical Statistics*, 1951, 22: 400-407

[13] Wriggers W, Milligan R, McCammon J. Situs: a package for docking crystal structures into low-resolution maps from electron microscopy[J]. *Journal of structural biology*, 1999, 125(2-3): 185-195

[14] Hrabe T, Chen Y X, Pfeffer S, et al. PyTom: a python-based toolbox for localization of macromolecules in cryo-electron tomograms and subtomogram analysis[J]. *Journal of Structural Biology*, 2012,178(2): 177-188

[15] Zhao Y X, Zeng X R, Guo Q, et al. An integration of

fast alignment and maximum-likelihood methods for electron subtomogram averaging and classification[J]. *Bioinformatics*,2018, 34(13):i227-i236

[16] Scheres S H, Melero R, Valle M, et al. Averaging of electron subtomograms and random conical tilt reconstructions through likelihood optimization[J]. *Structure*,2009, 17(12): 1563-1572

Subtomogram refined alignment based on mini-batch gradient descent and Spark distribution

Lü Yongchun[*,**,***], Zhao Xiaofang[*,**], Li Hua[*,**,***], Zeng Xiangrui[****], Xu Min[****]

([*]Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190)

([**]University of Chinese Academy of Sciences, Beijing 100190)

([***]Key Laboratory of Intelligent Information Processing, Chinese Academy of Sciences, Beijing 100190)

([****]School of Computer Science, Carnegie Mellon University, Pittsburgh 15213, USA)

Abstract

The radiation damage, lack of information and low signal-to-noise ratio (SNR) of biological samples when acquiring electron cryotomography (cryo-ET) images limit the recovery of three-dimensional structural information from tomographic data. Similar to the single particle three-dimensional reconstruction technique of electron cryomicroscopy (cryo-EM), subtomograms datasets are aligned and averaged to produce high-precision three-dimensional structure of macromolecular complexes. The existing subtomogram alignment techniques involve 6 degrees of freedom (3 rotation parameters and 3 translation parameters). Therefore, they process the whole three-dimensional volume image to calculate 6 degrees of freedom in each alignment of iteration, which is computationally intensive. To solve the above problems, this paper proposes a method based on mini-batch gradient descent (MBGD) to achieve subtomogram refined alignment of three-dimensional data, and uses Spark distributed framework for the first time to achieve global optimization of subtomogram refined alignment. Through the alignment of simulation data and experimental data, subtomogram refined alignment algorithm based on mini-batch gradient descent has achieved improvement in alignment accuracy and speed compared to the popular baseline method.

Key words: mini-batch gradient descent (MBGD), Spark, subtomogram refined alignment, electron cryotomography (cryo-ET)