REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
12-12-2014	Final Report	1-May-2011 - 30-Sep-2014

4. TITLE AND SUBTITLE
Final Report: High Performance Computing and Enabling Technologies for Nano and Bio Systems and Interfaces

5a. CONTRACT NUMBER
W911NF-11-1-0168

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
633002

6. AUTHORS
Ajit Kelkar, Ram Mohan, Roy George

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
North Carolina A&T State University
1601 East Market Street

Greensboro, NC 27411 -0001

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
60143-ST-H.23

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT
The understanding bio-nano interactions through high performance computational modeling is essential for molecular discovery and health monitoring in extreme environments; military medicine. Building upon the presence of molecular level features at the nano level, the project efforts focused on the molecular and multi-scale modeling of biological relevance. These investigations encompassed: Computational molecular dynamics modeling of protein – aptamer interactions and their relevance in aptamer based biosensors for detection of disease biomarkers; Development of coarse grain modeling approaches demonstrated via high performance computational

15. SUBJECT TERMS
computational modeling, bio-nano interfaces, bio systems, enabling technologies

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Ajit Kelkar
UU	UU	UU	UU		19b. TELEPHONE NUMBER 336-285-2864

Report Title

Final Report: High Performance Computing and Enabling Technologies for Nano and Bio Systems and Interfaces

ABSTRACT

The understanding bio-nano interactions through high performance computational modeling is essential for molecular discovery and health monitoring in extreme environments; military medicine. Building upon the presence of molecular level features at the nano level, the project efforts focused on the molecular and multi-scale modeling of biological relevance. These investigations encompassed: Computational molecular dynamics modeling of protein – aptamer interactions and their relevance in aptamer based biosensors for detection of disease biomarkers; Development of coarse grain modeling approaches demonstrated via high performance computational modeling of micellar nanocarriers for medicinal drug delivery; Computational modeling studies of fullerene based molecular therapeutic agents; Molecular Dynamics modeling analysis of biomolecules in binary solvents; computational methods for peptide-aptamer binding. The volume, complexity and speed of generation of data in multi-scale modeling and complex large scale high performance computing demand of automated techniques to understand the use of this data. Enabling technologies for intelligent data mining and algorithms for data analysis of protein – aptamer interaction systems were developed. All research investigations contributed to the research education, and training of several graduate students and researchers resulting in technical publications, MS thesis and Ph.D. dissertation. The project efforts and the detailed discussions in this report are aligned along these lines.

Enter List of papers submitted or published that acknowledge ARO support from the start of the project to the date of this printing. List the papers, including journal references, in the following categories:

(a) Papers published in peer-reviewed journals (N/A for none)

Received		Paper
08/31/2013	9.00	Ram V. Mohan, Ajit D. Kelkar, Goundla Srinivas. Polymer Micelle Assisted Transport and Delivery of Model Hydrophilic Components inside a Biological Lipid Vesicle: A Coarse-Grain Simulation Study, The Journal of Physical Chemistry B, (08 2013): 0. doi: 10.1021/jp405381k
TOTAL:	**1**	

Number of Papers published in peer-reviewed journals:

(b) Papers published in non-peer-reviewed journals (N/A for none)

Received		Paper
TOTAL:		

Number of Papers published in non peer-reviewed journals:

(c) Presentations

Number of Presentations: 3.00

Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received		Paper
12/12/2014	20.00	Roy George, H. Nair, K. A. Shujaee, D. A. Krooks, C. M. Armstrong. Automated Annotation of Text Using the Classification Based Annotation Workbench, International conference on Intelligent Systems and Applications. 10-APR-13, . : ,

TOTAL: **1**

Number of Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received		Paper
08/30/2013	6.00	K. P. Liyanage, R. George, K. Shujaee. Outliner Detection in Spatial Data using the m-snn Algorithm, IEEE South Eastern Conference. 08-APR-13, . : ,
08/31/2012	2.00	Krishani Liyanage, Roy George, Khalil Shujaee. Evaluation of Outlier Detection in Spatial Data using the SNN Algorithm, DATA 2012: International Conference on Data Technologies and Applications. 25-JUL-12, . : ,
08/31/2013	7.00	R. George, H. Nair, K. A. Shujaee, D. A. Crooks, C. M. Armstrong. Automated Annotation of Text Using the Classification-based Annotation Workbench (CLAW), 2nd International Conference on Intelligent Systems and Applications. 21-APR-13, . : ,
12/12/2014	16.00	Zeynab Bidoni, Roy George. Network Service Quality Ranking: A Network Selection Algorithm for Heterogeneous Networks, ACM/IEEE Symposium on Architectures for Networking and Communications Systems. 10-OCT-14, . : ,
12/12/2014	18.00	Bahrami Bidoni, Roy George, K. A. Shujaee. A Generalization of the PageRank Algorithm, Eighth International Conference on Digital Society. 10-APR-14, . : ,
12/12/2014	19.00	Zeynab Bidoni, Roy George. Discovering Community Structure in Dynamic Social Networks using the Correlation Density Rank, ASE International Conference on Social Computing. 10-MAY-14, . : ,
12/12/2014	21.00	K. P. Liyanage, R. George, K. Shujaee. Outlier Detection in Spatial Data using the m--SNN Algorithm, IEEE Southeastern Conference. 22-APR-13, . : ,

TOTAL: **7**

Number of Peer-Reviewed Conference Proceeding publications (other than abstracts):

(d) Manuscripts

Received	Paper
12/12/2014 10.00	Gounda Srinivas, Ram Mohan, Ajit Kelkar. Polymer Micelle Assisted Transport and Delivery of Model Hydrophylic Components inside a Biological Lipid Vesicle: A Coarse-Grain Simulation Study, The Journal of Physical Chemistry B (08 2013)
12/12/2014 11.00	Goundla Srinivas, Henry Ochije, Ram Mohan. Biomolecules in Binary Solvent: Computer Simulation Study of Lysozyme Protein in Ethanol-Water Mixed Solvent Environment, JSM Nantechnology & Nanomedicine (08 2014)
12/12/2014 15.00	Kristen Rhinehardt, Ram Mohan, Goundla Srinivas, Ajit Kelkar. Computational Modeling of Peptide - Aptamer Binding in Biosensor Applications, International Journal of Bioscience, Biochemistry, and Bioinformatics (03 2013)
TOTAL:	**3**

Number of Manuscripts:

Books

Received	Book
TOTAL:	

Received	Book Chapter
12/12/2014 12.00	Kristen Rhinehardt, Ram Mohan, Goundla Srinivas. Computational Modeling of Peptide-Aptamer Binding, New York: Springer Science, (01 2015)
12/12/2014 13.00	Goundla Srinivas, Ram Mohan, Ajit Kelkar. Computational Modeling of Nano-Bio Interfaces, Boca Raton: CRC Press, (04 2014)
TOTAL:	**2**

Patents Submitted

Patents Awarded

Awards

R. Mohan, Senior Researcher of the Year Award 2012
A. Kelkar, Intellectual Property Award, 2012
R. Mohan, Track Chair, IMECE 2014, ASME
R. Mohan, Keynote Invited Presentation, International Conference on Nanoscience and Engineering, Jawaharlal Nehru Technical University, June 2014
R. Mohan, Invited Keynote, World Congress on Computational Mechanics, July 2014
A. Kelkar, Plenary Speaker, NANOCON 2014, October 2014
R. Mohan, Promoted to Full Professor at North Carolina A&T State University
A. Kelkar, Conference Orgainzer, NANOCON 2014
R. Mohan, Invited Speaker, University of Manchester, UK, July 2014.

Graduate Students

NAME	PERCENT_SUPPORTED	Discipline
Kristen Rhinehardt	0.50	
Henry Ochije	0.50	
Aaron Jordan	0.50	
K. P. Liyanage	0.50	
H. B. Shashikala	0.50	
O. Eltayeby	0.50	
Z. Bidoni	0.50	
FTE Equivalent:	**3.50**	
Total Number:	**7**	

Names of Post Doctorates

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Names of Faculty Supported

NAME	PERCENT_SUPPORTED	National Academy Member
Ajit Kelkar	0.16	
Ram Mohan	0.10	
Roy George	0.20	
FTE Equivalent:	**0.46**	
Total Number:	**3**	

Names of Under Graduate students supported

NAME	PERCENT_SUPPORTED	Discipline
S. Reed	0.25	Computer Science
M. Brooks	0.25	Computer Science
FTE Equivalent:	**0.50**	
Total Number:	**2**	

Student Metrics

This section only applies to graduating undergraduates supported by this agreement in this reporting period

The number of undergraduates funded by this agreement who graduated during this period: 0.00

The number of undergraduates funded by this agreement who graduated during this period with a degree in science, mathematics, engineering, or technology fields:..... 2.00

The number of undergraduates funded by your agreement who graduated during this period and will continue to pursue a graduate or Ph.D. degree in science, mathematics, engineering, or technology fields:..... 1.00

Number of graduating undergraduates who achieved a 3.5 GPA to 4.0 (4.0 max scale):..... 2.00

Number of graduating undergraduates funded by a DoD funded Center of Excellence grant for Education, Research and Engineering:..... 0.00

The number of undergraduates funded by your agreement who graduated during this period and intend to work for the Department of Defense 1.00

The number of undergraduates funded by your agreement who graduated during this period and will receive scholarships or fellowships for further studies in science, mathematics, engineering or technology fields:..... 0.00

Names of Personnel receiving masters degrees

NAME	
Omar El Tayebey	
Henry Ochije	
Total Number:	**2**

Names of personnel receiving PHDs

NAME	
New Entry	
Total Number:	**1**

Names of other research staff

NAME	PERCENT_SUPPORTED
G. Srinivas (Research Scientist)	1.00
K. A. Shujaee	0.50
M. A. Sazegar	0.50
FTE Equivalent:	**2.00**
Total Number:	**3**

Sub Contractors (DD882)

Inventions (DD882)

Scientific Progress

See Attachment

Technology Transfer

Presented research on bio-nano modeling to Dr. Stephen Lee
Presented current work on bio-nano modeling to TransTech Pharma
Roy George interacted with ARL and CERL researchers
Research interaction on bio-modeling with ERDC

Final Technical Report
May 1, 2011 – September 30, 2014
ARO Award No: W911NF-11-1-0168
Proposal Number: 60143-ST-H

High Performance Computing and Enabling Technologies for Nano and Bio Systems and Interfaces

Lead Institution

North Carolina A & T State University
Greensboro, NC
PI: Dr. Ajit Kelkar
Co-PI: Dr. Ram Mohan

Partnering Institution

Clark Atlanta University
Atlanta, GA
PI: Dr. R. George

Technical Contact

Dr. Ajit Kelkar
Professor and Chair
Nanoengineering
Joint School of Nanoscience and Nanoengineering
2907 E Lee Street
North Carolina A & T State University
Greensboro, NC 27401
Phone: 336-285-2864
Fax: 336-500-0115
Email: **kelkar@ncat.edu**

Table of Contents

Foreword

The biological responses of a human body are based on sensing of the signals within the body. These could be chemical (variation in the plasma glucose concentration), biochemical (presence of virus), mechanical (accelerations such as impacts or pressure drops) or thermal. The biological response to these events are triggering of associated responses that are characterized by molecular changes in the composition of blood protein and lipid structures. Nano sensors detecting such variations are based on molecular medicine. The development and design of such bio-sensors require a molecular level interaction of abiotic – biotic systems (material systems such as silicon, metals, polymers) interacting with biological constituents; and their molecular diffusion. From an engineering perspective, these interactions of engineered material systems (nano) and biological systems (bio) and nano-bio interfaces are based on complex chemical-physical interactions and exhibit multi-scale, multi-physics characteristics. The characteristics of electron, thermo and mass transportation and their deformation behavior play a critical role in nano systems. In bio-systems, lipid layers encapsulate cells and control fluidity and surface tensions, which are macroscopic/hydrodynamic quantities. The biotic-abiotic systems bring such complex systems together and research in such nano-bio systems have significant potential for advancement in medical and health science applications. Complete understanding of these systems requires multi-scale, multi-physics simulation methods by overcoming the limitations of time and length scales. Computational modeling and simulations enabled through high performance computing have become essential tools in several scientific and engineering disciplines. Such techniques are essential to extend current simulation methods to provide understanding of complex nano-bio systems that are at present are generally based on experimental trial and error approaches. The overall objective of this proposal is to focus on such high performance computing, multi-scale modeling and enabling technologies for nano-bio systems and interfaces. The understanding of such nano-bio interactions through high performance computational modeling is essential to invoke more complex, closed loop systems for molecular discovery and health monitoring in extreme environments, for military medicine (for example, detection of mild traumatic brain injuries), athletics, etc.

There is a clear commonality of relevant features and structures at different length scales in both engineering and biological systems that govern the influence and behavior at the system. At the nano scale, the basic constituents of all condensed matter are the atoms bound together in chemical bonds. Processes occurring at certain scale govern and influence the behavior at the disparate length scales present in the system conforming to the structure – property paradigm in material science. In both the engineering and biological material world and in the interactions between the biotic – abiotic systems, the basic microscopic constituents of the material are atoms. The interaction of atoms at the microscopic level (length scale: nano meters; time scale: femto seconds) determine the behavior at the macroscopic scale (length scale: centimeters and beyond; time scale: milli seconds and beyond).

Building upon the presence of molecular level features at the nano level, the project efforts at North Carolina A&T State University focused on the molecular and multi-scale modeling of biological relevance. These investigations encompassed:

- Computational molecular dynamics modeling of protein – aptamer interactions. Aptamer based biosensors are highly efficient with high specificity and reusability. Within the biosensor the aptamers are immobilized to maximize their access to target molecules. Understanding orientation and location of the binding region for a peptide aptamer complex is critical in their applicability in a biosensor.
- Development of coarse grain modeling approaches demonstrated via high performance computational modeling of micellar nanocarriers for medicinal drug delivery. Coarse grain modeling approaches are essential for effective computational modeling of large scale biological systems and multi-scale modeling approaches to provide mesoscopic understanding of the biological processes.
- Computational modeling studies of fullerene based molecular therapeutic agents. Coarse grain modeling approaches for fullerene based molecular therapeutic agents for assisted drug delivery were investigated in the present work.
- Molecular Dynamics modeling analysis of biomolecules in binary solvents. Proteins are building blocks of biological systems and play an important role from the health and medical perspective in the drug reactions and efficiency. Project efforts focused on molecular level understanding of solvent effect on lysozyme protein in a water-ethanol binary solvent.
- Computational modeling work on peptide-aptamer binding leading to a literature review research resulting on an invited book chapter on the various methods leading to molecular dynamics that are applicable to such biological processes, in particular peptide-aptamer binding.

The volume, complexity and speed of generation of data in multi-scale modeling and complex large scale high performance computing demand the development of automated techniques to understand the use of this data. Conforming to this need, the project efforts at Clark Atlanta University focused on the enabling technologies for intelligent data mining and algorithms for data analysis. The project efforts and the detailed discussions in this technical report are aligned along these lines.

1. **Binding of Anti-MUC1 Aptamer and Mucin 1 Peptides: Molecular Dynamics Analysis and Relevance towards Biosensor Development** *(North Carolina A&T State University)*

The biological response of the human body to diseases, traumatic and post-traumatic events is characterized by molecular changes in the composition of blood protein and lipid structures (biomarkers). Nano sensors detecting such variations require a molecular level interaction of abiotic-biotic systems. Research, education and training of future US work force in such nano-bio systems have significant potential for advancement in medical and health science applications. High performance computational modeling is essential to invoke more complex, closed loop systems for molecular discovery, targeted drug delivery, diagnosis and prognosis for diseases, traumatic and post-traumatic events in military medicine (for example, detection of mild traumatic brain injuries, post-traumatic stress disorders) as well as athletics, etc. A summary of research conducted in this area focusing on the molecular dynamics modeling of the MUC 1 peptide and Anti-MUC1 aptamer binding and their relevance towards biosensor development is presented next

Aptamer based biosensors are highly efficient, with high specificity and reusability. Within the biosensor the aptamers are immobilized to maximize their access to target molecules. Knowledge of the orientation and location of the aptamer and peptide during binding could be gained through computational modeling. Experimentally, the aptamer (anti-MUC1 S2.2) has been identified as a breast cancer biomarker mucin 1 (MUC1) protein. However, within this protein lie several peptide variants with the common sequence ADPTRPAP that are targeted by the aptamer. Understanding orientation and location of the binding region for a peptide-aptamer complex is critical in their applicability in a biosensor. In this study, we explore the use of computational modeling to determine how this peptide sequence and its minor variants affect binding. We use molecular dynamics simulations to study multiple peptide-aptamer systems consisting of MUC1 (APDTRPAP) and MUC1-G (APDTRPAPG) peptides with the anti-MUC1 aptamer under similar physiological conditions reported experimentally. In the case of the MUC1 peptide and aptamer, simulations reveal that the aptamer and peptide interact between 3' and 5' ends but do not fully bind. Multiple simulations of the MUC1-G peptide indicate consistent binding with the thymine loop of the aptamer, initiated by the arginine residue of the peptide. We find that the binding event induces structural changes in the aptamer by altering the number of hydrogen bonds within the aptamer and establishing a stable peptide-aptamer complex. In all MUC1-G cases the binding event was confirmed by systematically studying the distance distributions between peptide and aptamers. We observe that the MUC1 peptide-aptamer binding is less stable compared to MUC1-G peptide-aptamer combination. These results are found to corroborate well with the experimental findings. Simulations highlight the role of the arginine residue in initiating the binding. The addition of the glycine residue to the peptide, as in the case of MUC1-G, provided a more stable binding event. Our study demonstrates the ability of MD simulations to obtain molecular insights for peptide-aptamer binding. Details on

the orientation and location of binding between the peptide-aptamer can be instrumental in biosensor development.

2. High Performance and Multi-Scale Computational Modeling in Bio-Nano Systems *(North Carolina A&T State University)*

Computational modeling and simulation play integral role in the scientific and technological advances. In particular, its application in biomedical research and nanotechnology revolutionized drug discovery. A fundamental assumption for computational modeling and simulation is that insight into system behavior can be developed or enhanced from a model that adequately represents a selected subset of the system's attributes. There is a wide variety of simulation techniques are available. Depending on properties and applicability, they can be broadly categorized into three classes: 1. Quantum simulations 2. Classical simulations and 3. Mesoscopic simulations. Each method has its merits and limitations as they have been designed to efficiently handle different spatial and temporal scales. For example, for studying chemical bond formation and bond dissociation between atoms and molecules involving electronic interactions, a simulation method based on quantum mechanics is an appropriate choice. However, due to the complex interactions and computationally intense numerical calculations, this method can only be applicable for systems with relatively smaller length and time scales. Nevertheless, many interesting phenomena relevant to biological and material sciences involves time and length scales over micrometers and microseconds and beyond. Such problems cannot be directly dealt with methods based on quantum mechanics. On the other hand, classical simulations based on Newtonian mechanics, are highly suitable to deal with such large time and length scales appropriately. For this very reason, classical molecular dynamics (MD) is the most commonly used simulation method. The classical MD approach is routinely used to explore tens of nanosecond times and nanometer length scales. However, such efficiency over quantum simulations comes with a penalty of relatively lower accuracy. Naturally, more efficient methods such as mescoscopic simulations that explore time and length scales beyond classical/atomistic simulation capabilities are relatively less accurate. Bridging these desperate spatial and temporal scales is essential to explore complex biomedical engineering problems in detail. Details of the work in this area during the project efforts are discussed.

2.1 Polymer Micelle Assisted Transport and Delivery of Model Hydrophilic Components inside a Biological Lipid Vesicle: A Coarse Grain Simulation Study *(North Carolina A&T State University)*

Understanding drug transportation and delivery mechanism from a molecular viewpoint is essential to find better treatment pathways. Despite the fact many significant drugs such as anticancer doxorubicin and mitoxantrone are predominantly hydrophilic, efficient methodology

to deliver hydrophilic drug components is not well established. Here we explore this problem by studying "patchy" polymeric micelle assisted hydrophilic component transportation across a lipid membrane and delivery inside a biological lipid vesicle. Using the MARTINI force field as the basis, we study the interaction of polymeric micelle with POPC lipid vesicles in detail. In order to facilitate hydrophilic drug transportation study, a primitive CG model for hydrophilic drug component is used. Extensive simulations carried out over hundreds of nanoseconds demonstrate successful encapsulation, transportation of hydrophilic components by patchy polymeric micelle. Results show the polymeric micelle releases significant portion of hydrophilic contents inside the lipid vesicle. Present simulation study also reveals a possible mechanism for efficient hydrophilic component transportation and delivery. Insights from this study could potentially help experimental community to design better delivery vehicles, especially for hydrophilic drug molecules.

2.2 Computational Modeling Studies of Fullerene Based Molecular Therapeutic Agents *(North Carolina A&T State University)*

In addition to the polymeric micelle carriers, many nano-material based nanocarriers have been explored for drug transportation and delivery. For example, carbon nanotubes and nanoparticles have been aggressively explored for this purpose. Among these nanomaterials, fullerene based molecular therapeutic agents have been highly promising. Encouraged by the success of Coarse Grain (CG) molecular dynamics modeling in the case of polymeric micelle assisted drug-delivery; we explored fullerene nanomaterial based molecular therapeutic agents. In order to design better drug delivery vehicles based on fullerenes, it is highly important to understand fullerenes and its derivatives interaction with cell membrane from a molecular viewpoint. For this purpose, we studied the interaction of fullerenes and amphiphilic polymer-fullerene complexes with biological lipid molecules. In a next step, we also plan to study the fullerene assisted self-assembly of liposomes.

3. Biomolecules in Binary Solvents: Computer Simulation Study of Lysozyme Protein in Ethanol-Water Mixed Solvent Environment *(North Carolina A&T State University)*

Proteins are building blocks of biological systems and play an important role from the health and medical perspective in the drug reactions and efficiency. Proteins function well in their natural water solvent environments and are influenced by modified solvent environments such as alcohol. Effect of protein-solvent interaction on the protein structure is widely studied with experimental and computational techniques. However, molecular level understanding of proteins interaction with many solvents is still not fully understood. The present work aims to obtain a detailed understanding of solvent effect on lysozyme protein, using water, ethanol, and different concentrations of water-ethanol mixtures as solvents. We use detailed atomistic molecular

dynamics simulations to study using GROMACS code. Compared to neat water environment, the lysozome structure shows remarkable changes in water-ethanol mixed solvent, with increasing ethanol concentration. Significant changes were observed in the protein secondary structure involving alpha helices. We found that increasing ethanol concentration results in a systematic increase in total energy, enthalpy, root mean square deviation (RMSD), and radius of gyration of lysozyme protein. A polynomial interpolation approach is presented to determine these quantities for any intermediate alcohol percentage, and compared with the values obtained from a full MD simulation. Results from MD simulation were in good agreement with those obtained from the interpolation approach. The polynomial approach eliminates the need for computationally intensive full MD analysis for the concentrations within the range (0-12%) studied.

4. **Computational Modeling of Peptide-Aptamer Binding** *(North Carolina A&T State University)*

Evolution is the progressive process that holds each living creature in its grasp. From strands of DNA evolution shapes life with response to our ever changing environment and time. It is the continued study of this most primitive process that has led to the advancement of modern biology. The success and failure in the reading, processing, replication and expression of genetic code and its resulting biomolecules keep the delicate balance of life. Investigations into these fundamental processes continue to make headlines as science continues to explore smaller scale interactions with increasing complexity. New applications and advanced understanding of DNA, RNA, peptides and proteins are pushing technology and science forward and together. Today the addition of computers and advances in science has led to the fields of computational biology and chemistry. Through these computational advances it is now possible not only to quantify the end results, but also visualize, analyze, and fully understand mechanisms by gaining deeper insights. The bio-molecular motion that exists governing the physical and chemical phenomena can now be analyzed with the advent of computational modeling. Ever increasing computational power combined with efficient algorithms and components are further expanding the fidelity and scope of such modeling and simulations. This focus here is on computational methods that apply biological processes, in particular computational modeling of peptide - aptamer binding.

5. **Enabling Technologies for Intelligent Data Mining and Algorithms for Data Analysis with Large Variations and Outliers** *(Clark Atlanta University (CAU))*

The volume, complexity and velocity of generation of the data in multi-scale modeling and large scale high performance computing demand the development of automated techniques to understand and use this data. While numerous algorithms have been developed to mine, and visualize data, techniques for discovering and visualizing outliers and rare events in these types of large data sets have not similarly been fully explored. The project research work focused on

the development of enabling technologies for such large scale data mining for application to nano-bio systems. This research effort focused on two different but related enabling technologies:

1. Outlier Detection Algorithms

Outlier detection is critically dependent on the nature of the data that is being analyzed. The type of the data, its distribution and dependencies require tuning of the approach, so that the outliers are detected appropriately. In this research project, we examined two applications, high dimensional spatio-temporal data, and network data and develop techniques for conducting outlier detection in each case. Both these scenarios are relevant to the nano-bio application and may be used in multiple contexts. Algorithms and techniques were developed and employed in several applications.

2. Visualization and Analysis of Large Scale Network Structures

In this project work, CAU team investigated network analysis techniques comprehensively, focusing on the development of new algorithms for this purpose. These algorithms will be integrated with outlier detection algorithms developed in our Year 1 work to give users a useful tool in the analysis of large scale multidimensional data. Application to visualizing outliers in multi-dimensional data.

Contribution to Research and Education Programs:

The infrastructure and results from this grant have been useful to other organization within the US Army. The CAU team was able to leverage this funding to obtain additional funding through contracts from ERDC, ARL, and DoE. The contracts relate to spatio-temporal modeling and analysis of communication data, both of which are topics that directly explored in this research effort. We have two publications resulting from this grant. Six students have been involved with this grant and the associated leveraged projects.

The project efforts enabling the development of research capability in the new direction of computational modeling of bio-nano systems and interfaces at North Carolina A&T State University benefiting and supporting the newly formed MS and Ph.D. program in nanoengineering, first such program at a HBCU in the nation. The project efforts have enabled the strengthening of the computational nanoengineering focus area. Computational nanoengineering, a highly interdisciplinary field with a significant potential to contribute and enhance the field of nanotechnology and understanding of the bio-nano interfaces from an empirical science to a quantitative engineering field. The present project efforts align and further strengthen this. The project efforts are facilitating the research, education and training of graduate students, many of them underrepresented at both the institutions. In addition, the project funding was leveraged towards the acquisition of a multi-processor Cray XC-30 system to meet

the computing analysis needs for computational nanoengineering focus area and is now in operational. These would not have been possible without this project funding and resources.

1. Binding of Anti-MUC1 Aptamer and Mucin 1 Peptides: Molecular Dynamics Analysis and Relevance towards Biosensor Development

K. Rhinehardt, G. Srinivas, R. Mohan, North Carolina A&T State University

The biological response of the human body to diseases, traumatic and post-traumatic events is characterized by molecular changes in the composition of blood protein and lipid structures (biomarkers). Nano sensors detecting such variations require a molecular level interaction of abiotic-biotic systems. Research, education and training of future US work force in such nano-bio systems have significant potential for advancement in medical and health science applications. High performance computational modeling is essential to invoke more complex, closed loop systems for molecular discovery, targeted drug delivery, diagnosis and prognosis for diseases, traumatic and post-traumatic events in military medicine (for example, detection of mild traumatic brain injuries, post-traumatic stress disorders) as well as in athletics, etc. Current research in this area focusing on the molecular dynamics modeling of the Anti-MUC1 Aptamer and MUC1 Peptide Binding and their relevance in biosensor development is presented next.

Abstract

Aptamer based biosensors are highly efficient, with high specificity and reusability. Within the biosensor the aptamers are immobilized to maximize their access to target molecules. Knowledge of the orientation and location of the aptamer and peptide during binding could be gained through computational modeling. Experimentally, the aptamer (anti-MUC1 S2.2) has been identified as a breast cancer biomarker mucin 1 (MUC1) protein. However, within this protein lie several peptide variants with the common sequence ADPTRPAP that are targeted by the aptamer. Understanding orientation and location of the binding region for a peptide-aptamer complex is critical in their applicability in a biosensor. In this study, we explore the use of computational modeling to determine how this peptide sequence and its minor variants affect binding. We use molecular dynamics simulations to study multiple peptide-aptamer systems consisting of MUC1 (APDTRPAP) and MUC1-G (APDTRPAPG) peptides with the anti-MUC1 aptamer under similar physiological conditions reported experimentally. In the case of the MUC1 peptide and aptamer, simulations reveal that the aptamer and peptide interact between 3' and 5' ends but do not fully bind. Multiple simulations of the MUC1-G peptide indicate consistent binding with the thymine loop of the aptamer, initiated by the arginine residue of the peptide. We find that the binding event induces structural changes in the aptamer by altering the number of hydrogen bonds within the aptamer and establishing a stable peptide-aptamer complex. In all MUC1-G cases the binding event was confirmed by systematically studying the

distance distributions between peptide and aptamers. We observe that the MUC1 peptide-aptamer binding is less stable compared to MUC1-G peptide-aptamer combination. These results are found to corroborate well with the experimental findings. Simulations highlight the role of the arginine residue in initiating the binding. The addition of the glycine residue to the peptide, as in the case of MUC1-G, provided a more stable binding event. Our study demonstrates the ability of MD simulations to obtain molecular insights for peptide-aptamer binding. Details on the orientation and location of binding between the peptide-aptamer can be instrumental in biosensor development.

Introduction

Biomarkers are molecules that correspond with or produced by bio-chemical changes in concentration, physiology and morphology (1). These molecules are effective sources as discerning diagnostic tools for detecting and tracking disease progression. Recently, biosensor developments have garnished a lot of interest (2-5). At the genetic level, one gene in particular that is prevalent in breast cancer cells is the mucin 1 (MUC1) gene (6). The MUC1 gene encodes transmembrane mucin proteins (7). In breast carcinomas the MUC1 protein is no longer transmembrane, but rather up-regulated and free floating in the blood stream (7). Carcinogenic isoform of these proteins lack the tandem repeat regions (6, 7) and have O-glycan's which are shorter and expose the core of the protein that contains a variety of peptide epitopes (8-10). Mucin 1 peptides often contain the sequence APDTRPAP. The antibody, SM3, has a high affinity for this cancerous version of MUC1 (7, 11, 12) and has been crystallized with a peptide antigen APDTRPAP (8). The concentration of the MUC1 isoform can be related to disease severity (7, 13). Such concentrations could be detected and quantified within a biosensor.

A biosensor is a receptor-transducer device that provides quantitative information using a bio-recognition element and a transducer (14). The transducer is generally based on changes in electrochemical, mass, optical or thermal properties while the bio-recognition element acts on biochemical mechanism (14, 15). When a biological sample is loaded into the sensor, the bio-recognition element recognizes the target within the sample and binds to it. The transducer registers the change which is quantified. Selecting efficient bio-recognition elements presents a major challenge for biosensor development (14). Antibodies are commonly used bio-recognition elements (16). Antibodies are large molecules that are specific in binding. However, they are not always readily synthesized and can be chemically unstable commonly limiting their application to single use biosensors (17). The large size of the antibodies often limits their application to low density biosensors (17, 18).

Such limitations and difficulties with antibodies motivated the search for efficient alternative bio-recognition elements. Studies have shown that aptamers can be potential bio-recognition elements (16, 17, 19-21). Aptamers are broadly considered as oligonucleotide sequences made of single stranded DNA or RNA (22). Aptamers are advantageous as bio-recognition elements since they are relatively small, chemically more stable and have a high binding affinity that rival or better than that of antibodies (22). Such high binding affinity of aptamers is not only due to sequence level binding but also their ability to reorient in correspondence to that binding. In addition, aptamers can be easily functionalized and immobilized on a surface to create highly ordered receptor layers (14).

Due to the complexity and vast possibilities of combinations, compilations of RNA and DNA oligonucleotides have been compiled into aptamer libraries. For example, a standard 25-mer library compilation has 10^{15} available aptamers (23). Clearly, selecting an appropriate aptamer from such a large pool is not practical. In 1990, the SELEX (Systematic Evolution of Ligands by Exponential Enrichment) process provided an experimental solution for aptamer selection. In this process, small pools of developed aptamer libraries undergo incubation with the desired target molecule (23, 24). Successfully bound aptamers from the incubation process are enhaced through polymerase chain reactions (PCR) (25-27); the incubation and subsequent steps are repeated until a small group of high affinity aptamers are achieved (24, 28). Upon completion of the SELEX cycle these high affinity aptamers are sequenced to uncover their makeup.

After successful selection and sequencing of an aptamer, it can be immobilized in the biosensor. The method of immobilization (2) and orientation of the aptamer within a biosensor directly affects its efficiency (29-31). Biosensor efficiency is greatly increased when the bio-recognition element is oriented in such a way that the binding site is easily accessible (2, 31). Experimental studies have shown that the anti-MUC1 aptamer binds to the mucin 1 protein, in particular the exposed peptide eptiopes that distinguish this carcinagenic isoform (32). Isolating and understanding peptide mechanisms and actions can provide key information on the larger proteins they embody (33). Given that these peptide epitopes are the targets of the aptamer, they can be used in lieu of the proteins for experimentation (32). However, details of binding process including the aptamer orientation and location of binding between unbound MUC1 peptides and the aptamer are not known. These details about the aptamer configuration play an important role in the function of an aptamer based biosensor as small changes in conformation or structure can have significant effects on the detection capability of the biosensor (32, 34, 35). Molecular level insights into such complex phenomenon can be achieved with the aid of suitable computational modeling to improve our understanding of aptamer binding.

In this work computational modeling is employed and shown to be an avenue to test, analyze and visualize the progression of the peptide-aptamer binding. In particular as discussed before binding is the foundation for the aptamer selection process and biosensor detection. Molecular

dynamics is a computational methodology commonly used with biomolecules (33, 36-38). A detailed molecular dynamics modeling analysis of individual MUC1 peptide, MUC1-G peptide, anti-MUC1 aptamer systems in solvent and their interactions are presented in this work. A preliminary study of molecular dynamics modeling was previously discussed for a similar system based on MUC1 peptide (39).

Modeling of the natural progression of peptide-aptamer binding in a solvent identifies key features such as orientation and location of the aptamer in the binding event providing insights for biosensor development. A series of extensive molecular dynamics (MD) simulations of MUC1 and MUC1-G peptides and aptamer combinations in solvent were conducted. Quantification of the results defining the specifics of aptamer-peptide binding and their relevance in biosensor development are discussed.

Methods

The atomistic representation of MUC1 aptamer was used in the present computational modeling study. Initial configurations for the systems studied were obtained from the protein data bank. The 115 atom MUC1 peptide sequence APDTRPAP (shown in Figure 1A) was isolated from the antibody complex using the visualization and analysis software Pymol (40). In order to create a variant of the MUC1 peptide, the MUC1 sequence was further modified by adding a glycine residue forming the MUC1-G peptide (APDTRPAPG) with 122 atoms (shown in Figure 1B). The NMR configuration of the anti-MUC1 aptamer S2.2 contained 728 atoms (Figure 1C). Pymol was used to create the initial configurations of the MUC1 peptide-aptamer and MUC1-G peptide-aptamer molecular systems by combining the original PDB files.

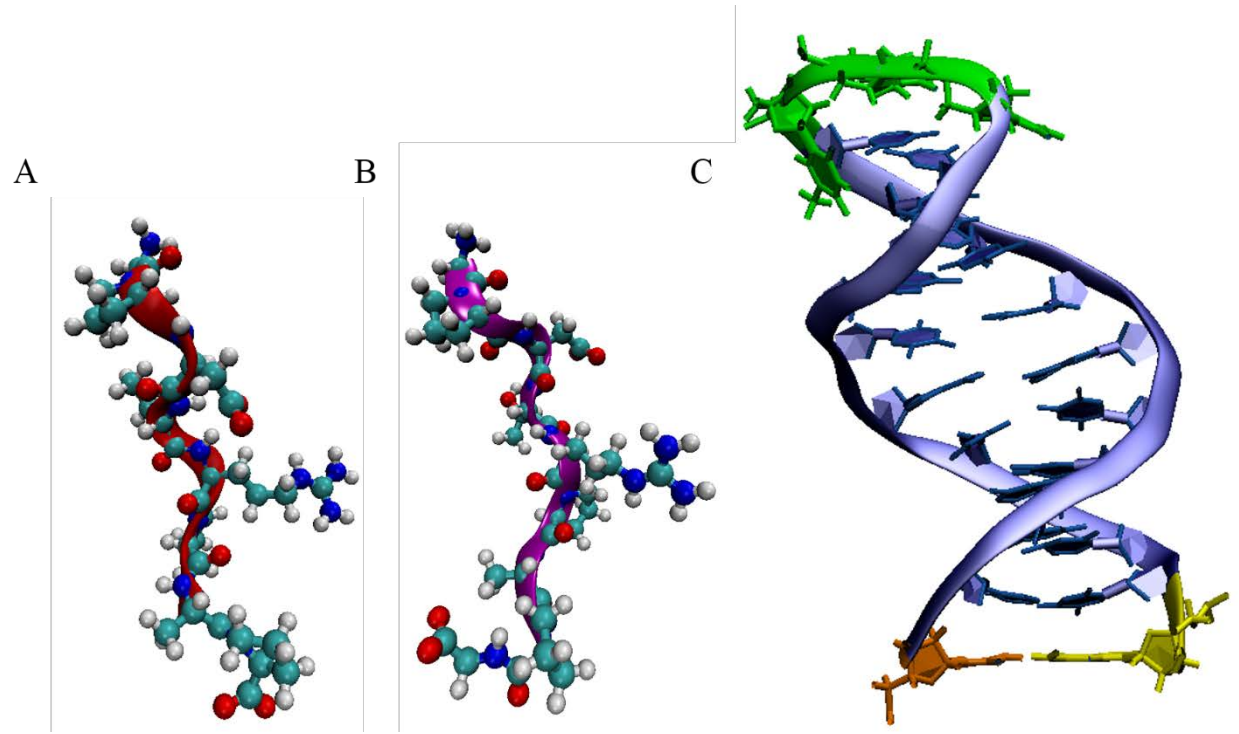


Figure 1: Structures of molecules used in the present study are shown: A) MUC1 peptide B) MUC1-G peptide and C) anti-MUC1 S2.2 23-mer Aptamer with the 3' (highlighted in orange) and 5' (highlighted in yellow) ends illustrated along with the open loop region of the aptamer consisting of 3 thymine nucleotides (highlighted in green). In (A) and (B) peptide backbones are represented in ribbon format, while the side chains are shown in a ball and stick representation. Color code: Oxygen, red; Carbon, cyan; Nitrogen, blue; and Hydrogen, white. In Figure (C), nucleic acids are shown as a ribbon backbone.

Detailed molecular dynamics simulations, using the structures shown in Figure 1, were conducted using GROMACS, an open source efficient MD code that is commonly used to simulate complex biological systems (41). Initially, individual simulations of the MUC1 peptide, MUC1-G peptide and anti-MUC1 aptamer in solvent were carried out. Subsequently, the peptide-aptamer combinations in solvent were simulated. In all the simulations we considered a 0.15M NaCl solvent at the room temperature (300K) and the atmospheric pressure (1bar) to closely mimic the general experimental conditions.

During MD analysis, each biomolecule was minimized to provide a minimal energy configuration using steepest descent technique and the potential energy was carefully analyzed to assure the simulation system reached a stable configuration. This configuration was further thermal and pressure equilibrated using NVT and NPT thermodynamic state ensembles, sequentially. The NVT thermal equilibration was done with all bonds within the biomolecule constrained and the temperature coupled by a velocity rescale thermostat specific to GROMACS

(42, 43). Subsequently, NPT pressure equilibration was applied with the same velocity-rescale temperature coupling in addition to the Parrinello-Rahmen pressure coupling (44). The fully temperature and pressure equilibrated system was then used as the starting configuration for the MD production dynamic analysis. In the production simulation, all the atoms were unconstrained and free to move in their most energetically favorable positions in a dynamic process. All simulations were conducted using a 2 femtoseconds (fs) time step. Individual biomolecular simulations were conducted for a time period of 10 nanoseconds to establish their molecular behavior. Later, multiple binding simulations of the MUC1 peptide-aptamer and MUC1-G peptide-aptamer combinations were conducted for a minimum of 250ns following the same MD procedure. All simulations were performed at standard temperature (300K) and pressures (1 bar) in 0.15M NaCl solvent environment. Details of each simulation system are described in Table 1.

Table 1: Simulation details for individual molecules and binding combinations in solvent.

	Individual Molecule Simulations			**Binding Combination Simulations**				
System	MUC1 Aptamer	MUC1 Peptide	MUC1-G Peptide	MUC1 Peptide and Aptamer	MUC1-G Peptide and Aptamer	MUC1-G_1 Peptide and Aptamer*	MUC1-G_2 Peptide and Aptamer*	MUC1-G_3 Peptide and Aptamer*
Number of Atoms in Biomolecules	728	115	122	843	850	850	850	850
Total Number of NaCl Atoms **	94	26	21	106	228	98	88	180
Water Molecules	12741	4792	3957	15158	37356	13423	11711	28509
Simulation Time (ns)	10	10	10	250	300	300	300	300

**The simulations involving G_1, G_2 and G_3 are similar to MUC1-G, except they have different starting configurations **The total number of NaCl atoms were based on 0.15M NaCl concentration.*

Following the simulation, final configurations of the molecular systems were examined for any structural reorganization using visual molecular dynamics (VMD) software and GROMACS (45). In order to analyze and quantify peptide-aptamer binding, distance between the aptamer and peptide atoms, root mean square deviation (RMSD) (46), radius of gyration (R_g) (47) and the number of hydrogen bonds were analyzed.

Results

A. Simulations of Individual Molecules in Solvent

Individual biomolecular system simulations give insight into the possible conformations and identify binding sites. Simulation snapshots of the individual MUC1 aptamer and peptides are shown in Figure 2. The peptide showed little change in the conformation during the initial stages of transient dynamic simulation. However, its conformation significantly altered as the simulation and transient time proceeded. It was observed that in the initial stages of simulation, peptide structure became compact but in the later stages the structure returned to a conformation similar to that of the starting structure (Figure 2A). Further, the peptide frequently altered its conformation during the simulation due to the flexible nature of the backbone. Conversely, the addition of the glycine residue in the MUC1-G peptide was found to decrease the flexibility of the backbone. For this reason, the backbone appeared to be less flexible compared to the MUC1 peptide although it has free rotational motion (Figure 2B). However, the addition of the glycine does not make the peptide completely rigid, which was evident in the later stages of the transient dynamic simulation analysis.

Unlike the peptides, the aptamer structure showed little or no change during the transient dynamic simulation. MD analysis results revealed two distinct open regions within the aptamer where binding may occur. The first being the loop region of the aptamer that has three non-base paired thymine nucleotides and the second open region include the 3’ and 5’ ends (highlighted regions in Figure 1C). Simulation also showed rotation of the three unbound thymine nucleotides and a relatively less mobility of the nucleotides at the open ends (Figure 2C).

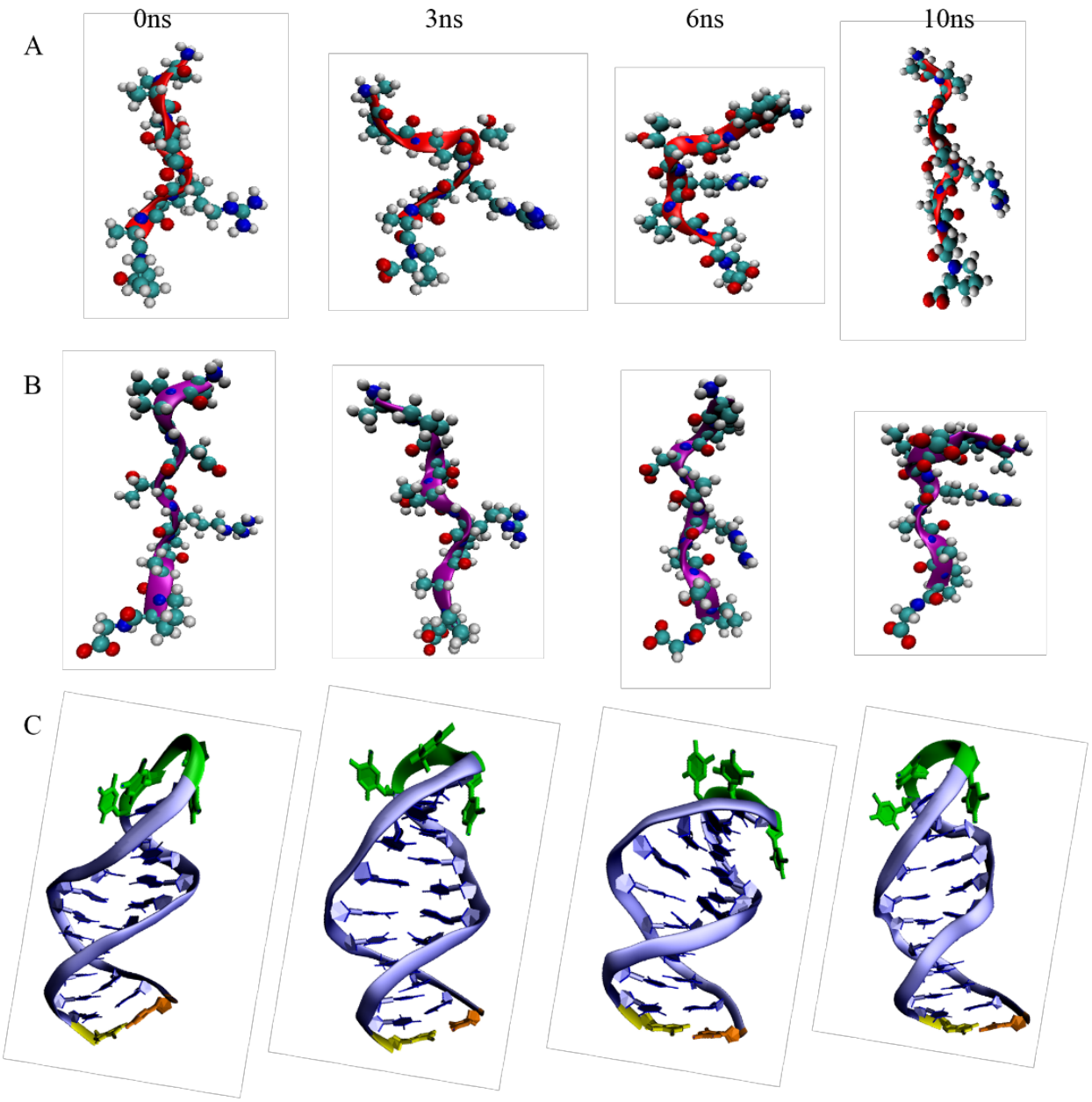


Figure 2: Snapshots at 0ns, 3ns, 6ns, and 10ns obtained from MD simulations of (A) MUC1 peptide (B) MUC1-G peptide and (C) anti-MUC1 aptamer are shown. Water and ions are not shown for figure clarity.

In order to investigate the conformational changes, we calculated the structural quantities such as RMSD and R_g for the aptamer and peptides as shown in Figure 3. The RMSD and R_g of the MUC1 peptide showed a significant variation throughout the simulation (Figure 3). The flexible backbone of the peptide causes large fluctuations as atoms rattle and shift in the confined space.

The changes observed in the backbone movement are also reflected in the RMSD values which fluctuated between 0.12nm and 0.44nm as shown in Figure 3A. The flexibility of the peptide backbone is evident from the fluctuations in R_g as well (Figure 3B).

The addition of the glycine residue in the variant MUC1-G peptide illustrated different structural behavior compared to the MUC1 peptide. Overall, the MUC1-G peptide showed a reduction in fluctuations in the R_g and RMSD (Figure 3). This indicates decreased flexibility of the peptide backbone. Though the MUC1-G peptide appears to be less flexible it is still capable of obtaining the conformations mirrored by the MUC1 peptide. However this peptide did not exhibit these conformations as frequently as was observed in case of the MUC1 peptide. The higher RMSD values shown in figure 3A for the MUC1-G peptide compared to that of MUC1 peptide can be attributed to the addition of the glycine residue. Similarly, the starting value of R_g of 0.8nm for the MUC1-G peptide was higher than that of the MUC1 peptide, which can also be attributed to the addition of this glycine residue. Later stages of simulations indicated that the peptide obtains a more compact motif resulting in a steep reduction in the R_g (Figure 3B). While both peptides can adapt similar conformations the MUC1-G peptide does so with less frequency due to the addition of the glycine residue.

The aptamer showed significantly different structural behavior compared to the peptides. The RMSD of the aptamer initially was at 0.25nm. Throughout the dynamic simulation, the RMSD fluctuates before stabilizing to a value around 0.18nm (Figure 3A). This can be attributed to the flexible movement of the unbound thymine nucelotides in the loop region. The R_g value for the aptamer was found to be consistently around 1.3nm (Figure 3B). This demonstrates that the aptamer has a relatively stable structure with little variation.

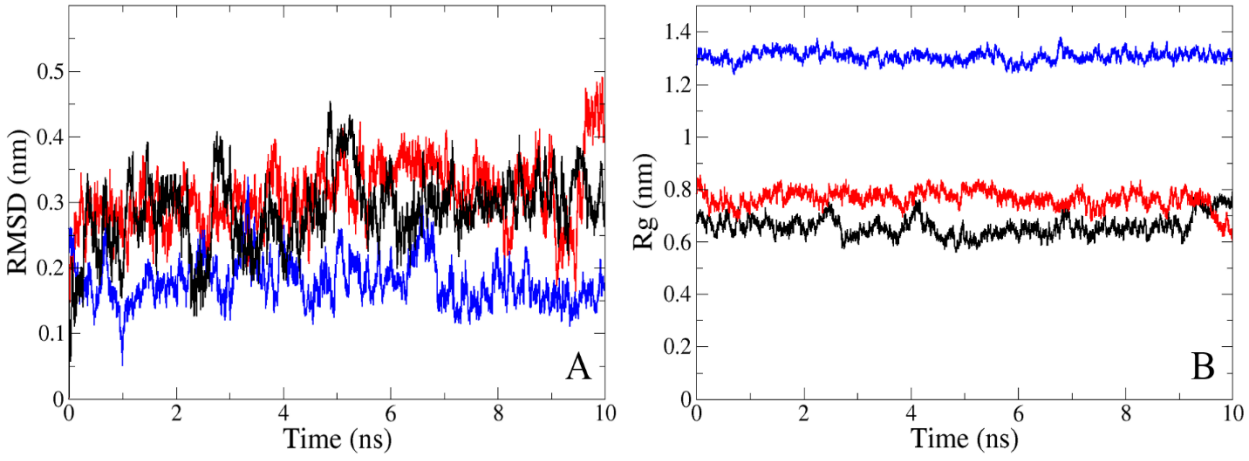


Figure 3: Structural analysis of MUC1 peptide (black), MUC1-G peptide (red) and anti-MUC1 aptamer (blue), obtained from individual biomolecule simulations in solvent. A) RMSD and B) R_g during 10ns simulation are plotted.

B. MD Analysis of Peptide-Aptamer Binding

I. MUC1 Peptide-Aptamer Simulations

After establishing the individual biomolecule behavior, we proceed to examine the peptide-aptamer systems. The binding of anti-MUC1 aptamer and MUC1 peptide was considered first. Snapshots from the molecular dynamics simulation of this peptide-aptamer combination are shown in Figure 4. During the simulation the aptamer and MUC1 peptide move apart compared to their initial configuration (Figure 4A). During the initial stages of the MD progression, the peptide associated briefly with the 12^{th} thymine nucleotide in the aptamer loop (Figure 4B). The peptide then was found to disassociate before a quick re-association (aided by aspartic acid and arginine residues of the peptide) with the 5'end of the aptamer (Figure 4C-D). The peptide was then found to interact via 5' end of the aptamer in a parallel conformation to that of the aptamer. Interestingly, after 10ns of binding the peptide deviates from its parallel orientation. However, the arginine residue of the peptide (5^{th} residue in the peptide sequence) still appears to be interacting with the aptamer at this stage (Figure 4E). Extending the MD simulation analysis to 250ns showed the peptide continued to remain near the 5'and 3' open ends of the aptamer. However, beyond 100ns no direct interaction with the key arginine or aspartic acid residues of the peptide was observed with the open ends of the aptamer (Figure 4F-G).

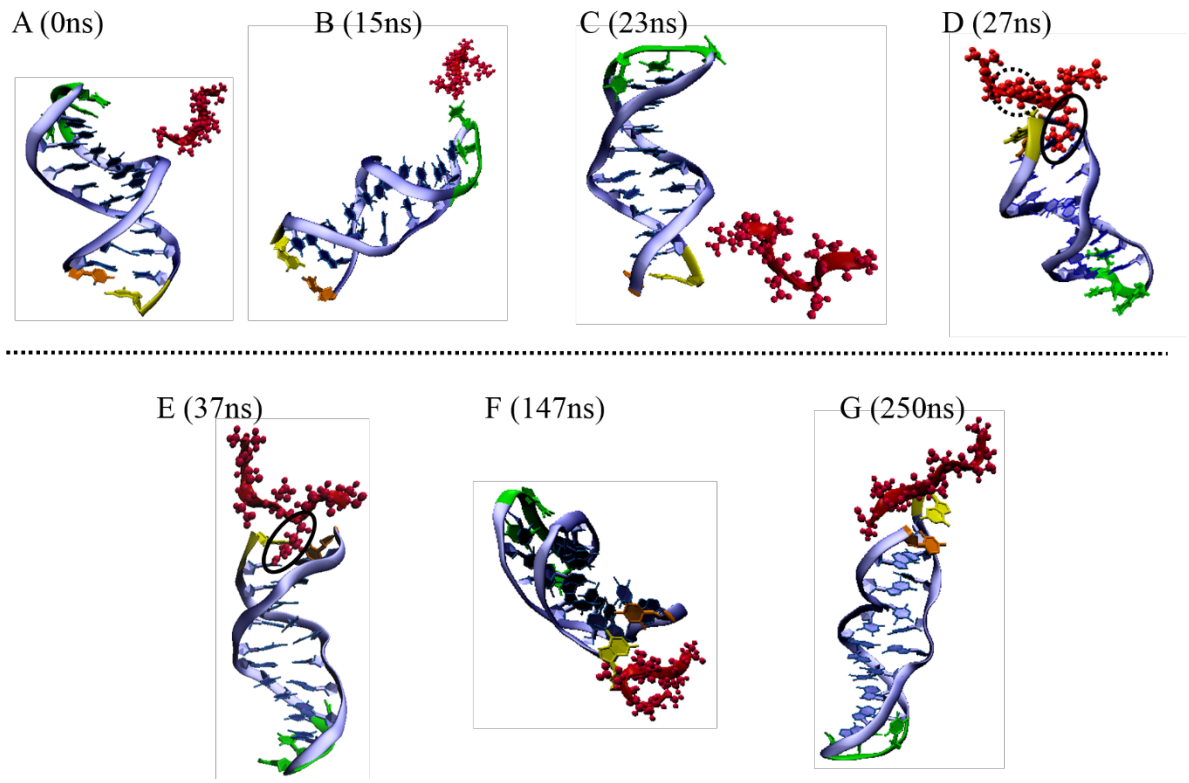


Figure 4: Anti-MUC1 aptamer (blue backbone) and MUC1 peptide (red backbone and residues) snapshots as observed in a 250ns MD simulation are shown (without water or ions). (A) The starting structure (B) A brief association event of aptamer-peptide (C) Dissociation of the aptamer and peptide (D) Initial association between the arginine residue (circled in solid line) and aspartic acid residue (circled in dashed line) of the peptide backbone and 5' end (yellow) of the aptamer are indicated (E) Peptide-aptamer snapshot taken at 10ns after initial association (F) Peptide-aptamer snapshot 120ns after initial association (G) Peptide-aptamer combination at 250ns.

II. MUC1-G Peptide-Aptamer Simulations

Snapshots from the simulation of the MUC1-G peptide-aptamer combination are shown in Figure 5. Analyses show that this combination exhibits markedly different structural behavior from that of the MUC1 peptide-aptamer combination. The addition of the glycine residue to the peptide caused significant changes in MUC1-G peptide-aptamer binding behavior. Unlike the previous combination, the peptide had associated with the open loop region of the aptamer in this case (Figure 5B). The arginine residue of the peptide was found to interact with the 11^{th} and 13^{th} thymine nucleotides of the aptamer (Figure 5C). Extended simulation to 300ns, reveal that the aptamer and peptide were still in the bound state.

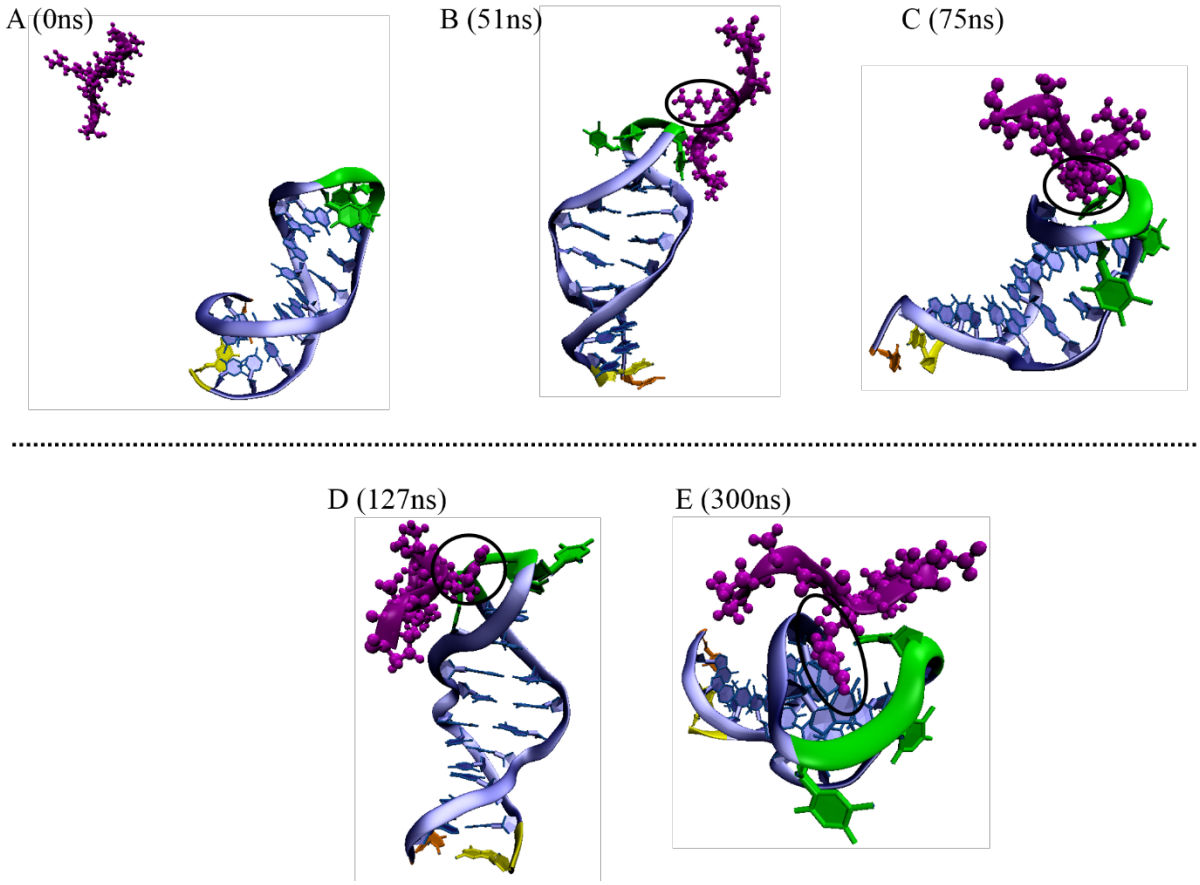


Figure 5: Anti-MUC1 aptamer (blue backbone) and MUC1-G peptide (purple backbone and residues) snapshots obtained from a 300ns MD simulation are shown (without water or ions). (A) Initial peptide-aptamer configuration. (B) Initial interaction of the aptamer and the arginine residue of the peptide (circled in black). (C) Peptide-aptamer binding between the open loop region of the aptamer and the arginine residue of the peptide at 25ns after initial binding. (D) Peptide-aptamer conformation at 76ns after binding (E) Peptide-aptamer conformation at the end of the 300ns simulation.

Quantitative Analysis

It was not always clear from the visual analysis if the peptide and aptamer were in a bound configuration or in the physical vicinity of the binding regions. In order to elucidate peptide-aptamer binding we carried out detailed quantitative analysis. The calculated center of mass distance, distance between the atoms within binding region and R_g of the aptamer and peptide complexes are shown in Figure 6. These analytical quantities together help monitor and quantify binding events.

In order to verify the binding event the distance between the peptide and the aptamer was evaluated. The distance between the aptamer and peptide was obtained by considering the center of mass (COM) of the individual biomolecules. The distance between the aptamer and peptide decreases and remain consistently near 2nm in both MUC1 and MUC1-G peptide-aptamer combinations (Figure 6A). In order to further examine aptamer and peptide binding, the distance between the aptamer and peptide atoms that are specific to the binding region were also calculated. The binding region was defined by analyzing the final binding configurations of the peptide-aptamer systems and identifying the atoms specifically and consistently involved in binding. A successful peptide-aptamer binding is known to be non-covalent consisting of hydrogen bonds, Van der Waals forces and electrostatic interactions (48, 49). Van der Waals interactions play significant role in such noncovalent binding, which typically extend to 6Å for large biomolecules. In the present simulation systems, we conservatively used a distance of 4.5Å as the cut-off distance to define non-covalent peptide-aptamer binding (49-51). Accordingly, the peptide–aptamer combination was considered to be bound if the distance between the atoms within the binding region is less than 4.5Å for a minimum of 15ns.

As stated previously, the MUC1 peptide and aptamer combination visually showed the formation of a peptide-aptamer complex (Figure 4D). The simulation trajectory analysis shows that the distance between the binding atoms of the MUC1 peptide–aptamer combination falls below 4.5Å briefly (for less than 1ns) during the simulation (Figure 6B). As the simulation continued this distance increases and remains above 4.5Å indicating that the MUC1 peptide and aptamer molecules were not in a bound configuration. Also, these values were consistently higher than 6Å which is the upper threshold used for large biomolecules (49). Even after the 250ns, there was no binding observed as shown in the final structure of the MUC1 peptide and aptamer (Figure 4G) with the binding distance consistently remaining higher than 6Å (Figure 6B).

Unlike the MUC1 peptide-aptamer combination, simulations of MUC1-G peptide-aptamer combination exhibit a strikingly different binding behavior. The binding distance in this case falls below 4.5Å during the initial association and remains below this threshold for the remainder of the simulation (Figure 6B). The atoms in the binding region of the MUC1-G peptide-aptamer complex averaged a distance of 3.5Å after binding. This clearly indicates that the MUC1-G peptide and aptamer bind and remain in that conformation throughout the simulation.

We also examined the changes in the structural conformation throughout binding in all the cases. Radius of gyration (R_g) of the peptide-aptamer complexes decreased during the MD analysis in both the peptide-aptamer combinations to approximately 2nm (Figure 6C). While both MD analyses had R_g values of 1.4nm for an extended time, the value of R_g in MUC1 peptide-aptamer combination later increased to 1.7nm. This increase in the R_g indicates additional separation between the MUC1 peptide and aptamer molecules.

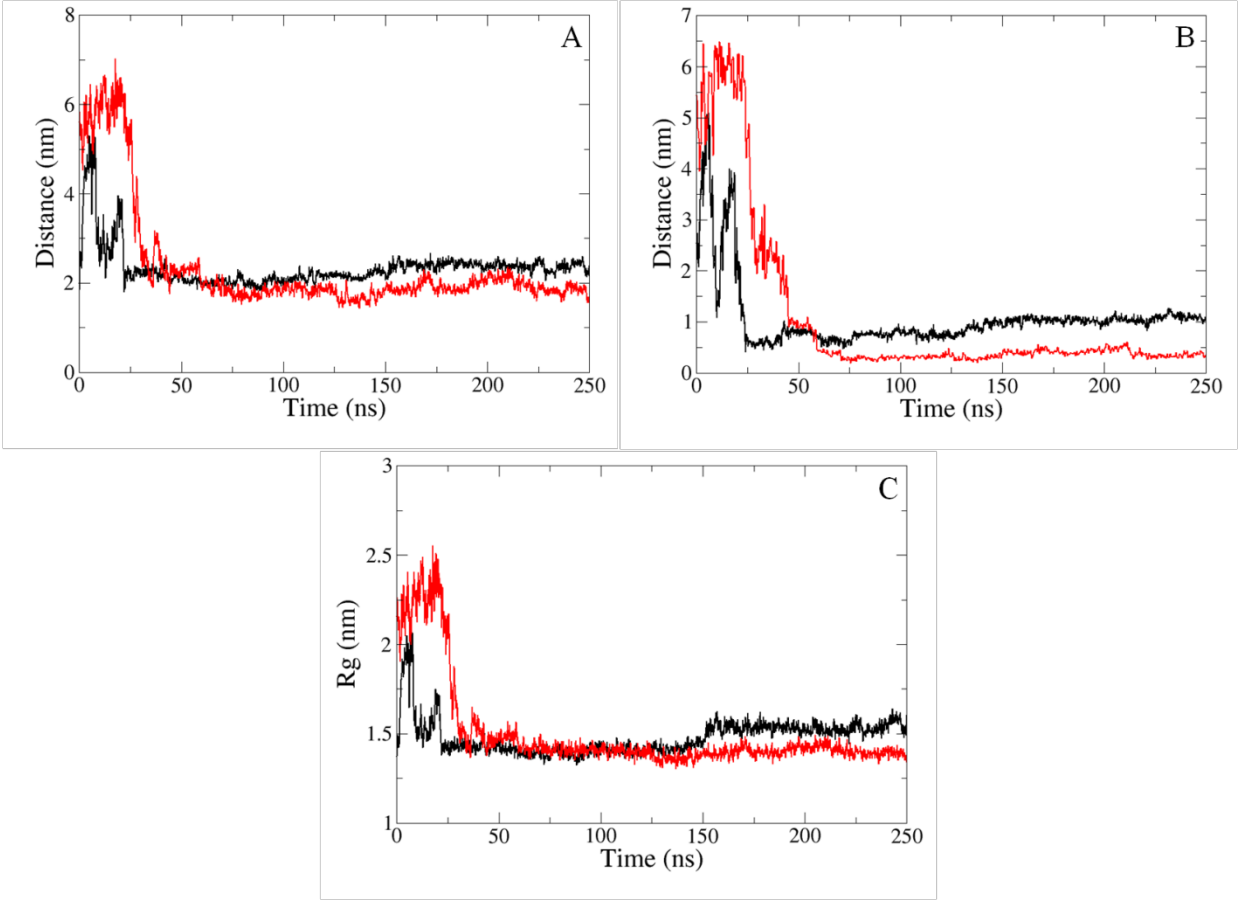


Figure 6: Comparison of anti-MUC1 aptamer binding with MUC1 peptide (black) and MUC1-G peptide (red) (A) Center of mass distance of the aptamer and peptide molecules (B) Distance between the atoms within the binding region of the aptamer and peptides and (C) Rg of the peptide-aptamer complex.

C. Multiple MUC1-G Peptide-Aptamer Simulations

Wet lab experiments with both MUC1 and MUC1-G peptides have shown that the MUC1-G peptide-aptamer combination has a higher binding affinity and is the most favorable binding combination (32). Our simulations and analysis discussed earlier support these findings as binding occurred only with the MUC1-G peptide–aptamer combination. In order to validate the consistency and repeatability of MUC1-G peptide-aptamer binding, we have carried out multiple additional simulations of this combination in solvent. For this purpose, we have constructed 3 additional MUC1-G peptide-aptamer systems under the same simulation conditions detailed in Table 1. This is done by using different starting configurations of MUC1-G peptide-aptamers in solvent. These simulations represented henceforth as MUC1-G_1, MUC1-G_2, and MUC1-G_3. Each of these simulations was carried out for at least 300ns.

Simulations show that these combinations varied in binding time but indicated that the MUC1-G peptide consistently binds with the loop region of the aptamer as before. However, the final binding conformation of the peptide in the loop region differed in each MUC1-G peptide-aptamer combination. Notably, as seen in the MUC1-G peptide-aptamer simulation, the arginine residue of the peptide plays a significant role in these binding combinations. Snapshots of the binding events from each of these simulations are shown in Figure 7. Together these simulation snapshots indicate the consistency and repeatability of the MUC1-G peptide-aptamer binding event. All the simulations indicate that the arginine residue consistently plays an integral role in the binding. The orientation of the peptide differs in each binding event, but the location in the loop region is consistent. The interaction between the arginine residue and the 13^{th} thymine nucleotide is present during the binding in all cases. The remaining peptide residues interact with the 11^{th} and 12^{th} thymine nucleotides in the loop region and surrounding backbone of the aptamer during binding. In one case (MUC1-G_3), in addition to the arginine and 13^{th} thymine nucleotide interaction, the peptide was also found to interact with helical region of the aptamer. Detailed analysis shows that the peptide backbone makes an arching conformation (which was not observed in other simulations) around the loop in this case to create a stable binding conformation (Figure 7D).

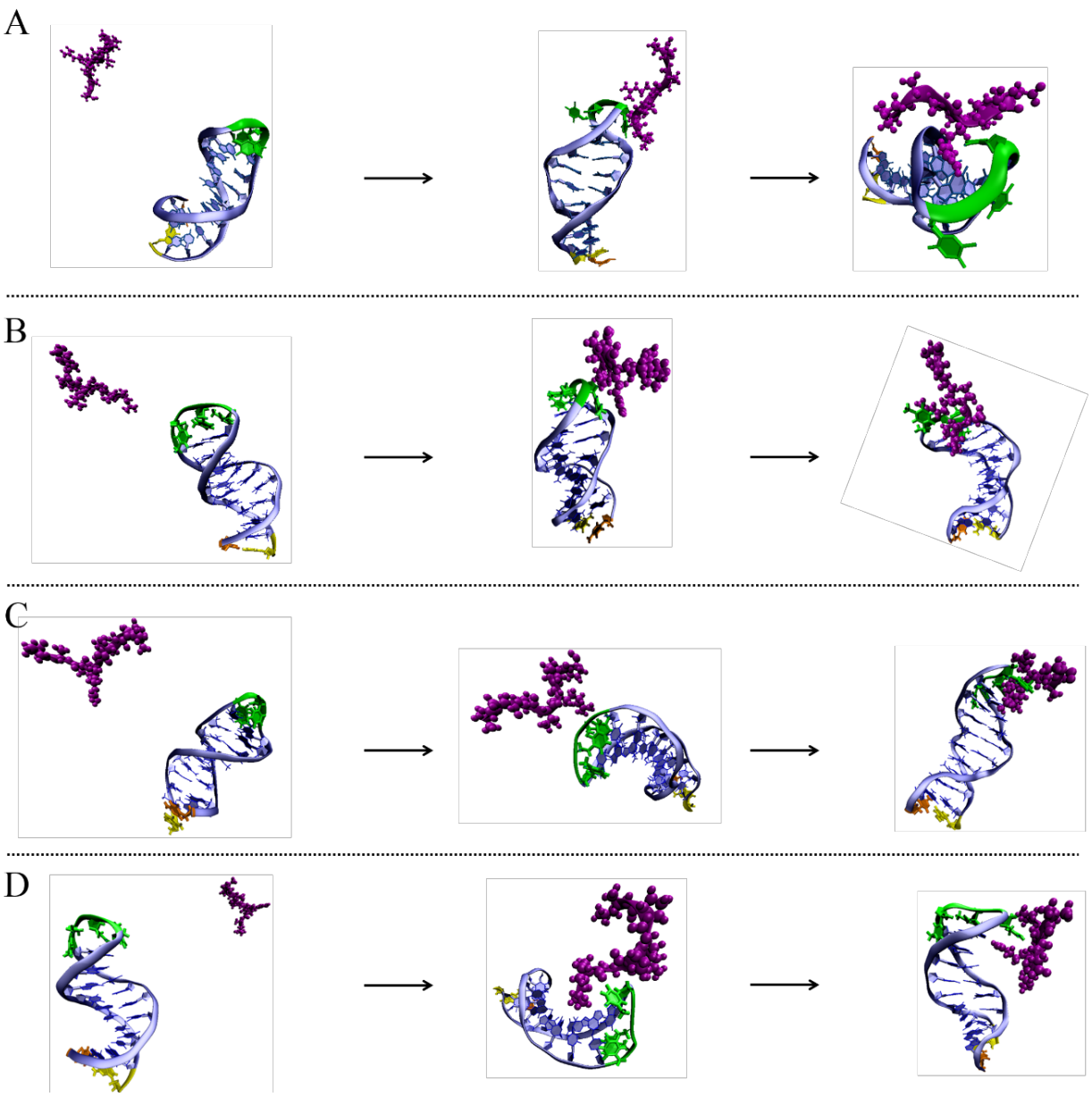


Figure 7: The simulation snapshots of initial, binding and final configurations (shown without water or ions for clarity). (A) MUC1-G (B) MUC1-G_1 (C) MUC1-G_2 and (D) MUC1-G_3.

During the structural analysis, we observed that the R_g for all MUC1-G peptide-aptamer simulations show similar behavior. Figure 8 shows the distribution of the R_g for the initial (blue) and final (red) 1-ns interval of MUC1-G peptide and aptamers from all the simulations. The initial configurations show a wide spread of R_g values. Though the time for the binding event varied in each case, after the binding event all the R_g values converged around 1.4nm as shown

in Figure 8. The convergence of the R_g further substantiates the consistency and repeatability of the MUC1-G binding.

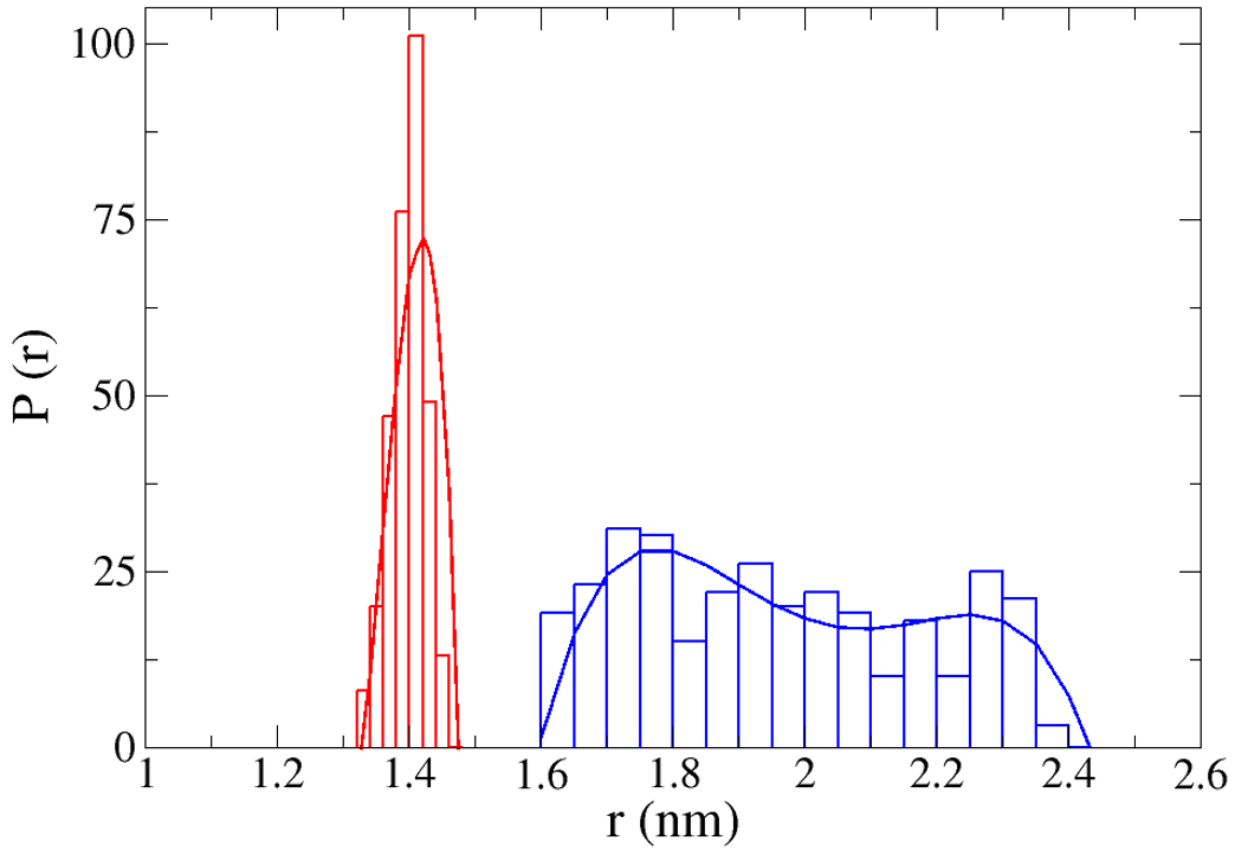


Figure 8: Distribution of the radius of gyration of MUC1-G peptide-aptamer combinations for the initial (blue) and final (red) nanosecond of all MUC1-G peptide and aptamer simulations. Corresponding polynomial fits obtained for the distributions are shown as lines.

We proceed to analyze the binding events based on the distance between the MUC1-G peptide and aptamer molecules. Figure 9 shows the average COM distance and binding atom distance results at key time intervals in the simulation. The average distance values are shown with symbols along with the corresponding variations calculated for the 4 MUC1-G simulations. Both the distance values and variations progressively decrease with the binding event. We find that in all the simulations the COM distance between the MUC1-G peptide and aptamer converged to at least 2nm after the binding. Although each simulation was started with a different initial configuration for peptide-aptamer combination, we observe successful binding in every case. This is also confirmed by calculating the distance between the atoms in the binding region of the peptide-aptamer complex, which converged to a distance below the considered binding threshold of 4.5Å. The average binding atom distance shown in the plot indicates

minimal variation in each simulation in the bound complex. Together these results indicate the consistency and repeatability of the MUC1-G peptide-aptamer binding.

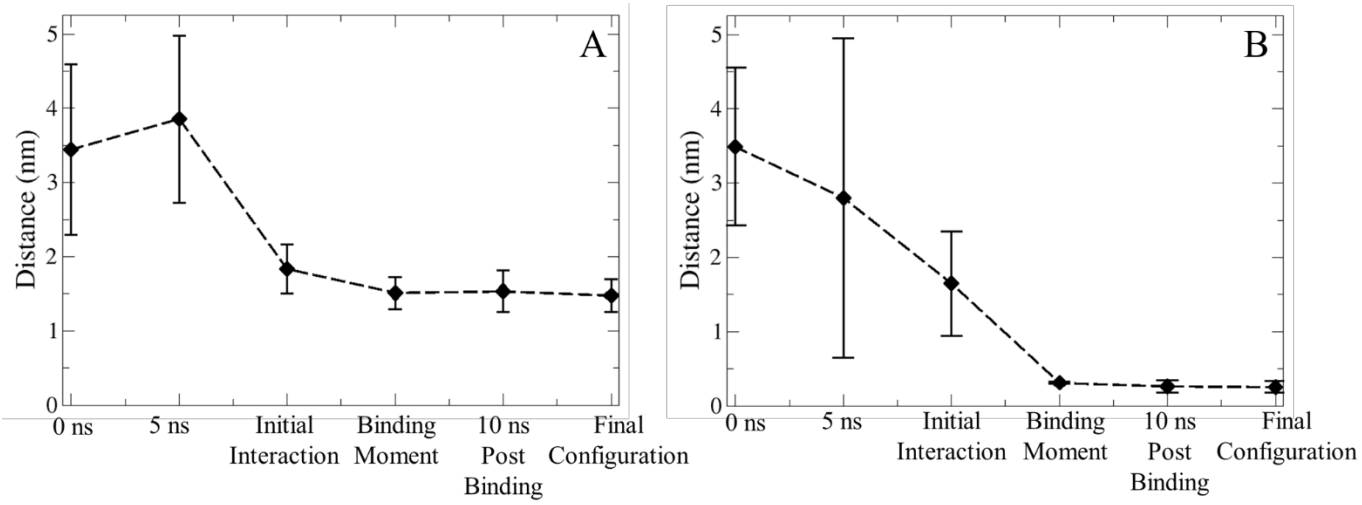


Figure 9: Distance plots at significant times in the binding simulations of multiple MUC1-G peptide-aptamer combinations. The average distance values are shown as symbols along with corresponding variations. A) COM distance and B) Binding atom distance.

In order to assess the overall behavior of the peptide-aptamer binding, we further analyze the distances obtained from multiple binding simulations. Figure 10 shows the distribution of the distances for the initial (blue) and final (red) 1-ns interval of MUC1-G peptide and aptamers from all the simulations. As can be seen from the figure, the distributions corresponding to starting configurations are relatively wide spread. This is due to the fact that the initial configurations of the peptide and aptamer structures were created to be separated by a minimum of 3nm. On the other hand, the distances corresponding to the final configurations show a narrower distribution centered around 2nm with a spread significantly smaller compared to the initial distribution. We further analyze the distribution of the distances between the atoms in the binding region (shown in Figure 9B). The distribution of the binding atoms in the initial 1-ns configurations is similar to the case of COM distances. However, the final nanosecond of simulation show a much narrower distribution centered around 3Å which is well below the binding threshold of 4.5Å considered in this study. Together, these distributions show that the initially well separated MUC1-G peptide and aptamer molecules successfully and consistently bind.

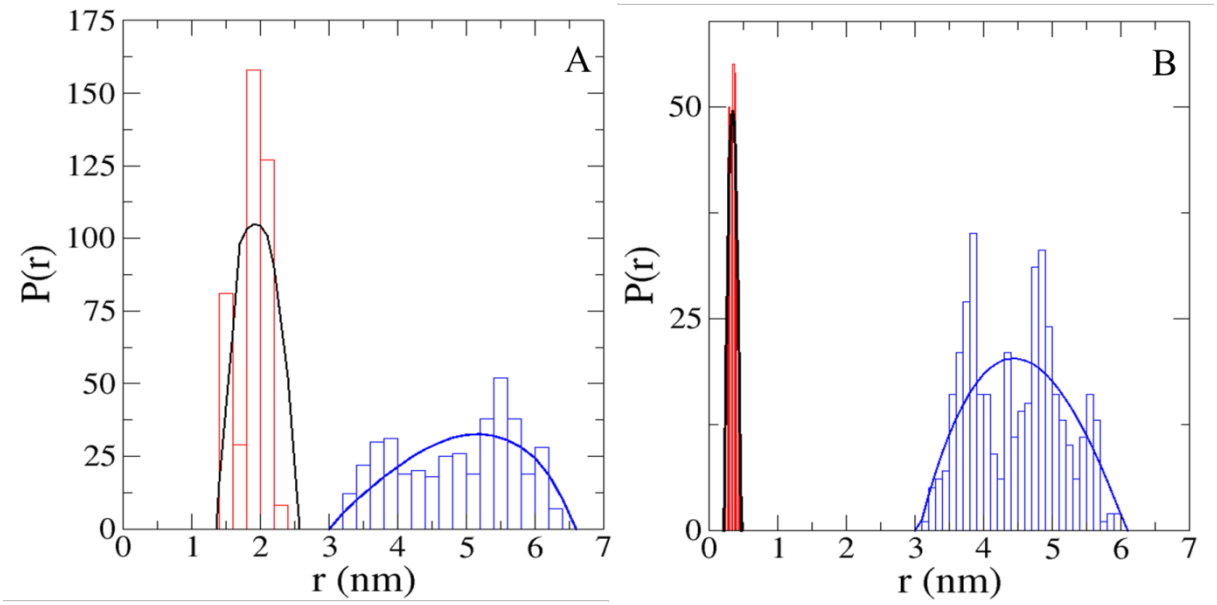


Figure 10: Distance distributions for the initial (blue) and final (red) nanosecond of all MUC1-G simulations. A) Distance distribution between the COM of MUC1-G peptide and aptamer and B) distance between the atoms within the binding region of the MUC1-G peptide and aptamer combination. Polynomial fits obtained for the corresponding distributions are shown as lines.

D. Hydrogen Bonding Analysis

From the simulation snapshots (shown in Figure 7) it is apparent that the aptamer undergoes structural reorganization during the binding event. Hydrogen bond analysis may be able to identify such structural changes (52), since the aptamer tertiary structure is held together by the formation of hydrogen bonds. Figure 11 shows the average number of intramolecular hydrogen bonds for all MUC1-G peptide and aptamer combinations at key intervals. Consistent with the GROMACS analysis, we consider the hydrogen bonds to have a maximum distance of 3.5Å and a corrresonding angle of 30° or less between the donor and acceptor atoms (53). In all cases at the start of the simulation several hydrogen bonds were found to hold the structure together except in the loop and helical regions of the aptamer. As the simulation continues we observe an increase in the number of hydrogen bonds as the system stablizes. The inital interaction with the peptide causes a decrease in the number of hydrogen bonds in the aptamer. The binding event triggered by the arginine residue causes the distruption of the hydrogen bonds in the open ends of the aptamer structure. This is evident from the reduction in the average number of hydrogen bonds, as shown in Figure 11. However, after the binding the number of hydrogen bonds gradually increases as the aptamer structure reorganizes to a stable confirmation. Towards the end of the simulations the variation within the number of hydrogen bonds in the aptamer

structure was found to be minimal. This indicated that the final binding combination in all the simulations obtains a similar compact binding structure.

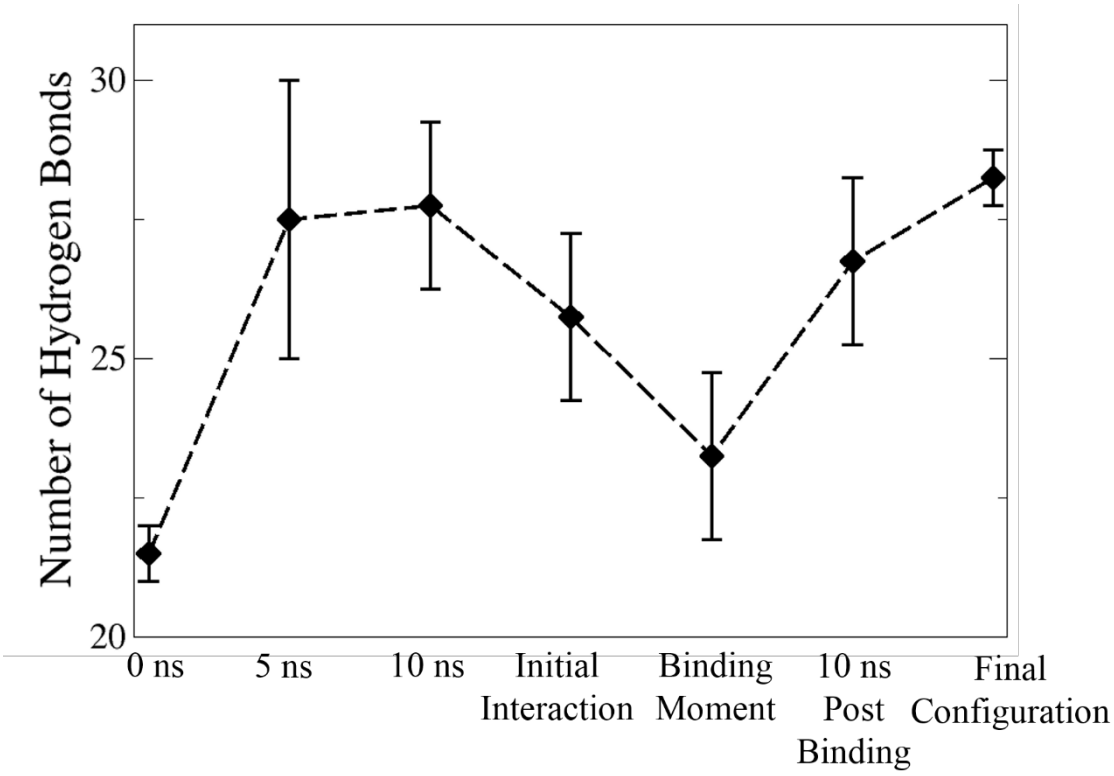


Figure 11: Average number of hydrogen bonds at significant times observed in the multiple peptide-aptamer binding simulations. The average numbers of hydrogen bonds are shown as symbols along with corresponding variations.

Discussion

Simulations of individual molecules established the stability of aptamer and peptide molecules in the solvent. Simulations with MUC1 peptide and the anti-MUC1 aptamer showed temporary association of these molecules. In particular, we observed that this association occurs at the 5' and 3' end region. However, longer time duration analysis revealed peptide reorientation, followed by subsequent separation of these molecules. This weak association between the MUC1 peptide and aptamer was due to the increased flexibility of the peptide backbone as indicated by the structural analysis. Binding atom distance analysis during the brief association fell below 4.5Å. However this distance continued to increase to average values greater than 6Å indicating that the MUC1 peptide and aptamer are not in a bound state.

Multiple simulations showed that the MUC1-G combination consistently interacts with the loop region of the aptamer and binds with that specific region. The arginine residue of the peptide initiated the interaction with the aptamer. Throughout binding the arginine residue consistently interacted with the 13th thymine nucleotide of the aptamer in all MUC1-G peptide–aptamer simulations. In each simulation the binding successfully occurred, with the binding time that differed in each case. Binding was further confirmed by the calculating the distance between the atoms in the binding region for all MUC1-G combinations studied. This distance was found to remain consistently lower than 4.5Å after binding which implies that the MUC1-G peptide and aptamer are non-covalently bound.

Detailed structural analysis of the multiple MUC1-G peptide-aptamer simulations showed that the aptamer structure is altered during the binding process. This was evident from the changes in the number of hydrogen bonds within the aptamer as it interacted with the peptide. In particular, the hydrogen bonds within the aptamer structure showed distinct changes during its binding with the MUC1-G peptide. Upon binding with the peptide, the aptamer internal structure is significantly altered and the number of hydrogen bonds decreased as the hydrogen bonds holding the helix and open ends together are disrupted. As the binding continued, the aptamer tries to reestablish its hydrogen bonds by forming a stable peptide-aptamer binding complex.

A previous wet lab experimental study by Ferreira *et al* indicated a strong binding of anti-MUC1 aptamer and MUC1-G peptide (32). By using Surface Plasmon Resonance imaging (SPRi) they found that the combination of MUC1-G peptide and anti-MUC1 aptamer showed a relatively faster association compared to other mucin 1 peptides. The MUC1-G peptide-aptamer complex was shown to be a stronger binding combination with a low dissociation constant. In their work (32), alternative combinations of MUC1 peptide and anti-MUC1 aptamer were also studied and were consistently shown to be weaker binding combinations compared to the MUC1-G peptide and aptamer. This concurs well with our multiple simulation study results that consistently showed a stronger binding between the MUC1-G peptide and anti-MUC1 aptamer.

The present study showed that the computational modeling can be effectively used to determine the binding behavior, location and orientation for peptide-aptamer binding combinations. Such findings have important implications for aptamer based biosensor development. In practice, the aptamers will be immobilized on the biosensor surface (2, 54, 55). To optimize sensor efficiency there is a need to immobilize the aptamers in a way that maximizes access to their binding region. Aptamers and their target ligands (like the MUC1-G peptide) can have multiple binding regions. By identifying the binding region and the aptamer orientation within the binding complex, it is possible to immobilize the aptamer within the biosensor in a manner that will provide greater accessibility to the binding region (2, 29). This in turn aids in efficient biomarker detection within a biosensor. The insights such as the binding conformation, binding location and orientation of the peptide-aptamer combination obtained from simulations could potentially aid in accelerating the development of biosensors.

Concluding Remarks

Computational modeling and molecular dynamics simulations could provide effective means to study and understand complex biological phenomenon such as biomarker-aptamer binding combinations. In this work extensive molecular dynamics simulations of the anti-MUC1 aptamer with MUC1 peptide and multiple MUC1-G peptide binding combinations were carried out using GROMACS. MUC1-G peptide, a variant of MUC1 peptide was generated by adding a glycine residue. After exploring the structure and dynamics of individual molecules, aptamer binding with MUC1 peptide and MUC1-G peptide were simulated separately. Multiple simulation results showed that the common peptide sequence ADPTRPAP, in particular the arginine residue plays an important role in facilitating binding. The addition of residues like glycine, in the MUC1-G peptide, help maintain binding to the anti-MUC1 aptamer. Binding events were monitored by calculating the distance between peptide and aptamer molecules. We find that the distance between the atoms within the binding region decreased due to binding. The formation of the binding complex was further supported by the calculation of the number of hydrogen bonds within the aptamer structure. Multiple simulations demonstrated that the MUC1-G binding event occurs consistently in the loop region of the aptamer. Further, simulations reveal that the arginine residue of the peptide played a key role in peptide-aptamer binding.

The simulation results of the present study demonstrate minute changes in peptide structure can impact molecular binding. Simulations also reveal a natural progression of the entire aptamer and peptide binding event in solvent. Insights from this study could aid in biosensor developments by providing pointers on specificities of binding such as aptamer orientation and location that are critical to enhance detection mechanisms.

The methods used in this study can be adapted to analyze aptamer-peptide combinations. The insights obtained from this study may also be used to simplify the complex and involved SELEX process. However, this involves studying a substantial number of biomarker-aptamer combinations which would be near impossible to study with the computational methods based on atomistic details. For this purpose, we intend to develop intermediate coarse-grain models, which are known to increase the simulation efficiency by decreasing the number of degrees of freedom (56, 57). Further work in this direction is under progress.

Acknowledgments

This work was supported in part by the U. S. Army Research Office via award/contract no. W911NF-11-1-0168. We thank Dr. M. Sandros for scientific discussions during the course of

this work. We would like to acknowledge the use of high performance computational facilities at the North Carolina State A&T University during the course of the present study.

References

1. Jain, K. K. 2010. The handbook of biomarkers. Springer, New York.
2. Wan, Y., Y. Su, X. Zhu, G. Liu, and C. Fan. 2013. Development of electrochemical immunosensors towards point of care diagnostics. Biosens Bioelectron 47:1-11.
3. Scognamiglio, V. 2013. Nanotechnology in glucose monitoring: Advances and challenges in the last 10 years. Biosensors and Bioelectronics 47:12-25.
4. Chang, K., Y. Pi, W. Lu, F. Wang, F. Pan, F. Li, S. Jia, J. Shi, S. Deng, and M. Chen. 2014. Label-free and high-sensitive detection of human breast cancer cells by aptamer-based leaky surface acoustic wave biosensor array. Biosensors and Bioelectronics 60:318-324.
5. Ma, F., C. Ho, A. K. H. Cheng, and H.-Z. Yu. 2013. Immobilization of redox-labeled hairpin DNA aptamers on gold: Electrochemical quantitation of epithelial tumor marker mucin 1. Electrochimica Acta 110:139-145.
6. Baruch, A., M. Hartmann, M. Yoeli, and e. al. 1999. The Breast Cancer-associated MUC1 Gene Generates Both a Receptor and Its Cognate Binding Protein. Cancer Research 59:1552-1561.
7. Gendler, S. J. 2001. MUC1, The Renaissance Molecule. Journal of Mammary Gland Biology and Neoplasia 6:339-353.
8. Dokurno, P., P. A. Bates, H. A. Band, L. M. D. Stewart, J. M. Lally, J. M. Burchell, J. Taylor-Papadimitriou, D. Snary, M. J. E. Sternberg, and P. S. Freemont. 1998. Crystal structure at 1.95 å resolution of the breast tumour-specific antibody SM3 complexed with its peptide epitope reveals novel hypervariable loop recognition. Journal of Molecular Biology 284:713-728.
9. Taylor-Papadimitriou, J., J. M. Burchell, T. Plunkett, R. Graham, I. Correa, D. Miles, and M. Smith. 2002. MUC1 and the Immunobiology of Cancer. Journal of Mammary Gland Biology and Neoplasia 7:209-221.
10. Kirnarsky, L., O. Prakash, S. M. Vogen, M. Nomoto, M. A. Hollingsworth, and S. Sherman. 2000. Structural Effects of O-Glycosylation on a 15-Residue Peptide from the Mucin (MUC1) Core Protein. Biochemistry 39:12076-12082.
11. von Mensdorff-Pouilly, S., M. M. Gourevitch, P. Kenemans, A. A. Verstraeten, G. J. van Kamp, A. Kok, K. van Uffelen, F. G. M. Snijdewint, M. A. Paul, S. Meijer, and J. Hilgers. 1998. An Enzyme-Linked Immunosorbent Assay for the Measurement of Circulating Antibodies to Polymorphic Epithelial Mucin (MUC1). Tumor Biology 19:186-195.

12. Pichinuk, E., I. Benhar, O. Jacobi, M. Chalik, L. Weiss, R. Ziv, C. Sympson, A. Karwa, N. I. Smorodinsky, D. B. Rubinstein, and D. H. Wreschner. 2012. Antibody targeting of cell-bound MUC1 SEA domain kills tumor cells. Cancer Research.
13. Hamanaka, Y., Y. Suehiro, M. Fukui, K. Shikichi, K. Imai, and Y. Hinoda. 2003. Circulating anti-MUC1 IgG antibodies as a favorable prognostic factor for pancreatic cancer. International Journal of Cancer 103:97-100.
14. Strehlitz, B., N. Nikolaus, and R. Stoltenburg. 2008. Protein Detection with Aptamer Biosensors. Sensors 8:4296-4307.
15. Erickson, D., S. Mandal, A. Yang, and B. Cordovez. 2008. Nanobiosensors: optofluidic, electrical and mechanical approaches to biomolecular detection at the nanoscale. Microfluidics and Nanofluidics 4:33-52.
16. Han, K., Z. Liang, and N. Zhou. 2010. Design strategies for aptamer-based biosensors. Sensors 10:4541-4557.
17. Wang, J. 2000. From DNA biosensors to gene chips. Nucleic Acid Research 28:3011-3016.
18. Song, S., L. Wang, J. Li, C. Fan, and J. Zhao. 2008. Aptamer-based biosensors. TrAC Trends in Analytical Chemistry 27:108-117.
19. McCauley, T. G., N. Hamaguchi, and M. Stanton. 2003. Aptamer-based biosensor arrays for detection and quantification of biological macromolecules. Analytical Biochemistry 319:244-250.
20. He, P., V. Oncescu, S. Lee, I. Choi, and D. Erickson. 2013. Label-free electrochemical monitoring of vasopressin in aptamer-based microfluidic biosensors. Analytica Chimica Acta 759:74-80.
21. Palchetti, I., and M. Mascini. 2012. Aptamer-based Biosensors for Cancer Studies. Biosensors and Cancer:85.
22. Clark, S. L., and V. T. Remcho. 2002. Aptamers as analytical reagents. ELECTROPHORESIS 23:1335-1340.
23. Stoltenburg, R., C. Reinemann, and B. Strehlitz. 2007. SELEX--a (r)evolutionary method to generate high-affinity nucleic acid ligands. Biomolecular engineering 24:381-403.
24. Tuerk, C., and L. Gold. 1990. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science 249:505-510.
25. Kolmodin, L. A., and D. E. Birch. 2002. Polymerase chain reaction. Basic principles and routine practice. Methods Mol Biol 192:3-18.
26. Bartlett, J. S., and D. Stirling. 2003. A Short History of the Polymerase Chain Reaction. In PCR Protocols. J. S. Bartlett, and D. Stirling, editors. Humana Press. 3-6.
27. Saiki, R., D. Gelfand, S. Stoffel, S. Scharf, R. Higuchi, G. Horn, K. Mullis, and H. Erlich. 1988. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science 239:487-491.

28. Chai, C., Z. Xie, and E. Grotewold. 2011. SELEX (Systematic Evolution of Ligands by EXponential Enrichment), as a powerful tool for deciphering the protein-DNA interaction space. Methods Mol Biol 754:249-258.
29. Ziegler, C., and W. Göpel. 1998. Biosensor development. Current Opinion in Chemical Biology 2:585-591.
30. Hosseini, S., F. Ibrahim, I. Djordjevic, and L. H. Koole. 2014. Recent advances in surface functionalization techniques on polymethacrylate materials for optical biosensor applications. Analyst 139:2933-2943.
31. Chen, H., J. Huang, J. Lee, S. Hwang, and K. Koh. 2010. Surface plasmon resonance spectroscopic characterization of antibody orientation and activity on the calixarene monolayer. Sensors and Actuators B: Chemical 147:548-553.
32. Ferreira, C. S., C. S. Matthews, and S. Missailidis. 2006. DNA aptamers that bind to MUC1 tumour marker: design and characterization of MUC1-binding single-stranded DNA aptamers. Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine 27:289-301.
33. Nguyen, Thuy Hien T., Z. Liu, and Preston B. Moore. 2013. Molecular Dynamics Simulations of Homo-oligomeric Bundles Embedded Within a Lipid Bilayer. Biophysical Journal 105:1569-1580.
34. Kim, S., J. Lee, S. J. Lee, and H. J. Lee. 2010. Ultra-sensitive detection of IgE using biofunctionalized nanoparticle-enhanced SPR. Talanta 81:1755-1759.
35. Ozsoz, M. S. 2012. Electrochemical DNA Biosensors. Pan Stanford.
36. Cheng, X., and I. Ivanov. 2012. Molecular Dynamics. In Computational Toxicology. B. Reisfeld, and A. N. Mayeno, editors. Humana Press. 243-285.
37. Ivanov, I., B. R. Chapados, J. A. McCammon, and J. A. Tainer. 2006. Proliferating cell nuclear antigen loaded onto double-stranded DNA: dynamics, minor groove interactions and functional implications. Nucleic Acids Research 34:6023-6033.
38. Palermo, G., U. Rothlisberger, A. Cavalli, and M. De Vivo. Computational insights into function and inhibition of fatty acid amide hydrolase. European Journal of Medicinal Chemistry.
39. Rhinehardt, K., R. Mohan, G. Srinivas, and A. Kelkar. 2013. Computational Modeling of Peptide - Aptamer Binding in Biosensor Applications. International Journal of Bioscience, Biochemistry and Bioinformatics 3:639-642.
40. Schrödinger, L. The PyMOL Molecular Graphics System.
41. Van Der Spoel, D., E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen. 2005. GROMACS: Fast, flexible, and free. Journal of Computational Chemistry 26:1701-1718.
42. Berendsen, H. J. C., D. van der Spoel, and R. van Drunen. 1995. GROMACS: A message-passing parallel molecular dynamics implementation. Computer Physics Communications 91:43-56.

43. Hess, B., C. Kutzner, D. van der Spoel, and E. Lindahl. 2008. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. Journal of Chemical Theory and Computation 4:435-447.
44. Nosé, S., and M. L. Klein. 1983. Constant pressure molecular dynamics for molecular systems. Molecular Physics 50:1055-1076.
45. Humphrey, W., A. Dalke, and K. Schulten. 1996. VMD: Visual Molecular Dynamics. Journal of Molecular Graphics 14:33-38.
46. Carugo, O. 2003. How root-mean-square distance (r.m.s.d.) values depend on the resolution of protein structures that are compared. Journal of Applied Crystallography 36:125-128.
47. de Gennes, P.-G. 1979. Scaling concepts in polymer physics. Cornel University Press.
48. Berg, J. M., F. H. Deis, J. L. Tymoczko, L. Stryer, N. C. Gerber, R. Gumport, and R. E. Koeppe. 2011. Biochemistry Student Companion. W. H. Freeman.
49. Farrell, S., and L. Taylor. 2005. Experiments in biochemistry: A hands-on approach. Cengage Learning.
50. Schalley, C. A. 2012. Analytical Methods in Supramolecular Chemistry: Vol. 1. Wiley-VCH Verlag GmbH & Company KGaA.
51. Bondi, A. 1964. van der Waals volumes and radii. The Journal of Physical Chemistry 68:441-451.
52. Pal, S., P. K. Maiti, and B. Bagchi. 2006. Exploring DNA groove water dynamics through hydrogen bond lifetime and orientational relaxation. J Chem Phys 125:234903.
53. D. van der Spoel, E. L., B. Hess, A. R. van Buuren, E. Apol, P. J. Meulenhoff,, A. L. T. M. S. D. P. Tieleman, K. A. Feenstra, R. van Drunen and H. J. C., and Berendsen. 2010. Gromacs User Manual version 4.5.4. www.gromacs.org.
54. Vance, S. A., and M. G. Sandros. 2014. Zeptomole Detection of C-Reactive Protein in Serum by a Nanoparticle Amplified Surface Plasmon Resonance Imaging Aptasensor. Sci. Rep. 4.
55. Pilolli, R., L. Monaci, and A. Visconti. 2013. Advances in biosensor development based on integrating nanotechnology and applied to food-allergen management. TrAC Trends in Analytical Chemistry 47:12-26.
56. Marrink, S. J., H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. de Vries. 2007. The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations. The Journal of Physical Chemistry B 111:7812-7824.
57. Sim, A. Y. L., P. Minary, and M. Levitt. 2012. Modeling nucleic acids. Current Opinion in Structural Biology 22:273-278.

2. High Performance and Multi-Scale Computational Modeling in Bio Systems

G. Srinivas, R. Mohan, A. Kelkar, North Carolina A&T State University

Introduction

Computational modeling and simulation play integral role in the scientific and technological advances. In particular, its application in biomedical research and nanotechnology revolutionized drug discovery. A fundamental assumption for computational modeling and simulation is that insight into system behavior can be developed or enhanced from a model that adequately represents a selected subset of the system's attributes. There is a wide variety of simulation techniques are available. Depending on properties and applicability, they can be broadly categorized into three classes: 1. Quantum simulations 2. Classical simulations, and 3. Mesoscopic simulations. Each method has its merits and limitations as they have been designed to efficiently handle different spatial and temporal scales. For example, for studying chemical bond formation and bond dissociation between atoms and molecules involving electronic interactions, a simulation method based on quantum mechanics is an appropriate choice. However, due to the complex interactions and computationally intense numerical calculations, this method can only be applicable for systems with relatively smaller length and time scales. Nevertheless, many interesting phenomena relevant to biological and material sciences involves time and length scales over micrometers and microseconds and beyond. Such problems cannot be directly dealt with methods based on quantum mechanics. On the other hand, classical simulations based on Newtonian mechanics, are highly suitable to deal with such large time and length scales appropriately. For this very reason, classical molecular dynamics (MD) is the most commonly used simulation method. The classical MD approach is routinely used to explore tens of nanosecond times and nanometer length scales. However, such efficiency over quantum simulations comes with a penalty of relatively lower accuracy. Naturally, more efficient methods such as mescoscopic simulations that explore time and length scales beyond classical/atomistic simulation capabilities are relatively less accurate. Bridging these desperate spatial and temporal scales is essential to explore complex biomedical engineering problems in detail.

Bridging Experiments and Simulations through High Performance Computing

Experimental work on complex biomedical and material systems spans a broad range of temporal and spatial scales, from femtosecond dynamics and atomistic detail to real-time macroscopic phenomena. Simulation methods in which each atom is explicitly represented are well established but have difficulty addressing many cooperative effects of experimental and theoretical interest. There is simply too large a gap between the timescale and spatial scale that govern typical intramolecular events and those which are relevant for collective motions. Often, the timescale gap between simulation and experiment is about six orders of magnitude [1].

Existing simulation techniques for specific timescales and spatial scales are illustrated schematically in figure 1. These techniques take a variety of approaches to reduce the level of detail in the representation of the system under study as the timescale and/or length scale grows. Bridging these disparate scales is possible with multiscale modeling [2] in which the various levels of treatment are coupled and fed back into one another. Reduced models which retain close connections to the underlying atomistic representation have been promising. Researchers working at the bio-nano interface aim to study events occurring on timescales of hundreds of nanoseconds to milliseconds and spatial scales of microns. Here we focus on the development and application of a coarse grain simulation method that has ready access to events on these scales; coarse grain models are gaining widespread usage in the material [3] and biophysical communities [4].

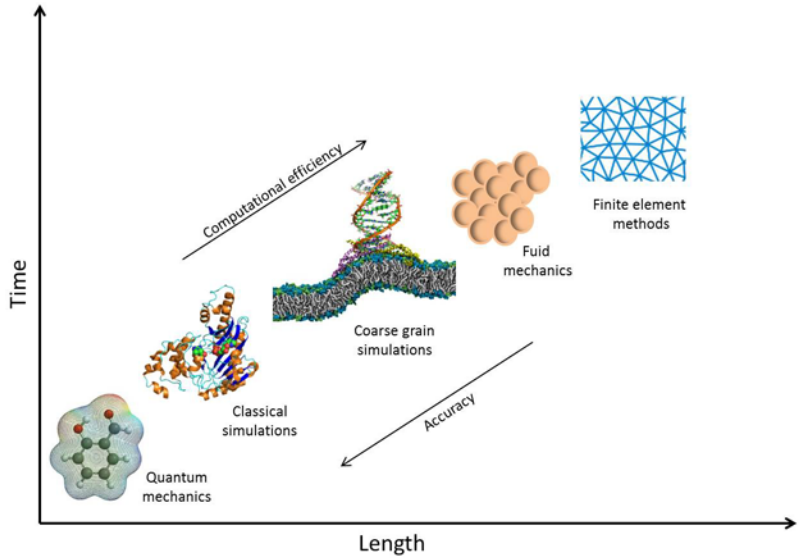


Figure 1: Various simulation techniques are depicted along with the representative systems that can be explored efficiently. Note the decrease in computational accuracy as the efficiency increase.

Challenges for Computational Modeling

There are many phenomena that lie within the mesoscopic spatio-temporal scale which might be explored with coarse grain (CG) methods. Examples of such phenomena from biology include protein–protein interactions, protein-lipid interactions, and membrane-membrane interactions. From a materials perspective, the optimal design of nanosyringes which penetrate membranes is of interest, as well as the design and properties of artificial polymer based membranes which can act as controlled release vesicles for drug delivery, which is the main focus of our current research project. Note that from a computational view point there exists lot of commonality between material and biological systems. In both the cases, underlying systems are still made of

atoms and molecules that are connected through chemical bonds. Broadly speaking, basic atomic and molecular structure and function determines large scale macroscopic properties in a given system. Note that the computational description of either system involves similar set of physical principles and mathematical equations. Moreover, both systems span similar spatial and length scales. Hence, we proceed to develop coarse-grain models for both systems using similar procedure.

Typical biological processes occur on mesoscopic level. An understanding of such phenomena from atomic level and connecting it to the corresponding mesoscopic properties is currently lacking. Experimental studies over the years have shown that vesicles similar to biological vesicles can be formed from organic superamphiphiles such as block copolymers [5]; diblock copolymers in particular have architecture similar to that of natural lipids. Many biological membrane processes such as protein integration, fusion, DNA encapsulation, and compatibility can be reliably mimicked by synthetic polymer vesicles. The overall copolymer molecular weight is considerably larger (3–20 kDa) than that of their natural lipid membrane counterparts (<1 kDa) [6]. In addition, the hydrophilic/hydrophobic ratio can be selected with ease. Block copolymers have the intrinsic ability to self-organize into membranes and offer fundamental insight into natural design principles for biomembranes. Unusually large system sizes (consisting typically of $>10^6$ atoms) makes this task computationally expensive. While current computational resources allow study of static components of such large systems, it becomes impractical to study dynamical phenomena such as polymer assisted drug delivery, which occurs typically on a multi-nanosecond to microsecond timescale. Existing simulation studies of simple prototype systems have been mostly carried out using arbitrary potentials. For example dissipative particle dynamics (DPD) and discontinuous molecular dynamics (DMD) have been used to study the self-assembly of block copolymers by micro-phase separation [7]. The coarse grain approach presented in this work has proven to be effective and reliable. In particular, a quantitative comparison with corresponding experiments is possible which in turn provides microscopic insights into the corresponding experimental system. In the following we explore interesting phenomena that occur at bio-nano interface. In particular, we study: 1. polymeric micellar nanocarriers for drug delivery, 2. Incorporation and aggregation of nanoparticles within polymer micelles. We aim to explore these topics using coarse-grain molecular dynamics simulations as described below.

2.1 Polymer Micelle Assisted Transport and Delivery of Model Hydrophilic Components inside a Biological Lipid Vesicle: A Coarse Grain Simulation Study

G. Srinivas, R. Mohan, A. Kelkar, Journal of Physical Chemistry B 2013
DOI: 10.1021/jp405381k.

ABSTRACT

Understanding drug transportation and delivery mechanism from a molecular viewpoint is essential to find better treatment pathways. Despite the fact many significant drugs such as anticancer doxorubicin and mitoxantrone are predominantly hydrophilic, efficient methodology to deliver hydrophilic drug components is not well established. Here we explore this problem by studying "patchy" polymeric micelle assisted hydrophilic component transportation across a lipid membrane and delivery inside a biological lipid vesicle. Using the MARTINI force field as the basis, we study the interaction of polymeric micelle with POPC lipid vesicles in detail. In order to facilitate hydrophilic drug transportation study, a primitive CG model for hydrophilic drug component is used. Extensive simulations carried out over hundreds of nanoseconds demonstrate successful encapsulation, transportation of hydrophilic components by patchy polymeric micelle. Results show the polymeric micelle releases significant portion of hydrophilic contents inside the lipid vesicle. Present simulation study also reveals a possible mechanism for efficient hydrophilic component transportation and delivery. Insights from this study could potentially help experimental community to design better delivery vehicles, especially for hydrophilic drug molecules.

KEYWORDS: drug delivery, copolymers, self-assembly, coarse grain simulations, lipid vesicle

I. INTRODUCTION

Efficient drug delivery is equally important as drug development[1-3]. Transporting drug components across cell membrane and delivering at the targeted location is critical for effective drug function. Many pharmaceutical and therapeutic studies have explored this problem with an aim to achieve efficient drug delivery across cell membrane[1-8]. Numerous experimental studies in the past have shown that the bare drug molecules cross complex cell membrane structure with relatively low efficiency[2,3].

Rapid growth in nanotechnology combined with biological applications is revolutionizing disease detection in addition to improving treatment methodologies. Recent studies suggest new pathways for drug delivery that utilize nanomaterials as supporting or embedding materials for

drug components[9-15]. For example, nanoparticle assisted drug transportation across cell membrane resulted in higher percentage of drug incorporation into intracellular domain[3,12-14]. Flexible materials such as self-assembled polymer micelles have also been shown to be promising candidates for not only efficient drug transportation but also for targeted delivery[16-18]. This method has several advantages over other nanomaterial based approaches. For example, the drug components can be incorporated into polymer micelle core, thereby protecting them from external environment[11,12]. In addition, polymers can be designed/selected so as to interact efficiently with cell membrane[15]. During the polymeric micelle interaction with cell membrane, the drug content from micelle core gets released inside the cell membrane, which eventually reaches intracellular targeted domains[10,15]. Clearly, the nature of polymer and its interaction with cell membrane plays crucial role in this case. For the efficient drug transportation and delivery, ideally, a polymer micelle should effectively incorporate the drug components, as well as transport across cell membrane and release them at the targeted site. In order to design such efficient polymer micelles, it is important to understand how polymer micelle interacts with cell membrane from a molecular view point. Despite numerous experimental studies in this area, efficient encapsulation and transportation of hydrophilic drug molecules such as anticancer doxorubicin and mitoxantrone is still a challenging task[10]. A deeper understanding of mechanism detailing how polymer micelle transports drug components across the cell membrane might help understand and address such questions.

Due to their size and structure similar to that of natural drug carriers such as liposomes and lipoproteins, polymeric micelles received ample attention for drug transportation and delivery studies[16]. Polymeric micelles also have an added advantage of long blood circulation times of micellar particles. Despite such advantages, polymeric micelles pose a challenge for the transportation of hydrophilic drug molecules. Typically, polymeric micelles used in such studies have a specific core-shell structure: hydrophobic inner core protected from the outer environment by hydrophilic shells[17]. Hence, hydrophobic drug molecules can be naturally incorporated within the hydrophobic inner core. The hydrophilic shell structure shields such incorporated drug components from the outer environment. Nevertheless, such hydrophobic core-hydrophilic shell structure makes it difficult to incorporate hydrophilic drug components inside polymeric micelle[18,19]. As a result, polymeric micelle assisted hydrophilic drug transportation remains a challenging task.

Such difficulties and shortcomings motivated computational modeling studies of drug transportation and delivery. However, limitations of computational efficiency in dealing with problems involving such large length and time scales often restricted the scope of such studies. For example, simulation studies typically consider a simple lipid bilayer in order to represent complex cell membrane structure. Such simple lipid bilayer models proved successful in studying various aspects of lipid membranes including membrane self-assembly, membrane rupture and stability etc. [20,21]. Even with the advent of present computational resources, it is

nearly impossible to simulate a complex cell membrane at atomistic detail over relevant time and length scales. In order to overcome such limitations, lower resolution methods such as coarse-grain (CG) simulation methods have been introduced[22-38]. In this approach, typically, a group of atoms is considered as single site instead of every atom in the system. As a result, the size of the system reduces at least ten times to that of the original system, resulting in a higher computational efficiency. Recent studies used CG approach to explore complex interdisciplinary problems from chemistry, biology and materials science. In an effort towards building a systematic CG approach, Marrink *et al.* developed the "MARTINI" coarse-grain force field to study biological lipid membranes and vesicles[23-28]. This force field has been used with some success in studies of protein folding problem as well. Klein *et al*. used a similar but different CG approach to study lipid membranes and polymer membranes and their self-assembly3[2-38]. They have also studied interaction of polymer bilayers and micelles with biological lipid membranes. In particular, they have studied the interaction of diblock copolymer micelle with a standalone biological lipid bilayer[39]. Recently, the Klein force field has been used to study the interaction of hydrophobic drugs with polymer micelles by conducting rational CG-MD simulations[40]. By carrying out extensive simulation studies, they showed that worm-like nanocarriers increase the amount of drug delivered to tumors thereby underlining the importance of nanocarrier shape in drug delivery[40].

In this work, we use CG approach to explore polymer micelle interaction with a lipid vesicle in water. In order to model polymer assisted drug transportation across cell membrane, we considered a 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine (POPC) lipid vesicle to represent a relatively simple cell membrane environment. In order to explore the drug transportation and delivery, we study the interaction of a binary diblock copolymer micelle with POPC lipid vesicle. The copolymer micelle used in this study was assembled from a binary mixture of diblock copolymers. The hydrophobic block was modeled to represent a simple alkane chain while the hydrophilic block model was based on polyethylene glycol (PEG). The binary diblock copolymer micelle has an added advantage for the transportation of hydrophilic molecules, as described below. In a subsequent analysis, in order to mimic the hydrophilic drug molecules, model hydrophilic components (HP) were incorporated inside polymer micelle at the beginning of the simulation. During the course of simulation, such loaded polymer micelle interacts with lipid vesicle, thereby establishing possible pathways for transporting hydrophilic contents across lipid membrane.

In the next Section, we describe CG simulation details, system setup and methods. Section III presents simulation results from polymer micelle interaction with lipid vesicle. Incorporation of model drug like molecules in polymer micelle and their transportation across lipid bilayer are described within the same section. We close the paper with few concluding remarks in Section IV.

II. SIMULATION DETAILS

Computer simulation studies at atomistic detail over the time and length scales of a drug molecule transportation and delivery across a cell membrane is near impossible even with the present day computational capabilities. Hence, in this work, we have used coarse-grain molecular dynamics (CGMD) simulation approach. We choose to simulate a model system of this problem, in which we explore the transportation of polymer micelle assisted molecular components across a biological lipid vesicle. We envisage such a model system provide insights for the efficient drug transportation across the cell membrane and delivery inside the cell.

For the present study, we employed the MARTINI coarse grain force field as the basis[22,23]. To this end we have constructed two different simulation systems. In the first step, we constructed a lipid vesicle and a polymer micelle system in water. For the later stage of simulations, model hydrophilic components were loaded inside the polymer micelle obtained in Step 1. In order to construct the simulation system, we used a preassembled POPC lipid vesicle in water (shown in Figure 1). The POPC lipid vesicle was minimized and equilibrated in water. The details of polymer micelle are described below.

In a prior work, Srinivas and Pitera have carried out systematic self-assembly simulation studies of binary diblock copolymer mixtures in water[41]. The hydrophilic polymer block was parameterized to represent polyethylene glycol (PEG) while the hydrophobic polymer parameters represent polyethylethylene (PEE) in that study[41]. By varying interaction between polymer blocks they could obtain polymer micelles with different morphological structures. For example, by selecting one of the hydrophilic blocks to be relatively more self-attractive, they obtained "patchy" copolymer micelles, in which the selective polymer blocks formed the "patches" on the surface of polymer micelle. The number and size of such "patches" in a given micelle can be controlled by the initial composition of diblock copolymer mixture. Further details of this system can be found elsewhere[41]. Such patchy particles have been the subject of other computational and experimental studies as well[42-44]4. Recently, Hammond *et al.* studied mixed "patchy" micelle nanocarriers for systemic tumor targeting[43].

In this study, we use similar diblock copolymer "patchy" micelles composing of *A*-*B* and *C*-*B* copolymers. However, in this study, *B* represents a simple alkane (hydrophobic polymer) chain while the hydrophilic polymers were represented by *A* and *C*. We use intermediate attraction parameters for *B* polymers. The polymer *C* is similar to *A*, except *C*-*C* self-interaction is the strongest compared to other polymer interactions in the system. CG interaction parameters adapted from the MARTINI force field for various CG sites are listed in Table 1. CG representation of individual *AB, CB* polymers and *POPC* lipid are shown in Figure 1. During the self-assembly process, due to highly specific attractive interaction, *C* polymer blocks form "patch" like structures on the surface of hydrophobic core of the micelle as shown in Figure 1.

The initial binary mixture involved 120 *AB* copolymers and 40 *CB* copolymers corresponding to a composition of 3:1 copolymer mixture. This ratio was chosen so as to obtain three "patchy" structures within the polymer micelle (number of patches increase with increasing *CB* copolymer composition).

The "patchy" micelle obtained from binary copolymer mixture self-assembly in water is shown in Figure 1. The self-assembled "patchy" polymer micelle was added to the previously equilibrated lipid vesicle-water system. While constructing such system, minimum distance between polymer micelle and lipid vesicle surfaces was ensured to be greater than 1.2nm, which is the cut-off distance for non-bonded interactions in the present simulations. The final system contained a total of 877 *POPC* lipids, 120 *AB* copolymers, 40 *CB* copolymers and 67364 CG water molecules, corresponding to a total number of 80768 CG sites. CG water molecules represent a loosely bound, three atomistic water molecules[41]. This system was minimized using conjugate-gradient method before equilibrating for 1ns. All the simulations were conducted using GROMACS software[45,46] in an NPT ensemble with a 10fs time step. Due to the relatively smoother potential employed, CG simulations allow the usage of such large time steps. All the simulations were carried out at 300K temperature and 1 atm pressure. Production simulation of this system was carried out for at least 400ns. Note that due to smoother potentials, CG timescales are typically two to three orders larger than the corresponding atomistic timescales. Nevertheless, all the times reported in this work are actual simulation duration in nanoseconds, not the adjusted coarse-grain timescales.

III. RESULTS AND DISCUSSION

The "patchy" micelle obtained from the self-assembly of binary copolymer mixture (*AB* and *CB*) in water is shown in Figure 1. This polymeric micelle has hydrophobic core formed by *B* polymers (shown in cyan), while the polymer *A* (shown yellow) constitutes the shell structure. The hydrophilic "patches" on the core surface were formed by the association of hydrophilic polymer *C*. These patches are depicted as blue in the Figure 1. Details of polymeric interaction parameters are described in Table 1.

A. Lipid vesicle and empty polymer micelle interaction

Here we present the results from CGMD simulations of binary copolymer micelle interaction with POPC lipid vesicle in water. Initially, polymer micelle and lipid vesicle were dissolved in water and separated by 11nm (center to center distance) as shown in Figure 2(A). During the course of CGMD simulation, polymeric micelle moved to lipid vesicle proximity. When the micelle moves within the interaction range of lipid vesicle, favorable interaction between partially charged lipid head groups and hydrophilic polymers initiates the association, as shown

in Figure 2(B). Due to this favorable interaction, hydrophilic polymers start penetrating into the lipid head region of the vesicle. Subsequently, hydrophobic polymer blocks, get associated with lipid tail region to penetrate as well. As the penetration progresses, a narrow path into the vesicle is formed (Figure 2(C)). As the path widens over time, polymer micelle further penetrates into lipid vesicle. Around 400ns micelle was found to fully penetrate inside the vesicle as revealed by the simulation snapshot (Figure 2(G)). The micelle penetration process was accompanied by significant structural changes in lipid vesicle morphology. We find that the spherically shaped lipid vesicle was transformed to near ellipsoidal shape during the micelle penetration process as shown in Figures 2(D) and 2(E). Clearly, this morphological change was driven by favorable lipid head and hydrophilic polymer interaction. A cross section of final snapshot at 400ns (Figure 2(H) reveals that most of the penetrated micelle resides inside the vesicle. We observe that during the penetration process, few individual polymer got detached from the polymer micelle and distributed into lipid vesicle. This observation is in accordance with recent experimental studies. For example, Zhang *et al.* observed the polymer disassembly event during the release of hydrophobic guest molecules that have been encapsulated inside the copolymer micelles[47]. Kataoka *et al.* observed similar behavior in the case of loaded anticancer doxorubicin (DOX) drug inside copolymer micelles during the slow release of the drug[11,14].

The average center-to-center distance between polymer micelle and lipid vesicles was monitored as a function of time during the above process. We observe a systematic decrease in the distance as shown in Figure 3. Few representative snapshots of fusion process are also shown in Figure 4 at the corresponding times. Interestingly, we find that the decrease in vesicle to polymer distance accelerates in the initial stages of fusion process. This also demonstrates that the fusion process nearly completes in less than 300ns and the reorganization process continues beyond 400ns.

In order to gain deeper insights into the fusion process, we further examined the simulation trajectory, in particular near the timeframe where the fusion process initiated. The simulation snapshots from this timeframe with a nanosecond time interval are shown in Figure 4. As can be seen from this figure, by 158 ns, the micelle and vesicle reach close proximity, but still remain as separate entities without any “molecular contact” established. At 159ns a first “molecular contact” was established between polymer micelle and lipid vesicle as shown in the Figure. In particular, the initial contact was facilitated via a contact between *CB* copolymer and lipid head group, as shown in the Figure. This initial contact acted as a bridge between polymer micelle and lipid vesicle to provide a “fusion pathway” for remaining molecules. The interaction between the molecules from both lipid vesicle and polymeric micelle along this pathway gradually lead to the “micelle-vesicle” fusion as demonstrated by the simulation snapshots in the Figure 4. Snapshots also reveal that most of the polymer micelle gets adsorbed inside the lipid vesicle by 300ns.

The energetics associated with the fusion process was examined in detail as well. In Figure 5, we plot the total energy of the simulation system as a function of time. As can be seen from the figure, the total energy of the system decreases as the fusion process continues, revealing the

process is energetically favored. In order to further examine the contribution of individual energetic components, we have calculated the individual energy contributions as well. Contribution from each of the components as a function of time is shown in Figure 6. We find that the favorable interaction between lipids and polymers favors the fusion process as shown in Figure 6. A close examination of this figure reveals, the lipid-*CB* copolymer interaction shows a relatively rapid decrease in the energy during the fusion process (50ns-250ns). Hence, it is fair to conclude that the patch forming copolymer (*CB*) interaction with lipids accelerates the fusion process. We have also calculated the polymer water interaction during the fusion process. As shown in Figures 6(C) and 6(D), unfavorable *CB* copolymer-water interactions nearly offset the energetic gain from favorable *AB* copolymer-water interaction. This observation further confirms that the polymer-lipid interaction is the primary driving force for the vesicle-micelle fusion process.

B. Loaded polymer micelle interaction with lipid vesicle

In the next step, we proceed to study incorporation of hydrophilic molecules in the binary polymeric micelle. To begin with, we use unconnected coarse grain beads to model representative hydrophilic drug like components. The incorporation of hydrophilic molecules inside the polymer micelle was done by replacing three selected copolymers within the hydrophilic patch of the micelle with unconnected CG beads. The interaction energy parameter (ε and σ) for these CG units was chosen to be the same as supra attractive groups (*C* polymers). In other words, they are hydrophilic like *C* polymers but they are single unconnected CG beads, not the polymers. Note that the micelle core is primarily hydrophobic and hence the preferential position for the hydrophilic components is the location of hydrophilic patches [41]. In the absence of such patchy structures, the hydrophilic contents get distributed within the micelle and may diffuse into the surrounding water environment as they are hydrophilic. The patchy structure serves two purposes in this case: (a) they accommodate hydrophilic components within the hydrophobic micelle core environment (b) they protect hydrophilic contents from water exposure during the micelle fusion and transportation process. The interaction parameters for such hydrophilic CG beads with rest of the CG system were obtained by using mixing rules. The complete list of interaction parameters are listed in Table 1. After incorporating hydrophilic contents, polymeric micelle was equilibrated for 1ns in water. The equilibrated polymer micelle contained 33 hydrophilic components residing in the patchy location. The weight percentage of the hydrophilic components corresponds to 14% of molecular weight of the polymer patch.

As before, at the beginning of the CGMD analysis, the equilibrated polymer micelle was placed at a distance greater than 1.2nm from the surface of the lipid vesicle. The simulation conditions including time step size, temperature, pressure and ensemble were chosen to be same as before. During the course of the CGMD analysis the polymer micelle interacts with lipid vesicle, as

before. During the micelle interaction with lipid vesicle, relative movements of hydrophilic components were closely monitored. As before, the initial molecular contact was facilitated by the favorable interaction between *CB* polymer blocks of the micelle with the lipid head groups of the vesicle. As a result, the polymer micelle progressively penetrated inside the lipid head group region of the vesicle. During the course of the interaction, the hydrophobic polymers of the micelle get exposed, as the micelle opens up. At this stage the hydrophobic polymer blocks penetrate towards the nonpolar lipid tail region due to favorable interactions as shown in Figure 7. This further exposed the polymer micelle core, thereby releasing the inner contents of the polymer micelle. As described before, polymer micelle was loaded with hydrophilic components (depicted as magenta beads in the Figure 7), which quickly diffuse into polar lipid head group region. However, the lipid vesicle has relatively thinner head group region (~1nm) compared to hydrophobic tail region (3-4nm). Hydrophilic contents do not favor the hydrophobic lipid tail region and move out of that region in less than few nanoseconds. We find that some of the hydrophilic components reversed the course and moved out of the vesicle (into the bulk water), while the remaining entered the inner core region of the vesicle, where the confined water exists (as shown in the final snapshots of Figure 7). In order to demonstrate typical pathways for the hydrophilic component transportation across the vesicle, two representative trajectories are shown in Figure 8. The trajectory in Figure 8(A) corresponds to the case where hydrophilic component successfully crossed the lipid membrane and reaches confined water inside the vesicle. On the other hand, the trajectory shown in Figure 8(B) corresponds to hydrophilic component that moved out of the lipid vesicle, after initial penetration, but reaches the bulk water outside the vesicle. Initial and final positions of the hydrophilic components are marked in the figure with white and yellow arrows, respectively. In order to determine the hydrophilic component transportation and delivery, we examined the density profiles of individual components. The normalized density profiles obtained by averaging over the last 100ns of simulation trajectory are shown in Figure 9. For the calculation convenience, the center of the lipid vesicle was shifted to origin. Hence, the “zero” on x-axis represents the center of the lipid vesicle. The density profiles clearly show a significant portion of hydrophilic components reached the confined water region inside the inner core of the vesicle. This is demonstrated by the presence of two hydrophilic component density peaks near the origin. We find approximately 50% of initial hydrophilic contents successfully transported across the lipid bilayer and delivered into the inner core of lipid vesicle. Present CGMD simulations thus demonstrate the incorporation, of hydrophilic components in polymer micelle and their efficient transportation across lipid bilayer to deliver inside lipid vesicle.

Together, the above coarse grain molecular dynamics simulation results reveal underlying molecular level picture for polymer micelle assisted hydrophilic content transportation across lipid membrane and delivery inside the lipid vesicle. Our CGMD analysis shows that the “patch” forming hydrophilic polymers interact with the lipid head groups of the vesicle to initiate micelle-vesicle fusion process. Consequently, the hydrophobic inner core of lipid vesicle gets

exposed to the hydrophobic micelle core. During this step, the hydrophilic contents get transported across lipid membrane via the pathways provided by the hydrophobic polymers.

IV. CONCLUSIONS

In conclusion, we have conducted a detailed coarse grain (CG) MD study of polymer assisted drug transportation and delivery inside a POPC lipid vesicle. For this purpose, we have modeled drug components as unconnected coarse grain beads with relatively hydrophilic nature. In order to address the incorporation and transportation of hydrophilic drug components, we considered a specifically designed "patch" forming polymer micelle, assembled from a special class of "hydrophilic blocks" that form patches on the micelle surface. Our simulations show that such "patchy" polymer micelles reliably accommodate hydrophilic contents.

We have performed this study in two stages. In the first stage, we studied the interaction of empty polymer micelle with POPC lipid vesicle. We find that the copolymer micelle interacts with lipid vesicle and gets fully absorbed, without destroying the lipid vesicle structure. In the second stage, we have incorporated hydrophilic contents inside the patchy polymer micelle. After the incorporation, equilibrated polymer micelle was placed in water along with the lipid vesicle. During the course of this simulation, the polymer micelle interacts with lipid vesicle due to favorable interactions between lipid head groups and hydrophilic polymer blocks. This opens up the polymer micelle, thereby exposing polymer hydrophobic core, which in turn preferably interacts with hydrophobic lipid tails. As the lipid vesicle reorganizes in an effort to accommodate polymer micelle, hydrophilic contents get released into the lipid vesicle head group region. Present CGMD analysis reveal that the hydrophilic contents get transported across the lipid bilayer in a specific polymer assisted pathway. Due to unfavorable apolar/polar interactions, hydrophilic contents quickly diffuse out of the lipid bilayer region and move into the inner core of lipid vesicle, where the confined water exists. Normalized density profiles averaged over the last 100ns of simulation trajectory reveal that a significant portion of hydrophilic components reached the confined water region inside the inner core of the vesicle. We find that nearly 50% of hydrophilic components were transported and delivered inside the lipid vesicle. Clearly, for the effective drug transportation and delivery, efficiency needs to be further improved in this case. Since this is a primitive model for drug-like components, there is opportunity and need for the improvement of the force field parameters. Nevertheless, insights obtained from the present study could potentially help both experimental and simulation community to understand and design better delivery vehicles both for hydrophobic and hydrophilic drugs.

Present study reveals polymer assisted model hydrophilic component transportation and delivery inside the lipid vesicle. Despite a detail modeling of polymers and lipid molecules, we have considered a simple single bead model to represent hydrophilic drug components in this study. We aim to incorporate specific hydrophilic drug molecules inside the polymer micelle, so as to study their transport and delivery mechanism inside lipid vesicles. However, this task involves and requires developing new CG parameters for the drug molecules as well as their cross interaction with polymers and lipid molecules, and is in progress.

Acknowledgement

This work was supported in parts by the U. S. Army Research Office under award/contract no. W911NF-11-1-0168. We would like to acknowledge the use of high performance computational facilities at the North Carolina State A&T University during the course of the present study.

Supporting Information

Some of the additional parameters, not listed in the Table 1, are presented in the supporting information. Additional figures to aid demonstration of hydrophilic component transportation mechanism are also presented. This information is available free of charge via the internet at http://pubs.acs.org/.

References:

1. Langer, R; Drug Delivery and Targeting, *Nature*, **1998**, *392*, 5–10
2. Hammond, P.T., Virtual Issue on Nanomaterials for Drug Delivery. *ACS Nano*, **2011**, *5*, 681-684.
3. Farokhzad, O. C.; Langer, R; Impact of Nanotechnology on Drug Delivery *ACS Nano* **2009**, *3*,16– 20.
4. Cheng, H.; Kastrup, C.; J.; Ramanathan, R.; Siegwart, D. J.; Ma, M.; Bogatyrev, S. R.; Xu,Q.; Whitehead, K. A.; Langer, R.; Anderson, D. G.Nanoparticulate Cellular Patches for Cell-Mediated Tumoritropic Delivery *ACS Nano* **2010**, *4*, 625– 631
5. Torchilin, V.P.; Recent Approaches to Intracellular Delivery of Drugs and DNA and Organelle Targeting, Annual Review of Biomedical Engineering, **2006**, *8*, 343–375.
6. McNeil, S.; E. Nanoparticle Therapeutics: A Personal Perspective *Nanomed. Nanobiotechnol.* **2009**, *1*, 264–271.

7. Lee, S.-M.; Ahn, R.; W.; Chen, F.; Fought, A. J.; O'Halloran, T. V.; Cryns, V. L.; Nguyen, S. T.; Biological Evaluation of pH-Responsive Polymer-Caged Nanobins for Breast Cancer Therapy *ACS Nano* **2008**, *4*, 4971– 4978.

8. Rios-Doria, J.; Carie, A.; Costich, T.; Burke, D.; Skaff, H.; Panicucci, R.; Sill, K.; A Versatile Polymer Micelle Drug Delivery System for Encapsulation and In Vivo Stabilization of Hydrophobic Anticancer Drugs, *Journal of Drug Delivery*, **2012**, Article ID 951741.

9. Kataoka, K.; Kwon,G.S.; Yokoyama, M.; Okano, T.; Sakurai, Y.; Block Copolymer Micelles a Vehicles for Drug Delivery *J. Control. Release*, **1993**, *24*, 119-132.

10. Kwon, G.; Naito, M.; Yokoyama, M.; Okano, T.; Sakurai, Y.; Kataoka, K.; Block Copolymer Micelles for Drug Delivery: loading and release of doxorubicin**,** *J. Control. Release*, **1997**, *48*, 195–201.

11. Harada, A.; Kataoka, K; Supramolecular Assemblies of Block Copolymers in Aqueous Media as Nanocontainers Relevant to Biological Applications, *Progress in Polymer Science, 2006, 31*, 949–982.

12. Attia, E.; Bte, A.; Ong, Z.Y.; Hedrick, J.L.; Lee, P.P.; Ee, P.L.R.; Hammond, P.T.; Yang, Y-Y.; Mixed Micelles Self-assembled from Block Copolymers for Drug Delivery. *Curr. Opin. Colloid Interface Sci.* **2011**, *16*, 182-194.

13. Kim, B.-S.; Lee, H.-i.; Min, Y.; Poon, Z.; Hammond, P. T.; "Hydrogen-Bonded Multilayer of Ph-Responsive Polymeric Micelles with Tannic Acid for Surface Drug Delivery". *Chem. Comm.* **2009**, *28*, 4194-4196.

14. Tyrrell Z.L.; Shen,Y; Radosz, M.; Fabrication of Micellar Nanoparticles for Drug Delivery Through the Self-assembly of Block Copolymers, *Progress in Polymer Science*, 2010, *35*, 1128–1143.

15. Liu, M.; Kono, K.; Frechet, M.J.; Water-Soluble Dendritic Unimolecular Micelles: Their Potential as Drug Delivery agents; *Journal of Controlled Release*, **2000**, *65*, 121-131.

16. Kataoka, K.; Matsumoto, T.; Yokoyama M; et al.; Doxorubicin-Loaded Poly(ethylene glycol)–Poly(β-benzyl-l-aspartate) Copolymer Micelles: Their Pharmaceutical Characteristics and Biological Significance *Journal of Controlled Release*, **2000,** *64*, 143–153.

17. Geng, Y.; Dalhaimer, P.; Cai, S.; Tsai, R.; Tewari, M.; Minko, T.; and Discher, D.; Shape Effects of Filaments Versus Spherical Particles in Flow and Drug Delivery,; *Nature Nanotechnology*, **2007**, *2*, 249-255.

18. Ahmed, F.; Pakunlu, R.I.; Srinivas, G.; Brannan, A.; Bates, F.; Klein, M.L.; Minko,T.; Discher, D. E.; Shrinkage of a Rapidly Growing Tumor by Drug-Loaded Polymersomes: pH-Triggered Release through Copolymer Degradation. *Molecular Pharmaceutics*, **2006,** *3*, 340-350.

19. Yih, T. C.; Al-Fandi, M.; Engineered Nanoparticles as Precise Drug Delivery Systems *J. Cell. Biochem.* **2006**, *97*, 1184– 1190;

20. Kwon, G.S.; Polymeric Micelles for Delivery of Poorly Water-Soluble Compounds**,** *Crit. Rev. Ther. Drug Carr. Syst.*, **2003**, *20*, 357.

21. Li P-C; Makarov D E; Theoretical Studies of the Mechanical Unfolding of the Muscle Protein Titin: Bridging the Time-Scale Gap Between Simulation and Experiment *J. Chem. Phys.* 2003, *119***,** 9260.

22. Venturoli, M.; Smit, B; *PhysChemComm* Simulating the Self-Assembly of Model Membranes **1999**, *2*, 10.

23. Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H.; The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations, *J. Phys. Chem. B* **2007**, 111, 7812;

24. http://md.chem.rug.nl/cgmartini/index.php/downloads/force-field-parameters

25. De Vries, A. H.; Mark, A. E.; Marrink, S. J.; Molecular Dynamics Simulation of the Spontaneous Formation of a Small DPPC Vesicle in Water in Atomistic Detail *J. Am. Chem. Soc.* **2004**, *126*, 4488.;

26. Tieleman, D. P.; Leontiadou, H.; Mark, A. E.; Marrink, S. J.; Simulation of Pore Formation in Lipid Bilayers by Mechanical Stress and Electric Fields *J. Am. Chem. Soc.* **2003**, *125*, 6382.

27. Marrink, S. J.; Mark, A. E.; The Mechanism of Vesicle Fusion as Revealed by Molecular Dynamics Simulations *J. Am. Chem. Soc* **2003**, *125*, 11144.

28. Marrink, S. J.; Mark, de Vries, A. H.; Mark, A. E.; Coarse Grained Model for Semiquantitative Lipid Simulations *J. Phys. Chem. B* **2004**, *108*, 750-760.

29. Wu, z.; Cui, Q. and Yethiraj, A.; A New Coarse-Grained Model for Water: The Importance of Electrostatic Interactions", *J. Phys. Chem. B* **2010,** *114***,** 10524-10529.;

30. Mondal, J.; Sung, B. J.; and Yethiraj, A.; Sequence Dependent Self-assembly of Beta-Peptides: Insight from a Coarse-Grained Model", *J. Chem. Phys.* **2010**, *132*, 065103.

31. Kranenburg, M.; Venturoli, M.; Smit, B. Phase Behavior and Induced Interdigitation in Bilayers Studied with Dissipative Particle Dynamics, *J. Phys. Chem. B* **2003**, *107*, 11491.

32. Shelley, J. C.; Shelley, M. Y.; Reeder, R. C.; Bandyopadhyay, S.; Klein, M. L.; A Coarse Grain Model for Phospholipid Simulations *J. Phys. Chem. B* **2001**, *105*, 4464.

33. Shelley, J. C.; Shelley, M. Y.; Reeder, R. C.; Bandyopadhyay, S.; Moore, P. B.; Klein, M. L.; Simulations of Phospholipids Using a Coarse Grain Model *J. Phys. Chem. B* **2001**, *105*, 9785.

34. Nielsen, S.; Klein, M. L.; In *Bridging time scales: Molecular simulations for the next decade*; Nielaba, P., Mareschali, M., Ciccotti, G., Eds.; Springer-Verlag: Berlin, Germany, **2002**; 25-63.

35. Nielsen, S.; Lopez, C. F.; Srinivas, G.; Klein, M. L.; A Coarse-Grain Model for n-Alkanes Parameterized from Surface Tension Data *J. Chem. Phys.* **2003**, *119*, 7043.

36. Nielsen, S. O.; Lopez, C. F.; Srinivas, G.; Klein, M. L.; Bridging the Time Scales by Coarse Grain Molecular Dynamics Simulations *J. Phys. Condens. Matter* **2004**, *16*, 481.

37. Srinivas, G.; Shelley, J.C.; Nielsen, S.; Discher, D.E. ; Klein, M. L.; Simulation of Diblock Copolymer Self-Assembly Using a Coarse-Grain Model, *J. Phys. Chem. B*, **2004**, *108* , 8153–8160.

38. Srinivas, G.; Discher,D.E.; and Klein, M.L. ; Self-assembly and Properties of Diblock Copolymers via Coarse Grain Molecular Dynamics Simulations., *Nature Materials*, 2004, *3*, 638.

39. Srinivas, G. and Klein, M. L.; Coarse-Grain Molecular Dynamics Simulations of Diblock Copolymer Surfactants Interacting With a Lipid Bilayer. *Molecular Physics*, 2004, *102*, 883.

40. Loverde, S.M.; Klein, M.L.; Discher, D.E.; Nanoparticle Shape Improves Delivery: Rational Coarse Grain Molecular Dynamics (rCG-MD) of Taxol in Worm-Like PEG-PCL Micelles, Advanced Materials, 2012, *24*, 2011.

41. Srinivas, G.; Pitera, J.; Soft Patchy Nanoparticles from Solution-Phase Self-Assembly of Binary Diblock Copolymers *Nanoletters*, **2008**, *8*, 611.

42. Zhang, A.; Glotzer, S. C.; Self-Assembly of Patchy Particles, *Nano Lett.* **2004**, *4*, 1407.

43. Poon, Z.; Lee, J.A.; Huang, S.; Prevost, R.J.; Hammond, P.T.; Highly Stable, Ligand-Clustered "Patchy" Micelle Nanocarriers for Systemic Tumor Targeting. *Nanomedicine: Nanotechnology, Biology and Medicine.* **2011**, *7*, 201-209.

44. Kim, S.H.; Tan, J.P.K.; Nederberg, F.; Fukushima, K.; Yang, Y.Y.; Waymouth R.M.; and Hedrick, J.L.; Mixed Micelle Formation Through Stereocomplexation Between Enantiomeric Poly(lactide) Block Copolymers *Macromolecules* **2009**, *42* , 25-29.

45. Berendsen, h.j.c.; VAN DER Spoel, D.; van Drunen, R.; . GROMACS: A Message-Passing Parallel Molecular Dynamics Implementation, *Comp. Phys. Comm.* **1995**, *91***,** 43-56.

46. Lindahl, E.; Hess, B.; van der Spoel, D.; GROMACS 3.0: A Package for Molecular Simulation and Trajectory Analysis, *J. Mol. Model.* **2001**, *7***:** 306-317.

47. Wei , H; Zhuo, R; Zhang, X; Design and Development of Polymeric Micelles with Cleavable Links for Intracellular Drug Delivery, *Progress in Polymer Science, In Press, Corrected Proof, Available online 16 July **2012**.*

Figure captions:

Figure 1: CG models of *AB, CB* copolymers and *POPC* lipid used in this study are shown in top panel. Bottom panel: The initial configuration of *AB* and *CB* binary copolymer micelle (right) is shown along with a POPC lipid vesicle used in the present coarse-grain molecular dynamics simulations. A cross-sectional view of the lipid vesicle is shown in order to reveal the internal structure. One of the lipid head groups are shown in bead representation while the rest of the lipid components are shown in stick (cyan) representation. Water is not shown for figure clarity. Polymer block color code: *A* (yellow), *B* (cyan (smaller beads)) and *C* (blue).

Figure 2: CGMD simulation snapshots of a binary *AB+CB* polymeric micelle interaction with POPC lipid vesicle are shown. Polymeric patches in micelle are shown in blue along with hydrophobic (cyan) and hydrophilic (yellow) polymers. One of the CG beads in lipid head group is shown in bead representation, while rest of the lipid is drawn in stick representation. Water is not shown for clarity. Initially separated polymeric micelle (at 0ns) gradually penetrates the lipid vesicle (400ns). The cross sectional view of final snapshot is shown in (H).

Figure 3: Distance between the centers of lipid vesicle and polymer micelle is shown as a function of simulation time during the micelle-vesicle fusion process. Few representative simulation snapshots at corresponding times are also shown in the figure.

Figure 4: CGMD simulation snapshots from the trajectory near the fusion process of binary *AB+CB* polymeric micelle with POPC lipid vesicle are shown with a 1ns interval. Polymeric patches in micelle are shown as blue beads while the lipids (cyan) and hydrophobic (cyan) and hydrophilic (yellow) polymers are shown in stick representation. The lipid molecule that initiated the "molecular contact" for the fusion process is highlighted in bead (cyan) representation. Water is not shown for clarity. Initially separated polymeric micelle (at 0ns) progressively penetrates the lipid vesicle as shown in final snapshots (250ns).

Figure 5: Total energy of the simulations system is plotted as a function of simulation time during the polymer micelle and lipid vesicle fusion process.

Figure 6: Individual interaction potential energies are plotted as a function of simulation time: (A) Lipid-*AB* copolymer (B) Lipid-*CB* copolymer (C) Water-*AB* copolymer and (D) Water-*CB* copolymer. Note the relatively faster decay in the lipid-*CB* polymer interaction energy during the initial stages of fusion process (50-250ns).

Figure 7: Hydrophilic component (shown in magenta) transportation across lipid membrane and release inside the inner core of the vesicle as observed in CGMD simulations. Lipid vesicle is represented in the transparent sticks (gray color) to demonstrate the hydrophilic content

penetration inside the vesicle. Polymers are shown in yellow and blue stick representation while water is not shown for clarity.

Figure 8: Representative trajectories (depicted as red zigzag lines) for two different hydrophilic components (shown in magenta) as observed in CGMD simulations: (A) shows the trajectory for a hydrophilic component that successfully got transported into the confined water and (B) shows the trajectory for the hydrophilic component that moves outside of the lipid vesicle, after the initial penetration. Inner and outer membrane locations of the vesicle are highlighted with dashed white and yellow circles, respectively. Initial and final positions of hydrophilic components are pointed by white and yellow arrows, respectively. Lipid vesicle is shown in the stick representation while water and polymers are not shown for clarity.

Figure 9: Normalized mass density profiles (plotted in arbitrary units) obtained by averaging over the final 100ns trajectory are plotted as a function of distance from the center of the lipid vesicle. Hydrophilic components density profile shows a significant portion have reached the vesicle center, as demonstrated by the two peaks in the region of -2.5nm to 2.5nm. However, the presence of peaks beyond 5nm distance from the vesicle center reveals that some of the hydrophilic components did not get transported to the vesicle center. Density profile for hydrophilic components magnified for the aid of visual clarity.

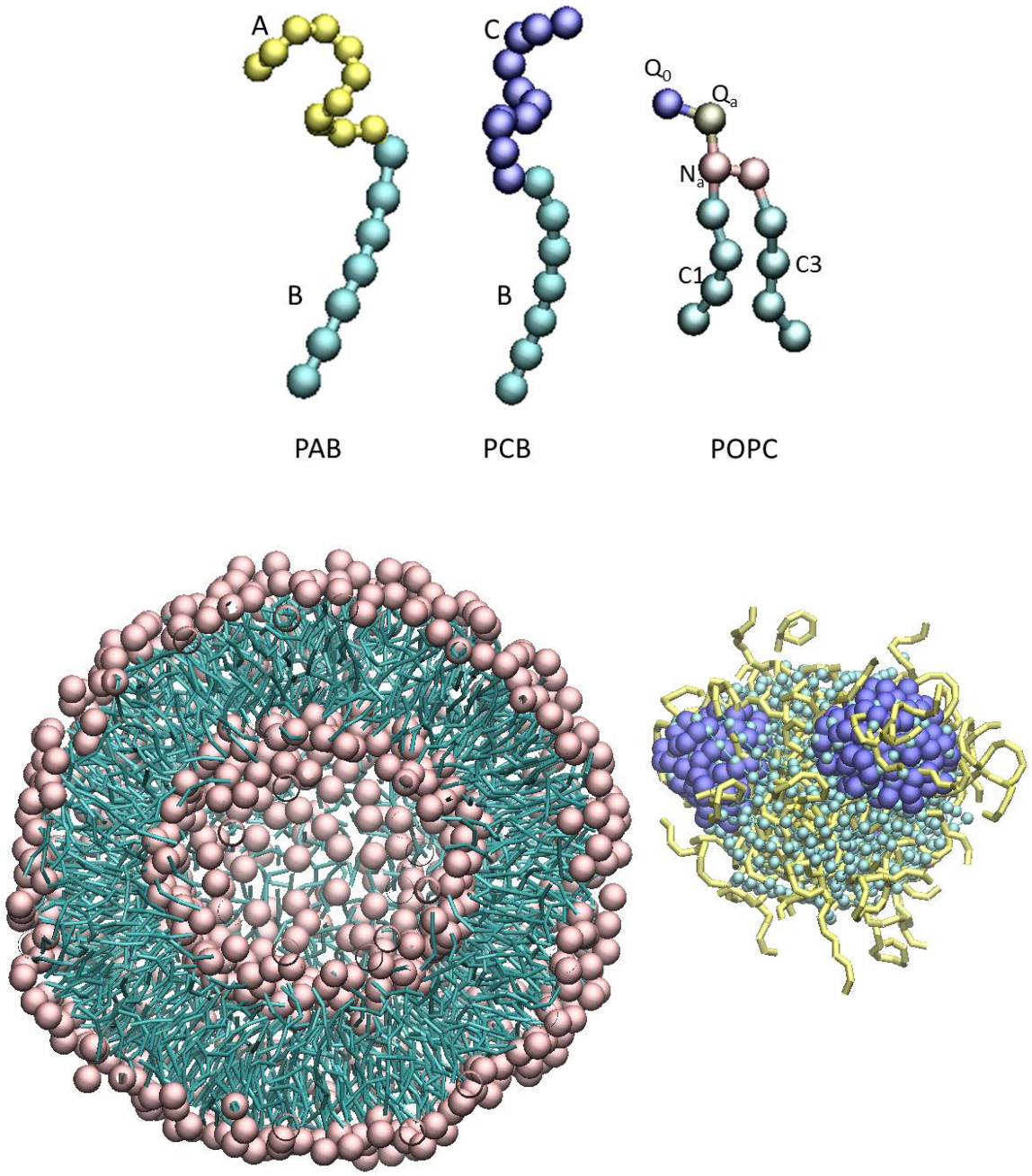


Figure 1: CG models of *AB*, *CB* copolymers and *POPC* lipid used in this study are shown in top panel. Bottom panel: The initial configuration of *AB* and *CB* binary copolymer micelle (right) is shown along with a POPC lipid vesicle used in the present coarse-grain molecular dynamics simulations. A cross-sectional view of the lipid vesicle is shown in order to reveal the internal structure. One of the lipid head groups are shown in bead representation while the rest of the lipid components are shown in stick (cyan) representation. Water is not shown for figure clarity. Polymer block color code: *A* (yellow), *B* (cyan (smaller beads)) and *C* (blue).

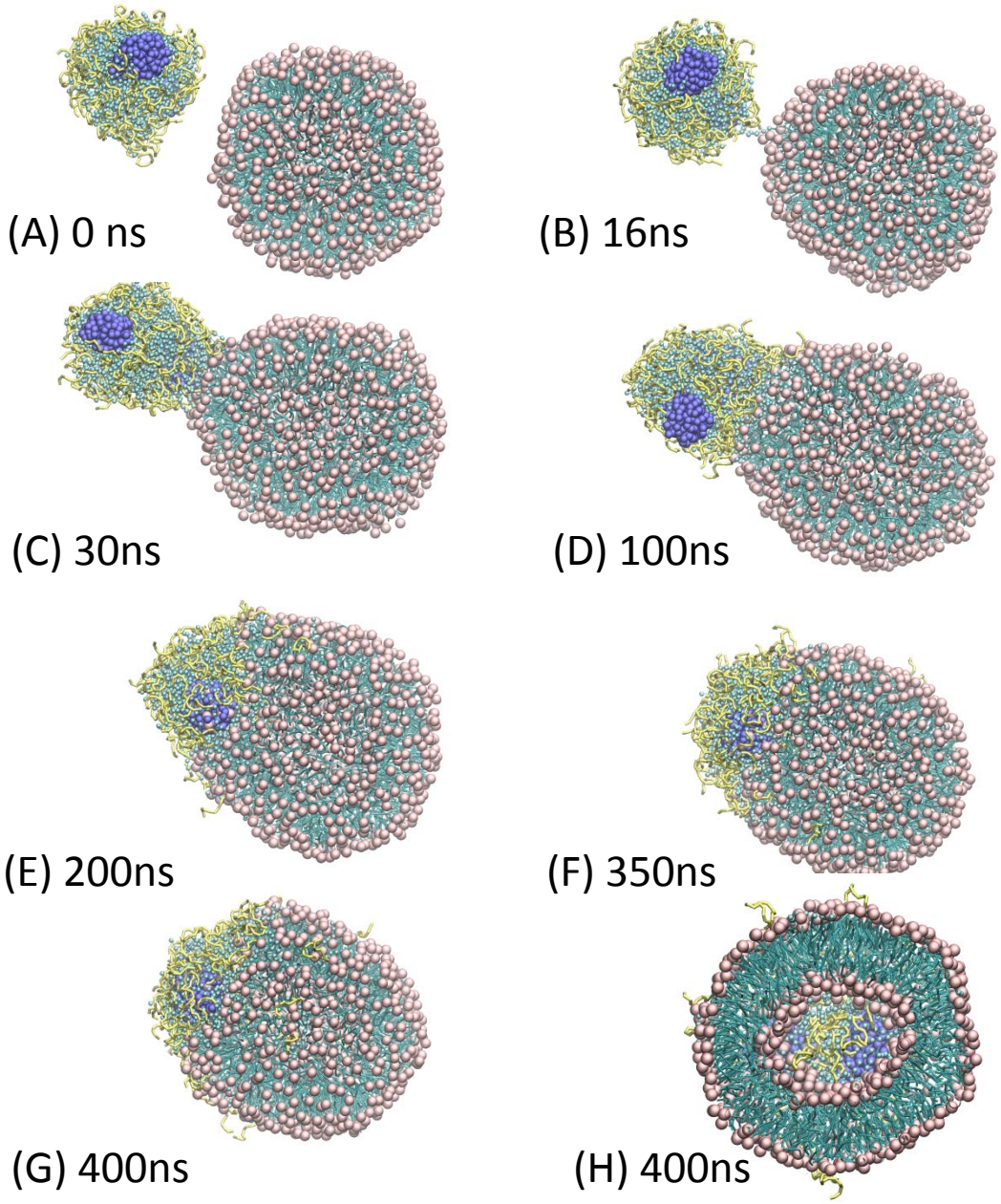


Figure 2: CGMD simulation snapshots of a binary *AB+CB* polymeric micelle interaction with POPC lipid vesicle are shown. Polymeric patches in micelle are shown in blue along with hydrophobic (cyan) and hydrophilic (yellow) polymers. One of the CG beads in lipid head group is shown in bead representation, while rest of the lipid is drawn in stick representation. Water is not shown for clarity. Initially separated polymeric micelle (at 0ns) gradually penetrates the lipid vesicle (400ns). The cross sectional view of final snapshot is shown in (H).

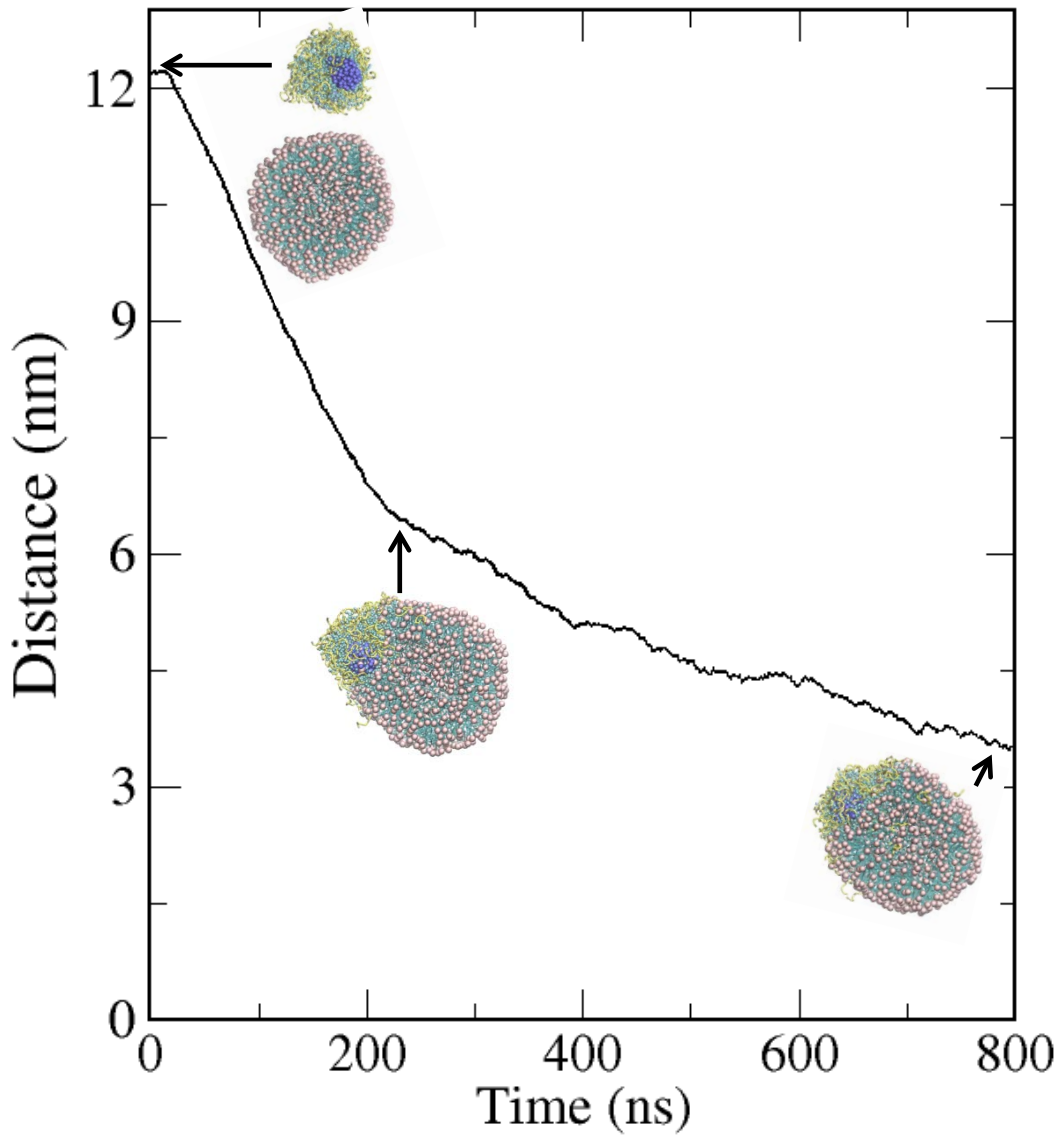


Figure 3: Distance between the centers of lipid vesicle and polymer micelle is shown as a function of simulation time during the micelle-vesicle fusion process. Few representative simulation snapshots at corresponding times are also shown in the figure.

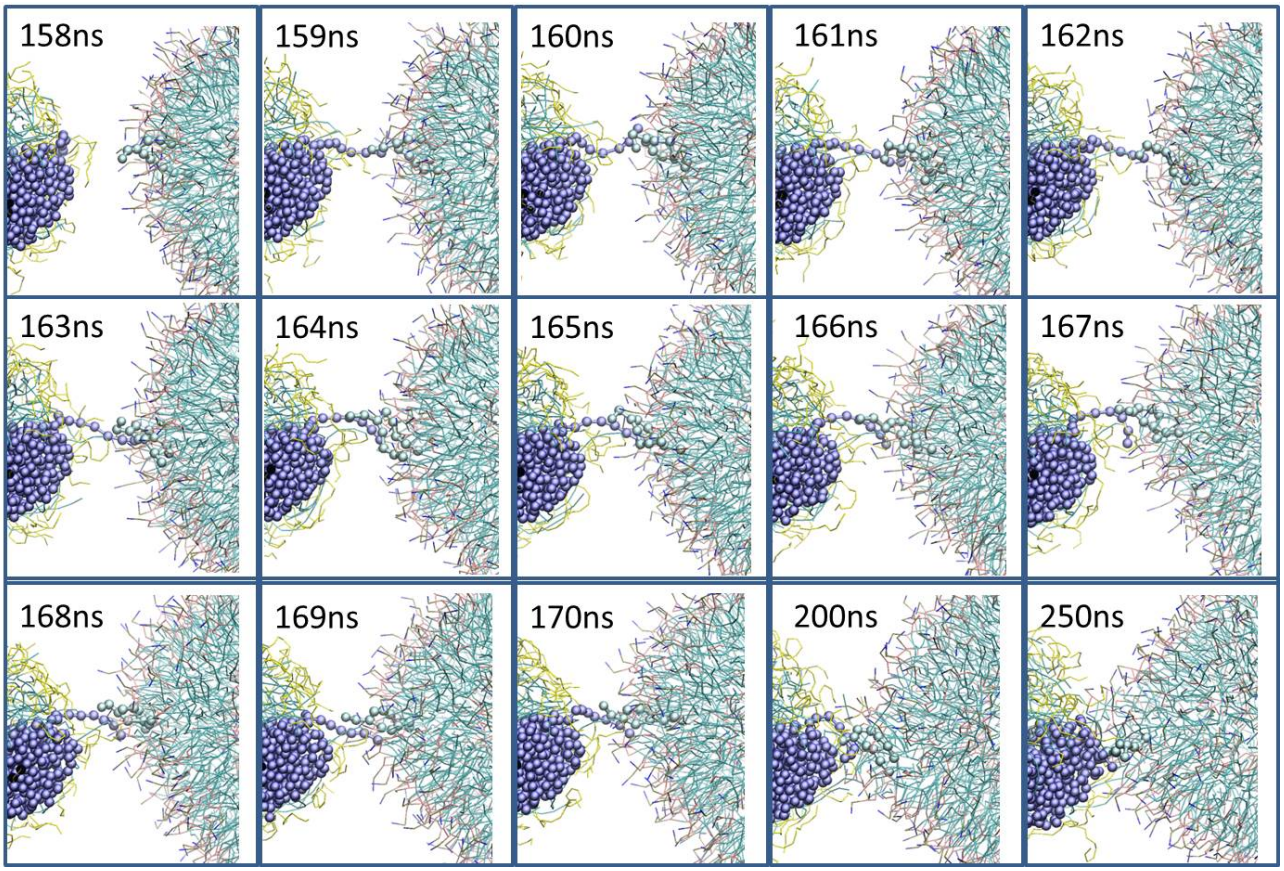


Figure 4: CGMD simulation snapshots from the trajectory near the fusion process of binary *AB+CB* polymeric micelle with POPC lipid vesicle are shown with a 1ns interval. Polymeric patches in micelle are shown as blue beads while the lipids (cyan) and hydrophobic (cyan) and hydrophilic (yellow) polymers are shown in stick representation. The lipid molecule that initiated the "molecular contact" for the fusion process is highlighted in bead (cyan) representation. Water is not shown for clarity. Initially separated polymeric micelle (at 0ns) progressively penetrates the lipid vesicle as shown in final snapshots (250ns).

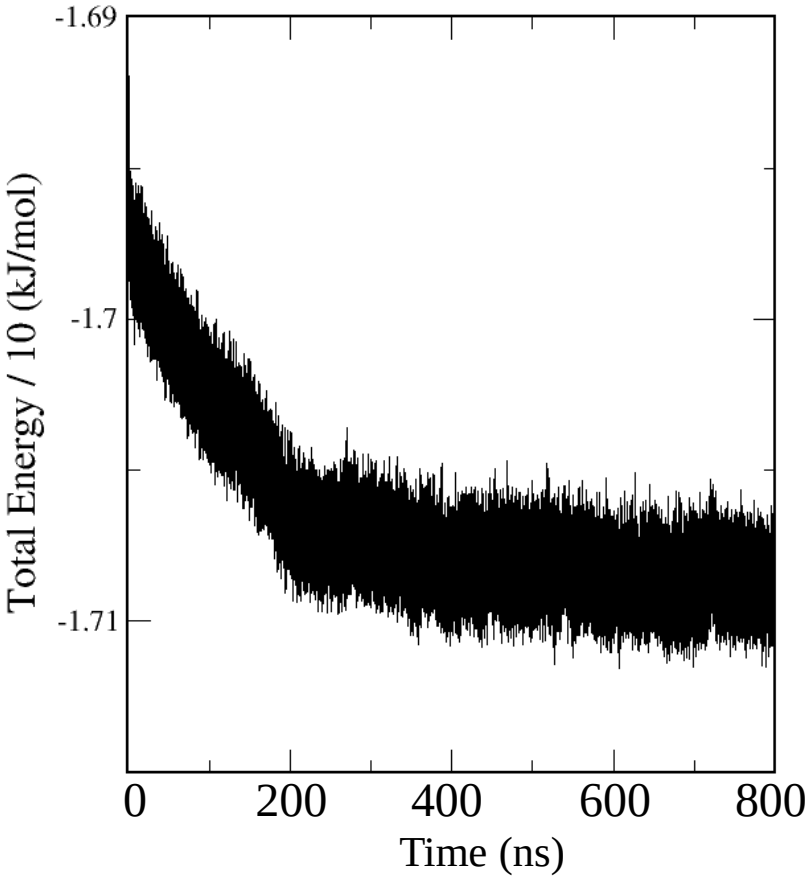


Figure 5: Total energy of the simulations system is plotted as a function of simulation time during the polymer micelle and lipid vesicle fusion process.

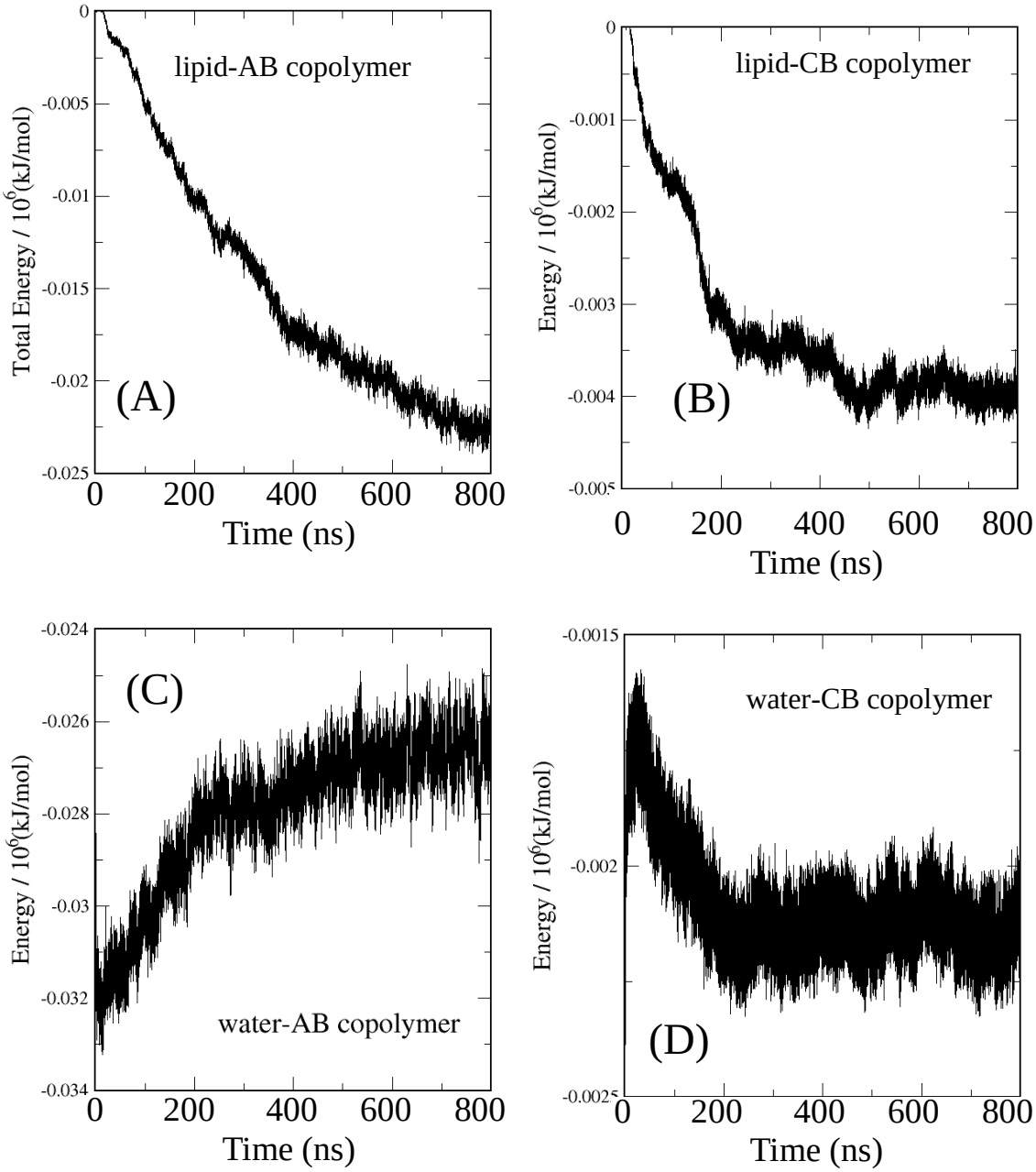


Figure 6: Individual interaction potential energies are plotted as a function of simulation time: (A) Lipid-*AB* copolymer (B) Lipid-*CB* copolymer (C) Water-*AB* copolymer and (D) Water-*CB* copolymer. Note the relatively faster decay in the lipid-*CB* polymer interaction energy during the initial stages of fusion process (50-250ns).

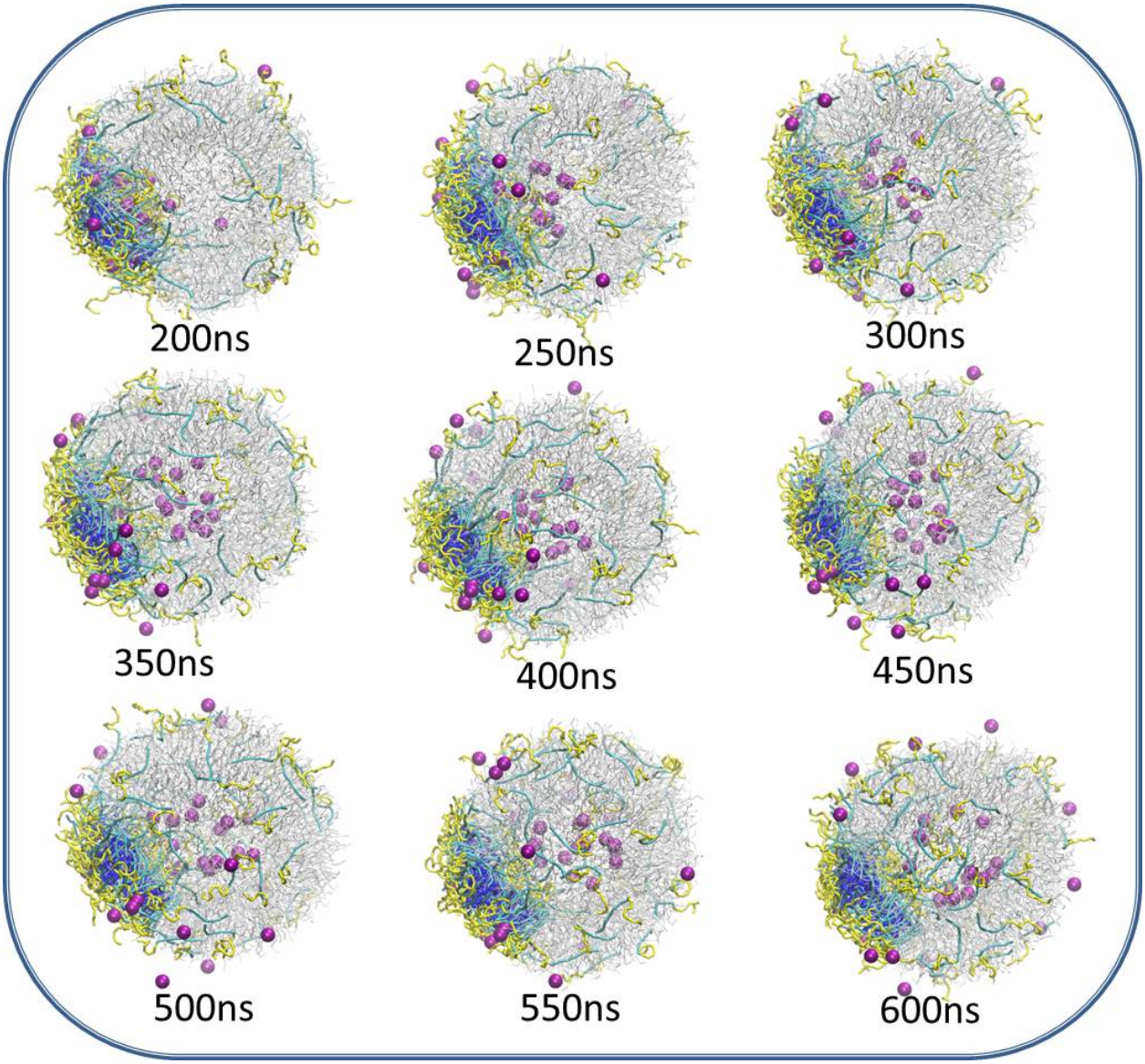


Figure 7: Hydrophilic component (shown in magenta) transportation across lipid membrane and release inside the inner core of the vesicle as observed in CGMD simulations. Lipid vesicle is represented in the transparent sticks (gray color) to demonstrate the hydrophilic content penetration inside the vesicle. Polymers are shown in yellow and blue stick representation while water is not shown for clarity.

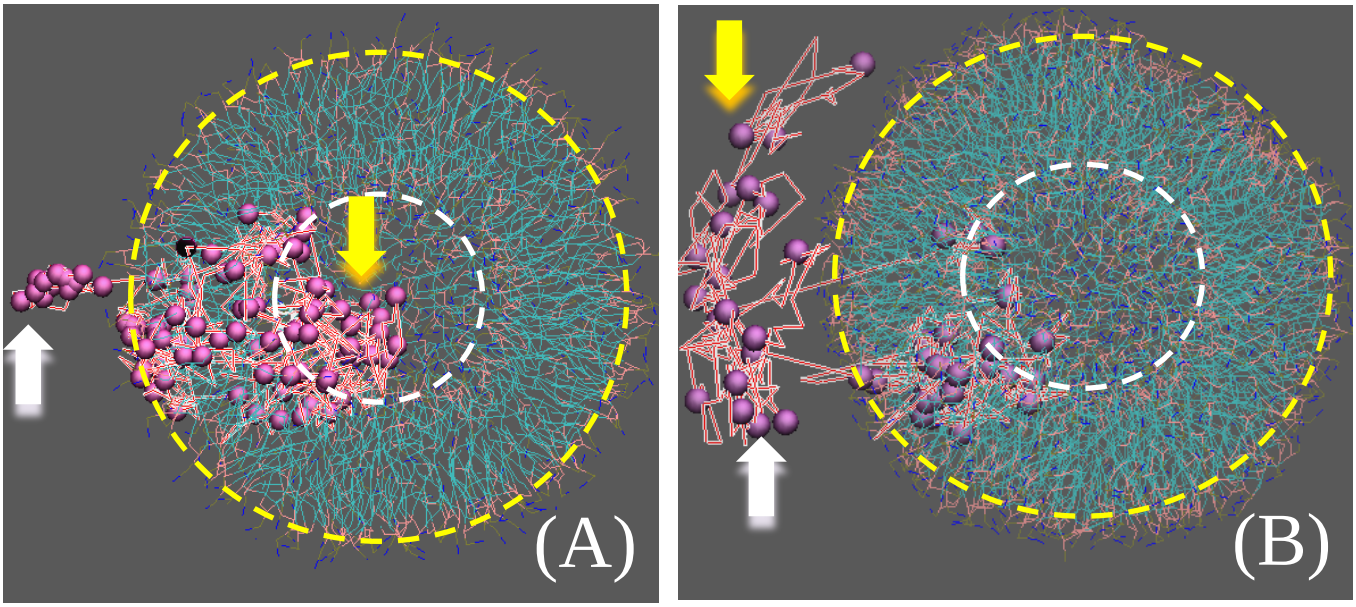


Figure 8: Representative trajectories (depicted as red zigzag lines) for two different hydrophilic components (shown in magenta) as observed in CGMD simulations: (A) shows the trajectory for a hydrophilic component that successfully got transported into the confined water and (B) shows the trajectory for the hydrophilic component that moves outside of the lipid vesicle, after the initial penetration. Inner and outer membrane locations of the vesicle are highlighted with dashed white and yellow circles, respectively. Initial and final positions of hydrophilic components are pointed by white and yellow arrows, respectively. Lipid vesicle is shown in the stick representation while water and polymers are not shown for clarity.

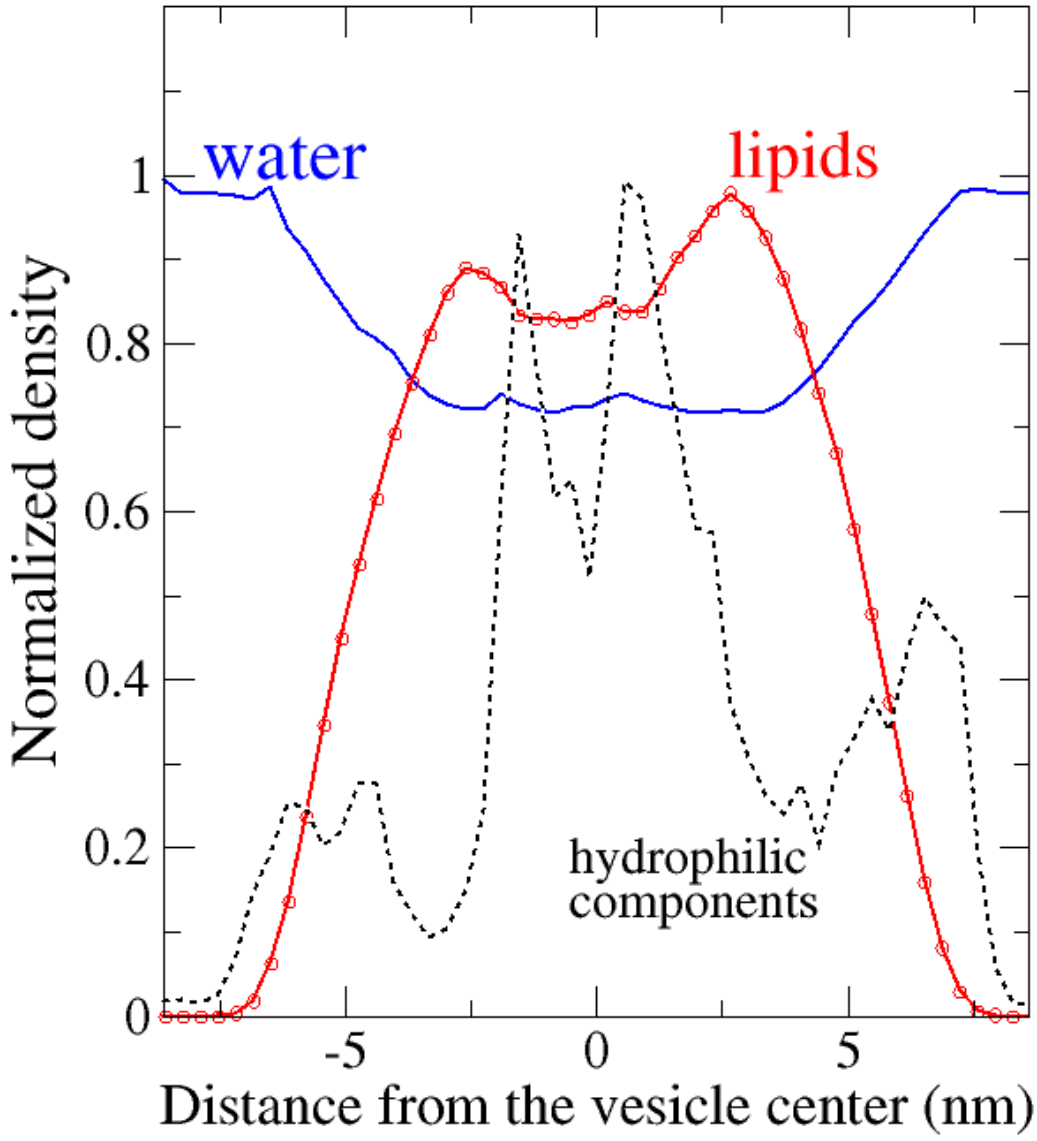


Figure 9: Normalized mass density profiles (plotted in arbitrary units) obtained by averaging over the final 100ns trajectory are plotted as a function of distance from the center of the lipid vesicle. Hydrophilic components density profile shows a significant portion have reached the vesicle center, as demonstrated by the two peaks in the region of -2.5nm to 2.5nm. However, the presence of peaks beyond 5nm distance from the vesicle center reveals that some of the hydrophilic components did not get transported to the vesicle center. Density profile for hydrophilic components magnified for the aid of visual clarity.

Table 1: Coarse-grain parameters for the polymer components used in this study are listed below. Additional parameters that are not listed below are either listed in supporting information or obtained by using mixing rule. HP indicates hydrophilic components.

Category	σ (nm)	ε (kJ/mol)	Pairs of interaction
Semi repulsive	0.470	2.691	A-C1, B-Q_a, B-Q_0, B-N_a
Repulsive	0.470	3.374	A-B, B-C
Super repulsive	0.620	1.997	A-C, A-HP, B-HP
Almost attractive	0.430	3.374	A-A, C-P4
Semi attractive	0.470	3.997	A-Q_0, A-Q_a, A-P4
Intermediate	0.470	3.497	B-B, B-C1, A-C3, A-N_a
Supra attractive	0.470	5.584	C-C, C-HP, HP-HP

Nanoparticle-incorporation and aggregation in cylyndrical polymer micelles:

We further explored the application of CG methodology to study nanomaterials and nano-composites as well. Nanoparticle-copolymer composites can result in a variety of novel structures that have potential in advanced catalytic activity, microelectronic, biomedical devices, optoelectronics, permeability and conductivity etc. [20]. Factors such as nanoparticle composition, volume fraction and their concentration play important role in determining the final morphology and the properties of such nano-composite materials. A challenging task in nano-lithography, is to fabricate highly regular and precise patterned arrays. Existing approaches such as optical lithographic techniques [21] are often complex and limited by the features sizes that can be achieved, while the electron beam lithography is time consuming and highly expensive. Recent studies suggest that such limitations can be overcome by using self-assembly of diblock copolymers (consist two chemically distinct polymer chains joined by a covalent bond at the interface) [22]. Dibock copolymers can self-assemble to form nanodomains that exhibit ordered morphologies at equilibrium and self-organize into ordered periodic structures on solid-surfaces on the tens of nanometer length scales.

Using the nano-composites such as mixing nanoparticles with block copolymers may lead to efficient alternative routes for pattern formation. For example, by using mean filed approaches, Balazas and coworkers [20] predicted a variety of mesophases of nanoparticle-copolymer (N-C) composites. They have theoretically shown that the interactions between mesophase forming copolymer and nanoscopic particles can lead to highly organized hybrid materials. Li and coworkers [23] presented a beautiful experimental demonstration of their prediction by using a combination of molecular designing strategy and synthetic expertise. They have developed a simple procedure to incorporate nanoparticles and control their location within different domains of diblock copolymer assemblies by controlling the surface chemistry of the particles. For example to localize the particles in domains A or B of A-B diblock copolymers, nanoparticles are coated with A or B type homopolymers, respectively. Furthermore, particles coated with a mixture of A- and B-type homopolymers localize the particles at the diblock interface. In another interesting experimental study [22], Russel and coworkers presented a simple and general route for fabricating nanostructured materials with hierarchical order. They could show that in the nanoparticle/copolymer mixtures, nanoparticles segregation to the interfaces mediates interfacial interactions and orients the copolymer domains normal to the surface from the parallel orientation. They have achieved the aggregation of both the particulate and polymer entities without the use of external field.

Despite such theoretical insights and experimental demonstration, an understanding of nanoparticle copolymer self-assembly mechanism is still lacking. The large time and length scales associated with N-C composite self-assembly, puts this problem beyond the reach of the traditional simulations based on atomistic details. Accordingly, in this work we aim to study N-C composite self-assembly phenomena using the above mentioned coarse-grain (CG) methodology. In addition we simulate the aggregation behavior of nanoparticles embedded within the preassembled cylindrical micelle, thereby providing further insights into the experimental designing strategies as well as a scope for future theoretical developments.

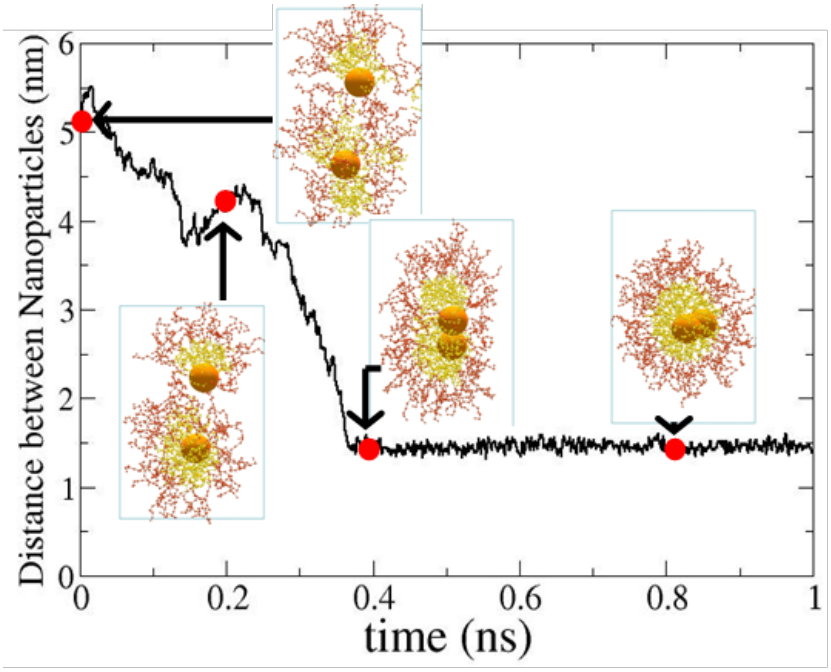


Figure 10: Distance between the nanoparticles during the nanoparticle-copolymer composite self-assembly is plotted as a function of simulation time. Snapshots corresponding to marked times are also shown. The nanoparticles are shown as spheres while the block copolymers are shown in hairy representation.

We study nanoparticle-copolymer composite self-assembly and aggregation in two stages (i) The N-C self-assembly from a random initial configuration and (ii) Aggregation of nanoparticles embedded within the preassembled cylindrical micelle. As mentioned above, we adopt coarse-grain (CG) simulation methodology. The force-field development and parameterization details for the block copolymers and solvent described elsewhere [24]. The block copolymers chosen in this study contain 50% hydrophilic fraction and are shown to form cylindrical micelles. Nanoparticles are chosen to be hydrophobic, 1.6nm in diameter, roughly four times larger than the CG sites representing rest of the simulation system.

The simulation started with an initial configuration with two nanoparticles and 196 diblock copolymers are randomly dispersed in 4000 CG waters. Figure 10 shows the distance between nanoparticles during the self-assembly process as a function of simulation time. Simulation snapshots corresponding to the positions marked in the figure are also shown. As can be seen from the figure, initially nanoparticles are placed at 5nm distance apart. Care has been taken in selecting the initial configuration such that the nanoparticles are sufficiently screened from each other by copolymers and/or solvent. The copolymers start aggregating around the nanoparticles thereby forming near spherical micelles, in which the nanoparticles reside in the hydrophobic core of the micelles. These initial aggregates explore favorable conditions for the fusion. This is shown in figure at 0.23ns where the nanoparticle separation is relatively increased. Nanoparticles show preferential interactions in the direction where there is a relatively smaller concentration of the adsorbed copolymers leading to a smaller screening. Further aggregation of the nanoparticles is assisted by the copolymer reorganization in the vicinity accordingly. As shown in Figure 1 the nanoparticle aggregation is not the final step of the self-assembly. During the next 0.5ns, copolymers self-organize and the self-assembly result in the formation of a nano-particle incorporated cylindrical micelle (or an N-C composite). Note that the nanoparticle separation nearly remains constant during late stage self-organization. The nanoparticle aggregation and copolymer rearrangement process lasts for about 0.15ns (40% of the selfassembly + aggregation time). Thus, simulations suggest that the nanoparticle aggregation in this case is driven by N-C composite fusion.

For comparison, we have carried out another simulation with the same initial configuration as above, without the nanoparticles. The self-assembly result in the cylindrical micelle formation in this case as well. Nevertheless, the dynamics are markedly different. Initially, the copolymers start forming local aggregates as in the case of N-C composites. Nevertheless, the aggregation process is found to be significantly slower in this case. We found that the copolymer aggregation without the nanoparticles take nearly four times longer to form cylindrical micelle. This is an interesting observation and highlights the role of nanoparticle in copolymer aggregation revealing that the presence of nanoparticles speeds up the self-assembly/aggregation process.

The above mentioned phenomenon is observed by starting from a random initial configuration in which the nanoparticles are dispersed in a copolymer/water mixture and separated by certain distance. In what follows, we examine the aggregation behavior of nanoparticles within a pre-assembled cylindrical micelle. In simulations, this is done by embedding the nanopartilcles in cylindrical micelle and placing them 3nm apart. The result is shown in Figure 11 where the distance between the two nanoparticles within the cylindrical micelle is shown as a function of simulation time. As shown in this figure, the dynamics of this system are very different from that observed for the self-assembly case. Although the nanoparticles separated by a relatively smaller distance, the time taken for the aggregation process is nearly an order of magnitude larger. Furthermore, the nanoparticle separation does not alter appreciably throughout the simulation

until the aggregation start to happen. We find that the major part of the aggregation time is devoted to the rearrangement of the copolymers surrounding the nanoparticles.

We have analyzed the density distribution of the copolymers surrounding the nanoparticles during the aggregation process to gain insights into the mechanism. The analysis suggests that the density of copolymers play a crucial role in triggering the nanoparticle aggregation process. Due to the course of dynamics, the copolymer density on the surface of the nanoparticles fluctuates. When the copolymer density on one of the nanoparticles gets depleted, the aggregation process triggers. We conclude that the local density of the copolymers surrounding the nanoparticles plays crucial role and drives the nanoparticle aggregation in such preassembled morphologies,

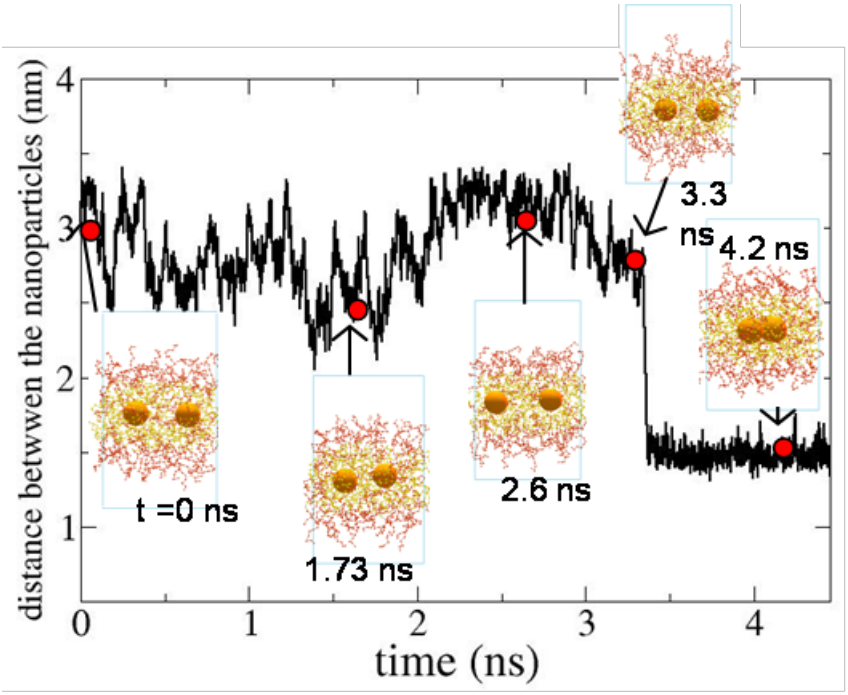


Figure 11: Distance between the nanoparticles embedded in a pre-assembled cylindrical micelle is plotted as a function of simulation time. Snapshots at five representative times are also shown. Nanoparticles are shown as spheres and polymers in hairy representation.

In order to verify the robustness of the current simulations and the system size effects, we have carried out further studies with a simulations system that is four times larger (8 nanoparticles and 784 copolymers). We found similar results in this case as well (details are presented in the supporting information).

In summary, using CG simulations we have studied the N-C composites self-assembly in water and pre-assembled morphologies. In addition to complimenting the existing theoretical

predictions and experimental observations, present simulations provide insights into nanoparticle-copolymer self-assembly mechanism. Such details may prove useful in furthering experimental strategies for the controlled patterned surfaces.

Need for the multi-scale simulation methods

It is important to note that due to the interdisciplinary nature of the bio-nano interface phenomena, as evident from our research work, neither atomistic nor coarse-grain simulations will be able to explore entirely. In order to explore such problems, we envision a hybrid atomistic/coarse-grain simulation approach based on the principles of quantum-mechanics/molecular-mechanics method. Such an approach may provide novel and efficient tools to study complex bio-nano interface phenomena on desired spatial and temporal scales.

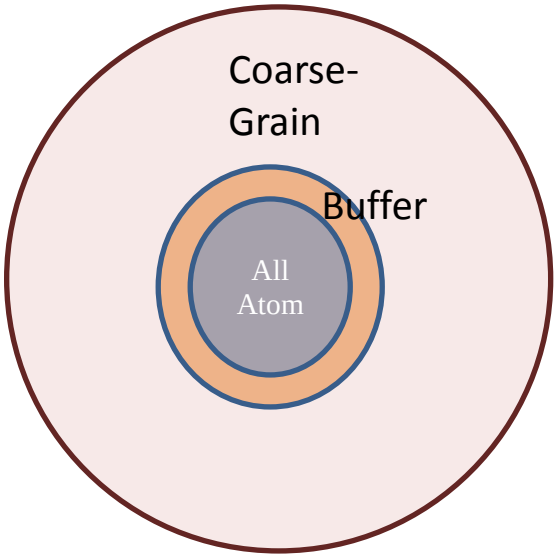


Figure 12: A schematic representation of hybrid AA/CG simulation approach with a defined buffer zone for the smooth interchange is shown.

As discussed before, atomistic simulations and coarse-grain simulations proven to be highly successful in exploring biomedical and material phenomena with suitable time and length scales. However, many processes that take place at the bio-nano interface cannot be fully studied using either one of the methods. For example, a large-scale protein or nanotubular motion occurs on time and length scales that can be efficiently studied using CG simulations while the highly important atomistic level interactions within these molecules can be best explored by classical

MD simulations. Hence, there is a need and opportunity for combining such disparate simulation methods using a hybrid multi-scale methodology. In principle, such method should be able to combine advantages of atomistic and coarse-grain simulations without sacrificing efficiency or losing the resolution. This is similar to the development of hybrid Quantum Mechanics/Molecular Mechanics (QM/MM) method which successfully combines atomistic and quantum mechanical simulations. In QM/MM simulations, a relatively small, highly specific region is defined where QM simulations will be carried out while the rest of the system will be treated using classical simulations. In addition, a buffer zone between QM and MM is created for smooth transition of QM and MM treatments as molecules and atoms moves in and out of either regions. A hybrid AA/CG method may utilize similar principles. However, unlike QM/MM methods, AA/CG hybrid method involves structural changes as well. In other words, in the buffer region between AA and CG regions, molecules will have part atomistic and part coarse-grain structure. Because of this reason, buffer region need to be relatively thicker compared to the case of QM/MM. A pictorial representation of possible hybrid AA/CG method is shown in Figure 12.

Recently, Voth *et al.* [25] proposed a Multiscale CG method and applied to 2-site and 4-site hexane models based on underlying atomistic data. Nevertheless, as the authors note, the method suffers from sampling issues. Moreover, its applicability for larger systems has not been tested, which is crucial to explore systems at bio-nano interface. Hence, we would like to further our present CG methodology to include hybrid AA/CG in an efficient fashion to explore bio-nano interface on relevant time and length scales.

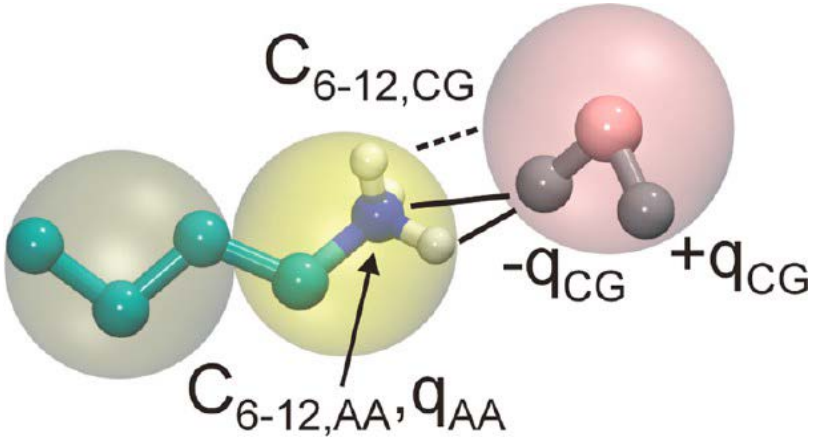


Figure 13: Concept of Gromos/MARTINI hybrid AA/CG simulations [26]. Lennard-Jones interactions between AA and CG subsystems (dashed line) are treated at the level of the CG force field by using virtual CG interaction sites (transparent beads). Charged particles in the AA and CG systems interact via Coulomb potentials (solid lines). The interactions within the AA and CG subsystems are described by the respective force fields.

In addition to the structural complexity, treatment of electrostatic interactions poses a tough challenge for the hybrid AA/CG simulations. In search of finding a solution to this problem, Schafer and coworkers explored the possibility of electrostatic coupling in hybrid AA/CG simulations by combining the Gromos atomistic force field with the MARTINI coarse-grained force field[10,11] (Figure 12). To enact electrostatic coupling, two recently developed CG water models with explicit electrostatic interactions were used: the polarizable MARTINI water model and the BMW model. The hybrid model was found to be sensitive to the strength of the AA−CG electrostatic coupling, which was adjusted through the relative dielectric permittivity εr(AA−CG). Potentials of mean force (PMFs) between pairs of amino acid side chain analogues in water and partitioning free enthalpies of uncharged amino acid side chain analogues between apolar solvent and water showed significant differences between the hybrid simulations and the fully AA or CG simulations, in particular for charged and polar molecules. For apolar molecules, they found that the hybrid AA/CG models are in better agreement with the fully atomistic results. Their research highlighted some key challenges on the way toward hybrid AA/CG models that are both computationally efficient and sufficiently accurate for biomolecular simulations.

Conclusion

In these research projects we explored intriguing phenomena that takes place at bio-nano interface. In particular we explored the interaction of drug-like component loaded polymer micelle with a biological lipid vesicle and nanoparticle incorporation and aggregation inside polymer micelles by using coarse-grain molecular dynamics simulation. For this purpose, we developed a coarse-grain model for polymers where necessary and used it in conjunction with existing MARTINI force-field for lipid molecules to study polymer micelle interaction with lipid vesicle.

In order to explore drug delivery mechanism, we have conducted a detailed coarse grain (CG) MD study of polymer assisted drug transportation and delivery inside a DPPC lipid vesicle. We used specifically designed “patch” forming polymer micelle to accommodate hydrophilic contents. In the first stage, we studied the interaction of empty polymer micelle with DPPC lipid vesicle. We find that the copolymer micelle interacts with lipid vesicle and gets fully absorbed, without destroying the lipid vesicle structure. In the second stage, we have incorporated hydrophilic contents inside the patchy polymer micelle. After the incorporation, equilibrated polymer micelle was placed in water along with the lipid vesicle. During the course of this simulation, the polymer micelle interacts with lipid vesicle due to favorable interactions between lipid head groups and hydrophilic polymer blocks. This opens up the polymer micelle, thereby exposing polymer hydrophobic core, which in turn preferably interacts with hydrophobic lipid tails. As the lipid vesicle reorganizes in an effort to accommodate polymer micelle, hydrophilic contents get released into the lipid vesicle head group region. Present CGMD analysis reveal that

the hydrophilic contents get transported across the lipid bilayer in a specific polymer assisted pathway. We find that nearly 50% of hydrophilic components were transported and delivered inside the lipid vesicle. Clearly, for the effective drug transportation and delivery, efficiency needs to be further improved in this case. Since this is a primitive model for drug-like components, there is opportunity and need for the improvement of the force field parameters. Nevertheless, insights obtained from the present study could potentially help both experimental and simulation community to understand and design better delivery vehicles both for hydrophobic and hydrophilic drugs.

However, despite a detail modeling of polymers and lipid molecules, we have considered a simple single bead model to represent hydrophilic drug components in this study. We aim to incorporate specific hydrophilic drug molecules inside the polymer micelle, so as to study their transport and delivery mechanism inside lipid vesicles. However, this task involves and requires developing new CG parameters for the drug molecules as well as their cross interaction with polymers and lipid molecules, and is in progress.

Present simulation studies clearly highlight the fact that due to the interdisciplinary nature of the bio-nano interface phenomena, neither atomistic nor coarse-grain simulations will be able to explore it entirely. In order to explore such problems, we need to develop hybrid atomistic/coarse-grain simulation approach based on the principles of quantum-mechanics/molecular-mechanics method. Such an approach may provide novel and efficient tools to study complex bio-nano interface phenomena on desired spatial and temporal scales. Work in this direction is under active progress.

Appendix

Coarse-grain Parameterization: As mentioned before, the coarse-grain parameterization fully relies on the information obtained from corresponding experimental and/or atomistic simulations. In this section we describe the development of CG parameters in detail. Similar to classical simulation force-field, CG force-field involves intramolecular and intermolecular interaction parameters. We describe developing each of these parameters separately.

Intramolecular Parameters: Typically, intramolecular parameters were determined based on atomistic information. For this purpose, an all-atom simulation with 144 polymer chains was carried out for 1 ns in an NPT ensemble at 298.15 K and 1 atm of pressure with a 1 fs time step. Data collected during the simulation was utilized to generate bond distance and bond angle distributions for the consecutive monomers in the polymer chain. These results were used as target observables for CG parameterization. The CG polymer chain was constructed using the center-of-mass position of each monomer of the corresponding AA polymer chain. Consecutive monomers in a polymer are bonded by harmonic bond,

$$V_b(r) = k_b(r - r_0)^2 \qquad \text{(A1)}$$

Where the equilibrium bond distance is denoted by r_0 and k_b represents the harmonic bond force constant. The bond distance distribution obtained from CG simulations was compared with the target distribution from atomistic simulations. r_0 and k_b were adjusted until AA and CG distributions show reasonable agreement. After each adjustment, CG simulation was repeated with the new bonded parameter set. After 4 to 5 iterations, a reasonable agreement between AA and CG distribution was obtained. Such iterations were carried out for each type of bonded interactions. Similarly, three consecutive monomers in a polymer chain were subjected to a harmonic bond angle potential,

$$V_a(\theta) = k_\theta(\theta - \theta_0)^2 \qquad \text{(A2)}$$

As before, the parameters, θ_0 the equilibrium angle and the force constant k_θ were determined by adjusting initial guesses until they satisfactorily reproduce the mean and variance of the bond bond angle distribution obtained from the AA simulations. Similar to the case of bond distance, parameter determination needed four to five CG simulation iterations.

Intermolecular Interactions:

Parameterization using experimental data: The nonbonded interactions among hydrophobic sites were modeled with a potential of the Lennard-Jones (9−6) form,

$$U(r_{ij}) = \frac{15}{4}\varepsilon\left[\left(\frac{\sigma}{r_{ij}}\right)^9 - \left(\frac{\sigma}{r_{ij}}\right)^6\right] \quad \text{(A3)}$$

where, ε and σ correspond to the well-depth and van der Waals diameter of the corresponding species. To determine these parameters, we have chosen experimental bulk density and surface tension of individual polymers as target observables (adapted from literature).

In order to determine ε and σ we constructed a relatively small CG simulation system of polymers in an NPT ensemble. The potential well-depth (ε) was adjusted until the simulation system yielded experimental density. While the calculation of density was straightforward, determining surface tension is complex and highly involved. For this purpose, we need to determine the van Der Walls radius (σ). To this end chose widely determined experimental quantity, namely surface tension as target observable. Calculating surface tension from simulations is not straight forward. For this purpose, we constructed a separate NVT simulation system where a slab of water was sandwiched between two polymer layers as shown in Figure 5. The box length in z-direction set to be nearly 10 times to that of other directions. Such arrangement produces two polymer-water interfaces and two air-polymer interfaces. The total tension of such system can be written as,

$$\tau_{tot} = 2\left[\gamma_{air-poly} + \gamma_{water-poly}\right] \quad \text{(A4)}$$

In the above equation, $\gamma_{air-poly}$ and $\gamma_{water-poly}$ represents air-polymer and water-polymer surface tension values, respectively. In simulation, the total surface tension can be calculated using the following formula,

$$\tau = \frac{L_z}{2}\left[P_{zz} - \left(\frac{P_{xx}+P_{yy}}{2}\right)\right] \quad \text{(A5)}$$

where L_z is the simulation box length in the longest (*z*) direction. P_{ij} represents pressure component along *ij* direction. The prefactor ½ in the above equation account for the factor that there are two such interfaces present in the simulation. σ was adjusted until simulation yielded a surface tension value such that $\left|\tau_{exp} - \tau_{sim}\right| \leq 2$dyn/cm. However, the experimental quantities may not be readily available for all the cases. In such cases, the distribution functions from underlying atomistic simulation were used as target observables, as describe below.

Parameterization using simulation data: Once the target data and the model have been established, the optimal non-bonded parameter set can be determined as follows. This procedure is numerically intensive and time-consuming but is largely devoid of physics. For the specific problem of using a tabulated potential to reproduce a target radial distribution function we follow a method suggested by Muller-Plathe [12], beginning with the potential $V_0(r)$ which is the

potential of mean force obtained by Boltzmann inversion of the target radial distribution function (g(r)), and subsequently iterating according to,

$$V_{n+1}(r) = V_n(r) + kT \ln \frac{g_n(r)}{g_{target}(r)} \quad \text{(A6)}$$

Fewer than ten iterations are normally required for convergence. Finally, we present a gradient based optimization method due to Lyubartsev and Laaksonen [13] which can be used to compute the full matrix of first partial derivatives, relating the changes in input parameters to the changes in observables, from a single simulation. We wish to establish the change in an observable A, here the radial distribution function g(r)) caused by a change in the parameter *a*. This relation allows for the parameter adjustment to be made optimally to first order. We assume the Hamiltonian,

$$H = \frac{P^2}{2M} + V(R; a) \quad \text{(A7)}$$

depends upon a parameter *a*, where the first term in equation (A2) is the kinetic energy of the system and the second term is the potential energy. For the observable $A \equiv A(R)$ depending only on the coordinates, its expectation value in the canonical ensemble is given by

$$\langle A \rangle = \frac{\int dR\, Ae^{-\beta H}}{\int dR e^{-\beta H}} \quad \text{(A8)}$$

where β is the inverse of the product of Boltzmann's constant and the temperature, so

$$\frac{\partial A}{\partial a} = -\beta \left[\langle A \frac{\partial H}{\partial a} \rangle - \langle A \rangle \langle \frac{\partial H}{\partial a} \rangle \right] \quad \text{(A9)}$$

This is the fundamental relation which links the change in input parameter to the change in output observable.

Parameterization using target observable from AA simulations: For the purpose of non-bonded parameterization, we chose radial distribution function from the AA as target observable. Initial CG simulations start with guess parameters that may utilize LJ type potentials for simplicity. The radial distribution function obtained from CG simulations then inserted in Eq.(A6) so as to obtain the updated potential $V_{n+1}(r)$. In the next step, the CG simulations utilize the updated potential. A new g(r) will be obtained from this simulation, which will be inserted in Eq.(A6) to update the potential. This iterative procedure continues until a reasonable agreement between CG and target g(r) is obtained. A comparison between radial distribution functions from AA and CG simulations after five such iterations is shown in Figure 6. More often than not, such iterative

procedure results in a potential that cannot be represented by LJ type potentials. Hence, a tabulated form, detailing a net potential value corresponding to each distance, was used to represent such potentials. Water is similarly parameterized, by modeling as a spherically symmetric interaction site (W), representing a loose grouping of three atomistic water molecules. Water sites interact via Lennard-Jones (6−4) potential.

References:

1. Li P-C and Makarov D E *J. Chem. Phys. 119*, 9260, **2003**.
2. Rafii-Tabar H and Chirazi A *Phys. Rep. 365***,** 145, **2002**.
3. Tries V, Paul W, Baschnagel J and Binder K *J. Chem. Phys. 106***,** 738,**1997.**
4. Stevens M J, Hoh J H and Woolf T B *Phys. Rev. Lett. 91* 188102,**2003.**
5. Cornelissen J J L M, Fischer M, Sommerdijk N A J M and Nolte R J M *Science 280***,** 1427,**1998.**
6. Zhang L and Eisenberg A *Science 268***,** 727,**1995.**
7. Pakula T, Karatasos T, Anastasiadis S H and Fytes G. *Macromolecules 30***,** 8463,**1997.**
8. J. Rios-Doria, A. Carie, T. Costich, B. Burke, H. Skaff, R. Panicucci, and K. Sill, Journal of Drug Delivery, Article ID 951741, **2012**.
9. K. Kataoka, T. Matsumoto, M. Yokoyama et al. *Journal of Controlled Release, 64*, 143–153, **2000**.
10. Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations. J. Phys. Chem. B **2007**, 111, 7812;
11. http://md.chem.rug.nl/cgmartini/index.php/downloads/force-field-parameters
12. Srinivas, G.; Pitera, J. Soft Patchy Nanoparticles from Solution-Phase Self-Assembly of Binary Diblock Copolymers. *Nano Lett.* **2008**, *8*, 611.
13. Zhang, A.; Glotzer, S. C. Self-Assembly of Patchy Particles. *Nano Lett.* **2004**, *4*, 1407.
14. Poon, Z.; Lee, J.A.; Huang, S.; Prevost, R.J.; Hammond, P.T. Highly Stable, Ligand-Clustered "Patchy" Micelle Nanocarriers for Systemic Tumor Targeting. *Nanome. Nanotech. Biol. Med.* **2011**, *7*, 201-209.
15. Berendsen, h.j.c.; VAN DER Spoel, D.; van Drunen, R. GROMACS: A Message-Passing Parallel Molecular Dynamics Implementation. *Comp. Phys. Comm.* **1995**, ***91*,** 43-56.
16. Lindahl, E.; Hess, B.; van der Spoel, D. GROMACS 3.0: A Package for Molecular Simulation and Trajectory Analysis. *J. Mol. Model.* **2001, *7*,** 306-317.
17. Wei, H.; Zhuo, R-X.; Zhang, X-Z. Design and Development of Polymeric Micelles with Cleavable Links for Intracellular Drug Delivery. *Prog. Poly. Sci.* **2012**, *38*, 503-535.
18. Harada, A.; Kataoka, K. Supramolecular Assemblies of Block Copolymers in Aqueous Media as Nanocontainers Relevant to Biological Applications. *Progress in Polymer Science*, **2006**, *31*, 949–982.
19. Tyrrell Z.L.; Shen,Y; Radosz, M. Fabrication of Micellar Nanoparticles for Drug Delivery Through the Self-assembly of Block Copolymers. *Progress in Polymer Science*, **2010**, *35*, 1128–1143.
20. Lin, Y. *et al. Nature*, **2005,** *434*, 55-59;
21. Bullen, H.A.; Garett, S.J.; *Nano Lett.* **2002,** *2*, 739-745.

22. Thurn-Albercht,T.; Schotter, J.; Kastle, G.A.; Emley, N.; Shibauchi. T.; Krusin-Elbaum, L.; Guarani, K.; Black, C.T.; Tuomine, M.T.; Russel, T.P.; *Science*, **2000,** *290*, 2126-2129.
23. Li, M.; Schnablegger, H.; Mann, S.; *Nature*, **1999,** *402*, 393-395.
24. Srinivas, G.; Discher,D.E.; and Klein, M.L. Self-assembly and Properties of Diblock Copolymers via Coarse Grain Molecular Dynamics Simulations. *Nature Materials*, **2004**, *3*, 638.
25. A.Das, L. Lu, H.C. Andersen, G.A.Voth, *J.Chem.Phys.* 194115, *196*, **2012**.
26. T.A. Wassenaar, H. I. Ingólfsson, M. Prieβ, S. J. Marrink, L. V. Schäfer. Mixing MARTINI: Electrostatic Coupling in Hybrid Atomistic–Coarse-Grained Biomolecular Simulations *J. Phys. Chem. B* **2012**, *117*,3516-3530.

2.2 Computational Modeling Studies of Fullerene Based Molecular Therapeutic Agents

G. Srinivas, R. Mohan, A. Kelkar, North Carolina A&T State University

Introduction

Drug delivery is equally important as the drug development process. For the drug molecule to be effective, it needs to be successfully transported across the cell membrane and need to be delivered at the targeted location within the cell. However, complexity of cell environment, the nature of drug molecules makes this goal complicated and involved. Broadly speaking, the drug molecules can be classified into hydrophobic and hydrophilic categories based on the nature of their interaction with water molecules. While the drug molecules are predominantly hydrophobic, a significant number of drug molecules such as anticancer doxorubicin and mitoxantrone are predominantly hydrophilic. Despite numerous experimental studies for drug delivery, efficient encapsulation and transportation of hydrophilic drug molecules such as anticancer doxorubicin and mitoxantrone is still a challenging task [1]. Rapid growth in nanotechnology combined with biological applications is revolutionizing drug delivery methodologies. Recent studies suggest new pathways for drug delivery that utilize nanomaterials as supporting or embedding materials for drug components [2-5]. For example, nanoparticle assisted drug transportation across cell membrane resulted in higher percentage of drug incorporation into intracellular domain [6-9]. Flexible materials such as self-assembled polymer micelles have also been shown to be promising candidates for not only efficient drug transportation but also for targeted delivery [10-12]. A schematic of polymer micelle assisted drug transportation is shown in Figure 1. This method has several advantages over other nanomaterial based approaches. For example, the drug components can be incorporated into polymer micelle core, thereby protecting them from external environment [3,4]. In addition, polymers can be designed/selected so as to interact efficiently with cell membrane [8]. During the polymeric micelle interaction with cell membrane, the drug content from micelle core gets released inside the cell membrane, which eventually reaches intracellular targeted domains [3,8]. Clearly, the nature of polymer and its interaction with cell membrane plays crucial role in this case. For the efficient drug transportation and delivery, ideally, a polymer micelle should effectively incorporate the drug components, as well as transport across cell membrane and release them at the targeted site. In order to design such efficient polymer micelles, it is important to understand how polymer micelle interacts with cell membrane from a molecular view point. Recently, we carried out a computational modeling study to obtain deeper understanding of how polymer micelle transports drug components across the cell membrane [13].

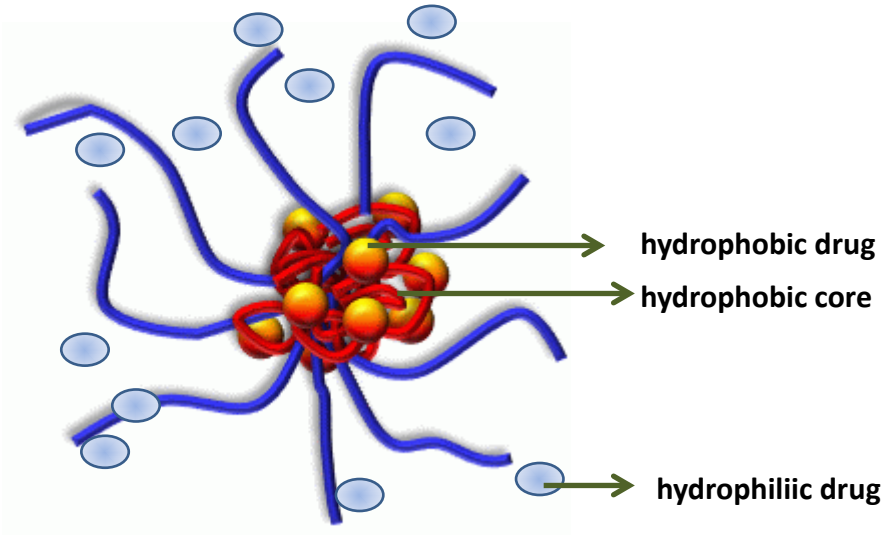


Figure 1: A schematic representation of polymer micelle carrying hydrophobic and hydrophilic drug components.

In particular, a detailed coarse grain (CG) MD study of polymer assisted drug transportation and delivery inside a DPPC lipid vesicle was studied by modeling drug components as unconnected coarse grain beads with relatively hydrophilic nature. In order to address the incorporation and transportation of hydrophilic drug components, a specifically designed "patch" forming polymer micelle was assembled from a special class of "hydrophilic blocks" that form patches on the micelle surface [14]. Simulations showed that such "patchy" polymer micelles reliably accommodate hydrophilic contents.

During the course of this simulation, the polymer micelle interacts with lipid vesicle due to favorable interactions between lipid head groups and hydrophilic polymer blocks. This opens up the polymer micelle, thereby exposing polymer hydrophobic core, which in turn preferably interacts with hydrophobic lipid tails. As the lipid vesicle reorganizes in an effort to accommodate polymer micelle, hydrophilic drugs from micelle get released into the lipid vesicle head group region. MD analysis reveals that the hydrophilic contents get transported across the lipid bilayer in a specific polymer assisted pathway. Due to unfavorable apolar/polar interactions, hydrophilic contents quickly diffuse out of the lipid bilayer region and move into the inner core of lipid vesicle, where the confined water exists. Nearly 50% of hydrophilic drugs were transported and delivered inside the lipid vesicle this way. Clearly, for the effective drug transportation and delivery, efficiency needs to be improved in this case. Since this is a primitive model for drug-like components, there is opportunity and need for the improvement of the force

field parameters. Nevertheless, insights obtained from the present study could potentially help both experimental and simulation community to understand and design better delivery vehicles both for hydrophobic and hydrophilic drugs.

In addition to the polymeric micelle carriers, many nano-material based nanocarriers have been explored for drug transportation and delivery. For example, carbon nanotubes and nanoparticles have been aggressively explored for this purpose. Among these nanomaterials, fullerene based molecular therapeutic agents have been highly promising.

A novel set of water soluble molecules termed "amphifullerene"compounds have been synthesized by Hirschand colleagues [15-20]. These amphifullerene nanostructures, based on a C60 core, contain both hydrophobic and hydrophilic moieties and self-assemble to form spherical vesicles referred to as "buckysomes" [17]. One such fullerene monomers is AF-1 which consists of a "buckyball" cage to which a Newkome-like dendrimer unit and ten lipophilic C12 chains positioned octahedrally to the dendrimer are attached. This globular amphiphile has a low critical micelle concentration and the polar dendrimer head group contains multiple carboxylic acid groups, resulting in pH sensitive assembly and release. The fullerene core in the amphifullerenes acts as an excellent carbon cage to which wide variety of hydrophilic and hydrophobic groups can be attached by well documented methodologies. The fullerene core along with the attached moieties determine the self-assembly process that leads to the formation of different nanostructures [21]. Fullerenes functionalized with different ionic groups have been shown to form aggregates [22], extended nanotubes [23], spheres [19,24,25], and vesicles [26]. Recently many therapeutic agents have been explored. Fullerene based therapeutic agents have been highly promising. However, for a drug delivery vehicle, high fullerene loading capacity is very important. Previous studies suffered from drawbacks including, low fullerene content of 3-7%. They have also shown to have damaged antioxidation bioactivity. Such drawbacks seriously dent the progress of fullerene based therapeutic agents. Recently, Kepley and coworkers introduced a novel fullerene delivery approach via liposomes encapsulated with amphiphilic C70 bisadducts that were designed to structurally mimic cellular membrane lipids. Due to the similar amphiphilic nature and structure, the strong association between fullerenes and auxiliary lipids allowed them to form dimensionally stable liposomes. Strikingly, the liposomes obtained this way contained as much as 65% (by weight) fullerene. Their approach provided a novel pathway to deliver fullerenes to sites where their antioxidant properties can be exploited. Their study also provided a novel platform for synthesizing a new class of fullerene based therapeutic agents that not only retain their biological activity, but can also be delivered via liposome carriers.

Encouraged by the success of CG molecular dynamics modeling in the case of polymeric micelle assisted drug-delivery; we explore fullerene nanomaterial based molecular therapeutic agents. In order to design better drug delivery vehicles based on fullerenes, it is highly important to understand fullerenes and its derivatives interaction with cell membrane from a molecular

viewpoint. For this purpose, we study the interaction of fullerenes and amphiphilic polymer-fullerene complexes with biological lipid molecules. In a next step, we also study the fullerene assisted self-assembly of liposomes.

Coarse-grain Modeling of Amphiphilic Fullerenes

Here we describe the coarse-grain modeling details of fullerene molecule and amphiphilic fullerenes. In this study we consider C60 fullerene molecules. An atomistic model of C60 fullerene is shown in Figure 1 along with a coarse grain representation. In the CG representation C60 was modeled as a 16-site spherical cage based on MARTINI coarse-grain force-field [27]. In order to design amphiphilic fullerene molecules, we have considered hydrophilic and hydrophobic polymer blocks. Simulation parameters for these polymers were obtained from our previous simulation studies [13,14]. These polymer models were successfully shown to reproduce experimentally observed micelle structures over a wide range of compositions.

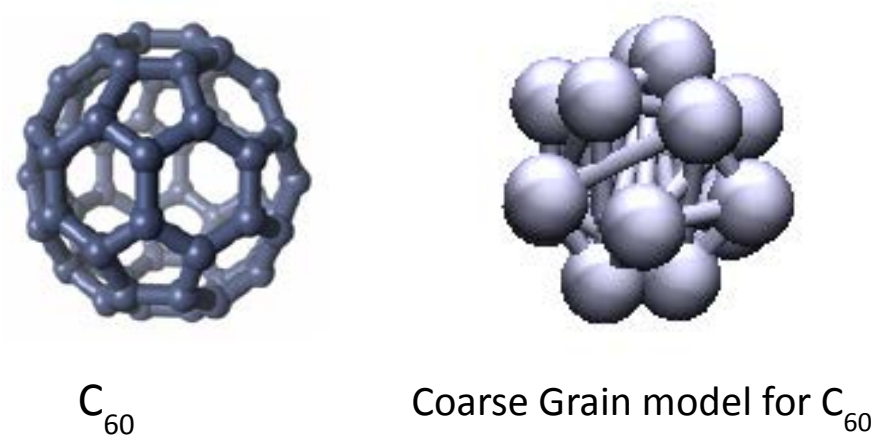


Figure 2: Atomistic and coarse-grain representations of C60 fullerene molecule are shown.

Amphiphilic Fullerene Design

For this purpose, motivated by similar amphiphilic nature of block copolymers, a series of C60 fullerene molecule based fullerene-polymer complexes were generated by linking one or more polymer blocks to the fullerene surface as show in Figure 3. The longest polymer chain had 9 monomers while the shortest polymer had 3 monomers. For simplicity, we represent fullerene by

F, hydrophobic and hydrophilic polymers as A and B, respectively. For example, F-A9 represents a hydrophilic polymer of 9 monomers linked to C60 fullerene molecules. Similarly, F-A3-B9 represents a trimer hydrophilic block and hydrophilic blocks of 9 monomers linked to the C60 fullerene surface. Details of all the designed molecules used in this study are given in Table 1.

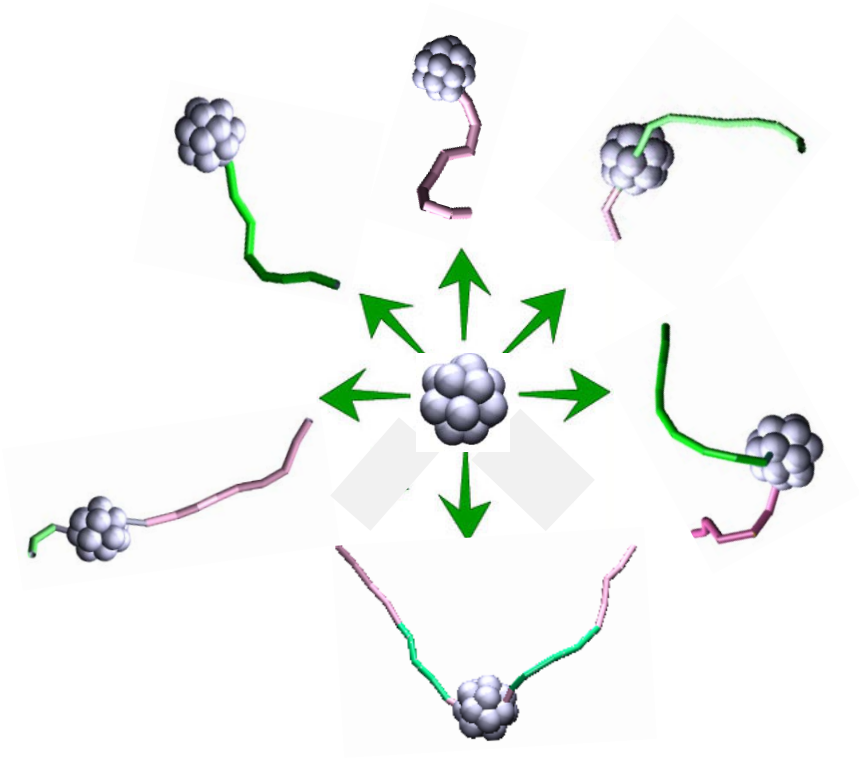


Figure 4: Various amphiphilic fullerene molecular complexes designed in this study are shown. Fullerene cage is represented as a spherical cage, while hydrophobic and hydrophilic blocks are shown in green and purple, respectively.

Simulation Details

Carrying out computer simulation studies at atomistic detail over the time and length scales of nanomaterial interaction with cell membrane is near impossible even with the advent of present day computational facilities. Hence, in this work, we have used coarse-grain molecular dynamics (CGMD) simulation approach. We choose to simulate a model system of this problem, in which we explore the interaction of amphiphilic fullerene molecular complexes with biological lipids in

water. We envisage such a model system provide insights for the design of better drug delivery vehicles.

For the present study, we employed "MARTINI" coarse grain force field as the basis [27]. To being with we have constructed several different simulation systems. All the systems contain DPPC lipids dispersed in water. From this system, we prepare several different systems by adding different type of fullerene-polymer complexes. As explained before, in the first step, we constructed a polymer-fullerene complex. In the later stage, fullerene-polymer complexes (described in the Table 1) were added to lipid-water system, separately. The details of polymer blocks are described below.

In a prior work, the hydrophilic polymer block was parameterized to represent polyethylene glycol (PEG) while the hydrophobic polymer parameters represent polyethylethylene (PEE) in that study [14]. In this study, we use similar diblock copolymers composing of *A and B* copolymers. *A* represents hydrophobic polymer while the hydrophilic polymers were represented by *B.*

During the construction of simulation system, molecules were randomly dispersed in the initial stage. Typical simulation system contained a total of 810 *DPPC* lipids, 27 fullerene-polymer complexes, and 56700 CG water molecules, corresponding to a total number of 67095 CG sites. CG water molecules represent a loosely bound, four atomistic water molecules [27]. This system was minimized using conjugate-gradient method before equilibrating for 1ns. All the simulations were conducted using GROMACS software [28,29] in an NPT ensemble with a 10fs time step. Due to the relatively smother potential employed, CG simulations allow the usage of such large timesteps. All the simulations were carried out at 300K temperature and 1 atm pressure. Production simulation of this system was carried out for at least 200ns. Previous simulation studies based on MARTINI force-field emphasized that the CG timescales are much larger than the corresponding atomistic timescales. Nevertheless, all the times reported in this work are actual simulation duration in nanoseconds, not the adjusted coarse-grain timescales.

Results

In the following we describe the simulation results for each of the system we have studied. We describe the results from single polymer attached fullerene systems followed by the amphiphilic fullerene molecules.

(i). Hydrophobic polymer-fullerene (F-A9) complex:

A single hydrophobic polymer attached fullerene molecules (F-A9) self-assembly and interaction with lipid molecules were simulated. This system initially formed a spherical micelle of lipid molecules. Simulations show that the hydrophobic polymer linked to fullerene favors the

interaction with hydrophobic lipid tails and places the fullerene molecules within the core of the lipid micelle as shown in Figure 5(A). However, this initial micelle structures was found to gradually transform into a stable bilayer structure as shown in Figure 5(B).

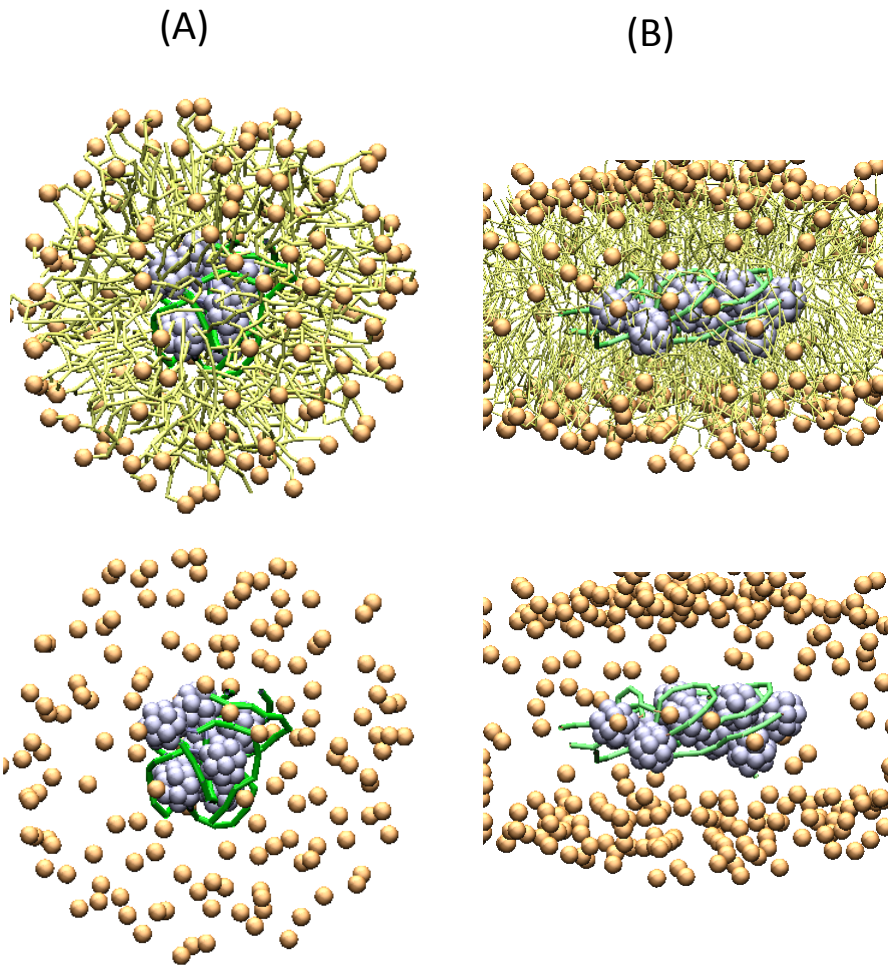


Figure 5: Hydrophobic polymer attached fullerene (F-A9) and DPPC lipid assembly snapshots are shown. Water is now shown for clarity. (A) Intermediate spherical micelle and (B) Final stable bilayer structure are shown. Color: lipids-yellow; fullerene-blue; polymer-green.

(ii). Hydrophilic polymer-fullerene (F-B9) complex:

Simulations show that a single hydrophilic polymer attached to fullerene molecule (F-B9) places it near the polar lipid-head group region of the lipid bilayer. This can be attributed to the favorable interaction between polar head groups and hydrophilic polymers. However, the

fullerene molecule, being hydrophobic, would want to move into the bilayer core. Nevertheless, a single polymer attachment was sufficient to stop the fullerene moving into the bilayer core. We found an initially formed spherical micelle gradually transforming into bilayer structures, as before. By varying the length of the hydrophilic polymer, we found that the length of the polymer has minimal effect on the fullerene molecule placement within the bilayer.

(iii). Amphiphilic-fullerene (F-AxBy) complexes:

Amphiphilic fullerenes were created by attaching a hydrophilic and a hydrophobic polymer to the either side of the fullerene cage. For this case, we have studied three different amphiphilic combinations: F-A3B9, F-A9B3 and F-A9B9. In all the cases we find, irrespective of the length of the individual polymer blocks, the fullerene molecules were found to be placed near the bilayer interface and away from the bilayer core center. These results indicate the repulsion between hydrophilic polymer and lipid bilayer core plays an important role in determining the location of the fullerene molecules within the bilayer. Final snapshots for each of these cases are shown in Figure 6.

(iv). Amphiphilic block copolymer-fullerene (F-(A5B4)2) complexes:

Instead of single polymers, in this case, amphiphilic diblock copolymers were linked to the either side of the fullerene surface. In other words, each attachment carried an amphiphilic diblock copolymer. The hydrophobic blocks were in direct contact with the fullerene surface. The self-assembly in this case resulted in the formation of bilayer as before. However, the fullerene molecules were found in the middle of the bilayer core and away from the bilayer interface. Despite the presence of hydrophilic polymers, the fullerenes were still able to move the bilayer core in this case as shown in Figure 6. A close examination reveals that the hydrophobic polymers within the diblock structure, acts as spacers between fullerenes hydrophilic blocks., thereby avoiding the unfavorable interaction between the lipid tails and hydrophilic polymers. This result also suggests, by adjusting the length of the the spacer (i.e. the hydrophobic block), fullerene placement within the bilayer can be precisely controlled. Such controllability may play important role in designing fullerene based drug delivery vehicles.

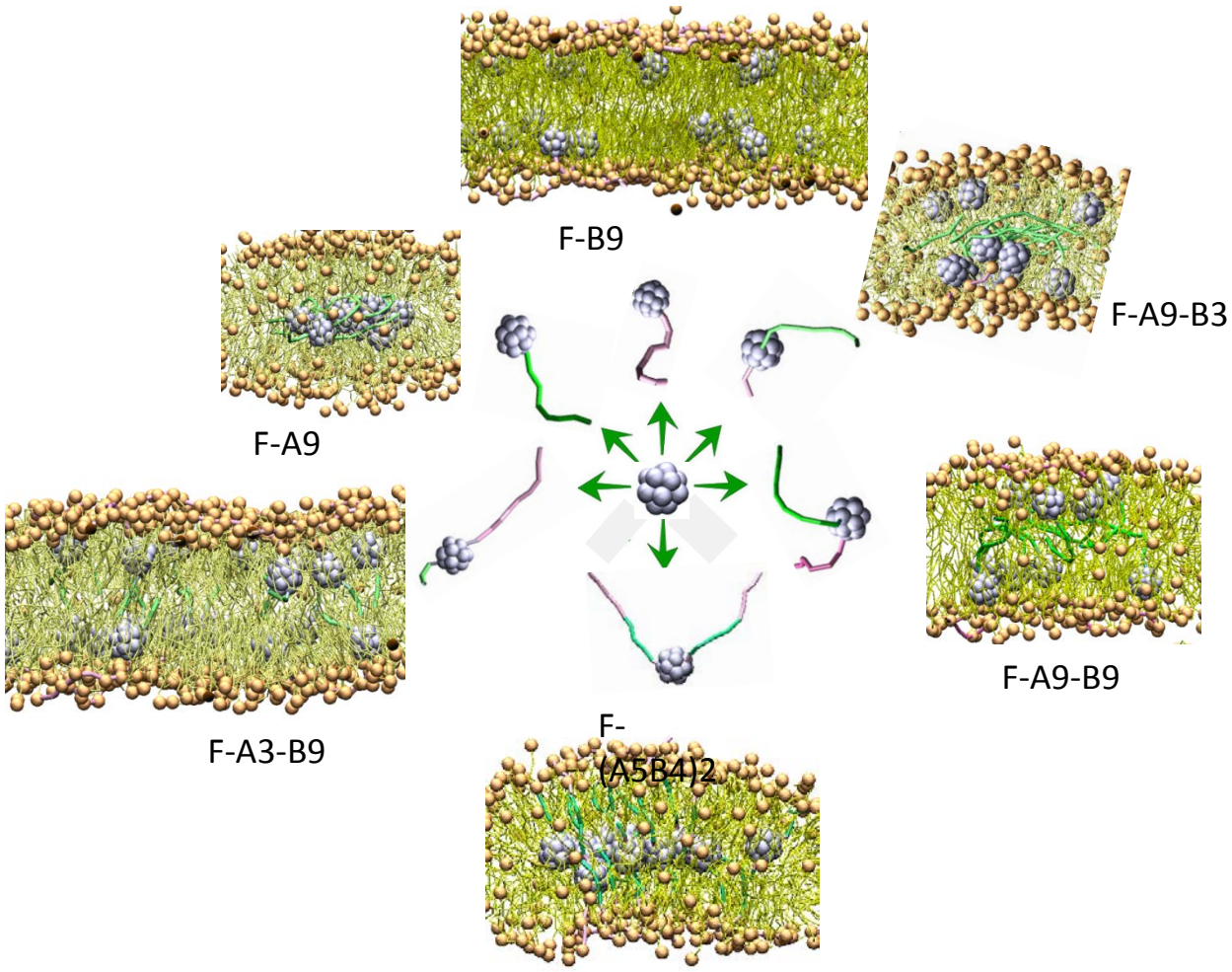


Figure 6: Final snapshots obtained from simulation studies of various polymer attached fullerene complexes. Water molecules are not shown. Color code is same as Figure 5.

Conclusions

In this work, we have conducted a detailed coarse grain (CG) MD simulation study of amphiphilic-fullerene insertion in a biological lipid bilayer. For this purpose, we have modeled fullerene molecule as a 16 coarse grain site spherical cage. The amphiphilic fullerene molecules were created by attaching one or more hydrophilic/hydrophobic polymers to the surfaces of the

fullerene molecule. By dispersing such amphiphilic fullerene molecules in water along with the DPPC lipids, their interaction and placement within the lipid assemblies were monitored. Previous experimental and simulation studies showed that the fullerene molecules tend reside in the hydrophobic core region of the lipid bilayer. However, present simulations reveal that by linking with the appropriate polymer blocks to the fullerene surface, it is possible to determine the placement of the fullerenes within the bilayer. For example, we found that by attaching a single hydrophilic block was sufficient to place the fullerene molecules near the lipid bilayer interface region, rather than in the core region. Multiple simulations with varying types of polymers attached fullerenes showed that the length of the hydrophilic polymer has no significant role in the fullerene placement. Both shorter and longer hydrophilic polymers were effective in placing the fullerenes near bilayer interface. On the other hand, when amphiphilic diblock copolymers were linked to the fullerene surface, hydrophobic polymer block acts as a spacer between fullerene and hydrophilic polymers. In this case, the fullerene molecules were placed in the hydrophobic core region of the bilayer. Simulations show that in this case, it is possible to precisely determine the placement of the fullerenes simply by adjusting the length of the hydrophobic block.

The present study establishes the basis for the computational studies of fullerene based drug delivery vehicles. Insights obtained from this study can be utilized to study transportation and delivery of real drug molecules. We are currently developing coarse grain models for anticancer drug molecule doxorubicin and mitoxantrone. In a next step, these drug molecules will be incorporated inside the amphiphilic fullerene-based drug carrier vehicles to study their transportation and delivery. We are currently investigating amphiphilic fullerene assisted lipid vesicle assembly as well. Together these studies will be able to provide deeper insights towards efficient and simultaneous delivery of multiple drug components.

Acknowledgement

This work was supported in parts by the U. S. Army Research Office under award/contract no. W911NF-11-1-0168. We would like to acknowledge the use of high performance computational facilities at the North Carolina State A&T University during the course of the present study.

References

1. G. Kwon, M. Naito, M. Yokoyama, T. Okano, Y. Sakurai, K. Kataoka; Block copolymer micelles for drug delivery: loading and release of doxorubicin**,** *J. Control. Release*, **1997**, *48*, 195–201.
2. K. Kataoka, G.S. Kwon, M. Yokoyama, T. Okano, Y. Sakurai, Block copolymer micelles a vehicles for drug delivery *J. Control. Release*, **1993**, *24*, 119-132.
3. Harada, A., Kataoka, K. Supramolecular assemblies of block copolymers in aqueous media as nanocontainers relevant to biological applications, *Progress in Polymer Science*, *2006, 31,* 949–982.
4. Attia, E.; Bte, A.; Ong, Z.Y.; Hedrick, J.L.; Lee, P.P.; Ee, P.L.R.; Hammond, P.T.; Yang, Y-Y., Mixed micelles self-assembled from block copolymers for drug delivery. *Curr. Opin. Colloid Interface Sci.* **2011**, *16*, 182-194.
5. Kim, B.-S.; Lee, H.-i.; Min, Y.; Poon, Z.; Hammond, P. T. "Hydrogen-Bonded Multilayer of Ph-Responsive Polymeric Micelles with Tannic Acid for Surface Drug Delivery". *Chem. Comm.* **2009**, 4194-4196.
6. Farokhzad, O. C.; Langer, R. Impact of Nanotechnology on Drug Delivery ACS Nano **2009**, *3*,16– 20.
7. Tyrrell Z.L., Shen,Y. Radosz, M. Fabrication of micellar nanoparticles for drug delivery through the self-assembly of block copolymers, *Progress in Polymer Science,* 2010, *35,* 1128–1143.
8. Liu, M.; Kono, K.; Frechet, M.J.; Water-soluble dendritic unimolecular micelles: Their potential as drug delivery agents; *Journal of Controlled Release*, **2000**, *65*, 121-131.
9. K. Kataoka, T. Matsumoto, M. Yokoyama et al., Doxorubicin-loaded poly(ethylene glycol)–poly(β-benzyl-l-aspartate) copolymer micelles: their pharmaceutical characteristics and biological significance *Journal of Controlled Release*, **2000,** *64*, 143–153.
10. Geng, Y., Dalhaimer, P., Cai, S.; Tsai, R.; Tewari, M.; Minko, T.; and Discher, D.; Shape effects of filaments versus spherical particles in flow and drug delivery,; *Nature Nanotechnology*, **2007**, *2*, 249-255.
11. Ahmed, F.; Pakunlu, R.I.; Srinivas, G.; Brannan, A.; Bates, F.; Klein, M.L.; Minko,T.; and Discher, D. E.; Shrinkage of a Rapidly Growing Tumor by Drug-Loaded Polymersomes: pH-Triggered Release through Copolymer Degradation. *Molecular Pharmaceutics*, **2006,** *3,* 340-350.
12. Yih, T. C.; Al-Fandi, M. Engineered Nanoparticles as Precise Drug Delivery Systems *J. Cell. Biochem.* **2006**, *97*, 1184– 1190; Kwon, G.S.; Polymeric micelles for delivery of poorly water-soluble compounds**,** Crit. Rev. Ther. Drug Carr. Syst., **2003**, 20, 357.

13. Polymer micelle assisted transport and delivery of model hydrophilic components inside a biological lipid vesicle: A coarse grain simulation study, G. Srinivas, R. Mohan, A. Kelkar, Journal of Physical Chemistry B, *117,* 12095 (2013).
14. Soft Patchy Nanoparticles from Solution-phase Self-assembly of Binary Diblock Copolymers G. Srinivas, and J.W. Pitera, Nanoletters, *8*, 611 (2008).
15. Brettreich M, Hirsch A: A highly water-soluble dendro[60]fullerene. *Tetrahedron Lett* 1998, 39:273-234. Brettreich M, Burghardt S, Bottcher C, Bayerl T, Bayerl S, Hirsch A:
16. Globuläre amphiphile: membranbildende Hexaaddukte von C60. *Angew Chem* 2000, 112:1915-1918.
17. Brettreich M, Burghardt S, Bottcher C, Bayerl T, Bayerl S, Hirsch A: Globular amphiphiles: membrane-forming hexaadducts of C(60). *Angew Chem Int Ed* 2000, 39:1845-1848.
18. Burghardt S, Hirsch A, Schade B, Ludwig K, Bottcher C: Switchable supramolecular organisation of tructurally defined micelles based on an amphiphilic fullerene. *Angew Chem Int Ed* 2005, 44:2976-2979.
27. Braun M, Atalick S, Guldi DM, Lanig H, Brettreich M, Burghardt S, Hatzimarinaki M, Ravanelli E, Prato M, Van Eldik R, Hirsch A: Electrostatic complexation and photoinduced electron transfer between Zn-cytochrome c and polyanionic fullerene dendrimers. *Chem Eur J* 2003, 9:3867-3875.
28. Maierhofer AP, Brettreich M, Burghardt S, Vostrowsky O, Hirsch A, Langridge S, Bayerl TM: Structure and electrostatic interaction properties of monolayers of amphiphilic molceules derived from C60 fullerenes: A film balance, neutron- and infrared reflection study. *Langmuir* 2000, 16:8884-8891.
19. Guldi DM, Zerbetto F, Georgakilas V, Prato M: Ordering Fullerene Materials at Nanometer Dimensions. *Acc Chem Res* 2005, 38:38-43.
20. Angelini G, De Maria P, Fontana A, Pierini M, Maggini M, Gasparrini F, Zappia G: Study of the Aggregation Properties of a Novel Amphiphilic C60 Fullerene Derivative. *Langmuir* 2001, 17:6404-6407.
21. Gan HY, Liu HB, Li YL, Gan LB, Jiang L, Jiu TG, Wang N, He XR, Zhu DB: Fabrication of fullerene nanotube arrays using a template technique. *Carbon* 2005, 43:205-208.
22. Liu Y, Xiao SQ, Li HM, Li YL, Liu HB, Lu FS, Zhuang JP, Zhu DB: Selfassembly and characterization of a novel hydrogen-bonded nanostructure. *J Phys Chem B* 2004, 108:6256-6260.
23. Georgakilas V, Pellarini VF, Prato M, Guldi DM, Melle-Franco M, Zerbetto F: Supramolecular self-assembled fullerene nanostructures. *Proc Natl Acad Sci* 2002, 99:5075.
24. Zhou S, Burger C, Chu B, Sawamura M, Nagahama N, Toganoh M, Hackler UE, Isobe H, Nakamura E: Spherical bilayer vesicles of fullerene-based surfactants in water: a laser light scattering study. *Science* 2001, 291:1944-1947.

25. Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. J. Phys. Chem. B **2007**, 111, 7812; http://md.chem.rug.nl/cgmartini/index.php/downloads/force-field-parameters
26. Berendsen, et al. *Comp. Phys. Comm.* **1995**, ***91,*** 43-56.
27. Lindahl, et al. *J. Mol. Model.* **2001, *7*:** 306-317.

3. Biomolecules in Binary Solvents: Computer Simulation Study of Lysozyme Protein in Ethanol-Water Mixed Solvent Environment

G. Srinivas, R. Mohan, A. Kelkar, North Carolina A&T State University

Journal Article: JSM Nanotechnology & Nanomedicine, Vol. 2, No. 2, 1029, 2014

ABSTRACT

Proteins are building blocks of biological systems and play an important role from the health and medical perspective in the drug reactions and efficiency. Proteins function well in their natural water solvent environments and are influenced by modified solvent environments such as alcohol. Effect of protein-solvent interaction on the protein structure is widely studied with experimental and computational techniques. However, molecular level understanding of proteins interaction with many solvents is still not fully understood. The present work aims to obtain a detailed understanding of solvent effect on lysozyme protein, using water, ethanol, and different concentrations of water-ethanol mixtures as solvents. We use detailed atomistic molecular dynamics simulations to study using GROMACS code. Compared to neat water environment, the lysozome structure shows remarkable changes in water-ethanol mixed solvent, with increasing ethanol concentration. Significant changes were observed in the protein secondary structure involving alpha helices. We found that increasing ethanol concentration results in a systematic increase in total energy, enthalpy, root mean square deviation (RMSD), and radius of gyration of lysozyme protein. A polynomial interpolation approach is presented to determine these quantities for any intermediate alcohol percentage, and compared with the values obtained from a full MD simulation. Results from MD simulation were in good agreement with those obtained from the interpolation approach. The polynomial approach eliminates the need for computationally intensive full MD analysis for the concentrations within the range (0-12%) studied.

1. Introduction

Protein molecules are important in the formation of most biological building blocks inside the cell. The structure of protein plays an important role in biological functions [1]. For example, proteins like collagen support the cell structure due to its coiled helical shape which is long, strong and stringy [2]. Hemoglobin, a globular protein with a folded and compact shape can maneuver through the blood vessel providing a function to supply oxygen to the body cells [3]. Proteins function well in their natural water solvent environments and other solvent environments such as alcohol alter the behavior of protein structures and influence the therapeutic drug interactions from a medical perspective, starting from their fundamental molecular structures. Previous experimental studies have shown that the alcohol exerts significant influence on the protein structure. Excess concentration of alcohol may denature the protein structure by disrupting the secondary structure of the protein. Such changes in the structure of the protein may also reduce the efficiency of drug reactions that target such proteins. For this very reason, alcohol intake is not advised along with medication. In order to gain further insights into the effect of alcohol on protein structure and dynamics, a deeper understanding of this phenomenon is needed. In this work, we try to obtain a molecular level picture of this problem by carrying out detailed molecular dynamics simulation studies.

Several experimental, theoretical and computational studies focused on understanding the structure-function relationship of proteins. Heat, acids and bases, reducing agents, alcohol etc. influence protein structure thereby its function. Ping *et al.* observed that the tertiary structure of lysozyme was destroyed as when the temperature increased to 80 ^{o}C that also affected protein's function [4]. Solvents also influence the protein structure to a great extent. For example, a protein adapts dissimilar structure in a hydrophobic solvent compared to that in water. Despite numerous experimental and simulation studies, protein behavior in mixed solvents such as ethanol-water mixtures has not been fully understood from a molecular viewpoint. In this work, we conduct extensive molecular dynamics simulation studies of a lysozyme protein in ethanol-water mixture at various concentrations of ethanol in water to understand the solvent influence on protein structure and dynamics. This provides insight into the fundamental understanding of the protein structure in an alcohol environment. The understanding of the variations in the protein molecular structure in an alcohol concentration environment can provide further insights into their molecular interactions and efficiency under therapeutic drugs. For this purpose we use GROMACS molecular dynamics simulation analysis code. We chose Lysozyme as it closely resembles protein structure found in humans. Secondly it is moderate in size for molecular dynamics simulation analysis compared to most other proteins. A cartoon representation of lysozyme protein (protein data bank file 1AKI.pdb) is shown in Figure 1.

Water plays an important role in maintaining cell membrane and enzyme activities acting as lubricant for protein movements in cells. More importantly, proteins need water to function and are their natural environment. Hence, we first simulate protein in water to understand the behavior of lysozyme in a water solvent environment. This is followed by the simulation of lysozyme in an ethanol-water mixed environment at different ethanol-water ratios. The hydroxyl group of ethanol (C_2H_5OH) can bond with hydrogen from other ethanol to make it less volatile and more viscous than lower polar organic compound with almost same molecular weight for example, propane. It is miscible with water and some other organic compounds. The focus of this work is to study and understand the effect of different solvent environments (ethanol-water mix) on protein. We anticipate an understanding of the effect of ethanol on lysozyme provides insights into similar other protein-solvent interactions as well. Several experimental studies have been carried out before to study the effect of ethanol on protein and the effect of water on protein with tangible results [5-8]. However, to our knowledge, the effect of ethanol and ethanol-water mixtures has not been studied computationally in detail. In this work, we explore the effect of ethanol on lysozyme by carrying out detailed atomistic molecular dynamics simulation studies.

In a series of simulation studies, Bagchi and coworkers examined transport properties of binary mixtures [9-11]. They have shown that the specific solute solvent interaction play important role in determining the properties of such solvents. Wensink *et al* [12] studied binary mixtures of alcohol and water using molecular dynamics simulation. They computed the shear viscosity using non-equilibrium molecular dynamics simulation. The diffusion constant was studied along the rotational correlation time, and was found that mobility correlates with viscosity data i.e. the viscosity is maximal at intermediate alcohol concentration [13]. It was found that at maximal viscosity, mobility was minimal. They combined viscosity and diffusion calculations to compute the effective hydrodynamic radius of the particles in the mixture using Stoke- Einstein relation [14]. The analysis indicated that there is no collective diffusion of molecular clusters in the mixture and the pure liquid. The present work examines lysozyme protein structure and dynamics in various alcohol-water mixtures by using series of molecular dynamics simulations at different alcohol-water solvent ratios. A brief background of the molecular dynamics analysis is presented next for completeness. The remainder of this article is organized as follows. Section 2 provides the details of the molecular dynamics modeling method employed in this study. Section 3 contains simulation results on protein structure in water, ethanol and water-ethanol mixed environments. Section 4 presents a detailed discussion and analysis of the results. We close the article with few concluding remarks in the summary section.

2. Molecular Dynamics Simulations

GROMACS is the MD analysis code employed in the present work. It is most commonly used open source software for molecular dynamics studies of materials and biological systems.

Importantly, it is one of the fastest MD codes among the open source codes for MD simulation that are currently available.

The force field and the initial configuration can be used to estimate or calculate the motion and position of the particles/atoms in a molecular system. The sum of the intermolecular interaction and intra-molecular interaction is equal to the total potential energy of the system. In the following, we describe these two interactions in detail.

The bond stretching and bond bending energy equation is based on Hooke's law [15].

$$E_b = \sum k_b (r - r_0)^2 \tag{1}$$

where E_b is bond energy, K_b is bond-interaction constant, and r_0 is the equilibrium position length between two bonded atoms.

$$E_\theta = \sum k_\theta (\theta - \theta_0)^2 \tag{2}$$

where E_θ is bond-angle energy, K_θ is the corresponding force constant, θ_0 is the equilibrium angle.

The interaction of the nonbonded molecules were modeled using the Lennard Jones potential and the columbic potential [16] as described in the following equation

$$v_{nonbonded(i,j)} = v_{lernad(i,j)} + v_{coulomb(i,j)} \tag{3}$$

where v is the potential for nonbonded interaction between atoms of different molecules i and j with the total potential is given as the sum of the coulomb potential and the Lennard Jones potential

$$v_{coulumb(i,j)} = \frac{1}{4\pi\varepsilon_0}\frac{q_i q_j}{r_{ij}} \tag{4}$$

$$v_{lernard(i,j)} = 4\varepsilon_{ij}\left(\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right) \tag{5}$$

ε_0 is the permittivity of free space, $q_{(i,j)}$ are charges, ε is the energy parameter with reference to the depth of the potential well, σ_{ij} is the finite distance for which inter-particle potential is zero. In the present study, we use the OPLS forcefield parameters that are widely accepted for the biomolecular simulation studies.

2.1 Molecular Dynamics Simulation of Lysozyme Protein in Water

We begin with simulating lysozyme protein in neat water environment, which has been extensively studied both by experimental and computational techniques. For this purpose, we analyze various thermodynamic quantities in addition to structural parameters such as radius of gyration and RMSD (Root Mean Square Deviation) for the molecular system. This initial study helps us verify the consistency of our simulation method and also serves as reference to compare and contrast the protein structure and its dynamics under other solvent environments such as the ethanol-water mixture focused in the present work.

The protein structure obtained from PDB was solvated with TIP3P [17] water and the whole system was minimized using steepest descent method. All computations were performed using in-house computation cluster on the NC A&T campus. Both NVT and NPT equilibration simulations were carried out for 200ps with a 1fs time step. We used temperature and pressure values of 300K and 1 bar, respectively. The final configuration from NPT equilibration was taken as the starting structure for the production MD.

The simulation system contained one lysozyme protein molecule of 129 residues with 1960 atoms and 12,365 water molecules (37,095 atoms). The system contained a total of 39,055 atoms. We used the Berendsen thermostat [18] to control the temperature. For the MD dynamics production simulation, the system was run for 50 million time steps corresponding a total duration of 50ns.

The structure of the protein was examined with commonly used structural properties such as Radius of gyration (R_g) root mean square deviation (RMSD) etc. which are described below.

Radius of gyration helps understand the compactness of the protein structure and is given by [15],

$$R_g = \left(\frac{\sum_i \|r_i\|^2 m_i}{\sum_i m_i} \right)^{\frac{1}{2}} \qquad (6)$$

m is the mass of the atom r_i and i are the positions specific to a particular atom with the reference point being the center of mass [19]. The R_g value for the lysozyme protein obtained from present simulation study is shown as a function of time in figure 2(A). As shown, the R_g value fluctuates around an average value of 1.42nm throughout the simulation, suggesting that the protein maintains relative compact size during the entire dynamics simulation time.

The RMSD (root-mean square deviation) is the measure of the average distance between the atoms of the back bone of superimposed proteins. The RMSD can be calculated with the equation [20],

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2} \quad (7)$$

RMSD could be used for quantitative comparison between the structure of the native state of protein and its partially folded state. For the present lysozyme-water system, RMSD value for the entire simulation time (50ns) is shown in Figure 2(B). The average RMSD value is less than 1nm, indicating a relatively folded structure for the protein in this natural water environment, consistent with the existing simulation studies [21]. Combined with the result of R_g for the protein in the present study indicates a compact structure for lysozyme in water during the entire simulation, in good agreement with earlier simulation analysis reported in the literature [22].

The next sections discuss the lysozyme protein in different solvent conditions that include ethanol, and ethanol-water mixture and comparison with the data from pure water environment.

2.2: Molecular Dynamic Simulation of Protein in Ethanol

This section focuses on the analysis of lysozyme in ethanol solvent environment. Ethanol is miscible with water. It is also known to have profound effect on proteins and biomolecules. This in turn could have significant impact on the interaction with and efficiency of therapeutic drugs. The behavior of lysozyme in a 100% ethanol solvent condition is studied here. For this purpose, lysozyme protein was solvated with 100% ethanol. The lysozyme-ethanol molecular system employed in the present study contained one lysozyme protein (1,960 atoms) and 2,289 ethanol molecules (20,601 atoms), resulting in a total of 22,561 atoms in the simulation system. This system was simulated for 50ns with a 1fs time step size.

At the end of the simulation we analyzed and compared the thermodynamic quantities following the same approach as in the case of water environment. The initial and final structures of lysozyme in ethanol are shown in figure 3(A) and 3(B), respectively. As shown, the protein structure showed significant changes at the end of the dynamics simulation in a full ethanol environment. The final structure appears expanded/swollen compared to the initial structure. For comparison, in Figure 3(C), we also show the final structure obtained from lysozyme in water simulations.

From Figure 3, lysozyme in ethanol appears to be swollen compared to that of water. A closer examination reveals significant difference in the protein alpha-helix content in ethanol compared to water. The alpha-helix structures were broken into relatively shorter helices in ethanol compared to water. In other words, longer helix components were no longer stabilized

when solvent environment was changed to ethanol. One of the factors that stabilize protein secondary structure is the hydrogen bonding between protein and solvent. As the solvent changed from water to ethanol, the hydrogen bond network between protein and water, responsible for stabilizing alpha-helix structure was broken in case of ethanol. This led to the instability of protein secondary structure in ethanol solvent. To further confirm our findings, we calculated and compared thermodynamic quantities such as total energy and enthalpy and structural quantities such as radius of gyration (R_g) and RMSD. The results of these quantities are compared with the results from water environment.

The total energy of the lysozyme protein system in water was -443,901 (KJ/Mol) and that of ethanol is -59,164(KJ/Mol). This significant energy difference clearly indicates that the lysozyme protein in water system is more stable compared to the same protein in the ethanol system. This is in accordance with the significant change in the final protein structure in ethanol and water as shown in figures 3(B) and 3(C). To further verify our results we analyzed the compactness of the protein structure by plotting the radius of gyration as a function of time as shown in the Figure 5.

We find the radius of gyration of protein in water is approximately 1.4 nm as shown in figure 4(B) and that of ethanol in figure 4(A) is approximately 3.12 nm. The significant difference between the R_g values indicates a significant change in the compactness of the protein. Hence, it can be inferred that the protein molecule is swollen in ethanol compared to that in water. The structural stability of protein molecule in water and ethanol was compared by analyzing the root mean square deviation (RMSD) as shown in Figure 5.

From figure 5, the RMSD for protein in water and ethanol were found to be approximately 0.162 nm and 4.15nm respectively. This is a significant change in RMSD value of the lysozyme in ethanol compared to that of water. In their folded native structure, proteins typically have RMSD values of about 0.1 to 0. 2 nm [23-24]. RMSD value as high as 4.15nm shows a significant deviation of the lysozyme protein from its native structure in an ethanol solvent environment.

The simulation results showed an increase in total energy, enthalpy, radius of gyration and root mean square deviation for lysozyme in ethanol compared to water. The significant increase in the calculated quantities affirms the changes in the stability of the system and compactness of the protein structure. Simulations thus reveal marked changes in protein structure and energy when solvated in ethanol compared to that in water. In order to quantify such changes in a systematic fashion, we proceed to study the effect of ethanol and water mixtures on lysozyme protein. For this purpose we use similar simulation setup as discussed in the previous chapters to create and study lysozyme protein at different ethanol concentrations in water. In the present study, low ethanol concentrations in the range of 0–12% in the ethanol–water mixtures are considered. Lysozyme protein behavior in the ethanol-water mixture environment at various percentages is compared to that of pure water and 100% ethanol environments, and is presented next.

2.3. Molecular Dynamics Simulation of Lysozyme Protein in Ethanol-Water Mixtures

In the previous sections, we have described the effect of pure water and pure ethanol on lysozyme protein. Here we conduct MD simulation of protein in ethanol-water mixtures over a range of ethanol concentration in a systematic fashion. For this purpose we set up six different ethanol concentrations of approximately 2%, 4%, 6%, 8%, 10% and 12%. This concentration range of ~2 to ~12% was selected based on the prior experimental investigations in the literature [25]. We aim to qualitatively analyze the lysozyme protein behavior as a function of ethanol concentration and to understand the changes with the addition of low concentration of ethanol to its natural water environment. We begin the simulation set up as before but solvating the protein with both water and ethanol in specified compositions.

MD analysis was conducted as before by solvating the lysozyme protein in both ethanol and water molecules in appropriate ratios so as to obtain different concentrations of ethanol. The details are shown in Table 1. We chose approximate percentage based on the mass of ethanol molecules and water molecules for the required percentages of water and ethanol mixture. We equilibrated the lysozyme in each ethanol-water mixture system and conducted full simulation for 50 ns with a time step of 1 fs, at a pressure of 1 bar, and a temperature of 300K. At the end of the simulation process, we calculated and compared the thermodynamic and structural quantities similar to the quantities calculated as before. The initial and final structures after the MD analysis of the lysozyme protein at different ethanol-water mixture concentrations are shown in figure 6. In this Figure only the protein is shown while water and ethanol are not shown for the clarity purpose. As shown in Figure 6, noticeable changes in the initial and final structure of protein at different water -ethanol concentrations can be found.

In order to understand the role of solvent, in particular that of ethanol, we have tracked the ethanol molecules motion over the course of simulation. Corresponding snapshots showing initial and final configurations for different concentration of ethanol-water mixture are shown in Figure 7. This Figure shows that the diffusion of ethanol molecules into the lysozyme causing the protein molecule to swell as the protein gets increasingly destabilized with increase in ethanol concentration. Both lysozyme and ethanol molecules are shown in this figure. Together from Figures 6 and 7, we notice that the protein secondary structure gets altered as ethanol progressively replaces water molecules with increasing ethanol concentration. This leads to the decrease in alpha helical content of the protein in the same direction. Together, these observations reveal that the protein increasingly gets deviated from its native folded structure as the ethanol concentration increases. To further verify our observation, we calculated and compared the thermodynamic quantities starting with the total energy as presented in Figure 8. An increase in the total energy of the system with increase in the concentration of ethanol is

noticed. This also indicates decrease in the stability of the system as the concentration of ethanol increases.

We observed a significant difference in the key structural and thermodynamic quantities of protein in pure water and that in ethanol. This is shown in Figure 8. In the range of ethanol-water mixture concentrations studied, we observed a trend of increase in the energy and enthalpy with increase in ethanol concentration. This increase in thermodynamic quantities further supports the observed swelling in protein structure. The dynamic variation in the key parameters over the MD analysis time duration is plotted in Figure 8. The dynamic changes also showed similar behavior that of the time averaged values for the key parameters studied.

Figure 8 shows the average thermodynamic quantities as a function of mole fraction. From this figure we notice a significant increase in all the calculated quantities during the progressive passage from pure water to pure ethanol with higher absolute values at 100% ethanol. These changes further confirm the observed decrease in protein stability with increase in ethanol concentration. The summarized time average values of the thermodynamic quantities of protein in different concentration of ethanol are presented in Table 2.

3. Polynomial interpolation method

The present computational analysis took an average of eighty four hours of computing time for each concentration, with thirty six processors, using GROMACS software on the multi-processor computing system at North Carolina A&T State University (HERMES). To obtain the thermodynamic and structural quantities for any new percentages, a complete simulation will be required. Such a complete simulation would require a significant computing time and resources. However, based on our present analysis, a relatively smooth variation of the time averaged values of the key parameters is clearly noticed. This could allow one to potentially interpolate the required values from the present generated data for a different ethanol concentration within the range of ethanol concentrations studied. Based on these observations, we propose and present an interpolation methodology for the quantitative key parameters studied as an alternative way by which the need for additional computer simulation and/or experiments can be avoided. The interpolation approach uses our simulation data and interpolating to obtain the unknown values for another ethanol concentration percentage that is within the range of present study (0 to 12%). The effectiveness of this interpolation methodology was tested as follows.

1. Select an intermediate percentage that was not used in the simulations. For this purpose, we selected ~7% ethanol concentration.

2. Using an appropriate interpolation method and exiting values in Table 2, calculate the interpolated values of the key parameters at this intermediate ethanol – water concentration range (7% ethanol case).
3. Subsequently, a complete MD analysis was performed at this concentration level and compared to the obtained interpolated values.

3.1. Using Polynomial method to interpolate the intermediate ~7% ethanol

In order to quantitatively determine the behavior of simulated quantities as a function of ethanol concentration, we have used a polynomial fitting as explained below. The following second order polynomial equation was used to fit the simulated quantities,

$$C_2 x^2 + C_1 x + C_0 = f(x) \quad (8)$$

where C_0, C_1, and C_2 are constants. The resulting fits for simulation results are shown in Figure 9. The values of these constants obtained by fitting to the simulation results for all thermodynamic and structural quantities are presented in Table 3.

We generated our data using equation 8.

Using the coefficients C_0, C_1 and C_2 one can determine the above mentioned thermodynamic and structural quantities at any ethanol concentration (from 0-12%), without carrying out the full scale additional actual simulations. In order to verify this approach, we selected an intermediate ethanol concentration that has not been used in fitting procedure. By using the polynomial equation (Eq 8) we first theoretically determine the thermodynamics and structural quantities. In the next step, we conducted a complete simulation study as before for the selected ethanol concentration. The simulation results were then compared with that obtained from polynomial fitting procedure. We have selected ~7% ethanol as our test case, which was not used in simulations.

Figure 9 shows the fitted interpolation of simulated quantities as a function of ethanol concentration. For the analyzed quantities: total energy, enthalpy, radius of gyration and RMSD. For the intermediate percentage of ethanol (~7%) chosen is also plotted in each figure, indicated by the red symbol (Figure 9). Table 4 shows the interpolated values obtained for the intermediate concentration of 7% ethanol using polynomial method.

3.2 Validating the Results from Interpolation Method

A full MD analysis run at approximately 7% ethanol in ethanol – water mixture was employed to compute the key parameters of energy, enthalpy, R_g and RMSD and compared with the corresponding interpolated value at this intermediate percentage. Actual percentage of 7.153 percent by mass of ethanol was employed. The system was set up similar to the other percentages we studied before. The details of the lysozyme-ethanol-water mixture simulation system are presented in Table 4. The comparison between calculated thermodynamic quantities and key parameters from the present simulation is presented in Table 4. To further validate the interpolation result for 7.153%, we computed the percentage error between the interpolated values and the simulated values for the 7.153% ethanol concentration as shown in Table 4. The values from the interpolated result and the actual simulation are in good agreement. The error margin found to be within the acceptable range. Such good agreement validates the interpolation method for any other percentage within our initial percentage range which is from 0 to 12 percent ethanol. This is a remarkable result, since it can potentially avoids the need for computational or experimental procedures for other concentration values within the concentration range studied in this work.

The simulations of protein at different concentration of ethanol in the range of ~2% - 12% ethanol in ethanol-water mixture reveal that the most of thermodynamics and structural quantities show an increase with increasing ethanol concentration, thereby revealing the destabilization of folded native protein structure with increasing ethanol concentration. The analysis of structural quantities such as RMSD and radius of gyration revealed the protein structural deviation from folded state. Together, these results demonstrate a uniform trend in increase with increasing ethanol concentration. Hence, we proceed to quantify this behavior as a function of ethanol concentration. For this purpose, we obtained optimal polynomial fit for each thermodynamic and structural quantity as a function of ethanol concentration. By using such polynomial expression along with the determined coefficients, we could obtain the results for any arbitrary concentration that is within the range of 0-12% ethanol without a need for additional full scale simulation analysis.

4. Conclusion

In the present work, we have presented a detailed study of structure and dynamics of lysozyme protein in water, ethanol and water-ethanol binary mixtures by conducting extensive computational molecular dynamics simulation studies. In each case, we performed detailed molecular dynamic simulation and analysis on the following thermodynamic and structural quantities of lysozyme: total energy, enthalpy, radius of gyration and RMSD. MD analysis

studies were carried out using GROMACS MD simulation code. We carried out the simulation process for protein in water environment and performed thermodynamic and structural analysis by calculating the total energy, enthalpy, radius of gyration, and root mean square deviation. The results from the analysis showed that protein was relatively stable in water environment, without showing significant deviations from its native folded structure. We proceed to simulate and analyze protein in pure ethanol under the same thermodynamic conditions. At the end of the analysis, we observed a significant change in protein structure between the water environment and the ethanol environment. We find the protein molecule relatively swollen in ethanol solvent compared to that in water environment. To further understand the effect of solvent on protein structure in more systematic fashion, we performed full simulations on different percentages of ethanol-water mixture (2%, 4%, 6%, 8%, 10%, and 12%) and carried out similar thermodynamic analysis as before. We observed changes in protein molecule with increase in ethanol concentration as the molecule seems to increase in size based on the visual structure of protein obtained from VMD and thermodynamic analysis. We observed a trend in the variation of the derived thermodynamic quantities analyzed for the various ethanol percentages studied. Based on this variation, we proceed to check the possibility of utilizing interpolation method for intermediate percentage within our range of percentages. We chose an intermediate percentage of seven percent for this purpose. With polynomial interpolation method, we were able to calculate the values of the thermodynamic quantities for seven percent using existing data from our previous simulation. In order to verify these results, we constructed a system of lysozyme protein in ~7% ethanol and performed a full scale molecular dynamic simulation. The results from the full simulation were compared with the interpolated results. We found both were in good agreement within the level of acceptable error.

Based on our simulation findings we conclude that the ethanol has a significant effect on lysozyme protein structure. The deviation of protein structure from its native environment suggests that the protein molecule is likely to function better in water environment compared to ethanol environment. We envisage, such molecular level insights into protein-solvent interactions can be used as guidelines in studying similar other protein-solvent interactions as well.

One of the main contributions of this work is that for subsequent percentages within a known range of percentage values, interpolation method can be successfully used to obtain the results. As demonstrated in the present work, it eliminates need for the large computing time and resources involved in such full scale simulations.

In the present work we have simulated and analyzed the dynamics of protein in ethanol-water solvent. With the knowledge acquired from this work, future studies can explore the effect of other components on proteins and other biological molecules employing the MD analysis methodology. Along these lines, one can study the effect of anesthetic molecules on proteins as well. Despite the routine usage of anesthetic molecules in medical and surgical procedures, this

procedure is not devoid of side-effects. Understanding the molecular level interaction between anesthetic molecules and proteins may provide further insights on the side effects. Similar to the study here, nature and concentration of anesthetic molecules may affect the protein structure thereby its function and provide practical extension of the present work.

Acknowledgments:

The financial support in part by U.S. Army Research Office through award/contract no. W911NF-11-1-0168 is acknowledged. We would also like to acknowledge the use of high performance computational facilities at North Carolina A&T state University during this study.

REFERENCES

1. Hunt, N.T., et al., *The dynamics of water-protein interaction studied by ultrafast optical Kerr-effect spectroscopy.* Journal of the American Chemical Society, 2007. **129**(11): p. 3168-3172.
2. Branden, C. and J. Tooze, *Introduction to protein structure*. Vol. 2. 1991: Garland New York.
3. Dunker, A.K., et al., *Intrinsic disorder and protein function.* BIOCHEMISTRY-PENNSYLVANIA THEN WASHINGTON-, 2002. **41**(21): p. 6573-6582.
4. Yanni, N.P.C.B.W.Y.W.J.L., *Analysis and Simulation of Molecular Dynamics of Lysozyme in Water Cluster System.* DOI 10.1007/s12209-012-1775-9, 2012. **Vol.18 No.1 2012**(DOI 10.1007/s12209-012-1775-9): p. 001-007.
5. Preedy, V. and T. Peters, *Acute effects of ethanol on protein synthesis in different muscles and muscle protein fractions of the rat.* Clinical science (London, England: 1979), 1988. **74**(5): p. 461.
6. 8. TIERNAN, J.M. and L.C. WARD, *Acute effects of ethanol on protein synthesis in the rat.* Alcohol and Alcoholism, 1986. **21**(2): p. 171-179.
7. 9. Zaks, A. and A.M. Klibanov, *The effect of water on enzyme action in organic media.* Journal of Biological Chemistry, 1988. **263**(17): p. 8017-8021.
8. 10. Wang, J.H., *Theory of the self-diffusion of water in protein solutions. A new method for studying the hydration and shape of protein molecules.* Journal of the American Chemical Society, 1954. **76**(19): p. 4755-4763.
9. Srinivas, G.; Mukherjee, A.; Bagchi, B.; Non-ideality in the composition dependence of viscosity in binary mixtures, *Journal of Chemical Physics,* 2001, *114*, 6220.
10. Mukherjee, A., Srinivas, G., Bagchi, B.; Reentrant behavior of relaxation time with viscosity at varying composition in binary mixtures, *Physical Review Letters,* 2001, *86,* 5926.
11. Srinivas, G, Bhattacharya, S., Bagchi, B.; Computer simulation and mode coupling theory study of the effect of specific solute-solvent interactions on diffusion: Crossover from a sub-slip to super stick limit of diffusion. *Journal of Chemical Physics. 1999,***110**; p. 4477.
12. Wensink, E.J.W., et al., *Dynamic properties of water/alcohol mixtures studied by computer simulation.* The Journal of chemical physics, 2003. **119**: p. 7308.
13. Wensink, E.J.W., et al., *Dynamic properties of water/alcohol mixtures studied by computer simulation.* The Journal of chemical physics, 2003. **119**(14): p. 7308-7317.
14. Bagchi, B.; *Molecular relaxation in liquids,* Oxford University Press Inc. Oxford Newyork. 2012.

15. Müller-Plathe, F., *YASP: A molecular simulation package.* Computer physics communications, 1993. **78**(1-2): p. 77-94.
16. Jorgensen, W.L. and C. Jenson, *Temperature dependence of TIP3P, SPC, and TIP4P water from NPT Monte Carlo simulations: Seeking temperatures of maximum density.* Journal of Computational Chemistry, 1998. **19**(10): p. 1179-1186.
17. Lemak, A. and N. Balabaev, *On the Berendsen thermostat.* Molecular Simulation, 1994. **13**(3): p. 177-187.
18. Newcomer, M., B. Lewis, and F. Quiocho, *The radius of gyration of L-arabinose-binding protein decreases upon binding of ligand.* Journal of Biological Chemistry, 1981. **256**(24): p. 13218-13222.
19. Lobanov, M.Y., N. Bogatyreva, and O. Galzitskaya, *Radius of gyration as an indicator of protein structure compactness.* Molecular Biology, 2008. **42**(4): p. 623-628.
20. Carugo, O. and S. Pongor, *A normalized root-mean-spuare distance for comparing protein three-dimensional structures.* Protein Science, 2008. **10**(7): p. 1470-1473.
21. Bowman, G.R. and V.S. Pande, *The roles of entropy and kinetics in structure prediction.* PloS one, 2009. **4**(6): p. e5840.
22. Dadarlat, V.M. and C.B. Post, *Insights into protein compressibility from molecular dynamics simulations.* The Journal of Physical Chemistry B, 2001. **105**(3): p. 715-724.
23. García, A.E. and J.N. Onuchic, *Folding a protein in a computer: an atomic description of the folding/unfolding of protein A.* Proceedings of the National Academy of Sciences, 2003. **100**(24): p. 13898-13903.
24. Dokholyan, N.V., et al., *Topological determinants of protein folding.* Proceedings of the National Academy of Sciences, 2002. **99**(13): p. 8637-8641.
25. Onori, G. and A. Santucci, *Dynamical and structural properties of water/alcohol mixtures.* Journal of molecular liquids, 1996. **69**: p. 161-181.

Table 1(A): Details of lysozyme – ethanol – water system simulated in the present study. Binary mixtures with approximately 2%, 4%, 6%, 8%, 10% and 12% ethanol concentration in water are listed in the Table.

Ethanol concentration	Number of water molecules	Molecular weight of water (g/mol)	Number of ethanol molecules	Molecular weight of ethanol (g/mol)	Total number of atoms in simulation
2.01%	6977	125586	56	2576	23395
3.93%	6928	124705	111	5113	23743
6.13%	6612	119016	169	7774	23317
8.212%	6455	116190	226	10396	23359
10.262%	6302	113436	282	12972	23404
12.333%	6140	110520	338	15548	23422

Table 2: Time average values for the parameters calculated for various concentrations of ethanol in ethanol-water binary mixtures.

Ethanol Concentration	Radius of gyration (nm)	Total Energy (kJ/Mol)	Enthalphy (kJ/Mol)	RMSD (nm)
0%	1.42	-443865.8	-443842.0	0.17
2.01%	2.03	-296529.6	-296515.7	1.94
3.933%	2.36	-295755.5	-295741.4	2.47
6.131%	2.42	-284809.8	-284796.0	2.86
8.212%	2.56	-279755.0	-279741.2	3.00
10.262%	2.54	-275197.3	-275183.4	3.08
12.33%	2.62	-270054.9	-269361.5	3.10
100%	3.12	-54189.38	-59108.76	4.17

Table 3: Table of constants for polynomial in equation 5.1 for approximately 7% ethanol

Property	C_0	C_1	C_2
Total energy	-596.247	10980.613	-329718.709
Enthalpy	-596.073	10978.794	-329700.3
Radius of gyration	2.477	0.066	0.009
RMSD	3.669	-0.564	0.058

Table 4: Simulated results vs Interpolated results for the 7.153% concentration of ethanol in water

Property	Calculated	Simulated	Percentage error
Radius of gyration(nm)	2.51	2.53	0.79
RMSD(nm)	2.86	2.84	0.63
Total energy(KJ/Mol)	-283714.93	-282421.43	0.46
Enthalpy (KJ/Mol)	-282313.54	-276895.71	1.91

Figure Captions:

Figure 1: A cartoon depiction of lysozyme protein structure is shown. Distinct colors represent different residues present in the protein.

Figure 2: Structural properties of lysozyme protein in water as observed in molecular dynamics simulations (A) Radius of gyration and (B) RMSD of lysozyme protein in water.

Figure 3: Comparing the protein structures: (A) initial structure (B) final structure in ethanol and (C) final structure in water.

Figure 4: Effect of solvent on protein as reflected in the structural property radius of gyration (R_g) of the protein: (A) R_g of protein in ethanol and (B) R_g of protein in water.

Figure 5: RMSD values for lysozyme protein: (A) in ethanol and (B) in water.

Figure 6: (A) initial and (B) final structures of protein in ethanol-water mixture at different concentrations of ethanol approximated at (I) 2% (II) 4% (III) 6% (IV) 8% (V) 10% and (VI) 12%

Figure 7: (A) Initial structure and (B) final structure of protein at various concentration of ethanol. Diffusion of ethanol molecules into the protein increases with increasing ethanol concentration as shown in the figure.

Figure 8: Time average of (A) total energy (B) enthalpy (C) radius of gyration and (D) RMSD of lysozyme protein in ethanol-water mixture is shown as a function of ethanol mole fraction.

Figure 9: Values obtained from interpolation method for (A) total energy (B) enthalpy (C) radius of gyration and (D) RMSD for approximately 7% of ethanol. (in each figure red symbol corresponds to ~ 7% ethanol case) are shown along with the values obtained from simulations.

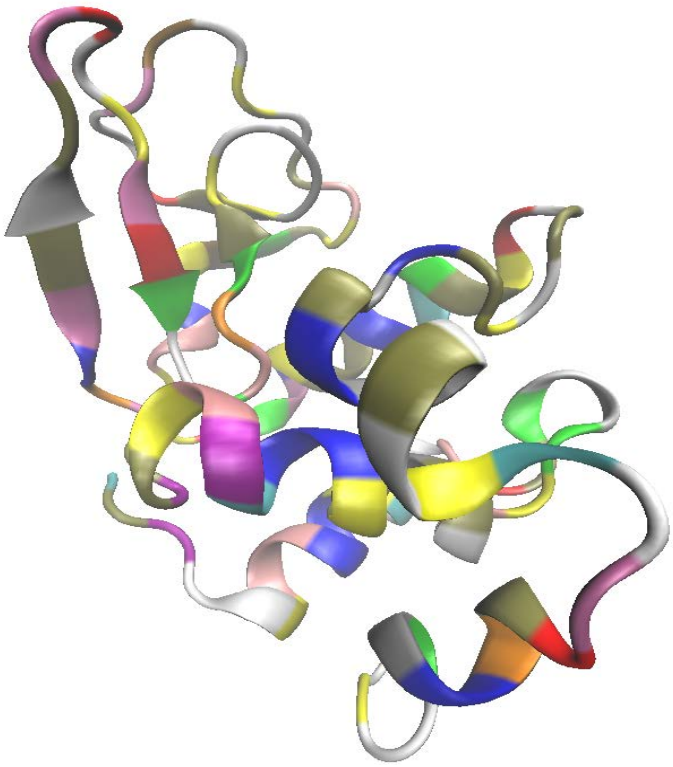


Figure 1: A cartoon depiction of lysozyme protein structure is shown. Distinct colors represent different residues present in the protein.

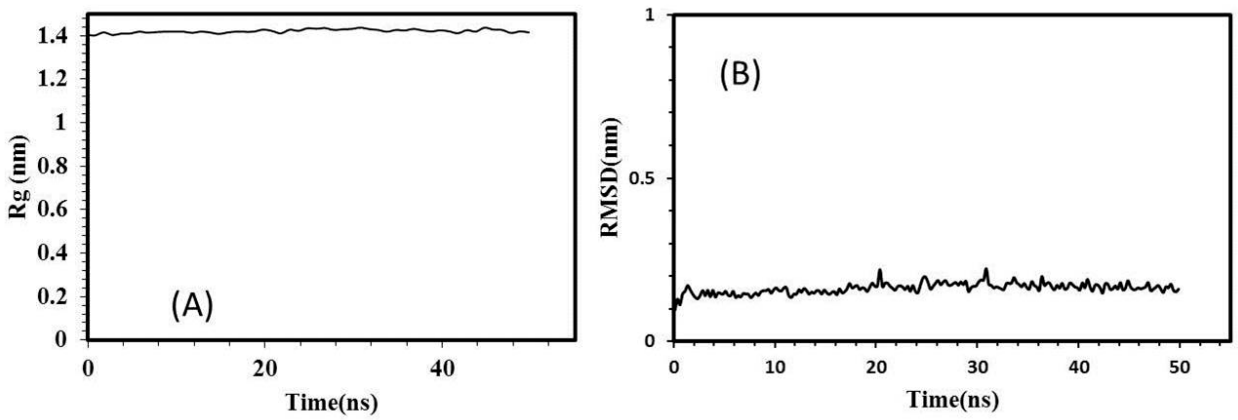


Figure 2: Structural properties of lysozyme protein in water as observed in molecular dynamics simulations (A) Radius of gyration and (B) RMSD of lysozyme protein in water.

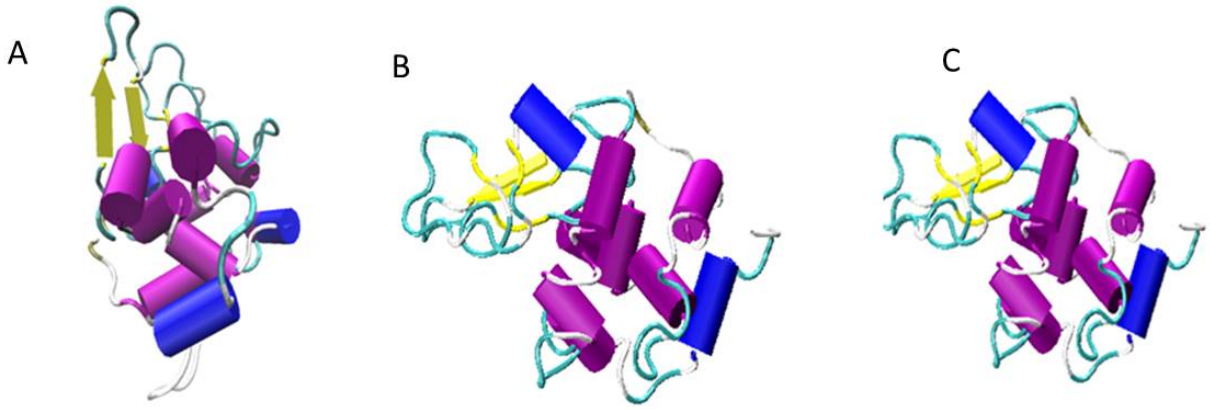


Figure 3: Comparing the protein structures: (A) initial structure (B) final structure in ethanol and (C) final structure in water.

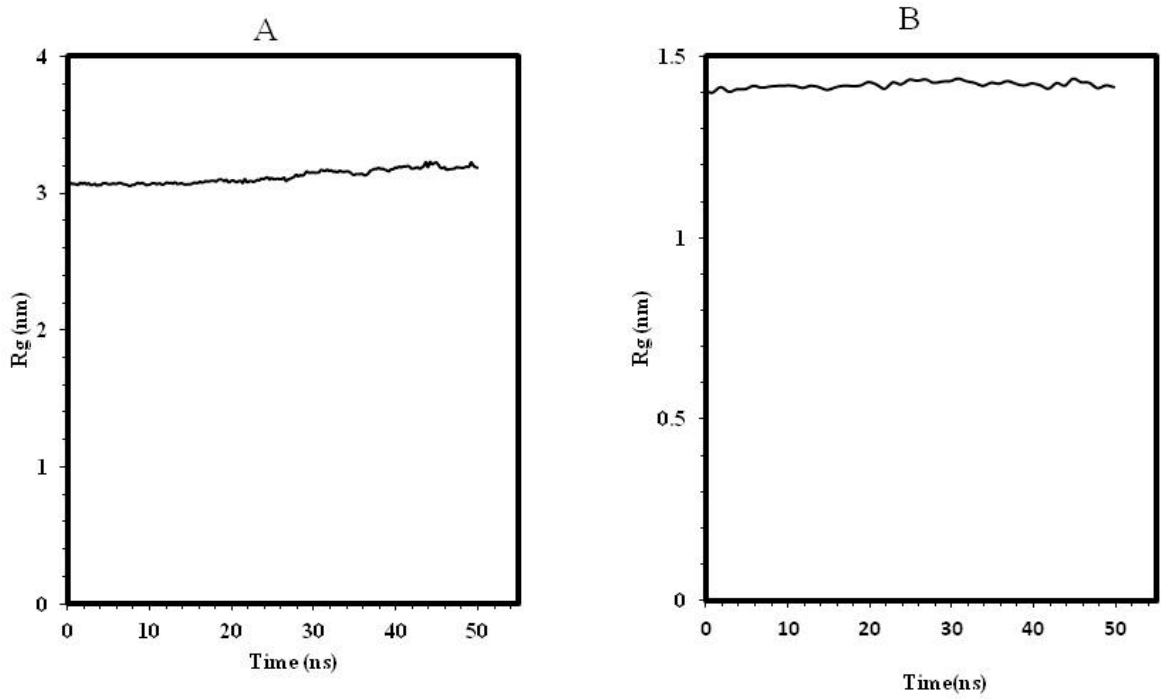


Figure 4: Effect of solvent on protein as reflected in the structural property radius of gyration (R_g) of the protein: (A) R_g of protein in ethanol and (B) R_g of protein in water.

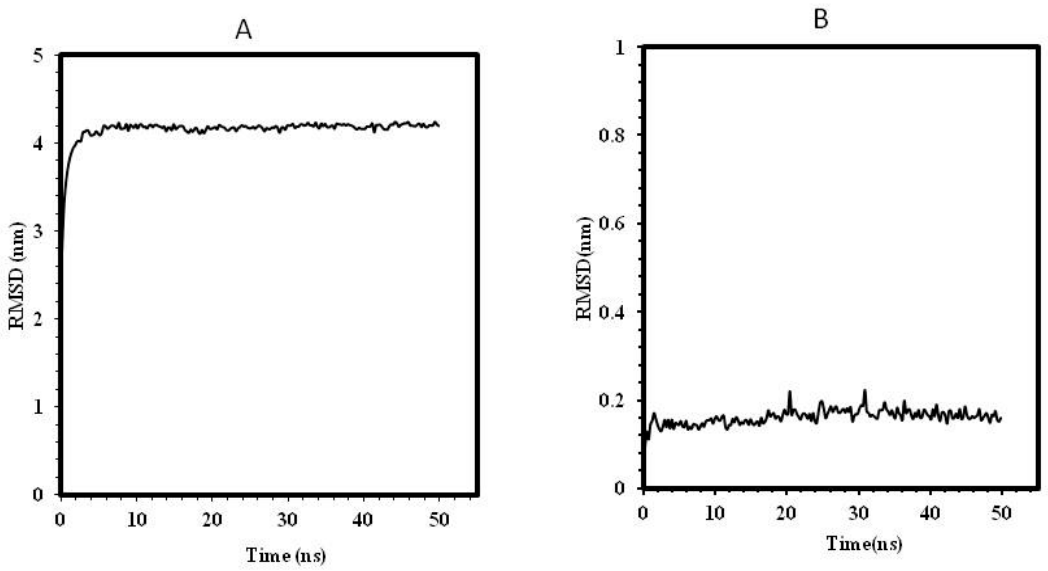


Figure 5: RMSD values for lysozyme protein: (A) in ethanol and (B) in water.

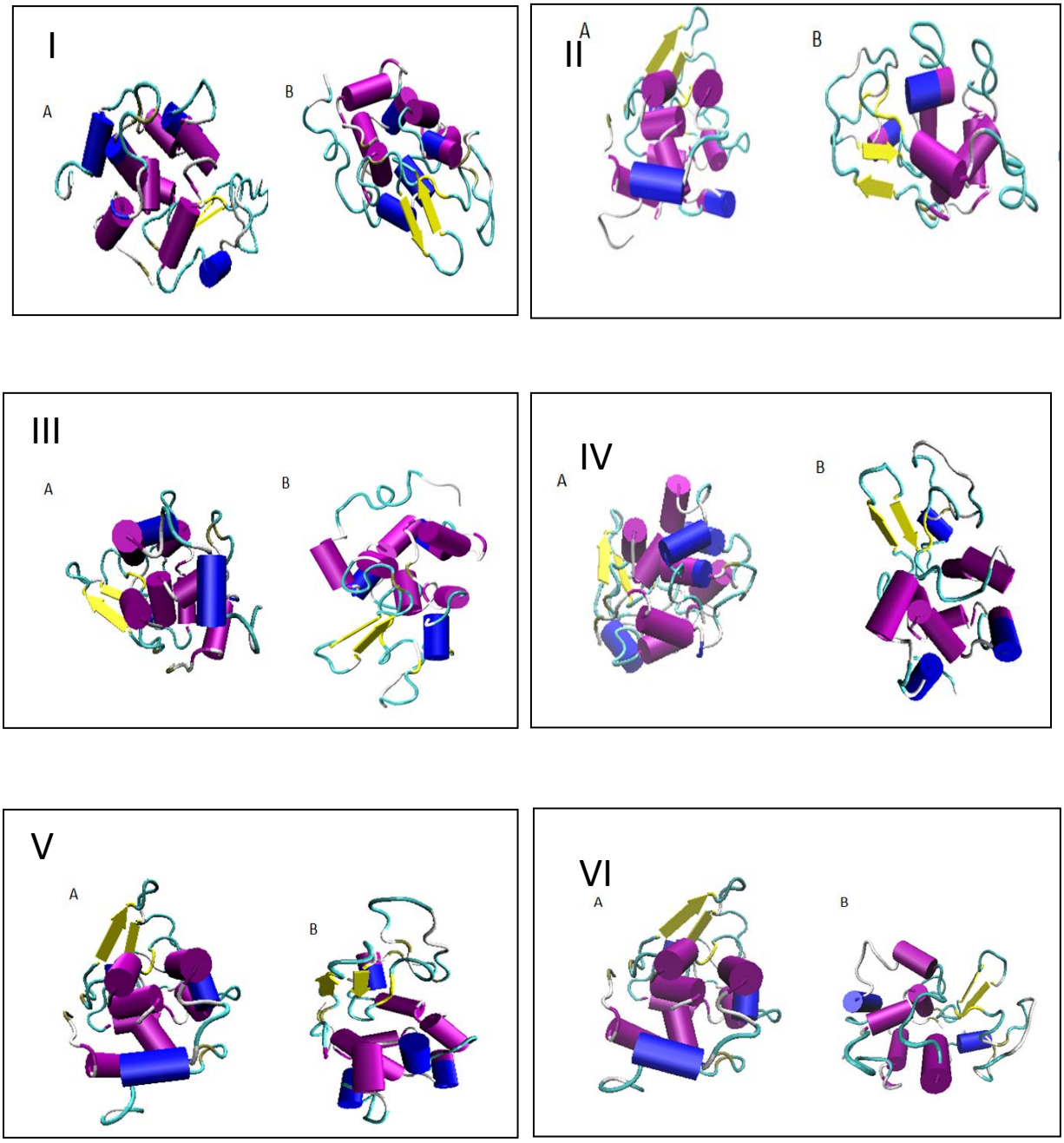


Figure 6: (A) initial and (B) final structures of protein in ethanol-water mixture at different concentrations of ethanol approximated at (I) 2% (II) 4% (III) 6% (IV) 8% (V) 10% and (VI) 12%

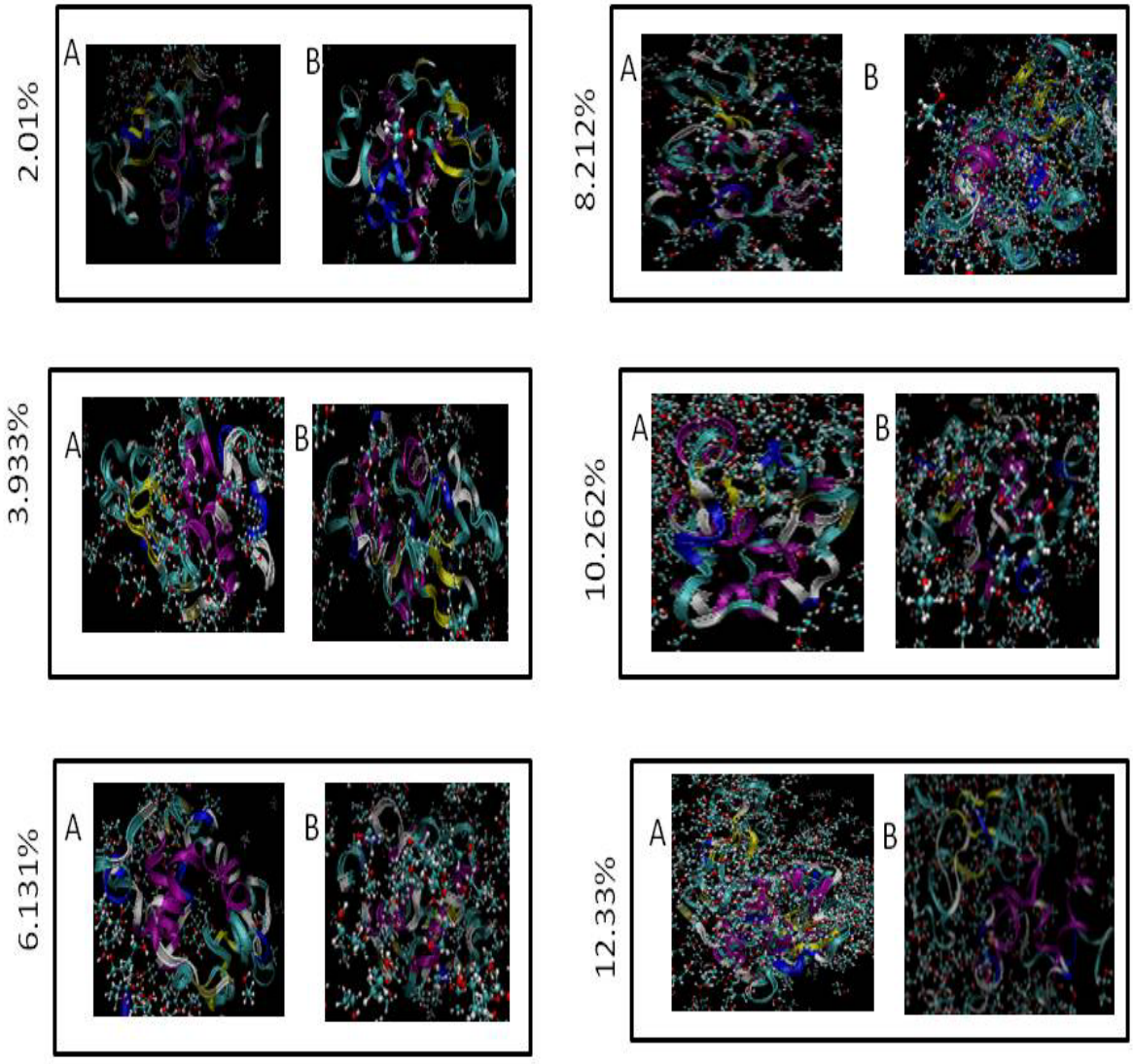


Figure 7: (A) Initial structure and (B) final structure of protein at various concentration of ethanol. Diffusion of ethanol molecules into the protein increases with increasing ethanol concentration as shown in the figure.

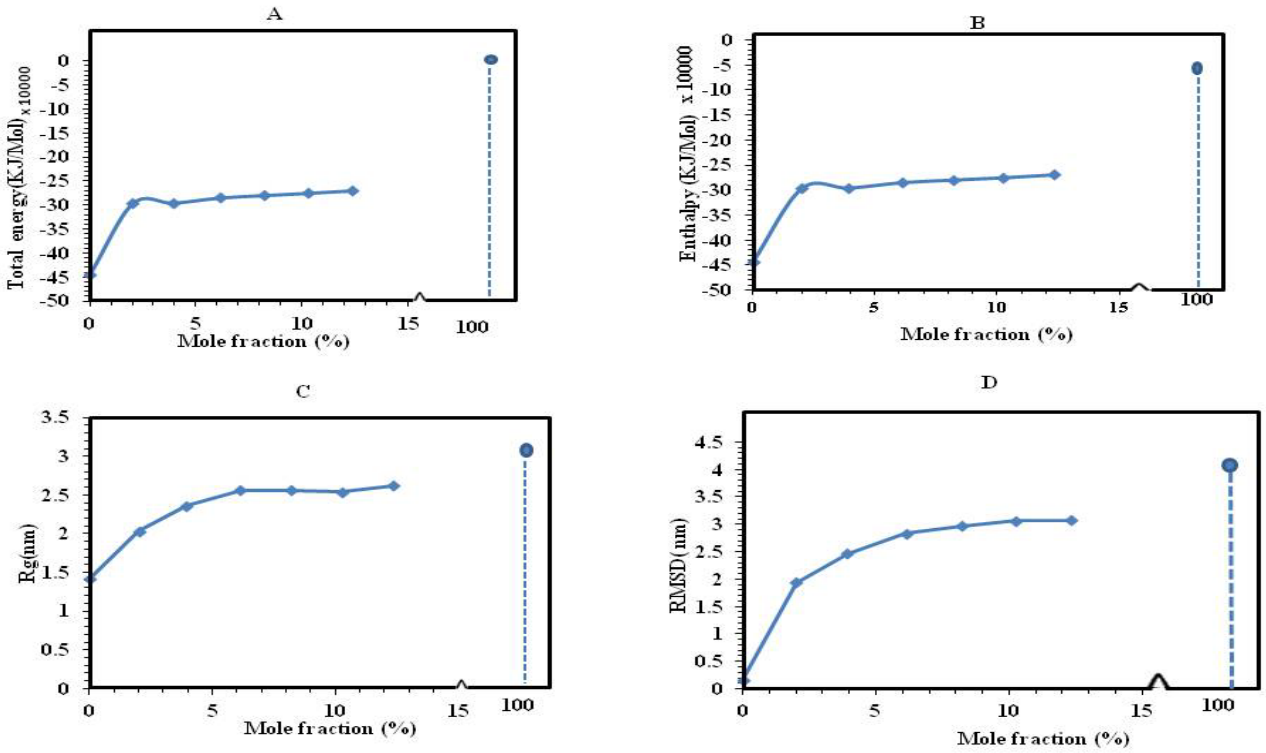


Figure 8: Time average of (A) total energy (B) enthalpy (C) radius of gyration and (D) RMSD of lysozyme protein in ethanol-water mixture is shown as a function of ethanol mole fraction.

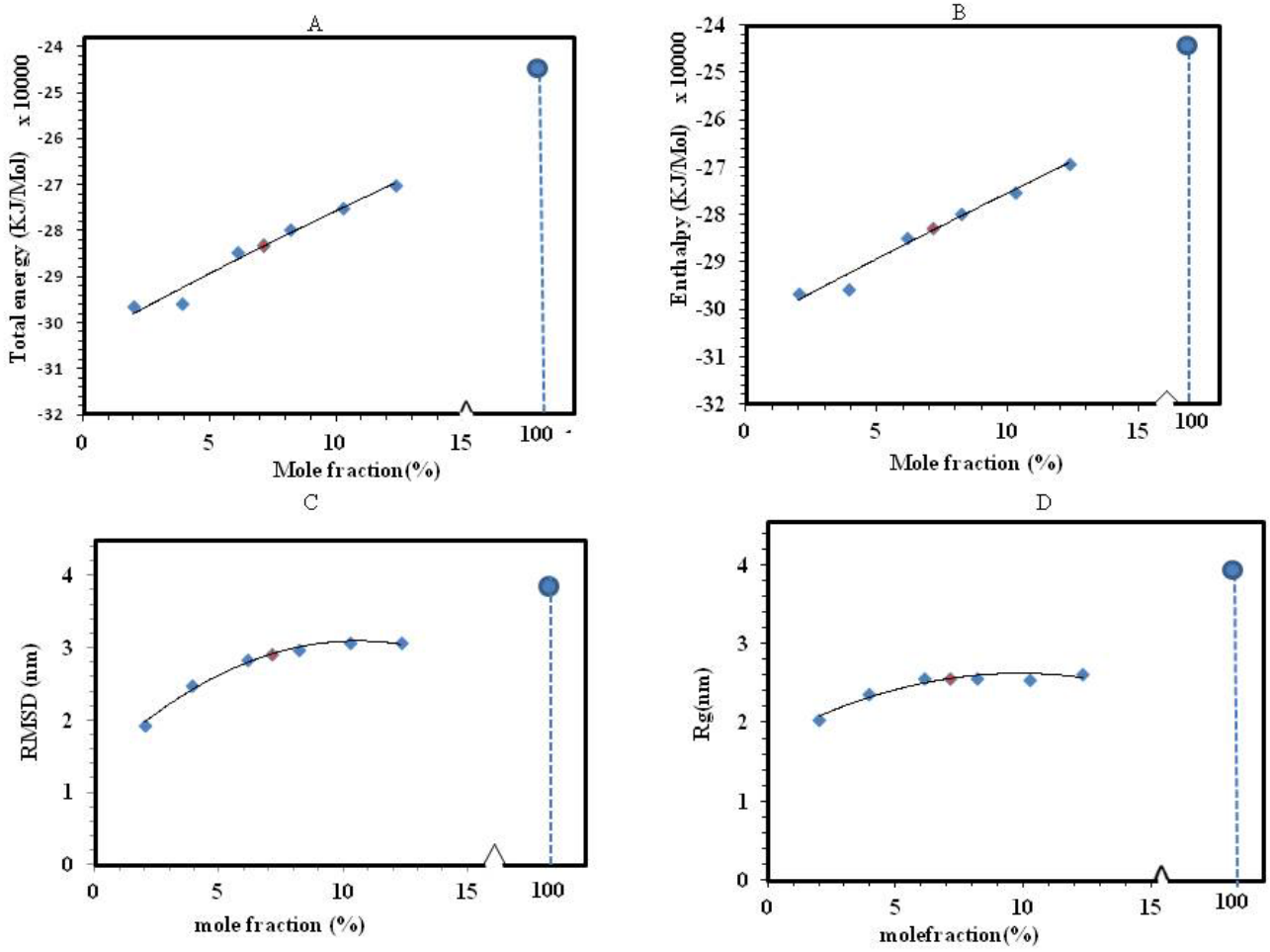


Figure 9: Values obtained from interpolation method for (A) total energy (B) enthalpy (C) radius of gyration and (D) RMSD for approximately 7% of ethanol. (in each figure red symbol corresponds to ~ 7% ethanol case) are shown along with the values obtained from simulations.

4. Computational Modeling of Peptide-Aptamer Binding

K. Rhinehardt, R. Mohan, G. Srinivas, North Carolina A&T State University

To be published in: Computational Peptidology, Methods in Molecular Biology, Vol. 1268, DOI 10.1007/978-1-4939-2285-7_14, Springer Science+Businesss Media, New York, 2015, Peng Zhou and Jian Huang (eds.)

Prelude

Evolution is the progressive process that holds each living creature in its grasp. From strands of DNA evolution shapes life with response to our ever changing environment and time. It is the continued study of this most primitive process that has led to the advancement of modern biology. The success and failure in the reading, processing, replication and expression of genetic code and its resulting biomolecules keep the delicate balance of life. Investigations into these fundamental processes continue to make headlines as science continues to explore smaller scale interactions with increasing complexity. New applications and advanced understanding of DNA, RNA, peptides and proteins are pushing technology and science forward and together. Today the addition of computers and advances in science has led to the fields of computational biology and chemistry. Through these computational advances it is now possible not only to quantify the end results, but also visualize, analyze, and fully understand mechanisms by gaining deeper insights. The bio-molecular motion that exists governing the physical and chemical phenomena can now be analyzed with the advent of computational modeling. Ever increasing computational power combined with efficient algorithms and components are further expanding the fidelity and scope of such modeling and simulations. This chapter discusses computational methods that apply biological processes, in particular computational modeling of peptide - aptamer binding.

Keywords: Molecular Dynamics, Aptamers, Peptides, Docking, Computational Peptidology

Introduction

Short sequences of amino acids and peptides have many applications that include self-assembly, cell signaling, nutritional enhancement and biomarker research. The application of peptides as biomarkers is of particular interest due to their use in disease detection. Biomarkers are molecules that correspond with bio-chemical changes in the body. Changes in concentration, physiology and morphology are indicators that allow tracking of disease progression and drug effectiveness in the body (1). For some diseases the biomarker is an individual peptide, but peptide sequences can be harvested from within protein biomarkers. As proteins are poly-peptides, one can look at peptide segments of proteins that are important in the binding of the

system. Experimentally targeting the peptides gives specific insights into the binding as well as provides a cost effective alternative to using entire proteins. The concentration of these molecules coincides with the severity of the disease. This corollary has opened a window of opportunity in the diagnosis of many diseases. However, discovering a suitable biomarker alone is not sufficient to create a viable diagnostic platform. Selecting a bio-receptor for the biomarker is essential as it specifically recognizes the biomarker among millions of other molecules. Bio-receptor-biomarker combinations act as a lock and key mechanism to create a biological complex which can be interpreted by a biosensor. A biosensor is a receptor-transducer device that provides quantitative information using a bio-recognition element/bio-receptor and a transducer (2). The transducer is generally based on electrochemical, mass, optical or thermal properties while the bio-recognition element or bio-receptor action involves biochemical mechanism (2; 3). When a biological sample is loaded into the sensor, the bio-recognition element/bio-receptor recognizes the target or key in the sample and binds to it. The transducer registers the change which is quantified and displayed on the device (see Figure 1).

The potential of diagnostic devices for various diseases are hinged on the ability to gather the appropriate bio-receptors (2). The most common method of making biosensors involves using antibodies as the bio-receptor (2). Antibodies accompany the biological response to disease and injury which facilitates their use in biosensors. Glucose meters and pregnancy tests are well known examples of biosensors because of their ease and immediacy of use.

A comprehensive sensor for many diseases would require a multi biomarker platform. Antibody based biosensors are common; however there are distinct disadvantages of using antibodies in a multi biomarker biosensor. Antibodies are large molecules that are not readily synthesized and can be chemically unstable (4; 5). Instability can cause errors and inaccuracies in readings of the biosensor. Their relatively larger size limits the number of antibodies that can be placed on the surface of the biosensor. Not only are antibodies large, they are often good for a single usage in a biosensor (5). These challenges motivated the investigation for better bio-recognition elements. Aptamers are one such bio-recognition element (6).

Aptamer Selection

Aptamers are broadly classified as either nucleic or peptide aptamers. Nucleic aptamers are synthetic oligonucleotides sequences made of single stranded DNA or RNA (7). Peptide aptamers are combinatorial protein molecules consisting of a variable peptide sequence inserted within a constant scaffold protein (8; 9). Aptamers are advantageous as bio-receptors since they are relatively small, chemically stable and have a high binding affinity (4). The aptamers have similar or better binding affinity compared to antibodies (7). Such binding affinity is not only due to the aptamers ability to bind to a specific structure but also to adapt conformation that favors the binding.

Nucleic and peptide aptamer types have distinct advantages as bio-recognition elements. Peptide aptamers have added chemical diversity compared to nucleic acid aptamers as binding does not occur on the sequence level (10). Both RNA and DNA aptamers are reusable. However RNA aptamers are susceptible to ribonuclease degradation limiting their reusability (2; 4). Due to small size, unlike antibodies, it is possible to affix large number of aptamers in a single location, creating a high density receptor area. Aptamers can also be easily functionalized and immobilized to surfaces to create highly ordered receptor layers (2).

Over the years, a compilation of oligonucleotides and peptides have been made into aptamer libraries. A standard nucleic 25-mer library compilation currently stands at 10^{15} available aptamers (11). In solution, these aptamers are quite flexible and adopts a tertiary conformation that complements the target molecule (7). In 1990, a reasonable experimental solution for nucleic aptamer selection was provided by the Systematic Evolution of Ligands by Exponential Enrichment (SELEX) process, in which developed libraries undergo incubation with the desired target molecule (11). Aptamers that do not bind to the target are removed, and bound aptamers are separated from the target and amplified using polymerase chain reactions (PCR) (11). In the PCR process primers are added to the aptamers and are replicated making many double stranded copies. These double strands are then separated, transcribed and purified into single stranded DNA (ssDNA) (11). This pool of aptamers goes through several more rounds of SELEX until the pool is reduced to a handful of sequences (see Figure 2). Target features, concentration, design of the initial library, experimental environment, and specificity of the binding are all determinants for the number of SELEX rounds that need to be done (11). After the SELEX process the pool of aptamers must be sequenced for identification. The resulting SELEX aptamers should contain a select group that has the highest binding affinity for the target molecule.

Several methods have evolved for the selection of peptide aptamers. One method is using phage display. In this process peptide libraries constrained in loops of capsid protein are presented to a filamentous bacteriophage (8; 10; 12). The gene needed for the capsid protein is contained within the phage effectively isolating the target and its aptamer. A second approach, also named "two-hybrid" approach, is to select peptide aptamers that bind to their targets within the cytoplasm of living cells (10). In this process, a protein is used as a scaffold for the display of a random library. A transcription factor is attached to the target protein within a cell containing a marker that is dependent on the expression of the transcription factor. Peptide aptamers are selected based on the inhibition of the selected protein. Alternative methods have been introduced using ribosomes and mRNA displays (10). These methods still follow the same protocols but utilize DNA and mRNA libraries. Though all of these methods are constrained by the size of the library, they are still effective in generating peptide aptamers.

Once identified, these aptamers must go through optimization and validation experiments. Aptamer binding can be optimized by examining slight variations in the sequenced aptamer and varying the solvent environments. Variations in solvents and ion concentrations are known to influence the binding event (13). Consideration must also be given to the target molecule source. For example, if the target will be introduced from a blood or urine sample, one needs to make sure that the aptamer is also viable in the associated environment. Aptamer binding validation is generally done using ELISA, microarrays and surface plasmon resonance imaging (SPRi) (14). These methods allow one to find binding affinity, association and dissociation constants which determine the strength of the binding combination.

Experimental Analysis of Peptide – Aptamer Binding: Challenges and Limitations

Despite several experimental studies for aptamer selection and binding, there are still challenges and limitations. Since its introduction, the SELEX process has evolved; however, for this process to be successful, sequencing of the aptamers is still required. Due to the massive size of the aptamer library, SELEX must be done in small batches and there are risks of damaging the aptamers during the process. It is worth mentioning that during the PCR process of SELEX, the aptamers are amplified with the addition of primers and extension regions. Although primers are later removed, any residual nucleotides would alter the sequence. This addition could also cause a change in the binding characteristics or location. Though this process can select a group of aptamers over time, a major drawback is its inability to identify the specific binding site or natural progression of binding. Experimentally, Nuclear Magnetic Resonance (NMR) and X-Ray crystallography are considered to be the current best tools to obtain molecular structures. These techniques provide a snap shot of stable molecular structures in a solution, but unable to provide insights into the natural progression of binding.

On the other hand, validation methods such as SPRi and microarrays have their own shortcomings. Both methods are efficient in determining binding but they do not provide structural details of binding events. Many larger aptamers, peptides and proteins will have multiple binding sites. SPRi and microarray results give authentication to the formation of a binding complex but do not reveal explicitly where the binding occurs. Similar shortcomings are associated with the methods used in binding of the aptamer targets to a surface. It is important to note that while dealing with small molecules like peptides and aptamers, even minute changes can impact the binding. For example, surface chemical methods applied to aptamers and peptides can cause changes in the molecular structure or interrupt possible binding sites.

The understanding of the natural progression of binding that is a currently a limitation in experimental studies can be further enhanced through computational modeling. A better understanding of biomarker or bio-receptor interactions can be obtained by developing computational models based upon their associated molecular systems. Computational modeling can facilitate the selection of target molecules for any biomarker. Such modeling can also aid the

SELEX process as it enables one to analyze and understand the progression of the binding process, and not just the end outcome. Computational peptide - aptamer binding experiments can help identify binding sites and structural motifs obtained under various conditions.

Computational Modeling

The power of today's computational modeling can be an avenue to test, analyze and visualize the peptide - aptamer binding that forms the basis of the aptamer selection process. The size of a nucleic acid aptamer library depends on the length of the variable region and can be approximated as

$$library\ size = 4^n \quad (1)$$

where *n* is the length of the variable region in the aptamer (2). Going through each sequence of such a library is nearly impossible using the current wet lab procedures. As explained before, using aptamers as bio-recognition elements offers high selectivity and specificity. Reducing the multi-trillion aptamer pool through computational modeling and analysis can potentially cut down the time and resources needed for optimal binding aptamer selection. It is now known that the open regions of the aptamer's 3D structure provide the binding sites for peptides. There can be several of these sites but it is unclear which site is used, and whether the site changes under varying conditions. Using aptamers in a biosensor has the additional challenge of identifying the optimum orientation that favors binding. Generally, aptamers are bound to a surface within the device. Surface chemistry and target orientation influence the effectiveness of biosensors. Modeling and understanding how binding occurs will aid in the device development providing insights into binding specificities.

Computational modeling includes the effect of the associated processes and physical phenomena and can provide an emulation of the real behavior through relevant mathematical and computational formulations. In the case of molecular structures involving peptide – aptamer interactions, computational simulations can provide insight into the progression of the binding process. The efficacy, applicability and insight from computational analysis of peptide – aptamer systems depend upon the fidelity of associated molecular models.

For the computational analysis of peptide and aptamer binding, docking and transient dynamic simulations of the relevant molecular systems are general approaches that can be employed. Simulation modeling and docking cannot be performed without a structure. Having a structure allows one to explore anomalies, mutations and destruction of molecules and evaluate how these could lead to changes in molecular function and phenotype. These molecular structures are typically generated from NMR or X-Ray crystallographic solutions of the associated molecular systems. If the structures are not available, structure prediction analysis can help predict the structure (15). Structure prediction methods start from the primary structures and compute

secondary or tertiary structures based on other closely related known structures or de novo physics (15). It is important to examine the molecular makeup prior to simulation analysis to identify key features and areas of interest. For peptides (short amino acid chains), one need to look at the sequence of residues as well as their length. Amino acids can be characterized into subgroups defined by their residues (Figure 3). Each residue serves a specific function in peptide and protein structure. Such features are important as they determine their role in binding. For example, proline tends to make kinks in peptide chains due to its formation of a nitrogen ring on a peptide backbone (15). Therefore a proline heavy peptide tends to be more rigid. Identification of such peptide characteristics aids in the simulation analysis and design. For aptamers, one must consider the tertiary structure, as well as the open regions that may act as potential areas for ligand interactions.

The function of active biological molecules leads to the understanding of biological pathways and mechanisms (16). The ability to readily change molecules and obtain a corresponding binding affinity of each combination can provide a preemptive look into the efficiency and efficacy of a binding combination. Testing different ligands with specific target molecules can be performed with docking method. Docking is a method of bringing a target and ligand together to assess binding and its ability to form a stable structure (17). This is done by bringing a biomolecule into a receptor binding site and moving the ligand to ascertain the location and conformation that is most advantageous for this binding to occur. Docking analysis can score each conformation to express the quality (17). As aptamers and peptides can have multiple binding sites; using methods like docking is one avenue of evaluating the quality of peptide – aptamer binding. Docking method and its application to a specific case of peptide - aptamer binding are discussed in detail in a later section.

In addition to the docking, computational modeling based on transient dynamic analysis methods provide a detailed insight of the progression of peptide – aptamer binding process. Transient dynamic models are based on physics based mathematical formulations involving different length scales and features of interest. Based on the time and length scales involved, computational modeling and analysis approaches can be broadly classified into three main categories; Quantum mechanics, atomistic modeling, and mesoscale dynamics (18).

Mesoscale Dynamics focuses on systems that involve billions of atoms and are generally based on larger geometrical sizes represented by appropriate physical laws (18). The algorithms in this model are generally based on Newtonian Physics. In biological systems this type of modeling can be used for organ, large biomolecules, and their interface dynamics. However, modeling small atomic scale interactions using the mesoscale dynamics would cause inaccuracy with singular or small groups of atoms.

Atomistic Modeling is suitable for small systems where individual atoms and or small clusters of atoms are involved and the phenomena is influenced by the motion of individual atoms (18).

Molecular dynamics and Monte Carlo simulations are common examples of atomistic modeling (19). These models can routinely explore time scales of picoseconds (10^{-12}s) to hundreds of nanoseconds. Although both approaches are based on interatomic potentials, they are inherently different. Monte Carlo modeling uses probabilistic approach to determine the lowest energy (20; 21). On the other hand, the governing equations in molecular dynamics follow classical Newtonian mechanics (22). This method is derived from Newton's equation of motion based on the selected force fields that defines the associated forces of the molecular interactions. This method is suitable to study dynamics and have been effectively used to model biological structures and interactions (19; 23; 24).

Quantum mechanics (QM) methods are highly suitable for simulating the electronic structure and properties of the system. Generally, chemical bond formation/breakage involve electron interactions between atoms (18). Such bond formation and breakage are accurately modeled in this approach. This method is the most accurate of the three methods for estimating the properties. However, this method is computationally expensive and is well suited only for extremely small systems. The high accuracy of this method is due to its ability to account for electron interactions through appropriate quantum mechanical equations. However, peptide-aptamer binding does not involve such electron interactions. Quantum mechanics methods are better suited for studying enzyme reactions, charge transfer, and analysis of chemically active regions in biomolecules (25; 26). For this reason, atomistic based simulations are suitable for transient dynamics analysis of peptide – aptamer binding. In the following sections, we describe docking and transient dynamics approaches involving atomistic modeling of peptide-aptamer binding.

Docking Methods

Docking methods can be used in peptide – aptamer binding models for determining locations of binding, possible high affinity sites and understanding structural isoforms. As described previously, binding is often depicted as "lock and key" mechanics where target molecules are considered the lock and its corresponding ligand is the key (27). Docking is a computational method of predicting the correct ligand as well as determining the structure and orientation of that ligand for a specific target molecule (28; 29). The goal of docking is to optimize the binding event by considering the best fit between the ligand and target molecule. Two main approaches to docking are geometric and flexible. Geometric methods consider the structural geometry, sterics and shape of the ligand and its binding sites. This structure based method analyzes the binding site surface and its chemistry between the ligand and target molecule to determine the most complementary combination (30). Investigating the binding site, one can define features that are distinct and necessary for docking before introducing a series of appropriate ligands. Though this method is sufficient in investigating the binding area, it considers molecules as rigid bodies. On the other hand, peptides and aptamers are flexible which means multiple

conformations as well as binding areas are possible. There can be many ligands that fit into the binding area and vary in shape. However, such shape variations are not well accounted in geometric docking approach. Flexible docking is useful in investigating shape variations and customization beyond geometric constraints.

X-ray crystallographic inspection of proteins and ligands have shown that high affinity ligands conform to the binding cavity to take advantage of the hydrogen bonding possibilities and electrostatic interactions (27). Flexible docking considers flexibility of the molecules instead of treating them as completely rigid bodies (29). Flexible docking analysis simulate the ligand near a target molecule active site and allow them to move based on energy minimization (31). This allows for binding to occur in the most favorable conformation. Upon binding, the affinity of each conformation is scored and the confirmation with the best score is considered to be the optimized state of the biological complex. While this method looks at the most favorable conformation, there is also margin of error in the calculation when compared to wet lab experimentation (32). The scoring and energy function calculation of a molecule or complex without solvent limits the method's accuracy in the interpretation of the experimental results.

Docking Studies of Peptide – Aptamer Binding

Docking methods have been applied to the discovery of peptide - aptamer binding of HIV Rev-RBE and BIV Tat-TAR complexes (33). Bovine immunodeficiency virus (BIV) Tat peptides bound to the BIV TAR element (forming BIV tat peptide-TAR complexes) were solved using multidimensional NMR analysis. This combination was used as a control for the 17 amino acid peptide from the 34-50 residue of the human immunodeficiency virus (HIV) Rev protein that binds to a 30 nucleotide RNA aptamer. This aptamer, aptly named Rev binding element (RBE), was previously modeled with NMR constraints. Cedergren *et al.* provided a detailed strategy for the docking and modelling of the $Rev_{34\text{-}50}$ peptide-RBE aptamer complex interactions (33). In their computational study, initial structures were treated as rigid bodies and individually minimized. Complexes were formed by combining the molecules into various binding conformations. Energetically most favorable conformations were determined by electrostatics and Van der Waals interactions. The side chains and binding partners of those conformations were extensively analyzed. This process was applied to the known complex of BIV Tat-TAR to determine its efficacy by correctly identifying the orientation (5'-3' relative to N and C termini) and register (juxtaposition) of two binding molecules. Validation was done by examining the binding free energy of the BIV Tat-TAR complex. Additional complexes were derived from the NMR solution of the BIV Tat-TAR complex by rotating the peptide and changing the register of the BIV Tat molecule in the binding site of the TAR element. It was found that the conformation identical to the NMR solution had the lowest binding energy. However the calculations also showed that the docking method employed is insensitive to small changes in the peptide register.

Nevertheless, the docking method can be used effectively in determining the global register and orientations of peptides docked with RNA aptamers.

After validating with the known BIV Tat-TAR complex the docking method was applied to the unknown structure of the Rev_{34-50}-RBE complex. Docking was guided by the electrostatic potentials and experimental data. Experimental work and the electrostatic potential surfaces indicated the major groove of the RBE aptamer to be the site of binding between key arginine residues (Arg2, Arg5, Arg6, Arg9 and Arg13) in the peptide (Figure 4). Five initial models were generated from the *anti (A)* and *syn (S)* conformations of the peptide with the open end (NO) and tetraloop (NL) regions of the aptamer. The local minimum energy was calculated for each complex model in various conformations. The lowest energy complex was formed when the N terminus of the *anti* (A) or *syn* (S) peptide points towards the UUGG tetraloop region of the RBE. Further investigation of the binding energy showed that the RBE NL model with the *anti*-form of the peptide was the best model for the complex (Figure 5). Based on the interaction of the peptide and RBE, the roles of the arginine residues and other side chains were also identified. The results indicated that the Arg2, Arg5 and Arg11 residues are important in binding but no single arginine side chain is singularly responsible.

The prediction of the optimal intermolecular geometry and interaction energy provides details of the binding area and the residues essential for the binding events. This work also showed that the major grove was the site of binding in the RNE RNA aptamer and predicted a possible structure based on the binding energy of peptides and aptamers (33).

Advances in docking methods in the mid 1990's through the 2000's led to the development of efficient docking analysis codes. More sophisticated analysis methods such as quantum mechanics and molecular dynamics were available but restricted by computing power at that time (34). The increase in computer power through parallel processing introduces allows for more detailed and accurate analysis of larger and complex problems. Docking can be improved with transient dynamics analysis, in particular molecular dynamics, which include fully solvated systems and more accurate models. As stated before, docking methods do not fully consider the flexibility of the molecules during binding. Movements such as the relaxation of active site around the ligand are still not considered in flexible docking. Further, such contributions make calculations of binding energy less reliable (35). Such factors not considered in docking, can be modeled in molecular dynamics and other transient dynamics methods. Peptide – aptamer modeling has now moved toward the more sophisticated analyses of molecular dynamics.

Transient Dynamic Analysis: Molecular Dynamics Methodology

One computational modeling technique applicable for the analysis of biomolecular motion and interactions is based on molecular dynamics (MD) modeling (36; 37). MD methodology allows for the natural progression of the biomolecules in solution (38). MD method has been applied to

determine the chemical, physical and mechanical properties of materials based on their molecular structures.

In the case of peptide – aptamer binding, the use of molecular dynamics (MD) analysis is most suitable for studying proteins and aptamers from a molecular level. The important question to be addressed is the behavior of the atoms within the macromolecules and how binding occurs under given conditions. Using the Newtonian equations of motion (Equation #.2), MD simulations can predict the movement of the atomic behavior of molecules as closely as possible to that of laboratory conditions (39).

$$M_I \frac{d^2R}{dt^2} = f_I(R,t) \tag{2}$$

This equation relates the mass (M_I) and change in position (R) over time (t) to the force at each point in time (f_I). This calculation is done for every atom present in the simulation system at each time step. For accuracy of such simulations along with the structure, we must consider a suitable force field that is used for calculating the energy changes in the system.

For MD simulations, force fields are an important and influence the accuracy of the simulations (40; 41). The energy summation (Equation #.3) is written as

$$E = E_{bonded} + E_{nonbonded} + E_{other} \tag{3}$$

Bonded energies include bond stretching, bond angle and torsional energy (42). Non-bonded energy includes interactions such as van der Waals and electrostatic forces. Energetic contributions from other interactions are included in the E_{other} term. Over the years, several MD analyses codes and packages have been developed (43-45). The choice of MD simulation analysis package depends upon the system studied and the availability of associated force field for the corresponding system of interest. Various force-fields such as CHARMM, AMBER, OPLS-AA and GROMOS etc. have been shown to be successful in simulating various physical, biological and material systems (40; 46-48). The AMBER force field in particular has often been used with peptides as well as RNA and DNA based molecules (49).

The molecular system configurations for the simulation studies are generally derived from a structure file in which the individual atoms, atom types, bonds and positions are defined to form the initial molecular structure. In general, the starting structure files need to be converted into a readable topology files in the respective data format for the chosen MD simulation analysis package. From the starting molecular conformations, the enclosed work space must be defined, solvated, minimized and equilibrated via different established methods before simulation (50). These preliminary steps are essential in establishing a stable initial system that best represents the real physical problem. The ability to follow a natural progression of a system is the advantage that MD gives over the docking method. Both quantitative parameters and visual

analysis are used to assess and analyze the results from the dynamic simulation study. A variety of visualization data analysis software tools are available that are compatible with the output files of commonly used computational molecular modeling packages. Visualization analysis of the simulation trajectory provides guidance into the quantitative analysis.

One can also quantify structural changes that may occur in the simulation process. For example, RNA and DNA structures are held together through hydrogen bonds and it is possible to track these bonds throughout the simulation. Solvent interaction can also be investigated and quantified. The distribution of the solvent surrounding the molecules can also be considered as a method to understand the influence of the molecules on the environment.

This well-defined and constantly growing method of modeling biological molecules has become more prevalent over the years. In 1977 modeling globular protein dynamics in vacuum for short moments (10ps) in time was a huge leap in computational applications (51). The advances in parallel processing and dynamic algorithms in the late 1990's and 2000's allowed moderate scale molecular dynamic simulations of nucleic acids and small proteins in solution to be explored (52-55). Today super, high end computing, greater dynamic algorithms and software have pushed the effectiveness of MD simulations to better understand, visualize and analyze bimolecular events.

Molecular Dynamics of Peptide - Aptamer Binding

One recent study focused on molecular dynamics approach to study peptide - aptamer binding (56). The initial work was done using the breast cancer aptamer – peptide binding combination of S2.2 Anti-Mucin 1 (MUC1) aptamer and a 9 amino acid Mucin 1 peptide. This combination was simulated using the GROMACS molecular dynamics analysis package. To parallel wet lab experiments, a 9 amino acid peptide - aptamer binding was simulated in 0.15M NaCl solution at standard temperature and pressure. Visual analysis of this work showed the transient progression of peptide - aptamer binding as well as the identification of conformational changes (see Figure 6). Simulation and analysis of the MUC1-G peptide and Anti-MUC1 aptamer show that binding occurs in the open loop region of the aptamer after 51ns of simulation. In the loop region the thymine residue locks onto the arginine residue. As the simulation continues the thymine residue rotates and interacts with peptide backbone which helps the peptide and aptamer stay bound. Repeated simulation of this peptide - aptamer combination with different initial configurations each converged and bound with the peptide interacting with the thymine loop region of the aptamer.

Quantitative analysis of this combination further reiterated the visual analysis results. As biomolecules bind using non-covalent interactions, we must consider electrostatics interactions, Van der Waals forces, and hydrophobic interactions. At this distance the atoms in the binding region should be less than 4.5Å to indicate that the aptamer and peptide are noncovalently bound

(57). The distance between the aptamer and peptide in the binding site averaged 3.5Å indicating binding has occurred (Figure 7A).

The root mean squared deviation (RMSD) at the open 5' and 3' ends showed significant conformational changes in the aptamer that corresponded to binding events (Figure 7B). Interactions of the peptide and aptamer residues during binding induce conformational changes along the backbone of the aptamer molecule. Though binding happens at the loop and open ends of the aptamer, changes in the aptamer appear to manifest in the 5' and 3' ends of the aptamer as it is the most loosely associated area of the aptamer backbone. The increase in the RMSD of the aptamer in this open region is due to the initial interaction between the MUC1 peptide and the aptamer. As the simulation continues and the peptide begins to interact more with the aptamer there are distinctive bouts of stability before the aptamer open ends settle. As the peptide forms a butterfly-like motif with the loop region of the aptamer the open ends begin to settle to an RMSD value closer to that seen at the start of the simulation.

RMSD along with the visual analysis of the aptamer structure indicate changes in confromation that correspond to peptide - aptamer binding. Figure 8 shows the number of hydrogen bonds in one case of the aptamer with the MUC1 peptide as a function of simulation time studied. As the aptamer structure is held together by the formation of hydrogen bonds, disruptions in the structure would be evident in number of hydrogen bonds within the aptamer. In the beginning several hydrgen bonds were found to hold the structure together except in the loop and helical regions of the aptamer. The decrease in the number of hydrogen bonds shown at 52ns (1ns after the interaction) is due to the arginene residue of the peptide disrupting the bonds at the open ends. Extended time after the binding event indicates that the number of hydrogen bonds increase as the hydrogen bonds in the open ends reform.

The simulation results were quantitatively analyzed for the conformational changes and the overall behavior of aptamer and peptide system was observed from a molecular view point, which is not always possible in the wet lab experiments. The observed changes in structure are reflected in the atomic distance, RMSD values, and hydrogen bonds. The changes in hydrogen bonds show direct correspondence to structural changes resulting from binding events. The present MD simulations, analysis and discussions clearly show that the aptamer and peptide binding could be efficiently simulated and analyzed using computational methods.

Summary

Computational modeling provides effective means to understand peptide-aptamer bindings. This method can offer additional insights into aptamer selection, and binding processes by providing a visual and quantitative scope through biomolecular modeling. Computational docking was used to study peptide-aptamer combination. Its application provides a fundamental view into the binding site and ligand identification and interaction. However the natural movement and

behavior of the aptamer and peptide cannot fully be investigated with docking alone. To interpret the experimental peptide - aptamer complexes and gain a greater insight into their binding interactions one needs to look into the transient dynamic analysis. Molecular dynamics provides more in-depth look into the natural progression as demonstrated by an example case of peptide - aptamer binding discussed in this chapter.

Figure Captions

Figure 12: Schematic representation of biosensor working principle.

Figure 13: Schematic diagram of the SELEX process (reproduced from Reference (11) with permission).

Figure 3: Chemical structures of amino acids shown in the stick representation. Color code: Carbon-green, oxygen-red, nitrogen-blue, hydrogen-white, and sulfur-yellow.

Figure 4: Electrostatic images of the Rev Peptide and RBE. A) Front and back views of the electrostatic surface of the Rev 34-50 peptide. Amino acid residues are indicated by single letter code and number B) Electrostatic surface of the RBE with groove and loop features indicated. Phosphate groups that are protected from chemical modification due to complex formation are indicated.(reproduced from Reference (33) with permission).

Figure 5: Stereo view of the A-NL model of the Rev34-50 - RBE complex (reproduced from Reference (33) with permission).

Figure 6: Visualization of Anti-MUC1 aptamer (blue) and MUC1-G peptide (purple) binding 300ns Simulation. (A) Starting peptide - aptamer configuration. (B) Peptide - aptamer configuration after 27ns. (C) Interation of the 11^{th} thymine residue of the aptamer and the peptide backbone after 51ns of simulation. (D) Continued peptide - aptamer interaction at the open loop region of the aptamer and the arginine residue after 127ns. (e) Magnified image of the peptide - aptamer interaction at the end of simulation.

Figure 7: Quantitative analysis of MUC1-G peptide and aptamer for 300ns. A) Atom Distance between the aptamer and peptide in the binding region B) RMSD of the aptamer 5' – 3' ends during simulation.

Figure 8: Number of hydrogen bonds in the Anti-MUC1 Aptamer (blue) during simulation with the MUC1-G peptide (purple). Corresponding simulation snapshots at 25ns, 52ns, 104ns and 300ns of the anti-MUC1 Aptamer with the peptide and hydrogen bonds (red dash lines) are shown.

References

1. Jain KK. 2010. *The handbook of biomarkers*. New York: Springer
2. Strehlitz B, Nikolaus N, Stoltenburg R. 2008. Protein Detection with Aptamer Biosensors. *Sensors* 8:4296-307
3. Erickson D, Mandal S, Yang A, Cordovez B. 2008. Nanobiosensors: optofluidic, electrical and mechanical approaches to biomolecular detection at the nanoscale. *Microfluidics and Nanofluidics* 4:33-52
4. Song S, Wang L, Li J, Fan C, Zhao J. 2008. Aptamer-based biosensors. *TrAC Trends in Analytical Chemistry* 27:108-17
5. Wang J. 2000. From DNA biosensors to gene chips. *Nucleic Acid Research* 28:3011-6
6. McCauley TG, Hamaguchi N, Stanton M. 2003. Aptamer-based biosensor arrays for detection and quantification of biological macromolecules. *Analytical Biochemistry* 319:244-50
7. Clark SL, Remcho VT. 2002. Aptamers as analytical reagents. *ELECTROPHORESIS* 23:1335-40
8. Mascini M, Palchetti I, Tombelli S. 2012. Nucleic Acid and Peptide Aptamers: Fundamentals and Bioanalytical Aspects. *Angewandte Chemie International Edition* 51:1316-32
9. Colas P, Cohen B, Jessen T, Grishina I, McCoy J, Brent R. 1996. Genetic selection of peptide aptamers that recognize and inhibit cyclin-dependent kinase 2. *Nature* 380:548-50
10. James W. 2001. Nucleic acid and polypeptide aptamers: a powerful approach to ligand discovery. *Current Opinion in Pharmacology* 1:540-6
11. Stoltenburg R, Reinemann C, Strehlitz B. 2007. SELEX--a (r)evolutionary method to generate high-affinity nucleic acid ligands. *Biomolecular engineering* 24:381-403
12. Baines IC, Colas P. 2006. Peptide aptamers as guides for small-molecule drug discovery. *Drug Discovery Today* 11:334-41
13. Ferreira CS, Matthews CS, Missailidis S. 2006. DNA aptamers that bind to MUC1 tumour marker: design and characterization of MUC1-binding single-stranded DNA aptamers. *Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine* 27:289-301
14. Ferreira C, Papamichael K, Guilbault G, Schwarzacher T, Gariepy J, Missailidis S. 2008. DNA aptamers against the MUC1 tumour marker: design of aptamer–antibody sandwich ELISA for the early diagnosis of epithelial tumours. *Analytical and Bioanalytical Chemistry* 390:1039-50
15. Tramontano A. 2006. *Protein structure prediction: concepts and applications*. Wiley-VCH
16. Bader DA. 2004. Computational biology and high-performance computing. *Commun. ACM* 47:34-41
17. Schneider G, Baringhaus K-H. 2008. *Molecular design: concepts and applications*. John Wiley & Sons
18. Gomperts R, Renner E, Mehta M. 2005. Enabling Technologies for Innovative New Materials. *American Laboratory* 37:12-4

19. Sim AYL, Minary P, Levitt M. 2012. Modeling nucleic acids. *Current Opinion in Structural Biology* 22:273-8
20. Berg BA. 2004. *Markov Chain Monte Carlo Simulations and Their Statistical Analysis: With Web-based Fortran Code*. World Scientific
21. Scherer POJ. 2010. *Computational Physics: Simulation of Classical and Quantum Systems*. Springer
22. Rapaport DC. 2004. *The Art of Molecular Dynamics Simulation*. Cambridge University Press
23. Karplus M, McCammon JA. 2002. Molecular dynamics simulations of biomolecules. *Nat Struct Biol* 9:646-52
24. Karplus M, Petsko GA. 1990. Molecular dynamics simulations in biology. *Nature* 347:631-9
25. Senn HM, Thiel W. 2009. QM/MM methods for biomolecular systems. *Angew Chem Int Ed Engl* 48:1198-229
26. Náray-Szabó G, Oláh J, Krámos B. 2013. Quantum Mechanical Modeling: A Tool for the Understanding of Enzyme Reactions. *Biomolecules* 3:662-702
27. Jones G, Willett P, Glen RC, Leach AR, Taylor R. 1997. Development and validation of a genetic algorithm for flexible docking. *Journal of Molecular Biology* 267:727-48
28. Brooijmans N, Kuntz ID. 2003. Molecular recognition and docking algorithms. *Annual Review of Biophysics and Biomolecular Structure* 32:335
29. Knegtel RMA, Kuntz ID, Oshiro CM. 1997. Molecular docking to ensembles of protein structures. *Journal of Molecular Biology* 266:424-40
30. Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE. 1982. A geometric approach to macromolecule-ligand interactions. *Journal of Molecular Biology* 161:269-88
31. Österberg F, Åqvist J. 2005. Exploring blocker binding to a homology model of the open hERG K+ channel using docking and molecular dynamics methods. *FEBS Letters* 579:2939-44
32. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, et al. 2004. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *Journal of Medicinal Chemistry* 47:1739-49
33. Srinivasan J, Leclerc F, Xu W, Ellington AD, Cedergren R. 1996. A docking and modelling strategy for peptide–RNA complexes: applications to BIV Tat–TAR and HIV Rev–RBE. *Folding and Design* 1:463-72
34. Karplus M, McCammon JA. 2002. Molecular dynamics simulations of biomolecules. *Nature Structural Biology* 9:646-52
35. Okimoto N, Futatsugi N, Fuji H, Suenaga A, Morimoto G, et al. 2009. High-Performance Drug Discovery: Computational Screening by Combining Docking and Molecular Dynamics Simulations. *PLoS Comput Biol* 5:e1000528
36. Auffinger P, Westhof E. 1998. Simulations of the molecular dynamics of nucleic acids. *Current Opinion in Structural Biology* 8:227-36
37. Jayapal P, Mayer G, Heckel A, Wennmohs F. 2009. Structure–activity relationships of a caged thrombin binding DNA aptamer: Insight gained from molecular dynamics simulation studies. *Journal of Structural Biology* 166:241-50
38. Hansson T, Oostenbrink C, van Gunsteren W. 2002. Molecular dynamics simulations. *Current Opinion in Structural Biology* 12:190-6

39. Stavrakoudis A, Tsoulos I, Uray K, Hudecz F, Apostolopoulos V. 2011. Homology modeling and molecular dynamics simulations of MUC1-9/H-2K(b) complex suggest novel binding interactions. *Journal of molecular modeling* 17:1817-29
40. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, et al. 1995. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society* 117:5179-97
41. Karplus M, Petsko GA. 1990. Molecular Dynamics Simulations in Biology. *Nature* 347:631-9
42. Guvench O, MacKerell AD. 2008. Comparison of Protein Force Fields for Molecular Dynamics Simulations Molecular Modeling of Proteins. ed. A Kukol, 443:63-88: Humana Press. Number of 63-88 pp.
43. Plimpton S. 1995. Fast Parallel Algorithms for Short-Range Molecular Dynamics. *Journal of Computational Physics* 117:1-19 http://lammps.sandia.gov
44. Berendsen HJC, van der Spoel D, van Drunen R. 1995. GROMACS: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications* 91:43-56
45. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, et al. 2005. Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry* 26:1781-802
46. MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, et al. 1998. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins†. *The Journal of Physical Chemistry B* 102:3586-616
47. Oostenbrink C, Villa A, Mark AE, van Gunsteren WF. 2004. A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. *J Comput Chem* 25:1656-76
48. Jorgensen WL, Maxwell DS, Tirado-Rives J. 1996. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *Journal of the American Chemical Society* 118:11225-36
49. Guvench O, MacKerell A, Jr. 2008. Comparison of Protein Force Fields for Molecular Dynamics Simulations. In *Molecular Modeling of Proteins*, ed. A Kukol, 443:63-88: Humana Press. Number of 63-88 pp.
50. Hünenberger P. 2005. Thermostat Algorithms for Molecular Dynamics Simulations. In *Advanced Computer Simulation*, ed. C Dr. Holm, K Prof. Dr. Kremer, 173:105-49: Springer Berlin Heidelberg. Number of 105-49 pp.
51. McCammon JA, Gelin BR, Karplus M. 1977. Dynamics of folded proteins. *Nature* 267:585-90
52. Freddolino PL, Liu F, Gruebele M, Schulten K. 2008. Ten-Microsecond Molecular Dynamics Simulation of a Fast-Folding WW Domain. *Biophysical Journal* 94:L75-L7
53. Schaeffer RD, Fersht A, Daggett V. 2008. Combining experiment and simulation in protein folding: closing the gap for small model systems. *Current Opinion in Structural Biology* 18:4-9
54. Duan Y, Kollman PA. 1998. Pathways to a Protein Folding Intermediate Observed in a 1-Microsecond Simulation in Aqueous Solution. *Science* 282:740-4
55. Pérez A, Luque FJ, Orozco M. 2007. Dynamics of B-DNA on the Microsecond Time Scale. *Journal of the American Chemical Society* 129:14739-45

56. Rhinehardt K, Mohan R, Srinivas G, Kelkar A. 2013. Computational Modeling of Peptide - Aptamer Binding in Biosensor Applications. *International Journal of Bioscience, Biochemistry and Bioinformatics* 3:639-42
57. Schalley CA. 2012. *Analytical Methods in Supramolecular Chemistry: Vol. 1*. Wiley-VCH Verlag GmbH & Company KGaA

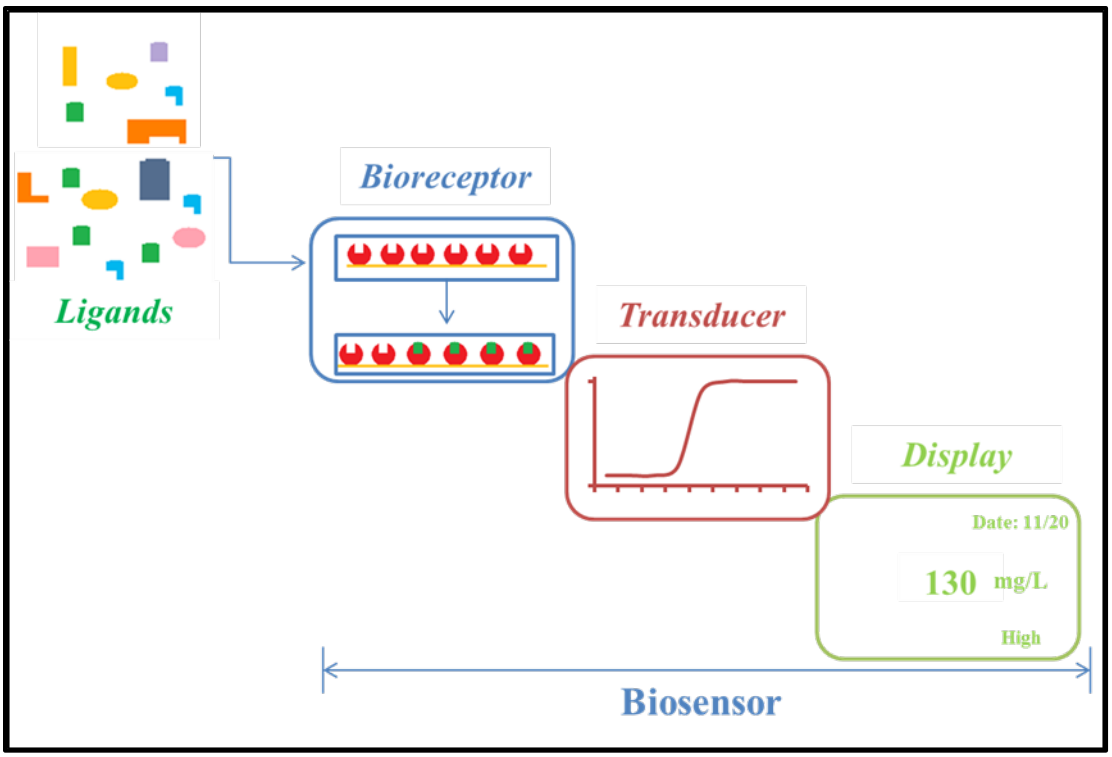


Figure 14: Schematic representation of biosensor working principle.

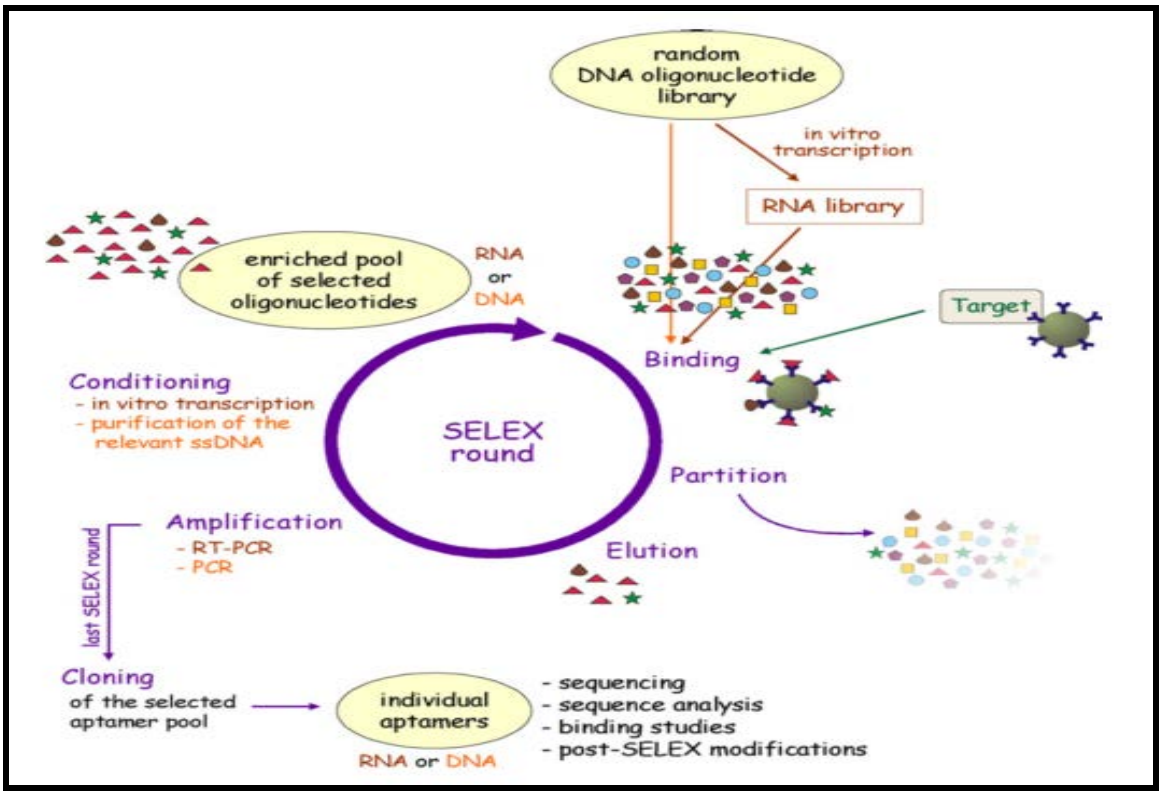


Figure 15: Schematic diagram of the SELEX process (reproduced from Reference (11) with permission).

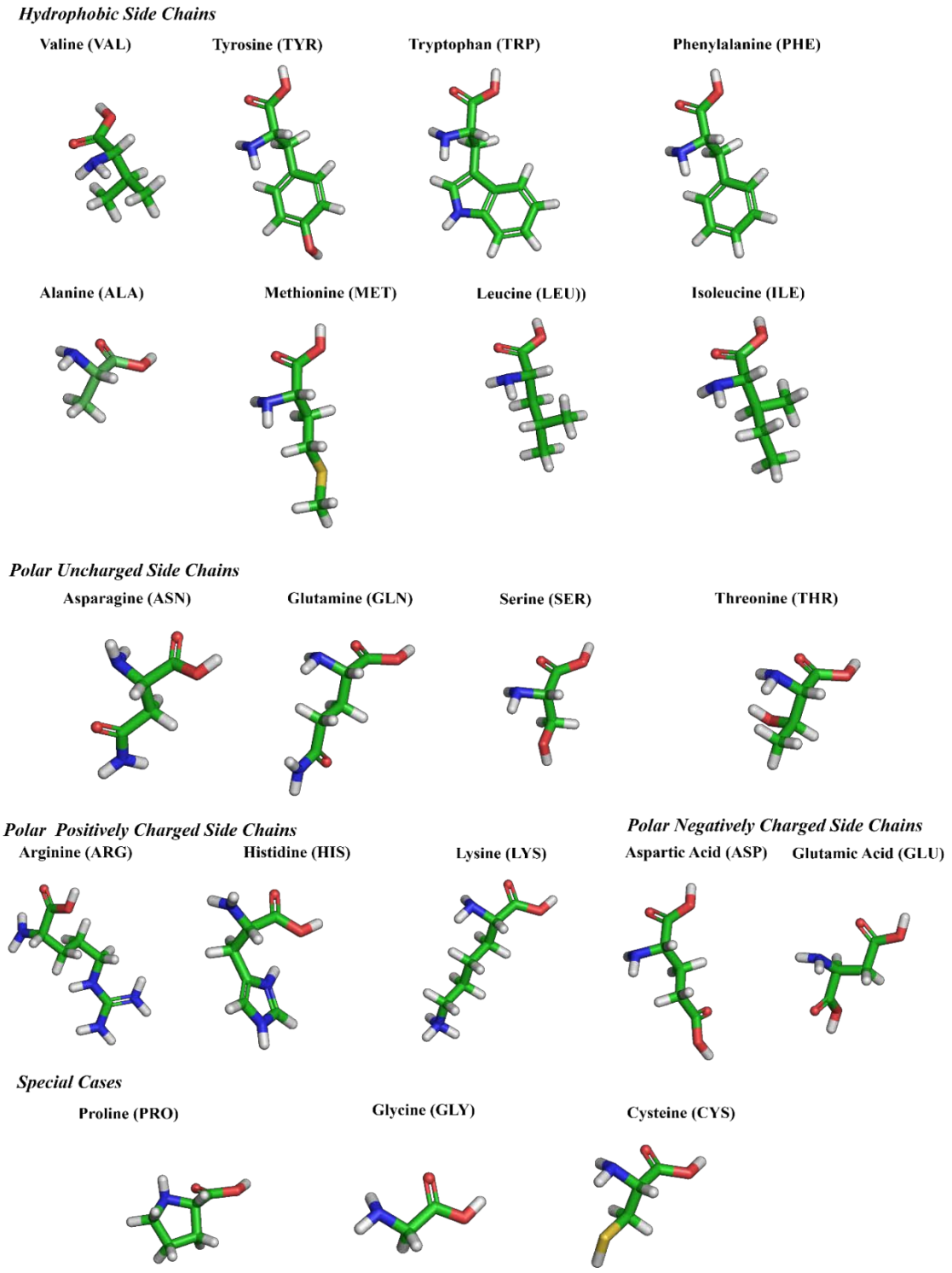


Figure 3: Chemical structures of amino acids shown in the stick representation. Color code: Carbon-green, oxygen-red, nitrogen-blue, hydrogen-white, and sulfur-yellow.

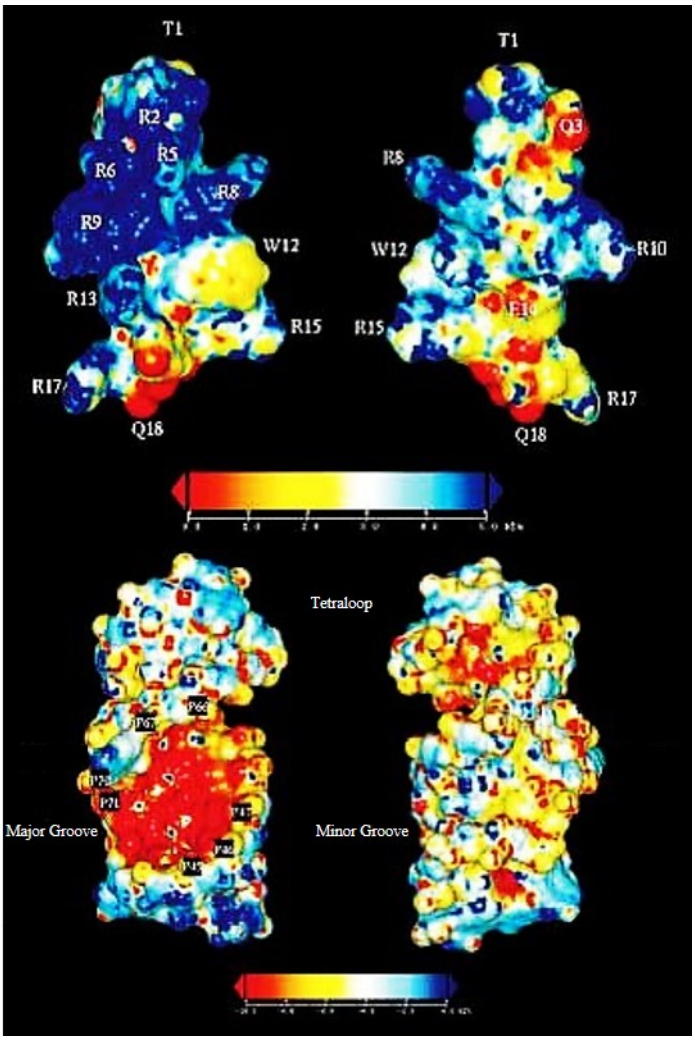


Figure 4: Electrostatic images of the Rev Peptide and RBE. A) Front and back views of the electrostatic surface of the Rev 34-50 peptide. Amino acid residues are indicated by single letter code and number B) Electrostatic surface of the RBE with groove and loop features indicated. Phosphate groups that are protected from chemical modification due to complex formation are indicated.(reproduced from Reference (33) with permission).

.

Figure 5: Stereo view of the A-NL model of the Rev34-50 - RBE complex (reproduced from Reference (33) with permission).

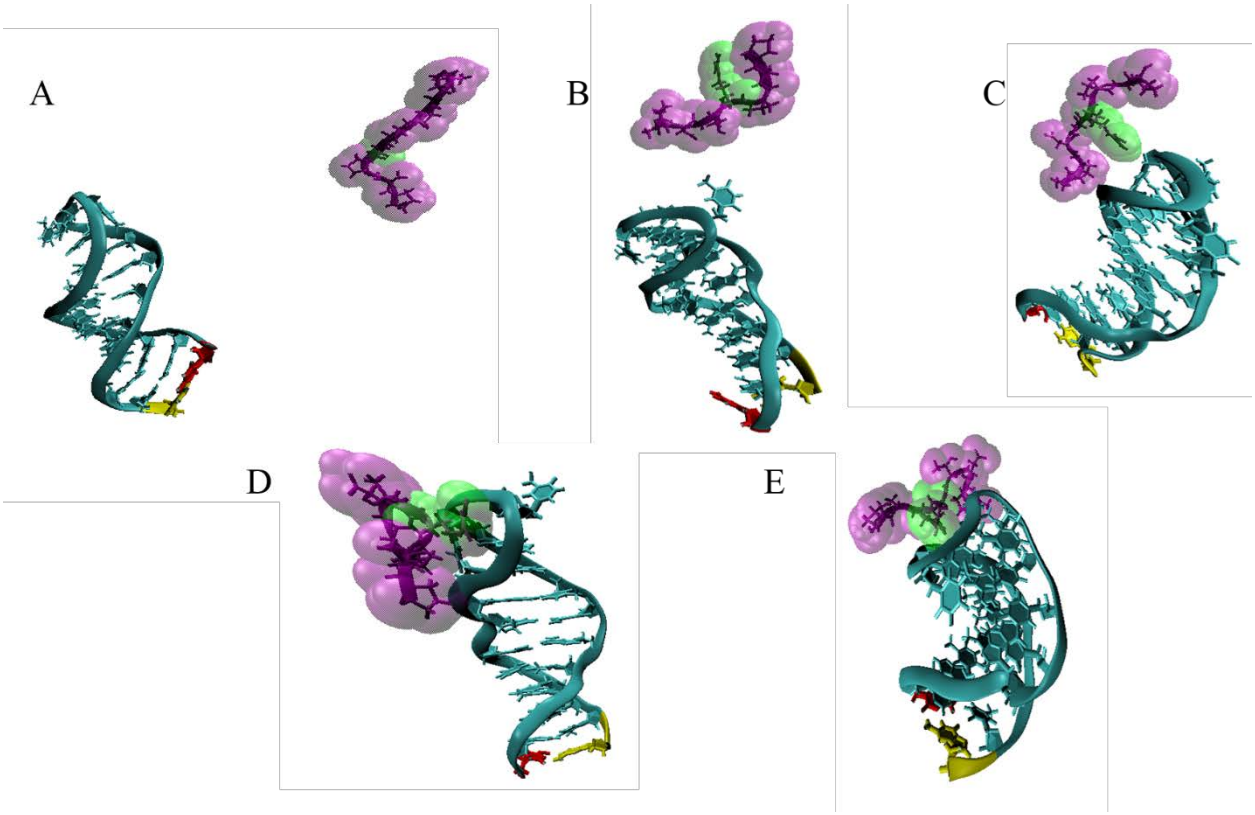


Figure 6: Visualization of Anti-MUC1 aptamer (blue) and MUC1-G peptide (purple) binding 300ns Simulation. (A) Starting peptide - aptamer configuration. (B) Peptide - aptamer configuration after 27ns. (C) Interation of the 11th thymine residue of the aptamer and the peptide backbone after 51ns of simulation. (D) Continued peptide - aptamer interaction at the open loop region of the aptamer and the arginine residue after 127ns. (E) Magnified image of the peptide - aptamer interaction at the end of simulation.

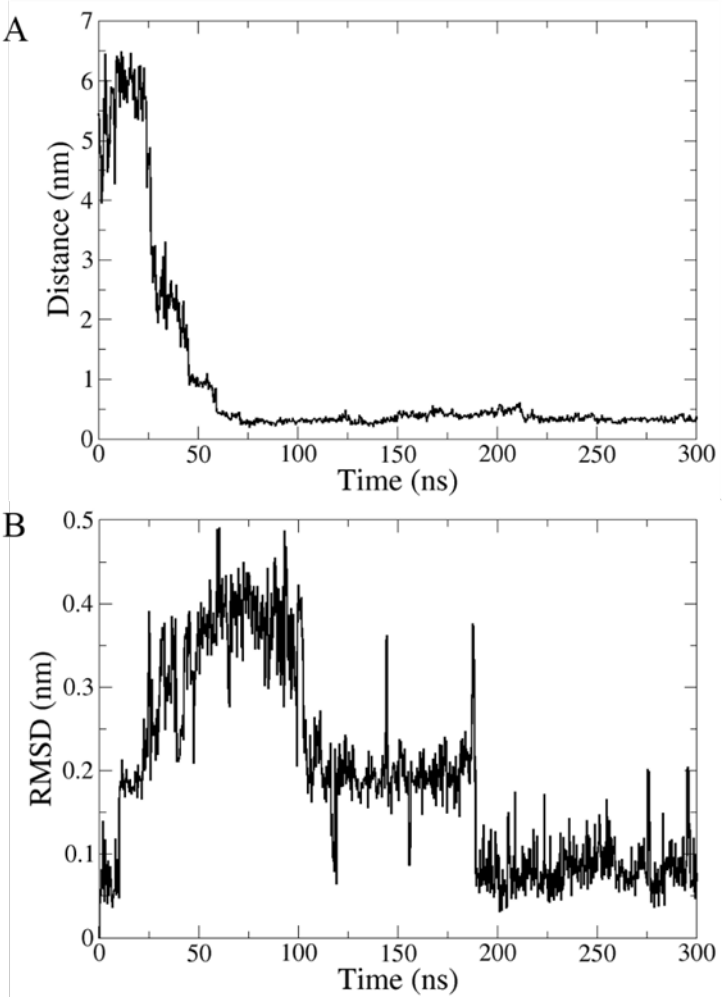
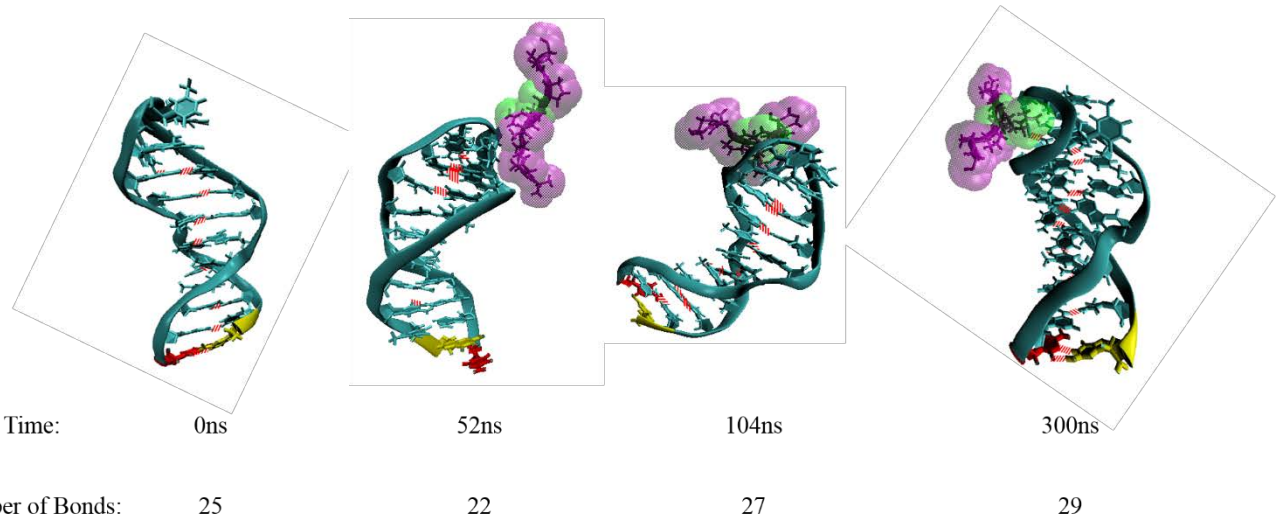


Figure 7: Quantitative analysis of MUC1-G peptide and aptamer for 300ns. A) Atom Distance between the aptamer and peptide in the binding region B) RMSD of the aptamer 5' – 3' ends during simulation.



Time:	0ns	52ns	104ns	300ns
Number of Bonds:	25	22	27	29

Figure 8: Number of hydrogen bonds in the Anti-MUC1 Aptamer (blue) during simulation with the MUC1-G peptide (purple). Corresponding simulation snapshots at 25ns, 52ns, 104ns and 300ns of the anti-MUC1 Aptamer with the peptide and hydrogen bonds (red dash lines) are shown.

5. Enabling Technologies for Intelligent Data Mining and Algorithms for Data Analysis with Large Variations and Outliers

R. George, Clark Atlanta University

Introduction

The volume, complexity and velocity of generation of the data in multi-scale modeling and complex large scale high performance computing demand the development of automated techniques to understand and use this data. While numerous algorithms have been developed to mine, and visualize data, techniques for discovering outliers and rare events in these types of large data sets have not similarly been fully developed. The research work under this grant focuses on the development of enabling technologies for such large scale data mining for application to nano-bio systems . Building on work performed during 2011-2012, this research effort focused on two different but related data mining tasks:

1. Visualization of Outlier Detection in Multi-dimensional Data
2. Development of Tools for Nano-Bio Applications

Details on the research work performed are described in Section I and Section II.

Section I: Outlier Detection Algorithms

Outlier detection is critically dependent on the nature of the data that is being analyzed. The type of the data, its distribution and dependencies require tuning of the approach, so that the outliers are detected appropriately. In this research we examine two applications, high dimensional spatio-temporal data, and network data and develop techniques for conducting outlier detection in each case. Both these scenarios are relevant to the nano-bio application and may be used in multiple contexts.

Section IA: Outlier Detection in High Dimensional Spatio-Temporal Data

Background

An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism [2]. In other words, for a given sample, a point is called an outlier if its behavior significantly deviates from the rest of the members of the sample. Outlier detection plays a vital role in several fields such as credit card fraud detection, severe weather prediction, image processing; pattern recognition and computer intrusions.

Substantial research has been done in outlier detection and these are classified into different types with respect to the detection approach being used. Exemplar techniques include Classification based methods, Nearest Neighbor based methods, Cluster based methods and Statistical based methods [16]. In the Classification-based approach [28], [29] a model is learnt from a set of labeled data points and then a test point is classified into one of the classes using appropriate testing. Support Vector Machine (SVM) based methods [27], methods based on Neural Networks [30] and Bayesian Networks based methods [22],[23],[31] belong to Classification based technique. The testing phase of this method performs fast since each test data is compared against the pre-built model. The accuracy of classification based methods relies on the availability of accurate pre classified examples for different normal classes, which is not always possible. Nearest Neighbor based methods [24],[26],[32] involve a distance or similarity measure defined between two data points. These techniques assume that typical data points lie in dense neighborhoods, whereas outliers are located distant from their closest neighbors. The main advantage of the Nearest Neighbor based method is that it does not make any assumptions about the distribution of data. Therefore having an appropriate distance measure helps to apply this method for any type of data sets. The LOF (Local Outlier Factor) method [13] is a Nearest Neighbor based approach. LOF gives each data point a *degree* of being an outlier via a relative Density Nearest Neighbor technique. A drawback of Nearest Neighbor based approach is the $O(n^2)$ computational complexity of the testing phase. Clustering based methods [19],[20],[21] use the approach of grouping similar data points into clusters. The performance of clustering

based techniques depends on the success of clustering algorithm; how accurately it gathers typical data into clusters. According to the basis of Statistical based methods [17],[18],[25] a point is an anomaly, because it is not generated by the stochastic model assumed. Here the normal data points are taken place in high probability regions of the stochastic model, while outliers are in the low probability regions of the model [16]. Both parametric [17] as well as non-parametric [25] methods are applied under statistical techniques. Therefore outlier detection using statistical methods is more accurate if the assumptions regarding the underlying data distribution are true. But the assumption that the data is generated from a particular distribution is often not correct.

The technique proposed in this research effort, the m-SNN (modified-Shared Nearest Neighbor) method is based on the non-parametric clustering algorithm Shared Nearest Neighbor (SNN) Approach developed by Ertöz et al. [6]. In contrast to parametric methods this technique does not assume an underlying probability distribution model for the data. m-SNN can also be regarded as a variant of nearest neighbor method. In this method, we consider the ratio between the summation of Euclidean distances to shared nearest neighbors and total number of shared neighbors. To differentiate between outliers and normal points hypothesis testing is used. m-SNN does not require any assumption about the data and does not require a threshold for declaring outliers. The number of nearest neighbors and the confidence level used are the only inputs required by m-SNN. We compare m-SNN approach with LOF method and Gaussian as a baseline parametric method and show that the algorithm presented can be used to detect outliers in distributions with different shapes and different densities. It is seen that the m-SNN compares well with the LOF in standard spatial data distributions and outperforms LOF in complex spatial data distributions. Finally, we show in a real situation, the m-SNN algorithm is effective in tracking of the progress of Hurricanes Katrina through the Gulf of Mexico.

Related Terms and Definitions

We introduce and define some of the terms which we use. First we define Neighbor Similarity, Density, Neighbor Similarity Distance and Sparseness; then terms local and global outliers. The definitions of Similarity and Density are based on the notions given in [6] and we define the terms p-value, null hypothesis (H_0) and alternative hypothesis (H_a) relating to the proposed technique.

Definition 1: Neighbor Similarity – For a given data point *u* the neighbor similarity is defined as the number of nearest neighbors being shared between *u* and its corresponding nearest neighbor. For example, as shown in Figure1 considering only three nearest neighbors, *u*'s nearest neighbors are *A*, *B* and *C* while *A*'s nearest neighbors are *B*, *D* and *E*. Hence, *B* is shared by both *u* and *A*. Therefore the neighbor similarity between *u* and *A* is 1.

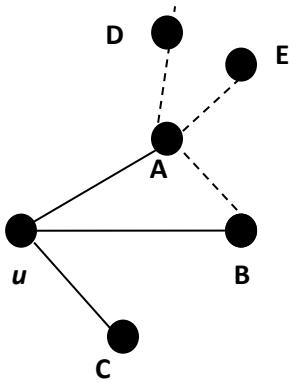


Figure I.1: Shared Nearest Neighbors

Definition 2: Density – For a given data point 'u' the density is defined as total number of neighbor similarities between u and its nearest neighbors (v_i).

$$density(u) = \sum_{i=1}^{K} NeighborSimilarity(u, v_i)$$

Definition 3: Neighbor Similarity Distance – For a given data point 'u' the neighbor similarity distance is defined as sum of Euclidean distances between u and its all shared neighbors.

Definition 4: Sparseness – For a given data point 'u' the neighbor sparseness is the ratio between Neighbor Similarity Distance and Density.

$$sparseness(u) = \frac{Neighbor\ Similairty\ Distance}{Density}$$

In complex real world situations both global and local outliers may be found [1]. A global metric would be unsuccessful of detecting local outliers. Locally an outlier could be discovered relative to a dense area of points. Conversely, a point with higher *sparseness* might not be considered an outlier, if it is in a neighborhood of a sparse set of data.

Definition 5: Null and Alternate Hypothesis

The null hypothesis and alternative hypothesis statements for m-SNN method are expressed below.

Null Hypothesis = H_0: Point u is not an outlier (p-value >= $\boldsymbol{\tau}$)

Alternative Hypothesis = H_a: Point u is an outlier (p-value < $\boldsymbol{\tau}$)

p-value is the maximum probability of observing a test statistic as the null hypothesis is true. p-value is also known as observed level of significance while $\boldsymbol{\tau}$ is the actual significance level. In [14], the p-value is obtained as the fraction of points in the class that have strangeness greater than or equal to that of the point. According to m-SNN algorithm, p-value of a point is calculated as the fraction of points in the class that have sparseness less than to that of the corresponding point. Therefore larger p-value indicates the high probability of accepting null hypothesis whereas smaller p-value implies the high probability of rejecting null hypothesis and accepting the alternative hypothesis.

Technical Approach

The algorithm of m-SNN method is based on shared nearest neighbor approach and p-value technique of hypothesis testing for finding outliers. Methodologically, for each data point we calculate its ***k*** nearest neighbors by using the Euclidean distance measure. Next, we calculate the Neighbor Similarity of the corresponding data instance i.e., for each data point we calculate the number of neighbors being shared between current node and it nearest neighbors. Subsequently, we calculate the Euclidean distance between current node and shared neighbors. Then the Density of the point is calculated by summing up its all Neighbor similarities. Finally, Sparseness is calculated by taking the ratio between sum of Euclidean distances to shared neighbors and density.

The following pseudo code explains our algorithm. Here *knn* is to store k nearest neighbors for a given data point, and findkNN finds the k nearest neighbors using Euclidean distance. In the pseudo code, knnI and knnJ are the k nearest neighbors for i^{th} data point and its j^{th} neighbor respectively. For a given data point Euclidean distance to shared neighbors and number of such nodes are stored in temporary variables distance and density respectively. The calculated sparseness is stored in *sparseness*. We take *n* as the number of data points in the sample. Table 1 shows the pseudo code of m-SNN algorithm.

As our method needs to find the *k* nearest neighbors for each data point, it is required to calculate the Euclidean distance between each other data points. Hence, since we have *n* data points the complexity of calculating Euclidean distance is equal to $O(n^2)$. Finding *k* nearest neighbors for a given data point based on Euclidean distance can be done in a constant time by finding the *k* shortest distanced points to the original point. This does not require sorting all the data points. Finally to find outliers we need to compare each data point with each other remaining data points, thus resulting $O(n^2)$ complexity.

Procedure: m-SNN Based Outlier Detection
Inputs: data[], a set of data points; k, the number of nearest neighbors; **τ**, the confidence level
Output: Print Outliers
Assume knn[] stores the k nearest neighbors for the data point, density [],To store shared neighbor density;

```
// Finding k-nearest neighbors for all the data points
for i = 1 to n
   knn[i] = findkNN(data[i])  //Find k-nearest neighbors for data point i and store in the array
end for

//Finding the shared neighbor nodes and distances to them
for i = 1 to n
     distance = 0
     density = 0
     knnI [] = knn[data[i]] // Get neighbors for data point i

      // Find the shared neighbors of data point i
     for j = 1 to k
          knnJ [] = knn[knnI[j]]   // Get neighbors for jth neighbor of data point i
          for x = 1 to k
               for  y = 1 to k
                    if knnJ[y] == knnI[x]   // checking for overlapping
                         // Calculate the distance to the overlapping data points
                         distance = distance + euclidean_distance(data[i],   knnJ[y])
                          density = density + 1
                    end if
               end for
          end for
      end for
      sparseness[i] = distance/density   // Calculate sparseness for data point i
end for

// Printing outliers
for i = 1 to n
     count = 0
     for j = 1 to n
          if  sparseness [i] >=  sparseness [j] then
               count = count + 1
           end if
          p-val = 1 - (count -1)/n
           if p-val < τ then   // If p-value less than τ, then point i is an outlier
              data[i] is an outlier
          else
              data[i]  is not an outlier
           end if
     end for
end for
```

Table I.1: The Modified Shared Nearest Neighbor Algorithm

Experimental Results

This section describes experiments and results with synthetic data sets followed by how the data was generated. We ran the experiments where τ was taken as 0.05. i.e., these experimental results are with 95% confidence.

To cover the broad range of applications we generated three main categories of spatial data sets, 1.clusters, 2. Complex spatial paths. In each case we use probabilistic distribution based data generation which takes user inputs to decide parameters of the data pattern. i.e., identify variables and then use a probabilistic model to generate the required number of data points and outliers.

After generating data, each set of data points with feature scaling was tested both with proposed outlier detection method and with the LOF technique. Since LOF technique gives a degree of being an outlier of a point, there is no clear cutoff value differentiating normal points and outliers. For comparison purposes, we considered a data point with LOF value greater than 2.0 as an outlier (a stricter standard than suggested in the original paper) and the results are tabulated.

Cluster Data Analysis

In our analysis we generated data set with two clusters with 1015 total data points where 15 of them were generated as global outliers. After applying m-SNN technique with tau 0.05, all the expected global outliers were detected and 35 additional points were detected where some of those can be considered as local outliers as shown in Figure 2. LOF approach also was able to detect all above labeled outliers correctly producing all LOF values corresponding to outliers greater than 2.0.

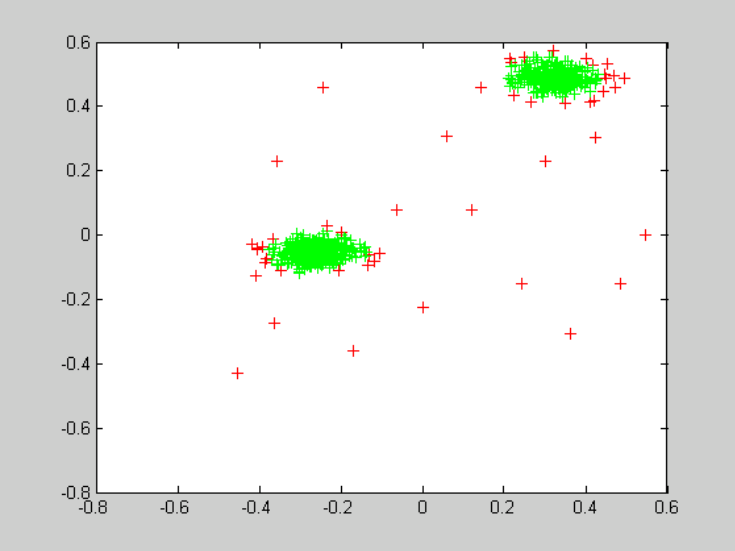


Figure I.2. Outlier detection: m-SNN Technique - Two clusters with 1015 data points

Path Data Analysis

To check how effective our proposed method, we generated data sets with different behaviors. Here we have a set of points that are located on curved paths and some deviated points as well. This set consists with 1000 normal data and 23 significantly deviated points. Figure 3 represents the output results of outlier detection using the proposed method. Generating equivalent results to m-SNN approach, LOF technique also detected 22 outliers with LOF values greater than 2.0.

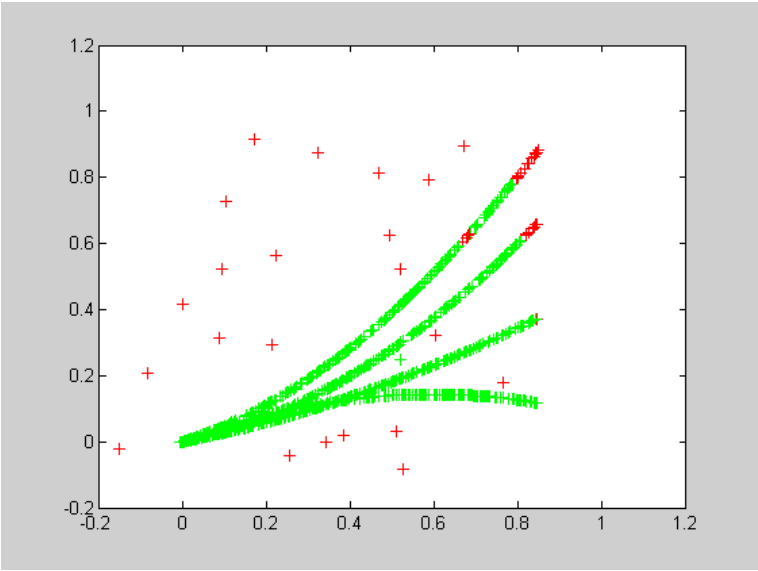


Figure I.3. Outlier detection: m-SNN Technique - Four curved paths with 1023 data points

Spiral Path Data Analysis

A synthetic spiral data with 1010 total data points and including 10 possible outliers was generated. As shown in Figure 4, m-SNN algorithm could detect 10 of those expected deviated points as outliers and it detected 50 points altogether as outliers. LOF approach detected 8 outliers correctly having Local Outlier Factor greater than 2.0.

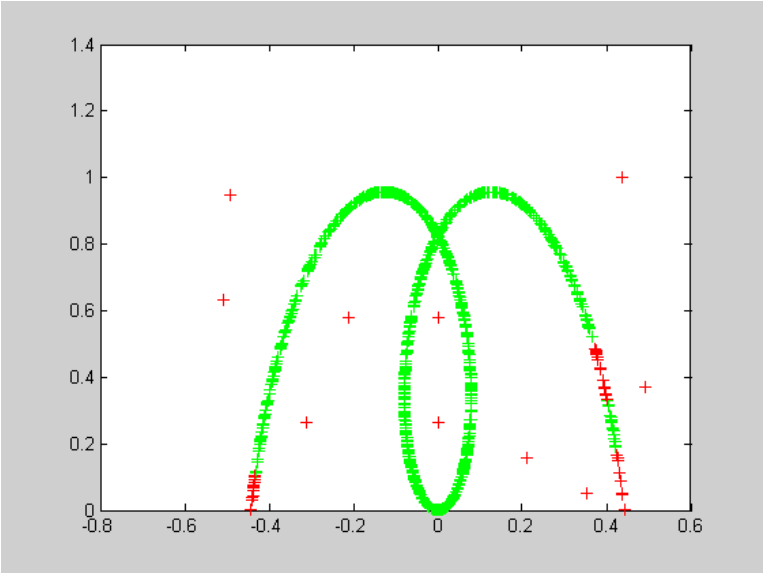


Figure I.4. Outlier detection: m-SNN Technique - Spiral path with 1010 data points

Circular Path Analysis

As the next step, two circled paths were generated which contains total of 1035 data points. This includes 35 points that can be regarded as outliers and 1000 typical data. The proposed method was successful to detect 51 points as outliers including all 35 expected outlier points which is graphically illustrated in Figure I-5. Only 29 points were detected correctly as anomalous data by LOF with minimum LOF of those being 2.0.

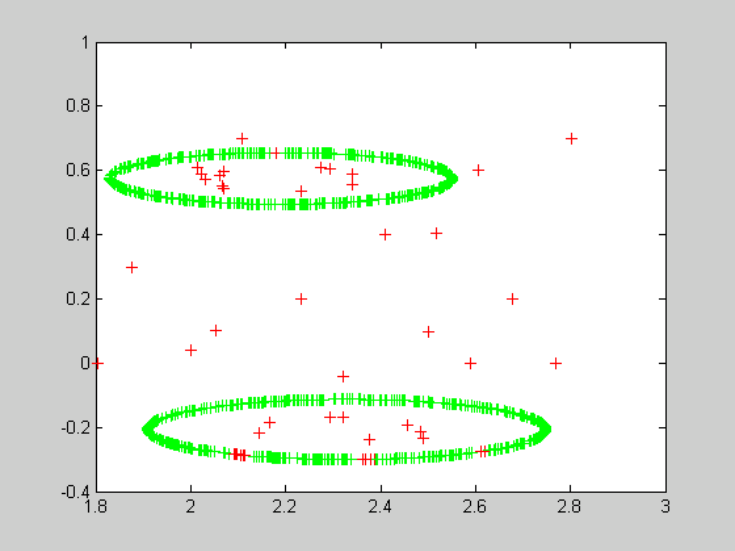


Figure I.5. Outlier detection: m-SNN Technique - Two circular paths with 1035 data points

As a control method above data sets were tested with Gaussian approach too. All the results obtained are summarized in Table 2 to demonstrate True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) values as percentages. FP is also known as Type 1 error and FN is also known as Type 2 error in statistics. Moreover in some circumstances TP is entitled Sensitivity and TN is named as Specificity.

Table I.2: Summarized results for the experiments

		TP(%)	FP(%)	TN(%)	FN(%)
Clusters	m-SNN	100.0	3.5	96.5	0.0
	LOF	100.0	0.08	99.2	0.0
	Gaussian	7.1	0.0	100.0	92.9
Curved Paths	m-SNN	95.7	2.9	97.1	4.3
	LOF	95.7	0	100	4.3
	Gaussian	30.4	6.1	93.9	69.6
Spiral Path	m-SNN	100.0	4.0	96.0	0.0
	LOF	80.0	0.0	100.0	20.0
	Gaussian	40.0	3.8	96.2	60.0
Circular Paths	m-SNN	100.0	1.6	98.4	0.0
	LOF	82.9	0.0	100.0	17.1
	Gaussian	2.9	0.0	100.0	97.1

According to the tabulated results of Table 2, it is clear that m-SNN has very high TP, TN percentages and very low FP, FN percentages. Therefore m-SNN is enhanced in accuracy and performance when detecting outliers comparing to Gaussian approach. Also the proposed method performs at least as equivalent to LOF approach. In particular comparing the values corresponding to non clustering data sets it is evident that m-SNN is more robust when finding outliers. This shows that Gaussian approach failed to find anomalies correctly as it gave low TPs and high FNs. m-SNN can be successfully used to detect outliers when the data distribution model is not known exactly. In addition to that, Table 2 shows that m-SNN is stronger in anomaly detection of datasets having arbitrary shapes and densities.

Application 1: Tracking Hurricane Katrina Using m-SNN

In this section we discuss experiments and results of applying m-SNN technique to real data sets. For this experimenting purpose, we chose data recorded from buoys located in Gulf of Mexico. Description of datasets, Experiments and results are presented below. There are several buoys located in Gulf of Mexico area and data recorded from those buoys are used for purposes including weather forecasts, marine forecasts and climate predictions. The buoys record data by

making a number of routine measurements such as wind direction, wind speed, wave height, barometric pressure, air temperature, sea surface temperature and dew point temperature.

For experimenting with real datasets and detecting outliers, we chose 17 datasets, which contain hourly basis weather data from 17 buoys located at specific geographic locations in the Gulf of Mexico during year 2005. From original data sets [1], we selected five features wind direction, wind speed, barometric pressure, air temperature and water temperature at each hour.

Experiments

We conducted experiments to detect outliers in two different ways i.e. detecting outliers at specific buoy location and detecting outliers at specific time.

1. **Detecting spatial outliers**

To find outliers which occurred at each buoy location, first each data set was tested with m-SNN algorithm to detect outliers, taking tau as 0.01. Then the detected outliers are analyzed and grouped into time durations when that have been occurred. These results clearly show that outliers are associated to time periods when major hurricanes (Katrina, Rita, Wilma) occurred. The time periods when outliers appeared differ from one buoy location to another, according to the hurricane track. Therefore the occurrence of outliers correlated with actual track's of hurricanes. Table 1 shows basic details of buoy datasets and how many outliers were detected at each buoy location. While Table 2 shows detected major outlier time intervals resulted from hurricane Katrina accordance with corresponding buoys, Table 3 includes detected major outlier time periods resulted from hurricane Rita and those corresponding buoys.

Table I.3: Summarized results for the experiments with Buoy Data

Buoy ID	Geographic Location		No. of instances	No. of detected outliers
	Latitude	Longitude		
42001	26.0 N	90.0 W	8741	87
42002	26.0 N	94.0 W	8729	87
42007	30.1 N	88.9 W	6705	67
42019	27.9 N	95.0 W	8676	86
42020	27.0 N	96.5 W	8685	86
42035	29.2 N	94.4 W	8738	87
42036	28.5 N	84.5 W	8346	83
42038	27.4 N	92.6 W	8095	80
42039	28.8 N	86.0 W	5091	50
42040	29.2 N	88.3 W	8251	82
FWYF1	25.6 N	80.1 W	8153	81
GDIL1	29.3 N	90.0 W	6264	62
LONF1	24.8 N	80.9 W	8750	87
MLRF1	25.0 N	80.4 W	8313	83
PTAT2	27.8 N	97.1 W	8746	87
SANF1	24.5 N	81.9 W	6295	62
SMKF1	24.6 N	81.1 W	6543	65

Table I.4: Outlier buoys and time periods resulting from Katrina

Buoy ID	Major outlier time periods resulted from hurricane Katrina
42001	08/28 10:00 to 08/29 12:00
42035	08/28 20:00 to 08/29/06:00
42040	08/29/04:00 to 08/30/03:00
FWYF1	08/25 16:00 to 08/26 11:00
GDIL1	08/28/23:00 to 08/29/12:00
LONF1	08/26 02:00 to 08/26 15:00
MLRF1	08/26 01:00 to 08/26 15:00
SANF1	08/26/05:00 to 08/26/23:00

Table I.5: Outlier buoys and time periods resulting from Rita

Buoy ID	Major outlier time periods resulted from hurricane Rita
42001	09/22 07:00 to 09/23 15:00
42002	09/23/ 13:00 to 09/24 00:00
42019	09/23 17:00 to 09/24 14:00
42038	09/27 19:00 to 09/28 17:00
LONF1	09/20 08:00 to 09/20 17:00
MLRF1	08/26 01:00 to 08/26 15:00
PTAT2	09/24/15:00 to 09/24 19:00
SANF1	09/20/13:00 to 09/20/22:00 (No data available after 09/20/22:00)
SMKF1	09/20 12:00 to 09/20 17:00

2. Detecting temporal outliers

To capture the outliers among 17 buoys at specific time, first we selected data corresponding to that specific time from each and every buoy. Then a new dataset was created by adding selected data into one set. Next new dataset was tested with m-SNN technique to detect outliers. This procedure was conducted to each new dataset corresponding to specific times. During this experiment tau was taken as 0.1, which means all the results obtained here is with 90% confidence level. We used m-SNN algorithm to detect outliers from August 26, 2005 at 2:00 am until August 29, 2005 at 2:00 pm on each 12 hours basis. Following Figure 1 to Figure 8 shows these resulting locations of outliers at different times due to hurricane Katrina.

Note: The green markers indicate location of the buoys, the orange marker buoys with outlier readings, and the red tailed ellipse the eye of the hurricane.

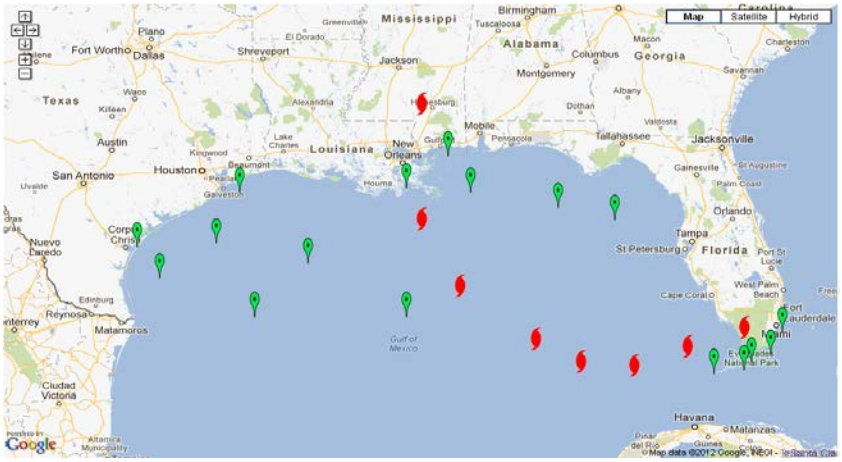


Figure I.6: Path of Hurricane Katrina (8/26/2005 – 8/29/2005)

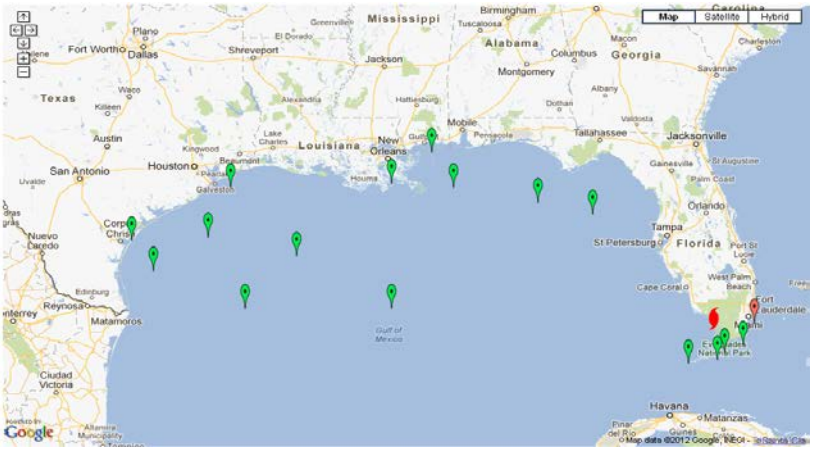


Figure I.7: Hurricane Katrina and Buoy Outlier (8/26/2005: 2AM)

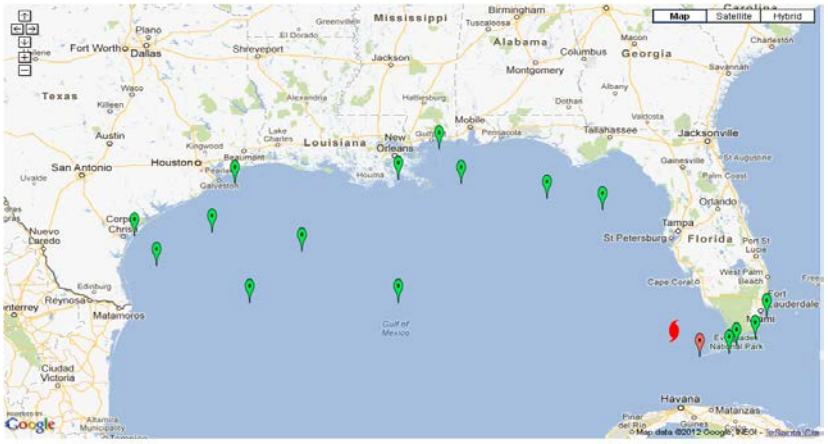


Figure I.8: Hurricane Katrina and Buoy Outlier (8/26/2005: 2PM)

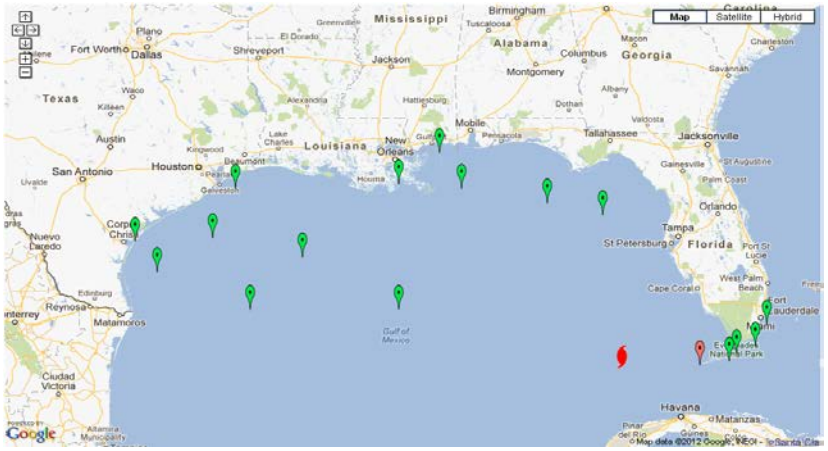


Figure I.9: Hurricane Katrina and Buoy Outlier (8/27/2005: 2AM)

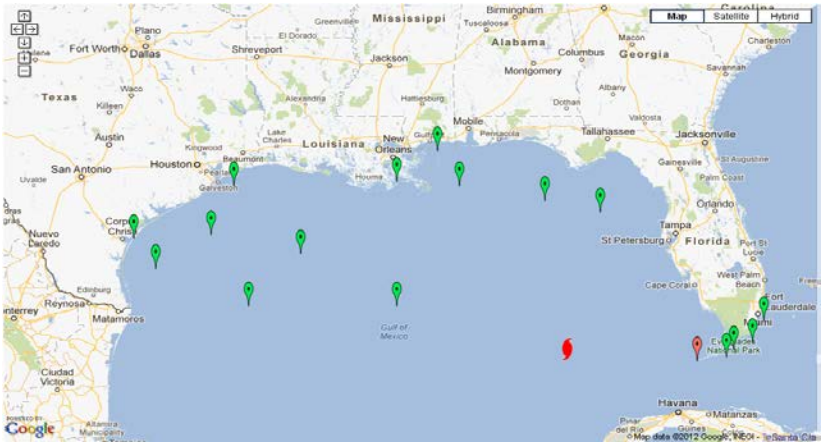


Figure I.10: Hurricane Katrina and Buoy Outlier(8/27/2005: 2PM)

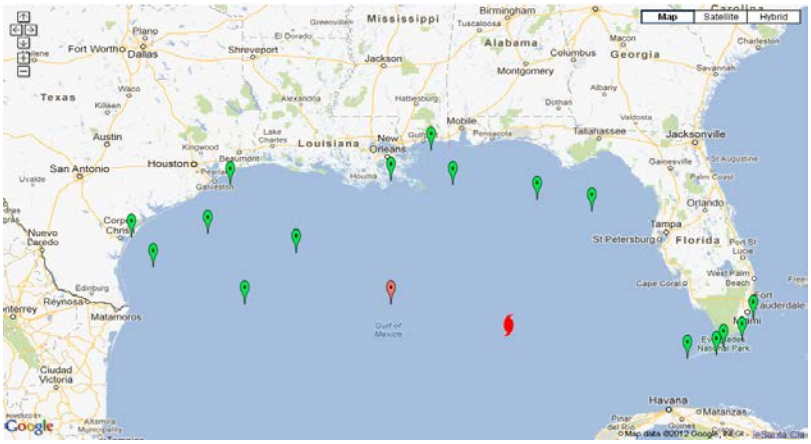


Figure I.11: Hurricane Katrina and Buoy Outlier(8/28/2005: 2AM)

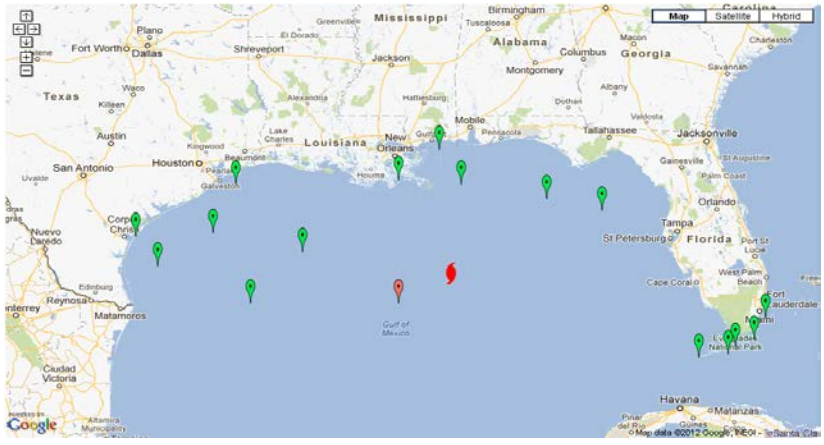


Figure I.12: Hurricane Katrina and Buoy Outlier (8/28/2005: 2PM)

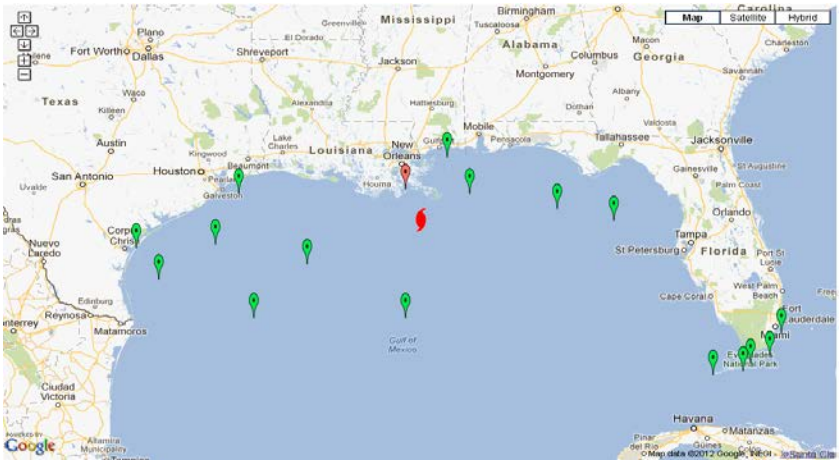


Figure I.13: Hurricane Katrina and Buoy Outlier (8/29/2005: 2AM)

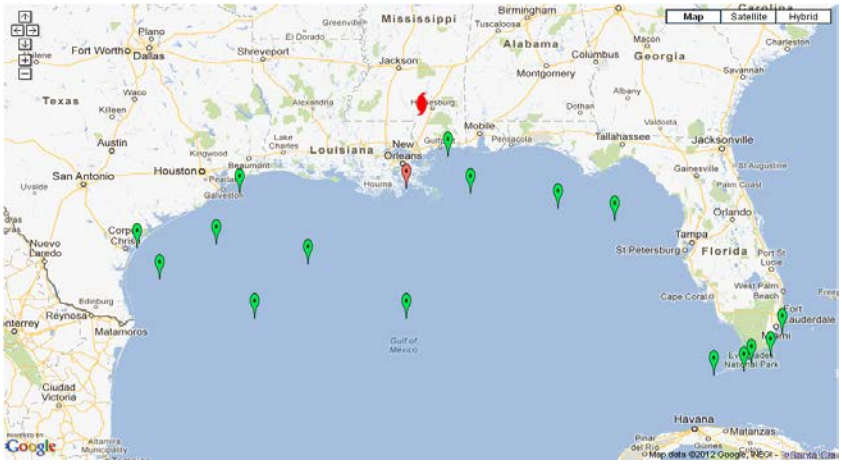


Figure I.14: Hurricane Katrina and Buoy Outlier (8/29/2005: 2PM)

Discussion of Results

In this research effort, we have described an algorithm m-SNN which is capable of detecting outliers in different types of data sets. This method is a combination of Shared Nearest Neighbor and distance based methods that avoids assumptions about data distributions and uses hypothesis testing to detect outliers. We compared the proposed technique with LOF and also with a baseline Gaussian approach on several data sets that containing different patterns of data distributions in two dimension environment. The proposed m-SNN technique results very high true positive and true negative values as well as very low false positive and false negative values. According to the experimentation results this technique provides good results on a variety of synthetic datasets when detecting both global and local outliers. The m-SNN approach produces

outlier detection results equivalent or better than other two comparative methods. The m-SNN was tested on sensor data from buoys in the Gulf of Mexico for 2005. It is seen that the technique can accurately capture outliers from key events, Hurricanes Katrina and Rita that occurred during this time period. Further, the outliers can be used to track the path of the Hurricanes consistently.

We have developed and are testing a version of the m-SNN algorithm on streaming data. This algorithm would help in situations where the data is being analyzed in real time. Future work is also being planned on a more version of m-SNN. The current algorithm is of order $O(n^2)$, which is not well suited to evaluation of very large data sets. We are currently formulating an approach where this might be improved to O(n lg(n)) and also the application of HPC to this problem.

The m-SNN is a general purpose algorithm which may be applied in the nano-bio field. While, our work has focused on spatial data it may be extended to 3 or higher dimensional spatial data. The algorithm as formulated currently uses kernels to characterize the various dimensions, for example, position, temperature, windspeed, etc. By appropriately weighting the kernels, the problem may be aligned to particular domains, thus domain independent permitting outlier detection. .

Application II: Outlier Detection in Network Data

Outlier detection is an important data mining task that is focused on the discovery of objects that are exceptional when compared with a set of observations that are considered typical. In many data analysis tasks a large number of variables are being recorded or sampled. One of the first steps towards obtaining a coherent analysis is the detection of outlaying observations. Although outliers are often considered as an error or noise, they may carry important information. Detected outliers are candidates for aberrant data that may otherwise adversely lead to model misspecification, biased parameter estimation and incorrect results. These objects are important since they often lead to the discovery of exceptional events. Substantial research has been done in outlier detection and these are classified into different types with respect to the detection approach being used. Exemplar techniques include Classification based methods, Nearest Neighbor based methods, Cluster based methods and Statistical based methods [19]. In the Classification-based approach [31], [32] a model is created from a set of labeled data points and then a test point is classified into one of the classes using appropriate testing. Support Vector Machine (SVM) based methods [30], methods based on Neural Networks [33] and Bayesian Networks based methods [25],[28],[34] belong to Classification based technique. The testing phase of this method performs quickly as each test data is compared against the pre-built model. The accuracy of classification based methods rely on the availability of accurate pre classified examples for different normal classes, which is rarely found. Nearest Neighbor based methods [27], [29], [35] involve distance or similarity measures which is defined between data points. In this researcg, we discuss new method to find out outlier that is based on graph. This method is

efficiently reduces the search space by finding a candidate set of vertices whose betweenness centralities can be computes their betweenness centeralities using candidate vertices only.

The Betweenness Centrality (BEC) is a measure that computes the relative importance of a vertex in a graph, and it is widely used in network analyses such as a social network analysis, biological graph analysis, and road network analysis. In the social network analysis, a vertex with higher centrality can be viewed as a more important vertex than a vertex with lower centrality. The BEC of a vertex in a graph is a measure used for the participation of the vertex in the shortest paths in the graph. There are many previous works on the BEC problem. The concept of the BEC is proposed in [35], but the definition proposed in [40] is more widely used. Recently, many variants of the definition are proposed in [38]. [37] improves the computation time of the BEC based on a modified breadth-first search algorithm and the dependency of a vertex, and it is the fastest known algorithm that computes the exact BEC of all the vertices in a graph. The computations of the shortest paths between all pairs of vertices are time consuming. Therefore, another definition of BEC is proposed [22]; this based on a random walk. In [42], each vertex has a probability of visiting its neighbor vertices. Also, [39], [36] and [41] propose approximation algorithms for computing the betweenness centrality. [43] and [44] adopt the betweenness centrality for detecting communities in a social network.

Although many works on calculating the BEC exist and the BEC is one of the major methods used in analyzing social network graphs, none of the works for computing the BEC address the problem of updating BEC. In this research we propose betweenness centrality to find out outliers for network type data.

The next section describes related terms and definitions which are used throughout the report. Furthermore, it outlines the approach that explains the algorithm behind the BEC approach. To get a better understanding and to demonstrate the accuracy of BEC, several experiments were conducted with different kinds of synthetic data sets those are described in more detail in experimental results section. By applying BEC technique to find outliers in synthetic data sets and compare with another method which is call as modified-Shared Nearest Neighbor. Finally concludes the report with a discussion of the performance, accuracy and the importance of the proposed technique. From the results of experiments, it is clear that this technique gives better results in comparison to modified-Shared Nearest Neighbor by giving higher true positive and true negative values and very low false positive and false negative values for the network type data.

The m-SNN (modified-Shared Nearest Neighbor) method [3] is based on the non-parametric clustering algorithm the Shared Nearest Neighbor (SNN) Approach developed by Ertöz et al. [9]. In this method, consider the ratio between the summation of Euclidean distances to shared nearest neighbors and their total number of shared neighbors. To differentiate between outliers

and normal point's hypothesis testing is used, which is the similar technique used by Babara et al [18] and Rogers [4].

Related Terms and Definitions

Betweenness Centrality

A measure that computes the relative importance of a vertex in a graph. The formal definition is presented below.

A graph is represented by $G=(V,E)$, where V is the set of vertices, and $E \subseteq V \times V$ is the set of edges. A path in a graph is represented by a sequence of vertices, $(v_1,...,v_n)$ where $v_i, v_j \in V$ for $1 \leq i, j \leq n, i \neq j$, except possible $1=n$.

Definition 1 (Betweenness Centrality). The betweenness centrality of a vertex $v_j \in G$ is:

$$c(v_j)=\sum_{i,k}\frac{\sigma_{v_i,v_k}(v_j)}{\sigma_{v_i,v_k}(1)}$$

Where, $v_i, v_j, v_k \in V$, $i \neq j \neq k$, $\sigma_{v_i,v_k}(v_j)$ is the number of shortest paths between v_i and v_k that include v_j, and σ_{v_i,v_k} is the number of shortest paths between v_i and v_k. The betweenness centrality can be computed as follows:

1 For each pair of vertices (v_s and v_t), compute the shortest paths between the two vertices.

2. For each pair of vertices, compute the ratio of each vertex participating in the shortest path(s). The ratio is the number of shortest paths between v_s and v_t that go through v_j divided by the number of shortest paths between v_s and v_t.

3. Accumulate the ratio for all pairs of vertices.

Definition 2 (Adjacency Matrices).

The adjacency matrix of a finite graph *G* on n vertices is the $n \times n$ matrix where the non-diagonal entry a_{ij} is the number of edges from vertex *i* to vertex *j*, and the diagonal entry a_{ii}, depending on the convention, is either once or twice the number of edges (loops) from vertex *i* to

itself. Undirected graphs often use the latter convention of counting loops twice, whereas directed graphs typically use the former convention

	A	B	C	D	E	F
A	0	1	1	1	1	-
B	1	0	-	1	1	-
C	1	-	0	-	-	1
D	1	1	-	0	-	1
E	1	1	-	-	0	1
F	-	-	1	1	1	0

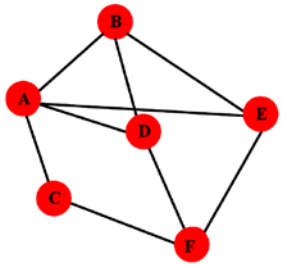


Figure I.15:Undirected graph with adjacency matrix.

In figure 1 shows the adjacency matrix for undirected graph. A, B, C, D, E, and F are represent the node. In diagonal all are zero and if two nodes are connected matrix denote by the value of 1.

Technical Approach

This outlier detection method is based on BEC for network data and p-value technique of hypothesis testing for finding outliers. For each data point we calculate its BEC by using adjacency matrix for network data. To find out the adjacency matrix for the data set, we calculate the shortest paths through nodes in destination IP addresses. The shortest path that was calculated creates adjacency matrix from friend nominations by utilizing sparce matrices in order to increase computational speed. Our calculation based on undirected network type data. The calculation for adjacency matrix yields an adjacency matrix from friendship nominations stored as a sparce matrix. The resulting adjacency matrix will include Id numbers in the first row and first column. To find the BEC, the calculation of the influence domain of each ego in a given adjacency matrix for a given step, returns the undirected BEC for each node of undirected adjacency matrix 'adj'. Matrix 'adj' must be an undirected network and may or may not be sparse. Simple to change if directed.'adj' is assumed to have id numbers in the first row and also, this code could probably be more vectorized to speed up calculations for large adjacency matrices.

As our method needs to find the adjacency matrix for each data point, it is required to calculate the shortest path between each other data points. Since we have n data points the complexity of calculating shortest path is equal to $O(n^2)$. Finally to find outliers we need to compare each data point with each other remaining data points, thus resulting $O(n^2)$ complexity.

Experimental Results

This section describes the experiments and the results with synthetic data sets followed by how the data was generated. The experiment was run where τ was taken as 0.05. i.e., these experimental results are with 95% confidence.

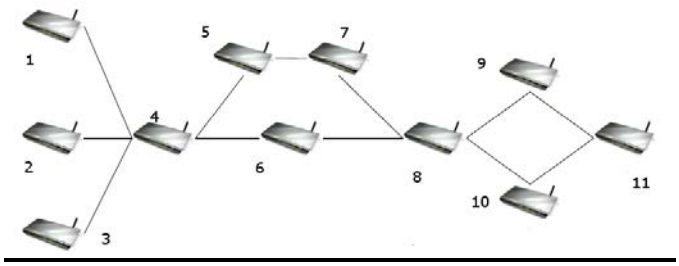


Figure I.16: Shortest paths through nodes in destination IP addresses.

To cover the broad range of applications, network type data sets were generated. We apply a rigorous set of tests to the path data to understand the strength (or weakness) of the method. In all cases we use probabilistic distribution based data generation which takes user inputs to decide parameters of the data pattern. i.e., identify variables and then use a probabilistic model to generate the required number of data points and outliers.

After generating data, each set of data points with feature scaling were tested with using both the BEC method and m-SNN [3] outlier detection method. The m-SNN method is a modification of the SNN (Shared Nearest Neighbor) for use in outlier detection.

In this analysis we generated data set with three different sizes which are small, medium and large. An example for small data set is a set with 56 total data points where 6 of them were generated as global outliers. After applying our new BEC method and m-SNN method with τ 0.05, all the expected global outliers were detected for the BEC method and the m-SNN approach was able to detect all above labeled outliers correctly.

The results obtained are summarized in Table II to demonstrate True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) values as percentages. It shows the average results for three different sizes of data set. From the results, it is clear that BEC has very high TP and TN percentages and very low FP, FN percentages compared to the m-SNN approach. Also the proposed method has the best results for the network type data. In comparing

the results of complex path data sets it is evident that BEC is more robust in finding outliers (compared to m-SNN) particularly with respect to true positives and minimizing false negatives.

Procedure: Betweenness centrality Based Outlier Detection

Inputs: data[], a set of network data points;

Output: List of Outliers

// Finding Adjecency matrix for all the data points

Inputs: data[], Adjacency matrix for data points;

Output: List of Betweenness centrality for all data points

// Finding Betweenness centrality for all the data points

Inputs: data[], Betweenness Centrality for data points;

Output: List of Outliers

//Finding the outliers based on p-value method

Table I.5: Betweenness centrality Based Outlier Detection Algorithm

	TP(%)	**FP(%)**	**TN(%)**	**FN(%)**
BEC	100.0	0.5	99.2	0.2
m-SNN	100.0	3.5	96.5	2.3

Table I.6: Experimental results for BEC and m-SNN.

Discussion of Results

We have described an algorithm and the graph theory capable of detecting outliers in different types of network type data sets. This method is a combination of adjacency matrix and BEC which avoids assumptions about data distributions and uses hypothesis testing to detect outliers. Through a series of experiments, we have shown this method achieve good results with very high true positive and true negative values with the BEC approach producing outlier detection

results equivalent or better than m-SNN methods. Furthermore this method can be used to identify outlier to update social network graph. Currently we are reformulating the algorithm to improve the run time efficiencies and also to parallelize the code to make it amenable for massively large data set

References

[1] Rogers J.P. "Detection of Outliers in Spatial-temporal Data", PhD Thesis.

[2] Hawkins, D.: "Identification of Outliers", Chapman and Hall, London, 1980.

[3] Johnson T., Kwok I., Ng R.: "Fast Computation of 2- Dimensional Depth Contours", Proc. 4th Int. Conf. on Knowledge Discovery and Data Mining, New York, NY, AAAI Press, 1998, pp. 224-228.

[4] Knorr E. M., Ng R. T.: "Algorithms for Mining Distance- Based Outliers in Large Datasets", Proc. 24th Int. Conf. on Very Large Data Bases, New York, NY, 1998, pp. 293-298.

[5] Knorr E. M., Ng R. T.: "Finding Intensional Knowledge of Distance-based Outliers", Proc. 25th Int. Conf. on Very Large Data Bases, Edinburgh, Scotland, 1999, pp. 211-222.

[6] Levent Ertöz, Michael Steinbach, Vipin Kumar.: "A New Shared Nearest Neighbor Clustering Algorithm and its Applications".

[7] Wang W., Yang J., Muntz R.: "STING: A Statistical Information Grid Approach to Spatial Data Mining", Proc. 23th Int. Conf. on Very Large Data Bases, Athens, Greece, Morgan Kaufmann Publishers, San Francisco, CA, 1997, pp. 186-195.

[8] Zhang T., Ramakrishnan R., Linvy M.: "BIRCH: An Efficient Data Clustering Method for Very Large Databases", Proc. ACM SIGMOD Int. Conf. on Management of Data, ACM Press, New York, 1996, pp.103-114.

[9] Angiulli, F. and Pizzuti, C. (2005) Outlier mining in large high-dimensional data sets. IEEE Transactions on Knowledge and Data Engineering, 17(2): 203-215.

[10] Gammerman, A., and Vovk, V. (2002) Prediction algorithms and confidence measures based on algorithmic randomness theory. Theoretical Computer Science. 287: 209-217.

[11] UCI Machine Learning Repository. http://www.ics.uci.edu/ mlearn/MLRepository.html

[12] Vapnik. V. (1998) Statistical Learning Theory, New York: Wiley.

[13] Breunig, M., Kriegel, H., Ng, R., Sander, J. (2000) LOF: Identifying Density-Based Local Outliers. Proc. of the ACM SIGMOD Conference on Management of Data, 427-438.

[14] Proedru, K., Nouretdinov, I., Vovk, V., Gammerman, A. (2002) Transductive confidence machine for pattern recognition. Proc. 13th European conference on Machine Learning. 2430:381-390.

[15] Barbara D., Domeniconi C., Rogers J.P.:"Detecting Outliers using Transduction and Statistical Testing", KDD'06, Philadelphia, Pennsylvania, 2006

[16] Velegrakis D., “Outlier Detection over Data Streams using Statistical Modeling and Density Neighborhoods”, Masters Thesis

[17] Eleazar Eskin. Anomaly detection over noisy data using learned probability distributions. Pages 255-262. Morgan Kaufmann, 2000.

[18] Eleazar Eskin, Wenke Lee, and Salvatore J. Stolfo. Modeling system calls for intrusion detection with dynamic window sizes. In In proceedings of DARPA information Survivability Conference and Exposition 2 (DISCEX, 2001).

[19] Martin Ester, Hans-Peter Kriegel, Jorg Sander, and Xiaowei Xu. A density based algorithm for discovering clusters in large spatial databases with noise. In Proc. Of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), pages 226-231, 1996.

[20] Levent Ertoz, Michael Steinbach, and Vipin Kumar. Finding topics in collections of documents: A shared nearest neighbor approach. In Workshop on Text Mining, held in conjunction with the First SIAM International Conference on Data Mining (SDM 2001). Society for Industrial and Applied Mathematics, 2003.

[21] Studipto Guha, Rajeev Rastogi, and Kyuseok shim. Rock: A robust clustering algorithm for categorical attributes. Inf. Syst., 25(5): 345-366, 2000.

[22] Daniel Barbara, Ningning Wu, and Sushil Jajodia. Detecting novel network intrusions using bayes estimators. In Proceedings of the First SIAM Conference on Data Mining, April 2001.

[23] Re Bronstein, Joydip Das, Marsha Duro, Rich Friedrich, Gary Kleyner, Martin Muller, Sharad Singhal and Ira Cohen. Self-aware services: Using Bayesian networks for detecting anomalies in internet-based services. In Northwestern University and Stanford University Gary Igor, Pages 623-638, 2001.

[24] Edwin M. Knorr, Raymond T. Ng, and Vladimir Tucakov. Distance-based outliers: algorithms and applications. The VLDB Journal, 8(3-4): 237-253, 2000.

[25] Markos Markou and Sameer Singh, Novelty detection: A review – part 1: Statistical approaches. Signal Processing, 83: 2003, 2003.

[26] Matthew Eric Otey, Amol Ghoting, and Srinivasan Parthasarathy. Fast distributed outlier detection in mixed-attribute data sets. Data Min. Knoewl. Discov., 12(2-3): 203-228, 2006.

[27] Gunnar Ratsch, Sebastian Mika, Bernhard Schkopf, and Klaus-Robert Muller. Constructing boosting algorithms for svms: an application to one-class classification, 2002.

[28] Volker Roth. Kernal fisher discriminants for outlier detection. Neural Computing, 18(4): 942-960, 2006.

[29] Bernhard Schlkopf, John C. Platt, John C. Shawe-Taylor, Alex J Smola, and Robert C. Williamson. Estimating the support of a high-dimensional distribution. Neural Computation, 13(7): 1443-1471, 2001.

[30] Graham Williams, Rohan Baxter, Hongxing He, Simaon Hawkins, and Lifang Gu. Comparative study of rnn for outlier detection in data mining. In ICDM, page 709, 2002.

[31] Abdallah Abbey Sebyala, Temitope Olukemi, and Dr. Lionel Sacks. Active platform security through intrusion detection using naïve Bayesian network for anomaly detection. In London Communications Symposium, 2002.

[32] Ji Zhang and Hai Wang. Detecting outlying subspaces for high-dimensional data: the new task, algorithms and performance. Knowl. Inf. Syst., 10 (3): 333-355, 2006.

[33] Alisdair McDiarmid, Stephen Bell, James Irvine, Jamie Banford: Nodobo: "Detailed Mobile Phone Usage Dataset".

[34] National Data Buoy Center, Data Availability Summary for NDBC Platforms, http://www.ndbc.noaa.gov/data_availability/data_avail.php

[35] K. P. Liyanage, R. George and K. Shujaee, "Outlier Detection in Spatial Data using the m-SNN Algorithm", IEEE southeastCon 2013.

[36] A. Bader, S. Kintali, K. Madduri, and M. Mihail, "Approximating betweenness centrality", In Proceedings of the 5th international conference on Algorithms and models for the web-graph, WAW'07, pages 124–137, Berlin, Heidelberg, 2007. Springer-Verlag.

[37] U. Brandes, "A faster algorithm for betweenness centrality". Journal of Mathematical Sociology, 25(1994):163–177, 2001.

[38] U. Brandes, "On variants of shortest-path betweenness centrality and their generic computation", Social Networks, 30(2):136–145, 2008.

[39] U. Brandes and C. Pich, "Centrality estimation in large networks" International Journal Of Bifurcation And Chaos, 17(7):2303, 2007.

[40] L. C. Freeman, "A set of measures of centrality based on betweenness", Sociometry, 40(1):35–41, 1977.

[41] R. Geisberger, P. Sanders, and D. Schultes, "Better approximation of betweenness centrality", In J. I. Munro and D. Wagner, editors, ALENEX, pages 90–100. SIAM, 2008

[42] . M. E. J. Newman. "A measure of betweenness centrality based on random walks", Social Networks, 27(1):39–54, 2005.

[43] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks", Physical Review E, 69(2): 26113, 2004.

Section II: Visualization and Analysis of Large Scale Network Structures

Section IIA: Visualizing Outliers in Multi-dimensional Data

Visualization of Outliers in Network Data

An outlier is an observation that deviates from other observations as to arouse suspicion that it was generated by a different mechanism [2]. In other words, for a given sample, a point is called an outlier if its behavior significantly deviates from the rest of the members of the sample. Outlier detection plays a vital role in several fields such as credit card fraud detection, severe weather prediction, image processing, pattern recognition and computer intrusions. In this reporting year we continued work on outlier detection particularly in integrating the visualization of data with detection techniques in large data.

Centrality of a Node determines the importance of an entity with respect to other entities. Measures affecting centrality include eccentricity, radiality, and degree and centroid measures. Based on the centrality measures we are able to plot the sample data. In the following analysis, we use communication data as the analogy in visualizing the relationships between entities.

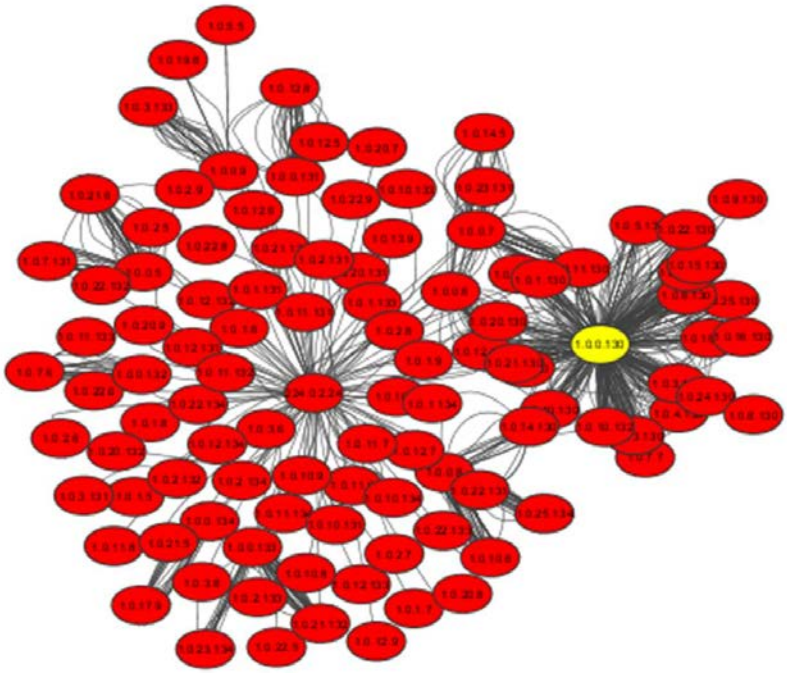


Figure I1-1: Plot of Node Centrality

Additional measures may be computed to provide a statistical analysis of each node

(Figure II-2).

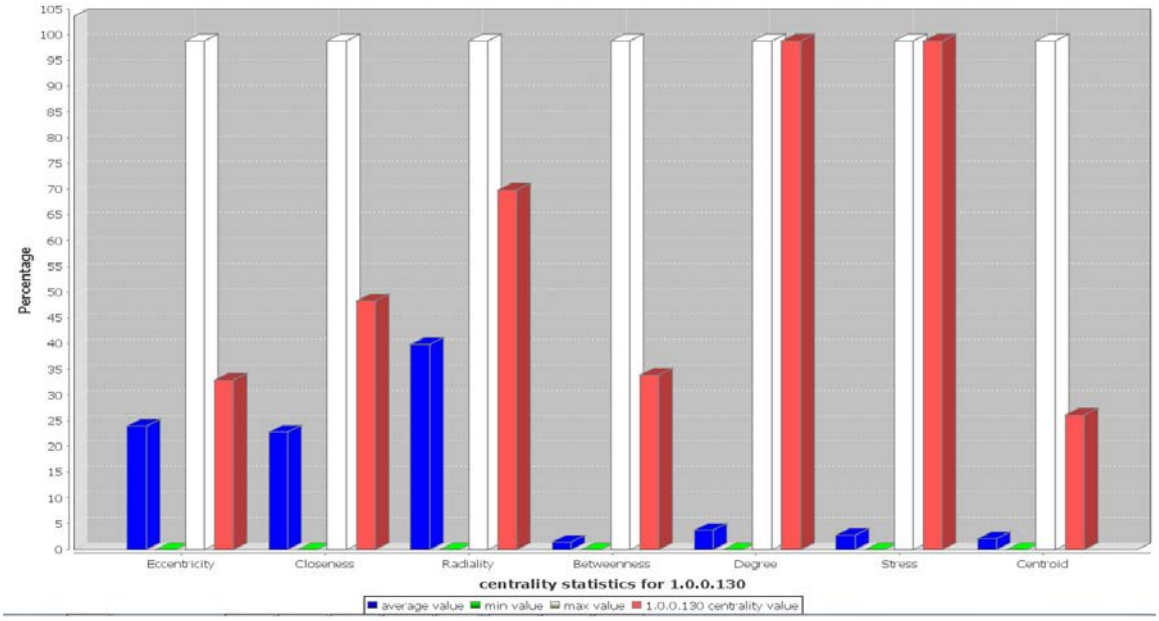


Figure I1-2: Centrality Statistics for entity 10.0.0.130

From these measures it is clearly seen the entity colored in yellow in Figure 1 occupies a central position in the sample data. Figures II-(3-6) show the centrality of alternate nodes, their corresponding statistical measures.

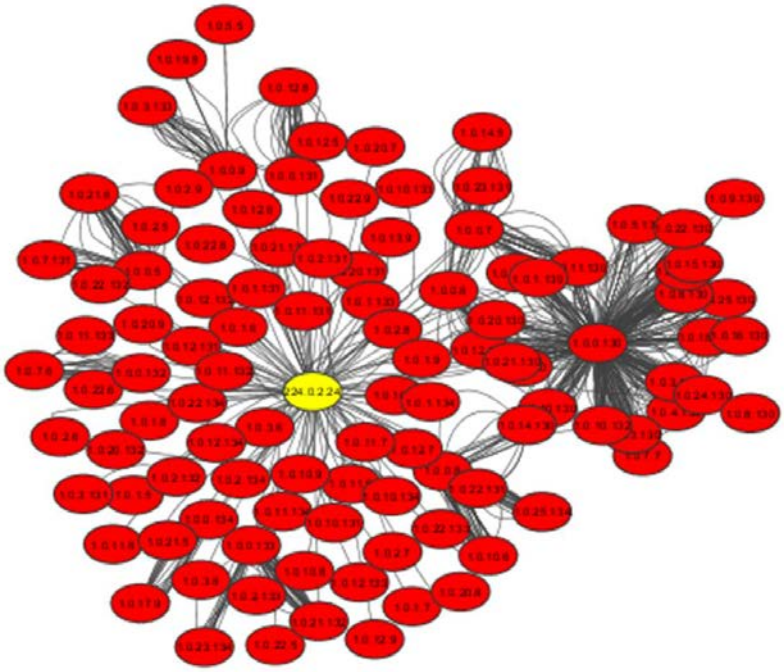


Figure I1-3: Node Centrality of Entity 224.0.2.24

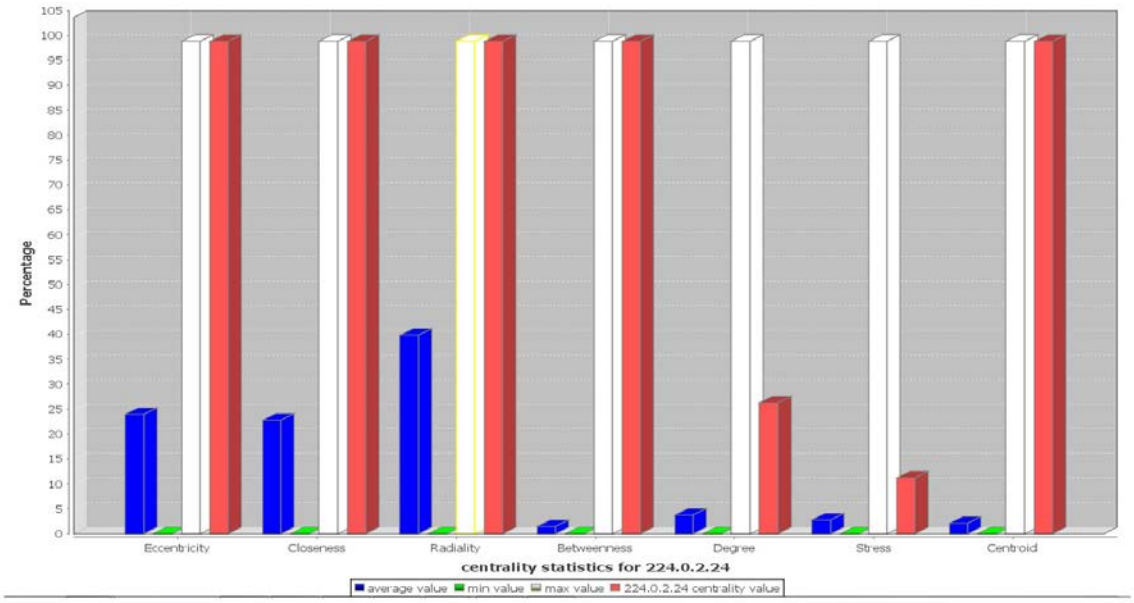


Figure II-4: Centrality Statistics for Entity 224.0.2.24

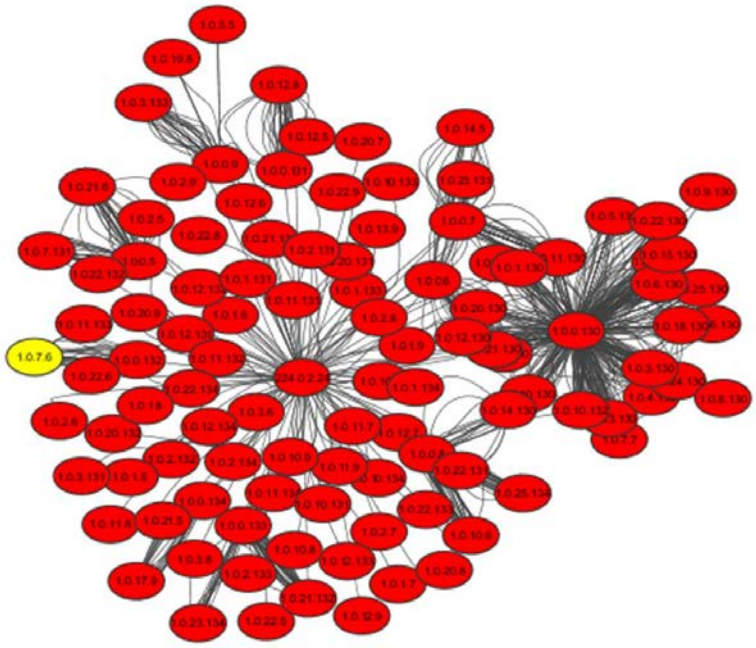


Figure II-5: Node Centrality of Entity 1.0.7.6

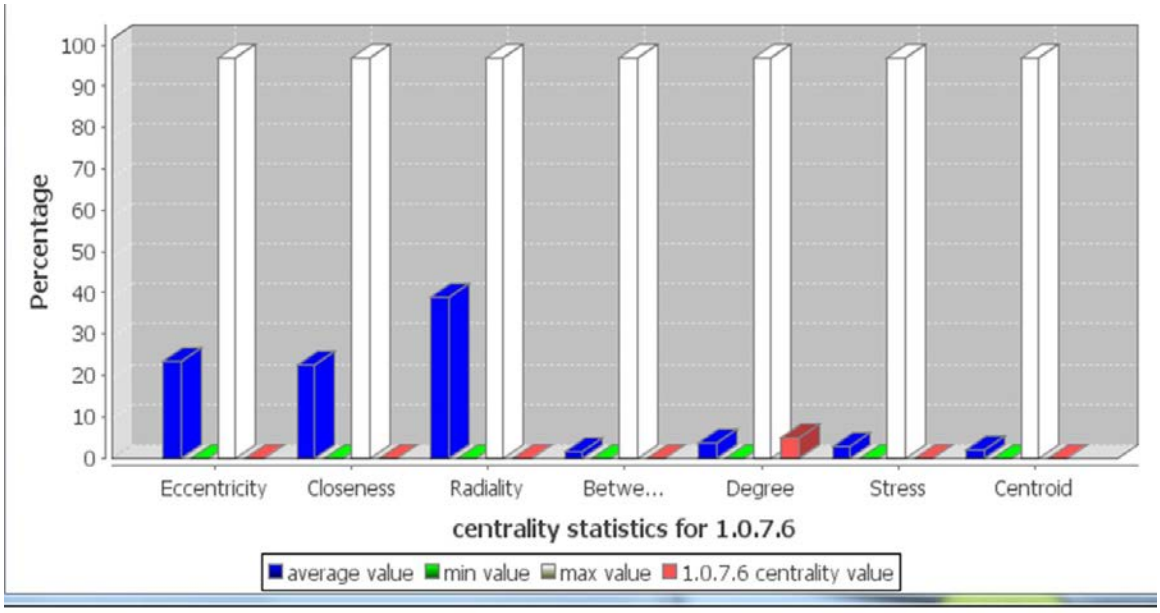


Figure II-6: Centrality Statistics for Entity 1.0.7.6

Using this as an initial study, the team developed the aggregate visualization of the structure seen in Figure II-7. A circular layout that emphasizes group and tree structures within a network is

adopted for this visualization. It partitions the network by analyzing its connectivity structure, and arranges the partitions as separate circles. The circles themselves are arranged in a radial tree layout fashion.

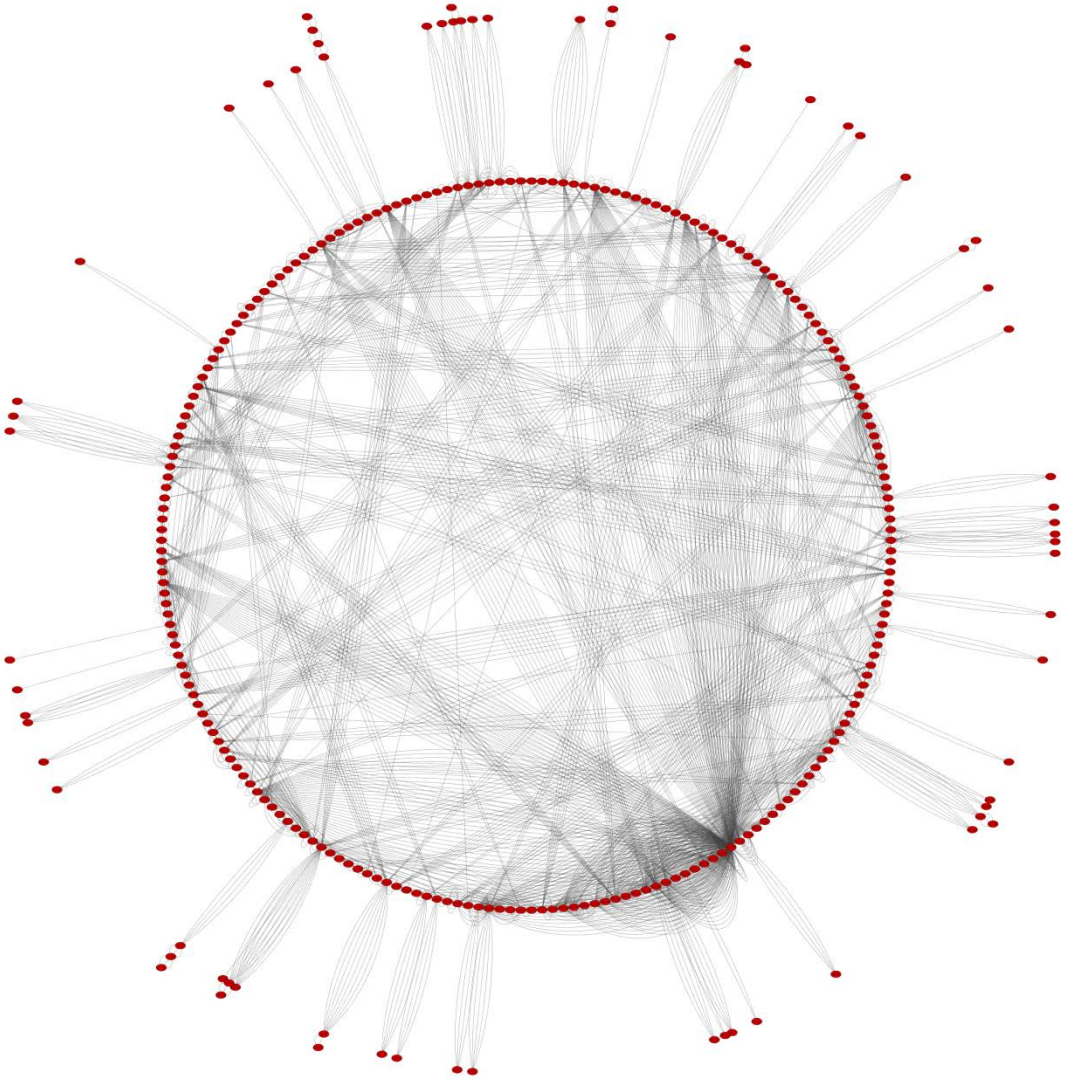


Figure II-7: Aggregate Visualization of the Network Elements

We have developed two software components for data visualization of the entity structure: a. statistical visualization module, and b. interactive visualization module for the network.

Statistical Visualization Module

This module is aimed for statistical analysis and charting/plotting of the network packets using different types of charts and filtering tools. The following figure is a snapshot of this module in action showing different aspect of the captured network data.

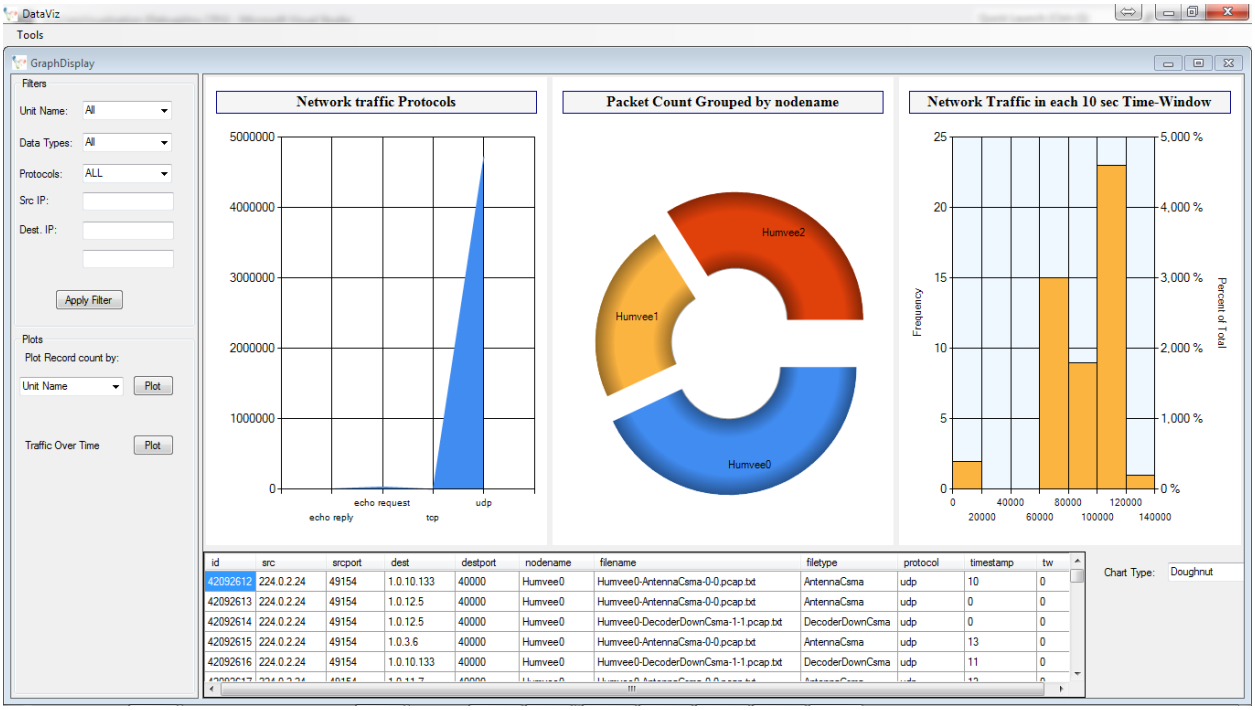


Figure II-8: Statistical Visualization of Data

In addition, this module has the capability of reading archive data in parallel depending on number of CPUs available and insert them into database.

Interactive Visualization Module

An interactive visualization module has been built using the GraphStream library. In this module the library is used to illustrate and represent relationships between the entities. The main components which are utilized in this module to reflect graph based analysis are grouped as follow:

- Graph Import:

- This module uses a custom parser which reads a flat text file consists of a set of source, destination and weight in each line forming an edge of graph. The imported files are produced by extracting network packets/segments from pcap files and transformed them into graph definition format.

- Computational functions:
  - Community detection based on length of edges with respect to the product of average length of all edges and a cut-off threshold set manually.
  - Measurement of network modularity to determined quality of communities by comparing the ratio of internal links within each community to the number of edges in each community with the expected values in the same network. Also all the links of the same network are randomly rewired. This yields the modularity Q as follow:

$$Q = \sum_c \frac{I_c}{m} - \sum_c \frac{(2I_c + O_c)^2}{(2m)^2}$$

I_c: Internal link in each community C
m: number of edges
O_c: Outgoing links from community C

- Graph Layout:

  This module enables the user to layout graphs using two different algorithms. Majority of the layout algorithms are force based using a repulsive such as Spring-Box and Lin-Log layout. They use the attraction (*a*) and repulsion(*r*) force factors of energy model. In Dynamic graph layout each time there is a change in the graph, the layout will compute its new state from its previous energy equilibrium. The two layout algorithms that are supported currently are:

  - The Spring-Box is the default layout algorithm used in the GraphStream. This algorithm is based on the Frutcherman-Reingold force layout with modified attraction force.
  - Lin-Log layout groups densely connected nodes and places them nearby each other, and puts the weakly connected nodes at distance location.

There is configuration file in which users can set location of the import file, filtering the source or destination node, switching layout and etc.

By using this module the end user is able to visualize a dynamic graph, which in this case, representing the extracted entity relations.

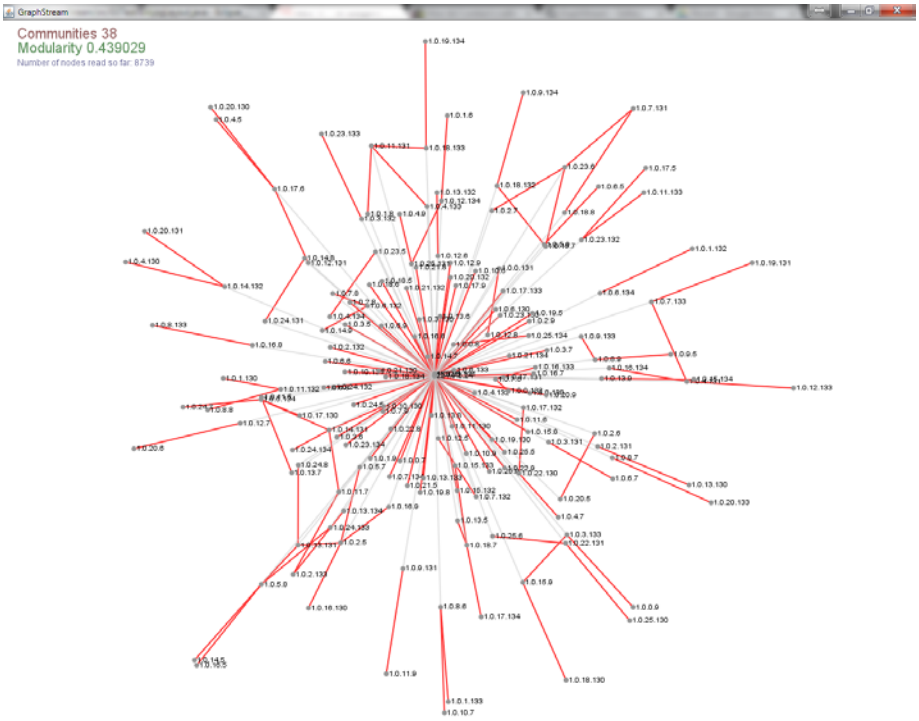


Figure II-9: Dynamic Graph Generation (Intermediate)

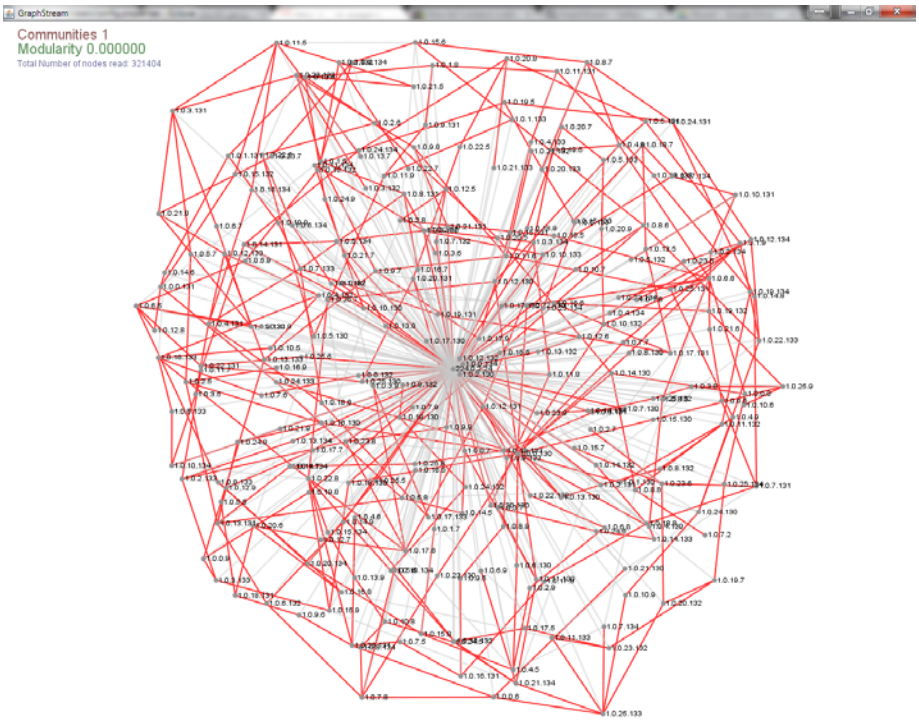


Figure II-10: Dynamic Graph Generation (Final)

The utility functions for data extraction and visualization developed during 2012-2013, have revealed new insights into the data. First, centrality statistics can provide a measure of the relative importance of a node at any given time. The centrality statistics may be used as a measure to determine outlier behavior in a node or across the network. A possible approach would be to utilize the m-SNN outlier predication approach [35]. An extension to this approach would be to apply these algorithms in conjunction with geo-location as a predictor for network congestion. We will also investigate a graph-based data mining using a graph representation of the data to extract knowledge. Graph based approaches are useful in extracting knowledge that might be deeply embedded in the data in the form of substructures. Network data may be naturally represented as graphs and therefore graph-based data mining approaches are appropriate investigative mechanisms. Graph based data mining have seen use in in several areas for the discovery of substructures and paths and show great promise in the extraction of useful knowledge from related data. The outlier detection approach may be used to understand data that is anomalous, and would provide further understanding of the large scale data.

Section IIB: Visualization of Outliers in Network Data

Background

With the rapid growth of the Web, users can get easily lost in the massive, dynamic and mostly unstructured network topology. Finding users' needs and providing useful information are the

primary goals of website owners. Web structure mining [44],[45],[46] is an approach used to categorize users and pages. It does so by analyzing the users' patterns of behavior, the content of the pages, and the order of the Uniform Resource Locator (URL) that tend to be accessed. In particular, Web structure mining plays an important role in guiding the users through the maze. The pages and hyperlinks of the World-Wide Web may be viewed as nodes and arcs in a directed graph. The problem is that this graph is massive, with more than a trillion nodes, several billion links, and growing exponentially with time. A classical approach used to characterize the structure of the Web graph through PageRank algorithm, which is the method of finding page importance.

The original PageRank algorithm [46],[47],[48] one of the most widely used structuring algorithms, states that a page has a high rank if the sum of the ranks of its backlinks is high. Google effectively applied the PageRank algorithm, to the Google search engine [47]. Xing and Ghorbani [49] enhanced the basic algorithm through a Weighted PageRank (WPR) algorithm, which assigns a larger rank values to the more important pages rather than dividing the rank value of a page evenly among its outgoing linked pages. Each outgoing link page gets a value proportional to its popularity (its number of in-links and out-links). Kleinberg [50] identifies two different forms of Web pages called hubs and authorities, which lead to the definition of an iterative algorithm called Hyperlink Induced Topic Search (HITS) [518].

Bidoki and Yazdani [52] proposed a novel recursive method based on reinforcement learning [53] that considers distance between pages as punishment, called "DistanceRank" to compute ranks of web pages in which the algorithm is less sensitive to the "rich-get-richer" problem [52],[54] and finds important pages faster than others. The DirichletRank algorithm has been proposed by X. Wang et al [56] to eliminate the zero-one gap problem found in the PageRank algorithm proposed by Brin and Page [47]. The zero-one gap problem occurs due to the ad hoc way of computing transition probabilities. They have also proved that this algorithm is more robust against several common link spams and is more stable under link perturbations. Singh and Kumar [57] provide a review and comparison of important PageRank based algorithms.

As search engines are used to find the way around the Web, there is an opportunity to fool search engines into leading people to particular page. This is the problem of web spamming [57], which is a method to maliciously induce bias to search engines so that certain target pages will be ranked much higher than they deserve. This leads to poor quality of search results and in turn reduces the trust in the search engine. Consequently, anti-spamming is a big challenge for all the search engines. Earlier Web spamming was done by adding a variety of query keywords on page contents regardless of their relevance. In link spamming [58], the spammers intentionally set up link structures, involving a lot of interconnected pages to boost the PageRank scores of a small number of target pages. This link spamming does not only increasing the rank gains, but also makes it harder to detect by the search engines. It is important to point out that link spamming is a special case of the spider-traps [59]. At the present time, the Taxation method [59] is the most

significant way to diminish the influence of the spider-traps and dead-ends by teleporting the random surfer to a random page in each iteration.

This research has two main contributions: First, we present a generalized formulation of the PageRank algorithm based on transition probabilities, which takes both in-link and out-links of node and their influence rates into account in order to calculate PageRank. This would permit the application of this approach to a wide variety of network problems that require consideration of the current state values (and PageRank) as a function of past state transitions. Second, we describe a novel approach of adding virtual edges to a graph that permits more realistic computations of PageRank, negating the effect of network anomalies such as spider-traps and dead-ends.

A Generalized Method

In web arena, a link by important pages will impact on significance of a page. However, there are other networks in which not just in-link but out-links are also weighty. For instance, in social networks, connecting to eminent people (out-link) is as crucial as being connected by key persons (in-link) in evaluating the degree of prominence of a member. Therefore, sometimes sorting and grading nodes of a graph only based on in-links will result in an incorrect evaluation. So, we take out-links and the rate of their impacts with respect to in-links into our computations.

Algorithm

Suppose we start as a random surfer at any of the *n* pages of the Web with equal probability. Then the initial vector will have *1/n* for each component. If M_f is the forward transition matrix of the Web, then after one forward step, the probability distribution of the next surfer place will be $M_f v_0$ and if M_b is the backward transition matrix of the Web, then after one backward step, the probability distribution of the previous surfer place will became $M_b v_0$. Also, we consider the importance weight factor of both in-links (β) and out-links ($1-\beta$).

Note that equation $\left(\beta M_f + (1-\beta) M_b\right)$ is the linear combination of both next and previous surfer place, and it is also stochastic because it is a linear combination of two stochastic matrices. So its eigenvalue associated with the principal eigenvector will be 1. The principal eigenvector of $\left(\beta M_f + (1-\beta) M_b\right)$ tells us where the surfer is most likely to be after a long time. Recall that the intuition behind PageRank is that the more likely a surfer is to be at a page, the more important the page is. We can compute the principal eigenvector of $\left(\beta M_f + (1-\beta) M_b\right)$ by starting with the initial vector v_0 and multiplying by $\left(\beta M_f + (1-\beta) M_b\right)$ some number of times, until the vector we get shows little change at each round. Considering this matrix instead of M_f has two

advantages: First, in computing PageRank of a node, the importance of its neighbors with both types of relationship (out-link and in-link) and their arbitrary impact rates (parameter β) have taken into account. Second, by using this method, we do not have the problems about dead-ends and spider-traps because we take the linear combination of entering probability from and exiting probability to other nodes in our computation. Therefore, in case $\beta \neq 0$ and $\beta \neq 1$, the columns related to dead-ends are not completely zero. Likewise, for the spider-trap columns, probabilities related to other nodes are not zero and they cannot absorb more unreasonable rank to themselves. About cases $\beta = 1$ or $\beta = 0$, in the following, we proposed another idea (adding virtual edges) by which the random surfer can exit from dead-ends and spider-traps.

The algorithm is as follows:

Step 1: Finding Forward and Backward transition matrices.

Step 2: considering appropriate formula and keep iterating until it gets converged.

In this step, three possible conditions can exist which are characterized as following:

- **Case 1:** $\beta \neq 0$ **and** $\beta \neq 1$**.** It means that both forward and backward trends are important to calculate PageRanks. Thus, we only need to calculate the eigenvector of matrix $\left(\beta M_f + (1-\beta) M_b\right)$.
- **Case 2:** $\beta = 1$ So, we need only the forward matrix to calculate PageRanks. If there are not a dead-end or a spider-trap in the graph, the vector of PageRanks is the eigenvector of M_f. If there are dead-ends or spider-traps, the eigenvector of M_f assigns most of PageRank to spider-traps and dead-ends that is not real. Thus we add enough virtual out-links to remove these spider and dead-end situations. For each dead-end and spider-trap, we will consider a virtual edge in which source of them are dead-ends and one member of each spider-traps, respectively. Also, their destinations can be any arbitrary nodes, excepting those of dead-end and spider-traps (see Figure 3. Green color edges). Hence, If assumed *v* is eigenvector of matrix $M_f^{'}$ (forward transition matrix after adding virtual links), in order to find final PageRanks of vertices, we have to remove effect of these virtual links on PageRanks by calculating the following equation $v - (M_f^{'} - M_f)v$.
- **Case 3:** $\beta = 0$. Here only backward trend (out-links) is important to consider for calculation of PageRanks. So we only need backward matrix to determine PageRanks. If there are not in-component or in-tendril vertices in the graph, vector of PageRanks is eigenvector of M_b. If there are in-component or in-tendril vertices, eigenvector of M_b assigns most of PageRank to in-component and in-tendril vertices, which is not real. Thus we add enough virtual in-links to remove these in-component and in-tendril situations then after computing eigenvector of new backward matrix $M_b^{'}$, we have to remove effect of these virtual links on PageRanks (see Figure 3. Red color edges). If suppose *v* is eigenvector of matrix $M_b^{'}$

(backward transition matrix after adding virtual links). The final PageRanks of vertices would be $v - (M_b^{'} - M_b)v$.

Step 3: normalize PageRank vector to find distribution probability of vertices.

As shown below, if we consider a matrix include the importance of pairwise comparison of vertices (A), eigenvector of this matrix would be distribution probability of vertices.

Note that, W is vector distribution probability of vertices that sum of its components is 1 and also w_i is amount of vertex i's importance. So, instead of w_i / w_j in matrix A, we let p_i / p_j , which p_i , p_j are PageRanks of nodes i, j. We calculate eigenvector of matrix A and to get the distribution probability of vertices.

$$AW = \begin{bmatrix} w_1/w_1 & w_1/w_2 & \cdots & w_1/w_n \\ w_2/w_1 & w_2/w_2 & \cdots & w_2/w_n \\ \vdots & \vdots & \cdots & \vdots \\ w_n/w_1 & w_n/w_2 & \cdots & w_n/w_n \end{bmatrix} * \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} = n * \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} = nW$$

Biased Random Walk

In order to bias the rank of all nodes with respect to a special subset of nodes, we use the Biased Random Walk method in which the random surfer, in each iteration, will jump on one of the member of the subset with equal probability. Its most important application is topic-sensitive PageRank [19] in search engines. The consequence of this approach is that random surfers are likely to be at an identified page, or a page reachable along a short path from one of these known pages, because the pages they link to are also likely to be about the same topic. The mathematical formulation for the iteration that yields topic-sensitive PageRank is similar to the equation we used for general PageRank. The only difference is how we add the new surfers. Suppose S is a set of integers consisting of the row/column numbers for the pages we have identified as belonging to a certain topic (called the teleport set). Let e_s be a vector that has 1 in the components in *S* and 0 in other components. Then the topic-sensitive PageRank for *S* is the limit of the iteration

$$v^{'} = \alpha(\beta M_f + (1-\beta)M_b)v + (1-\alpha)e_s / |s| \quad 0.8 \le \alpha \le o.9$$

Here, as usual, *M* is the transition matrix of the Web, and |*S*| is the size of set *S*.

Experimental Results

Figure II-11. is a graph with 20 vertices that include all kinds of network artifacts mentioned previously.

SCC:{1,2,4,5,7,8,9,10,15,17,18,20} tube:{16-6}

Out-component:{6,11,12} In-component:{3,13,16} Out-tendril:{14} In-tendril:{19}

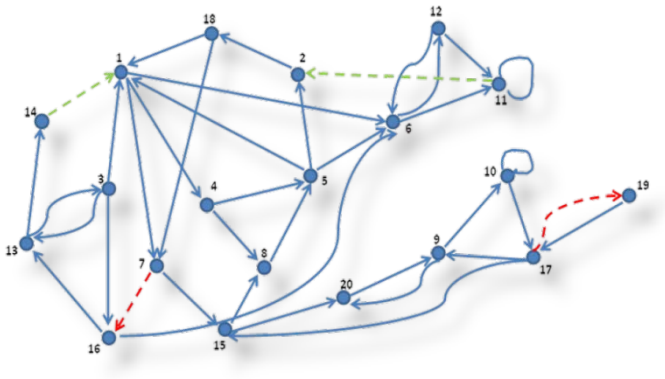


Figure II-11: Synthetic Graph Example

In case 2 ($\beta = 1$), there are a dead-end situation on vertex 14 and a spider-trap situation on set of vertices {6, 11, 12}, and in order to remove the dead-end and the spider-trap consider 2 virtual out-link (green edges) on these vertices. Also in case 3 ($\beta = 0$), there are in-component situation on set of vertices {3, 13, 16}, and in order to remove negative PageRank consider 2 virtual in-link (red edges) on these vertices. For completeness, we also compute the biased random walk on case1. Comparing the results with case1, TABLE I. , it is clear that PageRanks are biased on set S={2, 4, 7, 18}. As we expect, rank of nodes of set *S* and nodes that are pointed by set *S* get higher ranks.

Table II.1 : PageRank vector AT cases 1, 3, and biased random walk.

Results of case 1 ($\beta = 0.7$)		**Results of the biased random walk on case1**		**Results of case 3 ($\beta = 0$)**	
Nodes number	***PageRank***	***Nodes number***	***PageRank***	***Nodes number***	***PageRank***
11	0.945	5	0.9937	17	0.57916
12	0.2177	11	0.9878	10	0.38611
6	0.1767	18	0.9703	13	0.36037
9	0.0703	1	0.9432	1	0.27028
10	0.0632	7	0.9013	3	0.27028
5	0.0601	15	0.8513	5	0.25741
1	0.0543	2	0.7444	9	0.25741
20	0.0527	4	0.6847	7	0.24454
15	0.0495	6	0.65	4	0.19305
17	0.045	8	0.6414	19	0.19305
8	0.036	9	0.5045	16	0.18018
7	0.029	20	0.4878	2	0.16731
4	0.0272	12	0.3659	18	0.16731
18	0.025	10	0.3204	8	0.1287
3	0.0237	17	0.2976	15	0.1287
13	0.023	3	0.1628	20	0.1287
16	0.0223	13	0.1144	12	1.14E-17
2	0.0216	16	0.0923	6	7.34E-18
14	0.0081	19	0.0386	11	0
19	0.0068	14	0.035	14	0

Table 11.2 : Comparing results of the Algorithm and taxation method to avoid anomalies in Case 2

($\beta = 1$)

Using virtual edges		**Taxation**	
nodes no	***PageRank***	***nodes no***	***PageRank***
9	0.508068237	11	0.83086
10	0.508068237	9	0.25352
20	0.381051178	10	0.22903
2	0.265581124	20	0.19944
17	0.254034118	15	0.15968
15	0.254034118	6	0.1495
5	0.173205081	5	0.14569
18	0.161658075	17	0.14155
8	0.15011107	8	0.11547
1	0.138564065	1	0.11197
6	0.138564065	7	0.08907
7	0.127017059	12	0.08748
11	0.103923048	18	0.07921
12	0.069282032	2	0.06521
4	0.046188022	4	0.05567
3	7.50E-17	13	0.0528
13	2.12E-17	3	0.04612
16	1.16E-17	14	0.04612
14	1.02E-17	16	0.0369
19	0	19	0.02386

Comparing the results of the Taxation method and our method, Table II.A, obviously we can realize that our approach produces more reasonable outcomes. Because, as it is shown in the TABLE II.1, node 9 is the junction of two cycles; all nodes of these cycles are from SCC part of the graph, so the random surfer is most likely on it. The nodes 10 and 20 have higher rank after 9, because they have in-link from the node 9. The rank of node 5 cannot be higher than 17 because the node 17 is a member of the cycle consist of node 9 and 10. In Taxation result, the nodes with spider-trap situation such as 6 and 11 got higher and vertices 2 and 18 got lower PageRank than our approach results. Also, for other vertices, their ranks are either the same or very close to each other's.

Discussion of Results

In this research, the fundamental idea of Network Structure mining is explained in detail to have a generic understanding of the data structure used in web. The main purpose of this research is to present the new PageRank based network analysis algorithms and compare that with the previous algorithms.

The method generalizes the approach of finding PageRank based on transition probabilities by considering the arbitrary impact rates of both out-links and in-links, in order to include all possible cases because there are some conditions in which out-links have also an influence on PageRank of nodes. Moreover, it prevents that spider-traps and dead-ends have a high unreasonable rank and assign higher PageRanks to themselves. The noticeable weak point of previous method is that it assigns more unreasonable PageRank to spider-traps and dead-ends, and also reduces PageRank of SCC vertices. But in our approach this problem has been solved, because by adding virtual edges, random surfers will not stop on spider-traps and dead-ends. According to [57], DirichletRank has been so far the best method amongst previous methods, capable of diminishing the impact of link spamming (a special case of spider-traps) and dead-end problem that is, however, only applicable to backward analysis. Our approach in comparison with their method is general for more types of networks and simpler to understand and implement. Also, by using ideas suggested in this research, in any possible cases, PageRanks is insulated from the influence of anomalies including in/out-tendrils and in/out-components.

The generalization of the PageRank algorithm to include forward and backward links into a node makes this approach applicable to new domains beyond web mining and search engines. We are currently exploring the application of the new generalized algorithm to the analysis of network data for instance using PageRank as a measurement of node's activity score to find communities. The work has applications to biological systems which mimic networks, and this is a current area of analysis.

References:

[44] R. Kosala and H. Blockeel, “Web mining research: A survey,” ACM SIGKDD Explorations, 2(1), 2000, pp. 1–15.

[45] S. Madria, S. S. Bhowmick, W. K. Ng, and E.-P. Lim, “Research issues in web data mining,” In Proceedings of the Conference on Data Warehousing and Knowledge Discovery, 1999, pp. 303–319.

[46] S. Pal, V. Talwar, and P. Mitra, “Web mining in soft computing framework : Relevance, state of the art and future directions,” IEEE Trans. Neural Networks, 13(5), 2002, pp. 1163–1177.

[47] L. Page, S. Brin, R. Motwani, and T. Winograd, “The PageRank citation ranking: Bringing order to the web,” Technical report,Stanford Digital Libraries SIDL-WP-1999-0120, 1999.

[48] C. Ridings and M. Shishigin, “Pagerank uncovered,” Technical report, 2002.

[49] W. Xing and A. Ghorbani, “Weighted PageRank Algorithm,” Proc. of the Second Annual Conference on Communication Networks and Services Research (CNSR ’04) IEEE, 2004 , pp. 305-314, 0-7695-2096-0/04 .

[50] J. Kleinberg, “Authoritative Sources in a Hyper-Linked Environment”, Journal of the ACM 46(5), 1999, pp. 604-632.

[51] S. Chakrabarti, et al. "Mining the Web's link structure." Computer 32.8 ,1999, pp. 60-67.

[52] A. M. Zareh Bidoki and N. Yazdani, “DistanceRank: An intelligent ranking algorithm for web pages,” Information Processing and Management, Vol 44, No. 2, 2008, pp. 877-892.

[53] R.S. Sutton and A.G. Barto, “Reinforcement Learning: An Introduction,” Cambridge, MA: MIT Press, 1998.

[54] J. Cho, S. Roy and R. E. Adams, “Page Quality: In search of an unbiased web ranking,” Proc. of ACM International Conference on Management of Data,”. 2005, pp. 551-562.

[55] X. Wang, T. Tao, J. T. Sun, A. Shakery, and C. Zhai, “DirichletRank: Solving the Zero-One Gap Problem of PageRank,” ACM Transaction on Information Systems, Vol. 26, Issue 2, 2008.

[56] A. K. Singh and P. Ravi Kumar. "A Comparative Study of Page Ranking Algorithms for Information Retrieval," International Journal of Electrical and Computer Engineering 4, no. 7 (2009), pp. 469-480.

[57] Z. Gyongyi and H. Garcia-Molina, “Web Spam Taxonomy,” First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb 2005), 2005.

[58] Z..Gyongyi and H. Garcia-Molina, “Link Spam Alliances,” Proc. of the 31st International Conference on Very Large DataBases (VLDB), 2005, pp. 517-528.

[59] A. Rajaraman, J. Leskovec, and J. D. Ullman, “Mining of Massive Datasets,” 2013, pp.161-198.

[60] S. Brin and L. Page, “Anatomy of a large-scale hypertextual web search engine,” Proc. 7th Intl. World-Wide-Web Conference, 1998, pp. 107–117.

[61] A. Broder, et al. "Graph structure in the web," Computer networks 33.1 , 2000, pp. 309-320.

[62] T.H. Haveliwala, “Topic-sensitive PageRank,” Proc. 11th Intl. World-Wide-Web Conference, 2002, pp. 517–526.

[63] J. Qiu and Z. Lin, “A framework for exploring organizational structure in dynamic social networks,” Decision Support Systems, 51, 2011, pp.760–771.