Molecular Dynamics Simulations of Microtubule-associated protein 1A/1B-light chain 3 (LC3) and its membrane associated form (LC3-II)

Shyno Mathew

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2017
ABSTRACT

Molecular Dynamics Simulations of Microtubule-associated protein 1A/1B-light chain 3 (LC3) and its membrane associated form(LC3-II)

Shyno Mathew

Autophagy is the process by which cells eliminate its unwanted or dysfunctional components. A major step in autophagy is the formation of autophagosome, the double membrane that engulfs the unwanted cellular components. Dysregulation of autophagy affects neurodegenerative disorders, infectious diseases, cancer, and aging. In yeast, Atg8 protein is considered to play a crucial role in autophagosome maturation. Studies have shown that yeast lacking Atg8 protein form extremely small autophagosomes. Similarly, mammalian cells lacking Atg8 homologues produced open autophagosomes. Microtubule-associated protein (MAP) light chain3 (LC3), a human homologue of Atg8 protein is considered to play a major role in autophagosome maturation. However the exact mechanism by which Atg8/LC3 affects the autophagosome maturation is not completely known. A possible mechanism evolving from various studies is the following: Upon binding to the autophagosome, Atg8 family undergoes a conformational transition, which allows it to associate with another membrane-bound Atg8 in a trans-fashion. The proposed goals of this research include testing this hypothesis, identifying the stable conformations of LC3 and LC3-II (membrane bound LC3) and getting insights into the molecular mechanism by which LC3 influence autophagosome maturation. To accomplish this, we are performing Hamiltonian replica exchange molecular dynamics (HREMD) simulations on LC3 and on LC3-II. The most stable conformations of LC3, and LC3-II are identified via clustering analysis. As autophagy modulation is considered as a potential therapeutic target for various diseases, understanding the molecular mechanisms of different stages of autophagy is very important.
Table of Contents

List of Figures iv
List of Tables x

1 Introduction 1
  1.1 Autophagy ............................................. 2
    1.1.1 Different Stages of Autophagy ...................... 4
    1.1.2 Atg8 proteins ...................................... 4
    1.1.3 Human Homologues of Atg8 proteins ................. 5
    1.1.4 LC3 and LC3-II ................................... 6
    1.1.5 Current Literature ................................ 7
  1.2 Hypothesis ............................................ 11
    1.2.1 Our Approach .................................... 11

2 Methodology 13
  2.1 Molecular Dynamics Simulations ........................... 13
    2.1.1 Initialization .................................... 15
    2.1.2 Force Calculation .................................. 16
    2.1.3 Integrating the Equations of Motion ................ 24
    2.1.4 Experimental Conditions ............................ 29
    2.1.5 Limitations of MD simulations ...................... 33
  2.2 Improved Sampling Techniques ............................ 33
    2.2.1 REMD ............................................. 34
2.2.2 HREMD ..................................................... 35
2.3 Analysis of Trajectories ..................................... 39
  2.3.1 Cluster Analysis .......................................... 39
  2.3.2 Root Mean Square Deviation (RMSD) ................. 40

3 Comparison Study: LC3 vs. GATE-16 (in solution) 41
  3.1 Clustering Analysis ......................................... 41
    3.1.1 Criterion 1 ........................................... 46
    3.1.2 Criterion 2 ........................................... 48
    3.1.3 Criterion 3 ........................................... 50
    3.1.4 Criterion 4 ........................................... 53
  3.2 Results and Discussion .................................... 56
  3.3 Conclusions ................................................ 61

4 LC3 in solution vs. membrane bound LC3 62
  4.1 Clustering Method I ....................................... 62
    4.1.1 Results and Discussion ................................ 66
  4.2 Clustering Method II ...................................... 69
    4.2.1 Criterion 1 ........................................... 70
    4.2.2 Criterion 2 ........................................... 71
    4.2.3 Criterion 3 ........................................... 72
    4.2.4 Criterion 4 ........................................... 74
    4.2.5 Results and Discussion ................................ 76
    4.2.6 Conclusion and Future Work ......................... 84

Bibliography ..................................................... 84

A Appendix ....................................................... 93
  A.1 Abbreviations of amino acids .......................... 93
  A.2 Other abbreviations ...................................... 94

B Simulation details ........................................... 95
B.1 Regular molecular dynamics- simulation details ................................ 95
B.2 Hamiltonian replica exchange molecular dynamics- simulation details .... 95
List of Figures

1.1 Simple overview of “autophagy” adapted from reference [18]. .......................... 3
1.2 Simple overview of “autophagy” adapted from reference [18]. The focus of this study is encircled. .......................................................... 3
1.3 Schematic representation of autophagy including the key factors in different stages adapted from reference [17]. ................................. 4
1.4 Classification of human homologues of Atg8 proteins. ................................. 5
1.5 Schematic representation of the formation of LC3-II, adapted from reference [56] 6
1.6 LC3-I colored according to N-terminal (green) and C-terminal (blue) subdomains .......................................................... 7
1.7 Sketch of the proposed mechanism for Atg8-driven liposome tethering. The peptide has been colored in a red-white-blue scale according to residue number, starting with the N-terminal in red and ending with the C-terminal in blue .......................................................... 11

2.1 Schematic diagram of a basic MD code .................................................. 14
2.2 Representation of a two-dimensional periodic system. The central box is the system we are simulating, here molecules are numbered. Molecules can enter and leave each box through any of the four edges. For the 3-D system, molecules are free to cross any of the six cube faces, adapted from reference [6] 23
2.3 (a) Simplified representation of a free energy landscape of a protein (b) Schematic representation of free energy landscape for a protein in regular MD .................................................. 33
2.4 Simplified representation of replica exchange molecular dynamics (REMD). 34
2.5 Schematic representation of a generic cluster analysis, adapted from reference [19]. 40
3.1 Hydrophobic pocket of GATE-16 formed by Met1, Val36, Phe79, and aliphatic portion of Lys 82. 42
3.2 Aligned residues of LC3 compared to Met1 of GATE-16. 43
3.3 Aligned residues of LC3 compared to Val36 of GATE-16. 44
3.4 Aligned residues of LC3 compared to Phe79 of GATE-16. 44
3.5 Aligned residues of LC3 compared to Lys82 of GATE-16. 44
3.6 (a) GATE-16 structure highlighting hydrophobic pocket (HP) formed by Met1, Val36, Phe79, and aliphatic portion of Lys 82. HP is shown in red surface representation. (b) LC3 structure highlighting hydrophobic pocket (HP) formed by Ser3, Tyr38, Leu82, and His86. HP is shown in red surface representation. 45
3.7 Aligned residues of LC3 compared to Phe115 of GATE-16. 46
3.8 (a) Histogram for the number of contacts of the side chain of Phe115 with hydrophobic pocket (HP) formed by Met1, Val36, Phe79, and aliphatic portion of Lys 82 in GATE-16. The cutoff value 8, is indicated by the dashed line, adapted from reference [45] (b) Histogram for the number of contacts of the side chain of Phe119 with hydrophobic pocket (HP) formed by Ser3, Tyr38, Leu82, and His86 in LC3. The cutoff value 9, is indicated by the dashed line. 47
3.9 (a) GATE-16 structure highlighting criterion 1: coverage of hydrophobic pocket (HP) formed by Met1, Val36, Phe79, and aliphatic portion of Lys 82 by Phe115. HP is shown in red surface representation and Phe115 in yellow CPK representation. (b) LC3 structure highlighting criterion 1: coverage of hydrophobic pocket (HP) formed by Ser3, Tyr38, Leu82, and His86 by Phe119. HP is shown in red surface representation and Phe119 in yellow CPK representation. 48
3.10 Aligned residues of LC3 compared to Trp3 of GATE-16. 49
3.11 (a) GATE-16: distance distribution between C_ζ atom of Phe115 and any of Trp3 atoms. The cutoff value 7.2 Å, is indicated by the dashed line, adapted from reference [45] (b) LC3: distance distribution between C_ζ atom of Phe119 and any of Lys5 atoms. The cutoff value 15.29 Å, is indicated by the dashed line.

3.12 (a) GATE-16 structure highlighting criterion 2: orientation of Phe115 toward the N-terminus. HP is shown in red surface representation, residues Phe115 (yellow) and Trp3 (blue) in CPK representations. (b) LC3 structure highlighting criterion 2: orientation of Phe119 toward the N-terminus. HP is shown in red surface representation, residues Phe119 (yellow) and Lys5 (blue) in CPK representations.

3.13 Aligned residues of LC3 compared to Phe117 of GATE-16.

3.14 (a) Histogram for the number of contacts of the side chain of Phe117 with hydrophobic pocket (HP) formed by Met1, Val36, Phe79, and aliphatic portion of Lys 82 in GATE-16. The cutoff value 8, is indicated by the dashed line, adapted from reference [45] (b) Histogram for the number of contacts of the side chain of Gly120 with hydrophobic pocket (HP) formed by Ser3, Tyr38, Leu82, and His86 in LC3. The cutoff value 6, is indicated by the dashed line.

3.15 (a) GATE-16 structure highlighting criterion 3: coverage of hydrophobic pocket (HP) formed by Met1, Val36, Phe79, and aliphatic portion of Lys 82 by Phe117. HP is shown in red surface representation and Phe117 in yellow CPK representation. (b) LC3 structure highlighting criterion 3: coverage of hydrophobic pocket (HP) formed by Ser3, Tyr38, Leu82, and His86 by Gly120. HP is shown in red surface representation and Gly120 in yellow CPK representation.
3.16 (a) Distance distribution between N atom of Met1 and the closest of the two carboxylate oxygens of Phe117 in GATE-16. The cutoff distance 4 Å, is indicated by the dashed line, adapted from reference [45] (b) Distance distribution between N atom of Ser3 and the closest of the two carboxylate oxygens of Gly120 in GATE-16 The cutoff distance 18.49 Å, is indicated by the dashed line.

3.17 (a) GATE-16 structure highlighting criterion 4: salt bridge between N and C termini formed by Met1 and Phe117. HP is shown in red surface representation, Met1 (blue) and Phe117 (yellow) CPK representations. (b) LC3 structure highlighting criterion 4: salt bridge between N and C termini formed by Ser3 and Gly120. HP is shown in red surface representation, Ser3 (blue) and Gly120 (yellow) CPK representations.

3.18 (a) GATE-16 preferred state ‘yes no yes no’, along with criteria: HP is shown in red surface representation, Met1 (blue, part of HP), Trp3 (blue), Phe115(yellow) and Phe117 (yellow) CPK representations. (b) LC3 preferred state ‘yes yes yes yes’, along with criteria: HP is shown in red surface representation, Ser3 (blue, part of HP), Lys5 (blue), Phe119 (yellow) and Gly120 (yellow) CPK representations.

3.19 (a) GATE-16 and (b) LC3 structures highlighting N-terminal subdomain in green and C-terminal subdomain in blue. N-terminal subdomain consists of two α helices.

4.1 (a) LC3 and (b) LC3_mem structures highlighting criterion1: coverage of hydrophobic pocket (HP) by residue Phe119. HP shown in red surface representation, is formed by residues Ser3, Tyr38, Leu82, and His86. Phe119 is shown in yellow CPK representation.

4.2 (a) LC3 and (b) LC3_mem structures highlighting criterion2: orientation of Phe119 toward the N-terminus. HP is shown in red surface representation, residues Phe119 (yellow) and Lys5 (blue) in CPK representations.
4.3 (a) LC3 and (b) LC3_mem structures highlighting criterion3: coverage of hydrophobic pocket (HP) by residue Gly120. HP shown in red surface representation is formed by residues Ser3, Tyr38, Leu82, and His86. Gly120 is shown in yellow CPK representation.

4.4 (a) LC3 and (b) LC3_mem structures highlighting criterion4: salt bridge between N and C termini formed by residues Ser3 and Gly120. HP is shown in red surface, Ser3 in blue CPK and Gly120 in yellow CPK representations.

4.5 (a) LC3 preferred state ‘yes, yes, yes, yes’ (b) LC3_mem preferred state ‘no, n/a, no, yes’, along with the four criteria: HP formed by Ser3, Tyr38, Leu82, and His86 is shown in red surface representation. Residues covering HP, Phe119, and Gly120 are shown in yellow CPK representations. Lys5 (for criterion2) and Ser3 (for criterion 4) are shown in blue CPK representations.

4.6 a) LC3 and (b) LC3_mem structures highlighting criterion1: coverage of hydrophobic pocket (HP) formed by Phe7, Phe108, and Tyr110 by Ile34. HP is shown in red surface and Ile34 in yellow CPK representations.

4.7 Histograms for the number of contacts for Ile34 side chain atoms with hydrophobic pocket formed by Phe7, Phe108, and Tyr110, for (a) LC3 and in (b) LC3_mem. The cutoff value 11, is indicated by the dashed line.

4.8 a) LC3 and (b) LC3_mem structures highlighting criterion2: salt bridge between N and C termini, formed among residues Asp19 and Lys51. HP is shown in red surface, Asp19 in red CPK, and Lys51 in blue CPK representations.

4.9 Histograms for the Asp19-Lys51 salt bridge distance for (a) LC3 and in (b) LC3_mem. The cutoff value 3.6 Å, is indicated by the dashed line.

4.10 a) LC3 and (b) LC3_mem structures highlighting criterion3: salt bridge between N and C termini, formed between Glu105 and Arg16. HP is shown in red surface, Glu105 in red CPK and Arg16 in blue CPK representations.

4.11 Histogram for the Glu105-Arg16 salt bridge distance (a) LC3 and in (b) LC3_mem. The cutoff value 6.1 Å, is indicated by the dashed line.
4.12 a) LC3 and (b) LC3_mem structures highlighting criterion 4: salt bridge between N and C termini, formed between Asp106 and Arg11. HP is shown in red surface, Asp106 in red CPK and Arg11 in blue CPK representations.

4.13 Histogram for the Asp106-Arg11 salt bridge distance (a) LC3 and in (b) LC3_mem. The cutoff value 9.5 Å, is indicated by the dashed line.

4.14 Most populated states for (a) LC3 - ‘yes, no, yes, yes’ and in (b) LC3_mem - ‘yes, no, no, no’. The N-terminal subdomain is colored in green and C-terminal subdomain is colored in blue.

4.15 The second most populated states for (a) LC3 - ‘yes, no, no, no’ and in (b) LC3_mem - ‘yes, yes, no, no’. The N-terminal subdomain is colored in green and C-terminal subdomain is colored in blue.

4.16 The third most populated states for (a) LC3 - ‘yes, yes, yes, yes’ and in (b) LC3_mem - ‘yes, no, yes, no’. The N-terminal subdomain is colored in green and C-terminal subdomain is colored in blue.

4.17 The fourth most populated states for (a) LC3 - ‘yes, yes, yes, yes’ and in (b) LC3_mem - ‘yes, yes, yes, no’. The N-terminal subdomain is colored in green and C-terminal subdomain is colored in blue.

4.18 The salt bridges between N and C-terminal regions of (a) LC3 - 37 salt bridges and (b) LC3_mem - 13 salt bridges. The acidic residues (Asp, Glu) are shown in red CPK representation and basic residues (Arg, His, Lys) are shown in blue CPK representation.
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Criteria for GATE-16 clustering analysis</td>
<td>42</td>
</tr>
<tr>
<td>3.2</td>
<td>Criteria for Clustering Analysis</td>
<td>45</td>
</tr>
<tr>
<td>3.3</td>
<td>Criteria for Clustering Analysis</td>
<td>55</td>
</tr>
<tr>
<td>3.4</td>
<td>Comparison of states: GATE-16 vs. LC3</td>
<td>57</td>
</tr>
<tr>
<td>3.5</td>
<td>Preferred States</td>
<td>60</td>
</tr>
<tr>
<td>4.1</td>
<td>Criteria for Clustering Method I</td>
<td>63</td>
</tr>
<tr>
<td>4.2</td>
<td>Comparison of states: LC3 vs. LC3_mem</td>
<td>66</td>
</tr>
<tr>
<td>4.3</td>
<td>Criteria for Clustering Method II</td>
<td>76</td>
</tr>
<tr>
<td>4.4</td>
<td>Comparison of states - LC3 in solution (LC3) vs. membrane bound LC3 (LC3_mem)</td>
<td>77</td>
</tr>
</tbody>
</table>
Acknowledgments

Like the saying goes ‘Gratitude is the best attitude’. I am truly grateful for so many people who helped me to finish my PhD, including my advisor, Chemical Engineering department, Office of Graduate Student Affairs, teachers, mentors, friends and my family. I would like to thank each one of them for their support and generosity.

First and foremost I like to thank my adviser, Prof. Vanessa Ortiz for her support in guiding me throughout the project. With her intelligence, inspiration and ability to explain things simply and clearly, she helped me to have a good learning and research experience. Prof. Ortiz is truly a good scientist.

I sincerely acknowledge the efforts of my dissertation committee members: Prof. Sanat Kumar, Prof. Alan West, Prof. Scott A. Banta, Prof. Oleg Gang, and Prof. Allie Obermeyer. I thoroughly enjoyed the conversations with Prof. Edward F. Leonard, Prof. Ani Bedrossian-Omer, Prof. V. Faye McNeill, Prof. Alissa Park and Prof. Venkat Venktasubramanian. I would like to express my sincere gratitude to Chemical Engineering department for graciously supporting me with 7 teaching assistant positions. Through these TA assignments, I had the privilege to assist Prof. Alan C. West, Prof. Aghavni Bedrossian-Omer, Prof. Sanat K. Kumar, Prof. Ben O’Shaughnessy, Prof. Michael Hill, and Prof. Jeffrey T. Koberstein. These TA positions were excellent learning opportunities, learned a lot from the faculty, my fellow TAs and from my students. My co-TAs were Alysha Peters, Dr. Greesha Gadikota, Dr. Maxim Stonor, Dr. Thu Vi, and Nicholas Brady. The support gave by Kathy Marte, director of financial operations at our department was truly amazing. Once again, I like to thank all the faculty of Chemical Engineering department: Prof. Bishop, Prof. Chen, Prof. Durning, and Prof. Esposito. I like to thank all the current and previous staff at Chemical Engineering department including Teresa Colaizo, Mary Ko,
Ariel Sanchez, Krystal Paulino, Michele McCormack, and Rezarta Binaj.

For my defense talk practices so many PhD students from Chemical Engineering came to support me. Their feedback helped me make the best version of the defense talk. Thank you so much to all of you and so happy for the incredible work you all are doing- Thi Vo, Ahmet Hamdi Cavusoglu, Alison Fankhauser, Porakrit Leophairatana, Chathuranga De Silva, Sebastian Thomas Russel, Brian Michael Tackett, Natalie Labrador, Dr. Kevin Knehr. I would like to thank the current and past graduate students, research scientists, and postdocs from Chemical Engineering department - Dr. Behnaz, Dr. Bob, Dr. Michael, Dr. Manish, Dr. Jagan, Andrew, John, Alex, Cristianna, Kevin Lee, and Kenneth.

The conversations with Prof. Giovani Bussi really helped in setting up the advanced simulations for this study. Also communications with Oliver Schillinger, PhD student at Institute of Complex Systems and Dr. Francesco Musiani helped me to compare our results to the GATE-16 protein study, discussed in chapter 3 of this dissertation.

I would like to thank Office of Graduate Student Affairs specially the efforts of Vice Dean Soulymane Kachani and Assistant Dean of Graduate Student Affairs, Ellie Bastani for offering me the graduate assistant position when faced with funding difficulty. During the two years working as a graduate assistant, I learned a lot specially from Ellie Bastani, about organizing small events to large scale events. We created events like Engineering Your Style, and SEAS Women in Entrepreneurship hackathon. I consider Ellie Bastani as a great mentor. I would like to thank the efforts of Ellie and her husband Prof. Michael Collins in reviewing my dissertation. A special thanks to Associate Dean Tiffany Simon for always answering questions amidst of her busy schedule. Also, learned a lot while working at GSA and want to thank all the current and past GSA staff Jonathan Stark, Leora Brovman, Dr. Jennifer Piro, Jessie Jones, Fred Sanchez, Sam Ortiz and Srishti Sardana. I would like to thank Prof. Siu-Wai Chan for always motivating and sharing stories about perseverance. Thanks for always having a great smile Prof. Chan.

Apart from my parents and siblings, there are three people who really inspired me to a PhD: 1) Late Prof. Alan Bandy - the first person who told me about doing PhD and even offered a position to do PhD as his student in atmospheric chemistry. Truly grateful for
Dr. Bandy’s help, and for giving me the first job in US. I loved the time working in his lab in my freshmen year at Drexel Univeristy. I wish Prof. Bandy was around this time, but I hope he knew how much I respected him. 2) Dr. Arthur Kravchenko - Arthur was my coworker at Rohm and Haas while I was doing my internship and graciously gave many car rides back to the city. As Arthur used to be an international student, we always had great conversations, learned a lot from those car rides. Arthur always encouraged me to do a PhD. 3) Prof. Cameron Abrams - Prof. Abrams was my MS thesis adviser during the BSMS program at Drexel Univeristy. Working with Prof. Abrams was a great experience. He is one of the best teachers I ever had and helped me a lot in applying to PhD programs.

I like to express my sincere gratitude to Mindy Grossman and Neil Grossman for providing me the Grossman scholarship for the past 2 years, this really helped me. You both are a true motivation and I always cherish the annual lunches we had.

Columbia University is truly my second home. I met many inspiring people here through various events. I had the opportunity to be part of TEDxColumbiaEngineering twice; in 2013, I was the co-curator with Prof. Sheena Iyengar. Though I met Prof. Sheena only few times, she was always a true inspiration. I had the privilege to work with an amazing interdisciplinary team for TEDx event and we brought many great speakers including Prof. Shree Nayar, Prof. Michael Massimino, Dr. Jovana, Vince, Sean Kelly and many others.

Through one of the GSAS alumni events, I met George Khouri here at Columbia and ever since I am part of their family. George Khouri, and his wonderful wife Margee Khouri are both Columbia alumni, and was always supportive of my PhD work and startup project. I would like to thank their daughter Hope Khouri and her boy friend Miguel. Margee Khouri introduced me to Philanthropic Educational Organization (PEO), an international women’s organization that supports education. Truly thankful to all my PEO sisters specially Dr. Marti, Holly, Melanie, Bri, Erin and for the scholarship they gave me.

I like to thank the efforts of International Students and Scholars Office here at Columbia specially Jennifer Soler. Jennifer always gave the best advice. Columbia University Center for Career Education was always super helpful specially Vyju Manian, and Dr. Rachel Bernard. Columbia Athletics department is very dear to me, as I am a tutor for student
athletes for the past 2.5 years. I learned a lot while tutoring science and math courses. I like to thank Yana Rivers for the opportunity and for always giving me many students.

Columbia University has opened many doors for me; I met my startup founding team here. The time at Columbia inspired me to cofound qLogiX Entertainment with the mission to increase interest about STEM (Science Technology Engineering Mathematics) among young children. I want to thank my co-founder and a great friend/sister Dr. Jennifer Appawu for all her words of wisdom. My good friend Irene Chang also played a crucial role in the initial stages of qLogiX Entertainment. Starting June, qLogiX Entertainment will be at Columbia Startup Lab, another great initiative from the University. I like to thank Ivy Schultz, Chris McGarry, Alice and Adam from the Entrepreneurship office. Emily Ford from the Columbia Engineering Outreach Programs, has tremendously supported my startup, so a big thanks to her. Throughout my PhD journey, GSAS office was extremely helpful. I really admire the patience of Esmerlada McCormick for answering all my questions regarding dissertation deposit. I like to thank Julissa Pena for being very kind and helpful. A big thanks to Nikisha Alcindor and Mark Blacher from School of Professional Studies, I am looking forward to teaching this summer here at Columbia as well as at the GirlsinSTEM program.

Columbia presented me many remarkable opportunities; I felt it was a true blessing to meet His Holiness Ignatius Aphrem karim II, the patriarch of Syriac orthodox church of Antioch when he was the Archbishop of the Eastern Archdiocese in US. I am extremely thankful for email communications and the Holy father’s interest in supporting education. As the daughter of a Syriac orthodox priest, this was even more special. I like to thank Rev. Fr. Dan, members of Columbia Catholic Ministry, as well as the members of Orthodox Christian fellowship. A special thanks to Rev. Fr. Daniel Kearney of Ascension church, I was very touched when they selected me as one of the participants for washing of the feet during this year’s Holy Thursday service. I am extremely grateful for the support of Fr. Kearney and Robin Klueber, for my PhD as well as my startup venture.

I came to US for my undergraduate program in 2005. There were many family members and my parents friends who supported me through this journey. During my initial undergrad days, the help offered by Mathew Idichandy, his wife Jessie aunty and their daugther
Anniemol was superb. My mom’s cousin Kunjumol aunty helped me and my roommates a lot. A special thanks to kunjumol aunty, her husband Jose uncle, and their children Jeff, and Julia. I like to thank my uncle Sunny, his wife Suma aunty, Tanya, Charlie and Sarah. My dad’s cousins Korah uncle and his wife always gave good advice. I like to thank our relatives - Vipin achachan, his dad, Suju chechy and their kids. A special thanks to my parent’s friends- Late Dr. Nebu uncle, his wife Susan aunty, Lakshmi chechy, and her husband.

There are two family friends who incredibly supported me during my PhD: 1. Sajitha Chechy, her husband Jimmy Jacob - Words are not enough to thank this couple, they helped me many times including my relocation to NYC. They were the ones who kept all my belongings at their home, when I went to my parents in India after my undergrad graduation, and then drove all the way from philly to bring those things. Their children little Sam and little Danny truly cheered me during my graduation ceremony. 2. Martin Uncle, and his wife: Truly appreciate their help, always grateful to Martin Uncle for encouraging me to do a startup.

Apart from my parents support, it is the education I received as a child that made my journey possible, from a small town in India to my undergrad at Drexel University and then to PhD at Columbia University. I was the recipient of Navodaya academic scholarship, which provided me free yet the best education in the country from 6th grade to 12th grade. Students and teachers were selected to Navodaya schools based on a national level entrance examination and interviews. I was fortunate to have some of the best teachers at Navodaya school- Sreeja, Raziya, Valsamma, Vikraman, Palani Velu, Kurup, Minikumari, Sreekala, Renuka, Raji, Lisamma, Sunitha, Mini, Benny, George, Late. Velu, Geetha, Mary, Unnikrishnan, Raveendran, Sreekala, Uziya, Leelamma and many others. Our physical education instructor Cyril sir and family moved to US, so I really appreciate Cyril sir’s inquiries about my dissertation progress.

In terms of friendship, I think I am one of the richest people in the world. One of my best friends since my boarding school days at Navodaya, Steffi Jose has been a great motivation through out my PhD journey. She helped me a lot in many ways and words are not enough
to thank her. She is my non-biological sister. A huge thanks to Steffi, her husband Ajeesh and their son Johnykuttan. Another friend I met at Columbia during my orientation is Sonal Shetkar. A big thanks to you Sonal for always being there. I know I even missed your wedding due to my PhD defense, but a big thanks to you and your husband Sidd. There are some friends you can’t stop talking about, thanks to my biggest friends and supporters Wei and his wife Xiaoyu, Dimitris and Mario, Jeness and her husband Davis and their two wonderful children, David Tal, Anderi Papancea, Dr. Greeshma and her husband, Patricia Di Lima, Marissa, Sonali, Stamatis, Evan, Leah, Dr. Luc, Dr. Juan, Dr. Miguel and Dr. Mithun Radhakrishna. Thanks to many other friends I met during grad school like Alex, Tatyana, Jason, Nanfang, and Dean. Always learned something from each of you.

Since my childhood I dreamt of playing tennis everyday, this was not the case where I grew up. I am truly blessed to live near Columbia University with the access to two different sets of tennis courts: a) Riverside clay tennis association- the only real clay tennis courts in NYC and b) 119th street tennis association. My tennis buddies turned to be some of my best pals like Andrei, Christine Sala, Elyse, and KK. One of my initial tennis partners was Dr. Joseph Moll. A special thanks to Dr. Joe and his partner Kabir. My time at Drexel University brought me many wonderful friends like Dr. An Du and her husband Zhenhua Cai, Meyliana Wu, May Hnin, Thiri Hn, Dr. Ranjita Bose, Pooja Kapoor, Erin, Dr. Devika and Nala. Thanks you for always keeping in touch. Also, want to thank some of my childhood friends - Angel Mary, Tanya, Jerrin, Jintu, Joice, Tony, Boney, Sharon, Namitha, and Vinu.

Jeremy Balkin and his wife Rebecaa are very dear to me. I met Jeremy while organizing TEDxColumbiaEngineering, and had the privilege of having Jeremy as one of my life mentors. Emel Gunduc is one of the wisest women I know, thanks to your friendship and support, I consider you as my aunt.

It is because of my dad’s brother, Kunjachan I was able to do well in Math for the Navodaya Entrance examination. Those amazing math tricks you taught me years ago is still handy. I use some of it for BooMEE, first product of qLogiX Entertainment. I would like to thank Kunjachan, his wife Mini aunty, and their children- Bichu, Molu. I like to say thanks to
my dad’s sister- Sheeba aunty, her husband Monichan uncle, and their kids- Chesiya and Cheshiq. My dad’s mother (Ammachi) passed away few months ago, I bet she would have been very proud to see my dissertation. When I wrote my first article in the Navodaya school magazine, Ammachi was the first one who read it. Thanks to your love for reading, I guess you have passed it to many generations. A special thanks to Late appachi, my dad’s father who showed most of us grand children some interesting self defense mechanisms. I only smoked once in my life, when I was in Kindergarten or also, apparently sharing a beedi ‘Indian cigaratte ‘with my grandpa. I like to thank my grandfather’s brother whom we fondly call ‘daddy’, his wife, and all their children. Also would like to thank all the brothers, sisters of my grandparent’s and their families.

Now, it’s the turn of mum’s side. First I like to thank my mom’s sister, Illickal amma, and appachan. Illickal amma has cooked me many delicious meals while I was studying for my entrance examinations. I like to thank their children: 1) Molamma chechy and family- always a pleasure talking to Molamma chechy, a super smart and honest person I know. A special thanks to Molamma chechy, her husband aniyachayan, their kids Jomy and Jia. 2) thanks to Kochumol chechy, her husband Jojichachan, and their two daughters. Next I like to thank Thiruvanchur amma, my mum’s sister and her husband Late. Thiruvanchur appachan, their children: Tomachan chettan and family, Shijuchachan and family. I like to sincerely thank my mom’s brothers- 1) Rev. Fr. Philipose Corepiscopa Kottapurathu whom I call fondly as ‘Ajuapappa’, his wife aujammayi, their children- Aju achan and family, Anu Chechy and family and 2) Babychayan, his wife shaila aunty, and their kids- Biju and Bipin. My mom’s mother passed away before I was a year old. But I know she would be very proud see this achievement. I used to call my mum’s dad ‘Ajuappachan’; thanks so much aujappachan for showing us a healthy life style, I still remember the fresh orange coconuts he used to give us. Each time I pick a fresh young coconut from WholeFoods, I remeber those childhood days. Always admire the ayurvedic discipline aujappachan kept in his life, he would have been proud to hear that I am switching to more of an Ayurvedic life style. I like to thank my mum’s cousin- Leelamma aunty, her husband Late. Thankachayan, and their children- Vipin, and Vidya. My mom’s relatives, kumarakathamma, her husband, their daughter Asha chechy and family encouraged me throughout my studies.
A very special thanks to Late Very Rev. Fr. Joseph Pulickaparambil, my dad’s mentor. Rev. Fr. Joseph was the first person who taught me how to write. Another special thanks to my dad’s best friend, Rev. Fr. John achan, his wife and children: Sonu and family, Shimon. Thanks again for celebrating my PhD defense with my parents and siblings, they were so happy. Special thanks to Late. Thuruthel achan, his wife and their son, Jacob and family for their support. Another heartfelt thanks to Suresh sir and family for their kindness. Our family friends- Kummanathappachan and amma was always curious about my work, thanks to your prayers and support.

A sincere thanks to Diego Lynch for his tremendous help specially during the final days of my dissertation: really appreciate Diego reviewing my dissertation amidst of his submissions to major publications, tolerating my final practice before the defense and helping me so much on the defense day, truly grateful for this. Always had the best time with you, Diego.

Finally I want to thank my parents, Rev. Fr. A.T Mathew and Mrs. Annamma Mathew for inspiring me every single day, and for believing in all my dreams however crazy they are. My dad was supposed to do a PhD in US but didn’t work out for several reasons. I know two strokes had tired you a lot, but your ideas are still powerful. I still remember how happy and excited you were hearing my admission offer to Columbia. So, pappa this is for you. Truly grateful for my younger siblings Chelcy Smoony Mathew and Elias Tom Mathew for all the love and encouragement, and for always tolerating my gazillion startup ideas. Although you guys are younger than me, technically you both finished school before me :)

There are many people including the doorman to many students I taught who made my graduate student life a great experience. Please forgive me if I forgot mention your names. Once again thank you everyone for making this journey possible, and for making the dream of an ordinary girl come true :)
I have become my own version of an optimist. If I can’t make it through one door, I’ll go through another door - or I’ll make a door. Something terrific will come no matter how dark the present

- Rabindranath Tagore
For Pappa, Amma, Sonu, Appu

Steffi chechy, Diego

and Late. Prof. Alan Bandy
Chapter 1

Introduction

Autophagy is a cell’s cleansing mechanism and it occurs in all living cells. In Greek auto means self and phagy means eat, so autophagy is “self eat”. Belgian biochemist Christian de Duve coined the term autophagy [37]. It was first discovered in yeast in the 1950s, as a form of cellular self-digestion to help cells survive starvation. Highly conserved from yeast to human, autophagy is the process by which cells eliminate their unwanted or dysfunctional components, recycling useful components.

Autophagy is a very important cellular mechanism as it enables cells to sustain their function under conditions of stress such as nutrient deprivation, and the presence of pathogens. Many studies report that the dysregulation of autophagy affects diabetes, neurodegenerative disorders, infectious diseases, cancer, inflammatory bowel disease (IBD) and aging [33,56,57,74,75]. With type 2 diabetes, autophagy is slowed down which results in the accumulation of dysfunctional mitochondria which produces reactive oxygen species. Higher concentrations of reactive oxygen species leads to insulin resistance [41].

Researchers like Prof. Yoshinori Ohsumi, 2016 receipient of the Nobel Prize in medicine, have tremendously increased understanding of autophagy, for which he won the award. He used bakers yeast to identify genes essential for autophagy, and elucidated the underlying mechanisms for autophagy in yeast [1].

Autophagy is classified into three major types: macroautophagy, microautophagy and
chaperone-mediated autophagy (CMA) [66]. CMA has only been observed in mammals whereas macro and microautophagy can occur in plants, fungi, and eukaryotes [33]. Among these three types, macroautophagy is the most common [74]. This dissertation focuses on macroautophagy as microtubule-associated protein 1A/1B-light chain 3 or simply “LC3”, the protein under study is associated with macroautophagy. Hereafter, macroautophagy is referred to as autophagy.

Autophagy starts with the formation of a cup-like structure or “phagophore” that captures unwanted materials from the cell. The closure of these cup-like structures is necessary for removal of unwanted materials. Among the various proteins related to autophagy, LC3 is considered to be the marker of autophagy. It is a well established fact that, LC3 and its membrane associated form known as “LC3-II” plays a crucial role in the elongation of phagophore [59,72]. However, our hypothesis is that the membrane associated form of LC3 plays a crucial role in the closure of phagophores. Using simulations, we are identifying the molecular mechanism by which membrane associated form of LC3 drives the closure of these cup-like structures.

1.1 Autophagy

Autophagy can be considered as cells cleansing mechanism, it is the process by which cells eliminate its dysfunctional components. As shown in Figure 1.1, autophagy starts with the formation of a cup-like structure called “phagophore”, that engulfs the dysfunctional components.
Since LC3 and its membrane associated form (LC3-II) are associated with the elongation of phagophore, our study focuses on the area encircled in Figure 1.2.

Figure 1.2: Simple overview of “autophagy” adapted from reference [18]. The focus of this study is encircled.
1.1.1 Different Stages of Autophagy

The process of autophagy can be divided into three main stages: initiation, elongation and maturation. Figure 1.3 shows a more detailed view of autophagy including the key factors in different stages. As shown in Figures 1.1, 1.2, and 1.3 autophagy is initiated by the formation of phagophore (isolation membrane), which can selectively as well as non-selectively engulf cytosolic components. Cytosol is the fluid where organelles of cells reside.

![Figure 1.3: Schematic representation of autophagy including the key factors in different stages adapted from reference [17].](image)

As autophagy progresses, the phagophore expands and grows, resulting in a double membrane structure that finally closes to sequester its content from the cytosol [76]. These closed double membrane structures are termed as autophagosomes.

1.1.2 Atg8 proteins

The autophagy related genes essential for the assembly of autophagosomes are highly conserved between yeasts and mammals [66]. In yeast, about 15 autophagy-related (Atg) proteins are identified to accomplish autophagosome formation and maturation. Among these 15 Atg proteins, Atg8 plays a crucial role in autophagosome maturation and the spe-
cific recruitment of cargo proteins destined for autophagic degradation. Also, Atg8 is the only known stable, membrane-bound protein associated with the autophagosome, providing a reliable marker for the study of autophagy [35].

1.1.3 Human Homologues of Atg8 proteins

There are eight human homologues of Atg8, and based on amino acid sequence similarity, these eight proteins are classified into two groups as shown in Figure 1.4.

\[ \text{Figure 1.4: Classification of human homologues of Atg8 proteins.} \]

The first group, microtubule associated protein 1 light chain 3 (MAP1LC3) subfamily or simply LC3 subfamily consists of LC3A, LC3B, LC3B2 and LC3C. The second group, gammaaminobutyric acid (GABA) receptor-associated protein (GABARAP) subfamily contains GABARAP itself with several GABARAP-like (GABARAPL) proteins: GABARAPL1 (GEC1, glandular epithelial cell protein 1), GABARAPL2 (GATE-16, Golgi-associated ATPase enhancer of 16kDa) and GABARAPL3, which corresponds to a putative pseudogene [59]. Pseudogenes are relics of former genes that no longer possess biological functions [24]. The protein under study, LC3 is the first identified human homologue of Atg8. LC3 is known to play a role in the elongation of autophagosomes, whereas GABARAP family plays a role in the closure of autophagosomes [72].
1.1.4 LC3 and LC3-II

Microtubule-associated protein (MAP) light chain 3 (LC3) co-purifies with MAP1A and MAP1B, thus known as Microtubule-associated protein (MAP) 1A/1B-light chain 3 (LC3) [38, 40]. MAP-LC3 is a mammalian homologue of Atg8 protein. Unmodified and modified forms of MAP-LC3 is referred to as LC3-I and LC3-II (membrane bound LC3). The formation of LC3-II from LC3-I is shown in Figure 1.5 [56].

Studies have shown that all the Atg8 homologues undergo the same post translational modifications. For instance, let us consider LC3 subfamily: LC3 translated from mRNA is called Pro-LC3 and it possess an unnecessary C-terminal tail [76]. Atg4 protease cleaves this tail sequence to generate form-I LC3, which has a conserved C-terminal Glycine. Atg7, a ubiquitin activating like enzyme (E1) binds to the LC3-I C-terminal and then recruits Atg3 via its N-terminal domain. During the Atg3-Atg7 interaction, LC3-I is transferred to the Atg3 enzyme and Atg7 dissociates from the LC3-I complex. LC3-II (membrane associated form of LC3) is formed when LC3-I is conjugated to the lipid phosphatidylethanolamine (PE) in a reaction catalyzed by Atg3. The lipidation of LC3-I is enhanced by the binding
of Atg5-Atg12-Atg16L1 to Atg3. Also, Atg5-Atg12 complex has the ability to bind to PE already inserted into the isolation membrane. This ensures the incorporation of newly formed LC3-II into the expanding isolation membrane. It is known that the conversion of LC3-I to LC3-II is reversible. Atg4 can cleave PE from LC3-II and LC3 is recycled [76]. Experiments suggest LC3 attached to PE (LC3-II) binds tightly to autophagosomes [38].

1.1.5 Current Literature

The structure of LC3-I can be divided into two subdomains: N-terminal subdomain (residues 1 to 29) and C-terminal subdomain (residues 30 to 120). LC3-I structure colored according to these subdomains is shown in Figure 1.6.

![Figure 1.6: LC3-I colored according to N-terminal (green) and C-terminal (blue) subdomains](image)

The N-terminal subdomain consists of a short (H1) and long (H2) α-helices. The C-terminal
contains a central \( \beta \)-sheet composed of four strands S1, S2, S3, S4 and \( \alpha \)-helices H3, H4. Another term used for C-terminal domain of LC3 is the ubiquitin like domain (UBD) [38]. Other Atg8 homologues like GABARAP and GATE-16 proteins exhibit similar subdomains [51]. The N-terminal and C-terminal subdomains of LC3 vary in their function. Experiments show that N-terminal subdomain interacts with tubulin and microtubules via electrostatic interactions. On the other hand, C-terminal subdomain makes contact with membrane surface of autophagosomes [38]. So the tight binding of MAP-LC3 to autophagosomes occurs via C-terminal subdomain. The different features of the subdomains might allow LC3 to act as an adaptor protein between microtubules and autophagosomes [36].

Nakatogawa et al. provided the first experimental evidence suggesting Atg8 protein directly drives the growth of autophagosomes [51]. Their results showed that in vitro Atg8-PE (membrane bound Atg8) molecules function to tether together membranes to which they are attached. They also reported Atg8-PE mediating the fusion of these membranes. Further they identified the membrane fusion mediated by Atg8 is a hemifusion [15]. Hemifusion occurs when the outer/contacting leaflets of two apposed lipid bilayers merge while the inner/distal leaflets remain intact [40]. Another major observation made by Nakatogawa et al. is that “naked” membranes do not associate with membranes carrying Atg-PE, indicating that tethering is achieved due to interactions between Atg8-PE molecules on different membranes. To test the validity of the above mentioned results in vivo, Nakatogawa et al performed systematic mutational analyses. As mentioned before, Atg family proteins consists of two domains: an N-terminal helical domain (NHD) and a C-terminal ubiquitin-like domain (ULD). Deleting the first alpha helix of NHD resulted in a 30% decrease in autophagic activity. Deleting both alpha helices in NHD reduced the autophagic activity by 40%. These results suggested that NHD acts co-operatively with ULD in tethering and hemifusion. The study also identified a specific region on ULD that is important for clustering and hemifusion of liposomes. However this region is covered by NHD in the crystal structures of mammalian homologs. Thus they propose Atg8 adopts an “active” conformation upon PE binding, in which NHD is dissociated from the ULD, exposing the specific region in ULD, followed by multimerization, tethering and hemifusion of membranes [51].
Another study investigating the role of LC3 and GATE-16 N-termini in autophagosome maturation is done by Elazar and co-workers [73]. Here they used a liposome-based cell-free system based on a chemical crosslinking of the C terminus of LC3 and GATE-16 to liposomes, thereby mimicking the conjugation of these proteins to PE in vivo. By this chemical conjugation, they were able to bypass the requirement for conjugation machinery. The results showed LC3 and GATE-16 in their lipidated form promote membrane tethering and fusion. They further showed the fusion activity for both LC3 and GATE-16 is mediated by their N-terminal helices. However the membrane tethering activity depends on both N-terminus and ubiquitin core. Another key finding is that first -helices of GATE-16, and LC3 are enough for their fusion activity [73]. One striking difference compared to the results from Nakatogawa et al. is regarding membrane fusion. As opposed to the hemifusion reported by Nakatogawa et al., this study identified full membrane fusion. The authors relate the difference to the differences between experimental systems. The N-terminal -helices of GATE-16 and LC3 significantly differ in their amino acid sequence: GATE-16 N-terminus contains more hydrophobic residues and LC3 N-terminus has more positively charged residues. This study shows that for each protein, the different N-terminal residues present are essential for its tethering and fusion activities.

A study by Klionsky and co-workers contradicts some of the results shown by Nakatogawa et al. [50]. This study reports hemifusion of membranes mediated by At8/LC3 is not possible in the presence of physiological PE concentrations. They claim Nakatogawa et al. used non-physiological PE concentrations for their study. However the exact concentration of the PE at the phagophore is not known [25]. The study also provides some evidence for the involvement of SNARE proteins during autophagy. Also SNARE (Soluble NSF Attachment REceoptor) proteins are known to facilitate most of the fusion events in eukaryotic cells [62].

A recent study identified the conformational polymorphism in GATE-16 [45]. They chose GATE-16 as a representative of At8 protein and indicated LC3 might also show similar conformations as it belongs to the same Atg8 family. We are comparing our results to this study later in Chapter 3. Weidberg et al. showed that LC3 and GABARAP subfamilies are essential for autophagy [72]. They reached this conclusion by showing LC3 and GATE-16
(member of GABARAP subfamily) can undergo lipidation in the absence of the reciprocal subfamily whereas the overall autophagy is inhibited. They further showed LC3s are involved in the elongation of phagophore membrane and GATE-16 is involved in the closure of autophagosomes in the following manner: Using Atg4A, a protease that specifically recognizes GABARAPs, they removed GATE-16, which resulted in the accumulation of open autophagic membranes. As the Atg4A does not affect LC3s, the accumulation of larger phagophores under these conditions further emphasizes the role of LC3s in the elongation process.

Experimental evidence suggest a major function of Atg8 to be the regulation of membrane expansion as yeast lacking Atg8 form very small autophagosomes [36]. Mammalian cells lacking Atg8 homologues produced omegasome-like structures or “open” autophagosomes [72]. Another study reported the multimerization of Atg8-PE in vitro. They also showed Atg8 can tether liposomes thus bringing membranes into close proximity [51]. However the exact mechanism by which Atg8 affects the autophagosome maturation is not known. One possible mechanism evolving from various studies is the following: Upon binding to the autophagosome, Atg8 family undergoes a conformational transition, which allows it to associate with another membrane-bound Atg8 in a trans-fashion. Additionally, our experimental collaborator Prof. Ai Yamamoto observed LC3 attached to membrane or LC3-II interacts with another LC3-II via N-terminal. A study by Fitujita et al. indicate LC3 paralogues are involved in the closure of autophagosomes [22]. Experiments suggest the role of Atg8 as a cargo-capturing platform via its interactions with Atg19. Another proposed mechanism for Atg8-mediated membrane dynamics is the following: same interfaces that causes trans-interaction might capture proteins essential for autophagosome maturation including closure [14, 32]. More experiments and simulations are required to test these hypotheses.
1.2 Hypothesis

Our hypothesis is that the membrane associated form of LC3 or LC3-II plays a role in the closure of autophagosomes. We think, upon binding to the autophagosome, LC3 undergoes a conformational transition, which allows it to associate with another membrane-bound LC3 in a trans-fashion. Figure 1.6 depicts this proposed mechanism.

![Figure 1.6](image)

**Figure 1.6:** Sketch of the proposed mechanism for Atg8-driven liposome tethering. The peptide has been colored in a red-white-blue scale according to residue number, starting with the N-terminal in red and ending with the C-terminal in blue.

1.2.1 Our Approach

Our goal is to test the above mentioned hypothesis and gain insights into the molecular mechanism by which membrane bound LC3 influences autophagosome maturation. To accomplish this, we are performing Hamiltonian replica exchange molecular dynamics (HREMD) simulations of LC3 and LC3-II (membrane bound LC3). The most stable conformations of LC3, LC3-II are identified via clustering analysis. We are also compar-
ing our results for HREMD simulations of LC3 to a recent study that identified different conformations for GATE-16 protein, another human homologue of Atg8.
Chapter 2

Methodology

Computer simulations offer a direct route from the microscopic details of a system to macroscopic properties of experimental interest. Molecular simulation is a powerful approach to determine thermodynamic properties based on molecular models and is widely used in studying biological systems. The two most common molecular simulation methods are monte Carlo (MC) and molecular dynamics (MD). Two main advantages of MD over MC are: in MD, molecular motions are natural and it is possible to calculate dynamic properties [7]. We conducted MD simulations of LC3 using GROningen MAchine for Chemical Simulations (GROMACS) package [26]. GROMACS is a parallel molecular dynamics code primarily developed for simulating biochemical molecules like proteins, lipids and nucleic acids. GROMACS is considered to be the fastest MD code in the world [10,26].

2.1 Molecular Dynamics Simulations

Molecular Dynamics (MD) is a powerful computer simulation technique for studying many-body systems, as it allows one to predict the time evolution of a system of interacting particles. MD complement conventional experiments as it allows one to learn something new, something that cannot be identified through other ways. In many aspects MD simulations are similar to real experiments. For instance, in a real experiment we first prepare
a sample and then connect it to a measuring instrument. Then we measure the property of interest at a particular time interval. These measurements will become more accurate if we average them longer, as we can reduce the statistical noise. Similarly in MD simulations, we start by preparing the sample, selecting a model system consisting of N particles and then solving Newton’s equations of motion for the system until it reaches equilibrium. The actual measurements are taken at equilibrium. Figure 2.1 shows the structure of a basic MD code.

![Diagram of MD code](image)

**Figure 2.1:** Schematic diagram of a basic MD code

MD simulations can be connected to the ensemble average approach of elementary statistical thermodynamics through the ergodic hypothesis, which states that ensemble average and time average are equal,

\[ \langle f \rangle = \sum p_i f_i = \frac{1}{T} \int_{t}^{t+T} f(x(t)) dt \] (2.1)

The basic idea of ergodic hypothesis is that if we simulate the system for a longer time period, all micro-states of the system are equally accessible according to their equilibrium
probability [21]. Also, MD simulations can sample from a variety of statistical ensembles: microcanonical ensemble (NVE) which has constant number of particles (N), volume (V) and energy (E); canonical ensemble (NVT) where T stands for constant temperature; isothermal-isobaric (NPT) where P denotes constant pressure; grand canonical ensemble (µVT) where µ denotes chemical potential [21].

### 2.1.1 Initialization

The first requirement for starting an MD simulation is to generate the system’s initial configuration [21]. Here the initial configuration means the x, y, z coordinates of all particles of the system. For proteins, the initial positions are usually obtained from the published 3D structures available at the protein data bank or pdb.org [11]. For systems where the initial coordinates are not available, one must follow these rules [6,21]:

- Atoms must be arranged such that they don’t overlap
- All atoms must fit within the boundaries of simulation cell.

The next step is to generate initial velocities for the atoms. Initial velocities should satisfy the following rules [6,21]:

- The velocities are scaled so as to match an initial temperature.
- The sum of momenta is initialized to zero for momentum conservation.

\[
P = \sum_{i=1}^{N} m_i v_i = 0 \quad (2.2)
\]

Here \( P \) is the momentum, \( m_i \) and \( v_i \) are the mass and velocity of atom \( i \).

At a given temperature, the initial velocities are drawn randomly from a Maxwell-Boltzmann or Gaussian distribution [6]:

\[
p(v_{ix}) = \left(\frac{m_i}{2\pi k_B T}\right)^{1/2} \exp\left[-\frac{m_i v_{ix}^2}{2k_B T}\right] \quad (2.3)
\]

This equation shows the probability of atom \( i \) with mass \( m_i \) to have a velocity \( v_i \) in the x-direction. Again, the initial velocities are shifted so as to make the total momentum zero.
The instantaneous temperature \( T \) can be estimated using the following equation,

\[
T(t) = \frac{1}{k_B} \left( \sum_{i=1}^{N} \frac{m_i v_i^2(t)}{N_f} \right)
\]

(2.4)

where \( N_f \) is the number of degrees of freedom. For a system of \( N \) particles

\[
N_f = 3N - 3
\]

(2.5)

The evolution of system will drive the temperature to other values, meaning the average temperature over simulation time will not coincide with the desired temperature \([6, 21]\). This is because the starting atomic positions are not necessarily consistent with the desired temperature. Therefore it is important to drive the system from its initial conditions to other conditions that are compatible with the desired temperature. This is done by scaling the velocities during an initial period of the simulation, known as "equilibration". During equilibration each atomic velocity is scaled by a factor \( \sqrt{\frac{T_{ext}}{T_{inst}}} \), where \( T_{inst} \) is the instantaneous temperature and \( T_{ext} \) is the desired equilibrium temperature. Although the scaling will slowly drive the system to desired conditions, this creates a lack of energy conservation. Basically the dynamics is artificial and equilibration only prepares the system in conditions from which the real simulation can start. As a result, we cannot include any information obtained during equilibration in the analysis of results. The length of equilibration period varies from system to system. Ideally one should run equilibration for long enough such that system has lost memory of its initial conditions, and is fully at equilibrium at the desired temperature. Once the system reaches equilibrium, the average temperature would be close the desired temperature \([21]\).

### 2.1.2 Force Calculation

Everything moves because of forces, and these forces originate from the interactions among atoms and molecules. One can model molecular interactions through two different ways: quantum mechanics calculations and empirical forms. This leads to two different classes of molecular dynamic simulations, ab-initio MD and classical MD. For the purpose of this thesis, we are concentrating on classical MD and so from now on MD simulations stand for
classical MD. The MD simulation method is based on Newton’s second law or the equation of motion,

\[ F_i = m_i a_i \]  \hspace{1cm} (2.6)

where \( F_i \) is the force exerted on \( i \)th particle, \( m_i \) is its mass and \( a_i \) is the acceleration. We can write the above equation in terms of velocity \( v_i \) and position \( r_i \) of the \( i \)th particle.

\[ F_i = m_i \frac{dv_i}{dt} = m_i \frac{d^2 v_i}{dt^2} \]  \hspace{1cm} (2.7)

We know that force can also be expressed as the negative gradient of potential energy,

\[ F(\vec{r}) = -\nabla U(\vec{r}) \]  \hspace{1cm} (2.8)

If we consider atoms as particles of interest and assume the system has a total of \( N_{\text{at}} \) atoms, the force acting on \( i \)th atom at a given time can be obtained from the interatomic potential \( U \), which is a function of the positions of all atoms.

\[ \vec{F}_i = -\nabla_i U(\vec{r}_1, \vec{r}_2, \vec{r}_3, ..., \vec{r}_{N_{\text{at}}}) \]  \hspace{1cm} (2.9)

Once the initial conditions and the interaction potential are specified, we can solve the equations of motion numerically. The results will be the positions and velocities of all the atoms as a function of time. Therefore MD simulation is a deterministic approach, meaning any future or past state of the system can be generally predicted from the current state [21].

The potential energy can be expressed as the sum of potential energies due to bonded and non-bonded terms [6,21]:

\[ U(\vec{r}) = \sum U_{\text{bonded}}(\vec{r}) + \sum U_{\text{non-bonded}}(\vec{r}) \]  \hspace{1cm} (2.10)

### 2.1.2.1 CHARMM force field

Our MD simulations use the CHARMM (Chemistry at HARvard Macromolecular Mechanics) force field to describe the potential energy of particles, specifically CHARMM27 [46,47]. In this force field, the bonded interactions are handled by bond stretching, angle bending, dihedral, and improper dihedral energy terms as shown below:

\[ U_{\text{bonded}} = U_{\text{bonds}} + U_{\text{angles}} + U_{\text{Urey-Bradley}} + U_{\text{dihedrals}} + U_{\text{impropers}} + U_{\text{CMAP}} \]  \hspace{1cm} (2.11)
In the above equation, the first term takes care of the bond stretches, where $k_b$ is the bond force constant, $b_0$ is the equilibrium bond length. So, $b - b_0$ is the distance atom has moved from equilibrium. The force constant $k_b$ determines the strength of the bond. The second terms accounts for the bond angles, where $k_\theta$ is the angle force constant, and $\theta - \theta_0$ represents the deviation of angle from equilibrium between three bonded atoms. Similar to the first term, this term is also represented by a harmonic potential. The third term, "Urey-Bradley" term is an interaction based on distance between atoms separated by two bond (1,3 interaction). This term was introduced on a case by case basis to improve agreement with vibrational spectra. The Urey-Bradley term is considered to be crucial for in plane deformations and for separating symmetric and asymmetric bond stretching modes. Here $k_{UB}$ is the force constant, and $u$ is the distance between 1,3 atoms. The fourth term is for dihedrals or torsion angles, where $k_\varphi$ is the dihedral force constant, $n$ is the multiplicity of the function, $\varphi$ is the dihedral angle, and $\delta$ is the phase shift. This term models the presence of steric barriers between atoms separated by 3 covalent bonds (1,4 pairs). Here, the associated motion is rotation, described by coefficient of symmetry $n$, and dihedral angle around the middle bond. The torsion angle potential is assumed to be periodic and hence expressed as a cosine function. The out of plane bending or impropers is handled by the fifth term, where $k_\omega$ is the improper force constant, and $\omega - \omega_0$ denotes the out of plane angle [46]. The final term (CMAP term) is used for improving conformational properties of protein backbones. This is a cross-term for the $\Phi, \Psi$ (backbone dihedral angle) values, determined by grid based energy correction maps [47].

In the CHARMM force field, the non-bonded interactions are estimated using the Lennard-
Jones and Coulomb interactions [46,47]:

\[ U_{\text{non-bonded}} = U_{\text{LJ}} + U_{\text{elec}} \]  
\[ (2.13) \]

\[ U_{\text{non-bonded}}(\vec{r}) = \sum_{\text{nonb.pairs}} \epsilon_{ij} \left[ \left( \frac{R_{\text{min},ij}}{r_{ij}} \right)^{12} - 2 \left( \frac{R_{\text{min},ij}}{r_{ij}} \right)^{6} \right] + \sum_{\text{nonb.pairs}} \frac{q_i q_j}{\epsilon r_{ij}} \]  
\[ (2.14) \]

In equation 2.14, the first term or LJ term models the Van der Waals interactions. Here \( \epsilon_{ij} \) is the LJ well depth, \( r_{ij} \) is the distance between pair of atoms, and \( R_{\text{min},ij} \) is the minimum interaction radius. Usually, \( \epsilon_i, R_{\text{min},i} \) are obtained for individual atom types, and they are combined with appropriate mixing rules to yield the corresponding parameters for atom i-atom j interactions. For instance, the \( \epsilon_{ij} \) values are obtained via the geometric mean:

\[ \epsilon_{ij} = \sqrt{\epsilon_i \epsilon_j} \]  
\[ (2.15) \]

Similarly, \( R_{\text{min},ij} \) values are obtained via the arithmetic mean. Again, \( R_{\text{min}} \) represents the radius at which LJ function is a minimum:

\[ R_{\text{min},ij} = \frac{R_{\text{min},i} + R_{\text{min},j}}{2} \]  
\[ (2.16) \]

The final term of equation 2.14 is the Coulombic term where \( q \) is the partial atomic charge, \( \epsilon \) is the dielectric constant, and \( r_{ij} \) is the distance between atoms i and j.

As a whole, the energy function used in CHARMM27 is [46,47]:

\[ U(\vec{r}) = \sum_{\text{bonds}} k_b (b - b_0)^2 + \sum_{\text{angles}} k_\theta (\theta - \theta_0)^2 + \sum_{\text{Urey-Bradley}} k_{UB} (u - u_0)^2 \]
\[ + \sum_{\text{dihedrals}} k_\varphi [1 + \cos(n\varphi - \delta)] + \sum_{\text{impropers}} k_\omega (\omega - \omega_0)^2 \]
\[ + \sum_{\text{residues}} u_{\text{CMAP}}(\Phi, \Psi) + \sum_{\text{nonb.pairs}} \epsilon_{ij} \left[ \left( \frac{R_{\text{min},ij}}{r_{ij}} \right)^{12} - 2 \left( \frac{R_{\text{min},ij}}{r_{ij}} \right)^{6} \right] \]
\[ + \sum_{\text{nonb.pairs}} \frac{q_i q_j}{\epsilon r_{ij}} \]  
\[ (2.17) \]
Calculating force is the most computationally demanding part in MD simulations [21]. Among the force calculation, evaluating the non-bonded interactions is the most time consuming part as it involves almost all pairs of particles. For example, in a pairwise model to compute the interaction of N atoms, we need $N^2$ steps. One method to avoid this issue is to use a cut-off beyond which interactions are ignored. For LJ interactions, this is a reasonable approximation as it decay rapidly for large distances. However this is not a good option for electrostatic interactions, as it decays slowly as a function of $1/r$ [21].

2.1.2.2 Particle-Mesh-Ewald

Particle-Mesh-Ewald (PME), a fast incarnation of Ewald summation is the frequently used algorithm to solve electrostatic interactions [55]. To understand PME, let us first understand Ewald sums. Let us assume the system as a collection of charged particles in a cubic box with side length L. The system is electrically neutral, meaning we have equal number of positive and negative charges. We assume all particles repel each other at sufficiently short distances and periodic boundary conditions (explained in the following section). So our goal is to compute the Coulomb contribution to the potential energy of this N-particle system [6,21],

$$U_{Coul} = \frac{1}{2} \sum_{i=1}^{N} q_i \phi(r_i)$$  \hspace{1cm} (2.18)

Here $q_i$ is the charge, and $\phi(r_i)$ is the electrostatic potential at the position of ion i:

$$U(\vec{r}) = \sum_{j=1}^{N} \sum_{n \in \mathbb{Z}^d} \frac{q_j}{|r_{ij} + nL|}$$  \hspace{1cm} (2.19)

Here $n$ is a three dimensional integer vector. The prime on the first summation indicates that we will not include the term for which $j = i$ if $n = (0,0,0)$. By incorporating this restriction, we allow each particle to interact with its periodic images, but not with itself. We cannot use equation 2.19 to compute electrostatic energy in a simulation, because it contains a poorly converging sum [6,21]. Basically this sum decays like $1/|\vec{n}|$ and is a conditionally converging series, meaning that it converges ($\sum_{i=1}^{\infty} a_i < \infty$) but does not converges absolutely ($\sum_{i=1}^{\infty} |a_i|$ cannot be summed up). The idea of Ewald methods is to convert this difficult convergent series to a sum of two series, both of which converge much rapidly. So to evaluate
$U_{\text{Coul}}$ effectively, we can break it into two parts: 1) a short-ranged potential treated with a cut-off 2) a long-ranged potential that is periodic and slowly varying. This potential can be represented by a finite Fourier series [3].

In Ewald methods, we do the following: first screen each point charge using a diffuse cloud of opposite charge around each point charge, then compensate for the screening charges using a smoothly varying periodic charge density. This screening charge is constructed so as to make the electrostatic potential due to a charge at position $r_j$ decays rapidly to near zero at a particular distance. These interactions are treated in real space. Then the compensating charge density, which is the sum of all screening densities except with opposite charges are calculated using Fourier series. The standard choice for the screening potential is Gaussian [6,21]:

$$\rho_s(r) = -q_i (\alpha/\pi)^{3/2} \exp (-\alpha r^2) \quad (2.20)$$

At the end, there are three energy contributions to the Ewald sum:

$$U_{\text{Coul}} = U_{\text{real}} + U_{\text{rec}} - U_{\text{self}} \quad (2.21)$$

Here $U_{\text{real}}$ deals with the short-range contribution and $U_{\text{rec}}$ or the reciprocal-space contribution handles the long-range interactions. These long-range interactions become short-ranged in reciprocal space. Here, we have to subtract the self-interaction energy, $U_{\text{self}}$ because we broadened the charged distributions while substituting point charges with gaussian charge distributions. Equation 2.21 can be further written as:

$$U_{\text{Coul}}(\vec{r}) = \frac{1}{2} \sum_{i \neq j}^N q_i q_j \frac{\text{erfc}(\sqrt{\alpha r_{ij}})}{r_{ij}} + \frac{1}{2V} \sum_{k \neq 0} \frac{4\pi}{k^2} |\rho(k)|^2 \exp(-k^2/4\alpha) \quad (2.22)$$

$$- \left(\frac{\alpha}{\pi}\right)^{1/2} \sum_{i=1}^N q_i^2$$

where

$$\rho(k) = - \sum_{i=1}^N q_i \exp(ik \cdot r_i) \quad (2.23)$$
In equation 2.22, \( q_i, q_i \) are the two charges, \( \alpha \) is the Ewald parameter, and \( V \) is the volume of the system. Here, \( k \) is the reciprocal lattice vector:

\[
k = (2\pi/L)l
\]

where \( l = (l_x, l_y, l_z) \) are the lattice vectors in Fourier space.

For a fully optimized Ewald summation, the computational time scales with the number of particles as \( O(N^{3/2}) \). PME is based on formulating equation 2.18 so as to use Fast Fourier Transform to obtain better scaling [16,21]. In PME, charges are mapped on to grid points. Basically, a charge is distributed over grid points with weighting functions chosen according to the distance of the corresponding grid point to the charge. This idea of positioning all charges on a grid enables the application of FFT method and reduces the computational cost. The computer time associated with FFT technique scales as \( O(N \log N) \), where \( N \) is the number of points of discrete Fourier transform [16]. This method has been improved over the years by splitting the calculation into two groups, short-range and long-range contributions. The short-range contributions are calculated directly from particle-particle interactions and PME is used for the long-range contribution [21]. PME is used in conjunction with periodic boundary conditions (PBC).

### 2.1.2.3 Periodic Boundary Conditions

The main purpose of periodic boundary conditions (PBC) is that it mimics the presence of an infinite bulk surrounding the \( N \) particle system. Let’s think about this little more: we do MD simulations with the purpose of obtaining information about the properties of a macroscopic sample. No matter how large our system is, its number of atoms \( N \) is negligible when compared with the number of atoms contained in a macroscopic piece of matter (order of \( 10^{23} \)). Also, the ratio of number of surface atoms to total number of atoms will be much higher than in reality, making surface effects to be much more relevant than they should be [21]. To solve this issue we can use PBC. Here is the basic idea of PBC: the volume containing \( N \) particles is considered as the primitive cell of an infinite periodic lattice of identical cells (boxes) as shown in the following figure:
During the simulation, as a molecule moves in the original box, its periodic image in each of the neighboring boxes moves in the same way. So if a particle leaves the central box to the right, it is replaced by the corresponding particle from the image of the central box on the left. In other words, there are no walls at the boundary of central box, and no surface molecules. The number density in the central box (and entire system) is conserved.

In PBC, each particle interacts with all other particles in the infinite periodic system, meaning each particle interacts with all other particles in the same periodic cell and all particles in all other cells including its own periodic image. As you can see this would
increase the computational cost. So to make the simulation more efficient, we need to truncate these interactions beyond a certain distance [21]. For this purpose, GROMACS uses minimum image convention, where only one (the nearest image) image of each particle is considered for short range non-bonded interactions. The minimum image convention also puts another restriction: the cut-off radius used for truncating non-bonded interactions may not exceed half the shortest box vector, otherwise more than one image will be present within the cut-off.

Another factor that influences the electrostatic interactions is the solvent.

2.1.3 Integrating the Equations of Motion

As discussed before, MD simulation is based on Newton’s second law, \( F = ma \), and we know, velocity is the derivative of position and acceleration is the derivative of velocity. So we can write the equations of motion as:

\[
v = \frac{dr}{dt} \quad (2.25)
\]

\[
\frac{F(r)}{m} = \frac{dv}{dt} \quad (2.26)
\]

So, computing the trajectories involves solving ordinary differential equations. For a system \( N \) atoms, we have \( 3N \) position coordinates and \( 3N \) velocity coordinates. For such a system, we need to consider numerical solution as analytical solution is impossible [6, 21]. Various numerical algorithms have been developed for integrating the equations of motion. Many of these algorithms are finite difference methods. The fundamental principle of finite differences is to break the integration into small steps. At time \( t \), the force acting on a particle can be calculated as a vector sum of its interactions with other particles in the system. Then using equation 2.26, acceleration \( a = \frac{dv}{dt} \) can be determined. All these algorithms assume positions, velocities and accelerations can be approximated by a Taylor series expansion [6, 21]:

\[
\vec{r}(t + \delta t) = \vec{r}(t) + \vec{v}(t)\delta t + \frac{1}{2}\vec{a}(t)\delta t^2 + \frac{1}{3!}\vec{b}(t)\delta t^3 + O(\delta t^4) \quad (2.27)
\]

\[
\vec{v}(t + \delta t) = \vec{v}(t) + \vec{a}(t)\delta t + \frac{1}{2}\vec{b}(t)\delta t^2 + \frac{1}{3!}\vec{c}(t)\delta t^3 + O(\delta t^4) \quad (2.28)
\]
\[ \vec{a}(t + \delta t) = \vec{a}(t) + \vec{b}(t)\delta t + \frac{1}{2} \vec{c}(t)\delta t^2 + \frac{1}{3!} \vec{d}(t)\delta t^3 + O(\delta t^4) \] (2.29)

where \( \vec{r}(t), \vec{v}(t), \vec{a}(t) \), are the position, velocity, and acceleration. The remaining terms, \( \vec{b}(t), \vec{c}(t), \text{ and } \vec{d}(t) \) are first, second and third derivatives of the acceleration respectively. Although there are many criteria for choosing a good integrator, the following three are the most important ones [21]:

1. Energy and momentum Conservation

We know Newton’s equations of motion conserve energy. In the case of MD simulations, we need to think about long time and short time (from step to step) energy conservation. It would be ideal to have an algorithm that has good energy conservation for both short and long times, but no such algorithm exists. The more sophisticated higher order algorithms have very good short time energy conservation but significant drifts in energy for long times. The Verlet-style algorithms can come to the rescue here; they have moderate short time energy conservation but little long-term drifts [21].

When using MD simulations, the trajectory of any system through phase space depends sensitively on initial conditions. Here phase space stands for a \( 6N \) dimensional space in which there is an axis for every position and every conjugate momentum. So for Cartesian coordinates, that means there are \( x, y, z \) and \( p_x, p_y, p_z \) axes for each atom. The dependency on initial conditions means as time progresses, two trajectories that are initially very close will diverge exponentially [21]. Let us consider the position of one of the \( N \) particles at time \( t \). We know that this position is a function of the initial positions and momenta at time, \( t = 0: \)

\[ r(t) = f \left[ r^N(0), p^N(0) \right] \] (2.30)

If we consider the same position after perturbing the initial conditions, say the momenta is shifted by a small amount \( \epsilon \), we get a different value for \( r \):

\[ r'(t) = f \left[ r^N(0), p^N(0) + \epsilon \right] \] (2.31)

Now, let us say \( \Delta r(t) \) is the difference:

\[ \Delta r(t) = r(t) - r'(t) \] (2.32)
For shorter times, $\Delta r(t)$ is linear in $\epsilon$. But for longer times, the coefficient of linear dependence diverges exponentially as shown below:

$$|\Delta r(t)| \sim \epsilon \exp(\lambda t)$$  \hspace{1cm} (2.33)

Here $\lambda$ is called the Lyapunov exponent. The above relationship is called the Lyapunov instability, and it shows even small deviations in the starting conditions of trajectory results in large differences between trajectories at long times. That means, no matter how small the integration error, it will always cause the simulated trajectory to diverge exponentially from the true trajectory compatible with the same initial conditions. However, this is not an issue, because in MD simulations, we are interested in the average behavior of a system that was prepared in an initial state of which we know something not everything [21]. Basically we are interested in statistical properties of long simulations as ultimately it is the statistics that relate our data to experiments and to thermodynamics. Regardless of the intentions, we cannot fully justify the use of inaccurate trajectories unless they are close to true trajectories in some sense [21]. Evidence shows that there exist shadow orbits, which are true trajectories of a many-body system that closely follows the numerical trajectory for a time that is long compared to the time it takes the Lyapunov instability to develop. Another way to think of shadow orbit is that, it is the true trajectory to which the numerical trajectory overlaps for a certain period of time. Once the Lyapunov instability grows, the numerical trajectory will get far from that specific shadow orbit, but there will be another one to which it is superimposed.

2. Time step

Although the speed of the integrator seems important at first thought, this is insignificant when compared to the calculation of all forces within the system. That means it is beneficial to minimize the number of force calculations during the simulation. One way to do this is to make the force calculations less frequent, in other words choose a longer time step [21]. So it is crucial we choose an algorithm that guarantees accuracy for a long time step. Algorithms permit the use of a long time step by storing information on higher-order derivatives of particle coordinates. This in turn increases the memory storage requirements. Also, to conserve energy we need to pick a time step that is at least an order of magnitude lesser
than the fastest motions in the system, usually the hydrogen bond vibrations.

3. Time reversibility and Area preserving nature

Newton’s equations are time reversible, and so the integrators should preserve this property. That means if the velocities of all atoms were swapped in sign, the simulation would run exactly in the reverse direction. Basically, the past and future phase space coordinates play a symmetric role [21]. To understand the area preserving nature, let us assume, the trajectories that have a particular energy \( E \) are contained in a hyper volume \( \Omega \) in phase space. If we let the volume evolve in time according to Hamilton’s equations of motion, we will get exactly the same volume. But a non-area-preserving algorithm will usually map the volume to a larger volume say \( \Omega' \). Over time, this would cause the system to expand significantly, thus violating the energy conservation. So it is important to have an integrator, that follows the area preserving nature of Hamiltonian dynamics [21].

The most commonly used integrators, Verlet, Velocity-Verlet and Leap-frog algorithms are explained in the following sections.

### 2.1.3.1 Verlet algorithm

The Verlet algorithm uses positions and accelerations at time \( t \) and the positions from \( t - \delta t \) to calculate new positions at time \( t + \delta t \). To derive the Verlet algorithm, we can write [21]:

\[
\begin{align*}
    r(t + \delta t) &= r(t) + \frac{dr(t)}{dt} \delta t + \frac{1}{2} \frac{d^2 r(t)}{dt^2} \delta t^2 + \frac{1}{3!} \frac{d^3 r(t)}{dt^3} \delta t^3 + O(\delta t^4) \\
    r(t - \delta t) &= r(t) - \frac{dr(t)}{dt} \delta t + \frac{1}{2} \frac{d^2 r(t)}{dt^2} \delta t^2 - \frac{1}{3!} \frac{d^3 r(t)}{dt^3} \delta t^3 + O(\delta t^4)
\end{align*}
\]

Adding the above two equations, we get:

\[
r(t + \delta t) = 2r(t) - r(t - \delta t) + \frac{1}{2} \frac{d^2 r(t)}{dt^2} \delta t^2 + O(\delta t^4)
\]

In Verlet algorithm, velocities are not explicitly solved, they are usually calculated from first order central difference as shown below:

\[
v(t) = \frac{r(t + \delta t) - r(t - \delta t)}{2\delta t}
\]
The advantages of verlet algorithm are a) it is straightforward and b) it has moderate storage requirements. However this algorithm is of moderate precision [21].

2.1.3.2 Velocity-Verlet Algorithm

Velocity-Verlet algorithm has an improved accuracy when compared to the standard Verlet algorithm. Since this algorithm yields positions, velocities and accelerations at the same time t, it minimizes round-off errors. Here, we first calculate the velocities at mid time step [21].

\[
v(t + \frac{\delta t}{2}) = v(t) + \frac{1}{2} a(t) \delta t
\]  

(2.38)

Then using this velocity, we calculate positions at the next step:

\[
r(t + \delta t) = r(t) + v \left( t + \frac{\delta t}{2} \right) \delta t
\]  

(2.39)

Using the potential, we calculate accelerations at the next step. Then we use these accelerations to update the velocities at next step:

\[
v(t + \delta t) = v \left( t + \frac{\delta t}{2} \right) + \frac{1}{2} a(t + \delta t) \delta t
\]  

(2.40)

Other popular algorithms that are closely related to Verlet algorithm are Beeman algorithm and Velocity-corrected Verlet algorithm [21]. Both of these algorithms use more accurate expressions for velocities. However the complex expressions make the computational cost even more expensive.

2.1.3.3 Leap-frog Algorithm

Leap-frog algorithm is also derived from the Verlet algorithm. In this algorithm, we first calculate velocities at time \( t + \frac{1}{2} \delta t \) [21]:

\[
v \left( t + \frac{\delta t}{2} \right) = v \left( t - \frac{\delta t}{2} \right) + a(t) \delta t
\]  

(2.41)

Then these velocities are used to calculate the positions, \( r \), at time \( t + \delta t \):

\[
r(t + \delta t) = r(t) + v \left( t + \frac{\delta t}{2} \right) \delta t
\]  

(2.42)
So in this algorithm, velocities leap over positions and then the positions leap over the velocities. Here velocities are explicitly calculated, however they are not calculated at the same time as the positions. The velocities at time $t$ can be estimated from the equation:

$$v(t) = \frac{1}{2} \left[ v \left( t - \frac{\delta t}{2} \right) + v \left( t + \frac{\delta t}{2} \right) \right]$$ (2.43)

Studies show that leap-frog algorithm is more accurate than standard Verlet algorithm [5]. Similar to Verlet algorithm, it has moderate memory requirements. We used the md integrator (leap-frog algorithm) of GROMACS. According to GROMACS, for almost all production runs, md (Leap-frog) integrator is accurate enough when compared to md-vv (Velocity-Verlet) integrator [68].

There exists many higher order integration schemes, where Taylor expansions are taken to higher terms. One of the most popular class of higher-order algorithms is predictor-corrector algorithms. It requires more storage and is neither time reversible nor area preserving [21].

### 2.1.4 Experimental Conditions

In MD simulations, it is crucial to accurately simulate the experimental conditions that need to be replicated. In our case, when we investigate the conformations of a protein, we need to mimic physiological conditions as best as we can. That means we need to consider pressure, temperature and the protein’s environment.

#### 2.1.4.1 Ensembles

An ensemble is a collection of all possible systems that have different microscopic states but belong to a single macroscopic or thermodynamic state [21]. There are different types of ensembles of which the most common ones are: a) The microcanonical ensemble (NVE), which has constant number of particles (N), constant volume (V), and constant energy (E). Since energy is conserved, this is a closed system. b) The canonical ensemble (NVT), characterized by fixed number of particles (N), fixed volume (V), and fixed temperature (T). c) The grand canonical ensemble ($\mu$VT): the thermodynamic state characterized by a
constant chemical potential ($\mu$), constant volume (V), and constant temperature (T). d) The isobaric-isoenthalpic ensemble (NPH), which has constant number of particles (N), constant pressure (P), and constant enthalpy (H). e) The isobaric-isothermal ensemble, characterized by fixed number of particles (N), fixed pressure (P), and fixed temperature (T) [21].

Most of the early MD simulations were performed in the NVE ensemble. This is because without any external control, ideally MD generates a microcanonical (NVE) ensemble [6,21]. However, as most of the experiments are performed at constant temperature and volume or at constant temperature and pressure, NVT and NPT ensembles became more popular for MD simulations. For our study reported here, we use NPT ensemble.

In NPT simulations, to make sure the average temperature of a system is maintained correctly, we need to use a thermostat. The most commonly used temperature control schemes in MD are Berendsen, Nosé-Hoover and Langevin dynamics [21]. We used Nosé-Hoover thermostat.

2.1.4.2 Nosé-Hoover Thermostat

This popular thermostat retains a Boltzmann equilibrium distribution. The basic idea of this approach is to modify the system’s Hamiltonian by incorporating a thermal reservoir and a friction in the equations of motion. Now this friction, $\zeta$ slows down or accelerate the particles until the temperature (measured through kinetic energy and the equipartition function) is equal to the desired value. Now the equations of motion are [27,28]:

$$m_i \frac{d^2 r_i}{dt^2} = F_i - \zeta m_i v_i$$  \hspace{1cm} (2.44)

$$\frac{d\zeta(t)}{dt} = \frac{1}{Q} \left[ \sum_{i=1}^{N} m_i \frac{v_i^2}{2} - \frac{3N+1}{2} k_B T \right]$$  \hspace{1cm} (2.45)

In the above equation $Q$ determines the relaxation of the dynamics of the friction $\zeta(t)$, while $T$ denotes the target temperature. So, at steady state when $\frac{d\zeta}{dt} = 0$, the kinetic energy is given by $\frac{3N+1}{2} k_B T$ as required by the equipartition theorem. Here, instead of $3N$ there is a factor of $3N + 1$ because of the extra degree of freedom, $\zeta$. Again, it is important to recognize the temperature is not fixed, rather it tends to the desired value.
Although Nosé-Hoover thermostat is widely used in the MD simulation community, at times it can exhibit non-ergodic behavior for specific systems [21]. One of the methods to improve this is to use the Nosé-Hoover chains. In this approach, we control the extra degree of freedom by adding another Nosé-Hoover thermostat...and then to control that degree of freedom with another thermostat, etc. [21,68].

Now to perform the NPT simulation, we need to control the pressure using a barostat. Some of the most popular barostats include Berendsen, Andersen and Parrinello-Rahman [21,68]. The combination of Nose-Hoover thermostat and Parinello-Rahman barostat is considered more reliable for simulation at equilibrium and for predicting thermodynamic properties [20]. So, we used Parinello-Rahman barostat for our simulations of LC3 protein.

2.1.4.3 Parrinello-Rahman Barostat

In thermodynamics, systems at constant pressure exchange volume with their surroundings using a piston. Similarly, constant pressure simulations involve volume fluctuations. According to Parrinello-Rahman barostat, the box vectors are represented by the matrix $b$, and it obeys the equation of motion [53,54,68]:

$$\frac{db^2}{dt^2} = VW^{-1}b'^{-1}(P - P_{ref})$$ (2.46)

Here, $V$ is the volume of the box, $W$ is a matrix parameter that estimates the strength of coupling, $P$ is the current pressure matrix and $P_{ref}$ is the reference pressure matrix.

The equations of motion now become:

$$m_i \frac{d^2 r_i}{dt^2} = F_i - m_i M \frac{dr_i}{dt}$$ (2.47)

$$M = b^{-1} \left[ \frac{db'}{dt} + \frac{dt}{dt} b' \right] b'^{-1}$$ (2.48)

The inverse mass parameter matrix $W^{-1}$ is calculated automatically in GROMACS using the following equation:

$$(W^{-1})_{ij} = \frac{4\pi^2 \beta_{ij}}{3\tau_p^2 L}$$ (2.49)
In the above equation, $\beta$ is the isothermal compressibility, $\tau_p$ is the pressure time constant, and $L$ is the largest box matrix element. While doing the simulation in GROMACS, the user has to give the approximate isothermal compressibilities and the pressure time constant [68].

2.1.4.4 Solvent

To make the protein simulations more realistic, we need to replicate the protein’s environment more accurately. As most proteins exist in aqueous environments, in MD we use either the implicit solvation method or the explicit water models. The basic idea of implicit solvent model is to represent the electrostatic effects of solvent, while saving the computational cost associated with doing explicit solvent model simulations [68]. However, the implicit solvent models cannot reproduce certain microscopic features of the solvent [60]. For our study, we used explicit water models, where water is represented by all-atom force fields.

Among the several water models available, the TIP3P (transferable intermolecular potential 3P) and SPC/E (extended simple point charge) are the most water models for protein simulations. Both these water models use three sites for electrostatic interactions; the negative charge on oxygen atom balances the partial positive charges on the hydrogen atoms. In this model, the intermolecular interaction between two water molecules is computed using a Lennard-Jones type potential with interaction points centered on the oxygen atoms. This Lennard-Jones term along with the Coulombic interactions between all intermolecular pairs of charges defines the dimerization energy for monomer $m$, and $n$ as shown below [31]:

$$
\epsilon_{mn} = \sum_i^n \sum_j^m \frac{q_i q_j e^2}{r_{ij}} + \frac{A}{r_{oo}^{12}} - \frac{C}{r_{oo}^6} 
$$

(2.50)

Here, $\epsilon_{mn}$ is the dimerization energy, $q$ is the charge, $r_{ij}$ is the distance between pair of atoms, and $A, C$ are the Lennard-Jones parameters. The simplicity and high computational efficiency makes the three site water models highly popular for simulations. For this study, we used TIP3P water model as CHARMM parameters have been designed for solvation with TIP3P [48]. The SPC/E and TIP3P models differ in parameters.
2.1.5 Limitations of MD simulations

Sampling conformational space of proteins is one of the most challenging problems in structural biology. This is due to the fact that proteins don’t have a smooth energy landscape, rather the energy landscape is locally rugged, hence escaping from sub-valleys (local minima) is extremely difficult [23]. Although molecular dynamics (MD) is a powerful tool to simulate the time evolution of molecular systems, it suffers from time scale issues when sampling conformational space of proteins [44,65,70]. Basically in regular MD, proteins can get caught in the local minima and thus not accessing the global minima as shown in the following representation:

![Free Energy Landscape](image)

**Figure 2.3:** (a) Simplified representation of a free energy landscape of a protein (b) Schematic representation of free energy landscape for a protein in regular MD

2.2 Improved Sampling Techniques

Many methods have been developed over the years to effectively accelerate MD, so it can access all possible conformations within a reasonable time and with relatively low computational cost. These methods can be broadly classified into two groups: collective-variable-based methods and parallel tempering methods [2]. The first method is based on the idea of choosing, a priori, a set of collective varaiables which are then biased during the simulation. Examples of this method include metadynamics [42] and umbrella sampling [67]. The main drawback of this method is that the efficiency and accuracy are decided by the choice of collective variables. For systems with rough energy landscapes such as proteins, deciding proper collective variables is even more challenging [44,70]. The second method, parallel
tempering is based on the idea of raising temperature so as to accelerate sampling. One of the most popular advanced MD techniques for protein conformational sampling, replica exchange molecular dynamics (REMD) is based on the parallel tempering method [2].

2.2.1 REMD

The replica exchange molecular dynamics (REMD) was developed by Yuji Sugita and Yuko Okamoto to overcome protein folding issues with multiple minima [63]. In this method, several non-interacting replicas of the same system are simulated at different temperatures. Then using the Metropolis acceptance criterion, coordinates of the neighboring replicas are exchanged at specific intervals. The Metropolis criterion used for replica exchanges is the standard Boltzmann distribution criterion, so the coordinates are swapped between two replicas with the acceptance probability [63]:

$$\alpha = \min\left(1, \exp\left(\frac{1}{k_B T_2} - \frac{1}{k_B T_1}\right)[U(x_1) - U(x_2)]\right)$$

(2.51)

Here, $k_B$ is the Boltzmann constant, $T_1$ and $T_2$ are the temperatures of replica 1, 2 and $U(x_1), U(x_2)$ are the potential energies of replicas 1, 2.

A pictorial representation of REMD is shown below. In this sample REMD, we have isothermal MD simulations running in parallel at four different temperatures:

![Simplified representation of replica exchange molecular dynamics (REMD).](image)

Although REMD improves the protein’s conformational sampling [2, 63], the number of replicas scales as $O(f^{\frac{1}{2}})$, where $f$ is the number of degrees of freedom for the whole system. Also, with REMD membranes fall apart at higher temperatures [44]. Since we are simulating LC3-II (membrane associated form of LC3), REMD is not a viable option for our study.
2.2.2 HREMD

Another class of methods, which is sort of in an intermediate position between collective-variable-based methods and parallel tempering methods, is the Hamiltonian Replica Exchange (HREX) methods [13], where different replicas evolve according to different Hamiltonians. In HREX, the dependence of results on the choice of modified Hamiltonian is smaller than the dependence of umbrella-sampling efficiency on the choice of collective variables [13].

We can connect Hamiltonian replica exchange methods (HREX) and replica exchange methods (REM) in the following way: HREX (similarly HREMD or Hamiltonian replica exchange molecular dynamics) is more of a generalized approach and REM (similarly REMD) is a specific case. Let us try to understand this better. If we consider a system with potential energy $U(r)$, and coupled to a thermal bath with temperature $T$, we know that the probability of exploring a particular configuration is governed by the following relationship [13]:

$$P(r) \propto \exp\left(-\frac{U(r)}{k_B T}\right)$$  \hspace{1cm} (2.52)

where $r$ is the position coordinate, and $k_B$ is the Boltzmann constant.

The basic idea of replica exchange methods is to sample from one "cold" replica for unbiased statistics and use several "hot" replicas to accelerate sampling. Ideally, the hottest replica should explore the phase space quick enough to overcome the barriers for the system under study, and the intermediate replicas are designed to ensure a smooth transition from the "hottest" replica to the "coldest" replica [63]. This "hot" and "cold" can have different meanings in different methods. For example, in parallel tempering methods like REMD, "hot" and "cold" stands for the actual physical temperature controlled by a thermostat, whereas in HREX methods "hot" replicas can be biased in an arbitrary manner [13, 44, 64]. So, we can think of the most general set up as the following: each replica simulated at a different temperature using a different Hamiltonian. If we consider a system of $N$ number of replicas, the product ensemble now becomes:

$$P(r_1) \times P(r_2) \times ... \times P(r_N) \propto \exp\left(-\frac{U_1(r_1)}{k_B T_1} - \frac{U_2(r_2)}{k_B T_2} - ... - \frac{U_N(r_N)}{k_B T_N}\right)$$  \hspace{1cm} (2.53)
CHAPTER 2. METHODOLOGY

For the specific case when, $U_1 = U_2 = ... = U_N$, we get REMD. Also, in the equilibrium thermodynamics, changing the temperature of a system is equivalent to scaling the entire Hamiltonian of the system, and so temperature REM corresponds to a special case of Hamiltonian replica exchange methods [13,23].

The benefit of using potential energy for scaling instead of temperature is related to it’s property. Potential energy is an extensive property, meaning it depends on the size of the system. That means, we can select a portion of the system and specific parts of the Hamiltonian to be "heated". However, temperature is an intensive property, means it is independent of the size of the system.

Compared to parallel tempering methods, Hamiltonian replica exchange method is more efficient as the number of required replicas are much less [13,64]. To understand this better, let us go through the basics of HREX in detail. If we consider a system similar to a protein, it consists of relatively few relevant coordinates $r$ and many other coordinates $x$ weakly coupled to $r$. Examples of $r$ include protein atomic coordinates, and torsional degree of freedom. The solvent coordinates, protein bond length and protein bond angle degrees of freedom can be considered as $x$. Although including $x$ is crucial for the simulation, it increases the degree of freedom and further increase the number of replicas in the parallel tempering methods. Now, for our current system, the Hamiltonian is written as [23]:

$$E(r,x) = E_s(r) + E_{sb}(r,x) + E_b(x)$$ \hspace{1cm} (2.54)

where $s$ denotes system, $b$ stands for bath. Now, the idea is to introduce a series of replicas with following Hamiltonian:

$$E_m(r,x) = s_m[E_s(r) + E_{sb}(r,x)] + E_b(x)$$ \hspace{1cm} (2.55)

Here, $s_m = \frac{\beta_m}{\beta}$ is the "effective temperature" scaling for the degree of freedom $r$. As we can see the "effective temperature" of $x$ coordinates is unchanged. Therefore, the appropriate spacing in $\beta_m$ is controlled by the number of degrees of freedom for $r$, which is much smaller compared to the total degrees of freedom usually. Thus, the number of replicas required is significantly reduced in HREX [13,23].
Among the various HREX methods developed over the years, the most successful one is the replica exchange solute tempering (REST), with its REST2 variant [44, 70]. REST is developed by Prof. Bruce Berne and coworkers here at Columbia University in 2005 and the group developed the more efficient version of REST called REST2 in 2011. Basically, REST2 is REST with new scaling parameters for the Hamiltonians [70] and the following difference: both temperature and potential energy is different for each replica in REST1, whereas in REST2 all replicas run at same temperature but with different potential energy.

The basic idea of REST is to divide the system into two parts: protein as one part (central group) and water as the other part (bath). The choice of central group, and bath are flexible. For instance, one can use part of the protein as central group. With this approach, interaction energy of the system, $E(X)$ can be written as [44]:

$$E(X) = E_{pp}(X) + E_{pw}(X) + E_{ww}(X)$$  \hspace{1cm} (2.56)

Here, $X$ is the configuration of the whole system, $E_{pp}$ is the protein intermolecular energy, $E_{pw}$ is the interaction energy between protein and water, and $E_{ww}$ is the self-interaction energy between water molecules. As mentioned before, in REST2, all replicas run at same temperature but with different potential energy. For example, replica $m$ has the potential energy:

$$E(X) = \frac{\beta_m}{\beta_0} E_{pp}(X) + \sqrt{\frac{\beta_m}{\beta_0}} E_{pw}(X) + E_{ww}(X)$$  \hspace{1cm} (2.57)

where $\beta_0 = \frac{1}{k_B T_0}$, and $\beta_m = \frac{1}{k_B T_m}$. Here $T_0$ is the desired temperature and $T_m$ is the effective temperature of the protein with the unscaled potential energy. We can see that the equation 2.57 reduces to equation 2.56, for replica 0. As shown in the above equation, in REST2, improved sampling is obtained by scaling the intramolecular potential energy of the protein by $\frac{\beta_m}{\beta_0}$, a number smaller than 1, so that the barriers separating various conformations are reduced. The scaling factor for the coupling term or the interaction energy between solute and water is $\sqrt{\frac{\beta_m}{\beta_0}}$. Since, scaling the bond angle and bond stretch does not help the sampling, only dihedral angle terms in the bonded interaction of the solute are scaled in REST2 [70].

Now, to understand the acceptance probability for the exchange of configurations, let us
consider the exchange between \(mth\) and \(nth\) replicas [70]:

\[
(X_m, E_m(X_m), T_m) \rightarrow (X_n, E_m(X_n), T_m)(X_n, E_n(X_n), T_n) \rightarrow (X_m, E_n(X_m), T_n) \quad (2.58)
\]

Here, \(X_m, E_m(X_m), T_m\) are the configuration, energy and temperature of the \(mth\) replica right before an exchange is attempted. Similarly, expressions with \(n\) subscript correspond to the \(nth\) replica. Now, the equilibrium probability for the \(mth\) state is [23, 44, 70]:

\[
P_m = \frac{1}{Z_m} \exp(-\beta_m E_m(X_m)) \quad (2.59)
\]

where \(Z_m\) is the configurational partition function, and \(\beta_m = 1/(k_BT_m)\). Let us represent the transition probability for the exchange \(m \rightarrow n\) by \(T(m \rightarrow n)\), and for the reverse exchange by \(T(n \rightarrow m)\). Now, applying the detailed balance condition:

\[
P_m(X_m)P_n(X_n)T(m \rightarrow n) = P_n(X_m)P_m(X_n)T(n \rightarrow m) \quad (2.60)
\]

we can obtain the ratio of transition probabilities

\[
\frac{T(m \rightarrow n)}{T(n \rightarrow m)} = \exp(-\Delta_{mn}) \quad (2.61)
\]

where,

\[
\Delta_{mn} = (\beta_m - \beta_n) \left[ (E_{pp}(X_n) - E_{pp}(X_m)) + \frac{\sqrt{\beta_0}}{\sqrt{\beta_m} + \sqrt{\beta_n}} (E_{pw}(X_n) - E_{pw}(X_m)) \right] \quad (2.62)
\]

In the above equation, there is no \(E_{ww}\) term, the term that caused poor scaling with system size in the standard replica exchange methods like REMD. Now, if we apply the Metropolis criteria, we can obtain the acceptance probability as [23, 44, 70]:

\[
T(m \rightarrow n) = \begin{cases} 
1, & \text{if } \Delta_{mn} \leq 0 \\
\exp(-\Delta_{mn}), & \text{if } \Delta_{mn} > 0 
\end{cases}
\]

Since the potential energy surfaces of all replicas other than the lowest one are deformed, REST2 only samples the correct distribution for the lowest replica or replica 0 [44, 70]. This method can be used to study complex biological systems like protein attached to a membrane. Here one can consider protein as the central part and water along with membrane as the bath [70]. So for our simulations of LC3 protein attached to the membrane,
we considered LC3 protein as the central part and membrane with water as the bath. For the simulations of LC3 protein in solution, we treated LC3 protein as the central part and water as the bath. REST2 offers a clear advantage for complex bio-molecular systems, as membranes fall apart in regular replica exchange methods [44]. Recently, Prof. Bussi and co-workers came up with the implementation of REST2 in popular MD software GROMACS [13, 26]. They implemented the replica exchange methodology patched with the PLUMED plug-in [12]. This approach was constructed so that it could be used in a partial tempering scheme, meaning only the portion of a solute is heated. However, when used for the entire solute, it exactly reproduces REST2 [13], which is what we did for LC3 simulations. This implementation (HREX with PLUMED in GROMACS) was validated on alanine dipeptide in water and for the stability study of an RNA tetraloop [13]. A recent study also used the same HREMD implementation to study the conformational polymorphism in GATE-16 protein [45]. For our HREMD simulations, we used the solution structure of human LC3 (pdb code: 1v49) [11].

2.3 Analysis of Trajectories

To analyze the HREMD simulation trajectories of LC3 protein and its membrane associated form or simply “LC3_mem”, we used the following two methods.

2.3.1 Cluster Analysis

Cluster analysis identifies groups in data and separates them into different clusters [19, 61]. Here, members of the same cluster are similar whereas clusters vary from each other. Figure 2.5 shows schematic representation of a generic cluster analysis [19]. In our case, similar protein conformations are grouped into one cluster.
CHAPTER 2. METHODOLOGY

Figure 2.5: Schematic representation of a generic cluster analysis, adapted from reference [19].

We used two different types of cluster analysis namely, “Clustering Method I” and “clustering method II” to compare LC3, and LC3_mem. These approaches are explained in detail in chapter 4. Another thing to keep in mind is that, “Clustering Method I” is very similar to the clustering analysis we used for comparing LC3 and GATE-16 proteins in chapter 3. For details, please refer to chapters 3, and 4.

2.3.2 Root Mean Square Deviation (RMSD)

Root mean square deviation or RMSD measures the average distance in displacement for a selection of atoms of a particular time frame with respect to a reference time frame. RMSD is calculated by the following equation [10,39]:

\[
RMSD = \frac{1}{N} \sum_{i=1}^{N} (r_i(t) - r_i(t_{ref}))^2
\]

In the above equation, \(N\) is the number of atoms, \(r_i(t)\) is the position of \(i\)th atom at time \(t\) and \(r_i(t_{ref})\) is the position of \(i\)th atom at the reference time, \(t_{ref}\).
Chapter 3

Comparison Study: LC3 vs. GATE-16 (in solution)

As mentioned in Chapter 1, the eight human homologues of Atg8 are classified into two subfamilies: LC3 and GABARAP. One of the GABARAP subfamily members is GATE-16. In this Chapter, we are comparing our results to a published study, that used GATE-16 as a model to identify the conformational polymorphism in Atg8-type proteins via HREMD simulations [45].

3.1 Clustering Analysis

Ma and co-workers identified the stable conformations of GATE-16 (in solution) from HREMD simulations using clustering analysis. Similar to GATE-16 study, we are using clustering analysis to identify stable conformations of LC3 in solution. The criteria used for GATE-16 analysis is shown in Table 3.1, where HP stands for hydrophobic pocket formed by the following residues: methionine 1 (Met1), valine 36 (Val36), phenylalanine 79 (Phe79) and the aliphatic portion of lysine 82 (Lys82). Figure 3.1 shows this hydrophobic pocket of GATE-16. In order to compare the HREMD simulation results of LC3 to GATE-16, our first task was to identify a hydrophobic pocket in LC3, which is similar to the hydrophobic
pocket in GATE-16.

Table 3.1: Criteria for GATE-16 clustering analysis

<table>
<thead>
<tr>
<th>Criterion#</th>
<th>Criterion Details</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Residue for HP coverage</td>
<td>Phe115</td>
</tr>
<tr>
<td></td>
<td>Cutoff number of contacts</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>Orientation toward the N-terminus: residues for min. distance</td>
<td>Phe115ζ atom, Trp3</td>
</tr>
<tr>
<td></td>
<td>Cutoff distance</td>
<td>$\leq 7.2 , \text{Å}$</td>
</tr>
<tr>
<td>3</td>
<td>Residue for HP coverage</td>
<td>Phe117</td>
</tr>
<tr>
<td></td>
<td>Cutoff number of contacts</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>Residues for salt bridge</td>
<td>Met1, Phe117</td>
</tr>
<tr>
<td></td>
<td>Cutoff distance</td>
<td>$\leq 4 , \text{Å}$</td>
</tr>
</tbody>
</table>

Figure 3.1: Hydrophobic pocket of GATE-16 formed by Met1, Val36, Phe79, and aliphatic portion of Lys 82
CHAPTER 3. COMPARISON STUDY: LC3 VS. GATE-16 (IN SOLUTION)

Using DeepAlign, a structural alignment tool for proteins, we identified the hydrophobic pocket of LC3. We chose DeepAlign for the protein structure alignment due to the following reasons [71]:

1. Most of the protein structural alignment tools align two protein structures purely based on geometric similarity without considering functional and evolutionary relationship. However, DeepAlign takes into account the evolutionary information at the sequence and local substructure levels as well as the hydrogen bonding similarity in addition to the 3D geometric similarity.

2. Studies show DeepAlign generated structural alignments highly consistent with manually-curated alignments when compared to popular tools like DALI, MATT, Formatt, and TMalign.

To identify the hydrophobic pocket of LC3, we looked at the residues that are aligned with residues of GATE-16 hydrophobic pocket (Met1, Val36, Phe79 and Lys82), throughout the simulation. As shown in Figure 3.2, out of the many residues of LC3 that aligns with Met1 of GATE-16, serine 3 (Ser3) aligns the most, about 57% of the time.

![Figure 3.2: Aligned residues of LC3 compared to Met1 of GATE-16](image)

Figures 3.3, 3.4 show that LC3 residues tyrosine 38 (Tyr38), and leucine 82 (Leu82) align with GAET-16 residues Val36, and Phe 79 respectively throughout the simulation. Figure 3.5 captures the residues of LC3 that align with Lys82 of GATE-16; here about 78% during
the simulation, histidine 86 (His86) aligns with Lys82.

**Figure 3.3:** Aligned residues of LC3 compared to Val36 of GATE-16

**Figure 3.4:** Aligned residues of LC3 compared to Phe79 of GATE-16

**Figure 3.5:** Aligned residues of LC3 compared to Lys82 of GATE-16

Using Figures 3.2, 3.3, 3.4, and 3.5, we decided the candidates of hydrophobic pocket for LC3: Ser3, Tyr38, Leu82, and His86. Strictly speaking, we can’t call this a hydrophobic pocket as Ser, His are hydrophilic residues. However for the purpose of comparison, we are still going to designate it as hydrophobic pocket although only two out of four residues (Tyr, Leu) are hydrophobic. Figure 3.6b shows the visual representation of this identified hydrophobic pocket of LC3.
Our next objective was to identify the criteria for clustering method I, i.e. GATE-16 clustering analysis type criteria for clustering analysis of LC3 as shown in Table 3.2.

**Figure 3.6:** (a) GATE-16 structure highlighting hydrophobic pocket (HP) formed by Met1, Val36, Phe79, and aliphatic portion of Lys 82. HP is shown in red surface representation. (b) LC3 structure highlighting hydrophobic pocket (HP) formed by Ser3, Tyr38, Leu82, and His86. HP is shown in red surface representation.

<table>
<thead>
<tr>
<th>Table 3.2: Criteria for Clustering Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Criterion#</strong></td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
3.1.1 Criterion 1

For GATE-16 clustering analysis, criterion 1 is about the hydrophobic pocket coverage by Phe115 residue. To identify the corresponding residue for hydrophobic pocket coverage in LC3, we aligned the structures of LC3, and GATE-16 using DeepAlign. As shown in Figure 3.7, Phe119 of LC3 aligns with Phe115 of GATE-16 more than 99% of the time during the simulation.

In GATE-16, Phe115 is considered to be in the hydrophobic pocket (formed by Met1, Val36, Phe79 and Lys82) if at least 8 of the 14 side chain atoms of Phe115 are in contact with these four hydrophobic pocket residues. A contact is defined on the basis of minimal distance between each of Phe115’s side chain atoms and any of the atoms belonging to the residues forming the hydrophobic pocket. If this minimal distance is \( \leq 4 \, \text{Å} \), a contact is present [45].

Here, multiple contacts between a particular side chain atom of Phe115 and the hydrophobic pocket were counted as one, meaning the maximum number of hydrophobic pocket contacts for Phe115 is 14. The cutoff number of contacts 8, is determined from the distribution plot shown in Figure 3.8a. For LC3, we use similar definition to determine if Phe119 is inside the hydrophobic pocket formed by Ser3, Tyr38, Leu82, and His86. To identify the cutoff number of contacts for Phe119 of LC3, we did some image analysis using ImageJ [30, 58] on Figure 3.8a; we found the ratio of area under curve before and after cutoff to be 0.25. Using an in-house tcl script, we identified the cutoff that gives a closer ratio (0.28, within 15%) of area under curve before and after cutoff for Figure 3.8b. As shown in Figure 3.8b, this cutoff is 9.
Figure 3.8: (a) Histogram for the number of contacts of the side chain of Phe115 with hydrophobic pocket (HP) formed by Met1, Val36, Phe79, and aliphatic portion of Lys 82 in GATE-16. The cutoff value 8, is indicated by the dashed line, adapted from reference [45] (b) Histogram for the number of contacts of the side chain of Phe119 with hydrophobic pocket (HP) formed by Ser3, Tyr38, Leu82, and His86 in LC3. The cutoff value 9, is indicated by the dashed line.

Figure 3.9 shows the visual representation of identified criterion 1 for LC3 along with criterion 1 of GATE-16.
3.1.2 Criterion 2

For GATE-16 clustering analysis, criterion 2 is about the orientation of Phe115 when residing inside the hydrophobic pocket. Once inside the hydrophobic pocket, Phe115 can point in two different directions: towards the N-terminus or away from the N-terminus. Ma and coworkers distinguished these directions on the basis of minimum distance between the Cζ atom of Phe115 and any of the tryptophan 3 (Trp3) atoms [45]. They concluded Phe115 is oriented toward the N-terminus if this distance is \( \leq 7.2 \) Å based on Figure 3.11a. Similarly, to identify the orientation of Phe119 when residing inside the hydrophobic pocket, we need to select the corresponding residue in LC3 to that of Trp3 in GATE-16. For this, we aligned the structures of LC3, and GATE-16 using DeepAlign. As shown in Figure 3.10, Lys5 of
LC3 aligns with Trp3 of GATE-16 more than 61% of the time during the simulation.

![Figure 3.10: Aligned residues of LC3 compared to Trp3 of GATE-16](image)

To identify the cutoff distance for this criterion, we did some image analysis using ImageJ \[30, 58\] on Figure 3.11a; we found the ratio of area under curve before and after cutoff to be 1.65. Using an in-house tcl script, we identified the cutoff that gives a closer ratio (1.69, within 5%) of area under curve before and after cutoff for Figure 3.11b. As shown in Figure 3.11b, this cutoff is 15.29Å.

![Figure 3.11: (a) GATE-16: distance distribution between Cζ atom of Phe115 and any of Trp3 atoms. The cutoff value 7.2Å, is indicated by the dashed line, adapted from reference [45] (b) LC3: distance distribution between Cζ atom of Phe119 and any of Lys5 atoms. The cutoff value 15.29Å, is indicated by the dashed line.](image)

Figure 3.12 shows the visual representation of identified criterion 2 for LC3 along with criterion 2 of GATE-16.
CHAPTER 3. COMPARISON STUDY: LC3 VS. GATE-16 (IN SOLUTION)

Figure 3.12: (a) GATE-16 structure highlighting criterion 2: orientation of Phe115 toward the N-terminus. HP is shown in red surface representation, residues Phe115 (yellow) and Trp3 (blue) in CPK representations. (b) LC3 structure highlighting criterion 2: orientation of Phe119 toward the N-terminus. HP is shown in red surface representation, residues Phe119 (yellow) and Lys5 (blue) in CPK representations.

3.1.3 Criterion 3

For GATE-16 clustering analysis, criterion 3 is very similar to that of Criterion1, except the residue for hydrophobic pocket coverage is different, Phe117 instead of Phe115. To identify the corresponding residue for hydrophobic pocket coverage in LC3, we aligned the structures of LC3, and GATE-16 using DeepAlign. As shown in Figure 3.13, the only residue that aligns with Phe117 of GATE-16 is Gly120, although the probability is very low, 0.11% of the time during the simulation.
In GATE-16, Phe117 is considered to be in the hydrophobic pocket (formed by Met1, Val36, Phe79 and Lys82) if at least 8 of the 14 side chain atoms of Phe115 are in contact with these four hydrophobic pocket residues. A contact is defined on the basis of minimal distance between each of Phe117’s side chain atoms and any of the atoms belonging to the residues forming the hydrophobic pocket. If this minimal distance is \( \leq 4 \) Å, a contact is present. Here, multiple contacts between a particular side chain atom of Phe117 and the hydrophobic pocket were counted as one, meaning the maximum number of hydrophobic pocket contacts for Phe115 is 14 [45]. The cutoff number of contacts 8, is determined from the distribution plot shown in Figure 3.14a. For LC3, we use similar definition to determine if Gly120 is inside the hydrophobic pocket formed by Ser3, Tyr38, Leu82, and His86. To identify the cutoff number of contacts for Gly120 of LC3, we did some image analysis using ImageJ [30, 58] on Figure 3.14a; we found the ratio of area under curve before and after cutoff to be 6.31. Using an in-house tcl script, we identified the cutoff that gives a closer ratio (3.77, within 45%) of area under curve before and after cutoff for Figure 3.14b. As shown in Figure 3.14b, this cutoff is 6.
Figure 3.14: (a) Histogram for the number of contacts of the side chain of Phe117 with hydrophobic pocket (HP) formed by Met1, Val36, Phe79, and aliphatic portion of Lys 82 in GATE-16. The cutoff value 8, is indicated by the dashed line, adapted from reference [45] (b) Histogram for the number of contacts of the side chain of Gly120 with hydrophobic pocket (HP) formed by Ser3, Tyr38, Leu82, and His82 in LC3. The cutoff value 6, is indicated by the dashed line.

Figure 3.15 shows the visual representation of identified criterion 3 for LC3 along with criterion 3 of GATE-16.
For GATE-16 clustering analysis, criterion 4 is the salt bridge between N and C termini formed by Met1 and Phe117. Here is an interesting thing to keep in mind is that in general, salt bridges are formed if the distance between nitrogens in the basic side chain and oxygens in the acidic side chain is $\leq 4\text{Å}$ [8]. Both Met, Phe are neutral residues. So more than the actual definition, here the authors are trying to distinguish between open and closed conformations. To identify the corresponding residues for salt bridge between N and C termini in LC3, we aligned the structures of LC3, and GATE-16 using DeepAlign. As shown before, in Figures 3.2, 3.13, the corresponding residues for Met1, Phe117 of GATE-16 in LC3 are Ser3, and Gly120 respectively.

In GATE-16, the presence of salt bridge between N and C termini is evaluated by the
distance between N atom of Met1 and the closest of the two carboxylate oxygens of Phe117. If this distance is $\leq 4$ Å, a salt bridge is present. The cutoff distance of 4 Å, is determined from the distribution plot shown in Figure 3.16a. To identify the cutoff distance for the salt bridge of LC3, we did some image analysis using ImageJ [30, 58] on Figure 3.16a; we found the ratio of area under curve before and after cutoff to be 1.22. Using an in-house tcl script, we identified the cutoff that gives a closer ratio (1.22, within 5%) of area under curve before and after cutoff for Figure 3.16b. As shown in Figure 3.16b, this cutoff is 18.49 Å.

**Figure 3.16:** (a) Distance distribution between N atom of Met1 and the closest of the two carboxylate oxygens of Phe117 in GATE-16. The cutoff distance 4 Å, is indicated by the dashed line, adapted from reference [45] (b) Distance distribution between N atom of Ser3 and the closest of the two carboxylate oxygens of Gly120 in GATE-16 The cutoff distance 18.49 Å, is indicated by the dashed line.

Figure 3.17 shows the visual representation of identified criterion 4 for LC3 along with criterion 4 of GATE-16.
Figure 3.17: (a) GATE-16 structure highlighting criterion 4: salt bridge between N and C termini formed by Met1 and Phe117. HP is shown in red surface representation, Met1 (blue) and Phe117 (yellow) CPK representations. (b) LC3 structure highlighting criterion 4: salt bridge between N and C termini formed by Ser3 and Gly120. HP is shown in red surface representation, Ser3 (blue) and Gly120 (yellow) CPK representations.

Table 3.3 shows all the identified criteria for clustering method I, i.e. criteria for LC3 along with the criteria for GATE-16.

Table 3.3: Criteria for Clustering Analysis

<table>
<thead>
<tr>
<th>Criterion#</th>
<th>Criterion Details</th>
<th>GATE-16</th>
<th>LC3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Residue for HP coverage</td>
<td>Phe115</td>
<td>Phe119</td>
</tr>
<tr>
<td></td>
<td>Cutoff number of contacts</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>Orientation towards N-terminus: residues for min. distance</td>
<td>Phe115Cζ atom, Trp3</td>
<td>Phe119Cζ atom, Lys5</td>
</tr>
<tr>
<td></td>
<td>Cutoff distance</td>
<td>≤7.2 Å</td>
<td>≤15.29 Å</td>
</tr>
<tr>
<td>3</td>
<td>Residue for HP coverage</td>
<td>Phe117</td>
<td>Gly120</td>
</tr>
<tr>
<td></td>
<td>Cutoff number of contacts</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>Residues for salt bridge</td>
<td>Met1, Phe117</td>
<td>Ser3, Gly120</td>
</tr>
<tr>
<td></td>
<td>Cutoff distance</td>
<td>≤4 Å</td>
<td>≤18.49 Å</td>
</tr>
</tbody>
</table>
Before, we compare results from our LC3 simulations to GATE-16 results, another aspect to consider is the difference in sampling. For the GATE-16 study, the authors simulated 20 replicas with Hamiltonian scaling factors between 1.00 and 0.66. The exact scaling factors were 1.00, 0.98, 0.96, 0.94, 0.91, 0.89, 0.88, 0.86, 0.84, 0.82, 0.80, 0.78, 0.77, 0.75, 0.73, 0.72, 0.70, 0.69, 0.67, and 0.66, which correspond to temperatures between 298 K and 457 K [45]. In our case, we simulated 32 replicas with Hamiltonian scaling factors between 1.00 and 0.55. For LC3 simulations, the exact scaling factors were 1.00, 0.98, 0.96, 0.94, 0.92, 0.91, 0.89, 0.87, 0.86, 0.84, 0.82, 0.81, 0.79, 0.78, 0.76, 0.75, 0.73, 0.72, 0.70, 0.79, 0.68, 0.66, 0.65, 0.64, 0.63, 0.61, 0.60, 0.59, 0.58, 0.57, 0.56, and 0.55. These scaling factors correspond to temperatures between 300 K to 550 K. The exchange between neighboring replicas was attempted every 5 ps for both GATE-16 and LC3. For GATE-16, the total simulation time per replica was 500 ns, whereas for LC3 the total simulation time per replica was 390 ns. Also, the starting structure for HREMD simulations of GATE-16 was a structure that was equilibrated for 5 ns. However, for LC3 simulations we first equilibrated the structure using regular MD simulations for 200 ns and then it was used as the starting structure for HREMD simulations. Another difference comes from the force-field, GATE-16 study used Amber99SB-ILDN force field, and we used CHARMM27 forcefield. Both these forcefields are good for protein simulations [9, 43]. These differences in sampling might cause some observed differences in our results. For further details about our LC3 simulations, please refer to Appendix B.

3.2 Results and Discussion

We clustered LC3 structures into 12 different states using the identified criteria shown in Table 3.3. Then these LC3 states were compared to the 12 GATE-16 states identified by Ma and coworkers as shown in Table 3.4. In this table, the states are presented in descending order, from the most populated state to the least populated state in the case of GATE-16 in solution.

As shown in Table 3.4, five out of the six most populated states for GATE-16 are similar to
Table 3.4: Comparison of states: GATE-16 vs. LC3

<table>
<thead>
<tr>
<th>GATE-16 state</th>
<th>GATE-16 population</th>
<th>LC3 state</th>
<th>LC3 population</th>
<th>Criterion1: Phe119/Phe115 contacts HP</th>
<th>Criterion2: Orientation of Phe119/115 towards N-terminus</th>
<th>Criterion3: Gly120/Phe117 contacts HP</th>
<th>Criterion4: Ser3-Gly120/Met1-Phe117 salt bridge</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>41.2%</td>
<td>1</td>
<td>20.4%</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>2</td>
<td>15.7%</td>
<td>2</td>
<td>16.7%</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>3</td>
<td>14.7%</td>
<td>5</td>
<td>9.7%</td>
<td>no</td>
<td>n/a</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>4</td>
<td>8.4%</td>
<td>4</td>
<td>14.0%</td>
<td>yes</td>
<td>n/a</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>5</td>
<td>4.3%</td>
<td>10</td>
<td>1.8%</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>6</td>
<td>4.1%</td>
<td>6</td>
<td>7.5%</td>
<td>yes</td>
<td>n/a</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>7</td>
<td>3.5%</td>
<td>11</td>
<td>1.5%</td>
<td>no</td>
<td>n/a</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>8</td>
<td>2.5%</td>
<td>7</td>
<td>5.6%</td>
<td>yes</td>
<td>n/a</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>9</td>
<td>2.3%</td>
<td>12</td>
<td>0.5%</td>
<td>no</td>
<td>n/a</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>10</td>
<td>1.5%</td>
<td>9</td>
<td>2.5%</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>11</td>
<td>1.1%</td>
<td>3</td>
<td>15.6%</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>12</td>
<td>0.8%</td>
<td>8</td>
<td>4.3%</td>
<td>no</td>
<td>n/a</td>
<td>no</td>
<td>yes</td>
</tr>
</tbody>
</table>

that of LC3: the states 1, 2, 4 and 6 matches for GATE-16, and LC3. State 3 of GATE-16 is state 5 for LC3. The difference comes from the following: GATE-16 state 5 or 5th most populated state in GATE-16 ‘yes, no, yes, no’ is state 10 for LC3. Also, state 11 for GATE-16 is LC3 state 3, which is ‘yes, yes, yes, yes’. So for GATE-16, the preferred state is ‘yes, no, yes, no’: the one in which Phe115 contacts the hydrophobic pocket formed by Met1, Val36, Phe79 and Lys82. Here, once inside the hydrophobic pocket, Phe119 point away from the N-terminus. In this state, Phe117 contacts the hydrophobic pocket, and the salt bridge between N and C termini doesn’t exist. In the case of LC3, the preferred state is ‘yes, yes, yes, yes’: the one in which Phe119 contacts the hydrophobic pocket formed by Ser3, Tyr38, Leu82 and His86. Here, once inside the hydrophobic pocket, Phe119 point towards the N-terminus. In this state, Gly120 contacts the hydrophobic pocket, and the salt bridge
between N and C termini exists. Figure 3.18 shows the GATE-16 and LC3 preferred states.

Figure 3.18: (a) GATE-16 preferred state ‘yes no yes no’, along with criteria: HP is shown in red surface representation, Met1 (blue, part of HP), Trp3 (blue), Phe115 (yellow) and Phe117 (yellow) CPK representations. (b) LC3 preferred state ‘yes yes yes yes’, along with criteria: HP is shown in red surface representation, Ser3 (blue, part of HP), Lys5 (blue), Phe119 (yellow) and Gly120 (yellow) CPK representations.

By visually inspecting the trajectory and by referring to GATE-16 study [45], we know that having ‘yes’ for any of the criterion makes the structure more compact. Since LC3 prefers the state ‘yes, yes, yes, yes’ and GATE-16 prefers the state ‘yes, no, yes, no’, our results indicate LC3 prefers conformations that are more closed compared to GATE-16.

In Table 3.4, if we consider the top 4 most populated states of GATE-16, they constitute 80% (41.2%, 15.7%, 14.7%, 8.4%) of all the available conformations. However, for LC3 the top 4 most populated states make up 66.7% (20.4%, 16.7%, 15.6%, 14.0%) of the possible conformations. This trend indicates that LC3 in solution explores more conformations compared to GATE-16 in solution.

To understand these trends better, let us take a closer look at the GATE-16 study. Ma and co-workers developed the four criteria to study the conformational dynamics of C-terminal region of GATE-16 protein [45]. Although these criteria was developed to identify C-terminal dynamics, they are influenced by residues in N-terminal. Criteria 1, and 3 focus
entirely on C-terminal residues, whereas criteria 2, and 4 are influenced by N-terminal dynamics. For example consider criterion 2 for GATE-16: here, the orientation of a C-terminal residue, Phe115 when residing inside the hydrophobic pocket was determined using the minimum distance between the Cζ atom of Phe115 and any of the atoms of Trp3, an N-terminal residue. Also, criterion 4 identifies a salt bridge between N and C termini, specifically between residues Met1 and Phe117. We developed similar criteria for LC3 primarily using DeeapAlign [71], a structural alignment tool for proteins and other analysis as explained in detail earlier in this chapter. So for LC3, we looked at the orientation of Phe119, when residing inside the hydrophobic pocket towards the N-terminus by estimating the minimum distance between the Cζ atom of Phe119 and any of the atoms of Lys5, an N-terminal residue. Criterion 4 identifies a salt bridge between N and C termini formed by residues Met1 and Phe117.

Many studies suggest C-terminal region is conserved among all the Atg8 families, it is the N-terminus region that vary between families [45, 59, 69]. For instance, the N-terminal subdomain consists of two α helices; the first α-helix in GATE-16 is acidic whereas the first α-helix in LC3 is basic. In GATE-16, the second α-helix is neutral and in LC3 the α-helix is acidic. The N-terminal subdomain along with α-helices and the C-terminal subdomain is depicted clearly in Figure 3.19.
CHAPTER 3. COMPARISON STUDY: LC3 VS. GATE-16 (IN SOLUTION)

(a) GATE-16

(b) LC3

Figure 3.19: (a) GATE-16 and (b) LC3 structures highlighting N-terminal subdomain in green and C-terminal subdomain in blue. N-terminal subdomain consists of two $\alpha$ helices

Our results can be explained better considering the fact N-terminal domain varies among Atg8 families. As discussed before, our results show five out of the six most populated states in LC3 are similar to that of GATE-16. The one state that differs is shown explicitly in Table 3.5. It is clear these states differ in criteria 2, 4 which are influenced by N-terminal residues. Again, since N-terminal region differs among Atg8 families, it is expected to behave differently.

<table>
<thead>
<tr>
<th>Molecule</th>
<th>Criterion1: Phe115/119 contacts HP</th>
<th>Criterion2: Orientation of Phe115/119 towards N-terminus</th>
<th>Criterion3: Phe117/Gly120 contacts HP</th>
<th>Criterion4: Met1-Phe117/Ser3-Gly120 salt bridge</th>
</tr>
</thead>
<tbody>
<tr>
<td>GATE-16</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>LC3</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
</tbody>
</table>

Table 3.5: Preferred States
3.3 Conclusions

The clustering analysis shows LC3 conformations are similar to GATE-16 but they differ in motion of the N-terminal region. This is supported by studies that show that N-terminal region of Atg8 proteins varies between families. We found that LC3 prefers compact structures compared to GATE-16. This is evident from the preferred states: 1) For GATE-16 ‘yes, no, yes, no’- this is the state where both residues Phe115, and Phe117 contacts the hydrophobic pocket formed by Met1, Val36, Phe79 and aliphatic portion of Lys82. Once inside the hydrophobic pocket Phe115 points away from the N-terminus, and there is no salt bridge between Met1 and Phe117. 2) For LC3 ‘yes, yes, yes, yes’- the state in which both residues Phe119, and Gly120 contacts the hydrophobic pocket formed by Ser3, Tyr38, Leu82 and His86. Once inside the hydrophobic pocket Phe119 orient towards from the N-terminus, and there is a salt bridge between N and C-terminal subdomains, specifically between Ser3 and Gly120. Finally, observed differences might be due to differences in sampling: total simulation time of 500ns for GATE-16 vs. total simulation time of 390 ns for LC3.
Chapter 4

LC3 in solution vs. membrane bound LC3

In this chapter, our goal is to compare results for the HREMD simulations of LC3 in solution (LC3 or LC3-I) and membrane bound LC3 (LC3_mem or LC3-II). This comparison study is done using two methods: 1) Clustering Method I and 2) Clustering Method II. Here, ‘Clustering Method I’ refers to the same approach we used to compare LC3 and GATE-16 in chapter 3 of this dissertation, and ‘Clustering Method II’ is another approach we developed, which will be discussed in detail later on this chapter.

4.1 Clustering Method I

Here we are using the same criteria identified for LC3 in solution as mentioned in chapter 3. For example, criterion1 is about the hydrophobic pocket coverage by residue Phe119. This residue is considered to be inside the hydrophobic pocket formed by Ser3, Tyr38, Leu82, and His86 if at least 9 of the 14 side chain atoms of Phe119 are in contact with these four hydrophobic pocket residues. As mentioned in Chapter 3, we are still calling the pocket a ‘hydrophobic pocket’ although only two out of the four residues (Tyr, Leu) are hydrophobic. Criterion2 is about the orientation of residue Phe119 once inside the hydrophobic pocket.
Phe119 is oriented towards N-terminus if the distance between Phe119CZ atom and Lys5 is $\leq 15.29$ Å. Criterion 3 is about the hydrophobic pocket coverage by another residue, Gly120, which is considered to be inside the hydrophobic pocket, if at least 6 of the 8 side chain atoms of Gly120 are in contact with the four hydrophobic pocket residues. Finally, criterion 4 is about the existence of salt bridge between N and C terminal regions, specifically between residues Gly120 and Ser3. This salt bridge exists if the distance between Gly120 and Ser3 is $\leq 18.49$ Å. The criteria for clustering method I are shown in the Table 4.1 and through Figures 4.1, 4.2, 4.3, and 4.4. For additional details regarding these criteria, please refer to chapter 3.

**Table 4.1:** Criteria for Clustering Method I

<table>
<thead>
<tr>
<th>Criterion#</th>
<th>Criterion Details</th>
<th>LC3</th>
<th>LC3_mem</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Residue for HP coverage: Phe119</td>
<td>Phe119</td>
<td>Phe119</td>
</tr>
<tr>
<td></td>
<td>Cutoff number of contacts: 9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>Orientation towards N-terminus: residues for min. distance: Phe119CZ atom, Lys5</td>
<td>$\leq 15.29$ Å</td>
<td>$\leq 15.29$ Å</td>
</tr>
<tr>
<td></td>
<td>Cutoff distance</td>
<td>$\leq 15.29$ Å</td>
<td>$\leq 15.29$ Å</td>
</tr>
<tr>
<td>3</td>
<td>Residue for HP coverage: Gly120</td>
<td>Gly120</td>
<td>Gly120</td>
</tr>
<tr>
<td></td>
<td>Cutoff number of contacts: 6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>Residues for salt bridge: Ser3, Gly120</td>
<td>Ser3, Gly120</td>
<td>Ser3, Gly120</td>
</tr>
<tr>
<td></td>
<td>Cutoff distance</td>
<td>$\leq 18.49$ Å</td>
<td>$\leq 18.49$ Å</td>
</tr>
</tbody>
</table>
Figure 4.1: (a) LC3 and (b) LC3_mem structures highlighting criterion 1: coverage of hydrophobic pocket (HP) by residue Phe119. HP shown in red surface representation, is formed by residues Ser3, Tyr38, Leu82, and His86. Phe119 is shown in yellow CPK representation.

Figure 4.2: (a) LC3 and (b) LC3_mem structures highlighting criterion 2: orientation of Phe119 toward the N-terminus. HP is shown in red surface representation, residues Phe119 (yellow) and Lys5 (blue) in CPK representations.
Figure 4.3: (a) LC3 and (b) LC3_mem structures highlighting criterion 3: coverage of hydrophobic pocket (HP) by residue Gly120. HP shown in red surface representation is formed by residues Ser3, Tyr38, Leu82, and His86. Gly120 is shown in yellow CPK representation.

Figure 4.4: (a) LC3 and (b) LC3_mem structures highlighting criterion 4: salt bridge between N and C termini formed by residues Ser3 and Gly120. HP is shown in red surface, Ser3 in blue CPK and Gly120 in yellow CPK representations.

Before we move into results, one thing to keep in mind is the difference in sampling, specifically the total simulation time. For LC3, the total simulation time was 390 ns and for
LC3\textsubscript{mem}, the total simulation time was 215 ns. For further details about our simulations, please refer to Appendix B.

### 4.1.1 Results and Discussion

With the criteria shown in Table 4.1, we clustered LC3\textsubscript{mem} (membrane associated form of LC3) structures into 12 distinct states as shown in Table 4.2. The states are presented in descending order, from the most populated state to the least populated LC3 state. This table also compares the LC3\textsubscript{mem} states to corresponding states in LC3.

<table>
<thead>
<tr>
<th>LC3 state</th>
<th>LC3 population</th>
<th>LC3\textsubscript{mem} state</th>
<th>LC3\textsubscript{mem} population</th>
<th>Criterion1: Phe119 contacts HP</th>
<th>Criterion2: Orientation of Phe119 towards N-terminus</th>
<th>Criterion3: Gly120 contacts HP</th>
<th>Criterion4: Ser3-Gly120 salt bridge</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20.4%</td>
<td>1</td>
<td>35.8%</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>2</td>
<td>16.7%</td>
<td>2</td>
<td>21.6%</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>3</td>
<td>15.6%</td>
<td>7</td>
<td>1.3%</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>4</td>
<td>14.0%</td>
<td>3</td>
<td>16.7%</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>5</td>
<td>9.7%</td>
<td>4</td>
<td>10.9%</td>
<td>no</td>
<td>n/a</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>6</td>
<td>7.5%</td>
<td>5</td>
<td>5.8%</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>7</td>
<td>5.6%</td>
<td>10</td>
<td>0.7%</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>8</td>
<td>4.3%</td>
<td>6</td>
<td>4.6%</td>
<td>no</td>
<td>n/a</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>9</td>
<td>2.5%</td>
<td>9</td>
<td>0.9%</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>10</td>
<td>1.8%</td>
<td>8</td>
<td>1.1%</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>11</td>
<td>1.5%</td>
<td>11</td>
<td>0.3%</td>
<td>no</td>
<td>n/a</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>12</td>
<td>0.5%</td>
<td>12</td>
<td>0.2%</td>
<td>no</td>
<td>n/a</td>
<td>yes</td>
<td>no</td>
</tr>
</tbody>
</table>

From the above table, we see that the top four most populated states in the descending order for LC3 are ‘yes, yes, no, yes’, ‘yes, yes, no, no’, ‘yes, yes, yes, yes’, and ‘yes, no,
no, no’. In the case of membrane bound LC3 or LC3_mem, the top four most populated states are ‘yes, yes, no, yes’, ‘yes, yes, no, no’, ‘yes, no, no, no’, and ‘no, n/a, no, no’. To understand this better, let us look into more details related to the criteria. As mentioned in chapter 3, whenever a criterion is satisfied, meaning we have ‘yes’in the above table, it makes the structure more compact. So, a ‘no’ results in more open structures. It is very evident by visually inspecting conformations of all the 12 states. Now, if we consider the four most populated states of LC3 or LC3_mem, we have a total of 16 possibilities for all the four criteria. Out of these 16 options, there are 10 ‘yes’, and 6 ‘no’ for the four most populated states of LC3 (constitutes 66.7% of structures). Similarly for LC3_mem, among the most populated 4 states (constitutes 85% of structures), we have 6 ‘yes’, 9 ‘no’ and 1 ‘n/a’. Clearly, LC3 has more ‘yes’ and so LC3 prefers more compact structures compared to LC3_mem.

As mentioned in Chapter 3, Ma and co-workers developed the four criteria to study the conformational dynamics of C-terminal of GATE-16 protein. We developed similar criteria for LC3 primarily using DeeapAlign [71], a structural alignment tool for proteins and other analysis as explained in detail in Chapter 3. Although, the criteria was developed to identify C-terminal dynamics, some of them are influenced by residues in the N-terminal subdomain. For example, the criterion 2 of LC3 is about the orientation of a C-terminal residue, Phe119 when residing inside the hydrophobic pocket. We determined if Phe119 is oriented towards the N-terminus or not by estimating the minimum distance between the Cζ atom of Ph119 and any of the atoms of Lys5, an N-terminal residue. Also, criterion 4 identifies a salt bridge between N and C terminal subdomains, formed by residues Ser3 and Gly120.

Now, let us see if the criteria influenced by N-terminal residues namely criteria 2, and 4 explicitly supports our hypothesis. To refresh your memory, following is our hypothesis: upon binding to the lipid membrane LC3 undergoes a conformational transition, which allows it to associate with another membrane-bound LC3 in a trans fashion. Let us focus again on the four most populated structures of LC3_mem and LC3 shown in Table 4.2. Criterion 2 shows the following trend: for LC3, 3 out of 4 or 75% of structures have ‘yes’ or satisfies the criterion and for LC3_mem, 2 out of 4 or 50% of structures have ‘yes’ or satisfies the criterion.
Now, for criterion 4, we observe the following trend: for LC3, 2 out of 4 or 50% of structures have ‘yes’ or satisfies the criterion and for LC3_mem, 1 out of 4 or 25% of structures have ‘yes’ or satisfies the criterion. Here, both criteria are satisfied by a higher percent of structures in LC3 compared to LC3_mem, and so LC3 prefers closed structures compared to LC3_mem. To a certain extend, this indicates the N-terminal region of LC3_mem has more open conformations compared to LC3, and thus supports our hypothesis, that N-terminal region of membrane associated LC3 will adopt more open conformations as it interacts with another membrane associated LC3 via its N-terminal region.

So far, we have been considering only the four most populated states of LC3 and its membrane associated form. If we consider, the six most populated states of LC3 and LC3_mem, we notice five out of these six states are similar: states 1, and 2 matches for LC3, and LC3_mem. States 4, 5 and 6 of LC3 are states 3, 4, and 5 respectively for LC3_mem. The difference comes from the following: The third most populated state in LC3 or simply LC3 state 3 ‘yes, yes, yes, yes’ is the 7th most populated state or state 7 for LC3_mem. Also, LC3 state 8 ‘no, n/a, no, yes’ is state 6 for LC3_mem. Thus, LC3 preferred state is ‘yes, yes, yes, yes’: the one in which Phe119 contacts the hydrophobic pocket formed by Ser3, Tyr38, Leu82 and His86. Here, once inside the hydrophobic pocket, Phe119 point towards the N-terminus. In this state, Gly120 contacts the hydrophobic pocket, and the salt bridge between N and C termini or Ser3-Gly120 salt bridge exists. For LC3_mem, the preferred state is ‘no, n/a, no, yes’: the one in which residues Phe119 and Gly120 doesn’t contact the hydrophobic pocket formed by Ser3, Tyr38, Leu82 and His86. In this state, there is the presence of salt bridge between N and C termini. Once again, ‘yes’ for a criterion makes the structure more compact, so LC3 prefers a more closed conformation compared to LC3_mem. Figure 4.5 shows the preferred states for LC3 and LC3_mem.
CHAPTER 4. LC3 IN SOLUTION VS. MEMBRANE BOUND LC3

Figure 4.5: (a) LC3 preferred state ‘yes, yes, yes, yes’ (b) LC3_mem preferred state ‘no, n/a, no, yes’, along with the four criteria: HP formed by Ser3, Tyr38, Leu82, and His86 is shown in red surface representation. Residues covering HP, Phe119, and Gly120 are shown in yellow CPK representations. Lys5 (for criterion 2) and Ser3 (for criterion 4) are shown in blue CPK representations.

Another interesting trend, we can see in table 4.2 is the following: the four most populated states of LC3 ‘yes, yes, no, yes’, ‘yes, yes, no, no’, ‘yes, yes, yes, yes’ and ‘yes, no, no, no’ constitute 66.7% of all structures. However, for LC3_mem, the four most populated states, ‘yes, yes, no, yes’, ‘yes, yes, no, no’, ‘yes, yes, no, no’ and ‘no, n/a, no, no’ constitute 85% of all structures. This shows that LC3 in solution explore much more conformations compared to LC3_mem.

4.2 Clustering Method II

Clustering Method II refers to a method we created that can explicitly identify the conformational dynamics of N-terminal region of LC3. Using this method, we identified the conformational preferences for LC3 and LC3_mem. This approach is similar to ‘clustering method I’ in the sense we are identifying four criteria that can capture the N-terminal dynamics.
4.2.1 Criterion1

We identified a hydrophobic pocket that can capture the motion of N-terminal region. This hydrophobic pocket is formed by residues phenylalanine 7 (Phe7), phenylalanine 108 (Phe108) and tyrosine 110 (Tyr110). Criterion1 is about the presence of a residue isoleucine34 (Ile34) in this hydrophobic pocket (HP) as shown in Figure 4.6. Ile34 is considered to be in the hydrophobic pocket if at least 11 of the 13 side chain atoms of Ile34 are in contact with the four HP residues. A contact is defined on the basis of minimal distance between each of Ile34’s side chain atoms and any of the atoms belonging to the residues forming the hydrophobic pocket. If this minimal distance is $\leq$4 Å, a contact is present. Here, multiple contacts between a particular side chain atom of Ile34 and the hydrophobic pocket were counted as one, meaning the maximum number of hydrophobic pocket contacts for Ile34 is 13. The cutoff number of contacts 11, is determined from the distribution plot shown in Figure 4.7.

![Figure 4.6: a) LC3 and (b) LC3_mem structures highlighting criterion1: coverage of hydrophobic pocket (HP) formed by Phe7, Phe108, and Tyr110 by Ile34. HP is shown in red surface and Ile34 in yellow CPK representations.](image)
4.2.2 Criterion 2

Criterion 2 is a salt bridge between N and C termini formed among residues aspartic acid 19 (Asp19) and Lysine 51 (Lys51) as shown in Figure 4.8. The presence of this salt bridge between N and C termini is evaluated by the distance between nitrogen atom of Lys51 and the closest of the two carboxylate oxygens of Asp19. If this distance is \( \leq 3.6 \, \text{Å} \), a salt bridge is present. The cutoff distance of 3.6 Å, is determined from the distribution plot shown in Figure 4.9. Strictly speaking, salt bridges are formed if the distance between nitrogens in the basic side chain and oxygens in the acidic side chain is \( \leq 4 \, \text{Å} \). We do satisfy the definition of salt bridge in the case of residues; Asp is an acidic amino acid and Lys is a basic amino acid. However, we are not strictly following the cutoff distance of 4Å, because our intention here is to use criteria that can distinguish between open and closed conformations of LC3 and its membrane associated form.
CHAPTER 4. LC3 IN SOLUTION VS. MEMBRANE BOUND LC3

72

(a) LC3
(b) LC3_mem

Figure 4.8: a) LC3 and (b) LC3_mem structures highlighting criterion 2: salt bridge between N and C termini, formed among residues Asp19 and Lys51. HP is shown in red surface, Asp19 in red CPK, and Lys51 in blue CPK representations.

4.2.3 Criterion 3

Criterion 3 is another salt bridge between N and C termini formed with residues glutamic acid 105 (Glu105) and arginine 16 (Arg16) as shown in Figure 4.10. The presence of this salt bridge between N and C termini is evaluated by the distance between closest of the three nitrogen atoms of Arg16 and the closest of the two carboxylate oxygens of Glu105. If

Figure 4.9: Histograms for the Asp19-Lys51 salt bridge distance for (a) LC3 and in (b) LC3_mem. The cutoff value 3.6Å, is indicated by the dashed line.
this distance is $\leq 6.1$ Å, a salt bridge is present. The cutoff distance of 6.1 Å, is determined from the distribution plot shown in Figure 4.11. Strictly speaking, salt bridges are formed if the distance between nitrogens in the basic side chain and oxygens in the acidic side chain is $\leq 4$ Å. We do satisfy the definition of salt bridge in the case of residues; Glu is an acidic amino acid and Arg is a basic amino acid. Similar to criterion 2, we are not strictly following the cutoff distance of 4 Å, as our intention is to use a criteria that can distinguish between open and closed conformations of LC3 and its membrane associated form.

![Figure 4.10: a) LC3 and (b) LC3_mem structures highlighting criterion 3: salt bridge between N and C termini, formed between Glu105 and Arg16. HP is shown in red surface, Glu105 in red CPK and Arg16 in blue CPK representations.](image)
Figure 4.11: Histogram for the Glu105-Arg16 salt bridge distance (a) LC3 and in (b) LC3_mem. The cutoff value 6.1 Å, is indicated by the dashed line.

4.2.4 Criterion 4

Criterion 4 is a salt bridge between N and C termini formed between residues aspartic acid 106 (Asp106) and arginine 11 (Arg11) as shown in Figure 4.12. The presence of this salt bridge between N and C termini is evaluated by the distance between closest of the three nitrogen atoms of Arg16 and the closest of the two carboxylate oxygens of Asp106. If this distance is ≤ 9.5 Å, a salt bridge is present. The cutoff distance of 9.5 Å, is determined from the distribution plot shown in Figure 4.13. Strictly speaking, salt bridges are formed if the distance between nitrogens in the basic side chain and oxygens in the acidic side chain is ≤ 4 Å. Similar to criteria 2, and 3, we do satisfy the definition of salt bridge in the case of residues; Asp is an acidic amino acid and Arg is a basic amino acid. Similar to criterion 2, and 3, here we are not strictly following the cutoff distance of 4 Å, as our goal is to identify criteria that can distinguish between open and closed conformations of LC3 and its membrane associated form.
Figure 4.12: a) LC3 and (b) LC3_mem structures highlighting criterion 4: salt bridge between N and C termini, formed between Asp106 and Arg11. HP is shown in red surface, Asp106 in red CPK and Arg11 in blue CPK representations.

Figure 4.13: Histogram for the Asp106-Arg11 salt bridge distance (a) LC3 and in (b) LC3_mem. The cutoff value 9.5 Å, is indicated by the dashed line.

Table 4.3 shows all the identified criteria for clustering method II.
Table 4.3: Criteria for Clustering Method II

<table>
<thead>
<tr>
<th>Criterion#</th>
<th>Criterion Details</th>
<th>LC3</th>
<th>LC3_mem</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Residue for HP coverage</td>
<td>Ile34</td>
<td>Ile34</td>
</tr>
<tr>
<td></td>
<td>Cutoff number of contacts</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>2</td>
<td>Residues for salt bridge</td>
<td>Asp19, Lys51</td>
<td>Asp19, Lys51</td>
</tr>
<tr>
<td></td>
<td>Cutoff distance</td>
<td>≤3.6 Å</td>
<td>≤3.6 Å</td>
</tr>
<tr>
<td>3</td>
<td>Residues for salt bridge</td>
<td>Glu105, Arg16</td>
<td>Glu105, Arg16</td>
</tr>
<tr>
<td></td>
<td>Cutoff distance</td>
<td>≤6.1 Å</td>
<td>≤6.1 Å</td>
</tr>
<tr>
<td>4</td>
<td>Residues for salt bridge</td>
<td>Asp106, Arg11</td>
<td>Asp106, Arg11</td>
</tr>
<tr>
<td></td>
<td>Cutoff distance</td>
<td>≤9.5 Å</td>
<td>≤9.5 Å</td>
</tr>
</tbody>
</table>

4.2.5 Results and Discussion

Using the criteria shown in Table 4.3, we grouped LC3 protein and its membrane associated form into 16 different states as shown in Table 4.4. Here, the states are shown in the descending order, from the most populated state to the least populated state for LC3 in solution. Table 4.4 also compares the states for membrane associated LC3 or LC3_mem to corresponding states for LC3 in solution.

From Table 4.4, we see that the top four most populated states in the descending order for LC3 are ‘yes, no, yes, yes’, ‘yes, no, no, no’, ‘yes, yes, no, yes’, and ‘yes, yes, yes, yes’. In the case of membrane bound LC3 or LC3_mem, the top four most populated states are ‘yes, no, no, no’, ‘yes, yes, no, no’, ‘yes, no, yes, no’, and ‘yes, yes, yes, no’. To understand this better, let us look into more details related to the criteria. As mentioned in chapter 3, whenever a criterion is satisfied, meaning we have ‘yes’in the above table, it makes the structure more compact. So, a ‘no’results in more open structures. It is very evident by visually inspecting conformations of all the 16 states. Now, if we consider the four most populated states of LC3 or LC3_mem, we have a total of 16 possibilities for all the four criteria. Among the four most populated states of LC3 (constitutes 50.79% of structures), there are 11 ‘yes’, and 5 ‘no’. In the case of LC3_mem, for the most populated four states (constitutes 85% of structures), we have 8 ‘yes’, and 8 ‘no’. Since LC3 has more ‘yes’for the criteria, LC3 prefers more compact structures compared to LC3_mem.
## Table 4.4: Comparison of states - LC3 in solution (LC3) vs. membrane bound LC3 (LC3\text{\_mem})

<table>
<thead>
<tr>
<th>LC3 state</th>
<th>LC3 population</th>
<th>LC3\text{_mem} state</th>
<th>LC3\text{_mem} population</th>
<th>Criterion1: Ile34 contacts HP</th>
<th>Criterion2: Asp19-Lys51 salt bridge</th>
<th>Criterion3: Glu105-Arg16 salt bridge</th>
<th>Criterion4: Asp106-Arg11 salt bridge</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15.59%</td>
<td>13</td>
<td>0.06%</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>2</td>
<td>13.72%</td>
<td>1</td>
<td>47.77%</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>3</td>
<td>11.88%</td>
<td>6</td>
<td>3.27%</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>4</td>
<td>9.60%</td>
<td>9</td>
<td>0.40%</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>5</td>
<td>9.33%</td>
<td>16</td>
<td>0.004%</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>6</td>
<td>9.05%</td>
<td>2</td>
<td>33.21%</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>7</td>
<td>6.55%</td>
<td>14</td>
<td>0.02%</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>8</td>
<td>4.91%</td>
<td>8</td>
<td>0.41%</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>9</td>
<td>3.64%</td>
<td>12</td>
<td>0.07%</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>10</td>
<td>3.43%</td>
<td>3</td>
<td>4.48%</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>11</td>
<td>3.24%</td>
<td>5</td>
<td>3.66%</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>12</td>
<td>2.72%</td>
<td>7</td>
<td>2.29%</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>13</td>
<td>2.52%</td>
<td>15</td>
<td>0.01%</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>14</td>
<td>2.20%</td>
<td>4</td>
<td>3.88%</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>15</td>
<td>1.00%</td>
<td>10</td>
<td>0.29%</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>16</td>
<td>0.59%</td>
<td>11</td>
<td>0.20%</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
</tbody>
</table>

Another major trend we notice in Table 4.4 is that the top six states are different for LC3 and LC3\text{\_mem}. For example, let us take a closer look at the six most populated states of LC3 and LC3\text{\_mem}: no states match for LC3, and LC3\text{\_mem}. The most populated state of LC3 ‘yes, no, yes, yes’ is state 13 for LC3\text{\_mem}. LC3 state 2, ‘yes, no, no, no’ is state 1 for LC3\text{\_mem}. State 3 of LC3, ‘yes, yes, no, yes’ is state 6 for LC3\text{\_mem}. The fourth most populated LC3 state, ‘yes, yes, yes, yes’ is state 9 for LC3\text{\_mem}. The state ‘no, no, yes, yes’ is ranked 5 for LC3 and for LC3\text{\_mem}, this is the least populated state. LC3 state 6, ‘yes, yes, no, no’ corresponds to state 2 in LC3\text{\_mem}. This is substantially different from...
the trend we noticed for clustering method I results earlier in this chapter, where five out
of top 6 states were similar between LC3, and LC3\_mem. Again, clustering method I was
developed to capture the C-terminal dynamics though it has influence from N-terminal
residues. However, clustering method II captures the N-terminal dynamics. Therefore, this
trend suggests N-terminal regions of LC3, LC3\_mem adopts different conformations. To
understand if N-terminal region adopts closed vs. open conformations, we can count the
number of ‘yes’, and ‘no’ for the top 6 states. In LC3, there are 15 ‘yes’ and 9 ‘no’ and for
LC3\_mem there are 11 ‘yes’ and 13 ‘no’. As ‘yes’ for a criterion makes the structure more
compact, LC3 prefers more compact conformations compared to LC3\_mem.

In order to identify the preference of LC3, and LC3\_mem for different conformations, we
can take a closer look at the visual representations of the 4 most populated states in LC3,
and LC3\_mem. Figure 4.14 shows the most populated states in LC3 and LC3\_mem.

![Figure 4.14: Most populated states for (a) LC3 - 'yes, no, yes, yes' and in (b) LC3\_mem -
'yes, no, no, no'. The N-terminal subdomain is colored in green and C-terminal subdomain
is colored in blue.](image)

The most populated LC3 state is ‘yes, no, yes, yes’, the one in which Ile34 contacts the
hydrophobic pocket formed by residues Phe7, Phe108, and Tyr110. The remaining three
criteria involves the presence of salt bridges between N and C-terminal regions. In this
state, Asp19-Lys51 saltbridge doesn’t exist. However, Glu105-Arg16 and Asp106-Arg11 saltbridges does exist. For LC3_mem the most populated state is ‘yes, no, no, no’, the one in which Ile34 contacts the hydrophobic pocket formed by residues Phe7, Phe108, and Tyr110. In this state, none of the salt bridges Asp19-Lys51, Glu105-Arg16 and Asp106-Arg11 exist. Figure 4.14 clearly shows that the N-terminal subdomain adopts a more open conformation compared to LC3. Once again, this is evident from the state labels, the LC3 state has 3 ‘yes’, 1 ‘no’and LC3_mem has a single ‘yes’and 3 ‘no’. As ‘yes’for a criterion makes teh structure more compact, this means LC3 state is a more closed compared to LC3_mem. To further understand the motion of N-terminal region, we can calculate root mean square deviation (RMSD) and see how much these states particularly the N-terminal subdomain deviated from the starting conformations. For details about RMSD, please refer to chapter 2 of this dissertation. The RMSD values for the most populated states in LC3, and LC3_mem are 2.58Åand 5.37Årespectively. Here, we are calculating the RMSD specifically for the N-terminal region. The higher RMSD value for LC3_mem compared to LC3, shows that N-terminal region is more mobile for LC3_mem.

Figure 4.15 shows the second most populated states in LC3, and LC3_mem. As denoted, the second most populated state for LC3 is ‘yes, no, no, no’, the one in which Ile34 contacts the hydrophobic pocket and none of the salt bridges Asp19-Lys51, Glu105-Arg16 and Asp106-Arg11 exist. But for LC3_mem, the second most populated state is ‘yes, yes, no, no’, the one in which Ile34 contacts the hydrophobic pocket, and Asp19-Lys51 salt bridge exists. However, the other two salt bridges (Glu105-Arg16 and Asp106-Arg11 ) between N and C-terminal regions doesn’t exist. Now referring back to our conclusion, having ‘yes’makes the structure more compact, here we have 1 ‘yes’, 3 ‘no’for LC3 and 2 ‘yes’, 2 ‘no’for LC3_mem. Although this makes the overall structure of LC3_mem more compact than LC3, there is an interesting trend we see in Figure 4.15. The first α helix of LC3_mem is more open compared to LC3. The first α helix consists of residues 1 to 11 and it is the short helix in the green colored N-terminal subdomain. Additionally, LC3 state 2 represents a smaller population compared to state 2 of LC3_mem. For example, state 2 of LC3 represents 13.72% of all structures whereas LC3_mem state 2 constitute 33.21% of all structures. Here is what we observe by comparing RMSD values of the N-terminal regions: for LC3 the RMSD is
2.01Å and for LC3\textsubscript{mem} it is 4.01Å. Again, higher RMSD value in the case of LC3\textsubscript{mem} shows that N-terminal region is more mobile for LC3\textsubscript{mem} than LC3.

![Image](image1.png)

Figure 4.15: The second most populated states for (a) LC3 - ‘yes, no, no, no’ and in (b) LC3\textsubscript{mem} - ‘yes, yes, no, no’. The N-terminal subdomain is colored in green and C-terminal subdomain is colored in blue.

Now, let us take a closer look at the third most populated states for LC3, and LC3\textsubscript{mem}, shown in Figure 4.16. The third most populated LC3 state is ‘yes, yes, no, yes’, the one in which Ile34 contacts the hydrophobic pocket formed by residues Phe7, Phe108, and Tyr110 and there is the presence of two saltbridges between N and C-terminal regions, namely Asp19-Lys51 and Asp106-Arg11 saltbridges. However, Glu105-Arg16 salt bridge, another salt bridge between N and C terminal regions doesn’t exist. For LC3\textsubscript{mem}, the third most populated state is ‘yes, no, yes, no’, the one in which Ile34 contacts the hydrophobic pocket formed by residues Phe7, Phe108, and Tyr110. In this state, Asp19-Lys51 salt bridge does not exist, but Glu105-Arg16 salt bridge exists and finally Asp106-Arg11 salt bridge does not exist. Looking at Figure 4.16, we can see that N-terminal subdomain adopts a more open conformation compared to LC3. Once again, this is evident from the state labels: LC3 state has 3 ‘yes’, 1 ‘no’ and LC3\textsubscript{mem} has 2 ‘yes’, 2 ‘no’. This means LC3 prefers a
more closed structure compared to LC3_mem. To understand more about the motion of N-terminal subdomains between LC3, and LC3_mem, we can calculate the RMSD values of the N-terminal regions; RMSD for LC3 is 1.98 Å, whereas RMSD for LC3_mem is 4.46 Å. Once again, we are calculating the RMSD for N-terminal regions. The higher RMSD value for LC3_mem compared to LC3, shows that N-terminal region is more mobile for LC3_mem.

![Figure 4.16: The third most populated states for (a) LC3 - 'yes, yes, no, yes' and in (b) LC3_mem - 'yes, no, yes, no'. The N-terminal subdomain is colored in green and C-terminal subdomain is colored in blue.](image)

The fourth most populated states of LC3, LC3_mem are shown in Figure 4.17. For LC3, state 4 is 'yes, yes, yes, yes', the one in which Ile34 contacts the hydrophobic pocket formed by residues Phe7, Phe108, and Tyr110 and there is the presence of three saltbridges between N and C-terminal regions, namely Asp19-Lys51, Glu105-Arg16, and Asp106-Arg11 saltbridges. However, for LC3_mem, state 4 is 'yes, yes, yes, no', the one in which Ile34 contacts the hydrophobic pocket formed by residues Phe7, Phe108, and Tyr110. In this state, salt bridges Asp19-Lys51 and Glu105-Arg16 exist but salt bridge Asp106-Arg11 does not exist. In Figure 4.16, we can see that N-terminal subdomain adopts a more open conformation compared to LC3. This is clear from the state labels: LC3 state has 4 'yes
'and LC3_mem has 3 'yes', 1 'no'. Thus LC3 prefers a more closed structure compared to LC3_mem. Again, to understand more about the motion of N-terminal subdomains, we can calculate the RMSD values of the N-terminal regions; RMSD for LC3 is 1.55 Å, whereas RMSD for LC3_mem is 4.55 Å. The higher RMSD value for LC3_mem compared to LC3, shows that N-terminal region is more mobile for LC3_mem.

![Figure 4.17](image)

**Figure 4.17:** The fourth most populated states for (a) LC3 - 'yes, yes, yes, yes' and in (b) LC3_mem - 'yes, yes, yes, no'. The N-terminal subdomain is colored in green and C-terminal subdomain is colored in blue.

### 4.2.5.1 salt bridge Analysis

Since clustering method II has a strong influence from salt bridges as three out of the four criteria deals with salt bridges, we did some basic salt bridge analysis for LC3, and LC3_mem. Using the real definition of salt bridges, which states that salt bridges are formed when two oppositely charged functional groups are within a distance of 4 Å. In the case of proteins, this definition is: salt bridges are formed if the distance between nitrogens in the basic side chain and oxygens in the acidic side chain is ≤ 4 Å [8, 29]. The acidic amino acids include Aspartic acid and Glutamic acid. Arginine, Histidine, and Lysine are basic amino
acids. Studies show that salt bridges between Asp or Glu and His, Arg, or Lys display well defined conformational preferences [29]. By visually inspecting the different conformations and through the methods discussed earlier in this dissertation, we know that salt bridges between N and C terminal makes the structure more compact. Using an in house script, we found that there are 37 salt bridges between N and C-terminal subdomains for LC3 and for LC3\_mem, there are 13 salt bridges. That means LC3 adopts conformations that are more compact compared to LC3\_mem. The salt bridges between N and C-terminal regions for LC3, and LC3\_mem are shown in Figure 4.18.

![Figure 4.18](image1.png)

**Figure 4.18:** The salt bridges between N and C-terminal regions of (a) LC3 - 37 salt bridges and (b) LC3\_mem - 13 salt bridges. The acidic residues (Asp, Glu) are shown in red CPK representation and basic residues (Arg, His, Lys) are shown in blue CPK representation.

Another interesting thing to note is that, the initial crystal structures of both LC3, and LC3\_mem has fewer number of salt bridges between N and C-terminal subdomains. LC3 had 3 and LC3\_mem had 2 salt bridges. Other studies have also found salt bridges during the simulation that are not present in the initial X-ray crystal structures of proteins [29,34].
4.2.6 Conclusion and Future Work

Both clustering methods, clustering method I and II, show that LC3 prefers compact conformations compared to membrane associated LC3. Also, both methods show that, when in solution LC3 can explore more conformations than when associated with the membrane. Clustering Method I gave similar conformations for LC3 and LC3\_mem. However in clustering method II, the conformations adopted by LC3 are different from those when associated with the membrane. Additionally, we saw the N-terminal region of LC3\_mem is more open compared to LC3. The salt bridges analysis also support this, as there are about three times the number of salt bridges between N and C terminal region for LC3 compared to LC3\_mem. These patterns support our hypothesis that upon binding to autophagosome, LC3 undergoes a conformational transition, that allows it to associate with another membrane bound LC3 via the N-terminal region.

The openness we see in the N-terminal domain of LC3\_mem is one of the first steps in understanding how membrane associated LC3 proteins interact with each other and if they play a role in the closure of the isolation membrane. The next thing to look at is how exactly does the N-terminal subdomain interacts with the another membrane associate LC3. Will the hydrophobic residues present in one LC3\_mem will act as hydrophobic pocket for residues from another LC3\_mem. We might be able to find answers to these questions if we do simulations with two membrane associated LC3s. There are studies that looked at the effects of neighboring molecules by performing multimolecule MD simulations [4,49]. Nath et al. showed that lipidation of LC3/GABARAP happens on membranes possessing local packing defects [52]. This packing defect might be an interesting aspect to include in the simulations with two membrane associated LC3s. As we find the exact mechanism of how membrane associated LC3s interact with each other, another important aspect to consider is the role of this behavior: as the isolation membrane or phagophore expands, these interactions can recruit other proteins that help in the expansion of isolation membrane or they can close the mebrane or they can do a combination of both. More experiments and simulations are needed to understand these areas of autophagy.
Bibliography


[20] Finnerty, J. Molecular dynamics meets the physical world: Thermostats and barostats.


[48] MARK, P., and NILSSON, L. Structure and dynamics of liquid water with different long-range interaction truncation and temperature control methods in molecular


[61] Statistics, D. Defining initial clusters.


[76] **Yang, Z., and Klionsky, D. J.** An overview of the molecular mechanism of autophagy. *Current Topics in Microbiology and Immunology* 335 (2009), 132.
Appendix A

Appendix

A.1 Abbreviations of amino acids

Ala - Alanine  
Arg - Arginine  
Asn - Asparagine  
Asp - Aspartic Acid  
Cys - Cysteine  
Glu - Glutamic Acid  
Gln - Glutamine  
Gly - Glycine  
His - Histidine  
Ile - Isoleucine  
Leu - Leucine  
Lys - Lysine  
Met - Methionine  
Phe - Phenylalanine  
Pro - Proline  
Ser - Serine
Thr-Threonine
Tyr - Tyrosine
Trp - Tryptophan
Tyr - Tyrosine
Val - Valine

A.2 Other abbreviations

Atg - autophagy-related
ATP - adenosine triphosphate, a coenzyme cells use for energy storage.
kDa- kilo Dalton, Dalton is defined as 1/12th mass of a carbon atom.
GABARAP - gamma-aminobutyric acid type A receptor-associated protein
GATE-16 - Golgi-associated ATPase enhancer of 16kDa
HP - Hydrophobic pocket. In chapter 3, the HP of LC3 is formed by residues Ser3,Tyr38, Leu82, and His86. In reality, only two out of four residues (Tyr, Leu) are hydrophobic. But still we called the pocket, a hydrophobic pocket or HP. The same definition applies for HP for ‘Clustering Method I’discussed in chapter 4.
HREMD - Hamiltonian replica exchange molecular dynamics
LC3 - microtubule-associated protein 1A/1B light chain 3 (LC3)
LC3-II - membrane associated form of LC3
LC3_mem - membrane associated form of LC3
MD - molecular dynamics
REMD - replica exchange molecular dynamics
Appendix B

Simulation details

Both LC3 and LC3_mem systems were simulated for 200 ns before starting the HREMD simulations.

B.1 Regular molecular dynamics- simulation details

Systems were run with GROMACS engine using CHARMM 27 force field with TIP3P parameters. Each system was minimized and equilibrated to 1 atm and 300 K and was simulated for 200 ns. Particle-Mesh Ewald summations were used to treat electrostatics, with a grid size of 1Å. The cutoff for nonbonded interactions was 12Å. The integration time step was 1 fs for equilibration and 2 fs for the production run. Pressure was controlled using a Parinello-Rahman barostat and temperature was maintained using a Nose-Hoover thermostat.

B.2 Hamiltonian replica exchange molecular dynamics- simulation details

For HREMD, we used similar set up to that of regular MD. Here we have 32 replicas with scaling factors between 1.00 and 0.55. The exact scaling factors were 1.00, 0.98, 0.96, 0.94,
0.92, 0.91, 0.89, 0.87, 0.86, 0.84, 0.82, 0.81, 0.79, 0.78, 0.76, 0.75, 0.73, 0.72, 0.70, 0.79, 0.68, 0.66, 0.65, 0.64, 0.63, 0.61, 0.60, 0.59, 0.58, 0.57, 0.56, and 0.55. These scaling factors correspond to temperatures between 300 K to 550 K. For details see chapter methods.