Molecular dynamics simulation of the self-assembly of icosahedral virus

Jingzhi Chen

UNIVERSITÉ PARIS SUD
Comprendre le monde, construire l'avenir

Molecular Dynamics Simulation of the Self-assembly of Icosahedral Virus

Thèse de doctorat de l'Université Paris-Saclay
préparée à l'Université Paris-Sud

École doctorale n◦ 564: Physique en Île de France (PIF)
Spécialité de doctorat: Physique

Thèse présentée et soutenue à Orsay, le 24 Septembre 2019, par

M. Jingzhi CHEN

Composition du Jury :

René Messina Université de Lorraine	Professeur	Président
Martin Castelnovo ENS Lyon	Chargé de Recherche	Rapporteur
Catherine Etchebest Université Paris-Diderot	Professeur	Rapporteur
Yves Boulard CEA Saclay	Directeur de Recherche	Examinateur
Guillaume Tresset Laboratoire de Physique des Solides, Université Paris-Sud	Chargé de Recherche	Directeur de thèse
Yves Lansac GREMAN, Université de Tours	Maître de Conférences	Co-Directeur de thèse

Abstract

Viruses are known for infecting all classes of living organisms on Earth, whether vegetal or animal. Virions consist of a nucleic acid genome protected by a single or multilayered protein shell called capsid, and in some cases by an envelope of lipids. The viral capsid is generally made of hundreds or thousands of proteins forming ordered structures. Half of all known viruses exhibit an icosahedral symmetry, the rest being helical, prolate or having a complex irregular structure. Recently, viral particles have attracted an increasing attention due to their extremely regular structure and their potential use for fabricating nanostructures with various functions. Therefore, understanding the assembly mechanisms underlying the production of viral particles is not only helpful to the development of inhibitors for therapeutic purpose, but it should also open new routes for the self-assembly of complex supramolecular materials.

To date, numerous experimental and theoretical investigations on virus assembly have been performed. Through experimental investigations, a lot of information have been obtained on virus assembly, including the proper conditions required for the assembly and the kinetic pathways. Combining those information and theoretical methods, an initial understanding of the assembly mechanism of viruses has been worked out. However, information coming purely from experiments cannot give the whole picture, in particular at a microscopic scale. Therefore, in this thesis, we employed computer simulations, including Monte Carlo and molecular dynamics techniques, to probe the assembly of virus, with the expectation to gain new insights into the molecular mechanisms at play.

This thesis is organized as followed:

Chapter 1 is a literature review on viruses with a special emphasis on Cowpea Chlorotic Mottle Virus (CCMV), the viral model that we selected as a study target. A brief introduction on viruses is firstly given, including their structures, the mechanisms involved during their assembly, and the typical experimental techniques and the theoretical approaches employed so far. Secondly, we introduce the CCMV in details, including its structure and its frontier researches.

In Chapter 2, the principles of the methods used in this thesis, that is, Monte Carlo and molecular dynamics simulations, are presented.

In Chapter 3, a method to obtain the interactions between subunits is proposed. Fluorescence thermal shift assay was used to measure the thermal dissociation temperature of CCMV capsids, which subsequently was employed to extract the interaction strength between subunits with the assistance of a theoretical lattice model that we have developed. The effect of the relative strength of the interactions on the stability of viruses is analyzed as well. In addition, the thermal shrinkage phenomenon of CCMV capsids is also studied.

In Chapter 4, the early steps of the assembly process of CCMV are investigated by atomistic molecular dynamics simulations to get insight into the interactions involved at the molecular level. Both the interaction between dimers and the interaction between a dimer and RNA are probed. Besides, the interactions of a dimer within different subassemblies are also compared.

Résumé en français

Introduction

Un virus est une capsule biologique minuscule et simple, qui ne peut vivre que dans les cellules hôtes et proliférer par auto-réplication. Il a une taille allant de quelques dizaines de nanomètres à plusieurs centaines de nanomètres et plus de la moitié des virus connus ont une forme icosahédrique, le reste pouvant être hélicoïdal, prolongé, etc. Les particules virales les plus simples ont juste une nucléocapside, une association de génome viral avec une capside protectrice de polypeptide, comme le montre la Figure F1 (A). La capside est constituée de centaines voire de milliers de protéines identiques, codées par le génome viral. Ces protéines identiques agissent comme des sous-unités d'assemblage de base appelées capsomères et sont organisées dans un ordre régulier pour former une capside avec ou sans l'aide du génome viral. Cependant, ce qui est plus général, c'est que les virus ont une structure beaucoup plus compliquée, telle qu'une enveloppe lipidique supplémentaire dérivée de la membrane de la cellule hôte avec un certain nombre de protéines fonctionnelles à la surface, comme le montre la Figure F1 (B). Les dommages potentiels causés par des virus à des créatures ont suscité une attention considérable de la part des chercheurs en médecine. Cependant, les particules virales trouvent une autre application potentielle dans les recherches sur les matériaux en raison de la structure délicate de ses capsides, qui permettra de mieux comprendre l'auto-assemblage des nanomatériaux.

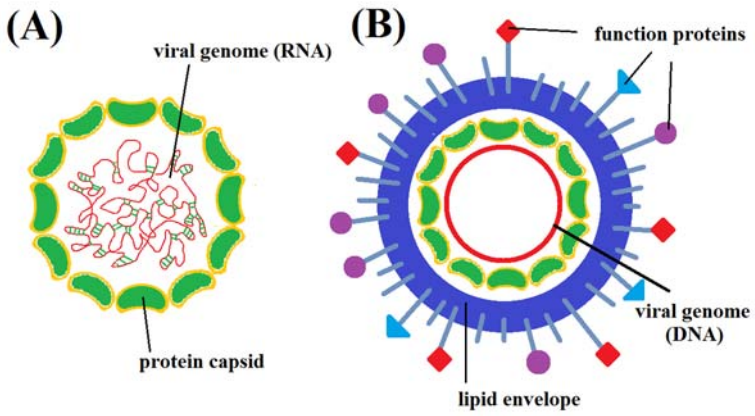

Figure F1. Schéma de la structure des virus. (A) Les virus les plus simples consistent en une capside de virus et le génome viral enrobé. (B) Virus avancé avec une enveloppe lipidique supplémentaire.

À ce jour, un certain nombre de recherches ont été effectuées pour révéler le mécanisme à l'origine de la formation de capsides virales à structure délicate. Cependant, le mécanisme d'assemblage des capsides virales reste controversé. Il n'existe aucun mécanisme générique capable d'expliquer de manière exhaustive l'assemblage de divers virus, ce qui indique la complexité inhérente à la réaction d'assemblage. Par conséquent, pour bien comprendre

l’assemblage du virus, nous devrions commencer par des virus simples. Deux mécanismes sont fréquemment signalés pour l'auto-assemblage de virus simples, le mécanisme de nucléation-croissance et le mécanisme de coopération. Dans le mécanisme de nucléation-croissance, un seul noyau se forme lentement par agrégation de sous-unités, puis croît rapidement par addition successive de dimères libres jusqu'à l'achèvement de la capside, tandis qu'une nucléation rapide suivie d'une lente association coopérante d'intermédiaires pour former une capside ordonnée dans le mécanisme de coopération. Cependant, le mécanisme d'assemblage variera en fonction du ratio de résistance entre l'interaction sous-unité-sous-unité et l'interaction sous-unité-génome, comme l'indiquent les simulations à grain grossier de Hagan et al.

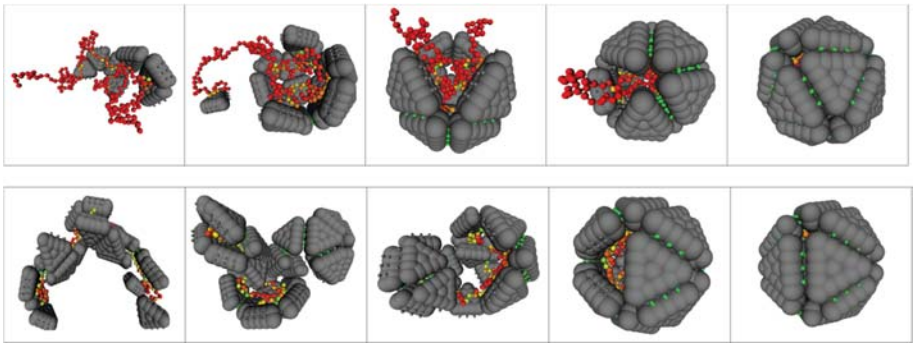

Figure F2. Deux mécanismes d'assemblage autour d'un polymère. (Rangée du haut) Mécanisme de nucléation-croissance lorsque les interactions sous-unité-sous-unité sont fortes, tandis que les interactions sous-unité-polymère sont relativement faibles. (Rangée inférieure) Un mécanisme d'assemblage désordonné ou un mécanisme coopératif lorsque les interactions sous-unité-sous-unité sont plus faibles et que les interactions sous-unité-polymère sont plus fortes. Dans ce mécanisme, l'assemblage commence par la formation d'un complexe désordonné composé de plusieurs sous-unités et d'un polymère au premier stade, suivi d'un recuit de multiples intermédiaires et enfin de son achèvement.

Cowpea Chlorotic Mottle Virus (CCMV)

CCMV est un virus icosaédrique T = 3 consistant en une couche de protéine monocouche recouvrant son génome d'ARN à l'intérieur. Comparé aux virus animaux, le CCMV est structurellement plus simple et plus sûr à manipuler, ce qui démontre un bon potentiel en tant que modèle initial pour l’étude de l’assemblage du virus. Le CCMV a un diamètre de 28 nm et une structure typique comme la majorité des virus, une chaîne protéique encapsulant le capuchon de la protéine (chaînes d'ARN pour le CCMV). La capside du CCMV a la forme d'un icosaèdre avec 180 molécules de protéines identiques uniformément réparties à la surface. Selon leur différence environnementale spécifique (située dans les pentamères ou les hexamères), ces protéines sont classées en trois types, A, B et C. La capside peut être divisée en différents sous-domaines, 12 pentagones et 20 hexagones, et présente un degré élevé de symétrie comme le montre la Figure F3. Le génome de CCMV est un ARN simple brin composé de quatre types de chaînes d'ARN, ARN1, ARN2, ARN3 et ARN4 dans trois formats d'ARN1, ARN2 et ARN3 / ARN4.

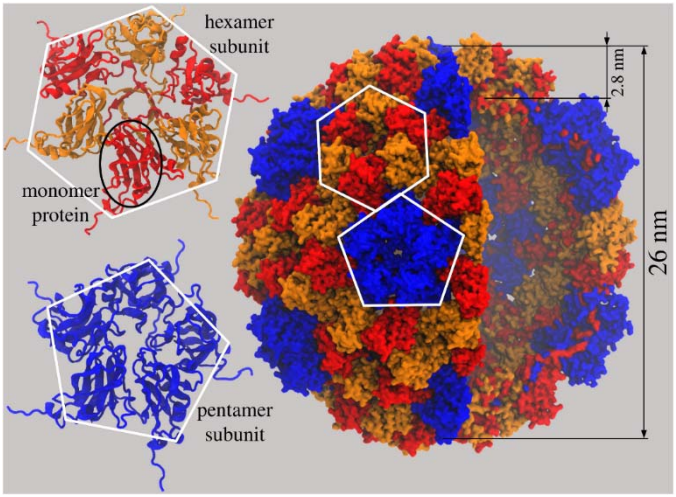

Figure F3. La structure du CCMV. Les protéines A, B et C sont en bleu, rouge et orange, respectivement. Les hexamères et les pentamères du dimère sont marqués et affichés à gauche.

Le montage ou le démontage des CCMV in vitro peut être facilement déclenché par une manipulation des facteurs environnementaux, tels que le pH et les forces ioniques. La sous-unité d'assemblage de base du CCMV est un dimère de protéines de capside. Sans la présence de génomes viraux, ces protéines de capside sont encore capables de s'auto-assembler en capsides vides selon un mécanisme classique de croissance par nucléation, comme le montre la Figure F4. En phase de nucléation, l'ajout de dimères de protéines de capside est défavorable jusqu'à ce que le noyau critique soit atteint. Les ajouts ultérieurs (phase de croissance) sont relativement favorables, bien que toujours réversibles, jusqu'à ce que la capside soit terminée.

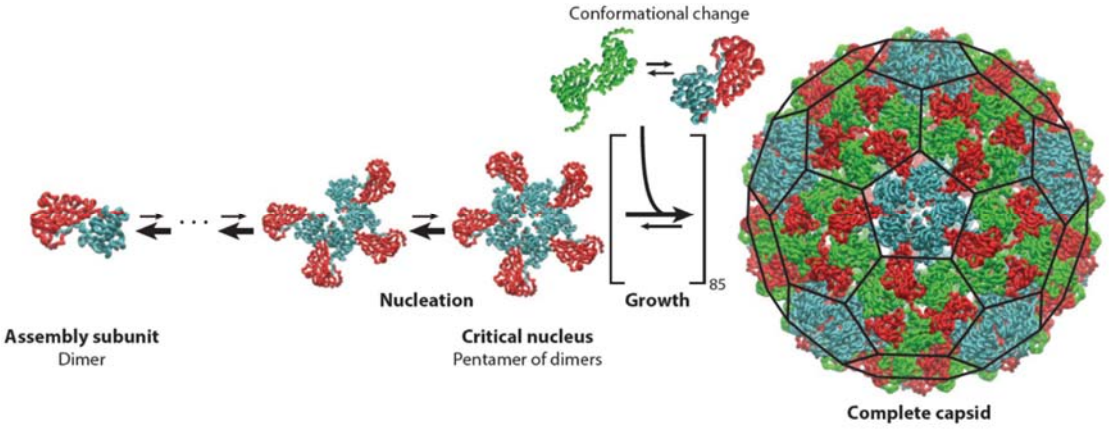

Figure F4. Illustration schématique du mécanisme d'assemblage pour CCMV.

En raison de la résolution limitée des techniques expérimentales, il est difficile de surveiller la réaction en phase de nucléation. Par conséquent, la plupart des expériences ont été engagées dans la phase de croissance et de nombreuses informations, telles que les interactions entre sous-unités, en phase de nucléation deviennent mal connues. Cependant, ces informations sont fondamentales pour nous permettre de comprendre parfaitement l'auto-assemblage des capsides virales. De plus, la précision des modèles développés pour décrire les réactions physiques des capsides virales dépend également fortement de la détermination précise des paramètres d'interaction entre les sous-unités. Dans cette thèse, nous nous sommes engagés à

révéler les interactions entre les différents composants formés au début de la réaction d'auto-assemblage par des approches informatiques.

Étude principale

Démontage thermique des capsides CCMV

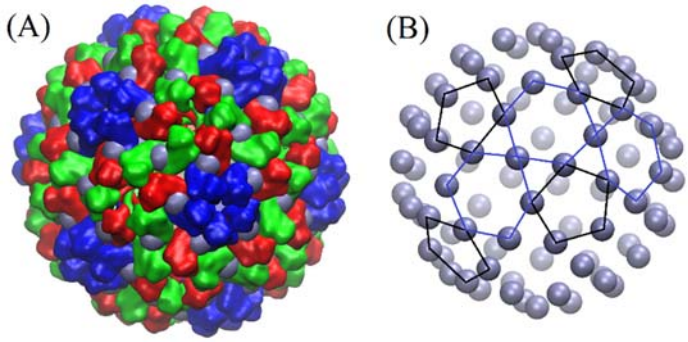

Figure F5. (A) Schéma de maillage d'une capside CCMV dans un modèle de réseau. Les protéines de la capside A, B et C sont respectivement en bleu, en rouge et en vert. Les sites de réseau sont situés au centre de masse de chaque dimère et représentés par une particule grise. (B) Les sites de réseau d'une capside du CCMV avec les dimères formant les pentamères et les hexamères sont liés par des lignes noires et bleues, respectivement.

Pour interpréter le désassemblage thermique des CCMV, nous avons proposé un modèle de réseau (voir Figure F5 (A) et (B)) où le dimère est représenté par un cordon qui interagit par le biais d'une attraction hydrophobe à courte portée et d'une répulsion électrostatique de Yukawa par paire. Avec ce modèle, nous avons estimé une énergie d'interaction hydrophobe de -4,38 k_BT_r pour les dimères CCMV, proche de la valeur de -5 ~ -10 k_BT_r prédit par l'analyse de la cinétique d'auto-assemblage. De plus, la charge effective de chaque dimère peut également être facilement estimée par le modèle. En considérant la dépendance en température des interactions hydrophobes, la charge effective peut être correctement estimée à -2 pour le dimère CCMV, ce qui correspond à la valeur estimée par la mobilité électrophorétique des virions CCMV.

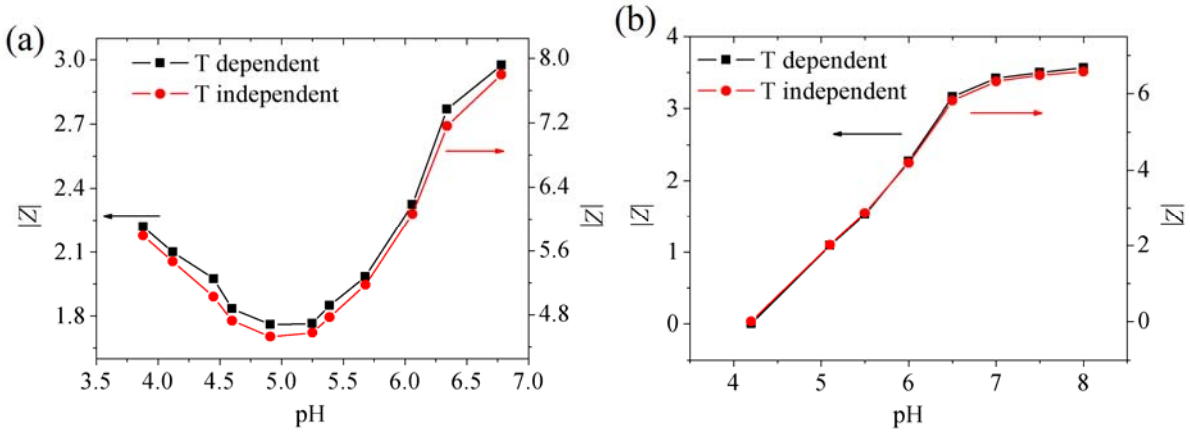

Figure F6. Valeur absolue de la charge effective | Z | d'une sous-unité constituant soit une capside CCMV vide (a), soit un virion CCMV (b) en fonction du pH estimé par les simulations MC à une salinité de 0,5 M avec les températures de fusion expérimentalement mesurées.

L'influence de la dépendance en température de l'interaction hydrophobe est analysée dans les deux cas: indépendante de la température (puce rouge) et dépendante de la température (carré noir).

La réponse du pH de la charge de dimère peut facilement être déterminée à l'aide de ce modèle, comme indiqué sur la Figure F6 (A) pour les capsides vides et (B) pour les capsides chargées d'ARN. Apparemment, la capside vide et la capside chargée d'ARN ont démontré une réponse totalement différente au pH. Dans le cas de capsides vides, le dimère présente un minimum différent de zéro autour de son point isoélectrique, ce qui indique que la répartition de la charge sur un dimère est de type dipolaire. Cependant, ce schéma disparaît dans le cas des capsides chargées en ARN, ce qui reflète le fait que les charges positives des queues riches en arginine ont été neutralisées par la charge négative de l'ARN.
Pour tenir compte de l'interaction asymétrique entre les sous-unités de la capside, nous avons proposé un modèle de réseau hétérogène, dans lequel deux composants ont été introduits. Comme le montre la Figure F7 (A), la symétrie de l'interaction entre les sous-unités affecte la voie de désassemblage des capsides virales. Lorsque l'interaction hydrophobe entre les sous-unités AB et CC était beaucoup plus faible que celle entre les sous-unités AB, la capside se dissociait par un processus en deux étapes, les sous-unités CC se dissociant d'abord à une température inférieure à celle des sous-unités AB (voir Figure F7 (B)). À son tour, la capside s'est dissociée en un processus précis en une étape lorsque les interactions hydrophobes étaient toutes comparables et que les deux types de sous-unités se sont dissociés simultanément.

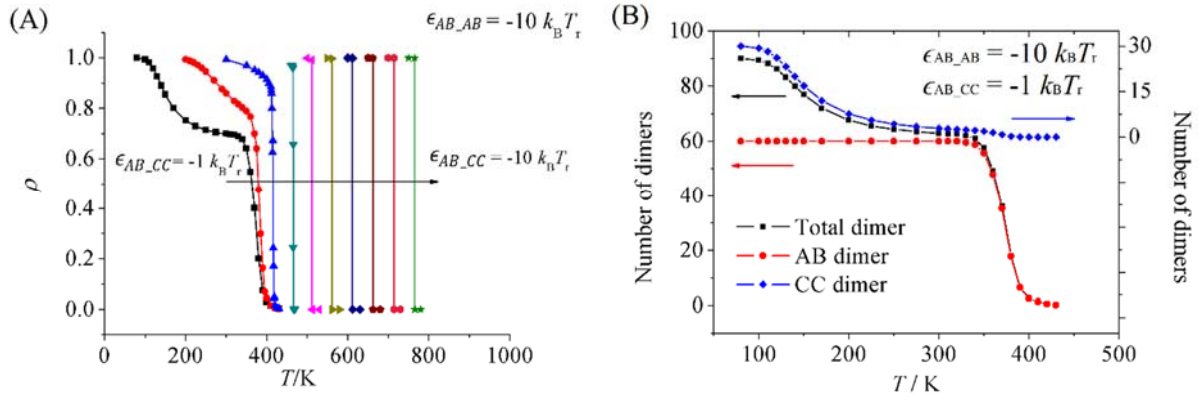

Figure F7. Courbes de fusion de la capside CCMV vide calculées par des simulations MC pour les sous-unités avec | Z | = 4.2. (A) Le changement de régime de fusion d'une transition en une étape à une transition en deux étapes en faisant passer ϵ_{AB_CC} de -10 $k_B T_r$ à -1 $k_B T_r$ avec ϵ_{AB_AB} fixé à -10 $k_B T_r$,. (B) Courbe de fusion de chaque composante du réseau hétérogène avec ϵ_{AB_AB} et ϵ_{AB_CC} égaux à -10 $k_B T_r$ et -1 $k_B T_r$, respectivement.

Interactions entre les composants moléculaires du CCMV

En utilisant des simulations atomistiques MD, nous avons étudié les interactions entre un dimère et divers sous-assemblages. La structure d'un dimère était plus lâche en solution que dans une capside et la conformation de ses queues N-terminales était sensible à la force ionique. A faible salinité, les queues N-terminales présentaient une conformation étirée et légèrement liée sur le corps du dimère, recouvrant une partie des domaines hydrophobes. Par la suite, la liaison d'un autre dimère a nécessité un changement de conformation des queues de manière à

exposer les domaines hydrophobes. Ce changement pourrait être déclenché par la présence d'ARN sur lequel les queues se lieraient préférentiellement, permettant ainsi l'association de dimères. Les queues N-terminales ont été impliquées dans l'auto-assemblage de dimères à faible force ionique: d'une part, elles ont accéléré l'association en pontant les dimères; mais d'autre part, ils pourraient conduire à des pièges cinétiques et à des erreurs d'assemblage causées par une interaction forte avec le corps de l'autre dimère.

La Figure F8 est une comparaison de la force d'interaction d'un dimère avec différents composants. Nous avons observé que le pentamère et l'hexamère de dimères étaient des sous-ensembles stables avec un dimère nécessitant une forte force de rupture pour être extraits de ceux-ci. Ce résultat est en accord avec le fait qu'une capside icosaédrique est constituée de pentamères et d'hexamères de dimères et que l'intégrité de la capside est préservée grâce à la solidité de ces sous-assemblages. Il est intéressant de noter que l'interaction entre l'ARN et un dimère était modérée, ce qui concordait avec nos mesures expérimentales. Cela peut être nécessaire à la survie du virus car il doit libérer facilement son génome dans la cellule hôte. Cela peut également être utile pour la sélection du génome: avec une énergie de liaison faible, les dimères sont en mesure de dissocier rapidement l'ARN s'il n'appartient pas au génome viral et ont ainsi plus de chances de regrouper les segments appropriés pour assurer la prolifération du virus. Un dimère se lie également préférentiellement aux domaines double brin de l'ARN car la densité de charge est plus élevée.

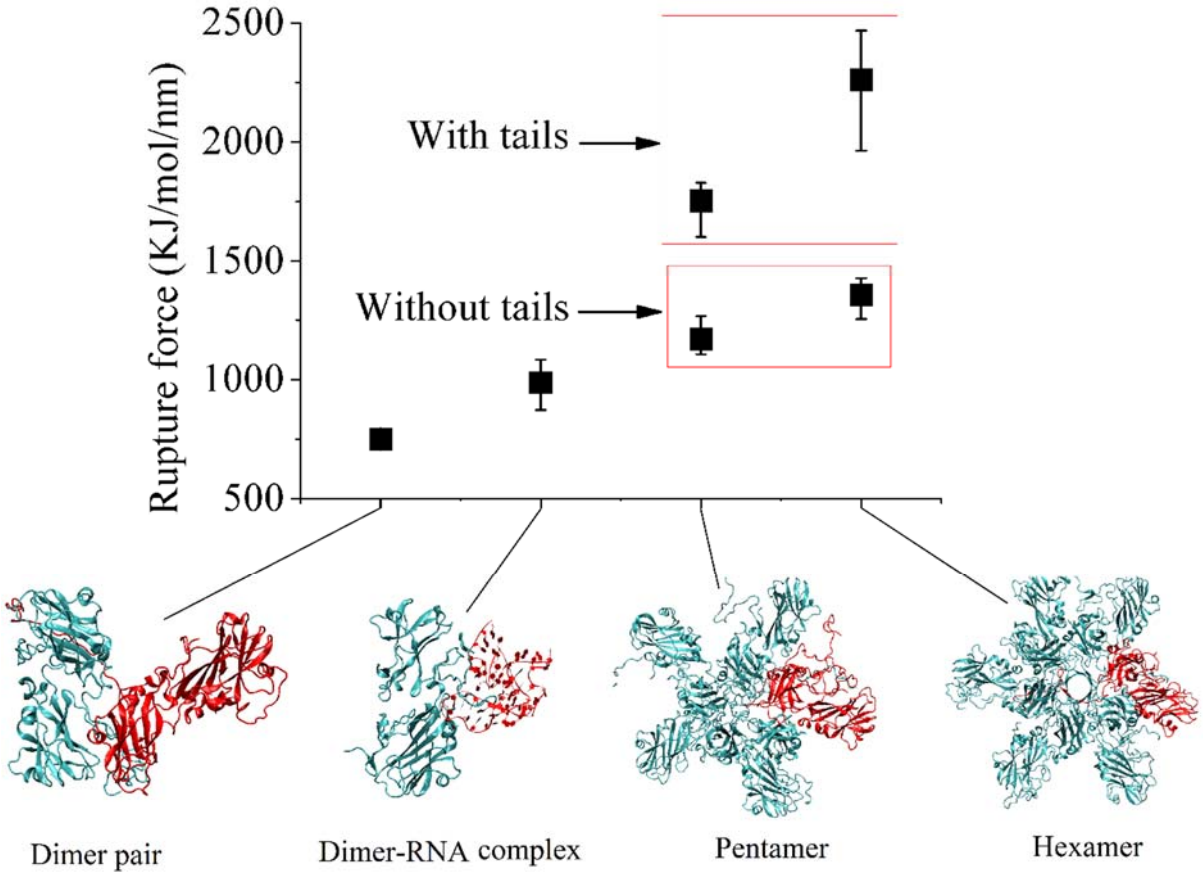

Figure F8. Force de rupture dans différents sous-ensembles. Les nombres sur l'axe des abscisses correspondent aux sous-ensembles: une paire de dimères de conformation native dans une capside, un complexe constitué d'un dimère et d'une pièce d'ARN flexible, un pentamère

de dimères et un hexamère de dimères. Le composant tiré est coloré en rouge.

Malgré la taille et la complexité du virus CCMV, nous avons pu éclaircir les interactions entre les dimères et les sous-ensembles au niveau atomistique. Les structures, les forces, les énergies et les échelles de temps deviennent de plus en plus importantes pour comprendre les phénomènes dynamiques à la base de l'autoassemblage du virus. Quelques-unes seulement de ces quantités sont accessibles par des expériences et les simulations atomistiques par ordinateur peuvent apporter des informations complémentaires réalistes. Nous pensons que les expériences, la théorie et les simulations atomistiques se nourriront et éclaireront les processus multi-échelles se produisant dans les systèmes biologiques.

Acknowledgments

I want to thank first of all my thesis director, Guillaume Tresset, who introduces me into the fantastic world of virus particles and supports me finish my PhD carrier especially at my latest period. He not only helped me to understand the world of viruses, but also gave me many suggestions on how to design my computer simulations to make them strongly linking to experimental researches. In addition, I also want to thank his great effort to find any available computer resources for me and help me to learn the experiment procedures and operations of the self-assembly of viruses.

Secondly, I must thank the co-director of my thesis, Yves Lansac. He provided me numerous support on my simulation skills. With his assistance, my skills become them more mature and more rigorous. Many thanks to Yves for helping me to obtain the right to use computer cluster at CCRS-Orléans.

The completion of this thesis greatly attributes to the continuous encouragement of my two directors and their deliberate corrections.

My thanks would also go to all of examination panel members: Martin Castelnovo from ENS Lyon, Catherine Etchebest from Université Paris-Diderot, René Messina from Université de Lorraine, and Yves Boulard from CEA Saclay. Thanks all of them for their time to review my manuscript and their scientific suggestions on this thesis.

I would like to thank the China Scholarship Council (CSC) for their financial support of my PhD study. I also thank the education office, embassy of the People's Republic of China in the French Republic for their assistance for me to have a better life in France.

I thanks all members of the group SOBIO: Amélie Leforestier, Jéril Degrouard, Maelenn Chevreuil, Laurent Marichal, Laetitia Poncet, and all non-permanent members that ever stayed in the group. It is a warm and friendly family. Thanks all of them for trying to help me to integrate into life and research atmosphere of France.

In addition, I would like to express my sincere gratitude to all members of LPS. Thank you for providing us a convenient and comfortable environment for our research life.

I would like to deliver my gratitude to my parents and my brother as well as his two lovely children. You are my joyful hours in the weekend.

Finally, I want to thank all those who ever helped me, inspired me, and corrected me.

Contents

Chapter 1

Introduction

1.1 What is a virus?

1.1.1 Overview

A virus is a tiny and simple biological capsule, which can only live in host cells and proliferate by self-replication. A virus is a protein and nucleic acids complex in a well-defined order, and was firstly discovered by Martinus Beijerinck in 1898 (tobacco mosaic virus, TMV), which triggered discoveries of virus species in the following century. To date, more than millions species of virus were discovered and over 5,000 of them were studied in details [1]. Even though some viruses are useful to human beings, such as bacteriophages, they bring larger damage by infecting all kind of life including animals, plants and even bacteria, and cause viral diseases, part of which cannot be cured appropriately so far, such as influenza virus, HIV and rabies.

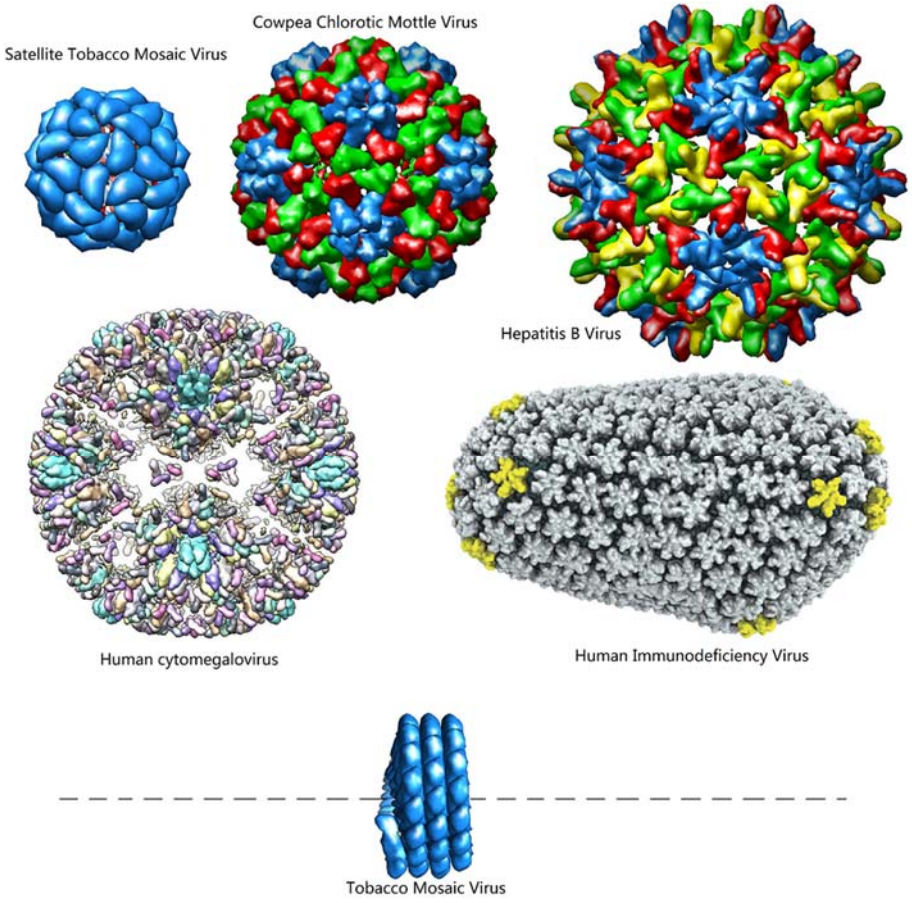

Figure 1-1. Structures of various viruses. Snapshots of viruses are taken from website of viperdb.

To some extent, virus can be viewed as a by-product of the metabolism of cells. Viruses outside a host cell exist as viral particles with a size ranging from 10 nm to 500 nm, and are known as virions. These viral particles are generally made of three parts: the viral genome in the form of DNA or RNA carrying genetic information; a single layered or multilayered protein shell encasing the viral genome; and in some cases an envelope of lipids that surrounds the capsid.

1.1.2 Structure

The simplest virions have just a nucleocapsid, an association of viral genome with a protective capsid of polypeptide, as shown in Figure 1-2(A). The capsid consists of hundreds or even thousands of identical proteins that were encoded by the viral genome. These identical proteins act as basic assembly subunits called capsomers and are organized in a regular order to form a capsid with or without the assistance of the viral genome. Yet what is more general is that viruses have a much more complicated structure, such as an additional lipid envelope derived from the host cell membrane with a number of functional proteins on the surface as revealed in Figure 1-2(B). The homology of the lipid layer with the membrane of the host cell grants viruses an unstoppable capacity to invade the healthy cells to proceed their replication. For some viruses, such as all herpesviruses [2], the gap between the lipid layer and the nucleocapsid is filled by a cluster of proteins that is known as viral tegument or more commonly as viral matrix and that will be released to inhibit the response of immune system and help the replication of viral genome after invasion.

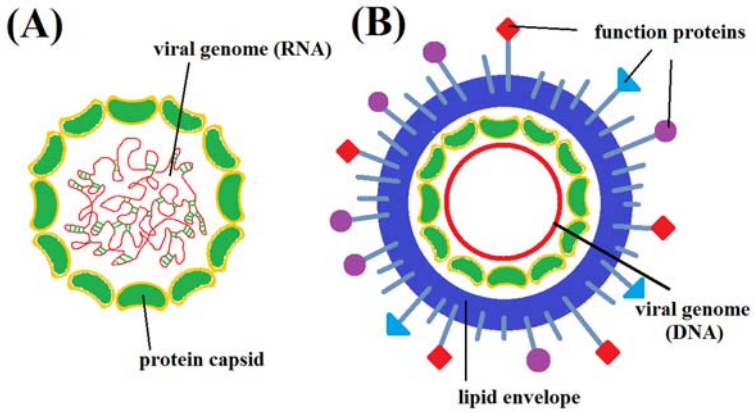

Figure 1-2. Schematic of the structure of viruses. (A) The simplest viruses consist of a virus capsid and the encased viral genome. (B) Advanced virus with an additional lipid envelope.

As shown in Figure 1-1, the virions have a wide difference in size and shape due to diverse species. The size of virions is normally smaller than 1 micrometer and thus cannot be observed by optical microscopy. Yet the size distribution is wide, ranging from 17 nm for Porcine circovirus, the smallest virus [3], to 440 nm for Megavirus chilensis [4], 1 μm in length and 0.5 μm in diameter for Pandoravirus [5] and 1.5 μm in length and 0.5 μm in diameter for Pithovirus [6], the largest virus discovered so far. The shape of viral particles normally means the shape of viral nucleocapsid. More than half of all known viruses have an icosahedral shape [7] and the rest can be helical, prolate and so on. Those viral capsids are generally made by hundreds

of one or several type of identical proteins, giving a biological advantage in the sense that less genetic information needs to be carried by the viral genome and thereby making the genome packaging easier. Given this fact, Crick and Watson [8] mathematically argued that viral capsids can only exist in two forms, helical rod-like structure and polyhedron structure, in order to leave enough room to package the viral genome.

Caspar and Klug [9] did a groundbreaking work on the description of the quaternary structural organization of the proteins in icosahedral capsids. They proposed a so-called quasi-equivalence principle where all capsid proteins are packed in slightly different environments on an icosahedron lattice. By unfolding an icosahedron on a flat surface along all five-fold symmetric units, one would find that an icosahedron is an assembly of 20 identical triangular facets as plotted in Figure 1-3(A). Conversely, an icosahedron can also be assembled by gluing 20 identical triangular facets with vertices located at the 12 five-fold symmetric units of an icosahedron, indicating that 12 pentamers are geometrically required for the formation of an icosahedron. Given that, a triangulation number of T, which denotes the number of monomers in each triangle, was introduced and the total number of capsid proteins can be simply described by $60T$ with $T = h^2+hk+k^2$ where h and k are integers, with $h \geqslant 1$ and $k \geqslant 0$. The icosahedral capsid with different T is depicted in Figure 1-3(B).

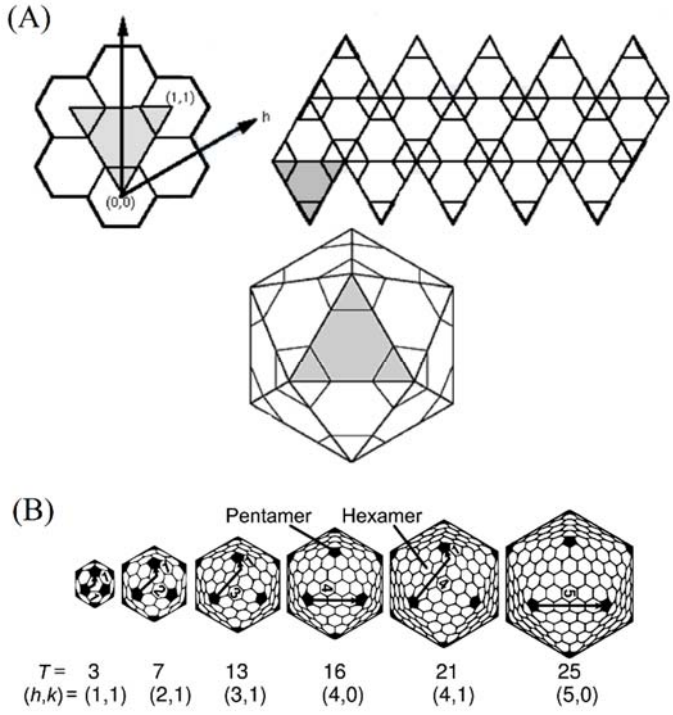

Figure 1-3. Schematics of a viral capsid. (A) Unfolded structure of an icosahedron on a flat surface. (B) Evolution of the viral capsid with triangulation numbers T from 3 to 25. Adapted from website of viperdb.

To explore the origin of icosahedral symmetry and the Caspar-Klug sequence of T numbers, a simple model [10] was devised: The viral capsid was geometrically divided into different domains represented by two different types of disklike capsomers (pentamers and hexamers) which interact with each other by a set of simple and well-defined force fields. It was found

that the icosahedral symmetry is not a generic consequence of free energy minimization, but an optimization required by the structural nature of capsomers [11]. Furthermore, as illustrated in Figure 1-4, dispersed energy minima (the number of subunits of 12, 24, 32, ⋯, see Figure 1-4) were found by minimizing the system energy with the number of capsomers composing a viral capsid through Monte Carlo simulation, consistent with the Caspar-Klug sequence and its thermodynamic origin.

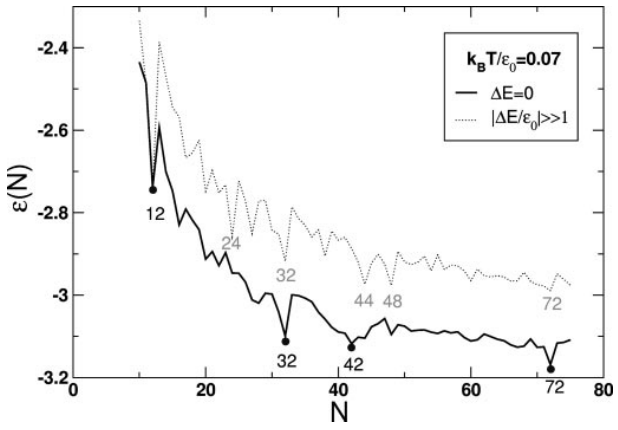

Figure 1-4. Energy per capsomer when $\Delta E = 0$ and $|\Delta E/\varepsilon_0|$ much larger than one with ΔE the energy difference between a pentamer and an hexamer capsomer, and ε_0 the capsomer-capsomer binding energy. Adapted from [10].

1.1.3 Applications

There is no doubt that viruses can cause a huge economic and health damage to the society. The history of human beings consciously fighting against viral diseases has lasted for hundreds of years, from Black Death in the early 15th century to AIDS and various influenza nowadays. The increasingly frequent communications of human beings and goods nowadays greatly rise the spread probability of viruses between countries. Means to develop an effective way to inhibit the spread of virus infection faster and cheaper is not only becoming important but also more and more challenging.

Vaccine, a term derived from the French name of smallpox of the cow (Variolae vaccinae) is the most widely used approach to prevent from being infected by viruses since its first application by Edward Jenner [12]. Therefore the better we understand viruses, the easier and faster we can develop appropriate vaccines. In addition, viruses can also act as an excellent system for the investigation of the functions of the cell and for various therapies [13,14]. For instance, viruses can act as a carrier to introduce some external genetic information into cancer cells where the replication of viruses takes place and ultimately results in the death of those cells by various mechanisms, such as necrosis [15] and mitotic catastrophe [16]. The production of pharmaceutical proteins can also rely on the assistance of viruses which will drive host cells to replicate the proteins by introducing specific genetic information [17].

Due to the good biocompatibility and cell penetration capability of viral capsids, viral particles have found a tremendous potential of application in nanotechnology [18-22]. Viral particle itself is an organic nanoparticle with many function proteins on the surface which grant

it a favorable ability to penetrate cell wall and enter the cell. Furthermore, for any specific species of viral particle, its shape and size as well as its local structure are constant. Those identical properties facilitate the manipulation of the load of drug inside the capsids and the fabrication of functional nanoparticles with specific external groups introduced by chemical modification. Moreover, viruses are able to evolve themselves to broaden the spectrum of viral particles as carriers for various materials.

Virus particles can also be employed as templates for the preparation of functional hybrid materials. For instance, cowpea mosaic virus (CPMV) particles were utilized to amplify the sensation signal of the DNA micro-assay where an engineered virus particle was used as a scaffold to attach fluorescent molecules so that fluorescence quenching can be efficiently obviated and the fluorescence tension was significantly improved [23]. In addition, CPMV particles can also be used for nanoscale molecular electric device. It was reported that virus particles can also be employed to prepare rechargeable batteries [24] or for piezoelectric applications [25].

1.2 How do virus particles form?

Viruses can infect all types of life forms, from animals and plants to bacteria, by replicating themselves inside the living cells. The infection process can be divided into several generic steps as shown in Figure 1-5, typically including attachment to the host cell, penetration into the host cell, uncoating of the virions inside the cell, replication of the genetic information and capsid proteins, assembly of capsid proteins and nucleic-acids chains into full capsids, and finally escape from the host cell.

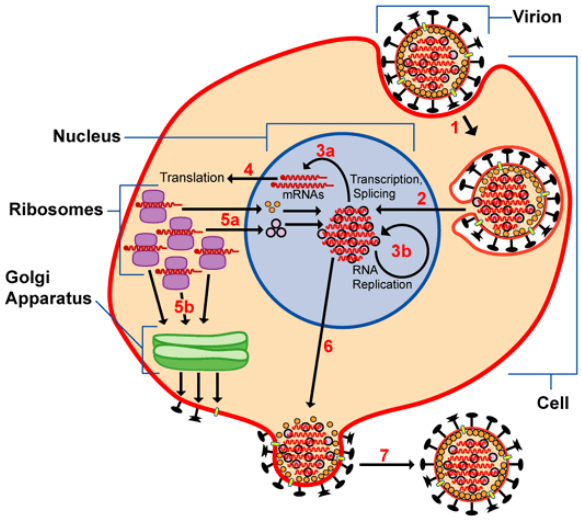

Figure 1-5. Life cycle of viruses from Wikipedia. The reproduction procedure goes as: (1) adsorption, (2) entry, (3) uncoating, (4) transcription/mRNA production, (5) synthesis of virus components, (6) virion assembly and (7) release.

The above procedure is the production process of viruses *in vivo*. Yet any reactions *in vivo* is difficult to probe experimentally and thus the assembly mechanism behind the formation of virus particle *in vivo* is elusive to date. Fortunately, numerous experiments confirm that most

of viruses can also be disassembled and reassembled *in vitro* under specific conditions which are different from those leading to the *in vivo* assembly [26]. In spite of the latent difference on kinetics and mechanism between the assembly *in vitro* and *in vivo* [27], it is a good start to understand the construction mechanism of these delicate structures.

1.2.1 Thermodynamics of self-assembly

Driving force for self-assembly

The formation of viral capsids is a process to assemble hundreds of dispersed and disordered proteins into a regular and compact structure. As indicated in the section 1.1.2, virus capsid itself is thermodynamically stable with the advantage of their geometric architecture. What accompanies the assembly is the loss of translational and rotational freedom of the proteins, which results in the increase of free energy and the assembly reaction becomes unfavorable. To keep the stability of viral capsids, an overall energy optimization should be a prerequisite for the assembly reaction, as indicated by Figure 1-6. Therefore, the association of the proteins should release enough energy to compensate the increasing energy caused by the loss of translational and rotational freedom, contributing as driving force for the assembly.

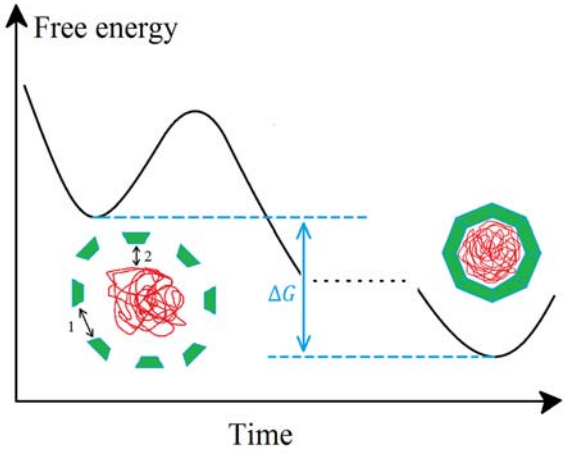

Figure 1-6. Evolution of free energy to assemble the dispersed capsid protein subunits and viral genome into a complete virion through the hydrophobic and electrostatic interactions between subunits (1) and electrostatic interaction between subunit and viral genome (2).

For most non-enveloped spherical viruses, their proteins share a very exquisite structure. Hydrophobic residues generally fold into a typical eight-strand antiparallel β-barrel structure [28] and are buried by the polar and charged residues which in some case form electrostatic tails interacting with their genetic materials. In general, the proteins tend to associate into various assembly subunits [26], such as dimer of proteins for both cowpea chlorotic mottle virus (CCMV) and Hepatitis B virus (HBV). The interaction between these subunits can be classified into two sections: a long-range electrostatic repulsion and a short-range hydrophobic interaction. In most cases, the driving force for the assembly of viral capsid is ascribed to the hydrophobic interactions [29] which is consistent with the observation that the assembly is driven by the increase of the entropy caused by the release of solvent molecules from the hydrophobic domains [30]. There have been a number of experiments reporting that association energies of

subunits are intermediate, ranging from -5 k_BT to -10 k_BT [26,31-33]. This intermediate association energy is an imperative condition for the formation of the capsid with regular structure [26,32]. If the energy is too high, the subunits will be misassembled and the assembly will fall into kinetics traps, while if the energy is too low, the assembly will not happen and the subunits will remain in a dispersed state.

Equilibrium model for virus self-assembly

To thermodynamically describe the self-assembly of virus capsids, Zlotnick proposed a model based on a T =1 capsid in 1994 [34]. Considering the complexity of the original model, a simplified version by Hagan [35] is presented here. For empty capsids assembled from N identical protein subunits, the model simply assumes that for each specific number of subunits there is just one dominant intermediate formed. The total free energy for a system of subunits, intermediates, and capsids in solution can be described by

$$F_{EC} = \sum_{n=1}^{N} k_B T \rho_n [\log(\rho_n v_0) - 1] + \rho_n G_n \quad (1)$$

where k_B is the Boltzmann constant, T the temperature, v_0 a standard state volume, ρ_n the density of subassemblies with n subunits, and G_n denotes the free energy of such a subassembly (for more details about G_n, see Ref [35]). The density of the intermediate with n subunits at equilibrium can be evaluated by minimizing F_{EC} with a constraint that the total number of subunits is conserved during the reaction and writes as:

$$\rho_n v_0 = (v_0 \rho_1)^n e^{-\beta G_n} \quad (2)$$

where $\beta = 1/k_B T$.

Given the experimental observations that the fraction of intermediates is so small that it can be reasonably neglected [33], the system can be simplified into a two-component system, containing subunits and full capsids in the solution. The fraction of subunits in capsids (defined as $f_c = N\rho_N/\rho_T$) can be reduced to:

$$\frac{f_c}{1-f_c} = (v_0 \rho_T)^{N-1} e^{-\beta G_N} \quad (3)$$

where ρ_T is the total density of subunits, with a constraint that the total number of subunits is conserved, that is $\rho_T = \rho_1 + N\rho_N$. For most of viruses, $N \gg 1$, then equation (2) gives,

$$\frac{f_c^{1/N}}{1-f_c} = \frac{\rho_T}{\rho^*}$$

$$\rho^* v_0 = \exp(\beta \frac{G_N}{N-1}) N^{-1/(N-1)} \approx \exp(\beta G_N/N) \quad (4)$$

with ρ^* the pseudocritical subunit concentration to trigger the assembly. The solution of equation (4) is depicted as Figure 1-7.

It is indicated on Figure 1-7 that the higher the initial concentration of subunits, the more subunits will assemble into capsids. However, this is inconsistent with the experimental observation. It is because the assembly will be interrupted by kinetic trap at high subunit concentration [36], when new nuclei form faster than the timescale of a complete assembly, and free subunits are depleted before most capsids finish to assemble.

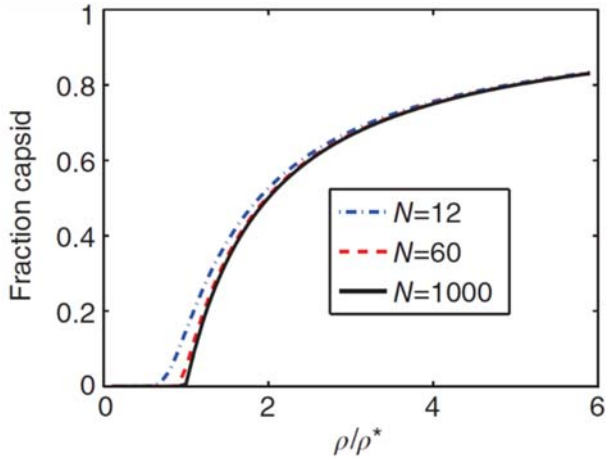

Figure 1-7. Fraction of capsids f_c as a function of subunit oversaturation ρ/ρ^* predicted by equation (4) for the number of subunits in a complete capsid N = 12, 60, and 1000. Adapted from [35].

Self-assembly mechanisms

To date, the assembly mechanism of viral capsids is still controversial. There is no generic mechanism capable of comprehensively explaining the assembly for various viruses, indicating the inherent complexity inside the assembly reaction. Therefore, to have a thorough understanding of the assembly of virus, we should begin with simple viruses.

Nucleation-elongation mechanism

Hepatitis B virus (HBV) can be a simple and well-documented case to cut in. As indicated by Figure 1-8(A) and (B), the assembly kinetics of HBV capsid protein shares a highly similar sigmoidal shape pattern with the polymerization of an infinite polymer, such as fibers or helices in one dimension, sheets in two dimensions and crystals in three dimensions, which suggests the applicability of the classical nucleation model to the assembly of virus capsid. There have been indeed many observations consistent with the assembly of virus capsid by the classical nucleation mechanism, such as HBV [37], the minute virus of mice (MVM) [38], CCMV [39] and Alphavirus [31].

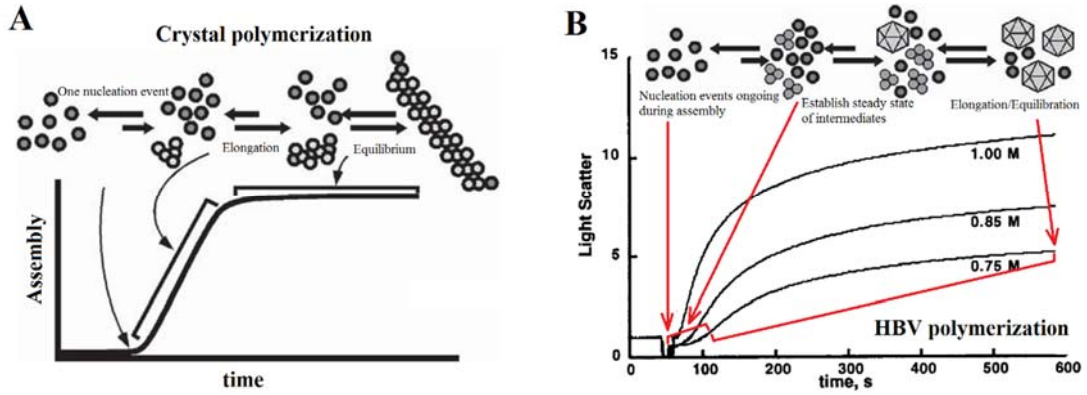

Figure 1-8. The assembly kinetics of crystals and HBV capsids. (A) Classical nucleation-growth model where a single infinite polymer can be generated by a single nucleus. (B) The

assembly kinetics of HBV capsid proteins at different protein concentrations where each discrete capsid must arise from an individual nucleus. Adapted from [40].

For the polymerization of virus capsids, the contacts between subunits in the initial stage of reaction are so few that the gain in energy is not sufficient to compensate for the mixing and rotational entropy penalties incurred upon association. The initially formed structures are not thermodynamically stable and undergo a repeating association-dissociation process until the formation of a nucleus which possesses a 50% probability to grow into a complete capsid [35]. The probability for the formation of stable nuclei can be improved by increasing the subunit concentration. Interestingly, the nucleus varies with viruses, a trimer of dimers for HBV [37] and MVM [38] while a pentamer of dimers for CCMV [39] and Alphavirus [31]. The nucleation stage is followed by an elongation stage where the subunits will subsequently add to the stable nucleus and form a complete capsid. The assembly process of HBV is depicted in Figure 1-9.

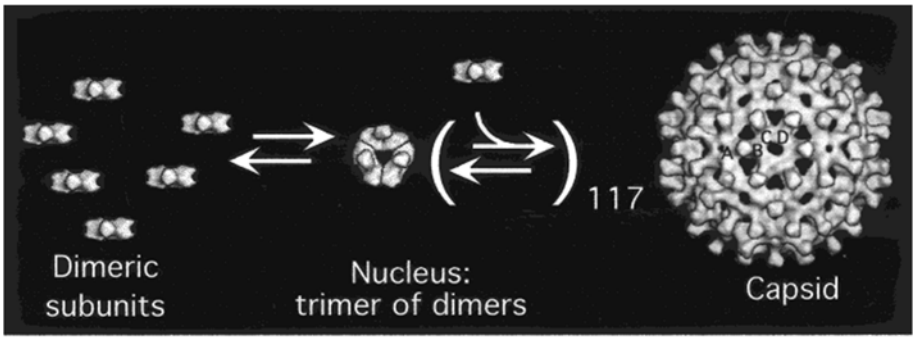

Figure 1-9. Assembly mechanism of HBV capsid. The assembly begins with the formation of a trimeric nucleus and is followed by a sequential addition of dimeric subunits into a complete *T*=4 capsid in the elongation step. Adapted from [37].

In spite of the pattern similarity in assembly kinetics, there is an inherent difference between the polymerization and the virus assembly. The polymerization of polymers is infinite while that of capsids is finite. In polymerization reaction, monomers can be infinitely polymerized from a single nucleus into an infinitely long polymer. The lag phase corresponds to the time required to generate the stable nucleus and a steady equilibrium plateau forms after depletion of subunits and the free subunit concentration reaches the critical polymerization concentration, as displayed by Figure 1-8(A). In contrast, virus capsid is a closed structure and is made of a finite number of subunits. The completion of a capsid indicates the termination of polymerization. Thus, nucleation occurs continuously and a number of polymerizations take place concurrently but not synchronously throughout the assembly reaction. In addition, unlike the constant number of polymer end points available for free monomer binding in a classical polymerization, the average number of binding sites varies with the progress of assembly, leading to an apparent cooperativity of the assembly. Those differences are clearly reflected in Figure 1-8(B).

For a single polymerization of a capsid, the reaction can be simplified as:

$$\text{subunits} \Leftrightarrow \text{nuclei} + \text{subunits} \Leftrightarrow \text{capsids}.$$

In the lag phase, a population of intermediates is generated. Given the concurrent but not synchronous features of the capsid assembly, the rate of accumulation of an intermediate of a

specific size can be expressed as:

$$\frac{\mathrm{d[nuc}+n]}{\mathrm{dt}} = k_{elong,n-1}[\mathrm{nuc}+n-1][\mathrm{subunit}] + k_{dissoc,n+1}[\mathrm{nuc}+n+1] - (k_{elong,n}[\mathrm{subunit}] + k_{dissoc,n})[\mathrm{nuc}+n].$$

As the reaction proceeds, those concurrent and non-synchronous reactions reach a delicate equilibrium and a dynamically steady population of intermediates is formed. The final concentration of capsids can be described in terms of a global association constant expressed as:

$$K_{\mathrm{capsid}} = [\mathrm{capsid}]/[\mathrm{subunit}]^N. \quad (5)$$

Since K_{capsid} is too complex to be calculated, one expects to derive other quantities that are more convenient. The average association energy between two subunits is one of them. Assuming that all contacts in the virus capsid are identical and each subunit has *c* contact surfaces, then K_{capsid} can be expressed in terms of a statistical factor accounting for reaction degeneracy and some power of the single contact equilibrium constant, K_{contact}. $cN/2$ contacts are required to form a capsid. The statistical factor accounting for degeneracy degree of *j* with *N* subunits is of a general form j^{N-1}/N. Thus, K_{capsid} can be statistically expressed by K_{contact} as followed:

$$K_{\mathrm{capsid}} = (j^{N-1}/N)(K_{\mathrm{contact}})^{cN/2}. \quad (6)$$

Furthermore, K_{contact} can also be related to the association energy per contact $\Delta G_{\mathrm{contact}}$ as:

$$\Delta G_{\mathrm{contact}} = -RT\ln(\Delta K_{\mathrm{contact}}). \quad (7)$$

With this, one can dissect the driving force of virus capsid assembly by van't Hoff analysis.

Cooperative mechanism

Besides the classical nucleation-growth mechanism, there are also some viruses demonstrating different assembly mechanisms. Cooperative mechanism, or *en masse* mechanism, was frequently observed in virus assembly, in particular with the presence of loaded cargo. For example, norovirus was found to assemble via a two-step process [41] as shown in Figure 1-10: a fast nucleation followed by a slow cooperative association of intermediates to form an ordered capsid.

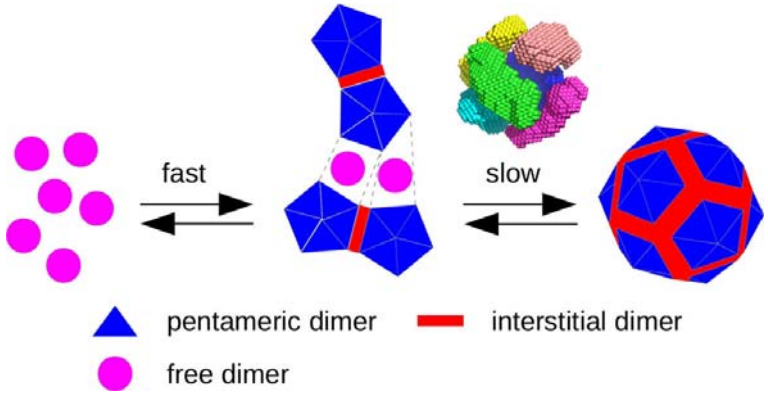

Figure 1-10. A two-step assembly mechanism of norovirus capsid. In the first step, free dimers are rapidly associated into intermediates which subsequently are glued together by the interstitial dimers and relax into capsids. Adapted from [41].

Assembly with scaffolding proteins

However, the factors that impact the assembly and stability of virus are getting intricate with the increased complexity of the viral particles. As indicated by Li et al. [42], scaffolding proteins will effectively lower energy barrier and increase radius of curvature by scaffolding proteins, and hence are necessary for the formation of large icosahedral viruses. P22 phage assembly [43-48] is a well-documented example demonstrating the important function of the scaffolding proteins to guide the assembly to a lower energy landscape pathway and to ultimately guarantee a homogenous population of completed virus particles.

In addition, there are several kinds of scaffolding proteins for some viruses [49,50]. Bacteriophage ΦX174 is a simple *T*=1 virus [51], but it only assembles by the assistance of two different scaffolding proteins via the procedure shown on Figure 1-11. The assembly proceeds via a pentameric intermediate [52], followed by an 9S intermediate (see Figure 1-11) stabilized by inner scaffolding proteins into a protein complex 12S which subsequently grows into a procapsid by being glued by external scaffolding protein D [53]. Considering the fundamental importance of the simple assembly system and the complication in the assembly of advanced viruses, we just focus on the basic assembly mechanism hereafter.

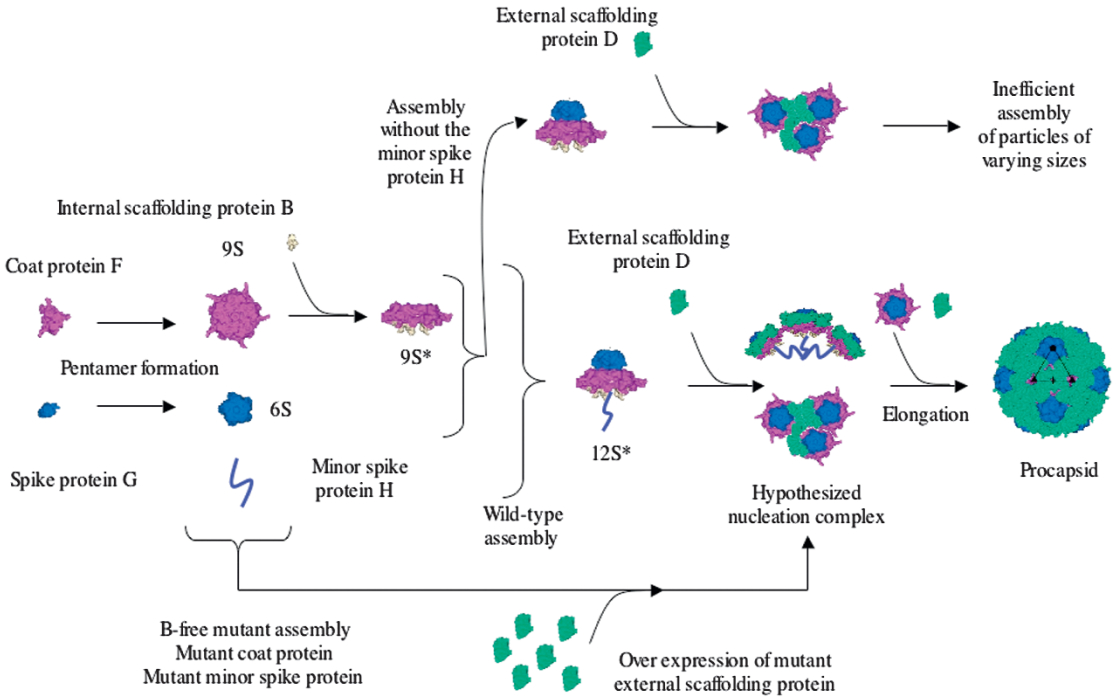

Figure 1-11. Schematic of the assembly pathway of bacteriophage ΦX174. Adapted from [53]

1.2.2 Experimental methods

Hitherto, many experimental methodologies have been employed to investigate the various properties of viral particles. With the introduction of new techniques, more and more information on viruses is revealed, motivating the development of theoretical methods.

Characterizing the structure and properties of viruses

To have an intuitive knowledge on the virus structure, we can probe the size and shape of viruses by Electron Microscopy (EM) [31,54,55] and their atomic information can also be revealed by solving the X-ray crystallography of virus crystals [56-58]. With this technique, the atomic structures of a number of viruses were disclosed no matter the size, from viruses as small as CCMV with a diameter less than 30 nm [57] to large viruses as big as Human cytomegalovirus with a diameter of over 130 nm [58]. Many researches on other aspects of viruses will be accelerated with the understanding of the atomic structures. For example, by atomic force microscopy (AFM), we are able to evaluate the mechanical strength of the virus particles [59] and the stability of assembly intermediates [60].

Insight in the assembly mechanism of virus capsid

What is more attractive to researchers, especially physicists, is the mechanism behind these well-defined virus capsids. Since many viruses are able to reassemble *in vitro*, which makes many complementary experimental techniques applicable and useful, the gate to the understanding of the assembly mechanism is gradually opening.

Light scattering and exclusion chromatography

In the early stage, the assembly kinetics of diverse viruses were analyzed by combining light scattering (LS) and exclusion chromatography (SEC). The fraction of assembly components can be monitored using SEC, while the mass-averaged molecular weight can be estimated with LS. The basic assembly subunits for each species of virus have identical size and molecular weight. Supposing that the fraction of assembly intermediates is small, then the consumption of subunits and the production of capsid can be tracked by light scattering and SEC, as shown in Figure 1-12(A). Zlotnick's group carried out a systematic work on spherical viruses, including HBV [32] and CCMV [39]. They found that the assembly of empty capsids followed the nucleation-growth mechanism, as was proposed by Prevelige et al. [43]. For CCMV the subunits formed a stable nucleus (pentamer of dimers) subsequently proceeding with a sequential addition of subunits. By fitting data (Figure 1-12(B)) extracted by combining LS and SEC with the thermodynamic model mentioned in section 1.2.1.3.1 (equations (5) to (7)), the apparent association energy between the self-assembling subunits $\Delta G_{\mathrm{contact}}$ can be evaluated, as displayed in Figure 1-12(C). It was indicated that weak interactions between subunits are nevertheless strong enough for the formation of a globally stable capsid [26,32]. The low association energy enables an efficient self-assembly by avoiding kinetic traps, which facilitates the formation of regular structures.

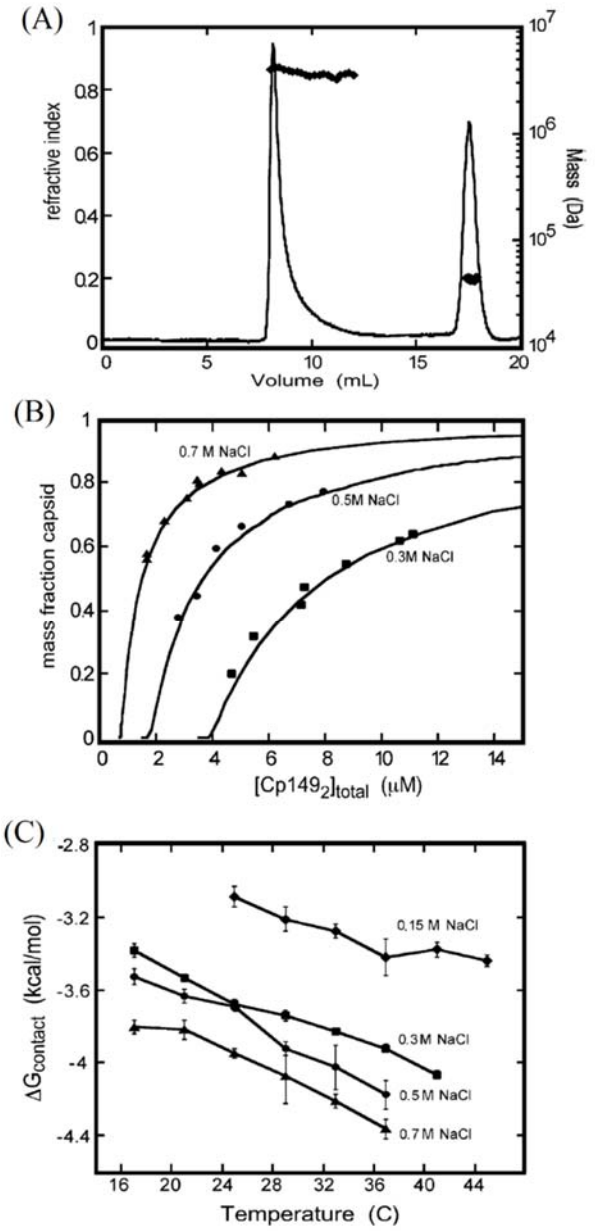

Figure 1-12. (A) Chromatography of HBV capsid protein assembly. Two major assembly components were 4 MDa capsid and 35 kDa dimer of proteins. (B) Fraction of capsids measured for assembly of empty HBV capsids from capsid proteins using SEC as a function of total dimer subunit concentration at different ionic strengths. The lines are fitted with equation (5). (C) Estimated values of $\Delta G_{\text{contact}}$ as a function of temperature and ionic strength. Adapted from [32].

Mass spectrometry

Mass spectrometry (MS) was also used to study the assembly of viruses [61]. Comparing with other techniques, MS demonstrates a great accuracy in distinguishing the intermediates formed during the assembly process, which gives deep insights into the assembly pathways [62,63]. Figure 1-13 is the mass spectrograph of the assembly of HBV where a number of intermediates generated in the assembly process can be clearly probed, facilitating the study of assembly

reaction in detail. Lutomski et al. carried out a detailed investigation on the assembly of HBV capsids by charge detection mass spectrometry (CDMS). They found that there are multiple pathways for the assembly depending on the association energy between subunits which varies with the assembly condition [36]. In addition, in the study of the final step of assembly, it was found that the subunits formed many virus particles that are defective and overgrown in the initial stage of the assembly, and then those particles going through a slow correction process to eventually form complete capsids [64]. Uetrecht et al. found that the full HBV capsid will exchange subunits with the soluble pool by native tandem MS [65].

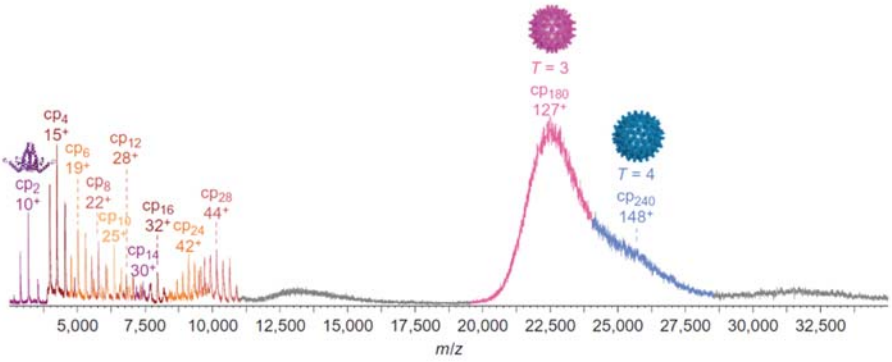

Figure 1-13. Mass spectrometry of the assembly intermediates of HBV. Adapted from [66].

Small-angle scattering

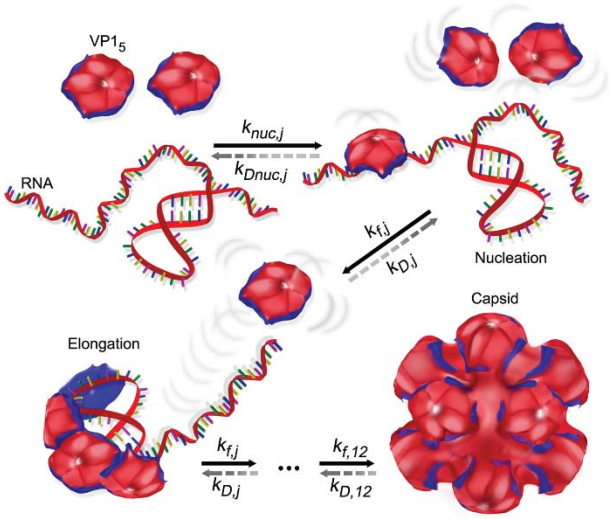

Figure 1-14. The assembly pathway of SV40 detected by TR-SAXS. Adapted from [69].

Given MS is able to provide precise assembly information, the vacuum environment in MS might lead to variations of assembled structures, which might cause artefacts in the final results. Yet the solvent accompanying assembly subunits hinders measurements in many techniques and thus becomes a major obstacle to obtain detailed information on the assembly. Recently, time-resolved small-angle scattering (TR-SAS) increasingly demonstrates its advantage on virus self-assembly. TR-SAS of X-rays and neutrons is able to probe the structure or distribution of nanoparticles in bulk solution with a relatively high resolution (~ 1 nm) regardless whether they possess a regular crystal-like structure or not. With TR-SAS the potential intermediates formed during the assembly of virus capsids can be captured and their size and three-

dimensional shape can be reconstructed by *ab initio* calculations based on scattering data [41]. By TR-SAXS, Tresset et al. revealed the self-assembly mechanism of norovirus capsids [41] and CCMV [67], meanwhile, the disassembly pathway of CCMV was also disclosed [68]. Similarly, a cooperative assembly mechanism was also observed for the assembly of SV40 by TR-SAXS [69], as illustrated in Figure 1-14.

Other techniques

Except from the methods above, many other conventional or newly invented techniques are introduced in the study of virus assembly. Recently, Medrano et al. [38] tentatively examined the self-assembly pathway of the minute virus of mice (MVM) by combining Transmission Electronic Microscopy (TEM) and high resolution AFM. A succession of transient intermediates formed at different assembly stage was captured and depicted in Figure 1-15(A) (TEM) and (B) (AFM). It is indicated that the assembly of MVM capsid obeys a nucleation-enlongation mechanism with trimer of proteins as the basic assembly subunit, preceding the formation of pentamer of trimers followed by sequentially additions of trimers to the pentamers leading finally to a complete MVM capsid.

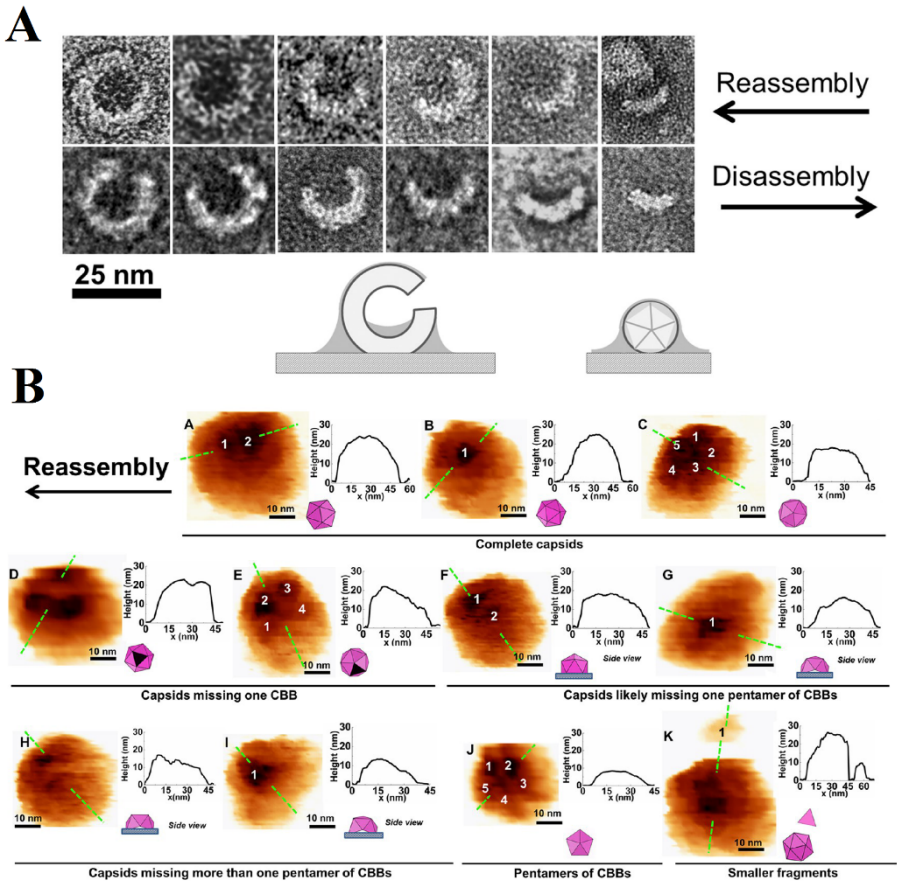

Figure 1-15. (A) The assembly and disassembly conformations detected by TEM at different reaction stages. (B) Reassembly process probed by AFM at different reaction stages. The inserted curves are the height profiles of the cross-sections marked by the green dashed lines and the purple polyhedrons are the estimated conformations of the intermediates. Adapted from [38].

In addition, Rogen et al. [70] employed NMR to monitor the disassembly of virus-like particles, while Harms et al. [71] designed a glass nanofluidic device to detect the assembly of HBV capsid. In order to trace the assembly of a single virion, Garmann et al. [72] developed a DNA-coated surface where RNA genome is able to bind to and detect the assembly reaction by TEM, providing a straightforward route for monofunctionalizing virus-like particles (VLPs). Zhou et al. [73] employed a resistive-pulse sensor with pores of 40 nm in diameter capable to detect the size of virus particles. Site-directed spin labeling (SDSL) electron paramagnetic resonance (EPR) found application in the structural transition of virus capsids as well [74].

1.2.3 Theoretical methods

With the tremendous amount of information produced by various experimental techniques, many theoretical works were performed to explain and refine the findings on the assembly of viruses [35]. In order to answer various questions on the properties of virus, some old models applied in other assembly systems have been tested in viruses, typically the nucleation-growth theory [43], and meanwhile a wide variety of new models or theories have been developed. Overall, those methods can be classified into analytical and numerical methods.

Analytical methods

To understand the secret behind the Caspar-Klug principle for icosahedral capsids as mentioned in section 1.1.2 and the reason for the shapes that viruses adopt, many models have been built [10,11,75,76]. These models usually begin with basic assembly units, generally the stable intermediates, which may subsequently assemble into a complete capsid, such as pentamer or hexamer of dimers for CCMV and trimer for HBV. The possible stable virus structures are well summarized and predicted [10].

On the other hand, to describe the assembly of virus capsid, Zlotnick et al. have developed two assembly kinetic models, both of them based on rate equations and nucleation-elongation model. The first one is an equilibrium model [34], whose simplified version can be found in section 1.2.1. The other is a kinetically limiting (KL) model [37] as described in section 1.2.1.3.1. Given how simple these two models are, both models are capable to reproduce the typical sigmoidal kinetics of capsid assembly. KL model have shown good agreement with many features of experimental assembly kinetics data, such as HBV [37], CCMV [33], BMV [77], SV40 [78], the impact of RNA on MS2 assembly [79] as well as the assembly of HIV capsid protein into tubes [80].

Coarse-grained modeling

Thanks to the experimental information reducing or narrowing the varieties in the virus assembly significantly, the difficulties to construct a coarse-grained (CG) model to reveal the molecular or even atomic level interaction on the assembly has been reduced a lot.

There are some highly CG models that build virus capsids with various ‘bricks’ with well-designed structure and interaction potentials between them. These CG models are generally applied with dynamic simulations and thus are able to disclose the interaction details of subunits during assembly. Hagan et al. conducted a solid work on this aspect and obtained a number of

interesting findings on the mechanism of virus assembly. For instance, Hagan et al. [81] found that the assembly mechanism depends on the strength ratio between subunit-subunit interaction and subunit-genome interaction. When subunit-genome interaction is much stronger than subunit-subunit interaction, the capsids will assemble with a cooperative mechanism, otherwise the assembly will be carried out with a nucleation-growth mechanism. The details can be found in Figure 1-16. In addition, this CG model was also applied to investigate the size control mechanism of viral capsids and its assembly polymorphism [82].

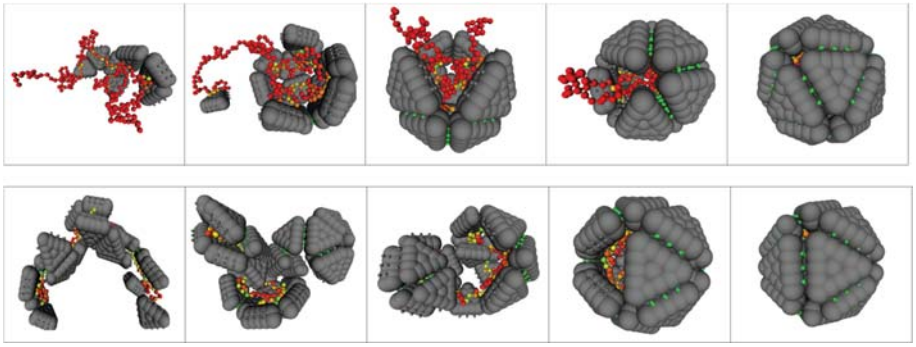

Figure 1-16. Two mechanisms for assembly around a polymer. (Top row) A nucleation-growth mechanism when subunit-subunit interactions are strong while subunit-polymer interactions are relatively weak. (Bottom row) A disordered assembly mechanism or a cooperative mechanism when subunit-subunit interactions are weaker and subunit-polymer interactions are stronger. In this mechanism, the assembly begins with the formation of a disordered complex composed by multi-subunits and polymer at the first stage, followed by annealing of multiple intermediates and finally completion. Adapted from [81].

Another similar model performed with Langevin dynamics simulation was proposed by Mahalik to probe how polymer would assist in the assembly of VLPs [83]. Nguyen et al. [84] developed a model to capture the assembly patterns of icosahedral capsids. It was found that the formation of full capsids is actually competing with the formation of misassembled structures, the protein concentration and temperature playing a determinant role in this competition.

1.3 Cowpea chlorotic mottle virus

Cowpea chlorotic mottle virus (CCMV) is a *T*=3 icosahedral virus consisting of a single-layered protein shell coating its RNA genome inside. CCMV is found in the cowpea plant, or black-eye pea, whose leaves develop yellow spots if infected, as shown in Figure 1-17, and was first experimentally isolated and characterized by Bancroft et al in 1967. Thanks to its simple but exquisite structure as well as series of excellent physical properties, CCMV has found a great potential of applications in nanotechnology, including nanocages [85-88], 3D multifunctional crystals [89], and drug targeting and delivery [19]. Compared with animal viruses, CCMV is structurally simpler and safer to handle, demonstrating a good potential as an initial template for the study of the assembly of virus.

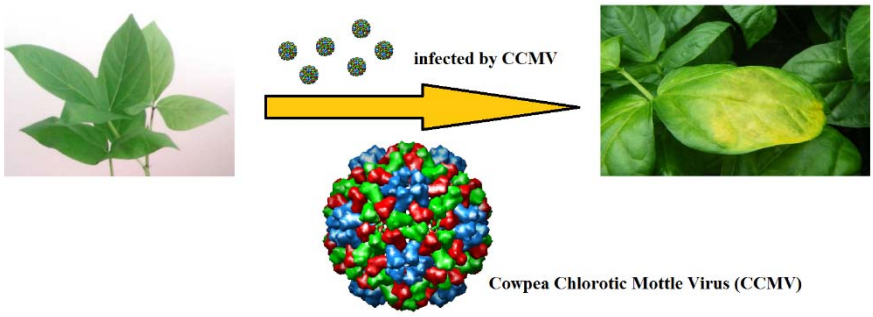

Figure 1-17. Leaves of black-eyed pea before and after infection by CCMV.

1.3.1 Structure and properties

CCMV is a *T*=3 virus and has a diameter of 28 nm. CCMV has a typical structure as the majority of viruses, a protein capsid encapsulating genome chains (RNA chains for CCMV). The atomic structure of CCMV capsid was first determined in 1995 by X-ray crystallography and cryo-electron microscopy [57]. As experimentally revealed, the capsid of CCMV has a shape of icosahedron with 180 identical protein molecules uniformly distributed on the surface. According to their specific environmental difference (located in pentamers or hexamers), these proteins are classified into three kinds, A, B and C. The capsid can be divided into different sub-domains, 12 pentagons and 20 hexagons, and exhibits a high degree of symmetry as shown in Figure 1-18. This specific construction has attracted many researchers' attention and has given inspiration on the understanding of its assembly pathway.

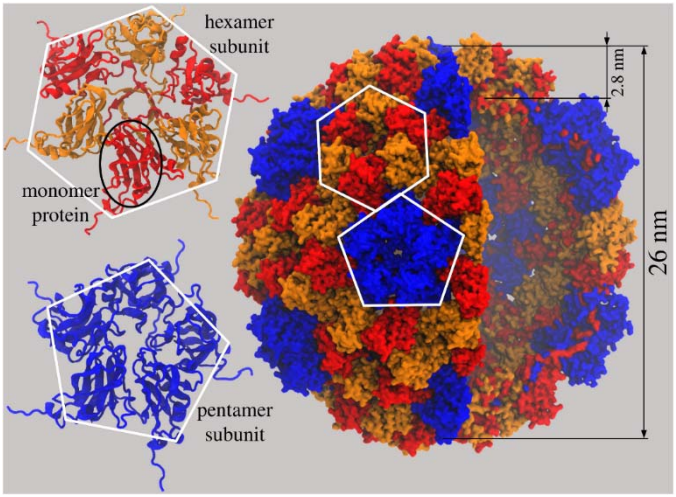

Figure 1-18. The structure of CCMV. The proteins A, B and C are in blue, red and orange, respectively. The hexamers and pentamers of dimer are marked and displayed on the left. Adapted from [90].

The basic assembly subunit for CCMV is a dimer of capsid proteins (see Figure 1-19(A)). Each dimer contains two N-terminal tails comprising basic residues and exhibiting a positive charge over a large range of pH, and a core with various domains (charged and hydrophobic) on the surface. If we classify the residues into basic, acidic, polar and nonpolar purely according to the residue type, the charge distribution of the dimer is very interesting as indicated by Figure

1-19(B)-(D): positive charges mainly concentrate on the inner side of dimer while the outer side displays an overall negative charge. Meanwhile, the hydrophobic domains are distributed on the two sides of dimer and act as the binding domains for the assembly of dimers. This charge distribution grants the capability for the dimer to assemble into a full CCMV capsid and even a multi-layered capsid as some experiments revealed [76].

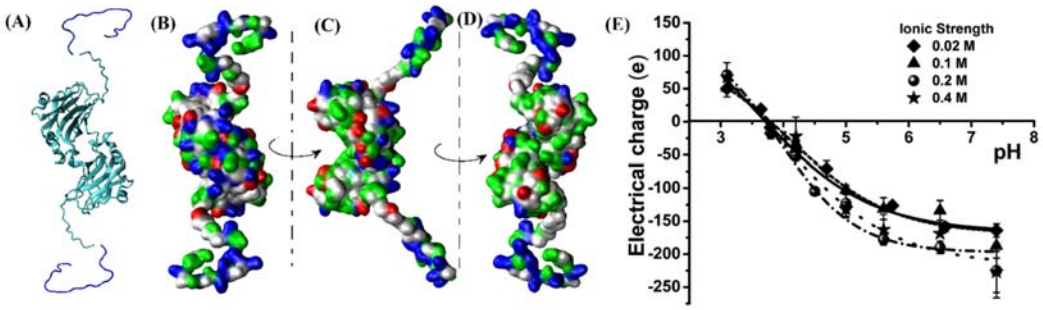

Figure 1-19. (A) The conformation of a CCMV capsid dimer. The positively charged N-terminal tails are in blue. (B), (C) and (D) are the inner, lateral and outer side of the residue distribution on the surface of dimer, respectively. The residues are colored according to their types, basic (blue), acidic (red), polar (green) and nonpolar (white). (C) The variation of the overall net charge of CCMV capsid with pH estimated by electrophoresis experiments. Adapted from [91].

The precise charge information of a dimer is fundamental for understanding the interaction between them. The isoelectric point of CCMV protein is ~4.8 and the charge of the hydrophobic core will strongly vary with pH. Yet the real overall charge of a dimer at specific pH is still ill-known [92,93]. Given that we have the atomic structure for a whole polypeptide chain, the charged condition for a dimer can be estimated theoretically via ionization equation by considering the environment effect or not. But the estimated charge differs from that measured by experiments, in particular for proteins with a condensed structure, which makes the ionization of groups more complex. To date, there has been electrophoresis experiments reporting that the overall charge for a whole CCMV virion varies from +60*e* to -160*e* when increasing pH from 3 to 7.5 [91], as shown in Figure 1-19(E). Thus, it is easy to estimate the net charges for a single dimer by simply dividing the overall charge by 90 and giving a value ranging from +0.6 *e* to -1.8 *e* at different pHs. Yet this is questionable because the presence of RNAs inside the capsid will neutralize the inner charges [94]. Furthermore, it is difficult to measure the net charge of a single dimer due to its irregular shape and unknown conformation in solutions.

The genome of CCMV is single-stranded RNA. Interestingly, there are 4 kinds of RNA chains with different lengths for CCMV, noted as RNA1 [95], RNA2, RNA3 [96] and RNA4 [97] with a number of nucleotides (nt) of 3171, 2774, 2173 and 824, respectively. The first two genomic RNAs are packaged alone, and the third genomic RNA is copackaged with a subgenomic RNA4 of 824 nt, so that each CCMV virion contains about 3000 nt [98]. The secondary structure determination of RNA is still a tough challenge and almost no information is available so far. Gopal et al. [99] attempted to visualize the secondary structure of CCMV capsid by various techniques as demonstrated in Figure 1-20, and a stretched conformation with a size slightly larger than a capsid was observed. Due to its high negative charge density, RNA

demonstrates a strong affinity for being bound to the N-terminal tails of dimers, effectively lowering the energy landscape of the capsid assembly [98].

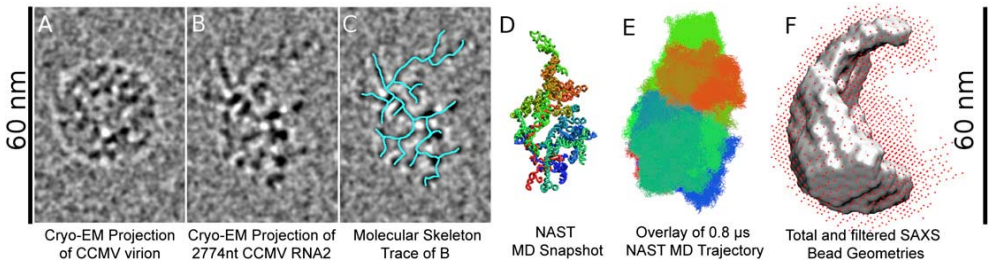

Figure 1-20. Secondary structure of CCMV RNA2 solved by various techniques: Cryo-electron microscopy ((A) to (C)), molecular dynamics simulation after energy minimization (D), over 0.8 µs (E), and reconstructed by SAXS (F). Adapted from [99].

1.3.2 Self-assembly of empty capsids

CCMV is one of the most popular viruses employed as a model system to investigate the assembly mechanisms of virus due to its numerous properties. As mentioned in the last section, the interaction between CCMV capsid dimers is weak while the interaction between dimer and RNA is mild, which makes CCMV easy to disassemble and reassemble *in vitro*. The presence of viral genome is not necessary for the formation of a full CCMV capsid. The CCMV capsid proteins themselves can self-assemble into an empty CCMV capsid under certain conditions [32,33,39], which grants it a number of potential applications in nanotechnology.

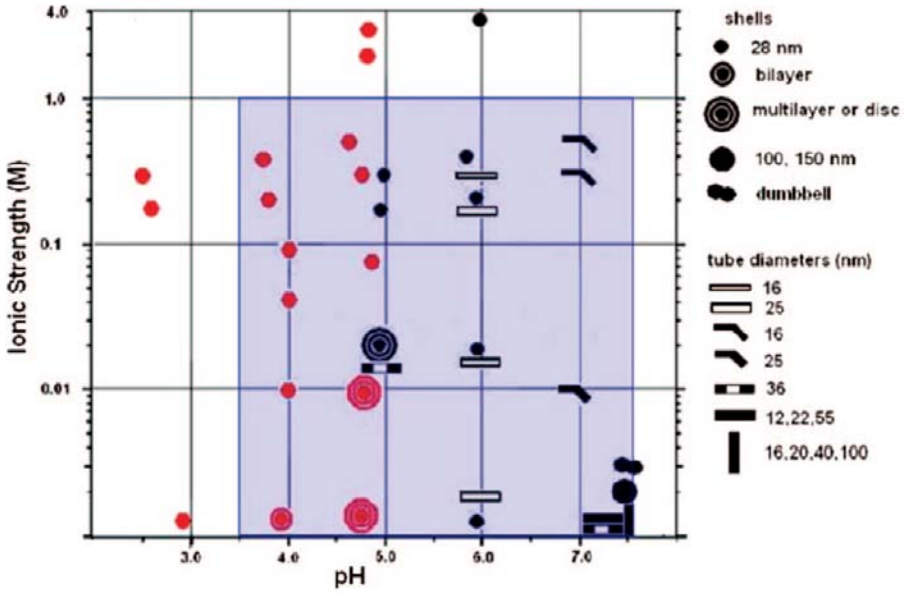

Figure 1-21. Experimental phase diagram of CCMV capsid protein as a function of pH and ionic strength. Adapted from [100].

The interactions between capsid dimers can be roughly divided into a short-ranged hydrophobic interaction and a long-ranged electrostatic interaction. The hydrophobic interaction is mainly a function of the temperature and of the buried area after the association of dimers, which is almost constant for a specific system. Thus, the only approach to manipulate

the interaction strength of dimers is based on adjustment of the electrostatic interaction by simply controlling the ionic strength and pH of the solution [101]. The phase diagram of the assembly of CCMV empty capsids as a function of ionic strength and pH is displayed in Figure 1-21, and it can be found that different kinds of capsid form under different environment conditions. Firstly, the most common capsid (single-layered and spherical, see in Figure 1-22(A)) is usually observed in an acid solution with an ionic strength over 0.3 M, while surprisingly a multi-layered capsid (Figure 1-22(B)) forms by decreasing the ionic strength to a very low level due to the typical charge distribution of the dimer (negatively charged for the outer domains while positively charged for inner domains). In addition, at around neutral pH a capsid in the form of tube is always found when salt concentration approaches zero (Figure 1-22(C)). Given that a simple and clear phase diagram on the assembly of empty capsids had been proposed, the final assembled products are much more complicated. In many cases, different types of capsid coexist in the same solution (see Figure 1-23(A) and (B)). Moreover, unlike the assembly *in vivo,* many assembled capsids are full of various defects (see Figure 1-23(C)). To simplify the assembly system, the final product should be as homogeneous as possible and therefore more attention has been paid on the single-layered spherical capsid.

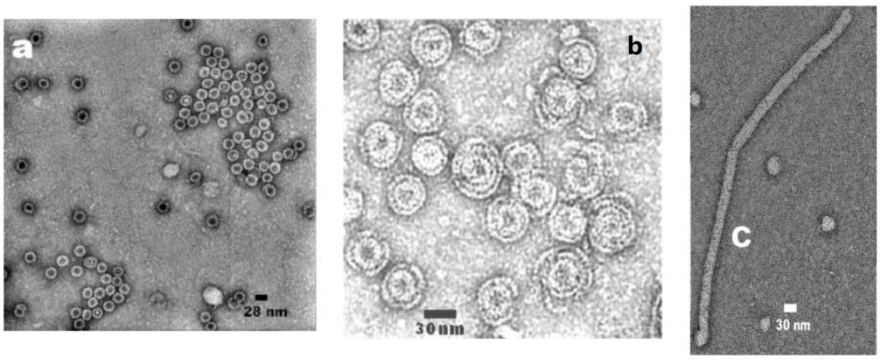

Figure 1-22. Different assembled structures of CCMV capsid proteins, (a) single-layered spherical capsids, (b) multi-layered spherical capsids and (c) rodlike capsids. Adapted from [100].

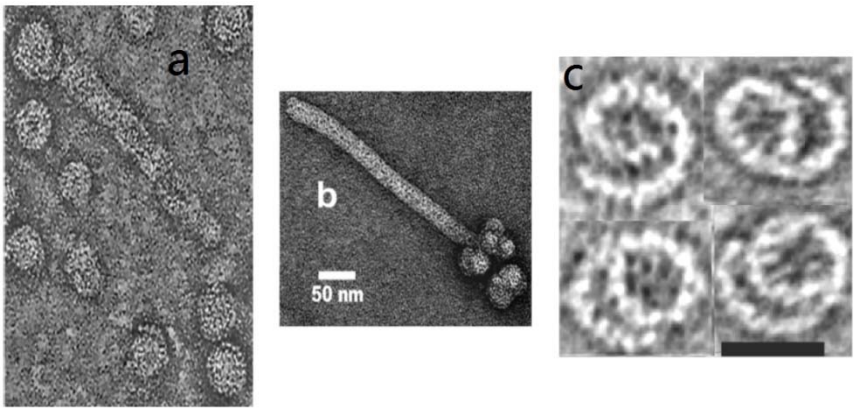

Figure 1-23. Coexistence of different structures under specific conditions (a) (pH = 5.65, 0.01 M sodium cacodylate buffer, and no added salt) and (b) (pH = 6.0, 0.01 M sodium cacodylate buffer, no added salt), and assembly defects (c). Adapted from [100].

Even for a simple structure like single-layered spherical empty CCMV capsids, the assembly process is sophisticated and the mechanism behind it is still controversial [67]. During the assembly process, the components in solution can be multivarious. To reveal the assembly mechanism, one must extract important components from them. It was proposed that the capsid proteins of CCMV normally exist in the form of dimer, which acts as the basic assembly subunit. However, the following reactions between dimers are unclear. Based on high-resolution conformation observed on swelling of CCMV capsid, Speir et al. proposed that the dimers form a hexamer of dimers stabilized by a beta-barrel structure and playing a role of nucleus [57]. Other dimers subsequently add to this nucleus and lead to the formation of a full capsid. Yet in the investigation of the assembly kinetics of CCMV by light scattering [39], Zlotnick et al. found that the assembly reaction consisted of two concurrent reactions, that is, a rapid reaction on oligomer formation (steep slope region) and a slow reaction on capsid formation (slow slope region), as indicated by Figure 1-24(A). Moreover, the reaction rate for the rapid reaction suggests a fifth-order dependence on the initial concentration of capsid proteins, suggesting the formation of a pentamer of dimers as nucleus (Figure 1-24(B)). The mechanism proposed by Zlotnick et al. is now widely accepted. With this consensus, Zlotnick et al. speculated that the free dimers in the solution add cooperatively to the formed pentamer to yield a T=3 capsid, as illustrated in Figure 1-24(C). But if the initial concentration of capsid proteins is too high, the formed pentamer will subsequently associate together to lead to a T=2 capsid [33].

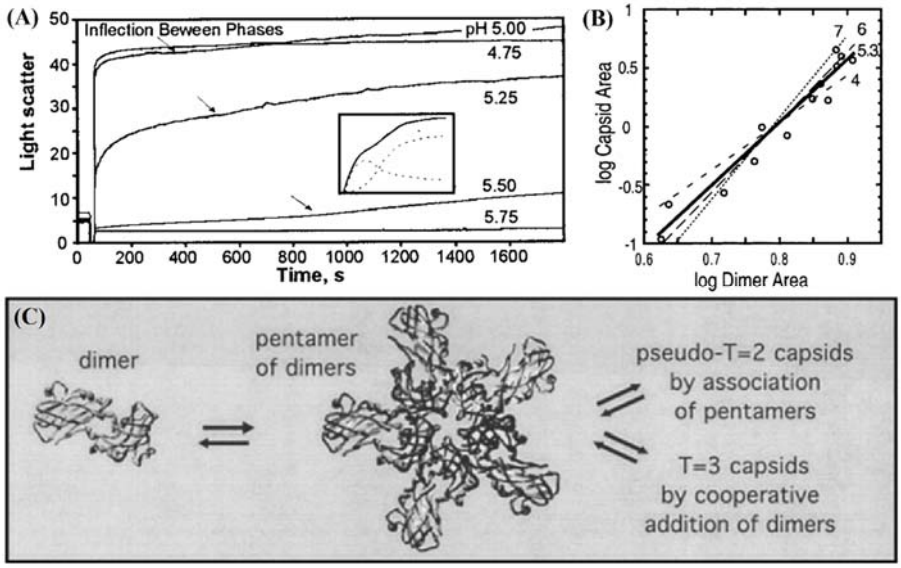

Figure 1-24. Assembly mechanism of CCMV capsid proteins proposed by Zlotnick et al. (A) Static light scattering (SLS) measurements of CCMV assembly at different pHs. (B) The assembly kinetics analysis on the concentration dependence of assembly basing on theoretical assembly models where a 5.3 power concentration dependence was found, suggesting the formation of pentamer of dimers. (C) Scheme of assembly mechanism. Adapted from [39].

1.3.3 Role of viral genome

Although CCMV capsids can be fabricated *in vitro* without viral genome, empty capsids are not thermodynamically stable at neutral pH, which limits their applications. Conversely, the introduction of viral genome will greatly improve the capsid stability [102] while also

complicating the assembly process. Apart from the interaction between capsid proteins (CPs), the interaction between CP and RNA is another decisive factor on the formation of viral particles. For the case of CCMV, the CPs-RNA interaction mainly originates from the non-specific electrostatic interaction between the positively charged N-terminal tail of the CPs and the negatively charged RNA phosphate groups [103]. The N-terminal tail (residues from 1 to 26) contributes +10 e and is mostly due to the charges carried by arginines (pKa 12.10) and lysines (pKa 10.67), whose charge dissociation both demonstrates a weak sensitivity to pH under normal range of value. Thus, the CP-RNA interaction can be simply manipulated through modulating the ionic strength of the solution [104]. At high ionic strength, the CP-RNA interaction will be turned off and both CP and RNA will remain dispersed in the solution, while the CP-RNA interaction will be turned on at low ionic strength and they start to associate together. Regarding the interaction strength between CP and RNA, it was estimated by Chevreuil et al. to be ~ -7 k_BT_0,[105] which is slightly higher than the CP-CP interaction (~ -4 to -5 k_BT_0).

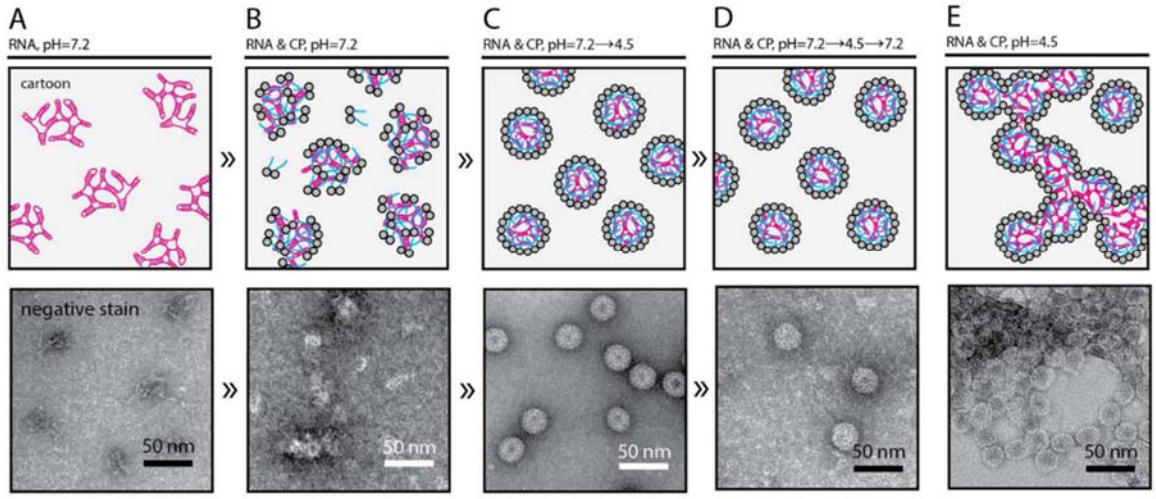

Figure 1-25. CCMV virions assembled by two pathways: A two-step assembly A→B→C and a one-step assembly A→E. (A) The CCMV RNA solution in a dispersed state at neutral pH. (B) Formation of RNA-protein complex after introducing capsid proteins into the RNA solution. (C) Complete CCMV capsids formed after triggering the structural reorganization of the RNA-protein complex at pH 4.5. (D) The reassembled capsids formed by an "annealed" assembly process (dialyzing the capsids with buffer solutions at different pHs (pH 7.2 → pH 4.5 → pH 7.2)). (E) The irregular structure formed by directly mixing the RNA solution with capsid protein solution at pH 4.5. Adapted from [98].

Firstly, the order of mixing CPs solution and RNA solution affects the final products, that is, the so-called one-step assembly and two-step assembly [104], as illustrated by Figure 1-25. The general assembly protocol begins with mixing RNA and CP at neutral pH and high ionic strength, followed by subsequent dialysis and equilibration. At this moment, both the CP-CP and CP-RNA interactions are turned off and the reactants keep dispersed. Next, we can start the assembly by modulating pH (turn on CP-CP interactions) and ionic strength (turning on CP-RNA interactions) to be in a specific condition. For a one-step assembly, these two parameters are simultaneously modulated in one step, while in the two-step assembly they are modulated one after another. The products formed by one-step assembly are generally clusters of irregular

spherical capsids connecting each other by the RNA chains, also known as multiplet structures. In those clusters, each RNA chain is shared by two or three capsids. The formation of multiplet structures was attributed to the strong attraction between CPs which results in the assembly falling into kinetic traps as revealed by coarse-grained simulations. The general two-step assembly firstly switches on the CP-RNA interactions to allow the CPs to fully bind the RNA and to generate a CP-RNA disordered complex, followed by an acidification of the solution to turn on the CP-CP interactions and trigger the assembly of CPs on the surface of RNA. The CP-RNA complex will then undergo a structural relaxation which can effectively avoid the kinetic traps and form a regular capsid.

In order to improve the packaging efficiency of RNA, one should carefully modulate the ratio between RNA and CP. An excess of CP is necessary to ensure that all RNA is effectively packaged [103,106-108]. In addition, the length of RNA will also affect the structure of the capsid [109]. A long RNA may lead to a larger size of capsid, but the overlong RNA might also act as a tether linking multiple capsids [107], while for short RNAs, several of them will be packaged cooperatively into one capsid [110]. It is interesting to find that the CP-RNA disordered complexes demonstrate a selectivity on the packaged RNA by exchanging the packaged one with a newly introduced one depending on the length and sequence of the RNA [111].

1.3.4 Perspectives

CCMV is an ideal model virus for uncovering the physical principles of virus assembly thanks to its simple but typical structure, the convenience in preparation as well as the safety on manipulation. Understanding the assembly mechanism as well as the mechanism of genome encapsulation will facilitate the development of nano-engineered objects such as vectors for drug delivery.

Cargo encapsulation and delivery is a hot topic in the field of materials science. How to deliver the cargo that is needed to the target, safely and precisely, is still a challenging work for drug delivery. The structurally regular virus capsids demonstrate a great potential in this field for their intrinsic encapsulation capability, and CCMV becomes a typical model due to its brilliant properties and to the possibility of large-scale production. Those modified virus capsids were called virus-like particles (VLPs) and granted various additional functions. Schoonen et al. [86] fabricated a nanoreactor by chemically modifying CCMV capsids to grant them a responsiveness to physiological conditions. Brasch et al. [85] manufactured nanocages assembled by CCMV capsid proteins modified by using dual-tasking nucleic acid tags, which are capable of coating enzyme-DNA hybrids inside CCMV capsid.

Recently, virus particles also found applications in other fields. For example, Liljestrom et al. [89] assembled CCMV particles and oppositely charged proteins together into three-dimensional crystals with an adjustable structure. Meanwhile, a two-dimensional cluster of nanocarriers [112] was also tentatively prepared by gluing CCMV VLPs that encapsulate functionalized nanoparticles with soft macromolecules, or by functionalizing a two-dimensional surface with CCMV VLPs. Kostiainen et al. [113] employed CCMV VLPs to prepare temperature-switchable virus-polymer complexes by chemically modifying the surface

of the VLPs with thermal response polymers to form a branched nanocages. CCMV VLPs were also introduced into hydrogels to grant encapsulation capability to the hydrogels [114].

Studies on the applications of CCMV and that on the assembly mechanism will complement each other. The assembly mechanism of CCMV obtains a thorough examination during the application development; meanwhile the properties of CCMV will also benefit from further investigations. On the other hand, the study of mechanisms will fertilize the development of new applications, such as improving the preparation process. With more and more experimental findings, the thirst of obtaining atomic details becomes increasingly stronger. Higher resolution techniques in aqueous environment or an atom-scale theoretical method with a time-resolved capability will be helpful on virus self-assembly, and molecular dynamics simulation is one of the best options.

1.4 References

[1] M. Breitbart and F. Rohwer, Trends in Microbiology **13**, 278 (2005).

[2] X. K. Yu, S. Shah, M. Lee, W. Dai, P. Lo, W. Britt, H. Zhu, F. Y. Liu, and Z. H. Zhou, Journal of Structural Biology **174**, 451 (2011).

[3] A. J. A. v. M. b. Mankertz, 355 (2008).

[4] D. Arslan, M. Legendre, V. Seltzer, C. Abergel, and J. M. Claverie, Proceedings of the National Academy of Sciences of the United States of America **108**, 17486 (2011).

[5] N. Philippe *et al.*, Science **341**, 281 (2013).

[6] M. Legendre *et al.*, Proceedings of the National Academy of Sciences of the United States of America **111**, 4274 (2014).

[7] W. H. Roos, R. Bruinsma, and G. J. L. Wuite, Nature Physics **6**, 733 (2010).

[8] F. H. C. Crick and J. D. Watson, Nature **177**, 473 (1956).

[9] D. L. Caspar and A. Klug, in *Cold Spring Harbor symposia on quantitative biology* (Cold Spring Harbor Laboratory Press, 1962), pp. 1.

[10] R. Zandi, D. Reguera, R. F. Bruinsma, W. M. Gelbart, and J. Rudnick, Proceedings of the National Academy of Sciences of the United States of America **101**, 15556 (2004).

[11] R. F. Bruinsma, W. M. Gelbart, D. Reguera, J. Rudnick, and R. Zandi, Physical Review Letters **90**, 248101 (2003).

[12] D. Baxby, Vaccine **17**, 301 (1999).

[13] A. Jefferson, V. E. Cadet, and A. Hielscher, Critical Reviews in Oncology Hematology **95**, 407 (2015).

[14] H. Lodish, A. Berk, S. L. Zipursky, P. Matsudaira, D. Baltimore, and J. Darnell, *Molecular cell biology* (WH Freeman New York, 1995), Vol. 3.

[15] Z. S. Guo *et al.*, Cancer Research **65**, 9991 (2005).

[16] M. Castedo *et al.*, Oncogene **23**, 4362 (2004).

[17] C. Caranta, *Recent advances in plant virology* (Horizon Scientific Press, 2011).

[18] G. C. Wang *et al.*, Acs Nano **9**, 799 (2015).

[19] H. Masarapu, B. K. Patel, P. L. Chariou, H. Hu, N. M. Gulati, B. L. Carpenter, R. A. Ghiladi, S. Shukla, and N. F. Steinmetz, Biomacromolecules **18**, 4141 (2017).

[20] Y. Azuma, M. Herger, and D. Hilvert, Journal of the American Chemical Society **140**, 558

(2018).
[21] K. Zhou, Y. G. Ke, and Q. B. Wang, Journal of the American Chemical Society **140**, 8074 (2018).
[22] E. M. Plummer and M. Manchester, Wiley Interdisciplinary Reviews-Nanomedicine and Nanobiotechnology **3**, 174 (2011).
[23] C. M. Soto, A. S. Blum, G. J. Vora, N. Lebedev, C. E. Meador, A. P. Won, A. Chatterji, J. E. Johnson, and B. R. Ratna, Journal of the American Chemical Society **128**, 5184 (2006).
[24] A. S. Blum *et al.*, Small **1**, 702 (2005).
[25] S. M. Yu, Nature Nanotechnology **7**, 343 (2012).
[26] A. Zlotnick, Virology **315**, 269 (2003).
[27] A. Oppenheim, O. Ben-Nun-Shaul, S. Mukherjee, and M. Abd-El-Latif, Computational and Mathematical Methods in Medicine **9**, 265 (2008).
[28] M. G. Rossmann and J. E. Johnson, Annual Review of Biochemistry **58**, 533 (1989).
[29] W. K. Kegel and P. van der Schoot, Biophysical Journal **86**, 3905 (2004).
[30] P. E. Prevelige, J. King, and J. L. Silva, Biophysical Journal **66**, 1631 (1994).
[31] J. C. Y. Wang, C. Chen, V. Rayaprolu, S. Mukhopadhyay, and A. Zlotnick, Acs Nano **9**, 8898 (2015).
[32] P. Ceres and A. Zlotnick, Biochemistry **41**, 11525 (2002).
[33] J. M. Johnson, J. H. Tang, Y. Nyame, D. Willits, M. J. Young, and A. Zlotnick, Nano Letters **5**, 765 (2005).
[34] A. Zlotnick, Journal of Molecular Biology **241**, 59 (1994).
[35] M. F. Hagan, in *Advances in Chemical Physics, Vol 155*, edited by S. A. Rice, and A. R. Dinner2014), pp. 1.
[36] C. A. Lutomski, N. A. Lyktey, E. E. Pierson, Z. C. Zhao, A. Zlotnick, and M. F. Jarrold, Journal of the American Chemical Society **140**, 5784 (2018).
[37] A. Zlotnick, J. M. Johnson, P. W. Wingfield, S. J. Stahl, and D. Endres, Biochemistry **38**, 14644 (1999).
[38] M. Medrano, M. A. Fuertes, A. Valbuena, P. J. P. Carrillo, A. Rodriguez-Huete, and M. G. Mateu, Journal of the American Chemical Society **138**, 15385 (2016).
[39] A. Zlotnick, R. Aldrich, J. M. Johnson, P. Ceres, and M. J. Young, Virology **277**, 450 (2000).
[40] Y. Khudyakov and P. Pumpens, *Viral Nanotechnology* (CRC Press, 2015).
[41] G. Tresset, C. Le Coeur, J. F. Bryche, M. Tatou, M. Zeghal, A. Charpilienne, D. Poncet, D. Constantin, and S. Bressanelli, Journal of the American Chemical Society **135**, 15373 (2013).
[42] S. Y. Li, P. Roy, A. Travesset, and R. Zandi, Proceedings of the National Academy of Sciences of the United States of America **115**, 10971 (2018).
[43] P. E. Prevelige, D. Thomas, and J. King, Biophysical Journal **64**, 824 (1993).
[44] R. Tuma, M. H. Parker, P. Weigele, L. Sampson, Y. H. Sun, N. R. Krishna, S. Casjens, G. J. Thomas, and P. E. Prevelige, Journal of Molecular Biology **281**, 81 (1998).
[45] Y. H. Sun, M. H. Parker, P. Weigele, S. Casjens, P. E. Prevelige, and N. R. Krishna, Journal of Molecular Biology **297**, 1195 (2000).
[46] K. N. Parent, S. M. Doyle, E. Anderson, and C. M. Teschke, Virology **340**, 33 (2005).
[47] M. H. Parker, W. F. Stafford, and P. E. Prevelige, Journal of Molecular Biology **268**, 655 (1997).
[48] R. Tuma, H. Tsuruta, K. H. French, and P. E. Prevelige, Journal of Molecular Biology **381**,

1395 (2008).
[49] T. Dokland, R. McKenna, L. L. Ilag, B. R. Bowman, N. L. Incardona, B. A. Fane, and M. G. Rossmann, Nature **389**, 308 (1997).
[50] J. E. Cherwa, A. Uchiyama, and B. A. Fane, Journal of Virology **82**, 5774 (2008).
[51] R. McKenna, D. Xia, P. Willingmann, L. L. Ilag, S. Krishnaswamy, M. G. Rossmann, N. H. Olson, T. S. Baker, and N. L. Incardona, Nature **355**, 137 (1992).
[52] B. A. Fane and P. E. Prevelige, Virus Structure **64**, 259 (2003).
[53] A. Zlotnick and B. A. Fane, in *Structural Virology*, edited by M. AgbandjeMckenna, and R. McKenna2011), pp. 180.
[54] S. Manuguri *et al.*, Nano Letters **18**, 5138 (2018).
[55] C. R. Bourne, S. P. Katen, M. R. Fulz, C. Packianathan, and A. Zlotnick, Biochemistry **48**, 1736 (2009).
[56] X. K. Yu, L. Jin, J. Jih, C. H. Shih, and Z. H. Zhou, Plos One **8**, UNSP e69729 (2013).
[57] J. A. Speir, S. Munshi, G. J. Wang, T. S. Baker, and J. E. Johnson, Structure **3**, 63 (1995).
[58] X. K. Yu, J. Jih, J. S. Jiang, and Z. H. Zhou, Science **356**, eaam6892 (2017).
[59] W. H. Roos *et al.*, Biophysical Journal **99**, 1175 (2010).
[60] M. Castellanos, R. Perez, P. J. P. Carrillo, P. J. de Pablo, and M. G. Mateu, Biophysical Journal **102**, 2615 (2012).
[61] G. Siuzdak, B. Bothner, M. Yeager, C. Brugidou, C. M. Fauquet, K. Hoey, and C. M. Chang, Chemistry & Biology **3**, 45 (1996).
[62] J. Snijder, R. J. Rose, D. Veesler, J. E. Johnson, and A. J. R. Heck, Angewandte Chemie-International Edition **52**, 4020 (2013).
[63] C. Uetrecht and A. J. R. Heck, Angewandte Chemie-International Edition **50**, 8248 (2011).
[64] E. E. Pierson, D. Z. Keifer, L. Selzer, L. S. Lee, N. C. Contino, J. C. Y. Wang, A. Zlotnick, and M. F. Jarrold, Journal of the American Chemical Society **136**, 3536 (2014).
[65] C. Uetrecht, N. R. Watts, S. J. Stahl, P. T. Wingfield, A. C. Steven, and A. J. R. Heck, Physical Chemistry Chemical Physics **12**, 13368 (2010).
[66] C. Uetrecht, I. M. Barbu, G. K. Shoemaker, E. van Duijn, and A. J. R. Heck, Nature Chemistry **3**, 126 (2011).
[67] D. Law-Hine, M. Zeghal, S. Bressanelli, D. Constantin, and G. Tresset, Soft Matter **12**, 6728 (2016).
[68] D. Law-Hine, A. K. Sahoo, V. Bailleux, M. Zeghal, S. Prevost, P. K. Maiti, S. Bressanelli, D. Constantin, and G. Tresset, Journal of Physical Chemistry Letters **6**, 3471 (2015).
[69] S. Kler, R. Asor, C. L. Li, A. Ginsburg, D. Harries, A. Oppenheim, A. Zlotnick, and U. Raviv, Journal of the American Chemical Society **134**, 8823 (2012).
[70] R. L. C. Leung *et al.*, Journal of the American Chemical Society **139**, 5277 (2017).
[71] Z. D. Harms, L. Selzer, A. Zlotnick, and S. C. Jacobson, Acs Nano **9**, 9087 (2015).
[72] R. F. Garmann, R. Sportsman, C. Beren, V. N. Manoharan, C. M. Knobler, and W. M. Gelbart, Journal of the American Chemical Society **137**, 7584 (2015).
[73] K. M. Zhou, L. C. Li, Z. N. Tan, A. Zlotnick, and S. C. Jacobson, Journal of the American Chemical Society **133**, 1618 (2011).
[74] R. J. Usselman, E. D. Walter, D. Willits, T. Douglas, M. Young, and D. J. Singel, Journal of the American Chemical Society **133**, 4156 (2011).
[75] M. Mosayebi, D. K. Shoemark, J. M. Fletcher, R. B. Sessions, N. Linden, D. N. Woolfson,

and T. B. Liverpool, Proceedings of the National Academy of Sciences of the United States of America **114**, 9014 (2017).
[76] R. F. Bruinsma, M. Comas-Garcia, R. F. Garmann, and A. Y. Grosberg, Physical Review E **93**, 032405 (2016).
[77] C. Chen, C. C. Kao, and B. Dragnea, Journal of Physical Chemistry A **112**, 9405 (2008).
[78] T. Keef, C. Micheletti, and R. Twarock, Journal of Theoretical Biology **242**, 713 (2006).
[79] V. L. Morton, E. C. Dykeman, N. J. Stonehouse, A. E. Ashcroft, R. Twarock, and P. G. Stockley, Journal of Molecular Biology **401**, 298 (2010).
[80] M. Tsiang, A. Niedziela-Majka, M. Hung, D. B. Jin, E. Hu, S. Yant, D. Samuel, X. H. Liu, and R. Sakowicz, Biochemistry **51**, 4416 (2012).
[81] O. M. Elrad and M. F. Hagan, Physical Biology **7**, 045003 (2010).
[82] O. M. Elrad and M. F. Hagan, Nano Letters **8**, 3850 (2008).
[83] J. P. Mahalik and M. Muthukumar, Journal of Chemical Physics **136**, 135101 (2012).
[84] H. D. Nguyen, V. S. Reddy, and C. L. Brooks, Nano Letters **7**, 338 (2007).
[85] M. Brasch, R. M. Putri, M. V. de Ruiter, D. Luque, M. S. T. Koay, J. R. Caston, and J. Cornelissen, Journal of the American Chemical Society **139**, 1512 (2017).
[86] L. Schoonen, S. Maassen, R. J. M. Nolte, and J. C. M. van Hest, Biomacromolecules **18**, 3492 (2017).
[87] I. J. Minten, L. J. A. Hendriks, R. J. M. Nolte, and J. Cornelissen, Journal of the American Chemical Society **131**, 17771 (2009).
[88] M. Comellas-Aragones, A. de la Escosura, A. J. Dirks, A. van der Ham, A. Fuste-Cune, J. Cornelissen, and R. J. M. Nolte, Biomacromolecules **10**, 3141 (2009).
[89] V. Liljestrom, J. Mikkila, and M. A. Kostiainen, Nature Communications **5**, 4445 (2014).
[90] O. Kononova, J. Snijder, Y. Kholodov, K. A. Marx, G. J. L. Wuite, W. H. Roos, and V. Barsegov, Plos Computational Biology **12**, e1004729 (2016).
[91] J. R. Vega-Acosta, R. D. Cadena-Nava, W. M. Gelbar, C. M. Knobler, and J. Ruiz-Garcia, Journal of Physical Chemistry B **118**, 1984 (2014).
[92] M. Hernando-Perez, A. X. Cartagena-Rivera, A. L. Bozic, P. J. P. Carrillo, C. S. Martin, M. G. Mateu, A. Raman, R. Podgornik, and P. J. de Pablo, Nanoscale **7**, 17289 (2015).
[93] D. Q. Zhang, R. Konecny, N. A. Baker, and J. A. McCammon, Biopolymers **75**, 325 (2004).
[94] J. M. Johnson, D. A. Willits, M. J. Young, and A. Zlotnick, Journal of Molecular Biology **335**, 455 (2004).
[95] A. M. Dzianott and J. I. J. V. Bujarski, **185**, 553 (1991).
[96] R. F. Allison, M. Janda, and P. Ahlquist, Virology **172**, 321 (1989).
[97] R. Dasgupta and P. Kaesberg, Nucleic Acids Research **10**, 703 (1982).
[98] R. F. Garmann, M. Comas-Garcia, C. M. Knobler, and W. M. Gelbart, Accounts of Chemical Research **49**, 48 (2016).
[99] A. Gopal, Z. H. Zhou, C. M. Knobler, and W. M. Gelbart, Rna **18**, 284 (2012).
[100] L. Lavelle, M. Gingery, M. Phillips, W. M. Gelbart, C. M. Knobler, R. D. Cadena-Nava, J. R. Vega-Acosta, L. A. Pinedo-Torres, and J. Ruiz-Garcia, Journal of Physical Chemistry B **113**, 3813 (2009).
[101] L. Lavelle, J. P. Michel, and M. Gingery, Journal of Virological Methods **146**, 311 (2007).
[102] B. D. Wilts, I. A. T. Schaap, and C. F. Schmidt, Biophysical Journal **108**, 2541 (2015).

[103] R. F. Garmann, M. Comas-Garcia, M. S. T. Koay, J. Cornelissen, C. M. Knobler, and W. M. Gelbart, Journal of Virology **88**, 10472 (2014).
[104] R. F. Garmann, M. Comas-Garcia, A. Gopal, C. M. Knobler, and W. M. Gelbart, Journal of Molecular Biology **426**, 1050 (2014).
[105] M. Chevreui *et al.*, Nature Communications **9**, 3071 (2018).
[106] G. Erdemci-Tandogan, H. Orland, and R. Zandi, Physical Review Letters **119**, 188102 (2017).
[107] R. D. Cadena-Nava, M. Comas-Garcia, R. F. Garmann, A. L. N. Rao, C. M. Knobler, and W. M. Gelbart, Journal of Virology **86**, 3318 (2012).
[108] G. Tresset, M. Tatou, C. Le Coeur, M. Zeghal, V. Bailleux, A. Lecchi, K. Brach, M. Klekotko, and L. Porcar, Physical Review Letters **113**, 128305 (2014).
[109] C. Beren, L. L. Dreesens, K. N. Liu, C. M. Knobler, and W. M. Gelbart, Biophysical Journal **113**, 339 (2017).
[110] M. Comas-Garcia, R. F. Garmann, S. W. Singaram, A. Ben-Shaul, C. M. Knobler, and W. M. Gebart, Journal of Physical Chemistry B **118**, 7510 (2014).
[111] M. Comas-Garcia, R. D. Cadena-Nava, A. L. N. Rao, C. M. Knobler, and W. M. Gelbart, Journal of Virology **86**, 12271 (2012).
[112] J. Mikkila, H. Rosilo, S. Nummelin, J. Seitsonen, J. Ruokolainen, and M. A. Kostiainen, Acs Macro Letters **2**, 720 (2013).
[113] M. A. Kostiainen, C. Pietsch, R. Hoogenboom, R. J. M. Nolte, and J. Cornelissen, Advanced Functional Materials **21**, 2012 (2011).
[114] L. L. Yang, A. J. Liu, M. V. de Ruiter, C. A. Hommersom, N. Katsonis, P. Jonkheijm, and J. Cornelissen, Nanoscale **10**, 4123 (2018).

Chapter 2

Methodology

In this thesis, Monte Carlo and molecular dynamics simulations are the main approaches of investigation. In this chapter, the principles of these two methods would be introduced respectively.

2.1 Monte Carlo simulations

2.1.1 Principles

Monte Carlo (MC) simulation or experiment is a computational algorithm to solve deterministic problems by statistically sampling to obtain numerical results. Monte Carlo, a place famous for gambling, was suggested and used as the name of MC method by Nicholas Metropolis, a colleague of John von Neumann and Stanisław Marcin Ulam who made a great contribution on the development of MC simulation to study problems in fission devices. MC simulation can nowadays find applications in any problems that can be interpreted probabilistically, ranging from estimation of multi-dimensional integral or optimization problems to complex systems (in physical, chemical, biological, social sciences) involving a huge number of particles.

For a system with N particles, its partition function Z can be classically given as

$$Z = c \int d\boldsymbol{p}^N d\boldsymbol{r}^N \exp[-\beta \mathcal{H}(\boldsymbol{p}^N, \boldsymbol{r}^N)], \tag{2.1}$$

where r^N stands for the coordinates of all N particles, and p^N for the corresponding momenta, and $\beta = 1/k_B T$ with k_B the Boltzmann constant and T the temperature of the system. $H(p^N, r^N)$ is the Hamiltonian of the system with $H = K + U$, K and U being the kinetic energy and the potential energy of the system, respectively. c is a constant of proportionality, for N identical particles, $c = 1/N!\,h^{3N}$. Then the probability density of finding the system in a configuration (p^N, r^N) is denoted as

$$\mathcal{N}(\boldsymbol{r}^N) \equiv \frac{exp[-\beta \mathcal{H}(\boldsymbol{p}^N, \boldsymbol{r}^N)]}{Z}. \tag{2.2}$$

The average of an observable A as a function of coordinates (r) and momenta (p) can be expressed by classical statistical mechanics in the canonical statistical ensemble as

$$\langle A \rangle = \frac{\int dp^N dr^N A(p^N, r^N) exp[-\beta H(p^N, r^N)]}{\int dp^N dr^N exp[-\beta H(p^N, r^N)]}. \tag{2.3}$$

Since K is a quadratic function of the momenta, the integration over momenta can be solved analytically. If the observable A does not depend on momenta, equation (2.1) can be simplified into

$$\langle A \rangle = \frac{\int dr^N exp[-\beta U(r^N)] A(r^N)}{\int dr^N exp[-\beta U(r^N)]}. \quad (2.4)$$

By equation (2.4), we can evaluate the observable *A*. However, for most of cases the multidimensional integral over r^N is impossible to be calculated analytically. Thus numerical techniques must be introduced and Monte Carlo simulation is one of them.

The Metropolis Method

To avoid the calculation of the multidimensional integral over particle coordinates which is not possible even if using MC method, Metropolis and coworkers provided an approach to know the ratio of the two integrals in equation 2.4.

Unlike the importance sampling where we know a priori the probability density $N(r^N)$, we can know only the Boltzmann factor $exp[-\beta U(r^N)]$, in other words, the relative but not the absolute probability density. Let's firstly consider a simple case that an old configuration *o* with a Boltzmann factor $exp[-\beta U(o)]$ performs a small random evolution to reach a new configuration *n* with a Boltzmann factor $exp[-\beta U(n)]$. Now we need to decide whether we will accept or reject this trial evolution. Here we introduce the Metropolis scheme [1] to answer this question.

For a system in equilibrium, the probability distribution of configurations should remain constant, that is, the average number of accepted trial moves that result in the system leaving configuration *o* must be exactly equal to the number of accepted trial moves from all other configurations *n* to configuration *o*. This detailed balance condition can be described by

$$N(o)\pi(o \to n) = N(n)\pi(n \to o), \quad (2.5)$$

where $\pi(o \to n)$ is the transition probability from state *o* to state *n*, which can be equivalently evaluated by the probability of accepting a trial move (acc) in MC simulation due to the property that the average chance to perform a trial move from *o* to *n* is equal to that from *n* to *o*. Then equation (2.5) can be transformed into

$$\frac{acc(o \to n)}{acc(n \to o)} = \frac{N(n)}{N(o)} = exp\{-\beta[U(n) - U(o)]\}. \quad (2.6)$$

Of course, since the accepting probability cannot be greater than 1, equation (2.6) can be rewritten as

$$acc(o \to n) = \begin{cases} exp\{-\beta[\mathcal{U}(n) - \mathcal{U}(o)]\}, & if\ \mathcal{U}(n) - \mathcal{U}(o) > 0 \\ 1, & if\ \mathcal{U}(n) - \mathcal{U}(o) \le 0 \end{cases}. \quad (2.7)$$

By the Metropolis method, we can effectively avoid calculating the partition function and the evolution of the MC simulation can be easily carried out.

Ensembles

Different physical systems have different coupling with their environment. Some of them are isolated, some are closed systems exchanging only heat with a reservoir while others are open

systems able to also exchange (besides heat) particles with a reservoir. These distinctive coupling interactions with environment are called ensembles in statistical mechanics. Performing simulations in different ensembles will yield different statistical expressions for the average of observables. In a conventional MC simulation, the native ensemble is the canonical ensemble with constant number of particles N, constant volume V and constant temperature T. For regular simulations, there are four ensembles used widely, that is, the canonical ensemble, microcanonical ensemble, isobaric-isothermal ensemble and grand canonical ensemble.

The canonical ensemble is a closed ensemble where a system evolves in a constant temperature (T) without particles exchange (N) and without variation of volume (V). Its basic principles and simulation process can be found in the sections above. The canonical ensemble is highly compatible with Metropolis scheme.

Comparing with other ensembles, the microcanonical ensemble where the number of particles (N), the volume (V) and the energy (E) are constant (isolated system) is seldom used in simulations. It is hard to find applications in the microcanonical ensemble, but is the default ensemble for molecular dynamics simulations.

Isobaric isothermal ensemble

An isobaric-isothermal ensemble is a collection of systems whose particle number (N), pressure (P) and temperature (T) are constant, a common situation encountered in most experimental observations. Given this, the NPT ensemble is widely used in MC simulations. The first constant-pressure MC simulations were carried out by Wood to study a system of two-dimensional hard disks [2], however, the employed method is difficult to be extended to arbitrary continuous potentials. The method introduced below is based on the principle developed by McDonald for the NPT simulations of a Lennard-Jones fluid [3], and the basic idea of this method is schemed in Figure 2-1.

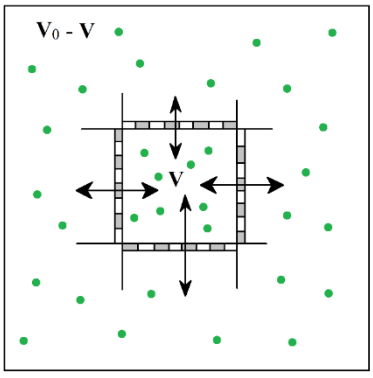

Figure 2-1. Schematic of an infinite NVT system with volume of V_0 divided into two pressure equilibrium subsystems: a smaller NPT subsystem with variable volume V and an ideal gas reservoir.

The partition function of the NPT ensemble is derived from the NVT ensemble. McDonald supposed an infinitely large NVT system with M particles in a cubic box of volume V_0 and side length L. The NVT system is considered as a combined system consisting of a much smaller subsystem with a variable volume V and an ideal gas reservoir surrounding the former. The subsystem is able to exchange heat and volume but not particles with the reservoir, so that the

N, P, and T condition can be maintained stable in the subsystem. By this, the partition function of this subsystem was derived and written in scaled coordinates s^N with $\boldsymbol{s}_i = \boldsymbol{r}_i/\mathrm{L}$ for a given particle *i* as

$$Q(N,P,T) \equiv \frac{\beta P}{\Lambda^{3N} N!} \int dV V^N exp(-\beta PV) \int ds^N exp[-\beta U(s^N, L)], \quad (2.8)$$

where $\Lambda = \sqrt{h^2/2\pi m k_B T}$ is the thermal de Broglie wavelength with Planck constant *h* and particle mass *m*. The probability density of a given microstate of volume V is given as

$$N_{N,P,T}(V) = \frac{V^N exp(-\beta PV) \int ds^N exp[-\beta U(s^N, L_\prime)]}{\int_0^{V_0} dV_\prime V_\prime^N exp(-\beta PV_\prime) \int ds^N exp[-\beta U(s^N, L_\prime)]}. \quad (2.9)$$

By carrying out NPT MC simulations with Metropolis scheme, the acceptance probability for a trial move of changing the volume of the system from V to $V' = V + \Delta V$, where ΔV is a small random change in volume, is expressed as

$$\mathrm{acc(o \rightarrow n)} = \min\left(1, \exp\left\{-\beta\left[\mathcal{U}(\boldsymbol{s}^N, V') - \mathcal{U}(\boldsymbol{s}^N, V) + P(V' - V) - N\beta^{-1}\ln\left(\frac{V'}{V}\right)\right]\right\}\right). \quad (2.10)$$

For NPT MC simulations of a molecular system, it should be noted that it is the center-of-mass positions of the molecules that should be scaled with the update of volume and not the positions of each atom in the molecules in order to keep the same bond lengths.

Grand canonical ensemble

A grand canonical ensemble where the temperature (T), volume (V) and chemical potential (μ) are fixed, is a situation occasionally occurring in experiments, such as the adsorbent in contact with a reservoir with constant chemical potential and temperature. The partition function of the grand canonical ensemble can be derived from NVT ensemble as well in a similar way as the derivation in NPT ensemble. It was first implemented by Norman and Filinov [4] and improved by other groups. In the μVT ensemble, the exchange between the subsystem and the ideal gas reservoir involves (in addition of heat) particles instead of volume. When the size of NVT system and the number of particles in the NVT system are getting close to infinity, the partition function of the μVT ensemble is expressed as

$$Q(\mu, V, T) = \sum_{N=0}^{\infty} \frac{exp(\beta\mu N)}{\Lambda^{3N} N!} \int ds^N exp[-\beta U(s^N)], \quad (2.11)$$

where μ is the chemical potential of the reservoir, given for an ideal gas by $\mu = k_B T ln\Lambda^3\rho$. Then the corresponding probability density of a microstate with N particles is given as

$$N_{\mu VT}(s^N; N) \propto \frac{exp(\beta\mu N) V^N}{\Lambda^{3N} N!} exp[-\beta U(s^N)]. \quad (2.12)$$

For the evolution of a grand canonical simulation, two trial moves are necessary for each MC step, one for the displacement of particles and another one for the insertion or removal of particles. The acceptance probabilities are

$$acc_{displacment}(s \to s') = min(1, exp\{-\beta[U(s'^N) - U(s^N)]\}) \quad (2.13)$$

$$acc_{insertion}(N \to N+1) = min\left(1, \frac{V}{\Lambda^3(N+1)} exp\{-\beta[\mu - U(N+1) + U(N)]\}\right) \quad (2.14)$$

$$acc_{removal}(N \to N-1) = min\left(1, \frac{\Lambda^3 N}{V} exp\{-\beta[\mu + U(N) - U(N-1)]\}\right) \quad (2.15)$$

Simulation protocol

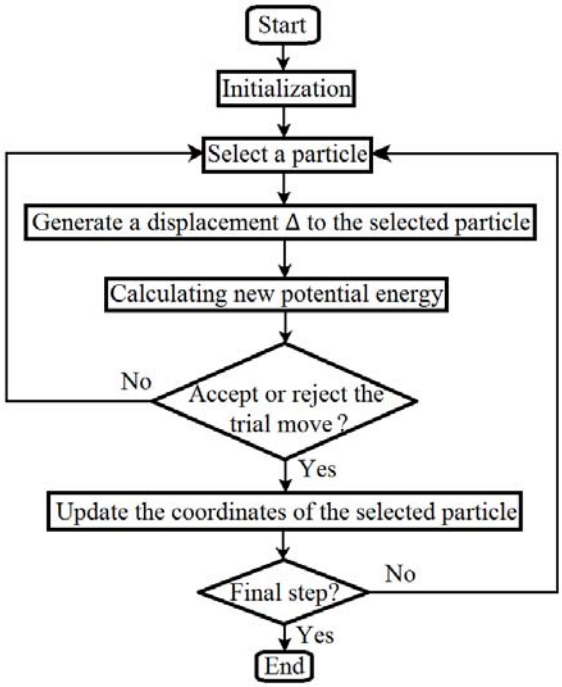

Figure 2-2. The workflow to perform a multi-particle MC simulation.

For simulations on a particle system, the workflow of MC simulation can be summarized as in Figure 2-2. The simulation starts by inserting the particles into a simulation box and initializing their positions. By uniform random sampling, a particle is selected and either translationally displaced by a small distance ± Δ/2 to a new position and/or rotated by a small amount (for molecules). The potential energy of the system is then updated and the trial move is accepted or rejected according to the Metropolis criterion. This procedure is repeated until the termination of the simulation.

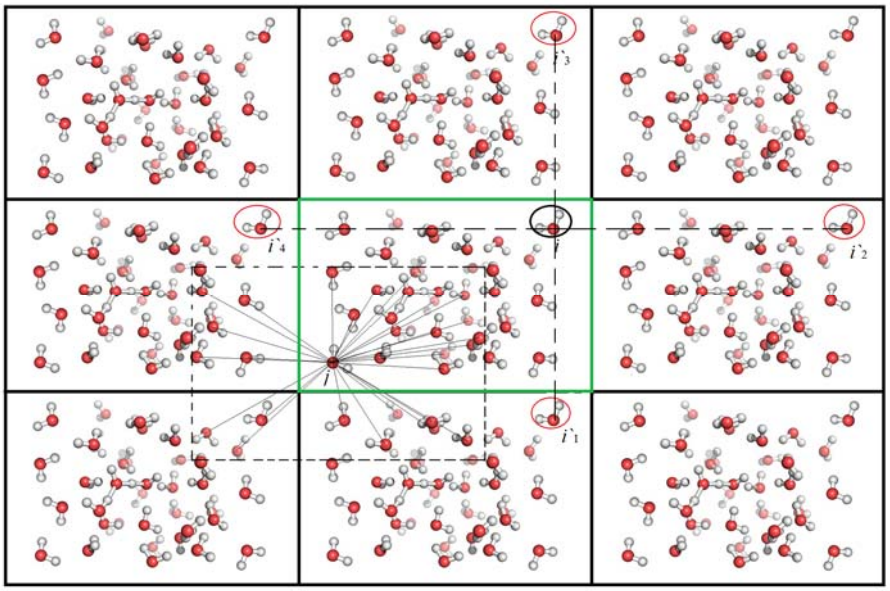

Figure 2-3. Schematic of periodic boundary conditions and minimum image in 2D. The rectangle in green denotes the simulation box while those in black are its six nearest neighbor images. Each water molecule *i* can found its periodic image particles (*i*`1, *i*`2, *i*`3, and *i*`4) in each directions. The rectangle in dashed line is centered at a given water molecule *j*, which interacts with the closest periodic images of all the other molecules present in the system.

In order to remove the boundary effect of the simulation box and finally mimic a large (or quasi-infinite system with a small number of particles), periodic boundary conditions (PBC) are usually applied in the calculation of potential energies. In spite of this artificial regulation, the simulation can still effectively reproduce the property of a homogeneous bulk system. The operation of PBC is shown in Figure 2-3 where the selected molecule (marked by a black circle) that is going to escape simulation box will re-enter the box from an opposite side. However, for systems where long (or infinite) wavelength fluctuations are of major importance, PBC will still be problematic since the largest fluctuations are of the size of the computational box. When applying PBC, each molecule have a number of corresponding periodic image molecules (see Figure 2-3), but only the image molecule that is the closest to the given molecule will be taken into account when calculating the potentials. Depending on the geometry of the real system simulated, PBC can be used along a single dimension or along several dimensions.

In most of the cases, the interaction potential function between particles is continuous with distance. For short-range interactions, such as the Lennard-Jones (LJ) potential [5], the total potential energy of a given particle is dominated by the interactions with its nearest particles, and thus can be truncated at a reasonable distance to save the computation time. The error introduced by ignoring the potential energy at larger distances can be regulated to small enough values by choosing a cutoff radius sufficiently large, a value above 2.5σ (where σ is the radius of soft core in the LJ potential) being generally used. Often used methods to truncate the potential can be: a) simple truncation; b) truncation and shift, where the potential is subtracted its value at cutoff radius so that the potential can continuously transit to zero at cutoff radius; and c) minimum image convention in which the interactions with the nearest image (see Figure 2-3) of all the particles in the simulation box is calculated. To avoid artifacts introduced by the

previously mentioned scheme, the truncated potential can be, to some extent, corrected by adding a tail contribution calculated analytically. However, for long-range interactions, such as Coulomb interaction, their potential function will not converge at a short distance, thus evaluating the long-range interactions by a simply truncation at cutoff radius will lead to a large error. There are several means to solve this problem. The simplest one is to introduce a relative dielectric constant for the interactions beyond the cutoff radius where the medium is uniform and has a uniform dielectric constant, while the interactions within the cutoff radius are explicitly calculated. This is the reaction field method. Another approach is to split the potential into a short part rapidly converging in the direct space and a long-range part rapidly converging in the reciprocal part (Ewald method, see section below).

2.1.2 Ising model

Although the Ising model is a basic mathematical model to describe ferromagnetism, many other interesting models can be derived from it. The Ising model can be intuitively understood as describing the magnetic dipole moment of atomic spins by considering a set of interacting spin variables located on a fixed lattice, each spin assuming only two states (up or down, represented by +1 or -1). It can be carried out on one-, two-, or three-dimensional lattices or even on a lattice of dimension greater than three. First of all, the one-dimensional Ising model is considered.

One-dimensional Ising model

For an infinitely long one-dimensional lattice, as shown in Figure 2-4, or a lattice ring, the spins $\boldsymbol{s}_i$ on the lattice can only interact with their nearest neighbors with a coupling strength J. Meanwhile, there is an external field h influencing the orientation of the spins with $\boldsymbol{s}_i = +1$ if parallel to the field, $\boldsymbol{s}_i = -1$ if antiparallel to the field.

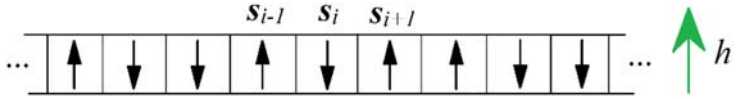

Figure 2-4. Scheme of an infinitely long one-dimensional Ising model in an external field h.

The Hamiltonian of the Ising model is

$$H(\{s_i\}) = -J\sum_{\langle i,j\rangle} s_i s_j - h\sum_i s_i, \tag{2.16}$$

where $\langle i,j\rangle$ indicates the sum over nearest neighbors ($j = i \pm 1$). In the canonical ensemble, the probability of finding a spin configuration $\{s_i\}$ is,

$$p(\{s_i\}) = \frac{1}{Z} exp[-\beta H(\{s_i\})], \tag{2.17}$$

where $Z = \sum_{\{s_i\}} exp[-\beta H(\{s_i\})]$ is the partition function. It is indicated by equation (2.17) that the spins prefer to take a homogeneous orientation for $J>0$ while an opposite orientation

with their neighbors for J<0. In addition, the spins will be interacting with the external field h and favored to take the orientation as h.

The free energy of the Ising model on a one-dimensional lattice with periodic boundary conditions can be solved analytically giving:

$$F(\beta,h)=\frac{-1}{\beta}ln\left\{e^{\beta J}cosh(\beta h)+\sqrt{e^{2\beta J}[sinh(\beta h)]^2+e^{-2\beta J}}\right\}. \quad (2.18)$$

When $h=0$ equation (2.18) can be simplified as,

$$F(\beta,0)=\frac{-1}{\beta}ln\left(e^{\beta J}+e^{-\beta J}\right) \quad (2.19)$$

It is found that no phase transition is observed in a one-dimensional Ising model.

Two-dimensional Ising model

The one-dimensional Ising model can be readily extended into a two-dimensional version, as displayed in Figure 2-5.

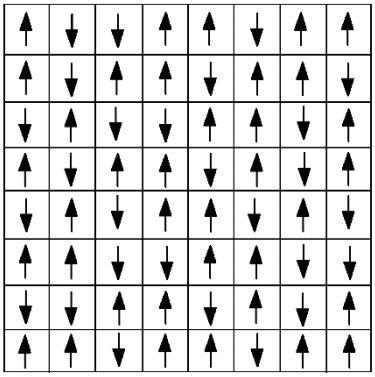

Figure 2-5. Scheme of a two-dimensional Ising model on a square lattice.

Comparing with the one-dimensional Ising model, the two-dimensional Ising model is much more complex. There was no analytical solution until 1944 when Onsager deduced an exact analytical expression for the free energy of the anisotropic Ising model on a square lattice in the absence of external field by using the transfer matrix method [6]. For an isotropic lattice, the expression of the free energy is given as

$$-\beta F=ln2+\frac{1}{8\pi^2}\int_0^{2\pi}d\theta_1\int_0^{2\pi}d\theta_2 ln\{cosh^2(2\beta J)-sinh(2\beta J)[cos(\theta_1)+cos(\theta_2)]\}. \quad (2.20)$$

With this equation, Onsager predicted a phase transition at a critical temperature T_c. This critical phase transition temperature is analytically given as

$$k_BT_c=\frac{2J}{ln(1+\sqrt{2})}\approx 2.269J. \quad (2.21)$$

By MC simulations, we can also observe the phase transition of 2D Ising model. As shown in Figure 2-6, the 2D (isotropic) Ising model demonstrates a clear phase transition at $k_BT=2.1-2.5J$.

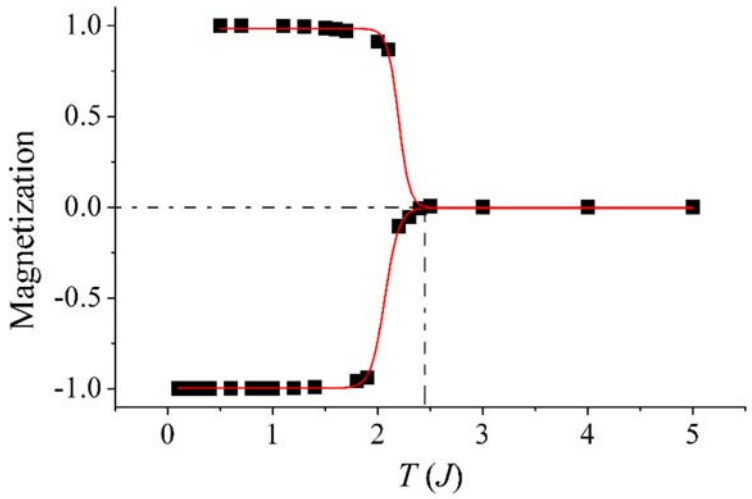

Figure 2-6. The magnetization of a 2D Ising model as a function of temperature without the influence of external field, calculated by MC simulations on a 25*25 lattice. The red line is used to guide the eyes. A phase transition was found at $k_BT = 2.1 - 2.5J$.

2.1.3 Lattice gas model

The lattice gas (LG) model is a derivative of the Ising model. The LG model describes the gas-liquid transition phenomenon and requires a variable number of particles. Thus the LG model calculation is usually carried out in the grand canonical ensemble. The basic idea of the LG model assumes that the dividing microscopic cells of the system are sufficiently small to contain at most one gas particles interacting only with their nearest neighbor particles with a strength λ, the kinetic energy of the particles being neglected. Each cell has only two possible states of occupancy, empty (represented by a value n=0) and filled by a gas particle (represented by a value n=1). A two-dimensional square LG model is shown in Figure 2-7.

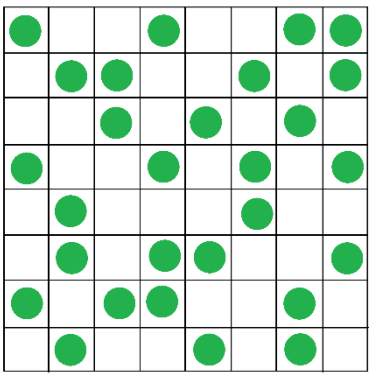

Figure 2-7. Schematic of a two-dimensional square lattice gas model.

The Hamiltonian of the square LG model in the grand canonical ensemble is

$$H - \mu N = -\lambda \sum_{\langle i,j \rangle} n_i n_j - \mu \sum_i n_i, \tag{2.22}$$

where n_i is the state of occupancy of the lattice with possible values of 0 and 1 and μ is the chemical potential of the reservoir. The Hamiltonian can be expressed as a function of the

Hamiltonian of the Ising model as

$$H - \mu N = -J\sum_{\langle i,j\rangle} \boldsymbol{s}_i \boldsymbol{s}_j - h\sum_i \boldsymbol{s}_i - \left(h - \frac{\gamma}{2}J\right) N_{\mathrm{L}} = H_I - \left(h - \frac{\gamma}{2}J\right) N_{\mathrm{L}},$$

$$J = \frac{h}{4}, \quad \boldsymbol{s}_i = 2n_i - 1, \quad h = \frac{\lambda}{4}\gamma + \frac{\mu}{2}, \tag{2.23}$$

where N_L is the total number of lattice sites and γ is the coordination number of each lattice (for a two-dimensional lattice, $\gamma = 4$). The partition function of the grand canonical ensemble Z_G can also be expressed as a function of the partition function Z_I of Ising model in the canonical ensemble as

$$Z_G = Z_I e^{\beta\left(\frac{\lambda}{8}\gamma + \frac{\mu}{2}\right)N_L}. \tag{2.24}$$

Correspondingly, the grand potential of the grand canonical ensemble can be written as

$$\Omega(\beta, \mu, N_L) = \frac{-1}{\beta} ln Z_G = F_I(\beta, \mu, N_L) - \left(\frac{\lambda}{8}\gamma + \frac{\mu}{2}\right) N_L, \tag{2.25}$$

where F_I is the Helmholtz free energy of the Ising model. The equilibrium density of gas particles $\rho = n/N_L$ is obtained by minimizing $\Omega(\beta, \mu, N_L)$. Unfortunately, ρ cannot be explicitly expressed but it is solution of the equation

$$\mu = \frac{1}{\beta} ln \frac{\rho}{1-\rho} - \lambda\gamma(1 - \rho). \tag{2.26}$$

Like for the Ising model, we also define a phase transition temperature T_c and a critical chemical potential μ_0 when $h = 0$, that is $\mu_0 = -2\lambda$. In particular, a two phase coexistence appears (reflected by $\rho = 0.5$) at critical temperature T_c and critical chemical potential μ_0. At this moment, T_c can be simplified as $k_B T_c = \lambda$. The LG model is able to describe both the liquid-gas transition and the crystallization transition, depending on the value of λ. When $\lambda > 0$, the particles interact with each other by an attractive potential and a liquid-gas transition can be reproduced, while a repulsive potential is present for $\lambda < 0$ and a crystallization transition exists.

The plot of equation (2.26) is displayed in Figure 2-8 (A). When the temperature is above T_c the chemical potential demonstrates a pairwise relation with ρ, but when the temperature is below T_c there are three corresponding ρ solutions for each chemical potential value, a thermodynamically stable solution, a thermodynamically metastable solution, and a thermodynamically instable solution. Figure 2-8 (B) shows a symmetrical behavior along $\rho = 0.5$ and reproduces a phase transition similar to the one observed in the 2D Ising model when the chemical potential is getting close to -2.0 $k_B T_c$.

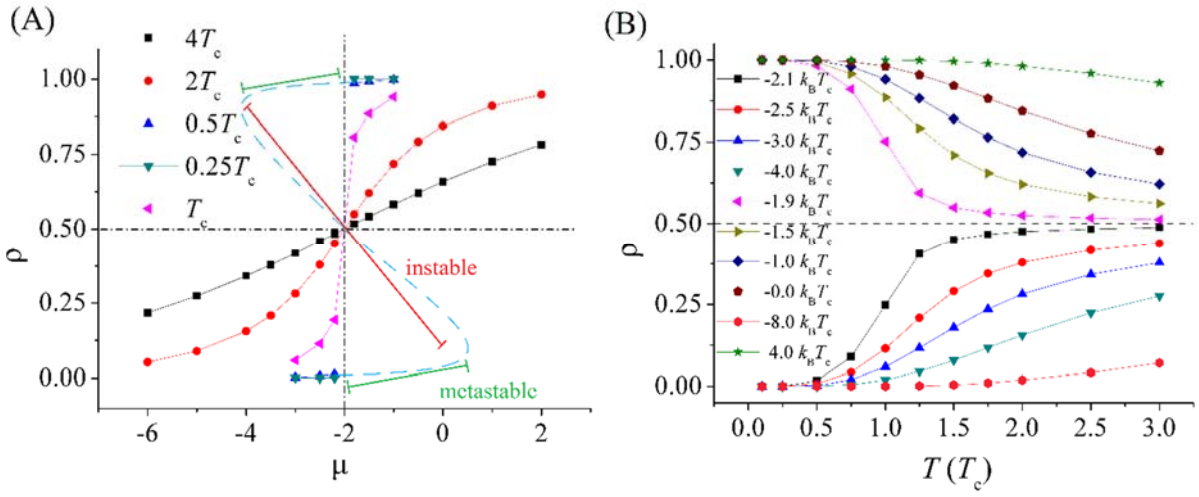

Figure 2-8. (A) Density of particles as a function of the chemical potential at different temperatures (4 T_c in black, 2 T_c in red, 0.5 T_c in blue, 0.25 T_c in cyan, and T_c in pink). All the curves intersect at $\rho = 0.5$ and $\mu = -2$. When temperature is below T_c, the curves exhibit thermodynamically instable (marked in red) and metastable (marked in green) branches. (B) Density of particles as a function of temperature at different chemical potentials.

2.2 Molecular dynamics simulations

Molecular dynamics (MD) simulation is a computational technique for studying the equilibrium properties of a multi-particle system by driving the physical movements of interacting particles according to the laws of classical mechanics or quantum mechanics. Nowadays, MD method has been developed into an indispensable and flourishing branch of numerical techniques. Many methods have been merged into MD technique leading to new sub-branches, such as *ab initio*, hybrid quantum-mechanical and molecular mechanics (QM/MM), coarse grained (CG) and multiscale simulations. Given that the investigations conducted in this thesis neglect the electronic (fast) degrees of freedom, we just focus in the following sections on the classical MD simulation method describing the nuclear (slow) degrees of freedom. Comparing with Monte Carlo simulation, MD simulations are able to monitor the dynamically evolution of the investigated system with time, which is particularly interesting for various physical reactions.

2.2.1 Basic ideas

In general, the systems studied by MD simulations consist of a very large number of particles, whose properties cannot be solved analytically. MD simulations use a numerical iteration method. The general procedure of performing a MD simulation can be simplified as follows: (1) prepare a system with the target investigated molecules under a well-designed environment and well-defined interaction potentials; (2) iteratively move the particles by solving Newton's equations of motion for short timesteps until the system reaches equilibrium; (3) compute the equilibrium properties using the statistical relationships between the macroscopic properties and the microscopic variables of the particles. A detailed procedure can be found in Figure 2-9.

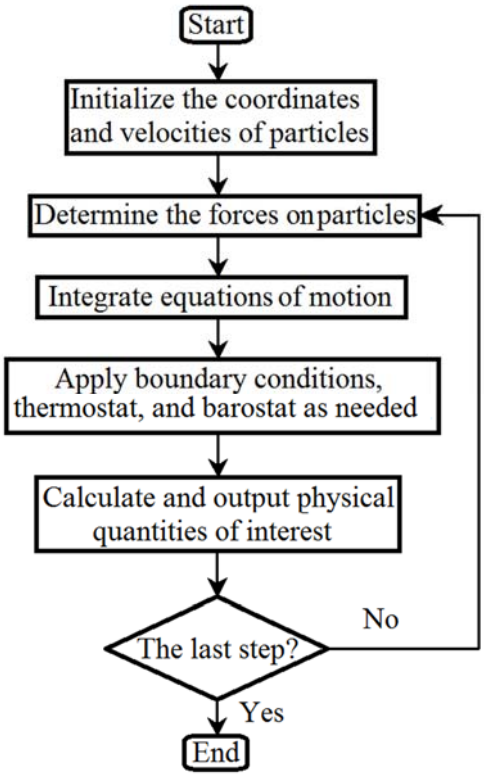

Figure 2-9. Flow chart of a basic MD simulation.

2.2.2 Integration algorithms

To drive the displacement of the particles, their coordinates and their velocities have to be calculated and updated at each MD step by integrating the Newton's equations of motion. Many integration algorithms have been proposed, which evaluate the trajectory and the velocities with different degrees of accuracy. For example, the Verlet algorithm [7] will yield a high accurate trajectory; the velocity-Verlet algorithm [8] is convenient to calculate velocity at a cost of a compromised accuracy of the trajectory; the Leap-frog algorithm [9] and the Beeman algorithm [10,11] will generate better estimates of the velocity, and other higher-order algorithms [12] will make it possible to use a longer time step without loss of accuracy but at the cost of more evaluations of the forces acting upon each particle at each timestep.

The following is the derivation of the Verlet algorithm. The coordinate of a particle can be expanded by Taylor expansion as

$$r(t+\Delta t) = r(t) + v(t)\Delta t + \frac{f(t)}{2m}\Delta t^2 + \frac{\Delta t^3}{3!}\dddot{r} + O(\Delta t^4),$$

$$r(t-\Delta t) = r(t) - v(t)\Delta t + \frac{f(t)}{2m}\Delta t^2 - \frac{\Delta t^3}{3!}\dddot{r} + O(\Delta t^4),$$

where r, v, and f denote the coordinate, velocity, and force of a particle, respectively. Summing these two equations, the trajectory integration equation of the Verlet algorithm is derived as

$$r(t+\Delta t) + r(t-\Delta t) = 2r(t) + \frac{f(t)}{2m}\Delta t^2 + O(\Delta t^4).$$

The velocity expression can also be derived as

$$v(t) = \frac{r(t+\Delta t)-r(t-\Delta t)}{2\Delta t} + O(\Delta t^2).$$

Although Verlet algorithm is able to generate accurate coordinates, the accuracy of the generated velocities are low, moreover, more memory is demanded to store the coordinates of the particles at three steps.

In the velocity-Verlet algorithm, the coordinate of a particle is simplified as

$$r(t + \Delta t) = r(t) + v(t)\Delta t + \frac{f(t)}{2m}\Delta t^2 + O(\Delta t^3),$$

where terms above $O(\Delta t^3)$ are neglected. Similarly, the coordinate at $t + 2\Delta t$ has expression of

$$r(t + 2\Delta t) = r(t + \Delta t) + v(t + \Delta t)\Delta t + \frac{f(t+\Delta t)}{2m}\Delta t^2 + O(\Delta t^3).$$

By subtraction we get

$$r(t + 2\Delta t) + r(t) = 2r(t + \Delta t) + [v(t + \Delta t) - v(t)]\Delta t + \frac{f(t+\Delta t)-f(t)}{2m}\Delta t^2 + O(\Delta t^3).$$

With the coordinate expression in Verlet algorithm, we can obtain the velocity at $t + \Delta t$,

$$v(t + \Delta t) = v(t) + \frac{f(t+\Delta t)+f(t)}{2m}\Delta t^2 + O(\Delta t^3).$$

The velocity-Verlet algorithm is able to provide the coordinates and the velocities of particles at a same step with a medium accuracy, which is convenient for simulations. Therefore, the choice of integration algorithm depends on the demanded accuracy of the trajectory and the velocity, and in some case, the compatibility to the algorithms in other respects (for example, velocity Verlet algorithm is the only compatible algorithm for RATTLE algorithm which is used to constrain the bond length, and Nose-Hoover thermostat).

2.2.3 Force fields

Force fields are developed to describe the interactions between specific particles with an expectation to reproduce the macroscopic properties of the investigated system. Although nowadays the bonded potentials in all-atom force fields are usually derived from ab initio/density functional theory (DFT) calculations, most of non-bonded interaction potentials are derived from various experimental data with many approximations, thus the all-atom force fields are overall empirical. Those potentials can be simply classified into non-bonded interactions and bonded interactions. The non-bonded potentials include for example the Lennard-Jones potential, one of the forms widely employed to describe the van der Waals forces, and the Coulomb potential, while bonded potentials involve the interactions of chemical bonds, bond angles, and bond dihedrals.

In MD simulations, the calculation of the forces on all the particles generally represents

the largest time. For the non-bonded potentials, these forces exhibit different convergences and are classified into short-range interactions and long-range interactions. To save CPU time, some techniques have been proposed to speed up the evaluation of both short-range and long-range interaction, such as Verlet list, cell list, and a combination of Verlet and cell lists. The principle of these techniques consists of using a list to store the serial number of particles that potentially interact with a given particle for each particle, so that the scale of interacting pairs is able to be efficiently reduced. Figure 2-10 is a schematic for Verlet list where a list is used to store the serial number of particles within r_v, a distance slightly larger than cutoff radius r_c, from particle i.

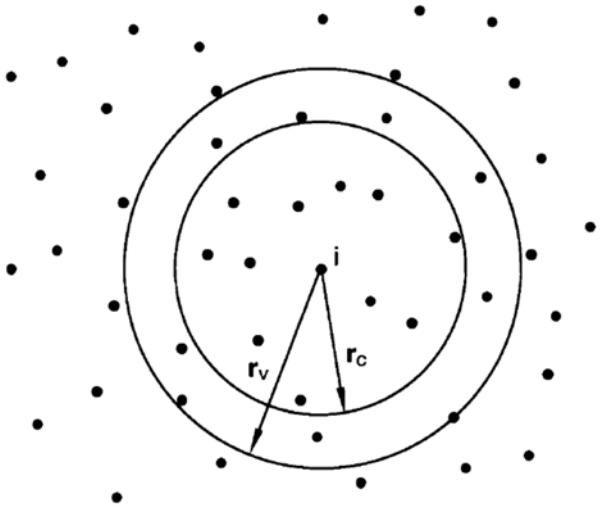

Figure 2-10. Schematic of Verlet list. Adapted from [12].

As mentioned in the previous section, the short-range interactions can be simply truncated at an appropriate distance. Using the accelerating algorithms, the computational effort for short-range interactions is low and scales as $O(N)$ with N the number of particles in the system. However, the computational effort for the long-range interactions scales more than $O(N^2)$ if calculating directly for a system with PBCs, indicating that it takes much longer time than it needs for the calculation of short-range interactions. Thus, a high efficient approach should be taken for such calculations. Reaction field (RF) method [13,14] is one option as mentioned in the above section, where a truncation scheme is also applied for the long-range interactions and a relative dielectric constant is introduced for the calculation beyond the cutoff radius. The RF method is a fast calculation scheme at the cost of a poor accuracy. Usually the long-range interactions are calculated with sophisticated algorithms of much better controlled accuracy (at the expense of a slightly decreased efficiency) such as Ewald summation [15], fast multipole methods [16-18], and particles-mesh-based (PME) techniques [19,20].

Ewald summation

The Ewald summation is a computational method to calculate long-range interactions in a periodic system. Here the Coulomb electrostatic interactions of a set of point charges are taken as an example. The long-range interactions cannot be evaluated by a simple cutoff scheme due to the bad convergence of their potential. If we could find a way to improve their convergence, we can also evaluate the long-range potential by the cutoff scheme.

Supposing that each point charge *i* is surrounded by a diffuse charge distribution of the opposite sign, such that the total charge of this cloud exactly cancels the charge of the point charge q_i. By this treatment the electrostatic potential due to this cloud can rapidly goes to 0 at large distances, thus the electrostatic potential of the charge clouds can be easily computed by the cutoff scheme. In order to obtain the contribution to electrostatic potential exclusively due to point charges, we need to remove the contribution due to the diffuse charge. To this end, we introduce another set of diffuse charges with charge sign as the point charges, denoted as compensating charges. The whole scheme of Ewald summation on treating the point charges is displayed in Figure 2-11 [12]. Overall, the electrostatic potential of a set of point charges can be equivalently expressed into two sections: the electrostatic potential due to a set of charge clouds (each cloud consists of a point charge and a screened charge cloud surrounding the point charge) minus the electrostatic potential due to a set of compensating charge clouds. The potential due to the charge clouds can be computed as short-range interactions, while that due to the compensating charge clouds has to be evaluated in reciprocal space.

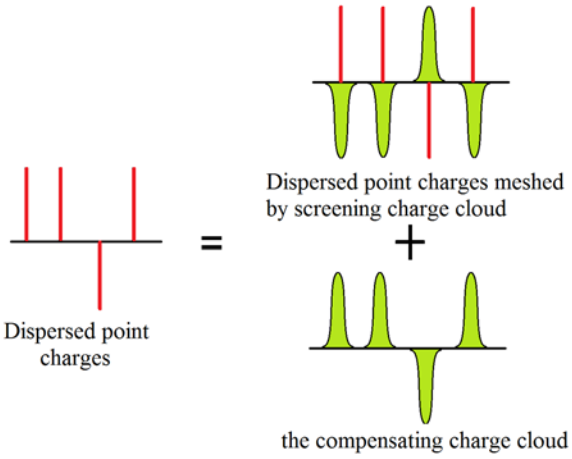

Figure 2-11. Schematic of the treatment of a set of point charges (in red lines) in Ewald. The point charges are considered as a set of screened charges (a set of charge clouds, each of them consists of a point charges and a diffuse charge layer (in green) surrounding it) minus the smoothly varying screening background (in green).

The compensating charge distribution surrounding the point charges with charge q_i can typically be taken as a Gaussian distribution with a width of $\sqrt{2/\alpha}$ and written as

$$\rho_{lr}(r) = -q_i(\alpha/\pi)^{\frac{3}{2}}exp(-\alpha r^2). \tag{2.27}$$

For a system with *N* point charges in a periodic box of side length *L* and volume *V*, equation (2.27) can be expressed in the reciprocal space by Fourier transform

$$\rho_{lr}(k) = \sum_{j=1}^{N} q_j exp\left(-ik \cdot r_j\right)exp(-k^2/4\alpha), \tag{2.28}$$

where $k = \frac{2\pi}{L}n$ with $n = \left(n_x, n_y, n_z\right)$ are the lattice vectors in the reciprocal space. By

inserting equation (2.28) into Poisson's equation, the electrostatic potential due to the compensating charge cloud is given as

$$\phi_{lr}(r) = \sum_{k\neq 0}\sum_{j=1}^{N}\frac{4\pi q_j}{k^2}exp\left[ik\cdot\left(r-r_j\right)\right]exp(-k^2/4\alpha), \tag{2.29}$$

where $k \neq 0$ is set because of the conditional convergence of the Ewald sum. The corresponding electrostatic potential energy is obtained with the correction for the spurious self-interaction between the compensating charge cloud and point charges and yields

$$\mathcal{U}_{\mathrm{lr}} = \frac{1}{2}\sum_i q_i\phi_{\mathrm{lr}}(\boldsymbol{r}_i) - \mathcal{U}_{\mathrm{self}} = \frac{1}{2V}\sum_{\boldsymbol{k}\neq 0}\frac{4\pi}{\boldsymbol{k}^2}\left|\sum_{i=1}^{N} q_i \exp(\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}_i)\right|^2 \exp(-\frac{k^2}{4\alpha}) - \left(\frac{\alpha}{\pi}\right)^{\frac{1}{2}}\sum_{i=1}^{N} q_i^2. \tag{2.30}$$

The short-range contribution is given by

$$U_{sr} = \frac{1}{2}\sum_{i\neq j}^{N}\frac{q_i q_j erfc(\sqrt{\alpha}r_{ij})}{r_{ij}} \tag{2.31}$$

and is calculated in the real space as any other short-ranged potentials.

The total electrostatic potential energy is rendered as $U_{Coul} = U_{lr} + U_{sr}$. On the calculation of U_{lr} in reciprocal space, the variable $k = \frac{2\pi}{L}n$ will also be truncated within a specific value, denoted as n_c. Thus, to carry out Ewald summation, three parameters have to be selected for a given accuracy ϵ, namely n_c the cutoff in reciprocal space, r_c the cutoff radius in real space, and α the distribution width of the Gaussian function. The relationships between them have been investigated in Ref [21-23], and are given as

$$\alpha^2 {r_c}^2 = \epsilon$$
$$n_c = \frac{2\epsilon}{r_c} = 2\alpha\sqrt{\epsilon}$$

With the optimized Ewald summation method, the CPU time required for the computation of long-range interactions is scaling as $O\left(N^{3/2}\right)$, which is still not efficient enough for large systems. Thus, based on the Ewald summation, alternative techniques with higher efficiency at the cost of a slightly decreased accuracy were developed. These techniques, namely the particle mesh Ewald (PME) method [19,20] or the particle-particle particle-mesh (PPPM) [9,24] method interpolate the charges onto a mesh and use the fast Fourier transforms to perform the calculations in the reciprocal space to achieve a $O\left(Nlog(N)\right)$ scaling behavior.

Bonded interactions

In order to describe the molecular conformations, some specific empirical potentials are employed between the atoms constituting the molecule. More specifically, these potentials are the bond potential describing the vibration between covalently bonded atoms, the angle potential describing the stretching of the angle between bonds, and the proper dihedral potential describing the torsion around the dihedral angle as well as the improper dihedral potential used to control the (non)-planarity of some groups of atoms. Figure 2-12 shows intuitive

representations.

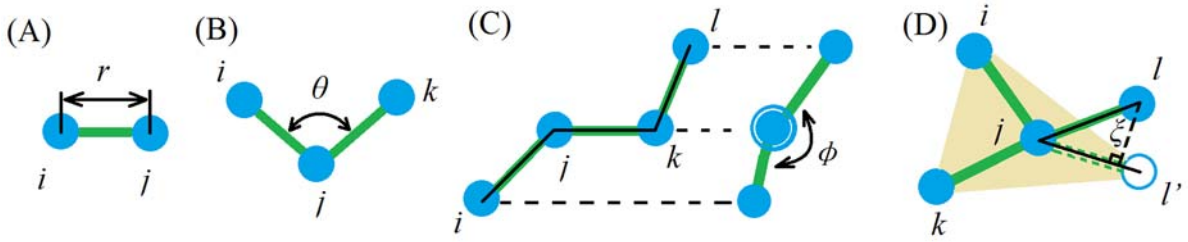

Figure 2-12. Schematics of bonded interactions: (A) bond length, (B) bond angle, (C) proper dihedral angle, and (D) improper dihedral angle.

However, there are numerous mathematical expressions describing the bonded interactions for various force fields. Here, we just list typical expressions as following:
The bond potential V_b describing the distance r between two covalently bonded atoms is often represented by a harmonic expression with a harmonic spring constant k_b and an equilibrium bond length r_b,

$$V_b = \frac{1}{2}k_b(r - r_b)^2. \tag{2.32}$$

The three-body angle potential V_a is often represented by a cosine harmonic potential with a force constant k_a and a bond angle θ with equilibrium value at θ_a,

$$V_a = \frac{1}{2}k_a(cos\theta - cos\theta_a)^2. \tag{2.33}$$

The four-body proper dihedral potential V_d can be expressed with the following form with a force constant $k_{d,n}$ with the dihedral angle ϕ assuming an equilibrium value ϕ_d

$$V_d = \sum_{n=1} k_{d,n}[1 + cos(n\phi - \phi_d)], \tag{2.34}$$

where *n* is the dihedral periodicity, used to distinguish the *cis-trans* isomerism of the dihedral angle. In order to maintain the structural and conformational stability of molecules, the simulations must be performed with a tiny timestep allowing to describe the rapid vibrations occurring within the molecules (normally 10-50 faster than the fastest intramolecular motions [12]). Therefore, to increase this timestep and achieve better efficiency, the multi-atom molecules are generally treated as rigid bodies with fixed bond lengths by constraint algorithms. Using these algorithms, MD simulations can be normally and stably conducted with a timestep in the order of the femtosecond. These algorithms include SHAKE [25], RATTLE [26], SETTLE [27], LINCS [28] etc. In some case, those algorithms will be used in a hybrid form to obtain a compromised optimization for the accuracy and efficiency, for instance LINCS for proteins while SETTLE for water molecules.

2.2.4 Thermostat and barostat

As mentioned above, the default ensemble of MD simulations is the microcanonical (NVE) [29]

ensemble where the exchange of particles, volume and energy is forbidden. However, most experiments are usually performed under constant temperature and pressure conditions or occasionally under constant temperature only. To mimic these experiments, MD simulations have to be conducted in NPT or NVT ensemble, and thus it is necessary to understand the principles of temperature and pressure regulations in MD simulations.

Temperature

From a classical statistical mechanical point of view, temperature is a function of the velocity of particles and is given by the relation

$$k_B T = \frac{\sum_{i=1}^{N} m_i {v_i}^2}{n_{free}}, \tag{2.35}$$

where m_i and v_i are the mass and velocity of particle *i*, n_{free} is the total number of degrees of freedom of the system with $n_{free} = 3N$ for a system constituted by *N* atoms. To ensure that the simulation is running at a constant temperature, we must find a way to control the velocity of the particles. The simplest way can be to rescale the velocity of each particle at each step to produce an absolutely constant-temperature simulation without any fluctuations on temperature, which is so-called the velocity scaling scheme. However, such a simple scheme does not reproduce a true constant-temperature ensemble and is problematic to measure equilibrium properties sensitive to fluctuations. Berendsen thermostat [30] is similar to the velocity scaling scheme, but the temperature is corrected in a slower way. Thus, it cannot reproduce a true canonical ensemble as well. In Berendsen thermostat it is assumed that the system is weakly coupled to a heat bath with a coupling strength or constant τ, the velocity scale factor being expressed as

$$\lambda = \frac{v_n}{v_o} = \sqrt{1 + \frac{\Delta t}{\tau}\left(\frac{T_e}{T_{rt}} - 1\right)}, \tag{2.36}$$

where v_n and v_o are the updated velocity and the velocity to be updated, respectively, T_e and T_{rt} are the expected temperature and the instantaneous temperature, respectively, and Δt the timestep of the MD simulation.

Another scheme is to contact the system to a large heat bath which imposes temperature by heat exchange. This constant-temperature scheme is able to render a true canonical ensemble and to generate a correct canonical velocity distribution for the particles. The most widely used canonical MD schemes include the Andersen thermostat [31] and the Langevin thermostat. In the Andersen thermostat, randomly selected particles in the system randomly exchange velocity with the heat bath by spurious stochastic collisions with a frequency of *v*, which, to some extent, is a kind of Monte Carlo operation. Similarly, Langevin thermostat uses an analogous scheme to control the temperature of the system. In Langevin dynamics (LD), collisions (or so-called "collisions and frictions") between investigating molecules and solvent molecules, are mimicked by introducing a variable to evaluate the viscous aspect of the solvent, thus it is convenient to control the velocities of molecules in LD simulations, that is LD can naturally be used as a thermostat.

We can also augment the degrees of freedom of the system by including variables describing the coupling to a thermostat and even to a barostat, these additional degrees of

freedom being handled through an extended Lagrangian. Nose-Hoover thermostat is a deterministic temperature controlling algorithm able to generate a canonical ensemble. It was firstly proposed by Nose [32] and later improved by Hoover [33]. The derivation of Nose-Hoover thermostat is based on the extended-Lagrangian formulation, using a Lagrangian that contains extra and artificial coordinates and velocities for heat bath. By the principle of the Liouville theorem and the conservation nature of the Hamiltonian, the partition function of the non-Hamiltonian system can be obtained. The derivation of Nose-Hoover algorithm is sophisticated, and more detailed explanations can be found in [32,33]. Martyna et al. [34] introduced Nose-Hoover chains to extend the Nose-Hoover algorithm to compatible multi-thermostat, allowing the temperature of the system to be more smoothly controlled and meanwhile, leading to more flexible and convenient ways to implement the algorithm.

Pressure

The expression of pressure in statistical mechanics is

$$P = \frac{2}{3V}\left(E_{kin} + \frac{1}{2}\sum_{i=1}^{N} r_i \cdot f_i\right), \tag{2.37}$$

where E_{kin} is the kinetic energy of the system. Equation (2.37) indicates that pressure is a function of the volume of the system. A constant-pressure MD simulation can be achieved by regulating the volume of the system. For the simplest case, the system can be considered, again, coupled with a "pressure bath". Based on this idea, the Berendsen algorithm can also be extended to the control of pressure by simply scaling the coordinates of molecules and box vectors every step. Unfortunately again such an algorithm cannot generate a true NPT ensemble. Nose-Hoover algorithm can also be extended to constant-pressure MD simulations with the improvement made by Martyna, Tuckerman, Tobias and Klein (MTTK) [34,35]. Another pressure control scheme is the Parrinello-Rahman extended-ensemble algorithm [36] where, like in the Nose-Hoover algorithm, some additional degrees of freedom in the equations of motion are introduced in a more complex way. More details about the Parrinello-Rahman algorithm can be found in reference [36]. Both the Nose-Hoover algorithm and the Parrinello-Rahman algorithm yield a correct NPT ensemble, but only Berendsen and the Parrinello-Rahman pressure control algorithm have a broader compatibility with any temperature control algorithms mentioned above. Moreover, the Parrinello-Rahman algorithm can easily handle anisotropic box deformations as well as non-orthorhombic boxes. Those advantages make the Parrinello-Rahman algorithm one of the most widely used pressure control algorithm.

2.2.5 Free energy calculation

For some aggregation reactions, the binding energy between the particles is an important parameter. We can of course estimate this parameter by experimental approaches, but it is not accurate. A relatively precise scheme to obtain this parameter is to rely on its estimation through molecular simulations. The binding energy can be readily extracted from the potential of mean force (PMF) which represents the free energy changes along the direction of the intermolecular separation. In MC or MD simulations, the PMF between molecules can be calculated by using

the umbrella sampling method.

Umbrella sampling is a method to improve the sampling of a system. Due to the energy landscape of the system, a uniform sampling of the configuration space is difficult to achieve. Most of the sampling methods generally concentrate in regions with a low energy or around a local minimum, while the sampling of the states of high energy is poor, or even entirely unsampled because of the low probability of overcoming the potential energy barriers. In umbrella sampling, an artificial biasing potential $V(r^N)$ is introduced, playing a role equivalent to a weighting function $w(r^N)$:

$$V(r^N) = -k_B T ln w(r^N).$$

Then the probability to find the system in a configuration r^N is given by

$$\pi(r^N) = \frac{w(r^N) exp[-\beta U(r^N)]}{\int w(r'^N) exp[-\beta U(r'^N)] dr'^N},$$

where U is the potential energy of the system, and the weighting function $w(r^N)$ will take high values at poorly sampled configurations and low values at fully sampled configurations, so that the presence of energy barriers can be compensated and a uniform probability distribution along the reaction coordinate can be obtained. As a result, a thermodynamic property *A* can be obtained in the canonical-ensemble by applying the formula:

$$\langle A \rangle = \frac{\langle A/w \rangle_\pi}{\langle 1/w \rangle_\pi}$$

with $\langle \cdots \rangle_\pi$ indicating an average over the probability distribution proportional to π.

If the biasing potential is strictly a function of a reaction coordinate, then the (unbiased) difference free energy profile ΔF_0 along the reaction coordinate can be calculated by subtracting the biasing potential V from the biased difference free energy profile ΔF_π.

$$\Delta F_0 = \Delta F_\pi - V$$

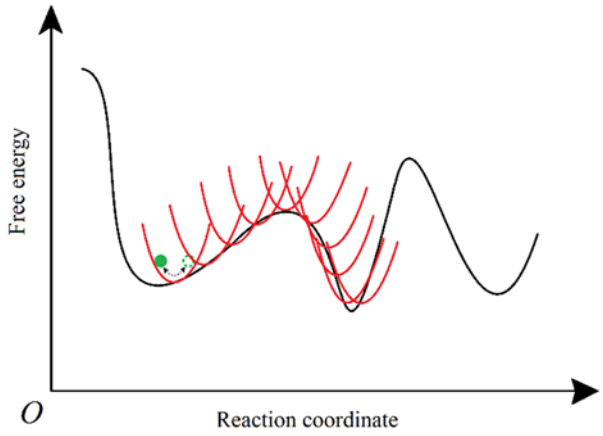

Figure 2-13. Schematic of the umbrella sampling method. Each umbrella allows to sample the configurations located around the equilibrium position of the biased potential (in red). A series of umbrella samplings are conducted to obtain the free energy difference along a given reaction coordinate.

In order to obtain the free energy difference along a given reaction coordinate, a series of umbrella sampling simulations are needed, as shown in Figure 2-13. The generated data can be subsequently analyzed using the weighted histogram analysis method (WHAM) [37] or its generalization [38]. WHAM can be derived using the Maximum likelihood method under the principle that if you have a discrete number of states, you can create a histogram with discrete bins along whatever reaction coordinate you have selected that provide you a relative probability of observing the states of interest.

2.2.6 Simulations at different scales

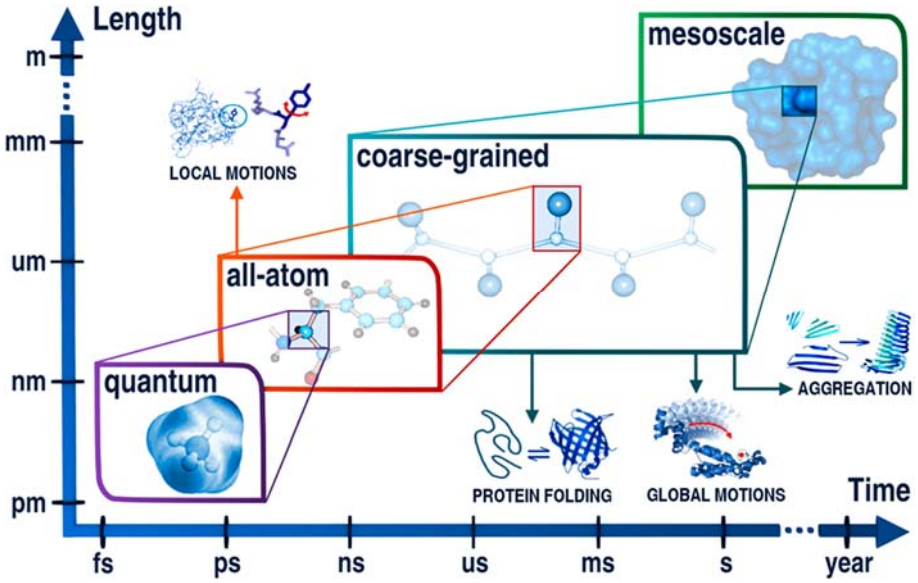

Figure 2-14. Application ranges of computational simulations at different scales. Adapted from [39].

For most of biomacromolecules, their interactions can be comprehensively described by all-atom force fields. Due to the limits in computational resource and in tolerable computation time, all-atom MD simulations are not able to reproduce the physical phenomena requiring a long timescale (above the order of second, for example) or too large a length scale to occur. In spite of the assistance of other techniques, such as replica exchange molecular dynamics [40], Markov multiple-state models [41], the maximum simulation time is normally less than 1 ms. Therefore, before starting MD simulations, an expected completion time should be estimated. The typical simulation time and length scales suitable to simulations operating at different resolutions are presented in Figure 2-14. From high to low resolution, MD simulations can be carried out at a quantum, an all-atom, a coarse-grained, and a mescoscale level, for typically picoseconds, nanoseconds, microseconds, and seconds, respectively.

The all-atom MD simulations on the self-assembly of virus capsid presented in the following chapters of this thesis are limited because a huge size of the computational box (above 100 nm) and a long simulation time (at least on a millisecond scale) are required for the assembly process to take place. In general, all-atom MD simulations were applied to investigate the properties of part of a virus capsid, for instance a protein subunit [42] or an oligomer of

capsid protein [43], and in some case the phase transition [44] or the properties of a completed capsid [45-48]. The understanding of details of the self-assembly process at an atomic-scale level can be a significant complement to experimental observations.

Due to the aforementioned limitations of all-atom simulations some complementary models have been devised. In coarse-grained (CG) models some functional groups of a macromolecule are represented by CG particles. The CG method accelerates the simulations in two manners: first of all, CG methods reduce the number of particles, decreasing the complexity of computations; secondly, the smoother CG potentials decrease the energy barrier and shorten the time needed for the equilibration of the simulated system. Since the pioneering work by Levitt and Warshel [49,50], a large diversity of CG models have been proposed. Most of them are dedicated to specific molecules [51], for example specific proteins and their CG potentials will be questionable if transferred to other types of molecules. However some of the CG models have a transferable capability for specific classes of molecules, such as AWSEM-MD [52], PaLaCe [53], and PACSAB [54] for proteins; and SimRNA [55] and many other CG models for nucleic acids. Some CG potentials possess the transferability across different classes of molecules, such as CafeMol [56], OPEP [57], SIRAH [58] for protein and DNA; MARTINI [59-63] for membrane lipids, proteins, and a large set of small organic molecules.

In general, the CG force fields describing these models can be obtained by three approaches [39]. Firstly, the CG force fields can be constructed by assembling a set of classical potentials in atomic force fields, called physics-based force fields. The parameters in these force fields can be derived from some experimental data, using several schemes, including the force matching approach [64], the iterative Boltzmann inversion approach [65] and the conditional reversible work approach [66]. A remarkable example of physics-based force fields is the MARTINI CG force field, whose parameters were extracted by reproducing the partitioning of free energies between polar and apolar phases for a large number of chemical compounds [59]. The other two schemes are structure-based force fields and knowledge-based statistical force fields, both of which are based on the molecular structures determined in experiments. Structure-based models, like the Go-type models [67,68], employ a specific force field approximation where only native-like interaction patterns are taken into account, while knowledge-based statistical force fields [69,70] rely on the statistical analysis of the conformational features and atomic packing of protein structures.

MARTINI coarse-grained method

As mentioned above, the MARTINI force field is a highly transferable CG model and is capable of simulating lipids, proteins, and DNA. It was developed by the groups of Marrink and Tieleman originally for the simulations of membrane lipids in 2006 [59], and extended to include proteins one year later [61], DNA in 2015 [62] and RNA in 2017 [63]. Thus far, the model includes also other molecules, such as various polymers [71-75], fullerene molecules [76], and hydrocarbon molecules [77,78]. The parameters in the force fields were adjusted by extensive calibration of the non-bonded interactions of the chemical building blocks against experimental data, in particular thermodynamic data such as oil/water partitioning coefficients. Due to its physics-based force field nature, the MARTINI CG model demonstrates a high compatibility to MD simulation packages, including GROMACS, Amber, and NAMD.

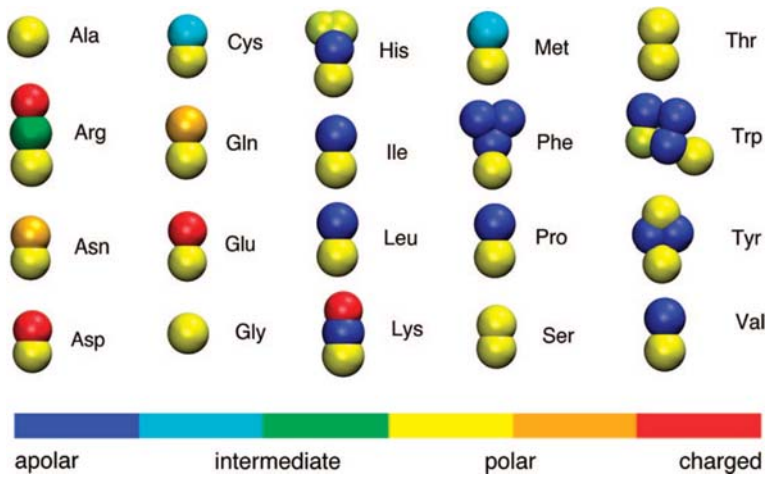

Figure 2-15. The mapping schemes of 20 peptides in the Martini CG model. The CG particles are colored according to their main types as indicated by the color band. Adapted from [63]

Mapping atomic structures onto CG particles and going backward is very convenient in Martini CG model with the assistance of embedded tools [79]. Figure 2-15 shows the CG schemes used for the 20 amino acids, each of which is represented by one to five CG particles. Overall, the Martini CG model takes a four-to-one mapping scheme, namely four heavy atoms are mapped onto a single CG particle. For solvent molecules, four water molecules are also represented by a single CG particle. The single ions and their first hydration shell are combined together and mapped onto a single CG particle.

Those CG particles interact through a potential consisting of four terms [61], namely a shifted Lennard-Jones potential, a shifted Coulomb electrostatic potential, a harmonic bonded potential and a weak harmonic angle potential, written as

$$U_{\mathrm{M}} = 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r}\right)^{12} - \left(\frac{\sigma_{ij}}{r}\right)^{6}\right] + \frac{q_i q_j}{4\pi\epsilon_0\epsilon_r r} + \frac{k_{\mathrm{b}}}{2}(r - r_{\mathrm{b}})^2 + \frac{k_{\mathrm{a}}}{2}[\cos(\theta) - \cos(\theta_{\mathrm{a}})]^2 + k_{\mathrm{d}}\left[1 + \cos\left(n\phi - \phi_{\mathrm{d}}\right)\right] + k_{\mathrm{id}}\left(\psi - \psi_{\mathrm{id}}\right)^2.$$

The default value for the parameter σ of the Lennard-Jones potential is 0.47 nm except for the interaction pair forming a hydrogen bond or located at ring structures, such as benzene ring. The default scheme for the calculation of the Lennard-Jones interactions is to truncate them at r = 0.9 nm and shift them from 0.9 nm to 1.2 nm to have them going smoothly to zero. Depending on the desired computation speed or accuracy, the calculation of the electrostatic interactions can be performed in several ways: using a simple cut-off value, the reaction-field theory, or Ewald-based approaches to take into account their long-range nature.

To identify the nature of different groups, the CG particles are classified into four main types [59]: polar (denoted as P), nonpolar (N), apolar (C), and charged (Q) (Figure 2-15). Each main type has four or five subtypes to indicate either the degree of polarity for P and C CG particle types (denoted by the integers 1 to 5), or the hydrogen-binding capabilities for N and Q CG particle types (with "0" for no hydrogen-binding capabilities, "a" for groups acting as hydrogen bond acceptor, "d" for groups acting as hydrogen bond donor, and "da" for groups with both donor and acceptor options). The interactions between the CG particles belonging to

different subtypes demonstrate diverse interaction strengths, reflected by the parameters of the Lennard-Jones potential. The interaction strength is classified into 10 levels, from the weakest IX level to the strongest O level, and the interaction matrix between them is presented in Table 2-1.

Table 2-1. The interaction matrix of CG particles in the Martini CG model

		Q				P					N				C				
	sub	da	d	a	0	5	4	3	2	1	da	d	a	0	5	4	3	2	1
Q	da	O	O	O	II	O	O	O	I	I	I	I	I	IV	V	VI	VII	IX	IX
	d	O	I	O	II	O	O	O	I	I	I	III	I	IV	V	VI	VII	IX	IX
	a	O	O	I	II	O	O	O	I	I	I	I	III	IV	V	VI	VII	IX	IX
	0	II	II	II	IV	I	O	I	II	III	III	III	III	IV	V	VI	VII	IX	IX
P	5	O	O	O	I	O	O	O	O	O	I	I	I	IV	V	VI	VI	VII	VIII
	4	O	O	O	O	O	I	I	II	II	III	III	III	IV	V	VI	VI	VII	VIII
	3	O	O	O	I	O	I	I	II	II	II	II	II	IV	IV	V	V	VI	VII
	2	I	I	I	II	O	II	II	II	II	II	II	II	III	IV	IV	V	VI	VII
	1	I	I	I	III	O	II	II	II	II	II	II	II	III	IV	IV	IV	V	VI
N	da	I	I	I	III	I	III	II	II	II	II	II	II	IV	IV	V	VI	VI	VI
	d	I	III	I	III	I	III	II	II	II	II	III	II	IV	IV	V	VI	VI	VI
	a	I	I	III	III	I	III	II	II	II	II	II	III	IV	IV	V	VI	VI	VI
	0	IV	IV	IV	IV	IV	IV	IV	III	III	IV	IV	IV	IV	IV	IV	IV	V	VI
C	5	V	V	V	V	V	V	IV	IV	IV	IV	IV	IV	IV	IV	IV	IV	V	V
	4	VI	VI	VI	VI	VI	VI	V	IV	IV	V	V	V	IV	IV	IV	IV	V	V
	3	VII	VII	VII	VII	VI	VI	V	V	IV	VI	VI	VI	IV	IV	IV	IV	IV	IV
	2	IX	IX	IX	IX	VII	VII	VI	VI	V	VI	VI	VI	V	V	V	IV	IV	IV
	1	IX	IX	IX	IX	VIII	VIII	VII	VII	VI	VI	VI	VI	VI	V	V	IV	IV	IV

Level of interaction	ϵ (kJ/mol)	σ (nm)
O	5.6	0.47
I	5.0	0.47
II	4.5	0.47
III	4.0	0.47
IV	3.5	0.47
V	3.1	0.47
VI	2.7	0.47
VII	2.3	0.47
VIII	2.0	0.47
IX	2.0	0.62

For atoms in ring structures, such as benzene and pyridine, a four-to-one mapping scheme is inadequate. Therefore, the ring atoms are mapped with a 2 or 3 to 1 scheme to ensure preservation of enough geometrical details. Correspondingly, the parameter σ of the Lennard-Jones potential is set to 0.43 nm and the interaction strength ϵ is scaled to 75% of its original value. In addition, an improper dihedral angle potential is applied on ring atoms to prevent out-of-plane distortions.

With the Martini CG model presented above, MD simulations can be carried out with a simulation timestep above 20 fs achieving a computational efficiency approximately three orders of magnitude faster than traditional all-atom simulations [59]. For many systems, such as lipid systems, solvent molecules represent a large fraction of the total amount of particles. Thus, a large computation time is spent on calculating the trajectory of such solvent molecules which, however, is not too important for some studied systems. Therefore, a Martini CG model has been also developed for lipid systems in implicit solvent, known as dry Martini [80].

Given that the Martini CG model was designed originally for membrane lipids, its application to simulate proteins presents many shortcomings. The structure and conformation of proteins are constrained by an elastic network to ensure their stability in the Martini CG model, meanwhile prohibiting the ability of the proteins to adjust their conformation. Thus, it is not advisable to implement the Martini CG model to study physical processes where the proteins undergo significant internal reorganizations. A polarizable water model has been also introduced to be able to describe biomolecular processes involving charged species moving between regions of high polarity, such as the water phase, and regions of lower polarity, such as the lipid membrane [81]. By combining the polarizable water model and an appropriate calculation algorithm for long-range interactions, the long-range effect can be better evaluated to some extent. Finally, the Martini CG model also provides a hybrid AA (all-atom)/CG scheme [82] for cases where a small part of the system needs to be treated with atomistic details.

Despite its shortcomings, the Martini CG model is still one of the most successful CG models ever constructed, thanks to its excellent transferability across a number of systems, its remarkable compatibility with various simulation packages and a good user support. It has been

extremely successful in the investigation of a wide-range of biological systems as well as systems such as polymers, carbohydrates, carbon-based nanoparticles. Recently, the Martini CG model also found applications to study viral capsids, including the deformations of capsids [83,84] and the interactions of capsid proteins with lipid layers [85-87].

2.2.7 Molecular dynamics simulation packages

There is not a unique general simulation package to cover the implementation of all the CG (as well as all-atom) models due to their wide diversity. Thus, most models are simulated either by specifically "home-made" codes or MD simulation packages developed originally for some all-atom models. Study of systems requiring the implementation of exotic models and (or) extremely specialized algorithms will have to be tackled by in-house developed codes, sometimes at the expense of performance. However for the vast majority of the investigations, very efficient massively parallel packages can be used. Several excellent free as well as commercial MD simulation packages are available to the community. For the simulations of biological systems, the best applicable free packages include GROMACS [88], NAMD [89], LAMMPS [90], while CHARMM [86], AMBER [91], are among the most popular commercial options.

Both CHARMM (Chemistry at Harvard Macromolecular Mechanics) and AMBER (Assisted Model Building with Energy Refinement) are without any doubt the two most widely used force field in simulations of biological systems. The two initial codes using these force fields have been built at early stages of the development of MD simulations (around 1970) and thus the codes (especially the parallel implementation) are not as efficient as those developed in the later stages.

Nanoscale Molecular Dynamics (NAMD) is developed by the theoretical and computational biophysics group, a famous MD simulation group, at the University of Illinois Urbana-Champaign. Because it was written with a parallel programming model, NAMD demonstrates brilliant performances for parallel computation and excellent scalable capability over both CPU processors and GPU processors [92]. Similarly, LAMMPS (Large-scale Atomic/Molecular Massively Parallel Simulator) is another efficient MD simulation package, in particular for simulations of homogeneous systems in a 3D orthorombic box. LAMMPS is not a package specific for biomacromolecules, and can handle a variety of force-fields to study for example inorganic molecules [93], metals [94], and a series of CG models due to its remarkable expandability. There are a significant number of MD simulations using CG force fields conducted via LAMMPS [95-97].

GROMACS (the abbreviation of GROningen Machine for Chemical Simulations) was firstly designed for simulations of proteins, but has been extended to other biological or organic molecules afterward, such as membrane lipids and polymers. Thanks to algorithmic and processor-specific optimizations, GROMACS is typically running 3-10 times faster than many other simulation programs. The embedded force fields in GROMACS including GROMOS [98], OPLS/AA [99], AMBER [100], and CHARMM [101] allowing its use in a large variety of all-atom simulations. All these typical force fields share a same expression for the non-bonded potentials, i.e. a Lennard-Jones potential plus a Coulomb potential, and various expressions for

the bonded potentials.

In addition, GROMACS is also compatible with some CG force fields that possess a similar form of the interaction potentials, such as the MARTINI CG force field . It is also possible to run a simulation with user-specified potential functions via GROMACS using look-up tables which store the values of the potentials as a function of the separation distances. Due to its versatility, GROMACS is the simulation package adopted in this thesis to perform all-atom simulations of virus capsid self-assembly.

2.3 Appendix

2.3.1 The assembled functions in GROMACS

Many macroscopic properties can be calculated by the functions implemented in GROMACS. There are several schemes provided by GROMACS to calculate the free energy of a system, including equilibrium techniques such as thermodynamic integration, and the potential of mean force using umbrella sampling and WHAM. Some non-equilibrium techniques are also present. The basic idea of steered molecular dynamics simulations is similar to atomic force microscopy experiments where the mechanical properties of molecules are studied by applying an external force. Such technique is able to evaluate either intramolecular free energy or intermolecular free energy (e.g. potential of mean force) by making use of the non-equilibrium relationship derived by Jarzynski and Crooks [102].

Sometimes, the simulated system cannot reach its global energy minimum because it is falling into kinetic traps. The simulated annealing [103] is typically a probabilistic technique to solve this problem via a global optimization at a higher temperature. In GROMACS, the simulated annealing technique can even allow multiple groups of atoms to be separately coupled to different reference temperature baths.

Replica exchange molecular dynamics (REMD) [40] is another efficient technique to avoid the kinetic traps caused by energy barriers. It involves simulating multiple replicas of the same system at different temperatures and randomly exchanging the complete state of two adjacent replicas at regular intervals by a Metropolis scheme. In GROMACS, temperature REMD has been generalized for various purposes, resulting in REMD in the isobaric-isothermal ensemble [104], Hamiltonian replica exchange, and Gibbs sampling replica exchange [105].

A hybrid quantum mechanical/molecular mechanics (QM/MM) scheme [106-108] has been developed for some systems, chemical reactions for instance, in which electrons can no longer be ignored and a quantum mechanical description is required.

Implicit solvent models are also available in GROMACS. For a large aqueous system, most of time is devoted to the calculation of the information about the water molecules surrounding the investigated molecules, which makes MD simulations incapable of dealing with large systems. An implicit solvent model is an available approach to address this problem. In GROMACS, implicit solvent calculations can be performed using the generalized Born-formalism with the Still [109], HCT [110], or OBC [111] model for the calculation of the Born radii.

2.3.2 Procedure for performing a MD simulation with GROMACS

It is convenient to employ GROMACS to carry out MD simulations. A set of command lines is used in GROMACS to control the procedure of MD simulations. The workflow to perform a MD simulation is shown in Figure 2-16, where the basic control command lines are indicated. A typical biological system will be prepared according to the following steps. First of all, we must convert the initial molecular structures (usually in the Protein Data Bank, pdb format) into topology files that are readable by GROMACS with *gmx pdb2gmx*, meanwhile selecting the desired force field and the water model. The generated topology files include the coordinates of each atom and the topology files (topolo.top) contain the connectivity information of the molecules. Next, we use the command *gmx editconf* to set the shape and the size of the simulation box. The shape of the box can be orthorhombic, triclinic, dodecahedronal, or octahedronal. The system is then solvated by filling up the simulation box with solvent molecules while preventing the occurrence of overlaps between atoms using the *gmx solvate* command. At the end of this step the topology file will be updated to include the information about the solvent molecules. The ions are subsequently added to the system by substituting some solvent molecules so that the system reaches charge neutralization and the desired salt concentration. It is readily carried out by using the command *gmx genion* in GROMACS. The topology informations of ions need to be added to the topology files as well. The initial state built using the previous mentioned steps needs to undergo an energy minimization to release the internal stress induced by the construction procedure. GROMACS provides three approaches - steepest descent, conjugate gradient, and L-BFGS [112], -, for energy minimization depending on the optimization quality that is expected. *gmx mdrun* is the main command line within GROMACS, allowing to perform different actions. We can use it to perform energy minimization, run a MD or a stochastic dynamics simulation, perform a test particle insertion or calculate some energy contributions. In general, the system is gently brought to the proper equilibrated conditions by performing a short NVT simulation followed by a short NPT run. Those simulations will generate the trajectory files in the format .trr or .xtc, containing the positions of each atoms for .xtc file and the positions, forces, and velocities for .trr file. After the equilibration under NVT and NPT conditions, the system undergoes a long production NPT simulation from which dynamical and equilibrium properties will be evaluated.

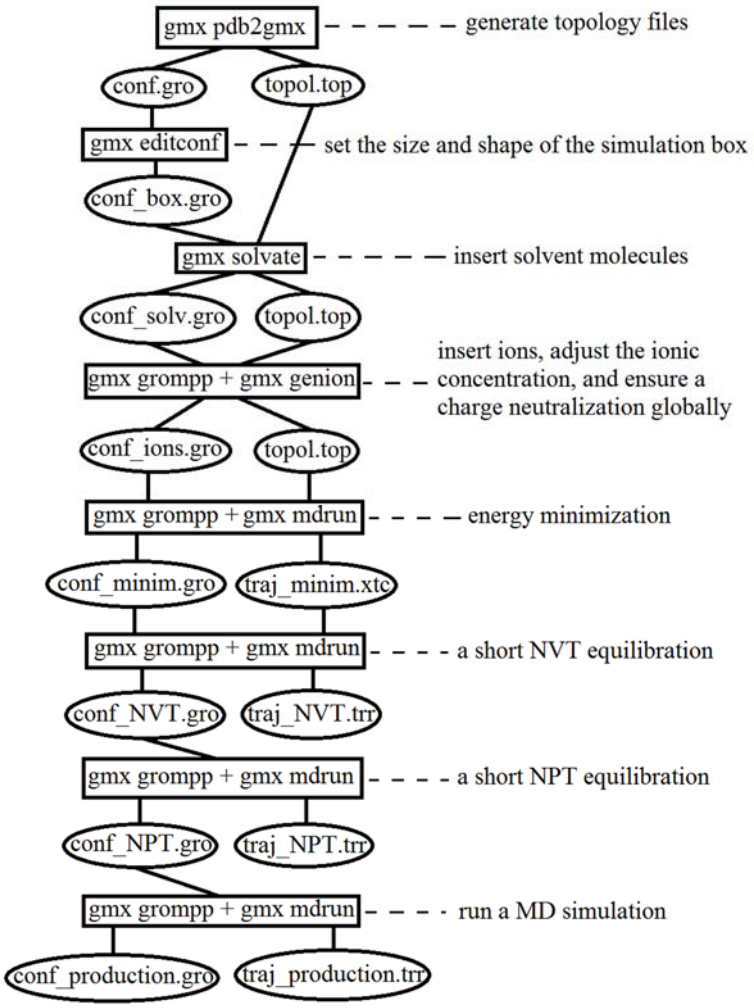

Figure 2-16. Schematic of the general workflow to perform a MD simulation with GROMACS.

Reference

[1] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, Journal of Chemical Physics **21**, 1087 (1953).
[2] W. W. Wood, Journal of Chemical Physics **48**, 415 (1968).
[3] I. R. McDonald, Molecular Physics **23**, 41 (1972).
[4] G. Norman and V. J. H. T. Filinov, **7**, 216 (1969).
[5] J. E. Jones, Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical **106**, 463 (1924).
[6] L. Onsager, Physical Review **65**, 117 (1944).
[7] L. Verlet, Physical Review **159**, 98 (1967).
[8] W. C. Swope, H. C. Andersen, P. H. Berens, and K. R. Wilson, Journal of Chemical Physics **76**, 637 (1982).
[9] R. W. Hockney and J. W. Eastwood, *Computer simulation using particles* (crc Press, 1988).
[10] D. Beeman, Journal of Computational Physics **20**, 130 (1976).
[11] P. Schofield, Computer physics communications **5**, 17 (1973).

[12] D. Frenkel and B. Smit, *Understanding molecular simulation: from algorithms to applications* (Elsevier, 2001), Vol. 1.
[13] J. A. Barker and R. O. Watts, Molecular Physics **26**, 789 (1973).
[14] R. O. Watts, Molecular Physics **28**, 1069 (1974).
[15] P. P. Ewald, Annalen der physik **369**, 253 (1921).
[16] A. W. Appel, Siam Journal on Scientific and Statistical Computing **6**, 85 (1985).
[17] J. Barnes and P. Hut, Nature **324**, 446 (1986).
[18] K. Esselink, Information Processing Letters **41**, 141 (1992).
[19] T. Darden, D. York, and L. Pedersen, Journal of Chemical Physics **98**, 10089 (1993).
[20] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, Journal of Chemical Physics **103**, 8577 (1995).
[21] H. G. Petersen, Journal of Chemical Physics **103**, 3668 (1995).
[22] D. Fincham, Molecular Simulation **13**, 1 (1994).
[23] J. Kolafa and J. W. Perram, Molecular Simulation **9**, 351 (1992).
[24] M. Deserno and C. Holm, The Journal of Chemical Physics **109**, 7694 (1998).
[25] J.-P. Ryckaert, G. Ciccotti, and H. J. J. J. o. C. P. Berendsen, **23**, 327 (1977).
[26] H. C. Andersen, Journal of Computational Physics **52**, 24 (1983).
[27] S. Miyamoto and P. A. Kollman, Journal of Computational Chemistry **13**, 952 (1992).
[28] B. Hess, H. Bekker, H. J. C. Berendsen, and J. Fraaije, Journal of Computational Chemistry **18**, 1463 (1997).
[29] J. J. Erpenbeck and W. W. Wood, in *Statistical Mechanics, Part B*1977).
[30] H. J. C. Berendsen, J. P. M. Postma, W. F. Vangunsteren, A. Dinola, and J. R. Haak, Journal of Chemical Physics **81**, 3684 (1984).
[31] H. C. Andersen, Journal of Chemical Physics **72**, 2384 (1980).
[32] S. Nose, Journal of Chemical Physics **81**, 511 (1984).
[33] W. G. Hoover, Physical Review A **31**, 1695 (1985).
[34] G. J. Martyna, M. L. Klein, and M. Tuckerman, Journal of Chemical Physics **97**, 2635 (1992).
[35] M. Tuckerman, B. J. Berne, and G. J. Martyna, Journal of Chemical Physics **97**, 1990 (1992).
[36] M. Parrinello and A. Rahman, Journal of Applied Physics **52**, 7182 (1981).
[37] S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg, Journal of Computational Chemistry **13**, 1011 (1992).
[38] C. Bartels, Chemical Physics Letters **331**, 446 (2000).
[39] S. Kmiecik, D. Gront, M. Kolinski, L. Wieteska, A. E. Dawid, and A. Kolinski, Chemical Reviews **116**, 7898 (2016).
[40] Y. Sugita and Y. Okamoto, Chemical Physics Letters **314**, 141 (1999).
[41] G. R. Bowman, X. H. Huang, and V. S. Pande, Methods **49**, 197 (2009).
[42] Z. Antal, J. Szoverfi, and S. N. Fejer, Journal of Chemical Information and Modeling **57**, 910 (2017).
[43] A. J. Pak *et al.*, Proceedings of the National Academy of Sciences of the United States of America **114**, E10056 (2017).
[44] Y. L. Miao, J. E. Johnson, and P. J. Ortoleva, Journal of Physical Chemistry B **114**, 11181 (2010).

[45] Y. Andoh, N. Yoshii, A. Yamada, K. Fujimoto, H. Kojima, K. Mizutani, A. Nakagawa, A. Nomoto, and S. Okazaki, Journal of Chemical Physics **141**, 165101 (2014).
[46] J. R. Perilla, J. A. Hadden, B. C. Goh, C. G. Mayne, and K. Schulten, Journal of Physical Chemistry Letters **7**, 1836 (2016).
[47] J. R. Perilla and K. Schulten, Nature Communications **8**, 15959 (2017).
[48] E. Tarasova, V. Farafonov, R. Khayat, N. Okimoto, T. S. Komatsu, M. Taiji, and D. Nerukh, Journal of Physical Chemistry Letters **8**, 779 (2017).
[49] M. Levitt and A. Warshel, Nature **253**, 694 (1975).
[50] A. Warshel and M. J. J. o. m. b. Levitt, **103**, 227 (1976).
[51] J. M. A. Grime, J. F. Dama, B. K. Ganser-Pornillos, C. L. Woodward, G. J. Jensen, M. Yeager, and G. A. Voth, Nature Communications **7**, 11568 (2016).
[52] A. Davtyan, N. P. Schafer, W. H. Zheng, C. Clementi, P. G. Wolynes, and G. A. Papoian, Journal of Physical Chemistry B **116**, 8494 (2012).
[53] M. Pasi, R. Lavery, and N. Ceres, Journal of Chemical Theory and Computation **9**, 785 (2013).
[54] A. Emperador, P. Sfriso, M. A. Villarreal, J. L. Gelpi, and M. Orozco, Journal of Chemical Theory and Computation **11**, 5929 (2015).
[55] M. J. Boniecki, G. Lach, W. K. Dawson, K. Tomala, P. Lukasz, T. Soltysinski, K. M. Rother, and J. M. Bujnicki, Nucleic Acids Research **44**, e63 (2016).
[56] H. Kenzaki, N. Koga, N. Hori, R. Kanada, W. F. Li, K. Okazaki, X. Q. Yao, and S. Takada, Journal of Chemical Theory and Computation **7**, 1979 (2011).
[57] F. Sterpone *et al.*, Chemical Society Reviews **43**, 4871 (2014).
[58] L. Darre, M. R. Machado, A. F. Brandner, H. C. Gonzalez, S. Ferreira, and S. Pantano, Journal of Chemical Theory and Computation **11**, 723 (2015).
[59] S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. de Vries, Journal of Physical Chemistry B **111**, 7812 (2007).
[60] S. J. Marrink and D. P. Tieleman, Chemical Society Reviews **42**, 6801 (2013).
[61] L. Monticelli, S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman, and S. J. Marrink, Journal of Chemical Theory and Computation **4**, 819 (2008).
[62] J. J. Uusitalo, H. I. Ingolfsson, P. Akhshi, D. P. Tieleman, and S. J. Marrink, Journal of Chemical Theory and Computation **11**, 3932 (2015).
[63] J. J. Uusitalo, H. I. Ingolfsson, S. J. Marrink, and I. Faustino, Biophysical Journal **113**, 246 (2017).
[64] F. Ercolessi and J. B. Adams, EPL **26**, 583 (1994).
[65] W. Schommers, Physical Review A **28**, 3599 (1983).
[66] E. Brini and N. F. A. Van der Vegt, The Journal of chemical physics **137**, 154113 (2012).
[67] N. Go and H. Taketomi, International Journal of Peptide and Protein Research **13**, 447 (1979).
[68] H. Taketomi, Y. Ueda, and N. Go, International Journal of Peptide and Protein Research **7**, 445 (1975).
[69] R. Samudrala and J. Moult, Journal of Molecular Biology **275**, 895 (1998).
[70] S. Tanaka and H. A. Scheraga, Macromolecules **9**, 945 (1976).
[71] R. Alessandri, J. J. Uusitalo, A. H. de Vries, R. W. A. Havenith, and S. J. Marrink, Journal of the American Chemical Society **139**, 3697 (2017).

[72] F. Grunewald, G. Rossi, A. H. de Vries, S. J. Marrink, and L. Monticelli, Journal of Physical Chemistry B **122**, 7436 (2018).
[73] E. Panizon, D. Bochicchio, L. Monticelli, and G. Rossi, Journal of Physical Chemistry B **119**, 8209 (2015).
[74] G. Rossi, L. Monticelli, S. R. Puisto, I. Vattulainen, and T. Ala-Nissila, Soft Matter **7**, 698 (2011).
[75] M. Vogele, C. Holm, and J. Smiatek, Journal of Chemical Physics **143**, 243151 (2015).
[76] J. Wong-Ekkabut, S. Baoukina, W. Triampo, I. M. Tang, D. P. Tieleman, and L. Monticelli, Nature Nanotechnology **3**, 363 (2008).
[77] C. A. Lopez, A. H. de Vries, and S. J. Marrink, Scientific Reports **3**, 2071 (2013).
[78] C. A. Lopez, A. J. Rzepiela, A. H. de Vries, L. Dijkhuizen, P. H. Hunenberger, and S. J. Marrink, Journal of Chemical Theory and Computation **5**, 3195 (2009).
[79] T. A. Wassenaar, K. Pluhackova, R. A. Bockmann, S. J. Marrink, and D. P. Tieleman, Journal of Chemical Theory and Computation **10**, 676 (2014).
[80] C. Arnarez, J. J. Uusitalo, M. F. Masman, H. I. Ingolfsson, D. H. de Jong, M. N. Melo, X. Periole, A. H. de Vries, and S. J. Marrink, Journal of Chemical Theory and Computation **11**, 260 (2015).
[81] S. O. Yesylevskyy, L. V. Schafer, D. Sengupta, and S. J. Marrink, Plos Computational Biology **6**, e1000810 (2010).
[82] A. J. Rzepiela, M. Louhivuori, C. Peter, and S. J. Marrink, Physical Chemistry Chemical Physics **13**, 10437 (2011).
[83] O. Kononova, J. Snijder, Y. Kholodov, K. A. Marx, G. J. L. Wuite, W. H. Roos, and V. Barsegov, Plos Computational Biology **12**, e1004729 (2016).
[84] V. Krishnamani, C. Globisch, C. Peter, and M. Deserno, European Physical Journal-Special Topics **225**, 1757 (2016).
[85] J. K. Marzinek, N. Bag, R. G. Huber, D. A. Holdbrook, T. Wohland, C. S. Verma, and P. J. Bond, Journal of Chemical Theory and Computation **14**, 3920 (2018).
[86] M. A. M. Yusoff, A. A. A. Hamid, N. M. Bunori, and K. B. Abd Halim, Journal of Molecular Graphics & Modelling **82**, 137 (2018).
[87] J. B. Gc, B. S. Gerstman, R. V. Stahelin, and P. P. Chapagain, Physical Chemistry Chemical Physics **18**, 28409 (2016).
[88] D. Van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, Journal of Computational Chemistry **26**, 1701 (2005).
[89] J. C. Phillips *et al.*, Journal of Computational Chemistry **26**, 1781 (2005).
[90] S. Plimpton, Journal of Computational Physics **117**, 1 (1995).
[91] D. A. Pearlman, D. A. Case, J. W. Caldwell, W. S. Ross, T. E. Cheatham, S. Debolt, D. Ferguson, G. Seibel, and P. Kollman, Computer Physics Communications **91**, 1 (1995).
[92] W. Jiang, D. J. Hardy, J. C. Phillips, A. D. MacKerell, K. Schulten, and B. Roux, Journal of Physical Chemistry Letters **2**, 87 (2011).
[93] Y. J. Wei, J. T. Wu, H. Q. Yin, X. H. Shi, R. G. Yang, and M. Dresselhaus, Nature Materials **11**, 759 (2012).
[94] W. P. Zhu, H. T. Wang, and W. Yang, Acta Materialia **60**, 7112 (2012).
[95] D. M. Huang, R. Faller, K. Do, and A. J. Moule, Journal of Chemical Theory and Computation **6**, 526 (2010).

[96] K. N. Schwarz, T. W. Kee, and D. M. Huang, Nanoscale **5**, 2017 (2013).
[97] W. Shinoda, R. DeVane, and M. L. Klein, Soft Matter **4**, 2454 (2008).
[98] C. Oostenbrink, A. Villa, A. E. Mark, and W. F. Van Gunsteren, Journal of Computational Chemistry **25**, 1656 (2004).
[99] W. L. Jorgensen, D. S. Maxwell, and J. TiradoRives, Journal of the American Chemical Society **118**, 11225 (1996).
[100] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling, Journal of Chemical Theory and Computation **11**, 3696 (2015).
[101] A. D. MacKerell *et al.*, Journal of Physical Chemistry B **102**, 3586 (1998).
[102] C. Jarzynski, Physical Review Letters **78**, 2690 (1997).
[103] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, Science **220**, 671 (1983).
[104] T. Okabe, M. Kawata, Y. Okamoto, and M. Mikami, Chemical Physics Letters **335**, 435 (2001).
[105] J. D. Chodera and M. R. Shirts, Journal of Chemical Physics **135**, 194110 (2011).
[106] M. J. Field, P. A. Bash, and M. Karplus, Journal of Computational Chemistry **11**, 700 (1990).
[107] F. Maseras and K. Morokuma, Journal of Computational Chemistry **16**, 1170 (1995).
[108] M. Svensson, S. Humbel, R. D. J. Froese, T. Matsubara, S. Sieber, and K. Morokuma, Journal of Physical Chemistry **100**, 19357 (1996).
[109] D. Qiu, P. S. Shenkin, F. P. Hollinger, and W. C. Still, Journal of Physical Chemistry A **101**, 3005 (1997).
[110] G. D. Hawkins, C. J. Cramer, and D. G. Truhlar, Journal of Physical Chemistry **100**, 19824 (1996).
[111] A. Onufriev, D. Bashford, and D. A. Case, Proteins-Structure Function and Bioinformatics **55**, 383 (2004).
[112] R. H. Byrd, P. H. Lu, J. Nocedal, and C. Y. Zhu, SIAM J. Sci. Comput. **16**, 1190 (1995).

Chapter 3

Disassembly of CCMV capsids

3.1 Introduction

Several virions can be readily dissociated into protein subunits (e.g. dimers) and reassembled *in vitro* by adjusting pH, ionic strength and temperature [1-3]. This makes it possible to investigate the self-assembly mechanism of capsid proteins by various experimental techniques including small-angle X-ray scattering (SAXS) and static light scattering (SLS) [4-9], mass spectrometry (MS) [10-12], atomic force microscopy (AFM) [13,14] and nanofluidic devices [15,16]. On the other hand, a number of large-scale coarse-grained (CG) models [17], which represent one subunit with one or several beads, were used to explore the possible initial formation of metastable intermediates and their subsequent assembly to form a full capsid [18-21]. To achieve this, the CG beads interacted with each other through various pairwise potentials. However, the parameters entering these potentials, such as the association energy and the effective charge assigned to the CG beads, were rough estimates and their accuracy might be questionable in many cases. Evaluating reliably the interaction energy between self-assembling subunits from experimental data is therefore a crucial task for the construction of theoretical models bearing a predictive capability.

The investigations devoted to evaluating the interaction strength between capsid subunits are limited. A common method used by Ceres and Zlotnick [22] is based on the analysis of self-assembly kinetics. By fitting SLS data with a model of self-assembly kinetics derived from a nucleation-elongation growth process, the investigators were able to compute the apparent association energy between the self-assembling subunits of empty capsids. These experiments reported a rather low association energy of -5 ~ -10 $k_B T_r$ [22] for the capsid dimers of an icosahedral plant virus, where k_B is the Boltzmann constant and T_r is the room temperature. The low association energy enables an efficient self-assembly by avoiding kinetic traps, which facilitates the formation of regular structures. In principle, one of the most promising methods to obtain the interaction strength between subunits with a satisfactory accuracy is the calculation of the potential of mean force (PMF) through all-atom computer simulations. However such simulations require a high-accuracy parameterization and are in addition still costly in computation time. More importantly, the ionic conditions play a crucial role, but the precise dynamical relationships between the ionized groups, the change of conformations and the configuration of the protein occurring in the course of the self-assembly process remain a formidable challenge [23,24].

As reverse process of self-assembly, the disassembly of viral capsids also involve

interactions between subunits, indicating that we might also be able to extract the interaction information of subunits. In a compact capsid structure, each subunit shares a similar physicochemical environment. For CCMV, each dimer has four nearest neighbors and strong hydrophobic interactions with them. The disassembly of a capsid must overcome those interactions and result in a dispersed state where the subunits have no or weak interactions with each other. This process is analogous to the conventional liquid-gas phase transition. Therefore, it is possible to apply the models that account for the liquid-gas phase transition to the disassembly of viral capsids, and the lattice model is one of them.

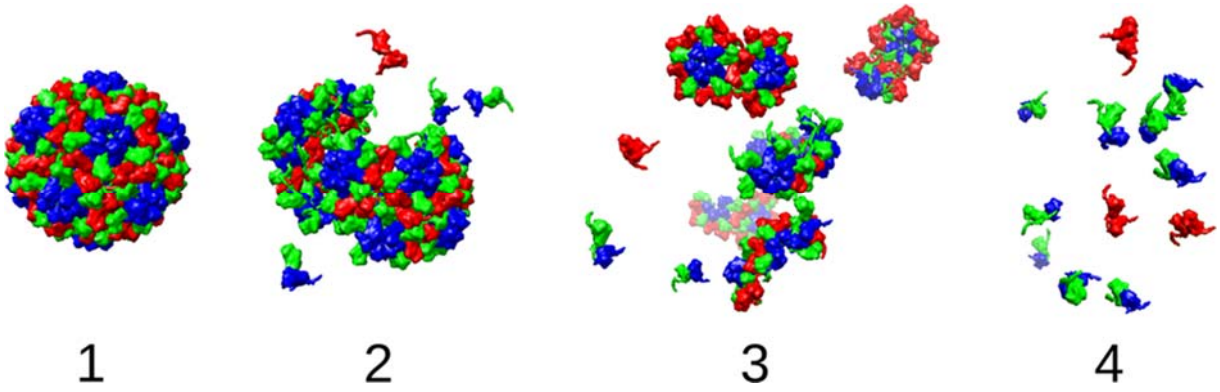

Figure 3-1. Disassembly pathway of empty CCMV capsids at pH 7.5 and ionic strength of 0.5 M detected by TR-SAXS. The disassembly pathway evolves from 1 to 4. Capsid proteins A, B, and C are colored into blue, green, and red, respectively. Adapted from [5].

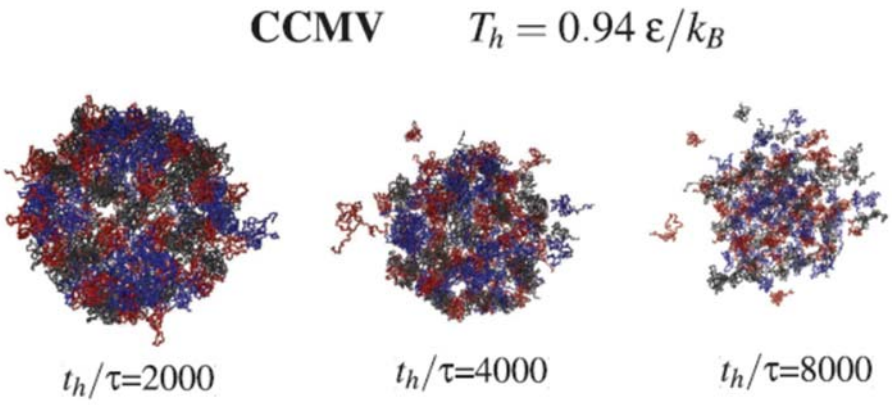

Figure 3-2. Disassembly of empty CCMV capsids at a high temperature $T_h = 0.94\varepsilon/k_B$, with ε the Lennard-Jones potential depth between the α–C atoms of proteins. The snapshots were taken at different dissociation time t_h and τ is the characteristic time scale needed for the effective particles to cover a distance of 5 Å through diffusion. Adapted from [25].

There are many methods to disassemble a capsid, based on different principles [5,25,26]. Like the way that we usually trigger the self-assembly of viral capsids, we can also trigger the disassembly of viral capsids by changing the interaction strength between subunits. In the case of CCMV for example, it can be achieved by adjusting the ionic strength and pH of the solution so that the structure of full capsids becomes thermodynamically unstable. This approach can be viewed as an enthalpy-driven disassembly. Disassembling CCMV capsids with this approach

has been probed by time-resolved small-angle X-ray scattering (TR-SAXS) technique [5]. The dissociation follows a pathway shown in Figure 3-1. The CCMV capsids firstly broke into two pieces, followed by a gradual dissociation of dimers from each piece and forming some metastable intermediates, before leaving only dispersed dimers. On the other hand, we can also disassemble a capsid in a way analogous to the evaporation of a liquid into a gas phase by rising the temperature. Hence, it is an entropy-driven process. With the increase of temperature, entropy is large enough to overcome the contribution of enthalpy, and dimers prefer to remain in a dispersed state to maximize their translational and rotational entropy. The pathway for the thermal disassembly of empty CCMV capsids determined by computer simulations [25] is depicted in Figure 3-2. Unlike the stepwise pattern observed in the disassembly triggered by tuning pH and ionic strength, the thermal disassembly of capsids exhibited a continuous and gradual pattern: the separation between proteins gradually increases until the capsids are entirely dissociated and form a dispersed state. Additionally, the presence of viral genome does not affect the pathway of thermal disassembly [25].

With the understanding of the disassembly of viral capsids above, we proposed a new approach for estimating the interaction strength between subunits based on the thermal disassembly of viral capsids. We dissociated the capsid by heating up the solution and we monitored the melting temperature by fluorescence thermal shift assay. A mean field (MF) model was constructed to describe the dissociation of a viral capsid, whose subunits interact through a short-range hydrophobic attraction and a pairwise electrostatic Yukawa repulsion.

3.2 Fluorescence thermal shift assay

As revealed by the computational simulations on the thermal disassembly of empty CCMV capsids above, we can expect that the hydrophobic domains become exposed when the gap between subunits is gradually widened with the disassembly. The exposed area of the hydrophobic domains, to some extent, reflects the disassembly progress. Those exposed hydrophobic domains can be readily monitored using hydrophobic fluorescent dyes which tend to bind on hydrophobic domains and can be quantitatively monitored. The schematic of the principle of fluorescence thermal shift assay experiment is plotted in Figure 3-3(A).

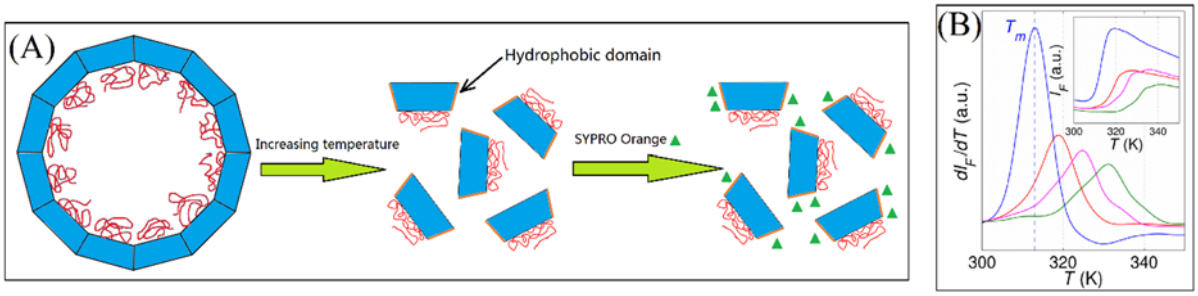

Figure 3-3. Schematic of the principle of fluorescence thermal shift assay experiments (A), and an example of the disassembly signals, with fluorescence intensities (inset) and their corresponding derivatives with respect to the temperature (B). T_m is defined as the temperature that maximizes the derivative of the intensity. The body of subunits is colored in blue while the tails of subunits are in red. Besides, the exposed hydrophobic domains are highlighted in orange

and SYPRO Orange dyes are denoted by green triangle.

The melting temperatures T_m of CCMV capsids are shown in Figure 3-4 and the effect of ionic strength and pH are probed. The interaction strength between dimers varies with the environment around them. The Coulomb potential between dimers is a function of both the ionic strength and pH of the solution. A high ionic strength weakens the Coulomb potential while the ionization of dimers is sensitive to the pH. These effects were reflected in the variation of the melting temperature T_m, a higher T_m indicating the strengthening of the interaction between subunits.

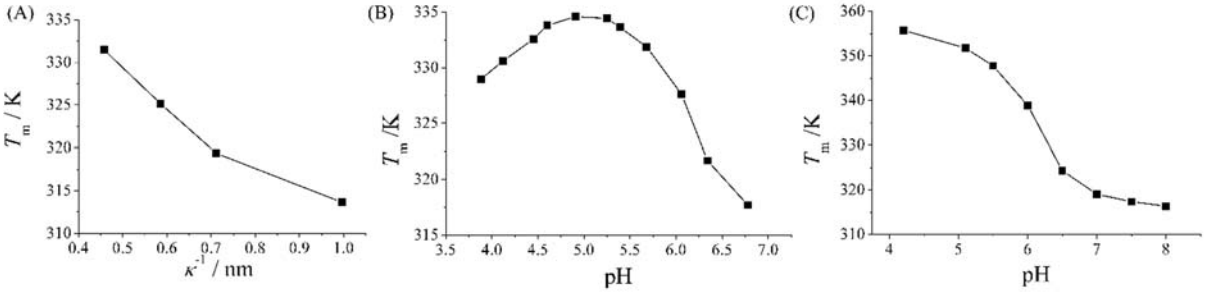

Figure 3-4. Melting temperature of empty CCMV capsids as a function of the Debye length at a fixed pH of 4.8 (A), of pH at a fixed NaCl concentration of 0.5 M (B), and melting temperature of CCMV virions as a function of pH at a fixed NaCl concentration of 0.5M (C).

3.3 Lattice model

CCMV capsid is a typical icosahedral protein shell formed by 180 proteins. Although the proteins are all chemically identical, they have slight conformational differences and are accordingly categorized into A, B or C depending on their environment within the icosahedral architecture. A $\mathcal{T}$ = 3 CCMV capsid comprises 60 dimers AB and 30 dimers CC as shown in Figure 3-5(A). These dimers are the basic self-assembly subunits and contact each other in two forms hereafter called AB-AB and AB-CC contacts.

In lattice model, a viral capsid was meshed into a set of homogeneous or heterogeneous sites located at the centers of mass of the subunits and separated by a distance of $2R_d$. Figure 3-5(A) shows a meshing scheme of a CCMV capsid for a homogeneous lattice model. Considering the short-range nature of the hydrophobic interaction between subunits, the sites were set to interact with their nearest neighbors by a pairwise potential ϵ_{hp} with subscript "hp" denoting hydrophobic interaction. To estimate the electrostatic interactions, we assumed that the subunits on the lattice sites met the condition $e\phi/k_{\mathrm{B}}T < 1$, where ϕ is the surface electrostatic potential of the subunits, and then the electrostatic interaction between the occupied sites was simply evaluated by a linearized Poisson-Boltzmann equation. For two occupied sites carrying an effective charge Z and separated by a distance r, the pairwise electrostatic potential is given by [27]

$$U(r) = \frac{Z^2 e^2}{4\pi\varepsilon_0\varepsilon_r(1+R_d\kappa)^2}\frac{\exp[-\kappa(r-2R_d)]}{r}, \quad (1)$$

where ε_0 and ε_r are the vacuum permittivity and the dielectric constant (80 for water) respectively, while the Debye screening length $\kappa^{-1} = \sqrt{\varepsilon_0\varepsilon_r k_B T/2e^2 c_s}$ is related to temperature T and solution salinity c_s.

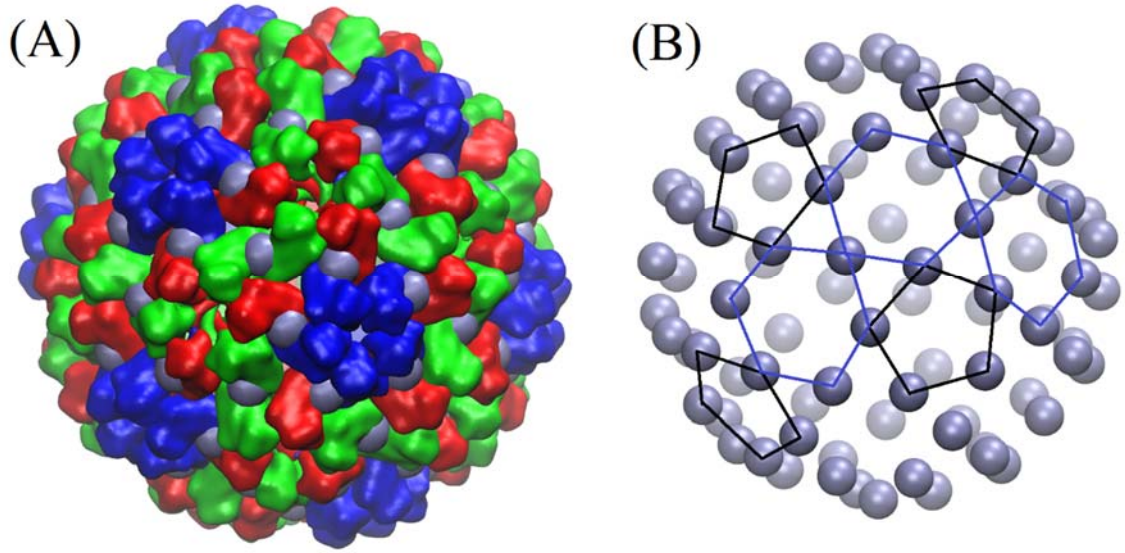

Figure 3-5. (A) Schematic of meshing a CCMV capsid into a lattice model. Capsid proteins A, B, and C are in blue, red, and green respectively. The lattice sites are located at the mass center of each dimer and represented with a gray particle. (B) Lattice sites of a CCMV capsid with the dimers forming pentamers and hexamers are linked by black and blue lines, respectively.

3.3.1 Homogenous lattice

In the homogenous lattice, each lattice site represents one subunit and is not distinguishable (see Figure 3-5(B)). The total free energy for our lattice in the grand canonical ensemble is given by the expression

$$\frac{F(\rho,T)}{M} = 2\epsilon_{hp}\rho^2 + \frac{Z^2 e^2 \rho^2}{8\pi\varepsilon_0\varepsilon_r R_d(1+\kappa R_d)^2} - \rho\mu_0 + k_B T[\rho\ln\rho + (1-\rho)\ln(1-\rho)], \quad (2)$$

where M is the total number of lattice sites and $\rho = N/M$ is the density of subunits. μ_0 is the chemical potential and, in the case of empty capsids, it can be written as $\mu_0 = k_B T\ln[\rho_0/(1-\rho_0)]$ [28], where ρ_0 is the density of free subunits in the reservoir and was set to 4.8×10^{-4} to be consistent with our experimental data for CCMV empty capsid [29]. In the case of RNA-loaded virions, since the dissociated subunits remain bound to RNA, the chemical potential arises from the surface tension γ_g of the protein-RNA globule complex expressed through its correlation length ξ_g, $\gamma_g = k_B T/\xi_g^2$. Then the chemical potential takes a new form, $\mu_0 = -\gamma_g D^2 = -k_B T\left(D/\xi_g\right)^2$ [30], where D^2 is the excluded area of subunits on the capsid surface.

At equilibrium, the free energy is minimal with respect to the density of subunits,

$$\frac{\partial F}{\partial \rho} = \left(4\epsilon_{\mathrm{hp}} + \frac{Z^2 e^2}{4\pi\varepsilon_0\varepsilon_r R_d(1+\kappa R_d)^2}\right)\rho - \mu_0 + k_{\mathrm{B}}T\ln\left(\frac{\rho}{1-\rho}\right) = 0,$$

and we obtain an equation of state relating ρ and T, which exhibits a first-order phase transition at a particular temperature T_{m}. At T_{m}, the total free energy of an empty capsid is a symmetrical function of ρ, possessing a pair of symmetrical minima about $\rho = 0.5$, one at high subunit density and another at low subunit density, as shown in Figure 3-6. When the temperature exceeds T_{m}, the minimum of F will shift from the high subunit density to the low density indicating the occurrence of a phase transition at T_{m}. Thus, we have the following relationship

$$k_{\mathrm{B}}T_{\mathrm{m}}\ln\left(\frac{\rho_0}{1-\rho_0}\right) = -2\epsilon_{\mathrm{hp}} + \frac{Z^2 e^2}{4\pi\varepsilon_0\varepsilon_r R_d\left[1+R_d\sqrt{\frac{2e^2 c_S}{\varepsilon_0\varepsilon_r k_{\mathrm{B}}T_{\mathrm{m}}}}\right]^2}. \quad (3)$$

Like the van der Waals liquid-gas phase transition, the equation of state exhibits stable, metastable, and unstable branches, as shown in the Figure 2-9 of Chapter 2.

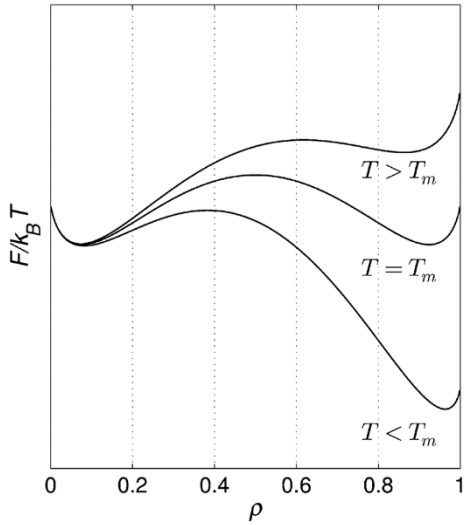

Figure 3-6. Mean-field free energy F as a function of subunit density ρ for various temperatures T around the mean-field melting temperature T_{m}. As the temperature increases, the stable point, i.e., the one with the lowest free energy, shifts from the high to the low subunit density, giving rise to a first-order phase transition. At T_{m}, the free energy is a symmetrical function about ρ = 0.5. Adapted from [29].

Temperature dependence of the hydrophobic interaction

The hydrophobic interaction is the main driving force for the association of viral subunits [19,22]. However, the hydrophobic interaction is considered to be the manifestation of entropic effects [31] and thus depends strongly on temperature. Some experiments on the self-assembly of different viruses showed that temperature could impact the self-assembly kinetics of viral capsids, yet the effect of temperature varies with the viral species. For instance, hepatitis B virus (HBV) was found to self-assemble faster at high temperature [22], whereas the self-assembly of CCMV capsid is more efficient at lower temperature [4]. To elucidate the reason of this discrepancy, it is relevant to study the temperature dependence of the hydrophobic interaction between capsid subunits.

The dependence of the hydrophobic energy ϵ_{hp} on temperature T can be accounted for by a first-order expansion as follows [32],

$$\epsilon_{\mathrm{hp}} \approx -2A_c[\gamma_0 - s_0(T - T_0)] \tag{4}$$

where A_c is the contact area, whereas γ_0 and s_0 are the surface tension and the surface excess entropy respectively at the reference temperature T_0, $T_0 = 273.15$ K here.

Monte Carlo (MC) simulations in the grand canonical ensemble were carried out to study the melting of viral capsid lattice. The total energy of a capsid comprising *N* subunits is

$$U(N) = \epsilon_{\mathrm{hp}} \sum_{i,j} n_i n_j + \frac{Z^2 e^2}{4\pi\varepsilon_0\varepsilon_r k_{\mathrm{B}} T}\left[\frac{\exp(\kappa R_d)}{1+\kappa R_d}\right]^2 \sum_{i,j} \frac{\exp(-\kappa r_{ij})}{r_{ij}} \tag{5}$$

where n_i and n_j equate to 0 or 1. On average, each dimer undergoes a removal or insertion trial from or into the available sites of the lattice according to the Metropolis algorithm. Thermal annealing method was used to accelerate sampling and to avoid the system to become trapped in metastable states.

Results

As shown in Figure 3-7(a), the lattice model was able to reproduce a first-order phase transition. By raising temperature, a full icosahedral lattice sharply melted at a specific temperature, T_{m}. We can also see that the melting temperature was more sensitive to the hydrophobic interaction than to the effective charge per subunit. Figure 3-7(b) clearly shows that the melting temperature determined by MC simulations was in perfect agreement with that given by MF prediction when the thermal annealing method was applied.

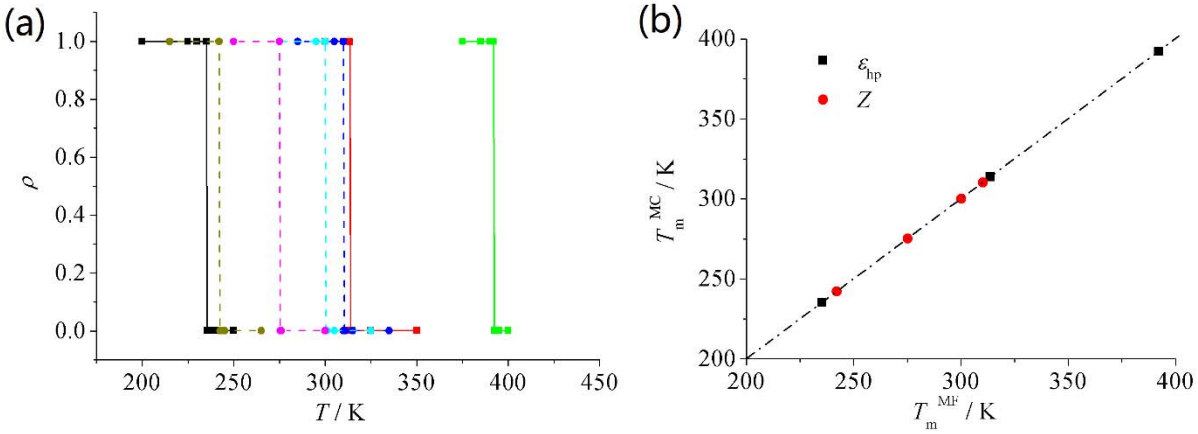

Figure 3-7. (a) Melting curves of a homogenous icosahedral lattice calculated by MC simulations for different hydrophobic interaction strengths at $Z = 0$ (solid lines): ϵ_{hp}= -$2k_{\mathrm{B}}T_{\mathrm{r}}$ (black), -$3k_{\mathrm{B}}T_{\mathrm{r}}$ (red) and -$4k_{\mathrm{B}}T_{\mathrm{r}}$ (green); and effective charges of subunit with $\epsilon_{\mathrm{hp}} = -2k_{\mathrm{B}}T_{\mathrm{r}}$ (dash lines): Z = -2 (blue), -4 (cyan), -7 (magenta) and -10 (dark yellow). The density of free subunits is $\rho_0 = 4.8 \times 10^{-4}$. (b) Comparison between melting temperatures T_{m} calculated by MC simulation (same data as in (a)) and by MF theory. The dot-dashed line is used to guide the eye. Adapted from [33].

Fitting parameters

In the homogeneous lattice model, the unknown parameters (ϵ_{hp} and Z for the case where the hydrophobic interaction is independent of temperature; s_0, γ_0 and Z for the case when a temperature dependence on the hydrophobic interaction was considered), could be determined by fitting the melting temperature measured by fluorescence thermal shift assays at different salinities [29] with Equation (3). For CCMV subunits, $A_c = 2.943\times10^{-17}$ m^2 [34]. The fitting parameters are given in Table 1. With these fitting parameters, the reproduced melting temperatures are shown in Figure 3-8. For the temperature-independent hydrophobic interaction, ϵ_{hp} was evaluated to be $-4.38k_BT_r$ with a room temperature T_r of 300 K, close to the value of -5 ~ $-10k_BT_r$ predicted by the analysis of self-assembly kinetics [22]. However, the effective charge of dimer $|Z|$ was predicted to be larger than 5, a value twice as high as that determined by electrophoretic mobility experiments [35]. The discrepancy between experimental and calculated melting temperatures at κ^{-1}=1 nm (Figure 3-8) was due to the fact that the linear approximation of the Poisson-Boltzmann equation was no longer valid for such a low ionic strength.

Table 1. Parameters of the interaction energies between subunits obtained by fitting experimental melting temperatures (see Figure 3-8) when the temperature dependence of the hydrophobic interaction is considered or not.

Temperature dependence	Parameter
independent	$\epsilon_{hp} = -4.38k_BT_r$
independent	$\|Z\| = 5.23$
dependent	$\gamma_0 = 0.248$ mN/m
dependent	$s_0 = -8.5\times10^{-4}$ mN/m•K
dependent	$\|Z\| = 2.0$

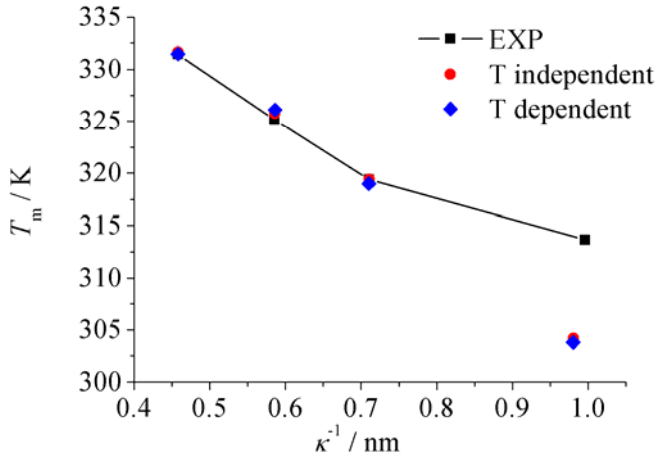

Figure 3-8. Reproduced melting temperatures T_m for empty CCMV capsid as a function of Debye length at pH 4.8 calculated by MC simulations with the fitting parameters given in Table 1 for temperature-independent (red bullet) and temperature-dependent (blue diamond)

hydrophobic interactions. The black squares are the melting temperatures measured by fluorescence thermal shift assay at pH 4.8 and different salinities. Adapted from [33].

In the case of a temperature-dependent hydrophobic interaction, the estimated negative value of s_0 indicates that the hydrophobic interaction becomes stronger with the increase of temperature, which was also reflected by the shrinkage phenomenon of loaded capsids at pD 7.5 observed by SANS (see section 4). The dependence of hydrophobic interaction on temperature can readily lead us to speculate that a high temperature would facilitate the self-assembly of CCMV capsids. However, there are experimental observations showing that the self-assembly of empty CCMV capsid is accelerated at low temperature [32]. These controversial conclusions reflect the complicate nature of the self-assembly process of viral capsids. In addition, by introducing a temperature dependence into the hydrophobic interaction, the effective charge of a CCMV subunit was reduced to an absolute value of 2, which is comparable to that estimated by the electrophoretic mobility of CCMV virions.

pH effect

CCMV capsid protein comprises a positively-charged flexible arm and a negatively-charged compact body. The high pK_a value of the charged residues in the arm (ARG pK_a 12.10, LYS pK_a 10.67 and terminal residue SER pK_a 9.05) leads to a constant net charge below pH 8.0. By contrast, some residues with medium pK_a values in the negatively-charged body, such as GLU (pK_a 4.15), will change their ionization state within the pH range of 3 ~ 8 commonly used in experiments.

Assuming that the hydrophobic interaction is independent of pH, the variations of the effective charge of subunits could be estimated by relating the MC simulations of the lattice model to the experimental melting temperatures (Figure 3-9). Despite the introduction of temperature dependence into the hydrophobic interaction, the variations of $|Z|$ with pH were very similar to those without temperature dependence, as shown in Figure 3-9(a). Both cases showed a minimum charge value around pH 4.8, which agrees with the isoelectric point of CCMV protein measured by electrophoretic mobility experiments [35] and predicted by the PDB2PQR package [36]. However the range of values was different for the two cases. For temperature-independent hydrophobic interaction, $|Z|$ varied from 4.5 to 8, whereas the values varied from 1.6 to 3 when taking the temperature effect into account.

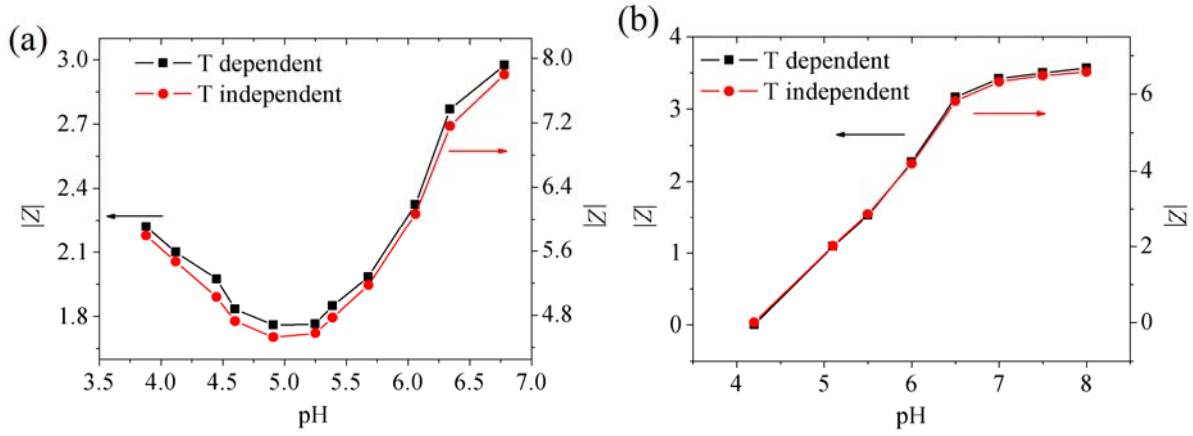

Figure 3-9. Absolute value of the effective charge $|Z|$ of a subunit constituting either an empty CCMV capsid (a) or a CCMV virion (b) as a function of pH estimated by MC simulations at a salinity of 0.5 M with the experimentally measured melting temperatures. The influence of the temperature dependence of the hydrophobic interaction is analyzed in both cases: temperature independent (red bullet) and temperature-dependent (black square). Adapted from [33].

In the case of RNA-loaded capsids, a charge compensation occurs between the cationic charges carried by the N-terminal arms of the dimer and the anionic charges carried by RNA, both of which being not sensitive to pH. Thus, the variation of the effective charge of the dimer can be mainly ascribed to the negatively charged body. In ref [29], we found that the dissociation temperature of CCMV virions was not sensitive to salinity. Besides, electrophoretic mobility experiments also revealed that the isoelectric point (pI) of 22- residues-cleaved CCMV protein in which the positive charge of the N-terminal arm was totally removed is around 4.1 [35], while predictions obtained by the PDB2PQR package [36] gave an isoelectric point around pH 4.3 for CCMV protein without the first 22 residues. Therefore, we made an approximation of $|Z| = 0$ at pH 4.2. If we assume that the hydrophobic interaction strength does not change in the presence of RNA, then we can have $\mu_0 = -k_B T \left(\frac{D}{\xi_g}\right)^2 = k_B T \ln[\rho_0/(1-\rho_0)]$ at equilibrium, and D/ξ_g can be estimated to be ~2.76. The variation of the absolute value of the effective charge of dimer with pH is presented in Figure 3-9(b). No matter whether the temperature effect was considered or not, the variation of the effective charge was similar: $|Z|$ increased with pH and eventually reached a plateau at pH higher than 7. This variation was comparable with the change of electrophoretic mobility with pH for cleaved proteins [35]. In addition, we also found that the effective charge in the case of temperature-dependent hydrophobic energy was half of that obtained in the case of temperature-independent hydrophobic energy.

3.3.2 Heterogeneous lattice

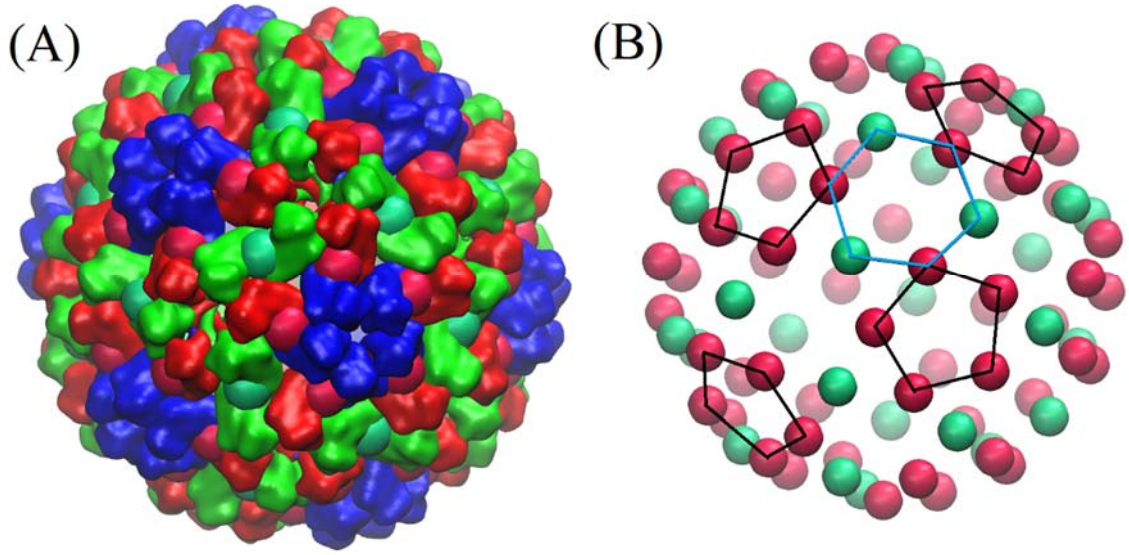

Figure 3-10. (A) Schematic of the meshing of a CCMV capsid into a heterogeneous lattice with capsid proteins A, B, and C in blue, red and green, respectively. (B) Heterogeneous lattice of a CCMV capsid where the dimers forming pentamers and hexamers are linked by black lines and blue lines, respectively. The lattice sites represented by red particles for AB dimers and green particles for CC dimers, are located at the center of mass of each dimer.

Up to date, the precise assembly and disassembly process of viral capsid is unclear. Experiments on HBV capsids by adjustment of pH revealed that capsid dissociation was independent of the dimer concentration, which suggested a possible first-order phase transition [26]. In addition, it has been shown in reference [29] as well as in the preceding section that the dissociation of a viral capsid occurs through a sharp phase transition in the homogeneous lattice model. However, experiments on the dissociation of viral capsids by AFM found that there were a lot of metastable subassemblies equilibrium with partially dissociated capsids [14], which suggested that the interaction strengths between the subunits were not all the same and that the dissociation of a capsid might take place in more than one step.

Indeed, as mentioned in the section 3, two kinds of dimer contact each other in two forms hereafter called AB-AB and AB-CC contacts. By analyzing the crystal structure of CCMV capsid [34], it was found that AB-AB contact has a larger contact area than AB-CC contact, indicating a stronger hydrophobic attraction between AB dimers and a weaker one between AB and CC dimers. The way how this asymmetrical interaction between dimers impacts the stability of viral capsid is however unclear.

Here, we assigned different values to the interaction strength between a pair of AB dimers $\epsilon_{\mathrm{AB\_AB}}$ and between an AB dimer and a CC dimer $\epsilon_{\mathrm{AB\_CC}}$. For the convenience of analysis, the temperature dependence of hydrophobic interaction was not considered in this model. Given that in a CCMV capsid, an AB dimer has four CC dimers as nearest neighbors, while the nearest neighbors of a CC dimer include two AB dimers and two CC dimers, the free energy of the two-component heterogeneous icosahedral lattice from mean field theory is given by

$$F = 4\epsilon_{\mathrm{AB\_CC}}\rho_{CC}\rho_{\mathrm{AB}}M_{\mathrm{CC}} + \epsilon_{\mathrm{AB\_AB}}{\rho_{\mathrm{AB}}}^2 M_{\mathrm{AB}} + \frac{Z^2e^2(M_{\mathrm{AB}}\rho_{\mathrm{AB}}+M_{\mathrm{CC}}\rho_{\mathrm{CC}})^2}{4\pi\varepsilon_0\varepsilon_r R_d(1+\kappa R_d)^2} - (M_{\mathrm{AB}}\rho_{\mathrm{AB}} + M_{\mathrm{CC}}\rho_{\mathrm{CC}})\mu_0 + k_{\mathrm{B}}TM_{\mathrm{AB}}[\rho_{\mathrm{AB}}\ln\rho_{\mathrm{AB}} + (1-\rho_{\mathrm{AB}})\ln(1-\rho_{\mathrm{AB}})] + k_{\mathrm{B}}TM_{\mathrm{CC}}[\rho_{\mathrm{CC}}\ln\rho_{\mathrm{CC}} + (1-\rho_{\mathrm{CC}})\ln(1-\rho_{\mathrm{CC}})], \quad (6)$$

where M_{AB} and M_{CC} are the number of lattice sites for AB dimer and CC dimer respectively (for CCMV capsid, $M_{\mathrm{AB}} = 60$ and $M_{\mathrm{CC}} = 30$) and ρ_{AB} and ρ_{CC} are the respective densities of AB and CC subunits.

Results

Figure 3-11 shows the evolution of the melting temperature T_{m} obtained from MC simulations at different $\epsilon_{\mathrm{AB\_AB}}$ and $\epsilon_{\mathrm{AB\_CC}}$. Interestingly, T_{m} showed a linear relationship with both $\epsilon_{\mathrm{AB\_AB}}$ and $\epsilon_{\mathrm{AB\_CC}}$ over a large range of values. However, in Figure 3-11(a) the stronger the interaction between AB-AB dimer, the higher the critical value of $\epsilon_{\mathrm{AB\_CC}}$ beyond which the melting temperature deviated from the linear regime. If we plot T_{m} as a function of $\epsilon_{\mathrm{AB\_AB}}$, we can clearly find a shift of the slope of T_{m} vs $\epsilon_{\mathrm{AB\_AB}}$ at low $\epsilon_{\mathrm{AB\_CC}}$ (see Figure 3-11(b)), reflecting the variation of the dependence of T_{m} on $\epsilon_{\mathrm{AB\_AB}}$.

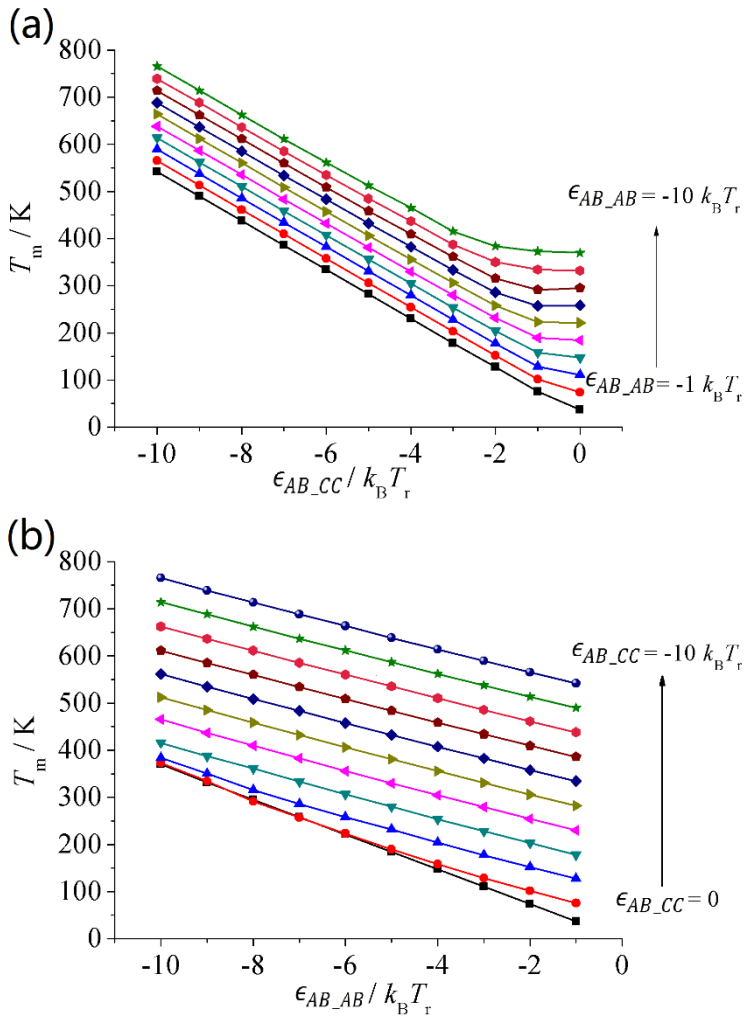

Figure 3-11. Melting temperatures T_m obtained by MC simulations with the heterogeneous lattice model for subunits with $|Z| = 4.2$. Both ϵ_{AB_CC} and ϵ_{AB_AB} vary from 0 to $-10k_BT_r$ with increment of $-1k_BT_r$. T_m versus ϵ_{AB_CC} (a) and T_m versus ϵ_{AB_AB} (b). Adapted from [33]

In order to understand the deviation from linearity for some ratios of the dimer interaction strengths, the melting curves of an empty capsid are plotted in Figure 3-12. Here, we used two typical sets of parameters. For ϵ_{AB_AB} set to $-10k_BT_r$ (Figure 3-12(a)), we can clearly see a transformation of the melting behavior from a sharp one-step transition to a gradual two-step transition and the corresponding threshold value of ϵ_{AB_CC} was found to be between $-4k_BT_r$ and $-3k_BT_r$. A clearer transformation process can be found in Figure 3-12(b) where ϵ_{AB_CC} was fixed at $-1k_BT_r$. Empty capsid underwent a partial melting prior to a full melting for $\epsilon_{AB\_AB} = -3k_BT_r$. With increasing values of ϵ_{AB_AB} a two-step melting gradually took place.

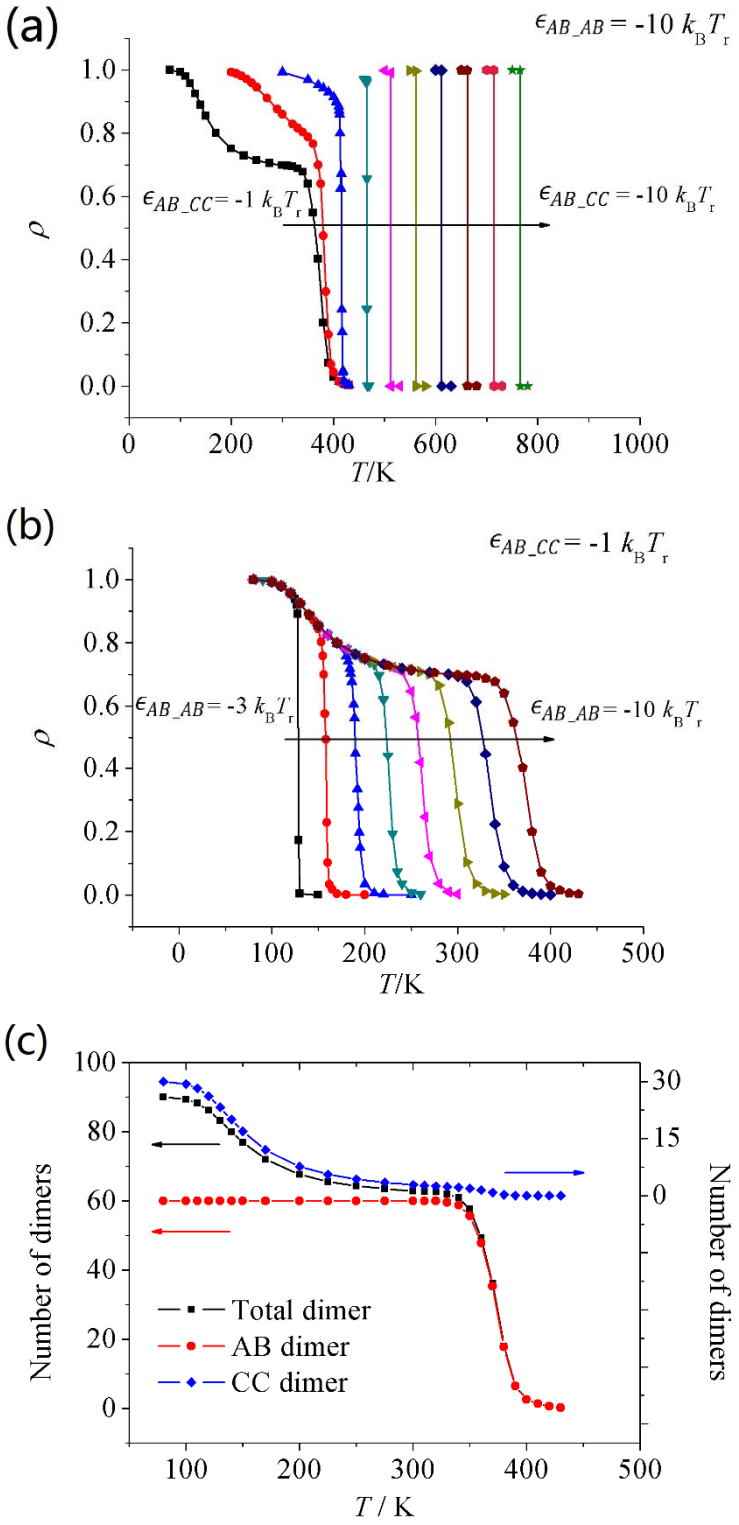

Figure 3-12. Melting curves of empty CCMV capsid calculated by MC simulations for subunits with $|Z| = 4.2$. (a) The change of melting regime from one-step transition to two-step transition by increasing ϵ_{AB_CC} from -10k_BT_r to -1k_BT_r with ϵ_{AB_AB} fixed at -10k_BT_r,. (b) Two-step melting at different ϵ_{AB_AB} (from -3 k_BT_r to -10 k_BT_r) with ϵ_{AB_CC} = -1 k_BT_r. (c) Melting curve of each component in the heterogeneous lattice with ϵ_{AB_AB} and ϵ_{AB_CC} equal to -10 k_BT_r and -1 k_BT_r, respectively. Adapted from [33].

We might get some insight into this phenomenon from the nature of the capsid. For CCMV capsid, 60 AB dimers form 12 pentamers of dimers located at the vertices of an icosahedron (see Figure 3-10(B)). The pentamer of dimers is often reported to be one of the most stable subassemblies in the CCMV capsid, whereas the 30 CC dimers act as glue between these pentamers to form a full capsid. Under this assumption, the weakly connected CC dimers will be the first to dissociate upon an increase in temperature, which is consistent with the fact that 30 subunits are dissociated during the first stage (Figure 3-12(c)). The second stage can be attributed to the dissociation of the more stable pentamers at higher temperature.

3.4 Thermal shrinkage of viral capsids

3.4.1 Small-Angle Neutron Scattering

As mentioned above, the disassembly pathway of viral capsids depend on the factor triggering the disassembly reaction. TR-SAXS revealed that the disassembly of CCMV capsids by changing pH and ionic strength follows a step-wise disassembly pathway, while the thermal disassembly pathway of CCMV capsids is still unclear even though a continuous and gradual pattern was observed by computational simulations. Similarly, to experimentally confirm the pathway predicted by computational simulations, small-angle neutron scattering (SANS) was applied here to monitor the thermal dissociation of CCMV capsids.

In SANS experiments, a buffer solution containing 68% of heavy water was used to contrast match RNA and to highlight the scattering intensity from proteins. The purified CCMV virions were dialyzed against buffer solutions at pD 7.5 (50 mM Tris-DCl, 0.1 M NaCl) and pD 4.8 (50 mM sodium acetate, 0.1 M NaCl). The typical virion concentration was around 9 g/L and the samples were stored at 4 °C. More details about the SANS experiments can be found in ref [37].

Results

CCMV has a capsid diameter of 28 ~ 32 nm depending on the pH [34]. By increasing pH from 4.8 to 7.5, the electrostatic repulsion between dimers becomes stronger due to the deprotonation of carboxyl groups (R-COOH → R-COO^-), which results in the swelling of CCMV particles at pH 7.5. Therefore, we found different radii of virions at 20 °C for pD 7.5 (see Figure 12(a)) and pD 4.8 (see Figure 12(b)).

In order to probe the phase transition of CCMV virions, we increased the temperature of the samples. The scattering intensities at pD 7.5 and pD 4.8 are displayed in Figure 3-13(c) and Figure 3-13(d), respectively. For pD 7.5, we observed that the first scattering peak slightly shifted to high q-values upon raising temperature from 20 °C to 65 °C (Figure 3-13(c)). By fitting the scattering intensities with a vesicle model, the radius of the capsid decreased from 14.5 nm at 20 °C to 12.5 nm at 65 °C (see Figure 3-13(a)). The shrinkage of CCMV capsid may result from the strengthening of hydrophobic interaction between capsid proteins due to its entropic nature. However, for pD 4.8, we did not found any significant variation for the

scattering patterns between 20 and 62 °C (see Figure 3-13(b)). It is probably because the subunits of the capsid were already in close contact at pD 4.8 and 20 °C and there was no further space left for shrinkage. Yet, at 72 °C the typical oscillations of the scattering patterns vanished, which indicated that the capsid proteins were probably in a disordered state.

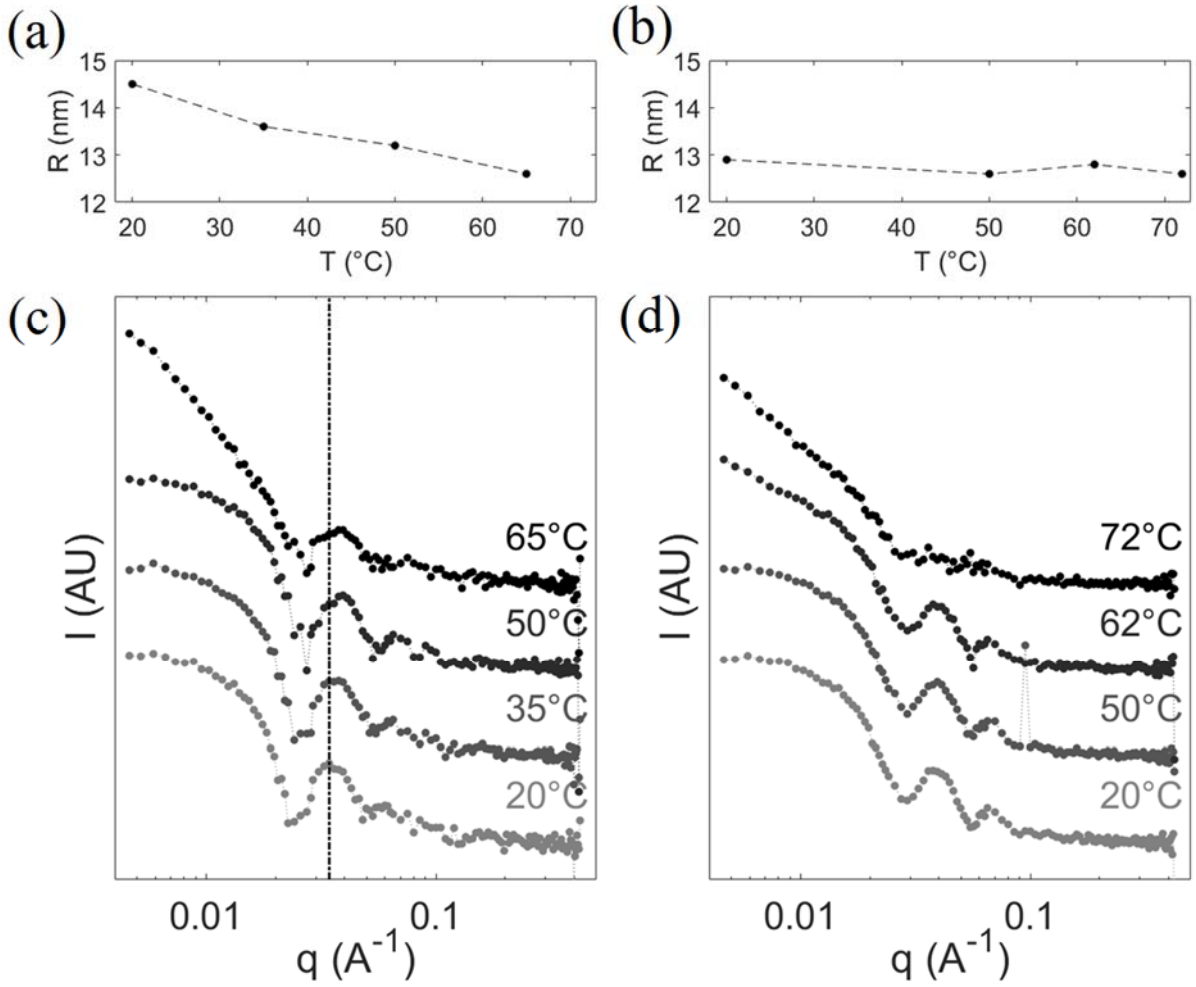

Figure 3-13. SANS patterns for the disassembly of CCMV capsids at pD 7.5 (a, c) and pD 4.8 (b, d) in 68% heavy water. The external radii of the CCMV virions in (a) and (b) are obtained by fitting the data with a vesicle model. Adapted from [33].

3.4.2 Molecular dynamics simulations

The major driving force for the capsid self-assembly is the hydrophobic interaction between subunits. Yet due to its nature, the temperature affects the strength of the hydrophobic interaction. In our previous thermal dissociation experiment of CCMV capsid by small-angle neutron scattering, we observed that the size of the capsid decreased from 14.5 nm at 20 ºC to 11.5 nm at 65 °C at pD 7.5 due to the strengthened hydrophobic interaction at high temperature. To gain a molecular insight into this phenomenon, a CG simulation using the MARTINI force field [38,39] was performed at a higher temperature (400 K) to accelerate the shrinkage process. Meanwhile, a reference simulation to probe the stability of the native capsid was performed at

room temperature (300 K).

Coarse-grained (CG) molecular dynamics simulation protocol

If a complete capsid is still too large to get statistically meaningful information with an all-atom model, CG models can help to improve our understanding of the capsid stability under various conditions and to allow comparisons with experiments [40,41]. The MARTINI CG force field is one of the optimal options, demonstrating a good compromise between molecular details and simulation speed, and it has been applied to investigate viral systems, including CCMV [42]. A 20-μs MD simulation at 300 K was first carried out to test the stability of a CCMV capsid modeled with the MARTINI CG force field. To mimic the thermal dissociation process of viral capsids, the temperature was increased up to 400 K for an additional 20 μs. The stability of the capsid was evaluated by analyzing the average radius of the capsid and the distance distribution between the monomers on the capsid.

CG simulations with MARTINI force field were performed with the GROMACS simulation package. The missing residues of the tails were not added and the CG particles were restrained by the ELNEDYN elastic network [43] with a spring constant of 200 kJ $mol^{-1}nm^{-2}$ to prevent the unfolding of the polypeptide chains. Temperature and pressure were maintained at 300 K or 400 K and 1 bar using the Berendsen weak-coupling method [44] with τ_T = 1 ps and Parrinello-Rahman barostat [45] with τ_p = 12 ps, respectively. The time step was 20 fs. The short-range van der Waals interactions were cut off and shifted at 1.1 nm, while Coulomb interactions were treated by the reaction field method [46] with a short-range truncation at 1.1 nm and the introduction of a relative permitivitty constant ε_r = 15. All bonds were constrained by the LINCS algorithm [47]. All CG simulations went through a 10,000-step energy optimization and a 25,000-step equilibration.

Results

The capsid was stable at 300 K with a radius around 11.4 nm (Figure 3-14(A)). The distribution of COM distances between monomers within the capsid shows a regular pattern with distinctive peaks with only a slight loss of long-range order (around a separation of 25 nm between monomers) due to thermal fluctuations (Figure 3-14(B)). In addition, the symmetry of the capsid, through the fivefold and threefold symmetry axes, was still conserved (see the snapshots in Figure 3-14(C)).

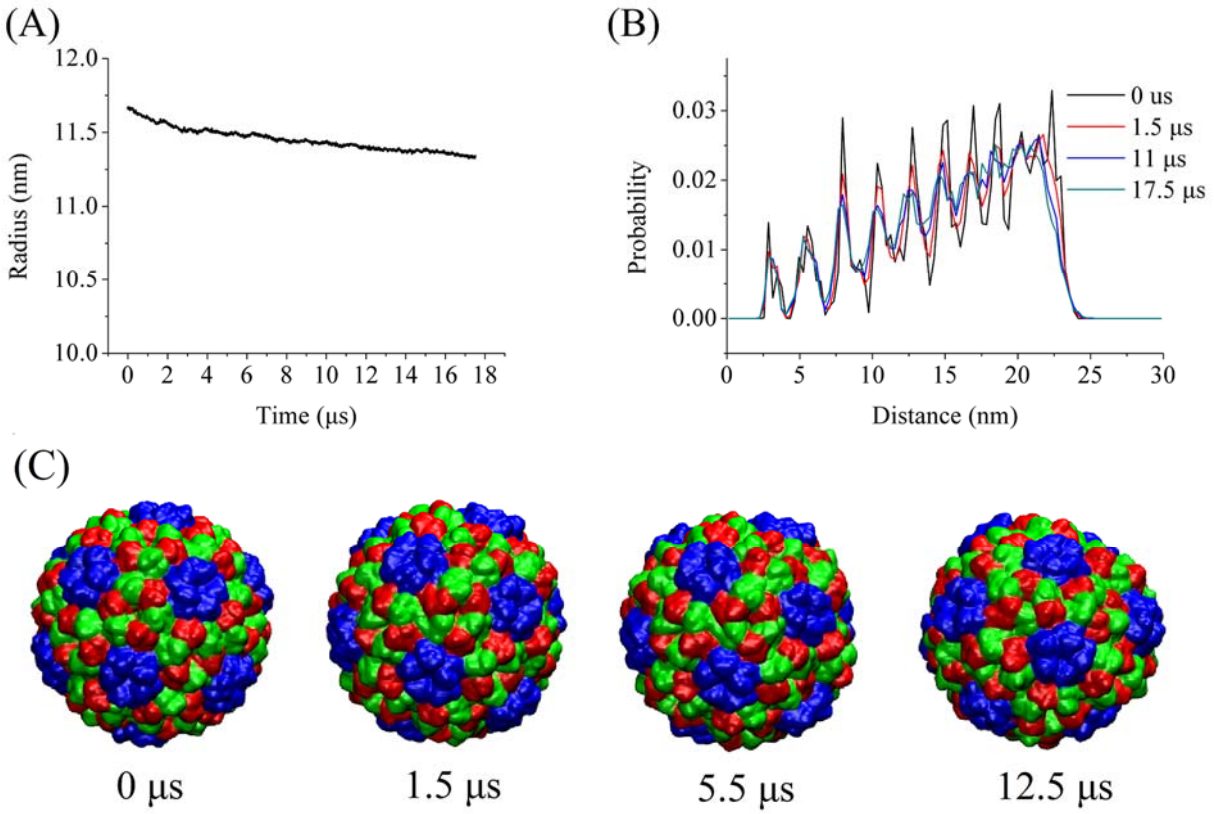

Figure 3-14. Variations of the radius of the capsid at 300 K (A) and distributions of COM distances between monomers within a capsid at different time steps (B). Snapshots of conformation at different time steps (C). Adapted from [37].

In contrast, at 400 K, a reduction of the apparent radius of the capsid from 11.8 nm to 10.2 nm was observed (Figure 3-15(A)). Moreover, the capsid quickly lost its symmetry with a distribution of distances between monomers fading away, leaving only a short-range order (Figure 3-15(B)). The strengthened hydrophobic interactions induced strong fluctuations of the monomers on the surface of the capsid, leading subsequently to a disordered structure which eventually broke apart. As shown in Figure 3-15(C), the fivefold symmetry axes were still, to some extent, conserved upon the increase of temperature, while the threefold symmetry axes disappeared entirely.

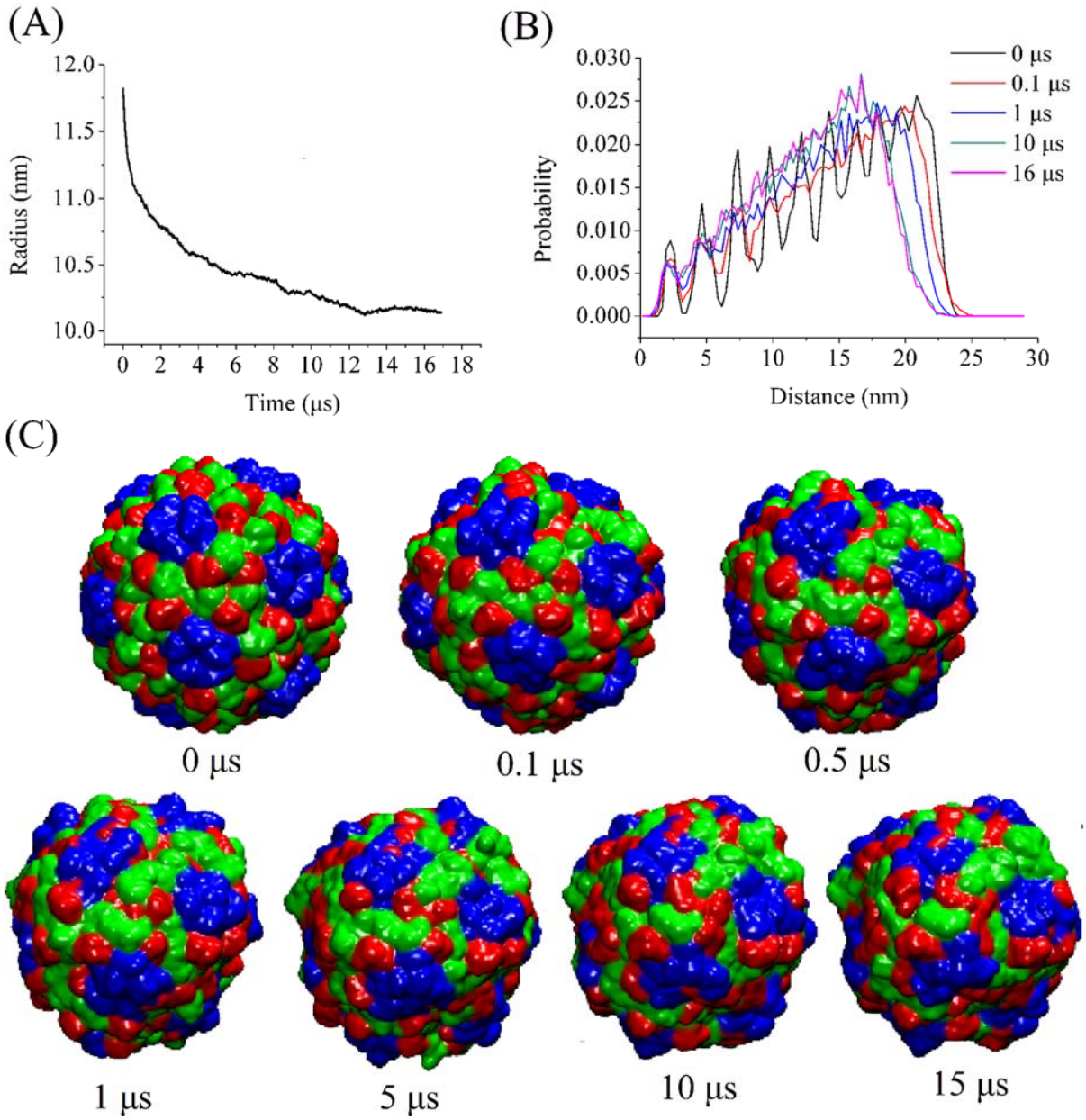

Figure 3-15. Shrinkage of CCMV capsid. Variations of the radius of the capsid at 400 K (A) and distributions of COM distances between monomers within a capsid at different time steps (B). Snapshots of conformation at different time steps (C). [37]

Discussion

Due to the time limit of MD simulation, we could only probe the evolution of the thermal disassembly in the early stage, which was consistent with observations in SANS experiments. By contrast, the evolution after the thermal shrinkage was unclear. According to the SANS experiments, the capsids continued to aggregate into a disordered cluster of capsid proteins and eventually disassembled, but the detailed pathway was difficult to extract. However, the thermal shrinkage phenomenon was not reported by the recent computational simulations on the disassembly of viral capsids performed by Wolek and Cieplak [25]. The reason might be due to the over-simplified force fields used for the interactions between the residues where the

temperature dependence was not properly taken into account. Moreover, the implicit solvent model in the simulations could not reflect the temperature dependence of the interaction as well since the hydrophobic interaction stems from the entropy of water molecules which is a function of their kinetic energy.

3.5 Conclusions

In this chapter, we captured the thermal dissociation of CCMV virions by SANS and we observed a slight shrinkage of the capsids upon heating at physiological pH. These findings point out the importance of the interaction strengths between capsid subunits and their dependence with temperature. We have therefore first implemented a temperature dependence of such interactions into a homogeneous lattice model that has been proven useful to study phase transition in viral capsids. The introduction of the temperature dependence on the hydrophobic interaction led to a lower and experimentally more realistic effective charge for CCMV capsid subunits regardless of the presence of genome.

In a heterogeneous lattice model, where two components were introduced to account for the asymmetric interaction between capsid subunits, we found a gradual and smooth dissociation process of empty capsids upon the increase of temperature. When the hydrophobic interaction between AB and CC subunits was much weaker than that between AB subunits, the capsid dissociated through a two-step process, in which CC subunits dissociated first at a lower temperature than AB subunits. In turn, the capsid dissociated in a sharp one-step process when the hydrophobic interactions were all comparable and both kinds of subunit dissociated simultaneously.

CG simulations have allowed us to probe the behavior of the whole capsid upon an increase in temperature. In qualitative agreement with experiments, we observed the shrinkage of the capsid due to the enhancement of hydrophobic interactions between proteins. Furthermore, the icosahedral symmetry was lost and the protein array became disordered.

3.6 Reference

[1] K. W. Adolph, Journal of General Virology **28**, 147 (1975).
[2] L. Lavelle, M. Gingery, M. Phillips, W. M. Gelbart, C. M. Knobler, R. D. Cadena-Nava, J. R. Vega-Acosta, L. A. Pinedo-Torres, and J. Ruiz-Garcia, Journal of Physical Chemistry B **113**, 3813 (2009).
[3] G. Tresset *et al.*, Archives of Biochemistry and Biophysics **537**, 144 (2013).
[4] J. M. Johnson, J. H. Tang, Y. Nyame, D. Willits, M. J. Young, and A. Zlotnick, Nano Letters **5**, 765 (2005).
[5] D. Law-Hine, A. K. Sahoo, V. Bailleux, M. Zeghal, S. Prevost, P. K. Maiti, S. Bressanelli, D. Constantin, and G. Tresset, Journal of Physical Chemistry Letters **6**, 3471 (2015).
[6] D. Law-Hine, M. Zeghal, S. Bressanelli, D. Constantin, and G. Tresset, Soft Matter **12**, 6728 (2016).

[7] G. Tresset, C. Le Coeur, J. F. Bryche, M. Tatou, M. Zeghal, A. Charpilienne, D. Poncet, D. Constantin, and S. Bressanelli, Journal of the American Chemical Society **135**, 15373 (2013).
[8] R. Tuma, H. Tsuruta, K. H. French, and P. E. Prevelige, Journal of Molecular Biology **381**, 1395 (2008).
[9] A. Zlotnick, R. Aldrich, J. M. Johnson, P. Ceres, and M. J. Young, Virology **277**, 450 (2000).
[10] E. E. Pierson, D. Z. Keifer, L. Selzer, L. S. Lee, N. C. Contino, J. C. Y. Wang, A. Zlotnick, and M. F. Jarrold, Journal of the American Chemical Society **136**, 3536 (2014).
[11] G. K. Shoemaker *et al.*, Molecular & Cellular Proteomics **9**, 1742 (2010).
[12] C. Uetrecht, I. M. Barbu, G. K. Shoemaker, E. van Duijn, and A. J. R. Heck, Nature Chemistry **3**, 126 (2011).
[13] M. Castellanos, R. Perez, P. J. P. Carrillo, P. J. de Pablo, and M. G. Mateu, Biophysical Journal **102**, 2615 (2012).
[14] M. Medrano, M. A. Fuertes, A. Valbuena, P. J. P. Carrillo, A. Rodriguez-Huete, and M. G. Mateu, Journal of the American Chemical Society **138**, 15385 (2016).
[15] Z. D. Harms, L. Selzer, A. Zlotnick, and S. C. Jacobson, Acs Nano **9**, 9087 (2015).
[16] P. Kondylis, J. S. Zhou, Z. D. Harms, A. R. Kneller, L. S. Lee, A. Zlotnick, and S. C. Jacobson, Analytical Chemistry **89**, 4855 (2017).
[17] M. F. Hagan and R. Zandi, Current Opinion in Virology **18**, 36 (2016).
[18] R. F. Bruinsma, W. M. Gelbart, D. Reguera, J. Rudnick, and R. Zandi, Physical Review Letters **90**, 248101 (2003).
[19] M. F. Hagan, in *Advances in Chemical Physics, Vol 155*, edited by S. A. Rice, and A. R. Dinner2014), pp. 1.
[20] G. R. Lazaro and M. F. Hagan, Journal of Physical Chemistry B **120**, 6306 (2016).
[21] H. D. Nguyen, V. S. Reddy, and C. L. Brooks, Nano Letters **7**, 338 (2007).
[22] P. Ceres and A. Zlotnick, Biochemistry **41**, 11525 (2002).
[23] A. M. Baptista, P. J. Martel, and S. B. Petersen, Proteins-Structure Function and Genetics **27**, 523 (1997).
[24] R. O. Soares, P. H. M. Torres, M. L. da Silva, and P. G. Pascutti, Journal of Structural Biology **195**, 216 (2016).
[25] K. Wolek and M. Cieplak, Journal of Physics-Condensed Matter **29**, 474003 (2017).
[26] S. Singh and A. Zlotnick, Journal of Biological Chemistry **278**, 18249 (2003).
[27] C. Yigit, J. Heyda, and J. Dzubiella, Journal of Chemical Physics **143**, 064904 (2015).
[28] D. Frenkel and B. Smit, *Understanding molecular simulation: from algorithms to applications* (Elsevier, 2001), Vol. 1.
[29] G. Tresset, J. Z. Chen, M. Chevreuil, N. Nhiri, E. Jacquet, and Y. Lansac, Physical Review Applied **7**, 014005 (2017).
[30] R. F. Bruinsma, M. Comas-Garcia, R. F. Garmann, and A. Y. Grosberg, Physical Review E **93**, 032405 (2016).
[31] J. N. Israelachvili, *Intermolecular and surface forces* (Academic press, 2011).
[32] W. K. Kegel and P. van der Schoot, Biophysical Journal **86**, 3905 (2004).
[33] J. Z. Chen, M. Chevreuil, S. Combet, Y. Lansac, and G. Tresset, Journal of Physics-Condensed Matter **29**, 474001 (2017).
[34] J. A. Speir, S. Munshi, G. J. Wang, T. S. Baker, and J. E. Johnson, Structure **3**, 63 (1995).
[35] J. R. Vega-Acosta, R. D. Cadena-Nava, W. M. Gelbar, C. M. Knobler, and J. Ruiz-Garcia,

Journal of Physical Chemistry B **118**, 1984 (2014).
[36] T. J. Dolinsky, J. E. Nielsen, J. A. McCammon, and N. A. Baker, Nucleic Acids Research **32**, W665 (2004).
[37] J. Z. Chen, Y. Lansac, and G. Tresset, Journal of Physical Chemistry B **122**, 9490 (2018).
[38] L. Monticelli, S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman, and S. J. Marrink, Journal of Chemical Theory and Computation **4**, 819 (2008).
[39] S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. de Vries, Journal of Physical Chemistry B **111**, 7812 (2007).
[40] T. Bereau, C. Globisch, M. Deserno, and C. Peter, Journal of Chemical Theory and Computation **8**, 3750 (2012).
[41] M. R. Machado, H. C. Gonzalez, and S. Pantano, Journal of Chemical Theory and Computation **13**, 5106 (2017).
[42] C. Globisch, V. Krishnamani, M. Deserno, and C. Peter, Plos One **8**, e60582 (2013).
[43] X. Periole, M. Cavalli, S. J. Marrink, and M. A. Ceruso, Journal of Chemical Theory and Computation **5**, 2531 (2009).
[44] H. J. C. Berendsen, J. P. M. Postma, W. F. Vangunsteren, A. Dinola, and J. R. Haak, Journal of Chemical Physics **81**, 3684 (1984).
[45] M. Parrinello and A. Rahman, Journal of Applied Physics **52**, 7182 (1981).
[46] I. G. Tironi, R. Sperb, P. E. Smith, and W. F. Vangunsteren, Journal of Chemical Physics **102**, 5451 (1995).
[47] B. Hess, H. Bekker, H. J. C. Berendsen, and J. Fraaije, Journal of Computational Chemistry **18**, 1463 (1997).

Chapter 4

Interactions between the molecular components of CCMV

The formation of a viral particle generally involves hundreds of proteins, making the assembly process intricate. Despite its intrinsic complexity, the production of a viral particle begins through the interactions between the basic assembly components. For the cowpea chlorotic mottle virus (CCMV), the first steps of the assembly process involve dimers of the capsid protein. Here, we carried out atomistic molecular dynamics (MD) simulations to investigate the initial assembly process of CCMV to get insight into the interactions at the molecular level.

4.1 Introduction

The assembly process of CCMV has been intensively investigated these last years, yet a number of issues remain unsolved, especially at the molecular level. The assembly of empty CCMV capsid is usually described by a nucleation-and-growth mechanism, where the assembly begins with the formation of a nucleus, and followed by the sequential addition of free dimers in the solution until the completion of a full capsid. The growth stage can be readily probed by experimental techniques [1-3], while less attention has been devoted to the nucleation stage because of its very transient nature.

The viral genome not only carries the genetic information, but also directs the assembly of viral particles. For CCMV in particular, the genome acts as a heterogeneous nucleus and effectively lowers the energy barrier to the capsid formation [4-6]. There is no doubt that the presence of the genome alters the assembly pathway [7] and makes it more complex. There are already several works on the influence of the genome on the formation of viral particles, notably through the charge ratio between the genome and the assembly subunits [8-10], or via the RNA secondary structure [11]. The interaction energy between the RNA genome and the subunits as well as the molecular processes taking place during assembly has been rarely investigated. The lack of these information causes difficulty on assigning proper values to the subunit-RNA and subunit-subunit interaction energies in many analytical models [4,12,13].

Molecular dynamics simulation is a helpful, and complementary method and has been widely used in the study of various viruses [14-17]. It can partly supply the missing information mentioned previously. To date, MD simulations have mainly focused on physical processes in subassemblies and capsids, including the interaction between capsid proteins and small molecules which might act as antiviral drugs [18-22], the properties of a whole capsid [14-

16,20,23-25] and the thermodynamic properties of the subunits [26]. However, investigations on the assembly of subunits are rather limited, even with coarse-grained (CG) methods.

In this chapter, we carry out all-atom MD simulations to shed light on the initial stage of the self-assembly of empty CCMV capsid and on the interaction strength between the relevant components. First, a single dimer with different salinities is investigated to obtain its equilibrium conformation, which is subsequently employed to study the interaction between a pair of dimers. We try next to elucidate the way dimers interact with RNA. Several simulations are conducted involving an infinitely long RNA rod and a short RNA strand. Meanwhile, the interaction strengths between different components are estimated via their rupture force determined by steered molecular dynamics (SMD) simulations.

4.2 Method

4.2.1 Force field

In this chapter, all MD simulations were performed with the Amber ff99sb-ILDN force field [27] which is an improved version of the Amber ff99SB force field. In the Amber ff99sh-ILDN force field, the side-chain torsion potentials had been improved and its expression can be written as

$$V_{AMBER} = \sum_{i>j} f_{ij}\left\{4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}\right\} + \sum_{bonds} k_b(r - r_0)^2 + \sum_{angles} k_a(\theta - \theta_0)^2 + k_0 + \sum_{dihedals}\sum_n k_d[1 + cos(n\phi - \phi_s)],$$

where k_0 is a new added constant.

In dozens of existing all-atom force fields, the Amber ff99 force field is famous for its remarkable capacity in reproducing the physicochemical properties of both protein and nucleic acid molecules. By improving the amino acid side-chain torsion potentials of the Amber ff99SB force field, we can have more accurate and realistic information, which is required in particular when comparing with NMR data.

4.2.2 Energy minimization

To release the inner stress in the initial conformations of molecules due to steric clashes or inappropriate geometry, an energy minimization is requested before performing MD simulations. Here we employed the steepest descent method (also known as the gradient descent method), one of the most popular minimization methods, for the energy minimization. To illustrate the principle of this method, we look for the minimum of the function $f(x) = x^2$ as a simple example. To find out the minimum of the function computationally, we need to conduct the following procedure.

1. Compute the gradient of $f(x)$, so that we have $\nabla f(x) = 2x$.
2. Calculate the new x by $x_n = x_{n-1} - \gamma\nabla f(x_{n-1})$.
3. Repeat the step 1 and 2 until we reach the required accuracy (that is, when the move $x_n - x_{n-1}$ is small enough). Then we obtain the minimum of $f(x)$.

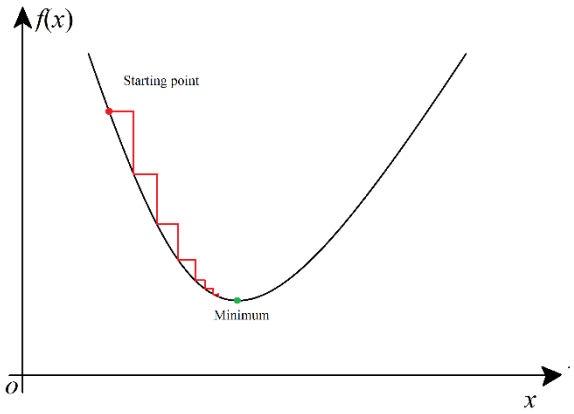

Figure 4-1. Illustration of the gradient descent on a function $f(x)$. The red line is the descent trajectory.

The process of finding the minimum of a function is shown on Figure 4-1. The steepest descent method is easy to implement in computer and thus becomes a widely used method in spite of some disadvantage inside it, such as a slow convergence, and the inability to locate a minimum in some cases [28].

In this chapter, the minimization step size of the steepest descent method is 0.01 nm and terminates automatically when the maximum force is smaller than 1000 kJ/mol/nm. The maximum number of minimization steps to perform is 50,000.

4.2.3 Constraint algorithm for bonds

In order to use a relatively large time step in the MD simulation we employed the LINCS (LINear Constraint Solver) constraint algorithm [29] for all bonds. The LINCS algorithm was developed by Hess, Bekker, Berendsen and Fraaije in 1997 based on the method of Edberg, Evans and Morriss (EEM) as proposed in 1986 and a modification thereof by Barayai and Evans (BE). The LINCS algorithm is compatible with the Leap-Frog and various other Verlet-type integration algorithms. It was reported that the LINCS algorithm is 3~4 times faster than the SHAKE constraint algorithm for the same accuracy. Thus, for a large simulation system, the LINCS algorithm is an advisable option.

In the LINCS algorithm, bonds are reset to their correct lengths after an unconstrained update. It is a non-iterative method, and always uses two steps per one time step, as illustrated in Figure 4-2. In the first step, the projections of the new bonds on the old bonds are set to zero. In the second step, a correction is applied for rotational lengthening, adjusting the position of atoms along the direction of the old bond so that the distance between atoms reaches the length of bond, as shown in the third picture in Figure 4-2. However, the LINCS algorithm becomes inefficient for simulations with angle constraint because these additional constraints greatly increase the computational complexity and therefore the computation time.

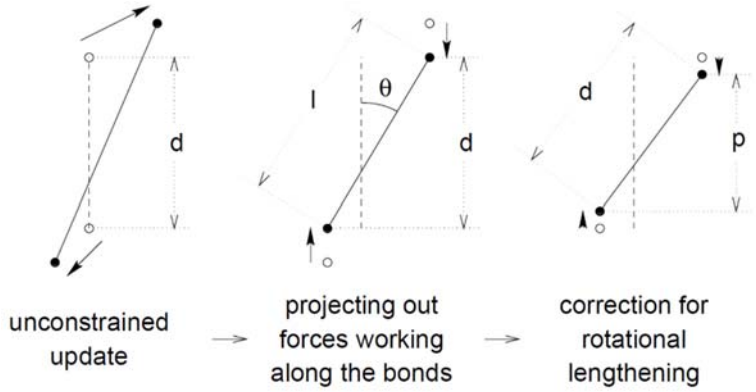

Figure 4-2. Schematic picture illustrating the procedure for updating the three positions for one time step in LINCS method. The dashed line is the old bond of length *d*, the solid line is the new bond. $l = dcos\theta$ and $p = (2d^2 - l^2)^{\frac{1}{2}}$ with θ the angle between the new bond and the old bond.

4.2.4 Integration algorithm

The MD simulations were integrated by the Leap-frog algorithm with a time step of 2 fs. To accelerate the simulation, a Verlet list with a grid method to search for neighbor particles was used in the simulation.

4.2.5 Simulation protocol

All simulations were performed with the GROMACS 5.1.4 simulation package [30], the atomistic Amber ff99sb-ILDN force field, and the TIP3P water model at pH 7. Simulations were carried out in an orthorombic box with a size large enough that the biomacromolecules in neighboring image boxes could not significantly interact with each other. The system was neutralized and adjusted to a specific ionic strength by adding NaCl salt. The simulations were conducted in NPT ensemble and the temperature was maintained at 300K by the Berendsen weak-coupling method [31] with coupling constants τ_T = 0.1 ps while the pressure was kept at 1 bar by the Parrinello-Rahman barostat [32] with time constant τ_P = 2.0 ps. For the nonbonded interactions, the short-range van der Waals and Coulomb interactions were cut off and shifted at 1.0 nm while the long-range Coulomb interactions were calculated by the Particle Mesh Ewald (PME) method [33] with a grid spacing of 0.16 nm. A time step of 2 fs was applied throughout the simulations, and all bonds were constrained by the LINCS algorithm.

4.3 Single dimer

4.3.1 Simulation protocol

The initial structures used in this article were extracted from the CCMV capsid crystal structure

resolved by Speir et al. [34] (PDB: 1CWP). There are three kinds of protein in this structure noted A, B and C. However, several residues making up the N-terminal tail of the proteins could not be resolved by X-ray crystallography (25 residues for proteins A and C, 39 residues for protein B) due to their flexibility. Considering that the charged N-terminal tails have an indispensable role in the interaction between dimers as reported by experiments [35], the structure of the missing residues were generated in the same way as in Ref [36] and manually added to the PDB structure by a package in VMD [37]. The modified structure for a single dimer is shown in Figure 4-3.

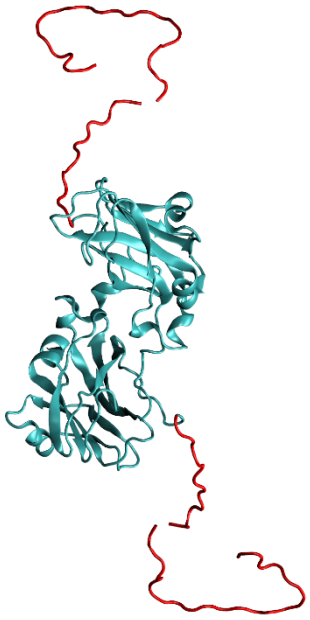

Figure 4-3. The modified conformation of a single dimer. The N-terminals (the first 40 residues) are colored in red. The broken short chains are the missing residues that cannot be resolved by X-ray crystallography, which are added artificially here.

The modified structure of a single dimer was used as the initial conformation to probe the behavior of a dimer in water. The effect of ionic strength on the conformation of dimer was analyzed by carrying out three independent simulations at different NaCl concentrations (0.1 M, 0.2 M and 0.5 M). An energy minimization was performed, followed by a 100 ps NVT run and a 100 ps NPT equilibration run. Finally each subsequent production run lasted for 500 ns. The relative separation and orientation of the proteins in a dimer were analyzed. The separation and the orientation were defined as Figure 4-4. Here the distance between the center of mass (COM) of the proteins was used to define the separation. Considering the stability of the beta-sheet structure of the protein in various environments, it was feasible to use the angle between the two proteins of a dimer to depict their orientation. To simplify, the orientation was defined as the angle between the nearest strands of the beta-sheet structures (the two strands highlighted in red) in the two proteins.

Root mean square deviation (RSMD) was calculated for both the N-terminal tail and the body of the protein (the rest of the residues excluding the N-terminal tail) independently by aligning the simulated structure with its initial conformation.

The radius of gyration of the N-terminal tails was determined using the GROMACS

analysis tools over the last 300 ns trajectory. Considering the fact that each dimer contains two N-terminal tails, the final value for the radius of gyration was the average of the two tails.

In addition, the contact map between the whole series of residues of a dimer was constructed by the GROMACS mdmat tools which established a distance matrix consisting of the smallest distance between residue pairs within 1.5 nm.

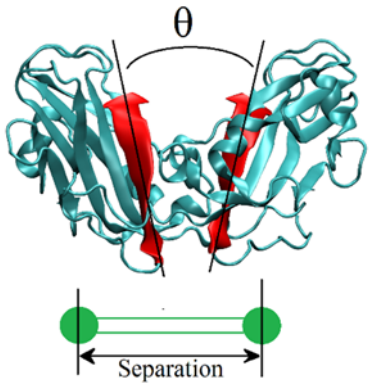

Figure 4-4. Definitions of the separation and the orientation of the proteins in a dimer. The COM of the proteins is represented with green particles at the bottom. The orientation between the two proteins in a dimer is determined by the angle between the nearest strands in the beta-sheet structure of the protein, which is highlighted in red.

4.3.2 Results

The conformations of a single dimer at different ionic strengths are displayed in Figure 4-5(A) where the N-terminals of dimers are colored in red. According to the calculations of the root mean square deviation (RMSD) in Figure 4-5(B) where both the positively charged N-terminal tails and the body of a dimer presented a structural change at different levels, the conformation of the body of the dimer remained stable during the simulation at all ionic strengths, which was in sharp contrast with a large conformational adjustment of the N-terminal tails.

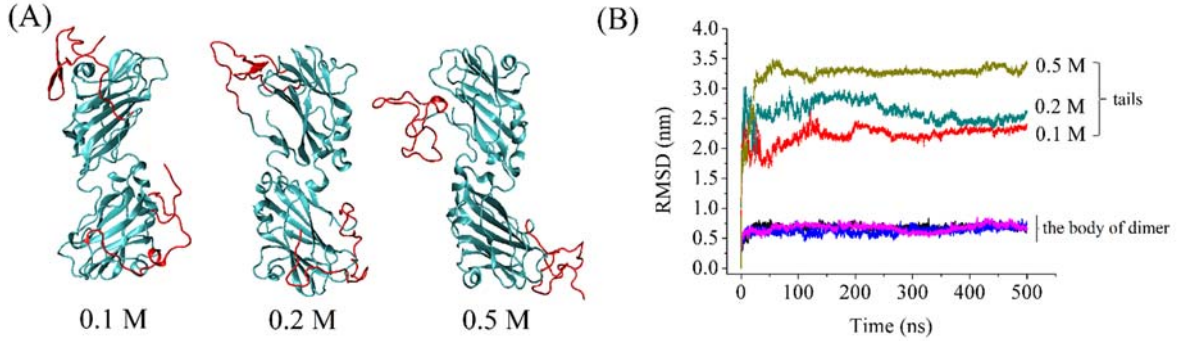

Figure 4-5. Conformations of a dimer for salinities of 0.1 M, 0.2 M and 0.5 M. (A) Snapshots of the conformations of a dimer. The tails of a dimer (residues 2-39) are in red. (B) RMSD of the tails (summed up over the two tails for each dimer) and of the body of a dimer with respect to their initial conformation at different ionic strengths.

To visually reveal the difference in the conformation, we directly fitted the simulated conformations of a dimer at different salinities to the native one and the conformation

superimposition is displayed in Figure 4-6(A). Compared to the native conformation which was taken from full capsids determined by X-ray crystallography, the hydrophobic cores remained stable in spite of the variation of salinities, while the orientation between the two proteins in the same dimer demonstrated a significant difference. The angle between the proteins in a dimer was more open than it was in the native conformation and the salinities had a weak influence on that phenomenon. As indicated on Figure 4-6(B) and (C), the orientation angle sharply increased from 37° to 81°, meanwhile the separation increased from 2.8 nm to 3.4 nm. This phenomenon was observed at all studied ionic strengths, suggesting that the structure of the free dimer was looser than within a capsid. The difference in the orientation might originate from the difference in the local environment where the dimer stayed.

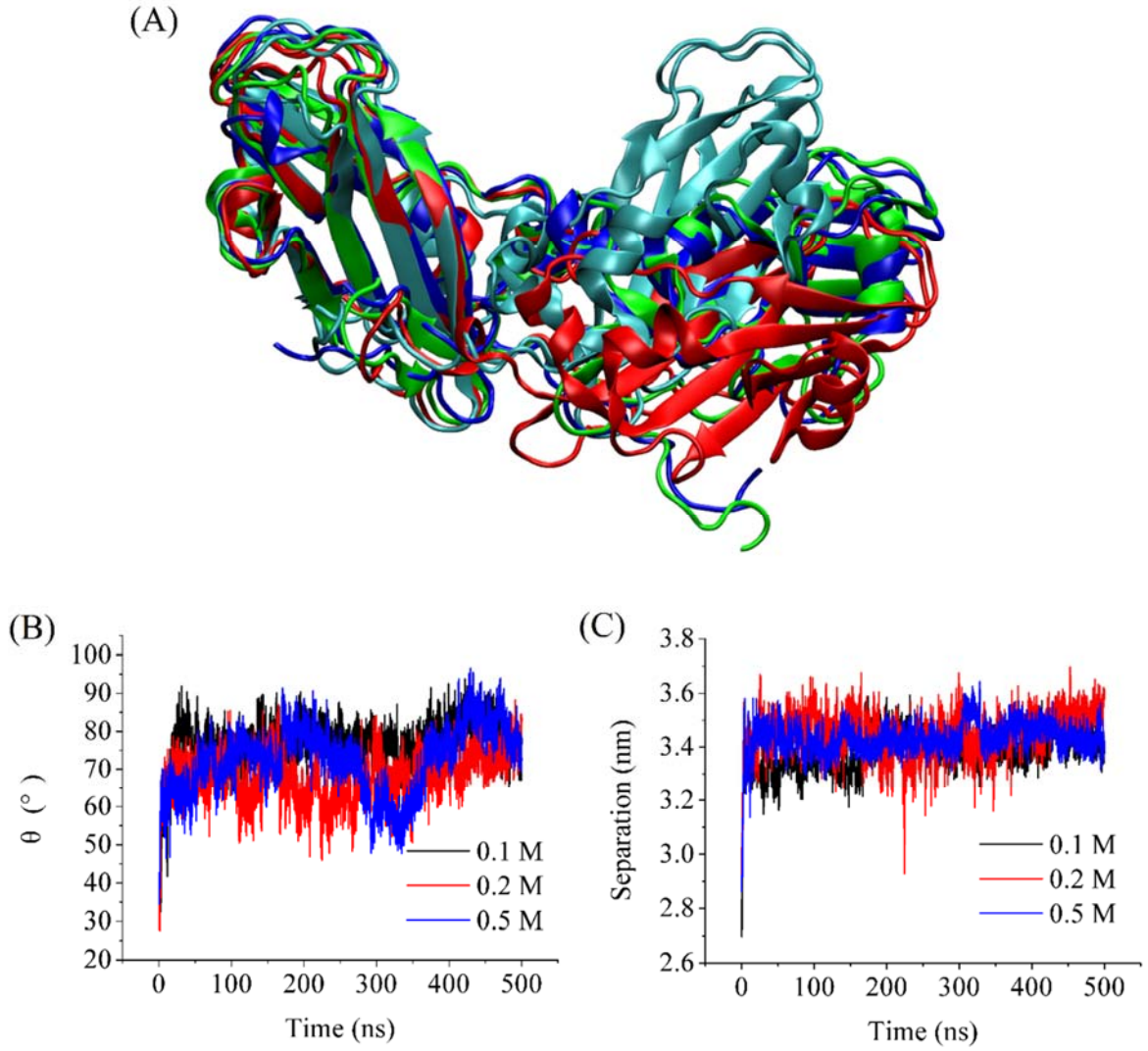

Figure 4-6. (A) Superimposition of the final conformations of the dimers at NaCl concentrations of 0.1 M (blue), 0.2 M (red), and 0.5 M (green), with the native (X-ray crystallography) conformation (cyan) after the 500-ns production run. All conformations are aligned by superimposing one protein of the dimer. The N-terminal tails are not shown. Evolutions of the angle θ (B) and the center of mass separation (C) between the two proteins within a dimer. The definitions of the angle θ and the separation are depicted in Figure 4-2.

As revealed by the RMSD, the N-terminal tails are more sensitive to salt concentration. To quantify this sensitivity, we firstly measured the radius of gyration of the N-terminal tails as a function of salinity and the measurements are illustrated in Figure 4-7. An obvious tendency was that a higher salinity led to a smaller radius of gyration. A smaller radius of gyration indicates a more compact conformation for the tails of dimers.

The contact map of the N-terminal tails with the bodies was calculated and plotted in Figure 4-8. Salinity had a strong impact on the conformation of the tails. At low salinities, the N-terminal tails presented a stretched conformation and easily bound on the mainly negatively charged body, as shown in Figure 4-5(A). There was not a specific part of the tails bound on the body, instead, the residues in contact were distributed evenly along the tails as indicated by Figure 4-8. In addition, it is surprising to observe that part of the N-terminal tails even covered the β-sheets where hydrophobic residues concentrate, leading to a decrease in the available hydrophobic area (see snapshot for 0.1 M in Figure 4-5(A)). Interestingly, this phenomenon disappeared by increasing the salinity up to 0.5 M where the N-terminal tails became more compact (Figure 4-7) and less likely to bind on the body (Figure 4-8). The response of the N-terminal tails to the salinity suggested that the change in the ionic strength might not only affect the electrostatic interactions between dimers, but also led to the adjustment of their conformation, especially that of the N-terminal segments, and subsequently altered the assembly pathway.

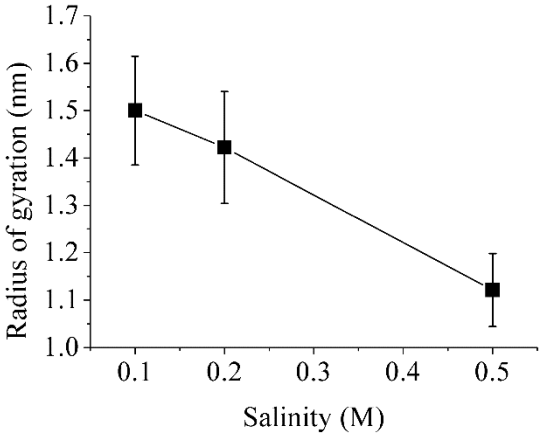

Figure 4-7. Radius of gyration of the N-terminal tails of dimers versus salinity. The error bars are calculated by averaging the radius of gyration over the two tails of the dimer.

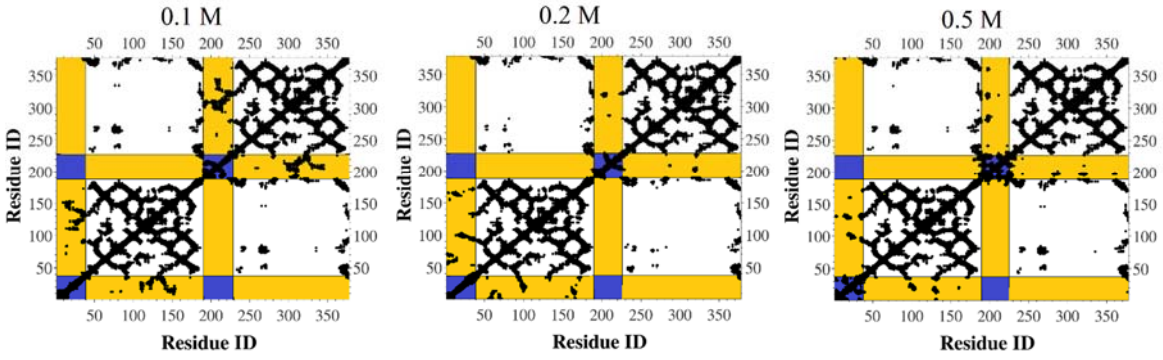

Figure 4-8. Contact map between the residue pairs in a dimer at three salinities. The yellow areas are the contacts between a residue in the tails (residues 2-39) and a residue in the body

(residues 40-190) while the blue and white area are the contacts between residues belonging to the tails only and to the body only, respectively.

With the MD simulations above, we can have some basic understanding of the conformation of dimers in an aqueous solution: the body of dimers is able to conserve its conformation except that the angle between the two proteins may increase but is not sensitive to the ionic strength of solution. By contrast, the N-terminal tails of the dimers exhibit diverse conformations which greatly depend on the ionic strength. The N-terminal tails cause the burying of hydrophobic domains at different levels, which might have a significant impact on the association of dimers.

4.4 Self-assembly of a dimer pair

The full understanding at an atomistic level of the interactions between two dimers is imperative to unveiling the self-assembly mechanisms behind the formation of viral capsids. The low spatial resolution inherent to most experimental techniques is far from being enough to gain insight into the interaction details during the assembly process. To date, the conformation of the N-terminal tails in the CCMV dimer is still unclear, even more so, when a dimer interacts with another one or with the genome. In this section, we focused on the interactions between a CCMV dimer pair, which is a very first step but also an indispensable one to comprehend the whole assembly process.

4.4.1 The native conformation of a pair of dimers

Simulation protocol

Two connected dimers including two proteins A and two proteins B were taken out from a full CCMV capsid and used as the initial conformation (denoted as native conformation) after the addition of the missing residues (residues 2-25). The modified conformation was put into a cuboid box with a margin of at least 3 nm along each direction, then the water molecules of TIP3P were filled into the box and the system was neutralized and adjusted to a salinity of 0.2 M by adding NaCl. The system subsequently underwent a routine equilibration consisting of an energy minimization by the steepest descent minimization algorithm with a minimization step size of 0.01 nm and a maximum force tolerance of 1000 kJ/mol/nm, a 100-ps NVT equilibration, and a 100-ps NPT equilibration. The temperature and pressure were constrained at 300K and 1 bar, respectively. After equilibration, a 500-ns production simulation was run.

In addition, another MD simulation over a pair of dimers without N-terminal tails (the first 41 residues were removed) in the native conformation was performed. The simulation procedure was the same as the simulation for the complete dimer.

The contact informations of the dimer pair were analyzed, including the orientation between the dimers, the separation and orientation between the proteins in each dimer (see the definition in Figure 4-4).

Results

Before we probe the interaction between a pair of dimers, we should have some basic understanding on the conformation of the dimer pair associating each other in a native form (the conformation in a full capsid). In order to investigate the stability of this conformation in an aqueous environment, we performed a 500-ns MD simulation in which a pair of dimers was associated in its native conformation and the missed N-terminal tails were added to the dimers. The MD simulation demonstrated that the dimers would not separate in the aqueous environment as reflected by an almost constant distance between the contacting proteins of the dimer pair (stabilizing at ~ 2.83 nm) as shown in Figure 4-9(A). Interestingly, the orientation angle of the dimers in the pair underwent an instantaneous change at the beginning of the simulation, decreasing from ~ 70° to ~ 50°. After this fast adjustment of the orientation, the angle was not entirely constant, but instead swung around 50° with an amplitude of 10° (see the inset in Figure 4-9(A)). In the other hand, the conformation of the dimer also underwent a relaxation. Figure 4-9(B) depicts the evolution of the separation between the centers of mass of the two proteins in each dimer of the dimer pair. The separations rose up from ~2.9 nm to ~3.5 nm shortly after the beginning of the simulation (Figure 4-9(B)). However, in overall the separation for dimer 1 was larger than that for dimer 2 due to a constrained force from the protein 2 over protein 3 and 4, which dragged protein 3 closer to protein 4 (see Figure 4-9(B) for the definition of proteins 1 to 4 and dimers 1 and 2). That phenomenon could also be confirmed by the orientation angle θ of the proteins in a dimer in the inset of Figure 4-9(B) where dimer 2 demonstrated an apparently smaller orientation angle than that of dimer 1 because of the existence of the constrained force from protein 2 (the peak value of the orientation angle for the isolated dimer at 0.2 M NaCl is ~ 68°, close to the value of dimer 1).

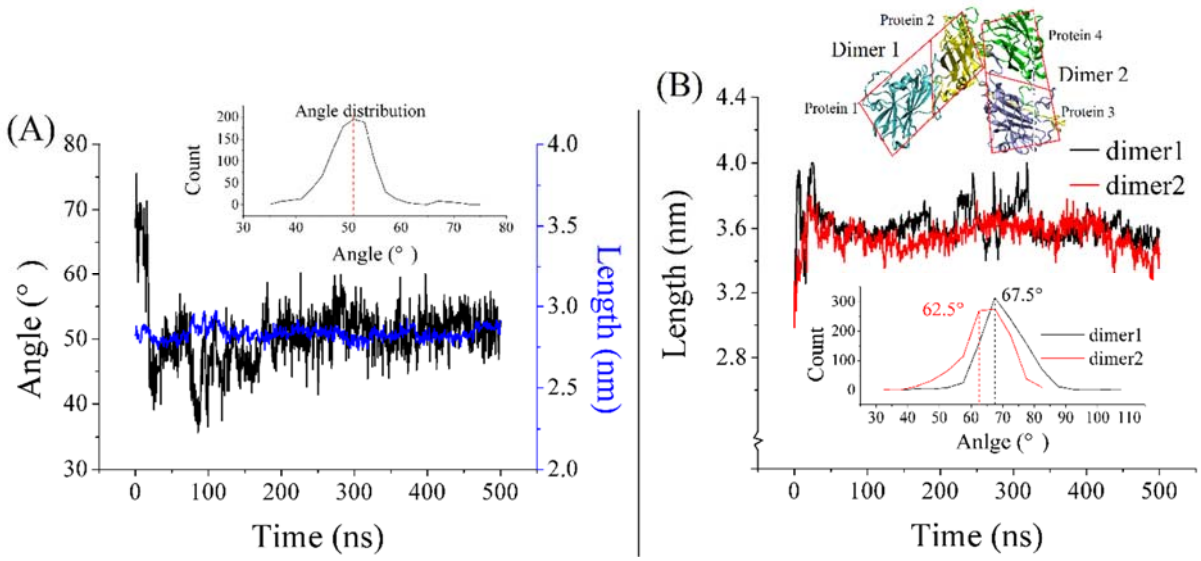

Figure 4-9. Evolution of a pair of dimers initially in their native conformation. (A) Variation of orientation angle and contact distance in a pair of dimers. The inset gives the distribution of the orientation angle. (B) Variation of the distance between the two proteins in each dimer. The inserted snapshot is the conformation of the dimer pair at the end of simulation where the proteins are colored differently and the dimers and proteins are separated with red lines. The inset below the curves is the distribution of the orientation angle θ of the proteins in each dimer.

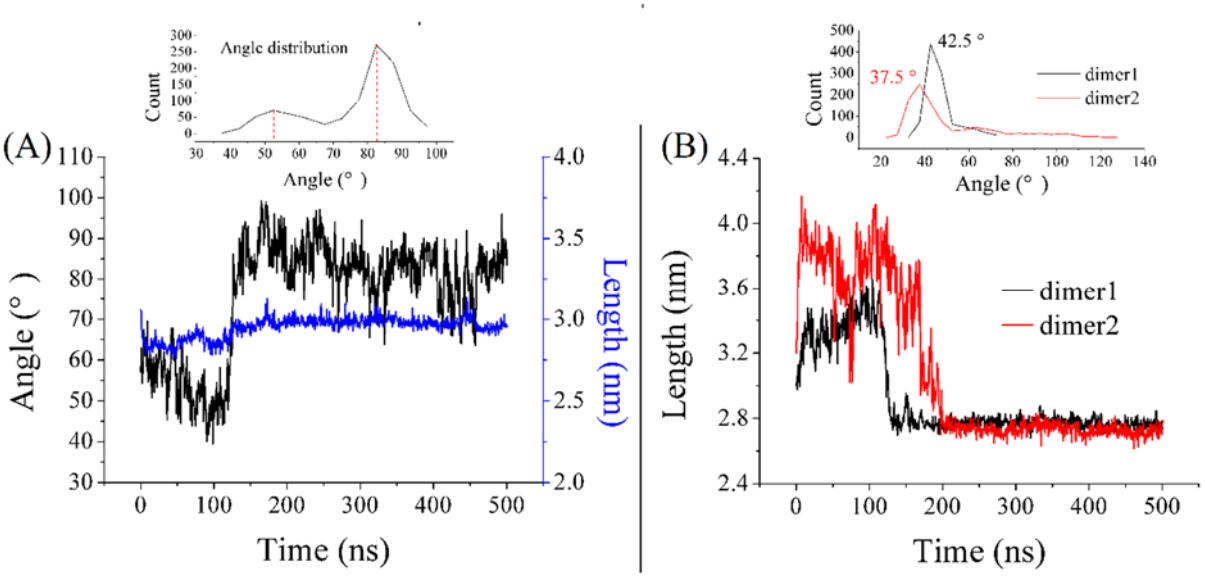

Figure 4-10. Evolution of a pair of dimers without N-terminal tails starting from its native conformation. (A) Variation of orientation and contact distance of a pair of dimers. The inset shows the distribution of the orientation angle. (B) Variation of the distance between the two proteins in each dimer of a pair. The inset below the curves gives the distribution of the orientation angle θ of the proteins in each dimer.

For the sake of comparison, we performed another MD simulation with a pair of dimers without N-terminal tails starting from a native conformation again. After removing the positively charged tails, the association of the dimers became slightly unstable, as revealed by a stronger fluctuation on contact distance and orientation angle in Figure 4-10(A). Interestingly, the orientation angle leaped from ~52° to ~82° instead of decreasing when the tails are present, as shown in the inset in Figure 4-10(A). In addition, the swing amplitude was larger, implying a weaker constraint from the other dimer. Unexpectedly, the proteins in each dimer sticked together in a aqueous environment without the protection of the N-terminal tails, indicated by a smaller separation between the proteins in a dimer (decreasing from over 3 nm to 2.8 nm, see Figure 4-10(B)) and a much smaller orientation angle between the proteins (~40° for the cleaved dimer versus over 60° for the complete dimer, see the insets in Figure 4-10(B) and Figure 4-9(B), respectively). There is no doubt that the removal of the N-terminal tails weakens the electrostatic interaction responsible for the repulsion between proteins, which in overall is equivalently expressed as a strengthening of the hydrophobic interaction. Thus, the orientation angle of the proteins in a dimer was getting smaller to further conceal the hydrophobic residues from the aqueous environment resulting in a weaker interaction strength.

4.4.2 Assembly of a pair of dimers

Simulation protocol

The equilibrated conformation of a single dimer in the previous section was employed as the initial conformation of dimers. Two dimers were aligned in parallel with a separation of 7 nm and the simulation box was set to 14×14×14 nm. Then the system was neutralized and adjusted

to a NaCl concentration of 0.2 M. Subsequently, a series of routine equilibrations were performed namely an energy minimization by steepest descent minimization algorithm, a 100-ps NVT equilibration, and a 100-ps NPT equilibration. The temperature and pressure were maintained at 300 K and 1 bar, respectively. Finally a 2.125 μs-production run was performed. Then the temperature of the system was increased to 400K in one step and kept for 1 μs to examine the stability of the structure formed at 300 K. The stability of the conformation of dimers was analyzed by calculating the RMSD of the body of monomers.

To investigate the effect of important factors on the assembly, additional three MD simulations were performed. The initial conformation of these three simulations was the same as the previous one. The first simulation was devoted to the assembly at NaCl concentration of 0.5 M and 300 K. This simulation lasted 1 μs. The second simulation dealt with the assembly of a pair of dimer without N-terminal tails (namely residues 1-40 removed) at NaCl concentration of 0.2 M and 300 K. This simulation lasted 415 ns. The last simulation was about the assembly of a pair of dimers with N-terminal tails partially cleaved (residues 1-25) at NaCl concentration of 0.2 M and 300 K. The simulation lasted 1 μs.

The evolution of the orientation and distance between the dimers was analyzed during the assembly.

Calculation of PMF by umbrella sampling

The first simulation was performed at a salinity of 0.2 M and the production run lasted 2.125 μs at 300K. The association energy between two dimers was determined by calculation of the potential of mean field (PMF). The PMF was calculated by umbrella sampling method and analyzed by the Weighted Histogram Analysis Method (WHAM) [38]. The details of the methodology can be found in the next paragraph, where one can find the initial conformation used to produce the conformations for sampling windows (Figure 4-11) and umbrella histograms (Figure 4-12).

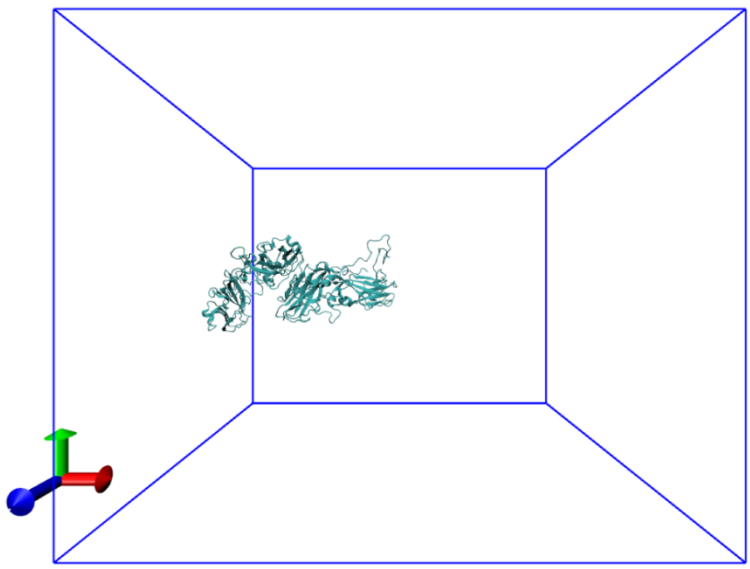

Figure 4-11. Initial conformation for the pulling simulation of a pair of dimers.

A pulling simulation on the final conformation of a pair of dimers at 300 K (see Figure 4-

11) was performed to generate the initial conformation for umbrella sampling. One dimer was pulled away from the other one along the x-axis over 400 ps, using a spring constant of 1000 kJ $mol^{-1}nm^{-1}$ and a pull rate of 0.01 nm ps^{-1}. 32 windows with separation of 0.15 nm apart were extracted from the trajectory as initial conformations for the umbrella sampling simulations. Each window was equilibrated at 300 K and 1 bar for 100 ps before a 10-ns production run was performed. During the sampling, the separation between two dimers was constrained by a spring constant of 1000 $kJ.mol^{-1}.nm^{-1}$. The final umbrella histograms for each window are shown in Figure 4-12. The Weighted Histogram Analysis Method (WHAM) was used to extract the PMF.

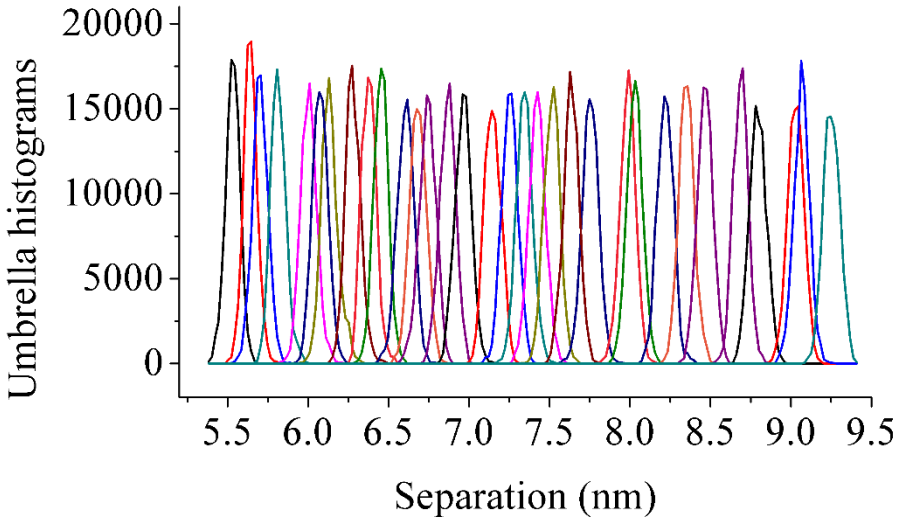

Figure 4-12. Umbrella histograms as a function of the separation of two dimers.

Results

The evolution of the separation and orientation between dimers at an ionic strength of 0.2 M are depicted in Figure 4-13(B). The dimer pair experienced a severe re-orientation before finding a "stable" bound conformation with an orientation angle between the dimers fluctuating around an angle of 120° and a separation between the centers of mass (COM) of the contacted proteins around 4 nm (see Figure 4-13(B)). The simulated conformation demonstrated that the N-terminal tails played a bridging role by connecting the two dimers in this case. However, this conformation was very different from the native conformation, whose orientation angle and COM separation were 60° and 2.9 nm respectively (Figure 4-13(A)). The difference was clear as indicated by the superimposition of the simulated conformation and the native conformation in Figure 4-16.

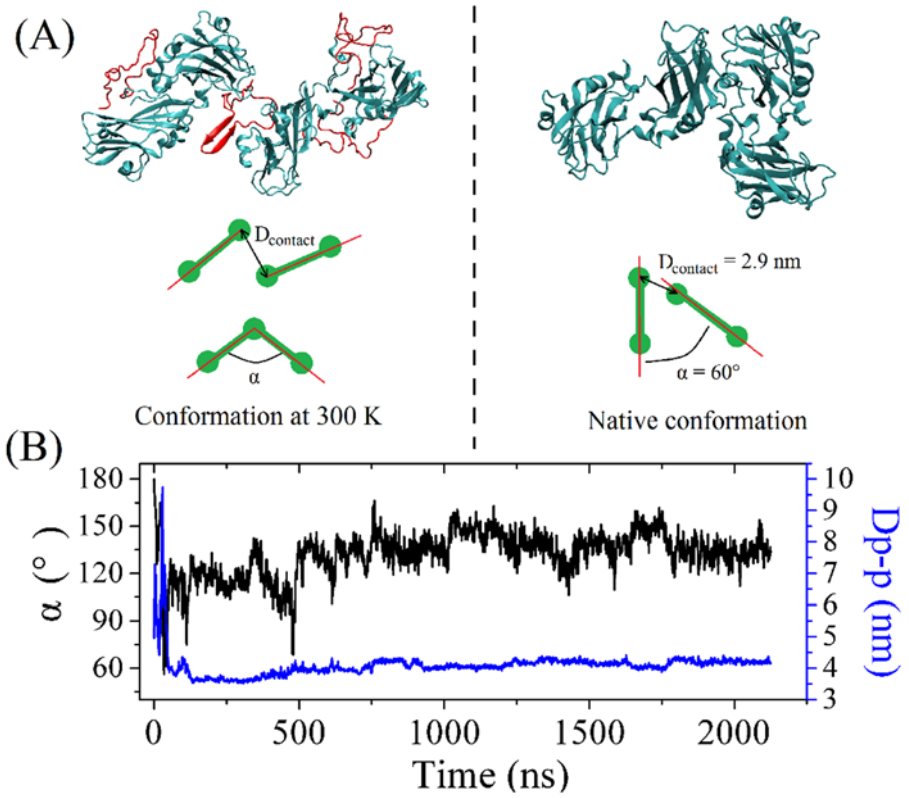

Figure 4-13. Simulations of a pair of dimers at an ionic strength of 0.2 M. (A) Snapshot of a dimer pair at the end of the simulation at 300K (left) and the native conformation of two dimers in the capsid (right). The rod-beads models below the snapshots illustrate the separation between the two dimers $D_{p\text{-}p}$ and orientation angle α. (B) Evolution of α and $D_{p\text{-}p}$ at 300K.

The association energy between two dimers was reported to range from -4 to -10 $k_B T_0$ [39-41], where k_B is the Boltzmann constant and T_0 is the room temperature. The association free energy was determined by calculating the PMF between the COM of the two dimers while their relative orientation was free. The calculation was carried out from the final conformation obtained at 300K. As shown in Figure 4-14, the association free energy of a pair of dimers was -8 $k_B T_0$, i.e., within the range reported experimentally. This finding makes it plausible our hypothesis that a strong association energy between the dimers leads to kinetically trapped conformations.

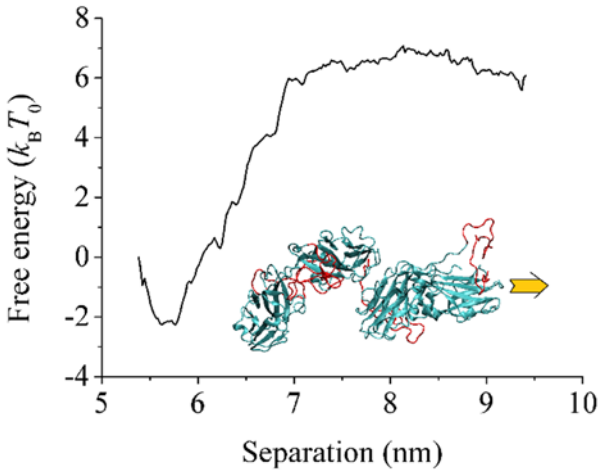

Figure 4-14. Potential of mean force (PMF) calculated by starting with the conformation obtained at the end of the simulation at 300K and 0.2 M salt concentration. The inserted snapshot points out the pulling direction.

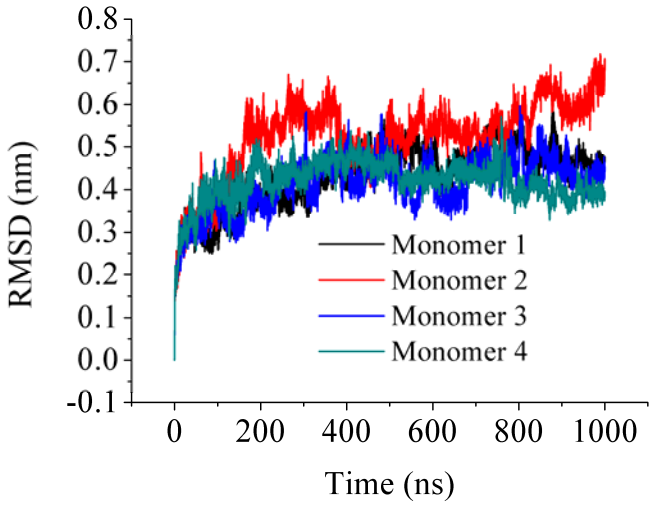

Figure 4-15. Root mean square difference (RMSD) of the body of each monomer in the dimer pair at 400K with respect to the final simulated conformation at 300K. It can be seen that the structure of the body remains stable for each monomer.

In an attempt of overcoming the energy barrier, the conformation simulated at 300K was brought to 400K for 1 μs in order to allow for larger energy fluctuations. After the increase of temperature, the RMSD of the body of dimers demonstrated a significantly small variation (~0.5 nm for all proteins, see Figure 4-15), indicating that the dimers were stable at 400 K. As shown in Figure 4-16, an immediate conformational adjustment occurred with the orientation angle decreasing from 120° to 90° and a closer contact formed (the separation goes from ~4 nm to ~3.5 nm). The snapshots in Figure 4-16 depicts a superimposition of the simulated conformation with the native conformation showing that the conformation obtained at 400K was closer from the native one even though some discrepancies persisted. It also demonstrates that the tails occupying the hydrophobic domains inhibited the access of other dimers, increasing the difficulty to reproduce the native conformation.

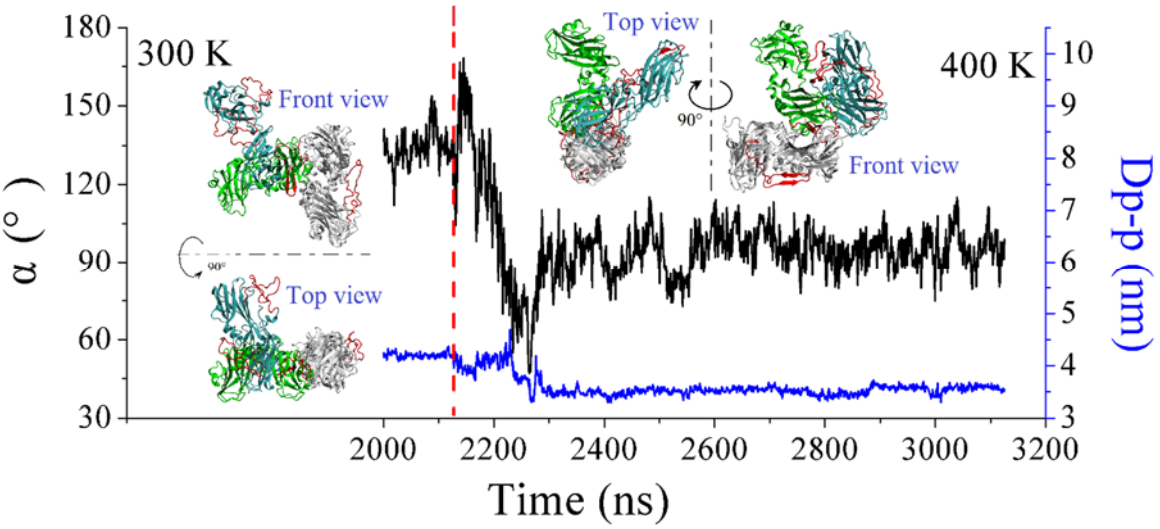

Figure 4-16. Evolution of the separation between the two dimers $D_{p\text{-}p}$ and orientation angle α at 300K (left of the red dot-dashed line) and at 400K (right of the red dot-dashed line). The inserted snapshots are the superimposition of the simulated conformation (in cyan for the body, in red for the tails, in gray for the aligned dimer) and the native conformation in the capsid (in green, tails are not shown) at 300K and 400K, from two different viewing angles, front view and top view.

Effect of ionic strength

The electrostatic interaction between dimers is important for the assembly of viral capsids, as reflected by the disassembly and reassembly capability of the capsids through modulation of the salinity and pH. Given that experiments indicated that the assembly of capsids is favorable at high ionic strengths, we performed a simulation on the assembly of a pair of dimers initially oriented parallel to each other with an initial separation distance of 12 nm and at a relative high NaCl concentration of 0.5 M. The evolution of the orientation angle and separation between the dimers was monitored and shown in Figure 4-17(A). Unlike the assembly at 0.2 M, the dimers associated together quickly after 300 ns, and the orientation angle was stable and fluctuated around 90°. Hereafter, the binding of the dimer pair underwent a slight conformational reorganization. Even though both orientation angle and contact separation were still different from the native ones, as shown in Figure 4-17(B), the difference became smaller than that at 0.2 M. By superimposing the simulated conformation with the native one, we can found that we were able to reproduce the native conformation simply via turning the dimer in cyan to a particular angle to arrive to the location of the dimer in green along the dimer in gray, as indicated by the bottom snapshot of Figure 4-17(B) where the arrow indicates the rotation angle.

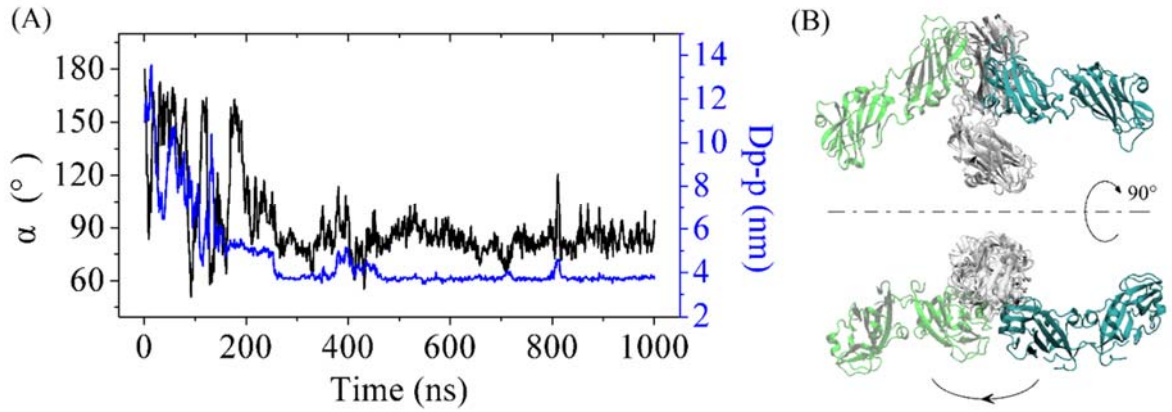

Figure 4-17. (A) Evolution of the separation $D_{p\text{-}p}$ and orientation angle α between the two dimers at 300K and NaCl concentration of 0.5 M. (B) Superimposition of the simulated and the native conformations in the capsid. The N-terminal tails are not shown.

There is no doubt that the electrostatic interaction between dimers is screened by the strengthening of ionic strength, which, on the other hand, increases the association energy of the dimers. Thus, the assembly of viral capsids is principally favored at a higher ionic strength. Even though there have been no experiments supporting this hypothesis at a neutral pH, the experimental observations at pH 4.8 are, to some extent, consistent with the simulation results.

Effect of the steric hindrance of the N-terminal tails

As indicated above, the tendency of the N-terminal tails to bind to the body of dimer causes a steric hindrance for the access of other dimers to the hydrophobic domains, which might mislead the assembly. Based on this inference, the assembly of dimers without N-terminal tails should be easy and fast. To test this assumption, we conducted MD simulations with a pair of initially parallel dimers, whose N-terminal tails were partially (residues 1-25) or entirely (residues 1-39) removed. The evolution of the orientation and separation between the dimers are displayed in Figure 4-18 and Figure 4-19.

For a dimer whose N-terminal tails were partially cleaved, the orientation angle was fluctuating around ~80° while contact separation was stabilized at ~3.6 nm, suggesting that the native conformation for a pair of dimers could not be reproduced. Meanwhile, the assembly of a dimer pair with partially cleaved tails did not show any significant difference from that with a complete residue sequence. However, as shown by the snapshot obtained after 1 μs in Figure 4-18, the partially cleaved tails guided the dimer into an abnormal location where one end accessed the middle of another dimer (T-shaped configuration) and the dimer pair contacted each other by one or two proteins (one of the two proteins in a dimer contacts with the two proteins of the another dimer). This unusual binding was quite different from the assembly of complete dimers. It suggested that the N-terminal tails played a significant role in guiding the assembly into a correct way [2].

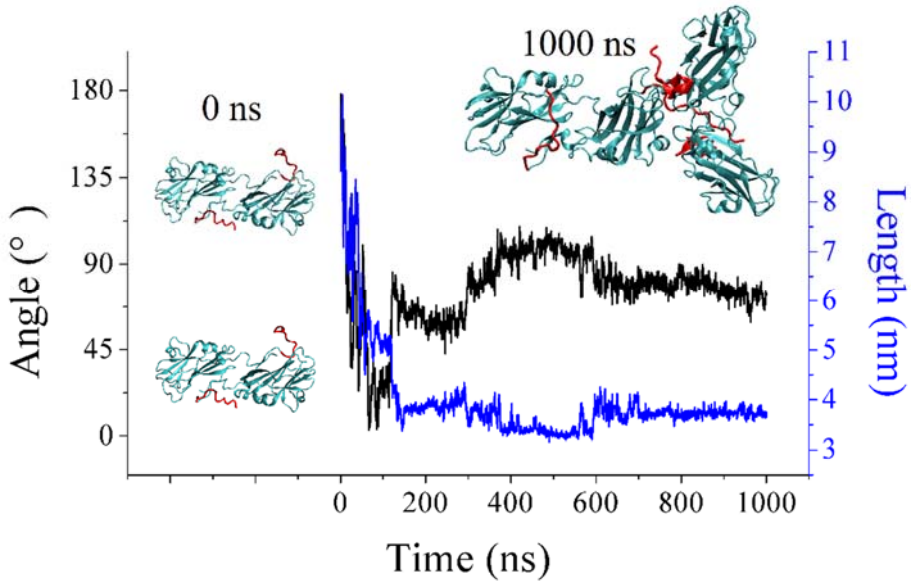

Figure 4-18. Evolution of the assembly of a pair of dimers with N-terminal tails partially removed. The inserted snapshots are the initial and simulated conformations of dimers. Residues 26-39 are in red.

Compared with the assembly of partially cleaved dimers, the assembly of entirely cleaved dimers demonstrated a significant difference. As shown in Figure 4-19, two cleaved dimers associated together shortly after the beginning of the simulation (around ~50 ns) and the association underwent a limited variation afterwards. The final orientation angle was ~90° with an amplitude of fluctuations of 20°, and the contact separation was stable at ~3.9 nm. Unlike the assembly of partially cleaved dimers, the dimer pair contacted each other through an end-to-end model, which is a common contact pathway for the assembly of dimers with complete residues. As mentioned in chapter 1, the dimer charge is negative on its outer surface but positive on its inner surface (see Figure 1-19). By entirely removing the N-terminal tails, the assembly of the cleaved dimers lost the guiding role of the N-terminal tails, and therefore, the association of dimers took a relatively random orientation. In addition, the hydrophobic interaction between dimers was strengthened while the electrostatic interaction weakened, which resulted in a stable binding after the first contact of the two dimers.

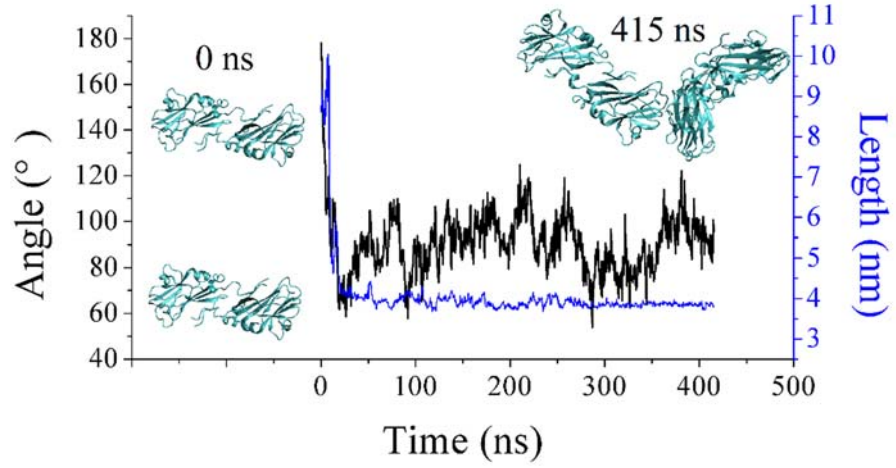

Figure 4-19. Evolution of the assembly of a pair of dimers with N-terminal tails entirely removed. The inserted snapshots are the initial and simulated conformations of dimers.

According to the simulations above, removing the N-terminal tails did not allow the assembly of dimers to reproduce the native conformation. There are several possible reasons for this failure. First of all, the assembly of cleaved dimers led to abnormal binding location in the absence of the guiding function of the N-terminal tails. Secondly, the hydrophobic interaction was strengthened to some extent while the electrostatic interaction was weakened after removing the N-terminal tails. Thus, the assembly of cleaved dimers gave rise to some misassembled viral capsids, as revealed by experimental observations [2].

Discussion

Despite the difficulties exhibited by the previous simulations to reproduce the assembly process of a pair of dimers into their native conformation in the full capsid, we can still extract some information on the interactions between them. There are several points to summarize on the assembly of a dimer pair. Firstly, the N-terminal tails of dimers play both a positive and a negative role on the assembly of dimers. On the one hand, the N-terminal tails assist the assembly of dimers via contributions on the electrostatic attraction between dimers at long distance and the guidance functions at close distances. On the other hand, the N-terminal tails cause steric hindrance preventing the dimer to accessing the hydrophobic domains and reaching the most stable conformation eventually. Secondly, when two dimers are getting closer, one dimer accesses to another one via its loop structure on the hydrophobic body. The residues on this loop are generally polar, suggesting a versatile capability of binding with any kind of charged and polar residues.

Given that we had performed all-atom MD simulations of the assembly of a dimer pair over 4 μs and in diverse environments, the dimer pair was not observed to assemble into a native conformation. The possible reason can be that a long time (longer that what is feasible in atomic scale simulations) is required to reorganize the pair of dimers to reach the minimum of energy (which might be corresponding to the native conformation). To obtain a complete assembly pathway, advanced simulation techniques will have to be devised in the future.

4.5 Interaction between RNA and dimers

The genome of viruses is a reservoir of negative charges, which can promote the assembly of viral particles in many cases. In our previous work on the self-assembly of CCMV particles in the presence of RNA genome, we found that the RNA genome captured free dimers in the solution and formed disordered nucleoprotein complexes, which subsequently relaxed into viral particles through a structural self-organization [42]. Despite the identification of the process, the way the dimers interact with the viral genome at the molecular level is still unclear. In this section, several simulations were carried out to attempt to clarify the mechanisms.

The viral genome is obviously too large to be studied by atomistic simulations, even considering the shortest segment, i.e., RNA4, which consists of 824 nucleotides [43]. Thus, to efficiently probe the interaction between RNA and dimers we can clip a small RNA segment

from RNA4 as RNA samples for simulations.

4.5.1 Interaction between a flexible RNA chain and a dimer

The viral genome has in fact a complex secondary structure and includes locally double-stranded segments and single-stranded loops. In order to identify the domains of the genome that the dimer prefers to bind on, a specific RNA chain possessing both double-stranded and single-stranded domains in its equilibrium conformation was extracted from the viral RNA4 segment and employed to interact with a dimer. The sequence of this RNA segment can be found in Figure 4-20. In this simulation, the RNA was permitted to move freely to sample its translational, orientational and conformational degrees of freedom without any position constraints. The RNA chain was first equilibrated for 200 ns leading to a conformation with two quasi-double-stranded segments formed at the two end of the chain connected by a single-stranded segment in the middle as shown in Figure 4-20. The interactions between the RNA segment and a dimer were probed for 500 ns.

GUGCCUAAUGAGCUAUCGUCCGAAAGGAACAAGCAGCUCAAGGUAG

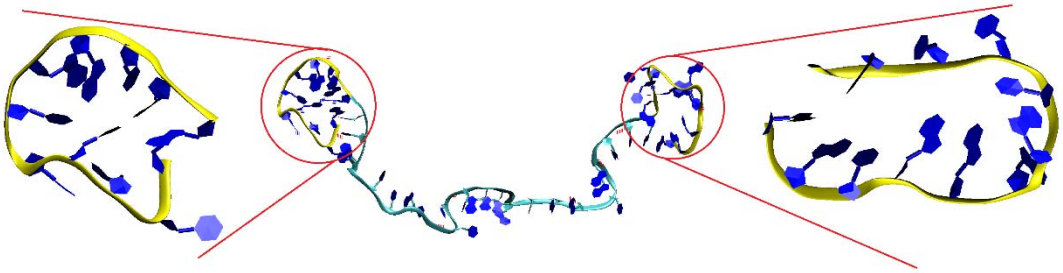

Figure 4-20. Snapshot of the conformation of a RNA segment after a 200-ns equilibration at 300K and 0.2 M. The yellow backbones stand for the quasi-double-stranded domains while the single-stranded domain is colored in cyan at the middle of the chain. The sequence of the RNA segment is above the snapshot.

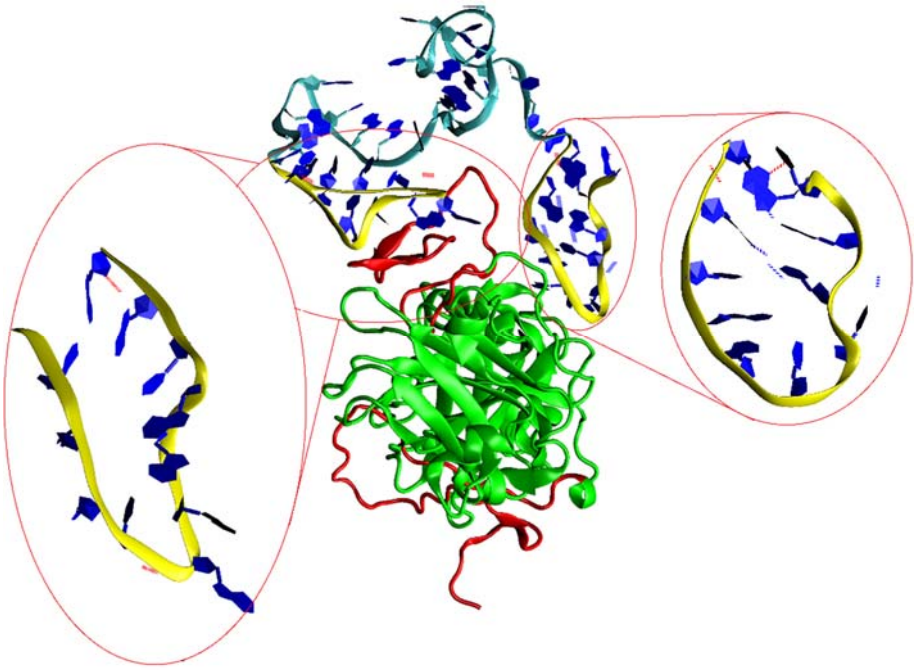

Figure 4-21. Snapshot of a dimer (in green for the body and in red for the tails) bound on a free RNA segment (46 nucleotides) with two quasi-double-stranded domains at the end of the chain (in yellow) and a single-stranded domain in the middle (in cyan). The initial conformation of the RNA segment is obtained through a 200-ns simulation.

The equilibrated RNA chain was then allowed to interact with a dimer and the resulting conformation of the complex is displayed in Figure 4-21. We found out that the dimer preferred to bind on the double-stranded domains via the N-terminal tails, which can be explained by the fact that double-stranded domains have a higher charge density.

4.5.2 Interaction between a fixed RNA chain and a dimer

Simulation protocol

A short RNA segment (50 nucleotides, see Figure 4-22 for its sequence) taken from the RNA4 in the form of either a single strand or a double strand was used to study the interactions with a single dimer. To further simplify the problem the heavy atoms of the RNA were fixed to enforce perfect helical rod conformations of these RNA strands during the simulations while the dimer was allowed to move freely. The dimer was located at a position of 9 nm away from the center of RNA rod to allow the dimer to adjust its orientation and conformation before being captured by the RNA. The added tails of the dimer were deliberately set in a stretched conformation. The initial conformation is shown in Figure 4-22.

AUGUCUACAGUCGGAACAGGGAAGUUAACUCGUGCACACGGAAGGGCUGC

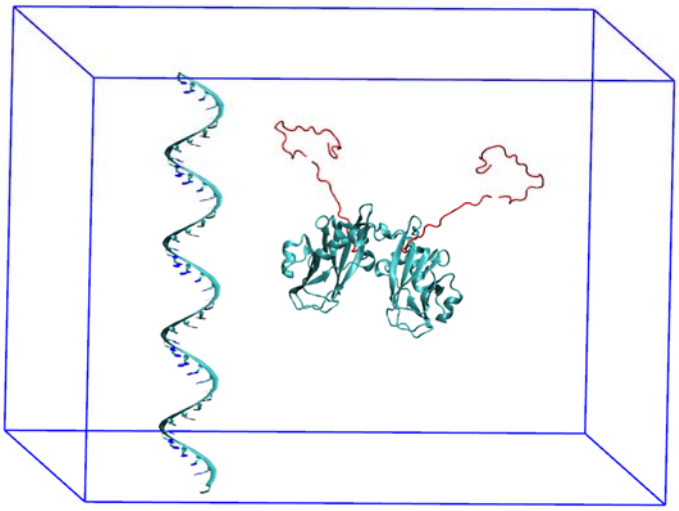

Figure 4-22. Initial conformation for the simulations probing the interaction between a dimer and a RNA rod, whose position was constrained. The sequence of the RNA rod is above the simulation box.

Three systems consisting of different components were studied at 0.1 M salt concentration. The first system consisted of a normal dimer and a single-stranded RNA rod and acted as a reference system (heavy atoms on the RNA rod were restrained during simulations). A mutated dimer without the N-terminal tails was implemented in the second system to investigate the role played by these tails in the interaction with RNA. At last, the influence of the RNA secondary structure form was evaluated by changing the RNA single-stranded rod into a double-stranded form. The simulations of the first two systems lasted for 1 μs, while the third one lasted for 800 ns. The final conformations were employed in steered molecular dynamics (SMD) simulations to measure the rupture force required to unbind the dimer from the RNA rod. The pulling force constant was 1000 $kJ.mol^{-1}.nm^{-1}$ and the pulling velocity was 10 $Å.ps^{-1}$. In addition, heavy atoms on the RNA rod were restrained during the pulling as well. The maximum force of pulling curves was used to estimate the rupture force to unbind the specific components from the subassemblies. Five pulling simulations with different initial velocities were performed for each system to obtain an average rupture force.

Results

As indicated by the snapshots of the simulated conformations in Figure 4-23, we found unexpectedly that the stretched N-terminal tails in the initial conformations preferred to bind to the body of dimers first, instead of the negatively charged RNA rods. It took a period of time for the dimer to reorient before binding to the RNA rod, namely, 600 ns for a dimer with a single-stranded RNA rod, over 750 ns for a cleaved dimer with a single-stranded RNA rod, and 500 ns for a dimer with a double-stranded RNA (see the separation and orientation evolutions in Figure 4-23). Obviously, the binding between the cleaved dimer and the single-stranded RNA rod was not as stable as that between the complete dimer and the single-stranded RNA rod, as revealed by the oscillation present after the first binding. It suggested a weak interaction

between the cleaved dimer and the RNA rod.

The simulated conformation of a dimer and a RNA rod can be seen in Figure 4-23. An interesting thing was that the dimer with tails preferred to contact the RNA rod with its inner surface which was more positively charged (see Figure 4-24), a phenomenon that was not observed either in the presence of a flexible RNA chain or by removing the N-terminal tails. This phenomenon was consistent with the physical nature of the dense negatively-charged phosphate groups of the RNA chain and the positively-charged residues on the inner surface of dimers. Such an orientation of the dimer increased the side-by-side contact probability between dimers, and given the fact that the side face of CCMV dimers is mostly dominated by hydrophobic domains, it suggested that this orientation facilitated the assembly of dimers.

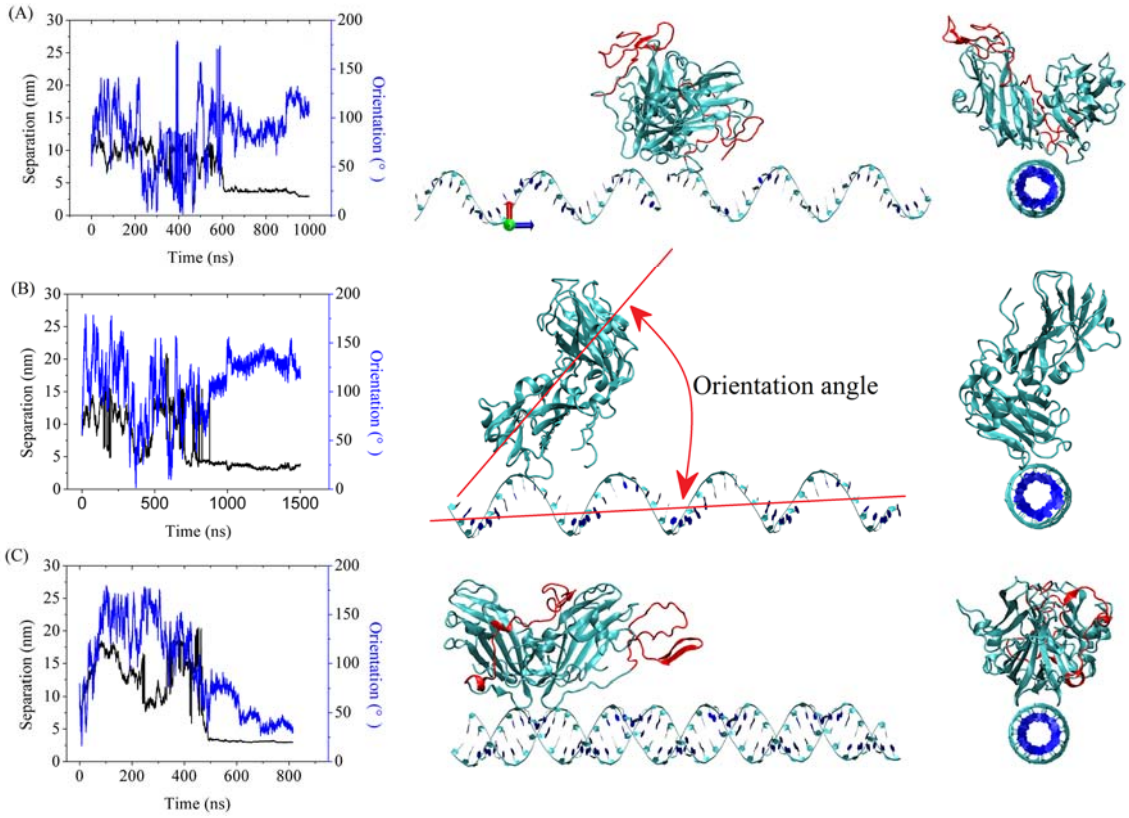

Figure 4-23. Evolution of the perpendicular separation (black lines) between the dimer and the RNA rod, and orientation angle of the dimer (blue lines) along the direction of the RNA rod at 0.1 M in different conditions: a single-stranded RNA rod with a dimer (A), a single-stranded RNA rod with a cleaved dimer (B), a double-stranded RNA rod with a dimer (C). The snapshots on the right side are the corresponding simulated conformations with the N-terminal tails in red.

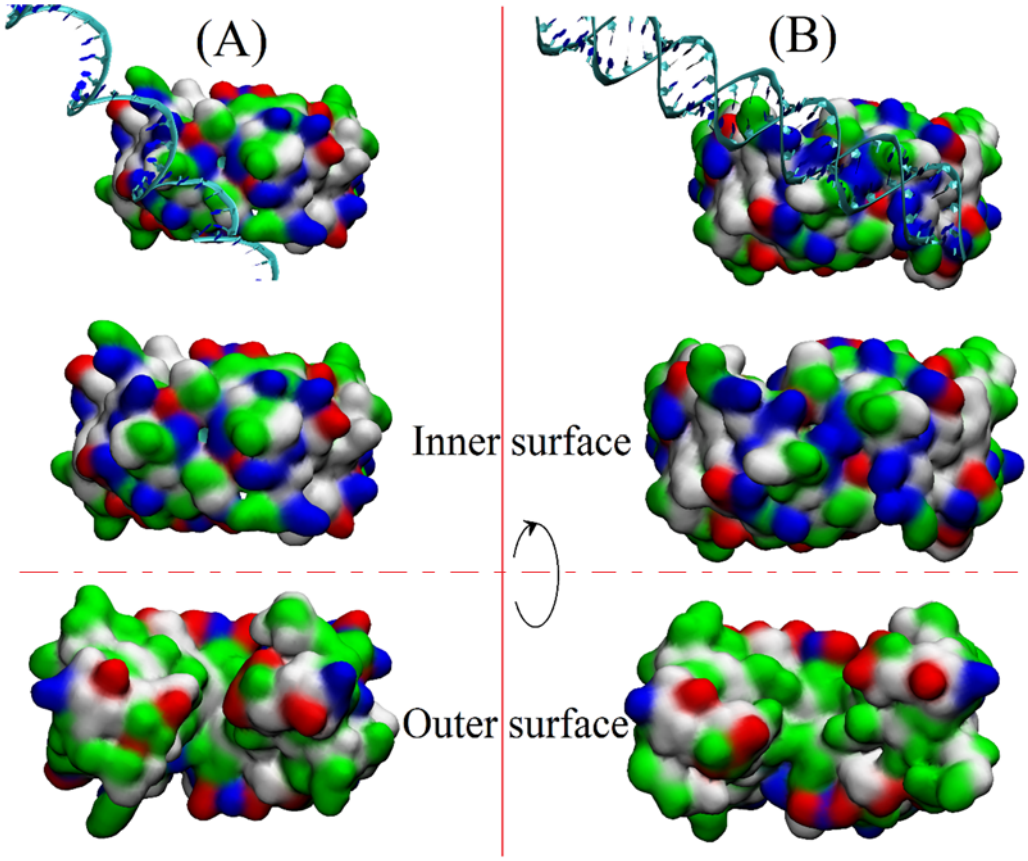

Figure 4-24. Surfaces of a complete dimer interacting with RNA rods in two forms: single-stranded (A) and double-stranded (B). The dimer is plotted using the QuickSurf method in VMD and the residues were colored according to their residue types: non-polar (white), polar (green), acidic (red) and basic (blue). The snapshots in the middle and bottom layer are the inner surface and outer surface of the dimer, respectively.

By SMD simulations, the binding strengths between the dimer and RNA were evaluated through the rupture force to separate the dimer from the RNA, and the results are shown in Figure 4-25. It was not surprising to find that a dimer had a stronger interaction (~1068 $kJ.mol^{-1}.nm^{-1}$) with a double-stranded RNA than with a single-stranded one (~750 $kJ.mol^{-1}.nm^{-1}$) owing to the stronger electrostatic attraction caused by a higher negative charge density. This finding agreed well with the simulation performed with a flexible RNA chain in Figure 4-21. However, the dimer-RNA interaction was greatly weakened to ~506 $kJ.mol^{-1}.nm^{-1}$ after removing the N-terminal tails which were highly positively charged and strongly involved in the interaction mechanism. The weaker rupture force for the cleaved dimer could also be confirmed by a longer association time and an unstable binding as revealed in Figure 4-23.

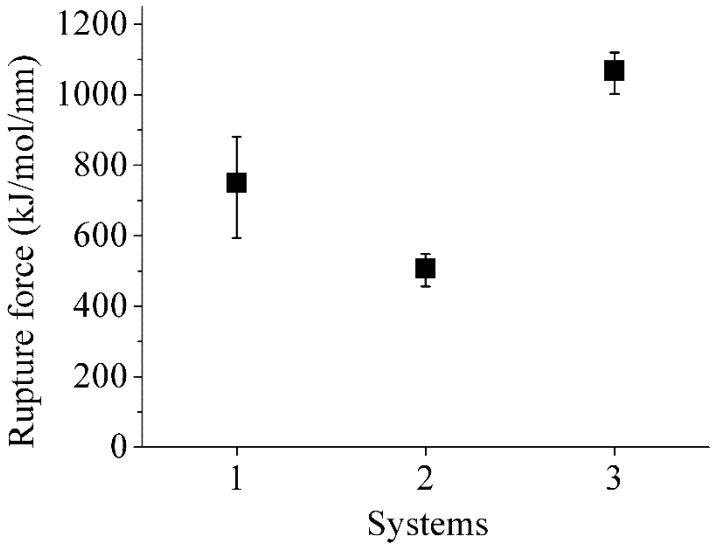

Figure 4-25. Rupture forces to unbind a dimer from a RNA rod at an ionic strength of 0.1 M. The numbers in x-axis represent the interacting systems: a dimer with a single-stranded RNA rod (1), a tail-cleaved dimer with a single-stranded RNA rod (2) and a dimer with a double-stranded RNA rod (3).

Discussion

The simulations above indicated that the dimer preferred to bind to the double-stranded domains of RNA chains with the inner surface of dimers. The simulation outcomes showed a high consistency with the structural characteristics of the components. Double-stranded RNA possesses a high negative charge density from phosphate groups, while the positively charged residues are concentrated at the inner surface of dimers. Thus it was natural to observe the phenomenon found in MD simulations. Overall, the genuine conformation of CCMV RNA is irregularly granular (see Figure 1-20). In principle, the double-stranded domains should be more water soluble due to their higher charge density and tend to be distributed at the surface of the RNA coil as a consequence. When dimers are getting close to the RNA coil, they should preferentially bind to the double-stranded domains forming a protein-RNA complex. It should be noted that since the dimers prefer to bind to the RNA with their inner surface, the probability for the hydrophobic domains located on the side surfaces of the dimers to come close to each other should subsequently increase, which will significantly improve the assembly of virons. Therefore, the presence of viral genome, to some extent, plays a role of an assembly template. The binding of dimers to RNA coil partially neutralizes the negative charge of RNA, resulting in the shrinkage of the size of the protein-RNA complex which subsequently self-organizes into a complete viral particle.

4.6 Interaction strengths within subassemblies

4.6.1 Simulation protocol

The initial conformations of subassemblies for SMD simulations were produced by various

ways. The conformation of a dimer-RNA complex was obtained from a simulation on the interaction of a dimer with a flexible RNA segment as described in the previous section. The conformation of the other subassemblies were obtained after 500 ns, 100 ns and 100 ns simulations for a pair, a pentamer and an hexamer of dimers, respectively. All these simulations were performed in NPT ensemble with a pressure of 1 bar and a temperature of 300 K.

The interaction strength within a pentamer and a hexamer was investigated both for intact proteins as well as for proteins missing N-terminal tails. In the latter case, all the monomers constituting the pentamer had their residues 2-39 cleaved while for the hexamer, the residues 2-25 of the "inner" monomers forming the barrel structure and the residues 2-39 of the outer monomers were cleaved (see Figure 4-26).

The pulling force constant was 1000 $kJ.mol^{-1}.nm^{-1}$ and the pulling velocity was 10 $Å.ps^{-1}$. In addition, the heavy atoms on the RNA segment were restrained during the pulling. The maximum force of the pulling curves was used to estimate the rupture force to unbind the specific components from the subassemblies. Five pulling simulations with different initial velocities were performed for each subassembly to obtain an average rupture force.

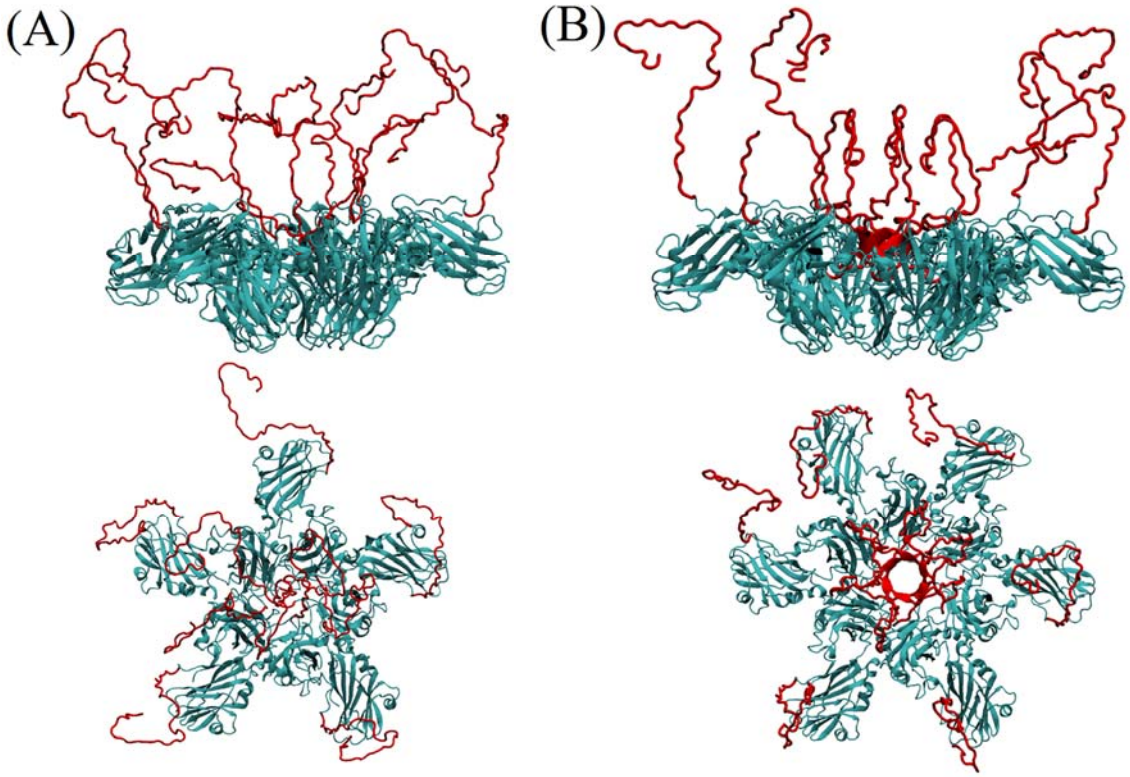

Figure 4-26. Initial conformation of a pentamer of dimers (A) and a hexamer of dimers (B) obtained by adding the missing residues. The N-terminal tails are colored in red.

4.6.2 Results

Quantifying the strength of interactions between a dimer and other viral subassemblies should be helpful to better understand the assembly process of viral particles. Due to the large size of the relevant subassemblies and the limitation in computational resources, it is difficult to determine the interaction energies directly by calculating their PMFs. Nevertheless, we can

obtain an estimate of the interactions by determining their rupture forces through SMD simulations.

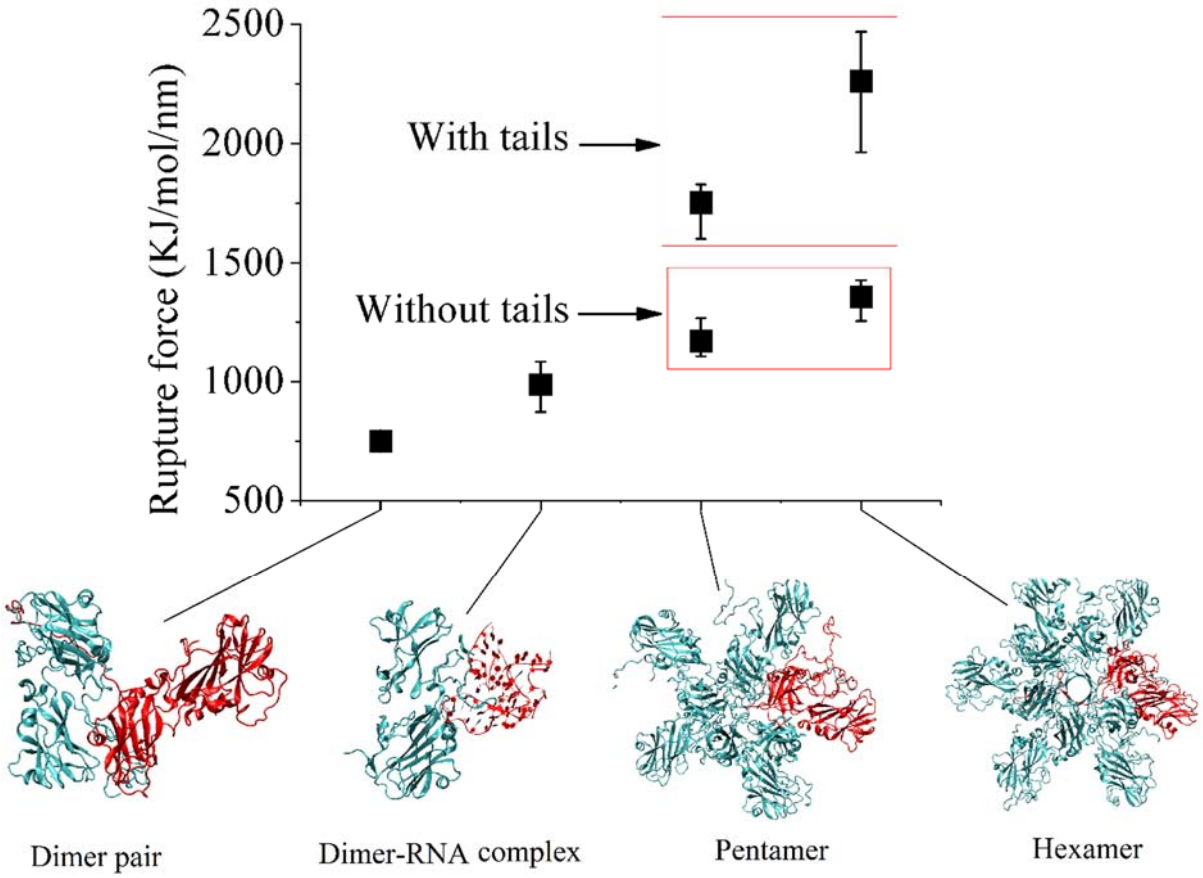

Figure 4-27. Rupture force in different subassemblies. The numbers on the x-axis correspond to the subassemblies: a pair of dimer in a native conformation within a capsid, a complex of a dimer and a flexible RNA piece, a pentamer of dimers and a hexamer of dimers. The pulled component is colored in red.

In this section, four typical subassemblies were analyzed: a pair of dimers, a dimer-RNA complex, a pentamer of dimers and a hexamer of dimers. Their rupture forces are given in Figure 4-27. The weakest rupture force (~750 $kJ.mol^{-1}.nm^{-1}$) was measured for separating two dimers, while the force to drag a dimer away from the flexible RNA segment comprising both double-stranded and single-stranded sections (see above) was a little higher and amounted to ~986 $kJ.mol^{-1}.nm^{-1}$. The dimer exhibited strong interactions within the pentamer (~1752 $kJ.mol^{-1}.nm^{-1}$) and the hexamer (~2262 $kJ.mol^{-1}.nm^{-1}$).The interaction strengths decreased sharply to a mediate level (~1170 $kJ.mol^{-1}.nm^{-1}$ for the pentamer and ~1354 $kJ.mol^{-1}.nm^{-1}$ for the hexamer) when the N-terminal tails were removed. The presence of the N-terminal tails linking the monomers in a pentamer or hexamer greatly increased the stability of these structures, emphasizing again the critical role played by the tails in stabilizing these structures. The high rupture forces showed that the pentamer and hexamer of dimers were stable subassemblies, which was consistent with their pivotal role in the icosahedral structure of a capsid. Interestingly, the rupture force between dimer and RNA was moderate. This fact can be rationalized by noticing that it is important for a virus to release easily its genetic material in the host cell and

therefore, a relatively weak binding energy between the dimers and the RNA genome may be necessary for its own survival.

4.7 Conclusions

By using atomistic MD simulations, we investigated the interactions between a dimer and various subassemblies. The structure of a dimer was looser in solution than within a capsid and the conformation of its N-terminal tails was sensitive to the ionic strength. At low salinity, the N-terminal tails presented a stretched conformation and slightly bound on the body of dimer, covering a part of the hydrophobic domains. Subsequently, the binding of another dimer required a conformational change of the tails so as to expose the hydrophobic domains. This change might be triggered by the presence of RNA on which the tails would preferentially bind allowing thus the association of dimers. N-terminal tails were involved in the self-assembly of dimers at low ionic strength: on the one hand, they accelerated the association by bridging the dimers; but on the other hand, they could lead to kinetic traps and misassembly caused by a strong interaction with the body of the other dimer.

We observed that pentamer and hexamer of dimers were stable subassemblies with a dimer requiring a strong rupture force to be extracted from them. This result is in agreement with the fact that an icosahedral capsid is made up of pentamers and hexamers of dimers, and the integrity of the capsid is preserved through the solidity of these subassemblies. Quite interestingly, the interaction between RNA and a dimer was moderate, which was consistent with our experimental measurements. It may be necessary for the survival of the virus since it needs to readily release its genome in the host cell. It may be also useful for the selection of the genome: with a weak binding energy, dimers are enabled to quickly unbind RNA if it does not belong to the viral genome and have thereby more chances to package the right segments to ensure the virus proliferation. A dimer binds also preferentially to the double-stranded domains of RNA because the charge density is higher.

Despite the size and complexity of the CCMV virus, we have been able to shed some light on the interactions between dimers and subassemblies at the atomistic level. Structures, forces, energies, and timescales become increasingly important to understand the dynamical phenomena underlying the virus self-assembly. Only a few of these quantities are accessible through experiments, and atomistic computer simulations can bring realistic, complementary information. We believe that experiments, theory and atomistic simulations will nurture one another and illuminate the multi-scale processes occurring in biological systems.

4.8 References

[1] Y. Hiragi, H. Inoue, Y. Sano, K. Kajiwara, T. Ueki, and H. Nakatani, Journal of Molecular Biology **213**, 495 (1990).
[2] J. M. Johnson, J. H. Tang, Y. Nyame, D. Willits, M. J. Young, and A. Zlotnick, Nano Letters **5**, 765 (2005).

[3] E. E. Pierson, D. Z. Keifer, L. Selzer, L. S. Lee, N. C. Contino, J. C. Y. Wang, A. Zlotnick, and M. F. Jarrold, Journal of the American Chemical Society **136**, 3536 (2014).
[4] O. M. Elrad and M. F. Hagan, Physical Biology **7**, 045003 (2010).
[5] A. Kivenson and M. F. Hagan, Biophysical Journal **99**, 619 (2010).
[6] M. A. Krol, N. H. Olson, J. Tate, J. E. Johnson, T. S. Baker, and P. Ahlquist, Proceedings of the National Academy of Sciences of the United States of America **96**, 13650 (1999).
[7] J. M. Johnson, D. A. Willits, M. J. Young, and A. Zlotnick, Journal of Molecular Biology **335**, 455 (2004).
[8] M. Comas-Garcia, R. F. Garmann, S. W. Singaram, A. Ben-Shaul, C. M. Knobler, and W. M. Gebart, Journal of Physical Chemistry B **118**, 7510 (2014).
[9] R. F. Garmann, R. Sportsman, C. Beren, V. N. Manoharan, C. M. Knobler, and W. M. Gelbart, Journal of the American Chemical Society **137**, 7584 (2015).
[10] G. Tresset, M. Tatou, C. Le Coeur, M. Zeghal, V. Bailleux, A. Lecchi, K. Brach, M. Klekotko, and L. Porcar, Physical Review Letters **113**, 128305 (2014).
[11] C. Beren, L. L. Dreesens, K. N. Liu, C. M. Knobler, and W. M. Gelbart, Biophysical Journal **113**, 339 (2017).
[12] G. Erdemci-Tandogan, H. Orland, and R. Zandi, Physical Review Letters **119**, 188102 (2017).
[13] J. D. Perlmutter, C. Qiao, and M. F. Hagan, Elife **2**, e00632 (2013).
[14] V. Krishnamani, C. Globisch, C. Peter, and M. Deserno, European Physical Journal-Special Topics **225**, 1757 (2016).
[15] J. R. Perilla and K. Schulten, Nature Communications **8**, 15959 (2017).
[16] D. Sala, S. Ciambellotti, A. Giachetti, P. Turano, and A. Rosato, Journal of Chemical Information and Modeling **57**, 2112 (2017).
[17] E. Tarasova, V. Farafonov, R. Khayat, N. Okimoto, T. S. Komatsu, M. Taiji, and D. Nerukh, Journal of Physical Chemistry Letters **8**, 779 (2017).
[18] R. Kong *et al.*, Science **352**, 828 (2016).
[19] A. Meeprasert, S. Hannongbua, and T. Rungrotmongkol, Journal of Chemical Information and Modeling **54**, 1208 (2014).
[20] J. R. Perilla, J. A. Hadden, B. C. Goh, C. G. Mayne, and K. Schulten, Journal of Physical Chemistry Letters **7**, 1836 (2016).
[21] K. L. Prachanronarong *et al.*, Journal of Chemical Theory and Computation **12**, 6098 (2016).
[22] A. Vergara-Jaque, H. Poblete, E. H. Lee, K. Schulten, F. Gonzalez-Nilo, and C. Chipot, Journal of Chemical Information and Modeling **52**, 2650 (2012).
[23] Y. Andoh, N. Yoshii, A. Yamada, K. Fujimoto, H. Kojima, K. Mizutani, A. Nakagawa, A. Nomoto, and S. Okazaki, Journal of Chemical Physics **141**, 165101 (2014).
[24] M. R. Machado, H. C. Gonzalez, and S. Pantano, Journal of Chemical Theory and Computation **13**, 5106 (2017).
[25] Y. L. Miao, J. E. Johnson, and P. J. Ortoleva, Journal of Physical Chemistry B **114**, 11181 (2010).
[26] Z. Antal, J. Szoverfi, and S. N. Fejer, Journal of Chemical Information and Modeling **57**, 910 (2017).
[27] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E.

Shaw, Proteins-Structure Function and Bioinformatics **78**, 1950 (2010).
[28] K. C. J. M. p. Kiwiel, **90**, 1 (2001).
[29] B. Hess, H. Bekker, H. J. C. Berendsen, and J. Fraaije, Journal of Computational Chemistry **18**, 1463 (1997).
[30] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. J. S. Lindahl, **1**, 19 (2015).
[31] H. J. C. Berendsen, J. P. M. Postma, W. F. Vangunsteren, A. Dinola, and J. R. Haak, Journal of Chemical Physics **81**, 3684 (1984).
[32] M. Parrinello and A. Rahman, Journal of Applied Physics **52**, 7182 (1981).
[33] T. Darden, D. York, and L. Pedersen, Journal of Chemical Physics **98**, 10089 (1993).
[34] J. A. Speir, S. Munshi, G. J. Wang, T. S. Baker, and J. E. Johnson, Structure **3**, 63 (1995).
[35] P. Annamalai, S. Apte, S. Wilkens, and A. L. N. Rao, Journal of Virology **79**, 3277 (2005).
[36] T. Bereau, C. Globisch, M. Deserno, and C. Peter, Journal of Chemical Theory and Computation **8**, 3750 (2012).
[37] W. Humphrey, A. Dalke, and K. Schulten, Journal of Molecular Graphics & Modelling **14**, 33 (1996).
[38] S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg, Journal of Computational Chemistry **13**, 1011 (1992).
[39] P. Ceres and A. Zlotnick, Biochemistry **41**, 11525 (2002).
[40] J. Z. Chen, M. Chevreuil, S. Combet, Y. Lansac, and G. Tresset, Journal of Physics-Condensed Matter **29**, 474001 (2017).
[41] G. Tresset, J. Z. Chen, M. Chevreuil, N. Nhiri, E. Jacquet, and Y. Lansac, Physical Review Applied **7**, 014005 (2017).
[42] M. Chevreui *et al.*, Nature Communications **9**, 3071 (2018).
[43] R. Dasgupta and P. Kaesberg, Nucleic Acids Research **10**, 703 (1982).

Perspective

In this thesis, a method to estimate the interaction between assembly subunits has been firstly proposed. Despite the simplicity of the proposed lattice model, the hydrophobic strength between subunits and the effective charge of subunits can readily be obtained by this approach, while this remains difficult by other techniques. There is no doubt that this model possesses a strong transferability to other viral systems.

Atomistic MD simulations have allowed us to gain a better understanding of the interactions between different components of the CCMV at a microscopic scale. However, these interactions are far from being fully understood due to the inherent complexity involved. Even for a pair of dimers, the possible association conformations are so diverse that it is challenging to entirely go through in a single MD simulation and to extract the most stable conformation. To solve this problem more advanced simulation techniques are required, such as Replica-Exchange Molecular Dynamics (REMD) which should be able to effectively avoid the problem of energy landscape traps during the assembly reactions. If the most stable association conformation for a pair of dimers in solution is not the native conformation obtained from a full capsid, it indicates that the association conformation will generally evolve with subsequent sequential additions of new dimers to the dimer pair to ultimately lead to the native conformation.

In this thesis, only the initial step in the assembly of viral capsids, namely, the assembly of a pair of dimers, was investigated, which is far from enough to reveal the whole assembly mechanism. The next step is to attempt to search for the possibly stable nuclei in the assembly reaction and to elucidate their formation pathway. Given the experimental observations that the assembly is affected by the concentration of capsid protein, there are two ways to investigate the subsequent assembly reactions by MD simulations: the first way is to mimic the assembly reaction with a high capsid protein concentration by introducing several dispersed dimers into a simulation box to observe how they interact with the others; the second way is to mimic the reaction at ultra-dilute capsid protein concentration by performing a series of MD simulations with an increasing number of dimers. More specifically, the first MD simulation begins with a pair of dimers. After the association reaction of these two dimers reaches an equilibrium, an additional dimer is introduced to the simulation system and another simulation is performed to probe the next association reaction. This step is repeated until the number of dimers in the box reaches the total number in a full capsid. By performing simulations in these two ways, we will be able to compare the reactions at low and high capsid protein concentrations. Since those systems comprised a very large number of atoms, it will be advisable to replace the atomistic force field by a coarse-grained (possibly working with an implicit solvent) force field.

However, modeling the assembly of viral particles by MD simulations while maintaining the details of interactions between subunits is still a formidable challenge. The atomistic simulations performed in this thesis to estimate the effective potential between a pair of dimers under various conditions constitute the first step towards the development of an accurate coarse-grained force field. Such a force field is an imperative prerequisite to any investigation of the molecular mechanisms involved in the assembly of complete viral capsids.

université PARIS-SACLAY
ÉCOLE DOCTORALE
Physique et ingénierie : électrons, photons, sciences du vivant (EOBE)

Titre : Simulations dynamique moléculaire de l'auto-assemblage de virus icosaédrique

Mots clés : Simulations numériques, virus, auto-assemblage

Résumé : Les virus sont connus pour infecter toutes les classes d'organismes vivants sur Terre, qu'elles soient végétales ou animales. Les virions consistent en un génome d'acide nucléique protégé par une enveloppe protéique unique ou multicouche appelée capside et, dans certains cas, par une enveloppe de lipides. La capside virale est généralement composée de centaines ou de milliers de protéines formant des structures ordonnées. La moitié des virus connus présentent une symétrie icosaédrique, les autres étant hélicoïdaux, prolats ou de structure irrégulière complexe. Récemment, les particules virales ont attiré une attention croissante en raison de leur structure extrêmement régulière et de leur utilisation potentielle pour la fabrication de nanostructures ayant diverses fonctions. Par conséquent, la compréhension des mécanismes d'assemblage sous-jacents à la production de particules virales est non seulement utile au développement d'inhibiteurs à des fins thérapeutiques, mais elle devrait également ouvrir de nouvelles voies pour l'auto-assemblage de matériaux supramoléculaires complexes.
À ce jour, de nombreuses études expérimentales et théoriques sur l'assemblage de virus ont été effectuées. Des recherches expérimentales ont permis d'obtenir de nombreuses informations sur l'assemblage du virus, y compris les conditions appropriées requises pour l'assemblage et les voies cinétiques. En combinant ces informations et méthodes théoriques, une première compréhension du mécanisme d'assemblage des virus a été élaborée. Cependant, les informations provenant uniquement d'expériences ne peuvent donner une image complète, en particulier à l'échelle microscopique. Par conséquent, dans cette thèse, nous avons utilisé des simulations informatiques, y compris des techniques de Monte Carlo et de la dynamique moléculaire, pour sonder l'assemblage du virus, dans l'espoir de mieux comprendre les mécanismes moléculaires en jeu.

Title : Molecular dynamics simulation of the self-assembly of icosahedral virus

Keywords : numerical simulations, virus, self-assembly

Abstract : Viruses are known for infecting all classes of living organisms on Earth, whether vegetal or animal. Virions consist of a nucleic acid genome protected by a single or multilayered protein shell called capsid, and in some cases by an envelope of lipids. The viral capsid is generally made of hundreds or thousands of proteins forming ordered structures. Half of all known viruses exhibit an icosahedral symmetry, the rest being helical, prolate or having a complex irregular structure. Recently, viral particles have attracted an increasing attention due to their extremely regular structure and their potential use for fabricating nanostructures with various functions. Therefore, understanding the assembly mechanisms underlying the production of viral particles is not only helpful to the development of inhibitors for therapeutic purpose, but it should also open new routes for the self-assembly of complex supramolecular materials.
To date, numerous experimental and theoretical investigations on virus assembly have been performed. Through experimental investigations, a lot of information have been obtained on virus assembly, including the proper conditions required for the assembly and the kinetic pathways. Combining those information and theoretical methods, an initial understanding of the assembly mechanism of viruses has been worked out. However, information coming purely from experiments cannot give the whole picture, in particular at a microscopic scale. Therefore, in this thesis, we employed computer simulations, including Monte Carlo and molecular dynamics techniques, to probe the assembly of virus, with the expectation to gain new insights into the molecular mechanisms at play.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France