Predicting the Assembly Order of Protein Complexes by Multi-Temperature Coarse-Grained Simulations

Yunxiao Lu[1], Xin Liu[2], Zhiyong Zhang[1,2,3*]

[1]MOE Key Laboratory for Cellular Dynamics and Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, Anhui 230026, P. R. China

[2]School of Data Science, University of Science and Technology of China, Hefei, Anhui 230026, P. R. China

[3]Department of Physics, University of Science and Technology of China, Hefei, Anhui 230026, P. R. China

*Corresponding author: Zhiyong Zhang, Tel: +86-551-63606134; Email: zzyzhang@ustc.edu.cn

Running title: The Assembly Order of Protein Complexes

Abstract

An increasing number of structures of protein complexes have been resolved. In addition to one or several static structures, the dynamic process of a protein complex, such as the assembly order of its protein subunits, is functionally very important. In addition to experimental techniques for studying subcomplexes in the assembly of a protein complex, computational methods can be used to predict the assembly order. Since sampling is a nontrivial issue in simulating the assembly process, coarse-grained simulations are more efficient than atomic simulations are. In this work, we developed a computational tool for predicting the disassembly and assembly order of protein complexes by coarse-grained simulations at multiple temperatures. The method was illustrated using two protein complexes, and the predicted assembly orders are consistent with available experimental data. Our studies also suggest that the assembly order of a protein complex may be determined not only by intersubunit interactions but also by the stability of the subunits.

Keywords

Protein complexes; Assembly order; Coarse-grained simulations; Native contacts

Significance

Protein complexes play important roles in various biological processes. The order in which a protein complex is assembled is critical for its proper formation and function. In the assembly process of a complex, the precise arrangement of protein subunits determines their stability, activity and interaction partners. This work provides a

computational tool for predicting the assembly order of protein complexes, with the expectation of revealing the dynamic mechanism behind the function of protein complexes.

Introduction

Protein complexes carry out the majority of catalytic, structural, and regulatory functions(1). Currently, cryo-electron microscopy (cryo-EM)(2) and X-ray crystallography(3) are the primary experimental techniques for obtaining high-resolution atomic structures of protein complexes. Additionally, computer techniques incorporating artificial intelligence enable predictive modeling of protein complexes(4). With the continuous advancement of experimental and computational methods, an increasing number of structures of protein complexes have been determined, providing opportunities to investigate their functions at the molecular level.

However, in addition to static structures, it is essential to study the dynamics of protein complexes, such as the assembly order(5). Many protein complexes have evolved to be assembled in a specific order(6). The stability and function of a protein complex rely on a multistep process of subunit assembly, which can provide valuable insights into the function and evolution of the protein complex. Numerous studies have demonstrated that improper protein assembly may result in negative biological effects(7). However, due to the high complexity of the assembly pathway, the current understanding of the assembly mechanism of protein complexes is limited. According

to Tompa and Rose's study(8), the assembly of proteins is hierarchical, and a protein complex must follow an ordered assembly pathway to achieve its biological function(9). Therefore, studying the assembly order of protein complexes contributes to the reconstruction of protein complexes in vitro, the design of artificial protein complexes, and the development of new drugs related to assembly(10).

The assembly order of a protein complex can be determined by studying the stabilized intermediates of two or more subunits using experimental techniques such as co-immunoprecipitation(11), mass spectrometry (MS)(12), and time-resolved cryo-electron microscopy(13) or by creating deletion mutants to study the effect of deletion on assembly(14). MS is a powerful technique for identifying subcomplexes during the assembly or disassembly of a protein complex(15). A typical MS measurement generally starts with ionization of the protein complex(16). The complex is excessively charged, and signals from a set of subcomplexes are subsequently detected via MS. Subunits that bind each other weakly tend to dissociate from the complex(17).

Computational modeling has the potential to predict the assembly order of protein complexes, making it a valuable tool in protein research. In early studies(18,19), the order of subcomplex formation was predicted by using the buried surface area (BSA), which is the difference between the solvent-accessible surface area (SASA) of the subunit in the complex and that of the isolated subunit. A higher BSA would correlate with a lower (more favorable) binding free energy, which may be used to determine the assembly order. The assembly orders were found to match the BSA predictions for 13

out of 16 homocomplexes and 7 out of 9 heterocomplexes. However, there was only a moderate correlation between BSA and binding free energy(20). Path-LZerD(21) uses knowledge-based potentials to evaluate binding preferences between subunits and can predict assembly paths even when the complex structure is unknown. The method involves constructing docked structures of subunits from single subunits via the Multi-LZerD method(22), and iterative optimizations are performed via a genetic algorithm(23). The predictions of Path-LZerD are consistent with the experimental data for three complexes with up to four subunits, but for larger complexes with five or more subunits, the number of correct predictions decreased. The hybrid Monte Carlo/molecular dynamics simulation (hMC/MD) method(24) uses a likelihood-based selection scheme to predict the disassembly order of a protein complex. The success rate of the method was greater than 0.9 for four tetrameric protein complexes.

When modeling large protein complexes, all-atom simulations are expensive for describing long time-scale dynamic processes. Therefore, other methods, such as coarse-grained (CG) simulations, which are computationally efficient, have been proposed to address this issue(25). Although a CG model does not describe protein interactions as accurately as an all-atom model, the former can still capture kinetic features of the protein complex(26). Furthermore, CG simulations can depict a biological process over a long timescale and have been employed to elucidate the physical mechanism of a variety of protein complexes(25-27).

This work developed a computational tool for predicting the assembly order of protein complexes. In this tool, the disassembly process of a protein complex is

simulated via CG simulations at different temperatures. This idea comes from MS experiments because standard electrospraying relies on the use of high temperatures, chemical cosolvents, and organic acids. The temperature can be used to simulate the degree of ionization in MS and induce the dissociation of protein complexes. The disassembly temperature of each subunit is estimated through multi-temperature CG simulations, after which the assembly order of the complex is predicted. Two protein complexes, the Arp2/3 complex and the archaeal RNA polymerase, were used to validate the method, and the predicted assembly order was consistent with the existing in vitro experimental data.

Methods

General workflow

To predict the assembly order of multiple subunits in a protein complex, a computational tool based on multi-temperature CG simulations was used (**Figure 1**). A high temperature breaks interactions between protein subunits, which may lead to subunit disassembly. This process can be more easily observed by CG simulations than by all-atom simulations.

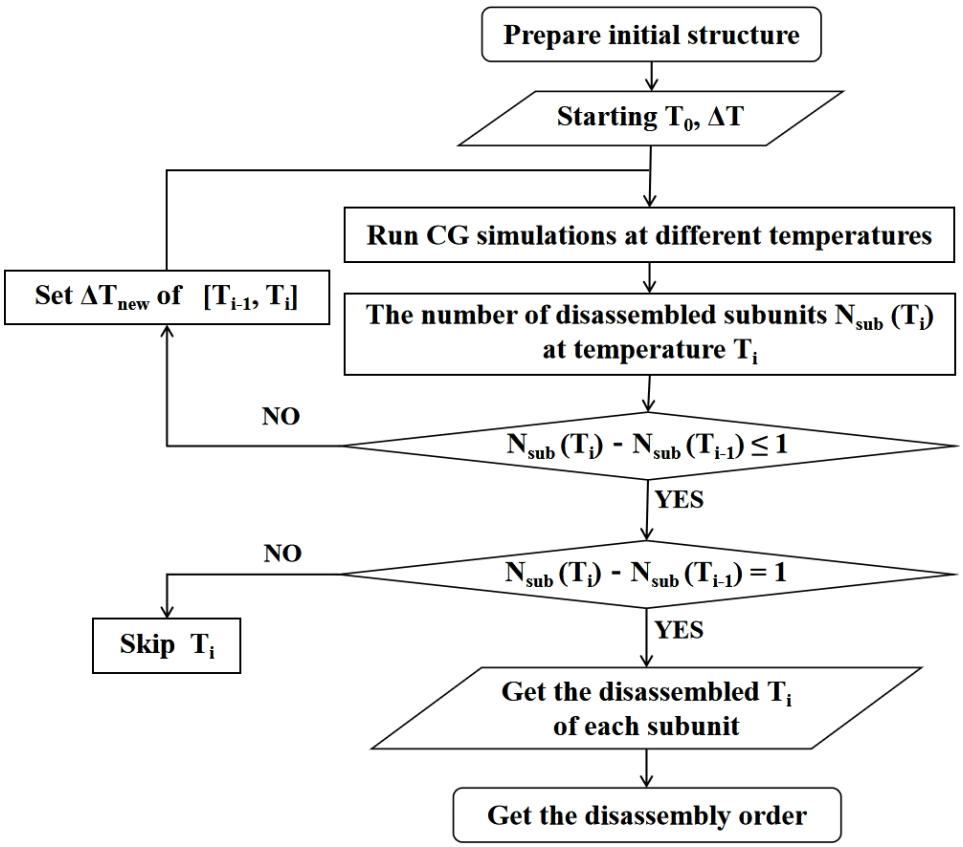

Figure 1 Flowchart of multi-temperature CG simulations for predicting the disassembly order of a protein complex.

The tool consists of the following steps:

1. The initial structure of the protein complex was prepared.

2. Run CG simulations at different temperatures. The starting temperature is 300 K, and 300 K + ΔT and 300 K + 2ΔT are used until all the subunits are disassembled at a certain temperature.

3. The CG trajectories from low to high temperatures were analyzed, and N_{sub} and ΔN_{sub} were obtained at each temperature. $N_{sub}(T_i)$ is the number of disassembled subunits at temperature T_i and can be obtained by analyzing the qscore in the CG simulation (details can be found in the next section). $\Delta N_{sub}(T_i)$ is the difference between N_{sub} at temperature T_i and temperature T_{i-1}.

$$\Delta N_{sub}(T_i) = N_{sub}(T_i) - N_{sub}(T_{i-1})$$

4. If $\Delta N_{sub}(T_i)=0$, the temperature T_i is skipped. If $\Delta N_{sub}(T_i)=1$, there is a disassembled subunit between T_{i-1} and T_i, and we set the disassembled temperature of this subunit as T_i. If $\Delta N_{sub}(T_i)>1$, there are two or more disassembled subunits between T_{i-1} and T_i. In this case, ΔT_{new} needs to be set in the temperature range [T_{i-1}, T_i]; similar to step 2, a series of CG simulations are run at $T_{i-1}+\Delta T_{new}$, $T_{i-1}+2\Delta T_{new}$, until $T_i-\Delta T_{new}$. These CG trajectories are analyzed as described in step 3.

5. The procedure was repeated until the disassembled temperature of each subunit was determined. We have written several scripts and programs to run the procedure automatically, which can be found at GitHub (zyzhangGroup/Path-mTCG (github.com)).

Determination of the disassembly of a subunit by the qscore

Native contacts are formed by pairs of amino acid residues that are physically close to each other in the native structure. The cutoff distance of the C_α atoms between two residues in native contact was set to 6.5 Å by default. The intersubunit native contacts of the complex structure were all counted.

qscore is the fraction of native contacts in a conformation. In the native structure of a protein complex, if two subunits have native contacts, the corresponding intersubunit qscore is 1.0. When the intersubunit qscore between two subunits is 0, there is no native contact between them. During a CG simulation, the qscores may change, which can be used to determine whether each subunit disassembles. If a subunit loses contact with other subunits (qscores become 0) in the CG simulation, the subunit is disassembled.

CG Models

The AICG2+ (the atomic interaction-based CG) protein model was used, which extends the classical Gō model to incorporate sequence-dependent local flexibility and nonlocal contacts(28). The energy function of the AICG2+ model is as follows:

$$V_{AICG2+}(R|R_0) = \sum_{ibd} K_{b,ibd}\left(b_{ibd} - b_{ibd,0}\right)^2 + V_{loc}^{flp} + \sum_{j=i+2} \varepsilon_{loc,ij} \exp\left(-\frac{\left(r_{ij} - r_{ij0}\right)^2}{2W_{ij}^2}\right) + \sum_{j=i+3} \varepsilon_{loc,ij} \exp\left(-\frac{\left(\phi_{ij} - \phi_{ij0}\right)^2}{2W_{\phi,ij}^2}\right) + \sum_{i<j-3}^{native} \varepsilon_{go,ij}\left[5\left(\frac{r_{ij0}}{r_{ij}}\right)^{12} - 6\left(\frac{r_{ij0}}{r_{ij}}\right)^{10}\right] + \sum_{i<j-3}^{non-native} \varepsilon_{ev}\left(\frac{d}{r_{ij}}\right)^{12}$$

R represents the $3n_{aa}$ Cartesian coordinates of the simulated protein, and R_0 represents those of the native structure. b_{ibd} is the ibd-th virtual bond length. r_{ij} is the distance between the *i*-th and *j*-th amino acids. V_{loc}^{flp} is a generic flexible local potential for virtual bond angles and dihedral angles constructed by analyzing loop structures in a protein structure database. The third term describes specific local interactions, and $\varepsilon_{loc,ij}$ and W_{ij} in the Gaussian function represent the strength and width of the local interaction between residues i and j, respectively. Φ_{ij} in the fourth term is the dihedral angle formed by the four consecutive residues i to i+3, and $W_{\Phi,ij}$ is the width of this local interaction. The fifth term is the structure-based native contact potential, and the last term represents the sum of interactions between non-native contact residue pairs.

CG Simulations

All CG simulations were run using the CafeMol 3.2 package, which is a general-purpose CG simulation software for simulating proteins and their complexes(29). We

used the AICG2+ model, excluded volume and electrostatic interactions for protein complexes. The electrostatic interactions were calculated using the Debye–Hückel equation with a dielectric constant of 78 F/m and an ion concentration of 0.15 M.

For each system, an energy minimization was initially performed at 300 K, running 1000 steps with the steepest descent method, followed by 2000 steps with the conjugate gradient algorithm. Subsequently, a production simulation was run using Langevin dynamics at a certain temperature. The time step is 0.4 Cafe-time (1 Cafe-time ~ 49 fs). In total, 10^8 steps were run for each simulation, and a snapshot was saved every 5000 steps. The neighbor list was updated every 100 steps.

The parameter go_unit determines the weight of the Gō potential in the energy function (the default value is 1). Different go_units from 0.5 to 1.0 were tested by running CG simulations of the complex at 300 K. We finally chose a go_unit, with which sufficient motion of the complex can be observed in the CG simulation while the complex is essentially stable.

Protein complexes for testing the method

We selected two asymmetric heteromeric protein complexes from the 3D Complex server(30), the Arp2/3 complex (Figure 2a) and the archaeal RNA polymerase (aRNAP, Figure 2b), for testing the method.

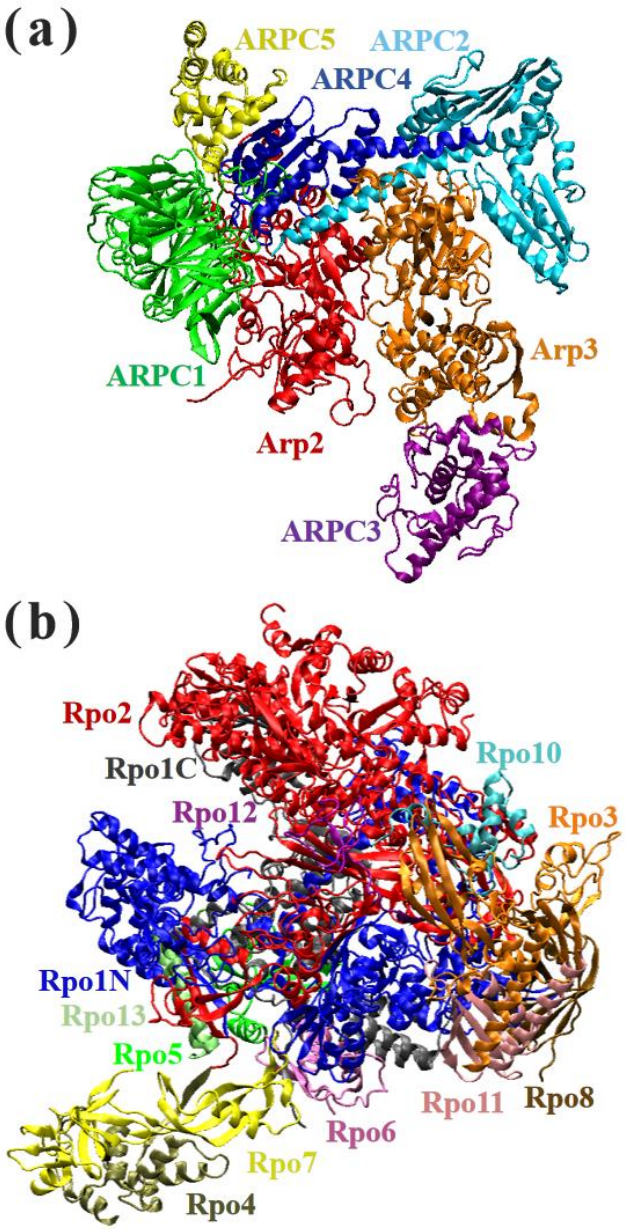

Figure 2 Structures of the Arp2/3 complex and aRNAP. (a) The inactive state of the Arp2/3 complex. The seven subunits are shown in different colors. The ATP/ADP molecules were excluded from the CG simulations. (b) aRNAP without DNA. The 13 subunits are shown in different colors.

The Arp2/3 complex. This complex is an important regulator of the cytoskeleton and a factor that enhances the mobility of cancer cells, showing a strong correlation with cancer development(31). The crystal structure of the inactive Arp2/3 complex (PDB 1K8K) from *bos taurus* was solved with a resolution of 2.0 Å(32), and the non-terminal missing residues were built by homologous modeling(33). The complex consists of seven subunits, including two actin-associated proteins (Arp2 and Arp3) and five other

proteins (named ARPC1-ARPC5 by descending order of molecular mass)(32). Both Arp2 and Arp3 are structurally similar to actin and may form a dimer with the ability to bind actin filaments to form branches(34). ARPC2 and ARPC4 hold the dimer together by anti-parallel binding of the C-terminal long helix. The α/β domain of ARPC4 interacts with Arp2, ARPC1, and ARPC5, while the two similar α/β domains of ARPC2 interact only with Arp3(35). ARPC3 is only in contact with Arp3(36).

aRNAP. The crystal structure of aRNAP from *Sulfolobus shibatae* (PDB ID: 4AYB) was solved with a resolution of 3.2 Å(37), and the non-terminal missing residues were built by homologous modeling(33). This complex consists of 13 protein subunits named Rpo1N, Rpo1C, Rpo2-Rpo8, and Rpo10-Rpo13(38). The three large subunits Rpo1N, Rpo1C and Rpo2 account for more than two-thirds of the molecular mass and form a large body similar to the shape of a crab claw (Figure 2b). The efficient assembly of large subunits requires an assembly platform comprising four small subunits (Rpo3, Rpo10, Rpo11, Rpo12) located on the opposite side of the claw. Rpo6 can stabilize the split of Rpo1N and Rpo1C and serve as an anchor for Rpo7/Rpo4(39). Rpo5 is located at the lower jaw of the crab claw, together with the large subunits Rpo1N and Rpo1C(40). Rpo13 is located between Rpo5 and the clamp-head domain of Rpo1N and functions as an aRNAP-DNA stabilizing factor(37). The two subunits Rpo4 and Rpo7 form a stalk that protrudes from the RNAP body. Rpo8 is located in the periphery and weakly interacts with Rpo1N(41).

Results and Discussion

Assembly order of the Arp2/3 complex

First, we obtained a matrix of native contacts of the Arp2/3 complex (Table 1) by using CafeMol(29). ARPC2 and ARPC4 have the largest number of intersubunit contacts, which is consistent with the experimental data showing that ARPC2 and ARPC4 serve as the core subunits of the complex(36).

Table 1 Intersubunit native contacts in the atomic structure of the Arp2/3 complex

Subunit	ARPC4	ARPC2	Arp3	Arp2	ARPC1	ARPC5	ARPC3	Sum
ARPC4 [c](167)		[a]244	72	102	188	107	0	[b]713
ARPC2 (282)			213	0	2	0	0	459
Arp3 (414)				79	0	0	88	452
Arp2 (378)					37	98	0	316
ARPC1 (368)						80	0	307
ARPC5 (143)							0	285
ARPC3 (173)								88

[a]Native contacts between two subunits; [b]The total number of intersubunit native contacts of the subunit; [c]The number in parentheses is the residue number of the subunit.

The go_unit was set to 0.65. This value allows for a significant conformational change in the complex in a CG simulation at 300 K, while the complex structure can be preserved. Then, we performed independent CG simulations under different temperatures with the scheme shown in Figure 1 to simulate the disassembly processes of the complex.

For each subunit i, a weighted average of intersubunit qscores (denoted as $<interq>_i$) was computed as follows:

$$< interq >_i = \frac{\sum_{j \neq i} (qscore)_{ij} \times (nat-con)_{ij}}{\sum_j (nat-con)_{ij}},$$

where $(qscore)_{ij}$ is the qscore between subunits i and j and $(nat\text{-}con)_{ij}$ is the number of native contacts between the two subunits. The time evolution of <interq> during the CG simulation at each temperature was investigated. At 315 K (Figure 3A), the <interq> values of Arp3 (orange), ARPC2 (cyan), and ARPC3 (purple) are essentially greater than 0.8. However, the <interq> of ARPC1 (green) drops to zero in the middle of the CG simulation, followed by that of ARPC5 (yellow). That is, ARPC1 and ARPC5 disassemble sequentially at 315 K (Figure 4). The <interq> values of Arp2 (Figure 3A, red) and ARPC4 (Figure 3A, blue) decrease since they have lost contact with the disassembled ARPC1 and ARPC5. Arp2 is unfolded after disassembly of ARPC1, although the former is still in contact with the core (Figure 4). In the CG simulation at 316 K (Figure 3B), Arp2 is disassembled since its <interq> decreases to zero. The <interq> values of Arp3 and ARPC4 decrease accordingly because they have lost contact with the disassembled Arp2. The <interq> values of ARPC2 and ARPC3 still remain high. The subcomplex ARPC2-ARPC4-Arp3-ARPC3 is well folded at this temperature (Figure 4). At 350 K (Figure 3C), ARPC3 disassembles. It should be noted that ARPC3 has only 88 native contacts with Arp3 (Table 1). However, the disassembly of ARPC3 requires a relatively high temperature, indicating that the binding of ARPC3 to Arp3 is stable. The <interq> of Arp3 further decreases at 350 K because its contacts with ARPC3 have broken, and that of ARPC2 starts to decrease. The subcomplex ARPC2-ARPC4-Arp3 is largely unfolded at this temperature (Figure 4). Finally, in the CG simulation at 352 K, Arp3 disassembles, followed by the ARPC2-ARPC4 core (Figure 3D). The core dimer (Figure 4) is the last to disassemble.

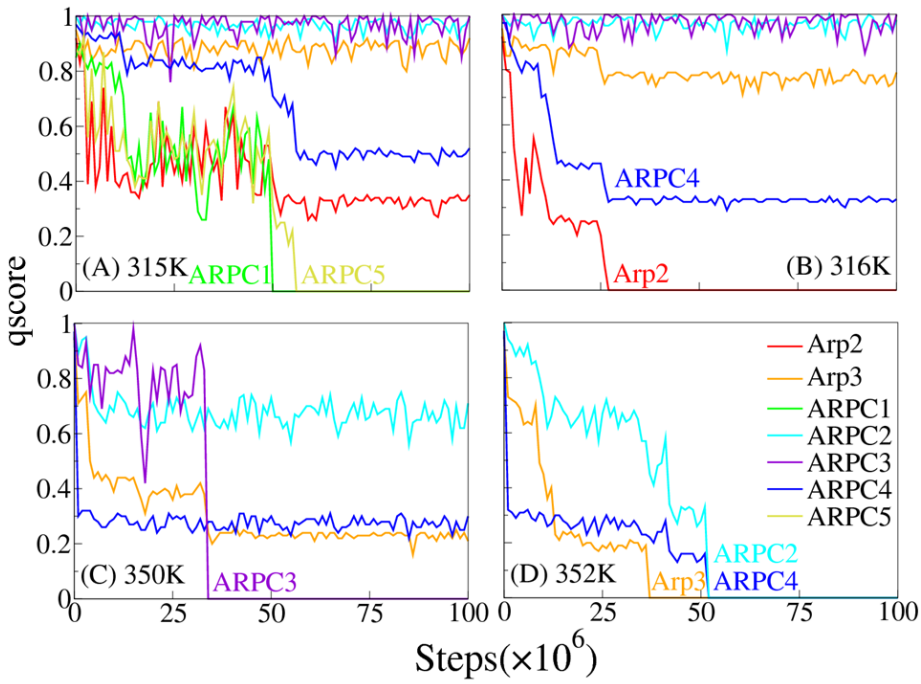

Figure 3 Time evolution of <interq> for every subunit in the Arp2/3 complex during the CG simulations at different temperatures. (A) 315 K, (B) 316 K, (C) 350 K, (D) 352 K.

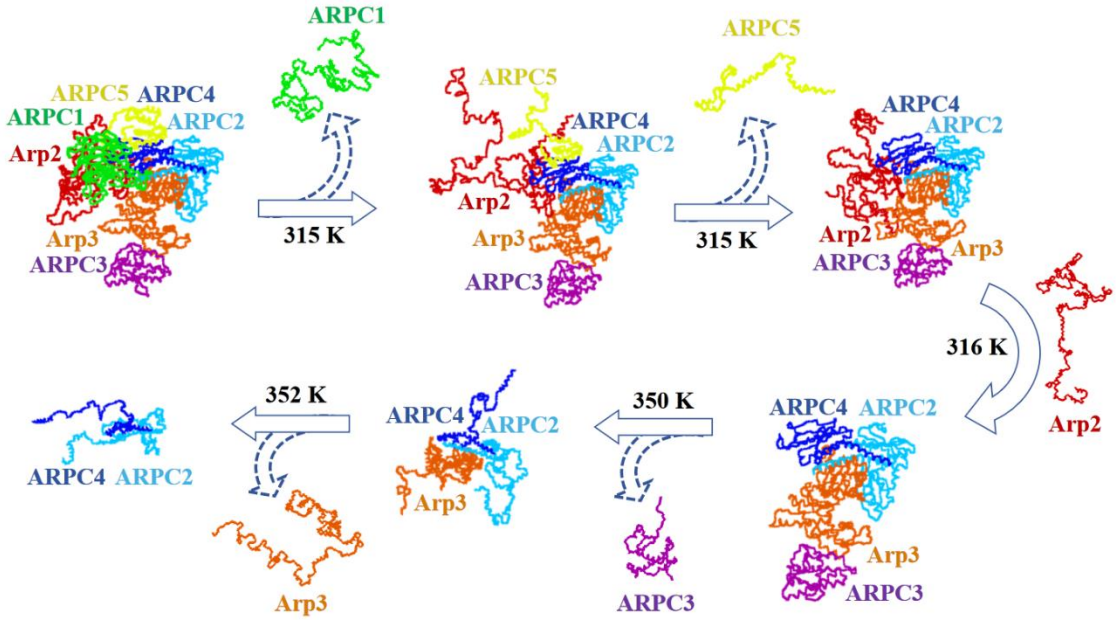

Figure 4 Snapshots during the disassembly process of the Arp2/3 complex at different temperatures.

According to the above analysis, the disassembly temperatures of the peripheral subunits are 315 K (ARPC1 and ARPC5), 316 K (Arp2), 350 K (ARPC3), and 352 K (Arp3). If two or more subunits had the same disassembly temperature, we determined

their disassembly order according to their <interq> plots (Figure 3). The procedure was repeated three times independently, and the same disassembly order of the subunits was obtained. Therefore, we predicted the assembly order of the subunits: ARPC2, ARPC4, Arp3, ARPC3, Arp2, ARPC5, and ARPC1.

The assembly order of the Arp2/3 complex predicted from the multi-temperature CG simulations is basically the same as that obtained from an in vitro experimental study. Reconstitution experiments indicated that ARPC2 and ARPC4 form a structural core of the complex(36), which defines the center of rotation and translation of two subcomplexes. One subcomplex consists of ARPC2, Arp3, and ARPC3, and the other consists of ARPC4, Arp2, ARPC1, and ARPC5(35). By purifying the Arp2/3 subcomplexes lacking one or more subunits, some information can be learned(36). Among the peripheral subunits, ARPC3 only interacts with Arp3. Arp2 is an independent peripheral subunit. ARPC1 and ARPC5 interact with each other but are independent of the other subunits.

An assembly pathway of the Arp2/3 complex

The assembly pathway of the Arp2/3 complex was investigated according to the predicted assembly order. Starting from the dimer of the core subunits ARPC2-ARPC4, a CG simulation at 300 K was conducted. The dimer is stable with an average RMSD of 2.1 Å.

In the first round, each of the peripheral subunits except ARPC3 was added to the core dimer, and the corresponding three-subunit subcomplex was used to run a CG simulation. ARPC3 was not considered in this round because it has no direct contact

with the core dimer. ARPC2-ARPC4-Arp3 is the most stable subcomplex in this round, with the smallest average RMSD of 2.0 Å (Figure 5a). Therefore, Arp3 may assemble first to form the core dimer, which is consistent with our predicted assembly order.

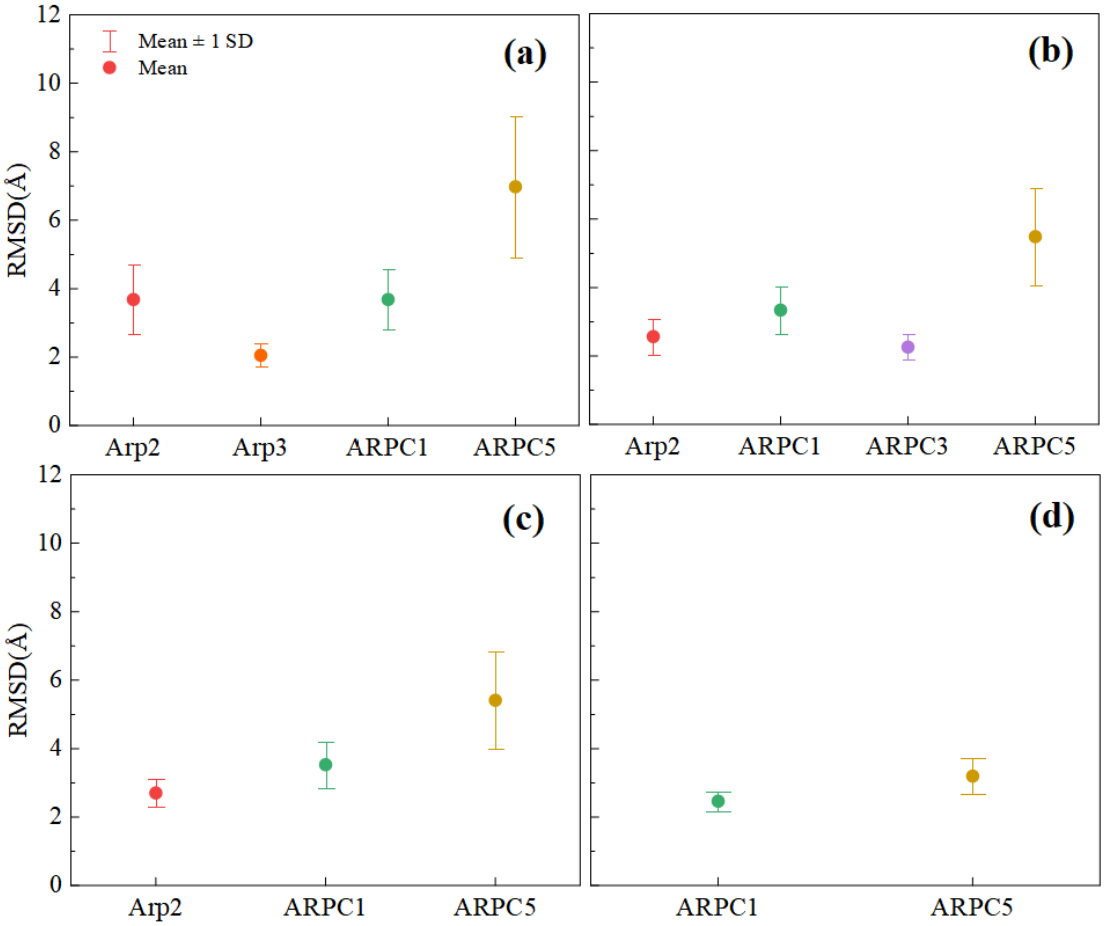

Figure 5 The stability of subcomplexes during the assembly of the Arp2/3 complex. The mean and standard deviation of the RMSD value of each subcomplex after adding a subunit from (a) the first to (d) the last round. The RMSD value of each CG simulation was calculated using the last 100 conformations of the trajectory.

In the second round, Arp2, ARPC1, ARPC3, and ARPC5 were added to the subcomplex ARPC2-ARPC4-Arp3, and a CG simulation was subsequently conducted for each four-subunit subcomplex. As shown in Figure 5b, the average RMSD of the subcomplex ARPC2-ARPC4-Arp3-ARPC3 is the smallest, which indicates that this complex has the highest stability. ARPC3 may assemble after Arp3.

We repeated the above procedure until a complete protein complex was formed.

In the third round, the most stable five-subunit subcomplex is identified as ARPC2-ARPC4-Arp3-ARPC3-Arp2 (Figure 5c). Arp2 may assemble after ARPC3. Finally, in the fourth round, ARPC5 may assemble before ARPC1 because the former can form a relatively more stable six-subunit subcomplex than the latter (Figure 5d). ARPC1 may be the last assembled subunit. Notably, although ARPC1 always has smaller average RMSD values than ARPC5 in rounds 1-3 (Figure 5a-5c), the latter has a smaller RMSD than the former after the addition of Arp2 to the subcomplex (Figure 5d). This may be because ARPC5 can form more native contacts with Arp2 (Table 1), thus increasing the stability of the subcomplex.

In summary, the most stable route for adding the subunits to the core dimer is Arp3, ARPC3, Arp2, ARPC5 and ARPC1, which is consistent with the assembly order predicted by CG simulations at different temperatures. This procedure can also be used to predict the assembly order of a protein complex, but it may require many more CG simulations than multi-temperature CG simulations.

Assembly order of the aRNAP complex

As another application of the method, the assembly order of the 13-subunit aRNAP was predicted. Table 2 shows the intersubunit native contact numbers of aRNAP. Rpo1N and Rpo2 has the largest number of contacts.

Table 2 Intersubunit native contacts in the atomic structure of aRNAP

Subunit	Rpo1N	Rpo2	Rpo1C	Rpo3	Rpo6	Rpo10	Rpo7	Rpo5	Rpo11	Rpo4	Rpo8	Rpo12	Rpo13	Sum
Rpo1N (872)		1081	588	1	147	0	6	73	93	1	183	0	37	2210
Rpo2 (1103)			129	200	8	223	24	0	32	8	0	87	0	1792

Rpo1C (376)	0	120	0	53	143	0	0	3	0	5	1041
Rpo3 (262)		0	132	0	0	171	0	0	69	0	573
Rpo6 (84)			0	56	40	0	0	0	0	0	371
Rpo10 (65)				0	0	0	0	0	13	0	368
Rpo7 (171)					0	0	189	0	0	0	328
Rpo5 (76)						0	0	0	0	68	324
Rpo11 (91)							0	0	0	0	296
Rpo4 (105)								0	0	0	198
Rpo8 (113)									0	0	186
Rpo12 (44)										0	169
Rpo13 (50)											110

During the optimization of the go_unit for aRNAP, Rpo4 was found to be highly flexible. Therefore, a go_unit of 1.0 was used in the CG simulations at the different temperatures.

At 400 K, only the <interq> of Rpo4 drops to zero (Figure 6A). Rpo4 mainly interacts with Rpo7 and has few contacts with other subunits (Table 2). Accordingly, the <interq> of Rpo7 is decreased significantly (Figure 6A), and the structure is unfolded after losing contact with Rpo4 (Figure 7). At 413 K, Rpo8, which is essentially in contact with Rpo1N, is disassembled (Figure 6B). The <interq> of Rpo1N is only decreased slightly because it has more than 2,200 contacts in total (Table 2). In the CG simulation at 420 K, Rpo7 eventually disassembles (Figure 6C). The remaining

subunits are nearly unaffected by the disassembly of the three subunits (Figure 7). Rpo12 and Rpo13 are then disassembled at 430 K and 450 K, respectively (Figure 6D and 6E). The subcomplex consisting of the remaining eight subunits is fairly stable (Figure 7) with relatively high qscores. Up to 475 K, Rpo5 is disassembled, and the <interq> values of some other subunits are decreased (Figure 6F). The subcomplex consisting of the remaining seven subunits becomes unstable at this temperature (Figure 7). At 494 K, Rpo6 and Rpo1C disassemble sequentially (Figure 6G). At 499 K, Rpo11 and Rpo3 are disassembled nearly at the same time (Figure 6H). The remaining three subunits in the subcomplex are further unfolded (Figure 7). At 510 K, Rpo1N disassembles (Figure 6I), and the last two subunits constitute the Rpo2-Rpo10 dimer (Figure 7). At 511 K, the dimer is ultimately disassembled.

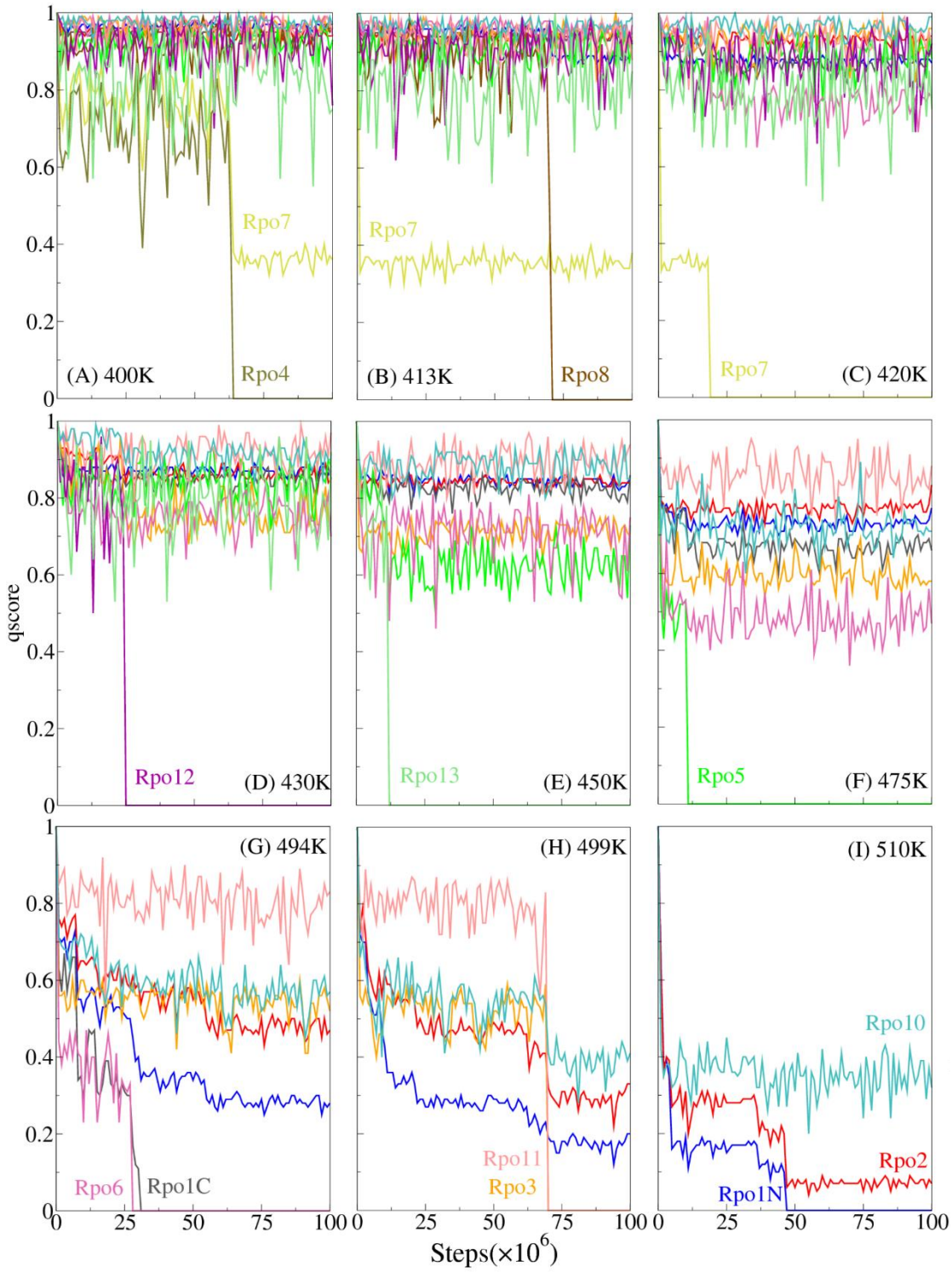

Figure 6 Time evolution of <interq> for every subunit in aRNAP during the CG simulations at different temperatures. (A) 400 K, (B) 413 K, (C) 420 K, (D) 430 K, (E) 450 K, (F) 475 K, (G) 494 K, (H) 499 K, (I) 510 K.

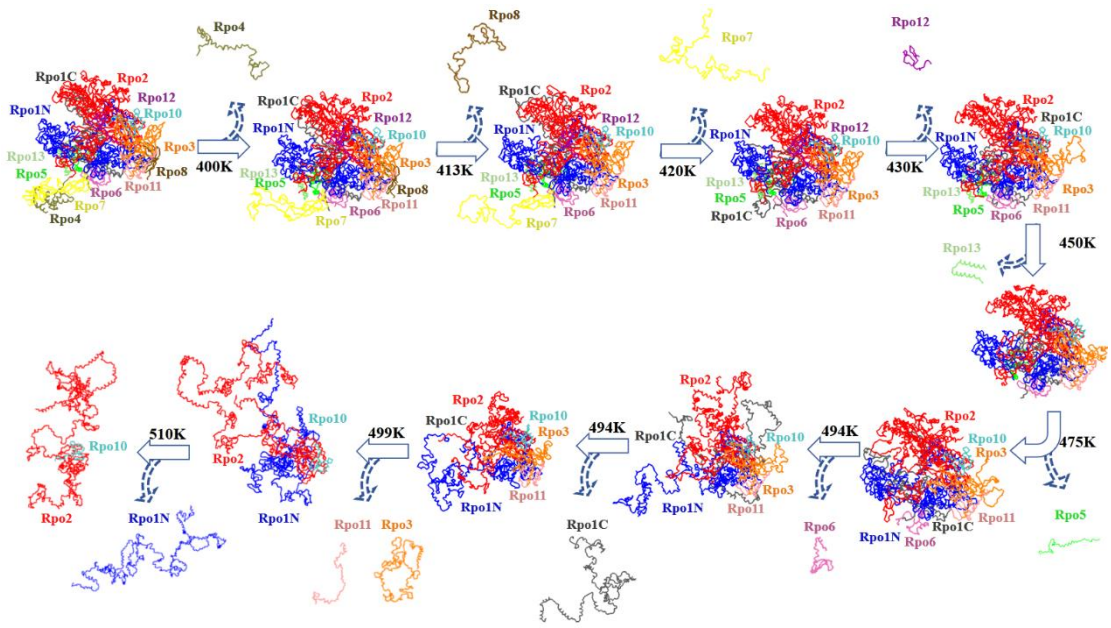

Figure 7 Snapshots during the disassembly process of aRNAP at different temperatures.

In summary, the disassembly temperatures of the subunits in aRNAP are 400 K (Rpo4), 413 K (Rpo8), 420 K (Rpo7), 430 K (Rpo12), 450 K (Rpo13), 475 K (Rpo5), 494 K (Rpo6 and Rpo1C), 499 K (Rpo11 and Rpo3), 510 K (Rpo1N), and 511 K (Rpo2 and Rpo10). If two or more subunits have the same disassembly temperature, their disassembly order is determined by their <interq> plots. The disassembly order of aRNAP was tested through three independent multi-temperature CG simulations, and the same results were obtained. Therefore, we predicted the following order of aRNAP assembly: Rpo2, Rpo10, Rpo1N, Rpo3, Rpo11, Rpo1C, Rpo6, Rpo5, Rpo13, Rpo12, Rpo7, Rpo8, and Rpo4.

In the native structure of aRNAP(37), Rpo1N and Rpo2 has the greatest number of contacts (Table 2). Taken together, our results indicate that Rpo2 binds to Rpo10 first and then assembles with Rpo1N. Rpo10 may somewhat stabilize the structure of Rpo2. Rpo1N and Rpo2 from a clamp, and the assembly of Rpo3 and Rpo11 may strengthen the clamp. Rpo10 forms a stable complex with Rpo3-Rpo11 to constitute an assembly

platform, which is vital for the efficient assembly and stability of archaeal RNA(39). Once Rpo1C is assembled, the catalytic core of aRNAP is formed. Rpo1 (including Rpo1N and Rpo1C) and Rpo2 account for more than two-thirds of the molecular mass of the protein complex and can bind to DNA to initiate transcription. Rpo6 stabilizes the split of Rpo1N and Rpo1C and serves as an anchor point for the Rpo4-Rpo7 complex. Rpo5 and a fragment of Rpo1 form the lower jaw domain of aRNAP(39). Table 2 shows that Rpo5 has more than 200 native contacts with Rpo1N and Rpo1C. Rpo13 forms a complex with Rpo5, and Rpo12 is located close to the catalytic core. The remaining three subunits (Rpo7, Rpo8, and Rpo4) are all peripheral subunits. It is known from the literature(37,39) that Rpo7 and Rpo4 form a stalk sticking out of the aRNAP structure. Several experimental analyses have indicated that the stalk is the most dynamic module and is anchored to aRNAP via Rpo6(38). Rpo7 is in contact with multiple subunits, while Rpo4 is mainly in contact with Rpo7 (Table 2). Rpo8 is located at the underside of aRNAP between the assembly platform and the core(39). (42)Although it interacts with Rpo1N, it is located in a peripheral position and has a partially disordered sequence (Figure 7).

Factors determining the assembly order of a protein complex

After predicting the assembly order of the two protein complexes, we discuss the factors that determine the assembly order of the subunits. One is the contact number of a subunit with other subunits. Generally, these subunits with more intersubunit contacts, such as ARPC2, ARPC4, and Arp3 in the Arp2/3 complex and Rpo2, Rpo1N, and Rpo1C in aRNAP, assemble early.

However, several exceptions were observed. Among the Arp2/3 complex members, ARPC3 has the least number of intersubunit contacts (Table 1), but it assembles before Arp2, ARPC5, and ARPC1. In aRNAP, Rpo10 has relatively few intersubunit contacts (Table 2), but it is the first subunit that binds to Rpo2. Rpo3 has 573 intersubunit contacts in total, and the number of contacts with Rpo11 is only 296; however, the two subunits can assemble before Rpo1C. Rpo12 and Rpo13 have the least numbers of intersubunit contacts, but they assemble before Rpo7, Rpo8, and Rpo4.

Early-assembly proteins without a large number of intersubunit contacts are generally stable, which means that their intrasubunit contacts are tolerant of relatively high temperatures. For the two protein complexes, we computed the T_f of each protein subunit (Table 3) using CafeMol(29), which is the temperature at which 50% of the intrasubunit contacts are broken. Among the seven subunits of the Arp2/3 complex, ARPC3 has the second highest T_f. In aRNAP, Rpo10 has the second highest T_f among all the subunits. The T_f of Rpo13 is the highest, and those of Rpo12, Rpo3, and Rpo11 are also relatively high.

Therefore, our results suggest that the assembly order of a protein subunit may be determined by both the number of intersubunit contacts and the stability of its intrasubunit contacts.

Table 3 T_f of subunits in the protein complexes

RNAP	T_f (K)	Arp2/3	T_f (K)
Rpo6	346	**Arp2**	395
Rpo8	355	**ARPC1**	405
Rpo4	365	**ARPC5**	407
Rpo7	370	**ARPC4**	418

Rpo5	373	**ARPC2**	430
Rpo11	384	**ARPC3**	445
Rpo1C	398	**Arp3**	455
Rpo3	415		
Rpo12	418		
Rpo1N	471		
Rpo2	484		
Rpo10	509		
Rpo13	649		

Conclusion

Protein complexes play crucial roles in nearly all biological processes. The objective of this work was to examine the order in which subunits assemble within a protein complex. We propose a framework for predicting assembly processes through CG simulations at multiple temperatures. This technique allows for a comprehensive examination of the disassembly order and dynamic procedure of a protein complex, comprising definable intermediate subcomplexes, and enables subsequent prediction of the assembly order. The results obtained for the Arp2/3 complex and aRNAP are consistent with previous experimental data.

Despite the simplicity of this method for predicting assembly order from a CG perspective, it aims to provide helpful insights into the assembly and disassembly of protein complexes and to investigate the potential molecular mechanism responsible for the assembly of protein complexes. It should be noted that the disassembly temperatures of subunits obtained by CG simulations do not correspond to the actual temperatures, which are used only to judge the disassembly order of the subunits. This is also the case for T_f (Table 3). For a homomeric protein complex, it has been found

that the subunits tend to disassemble simultaneously(6) in CG simulations; thus, our method may be more suitable for predicting the assembly order of heteromeric protein complexes.

In the future, we hope to expand the applicability of this method to additional biomolecular complexes, such as protein–DNA and protein–RNA complexes, by taking advantage of proper DNA/RNA CG models. For the two protein complexes we tested, a converged assembly order was obtained for each complex. Can the method predict multiple assembly paths when it is applied to other complexes? This is one of the future topics to be addressed.

Author Contributions

Z.Z. and Y.L. designed the project. Y.L. conducted the major simulations and analyzed the data. X.L. conducted the assembly simulations of the Arp2/3 complex. Y.L. and Z.Z. wrote the manuscript.

Conflicts of interest

The authors declare that they have no competing interests.

Acknowledgment

This work is supported by the National Key Research and Development Program of China (2021YFA1301504), the National Natural Science Foundation of China (91953101), and the Chinese Academy of Sciences Strategic Priority Research Program (XDB37040202). The Supercomputing Center of the USTC provided computer resources for this project, and we are grateful to Mr. Yundong Zhang for his technical support.

References

1. Sali, A., R. Glaeser, T. Earnest, and W. Baumeister. 2003. From words to literature in structural proteomics. *Nature*. 422:216-225.
2. Guaita, M., S. C. Watters, and S. Loerch. 2022. Recent advances and current trends in cryo-electron microscopy. *Curr Opin Struct Biol*. 77:102484.
3. Shi, Y. 2014. A glimpse of structural biology through X-ray crystallography. *Cell*. 159:995-1014.
4. Gao, M., D. Nakajima An, J. M. Parks, and J. Skolnick. 2022. AF2complex predicts direct physical interactions in multimeric proteins with deep learning. *Nat Commun*. 13:1744.
5. Marsh, J. A., and S. A. Teichmann. 2015. Structure, dynamics, assembly, and evolution of protein complexes. *Annu Rev Biochem*. 84:551-575.
6. Ahnert, S. E., J. A. Marsh, H. Hernandez, C. V. Robinson, and S. A. Teichmann. 2015. Principles of assembly reveal a periodic table of protein complexes. *Science*. 350:aaa2245.
7. Ellis, R. J. 2007. Protein misassembly. In Molecular aspects of the stress response: Chaperones, membranes and networks. P. Csermely, and L. Vígh, editors. Springer New York, New York, NY, pp. 1-13.
8. Tompa, P., and G. D. Rose. 2011. The levinthal paradox of the interactome. *Protein Sci*. 20:2074-2079.
9. Zhu, J., N. Avakyan, A. Kakkis, A. M. Hoffnagle, K. Han, Y. Li, Z. Zhang, T. S. Choi, Y. Na, C. J. Yu, and F. A. Tezcan. 2021. Protein assembly by design. *Chem Rev*. 121:13701-13796.
10. Bergendahl, L. T., L. Gerasimavicius, J. Miles, L. Macdonald, J. N. Wells, J. P. I. Welburn, and J. A. Marsh. 2019. The role of protein complexes in human genetic disease. *Protein Science*. 28:1400-1411.
11. Kennedy, K. A., E. G. Gachelet, and B. Traxler. 2004. Evidence for multiple pathways in the assembly of the escherichia coli maltose transport complex. *J Biol Chem*. 279:33290-33297.
12. Britt, H. M., T. Cragnolini, and K. Thalassinos. 2022. Integration of mass spectrometry data for structural biology. *Chemical Reviews*. 122:7952-7986.
13. Zhang, S., S. Zou, D. Yin, L. Zhao, D. Finley, Z. Wu, and Y. Mao. 2022. Usp14-regulated allostery of the human proteasome by time-resolved cryo-em. *Nature*. 605:567-574.
14. Bansal, P. K., R. Abdulle, and K. Kitagawa. 2004. Sgt1 associates with hsp90: An initial step of assembly of the core kinetochore complex. *Mol Cell Biol*. 24:8069-8079.
15. Ishii, K., M. Zhou, and S. Uchiyama. 2018. Native mass spectrometry for understanding dynamic protein complex. *Biochim Biophys Acta Gen Subj*. 1862:275-286.
16. Hernandez, H., and C. V. Robinson. 2007. Determining the stoichiometry and interactions of macromolecular assemblies from mass spectrometry. *Nat Protoc*.

2:715-726.
17. Chorev, D. S., L. A. Baker, D. Wu, V. Beilsten-Edmands, S. L. Rouse, T. Zeev-Ben-Mordehai, C. Jiko, F. Samsudin, C. Gerle, S. Khalid, A. G. Stewart, S. J. Matthews, K. Grunewald, and C. V. Robinson. 2018. Protein assemblies ejected directly from native membranes yield complexes for mass spectrometry. *Science*. 362:829-834.
18. Levy, E. D., E. Boeri Erba, C. V. Robinson, and S. A. Teichmann. 2008. Assembly reflects evolution of protein complexes. *Nature*. 453:1262-1265.
19. Marsh, J. A., H. Hernandez, Z. Hall, S. E. Ahnert, T. Perica, C. V. Robinson, and S. A. Teichmann. 2013. Protein complexes are under evolutionary selection to assemble via ordered pathways. *Cell*. 153:461-470.
20. Chen, J., N. Sawyer, and L. Regan. 2013. Protein–protein interactions: General trends in the relationship between binding affinity and interfacial buried surface area. *Protein Sci*. 22:510-515.
21. Peterson, L. X., Y. Togawa, J. Esquivel-Rodriguez, G. Terashi, C. Christoffer, A. Roy, W. H. Shin, and D. Kihara. 2018. Modeling the assembly order of multimeric heteroprotein complexes. *PLoS Comput Biol*. 14:e1005937.
22. Esquivel-Rodriguez, J., Y. D. Yang, and D. Kihara. 2012. Multi-lzerd: Multiple protein docking for asymmetric complexes. *Proteins*. 80:1818-1833.
23. Jones, G. 1998. Genetic and evolutionary algorithms. In Encyclopedia of computational chemistry.
24. Kurisaki, I., and S. Tanaka. 2022. Computational prediction of heteromeric protein complex disassembly order using hybrid monte carlo/molecular dynamics simulation. *Phys Chem Chem Phys*. 24:10575-10587.
25. Takada, S., R. Kanada, C. Tan, T. Terakawa, W. Li, and H. Kenzaki. 2015. Modeling structural dynamics of biomolecular complexes by coarse-grained molecular simulations. *Acc Chem Res*. 48:3026-3035.
26. Pak, A. J., and G. A. Voth. 2018. Advances in coarse-grained modeling of macromolecular complexes. *Curr Opin Struct Biol*. 52:119-126.
27. Takada, S. 2012. Coarse-grained molecular simulations of large biomolecules. *Curr Opin Struct Biol*. 22:130-137.
28. Li, W., W. Wang, and S. Takada. 2014. Energy landscape views for interplays among folding, binding, and allostery of calmodulin domains. *Proc Natl Acad Sci U S A*. 111:10550-10555.
29. Kenzaki, H., N. Koga, N. Hori, R. Kanada, W. Li, K. Okazaki, X. Q. Yao, and S. Takada. 2011. Cafemol: A coarse-grained biomolecular simulator for simulating proteins at work. *J Chem Theory Comput*. 7:1979-1989.
30. Levy, E. D., J. B. Pereira-Leal, C. Chothia, and S. A. Teichmann. 2006. 3d complex: A structural classification of protein complexes. *PLoS Comput Biol*. 2:e155.
31. Zheng, S., F. Qin, J. Yin, D. Li, Y. Huang, L. Hu, L. He, C. Lv, X. Li, S. Li, and W. Hu. 2023. Role and mechanism of actin-related protein 2/3 complex signaling in cancer invasion and metastasis: A review. *Medicine (Baltimore)*. 102:e33158.

32. Robinson, R. C., K. Turbedsky, D. A. Kaiser, J. B. Marchand, H. N. Higgs, S. Choe, and T. D. Pollard. 2001. Crystal structure of arp2/3 complex. *Science*. 294:1679-1684.
33. Webb, B., and A. Sali. 2016. Comparative protein structure modeling using modeler. *Current Protocols in Bioinformatics*. 54:5.6.1-5.6.37.
34. Pollard, T. D., and C. C. Beltzner. 2002. Structure and function of the arp2/3 complex. *Curr Opin Struct Biol*. 12:768-774.
35. Fassler, F., G. Dimchev, V. V. Hodirnau, W. Wan, and F. K. M. Schur. 2020. Cryo-electron tomography structure of arp2/3 complex in cells reveals new insights into the branch junction. *Nat Commun*. 11:6437.
36. Gournier, H., E. D. Goley, H. Niederstrasser, T. Trinh, and M. D. Welch. 2001. Reconstitution of human arp2/3 complex reveals critical roles of individual subunits in complex structure and activity. *Mol Cell*. 8:1041-1052.
37. Wojtas, M. N., M. Mogni, O. Millet, S. D. Bell, and N. G. Abrescia. 2012. Structural and functional analyses of the interaction of archaeal rna polymerase with DNA. *Nucleic Acids Res*. 40:9941-9952.
38. Werner, F., and D. Grohmann. 2011. Evolution of multisubunit rna polymerases in the three domains of life. *Nat Rev Microbiol*. 9:85-98.
39. Korkhin, Y., U. M. Unligil, O. Littlefield, P. J. Nelson, D. I. Stuart, P. B. Sigler, S. D. Bell, and N. G. Abrescia. 2009. Evolution of complex rna polymerases: The complete archaeal rna polymerase structure. *PLoS Biol*. 7:e1000102.
40. Hirata, A., B. J. Klein, and K. S. Murakami. 2008. The X-ray crystal structure of rna polymerase from archaea. *Nature*. 451:851-854.
41. Jun, S. H., M. J. Reichlen, M. Tajiri, and K. S. Murakami. 2011. Archaeal rna polymerase and transcription regulation. *Crit Rev Biochem Mol Biol*. 46:27-40.
42. Korkhin, Y., U. M. Unligil, O. Littlefield, P. J. Nelson, D. I. Stuart, P. B. Sigler, S. D. Bell, and N. G. A. Abrescia. 2009. Evolution of complex rna polymerases: The complete archaeal rna polymerase structure. *PLOS Biology*. 7:e1000102.