ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

IDRWalker: A Random Walk Based Tool for Generating Intrinsically Disordered Regions in Large Protein Complexes

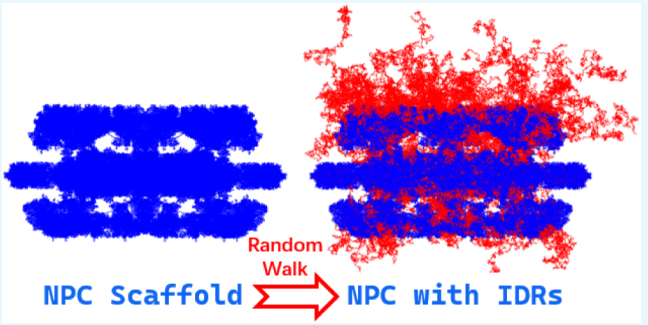
Guanglin Chen and Zhiyong Zhang*

Cite This: *ACS Omega* 2024, 9, 32059−32065

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Intrinsically disordered regions (IDRs), which may be functionally important, are common in proteins. However, the structures of IDRs are often missing due to their highly dynamic nature. In the study of IDRs, integrative modeling combining computational simulations and experimental data is a common approach, for which initial structures of the IDRs need to be built. However, applying this method to large protein complexes is challenging because existing structure generation tools are sometimes unsuitable for IDRs in large systems. To facilitate convenient and rapid structure generation of IDRs in large protein complexes, we developed a computational tool named IDRWalker based on self-avoiding random walks. Three protein complexes were used to illustrate the efficiency of the tool, and it was found that IDRs in more than 800 chains of the nuclear pore complex could be generated in minutes. These structures of large protein complexes with added IDRs can be further used to run computational simulations for integrative modeling.



INTRODUCTION

In many protein complexes, there are typically well-folded ordered structures and intrinsically disordered regions (IDRs) without stable structures. Although IDRs may also possess important biological functions,[1,2] their structures are frequently missing in protein complexes due to their highly dynamic nature and the limitations of experimental techniques. Cryo-electron microscopy (cryo-EM), which reconstructs 3D electron density maps with atomic-level resolution, is currently the most widely used technique for determining the structures of protein complexes.[3,4] In cryo-EM, high-resolution density maps are often limited to well-folded structures within a protein complex, with those of the IDRs remaining ambiguous.[5]

When high-resolution density maps are unavailable, integrative modeling combining computational simulations and experimental data has become a commonly employed approach for building structural models of large protein complexes. Using tools such as an integrative modeling platform (IMP),[6] Haddock,[7] and Assembline,[8] large protein complexes can be modeled through the following steps: first, obtaining structures of each subunit individually; and then guided by density maps and information from other experiments,[9] such as nuclear magnetic resonance (NMR),[10] cross-linking mass spectrometry (XL-MS),[11] and fluorescence resonance energy transfer (FRET),[12] the complete structure of the protein complex can be assembled via computational simulations. This approach has already achieved success in large systems such as the nuclear pore complex (NPC)[5,13] with a molecular weight of ∼120 MDa.

Integrative modeling not only facilitates the assembly of structured subunits but also enables the exploration of conformational ensembles of IDRs.[14] As mentioned earlier, the structures of IDRs are often absent, and their initial conformations need to be predicted. Current methods can already predict smaller IDRs quite well. For example, prediction of short loop (up to 32 amino acids) has been quite accurate with VSGB 2.0 model.[15] For other situations, there are also structure prediction methods including homology modeling, such as SWISS-MODEL[16] and MODELER,[17] as well as machine learning methods, such as AlphaFold2[18] and RoseTTAFold.[19] However, when studying IDRs within large protein complexes, integrative modeling may encounter some difficulties. If there are many missing IDRs in a large protein complex, structural predictions of all the IDRs simultaneously would involve immense computational demand.[20] On the other hand, predicting IDRs individually may produce serious spatial conflicts within the protein complex,

Received: April 30, 2024
Revised: June 16, 2024
Accepted: June 27, 2024
Published: July 10, 2024

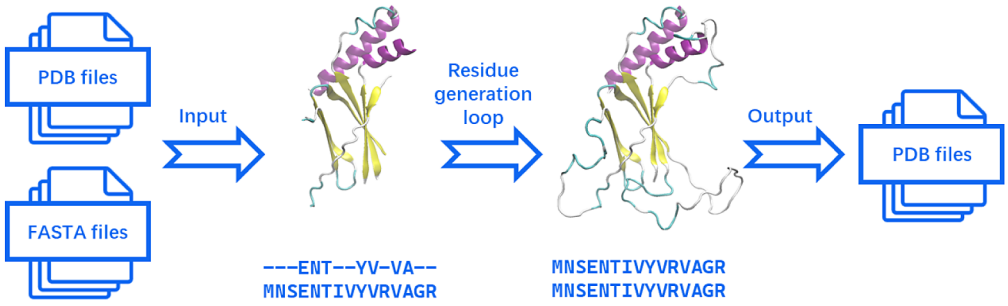


Figure 1. Workflow of IDRWalker.

making them unsuitable for subsequent computational simulations.

Given the persistent difficulties in modeling IDRs within large protein complexes,[2,21] we address issues of computational efficiency and spatial conflicts in this work. A self-avoiding random walk method is proposed with linear time complexity, which would enable rapid construction of IDR structures even within large protein complexes. Spatial conflicts of the generated IDRs are significantly eliminated because they are self-avoided. Therefore, starting from these initial conformations, computational simulations for integrative modeling can be performed smoothly.

While the implementation of the random walk method itself is easy, applying this method to large protein complexes can indeed be time-consuming, mainly due to tedious data preparation efforts. Therefore, a Python package called IDRWalker is written, which can automate the process of generating IDRs via random walks. Three protein complexes are used to illustrate the tool. With respect to the SARS-CoV-2 Omicron spike protein, we show the performance of IDRWalker in a scenario where knotting is likely to occur. In the calcium channel RyR1, we checked the chain connectivity of IDRs generated in the middle of the peptide chains. The final system is the human NPC, and the IDRs in more than 800 chains in this large protein machine can be generated with high efficiency.

In this paper, we first introduce the implementation of IDRWalker and some technical details. Then, the application of the program to the three protein complexes will be presented. Finally, we will discuss the prospects and future improvements of the method.

METHODS

Overview of IDRWalker. Figure 1 depicts the workflow of IDRWalker. First, the program reads sequence files and structure files, marking the positions of missing regions. Then, a residue generation loop is continuously repeated to generate residues one by one until the structures of all missing regions are constructed. Finally, the results are refined and output to the structure file.

Input Files Parsing. The generating process of a random walk is independent across different chains, so when reading input files, it is processed chain by chain. For each chain, the program, based on the read sequence, creates a chain containing the complete sequence by copying the amino acid residue templates, which are prepared in advance according to the AMBER ff19SB force field.[22] Subsequently, according to the residue numbers, the 3D coordinates in the input PDB file are written to the newly created chain.

To address spatial conflicts in the generating process at a lower cost, we partition the space where the protein exists into grid points. A matrix, referred to as the occupancy matrix in the subsequent text, is used to record whether each grid point is occupied by an atom. Before generation begins, the coordinates of each chain are written into the occupancy matrix. It is worth noting that IDRs are typically expansive, meaning that a very large occupancy matrix is needed, but a significant portion of the matrix is empty. To avoid the issue of memory waste from the large occupancy matrix, we apply periodic boundary conditions to the occupancy matrix. This allows IDRs to extend beyond the boundaries, enabling the use of a smaller occupancy matrix.

Residue Generation Loop. The residue generation loop is the core module of IDRWalker, and its function is to generate the coordinates of a new residue from a residue next to a missing region. This module can be implemented via the following steps:

The Coordinates of Backbone Atoms. Starting from a known residue, the coordinates of a new atom can be generated using the following equations:

$$\begin{cases} \boldsymbol{n}_2 = \boldsymbol{R}(\boldsymbol{v}_1, \theta_{\mathrm{d}})\boldsymbol{n}_1 \\ \boldsymbol{v}_2 = -\boldsymbol{R}(\boldsymbol{n}_2, \theta_{\mathrm{a}})\boldsymbol{v}_1 \\ \boldsymbol{r}_2 = \boldsymbol{r}_1 + l\cdot\boldsymbol{v}_2 \end{cases} \quad (1)$$

The formula involves a rotation matrix, $\boldsymbol{R}(\boldsymbol{axis}, \theta)$, where θ is the angle of rotation chosen around ***axis***. This matrix can be computed using the Rodrigues rotation formula. The meanings of the other terms are indicated in Figure 2a. In addition to the known residue, the formula also requires bond lengths, angles, and dihedrals. In protein structures, the bond lengths and bond angles are less variable and can therefore be determined directly, for example, using reference values in the AMBER ff19SB force field.[22] In contrast, dihedrals can be rotated relatively freely but with some regularities that can be depicted by the Ramachandran plot.[23] The dihedral ϕ in the diagram consists of C−N−C_α−C, and the dihedral ψ consists of N−C_α−C−N. Dihedral combinations located within the permissive region of the Ramachandran plot do not lead to spatial conflicts between neighboring residues. There is also a dihedral ω consisting of C_α−C−N−C_α in the main chain, and this dihedral is limited by the peptide bond and has a limited range of values, usually approximately 180° and, in a few cases, approximately 10°. Based on the described pattern, generating a suitable set of dihedral angles φ, ψ, and ω randomly and iterating eq 1 three times will yield the coordinates of the main chain atoms N, C_α, and C for the next residue. During the iteration process, it is essential not only to record the generated

https://doi.org/10.1021/acsomega.4c04161
ACS Omega 2024, 9, 32059−32065

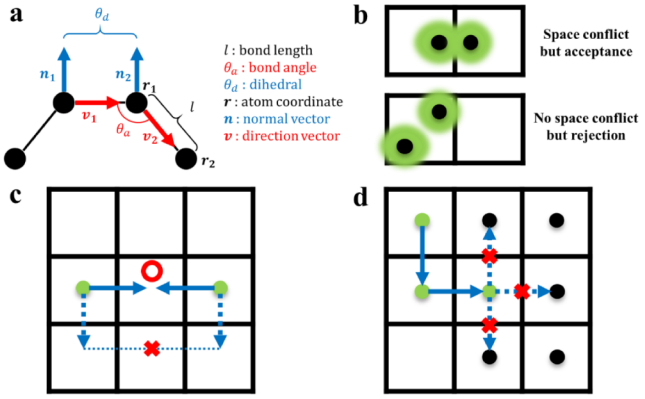


Figure 2. Some details of the residue generation loop. (a) Geometry of the four atoms. (b) Possible misjudging of space conflicts. (c) The connectivity of missing regions. (d) A dead-end of a self-avoiding random walk.

atomic coordinates but also to keep track of the normal vectors $\boldsymbol{n}$ and direction vectors $\boldsymbol{v}$ for use as inputs in the next iteration. Due to issues such as numerical precision, the lengths of the normal vectors and direction vectors may decrease slightly after each iteration. The accumulation of errors over multiple iterations can lead to false outputs. For this reason, it is necessary to normalize the normal vectors and direction vectors after a certain number of iterations to mitigate the impact of accumulated errors.

The Coordinates of the Side Chain Atoms. A random walk only generates coordinates for the main chain atoms C, C_α, and N, and the coordinates of the side chain atoms are unknown. Since the C_α atom is at the center of a tetrahedral configuration, the orientation of the side chain can be determined given the coordinates of the main chain atoms as well as the optical activity. The latter is implicit in the amino acid template. Accordingly, the coordinates of the side chain atoms can be obtained as follows: the amino acid template can be moved with the Kearsley algorithm,[24] and its main chain can be aligned with the newly generated main chain atoms. Subsequently, the coordinates of the template's side chain atoms are written into the new residue.

The Oxygen Atom in a Peptide Bond. Random walk also does not provide the coordinates of the oxygen atom in a peptide bond. The O atom in the peptide bond forms a double bond and is therefore coplanar with the neighboring C, C_α, and N atoms; thus, its coordinates can be expressed as a linear combination of the coordinates of these three atoms. Using the bond lengths and angles from the AMBER ff19SB force field,[22] the coordinates of the O atom can be calculated using the following formula:

$$\boldsymbol{r}_O = -0.757\cdot\boldsymbol{r}_{CA} - 0.886\cdot\boldsymbol{r}_N + 2.643\cdot\boldsymbol{r}_C \quad (2)$$

Due to the involvement of different atoms in bonding, the coordinates of the C-terminal oxygen atom cannot be calculated using eq 2. It can be assumed to lie in the same plane as the N, C_α, and C atoms of the same residue and then computed using the following formula:

$$\boldsymbol{r}_O = 0.783\cdot\boldsymbol{r}_N - 1.461\cdot\boldsymbol{r}_{CA} + 1.678\cdot\boldsymbol{r}_C \quad (3)$$

Unlike the situation where atoms determining the side chain position are within the same residue, the atoms related to the position of the oxygen atom in the peptide bond are in different residues. Due to this issue, the coordinates of the oxygen atom cannot be immediately determined after generating a new residue; they can only be finalized after generation of the entire chain.

Spatial Conflicts Check. Once the coordinates of all heavy atoms in the new residue, except for the oxygen atom in the peptide bond, are determined, it is crucial to determine their reasonability. The first step is to check for spatial conflicts. By comparing coordinates with the occupancy matrix and observing whether any atoms are positioned within already occupied grids, it is possible to make a rough assessment of spatial conflicts. Compared to calculating pair distances between atoms, this method involves significantly less computational effort. However, this approach can remove only large-scale overlaps and may not entirely eliminate the possibility of small-scale spatial conflicts. The latter can be resolved through energy minimization. Additionally, there is a risk of misjudging reasonable configurations as spatial conflicts (Figure 2b). A misjudgment involving the main chain atoms of adjacent residues could introduce false constraints to the main chain dihedrals. Considering that spatial conflicts between adjacent residues have already been addressed through the Ramachandran plot, when evaluating spatial conflicts, backbone atoms other than C_α are excluded.

Connectivity Check. Structural gaps can occur not only at the ends of peptide chains but also frequently within the middles of the chains. It is important to complete the models with these missing regions rather than leaving unconnected domains.[25] To generate a missing region in the middle of the peptide chain with a random walk, it is essential to ensure that the two ends of the missing region intersect at the same place after a certain number of random walk steps (Figure 2c). We used the Monte Carlo method to study the random walk law of the protein main chain. By repeating the random generation of a peptide chain with a length of 100 amino acid residues 1000 times, it was found that when the number of steps is large, the distribution of the distance from the starting point can be approximated by a normal distribution $N(\mu,\sigma)$ (Figure S1), and the relationship between $\langle\mu\rangle^2$,$\langle\sigma^2\rangle$ and the number of steps can be obtained by linear fitting (Figure S2). When the number of steps is small, although it cannot be described by the normal distribution, it can be directly described by the result of multiple walks. After a residue is generated by a random walk, the probability of connecting the peptide chain can be evaluated according to the above law; only results with a high likelihood of connecting pass the connectivity check.

Forward or Backward. If a newly generated residue passes the checks above, its coordinates can be written into the occupancy matrix, and the random walk moves forward. Attempts to generate a new residue are not always successful. If a generated residue fails the checks, the process needs to be repeated until success. If there are too many failed attempts, for example, exceeding 1000, the random walk might be stuck in a dead-end (Figure 2d). In such a case, moving backward is necessary, and the coordinates of the current residue should be removed from the occupancy matrix.

The above steps represent all the steps for generating a single residue. By iteratively running the residue generation loop, all missing IDRs can be generated. When filling multiple gaps, the filling process runs serially. If it were parallel, there might be cases where residues with spatial conflicts are created simultaneously. To prevent the order of generation from influencing the results, the missing region for generating residues is randomly selected each time during the process.

https://doi.org/10.1021/acsomega.4c04161
ACS Omega 2024, 9, 32059−32065

Refinement. Due to deficiencies in spatial conflict checks, a structure obtained through the above process needs refinement. IDRWalker does not have energy minimization capabilities, so structural corrections need to be performed using other tools such as the double-precision version of GROMACS.[26] Other details of energy minimization, such as molecular force fields and solvent effects, can also be configured in GROMACS. The structures that passed the energy minimization can be utilized for integrative modeling.

RESULTS AND DISCUSSION

The SARS-CoV-2 Omicron Spike Protein. The SARS-CoV-2 spike protein serves as a key player in the viral infection process, playing a central role in facilitating viral entry into host cells. Comprising two subunits, S1 and S2, the spike protein undergoes conformational changes that are crucial for mediating fusion between the viral and host cell membranes. The receptor-binding domain (RBD) within the S1 subunit directly interacts with the host cell receptor angiotensin converting enzyme 2 (ACE2), initiating the process of viral entry. Mutations, such as delta and omicron mutations, impact the affinity of a virus for ACE2, alter its transmissibility, and potentially confer immune escape.[27] In the solved structure of the spike protein (PDB ID: 7WK4),[28] many IDRs are missing, and some of them are near active structural domains, drug targets, and mutation sites.[27] Thus, generating missing IDRs in the spike protein is necessary.

When generating missing IDRs based on the solved structure of the SARS-CoV-2 Omicron spike protein, the orientation of terminal residues and the limited space may lead to the occurrence of knots between adjacent chains in the trimeric structure. Due to the rarity of naturally occurring knotted proteins,[29] knotting is generally considered irrational.

IDRWalker cannot avoid knots either, but due to the small amount of computation, randomness, and the convenience of adjusting the parameters, it is possible to try a large number of generations and select a reasonable result (Figure 3). Before performing computational simulations using this conformation, energy minimization is necessary. In the initial structure, the maximum force experienced by atoms reached 9.73×10^{17} kJ mol^{-1} nm^{-1}, indicating the potential presence of spatial conflicts. However, only a few hundred atoms exhibited significantly high forces, accounting for a small fraction of the total 34 994 atoms. This suggests that spatial conflicts do not occur extensively. After approximately 700 steps of energy minimization, the maximum force experienced by atoms decreased to below 1000 kJ mol^{-1} nm^{-1}.

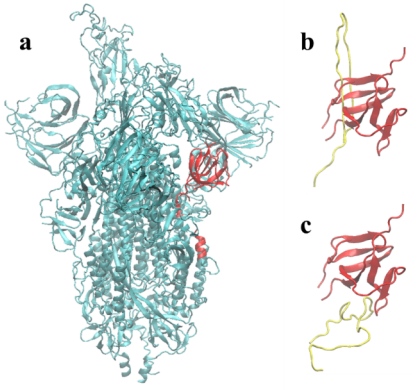


Figure 3. Generation of missing IDRs without knotting in the Omicron spike protein. (a) The known structure of the spike protein; an area of possible knots is marked in red. (b) A conformation with a knot. (c) A conformation without a knot.

Calcium Channel Protein RyR1. Ryanodine receptor 1 (RyR1) is a pivotal calcium channel protein located predominantly in the sarcoplasmic reticulum of skeletal muscle cells and plays a central role in excitation−contraction coupling.[30] Despite its physiological significance, obtaining a comprehensive and high-resolution structure of RyR1 has posed a significant challenge. Although the solved structures have provided valuable insights, certain regions are notably absent in each of the four symmetrical parts of RyR1, with some deletion regions exceeding 100 residues in length. In the structure of RyR1 with PDB ID 8SEU,[31] there are four gaps in each chain, and the numbers of missing amino acid residues in each gap are 115, 285, 38, and 22. These internal gaps in structural data often encompass highly flexible IDRs. When using random walks to generate these missing regions within chains, it is crucial to ensure that the generated residues can form a well-connected chain. Below, we examine whether the IDRWalker algorithm effectively ensures the connectivity of the peptide chains.

Based on the known structure of the A chain of RyR1, we generated conformations of the four IDRs using IDRWalker. In the generated IDRs (Figure 4a), the C-alpha distances between adjacent residues in all sequences are close, indicating the success of our algorithm in maintaining peptide chain connectivity, even in a long IDR with 285 residues.

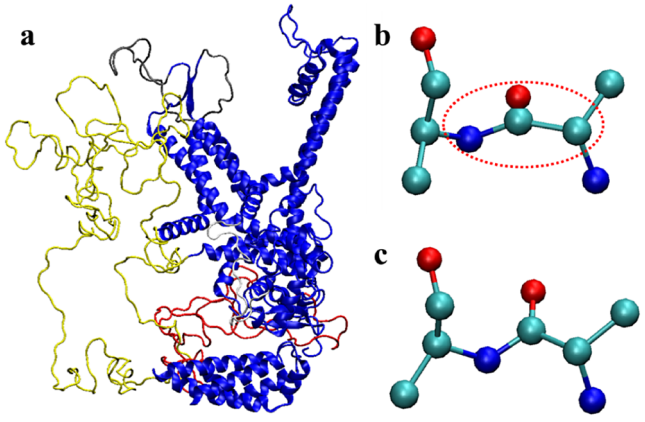


Figure 4. Generation of missing IDRs in the A chain of RyR1. (a) The generated result; the known regions are in blue, and the four generated IDRs are in red, yellow, gray, and white. (b) Incorrect bond angles and oxygen atom coordinates. (c) The corrected angles and oxygen atom coordinates after energy minimization.

Notably, issues with bond angles may arise, and as a result, the coordinates of some oxygen atoms could also be wrong (Figure 4b). This is because the algorithms ensuring connectivity impose restrictions only on distances without considering angles. Although the patterns of angle variation with the number of walking steps are similar to those of the distances, constraining angles significantly reduce the success rate of generation. Therefore, the problem of bond angle anomalies is not addressed during generation but is subsequently corrected through energy minimization. Before energy minimization, the maximum force experienced by atoms in the structure was 1.37×10^{15} kJ mol^{-1} nm^{-1}. After

https://doi.org/10.1021/acsomega.4c04161
ACS Omega 2024, 9, 32059−32065

672 steps of energy minimization, the maximum force decreased to below 1000 kJ mol^{-1} nm^{-1}, resolving both spatial conflicts and abnormal bond angles (Figure 4c).

The Human NPC. After successful testing of the two protein complexes, we applied the tool to a massive protein machine, NPC. The NPC is a crucial structure governing molecular trafficking between the nucleus and the cytoplasm. Notably, it is characterized by the significant presence of IDRs, such as phenylalanine-glycine (FG) repeats, within the central transport channel and cytoplasmic filaments.[5] These IDRs are believed to play a crucial role in controlling the entry and exit of substances into the cell nucleus.[13] However, the underlying mechanism can only be explained by some hypotheses that have not been completely verified. To further elucidate the underlying mechanism, it is essential to establish initial structures for these IDRs within the NPC.

Due to its enormous size, the NPC poses significant challenges for the generation of IDRs. Although IDRWalker automates the generating process, organizing the input files remains a nontrivial task. We downloaded the structure and sequence files of the human NPC (PDB ID: 7R5K),[13] which consists of a total of 808 chains. All the chains had missing regions, and the number of missing residues ranged from 1 to 2464, which resulted in a total of 248,928 missing residues. Subsequently, we saved the structure and sequence of each chain into separate PDB and FASTA files. These files were then sequentially read by IDRWalker to complete the generating process. The preprocessing of input files is not integrated into IDRWalker because it may be necessary to write different scripts to handle different systems.

Figure 5 shows the conformation of NPC with the missing IDRs added to the solved structure. Due to its immense size,

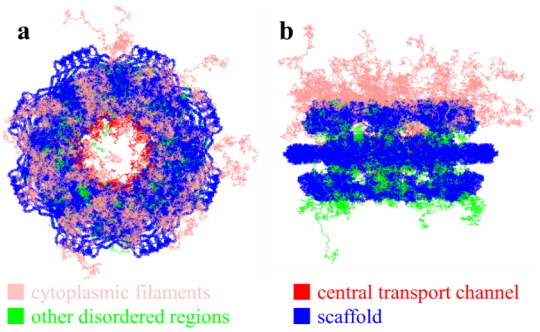


Figure 5. Generation of IDRs in the human NPC. (a) The generated IDRs together with the scaffold of the human NPC are shown on the cytoplasmic side. The different parts are color coded as indicated in the color bar. (b) Same as (a) but shown from a perspective parallel to the nuclear membrane.

energy minimization was performed using the Martini coarse-grained (CG) model[32] rather than the all-atom model. The maximum force experienced by CG beads in the system reached 1.50 × 10^{14} kJ mol^{-1} nm^{-1}, indicating spatial conflicts. After 250 steps of energy minimization in vacuum, the maximum force decreased to less than 1000 kJ mol^{-1} nm^{-1}. The result can be utilized as the initial conformation for computational simulations.

Efficiency of the Program. The motivation for designing the program is to address the efficiency issue associated with generating IDRs in large protein complexes, so it is important for the program to run at a sufficiently fast speed. Table 1 displays the time consumption and parameters for generation in different systems. IDRWalker can generate IDRs with fewer than 500 residues in less than 1 s. Even for a large protein complex such as NPC, the runtime is only a few minutes. The above results were obtained using only a single CPU core. If parallel processing is employed, it allows the simultaneous generation of multiple different conformations, thereby further enhancing efficiency. However, when generating large systems, parallel processing may face significant limitations due to substantial memory usage. This is an area that requires improvement.

Table 1. Running Details of IDRWalker

generated regions	PDB ID	generated residues	running time[a]	memory usage
internal gaps in omicron S protein	7WK4	157	<1 s	119.2 MB
internal gaps in A chain of RyR1	8SEU	460	<1 s	119.2 MB
missing IDRs in human NPC	7R5K	248928	81 s	2.21 GB

[a]Executed using Intel Core i7−12700KF processor in single-threaded mode.

CONCLUSION

Currently, there are challenges in modeling IDRs in large protein complexes, including high computational demands and spatial conflicts. For example, it is difficult for AlphaFold2 to predict IDRs in large protein complexes containing more than tens of thousands of residues. We address these issues by employing a self-avoiding random walk method, which allows for the generation of IDRs with low computational costs. Additionally, we developed a program named IDRWalker to simplify the generation steps, making it more efficient for application to large protein complexes. Using IDRWalker, the IDRs in large protein complexes such as NPC can be generated in several minutes.

We first generated IDRs in the Omicron spike protein and calcium channel protein RyR1 and verified that IDRWalker is capable of addressing issues that may arise during generation, such as knots and connectivity in peptide chains. Then, we generated the IDRs in the NPC. By examining the runtime and generated results, we validated the feasibility of the program in the application of large protein complexes.

It should be noted that, each run of IDRWalker will generate only one conformation. However, an IDR generally exists as an ensemble of rapidly interconverting conformations.[33] To better describe the IDR, it is necessary to generate a pool of many different conformations. Since IDRWalker is very efficient, the conformation pool can be obtained by running the program multiple times.

Even if a conformation pool is generated, the conformations may not fully represent the native conformation ensembles of the IDR. This is because the current program, to reduce the computational load, employs a simple approach and does not utilize any experimental data. To obtain high-quality models based on generated conformations, it is necessary to perform integrated modeling by running computational simulations with experimental data.

Incorporating experimental data during the generation of IDRs is also feasible. For instance, the density map of disordered regions can be used to control the generation. Another example is applying the algorithm used in IDRWalker,

https://doi.org/10.1021/acsomega.4c04161
ACS Omega 2024, 9, 32059−32065

which ensures chain continuity, to constrain the distance between residues, which can be determined through experiments such as NMR, FRET, and XL-MS. We will explore the implementation of these ideas in future research endeavors.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.4c04161.

Figure S1: Approximate description of distance distribution using normal distribution and Figure S2; $\langle\mu\rangle^2$ and $\langle\sigma^2\rangle$ vary with the number of steps (PDF)

Accession Codes

IDRWalker is an open-source Python package. Its source code is publicly accessible on GitHub (https://github.com/zyzhangGroup/IDRWalker).

■ AUTHOR INFORMATION

Corresponding Author

Zhiyong Zhang − *Department of Physics, University of Science and Technology of China, Hefei, Anhui 230026, PR China; MOE Key Laboratory for Cellular Dynamics, University of Science and Technology of China, Hefei, Anhui 230026, PR China;* orcid.org/0000-0001-8202-9672; Phone: +86-551-63606134; Email: zzyzhang@ustc.edu.cn

Author

Guanglin Chen − *Department of Physics, University of Science and Technology of China, Hefei, Anhui 230026, PR China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.4c04161

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work is supported by the National Key Research and Development Program of China (2021YFA1301504), the Chinese Academy of Sciences Strategic Priority Research Program (XDB37040202), and the National Natural Science Foundation of China (91953101). We are grateful to Mr. Yundong Zhang for technical support.

■ REFERENCES

(1) Dyson, H. J.; Wright, P. E. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* **2005**, *6* (3), 197−208.

(2) Liu, Y.; Wang, X.; Liu, B. A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction. *Brief. Bioinform.* **2019**, *20* (1), 330−346.

(3) Bai, X. C.; McMullan, G.; Scheres, S. H. How cryo-EM is revolutionizing structural biology. *Trends Biochem. Sci.* **2015**, *40* (1), 49−57.

(4) García-Nafría, J.; Tate, C. G. Structure determination of GPCRs: Cryo-EM compared with X-ray crystallography. *Biochem. Soc. Trans.* **2021**, *49* (5), 2345−2355.

(5) Lin, D. H.; Hoelz, A. The Structure of the Nuclear Pore Complex (An Update). *Annu. Rev. Biochem.* **2019**, *88*, 725−783.

(6) Russel, D.; Lasker, K.; Webb, B.; Velázquez-Muriel, J.; Tjioe, E.; Schneidman-Duhovny, D.; Peterson, B.; Sali, A. Putting the pieces together: Integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS Biol.* **2012**, *10* (1), No. e1001244.

(7) van Zundert, G. C. P.; Rodrigues, J. P. G. L. M.; Trellet, M.; Schmitz, C.; Kastritis, P. L.; Karaca, E.; Melquiond, A. S. J.; van Dijk, M.; de Vries, S. J.; Bonvin, A. M. J. J. The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *J. Mol. Biol.* **2016**, *428* (4), 720−725.

(8) Rantos, V.; Karius, K.; Kosinski, J. Integrative structural modeling of macromolecular complexes using Assembline. *Nat. Protoc.* **2022**, *17* (1), 152−176.

(9) Koukos, P. I.; Bonvin, A. Integrative Modelling of Biomolecular Complexes. *J. Mol. Biol.* **2020**, *432* (9), 2861−2881.

(10) Rout, M. P.; Sali, A. Principles for Integrative Structural Biology Studies. *Cell* **2019**, *177* (6), 1384−1403.

(11) Holding, A. N. XL-MS: Protein cross-linking coupled with mass spectrometry. *Methods* **2015**, *89*, 54−63.

(12) Dimura, M.; Peulen, T. O.; Hanke, C. A.; Prakash, A.; Gohlke, H.; Seidel, C. A. Quantitative FRET studies and integrative modeling unravel the structure and dynamics of biomolecular systems. *Curr. Opin. Struct. Biol.* **2016**, *40*, 163−185.

(13) Mosalaganti, S.; Obarska-Kosinska, A.; Siggel, M.; Taniguchi, R.; Turoňová, B.; Zimmerli, C. E.; Buczak, K.; Schmidt, F. H.; Margiotta, E.; Mackmull, M. T.; et al. AI-based structure prediction empowers integrative structural analysis of human nuclear pores. *Science* **2022**, *376* (6598), No. eabm9506.

(14) Chan-Yao-Chong, M.; Durand, D.; Ha-Duong, T. Molecular Dynamics Simulations Combined with Nuclear Magnetic Resonance and/or Small-Angle X-ray Scattering Data for Characterizing Intrinsically Disordered Protein Conformational Ensembles. *J. Chem. Inf. Model.* **2019**, *59* (5), 1743−1758.

(15) Li, J.; Abel, R.; Zhu, K.; Cao, Y.; Zhao, S.; Friesner, R. A. The VSGB 2.0 model: a next generation energy model for high resolution protein structure modeling. *Proteins* **2011**, *79* (10), 2794−2812.

(16) Waterhouse, A.; Bertoni, M.; Bienert, S.; Studer, G.; Tauriello, G.; Gumienny, R.; Heer, F. T.; de Beer, T. A. P.; Rempfer, C.; Bordoli, L.; et al. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* **2018**, *46* (W1), W296−W303.

(17) Webb, B.; Sali, A. Comparative Protein Structure Modeling Using MODELLER. *Curr. Protoc. Bioinformatics* **2016**, *54* (1), 5.6.1−5.6.37.

(18) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583−589.

(19) Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D.; et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021**, *373* (6557), 871−876.

(20) Bryant, P.; Pozzati, G.; Zhu, W.; Shenoy, A.; Kundrotas, P.; Elofsson, A. Predicting the structure of large protein complexes using AlphaFold and Monte Carlo tree search. *Nat. Commun.* **2022**, *13* (1), 6028.

(21) Ruff, K. M.; Pappu, R. V. AlphaFold and Implications for Intrinsically Disordered Proteins. *J. Mol. Biol.* **2021**, *433* (20), 167208.

(22) Tian, C.; Kasavajhala, K.; Belfon, K. A. A.; Raguette, L.; Huang, H.; Migues, A. N.; Bickel, J.; Wang, Y.; Pincay, J.; Wu, Q.; et al. ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. *J. Chem. Theory Comput.* **2020**, *16* (1), 528−552.

(23) Ramachandran, G. N.; Ramakrishnan, C.; Sasisekharan, V. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* **1963**, *7* (1), 95−99.

(24) Kearsley, S. On the orthogonal transformation used for structural comparisons. *Acta Crystallogr. A* **1989**, *45* (2), 208−210.

(25) Dror, R. O.; Arlow, D. H.; Borhani, D. W.; Jensen, M. Ø.; Piana, S.; Shaw, D. E. Identification of two distinct inactive conformations of the β2-adrenergic receptor reconciles structural and biochemical observations. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106* (12), 4689−4694.

(26) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1−2*, 19−25.
(27) Ebada, M. A.; Wadaa-Allah, A.; Bahbah, E.; Negida, A. An Updated Review on COVID-19. *Infect. Disord.: drug Targets* **2021**, *21* (8), No. e160921189190.
(28) Hong, Q.; Han, W.; Li, J.; Xu, S.; Wang, Y.; Xu, C.; Li, Z.; Wang, Y.; Zhang, C.; Huang, Z.; et al. Molecular basis of receptor binding and antibody neutralization of Omicron. *Nature* **2022**, *604* (7906), 546−552.
(29) Jamroz, M.; Niemyska, W.; Rawdon, E. J.; Stasiak, A.; Millett, K. C.; Sułkowski, P.; Sulkowska, J. I. KnotProt: A database of proteins with knots and slipknots. *Nucleic Acids Res.* **2015**, *43* (D1), D306−D314.
(30) Clarke, O. B.; Hendrickson, W. A. Structures of the colossal RyR1 calcium release channel. *Curr. Opin. Struct. Biol.* **2016**, *39*, 144−152.
(31) Cholak, S.; Saville, J. W.; Zhu, X.; Berezuk, A. M.; Tuttle, K. S.; Haji-Ghassemi, O.; Alvarado, F. J.; Van Petegem, F.; Subramaniam, S. Allosteric modulation of ryanodine receptor RyR1 by nucleotide derivatives. *Structure* **2023**, *31* (7), 790−800.e794.
(32) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations. *J. Phys. Chem. B* **2007**, *111* (27), 7812−7824.
(33) Tompa, P.; Han, K.-H. Intrinsically disordered proteins. *Phys. Today* **2012**, *65* (8), 64−65.

https://doi.org/10.1021/acsomega.4c04161
ACS Omega 2024, 9, 32059−32065