4-25-2013

Molecular Dynamic Analysis of Electrostatics of Single-Stranded DNA with a Prospective Towards Single Molecule Sequencing

Mohammad Jomah I. Abu Saude

MOLECULAR DYNAMIC ANALYSIS OF ELECTROSTATICS OF SINGLE-STRANDED DNA WITH A PROSPECTIVE TOWARDS SINGLE MOLECULE SEQUENCING

by

Mohammad Jomah I. Abu Saude

A Thesis

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

Major: Electrical Engineering and Computer Engineering

The University of Memphis

May, 2013

Acknowledgements

First and above all, praises and thanks to God, for His helpings and blessings to complete this thesis successfully.

I would like to thank my research supervisor, Dr. B. Morshed, professor in Electrical Engineering and Computer Engineering, University of Memphis, for giving me the opportunity to do research in the biomedical field (i.e. a new topic for me), providing great guidance throughout the work, and helping me to complete my degree.

I am also thankful for all the people in the ESARP lab, for helping me and giving me suggestions throughout my work.

I could have never achieved my thesis without my family. They have always supported me emotionally and financially throughout my degree. Special thanks for my wife and my parents.

ABSTRACT

Abu Saude, Mohammad J. M.S. The University of Memphis. May/2013. Molecular Dynamic Analysis of Electrostatics of Single-Stranded DNA with a Prospective Towards Single Molecule Sequencing. Major Professor: Dr. Bashir Morshed

Single molecule DNA sequencing, commonly referred as the third generation sequencing, requires new approaches to identify nucleotide bases. Based on molecular dynamics (MD) simulations, we investigate a prospective direct electronic sequencing approach of nucleotides using the electrostatics of single-stranded DNA (ssDNA).

To study the intrinsic electrostatic of ssDNA, electrostatic potentials have been calculated by solving the nonlinear Poisson-Boltzmann equation using visual molecular dynamics (VMD) for 25 base pair (bp) of polymers. The results show that the molecular electrostatic potential (MEP) differs for various bases within 3 nm from the center of the sugar backbone, with suitably differentiable variations at 1.4 nm distance. The MEP variations among four nucleotide bases are the most significant near ~33.7° and ~326.3° from the center of the nucleotide base, while the influences of the neighboring bases on MEPs become insignificant after the 3rd-nearest neighbors. With simulation of ssDNA under an applied electric field using NAMD MD simulator, the results suggest that the translocation rate of the ssDNA is dependent on the size of the nucleotides.

These results demonstrate the potential to develop novel single molecule electronic DNA sequencing technology. Ability of proximal probing, vibration of nucleotides, ionic interference and sequencing of polymers are some of the challenges to be resolved.

TABLE OF CONTENTS

Chapter Page

LIST OF FIGURES

Page

I. INTRODUCTION

A. Motivation

Deoxyribonucleic Acid (DNA) carries the genetic information of life. DNA is usually found in a double stranded form (dsDNA) of different nucleotides. Each nucleotide contains phosphate, sugar, and a base (Adenine (A), Thymine (T), Guanine (G), and Cytosine(C)). The DNA complementarity is one of its features where the A hydrogen bonds with T and C with G. Hence, sequencing of one strand of dsDNA is sufficient to know the complementary strand DNA sequence. Decoding this information (a complete order of bases in the human genome) is called genome sequencing, and is very important to scientists in many fields. The process of finding specific sequence of a strand of DNA is called DNA sequencing. The main challenges in DNA sequencing are cost, labor, time, read-length (number of bases that can be read from a single measurement), and base-calling accuracy. Many researchers are trying to resolve these challenges.

DNA sequencing started at the end of the 1970s with Sangers' chain-termination sequencing method. This method acted as a main sequencing method for more than 3 decades and was used in the first automated fluorescent project for sequencing of a genome. When the "Human Genome Project" launched in 1990 to sequence a whole genome, new sequencing technologies, known as next-generation sequencing, were produced with more throughput. One of these next-generation technologies is the massively parallel sequencing platforms by replicating the DNA and use sequencing methods in parallel to have more throughputs. Another technology produced is called single -molecule sequencing. The general goal of this technology is to sequence the DNA in real-time resulting technology development like SMRT, and nanopore sequencing. In nanopore sequencing, DNA passes through a fabricated or enzymatic nanopore and a reading is decoded by monitoring the blockage current produced by the nucleotides. Another nanopore technology studied by researchers is measuring the electrostatic potential that induced on nano-capacitor when the DNA passed through it and it has been shown through simulations that the DNA bases can be counted with this approach.

B. Objective

Our goal is to study the electrostatic characteristics around each type of the nucleotide to investigate if there is a significant difference among the bases such that they can be identified with this intrinsic electrostatics, which might lead to a prospective novel technology for single-molecule, based DNA sequencing.

C. Method

Investigating the electrostatic potential induced by DNA itself while translocating through a nanopore is largely uncharted territory. This work presents this unique prospective and investigates the feasibility of single molecule DNA sequencing based solely on the intrinsic electrostatics of DNA. The electrostatics around a molecule can be calculated using Poisson-Boltzmann equations. In this work we used Molecular Dynamic Simulation (NAMD) to simulate the molecule systems and a Poisson-Boltzmann equations solver (APBS) to study the electrostatic potential around the single stranded DNA (ssDNA). MD simulation can provide a virtual model of a biomolecular system that allows us to observe the properties of molecules over time. Using simulations, we build ssDNA polymers of four different nucleotide bases, and study its electrostatic characteristics.

D. Outcome

- The intrinsic electrostatics are unique to each nucleotide base until about 3 nm from the center of sugar backbone and suitably distinct at a distance of ~1.4 nm.
- The Molecular Electrostatic Potential variations among four nucleotide bases are the most significant near ~34° and ~326° from the center of the nucleotide base.
- The influences of the neighboring bases on MEPs become insignificant after the 3rd-nearest neighbors.
- The movement of the ssDNA under applied electric filed depends on the size of the nucleotides.

This work has resulted in following publications:

1. B. I. Morshed, M. J. I. A. Saude, “Molecular Dynamic Simulation Of Intrinsic Electrostatics Of Single-Stranded DNA”, NHGRI Advanced Sequencing Technology Development Meeting, (accepted), May 2013

2. M. J. I. A. Saude, B. I. Morshed, “Electrostatics of Single-Stranded DNA: A Prospective for Single Molecule Sequencing”, Applied Physics Letters, (submitted)

E. Outline

This thesis is organized in this format. In chapter 2, fundamentals of DNA sequencing and several methods of DNA sequencing techniques are discussed. In chapter 3, the general methods, the molecules modulation, and analysis procedures used in this thesis are discussed. In chapter 4 the final results and new approach that might help in DNA sequencing are presented and discussed. The last chapter concludes with major findings and indicates future research directions.

II. LITERATURE REVIEW

A. The structure of the DNA

Deoxyribonucleic acid (DNA) is a long chain of sequential order of four types of nucleotide bases: Adenine (A), Thymine (T), Cytosine (C), and Guanine (G). Fig.1 shows the chemical structure of DNA and the bases. The combination of these bases acts as information storage. These bases are bonded together to form a unit called base pairs, Adenine bonds with Thymine using two hydrogen bonds, while Cytosine bonds with Guanine using three hydrogen bonds. These bases are divided into two groups: Purine (A and G) and Pyrimidine (C and T). Each nucleotide contains phosphate-sugar backbone, and bases. The distance between each subsequent nucleotide base is ~3.4 Å, and the diameter of a DNA double-stranded helix is between 22 to 26 Å [1] .The genome is the complete set of DNA of organism [2]. A complete human genome contains 3 billion base pairs.

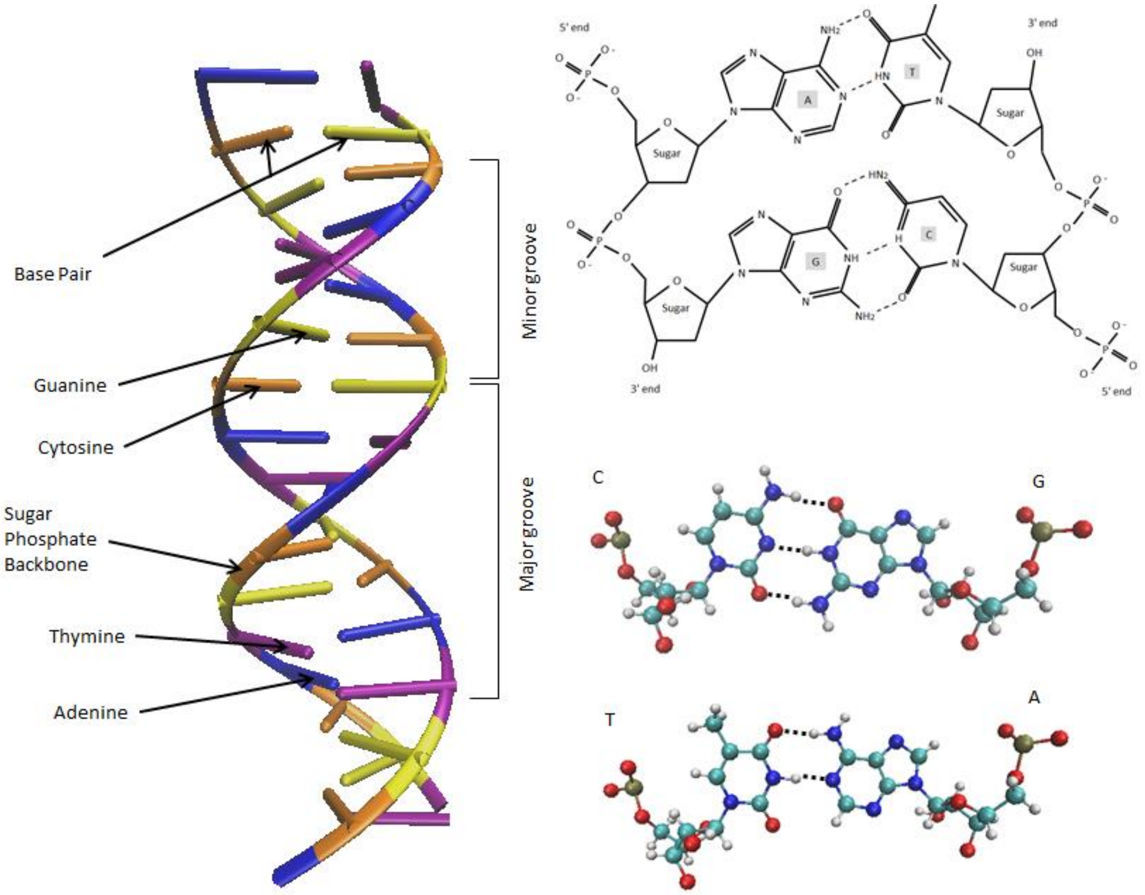


Figure 1. Left: the structure of the dsDNA. Right: the nucleotides of the DNA.

B. DNA SEQUENCING

1) Overview

Detecting DNA sequence of nucleotide bases is called DNA sequencing. Fred Sanger devised the first method of DNA sequencing in 1977 by using 2’, 3’-dideoxy nucleotides as chain terminating inhibitors of DNA polymerase [3]. Fig. 2 shows the basic concept of Sanger method. In 1986 and 1987 this method was developed by labeling the fragment with fluorescent dyes attached to the primer and to the dideoxy chain terminator to make the sequencing automated and faster [4][5]. By using dye-primer chemistry and dyeterminator chemistry, this allows all process to be carried out and run in single tube and single gel lane respectively. The laser would use to excite the fluorescent dyes and the detector collects the four different wavelengths.

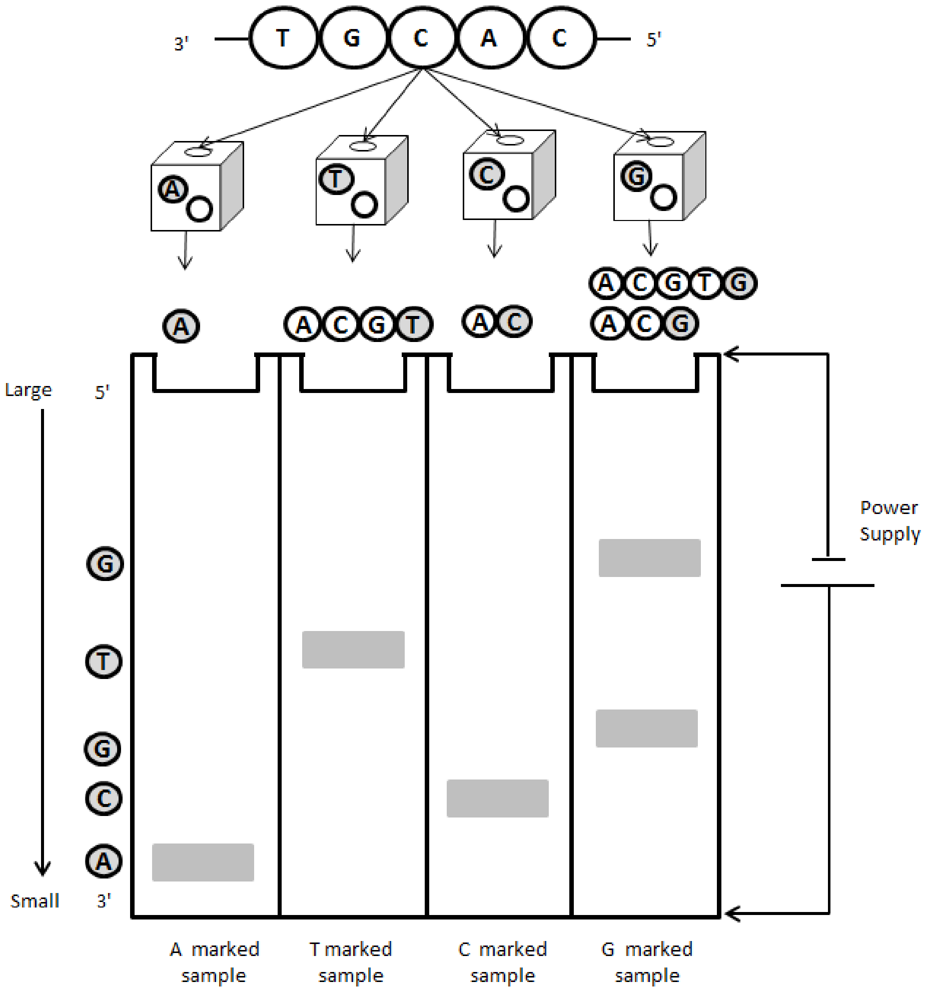


Figure 2 : Sanger sequencing method: Gray circles are terminator (ddNTP), white circles (dNTP)

According to the National Human Genome Research Institute (NHGRI), the DNA sequencing cost in 2001 was one hundred million dollars, but was drastically reduced to about 10 thousand dollars in 2012 [6], as researchers explored and developed new and improved DNA sequencing technologies. Since then, faster and more inexpensive DNA sequencing approaches have been developed, producing the second generation of sequencing technologies, such as DNA microarrays, and automated DNA sequencers [7]. The next generation of sequencing technology, often referred to as the third generation, is envisioned to be based on individual molecules of DNA with the aim of achieving an unprecedented sequencing rate in order to sequence a complete human genome in several hours with less than $1000 per genome [8]. This requires new sequencing technology development using an interdisciplinary approach. Two of such new single molecule based DNA sequencing technique demonstrating early promises are SMRT and nanopore technologies.

2) Single Molecule Real Time Sequencing (SMRT)

The SMRT technique utilizes a zero-mode waveguide (ZMW) to capture fluorescence emitted from individual tags as DNA is synthesized using DNA polymerase enzyme [9],[10]. This technology was developed by Pacific Biosciences to sequence the single molecule in real time by synthesis technology. It uses a silicon dioxide chips containing holes with femtoliter volume terminating to optical waveguide, called zero-mode waveguides (ZMWs). The ZMW is designed to create a nano-scale visualization volume to observe emissions of photons from only single labeled nucleotides. The sequencing is performed by attaching a single DNA polymerase to each ZMW, adding different dyed nucleotides to the concentrations, and allowing the polymerase to incorporate the complementary nucleotides (Fig.3). During the incorporation of each fluorescent tag that attached to the nucleotide generates a fluorescent signal in ZMW area. This signal can be detected using a detector and create a visualization chamber that allow to read the sequencing of the DNA. Massively parallel visualization of an array containing thousands of ZMWs enables rapid sequencing of DNA.

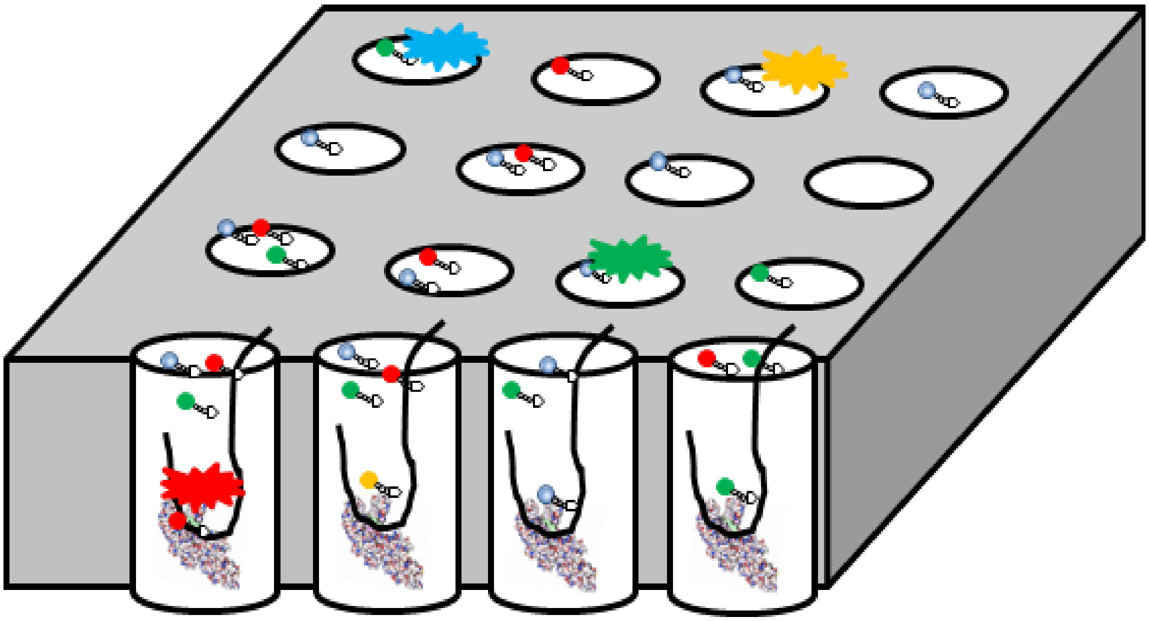


Figure 3. SMRT (single molecule real time) DNA sequencing. The DNA polymerase attached to the bottom of each of the ZMW. Once each nucleotide is incorporated, the tag on the terminal phosphate will split and diffuse in the ZMW volume, emitting corresponding fluorescence.

3) Nanopore sequencing

The concept of Nanopore sequencing started in 1996 when researchers showed that the single stranded DNA and RNA could be driven by electric field through 2.6nm diameter ion channel in a lipid bilayer membrane [11]. In nanopore technology, a nano-meter-sized pore is fabricated on a membrane through which a single stranded DNA molecule is passed under an applied voltage. An electric current is established through the nanopore that changes with the characteristic of the bases when a DNA strand is translocated through the pore from one side of the membrane to the other side (Fig.4). The changes of the current flow through the nanopore are monitored, and the sequence is decoded from the patterns of the relative current flow during this translocation process. By observing several independent parameters, such as time duration and temperature, along with the blocked current, the discrimination accuracy among various nucleotide types increases [12-14].

Nanopore sequencing has several challenges; one of these is the rate of translocation of the DNA is very fast which means the analytical resolution is not sufficient. By modifying the bases chemically, that allows slowing down the translocation and enables the detection of the bases of the ssDNA [15]. Another work done by molecular simulation shows that the movement of the ssDNA can be controlled using a transistor by trapping each base by one nucleotide spacing [16]. Fabricating the nanopore reliably is one of the major challenges, as more than nucleotide affects the blockage current in the nanopore making the detection more difficult. The nanopore sequencing offers the potential to read the longest lengths of all the single molecule methods.

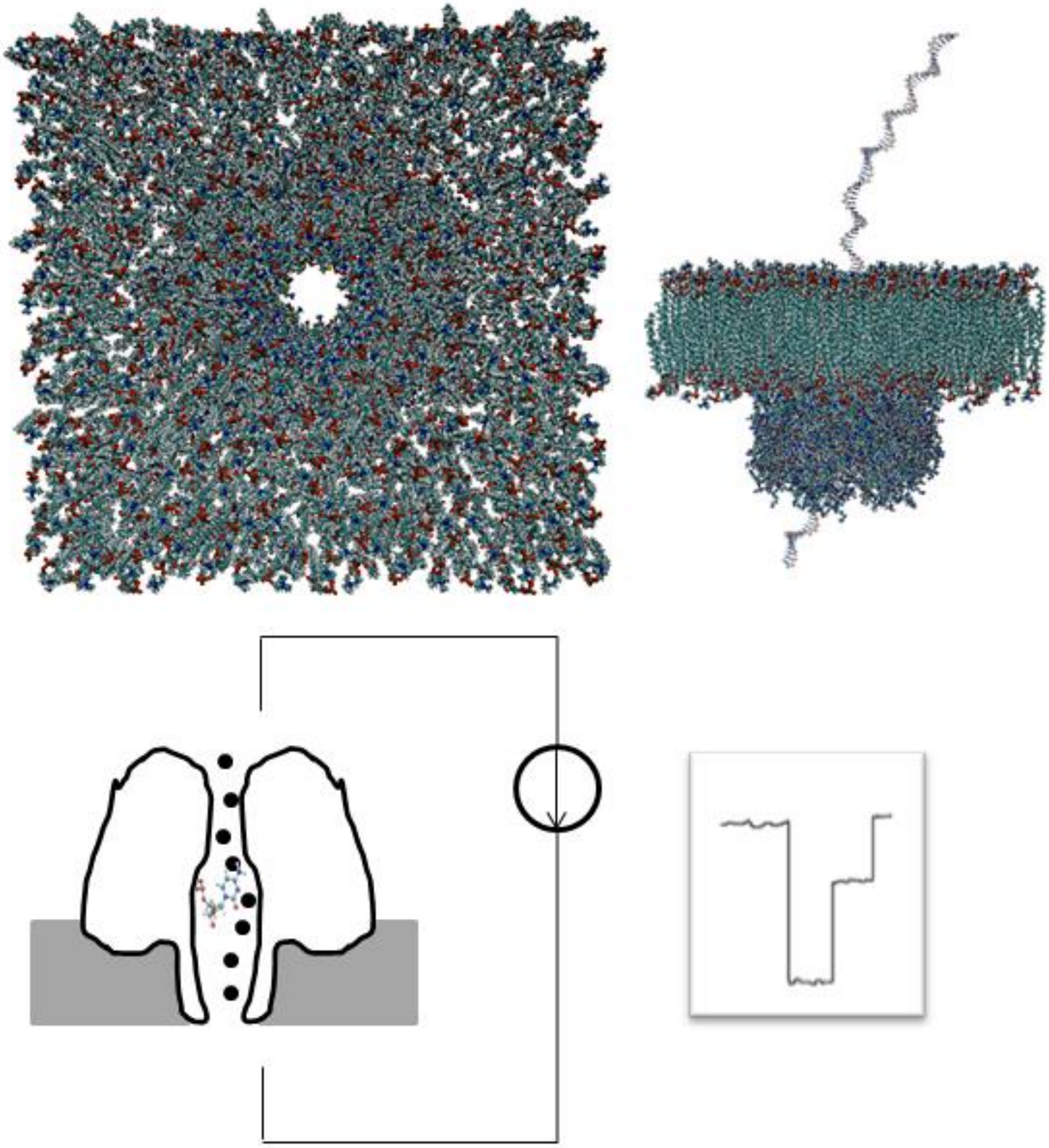


Figure 4. Nanopore sequencing. Top-left: the top view of the nanopore. Top-right is the nanopore while the DNA is translocating. Bottom: measuring the current while the nucleotide is blocking the ionic current.

4) Sequencing using the electrostatic characteristics

The electrostatics around a molecule can be calculated using Poisson-Boltzmann equations [17], and the charges, atomic arrangements, surrounding medium and shapes of the DNA and the dielectric boundary could affect the electrostatic potentials significantly [18]. It has been shown that the electrostatic interaction between two DNA molecules in a membrane depends on the charge density of the membrane and the distance between the two molecules [19]. Electrostatic interactions play a key role in various aspects of the structure and function of nucleic acids since they are highly charged molecules [18]. Using Molecular Electrostatic Potential (MEP), it has been shown that there are different patterns for paired and mispaired nucleotides located in the central plane between bases [20]. Several solid-state devices have been proposed to detect single molecules of DNA by studying the change of capacitances induced by the DNA molecule within the nanopore [21, 22]. These devices use a metal-oxide-semiconductor (MOS) capacitor that applies an electrical field to translocate the DNA molecule through the nanopore, while the response induced from the DNA backbone and its bases are recorded. Using a nanopore of 1 nm diameter in a MOS capacitor, the recorded voltage induced from different bases is between 2 to 9 mV (Fig. 5). These potentially could be exploited to detect the number of nucleotides and could lead to an electronic sequencing approach [23]. Furthermore, defective DNA can be detecting by calculating the electrostatic potential that induced by translocation the DNA through carbon nanotube at point located above the center of the CNT, probed at a distance of 2 nm. As the defective or mutated DNA has less charge than a normal base, the electrostatic signal that generated by the defective bases will be smaller than the normal bases [24].

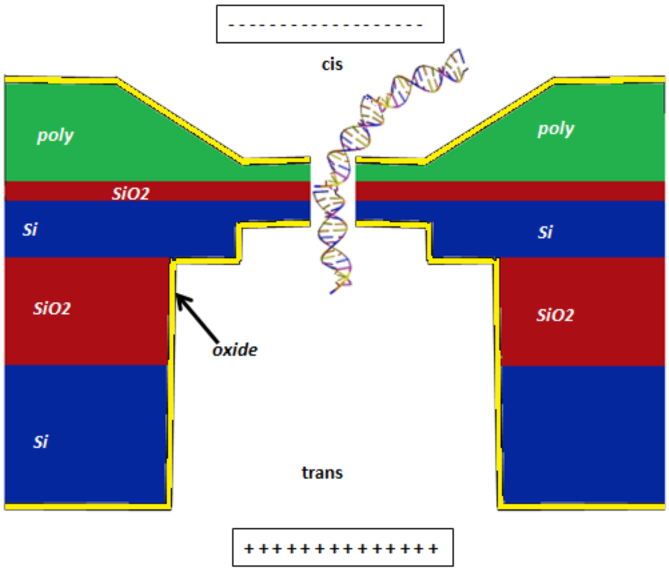


Figure 5. The schematic diagram of a nanopore in a capacitor membrane.

C. Molecular Dynamics

1) Overview

The meaning of computer simulation is to build a model of an actual system, using computer software and execute that model on a computing system. In the atomic level, the simulation could be used to explore the macroscopic properties of a system through microscopic simulations. Beside the real experiments, the simulation helps us to learn insightful information about the system behavior, as systems can have the same macroscopic properties, but quite different microscopic properties [25]. In other words, the simulation for the chemistry and biology is like a virtual laboratory. In the words of The Economist, October 17, 1998, the 1998 Nobel prizes: "*In the real world, this could eventually mean that most chemical experiments are conducted inside the silicon of chips* *instead of the glassware of laboratories. Turn off that Bunsen burner; it will not be wanted in ten years*."

There are two main groups of molecular simulation: molecular dynamics (MD) and Monte Carlo (MC). The Monte Carlo simulation is not time dependent and the output from the simulation depends only on the immediate state, while the molecular dynamic configuration depends on time so the state can be predicted at any time [26]. The MD goes in details for each motion of the individual molecules for all atoms over a small time step [27]. MD simulation is recommended for liquid whereas MC simulations are preferable for low density and gases system [27]. In this thesis, MD simulation is used. Five Steps to construct a simple MD program [28] as shown in Fig.6:

1. Input all parameters (Temperature, Density, Number of particles, Time step etc.)
2. Compute the forces on all particles.
3. Integrate Newton's equations of motion.
4. Perform statistical analysis on each new configuration of atoms.
5. After completion of the central loop, we compute and print the averages of calculated quantities, then stop.

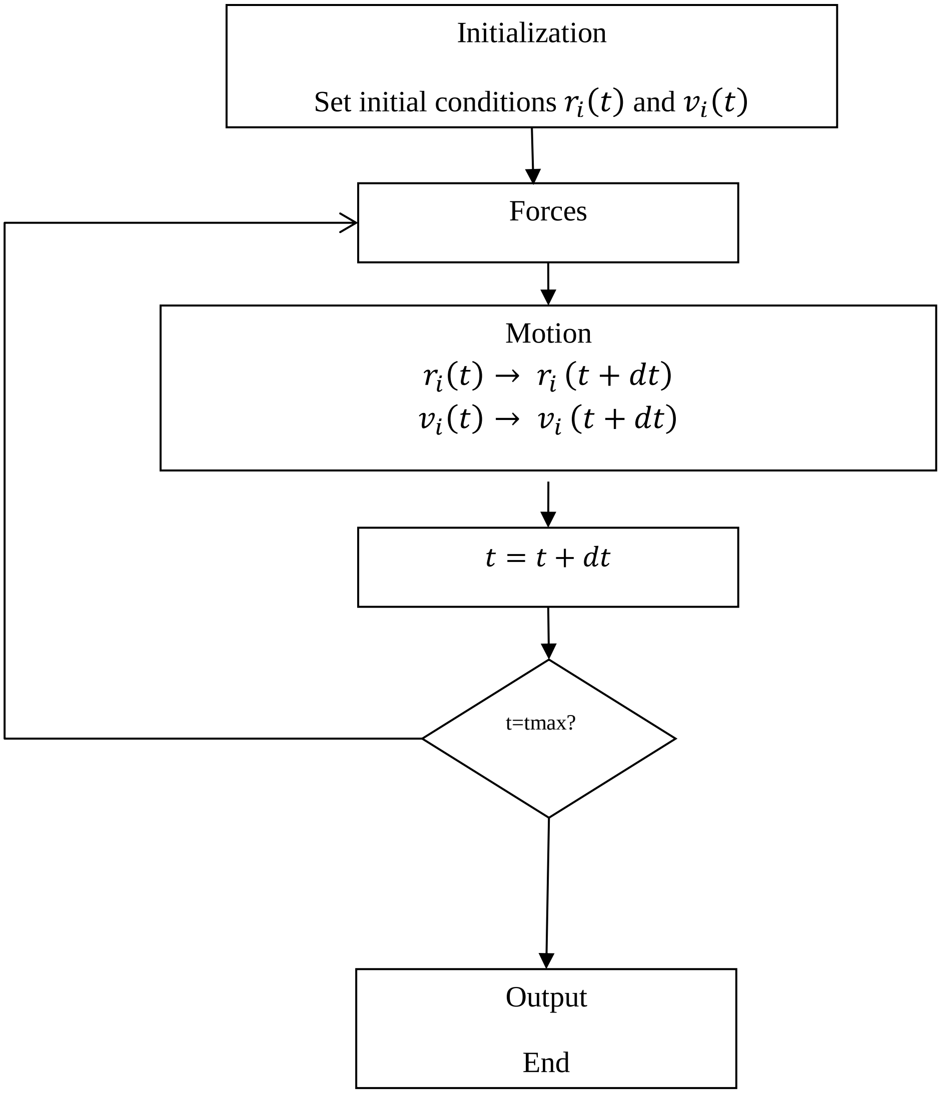

Initialization
Set initial conditions $r_i(t)$ and $v_i(t)$
Forces
Motion
$r_i(t) \rightarrow r_i\,(t+dt)$
$v_i(t) \rightarrow v_i\,(t+dt)$
$t = t + dt$
t=tmax?
Output
End

Figure 6. The flowchart of a simple MD program.

The MD simulation started in the late 1960s for a simple system. With increasing of computational system performance and capability, the demands on computer simulations have increased. Nowadays MD packages can simulate more than one million atoms using tools such as NAMD, GROMACS, AMPER, LAMMP, or DL POLY. All of these packages can be run on HPC (High performance computer) and parallel computing system. The author of [29] and [30] has done an experiment with different MD packages on different platforms, and his results show that NAMD is not the fastest software but shows the best scaling among the other. In this work we chose NAMD for all simulations, because it is recommended for simulating biological entities such as the proteins and DNA, and it is available on the HPC platform at university of Memphis.

The most important aspect for MD simulation is the actual runtime, for example some systems take one day or more to simulate hundred thousand of molecules for nanoseconds. To increase the performance of MD there are some approach that can be opted such as parallel computing, supercomputer, coarse-grained model and GPU. Graphics processing unit (GPU) usually shows better performance than CPU in these simulations [31]. HPC has developed CUDA-enabled version for GPU with MD simulation purposes and it was shown that a single CUDA-enabled version is about 26 times faster than its double precision CPU version.

The other approach to speed up the simulation is coarse-grained (CG) model. CG-MD is done by using “pseudo-atoms” to represent groups of atoms instead of using the trajectory for every atom (in contrast to All Atoms MD). CG model is important for proteins, nucleic acids and lipids simulations [32]. The disadvantage of CG models is not reproducing all features of the all-atom models.

2) CHARMM Force Field

The force field is mathematical functions that describe the potential energy of a system of particles (Fig. 7). The force field contains two main parts the topology (i.e. the mathematical function) and the parameters that used in the topology. The most popular force field which is used in molecular simulates are CHARMM, AMBER, and GROMOS. CHARMM (Chemistry at HARvard Macromolecular Mechanics) refer to the force field and MD simulation package. CHARMM force field was designed for proteins and it has three different versions: CHARMM19 which is used for united-atom, while CHARMM22 and CHARMM27 which are used for all-atom. CHARMM22 (1991) and CHARMM27 (1999) were designed for pure protein and nucleic acid, and they are equivalent except CHARMM27 was optimized for simulating DNA. CHARMM27 was chosen in this work to do all of the simulations, calculations, and naming schemes. The potential energy is the bonded energy (bonds, angles, and dihedral) in addition to the non-bonded energy (non-bonding van der Waals and electrostatic interactions between atoms). The expression for the potential energy is given mathematically as,

$$V = \sum_{bonds} k_b (b - b_0)^2 + \sum_{angles} k_\theta (\theta - \theta_0)^2 + \sum_{dihedral} k_\emptyset [1 + \cos(n\emptyset - \delta)] + \sum_{impropers} k_\omega (\omega - \omega_0)^2 + \sum_{Urey-Bradley} k_u (u - u_0)^2 + \sum_{residues} u_{CMAP}(\Phi, \Psi) + \sum_{nonbonded} \epsilon \left[\left(\frac{r_{ij}^{min}}{r_{ij}} \right)^{12} - \left(\frac{r_{ij}^{min}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{\epsilon r_{ij}} \qquad (1)$$

Here, V is the total potential energy. b is the bond length between two atoms. θ is the angle between three atoms. ∅ is the dihedral angle (i.e. between the plane containing the first three atoms and the plane containing the last three). ω is the improper angle of two planes of 4 atoms whereas the center atom bonds to the other 3. Urey-Bradley is a potential energy added to the angle which is a virtual bond between the first atom and the third atom where u is the distance between them. CMAP is a correction map to the dihedral energy and Φ and Ψ are the dihedral angles. r_{ij} is the distance between any non-bonded atoms. r_{ij}^{min} is the distance between any non-bonded atoms for which the energy is equal zero. q_i and q_j are the charges of atoms.

All the constant values such as the naught terms (e.g. b_0 in Ubond, θ_0 in Uangle etc.), the various force constants (k_b, k_θ, etc.), the partial charges q_i and LJ parameters (ε, $r_{min_{ij}}$) are taken from the force field parameters. (Source: CHARMM tutorial available at: http://www.charmmtutorial.org/index.php/The_Energy_Function).

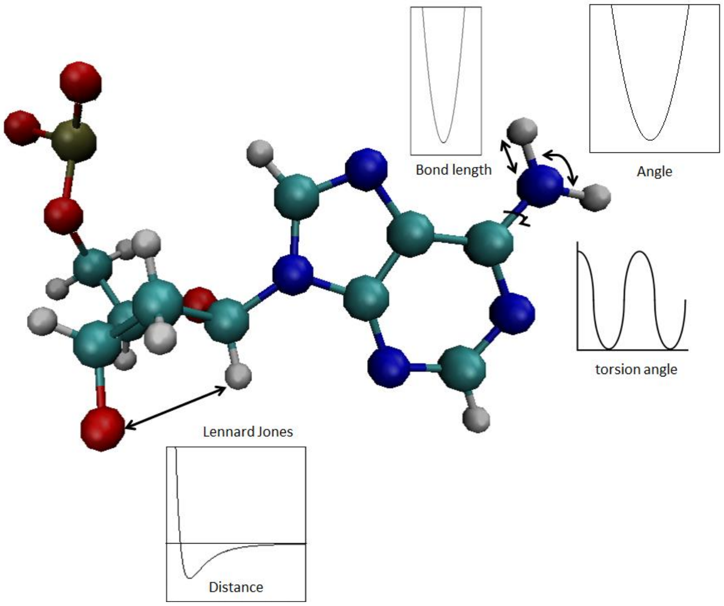


Figure 7: Molecular potential energies. Bond energy, angle energy, torsion angle energy and non-bonded energy (Lennard Jones).

III. SIMULATION SETUP AND METHOD

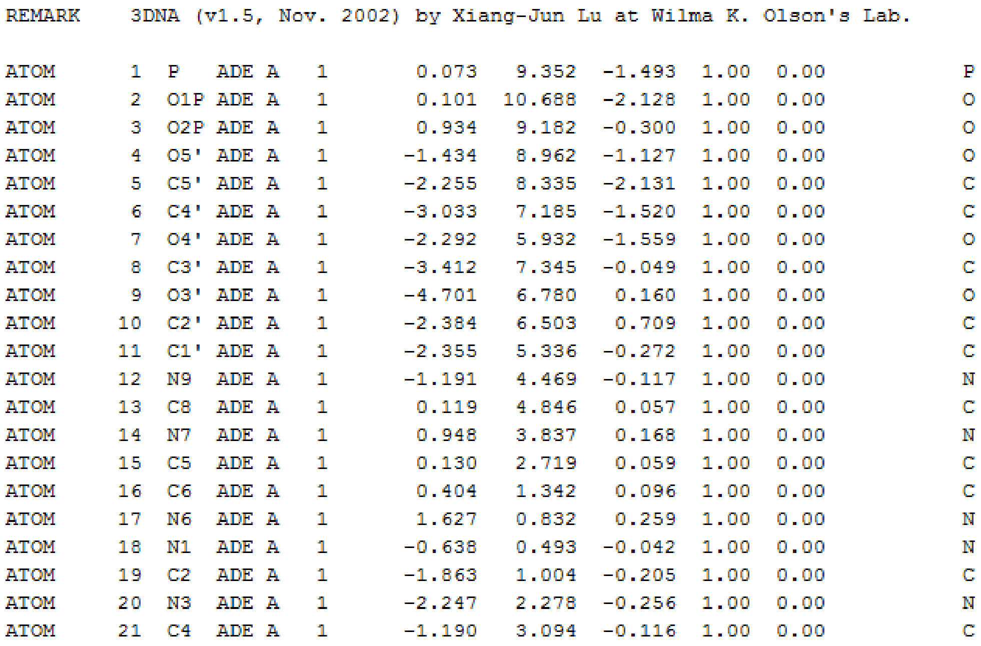
We utilize Molecular Dynamics (MD) simulation as the tool for this analysis. We investigated a single-stranded DNA (ssDNA) molecule, as such molecule will have a higher degree of electrostatic variation among the nucleotide bases compared to double-stranded DNA (dsDNA), as ssDNA contains nucleotide bases exposed for spatial probing without being bound to complementary bases while dsDNA nucleotides bases are bound to complementary bases and located inwards of the helix structure. Hence, we develop electrostatic models of ssDNA, measure and analyze the Molecular Electrostatic Potential (MEP) around the ssDNA, and explore detectable variations among four types of nucleotide bases (A, G, C, and T).

A. Preparing the Molecules

Initially, all of the molecules were in Protein Data Bank (format) so it needs some modifications to prepare more realistic scenario for simulation. Most of the preparation processes were done using VMD (Virtual Molecular Dynamics) and its plugin tools, along with pdb2pqr program.

1) Creating the PDB files of the dsDNA

The DNA molecules that are studied in this work are nucleic acids with conformation of B-DNA, which has been directly observed in functional organisms [33]. The pdb (Protein Data Bank) files of the DNA have been created from the publicly available tool at 3D-DART [34]. The pdb file is a text file that contains all atoms and its structure according to the X-Ray diffraction and NMR studies (Fig. 8).



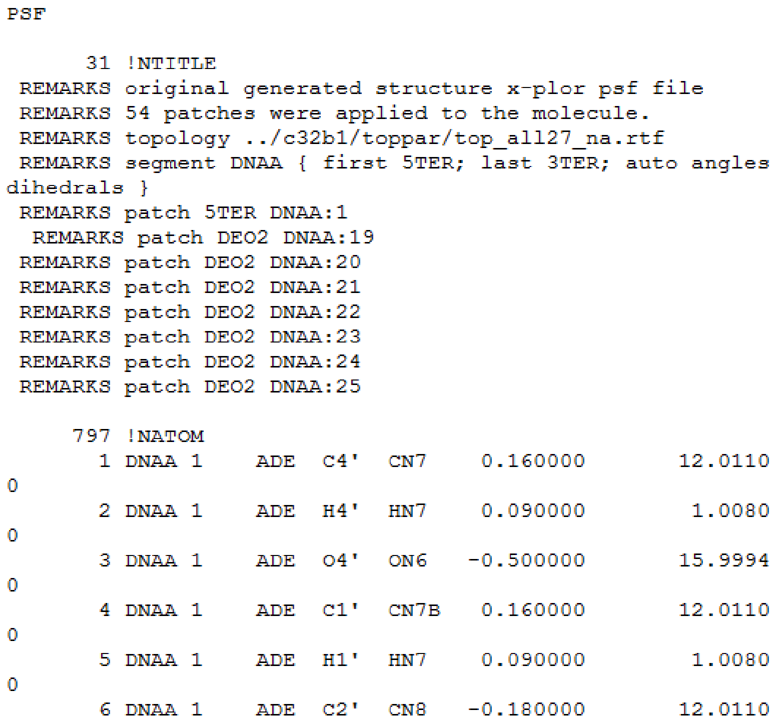
```
REMARK    3DNA (v1.5, Nov. 2002) by Xiang-Jun Lu at Wilma K. Olson's Lab.

ATOM      1  P   ADE A   1       0.073   9.352  -1.493  1.00  0.00           P
ATOM      2  O1P ADE A   1       0.101  10.688  -2.128  1.00  0.00           O
ATOM      3  O2P ADE A   1       0.934   9.182  -0.300  1.00  0.00           O
ATOM      4  O5' ADE A   1      -1.434   8.962  -1.127  1.00  0.00           O
ATOM      5  C5' ADE A   1      -2.255   8.335  -2.131  1.00  0.00           C
ATOM      6  C4' ADE A   1      -3.033   7.185  -1.520  1.00  0.00           C
ATOM      7  O4' ADE A   1      -2.292   5.932  -1.559  1.00  0.00           O
ATOM      8  C3' ADE A   1      -3.412   7.345  -0.049  1.00  0.00           C
ATOM      9  O3' ADE A   1      -4.701   6.780   0.160  1.00  0.00           O
ATOM     10  C2' ADE A   1      -2.384   6.503   0.709  1.00  0.00           C
ATOM     11  C1' ADE A   1      -2.355   5.336  -0.272  1.00  0.00           C
ATOM     12  N9  ADE A   1      -1.191   4.469  -0.117  1.00  0.00           N
ATOM     13  C8  ADE A   1       0.119   4.846   0.057  1.00  0.00           C
ATOM     14  N7  ADE A   1       0.948   3.837   0.168  1.00  0.00           N
ATOM     15  C5  ADE A   1       0.130   2.719   0.059  1.00  0.00           C
ATOM     16  C6  ADE A   1       0.404   1.342   0.096  1.00  0.00           C
ATOM     17  N6  ADE A   1       1.627   0.832   0.259  1.00  0.00           N
ATOM     18  N1  ADE A   1      -0.638   0.493  -0.042  1.00  0.00           N
ATOM     19  C2  ADE A   1      -1.863   1.004  -0.205  1.00  0.00           C
ATOM     20  N3  ADE A   1      -2.247   2.278  -0.256  1.00  0.00           N
ATOM     21  C4  ADE A   1      -1.190   3.094  -0.116  1.00  0.00           C
```

Figure 8. Part of a pdb file shows the atoms of the Adenine columns 7, 8 and 9 represent the location of the atoms in –x,-y, and –z coordinates.

2) Creating ssDNA and its PSF files

After downloading the pdb file, the dsDNA is separated into two chains and then converted to single-stranded DNA (ssDNA) molecules using Visual Molecular Dynamics (VMD) software (publicly available through the University of Illinois at Urbana-Champaign, IL, USA). Afterwards, the psf (Protein Structure File) files have been created using the “psfgen” plugin available in VMD by applying CHARMM27 force field. The psf file is a text files that contains all information about the atoms according to the topology file, such as charges, mass values, radius, etc. (Fig. 9).



```
PSF

      31 !NTITLE
 REMARKS original generated structure x-plor psf file
 REMARKS 54 patches were applied to the molecule.
 REMARKS topology ../c32b1/toppar/top_all27_na.rtf
 REMARKS segment DNAA { first 5TER; last 3TER; auto angles
dihedrals }
 REMARKS patch 5TER DNAA:1
  REMARKS patch DEO2 DNAA:19
 REMARKS patch DEO2 DNAA:20
 REMARKS patch DEO2 DNAA:21
 REMARKS patch DEO2 DNAA:22
 REMARKS patch DEO2 DNAA:23
 REMARKS patch DEO2 DNAA:24
 REMARKS patch DEO2 DNAA:25

     797 !NATOM
       1 DNAA 1    ADE  C4'  CN7    0.160000       12.0110
0
       2 DNAA 1    ADE  H4'  HN7    0.090000        1.0080
0
       3 DNAA 1    ADE  O4'  ON6   -0.500000       15.9994
0
       4 DNAA 1    ADE  C1'  CN7B   0.160000       12.0110
0
       5 DNAA 1    ADE  H1'  HN7    0.090000        1.0080
0
       6 DNAA 1    ADE  C2'  CN8   -0.180000       12.0110
```

Figure 9. Part of a psf file shows the atoms of the Adenine columns 8 and 9 represent the charge and mass respectively for each atom.

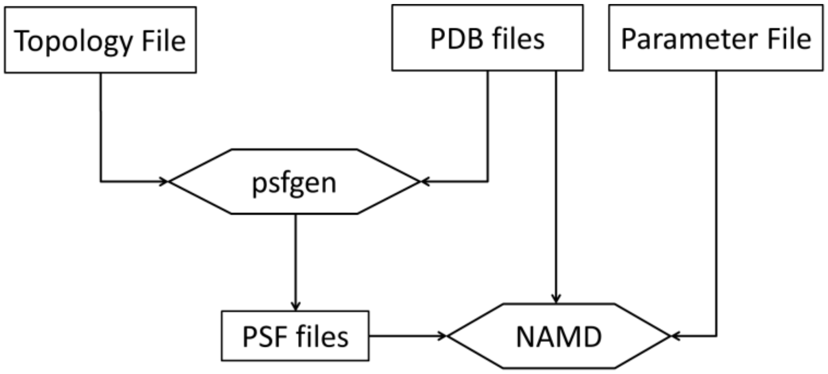


Figure 10. Sample structure preparation processes

3) Solvation and Ionization

The solvation is used to make system closely represent the real environment. Each ssDNA is solvated with a box of TIP3P water using “Solvate” plugin in VMD and the water molecules overlapping the DNA would be removed. The TIP3P is a computational model for the water that used in molecular dynamic simulation. The system was ionized with Na+ and Cl^- ions to achieve the desired concentration and to make the overall system neutral, using the “autoionize” plugin available in VMD. Fig.11 shows an example of an ssDNA after solvation and ionization. Solvation and ionization could be done using writing a TCL script to make it easy for applying the same parameters for all the simulations.

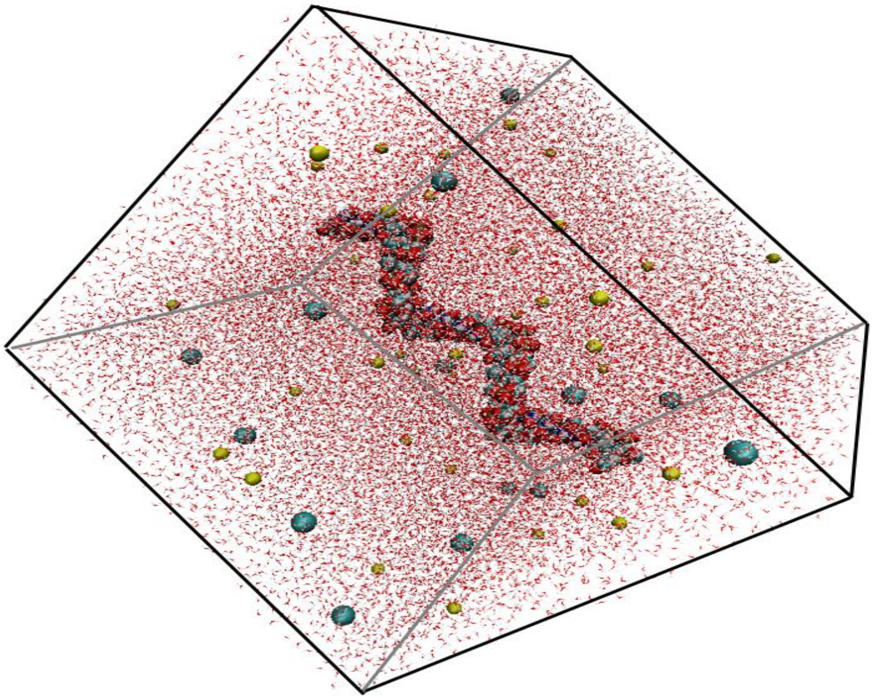


Figure 11. Simulation box after solvation and ionization

B. Analysis and Calculation

Each system, after the simulation, was visualized using the VMD by loading the psf files, pdb files and the trajectory files.

1) Molecular Electrostatic Potential (MEP)

The MEP is the electrostatic potential that is created from the molecules at a positive charge. MEP depends on the molecule charges, how they are distributed, and the location of the point (x,y,z). One of the most basic method to describe the interacts of the electrostatics between the molecules in ionic solution is Poisson-Boltzmann equation:

$$\vec{\nabla}.\left[\epsilon(\vec{r})\vec{\nabla}\psi(\vec{r})\right] = -4\pi\rho^{f}(\vec{r}) - 4\pi\sum_{i} c_i^{\infty} z_i q \exp\left[\frac{z_i q \psi(\vec{r})}{k_B T}\right]\lambda(\vec{r}) \qquad (2)$$

Where $\vec{\nabla}$ the divergence operator, $\epsilon(\vec{r})$ the dielectric, $\psi(\vec{r})$ the electrostatic potential, $\rho^{f}(\vec{r})$ the charge density of the solute, c_i^{∞} the concentration of ion i at an infinity distance from the molecule, z_i is its valency, q is the proton charge, k_B is the Boltzmann constant, T is the temperature and $\lambda(\vec{r})$ the accessibility to ions at point $\vec{r}$ [35].

To calculate the molecular electrostatic potential, all atoms must have corresponding charge and radius information. To achieve this, the pdb files of the ssDNA are converted to pqr files that have the corresponding atomic charge information, by using pdb2pqr software with the CHARMM force field. PDB2PQR is a Python software package to convert the PDB file format to PQR format (where P stand for PDB, Q for charge, and R for radius) by adding the missing atoms and hydrogen and assigning charge and radius parameters to the atoms from a specific force field (i.e. CHARMM). There are two procedures for using PDB2PQR: using PDB2PQR web server, or using the executable file on the local machine.

MEP is calculated using Adaptive Poisson-Boltzmann Solver (APBS) software (freely available macromolecular electrostatics calculation program governed by GPL). APBS is a software package use Poisson-Boltzmann equation for modeling bimolecular solution. FEtk (the Finite Element ToolKit) is used to solve the Poisson-Boltzmann equation numerically, where FEtk is libraries and tools written in an object-oriented of C and C++ for solving general coupled systems of nonlinear partial differential equations. APBS was written by Nathan Baker in collaboration with J. Andrew McCammon and Michael Holst and enhanced by contributions from several other authors. APBS VMD plugin provides a graphical user interface to perform electrostatic calculation while the APBS excitable file must be installed. APBS could be run using command-line. The APBS input file contains the information about the inputs, parameters, and the output for the calculation. The input files for electrostatic should contain the structure shown in Fig. 12.

```
READ
 ...
 ...
END

ELEC
 ...
 ...
END
QUIT
```

Figure 12 . The APBS input file structure.

Here, the read section should contain the input file in PQR format and the ELEC section should contain all information about the electrostatic parameters. This file should be saved with .in extension then execute the APBS with the following command:

exec apbs [--output-file=name | outputformat=type | ...] input-file

All of the outputs in our simulations are in OpenDX file format (i.e. with .dx extention). The OpenDX format is very flexible and is describe as shown in Fig. 13.

```
# Comments
  object 1 class gridpositions counts xPoints yPoints zPoints
  origin xO yO zO
  delta dx 0.0 0.0
  delta 0.0 dy 0.0
  delta 0.0 0.0 dz
  object 2 class gridconnections counts xPoints yPoints zPoints
  object 3 class array type double rank 0 items Points data follows
  u(0,0,0) u(0,0,1) u(0,0,2)
  u(0,0,3) u(0,0,4) u(0,0,5)
  ...
  u(0,0,nz-3) u(0,0,nz-2) u(0,0,nz-1)
  u(0,1,0) u(0,1,1) u(0,1,2)
  ...
  u(0,1,nz-3) u(0,1,nz-2) u(0,1,nz-1)
  ...
  u(0,ny-1,nz-3) u(0,ny-1,nz-2) u(0,ny-1,nz-1)
  u(1,0,0) u(1,0,1) u(1,0,2)
  ...
  attribute "dep" string "positions"
  object "regular positions regular connections" class field
  component "positions" value 1
  component "connections" value 2
  component "data" value 3
```

Figure 13. The structure of dx file (MEP file)

Here, xPoints yPoints zPoints is the number of the grid points in x, y and z direction, dx, dy and dz is the grid spacing in corresponding direction, xO,yO an zO is the origin of the box (i.e. the lower corner), Points is the total number of the grid points (i.e. equal to xPoints*yPoints*zPoints), and u(x,y,z) is the data values.

2) Analyzing the dx files

After calculating the MEP, the output files (dx files) was visualized using VMD software by loading it on the top of the molecule. There are four rendering methods to represent the MEP in VMD: volume slice, isosurface, field lines, and surface coloring. In the volume slice the data is mapped into two-dimensional slice where the colors represent the scalar data range with blue for low values, red for high values and green for the in between values (assuming the color scale BGR). The isosarface represent the data in 3-D surface corresponding to the data values within one scalar value. The field lines are used to represent the lines of the movement of the massless particles affected by the volume gradient vectors. While the surface coloring is by drawing the molecule using surf method according to the radii of each atoms and using coloring method by selecting the dx files to map the MEP to the surface. Fig.14 shows examples of four different types of rendering the data.

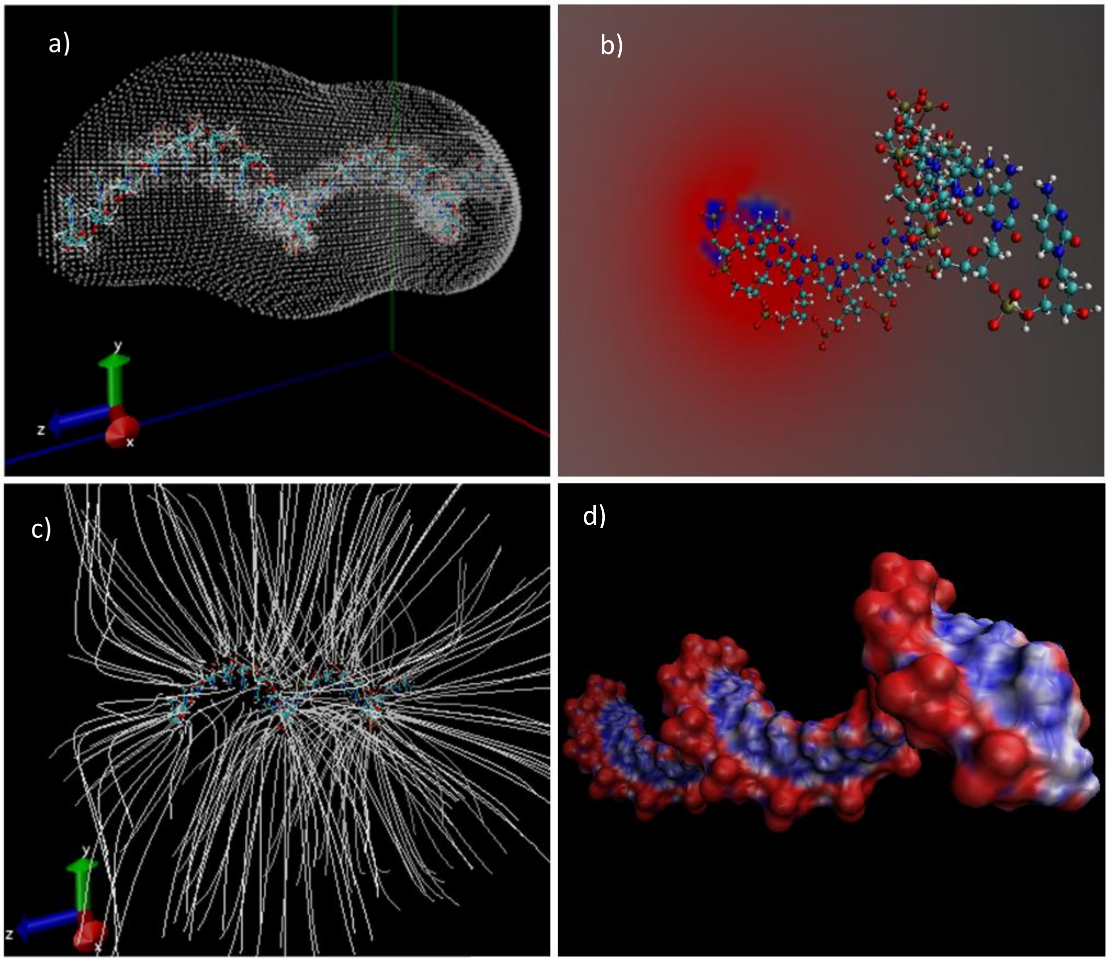


Figure 14. MEP rendering. a) isosurface, b) volume slice. c) field lines. d) surface coloring.

IV. SIMULATION RESULTS

The simulation results are rendered with VMD, and the electrostatic potentials have been visualized with VMD, analyzed and plotted with Matlab (MathWorks, MA, USA) and MicroSoft Excel (MicroSoft Corp., CA, USA).

A. The Results of modeling 25 bp long ssDNA

The length of ssDNA sequence studied in this work is 25 base-pair (bp) long, i.e. the polymers are poly(dA)$_{25}$, poly(dT)$_{25}$, poly(dG)$_{25}$ and poly(dC)$_{25}$, and have monitored electrostatics around the residue-13 nucleotide (13th base at the center of the ssDNA sequence). Later, we show that the length of 25 bp is sufficiently long to ignore the terminal effects on the bases under observation. Each system, consisting over 96,000 total atoms, is minimized using a scalable molecular dynamic simulation software, NAMD (publicly available through the University of Illinois at Urbana-Champaign, IL, USA). The minimization is simulated for 10,000 steps with 1 fs per step using CHARMM27 force field. The calculation presented in this work is based on the obtained system minimized to the stable state.

Each system was mapped onto a cube lattice of 161 × 161 × 161 points and the grid lengths are 132 Å in any of x-, y-, and z-directions (Fig. 15). MEP was calculated at each point inside the cube that can have a maximum of $\Delta/2$ error along x-, y-, or z-axis, where Δ (= 0.825 Å) is the distance between two measurement points (i.e. the grid spacing). The salt concentration is 0.025 mol/L, and the internal and external dielectric coefficients with respect to atomic volume are assumed to be 1 (e.g. vacuum within the atom) and 78.54 (e.g. surrounding water), respectively (Fig. 15).

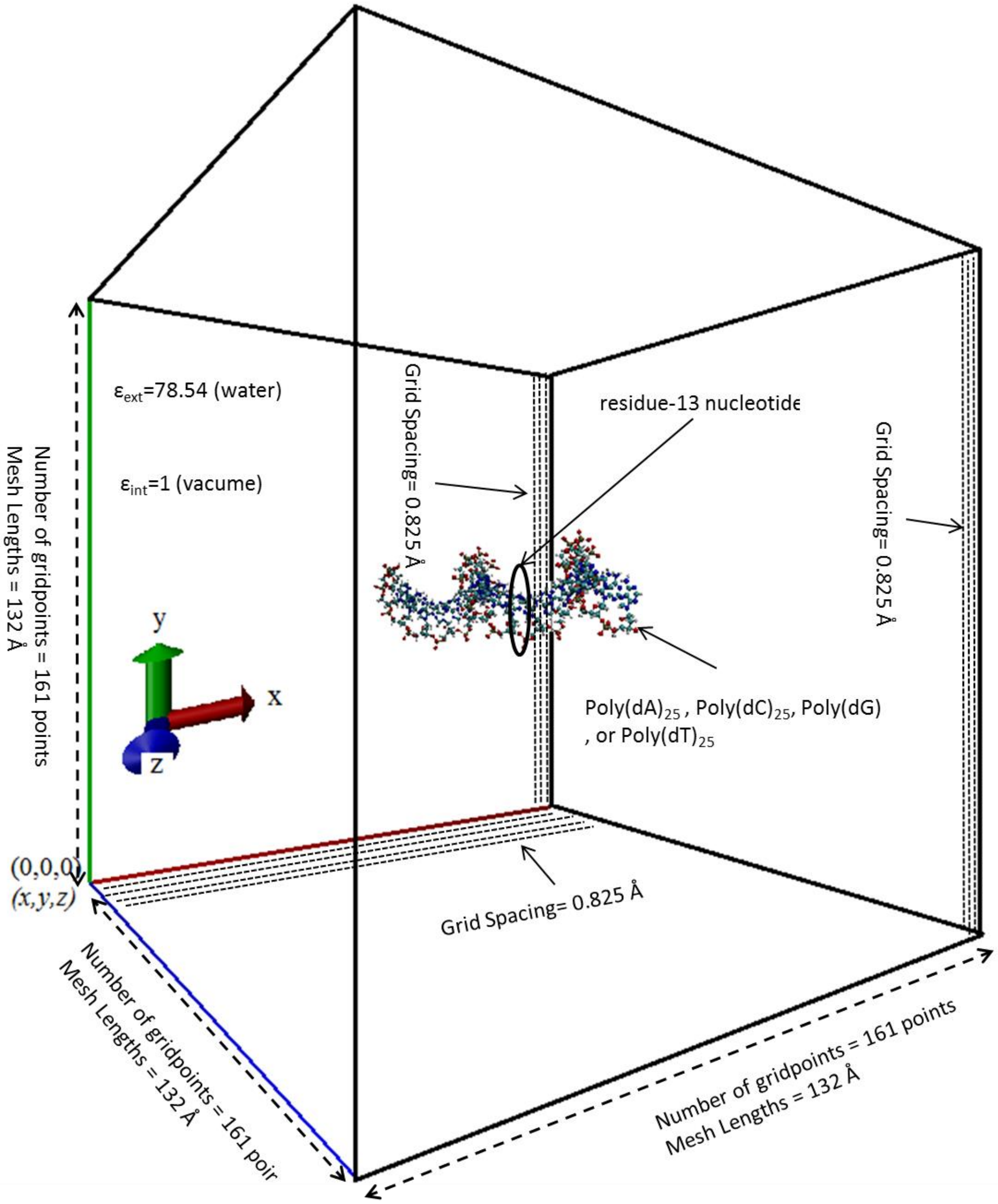

ε_{ext}=78.54 (water)
ε_{int}=1 (vacume)
Number of gridpoints = 161 points
Mesh Lengths = 132 Å
Grid Spacing= 0.825 Å
residue-13 nucleotide
Grid Spacing= 0.825 Å
y
x
z
Poly(dA)$_{25}$, Poly(dC)$_{25}$, Poly(dG)
, or Poly(dT)$_{25}$
(0,0,0)
(x,y,z)
Grid Spacing= 0.825 Å
Number of gridpoints = 161 poir
Mesh Lengths = 132 Å
Number of gridpoints = 161 points
Mesh Lengths = 132 Å

Figure 15. MEP box surrounded the ssDNA (poly(dA)$_{25}$).

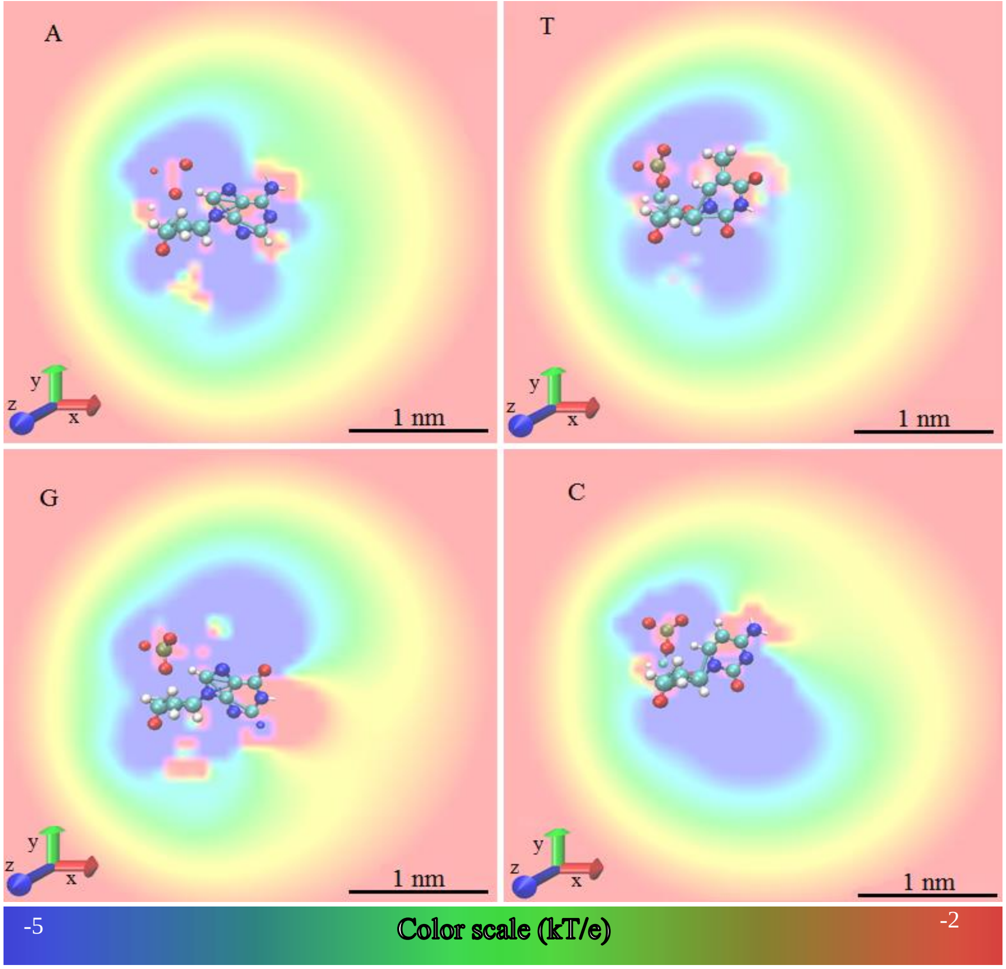


Figure 16. MEP distribution map around the residue-13 nucleotide (central base) in polymers of poly(dA)$_{25}$, poly(dT)$_{25}$, poly(dG)$_{25}$ and poly(dC)$_{25}$.

Fig. 16 shows two-dimensional electrostatic potential distribution map in a plane perpendicular to the DNA helix (oriented along the z-axis of the simulation box) around the residue-13 nucleotide (central base) of ssDNA polymers. The ssDNA is oriented such that the center of the sugar backbone and the center of the nucleotide base are on the x-axis. MEP evidently demonstrates slightly different patterns for each nucleotide. We note that the atomic electrostatic variations diminish within ~3 nm distance to bulk potentials.

Hence an effective electrostatic sequencing mechanism must probe within this distance from the ssDNA strand.

To study the variations around the ssDNA molecules, we analyzed MEP around the bases at four different radii (1 nm, 1.4 nm, 2 nm, 3 nm) on the xy-plane. The center of the sugar backbone of each nucleotide configuration is used as the reference, and the angle is measured counterclockwise from the center of the nucleotide along the x-axis as shown in Fig. 17. The values are plotted in Fig. 18 at every 2° for all four nucleotide bases.

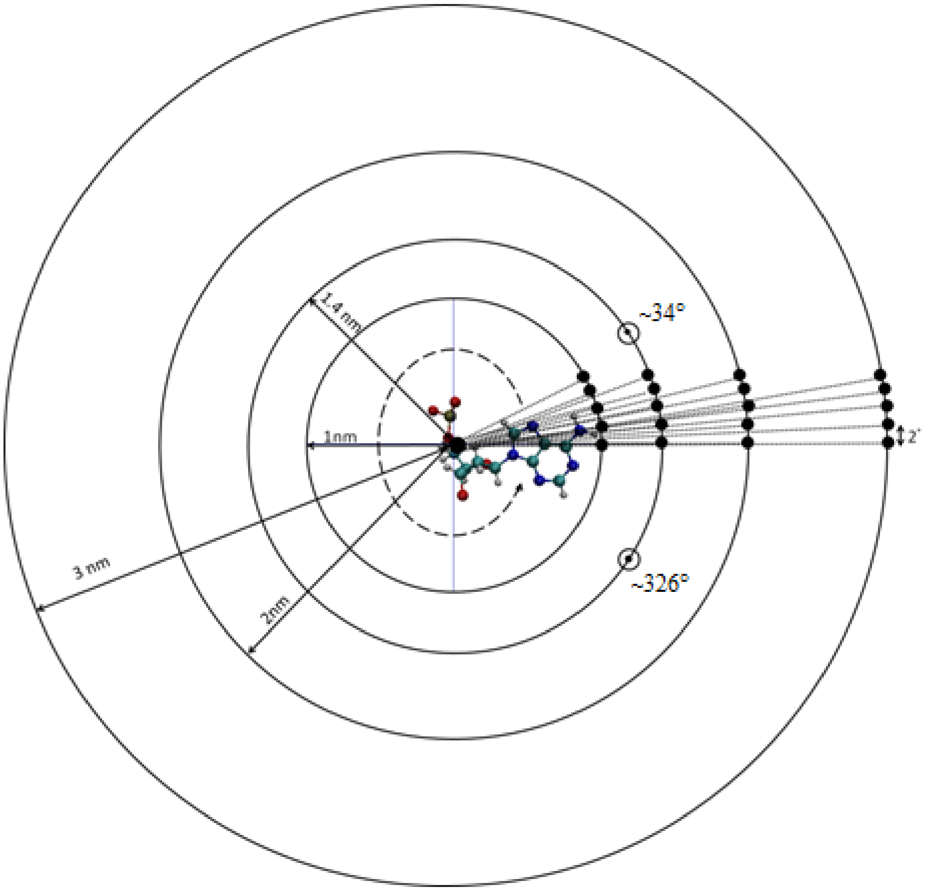


Figure 17. MEP surrounded the nucleotide 13.

Figure 18 depicts that the variations among the MEPs are the most diverse between radii 1 to 3 nm. We note that at distance of 1 nm, substantial variations are observed within 45° and after 315° due to the measurement points being inside the nucleotide atoms. The figure also indicates that there are two angular ranges where the variation among four bases peaks, one of them reside around ~34°, while the other is around ~326°. The corresponding data is presented in Table I. We have selected the radius of 1.4 nm from the center of the nucleotide backbone, a feasible probing distance, with an angular orientation of 34° from the x-axis for further analysis. Analysis along other radii and angular orientations are similar. The calculations are performed with various grid/lattice sizes with values ranging between ± 0.01 to ± 0.1 without any visually noticeable changes, indicating minimum sensitivity of the data on the grid/lattice size.

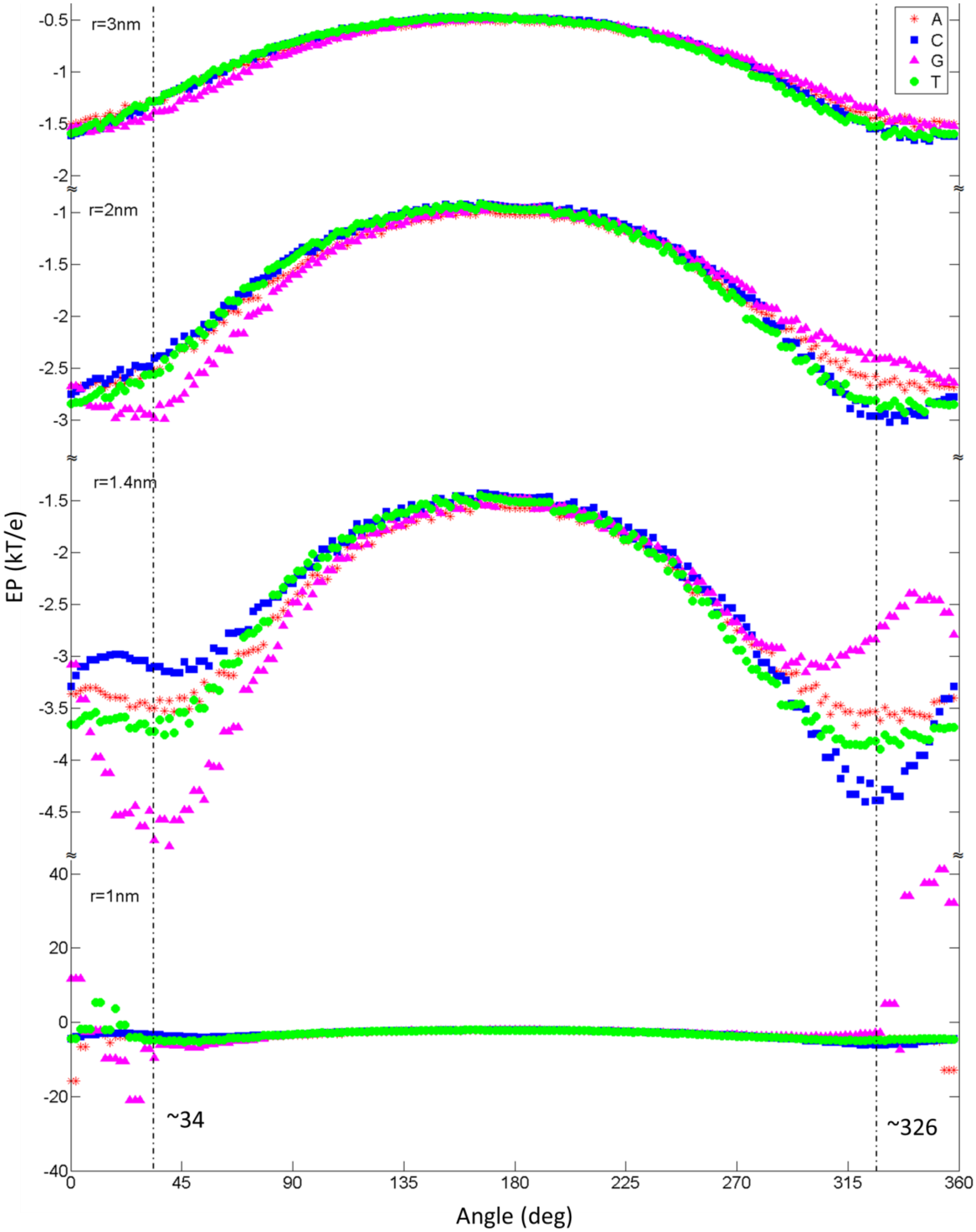


Figure 18. MEP on circles with various radii with the center at the suger backbone at various distances (r) for four types of nucleotides (A, C, G, and T). Data are plotted for every 2° counterclockwise from the x-axis.

Table 2: MEPs at r = 1.4 nm distance from the sugar backbone for residue-13 nucleotide (central base) in polymers of poly(dA)$_{25}$, poly(dT)$_{25}$, poly(dG)$_{25}$ and poly(dC)$_{25}$.

	MEPs for various nucleotide base types (kT/e)			
Angular orientation	poly(dG)	poly(dT)	poly(dA)	poly(dC)
~34°	-4.767	-3.722	-3.498	-3.098
~326°	-2.52	-3.816	-3.531	-4.386

Further, we investigate the influence on the electrostatic field distributions of the central nucleotide base from the neighboring nucleotides of the ssDNA strand. To reduce the number of combinations to be simulated, we limit the combinations of inserts to only identical bases symmetrically (with respect to residue-13) inserted at the middle of a polymer of ssDNA, replacing the corresponding bases of the polymer as shown in (Fig. 19). The 61 combinations are simulated and categorized into four groups based on the nucleotides being inserted (Fig. 20). In the first group, MEPs are calculated for 1, 3, 5, 7, and 9 base insertions of nucleotide base-A in the middle of polymers of poly(dA)$_{25}$, poly(dT)$_{25}$, poly(dG)$_{25}$, and poly(dC)$_{25}$. Similarly, other groups of simulations inserted nucleotide bases of T, G, and C into all four types of polymers.

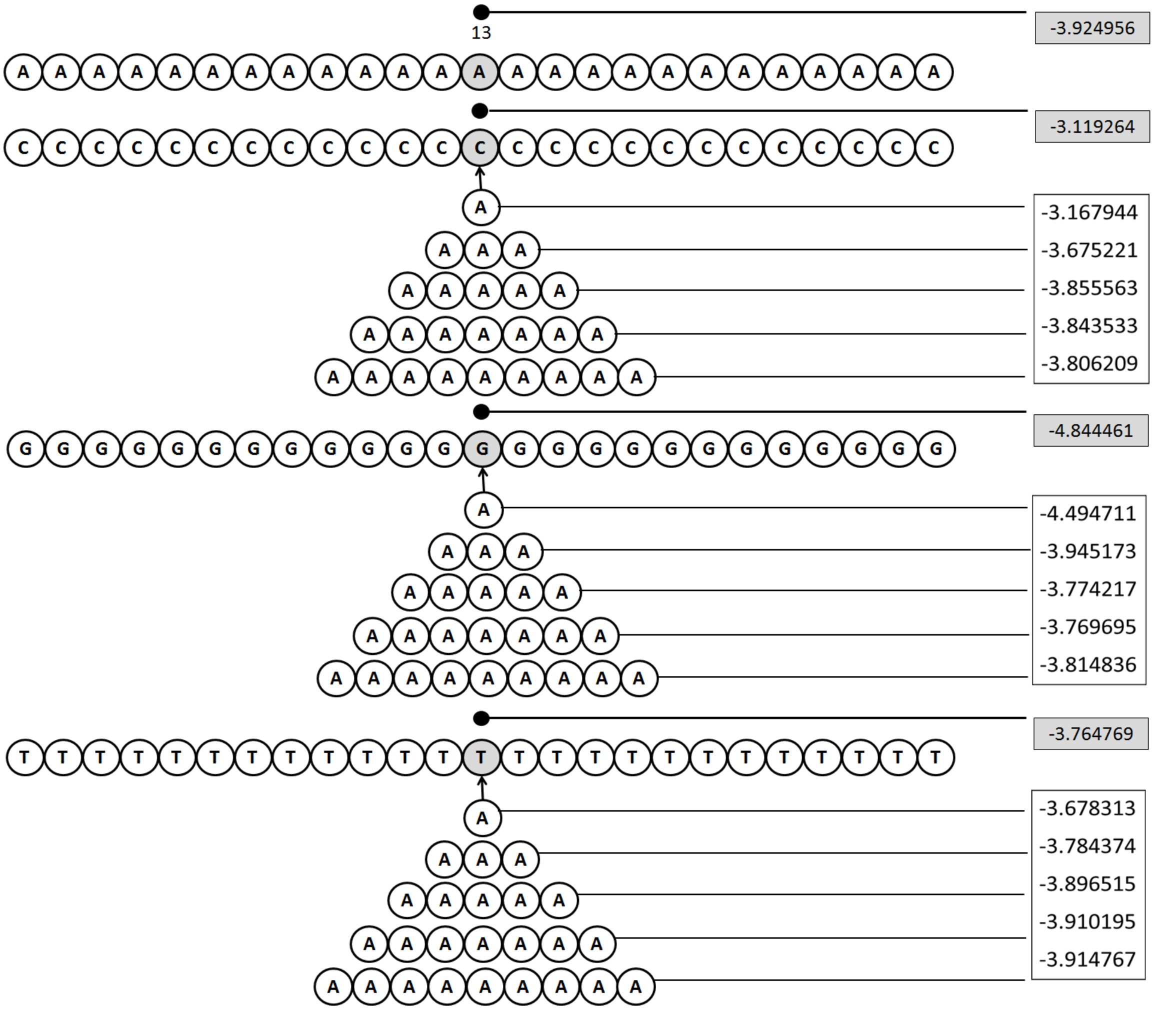


Figure 19: The procedure used to find the influence of the neighbor nucleotides. The gray circles are the measurement points (i.e. at 1.4 nm distance and ~34 degree angle). The scheme shown is for Adenine (A). The same procedures are applied for other bases (i.e. for G, C, and T).

An insertion of 1 base represents MEP influence by all adjacent bases of the same or different kinds (denoted as 1st neighbor influence). A symmetrical insertion of 3 bases yields the first neighbors on both sides to be of the same type, while from the 2nd neighbors to the terminals, original polymer bases are preserved (denoted as 2nd neighbor influence), and so on. Evidently the data shows the influence of neighboring bases diminishes for 5 or more inserted bases (3rd or higher neighbor influence). As the number of inserted base increases, MEP results trend to the MEP consisting of polymers of the inserted base. From this prospective, we note that simulations with 25 bp ssDNA for probing electrostatics at the residue-13 will have insignificant terminal effect on the analysis.

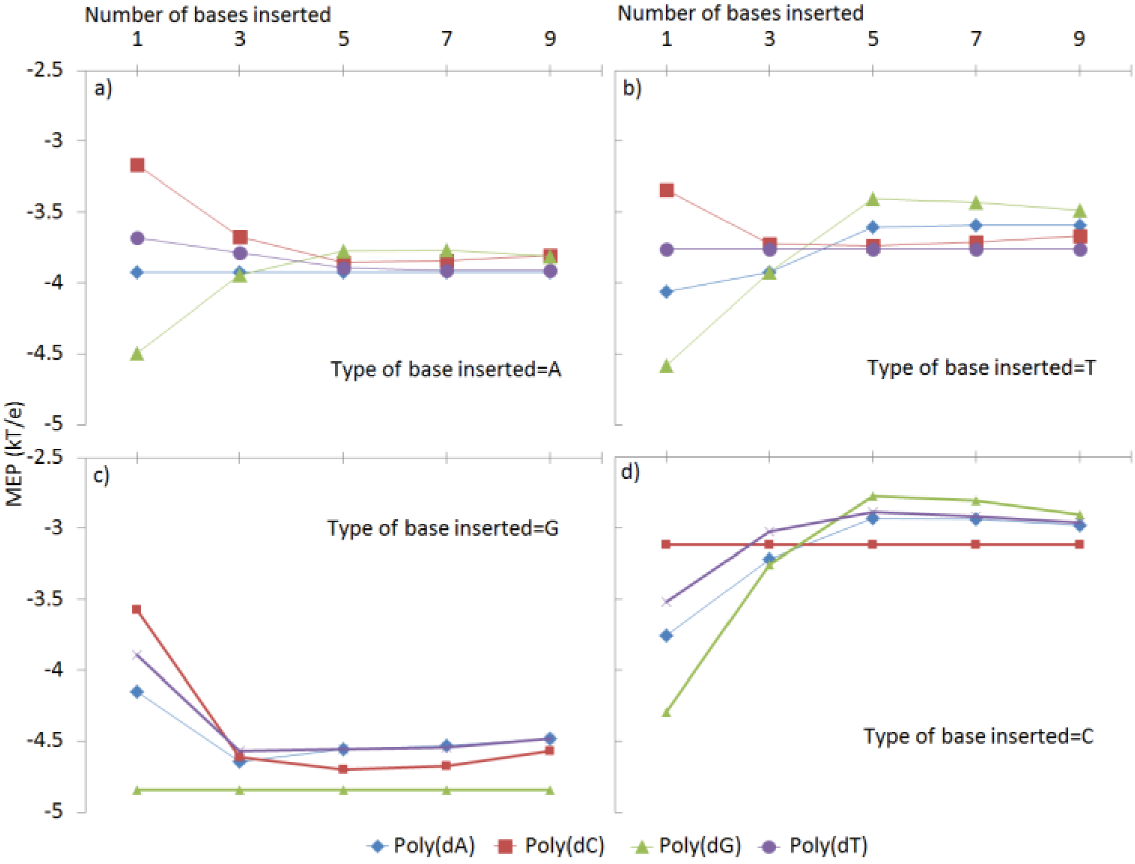


Figure 20. The influence of the neighboring nucleotides on MEP at the probing points (1.4 nm distance from the center of the sugar backbone of the ssDNA at 33.7° counterclockwise orientation from the x-axis). (a) Insertion of base-A. (b) Insertion of base-T. (c) Insertion of base-G. (d) Insertion of base-C. In all plots, the effect of inserting higher number of bases diminishes rapidly towards the polymer MEP of the inserted nucleotide type.

We investigate the MEP at the same point (i.e. 1.4 nm, with 34° angle) during the movement of the ssDNA, which occurs while minimization, heating, and equilibration. Each system was minimized for 10,000 steps, heated for 10,000 steps, equilibrated for 200,000 steps. The trajectory file was loaded to the VMD and we extracted the frames every 1 ps. The MEP was calculated for each frame and this procedure repeated for all nucleotides. Fig. 21 shows the resultant MEP variation over simulated time of 210 ps. The corresponding average and standard deviations of MEPs during equilibration are plotted in Fig. 22. We note that the averages are distinct for each base, however standard deviation for a few nucleotides are large, the largest being T due to sharp dips observed in the temporal MEP plot (Fig. 21).

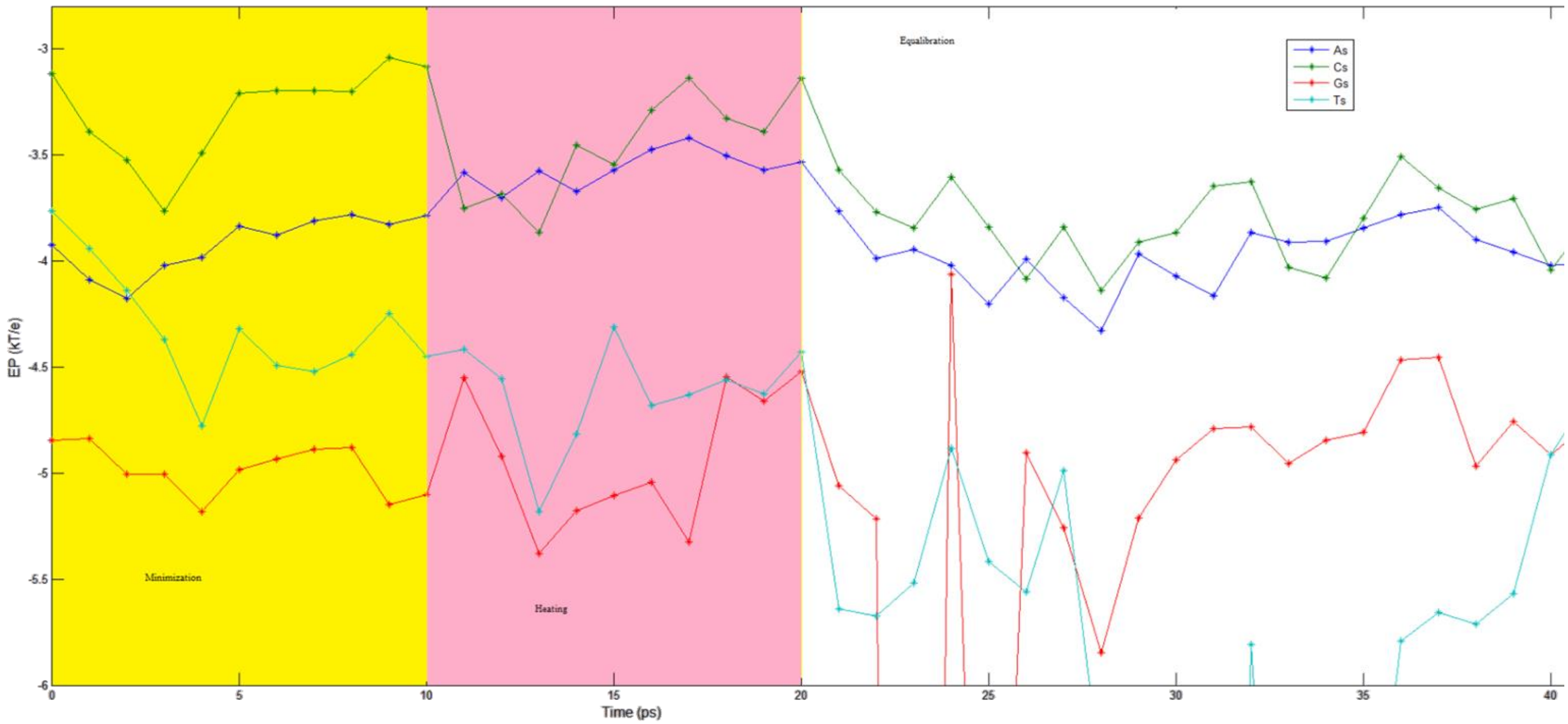


Figure 21. The MEP calculated every 1 ps for a duration of 210 ps.

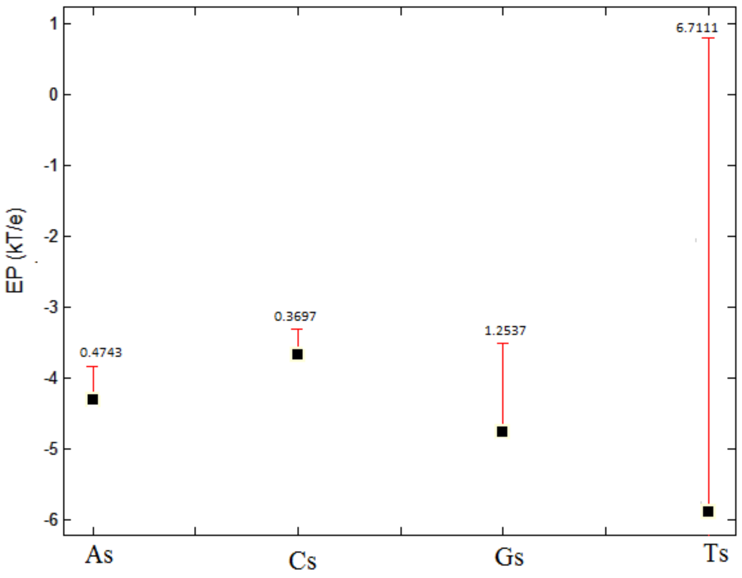


Figure 22. The average and standard deviation over the equilibration time for each nucleotide types.

B. The Results of modeling 5 long of ssDNA

In the second set of simulations, the length of ssDNA sequence studied is 5 base-pair (bp) long, i.e. the polymers are poly(dA)$_5$, poly(dT)$_5$, poly(dG)$_5$ and poly(dC)$_5$, and have monitored electrostatics around the 3rd nucleotide.

Table 2: The measurement of the 5 bp long ssDNA

Sequence	Number of Atoms	Length of polymer (Å)
poly(dA)$_5$	159	18.502
poly(dT)$_5$	159	18.499
poly(dG)$_5$	164	18.501
poly(dC)$_5$	149	18.507

MEP was calculated for each system with a cube lattice of 129 × 129 × 129 points, grid lengths are 21 × 28 × 29 Å in x-, y-, and z-directions respectively, grid spacing is 0.164, 0.21875, 0.22656 Å in x-, y-, and z-directions respectively, salt concentration is 0.025 mol/L, and the internal and external dielectric coefficients with respect to atomic volume are assumed to be 1 and 78.54 respectively, as shown in (Fig. 23) for Adenine.

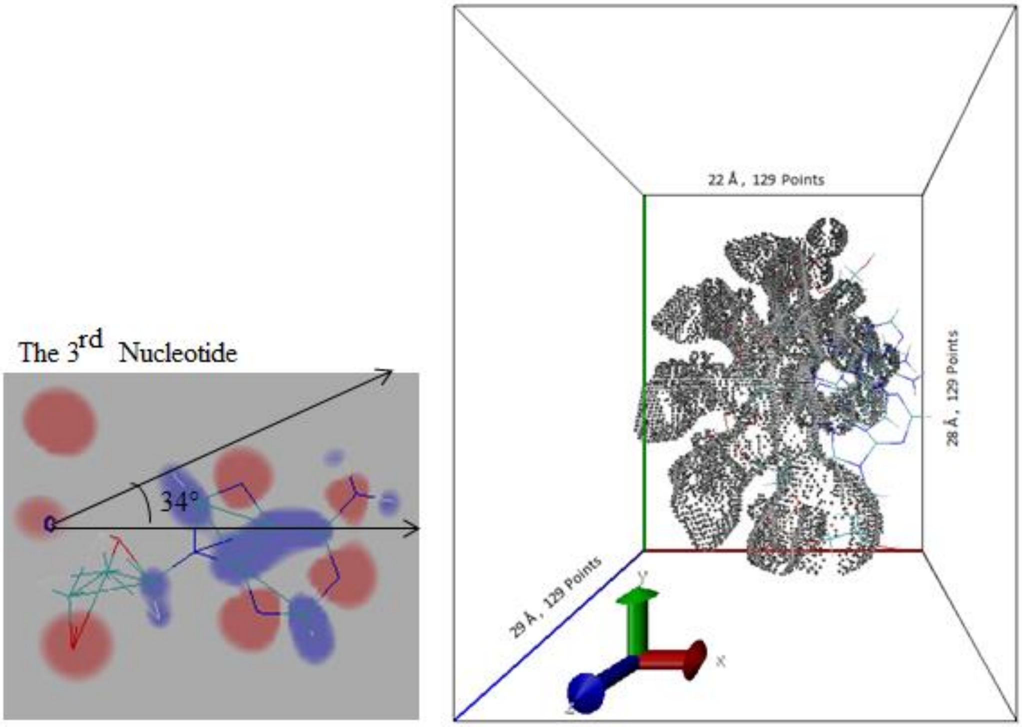


Figure 23. The Configuration of box for 5 long ssDNA. Left image is the measurement lines for every delta, while right image shows isosurface of a polymer, poly(dA)$_5$.

We analyzed MEP from the center of the sugar backbone of each nucleotide along the x-axis and along 33.69° in the anticlockwise direction from x-axis. The values are plotted in Fig. 24 and Fig. 25 for all four nucleotide bases. The variations of MEP were pronounced and monotonic along 33.69° in contrast to x-axis.

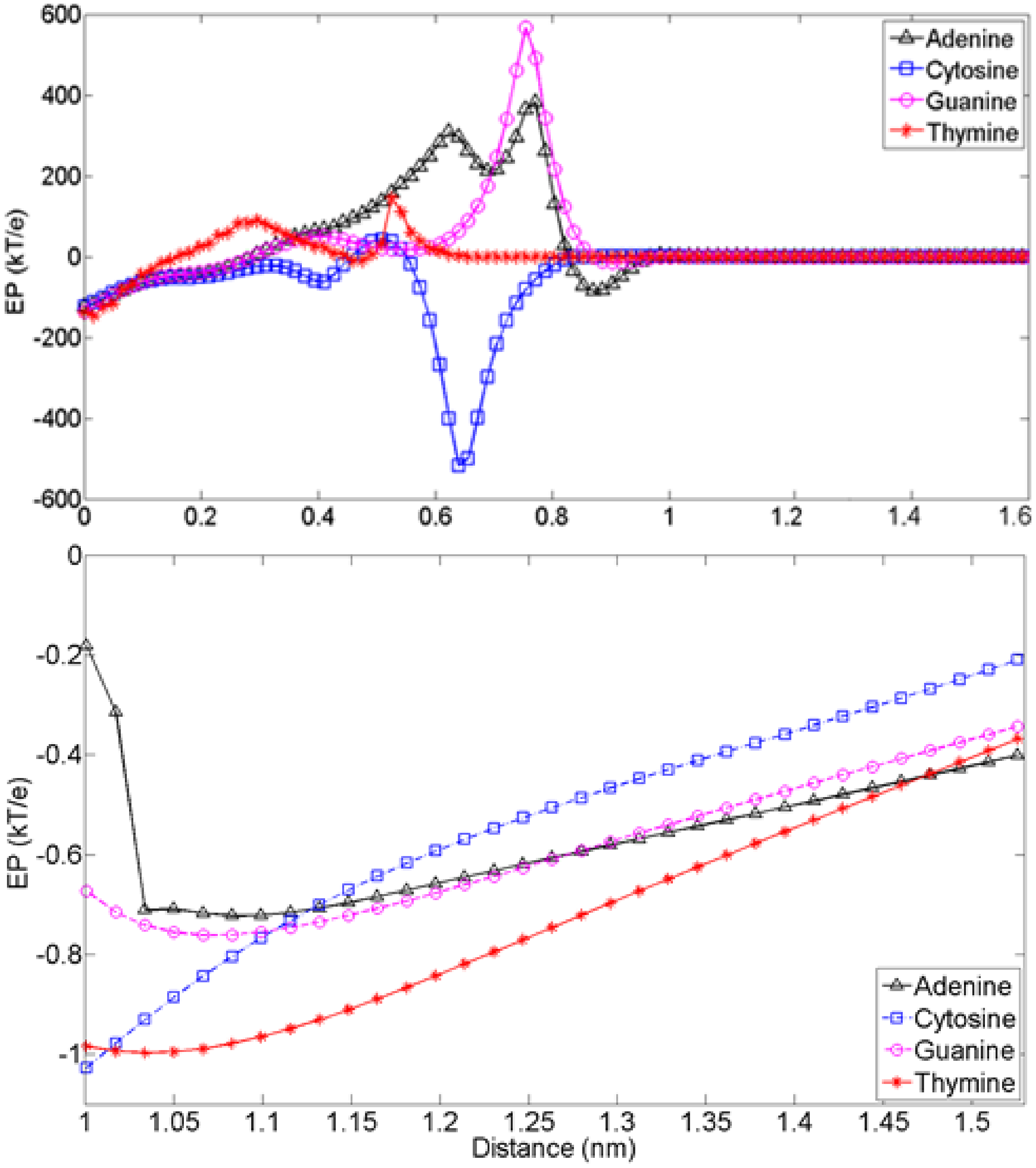


Figure 24. Electrostatic potential (EP) distributions with distance. EP distribution for different nucleotide bases along the x-axis.

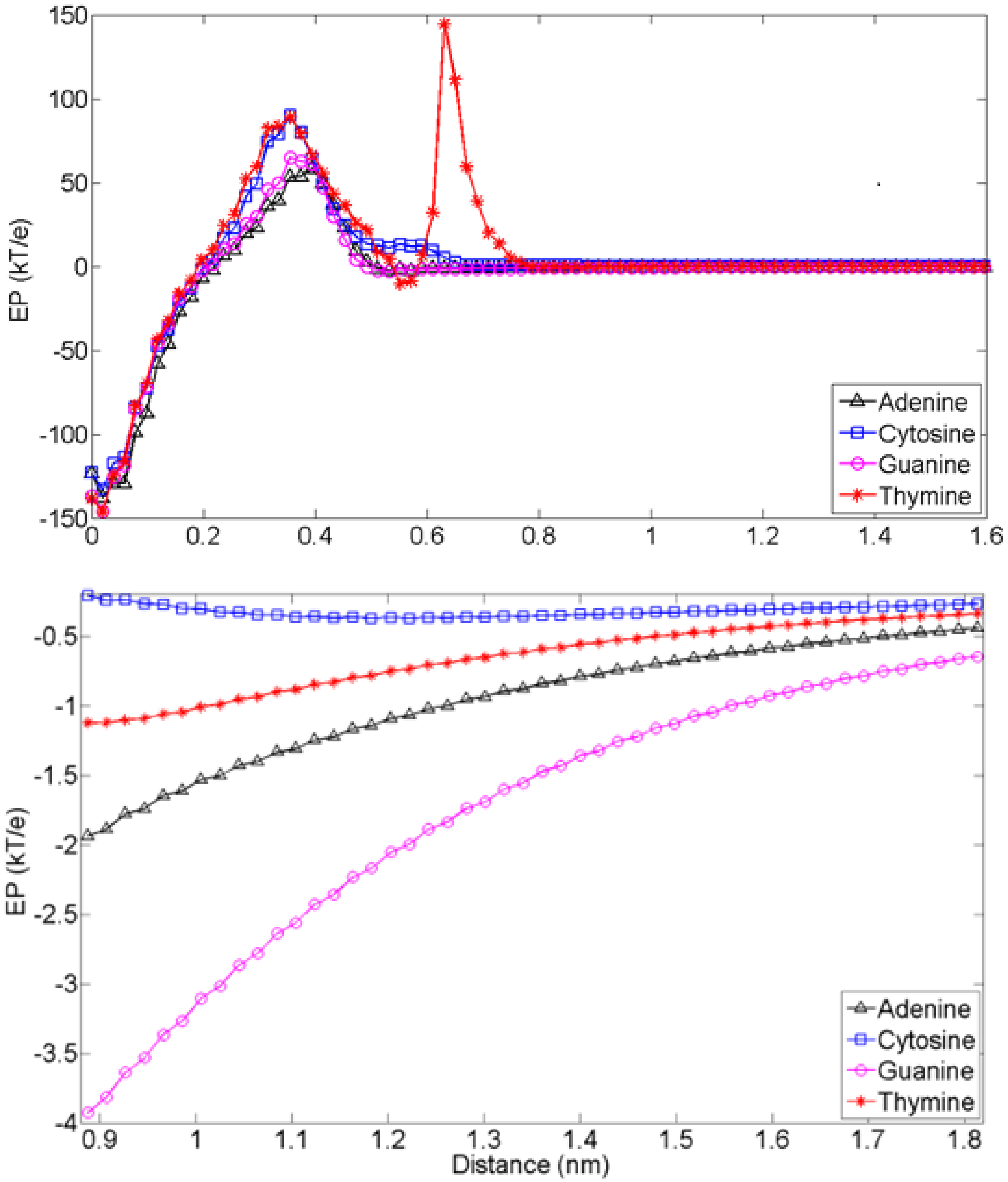


Figure 25. EP distribution for different nucleotide bases along 33.69° in the anticlockwise direction from x-axis. Bottom is after zooming around 1.4nm.This shows higher separations among EPs and monotonicity, with a sensitivity requirement of 0.01 kT/e (= ~258 µV) at 1.6 nm from the center of sugar-phosphate backbone.

C. ssDNA Movements Under an Applied Electric Field

In this simulation we observed the movement of ssDNA when applying a constant electric field. The length of sequences studied in this simulation is 40 base-pairs (bp) long, for each type of nucleotide. Each system was solvated by rectangular water box (TIP3 model) with lengths 60 x 60 x 180 Å, and the total atoms for each system was over 61,100. Then each system was minimized for 1,000 steps, heated for 10,000 steps, equilibrated for 100,00 steps, and simulated for 100,000 steps using NAMD with 1 fs per step using CHARMM27 force field. In the simulated section, we applied a constant electric field on the x-direction. Each polymer moved to the left with slightly different translocation rates, the order being: Thymine is the fastest, then Cytosine, then Adenine, while guanine is the slowest as shown in Fig. 26 and Fig. 27.

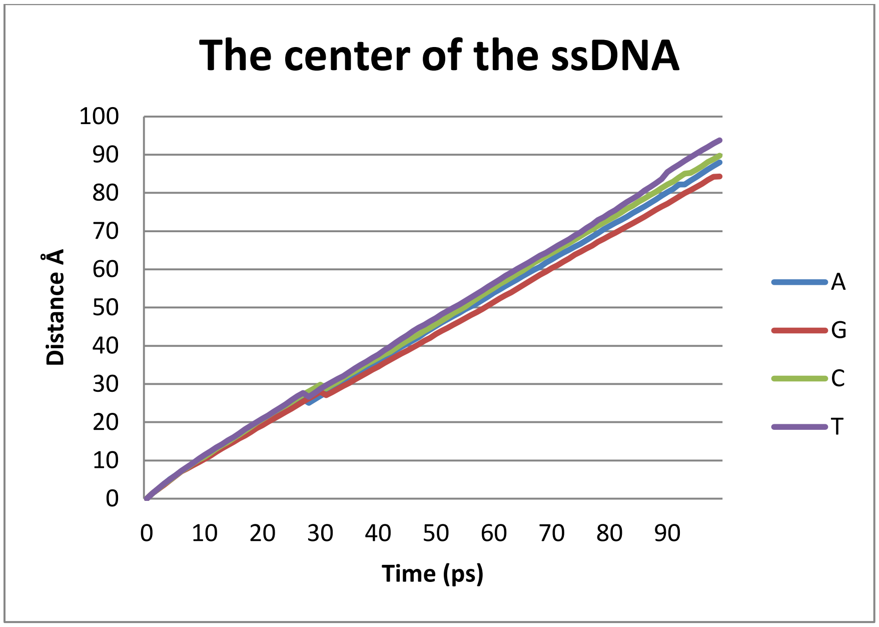


Figure 26. The movement of the center of each type of the DNA every 1 ps.

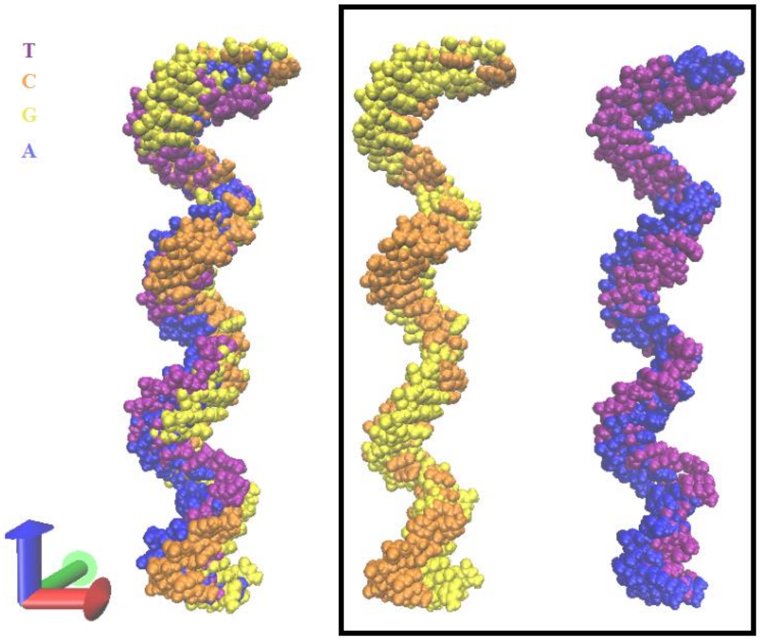


Figure 27. Snapshots captured during the movement of ssDNAs.

D. ssDNA with Carbon Nanotube Attached

In this simulation an armchair carbon nanotube (CNT) probe was built using a plugin in VMD of 5n length. The locations of atoms of the CNT were configured to be affixed at initial positions during all simulations. All of the steps in the previous simulations were applied here except the electric filed was along the z-axis and the carbon nanotube was placed under the ssDNA as shown in the Fig. 28. The simulation allows analyzing of the ssDNA while translocating near a CNT probe. The applied electric field strength was -2 *kcal/(mol Å e)* along x-direction and the translocation rate was 43.1 m/s. The distance from the tip of the CNT to the ssDNA was ~0.3nm.

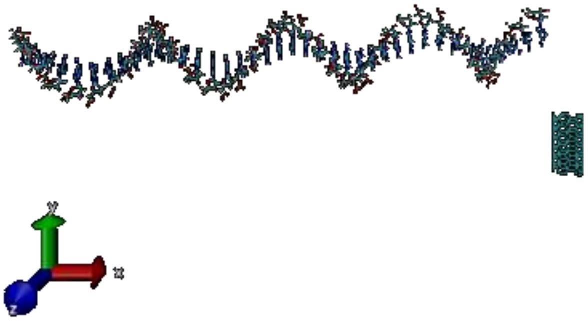


Figure 28. The ssDNA translocation alongside a CNT probe.

V. CONCLUSIONS

In this work, we have analyzed and studied the electrostatic distribution and MEP around ssDNA to demonstrate the feasiblity of direct indentification of neucleotide bases for single molecule DNA seuqneing. We show that a proximal probing location around 1.4 nm at ~34° might distinctly identify four types of nucleotide base polymers. However, the influence of at least 3rd neighbour must be taken into account, possibly using an adaptive or decision-directed sequencing algorithm. Furthermore, a simultanious probing of both ~34° and ~326° at proximities of ~1.4 nm would additionally improve the ability to uniquely identify each nucleotide base. We also analyzed the movement of 40 bases long ssDNA under a constant electric field and the results show that the movement is dependent primarily on the sizes of nucleotides.

Future research directions include studying the MEPs around ssDNA during heating and equilibration to explore the relative oscillations and rotations of the nucleotide bases within the DNA strand and the resultant sensitivity on the observed MEP at the probing point.

REFERENCES

[1] M. Mandelkern, J. Elias, D. Eden and D. Crothers, "The dimensions of DNA in solution.," *J Mol Biol*, vol. 152, no. 1, pp. 153-161, Oct. 1981

[2] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter, *Molecular Biology of the Cell*, 4th ed., Garland Science, New York, NY, 2002.

[3] F. Sanger, S. Nicklen and R. Coulson. "DNA sequencing with chain-terminating inhibitors," *Proc Natl Acad Sci*, vol. 74, no. 12, pp. 5463-5467, Dec. 1977.

[4] Smith L.M., Sanders J.Z., Kaiser R.J., Hughes P., Dodd C., Connell C.R., Heiner C., Kent S.B.H., Hood L.E. "Fluorescence detection in automated DNA sequence analysis," *Nature*, vol.321, no. 6071, pp. 674–679. Jun. 1986.

[5] Prober J.M., Trainor G.L., Dam R.J., Hobbs F.W., Robertson C.W., Zagursky R.J., Cocuzza A.J., Jensen M.A., Baumeister K. "A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides," *Science*, vol. 238, no. 4825, pp. 336–341, Oct. 1987.

[6] E. Mardis. "A decade's perspective on DNA sequencing technology," *Nature*,vol. 470, pp. 198-203, Feb. 2011.

[7] Collins, Francis. "Has the Revolution Arrived?," *Nature*, vol. 464, pp. 674-675, Apr. 2010.

[8] E. R. Mardis, "Anticipating the $1,000 genome," *Genome Biol*, vol.7, no. 7, pp. 112, Jul. 2006.

[9] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, et al."Real-time DNA sequencing from single polymerase molecules," *Science*, vol. 323, no. 5910, pp. 133-138. Jan. 2009.

[10] J. Korlach, K. P. Bjornson, B. P. Chauduri, R. L. Cicero, et al., "Real-time DNA sequencing from single polymerase molecules," *Methods Enzymol*, vol. 472, pp. 431-455 , 2010.

[11] J. J. Kasianowicz, E. Brandin, D. Branton and D. W. Deamer. "Characterization of individual polynucleotide molecules using a membrane channel," *Proc. Natl. Acad. Sci*. vol. 93,no. 24, pp. 13770–13773, Sep. 1996.

[12] Meller, A., Nivon, L., Brandin, E., Golovchenko, J. & Branton, D. “Rapid nanopore discrimination between single oligonucleotide molecules,” *Proc. Natl. Acad. Sci.* vol. 97, no. 3, pp. 1079–1084, Nov. 2000.

[13] Meller, A., Nivon, L. & Branton, D. “Voltage-driven DNA translocations through a nanopore,” *Phys. Rev. Lett*, vol. 86, no. 15, pp. 3435–3438, Apr. 2001.

[14] Deamer, D.W. & Branton, D. “Characterization of nucleic acids by nanopore analysis,” *Acc. Chem. Res*, vol. 35, no. 10, pp. 817–825, Sep. 2002.

[15] Nick Mitchell ,Stefan Howorka. “Chemical tags facilitate the sensing of individual dna strands with nanopores,” *Angewandte Chemie*, vol. 47, no. 30, pp. 5565-5568, Jun. 2008.

[16] Binquan Luan, Hongbo Peng, Stas Polonsky, Steve Rossnagel, Gustavo Stolovitzky, and Glenn Martyna, “Base-by-base ratcheting of single stranded DNA through a solid-state nanopore,” *Phys. Rev. Lett*, vol. 104, no. 23, pp. 238103, Jun. 2010.

[17] M. J. Holst, Ph.D. thesis, University of Illinois at Urbana-Champaign Urbana-Champaign, IL, USA, 1994.

[18] B. Jayaram, K. A. Sharp and B. Honig. “The electrostatic potential of B-DNA,” *Biopolymers*, vol. 28, no. 5, pp. 975–993, May. 1989.

[19] G. Caracciolo and R. Caminiti. “DNA–DNA electrostatic interactions within cationic lipid/DNA lamellar complexes,” *Phys. Lett.*, vol. 400, no.4, pp. 314–319, Dec. 2004.

[20] I. Otero-Navas and J. M. Seminario.”Molecular electrostatic potentials of DNA base–base pairing and mispairing” *J. Mol. Modeling*, vol. 18, no. 1, pp. 91-101, Jan. 2012.

[21] M. E. Gracheva, et al. “Simulation of the electric response of DNA translocation through a semiconductor nanopore–capacitor” *Nanotechnology*, vol. 17, no. 3, pp. 622–633, Jan. 2006.

[22] P. Mali and R. K. Lal, “The DNA SET: a novel device for single-molecule DNA sequencing,” *IEEE Trans Electron Devices*, vol. 51, no. 12, pp. 2004-2012, Dec. 2004.

[23] M. E. Gracheva, A. Aksimentiev and J. P. Leburton. “Electrical signatures of single-stranded DNA with single base mutations in a nanopore capacitor,” *Nanotechnology*, vol. 17,no.13, pp. 3160–3165, Jan. 2006.

[24] G. Sigalov, J. Comer, G. Timp and A. Aksimentiev. “Detection of DNA Sequences Using an Alternating Electric Field in a Nanopore Capacitor,” *Nano Lett.*, vol. 8, no.1, pp. 56–63, 2008.

[25] Allen, M., “Introduction to molecular dynamics simulation,”in *Computational Soft Matter-From Synthetic Polymers to Proteins,* vol. 23, Forschungszentrum Jülich, Germany: NIC, 2004, pp. 1–28.

[26] A.R. Leach.”Computer Simulation Methods,” in *Molecular Modelling. Principles and Applications*, 2nd ed. Harlow, England: Prentice-Hall, 2001, pp. 303-351.

[27] Jorgensen, W. L.,Tirado–Rives, J. “Monte Carlo vs Molecular Dynamics for Conformational Sampling,” *J Phys Chem*, , vol. 100, no. 34, pp. 14508-14513, Aug. 1996.

[28] Frenkel Daan, Smit Berend. “Molecular Dynamic Simulation,” in *Understanding Molecular Simulation: from algorithms to applications*. San Diego, California: Academic Press, 2001. Pp. 63-107.

[29] Loeffler, H. H., and M. D. Winn. “Large Biomolecular Simulation on HPC Platforms 1: Experiences with AMBER, Gromacs and NAMD,” *Science & Technology Facilities Council*, 2009.

[30] Loeffler, Hannes H., and Martyn D. Winna. "Large biomolecular simulation on HPC platforms II. DL POLY, Gromacs, LAMMPS and NAMD."

[31] Myung, Hun Joo, et al. "Accelerating molecular dynamics simulation using graphics processing unit," *Bulletin of the Korean Chemical Society,* vol. 31, no. 12, pp. 3639-3643, Oct. 2010.

[32] Takada, Shoji. "Coarse-grained molecular simulations of large biomolecules," *Current opinion in structural biology*, vol. 22, no. 2, pp. 130-137, Apr. 2012.

[33] A. Ghosh and M. Bansal, “A glossary of DNA structures from A to Z,” *Acta Crystallogr D*, vol. 59, no. 4, pp. 620–626, Apr. 2003.

[34] M. van Dijk and A. M. J. J. Bonvin, “3D-DART: a DNA structure modelling server,” *Nucl. Acids Res*., vol. 37, no. suppl 2 W235-W239, Apr. 2009.

[35] F. Fogolari, A. Brigo and H. Molinari. The Poisson–Boltzmann equation for biomolecular electrostatics: a tool for structural biology, *J. Mol. Recognit.*; vol. 15, no. 6, pp. 377–392, Dec. 2002.

APPENDIXES

APPENDIX A

TCL SCRIPT

1. The following script is used to separate the DNA to two chain, create ssDNA (psf and pdb), solvate the ssDNA , and add ions to the system.

```
#first run src.tcl in tcl/bin

#! c:/tcl/bin/tclsh

if { $argc != 2 } {
        puts "You have to enter the input file and the output prefex"
    } else {

    mol delete all

    package require psfgen
    resetpsf
    topology ../toppar/top_all27_na.rtf


    set in [lindex $argv 0]
    set out [lindex $argv 1]
    puts $in
    mol load pdb $in
    set a [atomselect top all]
    $a moveby "0 0 0"

# Seperate the DNA to two chains
    set temp [atomselect top "chain A"]
    $temp writepdb DNAchainA.pdb
    set temp [atomselect top "chain B"]
    $temp writepdb DNAchainB.pdb


# use the psfgen to guess the missing atom
# and create the psf for the dsDNA (from NAMD toturial)

foreach c {A B} {
    set sel [atomselect top "chain $c and name C1'"]
    set seg DNA$c
    segment $seg {
    first 5TER
    last 3TER
    pdb dsdna_$c.pdb
    }
    foreach resid [$sel get resid] resname [$sel get resname] {
    if { $resname eq "THY" || $resname eq "CYT" } {
        patch DEO1 $seg:$resid
    } elseif { $resname eq "ADE" || $resname eq "GUA" } {
        patch DEO2 $seg:$resid
    }
    }
    coordpdb dsdna_$c.pdb $seg
}
guesscoord
writepsf dsDNA.psf
writepdb dsDNA.pdb

#creating the psf and pdb files for the ssDNA
resetpsf
readpsf dsdna.psf
coordpdb dsdna.pdb
delatom DNAB
writepsf $out.psf
writepdb $out.pdb

#solvation

mol delete all
package require solvate
solvate $out.psf $out.pdb -minmax {{-40 -40 -70} {40 40 70}} -o
solvated$out

#ionization

autoionize -psf solvated$out.psf -pdb solvated$out.pdb -sc 0.025 -o
ionized$out

}
```

2. The following scrip is used to extract the frame from the simulation trajectory, convert all the extracted files to PQR and calculate the MEP.

```
set t1 [clock clicks -milliseconds ]
#--------------------------Extracting the pdb files
set NoFrames 311
set minFrames 100


puts "Extracting the pdb files"
file mkdir Results
set f [open "Results/centers.txt" w]
set x 0
while {$x <$NoFrames}  {

animate goto $x
#puts $f $x
set a [atomselect top "backbonetype nucleicback and resid 13"]
puts $f [measure center $a]
```

```
set n [atomselect top "nucleic"]
$n writepdb "Results/$x 1.pdb"
if {$x < $minFrames} {incr x 10} else {incr x 1}
after 10
}
close $f

cd Results
#-------------------------------Converting to pqr
set x 0
puts "Converting to pqr"
while {$x <$NoFrames}  {
set in [open "$x 1.pdb" r]
set out [open "$x.pdb" w]
while {[eof $in] !=1 } {
     gets $in line

     set l [string map {ADE DA CYT DC THY DT GUA DG} $line]
     puts $out $l
     }
close $in
close $out

file delete -force "$x 1.pdb"
exec python G:/MD/Softwares/pdb2pqr-1.8/pdb2pqr.py --ff=CHARMM "$x.pdb"
"$x.pqr"
if {$x < $minFrames} {incr x 10} else {incr x 1}
after 10
}
#----------------------------------Calculating Apbs
set x 0
puts "Calculating Apbs"
while {$x <$NoFrames} {

set f [open "$x.in" w]
puts $f "read
  mol pqr $x.pqr
end
elec
  mg-auto
  dime 161 161 161
  cglen 132 132 132
  cgcent mol 1
  fglen 132 132 132
  fgcent mol 1
  mol 1
  lpbe
  bcfl sdh
  srfm smol
  chgm spl2
  ion 1 0.025 2.0
  ion -1 0.025 2.0
  pdie  1.0
  sdie  78.54
  sdens  10.0
  srad  1.4
  swin  0.3
  temp  298.15
  gamma  0.105
  calcenergy no
  calcforce no
  write pot dx $x
end
quit
"
close $f

exec apbs --output-file=$x.out $x.in &
after 10000
if {$x < $minFrames} {incr x 10} else {incr x 1}
}
 set t2 [clock clicks -milliseconds ]

 puts "\nTime taken [expr $t2 - $t1] milliseconds"
cd ..
```

APPENDIX B

MATLAB CODE

- The following code used to analyze the MEP data around the nucleotide for different radii.

```
%Mohammed Abu Saude
%Plot circle

clear all;
close all;
clc
nFrames=11;
seq=['As9AAs';'Cs1CCs';'Gs1GGs';'Ts1TTs'];
anglestep=1;


radius=[10,  14, 20, 30];
results=zeros(4,360/anglestep,length(radius));
for l=1:1:length(radius)


for D=1:1:4

PATH=['g:\MD\25\',seq(D,:),'\Results\'];



mPoints=zeros(nFrames,3);
fCenters = fopen([PATH,'centers.txt'], 'r');
```

```
Centers = fscanf(fCenters, '%f',nFrames*3  );
Centers=reshape(Centers,3,nFrames);
Centers=Centers';
fclose(fCenters);
%----------------------------------------

for j=107:10:107
    disp(['Analayzing ',num2str(j),'.dx']);
    fid = fopen([PATH,num2str(j),'.dx'], 'r');
    delta=zeros(3,1);
    while(1)
        temp = fscanf(fid, '%s',1  );
        if (strcmp(temp,'origin'))
            origin=fscanf(fid, '%f',3  );
            origin=origin';
        end
        if (strcmp(temp,'delta'))
            delta=[delta fscanf(fid, '%f',3  )];
        end
        if (strcmp(temp,'counts'))
            dim=fscanf(fid, '%f',3  );
        end
        if (strcmp(temp,'follows'))
            data=fscanf(fid, '%f',prod(dim));
            break;
        end
    end
    delta(:,1)=[];
    delta=diag(delta);
    delta=delta';
    fclose(fid);
    data=data';

    data=reshape(data,161,161,161);  %reshape to 3D matrix (z,y,x)
%---------------------------------------------------------------
    cAtoBox=Centers(11,:)- origin;
    locOfCenter=(cAtoBox./delta); % Location of the center acourding to pot indices
    an=0;
    for c=1:1:360/anglestep
    angle=an*pi/180;

    mPoints(c,:)=round([cAtoBox(1)+radius(l)*cos(angle) cAtoBox(2)+radius(l)*sin(angle)
cAtoBox(3)]./delta);
    results(D,c,l)=data(mPoints(c,3),mPoints(c,2),mPoints(c,1))
    an=an+anglestep;
    end
end

end
end

for l=1:1:length(radius)
figure(l);
t=0:anglestep:359.5;
%old plot(t,results','*-');%,'Linewidth',1,'MarkerSize',11);
```

```
plot(t,results(1,:,l)','r*','MarkerSize',10,'MarkerFaceColor','r');
hold on
plot(t,results(2,:,l)','bs' ,'MarkerSize',10,'MarkerFaceColor','b');
plot(t,results(3,:,l)','m^','MarkerSize',10,'MarkerFaceColor','m');
plot(t,results(4,:,l)','go','MarkerSize',10,'MarkerFaceColor','g');
xlabel('Angle (deg)','FontSize',19);
ylabel('EP (kT/e) ','FontSize',19);
%legend(seq(1,:),seq(2,:),seq(3,:),seq(4,:));
legend('A','C','G','T',4);
title(['EP at r=', num2str(radius(l)/10),'nm'],'FontSize',19);
set(gca,'Fontsize',19);
set(gca,'XTick',[0:45:360])
%if (radius(l)>10)
axis([0 360 -40 100])
%end
end
```

- The following code used to analyze the MEP data along horizontal line and the line with 33.7 angle.

```
%Mohammed Abu Saude
%Plot EP A

clear all;
close all;
A=xlsread('pot.xlsx','A1:C715563'); % Read all data
A=A';
A=reshape(A,1,2146689);     %reshape the matrix to 1D array
A=reshape(A,129,129,129);   %reshape to 3D matrix (z,y,x)

A(66,104,106)
%chosing the plane
B=A(62,:,:);

B=reshape(B,129,129);



% C=diag(B);
% plot(C);
y=69;
step=2;
j=1;
for x=35:step:128

    D(j)=B(y,x);

    %for the middle pint
    j=j+1;
    D(j)=(B(y,x+1)+B(y+1,x))/2
    %---
```

```
     y=y+1;
     j=j+1;


end
xlswrite('Epot3.xlsx',D); %Epot3 for data with adding to the center of
the
%backbone
 plot(D(61:94),'rx');
figure (2);
plot(D,'rx');


%Mohammed Abu Saude
%Plot EP One Base 33.7 All

clear all;
close all;

A=xlsread('OneBaseEPall33.xlsx','A1:CL1');
C=xlsread('OneBaseEPall33.xlsx','A2:CH2');
G=xlsread('OneBaseEPall33.xlsx','A3:CJ3');
T=xlsread('OneBaseEPall33.xlsx','A4:CD4');



I=40:81;
I=I*0.19717858537693691446745741306479e-1;%nano

plot(I,A(41:82),'-k^','LineWidth',2,'MarkerSize',12);
hold on
plot(I,C(41:82),'-bs','LineWidth',2,'MarkerSize',12);
plot(I,G(41:82),'-mo','LineWidth',2,'MarkerSize',12);
plot(I,T(41:82),'-r*','LineWidth',2,'MarkerSize',12);
lgd1=legend('Adenine','Cytosine','Guanine','Thymine',4);
set(lgd1,'FontSize',28);
xlabel('Distance (nm)','FontSize',28);
ylabel('EP (kT/e) ','FontSize',28);
set(gca,'Fontsize',28);
axis([0.78 1.6 -1.2 0.62])

figure(3);
semilogy(I,A(41:82),'-k^','LineWidth',2,'MarkerSize',12);
hold on
 semilogy(I,C(41:82),'-bs','LineWidth',2,'MarkerSize',12);
semilogy(I,G(41:82),'-mo','LineWidth',2,'MarkerSize',12);
semilogy(I,T(41:82),'-r*','LineWidth',2,'MarkerSize',12);
lgd1=legend('Adenine','Cytosine','Guanine','Thymine',4);
set(lgd1,'FontSize',28);
xlabel('Distance (nm)','FontSize',28);
ylabel('EP (kT/e) ','FontSize',28);
set(gca,'Fontsize',28);
axis([0.88 1.82 -10 0])

J=0:81;
J=J*0.19717858537693691446745741306479e-1;%nano
```

```
figure(2);
plot(J,A(1:82),'-k^','LineWidth',2,'MarkerSize',11);
hold on
plot(J,C(1:82),'-bs','LineWidth',2,'MarkerSize',11);
plot(J,G(1:82),'-mo','LineWidth',2,'MarkerSize',11);
plot(J,T(1:82),'-r*','LineWidth',2,'MarkerSize',11);
legend('Adenine','Cytosine','Guanine','Thymine',4);
xlabel('Distance (nm)','FontSize',20);
ylabel('EP (kT/e) ','FontSize',20);
set(gca,'Fontsize',20);
```