ISSN 1007-7626
CN 11-3870/Q
中国生物化学与分子生物学报 http://cjbmb.bjmu.edu.cn
Chinese Journal of Biochemistry and Molecular Biology
2017 年 6 月
33(6):630~637

·技术与方法·

DOI: 10.13865/j.cnki.cjbmb.2017.06.14

缺失数据下蛋白质多结构叠加的迭代方法

路建波[1)]， 高华方[1)]， 张世华[2)]， 卢本卓[3)]， 马 旭[1)*]

([1)] 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; [2)] 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; [3)] 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 蛋白质三维结构叠加面临的主要问题是，参与叠加的目标蛋白质的氨基酸残基存在某些缺失，但是多结构叠加方法却大多数需要完整的氨基酸序列，而目前通用的方法是直接删去缺失的氨基酸序列，导致叠加结果不准确。由于同源蛋白质间结构的相似性，因此，一个蛋白质结构中缺失的某个区域，可能存在于另一个同源蛋白质结构中。基于此，本文提出一种新的、简单、有效的缺失数据下的蛋白质结构叠加方法(ITEMDM)。该方法采用缺失数据的迭代思想计算蛋白质的结构叠加，采用优化的最小二乘算法结合矩阵 SVD 分解方法，求旋转矩阵和平移向量。用该方法成功叠加了细胞色素 C 家族的蛋白质和标准 Fischer's 数据库的蛋白质(67 对蛋白质)，并且与其他方法进行了比较。数值实验表明，本算法有如下优点：①与 THESEUS 算法相比较，运行时间快，迭代次数少；②与 PSSM 算法相比较，结果准确，运算时间少。结果表明，该方法可以更好地叠加缺失数据的蛋白质三维结构。

关键词 蛋白质结构叠加；蛋白质结构比对；迭代算法；缺失数据

中图分类号 Q811

Iterative Method for Multiple Structure Superposition of Proteins with Missing Data

LU Jian-Bo[1)], GAO Hua-Fang[1)], ZHANG Shi-Hua[2)], LU Ben-Zhuo[3)], MA Xu[1)*],

([1)] *Human Genetics Resource Center, National Research Institute for Family Planning, Beijing* 100081, *China*; [2)] *Institute of Applied Mathematics, Academy of Mathematics and System Sciences, Chinese Academy of Sciences, Beijing* 100190, *China*; [3)] *LSEC, Institute of Computational Mathematics, Academy of Mathematics and System Sciences, Chinese Academy of Sciences, Beijing* 100190, *China*)

Abstract The main problem in three-dimensional protein superposition is that some amino acid residues are missing in the superimposed target protein structures. However, most multiple structure superposition methods require the complete amino acid sequence. Current superposition methods deal with this problem usually by excluding amino acid sequence from the proteins, which leads to inaccurate results. Due to the similarity of the homologous protein structures, one structure of a protein may omit a region that is present in another structure of the same protein. In this paper, we propose a noval, simple and effective method (ITEMDM) for superpositioning multiple proteins with missing data. This method uses the idea of the iterative of missing data to compute the protein superposition problem. The rotation matrix and the translation vector are obtained by using the optimized least squares algorithm combined with matrix SVD

收稿日期: 2017-03-22; **修回日期**: 2017-04-18; **接受日期**: 2017-05-08

国家重点研究发展计划(No. 2016YFC1000307 **和** No. 2016YFB0201304); **国家自然科学基金**(No. 21573274); **国家重点研究发展计划子课题**(No. 2016YFC1000307-10) **和国家卫生计生委科学技术研究所科技创新基金面上项目**(No. 2017GJM04) **资助**

* **通讯作者** Tel: 010-62179059; E-mail: genetic88@126.com

Received: March 22, 2017; Revised: April 18, 2017; Accepted: May 8, 2017

Supported by National Key Research and Development Program of China (No. 2016YFC1000307 and No. 2016YFB0201304); National Natural Science Foundation of China (No. 21573274); Sub-project of National Key Research and Development Program of China (No. 2016YFC1000307-10) and Program of National Research Institute for Family Planning(No. 2017GJM04)

* Corresponding author Tel: 010-62179059; E-mail: genetic88@126.com

decomposition method. We successfully superimpose the cytochrome C family and the standard Fischer's database (67 pairs of proteins) by using ITEMDM method, and compare them with other methods. Numerical experiments show that our algorithm has the following advantages: 1) The operation time is faster and the iterations' number is smaller when compared with the THESEUS algorithm. 2) The result is more accurate and the operation time is smaller than PSSM algorithm. The results show that ITEMDM can superimpose the three-dimensional structures of the protein with missing data.

Key words structure superposition; structure alignment; iterative algorithm; missing data

叠加问题在许多科研领域都存在,包括人类学、考古学、计算机可视化、遗传生物学、地质学、图像分析、医药、形态测量、古生物学、心理学和分子生物学等。本文采用叠加方法比较相似蛋白质的三维构象。蛋白质结构比较和叠加不同,比较是指两者或者更多蛋白质残基结构之间离散的映射。比较最常用的方法是行列矩阵,矩阵中的元素是蛋白质氨基酸残基的缩写[1]。然而,叠加问题是三维空间结构的特殊方位。另外,两个结构的叠加很容易推广到多个结构同时叠加。计算一个最优的蛋白质叠加通常需要不同结构之间原子的一一对应映射[2]。然而,在许多情况下,蛋白质结构中氨基酸残基常常存在缺失,例如一个蛋白质结构中缺失的一个环区域,可能存在于另一个同源的蛋白质结晶结构中,目前已有许多结构叠加的研究[3]。进行蛋白质结构叠加有助于理解蛋白质的保守性和分支的进化、功能预测[4]、自动对接和蛋白质数据库标准的建立[5],以及蛋白质结构预测[6-7]。因此,叠加相近蛋白质结构具有重要意义。绝大多数处理多序列结构叠加的问题都可以分解为 2 个子问题。第 1 个是识别对应结构元素的对应,适用于很多蛋白质家族需要比较叠加。传统方法有 2 种: 自动识别和手动识别。第 2 个子问题是计算合适的刚体变换以得到最优的叠加。两结构的比对和叠加方法可以推广为多结构比对和叠加的方法,如蒙特卡洛优化方法[8](combinatorial extension and monte carlo, CE-MC)、弹性比对方法[9](multiple alignment with translations and twists, MATT)、基于几何哈希法的方法[10-11]、偏序结构比对[12](partial order structure alignment, POSA)和基于沃罗诺伊分割的多重弹性蛋白质比对方法(vorolign)[13]。

现在的多结构叠加方法都需要完整的序列数据[14-16]。因此,目前的多结构叠加方法处理缺失的数据尚不成熟,通常是从蛋白质序列中删除缺失的片段[9]。基于最大似然估计的多重蛋白质叠加方法(THESEUS)[1]是第一个考虑缺失数据的软件,此算法基于期望最大算法(expectation maximization algorithm, EM)。然而,EM 算法收敛速度慢而且对初值依赖性较大,该算法需要知道序列一一对应。具体方法如下:

首先,将待比对的蛋白质表示为:

$$X_i = v_i X_i + u_i X_i$$

其中,$v_i X_i$ 是能观察到的数据,$u_i X_i$ 是不能观察到的数据即缺失数据。补充指标矩阵 v_i 和 u_i 是对称,对角方阵。

然后,每一个 X_i 被表示为一个高斯扰动模型:

$$X_i = (M + E_i) R_i' - l_k t_i'$$

其中,M 是参加比对的蛋白质的平均结构,t_i 是 3×1 的平移行向量,l_k 表示 $k \times 1$ 的由 1 构成的列向量,R_i 表示合适的 3×3 的正交旋转矩阵。E_i 服从高斯分布 $E_i \sim N_{k,3}(0, \Sigma, I_3)$。完整叠加(superposition) log-似然函数 $l(R, t, M, \Sigma \mid X) = l_s$ 如下:

$$l_s = -\frac{1}{2} \sum_i^r tr\{(Y_i - M)' \Sigma^{-1} (Y_i - M)\} - \frac{dr}{2} ln \mid \Sigma \mid$$

然后是 EM 算的 E 步和 M 步反复迭代。其中,E 步是求完整似然函数的期望,通过求期望,去掉完整似然函数中的变量。M 步是对 E 步计算得到的完整似然函数的期望求极大值,得到新的参数估计值[17]。每次参数更新会增加非完整似然值,反复迭代后,会收敛到似然的局部最大值。值得注意的是,此方法需要蛋白质的顺序结构以及不同结构之间的一一对应。

Lu 等[18]采用主成分分析(principal component analysis, PCA)和迭代就近点法(iterative closest point, ICP)算法相结合的方法,提出了 PSSM 算法(simple and efficient protein structure superposition method for addressing the cases with missing data),该算法不需要蛋白质的序列比对信息。本文提出一种新的缺失数据下的迭代算法(iterative of missing data method, ITEMDM),该算法简单和有效。数值实验表明,该算法有如下优点: 1) 与 THESEUS 算法相比较,运行时间快,迭代次数少; 2) 与 PSSM 算法相比较,结果准确,运算时间少。

1 方法

1.1 旋转矩阵和平移向量的最小二乘求解法

设 $X_1, X_2, \cdots, X_n$ 和 $Y_1, Y_2, \cdots, Y_n$ 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 的旋转矩阵，T 是平移向量，$S_{i,j}$ 是分配变量。

$$\min \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} S_{i,j} \mid X_i - R(Y_i - T) \mid^2 \quad (1)$$

$$s.t. \sum_{i=1}^{n_x} S_{i,j} \leqslant 1, j = 1, \cdots, n_y$$

$$\sum_{i=1}^{n_y} S_{i,j} \leqslant 1, j = 1, \cdots, n_x$$

$$S_{i,j} \in \{0,1\}$$

$\overline{X}$ 和 $\overline{Y}$ 定义为:

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \quad (2)$$

$$\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i \quad (3)$$

在公式(1)中的平移向量 T 可表示为 $T = \overline{Y} - R^t \overline{X}$。

旋转矩阵 R 可以用欧拉角 Φ,θ 和 ψ[19] 表示如下:

$$R = \begin{pmatrix} \cos\psi\cos\Phi - \cos\theta\sin\Phi\sin\psi & \cos\psi\sin\Phi + \cos\theta\cos\Phi\sin\psi & \sin\psi\sin\theta \\ \sin\psi\cos\Phi - \cos\theta\sin\Phi\cos\psi & -\sin\psi\sin\Phi + \cos\theta\cos\Phi\cos\psi & \cos\psi\sin\theta \\ \sin\theta\sin\Phi & -\sin\theta\cos\Phi & \cos\theta \end{pmatrix} \quad (4)$$

由文献[20]可以通过如下表达式定义四元数参数 ξ, η, ζ, χ:

$$\xi = \sin\frac{\theta}{2}\sin\frac{\psi - \Phi}{2}$$

$$\eta = \sin\frac{\theta}{2}\cos\frac{\psi - \Phi}{2}$$

$$\zeta = \cos\frac{\theta}{2}\sin\frac{\psi - \Phi}{2}$$

$$\chi = \cos\frac{\theta}{2}\cos\frac{\psi - \Phi}{2}$$

则旋转矩阵 R 可以表示如下:

$$R = \begin{pmatrix} -\xi^2 + \eta^2 - \zeta^2 + \chi^2 & 2(\xi\eta - \zeta\chi) & 2(\xi\chi + \eta\zeta) \\ -2(\xi\eta + \zeta\chi) & \xi^2 - \eta^2 - \zeta^2 + \chi^2 & 2(\eta\chi - \xi\zeta) \\ 2(\eta\zeta - \xi\chi) & -2(\xi\zeta + \eta\chi) & -\xi^2 - \eta^2 + \zeta^2 + \chi^2 \end{pmatrix} \quad (5)$$

四元数并不是独立的,它们有如下关系:

$$\xi^2 + \eta^2 + \zeta^2 + \chi^2 = 1 \quad (6)$$

定义如下协方差矩阵 C:

$$C = \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})^t \quad (7)$$

目标是最小化公式(1),通过计算 R 和 T 使目标函数最小。通过计算得到 $(\xi, \eta, \zeta, \chi)^t$ 是如下矩阵 P 的最大特征值对应的特征向量。

$$P = \begin{pmatrix} -C_{11} + C_{22} - C_{33} - C_{12} - C_{21} - C_{23} - C_{32} & C_{13} - C_{31} \\ -C_{12} - C_{21} & C_{11} - C_{22} - C_{33} & C_{13} + C_{31} & C_{23} - C_{32} \\ -C_{23} - C_{32} & C_{13} + C_{31} & -C_{11} - C_{22} + C_{33} & C_{12} - C_{21} \\ C_{13} - C_{31} & C_{23} - C_{32} & C_{12} - C_{21} & C_{11} + C_{22} + C_{33} \end{pmatrix} \quad (8)$$

可以用矩阵 SVD 分解[21]求最大的特征值和特征向量。

1.2 迭代方法的思想和方法

迭代方法的基本思想是利用已知的数据来循环求出缺失的数据。该方法同样需要蛋白质的一一对应即蛋白质结构的序列比对。蛋白质的多序列比对可以用很多软件实现,例如: MAFFT[22] (rapid multiple sequence alignment based on fast fourier transform)、Clustal W[23]、MultAlin[24]、T-Coffee[25] (multiple sequence alignment that provides a dramatic improvement in accuracy with a modest sacrifice in speed as compared to the most commonly used alternatives)。双序列比对常用的有欧洲生物信息研究所(European Bioinformatics Institute, EBI)的 FASTA 软件工具和 NCBI 的 BLAST 工具。假设参加比对的蛋白质为 P_a 和 P_b。首先,用已有软件进行蛋白质 P_a 和 P_b 的序列比对。之后,把蛋白质 P_a 链上的公共部分的三维坐标存放在矩阵 A_{m3} 中,与之对应的蛋白质 P_b 链的公共部分的三维坐标存放在矩阵 B_{n3} 中。本文假设蛋白质 P_b 链是有缺失数据的蛋白质链,存放在 l_b 中。与之对应的 P_a 链的部分三维坐标存放在 l_a 中。经过第 i 步计算,得到缺失数据后的 2 条整链存放在 P_i 和 Q_i 中。然后,利用最小二乘法求出旋转矩阵和平移向量。直到算法收敛到预先给定的数值。

1.3 缺失数据的蛋白质结构叠加迭代算法

本文用 Matlab 语言实现该算法。主要的步骤如下: 步骤 1) 输入蛋白质文件以及序列比对。选择最长的链作为模板链,其他的蛋白质与模板链做比较; 步骤 2) 根据序列比对删除模板链的缺失数据对应的部分,之后用最小二乘法求旋转矩阵 R 和平移向量 T。在这一步,把删掉的部分放到一个文件中,例如: ls. txt。用一个 $l \times 3$ 的矩阵 L 来存储; 步骤 3) 求矩阵 R 的逆,将其记做 R^{-1},和平移向量 T 的逆

T^{-1} 。然后,利用它们的逆和矩阵 L 初步求出缺失的数据; 步骤 4) 把缺失的数据和它的对应部分添加到原来的蛋白质上。得到 2 个新的蛋白质序列: 一个是模板蛋白质,另一个是缺失数据的蛋白质,它的缺失数据用步骤 3 的初始值来代替; 步骤 5) 返回到步骤 3 直到收敛(迭代步数小于阈值或者均方根偏差,root mean square deviation,RMSD,小于给定值) 。

2　结果与讨论

2.1　细胞色素 C 的叠加

利用缺失数据下的迭代算法给出细胞色素 C 的 10 个蛋白质的比较结果,从可视化的角度给出结果。10 个蛋白质分别是: d1cih、d1pcbb、d1lfma、d1crj、d1csu、d1csx、d1yeb、d1kyow、d1m60a 和 d1u74d。这里考虑蛋白质的主链目的是画出蛋白质的图片和相关的参数。10 个蛋白质是多序列结构细胞色素 C 的 pdb 文件。它们有着不同的氨基酸序列,包括很多缺失的数据,但是它们有相似的结构。选择 d1cih 作为模板蛋白质,Fig. 1 显示的是它们的结构图。其中,Fig. 1A 表明,大部分蛋白质是离开的。Fig. 1B 表明,经过缺失数据迭代方法计算以后,10 个蛋白质的三维结构的叠加。

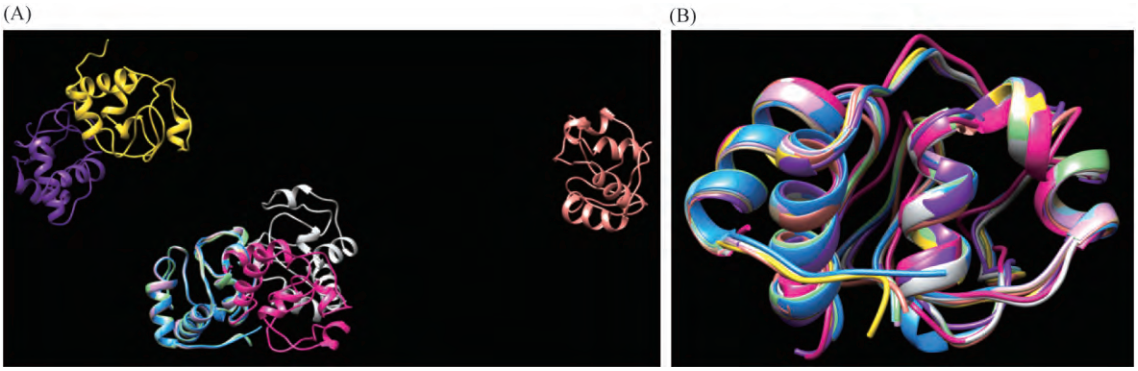

Fig. 1　Protein structures and positions before and after superposition by using iterative of missing data method for 10 cytochrome C proteins　(A) Protein structures before superposition. (B) Protein structures after superposition by using iterative of missing data method(ITEMDM) . The method used the idea of the iterative of missing data to compute the protein superposition problem. The rotation matrix and the translation vector were obtained by using the optimized least squares algorithm combined with matrix SVD decomposition method

2.2　与 THESEUS 方法比较

Table 1 表明,缺失数据迭代方法比 THESEUS 快。这是由于 THESEUS 算法需要计算矩阵的导数。这一步导致算法运算时间特别长,所以本文的算法程序迭代步数和程序的运行时间都比 THESEUS 算法少和短。下面具体说明迭代步数和运行花费的时间: 1) 蛋白质 d1cih 和 d1lfma,缺失数据迭代方法的运行的时间是 5.4 ms,而 THESEUS 需要 10 ms。缺失数据迭代方法的迭代次数是 7 次,而 THESEUS 的迭代次数是 37 次; 2) 蛋白质 d2pcbb 和 d1lfma 的比较,运行的时间是 4.7 ms,而 THESEUS 需要的时间是 10 ms。缺失数据迭代方法的迭代次数是 4 次,而 THESEUS 的迭代次数是 17 次; 3) 蛋白质 d1cih 和 d2pcbb 的比较,缺失数据迭代方法算法运行的时间是 3.6 ms,而 THESEUS 是 10 ms。缺失数据迭代方法的迭代次数是 6 次,而 THESEUS 的迭代次数是 27 次; 4) 蛋白质 d2pcbb 和 d1cih 的比较,缺失数据迭代方法运行的时间是 5.4 ms,而 THESEUS 需要的时间是 10 ms。缺失数据迭代方法的迭代次数是 8 次,而 THESEUS 的迭代次数是 27 次。从最后两个蛋白质的比较得出,同样的蛋白质相比较,模板选择不同,比较的时间和迭代次数也不一样。

2.3　与 PSSM 方法比较

Table 2 和 Table 3 是用缺失数据迭代方法和已经发表的 PSSM [17] 的方法相比较。其中,Table 1 是 C_α 碳原子的比较; Table 2 是全部主链原子的比较结果。Table 2 第一行表明: 蛋白质 d1cih 和 d1lfma 用缺失数据的迭代方法所花费的时间是 0.009 s,RMSD 是 0.5 Å; 而用 PSSM 算法的运行时间是 0.331 s,RMSD 是 0.6 Å。Table 2 的第 2 ~ 5 行表明用缺失数据下的迭代方法所花费的时间比 PSSM 方法短,RMSD 值相对 PSSM 方法比较小。然而,缺失数据下的迭代算法需要蛋白质的序列比对,而 PSSM 算法不需要序列的比对。Table 2 和 Table 3 说明缺失数据下的迭代算法和 PSSM 算法各有优缺点。当

蛋白质的序列比对难以进行时，可以用 PSSM 算法比较，这时所花费的时间和 RMSD 值相对较大。当能进行蛋白质序列比对时，可以用缺失数据下的迭代算法。

Table 1 The comparison of iterative method with THESEUS for the same error

Proteins names	ITEMDM time	THESEUS time	ITEMDM ite	THESEUS ite
d1cih and d1lfma	5.4 ms	10 ms	7	37
d2pcbb and d1lfma	4.7 ms	10 ms	4	17
d1cih and d2pcbb	3.6 ms	10 ms	6	27
d2pcbb and d1cih	5.4 ms	10 ms	8	27

ite: number of iterations

Table 2 The comparison of iterative method with PSSM method for missing data

Structure name id1(size) -id2(size)	Time (s)		RMSD (Å)	
	ITEMDM	PSSM	ITEMDM	PSSM
d1cih (108) -d1lfma (103)	0.009	0.3	0.5	0.6
d1cih (108) -d2pcbb (104)	0.027	9.4	0.7	0.8
d1cih (108) -d1m60a (104)	0.076	21.5	1.1	1.2
d2pcbb (104) -d1m60a (104)	0.097	7.4	1.2	1.3
d1cih (108) -d1kyow (108)	0.022	3.8	0.6	0.7

Table 3 The comparison of iterative method with PSSM method for missing data

Structure name id1(size) -id2(size)	Time (s)		RMSD (Å)	
	ITEMDM	PSSM	ITEMDM	PSSM
d1cih(835) -d1crj(847)	0.089	1.9	0.5	0.5
d1cih(835) -d1csu(846)	0.079	1.9	0.5	0.5
d1cih(835) -d1csx(846)	0.092	2.2	0.4	0.6
d1cih(835) -d1yeb(847)	0.088	2.9	0.8	0.8
d1cih(835) -d1u74d(847)	1.642	495.7	1.0	2.0

Fig. 2 表明，用缺失数据迭代方法旋转前和旋转后的蛋白质的位置。其中，Fig. 2A 是 d1cih 和 d1lfma 旋转前的位置，Fig. 2B 是 d1cih 和 d1lfma 叠加后的位置。Fig. 2C 是 d2pcbb 和 d1u74d 旋转前的位置，Fig. 2D 是 d2pcbb 和 d1u74d 叠加后的位置。

2.4 两种方法对 Fischer's 数据库数据处理结果的比较

Table 4 是该方法运用于标准 Fischer's 数据库(67 对)的两两叠加 RMSD 结果与 PSSM 方法的比较。

Table 4 显示，缺失数据迭代方法计算的总体 RMSD 值相对较小，而 PSSM 方法计算得到的 RMSD 值相对较大。这是由于 PSSM 方法不需要序列比对，而缺失数据迭代方法需要序列比对，已将没有比对上的氨基酸残基删除，所以 RMSD 值相对较小。两种方法各有优缺点。

本文提出一种有效的缺失数据下的蛋白质结构叠加的迭代算法，主要采用缺失的数据部分的反复迭代，该方法需要蛋白质的序列比对信息。数值实验表明对于同源蛋白质数据，缺失数据迭代算法比 THESEUS 算法快，而且迭代次数少。该算法的准确性比 PSSM 算法高，运行的时间短，但是需要序列比对信息。总之，数值试验表明该方法是有效、准确、稳定。有如下的优点：1) 运算结果较准确。2) 需要时间较少。3) 能处理缺失数据的蛋白质叠加。

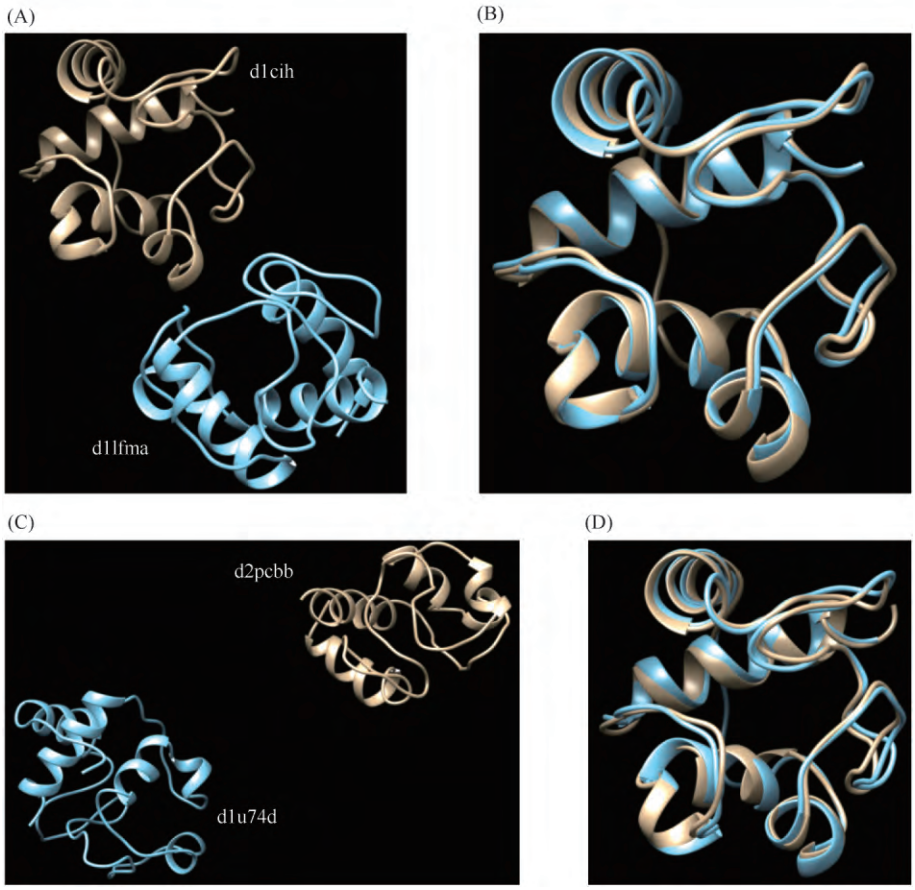

Fig. 2　Protein structures and positions before and after superposition by using iterative of missing data method for two sets　(A) Protein structures before superposition for d1cih and d1lfma. (B) Protein structures after superposition by using iterative of missing data method for d1cih and d1lfma. (C) Protein structures before superposition for d2pcbb and d1u74d. (D) Protein structures after superposition by using iterative of missing data method for d2pcbb and d1u74d. The method used the idea of the iterative of missing data to compute the protein superposition problem. The rotation matrix and the translation vector were obtained by using the optimized least squares algorithm combined with matrix SVD decomposition method

Table 4　The RMSD of pairwise superposition with PSSM for Fischer's dataset (67 pairs)

PDB-id1 (size)	——	PDB-id2 (size)	RMSD (Å)	
			PSSM	ITEMDM
1mdc (133)	–	1ifc (131)	1.7	0.9
1npx (447)	–	3grs (461)	2.8	1.6
1onc (103)	–	7rsa (124)	1.9	1.0
1osa (148)	–	4cpv (108)	4.1	2.2
1pfc (111)	–	3hlab (99)	2.4	1.3
2cmd (312)	–	6ldh (329)	2.4	2.2
2pna (104)	–	1shaa (103)	2.7	1.8
1bbha (262)	–	2ccya (127)	3.2	2.9
1c2ra (116)	–	1ycc (108)	2.9	1.7
1chra (370)	–	2mnr (357)	1.7	1.0
1dxtb (147)	–	1hbg (147)	1.9	1.2
2fbjl (213)	–	8fabb (214)	2.4	1.7
1gky (186)	–	3adk (194)	3.7	2.9
1hip (85)	–	2hipa (71)	1.6	0.8
2sas (185)	–	2scpa (348)	3.1	1.9
1fc1a (206)	–	2fb4h (229)	3.5	1.5
2hpda (457)	–	2cpp (405)	2.8	1.5

(**Continued Table 4**)

PDB-id1 (size)	——	PDB-id2 (size)	RMSD (Å)	
			PSSM	ITEMDM
1aba (87)	–	1ego (85)	2.8	1.9
1eaf (243)	–	4cla (213)	2.7	2.1
2sga (181)	–	5ptp (223)	2.9	1.8
2hhma (278)	–	1fbpa (316)	3.2	1.8
1aaj (105)	–	1paz (120)	2.8	1.5
5fd1 (106)	–	1iqz (81)	2.7	1.5
1isua (62)	–	2hipa (71)	2.7	1.9
1gal (581)	–	3cox (500)	3.2	1.6
1caub (184)	–	1caua (181)	3.6	3.1
1hom (68)	–	1lfb (77)	3.3	2.0
1tlk (103)	–	2rhe (114)	2.9	1.7
2omf (340)	–	2por (301)	3.2	1.6
1lgaa (343)	–	2cyp (293)	3.0	1.5
4sbva (199)	–	2tbva (287)	3.3	1.7
8ilb (146)	–	4fgf (124)	2.7	2.3
1hrha (125)	–	1rnh (151)	3.1	1.8
1mup (157)	–	1rbp (174)	3.3	1.5
1cpcl (172)	–	1cola (197)	3.3	1.7
2ak3a (226)	–	1gky (186)	3.7	1.4
1atna (372)	–	1atr (383)	4.1	3.5
1arb (263)	–	5ptp (223)	3.0	1.6
2pia (321)	–	1fnb (296)	2.7	1.3
3rubl (441)	–	6xia (387)	4.1	3.4
2sara (96)	–	9rnt (104)	2.9	1.4
3cd4 (178)	–	2rhe (114)	2.2	1.0
1aep (153)	–	256ba (106)	2.3	1.2
2mnr (357)	–	4enl (436)	2.3	1.3
1ltsd (103)	–	2xsc (69)	2.9	1.4
2gbp (309)	–	2liv (344)	2.4	1.2
1bbt (186)	–	2plv (288)	4.0	3.6
2mtac (147)	–	1ycc (108)	2.3	1.1
1taha (318)	–	1tca (317)	3.3	1.6
1rcb (129)	–	2gmfa (121)	3.1	2.1
1saca (204)	–	2ayh (214)	3.3	1.5
1dsba (188)	–	2trxa (109)	2.2	1.3
1st_ (98)	–	1mola (94)	2.6	1.4
2afna (331)	–	1aoza (552)	3.2	1.8
1fxia (96)	–	1ubq (76)	2.8	1.9
1bgeb (159)	–	2gmfa (121)	2.9	1.3
3hlab (99)	–	2rhe (114)	3.2	2.7
3chy (128)	–	2fox (138)	3.0	1.5
2azaa (129)	–	1paz (120)	2.8	1.1
1cew (108)	–	1mola (94)	2.9	1.8
1cid (177)	–	2rhe (114)	3.0	1.5
1crl (534)	–	1ede (310)	3.1	1.9
2sim (381)	–	1nsba (390)	3.4	2.8
1ten (89)	–	3hhrb (195)	2.9	1.4
1tie (166)	–	4fgf (124)	2.8	1.5
2snv (151)	–	5ptp (223)	2.8	1.2
1gp1a (432)	–	2trxa (109)	2.7	1.9

参考文献(References)

[1] Theobald DL, Steindel P A. Optimal simultaneous superpositioning of multiple structures with missing data [J]. Bioinformatics, 2012, **28**(15): 1972-1979

[2] Flower DR. Rotational superposition: a review of methods [J]. J Mol Graph Model, 1999, **17**(3-4): 238-244

[3] Diamond R. On the comparison of conformations using linear and quadratic transformations [J]. Acta Crystallogr, 1976, A32: 1-10

[4] Irving JA, Whisstock JC, Lesk AM. Protein structural alignments and functional genomics [J]. Proteins, 2001, **42**(3): 378-382

[5] Edgar RC, Batzoglou S. Multiple sequence alignment [J]. Curr Opin Struct Biol, 2006, **16**(3): 368-373

[6] Dunbrack RL Jr. Sequence comparison and protein structure prediction [J]. Curr Opin Struct Biol, 2006, **16**(3): 374-384

[7] Panchenko A, Marchler-Bauer A, Bryant SH. Threading with explicit models for evolutionary conservation of structure and sequence [J]. Proteins, 1999, Suppl **3**: 133-140

[8] Guda C, Lu S, Scheeff ED, *et al*. CE-MC: a multiple protein structure alignment server [J]. Nucleic Acids Res, 2004, **32**(Web Server Issue): W100-103

[9] Menke M, Berger B, Cowen L. Matt: Local flexibility aids protein multiple structure alignment [J]. PLoS Comput Biol, 2008, **4**(1): e10

[10] Leibowitz N, Nussinov R, Wolfson HJ. MUSTA-a general, efficient, automated method for multiple structure alignment and detection of common motifs: application to proteins [J]. J Comput Biol, 2001, **8**(2): 93-121

[11] Dror O, Benyamini H, Nussinov R, *et al*. Multiple structural alignment by secondary structures: algorithm and applications [J]. Protein Sci, 2003, **12**(11): 2492-2507

[12] Ye Y, Godzik A. Multiple flexible structure alignment using partial order graphs [J]. Bioinformatics, 2005, **21**(10): 2362-2369

[13] Birzele F, Gewehr JE, Csaba G, *et al*. Vorolign — fast structural alignment using Voronoi contacts [J]. Bioinformatics, 2007, **23**(2): e205-211

[14] Diamond R. On the multiple simultaneous superposition of molecular structures by rigid body transformations [J]. Protein Sci, 1992, **1**(10): 1279-1287

[15] Flower DR. Rotational superposition: a review of methods [J]. J Mol Graph Model, 1999, **17**(3-4): 238-244

[16] Theobald DL, Wuttke DS. Empirical Bayes hierarchical models for regularizing maximum likelihood estimation in the matrix Gaussian Procrustes problem [J]. Proc Natl Acad Sci U S A, 2006, **103**(49): 18521-18527

[17] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm [J]. J Roy Stat Soc, 1977, **39**(1): 1-38

[18] 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

[19] Goldstein H. Classical Mechanics [M]. Boston: Addison-Wesley, 1965

[20] Evans DJ. On the representation of orientation space [J]. Mol Phys, 1977, **34**(2): 317-325

[21] Gene H, Golub Charles F. Van Loan. Matrix Computations (Johns Hopkins Studies in Mathematical Sciences), 3rd Edition [M]. Baltimore: The Johns Hopkins University Press, 2011

[22] Katoh K, Kuma K, Toh H, *et al*. MAFFT version 5: improvement in accuracy of multiple sequence alignment [J]. Nucleic Acids Res, 2005, **33**(2): 511-518

[23] Larkin MA, Blackshields G, Brown NP, *et al*. Clustal W and Clustal X version 2.0 [J]. Bioinformatics, 2007, **23**(21): 2947-2948

[24] Corpet F. Multiple sequence alignment with hierarchical clustering [J]. Nucleic Acids Res, 1988, **16**(22): 10881-10890

[25] Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment [J]. J Mol Biol, 2000, **302**(1): 205-217