PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2022, 49(6): 1155~1164
www.pibb.ac.cn

A Comparative Evaluation of Several Matrix Completion Algorithms for Protein Structure Determination*

LI Zhi-Cheng[1,2)**], WEI Xian[3)], LI Jin-Ting[1,2)]

([1)]*Department of Physics, Taiyuan Normal University, Jinzhong* 030619, China;
[2)]*Institute of Computational and Applied Physics, Taiyuan Normal University, Jinzhong* 030619, China;
[3)]*Department of Science, Taiyuan Institute of Technology, Taiyuan* 030008, China)

Abstract **Objective** Nowadays, how to determine an accurate three-dimensional protein structure from nuclear magnetic resonance (NMR) spectroscopy experiments is a hot topic in biophysics, because understanding the spatial structure of a protein is crucial to research its function. However, this is a large challenge due to the serious lack of experimental data. **Methods** In this paper, the problem of protein structure determination was solved by matrix completion (MC) algorithms of recovering a distance matrix. Firstly, the initial distance matrix model was established, then its missing data were recovered by the MC algorithms at different sampling ratios. The subsequent stage involved adding the noise model to evaluate the noise resistance of the algorithms. Four proteins with different topological structures and 6 off-the-shelf MC algorithms were selected for testing. **Results** The results show that these algorithms have good performance in a certain range of sampling ratios and noises. More specifically, the advantages of different algorithms in the case of accurate sampling and noisy sampling are compared by analyzing the average and standard deviation of the root-mean-square deviation (RMSD) and computational time, which are two important indexes about algorithms. **Conclusion** We can conclude that 6 different MC algorithms have different performances and advantages for the problem of protein structure determination. These characteristics provide a basis for the development of a new MC algorithm. The results of this paper have potential promotion in the field of protein research based on MC algorithms.

Key words protein structure determination, distance matrix, matrix completion, noise resistance
DOI: 10.16476/j.pibb.2021.0278

Proteins are composed of ordered amino acid chains, which are vital macromolecules of cells in organisms. Determining the three-dimensional structure of proteins is central to biophysics and bioinformatics because it is important to understand the physical, chemical and biological properties of proteins and to analyze possible interactions with other proteins[1]. X-ray diffraction (XRD) crystallography was the main tool for obtaining protein information in the early period of protein structure determination[2]. However, the introduction of the nuclear magnetic resonance (NMR) technique is a breakthrough because NMR made it possible to obtain protein information in an aqueous environment much closer to the native state of a protein[3-4]. The protein NMR method conventionally involves sample preparation, peak picking, spectral assignment, nuclear Overhauser effect spectroscopy (NOESY)[5-6] assignment, structure calculation and refinement[7] as demonstrated in Figure 1. Because long distances (>5 Å, 1 Å=10^{-10} m) are difficult to be measured by NMR experiment[8], lacking sufficient distance information is the main challenge for this problem. To this end, the protein NMR methods rely profoundly on complex computational algorithms and techniques, *e.g.* distance geometry[9-12], molecular dynamics[13-15].

The proposed matrix completion (MC)

* This work was supported by grants from Scientific and Technological Innovation Programs (STIP) of Higher Education Institutions in Shanxi (2020L0513) and the Shanxi Province Science Foundation for Youths (202103021223328).

** Corresponding author.
Tel: 86-18101203596, E-mail: lizc@tynu.edu.cn

Received: September 17, 2021 Accepted: January 17, 2022

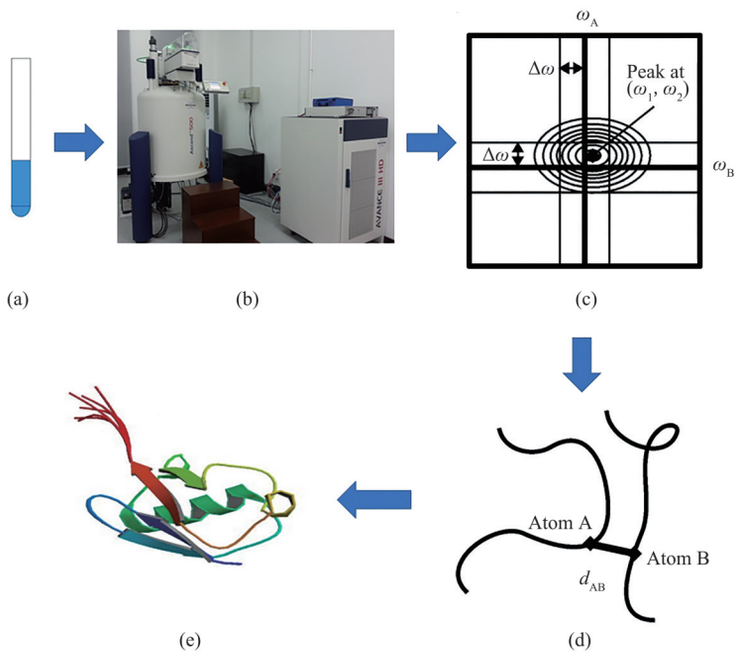

Fig. 1　The procedure of protein NMR structure determination

(a) Sample preparation: NMR experiments can directly measure protein samples in solution state; (b) NMR experiments: involving peak picking, spectral assignment, nuclear Overhauser effect spectroscopy (NOESY) assignment; (c) NMR spectroscopy: the sources of experimental data; (d) Distance constraints: obtaining a set of distance constraints from spectrums (the intensity of an NOE is the 6th power inversely proportional to the distance between two nuclei); (e) Structural calculation: the resulting geometric restraints are used as input for the structural calculation.

theory[16-17] provides a promising way to solve this problem. MC aims at recovering a low-rank matrix from a partial sampling of its entries. In the early stage of MC development, one typical example is the famous Netflix problem, which aims to predict the user's preferences for different types of movies based on a very sparse existing data set. In 2009, Candès and Recht[18] proposed that a matrix could be recovered with a very high probability if the number m of sampled entries obeys $m \geqslant Cn^{1.2}r\log n$, where C is a positive constant, n and r are dimension and rank of the matrix, respectively. Subsequently, Candès and Tao[19] improved this result to $m \geqslant Cnr\log n$. Afterward Gross[20] generalized the standard MC problem by proposing a simpler and more general method. In the meantime, the MC problem with noise was also proposed. Candès and Plan[21] proposed that a matrix could be recovered accurately with Gaussian random noise and bounded noise if m obeys $m \geqslant Cnr\log^6 n$. With the development of MC theory, it has received increasing interest and has been applied to various fields, such as protein structure calculation[22-25], image processing[26-27], traffic sensing[28].

In this paper, the protein NMR structure determination is addressed as an MC problem. In our previous work, the MC-based accelerated proximal gradient (APG) [29] algorithm and scaled alternating steepest descent (ScaledASD)[26] algorithm have been investigated in recent years. In 2017, we applied the APG algorithm to protein structure calculation and demonstrated the effectiveness of the algorithm by analyzing the accuracy and error of the calculation results[22]. In 2019, a new algorithm, ScaledASD, originally applied in image processing, was tested for protein structure calculation, the results show the algorithm overcomes the shortcomings of insufficient NMR data to a certain extent[23]. To further explore the effectiveness of MC in the field of protein structure estimation, 6 MC algorithms were selected to be tested and compared their performance in this paper. The remainder of the paper is organized as follows: Section **1** gives a detailed description of our method including the problem model, MC and quality assessment. In Section **2**, we evaluate the performance of 6 different MC algorithms under accurate sampling as well as noisy sampling, and make a detailed

analysis. Finally, the conclusions and some perspectives are elaborated in Section **3**.

1 Methods

1.1 Problem model

For a protein molecule, each atom can be considered as a three-dimensional point in space. We assume the target protein has n atoms, then the protein structure consists of a set of points $\{x_1, x_2, \cdots, x_n\}$, $x_i \in \mathrm{R}^3$. The coordinate matrix is defined as $X=[x_1; x_2; \cdots; x_n] \in \mathrm{R}^{n\times 3}$. Based on this, we can define a Euclidean distance matrix (EDM), whose elements stand for the distance between two atoms, such that

$$D_{ij} = \left\| x_i - x_j \right\|^2, \forall i,j \in \{1,2,\cdots,n\} \quad (1)$$

where $\| x \|$ is the Euclidean distance norm of vector x. Obviously, we can transform a coordinate matrix X into a Euclidean distance matrix D according to Equation (1). Conversely, a Euclidean distance matrix D can also be converted back into a coordinate matrix X by a general approach[30]. In detail, we induce the Gram matrix $G = XX^T$. Gram matrix G and distance matrix D have the following transformation relations:

$$G = -\frac{1}{2} HDH \quad (2)$$

where $H = I - \frac{1}{n}11^T$, I and 1 stand for the unit matrix and the all-ones vector, respectively. Consider the eigenvalue decomposition of G:

$$G = V\Lambda V^T \quad (3)$$

where V is an $n\times n$ square matrix, and Λ is a diagonal matrix in which the elements on the diagonal are the corresponding eigenvalues. Then the coordinate matrix corresponding to the protein structure is calculated as:

$$X = V\Lambda^{1/2} \quad (4)$$

That is to say, a protein structure can be determined as long as the complete distance matrix is known.

Although we can gain some partial short distances from NMR experiments and some covalent bond lengths information[31], the distance data are still too sparse to determine a protein structure. To ameliorate this situation, MC algorithms are used to recover the incomplete initial distance matrix. Taking into account the condition of the uniform sampling distribution, we will sample the remaining distances randomly. Once the distance matrix is recovered, the protein structure is determined in light of the previous discussion. The proposed framework is shown in Figure 2.

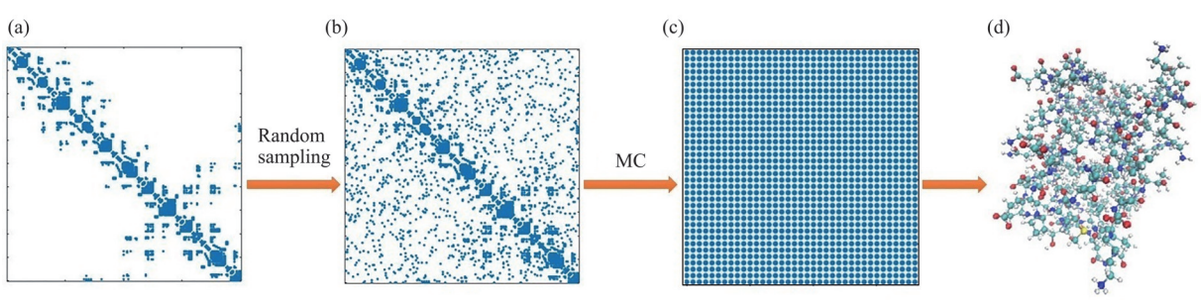

Fig. 2 The schematic diagram of protein structure determination based on MC

(a) The contour of the distance matrix containing only known short distances, including the atomic distances obtained from NMR experiments and covalent bond lengths; (b) The contour of the distance matrix after random sampling; (c) The contour of the complete distance matrix recovered by MC; (d) Transforming the complete distance matrix into a three-dimensional protein structure.

1.2 Matrix completion （MC）

As explained previously, MC is the problem of recovering a low-rank matrix from partial entries. A direct approach is to find a matrix D with the minimum rank that best approximates the underlying matrix D^0:

$$\text{minimize rank}(D), \quad \text{subjecte to } P_\Omega(D) = P_\Omega(D^0) \quad (5)$$

where Ω is a set of the indices for known elements, P_Ω denotes the sampling operator restricted to the entries indexed by Ω, that is, D has the same elements as D_0 for the entries in Ω. Solving the problem (5) is challenging because rank minimization is non-convex and generally NP-hard[18]. A convex and tractable approach is proposed to replace the rank objective

with the nuclear norm[32], then the problem (5) can be approximated by the following formulation:

$$\text{minimize } \|D\|_*, \text{ subject to } P_\Omega(D) = P_\Omega(D^0), \quad (6)$$

where $\|D\|_*$ denotes the nuclear norm of matrix D, namely, the sum of its singular values. To enhance the anti-noise performance of the problem, an alternative and stable approach is used by relaxing the equality constraint:

$$\text{minimize } \frac{1}{2}\left\|P_\Omega(D) - P_\Omega(D^0)\right\|_F^2 + \lambda\|D\|_* \quad (7)$$

where $\|\cdot\|_F$ denotes the Frobenius norm of the matrix, λ is a parameter that controls the rank of matrix D. In this process, it is necessary to compute a singular value decomposition (SVD)[33] in each iteration. There have been many algorithms proposed for the problem (7), such as the accelerated proximal gradient (APG) algorithm[29], the hard thresholding algorithms[17] and the scaled gradients on Grassmann manifolds (ScGrassMC) method[34].

Taking into account the computational complexity of SVD, an alternative method based on matrix factorization was proposed. In detail, for an n-dimensional distance matrix D, it can be written into a simple factorization form: $D=XY$ where $X\in\mathrm{R}^{n\times r}$ and $Y\in\mathrm{R}^{r\times n}$ (r is the rank of the matrix). Then the problem is transformed into solving the minimization of the following function:

$$\text{minimize } \frac{1}{2}\left\|P_\Omega(D^0) - P_\Omega(XY)\right\|_F^2 \quad (8)$$

Solving the problem (8) generally uses an alternating minimization approach, which is widely used for optimization problems. The algorithms based on the matrix factorization model include low-rank matrix fitting (LMaFit) algorithm[35], alternating steepest descent (ASD) algorithm and its scaled variant (ScaledASD)[26].

1.3 Quality assessment

In the field of structural biology, the accuracy of a molecular conformation is generally measured by the root-mean-square deviation (RMSD), which is a measure of the "average" deviation between the computed structure and the reference structure. Assume X denotes the computed configuration optimally aligned to the reference configuration X^* by the alignment procedure[36], then the RMSD is defined by the following formula:

$$\mathrm{RMSD} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\|x_i - x_i^*\right\|^2 \quad (9)$$

where n is the number of atoms, x_i denotes the coordinate of the ith atom. A more accurate structure corresponds to a smaller RMSD value. Typically, the RMSD value less than 2 Å represents a high-resolution model[37].

2 Results and discussion

In this paper, we studied the protein structure determination problem using 6 MC algorithms, listed in Table 1. Four proteins with different topological structures were selected as the target protein for testing in Table 2. For simplicity, we just took 1G6J as an example in this section. For the other proteins, similar results are obtained in **Supplementary**. All the tests in our work were carried out on a Windows 10 PC with a 3.1 GHz Intel Core i9-9900 CPU and 32 GB of memory. In our test, the information of the initial distance matrix includes short distances (less than 5 Å) between hydrogen atoms and covalent bond lengths. To reach the sampling conditions of MC, the remaining distances need to be randomly sampled. For testing purposes, we firstly presume the sampling distances are accurate, and the case of the sampling distances with noise is discussed in **2.2**.

Table 1 List of MC algorithms evaluated in this paper

Algorithm	Main techniques	References
APG	Accelerated Proximal Gradient	[29]
NIHT	Iterative Hard Thresholding	[17]
ScGrassMC	Grassmannian Manifolds	[34]
LMaFit	Low-rank Matrix Fitting	[35]
ASD	Alternating Steepest Descent	[26]
ScaledASD	Scaled Alternating Steepest Descent	[26]

Table 2　The information of four test proteins

PDB ID	Description	Topology	Atoms	Residues	References
1G6J	Ubiquitin	α+β	1 228	76	[38]
2M5Z	Antimicrobial protein	α	762	44	[39]
1B4R	PKD domain 1 from Liver fatty	β	1 114	87	[40]
1CN7	Ribosomal protein L30	α/β	1 648	105	[41]

2.1　Results for 6 MC algorithms under accurate sampling

To evaluate the performance of the aforementioned algorithms, we calculate the RMSD related to all atoms between the reconstructed structure and the reference structure in the case of accurate sampling under the sampling ratios ranging from 1% to 10%. The corresponding Protein Data Bank (PDB) [42] model is selected as the reference structure. For all the aforementioned algorithms, the number of the maximum iteration and the relative residual tolerance is set to 500 and 10^{-5}, respectively. For each algorithm and each sampling ratio, we calculated 100 times randomly and recorded the average value and standard deviation of the RMSD and the computational time. The standard deviation can reflect the stability of calculation results statistically, namely, a smaller standard deviation indicates a more stable calculation result. The results are shown in Figure 3.

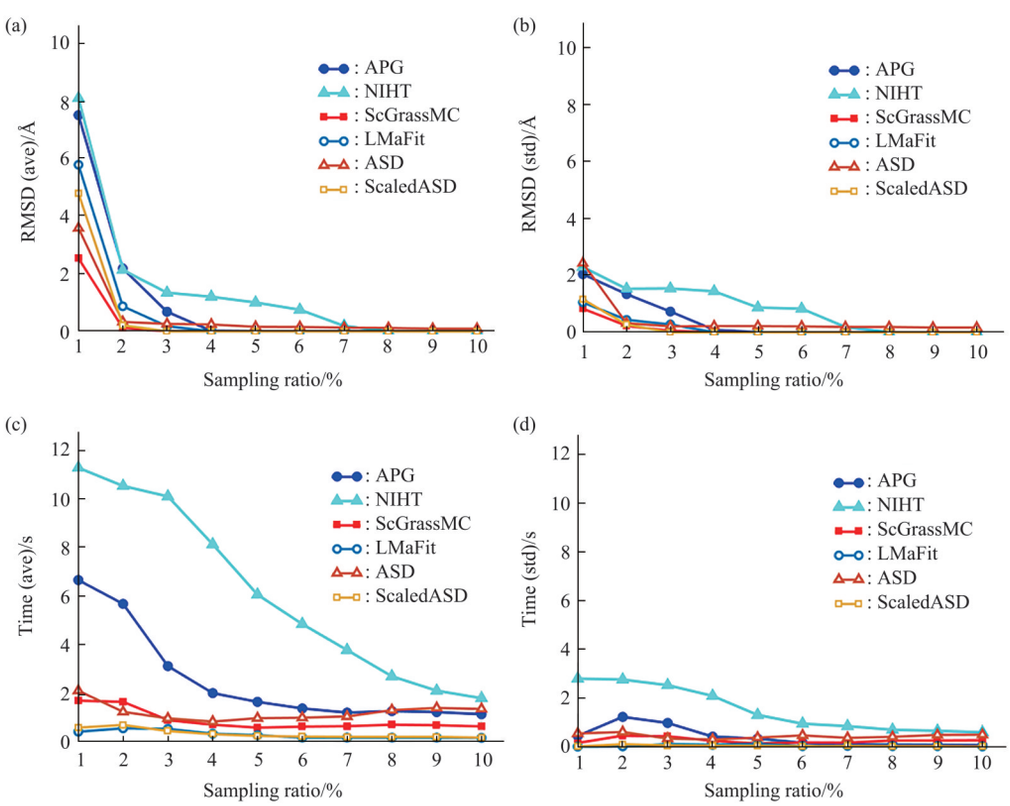

Fig. 3　The performance curve of 6 algorithms with different sampling ratios ranging from 1% to 10% for testing 1G6J

(a) The average RMSD value, denoted by RMSD (ave); (b) The standard deviation of RMSD, denoted by RMSD (std); (c) The average computational time, denoted by Time (ave); (d) The standard deviation of computational time, denoted by Time (std).

Figure 3 shows the visual comparison of the aforementioned 6 algorithms applied to protein structure determination under precise sampling. For comparison purposes, we adjusted the scale range of the average value and the standard deviation to be consistent. Figure 3a shows the RMSDs (ave) for 6 algorithms. Naturally, with the increase of sampling ratio, the RMSDs (ave) become smaller, indicating that the calculational accuracy of algorithms becomes higher. Although the RMSDs (ave) are relatively larger with low sampling ratios, when the sampling ratio exceeds 3%, almost RMSDs (ave) are below 2 Å for all the algorithms (except the NIHT for 1B4R shown in Figure S1), that is to say, high accuracy can be obtained. Notably, when the sampling ratio exceeds 7%, the RMSD values are almost close to 0, indicating that the recovered structures are high-resolution extremely. As can be seen in Figure 3b, the RMSD (std) tends to decrease as the sampling ratio increases. Especially under high sampling ratios (more than 7%), all the RMSDs (std) are almost close to 0, that is, the calculation of the algorithms turns to be more stable under high sampling ratios. According to Figure 3a, b, ScGrassMC is slightly more prominent than other algorithms, because its RMSDs (ave) and RMSDs (std) are relatively lower in most sampling ratios. Figure 3c shows the average computational time, by and large, the calculation costs less time with the increase of sampling ratio for all the algorithms. This seems to be natural because the scarcity of initial data can lead to more computations. NIHT and APG cost relatively more time, especially under low sampling ratios, while the computational times are significantly reduced under high sampling ratios. In Figure 3d, we can see that the Times (std) have lower values with high sampling ratios. On the whole, LMaFit and ScaledASD are more prominent in computational time because they cost less computational time and are more stable. Similar results with different sampling ratios for testing 1B4R, 2M5Z, 1CN7 are shown in Figure S1−S3, respectively.

2.2 Results for 6 MC algorithms under noisy sampling

In practical molecular conformation problems, the distance information tends to bring some noise, so it is instructive to consider the noise resistance ability of the algorithm in protein structure determination. To evaluate the anti-noise performance of the aforementioned algorithms, different levels of noise are added to the sampling distances. In this paper, we have tested with a “normal” noise model akin to Reference [43], *i.e.*, the noisy distances $\bar{d}_{ij}$ are given by:

$$\bar{d}_{ij} = \left(1 + \sigma z_{ij}\right) d_{ij} \tag{10}$$

where σ is a positive parameter, called noise factor, the value of z_{ij} accords with standard normal distribution N(0, 1), d_{ij} is the accurate sampling distance. Obviously, the noise level can be controlled by the noise factor σ, the noise becomes larger as the value of σ increases. In this section, the sampling ratio is set to 10%, and each test is still calculated 100 times randomly.

We can see in Figure 4a, the RMSDs (ave) become larger and larger following the increasing noise factor. All the algorithms are still able to produce a fairly accurate structure (RMSD<2 Å) when the noise factor is less than 15%. Meanwhile, within this noise range, the RMSDs (std) are also relatively stable according to Figure 4b. If the noise factor exceeds 15%, the RMSDs (std) become larger, that is, the stability of calculational accuracy is worse in high noise situations. Figure 4c shows the average computational time under different levels of noise. It is not immediately obvious that the computational time and noise have an explicit correlation. However, ScGrassMC and NIHT cost more computational time than other algorithms. Compared with Figure 3c, the computational time of the two algorithms is more sensitive to noise. We infer that the noisy distance data increases the number of iterations of calculation, which makes the computational time longer. Taking the ScGrassMC algorithm with a sampling ratio of 10% as an example, when there is no noise, the average number of iterations of calculation is 34, while when the noise factor reaches 10%, the average value of the number of iterations is 500, which is the maximum iteration set in our test. From Figure 4d, for most algorithms, Times (std) are small, indicating that the computational time is relatively stable. In comparison, the Times (std) of NIHT are slightly larger, and the Times (std) of ScGrassMC are larger only in high noise. Overall, in the case of noisy sampling distances, ASD and ScaledASD perform

superior both in terms of computational accuracy and computational time. Similar results in the case of sampling distances corrupted different levels of noise for testing 2M52, 1B4R, 1CN7 are shown in Figure S4−S6 respectively.

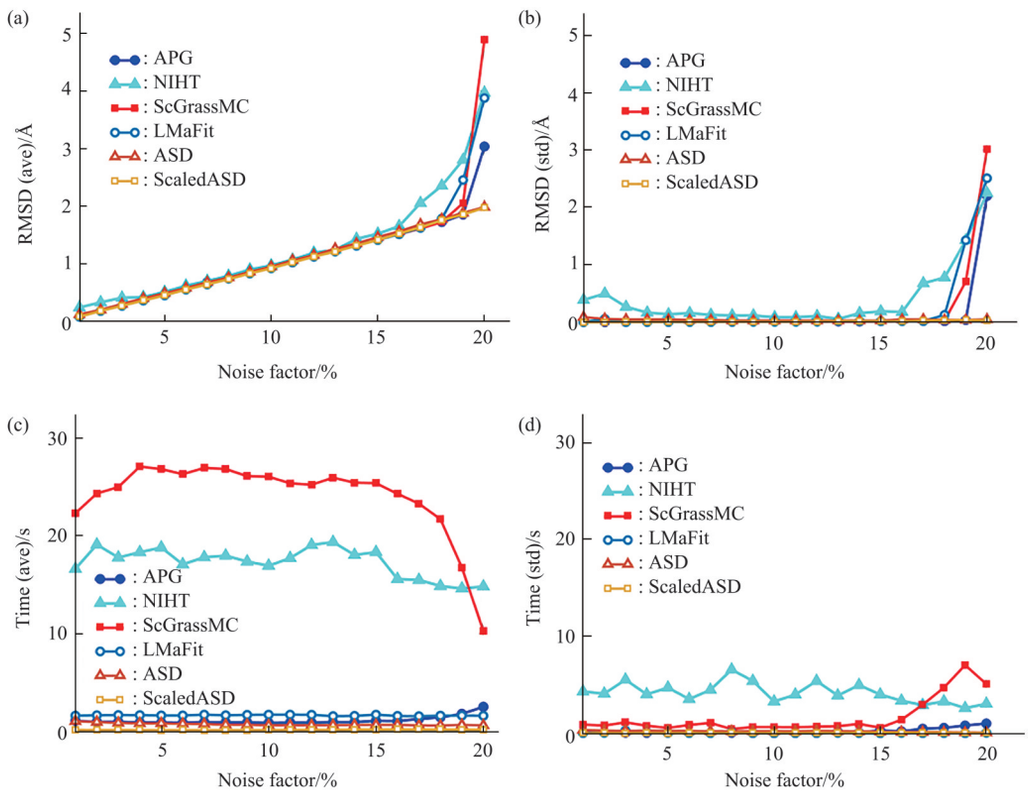

Fig. 4　The performance curve of six algorithms in the case of sampling distances corrupted different levels of noise for testing 1G6J

(a) The average RMSD value, denoted by RMSD (ave); (b) The standard deviation of RMSD, denoted by RMSD (std); (c) The average computational time, denoted by Time (ave); (d) The standard deviation of computational time, denoted by Time (std).

3　Conclusion and future work

In this paper, we have evaluated 6 MC algorithms applied to the protein structure determination problem. MC algorithms can effectively overcome the shortcomings of insufficient NMR experimental data. We took 4 proteins with different topological structures as examples to test the accuracy, computational time and noise resistance of these algorithms. Our test results show that the algorithms perform higher accuracy and shorter computational time with the increase of sampling ratio, and the stability of the algorithms is also higher. Especially, when the sampling ratio exceeds 3%, almost RMSDs are less than 2 Å, and when the sampling ratio exceeds 7%, the RMSDs are close to 0, which shows that all these algorithms can efficiently generate an accurate structure. A conclusion can be made that the algorithms perform remarkably well when there are enough exact distance data. By comparison, the ScGrassMC algorithm performs better in terms of computational accuracy, while LMaFit and ScaledASD are more advantageous in terms of computational time. Subsequently, we tested the noise resistance of the 6 algorithms by building the normal noise model. The computational accuracy decreases with the increase of noise, while when the noise factor is less than 15%, almost RMSDs are less than 2 Å, indicating all algorithms can produce relatively accurate structures in this situation. An interesting conclusion is that the NIHT and ScGrassMC algorithms cost significantly more computational time in the presence of noise,

indicating that the computational times of both algorithms are sensitive to noise. ASD and ScaledASD have better noise resistance performance, both in terms of computational accuracy and computational time. The results of this paper give us a degree of confidence that the MC algorithms are promising in the field of protein structure determination. In the future, a further study on algorithmic mechanisms is our next work, which would help us to establish a greater degree of accuracy and efficiency in this field. We also believe that the results of this paper can potentially promote the development of more effective new MC algorithms in the future.

Supplementary PIBB_20210278_Doc_S1. pdf is available online (http://www.pibb.ac.cn or http://www.cnki.net).

References

[1] Ma C, Yang D, Jiang Y, *et al*. The identification of intrinsically disordered proteins and their structural, functional, evolutionary features. Prog Biochem Biophys, 2015, **42**(1): 16-24
马冲, 杨冬, 姜颖, 等. 生物化学与生物物理进展, 2015, **42**(1): 16-24

[2] Kendrew J C. Architecture of a protein molecule. Nature, 1958, **182**(4638):764-767

[3] Williamson M P, Havel T F, Wüthrich K. Solution conformation of protein inhibitor IIA from bull seminal plasma by ^{1}H nuclear magnetic resonance and distance geometry. J Mol Biol, 1985, **182**(2): 295-315

[4] Dyson H J, Wright P E. Insights into the structure and dynamics of unfolded proteins from nuclear magnetic resonance. Adv Protein Chem, 2002, **62**(4): 311-340

[5] Pitner T P, Glickson J D, Dadok J, *et al*. Solvent exposure of specific nuclei of angiotensin II determined by NMR solvent saturation method. Nature, 1974, **250**(467): 582-584

[6] Kumar A, Ernst R R, Wüthrich K. A two-dimensional nuclear Overhauser enhancement (2D NOE) experiment for the elucidation of complete proton-proton cross-relaxation networks in biological macromolecules. Biochem Biophys Res Commun, 1980, **95**(1): 1-6

[7] Güntert P. Automated structure determination from NMR spectra. Eur Biophys J, 2009, **38**: 129-143

[8] Güntert P. Structure calculation of biological macromolecules from NMR data. Q Rev Biophys, 1998, **31**(2): 145-237

[9] Braun W, Bösch C, Brown L R, *et al*. Combined use of proton-proton Overhauser enhancements and a distance geometry algorithm for determination of polypeptide conformations. Biochim Biophys Acta, 1981, **667**(2): 377-396

[10] Moré J J, Wu Z. Global continuation for distance geometry problems. SIAM J Optim, 1997, **7**(3): 814-836

[11] Liberti L, Lavor C, Mucherino A, *et al*. Molecular distance geometry methods: from continuous to discrete. Int Trans Oper Res, 2010, **18**(1): 33-51

[12] Liberti L, Lavor C, Maculan N, *et al*. Euclidean distance geometry and applications. SIAM Rev, 2014, **56**(1): 3-69

[13] McCammon J A, Gelin B R, Karplus M. Dynamics of folded proteins. Nature, 1977, **267**(5612): 585-590

[14] Kaptein R, Zuiderweg E R, Scheek R M, *et al*. A protein structure from nuclear magnetic resonance data. lac repressor headpiece. J Mol Biol, 1985, **182**(1): 179-182

[15] Lindorff-Larsen K, Piana S, Palmo K, *et al*. Improved side-chain torsion potentials for the Amber ff99SB protein force field. Proteins, 2010, **78**(8): 1950-1958

[16] Keshavan R H, Montanari A, Oh S. Matrix completion from a few entries. IEEE Trans Inf Theory, 2010, **56**(6): 2980-2998

[17] Tanner J, Wei K. Normalized iterative hard thresholding for matrix completion. SIAM J Sci Comput, 2013, **35**(5): S104-S125

[18] Candès E J, Recht B. Exact matrix completion *via* convex optimization. Found Comput Math, 2009, **9**(6): 717-772

[19] Candès E J, Tao T. The power of convex relaxation: near-optimal matrix completion. IEEE Trans Inf Theory, 2010, **56**(5): 2053-2080

[20] Gross D. Recovering low-rank matrices from few coefficients in any basis. IEEE Trans Inf Theory, 2011, **57**(3): 1548-1566

[21] Candès E J, Plan Y. Matrix completion with noise. Proc IEEE, 2010, **98**(6): 925-936

[22] Li Z, Li Y, Lei Q, *et al*. Protein structure estimation from NMR data by matrix completion. Eur Biophys J, 2017, **46**(6): 525-532

[23] Li Z, Li S, Wei X, *et al*. Scaled alternating steepest descent algorithm applied for protein structure determination from nuclear magnetic resonance data. J Comput Biol, 2019, **26**(9): 1020-1029

[24] Wei X, Li Z, Li S, *et al*. Protein structure determination using a Riemannian approach. FEBS Lett, 2020, **594**(6): 1036-1051

[25] Li Z, Li S, Wei X, *et al*. Recovering the missing regions in crystal structures from the nuclear magnetic resonance measurement data using matrix completion method. J Comput Biol, 2020, **27**(5): 709-717

[26] Tanner J, Wei K. Low rank matrix completion by alternating steepest descent methods. Appl Comput Harmon Anal, 2016, **40**(2): 417-429

[27] Sobral A, Zahzah E. Matrix and tensor completion algorithms for background model initialization: a comparative evaluation. Pattern Recognit Lett, 2017, **96**: 22-33

[28] Du R, Chen C, Yang B, *et al*. Effective urban traffic monitoring by vehicular sensor networks. IEEE Trans Veh Technol, 2015, **64**(1): 273-286

[29] Toh K, Yun S. An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems. PAC J Optim, 2010, **6**(3): 615-640

[30] Crippen G M, Havel T F. Stable calculation of coordinates from

distance information. Acta Crystallogr A, 1978, **34**(2): 282-284

[31] Engh R A, Huber R. Accurate bond and ample parameters for X-ray protein structure refinement. Acta Crystallogr A, 1991, **47**(4): 392-400

[32] Recht B, Fazel M, Parrilo P A. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. SIAM Rev Soc Ind Appl Math, 2010, **52**(3): 471-501

[33] Alter O, Brown P O, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci USA, 2000, **97**(18): 10101-10106

[34] Ngo T T, Saad Y. Scaled gradient on Grassmann manifolds for matrix completion. Adv Neural Inf Process Syst, 2012: 1412-1420

[35] Wen Z, Yin W, Zhang Y. Solving a low-rank factorization model for matrix completion by a non-linear successive over-relaxation algorithm. Math Prog Comp, 2012, **4**(4): 333-361

[36] Leung N Z, Toh K. An SDP-based divide-and-conquer algorithm for large-scale noisy anchor-free graph realization. SIAM J Sci Comput, 2009, **31**(6): 4351-4372

[37] Kihara D, Chen H, Yang Y D. Quality assessment of protein structure models. Curr Protein Pept Sci, 2009, **10**(3): 216-228

[38] Babu C R, Flynn P F, Wand A J. Validation of protein structure preparations of encapsulated proteins dissolved in low viscosity fluids. J Am Chem Soc, 2001, **123**(11): 2691-2692

[39] Lohans C T, Towlw K M, Miskolzie M, *et al*. Solution structures of the linear leaderless bacteriocins enterocin 7A and 7B resemble carnocyclin A, a circular antimicrobial peptide. Biochemistry, 2013, **52**(23): 3987-3994

[40] Bycroft M, Bateman A, Clarke J, *et al*. The structure of a PKD domain from polycystin-1: implications for polycystic kidney disease. EMBO J, 1999, **18**(2): 297-305

[41] Mao H, Williamson J R. Local folding coupled to RNA binding in the yeast ribosomal protein L30. J Mol Biol, 1999, **292**(2): 345-359

[42] Berman H M, Westbrook J, Feng Z, *et al*. The protein data bank. Nucleic Acids Res, 2000, **28**(1): 235-242

[43] Wang Z, Zheng S, Ye Y, *et al*. Further relaxations of the semidefinite programming approach to sensor network localization. SIAM J Optim, 2008, **19**(2): 655-673

蛋白质结构确定领域中的几种矩阵填充算法的对比评估*

李志诚[1,2)**]　韦　仙[3)]　李晋婷[1,2)]

（[1)] 太原师范学院物理系，晋中 030619；

[2)] 太原师范学院计算物理与应用物理研究所，晋中 030619；[3)] 太原工业学院理学系，太原 030008）

摘要　目的　目前，如何从核磁共振（nuclear magnetic resonance，NMR）光谱实验中准确地确定蛋白质的三维结构是生物物理学中的一个热门课题，因为蛋白质是生物体的重要组成成分，了解蛋白质的空间结构对研究其功能至关重要，然而由于实验数据的严重缺乏使其成为一个很大的挑战。**方法**　在本文中，通过恢复距离矩阵的矩阵填充（matrix completion，MC）算法来解决蛋白质结构确定问题。首先，初始距离矩阵模型被建立，由于实验数据的缺乏，此时的初始距离矩阵为不完整矩阵，随后通过MC算法恢复初始距离矩阵的缺失数据，从而获得整个蛋白质三维结构。为了进一步测试算法的性能，本文选取了4种不同拓扑结构的蛋白质和6种现有的MC算法进行了测试，探究了算法在不同的采样率以及不同程度噪声的情况下算法的恢复效果。**结果**　通过分析均方根偏差（root-mean-square deviation，RMSD）和计算时间这两个重要指标的平均值及标准差评估了算法的性能，结果显示当采样率和噪声因子控制在一定范围内时，RMSD值和标准差都能达到很小的值。另外本文更加具体地比较了不同算法的特点和优势，在精确采样情况下，ScGrassMC算法计算的精度较高，LMaFit和ScaledASD算法则在计算时间上更具优势。在抗噪性方面，ASD和ScaledASD算法表现更为突出。**结论**　本文可以得出，MC算法应用在蛋白质结构确定领域具有很好的效果，而且不同的算法在计算中具有不同的特点和优势。这些结论为新的MC算法的开发提供了参考。本文的研究结果对基于MC算法的蛋白质结构确定领域具有潜在的推动作用。

关键词　蛋白质结构确定，距离矩阵，矩阵填充，抗噪性

中图分类号　Q615，Q71　　**DOI**：10.16476/j.pibb.2021.0278

* 山西省高等学校科技创新项目（2020L0513）和山西省青年科学研究项目（202103021223328）资助。

** 通讯联系人。

Tel：18101203596，E-mail：lizc@tynu.edu.cn

收稿日期：2021-09-17，接受日期：2022-01-17