APPLICATION OF *AB INITIO* CALCULATIONS TO COLLAGEN AND BROME MOSAIC VIRUS

A THESIS IN
Physics

Presented to the Faculty of the University
of Missouri-Kansas City in partial fulfillment of
the requirements for the degree

MASTER OF SCIENCE

by
JAY QUINSON EIFLER

Kansas City, Missouri
2014

APPLICATION OF *AB INITIO* CALCULATIONS TO
COLLAGEN AND BROME MOSAIC VIRUS

Jay Quinson Eifler, Candidate for the Master of Science Degree
University of Missouri-Kansas City, 2014

ABSTRACT

In bio-related research, large proteins are of important interest. We study two such proteins. Collagen contains one such protein, the collagen triple-helix, which forms part of the structural matrix for animals, such as in their bones and teeth. 1JS9 is another protein that is a component of the protein shell of the brome mosaic virus (BMV). And BMV is important for drug delivery and imaging. To better understand the properties of these proteins, quantum mechanically (QM) based results are needed, however computationally feasible methods are also necessary. The Orthogonalized Linear Combination of Atomic Orbitals (OLCAO) method is well-suited for application to such large proteins. However, a new approach to reduce the computational cost and increase the computational feasibility is required and this extension to the method we call the Amino-Acid Based Method (AAPM) of OLCAO. In brief, the AAPM calculates electronic, self-consistent field (scf) potentials for individual amino-acids with their neighboring amino-acids included as a boundary condition. This allows the costly scf part of the calculation to be skipped out. Additionally, the number of potentials used to describe the 1JS9 protein is also min-

imized. Results for effective charge and bond order are obtained and analyzed for Collagen and preliminary effective charge results are obtained for 1JS9. The effective charge results of the AAPM represent well those already obtained with the scf OLCAO result, but with reduced cost and preserved accuracy. The bond order results for Collagen also represent well the hydrogen bonding based on bond distances observed in experimentally-dervied images between the individual chains of the collagen triple-helix as well as the observed hydrogen bonding network.

APPROVAL PAGE

The faculty listed below, appointed by the Dean of the College of Arts and Sciences, have examined a thesis titled "Application of *ab initio* calculations to Collagen and Brome Mosaic Virus," presented by Jay Quinson Eifler, candidate for the Master of Science degree, and certify that in their opinion it is worthy of acceptance.

Supervisory Committee

Wai-Yim Ching, Ph.D., Committee Chair
Department of Physics

Paul Rulis, Ph.D.
Department of Physics

Fred Leibsle, Ph.D.
Department of Physics

CONTENTS

ILLUSTRATIONS

TABLES

ACKNOWLEDGMENTS

First, I would thank Professor W.Y. Ching for devoting himself and his time to physics research and its students. Without his supervision, focus, and interest my research efforts would not have come to fruition.

I would also like to thank my father for graciously preparing the thesis template and for providing pointers on typesetting. Also, I would like to thank Dr. Sitaram Aryal for pointers on preparing the manuscript as well as general discussions of interest.

Additionally, other members of the Electronic Structure Group have provided valuable stimulation in pursuing my research goals. It is good to belong to a research community.

The members of the thesis committee, Prof. Fred Leibsle, Prof. Paul Rulis, and Prof. Wai-Yim Ching, also deserve thanks for devoting their time and expertise in evaluating my work.

Finally, I would like to thank the University of Missouri-Kansas City and its Department of Physics and Astronomy for providing a good place for study and research in Physics.

DEDICATION

To my mother for inspiring me to be a physicist; to my father for explaining mathematics.

When I was young I wanted to be a scientist,
My mother wanted to be a physicist,
My father is a mathematician.

CHAPTER 1

INTRODUCTION

1.1 Motivation

Physics has been traditionally divided into the two practices of theory and experiment. However, with the advent of improved computing machines a third practice has emerged: computation. Today, many scientists espouse the concept of "shut-up and calculate over shut-up and let me think" as the most fruitful application of theoretical methods [1]. These calculated results can then be compared to experiment if this is possible, otherwise these results can be used as input into other calculations for measurable quantities. For example, calculated optical spectra can be compared to experiment [2] but effective charge cannot. Charge, though, can be used as input into a calculation [3] to determine the forces between two objects like proteins and then these forces can be used to partly describe the function of the protein.

Due to the large size of proteins, a more computationally efficient approach for obtaining the electronic structure is needed. First, bioscientists are primarily interested in studying large biosystems and complexes thereof. Second, for quantum mechanically based results the computational cost can be enormous, in fact, too much to be practical to bioscientists if they wished to perform such calculations themselves and to those specialized in these computations due to cost. With this in mind, we have developed a simplified approach or scheme to reduce computational cost and preserve the accuracy of the results so that the method can be practical to bioscientists. Lastly,

I should mention that these calculated results whether for charge or bonding can only be obtained from quantum mechanically based calculations [4]. Some quantum mechanical results however are semi-empirical and use a large number of parameters. The OLCAO method is an *ab initio* method since it uses very little such parameters itself and can be considered a first principles method. I should note that classically based results have been very useful in studying biomolecules for example [5, 6, 7, 8, 9].

In biophysics it is hoped that a more accurate picture of the physical properties of a biomolecule can be developed from *ab initio* quantum mechanical calculations. These methods are borrowed from solid-state and condensed matter physics but can be applied equally well to biological systems. To point this out, the OLCAO method has been successfully used to study the electronic and spectroscopic properties of many crystalline and non-crystalline materials including biomaterials and bioceramics with complex structures [10, 11, 12, 13, 14, 15, 16]. Ultimately, we want to understand the stability and formation of protein structures and the function of proteins interacting with other proteins and molecules. To do this a more accurate view of physical properties such as effective charge on atoms and the electronic potential on the surface of a molecule, as well as other properties, is necessary. Currently for many researchers, effective charge is often the isolated atom charge and electronic potential is qualitatively positive, neutral or negative on the molecular surface. But this can be improved as we will demonstrate in this thesis with our effective charge results, as well as bond order (an index of bond strength) and the more general density of states.

1.2 Summary

The computational method we are using is the Orthogonalized Linear Combination of Atomic Orbitals or OLCAO [17, 18]. This is a software suite implementing a Density Functional Theory (DFT) based method in the Local Density Approximation (LDA). This will be detailed in the methods section. The OLCAO package can calculate the electronic structure properties of biomolecules like proteins and DNA molecules and was originally designed for crystals and amorphous solids. In this work, we focus on proteins. Proteins fall into three classes: structural or fibrous, transmembrane and globular [19]. In fact most proteins are globular or close to spherical and these globular proteins are hard, compact structures that are only slightly deformable [20]. Schrödinger called proteins aperiodic crystals and solution NMR studies have revealed that the crystalline form examined in X-ray diffraction and the solvated form studied in NMR are very close. This means we can use a crystal structure for a protein and it is biologically as relevant as the protein in solution or *in vitro*, however no *in vivo* structures exist. However, there are disordered parts of proteins that do not have such stable structures too (for example the N-tails of the 1js9 protein) [21] and so are not often visible. The two proteins we have studied are collagen and 1js9. Collagen is the a frequently occurring structural protein found almost exclusively in animals. 1JS9 is a component of the protein capsid in the brome mosaic virus (BMV) and is a globular protein in structure even though it can be thought of as forming the structural shell or protein capsid.

1.3 About Proteins

Most of the general information in this subsection About Proteins and in the subsequent subsections of the introduction can be found in [19], however some specific citations are still given to other references when they are the source. Proteins are made of amino acids covalently bonded to one another in a linear sequence through what are called peptide bonds. In Figure 1, we can see an example of two amino acids coming together to form a dipeptide through a peptide bond with an eventual byproduct of water. Note the byproduct is not really water but the H and OH groups liberated from the amino acids which will reform into water molecules by reacting with other nearby water molecules. Note also that the peptide bond normally occurs through the action of a catalyst. Proteins are also described as having a primary, secondary, tertiary and quartenary structure. These structures are depicted in Figure 2, where for completeness I note that P13 is an accessory protein to the Human T cell leukemia virus type 1 (HTLV-1) which is not studied in this thesis. These structural designations are somewhat arbitrary, but the primary sequence is like (in a sense) the unfolded or denatured or randomly coiled protein but really just means the linear sequence of amino acids. The tertiary structure is the folded protein in its final three-dimensional form but retains the same linear sequence of amino acids, however the ends or terminals are never connected. The ends are called the N-terminal after the amino group end and the C-terminal after the carboxyl group end. Secondary structures have two definitions. The first is the structures that form while the protein is folding and often these secondary structures may be preserved in the tertiary or final structure. The second is the secondary structures that are present in

the tertiary structure but not necessarily present or important in the folding process. The connection to these two types of secondary structures are not yet well understood and will not be a concern of this thesis. Finally, quartenary structure results form the association of separate tertiary protein structures into one protein. These separate proteins within a protein are called (protein) domains and when they occur frequently in many proteins they are called protein modules.

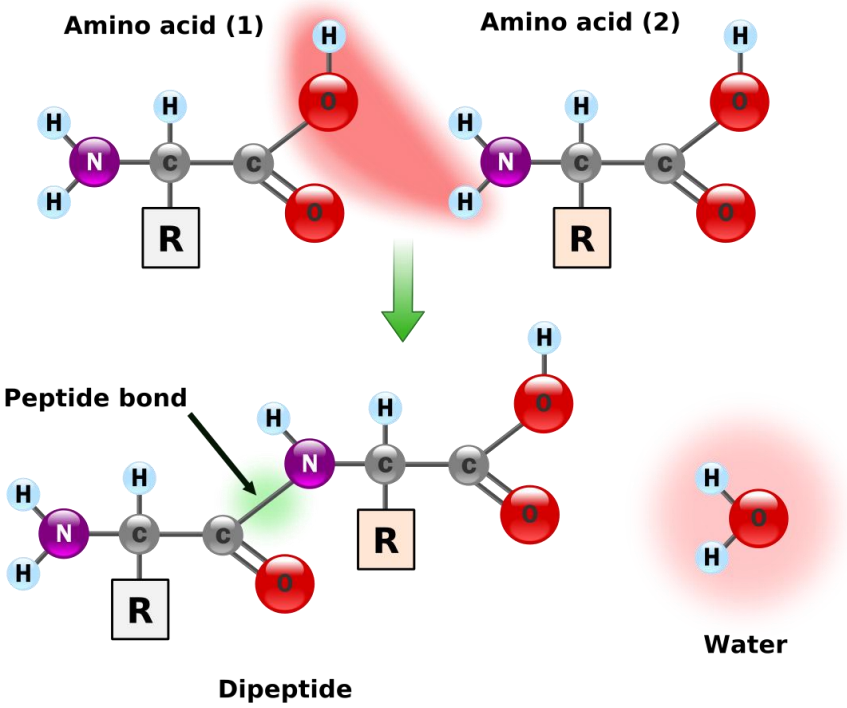

Figure 1. Two amino acids react to form a dipeptide through a peptide bond with an eventual byproduct of water. Adopted from Yassine Mrabet in WikiCommons of the amino-acid page.

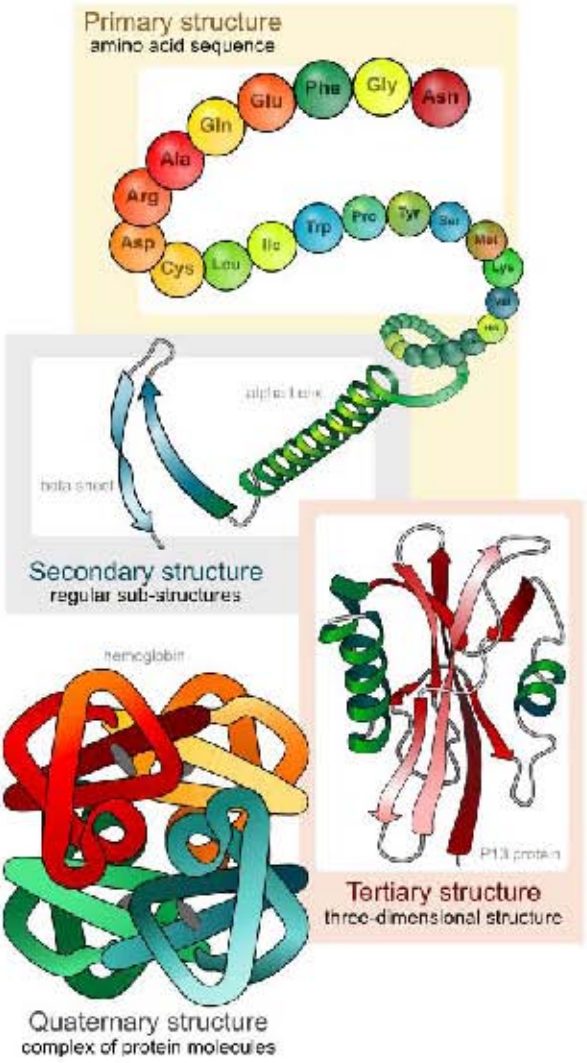

Figure 2. Depiction of the primary, secondary, tertiary, and quaternary structure of proteins. In the primary structure we see the linear chain or sequence of amino acids, in the secondary structure the geometry of smaller portions of the protein, in the tertiary structure of whole protein geometry, and in the quaternary structure the association of different smaller protein domains into a larger multi-domain protein. Adopted from LadyofHats in WikiCommons of the protein page.

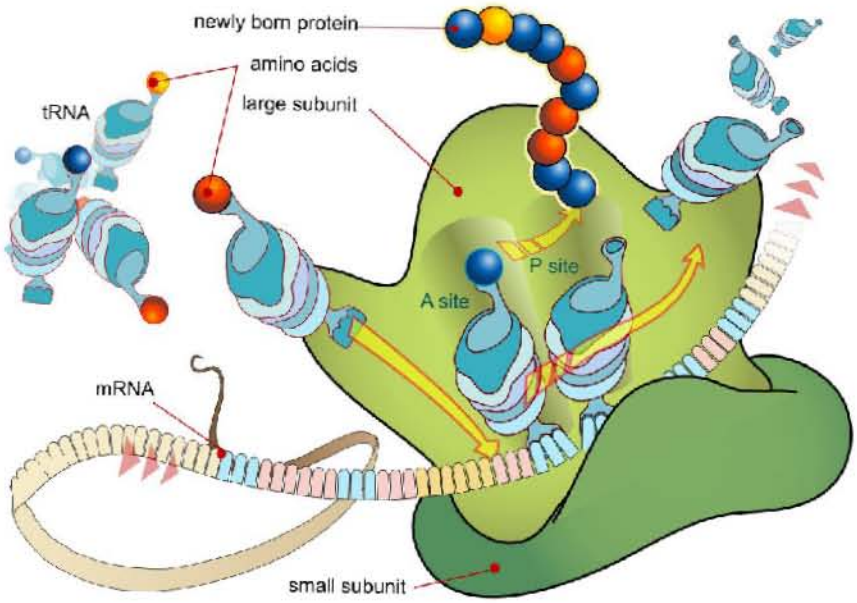

Figure 3. The synthesis of a protein with the several tRNA and mRNA shown interacting with the two subunits of a ribosome. Adopted from LadyofHats in WikiCommons of the protein page.

To describe where proteins originate we note that proteins are synthesized on ribosomes from messenger RNA and transfer RNA associated with amino acid triplets gathered from the cytoplasm as shown in Figure 3. The protein begins to fold immediately as it comes off the ribosome and is assisted in the process by other proteins, that is enzymes. And ultimately the RNAs are transcribed from the DNA which therefore encodes the protein. Although, not all DNA that encodes for proteins, called genes in molecular biology, produces a protein. So about 30,000 genes are in the human genome [22] and about 20,000 proteins are encoded for [23]. Of course, there are many more organisms on the Earth with some different genes and proteins, but ultimately the number of naturally occuring genes and proteins are finite, and often many if not most are shared. These proteins are basically called molecules which are just atoms that are covalently bound together. But the terminology is not consistently applied. The triple-helix is referred to as a molecule even though it is composed of three separately-encoded for molecules called chains or strands bonded together with hydrogen bonds. The chains are in a sense both primary and secondary structures, however, the triple-helix is thought of as a new secondary structure that replaces the old secondary structure of the individual chain molecules which were alpha chains. Secondary structures are usually referred to as alpha-helices or beta sheets formed by hydrogen bonds, though other secondary structures (like the triple-helix) exist. For example 1JS9 is described as a beta-barrel or beta jelly roll barrel but these are usually termed supersecondary structures. Note, again, these secondary structures may or may not necessarily be important in the formation of the tertiary structure.

1.4 About Amino Acids

Amino acids come in a number of forms. The amino acids occurring in nature are those that are encoded for and are also called proteinogenic amino acids. These are of 20 original or standard types or even the 21 type and 22 type (the 21st occurs is selenocysteine [24] and the 22nd is the rarely occurring pyrrolysine of Archaea [25]). They are levorotatary (L), alpha amino acids as opposed to the dextrorotatary (D) form which is not incorporated into proteins but does occur as a neurotransmitter. Amino acids are distinguished by their side chains. If the side chain is attached to the alpha carbon it is an alpha amino acid. If there are more carbon atom(s) between the amino group and the carboxylic group then beta, gamma, etc. forms can exist and the side chain is attached to the carbon for which it is named. For example, alpha amino-acids have their side-chains attached to the alpha carbon, see Figure 4, and so on. Side chains are of a variety of types with some containing only carbons and hydrogens (aliphatic) and others containing polar groups, benzene rings (aromatic) and ionizable groups (charged). In Table 1 are listed the 20 standard amino acids along with a few of their properties, including the three and one letter abbreviations that are commonly used, the earlier mentioned classifications of charged, aromatic, aliphatic and polar. Amino acids have been designed and over 40 of the types not occurring in nature exist (designed) and have been incorporated into proteins [26]. Amino acids can also be post-translationally modified. In collagen proline and lysine can be modified into hydroxyproline and hydroxylysine. These occur only in collagen. Other post-translational modifications also exist.

Table 1. Table of some standard amino acid properties

Name	3-letter	1-letter	Charged	Aromatic	Aliphatic	Polar
Alanine	Ala	A			X	
Arginine	Arg	R	Positive			X
Asparagine	Asn	N				X
Aspartic Acid	Asp	D	Negative			X
Cysteine	Cys	C				
Glutamic Acid	Glu	E	Negative			X
Glutamine	Gln	Q				X
Glycine	Gly	G			X	
Histidine	His	H	Positive	X		X
Isoleucine	Ile	I			X	
Leucine	Leu	L			X	
Lysine	Lys	K	Positive			X
Methionine	Met	M				
Phenylalanine	Phe	F		X		
Proline	Pro	P			X	
Serine	Ser	S				X
Threonine	Thr	T				X
Tryptophan	Trp	W		X		
Tyrosine	Tyr	Y		X		
Valine	Val	V			X	

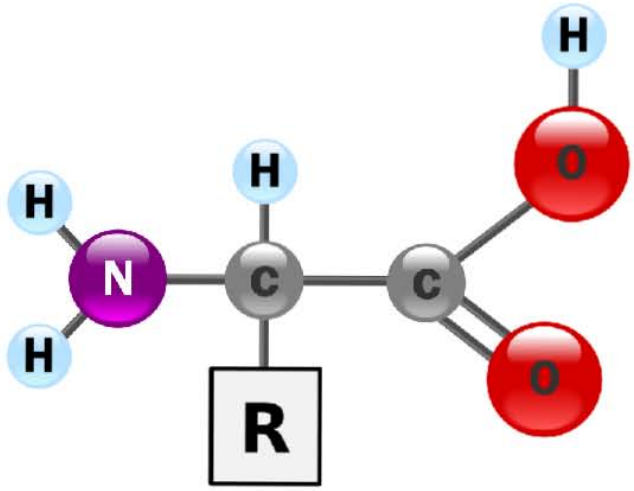

Figure 4. Structural diagram of an alpha-amino acid showing the side-chain or R-group connected to the alpha carbon which distinguishes alpha amino acids. Adopted from Yassine Mrabet in WikiCommons of the amino-acid page.

Amino acids can exist in gas, liquid or solid phases. Many studies exist in studying the properties of amino acids in gas phase, for example [27]. The amino acids have an inconsistent or unreliable melting temperature making the determination of this difficult [28]. However, amino acids do crystallize and therefore form crystal structures which are ionic solids held together by charges. A scanning electron microscope (SEM) image is shown in Figure 5. The microcrystalline L-Valine was obtained by evaporating an aqueous supersaturated solution to form the crystals. In solvated form the amino acid is not crystallized and is quite flexible. However, the properties of solvated amino acids are complicated by the nature of water. In air or water amino acids are zwitterions and are charged due to the interaction of the amino acid with its environment or ionizing elements.

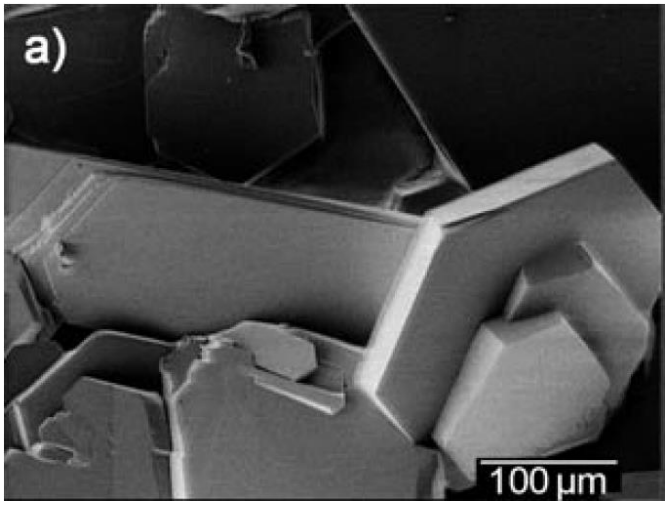

Figure 5. SEM image of microcrystalline L-valine obtained by evaporation of a supersaturated solution at 1atm. Adopted from [29].

1.5 Use of Neutral Amino Acids

In this study we examine the neutral form of the protein, meaning that all amino acids are in their neutral form and have no local charges as in the amino and carboxyl terminals and the side-chains of certain amino acids. This is due to limitations in DFT. In aqueous solution, amino acids exist in proton equilbria with their charged and neutral forms through the action of some ionizing agent like water (or air as another example). This is neglecting the influence of ions which are always present in water unless ultrapure. Isolated amino acids are charged at their amino and carboxylic ends (meaning they are zwitterionic) as well as on their side-chain if it is ionizable. However, the chance of an amino acid being charged is related to the pH and ultimately the pKa of the ionizable groups: side-chain, amino and carboxylic groups. Therefore, at a certain pH only a percentage of ionizable groups are not in their neutral form. Again, in this study all groups and the protein are in their neutral form.

OLCAO can calculate the electronic structure properties: density of states (DOS), effective charge (Q*) and bond order (ρ) as well as some others but these are the properties we will be interested in. Effective charge when compared to the isolated, neutral atom's number of electrons yields the partial charge (δ) or charge transferred (ΔQ*). Note, sometimes partial charge is negative for a gain in electrons (as we do) or positive. Also, we are referring only to the valence electrons. This partial charge on atoms can then be summed for a partial charge on the amino acids (or the whole protein even) [30] which can be used to formulate a charge distribution on the protein for the purpose of calculating electrostatic interactions between proteins. Although

in our neutral models the total net charge of the protein would be zero. We use the amino acid partial charge since the atomic resolution is not useful far enough from the molecule, such as would be used for van der Waals or electrostatic interactions. And I should note, the van der Waals and the electrostatic interactions are some of the important long-range interactions in biomolecular interactions [31, 32]. The problem of how to include the side-chain charge is usually handled by using a formal charge, that is, the amino acids are neutral and the side chain is charged such that a loss of a proton is -1 and the gain of a proton is +1 [3]. Although these formal charges are used often they are really a fiction of the chemist that is made up to fit some particular chemical model. And this is why the quantum mechanical results for partial charge are really important. Obviously, at a distance the protein with a net neutral charge would appear neutral since there are equal amounts of positive proton charges and negative electron charges. However not so obvious is that closer and closer to the protein the partial charges would dominate at the amino acid level and finally the atomic level if close enough.

The interaction of many ligands (ligands are atoms, ions, molecules or whole proteins that can bind to a site on a protein) with proteins is thought to occur at atomic resolution whereas interactions between proteins at a greater distance would have the partial charges at the level of the amino acids dominating and even further out the proteins charge is screened by other balancing charges like ions or charged molecules in the cytoplasm giving a neutral charge to that system. But in experimental methods like electrophoresis or gel electrophoresis the charge of the particle and ions can be acted on by an electric field so that the protein moves and then can

be separated based on charge as well as size (in going through an agar or matrix) even though the complex of proteins and screening charges produces a zero net charge [33, 34]. That is the protein and ions can move with some independence. Capillary electrophoresis can also be done on a chip now [4]. How to relate the partial charges of amino acids in a protein with the number of formal charges of the side-chains expected at various pH to give the charge of the protein is an outstanding problem that we will not directly address in this thesis. Adding water to the model will help and would alter the partial charges since water outcompetes with the amino acids for hydrogen bonding and thereby alters the partial charges as well as charges the whole protein. However, how to model the pH in our simulation box when adding water, such as a TIP3P model, is the outstanding problem. But likely the formal charges while useful are incorrect.

1.6 Collagen

First, no structure has been obtained for collagen since it is fibrous and this makes it apparently nearly impossible to extract. The knowledge of its structure comes from a variety of places. The primary sequences of many collagens are known so the amino acid composition and sequence are known and often this is used as a starting point for the a structural model (this is also the case for our model which will be detailed later in the Model section). In fact the GenBank was used [35] for this primary sequence. X-ray diffraction studies on the fibers of collagen (from, for example, a rat tendon) have yielded other information even if contradictory with other studies on whether the structure is 10-3 or 7-2 or even something else.

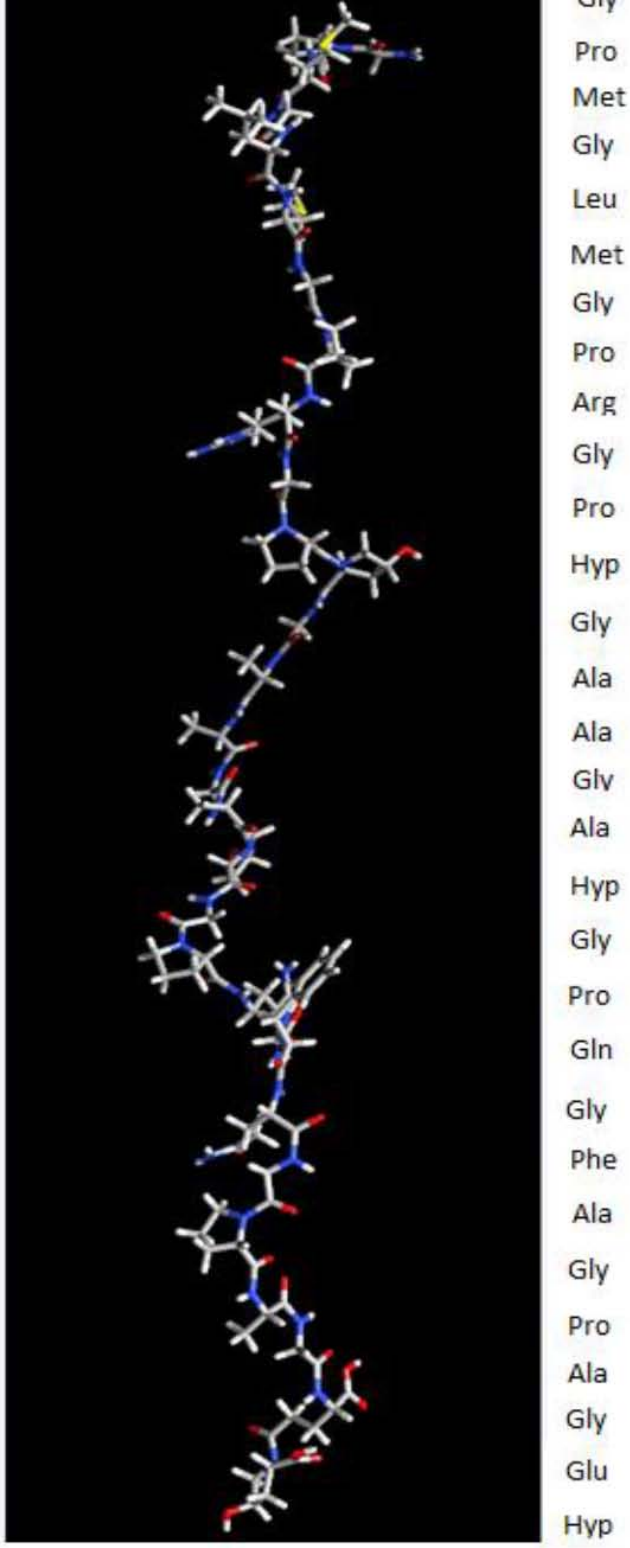

Figure 6. The $\alpha 2$(I) chain and its amino acid sequence and composition are shown above and every third amino acid is glycine.

Also, model collagen-like peptides have been synthesized and studied with X-ray diffraction to the level of atomic detail, although what the real collagen molecule is at atomic detail may be dissimilar. Since no complete structure is available structural segments can be studied instead, although some complete models exist [36]. Putting this altogether certain details are known about the collagen molecule and others consist of a few possibilities or are uncertain. One feature is that collagen consists of a repeating triplet of amino acids called a trimer of the form $(\text{Gly–X–Y})_n$ where n is the number of trimers.

Note the first amino acid or residue is always Glycine (Gly). In Figure 6, we can see the $\alpha 2$(I) chain with its amino acid sequence labelled at right. Any deviation from "every third residue is Gly" can destabilize the molecule by causing part of it to unwrap [37]. Collagen is composed of three individual strands or chains, this is known from its primary sequence and fiber diffraction studies as well as other molecular biology experiments. How these three chains are arranged together into the collagen triple-helix, as it is called, is not clear. In short, the 7-2 triple helix is most in favor [38], followed by the 10-3 triple-helix [39] (actually this article points out that whether the model should be 7-2 and 10-3 is inconclusive but this squarely puts 10-3 as still less favored in the literature due to the many articles of Okuyama who denies any mentioning of 10-3), but yet another model called close packed that has a central channel perhaps suitable for conduction has also been proposed [40]. And really there is no consensus on which triple helix model is correct. In this thesis, we study the 7-2 model. We will not discuss the close-packed structure. In the 7-2 and 10-3 models the three individual chains are twisted together.

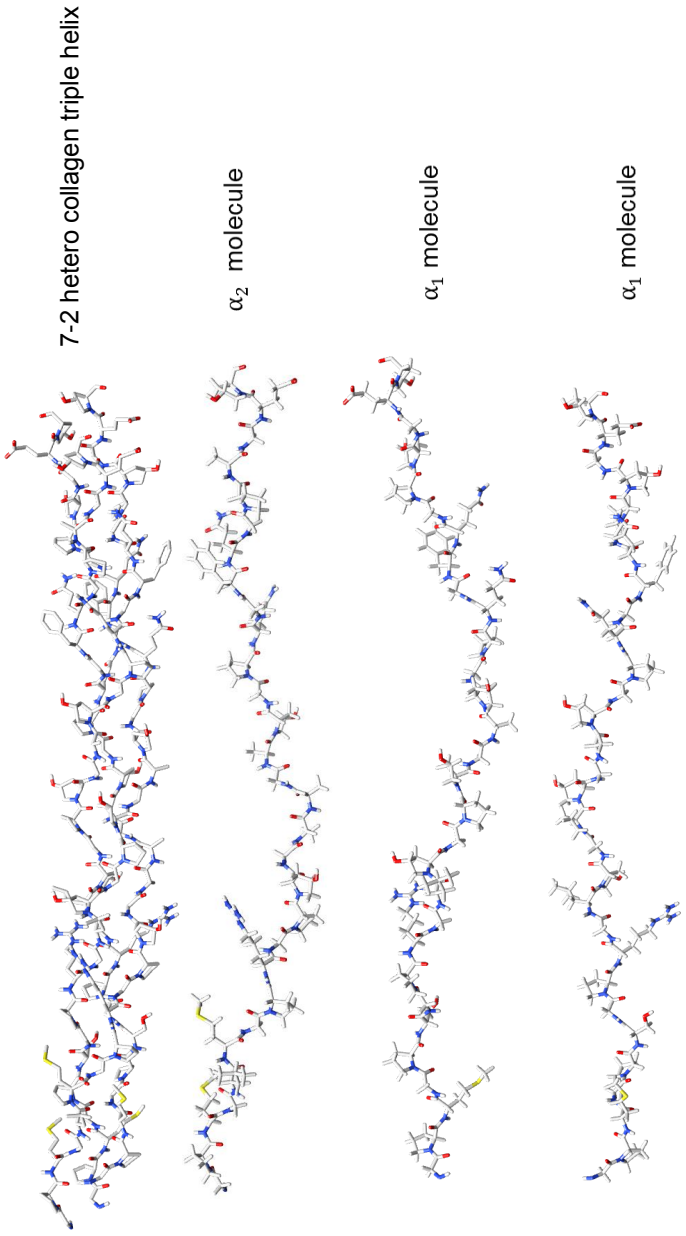

Figure 7. At top the 7-2 triple-helix is shown along with the component α-chains below with the α2(I)-chain at second top and the α1(I)-chains at next to bottom and bottom. Note not all hydrogens shown to aid viewing.

In Figure 7, the 7-2 triple-helix structure we are using is shown with the α-chains shown below. The 7-2 and 10-3 designations actually refer to X-ray diffraction patterns in the fiber studies. But in both cases the twist is related to the average number of amino acids in a turn.

1.7 Brome Mosaic Virus

The 1JS9 protein is a component of the capsid (or protein shell) for the brome mosaic virus (BMV) and is shown in Figure 8. BMV is a member of the Bromoviridae family of plant viruses and infects grasses.

Also, the capsid of the virus contains its genetic material in the form of RNA and so is called a RNA icosahedral virus. There is also a T or triangulation number designation [41] and BMV is a T=3 or truncated icosahedron. The T number is also the number of non-identical subunits. There are three subunits in a 1JS9 protein called A, B, and C each with an N-terminal tail that is disordered and interacts with the RNA within the capsid. In [21] it is reported that only the N-tail of the C subunit was visible enough to provide any structural information and this tail was modeled as a polyAlanine tail (meaning only alanines) while the other tails were not at all visible and so not included. The 1JS9 subunits assemble into capsomeres of five members (pentameric capsomeres) of only subunit A, and capsomeres of six members (hexameric capsomeres) of equal numbers of the B and C subunits. The subunits have identical amino acid sequences (except the part of the tails since they are missing in A and B subunits but these are really the same exact sequence) but different conformations since they are in different positions within the capsomeres and capsid as a whole. The capsomeres assemble into the final capsid. The details of the

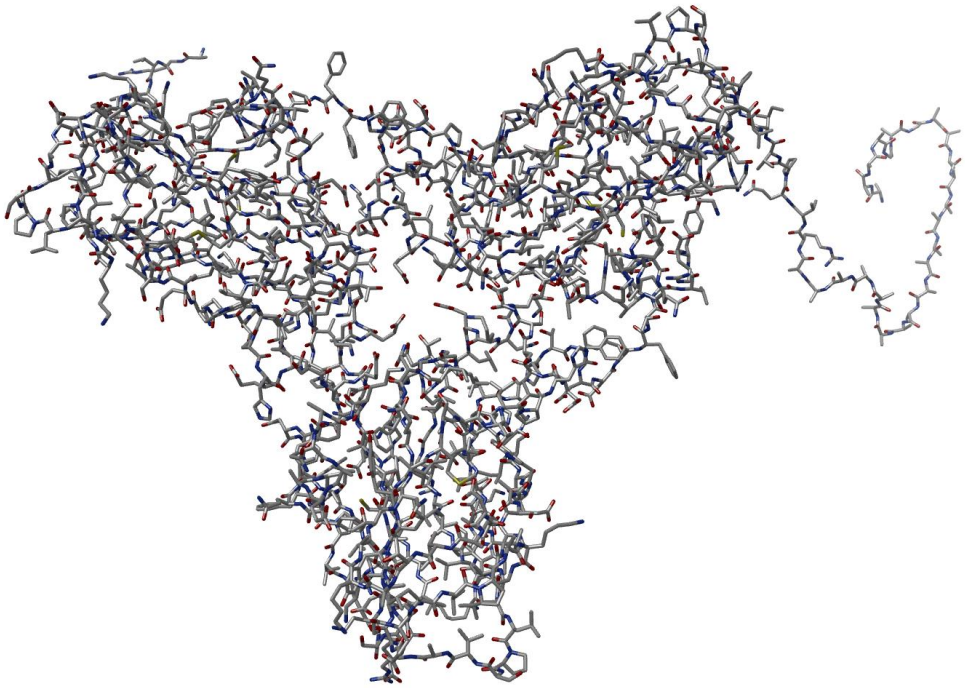

Figure 8. The atomic model of Brome Mosaic Virus composed of three subunits A, B, and C. Note not all hydrogens shown to aid in viewing the structure. Also, this is just 1/60th of the whole isocahedral capsid.

stability and assembly are still being explored [42] and we will not directly address those issues within this thesis, though our results are of some importance, we hope, to the question of how the BMV capsid is stabilized and at other times destabilizes to release its contents. Much research has been done on 1JS9 and it is hoped that it will be useful for drug or therapeutic materials to be delivered to target cells as well as in imaging [43].

CHAPTER 2

THEORETICAL BACKGROUND

Because the many-ion, many-electron Schrödinger wave equation cannot be solved explicitly it is necessary to use approximations. The first such approximation is called the Born Oppenheimer approximation and it treats the atomic nuclei as stationary or infinitely massive particles. Additionally, the nucleus is modelled as an electronic potential with which the electrons interact. This effectively reduces the Schrödinger wave equation (SWE) to a many-electron wave equation (Equation 2.1), to which more approximations shall be made.

$$\left[-\sum_{i}^{n}\frac{\hbar^2}{2m_i}\nabla_i^2-\sum_{i=1}^{n}\sum_{j=1}^{N}\frac{Z_j e}{|\mathbf{r}_i-\mathbf{R}_j|}+\sum_{i=1}^{n}\sum_{q>i}^{n}\frac{e^2}{|\mathbf{r}_i-\mathbf{r}_q|}\right]\psi(\mathbf{r}_1,\mathbf{r}_2,\ldots,\mathbf{r}_N) = E\psi(\mathbf{r}_1,\mathbf{r}_2,\ldots,\mathbf{r}_N) \qquad (2.1)$$

Historically, the Thomas-Fermi approximation was used to reduce the SWE to a tractable problem by using electron density instead of wavefunctions, thereby reducing 3N variables to just three. The problem was that the Thomas-Fermi model had an incorrect kinetic energy expression and it neglected the exchange energy arising from the Pauli exclusion principle. Note the Thomas Fermi model was the precursor to the more modern Density Functional Theory (DFT). The material in this section is drawn from [44] as well as lecture notes (W.Y. Ching, personal communication, January-May 2011).

The next approximation is the Hartree-Fock method which reduces the many-electron problem to a set of single-electron problems. The Hartree-Fock method exactly calculates the exchange energy but neglects correlations between electrons. Post Hartree-Fock methods partly overcome these limitations but do so by greatly increasing the computational cost. Hartree-Fock does not make use of electron density, rather it uses wavefunctions.

DFT includes both the exchange and correlation of electrons and reduces the computational cost considerably. The big difference between Thomas-Fermi and DFT comes from the Hohenberg-Kohn theorems and the practical means of the Kohn-Sham equations.

There are two Hohenberg-Kohn theorems. The first theorem states that the external potential is a unique functional of the electron density. The second theorem states that for a spatially-varying electron density the energy based on the electron density is greater than or equal to the true ground state energy. Now, the first theorem implies that the electron density determines the total energy and therefore it also determines the wavefunction. The second theorem implies the existence of a universal electron density functional which is of an unknown form.

The Hohenberg-Kohn theorems only give a theoretical justification for using the electron density of three variables in place of the 3N variables of the wavefunctions. The Kohn-Sham equations give a practical solution of the SWE using the theorems. Essentially, the system of N interacting electrons is reduced to a system of N non-interacting electrons. The energy functional has terms for the kinetic energy, the electron-electron potential, electron-ion potential, and a fourth term that includes the

exchange-correlation potential. In regards to the second Hohenberg-Kohn theorem, the total energy is minimized through the variation of the electron density. This results in an effective potential through which the SWE can be solved. The Kohn-Sham method is exact outside of the unknown parts of the exchange-correlation which is the main challenge in the Kohn-Sham approach to DFT.

There are various approximations to estimate the exchange-correlation energy functional. One such approximation is known as the local density approximation (LDA) which treats the electron as if it were in a homogenous electron gas. Surprisingly, the LDA is quite accurate for certain systems, but this is because of an overestimation of the exchange part and an underestimation of the correlation part such that they partially cancel. Although reasonably accurate, the LDA suffers when the electron density changes rapidly as in atoms and molecules, but despite this it is still used extensively in solid-state physics.

CHAPTER 3

METHOD

3.1 Selection of Method

We use the Orthgonalized Linear Combination of Atomic Oribitals (OLCAO) method in the thesis. This is a density functional theory based, all electron method applicable to crystals, amorphous solids, defect containing solids, liquids, molecules, etc. While there are a number of software packages for calculating quantum-mechanically based properties, some offer such accuracy that the calculation of properties, such as charge, are not feasible for large macromolecules. The OLCAO method is efficient for calculating the properties of large molecules and is one reason we choose it. Another reason we choose the OLCAO method is that it uses atomic orbitals and defining and calculating atom-specific charges is much easier than in other methods that do not use atomic orbitals as a basis, for example those that use a plane-wave basis. In this thesis, we use the method as a cluster-like method for biomolecules by placing the system of interest in the center of a simulation cell that is sufficiently large to eliminate interactions with the neighboring cell. The OLCAO method is an extension of the LCAO method wherein the core orbitals are orthogonalized against the valence orbitals to reduce the dimension of the secular equation. The justification for this is that the valence shells are the main determining factor of the chemical properties in the atom or molecule. The OLCAO method is explained, though not definitively, in [17, 18].

3.2 Overall Description

In the OLCAO method the solid-state wavefunction is expanded in terms of the Bloch functions as shown in Equation 3.1.

$$\Psi_{i\gamma}(\mathbf{r}) = \sum_{i,\gamma} C^n_{i\gamma}(\mathbf{k}) b_{i\gamma}(\mathbf{k}, \mathbf{r}) \tag{3.1}$$

In Equation 3.1 γ represents non-equivalent atomic sites and i is the orbital quantum number (l, m). However, in using the cluster method we will always use $\mathbf{k} = 0$ or the γ point which is in the center of the Brillioun zone (the Wigner-Seitz cell in reciprocal space). One k-point is sufficient because the cell size for a large molecule is of a comensurately large size. The Bloch functions $b_{i\gamma}$ in turn are an expansion in atomic orbitals $u_i(\mathbf{r})$ as shown in Equation 3.2.

$$b_{nk} = \frac{1}{\sqrt{N}} \sum_v e^{i\mathbf{kr}} u_i(\mathbf{r} - \mathbf{t}_\gamma - \mathbf{R}_v) \tag{3.2}$$

In Equation 3.2 t_γ is the position of the γ^{th} atom in the cell and R_v represents the lattice vector. The atomic orbitals have two parts: radial and angular. The angular part is given by spherical harmonics $Y_{lm}(\theta, \phi)$, and the radial part is expanded in terms of Gaussian type orbitals (GTOs) as shown in Equation 3.3.

$$u_i(\mathbf{r}) = \left[\sum_j^N C_j r^{n-1} e^{(-\alpha_j r^2)} \right] \cdot Y_{lm}(\theta, \phi) \tag{3.3}$$

In Equation 3.3 i represents the quantum numbers n, l, and m. Also, N is the number of GTOs and the set of α_j are predefined and usually guided by past experience and are distributed in geometric series ranging from α_{min} to α_{max}. Additionally,

the wave functions of the same atoms can share the same set of exponentials (α_j). Using GTOs for atomic orbitals puts all of the multi-center interaction integrals into an analytic form for faster and easier calculation. Also, the charge density and the one-electron potential are expressed as atom-centered Gaussian functions. The Kohn-Sham equation is solved in a self-consistent iterative cycle using the secular form of the equation shown below.

$$|H_{i\gamma,j\delta}(\mathbf{k}) - S_{i\gamma,j\delta}(\mathbf{k})E(\mathbf{k})| = 0 \tag{3.4}$$

3.3 Core Orthogonalization

One of the essential parts of OLCAO is that a mathematical procedure is applied to the interaction integral matrices such that the core orbitals are orthgonalized against the valence orbitals. Then, the core orbitals can be eliminated from the secular form of the Kohn-Sham equation. In the OLCAO method, the core orbitals are identified according to a rule-of-thumb that states that core orbitals are those orbitals that are deeper than an oxygen 2s orbital.

3.4 Basis Sets in OLCAO

Another feature of the OLCAO method is that one may choose the atomic orbital basis set with which to expand the solid state wave function, although there must be some cutoff and an overextended basis is not necessarily better. The Minimal Basis (MB) consists of the core orbitals (which are later orthogonalized against the valence orbitals) and the occupied valence orbitals and it is used for calculating the effective charge (Q*) and the bond order (ρ) using the Mulliken scheme. This scheme is used for the MB set because effective charge and bond order are relatively localized

properties. The Full Basis (FB) set consists of an extra shell beyond the MB and it is used for calculating, in periodic structures, the band gap and for crystals and other structures (such as proteins) the density of states (DOS). The third choice of basis is the extended basis (EB) which has an extra shell over the FB set and is used for optical and spectroscopic properties.

3.5 Calculating Properties within the OLCAO method

In this study we will use the density of states, partial density of states (PDOS), effective charge, and bond order calculations. The density of states is a count of the number of states available for an electron to occupy within a given energy range and it is shown below in Equation 3.5:

$$\begin{aligned} G(E) =& \frac{\Omega}{(2\pi)^3} \frac{d}{dE} \int_{BZ} d\mathbf{k} \\ =& \frac{\Omega}{(2\pi)^3} \int \frac{dS}{|\nabla E|} \end{aligned} \tag{3.5}$$

where Ω is volume of the unit cell and the integral is over the constant energy surface in the brillouin zone (BZ). More specifically, the total density of states (TDOS) can be resolved into its partial compenents (PDOS) that represent any subgroup of the total system. For example, in a protein we can obtain the PDOS of individual amino acids, atoms, and even atomic orbitals. The effective charge is the charge transferred to or from a particular atom from everything else in the system, and therefore it represents the (potentially fractional) number of electrons surrounding the nucleus. To calculate the effective charge, first we need an expression for the fractional charge which is given by Equations 3.6 and 3.7:

$$1 = \int |\psi_{n\mathbf{k}}(\mathbf{r})|^2 d\mathbf{r} = \sum_{i\alpha} \rho_{i\alpha}^{n\mathbf{k}} \tag{3.6}$$

$$\rho_{i\alpha}^{n\mathbf{k}} = \sum_{j\beta} C_{i\alpha}^{nk^*} C_{i\alpha}^{nk} S_{i\alpha,j\beta} \tag{3.7}$$

where, the $i\alpha$ is the ith orbital of the α^{th} atom. Now, the effective charge (Q*) can be found from Equation 3.8.

$$\begin{aligned} Q* &= \sum_{n\mathbf{k},occ} \sum_{i} \rho_{i\alpha}^{n\mathbf{k}} \\ &= \sum_{n\mathbf{k},occ} \sum_{i} \sum_{j\beta} C_{i\alpha}^{nk^*} C_{i\alpha}^{nk} S_{i\alpha,j\beta} \end{aligned} \tag{3.8}$$

The Q*, when subtracted from the neutral, isolated-atom charge, is sometimes called the partial charge (δ) in biology while in physics it is called charge transfer (ΔQ*). Note that the charge transfer is not specified as being between two specific atoms. Bond order (ρ) is an index of the bond strength between two atoms and, when accumlated across an entire system, it may be used to gauge the relative bond strength within the system studied. The bond order is given by Equation 3.9.

$$\rho_{\alpha\beta} = \sum_{n\mathbf{k},occ} \sum_{ij} C_{i\alpha}^{nk^*} C_{i\alpha}^{nk} S_{i\alpha,j\beta} \tag{3.9}$$

Positive bond orders represent bonding. These bonds can be classified into covalent and hydrogen bonds based on the bond order value (note: this is between two atoms α and β).

CHAPTER 4

MODELS FOR COLLAGEN AND THE AMINO ACID POTENTIAL METHOD

4.1 Models used for Collagen Triple-Helix and Brome Mosaic Virus

To start an *ab initio* calculation it is necessary to have good structural models. In this thesis, the triple-helix model of collagen was constructed by Dr. Simon Vesentini using the TripleHelicalBuilder program or its predecessors [45, 46, 47]. The method and results contained in this thesis were published in [48] for collagen. The 1JS9 results are preliminary and unpublished. Note that the model used for this research was dry. The triple-helix contains 90 amino acids or 30 trimers in the (Gly–X–Y) form. Each chain of the triple-helix contains 30 amino acids or 10 trimers. The whole structure is about 85 Å long and represents only a structural segment of the whole triple-helix molecule which is around 3000 Å long. The triple-helix model used is about 15 Å in diameter.

We put our triple-helix model into a 100 Å x 100 Å x 30 Å box so that there are at least 9 Å of space between the adjacent molecules in neighboring cells to avoid any interactions between molecules. The triple-helix contains 1135 atoms and 3246 valence electrons. Additonally, six hydrogen atoms (one for each end of the three α-chains) were added to the ends of the triple-helix molecule to eliminate the dangling bonds that are present because the model is only a structural segment.

The model we use for the protein component of the brome mosaic virus (BMV) comes from the protein database entry 1JS9 (various details are elaborated in the

pdb file itself) and was first reported in ref [21]. This means that the 1JS9 model is not a relaxed structure but is instead the best available structure that we currently have access to. 1JS9 consists of three subunits called A, B, and C. Each of these subunits was separately calculated because the whole 1JS9 is still beyond our current computational capacity. Each subunit was placed in a box of approximately 100 Å x 100 Å x 100 Å, although the exact dimensions varied for each subunit, especially for the C subunit with its long N-terminal tail.

Each subunit has about 2300-2800 atoms and the whole 1JS9 is about 7500 atoms. Each subunit has an identical amino acid sequence of 189 residues. As stated in the introduction, part of the N-terminal tails of the A and B subunits are missing because these were not visible from the experimental imaging. Residues 41-189 are present in the A subunit, residues 25-189 are present in the B subunit, and all 189 residues are present in the C subunit although the N-tail is actually modelled as a polyalanine tail. The same spacing between adjacent cells also applies here as it did earlier for the triple-helix.

4.2 Introduction to Amino Acid Potential Method (AAPM)

When site-specific atomic potentials are carefully constructed in a simple system they can be transferred to a more complex system. With this in mind to simplify the calculation of the properties for a large protein, potentials for individual amino acids were developed independently of the bulk protein, the results of which have been published in ref [48]. One of the difficulities encountered in doing this is that an amino acid in a protein is covalently bonded to its neighboring amino acids. That is, for the N-terminal and C-terminal, of the protein's chain of residues is there only

one amino acid to which another amino acid is covalently bonded. The exception to this is with disulfide bonds between cysteine amino acids (two cysteines bonded together by a disulfide bridge is known as a cystine amino acid), however in the two proteins we are studying (collagen and 1JS9) there are no disulfide bridges. There is one cysteine residue in each of the three subunits of 1JS9 but these have no disulfide bridges [49].

An attempt was made to replace the amino acids to which an amino acid is bonded to with hydrogens only using an isolated amino acid model. This is the least perturbing replacement that can be done. However this produced poor results in general. Also, using only the potential of an isolated amino acid was not much better. Arbitarily attaching two glycine amino acids to the end of an isolated amino acid model produced much better results, but still unsatisfactory as some N, C, and O atoms were not well represented. Even extracting the actual amino acid in the protein and attaching hydrogens to it produced similar results to simply using the isolated amino acid model.

Only by using the amino acid from the protein itself with the adjacent amino acids included (which themselves were hydrogen terminated) were we able to produce a satisfactory result. In fact, the AAPM results were better than a self-consistent reduce level 3 calculation. Note that in this approach all of the atoms in the amino acid had there own unique potential, this is called an all-type non-self consistent calculation. The AAPM can also be extended for use with disulfide bonds and the inclusion of a solvent, ions, and other smaller ligands in a manner similar to including the adjacent amino-acids.

More accurate, quantitative results can be obtained from state-of-the-art *ab-initio* quantum mechanical calculations but they are computationally costly. In our own OLCAO method, self-consistent (SCF) calculations on large proteins produce good results but are too costly. Therefore, we have developed a simplified scheme using the well-known result that proteins are composed of a linear sequence of amino acids of which there are twenty but with many conformations of the side-chains or χ angles as well as the backbone angles or ψ, ϕ and ω (the last of which is usually about 180 degress but does vary in real structures) angles.

In our simplified scheme, SCF calculations were done on each amino-acid in collagen to obtain a SCF atom-resolved electronic potential for each amino-acid. This scheme is a "divide and conquer" approach that is designed to lessen the computational cost by using SCF calculations only on individual amino acids with some reasonable boundary conditions. Subsequently, a non-self-consistent (non-scf) calculation is performed on the whole protein, in this case collagen or 1JS9 subunits, using the SCF potentials from individual amino acids. This calculation of the SCF potential for each amino-acid we call the Amino-acid based Potential Method (AAPM) from which we construct an amino-acid database for use in calculating the potential of large proteins.

To emphasize again, all proteins are made of a linear sequence of amino-acids where each amino-acid is connected to a preceding one or a subsequent amino-acid, except the first and last amino-acid (which are terminal) which are bonded to only one amino-acid. To base the calculation on more realistic boundary conditions, the amino-acids that are adjacent to the amino-acid of interest were included in the

AAPMs scf calculation for the amino-acids composing the protein.

4.3 Automation of AAPM

In the beginning, the AAPM was applied to smaller peptides of the collagen molecule by hand. This is a tedious process for a large protein and in fact would tend to produce too many errors and require too much time to construct to be useful. Therefore, the process of applying the AAPM to smaller peptides by hand is useful primarily for familiarizing oneself with the process of constructing the proteins potential from the amino acid potentials, and this leads easily enough into developing programs to construct the protein potential.

The programs for automating the AAPM were written in Python 3.1.1. Also, the molecular modelling package UCSF Chimera was used [50]. In fact the developers of Chimera wish the following to be included in referencing them: the molecular graphics and some of the analyses were done with UCSF Chimera; Chimera is developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (supported by NIGMS P41-GM103311). I choose Chimera because it rendered molecular images quickly and it was free and open source. However, there are numerous such packages.

Because the programs for automating the AAPM are not well documented or written for others to easily modify the programs themselves are not included in this thesis but they can me made available by the author upon request. Eventually this will be otherwise. The procedures in the appendices instruct how to actually use the scripts and may be somewhat technical and difficult to follow without actually processing a protein into its final result. However, these instructions are invalu-

able and illustrate the complexities involved. And finally note that almost all large biomolecules have at least one error in the data. I will not cite specific examples I have encountered in others published work but I have encountered them in almost all datasets.

CHAPTER 5

COLLAGEN RESULTS

5.1 Summary of Collagen Results

The AAPM was used to perform a non-SCF calculation to obtain effective charge (Q*) and total density of states (TDOS) for the collagen model (as well as 1JS9 subunits A, B and C later). All of these results for collagen are published in [48]. Then the calculated Q* for the collagen chains was used to assess the success of the method by comparing the results of the full all-type SCF calculations, a known good result, the non-SCF calculation that uses AAPMs, and the non-SCF calculation using the OLCAO database for atomic potentials which are simply the potentials obtained for each atom by single isolated atom SCF calculations. Qualitatively, the non-SCF calculation using AAPMs (non-SCF amino) for Q* was much closer to the SCF result than the non-SCF calculation using the atomic potentials (non-SCF atomic) which will be detailed later in this chapter.

When analyzing Q*, the total charge transfers for the atoms within an amino-acid when summed up for the total amino-acid Q*, resulted in almost no total charge transfer for the amino-acid. This implied that very little charge (one fiftieth to one hundredth of an electron) was transferred between amino-acids in the sequence except for at the terminal ends where about four times the charge was transferred. This extra charge accumlates on the ends from there being no additional amino acid for the terminal amino acid to transfer charge to or from. However, transfers due to

hydrogen bonds from amino acids not covalently bonded to the central amino acid of interest were not considered in the AAPM models because only the neighboring amino acids which are covalently bonded are included and not other surrounding amino acids which might form HBs. But, when this is done for the triple-helix and the subunits of 1JS9 we do get significant charge transfers for the amino acids. This may be since more amino acids surround each other and have more hydrogen bonding between them, however this has not yet been specifically analyzed.

For the individual collagen molecules, the computational cost was reduced by an order of magnitude. For the triple-helix this reduction would be much greater when compared to a costly SCF calculation for the triple-helix because the number of atoms are approximately tripled. For the 1JS9 subunits the reduction is even greater. This method then allows for the computationally feasible calculation of quantum-mechanically based electronic structure properties of proteins up to approximately 200 amino acids. Eventual extension to 500-700 amino acids should also be possible when the "reduce potential method" (see appendices) is extended and tested for structurally averaged potentials in the amino acid database.

5.2 Validation of the AAPM using Collagen

Q* and TDOS were calculated for the individual collagen molecules, as well as the entire triple-helix, using the amino-acid database, that is the AAPM. In Figure 9, the results for Q* belonging to molecule 1 are shown in detail and compared to the scf and the non-scf atomic results. Qualitatively, the non-scf amino result is seen to be much closer to the scf result than the non-scf atomic result. Note, that the dotted lines are only present to guide the eye to which atom's Q* result is next, since

otherwise the Q* result for some atoms are easily skipped over. Also, atoms are shown in the order they appear in the molecules amino-acid sequence.

Figure 9. Comparison of effective charges Q* (electrons) for each atom in chain α2(I) between: (1) Non-SCF calculation using atomic basis (green circles), (2) Non-SCF calculation using AAPM (red triangles), (3) Full SCF calculation (black squares).

Q* tends to fall into a few narrow ranges of values in each plot. For example, in the plot of Q*N, most values for scf and non-scf amino results fall into a range centered around 5.5 electrons. In [20], the peptide backbone has partial charges of +0.4 for the carboxylic carbon, -0.4 for the carboxylic oxygen, -0.2 for the amine nitrogen, +0.2 for the amine nitrogen's hydrogen, and zero for the alpha carbons giving an overall zero net charge to the peptide backbone. However, this is not a quantum mechanical result but shows our results are in the ballpark of theirs with significant enough deviations due to the local environment. Also in [51] where the partial charges are intended for classical molecular mechanics and dynamics modeling with water, the partial charges range from about -0.8 to +0.8 and this is again a range within which our partial charges also fit, but these are actually fitted charges and not Mulliken charges. I might note too that the Mulliken charges are not used in molecular dynamics since the partial charges are derived from amino acid triplets rather than the whole protein and are conformationally dependent.

In fact, for N, C and O most atoms gain about half an electron worth of charge. Also, several larger values for Q* can be seen in the Q*N plot: the first nitrogen which belongs to the N-terminal amino-acid Gly, two nitrogen atoms close together at about the tenth atom number belong to the side-chain of Arg, and two later peaks at about 25th and 30th atom number belong to the side-chains of Gln. Now, these are all amino acids with polar or potentially charged side-chains and would all be on the surface of the chains of the triple-helix since side-chains are projected out from the axis of the chain in general. And these represent the most highly partially charged electrophiles on the surface and would most likely form HBs with water or ligands or

the other chains of the triple-helix as will be seen later. Similar ranges of values and larger values either high or low can be characterized in the other plots for molecule 1 in Figure 9, as well as for the other two molecules of the triple-helix which are not shown. Note also when compared to the neutral atom charge that generally, as expected, H gave up charge, N and O gained charge, C both gained and lost charge, and S lost very little charge since S is not very acidic in the Met side-chain (S not shown).

With the observations of ranges of values and larger values of Q* for different atoms, atoms can be classified into bins of Q* values and therefore atoms of a specific element can be categorized into types based on their effective charge. This could also be compared to what amino-acid the element belongs to, but there will be variations depending on the local configuration of atoms, however we have not yet bothered with this kind of analysis. Figure 9 shows the validity of the simplified scheme in preserving the accuracy of the scf individual molecule results. Only in the Q*O figure is there much of a visible difference between the scf and non-scf results. Quantitatively, the gap is at its greatest about a twentieth of an electron charge implying that two digits can be included in the non-scf amino Q* results.

Figure 10. Comparison of effective charges Q* (electrons) from the SCF calculation of 3 individual chains and the triple-helix using AAPM. Red circles represent triple-helix values and black x's represent values from chain 1 (α2), 2 (α1) or 3 (α1) with atoms aligned with those of the triple-helix.

5.3 Effective charge results for the Collagen Triple-helix

Results for the non-scf amino Q* are compared between the triple-helix and individual molecules in Figure 10. The individual molecule results are aligned so that atoms of a molecule correspond correctly to atoms of the triple-helix. Red circles are for the triple-helix and black crosses are for the individual molecules. Since the non-scf amino Q* results for the individual molecules were shown in Figure 9 to be good approximations of the scf result and the non-scf Q* results for these individual molecules are repeated in Figure 10 and aligned with the triple-helix result to match atoms, we can see that the triple-helix result is somewhat qualitatively validated as an accurate calculation without performing the scf calculation on the whole protein. This partial validation is true, inasmuch, that the general outline of the Q* results is duplicated but the gap relates to the interaction between the molecules, so there could be some errors that have been washed out. Remember we mentioned comparing our AAPM molecule 1 result to a scf level 3 calculation (this level 3 is a reduction of the number of potential types used to describe the protein as opposed to all-type where every atom has a unique potential) as more accurate, and this means very likely the difference can be related to the interactions of the chains rather than error in the method producing the gap. Either way, the individual molecule results would not well represent the triple-helix and therefore shows the necessity of doing the whole protein calculation.

A small difference is noticeable between Q* for the triple-helix and Q* for the individual collagen molecules. The same ranges of values and larger values for Q* can be seen more easily and collectively in Figure 10 than Figure 9. Note, that the

non-scf amino result (red triangles) for Figure 9 is repeated in Figure 10 as the first third of each elements Q* (black crosses) but are pushed closer together. Sulfur is also shown in Figure 10.

To summarize the differences of Q* between the triple-helix and the individual molecules, we note which elements gain more or less charge or lose more or less charge. N gains less charge in the triple-helix Q* result when compared to the individual molecule results. C atoms that gain charge, gain less charge in the triple-helix result. C atoms that lose charge, lose more charge in the triple-helix. O gains more charge in the triple-helix. H loses less charge in the triple-helix.

These gains and losses of effective charge can also be seen for the different ranges of values of Q* for an element if one wanted. Since the individual molecules are not covalently linked, hydrogen bonding and other interactions (for collagen there was only one such other interaction) can be related to the differences in Q* seen in Figure 10 between the triple-helix and the individual molecules. Notice also that only O and H gain charge when comparing the triple-helix and molecule results and that N and C lose charge.

The individual molecules retain the structure they would have in the triple-helix and so do not represent the original α-helix structure one might think of the individual molecules or chains as having before being incorporated into the triple-helix. There are only intra-molecular interactions, that is hydrogen bonds, to relate to the Q* for the individual molecules. In the triple-helix, there are inter-molecular interactions and also a different set of intra-molecular interactions relating to Q* results.

The differences in Q* for the individual molecules and the triple-helix come in part from the change in inter-molecular and intra-molecular interactions nearly all of which can be thought of as hydrogen bonds. The hydrogen bonds and other interactions are thought to stabilize the individual molecules through intra-molecular interactions. One can then conjecture that the triple-helix is stabilized through hydrogen bonds and any other interactions that are inter-molecular or intra-molecular. These results can be seen partly as the differences of Q* between the triple-helix and individual molecule results in Figure 10 and could be thought of as holding the triple-helix together and providing its strength against extension (or tensile strength), although this can and would certainly be disputed as explaining collagen's mechanical properties. To make clear, the effective charge is the charge transferred to an atom from potentially all other atoms in the protein, and the bond order is a measure of bond strength between two atoms. So, the bond order cannot give us the effective charge or vice versa, but we can relate bond orders and effective charges or changes in effective charges in different but related calculations.

Figure 11. Comparison of total density of states for triple-helix and the sum of three individual molecules calculated using the AAPM.

5.4 Partial Charge of the Collagen Triple-helix

Q* was also converted to charge transferred (ΔQ*) or often in biology and chemistry called partial charge (δ). In Figure 12, the partial charges on the atoms of the collagen triple-helix are shown as colored-graded sphere, and this is one of the main results of this research and thesis. The size of the spheres is based on the covalent radii of the atoms which are in angstroms and are: 0.77 for C, 0.75 for N, 0.73 for O, 0.37 for H, and 1.02 for S. These covalent radii represent a charge surface (spherical) rather than a point. Later for 1JS9 we will use a solvent excluded surface instead to represent the interface with water which is not just a set of non-overlapping spheres of charge. Partial charges ranged from -0.88 (darkest blue) to 0.88 (darkest red) electrons and can be compared to the results in Figure 9 and Figure 10 by knowing the neutral atom charge to convert from the effective charge to the partial charge, but we haven't bothered too much to make this connection terribly explicit.

The colors for partial charge are coded so that white represents a neutral or nearly so charge. Pink and light blue are slightly positive or negative. Bluer and redder spheres are then more positively or negatively charged. And so dark red and dark blue are the most highly partially charged atoms. The ranges of values and larger values for Q* can be compared to the partial charge color coding if one wants.

From the view in Figure 12 we might suggest we can see what charges a test charge would see electrostatically (in a classical sense), as opposed to other interactions of less range, such as van der Waals or hydrogen bonding. We might note, a slightly bluer appearance to the left end and redder right end and a middle of about equal blue and red mixture and this might be due to the end amino-acids having

more charge transferred to them than amino-acids that are not terminal although this is slight. More importantly, we can see the distribution of partial charges within the structure of the triple-helix and that there is a mixture of colors so that there are no large groups of red or blue parts of the helix (other than the slightly bluer and redder ends).

5.5 Triple-helix Total Density of States

TDOS for the triple-helix and the sum of the TDOS of the three individual collagen molecules are shown in Figure 11 for the non-scf amino result. The red, thick line is for the triple-helix and the black, thin line is for the sum of the individual molecules. The spectra contain all the information of the electronic structure and can be resolved into PDOS for molecule, trimer, amino-acid, atom group, atom, and orbital. At energy levels below the top of the valance band (0 eV), the occupied states are fairly well represented by the sum of molecules only, but at higher unoccupied states the triple-helix and sum of molecules differ. Although in Figure 10 this difference is related to the interaction of the chains of the triple-helix and so the difference seen in the occupied states for the triple-helix and sum of molecules comes from the interaction of the chains in the triple-helix though this difference looks small.

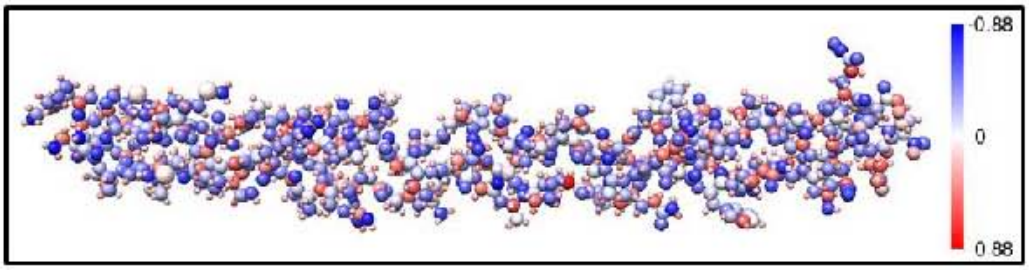

Figure 12. Partial charges on each atom in the 7-2 heterostructural model using. The size of the atoms is based on covalent radii used to compute an available surface area for the atom's partial charge. Partial charge is negative (positive) for a gain (loss) in fractional electron charge and colored blue (red). White color indicates no charge transfer for the atom.

5.6 Bond Order Analysis for Collagen Triple-Helix

In Figure 13 (triple-helix HBs), we see the hydrogen bonds (HB) above 0.002 bond order value displayed as hatched red-lines and green-lines between atoms in a stick model of the collagen triple-helix. The thicker, hatched green lines are the intermolecular HBs which are of more interest. Usually only certain of the existent HBs in a structure are discussed as being HBs though a network of many hydrogen bonds exists. This is usually known from the distances of possible acceptor and donor atoms and the bonds cannot really be visualized experimentally. In Figure 13 we see HBs that are also intramolecular both within an amino acid and between amino acids as redlines. Note the our HBs are not determined by bond distances but are based on a quantum mechanically calculated result.

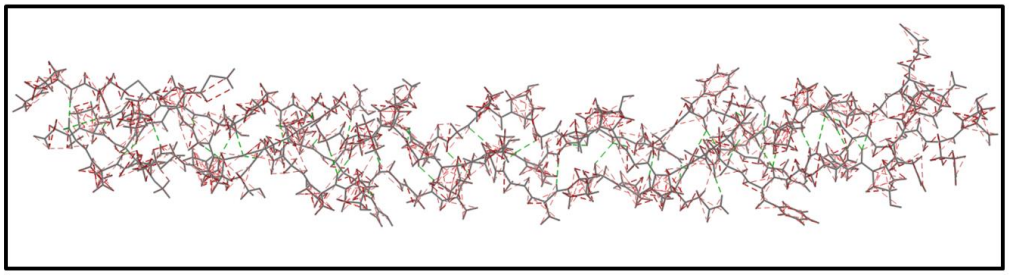

Figure 13. Sketch of H-bonds within the 7-2 heterostructural model. Green dashed lines for intermolecular H-bonding and red dashed line for the intra-molecular H bonding.

In Figure 14 is a graph of HBs classified into various acceptor/donor groups and plotted as bond length versus bond order value (so again we classify the bonds not just based on bond distance). Also the same red and green classification of the intermolecular and intramolecular HBs is used as seen in Figure 13. The graph goes up to 3.5 angstroms. As can be seen, the intermolecular HBs are all O—H except one N—H. The intramolecular HBs include O—H, N—H, as well as, C—H and H—H (not usually classified as a HB), however the vast majority of these are within amino acids (also not usually classfied as HBs) and the small remaining HBs would be intramolecular HBs between amino acids to clarify. The stronger (green) HBs may provide the necessary cohesion of the molecules and help to understand the stability and tensile properties of the collagen molecule as already mentioned earlier. It should be noted that the presence of water would alter the HBs and their bond order values in the molecule and so these results represent the dry collagen molecule.

In Figure 15 we see the BO shown as an associated bar height and the color of the bar indicating whether the bond is between backbone of chains or between a chain backbone and an amino-acid side-chain of another chain. As mentioned earlier, the most partially charged electrophiles are on the surface of the chains and would be available for HBs. And we in fact see several of these HBs as red bars from these kind of electrophiles. The BO values remain relatively constant along the length of the molecule and the distribution is fairly even for HBs between the chains. However, in the central region on the molecule's length there is a region where the only strong intermolecular HBs are between A and B. This indicates that the C chain may have a more flexible conformation.

Figure 14. Distribution of the calculated BO values for H-bonds in the 7-2 heterostructural model. Different types of the H-bonding are marked and colored as indicated.

Figure 15. Calculated H-bond location and relative strength between pairs of chains.

As well, it is known the the A chain is found in the hetero-trimeric form and leads to greater strength in collagen and this could indicate why there are more bonds between A and the other chains than just between B and C. Also, there are fewer HBs at the end on the right but this may just relate to the local conformation and ending of the molecule necessarily. These results shown in Figure 15 are consistent with the interpretation of the role of interchain H-bonding between different amino acids found in [39].

CHAPTER 6

BROME MOSAIC VIRUS RESULTS

6.1 Brome Mosaic Virus Partial Charges

The preliminary and unpublished results for 1JS9 include the partial charges on the amino acids for each of the subunits A, B, and C and are pictured together in a color graded scale just like they were for the earlier triple-helix results except that the partial charge is not atomic but summed up for the whole amino acid. Three views are provided of the front, back, and top of 1JS9. For the front view those amino acids that are visible on the surface and have the highest partial charges are pointed out in Figure 19. The back and top views are shown respectively in Figures 20 and 21. Note that in collagen there were covalent radii of atoms but here for 1JS9 we have the solvent excluded surface. Also, each of the three subunits are pointed out and one can see the long N-tail of the C subunit, and the the various amino acids that are more or less partially charged. The mixture of colors is similar to that in the triple-helix but just is not at the level of atomic detail. Also, in Figure **??** a tabular summary of all the amino acid partial charges is shown where the color in the table is the same as in the 1JS9 color-graded molecular images. This coloring of the table makes it easy to see the distribution of various partial charge amounts.

Partial Charges for 1JS9 Subunits A,B,C					40	GLY		0.036	0.07
Position	Amino Acid	A	B	C	41	LYN	-0.0674	-0.0895	-0.0619
1	ALA			-0.0658	42	ALA	-0.0614	-0.0074	-0.1005
2	ALA			-0.0151	43	ILE	-0.0192	0.0415	0.0583
3	ALA			0.0522	44	LYN	0.0264	-0.0668	0.0177
4	ALA			-0.1375	45	ALA	0.0123	0.0708	0.0056
5	ALA			0.1177	46	ILE	-0.0327	-0.0122	0.0355
6	ALA			-0.0451	47	ALA	-0.0212	-0.0561	-0.0462
7	ALA			-0.0176	48	GLY	-0.0024	0.1127	0.0115
8	ALA			-0.0039	49	TYR	0.083	-0.0309	0.0403
9	ALA			0.0487	50	SER	-0.0923	-0.1585	-0.0514
10	ALA			-0.0447	51	ILE	0.0591	0.1272	0.095
11	ALA			0.0183	52	SER	-0.2107	-0.1637	-0.1679
12	ALA			-0.0327	53	LYN	0.113	0.1255	0.1184
13	ALA			-0.0122	54	TRP	-0.0786	-0.0743	-0.1039
14	ALA			0.0384	55	GLH	0.0925	-0.001	0.1023
15	ALA			-0.0312	56	ALA	-0.0106	-0.0684	-0.027
16	ALA			0.0333	57	SER	0.0294	0.1093	0.0248
17	ALA			-0.0123	58	SER	-0.1179	-0.2281	-0.1277
18	ALA			0.0207	59	ASH	0.1918	0.2849	0.1435
19	ALA			-0.0141	60	ALA	-0.0832	-0.0681	-0.0358
20	ALA			-0.0202	61	ILE	-0.093	-0.001	-0.0567
21	ALA			0.0591	62	THR	0.1198	0.0486	0.1808
22	ALA			-0.0373	63	ALA	-0.0946	0.0584	-0.0014
23	ALA			-0.0308	64	LYN	-0.0121	-0.1396	-0.0723
24	ALA			0.0399	65	ALA	-0.131	-0.0386	-0.0689
25	ALA		-0.0555	0.0099	66	THR	0.1535	0.1059	0.1503
26	ARG		-0.0723	-0.024	67	ASN	-0.1174	-0.0653	-0.1644
27	VAL		0.0092	0.0026	68	ALA	0.0349	-0.0142	0.0859
28	GLN		-0.0182	-0.0665	69	MET	0.0116	0.0027	-0.0723
29	PRO		0.0309	0.0839	70	SER	-0.0825	-0.0301	-0.0488
30	VAL		-0.0232	-0.0557	71	ILE	0.0417	-0.0441	0.0276
31	ILE		0.0248	0.0035	72	THR	0.0431	0.1849	-0.0105
32	VAL		-0.0558	0.0209	73	LEU	-0.0798	-0.1106	0.0095
33	GLH		0.0464	-0.0334	74	PRO	0.011	0.0064	-0.0096
34	PRO		0.019	0.0504	75	HIS	0.0424	0.0175	-0.0033
35	LEU		-0.0287	0.0005	76	GLH	0.2029	0.1107	0.2297
36	ALA		-0.125	-0.1448	77	LEU	-0.0162	-0.0325	-0.1056
37	ALA		0.0957	0.0562	78	SER	-0.16	-0.0255	-0.0562
38	GLY		-0.0359	0.0477	79	SER	-0.0132	-0.0425	-0.0786
39	GLN		0.1191	0.044	80	GLH	0.0419	0.1012	0.2017

Figure 16. Part 1: Summary of the Brome Mosaic Virus amino acid partial charges with sequence number, amino acid name, and partial charge colored to match the 1JS9 color-graded molecular images.

81	LYN	0.132	0.1451	0.0902	122	ALA	-0.0445	-0.0451	-0.1081
82	ASN	-0.0963	-0.055	-0.1521	123	LEU	-0.0166	-0.0459	-0.0064
83	LYN	-0.0267	-0.0959	-0.0539	124	ALA	-0.0042	0.064	0.0738
84	GLH	0.0501	0.0565	0.1363	125	VAL	-0.0342	-0.0064	-0.0316
85	LEU	-0.0635	-0.2052	-0.1633	126	ALA	-0.0209	-0.0447	-0.0068
86	LYN	0.0827	0.2176	0.1	127	ASH	0.0238	-0.1059	0.0252
87	VAL	-0.0885	-0.1056	-0.0887	128	SER	0.1082	0.1352	0.072
88	GLY	0.1343	0.173	0.0999	129	SER	-0.1189	-0.208	-0.1078
89	ARG	-0.0928	-0.1159	-0.0944	130	LYN	0.0523	0.0186	0.0234
90	VAL	-0.0849	0.0058	0.036	131	GLH	-0.0494	-0.0864	-0.0642
91	LEU	0.0111	-0.0666	-0.036	132	VAL	0.0065	-0.0202	0.0319
92	LEU	0.033	0.0683	0.0143	133	VAL	0.0057	0.1185	-0.0091
93	TRP	0.0683	-0.0381	0.0268	134	ALA	-0.0653	-0.0831	-0.0413
94	LEU	0.1839	0.1855	0.1392	135	ALA	-0.0544	-0.0984	-0.0955
95	GLY	-0.0681	-0.1225	-0.1107	136	MET	0.0699	0.093	0.0545
96	LEU	0.1533	0.1363	0.1625	137	TYR	-0.1528	-0.1918	-0.1419
97	LEU	-0.1079	-0.0061	-0.0775	138	THR	0.1712	0.1979	0.1786
98	PRO	0.1209	-0.013	0.0957	139	ASH	-0.0977	-0.0412	-0.0505
99	SER	-0.0189	0.035	0.0075	140	ALA	-0.016	0.0526	-0.025
100	VAL	-0.006	-0.0648	-0.0574	141	PHE	-0.0057	-0.0504	-0.0042
101	ALA	0.0081	-0.0244	-0.0239	142	ARG	0.0451	-0.03	0.0569
102	GLY	-0.1427	0.0016	0.0646	143	GLY	0.0378	0.0074	-0.0136
103	ARG	0.1023	0.176	-0.0277	144	ALA	-0.0783	0.0058	-0.059
104	ILE	0.0198	0.031	-0.0696	145	THR	-0.0786	-0.1501	-0.2028
105	LYN	0.0303	-0.04	0.0683	146	LEU	0.2211	0.1233	0.1773
106	ALA	-0.105	-0.0247	-0.0554	147	GLY	-0.0495	-0.0007	0.0737
107	CYS	0.0435	-0.0398	-0.0121	148	ASH	-0.1201	-0.0697	-0.1038
108	VAL	0.0566	0.1167	0.0207	149	LEU	0.0288	-0.0708	0.0198
109	ALA	-0.0286	-0.0421	-0.0045	150	LEU	-0.0379	0.0654	-0.015
110	GLH	0.0459	0.0785	0.0228	151	ASN	-0.0174	-0.1188	0.0117
111	LYN	0.0723	-0.0248	0.1438	152	LEU	0.0191	0.1027	0.0055
112	GLN	-0.1433	-0.2006	-0.2593	153	GLN	0.0117	0.0313	-0.0832
113	ALA	0.0407	0.1008	0.1059	154	ILE	0.0311	0.0268	0.006
114	GLN	-0.0259	-0.1107	-0.065	155	TYR	-0.1753	-0.1601	-0.0826
115	ALA	0.059	0.1836	0.1477	156	LEU	0.232	0.1888	0.181
116	GLH	0.0554	0.0376	0.0706	157	TYR	-0.1037	-0.148	-0.0998
117	ALA	0.0009	0.0041	-0.0548	158	ALA	0.0738	0.0572	0.12
118	ALA	0.0261	0.01	0.0566	159	SER	-0.0136	-0.0477	-0.0196
119	PHE	-0.025	-0.0152	0.0541	160	GLH	0.1671	0.1868	0.0302
120	GLN	-0.0098	-0.077	-0.0694	161	ALA	-0.055	-0.0196	-0.0159
121	VAL	-0.0649	0.0162	-0.0002	162	VAL	-0.0561	-0.1065	-0.1006

Figure 17. Part 2: Summary of the Brome Mosaic Virus amino acid partial charges with sequence number, amino acid name, and partial charge colored to match the 1JS9 color-graded molecular images.

153	PRO	0.0267	-0.0088	-0.0012
154	ALA	0.1343	0.004	0.0891
155	TYR	-0.123	0.0194	-0.0275
156	ALA	-0.095	-0.1166	-0.1222
157	VAL	0.019	-0.0279	-0.0128
158	VAL	0.0146	-0.0071	0.0102
159	VAL	0.0116	0.0762	0.0121
170	HIS	0.0737	0.0871	0.0042
171	LEU	0.0513	0.1351	0.0425
172	GLH	-0.0354	-0.0353	-0.0478
173	VAL	-0.016	-0.0145	0.0122
174	GLH	0.037	0.2016	0.0792
175	HIS	-0.0718	-0.2352	-0.1046
176	VAL	0.1079	0.1706	0.175
177	ARG	0.0411	-0.0024	-0.0913
178	PRO	0.0826	0.1257	0.0699
179	THR	-0.0569	-0.0863	-0.0542
180	PHE	-0.0869	-0.0913	0.0038
181	ASH	0.1231	0.0763	0.0031
182	ASH	-0.0779	-0.0026	0.0008
183	PHE	0.0493	0.0032	0.0209
184	PHE	0.0155	0.0086	0.0006
185	THR	0.0149	0.0271	0.0103
186	PRO	-0.0636	-0.0559	-0.0766
187	VAL	-0.0565	0.0138	0.0217
188	TYR	0.1145	0.0261	0.0396
189	ARG	0.0838	0.0874	0.0926

Figure 18. Part 3: Summary of the Brome Mosaic Virus amino acid partial charges with sequence number, amino acid name, and partial charge colored to match the 1JS9 color-graded molecular images.

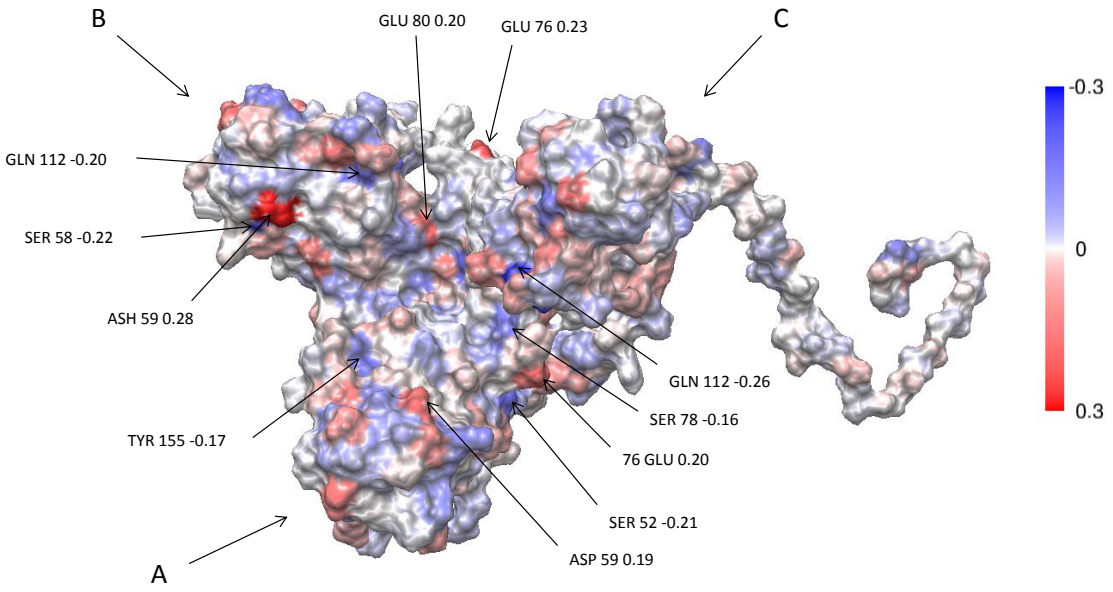

Figure 19. Front view of amino acid partial charge results for Brome Mosaic Virus subunits A, B and C with highest charges labeled and shown on the solvent excluded surface and color-graded such that blue is a gain of fractional electrons and red is a loss.

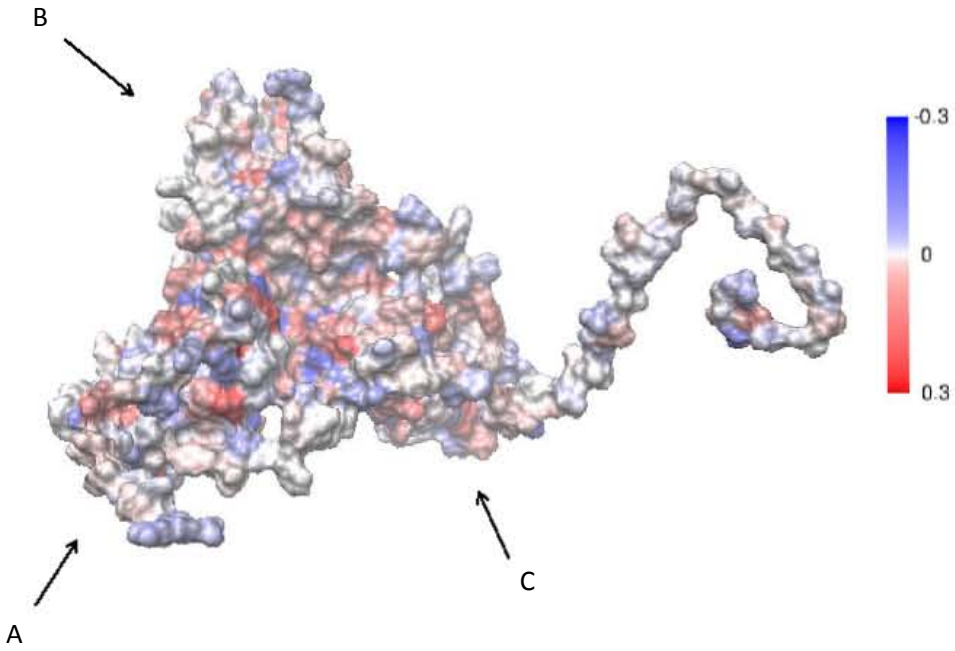

Figure 20. Back view of amino acid partial charge results for Brome Mosaic Virus subunits A, B and C shown on the solvent excluded surface and color-graded such that blue is a gain of fractional electrons and red is a loss.

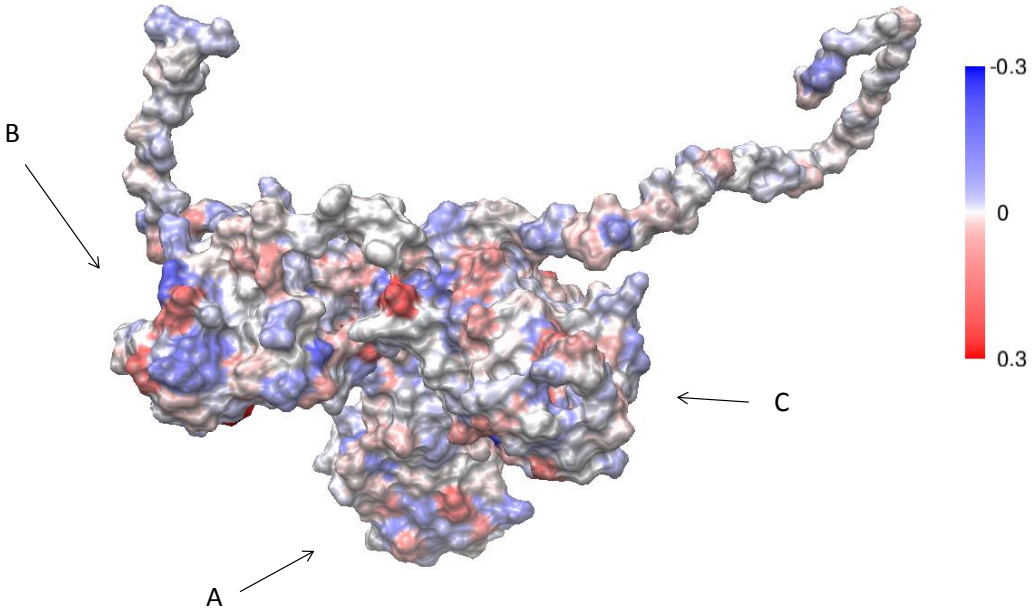

Figure 21. Top view of amino acid partial charge results for Brome Mosaic Virus subunits A, B and C shown on the solvent excluded surface and color-graded such that blue is a gain of fractional electrons and red is a loss.

6.2 Brome Mosaic Virus Most Partially Charged Amino Acids

The residues with the ten highest positive (red) or negative (blue) partial charges are shown in Table 2 with the amount of the charge and the amino acid name and sequence number alongside for each subunit A, B, and C. This is similar to the tabular data in Figure **??** but also included in parenthesis is whether the amino acids are on the surface (S), interface (I) or buried (B) in the core for the C subunit only, this was confirmed using the viperDB and through separate independent calculations.

Table 2. Table of highest partial charges in Brome Mosaic Virus subunits A, B and C

A			B			C		
Charge	Seq	Amino	Charge	Seq	Amino	Charge	Seq	Amino
−0.21	52	SER	−0.23	175	HIS	−0.25	112	GLN(S)
−0.17	155	TYR	−0.22	58	SER	−0.20	145	THR(I)
−0.16	78	SER	−0.20	129	SER	−0.16	145	THR(S)
−0.15	137	TYR	−0.20	85	LEU	−0.16	85	LEU(S)
−0.14	112	GLN	−0.20	112	GLN	−0.15	82	ASN(S)
+0.18	94	LEU	+0.18	156	LEU	+0.17	138	THR(I)
+0.19	59	ASP	+0.19	138	THR	+0.18	62	THR(S)
+0.20	76	GLU	+0.21	86	LYN	+0.18	156	LEU(B)
+0.22	146	LEU	+0.28	59	ASP	+0.20	80	GLU(I)
+0.23	156	LEU	+0.29	174	GLU	+0.22	76	GLU(S)

I should point out that the solvent excluded surface (SES) is also sometimes called the molecular surface and is related to the solvent accessible surface (SAS) in that it is also calculated with a ball rolling on the surface of the atoms. The ball is a water molecule but is really just an oxygen atom with hydrogens ignored or added in (which is negligble) to the radius of the oxygen atom. The SAS is taken at the center of the water molecule and the SES is where the water molecule contacts the van der Waals (vdW) surface of the protein atoms. The vdW surface would include sharp turning points but the SAS and SES would be smoothed. Really the SES uses the vdW radius of the protein atoms, whereas the SAS includes both the vdW radius of the water molecule and the protein atoms. So the SAS is a larger surface.

CHAPTER 7

FUTURE WORK

Partial charge is used in Molecular Dynamics (MD) and other theoretical methods to determine the nature of molecular interactions. This interaction can be divided into a number of components, one of which is the electrostatic interaction which acts over longer distances. However, in MD the partial charge used is not the Mulliken charge, but rather a partial charge fitted from the electrostatic potential. This is termed the Electro-Static Potential (ESP) method. There are a number of problems with these ESPs, for example, conformational dependence which has been accomondated for by the Restricted ESP (RESP) method. Again, these partial charges are not the actual partial charges of the atoms. Some have used this partial charge method with MD where periodically a new partial charge fit is introduced as the structure changes [52]. I should note that there are other terms in the simpliest force field that is used in MD than just partial charge: bond lengths, angles, torsions, van der Waals, repulsion due to orbital overlap, but really the electronic forces are the origin of all the major forces and are just accounted for in MD this way.

Our partial charge results are based on quantum mechanical calculations which can only give an accurate measure of the charge. Classical methods, of course, have been the frontrunner and of great help and can used in conjunction with QM methods. So, the primary question or future direction is how to interpret and use the partial charge results for determining the electrostatic interactions of biomolecules. This

also allows the QM result to be compared to a measureable phenomena of protein interaction, for example this can be helpful in understanding protein capsid assembly in viruses like 1JS9 into the BMV capsid.

Other future work includes extending the AAPM to include disulfide bonds that occur in some proteins (primarily extracellular proteins), for example Bovine Serum Albumin (BSA). However, whether this is necessary for the electronic structure has yet to be tested, meaning that it may not greatly improve the accuracy of the result. Also, testing the method on proteins with larger numbers of atoms, like BSA which hydrogenated would have approximately 18,000 atoms, will relate to future work. Addtionally, increasing the amount of potential reduction may be helpful as larger and larger structures are studied. Being able to determine the electronic properties from QM calculations for something as large as the entire BMV may seem remote since this can go to nearly a million atoms and with water even more. Currently, the Satelittle Tobacco Mosaic Virus (STMV) has been simulated classically. This has about 140,000 atoms in the protein shell, but about 900,000 water atoms [53]. Maybe in ten or twenty years this may be realizable.

APPENDIX A

INITIAL INPUT FILE FOR AAPM

The initial input file must be in protein data bank (pdb) file format. Basically, all atom records must start with ATOM, followed by an atom name such as 1HG2 or N. Sometimes the structure file will contain only coordinates and an element name. The coordinates must in xyz (cartesian) format, as opposed to fractional or spherical coordiantes, as well as, the xyz cooridinates must be shifted to be all positive (this is only currently necessary for my programs and not GULP or OLCAO and will be eventually changed). GULP is known as the General Lattice Utility Program (GULP) [54, 55, 56] and is used by OLCAO for the input structure file. The element name must also contain additional labels uniquely specifying the atom within an amino acid. Also, the atoms of the protein should be arranged in the amino acid sequence and provided with the sequence number and amino acid name, e.g. GLY 7, meaing the 7th amino acid is glycine. Details of the pdb file format can be found on at the Protein Databank itself [57, 58].

Once the input file is in pdb format some extra processing is necessary. If more than one molecule is in the file these need to be separated, that is, which ever molecule in the protein you want to examine you should put in a separate pdb file. A copy of this pdb file is made by replacing the atom names with a number code by the program atomcode-1js9 where the last three digits of the atom number replace the extra labels in the atom name, so 1HG2 could become 2H77, meaning it was atom 277 or 1277, etc. This numbering allows Chimera to properly add hydrogens to the amino acids, in other words, if the pdb atom name is there Chimera adds hydrogens as if the amino acid where charged (this may be fixed yet). This is done since the pdb file format is only 80 characters wide and to avoid having a second file indicating

the original identification number of the atom that must follow the pdb file.

So currently, only the last three digits of the actual original atom number is used to uniquely label the atom, and then the sequence number of the amino acid to which that atom belongs is then used to uniquely number or label the atom. This allows the programs to double check that we have the right atom. The program pre-secondpyprog selects only ATOM or HETATM records and converts them all to ATOM records, although this can be done with a text editor too.

These are the only two input files required. However, there are many intermediate files at this point necessary for the sequence of programs in the AAPM. Eventually this will all be integrated. There are two sets of output files that are the end products of the AAPM program sequence. The first output files are the gulp files of the individual amino acids models. If you have 100 amino acids in your protein you will have 100 models. These files can be transferred to the machine that OLCAO will be used on, and from there it is necessary to be familiar with OLCAO. This guide does not contain those instructions, but two example programs exist makeInFiles and makeOutFiles which automate the submission of the amino acid models and the collecting of the potential files into one directory for easy transfer back to the AAPM directory.

APPENDIX B

AAPM PROGRAM SEQUENCE

In Table 3, a list of the program sequence for the AAPM is given and afterwards a textual description of each program is given along with the input and output files used for each program. Note that the actual way the programs perform their operation is not included and would be gleaned from the actual code (not included in the thesis).

B.1 List of AAPM Program Sequence

Table 3. List of AAPM program sequence. Note that all programs end with the extension .py.

model-builder-adjacent-2
add-hydrogens-models-2-2-test-2
pdb-process-gulp-3
secondpyprog-4-python31-2
pdb-reorder-potential-ready-actual
triple-helix-potential-auto or 1js9a-potential-builder

B.2 Model-Builder-Adjacent-2

Two input files are necessary which are the two input files in pdb file format, one with numeric atom name labels that were mentioned earlier. Within the scripts the filein variable must contain the input file (you just type this in in the script itself). If the pdb-alphanumeric file is used a -2 must be typed in at three places as part of the output file names within the script itself. If the pdb-numeric file is used then a -1 must be typed. Remember the script atomcode-1js9 creates the pdb-numeric from the pdb-alphanumeric code. These files and which belong to -1 and -2 are shown at the top of the script. Again this will be improved in the future for ease of use.

Two types of output files are created: zzzzzz- + filetag + -1 or -2 + .pdb. Filetag consists of the first amino name, second, and third, followed by the molecule name, and then followed by the first, second and third amino acid sequence numbers, so: zzzzzz-GLY-GLY-GLY-A123-1.pdb is an example output file when using the pdb-numeric input file. If you have 100 aminos, you will have 200 zzzzzz files. These files must be placed in the Chimera python directory that you have picked. Eventually, better file handling will be included too.

The model builder program breaks the molecule, for example the collagen triple-helix, into the individual amino-acid submodels (of the model of the whole molecule) which also have the directly adjacent amino-acids attached as boundary conditions. Note, their is really no breaking though of bonds since the structure data is just position of atoms, and the submodels are just the rest of the structure ignored. So, the first amino acid would also have the next amino-acid included. The last amino-acid would also have the next-to-last amino-acid included. There are only

two amino-acids included in a terminal amino-acid submodel then.

For any non-terminal amino-acid the submodel would be three amino-acids, including the amino-acid of the submodel and the amino acid directly before and directly after the central amino-acid (the exception here would obviously be for a dipeptide that contains only two amino-acids). This notion all can be understood from the point of view that a tertiary or 3-dimensional protein structure still retains the original linear amino-acid sequence it had before folding, much like a ball of string. See the section in the introduction called About Proteins for more detail.

B.3 Add-Hydrogens-Models-2-2-Test-2

This program runs only in Chimera and is in Python2, accordingly. If all the scripts are converted to python 2 (an easy task actually for which a program can be written to perform) then the scripts can be all integrated. The program adds hydrogens to the ends of the amino acid model where bonds to adjacent amino acids were broken (again this means we just leave out the other atoms and we do not really break a bond). The input files are the output files of (this is also generally the case with all the programs) model-builder-adjacent-2 of the -1 and -2 varieties mentioned earlier. These, again, must be in the Chimera directory you are using.

The output files are the -1 file with an "a" at the beginning and end, so azzzzzz-GLY-GLY-GLY-A123-1.pdba is an example of this kind of output file. The azzzzzz file is a pdb file of the amino acid model with the hydrogens added where the bonds to adjacent amino acids have been broken, or in other words the adjacent amino-acids are replaced by a hydrogen atom. These files must be transferred back to the regular working python directory from Chimera's working python directory.

Note that there may be some problems with charged side-chain amino acids and the proper neutral hydrogenation. This can be checked when running the script by uncommenting out the "wait 1" print statement and wait commands near the very end of the file. The "zzzzzz" and such names are used to easily locate all the files and so that they are continuous and without other files in their midst. Also, no file handling is written in so one has to manually move and manage the files at this point.

In summary, this program adds hydrogens to the ends of the models broken apart in the model builder program and essentially replaces the adjacent amino-acid with a hydrogen atom. This is the least perturbing method we could use to create separate or individual amino-acid models from a protein where they are actually all linked together. This also means we ignore any hydrogen bonding to amino-acids in the vicinity. However, this method can be easily enough extended to include a ball of amino-acids around the central amino-acid of interest to include other bonds like disulfide bridges, hydrogen bonds, salt-bridges, or water. Also, including these extra amino-acids around the central amino-acid would increase accuracy but only marginally.

B.4 PDB-Process-Gulp-3 or PDB-Process-Gulp-2-Test

There are actually two scripts that perform the same operation in different ways and either can be used, but pdb-process-gulp-3 is suggested. The input file is the second pdb-numeric initial input file used at the very beginning and is for the whole protein. The other input files are the azzzzzz files that are to be placed the in python working directory from the Chimera python directory. The output files have an "A" at the beginning and a "-b" at the very end. So, Aazzzzzz-GLY-GLY-GLY-

A123-1.pda-b is an example of such a file. This program prepares the pdb file of the amino acid models to a modified pdb that can be changed into a gulp file as well as a reordered pdb for building the protein potential later.

Basically, the program removes everything but ATOM records. It also changes HETATM records to ATOM records. Also, the newly added hydrogen atoms are renamed properly and sorted to the end of the file.

B.5 Secondpyprog-4-Python31-2

This program creates the gulp file for each modified pdb file of the amino acid model for submission to OLCAO. The input files are the Aazzzzzz files of the preceding program in the sequence. Output files are preceded by a "B" and ended with a "-gulp", so an example would be, Baazzzzzz-GLY-GLY-GLY-A123-1.pdba-b-gulp. There should be a gulp for each amino-acid model, so if there are 100 amino acids in the protein there are 100 gulp files. Transfer these to the machine that OLCAO is run on. As mentioned earlier, there are two scripts to assist in running the gulps and collecting their output files for use with OLCAO.

The output files from OLCAO we need are the potential files called gs-scf-pot.dat in each gulp's directory created by the makeInFiles script. The potential output files have to be named with the input file followed with the "-gs-scf-pot.dat" tag, except, the ending is clipped to remove the "-pdba-b-gulp" part, so Baazzzzzz-GLY-GLY-GLY-A123-1-gs_scf-pot.dat is an examle of the potential file naming convention. These potential files must be named this way for potential builder program to make the protein potential and for them to be have different filenames and not be confused. The second script makeOutFiles collects the differently named potential files into one

directory for easy movement to another computer or directory.

B.6 PDB-Reorder-Potential-Ready-Actual

This reorders the pdb so that the potential builder program can align the records within the potential output files ("-gs-scf-pot.dat" files) to particular atoms listed in the reordered pdb, otherwise the wrong atom will be matched to the wrong potential since OLCAO puts output into an element ordering rather than amino acid ordering as we have used. The input files for this are the processed pdbs, for example, Aazzzzzz-GLY-GLY-GLY-A123-1.pda-b. The output files are the same with "-reorder" added, so Aazzzzzz-GLY-GLY-GLY-A123-1.pdb-b-reorder is an example of such a file. Note this script is not necessary to create the gulps for submission to OLCAO, but is necessary to build the protein potential which uses the potential output files end labelled with "-gs-scf-pot.dat".

B.7 Potential Builder Programs

Currently, there are two scripts to build the protein potential. One for the collagen triple-helix and one for 1js9. However, the difference between the scripts is that triple-helix one has only two digits for the amino-acid sequence number and so we use a three-digit atom name label of numbers to identify a particular atom. For the 1js9 program since it has a four-digit atom number, we use both a three digit amino-acid sequence number and a three-digit atom name label to determine the particular atom. This all stems from the 80 character wide limitation of the pdb file format and the desire not to split the pdb file into multiple associated files to know which atom was which.

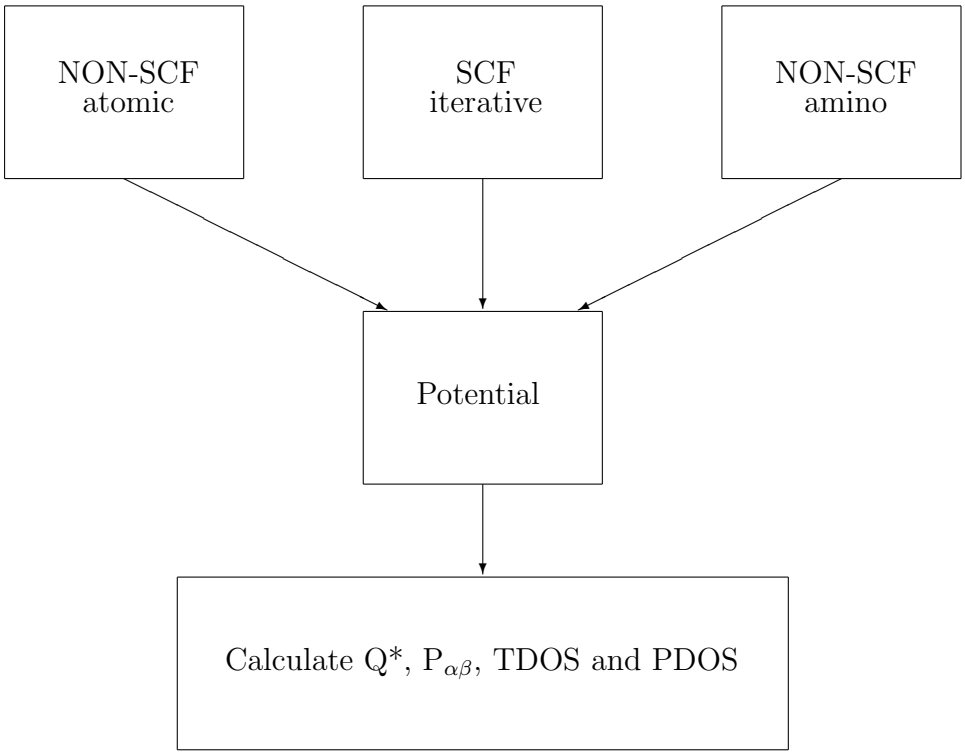

Figure 22. Any of three methods to construct a protein potential (atomic, scf, or aapm) can be used calculate electronic structure properties.

These programs are not yet unified. So use the triple-helix one for proteins of no more than 999 atoms and 99 amino acids, and the 1js9 for larger proteins. Also in Figure 22 we can see a flowchart diagramming how a protein potential is constructed in a general way. Either we use a potential directly from a scf interactive cycle (middle top) or substitute one of two types of database for the potentials. On the left top in the chart we see the atomic database option and on the right top the amino-acid database option.

The main point is that any of the protein potentials constructed from one of the three ways can all be used to calculate the properties of effective charge, bond order and density of states. The overall simplfied scheme is shown in Figure 23 from the initial models to the final calculation of the electronic structure properties (note that the bottom two boxes of Figure 22 correspond to the bottom two boxes of Figure 23 and the far right Non-scf amino box of Figure 22 corresponds to the second box from top in Figure 23).

First, the initial models must have no local charge centers since the DFT calculation will take too long to converge to be computationally feasible and it has been tested to converge. Also, the initial models do not usually come with hydrogen atoms since these are not yet capable of being resolved accurately in X-ray crystallographic methods and so must be hydrogenated (we use UCSF Chimera but there are other packages). Second, the amino acid database is constructed from scf calculations for the electronic potential on amino acids with adjacent amino acids in the sequence attached as a boundary condition.

In Figure 24, we can see two amino acid triplets excised from the $\alpha 2$(I) chain

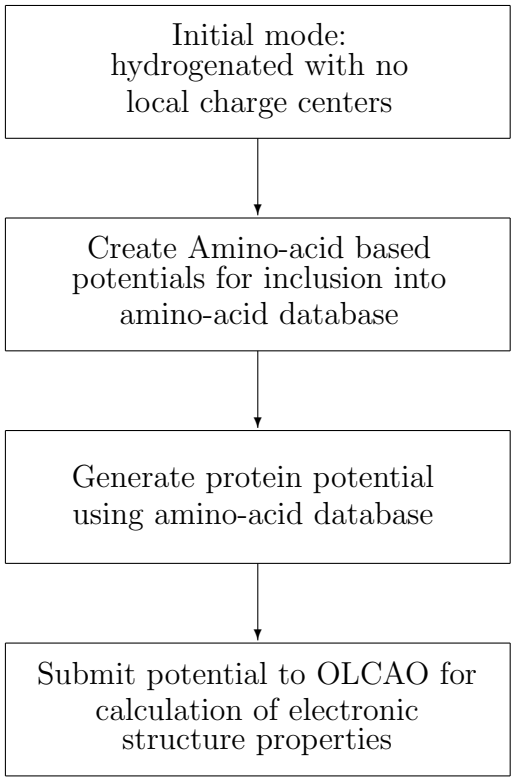

Figure 23. The flowchart shows the seuqence of operations performed by various programs to implement the simplfied scheme.

shown in Figure 6. The middle proline rings are the amino acid for which we are calculating a scf potential for inclusion to the amino acid database. The other amino acids in each triplet are the boundary conditions. Note, that for the terminal amino acids there would only be two amino acids in the excised submodel. Third, the protein potential is generated. And last, the complete protein potential along with the structure file is submitted to OLCAO for calculating the electronic structure properties.

B.8 Triple-Helix-Potential-Auto

The input file is a pdb file in the order of your gulp file which for what we have been discussing so far would be the same order. A program called secondpyprog-4-python-31, which is a version of the earlier script with the "-2" at the end, is just for doing a single file rather than a group of files with some file ending. You can use this program to convert your pdb into a gulp file for the whole protein structure file. Also, the pdb used is the pdb-numeric form as well as that the pdb must be in element order. The script trimmer-reorder-NCOHS element orders a pdb file.

The other input files used here are the potential files and the reordered pdb files. These are of the form Bazzzzzz-GLY-GLY-GLY-A123-1-gs_scf-pot.dat for the potential file and Aazzzzzz-GLY-GLY-GLY-A123-1.pdba-b-reorder for the reordered pdb. All these files must be in the python directory: pdb input file, reordered pdbs and potential files. The output file is a single potential file. This potential file and the gulp file are the input files for OLCAO which will generate a number of output files containing the results of the desired OLCAO calculations.

As mentioned earlier, this collagen protein potential builder uses just a three-

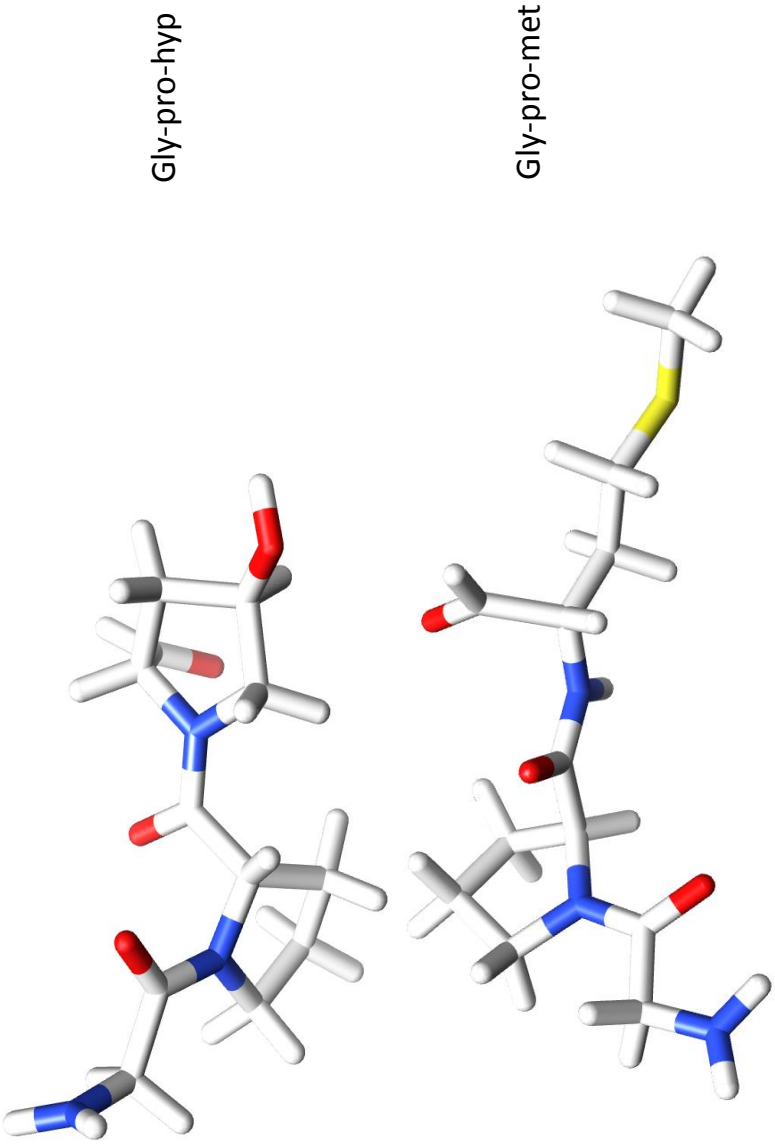

Figure 24. The fourth and first triplets of Figure 6 are shown. Note that Pro is the middle amino acid for which the potential is to be calculated and the other amino acids are the boundary conditions which have their own peptide bonded adjacent amino acids replaced by a hydrogen atom, except the Gly of Gly-pro-met since it is the N-terminal amino acid.

digit atom name to uniquely describe each atom which is sufficient. Unfortunately, if you need to use a four-digit name you cannot easily label the atom with just three digits. You can do this by using an alphanumeric code so that each digit place has more than ten possible numerals but this creates a host of other conversion issues. Again, this is all done since the pdb file format is 80 characers wide and only allows three-digits to name an atom, the fourth digit of the atom name being the element name.

B.9 1JS9a-Potential-Builder-2

As above, you must have a pdb in gulp order as well as all the reordered pdbs and potential files. This program otherwise operates in the above manner generating a potential file for the whole protein to submit to OLCAO along with the gulp file. As already mentioned, the 1JS9 protein potential building program uses a three-digit atom name and the amino-acid sequence number to uniquely identify atoms. This is done since the pdb file format is limited to 80 characters wide and only allows for three-digits to name an atom (not including the element name which is the fourth digit of the name).

Fortunately, even large proteins contain fewer than 10,000 atoms, as well as that the currently used box size in OLCAO does not extend well to include protein complexes containing 10,000 or more atoms. So the 1JS9 potential builder should work well for any protein we may study. I should note that the entire brome mosaic virus protein capsid (not including RNA which we are not studying in this thesis) contains about 400,000 atoms. Eventually, protein complexes of this size may be studied, say in ten years: so never say never!

B.10 Reduce Hydrogen Potential Types Programs

Currently, OLCAO accepts only up to 999 (or three-digits) element potential types (the total number of potential types for all the atoms is also limited to 5000 but this limit can be changed very easily). The change to the 999 limit per element has yet to be attempted and will likely be done only for the new OLCAO package being currently developed. So, if we have 1200 hydrogens in a protein then 201 of them must share potentials with other hydrogens although the number of hydrogens sharing a potential can be two or more. Note this reduce potential method is an extension to the AAPM method to allow for calculation of proteins containing more than 999 hydrogens. Additionally this is satisfying in that one would not think so many potential types, of hydrogen especially, should be necessary.

There are two methods to reduce the number of potential types. One method, avoids using a unique potential for each atom in calculating the scf result for the amino-acid models. The other method explores the data of the protein potential to match potentials. In future work we will attempt to use structurally related averaged potentials to extend the method to larger systems while still keeping the number of potential types low enough. So, one method reduces the types (the first) ahead of obtaining the amino-acid potentials and the other, second method reduces the potential types after obtaining the amino-acid potentials. Note, both methods can be used together or separately. The second method will only be used here.

Hydrogens have the least complex potential and are more numerous and for these two reasons are easier to match. Also, the hydrogens do not appear in the original structure file and are added by other software such as Chimera or other

packages and are ultimately calculated results (not experimentally derived) anyway. It is also more difficult to control the exact number of types reduced with the first method. Basically, the second method compares the potentials of two hydrogen atoms collectively and assigns a number to it. If this number falls below a certain cutoff then the hydrogens are matched and only one potential is necessary to describe them. A list of hydrogen atoms is made that fall below the cutoff.

The next step is to map what atoms map to what atoms. This can be approached in two ways: taking the closest matching hydrogens first or taking the hydrogen with the most possible hydrogens mapped to it first. The first approach minimizes the difference between potentials being matched and the second approach maximizes the number of hydrogens mapped or reduced in potential types. Sometimes these approaches can give the same result depending on the system and the amount of reduction. The most mapped approach is what has been used though since it generally in the two proteins that have been studied produces a result as good as the close reduction or better, that is, never worse.

The details of matching the hydrogens follows. Hydrogen atoms have six terms in their potential each with a coefficient that is adjusted in the scf iterative cycle. Between two hydrogen atoms, the coefficients for the first term of each hydrogen's potential are compared by taking the absolute value of their difference or absolute deviation. Note, we do not use root mean square since we are not taking derivatives or integrating or otherwise mathematically manipulating these results. This is done for the other five potentials. These potentials are summed and checked to be below a cutoff.

Alternatively, one could control how the potentials are grouped: each one below a cutoff or the first two, the second two, and last two each below a cutoff. Also, the cutoff can be set differently for each group. However, for hydrogen reduction the same cutoff is used and they are just all grouped together to allow for some variance. In carbon there are 16 coefficients to match so they are grouped into fours with the same cutoff, however we do not use carbon reduction in this thesis for the size of systems studied. This measuring of the differences between potentials may be one place where we could improve the reduction method.

The reduction scheme may well be improved by considering better the potential matching and also how the atoms are chosen to reduce or map to one another. Ultimately, this is a data exploration technique on a discrete data set and the best answer is not necessarily really better than a good answer since it will take too long to find the best to be feasible and the good answer gives an accurate estimate. In other words, it is a common mistake in the data exploration of discrete data sets to think one needs the best possible answer (private communication, Larry Eifler). Also, the number of potentials which can be reduced is probably limited and saves only a modest amount of computational time in comparison to skipping out the iterative scf calculation of the whole protein.

So, this reduction technique is best used to extend the size of system that can be studied only slightly yet. Currently, we have a 999 hydrogen potential limit, so then a protein of 2000 atoms is the largest that can be studied without any reduction of the potentials. Using the hydrogen reduction the size can be extended to about 3000 atoms without losing too much accuracy of the results. With possible improvements

the range could be extended up to perhaps 4000 atoms or a doubling of the size of the system. Additionally, we may average potentials (as already mentioned) which are structurally similar to increase the size of the system that can be studied. But this is only at the initial idea phase and has only been rudimentarily tested, although this averaging approach shows promise for doing the entire 1JS9 protein.

REFERENCES

1. P. Echenique-Robba, *Shut up and let me think! Or why you should work on the foundations of quantum mechanics as much as you please*, arXiv preprint arXiv:1308.5619 (2013).
2. R. Maul, *Electronic Excitations of Glycine, Alanine, and Cysteine Conformers from First-Principles Calculations*, Journal of Physical Chemistry A 111, 4370–4377 (2007).
3. Anze Losdorfer Bozic, Antonio Siber, Rudolf Podgornik. *How simple can a model of an empty viral capsid be? Charge distributions in viral capsids*, Journal of Biological Physics 38, 657–671 (2012).
4. Irina Gitlin, Jeffrey D. Carbeck, George M. Whitesides. *Why are Proteins Charged? Networks of Charge-Charge Interactions in Proteins Measured by Charge Ladders and Capillary Electrophoresis*, Angewandte Chemie 45, 3022–3060 (2006).
5. A Gautieri, MI Pate, S Vesentini, A Redaelli, MJ Buehler. *Hydration and distance dependence of intermolecular shearing between collagen molecules in a model microfibril*, Journal of Biomechanics 45, 2079–2083 (2012).
6. NH Rhys, L Dougan. *The emerging role of hydrogen bond interactions in polyglutamine structure, stability and association*, Soft Matter 9, 2359–2364 (2013).
7. A Gautieri, S Vesentini, A Redaelli, et al. *Intermolecular slip mechanism in tropocollagen nanofibrils*, International Journal of Materials Research 7, 921–925

(2009).

8. Z Qin, MJ Buehler. *Molecular Dynamics Simulation of the α-Helix to β-Sheet Transition in Coiled Protein Filaments: Evidence for a Critical Filament Length Scale*, Physical Review Letters 104, 198304–1–4 (2010).
9. CA Grant, MA Phillips, NH Thomson. *Dynamic mechanical analysis of collagen fibrils at the nanoscale*, Journal of the Mechanical Behavior of Biomedical Materials 5, 165–160 (2012).
10. L Ouyang, L Randaccio, P Rulis, EZ Kurmaez, A Moewes, WY Ching. *Electronic structure and bonding in Vitamin B12 cyanocobalamin*, Journal of Molecular Structure: THEOCHEM 622, 221–227 (2003).
11. L Ouyang, P Rulis, WY Ching, G Nadin, L Randaccio. *Electronic Structure and Bonding in Adenosylcobalamin*, Inorganic Chemistry 43, 1235–1241 (2004).
12. JB MacNaughton, A Moewes, JS Lee, SD Wetig, HB Kraatz, L Ouyang, WY Ching, EZ Kurrmaez. *Dependence of the DNA Electronic structure on environmental and structural variations*, Journal of Physical Chemistry 110, 15742–15748 (2006).
13. L Liang, P Rulis, B Kahr, WY Ching. *Theoretical study of the large linear dichroism of herapathite*, Physical Review B 80, 235132 (2009).
14. WY Ching, P Rulis, A Misra. *Ab initio elastic properties and tensile strength in hydroxyapatite crystal*, Acta Biomaterialia 5, 3067–3075 (2009).
15. L Liang, P Rulis, WY Ching. *Mechanical properties, electronic structure and bonding of α- and β-tri-calcium phosphates with surface characterization*, Acta Biomaterilia 6, 3763–3771 (2010).

16. L Liang, P Rulis, L Ouyang, WY Ching. *Ab initio investigation of hydrogen bonding and network structure in a supercooled model of water*, Physical Review B 83, 024201 (2011).

17. Wai-Yim Ching, Paul Rulis. *Electronic Structure Merthods for Complex Materials: The Orthogonalized Linear Combination of Atomic Orbitals*, Oxford University Press, 2012.

18. WY Ching. *Theoretical Studies of Electronic Properties of Ceramic Materials*, Journal of the American Ceramic Society 71, 3135–3160 (1990).

19. B. Alberts. *Molecular Biology of the Cell*, Garland, New York, 4th edition, 2002.

20. Alexei V. Finkelstein, Oleg Ptitsyn. *Protein physics: a course of lectures*, Academic Press, 2002.

21. Robert W. Lucas, Steven B. Larson, Alexander McPherson. *The crystallographic structure of brome mosaic virus*, Journal of molecular biology 317, 95–108 (2002).

22. FS Collins, et al. *Finishing the euchromatic sequence of the human genome*, Nature, 4311, 931–945 (2004).

23. Michele Clamp, et al. *Distinguishing protein-coding and noncoding genes in the human genome*, Proceedings of the National Academy of Sciences 104, 19428–19433 (2007).

24. A. Bock, et al. *Selenocysteine: the 21st amino acid*, Molecular microbiology 5, 515–520 (1991).

25. Gayathri Srinivasan, Carey M. James, Joseph A. Krzycki. *Pyrrolysine encoded by UAG in Archaea: charging of a UAG-decoding specialized tRNA*, Science 296, 1459–1462 (2002).

26. Q Wang, AR Parrish, L Wang. *Expanding the genetic code for biological studies*, Chemistry and Biology 16, 323–336 (2009).

27. Thomas R Rizzo, et al. *The electronic spectrum of the amino acid tryptophan in the gas phase*, The Journal of chemical physics 84, 2534–2541 (1986).

28. Max S. Dunn, T.W. Brophy. *Decomposition Points of the Amino Acids*, The Journal of Biological Chemistry 99, 221–229 (1932).

29. U. Meierhenrich, et. al. *Circular Dichroism of Amino Acids in the Vacuum-Ultraviolet Region*, Angewandte Chemie International Edition 49, 7799–7802 (2010).

30. Joel Janin, Susan Miller, Cyrus Chothia. *Surface, Subunit Interfaces and Interior of Oligomeric Proteins*, Journal of Molecular Biology 204, 155–164 (1988).

31. R. French, et al. *Long Range interactions in nanoscale science*, Reviews of Modern Physics 82, 1887.

32. V.A. Parsegian. *Van der Waals forces: a handbook for biologists, chemists, engineers, and physicists*, Cambridge University Press, 2006.

33. Jose Juan Lopez-Garcia, Jose Horno, Constantino Grosse. *Poisson-Boltzmann Description of the Electrical Double Layer including Ion Size Effects*, Langmuir 27, 13970–13974 (2011).

34. Dan Ben-Yaakov, David Andelman, Rudi Podgornick. *Dielectric decrement as a source of ion-specific effects*, The Journal of chemical physics 134, 074705–074705 (2011).

35. Dennis A. Benson, Ilene Karsch-Mizrachi, David L. Wheeler, et al. *GenBank*, Nucleic Acids Research, 33 (Database Issue), D34–D38 (2005).

36. Dale L. Bodian, et al. *Molecular Dynamics Simulations of the Full Triple Helical Region of Collagen Type I Provide An Atomic Scale Vie of the Protein's Regional Heterogeneity*, Pacific Symposium on Biocomputing, 193–204 (2011).

37. H Kuibaniemi, G Tormp, D Prockop. *Mutations in collagen genes: causes of rare and some common diseases in humans*, The FASEB Journal 5, 2052–2060 (1991).

38. K Okuyama, T Kawaguchi, M Shimura, K Noguchi, K Mizuno, HP Bachinger. *Crystal structure of the collagen model peptide (Pro-Pro-Gly)4-Hyp-Asp-Gly-(Pro-Pro-Gly)4 at 1.0 A resolution*, Bioploymers 99, 436–437 (2013).

39. Barabara Brodsky, John AM Ramshaw. *The collagen triple-helix structure*, Matrix Biology 15, 545–554 (1997).

40. Jakob Bohr, Kasper Olsen. *The close-packed triple helix as a possible new structural motif for collagen*, Theoterical Chemistry Accounts 130, 1095–1103 (2011).

41. DLD T Caspar, Aaron Klug. *Physical principles in the construction of regular viruses*, Cold Spring Harbor Symposia on Quantitative Biology 27, 1–24 (1962).

42. Anton Arkhipov, Peter L. Freddolino, Klaus Schulten. *Stability and Dynamics of Virus Capsids Descibed by Coarse-Grain Modeling*, Structure 14, 1767–1777 (2006).

43. Nicole F. Steinmetz, et al. *Structure-based engineering of an icosahedral virus for nanomedicine and nanotechnology*, Springer Berlin Heidelberg, 2009.

44. Robert G. Parr, Weitao Yang. *Density-functional theory of atoms and molecules*, Oxford University Press, 1989.

45. Jan K. Rainey, M. Cynthia Goh. *An interactive triple-helical builder*, Bioinfor-

matics 20, 2458–2459 (2004).

46. Jan K. Rainey, M. Cynthia Goh. *A statistically derived parameterization for the collagen triple-helix*, Protein Science 11, 2748–2754 (2002).

47. Jan K. Rainey, M. Cynthia Goh. *Statistically Based Reduced Representation of Amino Acid Side Chains*, Journal of chemical information and computer sciences 44, 817–830 (2004).

48. J. Eifler, P. Rulis, R. Tai, W. Y. Ching. *Computational Study of a Heterostructural Model of Type I Collagen and Implementation of an Amino Acid Potential Method Applicable to Large Proteins* Polymers 6, 491–514 (2014).

49. Masarapu Hema, et al. *Effects of amino-acid substitutions in the brome mosaic virus capsid protein on RNA encapsidation*, Molecular plant-microbe interactions 23, 1433–1447 (2010).

50. EF Petersen, TD Goddard, CC Huang, GS Couch, DM Greenblatt, EC Meng, TE Ferrin. *UCSF Chimera–a visualization system for exploratory research and analysis*, Journal of Computational Chemistry 25, 1605–1612 (2004).

51. David S. Cerutti, Julia E. Rice, William C. Swope, David A. Case. *Derivation of Fixed Partial Charges for Amino Acids Accommodating a Specific Water Model and Implicit Polarization*, The Journal of Physical Chemistry 117, 2328–2338 (2013).

52. S. Kimura, R. Rajamani, D.R. Langley. *Communication: Quantum polarized fluctuating charge model: A practical method to include ligand polarizability in biomolecular simulations* The Journal of Chemical Physics 135, 23101 (2001).

53. P. Freddolino, et al. *Molecular dynamics simulations of the complete satellite*

tobacco mosaic virus Structure 14, 437-449 (2006).

54. JD Gale. *GULP - a computer program for the symmetry adapted simulation of solids*, JCS Faraday Transactions 93, 629–637 (1997).
55. JD Gale. *Empirical potential derivation for ionic materials*, Philosophical Magazine B 73, 3–19 (1996).
56. JD Gale. *GULP (General Utility Lattice Program*, Royal Institution, London (1992).
57. www.rcsb.org
58. Helen M. Berman, et al. *The protein data bank*, Nucleic acids research 28, 235–242 (2000).

VITA

Jay Quinson Eifler was born on October 2, 1970, in Baton Rouge, Louisiana. In 1989 he graduated from William Chrisman High School in Independence, Missouri. He received a B.S. degree from the University of Missouri-Columbia in 1994.