蛋白质晶体结构刚体优化的新方法*

丁 玮[1)2)] 江 凡[1)†]

1)(中国科学院物理研究所,北京凝聚态物理国家实验室,北京 100190)

2)(中国科学院研究生院,北京 100049)

(2010 年 7 月 6 日收到;2010 年 12 月 7 日收到修改稿)

将多元函数的变尺度法与两电子密度图的相关系数相结合,得到一种能够对蛋白质晶体结构进行刚体优化的新方法. 初步测试结果表明,该方法能够明显地改善待测晶体的初始模型在晶胞中的取向和位置. 而与最大似然方法相比,该方法可能更适用于搜索空间中存在大量局域极值点的情况.

关键词: 蛋白质晶体结构, 刚体优化, 多元函数的变尺度方法, 电子密度图的相关系数

PACS: 61.50.Ah, 87.15.ad, 02.60.Pn

1. 引 言

分子置换法是当前测定蛋白质晶体结构的主要方法之一,近年来收录到国际蛋白质数据库中的新蛋白质晶体结构,约有 2/3 是使用上述方法测定的. 但由各种基于分子置换法原理设计的智能程序[1—7]求得的初始模型,虽然整体的结构和肽链的折叠方式基本正确,但原子坐标仍存在一定的错误或偏差,它们主要来自于衍射数据的测量误差和解析方法的近似. 为了获得更为精确的晶体结构数据,有必要对其进行优化.

常见的优化操作有,原子的坐标优化(coordinate refinement, CR),温度因子的优化(temperature factor refinement, TFR),占有率的优化(occupancy refinement, OR)等. 而刚体优化属于原子的坐标优化,其优化的基本思想是将模型当作刚体,仅对其空间取向和位置进行整体的调整. 它往往是对初始模型进行优化的首要步骤.

常用的优化方法主要有,1)最小二乘方法(least square method LSM)[8],它是以理论衍射振幅值$|F_c|$与实验衍射振幅值$|F_o|$之间差值的平方和作为优化的监控指标,并通过不断地修改模型的参数(如原子坐标、温度因子等),寻找其最小值的优化方法. 2)分子动力学方法(molecular dynamics method, MDM)[9],它是以系统能量值作为优化的监控指标,模拟冶金的升温降温过程,使整个系统重新收敛于一个新的能量最低状态的优化方法. 3)最大似然方法(maximum likelihood method, MLM)[10],它是以最大似然函数值作为优化的监控指标,在充分考虑实验值和模型参数值都存在误差的情况下,寻求最大似然的理论衍射振幅值$|F_c|$的优化方法. 最小二乘方法是过去较为常用的优化方法. 当初始模型质量较为理想时,它能较容易地得到理想的优化结果,但当初始模型的质量较差时,该方法往往难以跨越较高的能量壁垒而仅收敛于局域极值. 因此现在常将分子动力学方法和最大似然方法结合使用,以有效地减少局域极值以及过度优化,提高优化后模型的准确性.

基于上述优化思想,本文提出了一种新的优化方法,BFGS-MAPCC 方法. 它是以两电子密度图的相关系数(简称 MAPCC)作为优化的监控指标,并利用多元函数变尺度方法(variable metric methods in multidimensions, VMMM)[11]自动地调整模型的空间取向和质心位置,在一定范围内寻找 MAPCC 最大值的优化方法. 由于 MAPCC 值与初始模型和目标分子之间的平均相位差直接相关[12],因此它可以很好地充当优化的监控指标. 而多元函数变尺度方法除了具有良好的数值稳定性外,还可以人为地设定计算函数梯度时所使用的步长值,这将有助于

* 国家自然科学基金(批准号:10674172,10874229)资助的课题.

† 通讯联系人. E-mail: fjiang@aphy.iphy.ac.cn

避免函数收敛于局域极值或过度优化等问题.

PVMR-ROT[13,14]和 PHASER-BTF[6]是利用分子置换法求解晶体结构的两个程序,它们在获得分子置换法的旋转解和平移解上具有突出能力. 本文首先利用这两个程序求得 1J3F 晶体的平移解,进而根据平移解对搜索模型进行旋转和平移操作,得到待测晶体的初始模型,然后使用 BFGS-MAPCC 方法对这些模型进行刚体优化. 结果表明,经过 BFGS-MAPCC 方法的优化,模型与目标分子在空间取向和质心位置上更为接近,并且优化还使空间取向和质心位置都较为理想的模型的数目和排位得到了较大幅度的增加和提高. 此外本文还将 BFGS-MAPCC 方法与目前较为流行的最大似然方法在进行了对比测试,结果表明,BFGS-MAPCC 方法更易于跨越局域极值点的势垒,得到更为理想的优化效果. 因此 BFGS-MAPCC 方法有可能成为一种值得推广的适用于蛋白质晶体结构优化的新方法.

2. 方法与测试数据

2.1. MAPCC 值

如果待测晶体中的蛋白质分子与已知晶体中的蛋白质分子在结构上相同或相似,可以将已知晶体的蛋白质分子作为搜索模型,通过分子置换法求出待测晶体中类似分子的取向和位置,从而得到待测晶体分子结构的初始模型. 由初始模型可以计算出每个衍射点的相角和理论的衍射强度,进而可以根据电子密度的定义计算得到晶胞中的电子密度分布图[15],而若将理论的衍射振幅替换为实验的衍射振幅可以得到另一电子密度分布图,它们之间的相关系数简称 MAPCC,一般用 R 表示,其定义如下:

$$R = \frac{\langle \rho_1 \cdot \rho_2 \rangle - \langle \rho_1 \rangle \cdot \langle \rho_2 \rangle}{[\langle \rho_1^2 \rangle - \langle \rho_1 \rangle^2]^{1/2} \cdot [\langle \rho_2^2 \rangle - \langle \rho_2 \rangle^2]^{1/2}}, \tag{1}$$

ρ_1 和 ρ_2 分别为两电子密度图的电子密度值,$\langle\ \rangle$ 为平均值标记. R 的取值在 0—1 之间. 0 表示两电子密度图完全不符合,1 则表示完全符合. 由于 R 与初始模型和目标分子之间的平均相位差直接相关[12],因此 R 值的大小能较为准确地反映出初始模型与目标分子的符合情况. 本文使用文献[16]中的 OVERLAPMAP 程序对 R 值进行计算.

2.2. 多元函数变尺度方法

本文使用的多元函数变尺度方法[17]包含了用于确定各自变量搜索方向的 Broyden-Fletcher-Goldfarb-Shanno (BFGS) 秩 2 算法,它是求解无约束极小值点的拟牛顿方法中最有代表性的算法之一. 它要求函数可导,但是并不直接计算函数的 Hesse 矩阵,而是采用一阶梯度信息 $g_{(k)} = \nabla f(x^{(k)})$ 来构造一系列的正定矩阵 $H_{(k)}$ 来逼近 Hesse 矩阵. 由于本文的工作不存在具体的函数形式,函数梯度是通过数值差分法得到,即

$$g_{(k)} = \frac{f(x^{(k)} + \Delta x^{(k)}) - f(x^{(k)})}{\Delta x^{(k)}}, \tag{2}$$

其中的 $x^{(k)}$ 为分子置换法的平移解,$f(x^{(k)})$ 是该平移解对应的 MAPCC 与 1 之间的差值,即 (1-MAPCC). 因此通过该方法得到的函数最小值,即为 MAPCC 的最大值. $\Delta x^{(k)}$ 是人为选定的步长值. 一般而言,对于质量较好的搜索模型和实验数据,可以选择较小的步长,以避免过度优化,反之则应适当增加步长,以避免函数收敛于局域极值点. 本文在优化过程中用四元数[18]描述旋转,所使用的旋转步长是 0.01,而平移步长是 0.02 Å.

多元函数变尺度方法的基本流程如下:

1)给定初始点 $x^{(k)}$ ($k=1$) 以及允许误差 FTOL,并令 $\boldsymbol{H}_{(k)} = \boldsymbol{I}_n$ (单位矩阵).

2)计算出 $x^{(k)}$ 处的梯度

$$g_{(k)} = \nabla f(x^{(k)}). \tag{3}$$

3)令

$$d^{(k)} = -\boldsymbol{H}_{(k)} g_{(k)}. \tag{4}$$

4)从 $x^{(k)}$ 出发,沿方向 $d^{(k)}$ 作一维搜索,使得 $f(x^{(k)} + \lambda_{(k)} d^{(k)}) = \min\limits_{\lambda \geqslant 0} f(x^{(k)} + \lambda_{(k)} d^{(k)})$,以此确定步长因子 $\lambda_{(k)}$,并令

$$x^{(k+1)} = x^{(k)} + \lambda_{(k)} d^{(k)}. \tag{5}$$

5)检查是否满足收敛准则,若

$$\frac{|f(x^{(k+1)}) - f(x^{(k)})|}{|f(x^{(k+1)})| + |f(x^{(k)})|} < \frac{\mathrm{FTOL}}{2}, \tag{6}$$

则停止迭代,近似极小值点即为 $x^{(k+1)}$ 否则进行第 6)步.

6)令 $g_{(k+1)} = \nabla f(x^{(k+1)})$, $p^{(k)} = x^{(k+1)} - x^{(k)}$, $q^{(k)} = g_{(k+1)} - g_{(k)}$,然后利用 BFGS 公式计算

$$H_{(k+1)} = H_{(k)} + \frac{p^{(k)} p^{(k)\mathrm{T}}}{p^{(k)\mathrm{T}} q^{(k)}} - \frac{H_{(k)} q^{(k)} q^{(k)\mathrm{T}} H_{(k)}^{\mathrm{T}}}{q^{(k)\mathrm{T}} H_{(k)} q^{(k)}} + [q^{(k)\mathrm{T}} H_{(k)} q^{(k)}] u u^T, \tag{7}$$

其中 $u=\frac{p^{(k)}}{p^{(k)T}q^{(k)}}-\frac{H_kq^{(k)}}{q^{(k)T}H_kq^{(k)}}$.

让 $k=k+1$,若 k 已等于选定的最大迭代次数,迭代停止,否则返回第3)步.

利用 BFGS-MAPCC 方法,可以求出在当前平移解附近 MAPCC 最大值所对应的新平移解. 而更大的 MAPCC 值意味着由新平移解得到的初始模型与目标分子在空间取向和质心位置上可能更为接近. 此外,若按 MAPCC 值对初始模型进行排序将有利于提高空间取向和质心位置都较为理想的模型的排位.

2.3. 测试所使用的数据

测试使用的待测晶体为 1J3F 晶体[19],其空间群为 $P2_12_12_1$,晶胞参数为 $a=32.972$ Å, $b=58.787$ Å, $c=76.237$ Å,其中的蛋白质分子是含有 8 个 α 螺旋片段的肌红蛋白,共含 153 个残基. 测试使用的搜索模型源自 1A6M 分子[20]的 3 到 36 号残基,含 2 个由弯折联系起来的 α 螺旋 . 1A6M 与 1J3F 具有相似的三级结构,是一对同源蛋白. 1A6M 晶体的空间群为 $P2_1$,晶胞参数为 $a=63.80$ Å, $b=30.81$ Å, $c=34.35$ Å, $\beta=105.80°$,其中的蛋白质分子含有 151 个残基.

尽管目标分子共含 8 个 α 螺旋,但只存在唯一的一对 α 螺旋与搜索模型完全匹配,这意味着在搜索空间中只存在一个全局极值点. 但在这一对螺旋附近还分布着另外 3 个 α 螺旋(在此称为干扰螺旋),因此当搜索模型的质心位置和空间取向与目标分子的等同部分出现一定的偏差时,将会出现如下情况:搜索模型中的两个 α 螺旋的大部分残基与目标分子的等同部分重叠,但有一小部分残基与干扰螺旋发生部分重叠,而这一小部分重叠将有可能在一定程度上增加当前搜索模型与目标分子匹配的残基数目,这将使搜索空间出现局域极值点. 而优化过程中使用的监控指标(两电子密度图的相关系数)的性质特点也可能是使搜索空间出现局域极值点的一个原因,但它已是目前最好的优化监控指标之一[12].

测试中利用 PVMR-ROT 求搜索模型的旋转解时使用的帕特森图的分辨率为 3.0 Å; 利用 PHASER-BTF 求搜索模型的平移解时使用的衍射数据的分辨率为 3.0 Å.

2.4. 近似理想模型的判断标准

以已公布的 1J3F 晶体的分子结构作为标准,若由某一平移解得到初始模型与标准分子结构之间的角度距离在 25°(极角)以内、质心距离在 15 Å 以内,则称该初始模型为近似理想模型.

2.5. 平均相位差的计算

平均相位差是指由两不同的分子结构计算得到的各衍射点相角的差值的平均值. 为判断优化结果的好坏,本文计算了优化前后各近似理想模型与目标分子之间的平均相位差. 由于本文使用的搜索模型所含残基数目仅占目标分子残基数目的 22.2%,所以所得平均相位差均较大.

3. 结果和讨论

3.1. 测试步骤

本文工作的流程如下. 步骤 1. 利用 PVMR_ROT 和 PHASER_BTF 分别确定搜索模型在目标晶胞中的取向和位置. 首先利用 PVMR_ROT 求搜索模型的旋转解,并输出动态相关系数较大的 5000 个旋转解,然后利用 PHASER_BTF 求 5000 个旋转解对应的平移解,对每个旋转解输出 5 个平移解,并将这 25000 个平移解按平移函数的 Z-score (TFZ)的大小排序,然后输出 TFZ 较大的 5000 个平移解. 步骤 2. 根据平移解对搜索模型进行旋转和平移操作,得到 5000 个初始模型. 步骤 3. 利用 BFGS-MAPCC 方法对 5000 个初始模型进行优化. 步骤 4. 利用 PHASER_RNP[6]对 5000 个初始模型进行优化.

3.2. 结果分析

为了判断优化结果的好坏,本文选择了近似理想模型在所有初始模型中的数目、最佳排位、平均排位以及近似理想模型与目标分子之间的最小角度距离、平均角度距离、最小质心距离、平均质心距离、最小平均相位差、平均相位差的平均值共 9 项指标对优化的结果进行评价. 在理想情况下,优化的结果应该是增加近似理想模型的数目、提高近似理想模型的排位、减少近似理想模型与目标分子之间的角度距离、质心距离以及平均相位差.

此外,为了判断 BFGS-MAPCC 方法与目前较为

流行的最大似然方法在性能上的差异,本文还利用 PHASER_RNP 程序对 5000 个初始模型进行优化,并对这两组优化结果进行了比较. PHASER_RNP 是采用最大似然方法对结果进行刚体优化的程序,它以 Log Likelihood Gain(LLG)作为优化的监控指标,并且仅输出 LLG 大于 0 的解.

表 1 记录了 5000 个初始模型在经过 BFGS-MAPCC 方法优化前后与近似理想模型有关的数据. 结果显示,除了近似理想模型与目标分子之间的平均角度距离较优化前略有增加,其余指标均朝着理想的方向发展,这说明 BFGS-MAPCC 方法的确能够对待测晶体的初始模型进行优化. 而平均角度距离略为增加的原因是,5000 个初始模型中的一些非近似理想模型经过 BFGS-MAPCC 优化后成为了近似理想模型,但其空间取向和目标分子之间仍有一定的偏差,因此增加了近似理想模型与目标分子之间的平均角度距离.

表 2 记录了 5000 个初始模型在分别经过 BFGS-MAPCC 和 PHASER_RNP 优化后与近似理想模型有关的数据. 结果显示,经过 PHASER_RNP 优化,近似理想模型具有更为理想的排位,近似理想模型与目标分子之间的平均角度距离以及平均相位差的平均值也更为理想. 但由于本文使用的搜索模型较小,在旋转和平移空间中容易出现大量的局域极值点,而最大似然方法往往难以跨越其能量势垒而仅收敛于局域极值点中,因此经过 PHASER_RNP 优化,近似理想模型的数目较之前减少了 4 个,近似理想模型与目标分子的最小质心距离较优化前反而增加了 1.682 Å. 而经过 BFGS-MAPCC 方法的优化,近似理想模型的数目增加了 5 个,近似理想模型与目标分子之间的最小角度距离、最小质心距离、平均质心距离和最小平均相位差也更为理想. 这说明与最大似然方法相比较,BFGS-MAPCC 方法更易于跨越局域极值点的势垒,得到更为理想的优化效果,因此它可能更适用于搜索空间中存在大量局域极值点的情况.

表 1 5000 个初始模型在经过 BFGS-MAPCC 方法优化前后,与近似理想模型有关的数据[a)]

	数目	最佳排位	平均排位	最小角度距离/(°)	平均角度距离/(°)	最小质心距离/Å	平均质心距离/Å	最小平均相位差/(°)	平均相位差的平均值/(°)
不优化	9	292[b)]	2243	2.467	10.698	8.749	11.666	84.658	87.582
BFGS-MAPCC 优化	14	141[c)]	2194	0.638	13.574	6.277	10.954	82.330	87.344

a) 表格 1 从第二列起从左至右依次记录了在 5000 个初始模型中,近似理想模型的数目、最佳排位、平均排位以及近似理想模型与目标分子之间的最小角度距离、平均角度距离、最小质心距离、平均质心距离、最小平均相位差、平均相位差的平均值共 9 项数据.

b) 按 TFZ 的大小对优化前的模型进行排序得到的结果.

c) 按 MAPCC 的大小对优化后的模型进行排序得到的结果.

表 2 5000 个初始模型在分别经过 BFGS-MAPCC 和 PHASER_RNP 优化后,与近似理想模型有关的数据[a)]

	数目	最佳排位	平均排位	最小角度距离/(°)	平均角度距离/(°)	最小质心距离/Å	平均质心距离/Å	最小平均相位差/(°)	平均相位差的平均值/(°)
PHASER_RNP 优化	5	8[b)]	1110	1.466	8.932	10.431	11.700	82.879	86.451
BFGS-MAPCC 优化	14	141[c)]	2194	0.638	13.574	6.277	10.954	82.330	87.344

a) 表格 2 从第二列起从左至右依次记录了在 5000 个初始模型中,近似理想模型的数目、最佳排位、平均排位以及近似理想模型与目标分子之间的最小角度距离、平均角度距离、最小质心距离、平均质心距离、最小平均相位差、平均相位差的平均值共 9 项数据.

b) 按 LLG 的大小对优化后的模型进行排序得到的结果.

c) 按 MAPCC 的大小对优化后的模型进行排序得到的结果.

4. 结 论

BFGS-MAPCC 方法是将多元函数变尺度方法与 MAPCC 值结合得到的一种能够对蛋白质晶体结构进行刚体优化的新方法. 初步测试结果表明,BFGS-MAPCC 方法能够使 1J3F 晶体的初始模型在空间取向和质心位置上得到明显的改善,并且能够增加近似理想模型的数目、提高其在所有初始模型中的排位. 而与 PHASER_RNP 进行对比的测试结果则表明,与最大似然方法相比较,BFGS-MAPCC 方法可能更适用于模型较小、搜索空间中存在大量局域极值点的情况. 而优质的部分模型的获取将为待测晶体整体结构的解析奠定基础[21,22].

虽然初步的测试仅使用了一组待测晶体和搜索模型,但由于 BFGS-MAPCC 方法所使用的监控指标是 MAPCC 值,其大小仅与两电子密度图的符合程度有关而与待测晶体和搜索模型的种类无关,而两电子密度图符合程度的高低仅由搜索模型与目标分子等同部分重叠情况的好坏决定,因此可以预见 BFGS-MAPCC 方法将会具有良好的普适性.

尽管 BFGS-MAPCC 优化方法在初步的测试中得到令人鼓舞的结果,但使用该方法对单个模型的优化需要耗费计算机机时约 8 min(计算机的主频为 800 MHz). 而有研究表明[23,24],遗传算法和模拟退火算法可能具有更理想的优化效率. 因此,下一步的工作可以考虑进一步提高 BFGS-MAPCC 的计算效率.

[1] Trapani S, Navaza J 2008 *Acta Cryst.* D **64** 11

[2] Cohen S X, Ben Jelloul M, Long F, Vagin A, Knipscheer P, Lebbink J, Sixma T K, Lamzin V S, Murshudov G N, Perrakis A 2008 *Acta Cryst.* D **64** 49

[3] Long F, Vagin A A, Young P, Murshudov G N 2008 *Acta Cryst.* D **64** 125

[4] Vagin A, Teplyakov A 2010 *Acta Cryst.* D **66** 22

[5] Keegan R M, Winn M D 2008 *Acta Cryst.* D **64** 119

[6] McCoy A J, Grosse-Kunstleve R W, Adams P D, Winn M D, Storoni L C, Read R J 2007 *J. Appl. Cryst.* **40** 658

[7] Schwarzenbacher R, Godzik A, Jaroszewski L 2008 *Acta Cryst.* D **64** 133

[8] Konnert J 1976 *Acta Cryst.* A **32** 614

[9] Brunger A T, Adams P D, Rice L M 2006 *International Tables for Crystallography* F **375**

[10] Murshudov G N, Vagin A A, Dodson E J 1997 *Acta Cryst.* D **53** 240

[11] Press W, Teukolsky S, Vetterling W, Flannery B 1992 *Numerical recipes in C* Cambridge university press Cambridge p425

[12] Lunin V Y, Woolfson M M 1993 *Acta Cryst.* D **49** 530

[13] Jiang F 2008 *Acta Cryst.* D **64** 561

[14] Jiang F, Ding W 2010 *Chin. Phys.* B **19** 106101

[15] Drenth J, Mesters J 2007 *Principles of protein X-ray crystallography* 3rd ed. Springer New York p84

[16] Bailey S 1994 *Acta Cryst.* D **50** 760

[17] Upstill C 1988 *Nature* **333** 613

[18] Shoemake K 1985 *SIGGRAPH Comput. Graph.* **19** 245

[19] Ueno T, Koshiyama T, Ohashi M, Kondo K, Kono M, Suzuki A, Yamane T, Watanabe Y 2005 *J. Am. Chem. Soc* **127** 6556

[20] Vojtechovsky J, Chu K, Berendzen J, Sweet R M, Schlichting I 1999 *Biophys. J.* **77** 2153

[21] Zhang T, Wu L J, Gu Y X, Zheng C D, Fan H F 2010 *Chin. Phys.* B **19** 096101

[22] Zhang T, Wu L J, Gu Y X, Zheng C D, Fan H F 2010 *Chin. Phys.* B **19** 086103

[23] Bao W X, Zhu C C, Cui W Z 2005 *Acta Phys. Sin.* **54** 5281 (in Chinese) [保文星、朱长纯、崔万照 2005 物理学报 **54** 5281]

[24] Liang W X, Zhang J J, Lu J F, Liao R 2001 *Chin. Phys.* **10** 1129

A new method of rigid-body refinement for protein crystal structures*

Ding Wei[1)2)] Jiang Fan[1)†]

1) (*Beijing National Laboratory for Condensed Matter Physics, Institute of Physics, Chinese Academy of Sciences, Beijing* 100190, *China*)

2) (*Graduate School of the Chinese Academy of Sciences, Beijing* 100049, *China*)

(Received 6 July 2010; revised manuscript received 7 December 2010)

Abstract

A new rigid-body refinement method for protein crystal structures is described. It is based on "Variable Metric Methods in Multidimensions" and "the Electron Density Map Correlation Coefficient". Test shows that the orientation and translation of the initial models can be improved effectively by this new method. And it can escape from the local extremum and reach the global optimum more easily than the maximum likelihood method when the search space has a large number of local extremum points.

Keywords: protein crystal structures, rigid-body refinement, variable metric methods in multidimensions, the electron density map correlation coefficient

PACS: 61.50.Ah, 87.15.ad, 02.60.Pn

* Project supported by the National Natural Science Foundation of China (Grant Nos. 10674172, 10874229).

† Corresponding author. E-mail: fjiang@aphy.iphy.ac.cn