2003

Computational Analysis of the U2 Snrna-Intron Duplex

Darui Xu
THE FLORIDA STATE UNIVERSITY

COLLEGE OF ARTS AND SCIENCE

COMPUTATIONAL ANALYSIS OF THE U2 SNRNA-INTRON DUPLEX

By

DARUI XU

A Thesis submitted to the
Department of Chemistry and Biochemistry
in partial fulfillment of the
requirements for the degree of
Master of Science

Degree Awarded:
Fall Semester, 2003
The members of the Committee approve the Thesis of Darui Xu defended on November 21, 2003.

Nancy L. Greenbaum
Professor Directing Thesis

Hong Li
Committee Member

Igor Alabugin
Committee Member

The Office of Graduate Studies has verified and approved the above named committee members.
To my family who are behind me no matter what
ACKNOWLEDGEMENTS

I would like to thank Dr. Marcia Fenley, who deserves to be my co-major professor. But I am forbidden to do so because of administrative process. She introduced me into this exciting world of computational biochemistry. Together with Dr. Nancy Greenbaum, they have been supporting me academically, financially and emotionally. I want to thank Nancy especially for her understanding and encouragement. Without her, I never would have finished my Master in Biochemistry. I thank Dr. Maria Nagan at Truman State University. She provided the partial charge parameters for pseudouridine. She also traveled to FSU to help me get started with Molecular Dynamics simulation although it was not the most convenient time for her. She is still guiding me in the analysis of the trajectories.

I thank Drs. Hong Li and Igor Alabugin for serving on my committee. I would also like to thank the faculty and my fellow graduate students in both chemistry and computer science departments. They have taught me a lot directly or indirectly.

I thank Ms. Mimi Burbank in the School of Computational Science and Information Technology for providing the Latex template files and assistant, which makes writing this thesis less painful. I thank Smita Bhatia for her help with Pymol scripts and making figures.
# TABLE OF CONTENTS

List of Tables ........................................................................ vii
List of Figures ......................................................................... viii
Abstract .................................................................................. ix

1. INTRODUCTION ................................................................. 1
   1.1 The spliceosome and the U2-intron duplex ....................... 1
   1.2 Molecular Dynamics of RNA ...................................... 4
   1.3 Electrostatics of RNA ............................................. 7
   1.4 Surface Area and Free Energy Calculations .................. 10

2. METHODS ........................................................................... 14
   2.1 Structures of $\psi$BP and uBP ..................................... 14
   2.2 MD simulations ..................................................... 15
      2.2.1 System set-up ............................................ 15
      2.2.2 System equilibration and production ................. 15
      2.2.3 Analysis of the trajectories ............................. 17
   2.3 Numerical solution to the nonlinear PBE ...................... 17
   2.4 Visualization of the electrostatic surface potential .......... 18
   2.5 Calculation of solvent accessible surface area (SASA) .... 19
   2.6 Calculation of solvation energies ........................... 19

3. RESULTS AND DISCUSSIONS ............................................ 21
   3.1 Preliminary results of MD simulation ........................... 21
   3.2 Surface electrostatic potentials .................................. 23
      of the RNA duplexes ........................................ 23
      3.2.1 Surface electrostatic potentials using DelPhi ...... 24
      3.2.2 Surface electrostatic potentials using hybrid PBE approach .... 25
   3.3 Surface Areas and solvation free energy .................... 31

4. CONCLUSIONS .................................................................. 36

APPENDIX A: Perl scripts developed ................................. 37

REFERENCES ....................................................................... 45
LIST OF TABLES

3.1 Surface area calculations .......................................................... 33
3.2 Solvation free energy of $\psi$BP and uBP ...................................... 33
LIST OF FIGURES

1.1 Spliceosome Cycle. ......................................................... 2
1.2 Branch site sequences. .................................................. 4
1.3 A schematic view of the \( \psi \)BP (left) and uBP(right) duplexes. .............. 5
1.4 Definition of Molecular Surface Area .................................... 11
3.1 Time evolution of RMSD of the simulated structure (based on the all-atom model) with respect to the starting NMR structure .................. 22
3.2 Time evolution of RMSD of the simulated structure (based on the all-atom model) with respect to the average structure .................... 23
3.3 Time evolution of interphosphate distance (P4-P26) of the simulated structure 24
3.4 Surface electrostatic potentials of uBP calculated using DelPhi ............... 25
3.5 Surface electrostatic potentials of \( \psi \)BP calculated using DelPhi .......... 26
3.6 Surface electrostatic potential of an ideal A-RNA (major groove view) ......... 27
3.7 Surface electrostatic potential of an ideal A-RNA (minor groove view) ........ 28
3.8 Surface electrostatic potential map of uBP with Cornell \textit{et al.} force field parameters ......................................................... 29
3.9 Surface electrostatic potential of \( \psi \)BP with Cornell \textit{et al.} force field parameters 30
3.10 Surface electrostatic potential of \( \psi \)BP with Cornell \textit{et al.} force field combined with the structure of \( \psi \)BP ....................................................... 31
3.11 Surface electrostatic potential of uBP calculated with the Discover force field . 32
3.12 Surface electrostatic potential of \( \psi \)BP calculated with the Discover force field . 33
3.13 Surface electrostatic potential of \( \psi \)BP with only a -1e charge on the phosphate groups ......................................................... 34
3.14 Surface electrostatic potential of uBP with only a -1e charge on the phosphate groups ......................................................... 35
ABSTRACT

Pairing of a consensus sequence of the precursor (pre)-mRNA intron with a short region of the U2 small nuclear (sn)RNA during assembly of the eukaryotic spliceosome results in formation of a complementary helix of seven base pairs with a single unpaired adenosine, whose 2′ OH initiates the nucleophilic attack at the pre-mRNA 5′ splice site during the first step of splicing. The structure of the spliceosomal branch site solved by Newby and Greenbaum [4] showed that a highly conserved pseudouridine (ψ) residue in U2 snRNA induces a dramatically altered structure compared with that of its unmodified counterpart.

In this study, both modified and unmodified U2 snRNA-intron duplexes were analyzed using computer simulations including preliminary molecular dynamics (MD) simulations, electrostatic potential, surface area, and solvation free energy calculations. The preliminary MD simulations produce stable trajectories of the RNA duplexes in solution. The surface electrostatic potentials were calculated using finite difference Poisson-Boltzmann algorithm and a hybrid boundary element and finite difference Poisson-Boltzmann approach. Results show a region of exceptionally negative potential near the 2′OH of the branch site adenosine. The two RNA duplexes have similar solvent accessible surface areas, whereas the surface accessible area of the 2′OH of the branch site adenosine of the modified RNA duplex is considerably smaller than that of the unmodified RNA duplex. The solvation free energy calculation indicates that the unmodified RNA duplex is favored over the modified RNA duplex.
CHAPTER 1

INTRODUCTION

1.1 The spliceosome and the U2-intron duplex

Splicing is a part of the maturation process of precursor messenger (pre-m)RNA molecules. During splicing, the noncoding regions (introns) are excised from nascent mRNA transcripts and the flanking coding regions (exons) are ligated together. In eukaryotes, the splicing process is catalyzed by the spliceosome. The spliceosome is a dynamic complex of five small nuclear (sn)RNAs including U1, U2, U4/U6, and U5, which associate with numerous proteins forming the small nuclear ribonucleoproteins called snRNP (reviewed by Moore et.al [1]).

The spliceosome goes through a cycle during which snRNP components first assemble around the pre-mRNA, then the introns are removed and finally the spliceosome dissociates to an unknown extent and the snRNPs are recycled for the next round of splicing as shown in Figure 1.1. The earliest event in assembly is that U1 joins with the pre-mRNA through the base-pairing between the 5′ end of the U1 snRNA and the 5′ splice site consensus to form the commitment complex (CC). Subsequent to formation of the commitment complex, U1 snRNP promotes the binding of U2 snRNP to form complex A. A consensus sequence (GUAGUA) within U2 snRNA binds to the highly conserved branch site sequence in pre-mRNA. This presumably bulges out an adenosine residue (called the branch site adenosine) within the intron, generally located within 100 nucleotides of the 3′ splice site. The U4/U6.U5 then enters the spliceosome as a tri-snRNP complex to form complex B1. Shortly after entering the spliceosome, the extended RNA duplex between U4 and U6 snRNAs is dissociated, allowing U6 snRNA to base-pair with U2 snRNA. This forms the B2 complex. This U2/U6 pairing, together with the U2/branch site and U6/5′ splice site
interaction positions the 5′ splice site and the branch site for the initial catalytic reaction occurred in C1 complex. After the formation of complex C2 containing the products, the spliced exons from the spliceosome are released. The spliceosome is then dissolved and the snRNPs are released from the spliceosome for the next cycle. The half lives of the individual snRNAs are about 20 hours.

Splicing occurs in two steps as sequential transesterification reactions. The first step entails cleavage at the 5′ splice site to yield a 5′ exon intermediate with a free 3′ OH terminus. Concurrently, the 5′ end of the intron is joined via a 2′-5′ phosphodiester bond with the branch site adenosine forming the so-called lariat intermediate. This involves direct nucleophilic attack by the 2′OH of the branch site adenosine on the phosphate at the 5′ splice site. The intermediates formed are subsequently resolved by cleavage at the 3′ splice site and ligation of the two exons via a 3′-5′ phosphodiester bond. The pre-mRNA becomes two
parts: the mRNA and the excised intron in the lariat form. Both steps in splicing proceed through direct in-line $S_N2$ displacement mechanisms. ATP hydrolysis is required for intron excision, although none of the phosphates in the spliced products is derived from the ATPs.

Since there is obvious similarity between the intermediates and products of group II self-splicing introns and the spliceosome, the spliceosome has been proposed to be a ribozyme, i.e. the transfesterification reactions are actually RNA-catalyzed. Phylogenetic, mutational and biochemical studies have demonstrated the role of specific and highly conserved RNA sequences in assembly of the catalytic core [2]. The recent discovery of splicing activity by protein-free snRNAs further supports the proposal [3]. Therefore, much effort has been devoted to characterizing the networks of RNA:RNA interactions within the spliceosome. Among the critical RNA:RNA interactions is that of the U2 snRNA with the branch site intron. Understanding how the branch site is recognized by other components of spliceosome and how its 2$'$OH is positioned for the nucleophilic attack can help us understand the mechanism of splicing and provide insights into molecular recognition and structure-function relationships.

The structure of this U2 snRNA-intron branch site helix has been recently solved by Newby and Greenbaum [4] using high resolution nuclear magnetic resonance (NMR) methods. In their study, the structural role of yeast U2 snRNA $\psi$35, a pseudouridine ($\psi$) that pairs with the intron branch site directly opposite the A-A dinucleotide was investigated by exploring the architectural features of the pseudouridine modified ($\psi$BP) and unmodified (uBP) spliceosomal U2 snRNA-intron branch-site pairing (see Figure 1.2). $\psi$ residue is a post-transcriptional modification of uridine (U). The enzymatically isomerized residue differs from the canonical U only by covalent linkage of the ribose through the C5 instead of the N1 position and by an additional protonated ring nitrogen (see Figure 1.2a). $\psi$ residue has been identified in U2 snRNA of all eukaryotes. The Newby and Greenbaum study showed that the presence of the pseudouridine induces a dramatic change in the backbone of the intron strand. The structure of the unmodified U2-intron duplex adapts a continuous A-type helical RNA geometry. The branch site adenosine is stacked in the helix. The $\psi$ modified duplex, on the other hand, has a pronounced kink in the branch site region, which results in the branch site adenosine being extruded from the helix (see Figure 1.3). This extra-helical position of branch site adenosine may explain how the 2$'$OH is recognized and positioned for the first
Figure 1.2. Branch site sequences.
(a), Schematic structure of uridine (U) and pseudouridine (ψ) bases. As compared with U, ψ has an additional hydrogen bond donor (NH1) and is linked to its ribose by a carbon-carbon bond. (b), Sequences of unmodified (uBP) and ψ-modified (ψBP) duplexes representing the branch site topology. In each case, the upper strand corresponds to the U2 snRNA sequence, and the bottom strand to the intron. Residues shown in gray were added to the native sequences for stability. Solid circles indicate Watson-Crick base pairs documented by nuclear overhauser effect between exchangeable protons. Open circles indicate Watson-Crick base pairs inferred from stacking patterns involving non-exchangeable protons. Short lines in uBP indicate hydrogen bonding between U6 and both A23 and A24 suggested by structure calculation. [4]

step of splicing. In this thesis, preliminary studies of MD simulations were first carried out on the two RNA duplexes. The results showed that the models obtained from NMR study are in the globally energy-minimized form. Then electrostatic potential features of both RNA duplex were analyzed. The solvent accessible surface area and solvation free energies of the two duplexes were calculated. Results of these calculations will help us understand the interactions stabilizing ψBP structure and how the 2'OH of the branch site adenosine is recognized for the splicing reaction.

1.2 Molecular Dynamics of RNA

An understanding of the functional mechanism of a biological macromolecule requires knowledge not only of its precise molecular organization but also of its internal dynamics.

Molecular dynamics (MD) simulation is a powerful computational approach to study macromolecular structures and motions. In essence, MD simulation is a computer experiment
Figure 1.3. A schematic view of the \( \psi \text{BP} \) (left) and \( \text{uBP} \) (right) duplexes.

RNA duplexes were rendered using the DINO [38] program. The base C is in red, G in orange, U in blue and A in cyan. The branch site adenosine in \( \psi \text{BP} \) is extruding out to the major groove as compared with the \( \text{uBP} \).

in which the atoms of a stipulated molecular system undergo Newtonian dynamics on an assumed potential energy surface.

The simulation begins with an initial configuration, usually taken from a high resolution experimentally-determined molecular structure and an arbitrary arrangement of solvent molecules and mobile ions. The initial configuration first goes through energy minimization to relieve the major constraints in the system. During the MD stage, velocities on the atoms of the system are first increased. The system then seeks out a thermally bounded state, and proceeds to the production stage where the structural changes that occur in the system as a function of time are sampled. The position of each atom in the system over time is called the MD trajectory. The trajectory is obtained by solving the differential equations embodied in Newton’s second law, which describes the motion of a particle of mass \( m_i \) along
one coordinate $x_i$ with $F_{x_i}$ being the force on the particle in that direction:

$$\frac{d^2x_i}{dt^2} = \frac{F_{x_i}}{m_i} \quad (1.1)$$

The collected MD trajectories are first analyzed to determine the extent to which the calculated properties agree with corresponding experimental results. When sufficient agreement is established, the molecular details obtained can be used to interpret the experimental results, and to provide unique insights into the structure and dynamics of macromolecules that cannot be obtained by other computational and experimental techniques.

Before 1995, simulations of nucleic acids were plagued by instabilities, characterized by distortion of duplex structures, broken base pairing and misrepresentation of sequence-specific fine structure. This problem was due to a number of factors, such as simple truncation of long-range electrostatic interactions for highly charged systems like nucleic acids; the limit of computer power, and lack of reliable empirical potential functions (or “force fields”).

Since 1995, there has been significant advance in MD simulation of nucleic acids. There are three major factors which contribute to this progress (reviewed in [28]). First, the development of the fast Ewald methods such as Particle Mesh Ewald (PME)[29] allows a full Ewald treatment (which treat the system as an infinite crystal or periodic lattice to provide a complete treatment of the electrostatic interaction) with an efficient algorithm. Second, in the last decade, there has been enormous progress in development of computer hardware and software. The actual performance of computers increased by at least three orders of magnitude. Third, the appearance of the “second-generation” force fields are much improved in their representation of nucleic acid structure. For example, in the AMBER program [31], the potential function $U$ is as follows:

$$U = \sum_{bonds} k_b (r - r_{eq})^2 + \sum_{angles} k_{\theta} (\theta - \theta_{eq})^2 + \sum_{dihedrals} \sum_n \frac{V_n}{2} \times [1 + \cos(n\phi - \gamma)]$$

$$+ \sum_i \sum_{j>i} \left[ \left( \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6} \right) + \frac{q_i q_j}{\epsilon \cdot r_{ij}} \right]$$

where $k_b$ and $K_{\theta}$ are force constants, $r_{eq}$ and $\theta_{eq}$ are equilibrium bond length and angle, $V_n$ is the maximum potential energy, $\gamma$ is the phase angle, $n$ is the periodicity of the Fourier term, $A_{ij}$, $B_{ij}$ are the van der Waals parameters, $q_i$, $q_j$ are the partial charges and $\epsilon$ is
the dielectric constant. This molecular mechanics approach attempts to derive an effective potential for nuclear motion by fitting the above equation empirically and then using the calculated potential energy to solve the classical equations of motion for the system. The charges required for evaluation of the electrostatic term are derived by fitting the electrostatic potential (ESP) or restrained ESP and numerically adjusting the charges to reproduce a set of \textit{ab initio}-calculated interaction energies or liquid and solid properties. One example of such force fields is the Cornell \textit{et al.} [32] force field, which is routinely used in MD simulations.

With the use of PME to treat long-range electrostatics and the explicit solvent treatment, stable trajectories of nucleic acids can be obtained up to several nanoseconds. Details of the structural and dynamic properties such as nucleic acid hydration and ion association and energetics can be elucidated through MD studies. Recently, a number of studies were carried out using MD to analyze the stability, metal ion binding and hydration of RNA molecules. For example, the MD simulation of 5S rRNA loop E showed water bridges which complement non-Watson-Crick base pairing [33].

### 1.3 Electrostatics of RNA

The tertiary structures of nucleic acids are determined by the balance among a variety of forces: (1) electrostatic forces due to the negatively charged phosphates; (2) stacking interactions between bases due to hydrophobic and dispersion forces as well as to hydrogen bonding between bases; and (3) steric or repulsive interactions that control the conformational flexibility of the phosphate-sugar backbones. Electrostatic effects are among the most important factors in determining the conformations of biomolecules and their associations to other charged molecules. Classical electrostatics provides intuitive insights that help us to reason about the properties of molecules. Because of their strength and their long-range nature, electrostatic interactions play a central role in a variety of biological processes. The ability to model the electrostatic properties of nucleic acids accurately is particularly critical. Since nucleic acids are highly charged, the structure-function relationships and molecular recognition events involving nucleic acids are highly influenced by the electrostatic interactions. The role of electrostatics in molecular interactions of proteins has been well established in a number of studies [6]. These studies revealed that the surfaces of proteins
are covered with distinct patches of electrostatic potential which play important roles in molecular recognition. Unlike proteins, since nucleic acids contain mostly distributed negative charges, it is difficult to envision that there exists distinct regions of electrostatic potential on RNA surfaces, which may contribute to the interactions of nucleic acids. Recently, some calculations were carried out to examine the electrostatic potentials of RNA molecules [5], which show that there are also regions of electrostatic heterogenity on RNA surfaces corresponding to active sites.

Electrostatics of a biomolecule in aqueous solution can be studied using microscopic or macroscopic approaches. Ideally, a microscopic approach with explicit solvent gives a complete picture of the whole system at atomic level of detail. The macroscopic approach involves continuum models, in which solvent and mobile ions are treated as a continuum environment, characterized by the solvent dielectric constant and the salt concentration, while the solute is treated explicitly. Most continuum models use at least two dielectric constants to represent the system, as most solutes have lower dielectric constant than that of water. In practice, a microscopic approach is time consuming and computationally demanding. On the other hand, a continuum approach is much faster and can produce electrostatic results that are in exceptionally good agreement with experimental results [20]. Therefore, continuum approach is widely used to compute electrostatic properties such as free energies, pKa shifts, etc..

The fundamental problem of classical electrostatics is to find the electrostatic potential at every point in space for a given distribution of solute charges. The electrostatic potential $\phi$ at position $x$ can be described by the Poisson equation in terms of the fixed charge density of the solute $\rho_{\text{solute}}(x)$ and position dependent dielectric function $\epsilon(x)$:

$$\nabla \cdot [\epsilon(x)\nabla \phi(x)] = -4\pi \rho_{\text{solute}}(x)$$  \hspace{1cm} (1.2)

where $\nabla$ is the divergence operator. When there are mobile ions present in the solvent, the charge density can be described by the Boltzmann distribution as:

$$\rho(x) = \sum_{i=1}^{n_{i}} q_{i}c_{i}\exp[-\bar{E}_{i}(x)/k_{B}T]$$  \hspace{1cm} (1.3)

where $c_{i}$ is the bulk concentration of the ionic species $i$, $k_{B}$ is the Boltzmann constant, $T$ is the absolute temperature and $\bar{E}_{i}$ is the energy change required to bring the ion from
infinity to the position $x$. In practice, $\tilde{E}_i$ can be approximated by $\tilde{E}_i \approx q_i \phi(x)$. Thus by combining the Poisson equation with the Boltzmann charge density described by eq.1.3, the electrostatic potential is governed by the Poisson-Boltzmann equation (PBE):

$$\nabla \cdot [\epsilon(x) \nabla \phi(x)] = -4\pi \rho_{\text{solute}}(x) - 4\pi \sum_{i=1}^{n_i} q_i c_i \exp\left[-q_i \phi(x)/k_B T\right]$$ (1.4)

When $\phi(x)$ is small relative to $k_B T$, a linearized version of the PBE can be obtained by a Taylor expansion of the Boltzmann factor with a truncation beyond the first-order terms:

$$\exp[-q_i \phi(x)/k_B T] \approx 1 - [q_i \phi(x)/k_B T]$$ (1.5)

However, for highly charged systems such as nucleic acids, the full nonlinear PBE is needed to produce better agreement with the experimental results in computing the electrostatic potential [11].

Since the analytical solution to the PBE exists only for some simple geometries, more complicated cases can only be obtained by approximate numerical solutions. Most available programs use different versions of the finite difference method first proposed by Warwicker and Watson in 1982 [8]. The first step requires discretization, which places the system of interest into a cubic box divided into a regular Cartesian grid. All physical quantities are mapped onto the grid, replacing differential operators by grid value differences. To this end, charge is mapped to grid points composing the corners of the cube wherein the charge lies. Each of the small cubes is assigned a value of the dielectric constant that corresponds to that of the macromolecule or the solvent. The second step usually involves an iterative procedure, where an initial estimate of the solution is refined successively to solve the finite difference equation. During the 1980s, several software packages were developed using the finite difference algorithm including DelPhi [9], UHBD [19] and GRASP [20], for calculating the electrostatic properties of macromolecules in solution.

The finite difference procedure presents some disadvantages. For example, the free energy is largely dependent on the relative position of charges on the grid and on the dimension of the mesh; the mesh must be fine enough not to merge opposite charges at the same node; and the mesh must be fine enough to represent properly the macromolecule-solvent interfaces [21]. Also, the finite difference PBE algorithm is still computationally demanding, and there
is a dramatic increase of storage when the outer boundary must be increased in order to model highly charged systems properly at low salt conditions.

In response to these limitations, a variety of other numerical methods have been proposed to solve the PBE, such as the finite element method, and the boundary element method (BEM). BEM has several advantages over the finite difference method when used to solve the linear PBE: the far boundary conditions are implicitly and exactly imposed; only a surface mesh is needed, and accurate predictions of both the potential and the electric field are obtained at the molecular surface. However, when used to solve the nonlinear term, BEM requires much longer CPU times to achieve the same accuracy as a finite difference scheme.

Recently, a hybrid PBE method was developed by Boschitsch and Fenley [12] that combines the intrinsic advantages of BEM and finite difference schemes to solve the nonlinear PBE for arbitrary complex-shaped highly charged biomolecules. The fast BEM is used to solve the linearized PBE problem, which retains the intrinsic advantages of a BEM. In order to progress from the linear to nonlinear PBE solution, a correction potential is added. This algorithm involves two steps where the linear PBE is obtained first using the fast BEM, and then the correction for nonlinearity is produced by solving the correction term upon a finite difference mesh. Simulation using the hybrid approach showed very good convergence, even for highly charged systems in the limit of low salt concentration, without requiring the assignment of optimal parameters as the finite difference algorithm does. The hybrid approach takes much shorter CPU times compared with fast BEM. The computed electrostatic properties of nucleic acids are in good agreement with published results [12]. The fast BEM linear PBE approach has been used to study the electrostatic properties of some protein molecules. This study has shown that the fast BEM linear PBE approach accurately predicts surface electrostatic potentials and variations in electrostatic solvation free energies with salt concentrations for realistic protein structures [13].

### 1.4 Surface Area and Free Energy Calculations

Water molecules contribute to the overall stability of helical conformations of nucleic acids. Interaction with water affects both the 3-dimensional conformation and conforma-
tional flexibility. Study of solvent accessibility can provide valuable information as to the influence of water on the system of interest.

The concept of solvent accessible surface area was first introduced by Lee and Richards in 1971 [22] (see Figure 1.4a). The solvent accessible surface is traced out by the probe sphere center as it rolls over the macromolecule. It is analogous to an expanded van der Waals surface where the atomic radii are increased from its intrinsic value to the value corresponding to the sum of the radii of the atom and the probe solvent molecule (the latter is usually taken as 1.4 Å). Later, a concept called contact surface was introduced by Richmond and Richards [26]. The contact area of an atom is the surface area of the atom which may have contact with a specified probe molecule when placed over the solute atom without penetrating any other solute atom. Another parameter is called reentrant surface, which consists of the inward-facing part of the probe sphere when it is in contact with more than one atom. Reentrant surface together with contact surface form the molecular surface, which is the surface traced by the inward-facing surface of the solvent probe sphere (see Figure 1.4b). Interactions between molecules are most likely to be mediated by the

![Diagram](https://via.placeholder.com/150)

**Solvent accessible area**  
**Molecular surface**

*Figure 1.4. Definition of Molecular Surface Area*

(a)solvent accessible surface of a molecule (b)molecular surface of a molecule

properties of residues at their surfaces. Surface area calculations can be used to provide a quick intuitive understanding of the solute’s interaction with solvent. Monitoring the solvent accessible surface area can provide information about the binding and interactions of biological monomers to form a complex [23]. Quantitative calculations of surface area can
also be used in solvation free energy analysis to estimate the nonpolar contribution to the solvation free energy [9], [21].

The free energy of a process occurring in aqueous solution can be broken down into a series of distinct steps which can be treated separately using different theoretical approaches. For example, the free energy of a molecule can be calculated as:

\[ \Delta G_{\text{mol}} = < E_{\text{mm}} > + \Delta G_{\text{solv}} - T S_{\text{solute}} \]  

(1.6)

where \(< E_{\text{mm}} >\) denotes the total of molecular mechanics energies of a molecule in the gas phase, which can usually be obtained from molecular dynamics simulation. \(\Delta G_{\text{solv}}\) describes the free energy associated with transferring the molecule from the gas phase to aqueous solution. The \(-T S_{\text{solute}}\) term denotes an estimate of the solute entropies. The solvation free energy is divided into two parts:

\[ \Delta G_{\text{solv}} = \Delta G_{\text{solv}}^{\text{elec}} + \Delta G_{\text{solv}}^{\text{np}} \]

The \(\Delta G_{\text{solv}}^{\text{elec}}\) is the electrostatic contribution to the solvation free energy. This can be calculated using either the PB model or the generalized Born (GB) model. In the GB model, the electrostatic contribution to the solvation free energy is [17]:

\[ \Delta G_{\text{solv}}^{\text{elec}} = \frac{1}{2} \left( 1 - \frac{1}{\epsilon_w} \right) \sum_{i,j} \frac{q_i q_j}{f_{\text{GB}}} \]  

(1.7)

where \(q_i\) and \(q_j\) are partial charges, \(\epsilon_w\) is the solvent dielectric constant, and \(f_{\text{GB}}\) is a function that depends on an “effective Born radius” \(R_i\) and the distance \(r_{ij}\) between atoms. The effective Born radius roughly describes the average distance from a charge to the dielectric boundary, and depends on the positions and volumes of all other atoms in the solute.

In the PB model, the electrostatic solvation free energy is:

\[ \Delta G_{\text{solv}}^{\text{elec}} = G_{rf}(k, \epsilon_w) - G_{rf}(0, 1) \]

\[ G_{rf}(k, \epsilon) = \frac{1}{2} \sum_{q} \phi_{rf}(\rho_k) Q_k \]

where \(G_{rf}(k, \epsilon)\) is the reaction field energy with salt concentration \(k\) and dielectric constant \(\epsilon\). The electrostatic solvation free energy is the difference of reaction field energy between
the solvent and vacuum media [13]. The nonpolar contribution to the solvation free energy is dependent on the surface area of the macromolecule, which can be represented as:

\[ \Delta G_{np} = \gamma SA + b \]  

(1.8)

where \( SA \) is the surface accessible area, and \( \gamma \) and \( b \) are constants. Values for \( \gamma \) and \( b \) were extracted from a least-squares fit of a plot of experimental alkane transfer free energies versus accessible surface area. The best-fit values of \( \gamma \) and \( b \) varied slightly between radius sets, ranging between 5.0 ± 0.5 cal/Å² and 0.86 ± 0.1 kcal/mol for \( \gamma \) and \( b \), respectively [14].

The MM/PBSA or MM/GBSA approaches combine the molecular mechanics energy for the solute with a continuum solvation approach and normal mode analysis to evaluate the total free energy [27]. Depending on whether the solvation model is based on the PB or the GB equation, this approach is called MM/PBSA or MM/GBSA. MM/PBSA or MM/GBSA calculate the free energies of the end states directly, avoiding the time-consuming simulation of intermediate states required by traditional free energy perturbation and thermodynamic integration methods. This approach has been used to compute the binding free energies and compare the relative stabilities of different conformations of nucleic acids. For example: MM/PBSA approach has been used to calculate the absolute free energies of binding between cobalt (III) hexammine and the P5 helix of the group I ribozyme [15] and to analyze the stability of the DNA tetrameric I-motif [16].
CHAPTER 2

METHODS

2.1 Structures of $\psi$BP and uBP

The RNA structures used in this study are a consensus sequence of the intron with a short region of the U2 snRNA. This pairing forms a complementary helix of seven base pairs with a single unpaired adenosine, whose 2$'\text{'}$ hydroxyl is the nucleophile in the first transesterification reaction of the splicing process. To add stability to the structure, two base pairs were added to one end of the helix. The sequences of both duplexes are shown in Figure 1.2. Coordinates of the unmodified duplex (uBP, PDB: 1LMV) and the $\psi$-modified duplex ($\psi$BP PDB: 1LPW) were obtained from the Protein Data Bank (http://www.rcsb.org/pdb/). The structures of both RNA duplexes were determined using NMR spectroscopy by Newby and Greenbaum [4], [34]. The uBP adopts a typical A-type RNA helix. The bases around the branch site form normal Watson-Crick base pairing. Hydrogen bonds are formed between U6 and both A23 and A24. The A24 is stacked inside the helix. On the other hand, the $\psi$BP has a profound kink in the backbone of the intron strand, with the A24 extruding outside the helix. There are no hydrogen bonds between $\psi$6 and A23 and A24 that can be confirmed by NMR methods. Some conformational variability was observed in the backbones of the intron connecting the U22, A23 and A24, and also in the relative position of $\psi$6 with respect to A23. The stability of both duplexes was evaluated by measurement of melting transitions. The $\psi$BP was found to have a higher melting temperature than the uBP. The difference is about 4$^\circ$F. The free energies of formation was calculated and $\psi$BP is $-0.7$ kcal/mol more stable than uBP[34]. There are 9 lowest energy structures for $\psi$BP and 10 lowest energy structures for uBP deposited in the PDB. The PBE calculations in this study were applied to all the models in both duplexes.
For comparison purposes, an ideal A-type RNA structure (Arnott fiber diffraction model [35]) was generated using Amber’s NUCGEN utility. There are ten normal Watson-Crick base pairs with the exact same sequence as the uBP except that an extra uridine was added to form a Watson-Crick base pair with the branch site adenosine. Since the NUCGEN utility doesn’t give hydrogens in the resulting RNA, the Reduce program (version 2.13.5) was used to add hydrogens to the ideal RNA structure. Reduce is a C++ program which was developed at Duke University in the lab of David and Jane Richardson [37]. Hydrogens are added in standardized geometry.

2.2 MD simulations

The MD simulations were carried out using the Amber-6.0 program with the Cornell et al. force field. The starting structures were specified in section 2.1.

2.2.1 System set-up

The set-up was done in the LEaP module of the Amber program, which reads in force field topology and coordinate information and produces the files necessary for production MD calculations. For LEaP to read in the PDB files correctly, some naming and formatting was done using Perl scripts.

After the PDB file was loaded correctly, the RNA molecule was neutralized with sodium counterions by LEaP. A cubic box of the TIP3P water [46] (a rigid water model that represents water as three interaction sites associated with oxygen and proton positions) was added around the RNA to a depth of 10 Å on each side of the solute. The topology and coordinate files for the solute alone (called the gas phase) and with the explicit solvent were saved for later calculations.

2.2.2 System equilibration and production

Before production dynamics starts, potentially bad van der Waals and electrostatic interactions should be minimized. In particular, water was inserted from a pre-equilibrated water box and therefore, has not felt the influence of the solute. Moreover, there may be gaps between the solvent and solute and solvent and box edges. Therefore, before one runs
production dynamics, it is wise to let the water relax around the solute to come to an equilibrium density, etc.

In this study, the equilibration was done using the Sander-classic module of the Amber program. The first step for equilibration is a short minimization with the solute fixed. The minimization was turned on by setting IMIN=1. MAXCYC was set to be 1000 to run 1000 steps of steepest descent method. The solute was kept frozen by using the belly option. This means that the water and counterions are free to move while the solute is in a fixed position.

Minimizations serves to move the energy of the system down the hill. But it doesn’t really sample the accessible conformational space of the water. After the minimization, the dynamics was run, holding the solute fixed with belly option while the bulk of water relaxes. SHAKE method on hydrogens was turned on in a rigid water model like TIP3P so that the hydrogen atoms do not extend beyond the van der Waals sphere defined for the oxygen atoms. After that, the dynamics step was repeated with the PME method, since the production run will use PME to account for the long-range electrostatics interaction. PME is a grid-based implementation of the Ewald method that minimizes the computational expense of the Ewald approach, making it of comparable computational efficiency as truncation methods. After each dynamics step, the box size changes in an attempt to maintain constant pressure. Therefore, the reference coordinate should be the result coordinate file of the previous MD run.

This minimization and MD cycle was repeated twice. During this process, the constraints on the RNA solute were gradually lifted until the solute can move around freely. Then the system’s density, pressure and temperature were extracted from the output file with a Perl script. This contains information about the values of these properties as a function of time. The values were plotted to make sure that they become constant as time goes, implying that the system has reached equilibrium.

After equilibration, running the production stage is straightforward. The production run was carried out using the Sander module with PME and the coordinates from the result of equilibration stage as reference. The time step is 0.002 picoseconds. The trajectories were developed in a (N, P, T) ensemble. Each was extended into nanoseconds regime and saved every picosecond.
2.2.3 Analysis of the trajectories

Since the solvent was included explicitly in our simulation, water and ions were first stripped off from the trajectory files using the mdanal module. The input file to mdanal specifies the topology as reference and the trajectories we want to analyze. Carnal module was then used for calculating the root-mean standard deviation (RMSD) for all atoms of each frame in our trajectory to the starting NMR-calculated structure. In the input file, the gas-phase topology and trajectory files were specified as input and the file rms_to_start as output. After carnal was run, this file contained a table of RMS values. The RMS values were plotted against time using Xmgr to see the fluctuation of RMS as the simulation goes. An average structure from the trajectories was also calculated using the carnal module. The RMS to the average structure was analyzed using the same protocol.

The trajectories were visualized using the Visual Molecular Dynamics (VMD [47]) program. The RNA helical parameters were measured from MD snapshots using the 3DNA [36] program. The trajectories was first converted into PDB files. The PDB files were further formatted for 3DNA by a Perl script developed as a part of this thesis (see Appendix A).

2.3 Numerical solution to the nonlinear PBE

Due to the high charge density of nucleic acids, it is necessary to solve the complete nonlinear Poisson-Boltzmann (NLPB) equation. Numerical solutions to the NLPB were obtained from a finite difference PBE algorithm that has been incorporated into the DelPhi program and a hybrid boundary element and finite difference PBE algorithm developed by Boschitsch and Fenley.

A focusing procedure was used to run the DelPhi program. This was done in two steps. During the first step, the scale of the grid was set to be 1.4 grids/Å and the box fill was set to be 30%. A large percentage fill will provide a more detailed mapping of the molecular shape onto the lattice. On the other hand it will bring the dielectric boundary of the molecule closer to the lattice edge. This will cause larger errors arising from the boundary potential estimates. A smaller percentage fill will increase the accuracy of the boundary conditions, but results in a coarser representation of the molecule. So in the first step, this smaller value of box fill was chosen to get a more accurate boundary condition. As has been discussed
extensively with proteins, a dielectric constant of 2 was used to account for the electronic polarizability [5]. The interior dielectric constant of RNA was also set to be 2 and the exterior dielectric constant was left as the default 80. The salt was set as sodium chloride. The boundary between the molecule and solvent is defined by the solvent excluded surface using a 1.4 Å solvent probe. In the second run, the boundary condition for focusing was turned on and the potential map from the previous run was read in. The percent box fill was increased to be 90% to obtain a finer representation of the molecule. The univalent salt concentration of 0.1 mole/liter was set in both runs.

The AMBER and Discover parameter sets that describe atomic size and partial charges of the atoms in RNA molecules were employed. Most calculations of the PBE in this thesis were carried out using both the AMBER94 (Cornell et al.) force field and the Discover force field (http://www.pylelab.org). The partial charges of the pseudouridine base were provided by our collaborator (Dr. M. Nagan, Truman State University) and were obtained by high level molecular quantum mechanics calculations (Gaussian98 program).

2.4 Visualization of the electrostatic surface potential

The schematic view of the structures of each RNA duplex (Figure 1.3) was displayed with the Dino [38] program. The calculated electrostatic potentials obtained from DelPhi were also displayed using the Dino program. The molecular surface was calculated using the MSMS [39] program using atomic radii from the Cornell et al. force field. The PDB file, the surface file and the potential map obtained from DelPhi were all loaded into the program. Since RNA molecules have mostly a negative potential on the surface, the color mapping of the electrostatic potential was scaled as follows: yellow is the most negative region, red denotes potential from -15 kcal/mol to -3 kcal/mol, white represents the -3 kcal/mol to neutral region and blue is the neutral to 1 kcal/mol region and green is the most positive region.

The calculated electrostatic potentials from the Boschitsch and Fenley program were displayed using VRML 2.0 (virtual reality modeling language) [41]. Here a slightly different color scheme is employed where yellow is again the most negative region, followed by red, white is neutral potential and blue and green are regions of positive potential, with green
being more positive. To correlate the surface potential with atom position, the ribbons representation of the structure of the duplexes was incorporated into the electrostatic potential surface maps. The three-dimensional structure of the RNA duplex was displayed using the Ribbons program which was developed by Mike Carson [40]. The ribbons-data utility was used to generate the models. The original PDB file was first split into two pdb files for each strand of the double strands. The *.ss and *.cyl files were generated for each strand. The *.ribbons file was then created by combining the *.ss file for each strand. Since the Ribbons program needs phosphate groups to connect the backbones and the duplexes used didn’t have phosphates at the 5′ terminals, artificial phosphate groups were added to the 5′ terminals by mimicking the relative positions of sugar to phosphate of previous residues. The calculation was performed by a perl script written as part of this thesis (see Appendix A). The structures were saved as VRML 2.0 files, which were then incorporated into the electrostatic map file from the hybrid PBE program using the Inline command.

2.5 Calculation of solvent accessible surface area (SASA)

The overall polar and nonpolar surface area of each duplex was modeled with Naccess [42], which is an implementation of the surface area method of Lee and Richards. Since the surface area values calculated depend not only on the method, but also on the radii used for each atom. The van der Waals radii of each atom are taken from Alden and Kim [44] as suggested by Connolly [45] in his review. The probe radius is set to be 1.4 Å. The hydrogen atoms are included implicitly during the calculation.

2.6 Calculation of solvation energies

Nucleic acid oligomers are highly charged molecules, and a key aspect of their solvation energies lies in their electrostatic interaction with solvent. As discussed above in the introduction part, continuum approach using either PB or GB model can be used to evaluate the electrostatic component of the solvation free energy. In this study, the PB model was used as implemented in the UHBD (University of Houston Brownian Dynamics) program [43]. The dielectric boundary was taken as the molecular surface defined by a 1.4 Å probe
sphere and by spheres centered on each atom with radii taken from Alden and Kim and the partial charges were taken from Cornell et al. force fields in order to be consistent with the energies of the MD simulations we are analyzing. After the QCD files (files with coordinates, charge and radii) were generated using the PDB files and the charge and radii files, the NLPB equation was solved using a two-step focusing procedure. For the first run, $55^3$ grid points spaced by 1.6 Å was used. A finer grid of $110^3$ points spaced by 0.4 Å was used. The relative dielectrics of the solute and the solvent regions were set to be 1 and 78.5, respectively, while the ionic strength was set to 0.15 moles/liter.

The nonpolar contribution to solvation i.e. the free energy of solvation of the “discharged” molecule in which all the atom-centered partial charges are set to zero, was modeled by Equation 1.4, where $\gamma$ is taken as 0.00542 kcal/Å$^2$, and $b$ is 0.92 kcal/mol. The SASA was calculated using Naccess as described in Section 2.5.
CHAPTER 3
RESULTS AND DISCUSSIONS

3.1 Preliminary results of MD simulation

The MD simulations were carried out on the unmodified RNA duplex using AMBER-6.0 as described in Chapter 2. The MD simulations on the modified RNA duplex were carried out by our collaborator (Dr. Nagan). Therefore, only the results of the uBP MD simulations are reported here. The MD simulations were done on models one, four and six of the 10 NMR models deposited, which are the three lowest energy models. All the MD simulations were carried out beyond 2.5 nanoseconds.

The MD simulations produced fairly stable trajectories with low root mean square deviation (RMSD) with respect to the starting structure (see Figure 3.1) and the average structure (see Figure 3.2). The figures are the time evolutions of RMSD of the model four of the 10 NMR structures. The distance between the phosphate groups of U4 and A26 is also monitored. Figure 3.3 showed the evolution of the inter-phosphate distance of the simulated structure. The distance is fairly stable over the course of simulation. The average distance is 11.4 ± 1.3Å. The interphosphate distance of the NMR structure is 9.4 ± 1.07Å. These results indicate that a stable structure of the RNA duplex in solution has been reached for each simulation and the starting structures obtained by NMR methods are globally minimized.

The analysis of the MD trajectories is still continuing. With the detailed information about the structure and dynamics provided by the MD simulation, we are trying to achieve two major goals. The first one is to calculate free energies for the two RNA duplexes by using MM/PBSA and (or) MM/GBSA methods. This calculation has been applied to the relative energy calculations of A- and B- DNA by Srinivasan et al. [27], which found a preference for the B-form by a reasonable amount. By comparing the free energies of the uBP and ψBP,
Figure 3.1. Time evolution of RMSD of the simulated structure (based on the all-atom model) with respect to the starting NMR structure.

we should be able to determine what energy component(s) contribute to the overall stability of the ψBP reported by the Newby and Greenbaum study. The second goal is to study the hydration of the two RNA duplexes. Given that MD simulations give a time history for all atomic motions in the system, MD simulations with explicit solvent treatment will provide us the precise details for all specific ion and water interactions with the RNA solute. NMR experiments by Newby and Greenbaum [50] suggested the presence of a water-mediated hydrogen bond between the ψNH1 proton and the phosphate oxygens belonging to its own and the 5′ nucleotide’s phosphate group, which stabilize the extra-helical position of the branch site adenosine. The water residence time analysis will help us pinpoint the presence of the water-mediated hydrogen bond.
Figure 3.2. Time evolution of RMSD of the simulated structure (based on the all-atom model) with respect to the average structure

### 3.2 Surface electrostatic potentials of the RNA duplexes

The central aim in studying the electrostatic properties of macromolecules is to use the structural information from high-resolution experiments to obtain a realistic description of the electrostatic potential maps. The electrostatic potential distribution can be used in a variety of ways. One of the important applications is to use the graphical analysis of the electrostatic potential to reveal deeper aspects of the structure, and to help identify functionally important regions or active sites. Since interactions between molecules are most likely to be mediated by the properties on the surface, a useful way of visualizing electrostatic properties is to represent regions of electrostatic potentials on the molecular surface. This representation makes it possible to correlate certain regions of the surface with specific electrostatic characteristics. In this study, we calculated the electrostatic potentials of both unmodified (uBP) and \( \psi \) modified (\( \psi \)BP) branch site duplexes and compared the
Figure 3.3. Time evolution of interphosphate distance (P4-P26) of the simulated structure

surface potentials of both RNA duplexes to identify regions of heterogeneous electrostatic
surface potentials.

3.2.1 Surface electrostatic potentials using DelPhi

The surface electrostatic potentials of uBP and ψBP RNA duplexes were calculated
using the finite difference PBE algorithm implemented in DelPhi with the partial charges
and atomic radii taken from the Cornell et al. force field. The results are shown in Figure
3.4 and Figure 3.5. The electrostatic features of both RNA duplexes are similar to a typical
A-RNA helix. The backbones are mostly negative since the phosphate groups are negatively
charged and the grooves have patches of positive potentials due to some positively charged
base atoms. Surface potential in the major groove region of the ψBP is more negative than
that of the uBP to a small degree, but not significantly so to be conclusive.
3.2.2 Surface electrostatic potentials using hybrid PBE approach

The surface electrostatic potentials of uBP and \( \psi \)BP were calculated using the hybrid PBE approach as described in chapter 2 with the partial charges and atomic radii taken from the Cornell et al. force field. The electrostatic potentials of the ideal A-RNA as described in Chapter 2 were also calculated with the same force field parameters for comparison.

The electrostatic potential mapped on the surface of the unmodified duplex (Figure 3.8) has similar features to that of an ideal A-RNA helix (Figure 3.6 and Figure 3.7). In contrast to B-DNA, the surface potential of uBP in the major groove is much more negative than in
Figure 3.5. Surface electrostatic potentials of $\psi$BP calculated using DelPhi
The atomic radii and partial charges are taken from the AMBER-6.0 force field. The color scheme used in
the map is the same as in Figure 3.4.

the minor groove. This is consistent with the observation that divalent metal ions tend to
bind in the major groove of RNA[30].

The electrostatic surface potential of the $\psi$-modified duplex is more intriguing (see Figure
3.9). While the backbone region and the grooves have similar features to the uBP, there is a
distinctively negative region in the major groove that is not seen in the uBP. To identify the
location of this negative pocket, the structure was combined with the electrostatic potential
map (see Figure 3.10). As shown in the figure, this negative pocket is located near the 2$'$OH of
the branch site adenosine, which is the nucleophile in the cleavage reaction at the pre-mRNA
5$'$ splice site. Since the electrostatic properties calculated with continuum PBE model are
sensitive to the atomic radii and partial charges used, the electrostatic potentials calculations
were repeated with the Discover force field parameters in order to test the sensitivity of this
**Figure 3.6.** Surface electrostatic potential of an ideal A-RNA (major groove view)
This ideal A-RNA structure was generated by NUCGEN utility of AMBER-6.0. The sequence was similar to the uBP as described in Chapter 2. The atomic radii and partial charges are taken from the AMBER-6.0 force field. The color scheme used in the map is: from yellow to red to white to blue then to green. Yellow is the most negative (-3 kcal/mol) and green is the most positive (1 kcal/mol) and white is neutral. The calculation is carried out using the hybrid PBE approach.

potential hole to the charge and atomic radii distribution. The result shows very similar features in both uBP (see Figure 3.11) and ψBP (see Figure 3.12) as compared to the calculations with Cornell *et al.* force field parameters. In order to determine whether the negative potential region in ψBP comes from the phosphate-sugar backbone or the bases, calculations were performed where a -1e charge was placed only on the phosphates of the RNA duplexes. The negative potential hole in electrostatic surface potential map of ψBP is still present (see Figure 3.13), whereas it is not visible in the uBP (see Figure 3.14). This indicates that the partial negative charges on the bases and ribose are partially responsible for creating the exceptionally negative electrostatic potential region.
Figure 3.7. Surface electrostatic potential of an ideal A-RNA (minor groove view)
The sequence of this A-RNA is the same as in Figure 3.6. The atomic radii and partial charges are from the
AMBER-6.0 force field. The color scheme used in the map is the same as in Figure 3.6. The calculation is
carried out using the hybrid PBE approach.

The hybrid PBE approach produced more detailed surface potential map than the DelPhi
program because in the hybrid approach, the surface properties are more accurate since these
electrostatic properties at each surface element is provided directly, without the need for any
interpolation scheme, such as required in the finite difference PBE approach.

This heterogeneity is likely to play an important role leading to molecular recognition
of the branch site adenosine by other components of the spliceosome for nucleophilic
attack. Electrostatic surface potential studies have revealed that the surfaces of proteins are
covered by distinct patches of electrostatic potential that plays important roles in molecular
recognition and ligand binding events [6]. As the number of solved RNA structure deposited
in the PDB increases, electrostatic potential calculations have been carried out on a variety
of RNA molecules [5]. These calculations showed that on the surface of RNA molecules there
Figure 3.8. Surface electrostatic potential map of uBP with Cornell et al. force field parameters
The atomic radii and partial charges are taken from the AMBER-6.0 force field. The color scheme used in the map is the same as in Figure 3.6. The calculation is carried out using the hybrid PBE approach.

also exist sites of unusual electrostatic features that correspond to functionally important regions, especially metal ion binding sites.

Metal ions, especially Mg$^{2+}$, are crucial to stabilizing RNA structures and participate in catalytic reactions of some ribozymes [48]. Therefore, reliably locating the ion binding sites is important for understanding RNA structure and function. Identifying divalent metal ion binding sites in RNA molecules is a formidable task, especially when the RNAs reside within large, dynamic assemblies such as the spliceosome. The polyanionic nature of the RNA backbone leads to the binding of numerous metal ions, the vast majority of which have no catalytic role. Pinpointing the ligands for functionally important metal ions, active site or otherwise, can be a challenge given this background. If the structure of the RNA is known, then computational methods can be used to predict and characterize the ion binding
Figure 3.9. Surface electrostatic potential of \( \psi \)BP with Cornell \textit{et al.} force field parameters

The atomic radii and partial charges are taken from the AMBER-6.0 force field. The color scheme used in the map is the same as in Figure 3.6. The calculation is carried out using the hybrid PBE approach.

sites. For example, NLPB calculations reveal that in the center of the P645 domain from the \textit{Tetrahymena} group I intron (PDB: 1gid), there is a region with extremely negative electrostatic potential that corresponds precisely to the position of the metal ion core [5]. Other examples include the loop E of 5s rRNA (PDB: 364d) [5] and 23s rRNA [49]. Yean \textit{et al.} [51] showed that yeast U6 snRNA, which forms extensive base pairs with U2, specifically binds a divalent metal ion, probably Mg\(^{2+}\), that is required for catalysis of the first step of splicing. The Newby and Greenbaum study found that the presence of magnesium ion did not alter the NMR chemical shifts of the U2 snRNA-intron duplex, implying that there is no structurally important specific metal ion binding site. However, this finding does not exclude the possibility of involvement of hydrated ions during catalysis or of other cationic
Figure 3.10. Surface electrostatic potential of ψBP with Cornell et al. force field combined with the structure of ψBP.
The atomic radii and partial charges are taken from the AMBER-6.0 force field. The color scheme used in the map is the same as in Figure 3.6. The calculation is carried out using the hybrid PBE approach. The 2′OH of the branch site adenosine is shown as a pink sphere.

components involved in splicing activity. This site will be further investigated in future MD simulations.

3.3 Surface Areas and solvation free energy

Table 3.1 listed the results of the accessible surface area (ASA) calculation of the two RNA duplexes calculated with Naccess program as described in Chapter 2.

There is only a 2% difference in the total solvent accessible areas between the two RNA duplexes. This result was no surprising given that the two duplexes are so similar in their sequence and helix form. The ratio of polar/nonpolar exposure level of ψBP is slightly larger than that of uBP. Since the stable form of any polyanionic system such as RNA in solution
Figure 3.11. Surfac electrostatic potential of uBP calculated with the Discover force field. The atomic radii and partial charges are taken from the Discover force field. The color scheme used in the map is the same as in Figure 3.6. The calculation is carried out using the hybrid PBE approach.

may be the one with maximal polar part exposure and minimal nonpolar part exposure [52], this result showed that \( \psi \)BP is slightly more stable than uBP, which is consistent with the experimental results. There is a significant increase in the amount of nonpolar exposure of the branch site adenosine (A24). Since adenosine has the most nonpolar atoms among all the nucleotides, this is in line with the structure showing that the branch site adenosine of \( \psi \)BP is extruding into the solvent instead of stacking in the helix as the A24 of uBP.

The most intriguing result is the ASA of the A24’s 2’ hydroxyl of \( \psi \)BP, which is much smaller than that of uBP. This is contradictory to what we had expected, and implies that the 2’OH of \( \psi \)BP is embedded in the major groove.

Table 3.2 lists the result of solvation free energy calculation. The nonpolar contribution to the solvation energy is around 21 kcal/mol. There is no significant difference between the uBP and \( \psi \)BP RNA duplexes. This number is in good agreement with the calculations of
Figure 3.12. Surface electrostatic potential of $\psi$BP calculated with the Discover force field

The atomic radii and partial charges are taken from the Discover force field. The color scheme used in the map is the same as in Figure 3.6. The calculation is carried out using the hybrid PBE approach.

<table>
<thead>
<tr>
<th></th>
<th>Total ASA ($\text{Å}^2$)</th>
<th>polar/nonpolar</th>
<th>nonpolar ASA of A24 ($\text{Å}^2$)</th>
<th>ASA of 2’OH of A24 ($\text{Å}^2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\psi$BP</td>
<td>3803.34</td>
<td>2.5</td>
<td>99.54</td>
<td>0.45</td>
</tr>
<tr>
<td>uBP</td>
<td>3714.39</td>
<td>2.4</td>
<td>45.21</td>
<td>7.3</td>
</tr>
</tbody>
</table>

Table 3.2. Solvation free energy $\psi$BP and uBP

<table>
<thead>
<tr>
<th></th>
<th>Electrostatic solvation energy (kcal/mol)</th>
<th>Nonpolar solvation energy (kcal/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\psi$BP</td>
<td>-4862</td>
<td>21.53</td>
</tr>
<tr>
<td>uBP</td>
<td>-4997</td>
<td>21.05</td>
</tr>
</tbody>
</table>
Figure 3.13. Surface electrostatic potential of ψBP with only a -1e charge on the phosphate groups
The atomic radii are taken from the AMBER-6.0 force field. The color scheme used in this map is: from yellow to red to white to blue then to green. Yellow is the most negative (-3 kcal/mol) and green is the least negative -1 kcal/mol. The calculation is carried out using the hybrid PBE approach.

Srinivasan et al. [27]. The Srinivasan et al. study calculated the solvation free energies of A- and B-form RNA with ten base pairs. The nonpolar contribution to the solvation free energy for both A- and B-form RNAs is $24 \pm 0.1$ kcal/mol. Since the exposed surface area changes little with helix form, there is little difference between the two RNA duplexes in terms of the nonpolar solvation energy.

The uBP duplex is favored in terms of electrostatic contribution to the solvation free energy by 135 kcal/mol. This may seem to be a small amount compared with the total electrostatic solvation free energy. However, in the Srinivasan et al. study, the difference of electrostatic solvation free energy between A-RNA and B-RNA is 240 kcal/mol, with B-RNA disfavored, explaining why B-RNA is not stable in solution. By comparison with
Figure 3.14. Surface electrostatic potential of uBP with only a -1e charge on the phosphate groups
The atomic radii and are from the AMBER-6.0 force field. The color scheme used in the map is the same as in Figure 3.13. The calculation is carried out using the hybrid PBE approach.

the Srinivasan et al. study, we can see that solvation favors the uBP over ψBP by a fairly large amount. Since Newby and Greenbaum study showed that ψBP is more stable than uBP, we propose that the gas phase energy or the solute entropy would greatly favor ψBP to compensate the unfavorable solvation energy of ψBP. This hypothesis would be validated with the future MM/PBSA study.
CHAPTER 4

CONCLUSIONS

In this study, computer simulations were carried out on both modified and unmodified U2 snRNA-intron duplexes including preliminary molecular dynamics (MD) simulations, electrostatic potential, surface area, and solvation free energy.

The preliminary MD simulations results showed that the RMSD of the simulated structure is small and the major groove width is also close to the NMR structure. This indicates that the MD simulations have produced stable trajectories of the RNA duplexes in solution. In future work, the trajectories would be analyzed to study the energetics, hydration patterns and ion associations of both RNA duplexes. These studies will provide us insights into how the $\psi$BP is stabilized and what role water plays in the structure stabilization.

The surface electrostatic potentials were calculated with continuum solvent model using the nonlinear Poisson-Boltzmann equation (NLPB). The solutions to the NLPB were obtained with both finite difference algorithm implemented in DelPhi and a hybrid PBE approach developed by Boschitsch and Fenley. Results show that there is a region of exceptionally negative potential near the $2'\text{OH}$ of the branch site adenosine. This is proposed to be a potential metal ion binding site and will be further investigated in future MD simulations.

The two RNA duplexes have similar solvent accessible surface areas. Therefore, the nonpolar contribution to the solvation free energy is about the same. However, the electrostatic contribution to the solvation free energy favors the unmodified RNA duplex by 135 $kcal/mol$. Other components of the total free energy will be studied in the future work by MM/PBSA analysis.
APPENDIX A

PERL SCRIPTS DEVELOPED

#!/usr/bin/perl

########################################################
# This program will change the naming of residues from #
# AMBER format to 3DNA    #
# Then you should type "./amber2X3DNA1.pl input.pdb output.pdb" #
########################################################
$input = shift(@ARGV);
$output = shift(@ARGV);
format OUT=
@<<< @>>>> @<<<< @<<<< @<<<< @<<<< @<<<<
$a[0],$a[1],$a[2],$a[3],$a[4],
$a[5],$a[6], $a[7].

$strand_number = "A";
open(IN,"$input") || die "can't open $input";
open(OUT, ">$output") || die "can't create $output";
while ($string = <IN>) {
    @a = split (/\s+/, $string);
    $a[3] =~ s/RG5/RG/g;
    $a[3] =~ s/RA3/RA/g;
    $a[3] =~ s/RU5/RU/g;
    $a[3] =~ s/RC3/RC/g;
    if($a[4]>=10){
        $strand_number ="B";
}
#!/usr/bin/perl

# This program will renumber the amber output#
# This is step two in the process. #
# Run amber2X3DNA.pl first. #
# Then type "./amber2X3DNA.pl input output"#

$input = shift(@ARGV);
$output = shift(@ARGV);
open(IN, "$input") || die "can’t open $input";
open(OUT, ">$output") || die "can’t create $output";
while ($string = <IN>) {
    @a = split (/\s+/, $string);
    $a[3] =~ s/RA/A/g;
    $a[3] =~ s/RG/G/g;
    $a[3] =~ s/RC/C/g;
    $a[3] =~ s/RU/U/g;
    if($a[5] eq "B"){
        $a[4]=$a[4]-9;
    }
    if(substr($a[2],0,1)>0 && substr($a[2],0,1)<9){
        $symble=substr($a[2],0,1);
$a[2]=substr($a[2], 1);

# print "the symbol is $symbol\n";
printf OUT "%-6s%5d %1d%-4s%3s %1s%4d %8.3f%8.3f%8.3f\n",
$a[0],$a[1],$symble,$a[2],$a[3], $a[5],$a[4],$a[6],$a[7], $a[8], 0.00 ,0.00 ;
}
else{
    $symble=" ";
    printf OUT "%-6s%5d %1s%-4s%3s %1s%4d %8.3f%8.3f%8.3f\n",
    $a[0],$a[1],$symble,$a[2],$a[3], $a[5],$a[4],$a[6],$a[7], $a[8], 0.00 ,0.00 ;
}
}

close(IN);
close(OUT);
#!/usr/bin/perl
#extract the distance out of the 3DNA output.
$argu=@ARGV;
if($argu!=2){
    print "usage: extract.pl $x.out\n";
    exit;
}
$input = shift(@ARGV);
$output= shift(@ARGV);
$position=index($input, ".");
$time=substr($input, 0,$position);
open(IN,"$input") || die "can't open $input";
open OUT, ">>$output";
while ($string = <IN>) {
    if($string=~/\^\s*3\s*UG\/CA\s*\d+\s+---\s+\d+\s+\d+\s+\d+/){
        $distance=$1-5.8;
    }
print OUT " $time $distance \n";
}
}
close IN;
close OUT;
# /usr/bin/perl -w
# estimate the position of P based on the other relative info.
open PDB, "<$ARGV[0]";
open TEMP, ">$ARGV[1]";

while(<PDB>){
    if(!^-ATOM|HETATM/){
        if(!^-\w+\s+\d+/){
            if($2==1){
                ($dummy, $anum, $aname, $rname, $strand, $rnum, $coordx, $coordy, $coordz, $rest)=split;
                $ot[0]=$coordx;
                $ot[1]=$coordy;
                $ot[2]=$coordz;
            } else if($2==33){
                ($dummy, $anum, $aname, $rname, $strand, $rnum, $coordx, $coordy, $coordz, $rest)=split;
                $p[0]=$coordx;
                $p[1]=$coordy;
                $p[2]=$coordz;
            } else if($2==34){
                ($dummy, $anum, $aname, $rname, $strand, $rnum, $coordx, $coordy, $coordz, $rest)=split;
                $o1p[0]=$coordx;
                $o1p[1]=$coordy;
                $o1p[2]=$coordz;
            }
        } else if($2==33){
            ($dummy, $anum, $aname, $rname, $strand, $rnum, $coordx, $coordy, $coordz, $rest)=split;
            $p[0]=$coordx;
            $p[1]=$coordy;
            $p[2]=$coordz;
        } else if($2==34){
            ($dummy, $anum, $aname, $rname, $strand, $rnum, $coordx, $coordy, $coordz, $rest)=split;
            $o1p[0]=$coordx;
        }
    }
}
$o1p[1]=\text{coordy};
$o1p[2]=\text{coordz};
}

if($2==35){
    ($dummy, $anum, $aname, $rname, $strand,
    $rnum, $coordx, $coordy, $coordz, $rest)=\text{split};
    $o2p[0]=\text{coordx};
    $o2p[1]=\text{coordy};
    $o2p[2]=\text{coordz};
}

if($2==36){
    ($dummy, $anum, $aname, $rname, $strand,
    $rnum, $coordx, $coordy, $coordz, $rest)=\text{split};
    $o[0]=\text{coordx};
    $o[1]=\text{coordy};
    $o[2]=\text{coordz};
}

if($2==293){
    ($dummy, $anum, $aname, $rname, $strand,
    $rnum, $coordx, $coordy, $coordz, $rest)=\text{split};
    $otb[0]=\text{coordx};
    $otb[1]=\text{coordy};
    $otb[2]=\text{coordz};
}

if($2==321){
    ($dummy, $anum, $aname, $rname, $strand,
    $rnum, $coordx, $coordy, $coordz, $rest)=\text{split};
    $pb[0]=\text{coordx};
    $pb[1]=\text{coordy};
    $pb[2]=\text{coordz};
}
if($2==322){
    ($dummy, $anum, $aname, $rname, $strand,
    $rnum, $coordx, $coordy, $coordz, $rest)=split;
    $o1pb[0]=$coordx;
    $o1pb[1]=$coordy;
    $o1pb[2]=$coordz;
}
if($2==323){
    ($dummy, $anum, $aname, $rname, $strand,
    $rnum, $coordx, $coordy, $coordz, $rest)=split;
    $o2pb[0]=$coordx;
    $o2pb[1]=$coordy;
    $o2pb[2]=$coordz;
}
if($2==324){
    ($dummy, $anum, $aname, $rname, $strand,
    $rnum, $coordx, $coordy, $coordz, $rest)=split;
    $ob[0]=$coordx;
    $ob[1]=$coordy;
    $ob[2]=$coordz;
}
$anum=$2+3;
if($anum>293){
    $anum+=3;
}
}
printf TEMP "'$1$anum$'";
close TEMP;

for($i=0; $i<3; $i++){
    $pt[$i]=$ot[$i]-$o[$i]+$p[$i];
    $o1pt[$i]=$ot[$i]-$o[$i]+$o1p[$i];
    $o2pt[$i]=$ot[$i]-$o[$i]+$o2p[$i];
    $ptb[$i]=$otb[$i]-$ob[$i]+$pb[$i];
    $o1ptb[$i]=$otb[$i]-$ob[$i]+$o1pb[$i];
    $o2ptb[$i]=$otb[$i]-$ob[$i]+$o2pb[$i];
}
print "the coordinates are ", @ot, "\n";
print "the coordinates are ", @o1p, "\n";
print "the coordinates are ", @o2p, "\n";

open TEMP, "<$ARGV[1]";
open MPDB, ">$ARGV[2]";
printf MPDB "%-6s%5d %-4s %s %s%4d %8.3f%8.3f%8.3f%6.2f%6.2f %11s\n", 
    "ATOM",1,"P","G","A",1,$pt[0],$pt[1],$pt[2],1.00,0.00,"P";
printf MPDB "%-6s%5d %-4s %s %s%4d %8.3f%8.3f%8.3f%6.2f%6.2f %11s\n", 
    "ATOM",2,"O1P","G","A",1,$o1pt[0],$o1pt[1],$o1pt[2],1.00,0.00,"O";
printf MPDB "%-6s%5d %-4s %s %s%4d %8.3f%8.3f%8.3f%6.2f%6.2f %11s\n", 
    "ATOM",3,"O2P","G","A",1,$o2pt[0],$o2pt[1],$o2pt[2],1.00,0.00,"O";
while(<TEMP>){
    if(/\^ATOM|HETATM/){
        if(/\^\(w+s+(d+)/){
            if($2==299){
                printf MPDB "%-6s%5d %-4s %s %s%4d %8.3f%8.3f%8.3f%6.2f%6.2f %11s\n", 
                    "ATOM",296,"P","U","B",1,$ptb[0],$ptb[1],$ptb[2],1.00,0.00,"P";
printf MPDB "%-6s%5d %-4s %s %s%4d
%8.3f%8.3f%8.3f%6.2f%6.2f %11s
",
"ATOM",297,"O1P","U",1,$o1ptb[0],$o1ptb[1],$o1ptb[2],1.00,0.00,"0";
printf MPDB "%-6s%5d %-4s %s %s%4d
%8.3f%8.3f%8.3f%6.2f%6.2f %11s
",
"ATOM",298,"O2P","U",1,$o2ptb[0],$o2ptb[1],$o2ptb[2],1.00,0.00,"0";
}
REFERENCES


[30] Cate, J.H. and Doudna, J.A. Metal-ion bindings sites in the major groove of a large ribozyme domain Structure, 4:1221-1229, 1996


[34] Newby, M.I. and Greenbaum, N.L. A conserved pseudouridine modification in eukaryotic U2 snRNA induces a change in branch-site architecture RNA, 7:833-845, 2001


[38] Philippsen, A. DINO: Visualizing Structural Biology http://www.dino3d.org


[40] Carson, M. Ribbons http://sgce.cbse.uab.edu/ribbons/


[42] Hubbard, S. http://wolf.bi.umist.ac.uk/naccess


BIOGRAPHICAL SKETCH

Darui Xu

I was born in China in 1973. In 1996, I got my Bachelor of Science in Chemical Engineering and Bachelor of Arts in English from Dalian University of Technology. After that, I came to the Department of Chemistry and Biochemistry at the Florida State University to pursue my graduate study. After two years in Biochemistry, I felt that bench work in the lab was not my strength nor my interest, although I had no difficulty in understanding the theories. So I began to take courses in Computer Science and got my Master in Computer Science in 2002. Now I am working in the field of computer simulations of biomolecules, where I can use my knowledge in both biochemistry and computer science.